OPTIMAL RATES OF STATISTICAL SERIATION

By Nicolas Flammarion, Cheng Mao and Philippe Rigollet

Ecole Normale Supérieure and Massachusetts Institute of Technology

Given a matrix, the seriation problem consists in permuting its rows in such way that all its columns have the same shape, for example, they are monotone increasing. We propose a statistical approach to this problem where the matrix of interest is observed with noise and study the corresponding minimax rate of estimation of the matrices. Specifically, when the columns are either unimodal or monotone, we show that the least squares estimator is optimal up to logarithmic factors and adapts to matrices with a certain natural structure. Finally, we propose a computationally efficient estimator in the monotonic case and study its performance both theoretically and experimentally. Our work is at the intersection of shape constrained estimation and recent work that involves permutation learning, such as graph denoising and ranking.

1. Introduction. The *consecutive 1's problem* (C1P) [FG64] 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 $a \le j \le b$ for two integers a, b between 1 and n.

This problem has its roots in archeology and especially *sequence dating* where the goal is to recover the chronological order of sepultures based on artifacts found in these sepultures where the entry $A_{i,j}$ of matrix A indicates the presence of artifact 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. Already in the noiseless case, this problem presents an interesting algorithmic challenge and is reducible to the famous Travelling 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. C1P belongs to a more general class of so-called *seriation* problems that consist in optimizing various criteria over the discrete set of permutations. While such problems are hard in general, it can be shown that a subset of the these problems, including C1P, can be solve efficiently using spectral

AMS 2000 subject classifications: Primary 62G08; secondary 62C20

Keywords and phrases: Statistical Seriation, Permutation learning, Minimax estimation, Adaptation, Shape constraints, Matrix estimation.

method [ABH98] or convex optimization [FJBd13, LW14]. However, little is known about the robustness to 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 in 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 belong to a class of matrices 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 monotonic columns as a special case. These terms will be formally defined at the end of this section.

The rest of the paper is organized as follows. In Section 2 we formulate the model and discuss related work. Section 3 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 4, 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. Appendix A is devoted to the proofs of the upper bounds, which use the metric entropy bounds proved in Appendix B. The proofs of the information-theoretic lower bounds are presented in Appendix C. In Appendix D, we study the rate of estimation of the efficient estimator for the monotonic case. Appendix E contains a delayed proof of a trivial upper bound. Appendix F presents new bounds for unimodal regression implied by our analysis, which are minimax optimal up to logarithmic factors.

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 $u_n \leq C v_n$ 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 increasing[1] sequences in $\mathbb{R}^n$ by $\mathcal{S}_n = \{a \in \mathbb{R}^n : a_1 \leq \cdots \leq a_n\}$. 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

[1] Throughout the paper, we loosely use the terms "increasing" and "decreasing" to mean "monotonically non-decreasing" and "monotonically non-increasing" respectively.

columns.

For any $l \in [n]$, define the closed convex cone $\mathcal{C}_l = \{a \in \mathbb{R}^n : a_1 \leq \cdots \leq a_l\} \cap \{a \in \mathbb{R}^n : a_l \geq \cdots \geq a_n\}$, 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 $\mathcal{C}_{\mathbf{l}}^m = \mathcal{C}_{l_1} \times \cdots \times \mathcal{C}_{l_m}$. Then $\mathcal{U}^m$ is the union of the n^m closed convex cones $\mathcal{C}_{\mathbf{l}}^m, \mathbf{l} \in [n]^m$.

Finally, let $\mathfrak{S}_n$ be the set of $n \times n$ permutation matrices and define $\mathcal{M} = \bigcup_{\Pi \in \mathfrak{S}_n} \Pi \mathcal{U}^m$ 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 $\Pi \mathcal{C}_{\mathbf{l}}^m, \Pi \in \mathfrak{S}_n, \mathbf{l} \in [n]^m$.

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

2.1. *The seriation model.* Suppose that we observe a matrix $Y \in \mathbb{R}^{n \times m}$, $n \geq 2$ such that

$$Y = \Pi^* A^* + Z \,, \tag{2.1}$$

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

$$\mathbb{E}\big[\exp\big(\mathsf{Tr}(Z^\top M)\big)\big] \leq \exp\Big(\frac{\sigma^2 \|M\|_F^2}{2}\Big) \,,$$

where $\mathsf{Tr}(\cdot)$ is the trace operator. We write $Z \sim \mathrm{subG}_{n,m}(\sigma^2)$ or simply $Z \sim \mathrm{subG}(\sigma^2)$ 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.

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

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

and that

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

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

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

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

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

by the triangle inequality and the previous display. □

Lemma 2.1 guarantees that $\|\tilde{\Pi}A^* - \Pi^* A^*\|_F$ is a pertinent measure of the performance of $\tilde{\Pi}$. 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 4 (and Appendix D), the rest of this paper focuses on the least squares (LS) estimator defined by

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

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

$$\hat{M} \in \operatorname*{argmin}_{M \in \mathcal{M}} \|Y - M\|_F^2\,. \tag{2.3}$$

Note that in our case, the set of parameters $\mathcal{M}$ is not convex, but is a union of $n!n^m$ closed convex cones and 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 4. Nevertheless, the main focus of this paper 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.

2.2. *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^*.

2.2.1. *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 get 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 $z \sim \mathrm{subG}_{n,1}(\sigma^2)$.

When θ^* is assumed to be increasing we speak of isotonic regression [BBBB72]. The LS estimator defined by $\hat{\theta} = \mathrm{argmin}_{\theta \in \mathcal{S}_n} \|\theta - y\|_2^2$ 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

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

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* has it holds uniformly over the set of monotone vectors with variation $V(\theta^*)$. Recently, [CGS15b] 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 $k(\theta) = \mathsf{card}(\{\theta_1, \cdots, \theta_n\})$ denote the cardinality of entries of $\theta \in \mathbb{R}^n$. In this context, [CGS15b] showed that the LS estimator satisfies the adaptive bound

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

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 [CGS15b, 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:

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

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 current paper 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 F.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.

2.2.2. *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 $\min_{i\neq j}\|A_{i,\cdot}-A_{j,\cdot}\|_2 \geq c\sigma\big((\log n)^{1/2}\vee(m\log n)^{1/4}\big)$, 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 $Y_{i,j}\sim\text{Ber}(P_{i,j})$ 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 $A=\Pi^\top P\Pi$ satisfies $A_{1,k}\leq\cdots\leq A_{n,k}$ for all $k\in[n]$. Note that the NP-BTL model is a special case of our model (2.1) where $m=n$ and $Z\sim\text{subG}(1/4)$ 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 [SBGW15] for a variation of the estimator. This method does not yield individual estimators of Π or A, and [CM16] proposed estimators $\hat{\Pi}$ and $\hat{A}$ so that $\hat{\Pi}\hat{A}\hat{\Pi}^\top$ estimates P with the same rate $n^{-1/2}$ up to a logarithmic factor. The non-optimality of this rate has been observed in [SBGW15] 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 [SBW16] 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 $P = \Pi A \Pi^\top, \Pi \in \mathbb{R}^{\times n}$ is an unknown permutation matrix and A has a block structure, namely, there exist positive integers $n_1 < \ldots < n_k < n_{k+1} := n$, and k^2 real numbers $a_{s,t}, (s,t) \in [k]^2$ such that A has entries

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

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, Liu 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. Moreover, 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 $A^*_{1,j} \leq \cdots \leq A^*_{n,j}$ for all users $j \in [m]$. This model falls into the scope of the statistical seriation model (2.1).

