JMLR: Workshop and Conference Proceedings vol 40:1–20, 2015

Low Rank Matrix Completion with Exponential Family Noise

Jean Lafond JEAN.LAFOND@TELECOM-PARISTECH.FR
Institut Mines-Télécom, Télécom ParisTech, CNRS LTCI

Abstract

The matrix completion problem consists in reconstructing a matrix from a sample of entries, possibly observed with noise. A popular class of estimator, known as nuclear norm penalized estimators, are based on minimizing the sum of a data fitting term and a nuclear norm penalization. Here, we investigate the case where the noise distribution belongs to the exponential family and is sub-exponential. Our framework allows for a general sampling scheme. We first consider an estimator defined as the minimizer of the sum of a log-likelihood term and a nuclear norm penalization and prove an upper bound on the Frobenius prediction risk. The rate obtained improves on previous works on matrix completion for exponential family. When the sampling distribution is known, we propose another estimator and prove an oracle inequality *w.r.t.* the Kullback-Leibler prediction risk, which translates immediately into an upper bound on the Frobenius prediction risk. Finally, we show that all the rates obtained are minimax optimal up to a logarithmic factor.

Keywords: Low rank matrix estimation; matrix completion; exponential family model; nuclear norm

1. Introduction

In the matrix completion problem one aims at recovering a matrix, based on partial and noisy observations of its entries. This problem arises in a wide range of practical situations such as collaborative filtering or quantum tomography (see Srebro and Salakhutdinov (2010) or Gross (2011) for instance). In typical applications, the number of observations is usually much smaller than the total number of entries, so that some structural constraints are needed to recover the whole matrix efficiently.

More precisely, we consider an $m_1 \times m_2$ real matrix $\bar{X}$ and observe n samples of the form $(Y_i, \omega_i)_{i=1}^n$, with $(\omega_i)_{i=1}^n \in ([m_1] \times [m_2])^n$ an $i.i.d.$ sequence of indexes and $(Y_i)_{i=1}^n \in \mathbb{R}^n$ a sequence of observations which is assumed to be $i.i.d.$ conditionally to the entries $(\bar{X}_{\omega_i})_{i=1}^n$. To recover the unknown parameter matrix $\bar{X}$, a popular class of methods, known as penalized nuclear norm estimators, are based on minimizing the sum of a data fitting term and a nuclear norm penalization term. These estimators have been extensively studied over the past decade and strong statistical guarantees can be proved in some particular settings. When the conditional distribution $Y_i|\bar{X}_{\omega_i}$ is additive and sub-exponential it can be shown that the unknown matrix can be recovered efficiently, provided that it is low rank or approximately low rank, see Candès and Plan (2010); Keshavan et al. (2010); Koltchinskii et al. (2011); Negahban and Wainwright (2012); Cai and Zhou (2013a); Klopp (2014). In that case, the prediction error satisfies with high probability

$$\frac{\|\hat{X} - \bar{X}\|_{\sigma,2}^2}{m_1 m_2} = \mathcal{O}\left(\frac{(m_1 + m_2)\,\mathrm{rk}(\bar{X})\log(m_1 + m_2)}{n}\right), \tag{1}$$

with $\hat{X}$ denoting the estimator, $\|\cdot\|_{\sigma,2}$ the Frobenius norm and $\mathrm{rk}(\cdot)$ the rank of a matrix. It has been proved by Koltchinskii et al. (2011) that this rate is actually minimax optimal up to a logarithmic factor.

Although very common in practice, discrete distributions have received less attention. The analysis of a logistic noise was first addressed by Davenport et al. (2012). It was later considered by Cai and Zhou (2013b), Lafond et al. (2014) and Klopp et al. (2014) who have shown that the prediction error is also of the order of (1), for log-likelihood estimators, regularized with nuclear norm. Gunasekar et al. (2014) have investigated the case of distributions belonging to the exponential family, which is rich enough to encompass both continuous and discrete distributions (Gaussian, exponential, Poisson, logistic, etc.). They provide (see their Corollary 1) an upper bound for the prediction error when the noise is sub-Gaussian and the sampling uniform. However, this bound is of the form

$$\frac{\|\hat{X} - \bar{X}\|_{\sigma,2}^2}{m_1 m_2} = \mathcal{O}\left(\alpha^{*2} \frac{(m_1 + m_2)\,\mathrm{rk}(\bar{X})\log(m_1 + m_2)}{n}\right),$$

where α^{*2} is of the order $m_1 m_2$ (see Remark 7 below for more details). Therefore, the obtained rate does not match (1), which suggests that there may have some room for improvement.

In the present work, we further investigate the case of exponential family distributions and show that under some mild assumptions, the rate (1) holds and is minimax optimal up to a logarithmic factor. A matrix completion estimator, defined as the minimizer of the sum of a log-likelihood term and a nuclear norm penalization term, is first considered. Provided that the noise is sub-exponential and the sampling distribution satisfies some assumptions controlling its deviation from the uniform distribution, it is proved that with high probability, the prediction error is upper bounded by the same rate as in the Gaussian setting (1). It should be noticed that the sub-exponential assumption is satisfied by all the above mentioned distributions.

When the additional knowledge of the sampling distribution is available, we consider another estimator, which is inspired by the one proposed by Koltchinskii et al. (2011) in the additive sub-exponential noise setting. We adapt their proofs to the exponential family distributions and show that this estimator satisfies an oracle inequality with respect to the Kullback-Leibler prediction risk. The proof techniques involved are also closely related to the dual certificate analysis derived by Zhang and Zhang (2012). With high probability, an upper bound on the prediction error, still of the same order as in (1), is derived from the oracle inequality . Finally, it is proved that the previous upper bound order is in fact minimax-optimal up to a logarithmic factor.

The rest of the paper is organized as follows. In Section 2.1, the model is specified and some background on exponential family distributions is provided. Then we give an upper bound for log - likelihood matrix completion estimator in Section 2.2 and an oracle inequality (also yielding an upper bound) for the estimator with known sampling scheme in Section 2.3. Finally, the lower bound is provided in Section 2.4. The proofs of the main results are gathered in Section 3 and the most technical Lemmas and proofs are deferred to the Appendix.

Notation

Throughout the paper, the following notation will be used. 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 by $\mathbb{R}^{m_1 \times m_2}$) with the Hilbert-Schmidt inner product $\langle X | X' \rangle := \mathrm{tr}(X^\top X')$. 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 $s \geq 1$, we denote its Schatten s-norm (see Bhatia (1997)) by

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

with $\sigma_i(X)$ the singular values of X, ordered in decreasing order. We use the convention $\|X\|_{\sigma,\infty} = \sigma_1(X)$. For any vector $z := (z_i)_{i=1}^n$, $\mathrm{diag}(z)$ denotes the $\mathbb{R}^{n\times n}$ diagonal matrix whose diagonal entries are $z_1, \cdots, z_n$. For any convex differentiable function $G : \mathbb{R} \to \mathbb{R}$ and $x, x' \in \mathbb{R}$, the Bregman divergence of G is denoted by

$$d_G(x, x') := G(x) - G(x') - G'(x')(x - x') . \tag{2}$$

2. Main results

2.1. Model Specification

We consider an unknown parameter matrix $\bar{X} \in \mathbb{R}^{m_1\times m_2}$ that we aim at recovering. Assume that an $i.i.d.$ sequence of indexes $(\omega_i)_{i=1}^n \in ([m_1] \times [m_2])^n$ is sampled and denote by Π its distribution. The observations associated to this sequence are denoted by $(Y_i)_{i=1}^n$ and assumed to follow a natural exponential family distribution, conditionally on the $\bar{X}$ entries, that is:

$$Y_i|\bar{X}_{\omega_i} \sim \mathrm{Exp}_{h,G}(\bar{X}_{\omega_i}) := h(Y_i) \exp\left(\bar{X}_{\omega_i} Y_i - G(\bar{X}_{\omega_i})\right) , \tag{3}$$

where h and G are the base measure and log partition functions associated to the canonical representation. For ease of notation we often write $\bar{X}_i$ instead of $\bar{X}_{\omega_i}$.

