Optimal rates for ranking a permuted isotonic matrix in polynomial time

Emmanuel Pilliat, Alexandra Carpentier, Nicolas Verzelen

Optimal rates for ranking a permuted isotonic matrix in polynomial time

Emmanuel Pilliat, Alexandra Carpentier, and Nicolas Verzelen

Abstract: We consider a ranking problem where we have noisy observations from a matrix with isotonic columns whose rows have been permuted by some permutation π^*. This encompasses many models, including crowd-labeling and ranking in tournaments by pair-wise comparisons. In this work, we provide an optimal and polynomial-time procedure for recovering π^*, settling an open problem in [7]. As a byproduct, our procedure is used to improve the state-of-the art for ranking problems in the stochastically transitive model (SST). Our approach is based on iterative pairwise comparisons by suitable data-driven weighted means of the columns. These weights are built using a combination of spectral methods with new dimension-reduction techniques. In order to deal with the important case of missing data, we establish a new concentration inequality for sparse and centered rectangular Wishart-type matrices.

1. Introduction

Ranking problems have recently spurred a lot of interest in the statistical and computer science literature. This includes a variety of problems ranging from ranking experts/workers in crowd-sourced data, ranking players in a tournament or equivalently sorting objects based on pairwise comparisons.

To fix ideas, let us consider a problem where we have noisy partial observations from an unknown matrix $M \in [0,1]^{n\times d}$. In crowdsourcing problems, n stands for the number of experts (or workers), d stands for the number of questions (or tasks) and $M_{i,k}$ for the probability that expert i answers question k correctly. For tournament problems, we have $n = d$ players (or objects) and $M_{i,k}$ stands for the probability that player i wins against player k. Based on these noisy data, the general goal is to provide a full ranking of the experts or of the players.

Originally, these problems were tackled using parametric model for the matrix M. Notably, this includes the noisy sorting model [5] or Bradley-Luce-Terry model [4]. Still, it has been observed that these simple models are often unrealistic and do not tend to fit well.

This has spurred a recent line of literature where strong parametric assumptions are replaced by non-parametric assumptions [17, 18, 19, 20, 10, 9, 8, 7, 3, 16]. In particular, for tournament problems, the strong stochastically transitive (SST) model presumes that the square matrix M is, up to a common permutation π^* of the rows and of the columns, bi-isotonic and satisfies the skew symmetry condition $M_{i,k} + M_{k,i} = 1$. Although optimal rates for estimation of the permutation π^* have been pinpointed in the earlier paper of Shah et al. [18], there remains a large gap between these optimal rates and the best known performances of polynomial-time algorithms. This has led to conjecture the existence of a statistical-computational gap [10, 8].

For crowdsourcing data, the counterpart of the SST model is the so-called bi-isotonic model, where the rectangular matrix M is bi-isotonic, up to an unknown permutation π^* of its rows and an unknown permutation η^* of its columns. This model turns out to be really similar to the SST model and the existence of a statistical-computational gap has also been conjectured [10].

In this paper, we tackle a slightly different route and we consider the arguably more general isotonic model [7]. The only assumption is that all the columns of M are nondecreasing up to an unknown permutation of the rows, making the isotonic model more flexible than the bi-isotonic and SST models. It is in fact the most general model under which an unambiguous ranking

of the experts is well-defined. In this model as well, there is a gap between the (statistical) optimal rates, and the rate obtained by the (polynomial-time) algorithm in [7].

Our main contributions are as follows. For the isotonic model, we establish the optimal rate for recovering the permutation, and we introduce a polynomial-time procedure achieving this rate, thereby settling the absence of any computational gap in this model. Besides, our procedure and results have important consequences when applied to the SST and bi-isotonic model. More specifically, we achieve the best known guarantees in these two models [8, 9] and even improve them in some regimes.

1.1. Problem formulation

Let us further introduce our model. A bounded matrix $A \in [0,1]^{n\times d}$ is said to be isotonic if its columns are nondecreasing, that is $A_{i,k} \le A_{i+1,k}$ for any $i \in [n-1]$ and $k \in [d]$. Henceforth, we write $\mathbb{C}_{\text{iso}}$ for the collection of all $n \times d$ isotonic matrices taking values in $[0,1]$. In our model, we recall that we assume that the signal matrix M is isotonic up to an unknown permutation of its rows. In other words, there exists a permutation π^* of $[n]$ such that the matrix $M_{\pi^{*-1}}$ defined by $(M_{\pi^{*-1}})_{i,k} = (M_{\pi^{*-1}(i),k})$ has nondecreasing columns, that is

$$M_{\pi^{*-1}(i),k} \le M_{\pi^{*-1}(i+1),k} \ , \tag{1}$$

for any $i \in \{1,\dots,n-1\}$ and $k \in \{1,\dots,d\}$, or equivalently $M_{\pi^{*-1}} \in \mathbb{C}_{\text{iso}}$. Henceforth, π^* is called an oracle permutation. Using the terminology of crowdsourcing, we refer to i^{th} row of M as expert i and to k^{th} column as question k.

In this work, we have N partial and noisy observations of the matrix M of the form (x_t, y_t) where

$$y_t = M_{x_t} + \varepsilon_t \quad t = 1,\dots,N \ . \tag{2}$$

For each t, the position $x_t \in [n] \times [d]$ is sampled uniformly. The noise variables ε_t's are independent and their distributions only depend on the position x_t. We only assume that all these distributions are centered and are subGaussian with a subGaussian norm of at most 1 – see e.g. [23]. In particular, this encompasses the typical case where the y_t's follow Bernoulli distributions with parameters M_{x_t}.

As usual in the literature e.g. [14, 8, 10], we use, for technical convenience, the Poissonization trick which amounts to assuming that the number N of observations has been sampled according to a Poisson distribution with parameter λnd. We refer to $\lambda > 0$ as the sampling effort. When $\lambda > 1$, we have, in expectation, several independent observations per entry (i,j) - and $\lambda = 1$ means that there is on average one observation per entry. In this paper, we are especially interested in the sparse case where λ is much smaller than one, i.e. the case where we have missing observations for some entries. We refer to $\lambda = 1$ as the full observation regime at it bears some similarity to the case often considered in the literature –e.g. [18, 7], where we have a full observation of the matrix,

$$Y = M + E' \in \mathbb{R}^{n\times d} \ . \tag{3}$$

The entries of the noise matrix E' are independent, centered, and 1-subGaussian.

In this work, we are primarily interested in estimating the permutation π^*. Given an estimator $\hat{\pi}$, we use the square Frobenius norm $\|M_{\hat{\pi}^{-1}} - M_{\pi^{*-1}}\|_F^2$ as the loss. This loss quantifies the distance between the matrix M reordered according to the estimator $\hat{\pi}$ and the matrix M sorted according to the oracle permutation π^*. This loss is explicitly used in [8, 14] and is implicit in earlier works –see e.g. [18].

We define the associated optimal risk of permutation recovery as a function of the number n of experts, the number d of question and the sampling effort λ,

$$\mathcal{R}^*_{\text{perm}}(n,d,\lambda) = \inf_{\hat{\pi}} \sup_{\substack{\pi^* \in \Pi_n \\ M : M_{\pi^{*-1}} \in \mathbb{C}_{\text{iso}}}} \mathbb{E}_{(\pi^*, M)}\big[\|M_{\hat{\pi}^{-1}} - M_{\pi^{*-1}}\|_F^2\big] \ , \tag{4}$$

where the infimum is taken over all estimators. Here, Π_n stands for the collection of all permutations of $[n]$. If the main focus is not only to estimate π^*, but also to reconstruct the unknown matrix M, we also consider the optimal reconstruction rate

$$\mathcal{R}^*_{\text{reco}}(n,d,\lambda) = \inf_{\hat{M}} \sup_{\substack{\pi^* \in \Pi_n \\ M : M_{\pi^{*-1}} \in \mathbb{C}_{\text{iso}}}} \mathbb{E}\big[\|\hat{M} - M\|_F^2\big] \quad . \tag{5}$$

It turns out that reconstructing the matrix M is more challenging than estimating the permutation π^*. Considering both risks allows to disentangle the reconstruction of the matrix M: looking at both enables to distinguish the error that is due to estimating the permutation, from the error that comes from estimating an isotonic matrix.

1.2. Past results on the isotonic model and our contributions

In, the specific case where $d = 1$ (a single column), our model is equivalent to uncoupled isotonic regression and is motivated by optimal transport. Rigollet and Niles-Weed [15] have established that the reconstruction error of M is of the order of $n(\frac{\log\log(n)}{\log(n)})^2$.

For the general case $d \geq 1$, Flammarion et al. [7] have shown[1] that the optimal reconstruction error in the full observation model (3) is of the order of $n^{1/3}d + n$. However, the corresponding procedure is not efficient. They also introduce an efficient procedure that first estimates π^* using a score based on row comparisons on Y. Unfortunately, this method only achieves a reconstruction error of the order of $n^{1/3}d + n\sqrt{d}$ which is significantly slower than the optimal one. Whether or not there is a statistical-computationnal gap was therefore an open problem.

We prove in this work that there is no computational statistical gap in this model. More precisely, we introduce estimators that are both polynomial-time and minimax optimal up to some polylog factors. To that end, we characterize the optimal risks $\mathcal{R}^*_{\text{perm}}(n,d,\lambda)$ and $\mathcal{R}^*_{\text{reco}}(n,d,\lambda)$ of permutation estimation and matrix reconstruction, for all possible number of experts $n \geq 1$, number of questions $d \geq 1$ and all sampling efforts λ, up to some polylog factors in nd. Table 1 summarizes our findings in the arguably most interesting cases[2] $\lambda \in [1/(n \wedge d), 1]$.

	$n \leq d^{3/2}\sqrt{\lambda}$	$d^{3/2}\sqrt{\lambda} \leq n$
$\mathcal{R}^*_{\text{perm}}$	$n^{2/3}\sqrt{d}\lambda^{-5/6}$	n/λ
$\mathcal{R}^*_{\text{reco}}$	$n^{1/3}d\lambda^{-2/3}$	n/λ

,

TABLE 1

Optimal rates in our model, for all possible values of n, d and $\lambda \in [1/(n \wedge d), 1]$, up to a polylogarithmic factor in nd. These rates are achieved by polynomial-time estimators.

[1] The authors consider the isotonic model as a subcase of a seriation model, where each columns of $M_{\pi^{*-1}}$ is only assumed to be unimodal.

[2] We are indeed mostly interested in the more realistic sparse observation regime (meaning $\lambda \leq 1$). The case $\lambda \leq 1/d$ leads to the trivial minimax bound of order nd for both reconstruction and estimation, as in this case we have less than one observations per expert on average. As for the case $\lambda > 1/d$ but $\lambda \leq 1/n$, we have less than one observation per question on average, and this leads to a minimax risk of order $n\sqrt{d/\lambda}$ for permutation estimation and of order nd for matrix recontruction.

1.3. Implication for other models and connection to the literature

As discussed earlier, the isotonic model is quite general and encompasses both the bi-isotonic model for crowdsourcing problems as well the SST model for tournament problems.

Let us first focus on the SST model which corresponds to the case where $n = d$ together with a bi-isotonicity and a skew-symmetry assumption. In the full observation scheme (related to the case $\lambda = 1$) where one observes the noisy matrix $n \times n$, Shah et al. [18] have established that the optimal rates for estimating π^* and reconstructing the matrix M are of the order of n. In contrast, their efficient procedure which estimates π^* according to the row sums of Y only achieves the rate of $n^{3/2}$. In more recent years, there has been a lot of effort dedicated to improving this $\sqrt{n}$ statistical-computational gap. The SST model was also generalized to partial observations by [6], which corresponds to $\lambda \leq 1$. They introduced an efficient procedure that targets a specific sub-class of the SST model, and that achieves a rate of order $n^{3/2}\lambda^{-1/2}$ in the worst case for matrix reconstruction.

Recently, a few important contributions tackling both the bi-isotonic model and the SST model made important steps towards better understanding the statistical-computational gap. We first explain how their results translate in the SST model. Mao et al. [10, 9] introduced a polynomial-time procedure handling partial observation, achieving a rate of order $n^{5/4}\lambda^{-3/4}$ for matrix reconstruction. Nonetheless, [10] failed to exploit global information shared between the players/experts – as they only compare players/experts two by two – as pointed out by [8]. Building upon this remark, [8] managed to get the better rate $n^{7/6+o(1)}$ with a polynomial-time method in the case $\lambda = n^{o(1)}$.

Let us turn to the more general bi-isotonic model. Here, the rectangular matrix $M \in \mathbb{R}^{n\times d}$ is bi-isotonic up an unknown permutation π^* of the rows and an unknown permutation η^* of the columns. Since M is not necessarily square, this model can be used in more general crowdsourcing problems. The optimal rate for reconstruction in this model with partial observation has been established in [10] to be of order $\nu(n, d, \lambda) := (n \vee d)/\lambda + \sqrt{nd/\lambda} \wedge n^{1/3}d\lambda^{-2/3} \wedge d^{1/3}n\lambda^{-2/3}$ up to polylog factors, in the non-trivial regime where $\lambda \in [1/(n \wedge d), 1]$. However, the polynomial-time estimator provided by Mao et al. [10] only achieves the rate $n^{5/4}\lambda^{-3/4} + \nu(\lambda, n, d)$. In a nutshell, Mao et al. first compute column sums to give a first estimator of the permutation of the questions. Then, they compare the experts on aggregated blocks of questions, and finally compare the questions on aggregated blocks of experts. As explained in the previous paragraph for SST models, Liu and Moitra [8] improved this rate to $n^{7/6+o(1)}$ in the square case ($n = d$), with a subpolynomial number of observations per entry ($\lambda = n^{o(1)}$). Their estimators of the permutations π^*, η^* were based on hierarchical clustering and on local aggregation of high variation areas. Both [8, 10] made heavily use of the bi-isotonicity structure of M by alternatively sorting the columns and rows. As mentioned for the SST model, the order of magnitude $n^{7/6+o(1)}$ remains nevertheless suboptimal, and whether there exists an efficient algorithm achieving the optimal rate in this bi-isotonic model remains an open problem.

We now discuss the implications of our work concerning the bi-isotonic model and SST model. First, in the full observation setting ($\lambda = 1$) and square case for the bi-isotonic model ($n = d$), we reach in polynomial-time the upper bound $n^{7/6}$ up to polylog factors, for both permutation estimation and matrix reconstruction. In particular, we improve the rate in [8] by a subpolynomial factor in n, and we do not need a subpolynomial number of observation per entry. Moreover, our procedure being primarily designed for the isotonic model, it does not require any shape constraint on the rows in contrast to [8, 10]. Beyond the full observation regimes, we provide guarantees on our estimator of π^* for different values of λ. In particular, in Corollary 2.5, we derive an estimator of the matrix M that achieves a maximum reconstruction risk $\sup_{\pi^*,\eta^*,M} \mathbb{E}\left[\|\hat{M} - M_{\pi^{*-1}\eta^{*-1}}\|_F^2\right]$ of order less than $n^{7/6}\lambda^{-5/6}$ up to polylogs, thereby

improving the state-of-the-art polynomial-time methods in partial observation [10]. Lastly, we perform our analysis in the general rectangular case, giving guarantees for general values of d.

The optimal risks and the known polynomial-time upper bounds for the isotonic, bi-isotonic with two permutations and SST models are summarized in Table 2. For the sake of simplicity, we focus in the table to the specific case case $n = d$ and $\lambda \in [1/n, 1]$.

Different models, with $M \in \mathbb{R}^{n\times n}$		**Isotonic** $M_{\pi^{*-1}}$ **has nondecreasing columns**	Bi-isotonic(π^*, η^*) $M_{\pi^{*-1}\eta^{*-1}}$ has nondecreasing columns and rows	SST $M_{\pi^{*-1}\pi^{*-1}}$ has nondecreasing columns and rows, and $M_{ik} + M_{ki} = 1$
Permutation estimation	Poly. Time	$n^{7/6}\lambda^{-5/6}$ **[Th 2.2]**	$n^{7/6+o(1)}$ [8]($\lambda = n^{o(1)}$) $n^{7/6}\lambda^{-5/6}$ **[Th 2.2]**	$n^{7/6+o(1)}$ [8]($\lambda = n^{o(1)}$) $n^{7/6}\lambda^{-5/6}$ **[Th 2.2]**
	optimal rate	$n^{7/6}\lambda^{-5/6}$ **[Th 2.1]**	n/λ [10]	n/λ [10]
Matrix reconstruction	Poly. Time	$n^{3/2}$ ($\lambda = 1$)[7] $n^{4/3}\lambda^{-2/3}$ **[Cor 2.5]**	$n^{7/6+o(1)}$ [8]($\lambda = n^{o(1)}$) $n^{5/4}\lambda^{-3/4}$ [10] $n^{7/6}\lambda^{-5/6}$ **[Cor 2.5]**	$n^{7/6+o(1)}$ [8]($\lambda = n^{o(1)}$) $n^{5/4}\lambda^{-3/4}$ [10] $n^{7/6}\lambda^{-5/6}$ **[Cor 2.5]**
	optimal rate	$n^{4/3}\lambda^{-2/3}$ [7] (also [Prop 2.3])	n/λ [10]	n/λ [10]

TABLE 2

For the isotonic model, the optimal rate for permutation estimation (resp. matrix reconstruction) corresponds to $\mathcal{R}^*_{\text{perm}}$ (resp. $\mathcal{R}^*_{\text{reco}}$). For the two other columns, the optimal rates are similarly defined as minimax risk over the corresponding models. The Poly. Time rows correspond to state-of-the art rates achieved by polynomial-time methods. All the rates are given up to polylogarithmic factors in n.

Finally, we mention the even more specific model where the matrix M is bi-isotonic up to a single permutation π^* acting on the rows. This corresponds to the case where η^* is known in the previous paragraph [10, 14, 8]. Equivalently, this also corresponds to our isotonic model (2) with the additional assumption that all the rows are nondecreasing, that is $M_{i,k} \le M_{i,k+1}$. For this model, it is possible to leverage the shape constrains on the rows to build efficient and optimal estimators, this for all n, d, and λ – see [14].

1.4. Overview of our techniques

In this work, we introduce the iterative soft ranking (**ISR**) procedure, which gives an estimator $\hat{\pi}$ based on the observations. Informally, this method iteratively updates a weighted directed graph between experts, where the weight between any two experts quantifies the significance of their comparison. The procedure increases the weights at each step. After it stops, the final estimator is an arbitrary permutation $\hat{\pi}$ that agrees as well as possible with the final weighted directed graph.

As mentioned in [8], it is hopeless to use only local information between pairs of experts to obtain a rate of order $n^{7/6}$ up to polylogs, and we must exploit global information. Still, we do it in a completely different way of Liu and Moitra [8] who were building upon the bi-isotonicity of the matrix.

One first main ingredient of our procedure is a new dimension reduction technique. At a high level, suppose that we have partially ranked the rows in such a way that, for a given triplet (P, O, I) of subsets of $[n]$, we are already quite confident that experts in P are below those in I and above those in O. Relying on the shape constraint of the matrix M, it is therefore possible to build a high-probability confidence regions for rows in P based on the rows in O and the rows in I. If, for a question j, the confidence region is really narrow, this implies that all experts in P take almost the same value on this column. As a consequence,

this question is almost irrelevant for further comparing the experts in P. In summary, our dimension reduction technique selects the set of questions for which the confidence region of P is wide enough, and in that way reduces the dimension of the problem while keeping most of the relevant information.

The second main ingredient, once the dimension is reduced, is to use a spectral method to capture some global information shared between experts. That is why our procedure makes significant use of spectral methods to compute the updates of the weighted graph. Although this spectral scheme already appears in recent works [14, 8], those are used here for updating the weight of the comparison graph rather than performing a clustering as in [8]. Moreover, the analysis of the spectral step in the partial observation regime ($\lambda \ll 1$) leads to technical difficulties – see the discussion in Section 3.5.

Related to the latter problem, we need to establish a new tail bound on sparse rectangular matrices. More specifically, for a rectangular matrix X with centered independent entries that satisfy a Bernstein type condition, we provide a high-probability control of the operator norm of $XX^T - \mathbb{E}[XX^T]$. This result, based on non-commutative matrix Bernstein concentration inequality, may be of independent interest e.g. for controlling the spectral properties of a sparse bipartite random graph. We state it in Section 4, independently of the rest of the paper.

1.5. Notation

Given a vector u and $p \in [1,\infty]$, we write $\|u\|_p$ for its l_p norm. For a matrix A, $\|A\|_F$ and $\|A\|_{\mathrm{op}}$ stand for its Frobenius and its operator norm. We write $\lfloor x \rfloor$ (resp. $\lceil x \rceil$) for the largest (resp. smallest) integer smaller than (resp. larger than) or equal to x. Although M stands for an $n \times d$ matrix, we extend it sometimes in an infinite matrix defined for all $i \in \mathbb{N}, k \in \{1,\dots,d\}$ by setting $M_{ik} = 0$ when $i \le 0$ and $M_{ik} = 1$ when $i \ge n+1$. The corresponding infinite matrix $M_{\pi^{*(-1)}}$ which is obtained by permuting the n original rows is still isotonic and takes values in $[0,1]$. We shall often work with submatrices $M(P,Q)$ of M that are restricted to a subset $P \subset [n]$ and $Q \subset [d]$ of rows and columns. If A is any matrix in $\mathbb{R}^{P\times Q}$, we write $\overline{A}$ for the matrix whose rows are all equal to the average row of A, namely $\overline{A}_{ik} = \frac{1}{|P|}\sum_{j\in P} A_{jk}$.

2. Results

In this section, we first establish the statistical limit with a lower bound on $\mathcal{R}^*_{\mathrm{perm}}(n,d,\lambda)$. Then, we state the existence of a polynomial-time estimator that is minimax optimal up to polylog factors. More precisely, we prove that for all integers n, d and $\lambda \in [1/d, 8n^2]$, the optimal rate of permutation estimation $\mathcal{R}^*_{\mathrm{perm}}$ is of the order of

$$\rho_{\mathrm{perm}}(n,d,\lambda) := \frac{n^{2/3}\sqrt{d}}{\lambda^{5/6}} \bigwedge n\sqrt{\frac{d}{\lambda}} + \frac{n}{\lambda} \ , \tag{6}$$

up to some polylog factors. As a corollary, we then establish that the optimal rate of matrix reconstruction $\mathcal{R}^*_{\mathrm{reco}}$ is of order

$$\rho_{\mathrm{reco}}(n,d,\lambda) := \frac{n^{1/3}d}{\lambda^{2/3}} + \frac{n}{\lambda} \ , \tag{7}$$

up to polylog factors. We therefore establish that these two problems do not exhibit a computational-statistical gap.

2.1. Minimax lower bound for permutation estimation

Assume that $\lambda \in [1/d, 8n^2]$ is fixed and that we are given $N = Poi(\lambda nd)$ independent observations under model (2). Namely, we observe $(x_t, y_t)_{t=1,\dots,N}$ where x_t is sampled uniformly in $[n] \times [d]$ and $y_t = M_{x_t} + \varepsilon_t$ conditionally to x_t. The following theorem states that ρ_{perm} is a lower bound on the maximum risk of permutation estimation for all $n, d, \lambda \in [1/d, 8n^2]$, up to some numerical constant.

Theorem 2.1. *There exists a universal constant $c > 0$ such that, for any $n \geq 2$, $d \geq 1$, and $\lambda \in [1/d, 8n^2]$, we have*

$$\mathcal{R}^*_{\text{perm}}(n, d, \lambda) \geq c\rho_{\text{perm}}(n, d, \lambda) \ . \tag{8}$$

In the proof, we show a slightly stronger result that also covers the cases $\lambda < 1/d$ and $\lambda > 8n^2$, where $\mathcal{R}^*_{\text{perm}}(n, d, \lambda)$ is in fact respectively lower bounded by a quantity of order nd and $n\sqrt{d}/\lambda$. For the sake of readability, we chose to omit these arguably less interesting cases in the statement of Theorem 2.1 and of Theorem 2.2.

2.2. Optimal permutation estimation

Let us fix a quantity $\delta \in (0, 1)$ that will correspond to a small probability. We need to introduce some notation. We write

$$\phi_{\text{L}_1} = 10^4 \log\left(\frac{10^2 nd}{\delta}\right) \ . \tag{9}$$

Our procedure depends on a sequence of tuning parameters. For this reason, we introduce a subset $\Gamma \subset \mathbb{R}^+$, henceforth called a grid. The grid Γ is said to be valid if it contains a sequence $\gamma_0 \geq \cdots \geq \gamma_{2\lfloor \log_2(n) \rfloor + 2}$ of length $2\lfloor \log_2(n) \rfloor + 3$ such that that for all u,

$$\gamma_u - \gamma_{u+1} \geq \gamma_{2\lfloor \log_2(n) \rfloor + 2} + \phi_{\text{L}_1} \quad \text{and} \quad \gamma_{2\lfloor \log_2(n) \rfloor + 2} \geq \phi_{\text{L}_1} \ . \tag{10}$$

In light of this definition, we could simply choose the valid sequence $\Gamma = \{\phi_{\text{L}_1}, 2\phi_{\text{L}_1}, \dots, (2\lfloor \log_2(n) \rfloor + 3)\phi_{\text{L}_1}\}$ with a corresponding γ_0 that is polylogarithmic. Still, for practical purpose, we consider general grids; examples of such gris are discussed in more details in Section 3.6.

For any valid subset Γ, we define $\bar{\gamma}$ as the smallest possible value of γ_0 over all sequences that satisfy (10).

$$\bar{\gamma} = \min\{\gamma \ : \ \exists(\gamma_u) \text{ satisfying (10) s.t. } \gamma_0 = \gamma\} \ . \tag{11}$$

Our main procedure **ISR**, for iterative soft ranking, will be described in detail in Section 3. The only tuning parameters are the the number of steps T and the valid grid Γ.

Theorem 2.2. *There exists $C > 0$ such that the following holds. Let $\lambda \in [1/d, 8n^2]$ and $\delta > 0$. Assume that Γ is a valid grid and that $T \geq 4\bar{\gamma}^6$ with $\bar{\gamma}$ defined in (11). For any permutation $\pi^* \in \Pi_n$ and any matrix M such that $M_{\pi^{*-1}} \in \mathbb{C}_{\text{iso}}$, the estimator $\hat{\pi}$ from Algorithm* **ISR**(T, Γ) *defined in the next section satisfies*

$$\|M_{\hat{\pi}^{-1}} - M_{\pi^{*-1}}\|_F^2 \leq CT\bar{\gamma}^6 \rho_{\text{perm}}(n, d, \lambda) \ ,$$

with probability at least $1 - 10T\delta$.

In particular, if we suitably choose Γ (as discussed above) and $T = 4\lceil \bar{\gamma}^6 \rceil$ and $\delta = 1/(nd)^2$, we deduce from Theorem 2.2 that

$$\mathcal{R}^*_{\text{perm}}(n, d, \lambda) \leq C' \log^{C'}(nd) \rho_{\text{perm}}(n, d, \lambda) \ ,$$

for some numerical constant $C' > 0$. In the case where $\lambda = n^{o(1)}$ and $n = d$, this bound achieves the order of magnitude $n^{7/6}$, which aligns with the result presented in Theorem 2 of Liu and Moitra [8]. However, it is important to note that the analysis made in [8] focuses on the statistically easier bi-isotonic model, and their procedure heavily relies on the isotonicity structure imposed on the questions.

