

Matrix Estimation with Latent Permutations

by

Cheng Mao

B.S., M.A., University of California, Los Angeles (2013)

Submitted to the Department of Mathematics
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2018

© Massachusetts Institute of Technology 2018. All rights reserved.

Signature redacted

Author

Department of Mathematics

April 19, 2018


Signature redacted

Certified by


Philippe Rigollet

Associate Professor of Mathematics

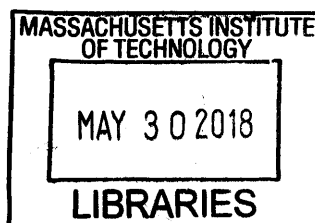
Thesis Supervisor


Signature redacted

Accepted by

 William P. Minicozzi II

Chairman, Department Committee on Graduate Theses



ARCHIVES

Matrix Estimation with Latent Permutations

by
Cheng Mao

Submitted to the Department of Mathematics
on April 19, 2018, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

Motivated by various applications such as seriation, network alignment and ranking from pairwise comparisons, we study the problem of estimating a structured matrix with rows and columns shuffled by latent permutations, given noisy and incomplete observations of its entries. This problem is at the intersection of shape constrained estimation which has a long history in statistics, and latent permutation learning which has driven a recent surge of interest in the machine learning community. Shape constraints on matrices, such as monotonicity and smoothness, are generally more robust than parametric assumptions, and often allow for adaptive and efficient estimation in high dimensions. On the other hand, latent permutations underlie many graph matching and assignment problems that are computationally intractable in the worst-case and not yet well-understood in the average-case. Therefore, it is of significant interest to both develop statistical approaches and design efficient algorithms for problems where shape constraints meet latent permutations.

In this work, we consider three specific models: the statistical seriation model, the noisy sorting model and the strong stochastic transitivity model. First, statistical seriation consists in permuting the rows of a noisy matrix in such a way that all its columns are approximately monotone, or more generally, unimodal. We study both global and adaptive rates of estimation for this model, and introduce an efficient algorithm for the monotone case.

Next, we move on to ranking from pairwise comparisons, and consider the noisy sorting model. We establish the minimax rates of estimation for noisy sorting, and propose a near-linear time multistage algorithm that achieves a near-optimal rate.

Finally, we study the strong stochastic transitivity model that significantly generalizes the noisy sorting model for estimation from pairwise comparisons. Our efficient algorithm achieves the rate $\tilde{O}(n^{-3/4})$, narrowing a gap between the statistically optimal rate $\tilde{\Theta}(n^{-1})$ and the state-of-the-art computationally efficient rate $\tilde{O}(n^{-1/2})$. In addition, we consider the scenario where a fixed subset of pairwise comparisons is given. A dichotomy exists between the worst-case design, where consistent estimation is often impossible, and an average-case design, where we show that the optimal rate of estimation depends on the degree sequence of the comparison topology.

Thesis Supervisor: Philippe Rigollet
Title: Associate Professor of Mathematics

Acknowledgments

First and foremost, I would like to thank my advisor Philippe Rigollet for his constant support, encouragement and patience. I could not have enjoyed my time in graduate school and become a research statistician without all the brilliant ideas and resources he has provided. Philippe's sharp insight and sagacious views both in and outside research have deeply influenced me.

My gratitude also goes to Martin Wainwright, who was generous with time and ideas during my visit at UC Berkeley. Collaboration with BLISS members has become a significant part of my work since then. I thank Ankur Moitra and Sasha Rakhlin for being on my thesis committee and offering numerous insights. I thank Larry Guth and Scott Sheffield for supervising me in my first year and a half of graduate study. Scott's beautiful vision in mathematics made working in his group a memorable experience for me, although I did not become a probabilist eventually. I also thank Kefeng Liu and Ciprian Manolescu for their support in my undergraduate years at UCLA, without which I could not have been able to start my journey in academia.

Collaboration with fellow students and junior researchers has been an honorable and delightful part of my graduate study. I greatly benefited from working with Xin Sun, Ewain Gwynne and Nina Holden in the first two years. Xin's wisdom has guided me through my life as a graduate student. Jan-Christian Hütter, Jonathan Weed and Nilin Abrahamsen have been the best peers in statistics I could have hoped for. It has always been enlightening to chat with them about anything. Working and having lunch with Victor-Emmanuel Brunel and Nicolas Flammarion was always a wonderful experience that I cherish. Ashwin Pananjady, Vidya Muthukumar and Nihar Shah made me feel at home while visiting UC Berkeley. Collaborating with Ashwin has been more productive and rewarding than I could have imagined. It is always pleasurable to catch up with Nick Strehlke and Tudor Pădurariu who both came to MIT after we spent wonderful years at UCLA together. Friendship with Nick has been full of inspiration.

I would like to thank all the brilliant minds above and also Florent Bekerman, Xue Chen, Yash Deshpande, Thao Do, Sam Elder, Justin Eldridge, Chenjie Fan, Teng Fei, Qiang Guang, James Hirst, Dax Koh, Fu Li, Haihao Lu, Marco Avella Medina, Amelia Perry, Andrej Risteski, Elina Robeva, Geoffrey Schiebinger, Yair Shenfeld, Irène Waldspurger, Xinan Wang, Fan Wei, Alex Wein, Ben Yang, Joy Yang, Yi Zeng and Tianyou Zhou, as well as many I forgot to name. I have been so fortunate to be surrounded by you.

The MIT Mathematics Department has provided a warm and productive environment for me throughout the years. I thank Barbara Peskin, Michele Gallarelli and all the staff in math and IDSS who have generously offered me help and made my graduate student life smoother.

Last but certainly not least, I am grateful to my friends Zirui Wang and Ruikun Hong who always stand by me. It was an incredible fifteen years. I am deeply indebted to my parents and wife for their everlasting love and support. None of this would have been possible without what they have given me.

Contents

1	Introduction	9
2	Optimal Rates of Statistical Seriation	13
2.1	Problem setup and related work	15
2.1.1	The seriation model	15
2.1.2	Related work	17
2.2	Main results	19
2.2.1	Adaptive oracle inequalities	19
2.2.2	Global oracle inequalities	20
2.2.3	Minimax lower bounds	22
2.3	Further results in the monotone case	23
2.3.1	RankScore: An efficient estimator and its performance	23
2.3.2	Simulations	26
2.4	Unimodal regression	29
2.5	Proofs	30
2.5.1	Proof of the upper bounds	30
2.5.2	Metric entropy	35
2.5.3	Proof of the lower bounds	36
2.5.4	Matrices with increasing columns	38
2.6	Discussion	39
2.7	Additional proofs	40
2.7.1	Proof of Proposition 2.2.7	40
2.7.2	Proof of Lemma 2.5.4	41
2.7.3	Proofs of lemmas in Section 2.5.2	42
2.7.4	Proofs of lemmas in Section 2.5.3	47
2.7.5	Proofs of lemmas in Section 2.5.4	50
2.7.6	Proof of Corollary 2.4.1	52
3	Minimax Rates and Efficient Algorithms for Noisy Sorting	53
3.1	Problem setup	56
3.1.1	Sampling models	56
3.1.2	Measures of performance	57
3.2	Main results	58
3.2.1	Minimax rates of noisy sorting	58
3.2.2	Efficient multistage sorting	59

3.3	Simulations	62
3.4	Discussion and open problems	63
3.5	The symmetric group and inversions	64
3.6	Proofs of the main results	66
3.6.1	Proof of Theorem 3.2.1	67
3.6.2	Proof of Theorem 3.2.2	73
4	Faster Rates for Permutation-based Models in Polynomial Time	77
4.1	Background and problem setup	80
4.1.1	Matrix models	80
4.1.2	Observation model	81
4.2	Main results	82
4.2.1	Statistical limits of estimation	82
4.2.2	Efficient algorithms	83
4.3	Applications	86
4.4	Proofs	87
4.4.1	Some preliminary lemmas	87
4.4.2	Proof of Theorem 4.2.1	89
4.4.3	Proof of Proposition 4.2.2	92
4.4.4	Proof of Theorem 4.2.3	92
4.4.5	Proof of Lemma 4.4.1	100
4.5	Discussion	104
4.6	Appendix: Poissonization reduction	105
4.7	Appendix: Truncation preserves sub-Gaussianity	106
5	Worst-case v.s. Average-case Design for Estimation from Fixed Pairwise Comparisons	109
5.1	Background and problem setup	112
5.1.1	Pairwise comparison models	112
5.1.2	Partial observation models	113
5.2	Main results	115
5.2.1	Worst-case design: minimax bounds	115
5.2.2	Average-case design: noisy sorting matrix estimation	116
5.2.3	Two random designs: SST matrix estimation	118
5.3	Dependence on graph topologies	120
5.4	Proofs	123
5.4.1	Proof of Theorem 5.2.1	123
5.4.2	Some useful lemmas for average-case proofs	125
5.4.3	Proof of Theorem 5.2.2	128
5.4.4	Proof of Theorem 5.2.3	133
5.4.5	Proof of Theorem 5.2.4	140
5.5	Discussion	147
5.6	Appendix: Bounds on the minimax denoising error	148
5.6.1	Proof of Theorem 5.6.1	148

Chapter 1

Introduction

The problem of matrix estimation and matrix completion has long been studied in mathematical statistics and machine learning [Faz02, AM07, RV07, CR09, KMO10, MHT10, RT11, Cha15]. Structural constraints, such as low-rankness [CP11, Kol11, KLT11, NW11, JNS13] and smoothness [GLZ15, HR16, KTV17, CGS18], allow for statistical inference of a high-dimensional matrix given noisy and incomplete observations of its entries. On the other hand, the past decade has witnessed a burgeoning literature on the problem of graph estimation and network reconstruction from noisy data [ACN08, BC09, GZFA10, Lov12, ACC13, GLZ15, KTV17, CLR17b]. Central tasks include, for example, clustering [AS15, LR15, ABH, BRS16, HWX16, MPW16, CLR17a], ranking [FV93, DKNS01, Alo06, Liu09, NOS12, RA14, SBB⁺16] and alignment [CFSV04, LJTKJ06, SXB08, ZBV09, Bur13, FQRM⁺16, ESDS16]. The current work lies at the intersection of the above two areas. More specifically, given noisy observations of entries of a structured matrix with rows and columns shuffled by latent permutations, we concern ourselves with the task of recovering the permutations and estimating the matrix. If the matrix in consideration is the adjacency matrix of a graph, then the task can be restated as reordering the nodes of the graph and estimating interactions between pairs of nodes. We study both statistical and computational aspects of the problem, by establishing minimax rates of estimation and designing provable polynomial-time algorithms, for several concrete models.

In slightly more formal terms, we consider the general model

$$Y = \Pi M \Sigma^\top + Z, \tag{1.1}$$

where M is an unknown $n_1 \times n_2$ real matrix with certain *shape constraints* (e.g. bivariate monotonicity), Π and Σ are unknown permutation matrices acting on the rows and columns of M respectively, and Z is a noise matrix which is assumed to be sub-Gaussian throughout this work. Given partial observations of the entries of Y , we aim to estimate the permutations Π and Σ , as well as the underlying matrix M . In the following chapters, we study three models—the statistical seriation model [FMR16], the noisy sorting model [BM08] and the strongly stochastic transitivity model [Cha15, SBGW17]—all of which can be written in the form (1.1).

Seriation. The first problem of interest is seriation, which finds its root in archaeology [Pet99, Rob51, Ken63, Ken69, OL99]. As the name suggests, the problem of seriation consists in ordering items in a series, so that adjacent items are more similar than distant items. In the setting of sequence dating in archaeology, the rows of the matrix M represent sepultures in chronologically order, and the columns represent artifacts. The sepultures are not ordered when they are discovered, which corresponds to that the rows of M are shuffled by an unknown permutation Π . We do not aim to order the artifacts, so Σ is assumed to be the identity. Moreover, entry $Y_{i,j}$ takes a binary value, indicating whether artifact j is present in sepulture i . It is a basic hypothesis [Pet99, Rob51] that the sets of artifacts in two chronologically close sepultures are similar, so we expect to see artifacts appearing consecutively if the sepultures are reordered chronologically. Equivalently, this says that reordering the rows of Y appropriately should bring it to having nearly consecutive ones along each column.

More generally, in Chapter 2, we consider a matrix M whose columns are *unimodal* (i.e., when we move down along a column, the entries first increase and then decrease). We refer to the model $Y = \Pi M + Z$ as the *statistical seriation* model, and study the corresponding minimax rates of estimation of the pair (Π, M) . Specifically, we demonstrate that the least squares estimator is optimal up to logarithmic factors and adapts to matrices whose columns have block structure. In addition, we propose and study a computationally efficient estimator in the case where the columns of M are monotone.

Noisy sorting. Next, we consider pairwise comparison data which arises naturally in various applications, such as social choice [CN91], tournament rankings [HMG06], web search [DKNS01] and recommender systems [BMR10]. Given noisy comparisons between pairs of items, the task is to recover the underlying ranks of the items (hence the name “noisy sorting”). Therefore, it is of importance to design and analyze robust models for ranking from pairwise comparisons.

More formally, suppose that items $1, \dots, n$ are associated with unknown ranks $\pi(1), \dots, \pi(n)$ according to their strength, where $\pi : [n] \rightarrow [n]$ is a permutation. Let M be an $n \times n$ matrix, whose entry $M_{k,\ell} \in [0, 1]$ denote the probability with which the k -th strongest item beats the ℓ -th strongest item in a comparison between them. The binary outcome of a comparison between items i and j is thus $Y_{i,j} = \text{Ber}(M_{\pi(i),\pi(j)})$, where $Y_{i,j} = 1$ if item i beats item j and $Y_{i,j} = 0$ if the opposite occurs. Therefore, the comparison model can be succinctly written as $Y = \text{Ber}(\Pi M \Pi^\top)$, where Π is the row permutation matrix corresponding to π , and the Bernoulli random variables are independent across the entries. Note that the model can be rewritten in the linearized form $Y = \Pi M \Pi^\top + Z$, which is a special case of (1.1).

A prototype of so-called *permutation-based* ranking models is the *noisy sorting* model [BM08, BM09]. In this model, a stronger item is assumed to beat a weaker item with probability $\frac{1}{2} + \lambda$ for a constant $\lambda \in [0, 1]$. Thus all the upper triangular entries of M are equal to $\frac{1}{2} + \lambda$, while all the lower triangular entries are equal to $\frac{1}{2} - \lambda$. It is easy to estimate the single parameter λ in this model, so the difficulty

entirely lies in recovering the unknown permutation π . In Chapter 3, we establish the minimax rates of learning the model, and provide a near-linear time algorithm to achieve near-optimal rates.

Stochastically transitive models. The noisy sorting model captures the discrete nature of ranking problems, yet falls short of properly modeling comparison probabilities in many situations, because it is unlikely that the outcomes of comparisons between two similar items and between two vastly different items follow the same distribution. Taking one step back, *parametric models* have been the mainstream in the statistics and machine learning literature for years [Hun04, NOS12, RA14, HOX14, SBB⁺16, NOS16, NOTX17]. In a parametric model, it is assumed that $M_{i,j} = F(w_i - w_j)$, where $F : \mathbb{R} \rightarrow [0, 1]$ an increasing link function known to the learner, and w is a decreasing vector whose entries represent the strength of the items. Modulo the nonlinearity F , the parametric model is essentially a rank-one model; therefore, while it allows polynomial-time rate-optimal estimation, the model again possesses weakness in modeling comparisons in certain scenarios [SBGW17].

Recently, a structurally richer permutation-based model, the strong stochastic transitivity (SST) model, has been proposed and shown to contain both the noisy sorting model and the parametric model as special cases [Cha15, SBGW17]. More precisely, the probability matrix M is assumed to be bivariate isotonic, i.e., to have nondecreasing rows and nonincreasing columns. Perhaps surprisingly, the minimax rate of estimation for the SST model is no slower than that for the noisy sorting or parametric model up to a logarithmic factor. However, this statistical advantage comes with a price: the parameter space of the SST model is highly non-convex, making efficient optimization unlikely. As a result, the statistically optimal rate for the SST model has not been achieved by efficient learning algorithms yet. In Chapter 4, we design and analyze polynomial-time algorithms that improve upon the state of the art. In particular, our results imply that for the SST model, a computationally efficient algorithm achieves the rate of estimation $\tilde{O}(n^{-3/4})$, narrowing the gap between $\tilde{\Theta}(n^{-1})$ and $\tilde{O}(n^{-1/2})$, which were hitherto the rates of the most statistically and computationally efficient methods respectively.

Fixed comparison topology The results discussed above for permutation-based models are achieved only when we have full or uniformly random observations of the entries of Y in equation (1.1). However, this is not necessarily a valid assumption for certain applications [HOX14, KO16, SBB⁺16], where we may observe comparisons of a *fixed* subset of pairs of items. The set of comparisons that we observe is referred to as the *comparison topology*. It is therefore of significant interest to study the dependence of the rate of estimation on the comparison topology for permutation-based models. In Chapter 5, we pursue this topic for both the noisy sorting model and the more general SST model.

More specifically, we show that when the assignment of items to the topology is arbitrary, these permutation-based models, unlike their parametric counterparts, do not admit consistent estimation for most comparison topologies used in practice. We

then demonstrate that consistent estimation is possible when the assignment of items to the topology is randomized, thus establishing a dichotomy between worst-case and average-case designs. We propose two estimators in the average-case setting and analyze their risk, showing that it depends on the comparison topology through the degree sequence of the topology. The rates achieved by these estimators are shown to be optimal for a large class of graphs.

Chapter 2 is based on joint work with Nicolas Flammarion and Philippe Rigollet [FMR16]. Chapter 3 is based on joint work with Jonathan Weed and Philippe Rigollet [MWR17]. Chapter 4 is based on joint work with Ashwin Pananjady and Martin J. Wainwright [MPW18]. Chapter 5 is based on joint work with Ashwin Pananjady, Vidya Muthukumar, Martin J. Wainwright and Thomas Courtade [PMM⁺17a].

Chapter 2

Optimal Rates of Statistical Seriation

Seriation has been a central technique for data analysis for over a century. It has roots in archaeology and especially *sequence dating* where the goal is to recover the chronological order of sepultures based on artifacts found in them [Pet99]. Since then seriation has found applications in a variety of disciplines ranging from anthropology [Cze09] to sociology [FK46], biology [Sok63] and marketing [ASDH88]. More recently, it was proposed as a method in computational biology for de novo DNA assembly [AS98]. See [Lii10] for a detailed account of seriation in data analysis. In modern language, seriation belongs to the class of *unsupervised learning* problems. Akin to clustering, it aims at rearranging heterogeneous data into a simple structure that is amenable to better interpretation and understanding. Actually, in his seminal work on clustering, Hartigan [Har72] advocates for a post-processing of direct clustering with seriation for better data visualization. However, unlike clustering methods that quantize the data into a pre-specified number of clusters, seriation methods are truly nonparametric and “non-destructive”, a term coined by Murtagh [Mur89], meaning that it does not discard information from the data. Perhaps one of the most spectacular successes of seriation was achieved in bioinformatics where it was used to display genome-wide expression patterns [ESBB98]. Despite its widespread use, seriation has not been the subject of statistical analysis. The main goal of this chapter is to propose a new model that is amenable to a statistical analysis of seriation.

To describe seriation in further details, we begin with a canonical problem, the *consecutive 1's problem* (C1P) [FG64] that is defined as follows. Given a binary matrix A the goal is to permute its rows in such a way that the resulting matrix enjoys the *consecutive 1's property*: each of its columns is a vector $v = (v_1, \dots, v_n)^\top$ where $v_j = 1$ if and only if $a \leq j \leq b$ for two integers a, b between 1 and n . This problem arises in the archaeology where the entry $A_{i,j}$ of matrix A indicates the presence of an artifact of type j in sepulture i . In his seminal work, egyptologist Flinders Petrie [Pet99] formulated the hypothesis that two sepultures should be close in the time domain if they present similar sets of artifacts, which indicate that the matrix A should be close to a matrix having the consecutive 1's property. In an influential follow-up work, Robinson [Rob51] generalized this problem to the case where $A_{i,j}$ counts the number of artifacts of type j in sepulture i . Robinson argues that “*types come into and get out of general use*” so that it is reasonable to assume that the

columns of A are, in fact unimodal: the count of a certain type of artifact increases as it comes into general use and decreases as it gets out. Note that matrices that satisfy the consecutive 1's property have, in particular, unimodal columns. More generally, seriation is used to rearrange matrices whose rows are permuted and whose columns satisfy a nonparametric *shape constraint*. For example the case where A has monotone columns arises in bipartite ranking under the strong stochastic transitivity assumption (see subsection 2.1.2). In the rest of this chapter we consider both the unimodal and the monotone setting.

Because of the presence of a latent permutation, the C1P exhibits interesting algorithmic challenges already in the noiseless case and that have motivated much of its study. In particular, it is reducible to the famous Traveling Salesman Problem [GG12] as observed by statistician David Kendall [Ken63, Ken69, Ken70, Ken71] who employed early tools from multidimensional scaling as a heuristic to solve it. The C1P belongs to a more general class of problems that consist in optimizing various criteria over the discrete set of permutations and that can be recast as examples of the notoriously hard *quadratic assignment problem* [LdABN⁺07]. While such problems are NP-hard in general, some examples, including C1P, may be solved efficiently using either combinatorial optimization [FG64], spectral methods [ABH98] or convex optimization [FJBd13, LW14]. However, little is known about the robustness to statistical noise of such methods.

In order to set the benchmark for the noisy case, we propose a *statistical seriation model* and study optimal rates of estimation for this model. Assume that we observe an $n \times m$ matrix $Y = \Pi A + Z$, where Π is an unknown $n \times n$ permutation matrix, Z is an $n \times m$ noise matrix and $A \in \mathbb{R}^{n \times m}$ is assumed to have columns that satisfy a certain shape constraint. Our goal is to give estimators $\hat{\Pi}$ and \hat{A} so that $\hat{\Pi}\hat{A}$ is close to ΠA . The shape constraint can be the consecutive 1's property, but more generally, we consider the class of matrices that have unimodal columns, which also include monotone columns as a special case. These terms will be formally defined at the end of this section.

The rest of the chapter is organized as follows. In Section 2.1 we formulate the model and discuss related work. Section 2.2 collects our main results, including uniform and adaptive upper bounds for the least squares estimator together with corresponding minimax lower bounds in the general unimodal case. In Section 2.3, for the special case of monotone columns, we propose a computationally efficient alternative to the least squares estimator and study its rates of convergence both theoretically and numerically. Section 2.4 presents new bounds for unimodal regression implied by our analysis, which are minimax optimal up to logarithmic factors. Section 2.5 is devoted to the proofs of the results. We conclude with a discussion in Section 2.6.

Notation. For a positive integer n , define $[n] = \{1, \dots, n\}$. For a matrix $A \in \mathbb{R}^{n \times m}$, let $\|A\|_F$ denote its Frobenius norm, and let $A_{i\cdot}$ be its i -th row and $A_{\cdot j}$ be its j -th column. Let $\mathcal{B}^n(a, t)$ denote the Euclidean ball of radius t centered at a in \mathbb{R}^n . We use C and c to denote positive constants that may change from line to line. For any two sequences $(u_n)_n$ and $(v_n)_n$, we write $u_n \lesssim v_n$ if there exists an absolute constant

$C > 0$ such that $u_n \leq Cv_n$ for all n . We define $u_n \gtrsim v_n$ analogously. Given two real numbers a, b , define $a \wedge b = \min(a, b)$ and $a \vee b = \max(a, b)$.

Denote the closed convex cone of increasing¹ sequences in \mathbb{R}^n by $\mathcal{S}_n = \{a \in \mathbb{R}^n : a_1 \leq \dots \leq a_n\}$. We define \mathcal{S}^m to be the Cartesian product of m copies of \mathcal{S}_n and we identify \mathcal{S}^m to the set of $n \times m$ matrices with increasing columns.

For any $l \in [n]$, define the closed convex cone $\mathcal{C}_l = \{a \in \mathbb{R}^n : a_1 \leq \dots \leq a_l\} \cap \{a \in \mathbb{R}^n : a_l \geq \dots \geq a_n\}$, which consists of vectors in \mathbb{R}^n that increase up to the l -th entry and then decrease. Define the set \mathcal{U} of unimodal sequences in \mathbb{R}^n by $\mathcal{U} = \bigcup_{l=1}^n \mathcal{C}_l$. We define \mathcal{U}^m to be the Cartesian product of m copies of \mathcal{U} and we identify \mathcal{U}^m to the set of $n \times m$ matrices with unimodal columns. It is also convenient to write \mathcal{U}^m as a union of closed convex cones as follows. For $\mathbf{l} = (l_1, \dots, l_m) \in [n]^m$, let $\mathcal{C}_1^m = \mathcal{C}_{l_1} \times \dots \times \mathcal{C}_{l_m}$. Then \mathcal{U}^m is the union of the n^m closed convex cones $\mathcal{C}_1^m, \mathbf{l} \in [n]^m$.

Finally, let \mathfrak{S}_n be the set of $n \times n$ permutation matrices and define the set $\mathcal{M} = \bigcup_{\Pi \in \mathfrak{S}_n} \Pi \mathcal{U}^m$ where $\Pi \mathcal{U}^m = \{\Pi A : A \in \mathcal{U}^m\}$, so that \mathcal{M} is the union of the $n!n^m$ closed convex cones $\Pi \mathcal{C}_1^m, \Pi \in \mathfrak{S}_n, \mathbf{l} \in [n]^m$.

2.1 Problem setup and related work

In this section, we formally state the problem of interest and discuss several lines of related work.

2.1.1 The seriation model

Suppose that we observe a matrix $Y \in \mathbb{R}^{n \times m}$, $n \geq 2$ such that

$$Y = \Pi^* A^* + Z, \quad (2.1)$$

where $A^* \in \mathcal{U}^m$, $\Pi \in \mathfrak{S}_n$ and Z is a centered sub-Gaussian noise matrix with variance proxy $\sigma^2 > 0$. Specifically, Z is a matrix such that $\mathbb{E}[Z] = 0$ and, for any $M \in \mathbb{R}^{n \times m}$,

$$\mathbb{E}[\exp(\text{Tr}(Z^T M))] \leq \exp\left(\frac{\sigma^2 \|M\|_F^2}{2}\right),$$

where $\text{Tr}(\cdot)$ is the trace operator. We write $Z \sim \text{subG}_{n,m}(\sigma^2)$ or simply $Z \sim \text{subG}(\sigma^2)$ when dimensions are clear from the context.

Given the observation Y , our goal is to estimate the unknown pair (Π^*, A^*) . The performance of an estimator $(\hat{\Pi}, \hat{A}) \in \mathfrak{S}_n \times \mathcal{U}^m$, is measured by the quadratic loss:

$$\frac{1}{nm} \|\hat{\Pi} \hat{A} - \Pi^* A^*\|_F^2.$$

In particular, its expectation is the mean squared error. Since we are interested in estimating $\Pi^* A^* \in \mathcal{M}$, we can also view \mathcal{M} as the parameter space.

¹Throughout the chapter, we loosely use the terms “increasing” and “decreasing” to mean “monotonically non-decreasing” and “monotonically non-increasing” respectively.

In the general unimodal case, upper bounds on the above quadratic loss do not imply individual upper bounds on estimation of the matrix Π^* or the matrix A^* due to lack of identifiability. Nevertheless, if we further assume that the columns of A^* are monotone increasing, that is $A^* \in \mathcal{S}^m$, then the following lemma holds.

Lemma 2.1.1. *If $A^*, \tilde{A} \in \mathcal{S}^m$, then for any $\Pi^*, \tilde{\Pi} \in \mathfrak{S}_n$, we have that*

$$\|\tilde{A} - A^*\|_F^2 \leq \|\tilde{\Pi}\tilde{A} - \Pi^*A^*\|_F^2,$$

and that

$$\|\tilde{\Pi}A^* - \Pi^*A^*\|_F^2 \leq 4\|\tilde{\Pi}\tilde{A} - \Pi^*A^*\|_F^2.$$

Proof. Let $a, b \in \mathcal{S}_n$ and $b_\pi = (b_{\pi(1)}, \dots, b_{\pi(n)})$ where $\pi : [n] \rightarrow [n]$ is a permutation. It is easy to check that $\sum_{i=1}^n a_i b_i \geq \sum_{i=1}^n a_i b_{\pi(i)}$, so $\|a - b\|_2^2 \leq \|a - b_\pi\|_2^2$. Applying this inequality to columns of matrices, we see that

$$\|\tilde{A} - A^*\|_F^2 \leq \|\tilde{A} - \tilde{\Pi}^{-1}\Pi^*A^*\|_F^2 = \|\tilde{\Pi}\tilde{A} - \Pi^*A^*\|_F^2,$$

since $A^*, \tilde{A} \in \mathcal{S}^m$. Moreover, $\|\tilde{\Pi}A^* - \tilde{\Pi}\tilde{A}\|_F = \|A^* - \tilde{A}\|_F$, so

$$\|\tilde{\Pi}A^* - \Pi^*A^*\|_F \leq \|A^* - \tilde{A}\|_F + \|\tilde{\Pi}\tilde{A} - \Pi^*A^*\|_F \leq 2\|\tilde{\Pi}\tilde{A} - \Pi^*A^*\|_F,$$

by the triangle inequality and the previous display. \square

Lemma 2.1.1 guarantees that $\|\tilde{\Pi}\tilde{A} - \Pi^*A^*\|_F$ is a pertinent measure of the performance of both $\tilde{\Pi}$ and \tilde{A} . Note further that $\|\tilde{\Pi}A^* - \Pi^*A^*\|_F$ is large if $\tilde{\Pi}$ misplaces rows of A^* that have large differences, and is small if $\tilde{\Pi}$ only misplaces rows of A^* that are close to each other. We argue that, in the seriation context, this measure of distance between permutations is more natural than ad hoc choices such as the trivial 0/1 distance or popular choices such as Kendall's τ or Spearman's ρ .

Apart from Section 2.3 (and Section 2.5.4), the rest of this chapter focuses on the least squares (LS) estimator defined by

$$(\hat{\Pi}, \hat{A}) \in \underset{(\Pi, A) \in \mathfrak{S}_n \times \mathcal{U}^m}{\operatorname{argmin}} \|Y - \Pi A\|_F^2. \quad (2.2)$$

Taking $\hat{M} = \hat{\Pi}\hat{A}$, we see that it is equivalent to define the LS estimator by

$$\hat{M} \in \underset{M \in \mathcal{M}}{\operatorname{argmin}} \|Y - M\|_F^2. \quad (2.3)$$

Note that in our case, the set of parameters \mathcal{M} is a union of $n!n^m$ closed convex cones but is not convex itself. Thus it is not clear how to compute the LS estimator efficiently. We discuss this aspect in further details in the context of monotone columns in Section 2.3. Nevertheless, the main focus of this chapter is the least squares estimator which, as we shall see, is near-optimal in a minimax sense and therefore serves as a benchmark for the statistical seriation model.

2.1.2 Related work

Our work falls broadly in the scope of statistical inference under shape constraints but presents a major twist: the unknown latent permutation Π^* .

Shape constrained regression

To set our goals, we first consider the case where the permutation is known and assume without loss of generality that $\Pi^* = I_n$. In this case, we can estimate individually each column $A^*_{\cdot,j}$ by an estimator $\hat{A}_{\cdot,j}$ and then obtain an estimator \hat{A} for the whole matrix by concatenating the columns $\hat{A}_{\cdot,j}$. Thus the task is reduced to estimation of a vector θ^* which satisfies a certain shape constraint from an observation $y = \theta^* + z$ where $z \sim \text{subG}_{n,1}(\sigma^2)$.

When θ^* is assumed to be increasing we speak of isotonic regression [BBBB72]. The LS estimator defined by $\hat{\theta} = \text{argmin}_{\theta \in \mathcal{S}_n} \|\theta - y\|_2^2$ can be computed in closed form in $O(n)$ using the Pool-Adjacent-Violators algorithm (PAVA) [ABE⁺55, BBBB72, RWD88] and its statistical performance has been studied by Zhang [Zha02] (see also [NPT85, Don90, vdG90, Mam91, vdG93] for similar bounds using empirical process theory) who showed in the Gaussian case $z \sim N(0, \sigma^2 I_n)$ that the mean squared error behaves like

$$\frac{1}{n} \mathbb{E} \|\hat{\theta} - \theta^*\|_2^2 \asymp \left(\frac{\sigma^2 V(\theta^*)}{n} \right)^{2/3}, \quad (2.4)$$

where $V(\theta) = \max_{i \in [n]} \theta_i - \min_{i \in [n]} \theta_i$ is the variation of $\theta \in \mathbb{R}^n$. Note that $2/3 = 2\beta/(2\beta + 1)$ for $\beta = 1$ so that this is the minimax rate of estimation of Lipschitz functions (see, e.g., [Tsy09]).

The rate in (2.4) is said to be *global* as it holds uniformly over the set of monotone vectors with variation $V(\theta^*)$. Recently, [CGS15] have initiated the study of *adaptive* bounds that may be better if θ^* has a simpler structure in some sense. To define this structure, let $k(\theta) = \text{card}(\{\theta_1, \dots, \theta_n\})$ denote the cardinality of entries of $\theta \in \mathbb{R}^n$. In this context, [CGS15] showed that the LS estimator satisfies the adaptive bound

$$\frac{1}{n} \mathbb{E} \|\hat{\theta} - \theta^*\|_2^2 \leq C \inf_{\theta \in \mathcal{S}_n} \left(\frac{\|\theta - \theta^*\|_2^2}{n} + \frac{\sigma^2 k(\theta)}{n} \log \frac{en}{k(\theta)} \right). \quad (2.5)$$

This result was extended in [Bel15] to a sharp oracle inequality where $C = 1$. This bound was also shown to be optimal in a minimax sense [CGS15, BT15].

Unlike its monotone counterpart, unimodal regression where $\theta^* \in \mathcal{U}$ has received sporadic attention [SZ01, KBI14, CL15]. This state of affairs is all the more surprising given that unimodal density estimation has been the subject of much more research [BF96, Bir97, EL00, DDS12, DDS⁺13, TG14]. It was recently shown in [CL15] that the LS estimator also adapts to $V(\theta^*)$ and $k(\theta^*)$ for unimodal regression:

$$\frac{1}{n} \|\hat{\theta} - \theta^*\|_2^2 \lesssim \min \left(\sigma^{4/3} \left(\frac{V(\theta^*) + \sigma}{n} \right)^{2/3}, \frac{\sigma^2}{n} k(\theta^*)^{3/2} (\log n)^{3/2} \right) \quad (2.6)$$

with probability at least $1 - n^{-\alpha}$ for some $\alpha > 0$. The exponent $3/2$ in the second

term was improved to 1 in the new version of [CL15] after the first version of our work [FMR16] was posted. Note that the exponents in (2.6) are different from the isotonic case. Our results will imply that they are not optimal and in fact the LS estimator achieves the same rate as in isotonic regression. See Corollary 2.4.1 for more details. The algorithmic aspect of unimodal regression has received more attention [Fri86, GS90, BS98, BMI06] and [Sto08] showed that the LS estimator can be computed with time complexity $O(n)$ using a modified version of PAVA. Hence there is little difference between isotonic and unimodal regressions from both computational and statistical points of views.

Latent permutation learning

When the permutation Π^* is unknown the estimation problem is more involved. Noisy permutation learning was explicitly addressed in [CD16] where the problem of matching two sets of noisy vectors was studied from a statistical point of view. Given $n \times m$ matrices $Y = A + Z$ and $\tilde{Y} = \Pi^* A + \tilde{Z}$, where $A \in \mathbb{R}^{n \times m}$ is an unknown matrix and $\Pi^* \in \mathbb{R}^{n \times n}$ is an unknown permutation matrix, the goal is to recover Π^* . It was shown in [CD16] that if $\min_{i \neq j} \|A_{i \cdot} - A_{j \cdot}\|_2 \geq c\sigma((\log n)^{1/2} \vee (m \log n)^{1/4})$, then the LS estimator defined by $\hat{\Pi} = \operatorname{argmin}_{\Pi \in \mathfrak{S}_n} \|\Pi Y - \tilde{Y}\|_F^2$ recovers the true permutation with high probability. However they did not directly study the behavior of $\|\hat{\Pi} A - \Pi^* A\|_F^2$.

In his celebrated paper on matrix estimation [Cha15], Sourav Chatterjee describes several noisy matrix models involving unknown latent permutations. One is the *nonparametric Bradley-Terry-Luce* (NP-BTL) model where we observe a matrix $Y \in \mathbb{R}^{n \times n}$ with independent entries $Y_{i,j} \sim \operatorname{Ber}(P_{i,j})$ for some unknown parameters $P = \{P_{i,j}\}_{1 \leq i,j \leq n}$ where $P_{i,j} \in [0, 1]$ is equal to the probability that item i is preferred over item j and $P_{j,i} = 1 - P_{i,j}$. Crucially, the NP-BTL model assumes the so-called *strong stochastic transitivity* (SST) [DM59, Fis73] assumption: there exists an unknown permutation matrix $\Pi \in \mathbb{R}^{n \times n}$ such that the ordered matrix $A = \Pi^T P \Pi$ satisfies $A_{1,k} \leq \dots \leq A_{n,k}$ for all $k \in [n]$. Note that the NP-BTL model is a special case of our model (2.1) where $m = n$ and $Z \sim \operatorname{subG}(1/4)$ is taken to be Bernoulli. Chatterjee proposed an estimator \hat{P} that leverages the fact that any matrix P in the NP-BTL model can be approximated by a low rank matrix and proved [Cha15, Theorem 2.11] that $n^{-2} \|\hat{P} - P\|_F^2 \lesssim n^{-1/4}$, which was improved to $n^{-1/2}$ by [SBGW17] for a variation of this estimator. This method does not yield individual estimators of Π or A . Instead [CM16] proposed estimators $\hat{\Pi}$ and \hat{A} so that $\hat{\Pi} \hat{A} \hat{\Pi}^T$ estimates P with the same rate $n^{-1/2}$ up to a logarithmic factor. The non-optimality of this rate has been observed in [SBGW17] who showed that the correct rate should be of order n^{-1} up to a possible $\log n$ factor. However, it is not known whether a computationally efficient estimator could achieve the fast rate. A recent work [SBW16b] explored a new notion of adaptivity for which the authors proved a computational lower bound, and also proposed an efficient estimator whose rate of estimation matches that lower bound.

Also mentioned in Chatterjee's paper is the so-called *stochastic block model* that has since received such extensive attention in various communities that it is futile

to attempt to establish a comprehensive list of references. Instead, we refer the reader to [GLZ15] and references therein. This paper establishes the minimax rates for this problem and its continuous limit, the graphon estimation problem and, as such, constitutes the state-of-the-art in the statistical literature. In the stochastic block model with $k \geq 2$ blocks, we assume that we observe a matrix $Y = P + Z$ where $P = \Pi A \Pi^\top$, $\Pi \in \mathbb{R}^{n \times n}$ is an unknown permutation matrix and A has a block structure, namely, there exist positive integers $n_1 < \dots < n_k < n_{k+1} := n$, and k^2 real numbers $a_{s,t}$, $(s, t) \in [k]^2$ such that A has entries

$$A_{i,j} = \sum_{(s,t) \in [k]^2} a_{s,t} \mathbb{1}\{n_s \leq i \leq n_{s+1}, n_t \leq j \leq n_{t+1}\}, \quad i, j \in [n].$$

While traditionally, the stochastic block model is a network model and therefore pertains only to Bernoulli observations, the more general case of sub-Gaussian additive error is also explicitly handled in [GLZ15]. For this problem, Gao, Lu and Zhou have established that the least squares estimator \hat{P} satisfies $n^{-2} \|\hat{P} - P\|_F^2 \lesssim k^2/n^2 + (\log k)/n$ together with a matching lower bound. Using piecewise constant approximation to bivariate Hölder functions, they also establish that this estimator with a correct choice of k leads to minimax optimal estimation of smooth graphons. Both results exploit extensively the fact that the matrix P is equal to or can be well approximated by a piecewise constant matrix and our results below take a similar route by observing that monotone and unimodal vectors are also well approximated by piecewise constant ones. In addition, we allow for rectangular matrices.

In fact, our result can be also formulated as a network estimation problem but on a bipartite graph, thus falling at the intersection of the above two examples. Assume that n left nodes represent items and that m right nodes represent users. Assume further that we observe the $n \times m$ adjacency matrix Y of a random graph where the presence of edge (i, j) indicates that user j has purchased or liked item i . Define $P = \mathbb{E}[Y]$ and assume SST across items in the sense that there exists an unknown $n \times n$ permutation matrix Π^* such that $P = \Pi^* A^*$ and A^* is such that $A^*_{1,j} \leq \dots \leq A^*_{n,j}$ for all users $j \in [m]$. This model of bipartite ranking falls into the scope of the statistical seriation model (2.1).

2.2 Main results

2.2.1 Adaptive oracle inequalities

For a matrix $A \in \mathcal{U}^m$, let $k(A_{\cdot,j}) = \text{card}(\{A_{1,j}, \dots, A_{n,j}\})$ be the number of values taken by the j -th column of A and define $K(A) = \sum_{j=1}^m k(A_{\cdot,j})$. Observe that $K(A) \geq m$. The first theorem shows that the LS estimator adapts to the complexity K .

Theorem 2.2.1. *For $A^* \in \mathbb{R}^{n \times m}$ and $Y = \Pi^* A^* + Z$, let $(\hat{\Pi}, \hat{A})$ be the LS estimator defined in (2.2). Then the following oracle inequality holds*

$$\frac{1}{nm} \|\hat{\Pi} \hat{A} - \Pi^* A^*\|_F^2 \lesssim \min_{A \in \mathcal{U}^m} \left(\frac{1}{nm} \|A - A^*\|_F^2 + \sigma^2 \frac{K(A)}{nm} \log \frac{enm}{K(A)} \right) + \sigma^2 \frac{\log n}{m} \quad (2.7)$$

with probability at least $1 - e^{-c(n+m)}$, $c > 0$. Moreover,

$$\frac{1}{nm} \mathbb{E} \|\hat{\Pi}\hat{A} - \Pi^*A^*\|_F^2 \lesssim \min_{A \in \mathcal{U}^m} \left(\frac{1}{nm} \|A - A^*\|_F^2 + \sigma^2 \frac{K(A)}{nm} \log \frac{enm}{K(A)} \right) + \sigma^2 \frac{\log n}{m}. \quad (2.8)$$

Note that while we assume that $A^* \in \mathcal{U}^m$ in (2.1), the above oracle inequalities hold in fact for any $A^* \in \mathbb{R}^{n \times m}$ even if its columns are *not* assumed to be unimodal. The oracle inequalities indicate that the LS estimator automatically trades off the approximation error $\|A - A^*\|_F^2$ for the stochastic error $\sigma^2 K(A) \log(enm/K(A))$. Moreover, 3 is the best constant we can achieve before the oracle approximation term when the error is expressed in the Frobenius norm, i.e.,

$$\|\hat{\Pi}\hat{A} - \Pi^*A^*\|_F \leq \min_{A \in \mathcal{U}^m} (3\|A - A^*\|_F + \text{stochastic error terms}).$$

This is the content of (2.21) in the proof of Theorem 2.2.1. Making (2.7) and (2.8) into sharp oracle inequalities remains an interesting open problem.

If A^* is assumed to have unimodal columns, then we can take $A = A^*$ in (2.7) and (2.8) to get the following corollary.

Corollary 2.2.2. *For $A^* \in \mathcal{U}^m$ and $Y = \Pi^*A^* + Z$, the LS estimator $(\hat{\Pi}, \hat{A})$ satisfies*

$$\frac{1}{nm} \|\hat{\Pi}\hat{A} - \Pi^*A^*\|_F^2 \lesssim \sigma^2 \left(\frac{K(A^*)}{nm} \log \frac{enm}{K(A^*)} + \frac{\log n}{m} \right)$$

with probability at least $1 - e^{-c(n+m)}$, $c > 0$. Moreover, the corresponding bound with the same rate holds in expectation.

The two terms in the adaptive bound can be understood as follows. The first term corresponds to the estimation of the matrix A^* with unimodal columns if the permutation Π^* is known. It can be viewed as a matrix version of the adaptive bound (2.5) for the vector case. The LS estimator adapts to the cardinality of entries of A^* as it achieves a provably better rate if $K(A^*)$ is smaller while not requiring knowledge of $K(A^*)$. The second term corresponds to the error due to the unknown permutation Π^* . As m grows to infinity this second term vanishes, because we have more samples to estimate Π^* better. If $m \geq n$, it is easy to check that the permutation term is dominated by the first term, so the rate of estimation is the same as if the permutation is known.

2.2.2 Global oracle inequalities

The bounds in Theorem 2.2.1 adapt to the cardinality of the oracle. In this subsection, we state another type of upper bounds for the LS estimator $(\hat{\Pi}, \hat{A})$. They are called global bounds because they hold uniformly over the class of matrices whose columns are unimodal and that have bounded variation. Recall that we call *variation* of a vector $a \in \mathbb{R}^n$ the scalar $V(a) \geq 0$ defined by

$$V(a) = \max_{1 \leq i \leq n} a_i - \min_{1 \leq i \leq n} a_i.$$

We extend this notion to a matrix $A \in \mathbb{R}^{n \times m}$ by defining

$$V(A) = \left(\frac{1}{m} \sum_{j=1}^m V(A_{\cdot,j})^{2/3} \right)^{3/2}.$$

While this $2/3$ -norm may seem odd at first sight, it turns out to be the correct extrapolation from vectors to matrices, at least in the context under consideration here. Indeed, the following upper bound, in which this quantity naturally appears, is matched by the lower bound of Theorem 2.2.6 up to logarithmic terms.

Theorem 2.2.3. *For $A^* \in \mathbb{R}^{n \times m}$ and $Y = \Pi^* A^* + Z$, let $(\hat{\Pi}, \hat{A})$ be the LS estimator defined in (2.2). Then it holds that*

$$\frac{1}{nm} \|\hat{\Pi} \hat{A} - \Pi^* A^*\|_F^2 \lesssim \min_{A \in \mathcal{U}^m} \left[\frac{1}{nm} \|A - A^*\|_F^2 + \left(\frac{\sigma^2 V(A) \log n}{n} \right)^{2/3} \right] + \sigma^2 \frac{\log n}{n \wedge m}. \quad (2.9)$$

with probability at least $1 - e^{-c(n+m)}$, $c > 0$. Moreover, the corresponding bound with the same rate holds in expectation.

If $A^* \in \mathcal{U}^m$, then taking $A = A^*$ in Theorem 2.2.3 leads to the following corollary that indicates that the LS estimator is adaptive to the quantity $V(A^*)$.

Corollary 2.2.4. *For $A^* \in \mathcal{U}^m$ and $Y = \Pi^* A^* + Z$, the LS estimator $(\hat{\Pi}, \hat{A})$ satisfies*

$$\frac{1}{nm} \|\hat{\Pi} \hat{A} - \Pi^* A^*\|_F^2 \lesssim \left(\frac{\sigma^2 V(A^*) \log n}{n} \right)^{2/3} + \sigma^2 \frac{\log n}{n \wedge m}$$

with probability at least $1 - e^{-c(n+m)}$, $c > 0$. Moreover, the corresponding bound with the same rate holds in expectation.

Akin to the adaptive bound, the above inequality can be viewed as a sum of a matrix version of (2.4) and an error due to estimation of the unknown permutation. Observe that if $\sigma = 1$, $m \geq n^{2/3}$ and all the entries are bounded by a universal constant, then the rate of estimation simplifies to $\tilde{O}(n^{-2/3})$. Since every monotone vector is unimodal, the rate $\tilde{O}(n^{-2/3})$ also holds for the case where columns of A^* are monotone, which will be discussed in detail in Section 2.3. Recently, rates of $\tilde{O}(n^{-1})$ have been established for bi-isotonic matrices with latent permutations [SBGW17, CM16], where bi-isotonicity means that the columns and the rows of the underlying matrix are both monotone. We emphasize that our rate is slower because only the columns of the matrix are assumed to be unimodal or monotone, while no constraints are imposed on the rows. The minimax lower bounds below in fact suggest that the rate $\tilde{O}(n^{-2/3})$ is optimal up to a logarithmic factor.

Having stated the main upper bounds, we digress a little to remark that the proofs of Theorem 2.2.1 and Theorem 2.2.3 also yield a minimax optimal rate of estimation (up to logarithmic factors) for unimodal regression, which improves the bound (2.6). We discuss the details in Section 2.4.

2.2.3 Minimax lower bounds

Given the model $Y = \Pi^*A^* + Z$ where entries of Z are i.i.d. $N(0, \sigma^2)$ random variables, let $(\hat{\Pi}, \hat{A})$ denote any estimator of (Π^*, A^*) , i.e., any pair in $\mathfrak{S}_n \times \mathbb{R}^{n \times m}$ that is measurable with respect to the observation Y . We will prove lower bounds that match the rates of estimation in Corollary 2.2.2 and Corollary 2.2.4 up to logarithmic factors. The combination of upper and lower bounds, implies simultaneous near optimality of the least squares estimator over a large scale of matrix classes.

For $m \leq K_0 \leq nm$ and $V_0 > 0$, define $\mathcal{U}_{K_0}^m = \{A \in \mathcal{U}^m : K(A) \leq K_0\}$ and $\mathcal{U}^m(V_0) = \{A \in \mathcal{U}^m : V(A) \leq V_0\}$. We present below two lower bounds, one for the adaptive rate uniformly over $\mathcal{U}_{K_0}^m$ and one for the global rate uniformly over $\mathcal{U}^m(V_0)$. This splitting into two cases is solely justified by better readability but it is worth noting that a stronger lower bound that holds on the intersection $\mathcal{U}_{K_0}^m \cap \mathcal{U}^m(V_0)$ can also be proved and is presented as Proposition 2.5.9.

Theorem 2.2.5. *There exists a constant $c \in (0, 1)$ such that for any $K_0 \geq m$, and any estimator $(\hat{\Pi}, \hat{A})$, it holds that*

$$\sup_{(\Pi, A) \in \mathfrak{S}_n \times \mathcal{U}_{K_0}^m} \mathbb{P}_{\Pi A} \left[\frac{1}{nm} \|\hat{\Pi}\hat{A} - \Pi A\|_F^2 \gtrsim \sigma^2 \left(\frac{K_0}{nm} + \frac{\log l}{m} \right) \right] \geq c,$$

where $l = \min(K_0 - m, m) + 1$ and $\mathbb{P}_{\Pi A}$ is the probability distribution of $Y = \Pi A + Z$. It follows that the lower bound with the same rate holds in expectation.

In fact, the lower bound holds for any estimator of the matrix Π^*A^* , not only those of the form $\hat{\Pi}\hat{A}$ with $\hat{A} \in \mathcal{U}^m$. The above lower bound matches the upper bound in Corollary 2.2.2 up to logarithmic factors.

Note the presence of a $\log l$ factor in the second term. If $l = 1$ then $K_0 = m$ which means that each column of A is simply a constant block, so $\Pi A = A$ for any $\Pi \in \mathfrak{S}_n$. In this case, the second term vanishes because the permutation does not play a role. More generally, the number $l - 1$ can be understood as the maximal number of columns of A on which the permutation does have an effect. The larger l , the harder the estimation. It is easy to check that if $l \geq n$ the second term in the lower bound will be dominated by the first term in the upper bound.

A lower bound corresponding to Corollary 2.2.4 also holds:

Theorem 2.2.6. *There exists a constant $c \in (0, 1)$ such that for any $V_0 \geq 0$, and any estimator $(\hat{\Pi}, \hat{A})$, it holds that*

$$\sup_{(\Pi, A) \in \mathfrak{S}_n \times \mathcal{U}^m(V_0)} \mathbb{P}_{\Pi A} \left[\frac{1}{nm} \|\hat{\Pi}\hat{A} - \Pi A\|_F^2 \gtrsim \left(\frac{\sigma^2 V_0}{n} \right)^{2/3} + \frac{\sigma^2}{n} + \frac{\sigma^2}{m} \wedge m^2 V_0^2 \right] \geq c,$$

where $\mathbb{P}_{\Pi A}$ is the probability distribution of $Y = \Pi A + Z$. The lower bound with the same rate also holds in expectation.

There is a slight mismatch between the upper bound of Corollary 2.2.4 and the lower bound of Theorem 2.2.6 above. Indeed the lower bound features a term $\frac{\sigma^2}{m} \wedge$

$m^2V_0^2$ instead of just $\frac{\sigma^2}{m}$. In the regime $m^2V_0^2 < \frac{\sigma^2}{m}$, where A has very small variation, the LS estimator may not be optimal. Proposition 2.2.7 below, whose proof can be found in Section 2.7, indicates that a matrix with constant columns obtained by averaging achieves optimality in this extreme regime.

Proposition 2.2.7. *For $Y = \Pi^*A^* + Z$ where $Z \sim \text{subG}(\sigma^2)$, let $\hat{\Pi} = I_n$ and \hat{A} be defined by $\hat{A}_{i,j} = \frac{1}{n} \sum_{k=1}^n Y_{k,j}$ for all $(i, j) \in [n] \times [m]$. Then,*

$$\frac{1}{nm} \|\hat{\Pi}\hat{A} - \Pi^*A^*\|_F^2 \lesssim \frac{\sigma^2}{n} + m^2V(A)^2$$

with probability at least $1 - \exp(-m)$ and the corresponding bound with the same rate holds in expectation.

2.3 Further results in the monotone case

A particularly interesting subset of unimodal matrices is \mathcal{S}^m , the set of $n \times m$ matrices with monotonically increasing columns. While it does not amount to the seriation problem in its full generality, this special case is of prime importance in the context of shape constrained estimation as illustrated by the discussion and references in Section 2.1.2. In fact, it covers the example of bipartite ranking discussed at the end of Section 2.1.2. In the rest of this section, we devote further investigation to this important case. To that end, consider the model (2.1) where we further assume that $A^* \in \mathcal{S}^m$. We refer to this model as the *monotone seriation model*. In this context, define the LS estimator by

$$(\hat{\Pi}, \hat{A}) \in \underset{(\Pi, A) \in \mathfrak{S}_n \times \mathcal{S}^m}{\operatorname{argmin}} \|Y - \Pi A\|_F^2.$$

Since \mathcal{S}^m is a convex subset of \mathcal{U}^m , it is easily seen that the upper bounds in Theorem 2.2.1 and 2.2.3 remain valid in this case. The lower bounds of Theorem 2.2.5 (with $\log l$ replaced by 1) and Theorem 2.2.6 also extend to this case; see Section 2.5.3.

Although for unimodal matrices the established error bounds do not imply any bounds on estimation of A^* or Π^* in general, for the monotonic case, however, Lemma 2.1.1 yields that

$$\|\hat{A} - A^*\|_F^2 \vee \frac{1}{4} \|(\hat{\Pi} - \Pi^*)A^*\|_F^2 \leq \|\hat{\Pi}\hat{A} - \Pi^*A^*\|_F^2.$$

so that the LS estimator $(\hat{\Pi}, \hat{A})$ also leads to good individual estimators of Π^* and A^* respectively.

2.3.1 RankScore: An efficient estimator and its performance

Because it requires optimizing over a union of $n!$ cones $\Pi\mathcal{S}^m$, no efficient way of computing the LS estimator is known since. As an alternative, we describe a simple and efficient algorithm to estimate (Π^*, A^*) and study its rate of estimation.

The main difficulty of the problem lies in providing an efficient estimator $\tilde{\Pi}$ of Π^* , because after determining $\tilde{\Pi}$ we may project Y onto the convex cone $\tilde{\Pi}\mathcal{S}^m$ efficiently to estimate A^* . Recovering the permutation Π^* is equivalent to sorting the rows of Π^*A^* from their noisy version Y . One simple method to aggregate information across columns, which we call **RankSum**, is to sort the rows of Y so that they have increasing row sums. However, it is easy to observe that this method fails if

$$A^* = \begin{bmatrix} 0 & 0 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & 0 \\ \sqrt{m} & 0 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ \sqrt{m} & 0 & \dots & 0 \end{bmatrix} \quad (2.10)$$

where the last $\lfloor \frac{n}{2} \rfloor$ entries in the first column of A^* are equal to \sqrt{m} and the entries of Z are i.i.d. standard Gaussian variables. Because the sum of noise in each row is of order \sqrt{m} which is no less than the gaps between row sums of A^* , **RankSum** will place a nonzero row before a zero row with a constant probability. Therefore, if $\tilde{\Pi}$ is the permutation given by **RankSum**, then $\|\tilde{\Pi}\tilde{A} - \Pi^*A^*\|_F^2$ will be of order nm regardless of the matrix $\tilde{A} \in \mathcal{S}^m$, so we have no hope of consistent estimation in general.

In fact, it is easy to distinguish the two types of rows of A^* even when noise is present, for example, by looking at the first entry of a row. To circumvent the issue raised by this A^* , we would like to combine the information from rows sums with that from each individual column. This motivates us to consider the following method called **RankScore**, which outperforms **RankSum** and yields consistent estimation.

For $A^* \in \mathbb{R}^{n \times m}$ and $i, i' \in [n]$, define

$$\Delta_{A^*}(i, i') = \max_{j \in [m]} (A_{i',j}^* - A_{i,j}^*) \vee \frac{1}{\sqrt{m}} \sum_{j=1}^m (A_{i',j}^* - A_{i,j}^*)$$

and define $\Delta_Y(i, i')$ analogously. The quantity $\Delta_{A^*}(i, i')$ measures the difference between row i and row i' of A^* by either the largest difference between two corresponding entries, or the difference between the row sums scaled by the effective noise level $m^{-1/2}$, whichever is larger. If the noisy version $\Delta_Y(i, i')$ is larger than some threshold τ , then with high probability row i of Y should be placed after row i' in the original order. The procedure **RankScore** aggregates the comparison results between all pairs of rows of Y as follows:

1. For each $i \in [n]$, define the score s_i of the i -th row of Y by

$$s_i = \sum_{l=1}^n \mathbb{1}(\Delta_Y(l, i) \geq 2\tau) \quad (2.11)$$

where $\tau := C\sigma\sqrt{\log(nm)}$ for some tuning constant C (see Section 2.5.4 for details).

2. Order the rows of Y so that their scores are increasing, with ties broken arbitrarily.

The score s_i is just the number of comparisons row i wins. Intuitively, rows with larger entries will win more comparisons and thus be placed after rows with smaller entries. Hence RankScore can be viewed as a variant of the classical counting-based method for ranking, Copeland's method [Cop51], with a counting rule designed specifically for the model under consideration.

The RankScore procedure recovers an order of the rows of Y , which leads to an estimator $\tilde{\Pi}$ of the permutation. Then we define $\tilde{A} \in \mathcal{S}^m$ so that $\tilde{\Pi}\tilde{A}$ is the projection of Y onto the convex cone $\tilde{\Pi}\mathcal{S}^m$.

To quantify the rate of estimation for the RankScore estimator $(\tilde{\Pi}, \tilde{A})$, we define a new quantity $R(A)$ for $A \in \mathcal{S}^m$ as follows:

$$R(A) = \frac{1}{n} \max_{\substack{\mathcal{I} \subset [n]^2 \\ |\mathcal{I}|=n}} \sum_{(i,j) \in \mathcal{I}} \left(\frac{\|A_{i,\cdot} - A_{j,\cdot}\|_2^2}{\|A_{i,\cdot} - A_{j,\cdot}\|_\infty^2} \wedge \frac{m\|A_{i,\cdot} - A_{j,\cdot}\|_2^2}{\|A_{i,\cdot} - A_{j,\cdot}\|_1^2} \right), \quad (2.12)$$

where the summand is understood to be 1 if the rows $A_{i,\cdot}$ and $A_{j,\cdot}$ are identical.

To understand what properties of A the quantity $R(A)$ captures, consider the difference between the rows $A_{i,\cdot}$ and $A_{j,\cdot}$, denoted by $u \in \mathbb{R}^m$. First, the quantity $\|u\|_2^2/\|u\|_\infty^2$ is small when u is sparse. We have $\|u\|_2^2/\|u\|_\infty^2 \geq 1$ with equality achieved when $\|u\|_0 = 1$. Second, the quantity $m\|u\|_2^2/\|u\|_1^2$ is small when u is dense. We have $m\|u\|_2^2/\|u\|_1^2 \geq 1$ with equality achieved when all entries of u are the same. In particular, it holds that $R(A) \geq 1$, and $R(A)$ is small when the differences between rows of A are either very sparse or very dense. For example, if A is the matrix in (2.10), then the difference between any two distinct rows is 1-sparse, so we have $R(A) = 1$. Another example is

$$A = \begin{bmatrix} \mathbf{0} \\ \mathbf{1} \end{bmatrix}, \quad (2.13)$$

where the lower $\lfloor n/2 \rfloor$ rows of A are all ones while the remaining entries are all zeros. For this matrix, the difference between any two distinct rows is the all ones vector, so again we have $R(A) = 1$.

Moreover, $\|u\|_2^2 \leq \|u\|_1\|u\|_\infty$ by Hölder's inequality, so $\frac{\|u\|_2^2}{\|u\|_\infty^2} \wedge \frac{m\|u\|_2^2}{\|u\|_1^2} \leq \sqrt{m}$ as the product of the two terms is no larger than m . The equality is achieved by $u = (1, \dots, 1, 0, \dots, 0)$ where the first \sqrt{m} entries are equal to one. Therefore we have

$$R(A) \in [1, \sqrt{m}]. \quad (2.14)$$

Roughly speaking, the quantity $R(A)$ is large if there exist $\Theta(n)$ pairs of rows for which the differences are \sqrt{m} -sparse. An example of such an A is the lower triangular matrix with all ones on the lower triangle. We can take pairs of rows that are \sqrt{m} positions apart, and their differences are exactly \sqrt{m} -sparse binary vectors. Thus we have $R(A) \asymp \sqrt{m}$.

Since RankScore makes use of entrywise differences between rows, together with the difference between row sums, we expect a better performance of RankScore when the

differences between rows of A^* are either very sparse or very dense, which is exactly what captured by the quantity $R(A^*)$. Therefore, it is natural that the estimator $(\tilde{\Pi}, \tilde{A})$ enjoys the following rate of estimation, characterized by $R(A^*)$ together with $K(A)$ defined in the previous section.

Theorem 2.3.1. *For $A^* \in \mathcal{S}^m$ and $Y = \Pi^* A^* + Z$, let $(\tilde{\Pi}, \tilde{A})$ be the estimator defined above using the RankScore procedure with threshold $\tau = 3\sigma\sqrt{(C+1)\log(nm)}$, $C > 0$. Then it holds that*

$$\begin{aligned} \frac{1}{nm} \|\tilde{\Pi}\tilde{A} - \Pi^* A^*\|_F^2 &\lesssim \min_{A \in \mathcal{S}^m} \left(\frac{1}{nm} \|A - A^*\|_F^2 + \sigma^2 \frac{K(A)}{nm} \log \frac{enm}{K(A)} \right) \\ &\quad + (C+1)\sigma^2 \frac{R(A^*) \log(nm)}{m}, \end{aligned}$$

with probability at least $1 - e^{-c(n+m)} - (nm)^{-C}$ for some constant $c > 0$.

The quantity $R(A^*)$ only depends on the matrix A^* . If $R(A^*)$ is bounded logarithmically, the estimator $(\tilde{\Pi}, \tilde{A})$ achieves the minimax rate up to logarithmic factors. In any case, $R(A^*) \leq \sqrt{m}$, so the estimator is still consistent with the permutation error (i.e. the last term) decaying at a rate $\tilde{O}(\frac{1}{\sqrt{m}})$. Furthermore, it is worth noting that $R(A^*)$ is not needed to construct $(\tilde{\Pi}, \tilde{A})$, so the estimator adapts to $R(A^*)$ automatically.

Remark 2.3.2. *In the same way that Theorem 2.2.3 follows from Theorem 2.2.1, we can deduce from Theorem 2.3.1 a global bound for the estimator $(\tilde{\Pi}, \tilde{A})$ which has rate*

$$\left(\frac{\sigma^2 V(A^*) \log n}{n} \right)^{2/3} + \sigma^2 \left(\frac{\log n}{n} + R(A^*) \frac{\log(nm)}{m} \right).$$

2.3.2 Simulations

We corroborate the theoretical results above with a numerical comparison between the RankSum and RankScore procedures.

Consider the model (2.1) with $A^* \in \mathcal{S}^m$ and assume without loss of generality that $\Pi^* = I_n$. For various $n \times m$ matrices A^* , we generate observations $Y = A^* + Z$ where entries of Z are i.i.d. standard Gaussian variables. The performance of the estimators given by RankScore and RankSum defined above is compared to the performance of the oracle \hat{A}^{oracle} defined by the projection of Y onto the cone \mathcal{S}^m . Note that we are not able to compute the LS estimator efficiently, so instead the oracle estimator is used as the benchmark. For the RankScore estimator we take $\tau = 6$. The curves are generated based on 30 equally spaced points on the base-10 logarithmic scale, and all results are averaged over 10 replications. The vertical axis represents the estimation error of an estimator $\hat{\Pi}\hat{A}$, measured by the sample mean of $\log_{10} \left(\frac{1}{nm} \|\hat{\Pi}\hat{A} - A^*\|_F^2 \right)$ unless otherwise specified.

