Bernoulli **20**(1), 2014, 282–303
DOI: 10.3150/12-BEJ486

Noisy low-rank matrix completion with general sampling distribution

OLGA KLOPP

MODAL'X, University Paris Ouest Nanterre and CREST, 200 avenue de la République, 92001 Nanterre, France. E-mail: kloppolga@math.cnrs.fr

In the present paper, we consider the problem of matrix completion with noise. Unlike previous works, we consider quite general sampling distribution and we do not need to know or to estimate the variance of the noise. Two new nuclear-norm penalized estimators are proposed, one of them of "square-root" type. We analyse their performance under high-dimensional scaling and provide non-asymptotic bounds on the Frobenius norm error. Up to a logarithmic factor, these performance guarantees are minimax optimal in a number of circumstances.

Keywords: high-dimensional sparse model; low rank matrix estimation; matrix completion; unknown variance

1. Introduction

This paper considers the problem of matrix recovery from a small set of noisy observations. Suppose that we observe a small set of entries of a matrix. The problem of inferring the many missing entries from this set of observations is the *matrix completion* problem. A usual assumption that allows to succeed such a completion is to suppose that the unknown matrix has low rank or has approximately low rank.

The problem of matrix completion comes up in many areas including collaborative filtering, multi-class learning in data analysis, system identification in control, global positioning from partial distance information and computer vision, to mention some of them. For instance, in computer vision, this problem arises as many pixels may be missing in digital images. In collaborative filtering, one wants to make automatic predictions about the preferences of a user by collecting information from many users. So, we have a data matrix where rows are users and columns are items. For each user, we have a partial list of his preferences. We would like to predict the missing rates in order to be able to recommend items that may interest each user.

The noiseless setting was first studied by Candès and Recht [5] using nuclear norm minimization. A tighter analysis of the same convex relaxation was carried out in [6]. For a simpler approach, see more recent papers of Recht [22] and Gross [10]. An alternative line of work was developed by Keshavan *et al.* in [12]. A more common situation in applications corresponds to the noisy setting in which the few available entries are corrupted by noise. This problem has been extensively studied recently. The most popular methods rely on nuclear norm minimization (see, e.g., [4,8,9,11,13,17,18,21,23]). One can also use rank penalization as it was done by Bunea *et al.* [3] and Klopp [14]. Typically, in the matrix completion problem, the sampling scheme is supposed to be uniform. However, in practice, the observed entries are not guaranteed to follow the uniform scheme and its distribution is not known exactly.

In the present paper, we consider nuclear norm penalized estimators and study the corresponding estimation error in Frobenius norm. We consider both cases when the variance of the noise is known or not. Our methods allow us to consider quite general sampling distribution: we only assume that the sampling distribution satisfies some mild "regularity" conditions (see Assumptions 1 and 2).

Let $A_0 \in \mathbb{R}^{m_1 \times m_2}$ be the unknown matrix. Our main results, Theorems 10 and 7, show the following bound on the normalized Frobenius error of the estimators $\hat{A}$ that we propose in this paper: with high probability

$$\frac{\|\hat{A} - A_0\|_2^2}{m_1 m_2} \lesssim \frac{\log(m_1 + m_2) \max(m_1, m_2) \operatorname{rank}(A_0)}{n},$$

where the symbol $\lesssim$ means that the inequality holds up to a multiplicative numerical constant. This theorem guarantees, that the prediction error of our estimator is small whenever $n \gtrsim \log(m_1 + m_2) \max(m_1, m_2) \operatorname{rank}(A_0)$. This quantifies the sample size necessary for successful matrix completion. Note that, when $\operatorname{rank}(A_0)$ is small, this is considerably smaller than $m_1 m_2$, the total number of entries. For large m_1, m_2 and small r, this is also quite close to the degree of freedom of a rank r matrix, which is $(m_1 + m_2)r - r^2$.

An important feature of our estimator is that its construction requires only an upper bound on the maximum absolute value of the entries of A_0. This condition is very mild. A bound on the maximum of the elements is often known in applications. For instance, if the entries of A_0 are some user's ratings it corresponds to the maximal rating. Previously, the estimators proposed by Koltchinskii *et al.* [18] and by Klopp [14] also require a bound on the maximum of the elements of the unknown matrix but their constructions use the uniform sampling and additionally require the knowledge of an upper bound on the variance of the noise. Other works on matrix completion require more involved conditions on the unknown matrix. For more details, see Section 3.

Sampling schemes more general than the uniform one were previously considered in [7,19,21]. Lounici [19] considers a different estimator and measures the prediction error in the spectral norm. In [7,21] the authors consider penalization using a weighted trace-norm, which was first introduced by Srebro *et al.* [24]. Negahban *et al.* in [21] assume that the sampling distribution is a product distribution, that is, the row index and the column index of the observed entries are selected independently. This assumption does not seem realistic in many cases (see discussion in [7]). An important advantage of our method is that the sampling distribution does not need to be equal to a product distribution. Foygel *et al.* in [7] propose a method based on the "smoothing" of the sampling distribution. This procedure may be applied to an arbitrary sampling distribution but requires a priori information on the rank of the unknown matrix. Moreover, unlike in the present paper, in [7] the prediction performances of the estimator are evaluated through a bound on the expected l-Lipschitz loss (where the expectation is taken with respect to the sampling distribution).

The weighted trace-norm, used in [7,21], corrects a specific situation where the standard trace-norm fails. This situation corresponds to a non-uniform distribution where the row/column marginal distribution is such that some columns or rows are sampled with very high probability (for a more thorough discussion see [7,24]). Unlike [7,21], we use the standard trace-norm penalization and our assumption on the sampling distribution (Assumption 1) guarantees that no row or column is sampled with very high probability.

Most of the existing methods of matrix completion rely on the knowledge or a pre-estimation of the standard deviation of the noise. The matrix completion problem with unknown variance of the noise was previously considered in [13] using a different estimator which requires uniform sampling. Note also that in [13] the bound on the prediction error is obtained under some additional condition on the rank and the "spikiness ratio" of the matrix. The construction of the present paper is valid for more general sampling distributions and does not require such an extra condition.

The remainder of this paper is organized as follows. In Section 2, we introduce our model and the assumptions on the sampling scheme. For the reader's convenience, we also collect notation which we use throughout the paper. In Section 3 we consider matrix completion in the case of known variance of the noise. We define our estimator and prove Theorem 3 which gives a general bound on its Frobenius error conditionally on bounds for the stochastic terms. Theorem 7, provides bounds on the Frobenius error of our estimator in closed form. Therefore, we use bounds on the stochastic terms that we derive in Section 5. To obtain such bounds, we use a non-commutative extension of the classical Bernstein inequality.

In Section 4, we consider the case when the variance of the noise is unknown. Our construction uses the idea of "square-root" estimators, first introduced by Belloni *et al.* [1] in the case of the square-root Lasso estimator. Theorem 10, shows that our estimator has the same performances as previously considered estimators which require the knowledge of the standard deviation of the noise and of the sampling distribution.

2. Preliminaries

2.1. Model and sampling scheme

Let $A_0 \in \mathbb{R}^{m_1 \times m_2}$ be an unknown matrix, and consider the observations (X_i, Y_i) satisfying the trace regression model

$$Y_i = \operatorname{tr}\big(X_i^T A_0\big) + \sigma \xi_i, \qquad i = 1, \dots, n. \tag{1}$$

The noise variables ξ_i are independent, with $\mathbb{E}(\xi_i) = 0$ and $\mathbb{E}(\xi_i^2) = 1$; X_i are random matrices of dimension $m_1 \times m_2$ and $\operatorname{tr}(A)$ denotes the trace of the matrix A. Assume that the design matrices X_i are i.i.d. copies of a random matrix X having distribution Π on the set

$$\mathcal{X} = \big\{ e_j(m_1) e_k^T(m_2), 1 \le j \le m_1, 1 \le k \le m_2 \big\}, \tag{2}$$

where $e_l(m)$ are the canonical basis vectors in $\mathbb{R}^m$. Then, the problem of estimating A_0 coincides with the problem of matrix completion with random sampling distribution Π.

One of the particular settings of this problem is the Uniform Sampling at Random (USR) matrix completion which corresponds to the uniform distribution Π. We consider a more general weighted sampling model. More precisely, let $\pi_{jk} = \mathbb{P}(X = e_j(m_1) e_k^T(m_2))$ be the probability to observe the (j,k)th entry. Let us denote by $C_k = \sum_{j=1}^{m_1} \pi_{jk}$ the probability to observe an element from the kth column and by $R_j = \sum_{k=1}^{m_2} \pi_{jk}$ the probability to observe an element from the jth row. Observe that $\max_{i,j}(C_i, R_j) \ge 1/\min(m_1, m_2)$.

