Bernoulli **24**(2), 2018, 1072–1100
DOI: 10.3150/16-BEJ865

On matrix estimation under monotonicity constraints

SABYASACHI CHATTERJEE[1], ADITYANAND GUNTUBOYINA[2] and BODHISATTVA SEN[3]

[1] *5747 S. Ellis Avenue, Chicago, IL 60637, USA. E-mail: sabyasachi@galton.uchicago.edu*
[2] *423 Evans Hall, Berkeley, CA 94720, USA. E-mail: aditya@stat.berkeley.edu*
[3] *1255 Amsterdam Avenue, New York, NY 10027, USA. E-mail: bodhi@stat.columbia.edu*

We consider the problem of estimating an unknown $n_1 \times n_2$ matrix $\boldsymbol{\theta}^*$ from noisy observations under the constraint that $\boldsymbol{\theta}^*$ is nondecreasing in both rows and columns. We consider the least squares estimator (LSE) in this setting and study its risk properties. We show that the worst case risk of the LSE is $n^{-1/2}$, up to multiplicative logarithmic factors, where $n = n_1 n_2$ and that the LSE is minimax rate optimal (up to logarithmic factors). We further prove that for some special $\boldsymbol{\theta}^*$, the risk of the LSE could be much smaller than $n^{-1/2}$; in fact, it could even be parametric, that is, n^{-1} up to logarithmic factors. Such parametric rates occur when the number of "rectangular" blocks of $\boldsymbol{\theta}^*$ is bounded from above by a constant. We also derive an interesting adaptation property of the LSE which we term variable adaptation – the LSE adapts to the "intrinsic dimension" of the problem and performs as well as the oracle estimator when estimating a matrix that is constant along each row/column. Our proofs, which borrow ideas from empirical process theory, approximation theory and convex geometry, are of independent interest.

Keywords: adaptation; bivariate isotonic regression; metric entropy bounds; minimax lower bound; oracle inequalities; tangent cone; variable adaptation

1. Introduction

This paper studies the problem of estimating an unknown $n_1 \times n_2$ matrix $\boldsymbol{\theta}^*$ under the constraint that $\boldsymbol{\theta}^*$ is nondecreasing in both rows and columns. In order to put this problem and our results in proper context, consider first the problem of estimating an unknown nondecreasing sequence under Gaussian measurements. Specifically, consider the problem of estimating $\theta^* = (\theta_1^*, \dots, \theta_n^*) \in \mathbb{R}^n$ from observations

$$y_i = \theta_i^* + \varepsilon_i \qquad \text{for } i = 1, \dots, n$$

under the constraint that the unknown sequence θ^* satisfies $\theta_1^* \le \cdots \le \theta_n^*$. Here the unobserved errors $\varepsilon_1, \dots, \varepsilon_n$ are i.i.d. $N(0, \sigma^2)$ with $\sigma > 0$ unknown. We refer to this as the vector isotonic estimation problem. This is a special case of univariate isotonic regression and has a long history; see, for example, Brunk [5], Ayer *et al.* [1], and van Eeden [21]. The most commonly used estimator here is the least squares estimator (LSE) defined as

$$\hat{\theta} := \operatorname*{argmin}_{\theta \in \mathcal{C}_n} \sum_{i=1}^{n} (y_i - \theta_i)^2 \qquad \text{where } \mathcal{C}_n := \{\theta \in \mathbb{R}^n : \theta_1 \le \cdots \le \theta_n\}. \tag{1.1}$$

The behavior of $\hat{\theta}$ as an estimator of θ^* is most naturally studied in terms of the risk:

$$R_{\text{vec}}(\theta^*, \hat{\theta}) := \frac{1}{n}\mathbb{E}_{\theta^*}\|\hat{\theta} - \theta^*\|^2,$$

where $\|\cdot\|$ denotes the usual Euclidean norm. The subscript vec is used to indicate that this denotes the risk in the vector estimation problem. This risk $R_{\text{vec}}(\theta^*, \hat{\theta})$ has been studied by a number of authors including van de Geer [22,23], Donoho [11], Birgé and Massart [4], Wang [28], Meyer and Woodroofe [16], Zhang [29] and Chatterjee *et al.* [8]. Zhang [29], among other things, showed the existence of a universal positive constant C such that

$$R_{\text{vec}}(\theta^*, \hat{\theta}) \le C\left\{\left(\frac{\sigma^2\sqrt{D(\theta^*)}}{n}\right)^{2/3} + \frac{\sigma^2 \log n}{n}\right\}. \tag{1.2}$$

with $D(\theta^*) := (\theta_n^* - \theta_1^*)^2$. This result shows that the risk of $\hat{\theta}$ is no more than $n^{-2/3}$ (ignoring constant factors) provided $D(\theta^*)$ is bounded from above by a constant. It can be proved that $n^{-2/3}$ is the minimax rate of estimation in this problem (see, e.g., Zhang [29]). Throughout the paper, C will denote a universal positive constant even though its exact value might change from place to place.

A complementary upper bound on $R_{\text{vec}}(\theta^*, \hat{\theta})$ has been proved recently by Bellec [3] who showed that

$$R_{\text{vec}}(\theta^*, \hat{\theta}) \le \inf_{\theta \in \mathcal{C}_n}\left(\frac{\|\theta^* - \theta\|^2}{n} + \frac{\sigma^2 k(\theta)}{n}(\log en)\right), \tag{1.3}$$

where $k(\theta)$ is the cardinality of the set $\{\theta_1, \dots, \theta_n\}$. This result is an improvement of a previous result by Chatterjee *et al.* [8] where inequality (1.3) was proved with an additional constant multiplicative factor.

The two bounds (1.2) and (1.3) provide a holistic understanding of the global accuracy of the LSE $\hat{\theta}$ in vector isotonic estimation: its risk can never be larger than the minimax rate $(\sigma^2\sqrt{D(\theta^*)}/n)^{2/3}$ while it can be the parametric rate σ^2/n, up to logarithmic multiplicative factors, if θ^* can be well approximated by $\theta \in \mathcal{C}_n$ with small $k(\theta)$. We refer to (1.2) as the worst case risk bound of the LSE and to (1.3) as the adaptive risk bound (adaptive because it states that the risk of the LSE is smaller than the worst case rate for certain special θ^*).

The goal of this paper is to extend both these worst case and adaptive risk bounds to the case of matrix isotonic estimation. Matrix isotonic estimation refers to the problem of estimating an unknown matrix $\boldsymbol{\theta}^* = (\boldsymbol{\theta}^*_{ij}) \in \mathbb{R}^{n_1 \times n_2}$ from observations

$$\boldsymbol{y}_{ij} = \boldsymbol{\theta}^*_{ij} + \boldsymbol{\varepsilon}_{ij} \qquad \text{for } i = 1, \dots, n_1,\ j = 1, \dots, n_2, \tag{1.4}$$

where $\boldsymbol{\theta}^*$ is constrained to lie in

$$\mathcal{M} := \{\boldsymbol{\theta} \in \mathbb{R}^{n_1 \times n_2} : \boldsymbol{\theta}_{ij} \le \boldsymbol{\theta}_{kl} \text{ whenever } i \le k \text{ and } j \le l\},$$

and the random errors $\boldsymbol{\varepsilon}_{ij}$'s are i.i.d. $N(0, \sigma^2)$, with $\sigma^2 > 0$ unknown. We refer to any matrix in $\mathcal{M}$ as an isotonic matrix. Throughout, we let $n := n_1 n_2$ denote the product of n_1 and n_2. As a notational convention, throughout the paper, we denote matrices in boldface and the (i, j)th entry of a matrix $\boldsymbol{A}$ will simply be denoted by $\boldsymbol{A}_{ij}$.

Monotonicity restrictions on matrices are increasingly being used as a key component of latent variable based models for the estimation of matrices and graphs. Two such examples are: (1) the estimation of graphons under monotonicity constraints (see Chatterjee and Mukherjee [10]), and (2) the nonparametric Bradley–Terry model (see Chatterjee [7], Shah *et al.* [20]). In both of these examples, the unknown matrix satisfies monotonicity constraints similar to the ones studied here. But the observation model is more complicated because of the presence of latent permutations. Nevertheless, studying the matrix isotonic estimation problem described above is a first step towards understanding the estimation properties in these more complicated models. The proof technique in this paper has been adapted and put to use in [10] and [20].

Specifically, in Chatterjee and Mukherjee [10], the parameter space considered is $\{\Pi^T \boldsymbol{\theta} \Pi : \boldsymbol{\theta} \in \mathcal{M}, \Pi \text{ is a permutation matrix}\}$ where $\mathcal{M}$ is the space of $n \times n$ isotonic matrices considered in this paper. The loss function considered is given by the usual Frobenius norm squared. This problem actually reduces to the problem studied in this paper when the latent permutation is known, albeit with Bernoulli noise. The estimator considered in [10] is a two step estimator, which first estimates the unknown permutation by sorting the row sums of the observed data matrix and then uses the projection operator on the space $\mathcal{M}$. The analysis of the two step estimator borrows heavily from the techniques used in this paper for the projection step.

The matrix isotonic estimation problem also has a direct connection to bivariate isotonic regression. Bivariate isotonic regression is the problem of estimating a regression function $f : [0, 1]^2 \to \mathbb{R}$ which is known to be coordinate-wise nondecreasing (i.e., if $s_1 \le t_1$ and $s_2 \le t_2$, where $(s_1, s_2), (t_1, t_2) \in [0, 1]^2$, then $f(s_1, s_2) \le f(t_1, t_2)$), from observations

$$\boldsymbol{y}_{ij} = f(i/n_1, j/n_2) + \boldsymbol{\varepsilon}_{ij} \qquad \text{for } i = 1, \dots, n_1, j = 1, \dots, n_2. \tag{1.5}$$

Identifying $f(i/n_1, j/n_2) \equiv \boldsymbol{\theta}^*_{ij}$ we see that (1.4) and (1.5) are equivalent problems. Equation (1.5) is possibly the simplest example of a multivariate shape constrained regression problem and arises quite often in production planning and inventory control; see, for example, the classical textbooks Barlow *et al.* [2] and Robertson *et al.* [19] on this subject.

Let us now introduce the LSE in matrix isotonic estimation. Let $\boldsymbol{y} = (\boldsymbol{y}_{ij})$ denote the matrix (of order $n_1 \times n_2$) of the observed response. The LSE, $\hat{\boldsymbol{\theta}}$, is defined as the minimizer of the squared Frobenius norm, $\|\boldsymbol{y} - \boldsymbol{\theta}\|^2$, over $\boldsymbol{\theta} \in \mathcal{M}$, that is,

$$\hat{\boldsymbol{\theta}} := \operatorname*{argmin}_{\boldsymbol{\theta} \in \mathcal{M}} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} (\boldsymbol{y}_{ij} - \boldsymbol{\theta}_{ij})^2. \tag{1.6}$$

Because $\mathcal{M}$ is a closed convex cone in $\mathbb{R}^{n_1 \times n_2}$ (which is the space of all $n_1 \times n_2$ matrices), the LSE $\hat{\boldsymbol{\theta}}$ exists uniquely. Further, it can be computed efficiently by an iterative algorithm (see, e.g., Gebhardt [13] and Robertson *et al.* [19], Chapter 1); this is in spite of the fact that it is defined as the solution of a quadratic program with $O(n^2)$ linear constraints.

It is fair to say that not much is known about the behavior of $\hat{\boldsymbol{\theta}}$ as an estimator of $\boldsymbol{\theta}^*$. The only result known in this direction is the consistency of $\hat{\boldsymbol{\theta}}$; see, for example, Hanson *et al.* [14], Makowski [15] and Robertson and Wright [18]. In this paper, we study the risk of $\hat{\boldsymbol{\theta}}$ as an estimator of $\boldsymbol{\theta}^*$, defined as

$$R(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}}) := \mathbb{E}_{\boldsymbol{\theta}^*} \ell^2(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}}) \qquad \text{where } \ell^2(\boldsymbol{\theta}^*, \boldsymbol{\theta}) := \frac{1}{n} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} (\boldsymbol{\theta}^*_{ij} - \boldsymbol{\theta}_{ij})^2.$$

Here $\mathbb{E}_{\boldsymbol{\theta}^*}$ denotes the expectation taken with respect to $\boldsymbol{y}$ having the distribution given by (1.4). Also, throughout the paper, we take $n = n_1 n_2$ and each of n_1 and n_2 to be strictly larger than one. We similarly define the risk $R(\boldsymbol{\theta}^*, \tilde{\boldsymbol{\theta}})$ for any other estimator $\tilde{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}^*$.

To the best of our knowledge, nothing is known in the literature about the risk $R(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}})$. The goal of this paper is to prove analogues of the inequalities (1.2) and (1.3) for $R(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}})$. The first result of this paper, Theorem 2.1, is the analogue of (1.2) for matrix isotonic estimation. Specifically, we prove in Theorem 2.1 that

$$R(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}}) \leq C\left(\sqrt{\frac{\sigma^2 D(\boldsymbol{\theta}^*)}{n}} (\log n)^4 + \frac{\sigma^2}{n} (\log n)^8\right) \tag{1.7}$$

for a universal positive constant C where $D(\boldsymbol{\theta}^*) := (\boldsymbol{\theta}^*_{n_1 n_2} - \boldsymbol{\theta}^*_{11})^2$.

Our second result proves that the minimax risk in this problem is bounded from below by $(\sigma^2 D(\boldsymbol{\theta}^*)/n)^{1/2}$, up to constant multiplicative factors. Specifically, we prove in Theorem 2.2 that

$$\inf_{\tilde{\boldsymbol{\theta}}} \sup_{\boldsymbol{\theta} \in \mathcal{M} : D(\boldsymbol{\theta}) \leq D} R(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}}) \geq \sqrt{\frac{\sigma^2 D}{192 n}} \tag{1.8}$$

under some conditions on n_1 and n_2 (see Theorem 2.2 for the precise statement). The above infimum is taken over all estimators $\tilde{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$. Combined with (1.7), this proves that $\hat{\boldsymbol{\theta}}$ is minimax, up to logarithmic multiplicative factors. Therefore, inequality (1.7) is the correct analogue of (1.2) for matrix isotonic estimation.

