Electronic Journal of Statistics
Vol. 9 (2015) 2950–2975
ISSN: 1935-7524
DOI: 10.1214/15-EJS1093

Adaptive multinomial matrix completion

Olga Klopp

CREST and MODAL'X, Université Paris Ouest
e-mail: olga.klopp@math.cnrs.fr

Jean Lafond

Institut Mines-Télécom, Télécom ParisTech, CNRS LTCI
e-mail: jean.lafond@telecom-paristech.fr

Éric Moulines

Institut Mines-Télécom, Télécom ParisTech, CNRS LTCI
e-mail: moulines@telecom-paristech.fr

and

Joseph Salmon

Institut Mines-Télécom, Télécom ParisTech, CNRS LTCI
e-mail: joseph.salmon@telecom-paristech.fr

Abstract: The task of estimating a matrix given a sample of observed entries is known as the *matrix completion problem*. Most works on matrix completion have focused on recovering an unknown real-valued low-rank matrix from a random sample of its entries. Here, we investigate the case of highly quantized observations when the measurements can take only a small number of values. These quantized outputs are generated according to a probability distribution parametrized by the unknown matrix of interest. This model corresponds, for example, to ratings in recommender systems or labels in multi-class classification. We consider a general, non-uniform, sampling scheme and give theoretical guarantees on the performance of a constrained, nuclear norm penalized maximum likelihood estimator. One important advantage of this estimator is that it does not require knowledge of the rank or an upper bound on the nuclear norm of the unknown matrix and, thus, it is adaptive. We provide lower bounds showing that our estimator is minimax optimal. An efficient algorithm based on lifted coordinate gradient descent is proposed to compute the estimator. A limited Monte-Carlo experiment, using both simulated and real data is provided to support our claims.

MSC 2010 subject classifications: Primary 62J02, 62J99; secondary 62H12,60B20.

Keywords and phrases: Low rank matrix estimation; matrix completion; multinomial model.

Received July 2014.

1. Introduction

The matrix completion problem arises in a wide range of applications such as image processing [14, 15, 27], quantum state tomography [12], seismic data reconstruction [28] or recommender systems [20, 2]. It consists in recovering all the entries of an unknown matrix, based on partial, random and, possibly, noisy observations of its entries. Of course, since only a small proportion of entries is observed, the problem of matrix completion is, in general, ill-posed and requires an additional structure on the unknown matrix. In the classical setting, the entries are assumed to be real valued and observed in presence of additive, homoscedastic Gaussian or sub-Gaussian noise. In this framework, the matrix completion problem can be solved provided that the unknown matrix is low rank, either exactly or approximately; see [6, 16, 19, 24, 4, 18] and the references therein. Most commonly used methods amount to solve a least square program under a rank constraint or its convex relaxation provided by the nuclear (or trace) norm [9].

In this paper, we consider a statistical model where instead of observing a real-valued entry of an unknown matrix we are now able to see only highly quantized outputs. These discrete observations are generated according to a probability distribution which is parameterized by the corresponding entry of the unknown low-rank matrix. This model is well suited to the analysis of voting patterns, preference ratings, or recovery of incomplete survey data, where typical survey responses are of the form “true/false”, “yes/no” or “agree/disagree/no opinion” for instance.

The problem of matrix completion over a finite alphabet has received much less attention than the traditional unquantized matrix completion. One-bit matrix completion, corresponding to the case of binary, i.e. yes/no, observations, was introduced by [7]. In this paper, the first theoretical guarantees on the performance of a nuclear-norm constrained maximum likelihood estimator are given. The sampling model considered in [7] assumes that the entries are sampled uniformly at random. Unfortunately, this condition is unrealistic for recommender system applications: in such a context some users are more active than others and popular items are rated more frequently. Another important issue is that the method of [7] requires the knowledge of an upper bound on the nuclear norm or on the rank of the unknown matrix. Such information is usually not available in applications. On the other hand, our estimator yields a faster rate of convergence than those obtained in [7]. The problems of matrix completion with binary, multiclass, and ordinal responses were studied before the work of [7]. These early approaches typically operated on explicit low-rank factorizations instead of imposing trace norm penalties (see, e.g. [29]).

One-bit matrix completion was further considered by [5] where a max-norm constrained maximum likelihood estimate is considered. This method allows more general non-uniform sampling schemes but still requires an upper bound on the max-norm of the unknown matrix. Here again, the rates of convergence obtained in [5] are slower than the rate of convergence of our estimator. Recently, [13] consider general exponential family distributions, which cover some distributions over finite sets. Their method, requires that for a known constant α^* the maximum absolute value of the entries of the unknown matrix is bounded by $\alpha^*/\sqrt{m_1 m_2}$ where $m_1 \times m_2$ is the dimension of the matrix and the uniform sampling scheme.

In the present paper, we consider a maximum likelihood estimator with nuclear-norm penalization. Our method allows us to consider general sampling scheme and only requires the knowledge of an upper bound on the maximum absolute value of the entries of the unknown matrix. All the previous works on this model also require the knowledge of this bound with sometimes the need of additional (and more difficult to obtain) information on the unknown matrix.

The paper is organized as follows. In Section 2.1, the one-bit matrix completion is first discussed and our estimator is introduced. We establish upper bounds both on the Frobenius norm between the unknown true matrix and the proposed estimator and on the associated Kullback-Leibler divergence. In Section 2.2 lower bounds are established, showing that our upper bounds are minimax optimal up to logarithmic factors. Then, the one-bit matrix completion problem is extended to the case of a more general finite alphabet. In Section 3 an implementation based on the lifted coordinate descent algorithm recently introduced in [8] is proposed. A limited Monte Carlo experiment supporting our claims is then presented in Section 4.

Notations

For any integers $n, m_1, m_2 > 0$, $[n] := \{1, \dots, n\}$, $m_1 \vee m_2 := \max(m_1, m_2)$ and $m_1 \wedge m_2 := \min(m_1, m_2)$. We equip the set of $m_1 \times m_2$ matrices with real entries (denoted $\mathbb{R}^{m_1 \times m_2}$) with the scalar product $\langle X | X' \rangle := \operatorname{tr}(X^\top X')$. For a given matrix $X \in \mathbb{R}^{m_1 \times m_2}$ we write $\|X\|_\infty := \max_{i,j} |X_{i,j}|$ and for any $\rho \geq 1$, we denote its Schatten ρ-norm (see [1]) by

$$\|X\|_{\sigma,\rho} := \left(\sum_{i=1}^{m_1 \wedge m_2} \sigma_i^\rho(X) \right)^{1/\rho} ,$$

with $\sigma_i(X)$ the singular values of X ordered in decreasing order. The operator norm of X is $\|X\|_{\sigma,\infty} := \sigma_1(X)$. For any integer $q > 0$, we denote by $\mathbb{R}^{m_1 \times m_2 \times q}$ the set of $m_1 \times m_2 \times q$ (3-way) tensors. A tensor $\mathcal{X}$ is of the form $\mathcal{X} = (X^l)_{l=1}^q$ where $X^l \in \mathbb{R}^{m_1 \times m_2}$ for any $l \in [q]$. For any integer $p > 0$, a function $f : \mathbb{R}^q \to \mathcal{S}_p$ is called a p-link function, where $\mathcal{S}_p$ is the $p-$dimensional probability simplex. Given a p-link function f and $\mathcal{X}, \mathcal{X}' \in \mathbb{R}^{m_1 \times m_2 \times q}$, we define the squared Hellinger distance

$$\mathrm{d}_{\mathrm{H}}^2 \left(f(\mathcal{X}), f(\mathcal{X}') \right) := \frac{1}{m_1 m_2} \sum_{k \in [m_1]} \sum_{k' \in [m_2]} \sum_{j \in [p]} \left[\left(\sqrt{f^j(\mathcal{X}_{k,k'})} - \sqrt{f^j(\mathcal{X}'_{k,k'})} \right)^2 \right] ,$$

where $\mathcal{X}_{k,k'}$ denotes the vector $(X_{k,k'}^j)_{j=1}^q$. The Kullback-Leibler divergence is

$$\mathrm{KL}\left(f(\mathcal{X}), f(\mathcal{X}')\right) := \frac{1}{m_1 m_2} \sum_{k\in[m_1]} \sum_{k'\in[m_2]} \sum_{j\in[p]} \left[f^j(\mathcal{X}_{k,k'}) \log\left(\frac{f^j(\mathcal{X}_{k,k'})}{f^j(\mathcal{X}'_{k,k'})}\right)\right] .$$

For any tensor $\mathcal{X} \in \mathbb{R}^{m_1\times m_2\times q}$ we define $\mathrm{rk}(\mathcal{X}) := \max_{l\in[q]} \mathrm{rk}(X^l)$, where $\mathrm{rk}(X^l)$ is the rank of the matrix X^l and its sup-norm by $\|\mathcal{X}\|_\infty := \max_{l\in[q]} \|X^l\|_\infty$.

2. Main results

2.1. One-bit matrix completion

Assume that the observations follow a Bernoulli distribution parametrized by a matrix $\bar{X} \in \mathbb{R}^{m_1\times m_2}$. Assume in addition that an *i.i.d.* sequence of coefficients $(\omega_i)_{i=1}^n \in ([m_1]\times[m_2])^n$ is revealed and denote by Π their distribution. The observations associated to these coefficients are denoted by $(Y_i)_{i=1}^n \in \{1,2\}^n$ and distributed as follows

$$\mathbb{P}(Y_i = j) = f^j(\bar{X}_{\omega_i}), \quad j \in \{1,2\} , \tag{1}$$

where $f = (f^j)_{j=1}^2$ is a 2−link function. For example, taking $f^1(x) = \frac{e^x}{1+e^x}$ and $f^2(x) = 1 - f^1(x)$ we get the usual logistic regression. Here, the corresponding entries of $\bar{X}$ represent the log odds of the Bernoulli distribution that governs our observations.

For ease of notation, we often write $\bar{X}_i$ instead of $\bar{X}_{\omega_i}$. Denote by Φ_{Y} the (normalized) negative log-likelihood of the observations:

$$\Phi_{\mathrm{Y}}(X) = -\frac{1}{n}\sum_{i=1}^n \left(\sum_{j=1}^2 \mathbb{1}_{\{Y_i=j\}} \log\left(f^j(X_i)\right)\right) . \tag{2}$$

Let $\gamma > 0$ be an upper bound of $\|\bar{X}\|_\infty$. We consider the following estimator of $\bar{X}$:

$$\hat{X} = \underset{X\in\mathbb{R}^{m_1\times m_2}, \|X\|_\infty\le\gamma}{\arg\min} \Phi_{\mathrm{Y}}^\lambda(X) , \quad \text{where} \quad \Phi_{\mathrm{Y}}^\lambda(X) = \Phi_{\mathrm{Y}}(X) + \lambda\|X\|_{\sigma,1} , \tag{3}$$

with $\lambda > 0$ being a regularization parameter. Consider the following assumptions.

H1. *The functions $x \mapsto -\ln(f^j(x))$, $j = 1, 2$ are convex. In addition, There exist positive constants H_γ, L_γ and K_γ such that:*

$$H_\gamma \ge 2 \sup_{|x|\le\gamma} \left(|\log(f^1(x))| \vee |\log(f^2(x))|\right) , \tag{4}$$

$$L_\gamma \ge \max\left(\sup_{|x|\le\gamma} \frac{|(f^1)'(x)|}{f^1(x)}, \sup_{|x|\le\gamma} \frac{|(f^2)'(x)|}{f^2(x)}\right) , \tag{5}$$

$$K_\gamma = \inf_{|x|\le\gamma} g(x) , \quad \textit{where } g(x) = \frac{(f^1)'(x)^2}{8 f^1(x)(1-f^1(x))} . \tag{6}$$

This assumption is quite mild, for example, it includes the logistic regression (with $f^1(x) = \frac{e^x}{1+e^x}$) and probit model (with $f^1(x) = \Phi(x/\sigma)$ where Φ is the cumulative distribution function of a standard Gaussian).

