Recovering Labels from Crowdsourced Data: An Optimal and Polynomial-Time Method

Emmanuel Pilliat EMMANUEL.PILLIAT@ENSAI.FR
ENSAI, 51 Rue Blaise Pascal, 35170 Bruz

Editors: Nika Haghtalab and Ankur Moitra

Abstract

Crowdsourcing involves aggregating meaningful information from partial and noisy data provided by a pool of n workers across d tasks. Traditional models, such as the Dawid-Skene model, assume that workers' abilities are independent of tasks, limiting their applicability in real-world scenarios where worker ability often varies significantly across tasks. Recent advances have proposed permutation-based models, which relax these assumptions by imposing only isotonicity constraints on worker abilities. In this work, we study a permutation-based model where each worker i has an ability M_{ik} to recover a binary label $x_k^* \in \{-1, 1\}$ for task k. The ability matrix M is assumed to be isotonic up to a permutation of its rows, and only a fraction λ of the worker-task pairs is observed.

We focus on three primary objectives: recovering the true labels, ranking the workers, and estimating the ability matrix M. We introduce a polynomial-time and minimax optimal procedure to recover the labels, contradicting a conjecture in the literature regarding the existence of a statistical-computational gap for this problem. Additionally, building on the literature on ranking, we further introduce a polynomial-time procedure to rank the workers and to estimate their abilities. Notably, we show that ranking the workers or estimating their abilities is no harder when the true labels are unknown than when they are known, within the main regimes of interest in the isotonic model.

Keywords: Crowdsourcing, Ranking, Isotonic, Minimax

1. Introduction

With the rise of data-driven technologies, crowdsourcing has become a critical tool for labeling large datasets efficiently and affordably. Crowdsourcing platforms enable the rapid collection of labels from a broad pool of workers, opening new avenues for machine learning and data analysis applications. However, this approach introduces unique challenges: the non-expert nature of many crowd workers leads to noisy labels, and there is often considerable variation in labeling quality across different workers and tasks. These factors make it essential to develop robust models and efficient algorithms that can aggregate and refine noisy, heterogeneous labels, ensuring reliable data quality for applications.

In this manuscript, we consider a crowdsourcing problem involving n *workers* and d *tasks*. Each task k requires providing an unknown binary *label* $x_k^* \in \{-1, 1\}$, while each worker i has an unknown ability $M_{ik} \in [0, 1]$ specific to task k. The main purpose of this manuscript is to recover the true labels x_k^* based on the workers' responses $Y_{ik} \in \mathbb{R}$, which will be modeled as partial and noisy observations of both the ability matrix M and the vector of true labels $x^\star$.

The celebrated Dawid and Skene (DS) model Dawid and Skene (1979) makes the assumption that the abilities of the workers do not depend on the task k, that is $M_{ik} = M_i$. Although the DS model and its variants have inspired a substantial body of literature, Ghosh et al. (2011); Dalvi et al. (2013); Karger et al. (2011), several authors have observed that these parametric models often fail

to fit the data accurately Welinder et al. (2010); Whitehill et al. (2009). This is because tasks are often not equivalent, and the ability of worker i can vary significantly depending on the task k.

Consequently, much attention has recently been directed toward developing permutation-based models Mao et al. (2020); Liu and Moitra (2020); Pananjady and Samworth (2022); Shah et al. (2016); Pilliat et al. (2024, 2023), in the simpler case where the true labels are known. In these models, the ability matrix M is assumed to satisfy isotonicity or bi-isotonicity constraints up to an unknown permutation of its rows and/or columns. It has been shown in Mao et al. (2020); Shah et al. (2016) that, in several non-parametric settings, the matrix M can be estimated at the same rate as in classical parametric models such as the DS model with known labels. Moreover, in the isotonic and bi-isotonic settings analyzed in Pilliat et al. (2024, 2023), polynomial-time methods were shown to achieve optimal rates for both estimating the permutation and the ability matrix M. The main analyses in the literature on the permutation-based models of optimal and polynomial methods make however heavily use of the knowledge of the labels to provide a ranking of the workers or to estimate the matrix M. This is not appropriate for crowdsourcing problems where one of the main objectives is to infer the labels x_k^* based on the workers' knowledge.

1.1. The Bi-Isotonic Permutation-Based Model with Unknown Labels

Relaxing the hypothesis of the knowledge of the labels, Shah et al. (2020) recently proposed a permutation-based model in which x^* is unknown and the ability matrix M is bi-isotonic up to an unknown permutation π^* of its rows and an unknown permutation σ^* of its columns. In Shah et al. (2020), each worker i considers a given task k with probability $\lambda \in [0,1]$. If worker i considers task k, they provide a response $Y_{ik} \in \{-1,1\}$, which is correct with probability $\frac{1+M_{ik}}{2}$. If worker i does not consider task k, which happens with probability $1-\lambda$, the response is $Y_{ik} = 0$.

$$Y_{ik} = \begin{cases} x_k^* \text{ with probability } \lambda\left(\frac{1+M_{ik}}{2}\right) \\ -x_k^* \text{ with probability } \lambda\left(\frac{1-M_{ik}}{2}\right) \\ 0 \text{ with probability } 1-\lambda \end{cases} . \tag{1}$$

In the model of Shah et al. (2020), there is an unknown and implicit ranking of the workers by ability, but also an unknown and implicit ranking of the tasks by difficulty. For any two workers i and j, one is uniformly better in expectation than the other on all tasks, and for any two tasks k and l, one is uniformly harder for all workers. This extends in particular the Strong Stochastically Transitive model (SST) introduced in Shah et al. (2016) to to the rectangular case where $n \leq d$ and to the case where the labels x^* are unknown. In the SST model, it is assumed that $n = d$, $\pi^* = \sigma^*$ and that x^* is known. Surprisingly, the authors established that it is not only possible to recover the labels at the same rate as in SST, it is also possible to recover the labels at the same rate as in much simpler classical parametric models, e.g. when M_{ik} is independent of k and when $n = d$. The optimal method of Shah et al. in model (3) is, however, conjectured not to be computationally tractable, leading them to conjecture the existence of an intrinsic computational-statistical gap in the problem of recovering the labels.

1.2. Our Contributions.

We identify three key tasks relevant to extracting information from crowdsourced data, which we informally describe as follows:

1. *Recovering* the labels x^*
2. *Ranking* the workers by estimating the permutation π^*
3. *Estimating* the ability matrix M of the workers

A major challenge is that these tasks are generally interdependent. For instance, knowing the workers' abilities could significantly help in recovering the labels, and having accurate labels would help evaluating and ranking the workers. This complexity is highlighted by the main optimization problem proposed in Shah et al. (2020) to recover the labels. In short, the approach consists in minimizing a least square objective over the non-convex set of all vector of labels x, over all permutations π and σ of the rows and columns and over all matrices M that are bi-isotonic up to π and σ. Implicitly, the least-square optimization procedure estimates the ability matrix M and the permutation π^* to recover the labels x^*. Despite being theoretically optimal (when $n = d$), the least square method is conjectured to be computationally infeasible.

In this manuscript, we address the three above objectives of recovering the labels, ranking the workers and estimating the abilities of the workers in polynomial time. We conduct our analysis within the isotonic model, where M is only assumed to have increasing columns up to a permutation of its rows. In contrast, Shah et al. (2020) assume the additional constraint that M has increasing rows up to a permutation of its columns.

Our main contribution is to provide an optimal and polynomial time algorithm, the Iterative Spectral Voting (**ISV**), to compute an estimator $\hat{x}$ of the labels x^*. Specifically, in terms of the risk $\mathbb{E}[\|M \operatorname{diag}(\hat{x} \neq x^*)\|_F^2]$ introduced by Shah et al. (2020), we prove that our estimator achieves an upper-bound of order d/λ, which is optimal when $\frac{1}{\lambda} \leq n \leq d$ up to polylogarithmic factors. In contrast to Shah et al. (2020), our procedure is computationally tractable and minimax optimal in both the isotonic model and the easier bi-isotonic model of Shah et al. We establish in particular that there is no computational-statistical gap for recovering the labels, contradicting the conjecture of Shah et al. (2020).

Secondly, we rely on our procedure, **ISV**, and on ranking techniques introduced in Pilliat et al. (2023), to provide polynomial-time estimators of the true ranking of the workers π^* and of the ability matrix M. We show that the corresponding estimators, $\hat{\pi}$ and $\hat{M}$, achieve optimal ability estimation risk $\mathbb{E}[\|\hat{M} - M\|_F^2]$ and optimal ranking risk $\mathbb{E}[\|M_{\hat{\pi}^{-1}} - M_{\pi^{*-1}}\|_F^2]$ in the case $n \gtrsim d^{3/4}\lambda^{-1/4}$ up to polylogarithmic factors. In particular, our procedures for ranking and estimating M are optimal in the square case $n = d$ and $\lambda \geq 1/n$, which is a standard setting in permutation-based models for ranking Liu and Moitra (2020); Shah et al. (2016); Mao et al. (2020). Surprisingly, a direct consequence of our results is that when $n \gtrsim d^{3/4}\lambda^{-1/4}$, ranking and estimation are not significantly more difficult when x^* is unknown than when it is known.

Lastly, we provide a numerical study comparing a slightly modified version of **ISV** with a naive majority voting approach and with the polynomial time method for recovering binary labels introduced in Shah et al. (2020), Obi-Wan. Our results suggest that our approach performs well in practice, in scenarios where M cannot be closely approximated by a matrix of rank 1.

1.3. Technical Overview

The algorithm **ISV** returns an estimator $\hat{x}$ of the labels, based on the matrix of responses Y of the workers. This involves subsampling the matrix Y by assigning each non-zero coefficient of Y uniformly to one of the T subsamples $(Y^{(1)}, \dots, Y^{(T)})$. At each step of the procedure, we use a spectral method to estimate weights for the n workers. These weights are then applied to

compute a weighted vote for each label that has not yet been determined. The true labels for which the corresponding weighted vote exceeds a certain threshold in absolute value are then estimated according to this same weighted vote.

Unlike the OBI-WAN method proposed in Shah et al. (2020), which performs only two spectral steps, **ISV** repeats a spectral step a polylogarithmic number of times. Performing two spectral steps allows to reach the minimax rate in the simpler DS model, where M is assumed to be of rank 1. This iterative approach allows to adapt to more general isotonic matrices, which are not necessarily of rank 1, and to reach the optimal minimax rate in polynomial time. Moreover, **ISV** simplifies the way the workers are aggregated to recover labels, compared to Shah et al. (2020). Instead of relying on a WAN step as in Shah et al. (2020), which first identifies top experts based on a leading singular vector before aggregating their responses through majority voting, **ISV** directly computes a weighted vote using a leading singular vector. Furthermore, **ISV** outputs a final estimator $\hat{x}$ that takes values in $\{-1, 0, 1\}$ rather than in $\{-1, 1\}$. We ensure in our proofs that the estimated labels that are in $-1, 1$ correspond to the correct labels with high probability. In particular, as we show in the proof of Theorem 3, this allows to restrict to correctly labeled tasks to optimally rank the workers and estimate M.

Another important aspect of this paper is the use of a more realistic partial Bernoulli observation scheme – where each coefficient is observed with probability λ – instead of the simpler partial Poisson observation scheme, where each coefficient (i, k) has a random number of observations that follows a Poisson distribution of parameter λ. Many previous works on the permutation-based model Mao et al. (2020); Liu and Moitra (2020); Pilliat et al. (2023) have focused on the Poisson observation scheme, which allows to define independent subsamples from the data. To the best of our knowledge, subsampling the data into independent subsamples $(Y^{(1)}, \dots, Y^{(T)})$ when Y_{ik} follows the Bernoulli model (1) is not possible in general. For this reason, we use a Bernoulli observation scheme to define subsamples that are only independent conditionally on the observation scheme. The main resulting technical difficulties lie in the dependence between the leading singular vector derived from the spectral method and the weighted vote. Consequently, we provide both concentration bounds on the observation scheme, and on the noise conditionally on the observation scheme.

2. Problem Formulation

In this section, we formally define the problem in a general sub-Gaussian setting that encompass the Bernoulli model (3) introduced in Shah et al. (2020). We also discuss the Poisson observation scheme that is usually analyzed for technical simplification (independent subsampling) in ranking models Liu and Moitra (2020); Mao et al. (2020); Pilliat et al. (2023).

2.1. General Setting for the Isotonic Permutation-Based Model

Assume that $M \in [0, 1]^{n \times d}$ is isotonic up to an unknown permutation π^* of its rows, that is, for all $i = 1, \dots, n-1$:

$$M_{\pi^{*-1}(i),k} \leq M_{\pi^{*-1}(i+1),k} \quad . \tag{2}$$

To define the partial Bernoulli observation scheme of Shah et al. (2020), let B_{ik} be a matrix of independent Bernoulli random variables with parameter $\lambda \in [0, 1]$ and let $\odot$ be the Hadamard product, that is $(A \odot A')_{ik} = A_{ik}A'_{ik}$ for any matrix A, A'. Worker i considers task k when $B_{ik} = 1$, and

does not consider task k when $B_{ik} = 0$. We observe the matrix of responses

$$Y = B \odot (M \operatorname{diag}(x^*) + E) \ , \tag{3}$$

where E is a matrix whose coefficients (E_{ik}) are independent, centered, and follow 1-sub-Gaussian distributions. This means that for any $x \in \mathbb{R}$, the moment generating function satisfies $\mathbb{E}[e^{xE_{ik}}] \leq e^{x^2/2}$.

Let us relate this model to general crowdsourcing problems. For each worker-task pair (i, k), worker i considers task k with probability $\lambda \in [0, 1]$. If worker i considers task k, the observed value is given by $Y_{ik} = M_{ik}x^*_k + E_{ik}$. In other words, the responses of the workers are biased toward the true label x^*_k, with the bias increasing as the ability M_{ik} becomes larger. Hence, a positive value of Y_{ik} indicates that worker i estimates the label of task k as 1, while a negative value indicates an estimation of the label as -1. The absolute value $|Y_{ik}|$ reflects worker i's confidence in their estimation.

Our model (3) encompasses the Bernoulli noise model defined in (1) and considered by Shah et al. Shah et al. (2020). Indeed, in (1), the responses Y_{ik} for which $B_{ik} = 1$ can be written $Y_{ik} = \big(2\mathcal{B}(\frac{1+M_{ik}}{2}) - 1\big)x^*_k = M_{ik}x^*_k + 2\big(\mathcal{B}(\frac{1+M_{ik}}{2}) - \frac{1+M_{ik}}{2}\big)x^*_k$ where the $\mathcal{B}(\frac{1+M_{ik}}{2})$ are independent Bernoulli random variables of parameters $\frac{1+M_{ik}}{2}$. The random variables $2\big(\mathcal{B}(\frac{1+M_{ik}}{2}) - \frac{1+M_{ik}}{2}\big)x^*_k$ are independent, centered and bounded by 2. In particular, they are 1-sub-Gaussian by Hoeffding's inequality.