Given two matrices $X^1, X^2 \in \mathbb{R}^{m_1\times m_2}$, we define the empirical and integrated Bregman divergences as follows

$$D_G^n(X^1, X^2) = \frac{1}{n} \sum_{i=1}^n d_G(X_i^1, X_i^2) \quad \text{and} \quad D_G^\Pi(X^1, X^2) = \mathbb{E}[D_G^n(X^1, X^2)] . \tag{4}$$

Note that for exponential family distributions, the Bregman divergence $d_G(\cdot, \cdot)$ corresponds to the Kullback-Leibler divergence. Let $\mathbb{P}_{X^1}$ (*resp.* $\mathbb{P}_{X^2}$) denote the distribution of (Y_1, ω_1) associated to the parameters X^1 (*resp.* X^2); then $D_G^n(X^1, X^2)$ is the Kullback-Leibler divergence between $\mathbb{P}_{X^1}$ and $\mathbb{P}_{X^2}$ conditionally to the sampling, whereas $D_G^\Pi(X^1, X^2)$ is the usual Kullback-Leibler divergence.

As mentioned in introduction, the exponential family encompasses a wide range of distributions, either discrete or continuous. Some information on the most commonly used is recalled below.

Remark 1 *On its domain G admits derivatives of all orders, which can be used to compute its moments (see (Wainwright and Jordan, 2008, Proposition 3.1)). In particular, $\mathbb{E}[Y_i|\bar{X}_i] = G'(\bar{X}_i)$ and $\mathrm{Var}[Y_i|\bar{X}_i] = G''(\bar{X}_i)$ hold.*

Distribution	Parameter x	$G(x)$
Gaussian: $\mathcal{N}(\mu, \sigma^2)$ (σ known)	μ/σ	$\sigma^2 x^2/2$
Binomial: $\mathcal{B}^N(p)$ (N known)	$\log(p/(1-p))$	$N\log(1+\mathrm{e}^x)$
Poisson: $\mathcal{P}(\lambda)$	$\log(\lambda)$	e^x
Exponential: $\mathcal{E}(\lambda)$	$-\lambda$	$-\log(-x)$

Table 1: Parametrization of some exponential family distributions

2.2. General Matrix Completion

In this section, we provide statistical guarantees on the prediction error of a matrix completion estimator, which is defined as the minimizer of the sum of a log-likelihood term and a nuclear norm penalization term. For any $X \in \mathbb{R}^{m_1 \times m_2}$, denote by $\Phi_{\mathrm{Y}}(X)$ the (normalized) conditional negative log-likelihood of the observations:

$$\Phi_{\mathrm{Y}}(X) = -\frac{1}{n}\sum_{i=1}^{n} \left(\log(h(Y_i)) + X_i Y_i - G(X_i)\right) \ . \tag{5}$$

For $\gamma > 0$ and $\lambda > 0$, the nuclear norm penalized estimator $\hat{X}$ is defined as follows:

$$\hat{X} = \operatorname*{arg\,min}_{X \in \mathbb{R}^{m_1 \times m_2}, \|X\|_\infty \leq \gamma} \Phi_{\mathrm{Y}}^{\lambda}(X) \ , \quad \text{where} \quad \Phi_{\mathrm{Y}}^{\lambda}(X) = \Phi_{\mathrm{Y}}(X) + \lambda \|X\|_{\sigma,1} \ . \tag{6}$$

The parameter λ controls the trade off between fitting the data and privileging a low rank solution: for large value of λ, the rank of $\hat{X}$ is expected to be small. The parameter γ is an upper bound on the absolute value of $\bar{X}$ entries. For example, in recommender system applications analyzed with a Gaussian distribution, γ is simply the maximum rating.

Before giving an upper bound on the prediction risk $\|\hat{X} - \bar{X}\|_{\sigma,2}^2$, the following assumptions on the noise and sampling distributions need to be introduced.

H1 *The function $x \mapsto G(x)$, is twice differentiable and strongly convex on $[-\gamma, \gamma]$, so that there exists constants $\underline{\sigma}_\gamma, \bar{\sigma}_\gamma > 0$ satisfying:*

$$\underline{\sigma}_\gamma^2 \leq G''(x) \leq \bar{\sigma}_\gamma^2 \ , \tag{7}$$

for any $x \in [-\gamma, \gamma]$.

Remark 2 *Under H 1, for any $x, x' \in [-\gamma, \gamma]$, the Bregman divergence satisfies $\underline{\sigma}_\gamma^2 (x - x')^2 \leq 2 d_G(x, x') \leq \bar{\sigma}_\gamma^2 (x - x')^2$.*

Remark 3 *If the observations follow a Gaussian distribution, the two convexity constants are equal to the standard deviation* i.e., *$\bar{\sigma}_\gamma = \underline{\sigma}_\gamma = \sigma$ (see Table 1).*

For the sampling distribution, one needs to ensure that each entry has a sampling probability, which is lower bounded by a strictly positive constant, that is:

H2 *There exists a constant $\mu \geq 1$ such that, for all m_1, m_2,*

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

Denote by $R_k = \sum_{l=1}^{m_2} \pi_{k,l}$ (*resp.* $C_l = \sum_{k=1}^{m_1} \pi_{k,l}$) the probability of sampling a coefficient from row k (*resp.* column l). The following assumption requires that no row nor column should be sampled far more frequently than the others.

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

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

Remark 4 *In the classical case of a uniform sampling, $\mu = \nu = 1$ holds.*

We define the sequence of matrices $(E_i)_{i=1}^n$, whose entries are all zeros except for the coefficient (ω_i) which is equal to one *i.e.,* $E_i := e_{k_i}(e'_{l_i})^\top$ with $(k_i, l_i) = \omega_i$ and $(e_k)_{k=1}^{m_1}$ (*resp.* $(e'_l)_{l=1}^{m_2}$) being the canonical basis of $\mathbb{R}^{m_1}$ (*resp.* $\mathbb{R}^{m_2}$). Furthermore, for $(\varepsilon_i)_{i=1}^n$ a Rademacher sequence independent from $(\omega_i, Y_i)_{i=1}^n$, we also define

$$\Sigma_R := \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i E_i \, , \tag{9}$$

and use the following notation

$$d = m_1 + m_2 \, , \quad M = m_1 \vee m_2, \quad m = m_1 \wedge m_2 \, . \tag{10}$$

With these assumptions and notation, we are now ready for stating our main results.

Theorem 5 *Assume H 1, H 2, $\|\bar{X}\|_\infty \leq \gamma$ and $\lambda \geq 2\|\nabla \Phi_{\mathrm{Y}}(\bar{X})\|_{\sigma,\infty}$. Then with probability at least $1 - 2d^{-1}$ the following holds:*

$$\frac{\|\hat{X} - \bar{X}\|_{\sigma,2}^2}{m_1 m_2} \leq C\mu^2 \max\left(m_1 m_2 \, \mathrm{rk}(\bar{X}) \left(\frac{\lambda^2}{\underline{\sigma}_\gamma^4} + (\mathbb{E}\|\Sigma_R\|_{\sigma,\infty})^2 \right), \frac{\gamma^2}{\mu} \sqrt{\frac{\log(d)}{n}} \right) \, ,$$

with Σ_R and d defined in (9) and (10) and C a numerical constant.

Proof See Section 3.1. ■

In Theorem 5, the term $\mathbb{E}\|\Sigma_R\|_{\sigma,\infty}$ only depends on the sampling distribution and can be upper bounded using assumption H 3. On the other hand, the gradient term $\|\nabla \Phi_{\mathrm{Y}}(\bar{X})\|_{\sigma,\infty}$ depends both on the sampling and on the observation distributions. In order to control this term with high probability, the noise is assumed to be sub-exponential.

H4 *There exist a constant $\delta_\gamma > 0$ such that for all $x \in [-\gamma, \gamma]$ and $Y \sim \mathrm{Exp}_{h,G}(x)$:*

$$\mathbb{E}\left[\exp\left(\frac{|Y - G'(x)|}{\delta_\gamma} \right) \right] \leq \mathrm{e} \quad . \tag{11}$$

Then Theorem 5, H3 and H4 yield together the following result.