Remark 1. As shown in [7, Lemma 2], K_γ satisfies

$$K_\gamma \leq \inf_{\substack{x,y \in \mathbb{R} \\ |x| \leq \gamma \\ |y| \leq \gamma}} \left(\sum_{j=1}^{2} \left(\sqrt{f^j(x)} - \sqrt{f^j(y)} \right)^2 / (x-y)^2 \right) . \tag{7}$$

Our framework allows a general distribution Π. We assume that Π satisfies the following assumptions introduced in [18] in the classical setting of unquantized matrix completion:

H2. *There exists a constant $\mu \geq 1$ such that, for any $m_1 > 0$ and $m_2 > 0$*

$$\min_{k \in [m_1], k' \in [m_2]} \pi_{k,k'} \geq 1/(\mu m_1 m_2) , \quad \textit{where } \pi_{k,k'} = \mathbb{P}(\omega_1 = (k,k')) . \tag{8}$$

Denote by $R_k = \sum_{k'=1}^{m_2} \pi_{k,k'}$ and $C_{k'} = \sum_{k=1}^{m_1} \pi_{k,k'}$ the probability of revealing a coefficient from row k and column k', respectively.

H3. *There exists a constant $\nu \geq 1$ such that, for all m_1, m_2,*

$$\max_{k,l}(R_k, C_l) \leq \frac{\nu}{m_1 \wedge m_2} ,$$

The first assumption ensures that every coefficient has a nonzero probability of being observed, whereas the second assumption requires that no column nor row is sampled with too high probability (see also [10, 18] for more details on these conditions). For instance, the uniform distribution yields $\mu = \nu = 1$. Define

$$d = m_1 + m_2 , \quad M = m_1 \vee m_2, \quad m = m_1 \wedge m_2 . \tag{9}$$

We are now ready to state our main results. We have two goals: the first is the recovery of the distribution of Y given by $f(\bar{X})$ and, the second, is to accurately recover the matrix $\bar{X}$ itself. The former is addressed by Theorem 1 which provides an upper bound on the KL divergence between $f(\bar{X})$ and $f(\hat{X})$, the latter is tackled by Corollary 2, which bounds the Frobenius norm of $\bar{X}$ estimation error.

Theorem 1. *Assume H1, H2, H3 and that $\|\bar{X}\|_\infty \leq \gamma$. Assume in addition that $n \geq 2m\log(d)/(9\nu)$. Take*

$$\lambda = 6L_\gamma \sqrt{\frac{2\nu \log(d)}{mn}} .$$

Then, with probability at least $1 - 3d^{-1}$ the Kullback-Leibler divergence is bounded by

$$\mathrm{KL}\left(f(\bar{X}), f(\hat{X})\right) \leq \mu \max\left(\bar{c}\mu\nu \frac{L_\gamma^2 \operatorname{rk}(\bar{X})}{K_\gamma} \frac{M \log(d)}{n}, \mathrm{e} H_\gamma \sqrt{\frac{\log(d)}{n}} \right) ,$$

with $\bar{c}$ a universal constant whose value is specified in the proof.

Proof. See Section 5.1.

This result immediately gives an upper bound on the estimation error of $\hat{X}$, measured in Frobenius norm:

Corollary 2. *Under the same assumptions and notations of Theorem 1 we have with probability at least* $1 - 3d^{-1}$

$$\frac{\|\bar{X} - \hat{X}\|_{\sigma,2}^2}{m_1 m_2} \leq \mu \max\left(\bar{c}\mu\nu \frac{L_\gamma^2 \operatorname{rk}(\bar{X})}{K_\gamma^2} \frac{M\log(d)}{n}, \frac{\mathrm{e}H_\gamma}{K_\gamma}\sqrt{\frac{\log(d)}{n}}\right) .$$

Proof. Using Lemma 9 and Theorem 1, the result follows.

Remark 2. Note that, up to the factor L_γ^2/K_γ^2, the rate of convergence given by Corollary 2, is the same as in the case of usual unquantized matrix completion, see, for example, [18] and [19]. For this usual matrix completion setting, it has been shown in [19, Theorem 3] that this rate is minimax optimal up to a logarithmic factor. Let us compare this rate of convergence with those obtained in previous works on 1-bit matrix completion. In [7], the parameter $\bar{X}$ is estimated by minimizing the negative log-likelihood under the constraints $\|X\|_\infty \leq \gamma$ and $\|X\|_{\sigma,1} \leq \gamma\sqrt{rm_1m_2}$ for some $r > 0$. Under the assumption that $\operatorname{rk}(\bar{X}) \leq r$, they could prove that

$$\frac{\|\bar{X} - \hat{X}\|_{\sigma,2}^2}{m_1 m_2} \leq C_\gamma \sqrt{\frac{rd}{n}} ,$$

where C_γ is a constant depending on γ (see [7, Theorem 1]). This rate of convergence is slower than the rate of convergence given by Corollary 2. [5] studied a max-norm constrained maximum likelihood estimate and obtain a rate of convergence similar to [7]. In [13], matrix completion was considered for a likelihood belonging to the exponential family. Note, for instance, that the logit distribution belongs to such a family. The following upper bound on the estimation error is provided (see [13, Theorem 1])

$$\frac{\|\bar{X} - \hat{X}\|_{\sigma,2}^2}{m_1 m_2} \leq C_\gamma \left(\alpha_*^2 \frac{\operatorname{rk}(\bar{X}) M \log(M)}{n}\right) . \tag{10}$$

Comparing with Corollary 2, (10) contains an additional term α_*^2 where α_* is an upper bound of $\sqrt{m_1 m_2}\|\bar{X}\|_\infty$.

2.2. Minimax lower bounds for one-bit matrix completion

Corollary 2 insures that our estimator achieves certain Frobenius norm errors. We now discuss the extent to which this result is optimal. A classical way to address this question is by determining minimax rates of convergence.

For any integer $0 \leq r \leq \min(m_1, m_2)$ and any $\gamma > 0$, we consider the following family of matrices

$$\mathcal{F}(r,\gamma) = \left\{\bar{X} \in \mathbb{R}^{m_1 \times m_2} : \operatorname{rank}(\bar{X}) \leq r, \|\bar{X}\|_\infty \leq \gamma\right\} .$$

We will denote by $\inf_{\hat{X}}$ the infimum over all estimators $\hat{X}$ that are functions of the data $(\omega_i, Y_i)_{i=1}^n$. For any $X \in \mathbb{R}^{m_1 \times m_2}$, let $\mathbb{P}_X$ denote the probability distribution of the observations $(\omega_i, Y_i)_{i=1}^n$ for a given $2-$link function f and sampling distribution Π. We establish a lower bound under an additional assumption on the function f^1:

H4. *$(f^1)'$ is decreasing on $\mathbb{R}_+$ and $K_\gamma = g(\gamma)$ where g and K_γ are defined in (6).*

In particular, H4 is satisfied in the case of logit or probit models. The following theorem establishes a lower bound on the minimax risk in squared Frobenius norm:

Theorem 3. *Assume H4. Let $\alpha \in (0, 1/8)$ Then there exists a constant $c > 0$ such that, for all $m_1, m_2 \geq 2$, $1 \leq r \leq m$, and $\gamma > 0$,*

$$\inf_{\hat{X}} \sup_{\bar{X} \in \mathcal{F}(r,\gamma)} \mathbb{P}_{\bar{X}} \left(\frac{\|\hat{X} - \bar{X}\|_2^2}{m_1 m_2} > c \min\left\{ \gamma^2, \frac{Mr}{n K_0} \right\} \right) \ \geq \ \delta(\alpha, M) \ ,$$

where

$$\delta(\alpha, M) = \frac{1}{1 + 2^{-rM/16}} \left(1 - 2\alpha - \frac{1}{2}\sqrt{\frac{\alpha}{\log(2)(rM)}} \right) \ . \tag{11}$$

Proof. See Section 5.3.

Note that the lower bound given by Theorem 3 is proportional to the rank multiplied by the maximum dimension of $\bar{X}$ and inversely proportional the sample size n. Therefore the lower bound matches the upper bound given by Corollary 2 up to a constant and a logarithmic factor. The lower bound does not capture the dependance on γ, note however that the upper and lower bound only differ by a factor L_γ^2 / K_γ.

2.3. Extension to multi-class problems

Let us now consider a more general setting where the observations follow a distribution over a finite set $\{1, \ldots, p\}$, parameterized by a tensor $\bar{\mathcal{X}} \in \mathbb{R}^{m_1 \times m_2 \times q}$. The distribution of the observations $(Y_i)_{i=1}^n \in [p]^n$ is

$$\mathbb{P}(Y_i = j) = f^j(\bar{\mathcal{X}}_{\omega_i}), \quad j \in [p] \ ,$$

where $f = (f^j)_{j=1}^p$ is now a p-link function and $\bar{\mathcal{X}}_{\omega_i}$ denotes the vector $(\bar{X}^l_{\omega_i})_{l=1}^q$. The negative log-likelihood of the observations is now given by:

$$\Phi_{\mathrm{Y}}(\mathcal{X}) = -\frac{1}{n} \sum_{i=1}^n \left(\sum_{j=1}^p \mathbb{1}_{\{Y_i = j\}} \log\left(f^j(\mathcal{X}_i) \right) \right) \ . \tag{12}$$

where we use the notation $\mathcal{X}_i = \mathcal{X}_{\omega_i}$. Our proposed the estimator is defined as:

$$\hat{\mathcal{X}} = \mathop{\arg\min}_{\substack{\mathcal{X} \in \mathbb{R}^{m_1 \times m_2 \times q} \\ \|\mathcal{X}\|_\infty \leq \gamma}} \Phi_Y^\lambda(\mathcal{X}) \ , \quad \text{where} \quad \Phi_Y^\lambda(\mathcal{X}) = \Phi_{\mathrm{Y}}(\mathcal{X}) + \lambda \sum_{j=1}^q \|X^j\|_{\sigma,1} \ , \tag{13}$$

In order to extend the results of the previous sections we make an additional assumption which allows to split the log-likelihood as a sum.

H5. *There exist functions* $(g_l^j)_{(l,j)\in[p]\times[q]}$ *such that the* p*-link function* f *can be factorized as follows*

$$f^j(x_1,\dots,x_q)=\prod_{l=1}^{q} g_l^j(x_l) \quad \textit{for } j\in[p]\ .$$

The model considered above covers many finite distributions including among others logistic binomial (see Section 2.1) and conditional logistic multinomial (see Section 3).