Next, we describe our analogue of inequality (1.3) for matrix isotonic estimation. The situation here is more subtle compared to the vector case. The most natural analogue of (1.3) in the matrix case is an inequality of the form:

$$R(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}}) \leq \inf_{\boldsymbol{\theta} \in \mathcal{M}} \left(\frac{\|\boldsymbol{\theta}^* - \boldsymbol{\theta}\|^2}{n} + \frac{\sigma^2 c(\boldsymbol{\theta}) p(\log n)}{n}\right), \tag{1.9}$$

where $p(\cdot)$ is some polynomial and $c(\boldsymbol{\theta})$ denotes the cardinality of the set $\{\boldsymbol{\theta}_{ij} : 1 \leq i \leq n_1, 1 \leq j \leq n_2\}$ and $\|\cdot\|$ refers to the Frobenius norm. Unfortunately, it turns out that this inequality cannot be true for every $\boldsymbol{\theta}^* \in \mathcal{M}$ because it contradicts the minimax lower bound (1.8). The argument for this is provided at the beginning of Section 2.2.

The fact that inequality (1.9) is false means that the LSE $\hat{\boldsymbol{\theta}}$ does not adapt to every $\boldsymbol{\theta}^* \in \mathcal{M}$ with small $c(\boldsymbol{\theta}^*)$. It turns out that inequality (1.9) can be proved for every $\boldsymbol{\theta}^* \in \mathcal{M}$ if the quantity $c(\boldsymbol{\theta})$ is replaced by a larger quantity. This quantity will be denoted by $k(\boldsymbol{\theta})$ (because it is the right analogue of $k(\theta)$ for the matrix case) and it is defined next after introducing some notation.

A subset A of $\{1,\dots,n_1\}\times\{1,\dots,n_2\}$ is called a rectangle if $A=\{(i,j): k_1\le i\le l_1, k_2\le j\le l_2\}$ for some $1\le k_1\le l_1\le n_1$ and $1\le k_2\le l_2\le n_2$. A rectangular partition of $\{1,\dots,n_1\}\times\{1,\dots,n_2\}$ is a collection of rectangles $\pi=(A_1,\dots,A_k)$ which are disjoint and whose union is $\{1,\dots,n_1\}\times\{1,\dots,n_2\}$. The cardinality of such a partition, $|\pi|$, is the number of rectangles in the partition. The collection of all rectangular partitions of $\{1,\dots,n_1\}\times\{1,\dots,n_2\}$ will be denoted by $\mathcal{P}$. For $\boldsymbol{\theta}\in\mathcal{M}$ and $\pi=(A_1,\dots,A_k)\in\mathcal{P}$, we say that $\boldsymbol{\theta}$ is constant on π if $\{\boldsymbol{\theta}_{ij}:(i,j)\in A_l\}$ is a singleton for each $l=1,\dots,k$. We are now ready to define $k(\boldsymbol{\theta})$ for $\boldsymbol{\theta}\in\mathcal{M}$. It is defined as the "number of rectangular blocks" of $\boldsymbol{\theta}$, that is, the smallest integer k for which there exists a partition $\pi\in\mathcal{P}$ with $|\pi|=k$ such that $\boldsymbol{\theta}$ is constant on π. It is trivial to see that $k(\boldsymbol{\theta})\ge c(\boldsymbol{\theta})$ for every $\boldsymbol{\theta}\in\mathcal{M}$. As a simple illustration, for $\boldsymbol{\theta}=\mathbf{1}\{i>1, j>1\}$, we have $c(\boldsymbol{\theta})=2$ and $k(\boldsymbol{\theta})=3$.

Inequality (1.9) becomes true for all $\boldsymbol{\theta}^*\in\mathcal{M}$ if $c(\boldsymbol{\theta})$ is replaced by $k(\boldsymbol{\theta})$. This is our adaptive risk bound for matrix isotonic estimation, proved in Theorem 2.4:

$$R(\boldsymbol{\theta}^*,\hat{\boldsymbol{\theta}})\le\inf_{\boldsymbol{\theta}\in\mathcal{M}}\left(\frac{\|\boldsymbol{\theta}^*-\boldsymbol{\theta}\|^2}{n}+\frac{C\sigma^2 k(\boldsymbol{\theta})}{n}(\log n)^8\right),\tag{1.10}$$

where C is a universal positive constant. As a consequence of this inequality, we obtain that the risk of the LSE converges to zero at the parametric rate σ^2/n, up to logarithmic multiplicative factors, provided $k(\boldsymbol{\theta}^*)$ is bounded from above by a constant.

We also establish a property of the LSE that we term variable adaptation. Let $\mathcal{C}_{n_1}:=\{\theta\in\mathbb{R}^{n_1}:\theta_1\le\cdots\le\theta_{n_1}\}$. Suppose $\boldsymbol{\theta}^*=(\boldsymbol{\theta}^*_{ij})\in\mathcal{M}$ has the property that $\boldsymbol{\theta}^*_{ij}$ only depends on i, that is, there exists $\theta^*\in\mathcal{C}_{n_1}$ such that $\boldsymbol{\theta}^*_{ij}=\theta^*_i$ for every i and j. If we knew this fact about $\boldsymbol{\theta}^*$, then the most natural way of estimating it would be to perform vector isotonic estimation based on the row-averages $\bar{y}:=(\bar{y}_1,\dots,\bar{y}_{n_1})$, where $\bar{y}_i:=\sum_{j=1}^{n_2}\boldsymbol{y}_{ij}/n_2$, resulting in an estimator $\breve{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}^*$. Using the vector isotonic risk bounds (1.2) and (1.3), it is easy to see then that the risk of $\breve{\boldsymbol{\theta}}$ has the following pair of bounds:

$$R(\boldsymbol{\theta}^*,\breve{\boldsymbol{\theta}})\le C\left\{\left(\frac{\sigma^2\sqrt{D(\boldsymbol{\theta}^*)}}{n}\right)^{2/3}+\frac{\sigma^2\log n_1}{n}\right\}\tag{1.11}$$

and

$$R(\boldsymbol{\theta}^*,\breve{\boldsymbol{\theta}})\le\inf_{\theta\in\mathcal{C}_{n_1}}\left(\frac{\|\theta^*-\theta\|^2}{n_1}+\frac{\sigma^2 k(\theta)}{n}\log n_1\right).\tag{1.12}$$

Note that the construction of $\breve{\boldsymbol{\theta}}$ requires the knowledge that all rows of $\boldsymbol{\theta}^*$ are constant. As a consequence of the adaptive risk bound (1.10), we shall show in Theorem 2.5 that the matrix isotonic LSE $\hat{\boldsymbol{\theta}}$ achieves the same risk bounds as $\breve{\boldsymbol{\theta}}$, up to additional logarithmic factors. This is remarkable because $\hat{\boldsymbol{\theta}}$ uses no special knowledge on $\boldsymbol{\theta}^*$; it automatically adapts to the additional structure present in $\boldsymbol{\theta}^*$.

Note that in the connection between matrix isotonic estimation and bivariate isotonic regression, the assumption that $\boldsymbol{\theta}^*_{ij}=f(i/n_1, j/n_2)$ does not depend on j is equivalent to assuming that f does not depend on its second variable. Thus, when estimating a bivariate isotonic regression function that only depends on one variable, the LSE automatically adapts and we get risk bounds that correspond to estimating a monotone function of one variable. This is the reason why we refer to this phenomenon as variable adaptation. To the best of our knowledge, such a result on automatic variable adaptation in multivariate nonparametric regression is very rare – most nonparametric regression techniques (e.g., kernel smoothing, splines) do not exhibit such automatic adaptation properties.

The proof techniques employed in this paper are quite different from the case of vector isotonic estimation. In the vector problem (1.1), the LSE has the closed form expression (see, e.g., Robertson *et al.* [19], Chapter 1):

$$\hat{\theta}_i := \min_{v \geq i} \max_{u \leq i} \frac{1}{v-u+1} \sum_{i=u}^{v} y_i. \tag{1.13}$$

This expression, along with some martingale maximal inequalities, are crucially used for the proofs of inequalities (1.2) and (1.3); see, for example, Zhang [29] and Chatterjee *et al.* [8]. The LSE (1.6) in the matrix estimation problem also has a closed form expression similar to (1.13):

$$\hat{\boldsymbol{\theta}}_{ij} = \min_{L \in \mathcal{L}:(i,j) \in L} \max_{U \in \mathcal{U}:(i,j) \in U} \bar{y}_{L \cap U}, \tag{1.14}$$

where $\mathcal{L}$ and $\mathcal{U}$ denote the collections of all lower sets and upper sets respectively and $\bar{y}_A$ is the average of $\{\boldsymbol{y}_{ij} : (i, j) \in A\}$; see Robertson *et al.* [19], Theorem 1.4.4, page 23, for the definitions of upper and lower sets and for a proof of (1.14). This unfortunately is a much more complicated expression to directly work with compared to (1.13). It is not clear to us if simple martingale techniques can be used in conjunction with the expression (1.14) to prove risk bounds for the LSE.

We therefore abandon the direct approach based on the expression (1.14) and instead resort to general techniques for LSEs in order to prove our results. Specifically, we use the standard empirical process based approach to prove the worst case bound (1.7). This approach relies on metric entropy calculations of the space of isotonic matrices. Metric entropy results for classes of isotonic matrices can be derived from those of bivariate coordinate-wise nondecreasing functions. However existing metric entropy results for classes of bivariate nondecreasing functions (as in Gao and Wellner [12]) require the functions to be uniformly bounded. Because of this reason, these results are not directly applicable to our setting. We suitably extend these results in order to allow for the lack of a uniform bound. On the other hand, for the adaptive risk bound (1.10), we use connections between the risk of LSEs and size measures of tangent cones. Thus, our proofs borrow ideas from empirical process theory, approximation theory and convex geometry and are of independent interest.

The rest of the paper is organized as follows. Our results are described in Section 2: Section 2.1 deals with the worst case risk bounds while Section 2.2 focuses on the adaptive bounds. In Section 3, we provide the necessary background on the general theory of the LSEs, prove our main metric entropy results and present the proof of our main worst case upper bound. In Section 4, we discuss connections between risk of LSEs and appropriate size measures of tangent cones, and also present the proof of our adaptive risk bounds. Additional discussion is provided in Section 5. We have also included an Appendix which contains the proofs of certain auxiliary technical results used in the paper.

2. Main results

In this section, we give risk bounds on the performance of the isotonic LSE $\hat{\boldsymbol{\theta}}$, defined in (1.6). We start with a generalization of (1.2) and then proceed to exhibit the adaptive risk behavior of $\hat{\boldsymbol{\theta}}$. We end this section with a result on the variable adaptation property of the LSE which shows that $\hat{\boldsymbol{\theta}}$ automatically adapts to the intrinsic dimension of the problem.

We would like to remark here that although our results give bounds on the risk $R(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}})$, they can easily be converted into high probability upper bounds on the loss $\ell^2(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}})$. The reason for this is the following: The mapping $\boldsymbol{\varepsilon} \mapsto \ell(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}})$ is $n^{-1/2}$-Lipschitz (under the Frobenius metric on the space of $n_1 \times n_2$ matrices) where $n = n_1 n_2$ and thus, by the usual concentration inequality for Lipschitz functions of Gaussian vectors, we have

$$\ell(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}}) \le \mathbb{E}\ell(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}}) + \sigma\sqrt{\frac{2t}{n}}$$

with probability at least $1 - \exp(-t)$. Note that this is a special of a general concentration result recently proved in van de Geer and Wainwright [24]. The elementary inequality $(x + y)^2 \le x^2(1 + a) + y^2(1 + 1/a)$ for every $a > 0$, then gives

$$\ell^2(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}}) \le (1 + a)R(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}}) + \frac{2(1 + 1/a)t\sigma^2}{n}$$

with probability at least $1 - \exp(-t)$. This inequality allows one to deduce high probability bounds on the loss $\ell^2(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}})$ from the risk bounds stated in this section.

2.1. Worst case risk bounds

Our first main result establishes inequality (1.7) which gives an upper bound on the worst case risk of the matrix isotonic LSE $\hat{\boldsymbol{\theta}}$. We will actually prove a slightly stronger bound than that given by inequality (1.7). We first need some notation. We define the *variance* of a matrix $\boldsymbol{\theta}$ as

$$V(\boldsymbol{\theta}) := \frac{1}{n}\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}(\boldsymbol{\theta}_{ij} - \overline{\theta})^2, \tag{2.1}$$

where $\overline{\theta} = \sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\boldsymbol{\theta}_{ij}/n$ is the mean of the entries of $\boldsymbol{\theta}$. Note that $V(\boldsymbol{\theta}) \le D(\boldsymbol{\theta})$ for every $\boldsymbol{\theta} \in \mathcal{M}$. We also denote the set $\{1, \dots, l\}$ by $[l]$ for positive integers l.

The following theorem, proved in Section 3.3, gives an upper bound on the risk $R(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}})$ in terms of the quantity $V(\boldsymbol{\theta}^*)$. Because $V(\boldsymbol{\theta}^*) \le D(\boldsymbol{\theta}^*)$, the conclusion of the theorem is stronger than inequality (1.7).

Theorem 2.1. *There exists a universal positive constant C such that for every $n_1, n_2 > 1$ with $n = n_1 n_2$ and $\boldsymbol{\theta}^* \in \mathcal{M}$,*

$$R(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}}) \le C\left(\frac{\sigma^2}{n}(\log n)^8 + \sqrt{\frac{\sigma^2 V(\boldsymbol{\theta}^*)}{n}}(\log n)^4\right).$$

Ignoring constants and logarithmic factors, Theorem 2.1 states that the risk of the LSE at $\boldsymbol{\theta}^*$ converges to zero at the rate $n^{-1/2}$ as long as $V(\boldsymbol{\theta}^*)$ is bounded away from zero. In the next result, proved in Appendix A.4, we argue that $n^{-1/2}$ is also a minimax lower bound in this problem. This implies that the rate $n^{-1/2}$ cannot be improved by any other estimator uniformly over the class $\{\boldsymbol{\theta}^* : V(\boldsymbol{\theta}^*) \le V\}$ for every constant V. The proof of the next result is done via Assouad's lemma. Note that it is also possible to prove a lower bound in probability as opposed to the expectation lower bound by using Fano's inequality. But we shall only provide the expectation lower bound for simplicity.