3. Main results.

3.1. *Adaptive oracle inequalities.* For a matrix $A \in \mathcal{U}^m$, let $k(A_{\cdot,j}) = \mathsf{card}(\{A_{1,j}, \dots, A_{n,j}\})$ 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 3.1. *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*

$$\frac{1}{nm}\|\hat{\Pi}\hat{A} - \Pi^* A^*\|_F^2 \lesssim \min_{A \in \mathcal{U}^m} \Big(\frac{1}{nm}\|A - A^*\|_F^2 + \sigma^2 \frac{K(A)}{nm} \log \frac{enm}{K(A)}\Big) + \sigma^2 \frac{\log n}{m} \tag{3.1}$$

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

$$\frac{1}{nm}\mathbb{E}\|\hat{\Pi}\hat{A} - \Pi^* A^*\|_F^2 \lesssim \min_{A \in \mathcal{U}^m} \Big(\frac{1}{nm}\|A - A^*\|_F^2 + \sigma^2 \frac{K(A)}{nm} \log \frac{enm}{K(A)}\Big) + \sigma^2 \frac{\log n}{m}\,. \tag{3.2}$$

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 above oracle inequalities indicate that the LS estimator automatically trades off the approximation error $\|A - A^*\|_F^2$ for the stochastic error $\sigma^2 K(A) \log(enm/K(A))$.

If A^* is assumed to have unimodal columns, then we can take $A = A^*$ in (3.1) and (3.2) to get the following corollary.

Corollary 3.2. *For $A^* \in \mathcal{U}^m$ and $Y = \Pi^* A^* + Z$, the LS estimator $(\hat{\Pi}, \hat{A})$ satisfies*

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

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 columuns if the permutation Π^* is known. It can be viewed as a matrix version of the adaptive bound (2.5) in 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.

3.2. *Global oracle inequalities.* The bounds in Theorem 3.1 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

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

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

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

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 3.6 up to logarithmic terms.

THEOREM 3.3. *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*

$$\frac{1}{nm}\|\hat\Pi\hat A - \Pi^* A^*\|_F^2 \lesssim \min_{A\in\mathcal{U}^m}\Big[\frac{1}{nm}\|A - A^*\|_F^2 + \Big(\frac{\sigma^2 V(A)\log n}{n}\Big)^{2/3}\Big] + \sigma^2\frac{\log n}{n\wedge m}\,. \tag{3.3}$$

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 3.3 leads to the following corollary that indicates that the LS estimator is adaptive to the quantity $V(A^*)$.

COROLLARY 3.4. *For $A^* \in \mathcal{U}^m$ and $Y = \Pi^* A^* + Z$, the LS estimator $(\hat\Pi, \hat A)$ satisfies*

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

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.

Having stated the main upper bounds, we digress a little to remark that the proofs of Theorem 3.1 and Theorem 3.3 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 Appendix F.

3.3. *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 3.2 and Corollary 3.4 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 $m \le K_0 \le nm$ and $V_0 > 0$, define $\mathcal{U}^m_{K_0} = \{A \in \mathcal{U}^m : K(A) \le K_0\}$ and $\mathcal{U}^m(V_0) = \{A \in \mathcal{U}^m : V(A) \le V_0\}$. We present below two lower bounds, one for the adaptive rate uniformly over $\mathcal{U}^m_{K_0}$ 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 $\mathcal{U}^m_{K_0} \cap \mathcal{U}^m(V_0)$ can also be proved and is presented as Proposition C.3.

THEOREM 3.5. *There exists a constant $c \in (0, 1)$ such that for any $K_0 \ge m$, and any estimator $(\hat\Pi, \hat A)$, it holds that*

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

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 $\Pi^* A^*$, 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 3.2 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 3.4 also holds:

Theorem 3.6. *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*

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

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 3.4 and the lower bound of Theorem 3.6 above. Indeed the lower bound features a term $\frac{\sigma^2}{m} \wedge m^2 V_0^2$ instead of just $\frac{\sigma^2}{m}$. In the regime $m^2 V_0^2 < \frac{\sigma^2}{m}$, where A has very small variation, the LS estimator may not be optimal. Proposition E.1 indicates that a matrix with constant columns obtained by averaging achieves optimality in this extreme regime.

4. 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 2.2. In fact, it covers the example of bipartite ranking discussed at the end of Section 2.2. 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

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

Since $\mathcal{S}^m$ is a convex subset of $\mathcal{U}^m$, it is easily seen that the upper bounds in Theorem 3.1 and 3.3 remain valid in this case. The lower bounds of Theorem 3.5 (with $\log l$ replaced by 1) and Theorem 3.6 also extend to this case; see Appendix C.

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 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.

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.

Let $K(A)$ and $V(A)$ be defined as before. Moreover, for a matrix $A \in \mathcal{S}^m$, let $\mathcal{J}$ denote the set of pairs of indices $(i, j) \in [n]^2$ such that $A_{i,\cdot}$ and $A_{j,\cdot}$ are not identical. Define the quantity $R(A)$ by

$$R(A) = \frac{1}{n} \max_{\substack{\mathcal{I} \subset [n]^2 \\ |\mathcal{I}| = n}} \sum_{(i,j) \in \mathcal{I} \cap \mathcal{J}} \Big(\frac{\|A_{i,\cdot} - A_{j,\cdot}\|_2^2}{\|A_{i,\cdot} - A_{j,\cdot}\|_\infty^2} \wedge \frac{m\|A_{i,\cdot} - A_{j,\cdot}\|_2^2}{\|A_{i,\cdot} - A_{j,\cdot}\|_1^2} \Big) \,. \tag{4.1}$$

It can be shown (see Appendix D) that $1 \leq R(A) \leq \sqrt{m}$. Intuitively, the quantity $R(A)$ is small if the difference u of any two rows of A is either very sparse ($\|u\|_2/\|u\|_\infty$ is small) or very dense ($m\|u\|_2/\|u\|_1$ is small). Indeed, for any nonzero vector $u \in \mathbb{R}^m$, $\|u\|_2^2/\|u\|_\infty^2 \geq 1$ with equality achieved when $\|u\|_0 = 1$, and $m\|u\|_2^2/\|u\|_1^2 \geq 1$ with equality achieved when all entries of u are the same.

For matrices with small $R(\cdot)$ values, it is possible to aggregate the information across each row to learn the unknown permutation Π^* in a simple fashion. Recovering the permutation Π^*, is equivalent to ordering (or ranking reversely) the rows of $\Pi^* A^*$ from their noisy version Y.

One simple method to achieve this goal, which we call $\mathsf{RankSum}$, is to permute the rows of Y so that they have increasing row sums. However, it is easy to observe that this method fails if

$$A^* = \begin{bmatrix} \sqrt{m} & 0 & \dots & 0 \\ 2\sqrt{m} & 0 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ n\sqrt{m} & 0 & \dots & 0 \end{bmatrix} \tag{4.2}$$

where $A^*_{i,1} = i\sqrt{m}$ and entries of Z are i.i.d. standard Gaussian variables, because the sum of noise in a row has order $\sqrt{m}$ which is no less than the gaps between row sums of A^*. In fact, $R(A^*) = 1$ and it should be easy to distinguish the two types of rows of A^*, for example, by looking at the first entry of a row. This motivates us to consider the following method called $\mathsf{RankScore}$.

For $i, i' \in [n]$, 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 $\mathsf{RankScore}$ procedure is defined as follows:

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

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

   where $\tau := C\sigma\sqrt{\log(nm)}$ for some tuning constant C (see Appendix D for more details).
2. Then order the rows of Y so that their scores are increasing, with ties broken arbitrarily.

The $\mathsf{RankScore}$ 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 $\tilde{\Pi}\mathcal{S}^m$. The estimator $(\tilde{\Pi}, \tilde{A})$ enjoys the following rate of estimation.