We begin with a simple example for which we set $n = m$. For each $\alpha \in [0, 1]$, define a matrix $A^* = A^*(\alpha) \in \mathbb{R}^{n \times n}$ by $A_{i,j}^* = m^{(1-\alpha)/2}$ for $n/2 \leq i \leq n, 1 \leq j \leq m^\alpha$ and $A_{i,j}^* = 0$ otherwise. Note that A^* is an interpolation between the matrix in (2.10)

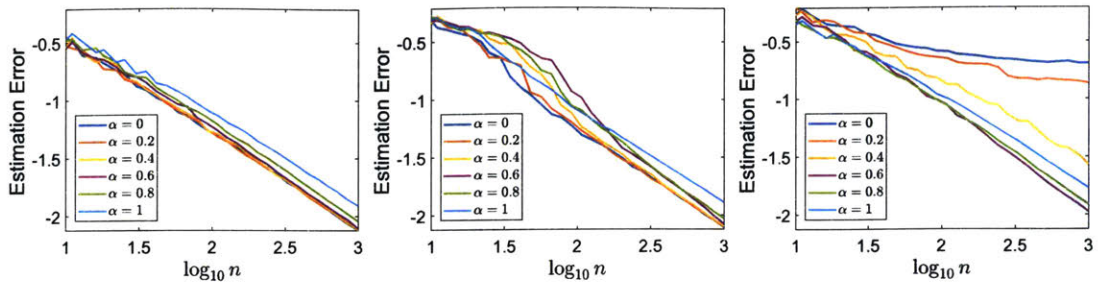


Figure 2-1: Estimation errors of the three estimators for $A^* = A^*(\alpha)$ where α ranges from 0 to 1. Left: the oracle estimator; Middle: the RankScore estimator; Right: the RankSum estimator.

(where $\alpha = 0$) and the matrix in (2.13) (where $\alpha = 1$). The nonzero rows of A^* have ℓ_2 -norm equal to \sqrt{m} for any $\alpha \in [0, 1]$.

In Figure 2-1, we plot the estimation errors of the oracle, RankScore and RankSum estimators for this A^* in the three plots respectively. As expected, RankSum has poor performance in estimating the true permutation when α is close to zero, because it fails to exploit the differences between rows along individual columns. When α is close to one, the weight of a nonzero row of A^* is distributed evenly across the columns, so it is appropriate to only consider row sums and thus RankSum behaves well. On the other hand, RankScore outperforms RankSum in recovering the permutation for any $\alpha \in [0, 1]$ when n is large, and it has roughly the same performance as the oracle. According to the discussion after (2.12), we have $R(A^*) = 1$ for $\alpha = 0$ or 1. Thus Theorem 2.3.1 predicts the fast rate, which is verified by the experiment. For α close to $1/2$, however, Theorem 2.3.1 only guarantees a rate $\tilde{O}(m^{-1/2})$ while the experiment suggests that RankScore still behaves as well as the oracle. Hence improving the adaptive bound in Theorem 2.3.1 remains an interesting problem for future research.

Note that the performance of each estimator for $\alpha = 0.6$ is slightly better than that for $\alpha = 1$. This is not inconsistent with our theoretical guarantees as the bounds we proved are up to logarithmic factors. Achieving sharper bounds to explain such a phenomenon also remains an interesting open question out of the scope of the present work.

In Figure 2-2, we compare the performance of RankScore to that of the oracle in three regimes of (n, m) . The matrices A^* are randomly generated for different values of n and m as follows. For the right plot, A^* is generated so that $V(A^*) \leq 1$, by sorting the columns of a matrix with i.i.d. $U(0, 1)$ entries. For the left plot, we further require that $K(A^*) = 5m$ by uniformly partitioning each column of A^* into five blocks and assigning each block the corresponding value from a sorted sample of five i.i.d. $U(0, 1)$ variables.

Since the oracle knows the true permutation, its behavior is independent of m , and its rates of estimation are bounded by $\frac{\log n}{n}$ for $K(A^*) = 5m$ and $(\frac{\log n}{n})^{\frac{2}{3}}$ for $V(A^*) = 1$ respectively by Theorem 2.2.1 and 2.2.3. (The difference is minor in the plots as n is not sufficiently large). For RankScore, the permutation term dominates

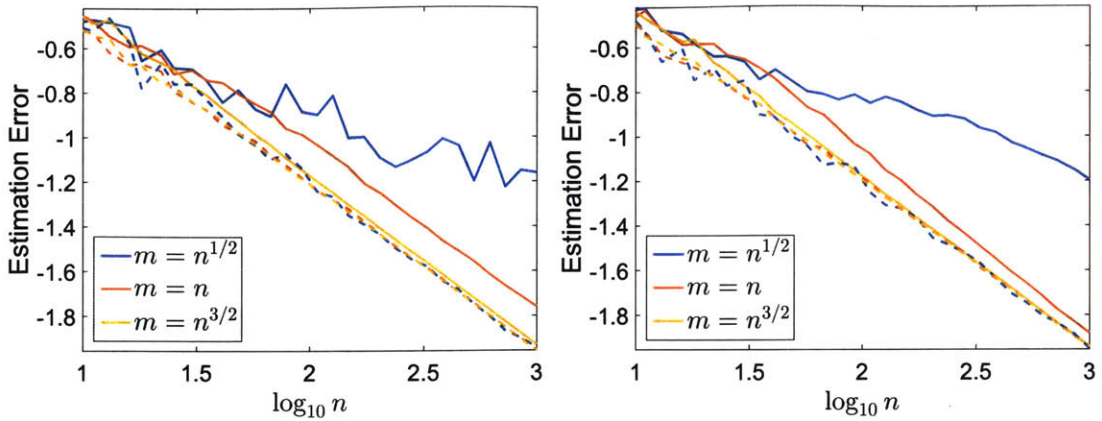


Figure 2-2: Estimation errors of the oracle (dashed lines) and RankScore (solid lines) for different regimes of (n, m) and randomly generated A^* of size $n \times m$. Left: $K(A^*) = 5m$; Right: $V(A^*) \leq 1$.

the estimation term when $m = n^{1/2}$ by Theorem 2.3.1. From the plots, the rates of estimation are better than $\tilde{O}(n^{-1/4})$ predicted by the worst-case analysis in both examples. For $m = n$, we also observe rates of estimation faster than the worst-case rate $\tilde{O}(n^{-1/2})$ and close to the oracle rates. We could explain this phenomenon by $R(A^*) < \sqrt{m}$, but such an interpretation may not be optimal since our analysis is based on worst-case deterministic A^* . Potential study of random designs of A^* is left open. Finally, for $m = n^{3/2}$, the permutation term is of order $\tilde{O}(n^{-3/4})$ theoretically, in between of the oracle rates for the two cases. Indeed RankScore has almost the same performance as the oracle experimentally. Overall Figure 2-2 illustrates the good behavior of RankScore in these random scenarios.

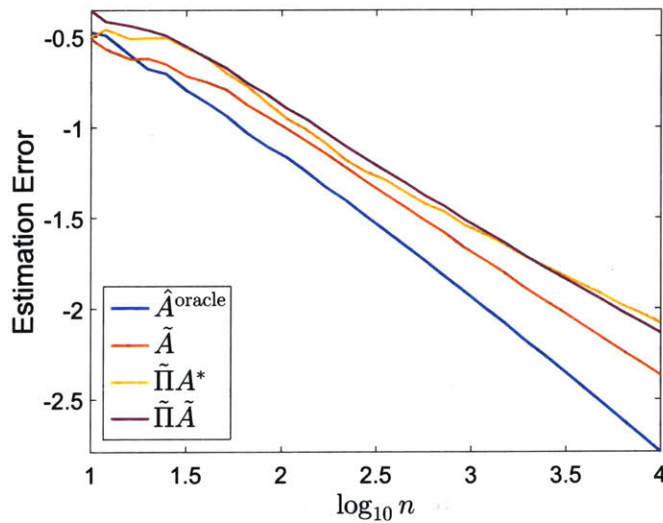


Figure 2-3: Various estimation errors of the oracle and RankScore for the triangular matrix.

To conclude our numerical experiments, we consider the $n \times n$ lower triangular matrix A^* defined by $A_{i,j}^* = \mathbb{1}(i \geq j)$. For this matrix, it is easy to check that $K(A^*) = 2n - 1$ and $R(A^*) \approx \sqrt{n}$. We plot in Figure 2-3 the estimation errors of $\tilde{\Pi}\tilde{A}$, $\tilde{\Pi}A^*$ and \tilde{A} given by RankScore, in addition to the oracle. By Theorem 2.3.1, the rate of estimation achieved by $\tilde{\Pi}\tilde{A}$ is of order $\tilde{O}(n^{-1/2})$, while that achieved by the oracle is of order $\tilde{O}(n^{-1})$ since there is no permutation term. The plot confirms this discrepancy. Moreover, $\frac{1}{n^2}\|\tilde{\Pi}A^* - A^*\|_F^2$ is an appropriate measure of the performance of $\tilde{\Pi}$ by Lemma 2.5.13 and 2.1.1, and the plot suggests that the rates of estimation achieved by $\tilde{\Pi}A^*$ and $\tilde{\Pi}\tilde{A}$ are about the same order. Finally \tilde{A} seems to have a slightly faster rate of estimation than $\tilde{\Pi}\tilde{A}$, so in practice \tilde{A} could be used to estimate A . However we refrain from making an explicit conjecture about the rate.

2.4 Unimodal regression

If the permutation in the main model (2.1) is known, then the estimation problem simply becomes a concatenation of m unimodal regressions. In fact, our proofs imply new oracle inequalities for unimodal regression. Recall that \mathcal{U} denotes the cone of unimodal vectors in \mathbb{R}^n . Suppose that we observe

$$y = \theta^* + z,$$

where $\theta^* \in \mathbb{R}^n$ and z is a sub-Gaussian vector with variance proxy σ^2 . Define the LS estimator $\hat{\theta}$ by

$$\hat{\theta} \in \underset{\theta \in \mathcal{U}}{\operatorname{argmin}} \|\theta - y\|_2^2.$$

Moreover let $k(\theta) = \operatorname{card}(\{\theta_1, \dots, \theta_n\})$ and $V(\theta) = \max_{i \in [n]} \theta_i - \min_{i \in [n]} \theta_i$.

Corollary 2.4.1. *There exists a constant $c > 0$ such that with probability at least $1 - n^{-\alpha}$, $\alpha \geq 1$,*

$$\frac{1}{n} \|\hat{\theta} - \theta^*\|_2^2 \lesssim \min_{\theta \in \mathcal{U}} \left(\frac{1}{n} \|\theta - \theta^*\|_2^2 + \sigma^2 \frac{k(\theta)}{n} \log \frac{en}{k(\theta)} \right) + \alpha \sigma^2 \frac{\log n}{n} \quad (2.15)$$

and

$$\frac{1}{n} \|\hat{\theta} - \theta^*\|_2^2 \lesssim \min_{\theta \in \mathcal{U}} \left[\frac{1}{n} \|\theta - \theta^*\|_2^2 + \left(\frac{\sigma^2 V(\theta) \log n}{n} \right)^{2/3} \right] + \alpha \sigma^2 \frac{\log n}{n}.$$

The corresponding bounds in expectation also hold.

The proof of Corollary 2.4.1 can be found in Section 2.7. Note that the bounds above match the minimax lower bounds for isotonic regression in [BT15] up to logarithmic factors. Since every monotone vector is unimodal, lower bounds for isotonic regression automatically hold for unimodal regression. Therefore, we have proved that the LS estimator is minimax optimal up to logarithmic factors for unimodal regression.

A result similar to (2.15) was obtained by Pierre C. Bellec in the revision of [Bel15] that was prepared independently and contemporaneously to our work [FMR16]. In

addition, Sabyasachi Chatterjee and John Lafferty also improved their bounds to having optimal exponents [CL15] after the first version of our work [FMR16] was posted. Interestingly Bellec employs bounds on the statistical dimension by leveraging results from [DA14], and Chatterjee and Lafferty use both the variational formula and the statistical dimension. Moreover, their results are presented in the well-specified case where $\theta^* \in \mathcal{U}$ and $\theta = \theta^*$.

2.5 Proofs

In this section, we provide the proofs of the main results.

2.5.1 Proof of the upper bounds

Before proving the main theorems, we discuss two methods adopted in recent works to bound the error of the LS estimator in shape constrained regression, in a general setting. Consider the least squares estimator $\hat{\theta}$ of the model $y = \theta^* + z$, where θ^* lies in a parameter space Θ and z is Gaussian noise. One way to study $\mathbb{E}\|\hat{\theta} - \theta^*\|_2^2$ is to use the *statistical dimension* [DA14] of a convex cone Θ defined by

$$\mathbb{E}\left[\left(\sup_{\theta \in \Theta, \|\theta\|_2 \leq 1} \langle \theta, z \rangle\right)^2\right].$$

This has been successfully applied to isotonic and more general shape constrained regression [CGS15, Bel15].

Another prominent approach is to express the error of the LS estimator via what is known as *Chatterjee's variational formula*, proved in [Cha14] and given by

$$\|\hat{\theta} - \theta^*\|_2 = \operatorname{argmax}_{t \geq 0} \left(\sup_{\theta \in \Theta, \|\theta - \theta^*\|_2 \leq t} \langle \theta - \theta^*, z \rangle - \frac{t^2}{2} \right). \quad (2.16)$$

Note that the first term is related to the *Gaussian width* (see, e.g., [CRPW12]) of Θ defined by $\mathbb{E}[\sup_{\theta \in \Theta} \langle \theta, z \rangle]$, whose connection to the statistical dimension was studied in [DA14]. The variational formula was first proposed for convex regression [Cha14], and later exploited in several different settings, including matrix estimation with shape constraints [CGS18] and unimodal regression [CL15]. Similar ideas have appeared in other works, for example, analysis of empirical risk minimization [Men15], ranking from pairwise comparison [SBGW17] and isotonic regression [Bel15]. In this latter work, Bellec has used the statistical dimension approach to prove spectacularly sharp oracle inequalities that seem to be currently out of reach for methods based on Chatterjee's variational formula (2.16). On the other hand, Chatterjee's variational formula seems more flexible as computations of the statistical dimension based on [DA14] are currently limited to convex sets Θ with a polyhedral structure. In this chapter, we use exclusively Chatterjee's variational formula.

A variational formula for the error of the LS estimator

We begin the proof by stating an extension of Chatterjee's variational formula. While we only need this lemma to hold for a union of closed convex sets we present a version that holds for all closed sets. The latter extension was suggested to us by Pierre C. Bellec in a private communication [Bel16].

Lemma 2.5.1. *Let \mathcal{C} be a closed subset of \mathbb{R}^d . Suppose that $y = a^* + z$ where $a^* \in \mathcal{C}$ and $z \in \mathbb{R}^d$. Let $\hat{a} \in \operatorname{argmin}_{a \in \mathcal{C}} \|y - a\|_2^2$ be a projection of y onto \mathcal{C} . Define the function $f_{a^*} : \mathbb{R}_+ \rightarrow \mathbb{R}$ by*

$$f_{a^*}(t) = \sup_{a \in \mathcal{C} \cap \mathcal{B}^d(a^*, t)} \langle a - a^*, z \rangle - \frac{t^2}{2}.$$

Then we have

$$\|\hat{a} - a^*\|_2 \in \operatorname{argmax}_{t \geq 0} f_{a^*}(t). \quad (2.17)$$

Moreover, if there exists $t^* > 0$ such that $f_{a^*}(t) < 0$ for all $t \geq t^*$, then $\|\hat{a} - a^*\|_2 \leq t^*$.

Proof. By definition,

$$\hat{a} \in \operatorname{argmin}_{a \in \mathcal{C}} \left(\|a - a^*\|_2^2 - 2\langle a - a^*, z \rangle + \|z\|_2^2 \right) = \operatorname{argmax}_{a \in \mathcal{C}} \left(\langle a - a^*, z \rangle - \frac{1}{2} \|a - a^*\|_2^2 \right).$$

Together with the definition of f_{a^*} , this implies that

$$\begin{aligned} f_{a^*}(\|\hat{a} - a^*\|_2) &\geq \langle \hat{a} - a^*, z \rangle - \frac{1}{2} \|\hat{a} - a^*\|_2^2 \\ &\geq \sup_{a \in \mathcal{C} \cap \mathcal{B}^d(a^*, t)} \left(\langle a - a^*, z \rangle - \frac{1}{2} \|a - a^*\|_2^2 \right) \\ &\geq \sup_{a \in \mathcal{C} \cap \mathcal{B}^d(a^*, t)} \langle a - a^*, z \rangle - \frac{t^2}{2} = f_{a^*}(t). \end{aligned}$$

Therefore (2.17) follows.

Furthermore, suppose that there is $t^* > 0$ such that $f_{a^*}(t) < 0$ for all $t \geq t^*$. Since $f_{a^*}(\|\hat{a} - a^*\|_2) \geq f_{a^*}(0) = 0$, we have $\|\hat{a} - a^*\|_2 \leq t^*$. \square

Note that this structural result holds for any error vector $z \in \mathbb{R}^d$ and any closed set \mathcal{C} which is not necessarily convex. In particular, this extends the results in [Cha14] and [CL15] which hold for convex sets and finite unions of convex sets respectively.

Proof of Theorem 2.2.1

For our purpose, we need a standard chaining bound on the supremum of a sub-Gaussian process that holds in high probability. The interested readers can find the proof, for example, in [vH14, Theorem 5.29], and refer to [LT91] for a more detailed account of the technique.

Lemma 2.5.2 (Chaining tail inequality). *Let $\Theta \subset \mathbb{R}^d$ and $z \sim \text{subG}(\sigma^2)$ in \mathbb{R}^d . For any $\theta_0 \in \Theta$, it holds that*

$$\sup_{\theta \in \Theta} \langle \theta - \theta_0, z \rangle \leq C\sigma \int_0^{\text{diam}(\Theta)} \sqrt{\log N(\Theta, \|\cdot\|_2, \varepsilon)} d\varepsilon + s$$

with probability at least $1 - C \exp(-\frac{cs^2}{\sigma^2 \text{diam}(\Theta)^2})$ where C and c are positive constants.

Let $\tilde{A} \in \mathcal{U}^m$. To lighten the notation, we define two rates of estimation:

$$R_1 = R_1(\tilde{A}, n) = \sigma \left(\sqrt{K(\tilde{A}) \log \frac{enm}{K(\tilde{A})}} + \sqrt{n \log n} \right) \quad (2.18)$$

and

$$R_2 = R_2(\tilde{A}, n) = \sigma^2 \left(K(\tilde{A}) \log \frac{enm}{K(\tilde{A})} + n \log n \right). \quad (2.19)$$

Note that $R_2 \leq R_1^2 \leq 2R_2$.

Lemma 2.5.3. *Suppose $Y = A^* + Z$ where $A^* \in \mathbb{R}^{n \times m}$ and $Z \sim \text{subG}(\sigma^2)$. For $\tilde{A} \in \mathcal{U}^m$ and all $t > 0$, define*

$$f_{\tilde{A}}(t) = \sup_{A \in \mathcal{M} \cap \mathcal{B}^{nm}(\tilde{A}, t)} \langle A - \tilde{A}, Y - \tilde{A} \rangle - \frac{t^2}{2}.$$

Then for any $s > 0$, it holds simultaneously for all $t > 0$ that

$$f_{\tilde{A}}(t) \leq CR_1 t + t \|A^* - \tilde{A}\|_F - \frac{t^2}{2} + st \quad (2.20)$$

with probability at least $1 - C \exp(-\frac{cs^2}{\sigma^2})$, where C and c are positive constants.

Proof. Define $\Theta = \Theta_{\mathcal{M}}(\tilde{A}, 1) = \bigcup_{\lambda \geq 0} \{B - \lambda \tilde{A} : B \in \mathcal{M} \cap \mathcal{B}^{nm}(\lambda \tilde{A}, 1)\}$ (see also Definition (2.24)). In particular, $\Theta \subset \mathcal{B}^{nm}(0, 1)$ and $0 \in \Theta$. Since \mathcal{M} is a finite union of convex cones and thus is star-shaped, by scaling invariance,

$$\sup_{A \in \mathcal{M} \cap \mathcal{B}^{nm}(\tilde{A}, t)} \langle A - \tilde{A}, Z \rangle = t \sup_{B \in \mathcal{M} \cap \mathcal{B}^{nm}(t^{-1}\tilde{A}, 1)} \langle B - t^{-1}\tilde{A}, Z \rangle \leq t \sup_{M \in \Theta} \langle M, Z \rangle.$$

By Lemma 2.5.2, with probability at least $1 - C \exp(-\frac{cs^2}{\sigma^2})$,

$$\sup_{M \in \Theta} \langle M, Z \rangle \leq C\sigma \int_0^2 \sqrt{\log N(\Theta, \|\cdot\|_F, \varepsilon)} d\varepsilon + s.$$

Moreover, it follows from Lemma 2.5.8 that

$$\log N(\Theta, \|\cdot\|_F, \varepsilon) \leq C\varepsilon^{-1} K(\tilde{A}) \log \frac{enm}{K(\tilde{A})} + n \log n.$$

Combining the previous three displays, we see that

$$\begin{aligned}
\sup_{A \in \mathcal{M} \cap \mathcal{B}^{nm}(\tilde{A}, t)} \langle A - \tilde{A}, Z \rangle &\leq C\sigma t \int_0^2 \sqrt{C\varepsilon^{-1}K(\tilde{A}) \log \frac{enm}{K(\tilde{A})} + n \log n} \, d\varepsilon + st \\
&\leq C\sigma t \sqrt{K(\tilde{A}) \log \frac{enm}{K(\tilde{A})}} + C\sigma t \sqrt{n \log n} + st \\
&= CR_1 t + st
\end{aligned}$$

with probability at least $1 - C \exp(-\frac{cs^2}{\sigma^2})$. Therefore

$$\begin{aligned}
f_{\tilde{A}}(t) &= \sup_{A \in \mathcal{M} \cap \mathcal{B}^{nm}(\tilde{A}, t)} \langle A - \tilde{A}, Y - \tilde{A} \rangle - \frac{t^2}{2} \\
&\leq \sup_{A \in \mathcal{M} \cap \mathcal{B}^{nm}(\tilde{A}, t)} \langle A - \tilde{A}, Z \rangle + \sup_{A \in \mathcal{M} \cap \mathcal{B}^{nm}(\tilde{A}, t)} \langle A - \tilde{A}, A^* - \tilde{A} \rangle - \frac{t^2}{2} \\
&\leq CR_1 t + st + t \|A^* - \tilde{A}\|_F - \frac{t^2}{2}
\end{aligned}$$

with probability at least $1 - C \exp(-\frac{cs^2}{\sigma^2})$ simultaneously for all $t > 0$. \square

We are now in a position to prove the adaptive oracle inequalities in Theorem 2.2.1. Recall that $(\hat{\Pi}, \hat{A})$ denotes the LS estimator defined in (2.2). Without loss of generality, assume that $\Pi^* = I_n$ and $Y = A^* + Z$.

Fix $\tilde{A} \in \mathcal{U}^m$ and define $f_{\tilde{A}}$ as in Lemma 2.5.3. We can apply Lemma 2.5.1 with $a^* = \tilde{A}$, $z = Y - \tilde{A}$, $y = Y$ and $\hat{a} = \hat{\Pi}\hat{A}$ to achieve an error bound on $\|\hat{\Pi}\hat{A} - \tilde{A}\|_F$, since $\hat{\Pi}\hat{A} \in \operatorname{argmin}_{M \in \mathcal{M}} \|Y - M\|_F^2$. To be more precise, for any $s > 0$ we define $t^* = 3C_1R_1 + 2\|A^* - \tilde{A}\|_F + 2s$ where C_1 is the constant in (2.20). Then it follows from Lemma 2.5.3 that with probability at least $1 - C \exp(-\frac{cs^2}{\sigma^2})$, it holds for all $t \geq t^*$ that

$$f_{\tilde{A}}(t) \leq C_1R_1t + t\|A^* - \tilde{A}\|_F - \frac{t^2}{2} + st < 0.$$

Therefore by Lemma 2.5.1,

$$\|\hat{\Pi}\hat{A} - \tilde{A}\|_F \leq t^* = 3C_1R_1 + 2\|A^* - \tilde{A}\|_F + 2s,$$

and thus

$$\|\hat{\Pi}\hat{A} - A^*\|_F \leq CR_1 + 3\|A^* - \tilde{A}\|_F + 2s \tag{2.21}$$

with probability at least $1 - C \exp(-\frac{cs^2}{\sigma^2})$.

In particular, if $s = R_1$, then $s \geq \sigma\sqrt{n+m}$ as $K(\tilde{A}) \geq m$. We see that with probability at least $1 - C \exp(-\frac{cs^2}{\sigma^2}) \geq 1 - e^{-c(n+m)}$,

$$\|\hat{\Pi}\hat{A} - A^*\|_F \lesssim R_1 + \|A^* - \tilde{A}\|_F$$

and thus

$$\|\hat{\Pi}\hat{A} - A^*\|_F^2 \lesssim \|A^* - \tilde{A}\|_F^2 + \sigma^2 K(\tilde{A}) \log \frac{enm}{K(\tilde{A})} + \sigma^2 n \log n.$$

Finally, (2.7) follows by taking the infimum over $\tilde{A} \in \mathcal{U}^m$ on the right-hand side and dividing both sides by nm .

Next, to prove the bound in expectation, observe that (2.21) yields

$$\mathbb{P}\left[\|\hat{\Pi}\hat{A} - A^*\|_F^2 - C(R_2 + \|A^* - \tilde{A}\|_F^2) \geq s\right] \leq C \exp\left(-\frac{cs}{\sigma^2}\right),$$

where R_2 is defined in (2.19). Integrating the tail probability, we get that

$$\mathbb{E}\|\hat{\Pi}\hat{A} - A^*\|_F^2 - C(R_2 + \|A^* - \tilde{A}\|_F^2) \lesssim \int_0^\infty \exp\left(-\frac{cs}{\sigma^2}\right) ds = \frac{\sigma^2}{c}$$

and therefore

$$\mathbb{E}\|\hat{\Pi}\hat{A} - A^*\|_F^2 \lesssim R_2 + \|A^* - \tilde{A}\|_F^2.$$

Dividing both sides by nm and minimizing over $\tilde{A} \in \mathcal{U}^m$ yields (2.8).

Proof of Theorem 2.2.3

In the setting of isotonic regression, [BT15] derived global bounds from adaptive bounds by a block approximation method, which also applies to our setting. The lemma below is a generalization of [BT15, Lemma 2] to the case of unimodal matrices.

For $k \in [n]$, let

$$\mathcal{U}_k = \{a \in \mathcal{U} : \text{card}(\{a_1, \dots, a_n\}) \leq k\}.$$

Define $k^* = \lceil \left(\frac{V(a)^2 n}{\sigma^2 \log(en)}\right)^{1/3} \rceil$. More generally, for $\mathbf{k} \in [n]^m$, we write $\mathbf{k} = (k_1, \dots, k_m)$ and let

$$\mathcal{U}_{\mathbf{k}}^m = \{A \in \mathcal{U}^m : \text{card}(\{A_{1,j}, \dots, A_{n,j}\}) = k_j \text{ for } 1 \leq j \leq m\}.$$

Then $K(A) = \sum_{j=1}^m k_j$ for $A \in \mathcal{U}_{\mathbf{k}}^m$. Define \mathbf{k}^* by

$$k_j^* = \lceil \left(\frac{V(A_{\cdot,j})^2 n}{\sigma^2 \log(en)}\right)^{1/3} \rceil.$$

Lemma 2.5.4. *For $A \in \mathcal{U}^m$, there exists $\tilde{A} \in \mathcal{U}_{\mathbf{k}^*}^m$ such that*

$$\frac{1}{nm} \|\tilde{A} - A\|_F^2 \leq \frac{1}{4} \left(\frac{\sigma^2 V(A) \log(en)}{n}\right)^{2/3} + \frac{\sigma^2}{4n} \log(en)$$

and

$$\frac{\sigma^2 K(\tilde{A})}{nm} \log(en) \leq 2 \left(\frac{\sigma^2 V(A) \log(en)}{n}\right)^{2/3} + \frac{2\sigma^2}{n} \log(en).$$

The proof of the lemma is provided in Section 2.7. To prove the theorem, for

$A \in \mathcal{U}^m$, choose $\tilde{A} \in \mathcal{U}_{\mathbf{k}^*}^m$ according to Lemma 2.5.4. Then

$$\begin{aligned} \frac{1}{nm} \|\tilde{A} - A^*\|_F^2 &\leq \frac{2}{nm} \|A - A^*\|_F^2 + \frac{2}{nm} \|\tilde{A} - A\|_F^2 \\ &\leq \frac{2}{nm} \|A - A^*\|_F^2 + \frac{5}{4} \left(\frac{\sigma^2 V(A) \log n}{n} \right)^{2/3} + \frac{5\sigma^2}{4n} \log n \end{aligned} \quad (2.22)$$

by noting that $\log(en) \leq 2.5 \log n$ for $n \geq 2$, and similarly

$$\frac{\sigma^2 K(\tilde{A})}{nm} \log(en) \leq 5 \left(\frac{\sigma^2 V(A) \log n}{n} \right)^{2/3} + \frac{5\sigma^2}{n} \log n. \quad (2.23)$$

Plugging (2.22) and (2.23) into the right-hand side of (2.7) and (2.8), and then minimizing over $A \in \mathcal{U}^m$, we complete the proof.

2.5.2 Metric entropy

This section is devoted to studying various *covering numbers* or *metric entropy* related to the parameter space of the model (2.1). The proofs of the lemmas in this section are provided in Section 2.7.

Recall that an ε -net of a subset $G \subset \mathbb{R}^n$ with respect to a norm $\|\cdot\|$ is a set $\{w_1, \dots, w_N\} \subset G$ such that for any $w \in G$, there exists $i \in [N]$ for which $\|w - w_i\| \leq \varepsilon$. The covering number $N(G, \|\cdot\|, \varepsilon)$ is the cardinality of the smallest ε -net with respect to the norm $\|\cdot\|$. Metric entropy is defined as the logarithm of a covering number. In the following, we will consider the Euclidean norm unless otherwise specified.

We start with a lemma bounding the metric entropy of a Cartesian product of convex cones. It is useful in later proofs and has its own interest. Let $\{I_i\}_{i=1}^m$ be a partition of $[n]$ with $|I_i| = n_i$ and $\sum_{i=1}^m n_i = n$. For $a \in \mathbb{R}^n$, the restriction of a to the coordinates in I_i is denoted by $a_{I_i} \in \mathbb{R}^{n_i}$. Let \mathcal{C}_i be a convex cone in \mathbb{R}^{n_i} and $\mathcal{C} = \mathcal{C}_1 \times \dots \times \mathcal{C}_m$.

Lemma 2.5.5. *With the notation above, suppose that $a_{I_i} \in \mathcal{C}_i \cap (-\mathcal{C}_i)$. Then for any $t > 0$ and $\varepsilon \in (0, t]$,*

$$\log N(\mathcal{C} \cap \mathcal{B}^n(a, t), \|\cdot\|_2, \varepsilon) \leq m \log \frac{Ct}{\varepsilon} + \sum_{i=1}^m \log N\left(\mathcal{C}_i \cap \mathcal{B}^{n_i}(a_{I_i}, t), \|\cdot\|_2, \frac{\varepsilon}{3}\right)$$

for some constant $C > 0$.

Recall that \mathcal{S}_n denotes the closed convex cone of increasing vectors in \mathbb{R}^n . First, we give a result on the metric entropy of \mathcal{S}_n intersecting with a ball.

Lemma 2.5.6. *Let $b \in \mathbb{R}^n$ be such that $b_1 = \dots = b_n$. Then for any $t > 0$ and $\varepsilon > 0$,*

$$\log N(\mathcal{S}_n \cap \mathcal{B}^n(b, t), \|\cdot\|_2, \varepsilon) \leq C\varepsilon^{-1}t \log(en).$$

Next, we study the metric entropy of the set of matrices with unimodal columns. Recall that $\mathcal{C}_l = \{a \in \mathbb{R}^n : a_1 \leq \dots \leq a_l\} \cap \{a \in \mathbb{R}^n : a_l \geq \dots \geq a_n\}$ for $l \in [n]$. For $\mathbf{l} = (l_1, \dots, l_m) \in [n]^m$, define $\mathcal{C}_1^m = \mathcal{C}_{l_1} \times \dots \times \mathcal{C}_{l_m}$. Moreover, for $A \in \mathbb{R}^{n \times m}$, $t > 0$ and $\mathcal{C} \subset \mathbb{R}^{n \times m}$, define

$$\begin{aligned} \Theta_{\mathcal{C}}(A, t) &= \bigcup_{\lambda \geq 0} \{B - \lambda A : B \in \mathcal{C} \cap \mathcal{B}^{nm}(\lambda A, t)\} \\ &= \bigcup_{\lambda \geq 0} (\mathcal{C} \cap \mathcal{B}^{nm}(\lambda A, t) - \lambda A). \end{aligned} \quad (2.24)$$

Note that in particular $\Theta_{\mathcal{C}}(A, t) \subset \mathcal{B}^{nm}(0, t)$.

Lemma 2.5.7. *Given $A \in \mathbb{R}^{n \times m}$ and $\mathbf{l} = (l_1, \dots, l_m) \in [n]^m$, we define the quantities $k(A_{\cdot j}) = \text{card}(\{A_{1,j}, \dots, A_{n,j}\})$ and $K(A) = \sum_{j=1}^m k(A_{\cdot j})$. Then for any $t > 0$ and $\varepsilon > 0$,*

$$\log N(\Theta_{\mathcal{C}_1^m}(A, t), \|\cdot\|_F, \varepsilon) \leq C\varepsilon^{-1}t K(A) \log \frac{enm}{K(A)}.$$

Finally, we consider the metric entropy of $\Theta_{\mathcal{M}}(A, t)$ for $A \in \mathbb{R}^{n \times m}$, $t > 0$ and $\mathcal{M} = \bigcup_{\Pi \in \mathfrak{S}_n} \Pi \mathcal{U}^m$. The above analysis culminates in the following lemma which we use to prove the main upper bounds.

Lemma 2.5.8. *Let $A \in \mathbb{R}^{n \times m}$ and $K(A)$ be defined as in the previous lemma. Then for any $\varepsilon > 0$ and $t > 0$,*

$$\log N(\Theta_{\mathcal{M}}(A, t), \|\cdot\|_F, \varepsilon) \leq C\varepsilon^{-1}t K(A) \log \frac{enm}{K(A)} + n \log n.$$

2.5.3 Proof of the lower bounds

For minimax lower bounds, we consider the model $Y = \Pi^* A^* + Z$ where entries of Z are i.i.d. $N(0, \sigma^2)$. Define $\mathcal{U}_{K_0}^m(V_0) = \mathcal{U}_{K_0}^m \cap \mathcal{U}^m(V_0)$ and $\mathcal{M}_{K_0}(V_0) = \bigcup_{\Pi \in \mathfrak{S}_n} \Pi \mathcal{U}_{K_0}^m(V_0)$. Define the subset of $\mathcal{M}_{K_0}(V_0)$ containing permutations of monotone matrices by $\mathcal{M}_{K_0}^S(V_0) = \{\Pi A \in \mathcal{M}_{K_0}(V_0) : \Pi \in \mathfrak{S}_n, A \in \mathcal{S}^m\}$. Since each estimator pair $(\hat{\Pi}, \hat{A})$ gives an estimator $\hat{M} = \hat{\Pi} \hat{A}$ of $M = \Pi A$, it suffices to prove a lower bound on $\|\hat{M} - M\|_F^2$. In fact, we prove a lower bound stronger than the one in Theorem 2.2.5. The proofs of the lemmas below can be found in Section 2.7.

Proposition 2.5.9. *Suppose that $K_0 \leq m(\frac{16n}{\sigma^2})^{1/3} V_0^{2/3} - m$. Then*

$$\begin{aligned} \inf_M \sup_{M \in \mathcal{M}_{K_0}(V_0)} \mathbb{P}_M \left[\frac{1}{nm} \|\hat{M} - M\|_F^2 \geq c\sigma^2 \frac{K_0}{nm} \right. \\ \left. + c \max_{1 \leq l \leq \min(K_0 - m, m) + 1} \min \left(\frac{\sigma^2}{m} \log l, m^2 l^{-3} V_0^2 \right) \right] \geq c' \end{aligned} \quad (2.25)$$

for some $c, c' > 0$, where \mathbb{P}_M is the probability with respect to $Y = M + Z$. This bound remains valid for the parameter subset $\mathcal{M}_{K_0}^S(V_0)$ if $l = 1$ or 2.

Note that the bound also holds for the larger parameter set $\mathcal{M}_{K_0} = \bigcup_{\Pi \in \mathfrak{S}_n} \Pi \mathcal{U}_{K_0}^m$. By taking $l = \min(K_0 - m, m) + 1$ and V_0 large enough, we see that the assumption in Proposition 2.5.9 is satisfied and the second term becomes simply $\frac{\sigma^2}{m} \log l$, so Theorem 2.2.5 follows. In the monotonic case, by the last statement of the proposition, if $K_0 \geq m + 1$ then taking $l = 2$ and V_0 large enough yields a lower bound of rate $\sigma^2(\frac{K_0}{nm} + \frac{1}{m})$ for the set of matrices A with increasing columns and $K(A) \leq K_0$.

The proof of Proposition 2.5.9 has two parts which correspond to the two terms respectively. First, the term $\sigma^2 \frac{K_0}{nm}$ is derived from the proof of lower bounds for isotonic regression in [BT15]. Then we derive the other term $\frac{\sigma^2}{m} \log l$ for any $1 \leq l \leq \min(K_0 - m, m) + 1$, which is due to the unknown permutation.

Lemma 2.5.10. *Suppose that $K_0 \leq m(\frac{16n}{\sigma^2})^{1/3} V_0^{2/3} - m$. For some $c, c' > 0$,*

$$\inf_{\hat{M}} \sup_{M \in \mathcal{M}_{K_0}^S(V_0)} \mathbb{P}_M \left[\|\hat{M} - M\|_F^2 \geq c\sigma^2 K_0 \right] \geq c,$$

where \mathbb{P}_M is the probability with respect to $Y = M + Z$.

For the second term in (2.25), we first note that the bound is trivial for $l = 1$ since $\log l = 0$. The next lemma deals with the case $l = 2$.

Lemma 2.5.11. *There exist constants $c, c' > 0$ such that for any $K_0 \geq m + 1$ and $V_0 \geq 0$,*

$$\inf_{\hat{M}} \sup_{M \in \mathcal{M}_{K_0}^S(V_0)} \mathbb{P}_M \left[\|\hat{M} - M\|_F^2 \geq cn \min(\sigma^2, m^3 V_0^2) \right] \geq c',$$

where \mathbb{P}_M is the probability with respect to $Y = M + Z$.

For the previous two lemmas, we have only used matrices with increasing columns. However, to achieve the second term in (2.25) for $l \geq 3$, we need matrices with unimodal columns.

Lemma 2.5.12. *There exist constants $c, c' > 0$ such that for any $K_0 \geq m$, $V_0 \geq 0$ and $3 \leq l \leq \min(K_0 - m, m) + 1$,*

$$\inf_{\hat{M}} \sup_{M \in \mathcal{M}_{K_0}(V_0)} \mathbb{P}_M \left[\|\hat{M} - M\|_F^2 \geq cn \min(\sigma^2 \log l, m^3 l^{-3} V_0^2) \right] \geq c',$$

where \mathbb{P}_M is the probability with respect to $Y = M + Z$.

Proof of Proposition 2.5.9. Combining Lemma 2.5.10, 2.5.11 and 2.5.12, and then dividing the bound by nm , we get (2.25) because the max of two terms is lower bounded by a half of their sum. The last statement in Proposition 2.5.9 holds since Lemma 2.5.10 and 2.5.11 are proved for matrices with increasing columns. \square

Furthermore, the proof of Theorem 2.2.6, provided in Section 2.7, only uses Lemma 2.5.10 and 2.5.11, so the lower bound of rate $(\frac{\sigma^2 V_0}{n})^{2/3} + \frac{\sigma^2}{n} + \min(\frac{\sigma^2}{m}, m^2 V_0^2)$ holds even if the matrices are required to have increasing columns.

2.5.4 Matrices with increasing columns

For the model $Y = \Pi^* A^* + Z$ where $A^* \in \mathcal{S}^m$ and $Z \sim \text{subG}(\sigma^2)$, a computationally efficient estimator $(\tilde{\Pi}, \tilde{A})$ has been constructed in Section 2.3 using the RankScore procedure. We will bound its rate of estimation in this section. Recall that the definition of $(\tilde{\Pi}, \tilde{A})$ consists of two steps. First, we recover an order (or a ranking) of the rows of Y , which leads to an estimator $\tilde{\Pi}$ of the permutation. Then define $\tilde{A} \in \mathcal{S}^m$ so that $\tilde{\Pi}\tilde{A}$ is the projection of Y onto the convex cone $\tilde{\Pi}\mathcal{S}^m$. For the analysis of the algorithm, we deal with the projection step first, and then turn to learning the permutation. The proofs of the results in the section can be found in Section 2.7.

In fact, for *any* estimator $\tilde{\Pi}$, if \tilde{A} is defined as above by the projection corresponding to $\tilde{\Pi}$, then the error $\|\tilde{\Pi}\tilde{A} - \Pi^* A^*\|_F^2$ can be split into two parts: the permutation error $\|(\tilde{\Pi} - \Pi^*)A^*\|_F^2$ and the estimation error of order $\tilde{O}(\sigma^2 K(A^*))$.

Lemma 2.5.13. *Consider the model $Y = \Pi^* A^* + Z$ where $A^* \in \mathcal{S}^m$ and $Z \sim \text{subG}(\sigma^2)$. For any $\tilde{\Pi} \in \mathfrak{S}_n$, define $\tilde{A} \in \mathcal{S}^m$ so that $\tilde{\Pi}\tilde{A}$ is the projection of Y onto $\tilde{\Pi}\mathcal{S}^m$. Then with probability at least $1 - e^{-c(n+m)}$, it holds simultaneously for all $\tilde{\Pi} \in \mathfrak{S}_n$ that*

$$\|\tilde{\Pi}\tilde{A} - \Pi^* A^*\|_F^2 \lesssim \min_{A \in \mathcal{S}^m} \left(\|A - A^*\|_F^2 + \sigma^2 K(A) \log \frac{enm}{K(A)} \right) + \sigma^2 n \log n + \|(\tilde{\Pi} - \Pi^*)A^*\|_F^2.$$

The idea of splitting the error into two terms as in Lemma 2.5.13 has appeared in [SBGW17, CM16].

By virtue of Lemma 2.5.13, it remains to control the permutation error $\|\tilde{\Pi}A^* - \Pi^* A^*\|_F^2$ where $\tilde{\Pi}$ is given by the RankScore procedure defined in Section 2.3. Recall that

$$\Delta_{A^*}(i, i') = \max_{j \in [m]} (A_{i',j}^* - A_{i,j}^*) \vee \frac{1}{\sqrt{m}} \sum_{j=1}^m (A_{i',j}^* - A_{i,j}^*)$$

for $i, i' \in [n]$ and $\Delta_Y(i, i')$ is defined analogously. Since columns of A^* are increasing,

$$|\Delta_{A^*}(i, i')| = \|A_{i',\cdot}^* - A_{i,\cdot}^*\|_\infty \vee \frac{1}{\sqrt{m}} \|A_{i',\cdot}^* - A_{i,\cdot}^*\|_1. \quad (2.26)$$

Recall that the RankScore procedure is defined as follows. First, for $i \in [n]$, we associate with the i -th row of Y a score s_i defined by $s_i = \sum_{l=1}^n \mathbb{1}(\Delta_Y(l, i) \geq 2\tau)$ for the threshold $\tau := 3\sigma \sqrt{\log(nm\delta^{-1})}$ where δ is the probability of failure. Then we order the rows of Y so that the scores are increasing with ties broken arbitrarily. This is equivalent to requiring that the corresponding permutation $\tilde{\pi} : [n] \rightarrow [n]$ satisfies that if $s_i < s_{i'}$ then $\tilde{\pi}^{-1}(i) < \tilde{\pi}^{-1}(i')$. Define $\tilde{\Pi}$ to be the $n \times n$ permutation matrix corresponding to $\tilde{\pi}$ so that $\tilde{\Pi}_{\tilde{\pi}(i),i} = 1$ for $i \in [n]$ and all other entries of $\tilde{\Pi}$ are zero. Moreover, let $\pi^* : [n] \rightarrow [n]$ be the permutation corresponding to Π^* .

To control the permutation error, we first state a lemma which asserts that if the gap between two rows of A^* is sufficiently large, then the permutation defined above will recover their relative order with high probability.

Lemma 2.5.14. *There is an event \mathcal{E} of probability at least $1 - \delta$ on which the following holds. For any $i, i' \in [n]$, if $\Delta_{A^*}(i, i') \geq 4\tau$, then $\tilde{\pi}^{-1} \circ \pi^*(i) < \tilde{\pi}^{-1} \circ \pi^*(i')$.*

Equipped with the above lemma, we are able to bound the permutation error in terms of the quantity $R(A^*)$ defined in (2.12).

Lemma 2.5.15. *There is an event \mathcal{E} of probability at least $1 - \delta$ on which*

$$\|\tilde{\Pi}A^* - \Pi^*A^*\|_F^2 \lesssim \sigma^2 R(A^*) n \log(nm\delta^{-1}).$$

Finally, the bound of Theorem 2.3.1 is an immediate consequence of Lemma 2.5.13 and Lemma 2.5.15 with $\delta = (nm)^{-C}$ for $C > 0$.

2.6 Discussion

While computational aspects of the seriation problem have received significant attention, the robustness of this problem to noise was still unknown to date. To overcome this limitation, we have introduced in this chapter the statistical seriation model and studied optimal rates of estimation by showing, in particular, that the least squares estimator enjoys several desirable statistical properties such as adaptivity and minimax optimality (up to logarithmic terms).

While this work paints a fairly complete statistical picture of the statistical seriation model, it also leaves many unanswered questions. There are several logarithmic gaps in the bounds. In the case of adaptive bounds, some logarithmic terms are unavoidable as illustrated by Theorem 2.2.5 (for the permutation term) and also by statistical dimension consideration explained in [Bel15] (for the estimation term). However, a more refined argument for the uniform bound, namely one that uses covering in ℓ_2 -norm rather than ℓ_∞ -norm, would allow us to remove the $\log n$ factor from the estimation term in the upper bound of Corollary 2.2.4. Such an argument can be found in [BS67, ABG⁺79, vdG91] for the larger class of vectors with bounded total variation (see [MvdG97]) but we do not pursue sharp logarithmic terms in this work. For the permutation term, $\log n$ in the upper bound of Corollary 2.2.2 and $\log l$ in the lower bound of Theorem 2.2.5 do not match if $l < n$. We do not seek answers to these questions in this chapter but note that their answers may be different for the unimodal and the monotone case.

Perhaps the most pressing question is that of computationally efficient estimators. Indeed, while statistically optimal, the least squares estimator requires searching through $n!$ permutations, which is not realistic even for problems of moderate size, let alone genomics applications. We gave a partial answer to this question in the specific context of monotone columns by proposing and studying the performance of a simple and efficient estimator called RankScore. This study reveals the existence of a potentially intrinsic gap between the statistical performance achievable by efficient estimators and that achievable by estimators with access to unbounded computation. A similar gap is also observed in the SST model for pairwise comparisons [SBGW17]. We conjecture that achieving optimal rates of estimation in the

seriation model is computationally hard in general but argue that the planted clique assumption that has been successfully used to establish statistical vs. computational gaps in [BR13, MW15, SBW16b] for example, is not the correct primitive. Instead, one has to seek for a primitive where hardness comes from searching through permutations rather than subsets.

2.7 Additional proofs

2.7.1 Proof of Proposition 2.2.7

Recall that $V(A) = (\frac{1}{m} \sum_{j=1}^m V_j(A)^{2/3})^{3/2}$. Since the ℓ_2 -norm of a vector is no larger than the $\ell_{\frac{2}{3}}$ -norm,

$$\sum_{j=1}^m V_j(A)^2 \leq \left(\sum_{j=1}^m V_j(A)^{2/3} \right)^3 = m^3 V(A)^2.$$

On the other hand,

$$\hat{A}_{i,j} = \frac{1}{n} \sum_{k=1}^n A_{k,j}^* + \frac{1}{n} \sum_{k=1}^n Z_{k,j},$$

so we have that

$$\begin{aligned} & \|\hat{\Pi}\hat{A} - \Pi^*A^*\|_F^2 \\ &= \sum_{i \in [n], j \in [m]} \left(\frac{1}{n} \sum_{k=1}^n A_{k,j}^* + \frac{1}{n} \sum_{k=1}^n Z_{k,j} - A_{i,j}^* \right)^2 \\ &\leq 2 \sum_{i \in [n], j \in [m]} \left(\frac{1}{n} \sum_{k=1}^n A_{k,j}^* - A_{i,j}^* \right)^2 + \frac{2}{n^2} \sum_{i \in [n], j \in [m]} \left(\sum_{k=1}^n Z_{k,j} \right)^2 \\ &\leq 2n \sum_{j \in [m]} V_j(A)^2 + \frac{2}{n} \sum_{j \in [m]} \left(\sum_{k=1}^n Z_{k,j} \right)^2 \\ &\leq 2nm^3 V(A)^2 + 2 \sum_{j \in [m]} g_j^2, \end{aligned}$$

where $g_j = \frac{1}{\sqrt{n}} \sum_{k=1}^n Z_{k,j}$ for $j \in [m]$ so that g_1, \dots, g_m are centered sub-Gaussian variables with variance proxy σ^2 . It is well-known that $\mathbb{E}g_j^2 \lesssim \sigma^2$, so

$$\mathbb{E}\|\hat{\Pi}\hat{A} - \Pi^*A^*\|_F^2 \lesssim nm^3 V(A)^2 + m\sigma^2.$$

Moreover, since (g_1, \dots, g_m) is a sub-Gaussian vector with variance proxy σ^2 , it follows from [HKZ12, Theorem 2.1] that $\sum_{j=1}^m g_j^2 \lesssim \sigma^2 m$ with probability at least $1 - \exp(-m)$. On this event,

$$\|\hat{\Pi}\hat{A} - \Pi^*A^*\|_F^2 \lesssim nm^3 V(A)^2 + m\sigma^2.$$

Dividing the previous two displays by nm completes the proof.

2.7.2 Proof of Lemma 2.5.4

Lemma 2 of [BT15] and its proof extend to the unimodal case with minor modifications. We provide the proof here for completeness.

Lemma 2.7.1. *For $a \in \mathcal{U}$ and $k \in [n]$, there exists $\tilde{a} \in \mathcal{U}_k$ such that*

$$\frac{1}{\sqrt{n}} \|\tilde{a} - a\|_2 \leq \frac{V(a)}{2k}. \quad (2.27)$$

In particular, there exists $\tilde{a} \in \mathcal{U}_{k^}$ such that*

$$\frac{1}{n} \|\tilde{a} - a\|_2^2 \leq \frac{1}{4} \max \left(\left(\frac{\sigma^2 V(a) \log(en)}{n} \right)^{2/3}, \frac{\sigma^2 \log(en)}{n} \right).$$

Moreover,

$$\frac{\sigma^2 k^*}{n} \log(en) \leq 2 \max \left(\left(\frac{\sigma^2 V(a) \log(en)}{n} \right)^{2/3}, \frac{\sigma^2 \log(en)}{n} \right).$$

Proof. Let $\underline{a} = \min(a_1, a_n)$, $\bar{a} = \max_{i \in [n]} a_i$ and $i_0 \in \operatorname{argmax}_{i \in [n]} a_i$. For $j \in [k-1]$, consider the intervals

$$I_j = \left[\underline{a} + \frac{j-1}{k} V(a), \underline{a} + \frac{j}{k} V(a) \right],$$

and $I_k = \left[\underline{a} + \frac{k-1}{k} V(a), \bar{a} \right]$. Also for $j \in [k]$, let $J_j = \{i \in [n] : a_i \in I_j\}$. We define the vector $\tilde{a} \in \mathbb{R}^n$ by $\tilde{a}_i = \underline{a} + \frac{j-1/2}{k} V(a)$ for $i \in [n]$, where j is uniquely determined by $i \in I_j$. Since a is increasing on $\{1, \dots, i_0\}$ and decreasing $\{i_0, \dots, n\}$, so is \tilde{a} . Thus $\tilde{a} \in \mathcal{U}_k$. Moreover, $|\tilde{a}_i - a_i| \leq \frac{V(a)}{2k}$ for $i \in [n]$, which implies (2.27).

Next we prove the latter two assertions. Since $k^* = \lceil (\frac{V(a)^2 n}{\sigma^2 \log(en)})^{1/3} \rceil$, if $\tilde{a} \in \mathcal{U}_{k^*}$ and $k^* = 1$ then

$$\frac{1}{n} \|\tilde{a} - a\|_2^2 \leq \frac{V(a)^2}{4} \leq \frac{\sigma^2}{4n} \log(en)$$

and

$$\frac{\sigma^2 k^*}{n} \log(en) = \frac{\sigma^2}{n} \log(en).$$

On the other hand, if $k^* > 1$, then

$$\frac{1}{n} \|\tilde{a} - a\|_2^2 \leq \frac{V(a)}{4(k^*)^2} \leq \frac{1}{4} \left(\frac{\sigma^2 V(a) \log(en)}{n} \right)^{2/3}$$

and

$$\frac{\sigma^2 k^*}{n} \log(en) \leq 2 \left(\frac{\sigma^2 V(a) \log(en)}{n} \right)^{2/3}.$$

□

Lemma 2.5.4 is then an easy generalization of the previous lemma to the matrix case. Applying Lemma 2.7.1 to columns of A , we see that there exists $\tilde{A} \in \mathcal{U}_{k^*}^m$ such that

$$\frac{1}{n} \|\tilde{A}_{\cdot,j} - A_{\cdot,j}\|_2^2 \leq \frac{1}{4} \max \left(\left(\frac{\sigma^2 V(A_{\cdot,j}) \log(en)}{n} \right)^{2/3}, \frac{\sigma^2}{n} \log(en) \right)$$

and

$$\frac{\sigma^2 k_j^*}{n} \log(en) \leq 2 \max \left(\left(\frac{\sigma^2 V(A_{\cdot,j}) \log(en)}{n} \right)^{2/3}, \frac{\sigma^2}{n} \log(en) \right).$$

Summing over $1 \leq j \leq m$, we get that

$$\begin{aligned} \frac{1}{nm} \|\tilde{A} - A\|_F^2 &\leq \frac{1}{4m} \left(\frac{\sigma^2 \log(en)}{n} \right)^{2/3} \sum_{j=1}^m V(A_{\cdot,j})^{2/3} + \frac{\sigma^2 \log(en)}{4n} \\ &= \frac{1}{4} \left(\frac{\sigma^2 V(A) \log(en)}{n} \right)^{2/3} + \frac{\sigma^2}{4n} \log(en), \end{aligned}$$

and similarly

$$\frac{\sigma^2 K(\tilde{A})}{nm} \log(en) \leq 2 \left(\frac{\sigma^2 V(A) \log(en)}{n} \right)^{2/3} + \frac{2\sigma^2}{n} \log(en).$$

2.7.3 Proofs of lemmas in Section 2.5.2

We start with a result on the metric entropy of a ball in one norm in \mathbb{R}^m with respect to another norm. This result is well-known for certain pairs of norms (e.g. [Mas07, Lemma 7.14]), and we use the general version from [Wai17, Lemma 5.2].

Lemma 2.7.2. *Let $\|\cdot\|$ and $\|\cdot\|'$ be a pair of norms on \mathbb{R}^m . Let \mathcal{B} and \mathcal{B}' denote the unit balls in $\|\cdot\|$ and $\|\cdot\|'$ respectively. Then for any $\varepsilon > 0$, it holds that*

$$N(\mathcal{B}, \|\cdot\|', \varepsilon) \leq \frac{\text{vol}(\frac{2}{\varepsilon}\mathcal{B} + \mathcal{B}')}{\text{vol}(\mathcal{B}')},$$

where $\text{vol}(\cdot)$ denotes the volume of the argument. In particular, for any $\varepsilon \in (0, 1]$,

$$N\left(\mathcal{B}^m(0, 1), \|\cdot\|_\infty, \frac{\varepsilon}{\sqrt{m}}\right) \leq (C/\varepsilon)^m,$$

where $\mathcal{B}^m(0, 1)$ is the unit ball in the ℓ_2 -norm and C is a positive constant.

Proof. The proof for the general bound is a standard volume argument and can be found in [Wai17]. To prove the second bound, note that the ℓ_∞ unit ball is contained in the ℓ_2 ball of radius \sqrt{m}/ε . Hence the general bound with $\|\cdot\| = \|\cdot\|_2$ and $\|\cdot\|' = \|\cdot\|_\infty$ implies that

$$N\left(\mathcal{B}^m(0, 1), \|\cdot\|_\infty, \frac{\varepsilon}{\sqrt{m}}\right) \leq \frac{\text{vol}(\frac{3\sqrt{m}}{\varepsilon}\mathcal{B})}{\text{vol}(\mathcal{B}')} \leq (C/\varepsilon)^m,$$

where the second inequality follows from the asymptotic formula for the volume of an Euclidean ball in \mathbb{R}^m : $\text{vol}(r\mathcal{B}) \sim \frac{1}{\sqrt{\pi m}} \left(\frac{2\pi e}{m}\right)^{m/2} r^m$ for $r > 0$. \square

Proof of Lemma 2.5.5. Since a product of balls $\mathcal{B}^{n_1}(0, \frac{\varepsilon}{\sqrt{m}}) \times \cdots \times \mathcal{B}^{n_m}(0, \frac{\varepsilon}{\sqrt{m}})$ is contained in $\mathcal{B}^n(0, \varepsilon)$, one could try to cover $\mathcal{C} \cap \mathcal{B}^n(a, t)$ by such products of balls. It turns out that this yields an upper bound of order $m^{3/2}$, which is too loose for our purpose. Fortunately, the following argument corrects this dependency.

Without loss of generality, we assume that $t = 1$. We construct a 3ε -net of $\mathcal{C} \cap \mathcal{B}^n(a, 1)$ as follows. First, let $\mathcal{N}_{\mathcal{B}}$ be a minimal $\frac{\varepsilon}{2\sqrt{m}}$ -net of $\mathcal{B}^m(0, 1)$ with respect to the ℓ_∞ -norm. Define

$$\mathcal{N}_{\mathcal{D}} = \left\{ \mu \in \mathcal{N}_{\mathcal{B}} : \min_{i \in [m]} \mu_i \geq -\frac{1}{2\sqrt{m}} \right\}.$$

Note that $\mu_i + \frac{1}{\sqrt{m}} > 0$ for $\mu \in \mathcal{N}_{\mathcal{D}}$, and let \mathcal{N}_{μ_i} be a minimal $(\mu_i + \frac{1}{\sqrt{m}})\varepsilon$ -net of $\mathcal{C}_i \cap \mathcal{B}^{n_i}(a_{I_i}, \mu_i + \frac{1}{\sqrt{m}})$. Define $\mathcal{N}_\mu = \mathcal{N}_{\mu_1} \times \cdots \times \mathcal{N}_{\mu_m}$, i.e.,

$$\mathcal{N}_\mu = \{w \in \mathbb{R}^n : w = (w_{I_1}, \dots, w_{I_m}), w_{I_i} \in \mathcal{N}_{\mu_i}\}.$$

We claim that $\bigcup_{\mu \in \mathcal{N}_{\mathcal{D}}} \mathcal{N}_\mu$ is a 3ε -net of $\mathcal{C} \cap \mathcal{B}^n(a, 1)$.

Fix $v \in \mathcal{C} \cap \mathcal{B}^n(a, 1)$. Let $v_{I_i} \in \mathbb{R}^{n_i}$ be the restriction of v to the component space \mathbb{R}^{n_i} . Then $v_{I_i} \in \mathcal{C}_i$. Let $\lambda \in \mathbb{R}^m$ be defined by $\lambda_i = \|v_{I_i} - a_{I_i}\|_2$, so $\|\lambda\|_2 = \|v - a\|_2 \leq 1$. Hence we can find $\mu \in \mathcal{N}_{\mathcal{B}}$ such that $\|\mu - \lambda\|_\infty \leq \frac{\varepsilon}{2\sqrt{m}}$. In particular, for all $i \in [m]$, $\mu_i \geq \lambda_i - \frac{\varepsilon}{2\sqrt{m}} \geq -\frac{1}{2\sqrt{m}}$, so $\mu \in \mathcal{N}_{\mathcal{D}}$. Moreover, $\|v_{I_i} - a_{I_i}\|_2 = \lambda_i < \mu_i + \frac{1}{\sqrt{m}}$ and $v_{I_i} \in \mathcal{C}_i$, so by definition of \mathcal{N}_{μ_i} , there exists $w_{I_i} \in \mathcal{N}_{\mu_i}$ such that $\|w_{I_i} - v_{I_i}\|_2 \leq (\mu_i + \frac{1}{\sqrt{m}})\varepsilon$. Let $w = (w_{I_1}, \dots, w_{I_m}) \in \mathcal{N}_\mu$. Since

$$\sum_{i=1}^m \mu_i^2 \leq \sum_{i=1}^m (\lambda_i + |\lambda_i - \mu_i|)^2 \leq \sum_{i=1}^m 2\lambda_i^2 + \frac{\varepsilon^2}{2} \leq \frac{5}{2},$$

we conclude that

$$\|w - v\|_2^2 \leq \sum_{i=1}^m \left(\mu_i + \frac{1}{\sqrt{m}}\right)^2 \varepsilon^2 \leq 7\varepsilon^2.$$

Therefore $\bigcup_{\mu \in \mathcal{N}_{\mathcal{D}}} \mathcal{N}_\mu$ is a 3ε -net of $\mathcal{C} \cap \mathcal{B}^n(a, 1)$.

It remains to bound the cardinality of this net. By Lemma 2.7.2, $|\mathcal{N}_{\mathcal{D}}| \leq |\mathcal{N}_{\mathcal{B}}| \leq (C/\varepsilon)^m$. Moreover, recall that \mathcal{N}_{μ_i} is a $(\mu_i + \frac{1}{\sqrt{m}})\varepsilon$ -net of $\mathcal{C}_i \cap \mathcal{B}^{n_i}(a_{I_i}, \mu_i + \frac{1}{\sqrt{m}})$. Since $a_{I_i} \in \mathcal{C}_i \cap (-\mathcal{C}_i)$, for any $t > 0$, $\mathcal{C}_i \cap \mathcal{B}^{n_i}(a_{I_i}, t) = \{x + a_{I_i} : x \in \mathcal{C}_i \cap \mathcal{B}^{n_i}(0, t)\}$. Hence we can choose the net so that

$$\begin{aligned} |\mathcal{N}_{\mu_i}| &= N\left(\mathcal{C}_i \cap \mathcal{B}^{n_i}\left(0, \mu_i + \frac{1}{\sqrt{m}}\right), \|\cdot\|_2, \left(\mu_i + \frac{1}{\sqrt{m}}\right)\varepsilon\right) \\ &= N(\mathcal{C}_i \cap \mathcal{B}^{n_i}(0, 1), \|\cdot\|_2, \varepsilon) \\ &= N(\mathcal{C}_i \cap \mathcal{B}^{n_i}(a_{I_i}, 1), \|\cdot\|_2, \varepsilon). \end{aligned}$$

As $|\mathcal{N}_\mu| \leq \prod_{i=1}^m |\mathcal{N}_{\mu_i}|$, therefore

$$\left| \bigcup_{\mu \in \mathcal{N}_D} \mathcal{N}_\mu \right| \leq \left(\frac{C}{\varepsilon} \right)^m \prod_{i=1}^m N(\mathcal{C}_i \cap \mathcal{B}^{n_i}(a_{I_i}, 1), \|\cdot\|_2, \varepsilon).$$

Taking the logarithm completes the proof. \square

Proof of Lemma 2.5.6. Part of this proof is due to Lemma 5.1 in an old version of [CL15], but we improve their result by a factor $\sqrt{\log n}$ and provide the whole proof for completeness. The technique we employ here is similar to that in the proof of Lemma 2.5.5. Roughly speaking, we shall construct a net of the original set by carefully combining nets of sub-blocks of vectors in the original set.

The bound holds trivially if $\varepsilon > t$, since the left-hand side is zero. Hence we can assume that $\varepsilon \leq t$. We also assume that $b = 0$ since the set of interest is translation invariant. Moreover, we assume that $t = 1$ and n is an even integer for simplicity, as the proof extends easily to the case where $t > 0$ or n is odd. First, let $n' = n/2$ and $I = [n']$. Define $\mathcal{S}' = \{(a_1, \dots, a_{n'}) \in \mathbb{R}^{n'} : a \in \mathcal{S}_n \cap \mathcal{B}^n(0, 1)\}$. Note that by splitting the vectors into two halves and using symmetry we have

$$\log N(\mathcal{S}_n \cap \mathcal{B}^n(0, 1), \|\cdot\|_2, \varepsilon) \leq 2 \log N(\mathcal{S}', \|\cdot\|_2, \varepsilon/\sqrt{2}). \quad (2.28)$$

To construct a net of \mathcal{S}' , we introduce some notation. Let k be the smallest integer for which $2^k > n'$, so that in particular $k \leq \log_2 n \leq C \log(en)$. We partition I into k blocks $I_j = I \cap [2^{j-1}, 2^j)$ for $j \in [k]$ and let $m_j = |I_j|$. Define a norm $\|\cdot\|$ on \mathbb{R}^k by

$$\|\mu\| = \left(\sum_{j=1}^k 2^j \mu_j^2 \right)^{1/2} \quad (2.29)$$

for $\mu \in \mathbb{R}^k$. Let $\mathcal{B}_{\|\cdot\|}^k(\mu, r)$ denote a ball in the norm $\|\cdot\|$ in \mathbb{R}^k with radius $r > 0$ centered at μ . Note that $\|\cdot\|$ is simply a weighted ℓ_2 -norm, and a ball in $\|\cdot\|$ is an ellipsoid (or more precisely, the set bounded by an ellipsoid).