As it was shown in [24], the trace-norm penalization fails in the specific situation when the row/column marginal distribution is such that some columns or rows are sampled with very high probability (for more details, see [7,24]). To avoid such a situation, we need the following assumption on the sampling distribution:

Assumption 1. *There exists a positive constant* $L \geq 1$ *such that*

$$\max_{i,j}(C_i, R_j) \leq L/\min(m_1, m_2).$$

In order to get bounds in the Frobenius norm, we suppose that each element is sampled with positive probability:

Assumption 2. *There exists a positive constant* $\mu \geq 1$ *such that*

$$\pi_{jk} \geq (\mu m_1 m_2)^{-1}.$$

In the case of uniform distribution $L = \mu = 1$. Let us set $\|A\|^2_{L_2(\Pi)} = \mathbb{E}(\langle A, X\rangle^2)$. Assumption 2 implies that

$$\|A\|^2_{L_2(\Pi)} \geq (m_1 m_2 \mu)^{-1} \|A\|_2^2. \tag{3}$$

2.2. Notation

We provide a brief summary of the notation used throughout this paper. Let A, B be matrices in $\mathbb{R}^{m_1 \times m_2}$.

- We define the *scalar product* $\langle A, B\rangle = \operatorname{tr}(A^T B)$.
- For $0 < q < \infty$ the *Schatten-q* (*quasi-*)*norm* of the matrix A is defined by

$$\|A\|_q = \left(\sum_{j=1}^{\min(m_1,m_2)} \sigma_j(A)^q \right)^{1/q} \quad \text{and} \quad \|A\| = \sigma_1(A),$$

  where $(\sigma_j(A))_j$ are the singular values of A ordered decreasingly.
- $\|A\|_\infty = \max_{i,j} |a_{ij}|$ where $A = (a_{ij})$.
- Let $\pi_{i,j} = \mathbb{P}(X = e_i(m_1)e_j^T(m_2))$ be the probability to observe the (i, j)th element.
- For $j = 1, \ldots, m_2$, $C_j = \sum_{i=1}^{m_1} \pi_{ij}$ and for $i = 1, \ldots, m_1$, $R_i = \sum_{j=1}^{m_2} \pi_{ij}$.
- $R = \operatorname{diag}(R_1, \ldots, R_{m_1})$ and $C = \operatorname{diag}(C_1, \ldots, C_{m_2})$.
- Let $M = \max(m_1, m_2)$, $m = \min(m_1, m_2)$ and $d = m_1 + m_2$.
- $\|A\|^2_{L_2(\Pi)} = \mathbb{E}(\langle A, X\rangle^2)$.
- Let $\{\varepsilon_i\}_{i=1}^n$ be an i.i.d. Rademacher sequence and we define

$$\Sigma_R = \frac{1}{n} \sum_{i=1}^n \varepsilon_i X_i \quad \text{and} \quad \Sigma = \frac{\sigma}{n} \sum_{i=1}^n \xi_i X_i. \tag{4}$$

- Define the observation operator $\Omega : \mathbb{R}^{m_1 \times m_2} \to \mathbb{R}^n$ as $(\Omega(A))_i = \langle X_i, A \rangle$.
- $Q(A) = \sqrt{\frac{1}{n}\sum_{i=1}^n (Y_i - \langle X_i, A\rangle)^2}$.

3. Matrix completion with known variance of the noise

In this section, we consider the matrix completion problem when the variance of the noise is known. We define the following estimator of A_0:

$$\hat{A} = \underset{\|A\|_\infty \le \mathbf{a}}{\arg\min} \left\{ \frac{1}{n} \sum_{i=1}^n \big(Y_i - \langle X_i, A\rangle\big)^2 + \lambda \|A\|_1 \right\}, \tag{5}$$

where $\lambda > 0$ is a regularization parameter and $\mathbf{a}$ is an upper bound on $\|A_0\|_\infty$. This is a restricted version of the matrix LASSO estimator. The matrix LASSO estimator is based on a trade-off between fitting the target matrix to the data using least squares and minimizing the nuclear norm and it has been studied by a number of authors (see, e.g., [4,20,23]).

A restricted version of a slightly different estimator, penalised by a weighted nuclear norm $\|\sqrt{R}A\sqrt{C}\|_1$, was first considered by Negahban and Wainwright in [21]. Here R and C are diagonal matrices with diagonal entries $\{R_j, j = 1, \dots, m_1\}$ and $\{C_k, k = 1, \dots, m_2\}$, respectively. In [21], the domain of optimization is the following one

$$\left\{ A : \|A\|_{\omega(\infty)} \le \frac{\alpha^*}{\sqrt{m_1 m_2}} \right\}, \tag{6}$$

where α^* is a bound on the "spikiness ratio" $\alpha_{sp} = \frac{\sqrt{m_1 m_2}\|A_0\|_{\omega(\infty)}}{\|A_0\|_{\omega(2)}}$ of the unknown matrix A_0. Here $\|A\|_{\omega(\infty)} = \|\sqrt{R}A\sqrt{C}\|_\infty$ and $\|A\|_{\omega(2)} = \|\sqrt{R}A\sqrt{C}\|_2$. In the particular setting of the uniform sampling (6) gives

$$\big\{ A : \|A\|_\infty \le \alpha \big\},$$

where α is an upper bound on the "spikiness ratio" $\frac{\sqrt{m_1 m_2}\|A_0\|_\infty}{\|A_0\|_2}$.

The following theorem gives a general upper bound on the prediction error of estimator $\hat{A}$ given by (5). Its proof is given in Appendix A. The stochastic terms $\|\Sigma\|$ and $\|\Sigma_R\|$ play a key role in what follows.

Theorem 3. *Let X_i be i.i.d. with distribution Π on $\mathcal{X}$ which satisfies Assumptions 1 and 2 and $\lambda > 3\|\Sigma\|$. Assume that $\|A_0\|_\infty \le \mathbf{a}$ for some constant $\mathbf{a}$. Then, there exist numerical constants (c_1, c_2) such that*

$$\frac{\|\hat{A} - A_0\|_2^2}{m_1 m_2} \le \max\left\{ c_1 \mu^2 m_1 m_2 \operatorname{rank}(A_0)\big(\lambda^2 + \mathbf{a}^2\big(\mathbb{E}\big(\|\Sigma_R\|\big)\big)^2\big), c_2 \mathbf{a}^2 \mu \sqrt{\frac{\log(d)}{n}} \right\}$$

with probability at least $1 - \frac{2}{d}$, where $d = m_1 + m_2$.

In order to get a bound in a closed form, we need to obtain suitable upper bounds on $\mathbb{E}(\|\Sigma_R\|)$ and, with probability close to 1, on $\|\Sigma\|$. We will obtain such bounds in the case of *sub-exponential noise*, that is, under the following assumption:

Assumption 4.

$$\max_{i=1,\dots,n} \mathbb{E}\exp\bigl(|\xi_i|/K\bigr) < \infty.$$

Let $K > 0$ be a constant such that $\max_{i=1,\dots,n} \mathbb{E}\exp(|\xi_i|/K) \le e$. The following two lemmas give bounds on $\|\Sigma\|$ and $\mathbb{E}(\|\Sigma_R\|)$. We prove them in Section 5 using the non-commutative Bernstein inequality.

Lemma 5. *Let X_i be i.i.d. with distribution Π on $\mathcal{X}$ which satisfies Assumptions 1 and 2. Assume that $(\zeta_i)_{i=1}^n$ are independent with $\mathbb{E}(\zeta_i) = 0$, $\mathbb{E}(\zeta_i^2) = 1$ and satisfy Assumption 4. Then, there exists an absolute constant $C^* > 0$ that depends only on K and such that, for all $t > 0$ with probability at least $1 - e^{-t}$ we have*

$$\left\| \frac{1}{n} \sum_{i=1}^n \zeta_i X_i \right\| \le C^* \max\left\{ \sqrt{\frac{L(t+\log(d))}{mn}}, \frac{\log(m)(t+\log(d))}{n} \right\}, \tag{7}$$

where $d = m_1 + m_2$.

Lemma 6. *Let X_i be i.i.d. with distribution Π on $\mathcal{X}$ which satisfies Assumptions 1 and 2. Assume that $(\zeta_i)_{i=1}^n$ are independent with $\mathbb{E}(\zeta_i) = 0$, $\mathbb{E}(\zeta_i^2) = 1$ and satisfy Assumption 4. Then, for $n \ge m \log^3(d)/L$, there exists an absolute constant $C^* > 0$ such that*

$$\mathbb{E}\left\| \frac{1}{n} \sum_{i=1}^n \zeta_i X_i \right\| \le C^* \sqrt{\frac{2eL\log(d)}{nm}},$$

where $d = m_1 + m_2$.

An optimal choice of the parameter t in these lemmas is $t = \log(d)$. Larger t leads to a slower rate of convergence and a smaller t does not improve the rate but makes the concentration probability smaller. With this choice of t the second terms in the maximum in (7) is negligible for $n > n^*$ where $n^* = 2\log^2(d)m/L$. Then, we can choose

$$\lambda = 3C^*\sigma \sqrt{\frac{2L\log(d)}{mn}}, \tag{8}$$

where C^* is an absolute numerical constant which depends only on K. If ξ_i are $N(0,1)$, then we can take $C^* = 6.5$ (see Lemma 4 in [13]). With this choice of λ, we obtain the following theorem.

