Submitted to the Annals of Statistics
arXiv: math.PR/0000000

NUCLEAR NORM PENALIZATION AND OPTIMAL RATES FOR NOISY LOW-RANK MATRIX COMPLETION

BY VLADIMIR KOLTCHINSKII[*,‡], KARIM LOUNICI[‡] AND ALEXANDRE B. TSYBAKOV[†,§]

Georgia Institute of Technology[‡] *and CREST*[§]

*Abstract*This paper deals with the trace regression model where n entries or linear combinations of entries of an unknown $m_1 \times m_2$ matrix A_0 corrupted by noise are observed. We propose a new nuclear norm penalized estimator of A_0 and establish a general sharp oracle inequality for this estimator for arbitrary values of n, m_1, m_2 under the condition of isometry in expectation. Then this method is applied to the matrix completion problem. In this case, the estimator admits a simple explicit form and we prove that it satisfies oracle inequalities with faster rates of convergence than in the previous works. They are valid, in particular, in the high-dimensional setting $m_1 m_2 \gg n$. We show that the obtained rates are optimal up to logarithmic factors in a minimax sense and also derive, for any fixed matrix A_0, a non-minimax lower bound on the rate of convergence of our estimator, which coincides with the upper bound up to a constant factor. Finally, we show that our procedure provides an exact recovery of the rank of A_0 with probability close to 1. We also discuss the statistical learning setting where there is no underlying model determined by A_0 and the aim is to find the best trace regression model approximating the data.

1. Introduction. Assume that we observe n independent random pairs $(X_i, Y_i), i = 1, \dots, n$, where X_i are random matrices with dimensions $m_1 \times m_2$ and Y_i are random variables in $\mathbb{R}$, satisfying the trace regression model:

$$\mathbb{E}(Y_i|X_i) = \mathrm{tr}(X_i^\top A_0), \quad i = 1, \dots, n, \tag{1.1}$$

where $A_0 \in \mathbb{R}^{m_1 \times m_2}$ is an unknown matrix, $\mathbb{E}(Y_i|X_i)$ is the conditional expectation of Y_i given X_i, and $\mathrm{tr}(B)$ denotes the trace of matrix B. We consider the problem of estimation of A_0 based on the observations $(X_i, Y_i), i = 1, \dots, n$. Though the results of this paper are obtained for general n, m_1, m_2,

[*]Supported in part by NSF grants DMS-09-06880 and CCF-0808863

[†]Supported in part by ANR "Parcimonie" and by PASCAL-2 Network of Excellence

AMS 2000 subject classifications: Primary 62J99,62H12; secondary 60B20, 60G15

Keywords and phrases: matrix completion, low-rank matrix estimation, recovery of the rank, statistical learning, optimal rate of convergence, noncommutative Bernstein inequality, Lasso

the main motivation is in the high-dimensional setting, which corresponds to $m_1 m_2 \gg n$, with low-rank matrices A_0.

It will be convenient to write the model (1.1) in the form

$$Y_i = \mathrm{tr}(X_i^\top A_0) + \xi_i, \quad i = 1, \dots, n, \tag{1.2}$$

where the noise variables $\xi_i = Y_i - \mathbb{E}(Y_i|X_i)$ are independent and have zero means.

For any matrices $A, B \in \mathbb{R}^{m_1 \times m_2}$, we define the scalar products

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

and

$$\langle A, B \rangle_{L_2(\Pi)} = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\big(\langle A, X_i \rangle \langle B, X_i \rangle\big) .$$

Here $\Pi = \frac{1}{n} \sum_{i=1}^{n} \Pi_i$, where Π_i denotes the distribution of X_i. The corresponding norm $\|A\|_{L_2(\Pi)}$ is given by

$$\|A\|_{L_2(\Pi)}^2 = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\big(\langle A, X_i \rangle^2\big) .$$