Let \mathcal{N}_ε be a minimal ε -net of $\mathcal{B}_{\|\cdot\|}^k(0, \sqrt{10}) \cap \mathbb{R}_{\geq 0}^k$ with respect to the norm $\|\cdot\|$, where $\mathbb{R}_{\geq 0}^k$ is the nonnegative orthant of \mathbb{R}^k . For each $\mu = (\mu_1, \dots, \mu_k) \in \mathcal{N}_\varepsilon$, let \mathcal{N}_{μ_j} be a minimal $\frac{\varepsilon}{\sqrt{k}}$ -net of $\mathcal{S}_{m_j} \cap [-\mu_j, \mu_j]^{m_j}$ with respect to the Euclidean distance. Then we define $\mathcal{N}_\mu = \mathcal{N}_{\mu_1} \times \dots \times \mathcal{N}_{\mu_k}$, and claim that $\bigcup_{\mu \in \mathcal{N}_\varepsilon} \mathcal{N}_\mu$ is a 2ε -net of \mathcal{S}' with respect to the Euclidean distance.

Fix $a \in \mathcal{S}'$ so that $a = \tilde{a}_I$ for some $\tilde{a} \in \mathcal{S}_n \cap \mathcal{B}^n(0, 1)$. For $j \in [k]$, let $\nu_j = \max_{i \in I_j} |a_i|$. For each block I_j where $j \geq 2$, the maximum ν_j is achieved either at the left boundary $i_1 = 2^{j-1}$ or the right boundary $i_2 = (2^j - 1) \wedge n'$. If we have $a_{i_1} \leq 0$ and $|a_{i_1}| = \nu_j$, then $|a_i| \geq \nu_j$ for all $i \leq i_1$ as a is increasing. Otherwise, we must have $a_{i_2} \geq 0$, and so $a_{i_2} = \nu_j$. In this case, $\tilde{a}_i \geq \nu_j$ for all $i \geq i_2$, and thus $\nu_j \leq 1/\sqrt{n - i_2 + 1} \leq \sqrt{1/n'}$ since $\|\tilde{a}\|_2 \leq 1$ and $i_2 \leq n' = n/2$. Combining the two cases, we obtain that $a_i^2 + 1/n' \geq \nu_j^2$ for any $i \in I_{j-1}$. Summing over all $i \in [n']$ yields

that

$$\sum_{j=2}^k 2^{j-2} \nu_j^2 \leq \sum_{j=2}^k \sum_{i \in I_{j-1}} (a_i^2 + 1/n') \leq 2.$$

Together with the trivial bound $\nu_1 \leq 1$, this implies $\sum_{j=1}^k 2^j \nu_j^2 \leq 10$, so we have that $\nu \in \mathcal{B}_{\|\cdot\|}^k(0, \sqrt{10}) \cap \mathbb{R}_{\geq 0}^k$. By the definition of \mathcal{N}_ε , there exists $\mu \in \mathcal{N}_\varepsilon$ such that $\|\mu - \nu\| \leq \varepsilon$. Moreover, define $a' \in \mathcal{S}'$ by $a'_i = (a_i \wedge \mu_j) \vee (-\mu_j)$ for any $i \in I_j$ where $j \in [k]$. Recall that $\nu_j = \max_{i \in I_j} |a_i|$, so it holds that

$$\|a' - a\|_2^2 = \sum_{j=1}^k \sum_{i \in I_j} (a'_i - a_i)^2 \leq \sum_{j=1}^k \sum_{i \in I_j} (\mu_j - \nu_j)^2 \leq \sum_{j=1}^k 2^{j-1} (\mu_j - \nu_j)^2 \leq \|\mu - \nu\|^2 \leq \varepsilon^2.$$

Note that $a'_{I_j} \in \mathcal{S}_{m_j} \cap [-\mu_j, \mu_j]^{m_j}$. By the definition of \mathcal{N}_{μ_j} , there exists a vector $x_{I_j} \in \mathcal{N}_{\mu_j}$ such that $\|x_{I_j} - a'_{I_j}\|_2 \leq \varepsilon/\sqrt{k}$. If we define $x \in \mathbb{R}^n$ by concatenating x_{I_j} for $j \in [k]$, then we have $x \in \mathcal{N}_\mu$ and $\|x - a'\|_2 \leq \varepsilon$. Finally, the triangle inequality gives that $\|x - a\|_2 \leq 2\varepsilon$, so $\bigcup_{\mu \in \mathcal{N}_\varepsilon} \mathcal{N}_\mu$ is indeed a 2ε -net of \mathcal{S}' .

It remains to bound the cardinality of $\bigcup_{\mu \in \mathcal{N}_\varepsilon} \mathcal{N}_\mu$. By the definition of \mathcal{N}_ε , its cardinality is bounded by $N(\mathcal{B}_{\|\cdot\|}^k(0, \sqrt{10}), \|\cdot\|, \varepsilon)$. Taking both norms in Lemma 2.7.2 to be the norm $\|\cdot\|$ defined in (2.29), we obtain that

$$\log |\mathcal{N}_\varepsilon| \leq \log N\left(\mathcal{B}_{\|\cdot\|}^k(0, \sqrt{10}), \|\cdot\|, \varepsilon\right) \leq \log \frac{\text{vol}(\mathcal{B}_{\|\cdot\|}^k(0, 9/\varepsilon))}{\text{vol}(\mathcal{B}_{\|\cdot\|}^k(0, 1))} = \frac{9}{\varepsilon} k, \quad (2.30)$$

because if we scale an ellipsoid in \mathbb{R}^k by ratio $r > 0$ then its volume scales as r^k . Next, we know from [Cha14, Lemma 4.20] that for any $d \geq c \geq 0$ and $n \geq 1$,

$$\log N(\mathcal{S}_n \cap [c, d]^n, \|\cdot\|_2, \varepsilon) \leq C\varepsilon^{-1} \sqrt{n}(d - c).$$

It follows that for any $\mu \in \mathcal{N}_\varepsilon$ and $j \in [k]$,

$$\log |\mathcal{N}_{\mu_j}| = \log N(\mathcal{S}_{m_j} \cap [-\mu_j, \mu_j]^{m_j}, \|\cdot\|_2, \varepsilon/\sqrt{k}) \leq C\varepsilon^{-1} \sqrt{k} 2^{j/2} \mu_j.$$

Summing over $j \in [k]$ and applying the Cauchy-Schwarz inequality, we get that

$$\log |\mathcal{N}_\mu| \leq \sum_{j=1}^k \log |\mathcal{N}_{\mu_j}| \leq \frac{C}{\varepsilon} \sqrt{k} \sum_{j=1}^k 2^{j/2} \mu_j \leq \frac{C}{\varepsilon} k \left(\sum_{j=1}^k 2^j \mu_j^2 \right)^{1/2} \leq \frac{C}{\varepsilon} k, \quad (2.31)$$

where the last inequality holds because $\|\mu\| \leq \sqrt{10}$ by definition. Since $k \leq C \log(en)$, (2.30) and (2.31) imply that $\log |\bigcup_{\mu \in \mathcal{N}_\varepsilon} \mathcal{N}_\mu| \leq C\varepsilon^{-1} \log(en)$. We complete the proof by combining this bound with (2.28). \square

Proof of Lemma 2.5.7. Assume that $\varepsilon \leq t$ since otherwise the left-hand side is zero and the bound holds trivially. For $j \in [m]$, define $I^{j,1} = [l_j]$ and $I^{j,2} = [n] \setminus [l_j]$. Define $k_{j,1} = k(A_{I^{j,1},j})$ and $k_{j,2} = k(A_{I^{j,2},j})$. Let $\varkappa = \sum_{j=1}^m (k_{j,1} + k_{j,2})$ and observe

that $K(A) \leq \varkappa \leq 2K(A)$. Moreover, let $\{I_1^{j,1}, \dots, I_{k_{j,1}}^{j,1}\}$ be the partition of $I^{j,1}$ such that $A_{I_i^{j,1},j}$ is a constant vector for $i \in [k_{j,1}]$. Note that elements of $I_i^{j,1}$ need not to be consecutive. Define the partition for $I^{j,2}$ analogously.

For $j \in [m]$ and $i \in [k_{j,1}]$ (resp. $[k_{j,2}]$), let $\mathcal{S}_{I_i^{j,1},j}$ (resp. $\mathcal{S}_{I_i^{j,2},j}$) denote the set of increasing (resp. decreasing) vectors in the component space $\mathbb{R}^{|I_i^{j,1}|}$ (resp. $\mathbb{R}^{|I_i^{j,2}|}$). Lemma 2.5.6 implies that

$$\log N(\mathcal{S}_{I_i^{j,r},j} \cap \mathcal{B}^{|I_i^{j,r}|}(A_{I_i^{j,r},j}, t), \|\cdot\|_F, \varepsilon) \leq C\varepsilon^{-1}t \log(e|I_i^{j,r}|).$$

As a matrix in $\mathbb{R}^{n \times m}$ can be viewed as a concatenation of $\varkappa = \sum_{j=1}^m (k_{j,1} + k_{j,2})$ vectors of length $|I_i^{j,r}|$, $r \in [2]$, $j \in [m]$, we define the cone \mathcal{S}^* in $\mathbb{R}^{n \times m}$ by $\mathcal{S}^* = \prod_{j=1}^m \prod_{r=1}^2 \prod_{i=1}^{k_{j,r}} \mathcal{S}_{I_i^{j,r},j}$, which is clearly a superset of \mathcal{C}_1^m . It also follows that $A \in \mathcal{S}^* \cap (-\mathcal{S}^*)$, and thus by Lemma 2.5.5 and the previous display,

$$\begin{aligned} \log N(\mathcal{S}^* \cap \mathcal{B}^{nm}(A, t), \|\cdot\|_F, \varepsilon) &\leq \varkappa \log \frac{Ct}{\varepsilon} + \sum_{j=1}^m \sum_{r=1}^2 \sum_{i=1}^{k_{j,r}} C\varepsilon^{-1}t \log(e|I_i^{j,r}|) \\ &\leq C\varepsilon^{-1}t \varkappa + C\varepsilon^{-1}t \varkappa \log \frac{e^{\sum_{j,r,i} |I_i^{j,r}|}}{\varkappa} \\ &\leq C\varepsilon^{-1}t K(A) \log \frac{enm}{K(A)}, \end{aligned}$$

where we used the concavity of the logarithm and Jensen's inequality in the second step, and that $K(A) \leq \varkappa \leq 2K(A)$ in the last step.

Since $A \in \mathcal{S}^* \cap (-\mathcal{S}^*)$ (the cone \mathcal{S}^* is pointed at A) we have that $\mathcal{S}^* \cap \mathcal{B}^{nm}(\lambda A, t) - \lambda A = \mathcal{S}^* \cap \mathcal{B}^{nm}(0, t)$ for any $\lambda \geq 0$. In view of the definition of Θ , it holds

$$\Theta_{\mathcal{S}^*}(A, t) = \bigcup_{\lambda \geq 0} \mathcal{S}^* \cap \mathcal{B}^{nm}(\lambda A, t) - \lambda A = \mathcal{S}^* \cap \mathcal{B}^{nm}(0, t) - \lambda A, \quad \forall \lambda \geq 0.$$

In particular, taking $\lambda = 1$, we get $\Theta_{\mathcal{S}^*}(A, t) = \mathcal{S}^* \cap \mathcal{B}^{nm}(A, t) - A$. Moreover, $\mathcal{C}_1^m \subset \mathcal{S}^*$, so that $\Theta_{\mathcal{C}_1^m}(A, t) \subset \Theta_{\mathcal{S}^*}(A, t) = \mathcal{S}^* \cap \mathcal{B}^{nm}(A, t) - A$. Thus the metric entropy of $\Theta_{\mathcal{C}_1^m}(A, t)$ is subject to the above bound as well. \square

Proof of Lemma 2.5.8. Assume that $\varepsilon \leq t$ since otherwise the left-hand side is zero and the bound holds trivially. Note that $\mathcal{U}^m = \bigcup_{I \in [n]^m} \mathcal{C}_1^m$, and $\mathcal{M} = \bigcup_{\Pi \in \mathfrak{S}_n} \Pi \mathcal{U}^m$. Thus \mathcal{M} is the union of $n^m n!$ cones of the form $\Pi \mathcal{C}_1^m$. By definition, $\Theta_{\mathcal{M}}(A, t)$ is also the union of $n^m n!$ sets $\Theta_{\Pi \mathcal{C}_1^m}(A, t)$, each having metric entropy subject to the bound in Lemma 2.5.7. Therefore, a union bound implies that

$$\begin{aligned} \log N(\Theta_{\mathcal{M}}(A, t), \|\cdot\|_F, \varepsilon) &\leq \log N(\Theta_{\mathcal{C}_1^m}(A, t), \|\cdot\|_F, \varepsilon) + \log(n^m n!) \\ &\leq C\varepsilon^{-1}t K(A) \log \frac{enm}{K(A)} + m \log n + n \log n \\ &\leq C\varepsilon^{-1}t K(A) \log \frac{enm}{K(A)} + n \log n, \end{aligned}$$

where the last step follows from that $K \log(enm/K) \geq m \log n$ for $m \leq K \leq nm$ and that $\varepsilon \leq t$. \square

2.7.4 Proofs of lemmas in Section 2.5.3

The Varshamov-Gilbert lemma [Mas07, Lemma 4.7] is a standard tool for proving lower bounds.

Lemma 2.7.3 (Varshamov-Gilbert). *Let δ denote the Hamming distance on $\{0, 1\}^d$ where $d \geq 2$. Then there exists a subset $\Omega \subset \{0, 1\}^d$ such that $\log |\Omega| \geq d/8$ and $\delta(\omega, \omega') \geq d/4$ for distinct $\omega, \omega' \in \Omega$.*

We also need the following useful lemma.

Lemma 2.7.4. *Consider the model $y = \theta + z$ where $\theta \in \Theta \subset \mathbb{R}^d$ and $z \sim N(0, \sigma^2 I_d)$. Suppose that $|\Theta| \geq 3$ and for distinct $\theta, \theta' \in \Theta$, $4\phi \leq \|\theta - \theta'\|_2^2 \leq \frac{\sigma^2}{8} \log |\Theta|$ where $\phi > 0$. Then there exists $c > 0$ such that*

$$\inf_{\hat{\theta}} \sup_{\theta \in \Theta} \mathbb{P}_{\theta} [\|\hat{\theta} - \theta\|_F^2 \geq \phi] \geq c.$$

Proof. Let \mathbb{P}_{θ} denote the probability with respect to $\theta + z$. Then the Kullback-Leibler divergence between \mathbb{P}_{θ} and $\mathbb{P}_{\theta'}$ satisfies that

$$\text{KL}(\mathbb{P}_{\theta}, \mathbb{P}_{\theta'}) = \frac{\|\theta - \theta'\|_F^2}{2\sigma^2} \leq \frac{\log |\Theta|}{16} \leq \frac{\log(|\Theta| - 1)}{10},$$

since $|\Theta| \geq 3$. Applying [Tsy09, Theorem 2.5] with $\alpha = \frac{1}{10}$ gives the conclusion. \square

Proof of Lemma 2.5.10. We adapt the proof of [BT15, Theorem 4] to the case of matrices. Let $V_j = V_0$ for all $j \in [m]$. Since

$$K_0 \leq m \left(\frac{16n}{\sigma^2} \right)^{1/3} V_0^{2/3} - m = \sum_{j=1}^m \left[\left(\frac{16n}{\sigma^2} \right)^{1/3} V_j^{2/3} - 1 \right],$$

we can choose $k_j \in [n]$ so that $k_j \leq \left(\frac{16n}{\sigma^2} \right)^{1/3} V_j^{2/3}$ and $K_0 = \sum_{j=1}^m k_j$. According to Lemma 2.7.3, there exists $\Omega \subset \{0, 1\}^{K_0}$ such that $\log |\Omega| \geq K_0/8$ and $\delta(\omega, \omega') \geq K_0/4$ for distinct $\omega, \omega' \in \Omega$. Consider the partition $[K_0] = \cup_{m=1}^j I_j$ with $|I_j| = k_j$. For each $\omega \in \Omega$, let $\omega^j \in \{0, 1\}^{k_j}$ be the restriction of ω to coordinates in I_j . Define $M^{\omega} \in \mathbb{R}^{n \times m}$ by

$$M_{i,j}^{\omega} = \frac{\lfloor (i-1)k_j/n \rfloor V_j}{2k_j} + \gamma_j \omega_{\lfloor (i-1)k_j/n \rfloor + 1},$$

where $\gamma_j = \frac{\sigma}{8} \sqrt{k_j/2n}$. It is straightforward to check that $k(M_{\cdot,j}) \leq k_j$, $V(M_{\cdot,j}) \leq V_j$ and $M_{\cdot,j}$ is increasing, so M is in the parameter space. Moreover, for distinct $\omega, \omega' \in \Omega$,

$$\|M^{\omega} - M^{\omega'}\|_F^2 \geq c \sum_{j=1}^m \frac{n}{k_j} \gamma_j^2 \delta(\omega^j, (\omega')^j) \geq c\sigma^2 \sum_{j=1}^m \delta(\omega^j, (\omega')^j) = c\sigma^2 K_0.$$

On the other hand,

$$\|M^\omega - M^{\omega'}\|_F^2 \leq 2 \sum_{j=1}^m \frac{n}{k_j} \gamma_j^2 \delta(\omega^j, (\omega')^j) \leq \frac{\sigma^2}{64} \delta(\omega, \omega') \leq \frac{\sigma^2 K_0}{64} \leq \frac{\sigma^2}{8} \log |\Omega|.$$

Applying Lemma 2.7.4 completes the proof. \square

Proof of Lemma 2.5.11. By Lemma 2.7.3, there exists $\Omega \subset \{0, 1\}^n$ such that $\log |\Omega| \geq n/8$ and $\delta(\omega, \omega') \geq n/4$ for distinct $\omega, \omega' \in \Omega$. For each $\omega \in \Omega$, define $M^\omega \in \mathbb{R}^{n \times m}$ by setting the first column of M^ω to be $\alpha\omega$ and all other entries to be zero, where $\alpha = \min(\frac{\sigma}{8}, m^{3/2}V_0)$. Then

1. $M^\omega \in \mathcal{M}_{K_0}^S(V_0)$ since $K(M) = m + 1 \leq K_0$, $V(M) \leq V_0$ and we can permute the rows of M^ω so that its first column is increasing;
2. $\|M^\omega - M^{\omega'}\|_F^2 \geq \min(\frac{\sigma^2}{64}, m^3V_0^2) \delta(\omega, \omega') \geq \min(\frac{n\sigma^2}{256}, \frac{n}{4}m^3V_0^2)$ for distinct $\omega, \omega' \in \Omega$;
3. $\|M^\omega - M^{\omega'}\|_F^2 \leq \frac{\sigma^2}{64} \delta(\omega, \omega') \leq \frac{\sigma^2}{64} n \leq \frac{\sigma^2}{8} \log |\Omega|$ for $\omega, \omega' \in \Omega$.

Applying Lemma 2.7.4 completes the proof. \square

The following packing lemma is the key to the proof of Lemma 2.5.12.

Lemma 2.7.5. *For $l \in [m]$, consider the set \mathfrak{M} of $n \times m$ matrices of the form*

$$M = \begin{cases} 1 & \text{for exactly one } j_i \in [l] \text{ for each } i \in [n], \\ 0 & \text{otherwise.} \end{cases}$$

For $\varepsilon > 0$, define $k = \lfloor \frac{\varepsilon^2 n}{2} \rfloor$. Then there exists an $\varepsilon\sqrt{n}$ -packing \mathcal{P} of \mathfrak{M} such that $|\mathcal{P}| \geq l^{n-k} (\frac{k}{en})^k$ if $k \geq 1$ and $|\mathcal{P}| = l^n$ if $k = 0$.

Proof. There are l choices of entries to put the one in each row of M , so $|\mathfrak{M}| = l^n$. Fix $M_0 \in \mathfrak{M}$. If $\|M - M_0\|_F \leq \varepsilon\sqrt{n}$ where $M \in \mathfrak{M}$, then M differs from M_0 in at most k rows. If $k = 0$, taking $\mathcal{P} = \mathfrak{M}$ gives the result. If $k \geq 1$ then

$$|\mathfrak{M} \cap B^{nm}(M_0, \varepsilon\sqrt{n})| \leq \binom{n}{k} l^k \leq \left(\frac{en}{k}\right)^k l^k.$$

Moreover, let \mathcal{P} be a maximal $\varepsilon\sqrt{n}$ -packing of \mathfrak{M} . Then \mathcal{P} is also an $\varepsilon\sqrt{n}$ -net, so $\mathfrak{M} \subset \bigcup_{M_0 \in \mathcal{P}} B^{nm}(M_0, \varepsilon\sqrt{n})$. It follows that

$$l^n = |\mathfrak{M}| \leq \sum_{M_0 \in \mathcal{P}} |\mathfrak{M} \cap B^{nm}(M_0, \varepsilon\sqrt{n})| \leq |\mathcal{P}| \cdot \left(\frac{en}{k}\right)^k l^k.$$

We conclude that $|\mathcal{P}| \geq l^{n-k} (\frac{k}{en})^k$. \square

Proof of Lemma 2.5.12. For notational simplicity, we consider $2 \leq l \leq \min(K_0 - m, m)$ instead of $3 \leq l \leq \min(K_0 - m, m) + 1$.

Set $\varepsilon = 1/2$ and let \mathcal{P} be the $\sqrt{n}/2$ -packing given by Lemma 2.7.5. If $k = \lfloor \frac{n}{8} \rfloor = 0$, then $\log |\mathcal{P}| = n \log l$. Now assume that $k \geq 1$. Since $(\frac{x}{en})^x$ is decreasing on $[1, n]$, we have that $|\mathcal{P}| \geq l^{7n/8} (\frac{1}{8e})^{n/8}$. Hence for $l \geq 2$,

$$\log |\mathcal{P}| \geq \frac{7n}{8} \log l - \frac{n}{8} \log(8e) \geq \frac{n}{4} \log l. \quad (2.32)$$

Moreover, for each $M_0 \in \mathcal{P}$, consider the rescaled matrix

$$M = \min \left(\frac{\sigma}{8} \sqrt{\frac{\log l}{2}}, \left(\frac{m}{l} \right)^{3/2} V_0 \right) M_0.$$

1. We can permute the rows of M_0 so that each column has consecutive ones (or all zeros), so $M \in \mathcal{M}$. Moreover,

$$K(M) = 2l + m - l \leq \min(m, K_0 - m) + m \leq K_0$$

and

$$V(M) \leq \left(\frac{1}{m} \sum_{j=1}^l \left((m/l)^{3/2} V_0 \right)^{2/3} \right)^{3/2} = V_0,$$

so $M \in \mathcal{M}_{K_0}(V_0)$ for $M_0 \in \mathcal{P}$.

2. For $M_0, M'_0 \in \mathcal{P}$, $\|M_0 - M'_0\|_F^2 \geq n/4$, so

$$\begin{aligned} \|M - M'\|_F^2 &= \min \left(\frac{\sigma^2 \log l}{128}, (m/l)^3 V_0^2 \right) \|M_0 - M'_0\|_F^2 \\ &\geq \min \left(\frac{\sigma^2}{512} n \log l, \frac{n}{4} \left(\frac{m}{l} \right)^3 V_0^2 \right). \end{aligned}$$

3. For $M_0, M'_0 \in \mathcal{P}$, $\|M_0 - M'_0\|_F^2 \leq 2\|M_0\|_F^2 + 2\|M'_0\|_F^2 \leq 4n$, so by (2.32),

$$\|M - M'\|_F^2 \leq \frac{\sigma^2 \log l}{128} \|M_0 - M'_0\|_F^2 \leq \frac{\sigma^2}{32} n \log l \leq \frac{\sigma^2}{8} \log |\mathcal{P}|.$$

Since $\log l \geq \frac{1}{2} \log(l+1)$ for $l \geq 2$, applying Lemma 2.7.4 completes the proof. \square

Proof of Theorem 2.2.6. The last term $\min(\frac{\sigma^2}{m}, m^2 V_0^2)$ is achieved by Lemma 2.5.11, so we focus on the trade-off between the first two terms. Suppose that $(\frac{16n}{\sigma^2})^{1/3} V_0^{2/3} \geq 3$, in which case the first term $(\frac{\sigma^2 V_0}{n})^{2/3}$ dominates the second term. Then we have $m(\frac{16n}{\sigma^2})^{1/3} V_0^{2/3} - m \geq 2m$. Setting

$$K_0 = \lfloor m \left(\frac{16n}{\sigma^2} \right)^{1/3} V_0^{2/3} - m \rfloor,$$

we see that $K_0 \geq \lfloor \frac{m}{2} (\frac{16n}{\sigma^2})^{1/3} V_0^{2/3} \rfloor$. Lemma 2.5.10 can be applied with this choice of K_0 . Then the term $c\sigma^2 \frac{K_0}{nm}$ is lower bounded by $c(\frac{\sigma^2 V_0}{n})^{2/3}$.

On the other hand, if $(\frac{16n}{\sigma^2})^{1/3} V_0^{2/3} \leq 3$, then the second term $\frac{\sigma^2}{n}$ dominates the first up to a constant. To deduce a lower bound of this rate, we apply Lemma 2.7.3 to get $\Omega \subset \{0, 1\}^m$ such that $\log |\Omega| \geq m/8$ and $\delta(\omega, \omega') \geq m/4$ for distinct $\omega, \omega' \in \Omega$. For each $\omega \in \Omega$, define $M^\omega \in \mathbb{R}^{n \times m}$ by setting every row of M^ω equal to $\frac{\sigma}{8\sqrt{n}} \omega^\top$. Then

1. $M^\omega \in \mathcal{U}^m(V_0)$ since $V(M^\omega) = 0$;
2. $\|M^\omega - M^{\omega'}\|_F^2 = \frac{\sigma^2}{64} \delta(\omega, \omega') \geq c\sigma^2 m$;
3. $\|M^\omega - M^{\omega'}\|_F^2 = \frac{\sigma^2}{64} \delta(\omega, \omega') \leq \frac{\sigma^2}{64} m \leq \frac{\sigma^2}{8} \log |\Omega|$.

Hence Lemma 2.7.4 implies a lower bound on $\frac{1}{nm} \|\hat{M} - M\|_F^2$ of rate $\frac{\sigma^2 m}{nm} = \frac{\sigma^2}{n}$. \square

2.7.5 Proofs of lemmas in Section 2.5.4

Proof of Lemma 2.5.13. The proof follows that of Theorem 2.2.1 with appropriate adaptation, so for simplicity we will not detail every step. Assume without loss of generality that $\Pi^* = I_n$. Fix $A \in \mathcal{S}^m$ and $\tilde{\Pi} \in \mathfrak{S}_n$. Define

$$f_{\tilde{\Pi}A}(t) = \sup_{M \in \tilde{\Pi} \mathcal{S}^m \cap \mathcal{B}^{nm}(\tilde{\Pi}A, t)} \langle M - \tilde{\Pi}A, Y - \tilde{\Pi}A \rangle - \frac{t^2}{2}.$$

Since $\mathcal{S}^m = \mathcal{C}_1^m$ with $\mathbf{1} = (n, \dots, n)$, by Lemma 2.5.7,

$$\log N(\Theta_{\tilde{\Pi} \mathcal{S}^m}(A, t), \|\cdot\|_F, \varepsilon) \leq C\varepsilon^{-1} t K(A) \log \frac{enm}{K(A)}.$$

Following the proof of Lemma 2.5.3, we see that

$$f_{\tilde{\Pi}A}(t) \leq C\sigma t \sqrt{K(A) \log \frac{enm}{K(A)}} + t \|\tilde{\Pi}A - A^*\|_F - \frac{t^2}{2} + st$$

with probability at least $1 - C \exp(-\frac{cs^2}{\sigma^2})$. Lemma 2.5.1 then implies that on this event

$$\|\tilde{\Pi}\tilde{A} - A^*\|_F \leq 2C\sigma \sqrt{K(A) \log \frac{enm}{K(A)}} + 3\|\tilde{\Pi}A - A^*\|_F + 2s. \quad (2.33)$$

Taking $s = \sigma \left[\sqrt{K(A) \log \frac{enm}{K(A)}} + C_2 \sqrt{n \log n} \right]$ for a sufficiently large constant $C_2 > 0$, we see that with probability at least $1 - \exp(-c(m+n) - n \log n)$,

$$\begin{aligned} \|\tilde{\Pi}\tilde{A} - A^*\|_F^2 &\lesssim \sigma^2 K(A) \log \frac{enm}{K(A)} + \sigma^2 n \log n + \|\tilde{\Pi}A - A^*\|_F^2 \\ &\lesssim \sigma^2 K(A) \log \frac{enm}{K(A)} + \sigma^2 n \log n + \|A - A^*\|_F^2 + \|\tilde{\Pi}A^* - A^*\|_F^2. \end{aligned}$$

Minimizing over $A \in \mathcal{S}^m$ yields the desired bound for a fixed $\tilde{\Pi}$. Finally, the bound holds simultaneously for all $\tilde{\Pi} \in \mathfrak{S}_n$ with probability at least $1 - e^{-c(m+n)}$ by a union bound since $n! < n^n = \exp(n \log n)$. \square

Proof of Lemma 2.5.14. Since $Z \sim \text{subG}(\sigma^2)$, $Z_{i,j}$ and $\frac{1}{\sqrt{m}} \sum_{j=1}^m Z_{i,j}$ are sub-Gaussian random variables with variance proxy σ^2 . A standard union bound yields that

$$\max \left(\max_{i \in [n], j \in [m]} |Z_{i,j}|, \max_{i \in [n]} \frac{1}{\sqrt{m}} \left| \sum_{j=1}^m Z_{i,j} \right| \right) \leq \tau = 3\sigma \sqrt{\log(nm\delta^{-1})}$$

on an event \mathcal{E} of probability at least $1 - 2(nm + n) \exp(-\frac{\tau^2}{2\sigma^2}) \geq 1 - \delta$.

In the sequel, we make statements that are valid on the event \mathcal{E} . Since $Y_{\pi^*(i),j} = A_{i,j}^* + Z_{i,j}$, by the triangle inequality,

$$|\Delta_Y(\pi^*(i), \pi^*(i')) - \Delta_{A^*}(i, i')| \leq 2\tau. \quad (2.34)$$

Suppose that $\Delta_{A^*}(i, i') \geq 4\tau$. We claim that $s_{\pi^*(i)} < s_{\pi^*(i')}$. If for $l \in [n]$ we have $\Delta_Y(\pi^*(l), \pi^*(i)) \geq 2\tau$, then $\Delta_{A^*}(l, i) \geq 0$ by (2.34). Since A^* has increasing columns, $\Delta_{A^*}(l, i') \geq 4\tau$. Again by (2.34), $\Delta_Y(\pi^*(l), \pi^*(i')) \geq 2\tau$. By the definition of the score, we see that $s_{\pi^*(i)} \leq s_{\pi^*(i')}$. Moreover, $\Delta_{A^*}(i, i') \geq 4\tau$ so $\Delta_Y(\pi^*(i), \pi^*(i')) \geq 2\tau$. Therefore $s_{\pi^*(i)} < s_{\pi^*(i')}$. According to the construction of $\tilde{\pi}$, $\tilde{\pi}^{-1} \circ \pi^*(i) < \tilde{\pi}^{-1} \circ \pi^*(i')$. \square

Proof of Lemma 2.5.15. Throughout the proof, we restrict ourselves to the event \mathcal{E} defined in Lemma 2.5.14. To simplify the notation, we define $\alpha_i = A_{\tilde{\pi}^{-1} \circ \pi^*(i), \cdot}^* - A_{i, \cdot}^*$. Then

$$\|\tilde{\Pi}A^* - \Pi^*A^*\|_F^2 = \sum_{i=1}^n \|A_{\tilde{\pi}^{-1} \circ \pi^*(i), \cdot}^* - A_{i, \cdot}^*\|_2^2 = \sum_{i \in I} \|\alpha_i\|_2^2, \quad (2.35)$$

where I is the set of indices i for which α_i is nonzero. For each $i \in I$,

$$\begin{aligned} \|\alpha_i\|_2^2 &= \min \left(\frac{\|\alpha_i\|_2^2}{\|\alpha_i\|_\infty^2}, \frac{m\|\alpha_i\|_2^2}{\|\alpha_i\|_1^2} \right) \cdot \max \left(\|\alpha_i\|_\infty^2, \frac{\|\alpha_i\|_1^2}{m} \right) \\ &= \min \left(\frac{\|\alpha_i\|_2^2}{\|\alpha_i\|_\infty^2}, \frac{m\|\alpha_i\|_2^2}{\|\alpha_i\|_1^2} \right) \cdot \Delta_{A^*}(i, \tilde{\pi}^{-1} \circ \pi^*(i))^2 \end{aligned} \quad (2.36)$$

by (2.26).

Next, we proceed to showing that $|\Delta_{A^*}(i, \nu(i))| \leq 4\tau$ for any $i \in [n]$, where $\nu = \tilde{\pi}^{-1} \circ \pi^*$. To that end, note that if $\Delta_{A^*}(i, \nu(i)) > 4\tau$, in which case $\Delta_{A^*}(i, i') > 4\tau$ for all $i' \in I' := \{i' \in [n] : i' \geq \nu(i)\}$, then it follows from Lemma 2.5.14 that on \mathcal{E} , $\nu(i) < \nu(i')$, $\forall i \in I'$. Note that $|\nu(I')| = |I'| = n - \nu(i) + 1$. Hence $\nu(i) < \nu(i')$, $\forall i \in I'$ implies that $\nu(i) \leq n - |\nu(I')| = \nu(i) - 1$, which is a contradiction. Therefore, there does not exist such $i \in [n]$ on \mathcal{E} . The case where $\Delta_{A^*}(i, \nu(i)) < -4\tau$ is treated in a symmetric manner.

Combining this bound with (2.35) and (2.36), we conclude that

$$\begin{aligned} \|\tilde{\Pi}A^* - \Pi^*A^*\|_F^2 &\lesssim \sum_{i \in I} \min\left(\frac{\|\alpha_i\|_2^2}{\|\alpha_i\|_\infty^2}, \frac{m\|\alpha_i\|_2^2}{\|\alpha_i\|_1^2}\right) \cdot \tau^2 \\ &\lesssim \sigma^2 R(A^*) n \log(nm\delta^{-1}). \end{aligned}$$

by the definitions of $R(A^*)$ and τ . □

2.7.6 Proof of Corollary 2.4.1

The proof closely follows that of Theorem 2.2.1 and Theorem 2.2.3.

First note that the term $n \log n$ in the bound of Lemma 2.5.8 comes from a union bound applied to the set of permutations, so it is not present if we consider only the set of unimodal matrices \mathcal{U}^m instead of \mathcal{M} . Hence taking $m = 1$ in the lemma yields that

$$\log N(\Theta_{\mathcal{U}}(\tilde{\theta}, t), \|\cdot\|_2, \varepsilon) \leq C\varepsilon^{-1} t k(\tilde{\theta}) \log \frac{en}{k(\tilde{\theta})}.$$

For $\tilde{\theta} \in \mathcal{U}$, define

$$f_{\tilde{\theta}}(t) = \sup_{\theta \in \mathcal{U} \cap \mathcal{B}^n(\tilde{\theta}, t)} \langle \theta - \tilde{\theta}, y - \tilde{\theta} \rangle - \frac{t^2}{2}.$$

Following the proof of Lemma 2.5.3 and using the above metric entropy bound, we see that

$$f_{\tilde{\theta}}(t) \leq C\sigma t \sqrt{k(\tilde{\theta}) \log \frac{en}{k(\tilde{\theta})}} + t\|\tilde{\theta} - \theta^*\|_2 - \frac{t^2}{2} + st$$

with probability at least $1 - C \exp(-\frac{cs^2}{\sigma^2})$. Then the proof of Theorem 2.2.1 gives that with probability at least $1 - C \exp(-\frac{cs^2}{\sigma^2})$,

$$\|\hat{\theta} - \theta^*\|_2 \leq C \left(\sigma \sqrt{k(\tilde{\theta}) \log \frac{en}{k(\tilde{\theta})}} + \|\tilde{\theta} - \theta^*\|_2 \right) + 2s.$$

Taking $s = C\sigma\sqrt{\alpha \log n}$ for $\alpha \geq 1$ and C sufficiently large, we get that with probability at least $1 - n^{-\alpha}$,

$$\|\hat{\theta} - \theta^*\|_2^2 \lesssim \sigma^2 k(\tilde{\theta}) \log \frac{en}{k(\tilde{\theta})} + \|\tilde{\theta} - \theta^*\|_2^2 + \alpha \sigma^2 \log n.$$

Minimizing over $\tilde{\theta} \in \mathcal{U}$ yields the first bound of the corollary. The corresponding bound in expectation follows from integrating the tail probability as in the proof of Theorem 2.2.1.

Finally, we can apply the proof of Theorem 2.2.3 with $m = 1$ to achieve the global bound.

Chapter 3

Minimax Rates and Efficient Algorithms for Noisy Sorting

Pairwise comparison data is frequently observed in various domains, including recommender systems, website ranking, voting and social choice [BMR10, DKNS01, Liu09, You88, CN91]. For these applications, it is of significant interest to produce a suitable ranking of the items by aggregating the outcomes of pairwise comparisons. The general problem of interest can be stated as follows. Suppose there are n items to be compared and an underlying matrix P of probability parameters, each entry $P_{i,j}$ of which represents the probability that item i beats item j if they are compared. Hence we have $P_{j,i} = 1 - P_{i,j}$ and the event that item i beats item j in a comparison can be viewed as a Bernoulli random variable with probability $P_{i,j}$. Observing the outcomes of N independent pairwise comparisons, we aim to estimate the absolute ranking of the items.

For the sake of consistency, one needs of course to impose some structure on the matrix $P = \{P_{i,j}\}_{1 \leq i,j \leq n}$. These structural assumptions are traditionally split between *parametric* and *nonparametric* ones. Classical *parametric models* include the Bradley-Terry-Luce model [BT52, Luc59] and the Thurstone model [Thu27]. These models can be recast as log-linear models, which enables the use of the statistical and computational machinery of maximum likelihood estimation in generalized linear models [Hun04, NOS12, RA14, HOX14, SBB⁺16, NOS16, NOTX17].

To allow richer structures on the probability matrix P beyond the scope of parametric models, *permutation-based models* such as the *noisy sorting model* [BM08, BM09] and the *strong stochastic transitivity* (SST) model [Cha15, SBGW17] have recently become more prevalent. These models only require shape constraints on the matrix P and are typically called *nonparametric*. In these models, the underlying ranking of items is determined by an unknown permutation π^* , and, additionally, the comparison probabilities are assumed to have a bi-isotonic structure when the items are aligned according to π^* . While permutation-based models provide ordering structures that are not captured by parametric models [Aga16, SBGW17], they introduce both statistical and computational barriers for estimation of the underlying ranking. These barriers are mainly due to the complexity of the discrete set of permutations. On the one hand, the complexity of the set of permutations is not well understood

(see the discussion following Theorem 8 of [CD16]), which leads to logarithmic gaps in the current statistical bounds for permutation-based models. On the other hand, it is computationally challenging to optimize over the set of permutations, so current algorithms either sacrifice nontrivial statistical performance or have impractical time complexity. In this chapter, we aim to address both questions for the noisy sorting model.

In practice, it is unlikely that all the items are compared to each other. To account for this limitation, a widely used scheme consists in assuming that each pairwise comparison is observed with probability $p \in (0, 1]$ [Cha15, SBGW17]. In addition to this model of missing comparisons, we study the model where N pairwise comparisons are sampled uniformly at random from the $\binom{n}{2}$ pairs, with replacement and independent of each other. It turns out that sampling with and without replacement yields the same rate of estimation up to a constant when the expected numbers of observations coincide.

Our contributions. We focus on the noisy sorting model with partial observations, under which a stronger item wins a comparison against a weaker item with probability at least $\frac{1}{2} + \lambda$ where $\lambda \in (0, \frac{1}{2})$. For sampling both with and without replacement, we establish the minimax rate of learning the underlying permutation. In particular, the rate does not involve a logarithmic term, and we explain this phenomenon through a careful analysis of the metric entropy of the set of permutations equipped with the Kendall tau distance, which is of independent theoretical interest.

Moreover, we propose a multistage sorting algorithm that has time complexity $\tilde{O}(n^2)$. For the sampling with replacement model, we prove a theoretical guarantee on the performance of the multistage sorting algorithm, which differs from the minimax rate by only a polylogarithmic factor. In addition, the algorithm is demonstrated to perform similarly for both sampling models using simulated examples.

Related work. The noisy sorting model was proposed by [BM08]. In the original paper, the optimal rate of estimation achieved by the maximum likelihood estimator (MLE) is established, and an algorithm with time complexity $O(n^C)$ is shown to find the MLE with high probability in the case of full observations¹, where $C = C(\lambda)$ is a large unknown constant. Moreover, their algorithm does not have a polynomial running time if only $o(n^2)$ random pairwise comparisons are observed. Our work generalizes the optimal rate to the partial observation settings by studying a variant of the MLE for the upper bound. In the model of sampling with replacement, our fast multistage sorting algorithm provably achieves near-optimal rate of estimation. Since finding the MLE for the noisy sorting model is an instance of the NP-hard feedback arc set problem [Alo06, KMS07, ACN08, BM08], our results indicate that, despite the NP-hardness of the worst-case problem, it is still possible to achieve (near-)optimal rates for the average-case statistical setting in polynomial time.

¹If the algorithm is allowed to actively choose the pairs to be compared, the sample complexity can be reduced to $O(n \log n)$. However, in the passive setting which we adopt throughout this chapter, the algorithm still needs $\Theta(n^2)$ pairwise comparisons.

The SST model generalizes the noisy sorting model, and minimax rates in the SST model have been studied by [SBGW17]. However, the upper bound specialized to noisy sorting contains an extra logarithmic factor, which this work shows to be unnecessary. Moreover, the lower bound there is based on noisy sorting models with λ shrinking to zero as $n \rightarrow \infty$, while we establish a matching lower bound at any fixed λ . In addition, algorithms of [WJJ13, SBGW17, CM16] are all statistically suboptimal for the noisy sorting model. This is partially addressed by our multistage sorting algorithm as discussed above.

In fact, both with- and without-replacement sampling models discussed in this chapter are restrictive for applications where the set of observed comparisons is subject to certain structural constraints [HOX14, SBB⁺16, NOTX17, PMM⁺17b]. Obtaining sharper rates of estimation for these more complex sampling models is of significant interest but is beyond the scope of the current work.

Finally, we mention a few other lines of related work. Besides permutation-based models, low-rank structures have also been proposed by [RA16] to generalize classical parametric models. Moreover, there is an extensive literature on active ranking from pairwise comparisons [JN11, HSRW16, AAK17], where the pairs to be compared are chosen actively and in a sequential fashion by the learner. The sequential nature of the models greatly reduces sample complexity, so we do not compare our results for passive observations to the literature on active learning. However, it is interesting to note that our multistage sorting algorithm is reminiscent of active algorithms, because it uses different batches of samples for different stages. Thus active learning algorithms could potentially be useful even for passive sampling models.

Organization. The noisy sorting model together with the two sampling models is formalized in Section 3.1. In Section 3.2, we present our main results, the minimax rate of estimation for the latent permutation and the near-optimal rate achieved by an efficient multistage sorting algorithm. To complement our theoretical findings, we inspect the empirical performance of the multistage sorting algorithm on numerical examples in Section 3.3. We discuss directions for future research in Section 3.4. Section 3.5 is devoted to the study of the set of permutations equipped with the Kendall tau distance. Proofs of the main results are provided in Section 3.6.

Notation. For a positive integer n , let $[n] = \{1, \dots, n\}$. For a finite set S , we denote its cardinality by $|S|$. Given $a, b \in \mathbb{R}$, let $a \wedge b = \min(a, b)$ and $a \vee b = \max(a, b)$. We use C and c , possibly with subscripts, to denote universal positive constants that may change at each appearance. For two sequences $\{u_n\}_{n=1}^\infty$ and $\{v_n\}_{n=1}^\infty$, we write $u_n \lesssim v_n$ if there exists a universal constant $C > 0$ such that $u_n \leq C v_n$ for all n . We define the relation $u_n \gtrsim v_n$ analogously, and write $u_n \asymp v_n$ if both $u_n \lesssim v_n$ and $u_n \gtrsim v_n$ hold. Let \mathfrak{S}_n denote the symmetric group on $[n]$, i.e., the set of permutations $\pi : [n] \rightarrow [n]$.

3.1 Problem setup

The noisy sorting model can be formulated as follows. Fix an unknown permutation $\pi^* \in \mathfrak{S}_n$ 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 the $\pi^*(i)$ -th weakest among the n items. For a fixed, possibly unknown $\lambda \in (0, 1/2)$, we define a class of matrices

$$\mathfrak{M}_n(\lambda) = \left\{ M \in [0, 1]^{n \times n} : M_{i,i} = \frac{1}{2}, M_{i,j} \geq \frac{1}{2} + \lambda \text{ if } i > j, M_{i,j} \leq \frac{1}{2} - \lambda \text{ if } i < j \right\},$$

where $\mathbf{1}_n$ is the n -dimensional all-ones vector. In addition, we define a special matrix $M_n^*(\lambda) \in \mathfrak{M}_n(\lambda)$ by

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

Note that $M_n^*(\lambda)$ satisfies strong stochastic transitivity but other matrices $M \in \mathfrak{M}_n(\lambda)$ may not. Though this observation plays a crucial role in the design of efficient algorithms, our statistical results hold for general matrices in $\mathfrak{M}_n(\lambda)$.

To model pairwise comparisons, fix $M \in \mathfrak{M}_n(\lambda)$ and let $M_{\pi^*(i), \pi^*(j)}$ denote the probability that items i beats item j when they are compared², so that a stronger item beats a weaker item with probability at least $\frac{1}{2} + \lambda$. As a result, λ captures the signal-to-noise ratio of our problem and our minimax results explicitly capture the dependence in this key parameter.

3.1.1 Sampling models

In the noisy sorting model, suppose that for each (unordered) pair (i, j) with $i \neq j$, we observe the outcomes of $N_{i,j}$ ($= N_{j,i}$) comparisons between them, and item i wins a comparison against item j with probability $M_{\pi^*(i), \pi^*(j)}$ independently. The set $\{N_{i,j}\}_{i < j}$ of $\binom{n}{2}$ nonnegative integers is determined by certain sampling models described below. We allow $N_{i,j}$ to be zero, which means that i and j are not compared. We collect sufficient statistics into a matrix $A \in \mathbb{R}^{n \times n}$ consisting of outcomes of pairwise comparisons, by defining $A_{i,j}$ to be the number of times item i beats item j among the $N_{i,j}$ comparisons between i and j . In particular, we have $A_{i,j} + A_{j,i} = N_{i,j} = N_{j,i}$ for $i \neq j$ and $A_{i,i} = 0$. Our goal is to aggregate the results of pairwise comparisons to estimate π^* , the underlying order of items.

In the full observation setup of [BM08], we have $N_{i,j} = 1$ for each pair (i, j) and the total number of observations is $N := \sum_{i < j} N_{i,j} = \binom{n}{2}$. Instead, we are interested here in the regime where the total number of observations N is much smaller than $\binom{n}{2}$. We study the following two sampling models in this chapter:

²The diagonal entries of M are inessential in the model as an item is not compared to itself, and they are set to $1/2$ only for concreteness.

(O_1) *Sampling without replacement.* In this sampling model, instead of observing all the pairwise comparisons, we observe each pair with probability $p \in (0, 1]$ independently. Hence each $N_{i,j} \sim \text{Ber}(p)$ is a Bernoulli random variable with parameter p , and in expectation we have $N' := p \binom{n}{2}$ observations in total.

(O_2) *Sampling with replacement.* We observe N pairwise comparisons between the items, sampled uniformly and independently with replacement from the $\binom{n}{2}$ pairs.

In the sequel, we study the noisy sorting model with either of the above two sampling models. In particular, the minimax rates of estimating π^* coincide for the two sampling models if $p \binom{n}{2} \asymp N$, i.e., if the expected number of observations are of the same order.

3.1.2 Measures of performance

Having discussed the sampling and comparison models, we turn to the distance used to measure the difference between the underlying permutation π^* and an estimated permutation $\hat{\pi}$. Among various distances defined on the symmetric group, we consider primarily the *Kendall tau distance*, i.e., the number of *inversions* (or discordant pairs) between permutations, defined as

$$d_{\text{KT}}(\pi, \sigma) = \sum_{(i,j): \sigma(i) < \sigma(j)} \mathbb{1}(\pi(i) > \pi(j))$$

for $\pi, \sigma \in \mathfrak{S}_n$. Note that $0 \leq d_{\text{KT}}(\pi, \sigma) \leq \binom{n}{2}$. The Kendall tau distance between two permutations is a natural metric on \mathfrak{S}_n , and it is equal to the minimum number of adjacent transpositions required to change from one permutation to another [Knu98]. A closely related distance on \mathfrak{S}_n is the ℓ_1 -distance, also known as Spearman's footrule, defined as

$$\|\pi - \sigma\|_1 = \sum_{i=1}^n |\pi(i) - \sigma(i)|$$

for $\pi, \sigma \in \mathfrak{S}_n$. It is well known [DG77] that

$$d_{\text{KT}}(\pi, \sigma) \leq \|\pi - \sigma\|_1 \leq 2d_{\text{KT}}(\pi, \sigma). \quad (3.1)$$

Hence the rates of estimation in the two distances coincide. Another distance on \mathfrak{S}_n we use is the ℓ_∞ -distance, defined as

$$\|\pi - \sigma\|_\infty = \max_{i \in [n]} |\pi(i) - \sigma(i)|.$$

Note that unlike existing literature on ranking from pairwise comparisons where metrics on the probability parameters are studied, we employ here distances that measure how far an item is from its true ranking.

3.2 Main results

In this section, we state our main results. Specifically, we establish the minimax rates of estimating π^* in the Kendall tau distance (and thus in ℓ_1 distance) for noisy sorting under both sampling models (O_1) and (O_2) . The minimax estimator that we propose is intractable in general and we complement our results with an efficient estimator of π^* which achieves near-optimal rates in both the Kendall tau and the ℓ_∞ -distance, under the sampling model (O_2) .

3.2.1 Minimax rates of noisy sorting

Under the noisy sorting model with latent permutation $\pi^* \in \mathfrak{S}_n$ and matrix of probabilities $M \in \mathfrak{M}_n(\lambda)$, we determine the minimax rate of estimating π^* in the following theorem. We assume that λ is given in this section for simplicity; an efficient procedure of estimating λ is presented in Section 3.2.2. Let $\mathbb{E}_{\pi^*, M}$ denote the expectation with respect to the probability distribution of the observations in the noisy sorting model with underlying permutation $\pi^* \in \mathfrak{S}_n$ and matrix of probabilities $M \in \mathfrak{M}_n(\lambda)$, in either sampling model.

Theorem 3.2.1. *Fix $\lambda \in (0, \frac{1}{2} - c]$ where c is a universal positive constant. It holds that*

$$\min_{\tilde{\pi}} \max_{\substack{\pi^* \in \mathfrak{S}_n \\ M \in \mathfrak{M}_n(\lambda)}} \mathbb{E}_{\pi^*, M}[d_{\text{KT}}(\tilde{\pi}, \pi^*)] \asymp \begin{cases} \frac{n^3}{N'\lambda^2} \wedge n^2, & \text{in sampling model } (O_1), \\ \frac{n^3}{N\lambda^2} \wedge n^2, & \text{in sampling model } (O_2), \end{cases}$$

where the minimum is taken minimized over all permutation estimators $\tilde{\pi} \in \mathfrak{S}_n$ that are measurable with respect to the observations.

The theorem establishes the minimax rates for noisy sorting, including the case of partial observations and weak signals. The upper bounds in fact hold with high probability as shown in Theorem 3.6.2. If the expected numbers of observations in the two sampling models (O_1) and (O_2) are of the same order, i.e., $N' = p \binom{n}{2} \asymp N$, then the two rates coincide. In this sense, the two sampling models are statistically equivalent. In sampling model (O_1) , if $p = 1$ and λ is larger than a constant, then the rate of order n recovers the upper bound proved by [BM08].

Note in particular the absence of logarithmic factor in the rates. Naively bounding the metric entropy of \mathfrak{S}_n by $\log |\mathfrak{S}_n| \simeq n \log n$ actually yields a superfluous logarithmic term in the upper bound. To avoid it, we employ the maximum likelihood estimator over an appropriately chosen ε -net of \mathfrak{S}_n , discussed in detail in Section 3.6.1. In addition, we study the doubling dimension of \mathfrak{S}_n ; see the discussion after Proposition 3.5.1. Closing this logarithmic gap for other problems involving latent permutations [CD16, FMR16, SBGW17, PWC17] remains an open question.

The technical assumption $\lambda \leq 1/2 - c$ in Theorem 3.2.1 is very mild, because we are interested in the “noisy” sorting model (meaning that the pairwise comparisons

are noisy, or equivalently that λ is not close to $\frac{1}{2}$). In fact the requirement that λ be bounded away from $\frac{1}{2}$ can be lifted, in which case we establish upper and lower bounds that match up to a logarithmic factor of order $\log(1/\Delta)$, where $\Delta = 1/2 - \lambda$ (see Section 3.6).

Finally, we note that the proof of Theorem 3.2.1 holds even in the so-called *semi-random* setting [BS95, MMV13], in which observations are generated by one of the random procedures described above, but a “helpful” adversary is allowed to reverse the outcome of any comparison in which a weaker item beat a stronger item. Though these reversals appear benign at first glance, the presence of such an adversary can in fact worsen statistical rates of estimation in more brittle models such as stochastic block models and the related broadcast tree model [MPW16]. Our results indicate that no such degradation occurs for the rates of estimation in the noisy sorting problem.

3.2.2 Efficient multistage sorting

The minimax upper bound in Theorem 3.2.1 is established using a computationally prohibitive estimator, so we now introduce an efficient estimator of the underlying permutation that can be computed in time $\tilde{O}(n^2)$. In this section, we prove theoretical guarantees for this estimator under the noisy sorting model with probability matrix $M = M_n^*(\lambda)$ and observations sampled with replacement according to (O_2) when λ is bounded away from zero by a universal constant. No polynomial-time algorithm was previously known to achieve near-optimal rates even in this simplified setting when $o(n^2)$ pairwise comparisons are observed.

Since we aim to prove guarantees up to constants, we may assume that we have $2N$ pairwise comparisons, and split them into two independent samples, each containing N pairwise comparisons. The first sample is used to estimate the parameter λ and the second one is used to estimate the permutation π^* .

First, we introduce a fairly simple estimator $\hat{\lambda}$ of λ that can be described informally as follows: first sort in increasing order the items according to the number of wins. Then for any pair (i, j) for which item i is ranked $n/2$ positions higher than item j , it is very likely that item i is stronger than item j so that it beats item j with probability $\frac{1}{2} + \lambda$. We then average the $\text{Ber}(\frac{1}{2} + \lambda)$ variables over all such pairs to obtain an estimator $\hat{\lambda}$ of λ . More formally, we further split the first sample into two subsamples, each containing $N/2$ pairwise comparisons. Denote by $A'_{i,j}$ and $A''_{i,j}$ the number of wins item i has against item j in the first and second subsample, respectively. The estimator $\hat{\lambda}$ is given by the following procedure:

1. For each $i \in [n]$, associate with item i a score $S_i = \sum_{j=1}^n A'_{i,j}$.
2. Construct a permutation $\tilde{\pi}$ by sorting the scores S_i in increasing order, i.e., $\tilde{\pi}$ is chosen so that $\tilde{\pi}(i) < \tilde{\pi}(j)$ if $S_i \leq S_j$, with ties broken arbitrarily.

3. Define $\hat{\lambda} = \frac{2}{N} \binom{n}{2} \binom{n/2}{2}^{-1} \sum_{\tilde{\pi}(i) - \tilde{\pi}(j) > \frac{n}{2}} A''_{i,j} - \frac{1}{2}$.

Given the estimator $\hat{\lambda}$, we now describe a multistage procedure to estimate the permutation π^* . To recover the underlying order of items, it is equivalent to estimate the row sums $\sum_{j=1}^n M_{\pi^*(i), \pi^*(j)}$ which we call scores of the items, because the scores are increasing linearly if the items are placed in order. Initially, for each $i \in [n]$, we estimate the score of item i by the number of wins item i has. If item i has a much higher score than item j in the first stage, then we are confident that item i is stronger than item j . Hence in the second stage, we can estimate $M_{\pi^*(i), \pi^*(j)}$ by $\frac{1}{2} + \hat{\lambda}$, which is very close to the truth. For those pairs that we are not certain about, $M_{\pi^*(i), \pi^*(j)}$ is still estimated by its empirical version. The variance of each score is thus greatly reduced in the second stage, thereby yielding a more accurate order of the items. Then we iterate this process to obtain finer and finer estimates of the scores and the underlying order.

To present the Multistage Sorting (MS) algorithm formally, let us fix a positive integer T which is the number of stages of the algorithm. We further split the second sample into T subsamples each containing N/T pairwise comparisons³. Similar to the data matrix A for the full sample, for $t \in [T]$ we define a matrix $A^{(t)} \in \mathbb{R}^{n \times n}$ by setting $A_{i,j}^{(t)}$ to be the number of wins item i has against item j in the t -th sample. The MS algorithm proceeds as follows:

1. For each $i \in [n]$, define $I^{(0)}(i) = [n]$, $I_-^{(0)}(i) = \emptyset$ and $I_+^{(0)}(i) = \emptyset$. For $0 \leq t \leq T$, we use $I^{(t)}(i)$ to denote the set of items j whose ranking relative to i has not been determined by the algorithm at stage t .
2. At the t -th stage where $t \in [T]$, compute the score $S_i^{(t)}$ of item i :

$$S_i^{(t)} = \frac{Tn(n-1)}{2N} \sum_{j \in I^{(t-1)}(i)} A_{i,j}^{(t)} + \sum_{j \in I_-^{(t-1)}(i)} \left(\frac{1}{2} + \hat{\lambda} \right) + \sum_{j \in I_+^{(t-1)}(i)} \left(\frac{1}{2} - \hat{\lambda} \right).$$

3. Let C_0 and C_1 be sufficiently large universal constants⁴. If it holds that

$$|I^{(t-1)}(i)| \geq C_1 n^2 \frac{T}{N} \log(nT), \quad (3.2)$$

then we set the threshold

$$\tau_i^{(t)} = (10 + 2C_0)n \sqrt{|I^{(t-1)}(i)| TN^{-1} \log(nT)},$$

and define the sets

$$\begin{aligned} I_-^{(t)}(i) &= \{j \in [n] : S_j^{(t)} - S_i^{(t)} < -\tau_i^{(t)}\}, \\ I_+^{(t)}(i) &= \{j \in [n] : S_j^{(t)} - S_i^{(t)} > \tau_i^{(t)}\}, \text{ and} \\ I^{(t)}(i) &= [n] \setminus (I_-^{(t)}(i) \cup I_+^{(t)}(i)). \end{aligned}$$

³We assume without loss of generality that T divides N to ease the notation.

⁴Determined according to Lemma 3.6.5 and Lemma 3.6.6 respectively.

If (3.2) does not hold, then we define $I^{(t)}(i) = I^{(t-1)}(i)$, $I_-^{(t)}(i) = I_-^{(t-1)}(i)$ and $I_+^{(t)}(i) = I_+^{(t-1)}(i)$.

4. After repeating Step 2 and 3 for $t = 1, \dots, T$, output a permutation $\hat{\pi}^{\text{MS}}$ by sorting the scores $S_i^{(T)}$ in increasing order, i.e., $\hat{\pi}^{\text{MS}}$ is chosen so that $\hat{\pi}^{\text{MS}}(i) < \hat{\pi}^{\text{MS}}(j)$ if $S_i^{(T)} \leq S_j^{(T)}$ with ties broken arbitrarily.

It is clear that the time complexity of each stage of the algorithm is $O(n^2)$. Take $T = \lfloor \log \log n \rfloor$ so that the overall time complexity of the MS algorithm is only $O(n^2 \log \log n)$. Our main result in this section is the following guarantee on the performance of the estimator $\hat{\pi}^{\text{MS}}$ given by the MS algorithm.

Theorem 3.2.2. *Suppose that $N \geq Cn \log n$ for a sufficiently large constant $C > 0$ and that $M = M_n^*(\lambda)$ where $\lambda \in [c, \frac{1}{2})$ for a constant $c > 0$. Then, under the noisy sorting model with sampling model (O_2), the following holds. With probability at least $1 - n^{-7}$, the MS algorithm with $T = \lfloor \log \log n \rfloor$ stages outputs an estimator $\hat{\pi}^{\text{MS}}$ that satisfies*

$$\|\hat{\pi}^{\text{MS}} - \pi^*\|_\infty \lesssim \frac{n^2}{N} (\log n) \log \log n$$

and

$$d_{\text{KT}}(\hat{\pi}^{\text{MS}}, \pi^*) \lesssim \frac{n^3}{N} (\log n) \log \log n.$$

Note that the second statement follows from the first one together with (3.1). Indeed, we have

$$d_{\text{KT}}(\hat{\pi}^{\text{MS}}, \pi^*) \leq \|\hat{\pi}^{\text{MS}} - \pi^*\|_1 \leq n \|\hat{\pi}^{\text{MS}} - \pi^*\|_\infty \lesssim \frac{n^3}{N} (\log n) \log \log n,$$

which is optimal up to a polylogarithmic factor in the regime where λ is bounded away from 0 according to Theorem 3.2.1 (and Theorem 3.6.3). Therefore, the MS algorithm achieves significant computational efficiency while sacrificing little in terms of statistical performance. On the downside, it is limited to the noisy sorting model where $M = M_n^*(\lambda)$ —this assumption is necessary to exploit strong stochastic transitivity—and our analysis does not account for the dependence in λ .

Furthermore, although we only consider model (O_2) of sampling with replacement in this section, the MS algorithm can be easily modified to handle model (O_1) of sampling without replacement. It is much more challenging to prove analogous theoretical guarantees in this case, because we cannot split the observations into independent samples. In Section 3.3, however, we provide empirical evidence showing that the MS estimator has very similar performance for the two sampling models.

Our algorithm bears comparison with the algorithm proposed by [BM08]. Their algorithm—which works in the full observation case $N = \binom{n}{2}$ —achieves the statistically optimal rate in time $O(n^C)$, where C is a large positive constant depending on λ . Though our algorithm’s statistical performance falls short of the optimal rate by a polylogarithmic factor, it runs in time $O(n^2 \log \log n)$ and works in the partial observation setting as long as $N \gtrsim n \log n$. Note by way of comparison that Theorem 3.6.3 indicates that no procedure achieves nontrivial recovery unless $N \gg n$.

3.3 Simulations

To support our theoretical findings in Section 3.2.2, we implement the MS algorithm on synthetic instances generated from the noisy sorting model. For simplicity, we take $\lambda = 0.25$ and set $\hat{\lambda} = \lambda$ in the algorithm. Theorem 3.2.2 predicts a scaling $n^3 N^{-1} (\log n) \log \log n$ of the estimation error in the Kendall tau distance for model (O_2) of sampling with replacement, where n is the number of items and N is the number of pairwise comparisons. This rate is optimal up to a polylogarithmic factor according to Theorem 3.6.3.

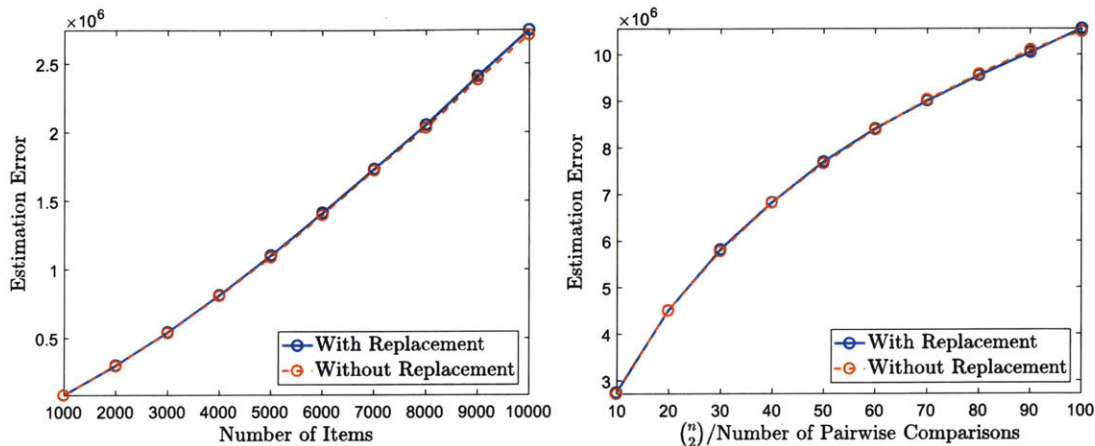


Figure 3-1: Estimation errors $d_{\text{KT}}(\hat{\pi}^{\text{MS}}, \pi^*)$ for the observations sampled with and without replacement. Left: $N = p \binom{n}{2} = 0.1 \binom{n}{2}$ and n ranging from 1,000 to 10,000; Right: $n = 10,000$ and $N = p \binom{n}{2}$ ranging from $0.1 \binom{n}{2}$ to $0.01 \binom{n}{2}$.