Theorem 7. *Let X_i be i.i.d. with distribution Π on $\mathcal{X}$ which satisfies Assumptions 1 and 2. Assume that $\|A_0\|_\infty \leq \mathbf{a}$ for some constant $\mathbf{a}$ and that Assumption 4 holds. Consider the regularization parameter λ satisfying (8). Then, there exist a numerical constant c', that depends only on K, such that*

$$\frac{\|\hat{A} - A_0\|_2^2}{m_1 m_2} \leq c' \max\left\{ \max\left(\sigma^2, \mathbf{a}^2\right) \mu^2 L \frac{\log(d)\,\mathrm{rank}(A_0) M}{n}, \mathbf{a}^2 \mu \sqrt{\frac{\log(d)}{n}} \right\} \tag{9}$$

with probability greater than $1 - 3/d$.

Remarks. *Comparison to other works*: An important feature of our estimator is that its construction requires only an upper bound on the maximum absolute value of the entries of A_0 (and an upper bound on the variance of the noise). This condition is very mild. Let us compare this matrix condition and the bound we obtain with some of the previous works on noisy matrix completion.

We will start with the paper of Keshavan *et al.* [11]. Their method requires a priori information on the rank of the unknown matrix as well as a matrix incoherence assumption (which is stated in terms of the singular vectors of A_0). Under a sampling scheme different from ours (uniform sampling without replacement) and sub-Gaussian errors, the estimator proposed in [11] satisfies, with high probability, the following bound

$$\frac{\|\hat{A} - A_0\|_2^2}{m_1 m_2} \lesssim k^4 \sqrt{\alpha} \frac{M}{n} \mathrm{rank}(A_0) \log n. \tag{10}$$

The symbol $\lesssim$ means that the inequality holds up to multiplicative numerical constants, $k = \sigma_{\max}(A_0)/\sigma_{\min}(A_0)$ is the condition number and $\alpha = (m_1 \vee m_2)/(m_1 \wedge m_2)$ is the aspect ratio. Comparing (10) and (9), we see that our bound is better: it does not involve the multiplicative coefficient $k^4\sqrt{\alpha}$ which can be big.

Wainwright *et al.* in [21] propose an estimator which uses a priori information on the "spikiness ratio" $\alpha_{sp} = \frac{\sqrt{m_1 m_2}\|A_0\|_\infty}{\|A_0\|_2}$ of A_0. This method requires α_{sp} bounded by a constant, say α_*, in which case the estimator proposed in [21] satisfies the following bound

$$\frac{\|\hat{A} - A_0\|_{\omega(2)}^2}{m_1 m_2} \lesssim \alpha_*^2 \frac{M}{n} \mathrm{rank}(A_0) \log m. \tag{11}$$

In the case of uniform sampling and bounded "spikiness ratio" this bound coincides with the bound given by Theorem 7. An important advantage of our method is that the sampling distribution does not need to be equal to a product distribution (i.e., π_{ij} need not be equal to $R_i C_j$) as is required in [21].

The methods proposed in [13,14,18] use the uniform sampling. Similarly to our construction, an a priori bound on $\|A_0\|_\infty$ is required. An important difference is that, in these papers, the bound on $\|A_0\|_\infty$ is used in the choice of the regularization parameter λ. This implies that the convex functional which is minimized in order to obtain $\hat{A}$ depends on $\mathbf{a}$. A too large bound may jeopardize the exactness of the estimation. In our construction, $\mathbf{a}$ determines the ball over which we are minimizing our convex functional, which itself is independent of $\mathbf{a}$. Our estimator achieves the same bound as the estimators proposed in these papers.

Minimax optimality: If we consider the matrix completion setting (i.e., $n \le m_1 m_2$), then, the maximum in (9) is given by its first therm. In the case of Gaussian errors and under the additional assumption that $\pi_{jk} \le \frac{\mu_1}{m_1 m_2}$ for some constant $\mu_1 \ge 1$ this rate of convergence is minimax optimal (cf. Theorem 5 of [18]). This optimality holds for the class of matrices $\mathcal{A}(r, a)$ defined as follows: for given r and a $A_0 \in \mathcal{A}(r, a)$ if and only if the rank of A_0 is not larger than r and all the entries of A_0 are bounded in absolute value by a.

Possible extensions: The techniques developed in this paper may also be used to analyse weighted trace norm penalty similar to one used in [7,21].

4. Matrix completion with unknown variance of the noise

In this section, we propose a new estimator for the matrix completion problem in the case when the variance of the noise σ is unknown. Our construction is inspired by the square-root Lasso estimator proposed in [1]. We define the following estimator of A_0:

$$\hat{A}_{\mathrm{SQ}} = \operatorname*{arg\,min}_{\|A\|_\infty \le \mathbf{a}} \left\{ \sqrt{\frac{1}{n} \sum_{i=1}^{n} \big(Y_i - \langle X_i, A \rangle\big)^2} + \lambda \|A\|_1 \right\}, \tag{12}$$

where $\lambda > 0$ is a regularization parameter and $\mathbf{a}$ is an upper bound on $\|A_0\|_\infty$. Note that the first term of this estimator is the square root of the data-dependent term of the estimator that we considered in Section 3. This is similar to the principle used to define the square-root Lasso estimator for the usual vector regression model.

Let us set $\rho = \frac{1}{16 \mu m_1 m_2 \operatorname{rank}(A_0)}$. The following theorem gives a general upper bound on the prediction error of the estimator $\hat{A}_{\mathrm{SQ}}$. Its proof is given in Appendix D.

Theorem 8. *Let X_i be i.i.d. with distribution Π on $\mathcal{X}$ which satisfies Assumptions 1 and 2. Assume that $\|A_0\|_\infty \le \mathbf{a}$ for some constant $\mathbf{a}$ and $\sqrt{\rho} \ge \lambda \ge 3\|\Sigma\|/Q(A_0)$. Then, there exist numerical constants c_1', that depends only on K, such that with probability at least $1 - \frac{2}{d}$*

$$\frac{\|\hat{A}_{\mathrm{SQ}} - A_0\|_2^2}{m_1 m_2} \le c_1' \max\left\{ \mu^2 m_1 m_2 \operatorname{rank}(A_0) \big(Q^2(A_0) \lambda^2 + \mathbf{a}^2 \big(\mathbb{E}\big(\|\Sigma_R\| \big) \big)^2 \big), \right.$$
$$\left. \mathbf{a}^2 \mu \sqrt{\frac{\log(d)}{n}} \right\},$$

where $Q(A_0) = \sigma \sqrt{\frac{1}{n} \sum_{i=1}^{n} \xi_i^2}$.

In order to get a bound on the prediction risk in a closed form, we use the bounds on $\|\Sigma\|$ and $\mathbb{E}(\|\Sigma_R\|)$ given by Lemmas 5 and 6 taking $t = \log(d)$. It remains to bound $Q(A_0) = \sigma \sqrt{\frac{1}{n} \sum_{i=1}^{n} \xi_i^2}$. We consider the case of sub-Gaussian noise:

Assumption 9. *There exists a constant K such that*

$$\mathbb{E}\big[\exp(t\xi_i)\big] \leq \exp\big(t^2/2K\big)$$

for all $t > 0$.

Note that condition $\mathbb{E}\xi_i^2 = 1$ implies that $K \leq 1$. Under Assumption 9, ξ_i^2 are sub-exponential random variables. Then, the Bernstein inequality for sub-exponential random variables implies that, there exists a numerical constant c_3 such that, with probability at least $1 - 2\exp\{-c_3 n\}$, one has

$$3\sigma/2 \geq Q(A_0) \geq \sigma/2. \tag{13}$$

Using Lemma 5 and the right-hand side of (13), for $n \geq 2\log^2(d)m/L$, we can take

$$\lambda = 6C^* \sqrt{\frac{2L\log(d)}{mn}}. \tag{14}$$

Note that λ *does not depend* on σ and satisfies the two conditions required in Theorem 8. We have that

$$\lambda \geq 3\|\Sigma\|/Q(A_0) \tag{15}$$

with probability greater then $1 - 1/d - 2\exp\{-c_3 n\}$ and

$$\lambda^2 \leq \frac{1}{16\mu m_1 m_2 \operatorname{rank}(A_0)} \tag{16}$$

for n large enough, more precisely, for n such that

$$n \geq c_4 \mu L M \operatorname{rank}(A_0)\log(d), \tag{17}$$

where $c_4 = 576(C^*)^2$. We obtain the following theorem.

