Bernoulli **26**(4), 2020, 3051–3080
https://doi.org/10.3150/20-BEJ1215

Estimation of Monge matrices

JAN-CHRISTIAN HÜTTER[1], CHENG MAO[2], PHILIPPE RIGOLLET[3] and ELINA ROBEVA[4]

[1]*Broad Institute, 415 Main Street, Cambridge, MA, 02142, USA. E-mail: jhuetter@broadinstitute.org*
[2]*School of Mathematics, Georgia Institute of Technology, Suite 117, 686 Cherry Street, Atlanta, GA, 30332-0160, USA. E-mail: cheng.mao@math.gatech.edu*
[3]*Department of Mathematics, Massachusetts Institute of Technology, Building 2, Room 106, 77 Massachusetts Avenue, Cambridge, MA, 02139-4307, USA. E-mail: rigollet@math.mit.edu*
[4]*Department of Mathematics, University of British Columbia, Room 121, 1984 Mathematics Road, Vancouver, BC, V6T 1Z2, Canada. E-mail: erobeva@math.ubc.ca*

Monge matrices and their permuted versions known as pre-Monge matrices naturally appear in many domains across science and engineering. While the rich structural properties of such matrices have long been leveraged for algorithmic purposes, little is known about their impact on statistical estimation. In this work, we propose to view this structure as a shape constraint and study the problem of estimating a Monge matrix subject to additive random noise. More specifically, we establish the minimax rates of estimation of Monge and pre-Monge matrices. In the case of pre-Monge matrices, the minimax-optimal least-squares estimator is not efficiently computable, and we propose two efficient estimators and establish their rates of convergence. Our theoretical findings are supported by numerical experiments.

Keywords: constrained least-squares estimation; Monge matrices; permuted matrix estimation; shape-constrained estimation

1. Introduction

A matrix $\theta \in \mathbb{R}^{n_1 \times n_2}$ is called a *Monge matrix* [37] or a *submodular matrix* [61], if

$$\theta_{i,j} + \theta_{k,\ell} \le \theta_{i,\ell} + \theta_{k,j} \quad \text{for all } 1 \le i \le k \le n_1, 1 \le j \le \ell \le n_2. \tag{1.1}$$

In addition, a matrix $\theta \in \mathbb{R}^{n_1 \times n_2}$ is called an *anti-Monge matrix* or a *supermodular matrix* if $-\theta$ is a Monge matrix. The Monge property dates back to Gaspard Monge's work on optimal transport [54]. Since then, it has been widely used and studied in optimization, discrete mathematics and computer science [1,9,12,13,37,64] as it allows for simple and fast algorithms in a variety of instances [10,12,37,59,60]. For example, if the cost matrix in the Hitchcock transportation problem [36] is a Monge matrix, then the so called north-west corner rule produces an optimal solution [12,37]. Other problems which become easier with a Monge cost matrix include, and are far from being limited to, the balanced max-cut problem [60] and the traveling salesman problem [10,59].

Additionally, many of these problems turn out to be invariant under relabeling of the rows and columns of the Monge matrix. It is therefore natural to introduce the following definition. A matrix $\theta \in \mathbb{R}^{n_1 \times n_2}$ is called *pre-Monge* if there exist permutations $\pi_1 : [n_1] \to [n_1]$ and $\pi_2 :$

$[n_2] \to [n_2]$ such that the matrix $\theta(\pi_1, \pi_2)$ defined by

$$\theta(\pi_1, \pi_2)_{i,j} = \theta\big(\pi_1(i), \pi_2(j)\big) \quad \text{for all } (i, j) \in [n_1] \times [n_2],$$

is Monge. Note that the terminology *permuted Monge* has also been used to define the same object [12]. A pre-anti-Monge matrix is defined analogously. Like Monge matrices, pre-Monge matrices have also been studied in the context of optimization [11,14], where the latent permutation yields new computational challenges. For example, even checking whether an $n_1 \times n_2$ matrix is pre-Monge is a nontrivial algorithmic task: Recognition of pre-Monge matrices can be done in $O(n_1 n_2 + n_1 \log n_1 + n_2 \log n_2)$ time [12,23,63], while recognition of pre-Monge matrices with missing entries is NP-complete [22].

1.1. Estimation of (pre-)Monge matrices

The aforementioned combinatorial optimization problems, such as the Hitchcock transportation and the traveling salesman problem, find wide applications in areas such as planning and logistics [12,58]. In these applications, the cost matrix typically consists of production or transportation costs that are approximately estimated in practice. As a result, it is more realistic to assume that the cost matrix is approximately, rather than exactly, (pre-)Monge. This motivates us to take a statistical approach in this work – we model the cost matrix as a (pre-)Monge matrix perturbed by random noise.

Furthermore, despite extensive study of (pre-)Monge matrices in various domains, previous work has focused on the noiseless setting, and existing algorithms typically fail when the cost matrix is not exactly (pre-)Monge. For example, it is known that the north-west corner rule produces an optimal solution to the Hitchcock transportation problem if and only if the cost matrix is exactly Monge (see Theorem 3.1 of [12]). Moreover, for the traveling salesman problem, efficient algorithms are based on pyramidal tours [59] or subtour patching [11], both of which crucially rely on the cost matrix being (pre-)Monge. To alleviate this problem when the cost matrix is a noisy version of a (pre-)Monge matrix θ, we may first use the cost matrix to estimate θ, and then any existing algorithm can be applied on θ downstream. Therefore, it is of practical interest to study estimation of a (pre-)Monge matrix in the presence of noise.

1.2. Geometric interpretation

In addition to the aforementioned applications in combinatorial optimization problems, the Monge property has observed strong ties with geometries of certain datasets, starting with the seminal work of Monge on optimal transport [54]. See also [12] for an example of an distance matrix with the Monge property. We now demonstrate how the Monge property arises in the context of *seriation* [3,29–31,45,46], where the goal is to recover the latent ordering of objects based on pairwise distances or correlations.

Let $X \in \mathbb{R}^{n \times d}$ be a data matrix with rows $x_1^\top, \dots, x_n^\top \in \mathbb{R}^d$. Suppose that

$$(x_{i+1} - x_i)^\top (x_{j+1} - x_j) \geq 0 \quad \text{for all } i, j \in [n-1]. \tag{1.2}$$

In this case, what property does the order-3 cumulant (tensor) have?
Does Monge tensor reflect the ordering properties among cumulant entries

In other words, the differences between consecutive points have a nonnegative correlation. It is then easy to check that the Gram matrix $\theta = XX^\top$ is an anti-Monge matrix, and the distance matrix D, defined by $D_{i,j} = \|x_i - x_j\|_2^2$ for $i, j \in [n]$, is a Monge matrix. Furthermore, in the context of seriation, we do not know the labels of the points a priori, so the Gram matrix and the distance matrix would be pre-anti-Monge and pre-Monge respectively.

sample covariance

Note that the Monge property (1.2) is a *local* condition in the sense that each inequality enforces one pair of consecutive differences to have a nonnegative correlation. Nevertheless, such a local property guarantees a *global* geometric structure of the data. Namely, the n points in fact approximately lie along a common direction, which can be found spectrally. Furthermore, applying principal component analysis on these points easily recovers the latent labeling, thereby solving the (noiseless) seriation problem.[1] This geometric structure, along with the spectral ordering method, is discussed in more detail in Appendix D.

1.3. Our contributions

In this work, we study the estimation of pre-(anti-)Monge matrices under additive sub-Gaussian noise. Statistically, we establish the minimax rates of estimation (up to logarithmic factors) for both Monge and pre-Monge matrices in Sections 2 and 3.1 respectively, where the upper bounds are achieved by the least-squares estimators.

Algorithmically, for estimating pre-Monge matrices, we further introduce two efficient estimators and study their rates of convergence. The Variance Sorting estimator introduced in Section 3.2, as the name suggests, employs second-order information to estimate the latent permutation. In Section 3.3, we study the singular value thresholding estimator based on approximation of pre-Monge matrices by low-rank ones (Proposition 7).

Furthermore, we provide various numerical experiments in Section 4 to corroborate the theoretically established rates of estimation. Using Dykstra's projection algorithm, we give a detailed implementation of the least-squares estimator for (anti-)Monge matrices, which is of practical interest.

1.4. Related work

This work connects to several lines of research.

Total positivity. The Monge property is closely related to the notion of *total positivity* [43]. An entrywise positive matrix $\theta \in \mathbb{R}^{n_1 \times n_2}$ is called totally positive (of order 2), if

multivariate totally positive of order 2 (MTP2)

$$\theta_{i,j}\theta_{k,\ell} \geq \theta_{i,\ell}\theta_{k,j} \quad \text{for all } 1 \leq i \leq k \leq n_1, 1 \leq j \leq \ell \leq n_2.$$

analogy for order K?

Therefore, an entrywise positive matrix θ is totally positive if and only if $\log(\theta)$ is anti-Monge, where $\log(\cdot)$ is applied to each entry of θ individually. As a result, total positivity is also known

[1] For a noisy seriation problem, more realistically we have additive noise on the data matrix X, rather than on the Gram matrix θ as assumed here. However, this is beyond the scope of the current work.

as *log-supermodularity*. Total positivity plays an essential role in statistical physics via the FKG inequality [32] and appears frequently in many other areas of probability and statistics [43,44]. More recently, there have been new developments in studying totally positive distributions and related estimation problems [27,48,62]. In a companion paper [38], we study minimax estimation of a totally positive distribution by employing mathematical tools that are closely related to those in the current paper.

Latent permutation learning. Estimating a pre-Monge matrix from its noisy version falls into the category of matrix learning with latent permutations, which has recently observed a surge of interest. Models involving latent permutations include noisy sorting [8], the strong stochastic transitivity model [17,66], feature matching [20], crowd labeling [67], statistical seriation [29] and graph matching [25,49], to name a few. Many of the previous approaches for learning latent permutations under such models are based on sorting row or column sums of the observed matrix (or equivalently, degrees of vertices) [19,57,68] or certain refinements [51,52]. However, since adding a constant to all entries in a row or column of a Monge matrix does not change its Monge property, first-order information such as row sums is uninformative for the Monge structure, and thus cannot be used to identify the latent permutation. Instead, we propose a new algorithm based on variance sorting. We show in Section 3.2 that this novel use of second order information is decisive when estimating pre-Monge matrices.

Graphon estimation. Another related, substantial body of literature is that on graphon estimation [6,15,33,72], where the goal is to estimate a bivariate function $f : [0,1]^2 \to \mathbb{R}$ from noisy observations of $\{f(X_i, Y_j) : 1 \le i \le n_1, 1 \le j \le n_2\}$. Unlike regression, the design points (X_i, Y_j) are not observed in graphon estimation, so the observations can be viewed as an $n_1 \times n_2$ matrix with latent permutations acting on its rows and columns.

There have been extensive studies on graphon estimation with various structures, including block models [2], smoothness [47] and low-rank structure [65]. The current work can be viewed as a study of denoising observations in graphon estimation with the Monge structure. More precisely, we say that $f : [0,1]^2 \to \mathbb{R}$ is a Monge graphon if

$$f(x_1, y_1) + f(x_2, y_2) \le f(x_1, y_2) + f(x_2, y_1) \quad \text{for all } 0 \le x_1 \le x_2 \le 1, 0 \le y_1 \le y_2 \le 1.$$

Therefore, denoising noisy observations of $\{f(X_i, Y_j) : 1 \le i \le n_1, 1 \le j \le n_2\}$ without the knowledge of the design points (X_i, Y_j) is precisely the problem of estimating a pre-Monge matrix studied in this work. If, for example, smoothness assumptions are imposed on the graphon f in addition, then our results can potentially be used to produce algorithms with theoretical guarantees for estimating f itself.

Shape-constrained estimation. Last but not least, estimation of a Monge matrix falls in the scope of shape-constrained estimation. In a work [71] studying estimation of monotone functions in Gaussian white noise, Wellner proposed the open problem of estimating a bivariate function $f : [-c, c]^2 \to \mathbb{R}$ satisfying

$$f(t_1, t_2) - f(t_1, s_2) - f(s_1, t_2) + f(s_1, s_2) \ge 0 \quad \text{for all } (s_1, s_2), (t_1, t_2) \in [-c, c]^2,$$

which is exactly the anti-Monge property in its continuous form. As the Gaussian white noise model considered in the aforementioned work is asymptotically equivalent to the Gaussian sequence model we adopt in this paper (see [41]), the current work therefore yields rates of estimation for Wellner's open problem.

Shortly before completing the current work, we became aware of a concurrent work by Fang, Guntuboyina and Sen [28] that studies multivariate extensions of isotonic regression. In the two-dimensional case, the notion of *entirely monotone* functions considered there almost coincides with the anti-Monge structure (without permutations) that we study, except that, additionally, the rows and columns of the matrix are assumed to be nondecreasing. Consequently, the rate achieved by the least-squares estimator specialized to dimension two, as expected, coincides with the main term of the rate given by Theorem 1 of our current paper (see also the discussion after Theorem 2). However, it is worth noting that the two proofs follow drastically different paths. While the proof in [28] relies on metric entropy estimates from [5,34], our proof is based on spectral decomposition of the difference operator D defined in (2.1), a technique which has been used for example to study the performance of total variation regularization [40,70]. Moreover, assuming $n = n_1 = n_2$, our upper bound given in Theorem 1 contains a log factor of order $\log(n)$, while the one in Theorem 4.1 of [28] potentially scales like $\log(n)^3$, a minor improvement which nonetheless shows the potential merits of our proof technique.

