Estimation of Monge Matrices

Jan-Christian Hütter, Cheng Mao, Philippe Rigollet and Elina Robeva

Massachusetts Institute of Technology and Yale University

Abstract. 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.

1. INTRODUCTION

A matrix $\theta \in \mathbb{R}^{n_1 \times n_2}$ is called a *Monge matrix* [38] or a *submodular matrix* [57], if

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

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 [51]. Since then, it has been widely used and studied in optimization, discrete mathematics and computer science [1, 8, 11, 13, 38, 60] as it allows for simple and fast algorithms in a variety of instances [9, 11, 38, 55, 56].

Many of these problems turn out to be invariant under relabeling of the rows and columns of the Monge matrix. Consequently, we 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

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

is Monge. Note that the terminology *permuted Monge* has also been used to define the same object [11]. A pre-anti-Monge matrix is defined analogously. Like Monge matrices, pre-Monge matrices have also been studied in the context of optimization [10, 14] where the latent permutation yields new computational challenges. For example, even checking that a matrix is pre-Monge is a nontrivial algorithmic task [22, 45].

However, previous work on pre-Monge matrices has focused on the noiseless setting and algorithms typically fail when the pre-Monge matrix is contaminated by random noise. This motivates us to take a statistical approach and study estimation of a pre-Monge matrix under random noise.

1.1 Geometric interpretation and spectral ordering.

The Monge property has strong ties with geometric properties of certain datasets, starting with the seminal work of Monge on optimal transport [51]. In this subsection, we demonstrate how the Monge property arises in the context of seriation [3, 30–32, 43, 44] where the goal is to recover the latent ordering of objects based on pairwise distances or inner products.

Let $X \in \mathbb{R}^{n\times d}$ be a data matrix with rows $x_1^\top, \dots, x_n^\top \in \mathbb{R}^d$. Moreover, suppose that for all $i, j \in [n-1]$, we have that $(x_{i+1} - x_i)^\top (x_{j+1} - x_j) \geq 0$. In other words, the differences between consecutive points form an acute angle so that the points $x_1, \dots, x_n$ may be ordered along a common direction. In this case, it is 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.

Suppose that we do not know the order of the points, or equivalently, we observe $x_{\pi(1)}, \dots, x_{\pi(n)}$, where there is an unknown permutation $\pi : [n] \to [n]$. How can we reorder the points in order to recover the above geometric structure (i.e. so that the differences between consecutive points form an acute angle)? Intuitively, assuming that such a reordering exists suggests that the n points should approximately lie along a hidden direction. Therefore, we can apply principal component analysis as follows. Let us assume without loss of generality that the points are centered so that $\sum_{i=1}^n x_i = 0$ and thus $\sum_{i=1}^n \theta_{i,j} = \sum_{j=1}^n \theta_{i,j} = 0$. Then the leading right singular vector of X gives the hidden direction, and the leading left singular vector v of X, i.e., the leading eigenvector of the Gram matrix θ is the first principal component of the data points. The entries of v then give a one-dimensional embedding of the points, from which we easily recover the original order.

Indeed, this intuition can be made rigorous using Corollary 2.6 of [28], which is a variant of the Perron-Frobenius theorem and states that the leading eigenvector of a doubly centered anti-Monge matrix (i.e. having row and column sums equal to zero) is monotone. Hence the leading eigenvector v of the Gram matrix θ is monotone. If the unknown permutation π relabels the points, then the leading eigenvector of the Gram matrix becomes v_π, defined by $(v_\pi)_i = v_{\pi(i)}$. As a result, sorting the entries of v_π recovers the permutation π and, therefore, the latent order of the points. The above method for spectral ordering is similar to the one for seriation proposed in [3].

1.2 Our contribution

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 our result (Proposition 6) on the approximation of pre-Monge matrices by low-rank ones.

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.

The proofs of all theorems and auxiliary lemmas can be found in Section 5.

1.3 Related work

This work connects to several lines of research that are described below.

Total positivity. The Monge property is closely related to the notion of *total positivity* [41]. An entrywise positive matrix $\theta \in \mathbb{R}^{n_1 \times n_2}$ is called totally positive (of order 2), if

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

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 as *log-supermodularity*. Total positivity plays an essential role in statistical physics via the FKG inequality [33] and appears frequently in many other areas of probability and statistics [41, 42]. More recently, there have been new developments in studying totally positive distributions and related estimation problems [26, 47, 58]. In a companion paper [39], 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 [7], the strong stochastic transitivity model [17, 62], feature matching [20], crowd labeling [64], statistical seriation [30] and graph matching [24, 48], 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, 54, 63] or certain refinements [49, 50]. 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 [5, 15, 34, 70], where the goal is to estimate a bivariate function $f : [0,1]^2 \to \mathbb{R}$ from samples $\{f(X_i, Y_j) : 1 \leq i \leq n_1,\, 1 \leq j \leq n_2\}$. 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 [46] and low-rank structure [61]. While our setup is not about recovering an underlying function f, the current work can be viewed as a study of denoising observations in graphon estimation with the Monge structure.

Shape-constrained estimation. Estimation of a Monge matrix, which we study in Section 2, falls in the scope of shape-constrained estimation. Closest to the present work is the estimation of a bivariate isotonic matrix under 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 7). 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 6 is different from that given by Lemma 4 of [62].

Shortly before completing the current work, we became aware of a concurrent work by Fang, Guntuboyina and Sen [27] that studies multivariate extensions of isotonic regression. The two-dimensional version almost coincides with the anti-Monge structure (without permutations) that we study, and 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 in our current paper.

However, it is worth noting that the two proofs follow drastically different paths. While the proof in [27] relies on metric entropy estimates from [4, 35], our proof is based on spectral decomposition of the difference operator D defined in (2.2), a technique which has been used for example to study the performance of total variation regularization [40, 69]. 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 [27] potentially scales like $\log(n)^3$, a minor improvement which nonetheless shows the potential merits of our proof technique.

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, without loss of generality, we also assume that $n_1 \ge n_2$.

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

$$D = \begin{bmatrix} -1 & 1 & 0 & \dots & 0 & 0 \\ 0 & -1 & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \dots & -1 & 1 \end{bmatrix}, \tag{2.2}$$

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

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

where the symbol $\ge$ denotes entrywise inequality.

For each $\theta \in \mathcal{M}$, we define the quantity

$$V(\theta) := \theta_{1,1} + \theta_{n_1,n_2} - \theta_{n_1,1} - \theta_{1,n_2} = \|D\theta\tilde{D}^\top\|_1\,, \tag{2.3}$$

where the last equality follows 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

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

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

$$\hat{\theta}^{\mathsf{ls}} := \operatorname*{argmin}_{\theta \in \mathcal{M}} \|\theta - y\|_F^2, \tag{2.4}$$

in terms of the mean squared error

$$\frac{1}{n_1 n_2} \|\hat{\theta} - \theta^*\|_F^2.$$

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 $\varepsilon \sim \mathsf{subG}_{n_1 \times n_2}(\sigma^2)$. Then, the least-squares estimator $\hat{\theta}^{\mathsf{ls}}$ achieves the rate*

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

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*

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

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

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

$$y = \theta^*(\pi_1^*, \pi_2^*) + \varepsilon, \tag{3.5}$$

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

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

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

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

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

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

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

$$\hat{\theta}^{\mathsf{gls}} \in \operatorname*{argmin}_{\theta \in \bar{\mathcal{M}}_{V_0}} \|\theta - y\|_F^2. \tag{3.6}$$

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 (3.6) achieves the rate*

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

with probability at least $1 - n_1^{-n_1}$. Moreover, 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).

We conjecture that a similar bound holds true for a version of the least-squares estimator where the projection onto $\bar{\mathcal{M}}_{V_0}$ is replaced by the unrestricted version $\bar{\mathcal{M}}$, but our current proof technique does not allow us to conclude this.

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 (3.6) 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.,

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

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 following algorithm is that if we knew the index $\pi_1^{-1}(1)$ corresponding to the first row of θ^*, the anti-Monge property would imply that the variances between any other row $i \in [n_1]$ and row 1 in the unpermuted matrix θ^*,

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

are monotonically increasing in i. Hence, given $\pi_1(1)$, we could estimate these variances and sort the rows accordingly. The precise method is given in the following $\mathsf{Variance\ Sorting}$ Subroutine.

Algorithm 1 Variance Sorting

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

$$\xi(i,j) = \sum_{k=1}^{n_2} \left[y_{i,k} - y_{j,k} - \frac{1}{n_2} \sum_{\ell=1}^{n_2} (y_{i,\ell} - y_{j,\ell}) \right]^2, \tag{3.7}$$

and define

$$(i_0, j_0) = \underset{(i,j)\in[n_1]^2,\, i<j}{\mathrm{argmax}} \ \xi(i,j). \tag{3.8}$$

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 $\hat{\pi}_1(n_1) = j_0$.

Note that in Algorithm 1, the pair (i_0, j_0) 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 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 $\pi_1^{\mathsf{r}} \in \mathcal{S}_{n_1}$ by $\pi_1^{\mathsf{r}}(i) = n_1 - i + 1$ for $i \in [n_1]$, and define similarly $\pi_2^{\mathsf{r}} \in \mathcal{S}_{n_2}$ by $\pi_2^{\mathsf{r}}(i) = n_2 - i + 1$ 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.

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

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

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

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

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 $\theta(\pi_1^{\mathsf{r}}, \pi_2^{\mathsf{r}}) \in \mathcal{M}_{V_0}$, and as a result

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

The estimator computed by the Main Algorithm achieves the following rate of estimation.

THEOREM 4. *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. Then it holds with probability at least* $1 - n_1^{-n_1}$ *that*

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

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 (3.5), we look for a possibly improper estimator $\tilde{\theta} \in \mathbb{R}^{n_1 \times n_2}$ so that $\big\|\tilde{\theta} - \theta^*(\pi_1^*, \pi_2^*)\big\|_F^2$ is small.

To this end, we consider the well-studied singular value thresholding (SVT) estimator [17, 36]. 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

$$\hat{\theta}^{\mathsf{svt}} := \sum_{i=1}^{n_2} \mathbb{1}\{\lambda_i > \rho\} \lambda_i u_i v_i^\top , \tag{3.9}$$

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 5. *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 the rate*

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

with probability at least $1 - \exp(-n_1)$. Moreover, 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 4. Note that for this result, the noise ε needs not be homoscedastic, and moreover, no knowledge of V_0 is required, i.e., the SVT estimator adapts to the quantity $V(\theta^*)$.

The proof technique leading to upper bounds for the SVT estimator is well developed [17, 62]. Our contribution mainly lies in the following low-rank approximation result for an anti-Monge matrix, which is of independent interest.

PROPOSITION 6. *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*

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

Note that using a similar proof, the same rate as in Theorem 5 can be obtained for a soft-thresholding estimator as well, that is, for

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

with a similar scaling for ρ.

As the rate given in Theorem 5 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 B, 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 6. This in turn gives evidence that the rate of convergence for the SVT estimator in Theorem 5 might be the best achievable by this method.

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 (2.4) can be calculated by a general purpose convex optimization software such as SCS [52, 53] or EOCS [25]. The most computationally intensive subroutine of these methods is usually solving 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 [6, 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 $\mathcal{M}_1, \dots, \mathcal{M}_m$ by iteratively projecting carefully chosen points to each individual set. This is similar to alternate projections of a point to each of the sets $\mathcal{M}_1, \dots, \mathcal{M}_m$, 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 $\bigcap_{j\in[m]} \mathcal{M}_j$.

Algorithm 3 Dykstra's algorithm

Input: $y \in \mathbb{R}^d$, the point to project; $\mathcal{M}_1, \dots, \mathcal{M}_m$ 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
```

To apply Dysktra's algorithm to the problem 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

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

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

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

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

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

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 $\mathcal{M}$.

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

