High dimensional matrix estimation with unknown variance of the noise

Olga Klopp, Stéphane Gaiffas

HIGH DIMENSIONAL MATRIX ESTIMATION WITH UNKNOWN VARIANCE OF THE NOISE

OLGA KLOPP

Abstract. Assume that we observe a small set of entries or linear combinations of entries of an unknown matrix A_0 corrupted by noise. We propose a new method for estimating A_0 which does not rely on the knowledge or an estimation of the standard deviation of the noise σ. Our estimator achieves, up to a logarithmic factor, optimal rates of convergence under the Frobenius risk and, thus, has the same prediction performance as previously proposed estimators which rely on the knowledge of σ.

1. Introduction

The problem of the recovery of a data matrix from incomplete and corrupted information appears in a variety of applications such as recommendation systems, system identification, global positioning, remote sensing (for more details see [3]). For instance, in the Netflix recommendation system, we observe a few movie ratings from a large data matrix in which rows are users and columns are movies. Each user only watches a few movies compared to the total database of movies available on Netflix. The goal is to predict the missing ratings in order to be able to recommend the movies to a person that he/she has not yet seen.

In the noiseless setting, if the unknown matrix has low rank and is "incoherent", then it can be reconstructed exactly with high probability from a small set of entries. This result was first proved by Candès and Recht [4] using nuclear norm minimization. A tighter analysis of the same convex relaxation was carried out in [5]. For a simpler approach see [16] and [8]. An alternative line of work was developed by Keshavan et al in [10]. More recent results of Gross [8] and Recht [16] provide sharper conditions. For example, Recht [16] showed that, if we observe n entries of a matrix $A_0 \in \mathbb{R}^{m_1 \times m_2}$ with locations uniformly sampled at random, then under "incoherence conditions" the exact recovery is possible with high probability if $n > Cr(m_1 + m_2)\log^2 m_2$ with some constant $C > 0$ and $r = \operatorname{rank}(A_0)$.

2000 *Mathematics Subject Classification.* 62J99, 62H12, 60B20, 60G05.

Key words and phrases. Matrix completion, matrix regression, low rank matrix estimation, recovery of the rank.

In a more realistic setting the observed entries are corrupted by noise. This question has been recently addressed by several authors (see, e.g., [3, 9, 17, 14, 15, 12, 13, 6, 11]). These methods rely on the knowledge or a pre-estimation of the standard deviation σ of the noise. Estimation of σ is non-trivial in the large-scaled problems. The estimator that we propose in the present paper eliminates the need to know or to pre-estimate σ. It is inspired (but leads to a different analysis) by the square-root lasso estimator proposed for the linear regression model by Belloni et al in [1]. We show that, up to a logarithmic factor, our estimator achieves optimal rates of convergence under the Frobenius risk. Thus, it has the same prediction performance as previously proposed estimators which rely on the knowledge of σ.

This paper is organized as follows. In Section 2 we set notations, introduce our model - the trace regression model and our estimator. In Section 3 (Theorem 2), we prove a general oracle inequality for the prediction error for the trace regression model.

In the Section 4, we apply Theorem 2 to the case of matrix completion under uniform sampling at random (USR). We propose a choice of the regularization parameter λ for our estimator which is independent of σ. The main result, Theorem 6, shows that in the case of USR matrix completion and under some mild conditions that link the rank and the "spikiness" of A_0, up to a constant, the prediction risk of our estimator is comparable to the sharpest bounds obtained until now. For more details see Section 4.

In Section 5, we apply our idea to the problem of matrix regression which is yet another special case of trace regression. Previously, the problem of matrix regression with unknown noise variance was considered in [2, 7]. These two papers study the rank-penalized estimators. Bunea et al [2], who first introduced the idea of such estimators, propose un unbiased estimator of σ which requires an assumption on the dimensions of the problem. This assumption excludes an interesting case, the case when the sample size is smaller than the number of covariates. The method proposed in [7] can be applied to this last case under a condition on the rank of the unknown matrix A_0. Our method, unlike the method of [2], can be applied to the case when the sample size is smaller than the number of covariates and our condition is weaker than the conditions obtained in [7]. For more details see Section 5.

2. Preliminaries

2.1. **Model.** 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

$$Y_i = \operatorname{tr}(X_i^T A_0) + \sigma \xi_i,\ i = 1, \dots, n. \tag{2.1}$$

Here, $\sigma > 0$ is the unknown standard deviation. The noise variables ξ_i are independent, identically distributed and having law Φ such that

$$\mathbb{E}_\Phi(\xi_i) = 0,\ \mathbb{E}_\Phi(\xi_i^2) = 1, \tag{2.2}$$

X_i are random matrices with dimension $m_1 \times m_2$ and $\mathrm{tr}(A)$ denotes the trace of the matrix A.

We consider the problem of estimating of A_0. Our main motivation is the high-dimensional setting, which corresponds to $m_1 m_2 \gg n$, with low rank matrices A_0.

The trace regression model is a quite general model which contains as particular cases a number of interesting problems. Let us give two examples which we will consider with more details in this paper.

- **Matrix Completion** Assume that the design matrices X_i are i.i.d uniformly distributed on the set

$$\mathcal{X} = \left\{ e_j(m_1) e_k^T(m_2), 1 \le j \le m_1, 1 \le k \le m_2 \right\}, \tag{2.3}$$

  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 under uniform sampling at random (USR).
- **Matrix regression** The matrix regression model is given by

$$U_i = V_i\, A_0 + E_i \qquad i = 1, \dots, l, \tag{2.4}$$

  where U_i are $1 \times m_2$ vectors of response variables, V_i are $1 \times m_1$ vectors of predictors, A_0 is an unknown $m_1 \times m_2$ matrix of regression coefficients and E_i are random $1 \times m_2$ vectors of noise with independent entries and mean zero.

  We can equivalently write this model as a trace regression model. Let $U_i = (U_{ik})_{k=1,\dots,m_2}$, $E_i = (E_{ik})_{k=1,\dots,m_2}$ and $Z_{ik}^T = e_k(m_2)\, V_i$, where $e_k(m_2)$ are the $m_2 \times 1$ vectors of the canonical basis of $\mathbb{R}^{m_2}$. Then, we can write (2.4) as

$$U_{ik} = \mathrm{tr}(Z_{ik}^T A_0) + E_{ik} \qquad i = 1, \dots, l \quad \text{and} \quad k = 1, \dots, m_2.$$

2.2. **Notation.** For any matrices $A, B \in \mathbb{R}^{m_1 \times m_2}$, we define the scalar product

$$\langle A, B \rangle = \mathrm{tr}(A^T B).$$

For $0 < q \le \infty$ the Schatten-q (quasi-)norm of the matrix A is defined by

$$\|A\|_q = \left(\sum_{j=1}^{\min(m_1,m_2)} \sigma_j(A)^q \right)^{1/q} \text{ for } 0 < q < \infty \text{ and } \|A\|_\infty = \sigma_1(A),$$

where $(\sigma_j(A))_j$ are the singular values of A ordered decreasingly.

We summarize the notations which we use throughout this paper

-

$$\mathbf{X} = \frac{\mu^2}{n} \sum_{i=1}^n Y_i X_i \ \text{ and } \ \mathbf{M} = \mu^{-2} (\mathbf{X} - A_0)\,; \tag{2.5}$$

•

$$\Delta = \frac{\|\mathbf{M}\|_\infty}{\|\mathbf{M}\|_2} \text{ and } \Delta_\infty = \|\mathbf{M}\|_\infty; \tag{2.6}$$

- ∂G is the subdifferential of G;
- $S^\perp$ is the orthogonal complement of S;
- P_S is the projector on the linear vector subspace S;
- $F(A) = \|A - \mathbf{X}\|_2 + \lambda\|A\|_1$.
- $\|A\|_{\sup} = \max_{i,j} \mid a_{ij} \mid$ where $A = (a_{ij})$.

2.3. **Estimator.** In [13], the authors propose the following estimator for the trace regression model

$$\hat{A}_{ML} = \operatorname*{arg\,min}_{A\in\mathbb{R}^{m_1\times m_2}} \left\{ \| A \|^2_{L_2(\Pi)} - \left\langle \frac{2}{n}\sum_{i=1}^n Y_i X_i, A \right\rangle + \lambda\|A\|_1 \right\}. \tag{2.7}$$

Here, $\lambda > 0$ is a regularization parameter, $\| A \|^2_{L_2(\Pi)} = \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left(\langle A, X_i\rangle^2\right)$ and $\Pi = \frac{1}{n}\sum_{i=1}^n \Pi_i$ where Π_i are the distributions of X_i.

If the following assumption of restricted isometry in expectation is satisfied

$$\text{for some } \mu > 0 \qquad \| A \|^2_{L_2(\Pi)} = \mu^{-2} \| A \|_2^2,$$

then (2.7) has a particularly simple form:

$$\hat{A}_{ML} = \operatorname*{arg\,min}_{A\in\mathbb{R}^{m_1\times m_2}} \left\{ \| A - \mathbf{X} \|_2^2 + \lambda\mu^2\|A\|_1 \right\} \tag{2.8}$$

where

$$\mathbf{X} = \frac{\mu^2}{n}\sum_{i=1}^n Y_i X_i. \tag{2.9}$$

In the first part of the present paper we study the following estimator

$$\hat{A}_{\lambda,\mu} = \operatorname*{arg\,min}_{A\in\mathbb{R}^{m_1\times m_2}} \left\{ \|A - \mathbf{X}\|_2 + \lambda\|A\|_1 \right\}. \tag{2.10}$$

In order to simplify our notations we will write $\hat{A} = \hat{A}_{\lambda,\mu}$. Note that the first part of our estimator coincides with the square root of the data-depending term in (2.8). This is similar to the principle used to define the square-root lasso for the usual vector regression model, see [1]. Theorem 2 gives an oracle bound on the prediction error of $\hat{A}$. This bound is obtained for an arbitrary μ and does not rely on the knowledge of the distributions of X_i. We apply Theorem 2 to matrix completion, taking $\mu^2 = m_1 m_2$.

In the second part of the present paper, dedicated to matrix regression problem, we consider a new estimator inspired by the same idea,

namely

$$\hat{A} = \underset{A\in\mathbb{R}^{m_1\times m_2}}{\arg\min} \left\{\|U - V\,A\|_2 + \lambda\|VA\|_1\right\}. \tag{2.11}$$

Note that in (2.11) we penalized by the nuclear norm of VA, rather the by the nuclear norm of A as in (2.7).

3. General oracle inequalities

In this section, in Theorem 2, we provide a general oracle inequality for the prediction error of our estimator. The proof of Theorem 2 is based on the ideas of the proof of Theorem 1 in [13]. However, as the statistical structure of our estimator is different from that of the estimator proposed in [13], the proof requires several modifications and additional information on the behavior of the estimator. This information is given in Lemmas 1 and 3. In particular, Lemma 1 provides a bound on the rank of our estimator in the general setting of the trace regression model.