Another shape-constrained estimation problem related to the present work is the estimation of a bivariate isotonic matrix in Gaussian noise [18]. In fact, every anti-Monge matrix can be written as the sum of a rank-two matrix and a bivariate isotonic matrix (Lemma A.1). However, our results suggest that the set of Monge matrices is in fact qualitatively different from the set of bivariate isotonic matrices. Particularly, the minimax rate of estimation in Theorem 1 is different from that given by Theorem 2.1 of [18], and the low-rank approximation rate in Proposition 7 is different from that given by Lemma 4 of [66].

Notation. For a positive integer n, let $[n] = \{1, 2, \dots, n\}$. For a finite set S, we use $|S|$ to denote its cardinality. For two sequences $\{a_n\}_{n=1}^{\infty}$ and $\{b_n\}_{n=1}^{\infty}$ of real numbers, we write $a_n \lesssim b_n$ if there is a universal constant C such that $a_n \le C b_n$ for all $n \ge 1$. The relation $a_n \gtrsim b_n$ is defined analogously. We use c and C (possibly with subscripts) to denote universal constants that may change from line to line. Let $\wedge$ and $\vee$ denote the min and the max operators between two real numbers respectively. Given a matrix $M \in \mathbb{R}^{n_1 \times n_2}$, we denote its i-th row by $M_{i,\cdot}$ and its j-th column by $M_{\cdot,j}$. We denote by $\|M\|_F$ and $\|M\|$ the Frobenius norm and the operator norm of M, and by $\|M\|_1$ and $\|M\|_\infty$ the ℓ^1 and ℓ^∞-norm of M when viewed as a vector in $\mathbb{R}^{n_1 n_2}$, respectively. We write $M^\dagger$ for the Moore–Penrose pseudoinverse of M. Finally, let $\mathcal{S}_n$ denote the set of permutations $\pi : [n] \to [n]$.

2. Anti-Monge matrix estimation

We start with estimation of a Monge matrix under sub-Gaussian noise, without latent permutations. It is mathematically equivalent to study estimation of an *anti*-Monge matrix $\theta^* \in \mathbb{R}^{n_1 \times n_2}$, which we find more convenient for the presentation. Throughout this work, we assume that $n_1 \ge n_2$ without loss of generality. In the case where $n_1 \le n_2$, our results and proofs remain valid with the roles of n_1 and n_2 swapped.

Consider the difference operator $D \in \mathbb{R}^{(n_1-1)\times n_1}$ defined by

$$D = \begin{bmatrix} -1 & 1 & 0 & \dots & 0 & 0 \\ 0 & -1 & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \dots & -1 & 1 \end{bmatrix}, \tag{2.1}$$

and we define $\tilde{D} \in \mathbb{R}^{(n_2-1)\times n_2}$ in the same way. Using a telescoping sum argument, it is easy to check that the set of anti-Monge matrices θ such that $-\theta$ satisfies (1.1) can be expressed as

$$\mathcal{M} = \mathcal{M}^{n_1,n_2} := \big\{\theta \in \mathbb{R}^{n_1\times n_2} : D\theta\tilde{D}^\top \geq 0\big\},$$

where the symbol $\geq$ denotes entrywise inequality. For each $\theta \in \mathcal{M}$, we define the quantity

$$V(\theta) := \theta_{1,1} + \theta_{n_1,n_2} - \theta_{n_1,1} - \theta_{1,n_2} = \big\|D\theta\tilde{D}^\top\big\|_1, \tag{2.2}$$

where the last equality follows again from a telescoping sum. We remark that $V(\theta)$ is a global seminorm of θ, and turns out to play a role in the rate of estimation.

In this work, we consider additive sub-Gaussian noise. Namely, for a zero-mean random matrix $\varepsilon \in \mathbb{R}^{n_1\times n_2}$, we say that ε is sub-Gaussian with variance proxy σ^2, or simply $\varepsilon \sim \mathsf{subG}_{n_1\times n_2}(\sigma^2)$, if for any matrix $M \in \mathbb{R}^{n_1\times n_2}$, it holds that

$$\mathbb{E}\big[\exp\big(\mathrm{Tr}\big(M^\top\varepsilon\big)\big)\big] \leq \exp\big(\sigma^2\|M\|_F^2/2\big).$$

Suppose that we observe

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

where $\varepsilon \sim \mathsf{subG}_{n_1\times n_2}(\sigma^2)$. We study the performance of the least-squares estimator

$$\hat{\theta}^{\mathsf{ls}} := \operatorname*{argmin}_{\theta\in\mathcal{M}} \|\theta - y\|_F^2, \tag{2.3}$$

in terms of the mean squared error

$$\frac{1}{n_1n_2}\big\|\hat{\theta} - \theta^*\big\|_F^2.$$

Our upper bound is stated in the following theorem.

Theorem 1. *Let $\theta^* \in \mathcal{M}^{n_1,n_2}$ be an anti-Monge matrix, and suppose that we observe $y = \theta^* + \varepsilon$ where $\varepsilon \sim \mathsf{subG}_{n_1\times n_2}(\sigma^2)$. Let the quantity $V(\theta^*)$ be defined by (2.2). Then the least-squares estimator $\hat{\theta}^{\mathsf{ls}}$ achieves the rate*

$$\frac{1}{n_1n_2}\big\|\hat{\theta}^{\mathsf{ls}} - \theta^*\big\|_F^2 \lesssim \left[\frac{\sigma^2}{n_2} + \left(\frac{\sigma^2 V(\theta^*)}{n_1n_2}\right)^{2/3}\log(n_1)^{1/3}\log(n_2)^{2/3}\right] \wedge \sigma^2$$

with probability at least $1 - \exp(-n_1)$. Moreover, the same bound holds in expectation.

Assuming Gaussian noise, the following theorem provides a lower bound that matches the above upper bound up to a logarithmic factor. For $V_0 \geq 0$, let us define

$$\mathcal{M}_{V_0} = \mathcal{M}_{V_0}^{n_1,n_2} := \{\theta \in \mathcal{M}^{n_1,n_2} : V(\theta) \leq V_0\}.$$

Theorem 2. *Consider the model $y = \theta^* + \varepsilon$, where $\theta^* \in \mathcal{M}_{V_0}^{n_1,n_2}$ and ε has i.i.d. $\mathcal{N}(0, \sigma^2)$ entries. For any $V_0 \geq 0$, it holds that*

$$\inf_{\tilde{\theta}} \sup_{\theta^* \in \mathcal{M}_{V_0}} \mathbb{E}\left[\frac{1}{n_1 n_2}\left\|\tilde{\theta} - \theta^*\right\|_F^2\right] \gtrsim \left[\frac{\sigma^2}{n_2} + \left(\frac{\sigma^2 V_0}{n_1 n_2}\right)^{2/3}\right] \wedge \sigma^2,$$

where the infimum is taken over all estimators measurable with respect to the observation y.

We now discuss the minimax rate of estimation with $\sigma = 1$, $n_1 = n_2 = n$ and the logarithmic factor omitted for simplicity. The first term $1/n$ is a boundary term due to the fact that the set of Monge matrices contains all matrices with constant rows or columns. If we further impose boundary conditions on the matrix θ^*, for example, by specifying its row sums and column sums, then this boundary term will vanish. The second term $V(\theta^*)^{2/3}/n^{4/3}$ is attributed to the Monge property and dominates when $V(\theta^*) \geq \sqrt{n}$. This is also the rate achieved by the concurrent work [28] for denoising an entirely monotone function up to a polylogarithmic factor. Moreover, the lower bound of [28], Theorem 4.4, for a class of functions with bounded Hardy–Krause variation consists of a $\log(n)^{2/3}$ factor, so it would be interesting if the logarithmic gap here could also be tightened.

3. Pre-anti-Monge matrix estimation

In this section, we move on to study the estimation of a pre-anti-Monge matrix, that is, an anti-Monge matrix whose rows and columns have been shuffled by latent permutations. Let $\mathcal{S}_n$ denote the set of permutations $\pi : [n] \to [n]$. For any matrix $\theta \in \mathbb{R}^{n_1 \times n_2}$ and permutations $\pi_1 \in \mathcal{S}_{n_1}, \pi_2 \in \mathcal{S}_{n_2}$, recall that $\theta(\pi_1, \pi_2)$ denotes the matrix defined by $\theta(\pi_1, \pi_2)_{i,j} = \theta(\pi_1(i), \pi_2(j))$. Define the sets

$$\mathcal{M}(\pi_1, \pi_2) := \{\theta(\pi_1, \pi_2) : \theta \in \mathcal{M}\} \quad \text{and} \quad \mathcal{M}_{V_0}(\pi_1, \pi_2) := \{\theta(\pi_1, \pi_2) : \theta \in \mathcal{M}_{V_0}\}$$

of anti-Monge matrices shuffled by fixed permutations.

Suppose that we observe

$$y = \theta^*(\pi_1^*, \pi_2^*) + \varepsilon, \tag{3.1}$$

where $(\pi_1^*, \pi_2^*, \theta^*) \in \mathcal{S}_{n_1} \times \mathcal{S}_{n_2} \times \mathcal{M}$ and $\varepsilon \sim \mathsf{subG}_{n_1 \times n_2}(\sigma^2)$. Our goal is to estimate the pre-anti-Monge matrix $\theta^*(\pi_1^*, \pi_2^*)$.

If two rows (or columns) of θ^* differ by a constant vector, then the matrix we obtain from switching these two rows is still anti-Monge. Therefore, even if the noise ε is zero, neither the pair of permutations (π_1^*, π_2^*) nor the matrix θ^* can be inferred from y. As a result, measures

of permutation and estimation errors such as $\|\theta^*(\hat{\pi}_1, \hat{\pi}_2) - \theta^*(\pi_1^*, \pi_2^*)\|_F$ and $\|\hat{\theta} - \theta^*\|_F$, may be not be pertinent. This is why, instead of studying identifiability of the permutations and the anti-Monge matrix, we focus on the denoising error

$$\left\|\tilde{\theta} - \theta^*\left(\pi_1^*, \pi_2^*\right)\right\|_F$$

for any estimator $\tilde{\theta}$ of the pre-anti-Monge matrix.

Depending on the application, it might be important to differentiate between *proper* and *improper* estimators $\tilde{\theta}$. In this context, a proper estimator is an estimator

$$\tilde{\theta} \in \overline{\mathcal{M}} := \bigcup_{\pi_1 \in \mathcal{S}_{n_1}, \pi_2 \in \mathcal{S}_{n_2}} \mathcal{M}(\pi_1, \pi_2),$$

that is, an estimator that needs to be a pre-anti-Monge matrix itself. By contrast, an improper estimator can be any matrix $\tilde{\theta} \in \mathbb{R}^{n_1 \times n_2}$.

The rest of this section is organized as follows. We first establish the minimax rate for estimating a pre-anti-Monge matrix in Section 3.1. It is achieved by the global least-squares estimator, which is proper by nature, but is likely to be computationally infeasible. Next, we give a computationally feasible proper estimator in Section 3.2 under additional assumptions. Finally, in Section 3.3, we present another computationally feasible estimator based on singular value thresholding that yields a better rate than the one in Section 3.2, but may be improper. This presents a shortcoming if one wants to leverage the Monge structure for downstream numerical computations.

3.1. Minimax rates of estimation

We work under the technical assumption that $\theta^* \in \mathcal{M}_{V_0}$ where V_0 is known. Define

$$\overline{\mathcal{M}}_{V_0} := \bigcup_{\pi_1 \in \mathcal{S}_{n_1}, \pi_2 \in \mathcal{S}_{n_2}} \mathcal{M}_{V_0}(\pi_1, \pi_2).$$

Our upper bound is achieved (up to a logarithmic factor) by the global least-squares estimator over the entire parameter space

$$\hat{\theta}^{\mathsf{gls}} \in \operatorname*{argmin}_{\theta \in \overline{\mathcal{M}}_{V_0}} \|\theta - y\|_F^2. \tag{3.2}$$

If the minimizer is not unique, an arbitrary one is chosen.

Theorem 3. *Suppose that we have* $y = \theta^*(\pi_1^*, \pi_2^*) + \varepsilon$, *where* $\theta^* \in \mathcal{M}_{V_0}^{n_1, n_2}$ *and* $\varepsilon \sim \mathsf{subG}(\sigma^2)$. *Then the global least-squares estimator* (3.2) *achieves*

$$\frac{1}{n_1 n_2}\left\|\hat{\theta}^{\mathsf{gls}} - \theta^*\left(\pi_1^*, \pi_2^*\right)\right\|_F^2 \lesssim \left[\frac{\sigma^2 \log(n_1)}{n_2} + \left(\frac{\sigma^2 V_0}{n_1 n_2}\right)^{2/3} \log(n_1)^{1/3} \log(n_2)^{2/3}\right] \wedge \sigma^2$$

with probability at least $1 - n_1^{-n_1}$. *The same bound holds in expectation.*

Note that this rate is the same (up to a logarithmic factor in the first term) as that for estimating an anti-Monge matrix without latent permutations in view of Theorem 1. Therefore, the lower bound of Theorem 2 for the smaller class implies minimax optimality of the above upper bound (up to a logarithmic factor).