Relation to the Poisson Observation Scheme. In the general model (3), we assume that for each worker-task pair, there is either 0 or 1 observation. However, many permutation-based models in the literature Mao et al. (2020, 2018); Liu and Moitra (2020); Pilliat et al. (2024) adopt a different approach, assuming that there are $N_{ik} \geq 0$ independent observations for each worker-task pair (i, k). The authors assume that the (N_{ik})'s are independent Poisson random variables with parameter λ', which allows them to obtain a polylogarithmic number of truly independent samples from the observations. This is known as the Poissonization trick.

The Poisson observation scheme can easily be reduced to the Bernoulli observation scheme (3). Given N_{ik} observations, set $Y_{ik} = 0$ if $N_{ik} = 0$ and select one of the observations among the N_{ik} observations for the value of Y_{ik} if $N_{ik} > 0$. Then, the probability of observing at least one sample in this case is $\mathbb{P}(N_{ik} \neq 0) = (1 - e^{-\lambda'}) \geq (1 - 1/e)\lambda'$. The converse, reducing from Poisson to Bernoulli observation scheme is not possible in general. There is no way to generate two independent Bernoulli random variables with an unknown parameter M_{ik} if only one is available. The assumption of having a Poisson number of observations can be unrealistic in crowdsourcing problems, as it implies that, on average, $O(nd\lambda'^2)$ entries of Y have at least two independent observations. This is questionable in practice, since a worker i typically does not provide independent answers for the same task k.

2.2. Error Measures

Our main task in this work is to recover the labels x^* from the responses of the workers Y. We define the loss of a given estimator $\hat{x}$ as the squared Frobenius norm of M, whose columns are restricted to incorrectly labeled tasks k. The risk is defined as the expectation of the loss.

$$\mathcal{L}_{M,x^*}(\hat{x}) := \|M \operatorname{diag}(\hat{x} \neq x^*)\|_F^2 \quad \text{and} \quad \mathcal{R}_{M,x^*}(\hat{x}) = \mathbb{E}[\mathcal{L}_{M,x^*}(\hat{x})] \ . \tag{4}$$

We use the notation $\mathrm{diag}(\hat{x} \neq x^*)$ for $d \times d$ the diagonal matrix where each diagonal coefficient is equal to 0 if $\hat{x}_k = x^*_k$ and 1 if $\hat{x}_k \neq x^*_k$, and the loss is equal to $\sum_{i,k} M^2_{ik} \mathbf{1}\{\hat{x}_k \neq x^*_k\}$. In the Bernoulli noise model (1), this loss matches the Q^*-loss introduced by Shah et al. Shah et al. (2020), up to a $4nd$ scaling factor. As noted in Shah et al. (2020), the quantity $\|M\|_F^2$ generalizes, up to a scaling factor, the *collective intelligence* of the workers. This terminology was originally introduced in classical crowdsourcing models such as those of Dawid and Skene Dawid and Skene (1979), for crowdsourcing problems where the ability of the experts does not depend on the tasks. The loss (4) reflects the collective intelligence of the workers, restricted to incorrectly labeled tasks. Hence, the greater the remaining collective intelligence on incorrectly predicted labels, the more the estimation is penalized.

The loss that is usually considered in the crowdsourcing literature is the Hamming loss, which consists in counting the number of incorrect label $\sum_{k=1}^{d} \mathbf{1}\{\hat{x}_k \neq x^*_k\}$. However, the Hamming loss harshly penalizes the evaluation of the estimator $\hat{x}$ when the workers perform poorly. For example, when $M_{ik} = 0$ for all (i,k), no estimator of the labels $\hat{x}$ from the workers' responses performs better than random guesses, and the Hamming loss is of order d – see also Theorem 4 (b) of Shah et al. (2020). In this situation, the loss based on collective intelligence (4) is equal to zero for any estimator $\hat{x}$. This is a key advantage for the loss (4) – and of the Q^* loss in Shah et al. (2020) – as it evaluates the performance of the estimator $\hat{x}$ itself, rather than the performances of the workers.

For a given estimator $\hat{x}$ in $\{-1,1\}^d$, we define the Maximum risk as the supremum of the risk over all $x^* \in \{-1,1\}^d$ and over all matrices M that are isotonic up to a permutation π^*. The minimax risk is then the infimum over all possible $\hat{x}$ of the maximum risk:

$$\mathcal{R}^*(n,d,\lambda) = \inf_{\hat{x}} \sup_{x^*,M,\pi^*} \mathbb{E}[\mathcal{L}_M(\hat{x},x^*)] \quad (\textit{Recovering Labels}) \ . \tag{5}$$

This minimax risk quantifies the minimal achievable risk for recovering the labels in the worst case. We also introduced the two minimax risks for ranking the workers and for estimating the ability matrix M:

$$\begin{aligned} \mathcal{R}^*_{(\mathrm{rk})}(n,d,\lambda) &= \inf_{\hat{x}} \sup_{x^*,M} \mathbb{E}[\|M_{\hat{\pi}} - M_{\pi^*}\|_F^2] \quad (\textit{Ranking Workers}) \ , \\ \mathcal{R}^*_{(\mathrm{est})}(n,d,\lambda) &= \inf_{\hat{M}} \sup_{x^*,M} \mathbb{E}[\|\hat{M} - M\|_F^2] \quad (\textit{Estimating Abilities}) \ . \end{aligned} \tag{6}$$

The ranking loss $\|M_{\hat{\pi}} - M_{\pi^*}\|_F^2$ was first introduced by Liu and Moitra (2020) in the case where x^* is known and when M is bi-isotonic up to two permutations. Minimizing this ranking loss was also the objective of Pilliat et al. (2024) when x^* is known and M is isotonic and in Pilliat et al. (2023) when M is bi-isotonic up to a permutation of its rows. The reason for considering this loss rather than a more common loss on permutations that does not depend on M, such as the Kendall-Tau distance $\sum_{\pi^*(i)<\pi^*(j)} \mathbf{1}\{\hat{\pi}(i) > \hat{\pi}(j)\}$, is analogous to the observation made above concerning the Hamming loss between $\hat{x}$ and x^*: when i and j are indistinguishable, meaning $M_{i\cdot} = M_{j\cdot}$, it is no better than random guessing to determine whether i should be ranked above or below j.

Minimizing the estimation loss $\|\hat{M} - M\|_F^2$ is the most common objective in the literature, both in classical models Dawid and Skene (1979); Ghosh et al. (2011) and permutation-based models Shah et al. (2020); Mao et al. (2020, 2018); Liu and Moitra (2020); Pilliat et al. (2024, 2023). In the model with Bernoulli noise and when M has constant rows, estimating M amounts to estimating what is classically called the *confusion matrices*, which are in this case the 2×2 symetric matrices

$\frac{1}{2} + \frac{1}{2}\begin{pmatrix} M_{ik} & -M_{ik} \\ -M_{ik} & M_{ik} \end{pmatrix}$. In contrast to classical models Dawid and Skene (1979); Ghosh et al. (2011), the confusion matrix depends here on task k.

3. Results

In this section, we present the results on label recovery, expert ranking, and ability estimation within the isotonic model. We first establish in Section 3.1 that the labels x^* can be recovered in polynomial-time, while achieving the minimax risk (5) up to a polylogarithmic factor. Then, in Section 3.2 we show how to combine the procedure **ISV** for recovering labels with other ranking and estimation methods to achieve the ranking and estimation minimax risks (6). The definition of our procedure **ISV** is postponed to Section 4.

3.1. Recovering the Labels

The procedure $\mathbf{ISV}(Y, T, \delta)$, takes as input the data Y, an integer T, and a parameter $\delta \in (0, 1)$. Ultimately, it outputs an estimator $\hat{x} \in \{-1, 0, 1\}^d$ of x^*, where 0 means that the estimation is uncertain. To avoid some unnecessary technical complications while maintaining clarity in the upper-bounds, we define explicitly ψ as a large quantity of order $\log(ndT/\delta)$: $\psi := \psi(n, d, T, \lambda, \delta) = 10^{10} \log\left(\frac{ndT}{\lambda\delta}\right)$ The following theorem establishes that, with high probability, the loss for recovering the labels (4) is at most of order d/λ. Moreover, it shows that, with high probability, the final estimator $\hat{x}_k$ given by **ISV** is either equal to 0 or to the true label x^*_k.

Theorem 1 (Upper Bound) *Assume that $\frac{1}{\lambda} \le n \le d$, and that $T \ge 1024e^2 \log^2(end)$. Then, the estimator $\hat{x}$ obtained from $\mathbf{ISV}(Y, T, \delta)$ (see Section 4) satisfies*

$$\hat{x}_k \in \{0, x^*_k\} \text{ for all } k \in [d] \quad \text{and} \quad \|M \operatorname{diag}(\hat{x} \neq x^*)\|_F^2 \le CT\psi^5 \frac{d}{\lambda} \ ,$$

for some numerical constant C, with probability at least $1 - 6\delta$.

The first point means that, with high probability, **ISV** recovers the correct label of each task k or returns 0 if uncertain. The second point implies that, if T is at least of order $\log^2(nd)$, the estimator $\hat{x}$ satisfies $\|M \operatorname{diag}(\hat{x} - x^*)\|_F^2 \lesssim \frac{d}{\lambda}$ up to a polylogarithmic factor. When $n \gg \frac{1}{\lambda}$, the rate $\frac{d}{\lambda}$ is significantly better than the current state-of-the-art polynomial-time method, OBI-WAN, which achieves only a rate of order $\sqrt{\frac{n}{\lambda}}d$ – see Proposition 1 in Shah et al. (2020). In fact, the rate of our method matches the rate achieved by the conjectured computationally hard least squares method which was established in Shah et al. (2020), to be optimal in the simpler bi-isotonic model in Bernoulli noise.

As mentioned above, a lower bound of order d/λ in the Bernoulli noise model can be found in Shah et al. (2020). For the sake of completeness, the following theorem presents a slightly refined lower bound in the Gaussian noise model with explicit constant.

Theorem 2 (Lower Bound) *Assume that $\frac{1}{\lambda} \le n$ and that E is a matrix with i.i.d. standard Gaussian entries. Let $g(n, \lambda) = \frac{9n\lambda}{9n\lambda+8}$. Then, there exists a matrix M with constant coefficients such that for any estimator $\hat{x}$ (that knows M),*

$$\sup_{x^* \in \{-1,1\}^d} \mathbb{E}[\|M \operatorname{diag}(x^* \neq \hat{x})\|_F^2] \ge g(n, \lambda) \frac{d}{27\lambda} \ , \tag{7}$$

Since $n\lambda \geq 1$, the function g satisfies $g(n,\lambda) \geq \frac{9}{17}$. The lower bound provided in Shah et al. (2020) considers a worst case where M is known but not necessarily with constant coefficients. In contrast, Theorem 2 establishes a lower bound in the easier model where M is known and constant. Since our lower bound matches the order $\frac{d}{\lambda}$ of the upper bound Theorem 1, this indicates that the model where M is known and constant is almost just as hard as the non-parametric isotonic model analyzed in this paper – up to polylogs. In fact, we prove the lower bound in the easier Poisson observation scheme defined at the end of Section 2.1. Our proof shows in particular that the Poisson observation scheme is not much easier than the Bernoulli observation scheme in the problem of recovering the labels.

3.2. Ranking the Workers and Estimating their Abilities

Label recovery, ranking workers, and estimating abilities are closely intertwined. Interestingly, as demonstrated in the definition of the **ISV** procedure – see Section 4, – achieving optimal label recovery in a minimax sense does not necessarily require explicit worker ranking or ability estimation. Conversely, building upon the **ISV** procedure and methods introduced by Pilliat et al. (2024), we can derive minimax approaches for both ranking workers and estimating their abilities. This section explores how we derive minimax optimal procedures from **ISV** and Pilliat et al. (2024) to rank the workers and estimate their abilities.

the results on ranking and estimation in Pilliat et al. (2024) are valid in the easier Poisson observation scheme – see the end of Section 2.1. To relate to and leverage the results in ranking literature applicable to the Poisson observation scheme, we assume in what follows that there are N_{ik} independent observations for each coefficient (i,k), where N_{ik} has parameter $\lambda' \in (0,1)$. Under this Poisson observation scheme, we can define $5T+2$ independent subsamples, as described in Equation (14) of Pilliat et al. (2024), for any integer $T \geq 1$. We describe our procedure to derive the estimators $\hat{\pi}$ and $\hat{M}$ as follows:

1. Use T independent subsamples to run **ISV** to obtain an estimator $\hat{x}$.
2. Use $5T$ subsamples restricted to columns k such that $\hat{x}_k \neq 0$, and run the procedure **ISR** defined in Pilliat et al. (2024), with a valid grid Γ chosen by default as in section 3.6 of Pilliat et al. (2024). It returns an estimator $\hat{\pi}$ of π^*.
3. Use the last subsample Y', also restricted to columns k such that $\hat{x}_k \neq 0$ to get an estimator $\hat{M}$ of M as described in Section 2.3 of Pilliat et al. (2024). This step is based on computing the minimizer of the quantity $\|\tilde{M} - Y'_{\hat{\pi}^{-1}}\|_F^2$, over all isotonic matrices $\tilde{M}$.

The steps described above can be computed in polynomial time, and allow to derive two estimators $\hat{\pi}$ and $\hat{M}$. Assume for simplicity that T is chosen as a large polylogarithmic factor, that is $T = \lceil 1024e^2 \log^2(end) \rceil \vee 4\lceil \overline{\gamma}^6 \rceil$, where $\overline{\gamma}$ is defined in equation (11) of Pilliat et al. (2024). The following result shows an upper-bound on the ranking and estimation risks.

Theorem 3 *Assume that $\frac{1}{\lambda} \leq n \leq d$. There exists a numerical constant C such that for any permutation π^* and any matrix M such that $M_{\pi^{*-1}}$ is isotonic,*

$$\mathbb{E}[\|M_{\hat{\pi}^{-1}} - M_{\pi^{*-1}}\|_F^2] \leq C \log^C(\tfrac{nd}{\lambda}) \left(\frac{d}{\lambda} \vee n^{2/3}\sqrt{d}\lambda^{-5/6} \right)$$
$$\mathbb{E}[\|\hat{M} - M\|_F^2] \leq C \log^C(\tfrac{nd}{\lambda}) \left(n^{1/3} d \lambda^{-2/3} \right) .$$

In the easier bi-isotonic model where M has increasing rows up to a permutation σ^*, if $n = d$, the upper bound on the estimation risk is much better:

$$\mathbb{E}[\|\hat{M} - M\|_F^2] \quad \leq C \log^C(\tfrac{nd}{\lambda}) \left(n^{7/6}\lambda^{-5/6}\right) \ .$$

The upper-bound of order $n^{7/6}\lambda^{-5/6}$ on the estimation risk, while not minimax optimal, aligns with the state of the art estimation rate in polynomial time in the isotonic and bi-isotonic models when x^* is known – see Pilliat et al. (2024) for general λ and Liu and Moitra (2020) in the special case $\lambda = 1$. Hence, the method for estimating M based on **ISV** achieves the best known performance in the harder case where x^* is unknown.