2.3. Optimal reconstruction of the matrix M

We now turn to the problem of estimating the signal matrix M. Obviously, the reconstruction of the matrix M from the observation of model in(2) is at least as hard as if we knew the permutation π^*. In this favorable situation, estimating M amounts to estimating d isotonic vectors from partial and noisy observations $Y_{ik} = \frac{1}{\lambda}\sum_t y_t \mathbf{1}_{x_t=(ik)}$. The isotonic regression problem is already well understood, and we state the following lower bound without proof since it directly follows from [10] (see in particular Theorem 3.1 therein). We recall that $\rho_{\mathrm{reco}}(n,d,\lambda)$ is defined in (7).

Proposition 2.3. *There exists a universal constant $c > 0$ such that, for any $n \geq 2$, any $d \geq 1$, and any $\lambda > 0$, we have*

$$\mathcal{R}^*_{\mathrm{reco}}(n,d,\lambda) \geq c\rho_{\mathrm{reco}}(n,d,\lambda) \ . \tag{12}$$

In particular, since $\rho_{\mathrm{perm}}(n,d,\lambda) \ll \rho_{\mathrm{reco}}(n,d,\lambda)$ in many regimes in n, d, λ, this proposition implies that the reconstruction of a permuted isotonic matrix is harder than the estimation of the permutation, namely that $\mathcal{R}^*_{\mathrm{perm}} \ll \mathcal{R}^*_{\mathrm{reco}}$.

To build an optimal estimator of M, we compute the estimated permutation $\hat{\pi}$ of Theorem 2.2 and estimate an isotononic matrix based on this ordering. This approach is similar to what is done in [10, 14], for related problems where a bi-isotonic assumption is done. For simplicity, set the tuning parameters T, Γ for Algorithm 1 so that $T = 4\lceil\bar{\gamma}^6\rceil$ and $\bar{\gamma}^6 \leq C' \log^{C'}(nd/\delta)$. We split the samples y_t defined in (2) into two independent sequences of samples $(y_t^{(1)})$, $(y_t^{(2)})$. First, we compute the estimator $\hat{\pi}$ of π^* with the first sub-samples $(y_t^{(1)})$. Then, we define $\hat{M}_{\mathrm{iso}}$ as the projection of $Y^{(2)}_{\hat{\pi}}$ onto the convex set of isotonic matrices, where $Y^{(2)}$ is the matrix defined by $Y^{(2)}_{ik} = \frac{1}{\lambda}\sum_t y_t^{(2)} \mathbf{1}_{x_t^{(2)}=(i,k)}$. More precisely, set

$$\hat{M}_{\mathrm{iso}} = \operatorname*{arg\,min}_{\tilde{M}\in\mathbb{C}_{\mathrm{iso}}} \|\tilde{M} - Y^{(2)}_{\hat{\pi}^{-1}}\|_2^2 \ .$$

The following corollary controls the risk of $\hat{M}_{\mathrm{iso}}$.

Corollary 2.4. *Assume that $\lambda \in [1/d, 8n^2]$. There exists a universal constant C'' such that the following holds for any permutation $\pi^* \in \Pi_n$ and any matrix $M \in \mathbb{C}_{\mathrm{iso}}$.*

$$\mathbb{E}[\|(\hat{M}_{\mathrm{iso}})_{\hat{\pi}} - M\|_F^2] \leq C'' \log^{C''}(nd)\rho_{\mathrm{reco}}(n,d,\lambda) \ .$$

As a consequence, the polynomial-time estimator $\hat{M}_{\mathrm{iso}}$ achieves the optimal risk for all values of n and d. For $\lambda = 1$, the optimal risk $\rho_{\mathrm{reco}}(n,d,1)$ is of the order of $n^{1/3}d + n$. In particular, our risk bound strictly improves over the one of Flammarion et al. [7] - e.g. their procedure achieves the estimation error $n\sqrt{d}$ for $n \geq d^{1/3}$. Their slower convergence rates are mainly due to the fact that their estimator of the permutation π^* is suboptimal in this regime.

2.4. Polynomial-time reconstruction in the bi-isotonic model

We now turn our attention to the problem of estimating the matrix M when M satisfies the additional assumption of being bi-isotonic up to unknown permutations π^* and η^* of its rows and columns respectively. In other words, the matrix $M_{\pi^{*-1}\eta^{*-1}}$ has non-decreasing entries. As explained in the introduction, this model has attracted a lot of attention in the last decade and encompasses the SST model for tournament problems.

To simplify the exposition, we focus in this section on the case $n = d$ and $\lambda \in [\frac{1}{n}, 1]$. Since the bi-isotonic model is a specific case of the isotonic model, we could rely on the estimator $\widehat{M}_{\text{iso}}$ introduced in the previous subsection. In fact, we can improve this estimation rate by relying on the bi-isotonicity of the matrix $M_{\pi^{*-1}\eta^{*-1}}$.

As previously, we choose the tuning parameters of Algorithm 1 in such a way that $T = 4\lceil\bar{\gamma}^6\rceil$ and $\bar{\gamma}^6 \le C' \log^{C'}(nd/\delta)$. Then, we use the following procedure:

1. Subsample the data into 3 independent samples $(y_t^{(1)})$, $(y_t^{(2)})$, $(y_t^{(3)})$.
2. Run our procedure Algorithm 1 to obtain an estimator $\hat{\pi}$ of the permutation π^* of the rows, using the first sample.
3. Run again Algorithm 1 to obtain an estimator $\hat{\eta}$ of the permutation η^* of the columns, using the second sample.
4. Compute the least-square estimator $\hat{M}_{\text{biso}} = \arg\min_{\tilde{M}\in\mathbb{C}_{\text{biso}}} \|\tilde{M} - Y^{(3)}_{\hat{\pi}^{-1}\hat{\eta}^{-1}}\|_2^2$, where $\mathbb{C}_{\text{biso}}$ is the set of all bi-isotonic matrices with entries in $[0,1]$ and $Y^{(3)}_{ik} = \frac{1}{\lambda}\sum_t y_t^{(3)} \mathbf{1}_{x_t^{(3)}=(i,k)}$.

The following corollary states that $\hat{M}_{\text{biso}}$ achieves a reconstruction rate of order $n^{7/6}\lambda^{-5/6}$ in the bi-isotonic model.

Corollary 2.5. *Assume that $\lambda \in [1/n, 8n^2]$. There exists a universal constant C'' such that*

$$\sup_{\substack{M:\, M_{\pi^{*-1}\eta^{*-1}}\in\mathbb{C}_{\text{biso}} \\ \pi^*,\eta^*\in\Pi_n}} \mathbb{E}\left[\|(\hat{M}_{\text{biso}})_{\hat{\pi}\hat{\eta}} - M\|_F^2\right] \le C'' \log^{C''}(n) n^{7/6}\lambda^{-5/6} \ .$$

Here, we have fixed $n = d$ to simplify the exposition but we could extend the analysis to general n and d. Our risk bound improves over the rate $n^{5/4}\lambda^{-3/4}$ of Mao et al. [10]. In [8], Liu and Moitra have introduced a procedure achieving the rate $n^{7/6}$ in the specific case where $\lambda = n^{o(1)}$. In some way, our procedure generalizes their results for general λ, while being applicable to the more general isotonic models.

Still, we recall that the optimal risk (without computational constraints) for estimating the matrix M is of the order n/λ – see e.g. [18, 10]. This remains an open problem to establish the existence of a computational-statistical gap or to construct a polynomial-time procedure achieving this risk on SST and bi-isotonic models.

3. Description of the ISR procedure

3.1. Weighted directed graph $\mathcal{W}$ and estimator $\hat{\pi}$

Our approach involves the iterative construction of a weighted directed graph $\mathcal{W}$, represented by an antisymmetric matrix in $\mathbb{R}^{n\times n}$. More formally, for any experts i, j in $[n]$, we have $\mathcal{W}(i,j) = -\mathcal{W}(j,i)$. In a nutshell, $\mathcal{W}(i,j)$ quantifies our evidence of the comparisons between expert i and expert j. If $\mathcal{W}(i,j)$ is large and positive (resp. negative), we are confident that the expert i is above (below) the expert j. Most of the procedure is dedicated to the construction of $\mathcal{W}$. Before this, let us explain how we deduce our estimator $\hat{\pi}$ from $\mathcal{W}$.

For a given weighted directed graph $\mathcal{W}$, we define its corresponding directed graph at threshold $\gamma > 0$ as

$$\mathcal{G}(\mathcal{W}, \gamma) = \{(i,j) \in [n]^2 \ : \ \mathcal{W}(i,j) > \gamma\} \ . \tag{13}$$

For any thresholds $\gamma < \gamma'$, it holds that $\mathcal{G}(\mathcal{W}, \gamma) \subset \mathcal{G}(\mathcal{W}, \gamma')$. In other words, the function $\gamma \to \mathcal{G}(\mathcal{W}, \gamma)$ is nondecreasing. When $\gamma \geq \max_{i,j} |\mathcal{W}(i,j)|$, $\mathcal{G}(\mathcal{W}, \gamma) = \varnothing$ is the trivial graph with no edges. Let $\hat{\gamma}$ be the smallest threshold γ such that $\mathcal{G}(\mathcal{W}, \gamma)$ is a directed acyclic graph (**DAG**). By monotonicity, $\mathcal{G}(\mathcal{W}, \hat{\gamma})$ is also the largest **DAG** among $\{\mathcal{G}(\mathcal{W}, \gamma), \gamma \geq \hat{\gamma}\}$. We then build the estimator $\hat{\pi}$ by picking any permutation that is consistent with the graph $\widehat{\mathcal{G}} := \mathcal{G}(\mathcal{W}, \hat{\gamma})$, that is if $(i,j) \in \widehat{\mathcal{G}} \cap [n]^2$ then $\hat{\pi}(i) \geq \hat{\pi}(j)$. To put it another way, the general idea of our procedure can be summarized into these three components:

1. Construct a weighted directed graph $\mathcal{W}$ between the experts.
2. Compute the largest directed acyclic graph $\widehat{\mathcal{G}}$ of $\mathcal{W}$.
3. Take any arbitrary permutation $\hat{\pi}$ that is consistent with $\widehat{\mathcal{G}}$.

The construction of $\mathcal{W}$ is at the core of this paper, and the computation of $\widehat{\mathcal{G}}$ and $\hat{\pi}$ will be discussed in Section 3.7. Still, we already point out that the third point can be dealt in polynomial time using Mirsky's algorithm [12].

3.2. Construction of $\mathcal{W}$ with ISR

3.2.1. Description of the subsampling

Let us now describe the construction of the weighted directed graph $\mathcal{W}$. Let $T \geq 1$ be an arbitrary integer, representing the number of steps of our procedure. In what follows, we explain how we subsample the data from (2) into $5T$ independent matrices $(Y^{(s)})_{s=1\dots 5T}$. Recall that we are given N observations (x_t, y_t), where N follows a Poisson distribution $\mathcal{P}(\lambda n d)$. Let us divide the observations into $5T$ batches $(N^{(s)})_{s=0,\dots,5T-1}$, aggregated into matrices of averaged observations $Y^{(s)}$. To that end, we let S_u be i.i.d. uniform random variables in $\{0, \dots, 5T-1\}$ representing a random batch for observation u and we define

$$N^{(s)} = \{u \in \{1 \dots, N\} \ : \ S_u = s\} \quad \text{and} \quad Y^{(s)}_{ik} = \sum_{t \in N^{(s)}} \frac{y_t}{\mathbf{r}^{(s)}_{ik} \vee 1} \mathbf{1}\{x_t = (i,k)\} \ , \tag{14}$$

where, for any $(i,k) \in [n] \times [d]$, $\mathbf{r}^{(s)}_{ik} = \sum_{t \in N^{(s)}} \mathbf{1}\{x_t = (i,k)\}$ is the number of times the coefficient position (i,k) is observed in batch s. $Y^{(s)}_{ik}$ is equal to 0 if (i,k) is not observed in batch s and it is equal to the average of the observations y_t for which $x_t = (i,k)$ otherwise. We also define the mask matrix $B^{(s)}$ as being equal to 0 at location (i,k) if the value is missing from batch s, and to 1 otherwise.

$$B^{(s)}_{ik} = \mathbf{1}\{\mathbf{r}^{(s)}_{ik} \geq 1\} \ . \tag{15}$$

Define $\lambda_0 = \lambda/5T$. In our sampling scheme, where the data is divided into $5T$ samples, each coefficient $B^{(s)}_{ik}$ has a probability of $1 - e^{-\lambda_0}$ of being equal to one. It is worth mentioning that a different subsampling scheme was performed in [14], consisting in aggregating consecutive columns. However, such a scheme is not applicable in our case as we do not assume the rows of M to be nondecreasing, unlike in [14].

3.2.2. Neighborhoods in comparison graphs

At each step $t = 0, \dots, T-1$ of the procedure, we aim to enrich our knowledge of the order of the experts, which we formally do by nondecreasing the weights of $\mathcal{W}$ in absolute value.

At $T = 0$, we start with the weights $\mathcal{W}_{ij}$ all being equal to zero. A meaningful update of $\mathcal{W}$ around a reference expert i can be done when we restrict ourselves to experts that are in a neighborhood of i. Broadly speaking, a neighborhood of i is a set made of all the experts j that are not comparable to i with respect to a given partial order.

More precisely, for any directed graph $\mathcal{G}$ and any experts $i, j \in \{1, \dots, n\}$, we say that i and j are $\mathcal{G}$-comparable if there is a path from i to j or from j to i in $\mathcal{G}$. The neighborhood $\mathcal{N}(\mathcal{G}, i)$ of i in $\mathcal{G}$ can then naturally be defined as the set of experts j that are not $\mathcal{G}$-comparable with i. Equipped with the concept of neighborhood, our overall strategy involves iterating over all possible thresholds $\gamma \in \Gamma$ such that $\mathcal{G}(\mathcal{W}, \gamma)$ is acyclic, as well as all possible experts i. At each iteration, we apply the soft local ranking procedure Algorithm 2 described in the next subsection. Algorithm 2 updates the weights between i and any expert j in the neighborhood $\mathcal{N}(\mathcal{G}(\mathcal{W}, \gamma), i)$ of i. Our approach can be summarized as follows:

1. Subsample the data - see Section 3.2.1.
2. Initialize $\mathcal{W}$ to be the directed graph with all weights set to 0.
3. For all $t = 0, \dots, T-1$ and $\gamma \in \Gamma$ such that $\mathcal{G}(\mathcal{W}, \gamma)$ is acyclic and all $i \in [n]$, update $\mathcal{W}$ with the soft local ranking procedure Algorithm 2.

Algorithm 1 **ISR**(T, Γ)

Require: N and observations $(x_t, y_t)_{t=1,\dots,N}$ according to (2), a number of steps T and a valid grid Γ as in (10)
Ensure: A weighted graph $\mathcal{W}$ and an estimator $\hat{\pi}$
1: Aggregate the observation into $5T$ matrices of observation $(Y^{(s)})$ as in (14)
2: Initialize $\mathcal{W}(i, j) = 0$ for all $(i, j) \in [n]^2$, and $\hat{\gamma} = 0$
3: **for** $t = 0, \dots, T-1$ **do**
4: **for** $\gamma \in \Gamma \cap [\hat{\gamma}, +\infty)$ **do**
5: Compute $\mathcal{G} = \mathcal{G}(\mathcal{W}, \gamma)$ the directed graph at threshold γ of $\mathcal{W}$ as in (13) and set $P = \mathcal{N}(\mathcal{G}, i)$.
6: Take 5 samples $\mathcal{Y} = (Y^{(5t)}, \dots, Y^{(5t+4)})$
7: **for** $i \in [n]$ **do**
8: Apply **SLR**$(\mathcal{Y}, \mathcal{W}, \gamma, i, \mathcal{G}, P)$ to update $\mathcal{W}$
9: **end for**
10: **end for**
11: Set $\hat{\gamma}$ as the smallest γ such that $\mathcal{G}(\mathcal{W}, \gamma)$ is acyclic
12: **end for**
13: Set $\widehat{\mathcal{G}} = \mathcal{G}(\mathcal{W}, \hat{\gamma})$ be the largest acyclic **DAG** (see (13))
14: Set $\hat{\pi}$ to be any arbitrary permutation that is consistent with $\widehat{\mathcal{G}}$
15: **return** $\mathcal{W}$ and $\hat{\pi}$

The main Line 8 of Algorithm 1 aims to provide a soft ranking of the neighborhood P of i by setting positive (resp. negative) weights $\mathcal{W}_{ij}$ to experts $j \in P$ that are significantly below (resp. above) i. Line 11 together with restricting $\gamma \geq \hat{\gamma}$ simply guarantees that all the considered graph $\mathcal{G}$ are acyclic. Finally, Lines 13 and 14 simply correspond to the construction of the final permutation, described in the second and third points of Section 3.1.

3.3. Description of the updating procedure

3.3.1. Local weighted sums

Let us describe the process of updating a given weighted graph $\mathcal{W}$, which will be used twice at each call of the soft local ranking Algorithm 2. Let us fix a weighted graph $\mathcal{W}$, an element $s \in \{0, \dots, 5T-1\}$ and $Y := Y^{(s)}$ the matrix defined in (14). We also let $i \in [n]$ be an arbitrary expert corresponding to Line 7 of Algorithm 1, and γ be any threshold in the grid Γ. We write

$P \coloneqq \mathcal{N}(\mathcal{G}(\mathcal{W},\gamma),i) \subset [n]$ for the neighborhood of i in $\mathcal{G}(\mathcal{W},\gamma)$, echoing the notation of the sets that are trisected in [14].

Since the matrix M is, up to a row-permutation, a column-wise isotonic matrix, it follows that, if the expert i is above j, then for any vector $w \in \mathbb{R}_+^d$, we have $\sum_{k=1}^d w_{ik}M_{ik} \geq \sum_{k=1}^d w_{jk}M_{jk}$. As a consequence, the crux of the algorithm is to find suitable data-driven weights w that allow to discriminate the experts. As explained in the introduction, earlier works focused on uniform weights $w = \mathbf{1}_{[d]}$ [18] which, unfortunately leads to suboptimal results. Before discussing the choice of the weights w in the following subsections, let us first formalize how we leverage on w to compare the experts and update the graph $\mathcal{W}$.

Given a subset $Q \subset [d]$ of columns and a non-zero vector $w \in \mathbb{R}_+^Q$, we first check whether the following condition is satisfied:

$$\lambda_0 \|w\|_2^2 \geq \|w\|_\infty^2 \ , \tag{16}$$

where we recall that $\lambda_0 = \lambda/5T$. This condition is always verified when $\lambda_0 \geq 1$, and it is equivalent to $\lambda_0|Q| \geq 1$ when $w = \mathbf{1}_Q$. Condition (16) ensures that w is not too sparse which could be harmful when many observations are lacking (λ_0 small).

If this condition is not satisfied, then we leave the weights of $\mathcal{W}$ unchanged. Otherwise, we define the (Y,P,w)-updating weights $\mathcal{U} \coloneqq \mathcal{U}(Y,P,w)$ around i as

$$\mathcal{U}_{ij} = \frac{1}{\sqrt{\frac{1}{\lambda_0} \wedge \lambda_0}} \cdot \langle Y_{i\cdot} - Y_{j\cdot}, \frac{w}{\|w\|_2} \rangle \ , \tag{17}$$

where, for all $w' \in \mathbb{R}^Q$ and $a \in \mathbb{R}^d$, we write $\langle a, w' \rangle = \sum_{k\in Q} a_k w'_k$. We can then update the weighted directed graph around i by setting, for all $i \in P$ such that $|\mathcal{U}_{ij}| \geq |\mathcal{W}_{ij}|$,

$$\mathcal{W}_{ij} = \mathcal{U}_{ij} \quad \text{and} \quad \mathcal{W}_{ji} = -\mathcal{U}_{ij} \ . \tag{18}$$

As explained above, if we replace $Y_{i\cdot}$ and $Y_{j\cdot}$ by $M_{i\cdot}$ and $M_{j\cdot}$ respectively in (17), then the corresponding value of the statistic is non-negative if expert i is above j. Hence, a large value for $\mathcal{U}_{ij}$ provides evidence that i is above j.

Computing $\mathcal{U}(Y,P,w)$ for suitable directions w is the basic brick or our procedure, since it is through the update (18) that we iteratively increase the weights of $\mathcal{W}$. This update shares some similarities to the pivoting algorithm introduced in [8] and also used in [14], in the sense that while we are fixing an arbitrary reference expert i to compute pairwise comparisons, they fix a set P and compute a pivot expert i_0 that would correspond to a quantile of the set $\{\langle Y_{j\cdot}, \frac{w}{\|w\|_2}\rangle, j \in P\}$ in the case $\lambda_0 = 1$.

Note that the orientation of a given weighted edge (i,j) can change during the procedure if it turns out that $|\mathcal{U}_{ij}| \geq |\mathcal{W}_{ij}|$ and that $\mathcal{U}_{ij}\mathcal{W}_{ij} \leq 0$. This simply means that if the direction w leads to a more significant weight between some experts i and j, then we are more confident to use the vector w and to revise the order between i and j.

For $Q \subset [d]$, choosing $w = \mathbf{1}_Q$ in (17) amounts to compute the average of the observations over all questions in Q. We now explain in the main sections how we iteratively build adaptive weights w that allow to improve over the naive global average given by $w = \mathbf{1}_{[d]}$.

3.3.2. Definitions of a rank in a **DAG**

We first introduce a few definitions on directed acyclic graphs $\mathcal{G}$, which we formally define as a set of directed edges $(i,j) \in [n]^2$ for which there is no cycle. We denote $\mathbf{path}(i,j) = \{(k_1,\dots,k_L) \ : \ L > 0 \text{ and } (i,k_1),\dots,(k_L,j) \in \mathcal{G}\}$ as the set of all possible paths from i to j, and we write $|s|$ for the length of any path s. We say that i and j are $\mathcal{G}$-comparable if

$\mathbf{path}(i,j) \cup \mathbf{path}(j,i) \neq \varnothing$, and we write $\mathcal{N}(i,\mathcal{G})$ for the set of all experts that are not $\mathcal{G}$-comparable with i. If i, j are $\mathcal{G}$-comparable, it either holds that $\mathbf{path}(i,j) = \varnothing$ or $\mathbf{path}(j,i) = \varnothing$. We say in the first case that i is $\mathcal{G}$-below j and that i is $\mathcal{G}$-above j in the second case. we also define the relative rank from i according to $\mathcal{G}$ as the length of the longest path in $\mathcal{G}$ from i to j, or minus the longest past from j to i depending on wether i is $\mathcal{G}$-above or $\mathcal{G}$-below j:

$$\mathbf{rk}_{\mathcal{G},i}(j) = \max\{|s| \ : \ s \in \mathbf{path}(i,j)\} - \max\{|s| \ : \ s \in \mathbf{path}(j,i)\} \ . \tag{19}$$

Here, we use the convention $\max \varnothing = 0$. With this definition, the neighborhood of a given expert i is equal to the set of experts whose relative rank is equal to 0, that is $\mathcal{N}(\mathcal{G},i) = \mathbf{rk}_{\mathcal{G},i}^{-1}(0)$. Moreover, an expert $j \in [n]$ is $\mathcal{G}$-above (resp. $\mathcal{G}$-below) i if and only if $\mathbf{rk}_{\mathcal{G},i}(j) \geq 1$ (resp. $\mathbf{rk}_{\mathcal{G},i}(j) \leq -1$). Although $\mathcal{G}$ stands for a finite set of edges with endpoints in $[n]$, we extend it to a set of edges with endpoints in $\mathbb{Z}^2$ by putting in $\mathcal{G}$ every $(i,j) \in \mathbb{Z}^2$ such that $i > j$ and $j \leq 0$ or $i \geq n+1$.

3.3.3. Description of the soft local ranking algorithm

To update the weighted directed graph $\mathcal{W}$ in Line 8 of Algorithm 1, we apply the soft local ranking procedure **SLR** to all experts $i \in [n]$ and all thresholds γ. To define our soft local ranking procedure, let us fix $\mathcal{W}$, an expert i and a threshold γ such that $\mathcal{G}(\mathcal{W},\gamma)$ is acyclic. As a shorthand, we write $\mathcal{G}$ and P respectively for the thresholded graph $\mathcal{G}(\mathcal{W},\gamma)$ and the neighborhood $\mathcal{N}(\mathcal{G},i)$ of i in $\mathcal{G}$.

We write $\mathcal{D}$ for the set of all dyadic numbers: $\mathcal{D} = \{2^k \ : \ k \in \mathbb{Z}\}$ and we define the set $\mathcal{H} = \mathcal{D} \cap \left[\frac{1}{nd}, 1\right]$. We denote $\overline{y}(P)$ as the mean of the vectors $Y_{j\cdot}$ over all $j \in P$, that is $\overline{y}_k(P) = \frac{1}{|P|}\sum_{j\in P} Y_{jk}$, for any $k \in [d]$. **SLR** relies on the following steps repeated over all height $h \in \mathcal{H}$. It is also described in Algorithm 2.

1. **Dimension reduction.** Using the first sample $Y^{(1)}$, we first reduce the dimension by selecting a subset $\widehat{Q}^h \subset [d]$ corresponding to wide confidence regions. Recall that $\mathbf{rk}_{\mathcal{G},i}$ is the relative rank to i defined in (19). For any $a > 0$, define the sets $\mathcal{N}_a := \mathcal{N}_a(\mathcal{G},i)$ (resp. $\mathcal{N}_{-a} := \mathcal{N}_{-a}(\mathcal{G},i)$) of experts j which are $\mathcal{G}$-above (resp. $\mathcal{G}$-below) all the experts of P and whose relative rank to any $i' \in P$ is at most a in absolute value:

$$\mathcal{N}_a = \bigcap_{i'\in P} \mathbf{rk}_{\mathcal{G},i'}^{-1}([1,a]) \quad \text{and} \quad \mathcal{N}_{-a} = \bigcap_{i'\in P} \mathbf{rk}_{\mathcal{G},i'}^{-1}([-1,-a]) \ . \tag{20}$$

   Secondly, we define for any question $k \in [d]$ and $a \geq 1$ the width statistic $\widehat{\mathbf{\Delta}}_k$ as the difference between the mean of the experts in $\mathcal{N}_a$ and the mean of the experts in $\mathcal{N}_{-a}$. Then, $\hat{a}_k$ is set to be the first value of $a \geq 1$ such that any $a' \geq a$ has a corresponding width statistic of at least $(\lambda_0 \wedge 1)h$:

$$\widehat{\mathbf{\Delta}}_k(a) = \overline{y}_k(\mathcal{N}_a) - \overline{y}_k(\mathcal{N}_{-a}) \quad \text{and} \quad \hat{a}_k(h) = \max\left\{a \geq 1 \ : \ \frac{1}{\lambda_0 \wedge 1}\widehat{\mathbf{\Delta}}_k(a) < h\right\} + 1 \ . \tag{21}$$

   Finally, we define $\widehat{Q}^h := \widehat{Q}^h(\mathcal{G},i)$ as the set of indices k such that $\hat{a}_k(h)$ is relatively small.

$$\widehat{Q}^h = \{k \in [d] \ : \ |\mathcal{N}_{\hat{a}_k(h)}| \wedge |\mathcal{N}_{-\hat{a}_k(h)}| \leq \frac{1}{\lambda_0 h^2}\} \ . \tag{22}$$

   Intuitively, if the experts above and below i vary by more than h on a specific question k, then this question should belong to $\widehat{Q}^h$. Conversely, if the experts below and above i are nearly equal on the question k, than $\hat{a}_k(h)$ will be large and k will not be selected in $\widehat{Q}^h$.