THEOREM 4.1. *For $A^* \in \mathcal{S}^m$ and $Y = \Pi^* A^* + Z$, let $(\tilde{\Pi}, \tilde{A})$ be the estimator defined above using the $\mathsf{RankScore}$ procedure with threshold $\tau =$*

$3\sigma\sqrt{(C+1)\log(nm)}$, $C>0$. Then it holds that

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

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, $R(A^*)\le\sqrt{m}$, so the estimator is still consistent with the permutation error (the last term) decaying at a rate no slower than $\tilde{O}(\frac{1}{\sqrt{m}})$. 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 4.2. *In the same way that Theorem 3.3 follows from Theorem 3.1, we can deduce from Theorem 4.1 a global bound for the estimator $(\tilde{\Pi},\tilde{A})$ which has rate*

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

We conclude this section 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 $\hat{A}^{\text{oracle}}$ defined by the projection of Y onto the cone $\mathcal{S}^m$. 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 $\log_{10}\big(\frac{1}{nm}\|\hat{\Pi}\hat{A}-A^*\|_F^2\big)$ unless otherwise specified.

We begin with two simple examples for which we set $n=m$. In the left plot of Figure 1, A^* is defined as in (4.2). As expected, RankSum fails to estimate the true permutation and performs very poorly. On the other hand, RankScore succeeds in recovering the correct permutation and has roughly the same performance as the oracle. Because the difference of any two rows of A^* is 1-sparse, $R(A^*)=1$ according to (4.1) and the discussion thereafter. Hence, Theorem 4.1 predicts the fast rate, which is verified by the

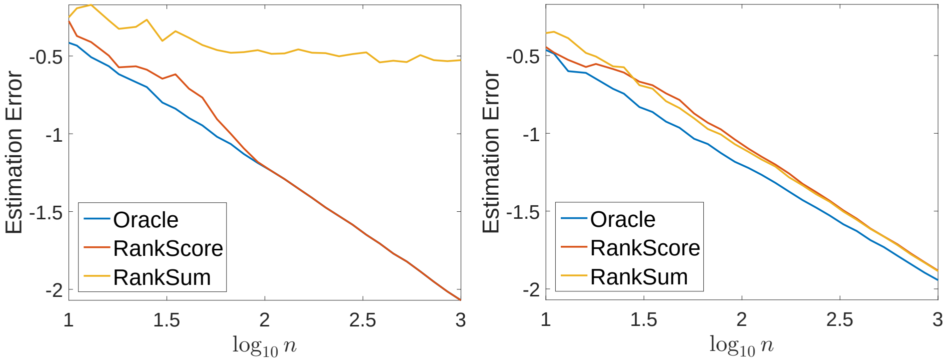

FIG 1. *Estimation errors of three estimators for two deterministic A^* of size $n \times n$. Left: rows of A^* are* 1*-sparse; Right: columns of A^* are identical.*

experiment. The right plot illustrates another extreme case; more precisely, we set A^* to be the matrix with all m columns equal to $\frac{1}{n}(1, \cdots, n)^\top$. The difference of any two rows of A^* is constant across all entries, so again we have $R(A^*) = 1$ by (4.1). Thus RankScore achieves the fast rate as expected. Note that RankSum also performs well in this case.

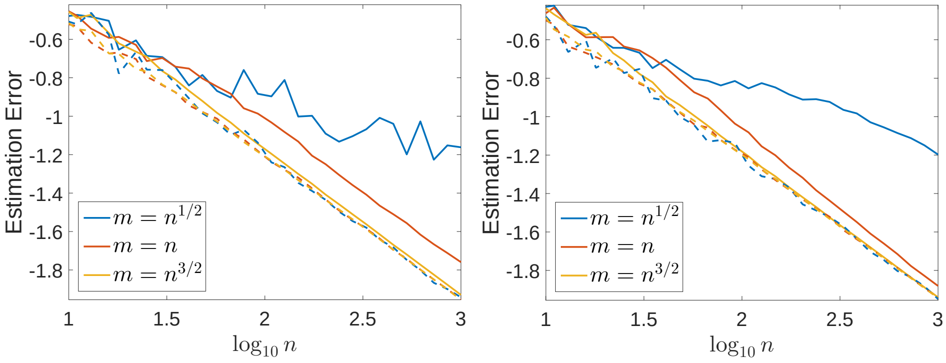

FIG 2. *Estimation errors of the oracle (dashed lines) and RankScore (solid lines) for different regimes of (n, m) and randomly generated A^* of size $n \times m$. Left: $K(A^*) = 5m$; Right: $V(A^*) \leq 1$.*

In Figure 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 3.1 and 3.3. (The difference is minor in the plots as n is not sufficiently large). For RankScore, the permutation term dominates the estimation term when $m = n^{1/2}$ by Theorem 4.1. 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 illustrates the good behavior of RankScore in this random scenario.

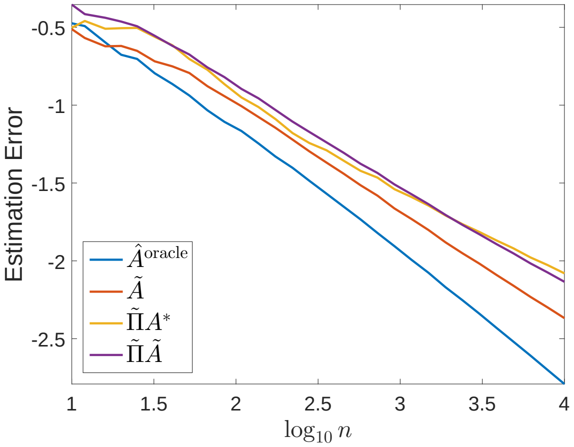

Fig 3. *Various estimation errors of the oracle and* RankScore *for the triangular matrix.*

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 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 4.1, 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 D.1 and 2.1, 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.

5. 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 paper 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 3.5 (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 3.4. 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 3.2 and $\log l$ in the lower bound of Theorem 3.5 do not match if $l < n$. We do not seek answers to these questions in this paper 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 [SBGW15]. 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, SBW16] 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.

Acknowledgments. P.R. is partially supported by NSF grant DMS-1317308 and NSF CAREER Award DMS-1053987. N.F. gratefully acknowledges partial support by NSF grant DMS-1317308, the Institute for Data Systems and Society and the Mathematics Department during his visit at MIT. C.M. acknowledges partial support by NSF CAREER Award DMS-1053987. P.R. also thanks Enno Mammen for his help in retracing the literature on sharp entropy bounds for monotone classes and Ramon van Handel for a stimulating discussion on entropy numbers. The authors thank Alexandre Tsybakov for bringing concurrent work on unimodal regression by Pierre C. Bellec to their attention. Finally, the authors thank Pierre C. Bellec for pointing out an error in an earlier version of Lemma A.1 and suggesting that it holds for all closed sets.

APPENDIX A: 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* [ALMT14] of a convex cone Θ defined by

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

This has been successfully applied to isotonic and more general shape constrained regression [CGS15b, 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

$$\|\hat{\theta} - \theta^*\|_2 = \operatorname*{argmax}_{t\ge 0}\Big(\sup_{\theta\in\Theta, \|\theta-\theta^*\|_2\le t} \langle \theta - \theta^*, z\rangle - \frac{t^2}{2}\Big)\,. \tag{A.1}$$

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 [ALMT14]. The variational formula was first proposed for convex regression [Cha14], and later exploited in several different settings, including matrix estimation with shape constraints [CGS15a] and unimodal regression [CL15]. Similar ideas have appeared in other works, for example, analysis of empirical risk minimization [Men15], ranking from pairwise comparison [SBGW15] 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 (A.1). On the other hand, Chatterjee's variational formula seems more flexible as computations of the statistical dimension based on [ALMT14] are currently limited to convex sets Θ with a polyhedral structure. In this paper, we use exclusively Chatterjee's variational formula.