Theorem 10. *Let X_i be i.i.d. with distribution Π on $\mathcal{X}$ which satisfies Assumptions 1 and 2. Assume that $\|A_0\|_\infty \leq \mathbf{a}$ for some constant $\mathbf{a}$ and that Assumption 9 holds. Consider the regularization parameter λ satisfying (14) and n satisfying (17). Then, there exist numerical constants (c'', c_3) such that,*

$$\frac{\|\hat{A}_{\mathrm{SQ}} - A_0\|_2^2}{m_1 m_2} \leq c'' \max\left\{\max\big(\sigma^2, \mathbf{a}^2\big)\mu^2 L \frac{\log(d)\operatorname{rank}(A_0)M}{n}, \mathbf{a}^2\mu\sqrt{\frac{\log(d)}{n}}\right\} \tag{18}$$

with probability greater than $1 - 3/d - 2\exp\{-c_3 n\}$.

Note that condition (17) is not restrictive: indeed the sampling sizes n satisfying condition (17) are of the same order of magnitude as those for which the normalized Frobenius error of our estimator is small. Thus, Theorem 10 shows, that $\hat{A}_{\mathrm{SQ}}$ has the same prediction performances as previously proposed estimators which rely on the knowledge of the standard deviation of the noise and of the sampling distribution.

5. Bounds on the stochastic errors

In this section, we will obtain the upper bounds for the stochastic errors $\|\Sigma_R\|$ and $\mathbb{E}(\|\Sigma_R\|)$ defined in (4). In order to obtain such bounds, we use the matrix version of Bernstein's inequality. The following proposition is obtained by an extension of Theorem 4 in [15] to rectangular matrices via self-adjoint dilation (cf., for example, 2.6 in [25]). Let $Z_1, \dots, Z_n$ be independent random matrices with dimensions $m_1 \times m_2$. Define

$$\sigma_Z = \max\left\{ \left\| \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\big(Z_i Z_i^T\big) \right\|^{1/2}, \left\| \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\big(Z_i^T Z_i\big) \right\|^{1/2} \right\}$$

and

$$U_i = \inf\big\{ K > 0 : \mathbb{E}\exp\big(\|Z_i\|/K\big) \le e \big\}.$$

Proposition 11. *Let $Z_1, \dots, Z_n$ be independent random matrices with dimensions $m_1 \times m_2$ that satisfy $\mathbb{E}(Z_i) = 0$. Suppose that $U_i < U$ for some constant U and all $i = 1, \dots, n$. Then, there exists an absolute constant c^*, such that, for all $t > 0$, with probability at least $1 - e^{-t}$ we have*

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

where $d = m_1 + m_2$.

5.1. Proof of Lemma 5

We apply Proposition 11 to $Z_i = \zeta_i X_i$. We first estimate σ_Z and U. Note that Z_i is a zero-mean random matrix which satisfies

$$\|Z_i\| \le |\zeta_i|.$$

Then, Assumption 4 implies that there exists a constant K such that $U_i \le K$ for all $i = 1, \dots, n$. We compute

$$\mathbb{E}\big(Z_i Z_i^T\big) = R \quad \text{and} \quad \mathbb{E}\big(Z_i^T Z_i\big) = C,$$

where C (resp., R) is the diagonal matrix with C_k (resp., R_j) on the diagonal. This and the fact that the X_i are i.i.d. imply that

$$\sigma_Z^2 = \max_{i,j}(C_i, R_j) \le L/m.$$

Note that $\max_{i,j}(C_i, R_j) \ge 1/m$ which implies that $\log(K/\sigma_Z) \le \log(Km)$ and the statement of Lemma 5 follows.

5.2. Proof of Lemma 6

The proof follows the lines of the proof of Lemma 7 in [14]. For sake of completeness, we give it here. Set $t^* = \frac{Ln}{m\log^2(m)} - \log(d)$. t^* is the value of t such that the two terms in (7) are equal. Note that Lemma 5 implies that

$$\mathbb{P}\left(\left\|\frac{1}{n}\sum_{i=1}^{n}\zeta_i X_i\right\| > t\right) \leq d\exp\left\{-t^2 nm/\left(\left(C^*\right)^2 L\right)\right\} \qquad \text{for } t \leq t^* \tag{19}$$

and

$$\mathbb{P}\left(\left\|\frac{1}{n}\sum_{i=1}^{n}\zeta_i X_i\right\| > t\right) \leq d\exp\left\{-tn/\left(C^*\log(m)\right)\right\} \qquad \text{for } t \geq t^*. \tag{20}$$

We set $\nu_1 = nm/((C^*)^2 L)$, $\nu_2 = n/(C^*\log(m))$. By Hölder's inequality, we get

$$\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^{n}\zeta_i X_i\right\| \leq \left(\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^{n}\zeta_i X_i\right\|^{2\log(d)}\right)^{1/(2\log(d))}.$$

The inequalities (19) and (20) imply that

$$\begin{aligned}
&\left(\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^{n}\zeta_i X_i\right\|^{2\log(d)}\right)^{1/2\log(d)} \\
&= \left(\int_0^{+\infty}\mathbb{P}\left(\left\|\frac{1}{n}\sum_{i=1}^{n}\zeta_i X_i\right\| > t^{1/(2\log(d))}\right)\mathrm{d}t\right)^{1/2\log(d)} \\
&\leq \left(d\int_0^{+\infty}\exp\left\{-t^{1/\log(d)}\nu_1\right\}\mathrm{d}t + d\int_0^{+\infty}\exp\left\{-t^{1/(2\log(d))}\nu_2\right\}\mathrm{d}t\right)^{1/2\log(d)} \\
&\leq \sqrt{e}\left(\log(d)\nu_1^{-\log(d)}\Gamma\left(\log(d)\right) + 2\log(d)\nu_2^{-2\log(d)}\Gamma\left(2\log(d)\right)\right)^{1/(2\log(d))}.
\end{aligned} \tag{21}$$

The Gamma-function satisfies the following bound:

$$\text{for } x \geq 2 \qquad \Gamma(x) \leq \left(\frac{x}{2}\right)^{x-1} \tag{22}$$

(see, e.g., [14]). Plugging this into (21), we compute

$$\begin{aligned}
&\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^{n}\zeta_i X_i\right\| \\
&\leq \sqrt{e}\left(\left(\log(d)\right)^{\log(d)}\nu_1^{-\log(d)}2^{1-\log(d)} + 2\left(\log(d)\right)^{2\log(d)}\nu_2^{-2\log(d)}\right)^{1/(2\log(d))}.
\end{aligned}$$

Observe that $n > n^*$ implies $\nu_1 \log(d) \le \nu_2^2$ and we obtain

$$\mathbb{E}\left\| \frac{1}{n} \sum_{i=1}^{n} \zeta_i X_i \right\| \le \sqrt{\frac{2e \log(d)}{\nu_1}}. \tag{23}$$

We conclude the proof by plugging $\nu_1 = nm/((C^*)^2 L)$ into (23).

Appendix A: Proof of Theorem 3

It follows from the definition of the estimator $\hat{A}$ that

$$\frac{1}{n} \sum_{i=1}^{n} \big(Y_i - \langle X_i, \hat{A} \rangle\big)^2 + \lambda \|\hat{A}\|_1 \le \frac{1}{n} \sum_{i=1}^{n} \big(Y_i - \langle X_i, A_0 \rangle\big)^2 + \lambda \|A_0\|_1,$$

which, using (1), implies

$$\frac{1}{n} \sum_{i=1}^{n} \big(\langle X_i, A_0 \rangle + \sigma \xi_i - \langle X_i, \hat{A} \rangle\big)^2 + \lambda \|\hat{A}\|_1 \le \frac{\sigma^2}{n} \sum_{i=1}^{n} \xi_i^2 + \lambda \|A_0\|_1.$$

Hence,

$$\frac{1}{n} \sum_{i=1}^{n} \langle X_i, A_0 - \hat{A} \rangle^2 + 2\langle \Sigma, A_0 - \hat{A} \rangle + \lambda \|\hat{A}\|_1 \le \lambda \|A_0\|_1,$$

where $\Sigma = \frac{\sigma}{n} \sum_{i=1}^{n} \xi_i X_i$. Then, by the duality between the nuclear and the operator norms, we obtain

$$\frac{1}{n} \big\|\Omega(A_0 - \hat{A})\big\|_2^2 + \lambda \|\hat{A}\|_1 \le 2\|\Sigma\| \|A_0 - \hat{A}\|_1 + \lambda \|A_0\|_1. \tag{24}$$

Let P_S be the projector on the linear vector subspace S and let $S^\perp$ be the orthogonal complement of S. Let $u_j(A)$ and $v_j(A)$ denote, respectively, the *left* and *right* orthonormal *singular vectors* of A. $S_1(A)$ is the linear span of $\{u_j(A)\}$, $S_2(A)$ is the linear span of $\{v_j(A)\}$. We set

$$\mathbf{P}_A^\perp(B) = P_{S_1^\perp(A)} B P_{S_2^\perp(A)} \quad \text{and} \quad \mathbf{P}_A(B) = B - \mathbf{P}_A^\perp(B). \tag{25}$$