Theorem 6 *Assume H1, H2, H3, H4,* $\|\bar{X}\|_\infty \leq \gamma$,

$$n \geq 2\log(d) m \nu^{-1} \max\left(\frac{\delta_\gamma^2}{\bar{\sigma}_\gamma^2} \log^2(\delta_\gamma \sqrt{\frac{m}{\underline{\sigma}_\gamma^2}}), 1/9 \right) ,$$

and take $\lambda = 2c_\gamma \bar{\sigma}_\gamma \sqrt{2\nu \log(d)/(mn)}$, *where* c_γ *is a constant which depends only on* δ_γ. *Then with probability at least* $1 - 3d^{-1}$ *the following holds:*

$$\frac{\|\hat{X} - \bar{X}\|_{\sigma,2}^2}{m_1 m_2} \leq \bar{C} \mu^2 \max\left[\left(\frac{c_\gamma \bar{\sigma}_\gamma^2}{\underline{\sigma}_\gamma^4} + 1 \right) \frac{\nu \, \mathrm{rk}(\bar{X}) M \log(d)}{n}, \frac{\gamma^2}{\mu} \sqrt{\frac{\log(d)}{n}} \right] ,$$

with $\bar{C}$ *a numerical constant.*

Proof See Section 3.2. ■

Remark 7 *When* γ *is treated as a constant and* n *is large, the order of the bound is*

$$\frac{\|\hat{X} - \bar{X}\|_{\sigma,2}^2}{m_1 m_2} = \mathcal{O}\left(\frac{\mathrm{rk}(\bar{X}) M \log(d)}{n} \right) ,$$

*which matches the rate obtained for Gaussian distributions (1). Matrix completion for exponential family distributions was considered in the case of uniform sampling (*i.e., $\mu = \nu = 1$*) and sub-Gaussian noise by Gunasekar et al. (2014). They provide the following upper bound on the estimation error*

$$\frac{\|\bar{X} - \hat{X}\|_{\sigma,2}^2}{m_1 m_2} = \mathcal{O}\left(\alpha^{*2} \frac{\mathrm{rk}(\bar{X}) M \log(d)}{n} \right) .$$

with α^* *satisfying* $\alpha^* \geq \sqrt{m_1 m_2} \|\bar{X}\|_\infty$. *Therefore, Theorem 6 improves this rate by a factor* $m_1 m_2$.

Remark 8 *In the proof, the noncommutative Bernstein inequality for sub-exponential noise is used to control* $\|\nabla \Phi_{\mathrm{Y}}(\bar{X})\|_{\sigma,\infty}$. *However, when the observations are uniformly bounded (*e.g., *logistic distribution), a uniform Bernstein inequality can be applied instead, leading in some cases to a sharper bound (see Koltchinskii et al. (2011) and Lafond et al. (2014) for instance).*

2.3. Matrix Completion with known sampling scheme

When the sampling distribution Π is known, the following estimator can be defined:

$$\check{X} := \underset{X \in \mathbb{R}^{m_1 \times m_2}, \|X\|_\infty \leq \gamma}{\arg\min} \Phi_{\mathrm{Y}}^{\Pi}(X) + \lambda \|X\|_{\sigma,1} \quad \text{with} , \tag{12}$$

$$\Phi_{\mathrm{Y}}^{\Pi}(X) := G^{\Pi}(X) - \frac{\sum_{i=1}^n X_i Y_i}{n} \quad \text{and} \quad G^{\Pi}(X) := \mathbb{E}\left[\frac{\sum_{i=1}^n G(X_i)}{n} \right] .$$

In the case of sub-exponential additive noise, Koltchinskii et al. (2011) proposed a similar estimator and have shown that it satisfies an oracle inequality *w.r.t.* the Frobenius prediction risk. Note that their estimator coincides with (12) for the particular setting of Gaussian noise. The main interest of

computing $\check{X}$ instead of $\hat{X}$, when the sampling distribution is known, lies in the fact that a sharp oracle inequality can be derived for $\check{X}$. This powerful tool allows to provide statistical guarantees on the prediction risk, even if the true parameter $\bar{X}$ does not belong to the class of estimators *i.e.,* when $\|\bar{X}\| \leq \gamma$ is not satisfied. In this section, it is proved that $\check{X}$ satisfies an oracle inequality *w.r.t.* the integrated Bregman divergence (see Definition (4)), which corresponds to the Kullback-Leibler divergence for exponential family distributions. An upper bound on the Frobenius prediction risk is then easily derived from this inequality.

Theorem 9 *Assume H1, H2 and* $\lambda \geq \|\nabla \Phi_{\mathrm{Y}}^{\Pi}(\bar{X})\|_{\sigma,\infty}$. *Then the following inequalities hold:*

$$D_G^{\Pi}(\check{X}, \bar{X}) \leq \inf_{X \in \mathbb{R}^{m_1 \times m_2}, \|X\|_{\infty} \leq \gamma} \left(D_G^{\Pi}(X, \bar{X}) + 2\lambda \|X\|_{\sigma,1} \right) \tag{13}$$

and

$$D_G^{\Pi}(\check{X}, \bar{X}) \leq \inf_{X \in \mathbb{R}^{m_1 \times m_2}, \|X\|_{\infty} \leq \gamma} \left(D_G^{\Pi}(X, \bar{X}) + \left(\frac{1+\sqrt{2}}{2} \right)^2 \frac{\mu}{\underline{\sigma}_{\gamma}^2} m_1 m_2 \lambda^2 \operatorname{rk}(X) \right) \tag{14}$$

Proof The proof of Theorem 9 is an adaptation (to exponential family distributions) of the proof by Koltchinskii et al. (2011), which uses the first order optimality conditions satisfied by $\check{X}$. Similar arguments are used by Zhang and Zhang (2012) to provide dual certificates for non smooth convex optimization problems. The detailed proof is given in Appendix C.1. ■

When $\|\bar{X}\|_{\infty} \leq \gamma$, the previous oracle inequalities imply the following upper bound on the prediction risk.

Theorem 10 *Assume H1, H2 and* $\lambda \geq \|\nabla \Phi_{\mathrm{Y}}^{\Pi}(\bar{X})\|_{\sigma,\infty}$ *and* $\|\bar{X}\|_{\infty} \leq \gamma$. *Then the following holds:*

$$\frac{\|\check{X} - \bar{X}\|_{\sigma,2}^2}{m_1 m_2} \leq \mu^2 \min \left(\frac{\left(1+\sqrt{2}\right)^2}{2} \frac{m_1 m_2}{\underline{\sigma}_{\gamma}^4} \lambda^2 \operatorname{rk}(\bar{X}), \frac{4}{\mu \underline{\sigma}_{\gamma}^2} \lambda \|\bar{X}\|_{\sigma,1} \right) . \tag{15}$$

Proof Applying Theorem 9 to $X = \bar{X}$ and using H2 and H1 yields the result. ■

As for the previous estimator, the term $\|\nabla \Phi_{\mathrm{Y}}^{\Pi}(\bar{X})\|_{\sigma,\infty}$ is stochastic and depends both on the sampling and observations. Assuming that the sampling distribution is uniform and that the noise is sub-exponential allows to control it with high probability. Before stating the result, let us define

$$L_{\gamma} := \sup_{x \in [-\gamma, \gamma]} |G'(x)| . \tag{16}$$

Theorem 11 *Assume that the sampling is* $i.i.d.$ *uniform and* $\|\bar{X}\|_{\infty} \leq \gamma$. *Suppose H1, H4, and*

$$n \geq 2 \log(d) m \max \left(\frac{\delta_{\gamma}^2}{\bar{\sigma}_{\gamma}^2} \log^2(\delta_{\gamma} \sqrt{\frac{m}{\underline{\sigma}_{\gamma}^2}}), 8/9 \right) .$$