Adaptive methods such as selection or aggregation [42,53] may be used to achieve a near-optimal upper bound without the knowledge of V_0. In fact, we conjecture that a comparable bound holds true for the version of the least-squares estimator where the projection onto $\overline{\mathcal{M}}_{V_0}$ is replaced by the unrestricted version $\overline{\mathcal{M}}$, but our current proof technique does not allow us to conclude this. Nevertheless, if we could choose V_0 such that $V_0 \lesssim V(\theta^*) \le V_0$, then our upper bound would be near-optimal at θ^*.

3.2. Efficient estimation via variance sorting

While the global least-squares estimator retains the minimax rate even in the presence of latent permutations, solving the optimization problem (3.2) is unlikely to be computationally efficient. Thus, we now discuss polynomial-time estimators. In this subsection, we assume that the noise matrix ε is homoscedastic with independent sub-Gaussian entries, i.e.,

$$\varepsilon_{i,j} \sim \mathsf{subG}\big(C\sigma^2\big) \quad \text{and} \quad \mathrm{Var}[\varepsilon_{i,j}] = \sigma^2.$$

As in the previous section, the estimator is based on projecting a permuted version of the observations onto $\mathcal{M}_{V_0}$, but we use an efficient method to find estimators of the permutations with respect to which we project on. Let us first focus on estimating the row permutation π_1. Since adding a constant to all entries in a row of the underlying matrix does not change its anti-Monge property, there is no first-order information that helps distinguish between the rows of y. Instead, we exploit second-order information, namely, the variance of row differences of y.

The intuition behind the Variance Sorting subroutine proposed below lies the following lemma, whose proof is deferred to Appendix A.8.2.

Lemma 4. *For an anti-Monge matrix $\theta^* \in \mathcal{M}^{n_1,n_2}$, the variance between row 1 and row i of θ^*, defined as*

$$\sum_{k=1}^{n_2}\left[\theta^*_{i,k} - \theta^*_{1,k} - \frac{1}{n_2}\sum_{\ell=1}^{n_2}\big(\theta^*_{i,\ell} - \theta^*_{1,\ell}\big)\right]^2 \quad \text{for } i \in [n_1],$$

is monotonically nondecreasing in i.

Therefore, if we knew the index $\pi_1^{-1}(1)$ corresponding to the first row of θ^*, we could estimate these variances by their empirical versions (defined in (3.3) below) and sort the rows accordingly. The precise method is given in the following Variance Sorting Subroutine.

Note that in Algorithm 1, the pair (i_0, j_0) defined in (3.4) is an estimator for the extremal rows $\pi_1^{-1}(1)$ and $\pi_1^{-1}(n_1)$, but the choice of which index corresponds to $\pi_1^{-1}(1)$ is broken arbitrarily by the constraint $i_0 < j_0$. In turn, the resulting estimator $\hat{\pi}_1$ can only be reliable up to a global flip of the coordinates. In order to obtain denoising rates, this indeterminacy can be overcome

Algorithm 1 Variance sorting

1. For each pair of rows (i, j) of y, compute the variance of their difference

$$\xi(i,j) := \sum_{k=1}^{n_2} \left[y_{i,k} - y_{j,k} - \frac{1}{n_2} \sum_{\ell=1}^{n_2} (y_{i,\ell} - y_{j,\ell}) \right]^2, \tag{3.3}$$

and define

$$(i_0, j_0) := \operatorname*{argmax}_{(i,j)\in[n_1]^2, i<j} \xi(i,j). \tag{3.4}$$

2. Define $\hat{\pi}_1 \in \mathcal{S}_{n_1}$ so that $\{\xi(i_0, \hat{\pi}_1^{-1}(i))\}_{i=1}^{n_1}$ is nondecreasing in i. In particular, we can pick $\hat{\pi}_1(1) = i_0$ and $\hat{\pi}_1(n_1) = j_0$.

by projecting y onto the set of anti-Monge matrices under both possible orientations and picking the best fit.

To facilitate our presentation, we define the reversal permutation $\pi_1^{\mathrm{r}} \in \mathcal{S}_{n_1}$ by $\pi_1^{\mathrm{r}}(i) = n_1 - i + 1$ for $i \in [n_1]$, and define similarly $\pi_2^{\mathrm{r}} \in \mathcal{S}_{n_2}$ by $\pi_2^{\mathrm{r}}(i) = n_2 - i + 1$ for $i \in [n_2]$. In short, Algorithm 2 below applies the Variance Sorting subroutine twice to estimate both row and column permutations, and then estimates θ by the (computationally efficient) least-squares estimator in the convex set of anti-Monge matrices along these estimated permutations.

Note that we only allowed a potential flip π_1^{r} for $\hat{\pi}_1$, although there is also such an ambiguity for $\hat{\pi}_2$. This suffices because if $\theta \in \mathcal{M}_{V_0}$, then $\theta(\pi_1^{\mathrm{r}}, \pi_2^{\mathrm{r}}) \in \mathcal{M}_{V_0}$, and as a result

$$\begin{aligned}
&\mathcal{M}_{V_0}(\hat{\pi}_1, \hat{\pi}_2) \cup \mathcal{M}_{V_0}\big(\pi_1^{\mathrm{r}} \circ \hat{\pi}_1, \hat{\pi}_2\big) \\
&\quad = \mathcal{M}_{V_0}(\hat{\pi}_1, \hat{\pi}_2) \cup \mathcal{M}_{V_0}\big(\pi_1^{\mathrm{r}} \circ \hat{\pi}_1, \hat{\pi}_2\big) \cup \mathcal{M}_{V_0}\big(\hat{\pi}_1, \pi_2^{\mathrm{r}} \circ \hat{\pi}_2\big) \cup \mathcal{M}_{V_0}\big(\pi_1^{\mathrm{r}} \circ \hat{\pi}_1, \pi_2^{\mathrm{r}} \circ \hat{\pi}_2\big).
\end{aligned}$$

Algorithm 2 Main algorithm

1. Find $\hat{\pi}_1$ using the Variance Sorting subroutine, Algorithm 1.
2. With y replaced by $y^\top$ and the roles of indices 1 and 2 switched, find $\hat{\pi}_2$ using the Variance Sorting subroutine, Algorithm 1.
3. Compute the least-squares estimator $\hat{\theta}$ as follows. If

$$\min_{\theta\in\mathcal{M}_{V_0}} \big\|\theta(\hat{\pi}_1, \hat{\pi}_2) - y\big\|_F^2 \le \min_{\theta\in\mathcal{M}_{V_0}} \big\|\theta\big(\pi_1^{\mathrm{r}} \circ \hat{\pi}_1, \hat{\pi}_2\big) - y\big\|_F^2,$$

then we define $\hat{\pi}_1' := \hat{\pi}_1$. Otherwise, we define $\hat{\pi}_1' := \pi_1^{\mathrm{r}} \circ \hat{\pi}_1$. Finally, we set

$$\hat{\theta} := \operatorname*{argmin}_{\theta\in\mathcal{M}_{V_0}(\hat{\pi}_1', \hat{\pi}_2)} \|\theta - y\|_F^2.$$

The estimator computed by the main algorithm achieves the following rate of estimation.

Theorem 5. *Suppose that* $y = \theta^*(\pi_1^*, \pi_2^*) + \varepsilon$, *where* $\theta^* \in \mathcal{M}_{V_0}^{n_1,n_2}$ *and* ε *has independent* $\mathsf{subG}(C\sigma^2)$ *entries with variance* σ^2. *Let the estimator* $\hat{\theta}$ *be given by the main algorithm. It holds with probability at least* $1 - n_1^{-n_1}$ *that*

$$\frac{1}{n_1 n_2}\left\|\hat{\theta} - \theta^*\left(\pi_1^*, \pi_2^*\right)\right\|_F^2 \lesssim \left(\sigma^2 + \sigma V_0\right)\left(\frac{\log n_1}{n_2}\right)^{1/2}.$$

Moreover, the same bound holds in expectation.

This rate achieved by our efficient estimator is consistent, but it is suboptimal in view of the minimax rate given by Theorem 3.

3.3. Denoising via singular value thresholding

While the Variance Sorting algorithm above yields efficient estimators of the latent permutations, the rate of convergence it achieves is suboptimal. We now aim for the easier task of *denoising* the pre-anti-Monge matrix without learning the latent permutations, in the hope of obtaining an efficient estimator with a faster rate of convergence. More precisely, under model (3.1), we look for a possibly *improper* estimator $\tilde{\theta} \in \mathbb{R}^{n_1 \times n_2}$ so that $\|\tilde{\theta} - \theta^*(\pi_1^*, \pi_2^*)\|_F^2$ is small.

To this end, we consider the well-studied singular value thresholding (SVT) estimator [17,35]. Let the singular value decomposition of y be

$$y = \sum_{i=1}^{n_2} \lambda_i u_i v_i^\top.$$

Then the SVT (hard-thresholding) estimator is defined as

$$\hat{\theta}^{\mathsf{svt}} := \sum_{i=1}^{n_2} \mathbb{1}\{\lambda_i > \rho\}\lambda_i u_i v_i^\top,$$

where we choose the threshold to be $\rho := C\sigma\sqrt{n_1}$ for a sufficiently large constant $C > 0$. The rate of estimation achieved by the SVT estimator is given in the following theorem.

Theorem 6. *Suppose that we have* $y = \theta^*(\pi_1^*, \pi_2^*) + \varepsilon$, *where* $\theta^* \in \mathcal{M}^{n_1,n_2}$ *and* $\varepsilon \sim \mathsf{subG}(\sigma^2)$. *The singular value thresholding estimator* $\hat{\theta}^{\mathsf{svt}}$ *achieves*

$$\frac{1}{n_1 n_2}\left\|\hat{\theta}^{\mathsf{svt}} - \theta^*\left(\pi_1^*, \pi_2^*\right)\right\|_F^2 \lesssim \left[\frac{\sigma^2}{n_2} + \frac{\sigma^{3/2} V(\theta^*)^{1/2}}{n_2^{3/4}}\right] \wedge \sigma^2$$

with probability at least $1 - \exp(-n_1)$. *The same bound holds in expectation.*

This rate sits between the minimax rate given by Theorem 3, and the rate for the Variance Sorting estimator given by Theorem 5. Note that for this result, the noise ε needs not be homoscedastic, and moreover, no knowledge of V_0 is required, that is, the SVT estimator adapts to the quantity $V(\theta^*)$.

The proof technique leading to upper bounds for the SVT estimator is well developed [17,66]. Our contribution lies in the following low-rank approximation result for an anti-Monge matrix, which is of independent interest.

Proposition 7. *For any $\theta \in \mathcal{M}^{n_1,n_2}$ and positive integer r, there exists a rank-$(3r+3)$ matrix $\tilde{\theta} \in \mathbb{R}^{n_1 \times n_2}$ such that*

$$\|\tilde{\theta} - \theta\|_F^2 \leq 2\frac{n_1 n_2}{r^3} V(\theta)^2.$$

Note that using a similar proof, the same rate as in Theorem 6 can be obtained for a soft-thresholding estimator as well, that is, for

$$\hat{\theta}^{\text{soft}} := \sum_{i=1}^{n_2} \big((\lambda_i - \rho) \vee 0\big) u_i v_i^\top,$$

with a similar scaling for ρ.

As the rate given in Theorem 6 does not match the minimax rate, it is natural to ask whether this suboptimality is an artifact of the proof or a true weakness of the SVT estimator. In Appendix C, we present a worst-case anti-Monge matrix which cannot be approximated by any low-rank matrix at a rate better than that given by Proposition 7. This in turn gives evidence that the rate of convergence for the SVT estimator in Theorem 6 might be the best achievable by this method. However, we do not have a rigorous proof for this statement, nor do we know whether the gap between the rates of Theorems 6 and 3 is truly a statistical-to-computational gap as those in other average-case problems with hidden structures [4,50].

4. Numerical experiments

In order to compare our theoretical guarantees with the empirical performance of the proposed estimators, we conducted experiments on synthetic data, using Dykstra's algorithm to project onto the cone of anti-Monge matrices.

We first present this projection algorithm in Section 4.1. We then show the experimental results of the projection onto the cone of anti-Monge matrices in Section 4.2 and of the two efficient strategies for denoising pre-anti-Monge matrices in Section 4.3.

