Multiple effects low-rank model with missing observations

Geneviève Robin[1], Olga Klopp[2,3], and Éric Moulines[1]

[1]CMAP, UMR 7641, École Polytechnique
[2]ESSEC Business School
[3]CREST-ENSAE, UMR CNRS 9194

March 3, 2018

Abstract

1 Introduction

Estimation of high-dimensional matrices has recently aroused interest in a number of applications where data are often incomplete, noisy and contain variables of different types. In health care in particular, databases collect measurements (usually arranged in columns) across patients (usually arranged in rows). These measurements can be numeric (blood pressure, hemoglobin level, etc), categorical (sex, type of illness, etc) or discrete (number of days spent in intensive care, etc.). Such collections of heterogeneous variables are often called *mixed data* [Pagès, 2015, Chapter 3]. Some entries in such data sets may be missing, for a variety of reasons including the impossibility to perform a medical examination due to technical issues or because the patient's state is too severe. Low-rank models - which embed rows and columns of the data set into low-dimensional spaces - are useful in this setting for tasks such as visualization, imputation and clustering.

To this end, low-rank models designed to analyze categorical, discrete, and possibly heterogeneous features have been proposed in the literature in the recent years. In their early contribution, [Collins et al., 2001] extended Principal Component Analysis (PCA) to exponential family probabilistic models, while [Gordon, 2002] developped Generalized^2 Linear^2 Models to extend PCA to losses derived from generalized Bregman divergences. Srebro [2004] later focused on regularized versions of this framework, adding even more general loss functions, including the hinge loss, and regularizers such as the nuclear norm. [Tropp, 2004, Chapter 8] linked the framework of low-rank modeling to classical clustering problems. Even more recently, Alquier et al. [2017] derived sharp oracle inequalities for a number of regularized problems including low-rank matrix completion with Lipschitz loss functions.

In the field of exploratory data analysis, Factorial Analysis of Mixed Data (FAMD) [Pagès, 2015] is an alternative to low-rank models which computes summary components for mixed data sets. The components of FAMD maximize the correlation with the quantitative variables and the square correlation ratio with the qualitative variables. Finally, Udell et al. [2016] proposed a framework for *generalized low-rank models*, to approximate heterogeneous data sets with missing values by the product of two low-dimensional factors through a range of minimization problems. In particular, they consider cases where the loss function depends on the column, allowing them to model both quantitative and qualitative data in a single data set. They developed algorithms which are applicable to a wide variety of settings. To the best of our knowledge, the statistical guarantees of the models considered by Udell et al. [2016] have not been studied so far.

In the present paper, we propose an extension of generalized low-rank models to cases where the signal comes from the superimposition of several "effects" in the data. Returning to our example in health care, medical centers are increasingly inclined to aggregate their data, giving a natural hierarchy where patients are nested within hospitals and which is usually called *multilevel* structure [Bryk and Raudenbush, 1992]. The hospitals potentially have a large effect on the measured features, due to lack of measurement standardization as well as social and geographical discrepancies between hospital populations. For the purpose of better understanding the population and practices, it can be of interest to model and estimate separately the "hospital effect" and the low-rank structure of the data, or "individual effect". Another example comes from ecological applications where species are counted across environments. Such data are classically analyzed using the log-linear model [Agresti, 2013] where the log of the expectations of species abundances are modeled as the sum of species and environment effects, plus a low-rank interaction term. Finally, the well-known *robust matrix completion* problem [Klopp et al., 2017] also falls in this landscape. This problem arises in recommendation systems where a few malicious users cohabitate with regular users.

Formally, we consider a data matrix $W \in \mathbb{R}^{m_1 \times m_2}$ generated by the following exponential family model. For all $(i,j) \in [\![m_1]\!] \times [\![m_2]\!]$ there exists a parameter X^0_{ij} such that

$$W_{ij}|X^0_{ij} \sim \mathrm{Exp}_{h_j,g_j}(X^0_{ij}), \tag{1}$$

where Exp_{h_j,g_j} denotes the exponential family with base measure h_j and link function g_j. The functions g_j model heterogeneous features (numeric, binary, discrete, etc.). For example, if the j-th variable is binary, g_j might be set to be the logistic link function; if the j-th variable is a count, g_j might be set to be the exponential link corresponding to a Poisson law; if it is quantitative, g_j might be set to be a quadratic function modeling a Gaussian variable.

We assume that X^0, appearing in (1) can be decomposed as

$$X^0 = \sum_{k=1}^{N} \alpha_k^0 U_k + L^0, \tag{2}$$

where $1 \leq N \leq m_1 m_2$, $\mathcal{U} = (U_1, \ldots, U_N)$ is a fixed dictionary of N matrices of $\mathbb{R}^{m_1 \times m_2}$, α^0 is a sparse vector with unknown support $\mathcal{I} = \{k \in [\![N]\!]; \alpha_k^0 \neq 0\}$ and L^0 has low-rank. We further assume the presence of missing values, *i.e* we observe a data frame

$$Y = W_\Omega := W \odot \Omega, \quad \Omega = (\omega_{ij})_{i,j} \in \{0,1\}^{m_1 \times m_2}. \tag{3}$$

Here, $\odot$ denotes the entry-wise product, $\omega_{ij} = 1$ if W_{ij} is observed and $\omega_{ij} = 0$ otherwise. The goal is to estimate L^0 and α^0 from these partial and noisy observations.

In the matrix completion literature a particular case of our model is the robust matrix completion problem, where one wants to estimate a matrix X^0 from partial observation of its entries, some of which have been corrupted. X^0 is usually modeled as the sum of a low-rank matrix L^0 and an entry-wise or column-wise sparse matrix S^0: $X^0 = S^0 + L^0$. This problem has raised a lot of attention in the past years, and many theoretical results were proved in both noisy and deterministic cases. In the noiseless and complete setting, Chandrasekaran et al. [2011] proved that exact recovery of (S^0, L^0) is possible with high probability, under appropriate identifiability conditions; Hsu et al. [2011] then proved the same result under milder conditions. Also in the noiseless setting but when entries are partially observed, Candès et al. [2011] proved exact recovery with high probability, and Xu et al. [2010] show a similar result when S^0 is column-wise sparse. Robust matrix completion from noisy observation is considered in Klopp et al. [2017], where the authors prove minimax convergence rates for the low-rank and sparse components. Agarwal et al. [2012] also prove non-asymptotic error bounds for matrix decomposition from noisy observations of a linear transformation of the initial matrix. In Elsener and van de Geer [2016] the authors tackle the robustness problem via absolute value and Huber loss functions. All these papers study the case of additive noise.

A closely related model was studied by Fithian and Mazumder [2013], who incorporate row and column covariates in their estimation procedure, by using a generalized nuclear norm penalty. Let R and C be matrices of row and column features respectively. Then Π_R and Π_C are orthogonal projection matrices on the linear span of R and C. The penalty used in Fithian and Mazumder [2013] is of the form $\left\|(I - \Pi_R)X(I - \Pi_C)^\top\right\|_*$, where $\|\cdot\|_*$ is the nuclear norm. In other words, only the directions orthogonal to R and C are penalized. The authors in Fithian and Mazumder [2013] mainly focus on optimization procedures, and to the best of our knowledge they did not provide statistical guarantees. In the present paper we consider a more gneral model and our estimation procedure is different.

Finally, let us mention the extensions of component methods to account for multiple effects in the data. Multilevel Simultaneous Component Analysis (MLSCA), introduced by Timmerman [2006], provides a component method for exploratory purposes in hierarchically structured data, by modeling the variance between and within groups separately. MLSCA operates by estimating simultaneously principal directions of variability for the between groups variability and for the within groups variability. It consists in performing two separate PCA, on a matrix containing the mean values of the features per group, and on the original data matrix centered by group.

The main contributions of the present paper are two-fold. First, we introduce a general for mixed features and multiple effects. Then, we propose an estimation procedure through the minimization of a doubly penalized negative log-likelihood. For this procedure, we derive upper bounds on the Frobenius estimation risks of both components α^0 and L^0. To the best of our knowledge such guarantees are not available in the literature. We also prove lower bounds that show that, in a number of situations, our upper bounds are near optimal.

The paper is organized as follows. In Section 2, we introduce our main assumptions and define our estimator through a convex program; we also state our main results. In Theorem 2, we provide upper bounds on the estimation error of both L^0 and α^0. We also present lower bounds in Theorem 3 and discuss cases in which they match the upper bounds. In Section 3, we specialize the general results to three examples of interest in applications, namely the multilevel model, the log-linear model and robust matrix completion. Finally, the proofs are in Section 4.

Along this article, we will use the following notation.

- $\vee$ and $\wedge$ denote the max and min operators respectively.
- $M = m_1 \vee m_2$, $m = m_1 \wedge m_2$ and $d = m_1 + m_2$.
- In $\mathbb{R}^{m_1 \times m_2}$, $\|\cdot\|_F$ is the Frobenius norm, $\|\cdot\|_*$ the nuclear norm (or trace norm, that is the sum of singular values), $\|\cdot\|$ the operator norm (the largest singular value) and $\|\cdot\|_\infty$ the infinity norm (the largest entry in absolute value).
- In $\mathbb{R}^N$, $\|\cdot\|_1$ is the ℓ_1-norm, $\|\cdot\|_2$ the Euclidean norm, $\|\cdot\|_\infty$ the infinity norm (the largest entry in absolute value), and $\|\cdot\|_0$ the ℓ_0-norm (the number of non zero coefficients).

- For some integer $n \geq 0$ we denote by $[\![n]\!]$ the set of integers $\{k \in \mathbb{N}, k \leq n\}$.
- For $(i,j) \in [\![m_1]\!] \times [\![m_2]\!]$, define $E_{ij} = e_i(m_1)e_j(m_2)'$, where $(e_k(l))$ are the canonical basis vectors of $\mathbb{R}^l$. That is, E_{ij}, $(i,j) \in [\![m_1]\!] \times [\![m_2]\!]$ is the canonical basis of $\mathbb{R}^{m_1 \times m_2}$.
- C is a numerical constant whose value might change from one instance to the other.
- $r = \operatorname{rank}\left(L^0\right)$ and $s = \left\|\alpha^0\right\|_0$.
- Let $\mathsf{f}_U : \mathbb{R}^N \to \mathbb{R}^{m_1 \times m_2}$ be the linear application such that for $\alpha \in \mathbb{R}^N$:

$$\mathsf{f}_U(\alpha) = \sum_{k=1}^{N} \alpha_k U_k.$$

- We denote by Π the distribution of the mask Ω. For all $(i,j) \in [\![m_1]\!] \times [\![m_2]\!]$, π_{ij} is the probability of observing entry Y_{ij}, *i.e.* $\pi_{ij} = \mathbb{P}\left(\omega_{ij} = 1\right)$.
- Let $G(\mathcal{U}) \in \mathbb{R}^{N \times N}$ be the Gram matrix of the dictionary $\mathcal{U}$ defined by $G(\mathcal{U})_{kl} = \langle U_k, U_l \rangle$ for all $(k,l) \in [\![N]\!] \times [\![N]\!]$.

2 Assumptions and main results

We start by introducing assumptions on our model. For $a \geq 0$ and a sparsity pattern $\mathcal{I} \subset N$ define two sets of matrices, with for $\alpha \in \mathbb{R}^N$, $\operatorname{supp}(\alpha) = \{k \in [\![N]\!], \alpha_k \neq 0\}$.

$$\mathcal{E}_{a,\mathcal{I}} = \Bigg\{ L \in \mathbb{R}^{m_1 \times m_2}, \alpha \in \mathbb{R}^N; \|L\|_\infty \leq a, \\ \|\alpha\|_\infty \leq a, \max_{k \in \mathcal{I}} |\langle L, U_k \rangle| = 0, \operatorname{supp}(\alpha) \subset \mathcal{I} \Bigg\}, \quad (4)$$

$$\mathcal{X}_{a,\mathcal{I}} = \left\{ X = \mathsf{f}_U(\alpha) + L; (L, \alpha) \in \mathcal{E}_{a,\mathcal{I}} \right\}. \quad (5)$$

Note that, we do not assume that $\mathcal{I}$ is known. First, we consider assumptions on the dictionary $\mathcal{U}$.

H 1. *For æ > 0, all $k \in [\![N]\!]$ and $(i,j) \in [\![m_1]\!] \times [\![m_2]\!]$, $(U_k)_{ij} \in [-1,1]$. Furthermore for all $(i,j) \in [\![m_1]\!] \times [\![m_2]\!]$, $\sum_{k=1}^{N} |(U_k)_{ij}| \leq$ æ.*

Assumption **H** 1 guarantees that for all $(L, \alpha) \in \mathcal{E}_{a,\mathcal{I}}$, $X = \mathsf{f}_U(\alpha) + L$ satisfies $\|X\|_\infty \leq (1 + æ)a$. In particular, this condition is true in the multilevel and log-linear models as well as in robust matrix completion.

H 2. *For $\kappa > 0$ and all $\alpha \in \mathbb{R}^N$, $\alpha^T G(\mathcal{U}) \alpha \geq \kappa^2 \|\alpha\|_2^2$.*

Assumption **H** 2 guarantees that for $\alpha \in \mathbb{R}^N$, we have $\|\alpha\|_2^2 \leq \kappa^{-2} \|\mathsf{f}_U(\alpha)\|_F^2$. In this paper we do not consider the case where $N > m_1 m_2$ and the Gram matrix $G(\mathcal{U})$ is singular.

H 3. $(L^0, \alpha^0) \in \mathcal{E}_{a,\mathcal{I}}$.

The condition $\langle L^0, U_k \rangle = 0$ for all $k \in \mathcal{I}$ and Assumption **H** 2 guarantee that every matrix $X \in \mathcal{X}_{a,\mathcal{I}}$ has a unique decomposition $X = \mathsf{f}_U(\alpha) + L$, $(L, \alpha) \in \mathcal{E}_{a,\mathcal{I}}$. All our results can also be derived in the case where the bound on $|L_{ij}^0|$ depends on i and j and the bound on $|\alpha_k^0|$ depends on k, but we stick to the definition of $\mathcal{E}_{a,\mathcal{I}}$ for simplicity.

For $0 < \sigma_-, \sigma_+, \gamma < +\infty$ consider the following assumption on the link functions g_j.