Lemma 1.

$$\mathrm{rank}(\hat{A}) \le 1/\lambda^2.$$

Proof. That $\hat{A}$ is the minimum of (2.10) implies that $0 \in \partial F(\hat{A})$. We will use the fact that the subdifferential of the convex function $A \to \|A\|_1$ is the following set of matrices (cf. [20])

$$\partial\|A\|_1 = \left\{ \sum_{j=1}^{\mathrm{rank}(A)} u_j(A)v_j^T(A) + P_{S_1^\perp(A)} W P_{S_2^\perp(A)} \ : \ \|W\|_\infty \le 1 \right\} \tag{3.1}$$

$u_j(A)$ and $v_j(A)$ are 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 $\{v_j(A)\}$. For $\hat{A} \neq \mathbf{X}$, this implies that there exists a matrix W such that $\|W\|_\infty \le 1$ and

$$\frac{\hat{A} - \mathbf{X}}{\|\hat{A} - \mathbf{X}\|_2} = -\lambda \sum_{j=1}^{\mathrm{rank}(\hat{A})} u_j(\hat{A})v_j^T(\hat{A}) - \lambda P_{S_1^\perp(\hat{A})} W P_{S_2^\perp(\hat{A})}. \tag{3.2}$$

Calculating the $\|\|_2^2$ norm of both sides of (3.2) we get that $1 \ge \lambda^2\mathrm{rank}(\hat{A})$. When $\hat{A} = \mathbf{X}$, instead of the differential of $\|\hat{A} - \mathbf{X}\|_2$ we use its subdifferential. In (3.2) the term $\dfrac{\hat{A} - \mathbf{X}}{\|\hat{A} - \mathbf{X}\|_2}$ is replaced by a matrix $\tilde{W}$ such that $\|\tilde{W}\|_2 \le 1$ and we get again $1 \ge \lambda^2\mathrm{rank}(\hat{A})$. □

Theorem 2. *Suppose that* $\dfrac{\rho}{\sqrt{2\mathrm{rank}(A_0)}} \ge \lambda \ge 3\Delta$ *for some* $\rho < 1$, *then*

$$\|\hat{A} - A_0\|_2^2 \le \inf_{\sqrt{2\mathrm{rank}(A)}\le\rho/\lambda} \left\{ (1-\rho)^{-1}\|A - A_0\|_2^2 + \left(\frac{2\lambda\mu^2}{1-\rho}\right)^2 \|\mathbf{M}\|_2^2\,\mathrm{rank}A \right\}$$

Proof. We need the following auxiliary result which is proven in the Appendix

Lemma 3. *Suppose that* $\dfrac{\rho}{\sqrt{\operatorname{rank}(A_0)}} \geq \lambda \geq 3\Delta$ *for some* $\rho < 1$, *then*

$$\|\hat{A} - \mathbf{X}\|_2 \geq \left(\frac{3 - \sqrt{1+\rho^2}}{3 + \sqrt{1+\rho^2}}\right) \|A_0 - \mathbf{X}\|_2 \tag{3.3}$$

If $\hat{A} = \mathbf{X}$, then (3.3) implies that $A_0 = \mathbf{X}$ and we get $\|\hat{A} - A_0\|_2 = 0$. If $\hat{A} \neq \mathbf{X}$, a necessary condition of extremum in (2.10) implies that there exists $\hat{V} \in \partial\|\hat{A}\|_1$ such that for any $A \in \mathbb{R}^{m_1 \times m_2}$

$$\frac{2\langle \hat{A} - \mathbf{X}, \hat{A} - A\rangle}{2\|\hat{A} - \mathbf{X}\|_2} + \lambda\langle \hat{V}, \hat{A} - A\rangle \leq 0$$

which yields

$$2\langle \hat{A} - A_0, \hat{A} - A\rangle - 2\langle \mathbf{X} - A_0, \hat{A} - A\rangle + 2\lambda\|\hat{A} - \mathbf{X}\|_2\langle \hat{V}, \hat{A} - A\rangle \leq 0. \tag{3.4}$$

By (3.1) we have the following representation for an arbitrary $V \in \partial\|A\|_1$

$$V = \sum_{j=1}^{r} u_j v_j^T + P_{S_1^\perp(A)} W P_{S_2^\perp(A)}, \tag{3.5}$$

for simplicity we write u_j and v_j instead of $u_j(A)$ and $v_j(A)$.

By the monotonicity of subdifferentials of convex functions we have that $\langle \hat{V} - V, \hat{A} - A\rangle \geq 0$. Then (3.4) and $2\langle \hat{A} - A_0, \hat{A} - A\rangle = \|\hat{A} - A_0\|_2^2 + \|\hat{A} - A\|_2^2 - \|A - A_0\|_2^2$ imply

(3.6)

$$\begin{aligned}
&\|\hat{A} - A_0\|_2^2 + \|\hat{A} - A\|_2^2 + 2\lambda\|\hat{A} - \mathbf{X}\|_2 \left\langle P_{S_1^\perp(A)} W P_{S_2^\perp(A)}, \hat{A} - A\right\rangle \\
&\quad \leq \|A - A_0\|_2^2 + 2\langle \mathbf{X} - A_0, \hat{A} - A\rangle - 2\lambda\|\hat{A} - \mathbf{X}\|_2 \left\langle \sum_{j=1}^{r} u_j v_j^T, \hat{A} - A\right\rangle.
\end{aligned}$$

From the trace duality we get that there exists W with $\|W\|_\infty \leq 1$ such that

$$\begin{aligned}
\left\langle P_{S_1^\perp(A)} W P_{S_2^\perp(A)}, \hat{A} - A\right\rangle &= \left\langle W, P_{S_1^\perp(A)}\left(\hat{A} - A\right) P_{S_2^\perp(A)}\right\rangle \\
&= \left\| P_{S_1^\perp(A)}\left(\hat{A} - A\right) P_{S_2^\perp(A)}\right\|_1.
\end{aligned} \tag{3.7}$$

For a $m_1 \times m_2$ matrix B let $\mathbf{P}_A(B) = B - P_{S_1^\perp(A)} B P_{S_2^\perp(A)}$. Since

$$\mathbf{P}_A(B) = P_{S_1^\perp(A)} B P_{S_2(A)} + P_{S_1(A)} B$$

and $\operatorname{rank}(P_{S_i(A)} B) \leq \operatorname{rank}(A)$ we have that $\operatorname{rank}(\mathbf{P}_A(B)) \leq 2\operatorname{rank}(A)$.

Using the trace duality and triangle inequality we get

$$\begin{aligned}\langle \mathbf{X}-A_0, \hat{A}-A\rangle &\le \|\mathbf{X}-A_0\|_\infty \|\hat{A}-A\|_1 \\ &\le \|\mathbf{X}-A_0\|_\infty \left\|\mathbf{P}_A\left(\hat{A}-A\right)\right\|_1 \\ &\quad + \|\mathbf{X}-A_0\|_\infty \left\|P_{S_1^\perp(A)}\left(\hat{A}-A\right)P_{S_2^\perp(A)}\right\|_1 .\end{aligned} \tag{3.8}$$

Note that $\left\|\sum_{j=1}^r u_j v_j^T\right\|_\infty = 1$. Then, the trace duality implies

$$\left\langle \sum_{j=1}^r u_j v_j^T, \hat{A}-A\right\rangle = \left\langle \sum_{j=1}^r u_j v_j^T, \mathbf{P}_A\left(\hat{A}-A\right)\right\rangle \le \left\|\mathbf{P}_A\left(\hat{A}-A\right)\right\|_1 . \tag{3.9}$$

Putting (3.7), (3.8), (3.9) into (3.6) we compute

$$\begin{aligned}\|\hat{A}-A_0\|_2^2 + \|\hat{A}-A\|_2^2 + 2\lambda\|\hat{A}-\mathbf{X}\|_2 \left\|P_{S_1^\perp(A)}\,\hat{A}\,P_{S_2^\perp(A)}\right\|_1 \\ \le \|A-A_0\|_2^2 + 2\|\mathbf{X}-A_0\|_\infty \left\|\mathbf{P}_A\left(\hat{A}-A\right)\right\|_1 \\ + 2\|\mathbf{X}-A_0\|_\infty \left\|P_{S_1^\perp(A)}\,\hat{A}\,P_{S_2^\perp(A)}\right\|_1 \\ + 2\lambda\|\hat{A}-\mathbf{X}\|_2 \left\|\mathbf{P}_A\left(\hat{A}-A\right)\right\|_1 .\end{aligned} \tag{3.10}$$

Using (3.3), from (3.10), we derive

$$\begin{aligned}\|\hat{A}-A_0\|_2^2 + \|\hat{A}-A\|_2^2 + 2\lambda\frac{3-\sqrt{1+\rho^2}}{3+\sqrt{1+\rho^2}}\|A_0-\mathbf{X}\|_2 \left\|P_{S_1^\perp(A)}\,\hat{A}\,P_{S_2^\perp(A)}\right\|_1 \\ \le \|A-A_0\|_2^2 + 2\|\mathbf{X}-A_0\|_\infty \left\|\mathbf{P}_A\left(\hat{A}-A\right)\right\|_1 \\ + 2\|\mathbf{X}-A_0\|_\infty \left\|P_{S_1^\perp(A)}\,\hat{A}\,P_{S_2^\perp(A)}\right\|_1 \\ + 2\lambda\|\hat{A}-\mathbf{X}\|_2 \left\|\mathbf{P}_A\left(\hat{A}-A\right)\right\|_1 .\end{aligned}$$

From the definition of λ we get

$$\begin{aligned}\|\hat{A}-A_0\|_2^2 + \|\hat{A}-A\|_2^2 + 6\frac{3-\sqrt{1+\rho^2}}{3+\sqrt{1+\rho^2}}\|A_0-\mathbf{X}\|_\infty \|P_{S_1^\perp(A)}\,\hat{A}\,P_{S_2^\perp(A)}\|_1 \\ \le \|A-A_0\|_2^2 + 2\|\mathbf{X}-A_0\|_\infty \|\mathbf{P}_A\left(\hat{A}-A\right)\|_1 \\ + 2\|\mathbf{X}-A_0\|_\infty \|P_{S_1^\perp(A)}\,\hat{A}\,P_{S_2^\perp(A)}\|_1 \\ + 2\lambda\|\hat{A}-\mathbf{X}\|_2 \|\mathbf{P}_A\left(\hat{A}-A\right)\|_1 .\end{aligned} \tag{3.11}$$

Note that $6\dfrac{3-\sqrt{1+\rho^2}}{3+\sqrt{1+\rho^2}} \geq 2$ for any $\rho < 1$. Thus, (3.11) yields