A.1. 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 A.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 $\hat a \in \mathrm{argmin}_{a \in \mathcal{C}} \|y - a\|_2^2$ be a projection of y onto $\mathcal{C}$. Define the function $f_{a^*} : \mathbb{R}_+ \to \mathbb{R}$ 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

$$\|\hat a - a^*\|_2 \in \operatorname*{argmax}_{t \geq 0} f_{a^*}(t). \tag{A.2}$$

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,

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

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

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

Therefore (A.2) follows.

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

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.

A.2. Proof of Theorem 3.1. 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 A.2 (Chaining tail inequality). *Let $\Theta \subset \mathbb{R}^d$ and $z \sim \mathrm{subG}(\sigma^2)$ in $\mathbb{R}^d$. For any $\theta_0 \in \Theta$, it holds that*

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

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

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

$$R_1 = R_1(\tilde{A}, n) = \sigma\Big(\sqrt{K(\tilde{A}) \log \frac{enm}{K(\tilde{A})}} + \sqrt{n \log n}\Big) \tag{A.3}$$

and

$$R_2 = R_2(\tilde{A}, n) = \sigma^2\Big(K(\tilde{A}) \log \frac{enm}{K(\tilde{A})} + n \log n\Big). \tag{A.4}$$

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

LEMMA A.3. *Suppose* $Y = A^* + Z$ *where* $A^* \in \mathbb{R}^{n\times m}$ *and* $Z \sim \mathrm{subG}(\sigma^2)$. *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*

$$f_{\tilde{A}}(t) \le CR_1 t + t\|A^* - \tilde{A}\|_F - \frac{t^2}{2} + st \tag{A.5}$$

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

PROOF. Define $\Theta = \Theta_{\mathcal{M}}(\tilde{A}, 1) = \bigcup_{\lambda \ge 0}\{B - \lambda\tilde{A} : B \in \mathcal{M} \cap \mathcal{B}^{nm}(\lambda\tilde{A}, 1)\}$ (see also Definition (B.2)). 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,

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

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

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

Moreover, it follows from Lemma B.5 that

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

Combining the previous three displays, we see that

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

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

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

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

We are now in a position to prove the adaptive oracle inequalities in Theorem 3.1. 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 A.3. We can apply Lemma A.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 (A.5). Then it follows from Lemma A.3 that with probability at least $1 - C\exp(-\frac{cs^2}{\sigma^2})$, it holds for all $t \ge t^*$ that

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

Therefore by Lemma A.1,

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

and thus

$$\|\hat{\Pi}\hat{A} - A^*\|_F \le C(R_1 + \|A^* - \tilde{A}\|_F) + 2s \tag{A.6}$$

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

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

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

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, (3.1) 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 (A.6) yields

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

where R_2 is defined in (A.4). Integrating the tail probability, we get that

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

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 (3.2).

A.3. Proof of Theorem 3.3. In the setting of isotonic regression, [BT15] derived global bounds from adaptive bounds by a block approximation method, which also applies to our setting. For $k \in [n]$, let

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

Define $k^* = \lceil (\frac{V(a)^2 n}{\sigma^2 \log(en)})^{1/3} \rceil$. The lemma below is very similar to [BT15, Lemma 2] and their proof also extends to the unimodal case with minor modifications. We present the result with proof for completeness.

Lemma A.4. *For $a \in \mathcal{U}$ and $k \in [n]$, there exists $\tilde{a} \in \mathcal{U}_k$ such that*

$$\frac{1}{\sqrt{n}}\|\tilde{a} - a\|_2 \leq \frac{V(a)}{2k}. \tag{A.7}$$

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

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

Moreover,

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

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

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

and $I_k = \Big[\underline{a} + \frac{k-1}{k}V(a), \bar{a}\Big]$. 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, $|\tilde{a}_i - a_i| \le \frac{V(a)}{2k}$ for $i \in [n]$, which implies (A.7).

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

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

and

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

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

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

and

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

□

It is straightforward to generalize the lemma to matrices. For $\mathbf{k} \in [n]^m$, we write $\mathbf{k} = (k_1, \dots, k_m)$ and let

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

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

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

LEMMA A.5. *For $A \in \mathcal{U}^m$, there exists $\tilde{A} \in \mathcal{U}_{\mathbf{k}^*}^m$ such that*

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

and

$$\frac{\sigma^2 K(\tilde{A})}{nm}\log(en) \le 2\Big(\frac{\sigma^2 V(A)\log(en)}{n}\Big)^{2/3} + \frac{2\sigma^2}{n}\log(en)\,.$$

Proof. Applying Lemma A.4 to columns of A, we see that there exists $\tilde{A} \in \mathcal{U}^m_{\mathbf{k}^*}$ such that

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

and

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

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

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

and similarly

$$\frac{\sigma^2 K(\tilde{A})}{nm}\log(en) \le 2\Big(\frac{\sigma^2 V(A)\log(en)}{n}\Big)^{2/3} + \frac{2\sigma^2}{n}\log(en)\,.$$

□

For $A \in \mathcal{U}^m$, choose $\tilde{A} \in \mathcal{U}^m_{\mathbf{k}^*}$ according to Lemma A.5. Then

$$\begin{aligned}\frac{1}{nm}\|\tilde{A} - A^*\|_F^2 &\le \frac{2}{nm}\|A - A^*\|_F^2 + \frac{2}{nm}\|\tilde{A} - A\|_F^2 \\ &\le \frac{2}{nm}\|A - A^*\|_F^2 + \frac{5}{4}\Big(\frac{\sigma^2 V(A)\log n}{n}\Big)^{2/3} + \frac{5\sigma^2}{4n}\log n \qquad \text{(A.8)}\end{aligned}$$

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

$$\frac{\sigma^2 K(\tilde{A})}{nm}\log(en) \le 5\Big(\frac{\sigma^2 V(A)\log n}{n}\Big)^{2/3} + \frac{5\sigma^2}{n}\log n\,. \qquad \text{(A.9)}$$

Plugging (A.8) and (A.9) into the right-hand side of (3.1) and (3.2), and then minimizing over $A \in \mathcal{U}^m$, we complete the proof.

APPENDIX B: METRIC ENTROPY

In this section, we study various *covering numbers* or *metric entropy* related to the parameter space of the model (2.1). First recall some standard definitions that date back at least to [KT61]. An ε-net of a subset $G \subset \mathbb{R}^n$ with respect to a norm $\|\cdot\|$ is a set $\{w_1, \cdots, w_N\} \subset G$ such that for any $w \in G$, there exists $i \in [N]$ for which $\|w - w_i\| \le \varepsilon$. 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.

B.1. Cartesian product of cones. Lemma B.2 below bounds covering numbers of product spaces and is useful in later proofs. We start with a well-known result on the covering number of a Euclidean ball with respect to the ℓ_∞-norm (see e.g. [Mas07, Lemma 7.14] for an analogous result).

Lemma B.1. *For any $\varepsilon \in (0,1]$,*

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

for some constant $C > 0$.