H 4. *The functions g_j are twice differentiable, and for all $x \in [-(1+æ)a - \gamma, (1+æ)a + \gamma]$, and $j \in [\![m_2]\!]$*

$$\sigma_-^2 \leq g_j''(x) \leq \sigma_+^2.$$

Assumption **H**4 implies the sub-exponentiality of the Y_{ij} through the following lemma.

Lemma 1. *Assume that Ξ follows an exponential family distribution $\mathrm{Exp}_{h,g}(x)$ with $x \in [-(1+æ)a, (1+æ)a]$. Assume that g is twice differentiable and that for all $x \in [-(1+æ)a - \gamma, (1+æ)a + \gamma]$,*

$$\sigma_-^2 \leq g''(x) \leq \sigma_+^2.$$

Then Ξ is a sub-exponential random variable with scale and variance parameters $1/\gamma$ and σ_+^2 respectively.

Proof. We compute the exponential moment of Ξ. For all z, $|z| < \gamma$

$$\mathbb{E}\left[\mathrm{e}^{z(\Xi - \mathbb{E}[\Xi])}\right] = \mathrm{e}^{-zg'(x)} \int h_j(\xi) \mathrm{e}^{\xi x - g(x)} \mathrm{e}^{z\xi} d\xi$$
$$= \mathrm{e}^{g(x+z) - g(x) - zg'(x)} \leq \mathrm{e}^{\frac{\sigma_+^2 x^2}{2}},$$

where we use that $(z + x, x) \in [-(1+æ)a - \gamma, (1+æ)a + \gamma]^2$ and the strong convexity of g on this interval. Now the characterization of sub-exponential variables [Vershynin, 2012, Section 5.2.4] implies the statement of the lemma. □

We finally state the assumptions regarding the sampling distribution. For $0 < p \leq 1$ consider the following assumption.

H 5. *The Bernoulli random variables ω_{ij}, $(i, j) \in [\![m_1]\!] \times [\![m_2]\!]$, defined in (3) are independent, and for all $(i, j) \in [\![m_1]\!] \times [\![m_2]\!]$, $\pi_{ij} \geq p$*

Assumption **H** 5 implies that for $A \in \mathbb{R}^{m_1 \times m_2}$, $\|A\|_{\Pi}^2 \geq p\|A\|_F^2$. For $j \in [\![m_2]\!]$, denote by $\pi_{.j} = \sum_{i=1}^{m_1} \pi_{ij}$, $j \in [\![m_2]\!]$ the probability of sampling an element in the j-th column. Similarly, for $i \in [\![m_1]\!]$, denote by $\pi_{i.} = \sum_{j=1}^{m_2} \pi_{ij}$ the probability of observing an element in the i-th row. We define the upper bound: $\max_{i,j}(\pi_{i.}, \pi_{.j}) \leq \beta$. For $B \in \mathbb{R}^{m_1 \times m_2}$ we define $\|B\|_{\Omega}^2 = \|B_{\Omega}\|_F^2$, and $\|B\|_{\Pi}^2 = \mathbb{E}\left[\|B\|_{\Omega}^2\right]$, where the expectation is taken with respect to Π.

We now define our estimator as the minimizer of a penalized negative log-likelihood. The negative log-likelihood of a matrix $X \in \mathbb{R}^{m_1 \times m_2}$ is defined as

$$\mathcal{L}(X; Y, \Omega) = \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} \omega_{ij} \left\{-Y_{ij} X_{ij} + g_j(X_{ij})\right\}. \tag{6}$$

We estimate L^0 and α^0 using the following convex program, where the nuclear norm and ℓ_1 norm penalties are convex relaxations of the rank and sparsity constraints respectively.

$$\hat{L}, \hat{\alpha} \quad \in \underset{\substack{\|L\|_{\infty} \leq a \\ \|\alpha\|_{\infty} \leq a}}{\operatorname{argmin}} \quad \left\{\mathcal{L}\left(\mathsf{f}_U(\alpha) + L; Y, \Omega\right) + \lambda_1 \|L\|_* + \lambda_2 \|\alpha\|_1\right\}. \tag{7}$$

The regularization parameters λ_1 and λ_2 control the trade-off between the data-fitting term and the low-rank and sparsity assumptions. In the sequel we denote by $\hat{X} = \mathsf{f}_U(\hat{\alpha}) + \hat{L}$. The gradient of the negative log-likelihood $\mathcal{L}$ with respect to X is given by

$$\nabla\mathcal{L}(X; Y, \Omega) = \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} \omega_{ij} \left\{-Y_{ij} + g_j'\left(X_{ij}\right)\right\} E_{ij}. \tag{8}$$

Note that, assumption **H**4 implies that for any pair of matrices satisfying $\|X\|_{\infty}, \left\|\tilde{X}\right\|_{\infty} \leq (1 + æ)a$ the two following inequalities hold:

$$\mathcal{L}(X; Y, \Omega) - \mathcal{L}(\tilde{X}; Y, \Omega) - \langle \nabla\mathcal{L}(\tilde{X}; Y, \Omega), X - \tilde{X}\rangle \geq \frac{\sigma_-^2}{2} \left\|X - \tilde{X}\right\|_{\Omega}^2, \tag{9}$$

$$\left\|\nabla\mathcal{L}(X; Y, \Omega) - \nabla\mathcal{L}(\tilde{X}; Y, \Omega)\right\|_F \leq \sigma_+^2 \left\|X - \tilde{X}\right\|_{\Omega}. \tag{10}$$

Upper bounds We now provide upper bounds for the Frobenius norm of the errors $\Delta X = X^0 - \hat{X}$, $\Delta L = L^0 - \hat{L}$ and $\Delta\alpha = \alpha^0 - \hat{\alpha}$. Let $\{\epsilon_{ij}\}$ be an i.i.d. Rademacher sequence independent of Y and Ω. We define

$$\Sigma_R = \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} \omega_{ij} \epsilon_{ij} E_{ij}.$$

In Theorem 1 we give a general result under some assumptions on the regularization parameters λ_1 and λ_2, which depend on the random matrices $\nabla\mathcal{L}(X^0; Y, \Omega)$

and Σ_R. Then, Lemma 2 and 3 allow us to compute values of λ_1 and λ_2 that satisfy the assumptions of Theorem 1 with high probability. Finally we combine these results in Theorem 2.

Define $u = \max_k \|U_k\|_1$ and the following quantities

$$\Phi_\alpha = \frac{\|\alpha^0\|_1}{p}\left\{\frac{\lambda_2}{\sigma_-^2} + a^2 u \mathbb{E}\left[\|\Sigma_R\|_\infty\right]\right\} + \left(\frac{a}{p}\right)^2 \log(d), \tag{11}$$

$$\Phi_L = \frac{r}{p^2}\mathbb{E}\left[\|\Sigma_R\|\right]^2 + \frac{\|\alpha^0\|_1}{p}\left\{\frac{\lambda_2}{(1+æ)a\lambda_1} + u\mathbb{E}\left[\|\Sigma_R\|_\infty\right]\right\} + \Phi_\alpha. \tag{12}$$

Theorem 1. *Let*

$$\lambda_1 \geq 2\left\|\nabla\mathcal{L}(X^0;Y,\Omega)\right\|, \quad \lambda_2 \geq 2u\left(\left\|\nabla\mathcal{L}(X^0;Y,\Omega)\right\|_\infty + 2\sigma_+^2(1+æ)a\right),$$

*and assumptions **H** 1-5 hold. Then, with probability at least* $1 - 8d^{-1}$

$$\begin{aligned} &(i) \qquad \|\mathsf{f}_U(\Delta\alpha)\|_F^2 \leq C\Phi_\alpha, \ \|\Delta\alpha\|_2^2 \leq \frac{C}{\kappa^2}\Phi_\alpha, \ \textit{and} \\ &(ii) \qquad \|\Delta L\|_F^2 \leq C\left\{\frac{r\lambda_1^2}{p^2\sigma_-^4} + (1+æ)a\Phi_L\right\}. \end{aligned} \tag{13}$$

Proof. See Section 4.1. □

We now give deterministic upper bounds on $\mathbb{E}\left[\|\Sigma_R\|\right]$ and $\mathbb{E}\left[\|\Sigma_R\|_\infty\right]$, and probabilistic upper bounds on $\left\|\nabla\mathcal{L}(X^0;Y,\Omega)\right\|$ and $\left\|\nabla\mathcal{L}(X^0;Y,\Omega)\right\|_\infty$. We will use them to select values of λ_1 and λ_2 which satisfy the assumptions of Theorem 1, and compute the corresponding upper bounds.

Lemma 2. *Let assumption **H** 5 hold. Then, there exists an absolute constant C^* such that the two following inequalities hold*

$$\mathbb{E}\left[\|\Sigma_R\|_\infty\right] \leq 1, \ \textit{and}$$

$$\mathbb{E}\left[\|\Sigma_R\|\right] \leq C^*\left\{\sqrt{\beta} + \sqrt{\log m}\right\}.$$

Proof. See Appendix D □

Lemma 3. *Let assumptions **H** 3-5 hold. Then, there exists an absolute constant c^* such the following two inequalities hold with probability at least $1 - d^{-1}$.*

$$\left\|\nabla\mathcal{L}(X^0;Y,\Omega)\right\|_\infty \leq 6\max\left\{\sigma_+\sqrt{\log d}, \frac{\log d}{\gamma}\right\}, \tag{14}$$

$$\left\|\nabla\mathcal{L}(X^0;Y,\Omega)\right\| \leq c^*\max\left\{\sigma_+\sqrt{\beta\log d}, \frac{\log d}{\gamma}\log\left(\frac{1}{\sigma_-}\sqrt{\frac{m_1 m_2}{\beta}}\right)\right\}, \tag{15}$$

*where $d = m_1 + m_2$, σ_+ and γ are defined in **H** 4, and β in **H** 5.*

Proof. See Appendix E. □

From Theorem 1, Lemma 2 and 3 combined with a union bound argument, we can deduce the following result. Define

$$\Phi'_\alpha = \frac{\left\|\alpha^0\right\|_1}{p}\left\{\frac{\log d}{\sigma_-^2\gamma} + u\right\} + \left(\frac{a}{p}\right)^2 \log d, \tag{16}$$

$$\begin{aligned}\Phi'_L = \Phi'_\alpha &+ \frac{r}{p^2}\left(\beta \vee \log m\right) \\ &+ \left\|\alpha^0\right\|_1 u \left\{\frac{\sigma_+ a}{p\sqrt{M \log d}} + \frac{1}{\gamma p(1+\ae)\sigma_+\sqrt{M\log d}} + 1\right\}. \end{aligned}\tag{17}$$

Let $m_1 + m_2 \geq \exp(\sigma_+^2/\gamma^2)$ and $M \geq \frac{4\sigma_+^2}{\gamma^6}\log\left(\frac{\sqrt{m}}{p\gamma\sigma_-}\right)^2$.

Theorem 2. *Assume **H** 1-5 hold. Let*

$$\lambda_1 = 2c^*\sigma_+\sqrt{\beta\log d}, \quad \lambda_2 = 12u\left(\frac{\log d}{\gamma} + 2\sigma_+^2(1+\ae)a\right),$$

where c_* is the absolute constant defined in Lemma 3. Then, with probability at least $1 - 9d^{-1}$,

$$\left\|\mathsf{f}_U(\Delta\alpha)\right\|_F^2 \leq C\Phi'_\alpha, \ \left\|\Delta\alpha\right\|_2^2 \leq \kappa^{-2}C\Phi'_\alpha \ \textit{and}$$

$$\left\|\Delta L\right\|_F^2 \leq C\Bigg\{\left(\frac{\sigma_+}{\sigma_-}\right)^2\frac{r\beta\log d}{p^2\sigma_-^2} + \frac{u\left\|\alpha^0\right\|_1}{p\sigma_-^2}\left(\frac{\log d}{\gamma}\right) + (1+\ae)a\Phi'_L\Bigg\}.$$

Denoting by $\lesssim$ the inequality up to constant and logarithmic factors and using we have $\Phi'_\alpha \lesssim su/p$ and $\Phi'_L \lesssim r\beta/p^2 + su/p$. The order of magnitude of the bounds are therefore:

$$\begin{aligned} \left\|\mathsf{f}_U(\Delta\alpha)\right\|_F^2 &\lesssim \frac{su}{p}, \\ \left\|\Delta\alpha\right\|_2^2 &\lesssim \frac{su}{p\kappa^2}, \\ \left\|\Delta L\right\|_F^2 &\lesssim \frac{r\beta}{p^2} + \frac{su}{p}. \end{aligned}$$

In the case of almost uniform sampling, *i.e.* $c_1 p \leq \pi_{ij} \leq c_2 p$ for all $(i,j) \in [\![m_1]\!] \times [\![m_2]\!]$ and two positive constants c_1 and c_2, we obtain that $\beta \leq c_2 M p$, which yields the following simplified bound:

$$\left\|\Delta L\right\|_F^2 \lesssim \frac{rM}{p} + \frac{su}{p}.$$

Lower bounds We now turn to the derivation of lower bounds on the Frobenius estimation error. We will need the two following assumptions.

H 6. *For all $k \in [\![N]\!]$, $\sum_{l \neq k} |\langle U_k, U_l, |\rangle \leq \tau$.*

Assumption **H** 6 implies that the U_k are almost orthogonal.