Theorem 2.2. *For every positive real number D,*

$$\inf_{\tilde{\boldsymbol{\theta}}} \sup_{\boldsymbol{\theta} \in \mathcal{M}: D(\boldsymbol{\theta}) \le D} R(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}}) \ge \sqrt{\frac{\sigma^2 D}{192 n}}, \tag{2.2}$$

where the infimum is over all estimators $\tilde{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$, provided the integers $n_1 \ge 1, n_2 \ge 1$ with $n = n_1 n_2$ satisfy $n \ge 9\sigma^2/D$ and

$$\min\left(\frac{n_1^3}{n_2}, \frac{n_2^3}{n_1}\right) \ge \frac{D}{9\sigma^2}. \tag{2.3}$$

Remark 2.1. The condition (2.3) is necessary to ensure that neither n_1 or n_2 are too small. Indeed, the inequality (2.2) is not true when, for example, $n_1 = 1, n_2 = n$ because in this case the problem reduces to vector isotonic estimation where the minimax risk is of the order $n^{-2/3} < n^{-1/2}$. When $n_1 = n_2 = \sqrt{n}$, the inequality (2.3) is equivalent to $n \ge D/(9\sigma^2)$ which is satisfied for all large n.

Remark 2.2. Recall the quantity $V(\boldsymbol{\theta})$ defined in (2.1). Because $V(\boldsymbol{\theta}) \le D(\boldsymbol{\theta})$, it follows that $\{\boldsymbol{\theta} : D(\boldsymbol{\theta}) \le D\} \subseteq \{\boldsymbol{\theta} : V(\boldsymbol{\theta}) \le D\}$. Therefore the bound (2.2) also holds if $\{\boldsymbol{\theta} : D(\boldsymbol{\theta}) \le D\}$ is replaced by the larger set $\{\boldsymbol{\theta} : V(\boldsymbol{\theta}) \le D\}$.

Note that without loss of generality we can assume $n_1 \le n_2$. In this case, (2.3) can be rewritten as $n_1 \ge (\frac{nD}{9\sigma^2})^{1/4}$. This raises the natural question about the minimax rate of our problem when $n_1 = o(n^{1/4})$. The following theorem (proved in Section A.5) answers this question.

Theorem 2.3. *For every positive real number D and for all sufficiently large n we have,*

$$\inf_{\tilde{\boldsymbol{\theta}}} \sup_{\boldsymbol{\theta} \in \mathcal{M}: D(\boldsymbol{\theta}) \le D} R(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}}) \ge \frac{(\sigma^2 \sqrt{D})^{2/3}}{64 n_2^{2/3}},$$

where the infimum is over all estimators $\tilde{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$, provided we have

$$n_1 \le \frac{(nD)^{1/4}}{(4\sigma)^{1/2}}.$$

Moreover in this case, for any $\boldsymbol{\theta}^*$ such that $D(\boldsymbol{\theta}^*) \le D$, there exists an estimator $\hat{\boldsymbol{\theta}}$ such that

$$R(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}^*) \le C \frac{(\sigma^2 \sqrt{D(\theta^*)})^{2/3}}{(n_2)^{2/3}} \tag{2.4}$$

for some universal constant $C > 0$.

Remark 2.3. Theorem 2.3 along with Theorems 2.1 and 2.2 characterize the minimax rates in our matrix estimation problem for different growth behaviours of $n_1 \le n_2$ with $n_1 n_2 = n$. The minimax rate of the problem scales like $O(1/\sqrt{n})$ in case $n_1 \ge O(n^{1/4})$. In case, $n_1 = o(n^{1/4})$ the minimax rate scales like $O(1/(n_2)^{2/3})$. Note that when $n_1 = 1$ we have $n_2 = n$ and we get back the standard univariate isotonic regression rate of $1/n^{2/3}$. In this way, the minimax rate increases smoothly from $1/n^{2/3}$ to $O(1/\sqrt{n})$ as n_1 increases from 1 to $n^{1/4}$.

In addition to proving that the LSE is minimax optimal up to logarithmic factors, another interesting aspect of Theorem 2.1 is that when $V(\boldsymbol{\theta}^*) = 0$, the upper bound on $R(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}})$ becomes the parametric rate σ^2/n up to a logarithmic factor. This rate is faster than the worst case rate $n^{-1/2}$. Thus, the LSE adapts to $\boldsymbol{\theta}^* \in \{\boldsymbol{\theta} : V(\boldsymbol{\theta}) = 0\}$. A more detailed description of the adaptation properties of the LSE is provided in the next theorem.

2.2. Adaptive risk bounds

The adaptation properties of the matrix isotonic LSE are more subtle compared to the vector case. In the latter case, adaptation of the LSE is described by inequality (1.3). The most natural analogue of (1.3) in the matrix case is an inequality of the form (1.9). Unfortunately, it turns out that this inequality cannot be true for every $\boldsymbol{\theta}^* \in \mathcal{M}$ because it contradicts the minimax lower bound proved in Theorem 2.2. The reason for this is the following. Fix $\boldsymbol{\theta}^* = (\boldsymbol{\theta}^*_{ij}) \in \mathcal{M}$ with $D := D(\boldsymbol{\theta}^*) = (\boldsymbol{\theta}^*_{n_1 n_2} - \boldsymbol{\theta}^*_{11})^2 > 0$. Now fix $c \ge 1$ and define $\boldsymbol{\theta} = (\boldsymbol{\theta}_{ij})$ by

$$\boldsymbol{\theta}_{ij} := \theta^*_{11} + \frac{\sqrt{D}}{c} \left\lfloor \frac{c(\boldsymbol{\theta}^*_{ij} - \boldsymbol{\theta}^*_{11})}{\sqrt{D}} \right\rfloor.$$

It is easy to see that $\boldsymbol{\theta} \in \mathcal{M}$ (because $\boldsymbol{\theta}_{ij}$ is a nondecreasing function of $\boldsymbol{\theta}^*_{ij}$). Also for every i, j, we have $\boldsymbol{\theta}^*_{ij} - \sqrt{D}/c \le \boldsymbol{\theta}_{ij} \le \boldsymbol{\theta}^*_{ij}$ which implies that $\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|^2 \le nD/c^2$. Finally, $c(\boldsymbol{\theta}) \le (c+1)$. Therefore if inequality (1.9) were true for every $\boldsymbol{\theta}^*$, we would obtain

$$R(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}}) \le p(\log n) \inf_{c \ge 1} \left(\frac{D}{c^2} + \frac{\sigma^2(c+1)}{n} \right).$$

Choosing $c = \lfloor (nD/\sigma^2)^{1/3} \rfloor$, we would obtain that $R(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}})$ converges to zero at the $n^{-2/3}$ rate. This obviously contradicts the minimax lower bound proved in Theorem 2.2. Therefore, one cannot hope to prove an inequality of the form (1.9) for every $\boldsymbol{\theta}^* \in \mathcal{M}$.

The fact that inequality (1.9) is false means that the LSE $\hat{\boldsymbol{\theta}}$ does not adapt to every $\boldsymbol{\theta}^* \in \mathcal{M}$ with small $c(\boldsymbol{\theta}^*)$. However, inequality (1.9) can be proved for every $\boldsymbol{\theta}^* \in \mathcal{M}$ if the quantity $c(\boldsymbol{\theta})$ is replaced by the larger quantity $k(\boldsymbol{\theta})$ – the number of rectangular blocks – as defined in the Introduction. We are now ready to state our main adaptive risk bound for the matrix LSE; see Section 4.2 for its proof.

Theorem 2.4. *There exists a universal constant $C > 0$ such that for every $n_1 \times n_2$ matrix $\boldsymbol{\theta}^*$ we have*

$$R(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}}) \leq \inf_{\boldsymbol{\theta} \in \mathcal{M}} \left\{ \frac{\|\boldsymbol{\theta}^* - \boldsymbol{\theta}\|^2}{n} + \frac{Ck(\boldsymbol{\theta})\sigma^2}{n} (\log n)^8 \right\}. \tag{2.5}$$

Remark 2.4. Note that $1 \leq k(\boldsymbol{\theta}) \leq n$ for all $\boldsymbol{\theta} \in \mathcal{M}$. There exist $\boldsymbol{\theta} \in \mathcal{M}$ for which $c(\boldsymbol{\theta}) = k(\boldsymbol{\theta})$. These are elements $\boldsymbol{\theta} \in \mathcal{M}$ whose level sets (level sets of $\boldsymbol{\theta}$ are non-empty sets of the form $\{(i, j) : \boldsymbol{\theta}_{ij} = a\}$ for some real number a) are all rectangular.

Remark 2.5. A simple consequence of Theorem 2.4 is that $R(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}})$ is bounded by the parametric rate (up to logarithmic factors) when $k(\boldsymbol{\theta}^*)$ is bounded from above by a constant. To see this, simply note that we can take $\boldsymbol{\theta} = \boldsymbol{\theta}^*$ in (2.5) to obtain

$$R(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}}) \leq C(\log n)^8 \frac{k(\boldsymbol{\theta}^*)\sigma^2}{n}.$$

The right-hand side above is just the parametric rate σ^2/n up to logarithmic factors provided $k(\boldsymbol{\theta}^*)$ is bounded by a constant (or a logarithmic factor of n).

Remark 2.6 (Model misspecification). The proof of Theorem 2.4 reveals that inequality (2.5) is actually true for every $n_1 \times n_2$ matrix $\boldsymbol{\theta}^*$ (it is not necessary that $\boldsymbol{\theta}^* \in \mathcal{M}$). Therefore, inequality (2.5) can also be treated as an oracle inequality for misspecification. Please see Bellec [3] for more background and general theory on such model misspecification oracle inequalities.

Remark 2.7. Inequality (2.5) sometimes gives near parametric bounds for $R(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}})$ even when $k(\boldsymbol{\theta}^*) = n$. This happens when $\boldsymbol{\theta}^*$ is well approximated by some $\boldsymbol{\theta} \in \mathcal{M}$ with small $k(\boldsymbol{\theta})$. An example of this is given below: Assume, for simplicity, that $n_1 = n_2 = \sqrt{n} = 2^k$ for some positive integer k. Define $\boldsymbol{\theta}^* \in \mathbb{R}^{n_1 \times n_2}$ by

$$\boldsymbol{\theta}^*_{ij} = -(2^{-i} + 2^{-j}) \qquad \text{for } 1 \leq i, j \leq n_1.$$

It should then be clear that $\boldsymbol{\theta}^* \in \mathcal{M}$ and $k(\boldsymbol{\theta}^*) = n$. Also, let us define $\boldsymbol{\theta} \in \mathcal{M}$ by

$$\boldsymbol{\theta}_{ij} = -(2^{-(i \wedge k)} + 2^{-(j \wedge k)}) \qquad \text{for } 1 \leq i, j \leq n_1,$$

where $a \wedge b := \min(a, b)$. Observe that $k(\boldsymbol{\theta}) \leq (k+1)^2 \leq C \log n$. Further

$$\frac{1}{n}\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|^2 \leq \max_{(i,j)}\big(\boldsymbol{\theta}_{ij} - \boldsymbol{\theta}^*_{ij}\big)^2 \leq 2\big(2^{-2k} + 2^{-2k}\big) = \frac{4}{n}.$$

Theorem 2.4 therefore gives

$$R\big(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}}\big) \leq C\left\{\frac{1}{n} + \frac{\sigma^2}{n}(\log n)^9\right\}.$$

This is the parametric bound up to logarithmic factors in n.

2.3. Variable adaptation

In this sub-section, we describe a very interesting property of the LSE which shows that $\hat{\boldsymbol{\theta}}$ adapts to the intrinsic dimension of the problem. Suppose that $\boldsymbol{\theta}^* \in \mathcal{M}$ is such that its value does not depend on the columns, that is, there exists $\theta^* \in \mathcal{C}_{n_1}$ (recall that $\mathcal{C}_{n_1} = \{\theta^* \in \mathbb{R}^{n_1} : \theta^*_1 \leq \cdots \leq \theta^*_{n_1}\}$) such that $\boldsymbol{\theta}^*_{ij} = \theta^*_i$ for every i and j. Note that in connection to bivariate isotonic regression, the assumption that $\boldsymbol{\theta}^*_{ij} := f(i/n_1, j/n_2)$ does not depend on j is equivalent to assuming that f does not depend on its second variable. If we knew this fact about $\boldsymbol{\theta}^*$, then the most natural way of estimating it would be to perform vector isotonic estimation based on the row-averages $\bar{y} := (\bar{y}_1, \ldots, \bar{y}_{n_1})$, where $\bar{y}_i := \sum_{j=1}^{n_2} \boldsymbol{y}_{ij}/n_2$, resulting in an estimator $\breve{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}^*$. This oracle estimator has risk bounds given in (1.11) and (1.12).

The following theorem, proved in Section 4.3, shows that the matrix isotonic LSE $\hat{\boldsymbol{\theta}}$ achieves the same risk bounds as $\breve{\boldsymbol{\theta}}$, up to additional multiplicative logarithmic factors. This is remarkable because $\hat{\boldsymbol{\theta}}$ uses no special knowledge on $\boldsymbol{\theta}^*$; it automatically adapts to the additional structure present in $\boldsymbol{\theta}^*$. Thus, when estimating a bivariate isotonic regression function that only depends on one variable, the LSE automatically adapts and we get risk bounds that correspond to estimating a monotone function in one variable. As mentioned in the Introduction, such a result on automatic variable adaptation in multivariate nonparametric regression is very rare.

Theorem 2.5. *Suppose $\boldsymbol{\theta}^* = (\boldsymbol{\theta}^*_{ij}) \in \mathcal{M}$ and $\theta^* = (\theta^*_i) \in \mathcal{C}_{n_1}$ are such that $\boldsymbol{\theta}^*_{ij} = \theta^*_i$ for all $1 \leq i \leq n_1$ and $1 \leq j \leq n_2$. Then the following pair of inequalities hold for a universal positive constant C:*

$$R\big(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}}\big) \leq \inf_{\theta \in \mathcal{C}_{n_1}}\left\{\frac{\|\theta^* - \theta\|^2}{n_1} + \frac{Ck(\theta)\sigma^2}{n}(\log n)^8\right\} \tag{2.6}$$

and

$$R\big(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}}\big) \leq C(\log n)^8\left(\frac{\sigma^2\sqrt{D(\boldsymbol{\theta}^*)}}{n}\right)^{2/3} \qquad \text{provided } nD\big(\boldsymbol{\theta}^*\big) \geq 2\sigma^2. \tag{2.7}$$

Remark 2.8. Although we only consider the case of Gaussian errors $(\boldsymbol{\varepsilon}_{ij})$ in Theorem 2.4 and Theorem 2.5 our results can be easily adapted to work for the case when the errors are *bounded* and have mean zero. This is possible via the use of a standard concentration inequality for Lipschitz convex functions of uniformly bounded random variables (see Theorem 6.2 and Proposition 6.1 in [10]).