```
Input: y ∈ R^{n1×n2}
Output: θ ≈ Π_M(y)
  function ProjetAntiMonge(y)
    η ← 0 ∈ R^{(n1−1)×(n2−1)}                              ▷ Initialize residuals
    θ ← y,                                                  ▷ Initialize iterates
    while not converged do
      for i1 = 1,...,n1 − 1, i2 = 1,...,n2 − 1 do
        η̃ ← max{ −Σ_{j1∈{0,1},j2∈{0,1}} (−1)^{j1+j2} θ_{i1+j1,i2+j2}/4 + η_{i1,i2}, 0 }   ▷ Compute new residuals
        for j1 ∈ {0,1}, j2 ∈ {0,1} do
          θ_{i1+j1,i2+j2} ← θ_{i1+j1,i2+j2} + (−1)^{j1+j2}(η̃ − η_{i1,i2})   ▷ Project shifted iterates
        end for
        η_{i1,i2} ← η̃                                       ▷ Store residuals
      end for
    end while
    return θ
  end function
```

The rate of convergence of Dykstra's method can be shown to be linearly exponential in the iterations [23], 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 [23] and matches our experience: simulations for larger values of n_1 and n_2 require more iterates before convergence.

In practice, convergence in Algorithm 4 can be checked by evaluating a measure of feasibility such as $\|D\theta\tilde{D}^\top\|_\infty$, 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

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

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 $\|\theta^{\mathsf{ls}} - \theta^*_{1,(n)}\|_F^2/n^2$ over 20 repetitions.

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

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

These were chosen because the singular value decay we proved in Proposition 6 is tight for these matrices (see Lemma 23). 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, i.e.,

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

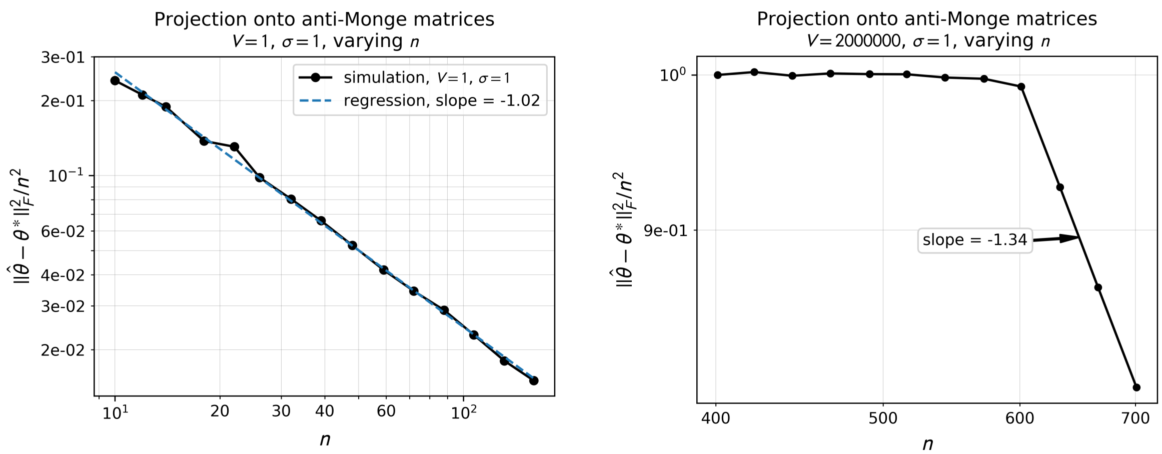
for $\theta^* = \theta^*_{2,(n)}$, averaged over 256 repetitions. This measure of the approximation quality of the estimated permutations corresponds to the upper bound used in the proof of Proposition 14 (see (5.34)) and is applicable since by construction, θ^* has row and column sums equal to zero. It is the dominating part in the error analysis, leading to the rate reported in Theorem 4, 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, this time setting $V = 1$, $\sigma = 0.1$, and varying n between 20 and 500. For this experiment, in Figure 3(b), we plotted the full denoising error,

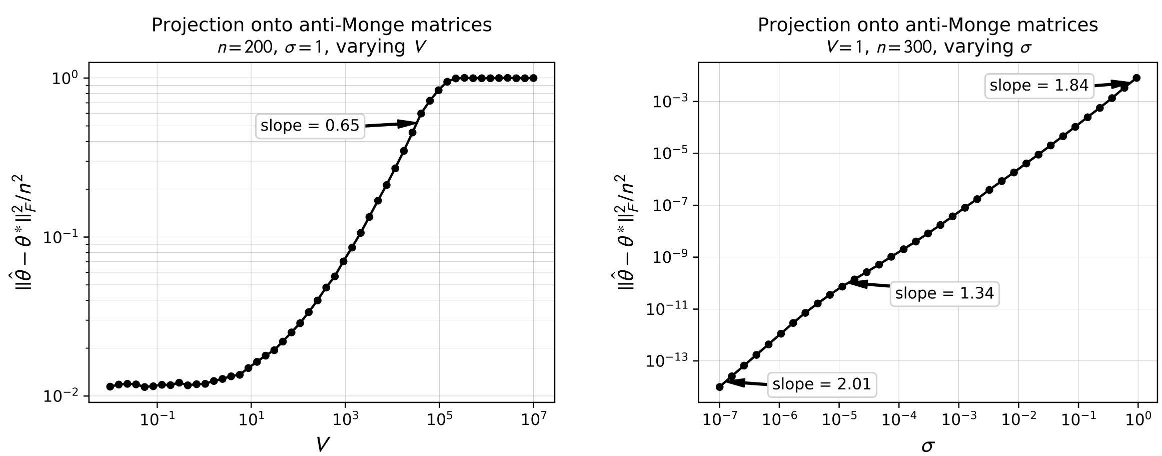
$$\frac{1}{n^2}\|\hat{\theta} - \theta^*\|_F^2,$$

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



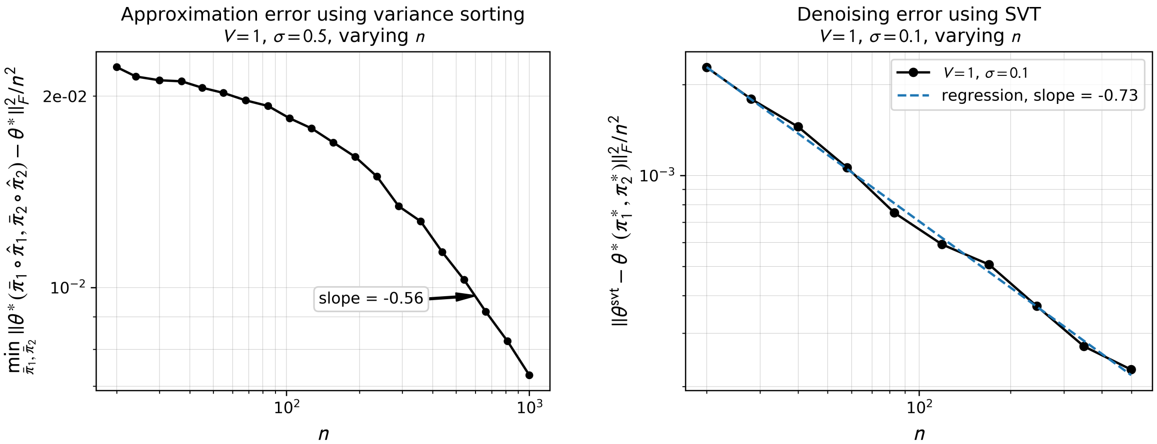
(a) n^{-1} scaling (b) $n^{-4/3}$ scaling

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



(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.



(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.

5. PROOFS

In this section, we provide the proofs of our results. Recall that D is defined in (2.2) 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 A structural lemma

To gain insight into the set of anti-Monge matrices, we first state and prove the following structural lemma, which says that any anti-Monge matrix can be written as the sum of a constant-row matrix, a constant-column matrix, and an anti-Monge, bivariate isotonic matrix. Several structural decompositions of this type are known, e.g., Lemma 2.1 of [11] and Lemma 2.5 of [29]. Our lemma is stated in a form that is convenient for application, and we provide a self-contained proof which also facilitates understanding the structure of an anti-Monge matrix.

LEMMA 7. *For each $\theta \in \mathcal{M}^{n_1,n_2}$, there exists a unique triple (R, S, B) of matrices in $\mathbb{R}^{n_1 \times n_2}$ such that*

- $\theta = R + S + B$;
- $R\tilde{D}^\top = 0$ *and* $DS = 0$, *i.e., R has constant rows and S has constant columns;*
- $S_{i,1} = B_{i,1} = B_{1,j} = 0$ *for* $i \in [n_1]$, $j \in [n_2]$.

Moreover, we have that

- *B is anti-Monge;*
- *B is bivariate isotonic, i.e., B has nondecreasing rows and columns;*
- $\max\{\|R\|_\infty, \|S\|_\infty, \|B\|_\infty\} \le 4\|\theta\|_\infty$;
- $\|B\|_\infty = B_{n_1,n_2} = V(\theta)$.

PROOF. By the condition $S_{i,1} = B_{i,1} = 0$ for $i \in [n_1]$, the first column of R is equal to that of θ. Since R has constant rows, it is uniquely defined by $R = \theta_{\cdot,1}\tilde{\mathbf{1}}^\top$ where $\tilde{\mathbf{1}}$ denotes the all-ones vector in $\mathbb{R}^{n_2}$. Moreover, by the condition $B_{1,j} = 0$ for $j \in [n_2]$, the first row of S is equal to that of $\theta - R$. Since S has constant columns, it is uniquely defined by $S = \mathbf{1}(\theta - R)_{1,\cdot}$ where $\mathbf{1}$ denotes the all-ones vector in $\mathbb{R}^{n_1}$. Finally, define $B = \theta - R - S$.

Since $DS = R\tilde{D}^\top = 0$, we have $DB\tilde{D}^\top = D(\theta - R - S)\tilde{D}^\top = D\theta\tilde{D}^\top \geq 0$. so that B is anti-Monge. Moreover, this implies

$$(DB)_{i,j} \leq (DB)_{i,j+1}, \quad \text{for } i \in [n_1],\ j \in [n_2 - 1], \quad \text{and} \tag{5.10}$$

$$(B\tilde{D}^\top)_{i,j} \leq (B\tilde{D}^\top)_{i+1,j}, \quad \text{for } i \in [n_1 - 1],\ j \in [n_2]. \tag{5.11}$$

The first column and first row of B are equal to 0 by construction. Consequently, we obtain that $(DB)_{\cdot,1} = 0$, and by (5.10), this gives $DB \geq 0$ and similarly $\theta\tilde{D}^\top \geq 0$ by (5.11) . Together, these facts yield that B is bivariate isotonic.

Additionally, from the triangle inequality and the way we constructed R, S, and B, we get

$$\|R\|_\infty \leq \|\theta\|_\infty, \quad \|S\|_\infty \leq 2\|\theta\|_\infty, \quad \text{and} \quad \|B\|_\infty \leq 4\|\theta\|_\infty.$$

Since B also inherits the variation of θ and is nonnegative, $\|B\|_\infty = V(B) = V(\theta)$, which completes the proof. □

5.2 Proof of Theorem 1

To control the performance of a least-squares estimator, we employ Chatterjee's variational formula [16] that we recall below. See, e.g., Lemma 6.1 of [30] 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*

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

Then we have

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

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 20.

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 $1 \le k \le n_1 n_2$ to be chosen later. Note that by Theorem 19 and Proposition 9, we obtain

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

where γ_2 denotes Talagrand's γ_2 functional. Therefore, Theorem 20 yields that with probability $1 - 4\exp(-s^2)$,

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

Let us define

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

we obtain that for any

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

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

$$f_{\theta^*}(t) < 0.$$

Therefore by Lemma 8, we obtain that with probability at least $1 - 4\exp(-s^2)$,

$$\frac{1}{n_1 n_2} \|\hat{\theta}^{\mathsf{ls}} - \theta^*\|_F^2 \le \frac{(t^*)^2}{n_1 n_2} \tag{5.12}$$
$$\lesssim \sigma^2 \left[\frac{1}{n_2} + \frac{k \log(n_2)}{n_1 n_2} + \frac{\log(n_1)\log(n_2)}{\sqrt{n_1 n_2 k}} \right] + \sigma \sqrt{\frac{\log(n_1)\log(n_2)}{n_1 n_2 k}}\, V(\theta^*) + \sigma^2 \frac{s^2}{n_1 n_2}.$$

We now choose $s = 2\sqrt{n_1}$. Balancing the terms 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 $1 \le k^* \le n_1 n_2$, then

$$\frac{1}{n_1 n_2} \|\hat{\theta}^{\mathsf{ls}} - \theta^*\|_F^2 \lesssim \frac{\sigma^2}{n_2} + \frac{\sigma^2 \log(n_2) \log(n_1)^{1/3} \big[\sqrt{\log(n_1)} + V(\theta^*) / \big(\sigma \sqrt{\log(n_2)} \big) \big]^{2/3}}{(n_1 n_2)^{2/3}}$$
$$\lesssim \frac{\sigma^2}{n_2} + \left(\frac{\sigma^2 V(\theta^*)}{n_1 n_2} \right)^{2/3} \log(n_1)^{1/3} \log(n_2)^{2/3}. \tag{5.13}$$

If $k^* < 1$, we replace it by 1, increasing the $k\log(n_2)/(n_1n_2)$ term while decreasing the ones with $1/\sqrt{k}$ in (5.12), hence leading to the same rate as in (5.13). If $k^* > n_1n_2$, note that the $k/(n_1n_2)$ term is already of the order σ^2, so a basic bound of σ^2 on the empirical process term yields the rate

$$\frac{1}{n_1n_2}\|\hat{\theta}^{\mathsf{ls}} - \theta^*\|_F^2 \leq \sigma^2.$$

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

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

To obtain the bound in expectation, we can first integrate the exponentially decaying tale of (5.12), and then choose the optimal k in the same way.

5.3 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 $W = \big[w_1 \,\big|\, \cdots \,\big|\, w_{n_1-1}\big]$.

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

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

With these two projections, we decompose

$$\mathbb{E}[\sup_{\substack{\theta\in\mathcal{M}\\ \|\theta-\theta^*\|_F\leq t}} \langle Z, \theta-\theta^*\rangle] \leq \mathbb{E}[\sup_{\substack{\theta\in\mathcal{M}\\ \|\theta-\theta^*\|_F\leq t}} \langle \Pi_1(Z), \theta-\theta^*\rangle] + \mathbb{E}[\sup_{\substack{\theta\in\mathcal{M}\\ \|\theta-\theta^*\|_F\leq t}} \langle (I-\Pi_1)\Pi_2(Z), \theta-\theta^*\rangle]$$
$$+\,\mathbb{E}[\sup_{\substack{\theta\in\mathcal{M}\\ \|\theta-\theta^*\|_F\leq t}} \langle (I-\Pi_1)(I-\Pi_2)(Z), \theta-\theta^*\rangle]. \tag{5.14}$$

We now bound the three terms in (5.14) separately.

Bounding the first term in (5.14). 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 $\theta \in \ker\mathcal{D}$, i.e., $D\theta\tilde{D}^\top = 0$, we apply Lemma 7 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

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

Bounding the second term in (5.14). Similarly, it suffices to compute the rank of Π_2, which is bounded as follows

$$|J| = \sum_{(l,r)\in J} 1 \le \sum_{r=1}^{n_2} k/r \le k\log(n_2). \tag{5.15}$$

Therefore, we obtain

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

Bounding the third term in (5.14). 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

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

Since Z has mean zero, it is sufficient to control

$$\mathbb{E}[\sup_{\substack{\theta\in\mathcal{M}\\ \|\theta-\theta^*\|_F\le t}} \langle (D^\dagger)^\top (I-\Pi_2)(Z)\tilde{D}^\dagger, D\theta\tilde{D}^\top\rangle]. \tag{5.16}$$

To bound this quantity, we need the following lemma, whose proof is deferred to Section 5.4.

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*

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

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

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

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

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

In addition, it holds for every θ that

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

by Hölder's inequality. We therefore obtain

$$\begin{aligned}&\mathbb{E}[\sup_{\substack{\theta\in\mathcal{M}\\ \|\theta-\theta^*\|_F\le t}} \langle (D^\dagger)^\top(I-\Pi_2)(Z)\tilde{D}^\dagger, D(\theta-\theta^*)\tilde{D}^\top\rangle] \\ &\le \mathbb{E}[\|(D^\dagger)^\top(I-\Pi_2)(Z)\tilde{D}^\dagger\oslash\Phi\|_\infty] \sup_{\substack{\theta\in\mathcal{M}\\ \|\theta-\theta^*\|_F\le t}} \|\Phi\odot D\theta\tilde{D}^\top\|_1 \\ &\lesssim \sup_{\substack{\theta\in\mathcal{M}\\ \|\theta-\theta^*\|_F\le t}} \|\Phi\odot(D\theta\tilde{D}^\top)\|_1. \end{aligned} \tag{5.17}$$

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

$$\|\Phi \odot (D\theta\tilde{D}^\top)\|_1 = \langle \Phi, D\theta\tilde{D}^\top \rangle = \langle \Phi, D(\theta - \theta^*)\tilde{D}^\top \rangle + \langle \Phi, D\theta^*\tilde{D}^\top \rangle. \tag{5.18}$$