Proof. We aim at bounding the covering number of a Euclidean ball by cubes. Let $\{x^1, \dots, x^M\}$ be a maximal $\frac{\varepsilon}{\sqrt{m}}$-packing of $\mathcal{B}^m(0,1)$ with respect to the ℓ_∞-norm, where a δ-packing of a set G with respect to a norm $\|\cdot\|$ is a set $\{w_1, \cdots, w_N\} \subset G$ such that $\|w_i - w_j\| \ge \delta$ for all distinct $i, j \in [N]$. Then this set is necessarily an $\frac{\varepsilon}{\sqrt{m}}$-net of $\mathcal{B}^m(0,1)$ by maximality, so $N(\mathcal{B}^m(0,1), \|\cdot\|_\infty, \frac{\varepsilon}{\sqrt{m}}) \le M$. Consider the cubes with side length $\frac{\varepsilon}{\sqrt{m}}$ centered at x^i for $1 \le i \le M$. These cubes are disjoint and contained in the set $\mathcal{B}^m(0,1) + Q^m(\frac{\varepsilon}{\sqrt{m}})$, where $Q^m(\frac{\varepsilon}{\sqrt{m}})$ is the cube with side length $\frac{\varepsilon}{\sqrt{m}}$ centered at the origin in $\mathbb{R}^m$. Since $Q^m(\frac{\varepsilon}{\sqrt{m}}) \subset \mathcal{B}^m(0,\varepsilon)$,

$$\begin{aligned} M \operatorname{Vol}\Big(Q^m(\frac{\varepsilon}{\sqrt{m}})\Big) &\le \operatorname{Vol}\Big(\mathcal{B}^m(0,1) + Q^m(\frac{\varepsilon}{\sqrt{m}})\Big) \\ &\le \operatorname{Vol}(\mathcal{B}^m(0,1+\varepsilon)) \\ &\le \operatorname{Vol}(\mathcal{B}^m(0,2)). \end{aligned}$$

This proves the following bound on the covering number in terms of a volume ratio:

$$N\Big(\mathcal{B}^m(0,1), \|\cdot\|_\infty, \frac{\varepsilon}{\sqrt{m}}\Big) \le \frac{\operatorname{Vol}(\mathcal{B}^m(0,2))}{\operatorname{Vol}(Q^m(\frac{\varepsilon}{\sqrt{m}}))} \le \frac{C^m m^{-m/2}}{\varepsilon^m m^{-m/2}} = (C/\varepsilon)^m.$$

□

Now we study the metric entropy of a Cartesian product of convex cones. 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 $\mathcal{C} = \mathcal{C}_1 \times \cdots \times \mathcal{C}_m$.

LEMMA B.2. *With the notation above, suppose that* $a_{I_i} \in \mathcal{C}_i \cap (-\mathcal{C}_i)$. *Then for any* $t > 0$ *and* $\varepsilon \in (0, t]$,

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

for some constant $C > 0$.

PROOF. 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 an $\frac{\varepsilon}{2\sqrt{m}}$-net of $\mathcal{B}^m(0,1)$ with respect to the ℓ_∞-norm. Define

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

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

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

We claim that $\bigcup_{\mu \in \mathcal{N}_{\mathcal{D}}} \mathcal{N}_\mu$ is an 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 $\|\lambda\|_2 = \|v - a\|_2 \le 1$. Hence we can find $\mu \in \mathcal{N}_{\mathcal{B}}$ such that $\|\mu - \lambda\|_\infty \le \frac{\varepsilon}{2\sqrt{m}}$. In particular, for all $i \in [m]$, $\mu_i \ge \lambda_i - \frac{\varepsilon}{2\sqrt{m}} \ge -\frac{1}{2\sqrt{m}}$, 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 $\mathcal{N}_{\mu_i}$, there exists $w_{I_i} \in \mathcal{N}_{\mu_i}$ such that $\|w_{I_i} - v_{I_i}\|_2 \le (\mu_i + \frac{1}{\sqrt{m}})\varepsilon$. Let $w = (w_{I_1}, \ldots, w_{I_m}) \in \mathcal{N}_\mu$. Since

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

we conclude that

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

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 B.1, $|\mathcal{N}_{\mathcal{D}}| \leq |\mathcal{N}_{\mathcal{B}}| \leq (C/\varepsilon)^m$. Moreover, recall that $\mathcal{N}_{\mu_i}$ 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

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

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

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

Taking the logarithm completes the proof. □

B.2. Unimodal vectors and matrices. Recall that $\mathcal{S}_n$ denotes the closed convex cone of increasing vectors in $\mathbb{R}^n$. First, we prove a result on the metric entropy of $\mathcal{S}_n$ intersecting with a ball using Lemma B.2.

LEMMA B.3. *Let $b \in \mathbb{R}^n$ be such that $b_1 = \cdots = b_n$. Then for any $t > 0$ and $\varepsilon > 0$,*

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

PROOF. The majority of the 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 bound holds trivially if $\varepsilon > t$, since the left-hand side is zero. It also clearly holds when $n = 1$. Hence we can assume without loss of generality that $\varepsilon \leq t$ and $n = 2n' \geq 2$. Moreover, assume that $t = 1$ for simplicity and the proof will work for any $t > 0$. Let $I = \{1, \ldots, n'\}$ and observe that

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

Let k be the smallest integer for which $2^k > n'$. We partition I into k blocks $A_j = I \cap [2^j, 2^{j+1})$ for $j \in [k]$ and let $m_j = |A_j|$. Since $\mathcal{S}_{n'} \subset \mathcal{S}_{m_1} \times \cdots \times \mathcal{S}_{m_k}$, Lemma B.2 yields that

$$\begin{aligned}&\log N(\mathcal{S}_{n'} \cap \mathcal{B}^{n'}(b_I,1), \|\cdot\|_2, \varepsilon/\sqrt{2}\,)\\ &\qquad\leq k \log\frac{C}{\varepsilon} + \sum_{j=1}^k \log N\Big(\mathcal{S}_{m_j} \cap \mathcal{B}^{m_j}(b_{A_j},1), \|\cdot\|_2, \frac{\varepsilon}{3\sqrt{2}}\Big)\,.\end{aligned} \tag{B.1}$$

We know from [Cha14, Lemma 4.20] that for any $c \le d$ and $n \ge 1$,

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

For each $a \in \mathcal{S}_n \cap \mathcal{B}^n(0,1)$, it holds that $|a_i| \le \frac{1}{\sqrt{i}}$ for $i \in I$ (since either $|a_l| \ge |a_i|$ for all $l \le i$ or $|a_l| \ge |a_i|$ for all $i \le l \le n$; see e.g. [DRXZ14]), so $\max_{i \in A_j} |a_i| \le 2^{-j/2}$. Also $m_j \le 2^j$, so we get that

$$\log N\Big(\mathcal{S}_{m_j} \cap \mathcal{B}^{m_j}(b_{A_j},1), \|\cdot\|_2, \frac{\varepsilon}{3\sqrt{2}}\Big) \le \frac{C}{\varepsilon}$$

for all $j \in [k]$. Substituting this bound into (B.1) and noting that $k \le \log_2 n$, we reach the conclusion

$$\log N(\mathcal{S}_{n'} \cap \mathcal{B}^{n'}(b_I,1), \|\cdot\|_2, \varepsilon/\sqrt{2}) \le C\varepsilon^{-1}\log(en)\,.$$

□

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

$$\begin{aligned}\text{(B.2)} \qquad \Theta_{\mathcal{C}}(A,t) &= \bigcup_{\lambda \ge 0} \{B - \lambda A : B \in \mathcal{C} \cap \mathcal{B}^{nm}(\lambda A, t)\} \\ &= \bigcup_{\lambda \ge 0} \Big(\mathcal{C} \cap \mathcal{B}^{nm}(\lambda A, t) - \lambda A\Big).\end{aligned}$$

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

Lemma B.4. *Given $A \in \mathbb{R}^{n\times m}$ and $\mathbf{l} = (l_1, \dots, l_m) \in [n]^m$, define $k(A_{\cdot,j}) = \mathsf{card}(\{A_{1,j}, \dots, A_{n,j}\})$ and $K(A) = \sum_{j=1}^m k(A_{\cdot,j})$. Then for any $t > 0$ and $\varepsilon > 0$,*

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

Proof. Assume that $\varepsilon \le t$ 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 $K(A) \le \varkappa \le 2K(A)$. 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 B.3 implies that

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

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 B.2 and the previous display,