3. General theory of LSEs, metric entropy calculations and the proof of Theorem 2.1

This section is mainly devoted to the proof of Theorem 2.1. The general theory of LSEs under convex constraints is crucially used to prove Theorem 2.1. Parts of this general theory that are relevant to the proof of Theorem 2.1 are recalled in the next subsection. Essentially, this general theory reduces the problem of bounding $R(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}})$ to certain metric entropy calculations of classes of isotonic matrices. In Section 3.2, we prove such results by extending appropriately existing metric entropy results for bivariate coordinate-wise nondecreasing functions due to Gao and Wellner [12]. Finally, in Section 3.3, we complete the proof of Theorem 2.1 by combining the metric entropy results with general results on LSEs.

3.1. General theory of LSEs

The following result due to Chatterjee ([6], Corollary 1.2), is a key technical tool for the proof of Theorem 2.1. It reduces the problem of bounding $R(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}})$ to controlling the expected supremum of an appropriate Gaussian process. This result is easier to apply in our setting compared to older results in empirical process theory described in Van de Geer [25] and Van der Vaart and Wellner [27].

Theorem 3.1 (Chatterjee). *Fix $\boldsymbol{\theta}^* \in \mathcal{M}$. Let us define the function $f_{\boldsymbol{\theta}^*} : \mathbb{R}_+ \to \mathbb{R}$ as*

$$f_{\boldsymbol{\theta}^*}(t) := \mathbb{E}\left(\sup_{\boldsymbol{\theta} \in \mathcal{M} : \|\boldsymbol{\theta}^* - \boldsymbol{\theta}\| \le t} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \boldsymbol{\varepsilon}_{ij} \left(\boldsymbol{\theta}_{ij} - \boldsymbol{\theta}^*_{ij} \right) \right) - \frac{t^2}{2}, \tag{3.1}$$

where $\boldsymbol{\varepsilon}_{ij}$'s are as in (1.4). Let $t_{\boldsymbol{\theta}^*}$ be the point in $[0, \infty)$ where $t \mapsto f_{\boldsymbol{\theta}^*}(t)$ attains its maximum (existence and uniqueness of $t_{\boldsymbol{\theta}^*}$ are proved in [6], Theorem 1.1). Then there exists a universal positive constant C such that

$$R\left(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}}\right) \le \frac{C}{n} \max\left(t^2_{\boldsymbol{\theta}^*}, \sigma^2\right). \tag{3.2}$$

The above theorem reduces the problem of bounding $R(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}})$ to that of bounding $t_{\boldsymbol{\theta}^*}$. For this latter problem, [6], Proposition 1.3, observed that

$$t_{\boldsymbol{\theta}^*} \le t^{**} \qquad \text{whenever } t^{**} > 0 \text{ and } f_{\boldsymbol{\theta}^*}\left(t^{**}\right) \le 0.$$

In order to bound $t_{\boldsymbol{\theta}^*}$, one therefore seeks $t^{**} > 0$ such that $f_{\boldsymbol{\theta}^*}(t^{**}) \le 0$. This now requires a bound on the expected supremum of the Gaussian process in the definition of $f_{\boldsymbol{\theta}^*}(t)$ in (3.1).

It will be convenient below to have the following notation. For $n_1 \times n_2$ matrices $\boldsymbol{M}, \boldsymbol{N} \in \mathbb{R}^{n_1 \times n_2}$, let $\|\boldsymbol{M} - \boldsymbol{N}\|$ denote the Frobenius distance between $\boldsymbol{M}$ and $\boldsymbol{N}$ defined by

$$\|\boldsymbol{M} - \boldsymbol{N}\|^2 := \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} (\boldsymbol{M}_{ij} - \boldsymbol{N}_{ij})^2.$$

For a subset $\mathcal{F} \subseteq \mathbb{R}^{n_1 \times n_2}$ and $\varepsilon > 0$, let $N(\varepsilon, \mathcal{F})$ denote the ε-covering number of $\mathcal{F}$ under the Frobenius metric $\|\cdot\|$ (i.e., $N(\varepsilon, \mathcal{F})$ is the minimum number of balls of radius ε required to cover $\mathcal{F}$). Also, for each $\boldsymbol{\theta}^* \in \mathcal{M}$ and $t > 0$, let

$$B(\boldsymbol{\theta}^*, t) := \{\boldsymbol{\theta} \in \mathcal{M} : \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\| \le t\} \tag{3.3}$$

denote the ball of radius t around $\boldsymbol{\theta}^*$. Observe that the supremum in the definition of (3.1) is over all $\boldsymbol{\theta} \in B(\boldsymbol{\theta}^*, t)$. Finally let

$$\langle \boldsymbol{\varepsilon}, \boldsymbol{\theta} - \boldsymbol{\theta}^* \rangle := \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \boldsymbol{\varepsilon}_{ij} (\boldsymbol{\theta}_{ij} - \boldsymbol{\theta}^*_{ij}).$$

The following chaining result gives an upper bound on the expected suprema of the above Gaussian process (see, e.g., Van de Geer [25]); see [9] for a proof.

Theorem 3.2 (Chaining). *For every $\boldsymbol{\theta}^* \in \mathcal{M}$ and $t > 0$,*

$$\mathbb{E}\Big[\sup_{\boldsymbol{\theta} \in B(\boldsymbol{\theta}^*, t)} \langle \boldsymbol{\varepsilon}, \boldsymbol{\theta} - \boldsymbol{\theta}^* \rangle\Big] \le \sigma \inf_{0 < \delta \le 2t} \left\{ 12 \int_{\delta}^{2t} \sqrt{\log N(\varepsilon, B(\boldsymbol{\theta}^*, t))}\, d\varepsilon + 4\delta\sqrt{n} \right\}.$$

The general results outlined here essentially reduce the problem of bounding $R(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}})$ to controlling the metric entropy of subsets of $\mathcal{M}$ of the form $B(\boldsymbol{\theta}^*, t)$. Such a metric entropy bound is proved in the next subsection. This is the key technical component in the proof of Theorem 2.1.

3.2. Main metric entropy result

Let $\mathbf{0}$ denote the $n_1 \times n_2$ matrix all of whose entries are equal to 0. According to the notation (3.3), we have

$$B(\mathbf{0}, 1) = \{\boldsymbol{\theta} \in \mathcal{M} : \|\boldsymbol{\theta} - \mathbf{0}\| \le 1\} = \left\{ \boldsymbol{\theta} \in \mathcal{M} : \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \boldsymbol{\theta}_{ij}^2 \le 1 \right\}. \tag{3.4}$$

The next theorem gives an upper bound on the ε-covering number of $B(\mathbf{0}, 1)$ (all covering numbers will be with respect to the Frobenius metric $\|\cdot\|$). It will be crucially used in our proof of Theorem 2.1.

Theorem 3.3. *There exists a universal positive constant C such that the following inequality holds for every $\varepsilon > 0$ and integers $n_1, n_2 > 1$:*

$$\log N\big(\varepsilon, B(\mathbf{0}, 1)\big) \le C \frac{(\log n_1)^2 (\log n_2)^2}{\varepsilon^2} \left[\log \frac{4\sqrt{\log n_1 \log n_2}}{\varepsilon}\right]^2. \tag{3.5}$$

Moreover for every $0 < \delta \le 1$,

$$\int_\delta^1 \sqrt{\log N\big(\varepsilon, B(\mathbf{0}, 1)\big)}\, d\varepsilon \le \frac{\sqrt{C}}{2} (\log n_1)(\log n_2) \left(\log \frac{4\sqrt{\log n_1 \log n_2}}{\delta}\right)^2. \tag{3.6}$$

In the rest of this section, we shall provide an overview of the proof of Theorem 3.3. We shall also state the main lemmas that are used in the proof of Theorem 3.3.

There is a close connection between metric entropy results for isotonic matrices and those for bivariate coordinate-wise nondecreasing functions. Indeed, for every isotonic matrix $\boldsymbol{\theta}$, we can associate a bivariate coordinate-wise nondecreasing function $\phi_{\boldsymbol{\theta}} : [0, 1]^2 \to \mathbb{R}$ via

$$\phi_{\boldsymbol{\theta}}(x_1, x_2) := \min\{\boldsymbol{\theta}_{ij} : n_1 x_1 \le i \le n_1, n_2 x_2 \le j \le n_2\}$$

for all $(x_1, x_2) \in [0, 1]^2$. It can then be directly verified that

$$\|\boldsymbol{\theta} - \boldsymbol{\nu}\|^2 = n \int_0^1 \int_0^1 \big(\phi_{\boldsymbol{\theta}}(x_1, x_2) - \phi_{\boldsymbol{\nu}}(x_1, x_2)\big)^2 \, dx_1 \, dx_2$$

for every pair $\boldsymbol{\theta}, \boldsymbol{\nu}$ of isotonic matrices. This means that metric entropy results for classes of isotonic matrices can be derived from those of bivariate nondecreasing functions. However existing metric entropy results for classes of bivariate nondecreasing functions (see Gao and Wellner [12]) require the functions to be uniformly bounded. If the average constraint in the definition (3.4) of $B(\mathbf{0}, 1)$ is replaced by a supremum constraint that is, if one considers the smaller set $B_\infty(\mathbf{0}, n^{-1/2}) := \{\boldsymbol{\theta} \in \mathcal{M} : \sup_{1 \le i \le n_1, 1 \le j \le n_2} |\boldsymbol{\theta}_{ij}| \le n^{-1/2}\}$, then the metric entropy of $B_\infty(\mathbf{0}, n^{-1/2})$ can be easily controlled via the results of Gao and Wellner [12]. This is the content of the following lemma where we actually consider the classes

$$B_\infty(\mathbf{0}, t) := \Big\{\boldsymbol{\theta} \in \mathcal{M} : \sup_{1 \le i \le n_1, 1 \le j \le n_2} |\boldsymbol{\theta}_{ij}| \le t\Big\}$$

for general $t > 0$.

Lemma 3.4. *There exists a universal positive constant C such that*

$$\log N\big(\varepsilon, B_\infty(\mathbf{0}, t)\big) \le C \left(\frac{t\sqrt{n}}{\varepsilon}\right)^2 \left[\log\left(\frac{t\sqrt{n}}{\varepsilon}\right)\right]^2$$

for every $t > 0$ and $\varepsilon > 0$.

Lemma 3.4 does not automatically imply Theorem 3.3 simply because the class $B(\mathbf{0}, 1)$ is much larger than $B_\infty(\mathbf{0}, n^{-1/2})$. Nevertheless, it turns out that the entries $\boldsymbol{\theta}_{ij}$ of a matrix $\boldsymbol{\theta}$ in $B(\mathbf{0}, 1)$ are bounded provided $\min(i-1, n_1-i)$ and $\min(j-1, n_2-j)$ are not too small. This is the content of Lemma 3.5 given below.

Lemma 3.5. *The following holds for every* $\boldsymbol{\theta} \in B(\mathbf{0}, 1)$ *and* $1 \le i \le n_1, 1 \le j \le n_2$:

$$|\boldsymbol{\theta}_{ij}| \le \max\left(\sqrt{\frac{1}{ij}}, \sqrt{\frac{1}{(n_1-i+1)(n_2-j+1)}}\right). \tag{3.7}$$

Using Lemma 3.5, we employ a peeling-type argument to prove Theorem 3.3 where we partition the entries of the matrix $\boldsymbol{\theta}$ into various subrectangles and use Lemma 3.4 in each subrectangle. The complete proof of Theorem 3.3 along with the proofs of Lemmas 3.4 and 3.5 are given in the Appendix.

3.3. Proof of Theorem 2.1

We provide the proof of Theorem 2.1 here using the results from the last two subsections.

Fix $\boldsymbol{\theta}^* \in \mathcal{M}$ and let $f_{\boldsymbol{\theta}^*}(\cdot)$ be defined as in (3.1) with $t_{\boldsymbol{\theta}^*}$ being the point in $[0, \infty)$ where $t \mapsto f_{\boldsymbol{\theta}^*}(t)$ attains its maximum.