H 7. *The sampling of entries is uniform, i.e. for all $(i,j) \in [\![m_1]\!] \times [\![m_2]\!]$, $\pi_{ij} = p$.*

Without loss of generality we assume $m_1 = m_1 \vee m_2 = M$. For all $X \in \mathbb{R}^{m_1 \times m_2}$ we denote $\mathbb{P}_X$ the product distribution of (Y, Ω) satisfying (1) and (3). Define the following set for two integers $s \leq m_1 m_2/2$ and $r \leq (m_1 \wedge m_2)/2$

$$\mathcal{F}(r,s) = \bigcup_{|\mathcal{I}| \leq s} \{(L, \alpha) \in \mathcal{E}_{a,\mathcal{I}}; \operatorname{rank}(L) \leq r\}\,. \tag{18}$$

Theorem 3. *Assume **H** 1-4 and consider two integers $s \leq (m_1 \wedge m_2)/2$ and $r \leq (m_1 \wedge m_2) \min(1/2, p)$ Then, there exists a constant $\delta > 0$ such that*

$$\inf_{\hat{L}, \hat{\alpha}} \sup_{(L^0, \alpha^0) \in \mathcal{F}} \mathbb{P}_{X^0}\left(\left\| L^0 - \hat{L} \right\|_F^2 + \left\| \alpha^0 - \hat{\alpha} \right\|_2^2 > \Psi_1 \frac{rM}{p} + \Psi_2 s \right) \geq \delta, \tag{19}$$

$$\begin{aligned} \Psi_1 &= C \min\left(\sigma_+^{-2}, \min(a, \sigma_+)^2\right), \\ \Psi_2 &= C \left(\frac{1}{\sigma_+^2 \left(\max_k \|U^k\|_F^2 + 2\tau\right)} \wedge (a \wedge \sigma_+)^2 \right). \end{aligned} \tag{20}$$

Proof. See Section 4.2. □

3 Examples

Multilevel model [**Timmerman, 2006**] In this model matrix $Y \in \mathbb{R}^{m_1 \times m_2}$ gathers m_2 measurements made on m_1 individuals, which we assume to have a multilevel structure. In other words, individuals are divided into H groups, such as different schools or hospitals, and the h-th group contains N_h individuals, $h \in [\![H]\!]$. There are $N = H \times m_2$ dictionary matrices defined as follows for all

$(h, j) \in [\![H]\!] \times [\![m_2]\!]$

$$U_{h,j} = \begin{array}{cc} \begin{array}{cccc} & & j & \end{array} & \\ \left(\begin{array}{cccc} 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 \\ \hline \vdots & \vdots & \vdots & \vdots \\ \hline 0 & 0 & 1 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 1 & 0 \\ \hline \vdots & \vdots & \vdots & \vdots \\ \hline 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 \end{array}\right) & \begin{array}{l} \\ \updownarrow N_1 \\ \\ \vdots \\ \\ \updownarrow N_h \\ \\ \vdots \\ \\ \updownarrow N_N \\ \\ \end{array} \end{array} \tag{21}$$

with 0 everywhere except on the $j-th$ column where there are 1 in the entries corresponding to individuals of group h, and $\mathcal{U} = (U_{h,j})_{h,j}$. By estimating α^0 we therefore estimate the "effect" of the groups on the variables. As detailed in the introduction, this example is useful in medical applications where patients are grouped in different hospitals, and hospitals potentially influence the mean value of the variables. In this case Assumption **H** 3 states that all variables have bounded means; **H** 1 is satisfied with constant æ $= 1$ and **H** 2 with constant $\kappa^2 = \min_h(N_h)$. Finally **H** 4 gives the sub-exponentiality of the cells Y_{ij} which is satisfied for a number of distributions which are often used to model medical data, such as the Gaussian, binomial and Poisson distributions. We emphasize that this model is different from the so-called column-wise sparse corruptions model described in Xu et al. [2010] and Klopp et al. [2017], where for s non-zero columns there are sM parameters; on the contrary here the entries of the non-zero columns take the same value therefore s non-zero columns yield s parameters.

Log-linear model [**Agresti, 2013**] Another classical model is the following log-linear model for count data analysis. $Y \in \mathbb{R}^{m_1 \times m_2}$ is an observation matrix of counts with independent cells, and the parameter matrix X^0, satisfying $\mathbb{E}[Y_{ij}] = \exp(X^0_{ij})$ for all $(i, j) \in [\![m_1]\!] \times [\![m_2]\!]$ is assumed to be decomposed as follows:

$$X^0_{ij} = \alpha^0_i + \beta^0_j + L^0_{ij},$$

where $\alpha^0 \in \mathbb{R}^{m_1}$, $\beta^0 \in \mathbb{R}^{m_2}$, and $L^0 \in \mathbb{R}^{m_1 \times m_2}$ has a low-rank structure. This model is often used in ecological applications where Y counts, for example, the abundance of m_1 species across m_2 environments. It can be re-written in our framework as

$$X^0 = \sum_{k=1}^{m_1+m_2} \alpha^0_k U_k + L^0,$$

with $N = m_1 + m_2$ and where for $k \in [\![m_1]\!]$ and $l \in [\![m_2]\!]$ we have

$$\begin{aligned} U_l &= \begin{array}{c} \begin{array}{cccc} & & l & \end{array} \\ \begin{pmatrix} 0 & 0 & 1 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 1 & 0 \end{pmatrix} \end{array}, \\ U_{m_2+k} &= \begin{array}{cc} \begin{pmatrix} 0 & \dots & 0 & 0 \\ 0 & \dots & 0 & 0 \\ 1 & \dots & 1 & 1 \\ 0 & \dots & 0 & 0 \end{pmatrix} & \begin{array}{c} \\ \\ k \\ \end{array} \end{array} . \end{aligned} \tag{22}$$

In this case Assumption **H** 3 states that all variables have bounded means; **H** 1 is satisfied with constant $æ = 2$ and **H** 2 with constant $\kappa^2 = m$. Finally **H** 4 gives the sub-exponentiality of the cells Y_{ij}, which is the case for the multinomial and Poisson distributions classically used in such models. Again, this model is different from the so-called column-wise sparse corruptions model.

Robust matrix completion Our framework also embeds the well-known robust matrix completion problem [Hsu et al., 2011, Candès et al., 2011, Klopp et al., 2017] which is of interest in recommendation systems for instance. In robust matrix completion we observe noisy and incomplete realizations of a low-rank matrix L^0 perturbed by corruptions denoted $\sum_{(i,j)\in\mathcal{I}} \alpha_k^0 U_{i,j}$, where the $U_{i,j}$, $(i,j) \in [\![m_1]\!] \times [\![m_2]\!]$, are the matrices of the canonical basis of $\mathbb{R}^{m_1 \times m_2}$, and $\mathcal{I}$ indicates the corrupted entries; i.e for all $(i,j) \in [\![m_1]\!] \times [\![m_2]\!]$

$$U_{i,j} = \begin{array}{cc} \begin{array}{cccc} & & j & \end{array} & \\ \begin{pmatrix} 0 & \dots & 0 & 0 \\ 0 & \dots & 0 & 0 \\ 0 & \dots & 1 & 0 \\ 0 & \dots & 0 & 0 \end{pmatrix} & \begin{array}{c} \\ \\ i \\ \end{array} \end{array} . \tag{23}$$

In recommendation applications, α^0 contains corruptions introduced by malicious users to manipulate the system. In this case we obtain $æ = 1$ and $\kappa^2 = 1$.

Replacing u, κ, Ψ_1 and Ψ_2 by their values in the three considered examples, we can specialize Theorem 2 and Theorem 3 to these particular cases. We denote for the multilevel model n_{max} and n_{min} the effectives of the largest and smallest group respectively.

Comparing Figure 1 and Figure 2 we see that the convergence rates obtained in Section 2 are minimax optimal whenever $s < r$. In the case of robust matrix completion we recover the minimax rates derived in Klopp et al. [2017].

Model	Multilevel	Log-linear	Robust MC
u	n_{max}	M	1
κ^2	n_{min}	m	1
$\lVert A(\Delta\alpha)\rVert_F^2$	sn_{max}/p	sM/p	s/p
$\lVert \Delta\alpha\rVert_2^2$	$sn_{\mathsf{max}}/(pn_{\mathsf{min}})$	$sM/(pm)$	s/p
$\lVert \Delta L\rVert_F^2$	$rM/p + sn_{\mathsf{max}}/p$	$rM/p + sM/p$	$rM/p + s/p$

Figure 1: Upper bounds for the multilevel and log-linear model and for robust matrix completion.

Model	Multilevel	Log-linear	Robust MC
u	n_{max}	M	1
$\max_k \lVert U_k\rVert_F^2$	n_{max}	M	1
$\lVert \Delta L\rVert_F^2 + \lVert \Delta\alpha\rVert_2^2$	$rM/p + s/n_{\mathsf{max}}$	$rM/p + s/M$	$rM/p + s$

Figure 2: Lower bounds for the multilevel and log-linear model and for robust matrix completion.

4 Proofs

4.1 Proof of Theorem 1

We first derive an upper bound on the Frobenius error restricted to the observed entries $\lVert \Delta X\rVert_\Omega^2$, then show that the expected Frobenius error $\lVert \Delta X\rVert_\Pi^2$ is upper bounded by $\lVert \Delta X\rVert_\Omega^2$ with high probability and up to a residual term defined later on. These two steps combined yield the upper bound on $\lVert \Delta X\rVert_\Pi^2$.

We now derive the upper bound on $\lVert \Delta X\rVert_\Omega^2$. By definition of $\hat{L}$ and $\hat{\alpha}$:

$$\mathcal{L}(\hat{X};Y,\Omega) - \mathcal{L}(X^0;Y,\Omega) \leq \lambda_1\left(\lVert L^0\rVert_* - \left\lVert \hat{L}\right\rVert_*\right) + \lambda_2\left(\lVert \alpha^0\rVert_1 - \lVert\hat{\alpha}\rVert_1\right).$$

Recall that for $\alpha \in \mathbb{R}^N$ we use the notation $\mathsf{f}_U(\alpha) = \sum_{k=1}^N \alpha_k U^k$. Adding $\langle \nabla\mathcal{L}(X;Y,\Omega), \Delta X\rangle$ on both sides of the last inequality, we get

$$\begin{aligned}\mathcal{L}(\hat{X};Y,\Omega) - \mathcal{L}(X^0;Y,\Omega) + \langle \nabla\mathcal{L}(X^0;Y,\Omega), \Delta X\rangle \leq \\ \lambda_1\left(\lVert L^0\rVert_* - \left\lVert \hat{L}\right\rVert_*\right) - \langle \nabla\mathcal{L}(X^0;Y,\Omega), \Delta L\rangle \\ + \lambda_2\left(\lVert \alpha^0\rVert_1 - \lVert\hat{\alpha}\rVert_1\right) - \langle \nabla\mathcal{L}(X^0;Y,\Omega), A(\Delta\alpha)\rangle.\end{aligned}$$

The strong convexity of the link functions g_j, $j \in \llbracket m_2 \rrbracket$, allows us to lower bound the left hand side term and obtain

$$\frac{\sigma_{\text{MIN}}^2}{2}\lVert \Delta X\rVert_\Omega^2 \leq \mathsf{A}_1 + \mathsf{A}_2;$$

$$\begin{aligned} \mathsf{A}_1 &= \lambda_1 \left(\left\|L^0\right\|_* - \left\|\hat{L}\right\|_* \right) + \left|\langle \nabla\mathcal{L}(X^0;Y,\Omega), \Delta L\rangle\right|, \\ \mathsf{A}_2 &= \lambda_2 \left(\left\|\alpha^0\right\|_1 - \|\hat{\alpha}\|_1 \right) + \left|\langle \nabla\mathcal{L}(X^0;Y,\Omega), A(\Delta\alpha)\rangle\right|. \end{aligned} \tag{24}$$

Let us upper bound A_1. The duality of the norms $\|\cdot\|_*$ and $\|\cdot\|$ implies that

$$\left|\langle \nabla\mathcal{L}(X^0;Y,\Omega), \Delta L\rangle\right| \le \left\|\nabla\mathcal{L}(X^0;Y,\Omega)\right\| \|\Delta L\|_*.$$

We need to introduce some additional notations. Denote by S_1 and S_2 the linear subspaces spanned respectively by the left and right singular vectors of L^0. Denote also by $P_{S_1^\perp}$ and $P_{S_2^\perp}$ the orthogonal projectors on the orthogonal of S_1 and S_2, $P_{L^{0\perp}} : X \mapsto P_{S_1^\perp} X P_{S_2^\perp}$ and $P_{L^0} : X \mapsto X - P_{S_1^\perp} X P_{S_2^\perp}$. The triangular inequality yields

$$\left\|\hat{L}\right\|_* = \left\|L^0 - P_{L^{0\perp}}(\Delta L) - P_{L^0}(\Delta L)\right\|_* \ge \left\|L^0 + P_{L^{0\perp}}(\Delta L)\right\|_* - \|P_{L^0}(\Delta L)\|_*. \tag{25}$$

Moreover, by definition of $P_{L^{0\perp}}$, the left and right singular vectors of $P_{L^{0\perp}}(\Delta L)$ are respectively orthogonal to the left and right singular spaces of L^0, implying $\left\|L^0 + P_{L^{0\perp}}(\Delta L)\right\|_* = \left\|L^0\right\|_* + \|P_{L^{0\perp}}(\Delta L)\|_*$. Plugging this identity into (25) we obtain

$$\left\|L^0\right\|_* - \left\|\hat{L}\right\|_* \le \|P_{L^0}(\Delta L)\|_* - \|P_{L^{0\perp}}(\Delta L)\|_*, \tag{26}$$

and

$$\mathsf{A}_1 \le \lambda_1 \left(\|P_{L^0}(\Delta L)\|_* - \|P_{L^{0\perp}}(\Delta L)\|_* \right) + \left\|\nabla\mathcal{L}(X^0;Y,\Omega)\right\| \|\Delta L\|_*.$$

Using $\|\Delta L\|_* \le \|P_{L^0}(\Delta L)\|_* + \|P_{L^{0\perp}}(\Delta L)\|_*$, the assumption $\lambda_1 \ge 2\left\|\nabla\mathcal{L}(X^0;Y,\Omega)\right\|$ we get

$$\mathsf{A}_1 \le \frac{3\lambda_1}{2} \|P_{L^0}(\Delta L)\|_*.$$

In addition, $\|P_{L^0}(\Delta L)\|_* \le \sqrt{\operatorname{rank}(P_{L^0}(\Delta L))} \|P_{L^0}(\Delta L)\|_F$, and Lemma 7 together with $\|P_{L^0}(\Delta L)\|_F \le \|\Delta L\|_F$ finally imply the following upper bound:

$$\mathsf{A}_1 \le \frac{3\lambda_1}{2} \sqrt{2r} \|\Delta L\|_F. \tag{27}$$

We now derive an upper bound for A_2. The duality between $\|\cdot\|_1$ and $\|\cdot\|_\infty$ ensures

$$\begin{aligned} \left|\langle \nabla\mathcal{L}(X^0;Y,\Omega), A(\Delta\alpha)\rangle\right| &\le \|\Delta\alpha\|_1 \max_k \left|\langle \nabla\mathcal{L}(X^0;Y,\Omega), U^k\rangle\right| \\ &\le \|\Delta\alpha\|_1 \left\|\nabla\mathcal{L}(X^0;Y,\Omega)\right\|_\infty u. \end{aligned} \tag{28}$$

The assumption $\lambda_2 \ge 2\left\|\nabla\mathcal{L}(X^0;Y,\Omega)\right\|_\infty u$ in conjunction with (28) and the triangular inequality $\|\Delta\alpha\|_1 \le \left\|\alpha^0\right\|_1 + \|\hat{\alpha}\|_1$ yield

$$\mathsf{A}_2 \le \frac{3\lambda_2}{2} \left\|\alpha^0\right\|_1. \tag{29}$$

Combining inequalities (24), (27) and (29) we obtain

$$\|\Delta X\|_{\Omega}^2 \leq \frac{3\lambda_1}{\sigma_-^2}\sqrt{2r}\,\|\Delta L\|_F + \frac{3\lambda_2}{\sigma_{\text{MIN}}^2}\left\|\alpha^0\right\|_1. \tag{30}$$

We now show that when the errors ΔL and $\Delta\alpha$ belong to a subspace $\mathcal{C}$ and for a residual D - both defined later on - the following holds with high probability:

$$\|\Delta X\|_{\Omega}^2 \geq \|\Delta X\|_{\Pi}^2 - \mathsf{D}. \tag{31}$$

The proof will follow the subsequent two steps: we start by defining our constrained set and prove that it contains the errors ΔL and $\Delta\alpha$ with high probability (Lemma 4-5); then we show that restricted strong convexity holds on this subspace (Lemma 6).