The proof of Theorem 3 consists in bounding the risk $\mathbb{E}[\|M_{\hat{\pi}^{-1}} - M_{\pi^{*-1}}\|_F^2]$ (resp. $\mathbb{E}[\|\hat{M} - M\|_F^2]$) on tasks k for which $x_k^* = 0$ and for which $x_k^* \neq 0$ separately. In the first case, we bound the risk by a quantity of order d/λ, using Theorem 1. In the second case, we use the results of Pilliat et al. (2024) to get an upper bound of order $n^{2/3}\sqrt{d}\lambda^{-5/6}$ (resp. $n^{1/3}d\lambda^{-2/3}$ and $n^{7/6}\lambda^{-5/6}$ when $n = d$ in the bi-isotonic model). In the following proposition, we restate the lower bound of Theorem 2.1 and of Proposition 2.3 of Pilliat et al. (2024), in our harder crowdsourcing model where x^* is unknown.

Proposition 4 *[Restatement of the Lower Bounds in Pilliat et al. (2024)] Assume that $1 < \frac{1}{\lambda} \leq n$ and that E is a matrix with i.i.d. standard Gaussian entries. Then for any estimator $\hat{\pi}$ and $\hat{M}$, the maximum ranking risks and estimation risks (6) satisfy the following lower bounds, for some numerical constant c:*

$$\sup_{M,x^*} \mathbb{E}[\|M_{\hat{\pi}^{-1}} - M_{\pi^{*-1}}\|_F^2] \geq c\left(n^{2/3}\sqrt{d}\lambda^{-5/6}\right) \quad \textit{and} \quad \sup_{M,x^*} \mathbb{E}[\|\hat{M} - M\|_F^2] \geq c\left(n^{1/3}d\lambda^{-2/3}\right) \ .$$

Hence, in the isotonic model, the polynomial-time estimator $\hat{M}$ is minimax optimal for all values of λ, n, d in the regime $\frac{1}{\lambda} \leq n \leq d$, up to polylogs. The estimator $\hat{\pi}$ is also optimal, in the case where $d/\lambda \lesssim n^{2/3}\sqrt{d}\lambda^{-5/6}$, that is when $n \gtrsim d^{3/4}\lambda^{-1/4}$. Hence, when $n \gtrsim d^{3/4}\lambda^{-1/4}$, ranking the workers or estimating their abilities in the isotonic model is not harder when labels are unknown than when the labels are known.

When $n \lesssim d^{3/4}\lambda^{-1/4}$, the minimax risk for ranking is not clear from our results. For example, if $n = 2$ and $\lambda = 1$, the lower bound on the ranking risk is of order $\sqrt{d}$ while the upper-bound is of order d, which is achieved by a trivial estimator. A lower bound of order d can be established in the model with Bernoulli noise. Indeed, if $M_{1k} = 1$ and $M_{2k} = 0$ for all k and if the labels are randomly chosen in $\{-1, 1\}$, the matrix Y is made of i.i.d. Bernoulli random variables of parameter $1/2$, and it is impossible to decipher which one of the two workers is the best with probability larger than $1/2$. This lower bound of order d is however only valid in the model with Bernoulli noise and Bernoulli observation scheme, and we conjecture that it is possible to improve this order of magnitude in the model with Gaussian noise or with Bernoulli noise with Poisson observation scheme, using signal detection techniques.

In the square regime with full observations, that is when $n = d$ and $\lambda = 1$, both $\hat{\pi}$ and $\hat{M}$ are optimal and achieve a risk of order $n^{7/6}$ and $n^{4/3}$, respectively. In particular, $\hat{\pi}$ achieves the state of the art minimax ranking rate in the easier bi-isotonic model analyzed by Liu and Moitra (2020). To the best of our knowledge, it is not known if it is possible to achieve a better rate than $n^{7/6}$ in polynomial time for the estimation of π^*, and a computational-statistical gap remains between n and $n^{7/6}$ in this simpler model.

4. Iterative Spectral Voting (ISV) for Recovering the Labels

Assume that $T \geq 1$ is an even integer. Our procedure consists in $T/2$ steps that each uses two subsamples of the data – that are independent conditionally on $(B^{(\tau)})_{\tau \geq 1}$. The whole procedure is also summarized in Algorithm 1. We first start by describing the subsampling method before detailing each step of spectral voting.

4.1. Subsampling

We subsample the matrix of observations Y given by (3) into T subsamples $Y^{(1)}, \dots, Y^{(T)}$ by assigning each coefficient (i,k) to one of the subsamples uniformly at random. More formally, let U be a matrix in $\mathbb{R}^{n \times d}$ with i.i.d. coefficients following a uniform distribution in $[0,1]$. We let, for any $\tau \in \{1, \dots, T\}$,

$$Y_{ik}^{(\tau)} = \mathbf{1}\{U_{ik} \in [\tfrac{\tau-1}{T}, \tfrac{\tau}{T})\} Y_{ik} \quad . \tag{8}$$

We call $B_{ik}^{(\tau)} = B_{ik}\mathbf{1}\{U_{ik} \in [\tfrac{\tau-1}{T}, \tfrac{\tau}{T}]\}$ the sub-sampling matrix. From (8) and the definition of the model (3), we have $Y_{ik}^{(\tau)} = B^{(\tau)} \odot (M \operatorname{diag}(x^*) + E)$, where $\odot$ denotes the Hadamard product, that is $(A \odot A')_{ik} = A_{ik}A'_{ik}$. The subsamples $Y^{(\tau)}$ are not independent, but they are independent conditionally on U (or equivalently to the subsampling matrices $B^{(\tau)}$). The entries of $B^{(\tau)}$ are i.i.d. Bernoulli random variables with parameter $\lambda_0 := \lambda/T$.

4.2. Spectral Voting

Let $Y^{(1)}, \dots, Y^{(T)}$ be the subsamples obtained with the subsampling procedure described in the previous section. At each step $t = 1, \dots, T/2$, we compute an estimator $\hat{x}^{(t)} \in \{-1, 0, 1\}^d$ of the labels x^*. Initially, we define $\hat{x}_k^{(0)} = 0$ for all $k = 1, \dots, d$. At each subsequent step $t \geq 1$, we use the two subsamples $Y^{(2t-1)}$ and $Y^{(2t)}$ of Y, and we perform the following steps to compute $\hat{x}^{(t)}$:

1. Principal Component Analysis. We compute the left singular vector $\hat{v}$ that corresponds to the largest singular value of $Y^{(2t-1)} \operatorname{diag}(\hat{x}_k^{(t-1)} = 0)$. Then, we take the absolute value of the entries of $\hat{v}$, and apply a threshold at $\sqrt{\lambda_0}$:

$$\hat{v} \in \operatorname*{arg\,max}_{\|v\|=1} \|v^T Y^{(2t-1)} \operatorname{diag}(\hat{x}_k^{(t-1)} = 0)\|_2^2 \quad \text{and} \quad \tilde{v}_i = |\hat{v}_i| \wedge \sqrt{\lambda_0} \, . \tag{9}$$

If $\tilde{v}_i > \tilde{v}_j$, it indicates that we have greater confidence in the prediction made by worker i compared to worker j. The key motivation behind taking a leading singular vector $\hat{v}$ stems from the observation that $(M \operatorname{diag}(x^*))(M \operatorname{diag}(x^*))^T = MM^T$ does not depend on the vector of unknown labels x^*. Let $M(t) = M \operatorname{diag}(\hat{x}_k^{(t-1)} = 0)$ be the matrix restricted to tasks with uncertain labels. The main idea in the proof is to show that, with high probability, $\hat{v}$ captures a significant part of the squared Frobenius norm of M, that is $\|\hat{v}^T M(t)\|_F^2 \gtrsim \|M(t)\|_F^2$ up to polylogs. One of the main ingredients to show this is that $\|M(t)\|_{\mathrm{op}}^2 \gtrsim \|M(t)\|_F^2$ when M_{π^*} is isotonic – we refer the reader to Theorem 8 for a proof of this fact.

2. Weighted Voting. The entries of $\tilde{v}$ are the weights assigned to each worker in this step. We compute the weighted vote $\tilde{v}^T Y_{\cdot,k}^{(2t)}$ for each task k such that $\hat{x}_k^{(t-1)} = 0$. We obtain a vector $\hat{w}$ that

has non-zero entries only on tasks k such that $\hat{x}_k^{(t-1)} = 0$:

$$\hat{w} = \tilde{v}^T Y^{(2t)} \operatorname{diag}(\hat{x}_k^{(t-1)} = 0) \,. \tag{10}$$

$\hat{v}^T Y^{(2t-1)} \operatorname{diag}(\hat{x}_k^{(t-1)} = 0)$ corresponds to the right singular vector of $Y^{(2t-1)} \operatorname{diag}(\hat{x}_k^{(t-1)} = 0)$. The idea in (10) is to use $\tilde{v}$ instead of $\hat{v}$ and the second sample $Y^{(2t)}$ to obtain better concentration bounds in the proof. Finally, we define the new label $x_k^{(t)}$ as the former label if task k has already been assigned to a label, and as the weighted vote if it is sufficiently significant in absolute value:

$$\hat{x}_k^{(t)} = \begin{cases} \hat{x}_k^{(t-1)} \text{ if } \hat{x}_k^{(t-1)} \neq 0 \\ \operatorname{sign}(\hat{w}_k) \text{ if } \hat{x}_k^{(t-1)} = 0 \text{ and } |\hat{w}_k| > \sqrt{2 \sum_{i=1}^n \tilde{v}_i^2 B_{ik}^{(2t)} \log(dT/\delta)} \\ 0 \text{ otherwise.} \end{cases} \tag{11}$$

In the proof, the key idea is to use the condition $|\hat{w}_k| \lesssim \sqrt{2 \sum_{i=1}^n \tilde{v}_i^2 B_{ik}^{(2t)}}$ on tasks k such that $\hat{x}_k^{(t)} = 0$, to show that $\|\tilde{v}^T M(t+1)\|_F^2 \lesssim \frac{d}{\lambda}$, where $M(t+1) = M \operatorname{diag}(\hat{x}_k^{(t)} = 0)$. From the two inequalities $\|\tilde{v}^T M(t)\|_F^2 \gtrsim \|M(t)\|_F^2$ and $\|\tilde{v}^T M(t+1)\|_F^2 \lesssim \frac{d}{\lambda}$, combined with $\|M(t)\|_F^2 - \|M(t+1)\|_F^2 \geq \|\tilde{v}^T M(t)\|^2 - \|\tilde{v}^T M(t+1)\|^2$, which follows from the Pythagorean theorem, we can establish that $\|M(t)\|_F^2$ decays exponentially as long as $\|M(t)\|_F^2 \gtrsim \frac{d}{\lambda}$. A highly technical issue that arises in our Bernoulli observation scheme is that $\tilde{v}$ is not independent of $B^{(2t)}$. To understand how we solve this problem, we refer the reader to the event ξ_Γ^- introduced in (15).

Algorithm 1 Iterative Spectral Voting (**ISV**)

Data: Y as in Equation (3), an even integer $T \geq 1$ and $\delta \in (0,1)$.
Result: An estimator $\hat{x}^{(T/2)}$ of x^*
Subsample Y into $Y^{(1)}, \ldots, Y^{(T)}$ as in Equation (8), and let $B^{(\tau)}$ be such that $Y_{ik}^{(\tau)} = B^{(\tau)} \odot (M \operatorname{diag}(x^*) + E)$.
$\hat{x}_k^{(0)} \leftarrow 0$ for all $k \in [d]$
for $t = 1, \ldots, T/2$: **do**
  $\hat{v} \leftarrow \arg\max_{\|v\|=1} \|v^T Y^{(2t-1)} \operatorname{diag}(\hat{x}_k^{(t-1)} = 0)\|_2^2$ as in (9)
  $\tilde{v}_i \leftarrow |\hat{v}_i| \wedge \sqrt{\lambda_0}$
  $\hat{w} \leftarrow \tilde{v}^T Y^{(2t)} \operatorname{diag}(\hat{x}_k^{(t-1)} = 0)$
  $\hat{x}_k^{(t)} \leftarrow \hat{x}_k^{(t-1)} + \operatorname{sign}(\hat{w}_k) \mathbf{1}\{|\hat{w}_k| > \sqrt{2 \sum_{i=1}^n \tilde{v}_i^2 B_{ik}^{(2t)} \log(dT/\delta)}\}$
end

4.3. Related Algorithms in the Literature

The polynomial time procedure closest to our Iterative Spectral Voting is the OBI-WAN method introduced in Shah et al. (2020), specifically within the bi-isotonic model in Bernoulli noise – see (1). OBI-WAN first generates two samples $Y^{(1)}$ and $Y^{(2)}$ from the observations by splitting the set of tasks $[d]$ uniformly into two groups. The OBI step then computes the left singular vector $\hat{v}$, corresponding to the largest singular value of $Y^{(1)}$, as described in (9). Finally, the WAN step performs a majority vote among the k_{WAN} best workers according to an order defined by $\hat{v}$. The

number k_{WAN} is chosen to maximize the number of tasks for which the majority vote is biased toward one of the labels in $\{-1, 1\}$. After these two steps, the OBI-WAN procedure assigns a label, either -1 or 1, to all tasks. The main differences between OBI-WAN and **ISV** are listed as follows:

1. Only significant labels are assigned a value in $\{-1, 1\}$ at the end of each step in **ISV**. This allows us in particular to distinguish the tasks k such that $\hat{x}_k \neq 0$ for which we can guarantee a perfect recovery of the labels with high probability.
2. **ISV** iterates over the matrix restricted to unknown tasks k, corresponding to $\hat{x}_k = 0$. This iteration is the main ingredient to achieve the minimax rate in polynomial time.
3. **ISV** performs a weighted voting instead of a majority voting over a set of workers. While it may seem intuitive to first identify the top workers before computing a majority vote, as done in the WAN step, this is not necessary as the vector $\tilde{v}$ can directly be used as weights to compute a weighted vote and accurately estimate the labels.