4.1. Dykstra's algorithm for projecting onto the set of anti-Monge matrices

Since the set $\mathcal{M}$ is a convex cone specified by $O(n_1 n_2)$ constraints, the least-squares estimator (2.3) can be calculated by a general purpose convex optimization software such as SCS [55,56] or ECOS [26]. The most computationally intensive subroutine of these methods is usually solving

Algorithm 3 Dykstra's algorithm

Input: $y \in \mathbb{R}^d$, the point to project; $\mathcal{M}_1, \ldots, \mathcal{M}_m$ a collection of cones
Output: θ, an approximation to the projection of y onto $\mathcal{M}_1 \cap \cdots \cap \mathcal{M}_m$

```
function ProjectDykstra(y)
    for i = 1, ..., m do
        p_i = 0_d                                   ▷ Initialize residuals
    end for
    θ_m = y                                         ▷ Initialize iterates
    while not converged do
        for i = 1, ..., m do
            θ_i ← Π_{M_i}(θ_{(i−2)%m+1} + p_i)      ▷ Project shifted iterates
            p_i ← θ_{(i−2)%m+1} + p_i − θ_i          ▷ Compute new residual
        end for
    end while
    return θ
end function
```

linear systems associated with the constraints specifying $\mathcal{M}$. Using direct methods to find these solutions results in a runtime that scales like $(n_1 n_2)^3$, rendering calculations relatively slow even for moderate values of n_1 and n_2. Hence, we chose to implement a specialized algorithm to calculate θ based on Dykstra's projection algorithm [7,21].

In its general form (see Algorithm 3), this algorithm is designed to calculate the projection of a vector $y \in \mathbb{R}^d$ onto the intersection of m convex sets $\mathcal{M}_1, \ldots, \mathcal{M}_m$ by iteratively projecting carefully chosen points to each individual set. This is similar to alternate projections of a point to each of the sets $\mathcal{M}_1, \ldots, \mathcal{M}_m$, but when initialized with $y \in \mathbb{R}^d$, Dykstra's algorithm not only finds a point in the intersection $\bigcap_{j\in[m]} \mathcal{M}_j$, but its iterates actually converge to the projection of y onto $\bigcap_{j\in[m]} \mathcal{M}_j$.

To apply Dysktra's algorithm to the task of projecting onto the cone of anti-Monge matrices, note that we can write $\mathcal{M} = \bigcap_{i_1=1}^{n_1-1} \bigcap_{i_2=1}^{n_2-1} \mathcal{M}_{i_1,i_2}$ with

$$\mathcal{M}_{i_1,i_2} := \left\{ \theta \in \mathbb{R}^{n_1,n_2} : \sum_{j_1\in\{0,1\}, j_2\in\{0,1\}} (-1)^{j_1+j_2} \theta_{i_1+j_1,i_2+j_2} \geq 0 \right\},$$

because a matrix is anti-Monge if and only if each contiguous 2×2 submatrix is anti-Monge. The projection of y onto $\mathcal{M}_{i_1,i_2}$ can be explicitly calculated to be the matrix with entries

$$\begin{aligned} &\left[\Pi_{\mathcal{M}_{i_1,i_2}}(y)\right]_{i_1+j_1,i_2+j_2} \\ &\quad = y_{i_1+j_1,i_2+j_2} + \frac{(-1)^{j_1+j_2}}{4} \max\left\{ - \sum_{k_1\in\{0,1\}, k_2\in\{0,1\}} (-1)^{k_1+k_2} y_{i_1+k_1,i_2+k_2}, 0 \right\} \end{aligned}$$

for $j_1, j_2 \in \{0, 1\}$, and

$$\left[\Pi_{\mathcal{M}_{i_1,i_2}}(y)\right]_{\ell_1,\ell_2} = y_{\ell_1,\ell_2},$$

Algorithm 4 Fast projection onto $\mathcal{M}$

Input: $y \in \mathbb{R}^{n_1 \times n_2}$
Output: $\theta \approx \Pi_{\mathcal{M}}(y)$
  function PROJETANTIMONGE(y)
    $\eta \leftarrow 0 \in \mathbb{R}^{(n_1-1)\times(n_2-1)}$ ▷ Initialize residuals
    $\theta \leftarrow y$, ▷ Initialize iterates
    while not converged **do**
      for $i_1 = 1, \dots, n_1 - 1, i_2 = 1, \dots, n_2 - 1$ **do**
        $\tilde{\eta} \leftarrow \max\{-\sum_{j_1 \in \{0,1\}, j_2 \in \{0,1\}} (-1)^{j_1+j_2} \theta_{i_1+j_1, i_2+j_2}/4 + \eta_{i_1,i_2}, 0\}$
        ▷ Compute new residuals
        for $j_1 \in \{0, 1\}$, $j_2 \in \{0, 1\}$ **do**
          $\theta_{i_1+j_1,i_2+j_2} \leftarrow \theta_{i_1+j_1,i_2+j_2} + (-1)^{j_1+j_2}(\tilde{\eta} - \eta_{i_1,i_2})$
          ▷ Project shifted iterates
        end for
        $\eta_{i_1,i_2} \leftarrow \tilde{\eta}$ ▷ Store residuals
      end for
    end while
    return θ
  end function

if $(\ell_1, \ell_2) \notin (i_1 + \{0, 1\}) \times (i_2 + \{0, 1\})$.

This leads to Algorithm 4 for projecting a matrix $y \in \mathbb{R}^{n_1 \times n_2}$ onto $\mathcal{M}$.

The rate of convergence of Dykstra's method can be shown to be linearly exponential in the iterations [24], that is, if we denote by $\theta^{(k)}$ the kth iterate of θ in Algorithm 4 and by $\theta^* = \Pi_{\mathcal{M}}(y)$, then $\|\theta^{(k)} - \theta^*\|_2 \lesssim c^k$ for a constant $c < 1$. However, note that the constant c may get closer to one with increasing n_1 and n_2, which is the case for isotonic regression as shown in [24] and matches our experience: simulations for larger values of n_1 and n_2 require more iterates before convergence. On the other hand, if the noise level is low, Algorithm 3 allows us to run simulations with up to $n_1 = n_2 = 700$ below, highlighting its practical applicability on large-scale data.

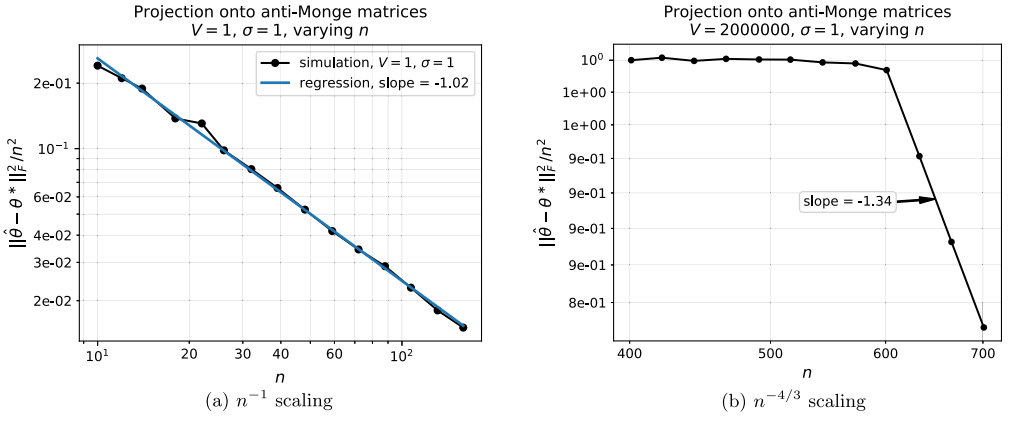
In practice, convergence in Algorithm 4 can be checked by evaluating a measure of feasibility such as $0 \wedge \min_{i,j}(D\theta \tilde{D})_{i,j}$, or by checking when the distance between two successive iterates is small.

4.2. Experiments for anti-Monge matrices

In the following two sections, we assume $n = n_1 = n_2$ for simplicity.

For the estimation of anti-Monge matrices, we consider the following family of ground truth signals, motivated by the construction of the lower bounds in the proof of Theorem 2. First, for $n \in \mathbb{N}$ and $V, \sigma > 0$, define $\theta_{1,(n)} \in \mathbb{R}^{n \times n}$ as

$$(\theta_{1,(n)})_{i,j} = \frac{V}{\lfloor k \rfloor^2} \left\lfloor \frac{(i-1)k}{n-1} \right\rfloor \left\lfloor \frac{(j-1)k}{n-1} \right\rfloor, \quad i \in [n], j \in [n],$$



(a) n^{-1} scaling (b) $n^{-4/3}$ scaling

Figure 1. Varying n for projection onto $\mathcal{M}$. When an arrow is present, "slope" indicates the slope between two consecutive points.

where $k = (Vn/\sigma)^{1/3}$. The ground truth $\theta^*_{1,(n)}$ is obtained by centering $\theta_{1,(n)}$ to have zero column and row sums. Finally, we set $y = \theta^*_{1,(n)} + \varepsilon$ where $\varepsilon_{i,j} \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma)$ and report the average denoising error $\|\theta^{\text{ls}} - \theta^*_{1,(n)}\|_F^2/n^2$ over 20 repetitions. For all experiments, Algorithm 4 is used to compute the projection onto $\mathcal{M}$ and we observe that convergence is fast provided the noise level is low, allowing us to consider examples with up to $n = 700$ within a few minutes.

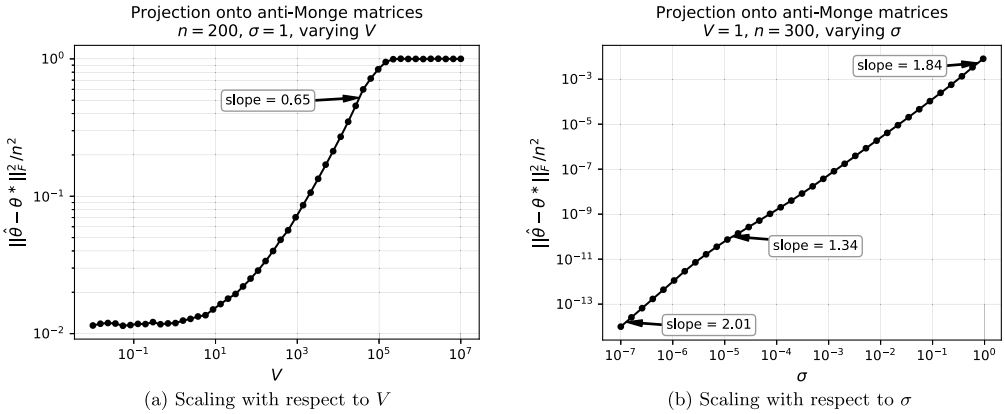
Our simulations recover the three regimes for n that appear in Theorem 1, although at different signal-to-noise ratios governed by V/σ. Namely, on the one hand, for $V = \sigma = 1$, we see in Figure 1(a) an error decay of $n^{-1.02} \approx n^{-1}$ for n between 10 and 160, obtained by linearly regressing the logarithm of the errors onto the logarithm of the n values. On the other hand, for $V = 2 \cdot 10^6$, we can see both a plateau when the trivial σ^2 error bound in Theorem 1 is active, as well as a decay of $n^{-1.34} \approx n^{-4/3}$ at the beginning of the decay becoming effective, where the slope in the doubly logarithmic plot is read off between two consecutive points as indicated in Figure 1(b).

Similarly, fixing $n = 200$, $\sigma = 1$, and varying V between 10^{-2} and 10^7, we can observe a $V^{0.65} \approx V^{2/3}$ scaling in Figure 2(a). The overall curve is shallower, plateauing both at the far low and high end of V, corresponding to the σ^2/n and σ^2 rates becoming active, respectively.

Finally, in Figure 2(b), when setting $n = 300$, $V = 1$, and varying σ between 10^{-7} and 1, we obtain slopes of $\sigma^{2.01}$ and $\sigma^{1.84}$ on the low and high end, while the lowest slope between consecutive points in the curve is $\sigma^{1.34}$, which matches the theoretical rates of σ^2, σ^2/n and $(V\sigma^2/n^2)^{2/3}$, respectively.

4.3. Experiments for pre-anti-Monge matrices

To illustrate the practical performance of the efficient methods presented for denoising a pre-anti-Monge matrix, Variance Sorting and singular value thresholding (see Sections 3.2 and 3.3,



(a) Scaling with respect to V (b) Scaling with respect to σ

Figure 2. Varying σ, and V individually for projection onto $\mathcal{M}$. When an arrow is present, "slope" indicates the slope between two consecutive points.

respectively), we further perform experiments by using both methods on the following family of ground truth matrices:

$$\theta^*_{2,(n)} = \frac{V}{n-1} D^\dagger \left(D^\dagger\right)^\top.$$

These were chosen because the singular value decay we proved in Proposition 7 is tight for these matrices (see Lemma C.1). By contrast, each ground truth example in the previous subsection, $\theta^*_{1,(n)}$, is a rank-one matrix, and hence should lead to an overall better performance of singular value thresholding that is independent of n.