Let $\overline{\boldsymbol{\theta}^*}$ denote the constant matrix taking the value $\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\boldsymbol{\theta}^*_{ij}/n$, i.e., $\overline{\boldsymbol{\theta}^*}_{kl} = \sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\boldsymbol{\theta}^*_{ij}/n$ for all $1 \le k \le n_1$ and $1 \le l \le n_2$. Writing $\boldsymbol{\theta} = \boldsymbol{\theta} - \overline{\boldsymbol{\theta}^*} + \overline{\boldsymbol{\theta}^*}$, we have

$$\sup_{\boldsymbol{\theta} \in B(\boldsymbol{\theta}^*, t)} \langle \boldsymbol{\varepsilon}, \boldsymbol{\theta} - \boldsymbol{\theta}^* \rangle = \sup_{\boldsymbol{\theta} \in B(\boldsymbol{\theta}^*, t)} \langle \boldsymbol{\varepsilon}, \boldsymbol{\theta} - \overline{\boldsymbol{\theta}^*} \rangle + \langle \boldsymbol{\varepsilon}, \overline{\boldsymbol{\theta}^*} - \boldsymbol{\theta}^* \rangle$$

for every $t \ge 0$. Taking expectations on both sides with respect to $\boldsymbol{\varepsilon}$, we obtain

$$\mathbb{E} \sup_{\boldsymbol{\theta} \in B(\boldsymbol{\theta}^*, t)} \langle \boldsymbol{\varepsilon}, \boldsymbol{\theta} - \boldsymbol{\theta}^* \rangle = \mathbb{E} \sup_{\boldsymbol{\theta} \in B(\boldsymbol{\theta}^*, t)} \langle \boldsymbol{\varepsilon}, \boldsymbol{\theta} - \overline{\boldsymbol{\theta}^*} \rangle. \tag{3.8}$$

Now by the triangle inequality, it is easy to see that

$$B(\boldsymbol{\theta}^*, t) \subseteq B(\overline{\boldsymbol{\theta}^*}, r_t) \qquad \text{where } r_t := t + \sqrt{nV(\boldsymbol{\theta}^*)}.$$

This and (3.8) together imply that

$$\mathbb{E} \sup_{\boldsymbol{\theta} \in B(\boldsymbol{\theta}^*, t)} \langle \boldsymbol{\varepsilon}, \boldsymbol{\theta} - \boldsymbol{\theta}^* \rangle \le \mathbb{E} \sup_{\boldsymbol{\theta} \in B(\overline{\boldsymbol{\theta}^*}, r_t)} \langle \boldsymbol{\varepsilon}, \boldsymbol{\theta} - \overline{\boldsymbol{\theta}^*} \rangle.$$

Because $\overline{\boldsymbol{\theta}^*}$ is a constant matrix, it is easy to see that

$$\sup_{\boldsymbol{\theta} \in B(\overline{\boldsymbol{\theta}^*}, r_t)} \langle \boldsymbol{\varepsilon}, \boldsymbol{\theta} - \overline{\boldsymbol{\theta}^*} \rangle = \sup_{\boldsymbol{\theta} \in B(\mathbf{0}, r_t)} \langle \boldsymbol{\varepsilon}, \boldsymbol{\theta} \rangle = r_t \sup_{\boldsymbol{\theta} \in B(\mathbf{0}, 1)} \langle \boldsymbol{\varepsilon}, \boldsymbol{\theta} \rangle,$$

where $\mathbf{0}$ denotes the constant matrix with all entries equal to 0.

As a consequence, we have

$$f_{\boldsymbol{\theta}^*}(t) \le r_t \mathbb{E} \sup_{\boldsymbol{\theta} \in B(\mathbf{0},1)} \langle \boldsymbol{\varepsilon}, \boldsymbol{\theta} \rangle - \frac{t^2}{2} \qquad \text{for all } t \ge 0. \tag{3.9}$$

We now use Theorem 3.2 with $\delta = 1/\sqrt{n}$ to obtain

$$\mathbb{E} \sup_{\boldsymbol{\theta} \in B(\mathbf{0},1)} \langle \boldsymbol{\varepsilon}, \boldsymbol{\theta} \rangle \le 12\sigma \int_{1/\sqrt{n}}^{2} \sqrt{\log N\big(\varepsilon, B(\mathbf{0}, 1)\big)}\, d\varepsilon + 4\sigma.$$

Inequality (3.6) with $\delta = n^{-1/2}$ then gives

$$\mathbb{E} \sup_{\boldsymbol{\theta} \in B(\mathbf{0},1)} \langle \boldsymbol{\varepsilon}, \boldsymbol{\theta} \rangle \le C\sigma \big(A\big(\log(B\sqrt{n})\big)^2 + 1\big)$$

with $A := (\log n_1)(\log n_2)$ and $B := 4\sqrt{(\log n_1)(\log n_2)}$.

Thus, letting $g(t) := C r_t \sigma (A(\log(B\sqrt{n}))^2 + 1)$, we obtain from (3.9) that

$$f_{\boldsymbol{\theta}^*}(t) \le g(t) - \frac{t^2}{2} \qquad \text{for all } t \ge 0.$$

It can now be directly verified that

$$f_{\boldsymbol{\theta}^*}\big(t^{**}\big) \le g\big(t^{**}\big) - \frac{1}{2}\big(t^{**}\big)^2 \le 0 \qquad \text{for } t^{**} := 2C\sqrt{\gamma^2 + \gamma\big(nV\big(\boldsymbol{\theta}^*\big)\big)^{1/2}},$$

where $\gamma := \sigma(A(\log(B\sqrt{n}))^2 + 1)$. Inequality (3.2) in Theorem 3.1 therefore gives

$$R\big(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}^*\big) \le \frac{C}{n} \max\big(\big(t^{**}\big)^2, \sigma^2\big). \tag{3.10}$$

Now $(t^{**})^2 = C(\gamma^2 + \gamma\sqrt{nV(\boldsymbol{\theta}^*)})$ and using the expressions for A and B, it is easy to see that (note that $n > 1$ because $n_1, n_2 > 1$)

$$\gamma = \sigma\big(A\big(\log(B\sqrt{n})\big)^2 + 1\big) \le C\sigma(\log n)^4.$$

This, along with (3.10), allows us to deduce

$$R\big(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}^*\big) \le C\bigg(\frac{\sigma^2}{n}(\log n)^8 + \sqrt{\frac{\sigma^2 V(\boldsymbol{\theta}^*)}{n}}(\log n)^4\bigg),$$

which proves Theorem 2.1.

4. Risk, tangent cones and the proofs of Theorems 2.4 and 2.5

This section is devoted to the proofs of Theorems 2.4 and 2.5. We use a recent result of Bellec [3] on the connection between the risk $R(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}})$ and certain size measures of tangent cones to $\mathcal{M}$ at $\boldsymbol{\theta}^*$. This result is recalled in the next subsection.

4.1. Risk and tangent cones

Fix $\boldsymbol{\theta} \in \mathcal{M}$. The tangent cone of $\mathcal{M}$ at $\boldsymbol{\theta}$ will be denoted by $T_{\mathcal{M}}(\boldsymbol{\theta})$ and is defined as the closure of the convex cone generated by $\boldsymbol{u} - \boldsymbol{\theta}$ as $\boldsymbol{u}$ varies over $\mathcal{M}$ that is,

$$T_{\mathcal{M}}(\boldsymbol{\theta}) := \text{closure}\bigl\{\alpha(\boldsymbol{u} - \boldsymbol{\theta}) : \alpha > 0 \text{ and } \boldsymbol{u} \in \mathcal{M}\bigr\}.$$

The tangent cone $T_{\mathcal{M}}(\boldsymbol{\theta})$ is a closed, convex subset of $\mathbb{R}^n = \mathbb{R}^{n_1 \times n_2}$. Observe that if $\boldsymbol{\theta}$ is a constant matrix (i.e., all entries of $\boldsymbol{\theta}$ are the same), then $T_{\mathcal{M}}(\boldsymbol{\theta})$ is simply equal to $\mathcal{M}$.

It turns out that the risk $R(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}})$ can be controlled by appropriate size measures of the tangent cones $T_{\mathcal{M}}(\boldsymbol{\theta}), \boldsymbol{\theta} \in \mathcal{M}$. This is formalized in the following lemma. This lemma is similar in spirit to results in Oymak and Hassibi [17]. More general such results involving model misspecification have recently appeared in Bellec [3].

Let $\boldsymbol{\varepsilon} = (\boldsymbol{\varepsilon}_{ij})$ denote the $n_1 \times n_2$ matrix all of whose entries are independent and normally distributed with zero mean and variance σ^2. The Euclidean projection of $\boldsymbol{\varepsilon}$ onto the tangent cone $T_{\mathcal{M}}(\boldsymbol{\theta})$ is defined in the usual way as

$$\Pi\bigl(\boldsymbol{\varepsilon}, T_{\mathcal{M}}(\boldsymbol{\theta})\bigr) := \operatorname*{argmin}_{\boldsymbol{u} \in T_{\mathcal{M}}(\boldsymbol{\theta})} \|\boldsymbol{\varepsilon} - \boldsymbol{u}\|^2.$$

Lemma 4.1. *For every $\boldsymbol{\theta}^* \in \mathcal{M}$, we have*

$$R\bigl(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}}\bigr) \le \frac{1}{n} \inf_{\boldsymbol{\theta} \in \mathcal{M}} \bigl(\bigl\|\boldsymbol{\theta}^* - \boldsymbol{\theta}\bigr\|^2 + \mathbb{E}\bigl\|\Pi\bigl(\boldsymbol{\varepsilon}, T_{\mathcal{M}}(\boldsymbol{\theta})\bigr)\bigr\|^2\bigr), \tag{4.1}$$

where the expectation on the right-hand side is with respect to $\boldsymbol{\varepsilon}$.

Proof. Recall that $\boldsymbol{y} = \boldsymbol{\theta}^* + \boldsymbol{\varepsilon}$ and that $\hat{\boldsymbol{\theta}}$ is the projection of the data matrix $\boldsymbol{y}$ onto $\mathcal{M}$. By the usual KKT conditions, this projection $\hat{\boldsymbol{\theta}}$ satisfies

$$\langle \boldsymbol{y} - \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\rangle \ge 0 \qquad \text{for every } \boldsymbol{\theta} \in \mathcal{M},$$

where $\langle A, B\rangle = \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} a_{ij} b_{ij}$ for $A = (a_{ij})$ and $B = (b_{ij})$. This inequality implies that

$$\|\boldsymbol{y} - \boldsymbol{\theta}\|^2 \ge \|\boldsymbol{y} - \hat{\boldsymbol{\theta}}\|^2 + \|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|^2 \qquad \text{for every } \boldsymbol{\theta} \in \mathcal{M}.$$

Writing $\boldsymbol{y} = \boldsymbol{\theta}^* + \boldsymbol{\varepsilon}$, expanding out the squares and rearranging terms, we obtain

$$\bigl\|\boldsymbol{\theta}^* - \boldsymbol{\theta}\bigr\|^2 + \|\boldsymbol{\varepsilon}\|^2 + 2\bigl\langle\boldsymbol{\theta}^* - \boldsymbol{\theta}, \boldsymbol{\varepsilon}\bigr\rangle \ge \bigl\|\boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}\bigr\|^2 + \|\boldsymbol{\varepsilon}\|^2 + 2\bigl\langle\boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}, \boldsymbol{\varepsilon}\bigr\rangle + \|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|^2,$$

$$\text{i.e.,}\qquad \|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|^2 \le 2\langle \hat{\boldsymbol{\theta}} - \boldsymbol{\theta}, \boldsymbol{\varepsilon}\rangle - \|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|^2 + \|\boldsymbol{\theta}^* - \boldsymbol{\theta}\|^2,$$

$$\text{i.e.,}\qquad \|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|^2 \le \|\boldsymbol{\theta}^* - \boldsymbol{\theta}\|^2 + \|\boldsymbol{\varepsilon}\|^2 - \|\boldsymbol{\varepsilon} - (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})\|^2.$$

Because $\hat{\boldsymbol{\theta}} \in \mathcal{M}$, the matrix $\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}$ belongs to the tangent cone $T_{\mathcal{M}}(\boldsymbol{\theta})$. We therefore get

$$\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|^2 \le \|\boldsymbol{\theta}^* - \boldsymbol{\theta}\|^2 + \|\boldsymbol{\varepsilon}\|^2 - \inf_{\boldsymbol{u} \in T_{\mathcal{M}}(\boldsymbol{\theta})} \|\boldsymbol{\varepsilon} - \boldsymbol{u}\|^2.$$

The infimum over $\boldsymbol{u}$ above is clearly achieved for $\boldsymbol{u} := \Pi(\boldsymbol{\varepsilon}, T_{\mathcal{M}}(\boldsymbol{\theta}))$ and hence

$$\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|^2 \le \|\boldsymbol{\theta}^* - \boldsymbol{\theta}\|^2 + \|\boldsymbol{\varepsilon}\|^2 - \|\boldsymbol{\varepsilon} - \Pi(\boldsymbol{\varepsilon}, T_{\mathcal{M}}(\boldsymbol{\theta}))\|^2. \tag{4.2}$$

Because $T_{\mathcal{M}}(\boldsymbol{\theta})$ is a closed convex cone, the projection $\Pi(\boldsymbol{\varepsilon}, T_{\mathcal{M}}(\boldsymbol{\theta}))$ satisfies (see, for example, [16], equation (4)):

$$\langle \boldsymbol{\varepsilon} - \Pi(\boldsymbol{\varepsilon}, T_{\mathcal{M}}(\boldsymbol{\theta})), \Pi(\boldsymbol{\varepsilon}, T_{\mathcal{M}}(\boldsymbol{\theta}))\rangle = 0.$$

The above equality and inequality (4.2) together imply that

$$\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|^2 \le \|\boldsymbol{\theta}^* - \boldsymbol{\theta}\|^2 + \|\Pi(\boldsymbol{\varepsilon}, T_{\mathcal{M}}(\boldsymbol{\theta}))\|^2.$$

The required inequality (4.1) now follows by taking expectations on both sides. □

Inequality (4.1) reduces the problem of bounding the risk to controlling the expected squared norm of the projection of $\boldsymbol{\varepsilon}$ onto the tangent cones $T_{\mathcal{M}}(\boldsymbol{\theta}), \boldsymbol{\theta} \in \mathcal{M}$. This will be crucially used in the proof of Theorem 2.4.

4.2. Proof of Theorem 2.4

We provide the proof of Theorem 2.4 in this subsection. The first step is to characterize the tangent cone $T_{\mathcal{M}}(\boldsymbol{\theta})$ for every $\boldsymbol{\theta} \in \mathcal{M}$. We need some notation here. For a subset S of $\{(i, j) : 1 \le i \le n_1, 1 \le j \le n_2\}$, let $\mathbb{R}^S$ denote the class of all real-valued functions from S to $\mathbb{R}$. Elements of $\mathbb{R}^S$ will be denoted by $(\boldsymbol{\theta}_{ij}, (i, j) \in S)$. We say that $(\boldsymbol{\theta}_{ij} : (i, j) \in S)$ is isotonic if

$$\boldsymbol{\theta}_{ij} \le \boldsymbol{\theta}_{kl} \qquad \text{whenever } (i, j), (k, l) \in S \text{ with } i \le k \text{ and } j \le l.$$

The set of such isotonic sequences in $\mathbb{R}^S$ will be denoted by $\mathcal{M}(S)$. Also for every two dimensional array $\boldsymbol{\theta} = (\boldsymbol{\theta}_{ij} : 1 \le i \le n_1, 1 \le j \le n_2)$, let

$$\boldsymbol{\theta}(S) := (\boldsymbol{\theta}_{ij} : (i, j) \in S).$$

Observe that $\boldsymbol{\theta}(S) \in \mathcal{M}(S)$ if $\boldsymbol{\theta} \in \mathcal{M}$. The following lemma provides a useful characterization of $T_{\mathcal{M}}(\boldsymbol{\theta})$ for $\boldsymbol{\theta} \in \mathcal{M}$. Recall that a rectangular partition of $[n_1] \times [n_2]$ is a partition of $[n_1] \times [n_2]$ into rectangles. The cardinality $|\pi|$ of a rectangular partition π equals the number of rectangles in the partition. The collection of all rectangular partitions of $[n_1] \times [n_2]$ is denoted by $\mathcal{P}$. We say that $\boldsymbol{\theta} \in \mathcal{M}$ is constant on $\pi = (A_1, \dots, A_k) \in \mathcal{P}$ if $\{\boldsymbol{\theta}_{ij} : (i, j) \in A_l\}$ is a singleton for each l.