2. **Average-based weighted sums.** Still using the first sample $Y^{(1)}$, we examine the corresponding submatrix $Y^{(1)}(P,\widehat{Q})$ restricted to questions in $\widehat{Q}$. If the row sums of Y are larger than the current edges, we update the weighted edges. More formally, we compute the $(Y^{(1)},P,\mathbf{1}_{\widehat{Q}})$-updating weighted edges $(\mathcal{U}_{\widehat{Q}})$ around i as defined in (17) and update $\mathcal{W}$ as in (18). We then also update $\mathcal{G}=\mathcal{G}(\mathcal{W},\gamma)$ and $P=\mathcal{N}(\mathcal{G},i)$.
3. **PCA-based weighted sums.** Relying on the samples $Y^{(2)}$, $Y^{(3)}$, $Y^{(4)}$, $Y^{(5)}$, we do a slight abuse of notation and write $Y^{(s)}$ for the restriction of $Y^{(s)}$ to the subset $P,\widehat{Q}^h$ for $s=2,3,4,5$. Ideally, we would get an informative direction w from the largest right singular vector of $\mathbb{E}[Y^{(2)}-\overline{Y}^{(2)}]\in\mathbb{R}^{P\times\widehat{Q}^h}$. Indeed, it is known (see the proofs for more details) that the entries of the first right singular vector of an isotonic matrix all share the same sign and are most informative to compare the experts. However, computing directly the empirical right-singular vector of $Y^{(2)}-\overline{Y}^{(2)}$ does not lead to the desired bounds because (i) this matrix is perhaps highly rectangular (ii) the noise is possibly heteroskedastic and (iii) this matrix is perhaps sparse because of the many missing observations when λ_0 is small. Here, we use a workaround which is reminiscent of that of [14] and discussed later. First, we compute $\hat{v}$ as a proxy for the first left singular vector of $\mathbb{E}[Y^{(2)}-\overline{Y}^{(2)}]$.

$$\hat{v}:=\hat{v}(P,\widehat{Q}^h)=\underset{v\in\mathbb{R}^P:\ \|v\|_2\le 1}{\arg\max}\left[\|v^T(Y^{(2)}-\overline{Y}^{(2)})\|_2^2-\frac{1}{2}\|v^T(Y^{(2)}-\overline{Y}^{(2)}-Y^{(3)}+\overline{Y}^{(3)})\|_2^2\right]. \quad (23)$$

The right-hand side term in (23) deals with the heteroskedasticity of the noise matrix E in (3). $\hat{v}$ in (23) can be computed efficiently since it corresponds to the leading eigenvector of a symetric matrix. For technical reasons occurring in the sparse observation regime (i.e. when λ_0 is small), we then threshold the largest absolute values of the coefficients of $\hat{v}$ at $\sqrt{\lambda_0}$ and define $(\hat{v}_-)_i=\hat{v}_i\mathbf{1}\{|\hat{v}_i|\le\sqrt{\lambda_0}\}$. After having calculated $\hat{v}_-$, we consider as in [14] the image $\hat{z}=\hat{v}_-^T(Y^{(4)}-\overline{Y}^{(4)})\in\mathbb{R}^{\widehat{Q}}$ of $\hat{v}_-$. We then threshold the smallest values of $\hat{z}$ and take the absolute values of the components. Thus, we get $\hat{w}^+\in\mathbb{R}^{\widehat{Q}}$ defined by $(\hat{w}^+)_l=|\hat{z}_l|\mathbf{1}\{|\hat{z}_l|\ge\gamma\sqrt{\lambda_0\wedge\frac{1}{\lambda_0}}\}$ for any $l\in\widehat{Q}$.

Finally, we consider the last submatrix $Y^{(5)}=Y^{(5)}(P,\widehat{Q})$. We apply these weights $\hat{w}^+$ to compute the row-wise weighted sums of $Y^{(5)}$ and update the weighted edges. More formally, we compute the $(Y^{(5)},P,\hat{w}^+)$-updating weighted edges $\mathcal{U}(Y^{(5)},P,\hat{w})$ around i as defined in (17). We finally update the weighted directed graph $\mathcal{W}$ with $\mathcal{U}(Y^{(5)},P,\hat{w}^+)$ as in (18).

Algorithm 2 $\mathbf{SLR}((Y^{(s)})_{s=1,\dots,5},\mathcal{W},\gamma,i,\mathcal{G},P)$

Require: 6 samples $(Y^{(s)})_{s=1,\dots,5}$, a weighted directed graph $\mathcal{W}$, a threshold γ such that $\mathcal{G}(\mathcal{W},\gamma)$ is acyclic and an expert $i\in[n]$. $\mathcal{G}$ and P are shorthands for the thresholded graph $\mathcal{G}(\mathcal{W},\gamma)$ and the neighborhood $\mathcal{N}(\mathcal{G},i)$.
Ensure: An update of $\mathcal{W}$

for $h\in\mathcal{H}$ **do**
  Compute $\widehat{Q}^h:=\widehat{Q}(\mathcal{G},i)$ as in (22) using sample $Y^{(1)}$
  Let $\mathcal{U}_{\widehat{Q}^h}$ be the $(Y^{(1)},P,\mathbf{1}_{\widehat{Q}^h})$-updating weighted edges around i as in (17), using again sample $Y^{(1)}$
  Update $\mathcal{W}$ with $\mathcal{U}(\widehat{Q}^h)$ as in (18) and update $\mathcal{G}=\mathcal{G}(\mathcal{W},\gamma)$, $P=\mathcal{N}(\mathcal{G},i)$
  Restrict the samples $(Y^{(s)})_{s=2,\dots,5}$ to $P,\widehat{Q}^h$ in the following remaining steps
  Compute the PCA-like direction $\hat{v}:=\hat{v}(P,\widehat{Q}^h)$ as in (23) and define $(\hat{v}_-)_i=\hat{v}_i\mathbf{1}\{|\hat{v}_i|\le\sqrt{\lambda_0}\}$
  Compute $\hat{z}=\hat{v}_-^T(Y^{(4)}-\overline{Y}^{(4)})$ and define $\hat{w}^+$ by $(\hat{w}^+)_l=|\hat{z}_l|\mathbf{1}\{|\hat{z}_l|\ge\gamma\sqrt{\lambda_0\wedge\frac{1}{\lambda_0}}\}$ for any $l\in\widehat{Q}^h$
  Let $\mathcal{U}(Y^{(5)},\hat{w}^+)$ be the $(Y^{(5)},P,\hat{w}^+)$-updating weighted edges around i as in (17)
  Update $\mathcal{W}$ with $\mathcal{U}(Y^{(5)},\hat{w}^+)$ as in (18)
end for

3.4. Toy example illustrating Algorithm 2

To understand why the steps described in Algorithm 2 are relevant, assume that $\pi^* = \mathrm{id}$ and consider the following simple example where $n = 204$, $d = 10$, and where the isotonic matrix $M_{\pi^{*-1}}$ can be decomposed into three blocks of rows as

$$M_{\pi^{*-1}} = \alpha + \frac{h}{2}\begin{pmatrix} \mathbf{0} & \mathbf{0} & \mathbf{1} & \mathbf{1} & \mathbf{0} & \mathbf{1} & \mathbf{1} & \mathbf{0} & \mathbf{1} & \mathbf{1} \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & -1 & 0 & -1 & 0 & 0 & -1 & -1 \\ 0 & 0 & 0 & -1 & 0 & -1 & 0 & 0 & -1 & -1 \\ \mathbf{0} & \mathbf{0} & \mathbf{-1} & \mathbf{-1} & \mathbf{0} & \mathbf{-1} & \mathbf{-1} & \mathbf{0} & \mathbf{-1} & \mathbf{-1} \end{pmatrix}.$$

In the above matrix, α is any number in $(h, 1-h)$, and $\mathbf{0}, \mathbf{1}$ are the columns in $\mathbb{R}^{100}$ whose coefficients are respectively all equal to 0 and 1. Assume that the statistician already knows that the first and the third blocks are made of experts that are respectively above and below the second block. If $\mathcal{W}, P, \gamma$ are the parameters fixed in Algorithm 2, the three blocks correspond respectively to the subsets $\mathcal{N}_1 \cup \mathcal{N}_2$, P and $\mathcal{N}_{-1} \cup \mathcal{N}_{-2}$ in our example. Provided that $\mathcal{N}_{-2}$ and $\mathcal{N}_2$ are large enough, the set $\widehat{Q}^h$ only keeps columns corresponding to indices k where $\widehat{\mathbf{\Delta}}_k(1)$ is large – those are highlighted in blue.

Then, we can work on the reduced subset $\widehat{Q}^h$ of columns highlighted in blue. As one may check, $\widehat{Q}^h$ contains all the relevant columns to decipher the experts in the block P. Besides, the expected matrix of observations restricted to the block P and to $\widehat{Q}^h$ is of rank one:

$$\mathbb{E}[Y - \overline{Y}] = \frac{h}{2}\begin{pmatrix} 0 & 1 & 1 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 & 1 & 1 \\ 0 & -1 & -1 & 0 & -1 & -1 \\ 0 & -1 & -1 & 0 & -1 & -1 \end{pmatrix}.$$

In particular, the right singular vector of this matrix is of the form $(0, 1, 1, 0, 1, 1)$ and provides suitable weights to decipher the two largest experts from the two lowest experts in the above matrix. The PCA-based weighted sums steps above precisely aims at estimating these weights.

3.5. Comments on the procedure and relation to the literature

Finding confidence regions $\widehat{Q}$ before computing weighted sums on the corresponding columns is at the core of our procedure. This idea generalizes the RankScore procedure of [7] which rather computes averages on the subsets $[d]$ or on the singletons $\{1\}, \dots, \{d\}$. As mentioned in the introduction, only using the subsets of the RankScore method in [7] does not allow to reach the optimal rate for permutation estimation or matrix reconstruction.

In Algorithm 2, the computation of subsets $\widehat{Q}^h$ is reminiscent of some aspects of the non oblivious trisection procedure used in [14] for the bi-isotonic model. In fact, the statistic $\widehat{\mathbf{\Delta}}_k$ corresponds to the statistic $\widehat{\mathbf{\Delta}}^{(\mathrm{ext})}_{k,1}$ in [14]. Apart from that, the selection of subsets of questions was quite different in [14] as it mostly involved change-point detection ideas as introduced in [8]. However, those ideas are irrelevant in our setting because the rows do not exhibit any specific structure in the isotonic model.

The high-level sorting method in [14] is based on a hierarchical sorting tree with memory. In contrast, our new algorithm is based on an iterative refinement of a weighted comparison graph. This new algorithm is more natural and benefits from the fact that it is almost free of any tuning parameter. Indeed, at the end of Algorithm 1, we simply use the threshold $\hat{\gamma}$ corresponding to the largest acyclic $\hat{\mathcal{G}}$ graph in $\mathcal{W}$. No significant threshold needs to be

chosen, since any permutation that is consistent with $\hat{\mathcal{G}}$ is also necessarily consistent with $\mathcal{W}$ thresholded at values larger than $\hat{\gamma}$.

The spectral step in [14] is quite similar to the third step of our procedure described in section 3.3.3, except for the first thresholding of $\hat{v}$ to obtain $\hat{v}_-$. In [14], this workaround was not needed mainly because in the bi-isotonic model, it is possible to aggregate sparse observations by merging consecutive columns – see [14] for further details. This is however not possible here.

As mentioned in the introduction, Liu and Moitra [8] obtain an upper bound of the permutation loss of the order of $n^{7/6}$ for the estimation of two unknown permutations in the case where $M \in \mathbb{R}^{n\times n}$ is bi-isotonic. Broadly speaking, their method involves iterating a clustering method called block-sorting over groups of rows or columns that are close with each other. Using this sorting method based on block-sorting, their whole approach alternates between row sorting and column sorting for a subpolynomial number of time. Besides, their procedure makes heavily use of bi-isotonicity of the matrix. It turns out that Algorithm 2 reaches the same rate in this bi-isotonic model by running only once on the rows, and once on the columns, as described in Section 2.4. Otherwise said, if the problem is to estimate only π^* in the bi-isotonic model, we proved that only the isotonicity of the columns is necessary to achieve the state-of-the-art polynomial-time upper bound of order $n^{7/6}$.

3.6. Examples of valid grids $\mathbf{\Gamma}$

Remark that the simple set $\{(u+1)\cdot\phi_{\mathrm{L}_1}, u \in \{0,\dots,2\lfloor\log_2(n)\rfloor+2\}\}$ is a valid grid of logarithmic size with $\bar{\gamma} \le (2\log_2(n)+3)\phi_{\mathrm{L}_1}$. This set is the smallest valid grid achieving the smallest possible value of $\bar{\gamma}$. However, it depends on the quantity ϕ_{L_1} which is perhaps a bit pessimistic in practice.

An other choice can be to take $\mathbb{R}^+$ itself, albeit infinite. Indeed, the set $\{\mathcal{G}(\mathcal{W},\gamma), \gamma \ge 0\}$ is made of at most n^2 possible directed graphs for any $\mathcal{W}$ during the whole procedure. Choosing $\mathbb{R}^+$ is convenient since it does not depend on the constants in ϕ_{L_1} that are likely to be overestimated. The drawback of choosing $\mathbb{R}^+$ though is that the number of tested γ in Algorithm 2 becomes quadratic in n.

Finally, a good compromise is to take the set $\{(1+\frac{1}{\log_2(n)})^{u'},\ u' \in \mathbb{Z}\}$. It is easy to check that it contains a sequence satisfying (10) whose length is at least $2\lfloor\log_2(n)\rfloor+3$ and whose maximum $\bar{\gamma}$ is a polylogarithmic function in nd/δ.

3.7. Discussion on the computation of $\widehat{\mathcal{G}}$ and $\hat{\boldsymbol{\pi}}$

Once we have suitable weighted graph $\mathcal{W}$, it remains to construct the permutation $\hat{\pi}$, as in the second and third point of Section 3.1.

For the second point, checking that a given directed graph is acyclic can be done through depth first search with a computational complexity less than n, so that computing $\hat{\gamma}$ can be done with less than $|\Gamma|n$ operations. As discussed in Section 3.3, it is possible to choose Γ to be of size of order less than $\log(n)$. If Γ is bounded and is such that any different thresholds γ,γ' in Γ satisfy $|\gamma-\gamma'| \ge \eta$ for some $\eta > 0$, the computation of $\hat{\gamma}$ can always be done with complexity of order less than $n\log(\max(\Gamma)/\eta)$.

Regarding the third point, a permutation $\hat{\pi}$ can be computed in polynomial time from the directed acyclic graph $\hat{\mathcal{G}}$ using Mirsky's algorithm [12] – see also [13]. It simply consists in finding the minimal experts i in $\hat{\mathcal{G}}$, removing them and repeat this process. This construction is in fact equivalent to ranking the experts according to the index $\mathbf{rk}_{\hat{\mathcal{G}},0}$ as defined in (19).

4. Concentration inequality for rectangular matrices

In this section, we state a concentration inequality for rectangular random matrices with independent entries satisfying a Bernstein-type condition. This section can be read independently of the rest of the paper. Let p and q be two positive integers and $X \in \mathbb{R}^{p\times q}$ be a random matrix with independent and mean zero coefficients. Assume that there exists $\sigma > 0$ and $K \geq 1$ such that for any $i = 1, \dots, p$ and $k = 1, \dots, q$,

$$\forall u \geq 1, \qquad \mathbb{E}[(X_{ik})^{2u}] \leq \frac{1}{2} u! \sigma^2 K^{2(u-1)} \ . \tag{24}$$

This Bernstein-type condition (24) is exactly the same as Assumption 1 in [2] – see [2] for a discussion. Let $\Lambda \in \mathbb{R}^{p\times p}$ be any orthogonal projection matrix, i.e. $\Lambda = \Lambda^T$ and $\Lambda^2 = \Lambda$. We write r_Λ for the rank of Λ.

Proposition 4.1. *There exists a positive numerical constant κ such that the following holds for any $\delta > 0$.*

$$\|\Lambda(XX^T - \mathbb{E}[XX^T])\Lambda\|_{\mathrm{op}} \leq \kappa \left[\sqrt{(\sigma^4 pq + \sigma^2 q)\log(p/\delta)} + (\sigma^2 r_\Lambda + K^2 \log(q))\log(p/\delta) \right] \ . \tag{25}$$

For the sake of the discussion, consider the particular case where $X_{ik} = B_{ik}E_{ik}$, with B_{ik} and E_{ik} being respectively independent Bernoulli random variable of parameter σ^2 and centered Gaussian random variable with variance 1. By a simple computation done e.g. in (77), X_{ik} satisfies condition (24) with K being of the order of a constant. Hence, if $K^2 \log(q) \leq \sigma^2 p$, applying Proposition 4.1 with the identity matrix Λ gives

$$\|XX^T - \mathbb{E}[XX^T]\|_{\mathrm{op}} \leq 2\kappa\sigma^2 \left[\sqrt{pq\log(p/\delta)} + p\log(p/\delta) \right] \ , \tag{26}$$

with probability at least $1 - \delta$.

Up to our knowledge, the inequality (26) is tighter than state-of-the-art result random rectangular sparse matrices in the regime where $q \gg p$ and $\sigma^2 \ll 1$. In fact, most of the results in the literature concerning random matrices state concentration inequalities for the non centered operator norm $\|XX^T\|_{\mathrm{op}}$ – see the survey of Tropp [21].

More specifically, Bandeira and Van Handel [1] provide tight non-asymptotic bounds for the spectral norm of a square symmetric random matrices with independent Gaussian entries, and derive tail bounds for the operator norm of XX^T. For instance, Corollary 3.11 in [1], implies that, for some numerical constant c, $\mathbb{E}[\|XX^T\|_{\mathrm{op}}^2] \leq c(\sigma^2(p \vee q) + \log(p \vee q))$. Together with a triangular inequality, Bandeira and Van Handel imply $\|XX^T - \mathbb{E}[XX^T]\|_{\mathrm{op}}^2 \leq c\sigma^2((p \vee q) + \log(\frac{p\vee q}{\delta}))$ with probability higher than $1 - \delta$.

While the order of magnitude $\sigma^2(p \vee q)$ is tight for controlling the operator norm $\|XX^T\|_{\mathrm{op}}^2$ of the non-centered Gram matrix with high probability, (26) implies that the right bound for $\|XX^T - \mathbb{E}[XX^T]\|_{\mathrm{op}}^2$ is rather $\sigma^2\sqrt{pq}$ which is significantly smaller in the regime $p \ll q$ and $\sigma^2 \ll 1$.

In the proof of Theorem 2.2, we could have used those previous results for controlling the matrices of the form $\|XX^T - \mathbb{E}[XX^T]\|_{\mathrm{op}}^2$. However, we would have then achieved a suboptimal risk upper bound. Indeed, Proposition 4.1 plays critical role in the proof of Theorem 2.2, when we need to handle matrices with partial observations that are possibly highly rectangular in the spectal step of the procedure (23).

The proof of Proposition 4.1 relies on the observation that the matrix $XX^T - \mathbb{E}[XX^T]$ is the sum of q centered rank 1 random matrices. This allows us to apply Matrix Bernstein-type concentration inequalities for controlling the operator norm of this sum – see [21] or Section 6 of [23].

Appendix A: Proof of Theorem 2.2

A.1. Notation and signal-noise decomposition

We first introduce some notation, and in particular the noise matrices on which we will apply concentration inequalities. In what follows, we define for any matrix $A \in \mathbb{R}^{n\times d}$, and any vector $w \in \mathbb{R}^d$:

$$\langle A_{i\cdot}, w\rangle = \sum_{k=1}^{d} A_{ik} w_k \quad . \tag{27}$$

If w belongs to $\mathbb{R}^Q$ where Q is some subset of $[d]$, we also write $< A_{i\cdot}, w >= \sum_{k\in Q}^{d} A_{ik} w_k$. The same notation stands for the scalar product on matrices, namely $\langle A, A'\rangle = \mathrm{Tr}(A^T A')$ if $A' \in \mathbb{R}^{n\times d}$. If A and A' are two matrices in $\mathbb{R}^{n\times d}$, then we write the coordinate-wise product $(A \odot A')_{ik} = A_{ik} A'_{ik}$. In what follows, we assume that $\pi^* = \mathrm{id}$. We make this assumption without loss of generality since we can reindex each expert i with $i' = \pi^{*-1}(i)$. Recalling that B is defined in (15) we define

$$\lambda_1 := \mathbb{P}(B_{ik}^{(s)} = 1) = 1 - e^{-\lambda_0} \quad . \tag{28}$$

If $\lambda_0 \le 1$, we have $\lambda_0 \ge \lambda_1 \ge (1 - \frac{1}{e})\lambda_0$. We assume in what follows that $\lambda_0 \le 1$, which corresponds to the case where there are potentially many unobserved coefficients. The case $\lambda_0 \ge 1$ will be treated in Appendix F. For an observation matrix $Y^{(s)}$ defined in (14), we make the difference between $\mathbb{E}[Y^{(s)}] = \lambda_1 M$, which is the unconditional expectation of $Y^{(s)}$, and $\mathbb{E}[Y^{(s)}|B^{(s)}] = B^{(s)} \odot M$, which is the expectation of $Y^{(s)}$ conditionally to the matrix B. We write the noise matrix

$$E^{(s)} = Y^{(s)} - \lambda_1 M \quad \text{and} \quad \widetilde{E}^{(s)} = Y^{(s)} - B^{(s)} \odot M \quad . \tag{29}$$

Recall that $\varepsilon_t = y_t - M_{x_t}$ is the subGaussian noise part in model (2), and that N_s is defined in (14). Each coefficient $\widetilde{E}_{ik}^{(s)}$ can be rewritten as the average of the noise ε_t. that are present in $N^{(s)}$ and that correspond to coefficient $x_t = (i,k)$.

$$\widetilde{E}_{ik}^{(s)} = \sum_{t\in N^{(s)}} \frac{\varepsilon_t}{\mathbf{r}_{ik}^{(s)} \vee 1} \mathbf{1}\{x_t = (i,k)\} \quad . \tag{30}$$

From now on, we often omit the dependence in s. We will extensively use the decomposition $Y = \lambda_1 M + E$, where λ_1 is defined in (28) and E in (29). Recalling that $B_{ik} = \mathbf{1}\{r_{ik} \ge 1\}$, we often rewrite E as the sum of two centered random variables:

$$E_{ik} = (B_{ik} - \lambda_1) M + B_{ik} \widetilde{E}_{ik} \quad .$$

Handling the concentration of the noise is more challenging in the case $\lambda_0 \le 1$ than in the full observation regime $\lambda_0 \ge 1$ discussed in Appendix F. Indeed, while subGaussian concentration inequalities are effective in the full observation regime $\lambda_0 \ge 1$, they lead to slower estimation rate in the case $\lambda_0 \le 1$, for instance in Lemma A.1. Indeed, it turns out that the variance of a coefficient ε_{ik} is of order $\lambda_0 \le 1$, while the hoeffding inequality only implies that $B_{ik} - \lambda_1$, and in particular ε_{ik} are c-subGaussian for some numerical constant c. To overcome this issue, one of the main ideas is to use Bernstein-type bounds on the coefficients of E and on the random matrix $EE^T - \mathbb{E}[EE^T]$- see Lemma B.1 and Proposition 4.1.

A.2. General property on $\mathcal{W}$

Recall that we assume that $\lambda_0 \le 1$, so that $\frac{1}{\lambda_0} \wedge \lambda_0 = \lambda_0$ in (18), and that ϕ_{L_1} is defined in (9) by $\phi_{\mathrm{L}_1} := 10^4 \log(10^2 nd/\delta)$. In the following, we let ξ be the event on which the noise concentrates well for all the pairs (Q, w) considered during the whole procedure. More precisely, we say that we are under event ξ, if for any $s = 0, \dots, 5T-1$ and for any pair (Q, w) that is used to compute a refinement as in (17) we have

$$\left|\langle E_{i\cdot}^{(s)} - E_{j\cdot}^{(s)}, w\rangle\right| \le \tfrac{1}{3}\phi_{\mathrm{L}_1}\sqrt{\lambda_0} \quad \text{for any } (i,j) \in [n]^2 \ . \tag{31}$$

Lemma A.1. *The event ξ holds true with probability at least $1 - 2T\delta$.*

The idea of Lemma A.1 is to apply a bernstein-type inequality and a union bound on all the possible dot products $\langle E_{i\cdot}^{(s)}, w\rangle$, for all the $5T$ possible s and the at most $2T$ possible w. The upper bound is of the order of the square of the variance of E_{ik} up to the polylogarithm factor ϕ_{L_1}. The crucial point is that if $\langle E_{i\cdot}^{(s)}, w\rangle$ is not λ_0-subGaussian, it satisfies the Bernstein's Condition [2.15 of [11]] with variance $\nu = \lambda_0$ and scaling factor $b = \|w\|_\infty$. We then obtain an upper bound of order $\sqrt{\lambda_0}$ since any w considered in the update step (18) must satisfy (16).Recall that $\bar{\gamma}$ is defined in (11). We fix in what follows a sequence $\overline{\gamma} = \gamma_0 > \gamma_1 > \gamma_2 > \cdots > \gamma_{\lfloor 2\log_2(n)\rfloor} = \gamma_{\min}$ in Γ satisfying property (10). We say that u is the level of the corresponding threshold γ_u. We say $\mathcal{W}$ and (γ_u) satisfies the property $\mathcal{C}(\mathcal{W}, (\gamma_u))$ if the following holds

1. **consistency:** For any $(i,j) \in \mathcal{G}(\mathcal{W}, \gamma_{\min})$ it holds that $\pi^*(i) > \pi^*(j)$.
2. **weak-transitivity:** Fix any $u \in \{0, \dots, \lfloor 2\log_2(n)\rfloor - 1\}$. For any experts i, j, k, if i is $\mathcal{G}(\mathcal{W}, \gamma_u)$-above j and $k \in \mathcal{N}(\mathcal{G}(\mathcal{W}, \gamma_{u+1}), j)$, then any $i' \ge i$ is also $\mathcal{G}(\mathcal{W}, \gamma_{\min})$-above k.

The first point of the above property means that at threshold $\gamma_{\min}$, there is no mistake in the directed graph $\mathcal{G}(\mathcal{W}, \gamma_{\min})$, meaning that if there is an edge from i to j in $\mathcal{G}(\mathcal{W}, \gamma_{\min})$, then i is truly above j. Moreover, we only state the consistency property of the graph $\mathcal{G}(\mathcal{W}, \gamma_{\min})$, but this property also implies that, for any more conservative threshold $\gamma \ge \gamma_{\min}$, any $(i,j) \in \mathcal{G}(\mathcal{W}, \gamma)$ satisfies $\pi^*(i) > \pi^*(j)$. This is due to the fact that $\mathcal{G}(\mathcal{W}, \gamma) \subset \mathcal{G}(\mathcal{W}, \gamma_{\min})$. The weak transitivity property states in particular that if there is a path from i to j in the more conservative graph $\mathcal{G}(\mathcal{W}, \gamma_u)$, then there is a path from i to any k in the neighborhood of j at the less conservative threshold $\gamma_{\min}$. The following lemma states that the above property remains true for the weighted graph $\mathcal{W}'$, after any update (18) of the whole procedure.