In Figure 3-1, we plot estimation errors $d_{\text{KT}}(\hat{\pi}^{\text{MS}}, \pi^*)$ averaged over 10 instances generated from the model. In the left plot, we let n range from 1,000 to 10,000 and set $N = 0.1 \binom{n}{2}$. For this choice of N , Theorem 3.2.2 predicts that $d_{\text{KT}}(\hat{\pi}^{\text{MS}}, \pi^*) = \tilde{O}_{\mathbb{P}}(n)$ and we indeed observe a near-linear scaling in that plot. In the right plot, we fix $n = 10,000$ and let the proportion of observed entries, $\alpha = N / \binom{n}{2}$ range from .01 to .1. For this choice of parameters, Theorem 3.2.2 predicts that $d_{\text{KT}}(\hat{\pi}^{\text{MS}}, \pi^*) \leq C_n \alpha^{-1}$ (recall that here n is fixed), and we clearly observe a sublinear relation between $d_{\text{KT}}(\hat{\pi}^{\text{MS}}, \pi^*)$ and α^{-1} . Note that this does not contradict the lower bound since the latter is stated up to constants.

Moreover, the MS algorithm can be easily modified to work for the without replacement model (O_1). Namely, given the partially observed pairwise comparisons, we assign each comparison to one of the samples $1, \dots, T$ uniformly at random, independent of all the other assignments. After splitting the whole sample into T subsamples, we execute the MS algorithm as in the previous case. In Figure 3-1, we take $p = N / \binom{n}{2}$ and plot the estimation errors for sampling without replacement, which closely follow the errors for observations sampled with replacement. Therefore, although it seems difficult to prove analogous guarantees on the performance of the MS algorithm applied to the without replacement model, empirically the algorithm performs very similarly for the two sampling models.

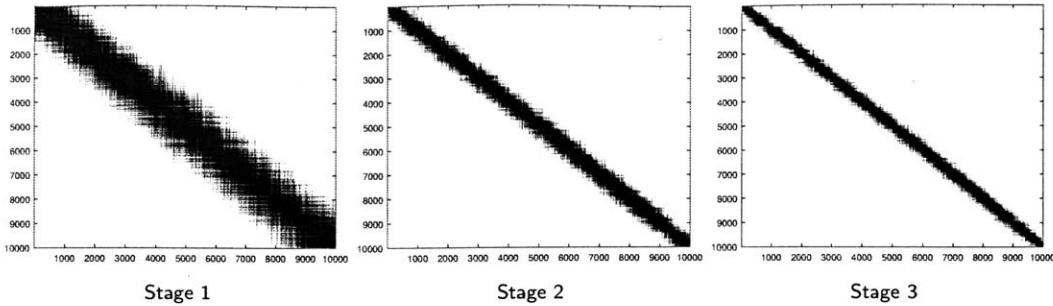


Figure 3-2: The uncertainty regions $\mathcal{R}^{(t)}$ at stages $t = 1, 2, 3$ of the MS algorithm. The two axes represent the indices of the items. A black pixel at (i, j) indicates that $(i, j) \in \mathcal{R}^{(t)}$, i.e., the algorithm is not certain about the relative order of item i and item j at stage t . A white pixel indicates the opposite.

To gain further intuition about the MS algorithm, we consider the set $I^{(t)}(i)$ defined in the algorithm. At stage t of the algorithm, the set $I^{(t)}(i)$ consists of all indices j for which we are not certain about the relative order of item i and item j . The proof of Theorem 3.2.2 essentially shows that the uncertainty set $I^{(t)}(i)$ is shrinking as the algorithm proceeds. To verify this intuition, in Figure 3-2 we plot the *uncertainty regions*

$$\mathcal{R}^{(t)} := \{(i, j) \in [n]^2 : i \in [n], j \in I^{(t)}(i)\}$$

at stages $t = 1, 2, 3$ of the MS algorithm, for $n = 10,000$ and $N = \binom{n}{2}$. The items are ordered according to $\pi^* = \text{id}$ for visibility of the region. As exhibited in the plots, the uncertainty region is indeed shrinking as the algorithm proceeds.

3.4 Discussion and open problems

In this chapter, we focused on minimax estimation of the latent permutation π^* . Viewing $M = \frac{1}{2}\mathbf{1}_n\mathbf{1}_n^\top$ as the null hypothesis and $M \in \mathfrak{M}_n(\lambda)$ as the alternative hypothesis, a natural question is to establish the minimax detection level of the signal strength λ in the hypothesis testing framework.

Moreover, we proved that the minimax rates for the noisy sorting problem do not involve any extra logarithmic factors even in the case of partial observations. For more complex models involving permutations [CD16, FMR16, SBGW17, PWC17, SBW17], however, there are logarithmic gaps between current upper and lower bounds. According to the discussion after Proposition 3.5.1, the logarithmic gaps do not necessarily stem from the unknown permutation, so it would be interesting to close these gaps or study whether they exist because of other aspects of the richer models.

For the MS algorithm, it remains an open question whether analogous upper bounds can be established for sampling without replacement. We conjecture that this is the case because of the empirical evidence in Section 3.3. More importantly, there are still statistical-computational gaps unresolved for the general noisy sorting

model where $M \in \mathfrak{M}_n(\lambda)$, for the SST model of [SBGW17] and for the seriation model of [FMR16]. It would be interesting to know if the ideas behind the MS algorithm could help tighten the gaps.

3.5 The symmetric group and inversions

Before proving the main results for the noisy sorting model, we study the metric entropy of the symmetric group \mathfrak{S}_n with respect to the Kendall tau distance. Counting permutations subject to constraints in terms of the Kendall tau distance is of theoretical importance and has interesting applications, e.g., in coding theory [BM10, MBZ13]. We present the results in terms of metric entropy, which easily applies to the noisy sorting problem and may find further applications in statistical problems involving permutations.

For $\varepsilon > 0$ and $S \subseteq \mathfrak{S}_n$, let $N(S, \varepsilon)$ and $D(S, \varepsilon)$ denote respectively the ε -covering number and the ε -packing number of S with respect to the Kendall tau distance. The following main result of this section provides bounds on the metric entropy of balls in \mathfrak{S}_n .

Proposition 3.5.1. *Consider the ball $\mathcal{B}(\pi, r) = \{\sigma \in \mathfrak{S}_n : d_{\text{KT}}(\pi, \sigma) \leq r\}$ centered at $\pi \in \mathfrak{S}_n$ with radius $r \in (0, \binom{n}{2})$. We have that for $\varepsilon \in (0, r)$,*

$$n \log \left(\frac{r}{n + \varepsilon} \right) - 2n \leq \log N(\mathcal{B}(\pi, r), \varepsilon) \leq \log D(\mathcal{B}(\pi, r), \varepsilon) \leq n \log \left(\frac{2n + 2r}{\varepsilon} \right) + 2n.$$

We now discuss some high-level implications of Proposition 3.5.1. Note that if $n \lesssim \varepsilon < r \leq \binom{n}{2}$, the lemma states that the ε -metric entropy of a ball of radius r in the Kendall tau distance scales as $n \log \frac{r}{\varepsilon}$. In other words, the symmetric group \mathfrak{S}_n equipped with the Kendall tau metric is a doubling space with doubling dimension $\Theta(n)$. One of the main messages of the current work is that although $\log |\mathfrak{S}_n| = \log(n!) \asymp n \log n$, the intrinsic dimension of \mathfrak{S}_n is $\Theta(n)$, which explains the absence of logarithmic factor in the minimax rate.

To start the proof, we first recall a useful tool for counting permutations, the *inversion table*. Formally, the inversion table b_1, \dots, b_n of a permutation $\pi \in \mathfrak{S}_n$ is defined by

$$b_i = \sum_{j:i < j} \mathbb{1}(\pi(i) > \pi(j))$$

for $i \in [n]$. Clearly, we have that $b_i \in \{0, 1, \dots, n - i\}$ and $d_{\text{KT}}(\pi, \text{id}) = \sum_{i=1}^n b_i$. It is easy to reconstruct a unique permutation using an inversion table with $b_i \in \{0, 1, \dots, n - i\}$, $i \in [n]$, so the set of inversion tables is bijective to \mathfrak{S}_n via this relation; see, e.g., [Mah00]. We use this bijection to bound the number of permutations that differ from the identity by at most k inversions. The following lemma appears in a different form in [BM10]. We provide a simple proof here for completeness.

Lemma 3.5.2. *For $0 \leq k \leq \binom{n}{2}$, we have that*

$$n \log(k/n) - n \leq \log |\{\pi \in \mathfrak{S}_n : d_{\text{KT}}(\pi, \text{id}) \leq k\}| \leq n \log(1 + k/n) + n.$$

Proof. According to the discussion above, the cardinality $|\{\pi \in \mathfrak{S}_n : d_{\text{KT}}(\pi, \text{id}) \leq k\}|$, which we denote by L , is equal to the number of inversion tables b_1, \dots, b_n where $b_i \in \{0, 1, \dots, n - i\}$ such that $\sum_{i=1}^n b_i \leq k$. On the one hand, if $b_i \leq \lfloor k/n \rfloor$ for all $i \in [n]$, then $\sum_{i=1}^n b_i \leq k$, so a lower bound on L is given by

$$\begin{aligned} L &\geq \prod_{i=1}^n (\lfloor k/n \rfloor + 1) \wedge (n - i + 1) \\ &\geq \prod_{i=1}^{n - \lfloor k/n \rfloor} (\lfloor k/n \rfloor + 1) \prod_{i=n - \lfloor k/n \rfloor + 1}^n (n - i + 1) \\ &\geq (k/n)^{n - k/n} \lfloor k/n \rfloor!. \end{aligned}$$

Using Stirling's approximation, we see that

$$\begin{aligned} \log L &\geq n \log(k/n) - (k/n) \log(k/n) + \lfloor k/n \rfloor \log \lfloor k/n \rfloor - \lfloor k/n \rfloor \\ &\geq n \log(k/n) - n. \end{aligned}$$

On the other hand, if b_i is only required to be a nonnegative integer for each $i \in [n]$, then we can use a standard "stars and bars" counting argument [Fel68] to get an upper bound of the form

$$L \leq \binom{n+k}{n} \leq e^n (1 + k/n)^n.$$

Taking the logarithm finishes the proof. \square

We are ready to prove Proposition 3.5.1.

of Proposition 3.5.1. The relation between the covering and the packing number is standard.

We employ a standard volume argument to control these numbers. Let \mathcal{P} be a 2ε -packing of $\mathcal{B}(\pi, r)$ so that the balls $\mathcal{B}(\sigma, \varepsilon)$ are disjoint for $\sigma \in \mathcal{P}$. Moreover, by the triangle inequality, $\mathcal{B}(\sigma, \varepsilon) \subseteq \mathcal{B}(\pi, r + \varepsilon)$ for each $\sigma \in \mathcal{P}$. By the invariance of the Kendall tau distance under composition, Lemma 3.5.2 yields

$$\begin{aligned} \log D(\mathcal{B}(\pi, r), 2\varepsilon) &\leq n \log(1 + r/n) + n - n \log(\varepsilon/n) + n \\ &= n \log\left(\frac{n+r}{\varepsilon}\right) + 2n. \end{aligned}$$

On the other hand, if \mathcal{N} is an ε -net of $\mathcal{B}(\pi, r)$, then the set of balls $\{\mathcal{B}(\sigma, \varepsilon)\}_{\sigma \in \mathcal{N}}$ covers $\mathcal{B}(\pi, r)$. By Lemma 3.5.2, we obtain

$$\begin{aligned} \log N(\mathcal{B}(\pi, r), \varepsilon) &\geq \log |\mathcal{B}(\pi, r)| - \log |\mathcal{B}(\sigma, \varepsilon)| \\ &\geq n \log(r/n) - n - n \log(1 + \varepsilon/n) - n \\ &= n \log\left(\frac{r}{n + \varepsilon}\right) - 2n, \end{aligned}$$

as claimed. \square

The lower bound on the packing number in Proposition 3.5.1 becomes vacuous when r and ε are smaller than n , so we complement it with the following result, which is useful for proving minimax lower bounds.

Lemma 3.5.3. *Consider the ball $\mathcal{B}(\pi, r)$ where $r < n/2$. We have that*

$$\log N(\mathcal{B}(\pi, r), r/4) \geq \frac{r}{5} \log \frac{n}{r}.$$

Proof. Without loss of generality, we may assume that $\pi = \text{id}$ and n is even. The sparse Varshamov-Gilbert bound (see Lemma 4.10 of [Mas07]) states that there exists a set \mathcal{S} of r -sparse vectors in $\{0, 1\}^{n/2}$, such that $\log |\mathcal{S}| \geq \frac{r}{5} \log \frac{n}{r}$ and any two distinct vectors in \mathcal{S} are separated by at least $r/2$ in the Hamming distance. We now map every $v \in \mathcal{S}$ to a permutation $\pi \in \mathcal{B}(\text{id}, r)$ by defining

1. $\pi(2i - 1) = 2i - 1$ and $\pi(2i) = 2i$ if $v(i) = 0$, and
2. $\pi(2i - 1) = 2i$ and $\pi(2i) = 2i - 1$ if $v(i) = 1$,

for $i \in [n]$. Note that $\pi \in \mathcal{B}(\text{id}, r)$ because π swaps at most r adjacent pairs. Denote by \mathcal{P} the image of \mathcal{S} under this mapping. Since the Hamming distance between any two distinct vectors in \mathcal{S} is lower bounded by $r/2$, we see that $d_{\text{KT}}(\pi, \sigma) \geq r/2$ for any distinct $\pi, \sigma \in \mathcal{P}$. Thus \mathcal{P} is an $r/2$ -packing of $\mathcal{B}(\text{id}, r)$. By construction, $|\mathcal{P}| = |\mathcal{S}| \geq \frac{r}{5} \log \frac{n}{r}$, so we can use the standard relation $D(\mathcal{B}(\text{id}, r), r/2) \leq N(\mathcal{B}(\text{id}, r), r/4)$ to complete the proof. \square

3.6 Proofs of the main results

This section is devoted to the proofs of our main results. We start with a lemma giving useful tail bounds for the binomial distribution.

Lemma 3.6.1. *Suppose that X has the Binomial distribution $\text{Bin}(N, p)$ where $N \in \mathbb{Z}_+$ and $p \in (0, 1)$. Then for $r \in (0, p)$ and $s \in (p, 1)$, we have*

1. $\mathbb{P}(X \leq rN) \leq \exp\left(-N \frac{(p-r)^2}{2p(1-r)}\right)$, and
2. $\mathbb{P}(X \geq sN) \leq \exp\left(-N \frac{(p-s)^2}{2s(1-p)}\right)$.

Proof. First, for $0 < q < p < 1$, by the definition of the Kullback-Leibler divergence, we have

$$\begin{aligned} \text{KL}(\text{Ber}(p) \parallel \text{Ber}(q)) &= p \log \frac{p}{q} + (1-p) \log \frac{1-p}{1-q} = \int_q^p \left(\frac{p}{x} - \frac{1-p}{1-x} \right) dx \\ &= \int_q^p \frac{p-x}{x(1-x)} dx \geq \int_q^p \frac{p-x}{p(1-q)} dx = \frac{(p-q)^2}{2p(1-q)}. \end{aligned} \quad (3.3)$$

Thus we also have

$$\text{KL}(\text{Ber}(q)\|\text{Ber}(p)) = \text{KL}(\text{Ber}(1-q)\|\text{Ber}(1-p)) \geq \frac{(p-q)^2}{2p(1-q)}. \quad (3.4)$$

Moreover, by Theorem 1 of [AG89] and symmetry, it holds that

1. $\mathbb{P}(X \leq rN) \leq \exp(-N\text{KL}(\text{Ber}(r)\|\text{Ber}(p)))$, and
2. $\mathbb{P}(X \geq sN) \leq \exp(-N\text{KL}(\text{Ber}(s)\|\text{Ber}(p)))$.

The claimed tail bounds hence follow from (3.3) and (3.4). \square

3.6.1 Proof of Theorem 3.2.1

First, to achieve optimal upper bounds, we consider a variant of maximum likelihood estimation. Fix $\lambda \in (0, 1/2)$, $p \in (0, 1]$ and define $\varphi = np^{-1}\lambda^{-2}$ in the case of sampling model (O_1) , and $\varphi = n^3N^{-1}\lambda^{-2}$ in the case of sampling model (O_2) . If λ or p is unknown, one may learn these scalar parameters easily from the observations and define φ using the estimated values. For readability, we assume that they are given to avoid these technical complications.

Let \mathcal{P} be a maximal φ -packing (and thus a φ -net) of the symmetric group \mathfrak{S}_n with respect to d_{KT} . Consider the following estimator:

$$\hat{\pi} \in \operatorname{argmax}_{\pi \in \mathcal{P}} \sum_{\pi(i) > \pi(j)} A_{i,j}. \quad (3.5)$$

It is easy to see that $\hat{\pi}$ is the MLE of π^* over \mathcal{P} . Such an estimator is often called *sieve estimator* (see, e.g., [LC86]) in the statistics literature. The estimator $\hat{\pi}$ satisfies the following upper bounds.

Theorem 3.6.2. *Consider the noisy sorting model with underlying permutation π^* and probability matrix $M \in \mathfrak{M}_n(\lambda)$ where $\lambda \in (0, \frac{1}{2})$. Then, with probability at least $1 - e^{-n/8}$, the estimator $\hat{\pi}$ defined in (3.5) satisfies*

$$d_{\text{KT}}(\hat{\pi}, \pi^*) \lesssim \begin{cases} \frac{n}{p\lambda^2} \wedge n^2 & \text{in model } (O_1) \\ \frac{n^3}{N\lambda^2} \wedge n^2 & \text{in model } (O_2). \end{cases}$$

By integrating the tail probabilities of the above bounds, we easily obtain bounds on the expectation $\mathbb{E}[d_{\text{KT}}(\hat{\pi}, \pi^*)]$ of the same order, which then prove the upper bounds in Theorem 3.2.1. One may wonder whether the rate in Theorem 3.6.2 can be achieved by the MLE $\tilde{\pi}$ over \mathfrak{S}_n defined by

$$\tilde{\pi} \in \operatorname{argmax}_{\pi \in \mathfrak{S}_n} \sum_{\pi(i) > \pi(j)} A_{i,j}.$$

Our current techniques only allow us to prove bounds on $d_{\text{KT}}(\tilde{\pi}, \pi^*)$ that incur an extra factor $\log(1/p\lambda)$ (resp. $\log(n^2/N\lambda)$) in model (O_1) (resp. (O_2)). It is unclear whether these logarithmic factors can be removed for the MLE.

of *Theorem 3.6.2*. We assume that n is lower bounded by a constant without loss of generality, and note that the bounds of order n^2 are trivial. The proof is split into four parts to improve readability.

Basic setup. Since \mathcal{P} is a maximal φ -packing of \mathfrak{S}_n , it is also a φ -net and thus there exists $\tilde{\pi} \in \mathcal{P}$ such that $\mathfrak{D} := d_{\text{KT}}(\tilde{\pi}, \pi^*) \leq \varphi$. By definition of $\hat{\pi}$, $\sum_{\hat{\pi}(i) < \hat{\pi}(j)} A_{i,j} \leq \sum_{\tilde{\pi}(i) < \tilde{\pi}(j)} A_{i,j}$. Canceling concordant pairs (i, j) under $\hat{\pi}$ and $\tilde{\pi}$, we see that

$$\sum_{\hat{\pi}(i) < \hat{\pi}(j), \tilde{\pi}(i) > \tilde{\pi}(j)} A_{i,j} \leq \sum_{\hat{\pi}(i) > \hat{\pi}(j), \tilde{\pi}(i) < \tilde{\pi}(j)} A_{i,j}.$$

Splitting the summands according to π^* yields that

$$\sum_{\substack{\hat{\pi}(i) < \hat{\pi}(j), \\ \tilde{\pi}(i) > \tilde{\pi}(j), \\ \pi^*(i) < \pi^*(j)}} A_{i,j} + \sum_{\substack{\hat{\pi}(i) < \hat{\pi}(j), \\ \tilde{\pi}(i) > \tilde{\pi}(j), \\ \pi^*(i) > \pi^*(j)}} A_{i,j} \leq \sum_{\substack{\hat{\pi}(i) > \hat{\pi}(j), \\ \tilde{\pi}(i) < \tilde{\pi}(j), \\ \pi^*(i) < \pi^*(j)}} A_{i,j} + \sum_{\substack{\hat{\pi}(i) > \hat{\pi}(j), \\ \tilde{\pi}(i) < \tilde{\pi}(j), \\ \pi^*(i) > \pi^*(j)}} A_{i,j}.$$

Since $A_{i,j} \geq 0$, we may drop the leftmost term and drop the condition $\hat{\pi}(i) > \hat{\pi}(j)$ in the rightmost term to obtain that

$$\sum_{\substack{\hat{\pi}(i) < \hat{\pi}(j), \\ \tilde{\pi}(i) > \tilde{\pi}(j), \\ \pi^*(i) > \pi^*(j)}} A_{i,j} \leq \sum_{\substack{\hat{\pi}(i) > \hat{\pi}(j), \\ \tilde{\pi}(i) < \tilde{\pi}(j), \\ \pi^*(i) < \pi^*(j)}} A_{i,j} + \sum_{\substack{\hat{\pi}(i) < \hat{\pi}(j), \\ \tilde{\pi}(i) < \tilde{\pi}(j), \\ \pi^*(i) > \pi^*(j)}} A_{i,j}. \quad (3.6)$$

This inequality is crucial to proving that $\hat{\pi}$ is close to π^* with high probability.

To set up the rest of the proof, we define, for $\pi \in \mathcal{P}$,

$$\begin{aligned} L_\pi &= |\{(i, j) \in [n]^2 : \pi(i) < \pi(j), \tilde{\pi}(i) > \tilde{\pi}(j), \pi^*(i) > \pi^*(j)\}| \\ &= |\{(i, j) \in [n]^2 : \pi(i) > \pi(j), \tilde{\pi}(i) < \tilde{\pi}(j), \pi^*(i) < \pi^*(j)\}|. \end{aligned}$$

Moreover, define the random variables

$$X_\pi = \sum_{\substack{\pi(i) < \pi(j), \\ \tilde{\pi}(i) > \tilde{\pi}(j), \\ \pi^*(i) > \pi^*(j)}} A_{i,j}, \quad Y_\pi = \sum_{\substack{\pi(i) > \pi(j), \\ \tilde{\pi}(i) < \tilde{\pi}(j), \\ \pi^*(i) < \pi^*(j)}} A_{i,j}, \quad \text{and} \quad Z = \sum_{\substack{\tilde{\pi}(i) < \tilde{\pi}(j), \\ \pi^*(i) > \pi^*(j)}} A_{i,j}.$$

We will prove that the random process $X_\pi - Y_\pi - Z$ is positive with high probability if π is too far from $\tilde{\pi}$. However, (3.6) says precisely that $X_{\hat{\pi}} - Y_{\hat{\pi}} - Z \leq 0$, so that π must be close to $\tilde{\pi}$ which is in turn close to π^* .

The case $M = M_n^*(\lambda)$ under sampling model (O_1) . Consider model (O_1) of sampling without replacement, and suppose that $M = M_n^*(\lambda)$ first. For a pair (i, j)

with $\pi^*(i) > \pi^*(j)$, the entry $A_{i,j}$ has distribution $\text{Ber}(p(\frac{1}{2} + \lambda))$, since item i and item j are compared with probability p and conditioned on them being compared, item i wins with probability $\frac{1}{2} + \lambda$. Moreover, $A_{i,j}$ is independent from any other $A_{k,\ell}$ with $\pi^*(k) > \pi^*(\ell)$. Hence X_π has distribution $\text{Bin}(L_\pi, p(\frac{1}{2} + \lambda))$. Similarly, Y_π has distribution $\text{Bin}(L_\pi, p(\frac{1}{2} - \lambda))$, and Z has distribution $\text{Bin}(\mathfrak{D}, p(\frac{1}{2} + \lambda))$. Therefore, Lemma 3.6.1 implies that

1. $\mathbb{P}(X_\pi \leq L_\pi p(\frac{1}{2} + \frac{1}{2}\lambda)) \leq \exp(-L_\pi p \lambda^2/8)$, and
2. $\mathbb{P}(Y_\pi \geq L_\pi p(\frac{1}{2} - \frac{1}{2}\lambda)) \leq \exp(-L_\pi p \lambda^2/8)$.

Then we have that

$$\mathbb{P}(X_\pi - Y_\pi \leq L_\pi p \lambda) \leq 2 \exp(-L_\pi p \lambda^2/8). \quad (3.7)$$

For an integer $r \in [C\varphi, \binom{n}{2}]$ where C is a sufficiently large constant to be chosen, consider the slice $\mathcal{S}_r = \{\pi \in \mathcal{P} : L_\pi = r\}$. Note that if $\pi \in \mathcal{S}_r$, then

$$\begin{aligned} d_{\text{KT}}(\pi, \pi^*) &= |\{(i, j) : \hat{\pi}(i) < \hat{\pi}(j), \pi^*(i) > \pi^*(j)\}| \\ &\leq |\{(i, j) : \hat{\pi}(i) < \hat{\pi}(j), \tilde{\pi}(i) > \tilde{\pi}(j), \pi^*(i) > \pi^*(j)\}| \\ &\quad + |\{(i, j) : \tilde{\pi}(i) < \tilde{\pi}(j), \pi^*(i) > \pi^*(j)\}| \\ &= L_\pi + d_{\text{KT}}(\tilde{\pi}, \pi^*) \leq r + \varphi. \end{aligned} \quad (3.8)$$

Since \mathcal{P} is a φ -packing of \mathfrak{S}_n and $\mathcal{S}_r \subseteq \mathcal{P}$, we see that $|\mathcal{S}_r|$ is bounded by the φ -packing number of the ball $\mathcal{B}(\pi^*, r + \varphi)$ in the Kendall tau distance. Therefore, Proposition 3.5.1 gives

$$\log |\mathcal{S}_r| \leq n \log \frac{2n + 2r + 2\varphi}{\varphi} + 2n \leq n \log \frac{45r}{\varphi}.$$

By (3.7) and a union bound over \mathcal{S}_r , we see that $\min_{\pi \in \mathcal{S}_r} (X_\pi - Y_\pi) > cL_\pi p$ with probability at least

$$\begin{aligned} &1 - \exp\left(n \log \frac{45r}{\varphi} + \log 2 - \frac{rp\lambda^2}{8}\right) \\ &= 1 - \exp\left(n \log \frac{45r}{\varphi} + \log 2 - \frac{rn}{8\varphi}\right) \geq 1 - \exp(-2n), \end{aligned}$$

where the inequality holds because $r/\varphi \geq C$ for a sufficiently large constant C . Then a union bound over integers $r \in [C\varphi, \binom{n}{2}]$ yields that $X_\pi - Y_\pi > cL_\pi p$ for all $\pi \in \mathcal{P}$ such that $L_\pi \geq C\varphi$ with probability at least $1 - e^{-n}$.

Furthermore, since $Z \sim \text{Bin}(\mathfrak{D}, p(\frac{1}{2} + \lambda))$ and $\mathfrak{D} \leq \varphi$, Lemma 3.6.1 gives that

$$\mathbb{P}(Z \geq 2\varphi p) \leq \exp(-\varphi p/4) \leq \exp(-n/4).$$

Combining the bounds on $X_\pi - Y_\pi$ and Z , we conclude that with probability at least $1 - e^{-n/8}$,

$$X_\pi - Y_\pi - Z > cC\varphi p - 2\varphi p > 0$$

for all $\pi \in \mathcal{P}$ with $L_\pi \geq C\varphi$, as long as $C > 2/c$.

We have seen in (3.6) that $X_{\hat{\pi}} - Y_{\hat{\pi}} - Z \leq 0$, so $L_{\hat{\pi}} \leq C\varphi$ on the above event. By (3.8), $d_{\text{KT}}(\hat{\pi}, \pi^*) \leq L_{\hat{\pi}} + \varphi$ on the same event, which completes the proof for the model (O_1) .

The general case under sampling model (O_1) . Let us continue to use X_π, Y_π and Z to denote the above random variables under the noisy sorting model \mathcal{P} with probability matrix $M_n^*(\lambda)$, and use $\tilde{X}_\pi, \tilde{Y}_\pi$ and \tilde{Z} to denote the corresponding random variables under a general noisy sorting model $\tilde{\mathcal{P}}$ with $M \in \mathfrak{M}_n(\lambda)$. We couple the two models such that:

1. The sets of pairs of items being compared are the same (and if a pair is compared multiple times, the multiplicity is also the same);
2. For each pair (i, j) with $\pi^*(i) > \pi^*(j)$, if item i beats item j in a comparison in the model \mathcal{P} , then it also beats item j in the corresponding comparison in the model $\tilde{\mathcal{P}}$.

The second statement can be satisfied because the results of comparisons are Bernoulli random variables and $M_{\pi^*(i), \pi^*(j)} \geq [M_n^*(\lambda)]_{\pi^*(i), \pi^*(j)}$ for all $\pi^*(i) > \pi^*(j)$, by definition. Under this coupling, we always have that $\tilde{X}_\pi \geq X_\pi$ and $\tilde{Y}_\pi \leq Y_\pi$, so the above high probability lower bound on $X_\pi - Y_\pi$ also holds on $\tilde{X}_\pi - \tilde{Y}_\pi$.

Moreover recall the definition $\tilde{Z} = \sum_{\substack{\tilde{\pi}(i) < \tilde{\pi}(j) \\ \pi^*(i) > \pi^*(j)}} A_{i,j}$ where we have that $A_{i,j} \sim \text{Ber}(p[M_n^*(\lambda)]_{\pi^*(i), \pi^*(j)})$. Since $[M_n^*(\lambda)]_{\pi^*(i), \pi^*(j)} \in (0, 1)$, we can couple a sequence of i.i.d. $B_{i,j} \sim \text{Ber}(p)$ with the $A_{i,j}$'s in such a way that $B_{i,j} = 1$ whenever $A_{i,j} = 1$. Define $W = \sum_{\substack{\tilde{\pi}(i) < \tilde{\pi}(j) \\ \pi^*(i) > \pi^*(j)}} B_{i,j}$. Then we see that $W \sim \text{Bin}(\mathfrak{D}, p)$ and $W \geq \tilde{Z}$. Since $\mathfrak{D} \leq \varphi$, Lemma 3.6.1 gives

$$\mathbb{P}(W \geq 2\varphi p) \leq \exp(-\varphi p/4) \leq \exp(-n/4).$$

Thus \tilde{Z} is subject to the same high probability upper bound as Z . Therefore, the proof for the model \mathcal{P} also works to show the desired bound for the model $\tilde{\mathcal{P}}$.

Sampling model (O_2) . The proof for model (O_2) of sampling with replacement is essentially the same, except the part of probability bounds where we assume $M = M_n^*(\lambda)$. We now demonstrate the differences in detail. For a single pairwise comparison sampled uniformly from the possible $\binom{n}{2}$ pairs, the probability that

1. the chosen pair (i, j) satisfies $\pi(i) < \pi(j)$, $\tilde{\pi}(i) > \tilde{\pi}(j)$ and $\pi^*(i) > \pi^*(j)$, and
2. item i wins the comparison,

is equal to $L_\pi \binom{n}{2}^{-1} (\frac{1}{2} + \lambda)$. By definition, X_π is the number of times the above event happens if N independent pairwise comparisons take place, so we have that $X_\pi \sim \text{Bin}(N, L_\pi \binom{n}{2}^{-1} (\frac{1}{2} + \lambda))$. Similarly, we have $Y_\pi \sim \text{Bin}(N, L_\pi \binom{n}{2}^{-1} (\frac{1}{2} - \lambda))$ and $Z \sim \text{Bin}(N, \mathfrak{D} \binom{n}{2}^{-1} (\frac{1}{2} + \lambda))$. Hence Lemma 3.6.1 gives that

1. $\mathbb{P}(X_\pi \leq L_\pi N \binom{n}{2}^{-1} (\frac{1}{2} + \frac{1}{2}\lambda)) \leq \exp(-L_\pi N \binom{n}{2}^{-1} \lambda^2/8)$,
2. $\mathbb{P}(Y_\pi \geq L_\pi N \binom{n}{2}^{-1} (\frac{1}{2} - \frac{1}{2}\lambda)) \leq \exp(-L_\pi N \binom{n}{2}^{-1} \lambda^2/8)$, and
3. $\mathbb{P}(Z \geq 2\varphi N \binom{n}{2}^{-1}) \leq \exp(-\varphi N \binom{n}{2}^{-1}/4)$.

Note that if we set $p = N \binom{n}{2}^{-1}$, then the tail bounds above are exactly the same as those for the model (O_1) . Therefore, replacing p by $N \binom{n}{2}^{-1}$ everywhere in the above proof, we then obtain the desired bound for the model (O_2) . \square

Next, we turn to the lower bounds. Let $\mathbb{P}_{\pi^*} = \mathbb{P}_{\pi^*, M_n^*(\lambda)}$ denote the probability distribution of the observations in the noisy sorting model with underlying permutation $\pi^* \in \mathfrak{S}_n$ and probability matrix $M_n^*(\lambda)$, where $\lambda \in (0, \frac{1}{2})$. We prove the following stronger statement which clearly implies the lower bounds in Theorem 3.2.1.

Theorem 3.6.3. *For the sampling model (O_1) , suppose we have $\lambda \in (0, \frac{1}{2})$ and $p \in (0, 1]$ such that $p \log \frac{1}{1-2\lambda} \leq C$ for some constant $C > 0$. Then it holds that*

$$\min_{\tilde{\pi}} \max_{\pi^* \in \mathfrak{S}_n} \mathbb{P}_{\pi^*} \left(d_{\text{KT}}(\tilde{\pi}, \pi^*) \gtrsim \frac{n}{p\lambda^2} \wedge \frac{n}{p \log \frac{1}{1-2\lambda}} \wedge n^2 \right) \geq c,$$

where the minimum is taken minimized over all permutation estimators $\tilde{\pi} \in \mathfrak{S}_n$ that are measurable with respect to the observations and c is a universal positive constant. Similarly, for the sampling model (O_2) , if we have $Nn^{-2} \log \frac{1}{1-2\lambda} \leq C$, then it holds that

$$\min_{\tilde{\pi}} \max_{\pi^* \in \mathfrak{S}_n} \mathbb{P}_{\pi^*} \left(d_{\text{KT}}(\tilde{\pi}, \pi^*) \gtrsim \frac{n^3}{N\lambda^2} \wedge \frac{n^3}{N \log \frac{1}{1-2\lambda}} \wedge n^2 \right) \geq c.$$

Compared to the lower bounds in Theorem 3.2.1, the above lower bounds hold in probability, weaken the condition that λ is bounded away from $1/2$ and only require maximizing π^* instead of both π^* and M , and are therefore stronger.

One key ingredient in proving lower bounds is to relate the Kullback-Leibler divergence between model distributions to the distance measuring the error (see, e.g., Chapter 2 of [Tsy09]). This is achieved in the following lemma for both sampling models.

Lemma 3.6.4. *Fix $\pi, \sigma \in \mathfrak{S}_n$ and $\lambda \in (0, \frac{1}{2})$. We denote by \mathbb{P}_π the probability distribution of the noisy sorting model with underlying permutation π . Then for the sampling model (O_1) we have*

$$\text{KL}(\mathbb{P}_\pi \| \mathbb{P}_\sigma) = 2 d_{\text{KT}}(\pi, \sigma) p \lambda \log \frac{1+2\lambda}{1-2\lambda},$$

and for the sampling model (O_2) we have

$$\text{KL}(\mathbb{P}_\pi \| \mathbb{P}_\sigma) = 2 d_{\text{KT}}(\pi, \sigma) N \binom{n}{2}^{-1} \lambda \log \frac{1+2\lambda}{1-2\lambda}.$$

Proof. First, we consider model (O_1) of sampling without replacement. For $i \neq j$, let $\mathbb{P}_\pi^{(i,j)}$ denote the distribution of outcomes between i and j , or more formally, the distribution of $N_{i,j}$ and $A_{i,j}$. For a pair (i, j) such that $\pi(i) > \pi(j)$ and $\sigma(i) > \sigma(j)$, the distributions $\mathbb{P}_\pi^{(i,j)}$ and $\mathbb{P}_\sigma^{(i,j)}$ are indistinguishable. For (i, j) such that $\pi(i) > \pi(j)$ and $\sigma(i) < \sigma(j)$, the probability that i and j are not compared stays the same, but the probability that they are compared and i wins the comparison is $p(\frac{1}{2} + \lambda)$ under $\mathbb{P}_\pi^{(i,j)}$ while it is $p(\frac{1}{2} - \lambda)$ under $\mathbb{P}_\sigma^{(i,j)}$. A symmetric statement holds for the probability that they are compared and j wins the comparison. Therefore, we obtain that

$$\begin{aligned} \text{KL}(\mathbb{P}_\pi^{(i,j)} \parallel \mathbb{P}_\sigma^{(i,j)}) &= p(1/2 + \lambda) \log \frac{1/2 + \lambda}{1/2 - \lambda} + p(1/2 - \lambda) \log \frac{1/2 - \lambda}{1/2 + \lambda} \\ &= 2p\lambda \log \frac{1 + 2\lambda}{1 - 2\lambda}. \end{aligned}$$

It follows from the chain rule that

$$\text{KL}(\mathbb{P}_\pi \parallel \mathbb{P}_\sigma) = \sum_{\pi(i) > \pi(j), \sigma(i) < \sigma(j)} \text{KL}(\mathbb{P}_\pi^{i,j} \parallel \mathbb{P}_\sigma^{i,j}) = 2 d_{\kappa\tau}(\pi, \sigma) p\lambda \log \frac{1 + 2\lambda}{1 - 2\lambda},$$

which proves the claimed bound.

Next, we move on to model (O_2) of sampling with replacement. In this case, for the noisy sorting model with underlying permutation π , we let \mathbb{Q}_π denote the distribution of the outcome of a single pairwise comparison chosen uniformly from the $\binom{n}{2}$ possible pairs. Conditioned on a pair (i, j) with $\pi(i) > \pi(j)$ and $\sigma(i) > \sigma(j)$ being chosen, the outcome is indistinguishable under \mathbb{Q}_π and \mathbb{Q}_σ . On the other hand, conditioned on having chosen (i, j) with $\pi(i) > \pi(j)$ and $\sigma(i) < \sigma(j)$, the probability that i wins the comparison is $p(\frac{1}{2} + \lambda)$ under \mathbb{Q}_π and is $p(\frac{1}{2} - \lambda)$ under \mathbb{Q}_σ . By the definition of the KL divergence, we have

$$\begin{aligned} \text{KL}(\mathbb{Q}_\pi \parallel \mathbb{Q}_\sigma) &= \sum_{\pi(i) > \pi(j), \sigma(i) < \sigma(j)} \left[\binom{n}{2}^{-1} (1/2 + \lambda) \log \frac{1/2 + \lambda}{1/2 - \lambda} \right. \\ &\quad \left. + \binom{n}{2}^{-1} (1/2 - \lambda) \log \frac{1/2 - \lambda}{1/2 + \lambda} \right] \\ &= 2 d_{\kappa\tau}(\pi, \sigma) \binom{n}{2}^{-1} \lambda \log \frac{1 + 2\lambda}{1 - 2\lambda}, \end{aligned}$$

where the bound holds similarly as above. Since N independent pairwise comparisons are observed and the KL divergence tensorizes, the conclusion follows. \square

We are ready to prove the minimax lower bound.

of Theorem 3.6.3. Consider the sampling model (O_1). We assume that n is lower bounded by a constant, and use the shorthand notation $\kappa = 4p\lambda \log \frac{1+2\lambda}{1-2\lambda}$. Note that $\kappa \leq C$ for some constant $C > 0$ by the assumption. Let $r = c_0 n \kappa^{-1} \wedge \binom{n}{2}$ and $\varepsilon = c_1 r$,

where c_0 and c_1 are constants to be chosen. Let \mathcal{P} be a maximal ε -packing of $\mathcal{B}(\text{id}, r)$, which is thus an ε -net by maximality. For any $\pi, \sigma \in \mathcal{P}$, we have $d_{\text{KT}}(\pi, \sigma) \leq 2r$, so Lemma 3.6.4 yields

$$\text{KL}(\mathbb{P}_\pi \| \mathbb{P}_\sigma) = \frac{1}{2} \kappa d_{\text{KT}}(\pi, \sigma) \leq \kappa r \leq c_0 n.$$

On one hand, if $\kappa \leq c_2$ for a sufficiently small constant $c_2 > 0$, then $r \geq c_0 c_2^{-1} n \wedge \binom{n}{2}$ and thus Proposition 3.5.1 implies that

$$\log |\mathcal{P}| \geq n \log \frac{r}{n + \varepsilon} - 2n \geq 10 c_0 n \geq 10 \text{KL}(\mathbb{P}_\pi \| \mathbb{P}_\sigma),$$

where we take $c_0 = 1$ and c_1, c_2 small enough for the inequalities to hold.

On the other hand, if $c_2 < \kappa \leq C$, then we take $c_1 = 1/8$ and c_0 sufficiently small so that $r \leq c_0 c_2^{-1} n < n/2$. Then we can apply Lemma 3.5.3 to obtain

$$\log |\mathcal{P}| \geq \frac{r}{5} \log \frac{n}{r} \geq \frac{c_0 n}{5C} \log \frac{c_2}{c_0} \geq 10 c_0 n \geq 10 \text{KL}(\mathbb{P}_\pi \| \mathbb{P}_\sigma),$$

where the second inequality holds since $c_0 C^{-1} n \leq r \leq c_0 c_2^{-1} n$ and the third inequality holds for c_0 small enough.

In either case, we have $\text{KL}(\mathbb{P}_\pi \| \mathbb{P}_\sigma) \leq 0.1 \log |\mathcal{P}|$. Therefore, using [Tsy09, Theorem 2.5] yields the lower bound of order $r \asymp n \kappa^{-1} \wedge n^2$. Considering the limiting behavior of κ as $\lambda \rightarrow 0$ and $\lambda \rightarrow \frac{1}{2}$ respectively, we see that $\kappa \lesssim p \lambda^2 \vee p \log \frac{1}{1-2\lambda}$, so the claimed lower bound follows.

For the sampling model (O_2) , the same argument follows if we replace p with $N \binom{n}{2}^{-1}$. \square

3.6.2 Proof of Theorem 3.2.2

Without loss of generality, assume that $\pi^* = \text{id}$ and n is even to simplify the notation. We define a score

$$s_i^* = \sum_{j \in [n] \setminus \{i\}} M_{i,j} = \lambda(2i - n - 1) + (n - 1)/2$$

for each $i \in [n]$, which is simply the i -th row sum of M minus $1/2$. Analogously, we define

$$\hat{s}_i = \sum_{j=1}^{i-1} \left(\frac{1}{2} + \hat{\lambda}\right) + \sum_{j=i+1}^n \left(\frac{1}{2} - \hat{\lambda}\right) = \hat{\lambda}(2i - n - 1) + (n - 1)/2$$

for each $i \in [n]$, which is a slightly perturbed version of s_i^* due to the difference between λ and $\hat{\lambda}$. The MS algorithm is designed to refine estimates for the scores s_i^* in multiple stages.

First, the estimator $\hat{\lambda}$ satisfies the following bound, which in particular implies that \hat{s}_i is close to s_i^* .

Lemma 3.6.5. *If $N \geq Cn \log n$, then we have $|\hat{\lambda} - \lambda| \leq C_0 \sqrt{N^{-1} \log n}$ with probability at least $1 - n^{-8}$, where C and C_0 are sufficiently large universal constants.*

Proof. Consider a single pairwise comparison chosen uniformly from the $\binom{n}{2}$ pairs. The probability that item i is chosen and wins the comparison is therefore equal to $(\sum_{j \in [n] \setminus \{i\}} M_{i,j}) / \binom{n}{2} = s_i^* / \binom{n}{2}$. Thus the random variable $S_i = \sum_{j=1}^n A'_{i,j}$ has distribution $\text{Bin}(N/2, s_i^* / \binom{n}{2})$. Hence Lemma 3.6.1 implies that

$$\mathbb{P}(|S_i - \mathbb{E}[S_i]| \geq c_1 \mathbb{E}[S_i]) \leq 2 \exp(-c_2 \mathbb{E}[S_i]) \leq n^{-10},$$

where the last inequality holds since $N \geq Cn \log n$, and we use c_1, c_2, \dots to denote sufficiently small constants. A union bound shows that with probability at least $1 - n^{-9}$, we have $|S_i - \mathbb{E}[S_i]| \leq c_1 \mathbb{E}[S_i]$ for all $i \in [n]$. Denote this high probability event by \mathcal{E} , and we condition on \mathcal{E} henceforth.

Recall that $s_i^* = 2\lambda i - \lambda(n+1) + (n-1)/2$. Using that λ is bounded away from zero, we can choose c_1 small enough so that if $i - j \geq n/4$, then $s_i^* - s_j^* > 2c_1 s_i^*$. Note that $E[S_i] = \frac{1}{2} N s_i^* / \binom{n}{2}$, so $E[S_i] - E[S_j] > 2c_1 E[S_i]$ if $i - j \geq n/4$. Therefore, on the event \mathcal{E} we have $S_i > S_j$ for all (i, j) with $i - j \geq n/4$. It follows that $\tilde{\pi}(i) > \tilde{\pi}(j)$ for these pairs (i, j) , as $\tilde{\pi}$ is defined by sorting the scores S_i .

Next consider (i, j) such that $\tilde{\pi}(i) - \tilde{\pi}(j) > n/2$. Suppose we have $i < j$. Then there exists $k \in [n]$ with $\tilde{\pi}(j) < \tilde{\pi}(k) < \tilde{\pi}(i)$ such that either $k - i \geq n/4$ or $j - k \geq n/4$, which gives a contradiction on the event \mathcal{E} . Therefore, it holds that $i > j$ for all pairs (i, j) with $\tilde{\pi}(i) - \tilde{\pi}(j) > n/2$.

Recall that $\hat{\lambda} = \frac{2}{N} \binom{n}{2} \binom{n/2}{2}^{-1} \sum_{(i,j) \in \mathcal{I}} A''_{i,j} - \frac{1}{2}$, where $\mathcal{I} = \{(i, j) \in [n]^2 : \tilde{\pi}(i) - \tilde{\pi}(j) > \frac{n}{2}\}$. Note that A'' is independent of \mathcal{E} , on which we have $i > j$ for all $(i, j) \in \mathcal{I}$. Similar to the argument at the beginning of the proof, the probability that a uniformly chosen pair falls in \mathcal{I} and i wins the comparison is $(\frac{1}{2} + \lambda) |\mathcal{I}| / \binom{n}{2}$. Hence the random variable $X := \sum_{(i,j) \in \mathcal{I}} A''_{i,j}$ has distribution $\text{Bin}(N/2, (\frac{1}{2} + \lambda) |\mathcal{I}| / \binom{n}{2})$. It follows that $\mathbb{E}[\hat{\lambda} | \mathcal{E}] = \lambda$ once we note that $|\mathcal{I}| = \binom{n/2}{2}$.

Moreover, Lemma 3.6.1 gives the bound

$$\mathbb{P}\left(|X - \mathbb{E}[X]| \geq C_2 \sqrt{N \log n} \mid \mathcal{E}\right) \leq 2 \exp(-c_3 \log n) \leq n^{-9},$$

and consequently $|\hat{\lambda} - \lambda| \leq C_0 \sqrt{N^{-1} \log n}$ with probability at least $1 - n^{-9}$ conditioned on the event \mathcal{E} , where C_2 and C_0 are sufficiently large constants. A union bound then completes the proof. \square

We condition on the high probability event of Lemma 3.6.5 throughout the rest of the proof, so that $|\hat{\lambda} - \lambda| \leq C_0 \sqrt{N^{-1} \log n}$ for a fixed constant $C_0 > 0$. In particular, $\hat{\lambda}$ is bounded away from zero by a universal constant since λ is and $N \geq Cn \log n$, and $\hat{s}_j < \hat{s}_i$ iff $j < i$. We proceed with the following key lemma.

Lemma 3.6.6. *Fix $t \in [T]$, $i \in [n]$ and $I \subseteq [n]$ with $i \in I$. Suppose that $|I| \geq$*

$C_1 \frac{n^2 T}{N} \log(nT)$ for a sufficiently large constant C . If we define

$$S = \frac{Tn(n-1)}{2N} \sum_{j \in I} A_{i,j}^{(t)} + \sum_{j \in [n] \setminus I, j < i} \left(\frac{1}{2} + \hat{\lambda}\right) + \sum_{j \in [n] \setminus I, j > i} \left(\frac{1}{2} - \hat{\lambda}\right),$$

then it holds with probability at least $1 - 2(nT)^{-9}$ that

$$|S - \hat{s}_i| \leq (5 + C_0)n\sqrt{|I|TN^{-1}\log(nT)}.$$

Proof. Consider a single pairwise comparison chosen uniformly from the $\binom{n}{2}$ pairs. The probability that the chosen pair consists of item i and an item in $I \setminus \{i\}$, and that item i wins the comparison, is equal to $q := (\sum_{j \in I \setminus \{i\}} M_{i,j}) / \binom{n}{2}$. Thus the random variable $X := \sum_{j \in I} A_{i,j}^{(t)}$ has distribution $\text{Bin}(N/T, q)$. In particular, we have $\mathbb{E}[X] = Nq/T = \frac{2N}{Tn(n-1)} \sum_{j \in I \setminus \{i\}} M_{i,j}$ and by Lemma 3.6.1,

$$\mathbb{P}\left(|X - \mathbb{E}[X]| \geq \frac{rN}{T}\right) \leq 2 \exp\left(-\frac{Nr^2}{2T(q+r)}\right).$$

Taking $r = 6\sqrt{\frac{Tq}{N} \log(nT)}$, we see that $r \leq q$ by the assumption $|I| \geq C_1 \frac{n^2 T}{N} \log(nT)$, so

$$\mathbb{P}\left(|X - \mathbb{E}[X]| \geq 6\sqrt{qNT^{-1}\log(nT)}\right) \leq 2(nT)^{-9}. \quad (3.9)$$

By the definitions of S and \hat{s}_i , it is straightforward to verify that

$$S - \hat{s}_i = \frac{Tn(n-1)}{2N} (X - \mathbb{E}[X]) + \sum_{j \in I, j < i} (\lambda - \hat{\lambda}) + \sum_{j \in I, j > i} (\hat{\lambda} - \lambda).$$

Therefore, we obtain from (3.9), the definition of q and the fact $|I| \leq n$ that

$$\begin{aligned} |S - \hat{s}_i| &\leq 3n(n-1)\sqrt{qTN^{-1}\log(nT)} + |I| |\hat{\lambda} - \lambda| \\ &\leq 5n\sqrt{|I|TN^{-1}\log(nT)} + C_0|I|\sqrt{N^{-1}\log n} \\ &\leq (5 + C_0)n\sqrt{|I|TN^{-1}\log(nT)} \end{aligned}$$

with probability at least $1 - 2(nT)^{-9}$. \square

To analyze the MS algorithm, we apply Lemma 3.6.6 inductively to each stage of the algorithm. Define $\mathcal{E}^{(0)}$ to be the full event. As the inductive hypothesis, we assume that on the event $\mathcal{E}^{(t-1)}$, it holds that $j < i$ for all $j \in I_-^{(t-1)}(i)$ and $j > i$ for all $j \in I_+^{(t-1)}(i)$. In particular, this holds trivially for $t = 1$.

On the event $\mathcal{E}^{(t-1)}$, the score $S_i^{(t)}$ is exactly the quantity S in Lemma 3.6.6 with $I = I^{(t-1)}(i)$. Thus the lemma shows that if $|I^{(t-1)}(i)| \geq C_1 \frac{n^2 T}{N} \log(nT)$ for a large enough constant C_1 , then

$$|S_i^{(t)} - \hat{s}_i| \leq (5 + C_0)n\sqrt{|I^{(t-1)}(i)|TN^{-1}\log(nT)} = \tau_i^{(t)}/2 \quad (3.10)$$

with probability at least $1 - 2(nT)^{-9}$ conditional on $\mathcal{E}^{(t-1)}$. We denote by $\mathcal{E}^{(t)}$ the sub-event of $\mathcal{E}^{(t-1)}$ that the above bound holds for all $i \in [n]$. Then $\mathbb{P}(\mathcal{E}^{(t)} | \mathcal{E}^{(t-1)}) \geq 1 - (nT)^{-8}$ and we condition on $\mathcal{E}^{(t)}$ henceforth.

For any $j \in I_-^{(t)}(i)$, by definition $S_j^{(t)} - S_i^{(t)} < -\tau_i^{(t)}$, so we have $\hat{s}_j < \hat{s}_i$ and thus $j < i$. Similarly, $j > i$ for any $j \in I_+^{(t)}(i)$ on the event $\mathcal{E}^{(t)}$. Hence the inductive hypothesis is verified. Moreover, note that $I^{(t)}(i) = \{j \in [n] : |S_j^{(t)} - S_i^{(t)}| \leq 2\tau_i^{(t)}\} \subseteq \{j \in [n] : |\hat{s}_j - \hat{s}_i| \leq 3\tau_i^{(t)}\}$. Since $\hat{s}_j - \hat{s}_i = 2\hat{\lambda}(j - i)$ and $\hat{\lambda}$ is bounded away from zero by a universal constant, we have

$$|I^{(t)}(i)| \leq C_2 \tau_i^{(t)} = C_3 n \sqrt{|I^{(t-1)}(i)| T N^{-1} \log(nT)}, \quad (3.11)$$

where we use C_2, C_3, \dots to denote sufficiently large constants.

Note that if we have $\alpha^{(0)} = n$ and the iterative relation $\alpha^{(t)} \leq \beta \sqrt{\alpha^{(t-1)}}$ where $\alpha^{(t)} > 0$ and $\beta > 0$, then it is easily seen that $\alpha^{(t)} \leq \beta^2 n^{2^{-t}}$. We would like to obtain such a bound from the relation (3.11). Note that $\mathcal{E}^{(T)} \subseteq \mathcal{E}^{(T-1)} \subseteq \dots \subseteq \mathcal{E}^{(0)}$ by definition and $\mathbb{P}(\mathcal{E}^{(T)}) = \prod_{t=1}^T \mathbb{P}(\mathcal{E}^{(t)} | \mathcal{E}^{(t-1)}) \geq 1 - n^{-8}$. Conditional on $\mathcal{E}^{(T)}$, the iterative relation (3.11) thus holds for all $t \in [T]$, and we have $|I^{(0)}(i)| = n$ by definition. Since $I^{(t)}(i)$ is not updated in the algorithm once $|I^{(t)}(i)| \leq C_1 \frac{n^{2^t}}{N} \log(nT)$, we obtain that

$$\begin{aligned} |I^{(T-1)}(i)| &\leq \left(C_3^2 \frac{n^{2T}}{N} \log(nT) n^{2^{-T+1}} \right) \vee \left(C_1 \frac{n^{2T}}{N} \log(nT) \right) \\ &\leq C_4 \frac{n^2}{N} (\log n) (\log \log n), \end{aligned}$$

where the last bound holds because we take $T = \lfloor \log \log n \rfloor$. Hence it follows from (3.10) that

$$|S_i^{(T)} - \hat{s}_i| \leq C_5 n^2 N^{-1} (\log n) (\log \log n),$$

and a similar argument as above shows that $S_i^{(T)} > S_j^{(T)}$ for all pairs (i, j) with $i - j > C_6 n^2 N^{-1} (\log n) (\log \log n) =: \delta$. As the permutation $\hat{\pi}^{\text{MS}}$ is defined by sorting the scores $S_i^{(T)}$ in increasing order, we see that $\hat{\pi}^{\text{MS}}(i) > \hat{\pi}^{\text{MS}}(j)$ for pairs (i, j) with $i - j > \delta$.

Finally, suppose that $\hat{\pi}^{\text{MS}}(i) - i < -\delta$ for some $i \in [n]$. Then there exists $j < i - \delta$ such that $\hat{\pi}^{\text{MS}}(j) > \hat{\pi}^{\text{MS}}(i)$, contradicting the guarantee we have just proved. A similar argument leads to a contradiction if $\hat{\pi}^{\text{MS}}(i) - i > \delta$. Therefore, we obtain that

$$|\hat{\pi}^{\text{MS}}(i) - i| \leq \delta = C_6 n^2 N^{-1} (\log n) (\log \log n)$$

for all $i \in [n]$, which completes the proof.

Chapter 4

Faster Rates for Permutation-based Models in Polynomial Time

Structured matrices with entries in the range $[0, 1]$ and unknown permutations acting on their rows and columns arise in multiple applications, including estimation from pairwise comparisons [BT52, SBGW17] and crowd-labeling [DS79, SBW16b]. Traditional parametric models [BT52, Luc59, Thu27, DS79] assume that these matrices are obtained from rank-one matrices via a known link function. Aided by tools such as maximum likelihood estimation and spectral methods, researchers have made significant progress in studying both statistical and computational aspects of these parametric models [HOX14, RA14, SBB⁺16, NOS16, ZCZJ16, GZ13, GLZ16, KOS11b, LPI12, DDKR13, GKM11] and their low-rank generalizations [RA16, NOTX17, KOS11a].

There has been evidence from empirical studies (e.g., [ML65, BW97]) that real-world data is not always well-captured by such parametric models. With the goal of increasing model flexibility, a recent line of work has studied the class of *permutation-based* models [Cha15, SBGW17, SBW16b]. Rather than imposing parametric conditions on the matrix entries, these models impose only shape constraints on the matrix, such as monotonicity, before unknown permutations act on the its rows and columns. This more flexible class reduces modeling bias compared to its parametric counterparts while, perhaps surprisingly, producing models that can be estimated at rates that differ only by logarithmic factors from parametric models. On the negative side, these advantages of permutation-based models are accompanied by significant computational challenges. The unknown permutations make the parameter space highly non-convex, so that efficient maximum likelihood estimation is unlikely. Moreover, spectral methods are often suboptimal in approximating shape-constrained sets of matrices [Cha15, SBGW17]. Consequently, results from many recent papers show a non-trivial statistical-computational gap in estimation rates for models with latent permutations [SBGW17, CM16, SBW16b, FMR16, PWC17].

Related work. While the main motivation of our work comes from nonparametric methods for aggregating pairwise comparisons, we begin by discussing a few other lines of related work. The current paper lies at the intersection of shape-constrained estimation and latent permutation learning. Shape-constrained estimation has long

been a major topic in nonparametric statistics, and of particular relevance to our work is the estimation of a bivariate isotonic matrix without latent permutations [CGS18]. There, it was shown that the minimax rate of estimating an $n \times n$ matrix from noisy observations of all its entries is $\tilde{\Theta}(n^{-1})$. The upper bound is achieved by the least squares estimator, which is efficiently computable due to the convexity of the parameter space.

Shape-constrained matrices with permuted rows or columns also arise in applications such as seriation [FJBd13, FMR16] and feature matching [CD16]. In particular, the monotone subclass of the statistical seriation model [FMR16] contains $n \times n$ matrices that have increasing columns, and an unknown row permutation. The authors established the minimax rate $\tilde{\Theta}(n^{-2/3})$ for estimating matrices in this class and proposed a computationally efficient algorithm with rate $\tilde{\mathcal{O}}(n^{-1/2})$. For the subclass of such matrices where in addition, the rows are also monotone, the results of the current paper improve the two rates to $\tilde{\mathcal{O}}(n^{-1})$ and $\tilde{\mathcal{O}}(n^{-3/4})$ respectively.

Another related model is that of noisy sorting [BM08], which involves a latent permutation but no shape-constraint. In this prototype of a permutation-based ranking model, we have an unknown, $n \times n$ matrix with constant upper and lower triangular portions whose rows and columns are acted upon by an unknown permutation. The hardness of recovering any such matrix in noise lies in estimating the unknown permutation. As it turns out, this class of matrices can be estimated efficiently at minimax optimal rate $\tilde{\Theta}(n^{-1})$ by multiple procedures: the original work by Braverman and Mossel [BM08] proposed an algorithm with time complexity $\mathcal{O}(n^c)$ for some unknown and large constant c , and recently, an $\tilde{\mathcal{O}}(n^2)$ -time algorithm was proposed by Mao et al. [MWR17]. These algorithms, however, do not generalize beyond the noisy sorting class, which constitutes a small subclass of an interesting class of matrices that we describe next.

The most relevant body of work to the current paper is that on estimating matrices satisfying the *strong stochastic transitivity* condition, or SST for short. This class of matrices contains all $n \times n$ bivariate isotonic matrices with unknown permutations acting on their rows and columns, with an additional skew-symmetry constraint. The first theoretical study of these matrices was carried out by Chatterjee [Cha15], who showed that a spectral algorithm achieved the rate $\tilde{\mathcal{O}}(n^{-1/4})$ in the normalized Frobenius norm. Shah et al. [SBGW17] then showed that the minimax rate of estimation is given by $\tilde{\Theta}(n^{-1})$, and also improved the analysis of the spectral estimator of Chatterjee [Cha15] to obtain the computationally efficient rate $\tilde{\mathcal{O}}(n^{-1/2})$. In follow-up work [SBW16a], they also showed a second CRL estimator based on the Borda count that achieved the same rate, but in near-linear time. In related work, Chatterjee and Mukherjee [CM16] analyzed a variant of the CRL estimator, showing that for subclasses of SST matrices, it achieved rates that were faster than $\mathcal{O}(n^{-1/2})$. In a complementary direction, a superset of the current authors [PMM⁺17a] analyzed the estimation problem under an observation model with structured missing data, and showed that for many observation patterns, a variant of the CRL estimator was minimax optimal.

Shah et al. [SBW16a] also showed that conditioned on the planted clique con-

jecture, it is impossible to improve upon a certain notion of adaptivity of the CRL estimator in polynomial time. Such results have prompted various authors [FMR16, SBW16a] to conjecture that a similar statistical-computational gap also exists when estimating SST matrices in the Frobenius norm.

Our contributions. Our main contribution in the current work is to tighten the aforementioned statistical-computational gap. More precisely, we study the problem of estimating a bivariate isotonic matrix with unknown permutations acting on its rows and columns, given noisy, partial observations of its entries; this matrix class strictly contains the SST model [Cha15, SBGW17] for ranking from pairwise comparisons. As a corollary of our results, we show that when the underlying matrix has dimension $n \times n$ and $\Theta(n^2)$ noisy entries are observed, our polynomial-time, two-dimensional sorting algorithm provably achieves the rate of estimation $\tilde{\mathcal{O}}(n^{-3/4})$ in the normalized Frobenius norm; thus, this result breaks the previously mentioned $\tilde{\mathcal{O}}(n^{-1/2})$ barrier [SBGW17, CM16]. Although the rate $\tilde{\mathcal{O}}(n^{-3/4})$ still differs from the minimax optimal rate $\tilde{\Theta}(n^{-1})$, our algorithm is, to the best of our knowledge, the first efficient procedure to obtain a rate faster than $\tilde{\mathcal{O}}(n^{-1/2})$ uniformly over the SST class. This guarantee, which is stated in slightly more technical terms below, can be significant in practice (see Figure 4-1).

Main theorem (informal) *There is an estimator \widehat{M} computable in time $\mathcal{O}(n^{2.5})$ such that for any $n \times n$ SST matrix M^* , given $\Theta(n^2)$ Bernoulli observations of its entries, we have*

$$\mathbb{E} \left[\frac{1}{n^2} \|\widehat{M} - M^*\|_F^2 \right] \leq C \left(\frac{\log n}{n} \right)^{3/4}.$$

Our algorithm is novel in the sense that it is neither spectral in nature, nor simple variations of the Borda count estimator that was previously employed. Our algorithm takes advantage of the fine monotonicity structure of the underlying matrix along both dimensions, and this allows us to prove tighter bounds than before. In addition to making algorithmic contributions, we also briefly revisit the minimax rates of estimation.

Organization. In Section 4.1, we formally introduce our estimation problem. Section 4.2 contains statements and discussions of our main results, and in Section 4.3, we describe in detail how the estimation problem that we study is connected to applications in crowd-labeling and ranking from pairwise comparisons. We provide the proofs of our main results in Section 4.4.