The second term in (5.18) can be bounded by

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

For the first term in (5.18), we need the following lemma, whose proof is deferred to Section 5.5.

Lemma 11. *We have the estimate*

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

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

$$\begin{aligned} &\langle \Phi, D(\theta - \theta^*)\tilde{D}^\top \rangle \\ &= \langle D^\top \Phi \tilde{D}, \theta - \theta^* \rangle \leq \|D^\top \Phi \tilde{D}\|_F \|\theta - \theta^*\|_F \\ &\lesssim t \left[\sqrt{\log(n_1)\log(n_2)} \left(\frac{n_1 n_2}{k}\right)^{1/4} + \log(n_1)\log(n_2) \right]. \end{aligned} \tag{5.19}$$

Combining (5.16)–(5.19), we conclude that

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

The bounds on the three terms of (5.14) together yield the desired result.

5.4 Proof of Lemma 10

By definition of Π_2, it holds that

$$\begin{aligned} (D^\dagger)^\top (I - \Pi_2)(Z)\tilde{D}^\dagger &= \sum_{(\ell,r) \in J^c} U\Sigma^\dagger W^\top w_\ell w_\ell^\top Z \tilde{w}_r \tilde{w}_r^\top \tilde{W} \tilde{\Sigma}^\dagger \tilde{U}^\top \\ &= \sum_{(\ell,r) \in J^c} (w_\ell^\top Z \tilde{w}_r) U \Sigma^\dagger e_\ell \tilde{e}_r^\top \tilde{\Sigma}^\dagger \tilde{U}^\top \\ &= \sum_{(\ell,r) \in J^c} (w_\ell^\top Z \tilde{w}_r) \Sigma_{\ell,\ell}^{-1} \tilde{\Sigma}_{r,r}^{-1} U e_\ell \tilde{e}_r^\top \tilde{U}^\top \\ &= \sum_{(\ell,r) \in J^c} (w_\ell^\top Z \tilde{w}_r) \Sigma_{\ell,\ell}^{-1} \tilde{\Sigma}_{r,r}^{-1} U_{\cdot,\ell} \tilde{U}_{\cdot,r}^\top. \end{aligned} \tag{5.20}$$

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

$$\begin{aligned}\mathbb{E}\exp\Big(\lambda \sum_{(\ell,r)\in J^c} (w_\ell^\top Z\tilde{w}_r)\Sigma_{\ell,\ell}^{-1}\tilde{\Sigma}_{r,r}^{-1}U_{i,\ell}\tilde{U}_{j,r}\Big) &= \mathbb{E}\exp\Big(\lambda \sum_{(\ell,r)\in J^c} \mathsf{Tr}(Z\tilde{w}_r w_\ell^\top)\Sigma_{\ell,\ell}^{-1}\tilde{\Sigma}_{r,r}^{-1}U_{i,\ell}\tilde{U}_{j,r}\Big)\\ &= \mathbb{E}\exp\Big\{\mathsf{Tr}\Big[Z\Big(\lambda \sum_{(\ell,r)\in J^c} \Sigma_{\ell,\ell}^{-1}\tilde{\Sigma}_{r,r}^{-1}U_{i,\ell}\tilde{U}_{j,r}\tilde{w}_r w_\ell^\top\Big)\Big]\Big\}\\ &\le \exp\Big\{\frac{\lambda^2}{2}\Big\|\sum_{(\ell,r)\in J^c} \Sigma_{\ell,\ell}^{-1}\tilde{\Sigma}_{r,r}^{-1}U_{i,\ell}\tilde{U}_{j,r}\tilde{w}_r w_\ell^\top\Big\|_F^2\Big\}. \qquad (5.21)\end{aligned}$$

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

$$\Big\|\sum_{(\ell,r)\in J^c} \Sigma_{\ell,\ell}^{-1}\tilde{\Sigma}_{r,r}^{-1}U_{i,\ell}\tilde{U}_{j,r}\tilde{w}_r w_\ell^\top\Big\|_F^2 = \sum_{(\ell,r)\in J^c} \Sigma_{\ell,\ell}^{-2}\tilde{\Sigma}_{r,r}^{-2}U_{i,\ell}^2\tilde{U}_{j,r}^2. \qquad (5.22)$$

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 [65] to be

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

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

$$\Sigma_{i,i} = 2\left|\sin\left(\frac{\pi i}{2n_1}\right)\right|, \quad i = 1,\dots,n_1-1. \qquad (5.23)$$

Note that because of the symmetry

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

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:

$$|\sin(x)| \le 1, \qquad \text{for all } x \in \mathbb{R}; \qquad (5.24)$$

$$\sin(x) \le x, \qquad \text{for } x \in [0,\infty); \qquad (5.25)$$

$$\sin(x) \ge \frac{2}{\pi}x \ge \frac{1}{2}x, \quad \text{for } x \in [0,\frac{\pi}{2}]. \qquad (5.26)$$

Plugging in the entries of U and Σ yields

$$
\begin{aligned}
\sum_{(\ell,r)\in J^c} \Sigma_{\ell,\ell}^{-2}\tilde{\Sigma}_{r,r}^{-2}U_{i,\ell}^2\tilde{U}_{j,r}^2 &= \sum_{(\ell,r)\in J^c} \frac{4\sin\left(\frac{\pi i\ell}{n_1}\right)^2\sin\left(\frac{\pi jr}{n_2}\right)^2}{16n_1n_2\sin\left(\frac{\pi\ell}{2n_1}\right)^2\sin\left(\frac{\pi r}{2n_2}\right)^2} \\
&\overset{(i)}{\lesssim} \frac{1}{n_1n_2}\sum_{(\ell,r)\in J^c}\frac{n_1^2n_2^2}{\ell^2r^2} \\
&\lesssim n_1n_2\sum_{r=1}^{n_2}\left(\frac{1}{r^2}\sum_{\ell=\lceil k/r\rceil}^{n_1}\frac{1}{\ell^2}\right) \\
&\lesssim n_1n_2\sum_{r=1}^{n_2}\left(\frac{1}{r^2}\sum_{\ell=\lceil k/r\rceil+1}^{n_1}\frac{1}{\ell^2}\right) + n_1n_2\sum_{r=1}^{n_2}\frac{1}{r^2}\frac{1}{\lceil k/r\rceil^2} \\
&\overset{(ii)}{\lesssim} n_1n_2\sum_{r=1}^{n_2}\frac{1}{r^2}\frac{r}{k} + n_1n_2\sum_{r=1}^{k}\frac{1}{k^2} + n_1n_2\sum_{r=k+1}^{n_2}\frac{1}{r^2} \\
&\overset{(iii)}{\lesssim} \frac{n_1n_2}{k}\log(n_2) + \frac{n_1n_2}{k} + \frac{n_1n_2}{k} \lesssim \frac{n_1n_2}{k}\log(n_2), \qquad (5.27)
\end{aligned}
$$

where we used (5.24) on the numerator and (5.26) on the denominator in (i) and the bound $\sum_{r=k+1}^{\infty}\frac{1}{r^2} \leq \frac{1}{k}$ for any $k \geq 1$ in (ii) and (iii).

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

$$
\begin{aligned}
\sum_{(\ell,r)\in J^c} \Sigma_{\ell,\ell}^{-2}\tilde{\Sigma}_{r,r}^{-2}U_{i,\ell}^2\tilde{U}_{j,r}^2 &\lesssim \sum_{\ell r\leq\frac{n_1n_2}{ij}} \frac{\sin\left(\frac{\pi i\ell}{n_1}\right)^2\sin\left(\frac{\pi jr}{n_2}\right)^2}{n_1n_2\sin\left(\frac{\pi\ell}{2n_1}\right)^2\sin\left(\frac{\pi r}{2n_2}\right)^2} + \sum_{\ell r>\frac{n_1n_2}{ij}} \frac{\sin\left(\frac{\pi i\ell}{n_1}\right)^2\sin\left(\frac{\pi jr}{n_2}\right)^2}{n_1n_2\sin\left(\frac{\pi\ell}{2n_1}\right)^2\sin\left(\frac{\pi r}{2n_2}\right)^2} \\
&\overset{(i)}{\lesssim} \sum_{\ell r\leq\frac{n_1n_2}{ij}} \frac{(i\ell jr)^2}{n_1n_2(\ell r)^2} + \frac{n_1n_2ij}{n_1n_2}\log(n_2) \\
&\lesssim \sum_{\ell r\leq\frac{n_1n_2}{ij}} \frac{(ij)^2}{n_1n_2} + ij\log(n_2) \\
&\overset{(ii)}{\lesssim} \frac{(ij)^2}{n_1n_2}\frac{n_1n_2}{ij}\log(n_2) + ij\log(n_2) \lesssim ij\log(n_2), \qquad (5.28)
\end{aligned}
$$

where in (i), we used (5.25) for the numerator and (5.26) for the denominator as well as (5.27) with k replaced by $\frac{n_1n_2}{ij}$, and (ii) follows from counting integer points in the set $\{(\ell, r) : \ell r \leq \frac{n_1n_2}{ij}\}$ as in (5.15).

A similar argument yields bounds with i replaced by $n_1 - i$, or j replaced by $n_2 - j$. Combining this observation with (5.20), (5.21), (5.22), (5.27) and (5.28) completes the proof.

5.5 Proof of Lemma 11

Define $\Phi' = \frac{\Phi}{\sqrt{\log(n_1)\log(n_2)}}$, i.e.,

$$\Phi'_{i,j} = \left(\sqrt{i} \wedge \sqrt{n_1 - i}\right)\left(\sqrt{j} \wedge \sqrt{n_2 - j}\right) \wedge \sqrt{\frac{n_1n_2}{k}}.$$

To simplify the notation, let $\Phi'_{i,0} = \Phi'_{i,n_2} = \Phi'_{0,j} = \Phi'_{n_1,j} = 0$ for all $i \in [n_1]$, $j \in [n_2]$. We need to bound

$$\|D^\top \Phi' \tilde{D}\|_F^2 = \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} (\Phi'_{i-1,j-1} + \Phi'_{i,j} - \Phi'_{i-1,j} - \Phi'_{i,j-1})^2.$$

By symmetry, it suffices to consider summation over $i \in [\frac{n_1-1}{2}]$, $j \in [\frac{n_2-1}{2}]$ where $\Phi' = \sqrt{ij} \wedge \sqrt{\frac{n_1 n_2}{k}}$. Moreover, note that the summand vanishes if $(i-1)(j-1) \geq \frac{n_1 n_2}{k}$. Hence we can further split the sum into two parts:

1. over $\{i \in [\frac{n_1-1}{2}], j \in [\frac{n_2-1}{2}] : (i-1)(j-1) < \frac{n_1 n_2}{k} < ij\}$, and
2. over $\{i \in [\frac{n_1-1}{2}], j \in [\frac{n_2-1}{2}] : ij \leq \frac{n_1 n_2}{k}\}$.

To bound the first part of the sum, first consider the case $i \leq j$, where we have

$$(i-1)(j-1) < (i-1)j \leq i(j-1) < ij.$$

Adjusting signs to ensure both differences in the following expression are positive, it is easily checked that

$$\left(\Phi'_{i-1,j-1} + \Phi'_{i,j} - \Phi'_{i-1,j} - \Phi'_{i,j-1}\right)^2 \leq \left[\sqrt{(i-1)j} - \sqrt{(i-1)(j-1)} + \sqrt{ij} - \sqrt{i(j-1)}\right]^2 \lesssim i/j.$$

By the conditions $(i-1)(j-1) < \frac{n_1 n_2}{k} < ij$ and $i \leq j$, we obtain

$$\begin{aligned} & i < \sqrt{\frac{n_1 n_2}{k}} + 1, \\ & \frac{n_1 n_2}{ki} < j < \frac{n_1 n_2}{k(i-1)} + 1, \text{ and} \\ & i/j < \frac{i^2 k}{n_1 n_2}. \end{aligned}$$

Therefore, it remains to bound

$$\sum_{i=1}^{\sqrt{\frac{n_1 n_2}{k}}} \sum_{j=\frac{n_1 n_2}{ki}}^{\frac{n_1 n_2}{k(i-1)}} \frac{i^2 k}{n_1 n_2} \lesssim \sum_{i=1}^{\sqrt{\frac{n_1 n_2}{k}}} \frac{n_1 n_2}{ki^2} \frac{i^2 k}{n_1 n_2} \leq \sqrt{\frac{n_1 n_2}{k}}.$$

An analogous argument yields the same bound for the case $i > j$.

Next, we consider the sum of $\left(\Phi'_{i-1,j-1} + \Phi'_{i,j} - \Phi'_{i-1,j} - \Phi'_{i,j-1}\right)^2$ over (i,j) such that $ij \leq \frac{n_1 n_2}{k}$, where we have

$$\begin{aligned} \left(\Phi'_{i-1,j-1} + \Phi'_{i,j} - \Phi'_{i-1,j} - \Phi'_{i,j-1}\right)^2 &= \left(\sqrt{(i-1)(j-1)} + \sqrt{ij} - \sqrt{(i-1)j} - \sqrt{i(j-1)}\right)^2 \\ &= \left(\sqrt{i} - \sqrt{i-1}\right)^2 \left(\sqrt{j} - \sqrt{j-1}\right)^2. \end{aligned}$$

Now even summing over all indices $i \in [n_1]$, $j \in [n_2]$ yields

$$\sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \left(\sqrt{i} - \sqrt{i-1}\right)^2 \left(\sqrt{j} - \sqrt{j-1}\right)^2 \lesssim \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \frac{1}{i} \frac{1}{j} \lesssim \log(n_1) \log(n_2).$$

Combining all the pieces, we obtain

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

5.6 Proof of Theorem 2

First, consider the set $\mathcal{C}$ of constant-row matrices, which is a subset of $\mathcal{M}$. Note that $V(\theta^*) = 0$ for any $\theta^* \in \mathcal{C}$. Since there are n_1 rows, it is not hard to see that the minimax rate of estimation over $\mathcal{C}$ is $\sigma^2 n_1$ in the squared Frobenius norm. The lower bound of order σ^2/n_2 then follows from normalization by $n_1 n_2$.

Next, we turn to the second term in the lower bound, which is based on Assouad's Lemma (Theorem 21). To this end, we construct an embedding of the hypercube into $\mathcal{M}_{V_0}$.

Consider integers $k_1 \in [n_1]$ and $k_2 \in [n_2]$, and let $m_1 = n_1/k_1$ and $m_2 = n_2/k_2$. Assume without loss of generality that m_1 and m_2 are integer. Denote the elements of the hypercube $\{-1,1\}^{k_1 \times k_2}$ by $(\tau_{u,v} : (u,v) \in [k_1] \times [k_2])$. For each $\tau \in \{-1,1\}^{k_1 \times k_2}$, define $\theta^\tau \in \mathcal{M}$ in the following way. For $i \in [n_1]$ and $j \in [n_2]$, first identify the unique $u \in [k_1], v \in [k_2]$ for which $(u-1)m_1 < i \le u m_1$ and $(v-1)m_2 < j \le v m_2$, and then take

$$\theta^\tau_{i,j} = V_0 \left(\frac{uv}{2k_1k_2} + \frac{\tau_{u,v}}{8k_1k_2} \right).$$

First, we check that $\theta^\tau \in \mathcal{M}$ and that $V(\theta^\tau) \le V_0$. For the former, it is enough to check that for each $1 \le i < n_1, 1 \le j < n_2$, we have that $\theta^\tau_{i,j} + \theta^\tau_{i+1,j+1} - \theta^\tau_{i,j+1} - \theta^\tau_{i+1,j} \ge 0$. We distinguish the following two cases:

1. There exists $u \in [k_1]$ such that $(u-1)m_1 < i < i+1 \le u m_1$, or there exists $v \in [k_2]$ such that $(v-1)m_2 < j < j+1 \le v m_2$. Then, either $\theta^\tau_{i,j} = \theta^\tau_{i+1,j}$, $\theta^\tau_{i,j+1} = \theta^\tau_{i+1,j+1}$ or $\theta^\tau_{i,j} = \theta^\tau_{i,j+1}$, $\theta^\tau_{i+1,j} = \theta^\tau_{i+1,j+1}$ respectively. In both cases, the difference above is 0.
2. There exist $u \in [k_1], v \in [k_2]$ such that $(u-1)m_1 < i \le u m_1 < i+1 \le (u+1)m_1$ and $(v-1)m_2 < j \le v m_2 < j+1 \le (v+1)m_2$. In this case, for any τ, we have

$$\begin{aligned}
&\theta^\tau_{i,j} + \theta^\tau_{i+1,j+1} - \theta^\tau_{i,j+1} - \theta^\tau_{i+1,j} \\
&= V_0 \left(\frac{uv + (u+1)(v+1) - u(v+1) - (u+1)v}{2k_1k_2} + \frac{\tau_{u,v} + \tau_{u+1,v+1} - \tau_{u,v+1} - \tau_{u+1,v}}{8k_1k_2} \right) \\