Lemma A.2. *Under ξ, the property $\mathcal{C}(\mathcal{W}', (\gamma_u))$ holds true for any directed weighted graph $\mathcal{W}'$ obtained at any stage of Algorithm 1 and Algorithm 2.*

We denote in the following $\mathcal{W}_t$ for the directed weighted graph at the begining of step t. For any $u \in [0, \lfloor 2\log_2(n)\rfloor]$, we also write as a short hand $\mathcal{G}_{t,u} = \mathcal{G}(\mathcal{W}_t, \gamma_u)$ for the directed graph at begining of step t and level u and $P_{t,u}(i) = \mathcal{N}(\mathcal{G}_{t,u}, i)$ for the set of experts that are not comparable with i according to $\mathcal{G}_{t,u}$. For any sequence of experts I, we write $\mathcal{P}_{t,u}(I)$ for the sequence of subsets $(P_{t,u}(i))_{i\in I}$. Let us now divide the T steps of the algorithm into $\tau_{\max} = \lfloor \log_2(n)\rfloor + 1$ epochs of $K = \lfloor T/\tau_{\max}\rfloor$ steps. For any $\tau \in [0, \tau_{\max}]$, we also write $\mathcal{G}^K_{\tau,u} = \mathcal{G}_{\tau K,u}$, $P^K_{\tau,u}(i) = P_{\tau K,u}(i)$ and $\mathcal{P}^K_{\tau,u}(I) = \mathcal{P}^K_{\tau,u}(I)$. Now we consider for each epoch τ a sequence of experts $I(\tau) = (i_1(\tau), \dots, i_{L_\tau}(\tau))$ defined by induction:

- $I(0)$ is the empty sequence
- For $\tau \ge 0$, let $(i_1, \dots, i_L)$ be the sequence ordered according to π^* and corresponding to the union of the already constructed sequences $\bigcup_{\tau' \le \tau} I(\tau')$, and $i = 0$, $i_{L+1} = n+1$. For any $l \in [0, L]$, let A_l be the set of experts that are $\mathcal{G}^K_{\tau+1,2\tau+1}$-below i_{l+1} but $\mathcal{G}^K_{\tau+1,2\tau+1}$-above i_l. For all l such that A_l is not empty, we define i'_l as the expert of A_l which is any expert closest to the median $\lfloor (i_l + i_{l+1})/2\rfloor$, and the new sequence $I(\tau+1) := (i'_l)$.

By definition, remark that $I(1)$ is equal to $(\lfloor (n+1)/2 \rfloor)$. The induction step aims at building a sequence $I(\tau+1)$ that is disjoint from $\cup_{\tau' \leq \tau} I(\tau')$, and that cuts each set A_l of experts that are above i_l and below i_{l+1} according to the graph at epoch $\tau+1$ and level $2\tau+1$. Given the already constructed collections of perfectly ordered experts $I(\tau')$ for $\tau' \leq \tau$, the idea of $I(\tau+1)$ is that it tends to fill the gaps between the neighborhoods in $\mathcal{G}_{\tau+1,2\tau+1}$ of any two successive experts in $\cup_{\tau' \leq \tau} I(\tau')$.

By monotonicity, it holds that for any expert i, epoch τ and level u that $P^K_{\tau+1,u+1}(i) \subset P^K_{\tau+1,u}(i) \subset P^K_{\tau,u}(i)$. We say that the sets $P^K_{\tau,2\tau}(i)$ and $P^K_{\tau,2\tau+1}(i)$ are the neighborhoods of i at the beginning of epoch τ and that the sets $P^K_{\tau+1,2\tau}(i), P^K_{\tau+1,2\tau+1}(i)$ are the neighborhood of i at the end of epoch τ. The neighborhoods at the end of a given epoch τ are obtained from the neighborhoods at the beginning the of epoch τ after K steps of the Algorithm 1. On the other hand, we say that the sets $P^K_{\tau,2\tau}, P^K_{\tau+1,2\tau}$ are the conservative subsets at epoch τ, since they correspond to a more conservative directed graph with threshold $\gamma_{2\tau} \geq \gamma_{2\tau+1}$. The following lemma states that, at any epoch τ, the conservative subsets at the beginning of epoch τ are well separated according to the true order $\pi^* = \mathrm{id}$:

Lemma A.3. *Under event ξ, for any $\tau \in [0, \tau_{\max}]$, letting $(i_1, \dots, i_L) = I(\tau)$, we have*

$$P^K_{\tau,2\tau}(i_1) < \cdots < P^K_{\tau,2\tau}(i_L).$$

In other words, Lemma A.3 implies that, for any $l < l'$, any expert in $P^K_{\tau,2\tau}(i_l)$ is π^*-below any expert in $P^K_{\tau,2\tau}(i_{l'})$. As a consequence, it holds that for any $l \in [1, L_\tau - 2]$,

$$P^K_{\tau,2\tau}(i_l) \overset{\mathcal{G}^K_{\tau,2\tau}}{<} P^K_{\tau,2\tau}(i_{l+2}) \ . \tag{32}$$

Namely, any expert in $P^K_{\tau,2\tau}(i_l)$ is $\mathcal{G}^K_{\tau,2\tau}$-below any expert in $P^K_{\tau,2\tau}(i_{l+2})$. Indeed, Lemma A.3 and first point of event ξ imply that any expert j in $P^K_{\tau,2\tau}(i_l)$ is $\mathcal{G}^K_{\tau,2\tau}$-below i_{l+1}, since j cannot be in $P^K_{\tau,2\tau}(i_l)$. On the other hand, i_{l+1} is itself $\mathcal{G}^K_{\tau,2\tau}$-below any expert of $P^K_{\tau,2\tau}(i_{l+2})$ for the same reason. The following lemma states that the ending less conservative subsets are covering the set of all experts.

Lemma A.4. *Under event ξ, it holds that*

$$[n] = \bigcup_{\tau=0}^{\tau_{\max}-1} \bigcup_{i \in I(\tau)} P^K_{\tau+1,2\tau+1}(i) \ .$$

Let $\hat{\pi}$ be the estimator obtained from the final weighted directed graph $\mathcal{W}$ at the end of the procedure, that is any permutation on $[n]$ that is consistent with the largest acyclic graph of the form $\mathcal{G}(\mathcal{W}, \gamma)$ for all $\gamma > 0$. For any sequence of subsets $\mathcal{P} = (P_1, \dots, P_L)$ we define

$$\mathrm{SN}(\mathcal{P}) = \sum_{P \in \mathcal{P}} \|M(P) - \overline{M}(P)\|_F^2 \ . \tag{33}$$

The following proposition that we can control the L_2 error of $\hat{\pi}$ by the maximum over all epoch τ of the sum over τ of the square norms of the groups in $\mathcal{P}^K_{\tau+1,2\tau+1}$.

Proposition A.5. *Under event ξ, it holds that*

$$\|M_{\hat{\pi}^{-1}} - M\|_F^2 \leq 4 \sum_{\tau=0}^{\tau_{\max}-1} \mathrm{SN}(\mathcal{P}^K_{\tau+1,2(\tau+1)}) \ . \tag{34}$$

Recall that $\bar{\gamma}$ is defined in (11), and that Γ can be taken to be a valid grid with $\bar{\gamma}$ smaller than a polylogarithm in n, d, δ. The final proposition states that at any level u and any step t, any sequence of subset that can be ordered according to the already constructed graph $\mathcal{G}_{t,u}$ as in (32) will either have a square norm smaller than the minimax rate ρ_{perm}, defined in (6) or almost exponentially decrease its square norm with high probability.

Proposition A.6. *Fix any $u \in [0, 2\tau_{\max}]$ and step $t < T$, and assume that $I = (i_1, \dots, i_L)$ is a sequence of experts that satisfies $P_{t,u}(i_1) \overset{\mathcal{G}_{t,u}}{\prec} \dots \overset{\mathcal{G}_{t,u}}{\prec} P_{t,u}(i_L)$. Then on the intersection of the event ξ (defined in (31)) and an event of probability higher than $1 - 5\delta$, it holds that*

$$\text{SN}(\mathcal{P}_{t+1,u}(I)) \le \left[C\bar{\gamma}^6 \rho_{\text{perm}}(n, d, \lambda_0)\right] \vee \left[\left(1 - \frac{1}{4\bar{\gamma}^2}\right) \text{SN}(\mathcal{P}_{t,u}(I))\right] \ ,$$

for some numerical constant C.

Let us fix $\tau \in \{0, \dots, \tau_{\max} - 1\}$. Applying Proposition A.6 for each $t = K\tau, \dots, K\tau + K - 1$ and $u = 2(\tau + 1)$ -the hypothesis of Proposition A.6 being satisfied by (32), we obtain with probability $1 - 5(K + T)\delta$ that

$$\begin{aligned} \text{SN}(\mathcal{P}_{\tau+1,2(\tau+1)}) &\le \left[C\bar{\gamma}^6 \rho_{\text{perm}}(n, d, \lambda_0)\right] \vee e^{-\frac{T}{4\tau_{\max}\bar{\gamma}^4}} nd \\ &\le CT\bar{\gamma}^6 \rho_{\text{perm}}(n, d, \lambda) \ , \end{aligned}$$

if T is larger than $4\bar{\gamma}^6 \ge 4\log^2(nd)\bar{\gamma}^4$. We conclude the proof of Theorem 2.2 with Proposition A.5, using that $4\tau_{\max} \le \bar{\gamma}$:

$$\|M_{\hat{\pi}^{-1}} - M\|_F^2 \le 4 \sum_{\tau=0}^{\tau_{\max}-1} \text{SN}(\mathcal{P}^K_{\tau+1,2(\tau+1)}) \le CT\bar{\gamma}^7 \rho_{\text{perm}}(n, d, \lambda) \ .$$

Appendix B: Proofs of the lemmas of Appendix A and of Proposition A.5

B.1. Proof of Proposition A.5

Let $\hat{\pi}$ be any arbitrary permutation that is consistent with the largest **DAG** $\mathcal{G}(\mathcal{W}, \bar{\gamma})$, as defined in Section 3.1. Recall that we assume in this proof that $\pi^* = \text{id}$. By Lemma A.4, for any $i \in [n]$ there exists $\tau \in [0, \tau_{\max} - 1]$ and $i_0 \in I(\tau)$ such that $i \in P^K_{\tau+1,2\tau+1}(i_0)$.

Let us define the interval $[a, b]$ as the maximal interval containing i_0 and that is included in the more conservative set $P^K_{\tau+1,2\tau}$. Now, if $j > b$, then by definition there exists j' such that $j \ge j' > b$ and $j' \notin P^K_{\tau+1,2\tau}$. Summarizing the properties, we have $j \ge j' \overset{\mathcal{G}_{\tau+1,2\tau}}{\succ} i_0$, and that i is in the neighborhood of i_0 in the graph $\mathcal{G}_{\tau+1,2\tau+1}$. Hence, applying the weak-transitivity property (first in $\mathcal{C}$), holding true on event ξ - see Lemma A.2, we obtain that j is $\mathcal{G}(\mathcal{W}_{K(\tau+1)}, \gamma_{\min})$-above i. By the consistency property (second point in $\mathcal{C}$), j is also necessarily $\mathcal{G}(\mathcal{W}, \gamma_{\min})$-above i, and this proves that all the $n - b$ experts j satisfying $j > b$ are $\mathcal{G}(\mathcal{W}, \gamma_{\min})$ above i. Hence it holds that $\hat{\pi}(i) \le b$. By symmetry, we also prove that $\hat{\pi}(i) \ge a$, so that

$$\hat{\pi}(i) \in [a, b] \subset P^K_{\tau+1,2\tau}(i_0). \tag{35}$$

Finally, we have

$$\begin{aligned}
\|M_{\hat{\pi}^{-1}} - M\|_F^2 &= \sum_{i=1}^{n} \|M_{\hat{\pi}(i)\cdot} - M_{i\cdot}\|_F^2 \\
&\le \sum_{\tau=0}^{\tau_{\max}} \sum_{i_0 \in I(\tau)} \sum_{i \in P^K_{\tau+1,2\tau+1}(i_0)} \|M_{\hat{\pi}(i)\cdot} - M_{i\cdot}\|^2 \\
&\le 2 \sum_{\tau=0}^{\tau_{\max}} \sum_{i_0 \in I(\tau)} \sum_{i \in P^K_{\tau+1,2\tau+1}(i_0)} \|M_{i\cdot} - \overline{m}(P^K_{\tau+1,2\tau}(i_0))\|^2 + \|M_{\hat{\pi}(i)\cdot} - \overline{m}(P^K_{\tau+1,2\tau}(i_0))\|^2 \\
&\le 4 \sum_{\tau=0}^{\tau_{\max}} \sum_{i_0 \in I(\tau)} \sum_{i \in P^K_{\tau+1,2\tau}(i_0)} \|M_{i\cdot} - \overline{m}(P^K_{\tau+1,2\tau}(i_0))\|^2 \ ,
\end{aligned}$$

where we used Lemma A.4 for the first inequality and (35) for the last inequality.

B.2. Proof of the lemmas of Appendix A

We postpone the proof of Lemma A.1 to the next subsection.

Proof of Lemma A.2. Recall that we consider the case $\lambda_0 \le 1$, so that $\lambda_0 \wedge 1/\lambda_0 = \lambda_0$ in (18).

Consider any substep of the whole procedure where the current directed weighted graph is $\mathcal{W}'$. For the first point, remark that i is $\mathcal{G}(\mathcal{W}', \gamma_{\min})$-above j only if there exists a previous substep during which we find out that $\langle Y_{i\cdot} - Y_{j\cdot}, w\rangle \ge \gamma_{\min}$ on some direction $w \in \mathbb{R}^Q$, where Y is the sample used to refine the edges (17). Since $\gamma_{\min} > \phi_{\mathrm{L}_1}$, then decomposing $Y = \lambda_1 M + E$ as in (29), we have

$$\lambda_1 \langle M_{i\cdot} - M_{j\cdot}, w\rangle \ge \langle Y_{i\cdot} - Y_{j\cdot}, w\rangle - \langle E_{i\cdot} - E_{j\cdot}, w\rangle > 0 \ , \tag{36}$$

where the last inequality comes from (31), using the notation (27). Since the coefficients of w are nonegative, we have proven that i is above j. For the second point, assume that i is $\mathcal{G}(\mathcal{W}, \gamma_u)$-above j, and take $i' \ge i$. As before, there exists a direction w used during the procedure such that $\langle Y_{i\cdot} - Y_{j\cdot}, w\rangle \ge \gamma_u$. Now consider any $k \in \mathcal{N}(\mathcal{G}(\mathcal{W}, \gamma_{u+1}), j)$. On the direction w, we have under the event ξ defined in (31) that

$$\begin{aligned}
\langle Y_{i'\cdot} - Y_{k\cdot}, w\rangle &\ge \lambda_1 \langle M_{i'\cdot} - M_{k\cdot}, w\rangle - \tfrac{1}{3}\phi_{\mathrm{L}_1}\sqrt{\lambda_0} \\
&\ge \lambda_1 \langle M_{i\cdot} - M_{k\cdot}, w\rangle - \tfrac{1}{3}\phi_{\mathrm{L}_1}\sqrt{\lambda_0} \\
&\ge \langle Y_{i\cdot} - Y_{j\cdot}, w\rangle - \langle Y_{k\cdot} - Y_{j\cdot}, w\rangle - \phi_{\mathrm{L}_1}\sqrt{\lambda_0} \\
&\ge (\gamma_u - \gamma_{u+1} - \phi_{\mathrm{L}_1})\sqrt{\lambda_0} \ge \gamma_{\min}\sqrt{\lambda_0} \ ,
\end{aligned}$$

where the last inequality comes from the assumption (10). We conclude that i' is $\mathcal{G}(\mathcal{W}', \gamma_{\min})$-above k. □

Proof of Lemma A.3. We proceed by induction over $\tau \ge 0$. The lemma is trivial for $\tau = 0, 1$ since $I(0)$ is empty and $I(1) = (\lfloor (n+1)/2 \rfloor)$. Let $\tau \ge 1$ and i_1, i_2, i_3 be three experts in $I(\tau) \cup \{0, n+1\}$ such that $i_1 < i_2 < i_3$. Let A be the set of experts that are $\mathcal{G}^K_{\tau+1,2\tau+1}$-above i_1 and $\mathcal{G}^K_{\tau+1,2\tau+1}$-below i_2, and A' be the set of experts that are $\mathcal{G}^K_{\tau+1,2\tau+1}$-above i_2 and $\mathcal{G}^K_{\tau+1,2\tau+1}$-below i_3. Assume that both sets A and A' are nonempty, and let $j \in A$ and $j' \in A'$. Let us apply the weak-transitivity of $\mathcal{W}, (\gamma_u)$ in Property $\mathcal{C}$ - which holds true under ξ from Lemma A.2 - with $u = 2\tau+1$. Since j is $\mathcal{G}^K_{\tau+1,2\tau+1}$-below i_2, any $k \in P^K_{\tau+1,2(\tau+1)}(j)$ is π^*-below i_2. We also prove that any $k' \in P^K_{\tau+1,2(\tau+1)}(j')$ is π^*-above i_2. We conclude that $P^K_{\tau+1,2(\tau+1)}(j) < P^K_{\tau+1,2(\tau+1)}(j')$, and the proof of the lemma follows. □

Proof of Lemma A.4. We prove that, by construction, any expert $i \in [n]$ is at distance less than $(n+1)/2^{\tau+1}$ of $\bigcup_{\tau'=0}^{\tau}\bigcup_{i\in I(\tau)} P^K_{\tau+1,2\tau+1}(i) \cup \{0, n+1\}$. This is obvious for $\tau = 0$ since any expert is at distance less than $(n+1)/2$ of 0 or $n+1$. Let $(i_1, \dots, i_L) = \bigcup_{\tau'\le\tau} I(\tau')$ be the collection of experts in the union of all possible $I(\tau')$ that is ordered according to π^*. If j is any expert in $[n]$, then we let $l \in [0, L]$ be such that $i_l \le j \le i_{l+1}$. We can assume that $j \notin P^K_{\tau+1,2\tau+1}(i_l)$ and $j \notin P^K_{\tau+1,2\tau+1}(i_{l+1})$ because otherwise the distance of j to $\bigcup_{\tau'=0}^{\tau}\bigcup_{i\in I(\tau)}^{L} P^K_{\tau+1,2\tau+1}(i)$ is 0. Using property $\mathcal{C}$ holding true from Lemma A.2, it holds that the set A of experts that are $\mathcal{G}^K_{\tau+1,2\tau+1}$-above i_l but $\mathcal{G}^K_{\tau+1,2\tau+1}$-below i_{l+1} contains j and therefore is nonempty. Now, let $m = \lfloor (i_l + i_{l+1})/2 \rfloor$ and i' be any expert closest to m in A, as defined in the construction of $I(\tau+1)$, and assume without loss of generality that $m \le i'$. We consider the following cases:

- $m \le i' \le j$: In that case, j is at distance less than $(i_{l+1} - m)/2$ of i' or i_{l+1}.
- $m \le j < i'$: This case is not possible since i' is the closest expert to m in A.
- $j < m < i'$: In that case, since i' minimizes the distance to m, we necessarily have that $m \in P^K_{\tau+1,2\tau+1}(i_l) \cup P^K_{\tau+1,2\tau+1}(i_{l+1})$. Hence j is at distance less than $(m - i_l)/2$ of m or i_l.

We have proved that the distance of any j to $\bigcup_{\tau'=0}^{\tau+1}\bigcup_{i\in I(\tau)}^{L} P^K_{\tau+1,2\tau+1}(i) \cup \{0, n+1\}$ is at most $(m - i_l)/2$ or $(i_{l+1} - m)/2$. Using the induction hypothese, we have that $m - i_l$ and $i_{l+1} - m$ are both less than $n/2^{\tau+1}$, which concludes the induction.

Finally, applying this property with $\tau_{\max} - 1 = \lfloor \log_2(n) \rfloor$ gives a distance strictly smaller than 1, which proves the result. □

B.3. Proof of Lemma A.1

Let us start with the following lemma, which gives a concentration bound when $\lambda_0 \le 1$:

Lemma B.1. *For any $\delta' > 0$ and for any matrix $W \in \mathbb{R}^{n\times d}$, the following inequality holds with probability at least $1 - \delta'$:*

$$|\langle E, W\rangle| \le \sqrt{4e^2 \|W\|_F^2 \lambda_0 \log\left(\frac{2}{\delta'}\right)} + \|W\|_\infty \log\left(\frac{2}{\delta'}\right) \ . \tag{37}$$

Now we apply Lemma B.1 with the matrix W with 0 coefficients except at line i where it is equal to the vector $\frac{w}{\|w\|_2}$ as defined in (17) and we deduce that

$$|\langle E_{i,\cdot}, \frac{w}{\|w\|_2}\rangle| \le \sqrt{4e^2 \lambda_0 \log\left(\frac{2}{\delta'}\right)} + \frac{\|w\|_\infty}{\|w\|_2} \log\left(\frac{2}{\delta'}\right) \le 11\sqrt{\lambda_0} \log(2/\delta') \ , \tag{38}$$

where the last inequality comes from Condition (16) on w. Now choosing $\delta' = \delta/(4Tn^6)$, a union bound over the at most $2n^2T|\mathcal{H}|(|\Gamma| \wedge n^2)$ pairs (Q, w) considered during the procedure, we deduce the bound of Lemma A.1 for all $\lambda_0 \le 1$.

Proof of Lemma B.1. Recall that $E, \widetilde{E}$ are defined in (29) and that we have in particular

$$E_{ik} = (B_{ik} - \lambda_1) M_{ik} + \widetilde{E}_{ik} \ .$$

Let $x > 0$. By Cauchy-Schwarz inequality, we have

$$\mathbb{E}[e^{xE_{ik}}] \le \sqrt{\mathbb{E}[e^{2x(B_{ik}-\lambda_1)M_{ik}}]}\sqrt{\mathbb{E}[e^{2x\widetilde{E}_{ik}}]} \ ,$$

where we recall that $\lambda_1 = 1 - e^{-\lambda_0} \le \lambda_0$. We have

$$\mathbb{E}[e^{2x(B_{ik}-\lambda_1)M_{ik}}] \le e^{-2\lambda_1 x M_{ik}}(\lambda_1(e^{2xM_{ik}} - 1) + 1) \le e^{\lambda_1 e^2 x^2} \ ,$$

and

$$\mathbb{E}[e^{2x\widetilde{E}_{ik}}] \le \lambda_1(e^{2x^2}-1)+1 \le e^{\lambda_1 e^2 x^2} \ ,$$

where we used the inequalities $e^{2x^2}-1 \le e^2x^2$ and $e^{2x}-1-2x \le e^2x^2$ for any $x \in [-1,1]$.

In particular, if $t>0$, a Chernoff bound with $x = \frac{t}{2\|W\|_F^2\lambda_0 e^2} \wedge 1$ gives

$$\mathbb{P}(\langle W,E\rangle \ge t) \le \exp(-(\tfrac{t^2}{4\|W\|_F^2\lambda_0 e^2} \wedge t)) \ ,$$

so that with probability at least $1-\delta'$:

$$|\langle W,E\rangle| \le \sqrt{4e^2\|W\|_F^2\lambda_0 \log\left(\frac{2}{\delta'}\right)} + \log\left(\frac{2}{\delta'}\right) \ .$$

□

Appendix C: Proof of Proposition A.6

Step 0 : general definitions

In this proof, we fix $u \in \{0,\dots,2\lfloor\log_2(n)\rfloor+2\}$ and a corresponding threshold γ_u in the sequence in Γ satisfying $\gamma_u \ge \phi_{\mathrm{L}_1}$ - see (10) - and a step $t<T$. We assume that $I=(i_1,\dots,i_L)$ is a fixed sequence of experts that satisfies $P_{t,u}(i_1) \overset{\mathcal{G}_{t,u}}{\prec} \dots \overset{\mathcal{G}_{t,u}}{\prec} P_{t,u}(i_L)$.

From now on, we ease the notation by omitting the dependence in t,u,γ_u and we write $\mathcal{G}=\mathcal{G}_{t,u}$, $\mathcal{G}'=\mathcal{G}_{t+1,u}$, $\mathcal{P}=(P_1,\dots,P_L)$ for $\mathcal{P}_{t,u}$ and $\mathcal{P}'$ for $\mathcal{P}_{t+1,u}$. We denote $\widetilde{\mathcal{G}}^h$ for the directed graph at threshold γ_u of the directed weighted graph $\widetilde{\mathcal{W}}^h$ obtained at the end the first update Line 3 of Algorithm 2. We also write $\widetilde{P}_l^h = \mathcal{N}(\widetilde{\mathcal{G}}^h, i_l)$ and $\widetilde{\mathcal{P}}^h = (\widetilde{P}_1^h,\dots,\widetilde{P}_L^h)$ for the corresponding sequence of subsets at height $h \in \mathcal{H}$. By monotonicity, it holds for any $h\in\mathcal{H}$ that

$$P_l' \subset \widetilde{P}_l^h \subset P_l \ .$$

C.1. Step 1: Analysis of the selected set $\widehat{Q}$

Recall the definition of the neighborhoods (20) of the set P_l in the graph $\mathcal{G}$:

$$\mathcal{N}_a(l) = \bigcap_{i\in P_l} \mathbf{rk}_{\mathcal{G},i_l}^{-1}([1,a]) \quad \text{and} \quad \mathcal{N}_{-a}(l) = \bigcap_{i\in P_l} \mathbf{rk}_{\mathcal{G},i_l}^{-1}([-1,-a]) \ ,$$

Define for $\kappa>0$ and $l\in[1,L]$ the population version $\mathbf{\Delta}_k^*$ of the width statistic $\widehat{\mathbf{\Delta}}_k$ - see (21) - as the difference of the best and worst expert of $P(i_l)$ if $a=0$ and as the difference of the average of the experts in $\mathcal{N}_a(l)$ and the average of the expert in $\mathcal{N}_{-a}(l)$:

$$\mathbf{\Delta}_k^*(0,l) = \max_{i,j\in P(i_l)} |M_{i,k}-M_{j,k}| \quad \text{and} \quad \mathbf{\Delta}_k^*(a,l) = \overline{m}_k(\mathcal{N}_a(l)) - \overline{m}_k(\mathcal{N}_{-a}(l)) \text{ if a} \ge 1. \quad (39)$$

We also define $a^*(h,l)$ as the minimum $a\ge 1$ such that there are at least $\frac{1}{\lambda_0 h^2}$ experts in $\mathcal{N}_a(l)$ and in $\mathcal{N}_{-a}(l)$:

$$a^*(h,l) = \min\{a\ge 1 \ : \ |\mathcal{N}_a(l)| \wedge |\mathcal{N}_{-a}(l)| \ge \frac{1}{\lambda_0 h^2}\} \ . \quad (40)$$

Now, define for $\phi \geq 1$:

$$\begin{aligned} Q_l^{*h}(\phi) &:= \{k \in [d] \ : \ \mathbf{\Delta}_k^*(0,l) \in [\phi h, 2\phi h]\} \\ \overline{Q}_l^{*h}(\phi) &:= \{k \in [d] \ : \ \mathbf{\Delta}_k^*(a^*(\phi^{-1}h, l), l) \geq h/2\} \ . \end{aligned} \tag{41}$$

The following lemma states that, for ϕ of order $\log(nd/\delta)$, we can sandwich $\widehat{Q}_l^h$ between the two fixed sets Q_l^{*h} and $\overline{Q}_l^{*h}$:

Lemma C.1. *Let l be a fixed index in $\{1, \dots, L\}$ and h a fixed height in $\mathcal{H}$. There exists a numerical constant $\kappa_0 > 0$ such that, with probability at least $1 - \delta/(L|\mathcal{H}|)$, we have*

$$Q_l^{*h}(\kappa_0 \log(nd/\delta)) \subset \widehat{Q}_l^h \subset \overline{Q}_l^{*h}(\kappa_0 \log(nd/\delta)) \ . \tag{42}$$

C.2. Step 2 : l1-control of the intermediary sets $\widetilde{\mathcal{P}}^h$

Recall that γ_u is a threshold corresponding to a sequence in Γ as defined in (10). For any sets $P \subset [n], Q \subset [d]$, we say that $M(P,Q)$ is indistinguishable in L_1-norm if it satisfies

$$\max_{i,j \in P} \|M_{i\cdot}(P,Q) - M_{j\cdot}(P,Q)\|_1 \leq 3\gamma_u \sqrt{\frac{|Q|}{\lambda_0}} \ . \tag{43}$$

For a fixed $l \in \{1, \dots, L\}$, let $\xi_{\mathrm{L}_1}(l,h)$ be the event under which $M(\widetilde{P}_l^h, \widehat{Q}_l^h)$ is indistinguishable in L_1-norm.