For non-negative constants d_1, d_Π, $\rho < m$ and ε that will be specified later on, define the two following sets where $\Delta\alpha$ and ΔL should lie:

$$\mathcal{A}(d_1, d_\Pi) = \left\{\alpha \in \mathbb{R}^N :\ \|\alpha\|_1 \leq d_1,\ \|\mathsf{f}_U(\alpha)\|_\Pi^2 \leq d_\Pi\right\}. \tag{32}$$

The constants d_1 and d_Π define the constraints on the ℓ_1 norm of α and weighted Frobenius norm of $\mathsf{f}_U(\alpha)$.

$$\begin{aligned}\mathcal{L}(\rho,\varepsilon) \quad &= \Bigg\{L \in \mathbb{R}^{m_1\times m_2}, \alpha \in \mathbb{R}^N : \|L + \mathsf{f}_U(\alpha)\|_\Pi^2 \geq \frac{72\log(d)}{p\log(6/5)}, \\ &\qquad\qquad \|L + \mathsf{f}_U(\alpha)\|_\infty \leq 1, \|L\|_* \leq \sqrt{\rho}\,\|L\|_F + \varepsilon\Bigg\}\end{aligned} \tag{33}$$

If $\|L + \mathsf{f}_U(\alpha)\|_\Pi^2$ is too small, the right hand side of (31) is negative. The first inequality in the definition of $\mathcal{L}(\rho,\varepsilon)$ prevents from this. Condition $\|L\|_* \leq \sqrt{\rho}\,\|L\|_F + \varepsilon$ is a relaxed form of the condition $\|L\|_* \leq \sqrt{\rho}\,\|L\|_F$ satisfied for matrices of rank ρ. Finally, we define the constrained set of interest:

$$\mathcal{C}(d_1, d_\Pi, \rho, \varepsilon) = \mathcal{L}(\rho,\varepsilon) \cap \left\{\mathbb{R}^{m_1\times m_2} \times \mathcal{A}(d_1, d_\Pi)\right\}.$$

Recall $u = \max_k \|U_k\|_1$ and let

$$\begin{aligned} d_1 \quad &= 4\left\|\alpha^0\right\|_1, \\ d_\Pi \quad &= \frac{3\lambda_2}{\sigma_-^2}\left\|\alpha^0\right\|_1 + 64a^2u\mathbb{E}\left[\|\Sigma_R\|_\infty\right]\left\|\alpha^0\right\|_1 + 3072a^2p^{-1} + \frac{72a^2\log(d)}{\log(6/5)}. \end{aligned}$$

The following Lemma states that with high probability, $\Delta\alpha \in \mathcal{A}(d_1, d_\Pi)$.

Lemma 4. *Let* $\lambda_2 \geq 2u\left(\left\|\nabla\mathcal{L}(X^0; Y, \Omega)\right\|_\infty + 2\sigma_+^2(1+u)a\right)$ *and assume* ***H*** *1-5 hold. Then, with probability at least* $1 - 8d^{-1}$*,*

$$\Delta\alpha \in \mathcal{A}(d_1, d_\Pi);$$

Proof. See Appendix A. □

Lemma 4 implies (i) of Theorem 1. Thus, we only need to prove (ii).

Lemma 5. *Let*

$$\lambda_1 \geq 2\left\|\nabla\mathcal{L}(X^0;Y,\Omega)\right\|, \quad \lambda_2 \geq 2u\left(\left\|\nabla\mathcal{L}(X^0;Y,\Omega)\right\|_\infty + 2\sigma_+^2(1+u)a\right),$$

and assumption ***H*** *4 hold. Then, for $\rho = 32r$ and $\varepsilon = 3\lambda_2/\lambda_1\left\|\alpha^0\right\|_1$,*

$$\|\Delta L\|_* \leq \sqrt{\rho}\,\|\Delta L\|_F + \varepsilon.$$

Proof. See Appendix B □

As a consequence, under the conditions on the regularization parameters λ_1 and λ_2 given in Lemma 5 and whenever

$$\|\Delta L + \mathsf{f}_U(\Delta\alpha)\|_\Pi^2 \geq \frac{72\log(d)}{p\log(6/5)},$$

the error terms $(\Delta L, \Delta\alpha)$ belong to the constrained set $\mathcal{C}(d_1, d_\Pi, \rho, \varepsilon)$ with high probability. We therefore consider the two possible cases: $\|\Delta L + \mathsf{f}_U(\Delta\alpha)\|_\Pi^2 < \frac{72\log(d)}{p\log(6/5)}$ and $\|\Delta L + \mathsf{f}_U(\Delta\alpha)\|_\Pi^2 \geq \frac{72\log(d)}{p\log(6/5)}$.

Case 1: Suppose $\|\Delta L + A(\Delta\alpha)\|_\Pi^2 < \frac{72\log(d)}{p\log(6/5)}$. Then, Lemma 4 combined with the fact that $\|M\|_F^2 \leq p^{-1}\|M\|_\Pi^2$ for all M, and the identity $(a+b)^2 \geq a^2/4 - 4b^2$ ensures that

$$\|\Delta L\|_F^2 \leq 4\|\Delta L + A(\Delta\alpha)\|_F^2 + 16\|A(\Delta\alpha)\|_F^2,$$

therefore

$$\|\Delta L\|_F^2 \leq \frac{288a^2\log(d)}{\log(6/5)} + 16\Phi_\alpha,$$

which implies (ii) of Theorem 1.

Case 2: Suppose $\|\Delta L + A(\Delta\alpha)\|_\Pi^2 \geq \frac{72\log(d)}{p\log(6/5)}$. Then, Lemma 4 and 5 yield that with probability at least $1 - 8d^{-1}$,

$$\left(\frac{\Delta L}{2(1+\text{æ})a}, \frac{\Delta\alpha}{2(1+\text{æ})a}\right) \in \mathcal{C}(d_1', d_\Pi', \rho', \varepsilon'), \text{ with}$$

$$\begin{aligned} d_1' &= \frac{d_1}{2(1+\text{æ})a}, & d_\Pi' &= \frac{d_\Pi}{4(1+\text{æ})^2a^2}, \\ \rho' &= \rho, & \varepsilon' &= \frac{\varepsilon}{2(1+\text{æ})a}, \end{aligned}$$

where d_1, d_Π, ρ and ε are defined in Lemma 4 and 5.

We use the following result, proven in Appendix C. Define the set $\tilde{\mathcal{A}}(d_1)$ as follows:

$$\tilde{\mathcal{A}}(d_1) = \left\{ \alpha \in \mathbb{R}^N : \quad \|\alpha\|_\infty \leq 1; \quad \|\alpha\|_1 \leq d_1; \quad \|\mathsf{f}_U(\alpha)\|_\Pi^2 \geq \frac{18\log(d)}{p\log(6/5)} \right\}.$$

Let d_1, d_Π, ρ and ε be positive constants, and

$$\begin{aligned} \mathsf{D}_\alpha \quad &= 8æd_1 u\mathbb{E}\left[\|\Sigma_R\|_\infty\right] + 768p^{-1}, \\ \mathsf{D}_X \quad &= \frac{76\rho}{p}\mathbb{E}\left[\|\Sigma_R\|\right]^2 + 8\varepsilon\mathbb{E}\left[\|\Sigma_R\|\right] + 8d_1u\mathbb{E}\left[\|\Sigma_R\|_\infty\right] + d_\Pi + 768p^{-1}. \end{aligned} \tag{34}$$

Lemma 6. *Assume **H** 5. Then, the following properties hold:*

(i) For any $\alpha \in \tilde{\mathcal{A}}(d_1)$, with probability at least $1 - 8d^{-1}$,

$$\|\mathsf{f}_U(\alpha)\|_\Omega^2 \geq \frac{1}{2}\|\mathsf{f}_U(\alpha)\|_\Pi^2 - \mathsf{D}_\alpha.$$

(ii) For any pair $(L, \alpha) \in \mathcal{C}(d_1, d_\Pi, \rho, \varepsilon)$, with probability at least $1 - 8d^{-1}$

$$\|L + \mathsf{f}_U(\alpha)\|_\Omega^2 \geq \frac{1}{2}\|L + \mathsf{f}_U(\alpha)\|_\Pi^2 - \mathsf{D}_X. \tag{35}$$

Proof. See Appendix C. □

We apply Lemma 6 (ii) to $\left(\frac{\Delta L}{2(1+æ)a}, \frac{\Delta\alpha}{2(1+æ)a}\right)$ which implies that with probability at least $1 - 8d^{-1}$, $\|\Delta X\|_\Pi^2 \leq 2\|\Delta X\|_\Omega^2 + 2(1+æ)a\Phi_L$. Combined with (30) and $\|\Delta X\|_F^2 \leq p^{-1}\|\Delta X\|_\Pi^2$, it implies that

$$\|\Delta X\|_F^2 \leq \frac{6\sqrt{2r}\lambda_1}{p\sigma_-^2}\|\Delta L\|_F + \frac{6\lambda_2}{p\sigma_-^2}\left\|\alpha^0\right\|_1 + 2(1+æ)a\Phi_L.$$

Now using $\|\Delta X\|_F^2 \geq \frac{\|\Delta L\|_F^2}{2} - \|A(\Delta\alpha)\|_F^2$ and $\frac{6\sqrt{2r}\lambda_1}{p\sigma_-^2}\|\Delta L\|_F \leq \frac{\|\Delta L\|_F^2}{4} + \frac{288r\lambda_1^2}{p^2\sigma_-^4}$, we obtain

$$\|\Delta L\|_F^2 \leq \frac{1152r\lambda_1^2}{p^2\sigma_-^4} + \frac{24\lambda_2\left\|\alpha^0\right\|_1}{p\sigma_-^2} + 2(1+æ)a\Phi_L + 4\Phi_\alpha,$$

which gives the result of Theorem 1 (ii).

4.2 Proof of Theorem 3

Define

$$\tilde{\mathcal{L}} = \left\{ \tilde{L} = (l_{ij}) \in \mathbb{R}^{m_1 \times r} : l_{ij} \in \left\{0, \eta\min(a, \sigma_+)\left(\frac{r}{pm}\right)^{1/2}\right\}, \right.$$
$$\left. \forall 1 \leq i \leq m_1, 1 \leq j \leq r \right\},$$

where $0 \leq \eta \leq 1$ will be chosen later, and the associated set of block matrices

$$\mathcal{L} = \left\{ L = (\tilde{L}|\dots|\tilde{L}|O) \in \mathbb{R}^{m_1 \times m_2} : \tilde{L} \in \tilde{\mathcal{L}} \right\},$$

where O denotes the $m_1 \times (m_2 - r\lfloor m_2/r \rfloor)$ zero matrix and $\lfloor x \rfloor$ is the integer part of x. Similarly, we define the set of vectors

$$\mathcal{A} = \left\{ \alpha = (\tilde{O}|\tilde{\alpha}) \in \mathbb{R}^N, \tilde{\alpha}_k \in \{0, \tilde{\eta} \min(a, \sigma_+)\} \; \forall 1 \leq k \leq s \right\},$$

where $\tilde{O} \in \mathbb{R}^{m_2 - s}$ is the null vector. Finally we set

$$\mathcal{X} = \left\{ X = L + \mathsf{f}_U(\alpha) \in \mathbb{R}^{m_1 \times m_2}, \alpha \in \mathcal{A}, L \in \mathcal{L} \right\}.$$

By construction any element of $\mathcal{X}$ as well as the difference between any two elements of $\mathcal{X}$ can be decomposed into a low-rank component L of rank at most r and a sparse component $\mathsf{f}_U(\alpha)$ with at most s non-zero coefficients α_k, $1 \leq k \leq N$. In addition, the entries of any matrix in $\mathcal{X}$ take values in $[0, (1+æ)a]$. Thus, $\mathcal{X} \subset \mathcal{X}^0$, where $\mathcal{X}^0$ is defined in (5).