&\ge V_0 \left(\frac{1}{2k_1k_2} - \frac{4}{8k_1k_2} \right) = 0.
\end{aligned}$$

Thus, $\theta^\tau \in \mathcal{M}$. We now check that $V(\theta^\tau) \le V_0$. Note that $V(\theta^\tau)$ can be written as the sum $\sum_{i,j} (\theta^\tau_{i,j} + \theta^\tau_{i+1,j+1} - \theta^\tau_{i,j+1} - \theta^\tau_{i+1,j})$. As we have seen above, this sum is nonzero in only $(k_1-1)(k_2-1)$ cases, and it equals exactly

$$\begin{aligned}
V(\theta^\tau) &= \sum_{u \in [k_1-1], v \in [k_2-1]} V_0 \left(\frac{1}{2k_1k_2} + \frac{\tau_{u,v} + \tau_{u+1,v+1} - \tau_{u,v+1} - \tau_{u+1,v}}{8k_1k_2} \right) \\
&= V_0 \left(\frac{(k_1-1)(k_2-1)}{2k_1k_2} + \frac{\tau_{11} + \tau_{k_1k_2} - \tau_{1k_2} - \tau_{k_11}}{8k_1k_2} \right) \\
&\le V_0 \left(\frac{(k_1-1)(k_2-1)}{2k_1k_2} + \frac{4}{8k_1k_2} \right) = V_0 \frac{(k_1-1)(k_2-1)+1}{2k_1k_2} \le V_0.
\end{aligned}$$

We now proceed to show our lower bound using Theorem 21, which states that

$$\inf_{\tilde{\theta}} \sup_{\theta \in \mathcal{M}_{V_0}} R(\tilde{\theta}, \theta^*) \ge \frac{d}{8} \min_{\tau \neq \tau'} \frac{\ell^2(\theta^\tau, \theta^{\tau'})}{d_H(\tau, \tau')} \min_{d_H(\tau,\tau')=1} (1 - \|\mathbb{P}_{\theta^\tau} - \mathbb{P}_{\theta^{\tau'}}\|_{TV}),$$

where d_H denotes the Hamming distance and $\ell^2(\theta, \theta') = \frac{1}{n_1 n_2}\|\theta - \theta'\|_F^2$. Note that

$$\begin{aligned}
\ell^2(\theta^\tau, \theta^{\tau'}) &= \frac{1}{n_1 n_2} \sum_{i,j} (\theta^\tau_{i,j} - \theta^{\tau'}_{i,j})^2 \\
&= \frac{1}{n_1 n_2} \sum_{u \in [k_1], v \in [k_2]} \sum_{(u-1)m_1 < i \le u m_1} \sum_{(v-1)m_2 < j \le v m_2} (\theta^\tau_{i,j} - \theta^{\tau'}_{i,j})^2 \\
&= \frac{V_0^2}{n_1 n_2} \sum_{u \in [k_1], v \in [k_2]} \sum_{(u-1)m_1 < i \le u m_1} \sum_{(v-1)m_2 < j \le v m_2} \left(\frac{\tau_{u,v} - \tau'_{u,v}}{8 k_1 k_2} \right)^2 \\
&= \frac{V_0^2}{n_1 n_2} \sum_{u \in [k_1], v \in [k_2]} \frac{m_1 m_2 (\tau_{u,v} - \tau'_{u,v})^2}{64 k_1^2 k_2^2} = \frac{V_0^2 m_1 m_2}{16 n_1 n_2 k_1^2 k_2^2} d_H(\tau, \tau') = \frac{V_0^2}{16 k_1^3 k_2^3} d_H(\tau, \tau').
\end{aligned}$$

Thus, we have

$$\min_{\tau \neq \tau'} \frac{\ell^2(\theta^\tau, \theta^{\tau'})}{d_H(\tau, \tau')} = \frac{V_0^2}{16 k_1^3 k_2^3}.$$

To bound $\|\mathbb{P}_{\theta^\tau} - \mathbb{P}_{\theta^{\tau'}}\|_{TV}$, we use Pinsker's inequality:

$$\|\mathbb{P}_{\theta^\tau} - \mathbb{P}_{\theta^{\tau'}}\|_{TV}^2 \le \frac{1}{2} D(\mathbb{P}_{\theta^\tau} \| \mathbb{P}_{\theta^{\tau'}}) = \frac{n_1 n_2}{4\sigma^2} \ell^2(\theta^\tau, \theta^{\tau'}) = \frac{V_0^2 n_1 n_2}{64 k_1^3 k_2^3 \sigma^2} d_H(\tau, \tau').$$

It follows that

$$\min_{d_H(\tau, \tau') = 1} (1 - \|\mathbb{P}_{\theta^\tau} - \mathbb{P}_{\theta^{\tau'}}\|_{TV}) \ge \left(1 - \frac{V_0}{8\sigma} \sqrt{\frac{n_1 n_2}{k_1^3 k_2^3}} \right).$$

Putting things together in Assouad's lemma, we obtain

$$\inf_{\tilde{\theta}} \sup_{\theta \in \mathcal{M}_{V_0}} R(\tilde{\theta}, \theta^*) \ge \frac{V_0^2}{128 k_1^2 k_2^2} \left(1 - \frac{V_0}{8\sigma} \sqrt{\frac{n_1 n_2}{k_1^3 k_2^3}} \right).$$

If $\frac{4\sigma}{\sqrt{n_1 n_2}} \le V_0 \le 4\sigma n_1 n_2$, then we can choose k_1 and k_2 such that $k_1 k_2$ is of order $\left(\frac{V_0 \sqrt{n_1 n_2}}{4\sigma}\right)^{\frac{2}{3}}$, which yields that

$$\inf_{\tilde{\theta}} \sup_{\theta \in \mathcal{M}_{V_0}} R(\tilde{\theta}, \theta^*) \gtrsim \left(\frac{\sigma^2 V_0}{n_1 n_2} \right)^{2/3}.$$

If $V_0 \le \frac{4\sigma}{\sqrt{n_1 n_2}}$, then $\left(\frac{\sigma^2 V_0}{n_1 n_2}\right)^{2/3} \lesssim \frac{\sigma^2}{n_1 n_2}$, so the second term in the statement of Theorem 2 is dominated by the first term. Finally, if $V_0 \ge 4\sigma n_1 n_2$, then $\left(\frac{\sigma^2 V_0}{n_1 n_2}\right)^{2/3} \ge \sigma^2$, so the rate becomes trivial. This completes the proof.

5.7 Proof of Theorem 3

Similar to the proof of Theorem 1, the main technical step in this proof is to bound the supremum of an empirical process. This is dealt with in the following proposition, whose proof is deferred to Section 5.8. Note that both the statement and the proof are very similar to Proposition 9, and that we can restrict our attention to noise matrices that are Gaussian.

PROPOSITION 12. *Fix any matrix $\theta^* \in \mathbb{R}^{n_1 \times n_2}$ and permutations $\pi_1 \in \mathcal{S}_{n_1}$ and $\pi_2 \in \mathcal{S}_{n_2}$. 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*

$$\mathbb{E}\Big[\sup_{\substack{\theta \in \mathcal{M}_{V_0} \\ \|\theta(\pi_1,\pi_2)-\theta^*\|_F \le t}} \langle Z, \theta(\pi_1,\pi_2) - \theta^* \rangle\Big] \lesssim t\left[\sqrt{n_1} + \sqrt{k \log(n_2)}\right] + \sqrt{\frac{n_1 n_2}{k}} \sqrt{\log(n_1)\log(n_2)}\, V_0,$$

where we use the convention that the supremum over the empty set is $-\infty$.

We assume without loss of generality that the underlying true permutations π_1^* and π_2^* are the identities throughout the proof. For fixed permutations $\pi_1 \in \mathcal{S}_{n_1}$ and $\pi_2 \in \mathcal{S}_{n_2}$, we obtain from Theorem 19 and Proposition 12 that

$$\begin{aligned}
&\gamma_2\big(\{\theta(\pi_1,\pi_2) - \theta^* : \theta \in \mathcal{M}_{V_0},\ \|\theta(\pi_1,\pi_2) - \theta^*\|_F \le t\}\big) \\
&\asymp \mathbb{E}\Big[\sup_{\substack{\theta \in \mathcal{M}_{V_0} \\ \|\theta(\pi_1,\pi_2)-\theta^*\|_F \le t}} \langle \varepsilon, \theta(\pi_1,\pi_2) - \theta^* \rangle\Big] \\
&\lesssim t\left[\sqrt{n_1} + \sqrt{k \log(n_2)}\right] + \sqrt{\frac{n_1 n_2}{k}} \sqrt{\log(n_1)\log(n_2)}\, V_0.
\end{aligned}$$

Therefore, Theorem 20 yields that with probability $1 - 4\exp(-s^2)$,

$$\sup_{\substack{\theta \in \mathcal{M}_{V_0} \\ \|\theta(\pi_1,\pi_2)-\theta^*\|_F \le t}} \langle \varepsilon, \theta(\pi_1,\pi_2) - \theta^* \rangle \lesssim t\sigma\left[\sqrt{n_1} + \sqrt{k \log(n_2)}\right] + \sigma\sqrt{\frac{n_1 n_2}{k}} \sqrt{\log(n_1)\log(n_2)}\, V_0 + \sigma s t.$$

Taking $s = 2\sqrt{n_1 \log(n_1)}$ and applying a union bound over all $(\pi_1, \pi_2) \in \mathcal{S}_{n_1} \times \mathcal{S}_{n_2}$ (which has log-cardinality $\log(n_1! n_2!) \le 2 n_1 \log n_1$), we get that with probability at least $1 - n_1^{-n_1}$,

$$\begin{aligned}
&\sup_{\substack{(\pi_1,\pi_2,\theta) \in \mathcal{S}_{n_1} \times \mathcal{S}_{n_2} \times \mathcal{M}_{V_0} \\ \|\theta(\pi_1,\pi_2)-\theta^*\|_F \le t}} \langle \varepsilon, \theta(\pi_1,\pi_2) - \theta^* \rangle \\
&\qquad \lesssim t\sigma\left[\sqrt{n_1 \log(n_1)} + \sqrt{k \log(n_2)}\right] + \sigma\sqrt{\frac{n_1 n_2}{k}} \sqrt{\log(n_1)\log(n_2)}\, V_0.
\end{aligned}$$

Let us define

$$f_{\theta^*}(t) = \sup_{\substack{(\pi_1,\pi_2,\theta) \in \mathcal{S}_{n_1} \times \mathcal{S}_{n_2} \times \mathcal{M}_{V_0} \\ \|\theta(\pi_1,\pi_2)-\theta^*\|_F \le t}} \langle \varepsilon, \theta(\pi_1,\pi_2) - \theta^* \rangle - \frac{t^2}{2}.$$

Then for any

$$t > t^* := C\sigma\left[\sqrt{n_1 \log(n_1)} + \sqrt{k \log(n_2)}\right] + C\left[\sigma\sqrt{\frac{n_1 n_2}{k}} \sqrt{\log(n_1)\log(n_2)}\, V_0\right]^{1/2}$$

where C is a sufficiently large constant, it holds with probability at least $1 - n_1^{-n_1}$ that

$$f_{\theta^*}(t) < 0.$$

Therefore by Lemma 8, we obtain

$$\frac{1}{n_1 n_2}\|\hat{\theta}^{\mathsf{gls}} - \theta^*\|_F^2 \le \frac{(t^*)^2}{n_1 n_2} \lesssim \sigma^2 \left[\frac{\log(n_1)}{n_2} + \frac{k \log(n_2)}{n_1 n_2}\right] + \sigma \frac{\sqrt{\log(n_1)\log(n_2)}}{\sqrt{n_1 n_2 k}} V_0,$$

noting that by assumption, $\theta^* \in \bigcup_{\substack{\pi_1 \in \mathcal{S}_{n_1} \\ \pi_2 \in \mathcal{S}_{n_2}}} \mathcal{M}_{V_0}(\pi_1, \pi_2)$.

Balancing the terms that depend on k leads to the choice

$$k^* = (n_1 n_2)^{1/3} \left(\frac{\log(n_1)}{\log(n_2)}\right)^{1/3} \left(\frac{V_0}{\sigma}\right)^{2/3},$$

and therefore we obtain that with probability $1 - n_1^{-n_1}$,

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

if $1 \le k^* \le n_1 n_2$. We conclude by arguing similar to the proof of Theorem 1 to handle the other possible cases of k^* and to get an error bound in expectation instead of with high probability.

5.8 Proof of Proposition 12

Note that Proposition 12 is very similar to Proposition 9, and so are their proofs. The difference is that Proposition 12 has extra complications arising from the presence of permutations, while at the same time it is simpler because we restrict θ to $\mathcal{M}_{V_0} \subset \mathcal{M}$. Hence, we focus only the differences to the proof of Proposition 9 here.

Since Z is equal in distribution to $Z(\pi_1^{-1}, \pi_2^{-1})$, we have

$$\langle Z, \theta(\pi_1, \pi_2) - \theta^* \rangle \stackrel{d}{=} \langle Z, \theta - \theta^*(\pi_1^{-1}, \pi_2^{-1}) \rangle$$

in distribution for any permutations $\pi_1 \in \mathcal{S}_{n_1}, \pi_2 \in \mathcal{S}_{n_2}$. Therefore, by replacing $\theta^*(\pi_1^{-1}, \pi_2^{-1})$ with θ^*, it suffices to prove that for any matrix $\theta^* \in \mathbb{R}^{n_1 \times n_2}$, it holds that

$$\mathbb{E}\Big[\sup_{\substack{\theta \in \mathcal{M}_{V_0} \\ \|\theta - \theta^*\|_F \le t}} \langle Z, \theta - \theta^* \rangle\Big] \lesssim t\left[\sqrt{n_1} + \sqrt{k \log(n_2)}\right] + \sqrt{\frac{n_1 n_2}{k}} \log(n_1) \log(n_2)\, V_0.$$

Note that this supremum is very similar to that studied in Proposition 9, with the main differences being that θ^* can be any matrix in $\mathbb{R}^{n_1 \times n_2}$ while θ is restricted to $\mathcal{M}_{V_0}$. As in the proof of Proposition 9, (5.14), the supremum can be split into three terms:

$$\begin{aligned}\mathbb{E}[\sup_{\substack{\theta \in \mathcal{M}_{V_0} \\ \|\theta - \theta^*\|_F \le t}} \langle Z, \theta - \theta^* \rangle] \le \mathbb{E}[&\sup_{\substack{\theta \in \mathcal{M}_{V_0} \\ \|\theta - \theta^*\|_F \le t}} \langle \Pi_1(Z), \theta - \theta^* \rangle] + \mathbb{E}[\sup_{\substack{\theta \in \mathcal{M}_{V_0} \\ \|\theta - \theta^*\|_F \le t}} \langle (I - \Pi_1)\Pi_2(Z), \theta - \theta^* \rangle] \\ &+ \mathbb{E}[\sup_{\substack{\theta \in \mathcal{M}_{V_0} \\ \|\theta - \theta^*\|_F \le t}} \langle (I - \Pi_1)(I - \Pi_2)(Z), \theta - \theta^* \rangle].\end{aligned}$$

The first two terms can be bounded exactly as before, because we only need the condition $\|\theta - \theta^*\|_F \le t$ but not any other property of θ^*. Up to a constant factor, the third term can be bounded by (recall (5.16) and (5.17) in the proof of Proposition 9)

$$\sup_{\substack{\theta \in \mathcal{M}_{V_0} \\ \|\theta - \theta^*\|_F \le t}} \|\Phi \odot (D\theta\tilde{D}^\top)\|_1.$$

Thanks to the constraint $\theta \in \mathcal{M}_{V_0}$ in this case, we immediately obtain

$$\|\Phi \odot (D\theta\tilde{D}^\top)\|_1 = \langle \Phi, D\theta\tilde{D}^\top \rangle \leq \|\Phi\|_\infty \|D\theta\tilde{D}^\top\|_1 \lesssim \sqrt{\frac{n_1 n_2}{k}} \sqrt{\log(n_1)\log(n_2)}\, V_0.$$

Hence, it holds

$$\mathbb{E}\Big[\sup_{\substack{\theta\in\mathcal{M} \\ \|\theta-\theta^*\|_F \leq t}} \langle (I-\Pi_1)(I-\Pi_2)(Z), \theta-\theta^* \rangle \Big] \lesssim \sqrt{\frac{n_1 n_2}{k}} \sqrt{\log(n_1)\log(n_2)}\, V_0.$$

Combining the bounds on the three terms completes the proof.

5.9 Proof of Theorem 4

We assume without loss of generality that the underlying true permutations π_1^* and π_2^* are the identities throughout the proof, except where the notations π_1^* and π_2^* are explicitly used. Recall that we defined the reversal permutation $\pi_1^{\mathsf{r}} \in \mathcal{S}_{n_1}$ by $\pi_1^{\mathsf{r}}(i) = n_1 - i + 1$ for $i \in [n_1]$. Given estimators $\hat{\pi}_1, \hat{\pi}_2$, let us define