Lemma C.2. *Let l be a fixed index in $\{1, \dots, L\}$ and $h \in \mathcal{H}$ such that $\lambda_0 |Q_l^{*h}| \geq 1$. The event $\xi_{\mathrm{L}_1}(l,h)$ holds true with probability at least $1 - \delta/(L|\mathcal{H}|)$.*

Let κ_0 be a numerical constant given by Lemma C.1 and let $\phi_0 = \kappa_0 \log(nd/\delta)$. In what follows, we write for simplicity $(Q_l^{*h}, \widehat{Q}_l^h, \overline{Q}_l^h) = (Q_l^{*h}(\phi_0), \widehat{Q}_l^h(\phi_0), \overline{Q}_l^h(\phi_0))$. Lemma C.2 provides an upper bound only on the L_1 distance between rows of M restricted to the subsets $\widetilde{P}_l^h$ and $\widehat{Q}_l^h$, while the square norm of a group (33) is defined with the L_2 distance. with (43). The idea is that for any k in Q^{*h}, and for any $i \in \widetilde{P}^h$, we have that $|M_{ik} - \overline{m}_k|^2 \leq 2\phi_0 h |M_{ik} - \overline{m}_k|$. In particular, $\|M_{i\cdot}(\widetilde{P}_l^h, Q_l^{*h}) - \overline{m}_\cdot(\widetilde{P}_l^h, Q_l^{*h})\|_2^2 \leq 2\phi_0 h \|M_{i\cdot}(\widetilde{P}_l^h, Q_l^{*h}) - \overline{m}_\cdot(\widetilde{P}_l^h, Q_l^{*h})\|_1$. Hence, it holds from Lemma C.1, Lemma C.2 and a union bound over all $l \in \{1, \dots, L\}$ and all $h \in \mathcal{H}$ satisfying $\lambda_0 |Q_l^{*h}| \geq 1$ that with probability at least $1 - 2\delta$,

$$\sum_{i \in \widetilde{P}_l^h} \|M_{i\cdot}(\widetilde{P}_l^h, Q_l^{*h}) - \overline{m}_\cdot(\widetilde{P}_l^h, Q_l^{*h})\|_2^2 \leq 6\phi_0 \gamma_u \left[h |\widetilde{P}_l^h| \sqrt{\frac{|\overline{Q}_l^{*h}|}{\lambda_0}} \right] \ , \tag{44}$$

simultaneously for all $l \in \{1, \dots, L\}$ and $h \in \mathcal{H}$ satisfying $\lambda_0 |Q_l^{*h}| \geq 1$.

Proof of Lemma C.2. Let l be a fixed index in $\{1, \dots, L\}$ and h be a fixed height in $\mathcal{H}$. If $a \geq 1$, the subset P_l is disjoint from the sets $\mathcal{N}_a(l) \cup \mathcal{N}_{-a}(l)$ so that $\widehat{Q}_l^h$ is independent of $Y^{(1)}(P_l)$. Remark also that condition (16) is satisfied since $\lambda_0 |Q_l^{*h}| \geq 1$ and $Q_l^{*h} \subset \widehat{Q}_l^h$.

Recall that we assume that $\lambda_0 \leq 1$. We write $w = \mathbf{1}_{\widehat{Q}_l^h}$ and we recall that $B = (B_{ik})$ is the matrix defined in (15). Let $i, j \in \widetilde{P}_l^h$ so that, by definition, we have that $|\langle Y_{i\cdot} - Y_{j\cdot}, w \rangle| \leq \gamma_u \sqrt{\lambda_0 |\widehat{Q}_l^h|}$. With probability at least $1 - \delta/L$, for all i, j in P_l we have that

$$\lambda_1 \, |\langle M_{i\cdot} - M_{j\cdot}, w \rangle| \leq |\langle Y_{i\cdot} - Y_{j\cdot}, w \rangle| + |\langle E_{i\cdot} - E_{j\cdot}, w \rangle| \leq (\gamma_u + \phi_{\mathrm{L}_1}/2) \sqrt{\lambda_0 |\widehat{Q}_l^h|} \ . \tag{45}$$

where the last inequality comes from Lemma B.1 applied with $\delta' = \delta/n^3$ and from the definition of ϕ_{L_1} (9). Recalling the two inequalities $\lambda_1 = 1 - e^{-\lambda_0} \geq \lambda_0/2$ and $\phi_{\mathrm{L}_1} \leq \gamma_u$, we obtain the result. □

C.3. Step 3 : Local square norm reduction

Henceforth we condition to the sample $Y^{(1)}$ of Algorithm 2 which allows us to assume that, for any $h \in \mathcal{H}$, the two sequences of sets $\widetilde{\mathcal{P}}^h$ and $\widehat{\mathcal{Q}}^h$ are fixed.

For $\kappa_1 > 0$, let $\xi_{\mathrm{loc}}(l,h,\kappa_1)$ be the event holding true if the local square norm of $M(P_l, \widehat{Q}_l^h)$ has decreased at the end of Algorithm 2, that is

$$\begin{aligned}\|M(P_l', \widehat{Q}_l^h) - \overline{M}(P_l', \widehat{Q}_l^h)\|_F^2 \le \kappa_1 \gamma_u^4 \left[\frac{1}{\lambda_0}\sqrt{|P_l||\widehat{Q}_l^h|} + \frac{|P_l|}{\lambda_0}\right] \\ \vee \left(1 - \frac{1}{4\gamma_u^2}\right)\|M(P_l, \widehat{Q}_l^h) - \overline{M}(P_l, \widehat{Q}_l^h)\|_F^2 \ .\end{aligned} \tag{46}$$

The following proposition states that given the fact that the experts in $\widetilde{P}_l^h$ are indistinguishable in L_1-norm and $\lambda_0(|\widetilde{P}_l^h| \wedge |Q_l^{*h}|) \ge 1$, the event ξ_{loc} holds true simultaneously for all l and h with high probability.

Proposition C.3. *There exists a numerical constant κ_1 such that the following holds, for any fixed index l in $\{1,\dots,L\}$, and fixed height h in $\mathcal{H}$. Conditionally to $Y^{(1)}$, the event $\xi_{\mathrm{L}_1}(l)$ and $\lambda_0(|\widetilde{P}_l^h| \wedge |Q_l^{*h}|) \ge 1$, the event $\xi_{\mathrm{loc}}(l,h,\kappa_1)$ holds true with probability at least $1 - 3\delta/(L|\mathcal{H}|)$.*

Proposition C.3 is at the core of the analysis, and its proof contains a significant part of the arguments. This proposition and its proof are similar to Proposition D.5 in [14], but the main difficulty with respect to [14] is that we do not achieve the optimal rate in $\lambda_0 \le 1$ using only the subgaussianity of the coefficients of the noise E. A key step in the proof of Proposition C.3 is Proposition 4.1, which implies Lemma E.2 and gives a concentration inequality of the operator norm of $EE^T - \mathbb{E}[EE^T]$. Proposition 4.1 is effective in that case since the coefficients of E will be proven to satisfy (24).

Then, the idea is that when a group P_l' has a square norm of order at least $\frac{1}{\lambda_0}\sqrt{|P_l||\widehat{Q}_l^h|} + \frac{|P_l|}{\lambda_0}$, the PCA-based procedure defined as in (23) will output a vector $\hat{v}$ that is well aligned with the first left singular vector of $M(\widetilde{P}_l^h, \widehat{Q}_l^h) - \overline{M}(\widetilde{P}_l^h, \widehat{Q}_l^h)$. Moreover, the isotonic structure of $M(\widetilde{P}_l^h, \widehat{Q}_l^h) - \overline{M}(\widetilde{P}_l^h, \widehat{Q}_l^h)$ implies in fact that its operator norm is greater than a polylogarithmic fraction of its Frobenius norm (see Lemma E.1 or Lemma E.4 in [14]], so that $\|\hat{v}^T(M(\widetilde{P}_l^h, \widehat{Q}_l^h) - \overline{M}(\widetilde{P}_l^h, \widehat{Q}_l^h))\|_2^2$ is of the same order as the square Frobenius norm. Hence after updating the edges, we can prove that the experts in $\widetilde{P}_l^h \setminus P_l'$ were contributing significantly to the Frobenius norm, which enforces the contraction part in the second term of the maximum in (46). All the details of the proof can be found in Appendix E.

C.4. Step 4 : Control of the size of the sets $\overline{Q}_l^{*h}$

For any $p \in [n] \cap \{2^k \ : \ k \in \mathbb{Z}^+\}$, let $\mathcal{L}(p)$ be the sets of indices $l = 1,\dots,L$ whose corresponding group size $|P_l|$ belongs to $[p, 2p)$. The two upper bounds implied by (44) and (46) both depend on the selected subset of columns, which is included in $\overline{Q}_l^{*h}$ under the event of Lemma C.1. The following lemma provides an upper bound on the sum over $l \in \mathcal{L}(p)$ of the size of the sets $\overline{Q}_l^{*h}(\phi)$ defined in (41), for any $\phi > 0$.

Lemma C.4. *For any $\phi \ge 1$ and any $h \in \mathcal{H}$, it holds that*

$$\sum_{l \in \mathcal{L}(p)} |\overline{Q}^{*h}(\phi)| \le 12\phi^2 \left(\frac{1}{p\lambda_0 h^2} \vee 1\right)\frac{d}{h} \ .$$

The proof of Lemma C.4 is mainly implied by the fact that the coefficients of M are bounded by 1. Then, the idea is that in the case where all the sets P_l are of size p, it is enough to take

a number of group a of order at most $\frac{1}{p\lambda_0 h^2} \vee 1$ above and below each P_l to ensure that the corresponding neighborhood of P_l has size $|\mathcal{N}_a(l)| \wedge |\mathcal{N}_{-a}(l)| \geq \frac{1}{\lambda_0 h^2}$.

C.5. Step 5 : Conclusion of the previous steps

We first decompose the square norm $\mathrm{SN}(\mathcal{P})$ as defined in (33) into two terms. Assume that the event of Lemma C.1, $\xi_{\mathrm{L}_1}(l)$ and $\xi_{\mathrm{loc}}(l,h,\kappa_1)$ - see Lemma C.2 and Proposition C.3 - hold true. Define $\mathcal{L}_-$ as the sequence of indices l such that the corresponding reduced subsets P'_l have low local square norm for all $h \in \mathcal{H}$. More precisely, we say that $l \in \mathcal{L}_-$ if for all $h \in \mathcal{H}$ we have

$$\begin{aligned}\|M(P'_l,\widehat{Q}^h_l) - \overline{M}(P'_l,\widehat{Q}^h_l)\|_F^2 \leq \kappa_1\gamma_u^4\left[\frac{1}{\lambda_0}\sqrt{|P_l\|\widehat{Q}^h_l|} + \frac{|P_l|}{\lambda_0}\right] \\ \vee \frac{1}{2|\mathcal{H}|}\|M(P_l) - \overline{M}(P_l)\|_F^2 \ .\end{aligned} \tag{47}$$

We also define the complementary $\mathcal{L}_+ = [1,L] \setminus \mathcal{L}_-$ and their corresponding subsets $\mathcal{P}'_+, \mathcal{P}'_-$ in $\mathcal{P}'$. We have the following decomposition:

$$\mathrm{SN}(\mathcal{P}') = \mathrm{SN}(\mathcal{P}'_+) + \mathrm{SN}(\mathcal{P}'_-) \ . \tag{48}$$

Let us now give an upper bound of $\mathrm{SN}(\mathcal{P}'_+)$. For any $l \in \mathcal{L}_+$, there exists by definition an element $h_l \in \mathcal{H}$ such that $\|M(P'_l,\widehat{Q}^{h_l}_l) - \overline{M}(P'_l,\widehat{Q}^{h_l}_l)\|_F^2 > \kappa_1\gamma_u^4\left[\frac{1}{\lambda_0}\sqrt{|P_l\|\widehat{Q}^{h_l}_l|} + \frac{|P_l|}{\lambda_0}\right] \vee \frac{1}{2|\mathcal{H}|}\|M(P_l,\widehat{Q}^h_l) - \overline{M}(P_l,\widehat{Q}^h_l)\|_F^2$. Hence applying (46) with $h = h_l$, we have that, for any $l \in \mathcal{L}_+$,

$$\begin{aligned}\|M(P'_l) - \overline{M}(P'_l)\|_F^2 &= \|M(P'_l,\widehat{Q}^{h_l}_l) - \overline{M}(P'_l,\widehat{Q}^{h_l}_l)\|_F^2 + \|M(P'_l,[d]\setminus\widehat{Q}^{h_l}_l) - \overline{M}(P'_l,[d]\setminus\widehat{Q}^{h_l}_l)\|_F^2 \\ &\leq \|M(P_l) - \overline{M}(P_l)\|_F^2 - \frac{1}{4\gamma_u^2}\|M(P_l,\widehat{Q}^{h_l}_l) - \overline{M}(P_l,\widehat{Q}^{h_l}_l)\|_F^2 \\ &\leq \left(1 - \frac{1}{\gamma_u^3}\right)\|M(P_l) - \overline{M}(P_l)\|_F^2 \ ,\end{aligned}$$

where the third inequality comes from the second term of (47) together with $P'_l \subset P_l$ and $\gamma_u \geq \phi_{L_1} \geq 8|\mathcal{H}|$, with ϕ_{L_1} defined in (9). Hence we obtain that

$$\mathrm{SN}(\mathcal{P}'_+) \leq \left(1 - \frac{1}{\gamma_u^3}\right)\mathrm{SN}(\mathcal{P}_+) \ . \tag{49}$$

Finally, we give an upper bound of $\mathrm{SN}(\mathcal{P}'_-)$. Let us write $\mathcal{D}_n = \{2^k \ : \ k \in \mathbb{Z}^+\} \cap [n]$ for the set of dyadic integer smaller than n. Given $p \in \mathcal{D}_n$, we write $\mathcal{L}_-(p) = \mathcal{L}(p) \cap \mathcal{L}_-$ for the set of indices in $\mathcal{L}_-$ such that $|P_l| \in [p,2p)$, and $\mathcal{P}'_-(p)$ for the corresponding sequence of subsets in $\mathcal{P}'_-(p)$. Let $\phi_0 = \kappa_0 \log(nd/\delta)$, where κ_0 is a numerical constant given by Lemma C.1. By definition of Q^{*h}_l, the square norm of a group P'_l restricted to questions that do not belong the set $\cup_{h\in\mathcal{H}} Q^{*h}_l$ is smaller than $\phi_0 nd \cdot \min(\mathcal{H}) \leq \phi_0$. Hence we have that

$$\mathrm{SN}(\mathcal{P}'_-) = \sum_{p\in\mathcal{D}_n}\mathrm{SN}(\mathcal{P}'_-(p)) \leq \phi_0 + \sum_{(p,h)\in\mathcal{D}_n\times\mathcal{H}}\ \sum_{l\in\mathcal{L}_-(p)}\|M(P'_l,Q^{*h}_l) - \overline{M}(P'_l,Q^{*h}_l)\|_F^2 \ . \tag{50}$$

If $\lambda_0|Q^{*h}_l| \leq 1$ then we use the trivial inequality $\|M(P'_l,Q^{*h}_l) - \overline{M}(P'_l,Q^{*h}_l)\|_F^2 \leq |P'_l\|Q^{*h}_l| \leq |P^h_l|/\lambda_0$,since the entries of M are bounded by one.

If $\lambda_0|Q_l^{*h}| \geq 1$ and $|\widetilde{P}_l^h|\lambda_0 \leq 1$, we have that $h|\widetilde{P}_l^h|\sqrt{\frac{|\overline{Q}_l^{*h}|}{\lambda_0}} \leq \sqrt{\frac{|\widetilde{P}_l^h||\overline{Q}_l^{*h}|}{\lambda_0^2}}$, using the fact that $h \leq 1$. Hence, since the experts in $P_l' \subset \widetilde{P}^h$ are indistinguishable in L_1 norm by Lemma C.2, (44) holds true and we have

$$\begin{aligned}\|M(P_l', Q_l^{*h}) - \overline{M}(P_l', Q_l^{*h})\|_F^2 &\leq 6\phi_0\gamma_u \left[h|\widetilde{P}_l^h|\sqrt{\frac{|\overline{Q}_l^{*h}|}{\lambda_0}} \right] \\ &\leq 6\phi_0\gamma_u \left[\sqrt{h^2|\widetilde{P}_l^h|^2\frac{|\overline{Q}_l^{*h}|}{\lambda_0}} \wedge \sqrt{\frac{|\widetilde{P}_l^h||\overline{Q}_l^{*h}|}{\lambda_0^2}} \right] \\ &\leq 12\phi_0\gamma_u \left[\sqrt{(h^2 p\lambda_0 \wedge 1)\frac{p|\overline{Q}_l^{*h}|}{\lambda_0^2}} + \frac{p}{\lambda_0} \right] .\end{aligned}$$

Finally, if $\lambda_0(|Q_l^{*h}| \vee |\widetilde{P}_l^h|) \geq 1$, we are in position to apply Proposition C.3. For all $l \in \mathcal{L}_-(p)$ and $h \in \mathcal{H}$ that $\|M(P_l', Q_l^{*h}) - \overline{M}(P_l', Q_l^{*h})\|_F^2$ is either smaller than $\frac{1}{2|\mathcal{H}|}\|M(P_l) - \overline{M}(P_l)\|_F^2$, or it is smaller than $\kappa_1\gamma_u^4\left[\frac{1}{\lambda_0}\sqrt{|P_l||\widehat{Q}_l^h|} + \frac{|P_l|}{\lambda_0}\right]$. From (44), it is also smaller than $6\phi_0\gamma_u h|\widetilde{P}_l^h|\sqrt{\frac{|\overline{Q}_l^{*h}|}{\lambda_0}}$. As a consequence, we obtain the following upper bound:

$$\begin{aligned}\|M(P_l', Q_l^{*h}) - \overline{M}(P_l', Q_l^{*h})\|_F^2 \leq &\kappa_2\gamma_u^4 \left[\sqrt{(h^2 p\lambda_0 \wedge 1)\frac{p|\overline{Q}_l^{*h}|}{\lambda_0^2}} + \frac{p}{\lambda_0} \right] \\ &\vee \frac{1}{2|\mathcal{H}|}\|M(P_l) - \overline{M}(P_l)\|_F^2 ,\end{aligned} \tag{51}$$

with $\kappa_2 = 12(\kappa_0 \vee \kappa_1)$, and using that $\phi_0 \leq \kappa_0\gamma_u$ and $|\widetilde{P}_l^h| \leq |P_l| \leq 2p$.

By the two previous cases on l, the inequality (51) is valid for any $l \in \mathcal{L}_-(p)$. Now, we decompose (50) into two terms, corresponding to the maximum in (51). First, since each P_l is in at most one $\mathcal{P}_-(p)$ for $p \in \mathcal{D}_n$, we have

$$\sum_{(p,h)\in\mathcal{D}_n\times\mathcal{H}} \sum_{l\in\mathcal{L}_-(p)} \frac{1}{2|\mathcal{H}|}\|M(P_l) - \overline{M}(P_l)\|_F^2 \leq \frac{1}{2}\,\mathrm{SN}(\mathcal{P}_-) . \tag{52}$$

Secondly, we have that

$$\begin{aligned}
\kappa_2\gamma_u^4 \sum_{(p,h)\in\mathcal{D}_n\times\mathcal{H}} \sum_{l\in\mathcal{L}_-(p)} \left[\sqrt{(h^2p\lambda_0\wedge 1)\frac{p|\overline{Q}_l^{*h}|}{\lambda_0^2}} + \frac{p}{\lambda_0}\right] \\
&\leq \kappa_2\gamma_u^6 \left[\max_{p,h} \sum_{l\in\mathcal{L}_-(p)} \sqrt{(h^2p\lambda_0\wedge 1)\frac{p|\overline{Q}_l^{h*}|}{\lambda_0^2}} + \frac{p}{\lambda_0}\right] \\
&\overset{(a)}{\leq} 2\kappa_2\gamma_u^6 \max_{p,h}\left[\frac{n}{\lambda_0} + \sqrt{(h^2p\lambda_0\wedge 1)\frac{p|\mathcal{L}(p)|\sum_{l\in\mathcal{L}(p)}|\overline{Q}_l^{h*}|}{\lambda_0^2}}\right] \\
&\overset{(b)}{\leq} 4\kappa_2^2\gamma_u^7 \max_{p,h}\left[\frac{n}{\lambda_0} + \sqrt{(h^2p\lambda_0\wedge 1)\left(\frac{n^2d}{\lambda_0^2p}\wedge\left(\frac{nd}{p\lambda_0^3h^3}\vee\frac{nd}{\lambda_0^2h}\right)\right)}\right] \\
&\leq 4\kappa_2^2\gamma_u^7 \max_{p,h}\left[\frac{n}{\lambda_0} + nh\sqrt{\frac{d}{\lambda_0}}\wedge\sqrt{\frac{n^2dh^2}{\lambda_0}}\wedge\frac{nd}{\lambda_0^2h}\right] \\
&\overset{(c)}{\leq} 4\kappa_2^2\gamma_u^7\left[\frac{n}{\lambda_0} + n\sqrt{\frac{d}{\lambda_0}}\wedge\frac{n^{2/3}\sqrt{d}}{\lambda_0^{5/6}}\right] .
\end{aligned}$$

where in (a) we used the Jensen inequality, in (b) we used Lemma C.4 with $\phi=\phi_0$ together with the trivial inequality $\sum_{l\in\mathcal{L}(p)}|\overline{Q}_l^{h*}|\leq nd/p$ and in (c) the fact that $x\wedge y\leq x^{2/3}y^{1/3}$ and $h\leq 1$.

Finally, combining this last inequality with (48), (49) and (52),
we obtain

$$\begin{aligned}
\mathrm{SN}(\mathcal{P}') &= \mathrm{SN}(\mathcal{P}'_+) + \mathrm{SN}(\mathcal{P}'_-) \\
&\leq \left(1-\frac{1}{\gamma_u^3}\right)\mathrm{SN}(\mathcal{P}_+) + 4\kappa_2^2\gamma_u^7\left[\frac{n}{\lambda_0} + n\sqrt{\frac{d}{\lambda_0}}\wedge\frac{n^{2/3}\sqrt{d}}{\lambda_0^{5/6}}\right]\vee\left[\frac{1}{2}\mathrm{SN}(\mathcal{P}_-)\right] \\
&\leq \left[C\bar{\gamma}^7\left(\frac{n}{\lambda_0} + n\sqrt{\frac{d}{\lambda_0}}\wedge\frac{n^{2/3}\sqrt{d}}{\lambda_0^{5/6}}\right)\right]\vee\left[\left(1-\frac{1}{\bar{\gamma}^3}\right)\mathrm{SN}(\mathcal{P})\right] ,
\end{aligned}$$

where we recall that $\bar{\gamma}$ is defined in (11) and satisfies $\bar{\gamma}\geq\gamma_u$. This concludes the proof of Proposition A.6.

Appendix D: Proof of the lemmas of Appendix C

Recall that we can write

$$E = (B-\mathbb{E}[B])\odot M + B\odot\widetilde{E} \ . \tag{53}$$

where we recall that $\widetilde{E}=Y-\mathbb{E}[Y|B]$ and that B is a matrix of Bernoulli random variables with parameter λ_1.

Proof of Lemma C.1. Assume first that $\lambda_0\leq 1$. Let us fix $l\in\{1,\dots,L\}$ and $h\in\mathcal{H}$. We omit the dependence in l in this proof to ease the notation and we write P for P_l. Let us define

$$E'_k(a) := \frac{1}{|\mathcal{N}_a|}\sum_{i\in\mathcal{N}_a}E_{ik} - \frac{1}{|\mathcal{N}_{-a}|}\sum_{i\in\mathcal{N}_{-a}}E_{ik} \quad \text{and} \quad \nu(a) := |\mathcal{N}_a|\wedge|\mathcal{N}_{-a}| \ . \tag{54}$$

Using Lemma B.1 with a column matrix W with coefficient in $\{0, \frac{1}{|\mathcal{N}_a|}, -\frac{1}{|\mathcal{N}_{-a}|}\}$ and a union bound over all $k \in [d]$ and $a \in [n]$, we have with probability at least $1 - \delta/L$ that:

$$\frac{1}{\lambda_0}\left|E'_k(a)\right| \le \kappa'_0 \log(nd/\delta)\left[\sqrt{\frac{1}{\lambda_0 \nu(a)}} + \frac{1}{\lambda_0 \nu(a)}\right] \ , \tag{55}$$

for some numerical constant κ'_0. In what follows, we work under that (55) holds true for all $a \in [n]$ and $k \in [d]$.

First inclusion. Let $k \in Q^*(\kappa_0 \log(nd/\delta)h)$ with numerical constant κ_0 to be fixed later. Let $a' \ge 1$ be any integer such that $\nu(a') \ge 1/(\lambda_0 h^2)$. We have

$$\frac{1}{\lambda_0}\left|E'_k(a')\right| \le 2\kappa'_0 \log(nd/\delta)h \ , \tag{56}$$

since we work under the event defined by (55) and since $h^2 \le h$. Then by consistency of the already constructed graph $\mathcal{G}_{t,u}$ at the beginning of step t, $\mathcal{N}_{a'}$ (resp. $\mathcal{N}_{-a'}$) contains by definition (20) only experts that are π^*-above (resp. below) all the experts of P. Since by assumption k is in Q^{*h}, it holds that $\mathbf{\Delta}^*_k(a') \ge \mathbf{\Delta}^*_k(0) \ge \kappa_0 \log(nd/\delta)h$ - see the definition (41) of Q^{*h}. Hence, recalling the signal-noise decomposition (53), we have that

$$\frac{1}{\lambda_0}\widehat{\mathbf{\Delta}}_k(a') = \frac{\lambda_1}{\lambda_0}\mathbf{\Delta}^*_k(a') + \frac{1}{\lambda_0}E'_k(a') \ge \log(nd/\delta)((1-1/e)\kappa_0 - 2\kappa'_0)h \ . \tag{57}$$

Choosing $\kappa_0 \ge 10\kappa'_0 + 1$, we obtain by definition (21) that $\nu(\hat{a}_k(h)) \le \frac{1}{\lambda_0 h^2}$ so that $k \in \widehat{Q}^h$.