Assumptions on constants depending on the link function are extended by

H6. *There exist positive constant* H_γ, L_γ *and* K_γ *such that:*

$$H_\gamma \geq \max_{(j,l)\in[p]\times[q]} \sup_{|x|\leq\gamma} 2|\log(g_l^j(x))|\ , \tag{14}$$

$$L_\gamma \geq \max_{(j,l)\in[p]\times[q]} \sup_{|x|\leq\gamma} \left|\frac{(g_l^j)'(x)}{g_l^j(x)}\right|\ , \tag{15}$$

$$K_\gamma \leq \inf_{\substack{x,y\in\mathbb{R}^q\\ \|x\|_\infty\leq\gamma\\ \|y\|_\infty\leq\gamma}} \left(\sum_{j=1}^{p}\left(\sqrt{f^j(x)}-\sqrt{f^j(y)}\right)^2/\|x-y\|_2^2\right)\ . \tag{16}$$

For any tensor $\mathcal{X}\in\mathbb{R}^{m_1\times m_2\times q}$, we write $\bar\Sigma := \nabla\,\Phi_{\mathrm{Y}}(\bar{\mathcal{X}})\in\mathbb{R}^{m_1\times m_2\times q}$. We also define the sequence of matrices $(E_i)_{i=1}^n$ associated to the revealed coefficients $(\omega_i)_{i=1}^n$ by $E_i := e_{k_i}(e'_{l_i})^\top$ where $(k_i,l_i)=\omega_i$ and with $(e_k)_{k=1}^{m_1}$ (*resp.* $((e'_l)_{l=1}^{m_2})$ being the canonical basis of $\mathbb{R}^{m_1}$ (*resp.* $\mathbb{R}^{m_2}$). Furthermore, if $(\varepsilon_i)_{1\leq i\leq n}$ is a Rademacher sequence independent from $(\omega_i)_{i=1}^n$ and $(Y_i)_{1\leq i\leq n}$ we define the matrix Σ_R as follow

$$\Sigma_R := \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i E_i\ .$$

We can now state the main results of this paper.

Theorem 4. *Assume H2, H5 and H6 hold,* $\lambda > 2\max_{l\in[q]}\|\bar\Sigma^l\|_{\sigma,\infty}$ *and* $\|\bar{\mathcal{X}}\|_\infty\leq\gamma$. *Then, with probability at least* $1-2d^{-1}$, *the Kullback-Leibler divergence is bounded by*

$$\mathrm{KL}\left(f(\bar{\mathcal{X}}),f(\hat{\mathcal{X}})\right)\leq$$
$$\mu\max\left(4\mu\frac{m_1m_2\,\mathrm{rk}(\bar{\mathcal{X}})}{K_\gamma}\left(\lambda^2+256\mathrm{e}(qL_\gamma\mathbb{E}\,\|\Sigma_R\|_{\sigma,\infty})^2\right),\mathrm{e}H_\gamma\sqrt{\frac{\log(d)}{n}}\right)\ .$$

with d *defined in* (9).

Proof. See Section 5.1.

Note that the lower bound of λ is stochastic and the expectation $\mathbb{E}\|\Sigma_R\|_{\sigma,\infty}$ is unknown. However, these quantities can be controlled using H3.

Theorem 5. *Assume H2, H3, H5 and H6 hold and that $\|\bar{\mathcal{X}}\|_\infty \le \gamma$. Assume in addition that $n \ge 2m\log(d)/(9\nu)$. Take*

$$\lambda = 6L_\gamma \sqrt{\frac{2\nu\log(d)}{mn}} \; .$$

Then, with probability at least $1-(2+q)d^{-1}$, the Kullback-Leibler divergence is bounded by

$$\mathrm{KL}\left(f(\bar{\mathcal{X}}), f(\hat{\mathcal{X}})\right) \;\le\; \mu \max\left(\bar{c}\mu\nu\frac{q^2L_\gamma^2\,\mathrm{rk}(\bar{\mathcal{X}})}{K_\gamma}\frac{M\log(d)}{n}, \mathrm{e}H_\gamma\sqrt{\frac{\log(d)}{n}}\right) \quad ,$$

with $\bar{c}$ a universal constant , d, m and M defined in (9).

Proof. See Section 5.2.

3. Implementation

In this section an implementation for the following p-class link function is given:

$$f^j(x^1,\dots,x^{p-1}) = \begin{cases} \exp(x^j)\left(\prod_{l=1}^{j}(1+\exp(x^l))\right)^{-1} & \text{if } j \in [p-1] \; , \\ \left(\prod_{l=1}^{p-1}(1+\exp(x^l))\right)^{-1} & \text{if } j = p \; . \end{cases}$$

This p-class link function boils down to parameterizing the distribution of the observation as follows:

$$\mathbb{P}(Y_i = 1) = \frac{\exp(\bar{X}_i^1)}{1+\exp(\bar{X}_i^1)} \; ,$$
$$\mathbb{P}(Y_i = j | Y_i > j-1) = \frac{\exp(\bar{X}_i^j)}{1+\exp(\bar{X}_i^j)} \quad \text{for } j \in \{2,\dots,p-1\} \; .$$

Assumption H5 is satisfied and the problem (13) is separable *w.r.t.* each matrix X^l. Following [7], we solve (13) without taking into account the constraint γ; as reported in [7] and confirmed by our experiments, the impact of this projection is negligible, whereas it increases significantly the computation burden.

Because the problem is separable, it suffices to solve in parallel each subproblem

$$\hat{X}^l = \operatorname*{arg\,min}_{X\in\mathbb{R}^{m_1\times m_2}} \Phi_\lambda^l(X) \; , \quad \text{where} \quad \Phi_\lambda^l(X) = \Phi^l(X) + \lambda\|X\|_{\sigma,1} \; . \tag{17}$$

This can be achieved by using the coordinate gradient descent algorithm introduced by [8]. To describe the algorithm, consider first the set of normalized rank one matrices

$$\mathcal{M} := \left\{ M \in \mathbb{R}^{m_1\times m_2} | M = uv^\top \mid \|u\|_2 = \|v\|_2 = 1, \; \right\} \; .$$

Define Θ the linear space of real-valued functions on $\mathcal{M}$ with finite support, *i.e.,* any $\theta \in \Theta$ satisfies $\theta(M) = 0$ except for a finite number of $M \in \mathcal{M}$. This space is equipped with the ℓ^1-norm $\|\theta\|_1 = \sum_{M\in\mathcal{M}} |\theta(M)|$. Define by Θ_+ the positive orthant, *i.e.,* the cone of functions $\theta \in \Theta$ such that $\theta(M) \geq 0$ for all $M \in \mathcal{M}$. Any matrix $X \in \mathbb{R}^{m_1\times m_2}$ can be associated to an element $\theta \in \Theta_+$ satisfying

$$X = \sum_{M\in\mathcal{M}} \theta(M)M \; . \tag{18}$$

Such function is not unique. Consider an SVD of X *i.e.,* $X = \sum_{i=1}^m \lambda_i u_i v_i^\top$, where $(\lambda_i)_{i=1}^m$ are the singular values and $(u_i)_{i=1}^m$, $(v_i)_{i=1}^m$ are left and right singular vectors, then $\theta_X = \sum_{i=1}^m \lambda_i \delta_{u_i v_i^\top}$ satisfies (18), with $\delta_M \in \Theta$ is the function on $\mathcal{M}$ satisfying $\delta_M(M) = 1$ and $\delta_M(M') = 0$ if $M' \neq M$. As seen below, the function θ_X plays a key role.

Conversely, for any $\theta \in \Theta_+$, define

$$W : \theta \to W_\theta := \sum_{M\in\mathcal{M}} \theta(M)M \; .$$

and the auxiliary objective function:

$$\tilde{\Phi}^l_\lambda : \theta \to \tilde{\Phi}^l_\lambda(\theta) := \lambda \sum_{M\in\mathcal{M}} \theta(M) + \Phi^l(W_\theta) \; . \tag{19}$$

The triangular inequality implies that for all $\theta \in \Theta_+$,

$$\|W_\theta\|_{\sigma,1} \leq \|\theta\|_1 \; .$$

For $\theta \in \Theta$ we denote by $\operatorname{supp}(\theta)$ the support of θ *i.e.,* the subset of $\mathcal{M}$ such that $\theta(M) \neq 0 \iff M \in \operatorname{supp}(\theta)$. If for any $M, M' \in \operatorname{supp}(\theta)$, $M \neq M'$, $\langle M \mid M' \rangle = 1$, then $\|\theta\|_1 = \|W_\theta\|_{\sigma,1}$. Indeed in such case $\sum_{M\in\mathcal{M}} \theta(M)M$ defines a SVD of W_θ. Therefore the minimization of (19) is actually equivalent to the minimization of (17); see [8, Theorem 3.2]. The minimization (19) can be implemented using a coordinate gradient descent algorithm which updates at each iteration the nonnegative finite support function θ.

The algorithm is summarized in Algorithm 1. Compared to the Soft-Impute [23] or the SVT [3] algorithms, this algorithm does not require the computation of a full SVD at each step of the main loop of an iterative (proximal) algorithm (recall that the proximal operator associated to the nuclear norm is the soft-thresholding operator of the singular values). The proposed algorithm requires only to compute the largest singular values and associated singular vectors.

Another interest of this algorithm is that it only requires to evaluate the coordinate of the gradient for the entries which have been actually observed. It is therefore memory efficient when the number of observations is smaller than the total number ocoefficients $m_1 m_2$, which is the typical setting in which matrix completion is used. Moreover, we use Arnoldi iterations to compute the top singular values and vector pairs (see [11, Section 10.5] for instance) which allows us to take full advantage of sparse structures, the minimizations in the

Algorithm 1: Lifted coordinate gradient descent

Initialization: initial parameter θ_0, precision ϵ
Loop:
Compute the top singular vector pair of $-\nabla\,\Phi^l(W_{\theta_k})$: u_k, v_k
$g_k \leftarrow \lambda + \langle \nabla\,\Phi^l(W_{\theta_k})\,|\,u_k v_k^\top \rangle$
if $g_k \leq -\epsilon/2$ **then**
  $\beta_k \leftarrow \underset{b\in\mathbb{R}_+}{\arg\min}\, \tilde{\Phi}^l_\lambda \left(\theta + b\delta_{u_k v_k^\top}\right)$
  $\theta_{k+1} \leftarrow \theta_k + \beta_k \delta_{u_k v_k^\top}$,
else
  $g_k^{\max} \leftarrow \max_{M\in\operatorname{supp}(\theta_k)} |\lambda + \langle \nabla\,\Phi^l(W_{\theta_k})\,|\,M\rangle|$
  if $g_k^{\max} \leq \epsilon$ **then**
    Break
  else
    $\theta_{k+1} \leftarrow \underset{\theta'\in\Theta_+,\operatorname{supp}(\theta')\subset\operatorname{supp}(\theta_k)}{\arg\min} \tilde{\Phi}^l_\lambda(\theta')$

TABLE 1
Execution time of the proposed algorithm for a simulated Bernouilli example

Parameter Size	1000×1000	3000×3000	10000×10000
Observations	$100\cdot10^3$	$1\cdot10^6$	$10\cdot10^6$
Execution Time (s.)	4.5	52	730

inner loop are carried out using the L-BFGS-B algorithm. Table 1 illustrates that the execution time (on a 3.07Ghz w3550 Xeon CPU with RAM 1.66 Go, Cache 8 Mo, C implementation) for the one-bit matrix completion estimation is roughly linear in the number of observations.

4. Numerical experiments

We have performed numerical experiments on both simulated and real data provided by the MovieLens project (http://grouplens.org). Both the one-bit matrix completion - $p = 2$, $q = 1$ - and the extended multi-class setting -$p = 5$, $q = 4$ - are considered; comparisons are also provided with the classical Gaussian matrix completion algorithm to assess the potential gain achieved by explicitly taking into account the facts that the observations belong to a finite alphabet. Only a limited part of the experiments are reported in this article; a more extensive assessment can be obtained upon authors request.