By definition of $\mathbf{P}_{A_0}^\perp$, for any matrix B, the singular vectors of $\mathbf{P}_{A_0}^\perp(B)$ are orthogonal to the space spanned by the singular vectors of A_0. This implies that $\|A_0 + \mathbf{P}_{A_0}^\perp(\hat{A} - A_0)\|_1 = \|A_0\|_1 + \|\mathbf{P}_{A_0}^\perp(\hat{A} - A_0)\|_1$. Then

$$\begin{aligned} \|\hat{A}\|_1 &= \|A_0 + \hat{A} - A_0\|_1 = \big\|A_0 + \mathbf{P}_{A_0}^\perp(\hat{A} - A_0) + \mathbf{P}_{A_0}(\hat{A} - A_0)\big\|_1 \\ &\ge \big\|A_0 + \mathbf{P}_{A_0}^\perp(\hat{A} - A_0)\big\|_1 - \big\|\mathbf{P}_{A_0}(\hat{A} - A_0)\big\|_1 \\ &= \|A_0\|_1 + \big\|\mathbf{P}_{A_0}^\perp(\hat{A} - A_0)\big\|_1 - \big\|\mathbf{P}_{A_0}(\hat{A} - A_0)\big\|_1. \end{aligned} \tag{26}$$

Note that from (26), we get

$$\|A_0\|_1 - \|\hat{A}\|_1 \le \big\|\mathbf{P}_{A_0}(A_0 - \hat{A})\big\|_1 - \big\|\mathbf{P}^{\perp}_{A_0}(A_0 - \hat{A})\big\|_1. \tag{27}$$

This, the triangle inequality and $\lambda \ge 3\|\Sigma\|$ lead to

$$\begin{aligned}\frac{1}{n}\big\|\Omega(A_0 - \hat{A})\big\|_2^2 &\le 2\|\Sigma\|\big\|\mathbf{P}_{A_0}(A_0 - \hat{A})\big\|_1 + \lambda\big\|\mathbf{P}_{A_0}(A_0 - \hat{A})\big\|_1 \\ &\le \frac{5}{3}\lambda\big\|\mathbf{P}_{A_0}(A_0 - \hat{A})\big\|_1.\end{aligned} \tag{28}$$

Since $\mathbf{P}_A(B) = P_{S_1^{\perp}(A)} B P_{S_2(A)} + P_{S_1(A)} B$ and $\operatorname{rank}(P_{S_i(A)} B) \le \operatorname{rank}(A)$ we have that $\operatorname{rank}(\mathbf{P}_A(B)) \le 2\operatorname{rank}(A)$. From (28), we compute

$$\frac{1}{n}\big\|\Omega(A_0 - \hat{A})\big\|_2^2 \le \frac{5}{3}\lambda\sqrt{2\operatorname{rank}(A_0)}\|\hat{A} - A_0\|_2. \tag{29}$$

For a $0 < r \le m$, we consider the following constrain set

$$\mathcal{C}(r) = \left\{A \in \mathbb{R}^{m_1 \times m_2} : \|A\|_\infty = 1, \|A\|_{L_2(\Pi)}^2 \ge \sqrt{\frac{64\log(d)}{\log(6/5)n}}, \|A\|_1 \le \sqrt{r}\|A\|_2\right\}. \tag{30}$$

Note that the condition $\|A\|_1 \le \sqrt{r}\|A\|_2$ is satisfied if $\operatorname{rank}(A) \le r$.

The following lemma shows that for matrices $A \in \mathcal{C}(r)$ the observation operator Ω satisfies some approximative restricted isometry. Its proof is given in Appendix B.

Lemma 12. *Let X_i be i.i.d. with distribution Π on $\mathcal{X}$ which satisfies Assumptions 1 and 2. Then, for all $A \in \mathcal{C}(r)$*

$$\frac{1}{n}\big\|\Omega(A)\big\|_2^2 \ge \frac{1}{2}\|A\|_{L_2(\Pi)}^2 - 44\mu r m_1 m_2\big(\mathbb{E}\big(\|\Sigma_R\|\big)\big)^2$$

with probability at least $1 - \frac{2}{d}$.

We need the following auxiliary lemma which is proven in Appendix E.

Lemma 13. *If $\lambda > 3\|\Sigma\|$*

$$\big\|\mathbf{P}^{\perp}_{A_0}(\hat{A} - A_0)\big\|_1 \le 5\big\|\mathbf{P}_{A_0}(\hat{A} - A_0)\big\|_1.$$

Lemma 13 implies that

$$\|\hat{A} - A_0\|_1 \le 6\big\|\mathbf{P}_{A_0}(\hat{A} - A_0)\big\|_1 \le \sqrt{72\operatorname{rank}(A_0)}\|\hat{A} - A_0\|_2. \tag{31}$$

Set $a = \|\hat{A} - A_0\|_\infty$. By definition of $\hat{A}$, we have that $a \le 2\mathbf{a}$. We now consider two cases, depending on whether the matrix $\frac{1}{a}(\hat{A} - A_0)$ belongs to the set $\mathcal{C}(72\operatorname{rank}(A_0))$ or not.

Case 1: Suppose first that $\|\hat{A} - A_0\|^2_{L_2(\Pi)} < a^2\sqrt{\frac{64\log(d)}{\log(6/5)n}}$, then (3) implies that

$$\frac{\|\hat{A} - A_0\|_2^2}{m_1 m_2} \le 4\mathbf{a}^2\mu\sqrt{\frac{64\log(d)}{\log(6/5)n}} \tag{32}$$

and we get the statement of Theorem 3 in this case.

Case 2: It remains to consider the case $\|\hat{A} - A_0\|^2_{L_2(\Pi)} \ge a^2\sqrt{\frac{64\log(d)}{\log(6/5)n}}$. Then (31) implies that $\frac{1}{a}(\hat{A} - A_0) \in \mathcal{C}(72\,\mathrm{rank}(A_0))$ and we can apply Lemma 12. From Lemma 12 and (29), we obtain that with probability at least $1 - \frac{2}{d}$ one has

$$\begin{aligned}\tfrac{1}{2}\|\hat{A} - A_0\|^2_{L_2(\Pi)} &\le \tfrac{5}{3}\lambda\sqrt{2\,\mathrm{rank}(A_0)}\|\hat{A} - A_0\|_2 + 3168\mu a^2\,\mathrm{rank}(A_0)m_1m_2\big(\mathbb{E}\big(\|\Sigma_R\|\big)\big)^2\\ &\le 6\lambda^2\mu m_1 m_2\,\mathrm{rank}(A_0) + \tfrac{1}{4}(m_1m_2\mu)^{-1}\|\hat{A} - A_0\|_2^2\\ &\quad + 3168\mu a^2\,\mathrm{rank}(A_0)m_1m_2\big(\mathbb{E}\big(\|\Sigma_R\|\big)\big)^2.\end{aligned}$$

Now (3) and $a \le 2\mathbf{a}$ imply that, there exist numerical constants c_1 such that

$$\|\hat{A} - A_0\|_2^2 \le c_1(\mu m_1 m_2)^2\,\mathrm{rank}(A_0)\big(\lambda^2 + \mathbf{a}^2\big(\mathbb{E}\big(\|\Sigma_R\|\big)\big)^2\big),$$

which, together with (32), leads to the statement of the Theorem 3.

Appendix B: Proof of Lemma 12

The main lines of this proof are close to those of the proof of Theorem 1 in [21]. Set $\mathcal{E} = 44\mu r m_1 m_2(\mathbb{E}(\|\Sigma_R\|))^2$. We will show that the probability of the following "bad" event is small

$$\mathcal{B} = \left\{\exists A \in \mathcal{C}(r) \text{ such that } \left|\frac{1}{n}\big\|\Omega(A)\big\|_2^2 - \|A\|^2_{L_2(\Pi)}\right| > \frac{1}{2}\|A\|^2_{L_2(\Pi)} + \mathcal{E}\right\}.$$

Note that $\mathcal{B}$ contains the complement of the event that we are interested in.