Example 1. Matrix Completion. Assume that the design matrices X_i are i.i.d. uniformly distributed on the set

$$\mathcal{X} = \Big\{ e_j(m_1) e_k^\top(m_2), 1 \le j \le m_1,\ 1 \le k \le m_2 \Big\} , \tag{1.3}$$

where $e_k(m)$ are the canonical basis vectors in $\mathbb{R}^m$. The set $\mathcal{X}$ forms an orthonormal basis in the space of $m_1 \times m_2$ matrices that will be called the matrix completion basis. Let also $n < m_1 m_2$. Then the problem of estimation of A_0 coincides with the problem of matrix completion under uniform sampling at random (USR) as studied in the non-noisy case ($\xi_i = 0$) in [14, 21], and in the noisy case in [13, 23]. Considering low-rank matrices A_0 is of a particular interest. Clearly, for such X_i we have the isometry

$$\|A\|_{L_2(\Pi)}^2 = \mu^{-2} \|A\|_2^2, \tag{1.4}$$

for all matrices $A \in \mathbb{R}^{m_1 \times m_2}$, where $\mu = \sqrt{m_1 m_2}$, and $\|A\|_2$ is the Frobenius norm of A. However, the restricted isometry property in the usual sense, i.e., "in probability", cf., e.g., [22], does not hold for matrix completion, since for $n < m_1 m_2$ there trivially exists a matrix of rank 1 in the null space of the sampling operator.

One can also consider more general matrix measurement models in which, for a given orthonormal basis in the space of matrices, a random sample of Fourier coefficients of the target matrix A_0 is observed subject to a random noise. For more discussion on matrix completion with other types of sampling, see [8, 10, 11, 15, 17] and references therein.

Example 2. Column masks. Assume that the design matrices X_i are i.i.d. replications of a random matrix X, which has only one nonzero column. For instance, let the distribution of X be such that all the columns have equal probability to be non-zero, and the random entries of nonzero column $x_{(j)}$ are such that $\mathbb{E}(x_{(j)}x_{(j)}^\top)$ is the identity matrix. Then $\|A\|_{L_2(\Pi)}^2 = \|A\|_2^2/m_2, \ \forall A \in \mathbb{R}^{m_1\times m_2}$, so that condition (1.4) is satisfied with $\mu = \sqrt{m_2}$. More generally, in view of application to multi-task learning, cf. [23], one can be interested in considering non-identically distributed X_i. The model can be then reformulated as a longitudinal regression model, with different distributions of X_i corresponding to different tasks.

Example 3. "Complete" subgaussian design. Assume that the design matrices X_i are i.i.d. replications of a random matrix X such that $\langle A, X\rangle$ is a subgaussian random variable for any $A \in \mathbb{R}^{m_1\times m_2}$. This approach has its roots in compressed sensing. The two major examples are given by the matrices X whose entries are either i.i.d. standard Gaussian or Rademacher random variables. In both cases, we have $\|A\|_{L_2(\Pi)}^2 = \|A\|_2^2, \ \forall A \in \mathbb{R}^{m_1\times m_2}$, so that condition (1.4) is satisfied with $\mu = 1$. The problem of exact reconstruction of A_0 under such a design in the non-noisy setting was studied in [9, 18, 22], whereas estimation of A_0 in the presence of noise is analyzed in [9, 18, 23], among which [9, 17, 23] treat the high-dimensional case $m_1m_2 > n$.

Example 4. Fixed design. Assume that all the Π_i are Dirac measures, so that the design matrices X_i are non-random. Then $\|A\|_{L_2(\Pi)}^2 = \frac{1}{n}\sum_{i=1}^n \langle A, X_i\rangle^2$, and we get the problem of trace regression with fixed design, cf. [23]. In particular, if $m_1 = m_2$, and A and X_i are diagonal matrices the trace regression model (1.2) becomes the usual linear regression model. Accordingly, the rank of A becomes the number of its non-zero diagonal elements. This observation will allow us to deduce, as a consequence of our general arguments, an oracle inequality for sparse linear regression with fixed design and the Lasso improving [6] in the sense that the inequality is sharp (cf. Theorem 2 and Section 5.4).

The general oracle inequalities that we will prove in Section 2 can be successfully applied to the above examples. The emphasis in this paper will be on the matrix completion problem (Example 1), for which the previously obtained results were suboptimal.

Statistical estimation of low-rank matrices has recently become a very active field with a rapidly growing literature. The most popular methods are based on penalized empirical risk minimization with nuclear norm penalty [2–5, 7–9, 13, 18, 23]. Estimators with other types of penalization, such as the Schatten-p norm [23], the von Neumann entropy [17], penalization by the rank [7, 12] or some combined penalties [13] are also discussed.