Second inclusion. Let $k \in \widehat{Q}^h$, and $a' = a^*((\kappa_0 \log(nd/\delta))^{-1}h)$ be as defined in (40). By definition, it holds that $\nu(a') \ge \kappa_0 \log(nd/\delta)/(\lambda_0 h^2) \ge \frac{1}{\lambda_0 h^2}$. Hence, since $k \in \widehat{Q}^h$, we have by definition (22) that $\nu(\hat{a}_k(h)) \le \frac{1}{\lambda_0 h^2} \le \nu(a')$, which implies in particular that $\hat{a}_k(h) \le a'$. Then, by definition (21) of $\hat{a}_k(h)$ we have that $\frac{1}{\lambda_0}\widehat{\mathbf{\Delta}}_k(a') \ge h$. Using the concentration inequality (55) with $h' = (\kappa_0 \log(nd/\delta))^{-1}h$ and the fact that $\lambda_0 \ge \lambda_1$ we obtain

$$\mathbf{\Delta}^*_k(a') \ge h - \frac{2\kappa'_0}{\kappa_0}h \ , \tag{58}$$

and we get the second inclusion by also choosing $\kappa_0 \ge 4(\kappa'_0 + 1)$. □

Proof of Lemma C.4. For simplicity, we renumber $\mathcal{L}(p) = (1, 2, \dots, L' := |\mathcal{L}(p)|)$. Let us write $\nu(a, l) = |\mathcal{N}_a(l)| \wedge |\mathcal{N}_{-a}(l)|$ and $\Lambda = \left\lfloor \frac{\phi^2}{p\lambda_0 h^2} \right\rfloor + 1$. We let $a^* := a^*(\phi^{-1}h, l)$ be as defined in (40) so that for any l, $\nu(a^*, l) \ge \frac{\phi^2}{\lambda_0 h^2}$.

By assumption of Proposition A.6, it holds that $P_1 \overset{\mathcal{G}}{\prec} P_2 \overset{\mathcal{G}}{\prec} \dots \overset{\mathcal{G}}{\prec} P_{|\mathcal{L}(p)|}$ where we recall $\mathcal{G} = \mathcal{G}_{t,u}$ is the already constructed graph - see Appendix C.1. Hence it holds that $\mathbf{rk}_{\mathcal{G},i}(j) \ge \Lambda$ for any $i \in P_l$ and $j \in P_{l+\Lambda}$ - see (19) for the definition of $\mathbf{rk}$. Since there are at least $p\Lambda \ge \frac{\phi^2}{\lambda_0 h^2}$ experts in the union $P_{l+1} \cup \dots \cup P_{l+\Lambda}$, we conclude that $a^* \le \Lambda$, and that any expert in $\mathcal{N}_{a^*}$ (resp. $\mathcal{N}_{-a^*}$) is below the maximal expert of $P_{l+\Gamma}$ (resp. above) the minimal expert of $P_{l-\Lambda}$. This implies that, upon writing $\overline{\mathbf{\Delta}}^*_k(l)$ for the difference of these maximal and minimal experts, we have by definition (41) of $\overline{Q}^{*h}$ that $\overline{\mathbf{\Delta}}^*_k(l) > h/2$ for all k in $\overline{Q}^{*h}$. This implies in particular that

$$\sum_{l \in \mathcal{L}(p)} |\overline{Q}^{*h}_l(h, \phi)| \le \sum_{k=1}^{d} \sum_{l \in \mathcal{L}(p)} \mathbf{1}\{\overline{\mathbf{\Delta}}^*_k(l) \ge h/2\} \le \frac{2}{h}\sum_{k=1}^{d} \sum_{l \in \mathcal{L}(p)} \overline{\mathbf{\Delta}}^*_k(l) \le (2\Lambda + 1)\frac{2d}{h} \le 6\frac{\Lambda d}{h} \ , \tag{59}$$

where in the last inequality we used the fact that $M_{i,k} \in [0, 1]$ and that the sequence $P_{l-\Lambda}, \dots, P_{l+\Lambda}$ is of length $2\Lambda + 1$, for any $l \in \mathcal{L}(p)$. □

Appendix E: Proof of Proposition C.3

Let us fix any $l \in \{1, \dots, L\}$ and $h \in \mathcal{H}$. Since l, h and $\widehat{Q}_l^h$ are fixed in this proof, we simplify the notation and we write $(P', \widetilde{P}, Q) = (P'_l, \widetilde{P}_l^h, \widehat{Q}_l^h)$ and $M := M(\widetilde{P}, Q)$ and $M(P') := M(P', Q)$. We also fix $\delta' = \delta/(L|\mathcal{H}|)$, where we recall that $L \le n$ is the number of groups.

Let us assume that

$$\|M(P') - \overline{M}(P')\|_F^2 \ge \kappa_1 \gamma_u^4 \left[\frac{1}{\lambda_0} \sqrt{|\widetilde{P}\|Q|} + \frac{|\widetilde{P}|}{\lambda_0} \right] , \tag{60}$$

for some constant κ_1 to be fixed later. In what follows, we show that under assumption (60) for some large enough numerical constant κ_1, we necessarily have that the square norm of P' is a contraction of the square norm of P, that is

$$\|M(P') - \overline{M}(P')\|_F^2 \le \left(1 - \frac{1}{4\gamma_u^2}\right) \|M - \overline{M}\|_F^2 \ . \tag{61}$$

Step 1: control of the vector $\hat{v}$

First, the following lemma states that the first singular value of $(M - \overline{M})$ is, up to polylogarithmic terms, of the same order as its Frobenius norm. This is mainly due to the fact that the entries of M lie in $[0, 1]$ and that $M - \overline{M}$ is an isotonic matrix.

Lemma E.1 (Lemma E.4 in [14])**.** *Assume that $\|M - \overline{M}\|_F \ge 2$. For any sets $\widetilde{P}$ and Q, we have*

$$\|M - \overline{M}\|_{\mathrm{op}}^2 \ge \frac{4}{\gamma_u^2} \|M - \overline{M}\|_F^2 \ .$$

This lemma was already stated and proved as Lemma E.4 in [14], recalling that $\gamma_u > \phi_{\mathrm{L}_1} \ge 8 \log(nd)$ – see (9) and (10).

Now, write $\hat{v} = \arg\max_{\|v\|_2 \le 1} \left[\|v^T (Y^{(2)} - \overline{Y}^{(2)})\|_2^2 - \frac{1}{2} \|v^T (Y^{(2)} - \overline{Y}^{(2)} - Y^{(3)} + \overline{Y}^{(3)})\|_2^2 \right]$, where the argmax is taken over all v in $\widetilde{P}$.

Lemma E.2. *Assume that $\lambda_0 |\widetilde{P}| \ge 1$. There exists a numerical constant κ_0' such that if*

$$\|M - \overline{M}\|_{\mathrm{op}}^2 \ge \kappa_0' \log^2(nd/\delta') \left(\frac{1}{\lambda_0} \sqrt{|Q\|\widetilde{P}|} + \frac{|\widetilde{P}|}{\lambda_0} \right) \ , \tag{62}$$

then, with probability higher than $1 - \delta'$, we have

$$\|\hat{v}^T \left(M - \overline{M}\right)\|_2^2 \ge \frac{1}{2} \|M - \overline{M}\|_{\mathrm{op}}^2 \ .$$

In light of Lemma E.1 and Condition (60), the Condition (62) in Lemma E.2 is valid if we choose κ_1 in Proposition C.3 such that $\kappa_1 \ge 16\kappa_0'$. Consequently, there exists an event of probability higher than $1 - \delta'$ such that

$$\|\hat{v}^T \left(M - \overline{M}\right)\|_2^2 \ge \frac{2}{\gamma_u^2} \|M - \overline{M}\|_F^2 \ . \tag{63}$$

Step 2: control of the vector $\hat{v}_-$

Now remark that since $\|\hat{v}_i\|_2 = 1$, there are at most $\frac{1}{\lambda_0}$ of experts i such that $\hat{v}_i > \sqrt{\lambda_0}$. Hence we have that

$$\begin{aligned}\|\hat{v}_-^T\left(M-\overline{M}\right)\|_2^2 &\geq \frac{2}{\gamma_u^2}\|M-\overline{M}\|_F^2 - \sum_{i\in\widetilde{P}}\mathbf{1}_{\hat{v}_i>\sqrt{\lambda_0}}\|M_{i\cdot}-\overline{m}\|_2^2 \\ &\overset{(a)}{\geq} \frac{2}{\gamma_u^2}\|M-\overline{M}\|_F^2 - \frac{3\gamma_u}{\lambda_0}\sqrt{\frac{|\widehat{Q}|}{\lambda_0}} \\ &\overset{(b)}{\geq} \frac{1}{\gamma_u^2}\|M-\overline{M}\|_F^2 \ .\end{aligned}$$

(a) comes from the fact that any expert in $\widetilde{P}$ satisfies (43) under the event of Lemma C.2. (b) comes from Condition (60) and the assumption that $\lambda_0|\widetilde{P}| \geq 1$.

Step 3: control of the vector $\hat{w}$

Next, we show that a thresholded version of $\hat{z} = (Y^{(4)} - \overline{Y}^{(4)})^T\hat{v}_-$ is almost aligned with $z^* = \lambda_1(M-\overline{M})^T\hat{v}_-$. We define the sets $S^* \subset Q$ and $\hat{S} \subset Q$ of questions by

$$S^* = \left\{k \in Q \ : \ |z_k^*| \geq 2\gamma_u\sqrt{\lambda_0}\right\} \ ; \quad \hat{S} = \left\{k \in Q \ : \ |\hat{z}_k| \geq \gamma_u\sqrt{\lambda_0}\right\} \ . \tag{64}$$

S^* stands for the collection of questions k such that z_k^* is large whereas $\hat{S}$ is the collection questions k with large $\hat{z}_k$. Finally, we consider the vectors w^* and $\hat{w}$ defined as theresholded versions of z^* and $\hat{z}$ respectively, that is $w_k^* = z_k^*\mathbf{1}_{k\in S^*}$ and $\hat{w}_k = \hat{z}_k\mathbf{1}_{k\in\hat{S}}$. Note that, up to the sign, $\hat{w}$ stands for the active coordinates computed in **SLR**, Line 7 of Algorithm 2.

Recall that we assume that $\lambda_0 \leq 1$. We write v for any unit vector in $\mathbb{R}^{|\widetilde{P}|}$. Let us apply Lemma B.1 for each column $k \in Q$ of the noise matrix E with the matrix W equal to $v - (\frac{1}{|\widetilde{P}|}\sum_{i\in\widetilde{P}}v_i)\mathbf{1}_{\widetilde{P}}$ at column k and 0 elsewhere. We deduce that, for any fixed matrix M, any subsets $\widetilde{P}$ and Q, and any unit vector $v \in \mathbb{R}^{\widetilde{P}}$ such that $\|v\|_\infty \leq 2\sqrt{\lambda_0}$, we have

$$\mathbb{P}\left[\max_{k\in Q}\left|(v^T(E^{(3)}-\overline{E}^{(3)}))_k\right| \leq 100\log(2|Q|/\delta')\sqrt{\lambda_0}\right] \geq 1-\delta' \ . \tag{65}$$

Observe that $\hat{z} = z^* + (E^{(3)}-\overline{E}^{(3)})^T\hat{v}_-$. Conditioning on $\hat{v}_-$, we deduce that, on an event of probability higher than $1-\delta'$, we have

$$\|\hat{z}-z^*\|_\infty \leq 100\log(2|Q|/\delta')\sqrt{\lambda_0} \ \leq \frac{\gamma_u}{2}\sqrt{\lambda_0} \ , \tag{66}$$

where the last inequality comes from $\gamma_u > \phi_{\mathrm{L}_1}$. Hence it holds that $S^* \subset \hat{S}$ and for $k \in \hat{S}$, we have $z_k^*/\hat{z}_k \in [1/2, 2]$. Next, we shall prove that, under this event, $\lambda_1\hat{v}_-^T(M-\overline{M})\hat{w}/\|\hat{w}\|_2$ is large (in absolute value):

$$\lambda_1\left|\hat{v}_-^T(M-\overline{M})\hat{w}\right| = \left|(z^*)^T\hat{w}\right| = \sum_{k\in\hat{S}}z_k^*\hat{z}_l \geq \frac{2}{5}\sum_{k\in\hat{S}}(z_k^*)^2 + (\hat{z}_l)^2 \geq \frac{2}{5}[\|w^*\|_2^2 + \|\hat{w}\|_2^2] \geq \frac{4}{5}\|\hat{w}\|_2\|w^*\|_2 \ ,$$

where we used in the first inequality that $z_k^*/\hat{z}_k \in [1/2, 2]$ and in the second inequality that $S^* \subset \hat{S}$. Thus, it holds that

$$\lambda_1^2\left|\hat{v}_-^T(M-\overline{M})\frac{\hat{w}}{\|\hat{w}\|_2}\right|^2 \geq \frac{16}{25}\|w^*\|_2^2 \ . \tag{67}$$

It remains to prove that $\|w^*\|_2$ is large enough. Writing S^{*c} for the complementary of S^* in Q, it holds that

$$\|w^*\|_2^2 = \|z^*\|_2^2 - \sum_{k\in S^{*c}} (z_k^*)^2 \ , \tag{68}$$

so that we need to upper bound the latter quantity. Write $z^*_{S^{*c}} = z^* - w^*$. Coming back to the definition of z^*,

$$\begin{aligned}
\left[\sum_{k\in S^{*c}} (z_k^*)^2\right]^2 &= \left[\sum_{k\in S^{*c}} \lambda_1 [\hat{v}_-^T (M-\overline{M})]_k z_k^*\right]^2 \\
&\le \|\lambda_1 \left(M-\overline{M}\right) z^*_{S^{*c}}\|_2^2 = \sum_{i\in\widetilde{P}} \left(\sum_{k\in S^{*c}} \lambda_1 (M_{ik} - \overline{m}_k) z_k^*\right)^2 \\
&\overset{(a)}{\le} \frac{4\gamma_u^2}{|\widetilde{P}|^2}\lambda_0 \sum_{i\in\widetilde{P}} \left(\sum_{k\in S^{*c}}\sum_{j\in\widetilde{P}} \lambda_1 |M_{ik} - M_{jk}|\right)^2 \\
&\le \frac{4\gamma_u^2}{|\widetilde{P}|^2}\lambda_0 \sum_{i\in\widetilde{P}} \left(\sum_{j\in\widetilde{P}} \lambda_1 \|M_{i\cdot} - M_{j\cdot}\|_1\right)^2 \\
&\overset{(b)}{\le} 40\gamma_u^4\lambda_0^2 |\widetilde{P}||Q| \\
&\le \left[7\gamma_u^2\lambda_0\sqrt{|\widetilde{P}||Q|}\right]^2 \le \left[\frac{1}{2\gamma_u^2}\lambda_0^2\|M-\overline{M}\|_F^2\right]^2 \ .
\end{aligned}$$

In (a), we used the definition of S^*. In (b), we used (43) that holds true since we are under the event Lemma C.1 and $\lambda_0|Q| \ge 1$. The last inequality comes from Condition (60), choosing $\kappa_1 \ge 14$.

Recall that $z^* = \hat{v}_-^T(M-\overline{M})$. Combining (63) and (68), we deduce that

$$\|w^*\|_2^2 \ge \frac{1}{2\gamma_u^2}\lambda_0^2\|M-\overline{M}\|_F^2 \ , \tag{69}$$

which, together with (67) and $\lambda_0 \ge \lambda_1$, yields

$$\left\|(M-\overline{M})\frac{\hat{w}}{\|\hat{w}\|_2}\right\|_2^2 \ge \left|\hat{v}_-^T(M-\overline{M})\frac{\hat{w}}{\|\hat{w}\|_2}\right|^2 \ge \frac{1}{2\gamma_u^2}\|M-\overline{M}\|_F^2 \ . \tag{70}$$

Write $\hat{w}^{(1)}$ and $\hat{w}^{(2)}$ the positive and negative parts of $\hat{w}$ respectively so that $\hat{w} = \hat{w}^{(1)} - \hat{w}^{(2)}$ and $\hat{w}^+ = \hat{w}^{(1)} + \hat{w}^{(2)}$. We obviously have $\|\hat{w}\|_2 = \|\hat{w}^+\|_2$. Besides, if the rows of M are ordered according to the oracle permutation, then $(M-\overline{M})\hat{w}^{(1)}$ and $(M-\overline{M})\hat{w}^{(2)}$ are nondecreasing vectors with mean zero. It then follows from Harris' inequality that these two vectors have a nonegative inner product. We have proved that

$$\left\|(M-\overline{M})\frac{\hat{w}^+}{\|\hat{w}^+\|_2}\right\|_2^2 \ge \left\|(M-\overline{M})\frac{\hat{w}}{\|\hat{w}\|_2}\right\|_2^2 \ge \frac{1}{2\gamma_u^2}\|M-\overline{M}\|_F^2 \ . \tag{71}$$

Step 4: Showing that $\hat{w}$ satisfies Condition (16)

Recall that we assume for simplicity that $\lambda_0 \le 1$. First we upper bound $\|w\|_\infty^2$ by using (a) that $\hat{z}$ is close to z^* with (66), (b) that for any $k \in Q$, $v^T M_{\cdot k} \le \|v\|_1$ and (c) that $\lambda_0|\widetilde{P}| \ge 1$:

$$\|\hat{w}\|_\infty^2 \overset{(a)}{\le} 2\|z^*\|_\infty^2 + \gamma_u^2\lambda_0 \overset{(b)}{\le} 2\lambda_0^2\|\hat{v}\|_1^2 + \gamma_u^2\lambda_0 \overset{(c)}{\le} 3\gamma_u^2\lambda_0^2|\widetilde{P}| \ . \tag{72}$$

Secondly, we lower bound $\|w\|_2^2$ by using (a) that $S^* \subset \hat{S}$ and that $z_k^*/\hat{z}_k \in [1/2, 2]$, (b) that $\|w^*\|_2^2$ captures a significant part of the L_2 norm -see (69), and (c) the Condition (60) with $\kappa_1 \geq 24$:

$$\|\hat{w}\|_2^2 \overset{(a)}{\geq} \frac{1}{4}\|w^*\|_2^2 \overset{(b)}{\geq} \frac{1}{8\gamma_u^2}\lambda_0^2\|M - \overline{M}\|_F^2 \overset{(c)}{\geq} 3\gamma_u^2\lambda_0|\widetilde{P}| \ . \tag{73}$$

We deduce that $\|\hat{w}\|_\infty^2 \leq \lambda_0\|\hat{w}\|_2^2$, which is exactly Condition (16). This shows that $\hat{w}^+$ is considered for the update (18) in the final step of the procedure Line 9 of Algorithm 2.

Step 5: upper bound of the Frobenius-norm restricted to P'

Equipped with this bound, we are now in position to show that the set P' of experts obtained from $\widetilde{P}$ when applying the pivoting algorithm with $\hat{w}^+/\|\hat{w}^+\|_2$ has a much smaller square norm. By Lemma B.1 used with the matrix W equal to 0 except at line i where it is equal to the vector $\hat{w}^+/\|\hat{w}^+\|_2$, there exists an event of probability higher than $1-\delta'$ such that

$$\max_{i,j\in P'}\left|\langle E_{i\cdot} - E_{j\cdot}, \frac{\hat{w}^+}{\|\hat{w}^+\|_2}\rangle\right| \leq \phi_{\mathrm{L}_1}\sqrt{\lambda_0} \leq \gamma_u\sqrt{\lambda_0} \ ,$$

where we recall that ϕ_{l_1} is defined in (9). Hence, since the vector $\hat{w}$ is considered in the update (18), we have $\max_{i,j\in P'}\left|\langle Y_{i\cdot} - Y_{j\cdot}, \frac{\hat{w}^+}{\|\hat{w}^+\|_2}\rangle\right| \leq \gamma_u\sqrt{\lambda_0}$ and

$$\max_{i,j\in P'}\left|\langle M_{i\cdot} - M_{j\cdot}, \frac{\hat{w}^+}{\|\hat{w}^+\|_2}\rangle\right| \leq 2\gamma_u\sqrt{\frac{1}{\lambda_0}} \ . \tag{74}$$

By convexity, it follows that

$$\left\|(M(P') - \overline{M}(P'))\tfrac{\hat{w}^+}{\|\hat{w}^+\|_2}\right\|_2^2 \leq 4\gamma_u^2\frac{1}{\lambda_0}|P'| \leq 4\gamma_u^2\frac{1}{\lambda_0}|\widetilde{P}| \ .$$

In light of Condition (60), this quantity is small compared to $\|M - \overline{M}\|_F^2$:

$$\|(M(P') - \overline{M}(P'))\tfrac{\hat{w}^+}{\|\hat{w}^+\|_2}\|_2^2 \leq \frac{1}{4\gamma_u^2}\|M - \overline{M}\|_F^2 \ , \tag{75}$$

which together with (71) leads to

$$\|(M - \overline{M})\tfrac{\hat{w}^+}{\|\hat{w}^+\|_2}\|_2^2 - \|(M(P') - \overline{M}(P'))\tfrac{\hat{w}^+}{\|\hat{w}^+\|_2}\|_2^2 \geq \frac{1}{4\gamma_u^2}\|M - \overline{M}\|_F^2 \ . \tag{76}$$

Since $P' \subset \widetilde{P}$, we deduce that, for any vector $w' \in \mathbb{R}^q$, we have $\|(M - \overline{M})w'\|_2^2 \geq \|(M(P') - \overline{M}(P'))w'\|^2$. It then follows from the Pythagorean theorem that

$$\|M - \overline{M}\|_F^2 - \|M(P') - \overline{M}(P')\|_F^2 \geq \|(M - \overline{M})\tfrac{\hat{w}^+}{\|\hat{w}^+\|_2}\|_2^2 - \|(M(P') - \overline{M}(P'))\tfrac{\hat{w}^+}{\|\hat{w}^+\|_2}\|_2^2 \ .$$

Then, together with (76), we arrive at

$$\|M(P') - \overline{M}(P')\|_F^2 \leq \left(1 - \frac{1}{4\gamma_u^2}\right)\|M - \overline{M}\|_F^2 \ .$$

We have shown that if (60) is satisfied, then there is a contraction in the sense of (61). This in turn gives the upper bound (46) and it concludes the proof of Proposition C.3.

Proof of Lemma E.2. Recall that we consider the case $\lambda_0 \le 1$ and that the case $\lambda_0 \ge 1$ is discussed in Appendix F. We start with the two following lemmas. To ease the notation, we assume in this proof that $\widetilde{P} = \{1, \dots, p\}$, that $Q = \{1, \dots q\}$. We only consider the matrices restricted to the sets $\widetilde{P}, Q$ and we write $E := E(\widetilde{P}, Q)$. Let us define $J = \mathbf{1}\mathbf{1}^T \in \mathbb{R}^{p\times p}$ the matrix with constant coefficients equals to 1 and $A = (\mathbf{I}_p - \frac{1}{p}J)$ be the projector on the orthogonal of $\mathbf{1}$, so that $E - \overline{E} = AE \in \mathbb{R}^{p\times q}$. The two following lemmas are direct consequences of Proposition 4.1, and a discussion of the corresponding concenration inequality on random rectangular matrices can be found in Section 4. We state weaker concentration inequalities than what is proven in Proposition 4.1 in order to factorize the polylogarithmic factors and to ease the reading of the proof.

Lemma E.3. *Assume that $\lambda_0 \le 1$ and that $\lambda_0(p \vee q) \ge 1$. It holds with probability larger than $1 - \delta'/4$ that*

$$\|EE^T - \mathbb{E}[EE^T]\|_{\mathrm{op}} \le \kappa_0'' \log^2(pq/\delta')\,[\lambda_0\sqrt{pq} + \lambda_0 p] \quad .$$

Lemma E.4. *Assume that $\lambda_0 \le 1$ and that $\lambda_0(p \vee q) \ge 1$. With probability larger than $1 - \delta'/4$, one has for any orthogonal projection $\Lambda \in \mathbb{R}^{q\times q}$ satisfying $\mathrm{rank}(\Lambda) \le p$ that*

$$\|\Lambda E^T E \Lambda\|_{\mathrm{op}} \le \kappa_1'' \log^2(pq/\delta')\,[\lambda_0\sqrt{pq} + \lambda_0 p] \quad ,$$

Proofs of Lemma E.3 and Lemma E.4. First, we recall that for any i, k, we have that $E_{ik} = (B_{ik} - \lambda_1)M_{ik} + \widetilde{E}_{ik}$, and that $\widetilde{E}$ is an average of 1-subGaussian random variables, as described in (30) For any $u \ge 0$ we have

$$\mathbb{E}[E_{ik}^{2u}] \le 3^u \mathbb{E}\left[B_{ik} + \lambda_0^{2u} + \widetilde{E}_{ik}^{2u}\right] \le 3^u \left(2\lambda_0 + u!\mathbb{E}[e^{\widetilde{E}_{ik}^2}]\right) \le \frac{1}{2}u!\lambda_0 1000^u \quad , \tag{77}$$

where for the last inequality we used the following inequalities:

$$\mathbb{E}[e^{\widetilde{E}_{ik}^2}] \le \sum_{u\ge 1} e^{-\lambda_0}\frac{\lambda_0^u}{u!}e^{1/u} \le \lambda_0 e \quad .$$

Hence condition (24) is satisfied with $K = 1000$ and $\sigma^2 = \lambda_0$ for the coefficients of E. We just apply Proposition 4.1 with $X = E$ for Lemma E.3. For Lemma E.4, we apply Proposition 4.1 with $X = E^T$ and we remark that $\|\Lambda E^T E \Lambda\|_{\mathrm{op}}^2 \le 2\|\Lambda E^T E - \mathbb{E}[E^T E]\Lambda\|_{\mathrm{op}}^2 + 2\|\mathbb{E}[E^T E]\|_{\mathrm{op}}^2$ together with the fact that $\|\mathbb{E}[E^T E]\|_{\mathrm{op}}^2 \le c'\lambda_0 p$ for some numerical constant c'. □

Remark that since we assume in Lemma E.2 that $\lambda_0 p \ge 1$, it holds that $\sqrt{\lambda_0 p} \le \lambda_0 p$ and $\sqrt{\lambda_0 q} \le \lambda_0^2\sqrt{pq}$, so that both upper bounds of Lemma E.3 and Lemma E.4 reduce - up to logarithmic factors - to $\lambda_0\sqrt{pq} + \lambda_0 p$. We write for short in the following

$$F := F(p, q, \lambda_0, \delta') = \log^2(pq/\delta')[\lambda_0\sqrt{pq} + \lambda_0 p] \quad , \tag{78}$$

and $\kappa_2'' = 8(\kappa_0'' \vee \kappa_1'')$.