In order to estimate the probability of $\mathcal{B}$, we use a standard peeling argument. Let $\nu = \sqrt{\frac{64\log(d)}{\log(6/5)n}}$ and $\alpha = \frac{6}{5}$. For $l \in \mathbb{N}$ set

$$S_l = \big\{A \in \mathcal{C}(r) : \alpha^{l-1}\nu \le \|A\|^2_{L_2(\Pi)} \le \alpha^l\nu\big\}.$$

If the event $\mathcal{B}$ holds for some matrix $A \in \mathcal{C}(r)$, then A belongs to some S_l and

$$\begin{aligned}\left|\frac{1}{n}\big\|\Omega(A)\big\|_2^2 - \|A\|^2_{L_2(\Pi)}\right| &> \frac{1}{2}\|A\|^2_{L_2(\Pi)} + \mathcal{E}\\ &> \frac{1}{2}\alpha^{l-1}\nu + \mathcal{E}\\ &= \frac{5}{12}\alpha^l\nu + \mathcal{E}.\end{aligned} \tag{33}$$

For each $T > \nu$ consider the following set of matrices

$$\mathcal{C}(r,T) = \left\{ A \in \mathcal{C}(r) : \|A\|^2_{L_2(\Pi)} \le T \right\}$$

and the following event

$$\mathcal{B}_l = \left\{ \exists A \in \mathcal{C}\big(r, \alpha^l \nu\big) : \left| \frac{1}{n} \big\|\Omega(A)\big\|_2^2 - \|A\|^2_{L_2(\Pi)} \right| > \frac{5}{12} \alpha^l \nu + \mathcal{E} \right\}.$$

Note that $A \in S_l$ implies that $A \in \mathcal{C}(r, \alpha^l \nu)$. Then (33) implies that $\mathcal{B}_l$ holds and we get $\mathcal{B} \subset \bigcup \mathcal{B}_l$. Thus, it is enough to estimate the probability of the simpler event $\mathcal{B}_l$ and then apply the union bound. Such an estimation is given by the following lemma. Its proof is given in Appendix C. Let

$$Z_T = \sup_{A \in \mathcal{C}(r,T)} \left| \frac{1}{n} \big\|\Omega(A)\big\|_2^2 - \|A\|^2_{L_2(\Pi)} \right|.$$

Lemma 14. *Let X_i be i.i.d. with distribution Π on $\mathcal{X}$ which satisfies Assumptions 1 and 2. Then,*

$$\mathbb{P}\big(Z_T > \tfrac{5}{12} T + 44 \mu r m_1 m_2 \big(\mathbb{E}\big(\|\Sigma_R\|\big)\big)^2\big) \le \exp\big(-c_5 n T^2\big),$$

where $c_5 = \frac{1}{128}$.

Lemma 14 implies that $\mathbb{P}(\mathcal{B}_l) \le \exp(-c_5 n \alpha^{2l} \nu^2)$. Using the union bound, we obtain

$$\begin{aligned} \mathbb{P}(\mathcal{B}) &\le \sum_{l=1}^{\infty} \mathbb{P}(\mathcal{B}_l) \\ &\le \sum_{l=1}^{\infty} \exp\big(-c_5 n \alpha^{2l} \nu^2\big) \\ &\le \sum_{l=1}^{\infty} \exp\big(-\big(2 c_5 n \log(\alpha) \nu^2\big) l\big), \end{aligned}$$

where we used $e^x \ge x$. We finally compute for $\nu = \sqrt{\frac{64 \log(d)}{\log(6/5) n}}$

$$\mathbb{P}(\mathcal{B}) \le \frac{\exp(-2 c_5 n \log(\alpha) \nu^2)}{1 - \exp(-2 c_5 n \log(\alpha) \nu^2)} = \frac{\exp(-\log(d))}{1 - \exp(-\log(d))}.$$

This completes the proof of Lemma 12.

Remark. As we mentioned in the beginning, the main lines of this proof are close to those of the proof of Theorem 1 in [21]. Let us briefly discuss the main differences between these two proofs.

Similarly to Theorem 1 in [21] we prove a kind of "restricted strong convexity" on a constrain set. However, our constrain set defined by (30) is quite different from the one introduced in [21]:

$$\mathcal{C}(n; c_0) = \left\{ A \in \mathbb{R}^{m_1 \times m_2} : \frac{\sqrt{m_1 m_2}\|A\|_1 \|A\|_\infty}{\|A\|_2^2} \le \frac{1}{c_0}\sqrt{\frac{n}{d\log(d)}} \right\}.$$

The present proof is also less involved (e.g., we do not need use the covering argument used in [21]). One important ingredient of our proof is a more efficient control of $\mathbb{E}\|\Sigma_R\|$ given by Lemma 6 (compare with Lemma 6 in [21]).

Appendix C: Proof of Lemma 14

Our approach is standard: first we show that Z_T concentrates around its expectation and then we upper bound the expectation.

By definition, $Z_T = \sup_{A\in\mathcal{C}(r,T)} |\frac{1}{n}\sum_{i=1}^n \langle X_i, A\rangle^2 - \mathbb{E}(\langle X, A\rangle^2)|$. Massart's concentration inequality (see, e.g., [2], Theorem 14.2) implies that

$$\mathbb{P}\big(Z_T \ge \mathbb{E}(Z_T) + \tfrac{1}{9}\big(\tfrac{5}{12}T\big)\big) \le \exp\big(-c_5 n T^2\big), \tag{34}$$

where $c_5 = \frac{1}{128}$. Next, we bound the expectation $\mathbb{E}(Z_T)$. Using a standard symmetrization argument (see, e.g., [16], Theorem 2.1), we obtain

$$\begin{aligned}\mathbb{E}(Z_T) &= \mathbb{E}\left(\sup_{A\in\mathcal{C}(r,T)} \left| \frac{1}{n}\sum_{i=1}^n \langle X_i, A\rangle^2 - \mathbb{E}\big(\langle X, A\rangle^2\big) \right| \right) \\ &\le 2\mathbb{E}\left(\sup_{A\in\mathcal{C}(r,T)} \left| \frac{1}{n}\sum_{i=1}^n \varepsilon_i \langle X_i, A\rangle^2 \right| \right),\end{aligned}$$

where $\{\varepsilon_i\}_{i=1}^n$ is an i.i.d. Rademacher sequence. The assumption $\|A\|_\infty = 1$ implies $|\langle X_i, A\rangle| \le 1$. Then, the contraction inequality (see, e.g., [16]) yields

$$\mathbb{E}(Z_T) \le 8\mathbb{E}\left(\sup_{A\in\mathcal{C}(r,T)} \left| \frac{1}{n}\sum_{i=1}^n \varepsilon_i \langle X_i, A\rangle \right| \right) = 8\mathbb{E}\Big(\sup_{A\in\mathcal{C}(r,T)} \big|\langle \Sigma_R, A\rangle\big| \Big),$$

where $\Sigma_R = \frac{1}{n}\sum_{i=1}^n \varepsilon_i X_i$. For $A \in \mathcal{C}(r, T)$, we have that

$$\begin{aligned}\|A\|_1 &\le \sqrt{r}\|A\|_2 \\ &\le \sqrt{\mu r m_1 m_2}\|A\|_{L_2(\Pi)} \\ &\le \sqrt{\mu m_1 m_2 r T},\end{aligned}$$

where we have used (3). Then, by the duality between nuclear and operator norms, we compute

$$\mathbb{E}(Z_T) \le 8\mathbb{E}\Big(\sup_{\|A\|_1 \le \sqrt{\mu m_1 m_2 r T}} \big|\langle \Sigma_R, A\rangle\big| \Big) \le 8\sqrt{\mu m_1 m_2 r T}\,\mathbb{E}\big(\|\Sigma_R\|\big).$$

Finally, using

$$\tfrac{1}{9}\big(\tfrac{5}{12}T\big)+8\sqrt{\mu m_1 m_2 r T}\,\mathbb{E}\big(\|\Sigma_R\|\big)\le\big(\tfrac{1}{9}+\tfrac{8}{9}\big)\tfrac{5}{12}T+44\mu r m_1 m_2\big(\mathbb{E}\big(\|\Sigma_R\|\big)\big)^2$$

and the concentration bound (34), we obtain that

$$\mathbb{P}\big(Z_T>\tfrac{5}{12}T+44\mu r m_1 m_2\big(\mathbb{E}\big(\|\Sigma_R\|\big)\big)^2\big)\le\exp\big(-c_5 n T^2\big)$$

with $c_5=\frac{1}{128}$ as stated.

Appendix D: Proof of Theorem 8

Let us set $\Delta=A_0-\hat{A}_{\mathrm{SQ}}$ and $Q(A)=\sqrt{\frac{1}{n}\sum_{i=1}^n(Y_i-\langle X_i,A\rangle)^2}$. We have that

$$\begin{aligned}Q^2(\hat{A}_{\mathrm{SQ}})-Q^2(A_0)&=\frac{1}{n}\big\|\Omega(\Delta)\big\|_2^2+2\left\langle\frac{\sigma}{n}\sum_{i=1}^n\xi_i X_i,\Delta\right\rangle\\&=\frac{1}{n}\big\|\Omega(\Delta)\big\|_2^2+2\langle\Sigma,\Delta\rangle,\end{aligned}$$

where $\Sigma=\frac{\sigma}{n}\sum_{i=1}^n\xi_i X_i$. This implies

$$\frac{1}{n}\big\|\Omega(\Delta)\big\|_2^2=-2\langle\Sigma,\Delta\rangle+\big(Q(\hat{A}_{\mathrm{SQ}})-Q(A_0)\big)\big(Q(\hat{A}_{\mathrm{SQ}})+Q(A_0)\big).\tag{35}$$

We need the following auxiliary lemma which is proven in Appendix F ($\mathbf{P}_{A_0}^{\perp}$ and $\mathbf{P}_{A_0}$ are defined in (25)).