It is worth pointing out that in many applications, such as in matrix completion, the distribution Π is known, and yet this information has not been exploited since the penalized estimation procedures considered in the literature involve the empirical risk $\frac{1}{n}\sum_{i=1}^{n}(Y_i - \mathrm{tr}(X_i^\top A))^2$ (cf. [13, 23]). In this paper we incorporate the knowledge of Π in the construction and we study the following estimator of A_0:

$$\hat{A}^\lambda = \mathrm{argmin}_{A \in \mathbb{A}} L_n(A), \tag{1.5}$$

where $\mathbb{A} \subseteq \mathbb{R}^{m_1 \times m_2}$ is a convex set of matrices,

$$L_n(A) = \|A\|_{L_2(\Pi)}^2 - \left\langle \frac{2}{n}\sum_{i=1}^{n} Y_i X_i, A \right\rangle + \lambda \|A\|_1, \tag{1.6}$$

$\lambda > 0$ is a regularization parameter, and $\|A\|_1$ is the nuclear norm of A. Note that if all X_i are non-random, $\hat{A}^\lambda$ coincides with the usual matrix Lasso estimator:

$$\hat{A}^\lambda = \mathrm{argmin}_{A \in \mathbb{A}} \left[n^{-1} \sum_{j=1}^{n} (Y_j - \langle A, X_j \rangle)^2 + \lambda \|A\|_1 \right]. \tag{1.7}$$

The emphasis in this paper is on the noisy matrix completion setting. Then the estimator $\hat{A}^\lambda$ has a particularly simple form; it is obtained from the matrix $\frac{m_1 m_2}{n}\sum_{i=1}^{n} Y_i X_i$ by soft thresholding of its singular values. One of the main results of this paper is to show that our estimators are rate optimal (up to logarithmic factors) under the Frobenius error for a simple class of matrices $\mathcal{A}(r, a)$ defined by two restrictions: the rank of A_0 is not larger than given r and all the entries of A_0 are bounded in absolute value by a constant a. This rather intuitive class has been first considered in [15]. However, the construction of the estimator in [15] requires the exact knowledge of $\mathrm{rank}(A_0)$ and the upper bound on the Frobenius error obtained in [15] is suboptimal (see the details in Section 3). The recent preprint [13] obtains suboptimal bounds of "slow rate" type for matrix completion. The analysis in [17] is focused on complex-valued Hermitian matrices with nuclear norm equal to 1 and motivated by density matrix estimation problem in quantum state tomography. These papers do not address the optimality issue. Optimal rates in noisy matrix completion are derived in [23], but on different classes of matrices and with the empirical prediction error rather than with the Frobenius error. Finally, a very recent preprint [19] discusses the optimality issue for the Frobenius error on the classes defined in terms of a "spikiness index" of A_0, which are not related to $\mathcal{A}(r,a)$, and proposes estimators that require prior information about this index.

The main contributions of this paper are the following. In Section 2 we derive a general oracle inequality for the prediction error $\|\hat{A}^\lambda - A_0\|^2_{L_2(\Pi)}$. This oracle inequality is sharp, i.e., with leading constant 1, both in the case of "slow rate" (for matrices A_0 with small nuclear norm) and in the case of "fast rate" (for matrices A_0 with small rank). As a particular instance of this general result, in Section 3 we obtain an oracle inequality for the matrix completion problem. In Section 4, we establish minimax lower bounds showing that the rates for matrix completion obtained in Section 3 are optimal up to a logarithmic factor. In Section 5, we briefly discuss some other implications and extensions of our method. Finally, Section 6 is devoted to the control of the stochastic term appearing in the proof of the upper bound.

2. General oracle inequalities.

We recall first some basic facts about matrices. Let $A \in \mathbb{R}^{m_1\times m_2}$ be a rectangular matrix, and let $r = \operatorname{rank}(A) \le \min(m_1, m_2)$ denote its rank. The singular value decomposition (SVD) of A has the form: $A = \sum_{j=1}^r \sigma_j(A) u_j v_j^\top$ with orthonormal vectors $u_1, \dots, u_r \in \mathbb{R}^{m_1}$, orthonormal vectors $v_1, \dots, v_r \in \mathbb{R}^{m_2}$ and real numbers $\sigma_1(A) \ge \cdots \ge \sigma_r(A) > 0$ (the singular values of A). The pair of linear vector spaces (S_1, S_2) where S_1 is the linear span of $\{u_1, \dots, u_r\}$ and S_2 is the linear span of $\{v_1, \dots, v_r\}$ will be called the *support* of A. We will denote by $S_j^\perp$ the orthogonal complements of S_j, $j = 1, 2$, and by P_S the projector on the linear vector subspace S of $\mathbb{R}^{m_j}$, $j = 1, 2$.