Notation. For a positive integer n , let $[n] := \{1, 2, \dots, n\}$. For a finite set S , we use $|S|$ to denote its cardinality. For two sequences $\{a_n\}_{n=1}^\infty$ and $\{b_n\}_{n=1}^\infty$, we write $a_n \lesssim b_n$ if there is a universal constant C such that $a_n \leq Cb_n$ for all $n \geq 1$. The

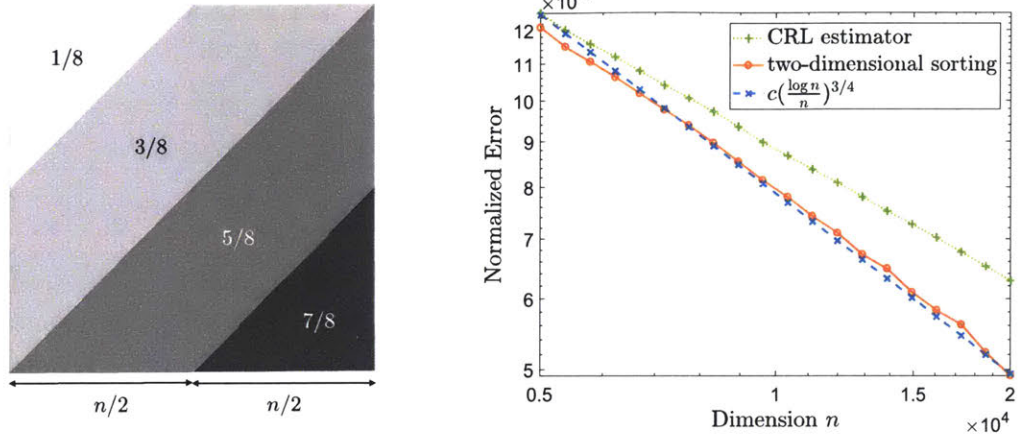


Figure 4-1: **Left:** A bivariate isotonic matrix; $M^* \in [0, 1]^{n \times n}$ is a row and column permuted version of such a matrix. **Right:** A log-log plot of the error $\frac{1}{n^2} \|\widehat{M} - M^*\|_F^2$ (averaged over 10 experiments each using n^2 Bernoulli observations) of our estimator and the CRL estimator [SBW16a].

relation $a_n \gtrsim b_n$ is defined analogously. We use c, C, c_1, c_2, \dots to denote universal constants that may change from line to line. We use $\text{Ber}(p)$ to denote the Bernoulli distribution with success probability p , the notation $\text{Bin}(n, p)$ to denote the binomial distribution with n trials and success probability p , and the notation $\text{Poi}(\lambda)$ to denote the Poisson distribution with parameter λ . Given a matrix $M \in \mathbb{R}^{n_1 \times n_2}$, its i -th row is denoted by M_i . For a vector $v \in \mathbb{R}^n$, define its variation as $\text{var}(v) = \max_i v_i - \min_i v_i$. Let \mathfrak{S}_n denote the set of all permutations $\pi : [n] \rightarrow [n]$. Let id denote the identity permutation, where the dimension can be inferred from context.

4.1 Background and problem setup

In this section, we present the relevant background and notation on permutation-based models, and introduce the observation model of interest.

4.1.1 Matrix models

Our main focus is on designing efficient algorithms for estimating a bivariate isotonic matrix with unknown permutations acting on its rows and columns. Formally, we define \mathbb{C}_{BISO} to be the class of matrices in $[0, 1]^{n_1 \times n_2}$ with nondecreasing rows and nondecreasing columns. For readability, we assume throughout that $n_1 \geq n_2$ unless otherwise stated; our results can be straightforwardly extended to the other case. Given a matrix $M \in \mathbb{R}^{n_1 \times n_2}$ and permutations $\pi \in \mathfrak{S}_{n_1}$ and $\sigma \in \mathfrak{S}_{n_2}$, we define the matrix $M(\pi, \sigma) \in \mathbb{R}^{n_1 \times n_2}$ by specifying its entries as

$$[M(\pi, \sigma)]_{i,j} = M_{\pi(i), \sigma(j)} \text{ for } i \in [n_1], j \in [n_2].$$

Also define the class $\mathbb{C}_{\text{BISO}}(\pi, \sigma) := \{M(\pi, \sigma) : M \in \mathbb{C}_{\text{BISO}}\}$ as the set of matrices that are bivariate isotonic when viewed along the row permutation π and column permutation σ , respectively.

The class of matrices that we are interested in estimating is given by

$$\mathbb{C}_{\text{Perm}} := \bigcup_{\substack{\pi \in \mathfrak{S}_{n_1} \\ \sigma \in \mathfrak{S}_{n_2}}} \mathbb{C}_{\text{BISO}}(\pi, \sigma).$$

In words, the class contains bivariate isotonic matrices with both rows and columns permuted.

4.1.2 Observation model

In order to study estimation from noisy observations of a matrix M^* in the class \mathbb{C}_{Perm} , we suppose that N noisy entries are sampled independently and uniformly with replacement from all entries of M^* . This sampling model is popular in the matrix completion literature, and is a special case of the *trace regression model* [NW12, KLT11]. It has also been used in the context of permutation models by Mao et al. [MWR17] to study the noisy sorting class.

More precisely, let $E^{(i,j)}$ denote the $n_1 \times n_2$ matrix with 1 in the (i, j) -th entry and 0 elsewhere, and suppose that X_ℓ is a random matrix sampled independently and uniformly from the set $\{E^{(i,j)} : i \in [n_1], j \in [n_2]\}$. We observe $N \leq n_1 n_2$ independent pairs $\{(X_\ell, y_\ell)\}_{\ell=1}^N$ from the model

$$y_\ell = \text{tr}(X_\ell^\top M^*) + z_\ell, \quad (4.1)$$

where the observations are contaminated by independent, centered, sub-Gaussian noise z_ℓ with variance parameter ζ^2 . Of particular interest is the noise model considered in applications such as crowd-labeling and ranking from pairwise comparisons. Here our samples take the form

$$y_\ell \sim \text{Ber}(\text{tr}(X_\ell^\top M)) \quad (4.2)$$

and consequently, the sub-Gaussian parameter ζ^2 is bounded; for a discussion of other regimes of noise in a related matrix model, see Gao [Gao17].

For analytical convenience, we employ the standard trick of Poissonization, so that we can assume throughout the paper that $N' = \text{Poi}(N)$ random samples are drawn according to the trace regression model (4.1). Upper and lower bounds derived under this model carry over with loss of constant factors to the model with exactly N samples; for a detailed discussion, see Appendix 4.6.

For notational convenience, denote the probability that an entry of the matrix is observed under Poissonized sampling by $p_{\text{obs}} = 1 - \exp(-N/n_1 n_2)$. Since we assume throughout that $N \leq n_1 n_2$, it can be verified that $\frac{N}{2n_1 n_2} \leq p_{\text{obs}} \leq \frac{N}{n_1 n_2}$.

Now given $N' = \text{Poi}(N)$ observations $\{(X_\ell, y_\ell)\}_{\ell=1}^{N'}$, let us define the matrix of

observations $Y = Y(\{(X_\ell, y_\ell)\}_{\ell=1}^{N'})$, with entry (i, j) given by

$$Y_{i,j} = \frac{1}{p_{\text{obs}}} \frac{1}{\mathbb{1} \vee \sum_{\ell=1}^{N'} \mathbb{1}\{X_\ell = E^{(i,j)}\}} \sum_{\ell=1}^{N'} y_\ell \mathbb{1}\{X_\ell = E^{(i,j)}\}. \quad (4.3)$$

In words, the rescaled entry $p_{\text{obs}} Y_{i,j}$ is the average of all the noisy realizations of $M_{i,j}^*$ that we have observed, or zero if the entry goes unobserved. Note that $\mathbb{E}[Y_{i,j}] = \frac{1}{p_{\text{obs}}} M_{i,j}^* \cdot p_{\text{obs}} = M_{i,j}^*$, so that $\mathbb{E}[Y] = M^*$. Moreover, we may write the model in the linearized form $Y = M^* + W$, where W is a matrix of additive noise having independent, zero-mean, sub-Gaussian entries.

4.2 Main results

In this section, we present our main results—we begin by briefly revisiting the fundamental limits of estimation, and then introduce our algorithms in Section 4.2.2. We assume throughout this section that as per the setup, we have $n_1 \geq n_2$ and $N \in [n_1 n_2]$.

4.2.1 Statistical limits of estimation

We begin by characterizing the fundamental limits of estimation under the trace regression observation model (4.1) with $N' = \text{Poi}(N)$ observations. We define the least squares estimator over the class of matrices \mathbb{C}_{Perm} as the projection

$$\widehat{M}_{\text{LS}}(Y) := \arg \min_{M \in \mathbb{C}_{\text{Perm}}} \|Y - M\|_F^2.$$

The projection is a non-convex problem, and is unlikely to be computable exactly in polynomial time. However, studying this estimator allows us to establish a baseline that characterizes the best achievable statistical rate. The following theorem characterizes its risk up to a logarithmic factor in the dimension; recall the shorthand $Y = Y(\{(X_\ell, y_\ell)\}_{\ell=1}^{N'})$.

Theorem 4.2.1. *For any matrix $M^* \in \mathbb{C}_{\text{Perm}}$, we have*

$$\frac{1}{n_1 n_2} \|\widehat{M}_{\text{LS}}(Y) - M^*\|_F^2 \lesssim (\zeta^2 \vee 1) \frac{n_1 \log^2 n_1}{N} \quad (4.4a)$$

with probability at least $1 - (n_1 n_2)^{-3}$.

Additionally, under the Bernoulli observation model (4.2), any estimator \widehat{M} satisfies

$$\sup_{M^* \in \mathbb{C}_{\text{Perm}}} \mathbb{E} \left[\frac{1}{n_1 n_2} \|\widehat{M} - M^*\|_F^2 \right] \gtrsim \frac{n_1}{N}. \quad (4.4b)$$

The factor $(\zeta^2 \vee 1)$ appears in the upper bound instead of the noise variance ζ^2 because even if the noise is zero, there are missing entries. The theorem characterizes the minimax rate of estimation for the class \mathbb{C}_{Perm} up to a logarithmic factor.

4.2.2 Efficient algorithms

Next, we propose polynomial-time algorithms for estimating the permutations (π, σ) and the matrix M^* . Our main algorithm relies on two distinct steps: first, we estimate the unknown permutations; we then project onto the class of matrices that are bivariate isotonic when viewed along the estimated permutations. The formal meta-algorithm is described below.

Algorithm 1 (meta-algorithm)

- Step 0: Split the observations into two disjoint parts, each containing $N'/2$ observations, and construct the matrices $Y^{(1)} = Y\left(\{X_\ell, y_\ell\}_{\ell=1}^{N'/2}\right)$ and $Y^{(2)} = Y\left(\{X_\ell, y_\ell\}_{\ell=N'/2+1}^{N'}\right)$.
- Step 1: Use $Y^{(1)}$ to obtain the permutation estimates $(\hat{\pi}, \hat{\sigma})$.
- Step 2: Return the matrix estimate $\widehat{M}(\hat{\pi}, \hat{\sigma}) := \arg \min_{M \in \mathbb{C}_{\text{BISO}}(\hat{\pi}, \hat{\sigma})} \|Y^{(2)} - M\|_F^2$.

Owing to the convexity of the set $\mathbb{C}_{\text{BISO}}(\hat{\pi}, \hat{\sigma})$, the projection operation in Step 2 of the algorithm can be computed in near linear time [BDPR84, KRS15]. The following result, a slight variant of Proposition 4.2 of Chatterjee and Mukherjee [CM16], allows us to characterize the error rate of any such meta-algorithm as a function of the permutation estimates $(\hat{\pi}, \hat{\sigma})$.

Proposition 4.2.2. *Suppose that $M^* \in \mathbb{C}_{\text{BISO}}(\pi, \sigma)$ where π and σ are unknown permutations in \mathfrak{S}_{n_1} and \mathfrak{S}_{n_2} respectively. Then with probability at least $1 - (n_1 n_2)^{-3}$, we have*

$$\begin{aligned} \frac{1}{n_1 n_2} \|\widehat{M}(\hat{\pi}, \hat{\sigma}) - M^*\|_F^2 &\lesssim (\zeta^2 \vee 1) \frac{n_1 \log^2 n_1}{N} + \frac{1}{n_1 n_2} \|M^*(\pi^{-1} \circ \hat{\pi}, \text{id}) - M^*\|_F^2 \\ &\quad + \frac{1}{n_1 n_2} \|M^*(\text{id}, \sigma^{-1} \circ \hat{\sigma}) - M^*\|_F^2. \end{aligned} \tag{4.5}$$

The first term on the right hand side of the bound (4.5) corresponds to an estimation error, if the true permutations π and σ were known a priori, and the latter two terms correspond to an approximation error that we incur as a result of having to estimate these permutations from data. Comparing the bound (4.5) to the minimax lower bound (4.4b), we see that up to a logarithmic factor, the first term of the bound (4.5) is unavoidable, and so we can restrict our attention to obtaining good permutation estimates $(\hat{\pi}, \hat{\sigma})$. We now present our main permutation estimation procedure that can be plugged into Step 1 of this meta-algorithm.

Two-dimensional sorting

To reorder the rows or columns of a matrix with monotonicity constraints, sorting row or column sums is perhaps the most natural approach popularly adopted in the literature [CM16, FMR16]. However, such a procedure does not take advantage of the fact that the underlying matrix is monotonic in *both* dimensions. To improve upon simply sorting row sums, we propose an algorithm that first sorts the columns of the matrix approximately, and then exploits this approximate ordering to sort the rows of the matrix.

We need more notation to facilitate the description of the algorithm. For a partition

$\mathbf{bl} = (\mathbf{bl}_1, \dots, \mathbf{bl}_K)$ of the set $[n_2]^1$, we group the columns of a matrix $Y \in \mathbb{R}^{n_1 \times n_2}$ into K blocks according to their indices in \mathbf{bl} , and refer to \mathbf{bl} as a partition or *blocking* of the columns of Y .

Given a data matrix $Y \in \mathbb{R}^{n_1 \times n_2}$, the following blocking subroutine returns a column partition $\mathbf{BL}(Y)$. In the main algorithm, partial row sums are computed on indices contained in each block.

Subroutine 1 (blocking)

- Step 1: Compute the column sums $\{C(j)\}_{j=1}^{n_2}$ of the matrix Y as

$$C(j) = \sum_{i=1}^{n_1} Y_{i,j}.$$

Let $\hat{\sigma}_{\text{pre}}$ be the permutation along which the sequence $\{C(\hat{\sigma}_{\text{pre}}(j))\}_{j=1}^{n_2}$ is nondecreasing.

- Step 2: Set $\tau = 16(\zeta + 1) \left(\sqrt{\frac{n_1^2 n_2}{N} \log(n_1 n_2)} + \frac{n_1 n_2}{N} \log(n_1 n_2) \right)$ and $K = \lceil n_2 / \tau \rceil$. Partition the columns of Y into K blocks by defining

$$\begin{aligned} \mathbf{bl}_1 &= \{j \in [n_2] : C(j) \in (-\infty, \tau)\}, \\ \mathbf{bl}_k &= \{j \in [n_2] : C(j) \in [(k-1)\tau, k\tau)\} \text{ for } 1 < k < K, \text{ and} \\ \mathbf{bl}_K &= \{j \in [n_2] : C(j) \in [(K-1)\tau, \infty)\}. \end{aligned}$$

Note that each block is contiguous when the columns are permuted by $\hat{\sigma}_{\text{pre}}$.

- Step 3 (aggregation): Set $\beta = n_2 \sqrt{\frac{n_1}{N} \log(n_1 n_2)}$. Call a block \mathbf{bl}_k “large” if $|\mathbf{bl}_k| \geq \beta$ and “small” otherwise. Aggregate small blocks in \mathbf{bl} while leaving the large blocks as they are, to obtain the final partition \mathbf{BL} .

More precisely, consider the matrix $Y' = Y(\text{id}, \hat{\sigma}_{\text{pre}})$ having nondecreasing column sums and contiguous blocks. Call two small blocks “adjacent” if there is no other small block between them. Take unions of adjacent small blocks to

¹ \mathbf{bl} is a partition of $[n_2]$ if $[n_2] = \cup_{k=1}^K \mathbf{bl}_k$ and $\mathbf{bl}_j \cap \mathbf{bl}_k = \emptyset$ for $j \neq k$

ensure that the size of each resulting block is in the range $[\frac{1}{2}\beta, 2\beta]$. If the union of all small blocks is smaller than $\frac{1}{2}\beta$, aggregate them all.

Return the resulting partition $\text{BL}(Y) = \text{BL}$.

The threshold τ is chosen to be a high probability bound on the perturbation of any column sum, so we are confident that columns in a block bl_j are in fact close to those in bl_j when the columns are sorted increasingly. It turns out that comparing partial row sums on these blocks aids us in reordering the rows of the matrix. Moreover, Step 3 aggregates small blocks into large enough ones to reduce noise in these partial row sums. We are now in a position to describe the two-dimensional sorting algorithm.

Algorithm 2 (two-dimensional sorting)

- Step 0: Split the observations into two independent subsamples of equal size, and form the corresponding matrices $Y^{(1)}$ and $Y^{(2)}$ according to equation (4.3).
- Step 1: Apply Subroutine 1 to the matrix $Y^{(1)}$ to obtain a partition $\text{BL} = \text{BL}(Y^{(1)})$ of the columns. Let K be the number of blocks in BL .
- Step 2: Using the second sample $Y^{(2)}$, compute the row sums

$$S(i) = \sum_{j \in [n_2]} Y_{i,j}^{(2)} \text{ for each } i \in [n_1],$$

and the partial row sums within each block

$$S_{\text{BL}_k}(i) = \sum_{j \in \text{BL}_k} Y_{i,j}^{(2)} \text{ for each } i \in [n_1], k \in [K].$$

Create a directed graph G with vertex set $[n_1]$, where an edge $u \rightarrow v$ is present if either

$$S(v) - S(u) > 16(\zeta + 1) \left(\sqrt{\frac{n_1 n_2^2}{N} \log(n_1 n_2)} + \frac{n_1 n_2}{N} \log(n_1 n_2) \right), \text{ or} \tag{4.6a}$$

$$S_{\text{BL}_k}(v) - S_{\text{BL}_k}(u) > 16(\zeta + 1) \left(\sqrt{\frac{n_1 n_2}{N} |\text{BL}_k| \log(n_1 n_2)} + \frac{n_1 n_2}{N} \log(n_1 n_2) \right) \text{ for some } k \in [K]. \tag{4.6b}$$

- Step 3: Compute a topological sort $\hat{\pi}_{\text{tds}}$ of the graph G ; if none exists, set $\hat{\pi}_{\text{tds}} = \text{id}$.
- Step 4: Repeat Steps 1–3 with $(Y^{(i)})^\top$ replacing $Y^{(i)}$ for $i = 1, 2$, the roles of n_1 and n_2 switched, and the roles of π and σ switched, to compute the permutation estimate $\hat{\sigma}_{\text{tds}}$.

- Step 5: Return the permutation estimates $(\widehat{\pi}_{\text{tds}}, \widehat{\sigma}_{\text{tds}})$.

Recall that a permutation π is called a topological sort of G if $\pi(u) < \pi(v)$ for every directed edge $u \rightarrow v$. The construction of the graph G in Step 2 dominates the computational complexity, and takes time $\mathcal{O}(n_1^2 n_2 / \beta) = \mathcal{O}(n_1^2 n_2^{1/2})$. We have the following guarantee for the two-dimensional sorting algorithm.

Theorem 4.2.3. *For any matrix $M^* \in \mathbb{C}_{\text{Perm}}$, we have*

$$\frac{1}{n_1 n_2} \left\| \widehat{M}(\widehat{\pi}_{\text{tds}}, \widehat{\sigma}_{\text{tds}}) - M^* \right\|_F^2 \lesssim (\zeta^2 \vee 1) \left[\left(\frac{n_1 \log n_1}{N} \right)^{3/4} + \frac{n_1 \log^2 n_1}{N} \right]$$

with probability at least $1 - 9(n_1 n_2)^{-3}$.

In particular, setting $N = n_1 n_2$, we have proved that our efficient estimator enjoys the rate

$$\frac{1}{n_1 n_2} \left\| \widehat{M}(\widehat{\pi}_{\text{tds}}, \widehat{\sigma}_{\text{tds}}) - M^* \right\|_F^2 = \widetilde{O} \left(n_2^{-3/4} \right),$$

which is the main theoretical guarantee established in this paper for permutation-based models.

4.3 Applications

We now discuss in detail how the matrix models studied in this paper arise in practice. The class \mathbb{C}_{Perm} was studied as a permutation-based model for crowd-labeling [SBW16b] in the case of binary questions, and was proposed as a strict generalization of the classical Dawid-Skene model [DS79, KOS11b, LPI12, DDKR13, GKM11]. Here there is a set of n_2 questions of a binary nature; the true answer to these questions can be represented by a vector $x^* \in \{0, 1\}^{n_2}$, and our goal is to estimate this vector by asking these questions to n_1 workers on a crowdsourcing platform. A key to this problem is being able to model the probabilities with which workers answer questions correctly, and we do so by collecting these probabilities within a matrix $M^* \in [0, 1]^{n_1 \times n_2}$. Assuming that workers have a strict ordering π of their abilities, and that questions have a strict ordering σ of their difficulties, the matrix M^* is bivariate isotonic when the rows are ordered in increasing order of worker ability, and columns are ordered in decreasing order of question difficulty. However, since worker abilities and question difficulties are unknown a priori, the matrix of probabilities obeys the inclusion $M^* \in \mathbb{C}_{\text{Perm}}$.

In the *calibration* problem, we would like to ask questions whose answers we know a priori, so that we can estimate worker abilities and question difficulties, or more generally, the entries of the matrix M^* . This corresponds to estimating matrices in the class \mathbb{C}_{Perm} from noisy observations of their entries, whose rate of estimation is our main result.

A subclass of \mathbb{C}_{Perm} specializes to the case $n_1 = n_2 = n$, and also imposes an additional skew symmetry constraint. More precisely, define $\mathbb{C}'_{\text{BISO}}$ analogously to the

class \mathbb{C}_{BISO} , except with matrices having columns that are nonincreasing instead of nondecreasing. Also define the class $\mathbb{C}_{\text{skew}}(n) := \{M \in [0, 1]^{n_1 \times n_2} : M + M^\top = \mathbf{1}\mathbf{1}^\top\}$, and the *strong stochastic transitivity* class

$$\mathbb{C}_{\text{SST}}(n) := \left(\bigcup_{\pi \in \mathfrak{S}_n} \mathbb{C}'_{\text{BISO}}(\pi, \pi) \right) \cap \mathbb{C}_{\text{skew}}(n).$$

The class $\mathbb{C}_{\text{SST}}(n)$ is useful as a model for estimation from pairwise comparisons [Cha15, SBGW17], and was proposed as a strict generalization of parametric models for this problem [BT52, NOS16, RA14]. In particular, given n items obeying some unknown underlying ranking π , entry (i, j) of a matrix $M^* \in \mathbb{C}_{\text{SST}}(n)$ represents the probability $\Pr(i \succ j)$ with which item i beats item j in a pairwise comparison between them. The shape constraint encodes the transitivity condition that for all triples (i, j, k) obeying $\pi(i) < \pi(j) < \pi(k)$, we must have

$$\Pr(i \succ k) \geq \max\{\Pr(i \succ j), \Pr(j \succ k)\}.$$

For a more classical introduction to these models, see the papers [Fis73, ML65, BW97] and the references therein. Our task is to estimate the underlying ranking from results of passively chosen pairwise comparisons² between the n items, or more generally, to estimate the underlying probabilities M^* that govern these comparisons³. All the results we obtain in this work clearly extend to the class $\mathbb{C}_{\text{SST}}(n)$ with minimal modifications; for example, either of the two estimates $\hat{\pi}_{\text{tds}}$ or $\hat{\sigma}_{\text{tds}}$ may be returned as an estimate of the permutation π . Consequently, the informal theorem stated in the introduction is an immediate corollary of Theorem 4.2.3 once these modifications are made to the algorithm.

4.4 Proofs

Throughout the proofs, we assume without loss of generality that $M^* \in \mathbb{C}_{\text{BISO}}(\text{id}, \text{id}) = \mathbb{C}_{\text{BISO}}$. Because we are interested in rates of estimation up to universal constants, we assume that each independent subsample contains $N' = \text{Poi}(N)$ observations (instead of $\text{Poi}(N)/2$ or $\text{Poi}(N)/4$). We use the shorthand $Y = Y(\{(X_\ell, y_\ell)\}_{\ell=1}^{N'})$, throughout.

4.4.1 Some preliminary lemmas

Before turning to the proof of Theorems 4.2.1 and 4.2.3, we provide three lemmas that underlie many of our arguments. The first lemma can be readily distilled from the proof of Theorem 5 of Shah et al. [SBGW17] with slight modifications. It is worth

²Such a passive, simultaneous setting should be contrasted with the *active* case (e.g., [HSRW16, FOPS17, AAK17]), where we may sequentially choose pairs of items to compare depending on the results of previous comparisons.

³Accurate, proper estimates of M^* translate to accurate estimates of the ranking π (see Shah et al. [SBGW17]).

mentioning that similar lemmas characterizing the estimation error of a bivariate isotonic matrix were also proved by [CGS18, CM16].

Lemma 4.4.1 ([SBGW17]). *Let $n_1 \geq n_2$, and let $M^* \in \mathbb{C}_{\text{Perm}}$. Assume that our observation model takes the form $Y = M^* + W$, where the noise matrix W satisfies the properties*

- (a) *the entries $W_{i,j}$ are independent, centered, $\frac{c_1}{p_{\text{obs}}}(\zeta \vee 1)$ -sub-Gaussian random variables;*
- (b) *the second moments are bounded as $\mathbb{E}[|W_{i,j}|^2] \leq \frac{c_2}{p_{\text{obs}}}(\zeta^2 \vee 1)$ for all $i \in [n_1], j \in [n_2]$.*

Then the least squares estimator $\widehat{M}_{\text{LS}}(Y)$ satisfies

$$\Pr \left\{ \left\| \widehat{M}_{\text{LS}}(Y) - M^* \right\|_F^2 \geq \frac{c_3}{p_{\text{obs}}}(\zeta^2 \vee 1)n_1 \log^2 n_1 \right\} \leq (n_1 n_2)^{-3}.$$

Moreover, the same result holds if the class \mathbb{C}_{Perm} is replaced by the class \mathbb{C}_{BISO} .

The proof follows that of Shah et al. [SBGW17, Theorem 5] very closely, and is postponed to Section 4.4.5. The next lemma establishes concentration of sums of our observations around their means.

Lemma 4.4.2. *For any nonempty subset $\mathcal{S} \subset [n_1] \times [n_2]$, it holds that*

$$\Pr \left\{ \left| \sum_{(i,j) \in \mathcal{S}} (Y_{i,j} - M_{i,j}^*) \right| \geq 8(\zeta + 1) \left(\sqrt{\frac{|\mathcal{S}|n_1 n_2}{N} \log(n_1 n_2)} + 2 \frac{n_1 n_2}{N} \log(n_1 n_2) \right) \right\} \leq 2(n_1 n_2)^{-4}.$$

Proof. According to definitions (4.1) and (4.3), we have

$$W_{i,j} = Y_{i,j} - M_{i,j}^* = \begin{cases} -M_{i,j}^* & \text{if entry } (i,j) \text{ is not observed, and} \\ M_{i,j}^*/p_{\text{obs}} - M_{i,j}^* + \frac{W'_{i,j}}{p_{\text{obs}}}, & \text{otherwise,} \end{cases}$$

where W' is a ζ -sub-Gaussian noise matrix with independent entries. Consequently, we can express the noise on each entry as $W_{i,j} = Z_{i,j}^{(1)} + Z_{i,j}^{(2)}$ where $\{Z_{i,j}^{(1)}\}_{i \in [n_1], j \in [n_2]}$ are independent, zero-mean random variables given by

$$Z_{i,j}^{(1)} = \begin{cases} M_{i,j}^*(p_{\text{obs}}^{-1} - 1) & \text{with probability } p_{\text{obs}}, \\ -M_{i,j}^* & \text{with probability } 1 - p_{\text{obs}}, \end{cases}$$

and $\{Z_{i,j}^{(2)}\}_{i \in [n_1], j \in [n_2]}$ are independent, zero-mean random variables such that

$$Z_{i,j}^{(2)} \text{ is } \begin{cases} \frac{\zeta}{p_{\text{obs}}}\text{-sub-Gaussian} & \text{with probability } p_{\text{obs}}, \\ 0 & \text{with probability } 1 - p_{\text{obs}}. \end{cases}$$

We control the two separately. First, we have $|Z_{i,j}^{(1)}| \leq 1/p_{\text{obs}}$ and the variance of each $Z_{i,j}^{(1)}$ is bounded by $(1 - p_{\text{obs}})^2/p_{\text{obs}} + (1 - p_{\text{obs}}) \leq 1/p_{\text{obs}}$. Hence Bernstein's inequality for bounded noise yields

$$\Pr \left\{ \left| \sum_{(i,j) \in \mathcal{S}} Z_{i,j}^{(1)} \right| \geq t \right\} \leq 2 \exp \left(- \frac{t^2/2}{|\mathcal{S}|/p_{\text{obs}} + t/(3p_{\text{obs}})} \right).$$

Taking $t = 4\sqrt{\frac{|\mathcal{S}|n_1n_2}{N} \log(n_1n_2)} + 6\frac{n_1n_2}{N} \log(n_1n_2)$ and recalling that $p_{\text{obs}} \geq \frac{N}{2n_1n_2}$, we obtain

$$\Pr \left\{ \left| \sum_{(i,j) \in \mathcal{S}} Z_{i,j}^{(1)} \right| \geq 4\sqrt{\frac{|\mathcal{S}|n_1n_2}{N} \log(n_1n_2)} + 6\frac{n_1n_2}{N} \log(n_1n_2) \right\} \leq (n_1n_2)^{-4}.$$

In order to control the deviation of the sum of $Z_{i,j}^{(2)}$, we note that the q -th moment of $Z_{i,j}^{(2)}$ is bounded by $\frac{N}{n_1n_2} \left(\frac{2\zeta}{p_{\text{obs}}} \sqrt{q} \right)^q \leq \frac{q!}{2} \frac{8\zeta^2 n_1n_2}{N} \left(\frac{4\zeta n_1n_2}{N} \right)^{q-2}$. Then another version of Bernstein's inequality [BLM13] yields

$$\Pr \left\{ \left| \sum_{(i,j) \in \mathcal{S}} Z_{i,j}^{(2)} \right| \geq \sqrt{\frac{16\zeta^2 |\mathcal{S}|n_1n_2}{N} t} + \frac{4\zeta n_1n_2}{N} t \right\} \leq 2 \exp(-t),$$

and setting $t = 4 \log(n_1n_2)$ gives

$$\Pr \left\{ \left| \sum_{(i,j) \in \mathcal{S}} Z_{i,j}^{(2)} \right| \geq 8\zeta \sqrt{\frac{|\mathcal{S}|n_1n_2}{N} \log(n_1n_2)} + 16\zeta \frac{n_1n_2}{N} \log(n_1n_2) \right\} \leq (n_1n_2)^{-4}.$$

Combining the above two deviation bounds completes the proof. \square

The last lemma is a deterministic result.

Lemma 4.4.3. *Let $\{a_i\}_{i=1}^n$ be a nondecreasing sequence of real numbers. If π is a permutation in \mathfrak{S}_n such that $\pi(i) < \pi(j)$ whenever $a_j - a_i > \tau$ where $\tau > 0$, then $|a_{\pi(i)} - a_i| \leq \tau$ for all $i \in [n]$.*

Proof. Suppose that $a_j - a_{\pi(j)} > \tau$ for some index $j \in [n]$. Since π is a bijection, there must exist an index $i \leq \pi(j)$ such that $\pi(i) > \pi(j)$. However, we then have $a_j - a_i \geq a_j - a_{\pi(j)} > \tau$, which contradicts the assumption. A similar argument shows that $a_{\pi(j)} - a_j > \tau$ also leads to a contradiction. Therefore, we obtain that $|a_{\pi(j)} - a_j| \leq \tau$ for every $j \in [n]$. \square

With these lemmas in hand, we are now ready to prove our main theorems.

4.4.2 Proof of Theorem 4.2.1

We split the proof into two parts by proving the upper and lower bounds separately.

Proof of upper bound

The upper bound follows from Lemma 4.4.1 once we check the conditions on the noise for our model. We have seen in the proof of Lemma 4.4.2 that the noise on each entry can be written as $W_{i,j} = Z_{i,j}^{(1)} + Z_{i,j}^{(2)}$. Again, $Z_{i,j}^{(1)}$ and $Z_{i,j}^{(2)}$ are $\frac{c}{p_{\text{obs}}}$ -sub-Gaussian and $\frac{c\zeta}{p_{\text{obs}}}$ -sub-Gaussian respectively, and have variances bounded by $\frac{1}{p_{\text{obs}}}$ and $\frac{c\zeta^2}{p_{\text{obs}}}$ respectively. Hence the conditions on W in Lemma 4.4.1 are satisfied. Then we can apply the lemma, recall the relation $p_{\text{obs}} \geq \frac{N}{2n_1n_2}$ and normalize the bound by $\frac{1}{n_1n_2}$ to complete the proof.

Proof of lower bound

The lower bound follows from an application of Fano's lemma. The technique is standard, and we briefly review it here. Suppose we wish to estimate a parameter θ over an indexed class of distributions $\mathcal{P} = \{\mathbb{P}_\theta \mid \theta \in \Theta\}$ in the square of a (pseudo-)metric ρ . We refer to a subset of parameters $\{\theta^1, \theta^2, \dots, \theta^K\}$ as a local (δ, ϵ) -packing set if

$$\min_{i,j \in [K], i \neq j} \rho(\theta^i, \theta^j) \geq \delta \quad \text{and} \quad \frac{1}{K(K-1)} \sum_{i,j \in [K], i \neq j} D(\mathbb{P}_{\theta^i} \parallel \mathbb{P}_{\theta^j}) \leq \epsilon.$$

Note that this set is a δ -packing in the metric ρ with the average KL-divergence bounded by ϵ . The following result is a straightforward consequence of Fano's inequality:

Lemma 4.4.4 (Local packing Fano lower bound). *For any (δ, ϵ) -packing set of cardinality K , we have*

$$\inf_{\hat{\theta}} \sup_{\theta^* \in \Theta} \mathbb{E} \left[\rho(\hat{\theta}, \theta^*)^2 \right] \geq \frac{\delta^2}{2} \left(1 - \frac{\epsilon + \log 2}{\log K} \right). \quad (4.7)$$

In addition, the Gilbert-Varshamov bound [Gil52, Var57] guarantees the existence of binary vectors $\{v^1, v^2, \dots, v^K\} \subseteq \{0, 1\}^{n_1}$ such that

$$K \geq 2^{c_1 n_1} \quad \text{and} \quad (4.8a)$$

$$\|v^i - v^j\|_2^2 \geq c_2 n_1 \quad \text{for each } i \neq j, \quad (4.8b)$$

for some fixed tuple of constants (c_1, c_2) . We use this guarantee to design a packing of matrices in the class \mathbb{C}_{Perm} . For each $i \in [K]$, fix some $\delta \in [0, 1/4]$ to be precisely set later, and define the matrix M^i having identical columns, with entries given by

$$M_{j,k}^i = \begin{cases} 1/2, & \text{if } v_j^i = 0 \\ 1/2 + \delta, & \text{otherwise.} \end{cases} \quad (4.9)$$

Clearly, each of these matrices $\{M^i\}_{i=1}^K$ is a member of the class \mathbb{C}_{Perm} , and each distinct pair of matrices (M^i, M^j) satisfies the inequality $\|M^i - M^j\|_F^2 \geq c_2 n_1 n_2 \delta^2$.

Let \mathbb{P}_M denote the probability distribution of the observations in the model (4.1) with underlying matrix $M \in \mathbb{C}_{\text{Perm}}$. Our observations are independent across entries of the matrix, and so the KL divergence tensorizes to yield

$$D(\mathbb{P}_{M^i} \|\mathbb{P}_{M^j}) = \sum_{\substack{k \in [n_1] \\ \ell \in [n_2]}} D(\mathbb{P}_{M_{k,\ell}^i} \|\mathbb{P}_{M_{k,\ell}^j}).$$

Let us now examine one term of this sum. We observe $T_{k,\ell} = \text{Poi}(\frac{N}{n_1 n_2})$ samples of entry (k, ℓ) ; conditioned on the event $T_{k,\ell} = m$, we have the distributions

$$\mathbb{P}_{M_{k,\ell}^i} = \text{Bin}(m, M_{k,\ell}^i), \quad \text{and} \quad \mathbb{P}_{M_{k,\ell}^j} = \text{Bin}(m, M_{k,\ell}^j).$$

Consequently, the KL divergence conditioned on $T_{k,\ell} = m$ is given by

$$D(\mathbb{P}_{M_{k,\ell}^i} \|\mathbb{P}_{M_{k,\ell}^j}) = mD(M_{k,\ell}^i \|\ M_{k,\ell}^j),$$

where we have used $D(p\|q) = p \log(\frac{p}{q}) + (1-p) \log(\frac{1-p}{1-q})$ to denote the KL divergence between the Bernoulli random variables $\text{Ber}(p)$ and $\text{Ber}(q)$.

Note that for $p, q \in [1/2, 3/4]$, we have

$$\begin{aligned} D(p\|q) &= p \log\left(\frac{p}{q}\right) + (1-p) \log\left(\frac{1-p}{1-q}\right) \\ &\stackrel{(i)}{\leq} p \left(\frac{p-q}{q}\right) + (1-p) \left(\frac{q-p}{1-q}\right) \\ &= \frac{(p-q)^2}{q(1-q)} \\ &\stackrel{(ii)}{\leq} \frac{16}{3}(p-q)^2. \end{aligned}$$

Here, step (i) follows from the inequality $\log x \leq x - 1$, and step (ii) from the assumption $q \in [\frac{1}{2}, \frac{3}{4}]$. Taking the expectation with respect to $T_{k,\ell}$, we have

$$D(\mathbb{P}_{M_{k,\ell}^i} \|\mathbb{P}_{M_{k,\ell}^j}) \leq \frac{16}{3} \frac{N}{n_1 n_2} (M_{k,\ell}^i - M_{k,\ell}^j)^2 \leq \frac{16}{3} \frac{N}{n_1 n_2} \delta^2,$$

Summing over $k \in [n_1], \ell \in [n_2]$ yields $D(\mathbb{P}_{M^i} \|\mathbb{P}_{M^j}) \leq \frac{16}{3} N \delta^2$.

Substituting into the Fano's inequality (4.7), we have

$$\inf_{\widehat{M}} \sup_{M^* \in \mathbb{C}_{\text{Perm}}} \mathbb{E} \left[\|\widehat{M} - M^*\|_F^2 \right] \geq \frac{c_2 n_1 n_2 \delta^2}{2} \left(1 - \frac{\frac{16}{3} N \delta^2 + \log 2}{c_3 n_1} \right).$$

Finally, choosing $\delta^2 = c \frac{n_1}{N}$ and normalizing by $n_1 n_2$ yields the claim.

4.4.3 Proof of Proposition 4.2.2

Recall the definition of $\widehat{M}(\widehat{\pi}, \widehat{\sigma})$ in the meta-algorithm, and additionally, define the projection of any matrix $M \in \mathbb{R}^{n_1 \times n_2}$, as

$$\mathcal{P}_{\pi, \sigma}(M) = \arg \min_{\widetilde{M} \in \mathbb{C}_{\text{BISO}}(\pi, \sigma)} \|M - \widetilde{M}\|_F^2.$$

and letting $W = Y^{(2)} - M^*$, we have

$$\begin{aligned} \|\widehat{M}(\widehat{\pi}, \widehat{\sigma}) - M^*\|_F^2 &\stackrel{(i)}{\leq} 2\|\mathcal{P}_{\widehat{\pi}, \widehat{\sigma}}(M^* + W) - \mathcal{P}_{\widehat{\pi}, \widehat{\sigma}}(M^*(\widehat{\pi}, \widehat{\sigma}) + W)\|_F^2 \\ &\quad + 2\|\mathcal{P}_{\widehat{\pi}, \widehat{\sigma}}(M^*(\widehat{\pi}, \widehat{\sigma}) + W) - M^*\|_F^2 \\ &\stackrel{(ii)}{\leq} 2\|M^*(\widehat{\pi}, \widehat{\sigma}) - M^*\|_F^2 + 2\|\mathcal{P}_{\widehat{\pi}, \widehat{\sigma}}(M^*(\widehat{\pi}, \widehat{\sigma}) + W) - M^*\|_F^2 \\ &\stackrel{(iii)}{\leq} 4\|\mathcal{P}_{\widehat{\pi}, \widehat{\sigma}}(M^*(\widehat{\pi}, \widehat{\sigma}) + W) - M^*(\widehat{\pi}, \widehat{\sigma})\|_F^2 + 6\|M^*(\widehat{\pi}, \widehat{\sigma}) - M^*\|_F^2, \end{aligned} \tag{4.10}$$

where step (ii) follows from the non-expansiveness of a projection onto a convex set, and steps (i) and (iii) from the triangle inequality.

The first term in (4.10) is the estimation error of a bivariate isotonic matrix with known permutations. Since the sample used to obtain $(\widehat{\pi}, \widehat{\sigma})$ is independent from the sample used in the projection step, it is equivalent to control the error $\|\mathcal{P}_{\text{id}, \text{id}}(M^* + W) - M^*\|_F^2$. As before, the noise matrix W satisfies the conditions of Lemma 4.4.1. Therefore, applying Lemma 4.4.1 in the case $M^* \in \mathbb{C}_{\text{BISO}}$ with $p_{\text{obs}} \geq \frac{N}{2n_1n_2}$ yields the desired bound of order $(\zeta^2 \vee 1) \frac{n_1 \log^2 n_1}{N}$.

It remains to bound the second term of (4.10), the approximation error of the permutation estimates. Note that the approximation error can be split into two components: one along the rows of the matrix, and the other along the columns. More explicitly, we have

$$\begin{aligned} \|M^* - M^*(\widehat{\pi}, \widehat{\sigma})\|_F^2 &\leq 2\|M^* - M^*(\widehat{\pi}, \text{id})\|_F^2 + 2\|M^*(\widehat{\pi}, \text{id}) - M^*(\widehat{\pi}, \widehat{\sigma})\|_F^2 \\ &= 2\|M^* - M^*(\widehat{\pi}, \text{id})\|_F^2 + 2\|M^* - M^*(\text{id}, \widehat{\sigma})\|_F^2. \end{aligned}$$

Recall that we assumed without loss of generality that the true permutations are identity permutations, so this completes the proof of Proposition 4.2.2. The proof readily extends to the general case by precomposing $\widehat{\pi}$ and $\widehat{\sigma}$ with π^{-1} and σ^{-1} respectively.

4.4.4 Proof of Theorem 4.2.3

Recall that according to Proposition 4.2.2, it suffices to bound the approximation error of our permutation estimate $\|M^* - M^*(\widehat{\pi}_{\text{tds}}, \text{id})\|_F^2$. To ease the notation, we use

the shorthand

$$\eta := 16(\zeta + 1) \left(\sqrt{\frac{n_1 n_2^2}{N} \log(n_1 n_2)} + 2 \frac{n_1 n_2}{N} \log(n_1 n_2) \right),$$

and for each block BL_k in Algorithm 2 where $k \in [K]$, we use the shorthand

$$\eta_k := 16(\zeta + 1) \left(\sqrt{\frac{|\text{BL}_k| n_1 n_2}{N} \log(n_1 n_2)} + 2 \frac{n_1 n_2}{N} \log(n_1 n_2) \right)$$

throughout the proof. Applying Lemma 4.4.2 with $\mathcal{S} = \{i\} \times [n_2]$ and then with $\mathcal{S} = \{i\} \times \text{BL}_k$ for each $i \in [n_1], k \in [K]$, we obtain that

$$\Pr \left\{ \left| S(i) - \sum_{\ell \in [n_2]} M_{i,\ell}^* \right| \geq \frac{\eta}{2} \right\} \leq 2(n_1 n_2)^{-4}, \quad (4.11a)$$

and that

$$\Pr \left\{ \left| S_{\text{BL}_k}(i) - \sum_{\ell \in \text{BL}_k} M_{i,\ell}^* \right| \geq \frac{\eta_k}{2} \right\} \leq 2(n_1 n_2)^{-4}. \quad (4.11b)$$

Note that $K \leq n_2/\beta \leq n_2^{1/2}$, so a union bound over all $n_1(K+1)$ events in inequalities (4.11a) and (4.11b) yields that $\Pr\{\mathcal{E}\} \geq 1 - 2(n_1 n_2)^{-3}$, where we define the event

$$\mathcal{E} := \left\{ \left| S(i) - \sum_{\ell \in [n_2]} M_{i,\ell}^* \right| \leq \frac{\eta}{2} \text{ and } \left| S_{\text{BL}_k}(i) - \sum_{\ell \in \text{BL}_k} M_{i,\ell}^* \right| \leq \frac{\eta_k}{2} \forall i \in [n_1], k \in [K] \right\}.$$

We now condition on event \mathcal{E} . Applying the triangle inequality yields that if

$$S(v) - S(u) > \eta \quad \text{or} \quad S_{\text{BL}_k}(v) - S_{\text{BL}_k}(u) > \eta_k,$$

then we have

$$\sum_{\ell \in [n_2]} M_{v,\ell}^* - \sum_{\ell \in [n_2]} M_{u,\ell}^* > 0 \quad \text{or} \quad \sum_{\ell \in \text{BL}_k} M_{v,\ell}^* - \sum_{\ell \in \text{BL}_k} M_{u,\ell}^* > 0.$$

It follows that $u < v$ since M^* has nondecreasing columns. Thus, by the choice of thresholds η and η_k in inequalities (4.6a) and (4.6b), we have guaranteed that every edge $u \rightarrow v$ in the graph G is consistent with the underlying permutation id , so a topological sort exists on event \mathcal{E} .

Conversely, if we have

$$\sum_{\ell \in [n_2]} M_{v,\ell}^* - \sum_{\ell \in [n_2]} M_{u,\ell}^* > 2\eta \quad \text{or} \quad \sum_{\ell \in \text{BL}_k} M_{v,\ell}^* - \sum_{\ell \in \text{BL}_k} M_{u,\ell}^* > 2\eta_k,$$

then the triangle inequality implies that

$$S(v) - S(u) > \eta \quad \text{or} \quad S_{\text{BL}_k}(v) - S_{\text{BL}_k}(u) > \eta_k.$$

Hence the edge $u \rightarrow v$ is present in the graph G , so the topological sort $\widehat{\pi}_{\text{tds}}(u)$ satisfies the relation $\widehat{\pi}_{\text{tds}}(u) < \widehat{\pi}_{\text{tds}}(v)$. Claim that this allows us to obtain the following bounds on event \mathcal{E} :

$$\left| \sum_{j \in [n_2]} (M_{\widehat{\pi}_{\text{tds}}(i),j}^* - M_{i,j}^*) \right| \leq 96(\zeta + 1) \sqrt{\frac{n_1 n_2^2}{N} \log(n_1 n_2)} \quad \text{for all } i \in [n_1], \text{ and} \quad (4.12a)$$

$$\left| \sum_{j \in \text{BL}_k} (M_{\widehat{\pi}_{\text{tds}}(i),j}^* - M_{i,j}^*) \right| \leq 96(\zeta + 1) \sqrt{\frac{n_1 n_2}{N} |\text{BL}_k| \log(n_1 n_2)} \quad \text{for all } i \in [n_1], k \in [K]. \quad (4.12b)$$

We now prove inequality (4.12b). The proof of inequality (4.12a) follows in the same fashion. We split the proof into two cases.

Case 1. First, suppose that $|\text{BL}_k| \geq \frac{n_1 n_2}{N} \log(n_1 n_2)$. Applying Lemma 4.4.3 with $a_i = \sum_{\ell \in \text{BL}_k} M_{i,\ell}^*$, $\pi = \widehat{\pi}_{\text{tds}}$ and $\tau = 2\eta_k$, we see that for all $i \in [n_1]$,

$$\left| \sum_{\ell \in \text{BL}_k} (M_{\widehat{\pi}_{\text{tds}}(i),\ell}^* - M_{i,\ell}^*) \right| \leq 2\eta_k \leq 96(\zeta + 1) \sqrt{\frac{n_1 n_2}{N} |\text{BL}_k| \log(n_1 n_2)}.$$

Case 2. Otherwise, we have $|\text{BL}_k| \leq \frac{n_1 n_2}{N} \log(n_1 n_2)$. It then follows that

$$\left| \sum_{\ell \in \text{BL}_k} (M_{\widehat{\pi}_{\text{tds}}(i),\ell}^* - M_{i,\ell}^*) \right| \leq 2|\text{BL}_k| \leq 2 \sqrt{\frac{n_1 n_2}{N} |\text{BL}_k| \log(n_1 n_2)},$$

where we have used the fact that $M \in [0, 1]^{n_1 \times n_2}$.

Next, we consider concentration of column sums of $Y^{(1)}$. Applying Lemma 4.4.2 again with $\mathcal{S} = [n_1] \times \{j\}$, we obtain that

$$\left| C(j) - \sum_{i=1}^{n_1} M_{i,j}^* \right| \leq 8(\zeta + 1) \left(\sqrt{\frac{n_1^2 n_2}{N} \log(n_1 n_2)} + 2 \frac{n_1 n_2}{N} \log(n_1 n_2) \right) \quad (4.13)$$

for all $j \in [n_2]$ with probability at least $1 - 2(n_1 n_2)^{-3}$. We carry out the remainder of the proof conditioned on the event of probability at least $1 - 4(n_1 n_2)^{-3}$ that inequalities (4.12a), (4.12b) and (4.13) hold.

Having stated the necessary bounds, we now split the remainder of the proof into two parts for convenience. In order to do so, we first split the set BL into two disjoint sets of blocks, depending on whether a block comes from an originally large block (of size larger than $\beta = n_2 \sqrt{\frac{n_1}{N} \log(n_1 n_2)}$ as in Step 3 of Subroutine 1) or from an

aggregation of small blocks. More formally, define the sets

$$\begin{aligned}\text{BL}^{\text{L}} &:= \{B \in \text{BL} : B \text{ was not obtained via aggregation}\}, \text{ and} \\ \text{BL}^{\text{S}} &:= \text{BL} \setminus \text{BL}^{\text{L}}.\end{aligned}$$

For a set of blocks \mathbf{B} , define the shorthand $\cup \mathbf{B} = \bigcup_{B \in \mathbf{B}} B$ for convenience. We begin by focusing on the blocks BL^{L} .

Error on columns indexed by $\cup \text{BL}^{\text{L}}$

Recall that when the columns of the matrix are ordered according to $\widehat{\sigma}_{\text{pre}}$, the blocks in BL^{L} are contiguous and thus have an intrinsic ordering. We index the blocks according to this ordering as B_1, B_2, \dots, B_ℓ where $\ell = |\text{BL}^{\text{L}}|$. Now define the disjoint sets

$$\begin{aligned}\text{BL}^{(1)} &:= \{B_k \in \text{BL}^{\text{L}} : k \equiv 0 \pmod{2}\}, \text{ and} \\ \text{BL}^{(2)} &:= \{B_k \in \text{BL}^{\text{L}} : k \equiv 1 \pmod{2}\}.\end{aligned}$$

Let $\ell_t = |\text{BL}^{(t)}|$ for each $t = 1, 2$.

Recall that each block B_k in BL^{L} remains unchanged after aggregation, and that the threshold we used to block the columns is $\tau = 16(\zeta + 1) \left(\sqrt{\frac{n_1^2 n_2}{N} \log(n_1 n_2)} + 2 \frac{n_1 n_2}{N} \log(n_1 n_2) \right)$. Hence, applying the concentration bound (4.13) together with the definition of blocks in Step 2 of Subroutine 1 yields

$$\left| \sum_{i=1}^{n_1} M_{i,j_1}^* - \sum_{i=1}^{n_1} M_{i,j_2}^* \right| \leq 96(\zeta + 1) \sqrt{\frac{n_1^2 n_2}{N} \log(n_1 n_2)} \quad \text{for all } j_1, j_2 \in B_k, \quad (4.15)$$

where we again used the argument leading to claim (4.12b) to combine the two terms. Moreover, since the threshold is twice the concentration bound, it holds that under the true ordering id , every index in B_k precedes every index in B_{k+2} for any $k \in [K - 2]$. By definition, we have thus ensured that the blocks in $\text{BL}^{(t)}$ do not “mix” with each other.

The rest of the argument hinges on the following lemma, which is proved in Section 4.4.4.

Lemma 4.4.5. *For $m \in \mathbb{Z}_+$, let $J_1 \sqcup \dots \sqcup J_\ell$ be a partition of $[m]$ such that each J_k is contiguous and J_k precedes J_{k+1} . Let $a_k = \min J_k$, $b_k = \max J_k$ and $m_k = |J_k|$. Let A be a matrix in $[0, 1]^{n \times m}$ with nondecreasing rows and nondecreasing columns. Suppose that*

$$\sum_{i=1}^n (A_{i,b_k} - A_{i,a_k}) \leq \tau \quad \text{for each } k \in [\ell] \text{ and some } \tau \geq 0.$$

Additionally, suppose that there are positive reals $\rho, \rho_1, \rho_2, \dots, \rho_\ell$, and a permutation π such that for any $i \in [n]$, we have (i) $\sum_{j=1}^m |A_{\pi(i),j} - A_{i,j}| \leq \rho$, and (ii) $\sum_{j \in J_k} |A_{\pi(i),j} -$

$A_{i,j} \leq \rho_k$ for each $k \in [\ell]$. Then it holds that

$$\sum_{i=1}^n \sum_{j=1}^m (A_{\pi(i),j} - A_{i,j})^2 \leq 2\tau \sum_{k=1}^{\ell} \rho_k + n\rho \max_{k \in [\ell]} \frac{\rho_k}{m_k}.$$

We apply the lemma as follows. For $t = 1, 2$, let the matrix $M^{(t)}$ be the submatrix of M^* restricted to the columns indexed by the indices in $\cup \text{BL}^{(t)}$. The matrix $M^{(t)}$ has nondecreasing rows and columns by assumption. We have shown that the blocks in $\text{BL}^{(t)}$ do not mix with each other, so they are contiguous and correctly ordered in $M^{(t)}$. Moreover, the inequality assumptions of the lemma correspond to (4.15), (4.12a) and (4.12b) respectively, with the substitutions

$$\begin{aligned} A &= M^{(t)}, & n &= n_1, & m &= |\cup \text{BL}^{(t)}|, & \tau &= 96(\zeta + 1) \sqrt{\frac{n_1^2 n_2}{N} \log(n_1 n_2)} \\ \rho &= 96(\zeta + 1) \sqrt{\frac{n_1 n_2^2}{N} \log(n_1 n_2)}, & \rho_k &= 96(\zeta + 1) \sqrt{\frac{n_1 n_2}{N} |J_k| \log(n_1 n_2)}, \end{aligned}$$

and setting J_1, \dots, J_ℓ to be the blocks in $\text{BL}^{(t)}$. Therefore, applying Lemma 4.4.5 yields

$$\begin{aligned} & \sum_{i \in [n_1]} \sum_{j \in \cup \text{BL}^{(t)}} (M_{\pi_{\text{ids}}(i),j}^* - M_{i,j}^*)^2 \\ & \lesssim (\zeta^2 \vee 1) \frac{n_1^{3/2} n_2}{N} \log(n_1 n_2) \sum_{B \in \text{BL}^{(t)}} \sqrt{|B|} + (\zeta^2 \vee 1) \frac{n_1^2 n_2^{3/2}}{N} \log(n_1 n_2) \max_{B \in \text{BL}^{(t)}} \frac{\sqrt{|B|}}{|B|} \\ & \stackrel{(i)}{\leq} (\zeta^2 \vee 1) \frac{n_1^{3/2} n_2}{N} \log(n_1 n_2) \sqrt{\sum_{B \in \text{BL}^{(t)}} |B| \sqrt{\ell_t}} + (\zeta^2 \vee 1) \frac{n_1^2 n_2^{3/2}}{N} \frac{\log(n_1 n_2)}{\min_{B \in \text{BL}^{(t)}} \sqrt{|B|}} \\ & \stackrel{(ii)}{\leq} \frac{(\zeta^2 \vee 1) n_1^{3/2} n_2^2}{\sqrt{\beta} N} \log(n_1 n_2) + \frac{(\zeta^2 \vee 1) n_1^2 n_2^{3/2}}{\sqrt{\beta} N} \log(n_1 n_2) \\ & \lesssim \frac{(\zeta^2 \vee 1)}{\sqrt{\beta}} (n_1 n_2)^{3/2} (n_1 \vee n_2)^{1/2} \frac{\log(n_1 n_2)}{N}, \end{aligned}$$

where step (i) follows from the Cauchy-Schwarz inequality, and step (ii) follows from the fact that $\min_{B \in \text{BL}^{(t)}} |B| \geq \beta = n_2 \sqrt{\frac{n_1}{N} \log(n_1 n_2)}$ so that $\ell_t \leq n_2/\beta$. Substituting for β and normalizing by $n_1 n_2$ yields

$$\frac{1}{n_1 n_2} \sum_{i \in [n_1]} \sum_{j \in \cup \text{BL}^{(t)}} (M_{\pi_{\text{ids}}(i),j}^* - M_{i,j}^*)^2 \lesssim (\zeta^2 \vee 1) n_1^{1/4} (n_1 \vee n_2)^{1/2} \left(\frac{\log(n_1 n_2)}{N} \right)^{3/4}. \quad (4.16)$$

This proves the required result for the set of blocks $\text{BL}^{(t)}$. Summing over $t = 1, 2$ then yields a bound of twice the size for columns of the matrix indexed by $\cup \text{BL}^{\mathbb{L}}$.

Error on columns indexed by $\cup\text{BL}^{\mathbb{S}}$

Next we bound the approximation error of each row of the matrix with column indices restricted to the union of all small blocks. In the easy case where $\text{BL}^{\mathbb{S}}$ contains a single block of size less than $\frac{1}{2}n_2\sqrt{\frac{n_1}{N}\log(n_1n_2)}$, we have

$$\begin{aligned}
\sum_{i \in [n_1]} \sum_{j \in \cup\text{BL}^{\mathbb{S}}} (M_{\hat{\pi}_{\text{tds}}(i),j}^* - M_{i,j}^*)^2 &\stackrel{(i)}{\leq} \sum_{i \in [n_1]} \sum_{j \in \cup\text{BL}^{\mathbb{S}}} |M_{\hat{\pi}_{\text{tds}}(i),j}^* - M_{i,j}^*| \\
&\stackrel{(ii)}{=} \sum_{i \in [n_1]} \left| \sum_{j \in \cup\text{BL}^{\mathbb{S}}} (M_{\hat{\pi}_{\text{tds}}(i),j}^* - M_{i,j}^*) \right| \\
&\stackrel{(iii)}{\leq} \sum_{i \in [n_1]} 96(\zeta + 1) \sqrt{\frac{n_1 n_2}{2N} n_2 \left[\frac{(n_1 \vee n_2)}{N} \right]^{1/2} \log^{3/2}(n_1 n_2)} \\
&= 48\sqrt{2}(\zeta + 1) \frac{n_1^{3/2} n_2 (n_1 \vee n_2)^{1/4}}{N^{3/4}} \log^{3/4}(n_1 n_2),
\end{aligned}$$

where step (i) follows from the Hölder's inequality and the fact that $M^* \in [0, 1]^{n_1 \times n_2}$, step (ii) from the monotonicity of the columns of M^* , and step (iii) from equation (4.12a).

Now we aim to prove a bound of the same order for the general case. Critical to our analysis is the following lemma:

Lemma 4.4.6. *For a vector $v \in \mathbb{R}^n$, define its variation as $\text{var}(v) = \max_i v_i - \min_i v_i$. Then we have*

$$\|v\|_2^2 \leq \text{var}(v) \|v\|_1 + \|v\|_1^2 / n.$$

See Section 4.4.4 for the proof of this claim.

For each $i \in [n_1]$, define Δ^i to be the restriction of the i -th row difference $M_{\hat{\pi}_{\text{tds}}(i)}^* - M_i^*$ to the union of blocks $\cup\text{BL}^{\mathbb{S}}$. For each block $B \in \text{BL}^{\mathbb{S}}$, denote the restriction of Δ^i to B by Δ_B^i . Lemma 4.4.6 applied with $v = \Delta^i$ yields

$$\begin{aligned}
\|\Delta^i\|_2^2 &= \sum_{B \in \text{BL}^{\mathbb{S}}} \|\Delta_B^i\|_2^2 \\
&\leq \sum_{B \in \text{BL}^{\mathbb{S}}} \text{var}(\Delta_B^i) \|\Delta_B^i\|_1 + \sum_{B \in \text{BL}^{\mathbb{S}}} \frac{\|\Delta_B^i\|_1^2}{|B|} \\
&\leq \left(\max_{B \in \text{BL}^{\mathbb{S}}} \|\Delta_B^i\|_1 \right) \sum_{B \in \text{BL}^{\mathbb{S}}} \text{var}(\Delta_B^i) + \frac{\max_{B \in \text{BL}^{\mathbb{S}}} \|\Delta_B^i\|_1}{\min_{B \in \text{BL}^{\mathbb{S}}} |B|} \sum_{B \in \text{BL}^{\mathbb{S}}} \|\Delta_B^i\|_1 \\
&\leq \left(\max_{B \in \text{BL}^{\mathbb{S}}} \|\Delta_B^i\|_1 \right) \left(\sum_{B \in \text{BL}^{\mathbb{S}}} \text{var}(\Delta_B^i) \right) + \frac{\max_{B \in \text{BL}^{\mathbb{S}}} \|\Delta_B^i\|_1}{\min_{B \in \text{BL}^{\mathbb{S}}} |B|} \sum_{B \in \text{BL}^{\mathbb{S}}} \|\Delta_B^i\|_1.
\end{aligned} \tag{4.17}$$

We now analyze the quantities in inequality (4.17). By the aggregation step of Subroutine 1, we have $\frac{1}{2}\beta \leq |B| \leq 2\beta$, where $\beta = n_2\sqrt{\frac{n_1}{N}\log(n_1n_2)}$. Additionally, the bounds (4.12a) and (4.12b) imply that

$$\begin{aligned} \sum_{B \in \text{BL}^S} \|\Delta_B^i\|_1 &= \|\Delta^i\|_1 \leq 96(\zeta + 1)\sqrt{\frac{n_1n_2^2}{N}\log(n_1n_2)} \lesssim (\zeta + 1)\beta, \quad \text{and} \\ \|\Delta_B^i\|_1 &\leq 96(\zeta + 1)\sqrt{\frac{n_1n_2}{N}|B|\log(n_1n_2)} \\ &\leq 96\sqrt{2}(\zeta + 1)\sqrt{\frac{n_1n_2}{N}\beta\log(n_1n_2)} \quad \text{for all } B \in \text{BL}^S. \end{aligned}$$

Moreover, to bound the quantity $\sum_{B \in \text{BL}^S} \text{var}(\Delta_B^i)$, we proceed as in the proof for the large blocks in BL^L . Recall that if we permute the columns by $\hat{\sigma}_{\text{pre}}$ according to the column sums, then the blocks in BL^S have an intrinsic ordering, even after adjacent small blocks are aggregated. Let us index the blocks in BL^S by B_1, B_2, \dots, B_m according to this ordering, where $m = |\text{BL}^S|$. As before, the odd-indexed (or even-indexed) blocks do not mix with each other under the true ordering id , because the threshold used to define the blocks is larger than twice the column sum perturbation. We thus have

$$\begin{aligned} \sum_{B \in \text{BL}^S} \text{var}(\Delta_B^i) &= \sum_{\substack{k \in [m] \\ k \text{ odd}}} \text{var}(\Delta_{B_k}^i) + \sum_{\substack{k \in [m] \\ k \text{ even}}} \text{var}(\Delta_{B_k}^i) \\ &\leq \sum_{\substack{k \in [m] \\ k \text{ odd}}} [\text{var}(M_{i, B_k}^*) + \text{var}(M_{\hat{\pi}_{\text{tds}}(i), B_k}^*)] \\ &\quad + \sum_{\substack{k \in [m] \\ k \text{ even}}} [\text{var}(M_{i, B_k}^*) + \text{var}(M_{\hat{\pi}_{\text{tds}}(i), B_k}^*)] \\ &\stackrel{(i)}{\leq} 2 \text{var}(M_i^*) + 2 \text{var}(M_{\hat{\pi}_{\text{tds}}(i)}^*) \stackrel{(ii)}{\leq} 4, \end{aligned}$$

where inequality (i) holds because the odd (or even) blocks do not mix, and inequality (ii) holds because M^* has monotone rows in $[0, 1]^{n_2}$.

Finally, putting together all the pieces, we can substitute for β , sum over the indices $i \in n_1$, and normalize by n_1n_2 to obtain

$$\frac{1}{n_1n_2} \sum_{i \in [n_1]} \|\Delta^i\|_2^2 \lesssim (\zeta^2 \vee 1) \left(\frac{n_1 \log(n_1n_2)}{N} \right)^{3/4}, \quad (4.18)$$

and so the error on columns indexed by the set $\cup \text{BL}^S$ is bounded as desired.