Take $\lambda = (c_\gamma \bar{\sigma}_\gamma + c^* L_\gamma)\sqrt{2\log(d)/(mn)}$*, where* c_γ *is a constant which depends only on* δ_γ*,* L_γ *is defined in* (16) *and* c^* *is a numerical constant. Then, with probability at least* $1-2d^{-1}$ *the following holds:*

$$\frac{\|\check{X} - \bar{X}\|_{\sigma,2}^2}{m_1 m_2} \leq \tilde{C}\left(\frac{c_\gamma \bar{\sigma}_\gamma + L_\gamma}{\underline{\sigma}_\gamma^2}\right)^2 \frac{\mathrm{rk}(\bar{X}) M \log(d)}{n} \lambda^2 \,,$$

with $\tilde{C}$ *a numerical constant.*

Remark 12 *For simplicity we have considered here only the case of uniform sampling distributions. However if we assume that the sampling satisfies H2, H3 and that there exists an absolute constant* ρ *such that* $\pi_{k,l} \leq \rho/\sqrt{m_1 m_2}$ *for any* $m_1, m_2 \in \mathbb{R}$*, then it is clear from the proof that the same bound still holds for a general* $i.i.d.$ *sampling, up to factors depending on* μ*,* ν *and* ρ*.*

Remark 13 *If* γ *is treated as a constant, the rate obtained for the Frobenius error is the same as in Theorem 6. If not, the two rates might differ because the rate of Theorem 11 depends on the constant* L_γ*, which does not appear in Theorem 6. Note in addition that Remark 8 also applies to Theorem 11.*

Proof The proof is similar to the one of Theorem 6, see Appendix C.2. ■

2.4. Lower Bound

It can be shown that the upper bounds obtained in Theorems 6 and 11 are in fact lower bounds (up to a logarithmic factor) when γ is treated as a constant. Before stating the result, let us first introduce the set $\mathcal{F}(r,\gamma)$ of matrices of rank at most r whose entries are bounded by γ:

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

The infimum over all estimators $\hat{X}$ that are measurable functions of the data $(\omega_i, Y_i)_{i=1}^n$ is denoted by $\inf_{\hat{X}}$.

Theorem 14 *There exists two constants* $c > 0$ *and* $\theta > 0$ *such that, for all* $m_1, m_2 \geq 2$*,* $1 \leq r \leq m_1 \wedge m_2$*, and* $\gamma > 0$*,*

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

Remark 15 *Theorem 14 provides a lower bound of order* $\mathcal{O}(Mr/(n\,\bar{\sigma}_\gamma^2))$*. The order of the ratio between this lower bound and the upper bounds of Theorem 6 is* $(c_\gamma(\bar{\sigma}_\gamma/\underline{\sigma}_\gamma)^4 \log(d) \vee \bar{\sigma}_\gamma^2)$*. If* γ *is treated as a constant, lower and upper bounds are therefore the same up to a logarithmic factor.*

Proof See Section 3.3. ■

3. Proofs of main results

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. Let $P_{\mathcal{S}_1^\perp(X)}$ (*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 $\mathbb{R}^{m_1 \times m_2}$

$$\mathcal{P}_X^\perp : \tilde{X} \mapsto P_{\mathcal{S}_1^\perp(X)} \tilde{X} P_{\mathcal{S}_2^\perp(X)} \text{ and } \mathcal{P}_X : \tilde{X} \mapsto \tilde{X} - \mathcal{P}_X^\perp(\tilde{X}) \, . \tag{17}$$

3.1. Proof of Theorem 5

From Definition (6), $\Phi_Y^\lambda(\hat{X}) \leq \Phi_Y^\lambda(\bar{X})$ holds, or equivalently

$$D_G^n(\hat{X}, \bar{X}) \leq \lambda(\|\bar{X}\|_{\sigma,1} - \|\hat{X}\|_{\sigma,1}) - \langle \nabla \Phi_Y(\bar{X}) \,|\, \hat{X} - \bar{X} \rangle \, ,$$

with $D_G^n(\cdot, \cdot)$ defined in (4). The first term of the right hand side can be upper bounded using Lemma 16-(iii) and the second by duality (between $\|\cdot\|_{\sigma,1}$ and $\|\cdot\|_{\sigma,\infty}$) and the assumption on λ, which yields

$$D_G^n(\hat{X}, \bar{X}) \leq \lambda \left(\| \mathcal{P}_{\bar{X}}(\hat{X} - \bar{X})\|_{\sigma,1} + \frac{1}{2} \|\hat{X} - \bar{X}\|_{\sigma,1} \right) \, .$$

Using Lemma 16-(ii) to bound the first term and Lemma 17-(ii) for the second, leads to

$$D_G^n(\hat{X}, \bar{X}) \leq 3\lambda \sqrt{2 \operatorname{rk}(\bar{X})} \|\hat{X} - \bar{X}\|_{\sigma,2} \, . \tag{18}$$

On the other hand, by strong convexity of G (H1), we get

$$\Delta_Y^2(\hat{X}, \bar{X}) := \frac{1}{n} \sum_{i=1}^n (\hat{X}_i - \bar{X}_i)^2 \leq \frac{2}{\underline{\sigma}_\gamma^2} D_G^n(\hat{X}, \bar{X}) \, . \tag{19}$$

We then define the threshold $\beta := 8\mathrm{e}\gamma^2 \sqrt{\log(d)/n}$ and distinguish the two following cases.

Case 1 If $\sum_{kl \in [m_1] \times [m_2]} \pi_{kl} (\hat{X}_{kl} - \bar{X}_{kl})^2 \leq \beta$, then Lemma 18 yields

$$\frac{\|\hat{X} - \bar{X}\|_{\sigma,2}^2}{m_1 m_2} \leq \mu\beta \, . \tag{20}$$

Case 2 If $\sum_{kl \in [m_1] \times [m_2]} \pi_{kl} (\hat{X}_{kl} - \bar{X}_{kl})^2 > \beta$, then Lemma 17-(ii) and Lemma 18 combined together give
$\hat{X} \in \mathcal{C}(\beta, 32\mu m_1 m_2 \operatorname{rk}(\bar{X}))$, where $\mathcal{C}(\cdot, \cdot)$ is the set defined as

$$\mathcal{C}(\beta, r) := \left\{ X \in \mathbb{R}^{m_1 \times m_2} \middle| \; \|X - \bar{X}\|_{\sigma,1} \leq \sqrt{r \mathbb{E}\left[\Delta_Y^2(X, \bar{X})\right]}; \mathbb{E}\left[\Delta_Y^2(X, \bar{X})\right] > \beta \right\} \, . \tag{21}$$

Hence, from Lemma 19 it holds, with probability at least $1 - (d-1)^{-1} \geq 1 - 2d^{-1}$, that

$$\Delta_Y^2(X, \bar{X}) \geq \frac{1}{2} \mathbb{E}\left[\Delta_Y^2(X, \bar{X})\right] - 512\mathrm{e}(\mathbb{E}\|\Sigma_R\|_{\sigma,\infty})^2 \mu m_1 m_2 \operatorname{rk}(\bar{X}) \, . \tag{22}$$

Combining (22) with (19), (18) and Lemma 18 leads to

$$\frac{\|\hat{X} - \bar{X}\|_{\sigma,2}^2}{2\mu m_1 m_2} - 512\mathrm{e}(\mathbb{E}\|\Sigma_R\|_{\sigma,\infty})^2 \mu m_1 m_2 \operatorname{rk}(\bar{X}) \leq \frac{6\lambda}{\underline{\sigma}_\gamma^2} \sqrt{2 m_1 m_2 \operatorname{rk}(\bar{X})} \frac{\|\hat{X} - \bar{X}\|_{\sigma,2}}{\sqrt{m_1 m_2}} \, . \tag{23}$$

Using the identity $ab \leq a^2 + b^2/4$ in (23) and combining with (20) achieves the proof of Theorem 5.

Lemma 16 *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) $\|\mathcal{P}_X(\tilde{X})\|_{\sigma,1} \leq \sqrt{2\,\mathrm{rk}(X)}\|\tilde{X}\|_{\sigma,2}$,

(iii) $\|X\|_{\sigma,1} - \|\tilde{X}\|_{\sigma,1} \leq \|\mathcal{P}_X(\tilde{X} - X)\|_{\sigma,1}$.