As mentioned in Shah et al. (2020), the spectral step is not new and has already been used in several other related parametric crowdsourcing model Khetan and Oh (2016); Ghosh et al. (2011). A similar approach, using first a right singular vector instead of a left singular vector has been used in other related non-parametric ranking problems Pilliat et al. (2023, 2024).

5. Numerical Study

In this final section, we present simulations using synthetic data to compare **ISV** to a naive majority voting method and to the Obi-Wan method proposed in Shah et al. (2020). We compare these three procedures using the normalized squared norm loss $\|M \operatorname{diag}(\hat{x} - x^*)\|_F^2 / \|M\|_F^2$ and the normalized Hamming loss $\frac{1}{d}\sum_{k=1}^{d} \mathbf{1}\{\hat{x}_k \neq x_k^*\}$. In practice, subsampling as in (8) turns out to result in significant underperformance. For that reason, we implemented a modified version of **ISV** in our simulations where the same sample Y is used at all steps instead of $Y^{(\tau)}$. Moreover, at the end of **ISV**, we choose to guess $\hat{x}_k^{(T/2)} = \operatorname{sign}(\hat{w}_k)$ for all k such that $\hat{x}_k^{(t-1)} = 0$. This prevents excessive penalties in the loss compared to leaving $\hat{x}_k^{(t-1)} = 0$. For $\alpha \in \mathbb{R}$, we define bi-isotonic matrices M with two values as

$$M_{ik} = \begin{cases} 0 \text{ if } \frac{k}{d} < g_\alpha(1 - \frac{i}{n}) \\ h \text{ otherwise} \end{cases},$$

where $g_\alpha(x) = \frac{1}{\alpha^2(x - \phi(\alpha))} + \phi(\alpha)$ and $\phi(\alpha) = \frac{\alpha - \sqrt{\alpha^2+4}}{2\alpha}$.

We choose this form for g_α for two key reasons. First, $g(0) = 1$, $g(1) = 0$, and the graph of the function g_α has the nice property of being symmetric according to the line $y = -x$. Second, varying α allows for a range of behaviors. When $\alpha \leq 0$, the bi-isotonic matrix M cannot be well approximated by a rank 1 matrix, which leads to underperformance of spectral methods that rely only on one singular vector. Conversely, when $\alpha \gg 1$, the leading singular vector of M captures a significant portion of $\|M\|_{\text{op}}^2$ making simple spectral methods good candidates for this scenario. We refer the reader to Figure 1 for an illustration of the chosen matrices M in our numerical study, when $\alpha \in \{-3, 0, 3\}$.

In Figure 2, we see that the modified version of **ISV** performs better than both the majority voting and the Obi-Wan methods when $\alpha \leq 0$ and when h becomes larger. When $\alpha \gg 1$, the matrix can be well approximated by a rank 1 matrix, and majority voting and Obi-Wan achieve comparable and better results than **ISV** for small h. A reproducible Julia Pluto notebook is available on the

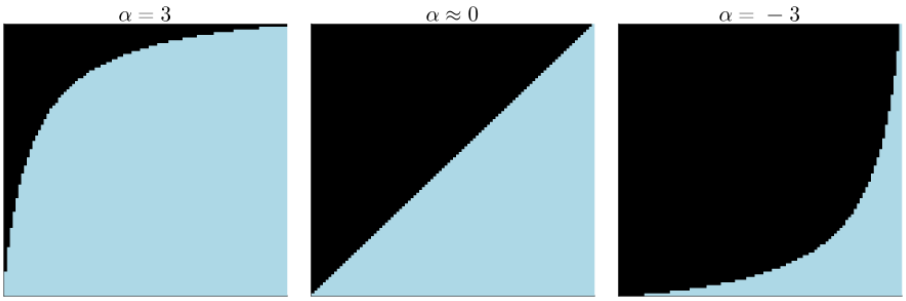

Figure 1: Matrices M defined for $\alpha \in \{3, 0, -3\}$. The black (resp. blue) area represents coefficients (i, k) for which $M_{ik} = 0$ (resp. $h \in [0, 1]$).

author's GitHub page. The Python code for Obi-Wan from Shah et al. (2020) was converted to Julia to enhance execution speed.

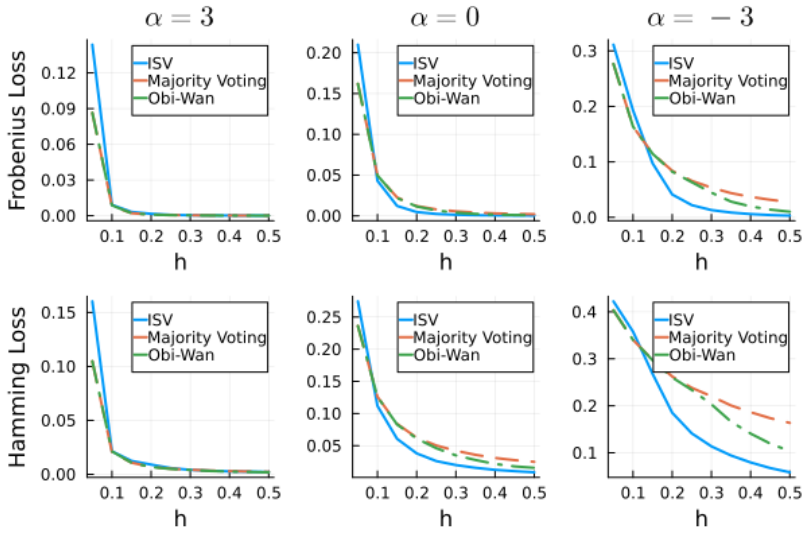

Figure 2: The Frobenius and Hamming Losses of the three methods as a function of $h \in \{0.05i, i \in [1 : 10]\}$. We chose $n = d = 1000$, $\lambda = 1$ and $T = 20$ for **ISV** without subsampling. Each line represents the mean of a Monte-Carlo simulation with 100 trials. For each trial, we uniformly generate labels $x_k^\star$ in $\{-1, 1\}$ and i.i.d. standard Gaussian noise.

Acknowledgments

The author is thankful to ENS Lyon for partially funding this work, and to the anonymous reviewers for their comments which helped improve this work.

References

T Tony Cai, Rungang Han, and Anru R Zhang. On the non-asymptotic concentration of heteroskedastic wishart-type matrix. Electronic Journal of Probability, 27:1–40, 2022.

Nilesh Dalvi, Anirban Dasgupta, Ravi Kumar, and Vibhor Rastogi. Aggregating crowdsourced binary ratings. In Proceedings of the 22nd international conference on World Wide Web, pages 285–294, 2013.

Alexander Philip Dawid and Allan M Skene. Maximum likelihood estimation of observer error-rates using the em algorithm. Journal of the Royal Statistical Society: Series C (Applied Statistics), 28 (1):20–28, 1979.

Arpita Ghosh, Satyen Kale, and Preston McAfee. Who moderates the moderators? crowdsourcing abuse detection in user-generated content. In Proceedings of the 12th ACM conference on Electronic commerce, pages 167–176, 2011.

David Karger, Sewoong Oh, and Devavrat Shah. Iterative learning for reliable crowdsourcing systems. Advances in neural information processing systems, 24, 2011.

Ashish Khetan and Sewoong Oh. Achieving budget-optimality with adaptive schemes in crowdsourcing. Advances in Neural Information Processing Systems, 29, 2016.

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

Cheng Mao, Ashwin Pananjady, and Martin 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.

Cheng Mao, Ashwin Pananjady, and Martin J Wainwright. Towards optimal estimation of bivariate isotonic matrices with unknown permutations. The Annals of Statistics, 48(6):3183–3205, 2020.

Pascal Massart. Concentration inequalities and model selection, volume 6. Springer, 2007.

Ashwin Pananjady and Richard J Samworth. Isotonic regression with unknown permutations: Statistics, computation and adaptation. The Annals of Statistics, 50(1):324–350, 2022.

Emmanuel Pilliat, Alexandra Carpentier, and Nicolas Verzelen. Optimal permutation estimation in crowdsourcing problems. The Annals of Statistics, 51(3):935–961, 2023.

Emmanuel Pilliat, Alexandra Carpentier, and Nicolas Verzelen. Optimal rates for ranking a permuted isotonic matrix in polynomial time. In Proceedings of the 2024 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA), pages 3236–3273. SIAM, 2024.

Yury Polyanskiy and Yihong Wu. Lecture notes on information theory. Lecture Notes for ECE563 (UIUC) and, 6(2012-2016):7, 2014.

Nihar B Shah, Sivaraman Balakrishnan, Adityanand Guntuboyina, and Martin J Wainwright. Stochastically transitive models for pairwise comparisons: Statistical and computational issues. IEEE Transactions on Information Theory, 63(2):934–959, 2016.

Nihar B Shah, Sivaraman Balakrishnan, and Martin J Wainwright. A permutation-based model for crowd labeling: Optimal estimation and robustness. IEEE Transactions on Information Theory, 67(6):4162–4184, 2020.

Martin J Wainwright. High-dimensional statistics: A non-asymptotic viewpoint, volume 48. Cambridge University Press, 2019.

Peter Welinder, Steve Branson, Pietro Perona, and Serge Belongie. The multidimensional wisdom of crowds. Advances in neural information processing systems, 23, 2010.

Jacob Whitehill, Ting-fan Wu, Jacob Bergsma, Javier Movellan, and Paul Ruvolo. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. Advances in neural information processing systems, 22, 2009.

Appendix A. Proof of the Upper Bounds (Theorem 1 and Theorem 3)

To prove Theorem 1, we first introduce some high probability events. Next, in Theorem 7, we establish that if all these events occur, then $\hat{x}_k \in \{0, x_k^*\}$ for all $k \in [d]$ and the loss $\|M \operatorname{diag}(\hat{x}^{(t)} - x^*)\|_F^2$ decreases exponentially until reaching the magnitude $\frac{d}{\lambda_0}$ up to a polylog. We provide a proof of the upper bounds for ranking and estimating M, Theorem 3 in Appendix A.3. Next sections, Appendix B and Appendix C, are dedicated to the proof of Theorem 7 and to the concentration of the events defined in this section. For clarity, we assume without loss of generality that $\pi^* = id$. This convention allows us to write $[l, r)$ for a contiguous set of rows in $1, \dots, n$ instead of $\pi^{*-1}([\pi^*(l), \pi^*(r)))$.

A.1. High probability events

Let us first define some events that will be proven to hold with high probability. We first start with events on the matrices $(B^{(\tau)} \odot E)_{\tau=1,\dots,T}$, which are independent conditionally on the matrices $(B^{(\tau)})_{\tau=1,\dots,T}$. Then, we introduce some events on the matrices $(B^{(\tau)})_{\tau=1,\dots,T}$.

High probability events on $B \odot E$

For a given step t in **ISV** and $C_{\text{op}} > 0$, we say that the event $\xi_{\text{op}}(C_{\text{op}}, t)$ holds if

$$\|(B^{(2t-1)} \odot E)(B^{(2t-1)} \odot E)^T\|_{\text{op}} \le C_{\text{op}} \log(\tfrac{ndT}{\delta}) \left(1 + \max_{i\in[n]} \sum_{k=1}^{d} B_{ik}^{(2t-1)} + \max_{k\in[d]} \sum_{i=1}^{n} B_{ik}^{(2t-1)}\right) . \tag{12}$$

In particular, under $\xi_{\text{op}}(t)$, the operator norm $\|(B^{(2t-1)} \odot E)(B^{(2t-1)} \odot E)^T\|_{\text{op}}$ is small if the maximum number of observations per row and columns are small. In our proof, we use the event $\xi_{\text{op}}(t)$ to analyze the concentration of $\hat{v}$ in **ISV**. For a given step t, we denote $\tilde{v}$ as the vector computed at step t in **ISV**. We also introduce the event $\xi_{\text{col}}(t)$, which holds when for all $k \in [d]$,

$$|\tilde{v}^T (B^{(2t)} \odot E)_{\cdot k}| \le \sqrt{2 \sum_{i=1}^{n} \tilde{v}_i^2 B_{ik}^{(2t)} \log(dT/\delta)} . \tag{13}$$

These two events hold with probability at least $1 - \delta$, as the following lemma shows:

Lemma 5 *There exists a numerical constant C_{op} such that the events $\xi_{\text{col}}(t)$ and $\xi_{\text{op}}(C_{\text{op}}, t)$ simultaneously hold for all $t = 1, \dots, T/2$, with probability larger than $1 - 2\delta$.*

In the proof of Theorem 5, for $\xi_{\text{op}}(C_{\text{op}}, t)$, we condition on $B^{(2t-1)}$ which is independent of E. For $\xi_{\text{col}}(t)$, we condition on both $B^{(2t)}$ and the vector $\hat{v}$. A crucial observation is that $\tilde{v}$ is independent of $B^{(2t)} \odot E$. We postpone the proof of Theorem 5 to Appendix C.1

High probability events on B

Let τ be any integer in $\{1, \dots, T\}$, and consider the event $\xi_{\text{RC}}^-(\tau)$ which holds when the two following inequalities are satisfied

$$\begin{cases} \sum_{k=1}^{d} B_{ik}^{(\tau)} \le \psi(1 + \lambda_0 d), \text{ for all } i \in [n] \\ \sum_{i=1}^{n} B_{ik}^{(\tau)} \le \psi(1 + \lambda_0 n), \text{ for all } k \in [d] \end{cases} \tag{14}$$

In other words, $\xi^-_{\mathrm{RC}}(\tau)$ holds true when there are at most a number of order $1 + (\lambda_0 d)$ (resp. $1 + (\lambda_0 n)$) observed coordinates in each row $i = 1, \dots, n$ (resp. in each columns $k = 1, \dots, d$). The order of magnitudes $\lambda_0 d$ and $\lambda_0 n$ respectively correspond to the expectation of $\sum_{k=1}^{d} B_{ik}^{(\tau)}$ and of $\sum_{i=1}^{n} B_{ik}^{(\tau)}$.

Let us now define the event $\xi^-_{\Gamma}(\tau)$, holding true if for all subset $\Gamma \subset [d]$ of size $|\Gamma| = \left\lceil 2\log(\frac{n}{\lambda_0})/\lambda_0 \right\rceil$ and all vector $v \in \mathbb{R}^n$ that satisfies $\|v\|_2 = 1$ and $\|v\|_\infty \leq \sqrt{\lambda_0}$,

$$\min_{k \in \Gamma} \sum_{i=1}^{n} v_i^2 B_{ik}^{(\tau)} \leq \psi^2 \lambda_0 \ , \tag{15}$$

The event $\xi^-_{\Gamma}(\tau)$ can be rephrased as follows. For any unit v satisfying $\|v\|_\infty \leq \sqrt{\lambda_0}$, we can remove the subset of columns Γ corresponding to the $\left\lceil 2\log(\frac{n}{\lambda_0})/\lambda_0 \right\rceil$ largest value of $\sum_{i=1}^{n} v_i^2 B_{ik}^{(\tau)}$ to ensure that for each remaining column k, $\sum_{i=1}^{n} v_i^2 B_{ik}^{(\tau)} \leq \min_{k \in \Gamma} \sum_{i=1}^{n} v_i^2 B_{ik}^{(\tau)} \leq \psi^2 \lambda_0$. Specifically, under $\xi^-_{\Gamma}(\tau)$, we can apply the above upperbound with the vector $v := \tilde{v}$ defined in **ISV**.