Combining the bounds (4.16) and (4.18), we conclude that

$$\frac{1}{n_1 n_2} \|M^*(\widehat{\pi}_{\text{tds}}, \text{id}) - M^*\|_F^2 \lesssim (\zeta^2 \vee 1) n_1^{1/4} (n_1 \vee n_2)^{1/2} \left(\frac{\log(n_1 n_2)}{N} \right)^{3/4}$$

with probability at least $1 - 4(n_1 n_2)^{-3}$. The same proof works with the roles of n_1 and n_2 switched and all the matrices transposed, so it holds with the same probability that

$$\frac{1}{n_1 n_2} \|M^*(\text{id}, \widehat{\sigma}_{\text{tds}}) - M^*\|_F^2 \lesssim (\zeta^2 \vee 1) n_2^{1/4} (n_1 \vee n_2)^{1/2} \left(\frac{\log(n_1 n_2)}{N} \right)^{3/4}.$$

Consequently,

$$\frac{1}{n_1 n_2} (\|M^*(\widehat{\pi}_{\text{tds}}, \text{id}) - M^*\|_F^2 + \|M^*(\text{id}, \widehat{\sigma}_{\text{tds}}) - M^*\|_F^2) \lesssim (\zeta^2 \vee 1) \left(\frac{n_1 \log n_1}{N} \right)^{3/4}$$

with probability at least $1 - 8(n_1 n_2)^{-3}$, where we have used the relation $n_1 \geq n_2$. Applying Proposition 4.2.2 completes the proof.

Proof of Lemma 4.4.5

Since A has increasing rows, for any $i, i_2 \in [n]$ with $i \leq i_2$ and any $j, j_2 \in J_k$, we have

$$\begin{aligned} A_{i_2, j} - A_{i, j} &= (A_{i_2, j} - A_{i_2, a_k}) + (A_{i_2, a_k} - A_{i, b_k}) + (A_{i, b_k} - A_{i, j}) \\ &\leq (A_{i_2, b_k} - A_{i_2, a_k}) + (A_{i_2, j_2} - A_{i, j_2}) + (A_{i, b_k} - A_{i, a_k}). \end{aligned}$$

Choosing $j_2 = \arg \min_{r \in J_k} (A_{i_2, r} - A_{i, r})$, we obtain

$$A_{i_2, j} - A_{i, j} \leq (A_{i_2, b_k} - A_{i_2, a_k}) + (A_{i, b_k} - A_{i, a_k}) + \frac{1}{m_k} \sum_{r \in J_k} (A_{i_2, r} - A_{i, r}).$$

Together with the assumption on π , this implies that

$$|A_{\pi(i), j} - A_{i, j}| \leq \underbrace{A_{\pi(i), b_k} - A_{\pi(i), a_k}}_{=: x_{i, k}} + \underbrace{A_{i, b_k} - A_{i, a_k}}_{=: y_{i, k}} + \frac{1}{m_k} \sum_{r \in J_k} \underbrace{|A_{\pi(i), r} - A_{i, r}|}_{=: z_{i, k}}.$$

Hence it follows that

$$\begin{aligned}
\sum_{i=1}^n \sum_{j=1}^m (A_{i,j} - A_{\pi(i),j})^2 &= \sum_{i=1}^n \sum_{k=1}^{\ell} \sum_{j \in J_k} (A_{i,j} - A_{\pi(i),j})^2 \\
&\leq \sum_{i=1}^n \sum_{k=1}^{\ell} \sum_{j \in J_k} |A_{i,j} - A_{\pi(i),j}| (x_{i,k} + y_{i,k} + z_{i,k}/m_k) \\
&= \sum_{i=1}^n \sum_{k=1}^{\ell} z_{i,k} (x_{i,k} + y_{i,k} + z_{i,k}/m_k).
\end{aligned}$$

According to the assumptions, we have

1. $\sum_{k=1}^{\ell} x_{i,k} \leq 1$ and $\sum_{i=1}^n x_{i,k} \leq \tau$ for any $i \in [n], k \in [\ell]$;
2. $\sum_{k=1}^{\ell} y_{i,k} \leq 1$ and $\sum_{i=1}^n y_{i,k} \leq \tau$ for any $i \in [n], k \in [\ell]$;
3. $z_{i,k} \leq \rho_k$ and $\sum_{k=1}^{\ell} z_{i,k} \leq \rho$ for any $i \in [n], k \in [\ell]$.

Consequently, the following bounds hold:

1. $\sum_{i=1}^n \sum_{k=1}^{\ell} z_{i,k} x_{i,k} \leq \sum_{i=1}^n \sum_{k=1}^{\ell} \rho_k x_{i,k} \leq \tau \sum_{k=1}^{\ell} \rho_k$;
2. $\sum_{i=1}^n \sum_{k=1}^{\ell} z_{i,k} y_{i,k} \leq \sum_{i=1}^n \sum_{k=1}^{\ell} \rho_k y_{i,k} \leq \tau \sum_{k=1}^{\ell} \rho_k$;
3. $\sum_{i=1}^n \sum_{k=1}^{\ell} z_{i,k}^2 / m_k \leq \sum_{i=1}^n \sum_{k=1}^{\ell} z_{i,k} \cdot \max_{k \in [\ell]} (\rho_k / m_k) \leq n \rho \max_{k \in [\ell]} (\rho_k / m_k)$.

Combining these inequalities yields the claim.

Proof of Lemma 4.4.6

Let $a = \min_{i \in [n]} v_i$ and $b = \max_{i \in [n]} v_i = a + \text{var}(v)$. Since the quantities in the inequality remain the same if we replace v by $-v$, we assume without loss of generality that $b \geq 0$. If $a \leq 0$, then $\|v\|_{\infty} \leq b - a = \text{var}(v)$. If $a > 0$, then $a \leq \|v\|_1/n$ and $\|v\|_{\infty} = b \leq \|v\|_1/n + \text{var}(v)$. Hence in any case we have $\|v\|_2^2 \leq \|v\|_{\infty} \|v\|_1 \leq [\|v\|_1/n + \text{var}(v)] \|v\|_1$.

4.4.5 Proof of Lemma 4.4.1

The proof parallels that of Shah et al. [SBGW17, Theorem 5(a)], so we only emphasize the differences and sketch the remaining argument. We may assume that $p_{\text{obs}} \geq \frac{1}{n_2}$, since otherwise the bound is trivial.

We first employ a truncation argument. Consider the event

$$\mathcal{E} := \left\{ |W_{i,j}| \leq \frac{c_3}{p_{\text{obs}}} (\zeta \vee 1) \sqrt{\log(n_1 n_2)} \text{ for all } i \in [n_1], j \in [n_2] \right\}.$$

If the universal constant c_3 is chosen to be sufficiently large, then it follows from the sub-Gaussianity of $W_{i,j}$ and a union bound over all index pairs $(i, j) \in [n_1] \times [n_2]$ that $\Pr\{\mathcal{E}\} \geq 1 - (n_1 n_2)^{-4}$. Now define the truncation operator

$$T_\lambda(x) := \begin{cases} x & \text{if } |x| \leq \lambda, \\ \lambda \cdot \text{sgn}(x) & \text{otherwise.} \end{cases} \quad (4.19)$$

With the choice $\lambda = \frac{c_3}{p_{\text{obs}}}(\zeta \vee 1)\sqrt{\log(n_1 n_2)}$, define the random variables $W_{i,j}^{(1)} = T_\lambda(W_{i,j})$ for each pair of indices $(i, j) \in [n_1] \times [n_2]$. Consider the model where we observe $M^* + W^{(1)}$ instead of $Y = M^* + W$. Then the new model and the original one are coupled so that they coincide on the event \mathcal{E} . Therefore, it suffices to prove a high probability bound assuming that the noise is given by $W^{(1)}$.

Let us define $\mu = \mathbb{E}[W^{(1)}]$ and $\widetilde{W} = W^{(1)} - \mu$. We claim that for any $i \in [n_1], j \in [n_2]$, the following relations hold:

1. $|\mu_{i,j}| \leq \frac{c}{p_{\text{obs}}}(\zeta \vee 1)(n_1 n_2)^{-4}$;
2. $\widetilde{W}_{i,j}$ are independent, centered and $\frac{c}{p_{\text{obs}}}(\zeta \vee 1)$ -sub-Gaussian;
3. $|\widetilde{W}_{i,j}| \leq \frac{c}{p_{\text{obs}}}(\zeta \vee 1)\sqrt{\log(n_1 n_2)}$;
4. $\mathbb{E}[|\widetilde{W}_{i,j}|^2] \leq \frac{c}{p_{\text{obs}}}(\zeta^2 \vee 1)$.

Taking these claims as given for the moment, we turn to the main argument assuming that our observations take the form $Y = M^* + \widetilde{W} + \mu$.

For any permutations $\pi \in \mathfrak{S}_{n_1}, \sigma \in \mathfrak{S}_{n_2}$, let $M_{\pi,\sigma} = \widehat{M}_{\text{LS}}(Y)$. We claim that for any fixed pair (π, σ) such that $\|Y - M_{\pi,\sigma}\|_F^2 \leq \|Y - M^*\|_F^2$, we have

$$\Pr \left\{ \|M_{\pi,\sigma} - M^*\|_F^2 \geq c_1(\zeta^2 \vee 1) \frac{n_1}{p_{\text{obs}}} \log^2(n_1) \right\} \leq n_1^{-3n_1}. \quad (4.20)$$

Treating claim (4.20) as true for the moment, we see that since the least squares estimator \widehat{M} is equal to $M_{\pi,\sigma}$ for some pair (π, σ) , a union bound over $\pi \in \mathfrak{S}_{n_1}, \sigma \in \mathfrak{S}_{n_2}$ yields

$$\Pr \left\{ \|\widehat{M} - M^*\|_F^2 \geq c_1(\zeta^2 \vee 1) \frac{n_1}{p_{\text{obs}}} \log^2 n_1 \right\} \leq n_1^{-n_1},$$

which completes the proof. Thus, to prove our result, it suffices to prove claim (4.20).

Let $\Delta_{\pi,\sigma} = M_{\pi,\sigma} - M^*$. The condition $\|Y - M_{\pi,\sigma}\|_F^2 \leq \|Y - M^*\|_F^2$ yields the basic inequality

$$\frac{1}{2} \|\Delta_{\pi,\sigma}\|_F^2 \leq \langle \Delta_{\pi,\sigma}, \widetilde{W} + \mu \rangle.$$

Since $\Delta_{\pi,\sigma} \in [-1, 1]^{n_1 \times n_2}$, we have $\langle \Delta_{\pi,\sigma}, \mu \rangle \leq \|\mu\|_1 \leq \frac{c}{p_{\text{obs}}}(\zeta \vee 1)n_1^{-6}$ by claim 1. If it holds that $\|\Delta_{\pi,\sigma}\|_F^2 \leq \frac{4c}{p_{\text{obs}}}(\zeta \vee 1)n_1^{-6}$, then the proof is immediate. Thus, we may

assume the opposite, from which it follows that

$$\frac{1}{4}\|\Delta_{\pi,\sigma}\|_F^2 \leq \langle \Delta_{\pi,\sigma}, \widetilde{W} \rangle. \quad (4.21)$$

Consider the set of matrices

$$\mathbb{C}_{\text{DIFF}}(\pi, \sigma) := \{\alpha(M - M^*) : M \in \mathbb{C}_{\text{BISO}}(\pi, \sigma), \alpha \in [0, 1]\}.$$

Additionally, for every $t > 0$, define the random variable

$$Z_{\pi,\sigma}(t) := \sup_{\substack{D \in \mathbb{C}_{\text{DIFF}}(\pi, \sigma), \\ \|D\|_F \leq t}} \langle D, \widetilde{W} \rangle.$$

For every $t > 0$, define the event

$$\mathcal{A}_t := \left\{ \exists D \in \mathbb{C}_{\text{DIFF}}(\pi, \sigma) \text{ s.t. } \|D\|_F \geq \sqrt{t\delta_n} \text{ and } \langle D, \widetilde{W} \rangle \geq 4\|D\|_F\sqrt{t\delta_n} \right\}.$$

For $t \geq \delta_n$, either we already have $\|\Delta_{\pi,\sigma}\|_F^2 \leq t\delta_n$, or we have $\|\Delta_{\pi,\sigma}\|_F > \sqrt{t\delta_n}$. In the latter case, on the complement of \mathcal{A}_t , we must have $\langle \Delta_{\pi,\sigma}, \widetilde{W} \rangle \leq 4\|\Delta_{\pi,\sigma}\|_F\sqrt{t\delta_n}$. Combining this with inequality (4.21) then yields $\|\Delta_{\pi,\sigma}\|_F^2 \leq ct\delta_n$. It thus remains to bound the probability $\Pr\{\mathcal{A}_t\}$.

Using the star-shaped nature of the set $\mathbb{C}_{\text{DIFF}}(\pi, \sigma)$, a rescaling argument yields

$$\Pr\{\mathcal{A}_t\} \leq \Pr\left\{ Z_{\pi,\sigma}(\delta_n) \geq 4\delta_n\sqrt{t\delta_n} \right\} \quad \text{for all } t \geq \delta_n.$$

The following lemma bounds the tail behavior of the random variable $Z_{\pi,\sigma}(\delta_n)$, and its proof is postponed to Section 4.4.5.

Lemma 4.4.7. *For any $\delta > 0$ and $u > 0$, we have*

$$\begin{aligned} \Pr\left\{ Z_{\pi,\sigma}(\delta) > \frac{c}{p_{\text{obs}}}(\zeta \vee 1)\sqrt{\log n_1} (n_1 \log^{1.5} n + u) \right\} \\ \leq \exp\left(\frac{-c_1 u^2}{p_{\text{obs}}\delta^2/(\log n_1) + n_1 \log^{1.5} n_1 + u} \right). \end{aligned}$$

Taking the lemma as given and setting $\delta_n^2 = \frac{c_2}{p_{\text{obs}}}(\zeta^2 \vee 1)n_1 \log^2 n_1$ and $u = c_3(\zeta \vee 1)n_1 \log^{1.5} n_1$, we see that for any $t \geq \delta_n$, we have

$$\begin{aligned} \Pr\{\mathcal{A}_t\} &\leq \Pr\left\{ Z_{\pi,\sigma}(\delta_n) \geq 4\delta_n\sqrt{t\delta_n} \right\} \\ &\leq \exp\left(\frac{-c_4(\zeta^2 \vee 1)n_1^2 \log^3 n_1}{(\zeta^2 \vee 1)n_1 \log n_1 + n_1 \log^{1.5} n_1} \right) \leq n_1^{-3n_2}. \end{aligned} \quad (4.22)$$

In particular, for $t = \delta_n$, on the complement of \mathcal{A}_t , we have

$$\|\Delta_{\pi,\sigma}\|_F^2 \leq \frac{c_5}{p_{\text{obs}}}(\zeta^2 \vee 1)n_1 \log^2 n_1,$$

which completes the proof. Note that the original proof sacrificed a logarithmic factor in proving the equivalent of equation (4.22), and this is why we recover the same logarithmic factors as in the bounded case in spite of the sub-Gaussian truncation argument.

In the setting where we know that $M^* \in \mathbb{C}_{\text{BISO}}$, the same proof clearly works, except that we do not even need to take a union bound over $\pi \in \mathfrak{S}_{n_1}, \sigma \in \mathfrak{S}_{n_2}$ as the columns and rows are ordered.

Proof of claims 1–4

We assume throughout that the constant c_3 is chosen to be sufficiently large. Claim 1 follows as a result of the following argument; we have

$$\begin{aligned} |\mu_{i,j}| &= \left| \mathbb{E}[W_{i,j}^{(1)}] \right| \\ &\leq \mathbb{E} \left[|W_{i,j}^{(1)} - W_{i,j}| \right] \\ &= \int_0^\infty \Pr\{|W_{i,j}^{(1)} - W_{i,j}| \geq t\} dt \\ &= \int_0^\infty \Pr\{|W_{i,j}| \geq \frac{c_3}{p_{\text{obs}}}(\zeta \vee 1)\sqrt{\log(n_1 n_2)} + t\} dt \\ &\leq (n_1 n_2)^{-5} \int_0^\infty \exp\left(\frac{-t^2}{c_4(\zeta^2 \vee 1)/p_{\text{obs}}^2}\right) dt \\ &\leq \frac{c_5}{p_{\text{obs}}}(\zeta \vee 1)(n_1 n_2)^{-4}. \end{aligned}$$

By definition, the random variables $W_{i,j}^{(1)} - \mu_{i,j}$ are independent and zero-mean, and applying Lemma 4.7.1 (see Appendix 4.7) yields that they are also sub-Gaussian with the claimed variance parameter, thus yielding claim 2. The triangle inequality together with the definition of $\widetilde{W}_{i,j}$ then yields claim 3.

Finally, since $|T(x)| \leq |x|$, we have

$$\mathbb{E}[|\widetilde{W}_{i,j}|^2] \leq \mathbb{E}[|W_{i,j}^{(1)}|^2] \leq \mathbb{E}[|W_{i,j}|^2] \leq \frac{c_6}{p_{\text{obs}}}(\zeta^2 \vee 1),$$

yielding claim 4.

Proof of Lemma 4.4.7

The chaining argument from the proof of Shah et al. [SBGW17, Lemma 10] can be applied to show that

$$\mathbb{E}[Z_{\pi,\sigma}(\delta)] \leq \frac{c_2}{p_{\text{obs}}}(\zeta \vee 1)n_1 \log^2 n_1,$$

as $\widetilde{W}_{i,j}$ is $\frac{c}{p_{\text{obs}}}(\zeta \vee 1)$ -sub-Gaussian by claim 2. Note that although we are considering a set of rectangular matrices $\mathbb{C}_{\text{DIFF}}(\pi, \sigma) \subset [-1, 1]^{n_1 \times n_2}$ instead of square matrices as in [SBGW17], we can augment each matrix by zeros to obtain an $n_1 \times n_1$ matrix, and so $\mathbb{C}_{\text{DIFF}}(\pi, \sigma)$ can be viewed as a subset of its counterpart consisting of $n_1 \times n_1$ matrices. Hence the entropy bound depending on n_1 can be employed so that the chaining argument indeed goes through.

In order to obtain the deviation bound, we apply Lemma 11 of [SBGW17] (i.e., Theorem 1.1(c) of Klein and Rio [KR05]) with $\mathcal{V} = \mathbb{C}_{\text{DIFF}}(\pi, \sigma) \cap \mathcal{B}_\delta$, $m = n_1 n_2$, $X = \frac{p_{\text{obs}}}{c(\zeta \vee 1)\sqrt{\log n_1}} \widetilde{W}$ and $X^\dagger = \frac{p_{\text{obs}}}{c(\zeta \vee 1)\sqrt{\log n_1}} Z_{\pi,\sigma}(\delta)$. Claim 3 guarantees that $|X|$ is uniformly bounded by 1. We also have $\mathbb{E}[\langle D, \widetilde{W} \rangle^2] \leq \frac{c}{p_{\text{obs}}}(\zeta^2 \vee 1)\delta^2$ by claim 4 for $\|D\|_F^2 \leq \delta^2$. Therefore, we conclude that

$$\begin{aligned} \Pr \left\{ Z_{\pi,\sigma}(\delta) > \mathbb{E}[Z_{\pi,\sigma}(\delta)] + \frac{c}{p_{\text{obs}}}(\zeta \vee 1)\sqrt{\log n_1} \cdot u \right\} \\ \leq \exp \left(\frac{-c_1 u^2}{p_{\text{obs}} \delta^2 / (\log n_1) + n_1 \log^{1.5} n_1 + u} \right). \end{aligned}$$

Combining the expectation and the deviation bounds completes the proof.

4.5 Discussion

While the current paper narrows the statistical-computational gap for estimation in permutation-based models with monotonicity constraints, several intriguing questions remain:

- Can Algorithm 2 be recursed so as to improve the rate of estimation, until we eventually achieve the statistically optimal rate (up to lower-order terms) in polynomial time?
- If not, does there exist a statistical-computational gap in this problem, and if so, what is the fastest rate achievable by computationally efficient estimators?
- Can the techniques from here be used to narrow statistical-computational gaps in other permutation-based models [SBW16b, FMR16, PWC17]?

As a partial answer to the first question, it can be shown that when our two-dimensional sorting algorithm is recursed in the natural way and applied to the noisy sorting subclass of the SST model, it yields another minimax optimal estimator for

noisy sorting, similar to the multistage algorithm of Mao et al. [MWR17]. However, showing that this same guarantee is preserved for the larger class of SST matrices seems out of the reach. In fact, we conjecture that any algorithm that only exploits partial row and column sums cannot achieve a rate faster than $O(n^{-3/4})$ for the SST class.

It is also worth noting that the model (4.1) allowed us to perform multiple sample-splitting steps while preserving the independence across observations. While our proofs also hold for the observation model where we have exactly 3 independent samples per entry of the matrix, handling the weak dependence of the original sampling model with one observation per entry is an interesting technical challenge that may also involve its own statistical-computational tradeoffs [Mon15].

4.6 Appendix: Poissonization reduction

In this section, we show that estimation error bounds proved under a Poissonized observation model are equivalent, up to constant factors, to bounds proved without Poissonization. Note that we can assume that $N \geq 4 \log(n_1 n_2)$, since otherwise, all the bounds in the theorems hold trivially.

In order to prove the upper bound, assume that we have an estimator $\widehat{M}_{\text{Poi}}(N)$, which is designed under $N' = \text{Poi}(N)$ observations $\{y_\ell\}_{\ell=1}^{N'}$. Now, given exactly N observations $\{y_\ell\}_{\ell=1}^N$ from the model (4.1), choose an integer $\tilde{N} = \text{Poi}(N/2)$, and output the estimator

$$\widehat{M}(N) = \begin{cases} \widehat{M}_{\text{Poi}}(N/2) & \text{if } \tilde{N} \leq N, \\ 0 & \text{otherwise.} \end{cases}$$

Recalling the assumption $N \geq 4 \log(n_1 n_2)$, we have

$$\Pr\{\tilde{N} \geq N\} \leq e^{-N/2} \leq (n_1 n_2)^{-2}.$$

Thus, the error of the estimator $\widehat{M}(N)$ is bounded by $\frac{1}{n_1 n_2} \|\widehat{M}_{\text{Poi}}(N/2) - M^*\|_F^2$ with probability greater than $1 - (n_1 n_2)^{-2}$, and moreover, we have

$$\mathbb{E} \left[\frac{1}{n_1 n_2} \|\widehat{M}(N) - M^*\|_F^2 \right] \leq \mathbb{E} \left[\frac{1}{n_1 n_2} \|\widehat{M}_{\text{Poi}}(N/2) - M^*\|_F^2 \right] + (n_1 n_2)^{-2}.$$

In order to prove a lower bound, we must show the reverse, that an estimator $\widehat{M}(N)$ designed using exactly N samples may be used to estimate M^* under a Poissonized observation model. Given $\tilde{N} = \text{Poi}(2N)$ samples, define the estimator

$$\widehat{M}_{\text{Poi}}(2N) = \begin{cases} \widehat{M}(N) & \text{if } \tilde{N} \geq N, \\ 0 & \text{otherwise,} \end{cases}$$

where in the former case, $\widehat{M}(N)$ is computed by discarding $\tilde{N} - N$ samples at random.

Again, using the fact that $N \geq 4 \log(n_1 n_2)$ yields

$$\Pr\{\tilde{N} \geq N\} \leq e^{-N} \leq (n_1 n_2)^{-4},$$

and so once again, the error of the estimator $\widehat{M}_{\text{Poi}}(2N)$ is bounded by $\frac{1}{n_1 n_2} \|\widehat{M}(N) - M^*\|_F^2$ with probability greater than $1 - (n_1 n_2)^{-4}$. A similar guarantee also holds in expectation.

4.7 Appendix: Truncation preserves sub-Gaussianity

In this appendix, we show that truncating a sub-Gaussian random variable preserves its sub-Gaussianity to within a constant factor.

Lemma 4.7.1. *Let X be a (not necessarily centered) σ -sub-Gaussian random variable, and for some choice $\lambda \geq 0$, let $T_\lambda(X)$ denote its truncation according to equation (4.19). Then $T_\lambda(X)$ is $\sqrt{2}\sigma$ -sub-Gaussian.*

Proof. The proof follows a symmetrization argument. Let X' denote an i.i.d. copy of X , and use the shorthand $Y = T_\lambda(X)$ and $Y' = T_\lambda(X')$. Let ε denote a Rademacher random variable that is independent of everything else. Then Y and Y' are i.i.d., and $\varepsilon(Y - Y') \stackrel{d}{=} Y - Y'$. Hence we have

$$\begin{aligned} \mathbb{E} [e^{t(Y - \mathbb{E}[Y])}] &= \mathbb{E} [e^{t(Y - \mathbb{E}[Y'])}] \\ &\leq \mathbb{E}_{Y, Y'} [e^{t(Y - Y')}] \\ &= \mathbb{E}_{Y, Y', \varepsilon} [e^{t\varepsilon(Y - Y')}] . \end{aligned}$$

Using the Taylor expansion of e^x , we have

$$\begin{aligned} \mathbb{E} [e^{t(Y - \mathbb{E}[Y])}] &\leq \mathbb{E}_{Y, Y', \varepsilon} \left[\sum_{i \geq 0} \frac{1}{i!} (t\varepsilon(Y - Y'))^i \right] \\ &= \mathbb{E}_{Y, Y'} \left[\sum_{j \geq 0} \frac{1}{(2j)!} (t(Y - Y'))^{2j} \right] , \end{aligned}$$

since only the even moments remain. Finally, since the map $T_\lambda : \mathbb{R} \rightarrow \mathbb{R}$ is 1-Lipschitz, we have $|Y - Y'| \leq |X - X'|$, and combining this with the fact that $X - X'$ has odd

moments equal to zero yields

$$\begin{aligned}\mathbb{E} [e^{t(Y-\mathbb{E}[Y])}] &\leq \mathbb{E}_{X,X'} \left[\sum_{j \geq 0} \frac{1}{(2j)!} (t(X - X'))^{2j} \right] \\ &= \mathbb{E}_{X,X'} \left[\sum_{i \geq 0} \frac{1}{i!} (t(X - X'))^i \right] \\ &= \mathbb{E}_{X,X'} \left[e^{t(X-X')} \right] \\ &\leq e^{t^2 \sigma^2},\end{aligned}$$

where the last step follows since the random variable $X - X'$ is zero-mean and $\sqrt{2}\sigma$ -sub-Gaussian. \square

Chapter 5

Worst-case v.s. Average-case Design for Estimation from Fixed Pairwise Comparisons

The problems of ranking and estimation from ordinal data arise in a variety of disciplines, including web search and information retrieval [DKNS01], crowdsourcing [CBCTH13], tournament play [HMG06], social choice theory [CN91] and recommender systems [BMR10]. The ubiquity of such datasets stems from the relative ease with which ordinal data can be obtained, and from the empirical observation that using pairwise comparisons as a means of data elicitation can lower the noise level in the observations [Bar03, SBC05].

Given that the number of items n to be compared can be very large, it is often difficult or impossible to obtain comparisons between all $\binom{n}{2}$ pairs of items. A subset of pairs to compare, which defines the *comparison topology*, must therefore be chosen. For example, such topologies arise from tournament formats in sports, experimental designs in psychology set up to aid interpretability, or properties of the elicitation process. For instance, in rating movies, pairwise comparisons between items of the same genre are typically more abundant than comparisons between items of dissimilar genres. For these reasons, studying the performance of ranking algorithms based on fixed comparison topologies is of interest. Fixed comparison topologies are also important in rank breaking [HOX14, KO16], and more generally in matrix completion based on structured observations [KTT15, PABN16].

An important problem in ranking is the design of accurate models for capturing uncertainty in pairwise comparisons. Given a collection of n items, the results of pairwise comparisons are completely characterized by the n -dimensional matrix of comparison probabilities,¹ and various models have been proposed for such matrices. The most classical models, among them the Bradley-Terry-Luce [BT52, Luc59] and Thurstone models [Thu27], assign a quality vector to the set of items, and assign pairwise probabilities by applying a cumulative distribution function to the difference of qualities as-

¹A comparison probability refers to the probability that item i beats item j in a comparison between them.

sociated to the pair. There is now a relatively large body of work on methods for ranking in such parametric models (e.g., see the papers [NOS16, HOX14, CS15, SBB⁺16] as well as references therein). In contrast, less attention has been paid to a richer class of models proposed decades ago in the sociology literature [Fis73, ML65], which impose a milder set of constraints on pairwise comparison matrix. Rather than positing a quality vector, these models impose constraints that are typically given in terms of a latent permutation that rearranges the matrix into a specified form, and hence can be referred to as *permutation-based* models. Two such models that have been recently analyzed are those of strong stochastic transitivity [SBGW17], as well as the special case of noisy sorting [BM08]. The strong stochastic transitivity (SST) model, in particular, has been shown to offer significant robustness guarantees and provide a good fit to many existing datasets [BW97], and this flexibility has driven recent interest in understanding its properties. Also, perhaps surprisingly, past work has shown that this additional flexibility comes at only a small price when one has access to all possible pairwise comparisons, or more generally, to comparisons chosen at random [SBGW17]; in particular, the rates of estimation in these SST models differ from those in parametric models by only logarithmic factors in the number of items. On a related note, permutation-based models have also recently been shown to be useful in other settings like crowd-labeling [SBW16b], statistical seriation [FMR16] and linear regression [PWC16].

Given pairwise comparison data from one of these models, the problem of estimating the comparison probabilities has applications in inferring customer preferences in recommender systems, advertisement placement, and sports, and is the main focus of this chapter.

Our Contributions: Our goal is to estimate the matrix of comparison probabilities for fixed comparison topologies, studying both the noisy sorting and SST classes of matrices. Focusing first on the worst-case setting in which the assignment of items to the topology may be arbitrary, we show in Theorem 5.2.1 that consistent estimation is impossible for many natural comparison topologies. This result stands in sharp contrast to parametric models, and may be interpreted as a “no free lunch” theorem: although it is possible to estimate SST models at rates comparable to parametric models when given a full set of observations [SBGW17], the setting of fixed comparison topologies is problematic for the SST class. This can be viewed as a price to be paid for the additional robustness afforded by the SST model.

Seeing as such a worst-case design may be too strong for permutation-based models, we turn to an average-case setting in which the items are assigned to a fixed graph topology in a randomized fashion. Under such an observation model, we propose and analyze two efficient estimators: Theorems 5.2.2 and 5.2.4 show that consistent estimation is possible under commonly used comparison topologies. Moreover, the error rates of these estimators depend only on the degree sequence of the comparison topology, and are shown to be unimprovable for a large class of graphs, in Theorem 5.2.3.

Our results therefore establish a sharp distinction between worst-case and average-case designs when using fixed comparison topologies in permutation-based models.

Such a phenomenon arises from the difference between minimax risk and Bayes risk under a uniform prior on the ranking, and may also be worth studying for other ranking models.

Related Work: The literature on ranking and estimation from pairwise comparisons is vast, and we refer the reader to some surveys [FV93, Mar96, Cat12] and references therein for a more detailed overview. Estimation from pairwise comparisons has been analyzed under various metrics like top- k ranking [CS15, SW15, JKSO16, CGMS17] and comparison probability or parameter estimation [HOX14, SBB⁺16, SBGW17]. There have been studies of these problems under active [JN11, HSRW16, MG15], passive [NOS16, RA16], and collaborative settings [PNZ⁺15, NOTX17], and also for fixed as well as random comparison topologies [WJJ13, SBGW17]. Here we focus on the subset of papers that are most relevant to the work described here.

The problem of comparison probability estimation under a passively chosen fixed topology has been analyzed for parametric models by Hajek et al. [HOX14] and Shah et al. [SBB⁺16]. Both papers analyze the worst-case design setting in which the assignment of items to the topology may be arbitrary, and derive bounds on the minimax risk of parameter (or equivalently, comparison probability) estimation. While their characterizations are not sharp in general, the rates are shown to depend on the spectrum of the Laplacian matrix of the topology. We point out an interesting consequence of both results: in the parametric model, provided that the comparison graph G is connected, the maximum likelihood solution, in the limit of infinite samples for each graph edge, allows for exact recovery of the quality vector, and hence matrix of comparison probabilities. We will see that this property no longer holds for the SST models considered in this chapter: there are comparison topologies and SST matrices for which it is impossible to recover the full matrix even given an infinite amount of data per graph edge. It is also worth mentioning that the top- k ranking problem has been analyzed for parametric models under fixed design assumptions [JKSO16], and here as well, asymptotic consistency is observed for connected comparison topologies.

Notation: Here we summarize some notation used throughout the remainder of this chapter. We use n to denote the number of items, and adopt the shorthand $[n] := \{1, 2, \dots, n\}$. We use $\text{Ber}(p)$ to denote a Bernoulli random variable with success probability p . For two sequences $\{a_n\}_{n=1}^\infty$ and $\{b_n\}_{n=1}^\infty$, we write $a_n \lesssim b_n$ if there is a universal constant C such that $a_n \leq Cb_n$ for all $n \geq 1$. The relation $a_n \gtrsim b_n$ is defined analogously, and we write $a_n \asymp b_n$ if the relations $a_n \lesssim b_n$ and $a_n \gtrsim b_n$ hold simultaneously. We use c, c_1, c_2 to denote universal constants that may change from line to line.

We use $\mathbf{e} \in \mathbb{R}^n$ to denote the all-ones vector in \mathbb{R}^n . Given a matrix $M \in \mathbb{R}^{n \times n}$, its i -th row is denoted by M_i . For a graph G with edge set E , let $M(G)$ denote the entries of the matrix M restricted to the edge set of G , and let $\|M\|_E^2 = \sum_{(i,j) \in E} M_{ij}^2$. For a matrix $M \in \mathbb{R}^{n \times n}$ and a permutation $\pi : [n] \rightarrow [n]$, we use the shorthand $\pi(M) = \Pi M \Pi^\top$, where Π represents the row permutation matrix corresponding to the permutation π . We let id denote the identity permutation. The Kendall tau

distance [Ken48] between two permutations π and π' is given by

$$\text{KT}(\pi, \pi') := \sum_{i,j \in [n]} \mathbf{1}\{\pi(i) < \pi(j), \pi'(i) > \pi'(j)\}.$$

Let $\mathcal{C}(G)$ represent the set of all connected, vertex-induced subgraphs of a graph G , and let $V(S)$ and $E(S)$ represent the vertex and edge set of a subgraph S , respectively. We let $\alpha(G)$ denote the size of the largest independent set of the graph G , which is a largest subset of vertices that have no edges among them. Define a biclique of a graph as two disjoint subsets of its vertices V_1 and V_2 such that $(u, v) \in E(G)$ for all $u \in V_1$ and $v \in V_2$. Define the biclique number $\beta(G)$ as the maximum number of edges in any such biclique, given by $\max_{V_1, V_2 \text{ biclique}} |V_1||V_2|$. Let d_v denote the degree of vertex $v \in V$.

5.1 Background and problem setup

Consider a collection of $n \geq 2$ items that obey a total ordering or ranking determined by a permutation $\pi^* : [n] \rightarrow [n]$. More precisely, item $i \in [n]$ is preferred to item $j \in [n]$ in the underlying ranking if and only if $\pi^*(i) < \pi^*(j)$. We are interested in observations arising from stochastic pairwise comparisons between items. We denote the matrix of underlying comparison probabilities by $M^* \in [0, 1]^{n \times n}$, with $M_{ij}^* = \Pr\{i \succ j\}$ representing the probability that item i beats item j in a comparison.

Each item i is associated with a *score*, given by the probability that item i beats another item chosen uniformly at random. More precisely, the score τ_i^* of item i is given by

$$\tau_i^* := [\tau(M^*)]_i := \frac{1}{n-1} \sum_{j \neq i} M_{ij}^*. \quad (5.1)$$

Arranging the scores in descending order naturally yields a ranking of items. In fact, for the models we define below, the ranking given by the scores is consistent with the ranking given by π^* , i.e., $\tau_i \geq \tau_j$ if $\pi^*(i) < \pi^*(j)$. The converse also holds if the scores are distinct.

5.1.1 Pairwise comparison models

We consider a permutation-based model for the comparison matrix M^* , one defined by the property of *strong stochastic transitivity* [Fis73, ML65], or the SST property for short. In particular, a matrix M^* of pairwise comparison probabilities is said to obey the SST property if for items i, j and k in the total ordering such that $\pi^*(i) < \pi^*(j) < \pi^*(k)$, it holds² that $\Pr(i \succ k) \geq \Pr(i \succ j)$. Alternatively, recalling that $\pi(M)$ denotes the matrix obtained from M by permuting its rows and columns

²We set $M_{ii}^* = 1/2$ by convention.

according to the permutation π , the SST matrix class can be defined in terms of permutations applied to the class \mathbb{C}_{BISO} of bivariate isotonic matrices as

$$\mathbb{C}_{\text{SST}} := \bigcup_{\pi} \pi(\mathbb{C}_{\text{BISO}}) = \bigcup_{\pi} \{\pi(M) : M \in \mathbb{C}_{\text{BISO}}\}. \quad (5.2)$$

Here the class \mathbb{C}_{BISO} of bivariate isotonic matrices is given by

$$\{M \in [0, 1]^{n \times n} : M + M^{\top} = \mathbf{e}\mathbf{e}^{\top} \text{ and } M \text{ has non-decreasing rows} \\ \text{and non-increasing columns}\},$$

where $\mathbf{e} \in \mathbb{R}^n$ denotes a vector of all ones.

As shown by Shah et al. [SBGW17], the SST class is substantially larger than commonly used class of *parametric* models, in which each item i is associated with a parameter $w_i \in \mathbb{R}$, and the probability that item i beats item j is given by $F(w_i - w_j)$, where $F : \mathbb{R} \mapsto [0, 1]$ is a smooth monotone function of its argument.

A special case of the SST model that we study in this chapter is the *noisy sorting* model [BM08], in which the all underlying probabilities are described with a single parameter $\lambda \in [0, 1/2]$. The matrix $M_{\text{NS}}(\pi, \lambda) \in [0, 1]^{n \times n}$ has entries

$$[M_{\text{NS}}(\pi, \lambda)]_{ij} = 1/2 + \lambda \cdot \text{sgn}(\pi(j) - \pi(i)),$$

and the noisy sorting classes are given by

$$\mathbb{C}_{\text{NS}}(\lambda) := \bigcup_{\pi} \{M_{\text{NS}}(\pi, \lambda)\}, \quad \text{and} \quad \mathbb{C}_{\text{NS}} := \bigcup_{\lambda \in [0, 1/2]} \mathbb{C}_{\text{NS}}(\lambda). \quad (5.3)$$

Here $\text{sgn}(x)$ is the sign operator, with the convention that $\text{sgn}(0) = 0$. In words, the noisy sorting class models the case where the probability $\Pr\{i \succ j\}$ depends only on the parameter λ and whether $\pi^*(i) < \pi^*(j)$. Although a noisy sorting model is a very special case of an SST model, apart from the degenerate case $\lambda^* = 1/2$, it cannot be represented by any parametric model with a smooth function F , and so captures the essential difficulty of learning in the SST class.

We now turn to describing the observation models that we consider in this chapter.

5.1.2 Partial observation models

Our goal is to provide guarantees on estimating the underlying comparison matrix M^* when the comparison topology is fixed. Suppose that we are given data for comparisons in the form of a graph $G = (V, E)$, where the vertices represent the n items and edges represent the comparisons made between items. We assume that the observations obey the probabilistic model

$$Y_{ij} = \begin{cases} \text{Ber}(M_{ij}^*) & \text{for } (i, j) \in E, \text{ independently} \\ \star & \text{otherwise,} \end{cases} \quad (5.4)$$

where \star indicates a missing observation. We set the diagonal entries of Y equal to $1/2$, and also specify that $Y_{ji} = 1 - Y_{ij}$ for $j > i$, so that $Y + Y^\top = \mathbf{e}\mathbf{e}^\top$. We consider two different instantiations of the edge set given the graph.

Worst-case setting

In this setting, we assume that the assignment of items to vertices of the comparison graph G is arbitrary. In other words, once the graph G and its edges E are fixed, we observe the entries of the matrix according to the observation model (5.4), and would like to provide uniform guarantees in the metric $\|\widehat{M} - M^*\|_F^2$ over all matrices M^* in our model class given this restricted set of observations.

This setting is of the worst-case type, since the adversary is allowed to choose the underlying matrix with knowledge of the edge set E . Providing guarantees against such an adversary is known to be possible for parametric models [HOX14, SBB⁺16]. However, as we show in Section 5.2.1, such a guarantee is impossible to obtain even over the the noisy sorting subclass of the full SST class. Consequently, the latter parts of our analysis apply to a less rigid, average-case setting.

Average-case setting

In this setting, we assume that the assignment of items to vertices of the comparison graph G is random. Equivalently, given a fixed comparison graph G having adjacency matrix A , the subset of the entries that we observe can be modeled by the operator $\mathcal{O} = \sigma(A)$ for a permutation $\sigma : [n] \rightarrow [n]$ chosen uniformly at random. For a fixed comparison matrix M^* , our observations themselves consist of a random subset of the entries of the matrix Y determined by the operator \mathcal{O} : a location where $\mathcal{O}_{ij} = 1$ (respectively $\mathcal{O}_{ij} = 0$) indicates that entry Y_{ij} is observed (respectively is not observed). Such a setting is reasonable when the graph topology is constrained, but we are still given the freedom to assign items to vertices of the comparison graph, e.g. in psychology experiments. A natural extension of such an observation model is the one of k random designs, consisting of multiple random observation operators $\{\mathcal{O}_i = \sigma_i(A)\}_{i=1}^k$, chosen with independent, random permutations $\{\sigma_i\}_{i=1}^k$.

Our guarantees in the one sample setting with the observation operator \mathcal{O} can be seen as a form of Bayes risk, where given a fixed observation pattern E (consisting of the entries of the comparison matrix Y determined by the adjacency matrix A of the graph G , with A_{ij} representing the indicator that entry Y_{ij} is observed), we want to estimate a matrix M^* under a uniform Bayesian prior on the ranking π^* . Studying this average-case setting is well-motivated, since given fixed comparisons between a set of items, there is no reason to assume a priori that the underlying ranking is generated adversarially.

We are now ready to state the goal of the chapter. We address the problems of recovering the ranking π^* and estimating the matrix M^* in the Frobenius norm. More precisely, given the observation matrix $Y = Y(E)$ (where the set E is random in the average-case observation model), we would like to output a matrix \widehat{M} that is function

of Y , and for which good control on the Frobenius norm error $\|\widehat{M} - M^*\|_F^2$ can be guaranteed.

5.2 Main results

In this section, we state our main results and discuss some of their consequences. Proofs are deferred to Section 5.4.

5.2.1 Worst-case design: minimax bounds

In the worst-case setting of Section 5.1.2, the performance of an estimator is measured in terms of the normalized minimax error

$$\mathcal{M}(G, \mathbb{C}) = \inf_{\widehat{M}=f(Y(G))} \sup_{M^* \in \mathbb{C}} \mathbb{E} \left[\frac{1}{n^2} \|\widehat{M} - M^*\|_F^2 \right],$$

where the expectation is taken over the randomness in the observations Y as well as any randomness in the estimator, and $\mathbb{C} \in \{\mathbb{C}_{\text{SST}}, \mathbb{C}_{\text{NS}}\}$ represents the model class. Our first result shows that for many comparison topologies, the minimax risk is prohibitively large even for the noisy sorting model.

Theorem 5.2.1. *For any graph G , the diameter of the set consistent with observations on the edges of G is lower bounded as*

$$\sup_{\substack{M_1, M_2 \in \mathbb{C}_{\text{NS}} \\ M_1(G) = M_2(G)}} \|M_1 - M_2\|_F^2 \geq \alpha(G)(\alpha(G) - 1) \vee \beta(G^c). \quad (5.5a)$$

Consequently, the minimax risk of the noisy sorting model is lower bounded as

$$\mathcal{M}(G, \mathbb{C}_{\text{NS}}) \geq \frac{1}{4n^2} [\alpha(G)(\alpha(G) - 1) \vee \beta(G^c)]. \quad (5.5b)$$

Note that via the inclusion $\mathbb{C}_{\text{NS}} \subset \mathbb{C}_{\text{SST}}$, Theorem 5.2.1 also implies the same lower bound (5.5b) on the risk $\mathcal{M}(G, \mathbb{C}_{\text{SST}})$. In addition to these bounds, the lower bounds for estimation in parametric models, known from past work [SBB⁺16], carry over directly to the SST model, since parametric models are subclasses of the SST class.

Theorem 5.2.1 is approximation-theoretic in nature: more precisely, (5.5a) is a statement purely about the size of the set of matrices consistent with observations on the graph. Consequently, it does not capture the uncertainty due to noise, and thus can be a loose characterization of the minimax risk for some graphs, with the complete graph being one example. The bound (5.5a) on the diameter of the set of consistent observations may be interpreted as the worst case error in the infinite sample limit of observations on G . Hence, Theorem 5.2.1 stands in sharp contrast to analogous results for parametric models [HOX14, SBB⁺16], in which it suffices for the graph to be connected in order to obtain consistent estimation in the infinite sample

limit. For example, connected graphs with large independent sets of order n do not admit consistent estimation over the noisy sorting and hence SST classes.

It is also worth mentioning that the connectivity properties of the graph that govern minimax estimation in the larger SST model are quite different from those appearing in parametric models. In particular, the minimax rates for parametric models are closely related (via the linear observation model) to the spectrum of the Laplacian matrix of the graph G . In Theorem 5.2.1, however, we see other functions of the graph appearing that are not directly related to the Laplacian spectrum. In Section 5.3, we evaluate these functions for commonly used graph topologies, showing that for many of them, the risk is lower bounded by a constant even for graphs admitting consistent parametric estimation.

Seeing as the minimax error in the worst-case setting can be prohibitively large, we now turn to evaluating practical estimators in the random observation models of Section 5.1.2.

5.2.2 Average-case design: noisy sorting matrix estimation

In the average-case setting described in Section 5.1.2, we measure the performance of an estimator using the risk

$$\sup_{M^* \in \mathcal{C}} \mathbb{E}_{\mathcal{O}, Y} \frac{1}{n^2} \|\widehat{M} - M^*\|_F^2.$$

It is important to note that the expectation is taken over both the comparison noise, as well as the random observation pattern \mathcal{O} (or equivalently, the underlying random permutation σ assigning items to vertices). We propose the Average-Sort-Project estimator (ASP for short) for matrix estimation in this metric, which is a natural generalization of the Borda count estimator [CM16, SBW16a]. It consists of three steps, described below for the noisy sorting model:

- (1) **Averaging step:** Compute the average $\widehat{\tau}_i = \frac{\sum_{j \neq i} Y_{ij} \mathcal{O}_{ij}}{\sum_{j \neq i} \mathcal{O}_{ij}}$, corresponding to the fraction of comparisons won by item i .
- (2) **Sorting step:** Choose the permutation $\widehat{\pi}_{\text{ASP}}$ such that the sequence $\{\widehat{\tau}_{\widehat{\pi}_{\text{ASP}}^{-1}(i)}\}_{i=1}^n$ is decreasing in i , with ties broken arbitrarily.
- (3) **Projection step:** Find the maximum likelihood estimate $\widehat{\lambda}$ by treating $\widehat{\pi}_{\text{ASP}}$ as the true permutation that sorts items in decreasing order. Output the matrix $\widehat{M}_{\text{ASP}} := M_{\text{NS}}(\widehat{\pi}_{\text{ASP}}, \widehat{\lambda})$.

We now state an upper bound on the mean-squared Frobenius error achievable using the ASP estimator. It involves the degree sequence $\{d_v\}_{v \in V}$ of a graph G without isolated vertices, meaning that $d_v \geq 1$ for all $v \in V$.

Theorem 5.2.2. *Let the observation process be given by \mathcal{O} . For any graph $G = (V, E)$ without isolated vertices and any matrix $M^* \in \mathbb{C}_{\text{NS}}(\lambda^*)$, we have*

$$\mathbb{E}_{\mathcal{O}, Y} \left[\frac{1}{n^2} \|\widehat{M}_{\text{ASP}} - M^*\|_F^2 \right] \lesssim \frac{1}{|E|} + \frac{n \log n}{|E|^2} + \frac{\lambda^*}{n} \sum_{v \in V} \frac{1}{\sqrt{d_v}}, \quad \text{and} \quad (5.6a)$$

$$\mathbb{E}_{\mathcal{O}, Y} [\text{KT}(\pi^*, \widehat{\pi}_{\text{ASP}})] \lesssim \frac{n}{\lambda^*} \sum_{v \in V} \frac{1}{\sqrt{d_v}}. \quad (5.6b)$$

A few comments are in order. First, while the results are stated in expectation, a high probability bound can be proved for permutation estimation—namely

$$\Pr_{\mathcal{O}, Y} \left\{ \text{KT}(\pi^*, \widehat{\pi}_{\text{ASP}}) \geq \frac{n \sqrt{\log n}}{\lambda^*} \sum_{v \in V} \frac{1}{\sqrt{d_v}} \right\} \leq n^{-10}.$$

Second, it can be verified that $\frac{1}{|E|} + \frac{n \log n}{|E|^2} \lesssim \frac{1}{n} \sum_{v \in V} \frac{1}{\sqrt{d_v}}$, so that taking a supremum over the parameter $\lambda^* \in [0, 1/2]$ guarantees that the mean-squared Frobenius error is upper bounded as $O\left(\frac{1}{n} \sum_{v \in V} \frac{1}{\sqrt{d_v}}\right)$, uniformly over the entire noisy sorting class \mathbb{C}_{NS} . Third, it is also interesting to note the dependence of the bounds on the noise parameter λ^* of the noisy sorting model. The “high-noise” regime $\lambda^* \approx 0$ is a good one for estimating the underlying matrix, since the true matrix M^* is largely unaffected by errors in estimating the true permutation. However, as captured by equation (5.6b), the permutation estimation problem is more challenging in this regime.

The bound (5.6a) can be specialized to the complete graph K_n and the Erdős-Rényi random graph with edge probability p to obtain the rates $1/\sqrt{n}$ and $1/\sqrt{np}$, respectively, for estimation in the mean-squared Frobenius norm. These rates are strictly sub-optimal for these graphs, since the minimax rates scale as $1/n$ and $1/(np)$, respectively; both are achieved by the global MLE [SBGW17]. Such a phenomenon is consistent with the gap observed between computationally constrained and unconstrained estimators in similar and related problems [SBGW17, FMR16, PWC17].

Interestingly, it turns out that the estimation rate (5.6a) is optimal in a certain sense, and we require some additional notions to state this precisely. Fix constants $C_1 = 10^{-2}$ and $C_2 = 10^2$ and two sequences $\{a_n\}_{n \geq 1}$ and $\{b_n\}_{n \geq 1}$ of (strictly) positive scalars. For each $n \geq 1$, define the family of graphs

$$\mathcal{G}_n(a_n, b_n) := \left\{ G(V, E) \text{ is connected} : |V| = n, \right. \\ \left. C_1 a_n \leq |E| \leq C_2 a_n, \text{ and } C_1 b_n \leq \sum_{v \in V} \frac{1}{\sqrt{d_v}} \leq C_2 b_n \right\}.$$

As noted in Section 5.1.2, the average-case design observation model is equivalent to choosing the matrix M^* from a random ensemble with the permutation π^* chosen uniformly at random, and observing fixed pairwise comparisons. Such a viewpoint is useful in order to state our lower bound. Expectations are taken over the randomness of both π^* and the Bernoulli observation noise.

Theorem 5.2.3. (a) Let $M^* = M_{\text{NS}}(\pi^*, 1/4)$, where the permutation π^* is chosen uniformly at random on the set $[n]$. For any pair of sequences $(\{a_n\}_{n \geq 1}, \{b_n\}_{n \geq 1})$ such that the set $\mathcal{G}_n(a_n, b_n)$ is non-empty for every $n \geq 1$, and for any estimators $(\widehat{M}, \widehat{\pi})$ that are measurable functions of the observations on G , we have

$$\sup_{G \in \mathcal{G}_n(a_n, b_n)} \mathbb{E} \left[\frac{1}{n^2} \|\widehat{M} - M^*\|_F^2 \right] \gtrsim \frac{b_n}{n}, \text{ and } \sup_{G \in \mathcal{G}_n(a_n, b_n)} \mathbb{E} [\text{KT}(\pi^*, \widehat{\pi})] \gtrsim nb_n.$$

(b) For any graph G , let $M^* = M_{\text{NS}}(\pi^*, c\sqrt{n/|E|})$, with the permutation π^* chosen uniformly at random and the constant c chosen sufficiently small. Then for any estimators $(\widehat{M}, \widehat{\pi})$ that are measurable functions of the observations on G , we have

$$\mathbb{E} \left[\frac{1}{n^2} \|\widehat{M} - M^*\|_F^2 \right] \gtrsim \frac{n}{|E|}.$$

Parts (a) and (b) of the lower bound may be interpreted respectively as the approximation error caused by having observations only on a subset of edges, and the estimation error arising from the Bernoulli observation noise. Note that part (b) applies to every graph, and is particularly noteworthy for sparse graphs. In particular, in the regime in which the graph has bounded average degree, it shows that the inconsistency exhibited by the ASP estimator is unavoidable for any estimator. A more detailed discussion for specific graphs may be found in Section 5.3.

Although part (a) of the theorem is stated for a supremum over graphs, we actually prove a stronger result that explicitly characterizes the class of graphs that attain these lower bounds. As an example, given the sequences $a_n = n^2$ and $b_n = \sqrt{n}$, we show that the ASP estimator is information-theoretically optimal for the sequence of graphs consisting of two disjoint cliques $K_{n/2} \cup K_{n/2}$, which can be verified to lie within the class $\mathcal{G}(a_n, b_n)$.

The ASP estimator for the SST model would replace step (iii), as stated, by a maximum likelihood estimate using the entries on the edges that we observe. However, analyzing such an estimator given only a single sample on the entries \mathcal{O} is a challenging problem due to dependencies between the different steps of the estimator, and the difficulty of solving the associated matrix completion problem. Consequently, we turn to an observation model consisting of two random designs, and design a different estimator that renders the matrix completion problem tractable.

5.2.3 Two random designs: SST matrix estimation

Recall the average-case setting with multiple random designs, as described in Section 5.1.2, in which the comparison topology is fixed ahead of time, but one can collect multiple observations by assigning items to the vertices of the underlying graph at random. In this section, we rely on two such independent observations \mathcal{O}_1 and \mathcal{O}_2 to design an estimator that is consistent over the SST class. In order to describe our estimator, we require some additional notation. For any matrix $X \in [0, 1]^{n \times n}$ such that $X + X^\top = \mathbf{e}\mathbf{e}^\top$, we use $r(X) := X\mathbf{e}$ to denote the vector of its row sums. Note

that this vector is related to the vector of scores, as defined in equation (5.1), via $r(X) = (n-1)\tau(X) + 1/2$.

Our estimator relies on the approximation of any matrix $M^* \in \mathbb{C}_{\text{SST}}$ by a block-wise constant matrix, and we require some more definitions to make this precise. For any vector $v \in \mathbb{R}_+^n$, fix some value $t \in (0, n)$ and define a block partition $\text{bl}_t(v)$ of v as

$$[\text{bl}_t(v)]_i = \{j \in [n] : v_j \in [[(i-1)t], [it] - 1]\}.$$

In particular, the blocking vector $\text{bl}_t(r(X))$ contains a partition of indices such that the row sums of the matrix within each block of the partition are within a gap t of each other. Denote the set of all possible partitions of the set $[n]$ by χ_n . For any partition $C \in \chi_n$ of the indices $[n]$, define the set of blocks $\mathcal{B}(C) = \{S \times T : S, T \in C\}$.

By definition, given a partition $C \in \chi_n$ of $[n]$, the set $\mathcal{B}(C)$ is a partition of the set $[n] \times [n]$ into blocks. We are now ready to describe the blocking operation. For indices $i, j \in [n]$, denote by $B_C(i, j)$ the block in $\mathcal{B}(C)$ that contains the tuple (i, j) . Given a matrix $X \in [0, 1]^{n \times n}$ satisfying $X + X^\top = \mathbf{e}\mathbf{e}^\top$, we define the blocked version of X depending on observations in a set $E \subseteq [n] \times [n]$ as

$$[\mathbf{B}(X, C, E)]_{ij} = \begin{cases} \frac{1}{|B_C(i, j) \cap E|} \sum_{(k, \ell) \in B_C(i, j) \cap E} X_{k\ell} & \text{if } B_C(i, j) \cap E \neq \emptyset \\ 1/2 & \text{otherwise.} \end{cases} \quad (5.7)$$

In words, this defines a projection of the matrix X onto the set of block-wise constant matrices, by block-wise averaging the entries of X over the observed set of entries E . We now turn to our estimator, called the Block-Average-Project estimator (BAP for short), of the underlying matrix $M^* \in \mathbb{C}_{\text{SST}}$. Given the observation matrix Y_1 , define

$$[Y_1']_{ij} = \begin{cases} \frac{n}{D_i} [Y_1]_{ij} & \text{if entry } (i, j) \text{ is observed,} \\ 0 & \text{otherwise,} \end{cases}$$

where $D_i = \sum_{j=1}^n [\mathcal{O}_1]_{ij}$ is the (random) degree of item i . We now perform three steps:

- (1) **Blocking step:** Fix $S = \sum_{v \in V} 1/\sqrt{d_v}$, and obtain the blocking vector $\hat{b} = \text{bl}_S(r(Y_1'))$ and permutation $\hat{\pi}_{\text{ASP}}$ as in step (2) of the ASP estimator.
- (2) **Averaging step:** Average the matrix Y_2 within each block to obtain the matrix $\widetilde{M} = \mathbf{B}(Y_2, \hat{b}, E_2)$.
- (3) **Projection step:** Project onto the space $\hat{\pi}_{\text{ASP}}(\mathbb{C}_{\text{BISO}}) = \{\hat{\pi}_{\text{ASP}}(M) : M \in \mathbb{C}_{\text{BISO}}\}$, to obtain the estimator $\widetilde{M}_{\text{BAP}}$.

The blocking and averaging steps of the estimator are the main ingredients that we use to bound the error of the associated matrix completion problem. Also, the projection step of the estimator can be computed in polynomial time via bivariate isotonic regression [BDPR84].

Theorem 5.2.4. *Let the observation process be given by $\mathcal{O}_1 \cup \mathcal{O}_2$. For any graph G*

without isolated vertices and any matrix $M^* \in \mathbb{C}_{\text{SST}}$, we have

$$\mathbb{E} \left[\frac{1}{n^2} \|\widehat{M}_{\text{BAP}} - M^*\|_F^2 \right] \lesssim \frac{1}{n} \sum_{v \in V} \frac{1}{\sqrt{d_v}},$$

where the expectation is taken over the noise, and observation patterns \mathcal{O}_1 and \mathcal{O}_2 .

To be clear, the blocking estimate \widehat{M}_{BAP} is well-defined even when we have just one sample \mathcal{O}_1 instead of two samples \mathcal{O}_1 and \mathcal{O}_2 , where step (2) is replaced by the estimate $\widetilde{M} = \text{B}(Y_1, \widehat{b}, E_1)$. In the simulations of Section 5.3, we see that for a large variety of graphs, using a single sample \mathcal{O}_1 enjoys similar performance to using two independent samples \mathcal{O}_1 and \mathcal{O}_2 . We require two independent samples of the observations in our theoretical analysis to decouple the randomness of the first step of the algorithm from the second. When using one sample \mathcal{O}_1 , the dependencies that are introduced between the different steps of the algorithm make the analysis challenging.

5.3 Dependence on graph topologies

In this section, we discuss implications of our results for some comparison topologies. Let us focus first on the worst-case design setting, and the lower bound of Theorem 5.2.1. For the star, path (or more generally, any graph with bounded average degree), and complete bipartite graphs, one can verify that we have $\alpha(G) \asymp n$, so $\mathcal{M}(G, \mathbb{C}_{\text{NS}}) \asymp 1$. If the graph is a union of disjoint cliques $K_{n/2} \cup K_{n/2}$ (or having a constant number of edges across the cliques, like a barbell graph), then we see that $\beta(G^c) \asymp n^2$, so $\mathcal{M}(G, \mathbb{C}_{\text{NS}}) \asymp 1$. Thus, our theory yields pessimistic results for many practically motivated comparison topologies under worst-case designs, even though all the connected graphs above admit consistent estimation for parametric models³ as the number of samples grows. In the average case-setting of Section 5.1.2, Theorems 5.2.2, 5.2.3 and 5.2.4 characterize the mean-squared Frobenius norm errors of the corresponding estimators (up to constants) as $\mathcal{D}(G) := \frac{1}{n} \sum_{v \in V} \frac{1}{\sqrt{d_v}}$.

In order to illustrate our results for the average-case setting, we present the results of simulations on data generated synthetically⁴ from two special cases of the SST model. We fix $\pi^* = \text{id}$ without loss of generality, and generate the ground truth comparison matrix M^* in one of two ways:

- (1) Noisy sorting with high SNR: We set $M^* = M_{\text{NS}}(\text{id}, 0.4)$.
- (2) SST with independent bands: We first set $M_{ii}^* = 1/2$ for every i . Entries on the diagonal band immediately above the diagonal (i.e. $M_{i,i+1}^*$ for $i \in [n-1]$) are chosen i.i.d. and uniformly at random from the set $[1/2, 1]$. The band above is then chosen uniformly at random from the allowable set, where every entry is

³The complete bipartite graph, for instance, admits optimal rates of estimation.

⁴Note that the SST model has been validated extensively on real data in past work (see, e.g. Ballinger and Wilcox [BW97]).

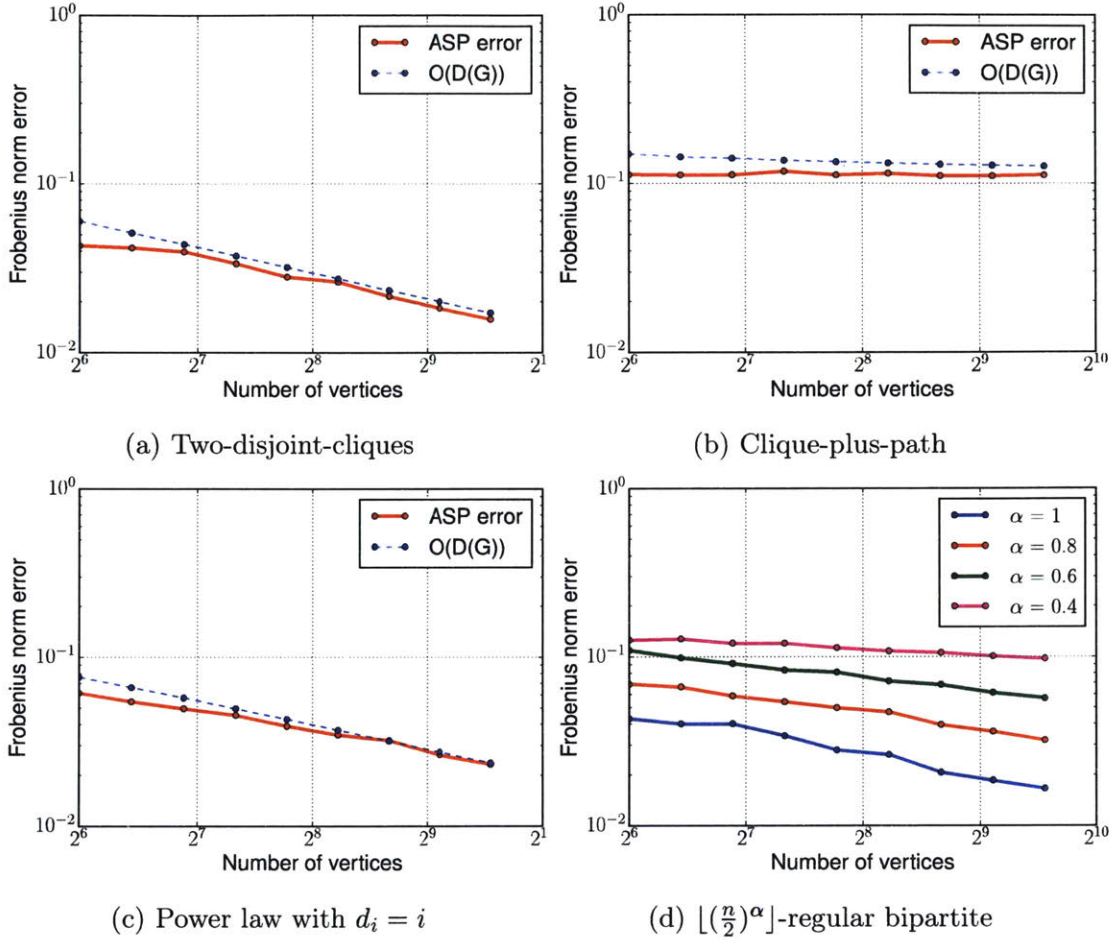


Figure 5-1: Normalized Frobenius norm error $\frac{1}{n^2} \|\widehat{M}_{\text{ASP}} - M^*\|_F^2$ with data generated using the noisy sorting model $M^* = M_{\text{NS}}(\text{id}, 0.4)$, averaged over 10 trials.

constrained to be upper bounded by 1 and lower bounded by the entries to its left and below. We also set $M_{ij}^* = 1 - M_{ji}^*$ to fill the rest of the matrix.

For each graph G with adjacency matrix A , the data is generated from ground truth by observing independent Bernoulli comparisons under the observation process $\mathcal{O} = \sigma(A)$, for a randomly generated permutation σ . For the SST model, we also generate data from two independent random observations \mathcal{O}_1 and \mathcal{O}_2 as required by the BAP estimator; however, we also simulate the behaviour of the estimator for one sample \mathcal{O}_1 and show that it closely tracks that of the two-sample estimator.

Recall that the estimation error rate was dictated by the degree functional $\mathcal{D}(G)$. While our graphs were chosen to illustrate scalings of $\mathcal{D}(G)$, some variants of these graphs also naturally arise as comparison topologies.