Lemma 15. *If* $\lambda>3\|\Sigma\|/Q(A_0)$, *then*

$$\big\|\mathbf{P}_{A_0}^{\perp}(\Delta)\big\|_1\le 2\big\|\mathbf{P}_{A_0}(\Delta)\big\|_1,$$

where $\Delta=\hat{A}_{\mathrm{SQ}}-A_0$.

Note that from (26) we get

$$\|A_0\|_1-\|\hat{A}_{\mathrm{SQ}}\|_1\le\big\|\mathbf{P}_{A_0}(\Delta)\big\|_1-\big\|\mathbf{P}_{A_0}^{\perp}(\Delta)\big\|_1.\tag{36}$$

The definition of $\hat{A}_{\mathrm{SQ}}$ and (36) imply that

$$\begin{aligned}Q(A_0)+Q(\hat{A}_{\mathrm{SQ}})&\le 2Q(A_0)+\lambda\big(\|A_0\|_1-\|\hat{A}_{\mathrm{SQ}}\|_1\big)\\&\le 2Q(A_0)+\lambda\big(\big\|\mathbf{P}_{A_0}(\Delta)\big\|_1-\big\|\mathbf{P}_{A_0}^{\perp}(\Delta)\big\|_1\big)\end{aligned}\tag{37}$$

and

$$\begin{aligned} Q(\hat{A}_{\mathrm{SQ}}) - Q(A_0) &\le \lambda\big(\|A_0\|_1 - \|\hat{A}_{\mathrm{SQ}}\|_1\big) \\ &\le \lambda\big(\big\|\mathbf{P}_{A_0}(\Delta)\big\|_1 - \big\|\mathbf{P}^{\perp}_{A_0}(\Delta)\big\|_1\big) \\ &\le \lambda\big(2\big\|\mathbf{P}_{A_0}(\Delta)\big\|_1 - \big\|\mathbf{P}^{\perp}_{A_0}(\Delta)\big\|_1\big). \end{aligned} \tag{38}$$

Lemma 15 implies that $2\|\mathbf{P}_{A_0}(\Delta)\|_1 - \|\mathbf{P}^{\perp}_{A_0}(\Delta)\|_1 \ge 0$. From (37) and (38), we compute

$$\begin{aligned} &\big(Q(\hat{A}_{\mathrm{SQ}}) - Q(A_0)\big)\big(Q(\hat{A}_{\mathrm{SQ}}) + Q(A_0)\big) \\ &\quad\le \lambda\big(2\big\|\mathbf{P}_{A_0}(\Delta)\big\|_1 - \big\|\mathbf{P}^{\perp}_{A_0}(\Delta)\big\|_1\big)\big(2Q(A_0) + \lambda\big(\big\|\mathbf{P}_{A_0}(\Delta)\big\|_1 - \big\|\mathbf{P}^{\perp}_{A_0}(\Delta)\big\|_1\big)\big) \\ &\quad= \lambda Q(A_0)\big\|\mathbf{P}_{A_0}(\Delta)\big\|_1 - 2\lambda Q(A_0)\big\|\mathbf{P}^{\perp}_{A_0}(\Delta)\big\|_1 \\ &\qquad+ 2\lambda^2\big\|\mathbf{P}_{A_0}(\Delta)\big\|_1^2 + \lambda^2\big\|\mathbf{P}^{\perp}_{A_0}(\Delta)\big\|_1^2 - 3\lambda^2\big\|\mathbf{P}_{A_0}(\Delta)\big\|_1\big\|\mathbf{P}^{\perp}_{A_0}(\Delta)\big\|_1. \end{aligned} \tag{39}$$

Lemma 15 implies that $\lambda^2\|\mathbf{P}^{\perp}_{A_0}(\Delta)\|_1^2 - 3\lambda^2\|\mathbf{P}_{A_0}(\Delta)\|_1\|\mathbf{P}^{\perp}_{A_0}(\Delta)\|_1 \le 0$ and we obtain from (39)

$$\begin{aligned} &\big(Q(\hat{A}_{\mathrm{SQ}}) - Q(A_0)\big)\big(Q(\hat{A}_{\mathrm{SQ}}) + Q(A_0)\big) \\ &\quad\le 4\lambda Q(A_0)\big\|\mathbf{P}_{A_0}(\Delta)\big\|_1 - 2\lambda Q(A_0)\big\|\mathbf{P}^{\perp}_{A_0}(\Delta)\big\|_1 + 2\lambda^2\big\|\mathbf{P}_{A_0}(\Delta)\big\|_1^2. \end{aligned} \tag{40}$$

Plugging (40) into (35), we get

$$\begin{aligned} \frac{1}{n}\big\|\Omega(\Delta)\big\|_2^2 \le{}& -2\langle\Sigma, \Delta\rangle + 4\lambda Q(A_0)\big\|\mathbf{P}_{A_0}(\Delta)\big\|_1 \\ &- 2\lambda Q(A_0)\big\|\mathbf{P}^{\perp}_{A_0}(\Delta)\big\|_1 + 2\lambda^2\big\|\mathbf{P}_{A_0}(\Delta)\big\|_1^2. \end{aligned}$$

Then, by the duality between the nuclear and the operator norms, we obtain

$$\begin{aligned} \frac{1}{n}\big\|\Omega(\Delta)\big\|_2^2 \le{}& 2\|\Sigma\|\big\|\mathbf{P}_{A_0}(\Delta)\big\|_1 + 2\|\Sigma\|\big\|\mathbf{P}^{\perp}_{A_0}(\Delta)\big\|_1 \\ &+ 4\lambda Q(A_0)\big\|\mathbf{P}_{A_0}(\Delta)\big\|_1 - 2\lambda Q(A_0)\big\|\mathbf{P}^{\perp}_{A_0}(\Delta)\big\|_1 \\ &+ 2\lambda^2\big\|\mathbf{P}_{A_0}(\Delta)\big\|_1^2. \end{aligned}$$

Using $\lambda Q(A_0) \ge 3\|\Sigma\|$ we compute

$$\frac{1}{n}\big\|\Omega(\Delta)\big\|_2^2 \le \frac{14}{3}\lambda Q(A_0)\big\|\mathbf{P}_{A_0}(\Delta)\big\|_1 + 2\lambda^2\big\|\mathbf{P}_{A_0}(\Delta)\big\|_1^2,$$

which leads to

$$\frac{1}{n}\big\|\Omega(\Delta)\big\|_2^2 \le \frac{14}{3}\lambda Q(A_0)\sqrt{2\operatorname{rank}(A_0)}\|\Delta\|_2 + 4\lambda^2\operatorname{rank}(A_0)\|\Delta\|_2^2.$$

The condition $4\mu m_1 m_2 \lambda^2 \operatorname{rank}(A_0) \le 1/4$ implies that

$$\frac{1}{n}\big\|\Omega(\Delta)\big\|_2^2 \le \frac{14}{3}\lambda Q(A_0)\sqrt{2\operatorname{rank}(A_0)}\|\Delta\|_2 + \frac{\|\Delta\|_2^2}{4\mu m_1 m_2}. \tag{41}$$

Set $a = \|\hat{A}_{\mathrm{SQ}} - A_0\|_\infty$. By the definition of $\hat{A}_{\mathrm{SQ}}$ we have that $a \le 2\mathbf{a}$. We now consider two cases, depending on whether the matrix $\frac{1}{a}(\hat{A}_{\mathrm{SQ}} - A_0)$ belongs or not to the set $\mathcal{C}(18\operatorname{rank}(A_0))$.

Case 1: Suppose first that $\|\hat{A}_{\mathrm{SQ}} - A_0\|_{L_2(\Pi)}^2 < a^2\sqrt{\frac{64\log(d)}{\log(6/5)n}}$, then (3) implies that

$$\frac{\|\hat{A}_{\mathrm{SQ}} - A_0\|_2^2}{m_1 m_2} \le 4\mathbf{a}^2\mu\sqrt{\frac{64\log(d)}{\log(6/5)n}} \tag{42}$$

and we get the statement of the Theorem 8 in this case.