We first establish a lower bound of the order rM/p. Let $\mathcal{X}_L \subset \mathcal{X}$ be such that for any $X = L + \mathsf{f}_U(\alpha) \in \mathcal{X}$, $\alpha = 0$. The Varshamov Gilbert bound [Tsybakov, 2008, Lemma 2.9] guarantees the existence of a subset $\mathcal{X}_L^0 \subset \mathcal{X}_L$ with cardinality $\mathrm{Card}(\mathcal{X}_L^0) \geq 2^{rM/8} + 1$ containing the zero $m_1 \times m_2$ matrix $\mathbf{0}$ and such that, for any two distinct elements X and X' of $\mathcal{X}_L^0$,

$$\|X - X'\|_F^2 \geq \frac{Mr}{8} \left(\eta^2 \min(a, \sigma_+)^2 \frac{r}{pm} \left\lfloor \frac{m_2}{r} \right\rfloor \right) \geq \frac{\eta^2}{16} \min(a^2, \sigma_+^2) \frac{rM}{p}. \quad (36)$$

For any $X \in \mathcal{X}_L^0$ the Kullback-Leibler divergence $\mathrm{KL}(\mathbb{P}_0, \mathbb{P}_X)$ between $\mathbb{P}_0$ and $\mathbb{P}_X$ satisfies

$$\begin{aligned} \mathrm{KL}(\mathbb{P}_0, \mathbb{P}_X) = \sum_{i,j} \pi_{ij} \left(g_j(X_{ij}) - g_j(0) - g_j'(0) X_{ij} \right) &\leq \frac{\sigma_+^2 \|X\|_\Pi^2}{2} \\ &\leq \frac{\sigma_+^2 \eta^2 \min(a, \sigma_+)^2 Mr}{2}, \quad (37) \end{aligned}$$

where we have used Assumption **H**4. From (37) we deduce that the condition

$$\frac{1}{\mathrm{Card}(\mathcal{X}_L^0) - 1} \sum_{X \in \mathcal{X}_L^0} \mathrm{KL}(\mathbb{P}_0, \mathbb{P}_X) \leq \frac{1}{16} \log(\mathrm{Card}(\mathcal{X}_L^0) - 1) \quad (38)$$

is satisfied by choosing $\tilde{\eta} = \min\left\{1, (8\sigma_+ \min(a, \sigma_+))^{-1}\right\}$. Then, conditions (36) and (37) guarantee that we can apply [Tsybakov, 2008, Theorem 2.5] and we obtain for some constant $\delta > 0$

$$\inf_{\hat{L}, \hat{\alpha}} \sup_{(L^0, \alpha^0) \in \mathcal{E}} \mathbb{P}_{X^0} \left(\|\Delta L\|_F^2 + \|\Delta \alpha\|_2^2 > \frac{C \min\left(\sigma_+^{-2}, \min(a, \sigma_+)^2\right) rM}{p} \right) \geq \delta. \quad (39)$$

We now turn to the proof of a lower bound of the order s. The Varshamov-Gilbert bound (Tsybakov [2008], Lemma 2.9) guarantees that there exists a subset $\mathcal{A}^0 \in \mathcal{A}$ with cardinality $\mathrm{Card}(\mathcal{A}^0) \geq 2^{s/8} + 1$ containing the null vector $\mathbf{0}$ of $\mathbb{R}^N$ and such that, for any two distinct elements α and α of $\mathcal{A}^0$,

$$\|\alpha - \alpha'\|_2^2 \geq \frac{s}{8}\tilde{\eta}^2 \min(a, \sigma_+)^2. \tag{40}$$

Define $\mathcal{X}_A \subset \mathcal{X}$ the set of matrices $X = \mathsf{f}_U(\alpha)$ such that $\alpha \in \mathcal{A}^0$ and $L = 0$. For any $X \in \mathcal{X}_A$ the Kullback-Leibler divergence $\mathrm{KL}(\mathbb{P}_0, \mathbb{P}_X)$ between $\mathbb{P}_0$ and $\mathbb{P}_X$ satisfies

$$\begin{aligned} \mathrm{KL}(\mathbb{P}_0, \mathbb{P}_X) &= \sum_{i,j} \pi_{ij}(g_j(X_{ij}) - g_j(0) - g_j'(0)X_{ij} \\ &\leq \sigma_+^2 \|\mathsf{f}_U(\alpha)\|_\Pi^2 \leq \sigma_+^2 p \|\mathsf{f}_U(\alpha)\|_F^2 . \end{aligned} \tag{41}$$

The U^k are almost orthogonal by Assumption **H**6 and we obtain, using Assumption **H**4

$$\begin{aligned} \mathrm{KL}(\mathbb{P}_0, \mathbb{P}_X) &\leq \sigma_+^2 p \max_k \left\|U^k\right\|_F^2 \|\alpha\|_2^2 \\ &\leq s\sigma_+^2 p \left(\max_k \left\|U^k\right\|_F^2 + 2\tau\right) \tilde{\eta}^2 \min(a, \sigma_+)^2. \end{aligned} \tag{42}$$

From (42) we deduce that

$$\frac{1}{\mathrm{Card}(\mathcal{A}^0) - 1} \sum_{\mathcal{A}^0} \mathrm{KL}(\mathbb{P}_0, \mathbb{P}_X) \leq s \left(\max_k \left\|U^k\right\|_F^2 + 2\tau\right) \sigma_+^2 \tilde{\eta}^2 \min(a, \sigma_+)^2. \tag{43}$$

Choosing $\tilde{\eta} = \min\left\{1, \left(\sigma_+ \max_k(\left\|U^k\right\|_F + 2\tau) \min(a, \sigma_+)\right)^{-1}\right\}$, we now use Tsybakov [2008], Theorem 2.5 which implies for some constant $\delta > 0$

$$\inf_{\hat{L}, \hat{\alpha}} \sup_{(L^0, \alpha^0) \in \mathcal{E}} \mathbb{P}_{X^0} \left\{\|\Delta L\|_F^2 + \|\Delta\alpha\|_F^2 > \Psi_2 s\right\} \geq \delta, \tag{44}$$

$$\Psi_2 = C \left(\frac{1}{\sigma_+^2 \left(\max_k \|U^k\|_F^2 + 2\tau\right)} \wedge (a \wedge \sigma_+)^2 \right).$$

We finally obtain the result by combining (39) and (44).

References

Alekh Agarwal, Sahand Negahban, and Martin J. Wainwright. Noisy matrix decomposition via convex relaxation: Optimal rates in high dimensions. *Ann. Statist.*, 40(2):1171–1197, 04 2012. doi: 10.1214/12-AOS1000. URL https://doi.org/10.1214/12-AOS1000.

Alan Agresti. *Categorical Data Analysis, 3rd Edition*. Wiley, 2013.

Pierre Alquier, Vincent Cottet, and Guillaume Lecué. Estimation bounds and sharp oracle inequalities of regularized procedures with Lipschitz loss functions. *ArXiv e-prints*, February 2017.

Afonso S. Bandeira and Ramon van Handel. Sharp nonasymptotic bounds on the norm of random matrices with independent entries. *Ann. Probab.*, 44(4): 2479–2506, 07 2016. doi: 10.1214/15-AOP1025. URL https://doi.org/10.1214/15-AOP1025.

Anthony S Bryk and Stephen W Raudenbush. *Hierarchical linear models : applications and data analysis methods*. Newbury Park : Sage Publications, 2nd ed edition, 1992.

Emmanuel J. Candès, Xiaodong Li, Yi Ma, and John Wright. Robust principal component analysis? *J. ACM*, 58(3):11:1–11:37, June 2011. ISSN 0004-5411. doi: 10.1145/1970392.1970395. URL http://doi.acm.org/10.1145/1970392.1970395.

Venkat Chandrasekaran, Sujay Sanghavi, Pablo A. Parrilo, and Alan S. Willsky. Rank-sparsity incoherence for matrix decomposition. *SIAM Journal on Optimization*, 21(2):572–596, 2011. doi: 10.1137/090761793. URL https://doi.org/10.1137/090761793.

Sourav Chatterjee. Matrix estimation by universal singular value thresholding. *Ann. Statist.*, 43(1):177–214, 02 2015. doi: 10.1214/14-AOS1272. URL https://doi.org/10.1214/14-AOS1272.

Michael Collins, Sanjoy Dasgupta, and Robert E. Schapire. A generalization of principal component analysis to the exponential family. In *Advances in Neural Information Processing Systems*. MIT Press, 2001.

Andreas Elsener and Sara van de Geer. Robust Low-Rank Matrix Estimation. *ArXiv e-prints*, March 2016.

William Fithian and Rahul Mazumder. Flexible Low-Rank Statistical Modeling with Side Information. *ArXiv e-prints*, August 2013.

Geoffrey J. Gordon. Generalized^2 Linear^2 models. *Advances in Neural Information Processing Systems*, pages 577–584, 2002.

Daniel Hsu, Sham M. Kakade, and Tong Zhang. Robust matrix decomposition with sparse corruptions. *EEE Transactions on Information Theory*, 57(11): 7221–7234, 2011.

Olga Klopp. Noisy low-rank matrix completion with general sampling distribution. *Bernoulli*, 20(1):282–303, 2014.

Olga Klopp. Matrix completion by singular value thresholding: sharp bounds. *Electronic journal of statistics* , 9(2):2348–2369, 2015. URL https://hal.archives-ouvertes.fr/hal-01111757.

Olga Klopp, Karim Lounici, and Alexandre B. Tsybakov. Robust matrix completion. *Probability Theory and Related Fields*, 169(1):523–564, Oct 2017. doi: 10.1007/s00440-016-0736-y. URL https://doi.org/10.1007/s00440-016-0736-y.

Vladimir Koltchinskii. *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery*. Springer, 2011.

Michel Ledoux. *The concentration of measure phenomenon*, volume 89 of *Mathematical Surveyx and Monographs*. American Mathematical Society, Providence, 2001.

Jérôme Pagès. *Multiple factor analysis by example using R*. Chapman & Hall/CRC the R series (CRC Press). Taylor & Francis Group, 2015.

Nathan Srebro. *Learning with Matrix Factorizations*. PhD thesis, Massachusetts Institute of Technology, 2004.

Michel Talagrand. A new look at independence. *Ann. Probab.*, 24(1):1–34, 01 1996. doi: 10.1214/aop/1042644705. URL https://doi.org/10.1214/aop/1042644705.

Marieke E. Timmerman. Multilevel component analysis. *British Journal of Mathematical and Statistical Psychology*, 59(2):301–320, 2006. ISSN 2044-8317. doi: 10.1348/000711005X67599. URL http://dx.doi.org/10.1348/000711005X67599.

Joel A. Tropp. *Topics in Sparse Approximation*. PhD thesis, The University of Texas at Austin, 2004.

Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Publishing Company, Incorporated, 1st edition, 2008. ISBN 0387790519, 9780387790510.

Madeleine Udell, Corinne Horn, Reza Zadeh, and Stephen Boyd. Generalized low rank models. *Foundations and Trends in Machine Learning*, 9(1), 2016. ISSN 1935-8237. doi: 10.1561/2200000055. URL http://dx.doi.org/10.1561/2200000055.

Roman Vershynin. *Compressed Sensing, Theory and Applications*. Cambridge University Press, 2012.

Huan Xu, Constantine Caramanis, and Sujay Sanghavi. Robust pca via outlier pursuit. In *Proceedings of the 23rd International Conference on Neural Information Processing Systems*, NIPS'10, pages 2496–2504, USA, 2010. Curran Associates Inc. URL http://dl.acm.org/citation.cfm?id=2997046.2997174.

A Proof of Lemma 4

We start by proving $\|\Delta\alpha\|_1 \le 4\left\|\alpha^0\right\|_1$. By the standard optimality conditions over a convex set, there exist two subgradients $\hat{f}_L$ in the subdifferential of $\|\cdot\|_*$ taken at $\hat{L}$ and $\hat{f}_\alpha$ in the subdifferential of $\|\cdot\|_1$ taken at $\hat{\alpha}$, such that for all feasible pairs (L,α) we have

$$\langle \nabla\mathcal{L}(\hat{X};Y,\Omega), L-\hat{L}+\sum_{k=1}^{N}(\alpha_k-\hat{\alpha}_k)U^k\rangle+\lambda_1\langle\hat{f}_L, L-\hat{L}\rangle+\lambda_2\langle\hat{f}_\alpha,\alpha-\hat{\alpha}\rangle \ge 0. \quad (45)$$

In particular for the pair $(\hat{L},\alpha^0)$ we obtain

$$\langle \nabla\mathcal{L}(\hat{X};Y,\Omega), \sum_{k=1}^{N}\Delta\alpha_k U^k\rangle + \lambda_2\langle\hat{f}_\alpha,\Delta\alpha\rangle \ge 0.$$

Denote $\tilde{X} = \hat{L}+\sum_{k=1}^{N}\alpha_k^0 U^k$. The last inequality is equivalent to

$$\underbrace{\langle \nabla\mathcal{L}(X^0;Y,\Omega), \sum_{k=1}^{N}\Delta\alpha_k U^k\rangle}_{\mathsf{B}_1} + \underbrace{\langle \nabla\mathcal{L}(\tilde{X};Y,\Omega)-\nabla\mathcal{L}(X^0;Y,\Omega), \sum_{k=1}^{N}\Delta\alpha_k U^k\rangle}_{\mathsf{B}_2}$$
$$+\underbrace{\langle \nabla\mathcal{L}(\hat{X};Y,\Omega)-\nabla\mathcal{L}(\tilde{X};Y,\Omega), \sum_{k=1}^{N}\Delta\alpha_k U^k\rangle}_{\mathsf{B}_3}+\lambda_2\langle\hat{f}_\alpha,\Delta\alpha\rangle \ge 0.$$

We now derive upper bounds on the three terms B_1, B_2 and B_3 separately. Recall that we denote $u=\max_k\left\|U^k\right\|_1$ and use (28) to bound B_1:

$$\mathsf{B}_1 \le \|\Delta\alpha\|_1\left\|\nabla\mathcal{L}(X^0;Y,\Omega)\right\|_\infty u. \quad (46)$$

Similarly, the duality between $\|\cdot\|_\infty$ and $\|\cdot\|_1$ gives

$$\mathsf{B}_2 \le \|\Delta\alpha\|_1\left\|\nabla\mathcal{L}(\tilde{X};Y,\Omega)-\nabla\mathcal{L}(X^0;Y,\Omega)\right\|_\infty u.$$

Moreover, $\nabla\mathcal{L}(\tilde{X};Y,\Omega)-\nabla\mathcal{L}(X^0;Y,\Omega)$ is a matrix with entries $g_j'(\tilde{X}_{ij})-g_j'(X_{ij}^0)$, therefore assumption **H** 4 ensures

$$\left\|\nabla\mathcal{L}(\tilde{X};Y,\Omega)-\nabla\mathcal{L}(X^0;Y,\Omega)\right\|_\infty \le 2\sigma_+^2(1+æ)a,$$

and finally we obtain

$$\mathsf{B}_2 \le \|\Delta\alpha\|_1 2\sigma_+^2(1+æ)au. \quad (47)$$

We finally bound B_3 as follows. We have that

$$\mathsf{B}_3 = \sum_{i=1}^{m_1}\sum_{j=1}^{m_2}\omega_{ij}\left(g_j'(\hat{X}_{ij})-g_j'(\tilde{X}_{ij})\right)\left(\tilde{X}_{ij}-\hat{X}_{ij}\right).$$

Now, for all $j \in [\![m_2]\!]$, g'_j is increasing therefore

$$\left(g'_j(\hat{X}_{ij}) - g'_j(\tilde{X}_{ij})\right)\left(\tilde{X}_{ij} - \hat{X}_{ij}\right) \leq 0,$$

which implies $\mathsf{B}_3 \leq 0$. Combined with (46) and (47) this yields

$$\lambda_2 \langle \hat{f}_\alpha, \hat{\alpha} - \alpha^0 \rangle \leq \|\Delta\alpha\|_1 \, u \left(\left\| \nabla \mathcal{L}(X^0; Y, \Omega) \right\|_\infty + 2\sigma_+^2 (1 + æ) a \right).$$

Besides, the convexity of $\|\cdot\|_1$ gives $\langle \hat{f}_\alpha, \hat{\alpha} - \alpha^0 \rangle \geq \|\hat{\alpha}\|_1 - \left\|\alpha^0\right\|_1$, therefore

$$\begin{aligned}\left\{\lambda_2 - u\left(\left\|\nabla\mathcal{L}(X^0;Y,\Omega)\right\|_\infty + 2\sigma_+^2(1+æ)a\right)\right\}\|\hat{\alpha}\|_1 \leq \\ \left\{\lambda_2 + u\left(\left\|\nabla\mathcal{L}(X^0;Y,\Omega)\right\|_\infty + 2\sigma_+^2(1+æ)a\right)\right\}\left\|\alpha^0\right\|_1,\end{aligned}$$

and the condition $\lambda_2 \geq 2\left\{u\left(\left\|\nabla\mathcal{L}(X^0;Y,\Omega)\right\|_\infty + 2\sigma_+^2(1+æ)a\right)\right\}$ gives $\|\hat{\alpha}\|_1 \leq 3\left\|\alpha^0\right\|_1$ and finally

$$\|\Delta\alpha\|_1 \leq 4\left\|\alpha^0\right\|_1. \tag{48}$$

We consider the two following cases.