For each matrix $\bar{X}^l$ we sampled uniformly five unitary (for the Euclidean norm) vector pairs $(u_k^l, v_k^l)_{k=1}^5$. The matrix $\bar{X}^l$ is then defined as

$$\bar{X}^l = \Gamma\sqrt{m_1 m_2}\sum_{k=1}^{5} \alpha_k u_k^l (v_k^l)^\top + \eta^l \mathbf{I}_{m_1\times m_2},$$

with $(\alpha_1,\ldots,\alpha_5) = (2, 1, 0.5, 0.25, 0.1)$, Γ a scaling factor and $\mathbf{I}_{m_1\times m_2}$ the $m_1 \times m_2$ matrix of ones. The term η_l has been fixed so that each class has the

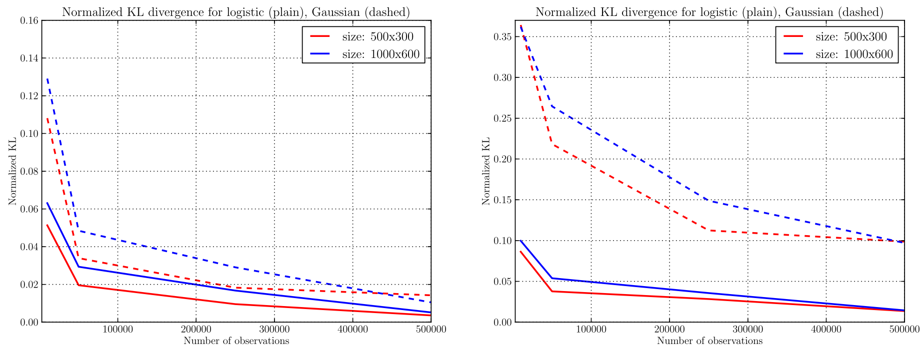

FIG 1. *Kullback-Leibler divergence between the estimated and the true model for different matrices sizes and sampling fraction, normalized by number of classes. Right figure: binomial and the Gaussian models; left figure: multinomial with five classes and Gaussian model.*

same average probability *i.e.,* $f^j((\mathbb{E}[\bar{X}^l])_{l=1}^{p-1}) = 1/p$ for $j \in [p]$. Note that the factor $\sqrt{m_1 m_2}$ implies that the variance of $\bar{X}^l$ coefficients does not depend on m_1 and m_2. The sizes investigated are $(m_1, m_2) \in \{(500, 300), (1000, 600)\}$.

The observations are sampled to the conditional multinomial logistic model introduced in Section 3. For comparison purposes we have also computed $\hat{X}^{\mathcal{N}}$, the classical Gaussian version, that is:

$$\hat{X}^{\mathcal{N}} := \underset{X \in \mathbb{R}^{m_1 \times m_2}}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} (Y_i - X_i)^2 + \lambda \|X\|_{\sigma,1} \ .$$

Contrary to the logit version, the Gaussian matrix completion does not directly recover the distribution of the observations $(Y_i)_{i=1}^n$. However, we can estimate $\mathbb{P}(Y_i = j)$ by the following quantity:

$$F_{\mathcal{N}(0,1)}(p_{j+1}) - F_{\mathcal{N}(0,1)}(p_j) \text{ with } p_j = \begin{cases} 0 & \text{if } j = 1 \ , \\ \frac{j - 0.5 - \hat{X}_i^{\mathcal{N}}}{\hat{\sigma}} & \text{if } 0 < j < p \\ 1 & \text{if } j = p \ , \end{cases}$$

where $F_{\mathcal{N}(0,1)}$ is the cdf of a zero-mean standard Gaussian random variable. Note that this setting is equivalent to considering a probit link function.

The choice of the regularization parameter λ has been solved for all methods by performing 5-fold cross-validation on a geometric grid of size $0.6 \log(n)$ (note that the estimators are null for λ greater than $\|\nabla \Phi_{\mathrm{Y}}(0)\|_{\sigma,\infty}$).

As evidenced in Figure 1, the Kullback-Leibler divergence for the logistic estimator is significantly lower than for the Gaussian estimator, for both the $p = 2$ and $p = 5$ cases. This was expected because the Gaussian model assume implicitly symmetric distributions with the same variance for all the ratings, These assumptions are of course avoided by the logistic model.

Regarding the prediction error, Table 2 and Table 3 summarize the results obtained for a 1000×600 matrix. The logistic model outperforms the Gaussian model (slightly for $p = 2$ and significantly for $p = 5$).

Table 2
Prediction errors for a binomial (2 classes) underlying model, for a 1000×600 matrix

Number of observations	$10 \cdot 10^3$	$50 \cdot 10^3$	$250 \cdot 10^3$	$500 \cdot 10^3$
Gaussian prediction error	0.50	0.38	0.32	0.32
Logistic prediction error	0.46	0.33	0.31	0.31

Table 3
Prediction Error for a multinomial (5 classes) distribution against a 1000×600 matrix

Number of observations	$10 \cdot 10^3$	$50 \cdot 10^3$	$250 \cdot 10^3$	$500 \cdot 10^3$
Gaussian prediction error	0.75	0.75	0.72	0.71
Logistic prediction error	0.75	0.67	0.58	0.57

We have also run the same estimators on the MovieLens $100k$ dataset. In this case, the Kullback-Leibler divergence cannot be computed. Therefore, to assess the prediction errors, we randomly select 20% of the entries as a test set, and the remaining entries are split between a training set (80%) and a validation set (20%).

For this dataset, ratings range from 1 to 5. To consider the benefit of a binomial model, we have tested each rating against the others (*e.g.,* ratings 5 are set to 0 and all others are set to 1).

These results are summarized in Table 4. For the multinomial case, we find a prediction error of 0.59 for the logistic model against a 0.63 for the Gaussian one (in that case a random choice yields 0.80 error rate).

Table 4
Binomial prediction error when performing one versus the others procedure on the MovieLens $100k$ dataset

Rating against the others	1	2	3	4	5
Gaussian prediction error	0.12	0.20	0.39	0.46	0.30
Logistic prediction error	0.06	0.11	0.27	0.34	0.20

5. Proofs of main results

5.1. Proof of Theorem 1 and Theorem 4

Proof. Since Theorem 1 is an application of Theorem 4 for $p = 2$ and $q = 1$ it suffices to prove Theorem 4.

We consider a tensor $\mathcal{X}$ which satisfies $\Phi^\lambda_{\mathrm{Y}}(\mathcal{X}) \leq \Phi^\lambda_{\mathrm{Y}}(\bar{\mathcal{X}})$, (*e.g.,* $\mathcal{X} = \hat{\mathcal{X}}$). We get from Lemma 6

$$\Phi_{\mathrm{Y}}(\mathcal{X}) - \Phi_{\mathrm{Y}}(\bar{\mathcal{X}}) \leq \lambda\sqrt{\bar{r}}\sqrt{\mathrm{KL}\left(f(\bar{\mathcal{X}}), f(\mathcal{X})\right)}\,, \tag{20}$$

where

$$\bar{r} = \frac{2m_1 m_2 \,\mathrm{rk}(\tilde{\mathcal{X}})}{K_\gamma}. \tag{21}$$

Let us define

$$\mathrm{D}\left(f(\bar{\mathcal{X}}), f(\mathcal{X})\right) := \mathbb{E}\left[\left(\Phi_{\mathrm{Y}}(\mathcal{X}) - \Phi_{\mathrm{Y}}(\bar{\mathcal{X}})\right)\right]\,, \tag{22}$$

where the expectation is taken both over the $(E_i)_{1\leq i\leq n}$ and $(Y_i)_{1\leq i\leq n}$. As stated in Lemma 11, H2 implies $\mu\,\mathrm{D}\left(f(\bar{\mathcal{X}}), f(\mathcal{X})\right) \geq \mathrm{KL}\left(f(\bar{\mathcal{X}}), f(\mathcal{X})\right)$. We now need to control the left hand side of (20) uniformly over X with high probability. Since we assume $\lambda > 2\max_{l\in[q]}\|\bar{\Sigma}^l\|_{\sigma,\infty}$ applying Lemma 10 (30) and then Lemma 11 yields

$$\sum_{l=1}^{q}\|X^l-\bar{X}^l\|_{\sigma,1} \leq 4\sqrt{\bar{r}}\sqrt{\mathrm{KL}\left(f(\bar{\mathcal{X}}), f(\mathcal{X})\right)} \leq 4\sqrt{\mu\bar{r}}\sqrt{\mathrm{D}\left(f(\bar{\mathcal{X}}), f(\mathcal{X})\right)}\ , \quad (23)$$

Consequently, if we define $\mathcal{C}(r)$ as

$$\mathcal{C}(r) := \left\{\mathcal{X}\in\mathbb{R}^{m_1\times m_2\times q}:\ \sum_{l=1}^{q}\|X^l-\bar{X}^l\|_{\sigma,1} \leq \sqrt{r\,\mathrm{D}\left(f(\bar{\mathcal{X}}), f(\mathcal{X})\right)}\right\}\ ,$$

we need to control $(\Phi_{\mathrm{Y}}(\mathcal{X})-\Phi_{\mathrm{Y}}(\bar{\mathcal{X}}))$ for $\mathcal{X}\in\mathcal{C}(16\mu\bar{r})$. We have to ensure that $\mathrm{D}\left(f(\bar{\mathcal{X}}), f(\mathcal{X})\right)$ is greater than a given threshold $\beta>0$ and therefore we define the following set

$$\mathcal{C}_\beta(r) = \left\{\mathcal{X}\in\mathcal{C}(r),\ \mathrm{D}\left(f(\bar{\mathcal{X}}), f(\mathcal{X})\right)\geq\beta\right\}\ . \quad (24)$$

We then consider the two following cases.
Case 1. If $\mathrm{D}\left(f(\bar{\mathcal{X}}), f(\mathcal{X})\right) > \beta$, (23) gives $X\in\mathcal{C}_\beta(16\mu\bar{r})$. Plugging Lemma 12 in (20) with $\beta = 2M_\gamma\sqrt{\log(d)}/(\eta\sqrt{n\log(\alpha)})$, $\alpha=e$ and $\eta=1/(4\alpha)$ then it holds with probability at least $1-2d^{-1}/(1-d^{-1})\geq 1-2/d$

$$\frac{\mathrm{D}\left(f(\bar{\mathcal{X}}), f(\mathcal{X})\right)}{2} - \epsilon(16\mu\bar{r},\alpha,\eta) \leq \lambda\sqrt{\bar{r}}\sqrt{\mathrm{KL}\left(f(\bar{\mathcal{X}}), f(\mathcal{X})\right)}\ ,$$

where ϵ is defined in Lemma 12. Recalling Lemma 11 we get

$$\frac{\mathrm{KL}\left(f(\bar{\mathcal{X}}), f(\mathcal{X})\right)}{2\mu} - \lambda\sqrt{\bar{r}}\sqrt{\mathrm{KL}\left(f(\bar{\mathcal{X}}, f(\mathcal{X})\right)} - \epsilon(16\mu\bar{r},\alpha,\eta) \leq 0\ .$$

An analysis of this second order polynomial and the relation $\epsilon(16\mu\bar{r},\alpha,\eta)/\mu = \epsilon(16\bar{r},\alpha,\eta)$ lead to

$$\sqrt{\mathrm{KL}\left(f(\bar{\mathcal{X}}), f(\mathcal{X})\right)} \leq \mu\left(\lambda\sqrt{\bar{r}}+\sqrt{\lambda^2\bar{r}+2\epsilon(16\bar{r},\alpha,\eta)}\right)\ . \quad (25)$$

Applying the inequality $(a+b)^2\leq 2(a^2+b^2)$ gives the bound of Theorem 4.
Case 2. If $\mathrm{D}\left(f(\bar{\mathcal{X}}), f(\mathcal{X})\right)\leq\beta$ then Lemma 11 yields

$$\mathrm{KL}\left(f(\bar{\mathcal{X}}), f(\mathcal{X})\right)\leq\mu\beta\ . \quad (26)$$

Combining (25) and (26) concludes the proof.