(3.12)
$$\|\hat{A}-A_0\|_2^2 + \|\hat{A}-A\|_2^2 \leq \|A-A_0\|_2^2 + 2\|\mathbf{X}-A_0\|_\infty \|\mathbf{P}_A\left(\hat{A}-A\right)\|_1 + 2\lambda\|\hat{A}-\mathbf{X}\|_2\|\mathbf{P}_A\left(\hat{A}-A\right)\|_1.$$

Now, using the triangle inequality and the fact that

$$\left\|\mathbf{P}_A\left(\hat{A}-A\right)\right\|_1 \leq \sqrt{2\mathrm{rank}(A)}\|\hat{A}-A\|_2$$

from (3.12) we get

(3.13)
$$\|\hat{A}-A_0\|_2^2 + \|\hat{A}-A\|_2^2 \leq \|A-A_0\|_2^2 + 2\Big(\|\mathbf{X}-A_0\|_\infty + \lambda\|\mathbf{X}-A_0\|_2\Big)\sqrt{2\mathrm{rank}(A)}\|\hat{A}-A\|_2 + 2\lambda\|\hat{A}-A_0\|_2\sqrt{2\mathrm{rank}(A)}\|\hat{A}-A\|_2.$$

From the definition of λ we get that $\|\mathbf{X}-A_0\|_\infty \leq \lambda\|\mathbf{X}-A_0\|_2/3$. For A such that $\lambda\sqrt{2\mathrm{rank}(A)} \leq \rho$, (3.13) implies

$$\|\hat{A}-A_0\|_2^2 + \|\hat{A}-A\|_2^2 \leq \|A-A_0\|_2^2 + \frac{8}{3}\lambda\|\mathbf{X}-A_0\|_2\sqrt{2\mathrm{rank}(A)}\|\hat{A}-A\|_2 + 2\rho\|\hat{A}-A_0\|_2\|\hat{A}-A\|_2.$$

Using $2ab \leq a^2+b^2$ twice we finally compute

$$(1-\rho)\|\hat{A}-A_0\|_2^2 + \|\hat{A}-A\|_2^2 \leq \|A-A_0\|_2^2 + \rho\|\hat{A}-A\|_2^2 + \frac{8}{3}\lambda\|\mathbf{X}-A_0\|_2\sqrt{2\mathrm{rank}(A)}\|\hat{A}-A\|_2$$

and

$$(1-\rho)\|\hat{A}-A_0\|_2^2 \leq \|A-A_0\|_2^2 + \frac{4}{1-\rho}\lambda^2\|\mathbf{X}-A_0\|_2^2\mathrm{rank}(A)$$

which implies the statement of Theorem 2. □

4. Matrix Completion

In this section we apply the general oracle inequality of Theorem 2 for the model of USR matrix completion. Assume that the design matrices X_i are i.i.d uniformly distributed on the set $\mathcal{X}$ defined in (2.3). This implies that

$$\frac{1}{n}\sum_{i=1}^n \mathbb{E}\left(\langle A, X_i\rangle^2\right) = (m_1m_2)^{-1}\|A\|_2^2,$$

for all matrices $A \in \mathbb{R}^{m_1 \times m_2}$ and we take $\mu^2 = m_1 m_2$.

We will consider the case of sub-Gaussian noise and matrices with uniformly bounded entries. We suppose that the noise variables ξ_i are such that

$$\mathbb{E}(\xi_i) = 0,\ \mathbb{E}(\xi_i^2) = 1 \tag{4.1}$$

and there exists a constant K such that

$$\mathbb{E}\left[\exp(t\xi_i)\right] \leq \exp\left(t^2/2K\right) \tag{4.2}$$

for all $t > 0$. Normal $N(0,1)$ random variables are sub-Gaussian with $K = 1$ and (4.2) implies that ξ_i has Gaussian type tails:

$$\mathbb{P}\left\{|\xi_i| > t\right\} \leq 2\exp\left\{-t^2/2K\right\}.$$

Note that condition $\mathbb{E}\xi_i^2 = 1$ implies that $K \leq 1$.

Let a denote a constant such that

$$\|A_0\|_{\sup} \leq a. \tag{4.3}$$

In order to specify the value of the regularization parameter λ, we need to estimate Δ (defined in (2.6)) with high probability. In what follows we will denote by c a numerical constant whose value can vary from one expression to the other and is independent from n, m_1, m_2. Set $m = m_1 + m_2$, $m_1 \wedge m_2 = \min(m_1, m_2)$ and $m_1 \vee m_2 = \max(m_1, m_2)$. The following bound is a consequence of Lemmas 2 and 3 in [13],

Lemma 4. *For $n > 8(m_1 \wedge m_2)\log^2 m$, with probability at least $1 - 3/m$, one has*

$$\Delta_\infty \leq (c_*\sigma + 2a)\sqrt{\frac{2\log(m)}{(m_1 \wedge m_2)n}} \tag{4.4}$$

where c_* is a numerical constant which depends only on K.
If ξ_i are $N(0,1)$, then we can take $c_* = 6.5$.

Proof. The bound (4.4) is stated in Lemmas 2 and 3 in [13]. A closer inspection of the proof of Proposition 2 in [12] gives an estimation on c_* in the case of Gaussian noise. For more details see the appendix. □

The following Lemma, proven in the appendix, provides bounds on $\|\mathbf{M}\|_2$.

Lemma 5. *Suppose that $4\,n \leq m_1 m_2$. Then, for $\mathbf{M}$ defined in (2.5), there exists absolute constants (c_1, c_2) such that, with probability at least $1 - 2/m_1 m_2 - c_1\exp\{-c_2 n\}$, one has*

(i)

$$2\left(\frac{\|A_0\|_2^2}{n m_1 m_2} + \frac{\sigma^2}{n}\right) \geq \|\mathbf{M}\|_2^2 \geq \frac{\sigma^2}{2\,n};$$

(ii)
$$\left\| \frac{1}{n} \sum_{i=1}^{n} Y_i X_i \right\|_2^2 \geq \frac{\|A_0\|_2^2}{n\, m_1 m_2} \geq \frac{4\, \|A_0\|_2^2}{(m_1 m_2)^2};$$

(iii)
$$\|\mathbf{M}\|_2 \geq \frac{1}{2} \left\| \frac{1}{n} \sum_{i=1}^{n} Y_i X_i \right\|_2 .$$

Recall that the condition on λ in Theorem 2 is that $\lambda \geq 3\Delta$. Using Lemma 4 and the lower bounds on $\|M\|_2$ given by Lemma 5 we can choose

$$\lambda = 2c_* \sqrt{\frac{\log m}{m_1 \wedge m_2}} + 4a \sqrt{\frac{2\, n \log m}{m_1 \wedge m_2}} \frac{1}{\left\| \sum_{i=1}^{n} Y_i X_i \right\|_2}. \tag{4.5}$$

With this choice of λ, the assumption of Theorem 2 that $\dfrac{\rho}{\sqrt{\operatorname{rank}(A_0)}} \geq \lambda$ takes the form

$$\frac{\rho}{\sqrt{\operatorname{rank}(A_0)}} \geq 2c_* \sqrt{\frac{\log m}{m_1 \wedge m_2}} + 4a \sqrt{\frac{2\, n \log m}{m_1 \wedge m_2}} \frac{1}{\left\| \sum_{i=1}^{n} Y_i X_i \right\|_2}. \tag{4.6}$$

Using (ii) of Lemma 5 we get that (4.6) is satisfied with a hight probability if

$$\frac{\rho}{\sqrt{\operatorname{rank}(A_0)}} \geq 2\, c_* \sqrt{\frac{\log m}{m_1 \wedge m_2}} + \frac{4\, a \sqrt{m_1 m_2}}{\|A_0\|_2} \sqrt{\frac{2 \log m}{m_1 \wedge m_2}}. \tag{4.7}$$

Note that as m_1 and m_2 are large, the first term in the rhs of (4.7) is small. Thus (4.7) is essentially equivalent to

$$\rho \geq 4 \sqrt{\frac{2 \log m}{(m_1 \wedge m_2)}} \sqrt{\operatorname{rank}(A_0)}\, \alpha_{sp} \tag{4.8}$$

where $\alpha_{sp} = \dfrac{\sqrt{m_1 m_2}\, \|A_0\|_{\sup}}{\|A_0\|_2}$ is the *spikiness ratio* of A_0. The notion of "spikiness" was introduced by Negahban and Wainwright in [15]. We have that $1 \leq \alpha_{sp} \leq \sqrt{m_1 m_2}$ and it is large for "spiky" matrices, i.e. matrices where some "large" coefficients emerge as spikes among very "small" coefficients. For instance, $\alpha_{sp} = 1$ if all the entries of A_0 are equal to some constant and $\alpha_{sp} = \sqrt{m_1 m_2}$ if A_0 has only one non-zero entry.

Condition (4.8) is a kind of trade-off between "spikiness" and rank. If α_{sp} is bounded by a constant, then, up to a logarithmic factor, $\operatorname{rank}(A_0)$ can be of the order $m_1 \wedge m_2$, which is its maximal possible value. If our matrix is "spiky", then we need low rank. To give some intuition let

us consider the case of square matrices. Typically, matrices with both high spikiness ratio and high rank look almost diagonal. Thus, under uniform sampling and if $n \ll m_1 m_2$, with high probability we do not observe diagonal (i.e. non-zero) elements.

Theorem 6. *Let the set of conditions (4.1) - (4.3) be satisfied and λ be as in (4.5). Assume that $8(m_1 \wedge m_2)\log^2 m < n \leq \dfrac{m_1 m_2}{4}$ and that (4.7) holds for some $\rho < 1$. Then, there exists absolute constants (c_1, c_2) such that, with probability at least $1 - 4/m - c_1 \exp\{-c_2 n\}$*

$$\frac{1}{m_1 m_2}\|\hat{A} - A_0\|_2^2 \leq C_* \frac{(m_1 \vee m_2)}{n}\mathrm{rank}(A_0)\log m \tag{4.9}$$

where $C_* = \dfrac{16\,(2c_*\sigma^2 + (18 + 2c_*)a^2)}{(1-\rho)^2}$.