Lemma 17 *Let $X, \tilde{X} \in \mathbb{R}^{m_1 \times m_2}$ satisfying $\|X\|_\infty \leq \gamma$ and $\|\tilde{X}\|_\infty \leq \gamma$. Assume that $\lambda > 2\|\nabla \Phi_Y(\bar{X})\|_{\sigma,\infty}$ and $\Phi_Y^\lambda(X) \leq \Phi_Y^\lambda(\tilde{X})$. Then*

(i) $\|\mathcal{P}_{\tilde{X}}^{\perp}(X - \tilde{X})\|_{\sigma,1} \leq 3\|\mathcal{P}_{\tilde{X}}(X - \tilde{X})\|_{\sigma,1}$,

(ii) $\|X - \tilde{X}\|_{\sigma,1} \leq 4\sqrt{2\,\mathrm{rk}(\tilde{X})}\|(X - \tilde{X})\|_{\sigma,2}$.

Lemma 18 *Under H2, for any $X \in \mathbb{R}^{m_1 \times m_2}$ it holds*

$$\sum_{kl \in [m_1] \times [m_2]} \pi_{kl}(X_{kl} - \bar{X}_{kl})^2 \geq \frac{1}{\mu m_1 m_2}\|X - \bar{X}\|_{\sigma,2}^2 .$$

Lemma 19 *For $\beta = 8\mathrm{e}\gamma^2\sqrt{\log(d)/n}$, with probability at least $1 - (d-1)^{-1}$, we have for all $X \in \mathcal{C}(\beta, r)$:*

$$\left|\Delta_Y^2(X, \bar{X}) - \mathbb{E}\left[\Delta_Y^2(X, \bar{X})\right]\right| \leq \frac{\mathbb{E}\left[\Delta_Y^2(X, \bar{X})\right]}{2} + 16\mathrm{e}(\mathbb{E}\|\Sigma_R\|_{\sigma,\infty})^2\, r ,$$

with $\mathcal{C}(\beta, r)$ defined in (21).

Proof Lemmas 16 and 17 are proved in Appendix A. Lemma 18 follows directly from H2. See Appendix B for the proof of Lemma 19. ■

3.2. Proof of Theorem 6

Starting from Theorem 5 one only needs to control $\mathbb{E}(\|\Sigma_R\|_{\sigma,\infty})$ and $\|\nabla \Phi_Y(\bar{X})\|_{\sigma,\infty}$ to obtain the result.

Control of $\mathbb{E}(\|\Sigma_R\|_{\sigma,\infty})$: One can write $\Sigma_R := n^{-1}\sum_{i=1}^n Z_i$, with $Z_i := \varepsilon_i E_i$ which satisfies $\mathbb{E}[Z_i] = 0$. Recalling the definitions $R_k = \sum_{l=1}^{m_2} \pi_{k,l}$ and $C_l = \sum_{k=1}^{m_1} \pi_{k,l}$ for any $k \in [m_1]$, $l \in [m_2]$, one obtains

$$\left\|\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n Z_i Z_i^\top\right]\right\|_{\sigma,\infty} \leq \left\|\mathrm{diag}((R_k)_{k=1}^{m_1})\right\|_{\sigma,\infty} \leq \frac{\nu}{m} , \tag{24}$$

where H3 was used for the last inequality. Using a similar argument one also gets $\|\mathbb{E}[\sum_{i=1}^n Z_i^\top Z_i]\|_{\sigma,\infty}/n \leq \nu/m$. Hence applying Lemma 20 with $U = 1$ and $\sigma_Z^2 = \nu/m$, for $n \geq m\log(d)/(9\nu)$ yields

$$\mathbb{E}\left[\|\Sigma_R\|_{\sigma,\infty}\right] \leq c^*\sqrt{\frac{2\mathrm{e}\nu\log(d)}{mn}} , \tag{25}$$

with c^* a numerical constant.

Control of $\|\nabla \Phi_{\mathrm{Y}}(\bar{X})\|_{\sigma,\infty}$: Let us define $Z'_i := (Y_i - G'(\bar{X}_i))E_i$, which satisfies $\nabla \Phi_{\mathrm{Y}}(\bar{X}) := n^{-1}\sum_{i=1}^n Z'_i$ and $\mathbb{E}[Z'_i] = 0$ (as any score function) and

$$\sigma^2_{Z'} := \max\left(\frac{1}{n}\|\mathbb{E}[\sum_{i=1}^n (Z'_i)^\top Z'_i]\|_{\sigma,\infty}\,,\ \frac{1}{n}\|\mathbb{E}[\sum_{i=1}^n Z'_i (Z'_i)^\top]\|_{\sigma,\infty}\right)\,.$$

Using H4, a similar analysis yields $\sigma^2_{Z'} \leq \bar{\sigma}^2_\gamma \nu/m$. On the other hand, $\max_{k,l}(R_k, C_l) \geq 1/m$ and $\mathbb{E}[(Y_i - G'(\bar{X}_i))^2] = G''(\bar{X}_i) \geq \underline{\sigma}^2_\gamma$ gives $\sigma^2_{Z'} \geq \underline{\sigma}^2_\gamma/m$. Applying Proposition 21 for $t = \log(d)$ gives with probability at least $1 - d^{-1}$

$$\|\nabla \Phi_{\mathrm{Y}}(\bar{X})\|_{\sigma,\infty} \leq c_\gamma \max\left\{\bar{\sigma}_\gamma\sqrt{\nu/m}\sqrt{\frac{2\log(d)}{n}}, \delta_\gamma \log(\frac{\delta_\gamma\sqrt{m}}{\underline{\sigma}_\gamma})\frac{2\log(d)}{n}\right\}, \tag{26}$$

with c_γ which depends only on δ_γ. By assumption on n, the left term dominates. Therefore taking λ as in Theorem 6 statement yields $\lambda \geq 2\|\nabla \Phi_{\mathrm{Y}}(\bar{X})\|_{\sigma,\infty}$ with probability at least $1 - d^{-1}$. A union bound argument combined to Theorem 5 achieves Theorem 6 proof.

Lemma 20 *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 $i = 1,\dots,n$ and define*

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

Then, for any $n \geq (U^2\log(d))/(9\sigma^2_Z)$ the following holds:

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

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

Proof See Klopp et al. (2014)[Lemma 15]. ■

Proposition 21 *Consider a finite sequence of independent random matrices $(Z_i)_{1\leq i\leq n} \in \mathbb{R}^{m_1\times m_2}$ satisfying $\mathbb{E}[Z_i] = 0$. For some $U > 0$, assume*

$$\inf\{\delta > 0 : \mathbb{E}[\exp(\|Z_i\|_{\sigma,\infty}/\delta)] \leq \mathrm{e}\} \leq U \quad \textit{for} \quad i = 1,\dots,n$$

and define σ_Z as in Lemma 20. Then for any $t > 0$, with probability at least $1 - \mathrm{e}^{-t}$

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

with c_U a constant which depends only on U.