$$\tilde{\theta} := \operatorname*{argmin}_{\theta \in \mathcal{M}_{V_0}(\hat{\pi}_1,\hat{\pi}_2) \cup \mathcal{M}_{V_0}(\pi_1^{\mathsf{r}} \circ \hat{\pi}_1, \hat{\pi}_2)} \|\theta - \theta^*\|_F^2. \tag{5.29}$$

The theorem follows from the next two propositions combined. The first proposition says that the final denoising error can be bounded by the sum of the minimax rate (the error rate incurred by the projection step of the algorithm), and the error incurred by the permutation estimators. The second proposition controls the error incurred by the permutation estimators.

Proposition 13. *Suppose that we have $y = \theta^* + \varepsilon$, where the noise matrix ε has independent* $\mathsf{subG}(C\sigma^2)$ *entries with $\mathrm{Var}[\varepsilon_{i,j}] = \sigma^2$. Let the estimators $(\hat{\pi}_1, \hat{\pi}_1', \hat{\pi}_2, \hat{\theta})$ be given by Algorithm 2, and define $\tilde{\theta}$ according to (5.29). Then it holds with probability at least $1 - n_1^{-n_1}$ that*

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

Proposition 14. *Suppose that we have $y = \theta^* + \varepsilon$, where $\theta^* \in \mathcal{M}_{V_0}^{n_1,n_2}$ and the noise matrix ε has independent* $\mathsf{subG}(C\sigma^2)$ *entries. For the permutation estimators $\hat{\pi}_1$ and $\hat{\pi}_2$ given by the* Variance Sorting *subroutine, Algorithm 1, let $\tilde{\theta}$ be defined by (5.29). Then it holds with probability at least $1 - n_1^{-9}$ that*

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

The error bound in Proposition 14 is clearly dominating, and combining the two propositions yields the statement of the theorem in probability. Taking into account that we can do the same analysis keeping track of the failure probability independently of n_1 and integrate, we obtain bounds in expectation instead of probability as well.

5.10 Proof of Proposition 13

To employ the variational formula in Lemma 8, let us view $\tilde{\theta}$ defined by (5.29) as the ground truth, and view $y - \tilde{\theta} = \varepsilon + \theta^* - \tilde{\theta}$ as the noise. Correspondingly, we define

$$f_{\tilde{\theta}}(t) := \sup_{\substack{\theta \in \mathcal{M}_{V_0}(\hat{\pi}_1, \hat{\pi}_2) \cup \mathcal{M}_{V_0}(\pi_1^{\mathsf{r}} \circ \hat{\pi}_1, \hat{\pi}_2) \\ \|\theta - \tilde{\theta}\|_F \le t}} \langle \varepsilon + \theta^* - \tilde{\theta}, \theta - \tilde{\theta} \rangle - \frac{t^2}{2}. \tag{5.30}$$

To facilitate our analysis, for each pair of permutations $(\pi_1, \pi_2) \in \mathcal{S}_{n_1} \times \mathcal{S}_{n_2}$, we define

$$\tilde{\theta}^{\pi_1, \pi_2} := \operatorname*{argmin}_{\theta \in \mathcal{M}_{V_0}(\pi_1, \pi_2)} \|\theta - \theta^*\|_F^2,$$

and note that we have that either $\tilde{\theta} = \tilde{\theta}^{\hat{\pi}_1, \hat{\pi}_2}$ or $\tilde{\theta} = \tilde{\theta}^{\pi_1^{\mathsf{r}} \circ \hat{\pi}_1, \hat{\pi}_2}$, so $\tilde{\theta} \in \mathcal{M}_{V_0}(\hat{\pi}_1, \hat{\pi}_2) \cup \mathcal{M}_{V_0}(\pi_1^{\mathsf{r}} \circ \hat{\pi}_1, \hat{\pi}_2)$ and Lemma 8 is applicable.

We further estimate the supremum in (5.30) by

$$\begin{aligned} f_{\tilde{\theta}}(t) \le & \sup_{\substack{\theta \in \mathcal{M}_{V_0}(\hat{\pi}_1, \hat{\pi}_2) \cup \mathcal{M}_{V_0}(\pi_1^{\mathsf{r}} \circ \hat{\pi}_1, \hat{\pi}_2) \\ \|\theta - \tilde{\theta}\|_F \le t}} \langle \varepsilon, \theta - \tilde{\theta} \rangle \\ & \quad + \sup_{\substack{\theta \in \mathcal{M}_{V_0}(\hat{\pi}_1, \hat{\pi}_2) \cup \mathcal{M}_{V_0}(\pi_1^{\mathsf{r}} \circ \hat{\pi}_1, \hat{\pi}_2) \\ \|\theta - \tilde{\theta}\|_F \le t}} \langle \theta^* - \tilde{\theta}, \theta - \tilde{\theta} \rangle - \frac{t^2}{2} & (5.31) \\ \le & \sup_{\substack{(\pi_1, \pi_1', \pi_2, \pi_2', \theta) \in \mathcal{S}_{n_1}^2 \times \mathcal{S}_{n_2}^2 \times \mathcal{M}_{V_0} \\ \|\theta(\pi_1', \pi_2') - \tilde{\theta}^{\pi_1, \pi_2}\|_F \le t}} \langle \varepsilon, \theta(\pi_1', \pi_2') - \tilde{\theta}^{\pi_1, \pi_2} \rangle + t\|\theta^* - \tilde{\theta}\|_F - \frac{t^2}{2}. & (5.32) \end{aligned}$$

Note that the random variables $\hat{\pi}_1, \hat{\pi}_2$ and $\tilde{\theta}$ depend on ε, so it is not clear how to control the first supremum in (5.31). Instead, in (5.32) we take a supremum over all $(\pi_1, \pi_1', \pi_2, \pi_2', \theta) \in \mathcal{S}_{n_1}^2 \times \mathcal{S}_{n_2}^2 \times \mathcal{M}_{V_0}$, where each individual quantity $\theta(\pi_1', \pi_2') - \tilde{\theta}^{\pi_1, \pi_2}$ is deterministic.

For fixed permutations $\pi_1, \pi_1' \in \mathcal{S}_{n_1}$ and $\pi_2, \pi_2' \in \mathcal{S}_{n_2}$, we obtain from Theorem 19 and Proposition 12 that

$$\begin{aligned} & \gamma_2\big(\{\theta(\pi_1', \pi_2') - \tilde{\theta}^{\pi_1, \pi_2} : \theta \in \mathcal{M}_{V_0}, \ \|\theta(\pi_1', \pi_2') - \tilde{\theta}^{\pi_1, \pi_2}\|_F \le t\}\big) \\ & \quad \asymp \mathbb{E}_{Z_{i,j} \overset{\text{iid}}{\sim} \mathcal{N}(0,1)} \Big[\sup_{\substack{\theta \in \mathcal{M}_{V_0} \\ \|\theta(\pi_1', \pi_2') - \tilde{\theta}^{\pi_1, \pi_2}\|_F \le t}} \langle Z, \theta(\pi_1', \pi_2') - \tilde{\theta}^{\pi_1, \pi_2} \rangle \Big] \\ & \quad \lesssim t \left[\sqrt{n_1} + \sqrt{k \log(n_2)} \right] + \sqrt{\frac{n_1 n_2}{k}} \sqrt{\log(n_1) \log(n_2)}\, V_0. \end{aligned}$$

Therefore, Theorem 20 yields that with probability $1 - 4\exp(-s^2)$,

$$\begin{aligned} & \sup_{\substack{\theta \in \mathcal{M}_{V_0} \\ \|\theta(\pi_1', \pi_2') - \tilde{\theta}^{\pi_1, \pi_2}\|_F \le t}} \langle \varepsilon, \theta(\pi_1', \pi_2') - \tilde{\theta}^{\pi_1, \pi_2} \rangle \\ & \lesssim t\sigma \left[\sqrt{n_1} + \sqrt{k \log(n_2)} \right] + \sigma \sqrt{\frac{n_1 n_2}{k}} \sqrt{\log(n_1) \log(n_2)}\, V_0 + \sigma s t. \end{aligned}$$

Taking $s = 3\sqrt{n_1 \log(n_1)}$ and applying a union bound over all $\pi_1, \pi_1' \in \mathcal{S}_{n_1}$ and $\pi_2, \pi_2' \in \mathcal{S}_{n_2}$, we see that with probability at least $1 - n_1^{-n_1}$,

$$\sup_{\substack{(\pi_1,\pi_1',\pi_2,\pi_2',\theta)\in\mathcal{S}_{n_1}^2\times\mathcal{S}_{n_2}^2\times\mathcal{M}_{V_0} \\ \|\theta(\pi_1',\pi_2')-\tilde{\theta}^{\pi_1,\pi_2}\|_F\le t}} \langle \varepsilon, \theta(\pi_1', \pi_2') - \tilde{\theta}^{\pi_1,\pi_2} \rangle$$
$$\lesssim t\sigma \left[\sqrt{n_1 \log(n_1)} + \sqrt{k \log(n_2)}\right] + \sigma \sqrt{\frac{n_1 n_2}{k}} \sqrt{\log(n_1)\log(n_2)}\, V_0.$$

This together with inequality (5.32) yields that for any

$$t > t^* := C\sigma \left[\sqrt{n_1 \log(n_1)} + \sqrt{k \log(n_2)}\right] + C\left[\sigma \sqrt{\frac{n_1 n_2}{k}} \sqrt{\log(n_1)\log(n_2)}\, V_0\right]^{1/2} + C\|\tilde{\theta} - \theta^*\|_F$$

where C is a sufficiently large constant, it holds with probability at least $1 - n_1^{-n_1}$ that

$$f_{\tilde{\theta}}(t) < 0.$$

Therefore, by Lemma 8 applied to the set $\mathcal{M}_{V_0}(\hat{\pi}_1, \hat{\pi}_2) \cup \mathcal{M}_{V_0}(\pi_1^{\mathsf{r}} \circ \hat{\pi}_1, \hat{\pi}_2)$, we obtain

$$\frac{1}{n_1 n_2}\|\hat{\theta} - \tilde{\theta}\|_F^2 \le \frac{(t^*)^2}{n_1 n_2} \lesssim \sigma^2 \left[\frac{\log(n_1)}{n_2} + \frac{k \log(n_2)}{n_1 n_2}\right] + \sigma \frac{\sqrt{\log(n_1)\log(n_2)}}{\sqrt{n_1 n_2 k}} V_0 + \frac{1}{n_1 n_2}\|\tilde{\theta} - \theta^*\|_F^2.$$

Balancing the terms that depend on k yields that with probability $1 - n_1^{-n_1}$,

$$\frac{1}{n_1 n_2}\|\hat{\theta} - \theta^*\|_F^2 \lesssim \frac{1}{n_1 n_2}\|\hat{\theta} - \tilde{\theta}\|_F^2 + \frac{1}{n_1 n_2}\|\tilde{\theta} - \theta^*\|_F^2$$
$$\lesssim \frac{\sigma^2 \log(n_1)}{n_2} + \left(\frac{\sigma^2 V_0}{n_1 n_2}\right)^{2/3} \log(n_1)^{1/3} \log(n_2)^{2/3} + \frac{1}{n_1 n_2}\|\tilde{\theta} - \theta^*\|_F^2,$$

if for the optimal k^*, we have $1 \le k^* \le n_1 n_2$. The other cases can be handled as in the proof of Theorem 1.

5.11 Proof of Proposition 14

5.11.1 Reduction to the row/column-centered case. First, we reduce the problem to the case where the underlying matrix θ^* has centered rows and columns. If θ^* is not centered, we may choose a matrix $R \in \mathbb{R}^{n_1 \times n_2}$ with constant rows and a matrix $S \in \mathbb{R}^{n_1 \times n_2}$ with constant columns, so that if $\bar{\theta} := \theta^* - R - S$, then for all $i \in [n_1]$, $j \in [n_2]$,

$$\sum_{k=1}^{n_1} \bar{\theta}_{k,j} = \sum_{\ell=1}^{n_2} \bar{\theta}_{i,\ell} = 0.$$

Since R and S have constant rows and columns respectively, we have $V(\bar{\theta}) = V(\theta^*)$ by definition (2.3), and thus $\bar{\theta} \in \mathcal{M}_{V_0}$. More importantly, according to the definition of $\xi(i,j)$ in (3.7), its value does not change if we replace y by $y - R - S$. Therefore, we may assume without loss of generality that

$$y = \bar{\theta} + \varepsilon,$$

which does not change the estimators $\hat{\pi}_1$, $\hat{\pi}_1'$ and $\hat{\pi}_2$ output by the algorithm.

Furthermore, we have

$$\min_{\theta \in \mathcal{M}_{V_0}(\hat{\pi}_1,\hat{\pi}_2) \cup \mathcal{M}_{V_0}(\pi_1^{\mathsf{r}} \circ \hat{\pi}_1,\hat{\pi}_2)} \|\theta - \theta^*\|_F^2 = \min_{\theta \in \mathcal{M}_{V_0}(\hat{\pi}_1,\hat{\pi}_2) \cup \mathcal{M}_{V_0}(\pi_1^{\mathsf{r}} \circ \hat{\pi}_1,\hat{\pi}_2)} \|\theta - \bar{\theta}\|_F^2, \tag{5.33}$$

since if $\tilde{\theta}$ minimizes the left-hand side, then $\tilde{\theta} - R - S$ minimizes the right-hand side. Hence it suffices to show that the right-hand side of (5.33) is bounded by the desired rate.

Note that by symmetry of the anti-Monge constraint,

$$\mathcal{M}_{V_0}(\pi_1^{\mathsf{r}} \circ \hat{\pi}_1, \hat{\pi}_2) = \mathcal{M}_{V_0}(\hat{\pi}_1, \pi_2^{\mathsf{r}} \circ \hat{\pi}_2) \quad \text{and}$$
$$\mathcal{M}_{V_0}(\hat{\pi}_1, \hat{\pi}_2) = \mathcal{M}_{V_0}(\pi_1^{\mathsf{r}} \circ \hat{\pi}_1, \pi_1^{\mathsf{r}} \circ \hat{\pi}_2).$$

In light of this, it is sufficient to show

$$\min_{\substack{\pi_1 \in \{\mathsf{id}, \pi_1^{\mathsf{r}}\} \\ \pi_2 \in \{\mathsf{id}, \pi_2^{\mathsf{r}}\}}} \|\bar{\theta}(\pi_1 \circ \hat{\pi}_1, \pi_2 \circ \hat{\pi}_2) - \bar{\theta}\|_F^2 \lesssim \big[\sigma^2 + \sigma V(\bar{\theta})\big] n_1 \sqrt{n_2 \log n_1}, \tag{5.34}$$

which we do in the sequel. This gives an upper bound on (5.33) and thus completes the proof.

5.11.2 Preliminaries. Before proceeding to proving (5.34), we start with some lemmas.

LEMMA 15. *For $\gamma \in \mathbb{R}^n$ such that $\sum_{k=1}^n \gamma_k = 0$, it holds that*

$$\sum_{k<\ell} (\gamma_k - \gamma_\ell)^2 = n \sum_{k=1}^n \gamma_k^2.$$

The proof follows by inspection.

LEMMA 16. *Let $f : [n] \times [n] \to \mathbb{R}$ be a symmetric bivariate function such that*

$$f(i, m) \vee f(m, j) \le f(i, j), \qquad \text{for all } 1 \le i \le m \le j \le n. \tag{5.35}$$

Let $\pi : [n] \to [n]$ be a permutation and $\tau \in \mathbb{R}$. Suppose that

$$f(i, j) \le \tau, \qquad \text{if } i < j \text{ and } \pi(i) > \pi(j). \tag{5.36}$$

Then we have $f(\pi(i), i) \le \tau$ for all $i \in [n]$.

PROOF. Suppose that $f(\pi(j), j) > \tau$ and $\pi(j) < j$ for some index $j \in [n]$. Since π is a bijection, there must exist an index $i \le \pi(j) < j$ such that $\pi(i) > \pi(j)$. However, by (5.35) we then have

$$f(i, j) \ge f(\pi(j), j) > \tau,$$

which contradicts assumption (5.36). A similar argument yields a contradiction in the case that $f(\pi(j), j) > \tau$ and $\pi(j) > j$. Therefore, we obtain that $f(\pi(j), j) \le \tau$ for all $j \in [n]$. □

Next, we study the quantity $\xi(i, j)$ used in the algorithm defined by (3.7). Throughout the rest of the proof, we use the notation