Lemma 4.2. *Fix $\boldsymbol{\theta} \in \mathcal{M}$ and $\pi = (A_1, \dots, A_k) \in \mathcal{P}$ such that $\boldsymbol{\theta}$ is constant on π. Then*

$$T_{\mathcal{M}}(\boldsymbol{\theta}) \subseteq \{\boldsymbol{v} \in \mathbb{R}^n : \boldsymbol{v}(A_i) \in \mathcal{M}(A_i) \text{ for each } i = 1, \dots, k\}. \tag{4.3}$$

Proof. Suppose that $\boldsymbol{v} = \alpha(\boldsymbol{t} - \boldsymbol{\theta})$ for some $\boldsymbol{t} \in \mathcal{M}$ and $\alpha > 0$. This means that $\boldsymbol{v}(A_i) = \alpha(\boldsymbol{t}(A_i) - \boldsymbol{\theta}(A_i))$ for each i. Because $\boldsymbol{t}(A_i) \in \mathcal{M}(A_i)$ and $\boldsymbol{\theta}(A_i)$ is a constant ($\boldsymbol{\theta}$ is constant on π), we now have $\boldsymbol{v}(A_i) \in \mathcal{M}(A_i)$. As the right-hand side of (4.3) is a closed set, and $T_{\mathcal{M}}(\boldsymbol{\theta})$ is the closure of all such $\boldsymbol{v}$'s, the desired result follows. □

Remark 4.1. Note that we did not use the fact that $A_1, \dots, A_k$ are rectangular in Lemma 4.2. We only used the fact that $\boldsymbol{\theta}$ is constant on each A_i. This means that (4.3) is true also when $A_1, \dots, A_k$ are the levels sets of $\boldsymbol{\theta}$ i.e., each $A_l = \{(i, j) : \boldsymbol{\theta}_{ij} = a\}$ for some real number a. In fact, when $A_1, \dots, A_k$ are the level sets of $\boldsymbol{\theta}$, we have equality in (4.3). This can be proved as follows.

Suppose that $\boldsymbol{v}(A_i) \in \mathcal{M}(A_i)$ for each i. We shall argue then that $\boldsymbol{\theta} + \alpha\boldsymbol{v} \in \mathcal{M}$ for some $\alpha > 0$ which, of course, proves that $\boldsymbol{v} \in T_{\mathcal{M}}(\boldsymbol{\theta})$. Observe first that $A_1, \dots, A_k$ form a partition of $[n_1] \times [n_2]$. Let D denote the collection of all pairs $((i, j), (k, l))$ such that $i \le j$ and $k \le l$ and $\boldsymbol{\theta}_{ij} \ne \boldsymbol{\theta}_{kl}$. Note, in particular, that (i, j) and (k, l) belong to different elements of the partition $A_1, \dots, A_k$ if $((i, j), (k, l)) \in D$. Let

$$\alpha := \min\left\{\frac{\boldsymbol{\theta}_{kl} - \boldsymbol{\theta}_{ij}}{\boldsymbol{v}_{ij} - \boldsymbol{v}_{kl}} : \big((i, j), (k, l)\big) \in D \text{ and } \boldsymbol{v}_{ij} > \boldsymbol{v}_{kl}\right\}.$$

By monotonicity of $\boldsymbol{\theta}$, it is clear that $\alpha > 0$. With this choice of α, it is elementary to check that $\boldsymbol{\theta} + \alpha\boldsymbol{v} \in \mathcal{M}$. This shows that (4.3) is true with equality when $A_1, \dots, A_k$ are the level sets of π.

We now have all the tools to complete the proof of Theorem 2.4.

Proof of Theorem 2.4. The first step is to observe via inequality (4.1) that it is enough to prove the existence of a universal positive constant C for which

$$\mathbb{E}\big\|\Pi\big(\boldsymbol{\varepsilon}, T_{\mathcal{M}}(\boldsymbol{\theta})\big)\big\|^2 \le Ck(\boldsymbol{\theta})\sigma^2(\log n)^8 \qquad \text{for all } \boldsymbol{\theta} \in \mathcal{M}.$$

From the definition of $k(\boldsymbol{\theta})$, it is enough of prove that

$$\mathbb{E}\big\|\Pi\big(\boldsymbol{\varepsilon}, T_{\mathcal{M}}(\boldsymbol{\theta})\big)\big\|^2 \le Ck\sigma^2(\log n)^8 \tag{4.4}$$

for every $\pi = (A_1, \dots, A_k) \in \mathcal{P}$ such that $\boldsymbol{\theta}$ is constant on π. To prove (4.4), use the characterization of $T_{\mathcal{M}}(\boldsymbol{\theta})$ in Lemma 4.2 to observe that

$$\mathbb{E}\big\|\Pi\big(\boldsymbol{\varepsilon}, T_{\mathcal{M}}(\boldsymbol{\theta})\big)\big\|^2 \le \sum_{i=1}^{k} \mathbb{E}\big\|\Pi\big(\boldsymbol{\varepsilon}(A_i), \mathcal{M}(A_i)\big)\big\|^2. \tag{4.5}$$

The task then reduces to that of bounding $\mathbb{E}\|\Pi(\boldsymbol{\varepsilon}(A_i), \mathcal{M}(A_i))\|^2$ for $i = 1, \dots, k$. It is crucial that each $A_1, \dots, A_k$ is a rectangle. Fix $1 \le i \le k$ and without loss of generality assume that

$A_i = [n_1'] \times [n_2']$ for some $1 \le n_1' \le n_1$ and $1 \le n_2' \le n_2$. It is then easy to see that Theorem 2.1 for $\boldsymbol{\theta}^* = \mathbf{0}$ and $n_1 = n_1'$, $n_2 = n_2'$ immediately gives

$$\mathbb{E}\big\|\Pi\big(\boldsymbol{\varepsilon}(A_i), \mathcal{M}(A_i)\big)\big\|^2 \le C\sigma^2\big(\log\big(2n_1'n_2'\big)\big)^8 \tag{4.6}$$

for a universal positive constant C as long as $n_1' > 1$ and $n_2' > 1$. When $n_1' = n_2' = 1$, it can be checked that the left-hand side of (4.6) equals σ^2 which means that (4.6) is still true provided C is changed accordingly. Finally when $\min(n_1', n_2') = 1$ and $\max(n_1', n_2') > 1$, one can use the result (1.2) from vector isotonic estimation to prove (4.6). We thus have

$$\mathbb{E}\big\|\Pi\big(\boldsymbol{\varepsilon}(A_i), \mathcal{M}(A_i)\big)\big\|^2 \le C\sigma^2(\log n)^8$$

for a universal constant C for all $n_1' \ge 1$ and $n_2' \ge 1$. This inequality together with inequality (4.5) implies (4.4) which completes the proof of Theorem 2.4. □

4.3. Proof of Theorem 2.5

We now give the proof of Theorem 2.5. Let us first prove inequality (2.6). For $\theta \in \mathcal{C}_{n_1}$, let $\Upsilon(\theta) \in \mathcal{M}$ be defined by $\Upsilon(\theta)_{ij} = \theta_i$ for all $1 \le i \le n_1$ and $1 \le j \le n_2$. Also let $\Upsilon(\mathcal{C}_{n_1}) := \{\Upsilon(\theta) : \theta \in \mathcal{C}_{n_1}\}$. Note first that all level sets of $\Upsilon(\theta)$ are rectangular for every $\theta \in \mathcal{C}_{n_1}$ which implies that $k(\Upsilon(\theta)) = k(\theta)$ for every $\theta \in \mathcal{C}_{n_1}$. Therefore, as a consequence of Theorem 2.4, we obtain that for every $\boldsymbol{\theta}^* \in \mathcal{M}$,

$$R\big(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}}\big) \le \inf_{\theta \in \mathcal{C}_{n_1}} \left(\frac{\|\boldsymbol{\theta}^* - \Upsilon(\theta)\|^2}{n} + \frac{Ck(\theta)\sigma^2}{n}(\log n)^8 \right).$$

Now if there exists $\theta^* \in \mathcal{C}_{n_1}$ such that $\Upsilon(\theta^*) = \boldsymbol{\theta}^*$, then it is obvious that $\|\boldsymbol{\theta}^* - \Upsilon(\theta)\|^2 = n_2\|\theta^* - \theta\|^2$ which proves (2.6).

Inequality (2.7) can now be derived from (2.6) by a standard approximation argument. For every $\theta^* \in \mathcal{C}_{n_1}$ with $D = D(\theta^*) = (\theta^*_{n_1} - \theta^*_1)^2$ and $0 \le \delta \le \sqrt{D}$, there exists $\theta \in \mathcal{C}_{n_1}$ with

$$\frac{\|\theta - \theta^*\|^2}{n_1} \le \delta^2 \quad \text{and} \quad k(\theta) \le \frac{2\sqrt{D}}{\delta}.$$

This fact is easy to prove and a proof can be found, for example, in Chatterjee *et al.* [8], Lemma B.1. Using this, it follows directly from (2.6) that

$$R\big(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}}\big) \le C \inf_{0<\delta\le\sqrt{D}} \left(\delta^2 + \frac{2\sigma^2\sqrt{D}}{n\delta}(\log n)^8 \right).$$

The choice $\delta = (2\sigma^2\sqrt{D}/n)^{1/3}$ now leads to inequality (2.7). This choice of δ satisfies $\delta \le \sqrt{D}$ provided $nD \ge 2\sigma^2$. This completes the proof of Theorem 2.5.

5. Discussion

In this paper, we have studied the risk behavior of the LSE of an unknown matrix $\boldsymbol{\theta}^*$, constrained to be nondecreasing in both rows and columns, when observed with errors. We prove both worst case and adaptive risk bounds for the LSE. A highlight of the adaptation properties of the LSE is that it adapts automatically to the intrinsic dimension of the problem.

Two further research questions are mentioned below.

The logarithmic factors in our risk bounds, for example, in Theorems 2.1 and 2.4, are probably not optimal. They arise as a consequence of (i) the presence of logarithmic factors in the covering number result in Gao and Wellner [12] (see the proof of Lemma 3.4), and (ii) the fact that the entropy integral in (3.6) in Theorem 3.3 diverges to $+\infty$ if $\delta \downarrow 0$. It is not clear to us at the moment how to remove or reduce these logarithmic factors.

In this paper, we deal with the estimation of an isotonic matrix. It is natural to ask how the results generalize to isotonic tensors of higher order, and more generally to estimating a multivariate isotonic regression function under general designs. It would be interesting to see whether such adaptation results hold in these situations.

Appendix

A.1. Proof of Lemma 3.4

For each $\boldsymbol{\theta} \in B_\infty(\mathbf{0}, t)$, we associate a bivariate coordinate-wise nondecreasing function $\phi_{\boldsymbol{\theta}} : [0,1]^2 \to \mathbb{R}$ via

$$\phi_{\boldsymbol{\theta}}(x_1, x_2) := \min\{\boldsymbol{\theta}_{ij} : n_1 x_1 \le i \le n_1, n_2 x_2 \le j \le n_2\}$$

for all $(x_1, x_2) \in [0,1]^2$. It can then be directly verified that

$$\|\boldsymbol{\theta} - \boldsymbol{\nu}\|^2 = n \int_0^1 \int_0^1 \big(\phi_{\boldsymbol{\theta}}(x_1, x_2) - \phi_{\boldsymbol{\nu}}(x_1, x_2)\big)^2 \, dx_1 \, dx_2$$

for every pair $\boldsymbol{\theta}, \boldsymbol{\nu} \in B_\infty(\mathbf{0}, t)$. Moreover, if $\mathcal{C}([0,1]^2, t)$ denotes the class of all bivariate coordinate-wise nondecreasing functions that are uniformly bounded by t, then it is straightforward to verify that $\phi_{\boldsymbol{\theta}} \in \mathcal{C}([0,1]^2, t)$ for every $\boldsymbol{\theta} \in B_\infty(\mathbf{0}, t)$. These two latter facts immediately imply that

$$N\big(\varepsilon, B_\infty(\mathbf{0}, t)\big) \le N\big(n^{-1/2}\varepsilon/2, \mathcal{C}\big([0,1]^2, t\big), L_2\big), \tag{A.1}$$

where $N(\varepsilon/2, \mathcal{C}([0,1]^2, t), L_2)$ denotes the $\varepsilon/2$-covering number of $\mathcal{C}([0,1]^2, t)$ under the L_2 metric $L_2(f, g) := (\int (f-g)^2)^{1/2}$. This latter covering number has been studied by Gao and Wellner [12] who proved that

$$N\big(\varepsilon/2, \mathcal{C}\big([0,1]^2, t\big), L_2\big) \le C\left(\frac{t}{\varepsilon}\right)^2 \left[\log\left(\frac{t}{\varepsilon}\right)\right]^2$$

for a universal positive constant C. This and (A.1) together complete the proof of Lemma 3.4.