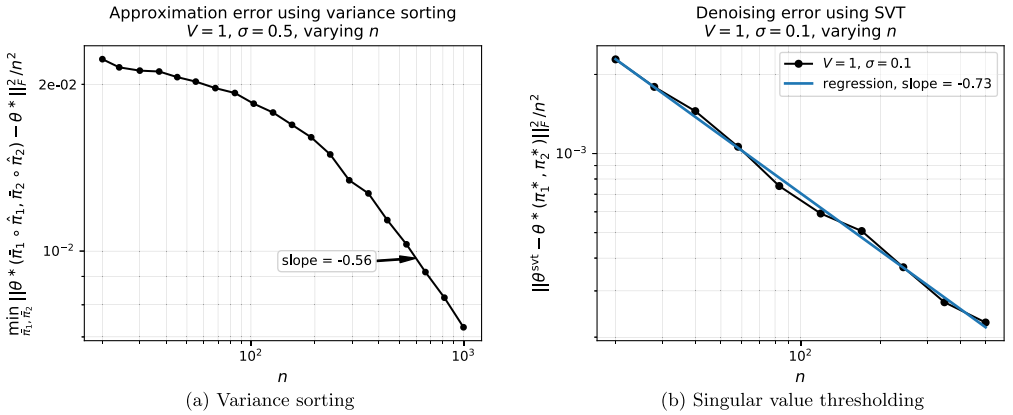
For the Variance Sorting algorithm, we set $V = 1$, $\sigma = 0.5$ and report the approximation error induced by the estimated permutations, that is,

$$\min_{\substack{\pi_1 \in \{\mathsf{id}, \pi_1^\mathsf{r}\} \\ \pi_2 \in \{\mathsf{id}, \pi_2^\mathsf{r}\}}} \frac{1}{n^2} \left\| \theta^*(\pi_1 \circ \hat{\pi}_1, \pi_2 \circ \hat{\pi}_2) - \theta^* \right\|_F^2$$

for $\theta^* = \theta^*_{2,(n)}$, averaged over 256 repetitions. This measure of the approximation quality of the estimated permutations corresponds to the upper bound used in the proof of Proposition A.4 (see (A.8)) and is applicable since by construction, θ^* has row and column sums equal to zero. It is the dominating part in the error analysis, leading to the rate reported in Theorem 5, and it allows us to study a larger range of n, avoiding the need for subsequent projection of the permuted y matrix.

In Figure 3(a), we observe that while for smaller n, we see a slower decay than predicted, for larger n, the decay scales like $n^{-0.47} \approx n^{-1/2}$, close to the predicted rate.

Finally, we perform singular value thresholding on the same set of ground truth matrices, setting $V = 1$, $\sigma = 0.1$, and varying n between 20 and 500. For this experiment, in Figure 3(b),



(a) Variance sorting (b) Singular value thresholding

Figure 3. Algorithms for denoising pre-anti-Monge matrix. When an arrow is present, "slope" indicates the slope between two consecutive points.

we plotted the full denoising error,

$$\frac{1}{n^2}\left\|\hat{\theta} - \theta^*\right\|_F^2,$$

averaged over 64 repetitions. As in the other experiments, we can see an error decay that is close to our theoretical guarantees, that is, $n^{-0.73} \approx n^{-3/4}$.

5. Proofs

We prove Theorem 1 in this section, and defer the remaining proofs of our results to the supplement [39]. Recall that D is defined in (2.1) and $\tilde{D}$ is defined analogously for dimension n_2. In the sequel, whenever we introduce notation in dimension n_1, the analogous object in dimension n_2 is denoted by the same symbol with a tilde.

5.1. Proof of Theorem 1

To analyze the performance of a least-squares estimator, we employ Chatterjee's variational formula [16] that we recall below. See, for example, Lemma 6.1 of [29] for this deterministic form.

Lemma 8 (Chatterjee's variational formula). *Let $\mathcal{M}$ be a closed subset of $\mathbb{R}^d$. Suppose that $y = \theta^* + \varepsilon$ where $\theta^* \in \mathcal{M}$ and $\varepsilon \in \mathbb{R}^d$. Let $\hat{\theta} \in \operatorname{argmin}_{\theta \in \mathcal{M}} \|y - \theta\|_2^2$ be a projection of y onto $\mathcal{M}$. Define the function $f_{\theta^*} : \mathbb{R}_+ \to \mathbb{R}$ by*

$$f_{\theta^*}(t) = \sup_{\theta \in \mathcal{M}, \|\theta - \theta^*\|_2 \le t} \langle \varepsilon, \theta - \theta^* \rangle - \frac{t^2}{2}.$$

Then we have

$$\left\|\hat{\theta} - \theta^*\right\|_2 \in \operatorname*{argmax}_{t \geq 0} f_{\theta^*}(t).$$

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

To control the supremum in Lemma 8, note that it suffices to consider Gaussian noise here, since the generalization to sub-Gaussian noise is taken care of by Theorem B.2.

Proposition 9. *Fix an anti-Monge matrix* $\theta^* \in \mathcal{M}$, *and suppose that* $Z \in \mathbb{R}^{n_1 \times n_2}$ *has i.i.d.* $\mathcal{N}(0,1)$ *entries. Then for any integer* $k \in [n_1 n_2]$ *and any* $t > 0$, *we have*

$$\mathbb{E}\Big[\sup_{\substack{\theta \in \mathcal{M} \\ \|\theta - \theta^*\|_F \leq t}} \langle Z, \theta - \theta^* \rangle\Big] \lesssim t\left[\sqrt{n_1} + \sqrt{k \log(n_2)} + \sqrt{\log(n_1)\log(n_2)}\left(\frac{n_1 n_2}{k}\right)^{1/4}\right] + \sqrt{\frac{n_1 n_2}{k}}\sqrt{\log(n_1)\log(n_2)}\, V(\theta^*).$$

To show Theorem 1 taking Proposition 9 as given, let $t > 0$ and $1 \leq k \leq n_1 n_2$ to be chosen later. Note that by Theorem B.1 and Proposition 9, we obtain

$$\begin{aligned}
&\gamma_2\big(\{\theta - \theta^* : \theta \in \mathcal{M}, \left\|\theta - \theta^*\right\|_F \leq t\}\big) \\
&\asymp \mathbb{E}_{Z_{i,j} \overset{\text{iid}}{\sim} \mathcal{N}(0,1)}\Big[\sup_{\substack{\theta \in \mathcal{M} \\ \|\theta - \theta^*\|_F \leq t}} \langle Z, \theta - \theta^* \rangle\Big] \\
&\lesssim t\left[\sqrt{n_1} + \sqrt{k \log(n_2)} + \sqrt{\log(n_1)\log(n_2)}\left(\frac{n_1 n_2}{k}\right)^{1/4}\right] + \sqrt{\frac{n_1 n_2}{k}}\sqrt{\log(n_1)\log(n_2)}\, V(\theta^*),
\end{aligned}$$

where γ_2 denotes Talagrand's γ_2 functional. Therefore, Theorem B.2 yields that with probability $1 - 4\exp(-s^2)$,

$$\sup_{\substack{\theta \in \mathcal{M} \\ \|\theta - \theta^*\|_F \leq t}} \langle \varepsilon, \theta - \theta^* \rangle \lesssim t\sigma\left[\sqrt{n_1} + \sqrt{k \log(n_2)} + \sqrt{\log(n_1)\log(n_2)}\left(\frac{n_1 n_2}{k}\right)^{1/4}\right] + \sigma\sqrt{\frac{n_1 n_2}{k}}\sqrt{\log(n_1)\log(n_2)}\, V(\theta^*) + \sigma s t.$$

Let us define

$$g(t) := \sup_{\substack{\theta \in \mathcal{M} \\ \|\theta - \theta^*\|_F \leq t}} \langle \varepsilon, \theta - \theta^* \rangle \quad \text{and} \quad f_{\theta^*}(t) := g(t) - \frac{t^2}{2}.$$

Then we obtain that for any fixed

$$t > t_s^* := C\sigma\left[\sqrt{n_1} + \sqrt{k\log(n_2)} + \sqrt{\log(n_1)\log(n_2)}\left(\frac{n_1 n_2}{k}\right)^{1/4}\right] + C\left[\sigma\sqrt{\frac{n_1 n_2}{k}}\sqrt{\log(n_1)\log(n_2)}V\left(\theta^*\right)\right]^{1/2} + C\sigma s,$$

where C is a sufficiently large constant, it holds with probability at least $1 - 4\exp(-s^2)$ that $g(t) < t^2/8$.

Note that this implies $f_{\theta^*}(t) < 0$ for any fixed $t > t_s^*$, but to apply Lemma 8, we would need negativity of $f_{\theta^*}(t)$ simultaneously for all $t > t_s^*$. Towards this end, let us define the event $\mathcal{E}_1 := \{\|\varepsilon\|_F^2 \le 5\sigma^2(n_1 n_2 + s\sqrt{n_1 n_2})\}$. Since $\varepsilon \sim \mathsf{subG}_{n_1\times n_2}(\sigma^2)$, Lemma B.4 implies that $\mathbb{P}\{\mathcal{E}_1\} \ge 1 - 2\exp[-c(s^2 \wedge s\sqrt{n_1 n_2})]$. We define $t_s^{\#} := 5\sigma[\sqrt{n_1 n_2} + \sqrt{s}(n_1 n_2)^{1/4}]$. It is then easily seen that on the event $\mathcal{E}_1$,

$$f_{\theta^*}(t) \le t\|\varepsilon\|_F - t^2/2 < 0$$

for all $t > t_s^{\#}$. Furthermore, if $t_s^* < t_s^{\#}$, we consider a discretization $T = \{t_1, \dots, t_k\}$ of the interval $[t_s^*, t_s^{\#}]$ such that $t_s^* = t_1 < \cdots < t_k = t_s^{\#}$, $2t_i \ge t_{i+1}$ for $i \in [k-1]$, and

$$k \le \log_2\left(t_s^{\#}/t_s^*\right) + 1 \le 5\log(n_1 \vee s).$$

A union bound over $t_i \in T$ then yields that on an event $\mathcal{E}_2$ of probability at least $1 - 20\log(n_1 \vee s)\exp(-s^2)$, we have $g(t_i) < t_i^2/8$ for all $i \in [k]$. Since the function $g(t)$ is nondecreasing, we have that for any $t \in [t_i, t_{i+1}]$,

$$f_{\theta^*}(t) \le g(t_{i+1}) - \frac{t_i^2}{2} \le g(t_{i+1}) - \frac{t_{i+1}^2}{8} < 0.$$

Combining the above results, we see that $f_{\theta^*}(t) < 0$ for all $t \ge t_s^*$ on the event $\mathcal{E}_1 \cap \mathcal{E}_2$.

Therefore by Lemma 8, we obtain that on $\mathcal{E}_1 \cap \mathcal{E}_2$,

$$\begin{aligned}
&\frac{1}{n_1 n_2}\left\|\hat{\theta}^{\mathsf{ls}} - \theta^*\right\|_F^2 \\
&\quad\le \frac{(t_s^*)^2}{n_1 n_2} \\
&\quad\lesssim \sigma^2\left[\frac{1}{n_2} + \frac{k\log(n_2)}{n_1 n_2} + \frac{\log(n_1)\log(n_2)}{\sqrt{n_1 n_2 k}}\right] \\
&\qquad + \sigma\sqrt{\frac{\log(n_1)\log(n_2)}{n_1 n_2 k}}V\left(\theta^*\right) + \sigma^2\frac{s^2}{n_1 n_2}.
\end{aligned} \tag{5.1}$$

In addition, it holds that $\mathbb{P}\{\mathcal{E}_1 \cap \mathcal{E}_2\} \ge 1 - 2\exp[-c(s^2 \wedge s\sqrt{n_1 n_2})] - 20\log(n_1 \vee s)\exp(-s^2)$.

We now choose $s = C\sqrt{n_1}$ for a sufficiently large constant C so that $\mathbb{P}\{\mathcal{E}_1 \cap \mathcal{E}_2\} \geq 1 - \exp(-n_1)$. Balancing the terms in (5.1) that depend on k leads to the choice

$$k^* := (n_1 n_2)^{1/3} \log(n_1)^{1/3} \left[\sqrt{\log(n_1)} + \frac{V(\theta^*)}{\sigma \sqrt{\log(n_2)}} \right]^{2/3},$$

in addition to possibly rounding k^* to an integer which we omitted to simplify the presentation. Therefore, we obtain that with probability $1 - \exp(-n_1)$, if $1 \leq k^* \leq n_1 n_2$, then

$$\begin{aligned} \frac{1}{n_1 n_2} \left\| \hat{\theta}^{\mathsf{ls}} - \theta^* \right\|_F^2 &\lesssim \frac{\sigma^2}{n_2} + \frac{\sigma^2 \log(n_2) \log(n_1)^{1/3} [\sqrt{\log(n_1)} + V(\theta^*)/(\sigma \sqrt{\log(n_2)})]^{2/3}}{(n_1 n_2)^{2/3}} \\ &\lesssim \frac{\sigma^2}{n_2} + \left(\frac{\sigma^2 V(\theta^*)}{n_1 n_2} \right)^{2/3} \log(n_1)^{1/3} \log(n_2)^{2/3}. \end{aligned} \tag{5.2}$$

If $k^* < 1$, we replace it by 1, increasing the $k \log(n_2)/(n_1 n_2)$ term while decreasing the ones with $1/\sqrt{k}$ in (5.1), hence leading to the same rate as in (5.2). If $k^* > n_1 n_2$, note that the $k/(n_1 n_2)$ term is already of the order σ^2, so a basic bound of σ^2 on the empirical process term yields the rate

$$\frac{1}{n_1 n_2} \left\| \hat{\theta}^{\mathsf{ls}} - \theta^* \right\|_F^2 \leq \sigma^2.$$

Combined, this yields that with probability at least $1 - \exp(-n_1)$,

$$\frac{1}{n_1 n_2} \left\| \hat{\theta}^{\mathsf{ls}} - \theta^* \right\|_F^2 \lesssim \left[\frac{\sigma^2}{n_2} + \left(\frac{\sigma^2 V(\theta^*)}{n_1 n_2} \right)^{2/3} \log(n_1)^{1/3} \log(n_2)^{2/3} \right] \wedge \sigma^2.$$

To obtain the bound in expectation, we can first integrate the exponentially decaying tale of (5.1), and then choose the optimal k in the same way.