For $X\in\mathbb{R}^{m_1\times m_2}$, denote by $\mathcal{S}_1(X)\subset\mathbb{R}^{m_1}$ (*resp.* $\mathcal{S}_2(X)\subset\mathbb{R}^{m_2}$) the linear spans generated by left (*resp.* right) singular vectors of X. $P_{\mathcal{S}_1^\perp(X)}$ (*resp.* $P_{\mathcal{S}_2^\perp(X)}$) denotes the orthogonal projections on $\mathcal{S}_1^\perp(X)$ (*resp.* $\mathcal{S}_2^\perp(X)$). We then define the following orthogonal projections on $\mathbb{R}^{m_1\times m_2}$

$$\mathcal{P}_X^\perp:\tilde{X}\mapsto P_{\mathcal{S}_1^\perp(X)}\tilde{X}P_{\mathcal{S}_2^\perp(X)}\ \text{and}\ \mathcal{P}_X:\tilde{X}\mapsto\tilde{X}-\mathcal{P}_X^\perp(\tilde{X})\ .$$

Lemma 6. *Let $\mathcal{X}, \tilde{\mathcal{X}} \in \mathbb{R}^{m_1 \times m_2 \times q}$ satisfying $\Phi_Y^\lambda(\mathcal{X}) \le \Phi_Y^\lambda(\tilde{\mathcal{X}})$, then*

$$\Phi_{\mathrm{Y}}(\mathcal{X}) - \Phi_{\mathrm{Y}}(\tilde{\mathcal{X}}) \le \lambda \bar{r}^{1/2} \sqrt{\mathrm{KL}\left(f(\tilde{\mathcal{X}}), f(\mathcal{X})\right)} \,,$$

where $\bar{r}$ is defined in (21).

Proof. Since $\Phi_Y^\lambda(\mathcal{X}) \le \Phi_Y^\lambda(\tilde{\mathcal{X}})$, we obtain

$$\begin{aligned}\Phi_{\mathrm{Y}}(\mathcal{X}) - \Phi_{\mathrm{Y}}(\tilde{\mathcal{X}}) \le& \lambda \sum_{l=1}^{q} (\|\tilde{X}^l\|_{\sigma,1} - \|X^l\|_{\sigma,1}) \le \lambda \sum_{l=1}^{q} \| \mathcal{P}_{\tilde{X}^l}(X - \tilde{X}^l)\|_{\sigma,1} \,, \\ \le& \lambda \sqrt{2\,\mathrm{rk}(\tilde{\mathcal{X}})} \left(\sum_{l=1}^{q} \|X - \tilde{X}\|_{\sigma,2} \right) \,,\end{aligned}$$

where we have used Lemma 7-(ii) and (iii) and for the last two lines and the definition of K_γ and Lemma 8 to get the result.

Lemma 7. *For any pair of matrices X, $\tilde{X} \in \mathbb{R}^{m_1 \times m_2}$ we have*

(i) $\|X + \mathcal{P}_X^\perp(\tilde{X})\|_{\sigma,1} = \|X\|_{\sigma,1} + \| \mathcal{P}_X^\perp(\tilde{X})\|_{\sigma,1}$,
(ii) $\| \mathcal{P}_X(\tilde{X})\|_{\sigma,1} \le \sqrt{2\,\mathrm{rk}(X)} \|\tilde{X}\|_{\sigma,2}$,
(iii) $\|X\|_{\sigma,1} - \|\tilde{X}\|_{\sigma,1} \le \| \mathcal{P}_X(\tilde{X} - X)\|_{\sigma,1}$.

Proof. If $A, B \in \mathbb{R}^{m_1 \times m_2}$ are two matrices satisfying $\mathcal{S}_i(A) \perp \mathcal{S}_i(B)$, $i = 1, 2$, then $\|A + B\|_{\sigma,1} = \|A\|_{\sigma,1} + \|B\|_{\sigma,1}$. Applying this identity with $A = X$ and $B = \mathcal{P}_X^\perp(\tilde{X})$, we obtain

$$\|X + \mathcal{P}_X^\perp(\tilde{X})\|_{\sigma,1} = \|X\|_{\sigma,1} + \| \mathcal{P}_X^\perp(\tilde{X})\|_{\sigma,1} \,,$$

showing (i).

It follows from the definition that $\mathcal{P}_X(\tilde{X}) = P_{\mathcal{S}_1(X)} \tilde{X} P_{\mathcal{S}_2^\perp(X)} + \tilde{X} P_{\mathcal{S}_2(X)}$. Note that $\mathcal{P}_X$ is an orthogonal projector on $\mathbb{R}^{m_1 \times m_2}$ equipped with the euclidean product $\langle \cdot \,|\, \cdot \rangle$. On the other hand, the Cauchy-Schwarz inequality implies that for any matrix C, $\|C\|_{\sigma,1} \le \sqrt{\mathrm{rk}(C)} \|C\|_{\sigma,2}$. Consequently (ii) follows from

$$\| \mathcal{P}_X(\tilde{X})\|_{\sigma,1} \le \sqrt{2\,\mathrm{rk}(X)} \| \mathcal{P}_X(\tilde{X})\|_{\sigma,2} \ \le \sqrt{2\,\mathrm{rk}(X)} \|\tilde{X}\|_{\sigma,2} \ .$$

Finally, since $\tilde{X} = X + \mathcal{P}_X^\perp(\tilde{X} - X) + \mathcal{P}_X(\tilde{X} - X)$ we have

$$\begin{aligned}\|\tilde{X}\|_{\sigma,1} &\ge \|X + \mathcal{P}_X^\perp(\tilde{X} - X)\|_{\sigma,1} - \| \mathcal{P}_X(\tilde{X} - X)\|_{\sigma,1} \,, \\ &= \|X\|_{\sigma,1} + \| \mathcal{P}_X^\perp(\tilde{X} - X)\|_{\sigma,1} - \| \mathcal{P}_X(\tilde{X} - X)\|_{\sigma,1} \,,\end{aligned}$$

leading to (iii).

Lemma 8. *For any tensor $\mathcal{X}, \tilde{\mathcal{X}} \in \mathbb{R}^{m_1 \times m_2 \times q}$ and p-link function f it holds:*

$$\mathrm{d}_{\mathrm{H}}^2 \left(f(\mathcal{X}), f(\tilde{\mathcal{X}}) \right) \le \mathrm{KL} \left(f(\mathcal{X}), f(\tilde{\mathcal{X}}) \right)$$

Proof. See [26, Lemma 4.2]

Lemma 9. *For any $p, q > 0$ and p-link function f and any $\mathcal{X}, \tilde{\mathcal{X}} \in \mathbb{R}^{m_1 \times m_2 \times q}$ satisfying $\|\mathcal{X}\|_\infty \le \gamma$ and $\|\tilde{\mathcal{X}}\|_\infty \le \gamma$, we get:*

$$\sum_{l=1}^{q} \|X^l - \tilde{X}^l\|_{\sigma,2}^2 \le \frac{m_1 m_2}{K_\gamma} \mathrm{d}_{\mathrm{H}}^2 \left(f(\mathcal{X}), f(\tilde{\mathcal{X}} \right) \le \frac{m_1 m_2}{K_\gamma} \mathrm{KL} \left(f(\mathcal{X}), f(\tilde{\mathcal{X}} \right) .$$

Proof. For $p = 2$ and $q = 1$, it is a consequence of Remark 1 and Lemma 8. Otherwise, the proof follows from the definition (16) of K_γ and Lemma 8.

Lemma 10. *Let $\mathcal{X}, \tilde{\mathcal{X}} \in \mathbb{R}^{m_1 \times m_2 \times q}$ satisfying $\|\mathcal{X}\|_\infty \le \gamma$ and $\|\tilde{\mathcal{X}}\|_\infty \le \gamma$. Assume that $\lambda > 2 \max_{l \in [q]} \|\Sigma_Y^l(\tilde{X})\|_{\sigma,\infty}$ and $\Phi_Y^\lambda(X) \le \Phi_Y^\lambda(\tilde{X})$. Then*

$$\sum_{l=1}^{q} \| \mathcal{P}_{\tilde{X}^l}^{\perp}(X^l - \tilde{X}^l)\|_{\sigma,1} \le 3 \sum_{l=1}^{q} \| \mathcal{P}_{\tilde{X}^l}(X^l - \tilde{X}^l)\|_{\sigma,1} \ , \tag{27}$$

$$\sum_{l=1}^{q} \|X^l - \tilde{X}^l\|_{\sigma,1} \le 4\sqrt{2\,\mathrm{rk}(\tilde{\mathcal{X}})} \sum_{l=1}^{q} \|(X^l - \tilde{X}^l)\|_{\sigma,2} \ , \tag{28}$$

$$\sum_{l=1}^{q} \|X^l - \tilde{X}^l\|_{\sigma,1} \le 4\sqrt{2 m_1 m_2\,\mathrm{rk}(\tilde{\mathcal{X}})/K_\gamma}\, \mathrm{d}_{\mathrm{h}} \left(f(\tilde{\mathcal{X}}), f(\mathcal{X}), \right), \tag{29}$$

$$\sum_{l=1}^{q} \|X^l - \tilde{X}^l\|_{\sigma,1} \le 4\sqrt{2 m_1 m_2\,\mathrm{rk}(\tilde{\mathcal{X}})/K_\gamma} \sqrt{\mathrm{KL} \left(f(\tilde{\mathcal{X}}), f(\mathcal{X}) \right)} . \tag{30}$$

Proof. Since $\Phi_Y^\lambda(\mathcal{X}) \le \Phi_Y^\lambda(\tilde{\mathcal{X}})$, we have

$$\Phi_{\mathrm{Y}}(\tilde{\mathcal{X}}) - \Phi_{\mathrm{Y}}(\mathcal{X}) \ge \lambda \sum_{l=1}^{q} (\|X^l\|_{\sigma,1} - \|\tilde{X}^l\|_{\sigma,1}).$$

For any $X \in \mathbb{R}^{m_1 \times m_2}$, using $X = \tilde{X} + \mathcal{P}_{\tilde{X}}^{\perp}(X - \tilde{X}) + \mathcal{P}_{\tilde{X}}(X - \tilde{X})$, Lemma 7-(i) and the triangular inequality, we get

$$\|X\|_{\sigma,1} \ge \|\tilde{X}\|_{\sigma,1} + \| \mathcal{P}_{\tilde{X}}^{\perp}(X - \tilde{X})\|_{\sigma,1} - \| \mathcal{P}_{\tilde{X}}(X - \tilde{X})\|_{\sigma,1} \ ,$$

which implies

$$\Phi_{\mathrm{Y}}(\tilde{\mathcal{X}}) - \Phi_{\mathrm{Y}}(\mathcal{X}) \ge \lambda \sum_{l=1}^{q} \left(\| \mathcal{P}_{\tilde{X}^l}^{\perp}(X^l - \tilde{X}^l)\|_{\sigma,1} - \| \mathcal{P}_{\tilde{X}^l}(X - \tilde{X}^l)\|_{\sigma,1} \right) . \tag{31}$$

Furthermore by concavity of Φ_{Y} we have

$$\Phi_{\mathrm{Y}}(\tilde{\mathcal{X}}) - \Phi_{\mathrm{Y}}(\mathcal{X}) \le \sum_{l=1}^{q} \langle \Sigma_Y^l(\tilde{\mathcal{X}}) \,|\, \tilde{X}^l - X^l \rangle \ .$$

The duality between $\|\cdot\|_{\sigma,1}$ and $\|\cdot\|_{\sigma,\infty}$ (see for instance [1, Corollary IV.2.6]) leads to

$$\begin{aligned}\Phi_{\mathrm{Y}}(\tilde{\mathcal{X}}) - \Phi_{\mathrm{Y}}(\mathcal{X}) &\le \max_{l\in[q]} \|\Sigma_Y^l(\tilde{X})\|_{\sigma,\infty} \sum_{l=1}^{q} \|\tilde{X}^l - X^l\|_{\sigma,1} \ , \\ &\le \frac{\lambda}{2}\sum_{l=1}^{q}\|\tilde{X}^l - X^l\|_{\sigma,1} \ , \\ &\le \frac{\lambda}{2}\sum_{l=1}^{q}(\|\mathcal{P}^{\perp}_{\tilde{X}^l}(X^l-\tilde{X}^l)\|_{\sigma,1} + \|\mathcal{P}_{\tilde{X}^l}(X^l-\tilde{X}^l)\|_{\sigma,1}) \ , \end{aligned} \tag{32}$$

where we used $\lambda > 2\max_{l\in[q]}\|\Sigma_Y^l(\tilde{X})\|_{\sigma,\infty}$ in the second line. Then combining (31) with (32) gives (27). Since for any $l \in [q]$, $X^l - \tilde{X}^l = \mathcal{P}^{\perp}_{\tilde{X}^l}(X^l - \tilde{X}^l) + \mathcal{P}_{\tilde{X}^l}(X^l - \tilde{X}^l)$, using the triangular inequality and (27) yields

$$\sum_{l=1}^{q}\|X^l - \tilde{X}^l\|_{\sigma,1} \le 4\|\mathcal{P}_{\tilde{X}^l}(X^l - \tilde{X}^l)\|_{\sigma,1}. \tag{33}$$

Combining (33) and (27) immediately leads to (28) and (29) is a consequence of (28) and the definition of K_γ. The statement (30) follows from (29) and Lemma 8.