A.2. Proof of Lemma 3.5

Fix $\boldsymbol{\theta} \in B(\mathbf{0}, 1)$ and $1 \le i \le n_1$, $1 \le j \le n_2$. Our proof of (3.7) involves considering the following two cases separately:

1. $\boldsymbol{\theta}_{ij} < 0$: Here, by monotonicity of $\boldsymbol{\theta}$, the inequality $\boldsymbol{\theta}_{kl} \le \boldsymbol{\theta}_{ij}$ must hold for all $1 \le k \le i$ and $1 \le l \le j$. Therefore, $|\boldsymbol{\theta}_{kl}| \ge |\boldsymbol{\theta}_{ij}|$ holds for all $(k, l) \in [1, i] \times [1, j]$. Finally because $\boldsymbol{\theta} \in B(\mathbf{0}, 1)$, we have
$$1 \ge \sum_{k=1}^{i} \sum_{l=1}^{j} \boldsymbol{\theta}_{kl}^2 \ge ij\boldsymbol{\theta}_{ij}^2.$$
This proves (3.7) when $\boldsymbol{\theta}_{ij} < 0$.
2. $\boldsymbol{\theta}_{ij} \ge 0$. Here by monotonicity of $\boldsymbol{\theta}$, the condition $\boldsymbol{\theta}_{kl} \ge \boldsymbol{\theta}_{ij}$ must hold for all $i \le k \le n_1$ and $j \le l \le n_2$. Therefore, by nonnegativity of $\boldsymbol{\theta}_{ij}$ and by virtue of $\boldsymbol{\theta} \in B(\mathbf{0}, 1)$ we have
$$1 \ge \sum_{k=i}^{n_1} \sum_{l=j}^{n_2} \boldsymbol{\theta}_{kl}^2 \ge (n_1 + 1 - i)(n_2 + 1 - j)\boldsymbol{\theta}_{ij}^2.$$
This proves (3.7) when $\boldsymbol{\theta}_{ij} \ge 0$.

A.3. Proof of Theorem 3.3

The basic idea behind this proof is the following. By Lemma 3.5, it is clear that for every matrix $\boldsymbol{\theta} \in B(\mathbf{0}, 1)$, the entries $\boldsymbol{\theta}_{ij}$ are bounded by constants provided $\min(i - 1, n_1 - i)$ and $\min(j - 1, n_2 - j)$ are not too small. Further, for bounded isotonic matrices, the metric entropy bounds can be obtained from Lemma 3.4. We shall therefore employ a peeling-type argument where we partition the entries of $\boldsymbol{\theta}$ into various subrectangles and use Lemma 3.4 in each subrectangle.

Let us introduce some notation. Let B denote the set $B(\mathbf{0}, 1)$ for simplicity. For a subset $S \subset \{(i, j) : 1 \le i \le n_1, 1 \le j \le n_2\}$ with cardinality $|S|$ and $\boldsymbol{\theta} \in \mathcal{M}$, let $\boldsymbol{\theta}(S) \in \mathbb{R}^{|S|}$ be defined as
$$\boldsymbol{\theta}(S) := \big(\boldsymbol{\theta}_{ij} : (i, j) \in S\big).$$
Further let B_S denote the collection of all $\boldsymbol{\theta}(S)$ as $\boldsymbol{\theta}$ ranges over B. The ε-metric entropy of B_S (under the Euclidean metric on $\mathbb{R}^{|S|}$) will be denoted by $N(\varepsilon, B_S)$.

We first prove inequality (3.5). Let $I_1 := \{i : 1 \le i \le n_1/2\}$ and $I_2 := \{i : n_1/2 < i \le n_1\}$. Also $J_1 := \{j : 1 \le j \le n_2/2\}$ and $J_2 := \{j : n_2/2 < j \le n_2\}$. Because
$$\|\boldsymbol{\theta} - \boldsymbol{\alpha}\|^2 = \sum_{1 \le k,l \le 2} \big\|\boldsymbol{\theta}(I_k \times J_l) - \boldsymbol{\alpha}(I_k \times J_l)\big\|^2$$
for all $\boldsymbol{\theta}$ and $\boldsymbol{\alpha}$, it follows that
$$\log N(\varepsilon, B) \le \sum_{k=1}^{2} \sum_{l=1}^{2} \log N(\varepsilon/2, B_{I_k \times J_l}).$$

We shall prove below that for every $1 \le k, l \le 2$ and $\varepsilon > 0$,

$$\log N(\varepsilon/2, B_{I_k \times J_l}) \le C \frac{(\log n_1)^2 (\log n_2)^2}{\varepsilon^2} \left[\log \frac{4\sqrt{\log n_1 \log n_2}}{\varepsilon}\right]^2 \tag{A.2}$$

for a universal positive constant C. This would then complete the proof of (3.5).

Let k_1 and k_2 denote the smallest integers for which $2^{k_1} > n_1/2$ and $2^{k_2} > n_2/2$. For every $0 \le u < k_1$ and $0 \le v < k_2$, let

$$N_u^1 := \{i \in I_1 : 2^u \le i \le \min(2^{u+1} - 1, n_1/2)\} \quad \text{and}$$
$$N_v^1 := \{j \in J_1 : 2^v \le j \le \min(2^{v+1} - 1, n_2/2)\}.$$

Similarly, let

$$N_u^2 := \{i \in I_2 : 2^u \le n_1 + 1 - i \le \min(2^{u+1} - 1, n_1/2)\} \quad \text{and}$$
$$N_v^2 := \{j \in J_2 : 2^v \le n_2 + 1 - j \le \min(2^{v+1} - 1, n_2/2)\}.$$

For each pair $1 \le k, l \le 2$, because

$$\|\boldsymbol{\theta}(I_k \times J_l) - \boldsymbol{\alpha}(I_k \times J_l)\|^2 = \sum_{u=0}^{k_1-1} \sum_{v=0}^{k_2-1} \|\boldsymbol{\theta}(N_u^k \times N_v^l) - \boldsymbol{\alpha}(N_u^k \times N_v^l)\|^2$$

it follows that

$$\log N(\varepsilon/2, B_{I_k \times J_l}) \le \sum_{u=0}^{k_1-1} \sum_{v=0}^{k_2-1} \log N\left(k_1^{-1/2} k_2^{-1/2} \varepsilon/2, B_{N_u^k \times N_v^l}\right). \tag{A.3}$$

Now fix $0 \le u < k_1, 0 \le v < k_2$ and $1 \le k, l \le 2$. We argue below that $N(k_1^{-1/2} k_2^{-1/2} \varepsilon/2, B_{N_u^k \times N_v^l})$ can be controlled using Lemmas 3.4 and 3.5. Note first that the cardinality of $N_u^k \times N_v^l$ is at most $|N_u^k||N_v^l| \le 2^{u+v}$. We also claim that

$$\max_{i \in N_u^k, j \in N_v^l} |\boldsymbol{\theta}_{ij}| \le 2^{-(u+v)/2} \qquad \text{for all } \boldsymbol{\theta} \in B. \tag{A.4}$$

We will prove the above claim a little later. Assuming for now that it is true, we can use Lemma 3.4 for $B_{N_u^k \times N_v^l}$ to deduce that

$$\log N\left(k_1^{-1/2} k_2^{-1/2} \varepsilon/2, B_{N_u^k \times N_v^l}\right) \le C \frac{k_1 k_2}{\varepsilon^2} \left(\log \frac{4 k_1^{1/2} k_2^{1/2}}{\varepsilon}\right)^2$$

for a universal positive constant C. Inequality (A.3) then gives

$$\log N(\varepsilon/2, B_{I_k \times J_l}) \le C \frac{k_1^2 k_2^2}{\varepsilon^2} \left(\log \frac{4 k_1^{1/2} k_2^{1/2}}{\varepsilon}\right)^2. \tag{A.5}$$

Because k_1 is the smallest integer for which $2^{k_1} > n_1/2$, we have $2^{k_1-1} \le n_1/2$ which means that $k_1 \le \log n_1$. Similarly $k_2 \le \log n_2$. This together with (A.5) implies (A.2) which completes the proof of (3.5). The only thing that remains now is to prove (A.4).

We first prove (A.4) for $k = l = 1$. By Lemma 3.5, we get that $|\boldsymbol{\theta}_{ij}| \le (ij)^{-1/2}$ for all for $\boldsymbol{\theta} \in B$ and $(i, j) \in I_1 \times J_1$. Clearly $\min_{i \in N_u^1} i = 2^u$ and $\min_{j \in N_v^1} i = 2^v$. This proves (A.4) for $k = l = 1$. A similar argument will also work for $k = l = 2$. For the case when $k = 1, l = 2$, note that

$$\max_{N_u^1 \times N_v^2} \boldsymbol{\theta}_{ij} \le \max_{N_u^2 \times N_v^2} \boldsymbol{\theta}_{ij} \le 2^{-(u+v)/2},$$

which follows from the monotonicity of $\boldsymbol{\theta}$ and (A.4) for $k = l = 2$. Similarly,

$$\min_{N_u^1 \times N_v^2} \boldsymbol{\theta}_{ij} \ge \min_{N_u^1 \times N_v^1} \boldsymbol{\theta}_{ij} \ge -2^{-(u+v)/2}.$$

Putting these together, we obtain (A.4) for $k = 1, l = 2$. A similar argument will work for $k = 2, l = 1$. This completes the proof of (3.5).

For (3.6), simply observe that by (3.5),

$$\begin{aligned}\int_\delta^1 \sqrt{\log N\big(\varepsilon, B(\mathbf{0}, 1)\big)}\, d\varepsilon &\le \sqrt{C}\sqrt{A} \int_\delta^1 \frac{1}{\varepsilon}\left(\log \frac{B}{\varepsilon}\right) d\varepsilon \\ &= \frac{\sqrt{C}\sqrt{A}}{2}\left[\left(\log \frac{B}{\delta}\right)^2 - (\log B)^2\right] \le \frac{\sqrt{C}\sqrt{A}}{2}\left(\log \frac{B}{\delta}\right)^2.\end{aligned}$$

This completes the proof of Theorem 3.3.

A.4. Proof of Theorem 2.2

We shall use Assouad's lemma to prove Theorem 2.2. The following version of Assouad's lemma is a consequence of Lemma 24.3 of van der Vaart [26], page 347.

Lemma A.1 (Assouad). *Fix $D > 0$ and a positive integer d. Suppose that, for each $\tau \in \{-1, 1\}^d$, there is an associated $\boldsymbol{g}^{\boldsymbol{\tau}}$ in $\mathcal{M}$ with $D(\boldsymbol{g}^{\boldsymbol{\tau}}) \le D$. Then*

$$\inf_{\tilde{\boldsymbol{\theta}}} \sup_{\boldsymbol{\theta} \in \mathcal{M}: D(\boldsymbol{\theta}) \le D} R(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}}) \ge \frac{d}{8} \min_{\tau \ne \tau'} \frac{\ell^2(\boldsymbol{g}^\tau, \boldsymbol{g}^{\tau'})}{\Upsilon(\tau, \tau')} \min_{\Upsilon(\tau, \tau') = 1} \big(1 - \|\mathbb{P}_{\boldsymbol{g}^\tau} - \mathbb{P}_{\boldsymbol{g}^{\tau'}}\|_{\mathrm{TV}}\big),$$

where $\Upsilon(\tau, \tau') := \sum_{i=1}^d I\{\tau_i \ne \tau_i'\}$ denotes the Hamming distance between τ and τ' and $\|\cdot\|_{\mathrm{TV}}$ denotes the total variation distance. The notation $\mathbb{P}_{\boldsymbol{g}}$ for $\boldsymbol{g} \in \mathcal{M}$ refers to the joint distribution of $\boldsymbol{y}_{ij} = \boldsymbol{g}_{ij} + \boldsymbol{\varepsilon}_{ij}$, for $i \in [n_1]$, $j \in [n_2]$ when $(\boldsymbol{\varepsilon}_{ij})$ are independent normally distributed random variables with mean zero and variance σ^2.

We are now ready to prove Theorem 2.2.

Fix $D > 0$ and an integer k with $1 \le k \le \min(n_1, n_2)$. Let m_1 and m_2 be defined so that $k = \lfloor n_1/m_1 \rfloor = \lfloor n_2/m_2 \rfloor$. Let $d = k^2$. We denote elements of $\{-1,1\}^d$ by $(\tau_{uv} : u, v \in [k] \times [k])$. For each such $\tau \in \{-1,1\}^d$, we define $\boldsymbol{g}^\tau \in \mathcal{M}$ in the following way. For $i \in [n_1]$ and $j \in [n_2]$, if there exist $u, v \in [k]$ for which $(u-1)m_1 < i \le um_1$ and $(v-1)m_2 < j \le vm_2$, we take

$$\boldsymbol{g}^\tau_{ij} = \sqrt{D}\left(\frac{u+v-2}{2k} + \frac{\tau_{uv}}{6k}\right).$$

Otherwise we take $\boldsymbol{g}^\tau_{ij} = \sqrt{D}$. One can check that $\boldsymbol{g}^\tau \in \mathcal{M}$ and $D(\boldsymbol{g}^\tau) \le D$ for every $\tau \in \{-1,1\}^d$.