Next, we define the event $\xi^+_C(\tau)$ that holds true if

$$\sum_{i=l}^{r-1} B_{ik}^{(\tau)} \geq \frac{\lambda_0}{2}(r-l) \ , \tag{16}$$

for all $k \in [d]$ and all $l \leq r-1$ in $[n]$ such that $\lambda_0(r-l) \geq \psi$. Recall that we assume that $\pi^* = id$ so that $[l, r)$ corresponds to a contiguous set of workers i. Under $\xi^+_C(\tau)$, the number of observation given by the subsampling matrix $B^{(\tau)}$ in $[l, r) \times \{k\}$ is at least equal to $\frac{\lambda_0}{2}(r-l)$.

Finally, we define the event $\xi^+_{\mathrm{op}}(C'_{\mathrm{op}}, \tau)$ by the inequality

$$\|(B^{(\tau)} \odot M - \lambda_0 M)(B^{(\tau)} \odot M - \lambda_0 M)^T\|_{\mathrm{op}} \leq C'_{\mathrm{op}} \log^2(\tfrac{ndT}{\delta})(\lambda_0 d + 1) \ . \tag{17}$$

Under $\xi^+_{\mathrm{op}}(C'_{\mathrm{op}}, \tau)$, the square distance of $B^{(\tau)} \odot M$ to its expected value $\lambda_0 M$ in operator norm is smaller than a quantity of order $\lambda_0 d + 1$. The following lemma, whose proof is postponed to Appendix C.2 establishes that all the events of this section hold with high probability.

Lemma 6 *There exists a numerical constant C'_{op} such that the four events $\xi^-_{\mathrm{RC}}(\tau)$, $\xi^-_{\Gamma}(\tau)$, $\xi^+_C(\tau)$, $\xi^+_{\mathrm{op}}(\tau)$ – which respectively correspond to Equations (14) to (17) – simultaneously hold for all $\tau = 1, \dots, T$, with probability at least $1 - 4\delta$.*

A.2. Exponential decay of the loss and conclusion of the proof

Assume that all the events from the previous section $\xi_{\mathrm{op}}(C_{\mathrm{op}}, t)$, $\xi_{\mathrm{col}}(t)$, $\xi^-_{\mathrm{RC}}(\tau)$, $\xi^-_{\Gamma}(\tau)$, $\xi^+_C(\tau)$, $\xi^+_{\mathrm{op}}(C'_{\mathrm{op}}, \tau)$ simultaneously hold true for all $t = 1, \dots T/2$ and $\tau = 1, \dots, T$. From Theorem 5 and Theorem 6, this happens with probability at least $1 - 5\delta$. Under these events, the following proposition states two properties. The first one is that $\hat{x}_k^{(t)}$ is either 0 or equal to the true label x_k^*, which directly implies the first part of Theorem 1. The second one is that the loss at step t is either of order d/λ_0 or is exponentially smaller than nd.

Proposition 7 *Under the events of the previous section Appendix A.1, it holds for all $t = 1, \dots, T/2$ that*

$$\hat{x}_k^{(t)} \in \{0, x_k^*\} \text{ for all } k \in [d] \tag{18}$$

and

$$\|M\,\mathrm{diag}(\hat{x}^{(t)} - x^*)\|_F^2 \le \left[\overline{C}\psi^5 \frac{d}{\lambda_0}\right] \vee \left[\left(1 - \frac{1}{512e^2\log^2(end)}\right)^t nd\right] \ , \tag{19}$$

where $\overline{C} = \max(C_{\mathrm{op}}, C'_{\mathrm{op}})$.

In particular, Theorem 7 implies that if $T \ge 1024e^2\log^2(end)$, then the final estimator $\hat{x}^{(T/2)}$ satisfies

$$\|M\,\mathrm{diag}(\hat{x}^{(T/2)} - x^*)\|_F^2 \le \overline{C}\psi^5 \frac{d}{\lambda_0} = \overline{C}T\psi^5 \frac{d}{\lambda} \ ,$$

which concludes the proof of Theorem 1. The proof of Theorem 7 is given in Appendix B. Before providing the proof of Theorem 7 in Appendix B, we give the proof of the upper bound of $\hat{\pi}$ and $\hat{M}$ defined in Section 3.2

A.3. Proof of Theorem 3

Let us choose $\delta = 1/(nd)$, so that from Theorem 1, $\mathbb{E}[M\,\mathrm{diag}(\hat{x} \neq x^*)] \le C_1 \log^{C_1}(\frac{nd}{\lambda})\frac{d}{\lambda}$ for some numerical constant C_1. We are also in position to apply Theorem 2.2 of Pilliat et al. (2024) to the isotonic matrix $M' = M\,\mathrm{diag}(\hat{x} = x^*)$, conditionally on the T first subsamples. We obtain that $\mathbb{E}[\|(M'_{\hat{\pi}^{-1}} - M'_{\pi^{*-1}})\|_F^2] \le C_2 \log^{C_2}(\frac{nd}{\lambda})n^{2/3}\sqrt{d}\lambda^{-5/6}$ in our regime $\frac{1}{\lambda} \le n \le d$. We obtain that

$$\mathbb{E}[\|M_{\hat{\pi}^{-1}} - M_{\pi^{*-1}}\|_F^2] \le 2C_1 \log^{C_1}(\tfrac{nd}{\lambda})\frac{d}{\lambda} + C_2 \log^{C_2}(\tfrac{nd}{\lambda})n^{2/3}\sqrt{d}\lambda^{-5/6} \ .$$

We conclude with a similar argument, based on Corollary 2.4 of Pilliat et al. (2024), that

$$\mathbb{E}[\|\hat{M} - M\|_F^2] \le C\log^C(\tfrac{nd}{\lambda})\left(\frac{d}{\lambda} \vee n^{1/3}d\lambda^{-2/3}\right) \le C\log^C(\tfrac{nd}{\lambda})n^{1/3}d\lambda^{-2/3} \ .$$

For the last upper-bound in the bi-isotonic model, we use Corollary 2.5 of Pilliat et al. (2024), and the fact that $d/\lambda = n/\lambda \le n^{7/6}\lambda^{-5/6}$.

Appendix B. Proof of Theorem 7

In this section, we assume that the events $\xi_{\mathrm{op}}(C_{\mathrm{op}}, t)$, $\xi_{\mathrm{col}}(t)$, $\xi^{-}_{\mathrm{RC}}(\tau)$, $\xi^{-}_{\Gamma}(\tau)$, $\xi^{+}_{C}(\tau)$, $\xi^{+}_{\mathrm{op}}(C'_{\mathrm{op}}, \tau)$ simultaneously hold true for all $t = 1, \dots T/2$ and $\tau = 1, \dots, T$, for some constant $C_{\mathrm{op}}, C'_{\mathrm{op}}$. We first show that $\hat{x}_k^{(t)} \in \{0, x_k^*\}$ for all $k \in [d]$ (18) before establishing the exponential decay of the loss (19).

B.1. Non-zero estimated labels correspond to the true labels

Let us show that at each step $t = 1, \dots T/2$, the estimated labels are either equal to 0 or to the true label x_k^*, that is for all $k \in [d]$,

$$\hat{x}_k^{(t)} \in \{0, x_k^*\} \ . \tag{20}$$

By definition, we have $\hat{x}_k^{(0)} = 0$ for all $k \in [d]$. Let $t \geq 1$ be a step of **ISV**. By definition, the vector $\hat{w}$ computed at step t, line Algorithm 1 has its coefficients equal to 0 on columns k such that $\hat{x}_k^{(t-1)} \in \{-1, 1\}$. These columns k correspond to the tasks that have already been labeled in a previous step. Let us assume that k rather corresponds to a task with an unknown label at the end of step $t-1$, that is $\hat{x}_k^{(t-1)} = 0$. By definition, $\hat{x}_k^{(t)} = \mathrm{sign}(\hat{w}_k)\mathbf{1}\{|\hat{w}_k| > \sqrt{2\sum_{i=1}^n \tilde{v}_i^2 B_{ik}^{(2t)} \log(dT/\delta)}\}$. Hence, if $|\hat{w}_k| \leq \sqrt{2\sum_{i=1}^n \tilde{v}_i^2 B_{ik}^{(2t)} \log(dT/\delta)}$, then $\hat{x}_k^{(t)} = 0$. On the other hand, assume that $|\hat{w}_k| > \sqrt{2\sum_{i=1}^n \tilde{v}_i^2 B_{ik}^{(2t)} \log(dT/\delta)}$. Since we are in the case where $\hat{x}_k^{(t-1)} = 0$,

$$\hat{w}_k = \left(\tilde{v}^T Y^{(2t)}\right)_k = \gamma x_k^* + \left(\tilde{v}^T (B^{(2t)} \odot E)\right)_k \ ,$$

where $\gamma = \sum_{i=1}^n \tilde{v}_i B_{ik}^{(2t)} M_{ik} \geq 0$. Using the event $\xi_{\mathrm{col}}(t)$ (13), we have that

$$|\tilde{v}^T (B^{(2t)} \odot E)_{\cdot k}| \leq \sqrt{2\sum_{i=1}^n \tilde{v}_i^2 B_{ik}^{(2t)} \log(dT/\delta)} < |\hat{w}_k| \ .$$

We deduce that γx_k^* is non-zero and has the same sign as $\hat{w}_k$, which proves in that case that $x_k^{(t)} = \mathrm{sign}(\hat{w}_k) = x_k^*$. Hence, all the events $\xi_{\mathrm{truth}}(t)$ – see (18) – simultaneously hold true for all $t = 1, \dots, T/2$.

B.2. Exponential decay of the loss until d/λ_0 (19)

As a shorthand, we write in this subsection $B' = B^{(2t-1)}$, $B'' = B^{(2t)}$ and the corresponding subsamples $Y' = Y^{(2t-1)}$ and $Y'' = Y^{(2t)}$. For any matrix $A \in \mathbb{R}^{n \times d}$, we denote

$$A(t) = A\,\mathrm{diag}(\hat{x}_k^{(t-1)} = 0) \ . \tag{21}$$

In particular, $M(t)$ is the matrix restricted to columns such that $\hat{x}_k^{(t-1)} = 0$ and $\|M(t)\|_F^2$ corresponds to the loss at the end of step $t-1$. Let us now establish by induction that for all $t = 0, \dots, T/2 - 1$,

$$\|M(t+1)\|_F^2 \leq \left\lceil \overline{C}\psi^5 \frac{d}{\lambda_0} \right\rceil \vee \left\lceil \left(1 - \frac{1}{512 e^2 \log^2(end)}\right)^t nd \right\rceil \ .$$

Since the coefficients of M are bounded by 1, $\|M \operatorname{diag}(\hat{x}^{(0)} - x^*)\|_F^2 = \|M\|_F^2 \leq nd$ and (19) is satisfied for $t = 0$. In what follows, we fix $t \geq 0$, and we assume that

$$\|M(t+1)\|_F^2 > \overline{C}\psi^5 \frac{d}{\lambda_0} \quad , \tag{22}$$

where $\overline{C} = \max(C_{\mathrm{op}}, C'_{\mathrm{op}})$. Since $t \mapsto \|M(t)\|_F^2$ is a non-increasing function, assumption (22) implies that $\|M(t)\|_F^2 > \overline{C}\psi^5 \frac{d}{\lambda_0}$.

Step 1: Lower bound on $\|\hat{v}^T(B' \odot M(t))\|_2^2$. Let $\hat{v}$ be the vector computed at step t in the algorithm. Let us prove that $\|\hat{v}^T(B' \odot M(t))\|_2^2 \gtrsim \|M(t)\|_F^2$, up to a polylogarithmic factor. We start with the following inequality:

$$\begin{aligned}\|\hat{v}^T(B' \odot M(t))\|_2^2 &= \|\hat{v}^T Y'(t)\|_2^2 + 2\langle \hat{v}^T Y'(t), \hat{v}^T(B' \odot E(t))\rangle + \|\hat{v}^T(B' \odot E(t))\|_2^2 \\ &\geq \frac{1}{2}\|\hat{v}^T Y'(t)\|_2^2 - \|\hat{v}^T(B' \odot E)\|_2^2 \quad ,\end{aligned}$$

where we used the inequality $2|\langle x, y\rangle| \leq \frac{1}{2}\|x\|_2^2 + 2\|y\|_2^2$. Next, we use the following Lemma to give a lower bound on $\|\hat{v}^T Y'(t)\|_2^2$. We postpone its proof to the end of this section.