Lemma 11. *Under H2 we have*

$$\mathrm{D}\left(f(\bar{\mathcal{X}}), f(\mathcal{X})\right) \ge \frac{1}{\mu}\,\mathrm{KL}\left(f(\bar{\mathcal{X}}), f(\mathcal{X})\right) \ .$$

where $\mathrm{D}(\cdot,\cdot)$ *is defined in* (22).

Proof. Follows from

$$\begin{aligned}\mathrm{D}\left(f(\bar{\mathcal{X}}), f(\mathcal{X})\right) &= \frac{1}{n}\sum_{i=1}^{n}\sum_{\substack{k\in[m_1]\\ l\in[m_2]}}\sum_{j\in[p]}\pi_{k,l}\left[f^j(\bar{\mathcal{X}}_{k,l})\log\left(\frac{f^j(\bar{\mathcal{X}}_{k,l})}{f^j(\mathcal{X}_{k,l})}\right)\right] \ , \\ &\ge \frac{1}{\mu m_1 m_2}\sum_{\substack{k\in[m_1]\\ l\in[m_2]}}\sum_{j\in[p]}\left[f^j(\bar{\mathcal{X}}_{k,l})\log\left(\frac{f^j(\bar{\mathcal{X}}_{k,l})}{f^j(\mathcal{X}_{k,l})}\right)\right] \ .\end{aligned}$$

Lemma 12. *Assume that* $\lambda \ge \bar{\Sigma}$. *Let* $\alpha > 1$, $\beta > 0$ *and* $0 < \eta < 1/2\alpha$. *Then with probability at least*

$$1 - 2(\exp(-n\eta^2\log(\alpha)\beta^2/(4M_\gamma^2))/(1-\exp(-n\eta^2\log(\alpha)\beta^2/(4M_\gamma^2)))$$

we have for all $\mathcal{X} \in \mathcal{C}_\beta(r)$:

$$|\,\Phi_{\mathrm{Y}}(\mathcal{X}) - \Phi_{\mathrm{Y}}(\bar{\mathcal{X}}) - \mathrm{D}\left(f(\bar{\mathcal{X}}), f(\mathcal{X})\right)| \le \frac{\mathrm{D}\left(f(\bar{\mathcal{X}}), f(\mathcal{X})\right)}{2} + \epsilon(r,\alpha,\eta) \ ,$$

where

$$\epsilon(r,\alpha,\eta) := \frac{4q^2 L_\gamma^2 r}{1/(2\alpha)-\eta}(\mathbb{E}\|\Sigma_R\|_{\sigma,\infty})^2 \ , \tag{34}$$

and $\mathcal{C}_\beta(r)$ *is defined in* (24).

Proof. The proof is adapted from [24, Theorem 1] and [18, Lemma 12]. We use a peeling argument combined with a sharp deviation inequality detailed in Lemma 13, Consider the events

$$\mathcal{B} := \Bigg\{\exists \mathcal{X} \in \mathcal{C}_\beta(r) \Bigg| \\ |\,\Phi_{\mathrm{Y}}(\mathcal{X}) - \Phi_{\mathrm{Y}}(\bar{\mathcal{X}}) - \mathrm{D}\left(f(\bar{\mathcal{X}}), f(\mathcal{X})\right)| > \frac{\mathrm{D}\left(f(\bar{\mathcal{X}}), f(\mathcal{X})\right)}{2} + \epsilon(r,\alpha,\eta)\Bigg\} \ ,$$

and

$$\mathcal{S}_l := \left\{\mathcal{X} \in \mathcal{C}_\beta(r) | \alpha^{l-1}\beta < \mathrm{D}\left(f(\bar{\mathcal{X}}), f(\mathcal{X})\right) < \alpha^l \beta\right\} \ .$$

Let us also define the set

$$\mathcal{C}_\beta(r,t) = \left\{\mathcal{X} \in \mathbb{R}^{m_1 \times m_2} |\ \mathcal{X} \in \mathcal{C}_\beta(r),\ \mathrm{D}\left(f(\bar{\mathcal{X}}), f(\mathcal{X})\right) \le t\right\} \ ,$$

and

$$Z_t := \sup_{\mathcal{X} \in \mathcal{C}_\beta(r,t)} |\,\Phi_{\mathrm{Y}}(\mathcal{X}) - \Phi_{\mathrm{Y}}(\bar{\mathcal{X}}) - \mathrm{D}\left(f(\bar{\mathcal{X}}), f(\mathcal{X})\right)| \ , \tag{35}$$

Then for any $\mathcal{X} \in \mathcal{B} \cap \mathcal{S}_l$ we have

$$|\,\Phi_{\mathrm{Y}}(\mathcal{X}) - \Phi_{\mathrm{Y}}(\bar{\mathcal{X}}) - \mathrm{D}\left(f(\bar{\mathcal{X}}), f(\mathcal{X})\right)| > \frac{1}{2}\alpha^{l-1}\beta + \epsilon(r,\alpha,\eta) \ ,$$

Moreover by definition of $\mathcal{S}_l$, $\mathcal{X} \in \mathcal{C}_\beta(r, \alpha^l\beta)$. Therefore

$$\mathcal{B} \cap \mathcal{S}_l \subset \mathcal{B}_l := \{Z_{\alpha^l\beta} > \frac{1}{2\alpha}\alpha^l\beta + \epsilon(r,\alpha,\eta)\} \ ,$$

If we now apply the union bound and Lemma 13 we get

$$\mathbb{P}(\mathcal{B}) \le \sum_{l=1}^{+\infty}\mathbb{P}(\mathcal{B}_l) \le \sum_{l=1}^{+\infty}\exp\left(-\frac{n\eta^2(\alpha^l\beta)^2}{8M_\gamma^2}\right) \le \frac{\exp(-\frac{n\eta^2\log(\alpha)\beta^2}{4M_\gamma^2})}{1-\exp(-\frac{n\eta^2\log(\alpha)\beta^2}{4M_\gamma^2})} \ ,$$

where we used $x \le \mathrm{e}^x$ in the second inequality.

Lemma 13. *Assume that* $\lambda \ge \bar{\Sigma}$. *Let* $\alpha > 1$ *and* $0 < \eta < \frac{1}{2\alpha}$. *Then we have*

$$\mathbb{P}\left(Z_t > t/(2\alpha) + \epsilon(r,\alpha,\beta)\right) \le \exp\left(-n\eta^2 t^2/(8M_\gamma^2)\right) \ , \tag{36}$$

where Z_t *and* $\epsilon(r,\alpha,\eta)$ *are defined in* (35) *and* (34), *respectively.*

Proof. Using Massart's inequality ([22, Theorem 9]) we get for $0 < \eta < 1/(2\alpha)$

$$\mathbb{P}(Z_t > \mathbb{E}[Z_t] + \eta t) \leq \exp\left(-\eta^2 n t^2/(8M_\gamma^2)\right) \ . \tag{37}$$

By using the standard symmetrization argument, we get

$$\mathbb{E}[Z_t] \leq 2\mathbb{E}\left[\sup_{\mathcal{X}\in\mathcal{C}_\beta(r,t)} \left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i \sum_{j=1}^p \mathbb{1}_{\{Y_i=j\}} \log\left(\frac{f^j(\mathcal{X}_i)}{f^j(\bar{\mathcal{X}}_i)}\right)\right|\right] \ ,$$

where $\boldsymbol{\varepsilon} := (\varepsilon_i)_{1\leq i\leq n}$ is a Rademacher sequence which is independent from $(Y_i)_{1\leq i\leq n}$ and $(E_i)_{1\leq i\leq n}$. H5 yields

$$\mathbb{E}[Z_t] \leq \sum_{l=1}^q 2\mathbb{E}\left[\sup_{\mathcal{X}\in\mathcal{C}_\beta(r,t)} \left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i \sum_{j=1}^p \mathbb{1}_{\{Y_i=j\}} \log\left(\frac{g_l^j(X_i^l)}{g_l^j(\bar{X}_i^l)}\right)\right|\right] \ .$$

Since for any $i \in [n]$, the function

$$\phi_i(x) := \frac{1}{L_\gamma}\sum_{j=1}^p \mathbb{1}_{\{Y_i=j\}} \log\left(\frac{g_l^j(x+\bar{X}_i^l)}{g_l^j(\bar{X}_i^l)}\right)$$

is a contraction satisfying $\phi_i(0) = 0$, the contraction principle ([21, Theorem 4.12]) and the fact that $(\varepsilon_i)_{i=1}^n$ is independent from $(Y_i)_{i=1}^n$ and $(\omega_i)_{i=1}^n$ yields

$$\mathbb{E}[Z_t] \leq 4L_\gamma \sum_{l=1}^q \mathbb{E}\left[\sup_{\mathcal{X}\in\mathcal{C}_\beta(r,t)} \left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i \langle X^l - \bar{X}^l \mid E_i\rangle\right|\right] =$$

Denoting $\Sigma_R := n^{-1}\sum_{i=1}^n \varepsilon_i E_i$ and the duality, the previous inequality implies

$$\begin{aligned}\mathbb{E}[Z_t] &\leq 4L_\gamma \sum_{l=1}^q \mathbb{E}\left[\sup_{\mathcal{X}\in\mathcal{C}_\beta(r,t)} \left|\langle X^l - \bar{X}^l \mid \Sigma_R\rangle\right|\right] \\ &\leq 4L_\gamma \sum_{l=1}^q \mathbb{E}\left[\sup_{\mathcal{X}\in\mathcal{C}_\beta(r,t)} \|X^l - \bar{X}^l\|_{\sigma,1}\|\Sigma_R\|_{\sigma,\infty}\right] \leq 4qL_\gamma \mathbb{E}[\|\Sigma_R\|_{\sigma,\infty}]\sqrt{rt} \ ,\end{aligned}$$

where we have the definition of $\mathcal{C}_\beta(r,t)$ for the last inequality. Plugging into (37) gives

$$\mathbb{P}(Z_t > 4qL_\gamma\mathbb{E}[\|\Sigma_R\|_{\sigma,\infty}]\sqrt{rt} + \eta t) \leq \exp\left(-\eta^2 n t^2/(8M_\gamma^2)\right) \ .$$