(1) **Two-disjoint-clique graph:** For this graph $K_{n/2} \cup K_{n/2}$, we have $d_v = \frac{n}{2} - 1$ for every $v \in V$, and simple calculations yield $\mathcal{D}(G) \asymp \frac{1}{\sqrt{n}}$. It is interesting to note that this graph has unfavorable guarantees for parametric estimation under the adversarial model, because it is disconnected (and thus has a Laplacian with zero spectral gap.)

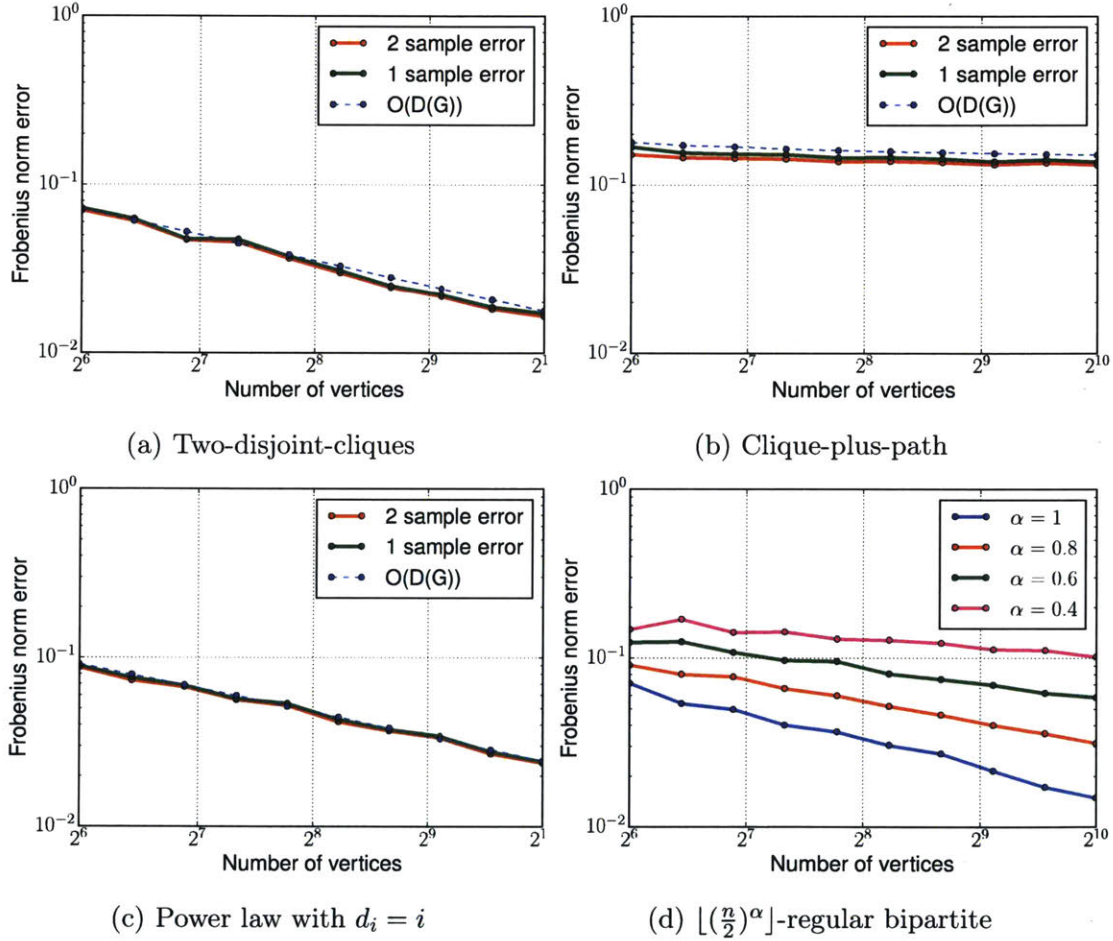


Figure 5-2: Normalized Frobenius norm error $\frac{1}{n^2} \|\widehat{M}_{\text{BAP}} - M^*\|_F^2$ with data generated using the SST model with independent bands, averaged over 10 trials, plotted for one and two samples.

We observe that this spectral property does not play a role in our analysis of the ASP or BAP estimator under the average-case observation model, and this behavior is corroborated by our simulations. Although we do not show it here, a similar behavior is observed for the stochastic block model, a practically motivated comparison topology when there are genres present among the items, which is a relaxation of the two-clique case allowing for sparser “communities” instead of cliques, and edges between the communities.

(2) **Clique-plus-path graph:** The nodes are partitioned into two sets of $n/2$ nodes each. The graph contains an edge between every two nodes in the first set, and a path starting from one of the nodes in the first set and chaining the other $n/2$ nodes. This is an example of a graph construction that has many ($\asymp n^2$) edges, but is unfavorable for noisy sorting or SST estimation. Simple calculations show that the degree functional is dominated by the constant degree terms and we obtain $\mathcal{D}(G) \asymp 1$.

(3) **Power law graph:** We consider the special power law graph [BA99] with degree sequence $d_i = i$ for $1 \leq i \leq n$, and construct it using the Havel-Hakimi algorithm [Hav55, Hak62]. For this graph, we have a disparate degree sequence, but $\mathcal{D}(G) \asymp \frac{1}{\sqrt{n}}$, and the simulated estimators are consistent.

(4) **$\lfloor (n/2)^\alpha \rfloor$ -regular bipartite graphs:** A final powerful illustration of our theoretical guarantees is provided by a regular bipartite graph construction in which the nodes are partitioned into two sets of $n/2$ nodes each, and each node in one set is (deterministically) connected to $\lfloor (n/2)^\alpha \rfloor$ nodes in the other set. This results in the degree sequence $d_v = \lfloor (n/2)^\alpha \rfloor$ for all $v \in V$, and the degree functional evaluates to $\mathcal{D}(G) \asymp n^{-\alpha/2}$. The value of α thus determines the scaling of the estimation error for the ASP estimator in the noisy sorting case, as well as the BAP estimator in the SST case, as seen from the slopes of the corresponding plots.

Some other graphs that were considered in parametric model environments by [SBB⁺16], such as the star, cycle, path and hypercube graphs, turn out to be unfavorable for permutation-based models even in the average-case setting, as corroborated by the lower bound of Theorem 5.2.3, part (b).

5.4 Proofs

In this section, we provide the proofs of our main results. We assume throughout that $n \geq 2$, and use c, c' to denote universal constants that may change from line to line.

5.4.1 Proof of Theorem 5.2.1

For each fixed graph G , define the quantity

$$\mathcal{A}(G) := \sup_{\substack{M, M' \in \mathbb{C}_{\text{NS}} \\ M(G) = M'(G)}} \frac{1}{n^2} \sum_{(i,j) \notin E} (M_{ij} - M'_{ij})^2$$

corresponding to the diameter quantity that is lower bounded in equation (5.5a). Taking the lower bound (5.5a) as given for the moment, we first prove the lower bound (5.5b) on the minimax risk. It suffices to show that the minimax risk is lower bounded in terms of $\mathcal{A}(G)$ as

$$\inf_{\widehat{M}=f(Y(G))} \sup_{M^* \in \mathbb{C}_{\text{NS}}} \mathbb{E} \left[\frac{1}{n^2} \|\widehat{M} - M^*\|_F^2 \right] \geq \frac{1}{4} \mathcal{A}(G). \quad (5.8)$$

In order to verify this claim, consider the two matrices $M^1, M^2 \in \mathbb{C}_{\text{SST}}$ that attain the supremum in the definition of $\mathcal{A}(G)$; note that such matrices exist due to the compactness of the space and the continuity of the squared loss. By construction, these two matrices satisfy the properties

$$M^1(G) = M^2(G), \quad \text{and} \quad \sum_{(i,j) \notin E} (M_{ij}^1 - M_{ij}^2)^2 = n^2 \mathcal{A}(G).$$

We can now reduce the problem to one of testing between the two matrices M^1 and M^2 , with the distribution of observations being identical for both alternatives. Consequently, any procedure can do no better than to make a random guess between the two, so we have

$$\inf_{\widehat{M}} \sup_{M^* \in \mathbb{C}_{\text{NS}}} \mathbb{E} \left[\|\widehat{M} - M^*\|_F^2 \right] \geq \frac{1}{4} \sum_{(i,j) \notin E} (M_{ij}^1 - M_{ij}^2)^2,$$

which proves the claim (5.8).

It remains to prove the claimed lower bound (5.5a) on $\mathcal{A}(G)$. This lower bound can be split into the following two claims:

$$\mathcal{A}(G) \geq \frac{1}{n^2} \alpha(G)(\alpha(G) - 1), \quad \text{and} \quad (5.9a)$$

$$\mathcal{A}(G) \geq \frac{1}{n^2} \beta(G^c). \quad (5.9b)$$

We use a different argument to establish each claim.

Proof of claim (5.9a): Recall the definition of the largest independent set. Without loss of generality, let the largest independent set be given by $I = \{v_1, \dots, v_\alpha\}$. Assign item i to vertex v_i for $i \in [\alpha]$. Now we choose permutations π and π' so that

- $\pi(i) = i$ for $i \in [\alpha]$,
- $\pi'(i) = \alpha - i + 1$ for $i \in [\alpha]$,
- π and π' agree on $\{\alpha + 1, \dots, n\}$.

Note that last step is possible because $\pi([\alpha]) = \pi'([\alpha])$. Moreover, define the matrices $M = M_{\text{NS}}(\pi, 1/2)$ and $M' = M_{\text{NS}}(\pi', 1/2)$. Note that by construction, we have

ensured that $M(G) = M'(G)$. However, it holds that

$$\sum_{(i,j) \notin E} (M_{ij} - M'_{ij})^2 = \|M - M'\|_F^2 = 2\text{KT}(\pi, \pi') = \alpha(\alpha - 1),$$

which completes the proof.

Proof of claim (5.9b): Recall the definition of a maximum biclique. Since the complement graph G^c has a biclique with $\beta(G^c)$ edges, the graph G has two disjoint sets of vertices V_1 and V_2 with $|V_1||V_2| = \beta(G^c)$ that do not have edges connecting one to the other. We now pick the two permutations π and π' so that

- the permutation π ranks items from V_1 as the top $|V_1|$ items, and ranks items from V_2 as the next $|V_2|$ items;
- the permutation π' ranks items from V_2 as the top $|V_2|$ items, and ranks items from V_1 as the next $|V_2|$ items;
- the permutations π and π' agree with each other apart from the above constraints.

As before, we define $M = M_{\text{NS}}(\pi, 1/2)$ and $M' = M_{\text{NS}}(\pi', 1/2)$, and again, we have $M(G) = M'(G)$. The relative orders of items have been interchanged across the biclique, so it holds that $2\text{KT}(\pi, \pi') = \beta(G^c)$, which completes the proof. \square

5.4.2 Some useful lemmas for average-case proofs

We now turn to proofs for the average-case setting. For convenience, we begin by stating two lemmas that are used in multiple proofs. The first lemma bounds the performance of the permutation estimator $\hat{\pi}_{\text{ASP}}$ for a general SST matrix, and is thus of independent interest.

Lemma 5.4.1. *For any matrix $M^* \in \mathbb{C}_{\text{SST}}$, the permutation estimator $\hat{\pi}_{\text{ASP}}$ satisfies*

$$\|\hat{\pi}_{\text{ASP}}(M^*) - M^*\|_F^2 \leq 4(n-1)\|\tau^* - \hat{\tau}\|_1, \quad (5.10a)$$

and if additionally, $M^* \in \mathbb{C}_{\text{NS}}(\lambda^*)$, we have

$$\|\hat{\pi}_{\text{ASP}}(M^*) - M^*\|_F^2 \leq 8\lambda^*(n-1)\|\tau^* - \hat{\tau}\|_1. \quad (5.10b)$$

In addition, the score estimates satisfy the bounds

$$\mathbb{E}\|\tau^* - \hat{\tau}\|_1 \leq c \sum_{v \in V} \frac{1}{\sqrt{d_v}}, \quad \text{and} \quad \Pr \left\{ \|\tau^* - \hat{\tau}\|_1 \geq c\sqrt{\log n} \sum_{v \in V} \frac{1}{\sqrt{d_v}} \right\} \leq n^{-10}.$$

Note that Lemma 5.4.1 implies the bound (5.6b), since for a matrix $M^* \in \mathbb{C}_{\text{NS}}(\lambda^*)$, we have $8\lambda^2\text{KT}(\hat{\pi}_{\text{ASP}}, \pi^*) = \|\hat{\pi}_{\text{ASP}}(M^*) - M^*\|_F^2$.

Our second lemma is a type of rearrangement inequality.

Lemma 5.4.2. *Let $\{a_u\}_{u=1}^n$ be an increasing sequence of positive numbers and let $\{b_u\}_{u=1}^n$ be a decreasing sequence of positive numbers. Then we have*

$$\left(\sum_{u=1}^n a_u\right)\left(\sum_{u=1}^n b_u\right) \geq n \sum_{u=1}^n a_u b_u.$$

Proof of Lemma 5.4.1

Assume without loss of generality that $\pi^* = \text{id}$. We begin by applying Hölder's inequality to obtain

$$\|\widehat{\pi}_{\text{ASP}}(M^*) - M^*\|_F^2 \leq \|\widehat{\pi}_{\text{ASP}}(M^*) - M^*\|_\infty \|\widehat{\pi}_{\text{ASP}}(M^*) - M^*\|_1.$$

In the case where $M^* \in \mathbb{C}_{\text{NS}}(\lambda^*)$, we have $\|M_{\widehat{\pi}_{\text{ASP}}(i)}^* - M_i^*\|_\infty \leq 2\lambda^*$; in the general case $M^* \in \mathbb{C}_{\text{SST}}$, we have $\|M_{\widehat{\pi}_{\text{ASP}}(i)}^* - M_i^*\|_\infty \leq 1$. Next, if $M_{\widehat{\pi}_{\text{ASP}}}^*$ denotes the matrix obtained from permuting the rows of M^* by $\widehat{\pi}_{\text{ASP}}$, then it holds that

$$\begin{aligned} \|\widehat{\pi}_{\text{ASP}}(M^*) - M^*\|_1 &\leq \|\widehat{\pi}_{\text{ASP}}(M^*) - M_{\widehat{\pi}_{\text{ASP}}}^*\|_1 + \|M_{\widehat{\pi}_{\text{ASP}}}^* - M^*\|_1 \\ &= 2 \sum_{i=1}^n \|M_{\widehat{\pi}_{\text{ASP}}(i)}^* - M_i^*\|_1, \end{aligned}$$

where the equality follows from the condition $M_{ij}^* + M_{ji}^* = 1$. We also have

$$\begin{aligned} \sum_{i=1}^n \|M_{\widehat{\pi}_{\text{ASP}}(i)}^* - M_i^*\|_1 &\stackrel{(i)}{=} (n-1) \sum_{i=1}^n |\tau_{\widehat{\pi}_{\text{ASP}}(i)}^* - \tau_i^*| \\ &= (n-1) \sum_{i=1}^n |\tau_i^* - \tau_{\widehat{\pi}_{\text{ASP}}^{-1}(i)}^*| \\ &\leq (n-1) \left[\sum_{i=1}^n |\tau_i^* - \widehat{\tau}_{\widehat{\pi}_{\text{ASP}}^{-1}(i)}| + \sum_{i=1}^n |\widehat{\tau}_{\widehat{\pi}_{\text{ASP}}^{-1}(i)} - \tau_{\widehat{\pi}_{\text{ASP}}^{-1}(i)}^*| \right] \\ &\stackrel{(ii)}{\leq} (n-1) \left[\sum_{i=1}^n |\tau_i^* - \widehat{\tau}_i| + \sum_{i=1}^n |\widehat{\tau}_i - \tau_i^*| \right] \\ &= 2(n-1) \|\tau^* - \widehat{\tau}\|_1, \end{aligned}$$

where step (i) is due to monotonicity along each column of M^* , and step (ii) follows from the ℓ_1 -rearrangement inequality (see, e.g., Example 2 in the paper [Vin90]), using the fact that both sequences $\{\tau_i^*\}_{i=1}^n$ and $\{\widehat{\tau}_{\widehat{\pi}_{\text{ASP}}^{-1}(i)}\}_{i=1}^n$ are sorted in decreasing order. Combining the last three displays yields the claimed bounds (5.10a) and (5.10b).

In order to prove the second part of the lemma, it suffices to show that the random variable $\|\tau^* - \widehat{\tau}\|_1$ is sub-Gaussian with parameter cS , where $S := \sum_{v \in V} 1/\sqrt{d_v}$. Let $\sigma : [n] \rightarrow V$ be the uniform random assignment of items to vertices with $\sigma(A) = \mathcal{O}$, and let D_i denote the random degree $d_{\sigma(i)} = \sum_{j \neq i} \mathcal{O}_{ij}$ of item i . Note that conditioned

on the event $\sigma(i) = v$, the difference between a score and its empirical version can be written as

$$\hat{\tau}_i - \tau_i^* = \left(\frac{1}{d_v} \sum_{j:\sigma(j)\sim v} M_{ij}^* - \frac{1}{n-1} \sum_{j\neq i} M_{ij}^* \right) + \frac{1}{d_v} \sum_{j:\sigma(j)\sim v} W_{ij},$$

where \sim denotes the presence of an edge between two vertices. Note that the term $\frac{1}{d_v} \sum_{j:\sigma(j)\sim v} M_{ij}^*$ is the empirical mean of d_v numbers chosen uniformly at random without replacement from the set $\{M_{ij}^*\}_{j\neq i}$, while $\frac{1}{n-1} \sum_{j\neq i} M_{ij}^*$ is the true expectation. Moreover, W_{ij} represents independent, zero-mean noise bounded within the interval $[-1, 1]$. Consequently, applying Hoeffding's inequality for sampling without replacement [BM15, Proposition 1.2] and the standard Hoeffding bound [Hoe63] to the two parts respectively, we obtain

$$\Pr \{ |\hat{\tau}_i - \tau_i^*| \geq t \mid \sigma(i) = v \} \leq 4 \exp(-c d_v t^2). \quad (5.11)$$

Replacing t by $t/\sqrt{d_v}$, we see that conditioned on the event $\sigma(i) = v$, the random variable $\sqrt{d_v}|\hat{\tau}_i - \tau_i^*|$ is sub-Gaussian with a constant parameter c' , or equivalently,

$$\mathbb{E} \left[\exp \left(t \sqrt{D_i} |\hat{\tau}_i - \tau_i^*| \right) \mid \sigma(i) = v \right] \leq \exp(ct^2). \quad (5.12)$$

Since $S = \sum_{i=1}^n 1/\sqrt{D_i}$, Jensen's inequality implies that

$$\begin{aligned} & \mathbb{E} \left[\exp \left(t \sum_{i=1}^n |\hat{\tau}_i - \tau_i^*| \right) \right] \\ & \leq \mathbb{E} \left[\sum_{i=1}^n \frac{1/\sqrt{D_i}}{S} \exp \left(t S \sqrt{D_i} |\hat{\tau}_i - \tau_i^*| \right) \right] \\ & = \sum_{i=1}^n \frac{1}{S} \sum_{v \in V} \Pr \{ \sigma(i) = v \} \mathbb{E} \left[\frac{1}{\sqrt{D_i}} \exp \left(t S \sqrt{D_i} |\hat{\tau}_i - \tau_i^*| \right) \mid \sigma(i) = v \right] \\ & \leq \sum_{i=1}^n \frac{1}{S} \sum_{v \in V} \frac{1}{n} \frac{1}{\sqrt{d_v}} \exp(cS^2 t^2) \\ & = \exp(cS^2 t^2), \end{aligned}$$

where the last inequality follows from equation (5.12). Therefore, the random variable $\|\hat{\tau} - \tau^*\|_1$ is sub-Gaussian with parameter cS , as claimed. \square

Proof of Lemma 5.4.2

For any increasing sequence $\{a_u\}$ and decreasing sequence $\{b_u\}$, the rearrangement inequality (see, e.g., Example 2 in the paper [Vin90]) guarantees that

$$\sum_{u=1}^n a_u b_u \leq \sum_{u=1}^n a_u b_{\pi(u)} \quad \text{for any permutation } \pi.$$

This inequality implies that

$$\frac{1}{n} \left(\sum_{u=1}^n a_u \right) \left(\sum_{u=1}^n b_u \right) = \frac{1}{n} \sum_{v=1}^n \sum_{u=1}^n a_u b_{\pi^{(v)}(u)} \geq \frac{1}{n} \sum_{v=1}^n \sum_{u=1}^n a_u b_u = \sum_{u=1}^n a_u b_u,$$

where we define $\pi^{(v)}(u) := (u+v) \bmod n$ and have used the rearrangement inequality for each of these permutations. \square

Equipped with these two lemmas, we are now ready to prove Theorem 5.2.2.

5.4.3 Proof of Theorem 5.2.2

Without loss of generality, reindexing as necessary, we may assume that the true permutation π^* is the identity id , thereby ensuring that $M^* = M_{\text{NS}}(\text{id}, \lambda^*)$. We begin by applying the triangle inequality to upper bound the error as a sum of two terms:

$$\frac{1}{2} \|\widehat{M}_{\text{ASP}} - M^*\|_F^2 \leq \underbrace{\|\widehat{M}_{\text{ASP}} - \widehat{\pi}_{\text{ASP}}(M^*)\|_F^2}_{\text{estimation error}} + \underbrace{\|\widehat{\pi}_{\text{ASP}}(M^*) - M^*\|_F^2}_{\text{approximation error}}.$$

Applying Lemma 5.4.1 yields bound on the approximation error. In particular, we have

$$\mathbb{E} [\|\widehat{\pi}_{\text{ASP}}(M^*) - M^*\|_F^2] \leq cn \sum_{v \in V} \frac{1}{\sqrt{d_v}}.$$

We now turn to the estimation error term, which evaluates to $n^2(\widehat{\lambda} - \lambda^*)^2$, with $\widehat{\lambda}$ representing the MLE of λ^* conditional on $\widehat{\pi}$ being the correct permutation. For each random set of edges E (we now let E be random in order to lighten notation) and permutation π , define the set

$$I_\pi(E) = \{(i, j) \in E \mid i < j, \pi(i) > \pi(j)\},$$

corresponding to the set of inversions that are also observed on the edge set E . We require that each ordered pair $(i, j) \in E$ obeys $i < j$. Therefore, the MLE takes the

form

$$\begin{aligned}
1/2 + \widehat{\lambda} &= \frac{1}{|E|} \left(\sum_{(i,j) \in E \setminus I_{\widehat{\pi}_{\text{ASP}}}(E)} Y_{ij} + \sum_{(i,j) \in I_{\widehat{\pi}_{\text{ASP}}}(E)} (1 - Y_{ij}) \right) \\
&= \frac{1}{|E|} \left(\sum_{(i,j) \in E} Y_{ij} + \sum_{(i,j) \in I_{\widehat{\pi}_{\text{ASP}}}(E)} (1 - 2Y_{ij}) \right) \\
&= 1/2 + \lambda^* + \frac{1}{|E|} \left(\sum_{(i,j) \in E} W_{ij} \right) + \frac{1}{|E|} \left(\sum_{(i,j) \in I_{\widehat{\pi}_{\text{ASP}}}(E)} -2\lambda^* - 2W_{ij} \right),
\end{aligned}$$

where we have written $Y_{ij} = M_{ij}^* + W_{ij}$. Consequently, the error obeys

$$\begin{aligned}
(\widehat{\lambda} - \lambda^*)^2 &\leq \frac{3}{|E|^2} \left(\sum_{(i,j) \in E} W_{ij} \right)^2 + \frac{12}{|E|^2} (\lambda^*)^2 |I_{\widehat{\pi}_{\text{ASP}}}(E)|^2 + \frac{12}{|E|^2} \left(\sum_{(i,j) \in I_{\widehat{\pi}_{\text{ASP}}}(E)} W_{ij} \right)^2 \\
&\stackrel{(i)}{\leq} \underbrace{\frac{3}{|E|^2} \left(\sum_{(i,j) \in E} W_{ij} \right)^2}_{T_1} + \underbrace{\frac{12}{|E|} (\lambda^*)^2 |I_{\widehat{\pi}_{\text{ASP}}}(E)|}_{T_2} + \underbrace{\frac{12}{|E|^2} \left(\sum_{(i,j) \in I_{\widehat{\pi}_{\text{ASP}}}(E)} W_{ij} \right)^2}_{T_3},
\end{aligned}$$

where step (i) follows since $|I_{\widehat{\pi}_{\text{ASP}}}(E)| \leq |E|$ pointwise. We now bound each of the terms T_1 , T_2 and T_3 separately. First, by standard sub-exponential tail bounds, and noting that $W_{ij} \in [-1, 1]$, we have

$$\mathbb{E}[T_1] \leq \frac{3}{|E|}, \quad \text{and} \quad \Pr \left\{ T_1 \geq \frac{6}{|E|} \right\} \leq e^{-|E|}.$$

We also have

$$\begin{aligned}
\frac{|E|}{12(\lambda^*)^2} \mathbb{E}[T_2] &= \mathbb{E}[|I_{\widehat{\pi}_{\text{ASP}}}(E)|] \\
&= \sum_{i < j} \sum_{(u,v) \in E} \Pr[\sigma(i) = u, \sigma(j) = v] \Pr[\widehat{\pi}_{\text{ASP}}(i) > \widehat{\pi}_{\text{ASP}}(j) | \sigma(i) = u, \sigma(j) = v] \\
&= \sum_{(u,v) \in E} \sum_{i < j} \frac{1}{n(n-1)} \Pr[\widehat{\pi}_{\text{ASP}}(i) > \widehat{\pi}_{\text{ASP}}(j) | \sigma(i) = u, \sigma(j) = v].
\end{aligned}$$

We now require the following lemma, which is proved at the end of this section.

Lemma 5.4.3. *For any pair of vertices $u \neq v$, we have*

$$\sum_{i < j} \frac{1}{n(n-1)} \Pr[\widehat{\pi}_{\text{ASP}}(i) > \widehat{\pi}_{\text{ASP}}(j) | \sigma(i) = u, \sigma(j) = v] \leq \frac{c}{\lambda^*} \left(\frac{1}{\sqrt{d_u}} + \frac{1}{\sqrt{d_v}} \right). \quad (5.13)$$

Using Lemma 5.4.3 in conjunction with our previous bounds yields

$$\mathbb{E}[T_2] \leq c \frac{\lambda^*}{|E|} \sum_{(u,v) \in E} \left(\frac{1}{\sqrt{d_u}} + \frac{1}{\sqrt{d_v}} \right) = c \lambda^* \frac{\sum_{u \in V} \sqrt{d_u}}{\sum_{u \in V} d_u}, \quad (5.14)$$

where the equality follows since each term $\frac{1}{\sqrt{d_u}}$ appears d_u times in the sum over all edges, and $2|E| = \sum_{u \in V} d_u$. Let $\{d_{(u)}\}_{u=1}^n$ represent the sequence of vertex degrees sorted in ascending order. An application of Lemma 5.4.2 with $a_u = d_{(u)}$ and $b_u = \frac{1}{\sqrt{d_{(u)}}}$ for $u \in [n]$ yields

$$\sum_{u \in V} \sqrt{d_u} \leq \frac{1}{n} \left(\sum_{u \in V} d_u \right) \left(\sum_{u \in V} \frac{1}{\sqrt{d_u}} \right).$$

Together with equation (5.14), we find that

$$\mathbb{E}[T_2] \leq \frac{c \lambda^*}{n} \sum_{u \in V} \frac{1}{\sqrt{d_u}}.$$

In order to complete the proof, it remains to bound $\mathbb{E}[T_3]$. Note that this step is non-trivial, since the noise terms W_{ij} for $(i, j) \in I_{\hat{\pi}_{\text{ASP}}}(E)$ depend on and are coupled through the data-dependent quantity $\hat{\pi}_{\text{ASP}}$. In order to circumvent this tricky dependency, consider some *fixed* permutation π , and let $T_3^\pi = \left(\sum_{(i,j) \in I_\pi(E)} W_{ij} \right)^2$. Note that T_3^π has two sources of randomness: randomness in the edge set E and randomness in observations. Since the observations $\{W_{ij}\}$ are independent and bounded and $|I_\pi(E)| \leq |E|$, the term

$$\sum_{(i,j) \in I_\pi(E)} W_{ij}$$

is sub-Gaussian with parameter at most $\sqrt{|E|}$. We then have the uniform sub-exponential tail bound

$$\Pr\{T_3^\pi \geq |E| + \delta\} \leq e^{-c\delta}. \quad (5.15)$$

Notice that for any $\alpha \in \mathbb{R}$, the inequality $T_3 \geq \alpha$ implies that the inequality $\frac{12}{|E|^2} T_3^\pi \geq \alpha$ holds for some *fixed* permutation π . Taking a union bound over all $n! \leq e^{n \log n}$ fixed permutations, and setting $\delta = cn \log n$ for a constant $c > 1$ yields

$$\Pr \left\{ T_3 \geq \frac{12}{|E|} + c \frac{n \log n}{|E|^2} \right\} \leq \exp \{n \log n - cn \log n\} \leq \exp \{-c'n \log n\}. \quad (5.16)$$

Noticing that $T_3 \leq 1$, we obtain

$$\begin{aligned}
\mathbb{E}[T_3] &\leq \Pr \left\{ T_3 \geq \frac{12}{|E|} + c \frac{n \log n}{|E|^2} \right\} \\
&\quad + \left(1 - \Pr \left\{ T_3 \geq \frac{12}{|E|} + c \frac{n \log n}{|E|^2} \right\} \right) \left(\frac{12}{|E|} + c \frac{n \log n}{|E|^2} \right) \\
&\leq \exp \{-c'n \log n\} + \frac{12}{|E|} + c \frac{n \log n}{|E|^2} \\
&\leq c' \left(\frac{1}{|E|} + \frac{n \log n}{|E|^2} \right).
\end{aligned}$$

Combining the pieces proves the claimed bound on the expectation. \square

The only remaining detail is to prove Lemma 5.4.3.

Proof of Lemma 5.4.3

We fix $i, j \in [n]$ with $i < j$ and condition on the event that $\sigma(i) = u$ and $\sigma(j) = v$ throughout the proof. First, note that the bound stated is trivially true if one of the vertices u or v has degree 1, by adjusting the constant appropriately. Hence, we assume for the rest of the proof that $d_u, d_v \geq 2$. Define the quantity

$$\tilde{\Delta}_{ji} = 2\lambda^* \frac{j-i-1}{n-2}. \quad (5.17)$$

We divide the rest of our analysis into two cases.

Case 1, $(u, v) \notin E(G)$: When the vertices u and v are not connected, we have

$$\begin{aligned}
\bar{\tau}_j &:= \mathbb{E}[\hat{\tau}_j] = \frac{1}{2} + \lambda^* \left(\frac{n-j}{n-2} - \frac{j-2}{n-2} \right) \text{ and} \\
\bar{\tau}_i &:= \mathbb{E}[\hat{\tau}_i] = \frac{1}{2} + \lambda^* \left(\frac{n-i-1}{n-2} - \frac{i-1}{n-2} \right),
\end{aligned}$$

and it can be verified that $\bar{\tau}_i - \bar{\tau}_j = \tilde{\Delta}_{ji}$. Consequently, we have

$$\begin{aligned}
&\Pr \{ \hat{\pi}_{\text{ASP}}(j) < \hat{\pi}_{\text{ASP}}(i) \mid \sigma(i) = u, \sigma(j) = v \} \\
&= \Pr \{ \hat{\tau}_j > \hat{\tau}_i \mid \sigma(i) = u, \sigma(j) = v \} \\
&\leq \Pr \left\{ |\hat{\tau}_j - \bar{\tau}_j| > \frac{\sqrt{d_u}}{\sqrt{d_v} + \sqrt{d_u}} \tilde{\Delta}_{ji} \mid \sigma(i) = u, \sigma(j) = v \right\} \\
&\quad + \Pr \left\{ |\hat{\tau}_i - \bar{\tau}_i| > \frac{\sqrt{d_v}}{\sqrt{d_v} + \sqrt{d_u}} \tilde{\Delta}_{ji} \mid \sigma(i) = u, \sigma(j) = v \right\} \\
&\leq 4 \exp \left\{ -c \frac{d_u d_v}{(\sqrt{d_u} + \sqrt{d_v})^2} \tilde{\Delta}_{ji}^2 \right\}, \quad (5.18)
\end{aligned}$$

where the last step follows from the Hoeffding bound for sampling without replacement in conjunction with the standard Hoeffding bound for bounded independent noise, by an argument similar to that of equation (5.11).

Case 2, $(u, v) \in E(G)$: When the vertices u and v are connected, we have

$$\begin{aligned}\bar{\tau}_j &:= \mathbb{E}[\widehat{\tau}_j] = \frac{1}{2} + \frac{d_v - 1}{d_v} \lambda^* \left(\frac{n-j}{n-2} - \frac{j-2}{n-2} \right) - \frac{1}{d_v} \lambda^* \text{ and} \\ \bar{\tau}_i &:= \mathbb{E}[\widehat{\tau}_i] = \frac{1}{2} + \frac{d_u - 1}{d_u} \lambda^* \left(\frac{n-i-1}{n-2} - \frac{i-1}{n-2} \right) + \frac{1}{d_u} \lambda^*,\end{aligned}$$

and it can be verified that $\bar{\tau}_i - \bar{\tau}_j \geq \widetilde{\Delta}_{ji}$.

Now, however, we must apply the Hoeffding bound for sampling without replacement to $d_u - 1$ and $d_v - 1$ random variables, respectively. Recalling that $d_u, d_v \geq 2$, we have

$$\begin{aligned}\Pr \{ \widehat{\pi}_{\text{ASP}}(j) < \widehat{\pi}_{\text{ASP}}(i) \mid \sigma(i) = u, \sigma(j) = v \} \\ &= \Pr \{ \widehat{\tau}_j > \widehat{\tau}_i \mid \sigma(i) = u, \sigma(j) = v \} \\ &\leq \Pr \left\{ |\widehat{\tau}_j - \bar{\tau}_j| > \frac{\sqrt{d_u}}{\sqrt{d_v} + \sqrt{d_u}} \widetilde{\Delta}_{ji} \mid \sigma(i) = u, \sigma(j) = v \right\} \\ &\quad + \Pr \left\{ |\widehat{\tau}_i - \bar{\tau}_i| > \frac{\sqrt{d_v}}{\sqrt{d_v} + \sqrt{d_u}} \widetilde{\Delta}_{ji} \mid \sigma(i) = u, \sigma(j) = v \right\} \\ &\leq 4 \exp \left\{ -c \frac{(d_u - 1)(d_v - 1)}{(\sqrt{d_u} - 1 + \sqrt{d_v} - 1)^2} \widetilde{\Delta}_{ji}^2 \right\} \\ &\leq 4 \exp \left\{ -c' \frac{d_u d_v}{(\sqrt{d_u} + \sqrt{d_v})^2} \widetilde{\Delta}_{ji}^2 \right\}.\end{aligned}\tag{5.19}$$

We use the shorthand L_{uv} to denote the LHS of equation (5.13). Having established the bounds (5.18) and (5.19), we now combine them to derive that

$$\begin{aligned}L_{uv} &\leq \frac{1}{n(n-1)} \sum_{j=2}^n \sum_{i < j} 4 \exp \left\{ -c \frac{d_u d_v}{(\sqrt{d_u} + \sqrt{d_v})^2} (j-i-1)^2 \frac{(\lambda^*)^2}{(n-2)^2} \right\} \\ &\leq \frac{4}{n(n-1)} (n-1) \sum_{m=1}^n \exp \left\{ -\frac{d_u d_v}{(\sqrt{d_u} + \sqrt{d_v})^2} m^2 \frac{(\lambda^*)^2}{(n-2)^2} \right\},\end{aligned}$$

where we have used $m = j - i$, and noted that there are at most $n - 1$ repetitions of each distinct value of $j - i$ in the sum over $j > i$.

Defining $\psi(q) = \sum_{m=1}^{\infty} q^{m^2}$, we recall the following theta function identity⁵ for $ab = \pi$ (see, for instance, equation (2.3) in Yi [Yi04]):

$$\sqrt{a} \left(1 + 2\psi(e^{-a^2}) \right) = \sqrt{b} \left(1 + 2\psi(e^{-b^2}) \right).$$

⁵For the rest of this subsection, π denotes the universal constant.

Using the identity by setting $a^2 = c \frac{d_u d_v}{(\sqrt{d_u} + \sqrt{d_v})^2} \frac{(\lambda^*)^2}{n^2}$ yields

$$\begin{aligned}
L_{uv} &\leq \frac{c}{n} \frac{n}{\lambda^*} \frac{\sqrt{d_u} + \sqrt{d_v}}{\sqrt{d_u d_v}} \left(1 + 2 \sum_{m=1}^{\infty} \exp \left\{ -\pi^2 \frac{(\sqrt{d_u} + \sqrt{d_v})^2}{d_u d_v} m^2 \frac{n^2}{(\lambda^*)^2} \right\} \right) \\
&\leq \frac{c}{\lambda^*} \frac{\sqrt{d_u} + \sqrt{d_v}}{\sqrt{d_u d_v}} \left(1 + 2 \sum_{m=1}^{\infty} \exp \left\{ -\pi^2 \frac{(\sqrt{d_u} + \sqrt{d_v})^2}{d_u d_v} m \frac{n^2}{(\lambda^*)^2} \right\} \right) \\
&\leq \frac{c}{\lambda^*} \frac{\sqrt{d_u} + \sqrt{d_v}}{\sqrt{d_u d_v}} \left(1 + \sum_{m=1}^{\infty} \exp \{ -16\pi^2 n m \} \right), \tag{5.20}
\end{aligned}$$

where in the last step, we have used the fact that $\lambda^* \leq 1/2$, and that $\frac{(\sqrt{d_u} + \sqrt{d_v})^2}{d_u d_v} \geq 4/n$. Bounding the geometric sum by a universal constant yields the required result.

5.4.4 Proof of Theorem 5.2.3

We prove the two parts of the theorem separately.

Proof of part (a)

The proof of part (a) is based on the following lemmas.

Lemma 5.4.4. *Consider a matrix of the form $M^* = M_{\text{NS}}(\pi^*, 1/4)$ where the permutation π^* is chosen uniformly at random. For any graph $G = K_1 \cup K_2 \cup \dots$ composed of multiple disjoint cliques with the number of vertices bounded as $C \leq |K_i| \leq n/5$ for all i , and for any estimators $(\widehat{M}, \widehat{\pi})$ that are measurable functions of the observations on G , we have*

$$\mathbb{E} \left[\frac{1}{n^2} \|\widehat{M} - M^*\|_F^2 \right] \geq \frac{c_2}{n} \sum_{v \in V} \frac{1}{\sqrt{d_v}}, \quad \text{and} \quad \mathbb{E} [\text{KT}(\pi^*, \widehat{\pi})] \geq c_2 n \sum_{v \in V} \frac{1}{\sqrt{d_v}}. \tag{5.21}$$

Lemma 5.4.5. *Given any graph G with degree sequence $\{d_v\}_{v \in V}$, there exists a graph G' consisting of multiple disjoint cliques with degree sequence $\{d'_v\}_{v \in V}$ such that*

$$|E| \asymp |E'| \quad \text{and} \quad \sum_{v \in V} \frac{1}{\sqrt{d_v}} \asymp \sum_{v \in V} \frac{1}{\sqrt{d'_v}}. \tag{5.22}$$

Part (a) follows by combining these two lemmas, so that it suffices to prove each of the lemmas individually.

Proof of Lemma 5.4.4: Our result is structural, and proved for permutation recovery. The bound for matrix recovery follows as a corollary. Assume we are given a graph on n vertices consisting of k disjoint cliques of sizes n_1, \dots, n_k . Let $N_0 = 0$ and $N_j = \sum_{i=1}^j n_i$ for $j \in [k]$. Without loss of generality, we let the j -th clique consist of the set of vertices V_j indexed by $\{N_{j-1} + 1, \dots, N_j\}$. By assumption, each n_j is upper bounded by $n/5$ and lower bounded by a universal constant.

Note that any estimator can only use the observations to construct the correct partial order within each clique, but not across cliques. We denote the induced partial order of a permutation π on the clique V_j by the permutation $\pi_j : [n_j] \rightarrow [n_j]$ ⁶. We will demonstrate that there exists a coupling of two marginally uniform random permutations $(\pi^*, \pi^\#)$ such that

$$\mathbb{E}[\text{KT}(\pi^*, \pi^\#)] \geq cn \sum_{j=1}^k \sqrt{n_j} = cn \sum_{v \in V} \frac{1}{\sqrt{d_v}},$$

and the partial order of π^* agrees with that of $\pi^\#$ on each clique, that is, $\pi_j^* = \pi_j^\#$ for all $j \in [k]$. Another way of stating this is that for every clique V_j and every two vertices $i_1, i_2 \in V_j$, we need that $\pi^\#(i_1) < \pi^\#(i_2)$ if and only if $\pi^*(i_1) < \pi^*(i_2)$.

Let $\mathbb{E}[\cdot \mid \pi^*]$ denote the expectation over the observations conditional on π^* . Given a pair of permutations $(\pi^*, \pi^\#)$ satisfying the above assumption, we view them as two hypotheses of the latent permutation. Then for any estimator $\hat{\pi}$, the Neyman-Pearson lemma [NP66] guarantees that

$$\mathbb{E}[\text{KT}(\hat{\pi}, \pi^*) \mid \pi^*] + \mathbb{E}[\text{KT}(\hat{\pi}, \pi^\#) \mid \pi^\#] \geq \text{KT}(\pi^\#, \pi^*)$$

for each instance of $(\pi^*, \pi^\#)$, because the observations are identical for π^* and $\pi^\#$. Taking expectation over $(\pi^*, \pi^\#)$, we obtain that

$$2 \mathbb{E}[\text{KT}(\hat{\pi}, \pi^*)] \geq \mathbb{E}[\text{KT}(\pi^*, \pi^\#)] \geq cn \sum_{v \in V} \frac{1}{\sqrt{d_v}}$$

since both π^* and $\pi^\#$ are marginally uniform.

To finish the proof, it remains to construct the required coupling $(\pi^*, \pi^\#)$. The construction is done as follows. First, permutations π^* and $\tilde{\pi}$ are generated uniformly at random and independently. Second, we sort the permutation $\tilde{\pi}$ on each clique according to π^* , and denote the resulting permutation by $\pi^\#$. Then the permutations π^* and $\pi^\#$ are marginally uniform and have common induced partial orders on the cliques, which we denote by $\{\pi_j^* : j \in [k]\}$.

With some extra notation, we can define the sorting step more formally for the interested reader. For a set of partial orders on the cliques $\{\pi_j : j \in [k]\}$, we define a special permutation that effectively orders vertices within each clique V_j according to its corresponding partial order π_j , but does not permute any vertices across cliques. We denote this special permutation by $\pi_{\text{par}}(\{\pi_j : j \in [k]\})$. For every clique V_j , we consider the permutation $\pi_{\text{sort},j} := \pi_j^* \circ (\tilde{\pi}_j)^{-1}$. Now, we can formally define the sorting step to generate $\pi^\#$ by

$$\pi^\# := \pi_{\text{par}}(\{\pi_{\text{sort},j} : j \in [k]\}) \circ \tilde{\pi}.$$

Next, we need to evaluate the expected Kendall's tau distance between these

⁶As an example, the identity permutation $\pi = \text{id}$ would yield $\pi_j = \text{id}$ on $[n_j]$ for all $j \in [k]$.

coupled permutations. By the tower property, we have

$$\mathbb{E}[\text{KT}(\pi^*, \pi^\#)] = \mathbb{E}[\mathbb{E}[\text{KT}(\pi^*, \pi^\#) \mid \{\pi_j^* : j \in [k]\}]].$$

The inner expectation can be simplified as follows. Pre-composing permutations π^* and $\pi^\#$ with any permutation does not change the Kendall's tau distance between them, so we have

$$\mathbb{E}[\text{KT}(\pi^*, \pi^\#) \mid \{\pi_j^* : j \in [k]\}] = \mathbb{E}[\text{KT}(\pi, \pi')]$$

where the permutations π and π' are drawn independently and uniformly at random from the set of permutations that are increasing on every clique. That is, for every clique V_j and every two vertices $i_1, i_2 \in V_j$, we have⁷ $\pi(i_1) < \pi(i_2)$ and $\pi'(i_1) < \pi'(i_2)$.

We now turn to computing the quantity $\mathbb{E}[\text{KT}(\pi, \pi')]$. It is well-known [DG77] that $2\text{KT}(\pi, \pi') \geq \|\pi - \pi'\|_1$. This fact together with Jensen's inequality implies that

$$\begin{aligned} 2\mathbb{E}[\text{KT}(\pi, \pi')] &\geq \sum_{i=1}^n \mathbb{E}[|\pi(i) - \pi'(i)|] \\ &\geq \sum_{i=1}^n \mathbb{E}\left[\left|\mathbb{E}[\pi(i) - \pi'(i) \mid \pi]\right|\right] \\ &= \sum_{i=1}^n \mathbb{E}\left[\left|\pi(i) - \mathbb{E}[\pi'(i)]\right|\right] \\ &= \mathbb{E}[\|\pi - \mathbb{E}[\pi]\|_1]. \end{aligned} \tag{5.23}$$

It therefore suffices to lower bound the quantity $\mathbb{E}[\|\pi - \mathbb{E}[\pi]\|_1]$.

Fix any $i \in [n]$. Then i is ℓ -th smallest index in the j -th clique for some $j \in [k]$ and $\ell \in [n_j]$, or succinctly, $i = N_{j-1} + \ell$. If we view π^{-1} as random draws from the n items, then $\pi(i)$ is equal to the number of draws needed to get the ℓ -th smallest element of V_j . Denoting $\mathbb{E}[\pi(i)]$ by μ , we have

$$\mu = \ell + \mathbb{E}\left[\sum_{r:\sigma(r) \notin V_j} \mathbf{1}\{r \text{ is drawn before } i\}\right] = \ell + (n - n_j) \frac{\ell}{n_j + 1} = \ell \frac{n + 1}{n_j + 1},$$

since the probability that an item not in V_j is drawn before the ℓ -th smallest element of V_j is $\ell/(n_j + 1)$. Furthermore, $\pi(i) = s$ if and only if $\ell - 1$ elements of V_j are selected in the first $s - 1$ draws and the s -th draw is from V_j , so

$$\Pr\{\pi(i) = s\} = \binom{n_j}{\ell - 1} \binom{n - n_j}{s - \ell} \binom{n}{s - 1}^{-1} \frac{n_j - \ell + 1}{n - s + 1}. \tag{5.24}$$

⁷To understand why π and π' can be chosen independently, note that the only dependency between the original permutations π^* and $\pi^\#$ is through the common induced partial orders $\{\pi_j^* : j \in [k]\}$. By conditioning and pre-composing, we are able to remove that dependency.

We claim that for all $\lceil 2n_j/5 \rceil \leq \ell \leq \lfloor 3n_j/5 \rfloor$ and $|s - \mu| \leq n/\sqrt{n_j}$, it holds that

$$\Pr\{\pi(i) = s\} \leq c\sqrt{n_j}/n \quad (5.25)$$

where c is a universal positive constant.

If the claim holds, then for any $0 \leq m \leq n/\sqrt{n_j}$, we have

$$\mathbb{E}[|\pi(i) - \mu|] \geq m \Pr\{|\pi(i) - \mu| \geq m\} \geq m[1 - c(2m + 1)\sqrt{n_j}/n]$$

by Markov's inequality. Choosing $m = \frac{n}{6c\sqrt{n_j}}$ yields

$$\mathbb{E}[|\pi(i) - \mu|] \geq c_2 n/\sqrt{n_j}$$

for some positive constant c_2 . Summing over ℓ in the given range, together with inequality (5.23), completes the proof.

Proof of claim (5.25): For $\ell \in [n_j]$ and $\ell \leq s \leq n - n_j + \ell$, define a bivariate function

$$p(\ell, s) := \binom{n_j}{\ell - 1} \binom{n - n_j}{s - \ell} \binom{n}{s - 1}^{-1}.$$

Note that for any fixed s , the function $\ell \mapsto p(\ell, s)$ is the probability mass function of the hypergeometric distribution that describes the probability of $\ell - 1$ successes in $s - 1$ draws without replacement from a population of size n with n_j successes. Hence, its maximum is attained at $\ell = \lfloor s \frac{n_j + 1}{n + 2} \rfloor$. Now we consider the index set

$$\begin{aligned} \mathcal{I} &= \left\{ (\ell, s) : \left\lceil \frac{n_j}{3} \right\rceil \leq \ell \leq \left\lceil \frac{2n_j}{3} \right\rceil, \left\lceil \frac{n_j}{3} \right\rceil \leq \left\lfloor s \frac{n_j + 1}{n + 2} \right\rfloor \leq \left\lceil \frac{2n_j}{3} \right\rceil \right\} \\ &\subset \left[\frac{n_j}{3}, \frac{2n_j}{3} \right] \times \left[\frac{n}{5}, \frac{4n}{5} \right]. \end{aligned}$$

In particular, the range of interest $\lceil 2n_j/5 \rceil \leq \ell \leq \lfloor 3n_j/5 \rfloor$ and $|s - \mu| \leq n/\sqrt{n_j}$, is contained within the set \mathcal{I} , since $\mu = \ell \frac{n_j + 1}{n_j + 1}$. Moreover, inequality (5.24) ensures that $\Pr\{\pi(i) = s\} \leq p(\ell, s) \frac{c_1 n_j}{n}$ for $(\ell, s) \in \mathcal{I}$. Thus, in order to complete the proof, it suffices to prove that $p(\ell, s) \leq c/\sqrt{n_j}$ for $(\ell, s) \in \mathcal{I}$, and it suffices to consider (ℓ, s) such that $\ell = \lfloor s \frac{n_j + 1}{n + 2} \rfloor$ since each function $\ell \mapsto p(\ell, s)$ attains its maximum at such a pair (ℓ, s) .

Toward this end, we use Stirling's approximation [DM56] to obtain

$$p(\ell, s) \leq c_2 \frac{\sqrt{n_j(n - n_j)(s - 1)(n - s + 1)}}{\sqrt{(\ell - 1)(n_j - \ell + 1)(s - \ell)(n - n_j - s + \ell)n}} \quad (5.26)$$

$$\frac{n_j^{n_j} (n - n_j)^{n - n_j} (s - 1)^{s - 1} (n - s + 1)^{n - s + 1}}{(\ell - 1)^{\ell - 1} (n_j - \ell + 1)^{n_j - \ell + 1} (s - \ell)^{s - \ell} (n - n_j - s + \ell)^{n - n_j - s + \ell} n^n}. \quad (5.27)$$

Since the factor in line (5.26) scales as $1/\sqrt{n_j}$ for $(\ell, s) \in \mathcal{I}$, it remains to bound the factor in line (5.27) by a universal constant. This follows from lengthy yet standard approximations which we briefly describe here. Assume that $s \frac{n_j+1}{n+2}$ is an integer for simplicity, so that ℓ is equal to this quantity and we have $s = \ell \frac{n+2}{n_j+1}$; the extension to the general case is easy. We first group together

$$\begin{aligned} \left[\frac{n_j(s-1)}{(\ell-1)n} \right]^{\ell-1} &= \left[\frac{n_j(n\ell + 2\ell - n_j - 1)/(n_j+1)}{(\ell-1)n} \right]^{\ell-1} \\ &= \left[1 + \frac{1 + (2\ell n_j - n_j^2 - n_j - \ell n)/(n_j n + n)}{\ell-1} \right]^{\ell-1}, \end{aligned}$$

which is bounded by a constant for $(\ell, s) \in \mathcal{I}$ considering that $\lim_{m \rightarrow \infty} (1 + \frac{a}{m})^m = e^a$. Then, we group together the terms

$$\left[\frac{n_j(n-s+1)}{(n_j-\ell+1)n} \right]^{n_j-\ell+1}, \left[\frac{(n-n_j)(s-1)}{(s-\ell)n} \right]^{s-\ell} \text{ and } \left[\frac{(n-n_j)(n-s+1)}{(n-n_j-s+\ell)n} \right]^{n-n_j-s+\ell}$$

respectively, and a similar argument yields that each term is bounded by a constant. \square

Proof of Lemma 5.4.5: Fix a graph G with degree sequence $\{d_v\}_{v \in V}$, and introduce the shorthand $S = \sum_{v \in V} 1/\sqrt{d_v}$. For some parameter k to be chosen, define the graph G' on the same vertex set to be the disjoint union of one clique of size $c_1 \lfloor \sqrt{|E|} \rfloor$, $c_2 k$ cliques of size $\lfloor n/k \rfloor$ and $c_3 S$ cliques of size 2, where c_1, c_2 and c_3 are constants to be determined such that the sizes of each clique are integers. The number of vertices remains the same, so that

$$n = c_1 \lfloor \sqrt{|E|} \rfloor + c_2 k \lfloor n/k \rfloor + 2c_3 S. \quad (5.28)$$

The number of edges of G' is

$$|E'| = \binom{c_1 \lfloor \sqrt{|E|} \rfloor}{2} + c_2 k \binom{\lfloor n/k \rfloor}{2} + c_3 S \asymp |E| + \frac{n^2}{k},$$

where the last approximation holds because $S \leq n \leq 2|E|$. Moreover, let

$$S' = \sum_{v \in V} \frac{1}{\sqrt{d'_v}} = \frac{c_1 \lfloor \sqrt{|E|} \rfloor}{\sqrt{c_1 \lfloor \sqrt{|E|} \rfloor - 1}} + \frac{c_2 k \lfloor n/k \rfloor}{\sqrt{\lfloor n/k \rfloor - 1}} + c_3 S \asymp \sqrt{nk} + S,$$

where the last approximation holds since $|E|^{1/4} \leq \sqrt{n} \leq S$.

In order to guarantee that $|E'| \asymp |E|$ and $S' \asymp S$, we need to choose an integer k so that $n^2/k \leq c|E|$ and $\sqrt{nk} \leq cS$, or equivalently

$$\frac{n^2}{c|E|} \leq k \leq c^2 \frac{S^2}{n}.$$

Such an integer k exists if $|E|S^2 \geq n^3$. Indeed, applying Lemma 5.4.2 twice (with $a_u = d_{(u)}$ and $b_u = 1/\sqrt{d_{(u)}}$ the first time and $a_u = \sqrt{d_{(u)}}$ and $b_u = 1/\sqrt{d_{(u)}}$ the second time, where $\{d_{(u)}\}_{u=1}^n$ is the degree sequence in ascending order), we obtain that

$$|E|S^2 = \left(\sum_{v \in V} d_v \right) \left(\sum_{v \in V} \frac{1}{\sqrt{d_v}} \right)^2 \geq n \left(\sum_{v \in V} \sqrt{d_v} \right) \left(\sum_{v \in V} \frac{1}{\sqrt{d_v}} \right) \geq n^3.$$

With k selected, it is easy to choose c_1, c_2 and c_3 so that inequality (5.28) holds, since each of $\sqrt{|E|}$, $k\lfloor n/k \rfloor$ and S is no larger than n . The issue of integrality can be taken care of by constant-order adjustment of these numbers, so the proof is complete. \square

Proof of part (b)

Given a parameter space Θ , a set $\mathcal{P} = \{\theta_1, \theta_2, \dots, \theta_{|\mathcal{P}|}\}$ is said to be a δ -packing in the metric ρ if $\rho(\theta_i, \theta_j) > \delta$ for all $i \neq j$. The lower bound of part (b) is based on the following packing lemma for the set of permutations in Kendall's tau distance. We note that a similar lemma was proved by Barg and Mazumdar [BM10].

Lemma 5.4.6. *For some positive constant c_1 , there exists an $c_1 n^2$ -packing \mathcal{P} of the set of permutations in the Kendall's tau distance such that $\log |\mathcal{P}| \geq n$.*

Consider the random observation model with graph $G = (V, E)$, where E denotes the random edge set of observations. We denote by \mathbb{Q}_M the law of the random observation noisy sorting model with underlying matrix $M = M_{\text{NS}}(\pi, \lambda)$. We require the following lemma.

Lemma 5.4.7. *Let $\mathbb{P}_{M,G}$ denote the law of the noisy sorting model with underlying matrix $M \in \mathbb{C}_{\text{NS}}(\lambda)$ for $\lambda \in [0, 1/4]$ and comparison graph G . Suppose that the entries of two matrices $M, M' \in \mathbb{C}_{\text{NS}}(\lambda)$ differ in s edges of the graph G . Then the KL divergence is bounded as*

$$\text{KL}(\mathbb{P}_{M,G}, \mathbb{P}_{M',G}) \leq 9\lambda^2 s. \quad (5.29)$$

Note that conditional on any instance of E , Lemma 5.4.7 guarantees that

$$\text{KL}(\mathbb{P}_{M,G}, \mathbb{P}_{M',G}) \leq 9\lambda^2 \left| \{(i, j) \in E : i < j, M_{i,j} \neq M'_{i,j}\} \right|,$$

where $\mathbb{P}_{M,G}$ denotes the model for fixed graph G . Hence taking expectation over the random edge set yields the upper bound

$$\text{KL}(\mathbb{Q}_M, \mathbb{Q}_{M'}) \leq 9\lambda^2 \sum_{i < j, M_{i,j} \neq M'_{i,j}} \Pr\{(i, j) \in E\} \leq 9\lambda^2 \sum_{i < j} \frac{2|E|}{n(n-1)} = 9\lambda^2 |E|,$$

valid for any $M, M' \in \mathbb{C}_{\text{NS}}(\lambda)$.

Note that $\|M - M'\|_F^2 = 8\lambda^2 \text{KT}(\pi, \pi')$ for $M = M_{\text{NS}}(\pi, \lambda)$ and $M' = M_{\text{NS}}(\pi', \lambda)$. Hence Fano's inequality applied to the packing given by Lemma 5.4.6 yields that

$$\inf_{\widehat{M}} \sup_{M^* \in \mathcal{C}_{\text{NS}}} \mathbb{E} \left[\|\widehat{M} - M^*\|_F^2 \right] \geq 8\lambda^2 c_1 n^2 \left(1 - \frac{9\lambda^2 |E| + \log 2}{n} \right).$$

The proof is completed by choosing $\lambda^2 = c_2 n / |E|$ for a sufficiently small constant c_2 . \square

It remains to prove Lemmas 5.4.6 and 5.4.7.

Proof of Lemma 5.4.6: The inversion table $b = (b_1, \dots, b_n)$ of a permutation π has entries defined by

$$b_i = \sum_{j=i+1}^n \mathbf{1}\{\pi(i) > \pi(j)\} \text{ for each } i \in [n].$$

We refer the reader to Mahmoud [Mah00] and references therein for background on inversion tables. By definition, we have $b_i \in \{0, 1, \dots, n - i\}$ and $\text{KT}(\pi, \text{id}) = \sum_{i=1}^n b_i$ where id denotes the identity permutation. In fact, the set of tables b satisfying $b_i \in \{0, 1, \dots, n - i\}$ is bijective to the set of permutations via this relation [Mah00]. This bijection aids in counting permutations with constraints.

Denote by $\mathcal{B}(\text{id}, r)$ the set of permutations that are within Kendall's tau distance r of the identity id . We seek an upper bound on $|\mathcal{B}(\text{id}, r)|$. Every $\pi \in \mathcal{B}(\text{id}, r)$ corresponds to an inversion table b such that $\sum_{i=1}^n b_i \leq r$. If b_i is only required to be a nonnegative integer, then the number of b satisfying $\sum_{i=1}^n b_i \leq r$ is bounded by $\binom{n+r}{n}$. After taking logarithms, this yields a bound

$$\log |\mathcal{B}(\text{id}, r)| \leq n \log(1 + r/n) + n.$$

Let \mathcal{P} be a maximal $c_1 n^2$ -packing of the set of permutations, which is necessarily also a $c_1 n^2$ -covering of that set. Then the family $\{\mathcal{B}(\pi, c_1 n^2)\}_{\pi \in \mathcal{P}}$ covers all permutations. By the right-invariance of the Kendall's tau distance under composition, the above bound yields $\log |\mathcal{B}(\pi, c_1 n^2)| \leq n \log(1 + c_1 n) + n$ for each π . Since there are $n!$ permutations in total, we conclude that $\log |\mathcal{P}| \geq \log(n!) - n \log(1 + c_1 n) - n \geq n$ for a sufficiently small constant c_1 . \square

Proof of Lemma 5.4.7

The KL divergence between Bernoulli observations has the form

$$\begin{aligned}
\text{KL}(\text{Ber}(1/2 + \lambda), \text{Ber}(1/2 - \lambda)) &= \text{KL}(\text{Ber}(1/2 - \lambda), \text{Ber}(1/2 + \lambda)) \\
&= (1/2 + \lambda) \log \frac{1/2 + \lambda}{1/2 - \lambda} + (1/2 - \lambda) \log \frac{1/2 - \lambda}{1/2 + \lambda} \\
&= 2\lambda \log \frac{1/2 + \lambda}{1/2 - \lambda} \\
&\leq 9\lambda^2 \quad \text{for all } \lambda \in [0, 1/4],
\end{aligned}$$

where the last inequality follows by some simple algebra,

Note that the KL divergence between a pair of product distributions is equal to the sum of the KL divergences between individual pairs. Since M and M' differ in s entries on the graph G and the Bernoulli observations are independent for different edges, we see that $\text{KL}(\mathbb{P}_{M,G}, \mathbb{P}_{M',G}) \leq 9\lambda^2 s$. \square

5.4.5 Proof of Theorem 5.2.4

For the purpose of the proof, it is helpful to think of the observation model in its linearized form. In particular, we have two random edge sets E_1 and E_2 and the observation matrices

$$Y_i := M^* + W_i$$

for each $i \in \{1, 2\}$. We also use the shorthand $\mathbf{B}(X, C) := \mathbf{B}(X, C, [n] \times [n])$, and recall the notation $\|M\|_B^2 := \sum_{(i,j) \in B} M_{ij}$.

By the triangle inequality, we have

$$\begin{aligned}
\|\widehat{M}_{\text{BAP}} - M^*\|_F^2 &\leq 2\|\widehat{M}_{\text{BAP}} - \widehat{\pi}_{\text{ASP}}(M^*)\|_F^2 + 2\|M^* - \widehat{\pi}_{\text{ASP}}(M^*)\|_F^2 \\
&\stackrel{(i)}{\leq} 2\|\widetilde{M} - \widehat{\pi}_{\text{ASP}}(M^*)\|_F^2 + 2\|M^* - \widehat{\pi}_{\text{ASP}}(M^*)\|_F^2 \\
&\leq 4\|\widetilde{M} - M^*\|_F^2 + 6\|M^* - \widehat{\pi}_{\text{ASP}}(M^*)\|_F^2,
\end{aligned} \tag{5.30}$$

where step (i) follows from the non-expansiveness of the projection operator. We know from Lemma 5.4.1 that the second term in inequality (5.30) is bounded in expectation by the quantity $nS = n \sum_{v \in V} 1/\sqrt{d_v}$ as desired, so it remains to bound the first term. Toward that end, again apply triangle inequality to write

$$\|\widetilde{M} - M^*\|_F^2 \leq 2\|\widetilde{M} - \mathbf{B}(M^*, \widehat{b})\|_F^2 + 2\|M^* - \mathbf{B}(M^*, \widehat{b})\|_F^2. \tag{5.31}$$

We now bound each of these terms separately. Starting with the first, let us define some notation. For a set $S \subseteq [n] \times [n]$ and a matrix $M \in \mathbb{R}^{n \times n}$, let $\|M\|_S^2 =$

$\sum_{(i,j) \in S} M_{ij}^2$. We have

$$\|\widetilde{M} - \mathbf{B}(M^*, \widehat{b})\|_F^2 = \sum_{B \in \mathcal{B}(\widehat{b})} \|\widetilde{M} - \mathbf{B}(M^*, \widehat{b})\|_B^2.$$

Note that it is sufficient to consider off diagonal blocks in the sum, since both \widetilde{M} and $\mathbf{B}(M^*, \widehat{b})$ are identically 1/2 in the diagonal blocks. Considering each block separately, we now split the analysis into two cases.

Case 1, $B \cap E_2 = \emptyset$: Because the entries of the error matrix are bounded within $[-1, 1]$, we have

$$\|\widetilde{M} - \mathbf{B}(M^*, \widehat{b})\|_B^2 \leq |B|.$$

Case 2, $B \cap E_2 \neq \emptyset$: Since both \widetilde{M} and $\mathbf{B}(M^*, \widehat{b})$ are constant on each block, we have

$$\begin{aligned} \|\widetilde{M} - \mathbf{B}(M^*, \widehat{b})\|_B^2 &= \frac{|B|}{|B \cap E_2|} \|\widetilde{M} - \mathbf{B}(M^*, \widehat{b})\|_{B \cap E_2}^2 \\ &= \frac{|B|}{|B \cap E_2|} \|\mathbf{B}(M^* + W_2, \widehat{b}, E_2) - \mathbf{B}(M^*, \widehat{b})\|_{B \cap E_2}^2 \\ &\leq 2 \frac{|B|}{|B \cap E_2|} \left(\|\mathbf{B}(M^* + W_2, \widehat{b}, E_2) - \mathbf{B}(\mathbf{B}(M^*, \widehat{b}) + W_2, \widehat{b}, E_2)\|_{B \cap E_2}^2 \right. \\ &\quad \left. + \|\mathbf{B}(\mathbf{B}(M^*, \widehat{b}) + W_2, \widehat{b}, E_2) - \mathbf{B}(M^*, \widehat{b})\|_{B \cap E_2}^2 \right). \end{aligned} \tag{5.32}$$

Let us handle each term on the RHS of the last inequality separately. First, by non-expansiveness of the projection operation defined by equation (5.7), we have

$$\|\mathbf{B}(M^* + W_2, \widehat{b}, E_2) - \mathbf{B}(\mathbf{B}(M^*, \widehat{b}) + W_2, \widehat{b}, E_2)\|_{B \cap E_2}^2 \leq \|M^* - \mathbf{B}(M^*, \widehat{b})\|_{B \cap E_2}^2. \tag{5.33}$$

We also require the following technical lemma:

Lemma 5.4.8. *For any block B and tuple $(i, j) \in B$, we have*

$$\Pr \left\{ (i, j) \in E_2 \mid |B \cap E_2| = k \right\} = \frac{k}{|B|}.$$

See Section 5.4.5 for the proof of this claim.