Proof. This is a consequence of Theorem 2 for $A = A_0$. From (4.5) we get

$$\|\hat{A} - A_0\|_2^2 \leq \frac{8(m_1 m_2)^2}{(1-\rho)^2}\left(c_*\sqrt{\frac{4\log m}{m_1 \wedge m_2}} + 2a\sqrt{\frac{2\,n\log m}{m_1 \wedge m_2}}\frac{1}{\left\|\sum_{i=1}^n Y_i X_i\right\|_2}\right)^2 \times \|\mathbf{M}\|_2^2\,\mathrm{rank}(A_0). \tag{4.10}$$

Using triangle inequality and (ii) of Lemma 5 we compute

$$\|\mathbf{M}\|_2 \leq \left\|\frac{1}{n}\sum_{i=1}^n Y_i X_i\right\|_2 + \frac{1}{m_1 m_2}\|A_0\|_2 \leq \frac{3}{2}\left\|\frac{1}{n}\sum_{i=1}^n Y_i X_i\right\|_2. \tag{4.11}$$

Using (i) of Lemma 5 and (4.11), from (4.10) we get

$$\|\hat{A} - A_0\|_2^2 \leq \frac{16\log(m)(m_1 m_2)^2}{(1-\rho)^2(m_1 \wedge m_2)}\left(2\,c_*\left(\frac{\|A_0\|_2^2}{n m_1 m_2} + \frac{\sigma^2}{n}\right) + \frac{18a^2}{n}\right)\mathrm{rank}(A_0).$$

We then use $\|A_0\|_2^2 \leq a^2 m_1 m_2$ to obtain

$$\frac{\|\hat{A} - A_0\|_2^2}{m_1 m_2} \leq \frac{16\log(m)(m_1 \vee m_2)}{(1-\rho)^2 n}\left(2c_*\sigma^2 + (18 + 2c_*)a^2\right)\mathrm{rank}(A_0).$$

This completes the proof of Theorem 6. □

Theorem 6 guarantees that the normalized Frobenius error $\dfrac{\|\hat{A} - A_0\|_2}{\sqrt{m_1 m_2}}$ of the estimator $\hat{A}$ is small whenever $n > C(m_1 \vee m_2)\log(m)\mathrm{rank}(A_0)$ with a constant C large enough. This quantifies the sample size n necessary for successful matrix completion from noisy data with unknown variance of the noise. This sampling size is the same as in the case of known variance of the noise.

In order to compare our bounds to those obtained in past works on noisy matrix completion, we will start with the paper of Keshavan et al [9]. Under a sampling scheme different from ours (sampling without replacement) and sub-Gaussian errors, the estimator proposed in [9] satisfies, with high probability, the following bound

$$\frac{1}{m_1 m_2}\|\hat{A} - A_0\|_2^2 \lesssim k^4\sqrt{\alpha}\frac{(m_1 \vee m_2)}{n}\mathrm{rank}(A_0)\log n. \tag{4.12}$$

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 (4.12) and (4.9), 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 [15] propose an estimator which, in the case of USR matrix completion and sub-exponential noise, satisfies

$$\frac{1}{m_1 m_2}\|\hat{A} - A_0\|_2^2 \lesssim \alpha_{sp}\frac{m}{n}\mathrm{rank}(A_0)\log m. \tag{4.13}$$

Here α_{sp} is the spikiness ratio of A_0. For α_{sp} bounded by a constant, (4.13) gives the same bound as Theorem 6. The construction of $\hat{A}$ in [15] requires a prior information on the spikiness ratio of A_0 and on σ. This is not the case for our estimator, which is completely data-driven.

The estimator proposed by Koltchinskii et al in [13] achieves the same bound as ours. In addition to prior information on $\|A_0\|_{\sup}$, their method also requires prior information on σ. In the case of Gaussian errors, this rate of convergence is optimal (cf. Theorem 6 of [13]) for the class of matrices $\mathcal{A}(r,a)$ defined as follows: for given r and a, for any $A_0 \in \mathcal{A}(r,a)$ the rank of A_0 is supposed not to be larger than r and all the entries of A_0 are supposed to be bounded in absolute value by a.

5. Matrix Regression

In this section we apply our method to matrix regression. Recall that the matrix regression model is given by

$$U_i = V_i A_0 + E_i \qquad i = 1, \dots, l, \tag{5.1}$$

where U_i are $1 \times m_2$ vectors of response variables; V_i are $1 \times m_1$ vectors of predictors; A_0 is an unknown $m_1 \times m_2$ matrix of regression coefficients; E_i are random $1 \times m_2$ noise vectors with independent entries E_{ij}. We suppose that E_{ij} has mean zero and unknown standard deviation σ. Set $V = \left(V_1^T, \dots, V_l^T\right)^T$, $U = \left(U_1^T, \dots, U_l^T\right)^T$ and $E = \left(E_1^T, \dots, E_l^T\right)^T$.

We define the following estimator of A_0:

$$\hat{A} = \underset{A \in \mathbb{R}^{m_1 \times m_2}}{\arg\min} \left\{\|U - V A\|_2 + \lambda\|VA\|_1\right\},$$

where $\lambda > 0$ is a regularization parameter.

Let $\mathcal{P}_V$ denote the orthogonal projector on the linear span of the columns of matrix V and let $\mathcal{P}_V^\perp = 1 - \mathcal{P}_V$. Note that

$$V^T \mathcal{P}_V^\perp = 0$$

and $A^T V = 0$ (which means that the columns of A are orthogonal to the columns of V) implies $\mathcal{P}_V A = 0$.

The following lemma is the counterpart of Lemma 1 in the present setting.

Lemma 7.

$$\operatorname{rank}(V\hat{A}) \leq 1/\lambda^2.$$

Proof. That $\hat{A}$ is the minimum of (2.11) implies that $0 \in \partial G(\hat{A})$ where

$$G = \|U - V\,A\|_2 + \lambda \|VA\|_1.$$

Note that the subdifferential of the convex function $A \to \|VA\|_1$ is the following set of matrices

$$\partial \|VA\|_1 = V^T \left\{ \sum_{j=1}^{\operatorname{rank}(VA)} u_j(VA) v_j^T(VA) + P_{S_1^\perp(VA)} W P_{S_2^\perp(VA)} \ : \ \|W\|_\infty \leq 1 \right\}$$

where $S_1(VA)$ is the linear span of $\{u_j(VA)\}$ and $S_2(VA)$ is the linear span of $\{v_j(VA)\}$.

If $\hat{A}$ is such that $V\hat{A} \neq U$, we obtain that there exists a matrix W such that $\|W\|_\infty \leq 1$ and

$$V^T \frac{V\hat{A} - U}{\|V\hat{A} - U\|_2} = -\lambda V^T \left\{ \sum_{j=1}^{\operatorname{rank}(VA)} u_j(VA) v_j^T(VA) + P_{S_1^\perp(VA)} W P_{S_2^\perp(VA)} \right\}$$

which implies

(5.2)

$$V^T \mathcal{P}_V \frac{V\hat{A} - U}{\|V\hat{A} - U\|_2} = -\lambda V^T \mathcal{P}_V \left\{ \sum_{j=1}^{\operatorname{rank}(VA)} u_j(VA) v_j^T(VA) + P_{S_1^\perp(VA)} W P_{S_2^\perp(VA)} \right\}.$$

Recall that $V^T \mathcal{P}_V B = 0$ implies that $\mathcal{P}_V^2 B = \mathcal{P}_V B = 0$ and we get from (5.2)

(5.3)

$$\mathcal{P}_V \frac{V\hat{A} - U}{\|V\hat{A} - U\|_2} = -\lambda \left\{ \sum_{j=1}^{\operatorname{rank}(VA)} \mathcal{P}_V u_j(VA) v_j^T(VA) + \mathcal{P}_V \left[P_{S_1^\perp(VA)} W P_{S_2^\perp(VA)} \right] \right\}.$$

By the definition of the singular vectors we have that $VA\,(v_j(VA)) = \sigma_j(VA) u_j(VA)$ and we compute

$$\mathcal{P}_V VA\,(v_j(VA)) = \sigma_j(VA) \mathcal{P}_V u_j(VA)$$

using $\mathcal{P}_V VA\,(v_j(VA)) = VA\,(v_j(VA)) = \sigma_j(VA) u_j(VA)$ and $\sigma_j \neq 0$ we get

$$\mathcal{P}_V u_j(VA) = u_j(VA) \tag{5.4}$$

and we obtain from (5.3)

(5.5)
$$\mathcal{P}_V \frac{V\hat{A}-U}{\|V\hat{A}-U\|_2} = -\lambda \left\{ \sum_{j=1}^{\mathrm{rank}(VA)} u_j(VA) v_j^T(VA) + \mathcal{P}_V \left[P_{S_1^\perp(VA)} W P_{S_2^\perp(VA)} \right] \right\}.$$

Note that for any w such that $\langle w, u_j(VA)\rangle = 0$ (5.4) implies that

$$\langle \mathcal{P}_V w, u_j(VA)\rangle = \langle w, u_j(VA)\rangle = 0. \tag{5.6}$$

By the definition, $P_{S_1^\perp(VA)}$ projects on the subspace orthogonal to the linear span of $\{u_j(VA)\}$. Thus, (5.6) implies that $\mathcal{P}_V P_{S_1^\perp(VA)}$ also projects on the subspace orthogonal to the linear span of $\{u_j(VA)\}$.

Calculating the $\|\|_2^2$ norm of both sides of (5.5) we get that $1 \geq \lambda^2 \mathrm{rank}(V\hat{A})$. When $V\hat{A} = U$, instead of the differential of $\|U - VA\|_2$ we use its subdiffential. □

Minor modifications in the proof of Theorem 2 yield the following result. We set $\Delta' = \dfrac{\|\mathcal{P}_V(E)\|_\infty}{\|E\|_2}$.

Theorem 8. *Suppose that* $\dfrac{\rho}{\sqrt{2\mathrm{rank}(VA_0)}} \geq \lambda \geq 3\Delta'$ *for some* $\rho < 1$, *then*

$$\left\| V\left(\hat{A} - A_0\right)\right\|_2^2 \leq \inf_{\sqrt{2\mathrm{rank}(VA)} \leq \rho/\lambda} \left\{ \frac{\|V(A-A_0)\|_2^2}{1-\rho} + \left(\frac{2\lambda}{1-\rho}\right)^2 \|E\|_2^2 \,\mathrm{rank}(VA) \right\}$$

Proof. The proof follows the lines of the proof of Theorem 2 and it is given in the appendix. □

To get the oracle inequality in a closed form it remains to specify the value of regularization parameter λ such that $\lambda \geq 3\Delta'$. This requires some assumptions on the distribution of the noise $(E_{ij})_{i,j}$. We will consider the case of Gaussian errors. Suppose that $E_{ij} = \sigma\xi_{ij}$ where ξ_{ij} are normal $N(0,1)$ random variables. In order to estimate $\|\mathcal{P}_V E\|_\infty$ we will use the following result proven in [2].