Proof This result is an extension of the sub-exponential noncommutative Bernstein inequality (Koltchinskii, 2013, Theorem 4), to rectangular matrices by dilation, see (Klopp, 2014, Proposition 11) for details. ■

3.3. Proof of Theorem 14

We start with a packing set construction, inspired by Koltchinskii et al. (2011). Assume *w.l.o.g.,* that $m_1 \geq m_2$. Let $\alpha \in (0, 1/8)$ and define $\kappa := \min(1/2, \sqrt{\alpha m_1 r}/(2\gamma\bar{\sigma}_\gamma^2\sqrt{n})$ and the set of matrices

$$\mathcal{L} = \left\{ L = (l_{ij}) \in \mathbb{R}^{m_1 \times r} : l_{ij} \in \{0, \kappa\gamma\} \ , \forall i \in [m_1], \ \forall j \in [r] \right\}.$$

Consider the associated set of block matrices

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

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. The Varshamov-Gilbert bound ((Tsybakov, 2009, Lemma 2.9)) guarantees the existence of a subset $\mathcal{A} \subset \mathcal{L}'$ with cardinality $\mathrm{Card}(\mathcal{A}) \geq 2^{(rm_1)/8} + 1$ containing the null matrix X^0 and such that, for any two distinct elements X^1 and X^2 of $\mathcal{A}$,

$$\|X^1 - X^2\|_2^2 \geq \frac{m_1 r\, \kappa^2\gamma^2}{8} \left\lfloor \frac{m_2}{r} \right\rfloor \geq \frac{m_1 m_2\, \kappa^2\gamma^2}{16} \,. \tag{27}$$

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 thus $\mathcal{A} \subset \mathcal{F}(r, \gamma)$. For some $X \in \mathcal{A}$, we now estimate the Kullback-Leibler divergence $D\left(\mathbb{P}_X \| \mathbb{P}_{X^0}\right)$ between probability measures $\mathbb{P}_{X^0}$ and $\mathbb{P}_X$. By independence of the observations $(Y_i, \omega_i)_{i=1}^n$ and since the distribution of $Y_i|\omega_i$ belongs to the exponential family one obtains

$$D\left(\mathbb{P}_X \| \mathbb{P}_{X^0}\right) = n\mathbb{E}_{\omega_1}\left[G'(X_{\omega_1})(X_{\omega_1} - X^0_{\omega_1}) - G(X_{\omega_1}) + G(X^0_{\omega_1}) \right] \,.$$

Since $X^0_{\omega_1} = 0$ and either $X_{\omega_1} = 0$ or $X_{\omega_1} = \kappa\gamma$, by strong convexity and by definition of κ one gets

$$D\left(\mathbb{P}_X \| \mathbb{P}_{X^0}\right) \leq n\frac{\bar{\sigma}_\gamma^2}{2}\kappa^2\gamma^2 \leq \frac{\alpha r m_1}{8} \leq \alpha \log_2(\mathrm{Card}(\mathcal{A}) - 1) \,,$$

which implies

$$\frac{1}{\mathrm{Card}(\mathcal{A}) - 1} \sum_{X \in \mathcal{A}} D\left(\mathbb{P}_{X^0} \| \mathbb{P}_X\right) \ \leq \ \alpha \log\left(\mathrm{Card}(\mathcal{A}) - 1\right) \,. \tag{28}$$

Using (27), (28) and (Tsybakov, 2009, Theorem 2.5) together gives

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

where

$$\delta(\alpha, M) = \frac{1}{1 + 2^{-rM/16}} \left(1 - 2\alpha - \frac{1}{2}\sqrt{\frac{\alpha}{rM\log(2)}} \right) \,, \tag{29}$$

and $\tilde{c}$ is a numerical constant. Since we are free to choose α as small as possible, this achieves the proof.

Acknowledgments

Jean Lafond is grateful for fundings from the Direction Générale de l'Armement (DGA) and to the labex LMH through the grant no ANR-11-LABX-0056-LMH in the framework of the "Programme des Investissements d'Avenir".

References

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

T. T. Cai and W-X. Zhou. Matrix completion via max-norm constrained optimization. *CoRR*, abs/1303.0341, 2013a.

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, 2013b.

E. J. Candès and Y. Plan. Matrix completion with noise. *Proceedings of the IEEE*, 98(6):925–936, 2010.

M. A. Davenport, Y. Plan, E. van den Berg, and M. Wootters. 1-bit matrix completion. *CoRR*, abs/1209.3672, 2012.

D. Gross. Recovering low-rank matrices from few coefficients in any basis. *Information Theory, IEEE Transactions on*, 57(3):1548–1566, 2011.

S. Gunasekar, P. Ravikumar, and J. Ghosh. Exponential family matrix completion under structural constraints. *ICML*, 2014.

R. H. Keshavan, A. Montanari, and S. Oh. Matrix completion from noisy entries. *J. Mach. Learn. Res.*, 11:2057–2078, 2010.

O. Klopp. Noisy low-rank matrix completion with general sampling distribution. *Bernoulli*, 2(1): 282–303, 02 2014.

O. Klopp, J. Lafond, E. Moulines, and J. Salmon. Adaptive Multinomial Matrix Completion. August 2014.

V. Koltchinskii. *A remark on low rank matrix recovery and noncommutative Bernstein type inequalities*, volume Volume 9 of *Collections*, pages 213–226. Institute of Mathematical Statistics, 2013.

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.

J. Lafond, O. Klopp, E. Moulines, and J. Salmon. Probabilistic low-rank matrix completion on finite alphabets. In *NIPS*. 2014.

M. Ledoux and M. Talagrand. *Probability in Banach spaces*, volume 23. Springer-Verlag, Berlin, 1991.

P. Massart. About the constants in Talagrand's concentration inequalities for empirical processes. *Ann. Probab.*, 28, 2000.

S. Negahban and M. J. Wainwright. Restricted strong convexity and weighted matrix completion: optimal bounds with noise. *J. Mach. Learn. Res.*, 13, 2012.

N. Srebro and R. R. Salakhutdinov. Collaborative filtering in a non-uniform world: Learning with the weighted trace norm. 2010.

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

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

M. J. Wainwright and M. I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1, 2008.

G. A. Watson. Characterization of the subdifferential of some matrix norms. *Linear Algebra and its Applications*, 170:33–45, 1992.

C. H. Zhang and T. Zhang. A General Framework of Dual Certificate Analysis for Structured Sparse Recovery Problems. *arXiv.org*, January 2012.

Appendix A. Proof of Lemma 16 and Lemma 17

Lemma 16

Proof If $A, B \in \mathbb{R}^{m_1 \times m_2}$ are two matrices satisfying $\mathcal{S}_i(A) \perp \mathcal{S}_i(B)$, $i = 1, 2$, (see Definition (17)) 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).

From the definition of $\mathcal{P}_X(\cdot)$, $\mathcal{P}_X(\tilde{X}) = P_{\mathcal{S}_1(X)}\tilde{X}P_{\mathcal{S}_2^{\perp}(X)} + \tilde{X}P_{\mathcal{S}_2(X)}$ holds and therefore $\mathrm{rk}(\mathcal{P}_X(\tilde{X})) \leq 2\,\mathrm{rk}(X)$. On the other hand, the Cauchy-Schwarz inequality implies that for any matrix A, $\|A\|_{\sigma,1} \leq \sqrt{\mathrm{rk}(A)}\|C\|_{\sigma,2}$. Consequently (ii) follows from

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

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

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

leading to (iii). ■

Lemma 17

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

$$\Phi_Y(\tilde{X}) - \Phi_Y(X) \geq \lambda(\|X\|_{\sigma,1} - \|\tilde{X}\|_{\sigma,1}).$$

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 16-(i) and the triangular inequality, we get

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

which implies

$$\Phi_{\mathrm{Y}}(\tilde{X}) - \Phi_{\mathrm{Y}}(X) \geq \lambda \left(\|\mathcal{P}_{\tilde{X}}^{\perp}(X - \tilde{X})\|_{\sigma,1} - \|\mathcal{P}_{\tilde{X}}(X - \tilde{X})\|_{\sigma,1} \right) . \tag{30}$$

Furthermore by convexity of Φ_{Y} we have

$$\Phi_{\mathrm{Y}}(\tilde{X}) - \Phi_{\mathrm{Y}}(X) \leq \langle \nabla \Phi_{\mathrm{Y}}(\tilde{X}) \,|\, \tilde{X} - X \rangle ,$$

which yields by duality

$$\begin{aligned} \Phi_{\mathrm{Y}}(\tilde{X}) - \Phi_{\mathrm{Y}}(X) &\leq \|\nabla \Phi_{\mathrm{Y}}(\tilde{X})\|_{\sigma,\infty} \|\tilde{X} - X\|_{\sigma,1} \leq \frac{\lambda}{2} \|\tilde{X} - X\|_{\sigma,1} , \\ &\leq \frac{\lambda}{2} (\|\mathcal{P}_{\tilde{X}}^{\perp}(X - \tilde{X})\|_{\sigma,1} + \|\mathcal{P}_{\tilde{X}}(X - \tilde{X})\|_{\sigma,1}) , \end{aligned} \tag{31}$$

where we used $\lambda > \|\nabla \Phi_{\mathrm{Y}}(\tilde{X})\|_{\sigma,\infty}$ in the second line. Then combining (30) with (31) gives (i). Since $X - \tilde{X} = \mathcal{P}_{\tilde{X}}^{\perp}(X - \tilde{X}) + \mathcal{P}_{\tilde{X}}(X - \tilde{X})$, using the triangular inequality and (i) yields

$$\|X - \tilde{X}|_{\sigma,1} \leq 4 \|\mathcal{P}_{\tilde{X}}(X - \tilde{X})\|_{\sigma,1} . \tag{32}$$