The proof is concluded by noting that, since for any $a, b \in \mathbb{R}$ and $c > 0$, $ab \leq (a^2/c + cb^2)/2$,

$$4qL_\gamma\mathbb{E}[\|\Sigma_R\|_{\sigma,\infty}]\sqrt{rt} \leq \frac{1}{1/(2\alpha)-\eta} 4q^2L_\gamma^2 r\mathbb{E}[\|\Sigma_R\|_{\sigma,\infty}]^2 + (1/(2\alpha)-\eta)t \ .$$

5.2. Proof of Theorem 5

Proof. By Theorem 5 it suffices to control $\|\bar{\Sigma}^l\|_{\sigma,\infty}$ and $\mathbb{E}[\|\Sigma_R\|_{\sigma,\infty}]$. For any $l \in [q]$, by definition

$$\bar{\Sigma}^l = -\frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j=1}^{p}\mathbb{1}_{\{Y_i=j\}}\frac{\partial_l f^j(\bar{\mathcal{X}}_i)}{f^j(\bar{\mathcal{X}}_i)}\right)E_i \ ,$$

with ∂_l designating the partial derivative against the l-th variable. The sequence of matrices

$$Z_i := \left(\sum_{j=1}^{p}\mathbb{1}_{\{Y_i=j\}}\frac{\partial_l f^j(\bar{\mathcal{X}}_i)}{f^j(\bar{\mathcal{X}}_i)}\right)E_i = \left(\sum_{j=1}^{p}\mathbb{1}_{\{Y_i=j\}}\frac{(g_l^j)'(\bar{X}_i^l)}{g_l^j(\bar{X}_i^l)}\right)E_i$$

satisfies $\mathbb{E}[Z_i]=0$ (as any score function) and $\|Z_i\|_{\sigma,\infty} \le L_\gamma$.
Noticing $e_k(e'_{k'})^\top(e_k(e'_{k'})^\top)^\top = e_k(e'_{k'})^\top$ we also get

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[Z_iZ_i^\top] = \sum_{k=1}^{m_1}\left(\sum_{k'=1}^{m_2}\pi_{k,k'}\left(\sum_{j=1}^{p}f^j(\bar{\mathcal{X}}_{k,k'})\left(\frac{\partial_l f^j(\bar{\mathcal{X}}_{k,k'})}{f^j(\bar{\mathcal{X}}_{k,k'})}\right)^2\right)\right)e_k(e'_k)^\top \ ,$$

which is diagonal. We recall the definition $C_{k'} = \sum_{k=1}^{m_1}\pi_{k,k'}$ and $R_k = \sum_{k'=1}^{m_2}\pi_{k,k'}$ for any $k' \in [m_2]$, $k \in [m_1]$. Since

$$\left(\frac{\partial_l f^j(\bar{\mathcal{X}}_{k,k'})}{f^j(\bar{\mathcal{X}}_{k,k'})}\right)^2 \le L_\gamma^2 \ ,$$

and $(f^j(\bar{\mathcal{X}}_{k,k'}))_{j=1}^p$ is a probability distribution, we obtain

$$\left\|\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}Z_iZ_i^\top\right]\right\|_{\sigma,\infty} \le L_\gamma^2\left\|\mathrm{diag}((R_k)_{k=1}^{m_1})\right\|_{\sigma,\infty} \le L_\gamma^2\frac{\nu}{m} \ ,$$

were we have H 3 for the last inequality. Using a similar argument we get $\|\mathbb{E}[\sum_{i=1}^n Z_i^\top Z_i]\|_{\sigma,\infty}/n \le L_\gamma^2\nu/m$. Therefore, Proposition 14 applied with $t = \log(d)$, $U = L_\gamma$ and $\sigma_Z^2 = L_\gamma^2\nu/m$ yields with at least probability $1-1/d$,

$$\left\|\Sigma_Y^l(\tilde{X})\right\|_{\sigma,\infty} \le (1+\sqrt{3})L_\gamma \max\left\{\sqrt{\frac{2\nu\log(d)}{mn}}, \frac{2}{3}\frac{\log(d)}{n}\right\} \ . \tag{38}$$

With the same analysis for $\Sigma_R := \frac{1}{n}\sum_{i=1}^n \varepsilon_i E_i$ and by applying Lemma 15 with $U=1$ and $\sigma_Z^2 = \frac{\nu}{m}$, for $n \ge n^* := m\log(d)/(9\nu)$ it holds:

$$\mathbb{E}\left[\|\Sigma_R\|_{\sigma,\infty}\right] \le c^*\sqrt{\frac{2e\nu\log(d)}{mn}} \ . \tag{39}$$

Assuming $n \geq 2m\log(d)/(9\nu)$, implies $n \geq n^*$ and (39) is therefore satisfied. Since it also implies $\sqrt{2\nu\log(d)/(mn)} \geq 2\log(d)/(3n)$, the second term of (38) is negligible. Consequently taking $\lambda \geq 2(1+\sqrt{3})L_\gamma\sqrt{2\nu\log(d)/(mn)}$, a union bound argument ensures that $\lambda > 2\max_{l\in[q]}\|\Sigma_Y^l(\tilde{X})\|_{\sigma,\infty}$ with probability at least $1-q/d$.

By taking λ, β and n as in Theorem 5 statement, with probability larger than $1-(2+q)/d$, Theorem 4 result holds when replacing $\mathbb{E}\|\Sigma_R\|_{\sigma,\infty}$ by its upper bound (39). Using the inequality $(a+b)^2 \leq 2(a^2+b^2)$ yields the result with $\bar{c} = 24832$.

Proposition 14. *Consider a finite sequence of independent random matrices $(Z_i)_{1\leq i\leq n} \in \mathbb{R}^{m_1\times m_2}$ satisfying $\mathbb{E}[Z_i]=0$ and for some $U>0$, $\|Z_i\|_{\sigma,\infty} \leq U$ for all $i=1,\dots,n$. Then for any $t>0$*

$$\mathbb{P}\left(\left\|\frac{1}{n}\sum_{i=1}^n Z_i\right\|_{\sigma,\infty} > t\right) \leq d\exp\left(-\frac{nt^2/2}{\sigma_Z^2+Ut/3}\right) ,$$

where $d = m_1+m_2$ and

$$\sigma_Z^2 := \max\left\{\left\|\frac{1}{n}\sum_{i=1}^n \mathbb{E}[Z_iZ_i^\top]\right\|_{\sigma,\infty}, \left\|\frac{1}{n}\sum_{i=1}^n \mathbb{E}[Z_i^\top Z_i]\right\|_{\sigma,\infty}\right\} .$$

In particular it implies that with at least probability $1-\mathrm{e}^{-t}$

$$\left\|\frac{1}{n}\sum_{i=1}^n Z_i\right\|_{\sigma,\infty} \leq c^*\max\left\{\sigma_Z\sqrt{\frac{t+\log(d)}{n}}, \frac{U(t+\log(d))}{3n}\right\} ,$$

with $c^* = 1+\sqrt{3}$.

Proof. The first claim of the proposition is Bernstein's inequality for random matrices (see for example [25, Theorem 1.6]). Solving the equation (in t) $-\frac{nt^2/2}{\sigma_Z^2+Ut/3}+\log(d) = -v$ gives with at least probability $1-\mathrm{e}^{-v}$

$$\left\|\frac{1}{n}\sum_{i=1}^n Z_i\right\|_{\sigma,\infty} \leq \frac{1}{n}\left[\frac{U}{3}(v+\log(d)) + \sqrt{\frac{U^2}{9}(v+\log(d))^2 + 2n\sigma_Z^2(v+\log(d))}\right] ,$$

we conclude the proof by distinguishing the two cases $n\sigma_Z^2 \leq (U^2/9)(v+\log(d))$ or $n\sigma_Z^2 > (U^2/9)(v+\log(d))$.

Lemma 15. *Let $h \geq 1$. With the same assumptions as Proposition 14, assume $n \geq (U^2\log(d))/(9\sigma_Z^2)$ then the following holds:*

$$\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n Z_i\right\|_{\sigma,\infty}^h\right] \leq \left(\frac{2ehc^{*2}\sigma_Z^2\log(d)}{n}\right)^{h/2} ,$$

with $c^* = 1+\sqrt{3}$.

Proof. The proof is adapted from [18, Lemma 6]. Define $t^* := (9n\sigma_Z^2)/U^2 - \log(d)$ the value of t for which the two bounds of Proposition 14 are equal. Let $\nu_1 := n/(\sigma_Z^2 c^{*2})$ and $\nu_2 := 3n/(Uc^*)$ then, from Proposition 14 we have

$$\mathbb{P}\left(\left\|\frac{1}{n}\sum_{i=1}^n Z_i\right\|_{\sigma,\infty} > t\right) \leq d\exp(-\nu_1 t^2) \text{ for } t \leq t^* ,$$
$$\mathbb{P}\left(\left\|\frac{1}{n}\sum_{i=1}^n Z_i\right\|_{\sigma,\infty} > t\right) \leq d\exp(-\nu_2 t) \text{ for } t \geq t^* ,$$

Let $h \geq 1$, then

$$\begin{aligned}
&\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n Z_i\right\|_{\sigma,\infty}^h\right] \leq \mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n Z_i\right\|_{\sigma,\infty}^{2h\log(d)}\right]^{1/(2\log(d))} ,\\
&\leq \left(\int_0^{+\infty} \mathbb{P}\left(\left\|\frac{1}{n}\sum_{i=1}^n Z_i\right\|_{\sigma,\infty} > t^{1/(2h\log(d))}\right)dt\right)^{1/(2\log(d))} ,\\
&\leq d^{(2h\log(d))^{-1}}\left(\int_0^{+\infty}\exp(-\nu_1 t^{2/(2h\log(d))}) + \exp(-\nu_2 t^{1/(2h\log(d))})dt\right)^{1/(2\log(d))} ,\\
&\leq \sqrt{e}\Big(h\log(d)\nu_1^{-h\log(d)}\Gamma(h\log(d)) + 2h\log(d)\nu_2^{-2h\log(d)}\Gamma(2h\log(d))\Big)^{1/(2\log(d))} ,
\end{aligned}$$

where we used Jensen's inequality for the first line. Since Gamma-function satisfies for $x \geq 2$, $\Gamma(x) \leq (\frac{x}{2})^{x-1}$ (see [17, Proposition 12]) we have

$$\begin{aligned}
&\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n Z_i\right\|_{\sigma,\infty}^h\right] \leq \\
&\sqrt{e}\Big((h\log(d))^{h\log(d)}\nu_1^{-h\log(d)}2^{1-h\log(d)} + 2(h\log(d))^{2h\log(d)}\nu_2^{-2h\log(d)}\Big)^{1/(2\log(d))} .
\end{aligned}$$

For $n \geq (U^2\log(d))/(9\sigma_Z^2)$ we have $\nu_1\log(d) \leq \nu_2^2$ and therefore we get

$$\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n Z_i\right\|_{\sigma,\infty}^h\right] \leq \left(\frac{2eh\log(d)}{\nu_1}\right)^{h/2} .$$

5.3. Proof of Theorem 3

Proof. Let h be the following function

$$h(\kappa) = \min\left\{1/2, \sqrt{\alpha\, rMK_{(1-\kappa)\gamma}^{-1}}/(8\gamma\sqrt{n})\right\} . \tag{40}$$

Since $0 < h(\kappa) \leq 1/2$ and h is continuous, there exists a fixed point $\kappa^* \in (0, 1/2]$:

$$h(\kappa_*) = \kappa_* \; . \tag{41}$$

For notational convenience, the dependence of κ_* in r, M and n is implicit. We start with a packing set construction, inspired by [7]. Assume *w.l.o.g.,* that $m_1 \geq m_2$. For $\kappa \leq 1$, define

$$\mathcal{L} = \left\{ L = (l_{ij}) \in \mathbb{R}^{m_1 \times r} : l_{ij} \in \left\{ -\frac{\kappa\gamma}{2}, \frac{\kappa\gamma}{2} \right\}, \forall i \in [m_1], \, \forall j \in [r] \right\},$$

and consider the associated set of block matrices

$$\mathcal{L}' = \left\{ L' = (\; L \mid \cdots \mid L \mid O \;) \in \mathbb{R}^{m_1 \times m_2} : L \in \mathcal{L} \right\},$$

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.