$$f(i, j) := \sum_{k=1}^{n_2} (\bar{\theta}_{i,k} - \bar{\theta}_{j,k})^2. \tag{5.37}$$

LEMMA 17. *Suppose that $y = \bar{\theta} + \varepsilon$, where $\bar{\theta}$ has centered rows and columns, and ε has independent $\mathsf{subG}(C\sigma^2)$ entries with $\mathrm{Var}[\varepsilon_{i,j}] = \sigma^2$. Then it holds that for all distinct $i, j \in [n_1]$,*

$$\mathbb{E}[\xi(i,j)] = f(i,j) + 2(n_2 - 1)\sigma^2, \tag{5.38}$$

and that with probability $1 - n_1^{-10}$, for all $i, j \in [n_1]$,

$$\big|\xi(i,j) - \mathbb{E}[\xi(i,j)]\big| \le \tau := C\big[\sigma^2 + \sigma V(\bar{\theta})\big]\sqrt{n_2 \log n_1}, \tag{5.39}$$

where $C > 0$ is a universal constant and $V(\bar{\theta})$ is defined as in (2.3).

The proof of the lemma is deferred to Section 5.12.

Next, we study properties of the expectation $\mathbb{E}[\xi(i,j)]$, or equivalently $f(i,j)$, thanks to (5.38).

LEMMA 18. *It holds for all $1 \le i \le m \le j \le n_1$ that $f(i,m) + f(m,j) \le f(i,j)$.*

PROOF. Fix indices $1 \le i \le m \le j \le n_1$. Since all the row sums of $\bar{\theta}$ are zero by assumption, it follows from Lemma 15 that

$$f(i,j) = \sum_{k=1}^{n_2} \big(\bar{\theta}_{i,k} - \bar{\theta}_{j,k}\big)^2 = \frac{1}{n_2} \sum_{k<\ell} \big(\bar{\theta}_{i,k} - \bar{\theta}_{j,k} - \bar{\theta}_{i,\ell} + \bar{\theta}_{j,\ell}\big)^2. \tag{5.40}$$

Note that we have

$$\big(\bar{\theta}_{i,k} - \bar{\theta}_{m,k} - \bar{\theta}_{i,\ell} + \bar{\theta}_{m,\ell}\big) + \big(\bar{\theta}_{m,k} - \bar{\theta}_{j,k} - \bar{\theta}_{m,\ell} + \bar{\theta}_{j,\ell}\big) = \big(\bar{\theta}_{i,k} - \bar{\theta}_{j,k} - \bar{\theta}_{i,\ell} + \bar{\theta}_{j,\ell}\big),$$

where each of the three bracketed terms is nonnegative because $\bar{\theta}$ is anti-Monge. Therefore, we obtain

$$\big(\bar{\theta}_{i,k} - \bar{\theta}_{m,k} - \bar{\theta}_{i,\ell} + \bar{\theta}_{m,\ell}\big)^2 + \big(\bar{\theta}_{m,k} - \bar{\theta}_{j,k} - \bar{\theta}_{m,\ell} + \bar{\theta}_{j,\ell}\big)^2 \le \big(\bar{\theta}_{i,k} - \bar{\theta}_{j,k} - \bar{\theta}_{i,\ell} + \bar{\theta}_{j,\ell}\big)^2.$$

This together with (5.40) completes the proof. □

5.11.3 Proof of main bound (5.34). We condition on the event of probability $1 - n_1^{-10}$ that (5.39) holds. Consider the permutation estimator $\hat{\pi}_1$ defined as in the algorithm so that $\{\xi(i_0, \hat{\pi}_1^{-1}(i))\}_{i=1}^{n_1}$ is nondecreasing. Note that we made the decision of whether to consider i_0 or j_0 as an estimator for $\hat{\pi}_1^{-1}(1)$ arbitrarily by demanding $i_0 < j_0$. In the following, we make use of the fact that this orientation aligns with the assumption of $\pi_1^* = \mathsf{id}$, that is, we use $\pi_1^*(i_0) < \pi_1^*(j_0)$. If the reverse inequality holds, then we may repeat the same proof with $\hat{\pi}_1$ replaced by $\pi_1^{\mathsf{r}} \circ \hat{\pi}_1$, to obtain permutation guarantees for the reversed permutation instead.

We claim that for f defined in (5.37),

$$f(i,j) \le 12\tau, \quad \text{if } i < j \text{ and } \hat{\pi}_1(i) > \hat{\pi}_1(j). \tag{5.41}$$

To establish the claim, we first consider any pair (i,j) for which $i_0 < i < j$ and $\hat{\pi}_1(i) > \hat{\pi}_1(j)$. Thus by the definition of $\hat{\pi}_1$, we have $\xi(i_0, i) > \xi(i_0, j)$. Then it follows from Lemma 18, (5.38) and (5.39) that

$$f(i,j) \le f(i_0,j) - f(i_0,i) = \mathbb{E}[\xi(i_0,j)] - \mathbb{E}[\xi(i_0,i)] \le \xi(i_0,j) - \xi(i_0,i) + 2\tau \le 2\tau.$$

Next, consider any pair (i, j) where $i \le i_0$ and $\hat{\pi}_1(i) > \hat{\pi}_1(j)$. By the definition of (i_0, j_0) in (3.8), we have that $i_0 < j_0$ and $\xi(i_0, j_0) \ge \xi(i, j_0)$. Together with Lemma 18, (5.38) and (5.39), this implies that

$$f(i, i_0) \le f(i, j_0) - f(i_0, j_0) = \mathbb{E}[\xi(i, j_0)] - \mathbb{E}[\xi(i_0, j_0)] \le \xi(i, j_0) - \xi(i_0, j_0) + 2\tau \le 2\tau. \tag{5.42}$$

Moreover, we have $\xi(i, i_0) \ge \xi(i_0, j)$ since $\hat{\pi}_1(i) > \hat{\pi}_1(j)$. Therefore, by (5.38) and (5.39) it holds

$$f(i_0, j) - f(i, i_0) = \mathbb{E}[\xi(i_0, j)] - \mathbb{E}[\xi(i, i_0)] \le \xi(i_0, j) - \xi(i, i_0) + 2\tau \le 2\tau. \tag{5.43}$$

Combining (5.42) and (5.43), we obtain

$$\begin{aligned} f(i, j) = \sum_{k=1}^{n_2} \big(\bar{\theta}_{j,k} - \bar{\theta}_{i,k}\big)^2 &\le 2 \sum_{k=1}^{n_2} \Big[\big(\bar{\theta}_{j,k} - \bar{\theta}_{i_0,k}\big)^2 + \big(\bar{\theta}_{i_0,k} - \bar{\theta}_{i,k}\big)^2\Big] \\ &= 2f(i_0, j) + 2f(i, i_0) \le 12\tau. \end{aligned}$$

Therefore, claim (5.41) is established.

Note that assumption (5.35) of Lemma 16 holds in view of Lemma 18 and the fact that $f(i, j) \ge 0$. Hence claim (5.41) and Lemma 16 together yield that for all $i \in [n_1]$,

$$f(\hat{\pi}_1(i), i) = \sum_{k=1}^{n_2} \big(\bar{\theta}_{\hat{\pi}_1(i),k} - \bar{\theta}_{i,k}\big)^2 \le 12\tau.$$

Summing over i, we conclude that

$$\|\bar{\theta}(\hat{\pi}_1, \mathsf{id}) - \bar{\theta}\|_F^2 \le 12\tau n_1,$$

contingent on the assumption that for the ground truth permutation π_1^*, we have $(\pi_1^*)^{-1}(i_0) < (\pi_1^*)^{-1}(j_0)$. If not, repeat the same proof with $\hat{\pi}_1$ replaced by $\pi_1^{\mathsf{r}} \circ \hat{\pi}_1$ to obtain

$$\|\bar{\theta}(\pi_1^{\mathsf{r}} \circ \hat{\pi}_1, \mathsf{id}) - \bar{\theta}\|_F^2 \le 12\tau n_1.$$

Finally, the proof remains valid if we replace $\bar{\theta}$ and y by their transposes, and switch the roles of row and column indices. Hence it also holds with probability $1 - n_1^{-10}$ that

$$\|\bar{\theta}(\mathsf{id}, \hat{\pi}_2) - \bar{\theta}\|_F^2 \wedge \|\bar{\theta}(\mathsf{id}, \pi_2^{\mathsf{r}} \circ \hat{\pi}_2) - \bar{\theta}\|_F^2 \lesssim \big[\sigma^2 + \sigma V(\bar{\theta})\big] n_2 \sqrt{n_1 \log n_1}.$$

We then complete the proof of (5.34) by using the triangle inequality to see that for some choice of $\bar{\pi}_1 \in \{\mathsf{id}, \pi_1^{\mathsf{r}}\}$, $\bar{\pi}_2 \in \{\mathsf{id}, \pi_2^{\mathsf{r}}\}$, it holds that

$$\begin{aligned} \|\bar{\theta}(\bar{\pi}_1 \circ \hat{\pi}_1, \bar{\pi}_2 \circ \hat{\pi}_2) - \bar{\theta}\|_F^2 &\le 2\|\bar{\theta}(\bar{\pi}_1 \circ \hat{\pi}_1, \bar{\pi}_2 \circ \hat{\pi}_2) - \bar{\theta}(\bar{\pi}_1 \circ \hat{\pi}_1, \mathsf{id})\|_F^2 + 2\|\bar{\theta}(\bar{\pi}_1 \circ \hat{\pi}_1, \mathsf{id}) - \bar{\theta}\|_F^2 \\ &= 2\|\bar{\theta}(\mathsf{id}, \bar{\pi}_2 \circ \hat{\pi}_2) - \bar{\theta}\|_F^2 + 2\|\bar{\theta}(\bar{\pi}_1 \circ \hat{\pi}_1, \mathsf{id}) - \bar{\theta}\|_F^2 \\ &\lesssim \big[\sigma^2 + \sigma V(\bar{\theta})\big] n_1 \sqrt{n_2 \log n_1}. \end{aligned}$$

5.12 Proof of Lemma 17

Recall that $y = \bar{\theta} + \varepsilon$ where $\mathrm{Var}[\varepsilon_{i,j}] = \sigma^2$, and recall the notation

$$\xi(i,j) = \sum_{k=1}^{n_2} \Big[y_{i,k} - y_{j,k} - \frac{1}{n_2} \sum_{\ell=1}^{n_2} (y_{i,\ell} - y_{j,\ell}) \Big]^2.$$

For the statement about its expectation, (5.38), we need to prove that for distinct $i, j \in [n_1]$,

$$\mathbb{E}[\xi(i,j)] = \sum_{k=1}^{n_2} (\bar{\theta}_{i,k} - \bar{\theta}_{j,k})^2 + 2(n_2 - 1)\sigma^2. \tag{5.44}$$

For the statement about its deviation, (5.39), we claim that it suffices to prove that with probability $1 - n_1^{-12}$,

$$\big|\xi(i,j) - \mathbb{E}[\xi(i,j)]\big| \lesssim \sigma^2 \sqrt{n_2 \log n_1} + \sigma \Big[\sum_{k=1}^{n_2} (\bar{\theta}_{i,k} - \bar{\theta}_{j,k})^2 \Big]^{1/2} \sqrt{\log n_1}. \tag{5.45}$$

To see this, note that Lemma 15 gives

$$\sum_{k=1}^{n_2} (\bar{\theta}_{i,k} - \bar{\theta}_{j,k})^2 = \frac{1}{n_2} \sum_{k<\ell} (\bar{\theta}_{i,k} - \bar{\theta}_{j,k} - \bar{\theta}_{i,\ell} + \bar{\theta}_{j,\ell})^2 \le \sum_{k<\ell} \frac{1}{n_2} V_0^2 \le n_2 V(\bar{\theta})^2,$$

where we used that $|\bar{\theta}_{i,k} + \bar{\theta}_{j,\ell} - \bar{\theta}_{j,k} - \bar{\theta}_{i,\ell}| \le V_0$. Plugging this bound into (5.45) and applying a union bound over all $i, j \in [n_1]$ then completes the proof.

The claims (5.44) and (5.45) can be simplified as follows. For distinct $i, j \in [n_1]$, we let

$$n = n_2, \quad x = y_{i,\cdot} - y_{j,\cdot}, \quad \gamma = \bar{\theta}_{i,\cdot} - \bar{\theta}_{j,\cdot}, \quad \delta = \varepsilon_{i,\cdot} - \varepsilon_{j,\cdot}, \quad \text{and} \quad \zeta = \sum_{k=1}^{n} \Big(x_k - \frac{1}{n} \sum_{\ell=1}^{n} x_\ell \Big)^2.$$

Note that δ has independent $\mathsf{subG}(C\sigma^2)$ entries and $\mathbb{E}[\delta_k^2] = 2\sigma^2$. We need to prove that

$$\mathbb{E}[\zeta] = \sum_{k=1}^{n} \gamma_k^2 + 2(n-1)\sigma^2,$$

and that with probability $1 - n_1^{-12}$,

$$\big|\zeta - \mathbb{E}[\zeta]\big| \lesssim \sigma^2 \sqrt{n \log n_1} + \sigma \Big(\sum_{k=1}^{n} \gamma_k^2 \Big)^{1/2} \sqrt{\log n_1}.$$

Recall that all the rows and columns of $\bar{\theta}$ are centered by assumption, so we have $\sum_{k=1}^{n} \gamma_k = 0$. Using this, we get the following Hoeffding decomposition

$$\zeta = \sum_{k=1}^{n} \Big(x_k - \frac{1}{n} \sum_{i=1}^{n} x_i \Big)^2 = \sum_{k=1}^{n} \gamma_k^2 + \sum_{k=1}^{n} \delta_k^2 - \frac{1}{n} \Big(\sum_{k=1}^{n} \delta_k \Big)^2 + 2 \sum_{k=1}^{n} \gamma_k \delta_k.$$

In particular, it follows that

$$\mathbb{E}[\zeta] = \|\gamma\|_2^2 + 2(n-1)\sigma^2.$$

Moreover, we have that with probability $1 - n_1^{-12}$,

$$\big|\zeta - \mathbb{E}[\zeta]\big| \le \Big|\sum_{k=1}^{n} \delta_k^2 - 2n\sigma^2\Big| + \Big|\frac{1}{n}\Big(\sum_{k=1}^{n} \delta_k\Big)^2 - 2\sigma^2\Big| + 2\Big|\sum_{k=1}^{n} \gamma_k \delta_k\Big|$$
$$\lesssim \sigma^2 \sqrt{n \log n_1} + \sigma \|\gamma\|_2 \sqrt{\log n_1},$$

where concentration of the first two terms is due to Lemma 22, and the last term is $\mathsf{subG}(C\sigma^2\|\gamma\|_2^2)$. This completes the proof.

5.13 Proof of Theorem 5

This proof technique was developed by [17, 62], and our presentation follows that of Theorem 2 of [62]. We assume without loss of generality that the underlying true permutations π_1^* and π_2^* are the identities throughout the proof. Let $\|\cdot\|$ denotes the operator norm of a matrix. It is well known (see Theorem 4.4.5 of [68]) that $\|\varepsilon\| \le C_1 \sigma \sqrt{n_1}$ with probability at least $1 - \exp(-n_1)$. We condition on this event in the sequel.

Recall that the singular value decomposition of y is

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

where $\lambda_1, \dots, \lambda_{n_2}$ are ordered non-increasingly, and the SVT 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 $\rho = 2C_1\sigma\sqrt{n_1}$. Moreover, write the singular value decomposition of θ^* as

$$\theta^* = \sum_{i=1}^{n_2} \lambda_i^* u_i^* (v_i^*)^\top,$$

where $\lambda_1^*, \dots, \lambda_{n_2}^*$ are ordered non-increasingly. Let s be the number of singular values of θ^* that are larger than $C_1\sigma\sqrt{n_1}$, and define

$$\theta_s = \sum_{i=1}^{s} \lambda_i^* u_i^* (v_i^*)^\top.$$

Note that the for each $i > s$, by Weyl's inequality, we have

$$\lambda_i \le \lambda_i^* + \|\varepsilon\| \le C_1\sigma\sqrt{n_1} + C_1\sigma\sqrt{n_1} = \rho.$$

Therefore, $\hat{\theta}^{\mathsf{svt}}$ has rank at most s, and so has θ_s. It follows that

$$\|\hat{\theta}^{\mathsf{svt}} - \theta^*\|_F \le \|\hat{\theta}^{\mathsf{svt}} - \theta_s\|_F + \|\theta_s - \theta^*\|_F \le \sqrt{2s}\, \|\hat{\theta}^{\mathsf{svt}} - \theta_s\| + \Big[\sum_{i=s+1}^{n_2} (\lambda_i^*)^2\Big]^{1/2}.$$

Moreover, it holds that

$$\|\hat{\theta}^{\mathsf{svt}} - \theta_s\| \le \|\hat{\theta}^{\mathsf{svt}} - y\| + \|\varepsilon\| + \|\theta^* - \theta_s\| \le 4C_1\sigma\sqrt{n_1}.$$