5.2. Proof of Proposition 9

Our main strategy consists in decomposing the noise matrix Z into three terms according to the spectral decomposition of the linear map $\mathcal{D}$, defined by $\mathcal{D}(A) = DA\tilde{D}^\top$ for $A \in \mathbb{R}^{n_1 \times n_2}$.

Spectral decomposition of the difference operator. Denote the (reduced) singular value decomposition of D by $D = U \Sigma W^\top$, where we order the singular values in Σ in ascending magnitude. In addition, we write $W = [w_1 | \cdots | w_{n_1}]$.

First, let Π_1 denote the projection onto $\ker \mathcal{D}$. Moreover, let $J = \{(l, r) \in [n_1] \times [n_2] : lr \leq k\}$ and $J^c = [n_1] \times [n_2] \setminus J$. Define the projection Π_2 by

$$\Pi_2(A) = \sum_{(l,r) \in J} w_l w_l^\top A \tilde{w}_r \tilde{w}_r^\top \quad \text{and so}$$

$$(I - \Pi_2)(A) = \sum_{(l,r) \in J^c} w_l w_l^\top A \tilde{w}_r \tilde{w}_r^\top.$$

With these two projections, we decompose

$$\begin{aligned}
&\mathbb{E}\Big[\sup_{\substack{\theta\in\mathcal{M}\\ \|\theta-\theta^*\|_F\le t}}\big\langle Z,\theta-\theta^*\big\rangle\Big]\\
&\quad\le \mathbb{E}\Big[\sup_{\substack{\theta\in\mathcal{M}\\ \|\theta-\theta^*\|_F\le t}}\big\langle \Pi_1(Z),\theta-\theta^*\big\rangle\Big]+\mathbb{E}\Big[\sup_{\substack{\theta\in\mathcal{M}\\ \|\theta-\theta^*\|_F\le t}}\big\langle (I-\Pi_1)\Pi_2(Z),\theta-\theta^*\big\rangle\Big]\\
&\qquad+\mathbb{E}\Big[\sup_{\substack{\theta\in\mathcal{M}\\ \|\theta-\theta^*\|_F\le t}}\big\langle (I-\Pi_1)(I-\Pi_2)(Z),\theta-\theta^*\big\rangle\Big]. \qquad (5.3)
\end{aligned}$$

We now bound the three terms in (5.3) separately.

Bounding the first term in (5.3). Recall that Π_1 be the projection onto $\ker\mathcal{D}$. We claim that $\dim(\ker\mathcal{D})=n_1+n_2-1$. Given a matrix $\theta\in\ker\mathcal{D}$, that is, $D\theta\tilde{D}^\top=0$, we apply Lemma A.1 to obtain the unique decomposition $\theta=R+S+B$, where $R\tilde{D}^\top=0$ and $DS=0$. It follows that $DB\tilde{D}^\top=0$. Since the first column and the first row of B are both identically zero, it is easy to see from an inductive argument that $B=0$ so that $\ker\mathcal{D}$ contains only matrices of the form $\theta=R+S$. The set of constant-row matrices R has dimension n_1; the set of constant-column matrices S with $S_{i,1}=0$ for $i\in[n_1]$ has dimension n_2-1. Thus, $\dim(\ker\mathcal{D})=n_1+n_2-1$. Consequently, we have

$$\mathbb{E}\Big[\sup_{\substack{\theta\in\mathcal{M}\\ \|\theta-\theta^*\|_F\le t}}\big\langle \Pi_1(Z),\theta-\theta^*\big\rangle\Big]\le t\mathbb{E}\big[\big\|\Pi_1(Z)\big\|_F\big]\le t\sqrt{n_1+n_2-1}\lesssim t\sqrt{n_1}.$$

Bounding the second term in (5.3). Similarly, it suffices to compute the rank of Π_2, which is bounded as follows

$$|J|=\sum_{(l,r)\in J}1\le\sum_{r=1}^{n_2}k/r\le k\log(n_2). \qquad (5.4)$$

Therefore, we obtain

$$\mathbb{E}\Big[\sup_{\substack{\theta\in\mathcal{M}\\ \|\theta-\theta^*\|_F\le t}}\big\langle (I-\Pi_1)\Pi_2(Z),\theta-\theta^*\big\rangle\Big]\le t\mathbb{E}\big[\big\|(I-\Pi_1)\Pi_2(Z)\big\|_F\big]\lesssim t\sqrt{k\log(n_2)}.$$

Bounding the third term in (5.3). Note that $I-\Pi_1$ is the projection onto the image of the linear map $\mathcal{D}^\top$, defined by $\mathcal{D}^\top(A)=D^\top A\tilde{D}$. Hence, we have

$$\begin{aligned}
\big\langle (I-\Pi_1)(I-\Pi_2)(Z),\theta-\theta^*\big\rangle&=\big\langle D^\top\big(D^\top\big)^\dagger(I-\Pi_2)(Z)\tilde{D}^\dagger\tilde{D},\theta-\theta^*\big\rangle\\
&=\big\langle\big(D^\dagger\big)^\top(I-\Pi_2)(Z)\tilde{D}^\dagger,D\big(\theta-\theta^*\big)\tilde{D}^\top\big\rangle.
\end{aligned}$$

Since Z has mean zero, it is sufficient to control

$$\mathbb{E}\Big[\sup_{\substack{\theta\in\mathcal{M}\\ \|\theta-\theta^*\|_F\le t}} \big\langle \big(D^\dagger\big)^\top (I-\Pi_2)(Z)\tilde{D}^\dagger, D\theta\tilde{D}^\top\big\rangle\Big]. \tag{5.5}$$

To bound this quantity, we need the following lemma, whose proof is deferred to Section 5.3.

Lemma 10. *For any $i\in[n_1]$, $j\in[n_2]$, the quantity $[(D^\dagger)^\top(I-\Pi_2)(Z)(\tilde{D}^\dagger)]_{i,j}$ is sub-Gaussian with variance proxy*

$$O\bigg(\log(n_2)\bigg[\big(i\wedge(n_1-i)\big)\big(j\wedge(n_2-j)\big)\wedge\frac{n_1n_2}{k}\bigg]\bigg).$$

Let us define $\Phi\in\mathbb{R}^{(n_1-1)\times(n_2-1)}$ by

$$\Phi_{i,j}=\sqrt{\log(n_1)\log(n_2)}\bigg[(\sqrt{i}\wedge\sqrt{n_1-i})(\sqrt{j}\wedge\sqrt{n_2-j})\wedge\sqrt{\frac{n_1n_2}{k}}\bigg],$$

and let $\oslash$ denote element-wise division. Lemma 10, together with a union bound readily yields

$$\mathbb{E}\big[\big\|\big(D^\dagger\big)^\top(I-\Pi_2)(Z)\tilde{D}^\dagger\oslash\Phi\big\|_\infty\big]\lesssim 1.$$

In addition, it holds for every θ that

$$\begin{aligned}\big\langle\big(D^\dagger\big)^\top(I-\Pi_2)(Z)\tilde{D}^\dagger, D\theta\tilde{D}^\top\big\rangle&=\big\langle\big(D^\dagger\big)^\top(I-\Pi_2)(Z)\tilde{D}^\dagger\oslash\Phi,\Phi\odot D\theta\tilde{D}^\top\big\rangle\\&\le\big\|\big(D^\dagger\big)^\top(I-\Pi_2)(Z)\tilde{D}^\dagger\oslash\Phi\big\|_\infty\big\|\Phi\odot D\theta\tilde{D}^\top\big\|_1\end{aligned}$$

by Hölder's inequality. We therefore obtain

$$\begin{aligned}&\mathbb{E}\Big[\sup_{\substack{\theta\in\mathcal{M}\\ \|\theta-\theta^*\|_F\le t}}\big\langle\big(D^\dagger\big)^\top(I-\Pi_2)(Z)\tilde{D}^\dagger, D\big(\theta-\theta^*\big)\tilde{D}^\top\big\rangle\Big]\\&\quad\le\mathbb{E}\big[\big\|\big(D^\dagger\big)^\top(I-\Pi_2)(Z)\tilde{D}^\dagger\oslash\Phi\big\|_\infty\big]\sup_{\substack{\theta\in\mathcal{M}\\ \|\theta-\theta^*\|_F\le t}}\big\|\Phi\odot D\theta\tilde{D}^\top\big\|_1\\&\quad\lesssim\sup_{\substack{\theta\in\mathcal{M}\\ \|\theta-\theta^*\|_F\le t}}\big\|\Phi\odot\big(D\theta\tilde{D}^\top\big)\big\|_1.\end{aligned}$$

It remains to bound this supremum. For $\theta\in\mathcal{M}$, because $D\theta\tilde{D}^\top\ge 0$ we can write

$$\big\|\Phi\odot\big(D\theta\tilde{D}^\top\big)\big\|_1=\big\langle\Phi, D\theta\tilde{D}^\top\big\rangle=\big\langle\Phi, D\big(\theta-\theta^*\big)\tilde{D}^\top\big\rangle+\big\langle\Phi, D\theta^*\tilde{D}^\top\big\rangle. \tag{5.6}$$

The second term in (5.6) can be bounded by

$$\big\langle\Phi, D\theta^*\tilde{D}^\top\big\rangle\le\|\Phi\|_\infty\big\|D\theta^*\tilde{D}^\top\big\|_1\lesssim\sqrt{\frac{n_1n_2}{k}}\sqrt{\log(n_1)\log(n_2)}V\big(\theta^*\big).$$

For the first term in (5.6), we need the following lemma, whose proof is deferred to Appendix A.2.

Lemma 11. *We have the estimate*

$$\left\| D^\top \Phi \tilde{D} \right\|_F^2 \lesssim \log(n_1)\log(n_2)\sqrt{\frac{n_1 n_2}{k}} + \log^2(n_1)\log^2(n_2).$$

If $\|\theta - \theta^*\|_F \le t$, then the above lemma together with the Cauchy–Schwarz inequality yields

$$\begin{aligned}
&\left\langle \Phi, D\big(\theta - \theta^*\big)\tilde{D}^\top \right\rangle \\
&\quad = \left\langle D^\top \Phi \tilde{D}, \theta - \theta^* \right\rangle \le \left\| D^\top \Phi \tilde{D} \right\|_F \left\| \theta - \theta^* \right\|_F \\
&\quad \lesssim t\left[\sqrt{\log(n_1)\log(n_2)} \left(\frac{n_1 n_2}{k} \right)^{1/4} + \log(n_1)\log(n_2) \right]. \qquad (5.7)
\end{aligned}$$

Combining (5.5)–(5.7), we conclude that

$$\begin{aligned}
&\mathbb{E}\Big[\sup_{\substack{\theta \in \mathcal{M} \\ \|\theta - \theta^*\|_F \le t}} \left\langle (I - \Pi_1)(I - \Pi_2)(Z), \theta - \theta^* \right\rangle \Big] \\
&\quad \lesssim \sqrt{\frac{n_1 n_2}{k}} \sqrt{\log(n_1)\log(n_2)}\, V\big(\theta^*\big) + t\left[\sqrt{\log(n_1)\log(n_2)} \left(\frac{n_1 n_2}{k} \right)^{1/4} + \log(n_1)\log(n_2) \right].
\end{aligned}$$

The bounds on the three terms of (5.3) together yield the desired result.