We shall now use Lemma A.1 with $d = k^2$ and this collection $\{\boldsymbol{g}^\tau : \tau \in \{-1,1\}^d\}$. Note first that

$$\begin{aligned}
\ell^2(\boldsymbol{g}^\tau, \boldsymbol{g}^{\tau'}) &= \frac{1}{n}\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}(\boldsymbol{g}^\tau_{ij} - \boldsymbol{g}^{\tau'}_{ij})^2 \\
&= \frac{1}{n}\sum_{u,v\in[k]}\sum_{i:(u-1)m_1<i\le um_1}\sum_{j:(v-1)m_2<j\le vm_2}(\boldsymbol{g}^\tau_{ij} - \boldsymbol{g}^{\tau'}_{ij})^2 \\
&= \frac{D}{n}\sum_{u,v\in[k]}\frac{m_1m_2}{36k^2}(\tau_{uv} - \tau'_{uv})^2 = \frac{m_1m_2D}{9nk^2}\Upsilon(\tau, \tau').
\end{aligned}$$

Therefore, this implies that

$$\min_{\tau\ne\tau'}\frac{\ell^2(\boldsymbol{g}^\tau, \boldsymbol{g}^{\tau'})}{\Upsilon(\tau,\tau')} = \frac{m_1m_2D}{9nk^2}.$$

To bound $\|\mathbb{P}_{\boldsymbol{g}^\tau} - \mathbb{P}_{\boldsymbol{g}^{\tau'}}\|_{\mathrm{TV}}$, we use Pinsker's inequality because the Kullback–Leibler divergence $D(\mathbb{P}_{\boldsymbol{g}^\tau}\|\mathbb{P}_{\boldsymbol{g}^{\tau'}})$ has a simple expression in terms of $\ell^2(\boldsymbol{g}^\tau, \boldsymbol{g}^{\tau'})$:

$$\|\mathbb{P}_{\boldsymbol{g}^\tau} - \mathbb{P}_{\boldsymbol{g}^{\tau'}}\|^2_{\mathrm{TV}} \le \frac{1}{2}D(\mathbb{P}_{\boldsymbol{g}^\tau}\|\mathbb{P}_{\boldsymbol{g}^{\tau'}}) = \frac{n}{4\sigma^2}\ell^2(\boldsymbol{g}^\tau, \boldsymbol{g}^{\tau'}) = \frac{m_1m_2D}{36\sigma^2k^2}\Upsilon(\tau,\tau').$$

This gives

$$\min_{\Upsilon(\tau,\tau')=1}(1 - \|\mathbb{P}_{\boldsymbol{g}^\tau} - \mathbb{P}_{\boldsymbol{g}^{\tau'}}\|_{\mathrm{TV}}) \ge \left(1 - \frac{\sqrt{m_1m_2D}}{6k\sigma}\right).$$

Lemma A.1 then gives the lower bound for $\Delta := \inf_{\tilde{\boldsymbol{\theta}}}\sup_{\boldsymbol{\theta}\in\mathcal{M}:D(\boldsymbol{\theta})\le D} R(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}})$ as given below:

$$\Delta \ge \frac{m_1m_2D}{72n}\left(1 - \frac{\sqrt{m_1m_2D}}{6k\sigma}\right).$$

Because $k = \lfloor n_i/m_i \rfloor$ for $i = 1, 2$, it follows that $n_i/(k+1) \le m_i \le n_i/k$ for $i = 1, 2$. This gives

$$\Delta \ge \frac{D}{72(k+1)^2}\left(1 - \frac{\sqrt{nD}}{6\sigma k^2}\right) \ge \frac{D}{288k^2}\left(1 - \frac{\sqrt{nD}}{6\sigma k^2}\right),$$

where we have also used that $k+1 \le 2k$. The choice $k = (nD/(9\sigma^2))^{1/4}$ now leads to $\Delta \ge \sigma\sqrt{D}/(192\sqrt{n})$. This gives what we wanted to prove provided our choice of k satisfies $1 \le k \le \min(n_1, n_2)$. For this, it suffices to simply note that $n \ge 9\sigma^2/D$ implies that $k \ge 1$ and (2.3) implies $k \le \min(n_1, n_2)$. This completes the proof of Theorem 2.2.

A.5. Proof of Theorem 2.3

Proof. Fix $D > 0$. Also fix integers $1 \le k_1 \le n_1$ and $1 \le k_2 \le n_2$. Let m_1 and m_2 be defined so that $k_1 = \lfloor n_1/m_1 \rfloor$ and $k_2 = \lfloor n_2/m_2 \rfloor$. Let $d = k_1 k_2$. We denote elements of $\{0,1\}^d$ by $(\tau_{uv} : u, v \in [k_1] \times [k_2])$. For each such $\tau \in \{0,1\}^d$, we define $\boldsymbol{g}^\tau \in \mathcal{M}$ in the following way. For $i \in [n_1]$ and $j \in [n_2]$, if there exist $u, v \in [k_1] \times [k_2]$ for which $(u-1)m_1 < i \le um_1$ and $(v-1)m_2 < j \le vm_2$, we take

$$\boldsymbol{g}_{ij}^\tau = \sqrt{D}\left(\frac{u-1}{k_1} + \frac{v-1}{k_2} + \alpha\tau_{uv}\right),$$

where $\alpha = \max\{\frac{1}{k_1}, \frac{1}{k_2}\}$. Otherwise we take $\boldsymbol{g}_{ij}^\tau = \sqrt{D}$. One can check that $\boldsymbol{g}^\tau \in \mathcal{M}$ and $D(\boldsymbol{g}^\tau) \le D$ for every $\tau \in \{0,1\}^d$.

We shall now use Lemma A.1 with $d = k_1 k_2$ and this collection $\{\boldsymbol{g}^\tau : \tau \in \{-1,1\}^d\}$. Note first that

$$\begin{aligned}\ell^2(\boldsymbol{g}^\tau, \boldsymbol{g}^{\tau'}) &= \frac{1}{n}\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}(\boldsymbol{g}_{ij}^\tau - \boldsymbol{g}_{ij}^{\tau'})^2 \\ &= \frac{1}{n}\sum_{u,v\in[k]}\sum_{i:(u-1)m_1<i\le um_1}\sum_{j:(v-1)m_2<j\le vm_2}(\boldsymbol{g}_{ij}^\tau - \boldsymbol{g}_{ij}^{\tau'})^2 \\ &= \frac{m_1 m_2 D}{n}\alpha^2\Upsilon(\tau,\tau').\end{aligned}$$

Therefore, this implies that

$$\min_{\tau\ne\tau'}\frac{\ell^2(\boldsymbol{g}^\tau, \boldsymbol{g}^{\tau'})}{\Upsilon(\tau,\tau')} = \frac{m_1 m_2 D\alpha^2}{n}.$$

To bound $\|\mathbb{P}_{\boldsymbol{g}^\tau} - \mathbb{P}_{\boldsymbol{g}^{\tau'}}\|_{\mathrm{TV}}$, we again use Pinsker's inequality to obtain

$$\|\mathbb{P}_{\boldsymbol{g}^\tau} - \mathbb{P}_{\boldsymbol{g}^{\tau'}}\|_{\mathrm{TV}}^2 \le \frac{1}{2}D(\mathbb{P}_{\boldsymbol{g}^\tau}\|\mathbb{P}_{\boldsymbol{g}^{\tau'}}) = \frac{n}{4\sigma^2}\ell^2(\boldsymbol{g}^\tau, \boldsymbol{g}^{\tau'}) = \frac{m_1 m_2 D\alpha^2}{4\sigma^2}\Upsilon(\tau,\tau').$$

This gives

$$\min_{\Upsilon(\tau,\tau')=1}\left(1 - \|\mathbb{P}_{\boldsymbol{g}^\tau} - \mathbb{P}_{\boldsymbol{g}^{\tau'}}\|_{\mathrm{TV}}\right) \ge \left(1 - \alpha\frac{\sqrt{m_1 m_2 D}}{2\sigma}\right).$$

Lemma A.1 then gives the lower bound for $\Delta := \inf_{\tilde{\boldsymbol{\theta}}} \sup_{\boldsymbol{\theta} \in \mathcal{M}: D(\boldsymbol{\theta}) \le D} R(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}})$ as given below:

$$\Delta \ge \frac{k_1 k_2 m_1 m_2 \alpha^2 D}{8n} \left(1 - \alpha \frac{\sqrt{m_1 m_2 D}}{2\sigma}\right).$$

Because $k_i = \lfloor n_i/m_i \rfloor$ for $i = 1, 2$, it follows that $m_i \le n_i/k_i$ for $i = 1, 2$. This gives

$$\Delta \ge \frac{k_1 k_2 m_1 m_2 \alpha^2 D}{8n} \left(1 - \alpha \frac{\sqrt{n_1 n_2 D}}{2\sigma \sqrt{k_1 k_2}}\right).$$

Now we set $k_1 = n_1$ and $k_2 = \lceil (\frac{D}{16\sigma^2})^{1/3} n_2^{1/3} \rceil$. Note that for sufficiently large n, we will have $(m_2 k_2)/n_2 \ge 1/2$ and hence we will have

$$\Delta \ge \frac{\alpha^2 D}{16n} \left(1 - \alpha \frac{\sqrt{n_1 n_2 D}}{2\sigma \sqrt{k_1 k_2}}\right).$$

One can check that in the regime $n_1 \le \frac{(nD)^{1/4}}{(4\sigma)^{1/2}}$ the choice of k_1, k_2 satisfies the inequality $k_2 \ge k_1$. Recall that $\alpha = 1/\max\{k_1, k_2\}$ and hence in this regime we have $\alpha = \frac{1}{k_2}$. Then the choice of k_1, k_2 implies the following minimax lower bound

$$\Delta \ge \frac{D^{1/3} \sigma^{4/3}}{64 n_2^{2/3}}.$$

This completes the demonstration of the minimax lower bound. Now consider the estimator which just performs vector isotonic regression in each row. Since this estimator need not satisfy the matrix monotonicity constraints, we then project this estimator onto $\mathcal{M}$ to obtain the final estimator $\hat{\boldsymbol{\theta}}$. Clearly the projection step can only decrease the squared frobenius distance to $\theta^* \in \mathcal{M}$ by the Pythagorus identity for projections onto cones. Therefore, an upper bound to the risk of this estimator can be obtained by an application of (1.2) which finally gives us (2.4). This completes the proof of Theorem 2.3. □

Acknowledgements

We thank the anonymous reviewers for their careful reading and constructive comments, which helped us to improve the manuscript. The second author would like to thank Sivaraman Balakrishnan for helpful discussions. A. Guntuboyina supported by NSF Grant DMS-13-09356. B. Sen supported by NSF Grant DMS-11-50435.

References

[1] Ayer, M., Brunk, H.D., Ewing, G.M., Reid, W.T. and Silverman, E. (1955). An empirical distribution function for sampling with incomplete information. *Ann. Math. Stat.* **26** 641–647. MR0073895

[2] Barlow, R.E., Bartholomew, D.J., Bremner, J.M. and Brunk, H.D. (1972). *Statistical Inference Under Order Restrictions*. *The Theory and Application of Isotonic Regression*. New York: Wiley. MR0326887

[3] Bellec, P. (2015). Sharp oracle inequalities for least squares estimators in shape restricted regression. Preprint. Available at arXiv:1510.08029.

[4] Birgé, L. and Massart, P. (1993). Rates of convergence for minimum contrast estimators. *Probab. Theory Related Fields* **97** 113–150. MR1240719

[5] Brunk, H.D. (1955). Maximum likelihood estimates of monotone parameters. *Ann*. *Math*. *Stat*. **26** 607–616. MR0073894

[6] Chatterjee, S. (2014). A new perspective on least squares under convex constraint. *Ann*. *Statist*. **42** 2340–2381. MR3269982

[7] Chatterjee, S. (2015). Matrix estimation by universal singular value thresholding. *Ann*. *Statist*. **43** 177–214. MR3285604

[8] Chatterjee, S., Guntuboyina, A. and Sen, B. (2015). On risk bounds in isotonic and other shape restricted regression problems. *Ann*. *Statist*. **43** 1774–1800. MR3357878

[9] Chatterjee, S., Guntuboyina, A. and Sen, B. (2015). On matrix estimation under monotonicity constraints. Available at http://arxiv.org/abs/1506.03430.

[10] Chatterjee, S. and Mukherjee, S. (2016). On estimation in tournaments and graphs under monotonicity constraints. Preprint. Available at arXiv:1603.04556.

[11] Donoho, D. (1991). Gelfand n-widths and the method of least squares Technical report, University of California, Berkeley.

[12] Gao, F. and Wellner, J.A. (2007). Entropy estimate for high-dimensional monotonic functions. *J*. *Multivariate Anal*. **98** 1751–1764. MR2392431

[13] Gebhardt, F. (1970). An algorithm for monotone regression with one or more independent variables. *Biometrika* **57** 263–271.

[14] Hanson, D.L., Pledger, G. and Wright, F.T. (1973). On consistency in monotonic regression. *Ann*. *Statist*. **1** 401–421. MR0353540

[15] Makowski, G.G. (1977). Consistency of an estimator of doubly nondecreasing regression functions. *Z*. *Wahrsch*. *Verw*. *Gebiete* **39** 263–268. MR0652723

[16] Meyer, M. and Woodroofe, M. (2000). On the degrees of freedom in shape-restricted regression. *Ann*. *Statist*. **28** 1083–1104. MR1810920

[17] Oymak, S. and Hassibi, B. (2016). Sharp MSE Bounds for Proximal Denoising. *Found*. *Comput*. *Math*. **16** 965–1029. MR3529131

[18] Robertson, T. and Wright, F.T. (1975). Consistency in generalized isotonic regression. *Ann*. *Statist*. **3** 350–362. MR0365871

[19] Robertson, T., Wright, F.T. and Dykstra, R.L. (1988). *Order Restricted Statistical Inference*. *Wiley Series in Probability and Mathematical Statistics*: *Probability and Mathematical Statistics*. Chichester: Wiley. MR0961262

[20] Shah, N.B., Balakrishnan, S., Guntuboyina, A. and Wainright, M.J. (2015). Stochastically transitive models for pairwise comparisons: Statistical and computational issues. Preprint. Available at arXiv:1510.05610.

[21] van Eeden, C. (1958). *Testing and Estimating Ordered Parameters of Probability Distributions* Amsterdam: Mathematical Centre. MR0102874

[22] van de Geer, S. (1990). Estimating a regression function. *Ann*. *Statist*. **18** 907–924. MR1056343

[23] van de Geer, S. (1993). Hellinger-consistency of certain nonparametric maximum likelihood estimators. *Ann*. *Statist*. **21** 14–44. MR1212164

[24] van de Geer, S. and Wainwright, M. (2015). On concentration for (regularized) empirical risk minimization. Preprint. Available at arXiv:1512.00677.

[25] van de Geer, S.A. (2000). *Applications of Empirical Process Theory*. *Cambridge Series in Statistical and Probabilistic Mathematics* **6**. Cambridge: Cambridge Univ. Press. MR1739079

[26] van der Vaart, A.W. (1998). *Asymptotic Statistics*. *Cambridge Series in Statistical and Probabilistic Mathematics* **3**. Cambridge: Cambridge Univ. Press. MR1652247

[27] van der Vaart, A.W. and Wellner, J.A. (1996). *Weak Convergence and Empirical Processes*: *With Applications to Statistics*. *Springer Series in Statistics*. New York: Springer. MR1385671

[28] Wang, Y. (1996). The L_2 risk of an isotonic estimate. *Comm*. *Statist*. *Theory Methods* **25** 281–294. MR1379445

[29] Zhang, C.-H. (2002). Risk bounds in isotonic regression. *Ann*. *Statist*. **30** 528–555. MR1902898

Received November 2015 and revised April 2016