Plugging this bound into the previous one, we obtain

$$\|\hat{\theta}^{\mathsf{svt}} - \theta^*\|_F^2 \lesssim s\sigma^2 n_1 + \sum_{i=s+1}^{n_2} (\lambda_i^*)^2 \lesssim \sum_{i=1}^{n_2} \big[\sigma^2 n_1 \wedge (\lambda_i^*)^2\big]. \tag{5.46}$$

For any integer $r \geq 6$, Proposition 6 yields a rank-r matrix $\tilde{\theta} \in \mathbb{R}^{n_1 \times n_2}$ such that

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

Since $\theta_r = \sum_{i=1}^r \lambda_i^* u_i^* (v_i^*)^\top$ is by definition the best rank-r approximation of θ^* in the Frobenius norm, we see that

$$\sum_{i=r+1}^{n_2} (\lambda_i^*)^2 = \|\theta_r - \theta^*\|_F^2 \lesssim \frac{n_1 n_2}{r^3} V(\theta^*)^2.$$

Hence it follows from (5.46) that

$$\|\hat{\theta}^{\mathsf{svt}} - \theta^*\|_F^2 \lesssim r\sigma^2 n_1 + \frac{n_1 n_2}{r^3} V(\theta^*)^2.$$

Choosing the optimal r^* and considering the boundary cases $r^* < 1$ and $r^* > n_2$ then yields

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

By repeating the same proof keeping track of the failure probability of the statement, we can obtain bounds in expectation as well.

5.14 Proof of Proposition 6

By rescaling, we may assume that $V(\theta) = 1$ without loss of generality. Lemma 7 yields that $\theta = R + S + B$, where R and S are rank-one matrices, and B is anti-Monge, bivariate isotonic (i.e., B has nondecreasing rows and columns). Additionally, we have $B_{i,1} = B_{1,j} = 0$ for $i \in [n_1], j \in [n_2]$, and $B_{n_1,n_2} = 1$. It suffices to find a low-rank approximation of the matrix B.

5.14.1 Subdivision. We claim that there exist two increasing sequences of indices $\{i_k\}_{k=1}^{r+1}$ and $\{j_\ell\}_{\ell=1}^{2r}$ such that

- $0 \leq i_k - i_{k-1} \leq n_1/r$ for $k \in [r+1]$;
- $0 \leq j_\ell - j_{\ell-1} \leq n_2/r$ and $B_{n_1,j_\ell} - B_{n_1,j_{\ell-1}+1} \leq 1/r$ for $\ell \in [2r]$.

For $\{i_k\}_{k=1}^{r+1}$, it suffices to choose $i_0 = 0$, $i_k = i_{k-1} + \lfloor n_1/r \rfloor$ for $k \in [r]$ and $i_{r+1} = n_1$. For $\{j_\ell\}_{\ell=1}^{2r}$, since $B_{n_1,1} = 0$, $B_{n_1,n_2} = 1$ and B has nondecreasing rows, there is an increasing sequence of indices $\{j'_\ell\}_{\ell=1}^{r}$ such that $B_{n_1,j'_\ell} - B_{n_1,j'_{\ell-1}+1} \leq 1/r$ for all $j \in [r]$. Moreover, by inserting (at most) another r indices between the indices j'_ℓ to obtain a new sequence $\{j_\ell\}_{\ell=1}^{2r}$, we can guarantee that not only $B_{n_1,j_\ell} - B_{n_1,j_{\ell-1}+1} \leq 1/r$, but also $j_\ell - j_{\ell-1} \leq n_2/r$.

5.14.2 Low-rank approximation. Let $\{i_k\}_{k=1}^{r+1}$ and $\{j_\ell\}_{\ell=1}^{2r}$ be chosen so that the above conditions are satisfied. We define a matrix $X \in \mathbb{R}^{n_1 \times n_2}$ by setting $X_{i,j} = B_{i,j_{\ell-1}+1}$ for all $i \in [n_1]$ and $j_{\ell-1} < j \leq j_\ell$ where $\ell \in [2r]$. By definition, all columns of X with indices in $(j_{\ell-1}, j_\ell]$ are the same, so X has rank at most $2r$.

Furthermore, we define a matrix $Y \in \mathbb{R}^{n_1 \times n_2}$ by setting $Y_{i,j} = (B - X)_{i_{k-1}+1,j}$ for all $j \in [n_2]$ and $i_{k-1} < i \le i_k$ where $k \in [r+1]$. Similarly, all rows of Y with indices in $(i_{k-1}, i_k]$ are the same, so Y has rank at most $r + 1$.

It remains to bound $\|\Delta\|_F^2$ where $\Delta = X + Y - B$. First, let us focus on a block with double indices in $(i_{k-1}, i_k] \times (j_{\ell-1}, j_\ell]$. *On each of these blocks*, by definition it holds that:

- X has constant rows, and Y has constant columns;
- the first column of Y is zero, and the first column and first row of Δ is zero.

Then by Lemma 7 (applied with the corresponding blocks of (B, X, Y, Δ) in place of (θ, R, S, B)), we see that Δ is bivariate isotonic on each of these block, and

$$\Delta_{i_k,j_\ell} = V\big(B_{i_{k-1}+1:i_k, j_{\ell-1}+1:j_\ell}\big) = B_{i_{k-1}+1, j_{\ell-1}+1} + B_{i_k,j_\ell} - B_{i_{k-1}+1,j_\ell} - B_{i_k, j_{\ell-1}+1}.$$

Therefore, it follows that

$$\sum_{i=i_{k-1}+1}^{i_k} \sum_{j=j_{\ell-1}+1}^{j_\ell} \Delta_{i,j}^2 \le (i_k - i_{k-1})(j_\ell - j_{\ell-1}) \Delta_{i_k,j_\ell}^2$$
$$\le \frac{n_1 n_2}{r^2} \Delta_{i_k,j_\ell}^2 = \frac{n_1 n_2}{r^2} V\big(B_{i_{k-1}+1:i_k, j_{\ell-1}+1:j_\ell}\big)^2,$$

where the second inequality holds thanks to the above choice of indices.

Summing over all the blocks, we obtain

$$\|\Delta\|_F^2 = \sum_{k=1}^{r+1} \sum_{\ell=1}^{2r} \sum_{i=i_{k-1}+1}^{i_k} \sum_{j=j_{\ell-1}+1}^{j_\ell} \Delta_{i,j}^2 \le \frac{n_1 n_2}{r^2} \sum_{k=1}^{r+1} \sum_{\ell=1}^{2r} V\big(B_{i_{k-1}+1:i_k, j_{\ell-1}+1:j_\ell}\big)^2.$$

5.14.3 Bounding the sum of variations. It remains to bound the above sum. Recall that a telescoping sum gives

$$\sum_{k=1}^{r+1} V\big(B_{i_{k-1}+1:i_k, j_{\ell-1}+1:j_\ell}\big) = V\big(B_{1:n_1, j_{\ell-1}+1:j_\ell}\big) \overset{(i)}{=} B_{n_1,j_\ell} - B_{n_1,j_{\ell-1}+1} \overset{(ii)}{\le} 1/r,$$

where (i) holds because the first row and first column of B are zero, and (ii) holds because of our choice of $\{j_\ell\}_{\ell=1}^{2r}$. As a result, it holds by Hölder's inequality that

$$\sum_{\ell=1}^{2r} \sum_{k=1}^{r+1} V\big(B_{i_{k-1}+1:i_k, j_{\ell-1}+1:j_\ell}\big)^2 \le \sum_{\ell=1}^{2r} V\big(B_{1:n_1, j_{\ell-1}+1:j_\ell}\big)^2 \le 2r(1/r)^2 = 2/r.$$

We therefore obtain $\|\Delta\|_F^2 \le \frac{2n_1n_2}{r^3}$. The proof is complete since $\Delta = X + Y - B = R + S + X + Y - \theta$, where the matrix $R + S + X + Y$ has rank at most $3r + 3$.

ACKNOWLEDGMENT

We thank Adityanand Guntuboyina for discussing their 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 an NSF MSPRF DMS-1703821.

APPENDIX A: EXISTING RESULTS

In this appendix, we state some existing results that are used in our proofs.

The following result is Talagrand's majorizing measure theorem [66]. See also Theorem 8.6.1 of [68].

THEOREM 19 (Talagrand's majorizing measure theorem). *For any set $\mathcal{M} \subset \mathbb{R}^d$ and a Gaussian random vector $\varepsilon \sim \mathcal{N}(0, I_d)$, we have*

$$\mathbb{E} \sup_{\theta \in \mathcal{M}} \langle \varepsilon, \theta \rangle \asymp \gamma_2(\mathcal{M}),$$

where $\gamma_2(\cdot)$ denotes Talagrand's γ_2 functional.

The following theorem gives a tail bound on the supremum of a sub-Gaussian process [66, 68].

THEOREM 20 (Generic chaining tail bound). *Consider a sub-Gaussian vector $\varepsilon \sim \mathsf{subG}_d(\sigma^2)$. For any set $\mathcal{M} \subset \mathbb{R}^d$ and $s > 0$, it holds with probability at least $1 - 2\exp(-s^2)$ that*

$$\sup_{\theta \in \mathcal{M}} \langle \varepsilon, \theta \rangle \lesssim \sigma \big[\gamma_2(\mathcal{M}) + s \cdot \sup_{\theta \in \mathcal{M}} \|\theta\|_2 \big].$$

PROOF. By Theorem 8.5.5 of [68], we have that for any $\theta^* \in \mathcal{M}$,

$$\sup_{\theta \in \mathcal{M}} |\langle \varepsilon, \theta - \theta^* \rangle| \lesssim \sigma \big[\gamma_2(\mathcal{M}) + s \cdot \mathsf{diam}(\mathcal{M}) \big]$$

with probability at least $1 - 2\exp(-s^2)$, where $\mathsf{diam}(\cdot)$ denotes the diameter of a set. Moreover, we have with probability $1 - 2\exp(-s^2)$ that

$$|\langle \varepsilon, \theta^* \rangle| \lesssim \sigma s \|\theta^*\|_2.$$

It follows that with probability at least $1 - 4\exp(-s^2)$,

$$\sup_{\theta \in \mathcal{M}} \langle \varepsilon, \theta \rangle \le \sup_{\theta \in \mathcal{M}} |\langle \varepsilon, \theta - \theta^* \rangle| + |\langle \varepsilon, \theta^* \rangle| \lesssim \sigma \big[\gamma_2(\mathcal{M}) + s \cdot \mathsf{diam}(\mathcal{M}) + s \|\theta^*\|_2 \big].$$

Since $\mathsf{diam}(\mathcal{M}) + \|\theta^*\|_2 \lesssim \sup_{\theta \in \mathcal{M}} \|\theta\|_2$, the proof is complete. □

Assouad's lemma is used to prove the lower bounds (see [67, Lemma 24.3]).

THEOREM 21 (Assouad's Lemma). *Consider a parameter space $\mathcal{M}$. Let $\mathbb{P}_\theta$ denote the distribution of the observation given that the true parameter is $\theta \in \mathcal{M}$. Let $\mathbb{E}_\theta$ denote the corresponding expectation. Suppose that for each $\tau \in \{-1, 1\}^d$, there is an associated $\theta^\tau \in \mathcal{M}$. Then it holds that*

$$\inf_{\tilde{\theta}} \sup_{\theta^* \in \mathcal{M}} \mathbb{E}_{\theta^*} \ell^2(\tilde{\theta}, \theta^*) \ge \frac{d}{8} \min_{\tau \ne \tau'} \frac{\ell^2(\theta^\tau, \theta^{\tau'})}{d_H(\tau, \tau')} \min_{d_H(\tau, \tau') = 1} (1 - \|\mathbb{P}_{\theta^\tau} - \mathbb{P}_{\theta^{\tau'}}\|_{TV}),$$

where ℓ denotes any distance function on $\mathcal{M}$, d_H denotes the Hamming distance, $\|\cdot\|_{TV}$ denotes the total variation distance, and the infimum is taken over all estimators $\tilde{\theta}$ measurable with respect to the observation y.

The following lemma can be proven with basic concentration theory, or follows as a special instance from the Hanson-Wright inequality [37, 59].

LEMMA 22. *Suppose that $\varepsilon \in \mathbb{R}^n$ is a random vector with independent centered $\mathsf{subG}(\sigma^2)$ entries. Then it holds that for all $t \geq 0$,*

$$\mathbb{P}\Big\{\Big|\sum_{i=1}^{n}\big(\varepsilon_i^2 - \mathbb{E}[\varepsilon_i^2]\big)\Big| \geq t\Big\} \leq 2\exp\Big[-c\min\Big(\frac{t^2}{\sigma^4 n}, \frac{t}{\sigma^2}\Big)\Big].$$

APPENDIX B: LOWER BOUNDS FOR THE SVT ESTIMATOR

Let us focus on the special case $n_1 = n_2 = n$, and study a worst-case matrix in $\mathcal{M} = \mathcal{M}^{n,n}$, for which the approximation rate given by Proposition 6 is tight.

LEMMA 23. *There exists a matrix $\theta \in \mathcal{M}^{n,n}$ such that*

$$\min_{\theta_r \text{ has rank } r} \|\theta_r - \theta\|_F^2 \gtrsim \frac{n^2}{r^3} V(\theta)^2.$$

PROOF. Consider the matrix

$$\theta := \frac{V_0}{n} D^\dagger (D^\dagger)^\top$$

for $V_0 > 0$. It is anti-Monge because

$$D\theta D^\top = \frac{V_0}{n} I \geq 0.$$

It also follows from (2.3) that

$$V(\theta) = \|D\theta D^\top\|_1 = \frac{V_0}{n}\|I\|_1 = V_0.$$

Moreover, the singular values of D are given in (5.23), so the eigenvalues of θ are

$$\mu_i = \frac{V_0}{4n}\Big(\sin\frac{\pi i}{2n}\Big)^{-2}.$$

Using the fact that $x/2 \leq \sin x \leq x$ for $x \in [0, \pi/2]$, we obtain

$$\frac{V_0 n}{\pi^2 i^2} \leq \mu_i \leq \frac{4 V_0 n}{\pi^2 i^2}.$$

If the spectral decomposition of θ is $\theta = \sum_{i=1}^n \mu_i w_i w_i^\top$, then the best rank-$r$ approximation of θ in the Frobenius norm is $\theta_r = \sum_{i=1}^r \mu_i w_i w_i^\top$, and

$$\|\theta_r - \theta\|_F^2 = \sum_{i=r+1}^{n} \mu_i^2 \geq \sum_{i=r+1}^{n} \frac{V_0^2 n^2}{\pi^4 i^4} \gtrsim \frac{V_0^2 n^2}{r^3},$$

which completes the proof. □

Since the anti-Monge matrix θ in the above proof cannot be approximated by a low-rank matrix at a better rate, we conjecture that for this choice of θ, the rate of convergence given by Theorem 5 for the SVT estimator (3.9) is tight. Intuitively, if we set threshold ρ in definition (3.9) to be larger, then the resulting estimator has a lower rank, thus incurring a larger bias according to the above lemma. On the other hand, setting threshold ρ to be smaller incurs a larger variance due to the noise ε.

More precisely, under the model $y = \theta + \varepsilon$ where ε has i.i.d. $\mathcal{N}(0, \sigma^2)$ entries, we conjecture that for any choice of threshold ρ in the estimator (3.9), it holds with constant probability that

$$\frac{1}{n^2}\|\hat{\theta}^{\mathsf{svt}} - \theta\|_F^2 \gtrsim \frac{\sigma^2}{n} + \frac{\sigma^{3/2}V(\theta)^{1/2}}{n^{3/4}}.$$