Case I: $\|\mathsf{f}_U(\Delta\alpha)\|_\Pi^2 < \frac{72a^2\log(d)}{p\log(6/5)}$. Then the result holds trivially.

Case II: $\|\mathsf{f}_U(\Delta\alpha)\|_\Pi^2 \geq \frac{72a^2\log(d)}{p\log(6/5)}$. For $d_1 > 0$ recall the definition of the set

$$\tilde{\mathcal{A}}(d_1) = \left\{\alpha \in \mathbb{R}^N : \quad \|\alpha\|_\infty \leq 1; \quad \|\alpha\|_1 \leq d_1; \quad \|\mathsf{f}_U(\alpha)\|_\Pi^2 \geq \frac{18\log(d)}{p\log(6/5)}\right\}.$$

Inequality (48) and $\|\Delta\alpha\|_\infty \leq 2a$ imply that

$$\frac{\Delta\alpha}{2a} \in \tilde{\mathcal{A}}\left(\frac{2\left\|\alpha^0\right\|_1}{a}\right).$$

Therefore we can apply Lemma 6(i) and obtain that with probability at least $1 - 8d^{-1}$,

$$\|\mathsf{f}_U(\Delta\alpha)\|_\Pi^2 \leq \|\mathsf{f}_U(\Delta\alpha)\|_\Omega^2 + 64æa\left\|\alpha^0\right\|_1 u\mathbb{E}\left[\|\Sigma_R\|_\infty\right] + 3072a^2p^{-1}. \tag{49}$$

We now must upper bound the quantity $\|\mathsf{f}_U(\Delta\alpha)\|_\Omega^2$. Recall that $\tilde{X} = \sum_{k=1}^N \alpha_k^0 U^k + \hat{X}$. By definition,

$$\mathcal{L}(\hat{X};Y,\Omega) + \lambda_1\left\|\hat{L}\right\|_* + \lambda_2\|\hat{\alpha}\|_1 \leq \mathcal{L}(\tilde{X};Y,\Omega) + \lambda_1\left\|\hat{L}\right\|_* + \lambda_2\left\|\alpha^0\right\|_1,$$

i.e.

$$\mathcal{L}(\hat{X};Y,\Omega) - \mathcal{L}(\tilde{X};Y,\Omega) \leq \lambda_2\left(\left\|\alpha^0\right\|_1 - \|\hat{\alpha}\|_1\right).$$

Substracting $\langle \nabla \mathcal{L}(\tilde{X}; Y, \Omega), \hat{X} - \tilde{X} \rangle$ on both sides and by strong convexity of $\mathcal{L}$ we obtain

$$\begin{aligned} \frac{\sigma_-^2}{2} \left\| \mathsf{f}_U(\Delta\alpha) \right\|_\Omega^2 &\leq \lambda_2 \left(\left\| \alpha^0 \right\|_1 - \left\| \hat{\alpha} \right\|_1 \right) + \langle \nabla \mathcal{L}(\tilde{X}; Y, \Omega), \mathsf{f}_U(\Delta\alpha) \rangle \\ &\leq \lambda_2 \left(\left\| \alpha^0 \right\|_1 - \left\| \hat{\alpha} \right\|_1 \right) + \underbrace{\left| \langle \nabla \mathcal{L}(X^0; Y, \Omega), \mathsf{f}_U(\Delta\alpha) \rangle \right|}_{\mathsf{C}_1} . \\ &+ \underbrace{\left| \langle \nabla \mathcal{L}(X^0; Y, \Omega) - \nabla \mathcal{L}(\tilde{X}; Y), \mathsf{f}_U(\Delta\alpha) \rangle \right|}_{\mathsf{C}_2} \end{aligned} \tag{50}$$

The duality of $\|\cdot\|_1$ and $\|\cdot\|_\infty$ yields $\mathsf{C}_1 \leq \left\| \nabla \mathcal{L}(X^0; Y, \Omega) \right\|_\infty u \left\| \Delta\alpha \right\|_1$, and

$$\mathsf{C}_2 \leq \left\| \nabla \mathcal{L}(X^0; Y, \Omega) - \nabla \mathcal{L}(\tilde{X}; Y, \Omega) \right\|_\infty u \left\| \Delta\alpha \right\|_1 .$$

Furthermore,

$$\left\| \nabla \mathcal{L}(X^0; Y, \Omega) - \nabla \mathcal{L}(\tilde{X}; Y, \Omega) \right\|_\infty \leq 2\sigma_+^2 a,$$

since for all $(i,j) \in [\![m_1]\!] \times [\![m_2]\!]$ $|\tilde{X}_{ij} - X_{ij}^0| \leq 2a$ and $g_j''(\tilde{X}_{ij}) \leq \sigma_+^2$. The last three inequalities plugged in (50) give

$$\frac{\sigma_-^2}{2} \left\| \mathsf{f}_U(\Delta\alpha) \right\|_\Omega^2 \quad \leq \lambda_2 \left(\left\| \alpha^0 \right\|_1 - \left\| \hat{\alpha} \right\|_1 \right) + u \left\| \Delta\alpha \right\|_1 \left\{ \left\| \nabla \mathcal{L}(X^0; Y, \Omega) \right\|_\infty + 2\sigma_+^2 a \right\} .$$

The triangular inequality gives

$$\begin{aligned} \frac{\sigma_-^2}{2} \left\| \mathsf{f}_U(\Delta\alpha) \right\|_\Omega^2 \quad &\leq \left\{ u \left(\left\| \nabla \mathcal{L}(X^0; Y, \Omega) \right\|_\infty + 2\sigma_+^2 a \right) + \lambda_2 \right\} \left\| \alpha^0 \right\|_1 \\ &+ \left\{ u \left(\left\| \nabla \mathcal{L}(X^0; Y, \Omega) \right\|_\infty + 2\sigma_+^2 a \right) - \lambda_2 \right\} \left\| \hat{\alpha} \right\|_1 . \end{aligned}$$

Then, the assumption $\lambda_2 \geq 2u \left(\left\| \nabla \mathcal{L}(X^0; Y, \Omega) \right\|_\infty + 2\sigma_+^2 (1 + æ) a \right)$ gives

$$\left\| \mathsf{f}_U(\Delta\alpha) \right\|_\Omega^2 \leq \frac{3\lambda_2}{\sigma_-^2} \left\| \alpha^0 \right\|_1 .$$

Plugged into (49), this last inequality implies that with probability at least $1 - 8d^{-1}$

$$\left\| \mathsf{f}_U(\Delta\alpha) \right\|_\Pi^2 \leq \frac{3\lambda_2}{\sigma_-^2} \left\| \alpha^0 \right\|_1 + 64æa \left\| \alpha^0 \right\|_1 u\mathbb{E}\left[\left\| \Sigma_R \right\|_\infty \right] + 3072a^2 p^{-1}. \tag{51}$$

Combining (48) and (51) gives the result.

B Proof of Lemma 5

Using (45) for $L = L^0$ and $\alpha = \alpha^0$ we obtain

$$\langle \nabla \mathcal{L}(\hat{X}; Y, \Omega), \Delta L + \sum_{k=1}^{N} (\Delta\alpha_k) U^k \rangle + \lambda_1 \langle \hat{f}_L, \Delta L \rangle + \lambda_2 \langle \hat{f}_\alpha, \Delta\alpha \rangle \geq 0.$$

Then, the convexity of $\|\cdot\|_*$ and $\|\cdot\|_1$ imply that

$$\begin{aligned}\left\|L^0\right\|_* &\geq \left\|\hat{L}\right\|_* + \langle\partial\left\|\hat{L}\right\|_*, \Delta L\rangle,\\ \left\|\alpha^0\right\|_1 &\geq \left\|\hat{\alpha}\right\|_* + \langle\partial\left\|\hat{\alpha}\right\|_1, \Delta\alpha\rangle.\end{aligned}$$

The last three inequalities yield

$$\begin{aligned}\lambda_1\left(\left\|\hat{L}\right\|_* - \left\|L^0\right\|_*\right) + \lambda_2\left(\left\|\hat{\alpha}\right\|_1 - \left\|\alpha^0\right\|_1\right) &\leq \langle\nabla\mathcal{L}(\hat{X}; Y, \Omega), \Delta L\rangle\\ &\quad + \langle\nabla\mathcal{L}(\hat{X}; Y, \Omega), \sum_{k=1}^N(\Delta\alpha_k)U^k\rangle\\ &\leq \left\|\nabla\mathcal{L}(\hat{X}; Y, \Omega)\right\|\left\|\Delta L\right\|_* + u\left\|\nabla\mathcal{L}(\hat{X}; Y, \Omega)\right\|_\infty\left\|\Delta\alpha\right\|_1.\end{aligned}$$

Using (26) and the conditions

$$\lambda_1 \geq 2\left\|\nabla\mathcal{L}(X^0; Y, \Omega)\right\|, \quad \lambda_2 \geq 2u\left\{\left\|\nabla\mathcal{L}(X^0; Y, \Omega)\right\|_\infty + 2\sigma_+^2(1+æ)a\right\},$$

we get

$$\begin{aligned}\lambda_1\left(\left\|P_{L^0}^\perp(\Delta L)\right\|_* - \left\|P_{L^0}(\Delta L)\right\|_*\right) + \lambda_2\left(\left\|\hat{\alpha}\right\|_1 - \left\|\alpha^0\right\|_1\right) \leq&\\ \frac{\lambda_1}{2}\left(\left\|P_{L^0}^\perp(\Delta L)\right\|_* + \left\|P_{L^0}(\Delta L)\right\|_*\right) + \frac{\lambda_2}{2}\left\|\Delta\alpha\right\|_1,&\end{aligned}$$

which implies

$$\left\|P_{L^0}^\perp(\Delta L)\right\|_* \leq 3\left\|P_{L^0}(\Delta L)\right\|_* + 3\lambda_2/\lambda_1\left\|\alpha^0\right\|_1.$$

Now, using

$$\left\|\Delta L\right\|_* \leq \left\|P_{L^0}^\perp(\Delta L)\right\|_* + \left\|P_{L^0}(\Delta L)\right\|_*, \quad \left\|P_{L^0}(\Delta L)\right\|_F \leq \left\|\Delta L\right\|_F$$

and Lemma 7, we get

$$\left\|\Delta L\right\|_* \leq \sqrt{32r}\left\|\Delta L\right\|_F + 3\lambda_2/\lambda_1\left\|\alpha^0\right\|_1.$$

This completes the proof of Lemma 5.