The Schatten-p (quasi-)norm $\|A\|_p$ of matrix A is defined by

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

Recall the well-known *trace duality* property:

$$\left|\operatorname{tr}(A^\top B)\right| \le \|A\|_1 \|B\|_\infty, \quad \forall A, B \in \mathbb{R}^{m_1\times m_2}.$$

We will also use the fact that the subdifferential of the convex function $A \mapsto \|A\|_1$ is the following set of matrices:

$$\partial\|A\|_1 = \Big\{\sum_{j=1}^{r} u_j v_j^\top + P_{S_1^\perp} W P_{S_2^\perp} : \ \|W\|_\infty \le 1\Big\} \tag{2.1}$$

(cf. [27]). Define the random matrix

$$\mathbf{M} = \frac{1}{n}\sum_{i=1}^{n}(Y_i X_i - \mathbb{E}(Y_i X_i)). \tag{2.2}$$

We will need the following assumption on the distribution of the matrices X_i.

Assumption 1. *There exists a constant $\mu > 0$ such that, for all matrices $A \in \mathbb{A} - \mathbb{A} := \{A_1 - A_2 : A_1, A_2 \in \mathbb{A}\}$,*

$$\|A\|_{L_2(\Pi)}^2 \ge \mu^{-2}\|A\|_2^2.$$

As discussed in the Introduction, Assumption 1 is satisfied, often with equality and for $\mathbb{A} = \mathbb{A} - \mathbb{A} = \mathbb{R}^{m_1\times m_2}$, in several interesting examples. The next theorem plays the key role in what follows.

Theorem 1. *Let $\mathbb{A} \subseteq \mathbb{R}^{m_1\times m_2}$ be any set of matrices. If $\lambda \ge 2\|\mathbf{M}\|_\infty$, then*

$$\|\hat{A}^\lambda - A_0\|_{L_2(\Pi)}^2 \le \inf_{A\in\mathbb{A}}\Big[\|A - A_0\|_{L_2(\Pi)}^2 + 2\lambda\|A\|_1\Big]. \tag{2.3}$$

If, in addition, $\mathbb{A}$ is a convex set and Assumption 1 is satisfied, then

$$\|\hat{A}^\lambda - A_0\|_{L_2(\Pi)}^2 \le \inf_{A\in\mathbb{A}}\Big[\|A - A_0\|_{L_2(\Pi)}^2 + \left(\frac{1+\sqrt{2}}{2}\right)^2 \mu^2\lambda^2 \mathrm{rank}(A)\Big]. \tag{2.4}$$

Furthermore, in this case for all $A \in \mathbb{A}$ with support (S_1, S_2),

$$\begin{aligned} &\|\hat{A}^\lambda - A_0\|_{L_2(\Pi)}^2 + (\lambda - 2\|\mathbf{M}\|_\infty)\|P_{S_1^\perp}\hat{A}^\lambda P_{S_2^\perp}\|_1 \\ &\quad \le \|A - A_0\|_{L_2(\Pi)}^2 + \left(\frac{1+\sqrt{2}}{2}\right)^2 \mu^2\lambda^2 \mathrm{rank}(A). \end{aligned} \tag{2.5}$$

PROOF. It follows from the definition of the estimator $\hat{A}$ that, for all $A \in \mathbb{A}$,

$$L_n(\hat{A}^\lambda) = \|\hat{A}^\lambda\|^2_{L_2(\Pi)} - \left\langle \frac{2}{n}\sum_{i=1}^n Y_i X_i, \hat{A}^\lambda \right\rangle + \lambda\|\hat{A}^\lambda\|_1 \leq$$

$$\|A\|^2_{L_2(\Pi)} - \left\langle \frac{2}{n}\sum_{i=1}^n Y_i X_i, A \right\rangle + \lambda\|A\|_1 = L_n(A).$$

Also, note that

$$\frac{1}{n}\sum_{i=1}^n \mathbb{E}(Y_i X_i) = \frac{1}{n}\sum_{i=1}^n \mathbb{E}\big(\langle A_0, X_i\rangle X_i\big) \text{ and } \frac{1}{n}\sum_{i=1}^n \langle \mathbb{E}(Y_i X_i), A\rangle = \langle A_0, A\rangle_{L_2(\Pi)}.$$