Lemma 9 ([2], Lemma 3)**.** *Let* $r = \mathrm{rank}(V)$ *and assume that* E_{ij} *are independent* $N(0,\sigma^2)$ *random variables. Then*

$$\mathbb{E}(\|\mathcal{P}_V E\|_\infty) \leq \sigma(\sqrt{m_2} + \sqrt{r})$$

and

$$\mathbb{P}\left\{\|\mathcal{P}_V E\|_\infty \geq \mathbb{E}(\|\mathcal{P}_V E\|_\infty) + \sigma t\right\} \leq \exp\left\{-t^2/2\right\}.$$

We use Bernstein's inequality to get a bound on $\|E\|_2$. Let $\alpha < 1$. With probability at least $1 - 2\exp\{-c\alpha^2 l m_2\}$, one has

$$(1+\alpha)\sigma\sqrt{lm_2} \geq \|E\|_2 \geq (1-\alpha)\sigma\sqrt{lm_2}. \tag{5.7}$$

Let $\beta > 0$ and take $t = \beta\left(\sqrt{m_2} + \sqrt{r}\right)$ in Lemma 9. Then, using (5.7) we can take

$$\lambda = \frac{(1+\beta)\left(\sqrt{m_2} + \sqrt{r}\right)}{(1-\alpha)\sqrt{lm_2}}. \tag{5.8}$$

Put $\gamma = \dfrac{1+\beta}{1-\alpha} > 1$. Thus, condition $\dfrac{\rho}{\sqrt{2\mathrm{rank}(VA_0)}} \geq \lambda$ gives

$$\mathrm{rank}(VA_0) \leq \frac{\rho^2 lm_2}{2\gamma^2\left(\sqrt{m_2} + \sqrt{r}\right)^2} \tag{5.9}$$

and we get the following result.

Theorem 10. *Assume that ξ_{ij} are independent $N(0,1)$. Pick λ as in (5.8). Assume (5.9) be satisfied for some $\rho < 1$, $\alpha < 1$ and $\beta > 0$. Then, with probability at least $1 - 2\exp\{-c(m_2 + r)\}$ we have that*

$$\left\|V\left(\hat{A} - A_0\right)\right\|_2^2 \lesssim \sigma^2(m_2 + r)\,\mathrm{rank}(VA_0).$$

The symbol $\lesssim$ means that inequality holds up to a multiplicative numerical constant and c denotes a numerical constant that depends on α and β.

Proof. This is a consequence of Theorem 8. □

Let us now compare condition (5.9) with the conditions obtained in [2, 7]. The method proposed in [2] requires $m_2(l-r)$ to be large, which holds whenever $l \gg r$ or $l - r \geq 1$ and m_2 is large. This condition excludes an interesting case $l = r \ll m_2$. On the other hand (5.9) is satisfied for $l = r \ll m_2$ if

$$\mathrm{rank}(A_0) \lesssim l$$

where we used $\mathrm{rank}(VA_0) \leq r \wedge \mathrm{rank}(A_0)$.

The method of [7] requires the following condition to be satisfied

$$\mathrm{rank}(A_0) \leq \frac{C_1(lm_2 - 1)}{C_2\left(\sqrt{m_2} + \sqrt{r}\right)^2} \tag{5.10}$$

with some constants $C_1 < 1$ and $C_2 > 1$. As $\mathrm{rank}(VA_0) \leq \mathrm{rank}(A_0)$, condition (5.9) is weaker then (5.10). Note also, that, to the opposite of [7], our results are valid for all A_0 provided that

$$r \leq \frac{\rho^2 lm_2}{2\gamma^2\left(\sqrt{m_2} + \sqrt{r}\right)^2}.$$

For large $m_2 \gg l$, this condition roughly mean that $l > cr$ for some constant c.

Appendix A. Proof of Lemma 3

If $A_0 = \mathbf{X}$, then we have trivially $\|\hat{A} - \mathbf{X}\|_2 \geq 0$.
If $A_0 \neq \mathbf{X}$, by the convexity of the function $A \to \|A - \mathbf{X}\|_2$, we have

(A.1)

$$\begin{aligned}\|\hat{A} - \mathbf{X}\|_2 - \|A_0 - \mathbf{X}\|_2 &\geq \frac{\langle A_0 - \mathbf{X}, \hat{A} - A_0\rangle}{\|A_0 - \mathbf{X}\|_2} \\ &\geq -\frac{\|A_0 - \mathbf{X}\|_\infty}{\|A_0 - \mathbf{X}\|_2}\|\hat{A} - A_0\|_1 \\ &\geq -\frac{\|A_0 - \mathbf{X}\|_\infty}{\|A_0 - \mathbf{X}\|_2}\sqrt{\operatorname{rank}(\hat{A}) + \operatorname{rank}(A_0)}\|\hat{A} - A_0\|_2.\end{aligned}$$

Using Lemma 1, the bound $\dfrac{\rho}{\sqrt{\operatorname{rank}(A_0)}} \geq \lambda$ and the triangle inequality, from (A.1) we get

(A.2)

$$\|\hat{A} - \mathbf{X}\|_2 - \|A_0 - \mathbf{X}\|_2 \geq - \frac{\sqrt{1+\rho^2}}{\lambda}\frac{\|A_0 - \mathbf{X}\|_\infty}{\|A_0 - \mathbf{X}\|_2}\left(\|\hat{A} - \mathbf{X}\|_2 + \|A_0 - \mathbf{X}\|_2\right).$$

Note that $\dfrac{\|A_0 - \mathbf{X}\|_\infty}{\lambda\|A_0 - \mathbf{X}\|_2} \leq 1/3$ which finally leads to

$$\left(1 + \frac{\sqrt{1+\rho^2}}{3}\right)\|\hat{A} - \mathbf{X}\|_2 \geq \left(1 - \frac{\sqrt{1+\rho^2}}{3}\right)\|A_0 - \mathbf{X}\|_2.$$

This completes the proof of Lemma 3.

Appendix B. Proof of Lemma 4

Our goal is to get a numerical estimation on c_* in the case of Gaussian noise. Let $Z_i = \xi_i\,(X_i - \mathbb{E}X_i)$ and

$$\sigma_Z = \max\left\{\left\|\frac{1}{n}\sum_{i=1}^n \mathbb{E}\left(Z_i Z_i^T\right)\right\|_\infty^{1/2}, \left\|\frac{1}{n}\sum_{i=1}^n \mathbb{E}\left(Z_i^T Z_i\right)\right\|_\infty^{1/2}\right\} = \frac{1}{m_1 \wedge m_2}.$$

The constant c_* comes up in the proof of Lemma 2 in [13] in the estimation of

$$\Delta_1 = \left\|\frac{1}{n}\sum_{i=1}^n \xi_i X_i\right\|_\infty \leq \left\|\frac{1}{n}\sum_{i=1}^n \xi_i\,(X_i - \mathbb{E}X_i)\right\|_\infty + \frac{1}{\sqrt{m_1 m_2}}\left|\frac{1}{n}\sum_{i=1}^n \xi_i\right|.$$

A standard application of Markov's inequality gives that, with probability at least $1 - 1/m$

$$\frac{1}{\sqrt{m_1 m_2}}\left|\frac{1}{n}\sum_{i=1}^n \xi_i\right| \leq 2\sqrt{\frac{\log m}{n m_1 m_2}}. \tag{B.1}$$

In [13], the authors estimate $\left\|\frac{1}{n}\sum_{i=1}^{n}\xi_i\left(X_i-\mathbb{E}X_i\right)\right\|_\infty$ using [12, Proposition 2]. To get a numerical estimation on c_* we follow the lines of the proof of [12, Proposition 2]. In order to simplify notations, we write $\|\ \|_\infty = \|\ \|$ and we consider the case of Hermitian matrices of size m'. Its extension to rectangular matrices is straightforward via self-adjoint dilation, cf., for example, 2.6 in [18].

Let $Y_n = \sum_{i=1}^{n} Z_i$. In the proof of [12, Proposition 2], after following the standard derivation of the classical Bernstein inequality and using the Golden-Thompson inequality, the author derives the following bound

$$\mathbb{P}\left(\|Y_n\| \geq t\right) \leq 2m' e^{-\lambda t}\|\mathbb{E}e^{\lambda Z_1}\|^n \tag{B.2}$$

and

$$\|\mathbb{E}e^{\lambda Z_1}\| \leq 1+\lambda^2\left\|\mathbb{E}Z_1^2\left[\frac{e^{\lambda\|Z_1\|}-1-\lambda\|Z_1\|}{\lambda^2\|Z_1\|^2}\right]\right\|. \tag{B.3}$$

Using that $\|Z_1\| \leq 2|\xi_i|$, from (B.3), we compute

(B.4)

$$\begin{aligned}\left\|\mathbb{E}e^{\lambda Z_1}\right\| &\leq 1+\lambda^2\left\|\mathbb{E}\left[\left(X_i-\mathbb{E}X_i\right)^2\right]\mathbb{E}\left(\xi_i^2\left[\frac{e^{2\lambda|\xi_i|}-1-2\lambda|\xi_i|}{4\lambda^2\xi_i^2}\right]\right)\right\| \\ &\leq 1+\lambda^2\sigma_Z^2\mathbb{E}\left(\frac{(2|\xi_i|)^2}{2!}+\frac{\lambda(2|\xi_i|)^3}{3!}+\cdots\right).\end{aligned}$$

Assume that $\lambda < 1$, then (B.4) implies

$$\left\|\mathbb{E}e^{\lambda Z_1}\right\| \leq 1+\lambda^2\sigma_Z^2\mathbb{E}e^{2|\xi_i|} \leq 1+2\lambda^2\sigma_Z^2e^2 \leq \exp\{2\lambda^2\sigma_Z^2e^2\}.$$

Using this bound, from (B.2) we get

$$\mathbb{P}\left(\|Y_n\| \geq t\right) \leq 2m'\exp\{-\lambda t+2\lambda^2\sigma_Z^2e^2\}.$$

It remains now to minimize the last bound with respect to $\lambda \in (0,1)$ to obtain that

$$\mathbb{P}\left(\|Y_n\| \geq t\right) \leq 2m'\exp\left\{-\frac{t^2}{4e^2\sigma_Z^2n}\right\}$$

where we supposed that n is large enough.

Putting $2m'\exp\left\{-\frac{t^2}{4\sigma_Z^2e^2n}\right\} = 1/(2m')$, we get $t = 2e\sqrt{\frac{2\log(2m')n}{m_1\wedge m_2}}$. Using (B.1) we compute the following bound on c_*

$$c_* \leq 2e+1 \leq 6.5.$$

This completes the proof of Lemma 4.

Appendix C. Proof of Lemma 5

Let $\epsilon_i = \sigma\xi_i$. To prove (i) we compute

(C.1)

$$\begin{aligned}\langle \mathbf{M}, \mathbf{M}\rangle = \frac{\|A_0\|_2^2}{(m_1m_2)^2} + \left(1 - \frac{2n}{m_1m_2}\right) \underbrace{\frac{1}{n^2}\sum_{i=1}^n \langle A_0, X_i\rangle^2}_{\mathbf{I}} + \underbrace{\frac{1}{n^2}\sum_{i=1}^n \epsilon_i^2}_{\mathbf{II}} \\ + \left(1 - \frac{n}{m_1m_2}\right) \underbrace{\frac{2}{n^2}\sum_{i=1}^n \langle A_0, X_i\rangle \epsilon_i}_{\mathbf{III}} + \underbrace{\frac{4}{n^2}\sum_{i<j} \epsilon_i \langle A_0, X_j\rangle \langle X_i, X_j\rangle}_{\mathbf{IV}} \\ + \underbrace{\frac{2}{n^2}\sum_{i<j} \epsilon_i\epsilon_j \langle X_i, X_j\rangle}_{\mathbf{V}} + \underbrace{\frac{1}{n^2}\sum_{i\neq j} \langle A_0, X_i\rangle \langle A_0, X_j\rangle \langle X_j, X_i\rangle}_{\mathbf{VI}}.\end{aligned}$$

We estimate each term in (C.1) separately with a good probability. The estimations we give on this probability involve an absolute constant $c > 0$.