Now let us write

$$AY = \lambda_1 AM + AE \quad ,$$

so that, for any $v \in \mathbb{R}^p$, recalling that $AY = Y - \overline{Y}$,

$$\|v^T AY\|_2^2 = \lambda_1^2\|v^T AM\|_2^2 + \|v^T AE\|_2^2 + 2\lambda_1\langle v^T AE, v^T AM\rangle \quad ,$$

which, in turn, implies that

$$\begin{aligned}\left|\|v^T AY\|_2^2 - \lambda_1^2\|v^T AM\|_2^2 - \mathbb{E}\left[\|v^T AE\|_2^2\right]\right| &\le \left|\|v^T AE\|_2^2 - \mathbb{E}\left[\|v^T AE\|_2^2\right]\right| + 2\lambda_1|v^T AME^T(Av)| \\ &\overset{(a)}{\le} \|A(EE^T - \mathbb{E}[EE^T])A\|_{\mathrm{op}} + 2\lambda_1\|AME^T E(AM)^T\|_{\mathrm{op}}^{1/2} \\ &\le \|EE^T - \mathbb{E}[EE^T]\|_{\mathrm{op}} + 2\lambda_1\|AM\|_{\mathrm{op}}\|\Lambda E^T E\Lambda\|_{\mathrm{op}}^{1/2} \quad ,\end{aligned}$$

Where we define $\Lambda \in \mathbb{R}^{d\times d}$ as the orthogonal projector on the image of $\ker(AM)^\perp$ which is of rank less than p. For (a), we used the fact that A is contracting the operator norm as an orthogonal projector so that $\|Av\|_2 \le 1$. We now apply Lemma E.3 and Lemma E.4 together with the fact that $\lambda_1 \le \lambda_0$, and we obtain with probability at least $1-\delta'/2$ that

$$\sup_{v\in\mathbb{R}^p,\|v\|=1} \left| \|v^T AY\|_2^2 - \lambda_1^2 \|v^T AM\|_2^2 - \mathbb{E}\left[\|v^T AE\|_2^2\right] \right| \le \kappa_2'' F + \lambda_1 \|AM\|_{\mathrm{op}} \sqrt{\kappa_2'' F} \ . \tag{79}$$

where F is defined in (78). In the same way, we have that, with probability larger than $1-\delta'/2$,

$$\sup_{v\in\mathbb{R}^p:\ \|v\|_2\le 1} \left| \frac{1}{2}\|v^T A(Y-Y')\|_2^2 - \mathbb{E}\left[\|v^T AE\|_2^2\right] \right| = \frac{1}{2} \sup_{v\in\mathbb{R}^p:\ \|v\|_2\le 1} \left| \|v^T A(Y-Y')\|_2^2 - \mathbb{E}\|v^T A(Y-Y')\|_2^2 \right| \le \kappa_3'' F \ ,$$

for some numerical constant κ_3''. Putting everything together we conclude that, on an event of probability higher than $1-\delta'$, we have simultaneously for all $v \in \mathbb{R}^p$ with $\|v\|_2 \le 1$ that

$$\left| \|v^T AY\|_2^2 - \|v^T AM\|_2^2 - \frac{1}{2}\|v^T A(Y-Y')\|_2^2 \right| \le \kappa_4'' F + \lambda_1 \|AM\|_{\mathrm{op}} \sqrt{\kappa_4'' F} \ ,$$

with $\kappa_4'' = \kappa_2'' \vee \kappa_3''$. Choosing the numerical constant κ_0' of Lemma E.2 such that $\kappa_0' \ge 4\cdot 16(1-1/e)^{-1}\kappa_4''$ we have

$$\lambda_1^2 \|AM\|_{\mathrm{op}}^2 \ge 4\cdot 16 \kappa_4'' F \ ,$$

since it holds that $\lambda_1 \ge (1-1/e)\lambda_0$. We deduce that on the same event:

$$\sup_{v\in\mathbb{R}^p:\ \|v\|_2\le 1} \left| \|v^T AY\|_2^2 - \|v^T AM\|_2^2 - \frac{1}{2}\|v^T A(Y-Y')\|_2^2 \right| \le \frac{1}{4}\|AM\|_{\mathrm{op}}^2 \ .$$

Writing $\psi(v) = \left| \|v^T(Y-\overline{Y})\|_2^2 - \frac{1}{2}\|v^T A(Y-Y')\|_2^2 \right|$, we deduce that, for v such that $\|v^T AM\|_2^2 = \|AM\|_{\mathrm{op}}^2$, we have $\Psi(v) \ge \frac{3}{4}\|AM\|_{\mathrm{op}}^2$, whereas, for v such that $\|v^T AM\|_2^2 < \frac{1}{2}\|AM\|_{\mathrm{op}}^2$, we have $\Psi(v) < \frac{3}{4}\|AM\|_{\mathrm{op}}^2$. We conclude that $\hat{v}$ satisfies $\|\hat{v}^T AM\|_2^2 > \frac{1}{2}\|AM\|_{\mathrm{op}}^2$ with probability at least $1-\delta'$.

□

Appendix F: Proof of Theorem 2.2 when $\lambda_0 \ge 1$

The aim of this section is to provide an extension of the proof of Theorem 2.2 to the case $\lambda_0 \ge 1$. Recall that we fix δ to be a small probability the proof of Theorem 2.2, and that E and $\widetilde{E}$ are the matrices defined in (29) and (30) by

$$\widetilde{E}_{ik}^{(s)} = \sum_{t\in N^{(s)}} \frac{\varepsilon_t}{\mathbf{r}_{ik}^{(s)} \vee 1} \mathbf{1}\{x_t = (i,k)\} \quad \text{and} \quad E_{ik}^{(s)} = (B_{ik}^{(s)} - \lambda_1)M + B_{ik}^{(s)} \widetilde{E}_{ik}^{(s)} \ .$$

In what follows, we consider the two subcases where $\lambda_0 > 16\log(5nd/\delta)$ or $\lambda_0 \le 16\log(5nd/\delta)$, which essentialy rely on the two following ideas:

- If $\lambda_0 \le 16\log(5nd/\delta)$, we use the fact that the coefficients of E defined in (29) are 5-subGaussian together with the same signal-noise decomposition $Y = \lambda_1 M + E$ as in the proofs when $\lambda_0 \le 1$. The difference from the case $\lambda_0 \le 1$ lies in the application of subGaussian inequalities of E_{ik} instead of Bernstein inequalities as in (37).

- If $\lambda_0 > 16\log(5nd/\delta)$, we show that the event $\{\mathbf{r}_{ik}^{(s)} \geq \lambda_0/2\}$ holds true for all i,k,s with high probability. Working conditionally to this event, we use the decomposition $Y = M + \widetilde{E}$ and we show that the noise $\widetilde{E}$ has $\frac{2}{\lambda_0}$-subGaussian independent coefficients. The rationale behind using $\widetilde{E}$ when λ_0 is large is that $\widetilde{E}_{ik}$ takes advantage of the mean of $2/\lambda_0$ subGaussian variables with high probability.

Let $\mathbf{r}_{\min}^{(s)} = \min_{i,k} \mathbf{r}_{ik}^{(s)}$ be the minimum number of observation at positions (i,k) in N_s - see (14). In the case $\lambda_0 > 16\log(5nd/\delta)$, the following lemma states that with high probability, we observe all the coefficients for all sample s in the full observation regime.

Lemma F.1. *Assume that $\lambda_0 \geq 16\log(5nd/\delta)$. The event $\{\mathbf{r}_{\min}^{(s)} \geq \lambda_0/2\}$ holds simultaneously for all sample s with probability at least $1 - 5T\delta$.*

Proof of Lemma F.1. We apply Chernoff's inequality - see e.g. section 2.2 of [11] - to derive that for any i,k

$$\mathbb{P}(\mathbf{r}_{ik}^{(s)} \leq \lambda_0/2) \leq \exp(-\frac{1}{8}\lambda_0) \leq \delta/(nd) \ , \tag{80}$$

where we use the inequality $(1 - \log(2))/2 \geq 1/8$. We conclude with a union bound over all coefficients in $[n] \times [d]$ and all $5T$ samples. □

Let us now omit the dependence of E and $\widetilde{E}$ in the sample s. In what follows, use that the coefficients of E are 5-subGaussian, which is a consequence of the fact that E_{ik} is the sum of a centered variable bounded by 1 and a 1-subgaussian random variable $\widetilde{E}_{ik}$, so that by Cauchy-Schwarz and the Hoeffding inequality we have

$$\mathbb{E}[\exp(xE_{ik})] \leq \sqrt{\exp(4x^2/8)}\sqrt{\exp(4x^2/2)} = \exp(5/4x^2) \ . \tag{81}$$

Under the event of Lemma F.1, we use that $\widetilde{E}_{ik}$ is $\lambda_0/2$-subGaussian, as an average of at least $2/\lambda_0$ random variables that are 1-subGaussians:

$$\mathbb{E}[\exp(x\widetilde{E}_{ik})] \leq \exp(\tfrac{1}{\lambda_0}x^2) \ , \tag{82}$$

F.1. Adjustements for the general analysis

We first make the changes that should be done in Appendix A to have a proper proof in the case $\lambda_0 \geq 1$.

If $\lambda_0 \in [1, 16\log(5nd/\delta)]$, we simply replace λ_0 by $1/\lambda_0$ in the upper bound of (31) for the event ξ in Lemma A.1. In the proof of the restated Lemma A.1, we can replace the inequality (37) by

$$|\langle E, W\rangle| \leq \sqrt{10\|W\|_F^2 \log\left(\frac{2}{\delta'}\right)} \ , \tag{83}$$

for any matrix $W \in \mathbb{R}^{n\times d}$, with probability at least $1 - \delta'$. We can then obtain $1/\lambda_0$ instead of λ_0 simply by using that $\phi_{\mathrm{L}_1}/\sqrt{\lambda_0} \geq \sqrt{\phi_{\mathrm{L}_1}}$, recalling that ϕ_{L_1} is defined in (9).

If $\lambda_0 > 16\log(5nd/\delta)$, we say that we are under event ξ if the event of Lemma F.1 holds and (31) holds for all pairs (Q,w), replacing E by $\widetilde{E}$, and λ_0 by $1/\lambda_0$. The proof of the new version of Lemma A.1 lies in the Hoeffding inequality applied to $\widetilde{E}$ under the event of Lemma F.1, leading to the subsequent equation:

$$|\langle \widetilde{E}, W\rangle| \leq \sqrt{\frac{4\|W\|_F^2}{\lambda_0} \log\left(\frac{2}{\delta'}\right)} \ , \tag{84}$$

for any matrix $W \in \mathbb{R}^{n\times d}$, with probability at least $1 - \delta'$. This equation then replaces (37).

F.2. Adjustments to the proofs of Proposition A.6

We now adapt the proofs in Appendix C of Proposition A.6 to the case $\lambda_0 \geq 1$.

All the lemmas of Appendix C can be stated as is for any $\lambda_0 \geq 1$, and the only adjustments concern the proofs of Lemma C.1, Lemma C.2 and Proposition C.3.

F.2.1. Adjustments in the proofs of Lemma C.1 and Lemma C.2

Consider the proof of Lemma C.1. First, if $\lambda_0 \geq 16\log(5nd/\delta)$, we place ourselves under the event Lemma F.1 and replace λ_1 by 1 and all the E by $\widetilde{E}$. Instead of inequality (55), we use the fact that the coefficients of $\widetilde{E}$ are $2/\lambda_0$-subGaussian - see (82) - leading to the following inequality with probability at least $1-\delta$:

$$\left|\widetilde{E}_k(a)\right| := \left|\frac{1}{|\mathcal{N}_a|}\sum_{i\in\mathcal{N}_a}\widetilde{E}_{ik} - \frac{1}{|\mathcal{N}_{-a}|}\sum_{i\in\mathcal{N}_{-a}}\widetilde{E}_{ik}\right| \leq \kappa_0' \log(nd/\delta)\sqrt{\frac{1}{\lambda_0\nu(a)}}\ , \tag{85}$$

for some numerical constant κ_0'. The rest of the proof remains unchanged.

If $\lambda_0 \in [1, 16\log(5nd/\delta)]$, we use the fact that E has 5-subGaussians coefficients - see (81) and we do not divide by λ_0 in (57) - see the definition of $\widehat{\mathbf{\Delta}}$ (21).

Concerning Lemma C.2, the adjustments are the same as for Lemma A.1, namely working under the event of Lemma F.1 and we replacing E by $\widetilde{E}$, λ_0 by $1/\lambda_0$ and λ_1 by 1 if $\lambda_0 \geq 16\log(5nd/\delta)$, and using the fact that the coefficient of E are 5-subGaussians - see (81) if $\lambda_0 \in [1, 16\log(5nd/\delta)]$.

F.2.2. Adjustments in the proof of Proposition C.3

We now adapt the proofs in Appendix E of Proposition C.3 to the case $\lambda_0 \geq 1$. First, Lemma E.2 can be stated as is, and its proof when $\lambda_0 \geq 1$ is directly implied by Lemma E.5 in [14] with $\Theta := M$ either conditionally on Lemma F.1 with noise $N := \widetilde{E}$ and $\zeta^2 := 2/\lambda_0$ when $\lambda_0 \geq 16\log(5nd/\delta)$ or with noise $N := E$ and $\zeta^2 := 5$ when $\lambda_0 \leq 16\log(5nd/\delta)$.

Secondly, remark that if $\lambda_0 \geq 1$, it holds that $\hat{v}_- = \hat{v}$ and that Condition (16) on $\hat{w}$ is automatically satisfied, so that step 2 and step 4 can be removed from the proof in that case. For Step 3 and 5, we do the following adjustments:

If $\lambda_0 \in [1, 16\log(5nd/\delta)]$, the proof remains unchanged except that we use that the coefficients of E are 5-subGaussian -see (81).

If $\lambda_0 \geq 16\log(5nd/\delta)$, we work conditionnally on the event of Lemma F.1 and we replace λ_1 by 1 and E by $\widetilde{E}$. The subgaussian concentration bound on $\widetilde{E}$ (84) allows us to replace λ_0 by $\frac{1}{\lambda_0}$ in the equations from (64) to (69).

Appendix G: Proof of Corollaries 2.4 and 2.5

Proof of Corollary 2.4. Assume that $\pi^* = \mathrm{id}$ for simplicity. Let P_{iso} be the projector on the set of isotonic matrices, and $E' = Y^{(2)}_{\hat{\pi}^{-1}} - M_{\hat{\pi}^{-1}}$ so that $\hat{M}_{\mathrm{iso}} = P_{\mathrm{iso}}(M_{\hat{\pi}^{-1}} + E')$. Remark that the loss can be decomposed as

$$\|(\hat{M}_{\mathrm{iso}})_{\hat{\pi}} - M\|_F^2 = \|P_{\mathrm{iso}}M_{\hat{\pi}^{-1}} - P_{\mathrm{iso}}M + P_{\mathrm{iso}}(M+E') - M + M - M_{\hat{\pi}^{-1}}\|_F^2\ .$$

Using the non-expansiveness of P_{iso} and the triangular inequality as in the proof of proposition 3.3 of [10], we deduce that

$$\|\hat{M}_{\mathrm{iso}} - M\|_F^2 \leq 4\|M_{\hat{\pi}^{-1}} - M\|_F^2 + 2\|P_{\mathrm{iso}}(M+E') - M\|_F^2\ . \tag{86}$$

Since the projection of $M + E'$ on isotonic matrices is equal to the columnwise projection on isotonic vectors, it holds that $\sup_{M\in\mathbb{C}_{\text{iso}}(n,d)} \mathbb{E}\|P_{\text{iso}}(M+E') - M\|_F^2 = d\sup_{M\in\mathbb{C}(n,1)} \mathbb{E}\|P_{\text{iso}}(M_{\cdot 1} + E'_{\cdot 1}) - M_{\cdot 1}\|_F^2$, where we also use the notation P_{iso} for the projector on isotonic vectors. The rate of estimation in L_2 norm of an isotonic vector with bounded total variation partial observation can be found in [24], with $V := 1$ and $\sigma^2 := 1/\lambda$. Hence, we obtain that $\sup_{M\in\mathbb{C}(n,1)} \mathbb{E}\|P_{\text{iso}}(M_{\cdot 1} + E'_{\cdot 1}) - M_{\cdot 1}\|_F^2 \le C_1 n^{1/3}/\lambda^{2/3}$. Upper bounding the first term in (86) with a quantity of order $\rho_{\text{perm}} \le 2\rho_{\text{reco}}$ by Theorem 2.2 concludes the proof.

□

Proof of Corollary 2.5. We follow the same steps as in Corollary 2.4. Assume that $\pi^* = \eta^* = \text{id}$, $E' = Y^{(3)}_{\hat\pi^{-1}\hat\eta^{-1}} - M$, and let P_{biso} be the projector on bi-isotonic matrices. We have that

$$\|(\hat M_{\text{biso}})_{\hat\pi\hat\eta} - M\|_F^2 \le 4\|M_{\hat\pi^{-1}\hat\eta^{-1}} - M\|_F^2 + 2\|P_{\text{biso}}(M+E') - M\|_F^2 \ . \tag{87}$$

M is isotonic in both directions so that we can apply Theorem 2.2 in rows and columns. After the first two steps of the above procedure, we obtain two estimator $\hat\pi, \hat\eta$ that satisfy

$$\sup_{\substack{\pi^*,\eta^*\in\Pi_n \\ M:\, M_{\pi^{*-1}\eta^{*-1}}\in\mathbb{C}_{\text{biso}}}} \mathbb{E}\left[\|M_{\hat\pi^{-1}\hat\eta^{-1}} - M_{\pi^{*-1}\eta^{*-1}}\|_F^2\right] \le C'' \log^{C''}(n) n^{7/6}\lambda^{-5/6} \ . \tag{88}$$

The second term of (87) is the risk of a bi-isotonic regression by least square, and is smaller than $n/\lambda \le n^{7/6}\lambda^{-5/6}$ - see e.g. [10]. □

Appendix H: Proof of the minimax lower bound

Proof of Theorem 2.1. Since $\rho_{\text{perm}}(n,d,\lambda)$ is nondecreasing with n and d, we can assume without loss of generality that both n and d express as a power of 2.

The following proof is strongly related to the proof of Theorem 4.1 in [14]. While a worst case distribution is defined on the set of matrices that have nondecreasing rows and nondecreasing columns in [14], we aim here at defining a worst case distribution on matrices only have nondecreasing columns. Since the isotonic model is less constrained than the bi-isotonic model studied in [14], the permutation estimation problem is statistically harder, and the lower bound has a greater order of magnitude.

As in [14], the general idea is first to build a collection of prior $\nu_{\mathbf{G}}$ indexed by some $\mathbf{G} \in \boldsymbol{\mathcal{G}}$ on M, then to reduce the problem to smaller problems and finally to specify the prior in function of the regime in n, d and λ. By assumption, the data y_t is distributed as a normal random variable with mean M_{x_t} and variance 1, conditionally on M and x_t. We write as in [14] $\mathbf{P}^{(\mathbf{full})}_{\mathbf{G}}$ and $\mathbf{E}^{(\mathbf{full})}_{\mathbf{G}}$ the corresponding marginal probability distributions and expectations on the data (x_t, y_t). Our starting point is the fact that the minimax risk (4) is higher than the worst Bayesian risk:

$$\mathcal{R}^*_{\text{perm}}(n,d,\lambda) \ge \inf_{\hat\pi} \sup_{\mathbf{G}\in\boldsymbol{\mathcal{G}}} \mathbf{E}^{\mathbf{full}}_{\mathbf{G}}\left[\|M_{\hat\pi^{-1}} - M_{\pi^{*-1}}\|_F^2\right] \ . \tag{89}$$

Step 1: Construction of the prior distribution on M

Let $p \in \{2,\dots,n\}$ and $q \in [d]$ be two powers of 2 to be fixed later, and $\overline{G}^{(\iota)} := [(\iota-1)p+1, \iota p]$, for $\iota \in \{1,\dots,n/p\}$. The general idea is to build a simple prior distribution on isotonic matrices in $\mathbb{R}^{\overline{G}^{(\iota)}\times d}$, and to derive a prior distribution on isotonic matrices in $\mathbb{R}^{n\times d}$ by combining n/p independent simple prior distributions defined on each strip $\mathbb{R}^{\overline{G}^{(\iota)}\times d}$.

Let $w \in \mathbb{R}^n$ be a vector that is constant on each group $\overline{G}^{(\iota)} = [(\iota-1)p+1, \iota p]$ and that has linearly nondecreasing steps:

$$w_i = \left\lfloor \frac{i}{p} \right\rfloor \frac{p}{4n} \in [0, 1/4] \ . \tag{90}$$

Letting $\mathbf{1}_{[d]}$ be constant equal to 1 in $\mathbb{R}^d$, we define

$$M = w\mathbf{1}_{[d]}^T + \frac{\upsilon}{\sqrt{p\lambda}} B^{(\mathbf{full})} \ , \tag{91}$$

where the random matrix $B^{(\mathbf{full})} \in \{0,1\}^{n\times d}$ is defined as in [14]. We recall the definition of its distribution in what follows for the sake of completeness.

Consider a collection $\mathcal{G}$ of subsets of $[p]$ with size $p/2$ that are well-separated in symmetric difference as defined by the following lemma.

Lemma H.1. *There exists a numerical constant c_0 such that the following holds for any even integer p. There exists a collection $\mathcal{G}$ of subsets of $[p]$ with size $p/2$ which satisfies $\log(|\mathcal{G}|) \geq c_0|p|$ and whose elements are $p/4$-separated, that is $|G_1 \Delta G_2| \geq p/4$ for any $G_1 \neq G_2$.*

The above result is stated as is in [14] and is a consequence of Varshamov-Gilbert's lemma - see e.g. [22].

For each $\iota \in [n/p]$, we fix a subset $G^{(\iota)}$ from $\mathcal{G}$, and its translation $G^{t(\iota)} = \{(\iota-1)p + x : x \in G^{(\iota)}\} \subset \overline{G}^{(\iota)}$. The experts of $G^{t(\iota)}$ will correspond the $p/2$ experts in $\overline{G}^{(\iota)}$ that are above the $p/2$ experts in $\overline{G}^{(\iota)} \setminus G^{t(\iota)}$. We write $\mathbf{G} = (G^{t(1)}, \dots, G^{t(n/p)})$ and $\boldsymbol{\mathcal{G}}$ the corresponding collection of all possible $\mathbf{G}$. Given any such $\mathbf{G}$, we shall define a distribution $\nu_{\mathbf{G}}$ of $B^{(\mathbf{full})}$, and equivalently of M by (91).

For $\iota \in [n/p]$, we sample uniformly a subset $Q^{(\iota)}$ of q questions among the d columns. In each of these q columns, the corresponding rows of $B^{(\mathbf{full})}$ are equal to one. More formally, we have

$$B^{(\mathbf{full})} = \sum_{\iota=1}^{n/p} \mathbf{1}_{G^{t(\iota)}} \mathbf{1}_{Q^{(\iota)}} \ . \tag{92}$$

As mentioned above, the definition of $B^{(\mathbf{full})}$ is the same as in [14], if $\tilde{d}$ is set to be equal to d. They define a block constant constant matrix when $\tilde{d} < d$ to get an appropriate prior distribution for bi-isotonic matrices, but we do not need to do that here since we do not put any constraint on the rows of M.

The matrix M defined in (92) is isotonic up to a permutation of its rows and has coefficients in $[0,1]$, if the following inequality is satisfied.

$$\frac{\upsilon}{\sqrt{p\lambda}} \leq \frac{p}{8n} \ . \tag{93}$$

This constraint is strictly weaker than its counterpart (149) in [14], and this is precisely what makes the lower bound in the isotonic setting larger than the lower bound in the bi-isotonic setting of [14]. Our purpose will be to wisely choose parameters p, q and $\upsilon > 0$ to maximize the Bayesian risk (89) with $\nu_{\mathbf{G}}$.

Step 2: Problem Reduction

In what follows, we use the same reduction arguments as in [14]. Using the notation of [14], we write $\mathbf{P}_{\mathbf{G}}^{(\mathbf{full})}$ and $\mathbf{E}_{\mathbf{G}}^{(\mathbf{full})}$ for the probability distribution and corresponding expectation of

the data (x_t, y_t), when M is sampled according to $\nu_{\mathbf{G}}$. Since the distribution of the rows of M in $\overline{G}^{t(\iota)}$ only depend on $G^{t(\iota)}$, we write $\nu_{G^{t(\iota)}}$ for the distribution of these rows. We also write $\mathbf{P}^{(\mathbf{full})}_{G^{t(\iota)}}$ and $\mathbf{E}^{(\mathbf{full})}_{G^{t(\iota)}}$ for the corresponding marginal distribution and corresponding expectation of the observations (x_t, y_t) such that $(x_t)_1 \in \overline{G}^{t(\iota)}$. By the poissonization trick, the distribution $\mathbf{P}^{(\mathbf{full})}_{\mathbf{G}}$ is a product measure of $\mathbf{P}^{(\mathbf{full})}_{G^{t(\iota)}}$ for $\iota = 1, \dots, n/p$.

Let $\tilde{\pi}$ be any estimator of π^*. Let us provide more details than [14] to prove that $\tilde{\pi}$ can be modified into an estimator $\hat{\pi}$ satisfying $\hat{\pi}(\overline{G}^{(\iota)}) = \overline{G}^{(\iota)}$ for all $\iota = 1, \dots, n/p$, and reducing the loss $\|M_{\hat{\pi}^{-1}} - M_{\pi^{*-1}}\|_F^2 \le \|M_{\tilde{\pi}^{-1}} - M_{\pi^{*-1}}\|_F^2$ almost surely, for all possible prior $\nu_{\mathbf{G}}$. For that purpose, we introduce

$$N(\pi) = \sum_{\iota=1}^{n/p} \sum_{i \in \overline{G}^{(\iota)}} \mathbf{1}\{i \notin \overline{G}^{(\iota)}\} \ .$$

If $N(\tilde{\pi}) > 0$, then there exists ι_0 and $i_0 \in \overline{G}^{(\iota_0)}$ such that $\tilde{\pi}(i_0) \in \overline{G}^{(\iota_1)}$ with $\iota_1 \neq \iota_0$. Then, $\tilde{\pi}$ being a permutation, we consider its associated cycle containing i_0, which we denote by $(i_1, \dots, i_K)$. Let $(i'_1, i'_2, \dots, i'_L)$ be the elements of this cycle such that $\tilde{\pi}(i'_l) \notin \overline{G}^{(\iota_l)}$, where ι_l satisfies $i'_l \in \overline{G}^{(\iota_l)}$. Then it holds that for any $l = 1, \dots, L-1$, $\tilde{\pi}(i'_l) \in \overline{G}^{(\iota_{l+1})}$, and $\tilde{\pi}(i'_L) \in \overline{G}^{(\iota_1)}$. We now define $\tilde{\pi}'(i) = \tilde{\pi}(i)$ for all i, except on the cycle $(i'_1, \dots, i'_L)$ where we set $\tilde{\pi}'(i'_l) = \tilde{\pi}(i'_{l-1})$. Then, we easily check that $N(\tilde{\pi}') = N(\tilde{\pi}) - L < N(\tilde{\pi})$, and that $\|M_{\tilde{\pi}'^{-1}} - M_{\pi^{*-1}}\|_F^2 \le \|M_{\tilde{\pi}^{-1}} - M_{\pi^{*-1}}\|_F^2$ if condition (93) is satisfied.