Therefore, we have

$$\|\hat{A}^\lambda\|^2_{L_2(\Pi)} - 2\langle \hat{A}^\lambda, A_0\rangle_{L_2(\Pi)} \leq \|A\|^2_{L_2(\Pi)} - 2\langle A, A_0\rangle_{L_2(\Pi)} +$$

$$\left\langle \frac{2}{n}\sum_{i=1}^n (Y_i X_i - \mathbb{E}(Y_i X_i)), \hat{A}^\lambda - A \right\rangle + \lambda(\|A\|_1 - \|\hat{A}^\lambda\|_1),$$

which implies, due to the trace duality,

$$\|\hat{A}^\lambda - A_0\|^2_{L_2(\Pi)} \leq \|A - A_0\|^2_{L_2(\Pi)} + 2\Delta\|\hat{A}^\lambda - A\|_1 + \lambda(\|A\|_1 - \|\hat{A}^\lambda\|_1),$$

where we set for brevity $\Delta = \|\mathbf{M}\|_\infty$. Under the assumption $\lambda \geq 2\Delta$ this yields

$$\|\hat{A}^\lambda - A_0\|^2_{L_2(\Pi)} \leq \|A - A_0\|^2_{L_2(\Pi)} + \lambda(\|\hat{A}^\lambda - A\|_1 + \|A\|_1 - \|\hat{A}^\lambda\|_1) \leq \|A - A_0\|^2_{L_2(\Pi)} + 2\lambda\|A\|_1,$$

and the bound (2.3) follows.

To prove the remaining bounds, note that a necessary condition of extremum in problem (1.5) implies that there exists $\hat{V} \in \partial\|\hat{A}^\lambda\|_1$ such that, for all $A \in \mathbb{A}$,

$$2\langle \hat{A}^\lambda, \hat{A}^\lambda - A\rangle_{L_2(\Pi)} - \left\langle \frac{2}{n}\sum_{i=1}^n Y_i X_i, \hat{A}^\lambda - A \right\rangle + \lambda\langle \hat{V}, \hat{A}^\lambda - A\rangle \leq 0. \tag{2.6}$$

Indeed, since $\hat{A}^\lambda$ is a minimizer of $L_n(A)$ in $\mathbb{A}$, there exists a matrix $B \in \partial L_n(\hat{A}^\lambda)$ such that $-B$ belongs to the normal cone of $\mathbb{A}$ at the point $\hat{A}^\lambda$. It is easy to see that B can be represented as follows

$$B = 2\int_{\mathbb{R}^{m_1\times m_2}} \langle \hat{A}^\lambda, X\rangle X \Pi(dX) - \frac{2}{n}\sum_{i=1}^n Y_i X_i + \lambda\hat{V},$$

where $\hat{V} \in \partial\|\hat{A}^\lambda\|_1$. The condition that $-B$ belongs to the normal cone at the point $\hat{A}^\lambda$ implies that $\langle B, \hat{A}^\lambda - A\rangle \le 0$, and (2.6) follows.

Consider an arbitrary $A \in \mathbb{A}$ of rank r with spectral representation $A = \sum_{j=1}^r \sigma_j u_j v_j^\top$ and with support (S_1, S_2). It follows from (2.6) that

$$2\langle \hat{A}^\lambda - A_0, \hat{A}^\lambda - A\rangle_{L_2(\Pi)} + \lambda\langle \hat{V} - V, \hat{A}^\lambda - A\rangle \le -\lambda\langle V, \hat{A}^\lambda - A\rangle + 2\langle \mathbf{M}, \hat{A}^\lambda - A\rangle \tag{2.7}$$

for an arbitrary $V \in \partial\|A\|_1$. By monotonicity of subdifferentials of convex functions, $\langle \hat{V} - V, \hat{A}^\lambda - A\rangle \ge 0$. On the other hand, by (2.1), the following representation holds

$$V = \sum_{j=1}^r u_j v_j^\top + P_{S_1^\perp} W P_{S_2^\perp},$$

where W is an arbitrary matrix with $\|W\|_\infty \le 1$. It follows from the trace duality that there exists W with $\|W\|_\infty \le 1$ such that