Lemma 8 *Let $A \in [0,1]^{n \times d}$ be any isotonic matrix, that is $A_{i,k} \leq A_{i+1,k}$. There exist $a > 0$, a subset $I \subset [n]$ of the form $[l, r)$ and a subset $K \subset [d]$ and such that $A_{ik} \geq a$ for all $(i,k) \in I \times K$ and*

$$a^2 |I||K| \geq \frac{1}{e^2 \log^2(end)}(\|A\|_F^2 - 1) \quad .$$

According to Theorem 8, there exist $a \in (0,1]$, I and K such that $M_{ik}(t) \geq a$ for all (i,k) in $I \times K$ and $a^2|I||K| \geq \frac{1}{e^2 \log^2(end)}(\|M(t)\|_F^2 - 1)$. Let u be the unit vector equal to $\frac{1}{\sqrt{|I|}}\mathbf{1}_I$. We have that

$$\begin{aligned}\|\hat{v}^T Y'(t)\|_2^2 &\geq \|u^T Y'(t)\|_2^2 \\ &\geq \frac{1}{2}\|u^T(B' \odot M(t))\|_2^2 - \|u^T(B' \odot E)\|_2^2 \\ &\geq \frac{1}{2}a^2 \sum_{k \in K} \frac{1}{|I|}\left(\sum_{i \in I} B'_{ik}\right)^2 - \|u^T(B' \odot E)\|_2^2 \quad .\end{aligned}$$

We used the definition of $\hat{v}$ in the first inequality, $2|\langle x, y\rangle| \leq \frac{1}{2}\|x\|_2^2 + 2\|y\|_2^2$ in the second inequality and the properties of a, I and K in the third inequality. Combining the two above inequalities, we obtain that

$$\|\hat{v}^T(B' \odot M(t))\|_2^2 \geq \frac{1}{4}a^2 \sum_{k \in K} \frac{1}{|I|}\left(\sum_{i \in I} B'_{ik}\right)^2 - \frac{1}{2}\|u^T(B' \odot E)\|_2^2 - \|\hat{v}^T(B' \odot E)\|_2^2 \quad . \tag{23}$$

From the event ξ_C^+ (16) with $[l,r) = I$, it holds that $\sum_{i\in I} B'_{ik} > \lambda_0|I|/2$ if $\lambda_0|I| > \psi$. Hence,

$$\begin{aligned}\frac{1}{4}a^2 \sum_{k\in K}\frac{1}{|I|}\left(\sum_{i\in I} B'_{ik}\right)^2 &\geq \frac{1}{16}\lambda_0^2 a^2|I||K| - \frac{1}{16}\psi\lambda_0|K| \\ &\geq \frac{1}{16e^2\log^2(end)}\lambda_0^2(\|M(t)\|_F^2 - 1) - \frac{1}{16}\psi\lambda_0|K| \\ &\geq \frac{1}{16e^2\log^2(end)}\lambda_0^2\|M(t)\|_F^2 - \frac{1}{8}\psi\lambda_0 d \ .\end{aligned}$$

In the first inequality, we used that $a \leq 1$ and that the right hand side is non positive if $\lambda_0|I| \leq \psi$. In the second and last inequalities, we used the definition of I, K – see Theorem 8 – and that $|K| \leq d$. From the events $\xi_{\mathrm{op}}(t)$ and $\xi_{\mathrm{RC}}^+(t)$ – see (12) and (14) – we have

$$\begin{aligned}\frac{1}{2}\|u^T(B'\odot E)\|_2^2 + \|\hat{v}^T(B'\odot E)\|_2^2 &\leq 2C_{\mathrm{op}}\log(\tfrac{ndT}{\delta})\left(1 + \max_{i\in[n]}\sum_{k=1}^d B'_{ik} + \max_{k\in[d]}\sum_{i=1}^n B'_{ik}\right) \\ &\leq 2C_{\mathrm{op}}\log(\tfrac{ndT}{\delta})\left(1 + \psi(\lambda_0 n + 1) + \psi(\lambda_0 d + 1)\right) \\ &\leq \frac{1}{2}C_{\mathrm{op}}\psi^2(\lambda_0 d + 1) \ .\end{aligned}$$

In the last inequality, we used that ψ is a large quantity of order $\log(ndT/\delta)$, see Section 3.1 for its definition. Plugging the two above inequalities in (23), we obtain

$$\begin{aligned}\|\hat{v}^T(B'\odot M(t))\|_2^2 &\geq \frac{1}{16e^2\log^2(end)}\lambda_0^2\|M(t)\|_F^2 - \left(\frac{1}{2}C_{\mathrm{op}} + \frac{1}{8}\right)\psi^2(\lambda_0 d + 1) \\ &\geq \frac{1}{32e^2\log^2(end)}\lambda_0^2\|M(t)\|_F^2 \ .\end{aligned}$$

where we used in the last inequality the assumptions $\lambda_0(n\wedge d) \geq 1$ and $\lambda_0^2\|M(t)\|_F^2 \geq C_{\mathrm{op}}\psi^5(\lambda_0 d)$ and the definition of ψ. To conclude this step, we have proven that under assumption (22),

$$\|\hat{v}^T(B'\odot M(t))\|_2^2 \geq \frac{1}{32e^2\log^2(end)}\lambda_0^2\|M(t)\|_F^2 \ . \tag{24}$$

Step 2: Lower bound on $\|\hat{v}^T M(t)\|_2^2$. From the event $\xi_{\mathrm{op}}^+(t)$ (17), we have that

$$\|\hat{v}^T(B'\odot M(t) - \lambda_0 M(t))\|_2^2 \leq \|\hat{v}^T(B'\odot M - \lambda_0 M)\|_2^2 \leq C'_{\mathrm{op}}\log^2(\tfrac{ndT}{\delta})\left[\lambda_0 d + 1\right] \ .$$

Hence, using the inequality $\|x-y\|_2^2 \geq \frac{1}{2}\|x\|_2^2 - \|y\|_2^2$ for any vectory x, y, we obtain that

$$\begin{aligned}\lambda_0^2\|\hat{v}^T M(t)\|_2^2 &\geq \frac{1}{2}\|\hat{v}^T(B'\odot M(t))\|_2^2 - C'_{\mathrm{op}}\log^2(\tfrac{ndT}{\delta})\left[\lambda_0 d + 1\right] \\ &\geq \frac{1}{64e^2\log^2(end)}\lambda_0^2\|M(t)\|_F^2 - C'_{\mathrm{op}}\log^2(\tfrac{ndT}{\delta})\left[\lambda_0 d + 1\right] \\ &\geq \frac{1}{128e^2\log^2(end)}\lambda_0^2\|M(t)\|_F^2 \ ,\end{aligned}$$

Where in the last inequality, we used again the assumption that $\lambda_0^2\|M(t)\|_F^2 \geq \overline{C}_{\mathrm{op}}\psi^5\lambda_0 d$, $\lambda_0(n \wedge d) \geq 1$ and the definition of ψ – see Section 3.1. We conclude that we can remove the term B' in (24) at the cost of a factor of order λ_0^2:

$$\lambda_0^2\|\hat{v}^T M(t)\|_2^2 \geq \frac{1}{128e^2\log^2(end)}\lambda_0^2\|M(t)\|_F^2 \ . \tag{25}$$

Step 3: Lower bound on $\|\tilde{v}^T M(t)\|_2^2$. Recall that $\tilde{v}_i = |\hat{v}_i| \wedge \sqrt{\lambda_0}$ and define $v'_i = |\hat{v}_i| - \tilde{v}_i \geq 0$. Since $\|\hat{v}\|_2^2 = 1$, v' has at most $1/\lambda_0$ nonzero coordinates. This implies by the Cauchy-Schwarz inequality that $\|v'\|_1 \leq \frac{1}{\sqrt{\lambda_0}}\|v'\|^2 \leq \frac{1}{\sqrt{\lambda_0}}$. Using the inequality $\|x+y\|_2^2 \geq \frac{1}{2}\|x\|_2^2 - \|y\|_2^2$, we have that

$$\|\tilde{v}^T M(t)\|_2^2 \geq \frac{1}{2}\||\hat{v}|^T M(t)\|_2^2 - \|(v')^T M(t)\|_2^2 \ . \tag{26}$$

Since the coordinates of M are non-negative, and from the lower bound from the previous step (25),

$$\||\hat{v}|^T M(t)\|_2^2 \geq \|\hat{v}^T M(t)\|_2^2 \geq \frac{1}{128e^2\log^2(end)}\|M(t)\|_F^2 \ . \tag{27}$$

For the second term of the right hand side in (26),

$$\|(v')^T M(t)\|_2^2 \leq \sum_{i,j\in[n]^2}\sum_{k=1}^{d} v'_i v'_j \leq d\|v'\|_1^2 \leq \frac{d}{\lambda_0} \ . \tag{28}$$

Plugging (27),(28) in (26) and using assumption (22), we obtain that

$$\|\tilde{v}^T M(t)\|_2^2 \geq \frac{1}{128e^2\log^2(end)}\|M(t)\|_F^2 - \frac{d}{\lambda_0} \geq \frac{1}{256e^2\log^2(end)}\|M(t)\|_F^2 \ . \tag{29}$$

Step 4: Upper bound on $\|\tilde{v}^T M(t+1)\|_2^2$. We recall that $\hat{w} = \tilde{v}^T Y''(t)$ and that $\hat{x}_k^{(t)} = \hat{x}_k^{(t-1)} + \mathrm{sign}(\hat{w}_k)\mathbf{1}\{|\hat{w}_k| > \sqrt{2\sum_{i=1}^n \tilde{v}_i^2 B''_{ik}\log(1/\delta)}\}$. We denote $Q := \{k : \hat{x}_k^{(t)} = 0\}$, so that the nonzero columns of $M(t+1) = M(t)\,\mathrm{diag}(\hat{x}_k^{(t)} = 0)$ correspond to the set Q. Since $\hat{w}$ is nonzero only for k such that $\hat{x}^{(t-1)} = 0$, we have for all $k \in Q$ that

$$|\hat{w}_k| = \left|\tilde{v}^T Y''_{\cdot k}\right| \leq \sqrt{2\sum_{i=1}^n \tilde{v}_i^2 B''_{ik}\log(dT/\delta)} \ .$$

From the event ξ_{col} (13), we have that for all $k \in Q$,

$$\left|\tilde{v}^T(B''\odot E)_{\cdot k}\right| \leq \sqrt{2\sum_{i=1}^n \tilde{v}_i^2 B''_{ik}\log(dT/\delta)} \ .$$

By definition of Q, and from the event $\xi_\Gamma^-(t)$ (15), there exists a subset Γ of size $|\Gamma| = \left\lceil 2\log(\frac{n}{\lambda_0})/\lambda_0 \right\rceil$ such that for all $k \in Q \setminus \Gamma$,

$$\sum_{i=1}^n \tilde{v}_i^2 B''_{ik} \leq \psi^2\lambda_0 \ .$$

Let us write M' for the matrix that has its coefficients equal to $M(t+1)$ on columns k that are not in Γ: $M'_{ik} = M(t+1)_{ik}\mathbf{1}\{k \notin \Gamma\}$. We have that

$$\begin{aligned}\|\tilde{v}^T(B'' \odot M')\|_2^2 &\leq 2\sum_{k\in Q\setminus\Gamma}|\tilde{v}^T Y''_{\cdot k}|^2 + 2\sum_{k\in Q\setminus\Gamma}|\tilde{v}^T(B''\odot E)_{\cdot k}|^2 \\ &\leq 8\sum_{k\in Q\setminus\Gamma}\sum_{i=1}^{n}\tilde{v}_i^2 B''_{ik}\log(dT/\delta) \leq \psi^3\lambda_0 d \ .\end{aligned}$$

Using the fact that $\|\tilde{v}\|_1 \leq \sqrt{n}$ and that $M_{ik} \in [0,1]$ for all i,k, we have

$$\|\tilde{v}^T M(t+1)\|_2^2 \leq \|\tilde{v}^T M'\|_2^2 + \sum_{k\in\Gamma}\|\tilde{v}\|_1^2 \leq \|\tilde{v}^T M'\|_2^2 + |\Gamma|n \ .$$

Hence, from the event $\xi^+_{\mathrm{op}}(t)$ (17), we obtain

$$\begin{aligned}\lambda_0^2\|\tilde{v}^T M(t+1)\|_2^2 &\leq \|\tilde{v}^T M'\|_2^2 + |\Gamma|n \\ &\leq 2\|\tilde{v}^T(B''\odot M')\|_2^2 + 2\|B''\odot M - \lambda_0 M\|_{\mathrm{op}}^2 + \lambda_0^2|\Gamma|n \\ &\leq 8\psi^2\lambda_0 d + C'_{\mathrm{op}}\log^2\!\left(\tfrac{ndT}{\delta}\right)(\lambda_0 d+1) + \psi\lambda_0 n \\ &\leq \psi^3 C'_{\mathrm{op}}\lambda_0(d) \ .\end{aligned}$$

For the first inequality, we used the fact that $\|x+y\|_2^2 \leq 2\|x\|_2^2 + 2\|y\|_2^2$ and that $\|B''\odot M' - \lambda_0 M'\|_{\mathrm{op}} \leq \|B''\odot M - \lambda_0 M\|_{\mathrm{op}}$. We used in the last inequality that $\lambda_0(n\wedge d) \geq 1$ and that ψ is a large polylog factor – see Section 3.1.

Conclusion of the Proof. From the Pythagorean theorem, if $M(t)$ satisfies assumption (22), then combining the last inequalities of Step 4 and Step 5 gives

$$\begin{aligned}\|M(t)\|_F^2 - \|M(t+1)\|_F^2 &\geq \|\tilde{v}^T M(t)\|_F^2 - \|\tilde{v}^T M(t+1)\|_F^2 \\ &\geq \frac{1}{256e^2\log^2(end)}\|M(t)\|_F^2 - \psi^3 C'_{\mathrm{op}}\frac{d}{\lambda_0} \\ &\geq \frac{1}{512e^2\log^2(end)}\|M(t)\|_F^2 \ .\end{aligned}$$

Hence,

$$\|M(t+1)\|_F^2 \leq \left[\left(1 - \frac{1}{512e^2\log^2(end)}\right)\|M(t)\|_F^2\right] \vee \left[\overline{C}_{\mathrm{op}}\psi^5\left(\frac{d}{\lambda_0}\right)\right] \ ,$$

which concludes the proof of Theorem 7.

Proof [Proof of Theorem 8] Let $\mathcal{A} = \{e^{-i} \ : \ i = 1,\dots,\lceil\log(nd)\rceil\}$. For any $x \in [0,1]$, there exists $a \in \mathcal{A}$ such that $x \leq e\cdot a\mathbf{1}\{x \geq a\} \vee \frac{1}{nd}$. Applying this to all the nd coefficents of A, we obtain that

$$\|A\|_F^2 \leq 1 + e^2\sum_{a\in\mathcal{A}}\sum_{i,k}a^2\mathbf{1}\{A_{ik} \geq a\} \ .$$

For each $k = 1,\dots,d$, define the set $R_k := \{i \ : \ A_{ik} \geq a\}$, and let $k_1,\dots,k_d$ be such that

$$|R_{k_1}| \geq \cdots \geq |R_{k_d}| \ .$$

Since A is assumed to be isotonic, R_{k_s} is of the form $[n - i_s + 1, n]$, where $i_1 \geq i_2 \cdots \geq i_d \geq 0$. We define

$$s^*(a) := s^*(t,a) = \operatorname*{arg\,max}_{s=1,\dots,d}(si_s) \quad \text{and} \quad I_a \times K_a = [n - i_{s^*(a)} + 1, n] \times \{k_1, \dots, k_{s^*(a)}\} \ .$$

Hence,

$$\begin{aligned}
\|A\|_F^2 &\leq 1 + e^2 \sum_{a \in \mathcal{A}} \sum_{s=1}^{d} a^2 i_s \\
&\leq 1 + e^2 \sum_{a \in \mathcal{A}} \sum_{s=1}^{d} a^2 \frac{s^*(a) i_{s^*(a)}}{s} \\
&\leq 1 + e^2 \log(ed) \sum_{a \in \mathcal{A}} a^2 s^*(a) i_{s^*(a)} \\
&= 1 + e^2 \log(ed) \sum_{a \in \mathcal{A}} a^2 |I_a||K_a| \\
&\leq 1 + e^2 \log^2(end)(a^*)^2 |I_{a^*}||K_{a^*}| \ ,
\end{aligned}$$

where $a^* = \arg\max_{a \in \mathcal{A}} a^2 |I_a||K_a|$. Since $M_{ik} \geq a$ for any $(i,k) \in I_a \times K_a$, this conclude the proof of Theorem 8 with $a = a^*, I = I_{a^*}$ and $K = K_{a^*}$.