Combining (32) and Lemma 16-(i) leads to (ii). ■

Appendix B. Proof of Lemma 19

Proof The proof is adapted from (Negahban and Wainwright, 2012, Theorem 1) and (Klopp, 2014, Lemma 12). We use a peeling argument combined with a sharp deviation inequality detailed in Theorem 22. For any $\alpha > 1$, $\beta > 0$ and $0 < \eta < 1/2\alpha$, define

$$\epsilon(r, \alpha, \eta) := \frac{4}{1/(2\alpha) - \eta} (\mathbb{E}\|\Sigma_R\|_{\sigma,\infty})^2 r , \tag{33}$$

and consider the events

$$\mathcal{B} := \left\{ \exists X \in \mathcal{C}(\beta, r) \middle| \left| \Delta_Y^2(X, \bar{X}) - \mathbb{E}\left[\Delta_Y^2(X, \bar{X})\right] \right| > \frac{\mathbb{E}\left[\Delta_Y^2(X, \bar{X})\right]}{2} + \epsilon(r, \alpha, \eta) \right\} ,$$

and

$$\mathcal{R}_l := \left\{ X \in \mathcal{C}(\beta, r) | \alpha^{l-1}\beta < \mathbb{E}\left[\Delta_Y^2(X, \bar{X})\right] < \alpha^l \beta \right\} .$$

Let us also define the set

$$\mathcal{C}(\beta, r, t) := \left\{ X \in \mathcal{C}(\beta, r) |\ \mathbb{E}\left[\Delta_Y^2(X, \bar{X})\right] \leq t \right\} ,$$

and

$$Z_t := \sup_{X \in \mathcal{C}(\beta,r,t)} |\Delta_Y^2(X, \bar{X}) - \mathbb{E}\left[\Delta_Y^2(X, \bar{X})\right]| \,. \tag{34}$$

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

$$|\Delta_Y^2(X, \bar{X}) - \mathbb{E}\left[\Delta_Y^2(X, \bar{X})\right]| > \frac{1}{2}\alpha^{l-1}\beta + \epsilon(r, \alpha, \eta) \,,$$

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

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

If we now apply a union bound argument combined to Lemma 22 we get

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

where we used $x \leq \mathrm{e}^x$ in the second inequality. Choosing $\alpha = \mathrm{e}$, $\eta = (4\mathrm{e})^{-1}$ and β as stated in the Lemma yields the result. ■

Lemma 22 *Let $\alpha > 1$ and $0 < \eta < \frac{1}{2\alpha}$. Then we have*

$$\mathbb{P}\left(Z_t > t/(2\alpha) + \epsilon(r, \alpha, \eta)\right) \leq \exp\left(-n\eta^2 t^2/(8\gamma^4)\right) \,, \tag{35}$$

where $\epsilon(r, \alpha, \eta)$ and Z_t are defined in (33) and (34).

Proof From Massart's inequality ((Massart, 2000, Theorem 9)) we get for $0 < \eta < 1/(2\alpha)$

$$\mathbb{P}(Z_t > \mathbb{E}[Z_t] + \eta t) \leq \exp\left(-\eta^2 n t^2/(8\gamma^4)\right) \,. \tag{36}$$

A symmetrization argument gives

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

where $\boldsymbol{\varepsilon} := (\varepsilon_i)_{1 \leq i \leq n}$ is a Rademacher sequence independent from $(Y_i, \omega_i)_{i=1}^n$. The contraction principle ((Ledoux and Talagrand, 1991, Theorem 4.12)) yields

$$\mathbb{E}[Z_t] \leq 4\mathbb{E}\left[\sup_{X \in \mathcal{C}(\beta,r,t)} \left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i (X_i - \bar{X}_i)\right|\right] = 4\mathbb{E}\left[\sup_{X \in \mathcal{C}(\beta,r,t)} \left|\langle \Sigma_R \,|\, X - \bar{X}\rangle\right|\right] \,,$$

where Σ_R is defined in (9). Applying the duality inequality and then plugging into (36) gives

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

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

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

■

Appendix C. Proof of Oracle inequalities and Bounds for Completion with known sampling

C.1. Proof of Theorem 9

Proof The proof is an extension (to the exponential family case) of the one proposed in (Koltchinskii et al., 2011, Theorem 1). For ease of notation, let us define $H := \nabla \Phi_{\mathrm{Y}}^{\Pi}(\bar{X})$ and the set $\Gamma := \{X \in \mathbb{R}^{m_1 \times m_2} | \ \|X\|_\infty \leq \gamma\}$. In view of Remark 1, one obtains

$$H = \frac{\sum_{i=1}^n Y_i E_i}{n} - \nabla G^{\Pi}(\bar{X}) = \frac{\sum_{i=1}^n (Y_i E_i - \mathbb{E}[Y_i E_i])}{n} \ .$$

From the definition of $\check{X}$, for any $X \in \Gamma$,

$$G^{\Pi}(\check{X}) - \frac{\sum_{i=1}^n \check{X}_i Y_i}{n} \leq G^{\Pi}(X) - \frac{\sum_{i=1}^n X_i Y_i}{n} + \lambda(\|X\|_{\sigma,1} - \|\check{X}\|_{\sigma,1})$$

or equivalently

$$\begin{aligned} &G^{\Pi}(\check{X}) - G^{\Pi}(\bar{X}) - \langle \nabla G^{\Pi}(\bar{X}) \,|\, \check{X} - \bar{X} \rangle \\ &\quad \leq G^{\Pi}(X) - G^{\Pi}(\bar{X}) - \langle \nabla G^{\Pi}(\bar{X}) \,|\, X - \bar{X} \rangle + \langle H \,|\, \check{X} - X \rangle + \lambda(\|X\|_{\sigma,1} - \|\check{X}\|_{\sigma,1}) \end{aligned}$$

Applying Lemma 16 (ii),(iii) and duality yields

$$D_G^{\Pi}(\check{X}, \bar{X}) - D_G^{\Pi}(X, \bar{X}) \leq \lambda(\|\check{X} - X\|_{\sigma,1} + \|X\|_{\sigma,1} - \|\check{X}\|_{\sigma,1}) \leq 2\lambda \|X\|_{\sigma,1} \ .$$

where we used the assumption $\lambda \geq \|H\|_{\sigma,\infty}$. This proves (13).