I : We have that $\mathbb{E}\left(\frac{1}{n^2}\sum_{i=1}^n \langle A_0, X_i\rangle^2\right) = \frac{\|A_0\|_2^2}{nm_1m_2}$ and $|\langle A_0, X_i\rangle| \leq a$.
Using Hoeffding's inequality , we get that, with probability at least
$1 - 2\exp\{-2\sigma^4 n/(8a)^2\}$

$$\frac{\|A_0\|_2^2}{nm_1m_2} + \frac{\sigma^2}{8\,n} \geq \frac{1}{n^2}\sum_{i=1}^n \langle A_0, X_i\rangle^2 \geq \frac{\|A_0\|_2^2}{nm_1m_2} - \frac{\sigma^2}{8\,n}.$$

II: ϵ_i^2 are sub-exponential random variables and $\mathbb{E}\left(\frac{1}{n^2}\sum_{i=1}^n \epsilon_i^2\right) = \frac{\sigma^2}{n}$. Using Bernstein inequality for sub-exponentials random variables (cf. [19, Proposition 16]) we get that, with probability at least
$1 - 2\exp\left\{-cn\min\left[\sigma^2K/8^2, \sigma\sqrt{K}/8\right]\right\}$

$$\frac{\sigma^2}{n} + \frac{\sigma^2}{8n} \geq \frac{1}{n^2}\sum_{i=1}^n \epsilon_i^2 \geq \frac{\sigma^2}{n} - \frac{\sigma^2}{8n}.$$

III: We have that $\mathbb{E}\left(\frac{2}{n^2}\sum_{i=1}^n \langle A_0, X_i\rangle \epsilon_i\right) = 0$, using Hoeffding's type inequality for sub-Gaussian random variables (cf. [19, Proposition 10]) we get that, with probability at least $1 - e\exp\{-c\sigma^2Kn/a^2\}$

$$\frac{\sigma^2}{8n} \geq \frac{2}{n^2}\sum_{i=1}^n \langle A_0, X_i\rangle \epsilon_i \geq -\frac{\sigma^2}{8n}.$$

IV: We compute $\mathbb{E}\left(\frac{4}{n^2}\sum_{i<j}\epsilon_i \langle A_0, X_j\rangle \langle X_i, X_j\rangle\right) = 0$. We use the following lemma which is proven in the Appendix.

Lemma 11. *Suppose that $n \leq m_1 m_2$. With probability at least $1 - \frac{2}{m_1 m_2}$*

$$\sum_{i<j} \langle X_i, X_j\rangle \leq n.$$

Lemma 11 and Hoeffding's type inequality imply that, with probability at least $1 - 2/m_1m_2 - e\exp\{-c\sigma^2 nK/a^2\}$

$$\frac{\sigma^2}{8n} \geq \frac{4}{n^2}\sum_{i<j}\epsilon_i \langle A_0, X_j\rangle \langle X_i, X_j\rangle \geq -\frac{\sigma^2}{8n}$$

V: We have that $\mathbb{E}\left(\frac{2}{n^2}\sum_{i<j}\epsilon_i\epsilon_j \langle X_i, X_j\rangle\right) = 0$. Using Bernstein type inequalities for sub-exponential random variables and Lemma 11 we get that, with probability at least $1-2\exp\left\{-cn\min\left[\sigma^2K/8^2, \sigma\sqrt{K}/8\right]\right\}$

$$\frac{\sigma^2}{8n} \geq \frac{2}{n^2}\sum_{i<j}\epsilon_i\epsilon_j \langle X_i, X_j\rangle \geq -\frac{\sigma^2}{8n}.$$

VI: We compute that

$$\begin{aligned}
&\mathbb{E}\left(\frac{1}{n^2}\sum_{i\neq j} \langle A_0, X_i\rangle \langle A_0, X_j\rangle \langle X_j, X_i\rangle\right) \\
&= \frac{1}{n^2}\sum_{i\neq j} \langle \mathbb{E}\left(\langle A_0, X_j\rangle X_j\right), \mathbb{E}\left(\langle A_0, X_i\rangle X_i\right)\rangle = \frac{1}{n^2}\sum_{i\neq j}\frac{\|A_0\|_2^2}{(m_1m_2)^2} \\
&\leq \frac{\|A_0\|_2^2}{(m_1m_2)^2}.
\end{aligned}$$

Using Lemma 11 and Hoeffding's inequality, we get that, with probability at least $1 - 2/m_1m_2 - 2\exp\{-2\sigma^4 n/(8a)^2\}$

$$\frac{1}{n^2}\sum_{i\neq j} \langle A_0, X_i\rangle \langle A_0, X_j\rangle \langle X_j, X_i\rangle \leq \frac{\|A_0\|_2^2}{(m_1m_2)^2} + \frac{\sigma^2}{8n}.$$

To obtain the lower bound, note that, for $i \neq j$, $\langle X_i, X_j\rangle \neq 0$ iff $X_i = X_j$. This implies that $\sum_{i\neq j} \langle A_0, X_i\rangle \langle A_0, X_j\rangle \langle X_j, X_i\rangle \geq 0$. We use that $2n < m_1m_2$ to get

$$\frac{\|A_0\|_2^2}{(m_1m_2)^2} + \left(1 - \frac{2n}{m_1m_2}\right)\frac{1}{n^2}\sum_{i=1}^{n} \langle A_0, X_i\rangle^2 \geq 0.$$

Putting the lower bounds in $\mathbf{II} - \mathbf{V}$ together we compute from (C.1)

$$\|\mathbf{M}\|_2^2 \geq \frac{\sigma^2}{2\,n}.$$

To obtain the upper bound, we use the upper bounds in $\mathbf{I}-\mathbf{VI}$. From (C.1) we get

$$\|\mathbf{M}\|_2^2 \le \frac{2\,\|A_0\|_2^2}{(m_1m_2)^2} + \frac{\|A_0\|_2^2}{nm_1m_2} + \frac{14\sigma^2}{8\,n} \le 2\left(\frac{\|A_0\|_2^2}{nm_1m_2} + \frac{\sigma^2}{n}\right)$$

where we used that $2n \le m_1m_2$. This completes the proof of part (i) in Lemma 5.

To prove (ii) we use that $\langle X_i, X_i\rangle = 1$ and $\langle X_i, X_j\rangle \neq 0$ iff $X_i = X_j$. We compute

$$\begin{aligned}\frac{1}{n^2}\left\langle \sum_{i=1}^n Y_iX_i, \sum_{i=1}^n Y_iX_i\right\rangle &= \frac{1}{n^2}\sum_{i=1}^n Y_i^2 + \frac{2}{n^2}\sum_{i<j} Y_iY_j\,\langle X_i, X_j\rangle \\ &= \frac{1}{n^2}\sum_{i=1}^n \left(\langle A_0, X_i\rangle^2 + \epsilon_i^2 + 2\,\langle A_0, X_i\rangle\,\epsilon_i\right) \\ &\quad + \frac{2}{n^2}\sum_{i<j}\langle A_0, X_i\rangle^2\,\langle X_i, X_j\rangle \\ &\quad + \frac{4}{n^2}\sum_{i<j}\epsilon_i\,\langle A_0, X_j\rangle\,\langle X_i, X_j\rangle + \frac{2}{n^2}\sum_{i<j}\epsilon_i\epsilon_j\,\langle X_i, X_j\rangle\,.\end{aligned}$$

This implies that

(C.2)

$$\begin{aligned}\frac{1}{n^2}\left\langle \sum_{i=1}^n Y_iX_i, \sum_{i=1}^n Y_iX_i\right\rangle \ge{}& \underbrace{\frac{1}{n^2}\sum_{i=1}^n \langle A_0, X_i\rangle^2}_{\mathbf{I}} + \underbrace{\frac{1}{n^2}\sum_{i=1}^n \epsilon_i^2}_{\mathbf{II}} + \underbrace{\frac{2}{n^2}\sum_{i=1}^n \langle A_0, X_i\rangle\,\epsilon_i}_{\mathbf{III}} \\ &+ \underbrace{\frac{4}{n^2}\sum_{i<j}\epsilon_i\,\langle A_0, X_j\rangle\,\langle X_i, X_j\rangle}_{\mathbf{IV}} + \underbrace{\frac{2}{n^2}\sum_{i<j}\epsilon_i\epsilon_j\,\langle X_i, X_j\rangle}_{\mathbf{V}}.\end{aligned}$$

Using the lower bounds for $\mathbf{I}-\mathbf{V}$ we get from (C.2)

$$\frac{1}{n^2}\left\langle \sum_{i=1}^n Y_iX_i, \sum_{i=1}^n Y_iX_i\right\rangle \ge \frac{\|A_0\|_2^2}{nm_1m_2}$$

which proves the part (ii) of Lemma 5.