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

where we used the concavity of the logarithm and Jensen's inequality in the second step, and that $K(A) \le \varkappa \le 2K(A)$ 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 $\lambda \ge 0$. In view of Definition (B.2), it holds

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

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. □

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 B.5. *Let $A \in \mathbb{R}^{n\times m}$ and $K(A)$ be defined as in the previous lemma. Then for any $\varepsilon > 0$ and $t > 0$,*

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

Proof. Assume that $\varepsilon \le t$ since otherwise the left-hand side is zero and the bound holds trivially. Note that $\mathcal{U}^m = \bigcup_{\mathbf{l}\in[n]^m} \mathcal{C}^m_{\mathbf{l}}$, and that $\mathcal{M} = \bigcup_{\Pi\in\mathfrak{S}_n} \Pi\mathcal{U}^m$. Thus $\mathcal{M}$ is the union of $n^m n!$ cones of the form $\Pi\mathcal{C}^m_{\mathbf{l}}$. By Definition (B.2), $\Theta_{\mathcal{M}}(A,t)$ is also the union of $n^m n!$ sets $\Theta_{\Pi\mathcal{C}^m_{\mathbf{l}}}(A,t)$, each having metric entropy subject to the bound in Lemma B.4. Therefore, a union bound implies that

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

where the last step follows from that $K\log(enm/K) \ge m\log n$ for $m \le K \le nm$ and that $\varepsilon \le t$. □

APPENDIX C: 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. $N(0,\sigma^2)$. The Varshamov-Gilbert lemma [Mas07, Lemma 4.7] is a standard tool for proving lower bounds.

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

We also need the following useful lemma.

Lemma C.2. *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 $|\Theta| \ge 3$ and for distinct $\theta,\theta' \in \Theta$, $4\phi \le \|\theta-\theta'\|_2^2 \le \frac{\sigma^2}{8}\log|\Theta|$ where $\phi > 0$. Then there exists $c > 0$ such that*

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

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

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

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

C.1. Proof of Theorem 3.5. We define $\mathcal{U}^m_{K_0}(V_0) = \mathcal{U}^m_{K_0} \cap \mathcal{U}^m(V_0)$ and $\mathcal{M}_{K_0}(V_0) = \bigcup_{\Pi \in \mathfrak{S}_n} \Pi \mathcal{U}^m_{K_0}(V_0)$. Define the subset of $\mathcal{M}_{K_0}(V_0)$ containing permutations of monotonic matrices by $\mathcal{M}^{\mathcal{S}}_{K_0}(V_0) = \{\Pi A \in \mathcal{M}_{K_0}(V_0) : \Pi \in \mathfrak{S}_n, A \in \mathcal{S}^m\}$. 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 stronger lower bound than the one in Theorem 3.5.

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

$$\inf_{\hat M} \sup_{M \in \mathcal{M}_{K_0}(V_0)} \mathbb{P}_M \Big[\frac{1}{nm} \|\hat M - M\|_F^2 \ge c\sigma^2 \frac{K_0}{nm} + c \max_{1 \le l \le \min(K_0 - m, m)+1} \min\Big(\frac{\sigma^2}{m}\log l, m^2 l^{-3} V_0^2\Big)\Big] \ge c' \tag{C.1}$$

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* $\mathcal{M}^{\mathcal{S}}_{K_0}(V_0)$ *if* $l = 1$ *or* 2.

Note that the bound clearly holds for the larger parameter space $\mathcal{M}_{K_0} = \bigcup_{\Pi \in \mathfrak{S}_n} \Pi \mathcal{U}^m_{K_0}$. By taking $l = \min(K_0 - m, m) + 1$ and V_0 large enough, we see that the assumption in Proposition C.3 is satisfied and the second term becomes simply $\frac{\sigma^2}{m} \log l$, so Theorem 3.5 follows. In the monotonic case, by the last statement of the proposition, if $K_0 \ge m+1$ 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 $K(A) \le K_0$.

The proof of Proposition C.3 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 $1 \le l \le \min(K_0 - m, m) + 1$, which is due to the unknown permutation.

Lemma C.4. *Suppose that* $K_0 \le m(\frac{16n}{\sigma^2})^{1/3} V_0^{2/3} - m$. *For some* $c, c' > 0$,

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

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

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

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

we can choose $k_j \in [n]$ so that $k_j \leq (\frac{16n}{\sigma^2})^{1/3} V_j^{2/3}$ and $K_0 = \sum_{j=1}^m k_j$. According to Lemma C.1, 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

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

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 C.2 completes the proof. □

For the second term in (C.1), 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 C.5. *There exist constants $c, c' > 0$ such that for any $K_0 \geq m+1$ and $V_0 \geq 0$,*

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

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

Proof. By Lemma C.1, 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 $\alpha = \min\big(\frac{\sigma}{8}, m^{3/2} V_0\big)$. Then

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

3. $\|M^\omega - M^{\omega'}\|_F^2 \le \frac{\sigma^2}{64}\delta(\omega,\omega') \le \frac{\sigma^2}{64} n \le \frac{\sigma^2}{8}\log|\Omega|$ for $\omega, \omega' \in \Omega$.

Applying Lemma C.2 completes the proof. □

For the previous two lemmas, we have only used matrices with increasing columns. However, to achieve the second term in (C.1) for $l \ge 3$, we need matrices with unimodal columns. The following packing lemma is the key.

Lemma C.6. *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 $|\mathcal{P}| \ge l^{n-k}(\frac{k}{en})^k$ if $k \ge 1$ 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 $\|M - M_0\|_F \le \varepsilon\sqrt{n}$ 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 $k \ge 1$ then

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

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

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

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

For notational simplicity, we now consider $2 \le l \le \min(K_0 - m, m)$ instead of $3 \le l \le \min(K_0 - m, m) + 1$.

Lemma C.7. *There exist constants $c, c' > 0$ such that for any $K_0 \ge m$, $V_0 \ge 0$ and $2 \le l \le \min(K_0 - m, m)$,*

$$\inf_{\hat{M}} \sup_{M \in \mathcal{M}_{K_0}(V_0)} \mathbb{P}_M \Big[\|\hat{M} - M\|_F^2 \ge cn \min\big(\sigma^2 \log(l+1), m^3(l+1)^{-3} V_0^2\big) \Big] \ge c' ,$$

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

Proof. Set $\varepsilon = 1/2$ and let $\mathcal{P}$ be the $\sqrt{n}/2$-packing given by Lemma C.6. 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 $l \geq 2$,

$$\log|\mathcal{P}| \geq \frac{7n}{8}\log l - \frac{n}{8}\log(8e) \geq \frac{n}{4}\log l\,. \tag{C.2}$$

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

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

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

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

   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

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

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 (C.2),

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

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

Combining Lemma C.4, C.5 and C.7, and dividing the bound by nm, we get (C.1) because the max of two terms is lower bounded by a half of their sum. The last statement in Proposition C.3 holds since Lemma C.4 and C.5 are proved for matrices with increasing columns.

C.2. Proof of Theorem 3.6. The proof will only use Lemma C.4 and C.5, 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.

The last term $\min(\frac{\sigma^2}{m}, m^2 V_0^2)$ is achieved by Lemma C.5, 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 $m(\frac{16n}{\sigma^2})^{1/3} V_0^{2/3} - m \geq 2m$. Setting

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

we see that $K_0 \geq \lfloor \frac{m}{2}(\frac{16n}{\sigma^2})^{1/3} V_0^{2/3} \rfloor$. Lemma C.4 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 C.1 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 C.2 implies a lower bound on $\frac{1}{nm}\|\hat{M} - M\|_F^2$ of rate $\frac{\sigma^2 m}{nm} = \frac{\sigma^2}{n}$.