5.3. Proof of Lemma 10

By definition of Π_2, it holds that

$$\begin{aligned}
\big(D^\dagger\big)^\top (I - \Pi_2)(Z)\tilde{D}^\dagger &= \sum_{(\ell,r)\in J^c} U \Sigma^\dagger W^\top w_\ell w_\ell^\top Z \tilde{w}_r \tilde{w}_r^\top \tilde{W} \tilde{\Sigma}^\dagger \tilde{U}^\top \\
&= \sum_{(\ell,r)\in J^c} \big(w_\ell^\top Z \tilde{w}_r\big) U \Sigma^\dagger e_\ell \tilde{e}_r^\top \tilde{\Sigma}^\dagger \tilde{U}^\top \\
&= \sum_{(\ell,r)\in J^c} \big(w_\ell^\top Z \tilde{w}_r\big) \Sigma_{\ell,\ell}^{-1} \tilde{\Sigma}_{r,r}^{-1} U e_\ell \tilde{e}_r^\top \tilde{U}^\top \\
&= \sum_{(\ell,r)\in J^c} \big(w_\ell^\top Z \tilde{w}_r\big) \Sigma_{\ell,\ell}^{-1} \tilde{\Sigma}_{r,r}^{-1} U_{\cdot,\ell} \tilde{U}_{\cdot,r}^\top. \qquad (5.8)
\end{aligned}$$

We now study the sub-Gaussianity of the (i, j)-th entry of this quantity. Since Z has i.i.d. $\mathcal{N}(0, 1)$ entries, it holds for each $\lambda > 0$ that

$$\begin{aligned}
&\mathbb{E}\exp\Bigg(\lambda \sum_{(\ell,r)\in J^c} \big(w_\ell^\top Z \tilde{w}_r\big)\Sigma_{\ell,\ell}^{-1}\tilde{\Sigma}_{r,r}^{-1}U_{i,\ell}\tilde{U}_{j,r}\Bigg)\\
&\quad= \mathbb{E}\exp\Bigg(\lambda \sum_{(\ell,r)\in J^c} \operatorname{Tr}\big(Z\tilde{w}_r w_\ell^\top\big)\Sigma_{\ell,\ell}^{-1}\tilde{\Sigma}_{r,r}^{-1}U_{i,\ell}\tilde{U}_{j,r}\Bigg)\\
&\quad= \mathbb{E}\exp\Bigg\{\operatorname{Tr}\Bigg[Z\Bigg(\lambda \sum_{(\ell,r)\in J^c} \Sigma_{\ell,\ell}^{-1}\tilde{\Sigma}_{r,r}^{-1}U_{i,\ell}\tilde{U}_{j,r}\tilde{w}_r w_\ell^\top\Bigg)\Bigg]\Bigg\}\\
&\quad\le \exp\Bigg\{\frac{\lambda^2}{2}\Bigg\| \sum_{(\ell,r)\in J^c} \Sigma_{\ell,\ell}^{-1}\tilde{\Sigma}_{r,r}^{-1}U_{i,\ell}\tilde{U}_{j,r}\tilde{w}_r w_\ell^\top\Bigg\|_F^2\Bigg\}. \qquad (5.9)
\end{aligned}$$

Note that $\|\tilde{w}_r w_\ell^\top\|_F = 1$, and $\langle \tilde{w}_r w_\ell^\top, \tilde{w}_{r'} w_{\ell'}^\top\rangle = 0$ for any pairs $(r, \ell) \neq (r', \ell')$, so we have

$$\Bigg\| \sum_{(\ell,r)\in J^c} \Sigma_{\ell,\ell}^{-1}\tilde{\Sigma}_{r,r}^{-1}U_{i,\ell}\tilde{U}_{j,r}\tilde{w}_r w_\ell^\top\Bigg\|_F^2 = \sum_{(\ell,r)\in J^c} \Sigma_{\ell,\ell}^{-2}\tilde{\Sigma}_{r,r}^{-2}U_{i,\ell}^2\tilde{U}_{j,r}^2. \qquad (5.10)$$

It remains to bound this quantity. Without loss of generality, assume that n_1 is odd, so $n_1 - 1$ is even. The matrix D has the same left-singular vectors as

$$DD^\top = \begin{bmatrix} 2 & -1 & 0 & \dots & 0 & 0\\ -1 & 2 & -1 & \dots & 0 & 0\\ \vdots & & \ddots & & \vdots & \\ 0 & 0 & \dots & -1 & 2 & -1\\ 0 & 0 & \dots & 0 & -1 & 2 \end{bmatrix},$$

which are known [69] to be

$$U_{i,j} = \sqrt{\frac{2}{n_1}}\sin\Big(\frac{\pi ij}{n_1}\Big), \quad i, j = 1, \dots, n_1 - 1.$$

Moreover, the matrix D has (non-zero) singular values

$$\Sigma_{i,i} = 2\Big|\sin\Big(\frac{\pi i}{2n_1}\Big)\Big|, \quad i = 1, \dots, n_1 - 1.$$

Note that because of the symmetry

$$\sin\Big(\frac{\pi ij}{n_1}\Big) = \sin\Big(\frac{\pi j(n_1 - i)}{n_1}\Big), \quad i = 1, \dots, n_1 - 1,$$

it is enough to consider $i = 1, \dots, \frac{n_1-1}{2}$. We make use of the following inequalities to control the sin terms involved:

$$\left|\sin(x)\right| \le 1 \quad \text{for all } x \in \mathbb{R}; \tag{5.11}$$

$$\sin(x) \le x \quad \text{for } x \in [0, \infty); \tag{5.12}$$

$$\sin(x) \ge \frac{2}{\pi}x \ge \frac{1}{2}x \quad \text{for } x \in \left[0, \frac{\pi}{2}\right]. \tag{5.13}$$

Plugging in the entries of U and Σ yields

$$\begin{aligned}
&\sum_{(\ell,r)\in J^c} \Sigma_{\ell,\ell}^{-2}\tilde{\Sigma}_{r,r}^{-2}U_{i,\ell}^2\tilde{U}_{j,r}^2 \\
&\quad= \sum_{(\ell,r)\in J^c} \frac{4\sin(\frac{\pi i\ell}{n_1})^2\sin(\frac{\pi jr}{n_2})^2}{16n_1n_2\sin(\frac{\pi\ell}{2n_1})^2\sin(\frac{\pi r}{2n_2})^2} \\
&\overset{(i)}{\lesssim} \frac{1}{n_1n_2}\sum_{(\ell,r)\in J^c}\frac{n_1^2n_2^2}{\ell^2r^2} \\
&\lesssim n_1n_2\sum_{r=1}^{n_2}\left(\frac{1}{r^2}\sum_{\ell=\lceil k/r\rceil}^{n_1}\frac{1}{\ell^2}\right) \\
&\lesssim n_1n_2\sum_{r=1}^{n_2}\left(\frac{1}{r^2}\sum_{\ell=\lceil k/r\rceil+1}^{n_1}\frac{1}{\ell^2}\right) + n_1n_2\sum_{r=1}^{n_2}\frac{1}{r^2}\frac{1}{\lceil k/r\rceil^2} \\
&\overset{(ii)}{\lesssim} n_1n_2\sum_{r=1}^{n_2}\frac{1}{r^2}\frac{r}{k} + n_1n_2\sum_{r=1}^{k}\frac{1}{k^2} + n_1n_2\sum_{r=k+1}^{n_2}\frac{1}{r^2} \\
&\overset{(iii)}{\lesssim} \frac{n_1n_2}{k}\log(n_2) + \frac{n_1n_2}{k} + \frac{n_1n_2}{k} \lesssim \frac{n_1n_2}{k}\log(n_2),
\end{aligned} \tag{5.14}$$

where we used (5.11) on the numerator and (5.13) on the denominator in (i) and the bound $\sum_{r=k+1}^{\infty}\frac{1}{r^2} \le \frac{1}{k}$ for any $k \ge 1$ in (ii) and (iii).

On the other hand, even without using the constraint $(\ell, r) \in J^c$, we have

$$\begin{aligned}
\sum_{(\ell,r)\in J^c}\Sigma_{\ell,\ell}^{-2}\tilde{\Sigma}_{r,r}^{-2}U_{i,\ell}^2\tilde{U}_{j,r}^2 &\lesssim \sum_{\ell r\le\frac{n_1n_2}{ij}}\frac{\sin(\frac{\pi i\ell}{n_1})^2\sin(\frac{\pi jr}{n_2})^2}{n_1n_2\sin(\frac{\pi\ell}{2n_1})^2\sin(\frac{\pi r}{2n_2})^2} + \sum_{\ell r>\frac{n_1n_2}{ij}}\frac{\sin(\frac{\pi i\ell}{n_1})^2\sin(\frac{\pi jr}{n_2})^2}{n_1n_2\sin(\frac{\pi\ell}{2n_1})^2\sin(\frac{\pi r}{2n_2})^2} \\
&\lesssim \sum_{\ell r\le\frac{n_1n_2}{ij}}\frac{(i\ell jr)^2}{n_1n_2(\ell r)^2} + \frac{n_1n_2ij}{n_1n_2}\log(n_2),
\end{aligned}$$

where we used (5.12) for the numerator and (5.13) for the denominator as well as (5.14) with k replaced by $\frac{n_1 n_2}{ij}$. Therefore,

$$\sum_{(\ell,r)\in J^c} \Sigma_{\ell,\ell}^{-2} \tilde{\Sigma}_{r,r}^{-2} U_{i,\ell}^2 \tilde{U}_{j,r}^2 \lesssim \sum_{\ell r \le \frac{n_1 n_2}{ij}} \frac{(ij)^2}{n_1 n_2} + ij \log(n_2)$$

$$\lesssim \frac{(ij)^2}{n_1 n_2} \frac{n_1 n_2}{ij} \log(n_2) + ij \log(n_2) \lesssim ij \log(n_2), \qquad (5.15)$$

by counting integer points in the set $\{(\ell, r) : \ell r \le \frac{n_1 n_2}{ij}\}$ as in (5.4).

A similar argument yields bounds with i replaced by $n_1 - i$, or j replaced by $n_2 - j$. Combining this observation with (5.8), (5.9), (5.10), (5.14) and (5.15) completes the proof.

Acknowledgements

We thank Adityanand Guntuboyina for discussing his concurrent work with us. PR was supported by NSF awards IIS-1838071, DMS-1712596 and DMS-TRIPODS-1740751; ONR grant N00014-17-1-2147 and grant 2018-182642 from the Chan Zuckerberg Initiative DAF. ER was supported by NSF MSPRF DMS-1703821.

Supplementary Material

Supplement to: Estimation of Monge matrices (DOI: 10.3150/20-BEJ1215SUPP; .pdf). In the supplementary material, we provide additional proofs of our results.

References

[1] Aggarwal, A., Klawe, M.M., Moran, S., Shor, P. and Wilber, R. (1987). Geometric applications of a matrix-searching algorithm. *Algorithmica* **2** 195–208. MR0895444 https://doi.org/10.1007/BF01840359

[2] Airoldi, E.M., Costa, T.B. and Chan, S.H. (2013). Stochastic blockmodel approximation of a graphon: Theory and consistent estimation. *Adv. Neural Inf. Process. Syst.* **26** 692–700.

[3] Atkins, J.E., Boman, E.G. and Hendrickson, B. (1999). A spectral algorithm for seriation and the consecutive ones problem. *SIAM J. Comput.* **28** 297–310. MR1630473 https://doi.org/10.1137/S0097539795285771

[4] Berthet, Q. and Rigollet, P. (2013). Complexity theoretic lower bounds for sparse principal component detection. In *COLT* 2013 – *The* 26*th Conference on Learning Theory* (S. Shalev-Shwartz and I. Steinwart, eds.). *JMLR W&CP* **30** 1046–1066.

[5] Blei, R., Gao, F. and Li, W.V. (2007). Metric entropy of high dimensional distributions. *Proc. Amer. Math. Soc.* **135** 4009–4018. MR2341952 https://doi.org/10.1090/S0002-9939-07-08935-6

[6] Borgs, C., Chayes, J.T., Cohn, H. and Ganguly, S. (2015). Consistent nonparametric estimation for heavy-tailed sparse graphs. Preprint. Available at arXiv:1508.06675.

[7] Boyle, J.P. and Dykstra, R.L. (1986). A method for finding projections onto the intersection of convex sets in Hilbert spaces. In *Advances in Order Restricted Statistical Inference* (*Iowa City, Iowa*, 1985). *Lect. Notes Stat.* **37** 28–47. Berlin: Springer. MR0875647 https://doi.org/10.1007/978-1-4613-9940-7_3

[8] Braverman, M. and Mossel, E. (2008). Noisy sorting without resampling. In *Proceedings of the Nineteenth Annual ACM-SIAM Symposium on Discrete Algorithms* 268–276. New York: ACM. MR2485312

[9] Burkard, R.E. (2007). Monge properties, discrete convexity and applications. *European J. Oper. Res.* **176** 1–14. MR2265130 https://doi.org/10.1016/j.ejor.2005.04.050

[10] Burkard, R.E., Deĭneko, V.G., van Dal, R., van der Veen, J.A.A. and Woeginger, G.J. (1998). Well-solvable special cases of the traveling salesman problem: A survey. *SIAM Rev.* **40** 496–546. MR1642799 https://doi.org/10.1137/S0036144596297514

[11] Burkard, R.E., Deĭneko, V.G. and Woeginger, G.J. (1999). The travelling salesman problem on permuted Monge matrices. *J. Comb. Optim.* **2** 333–350. MR1669311 https://doi.org/10.1023/A:1009768317347

[12] Burkard, R.E., Klinz, B. and Rudolf, R. (1996). Perspectives of Monge properties in optimization. *Discrete Appl. Math.* **70** 95–161. MR1403225 https://doi.org/10.1016/0166-218X(95)00103-X

[13] Cechlárová, K. and Szabó, P. (1990). On the Monge property of matrices. *Discrete Math.* **81** 123–128. MR1054969 https://doi.org/10.1016/0012-365X(90)90143-6

[14] Çela, E., Deineko, V. and Woeginger, G.J. (2018). New special cases of the quadratic assignment problem with diagonally structured coefficient matrices. *European J. Oper. Res.* **267** 818–834. MR3760806 https://doi.org/10.1016/j.ejor.2017.12.024

[15] Chan, S. and Airoldi, E. (2014). A consistent histogram estimator for exchangeable graph models. In *International Conference on Machine Learning* 208–216.