Lemma 7. $\operatorname{rank}(P_{L^0}(\Delta L)) \leq 2r$

Proof.

$$\begin{aligned}P_{L^0}(\Delta L) \quad &= X - P_{S_1^\perp}XP_{S_2^\perp} = (P_{S_1} + P_{S_1^\perp})X(P_{S_2} + P_{S_2^\perp}) - P_{S_1^\perp}XP_{S_2^\perp}\\ &= P_{S_1}XP_{S_2} + P_{S_1}XP_{S_2^\perp} + P_{S_1^\perp}XP_{S_2} + P_{S_1^\perp}XP_{S_2^\perp} - P_{S_1^\perp}XP_{S_2^\perp}\\ &= P_{S_1}X(P_{S_2} + P_{S_2^\perp}) + P_{S_1^\perp}XP_{S_2}\\ &= P_{S_1}X + P_{S_1^\perp}XP_{S_2}.\end{aligned}$$

$P_{L^0}(\Delta L)$ is the sum of two matrices of rank at most $\operatorname{rank}\left(L^0\right) = r$, which proves the lemma. □

C Proof of Lemma 6

Proof of (i): Recall

$$\mathsf{D}_\alpha = 8\text{æ} d_1 u \mathbb{E}\left[\|\Sigma_R\|_\infty\right] + 768 p^{-1}$$

and

$$\tilde{\mathcal{A}}(d_1) = \left\{\alpha \in \mathbb{R}^N : \quad \|\alpha\|_\infty \leq 1; \quad \|\alpha\|_1 \leq d_1; \quad \|\mathsf{f}_U(\alpha)\|_\Pi^2 \geq \frac{18\log(d)}{p\log(6/5)}\right\}.$$

We will show that the probability of the following event is small

$$\mathcal{B} = \left\{\exists \alpha \in \tilde{\mathcal{A}}(d_1) \text{ such that } \left|\|A(\alpha)\|_\Omega^2 - \|A(\alpha)\|_\Pi^2\right| > \frac{1}{2}\|A(\alpha)\|_\Pi^2 + \mathsf{D}_\alpha\right\}.$$

Indeed, $\mathcal{B}$ contains the complement of the event we are interested in. In order to prove that the probability of $\mathcal{B}$ is small, we use a standard peeling argument. Let $\nu = \frac{18\log(d)}{p\log(6/5)}$ and $\eta = 6/5$. For $l \in \mathbb{N}$ set

$$\mathcal{S}_l = \left\{\alpha \in \tilde{\mathcal{A}}(d_1) : \quad \eta^{l-1}\nu \leq \|A(\alpha)\|_\Pi^2 \leq \eta^l \nu\right\}.$$

If the event $\mathcal{B}$ holds, there exists $l \geq 1$ and $\alpha \in \tilde{\mathcal{A}}(d_1) \cap S_l$ such that

$$\begin{aligned} \left|\|A(\alpha)\|_\Omega^2 - \|A(\alpha)\|_\Pi^2\right| &> \frac{1}{2}\|A(\alpha)\|_\Pi^2 + \mathsf{D}_\alpha \\ &> \frac{1}{2}\eta^{l-1}\nu + \mathsf{D}_\alpha \\ &= \frac{5}{12}\eta^l \nu + \mathsf{D}_\alpha. \end{aligned} \tag{52}$$

For $T > \nu$, consider the set of vectors $\tilde{\mathcal{A}}(d_1, T) = \left\{\alpha \in \tilde{\mathcal{A}}(d_1) : \|A(\alpha)\|_\Pi^2 \leq T\right\}$ and the event

$$\mathcal{B}_l = \left\{\exists \alpha \in \tilde{\mathcal{A}}(d_1, \eta^l \nu) : \left|\|A(\alpha)\|_\Omega^2 - \|A(\alpha)\|_\Pi^2\right| > \frac{5}{12}\eta^l \nu + \mathsf{D}_\alpha\right\}.$$

Note that $S_l \subset \tilde{\mathcal{A}}(d_1, \eta^l \nu)$. Then, (52) implies that $\mathcal{B}_l$ holds, therefore $\mathcal{B} \subset \cup_{l=1}^{+\infty} \mathcal{B}_l$, thus it is enough to estimate the probability of the events $\mathcal{B}_l$ and then apply the union bound. Such an estimation is given in the following lemma, adapted from Klopp [2015], Lemma 10.

Lemma 8. *Define* $Z_T = \sup_{\alpha \in \tilde{\mathcal{A}}(d_1, T)} \left|\|A(\alpha)\|_\Omega^2 - \|A(\alpha)\|_\Pi^2\right|$. *Then,*

$$\mathbb{P}\left(Z_T \geq \mathsf{D}_\alpha + \frac{5}{12}T\right) \leq 4\mathrm{e}^{-pT/18}.$$

Proof. By definition,

$$Z_T = \sup_{\alpha\in\tilde{\mathcal{A}}(d_1,T)} \left| \sum_{(i,j)} \omega_{ij} A(\alpha)_{ij}^2 - \mathbb{E}\left[\sum_{(i,j)} \omega_{ij} A(\alpha)_{ij}^2 \right] \right| .$$

We use the following Talagrand's concentration inequality, proved in Chatterjee [2015] and Talagrand [1996].

Lemma 9. *Assume $f : [-1,1]^n \mapsto \mathbb{R}$ is a convex Lipschitz function with Lipschitz constant L. Let $\Xi_1, \dots, \Xi_n$ be independent random variables taking values in $[-1,1]$. Let $Z := f(\Xi_1, \dots, \Xi_n)$. Then, for any $t \geq 0$,*

$$\mathbb{P}\left(|Z - \mathbb{E}[Z]| \geq 16L + t\right) \leq 4\mathrm{e}^{-t^2/2L^2}.$$

Let $f(x_{11}, \dots, x_{m_1m_2}) = \sup_{\alpha\in\tilde{\mathcal{A}}(d_1,T)} \left|\sum_{(i,j)}(x_{ij} - \pi_{ij})A(\alpha)_{ij}^2\right|$. We have that f is Lipschitz with Lipschitz constant $\sqrt{p^{-1}T}$. Indeed, for any $(x_{11}, \dots, x_{m_1m_2}) \in \mathbb{R}^{m_1\times m_2}$ and $(z_{11}, \dots, z_{m_1m_2}) \in \mathbb{R}^{m_1\times m_2}$:

$$\begin{aligned}
&|f(x_{11}, \dots, x_{m_1m_2}) - f(z_{11}, \dots, z_{m_1m_2})| \\
&\quad= \left| \sup_{\alpha\in\tilde{\mathcal{A}}(d_1,T)} \left| \sum_{(i,j)} (x_{ij} - \pi_{ij}) A(\alpha)_{ij}^2 \right| - \sup_{\alpha\in\tilde{\mathcal{A}}(d_1,T)} \left| \sum_{(i,j)} (z_{ij} - \pi_{ij}) A(\alpha)_{ij}^2 \right| \right| \\
&\quad\leq \sup_{\alpha\in\tilde{\mathcal{A}}(d_1,T)} \left| \left| \sum_{(i,j)} (x_{ij} - \pi_{ij}) A(\alpha)_{ij}^2 \right| - \left| \sum_{(i,j)} (z_{ij} - \pi_{ij}) A(\alpha)_{ij}^2 \right| \right| \\
&\quad\leq \sup_{\alpha\in\tilde{\mathcal{A}}(d_1,T)} \left| \sum_{(i,j)} (x_{ij} - \pi_{ij}) A(\alpha)_{ij}^2 - \sum_{(i,j)} (z_{ij} - \pi_{ij}) A(\alpha)_{ij}^2 \right| \\
&\quad\leq \sup_{\alpha\in\tilde{\mathcal{A}}(d_1,T)} \left| \sum_{(i,j)} (x_{ij} - z_{ij}) A(\alpha)_{ij}^2 \right| \\
&\quad\leq \sup_{\alpha\in\tilde{\mathcal{A}}(d_1,T)} \sqrt{\sum_{(i,j)} \pi_{ij}^{-1} (x_{ij} - z_{ij})^2} \sqrt{\sum_{(i,j)} \pi_{ij} A(\alpha)_{ij}^4} \\
&\quad\leq \sup_{\alpha\in\tilde{\mathcal{A}}(d_1,T)} \sqrt{p^{-1}} \sqrt{\sum_{(i,j)} (x_{ij} - z_{ij})^2} \sqrt{\sum_{(i,j)} \pi_{ij} A(\alpha)_{ij}^2} \\
&\quad\leq \sqrt{p^{-1}T} \sqrt{\sum_{(i,j)} (x_{ij} - z_{ij})^2},
\end{aligned}$$

where we used $||a| - |b|| \leq |a-b|, \|A(\alpha)\|_\infty \leq 1$ and $\|A\|_\Pi^2 \leq T$. Now, Talagrand's concentration inequality and $\sqrt{p^{-1}T} \leq \frac{96p^{-1}}{2} + \frac{T}{2\times 96}$ imply

$$\mathbb{P}\left(|Z - \mathbb{E}[Z]| \geq 768p^{-1} + \frac{1}{12}T + t \right) \leq 4e^{-t^2 p/2T}.$$

Taking $t = T/3$ we get

$$\mathbb{P}\left(|Z - \mathbb{E}\left[Z\right]| \geq 768p^{-1} + \frac{5}{12}T\right) \leq 4e^{-pT/18}. \tag{53}$$

Now we must bound the expectation $\mathbb{E}\left[Z_T\right]$. To do so, we use a symmetrization argument [Ledoux, 2001] which gives

$$\begin{aligned} \mathbb{E}\left[Z_T\right] &= \mathbb{E}\left[\sup_{\alpha \in \tilde{\mathcal{A}}(d_1,T)} \left|\sum_{(i,j)} \omega_{ij} A(\alpha)_{ij}^2 - \mathbb{E}\left[\sum_{(i,j)} \omega_{ij} A(\alpha)_{ij}^2\right]\right|\right] \\ &\leq 2\mathbb{E}\left[\sup_{\alpha \in \tilde{\mathcal{A}}(d_1,T)} \left|\sum_{(i,j)} \epsilon_{ij}\omega_{ij} A(\alpha)_{ij}^2\right|\right], \end{aligned}$$

where $\{\epsilon_{ij}\}$ is an i.i.d. Rademacher sequence independent of $\{\omega_{ij}\}$. Talagrand's contraction inequality (see Koltchinskii [2011], Theorem 2.2) states that for $\mathsf{T} \subset \mathbb{R}^{m_1 \times m_2}$, and functions $\phi_{ij} : \mathbb{R} \to \mathbb{R}$, $(i,j) \in [\![m_1]\!] \times [\![m_2]\!]$ such that $\phi_{ij}(0) = 0$ and for all $u, v \in \mathbb{R}$,

$$|\phi_{ij}(u) - \phi_{ij}(v)| \leq |u - v|,$$

that is, ϕ_i are contractions. For all convex non-decreasing functions $\Phi : \mathbb{R}_+ \to \mathbb{R}_+$,

$$\mathbb{E}\left[\Phi\left(\frac{1}{2}\sup_{t \in \mathsf{T}}\left|\sum_{i,j}\epsilon_{ij}\phi_{ij}(t_{ij})\right|\right)\right] \leq \mathbb{E}\left[\Phi\left(\sup_{t \in \mathsf{T}}\left|\sum_{i,j}\epsilon_{ij}t_{ij}\right|\right)\right].$$

We apply the extension of this result to Lipschitz functions, with

$$\mathsf{T} = \left\{A \in \mathbb{R}^{m_1 \times m_2}; \exists \alpha \in \tilde{\mathcal{A}}(d_1, T), A = \sum_{k=1}^{N} \alpha_k U_k\right\},$$

$\Phi : x \mapsto x$, and $\phi_{ij} = \phi : x \mapsto x^2$ for all i, j. For all $A \in \mathsf{T}$, $|A_{ij}| \leq$ æ, and ϕ is Lipschitz with constant 2æ on [−æ, æ]. We obtain

$$\begin{aligned} \mathbb{E}\left[Z_T\right] = \mathbb{E}\left[\sup_{A \in \mathsf{T}}\left|\sum_{i,j}\epsilon_{ij}\omega_{ij}A_{ij}^2\right|\right] &\leq 4æ\mathbb{E}\left[\sup_{\alpha \in \tilde{\mathcal{A}}(d_1,T)}\left|\sum_{(i,j)}\epsilon_{ij}\omega_{ij}A_{ij}\right|\right] \\ &= 4æ\mathbb{E}\left[\sup_{\alpha \in \tilde{\mathcal{A}}(d_1,T)}|\langle \Sigma_R, A(\alpha)\rangle|\right], \end{aligned}$$

where $\Sigma_R = \sum_{(i,j)} \epsilon_{ij}\omega_{ij}E_{ij}$. Moreover, for $\alpha \in \tilde{\mathcal{A}}(d_1, T)$ we have

$$\begin{aligned} |\langle \Sigma_R, A(\alpha)\rangle| &= \left|\langle \Sigma_R, \sum_{k=1}^{N}\alpha_k U^k\rangle\right| \\ &\leq \|\alpha\|_1 u \|\Sigma_R\|_\infty. \end{aligned}$$

Finally, $\mathbb{E}\left[Z_T\right] \leq 4æd_1u\mathbb{E}\left[\|\Sigma_R\|_\infty\right]$. Combining this with the concentration inequality (53) we finally obtain

$$\mathbb{P}\left(Z_T \geq 8æd_1u\mathbb{E}\left[\|\Sigma_R\|_\infty\right] + 768p^{-1} + \frac{5}{12}T\right) \leq 4\mathrm{e}^{-pT/18}.$$

□

Lemma 8 gives that $\mathbb{P}\left(\mathcal{B}_l\right) \leq 4\exp(-p\eta^l\nu/8)$. Applying the union bound we obtain

$$\begin{aligned}\mathbb{P}\left(\mathcal{B}\right) &\leq \sum_{l=1}^{\infty}\mathbb{P}\left(\mathcal{B}_l\right)\\ &\leq 4\sum_{l=1}^{\infty}\exp(-p\eta^l\nu/8)\\ &\leq 4\sum_{l=1}^{\infty}\exp(-p\log(\eta)l\nu/8),\end{aligned}$$

where we used $e^x \geq x$. Finally, for $\nu = \frac{8\log(d)}{p\log(6/5)}$ we obtain

$$\mathbb{P}\left(\mathcal{B}\right) \leq \frac{4\exp(-p\nu\log(\eta)/8)}{1-\exp(-p\nu\log(\eta)/8)} \leq \frac{4\exp(-\log(d))}{1-\exp(-\log(d))} \leq \frac{8}{d},$$

since $d-1 \geq d/2$, which concludes the proof of (i).