$$\langle P_{S_1^\perp} W P_{S_2^\perp}, \hat{A}^\lambda - A\rangle = \langle P_{S_1^\perp} W P_{S_2^\perp}, \hat{A}^\lambda\rangle = \langle W, P_{S_1^\perp} \hat{A}^\lambda P_{S_2^\perp}\rangle = \|P_{S_1^\perp} \hat{A}^\lambda P_{S_2^\perp}\|_1,$$

where in the first equality we used that A has the support (S_1, S_2). For this particular choice of W, (2.7) implies that

$$2\langle \hat{A}^\lambda - A_0, \hat{A}^\lambda - A\rangle_{L_2(\Pi)} + \lambda\|P_{S_1^\perp} \hat{A}^\lambda P_{S_2^\perp}\|_1 \le -\lambda\left\langle \sum_{j=1}^r u_j v_j^\top, \hat{A}^\lambda - A\right\rangle + 2\langle \mathbf{M}, \hat{A}^\lambda - A\rangle. \tag{2.8}$$

Using the identity

$$2\langle \hat{A}^\lambda - A_0, \hat{A}^\lambda - A\rangle_{L_2(\Pi)} = \|\hat{A}^\lambda - A_0\|^2_{L_2(\Pi)} + \|\hat{A}^\lambda - A\|^2_{L_2(\Pi)} - \|A - A_0\|^2_{L_2(\Pi)} \tag{2.9}$$

and the facts that

$$\|\sum_{j=1}^r u_j v_j^\top\|_\infty = 1, \quad \left\langle \sum_{j=1}^r u_j v_j^\top, \hat{A}^\lambda - A\right\rangle = \left\langle \sum_{j=1}^r u_j v_j^\top, P_{S_1}(\hat{A}^\lambda - A)P_{S_2}\right\rangle \tag{2.10}$$

we deduce from (2.8) that

$$\begin{aligned} &\|\hat{A}^\lambda - A_0\|^2_{L_2(\Pi)} + \|\hat{A}^\lambda - A\|^2_{L_2(\Pi)} + \lambda\|P_{S_1^\perp} \hat{A}^\lambda P_{S_2^\perp}\|_1 \le \\ &\|A - A_0\|^2_{L_2(\Pi)} + \lambda\|P_{S_1}(\hat{A}^\lambda - A)P_{S_2}\|_1 + 2\langle \mathbf{M}, \hat{A}^\lambda - A\rangle. \end{aligned} \tag{2.11}$$

To provide an upper bound on $2\langle \mathbf{M}, \hat{A}^\lambda - A\rangle$ we use the following decomposition

$$\begin{aligned} \langle \mathbf{M}, \hat{A}^\lambda - A\rangle &= \langle \mathcal{P}_A(\mathbf{M}), \hat{A}^\lambda - A\rangle + \langle P_{S_1^\perp} \mathbf{M} P_{S_2^\perp}, \hat{A}^\lambda - A\rangle \\ &= \langle \mathcal{P}_A(\mathbf{M}), \mathcal{P}_A(\hat{A}^\lambda - A)\rangle + \langle P_{S_1^\perp} \mathbf{M} P_{S_2^\perp}, \hat{A}^\lambda\rangle, \end{aligned}$$

where $\mathcal{P}_A(\mathbf{M}) = \mathbf{M} - P_{S_1^\perp}\mathbf{M}P_{S_2^\perp}$. This implies, due to the trace duality,

$$2\big|\langle \mathbf{M}, \hat{A}^\lambda - A\rangle\big| \leq \Lambda\|\mathcal{P}_A(\hat{A}^\lambda - A)\|_2 + \Gamma\|P_{S_1^\perp}\hat{A}^\lambda P_{S_2^\perp}\|_1 \tag{2.12}$$

$$\leq \Lambda\|\hat{A}^\lambda - A\|_2 + \Gamma\|P_{S_1^\perp}\hat{A}^\lambda P_{S_2^\perp}\|_1, \tag{2.13}$$

where

$$\Lambda = 2\|\mathcal{P}_A(\mathbf{M})\|_2, \quad \Gamma = 2\|P_{S_1^\perp}\mathbf{M}P_{S_2^\perp}\|_\infty. \tag{2.14}$$