Remark 3. In the case $m_1 < m_2$, we only need to change the construction of the low rank component of the test set. We first build a matrix $\tilde{L} \in \mathbb{R}^{r \times m_2}$ with entries in $\left\{ -\frac{\kappa\gamma}{2}, \frac{\kappa\gamma}{2} \right\}$ and then we replicate this matrix to obtain a block matrix L of size $m_1 \times m_2$.

Let $\mathbf{I}_{m_1 \times m_2}$ denote the $m_1 \times m_2$ matrix of ones. The Varshamov-Gilbert bound ([26, Lemma 2.9]) guarantees the existence of a subset $\mathcal{L}'' \subset \mathcal{L}'$ with cardinality $\mathrm{Card}(\mathcal{L}'') \geq 2^{(rM)/8} + 1$ containing the matrix $(\kappa\gamma/2)\,\mathbf{I}_{m_1 \times m_2}$ and such that, for any two distinct elements X_1 and X_2 of $\mathcal{L}''$,

$$\|X_1 - X_2\|_2^2 \geq \frac{Mr\,\kappa^2\gamma^2}{8} \left\lfloor \frac{m_2}{r} \right\rfloor \geq \frac{m_1 m_2\,\kappa^2\gamma^2}{16} \, . \tag{42}$$

Then, we construct the packing set $\mathcal{A}$ by setting

$$\mathcal{A} = \left\{ L + \frac{(2-\kappa)\gamma}{2} \mathbf{I}_{m_1 \times m_2} \; : \; L \in \mathcal{L}'' \right\}.$$

By construction, any element of $\mathcal{A}$ as well as the difference of any two elements of $\mathcal{A}$ has rank at most r, the entries of any matrix in $\mathcal{A}$ take values in $[0, \gamma]$, and $X^0 = \gamma \mathbf{I}_{m_1 \times m_2}$ belongs to $\mathcal{A}$. Thus, $\mathcal{A} \subset \mathcal{F}(r, \gamma)$. Note that $\mathcal{A}$ has the same size as $\mathcal{L}''$ and it also satisfies the same bound on pairwise distances, i.e. for any two distinct elements X_1 and X_2 of $\mathcal{A}$, (42) is satisfied.

For some $X \in \mathcal{A}$, we now estimate the Kullback-Leibler divergence $D\left(\mathbb{P}_{X^0} \| \mathbb{P}_X\right)$ between probability measures $\mathbb{P}_{X^0}$ and $\mathbb{P}_X$. By independence of the observations $(Y_i, \omega_i)_{i=1}^n$,

$$D\left(\mathbb{P}_{X^0} \| \mathbb{P}_X\right) = n\mathbb{E}_{\omega_1}\left[\sum_{j=1}^{2} f^j(X^0_{\omega_1}) \log\left(\frac{f^j(X^0_{\omega_1})}{f^j(X_{\omega_1})} \right) \right] \, .$$

Since $X^0_{\omega_1} = \gamma$ and either $X_{\omega_1} = X^0_{\omega_1}$ or $X_{\omega_1} = (1-\kappa)\gamma$, by Lemma 16 we get

$$D\left(\mathbb{P}_{X^0} \| \mathbb{P}_X\right) \leq \frac{n\left[f^1(\gamma) - f^1\left((1-\kappa)\gamma\right) \right]^2}{f^1\left((1-\kappa)\gamma\right)\left[1 - f^1\left((1-\kappa)\gamma\right)\right]}.$$

From the mean value theorem, for some $\xi \in [(1-\kappa)\gamma, \gamma]$ we have

$$D\left(\mathbb{P}_{X^0} \| \mathbb{P}_X\right) \le \frac{n\{(f^1)'(\xi)\}^2(\kappa\gamma)^2}{f^1\left((1-\kappa)\gamma\right)\left[1 - f^1\left((1-\kappa)\gamma\right)\right]} .$$

Using H4, the function $(f^1)'$ is decreasing and the latter inequality implies

$$D\left(\mathbb{P}_{X^0} \| \mathbb{P}_X\right) \le 8\, n(\kappa\gamma)^2 g((1-\kappa)\gamma) , \tag{43}$$

where g is defined in (6). From (43) and plugging $\kappa = \kappa^*$ defined in eq. (41), we get

$$D\left(\mathbb{P}_{X^0} \| \mathbb{P}_X\right) \le \frac{\alpha r M}{8} \le \alpha \log_2(rM/8) ,$$

which implies that

$$\frac{1}{\operatorname{Card}(\mathcal{A}) - 1} \sum_{X \in \mathcal{A}} D\left(\mathbb{P}_{X^0} \| \mathbb{P}_X\right) \ \le\ \alpha \log\left(\operatorname{Card}(\mathcal{A}) - 1\right) . \tag{44}$$

Using (42) and (44), [26, Theorem 2.5] implies

$$\inf_{\hat{X}} \sup_{\bar{X} \in \mathcal{F}(r,\gamma)} \mathbb{P}_{\bar{X}} \left(\frac{\|\hat{X} - \bar{X}\|_2^2}{m_1 m_2} > c \min\left\{\gamma^2, \frac{Mr}{n\, K_{(1-\kappa^*)\gamma}}\right\} \right) \ \ge\ \delta \tag{45}$$

for some universal constants $c > 0$ and $\delta \in (0,1)$.

Lemma 16. *Let us consider $x, y \in (0,1)$ and*

$$k(x,y) := x \log(x/y) + (1-x)\log((1-x)/1-y) .$$

Then the following holds

$$k(x,y) \le \frac{(x-y)^2}{y(1-y)} .$$

Proof. The proof is taken from [7, Lemma 4]. Since $k(x,y) = k(1-x, 1-y)$, *w.l.o.g.,* we may assume $y > x$. The function $g(t) = k(x, x+t)$ satisfies $g'(t) = t/[(x+t)(1-x-t)]$ and $g''(t) \ge 0$. Therefore the mean value Theorem gives $g(y-x) - g(0) \le g'(y-x)(y-x)$ which yields the result.

References

[1] R. Bhatia. *Matrix analysis*, volume 169 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, 1997. MR1477662

[2] J. Bobadilla, F. Ortega, A. Hernando, and A. Gutiérrez. Recommender systems survey. *Knowledge-Based Systems*, 46(0):109 – 132, 2013.

[3] J-F. Cai, E. J. Candès, and Z. Shen. A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization*, 20(4):1956–1982, 2010. MR2600248

[4] T. T. Cai and W-X. Zhou. Matrix completion via max-norm constrained optimization. *CoRR*, abs/1303.0341, 2013.
[5] T. T. Cai and W-X. Zhou. A max-norm constrained minimization approach to 1-bit matrix completion. *J. Mach. Learn. Res.*, 14:3619–3647, 2013. MR3159403
[6] E. J. Candès and Y. Plan. Matrix completion with noise. *Proceedings of the IEEE*, 98(6):925–936, 2010.
[7] M. A. Davenport, Y. Plan, E. van den Berg, and M. Wootters. 1-bit matrix completion. *Information and Inference*, 3(3):189–223, 2014. MR3311452
[8] M. Dudík, Z. Harchaoui, and J. Malick. Lifted coordinate descent for learning with trace-norm regularization. In *AISTATS*, 2012.
[9] M. Fazel. *Matrix rank minimization with applications*. PhD thesis, Stanford University, 2002.
[10] R. Foygel, R. Salakhutdinov, O. Shamir, and N. Srebro. Learning with the weighted trace-norm under arbitrary sampling distributions. In *NIPS*, pages 2133–2141, 2011.
[11] G. H. Golub and C. F. van Loan. *Matrix computations*. Johns Hopkins University Press, Baltimore, MD, fourth edition, 2013. MR3024913
[12] D. Gross. Recovering low-rank matrices from few coefficients in any basis. *Information Theory, IEEE Transactions on*, 57(3):1548–1566, 2011. MR2815834
[13] S. Gunasekar, P. Ravikumar, and J. Ghosh. Exponential family matrix completion under structural constraints. *ICML*, 2014.
[14] J. Hui, L. Chaoqiang, S. Zuowei, and X. Yuhong. Robust video denoising using low rank matrix completion. *CVPR*, 0:1791–1798, 2010.
[15] L. Ji, P. Musialski, P. Wonka, and Y. Jieping. Tensor completion for estimating missing values in visual data. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(1):208–220, 2013.
[16] R. H. Keshavan, A. Montanari, and S. Oh. Matrix completion from noisy entries. *J. Mach. Learn. Res.*, 11:2057–2078, 2010. MR2678022
[17] O. Klopp. Rank penalized estimators for high-dimensional matrices. *Electron. J. Stat.*, 5:1161–1183, 2011. MR2842903
[18] O. Klopp. Noisy low-rank matrix completion with general sampling distribution. *Bernoulli*, 2(1):282–303, 02 2014. MR3160583
[19] V. Koltchinskii, A. B. Tsybakov, and K. Lounici. Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *Ann. Statist.*, 39(5):2302–2329, 2011. MR2906869
[20] Y. Koren, R. Bell, and C. Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37, 2009.
[21] M. Ledoux and M. Talagrand. *Probability in Banach spaces*, volume 23 of *Ergebnisse der Mathematik und ihrer Grenzgebiete (3) [Results in Mathematics and Related Areas (3)]*. Springer-Verlag, Berlin, 1991. Isoperimetry and processes. MR1102015
[22] P. Massart. About the constants in Talagrand's concentration inequalities for empirical processes. *Ann. Probab.*, 28(2):863–884, 2000. MR1782276

[23] R. Mazumder, T. Hastie, and R. Tibshirani. Spectral regularization algorithms for learning large incomplete matrices. *J. Mach. Learn. Res.*, 11:2287–2322, 2010. MR2719857

[24] S. Negahban and M. J. Wainwright. Restricted strong convexity and weighted matrix completion: optimal bounds with noise. *J. Mach. Learn. Res.*, 13:1665–1697, 2012. MR2930649

[25] J. A. Tropp. User-friendly tail bounds for sums of random matrices. *Found. Comput. Math.*, 12(4):389–434, 2012. MR2946459

[26] A. B. Tsybakov. *Introduction to nonparametric estimation.* Springer Series in Statistics. Springer, New York, 2009. MR2724359

[27] H. Xu, W. Jiasong, W. Lu, C. Yang, L. Senhadji, and H. Shu. Linear total variation approximate regularized nuclear norm optimization for matrix completion. *Abstr. Appl. Anal.*, pages Art. ID 765782, 8, 2014. MR3216075

[28] Y. Yang, J. Ma, and S. Osher. Seismic data reconstruction via matrix completion. *Inverse Probl. Imaging*, 7(4):1379–1392, 2013. MR3180685

[29] Y. Koren and J. Sill. Ordrec: An ordinal model for predicting personalized item rating distributions. In *Proceedings of the Fifth ACM Conference on Recommender Systems*, RecSys '11, pages 117–124, New York, NY, USA, 2011. ACM.