[16] Chatterjee, S. (2014). A new perspective on least squares under convex constraint. *Ann. Statist.* **42** 2340–2381. MR3269982 https://doi.org/10.1214/14-AOS1254

[17] Chatterjee, S. (2015). Matrix estimation by universal singular value thresholding. *Ann. Statist.* **43** 177–214. MR3285604 https://doi.org/10.1214/14-AOS1272

[18] Chatterjee, S., Guntuboyina, A. and Sen, B. (2018). On matrix estimation under monotonicity constraints. *Bernoulli* **24** 1072–1100. MR3706788 https://doi.org/10.3150/16-BEJ865

[19] Chatterjee, S. and Mukherjee, S. (2019). Estimation in tournaments and graphs under monotonicity constraints. *IEEE Trans. Inf. Theory* **65** 3525–3539. MR3959003 https://doi.org/10.1109/TIT.2019.2893911

[20] Collier, O. and Dalalyan, A.S. (2016). Minimax rates in permutation estimation for feature matching. *J. Mach. Learn. Res.* **17** Paper No. 6, 31. MR3482926

[21] Combettes, P.L. and Pesquet, J.-C. (2011). Proximal splitting methods in signal processing. In *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*. *Springer Optim. Appl.* **49** 185–212. New York: Springer. MR2858838 https://doi.org/10.1007/978-1-4419-9569-8_10

[22] Deineko, V., Rudolf, R. and Woeginger, G.J. (1996). On the recognition of permuted Supnick and incomplete Monge matrices. *Acta Inform.* **33** 559–569. MR1408047 https://doi.org/10.1007/s002360050058

[23] Deineko, V.G. and Filonenko, V.L. (1979). On the reconstruction of specially structured matrices. Aktualnyje Problemy EVM, programmirovanije, Dnepropetrovsk, DGU. (in Russian).

[24] Deutsch, F. and Hundal, H. (1994). The rate of convergence of Dykstra's cyclic projections algorithm: The polyhedral case. *Numer. Funct. Anal. Optim.* **15** 537–565. MR1281561 https://doi.org/10.1080/01630569408816580

[25] Ding, J., Ma, Z., Wu, Y. and Xu, J. (2018). Efficient random graph matching via degree profiles. Preprint. Available at arXiv:1811.07821.

[26] Domahidi, A., Chu, E. and ECOS, S.B. (2013). An SOCP solver for embedded systems. In *European Control Conference* (*ECC*) 3071–3076.
[27] Fallat, S., Lauritzen, S., Sadeghi, K., Uhler, C., Wermuth, N. and Zwiernik, P. (2017). Total positivity in Markov structures. *Ann. Statist.* **45** 1152–1184. MR3662451 https://doi.org/10.1214/16-AOS1478
[28] Fang, B., Guntuboyina, A. and Sen, B. (2019). Multivariate extensions of isotonic regression and total variation denoising via entire monotonicity and Hardy–Krause variation. Preprint. Available at arXiv:1903.01395.
[29] Flammarion, N., Mao, C. and Rigollet, P. (2019). Optimal rates of statistical seriation. *Bernoulli* **25** 623–653. MR3892331 https://doi.org/10.3150/17-bej1000
[30] Fogel, F., d'Aspremont, A. and Vojnovic, M. (2014). Serialrank: Spectral ranking using seriation. In *Advances in Neural Information Processing Systems* 27 (Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence and K. Weinberger, eds.) 900–908. Curran Associates,.
[31] Fogel, F., Jenatton, R., Bach, F. and d'Aspremont, A. (2013). Convex relaxations for permutation problems. In *Advances in Neural Information Processing Systems* 26 (C. Burges, L. Bottou, M. Welling, Z. Ghahramani and K. Weinberger, eds.) 1016–1024. Curran Associates.
[32] Fortuin, C.M., Kasteleyn, P.W. and Ginibre, J. (1971). Correlation inequalities on some partially ordered sets. *Comm. Math. Phys.* **22** 89–103. MR0309498
[33] Gao, C., Lu, Y. and Zhou, H.H. (2015). Rate-optimal graphon estimation. *Ann. Statist.* **43** 2624–2652. MR3405606 https://doi.org/10.1214/15-AOS1354
[34] Gao, F. (2013). Bracketing entropy of high dimensional distributions. In *High Dimensional Probability VI. Progress in Probability* **66** 3–17. Basel: Birkhäuser/Springer. MR3443489
[35] Gavish, M. and Donoho, D.L. (2014). The optimal hard threshold for singular values is $4/\sqrt{3}$. *IEEE Trans. Inf. Theory* **60** 5040–5053. MR3245370 https://doi.org/10.1109/TIT.2014.2323359
[36] Hitchcock, F.L. (1941). The distribution of a product from several sources to numerous localities. *J. Math. Phys. Mass. Inst. Tech.* **20** 224–230. MR0004469 https://doi.org/10.1002/sapm1941201224
[37] Hoffman, A.J. (1963). On simple linear programming problems. In *Proc. Sympos. Pure Math., Vol. VII* 317–327. Providence, RI: Amer. Math. Soc. MR0157778
[38] Hütter, J.-C., Mao, C., Rigollet, P. and Robeva, E. (2020). Optimal rates for estimation of two-dimensional totally positive distributions. In preparation.
[39] Hütter, J.-C., Mao, C., Rigollet, P. and Robeva, E. (2020). Supplement to "Estimation of Monge Matrices." https://doi.org/10.3150/20-BEJ1215SUPP
[40] Hütter, J.-C. and Rigollet, P. (2016). Optimal rates for total variation denoising. In *Conference on Learning Theory* 1115–1146.
[41] Johnstone, I.M. (2017). Gaussian estimation: Sequence and wavelet models. Unpublished manuscript.
[42] Juditsky, A., Rigollet, P. and Tsybakov, A.B. (2008). Learning by mirror averaging. *Ann. Statist.* **36** 2183–2206. MR2458184 https://doi.org/10.1214/07-AOS546
[43] Karlin, S. and Rinott, Y. (1980). Classes of orderings of measures and related correlation inequalities. I. Multivariate totally positive distributions. *J. Multivariate Anal.* **10** 467–498. MR0599685 https://doi.org/10.1016/0047-259X(80)90065-2
[44] Karlin, S. and Rinott, Y. (1980). Classes of orderings of measures and related correlation inequalities. II. Multivariate reverse rule distributions. *J. Multivariate Anal.* **10** 499–516. MR0599686 https://doi.org/10.1016/0047-259X(80)90066-4
[45] Kendall, D.G. (1969). Incidence matrices, interval graphs and seriation in archeology. *Pacific J. Math.* **28** 565–570. MR0239990
[46] Kendall, D.G. (1970). A mathematical approach to seriation. *Philos. Trans. R. Soc. Lond. Ser. A, Math. Phys. Sci.* **269** 125–134.
[47] Klopp, O., Tsybakov, A.B. and Verzelen, N. (2017). Oracle inequalities for network models and sparse graphon estimation. *Ann. Statist.* **45** 316–354. MR3611494 https://doi.org/10.1214/16-AOS1454

[48] Lauritzen, S., Uhler, C. and Zwiernik, P. (2019). Maximum likelihood estimation in Gaussian models under total positivity. *Ann. Statist.* **47** 1835–1863. MR3953437 https://doi.org/10.1214/17-AOS1668

[49] Livi, L. and Rizzi, A. (2013). The graph matching problem. *PAA Pattern Anal. Appl.* **16** 253–283. MR3084902 https://doi.org/10.1007/s10044-012-0284-8

[50] Ma, Z. and Wu, Y. (2015). Computational barriers in minimax submatrix detection. *Ann. Statist.* **43** 1089–1116. MR3346698 https://doi.org/10.1214/14-AOS1300

[51] Mao, C., Pananjady, A. and Wainwright, M.J. (2018). Breaking the $1/\sqrt{n}$ barrier: Faster rates for permutation-based models in polynomial time. Preprint. Available at arXiv:1802.09963.

[52] Mao, C., Weed, J. and Rigollet, P. (2018). Minimax rates and efficient algorithms for noisy sorting. In *AProceedings of the* 28*th International Conference on Algorithmic Learning Theory*.

[53] Massart, P. (2007). *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*. Berlin: Springer.

[54] Monge, G. (1781). Mémoire sur la théorie des déblais et des remblais. *Histoire de L'Académie Royale des Sciences de Paris*.

[55] O'Donoghue, B., Chu, E., Parikh, N. and Boyd, S. (2016). Conic optimization via operator splitting and homogeneous self-dual embedding. *J. Optim. Theory Appl.* **169** 1042–1068. MR3501397 https://doi.org/10.1007/s10957-016-0892-3

[56] O'Donoghue, B., Chu, E., Parikh, N. and SCS, S.B. (2017). Splitting conic solver, version 2.0.2. https://github.com/cvxgrp/scs.

[57] Pananjady, A., Mao, C., Muthukumar, V., Wainwright, M.J. and Courtade, T.A. (2017). Worst-case vs average-case design for estimation from fixed pairwise comparisons. Preprint. Available at arXiv:1707.06217.

[58] Park, J.K. (1991). The monge array: An abstraction and its applications. Ph.D. Thesis, Massachusetts Institute of Technology.

[59] Park, J.K. (1991). A special case of the n-vertex traveling-salesman problem that can be solved in $O(n)$ time. *Inform. Process. Lett.* **40** 247–254. MR1148465 https://doi.org/10.1016/0020-0190(91)90118-2

[60] Pferschy, U., Rudolf, R. and Woeginger, G.J. (1994). Monge matrices make maximization manageable. *Oper. Res. Lett.* **16** 245–254. MR1316146 https://doi.org/10.1016/0167-6377(94)90037-X

[61] Queyranne, M., Spieksma, F. and Tardella, F. (1998). A general class of greedily solvable linear programs. *Math. Oper. Res.* **23** 892–908. MR1662430 https://doi.org/10.1287/moor.23.4.892

[62] Robeva, E., Sturmfels, B., Tran, N. and Uhler, C. (2018). Maximum likelihood estimation for totally positive log-concave densities. Preprint. Available at arXiv:1806.10120.

[63] Rudolf, R. (1994). Recognition of d-dimensional Monge arrays. *Discrete Appl. Math.* **52** 71–82. MR1283245 https://doi.org/10.1016/0166-218X(92)00189-S

[64] Rudolf, R. and Woeginger, G.J. (1995). The cone of Monge matrices: Extremal rays and applications. *Math. Methods Oper. Res.* **42** 161–168. MR1353231 https://doi.org/10.1007/BF01415751

[65] Shah, D. and Lee, C. Reducing crowdsourcing to graphon estimation, statistically. In *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics* (A. Storkey and F. Perez-Cruz, eds.). *Proceedings of Machine Learning Research* **84** 1741–1750. Playa Blanca, Lanzarote, Canary Islands, 09–11 Apr 2018. PMLR.

[66] Shah, N.B., Balakrishnan, S., Guntuboyina, A. and Wainwright, M.J. (2017). Stochastically transitive models for pairwise comparisons: Statistical and computational issues. *IEEE Trans. Inf. Theory* **63** 934–959. MR3604649 https://doi.org/10.1109/TIT.2016.2634418

[67] Shah, N.B., Balakrishnan, S. and Wainwright, M.J. (2016). A permutation-based model for crowd labeling: Optimal estimation and robustness. Preprint. Available at arXiv:1606.09632.

[68] Shah, N.B., Balakrishnan, S. and Wainwright, M.J. (2019). Feeling the bern: Adaptive estimators for Bernoulli probabilities of pairwise comparisons. *IEEE Trans. Inf. Theory* **65** 4854–5874. MR3988527 https://doi.org/10.1109/TIT.2019.2903249

[69] Strang, G. (2007). *Computational Science and Engineering*. Wellesley, MA: Wellesley-Cambridge Press. MR2742791

[70] Wang, Y.-X., Sharpnack, J., Smola, A.J. and Tibshirani, R.J. (2015). Trend filtering on graphs. *Artificial Intelligence and Statistics* 1042–1050.

[71] Wellner, J.A. (2003). Gaussian white noise models: Some results for monotone functions. In *Crossing Boundaries*: *Statistical Essays in Honor of Jack Hall*. *Institute of Mathematical Statistics Lecture Notes – Monograph Series* **43** 87–104. Beachwood, OH: IMS. MR2125049 https://doi.org/10.1214/lnms/1215092392

[72] Wolfe, P.J. and Olhede, S.C. (2013). Nonparametric graphon estimation. Preprint. Available at arXiv:1309.5936.

Received July 2019 and revised March 2020