This is because we believe that the bias-variance trade-off in the proof of Theorem 5 is optimal. However, we are unable to prove a lower bound based on a similar argument in [62], since there, the authors are able to exploit a varying signal-to-noise ratio within the classes of matrices they consider. This allows them to employ a triangle inequality argument instead of a explicit bias-variance decomposition that holds with equality. Potential other approaches include analyzing an explicit unbiased estimate of the risk for SVT [12], and studying the exact asymptotic optimal choice of threshold ρ as in [36]. However, since any of these approaches require asymptotic random matrix theory, we consider them beyond the scope of the current work.

REFERENCES

[1] A. Aggarwal, M. M. Klawe, S. Moran, P. Shor, and R. Wilber. Geometric applications of a matrix-searching algorithm. *Algorithmica*, 2(1-4):195–208, 1987.

[2] E. M. Airoldi, T. B. Costa, and S. H. Chan. Stochastic blockmodel approximation of a graphon: Theory and consistent estimation. In *Advances in Neural Information Processing Systems 26*, pages 692–700, 2013.

[3] J. E. Atkins, E. G. Boman, and B. Hendrickson. A spectral algorithm for seriation and the consecutive ones problem. *SIAM Journal on Computing*, 28(1):297–310, 1998.

[4] R. Blei, F. Gao, and W. Li. Metric entropy of high dimensional distributions. *Proceedings of the American Mathematical Society*, 135(12):4009–4018, 2007.

[5] C. Borgs, J. T. Chayes, H. Cohn, and S. Ganguly. Consistent nonparametric estimation for heavy-tailed sparse graphs. *arXiv preprint arXiv:1508.06675*, 2015.

[6] J. P. Boyle and R. L. Dykstra. A method for finding projections onto the intersection of convex sets in Hilbert spaces. In *Advances in Order Restricted Statistical Inference*, pages 28–47. Springer, 1986.

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

[8] R. E. Burkard. Monge properties, discrete convexity and applications. *European journal of operational research*, 176(1):1–14, 2007.

[9] R. E. Burkard, V. G. Deĭneko, R. van Dal, J. A. van der Veen, and G. J. Woeginger. Well-solvable special cases of the traveling salesman problem: a survey. *SIAM review*, 40(3):496–546, 1998.

[10] R. E. Burkard, V. G. Deĭneko, and G. J. Woeginger. The travelling salesman problem on permuted monge matrices. *Journal of combinatorial optimization*, 2(4):333–350, 1998.

[11] R. E. Burkard, B. Klinz, and R. Rudolf. Perspectives of monge properties in optimization. *Discrete Applied Mathematics*, 70(2):95–161, 1996.

[12] E. J. Candes, C. A. Sing-Long, and J. D. Trzasko. Unbiased risk estimates for singular value thresholding and spectral estimators. *IEEE transactions on signal processing*, 61(19):4643–4657, 2013.

[13] K. Cechlárová and P. Szabó. On the monge property of matrices. *Discrete Mathematics*, 81(2):123–128, 1990.

[14] E. Cela, V. G. Deĭneko, and G. J. Woeginger. New special cases of the quadratic assignment problem with diagonally structured coefficient matrices. *European journal of operational research*, 267(3):818–834, 2018.

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

[16] S. Chatterjee. A new perspective on least squares under convex constraint. *The Annals of Statistics*, 42(6):2340–2381, Dec. 2014.

[17] S. Chatterjee. Matrix estimation by universal singular value thresholding. *The Annals of Statistics*, 43(1):177–214, 2015.

[18] S. Chatterjee, A. Guntuboyina, and B. Sen. On matrix estimation under monotonicity constraints. *Bernoulli*, 24(2):1072–1100, 2018.

[19] S. Chatterjee and S. Mukherjee. Estimation in tournaments and graphs under monotonicity constraints. *IEEE Transactions on Information Theory*, 2019.

[20] O. Collier and A. S. Dalalyan. Minimax rates in permutation estimation for feature matching. *Journal of Machine Learning Research*, 17(6):1–31, 2016.

[21] P. L. Combettes and J.-C. Pesquet. Proximal splitting methods in signal processing. In *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*, pages 185–212. Springer, 2011.

[22] V. G. Deĭneko, R. Rudolf, and G. J. Woeginger. On the recognition of permuted supnick and incomplete monge matrices. *Acta Informatica*, 33(5):559–569, 1996.

[23] F. Deutsch and H. Hundal. The rate of convergence of Dykstra's cyclic projections algorithm: The polyhedral case. *Numerical Functional Analysis and Optimization*, 15(5-6):537–565, 1994.

[24] J. Ding, Z. Ma, Y. Wu, and J. Xu. Efficient random graph matching via degree profiles. *arXiv preprint arXiv:1811.07821*, 2018.

[25] A. Domahidi, E. Chu, and S. Boyd. ECOS: An SOCP solver for embedded systems. In *European Control Conference (ECC)*, pages 3071–3076, 2013.

[26] S. Fallat, S. Lauritzen, K. Sadeghi, C. Uhler, N. Wermuth, and P. Zwiernik. Total positivity in Markov structures. *The Annals of Statistics*, 45(3):1152–1184, 2017.

[27] B. Fang, A. Guntuboyina, and B. Sen. Multivariate extensions of isotonic regression and total variation denoising via entire monotonicity and Hardy-Krause variation. *arXiv preprint arXiv:1903.01395*, 2019.

[28] M. Fiedler. Equilibrated anti-monge matrices. *Linear Algebra and Its Applications*, 335(1-3):151–156, 2001.

[29] M. Fiedler. Remarks on monge matrices. *Mathematica Bohemica*, 127(1):27–32, 2002.

[30] N. Flammarion, C. Mao, and P. Rigollet. Optimal rates of statistical seriation. *Bernoulli*, 25(1):623–653, 2019.

[31] F. Fogel, A. d'Aspremont, and M. Vojnovic. Serialrank: Spectral ranking using seriation. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 900–908. Curran Associates, Inc., 2014.

[32] F. Fogel, R. Jenatton, F. Bach, and A. d'Aspremont. Convex relaxations for permutation problems. In C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Weinberger, editors,

Advances in Neural Information Processing Systems 26, pages 1016–1024. Curran Associates, Inc., 2013.

[33] C. M. Fortuin, P. W. Kasteleyn, and J. Ginibre. Correlation inequalities on some partially ordered sets. *Communications in Mathematical Physics*, 22(2):89–103, 1971.

[34] C. Gao, Y. Lu, and H. H. Zhou. Rate-optimal graphon estimation. *The Annals of Statistics*, 43(6):2624–2652, 2015.

[35] F. Gao. Bracketing entropy of high dimensional distributions. In *High Dimensional Probability VI*, pages 3–17. Springer, 2013.

[36] M. Gavish and D. L. Donoho. The optimal hard threshold for singular values is $4/\sqrt{3}$. *IEEE Transactions on Information Theory*, 60(8):5040–5053, 2014.

[37] D. L. Hanson and F. T. Wright. A bound on tail probabilities for quadratic forms in independent random variables. *The Annals of Mathematical Statistics*, 42(3):1079–1083, 1971.

[38] A. J. Hoffman. On simple linear programming problems. In *Proceedings of Symposia in Pure Mathematics*, volume 7, pages 317–327, 1963.

[39] J.-C. Hütter, C. Mao, P. Rigollet, and E. Robeva. *In preparation*, 2019.

[40] J.-C. Hütter and P. Rigollet. Optimal rates for total variation denoising. In *Conference on Learning Theory*, pages 1115–1146, 2016.

[41] S. Karlin and Y. Rinott. Classes of orderings of measures and related correlation inequalities. i. multivariate totally positive distributions. *Journal of Multivariate Analysis*, 10(4):467–498, 1980.

[42] S. Karlin and Y. Rinott. Classes of orderings of measures and related correlation inequalities ii. multivariate reverse rule distributions. *Journal of Multivariate Analysis*, 10(4):499–516, 1980.

[43] D. G. Kendall. Incidence matrices, interval graphs and seriation in archeology. *Pacific J. Math.*, 28:565–570, 1969.

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

[45] B. Klinz, R. Rudolf, and G. J. Woeginger. On the recognition of permuted bottleneck monge matrices. *Discrete applied mathematics*, 63(1):43–74, 1995.

[46] O. Klopp, A. B. Tsybakov, and N. Verzelen. Oracle inequalities for network models and sparse graphon estimation. *The Annals of Statistics*, 45(1):316–354, 2017.

[47] S. Lauritzen, C. Uhler, and P. Zwiernik. Maximum likelihood estimation in Gaussian models under total positivity. *arXiv preprint arXiv:1702.04031*, 2017.

[48] L. Livi and A. Rizzi. The graph matching problem. *Pattern Analysis and Applications*, 16(3):253–283, 2013.

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

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

[51] G. Monge. Mémoire sur la théorie des déblais et des remblais. *Histoire de l'Académie Royale des Sciences de Paris*, 1781.

[52] B. O'Donoghue, E. Chu, N. Parikh, and S. Boyd. Conic optimization via operator splitting and homogeneous self-dual embedding. *Journal of Optimization Theory and Applications*, 169(3):1042–1068, June 2016.

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

[54] A. Pananjady, C. Mao, V. Muthukumar, M. J. Wainwright, and T. A. Courtade. Worst-

case vs average-case design for estimation from fixed pairwise comparisons. *arXiv preprint arXiv:1707.06217*, 2017.
[55] J. K. Park. A special case of the n-vertex traveling-salesman problem that can be solved in $O(n)$ time. *Inform. Process. Lett.*, 40(5):247–254, 1991.
[56] U. Pferschy, R. Rudolf, and G. J. Woeginger. Monge matrices make maximization manageable. *Operations Research Letters*, 16(5):245–254, 1994.
[57] M. Queyranne, F. Spieksma, and F. Tardella. A general class of greedily solvable linear programs. *Mathematics of Operations Research*, 23(4):892–908, 1998.
[58] E. Robeva, B. Sturmfels, N. Tran, and C. Uhler. Maximum likelihood estimation for totally positive log-concave densities. *arXiv preprint arXiv:1806.10120*, 2018.
[59] M. Rudelson and R. Vershynin. Hanson-wright inequality and sub-gaussian concentration. *Electronic Communications in Probability*, 18, 2013.
[60] R. Rudolf and G. J. Woeginger. The cone of monge matrices: extremal rays and applications. *Zeitschrift für Operations Research*, 42(2):161–168, 1995.
[61] D. Shah and C. Lee. Reducing crowdsourcing to graphon estimation, statistically. In A. Storkey and F. Perez-Cruz, editors, *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pages 1741–1750, Playa Blanca, Lanzarote, Canary Islands, 09–11 Apr 2018. PMLR.
[62] N. B. Shah, S. Balakrishnan, A. Guntuboyina, and M. J. Wainwright. Stochastically transitive models for pairwise comparisons: statistical and computational issues. *IEEE Trans. Inform. Theory*, 63(2):934–959, 2017.
[63] N. B. Shah, S. Balakrishnan, and M. J. Wainwright. Feeling the bern: Adaptive estimators for bernoulli probabilities of pairwise comparisons. In *2016 IEEE International Symposium on Information Theory (ISIT)*, pages 1153–1157. IEEE, 2016.
[64] N. B. Shah, S. Balakrishnan, and M. J. Wainwright. A permutation-based model for crowd labeling: Optimal estimation and robustness. *arXiv preprint arXiv:1606.09632*, 2016.
[65] G. Strang. *Computational science and engineering*, volume 791. Wellesley-Cambridge Press Wellesley, 2007.
[66] M. Talagrand. *Upper and lower bounds for stochastic processes*, volume 60. Springer-Verlag Berlin Heidelberg, 2014.
[67] A. W. v. d. Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998.
[68] R. Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge University Press, 2018.
[69] Y.-X. Wang, J. Sharpnack, A. Smola, and R. Tibshirani. Trend filtering on graphs. In *Artificial Intelligence and Statistics*, pages 1042–1050, 2015.
[70] P. J. Wolfe and S. C. Olhede. Nonparametric graphon estimation. *arXiv preprint arXiv:1309.5936*, 2013.