■

Appendix C. Proofs related to the high-probability events defined in Appendix A.1

C.1. Proof of Theorem 5

Let us first fix $t \in \{1, \dots, T/2\}$. From a union bound argument, it is enough to prove that each event $\xi_{\text{col}}(t)$, $\xi_{\text{op}}(t)$ holds with probability at least $1 - 2\delta/T$.

High probability control of ξ_{col} (13). The vector $\hat{v}^{(2t-1)}$ is measurable with respect to the matrices $(B^{(\tau)} \odot E)_{\tau \leq 2t-1}$. $\hat{v}^{(2t-1)}$ is therefore independent of $B^{(2t)} \odot E$ conditionally on the sampling matrices $B^{(\tau)}$, $\tau \in [T]$. Since the coefficients of $B_{ik}^{(2t)} \odot E_{ik}$ are $B_{ik}^{(2t)}$-sub-Gaussian conditionally on $B^{(2t)}$, we are in position to apply the Hoefding inequality – see e.g. Proposition 2.5 of Wainwright (2019). It holds for any $k \in [d]$ that with probability at least $1 - 2\delta/(dT)$,

$$\begin{aligned}
\left| |\hat{v}|^T (B^{(2t)} \odot E)_{\cdot k} \right| &= \left| \sum_{i=1}^{n} |\hat{v}_i^{(t)}| B^{(2t)} E_{ik} \right| \\
&\leq \sqrt{2 \sum_{i=1}^{n} (v_i)^2 B_{ik}^{(2t)} \log(dT/\delta)} \ .
\end{aligned}$$

We conclude with a union bound over all $k = 1, \dots, d$ that $\mathbb{P}(\xi_{\text{col}}(t)) \geq 1 - 2\delta/T$.

High probability control of ξ_{op} (12). Let us write $X = (B^{(2t-1)} \odot E)(B^{(2t-1)} \odot E)^T$ and apply the high probability tail bound provided in Theorem 3.8 of Cai et al. (2022). Our problem corresponds to the following authors' notation: $p_1, p_2 = n, d$, $\sigma_{ik} = B_{ik}$ and $Z = B^{(2t-1)} \odot E$,

$\sigma_C^2 = \max_{k\in[d]} \sum_{i=1}^n B_{ik}^{(2t-1)}$, $\sigma_R^2 = \max_{i\in[d]} \sum_{k=1}^n B_{ik}^{(2t-1)}$ and $\sigma_*^2 = \max_{i,k} B_{ik} \leq 1$. Theorem 3.8 of Cai et al. (2022) with $x = \sqrt{\log(T/\delta)}$ gives that with probability at least $1 - \delta/T$,

$$\|X - \mathbb{E}[X]\|_{\mathrm{op}} \leq C' \left(\sqrt{\max_{i\in[n]} \sum_{k=1}^d B_{ik}^{(2t-1)}} + \sqrt{\max_{k\in[d]} \sum_{i=1}^n B_{ik}^{(2t-1)}} + \sqrt{\log(n \wedge d)} + \sqrt{\log(T/\delta)} \right)^2$$
$$\leq C'' \log(ndT/\delta) \left(1 + \max_{i\in[n]} \sum_{k=1}^d B_{ik}^{(2t-1)} + \max_{k\in[d]} \sum_{i=1}^n B_{ik}^{(2t-1)} \right) ,$$

for some numerical constants C', C''. Moreover, since $\mathbb{E}[E_{ik}^2] = 1$ for all i, k, a simple computation gives that $\|\mathbb{E}[X]\|_{\mathrm{op}} = \sum_{k=1}^d B_{ik}$. Hence,

$$\|X\|_{\mathrm{op}} \leq \|X - \mathbb{E}[X]\|_{\mathrm{op}} + \|\mathbb{E}[X]\|_{\mathrm{op}}$$
$$= \|X - \mathbb{E}[X]\|_{\mathrm{op}} + \max_{i\in[n]} \sum_{k=1}^d B_{ik}^{(2t-1)} .$$

We conclude from the two above inequalities that $\mathbb{P}(\xi_{\mathrm{op}}(t)) \geq 1 - \delta/T$ for some numerical constant C_{op}.

One might argue however that Theorem 3.8 of Cai et al. (2022) requires that the coefficients of E are i.i.d. standard Gaussian random variables, which is not the case here since we just assume that the coefficients of E are independent and sub-Gaussian. This issue can however simply be solved by applying the sub-Gaussian comparison lemma 3.1 of Cai et al. (2022) instead of Lemma 2.4 of Cai et al. (2022) in the proof of Theorem 3.8 of Cai et al. (2022). In other words, the sub-Gaussian comparison lemma is used to prove an upperbound of the expectation of $\|X - \mathbb{E}[X]\|_{\mathrm{op}}$ – see corollary 3.3 of Cai et al. (2022) – but it can also be used to prove the same tail bound as in Theorem 3.8 in the sub-Gaussian case.

C.2. Proof of Theorem 6

Let us fix $\tau \in \{1, \dots, T\}$. From a union bound argument, it is enough to prove that each event $\xi_{\mathrm{RC}}^-(\tau)$, $\xi_{\Gamma}^-(\tau)$, $\xi_C^+(\tau)$, $\xi_{\mathrm{op}}^+(\tau)$ holds with probability at least $1 - \delta/T$.

High probability control of $\xi_{\mathrm{RC}}^-(\tau)$ (14). Let us fix a row $i \in [n]$. From Bernstein's inequality – see e.g. Massart (2007) – it holds with probability at least $1 - \delta/(2Tn)$ that

$$\sum_{k=1}^d B_{ik}^{(\tau)} \leq \lambda_0 d + \sqrt{2\lambda_0 d \log(\tfrac{2Tn}{\delta})} + \log(\tfrac{2Tn}{\delta}) \leq 4\log(\tfrac{2Tn}{\delta})(\lambda_0 d + 1) \leq \psi(\lambda_0 d + 1) .$$

A union bound over all distinct i, j in $[n]$ implies that $\sum_{k=1}^d B_{ik}^{(\tau)} \leq \psi(\lambda_0 d + 1)$ simultaneously for all possible $i \in [n]$, with probability at least $1 - \delta/T$. With a same arguments on the columns $k = 1, \dots, d$, we also have that $\sum_{i=1}^n B_{ik}^{(\tau)} \leq \psi(\lambda_0 n + 1)$ for all $k \in [d]$, with probability at least $1 - \delta/(2T)$. Hence, $\xi_{\mathrm{RC}}^-(\tau)$ holds with probability at least $1 - \delta$.

High probability control of $\xi_{\Gamma}^-(\tau)$ (15). The analysis of $\xi_{\Gamma}^-(\tau)$ is much more challenging. We first reduce the problem to sparse vectors v before applying a union bound over all sparse vectors

and all subsets Γ. For that purpose, let us fix a subset Γ of size $\left\lceil 2\log(\frac{n}{\lambda_0})/\lambda_0 \right\rceil$ and a vector v such that $\|v\|_2 = 1$ and $\|v\|_\infty \le \sqrt{\lambda_0}$. For any $s \in \{1,\dots,n\}$, we define the set of s-sparse vectors as

$$\mathcal{V}_s = \{v' : \ \|v'\|_2 = 1 \text{ and } |v'_i| \in \{0, \tfrac{1}{\sqrt{s}}\}\} \ . \tag{30}$$

In other words, v' is in $\mathcal{V}_s$ if it has exactly s nonzero coordinates which are either equal to $1/\sqrt{s}$ or $-1/\sqrt{s}$. We also define a set of levels $\mathcal{L} := \{\lambda_0 2^{-\alpha} \ : \ \alpha = 1,\dots,\lceil \log_2(n/\lambda_0)\rceil\}$, which satisfies $\min\mathcal{L} \le \lambda_0/n$, $\max\mathcal{L} = \lambda_0/2$ and $|\mathcal{L}| \le 2\log_2(n/\lambda_0)$. For a given $k \in [d]$, we have that

$$\sum_{i=1}^n v_i^2 B_{ik}^{(\tau)} \le \lambda_0 + \sum_{\ell\in\mathcal{L}}\sum_{i=1}^n 2\ell \mathbf{1}\{v_i^2 \in (\ell, 2\ell]\} B_{ik}^{(\tau)} \le \lambda_0 + 2|\mathcal{L}|\ell_k^* \sum_{i=1}^n \mathbf{1}\{v_i^2 \in (\ell_k^*, 2\ell_k^*]\} B_{ik}^{(\tau)} \ ,$$

where ℓ_k^* is the maximizer of $\sum_{i=1}^n 2\ell\mathbf{1}\{v_i \in [\ell, 2\ell)\} B_{ik}^{(\tau)}$ over all ℓ in $\mathcal{L}$. This implies that

$$\begin{aligned}
\min_{k\in\Gamma} \sum_{i=1}^n v_i^2 B_{ik}^{(\tau)} &\le \lambda_0 + 2|\mathcal{L}| \min_{k\in\Gamma} \ell_k^* \sum_{i=1}^n \mathbf{1}\{v_i^2 \in (\ell_k^*, 2\ell_k^*]\} B_{ik}^{(\tau)} \\
&= \lambda_0 + 2|\mathcal{L}| \min_{\ell'\in\mathcal{L}} \min_{\substack{k\in\Gamma \\ \ell_k^* = \ell'}} \ell' \sum_{i=1}^n \mathbf{1}\{v_i^2 \in (\ell', 2\ell']\} B_{ik}^{(\tau)} \\
&\le \lambda_0 + 2|\mathcal{L}| \min_{k\in\Gamma'} \ell^* \sum_{i=1}^n \mathbf{1}\{v_i^2 \in (\ell^*, 2\ell^*]\} B_{ik}^{(\tau)} \ ,
\end{aligned}$$

Where in the last inequality we set

$$\ell^* = \arg\max_{\ell'} |\{k \in \Gamma : \ \ell_k^* = \ell'\}| \quad \text{and} \quad \Gamma' = \{k \in \Gamma : \ \ell_k^* = \ell^*\} \ . \tag{31}$$

From the pingeonhole principle, it holds that $|\Gamma'| \ge |\Gamma|/|\mathcal{L}| \ge \lceil 1/\lambda_0 \rceil$. Since $\|v\| = 1$, there are at most $s^* := \lfloor 1/\ell^* \rfloor$ indices i such that $v_i^2 > \ell^*$. Moreover, each nonzero coefficient $\ell^*\mathbf{1}\{v_i^2 \in (\ell^*, 2\ell^*]\}$ is smaller than $1/s^*$, and from the definition of $\mathcal{L}$, $s^* = \lfloor 1/\ell^* \rfloor \ge 1/\lambda_0$. Hence, letting $\overline{\mathcal{V}} = \cup_{s\ge 1/\lambda_0} \mathcal{V}_s$, it holds that

$$\min_{k\in\Gamma} \sum_{i=1}^n v_i^2 B_{ik}^{(\tau)} \le \lambda_0 + 2|\mathcal{L}| \max_{\tilde{v}\in\overline{\mathcal{V}}} \min_{k\in\Gamma'} \sum_{i=1}^n \tilde{v}_i^2 B_{ik}^{(\tau)} \ . \tag{32}$$

We have reduced the problem of controlling the quantity $\min_{k\in\Gamma'} \sum_{i=1}^n v_i^2 B_{ik}^{(\tau)}$ over all unit vector v to controlling the same quantity over all s-sparse vector $\tilde{v} \in \mathcal{V}_s$, for $s \ge 1/\lambda_0$. Let us fix Γ' of size $\lceil 1/\lambda_0 \rceil$, $s \ge 1/\lambda_0$ and $\tilde{v} \in \mathcal{V}_s$. If $k \in \Gamma'$, then by the Bernstein's inequality, for any $\delta' \in (0,1)$,

$$\mathbb{P}\left(\sum_{i=1}^n \tilde{v}_i^2 B_{ik}^{(\tau)} \ge \lambda_0 + \sqrt{2\frac{\lambda_0}{s}\log(1/\delta')} + \frac{1}{s}\log(1/\delta') \right) \le \delta' \ .$$

Since the random variables $\sum_{i=1}^n \tilde{v}_i^2 B_{ik}^{(\tau)}$ are independent for all $k \in \Gamma'$,

$$\mathbb{P}\left(\min_{k\in\Gamma'} \sum_{i=1}^n \tilde{v}_i^2 B_{ik}^{(\tau)} \ge \lambda_0 + \sqrt{2\frac{\lambda_0}{s}\log(1/\delta')} + \frac{1}{s}\log(1/\delta') \right) \le (\delta')^{|\Gamma'|} \ .$$

Letting $\delta'' = (\delta')^{|\Gamma'|}$, it follows that with probability at least $1-\delta''$,

$$\min_{k\in\Gamma'} \sum_{i=1}^{n} \tilde{v}_i^2 B_{ik}^{(\tau)} \leq \lambda_0 + \sqrt{2\frac{\lambda_0}{s|\Gamma'|}\log(1/\delta'')} + \frac{1}{s|\Gamma'|}\log(1/\delta'') \ .$$

Remark that $|\mathcal{V}_s| = 2^s\binom{n}{s}$ and that the number of possible subset Γ' is $\binom{d}{\lceil 1/\lambda_0\rceil}$. We now choose δ'' equal to $\delta/(TN)$, where $N = n2^s\binom{n}{s}\binom{d}{\lceil 1/\lambda_0\rceil}$. A union bound gives that with probability at least $1-\delta/T$, it holds simultaneously for all $s = 1,\dots,n$, all $v \in \mathcal{V}_s$ and all $\Gamma' \subset [d]$ of size $|\Gamma'| = \lceil 1/\lambda_0\rceil$:

$$\begin{aligned}\min_{k\in\Gamma'} \sum_{i=1}^{n} \tilde{v}_i^2 B_{ik}^{(\tau)} &\leq \lambda_0 + \sqrt{4(\frac{\lambda_0}{s} + \frac{\lambda_0}{|\Gamma'|})\log(ndT/\delta)} + 2(\frac{1}{s} + \frac{1}{|\Gamma'|})\log(ndT/\delta) \\ &\leq 8\lambda_0 \log(ndT/\delta) \ ,\end{aligned}$$

where we used the rough upperbound $\log(N/\delta) \leq 2\log(ndT/\delta)(s+|\Gamma'|)$ in the first inequality and $s \geq 1/\lambda_0$, $|\Gamma'| \geq 1/\lambda_0$ in the second inequality. Injecting this latter upperbound in (32), we obtain that with probability at least $1-\delta$, for any unit vector v such that $\|v\|_\infty \leq \sqrt{\lambda_0}$ and any $\Gamma' \subset [d]$ of size $\lceil 1/\lambda_0\rceil$,

$$\min_{k\in\Gamma'} \sum_{i=1}^{n} v_i^2 B_{ik}^{(\tau)} \leq \lambda_0 + 16|\mathcal{L}|\lambda_0\log(ndT/\delta) \leq \psi^2\lambda_0 \ .$$

We conclude that $\mathbb{P}(\xi_\Gamma^-) \geq 1-\delta/T$.