Proof of (ii): The proof is very similar to that of (i); we recycle some of the notations for simplicity. Recall

$$\mathsf{D}_X = 448\rho p^{-1}\mathbb{E}\left[\|\Sigma_R\|\right]^2 + 8æ\varepsilon\mathbb{E}\left[\|\Sigma_R\|\right] + 8æd_1u\mathbb{E}\left[\|\Sigma_R\|_\infty\right] + d_\Pi + 768p^{-1}.$$

It is enough to show that the probability of the random event

$$\mathcal{B} = \Big\{\exists(L,\alpha) \in \mathcal{C}(d_1, d_\Pi, \rho, \varepsilon);\\ \left|\|L + A(\alpha)\|_\Omega^2 - \|L + A(\alpha)\|_\Pi^2\right| > \frac{1}{2}\|L + A(\alpha)\|_\Pi^2 + \mathcal{D}_2\Big\}$$

is small. Indeed, $\mathcal{B}$ contains the complement of the event we are interested in. In order to prove that the probability of $\mathcal{B}$ is small, we use a standard peeling argument. Let $\nu = \frac{72\log(d)}{p\log(6/5)}$ and $\eta = \frac{6}{5}$. For $l \in \mathbb{N}$ set

$$\mathcal{S}_l = \left\{(L,\alpha) \in \mathcal{C}(d_1, d_\Pi, \rho, \varepsilon) : \quad \eta^{l-1}\nu \leq \|L + A(\alpha)\|_\Pi^2 \leq \eta^l\nu\right\}.$$

If the event $\mathcal{B}$, then there exist $l \geq 2$ and $(L,\alpha) \in \mathcal{C}(d_1, d_\Pi, \rho, \varepsilon) \cap S_l$ such that

$$\begin{aligned}\left|\|L + A(\alpha)\|_\Omega^2 - \|L + A(\alpha)\|_\Pi^2\right| &> \frac{1}{2}\|L + A(\alpha)\|_\Pi^2 + \mathcal{D}_2\\ &> \frac{1}{2}\eta^{l-1}\nu + \mathcal{D}_2 \qquad (54)\\ &= \frac{5}{12}\eta^l\nu + \mathcal{D}_2.\end{aligned}$$

For $T > \nu$, consider the set - defined with a small abuse of notations for sake of clarity - $\tilde{\mathcal{C}}(T) = \left\{(L,\alpha) \in \mathcal{C}(d_1, d_\Pi, \rho, \varepsilon) : \|L + A(\alpha)\|_\Pi^2 \leq T\right\}$, and the event

$$\mathcal{B}_l = \left\{\exists (L,\alpha) \in \tilde{\mathcal{C}}(\eta^l \nu) : \quad \left| \|L + A(\alpha)\|_\Omega^2 - \|L + A(\alpha)\|_\Pi^2 \right| > \frac{5}{12}\eta^l \nu + \mathsf{D}_X \right\}.$$

Note that $S_l \subset \tilde{\mathcal{C}}(\eta^l \nu)$. Then, (54) implies that $\mathcal{B}_l$ holds and $\mathcal{B} \subset \cup_{l=1}^{+\infty} \mathcal{B}_l$. Thus, it is enough to estimate the probability of the events $\mathcal{B}_l$, and then apply the union bound. Such an estimation is given in the following lemma.

Lemma 10. *Let* $W_T = \sup_{(L,\alpha)\in\tilde{\mathcal{C}}(T)} \left| \|L + A(\alpha)\|_\Omega^2 - \|L + A(\alpha)\|_\Pi^2 \right|$.

$$\mathbb{P}\left(W_T \geq \mathsf{D}_X + \frac{5}{12}T\right) \leq 4\mathrm{e}^{-pT/72}.$$

Proof. By definition,

$$W_T = \sup_{(L,\alpha)\in\tilde{\mathcal{C}}(T)} \left| \sum_{(i,j)} \omega_{ij}(L_{ij} + A(\alpha)_{ij})^2 - \mathbb{E}\left[\sum_{(i,j)} \omega_{ij}(L_{ij} + A(\alpha)_{ij})^2\right] \right|.$$

We first show that W_T concentrates around its mean, then bound its expectation. The concentration proof is exactly similar to the proof in Lemma 8, but we choose $t = T/6$, and we obtain

$$\mathbb{P}\left(|W_T - \mathbb{E}[W_T]| \geq 768p^{-1} + \frac{1}{3}\left(\frac{5}{12}T\right)\right) \leq 4e^{-pT/72}. \tag{55}$$

Now, we must bound the expectation $\mathbb{E}[W_T]$. Again, we use a standard symmetrization argument [Ledoux, 2001] which gives

$$\begin{aligned}\mathbb{E}[W_T] \quad &= \mathbb{E}\left[\sup_{(L,\alpha)\in\tilde{\mathcal{C}}(T)} \left| \sum_{(i,j)} \omega_{ij}(L_{ij} + A(\alpha)_{ij})^2 - \mathbb{E}\left[\sum_{(i,j)} \omega_{ij}(L_{ij} + A(\alpha)_{ij})^2\right]\right|\right] \\ &\leq 2\mathbb{E}\left[\sup_{(L,\alpha)\in\tilde{\mathcal{C}}(T)} \left| \sum_{(i,j)} \epsilon_{ij}\omega_{ij}(L_{ij} + A(\alpha)_{ij})^2 \right|\right],\end{aligned}$$

where $\{\epsilon_{ij}\}$ is an i.i.d. Rademacher sequence. Then, the contraction inequality (see Koltchinskii [2011], Theorem 2.2) yields

$$\begin{aligned}\mathbb{E}[W_T] \leq 4æ\mathbb{E}&\left[\sup_{(L,\alpha)\in\tilde{\mathcal{C}}(T)} \left| \sum_{(i,j)} \epsilon_{ij}\omega_{ij}(L_{ij} + A(\alpha)_{ij}) \right|\right] \\ &\leq 4æ\mathbb{E}\left[\sup_{(L,\alpha)\in\tilde{\mathcal{C}}(T)} |\langle \Sigma_R, L + A(\alpha)\rangle|\right],\end{aligned}$$

where $\Sigma_R = \sum_{(i,j)} \epsilon_{ij}\omega_{ij}E_{ij}$. Moreover

$$\begin{aligned} |\langle \Sigma_R, L + A(\alpha)\rangle| \quad &\leq |\langle \Sigma_R, L\rangle| + |\langle \Sigma_R, A(\alpha)\rangle| \\ &\leq \|L\|_* \|\Sigma_R\| + \|\alpha\|_1 u \|\Sigma_R\|_\infty . \end{aligned}$$

For $(L,\alpha) \in \tilde{\mathcal{C}}(T)$ we have by assumption $\|\alpha\|_1 \leq d_1$, $\|A(\alpha)\|_\Pi \leq \sqrt{d_\Pi}$ and $\|L\|_* \leq \sqrt{\rho}\,\|L\|_F + \varepsilon$. We obtain

$$\begin{aligned} \|L\|_* \quad &\leq \sqrt{p^{-1}\rho}\,\|L\|_\Pi + \varepsilon \\ &\leq \sqrt{p^{-1}\rho}\,(\|L + A(\alpha)\|_\Pi + \|A(\alpha)\|_\Pi) + \varepsilon \\ &\leq \sqrt{p^{-1}\rho}\left(\sqrt{T} + \sqrt{d_\Pi}\right) + \varepsilon \end{aligned}$$

This gives

$$\begin{aligned} \mathbb{E}\left[W_T\right] \quad &\leq 4æ\left\{\sqrt{p^{-1}\rho}\left(\sqrt{T} + \sqrt{d_\Pi}\right) + \varepsilon\right\}\|\Sigma_R\| + 4æ d_1 u \|\Sigma_R\|_\infty \\ &\leq \frac{T}{12} + \frac{d_\Pi}{2} + 224æ^2 p^{-1}\rho\,\|\Sigma_R\|^2 + 4æ\varepsilon\,\|\Sigma_R\| + 4æ d_1 u \|\Sigma_R\|_\infty . \end{aligned}$$

Combining this with the concentration inequality (53) we finally obtain

$$\mathbb{P}\left(W_T \geq \mathsf{D}_X + \frac{5}{12}T\right) \leq 4\mathrm{e}^{-pT/72}.$$

□

Lemma 10 gives that $\mathbb{P}(\mathcal{B}_l) \leq 4\exp(-p\eta^l\nu/72)$. Applying the union bound we obtain

$$\begin{aligned} \mathbb{P}(\mathcal{B}) \quad &\leq \sum_{l=1}^{\infty} \mathbb{P}(\mathcal{B}_l) \\ &\leq 4\sum_{l=1}^{\infty} \exp(-p\eta^l\nu/72) \\ &\leq 4\sum_{l=1}^{\infty} \exp(-p\log(\eta)l\nu/72), \end{aligned}$$

where we used $e^x \geq x$. Finally, for $\nu = \frac{72\log(d)}{p\log(6/5)}$ we obtain

$$\mathbb{P}(\mathcal{B}) \leq \frac{4\exp(-p\nu\log(\eta)/72)}{1 - \exp(-p\nu\log(\eta)/72)} \leq \frac{4\exp(-\log(d))}{1 - \exp(-\log(d))} \leq 8d^{-1},$$

since $d - 1 \geq d/2$, which concludes the proof of (ii).

D Proof of Lemma 2

The first inequality is trivially true using that $\|\Sigma\|_\infty = \max_{i,j} |w_{ij}\epsilon_{ij}| \leq 1$.

We prove the second inequality with the following result taken from Klopp et al. [2017] and obtained by extension to rectangular matrices via self-adjoint dilation of Corollary 3.3 in Bandeira and van Handel [2016].

Proposition 1. *Let A be an $m_1 \times m_2$ rectangular matrix with A_{ij} independent centered bounded random variables. then, there exists a universal constant C^* such that*

$$\mathbb{E}\left[\|A\|\right] \leq C^* \left\{\sigma_1 \vee \sigma_2 + \sigma_* \sqrt{\log(m_1 \wedge m_2)}\right\},$$

$$\sigma_1 = \max_i \sqrt{\sum_j \mathbb{E}\left[A_{ij}^2\right]}, \quad \sigma_2 = \max_j \sqrt{\sum_i \mathbb{E}\left[A_{ij}^2\right]}, \quad \sigma_* = \max_{i,j} |A_{ij}|.$$

Applying Proposition 1 to Σ_R with $\sigma_1 \vee \sigma_2 \leq \sqrt{L}$ and $\sigma_* \leq 1$ we obtain

$$\mathbb{E}\left[\|\Sigma_R\|\right] \leq C^* \left\{\sqrt{L} + \sqrt{\log(m_1 \wedge m_2)}\right\}.$$

E Proof of Lemma 3

Denote $\Sigma = \nabla\mathcal{L}(X^0; Y, \Omega)$. Lemma 1 ensures that the random variables $Y_{ij} - g_j^{'}(X_{ij}^0)$ are sub-exponential for all i, j with scale and variance parameters $1/\gamma$ and σ_+^2 respectively. Then, noticing that $|\omega_{ij}| \leq 1$ implies that for all $t \geq 0$,

$$\mathbb{P}\left\{\left|\omega_{ij}\left(Y_{ij} - g_j^{'}(X_{ij}^0)\right)\right| \geq t\right\} \leq \mathbb{P}\left\{\left|Y_{ij} - g_j^{'}(X_{ij}^0)\right| \geq t\right\},$$

we obtain that the random variables $\Sigma_{ij} = \omega_{ij}\left(Y_{ij} - g_j^{'}(X_{ij}^0)\right)$ are also sub-exponential. Thus, for all i, j and for all $t \geq 0$ we have that $|\Sigma_{ij}| \leq t$ with probability at least $1 - \max\left\{2e^{-t^2/2\sigma_+^2}, 2e^{-\gamma t/2}\right\}$. A union bound argument then yields

$$\|\Sigma\|_\infty \leq t \quad \text{w. p. at least } 1 - \max\left\{2m_1m_2e^{-t^2/2\sigma_+^2}, 2m_1m_2e^{-\gamma t/2}\right\},$$

where γ and σ_+ are defined in **H** 4. Using $\log(m_1m_2) \leq 2\log d$, where $d = m_1 + m_2$ and setting

$$t = 6\max\left\{\sigma_+\sqrt{\log d}, \gamma^{-1}\log d\right\},$$

we obtain that with probability at least $1 - d^{-1}$,

$$\|\Sigma\|_\infty \leq 6\max\left\{\sigma_+\sqrt{\log d}, \gamma^{-1}\log d\right\},$$

which proves (14). The following result is taken from Klopp [2014], Proposition 11.

Proposition 2. *Let $W_1, \dots, W_n$ be independent random matrices with dimensions $m_1 \times m_2$ that satisfy $\mathbb{E}\left[W_i\right] = 0$. Suppose that*

$$\delta_* = \sup_{i \in [\![n]\!]} \inf_{\delta > 0} \left\{ \mathbb{E}\left[\exp\left(\|W_i\| / \delta\right)\right] \leq \mathrm{e} \right\} < +\infty. \tag{56}$$

Then, there exists an absolute constant c^* such that, for all $t > 0$ and with probability at least $1 - \mathrm{e}^{-t}$ we have

$$\left\| \frac{1}{n} \sum_{i=1}^{n} W_i \right\| \leq c^* \max\left\{ \sigma_W \sqrt{\frac{t + \log d}{n}}, \delta_* \left(\log \frac{\delta_*}{\sigma_W}\right) \frac{t + \log d}{n} \right\},$$

where

$$\sigma_W = \max\left\{ \left\| \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}(W_i W_i^T) \right\|^{1/2}, \left\| \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}(W_i^T W_i) \right\|^{1/2} \right\}.$$

For all $(i, j) \in [\![m_1]\!] \times [\![m_2]\!]$ define $Z_{ij} = -\omega_{ij}\left(Y_{ij} - g_j^{'}(X_{ij}^0)\right) E_{ij}$. The sub-exponentiality of the variables $\omega_{ij}\left(Y_{ij} - g_j^{'}(X_{ij}^0)\right)$ implies that for all $i, j \in [\![m_1]\!] \times [\![m_2]\!]$

$$\delta_{ij} = \inf_{\delta > 0} \quad \left\{ \mathbb{E}\left[\exp\left(\left|\omega_{ij}\left(Y_{ij} - g_j^{'}(X_{ij}^0)\right)\right| / \delta\right)\right] \leq \mathrm{e} \right\} \leq \frac{1}{\gamma}.$$

We can therefore apply Proposition 2 to the matrices Z_{ij} defined above, with the quantity

$$\sigma_Z = \max\left\{ \left\| \frac{1}{m_1 m_2} \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} \mathbb{E}(Z_{ij} Z_{ij}^T) \right\|^{1/2}, \right. \\ \left. \left\| \frac{1}{m_1 m_2} \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} \mathbb{E}(Z_{ij}^T Z_{ij}) \right\|^{1/2} \right\}. \tag{57}$$

We obtain that for all $t \geq 0$ and with probability at least $1 - \mathrm{e}^{-t}$,

$$\|\Sigma\| \leq c^* \max\left\{ \sigma_Z \sqrt{m_1 m_2 (t + \log d)}, \left(\log \frac{1}{\gamma \sigma_Z}\right) \frac{t + \log d}{\gamma} \right\}.$$

Using $\sigma_- \sqrt{\beta / (m_1 m_2)} \leq \sigma_Z \leq \sigma_+ \sqrt{\beta / (m_1 m_2)}$, and setting $t = \log d$, we further obtain for all $t \geq 0$ and with probability at least $1 - d^{-1}$:

$$\|\Sigma\| \leq c^* \max\left\{ \sigma_+ \sqrt{2 \beta \log d}, \frac{2 \log d}{\gamma} \log\left(\frac{1}{\sigma_-} \sqrt{\frac{m_1 m_2}{\beta}} \right) \right\},$$

which proves (15).