APPENDIX D: MATRICES WITH INCREASING COLUMNS

For the model $Y = \Pi^* A^* + Z$ where $A^* \in \mathcal{S}^m$ and $Z \sim \mathrm{subG}(\sigma^2)$, a computationally efficient estimator $(\tilde{\Pi}, \tilde{A})$ has been constructed in Section 4 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.

D.1. Projection. 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 $\tilde{O}(\sigma^2 K(A^*))$.

The proof of the following oracle inequality is very similar to that of Theorem 3.1, so we will sketch the proof without providing all the details.

Lemma D.1. *Consider the model* $Y = \Pi^* A^* + Z$ *where* $A^* \in \mathcal{S}^m$ *and* $Z \sim \mathrm{subG}(\sigma^2)$. *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* $1 - e^{-c(n+m)}$,

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

Proof. Assume without loss of generality that $\Pi^* = I_n$. Let $A \in \mathcal{S}^m$ and 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 $\mathcal{S}^m = \mathcal{C}_{\mathbf{l}}^m$ with $\mathbf{l} = (n, \dots, n)$, by Lemma B.4,

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

Using the proof of Lemma A.3, we see that

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

with probability at least $1 - C\exp(-\frac{cs^2}{\sigma^2})$. Then the proof of Theorem 3.1 implies that with probability at least $1 - e^{-c(n+m)}$,

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

Minimizing over $A \in \mathcal{S}^m$ yields the desired result. □

The idea of splitting the error into two terms as in Lemma D.1 has appeared in [SBGW15, CM16].

D.2. Permutation. By virtue of Lemma D.1, 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 4. Recall that for $i, i' \in [n]$,

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

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

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

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

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

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 $\tilde{\pi} : [n] \to [n]$ 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 $\pi^* : [n] \to [n]$ 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 D.2. *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 $\tilde{\pi}^{-1} \circ \pi^*(i) < \tilde{\pi}^{-1} \circ \pi^*(i')$.*

Proof. 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

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

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,

$$|\Delta_Y(\pi^*(i), \pi^*(i')) - \Delta_{A^*}(i, i')| \leq 2\tau. \tag{D.3}$$

Suppose that $\Delta_{A^*}(i, i') \geq 4\tau$. We claim that $s_{\pi^*(i)} < s_{\pi^*(i')}$. If for $l \in [n]$, $\Delta_Y(\pi^*(l), \pi^*(i)) \geq 2\tau$, then $\Delta_{A^*}(l, i) \geq 0$ by (D.3). Since A^* has increasing columns, $\Delta_{A^*}(l, i') \geq 4\tau$. Again by (D.3), $\Delta_Y(\pi^*(l), \pi^*(i')) \geq 2\tau$. By definition (D.2), 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 $\tilde{\pi}$, $\tilde{\pi}^{-1} \circ \pi^*(i) < \tilde{\pi}^{-1} \circ \pi^*(i')$. □

Next, recall that for a matrix $A \in \mathcal{S}^m$, $\mathcal{J}$ denotes the set of pairs of indices $(i,j) \in [n]^2$ such that $A_{i,\cdot}$ and $A_{j,\cdot}$ are not identical. The quantity $R(A)$ is defined by

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

For any nonzero vector $u \in \mathbb{R}^m$, $\|u\|_2^2/\|u\|_\infty^2 \geq 1$ with equality achieved when $\|u\|_0 = 1$, and $\|u\|_2^2/\|u\|_1^2 \geq m^{-1}$ with equality achieved when all entries of u are the same. Hence $R(A) \geq 1$. 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,

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

Intuitively, the quantity $R(A)$ is small if the difference of any two rows of A is either very sparse or very dense.

Lemma D.3. *There is an event $\mathcal{E}$ of probability at least $1 - \delta$ on which*

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

Proof. Throughout the proof, we restrict ourselves to the event $\mathcal{E}$ defined in Lemma D.2. To simplify the notation, we define $\alpha_i = A^*_{\tilde{\pi}^{-1} \circ \pi^*(i),\cdot} - A^*_{i,\cdot}$. Then

$$\|\tilde{\Pi} A^* - \Pi^* A^*\|_F^2 = \sum_{i=1}^n \|A^*_{\tilde{\pi}(i),\cdot} - A^*_{\pi^*(i),\cdot}\|_2^2 = \sum_{i \in I} \|\alpha_i\|_2^2 , \tag{D.4}$$

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

$$\begin{aligned} \|\alpha_i\|_2^2 &= \min\Big(\frac{\|\alpha_i\|_2^2}{\|\alpha_i\|_\infty^2}, \frac{m\|\alpha_i\|_2^2}{\|\alpha_i\|_1^2} \Big) \cdot \max\Big(\|\alpha_i\|_\infty^2, \frac{\|\alpha_i\|_1^2}{m} \Big) \\ &= \min\Big(\frac{\|\alpha_i\|_2^2}{\|\alpha_i\|_\infty^2}, \frac{m\|\alpha_i\|_2^2}{\|\alpha_i\|_1^2} \Big) \cdot \Delta_{A^*}\big(i, \tilde{\pi}^{-1} \circ \pi^*(i)\big)^2 \end{aligned} \tag{D.5}$$

by (D.1).

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 D.2 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 (D.4) and (D.5), we conclude that

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

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

D.3. Proof of Theorem 4.1. The bound is an immediate consequence of Lemma D.1 and Lemma D.3 with $\delta = (nm)^{-C}$ for $C > 0$.

APPENDIX E: UPPER BOUNDS IN A TRIVIAL CASE

In Theorem 3.6, we have observed the term $\frac{\sigma^2}{m} \wedge m^2 V(A)^2$, whereas the LS estimator only has $\frac{\sigma^2}{m} \log n$ in the upper bounds. The next proposition shows that in the case $m^2 V(A)^2 \leq \frac{\sigma^2}{m}$, we can simply use an averaging estimator that achieves the term $m^2 V(A)^2$.

PROPOSITION E.1. *For $Y = \Pi^* A^* + Z$ where $Z \sim \mathrm{subG}(\sigma^2)$, 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,*

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

with probability at least $1 - \exp(-m)$ and

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

PROOF. 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,

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

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

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

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. □

APPENDIX F: 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

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

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

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

$$\frac{1}{n}\|\hat\theta - \theta^*\|_2^2 \lesssim \min_{\theta\in\mathcal{U}} \Big(\frac{1}{n}\|\theta - \theta^*\|_2^2 + \sigma^2 \frac{k(\theta)}{n} \log \frac{en}{k(\theta)}\Big) + \alpha\sigma^2 \frac{\log n}{n} \tag{F.1}$$

and

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

The corresponding bounds in expectation also hold.

PROOF. The proof closely follows that of Theorem 3.1 and Theorem 3.3.

First note that the term $n \log n$ in the bound of Lemma B.5 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 A.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 3.1 gives that with probability at least $1 - C\exp(-\frac{cs^2}{\sigma^2})$,

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

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 (F.1). The corresponding bound in expectation follows from integrating the tail probability as in the proof of Theorem 3.1.

Finally, we can apply the proof of Theorem 3.3 with $m = 1$ to achieve the global bound. □

Note that the bounds in Corollary F.1 match the minimax lower bounds for isotonic regression in [BT15] up to logarithmic factors. Since every monotonic 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 (F.1) was obtained by Bellec in the revision of [Bel15] that was prepared independently and contemporaneously to this paper. Chatterjee and Lafferty also improved their bounds to having optimal exponents [CL15] after the first version of our current paper was posted. Interestingly Bellec employs bounds on the statistical dimension by leveraging results from [ALMT14], 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 $\theta = \theta^*$.

REFERENCES

[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, 12 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. Lokutsievskiĭ, E. V. Malinovskaya, I. F. Podlivaev, G. P. Prokopov, I. D. Sofronov, and R. P. Fedorenko. *Teoreticheskie osnovy i konstruirovanie chislennykh algoritmov zadach matematicheskoi fiziki.* "Nauka", Moscow, 1979.