(iii) is a consequence of (ii). For $4n \le m_1m_2$ (ii) implies

$$\frac{1}{4n^2}\left\langle \sum_{i=1}^n Y_iX_i, \sum_{i=1}^n Y_iX_i\right\rangle \ge \frac{\|A_0\|_2^2}{(m_1m_2)^2}.$$

Now we complete the proof of part (iii) of Lemma 5 using that

$$\|\mathbf{M}\|_2 \ge \left\|\frac{1}{n}\sum_{i=1}^n Y_iX_i\right\|_2 - \frac{\|A_0\|_2}{m_1m_2}.$$

Appendix D. Proof of Lemma 11

Recall that for $i \neq j$, X_i and X_j are independent. We compute the expectation

$$\mathbb{E}\left(\sum_{i<j}\langle X_i, X_j\rangle\right) = \sum_{i<j}\langle \mathbb{E}X_i, \mathbb{E}X_j\rangle = \frac{n(n-1)}{2m_1m_2}$$

and the variance

$$\mathbb{E}\left(\left(\sum_{i<j}\langle X_i, X_j\rangle\right)^2\right) - \left(\mathbb{E}\left(\sum_{i<j}\langle X_i, X_j\rangle\right)\right)^2 = \mathbb{E}\left(\sum_{\substack{i<j\\ i'<j'}}\langle X_i, X_j\rangle\langle X_{i'}, X_{j'}\rangle\right) - \sum_{\substack{i<j\\ i'<j'}}\mathbb{E}\left(\langle X_i, X_j\rangle\right)\mathbb{E}\left(\langle X_{i'}, X_{j'}\rangle\right).$$

When i, j, i', j' are all distinct, $\mathbb{E}\left(\langle X_i, X_j\rangle\langle X_{i'}, X_{j'}\rangle\right)$ is canceled by the corresponding term in $\sum_{\substack{i<j\\ i'<j'}}\mathbb{E}\left(\langle X_i, X_j\rangle\right)\mathbb{E}\left(\langle X_{i'}, X_{j'}\rangle\right)$. Then, it remains to consider the following five cases: (1) $i = i'$ and $j = j'$; (2) $i = i'$ and $j \neq j'$; (3) $i \neq i'$ and $j = j'$; (4) $i = j'$ and $j \neq i'$; (5) $i' = j$ and $j' \neq i$.

case (1) Note that $\langle X_i, X_j\rangle$ takes only two values 0 or 1, which implies that

$$\mathbb{E}\left(\langle X_i, X_j\rangle^2\right) = \mathbb{E}\left(\langle X_i, X_j\rangle\right) = \frac{1}{m_1m_2}.$$

cases (2)-(5) In these four cases, the calculation reduces to calculate $\mathbb{E}\left(\langle X_i, X_k\rangle\langle X_k, X_j\rangle\right)$ for $i \neq j$ and $k \notin \{i, j\}$. Note that $\mathcal{P}_{X_k} = \langle\cdot, X_k\rangle X_k$ is the orthogonal projector on the vector space spanned by X_k. We compute

$$\mathbb{E}\mathcal{P}_{X_k} = \frac{1}{m_1m_2}\mathrm{Id}$$

where Id is the identity application on $\mathbb{R}^{m_1\times m_2}$. Then, we get

$$\begin{aligned}\mathbb{E}\left(\langle\langle X_i, X_k\rangle X_k, X_j\rangle\right) &= \mathbb{E}\left(\langle\mathcal{P}_{X_k}(X_i), X_j\rangle\right)\\ &= \langle\mathbb{E}\left(\mathcal{P}_{X_k}\right)\left(\mathbb{E}X_i\right), \mathbb{E}X_j\rangle\\ &= \frac{1}{m_1m_2}\langle\mathbb{E}X_i, \mathbb{E}X_j\rangle = \frac{1}{\left(m_1m_2\right)^2}.\end{aligned}$$

These terms are canceled by the corresponding terms in $\sum_{\substack{i<j\\ i'<j'}}\mathbb{E}\left(\langle X_i, X_j\rangle\right)\mathbb{E}\left(\langle X_{i'}, X_{j'}\rangle\right)$ as $\mathbb{E}\left(\langle X_i, X_k\rangle\right)\mathbb{E}\left(\langle X_k, X_j\rangle\right) = \frac{1}{(m_1m_2)^2}$.

Finally we get that

$$\mathbb{E}\left(\left(\sum_{i<j}\langle X_i, X_j\rangle\right)^2\right) - \left(\mathbb{E}\left(\sum_{i<j}\langle X_i, X_j\rangle\right)\right)^2 \leq \frac{n(n-1)}{2m_1m_2}.$$

The Bienaymé-Tchebychev inequality implies that

$$\mathbb{P}\left(\sum_{i<j}\langle X_i, X_j\rangle \geq n\right) \leq \frac{n(n-1)}{2m_1m_2\left(n-\frac{n(n-1)}{2m_1m_2}\right)^2} \leq \frac{2}{m_1m_2}$$

when $m_1m_2 \geq n$. This completes the proof of Lemma 11.

Appendix E. Proof of Theorem 8

We need the following auxiliary result, which corresponds to Lemma 3, and which is proven in the appendix.

Lemma 12. *Suppose that* $\frac{\rho}{\sqrt{\operatorname{rank}(VA_0)}} \geq \lambda \geq 3\Delta'$ *for some* $\rho < 1$, *then*

$$\left\|V\hat{A}-U\right\|_2 \geq \left(\frac{3-\sqrt{1+\rho^2}}{3+\sqrt{1+\rho^2}}\right)\|E\|_2.$$

We resume the proof of Theorem 8. If $V\hat{A} = U$, then Lemma 12 implies that $VA_0 = U$ and we get $\left\|V\left(\hat{A}-A_0\right)\right\|_2 = 0$.
If $V\hat{A} \neq U$, a necessary condition of extremum in (2.11) implies that there exists a $\hat{W} \in \partial\|V\hat{A}\|_1$ such that for any $A \in \mathbb{R}^{m_1\times m_2}$

$$\frac{2\left\langle V\hat{A}-U, V\left(\hat{A}-A\right)\right\rangle}{2\left\|V\hat{A}-U\right\|_2} + \lambda\left\langle \hat{W}, V\left(\hat{A}-A\right)\right\rangle \leq 0$$

that is

(E.1)

$$2\left\langle V\left(\hat{A}-A_0\right), V\left(\hat{A}-A\right)\right\rangle - 2\left\langle U-VA_0, V\left(\hat{A}-A\right)\right\rangle + 2\lambda\left\|V\hat{A}-U\right\|_2\left\langle \hat{W}, V\left(\hat{A}-A\right)\right\rangle \leq 0.$$

For a $W_A \in \left\{\sum_{j=1}^{\operatorname{rank}(VA)} u_j(VA)v_j^T(VA) + P_{S_1^\perp(VA)}WP_{S_2^\perp(VA)} \ : \ \|W\|_\infty \leq 1\right\}$, by the monotonicity of subdifferentials of convex functions we have that

$$\left\langle \hat{W}-W_A, V\left(\hat{A}-A\right)\right\rangle \geq 0.$$

Then, (E.1) and

$$2\left\langle V\left(\hat{A}-A_0\right), V\left(\hat{A}-A\right)\right\rangle = \left\|V\left(\hat{A}-A_0\right)\right\|_2^2 + \left\|V\left(\hat{A}-A\right)\right\|_2^2 - \left\|V\left(A-A_0\right)\right\|_2^2$$

imply

(E.2)
$$\begin{aligned}\left\|V\left(\hat{A}-A_0\right)\right\|_2^2+\left\|V\left(\hat{A}-A\right)\right\|_2^2 &\\ +2\lambda\left\|V\hat{A}-U\right\|_2\left\langle P_{S_1^\perp(VA)}WP_{S_2^\perp(VA)},V\left(\hat{A}-A\right)\right\rangle &\\ \leq\left\|V\left(A-A_0\right)\right\|_2^2+2\left\langle E,V\left(\hat{A}-A\right)\right\rangle &\\ -2\lambda\left\|V\hat{A}-U\right\|_2\left\langle\sum_{j=1}^{\operatorname{rank}(VA)}u_j(VA)v_j(VA)^T,V\left(\hat{A}-A\right)\right\rangle.&\end{aligned}$$

By trace duality we can pick W with $\|W\|_\infty\leq 1$ such that

(E.3)
$$\begin{aligned}\left\langle P_{S_1^\perp(VA)}WP_{S_2^\perp(VA)},V\left(\hat{A}-A\right)\right\rangle&=\left\langle W,P_{S_1^\perp(VA)}V\left(\hat{A}-A\right)P_{S_2^\perp(VA)}\right\rangle\\&=\left\|P_{S_1^\perp(VA)}V\left(\hat{A}-A\right)P_{S_2^\perp(VA)}\right\|_1.\end{aligned}$$

Let $\mathbf{P}_{VA}(B)=B-P_{S_1^\perp(VA)}BP_{S_2^\perp(VA)}$. Then, using the trace duality and the triangle inequality we get

(E.4)
$$\begin{aligned}\left\langle E,V\left(\hat{A}-A\right)\right\rangle=\left\langle\mathcal{P}_VE,V\left(\hat{A}-A\right)\right\rangle&\leq\|\mathcal{P}_VE\|_\infty\left\|V\left(\hat{A}-A\right)\right\|_1\\&\leq\|\mathcal{P}_VE\|_\infty\left\|\mathbf{P}_{VA}\left[V\left(\hat{A}-A\right)\right]\right\|_1\\&\quad+\|\mathcal{P}_VE\|_\infty\left\|P_{S_1^\perp(VA)}V\left(\hat{A}-A\right)P_{S_2^\perp(VA)}\right\|_1.\end{aligned}$$

Note that $\left\|\sum_{j=1}^{\operatorname{rank}(VA)}u_j(VA)v_j^T(VA)\right\|_\infty=1$. Hence the trace duality implies

(E.5)
$$\begin{aligned}\left\langle\sum_{j=1}^{\operatorname{rank}(VA)}u_j(VA)v_j^T(VA),V\left(\hat{A}-A\right)\right\rangle&=\left\langle\sum_{j=1}^{\operatorname{rank}(VA)}u_j(VA)v_j(VA)^T,\mathbf{P}_{VA}\left[V\left(\hat{A}-A\right)\right]\right\rangle\\&\leq\left\|\mathbf{P}_{VA}\left[V\left(\hat{A}-A\right)\right]\right\|_1.\end{aligned}$$

Plugging (E.3), (E.4) and (E.5) into (E.2) we obtain

(E.6)
$$\begin{aligned}\left\|V\left(\hat{A}-A_0\right)\right\|_2^2+\left\|V\left(\hat{A}-A\right)\right\|_2^2+2\lambda\left\|V\hat{A}-U\right\|_2\left\|P_{S_1^\perp(VA)}V\left(\hat{A}-A\right)P_{S_2^\perp(VA)}\right\|_1&\\ \leq\left\|V\left(A-A_0\right)\right\|_2^2+2\|\mathcal{P}_VE\|_\infty\left\|P_{S_1^\perp(VA)}V\left(\hat{A}-A\right)P_{S_2^\perp(VA)}\right\|_1&\\ +2\|\mathcal{P}_VE\|_\infty\left\|\mathbf{P}_{VA}\left[V\left(\hat{A}-A\right)\right]\right\|_1+2\lambda\left\|V\hat{A}-U\right\|_2\left\|\mathbf{P}_{VA}\left[V\left(\hat{A}-A\right)\right]\right\|_1.&\end{aligned}$$

Using Lemma 12, from (E.6) we compute

$$\begin{aligned}
&\left\|V\left(\hat{A}-A_0\right)\right\|_2^2+\left\|V\left(\hat{A}-A\right)\right\|_2^2\\
&\qquad+2\lambda\left(\frac{3-\sqrt{1+\rho^2}}{3+\sqrt{1+\rho^2}}\right)\|E\|_2\left\|P_{S_1^\perp(VA)}V\left(\hat{A}-A\right)P_{S_2^\perp(VA)}\right\|_1\\
&\qquad\le\left\|V\left(A-A_0\right)\right\|_2^2+2\left\|\mathcal{P}_V E\right\|_\infty\left\|P_{S_1^\perp(VA)}V\left(\hat{A}-A\right)P_{S_2^\perp(VA)}\right\|_1\\
&\qquad+2\left\|\mathcal{P}_V E\right\|_\infty\left\|\mathbf{P}_{VA}\left[V\left(\hat{A}-A\right)\right]\right\|_1\\
&\qquad+2\lambda\left\|V\hat{A}-U\right\|_2\left\|\mathbf{P}_{VA}\left[V\left(\hat{A}-A\right)\right]\right\|_1.
\end{aligned}$$