Concentration of $\xi_C^+(\tau)$ (16) Let us fix $k \in [d]$ and l, r such that $\lambda_0(r-l) \geq \psi$. It holds from Bernstein's inequality that with probability at least $1-\delta/(n^2dT)$,

$$\sum_{i=l}^{r} B_{ik}^{(\tau)} \geq \lambda_0(r-l) - \sqrt{2\lambda_0(r-l)\log(n^2dT/\delta)} - \log(n^2dT/\delta) \geq \lambda_0(r-l)/2 \ .$$

The last inequality comes from the the fact that if $\lambda_0(r-l) \geq \psi \geq 32\log(n^2dT/\delta)$, then $\sqrt{2\lambda_0(r-l)\log(n^2dT/\delta)} + \log(n^2dT/\delta) \leq \lambda_0(r-l)/2$. We conclude by a union bound argument over all possible $l \leq r$ in $[n]^2$ and k in $[d]$ that $\mathbb{P}(\xi_{\mathrm{op}}^+(\tau)) \geq 1-\delta/T$.

Concentration of ξ_{op}^+ (17) The event ξ_{op}^+ consists in controlling the operator norm of the matrix $B^\tau \odot M - \lambda_0 M$ whose expectation is zero and whose entries are heteroskedastic. Unfortunately, we cannot apply the results in Cai et al. (2022) as we did in the proof of the concentration of ξ_{op} (12). Indeed, the Bernoulli random variables $B_{ik}^{(\tau)}$ are of variance of order λ_0 but are not λ_0-subGaussians.

For this reason, we rather apply the proposition 4.1 of Pilliat et al. (2024). The coefficients of $B^{(\tau)} \odot M - \lambda_0 M$ are independent, of mean 0 and satisfy the Bernstein condition (24) of Pilliat et al. (2024) with $\sigma^2 := 2\lambda_0$ and $K := 1$ since $\mathbb{E}[((B_{ik}-\lambda_0)M_{ik})^{2u}] \leq \lambda_0$. Let us write $X = B^{(\tau)} \odot M - \lambda_0 M$. Using the identity matrix for the orthogonal projection Λ in Pilliat et al. (2024), we obtain that with probability at least $1-\delta/T$,

$$\begin{aligned}\|XX^T - \mathbb{E}[XX^T]\|_{\mathrm{op}} &\leq \kappa\left[\sqrt{(\lambda_0^2 nd + \lambda_0 d)\log(nT/\delta)} + (\lambda_0 n + \log(d))\log(nT/\delta)\right] \\ &\leq \kappa' \log^2(\tfrac{ndT}{\delta})\left[\lambda_0 d + 1\right] \ ,\end{aligned}$$

for some numerical constants κ, κ'. In the second inequality, we used that $\sqrt{x+y} \le 2(\sqrt{x}+\sqrt{y})$ for any $x, y > 0$. Since $\mathbb{E}[XX^T]$ is a diagonal matrix with entrie less than $d\lambda_0$, we have that $\|\mathbb{E}[XX^T]\|_{\mathrm{op}} \le \lambda_0 d$. This implies that for some numerical constant C'_{op}.

$$\|(B^{(\tau)} \odot M - \lambda_0 M)(B^{(\tau)} \odot M - \lambda_0 M)^T\|_{\mathrm{op}} \le C'_{\mathrm{op}} \log^2\left(\tfrac{ndT}{\delta}\right)[\lambda_0 d + 1] \quad .$$

We conclude that $\mathbb{P}(\xi^+_{\mathrm{op}}(\tau)) \ge 1 - \delta/T$.

Appendix D. Proof of the Lower Bound (Theorem 2)

D.1. Proof of the Lower Bound (Theorem 2)

The proof is mainly based on sufficient statistic arguments – see also the proof of Lemma J.2 in Pilliat et al. (2023) for similar ideas and computations. Let M be a matrix whose coefficients are all equal to a known quantity $u > 0$, and $\hat{x}$ be any estimator of the labels. We denote $\mathbb{P}_{x^*}$ and $\mathbb{E}_{x^*}$ for the distribution and expectation of the matrix Y, and $\mathbb{P}_1^{(-k)}$ (resp. $\mathbb{P}_{-1}^{(-k)}$) for the distribution of the data when x^*_k is replaced by 1 (resp. -1).

Case $\lambda = 1$. It holds that

$$\begin{aligned}
\max_{x^* \in \{-1,1\}} \sum_{k=1}^d \mathbb{E}_{x^*}[M \operatorname{diag}(x^*_k \neq \hat{x}_k)] &\ge nu^2 \max_{x^* \in \{-1,1\}} \sum_{k=1}^d \mathbb{P}_{x^*_k}(x^*_k \neq \hat{x}_k) \\
&\ge nu^2 \max_{x^* \in \{-1,1\}} \sum_{k=1}^d \frac{1}{2}\left(\mathbb{P}_1^{(-k)}(\hat{x}_k = -1) + \mathbb{P}_{-1}^{(-k)}(\hat{x}_k = 1)\right) \\
&\ge nu^2 \max_{x^* \in \{-1,1\}} \sum_{k=1}^d \frac{1}{2}\left(1 - d_{TV}\left(\mathbb{P}_1^{(-k)}, \mathbb{P}_{-1}^{(-k)}\right)\right) \quad ,
\end{aligned}$$

Since the noise is Gaussian, $\frac{1}{n}\sum_{i=1}^n Y_{ik} \sim \mathcal{N}(x^*_k u, n)$ is a sufficient statistic for estimating x^*_k. Hence, letting $\mathbb{P}_1$ (resp. $\mathbb{P}_{-1}$) be its distribution when $x^*_k = 1$ (resp. -1), a data processing argument[1] gives

$$d_{TV}\left(\mathbb{P}_1^{(-k)}, \mathbb{P}_{-1}^{(-k)}\right) \le d_{TV}\left(\mathbb{P}_1, \mathbb{P}_{-1}\right) . \tag{33}$$

From the Pinsker's inequality, we obtain

$$\max_{x^* \in \{-1,1\}} \sum_{k=1}^d \mathbb{E}_{x^*}[M \operatorname{diag}(x^*_k \neq \hat{x}_k)] \ge \frac{ndu^2}{2}\left(1 - \sqrt{\frac{1}{2}\operatorname{KL}(\mathbb{P}_1, \mathbb{P}_{-1})}\right) = \frac{ndu^2}{2}\left(1 - \sqrt{nu^2}\right) \quad .$$

Choosing $u = \frac{2}{3\sqrt{n}}$, this gives

$$\max_{x^* \in \{-1,1\}} \sum_{k=1}^d \mathbb{E}_{x^*}[M \operatorname{diag}(x^*_k \neq \hat{x}_k)] \ge \frac{2}{27} d . \tag{34}$$

1. see e.g. Polyanskiy and Wu (2014)

Case $\lambda < 1$. In this case, we create an easier Poisson observation model from the Bernoulli observation model. Let λ' be such that $1 - e^{-\lambda'} = \lambda$, that is $\lambda' = -\log(1-\lambda)$. For each coefficient (i,k) such that $Y_{ik} \neq 0$, an oracle randomly generates random variable N_{ik} that follows a Poisson distribution of parameter λ', conditionally on the event $(N_{ik} \geq 1)$. Then, in addition to the first observation $Y_{ik}^{(1)} := Y_{ik}$, the oracle create $N_{ik} - 1$ new independent observations $(Y_{ik}^{(s)})_{s=2,\dots,N_{ik}}$ that are i.i.d. $\mathcal{N}(x^* u, 1)$. Consider $\mathbb{P}'_{x^*}$ and $\mathbb{E}'_{x^*}$ the distribution of the data when the statistican observes $(N_{ik}, (Y_{ik}^{(s)})_{s=1,\dots,N_{ik}})$ if $Y_{ik} \neq 0$ and $(0,0)$ if $Y_{ik} = 0$. Since an estimator $\hat{x}'$ can simply ignore the new observations generated by the oracle, we obtain that

$$\inf_{\hat{x}} \max_{x^* \in \{-1,1\}} \sum_{k=1}^{d} \mathbb{E}_{x^*}[M \operatorname{diag}(x_k^* \neq \hat{x}_k)] \geq \inf_{\hat{x}} \max_{x^* \in \{-1,1\}} \sum_{k=1}^{d} \mathbb{E}'_{x^*}[M \operatorname{diag}(x_k^* \neq \hat{x}_k)] \ . \tag{35}$$

Notice that, on the right hand side, the N_{ik}'s have parameter λ' while on the left hand side, the B_{ik}'s have parameter λ.

We now give a lower bound on the maximum risk of $\hat{x}$ in the Poisson observation scheme. Let $N_k = \sum_{i=1}^n N_{ik}$ and $Y_k^{\downarrow} = \frac{1}{N_k} \sum_i \sum_{s=1}^{N_{ik}} Y_{ik}^{(s)}$ be the average of all the N_k observations in column k. Since the observations are Gaussian, $Z_k = (N_k, Y_k^{\downarrow})$ is a sufficient statistic for estimating x_k^* – we set $Z_k = (0,0)$ if $N_k = 0$. Hence, letting $\mathbb{P}'_1$ (resp. $\mathbb{P}'_{-1}$) be the distribution of Z_k when $x_k^* = 1$ (resp. -1), we have, as in the case $\lambda = 1$, that

$$d_{TV}\left(\mathbb{P}_1'^{(-k)}, \mathbb{P}_{-1}'^{(-k)}\right) \leq d_{TV}\left(\mathbb{P}'_1, \mathbb{P}'_{-1}\right) \leq \sqrt{\frac{1}{2}\operatorname{KL}(\mathbb{P}'_1, \mathbb{P}'_{-1})} \leq \sqrt{\frac{1}{2}\log\left(1 + \chi^2(\mathbb{P}'_1, \mathbb{P}'_{-1})\right)} \ . \tag{36}$$

We used Jensen's inequality and the convexity of $\exp$ in the last inequality. Plugging (36) in the above inequality, we obtain

$$\max_{x^* \in \{-1,1\}} \sum_{k=1}^{d} \mathbb{E}_{x^*}[M \operatorname{diag}(x_k^* \neq \hat{x}_k)] \geq \frac{ndu^2}{2}\left(1 - \sqrt{\frac{1}{2}\log\left(1 + \chi^2(\mathbb{P}'_1, \mathbb{P}'_{-1})\right)}\right) \ . \tag{37}$$

Let ν be the Lebesgue measure on $\mathbb{R}$ and μ be the counting measure – $\mu(S) = +\infty$ if S is infinite and $\mu(S) = |S|$ otherwise. If $\epsilon \in \{-1,1\}$, the density of $\mathbb{P}'_\epsilon$ with respect to $\mu \otimes \nu$ is

$$\alpha_\epsilon(r,y) = \begin{cases} \frac{(n\lambda')^r}{r!} e^{-\lambda'} \frac{1}{\sqrt{2\pi r}} e^{-\frac{(y-\epsilon u r)^2}{2r}} & \text{if } x \geq 1, \\ e^{-\lambda'} \text{ if } r = 0 \text{ and } y = 0, \\ 0 \text{ otherwise.} \end{cases} \tag{38}$$

Hence, letting $\mathbb{N}^* = \mathbb{N} \setminus \{0\}$,

$$
\begin{aligned}
1+\chi^2(\mathbb{P}'_1,\mathbb{P}'_{-1}) &= e^{-n\lambda'} + \int_{\mathbb{N}^*\times\mathbb{R}} \frac{(\alpha_1(r,y))^2}{\alpha_{-1}(r,y)} d\mu(r)dy \\
&= e^{-n\lambda'} + \int_{\mathbb{N}^*\times\mathbb{R}} \frac{(n\lambda')^r}{r!} e^{-n\lambda'} \frac{1}{\sqrt{2\pi r}} \exp\left(-\frac{(y-ur)^2}{r} + \frac{(y+ur)^2}{2r}\right) d\mu(r)dy \\
&= e^{-n\lambda'} + \int_{\mathbb{N}^*\times\mathbb{R}} \frac{(n\lambda')^r}{r!} e^{-n\lambda'} \frac{1}{\sqrt{2\pi r}} \exp\left(-\frac{(y-3ur)^2}{2r} + 4u^2 r\right) d\mu(r)dy \\
&= \sum_{r\geq 0} \frac{(n\lambda')^r}{r!} e^{-n\lambda'} e^{4u^2 r} \\
&= \exp\left(n\lambda'(e^{4u^2}-1)\right) .
\end{aligned}
$$

Combining this last inequality with (37), we obtain

$$
\max_{x^*\in\{-1,1\}} \sum_{k=1}^{d} \mathbb{E}_{x^*}[M \operatorname{diag}(x_k^* \neq \hat{x}_k)] \geq \frac{ndu^2}{2}\left(1 - \sqrt{\frac{n\lambda'}{2}(e^{4u^2}-1)}\right) .
$$

Let us take $u = \frac{1}{2}\sqrt{\log(1+\nu')}$ with $\nu' = \frac{8}{9n\lambda'}$. Since $\lambda' = -\log(1-\lambda) \leq \lambda$, it holds that $\nu \geq \nu' = \frac{8}{9n\lambda'}$. We obtain

$$
\max_{x^*\in\{-1,1\}} \sum_{k=1}^{d} \mathbb{E}_{x^*}[M \operatorname{diag}(x_k^* \neq \hat{x}_k)] \geq \frac{nd}{3\cdot 8}\log(1+\nu') \geq \frac{nd}{3\cdot 8}\log(1+\nu) .
$$

Hence,

$$
\max_{x^*\in\{-1,1\}} \sum_{k=1}^{d} \mathbb{E}_{x^*}[M \operatorname{diag}(x_k^* \neq \hat{x}_k)] \geq \frac{\log(1+\nu)}{\nu}\frac{\nu nd}{3\cdot 8} \geq \frac{1}{1+\nu}\frac{d}{27\lambda} = \frac{9n\lambda}{9n\lambda+8}\frac{d}{27\lambda} ,
$$

where we used the inequality $\log(1+x)/x \geq 1/(x+1)$ in the last inequality. This concludes the proof of Theorem 2.