For (14), by definition

$$\check{X} = \operatorname*{arg\,min}_{X \in \mathbb{R}^{m_1 \times m2}} F(X) := G^{\Pi}(X) - \frac{\sum_{i=1}^n X_i Y_i}{n} + \lambda \|X\|_{\sigma,1} + \delta_\Gamma(X) \ ,$$

where δ_Γ is the indicatrice function of the bounded closed convex set Γ *i.e.,* $\delta_\Gamma(x) = 0$ if $x \in \Gamma$ and $\delta_\Gamma(x) = +\infty$ otherwise. Since F is convex, $\check{X}$ satisfies $0 \in \partial F(\check{X})$ with ∂F denoting the subdifferential of F. It is easily checked that the subdifferential $\partial \delta_\Gamma(\check{X})$ is the normal cone of Γ at the point $\check{X}$. Hence, $0 \in \partial F(\check{X})$ implies that there exists $\check{V} \in \partial \|\check{X}\|_{\sigma,1}$ such that for any $X \in \Gamma$,

$$\langle \nabla G^{\Pi}(\check{X}) \,|\, \check{X} - X \rangle - \left\langle \frac{\sum_{i=1}^n Y_i E_i}{n} \,|\, \check{X} - X \right\rangle + \lambda \langle \check{V} \,|\, \check{X} - X \rangle \leq 0 \ ,$$

or equivalently

$$\langle \nabla G^{\Pi}(\check{X}) - \nabla G^{\Pi}(\bar{X}) \,|\, \check{X} - X \rangle + \lambda \langle \check{V} \,|\, \check{X} - X \rangle \leq \langle H \,|\, \check{X} - X \rangle \ .$$

For any $\tilde{x}, \bar{x}, x \in \mathbb{R}$, from the Bregman divergence definition it holds

$$(G'(\check{x}) - G'(\bar{x}))(\check{x} - x) = d_G(x, \check{x}) + d_G(\check{x}, \bar{x}) - d_G(x, \bar{x}) \ . \tag{37}$$

In addition, for any $V \in \partial \|X\|_{\sigma,1}$, the subdifferential monotonicity yields $\langle \check{V} - V \,|\, \check{X} - X \rangle \geq 0$. Therefore

$$D_G^{\Pi}(X, \check{X}) + D_G^{\Pi}(\check{X}, \bar{X}) - D_G^{\Pi}(X, \bar{X}) \leq \langle H \,|\, \check{X} - X \rangle - \lambda \langle V \,|\, \check{X} - X \rangle \ . \tag{38}$$

In Watson (1992), it is shown that:

$$\partial\|X\|_{\sigma,1} = \{\sum_{i=1}^{r} u_i v_i^\top + \mathcal{P}_X^\perp W \;\; | W \in \mathbb{R}^{m_1 \times m_2},\; \|W\|_{\sigma,\infty} \leq 1\}\;, \tag{39}$$

where $r := \mathrm{rk}(X)$, u_i (*resp.* v_i) are the left (*resp.* right) singular vectors of X and $\mathcal{P}_X^\perp$ is defined in (17). Denote by $\mathcal{S}_1$ (*resp.* $\mathcal{S}_2$) the space of the left (*resp.* right) singular vectors of X. For $W \in \mathbb{R}^{m_1 \times m_2}$,

$$\left\langle \sum_{i=1}^{r} u_i v_i^\top + \mathcal{P}_X^\perp W \,|\, \check{X} - X \right\rangle = \left\langle \sum_{i=1}^{r} u_i v_i^\top \,|\, P_{\mathcal{S}_1}(\check{X} - X)P_{\mathcal{S}_1} \right\rangle + \left\langle W \,|\, \mathcal{P}_X^\perp(\check{X}) \right\rangle\;,$$

and W can be chosen such that $\langle W \,|\, \mathcal{P}_X^\perp(\check{X})\rangle = \|\,\mathcal{P}_X^\perp(\check{X})\|_{\sigma,1}$ and $\|W\|_{\sigma,\infty} \leq 1$. Taking $V \in \partial\|X\|_{\sigma,1}$ associated to this choice of W (in the sense of (39)) and $\|\sum_{i=1}^r u_i v_i^\top\|_{\sigma,\infty} = 1$ yield

$$\begin{aligned} D_G^\Pi(X, \check{X}) + D_G^\Pi(\check{X}, \bar{X}) - D_G^\Pi(X, \bar{X}) + \lambda\|\,\mathcal{P}_X^\perp(\check{X})\|_{\sigma,1} \\ \leq \langle H \,|\, \check{X} - X\rangle + \|P_{\mathcal{S}_1}(\check{X} - X)P_{\mathcal{S}_1}\|_{\sigma,1}\;. \end{aligned} \tag{40}$$

The first right hand side term can be upper bounded as follows

$$\begin{aligned} \langle H \,|\, \check{X} - X\rangle &= \langle H \,|\, \mathcal{P}_X(\check{X} - X)\rangle + \langle H \,|\, \mathcal{P}_X^\perp(\check{X})\rangle \\ &\leq \|H\|_{\sigma,\infty}(\sqrt{2\,\mathrm{rk}(X)}\|\check{X} - X\|_{\sigma,2} + \|\,\mathcal{P}_X^\perp(\check{X})\|_{\sigma,1})\;, \end{aligned} \tag{41}$$

where duality and Lemma 16-(ii) are used for the inequality. Since $\mathrm{rk}(P_{\mathcal{S}_1}(\check{X} - X)P_{\mathcal{S}_1}) \leq \mathrm{rk}(X)$, the second term satisfies

$$\|P_{\mathcal{S}_1}(\check{X} - X)P_{\mathcal{S}_1}\|_{\sigma,1} \leq \sqrt{\mathrm{rk}\,X}\|\check{X} - X\|_{\sigma,2}\;. \tag{42}$$

Using $\lambda \geq \|H\|_{\sigma,\infty}$, (40), (41) and (42) gives

$$\begin{aligned} D_G^\Pi(X, \check{X}) + D_G^\Pi(\check{X}, \bar{X}) + (\lambda - \|H\|_{\sigma,\infty})\|\,\mathcal{P}_X^\perp(\check{X})\|_{\sigma,1} \\ \leq D_G^\Pi(X, \bar{X}) + \lambda(1 + \sqrt{2})\sqrt{\mathrm{rk}(X)}\|\check{X} - X\|_{\sigma,2}\;. \end{aligned} \tag{43}$$

By H1 and H2, $\|\check{X} - X\|_{\sigma,2} \leq \underline{\sigma}_\gamma^{-1}\sqrt{2m_1m_2\mu D_G^\Pi(X, \check{X})}$, hence

$$\begin{aligned} D_G^\Pi(\check{X}, \bar{X}) + (\lambda - \|H\|_{\sigma,\infty})\|\,\mathcal{P}_X^\perp(\check{X})\|_{\sigma,1} \\ \leq D_G^\Pi(X, \bar{X}) + (\frac{1 + \sqrt{2}}{2})^2\underline{\sigma}_\gamma^{-2}m_1m_2\mu\lambda^2\,\mathrm{rk}(X)\;, \end{aligned} \tag{44}$$

proving (14). ■

C.2. proof of Theorem 11

Proof By the triangle inequality,

$$\|H\|_{\sigma,\infty} \leq \left\| \frac{\sum_{i=1}^n (Y_i - G'(X_i)E_i}{n} \right\|_{\sigma,\infty} + \left\| \frac{\sum_{i=1}^n G'(X_i)E_i}{n} - \mathbb{E}[G'(X_1)E_1 \right\|_{\sigma,\infty} , \tag{45}$$

holds. As seen in the proof of Theorem 6 (in Section 3.2), the first term of the right hand side satisfies (26) with probability at least $1 - d^{-1}$. If we define $Z_i = G'(X_i)E_i - \mathbb{E}[G'(X_1)E_1]$, then $\mathbb{E}[Z_i] = 0$ gives $\|Z_i\|_{\sigma,\infty} \leq 2L_\gamma$, with L_γ defined in (16). A similar argument to the one used to derive Equation (24) yields

$$\left\| \mathbb{E}\left[Z_i^\top Z_i \right] \right\|_{\sigma,\infty} \leq \|\mathbb{E}[(G'(X_i)E_i)(G'(X_i)E_i)^\top]\|_{\sigma,\infty} \leq L_\gamma^2 \frac{1}{m} ,$$

and the same bound holds for $\mathbb{E}[Z_i Z_i^\top]$. Therefore, the uniform version of the noncommutative Bernstein inequality (Proposition 23) ensures that with probability at least $1 - d^{-1}$

$$\left\| \frac{\sum_{i=1}^n G'(X_i)E_i}{n} - \mathbb{E}[G'(X_1)E_1 \right\|_{\sigma,\infty} \leq c^* \max\left(\frac{L_\gamma}{\sqrt{m}} \sqrt{\frac{2\log(d)}{n}}, 4L_\gamma \frac{\log(d)}{3n} \right) . \tag{46}$$

Combining (26), (46) with the assumption made on n in Theorem 11, achieves the proof. ■

Proposition 23 *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 $i = 1, \ldots, n$. 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

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

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 (Tropp, 2012, 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 $n\sigma_Z^2 > (U^2/9)(v + \log(d))$. ■