Returning to equation (5.33) and taking expectation over the randomness in E_2

(which, crucially, is independent of the randomness in \widehat{b}), we have

$$\begin{aligned}
\mathbb{E}_{E_2} \left[\left\| M^* - \mathbf{B}(M^*, \widehat{b}) \right\|_{B \cap E_2}^2 \mid |B \cap E_2| = k \right] \\
&= \sum_{(i,j) \in B} \Pr \left\{ (i,j) \in E_2 \mid |B \cap E_2| = k \right\} \cdot [M^* - \mathbf{B}(M^*, \widehat{b})]_{ij}^2 \\
&\stackrel{(ii)}{=} \sum_{(i,j) \in B} \frac{k}{|B|} [M^* - \mathbf{B}(M^*, \widehat{b})]_{ij}^2 \\
&= \frac{k}{|B|} \|M^* - \mathbf{B}(M^*, \widehat{b})\|_B^2,
\end{aligned} \tag{5.34}$$

where step (ii) follows from Lemma 5.4.8.

Additionally, notice that $[W_2]_{ij}$ for $(i,j) \in E_2$ is independent and bounded within the interval $[-1, 1]$. Consequently, we have

$$\mathbb{E}_{W_2} \left[\left\| \mathbf{B}(\mathbf{B}(M^*, \widehat{b}) + W_2, \widehat{b}, E_2) - \mathbf{B}(M^*, \widehat{b}) \right\|_{B \cap E_2}^2 \right] \leq 1, \tag{5.35}$$

where we have used the fact that the entries of the matrix $\mathbf{B}(M^*, \widehat{b})$ are constant on the set of indices $B \cap E_2$.

It follows from equations (5.32), (5.33), (5.34) and (5.35) that

$$\mathbb{E} \left[\left\| \widetilde{M} - \mathbf{B}(M^*, \widehat{b}) \right\|_B^2 \right] \leq 2\mathbb{E} \left[\frac{|B|}{|B \cap E_2|} \right] + 2\mathbb{E} \left[\left\| M^* - \mathbf{B}(M^*, \widehat{b}) \right\|_B^2 \right].$$

Combining the two cases and summing over the blocks, we obtain that

$$\mathbb{E} \left[\left\| \widetilde{M} - \mathbf{B}(M^*, \widehat{b}) \right\|_F^2 \right] \leq 2 \sum_{B \in \mathcal{B}(\widehat{b})} \mathbb{E} \left[\frac{|B|}{|B \cap E_2| \vee 1} \right] + 2\mathbb{E} \left[\left\| M^* - \mathbf{B}(M^*, \widehat{b}) \right\|_F^2 \right]. \tag{5.36}$$

Note that the second term above is the same as the second term on the RHS of inequality (5.31).

We now require the following definition, and two lemmas to complete the proof. Given a matrix M^* and a partition $C \in \chi_n$, define its row average as

$$[\mathbf{R}(M^*, C)]_i = \frac{1}{|C(i)|} \sum_{j \in C(i)} M_j^*.$$

Lemma 5.4.9. *With $S = \sum_{v \in V} 1/\sqrt{d_v}$ and for the partition $\widehat{b} = \text{bl}_t(r(Y'_1))$, we have*

$$\mathbb{E}_{E_2} \left[\sum_{B \in \mathcal{B}(\widehat{b})} \frac{|B|}{|B \cap E_2| \vee 1} \right] \leq nS.$$

Lemma 5.4.10. *Given any matrix $X \in [0, 1]^{n \times n}$ with monotone columns, a score vector $\hat{r} \in [0, n]^n$, and a value $t \in [0, n]$, we have*

$$\|X - R(X, \text{bl}_t(\hat{r}))\|_F^2 \leq nt + 2\|\hat{r} - r(X)\|_1.$$

Applying Lemma 5.4.9 with the expectation taken over the edge set E_2 yields the desired bound on the first term of inequality (5.36).

In order to bound the second term of inequality (5.36), note that by definition, we have

$$B(M^*, C) = R(R(M^*, C)^\top)^\top.$$

Consequently, it holds that

$$\begin{aligned} \|M^* - B(M^*, C)\|_F^2 &\leq 2\|M^* - R(M^*, C)\|_F^2 + 2\|R(M^*, C) - B(M^*, C)\|_F^2 \\ &= 2\|M^* - R(M^*, C)\|_F^2 + 2\|R(M^*, C)^\top - R(R(M^*, C)^\top, C)\|_F^2. \end{aligned}$$

Setting $C = \text{bl}_S(\hat{r})$ and applying Lemma 5.4.10 to both the terms, we obtain

$$\|M^* - B(M^*, \text{bl}_S(\hat{r}))\|_F^2 \leq 2nS + 4\|\hat{r} - r(M^*)\|_1.$$

Applying Lemma 5.4.1 yields a bound on the second term in expectation. This together with equations (5.31) and (5.36) completes the proof of Theorem 5.2.4 with the choice $t = \sum_{v \in V} 1/\sqrt{d_v}$.

It remains to prove Lemmas 5.4.8, 5.4.9 and 5.4.10.

Proof of Lemma 5.4.8

Our proof relies crucially on the fact that one of the two sets is a block.

For a fixed integer k , we condition on the event $\{|B \cap E_2| = k\}$. Note that E_2 is the random edge set defined by

$$E_2 = \pi(E) = \{(i, j) : (\pi(i), \pi(j)) \in E\},$$

where π is a uniform random permutation, and E is a fixed instance of E_2 . For any pair of tuples $(i, j), (k, \ell) \in B$, consider the permutation $\tilde{\pi}$ defined by

- $\tilde{\pi}(i) = k, \tilde{\pi}(k) = i, \tilde{\pi}(j) = \ell$ and $\tilde{\pi}(\ell) = j$;
- $\tilde{\pi}(m) = m$ for $m \neq i, j, k$ or ℓ .

Note that right-composition by $\tilde{\pi}$ is clearly a bijection between the sets $\{\pi : (i, j) \in \pi(E)\}$ and $\{\pi : (k, \ell) \in \pi(E)\}$. Therefore, we have $|\{\pi : (i, j) \in E_2\}| = |\{\pi : (k, \ell) \in E_2\}|$. A counting argument then completes the proof. Indeed, conditioned on the event $\{|B \cap E_2| = k\}$, we have

$$\sum_{(i,j) \in B} \Pr\{(i, j) \in E_2\} = \mathbb{E}\left[\sum_{(i,j) \in B} \mathbf{1}\{(i, j) \in E_2\} \right] = k,$$

which implies that $\Pr\{(i, j) \in E_2\} = \frac{k}{|B|}$.

Proof of Lemma 5.4.9

Fix an individual block B of dimensions $h \times w$, and let $E = E_2$ for notational convenience. Define the random variable $Y = |B \cap E| + 1$ so that $(|B \cap E| \vee 1)^{-1} \leq 2/Y$. Hence we require a bound on the quantity $\mathbb{E}[Y^{-1}]$. Toward this end, we write

$$Y = 1 + \sum_{(i,j) \in B} \mathbf{1}\{(i, j) \in E\}, \text{ and}$$

$$Y^2 = 1 + 2 \sum_{(i,j) \in B} \mathbf{1}\{(i, j) \in E\} + \sum_{(i,j), (i',j') \in B} \mathbf{1}\{(i, j), (i', j') \in E\}.$$

Note that for $(i, j), (i', j') \in B$ where $i \neq i'$ and $j \neq j'$, we have

$$\Pr\{(i, j) \in E\} = \frac{2|E|}{n(n-1)},$$

$$\Pr\{(i, j), (i, j') \in E\} = \frac{\sum_{v \in V} d_v(d_v - 1)}{n(n-1)(n-2)}, \text{ and}$$

$$\Pr\{(i, j), (i', j') \in E\} = \frac{4|E|^2 - 2 \sum_{v \in V} d_v(d_v - 1) - 2|E|}{n(n-1)(n-2)(n-3)}.$$

Hence, we can compute the first two moments of Y as

$$\mathbb{E}[Y] = 1 + \sum_{(i,j) \in B} \Pr\{(i, j) \in E\} = 1 + \frac{2hw|E|}{n(n-1)}, \text{ and}$$

$$\mathbb{E}[Y^2] = 1 + 2 \sum_{(i,j) \in B} \Pr\{(i, j) \in E\} + \sum_{(i,j), (i',j') \in B} \Pr\{(i, j), (i', j') \in E\}$$

$$= 1 + \frac{4hw|E|}{n(n-1)} + \frac{2hw|E|}{n(n-1)} + \left[hw(w-1) + wh(h-1) \right] \frac{\sum_{v \in V} d_v(d_v - 1)}{n(n-1)(n-2)}$$

$$+ h(h-1)w(w-1) \frac{4|E|^2 - 2 \sum_{v \in V} d_v(d_v - 1) - 2|E|}{n(n-1)(n-2)(n-3)}.$$

where for the last step we split into cases according to whether $i = i'$ or $j = j'$. Therefore, the variance $\text{var}(Y)$ is equal to

$$\mathbb{E}[Y^2] - \mathbb{E}[Y]^2 = \frac{2hw|E|}{n(n-1)} + \left[hw(w-1) + wh(h-1) \right] \frac{\sum_{v \in V} d_v(d_v - 1)}{n(n-1)(n-2)}$$

$$+ h(h-1)w(w-1) \frac{4|E|^2 - 2 \sum_{v \in V} d_v(d_v - 1) - 2|E|}{n(n-1)(n-2)(n-3)} - \frac{4h^2w^2|E|^2}{n^2(n-1)^2}.$$

We note that

$$\begin{aligned} \frac{h(h-1)w(w-1)}{n(n-1)(n-2)(n-3)} - \frac{h^2w^2}{n^2(n-1)^2} &= \frac{hw[hw(4n-6) - (h+w-1)n(n-1)]}{n(n-1)(n-2)(n-3)} \\ &\leq \frac{2h^2w^2}{n^2(n-1)^2(n-2)(n-3)}. \end{aligned}$$

where in the last step, we have used the fact that the quantity above is maximized when $h = w$, and that $2 \leq h + w \leq n$ by the construction of the blocks.

Combining the pieces, we conclude that $\text{var}(Y)$ is bounded by

$$c \frac{hw|E|}{n^2} + c(hw^2 + wh^2) \frac{\sum_{v \in V} d_v^2}{n^3} + c \frac{h^2w^2|E|^2}{n^6} \leq 2c \frac{hw|E|}{n^2} + c(hw^2 + wh^2) \frac{\sum_{v \in V} d_v^2}{n^3}$$

where the inequality holds because $h \leq n$, $w \leq n$ and $|E| \leq n^2$. Using the fact that $Y \geq 1$ and applying Chebyshev's inequality, we obtain

$$\begin{aligned} \mathbb{E}[Y^{-1}] &\leq \Pr \left\{ Y \leq \frac{\mathbb{E}[Y]}{2} \right\} + \frac{2}{\mathbb{E}[Y]} \\ &\leq \frac{4}{\mathbb{E}[Y]^2} \text{var}(Y) + \frac{2}{\mathbb{E}[Y]} \\ &\leq c \frac{n^4}{h^2w^2|E|^2} \left[\frac{hw|E|}{n^2} + (hw^2 + wh^2) \frac{\sum_{v \in V} d_v^2}{n^3} \right] + c \frac{n^2}{hw|E|} \\ &= 2c \frac{n^2}{hw|E|} + cn \frac{h+w}{hw} \frac{\sum_{v \in V} d_v^2}{|E|^2}. \end{aligned}$$

Now the above bound yields

$$\mathbb{E} \frac{|B|}{Y} \leq 2c \frac{n^2}{|E|} + cn(h+w) \frac{\sum_{v \in V} d_v^2}{|E|^2}.$$

Note that there are at most $m^2 = (n/S)^2$ blocks in total and the sum of h over $m-1$ off-diagonal blocks vertically is bounded by n (similarly for w). Thus we conclude that

$$\mathbb{E} \sum_{B \in \mathcal{B}(\delta)} \frac{|B|}{|B \cap E| \vee 1} \leq c \frac{m^2 n^2}{|E|} + cmn^2 \frac{\sum_{v \in V} d_v^2}{|E|^2}.$$

In order to complete the proof, it suffices to show that

$$\frac{n^2}{|E|} \left(\sum_{v \in V} \frac{1}{\sqrt{d_v}} \right)^{-2} + n \left(\sum_{v \in V} \frac{1}{\sqrt{d_v}} \right)^{-1} \frac{\sum_{v \in V} d_v^2}{|E|^2} \leq \frac{c}{n} \sum_{v \in V} \frac{1}{\sqrt{d_v}}.$$

Note that Lemma 5.4.2 implies that

$$2|E|\left(\sum_{v \in V} \frac{1}{\sqrt{d_v}}\right)^2 = \left(\sum_{v \in V} d_v\right)\left(\sum_{v \in V} \frac{1}{\sqrt{d_v}}\right)^2 \geq n^3.$$

It follows that

$$\frac{n^2}{|E|}\left(\sum_{v \in V} \frac{1}{\sqrt{d_v}}\right)^{-2} \leq \frac{2}{n} \leq \frac{2}{n} \sum_{v \in V} \frac{1}{\sqrt{d_v}},$$

and that

$$n\left(\sum_{v \in V} \frac{1}{\sqrt{d_v}}\right)^{-1} \frac{\sum_{v \in V} d_v^2}{|E|^2} \leq \frac{4}{n^2} \frac{\sum_{v \in V} d_v^2}{\sum_{v \in V} d_v} \left(\sum_{v \in V} \frac{1}{\sqrt{d_v}}\right) \leq \frac{4}{n} \sum_{v \in V} \frac{1}{\sqrt{d_v}}$$

since $d_v \leq n$.

Proof of Lemma 5.4.10

This lemma is a generalization of an approximation theorem due to Chatterjee [Cha15] and Shah et al. [SBGW17] to the noisy and two-dimensional setting.

We use the shorthand $\widehat{C}_t = \text{bl}_t(\widehat{\mathbf{r}})$ for the rest of the proof. Also define the set of placeholder elements in the partition \widehat{C}_t as

$$s(\widehat{C}_t) = \{i : i \text{ is smallest index in some set } I \in \widehat{C}_t\}.$$

We are now ready to prove the lemma. Begin by writing

$$\begin{aligned}
\|X - R(X, \widehat{C}_t)\|_F^2 &= \sum_{k=1}^n \left\| X_k - \frac{1}{|\widehat{C}_t(k)|} \sum_{j \in \widehat{C}_t(k)} X_j \right\|_2^2 \\
&\stackrel{(i)}{\leq} \sum_{k=1}^n \left\| X_k - \frac{1}{|\widehat{C}_t(k)|} \sum_{j \in \widehat{C}_t(k)} X_j \right\|_1 \\
&\stackrel{(ii)}{\leq} \sum_{k=1}^n \frac{1}{|\widehat{C}_t(k)|} \sum_{j \in \widehat{C}_t(k)} \|X_k - X_j\|_1 \\
&\stackrel{(iii)}{\leq} \sum_{k=1}^n \frac{1}{|\widehat{C}_t(k)|} \sum_{j \in \widehat{C}_t(k)} |r(X)_k - r(X)_j| \\
&= \sum_{k \in \mathfrak{S}(\widehat{C}_t)} \frac{1}{|\widehat{C}_t(k)|} \sum_{i \in \widehat{C}_t(k)} \sum_{j \in \widehat{C}_t(k)} |r(X)_i - r(X)_j| \\
&\leq \sum_{k \in \mathfrak{S}(\widehat{C}_t)} \frac{1}{|\widehat{C}_t(k)|} \sum_{i, j \in \widehat{C}_t(k)} (|\widehat{r}_i - r(X)_i| + |\widehat{r}_j - r(X)_j| + |\widehat{r}_i - \widehat{r}_j|) \\
&\stackrel{(iv)}{\leq} \|\widehat{r} - r(X)\|_1 + \|\widehat{r} - r(X)\|_1 + \sum_{k \in \mathfrak{S}(\widehat{C}_t)} t |\widehat{C}_t(k)| \\
&= 2\|\widehat{r} - r(X)\|_1 + nt.
\end{aligned}$$

Step (i) follows from the fact that each entry of the difference matrix $X - R(X, \widehat{C}_t)$ is bounded in the interval $[-1, 1]$; step (ii) follows from Jensen's inequality and convexity of the ℓ_1 norm; step (iii) uses the fact that for fixed k and j , the quantity $X_{k\ell} - X_{j\ell}$ has the same sign for all $\ell \in [n]$ due to the monotonicity of columns of the matrix X ; step (iv) uses the property of the blocking partition \widehat{C}_t , which ensures that $|\widehat{r}_i - \widehat{r}_j| \leq t$ when the inclusion $i, j \in \widehat{C}_t(k)$ is satisfied for some k . This completes the proof.

5.5 Discussion

In this chapter, we studied the problem of estimating the comparison probabilities from noisy pairwise comparisons under worst-case and average-case design assumptions. We exhibited a dichotomy between worst-case and average-case models for permutation-based models, which suggests that a similar distinction may exist even for their parametric counterparts. Our bounds leave a few interesting questions unresolved: Is there a sharp characterization of the diameter $\mathcal{A}(G)$ quantifying the approximation error of a comparison topology G ? The Borda count estimator, a variant of which we analyzed, is known to achieve a sub-optimal rate in the case of full observations; the estimator of Braverman and Mossel [BM08] achieves the optimal rate over the noisy sorting class. What is the analog of such an estimator in the average-case setting with partial pairwise comparisons? Is there a computa-

tional lower bound to show that our estimators are the best possible polynomial-time algorithms for SST matrix estimation in the average-case setting?

5.6 Appendix: Bounds on the minimax denoising error

As we saw in Theorem 5.2.1, the minimax risk of Frobenius norm estimation is prohibitively large for many comparison topologies. In some applications, however, it may be of interest to control the denoising error, which is the error we make on the observations seen on the edges of the graph. Accordingly, we define the quantity

$$\mathcal{E}(G, \mathbb{C}) = \inf_{\widehat{M}=f(Y(G))} \sup_{M^* \in \mathbb{C}} \mathbb{E} \left[\frac{1}{|E|} \|\widehat{M} - M^*\|_E^2 \right],$$

where we have used a normalization of $|E|$ to provide an average entry-wise bound on the denoising error. The following theorem provides bounds on the minimax denoising error for fixed topologies.

Theorem 5.6.1. *For any connected graph G , we have*

$$\mathcal{E}(G, \mathbb{C}_{\text{NS}}) \geq \frac{c_1}{|E|} \max_{S \in \mathcal{C}_G} \frac{|V(S)|^2}{|E(S)|}, \quad \text{and} \quad \mathcal{E}(G, \mathbb{C}_{\text{SST}}) \leq \frac{c_2 n \log^2 n}{|E|}. \quad (5.37)$$

Again, the lower bound on the error of the noisy sorting class provides a lower bound for the SST class. Conversely, the upper bound on the error for the SST class upper bounds the error for the noisy sorting class.

For many graphs used in practice, the lower bound can be evaluated to show that Theorem 5.6.1 provides a sharp characterization of the denoising error up to logarithmic factors.

The upper bound is obtained by the least squares estimator

$$\widehat{M}_{\text{LS}} = \arg \min_{\widehat{M} \in \mathbb{C}_{\text{SST}}} \|Y - \widehat{M}\|_E^2.$$

While we do not know yet whether such an estimator is computable in polynomial time, analyzing it provides a notion of the fundamental limits of the problem. In particular, it is clear that the denoising problem is easier than Frobenius norm estimation, and we obtain consistent rates provided that the number of edges in the graph satisfies $|E| = \omega(n \log^2 n)$.

5.6.1 Proof of Theorem 5.6.1

In this section, we prove Theorem 5.6.1 on the denoising error rate of the problem, splitting it into proofs of the lower and upper bounds.

Proof of lower bound

In order to prove the lower bound, we construct a suitable local packing \mathcal{P} of the parameter space \mathbb{C}_{NS} , and then apply Fano's inequality. For simpler presentation, we describe the packing \mathcal{P} by gradually putting constraints on its members. First, every matrix in \mathcal{P} is chosen to be $M_{\text{NS}}(\pi, \lambda)$ for a fixed λ and some permutation π , so we focus on selecting the permutations π .

Consider any connected subgraph $S \in \mathcal{C}_G$ with at least two vertices. Let the vertices of S form the top $|V(S)|$ items and choose the same ranking for the vertices of S^c for each instance in the packing. Then all the matrices in the packing \mathcal{P} have the same (i, j) -th entry if $i \in S^c$ or $j \in S^c$. Hence the KL divergence between any two models with underlying matrices in the packing \mathcal{P} is bounded by $9\lambda^2|E(S)|$, by Lemma 5.4.7.

Next, fix a spanning tree $T(S)$ of S which has $|V(S)| - 1$ edges. Note that all the $2^{|V(S)|-1}$ assignments of values to these edges

$$\{M_{ij} : (i, j) \in T(S), i < j\} \in \{1/2 + \lambda, 1/2 - \lambda\}^{|V(S)|-1}$$

are possible, since there are no cycle conflicts in the spanning tree. Using the Gilbert-Varshamov bound, we are guaranteed that there are constants a and b such that at least $2^{a|V(S)|}$ such assignments are separated pairwise by $b|V(S)|$ in the Hamming distance. We choose the packing \mathcal{P} consisting of matrices corresponding to these assignments, so that $\|M - M'\|_F^2 \geq 8b\lambda^2|V(S)|$ for any distinct $M, M' \in \mathcal{P}$.

Finally, Fano's inequality implies that

$$|E| \mathcal{E}(G, \mathbb{C}_{\text{NS}}) \geq 8b\lambda^2|V(S)| \left(1 - \frac{9\lambda^2|E(S)| + \log 2}{a|V(S)|}\right).$$

The proof then follows by choosing $\lambda^2 = c \frac{|V(S)|}{|E(S)|}$, for a sufficiently small constant c . \square

Proof of upper bound

As mentioned before, we obtain the upper bound by considering the estimator \widehat{M}_{LS} . The proof follows from previous results on the full observation case [SBGW17], but we provide it for completeness. Note that for each $(i, j) \in E$, the observation model takes the form

$$Y_{ij} = M_{ij}^* + W_{ij},$$

where W_{ij} is a zero-mean noise variable lying in the interval $[-1, 1]$.

The optimality of \widehat{M}_{LS} and feasibility of M^* imply that we must have the basic inequality $\|Y - \widehat{M}_{\text{LS}}\|_E^2 \leq \|Y - M^*\|_E^2$, which after simplification, leads to

$$\frac{1}{2}\|\Delta\|_E^2 \leq \langle \Delta, W \rangle_E, \quad (5.38)$$

where $\Delta = \widehat{M}_{\text{LS}} - M^*$, and $\langle A, B \rangle_E = \sum_{(i,j) \in E} A_{ij} B_{ij}$ denotes the trace inner product

restricted to the indices in E .

In order to establish the upper bound, we first define the class of difference matrices $\mathbb{C}_{\text{DIFF}} := \{M - M' \mid M, M' \in \mathbb{C}_{\text{SST}}\}$, as well as the associated random variable

$$Z(t) := \sup_{D \in \mathbb{C}_{\text{DIFF}}: \|D\|_E \leq t} \langle D, W \rangle_E.$$

With this notation, inequality (5.38) implies $\frac{1}{2} \|\Delta\|_E^2 \leq Z(\|\Delta\|_E)$. It follows from the star-shaped property⁸ of the set \mathbb{C}_{DIFF} that the following critical inequality is satisfied for some $\delta > 0$:

$$\mathbb{E}[Z(\delta)] \leq \frac{\delta^2}{2}.$$

We are interested in the smallest such value δ . In order to find it, we use Dudley's entropy integral, for which we require a bound on the covering number of the class \mathbb{C}_{DIFF} . Such a bound was calculated for the Frobenius norm by Shah et al. [SBGW17] using the results of Gao and Wellner [GW07]. Clearly, since $\|M_i - M_j\|_E^2 \leq \|M_i - M_j\|_F^2$, a δ -covering in the Frobenius norm automatically serves as a δ -covering in the edge norm $\|\cdot\|_E$. Thus, we have the following lemma.

Lemma 5.6.2. [SBGW17] *For every $\epsilon > 0$, we have the metric entropy bound*

$$\log N(\epsilon, \mathbb{C}_{\text{DIFF}}, \|\cdot\|_E) \leq \log N(\epsilon, \mathbb{C}_{\text{DIFF}}, \|\cdot\|_F) \leq 9 \frac{n^2}{\epsilon^2} \left(\log \frac{n}{\epsilon}\right)^2 + 9n \log n.$$

Dudley's entropy integral then yields that for all $t > 0$, we have

$$\begin{aligned} \mathbb{E}[Z(t)] &\leq c \inf_{\delta \in [0, n]} \left\{ n\delta + \int_{\delta/2}^t \sqrt{\log N(\epsilon, \mathbb{C}_{\text{DIFF}} \cap \mathbb{B}_E(t), \|\cdot\|_E)} d\epsilon \right\} \\ &\leq c \left\{ n^{-8} + \int_{n^{-9/2}}^t \sqrt{\log N(\epsilon, \mathbb{C}_{\text{DIFF}}, \|\cdot\|_E)} d\epsilon \right\}. \end{aligned}$$

After some algebra (for details, see Shah et al. [SBGW17]), we have

$$\mathbb{E}[Z(t)] \leq c \{ n \log^2 n + t \sqrt{n \log n} \}.$$

Setting $t = c\sqrt{n} \log n$ completes the proof. □

⁸A set S is said to be star-shaped if $t \in S$ implies that $\alpha t \in S$ for all $\alpha \in [0, 1]$

Bibliography

- [AAAK17] Arpit Agarwal, Shivani Agarwal, Sepehr Assadi, and Sanjeev Khanna. Learning with limited rounds of adaptivity: Coin tossing, multi-armed bandits, and ranking from pairwise comparisons. In *Conference on Learning Theory*, pages 39–75, 2017.
- [ABE⁺55] Miriam Ayer, H. D. Brunk, G. M. Ewing, W. T. Reid, and Edward Silverman. An empirical distribution function for sampling with incomplete information. *Ann. Math. Statist.*, 26(4):641–647, December 1955.
- [ABG⁺79] N. N. Anuchina, K. I. Babenko, S. K. Godunov, N. A. Dmitriev, L. V. Dmitrieva, V. F. D'yachenko, A. V. Zabrodin, O. V. Lokutsievskii, E. V. Malinovskaya, I. F. Podlivaev, G. P. Prokopov, I. D. Sofronov, and R. P. Fedorenko. *Teoreticheskie osnovy i konstruirovaniye chislennykh algoritmov zadach matematicheskoi fiziki*. “Nauka”, Moscow, 1979.
- [ABH] Emmanuel Abbe, Afonso S. Bandeira, and Georgina Hall. Exact recovery in the stochastic block model. 62(1):471–487.
- [ABH98] Jonathan E. Atkins, Erik G. Boman, and Bruce Hendrickson. A spectral algorithm for seriation and the consecutive ones problem. *SIAM Journal on Computing*, 28(1):297–310, 1998.
- [ACC13] Edo M. Airolidi, Thiago B Costa, and Stanley H Chan. Stochastic block-model approximation of a graphon: Theory and consistent estimation. In *Advances in Neural Information Processing Systems 26*, pages 692–700. 2013.
- [ACN08] Nir Ailon, Moses Charikar, and Alantha Newman. Aggregating inconsistent information: Ranking and clustering. *J. ACM*, 55(5):1–27, November 2008.
- [AG89] Richard Arratia and Louis Gordon. Tutorial on large deviations for the binomial distribution. *Bull. Math. Biol.*, 51(1):125–131, 1989.
- [Aga16] Shivani Agarwal. On ranking and choice models. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI’16*, pages 4050–4053. AAAI Press, 2016.

- [Alo06] Noga Alon. Ranking tournaments. *SIAM J. Discret. Math.*, 20(1):137–142, January 2006.
- [AM07] Dimitris Achlioptas and Frank McSherry. Fast computation of low-rank matrix approximations. *J. ACM*, 54(2), April 2007.
- [AS98] Fred Annexstein and Ram Swaminathan. On testing consecutive-ones property in parallel. *Discrete Appl. Math.*, 88(1-3):7–28, 1998.
- [AS15] Emmanuel Abbe and Colin Sandon. Community detection in general stochastic block models: Fundamental limits and efficient algorithms for recovery. In *Foundations of Computer Science (FOCS), 2015 IEEE 56th Annual Symposium on*, pages 670–688. IEEE, 2015.
- [ASDH88] P. Arabie, S. Schleutermann, J. Daws, and L. Hubert. Marketing applications of sequencing and partitioning of nonsymmetric and/or two-mode matrices. In Wolfgang Gaul and Martin Schader, editors, *Data, Expert Knowledge and Decisions: An Interdisciplinary Approach with Emphasis on Marketing Applications*, pages 215–224. Springer Berlin Heidelberg, Berlin, Heidelberg, 1988.
- [BA99] Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999.
- [Bar03] William Barnett. The modern theory of consumer behavior: Ordinal or cardinal? *Quarterly Journal of Austrian Economics*, 6(1):41–65, 2003.
- [BBBB72] R. E. Barlow, D. J. Bartholomew, J. M. Bremner, and H. D. Brunk. *Statistical inference under order restrictions. The theory and application of isotonic regression*. John Wiley & Sons, London-New York-Sydney, 1972.
- [BC09] Peter J. Bickel and Aiyou Chen. A nonparametric view of network models and newman–girvan and other modularities. *Proceedings of the National Academy of Sciences*, 106(50):21068–21073, 2009.
- [BDPR84] Gordon Bril, Richard Dykstra, Carolyn Pillers, and Tim Robertson. Algorithm AS 206: isotonic regression in two independent variables. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 33(3):352–357, 1984.
- [Bel15] Pierre C. Bellec. Sharp oracle inequalities for least squares estimators in shape restricted regression. *arXiv preprint arXiv:1510.08029*, 2015.
- [Bel16] Pierre C. Bellec. Private communication. 2016.
- [BF96] Peter J. Bickel and Jianqing Fan. Some problems on the estimation of unimodal densities. *Statist. Sinica*, 6(1), 1996.

- [Bir97] L. Birgé. Estimation of unimodal densities without smoothness assumptions. *Ann. Statist.*, 25(3), 1997.
- [BLM13] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.
- [BM08] Mark Braverman and Elchanan Mossel. Noisy sorting without resampling. In *Proceedings of the Nineteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 268–276. ACM, New York, 2008.
- [BM09] Mark Braverman and Elchanan Mossel. Sorting from noisy information. *arXiv preprint arXiv:0910.1191*, 2009.
- [BM10] Alexander Barg and Arya Mazumdar. Codes in permutations and error correction for rank modulation. *IEEE Transactions on Information Theory*, 56(7):3158–3165, 2010.
- [BM15] Rémi Bardenet and Odalric-Ambrym Maillard. Concentration inequalities for sampling without replacement. *Bernoulli*, 21(3):1361–1385, 2015.
- [BMI06] Victor Boyarshinov and Malik Magdon-Ismail. Linear time isotonic and unimodal regression in the L_1 and L_∞ norms. *J. Discrete Algorithms*, 4(4), 2006.
- [BMR10] Linas Baltrunas, Tadas Makcinskas, and Francesco Ricci. Group recommendations with rank aggregation and collaborative filtering. In *Proceedings of the Fourth ACM Conference on Recommender Systems, RecSys '10*, pages 119–126. ACM, 2010.
- [BR13] Quentin Berthet and Philippe Rigollet. Complexity theoretic lower bounds for sparse principal component detection. In Shai Shalev-Shwartz and Ingo Steinwart, editors, *COLT 2013 - The 26th Conference on Learning Theory, Princeton, NJ, June 12-14, 2013*, volume 30 of *JMLR W&CP*, pages 1046–1066, 2013.
- [BRS16] Quentin Berthet, Philippe Rigollet, and Piyush Srivastava. Exact recovery in the ising blockmodel. *arXiv preprint arXiv:1612.03880*, 2016.
- [BS67] Mikhail Shlemovich Birman and Mikhail Zakharovich Solomyak. Piecewise polynomial approximations of functions of classes W_p^α . *Mat. Sb. (N.S.)*, 73 (115):331–355, 1967.
- [BS95] Avrim Blum and Joel Spencer. Coloring random and semi-random k -colorable graphs. *J. Algorithms*, 19(2):204–234, 1995.

- [BS98] Rasmus Bro and Nikolaos D. Sidiropoulos. Least squares algorithms under unimodality and non-negativity constraints. *J. Chemometrics*, 12:223–247, 1998.
- [BT52] Ralph A. Bradley and Milton E. Terry. Rank analysis of incomplete block designs. I. The method of paired comparisons. *Biometrika*, 39:324–345, 1952.
- [BT15] Pierre C. Bellec and Alexandre B. Tsybakov. Sharp oracle bounds for monotone and convex regression through aggregation. *Journal of Machine Learning Research*, 16:1879–1892, 2015.
- [Bur13] Rainer E. Burkard. Quadratic assignment problems. *Handbook of combinatorial optimization*, pages 2741–2814, 2013.
- [BW97] T. Parker Ballinger and Nathaniel T. Wilcox. Decisions, error and heterogeneity. *The Economic Journal*, 107(443):1090–1105, 1997.
- [Cat12] Manuela Cattelan. Models for paired comparison data: A review with emphasis on dependent data. *Statistical Science*, pages 412–433, 2012.
- [CBCTH13] Xi Chen, Paul N. Bennett, Kevyn Collins-Thompson, and Eric Horvitz. Pairwise ranking aggregation in a crowdsourced setting. In *Proceedings of the sixth ACM international conference on Web search and data mining*, pages 193–202. ACM, 2013.
- [CD16] Olivier Collier and Arnak S. Dalalyan. Minimax rates in permutation estimation for feature matching. *Journal of Machine Learning Research*, 17(6):1–31, 2016.
- [CFSV04] Donatello Conte, Pasquale Foggia, Carlo Sansone, and Mario Vento. Thirty years of graph matching in pattern recognition. *International journal of pattern recognition and artificial intelligence*, 18(03):265–298, 2004.
- [CGMS17] Xi Chen, Sivakanth Gopi, Jieming Mao, and Jon Schneider. Competitive analysis of the top-k ranking problem. In *Proceedings of the Twenty-Eighth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1245–1264. SIAM, 2017.
- [CGS15] Sabyasachi Chatterjee, Adityanand Guntuboyina, and Bodhisattva Sen. On risk bounds in isotonic and other shape restricted regression problems. *Annals of Statistics*, 43(4):1774–1800, 2015.
- [CGS18] Sabyasachi Chatterjee, Adityanand Guntuboyina, and Bodhisattva Sen. On matrix estimation under monotonicity constraints. *Bernoulli*, (2):1072–1100, May 2018.

- [Cha14] Sourav Chatterjee. A new perspective on least squares under convex constraint. *Ann. Statist.*, 42(6):2340–2381, December 2014.
- [Cha15] Sourav Chatterjee. Matrix estimation by universal singular value thresholding. *Ann. Statist.*, 43(1):177–214, 2015.
- [CL15] Sabyasachi Chatterjee and John Lafferty. Adaptive risk bounds in unimodal regression. *arXiv preprint arXiv:1512.02956*, 2015.
- [CLR17a] T. Tony Cai, Tengyuan Liang, and Alexander Rakhlin. Weighted message passing and minimum energy flow for heterogeneous stochastic block models with side information. *arXiv preprint arXiv:1709.03907*, 2017.
- [CLR17b] Tony Cai, Tengyuan Liang, and Alexander Rakhlin. On detection and structural reconstruction of small-world random networks. *IEEE Transactions on Network Science and Engineering*, 4(3):165–176, 2017.
- [CM16] Sabyasachi Chatterjee and Sumit Mukherjee. On estimation in tournaments and graphs under monotonicity constraints. *arXiv preprint arXiv:1603.04556*, 2016.
- [CN91] Andrew Caplin and Barry Nalebuff. Aggregation and social choice: A mean voter theorem. *Econometrica*, 59(1):1–23, 1991.
- [Cop51] A. H. Copeland. A reasonable social welfare function. In *Mimeographed notes from a Seminar on Applications of Mathematics to the Social Sciences*, University of Michigan, 1951.
- [CP11] Emmanuel J Candes and Yaniv Plan. Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements. *IEEE Transactions on Information Theory*, 57(4):2342–2359, 2011.
- [CR09] Emmanuel J Candès and Benjamin Recht. Exact matrix completion via convex optimization. *Foundations of Computational mathematics*, 9(6):717, 2009.
- [CRPW12] Venkat Chandrasekaran, Benjamin Recht, Pablo A. Parrilo, and Alan S. Willsky. The convex geometry of linear inverse problems. *Found. Comput. Math.*, 12(6):805–849, 2012.
- [CS15] Yuxin Chen and Changho Suh. Spectral MLE: Top-k rank aggregation from pairwise comparisons. In *International Conference on Machine Learning*, pages 371–380, 2015.
- [Cze09] Jan Czekanowski. Zur differential diagnose der neandertalgruppe. korrespondenzblatt der deutschen gesellschaft für anthropologie. *Ethnologie und Urgeschichte*, 40:44–47, 1909.

- [DA14] Michael B. McCoy Joel A. Tropp Dennis Amelunxen, Martin Lotz. Living on the edge: phase transitions in convex programs with random data. *Information and Inference*, 2014.
- [DDKR13] Nilesh Dalvi, Anirban Dasgupta, Ravi Kumar, and Vibhor Rastogi. Aggregating crowdsourced binary ratings. In *Proceedings of the 22nd international conference on World Wide Web*, pages 285–294. ACM, 2013.
- [DDS12] Constantinos Daskalakis, Ilias Diakonikolas, and Rocco A. Servedio. Learning k-modal distributions via testing. In *Proceedings of the Twenty-third Annual ACM-SIAM Symposium on Discrete Algorithms, SODA '12*, pages 1371–1385. Society for Industrial and Applied Mathematics, 2012.
- [DDS+13] Constantinos Daskalakis, Ilias Diakonikolas, Rocco A. Servedio, Gregory Valiant, and Paul Valiant. Testing k-modal distributions: Optimal algorithms via reductions. In *Proceedings of the Twenty-Fourth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA '13*, pages 1833–1852. Society for Industrial and Applied Mathematics, 2013.
- [DG77] Persi Diaconis and Ronald L. Graham. Spearman’s footrule as a measure of disarray. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 262–268, 1977.
- [DKNS01] Cynthia Dwork, Ravi Kumar, Moni Naor, and Dandapani Sivakumar. Rank aggregation methods for the web. In *Proceedings of the 10th international conference on World Wide Web*, pages 613–622. ACM, 2001.
- [DM56] Abraham De Moivre. *The doctrine of chances: or, A method of calculating the probabilities of events in play*, volume 1. Chelsea Publishing Company, 1756.
- [DM59] Donald Davidson and Jacob Marschak. Experimental tests of a stochastic decision theory. *Measurement: Definitions and theories*, 1959.
- [Don90] David L. Donoho. Gel’fand n -widths and the method of least squares. Statistics Technical Report 282, University of California, Berkeley, December 1990.
- [DS79] Alexander Philip Dawid and Allan M. Skene. Maximum likelihood estimation of observer error-rates using the EM algorithm. *Applied Statistics*, pages 20–28, 1979.
- [EL00] P. P. B. Eggermont and V. N. LaRiccia. Maximum likelihood estimation of smooth monotone and unimodal densities. *Ann. Statist.*, 28(3), 2000.

- [ESBB98] Michael B. Eisen, Paul T. Spellman, Patrick O. Brown, and David Botstein. Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences*, 95(25):14863–14868, 1998.
- [ESDS16] Frank Emmert-Streib, Matthias Dehmer, and Yongtang Shi. Fifty years of graph matching, network alignment and network comparison. *Information Sciences*, 346:180–197, 2016.
- [Faz02] Maryam Fazel. *Matrix rank minimization with applications*. PhD thesis, 2002.
- [Fel68] William Feller. *An Introduction to Probability Theory and Its Applications*, volume 1. Wiley, 1968.
- [FG64] Delbert Ray Fulkerson and Oliver Alfred Gross. Incidence matrices with the consecutive 1’s property. *Bull. Amer. Math. Soc.*, 70:681–684, 1964.
- [Fis73] Peter C. Fishburn. Binary choice probabilities: on the varieties of stochastic transitivity. *Journal of Mathematical psychology*, 10(4):327–352, 1973.
- [FJBd13] Fajwel Fogel, Rodolphe Jenatton, Francis Bach, and Alexandre d’Aspremont. Convex relaxations for permutation problems. In *Advances in Neural Information Processing Systems 26*, pages 1016–1024. 2013.
- [FK46] Elaine Forsyth and Leo Katz. A matrix approach to the analysis of sociometric data: Preliminary report. *Sociometry*, 9(4):340–347, 1946.
- [FMR16] Nicolas Flammarion, Cheng Mao, and Philippe Rigollet. Optimal rates of statistical seriation. *arXiv preprint arXiv:1607.02435*, 2016.
- [FOPS17] Moein Falahatgar, Alon Orlitsky, Venkatadheeraj Pichapati, and Ananda Theertha Suresh. Maximum selection and ranking under noisy comparisons. *arXiv preprint arXiv:1705.05366*, 2017.
- [FQRM+16] Soheil Feizi, Gerald Quon, Mariana Recamonde-Mendoza, Muriel Medard, Manolis Kellis, and Ali Jadbabaie. Spectral alignment of graphs. *arXiv preprint arXiv:1602.04181*, 2016.
- [Fri86] M. Frisen. Unimodal regression. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 35(4):479–485, 1986.
- [FV93] Michael A. Fligner and Joseph S. Verducci. *Probability models and statistical analyses for ranking data*, volume 80. Springer, 1993.

- [Gao17] Chao Gao. Phase transitions in approximate ranking. *arXiv preprint arXiv:1711.11189*, 2017.
- [GG12] Thomas L. Gertzen and Martin Grötschel. Flinders Petrie, the traveling salesman problem, and the beginning of mathematical modeling in archaeology. *Doc. Math.*, X(Extra volume: Optimization stories):199–210, 2012.
- [Gil52] Edgar N. Gilbert. A comparison of signalling alphabets. *Bell Labs Technical Journal*, 31(3):504–522, 1952.
- [GKM11] Arpita Ghosh, Satyen Kale, and Preston McAfee. Who moderates the moderators?: Crowdsourcing abuse detection in user-generated content. In *Proceedings of the 12th ACM conference on Electronic commerce*, pages 167–176. ACM, 2011.
- [GLZ15] Chao Gao, Yu Lu, and Harrison H. Zhou. Rate-optimal graphon estimation. *Ann. Statist.*, 43(6):2624–2652, December 2015.
- [GLZ16] Chao Gao, Yu Lu, and Dengyong Zhou. Exact exponent in optimal rates for crowdsourcing. In *International Conference on Machine Learning*, pages 603–611, 2016.
- [GS90] Zhi Geng and Ning-Zhong Shi. Algorithm as 257: Isotonic regression for umbrella orderings. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 39(3):397–402, 1990.
- [GW07] Fuchang Gao and Jon A. Wellner. Entropy estimate for high-dimensional monotonic functions. *Journal of Multivariate Analysis*, 98(9):1751–1764, 2007.
- [GZ13] Chao Gao and Dengyong Zhou. Minimax optimal convergence rates for estimating ground truth from crowdsourced labels. *arXiv preprint arXiv:1310.5764*, 2013.
- [GZFA10] Anna Goldenberg, Alice X. Zheng, Stephen E. Fienberg, and Edoardo M. Airoldi. A survey of statistical network models. *Found. Trends Mach. Learn.*, 2(2):129–233, February 2010.
- [Hak62] S. Louis Hakimi. On realizability of a set of integers as degrees of the vertices of a linear graph. i. *Journal of the Society for Industrial and Applied Mathematics*, 10(3):496–506, 1962.
- [Har72] John A. Hartigan. Direct clustering of a data matrix. *Journal of the American Statistical Association*, 67:123–129, 1972.
- [Hav55] Václav Havel. A remark on the existence of finite graphs. *Casopis Pest. Mat.*, 80:477–480, 1955.

- [HKZ12] Daniel Hsu, Sham Kakade, and Tong Zhang. A tail inequality for quadratic forms of subgaussian random vectors. *Electron. Commun. Probab.*, 17, 2012.
- [HMG06] Ralf Herbrich, Tom Minka, and Thore Graepel. Trueskill: a Bayesian skill rating system. In *Proceedings of the 19th International Conference on Neural Information Processing Systems*, pages 569–576. MIT Press, 2006.
- [Hoe63] Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American statistical association*, 58(301):13–30, 1963.
- [HOX14] Bruce Hajek, Sewoong Oh, and Jiaming Xu. Minimax-optimal inference from partial rankings. In *Advances in Neural Information Processing Systems*, pages 1475–1483, 2014.
- [HR16] Jan-Christian Hütter and Philippe Rigollet. Optimal rates for total variation denoising. In *Conference on Learning Theory*, pages 1115–1146, 2016.
- [HSRW16] Reinhard Heckel, Nihar B. Shah, Kannan Ramchandran, and Martin J. Wainwright. Active ranking from pairwise comparisons and when parametric assumptions don’t help. *arXiv preprint arXiv:1606.08842*, 2016.
- [Hun04] David R. Hunter. MM algorithms for generalized Bradley-Terry models. *Annals of Statistics*, pages 384–406, 2004.
- [HWX16] Bruce Hajek, Yihong Wu, and Jiaming Xu. Information limits for recovering a hidden community. In *Information Theory (ISIT), 2016 IEEE International Symposium on*, pages 1894–1898. IEEE, 2016.
- [JKSO16] Minje Jang, Sunghyun Kim, Changho Suh, and Sewoong Oh. Top- k ranking from pairwise comparisons: When spectral ranking is optimal. *arXiv preprint arXiv:1603.04153*, 2016.
- [JN11] Kevin G. Jamieson and Robert D. Nowak. Active ranking using pairwise comparisons. In *Advances in Neural Information Processing Systems*, pages 2240–2248, 2011.
- [JNS13] Prateek Jain, Praneeth Netrapalli, and Sujay Sanghavi. Low-rank matrix completion using alternating minimization. In *Proceedings of the Forty-fifth Annual ACM Symposium on Theory of Computing, STOC ’13*, pages 665–674, New York, NY, USA, 2013. ACM.
- [KBI14] Claudia Köllmann, Björn Bornkamp, and Katja Ickstadt. Unimodal regression using Bernstein-Schoenberg splines and penalties. *Biometrics*, 70(4), 2014.

- [Ken48] Maurice G. Kendall. *Rank correlation methods*. Charles Griffin and Company, London, 1948.
- [Ken63] David G. Kendall. A statistical approach to Flinders Petrie’s sequence-dating. *Bull. Inst. Internat. Statist.*, 40:657–681, 1963.
- [Ken69] David G. Kendall. Incidence matrices, interval graphs and seriation in archeology. *Pacific J. Math.*, 28:565–570, 1969.
- [Ken70] David G. Kendall. A mathematical approach to seriation. *Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences*, 269(1193):125–134, 1970.
- [Ken71] David G. Kendall. Abundance matrices and seriation in archaeology. *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete*, 17:104–112, 1971.
- [KLT11] Vladimir Koltchinskii, Karim Lounici, and Alexandre B. Tsybakov. Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *The Annals of Statistics*, 39(5):2302–2329, 2011.
- [KMO10] Raghunandan H. Keshavan, Andrea Montanari, and Sewoong Oh. Matrix completion from a few entries. *IEEE Transactions on Information Theory*, 56(6):2980–2998, 2010.
- [KMS07] Claire Kenyon-Mathieu and Warren Schudy. How to rank with few errors. In *Proceedings of the thirty-ninth annual ACM symposium on Theory of computing*, pages 95–103. ACM, 2007.
- [Knu98] Donald E. Knuth. *The art of computer programming. Vol. 3*. Addison-Wesley, Reading, MA, 1998.
- [KO16] Ashish Khetan and Sewoong Oh. Data-driven rank breaking for efficient rank aggregation. *Journal of Machine Learning Research*, 17(193):1–54, 2016.
- [Kol11] Vladimir Koltchinskii. Von neumann entropy penalization and low-rank matrix estimation. *Ann. Statist.*, 39(6):2936–2973, 12 2011.
- [KOS11a] David R Karger, Sewoong Oh, and Devavrat Shah. Budget-optimal crowdsourcing using low-rank matrix approximations. In *Communication, Control, and Computing (Allerton), 2011 49th Annual Allerton Conference on*, pages 284–291. IEEE, 2011.
- [KOS11b] David R Karger, Sewoong Oh, and Devavrat Shah. Iterative learning for reliable crowdsourcing systems. In *Advances in neural information processing systems*, pages 1953–1961, 2011.
- [KR05] Thierry Klein and Emmanuel Rio. Concentration around the mean for maxima of empirical processes. *Ann. Probab.*, 33(3):1060–1077, May 2005.

- [KRS15] Rasmus Kyng, Anup Rao, and Sushant Sachdeva. Fast, provable algorithms for isotonic regression in all l_p -norms. In *Advances in Neural Information Processing Systems*, pages 2719–2727, 2015.
- [KTT15] Franz J Király, Louis Theran, and Ryota Tomioka. The algebraic combinatorial approach for low-rank matrix completion. *The Journal of Machine Learning Research*, 16(1):1391–1436, 2015.
- [KTV17] Olga Klopp, Alexandre B. Tsybakov, and Nicolas Verzelen. Oracle inequalities for network models and sparse graphon estimation. *Ann. Statist.*, 45(1):316–354, 02 2017.
- [LC86] Lucien Le Cam. *Asymptotic methods in statistical decision theory*. Springer Series in Statistics. Springer-Verlag, New York, 1986.
- [LdABN⁺07] Eliane Maria Loiola, Nair Maria Maia de Abreu, Paulo Oswaldo Boaventura-Netto, Peter Hahn, and Tania Querido. A survey for the quadratic assignment problem. *European Journal of Operational Research*, 176(2):657–690, 2007.
- [Lii10] Innar Liiv. Seriation and matrix reordering methods: An historical overview. *Statistical Analysis and Data Mining*, 3(2):70–91, 2010.
- [Liu09] Tie-Yan Liu. Learning to rank for information retrieval. *Foundations and Trends® in Information Retrieval*, 3(3):225–331, 2009.
- [LJTKJ06] Simon Lacoste-Julien, Ben Taskar, Dan Klein, and Michael I. Jordan. Word alignment via quadratic assignment. In *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, pages 112–119. Association for Computational Linguistics, 2006.
- [Lov12] László Lovász. *Large networks and graph limits*, volume 60. American Mathematical Soc., 2012.
- [LPI12] Qiang Liu, Jian Peng, and Alexander T Ihler. Variational inference for crowdsourcing. In *Advances in neural information processing systems*, pages 692–700, 2012.
- [LR15] Jing Lei and Alessandro Rinaldo. Consistency of spectral clustering in stochastic block models. *The Annals of Statistics*, 43(1):215–237, 2015.
- [LS17] Christina E. Lee and Devavrat Shah. Unifying framework for crowdsourcing via graphon estimation. *arXiv preprint arXiv:1703.08085*, 2017.
- [LT91] Michel Ledoux and Michel Talagrand. *Probability in Banach spaces*, volume 23 of *Ergebnisse der Mathematik und ihrer Grenzgebiete (3) [Results in Mathematics and Related Areas (3)]*. Springer-Verlag, Berlin, 1991. Isoperimetry and processes.

- [Luc59] R. Duncan Luce. *Individual choice behavior: A theoretical analysis*. John Wiley & Sons, Inc., New York; Chapman & Hall, Ltd., London, 1959.
- [LW14] Cong Han Lim and Stephen Wright. Beyond the birkhoff polytope: Convex relaxations for vector permutation problems. In *Advances in Neural Information Processing Systems 27*, pages 2168–2176. 2014.
- [Mah00] Hosam M. Mahmoud. *Sorting: A Distribution Theory*. Wiley Series in Discrete Mathematics and Optimization. Wiley, 2000.
- [Mam91] Enno Mammen. Estimating a smooth monotone regression function. *Ann. Statist.*, 19(2):724–740, June 1991.
- [Mar96] John I. Marden. *Analyzing and modeling rank data*. CRC Press, 1996.
- [Mas07] P. Massart. *Concentration inequalities and model selection: Ecole d’Eté de Probabilités de Saint-Flour XXXIII - 2003*. Number no. 1896 in *Ecole d’Eté de Probabilités de Saint-Flour*. Springer-Verlag, 2007.
- [MBZ13] Arya Mazumdar, Alexander Barg, and Gilles Zémor. Constructions of rank modulation codes. *IEEE transactions on information theory*, 59(2):1018–1029, 2013.
- [Men15] Shahar Mendelson. Learning without concentration. *J. ACM*, 62(3), June 2015.
- [MG15] Lucas Maystre and Matthias Grossglauser. Robust active ranking from sparse noisy comparisons. *arXiv preprint arXiv:1502.05556*, 2015.
- [MHT10] Rahul Mazumder, Trevor Hastie, and Robert Tibshirani. Spectral regularization algorithms for learning large incomplete matrices. *Journal of machine learning research*, 11(Aug):2287–2322, 2010.
- [ML65] Don H. McLaughlin and R. Duncan Luce. Stochastic transitivity and cancellation of preferences between bitter-sweet solutions. *Psychonomic Science*, 2(1-12):89–90, 1965.
- [MMV13] Konstantin Makarychev, Yury Makarychev, and Aravindan Vijayaraghavan. Sorting noisy data with partial information. In *Innovations in Theoretical Computer Science 2013, Berkeley, CA, USA, January 9-12, 2013*, pages 515–528, 2013.
- [Mon15] Andrea Montanari. Computational implications of reducing data to sufficient statistics. *Electronic Journal of Statistics*, 9(2):2370–2390, 2015.

- [MPW16] Ankur Moitra, William Perry, and Alexander S. Wein. How robust are reconstruction thresholds for community detection? In *Proceedings of the 48th Annual ACM SIGACT Symposium on Theory of Computing, STOC 2016, Cambridge, MA, USA, June 18-21, 2016*, pages 828–841, 2016.
- [MPW18] Cheng Mao, Ashwin Pananjady, and Martin J. Wainwright. Breaking the $1/\sqrt{n}$ barrier: Faster rates for permutation-based models in polynomial time. *arXiv preprint arXiv:1802.09963*, 2018.
- [Mur89] Fionn Murtagh. review of book data, expert knowledge and decisions, w. gaul and m. schader (eds.), springer-verlag, 1988. *Journal of Classification*, 6:129–132, 1989.
- [MvdG97] Enno Mammen and Sara van de Geer. Locally adaptive regression splines. *Ann. Statist.*, 25(1):387–413, February 1997.
- [MW15] Zongming Ma and Yihong Wu. Computational barriers in minimax submatrix detection. *Ann. Statist.*, 43(3):1089–1116, June 2015.
- [MWR17] Cheng Mao, Jonathan Weed, and Philippe Rigollet. Minimax rates and efficient algorithms for noisy sorting. *arXiv preprint arXiv:1710.10388*, 2017.
- [NOS12] Sahand Negahban, Sewoong Oh, and Devavrat Shah. Iterative ranking from pair-wise comparisons. In *Advances in Neural Information Processing Systems*, pages 2474–2482, 2012.
- [NOS16] Sahand Negahban, Sewoong Oh, and Devavrat Shah. Rank centrality: Ranking from pairwise comparisons. *Operations Research*, 65(1):266–287, 2016.
- [NOTX17] Sahand Negahban, Sewoong Oh, Kiran K. Thekumparampil, and Jiaming Xu. Learning from comparisons and choices. *arXiv preprint arXiv:1704.07228*, 2017.
- [NP66] Jerzy Neyman and Egon S. Pearson. *Joint statistical papers*. Univ of California Press, 1966.
- [NPT85] Arkadi S Nemirovski, Boris Teodorovich Polyak, and Aleksandr Borisovich Tsybakov. Convergence rate of nonparametric estimates of maximum-likelihood type. *Problemy Peredachi Informatsii*, 21(4):17–33, 1985.
- [NW11] Sahand Negahban and Martin J. Wainwright. Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *Ann. Statist.*, 39(2):1069–1097, 04 2011.

- [NW12] Sahand Negahban and Martin J. Wainwright. Restricted strong convexity and (weighted) matrix completion: Optimal bounds with noise. *Journal of Machine Learning Research*, 13:1665–1697, May 2012.
- [OL99] Michael J. O’Brien and R. Lee Lyman. *Seriation, stratigraphy, and index fossils: the backbone of archaeological dating*. Springer Science & Business Media, 1999.
- [PABN16] Daniel L. Pimentel-Alarcón, Nigel Boston, and Robert D. Nowak. A characterization of deterministic sampling patterns for low-rank matrix completion. *IEEE Journal of Selected Topics in Signal Processing*, 10(4):623–636, 2016.
- [Pet99] W. M. Flinders Petrie. Sequences in prehistoric remains. *The Journal of the Anthropological Institute of Great Britain and Ireland*, 29(3/4):295–301, 1899.
- [PMM⁺17a] Ashwin Pananjady, Cheng Mao, Vidya Muthukumar, Martin J. Wainwright, and Thomas A. Courtade. Worst-case vs average-case design for estimation from fixed pairwise comparisons. *arXiv preprint arXiv:1707.06217*, 2017.
- [PMM⁺17b] Ashwin Pananjady, Cheng Mao, Vidya Muthukumar, Martin J. Wainwright, and Thomas A. Courtade. Worst-case vs average-case design for estimation from fixed pairwise comparisons. *arXiv preprint arXiv:1707.06217*, 2017.
- [PNZ⁺15] Dohyung Park, Joe Neeman, Jin Zhang, Sujay Sanghavi, and Inderjit Dhillon. Preference completion: Large-scale collaborative ranking from pairwise comparisons. In *International Conference on Machine Learning*, pages 1907–1916, 2015.
- [PWC16] Ashwin Pananjady, Martin J. Wainwright, and Thomas A. Courtade. Linear regression with an unknown permutation: Statistical and computational limits. *arXiv preprint arXiv:1608.02902*, 2016.
- [PWC17] Ashwin Pananjady, Martin J. Wainwright, and Thomas A. Courtade. Denoising linear models with permuted data. In *Information Theory (ISIT), 2017 IEEE International Symposium on*, pages 446–450. IEEE, 2017.
- [RA14] Arun Rajkumar and Shivani Agarwal. A statistical convergence perspective of algorithms for rank aggregation from pairwise data. In *International Conference on Machine Learning*, pages 118–126, 2014.
- [RA16] Arun Rajkumar and Shivani Agarwal. When can we rank well from comparisons of $o(n \log(n))$ non-actively chosen pairs? In Vitaly Feldman, Alexander Rakhlin, and Ohad Shamir, editors, *29th Annual Conference*

on Learning Theory, volume 49 of *Proceedings of Machine Learning Research*, pages 1376–1401, Columbia University, New York, New York, USA, 2016. PMLR.

- [Rob51] William S. Robinson. A method for chronologically ordering archaeological deposits. *American Antiquity*, 16(4):293–301, 1951.
- [RT11] Angelika Rohde and Alexandre B. Tsybakov. Estimation of high-dimensional low-rank matrices. *Ann. Statist.*, 39(2):887–930, 04 2011.
- [RV07] Mark Rudelson and Roman Vershynin. Sampling from large matrices: An approach through geometric functional analysis. *Journal of the ACM (JACM)*, 54(4):21, 2007.
- [RWD88] Tim Robertson, Farrol T. Wright, and Richard Dykstra. *Order Restricted Statistical Inference*. Probability and Statistics Series. Wiley, 1988.
- [SBB⁺16] Nihar B. Shah, Sivaraman Balakrishnan, Joseph Bradley, Abhay Parekh, Kannan Ramchandran, and Martin J. Wainwright. Estimation from pairwise comparisons: sharp minimax bounds with topology dependence. *Journal of Machine Learning Research*, 17:Paper No. 58, 47, 2016.
- [SBC05] Neil Stewart, Gordon D. A. Brown, and Nick Chater. Absolute identification by relative judgment. *Psychological Review*, 112(4):881, 2005.
- [SBGW17] Nihar B. Shah, Sivaraman Balakrishnan, Adityanand Guntuboyina, and Martin J. Wainwright. Stochastically transitive models for pairwise comparisons: statistical and computational issues. *IEEE Trans. Inform. Theory*, 63(2):934–959, 2017.
- [SBW16a] Nihar B. Shah, Sivaraman Balakrishnan, and Martin J. Wainwright. Feeling the Bern: Adaptive estimators for bernoulli probabilities of pairwise comparisons. In *Information Theory (ISIT), 2016 IEEE International Symposium on*, pages 1153–1157. IEEE, 2016.
- [SBW16b] Nihar B. Shah, Sivaraman Balakrishnan, and Martin J. Wainwright. A permutation-based model for crowd labeling: Optimal estimation and robustness. *arXiv preprint arXiv:1606.09632*, 2016.
- [SBW17] Nihar B. Shah, Sivaraman Balakrishnan, and Martin J. Wainwright. Low permutation-rank matrices: Structural properties and noisy completion. *arXiv preprint arXiv:1709.00127*, 2017.
- [Sok63] Robert R. Sokal. The principles and practice of numerical taxonomy. *Taxon*, 12(5):190–199, 1963.

- [Sto08] Quentin F. Stout. Unimodal regression via prefix isotonic regression. *Comput. Statist. Data Anal.*, 53(2):289–297, 2008.
- [SW15] Nihar B. Shah and Martin J. Wainwright. Simple, robust and optimal ranking from pairwise comparisons. *arXiv preprint arXiv:1512.08949*, 2015.
- [SXB08] Rohit Singh, Jinbo Xu, and Bonnie Berger. Global alignment of multiple protein interaction networks with application to functional orthology detection. *Proceedings of the National Academy of Sciences*, 105(35):12763–12768, 2008.
- [SZ01] Jyh-Ming Shoung and Cun-Hui Zhang. Least squares estimators of the mode of a unimodal regression function. *Ann. Statist.*, 29(3), 2001.
- [TG14] Bradley C. Turnbull and Sujit K. Ghosh. Unimodal density estimation using bernstein polynomials. *Computational Statistics & Data Analysis*, 72:13–29, 2014.
- [Thu27] Louis L. Thurstone. A law of comparative judgment. *Psychological review*, 34(4):273, 1927.
- [Tsy09] Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Series in Statistics. Springer, 2009.
- [Var57] Rom R. Varshamov. Estimate of the number of signals in error correcting codes. In *Dokl. Akad. Nauk SSSR*, volume 117, pages 739–741, 1957.
- [vdG90] Sara van de Geer. Estimating a regression function. *Ann. Statist.*, 18(2), 1990.
- [vdG91] Sara van de Geer. The entropy bound for monotone functions. Technical Report 91-10, Leiden Univ., 1991.
- [vdG93] Sara van de Geer. Hellinger-consistency of certain nonparametric maximum likelihood estimators. *Ann. Statist.*, 21(1), 1993.
- [vH14] Ramon van Handel. Probability in high dimension. Lecture Notes (Princeton University), 2014.
- [Vin90] Andrew Vince. A rearrangement inequality and the permutahedron. *Amer. Math. Monthly*, 97(4):319–323, 1990.
- [Wai17] Martin J. Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*. In preparation, 2017.
- [WJJ13] Fabian Wauthier, Michael Jordan, and Nebojsa Jojic. Efficient ranking from pairwise comparisons. In *International Conference on Machine Learning*, pages 109–117, 2013.

- [Yi04] Jinhee Yi. Theta-function identities and the explicit formulas for Theta-function and their applications. *J. Math. Anal. Appl.*, 292(2):381–400, 2004.
- [You88] H. Peyton Young. Condorcet’s theory of voting. *American Political Science Review*, 82(4):1231–1244, 1988.
- [ZBV09] Mikhail Zaslavskiy, Francis Bach, and Jean-Philippe Vert. Global alignment of protein–protein interaction networks by graph matching methods. *Bioinformatics*, 25(12):i259–1267, 2009.
- [ZCZJ16] Yuchen Zhang, Xi Chen, Dengyong Zhou, and Michael I Jordan. Spectral methods meet EM: A provably optimal algorithm for crowdsourcing. *The Journal of Machine Learning Research*, 17(1):3537–3580, 2016.
- [Zha02] Cun-Hui Zhang. Risk bounds in isotonic regression. *Ann. Statist.*, 30(2):528–555, April 2002.