Case 2: It remains to consider the case $\|\hat{A}_{\mathrm{SQ}} - A_0\|_{L_2(\Pi)}^2 \ge a^2\sqrt{\frac{64\log(d)}{\log(6/5)n}}$. Lemma 15 implies that $\frac{1}{a}(\hat{A}_{\mathrm{SQ}} - A_0) \in \mathcal{C}(18\operatorname{rank}(A_0))$ and we can apply Lemma 12. From Lemma 12, (3) and (41) we obtain that, with probability at least $1 - \frac{2}{d}$ one has

$$\begin{aligned}\frac{\|\Delta\|_2^2}{2\mu m_1 m_2} &\le \frac{14}{3}\lambda Q(A_0)\sqrt{2\operatorname{rank}(A_0)}\|\Delta\|_2 + \frac{\|\Delta\|_2^2}{4\mu m_1 m_2} \\ &\quad + 792a^2\mu m_1 m_2 \operatorname{rank}(A_0)\big(\mathbb{E}\big(\|\Sigma_R\|\big)\big)^2.\end{aligned}$$

A simple calculation yields

$$\begin{aligned}&\left(\frac{\|\Delta\|_2}{2\sqrt{\mu m_1 m_2}} - \frac{14}{3}\lambda Q(A_0)\sqrt{2\operatorname{rank}(A_0)\mu m_1 m_2}\right)^2 \\ &\quad\le \left(\frac{14}{3}\lambda Q(A_0)\sqrt{2\operatorname{rank}(A_0)\mu m_1 m_2}\right)^2 + 792a^2\mu m_1 m_2 \operatorname{rank}(A_0)\big(\mathbb{E}\big(\|\Sigma_R\|\big)\big)^2\end{aligned}$$

and

$$\begin{aligned}\frac{\|\Delta\|_2}{2\sqrt{\mu m_1 m_2}} &\le \frac{28}{3}\lambda Q(A_0)\sqrt{2\operatorname{rank}(A_0)\mu m_1 m_2} \\ &\quad + \sqrt{792a^2\mu m_1 m_2 \operatorname{rank}(A_0)\big(\mathbb{E}\big(\|\Sigma_R\|\big)\big)^2}.\end{aligned} \tag{43}$$

This and $a \le 2\mathbf{a}$ imply that, there exist numerical constant c_1' such that

$$\frac{\|\hat{A}_{\mathrm{SQ}} - A_0\|_2^2}{m_1 m_2} \le c_1'\mu^2 m_1 m_2\big(Q^2(A_0)\lambda^2\operatorname{rank}(A_0) + \mathbf{a}^2\operatorname{rank}(A_0)\big(\mathbb{E}\big(\|\Sigma_R\|\big)\big)^2\big),$$

which, together with (42), leads to the statement of the Theorem 8.

Appendix E: Proof of Lemma 13

By the convexity of $Q^2(A)$ and using $\lambda \geq 3\Delta$ we have

$$\begin{aligned} Q^2(\hat{A}) - Q^2(A_0) &\geq -\frac{2}{n}\sum_{i=1}^{n}\big(Y_i - \langle X_i, A_0\rangle\big)\langle X_i, \hat{A} - A_0\rangle \\ &= -2\langle \Sigma, \hat{A} - A_0\rangle \\ &\geq -2\|\Sigma\|\|\hat{A} - A_0\|_1 \\ &\geq -\frac{2}{3}\lambda\|\hat{A} - A_0\|_1. \end{aligned}$$

Using the definition of $\hat{A}$, we compute

$$\begin{aligned} \lambda\|\hat{A}\|_1 - \lambda\|A_0\|_1 &\leq Q^2(A_0) - Q^2(\hat{A}) \\ &\leq \tfrac{2}{3}\lambda\|\hat{A} - A_0\|_1. \end{aligned}$$

This and (26) implies that

$$\big\|\mathbf{P}_{A_0}^{\perp}(\hat{A} - A_0)\big\|_1 \leq 5\big\|\mathbf{P}_{A_0}(\hat{A} - A_0)\big\|_1$$

as stated.

Appendix F: Proof of Lemma 15

By the convexity of $Q(A)$, we have

$$\begin{aligned} Q(\hat{A}_{\mathrm{SQ}}) - Q(A_0) &\geq \frac{-(\sum_{i=1}^{n}(Y_i - \langle X_i, A_0\rangle)\langle X_i, \hat{A}_{\mathrm{SQ}} - A_0\rangle)/n}{Q(A_0)} \\ &= \frac{-\langle \Sigma, \hat{A}_{\mathrm{SQ}} - A_0\rangle}{Q(A_0)} \\ &\geq -\frac{\|\Sigma\|}{Q(A_0)}\|\hat{A}_{\mathrm{SQ}} - A_0\|_1 \\ &\geq -\frac{1}{3}\lambda\|\hat{A}_{\mathrm{SQ}} - A_0\|_1. \end{aligned}$$

Using the definition of $\hat{A}_{\mathrm{SQ}}$, we compute

$$\begin{aligned} \lambda\|\hat{A}_{\mathrm{SQ}}\|_1 - \lambda\|A_0\|_1 &\leq Q(A_0) - Q(\hat{A}_{\mathrm{SQ}}) \\ &\leq \tfrac{1}{3}\lambda\|\hat{A}_{\mathrm{SQ}} - A_0\|_1. \end{aligned}$$

Then (26) and the triangle inequality imply

$$\left\|\mathbf{P}_{A_0}^{\perp}(\hat{A}-A_0)\right\|_1 - \left\|\mathbf{P}_{A_0}(\hat{A}-A_0)\right\|_1 \le \tfrac{1}{3}\left(\left\|\mathbf{P}_{A_0}^{\perp}(\hat{A}-A_0)\right\|_1 + \left\|\mathbf{P}_{A_0}(\hat{A}-A_0)\right\|_1\right)$$

and the statement of Lemma 15 follows.

Acknowledgement

I would like to thank Miao Weimin for his interesting comment.

References

[1] Belloni, A., Chernozhukov, V. and Wang, L. (2011). Square-root lasso: Pivotal recovery of sparse signals via conic programming. *Biometrika* **98** 791–806. MR2860324
[2] Bühlmann, P. and van de Geer, S. (2011). *Statistics for High-Dimensional Data*: *Methods*, *Theory and Applications*. *Springer Series in Statistics*. Heidelberg: Springer. MR2807761
[3] Bunea, F., She, Y. and Wegkamp, M.H. (2011). Optimal selection of reduced rank estimators of high-dimensional matrices. *Ann. Statist.* **39** 1282–1309. MR2816355
[4] Candès, E.J. and Plan, Y. (2009). Matrix completion with noise. *Proceedings of IEEE* **98** 925–936.
[5] Candès, E.J. and Recht, B. (2009). Exact matrix completion via convex optimization. *Found. Comput. Math.* **9** 717–772. MR2565240
[6] Candès, E.J. and Tao, T. (2010). The power of convex relaxation: Near-optimal matrix completion. *IEEE Trans. Inform. Theory* **56** 2053–2080. MR2723472
[7] Foygel, R., Salakhutdinov, R., Shamir, O. and Srebro, N. (2011). Learning with the weighted trace-norm under arbitrary sampling distributions. In *Advances in Neural Information Processing Systems* (*NIPS*) **24**.
[8] Foygel, R. and Srebro, N. (2011). Concentration-based guarantees for low-rank matrix reconstruction. In *24nd Annual Conference on Learning Theory* (*COLT*).
[9] Gaïffas, S. and Lecué, G. (2011). Sharp oracle inequalities for high-dimensional matrix prediction. *IEEE Trans. Inform. Theory* **57** 6942–6957. MR2882272
[10] Gross, D. (2011). Recovering low-rank matrices from few coefficients in any basis. *IEEE Trans. Inform. Theory* **57** 1548–1566. MR2815834
[11] Keshavan, R.H., Montanari, A. and Oh, S. (2010). Matrix completion from noisy entries. *J. Mach. Learn. Res.* **11** 2057–2078. MR2678022
[12] Keshavan, R.H., Montanari, A. and Oh, S. (2010). Matrix completion from a few entries. *IEEE Trans. Inform. Theory* **56** 2980–2998. MR2683452
[13] Klopp, O. (2011). Matrix completion with unknown variance of the noise. Available at http://arxiv.org/abs/1112.3055.
[14] Klopp, O. (2011). Rank penalized estimators for high-dimensional matrices. *Electron. J. Stat.* **5** 1161–1183. MR2842903
[15] Koltchinskii, V. (2011). A remark on low rank matrix recovery and noncommutative Bernstein type inequalities. In *IMS Collections*, *Festschrift in Honor of J. Wellner.*
[16] Koltchinskii, V. (2011). *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems*. *Lecture Notes in Math.* **2033**. Heidelberg: Springer. MR2829871
[17] Koltchinskii, V. (2011). Von Neumann entropy penalization and low rank matrix estimation. *Ann. Statist.* **39** 2936–2973.

[18] Koltchinskii, V., Lounici, K. and Tsybakov, A.B. (2011). Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *Ann. Statist.* **39** 2302–2329. MR2906869

[19] Lounici, K. (2011). Optimal spectral norm rates for noisy low-rank matrix completion. Available at http://arxiv.org/abs/1110.5346.

[20] Negahban, S. and Wainwright, M.J. (2011). Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *Ann. Statist.* **39** 1069–1097. MR2816348

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

[22] Recht, B. (2011). A simpler approach to matrix completion. *J. Mach. Learn. Res.* **12** 3413–3430. MR2877360

[23] Rohde, A. and Tsybakov, A.B. (2011). Estimation of high-dimensional low-rank matrices. *Ann. Statist.* **39** 887–930. MR2816342

[24] Salakhutdinov, R. and Srebro, N. (2010). Collaborative filtering in a non-uniform world: Learning with the weighted trace norm. In *Advances in Neural Information Processing Systems* (*NIPS*) **23**.

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

Received March 2012 and revised July 2012