[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.

[ALMT14] D. Amelunxen, M. Lotz, M. B. McCoy, and J. A. Tropp. Living on the edge: phase transitions in convex programs with random data. *Information and Inference*, 2014.

[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.

[Bel15] P. C. Bellec. Sharp oracle inequalities for least squares estimators in shape restricted regression. *arXiv preprint arXiv:1510.08029*, 2015.

[Bel16] P. C. Bellec. *Private communication*, July 2016.

[BF96] P. J. Bickel and J. 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.

[BMI06] V. Boyarshinov and M. Magdon-Ismail. Linear time isotonic and unimodal regression in the L_1 and L_∞ norms. *J. Discrete Algorithms*, 4(4), 2006.

[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.

[BS67] M. Š. Birman and M. Z. Solomjak. Piecewise polynomial approximations of functions of classes ${W_p}^\alpha$. *Mat. Sb. (N.S.)*, 73 (115):331–355, 1967.

[BS98] R. Bro and N. Sidiropoulos. Least squares algorithms under unimodality and non-negativity constraints. *J. Chemometrics*, 12:223–247, 1998.

[BT15] P. Bellec and A.B. Tsybakov. Sharp oracle bounds for monotone and convex regression through aggregation. *Journal of Machine Learning Research*, 16:1879–1892, 2015.

[CD16] O. Collier and A. S. Dalalyan. Minimax rates in permutation estimation for feature matching. *Journal of Machine Learning Research*, 17(6):1–32, 2016.

[CGS15a] Sabyasachi Chatterjee, Adityanand Guntuboyina, and Bodhisattva Sen. On matrix estimation under monotonicity constraints. *arXiv preprint arXiv:1506.03430*, 2015.

[CGS15b] 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.

[Cha14] Sourav Chatterjee. A new perspective on least squares under convex constraint. *Ann. Statist.*, 42(6):2340–2381, 12 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.

[CM16] Sabyasachi Chatterjee and Sumit Mukherjee. On estimation in tournaments and graphs under monotonicity constraints. *arXiv preprint arXiv:1603.04556*, 2016.

[CRPW12] V. Chandrasekaran, B. Recht, P. A. Parrilo, and A. S. Willsky. The convex geometry of linear inverse problems. *Found. Comput. Math.*, 12(6):805–849, 2012.

[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, Philadelphia, PA, USA, 2012. Society for Industrial and Applied Mathematics.

[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, Philadelphia, PA, USA, 2013. Society for Industrial and Applied Mathematics.

[DM59] D. Davidson and J. 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.

[DRXZ14] D. Dai, P. Rigollet, L. Xia, and T. Zhang. Aggregation of affine estimators. *Electron. J. Statist.*, 8(1):302–327, 2014.

[EL00] P. P. B. Eggermont and V. N. LaRiccia. Maximum likelihood estimation of smooth monotone and unimodal densities. *Ann. Statist.*, 28(3), 2000.

[FG64] D. R. Fulkerson and O. A. Gross. Incidence matrices with the consecutive 1's property. *Bull. Amer. Math. Soc.*, 70:681–684, 1964.

[Fis73] P. C. Fishburn. Binary choice probabilities: on the varieties of stochastic transitivity. *Journal of Mathematical psychology*, 10(4), 1973.

[FJBd13] Fajwel Fogel, Rodolphe Jenatton, Francis Bach, and Alexandre d'Aspremont. Convex relaxations for permutation problems. In C.J.C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 1016–1024. Curran Associates, Inc., 2013.

[Fri86] M. Frisen. Unimodal regression. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 35(4):479–485, 1986.

[GG12] Thomas L. Gertzen and Martin Grötschel. Flinders Petrie, the travelling salesman problem, and the beginning of mathematical modeling in archaeology. *Doc. Math.*, X(Extra volume: Optimization stories):199–210, 2012.

[GLZ15] Chao Gao, Yu Lu, and Harrison H. Zhou. Rate-optimal graphon estimation. *Ann. Statist.*, 43(6):2624–2652, 12 2015.

[GS90] Z. Geng and N. Z. Shi. Algorithm as 257: Isotonic regression for umbrella orderings. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 39(3):397–402, 1990.

[HKZ12] D. Hsu, S. Kakade, and T. Zhang. A tail inequality for quadratic forms of subgaussian random vectors. *Electron. Commun. Probab.*, 17, 2012.

[KBI14] C. Köllmann, B. Bornkamp, and K. Ickstadt. Unimodal regression using Bernstein-Schoenberg splines and penalties. *Biometrics*, 70(4), 2014.

[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] D. G. Kendall. A mathematical approach to seriation. *Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences*, 269(1193):pp. 125–134, 1970.

[Ken71] David G. Kendall. Abundance matrices and seriation in archaeology. *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete*, 17:104–112, 1971.

[KT61] A. N. Kolmogorov and V. M. Tihomirov. ε-entropy and ε-capacity of sets in functional space. *Amer. Math. Soc. Transl. (2)*, 17, 1961.

[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.

[LW14] Cong Han Lim and Stephen Wright. Beyond the birkhoff polytope: Convex relaxations for vector permutation problems. In Z. Ghahramani, M. Welling, C. Cortes, N.D. Lawrence, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 2168–2176. Curran Associates, Inc., 2014.

[Mam91] Enno Mammen. Estimating a smooth monotone regression function. *Ann. Statist.*, 19(2):724–740, 06 1991.

[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.

[Men15] S. Mendelson. Learning without concentration. *J. ACM*, 62(3), June 2015.

[MvdG97] Enno Mammen and Sara van de Geer. Locally adaptive regression splines. *Ann. Statist.*, 25(1):387–413, 02 1997.

[MW15] Zongming Ma and Yihong Wu. Computational barriers in minimax submatrix detection. *Ann. Statist.*, 43(3):1089–1116, 06 2015.

[NPT85] A. S. Nemirovskiĭ, B. T. Polyak, and A. B. Tsybakov. The rate of convergence of nonparametric estimates of maximum likelihood type. *Problemy Peredachi Informatsii*, 21(4):17–33, 1985.

[Pet99] W. M. Flinders Petrie. Sequences in prehistoric remains. *The Journal of the Anthropological Institute of Great Britain and Ireland*, 29(3/4):pp. 295–301, 1899.

[RWD88] T. Robertson, F.T. Wright, and R. Dykstra. *Order Restricted Statistical Inference.* Probability and Statistics Series. Wiley, 1988.

[SBGW15] N. B. Shah, S. Balakrishnan, A. Guntuboyina, and M. J. Wainright. Stochastically transitive models for pairwise comparisons: Statistical and computational issues. *arXiv preprint arXiv:1510.05610*, 2015.

[SBW16] N. B. Shah, S. Balakrishnan, and M. J. Wainwright. Feeling the bern: Adaptive estimators for bernoulli probabilities of pairwise comparisons. *arXiv preprint arXiv:1603.06881*, 2016.

[Sto08] Q. F. Stout. Unimodal regression via prefix isotonic regression. *Comput. Statist. Data Anal.*, 53(2):289–297, 2008.

[SZ01] J.M. Shoung and C.H. 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.

[Tsy09] A.B. Tsybakov. *Introduction to Nonparametric Estimation.* Springer Series in Statistics. Springer, 2009.

[vdG90] S. 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] S. van de Geer. Hellinger-consistency of certain nonparametric maximum likelihood estimators. *Ann. Statist.*, 21(1), 1993.

[vH14] R. van Handel. Probability in high dimension. Lecture Notes (Princeton University), 2014.

[Zha02] Cun-Hui Zhang. Risk bounds in isotonic regression. *Ann. Statist.*, 30(2):528–555, 04 2002.