We can therefore restrict ourselves to estimators $\hat{\pi}$ such that $\hat{\pi}(\overline{G}^{(\iota)}) = \overline{G}^{(\iota)}$ for all ι. There is however still another catch to obtain the same lines as in [14]. Indeed, the restriction $\hat{\pi}^{(\iota)}$ of $\hat{\pi}$ to $\overline{G}^{(\iota)}$ is measurable with respect to the observation Y, but not necessarily to $Y(\overline{G}^{(\iota)})$. Still, this restriction can be writen as $\hat{\pi}^{(\iota)} = \hat{\pi}^{(\iota)}(Y(\overline{G}^{(\iota)}), Y([n] \setminus \overline{G}^{(\iota)}))$, and, for any $\alpha > 0$, there exists $y^{*(\iota)}(\alpha)$ such that

$$\mathbf{E}^{(\mathbf{full})}_{\mathbf{G}}\left[\|M_{\hat{\pi}^{(\iota)-1}} - M_{\pi^{*-1}}\|_F^2\right] \ge \mathbf{E}^{(\mathbf{full})}_{\mathbf{G}}\left[\|M_{\bar{\pi}^{(\iota)-1}(\alpha)} - M_{\pi^{*-1}}\|_F^2\right] - \alpha \ ,$$

where $\bar{\pi}^{(\iota)} := \hat{\pi}^{(\iota)}(Y(\overline{G}^{(\iota)}), y^{*(\iota)}(\alpha))$ is measurable with respect to $Y(\overline{G}^{(\iota)})$. Since it is possible such a stable estimator for any $\alpha > 0$, we finally obtain the inequality

$$\begin{aligned}
\mathcal{R}^*_{\mathrm{perm}}(n, d, \lambda) &\ge \inf_{\hat{\pi}:\ \hat{\pi}(\overline{G}^{(\iota)}) = \overline{G}^{(\iota)}} \sup_{\mathbf{G} \in \mathcal{G}} \sum_{\iota=1}^{n/p} \mathbf{E}^{(\mathbf{full})}_{\mathbf{G}}\left[\|\left(M_{\hat{\pi}^{-1}} - M_{\pi^{*-1}}\right)_{\overline{G}^{(\iota)}}\|_F^2\right] \\
&\ge \sum_{\iota=1}^{n/p} \inf_{\hat{\pi}^{(\iota)}} \sup_{G^{t(\iota)}} \mathbf{E}^{(\mathbf{full})}_{G^{t(\iota)}}\left[\|\left(M_{\hat{\pi}^{(\iota)-1}} - M_{\pi^{*-1}}\right)_{\overline{G}^{(\iota)}}\|_F^2\right] \ .
\end{aligned}$$

The problem of estimating the permutation π^* is now broken down into the n/p smaller problems of estimating the subsets $G^{t(\iota)} \subset \overline{G}^{(\iota)}$. The square Euclidean distance between to experts in $\overline{G}^{(\iota)}$ of experts is 0 is they are both either in or not in $G^{t(\iota)}$ and it is equal to $\frac{q\upsilon^2}{p\lambda}$ otherwise. Let us focus on the easier problem of estimating the subsets $G^{t(\iota)}$ and define $\hat{G}^{t(\iota)}$ the set of the $p/2$ experts that are ranked above according $\hat{\pi}^{(\iota)}$. Then, we have that

$$\|\left(M_{\hat{\pi}^{(\iota)-1}} - M_{\pi^{*-1}}\right)_{\overline{G}^{(\iota)}}\|_F^2 = \frac{q\upsilon^2}{p\lambda}\left|\hat{G}^{(\iota)} \Delta G^{t(\iota)}\right| \ge \frac{q\upsilon^2}{4\lambda}\mathbf{1}\{\hat{G}^{(\iota)} \neq G^{t(\iota)}\} \ ,$$

where the last inequality comes from the construction of the sets $G^{t(\iota)}$ by Lemma H.1. Hence, we deduce that

$$\mathcal{R}^*_{\text{perm}}(n,d,\lambda) \geq \frac{q\upsilon^2}{4\lambda} \sum_{\iota=1}^{n/p} \inf_{\hat{\pi}^{(\iota)}} \sup_{G^{t(\iota)}} \mathbf{P}^{(\mathbf{full})}_{G^{t(\iota)}} \left[\hat{G}^{(\iota)} \neq G^{t(\iota)} \right] , \tag{94}$$

so that by symmetry,

$$\mathcal{R}^*_{\text{perm}}(n,d,\lambda) \geq \frac{nq\upsilon^2}{4p\lambda} \inf_{\hat{G}^{(1)}} \sup_{G^{t(1)}} \mathbf{P}^{(\mathbf{full})}_{G^{t(1)}} \left[\hat{G}^{(1)} \neq G^{t(1)} \right] .$$

Consider the $p \times d$ matrices N and $Y^{\downarrow}$ defined by

$$N_{ik} = \sum_t \mathbf{1}_{x_t=(i,k)} \; ; \qquad Y^{\downarrow}_{ik} = \sum_t \mathbf{1}_{x_t=(i,k)} (y_t - w_i) \; ,$$

where w is defined in (90). To simplify the notation, we write henceforth G and $\hat{G}$ for $G^{t(1)}$ and $\hat{G}^{(1)}$ respectively. Letting $\mathbf{P}_G$ for the corresponding marginal distribution of N and $Y^{\downarrow}$, the same sufficiency argument as in [14] gives that

$$\inf_{\hat{G}} \sup_{G} \mathbf{P}^{(\mathbf{full})}_{G} \left[\hat{G} \neq G \right] = \inf_{\hat{G}} \sup_{G} \mathbf{P}_G \left[\hat{G} \neq G \right] .$$

We finally obtain the following inequality:

$$\mathcal{R}^*_{\text{perm}}(n,d,\lambda) \geq \frac{nq\upsilon^2}{4p\lambda} \inf_{\hat{G}} \sup_{G} \mathbf{P}_G \left[\hat{G} \neq G \right] . \tag{95}$$

Let $\mathbf{P}_0$ be the distribution on N and $Y^{\downarrow}$ corresponding to the case $\upsilon = 0$. The entries of N of are independent and follow a poisson distribution of parameter λ. Conditionally to N_{ik}, we have $Y^{\downarrow}_{ik}$ is a gaussian variable with mean 0 and variance N_{ik}. Then, we deduce from Fano's inequality [22] that

$$\inf_{\hat{G}} \sup_{G \in \mathcal{G}} \mathbf{P}_G(\hat{G} \neq G) \geq 1 - \frac{1 + \max_{G \in \mathcal{G}} \mathrm{KL}(\mathbf{P}_G \| \mathbf{P}_0)}{\log(|\mathcal{G}|)} , \tag{96}$$

where $\mathrm{KL}(.\|.)$ stands for the Kullback-Leibler divergence. The following lemma gives an upper bound of these Kullback-Leibler divergences. It can be found in [14], with the slightly stronger assumption that $p\lambda \geq 1$.

Lemma H.2 (Lemma J.2 of [14]). *There exists a numerical constant c_1 such that the following holds true. If $\upsilon^2 \leq 1 \wedge p\lambda$, then for any $G \in \mathcal{G}$, we have*

$$\mathrm{KL}(\mathbf{P}_G \| \mathbf{P}_0) \leq c_1 \frac{\upsilon^2 q^2}{d} \quad .$$

The proof of Lemma H.2 can be found in [14], with $\tilde{n} := p$ and $\tilde{d} = d$. The slighlty stronger assumption that $p\lambda \geq 1$ made in Lemma J.2 in [14] is in fact not necessary. Indeed, it is only used to prove that $\mathcal{I} := \lambda p(e^{\upsilon^2/(\lambda p)} - 1) \leq c_1' \upsilon^2$ in the proof of Lemma J.2 in [14], and this inequality remains valid under the assumption of Lemma H.2, that is $u^2 := \upsilon^2/(\lambda p) \leq 1$.

Step 3: Choice of suitable parameters p, q and υ

By combining (95), (96), with Lemma H.2 and the different constraints on the parameters (93), we directly obtain the following proposition.

Proposition H.3. *There exists a numerical constant c such that if $p \in \{2, \dots, n\}$, $q \in \{1, \dots, d\}$ are dyadic integers, and v satisfy the following condition:*

$$v \leq c\left[1 \wedge \sqrt{p\lambda} \wedge \frac{\sqrt{pd}}{q} \wedge \sqrt{\lambda}\frac{p^{3/2}}{n}\right] , \tag{97}$$

then we have

$$\mathcal{R}^*_{\mathrm{perm}}(n, d, \lambda) \geq c\frac{nqv^2}{p\lambda} . \tag{98}$$

The above proposition being a direct consequence of what preceeds it, we consider that it does not require a proof. Let us now apply Proposition H.3 for different parameters p, q and v to conclude the proof of Theorem 2.1.

First, using the lower bound in the bi-isotonic case – see Theorem 4.1 of [14], we have for some constant c' that

$$\mathcal{R}^*_{\mathrm{perm}}(n, d, \lambda) \geq c'(n/\lambda \wedge nd) . \tag{99}$$

In what follows, we write $\lfloor x \rfloor_{\mathrm{dya}}$ for the greatest integer that is a power of two and smaller than x. Let us consider the following inequality:

$$\lambda \geq 1/d \vee n^2/d^3 . \tag{100}$$

In the case where (100) is not satisfied, then $n\sqrt{d/\lambda} \wedge n^{2/3}\sqrt{d}\lambda^{-5/6} \leq n/\lambda \wedge nd$ and the lower bound of Theorem 2.1 is proven by (99).

We subsequently assume that (100) is satisfied.

Case 1: $\lambda n \leq 1$. In this case, we choose $q = \left\lfloor \sqrt{\frac{d}{\lambda}} \right\rfloor_{\mathrm{dya}}$ and $p = n/2$. We have that $q \in \{1, \dots, d\}$ since $\lambda \leq 1$ in that case and by assumption (100), $\lambda \geq 1/d$. We deduce from Proposition H.3 applied with $v/c = \sqrt{p\lambda} = \sqrt{pd}/q$ that

$$\mathcal{R}^*_{\mathrm{perm}}(n, d, \lambda) \geq c'' n\sqrt{\frac{d}{\lambda}} .$$

Case 2: $\lambda \in [\frac{1}{n}, 8n^2]$. In this case, we choose $q = \left\lfloor \frac{n^{1/3}\sqrt{d}}{\lambda^{1/6}} \right\rfloor_{\mathrm{dya}}$ and $p = \left\lfloor \frac{n^{2/3}}{\lambda^{1/3}} \right\rfloor_{\mathrm{dya}}$. We deduce from (100) that $q \leq d$. Since $\lambda \in [\frac{1}{n}, 8n^2]$, we also necessarily have that that $q \geq 1, p \geq 2$ and $p \leq n$. Applying the above proposition with $v/c = 1 = \sqrt{pd}/q = \sqrt{\lambda}p^{3/2}/n$, we deduce that

$$\mathcal{R}^*_{\mathrm{perm}}(n, d, \lambda) \geq c'' \frac{n^{2/3}\sqrt{d}}{\lambda^{5/6}} .$$

Case 3: $\lambda \geq 8n^2$. When λ satisfies this condition that is out of the scope of Theorem 2.1 but discussed below Theorem 2.1, we choose $q = \left\lfloor \sqrt{d} \right\rfloor_{\mathrm{dya}}$ and $p = 2$. Applying the above proposition with $v/c = 1$, we deduce that

$$\mathcal{R}^*_{\mathrm{perm}}(n, d, \lambda) \geq c'' \frac{n\sqrt{d}}{\lambda} .$$

We have proved that for any n, d and λ, we have the lower bound

$$\mathcal{R}^*_{\mathrm{perm}}(n, d, \lambda) \geq c''\left[n\sqrt{\frac{d}{\lambda}} \wedge \frac{n^{2/3}\sqrt{d}}{\lambda^{5/6}} \wedge \frac{n\sqrt{d}}{\lambda} + n/\lambda\right] \wedge nd .$$

This concludes in particular the proof of Theorem 2.1, stated for $\lambda \in [1/d, 8n^2]$.

□

Appendix I: Proof of Proposition 4.1

Let us introduce $\mathbf{P}_k = \Lambda(X_{\cdot k}X_{\cdot k}^T - \mathbb{E}[X_{\cdot k}X_{\cdot k}^T])\Lambda \in \mathbb{R}^{p\times p}$, so that

$$\Lambda(XX^T - \mathbb{E}[XX^T])\Lambda = \sum_{k=1}^{q} \mathbf{P}_k \ . \tag{101}$$

Lemma I.1. *There exists a numerical constant κ_3''' such that for any $x \in [0, (\kappa_3'''(\sigma^2 r_\Lambda + K^2\log(q)))^{-1}]$, we have*

$$\|\mathbb{E}[e^{x\mathbf{P}_k}]\|_{\mathrm{op}} \le \exp(\kappa_3''' x^2(\sigma^2 + \sigma^4 p)) + \frac{1}{q} \ .$$

Moreover, applying the Matrix Chernoff techniques for the independent matrices $\mathbf{P}_k$ (see lemma 6.12 and 6.13 of [23]), we have for any $t > 0$ that

$$\begin{aligned}
\log(\mathbb{P}(\|\sum_{k=1}^{q}\mathbf{P}_k\|_{\mathrm{op}} \ge t)) &\le \log(\mathrm{tr}\left[\mathbb{E}[e^{x\sum_{k=1}^{q}\mathbf{P}_k}]\right]) - xt \\
&\le \log\left(\mathrm{tr}\left[\exp\left(\sum_{k=1}^{q}\log(\mathbb{E}[e^{x\mathbf{P}_k}])\right)\right]\right) - xt \\
&\le \log(p) + \sum_{k=1}^{q}\|\log(\mathbb{E}[e^{x\mathbf{P}_k}])\|_{\mathrm{op}} - xt \\
&= \log(p) + \sum_{k=1}^{q}\log(\|\mathbb{E}[e^{x\mathbf{P}_k}]\|_{\mathrm{op}}) - xt \ .
\end{aligned}$$

Applying Lemma I.1, it holds for any $x \in [0, (\kappa_3'''(\sigma^2 r_\Lambda + K^2\log(q)))^{-1}]$ that

$$\begin{aligned}
\sum_{k=1}^{q}\log(\|\mathbb{E}[e^{x\mathbf{P}_k}]\|_{\mathrm{op}}) &\le q\log\left(\exp\left(\kappa_3''' x^2(\sigma^2 + \sigma^4 p)\right) + \frac{1}{q}\right) \\
&\le \kappa_3''' x^2(\sigma^2 q + \sigma^4 pq) + 1 \ .
\end{aligned}$$

where in the last inequality we used the fact that for any $a \ge 1$ and $u > 0$, $\log(a+u) \le \log(a) + u/a$.

Hence we obtain

$$\log(\mathbb{P}(\|\sum_{k=1}^{q}\mathbf{P}_k\|_{\mathrm{op}} \ge t)) \le \log(ep) + \kappa_3''' x^2(\sigma^2 q + \sigma^4 pq) - xt \ .$$

Hence if $t \le 2\frac{\sigma^2 q + \sigma^4 pq}{\sigma^2 r_\Lambda + K^2\log(q)}$, we choose $x = \frac{t}{2\kappa_3'''(\sigma^2 q + \sigma^4 pq)}$ and if $t > 2\frac{\sigma^2 q + \sigma^4 pq}{\sigma^2 r_\Lambda + K^2\log(q)}$ we choose $x = \frac{1}{\kappa_3'''(\sigma^2 r_\Lambda + K^2\log(q))}$, which gives

$$\mathbb{P}(\|\sum_{k=1}^{q}\mathbf{P}_k\|_{\mathrm{op}} \ge t) \le ep\max\left[\exp\left(-\frac{1}{\kappa_3}\left(\frac{t^2}{4(\sigma^2 q + \sigma^4 pq)} \vee \frac{t}{2(\sigma^2 r_\Lambda + K^2\log(q))}\right)\right)\right] \ .$$

We deduce that with probability at least $1-\delta$, it holds that

$$\|\sum_{k=1}^{q}\mathbf{P}_k\|_{\mathrm{op}} \le \kappa\left[\sqrt{(\sigma^4 pq + \sigma^2 q)\log(p/\delta'')} + (\sigma^2 r_\Lambda + K^2\log(q))\log(p/\delta'')\right] \ ,$$

for some numerical constant κ.

Proof of Lemma I.1. Since $\|\Lambda X_{\cdot k} X_{\cdot k}^T \Lambda\|_{\mathrm{op}} = \|\Lambda X_{\cdot k}\|_2^2$, we state the following lemma controlling the moment generating function of the L_2 norm of the projection $\Lambda X_{\cdot k}$:

Lemma I.2. *There exists a numerical constant κ_0''' such that for any $x \le \frac{1}{\kappa_0''' K^2}$ we have*

$$\mathbb{E}[e^{x\|\Lambda X_{\cdot k}\|_2^2}] \le e^{\kappa_0''' \sigma^2 r_\Lambda x} \ .$$

Now we define the event $\xi_{\mathrm{op}} := \{\max_{k=1,\dots,d} \|\Lambda X_{\cdot k}\|_2^2 \le \kappa_0'''(\sigma^2 r_\Lambda + K^2 \log(q^3))\}$, where κ_0''' is the numerical constant given by Lemma I.2. Applying the same lemma together with the Chernoff bound, a union bound over all $k = 1 \dots d$ gives

$$\mathbb{P}(\xi_{\mathrm{op}}^c) \le \tfrac{1}{q^2} \ .$$

We consider in what follows the relation order $\preceq$ induced by the cone of nonegative symetric matrices $\mathbb{S}_n^+$, namely $X' \preceq X''$ if and only if $X'' - X' \in \mathbb{S}_n^+$. Under the event ξ_{op}, it holds that for any $u \ge 2$,

$$\begin{aligned}
\mathbf{P}_k^u &\preceq \|\mathbf{P}_k\|_{\mathrm{op}}^{u-2} \mathbf{P}_k^2 \\
&\preceq \|\Lambda(X_{\cdot k}^T X_{\cdot k} - \mathbb{E}[X_{\cdot k}^T X_{\cdot k}])\Lambda\|_{\mathrm{op}}^{u-2} \mathbf{P}_k^2 \\
&\preceq (\kappa_1'''(\sigma^2 r_\Lambda + K^2 \log(q)))^{u-2} \mathbf{P}_k^2 \ ,
\end{aligned}$$

for some numerical constant κ_1''' (depending on κ_0'''). In the third inequality we used the definition of ξ_{op} the fact that $\mathbb{E}[\|\Lambda X_{\cdot k}\|_2^2] \le \kappa_0''' \sigma^2 r_\Lambda$.

We now give an upper bound of $\|\mathbb{E}[\mathbf{P}_k^2]\|_{\mathrm{op}}$, which is the operator norm of the variance of $\mathbf{P}_k$ as defined in section 6 in [23]. Remark that since any matrix $U \in \mathbb{R}^{q\times q}$ satisfies $U\Lambda U^T \preceq UU^T$, we have that $\mathbf{P}_k \preceq \Lambda(X_{\cdot k}^T X_{\cdot k} - \mathbb{E}[X_{\cdot k}^T X_{\cdot k}])^2 \Lambda$.

Let us compute the expectation of $(X_{\cdot k}^T X_{\cdot k} - \mathbb{E}[X_{\cdot k}^T X_{\cdot k}])^2$:

$$\mathbb{E}[(X_{\cdot k}^T X_{\cdot k} - \mathbb{E}[X_{\cdot k}^T X_{\cdot k}])^2]_{ij} = \sum_{l \in P} \mathbb{E}[(X_{ik}X_{lk} - \mathbb{E}[X_{ik}X_{lk}])(X_{lk}X_{jk} - \mathbb{E}[X_{lk}X_{jk}])] \ .$$

The off diagonal terms are zero, and the i^{th} diagonal element satisfies:

$$\mathbb{E}[(X_{\cdot k}^T X_{\cdot k} - \mathbb{E}[X_{\cdot k}^T X_{\cdot k}])^2]_{ii} = \mathbb{E}[(X_{ik}^2 - \mathbb{E}[X_{ik}^2])^2] + \sum_{j \ne i} \mathbb{E}[X_{ik}^2]\mathbb{E}[X_{jk}^2] \ . \tag{102}$$

By assumption (24), the first term of (102) satisfies

$$\mathbb{E}[(X_{ik}^2 - \mathbb{E}[X_{ik}^2])^2] \le 4\mathbb{E}[(X_{ik}^4)] \le 48\sigma^2 K^2 \ .$$

The second term of (102) is smaller than $\sigma^4 p$, still by assumption (24). Hence we have some numerical constant κ_2''' that

$$\|\mathbb{E}[\mathbf{P}_k^2]\|_{\mathrm{op}} \le \|\mathbb{E}[(X_{ik}^2 - \mathbb{E}[X_{ik}^2])^2]\|_{\mathrm{op}} \le \kappa_2'''(\sigma^2 + \sigma^4 p) \ .$$

Now, by the definition of the exponential of matrices, the triangular inequality and the fact that $\mathbf{P}_k$ is centered, we have

$$\|\mathbb{E}[\exp(x\mathbf{P}_k)]\|_{\mathrm{op}} = 1 + \sum_{u \ge 2} \frac{x^u}{u!} \|\mathbb{E}[\mathbf{P}_k^u \mathbf{1}_{\xi_{\mathrm{op}}}]\|_{\mathrm{op}} + \sum_{u \ge 2} \frac{x^u}{u!} \|\mathbb{E}[\mathbf{P}_k^u \mathbf{1}_{\xi_{\mathrm{op}}^c}]\|_{\mathrm{op}} \ . \tag{103}$$

By definition of ξ_{op} together with the upper bound of the variance of $\mathbf{P}_k^2 \mathbf{1}_{\xi_{\mathrm{op}}} \preceq \mathbf{P}_k^2$, it holds for any $x \in [0, (\kappa_1'''(\sigma^2 r_\Lambda + K^2 \log(q)))^{-1}]$ that

$$\begin{aligned}\sum_{u\geq 2} x^u \|\mathbb{E}[\mathbf{P}_k^u \mathbf{1}_{\xi_{\mathrm{op}}}]\|_{\mathrm{op}} &\leq x^2 \|\mathbb{E}[\mathbf{P}_k^2]\|_{\mathrm{op}} \sum_{u\geq 2} \frac{x^{u-2}}{u!} (\kappa_1'''(\sigma^2 r_\Lambda + K^2 \log(q)))^{u-2} \\ &\leq x^2 \kappa_2'''(\sigma^2 + \sigma^4 p) \sum_{u\geq 0} \frac{x^u}{(u+2)!} (\kappa_1'''(\sigma^2 r_\Lambda + K^2 \log(q)))^u \\ &\leq \exp(\kappa_3''' x^2(\sigma^2 + \sigma^4 p)) - 1 \ ,\end{aligned}$$

for some numerical constant κ_3'''. We now control the second term of (103) under the complementary event ξ_{op}, for any $x \in [0, (2\kappa_0'''(\sigma^2 r_\Lambda + K^2 \log(q)))^{-1}]$:

$$\begin{aligned}\sum_{u\geq 2} \frac{x^u}{u!} \|\mathbb{E}[\mathbf{P}_k^u \mathbf{1}_{\xi_{\mathrm{op}}^c}] \ &\leq \mathbb{E}[\exp(x\|\mathbf{P}_k\|_{\mathrm{op}} \mathbf{1}_{\xi_{\mathrm{op}}^c}]] \\ &\overset{(a)}{\leq} \sqrt{\frac{1}{q^2}} \sqrt{\mathbb{E}[\exp(2x\|\mathbf{P}_k\|_{\mathrm{op}})]} \\ &\overset{(b)}{\leq} \frac{1}{q} \exp(x\kappa_0''' \sigma^2 r_\Lambda) \\ &\leq \frac{1}{q} \ ,\end{aligned}$$

where in (a) we used the cauchy-schwarz inequality for real random variables and in (b) we applied Lemma I.2.

□

Proof of Lemma I.2. We use the result of [2] which is a generalization of the Hanson-Wright inequality to random variables with coefficients with bernstein's moments.

[Assumption 1 of [2]] is satisfied with parameters σ^2 and K, and we have the following upper bound on the moment generating function of the quadratic form $\|\Lambda X_{\cdot k}^T\|_2^2 = |X_{\cdot k} \Lambda X_{\cdot k}^T|$:

$$\mathbb{E}[e^{x\|\Lambda X_{\cdot k}^T\|_2^2}] \leq e^{x\mathbb{E}[\|\Lambda X_{\cdot k}^T\|_2^2]} e^{\kappa_0''' x^2 K^2 \sigma^2 \|\Lambda\|_F^2} \leq e^{\kappa_1''' x \sigma^2 r_\Lambda} \ , \tag{104}$$

for any x satisfying condition (6) of [2], that is $128x\|\Lambda\|_{\mathrm{op}} K^2 \leq 1$. For the last inequality, we used the fact that $\|\Lambda\|_F^2 = \mathrm{rank}(\Lambda)$. We obtain the result by choosing $\kappa_2''' = \kappa_1''' \vee 128$. □

References

[1] A. S. Bandeira and R. Van Handel. Sharp nonasymptotic bounds on the norm of random matrices with independent entries. 2016.

[2] P. C. Bellec. Concentration of quadratic forms under a bernstein moment assumption. arXiv preprint arXiv:1901.08736, 2019.

[3] V. Bengs, R. Busa-Fekete, A. El Mesaoudi-Paul, and E. Hüllermeier. Preference-based online learning with dueling bandits: A survey. The Journal of Machine Learning Research, 22(1):278–385, 2021.

[4] R. A. Bradley and M. E. Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. Biometrika, 39(3/4):324–345, 1952.

[5] M. Braverman and E. Mossel. Noisy sorting without resampling. In Proceedings of the nineteenth annual ACM-SIAM symposium on Discrete algorithms, pages 268–276, 2008.

[6] S. Chatterjee and S. Mukherjee. Estimation in tournaments and graphs under monotonicity constraints. IEEE Transactions on Information Theory, 65(6):3525–3539, 2019.

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

[8] A. Liu and A. Moitra. Better algorithms for estimating non-parametric models in crowdsourcing and rank aggregation. In Conference on Learning Theory, pages 2780–2829. PMLR, 2020.

[9] C. Mao, A. Pananjady, and M. J. Wainwright. Breaking the $1/\sqrt{n}$ barrier: Faster rates for permutation-based models in polynomial time. In Conference On Learning Theory, pages 2037–2042. PMLR, 2018.

[10] C. Mao, A. Pananjady, and M. J. Wainwright. Towards optimal estimation of bivariate isotonic matrices with unknown permutations. The Annals of Statistics, 48(6):3183–3205, 2020.

[11] P. Massart. Concentration inequalities and model selection, volume 6. Springer, 2007.

[12] L. Mirsky. A dual of dilworth's decomposition theorem. Amer. Math. Monthly, 78(8):876–877, 1971.

[13] A. Pananjady and R. J. Samworth. Isotonic regression with unknown permutations: Statistics, computation and adaptation. The Annals of Statistics, 50(1):324–350, 2022.

[14] E. Pilliat, A. Carpentier, and N. Verzelen. Optimal permutation estimation in crowdsourcing problems. arXiv preprint arXiv:2211.04092, 2022.

[15] P. Rigollet and J. Weed. Uncoupled isotonic regression via minimum wasserstein deconvolution. Information and Inference: A Journal of the IMA, 8(4):691–717, 2019.

[16] E. M. Saad, N. Verzelen, and A. Carpentier. Active ranking of experts based on their performances in many tasks. arXiv preprint arXiv:2306.02628, 2023.

[17] N. Shah, S. Balakrishnan, J. Bradley, A. Parekh, K. Ramchandran, and M. Wainwright. Estimation from pairwise comparisons: Sharp minimax bounds with topology dependence. In Artificial intelligence and statistics, pages 856–865. PMLR, 2015.

[18] N. B. Shah, S. Balakrishnan, A. Guntuboyina, and M. J. Wainwright. Stochastically transitive models for pairwise comparisons: Statistical and computational issues. IEEE Transactions on Information Theory, 63(2):934–959, 2016.

[19] N. B. Shah, S. Balakrishnan, and M. J. Wainwright. Feeling the bern: Adaptive estimators for bernoulli probabilities of pairwise comparisons. IEEE Transactions on Information Theory, 65(8):4854–4874, 2019.

[20] N. B. Shah, S. Balakrishnan, and M. J. Wainwright. A permutation-based model for crowd labeling: Optimal estimation and robustness. IEEE Transactions on Information Theory, 67(6):4162–4184, 2020.

[21] J. A. Tropp. An Introduction to Matrix Concentration Inequalities. ArXiv e-prints, Jan 2015.

[22] A. B. Tsybakov. Introduction to Nonparametric Estimation. 2008.

[23] M. J. Wainwright. High-dimensional statistics: A non-asymptotic viewpoint, volume 48. Cambridge University Press, 2019.

[24] C.-H. Zhang. Risk bounds in isotonic regression. The Annals of Statistics, 30(2):528–555, 2002.