From the definition of λ we get

(E.7)
$$\begin{aligned}
&\left\|V\left(\hat{A}-A_0\right)\right\|_2^2+\left\|V\left(\hat{A}-A\right)\right\|_2^2\\
&\qquad+6\left(\frac{3-\sqrt{1+\rho^2}}{3+\sqrt{1+\rho^2}}\right)\left\|\mathcal{P}_V E\right\|_\infty\left\|P_{S_1^\perp(VA)}V\left(\hat{A}-A\right)P_{S_2^\perp(VA)}\right\|_1\\
&\qquad\le\left\|V\left(A-A_0\right)\right\|_2^2+2\left\|\mathcal{P}_V E\right\|_\infty\left\|P_{S_1^\perp(VA)}V\left(\hat{A}-A\right)P_{S_2^\perp(VA)}\right\|_1\\
&\qquad+2\left\|\mathcal{P}_V E\right\|_\infty\left\|\mathbf{P}_{VA}\left[V\left(\hat{A}-A\right)\right]\right\|_1\\
&\qquad+2\lambda\left\|V\hat{A}-U\right\|_2\left\|\mathbf{P}_{VA}\left[V\left(\hat{A}-A\right)\right]\right\|_1.
\end{aligned}$$

Note that $6\dfrac{3-\sqrt{1+\rho^2}}{3+\sqrt{1+\rho^2}}\ge 2$ for any $\rho<1$. Thus, (E.7) implies

(E.8)
$$\begin{aligned}
\left\|V\left(\hat{A}-A_0\right)\right\|_2^2+\left\|V\left(\hat{A}-A\right)\right\|_2^2&\le\left\|V\left(A-A_0\right)\right\|_2^2\\
&\quad+2\left\|\mathcal{P}_V E\right\|_\infty\left\|\mathbf{P}_{VA}\left[V\left(\hat{A}-A\right)\right]\right\|_1\\
&\quad+2\lambda\left\|V\hat{A}-U\right\|_2\left\|\mathbf{P}_{VA}\left[V\left(\hat{A}-A\right)\right]\right\|_1.
\end{aligned}$$

Using $\left\|V\hat{A}-U\right\|_2\le\left\|V\hat{A}-VA_0\right\|_2+\left\|VA_0-U\right\|_2$, and the fact that

$$\left\|\mathbf{P}_{VA}\left[V\left(\hat{A}-A\right)\right]\right\|_1\le\sqrt{2\mathrm{rank}(VA)}\left\|V\left(\hat{A}-A\right)\right\|_2$$

from (E.8) we compute

(E.9)
$$\begin{aligned}\left\|V\left(\hat{A}-A_0\right)\right\|_2^2+\left\|V\left(\hat{A}-A\right)\right\|_2^2 \leq \|V\left(A-A_0\right)\|_2^2 \\ +2\lambda\sqrt{2\operatorname{rank}(VA)}\left\|V\left(\hat{A}-A_0\right)\right\|_2\left\|V\left(\hat{A}-A\right)\right\|_2 \\ +2\lambda\|E\|_2\sqrt{2\operatorname{rank}(VA)}\left\|V\left(\hat{A}-A\right)\right\|_2 \\ +2\left\|\mathcal{P}_V E\right\|_\infty\sqrt{2\operatorname{rank}(VA)}\left\|V\left(\hat{A}-A\right)\right\|_2 .\end{aligned}$$

From the definition of λ we get that $\left\|\mathcal{P}_V E\right\|_\infty \leq \lambda\|E\|_2/3$ and $\lambda\sqrt{2\operatorname{rank}(VA)} \leq \rho$. This implies that

$$\begin{aligned}\left\|V\left(\hat{A}-A_0\right)\right\|_2^2+\left\|V\left(\hat{A}-A\right)\right\|_2^2 \leq \|V\left(A-A_0\right)\|_2^2 \\ +8/3\lambda\|E\|_2\sqrt{2\operatorname{rank}(VA)}\left\|V\left(\hat{A}-A\right)\right\|_2 \\ +2\rho\left\|V\left(\hat{A}-A_0\right)\right\|_2\left\|V\left(\hat{A}-A\right)\right\|_2 .\end{aligned}$$

Using $2ab \leq a^2+b^2$ twice we finally compute

$$\begin{aligned}(1-\rho)\left\|V\left(\hat{A}-A_0\right)\right\|_2^2+\left\|V\left(\hat{A}-A\right)\right\|_2^2 \leq \|V\left(A-A_0\right)\|_2^2+\rho\left\|V\left(\hat{A}-A\right)\right\|_2^2 \\ +8/3\lambda\|E\|_2\sqrt{2\operatorname{rank}(VA)}\left\|V\left(\hat{A}-A\right)\right\|_2\end{aligned}$$

and

$$(1-\rho)\left\|V\left(\hat{A}-A_0\right)\right\|_2^2 \leq \|V\left(A-A_0\right)\|_2^2+\frac{4\lambda^2}{1-\rho}\|E\|_2^2\operatorname{rank}(VA)$$

which implies the statement of Theorem 8.

Appendix F. Proof of Lemma 12

If $VA_0 = U$, then we have trivially $\|V\hat{A}-U\|_2 \geq 0$. If $VA_0 \neq U$, by the convexity of function $A \to \|VA-U\|_2$, we have

(F.1)
$$\begin{aligned}\left\|V\hat{A}-U\right\|_2-\|VA_0-U\|_2 &\geq \frac{\left\langle VA_0-U, V\left(\hat{A}-A_0\right)\right\rangle}{\|VA_0-U\|_2} \\ &= \frac{\left\langle\mathcal{P}_V(E), V\left(\hat{A}-A_0\right)\right\rangle}{\|VA_0-U\|_2} \geq -\frac{\|\mathcal{P}_V(E)\|_\infty}{\|E\|_2}\left\|V\left(\hat{A}-A_0\right)\right\|_1 \\ &\geq -\frac{\|\mathcal{P}_V(E)\|_\infty}{\|E\|_2}\sqrt{\operatorname{rank}(VA_0)+\operatorname{rank}(V\hat{A})}\left\|V\left(\hat{A}-A_0\right)\right\|_2 .\end{aligned}$$

Using the bound $\frac{\rho}{\sqrt{\operatorname{rank}(VA)}} \geq \lambda$, Lemma 7 and the triangle inequality from (F.1) we get

$$\left\|V\hat{A}-U\right\|_2 - \|VA_0-U\|_2 \geq \\ -\frac{\sqrt{1+\rho^2}}{\lambda}\frac{\|\mathcal{P}_V(E)\|_\infty}{\|E\|_2}\left(\|V\hat{A}-U\|_2+\|VA_0-U\|_2\right).$$

By the definition of λ we have $\frac{\|\mathcal{P}_V(E)\|_\infty}{\lambda\|E\|_2} \leq 1/3$ which finally leads to

$$\left(1+\sqrt{1+\rho^2}/3\right)\|V\hat{A}-U\|_2 \geq \left(1-\sqrt{1+\rho^2}/3\right)\|VA_0-U\|_2.$$

This completes the proof of Lemma 12.

Acknowledgements. It is a pleasure to thank A. Tsybakov for introducing me this problem and illuminating discussions.

References

[1] Belloni, A., Chernozhukov, V. and Wang, L. (2011) Square-root Lasso: Pivotal Recovery of Sparse Signals via Conic Programming. *Biometrika*, to appear.

[2] Bunea, F., She, Y. and Wegkamp,M. (2011) Optimal selection of reduced rank estimators of high-dimensional matrices. *Annals of Statistics*, **39**, 1282–1309.

[3] Candès, E. J. and Plan, Y. (2009). Matrix completion with noise. *Proceedings of IEEE*.

[4] Candès, E.J. and Recht, B. (2009) Exact matrix completion via convex optimization. *Fondations of Computational Mathematics*, **9(6)**, 717-772.

[5] Candès, E.J. and Tao, T. (2009) The power of convex relaxation: Near-optimal matrix completion. *IEEE Trans. Inform. Theory*, **56(5)**, 2053-2080.

[6] Gaïffas, S. and Lecué, G. (2010) Sharp oracle inequalities for the prediction of a high-dimensional matrix. *IEEE Transactions on Information Theory*, to appear.

[7] Giraud,C. (2011) Low rank Multivariate regression. *Electronic Journal of Statistics*, **5**, 775-799.

[8] Gross, D. (2011) Recovering low-rank matrices from few coefficients in any basis. *IEEE Transactions*, **57(3)**, 1548-1566.

[9] Keshavan, R.H., Montanari, A. and Oh, S. (2010) Matrix completion from noisy entries. *Journal of Machine Learning Research*, **11**, 2057-2078.

[10] Keshavan, R.H., Montanari, A. and Oh, S. (2010) Matrix completion from a few entries. *IEEE Trans. on Info. Th.*, **56(6)**, 2980-2998.

[11] Klopp, O. (2011) Rank penalized estimators for high-dimensional matrices. *Electronic Journal of Statistics*, **5**, 1161-1183.

[12] Koltchinskii, V. (2011) Von Neumann Entropy Penalization and Low Rank Matrix Estimation. *Annals of Statistics*, to appear.

[13] Koltchinskii, V., Lounici, K. and Tsybakov, A. (2011) Nuclear norm penalization and optimal rates for noisy low rank matrix completion. *Annals of Statistics*, to appear.

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

[15] Negahban, S. and Wainwright, M. J. (2010). Restricted strong convexity and weighted matrix completion: Optimal bounds with noise. `Arxiv`: 1009.2118.

[16] Recht, B. (2009) A simpler approach to matrix completion. *Journal of Machine Learning Research*, to appear.

[17] Rohde, A. and Tsybakov, A. (2011) Estimation of High-Dimensional Low-Rank Matrices. *Annals of Statistics*, **39**, 887-930.

[18] Tropp, J.A. (2011) User-friendly tail bounds for sums of random matrices. *Found. Comput. Math.*, to appear.

[19] Vershynin, R. (2012) *Introduction to the non-asymptotic analysis of random matrices.* Chapter 5 of the book Compressed Sensing, Theory and Applications, ed. Y. Eldar and G. Kutyniok. Cambridge University Press.

[20] Watson, G. A. (1992) Characterization of the subdifferential of some matrix norms. *Linear Algebra Appl.*, **170**, 33-45.

Laboratoire de Statistique, CREST and University Paris Dauphine, CREST 3, Av. Pierre Larousse 92240 Malakoff France

E-mail address: `olga.klopp@ensae.fr`