Note that

$$\Gamma \leq 2\|\mathbf{M}\|_\infty = 2\Delta. \tag{2.15}$$

Since

$$\mathcal{P}_A(\mathbf{M}) = P_{S_1^\perp}\mathbf{M}P_{S_2} + P_{S_1}\mathbf{M} \tag{2.16}$$

and $\mathrm{rank}(P_{S_j}) \leq \mathrm{rank}(A)$, $j = 1, 2$, we have

$$\Lambda \leq 2\sqrt{\mathrm{rank}(\mathcal{P}_A(\mathbf{M}))}\,\|\mathbf{M}\|_\infty \leq 2\sqrt{2\,\mathrm{rank}(A)}\,\Delta \leq \sqrt{2\,\mathrm{rank}(A)}\,\lambda.$$

Due to the fact that

$$\|P_{S_1}(\hat{A}^\lambda - A)P_{S_2}\|_1 \leq \sqrt{\mathrm{rank}(A)}\|P_{S_1}(\hat{A}^\lambda - A)P_{S_2}\|_2 \leq \sqrt{\mathrm{rank}(A)}\|\hat{A}^\lambda - A\|_2$$

and to the Assumption 1, it follows from (2.11) and (2.13) that

$$\begin{aligned} &\|\hat{A}^\lambda - A_0\|^2_{L_2(\Pi)} + \|\hat{A}^\lambda - A\|^2_{L_2(\Pi)} + \lambda\|P_{S_1^\perp}\hat{A}^\lambda P_{S_2^\perp}\|_1 \\ &\quad \leq \|A - A_0\|^2_{L_2(\Pi)} + \mu(\lambda\sqrt{\mathrm{rank}(A)} + \Lambda)\|\hat{A}^\lambda - A\|_{L_2(\Pi)} \\ &\qquad + \Gamma\|P_{S_1^\perp}\hat{A}^\lambda P_{S_2^\perp}\|_1. \end{aligned} \tag{2.17}$$

Using the above bounds on Λ and Γ, we obtain from (2.17) that

$$\begin{aligned} &\|\hat{A}^\lambda - A_0\|^2_{L_2(\Pi)} + \|\hat{A}^\lambda - A\|^2_{L_2(\Pi)} + (\lambda - 2\Delta)\|P_{S_1^\perp}\hat{A}^\lambda P_{S_2^\perp}\|_1 \\ &\quad \leq \|A - A_0\|^2_{L_2(\Pi)} + (1+\sqrt{2})\mu\lambda\sqrt{\mathrm{rank}(A)}\|\hat{A}^\lambda - A\|_{L_2(\Pi)} \end{aligned}$$

which implies

$$\begin{aligned} &\|\hat{A}^\lambda - A_0\|^2_{L_2(\Pi)} + (\lambda - 2\Delta)\|P_{S_1^\perp}\hat{A}^\lambda P_{S_2^\perp}\|_1 \\ &\quad \leq \|A - A_0\|^2_{L_2(\Pi)} + \frac{1}{4}(1+\sqrt{2})^2\mu^2\lambda^2\mathrm{rank}(A). \end{aligned}$$

□

The following immediate corollary of Theorem 1 provides a bound for the Frobenius error.

Corollary 1. *Let $\mathbb{A}$ be a convex subset of $m_1 \times m_2$ matrices containing A_0, and let Assumption 1 be satisfied. If $\lambda \geq 2\|\mathbf{M}\|_\infty$, then*

$$\|\hat{A}^\lambda - A_0\|_2^2 \leq \lambda\mu^2 \min\left\{2\|A_0\|_1,\ \left(\frac{1+\sqrt{2}}{2}\right)^2 \lambda\mu^2\,\mathrm{rank}(A_0)\right\}. \tag{2.18}$$

Next we consider a version of Theorem 1 under weaker assumptions than Assumption 1 that are akin to Restricted Eigenvalue condition in sparse estimation of vectors. For simplicity, we will do it only when the domain $\mathbb{A}$ of minimization in (1.5) is a linear subspace of $\mathbb{R}^{m_1\times m_2}$. Recall that, given $A \in \mathbb{A}$ with support (S_1, S_2), we denote

$$\mathcal{P}_A(B) := B - P_{S_1^\perp} B P_{S_2^\perp},\ \mathcal{P}_A^\perp(B) := P_{S_1^\perp} B P_{S_2^\perp},\ B \in \mathbb{R}^{m_1\times m_2},$$

and, for $c_0 \geq 0$, define the following cone of matrices:

$$\mathbb{C}_{A,c_0} := \Big\{B \in \mathbb{A} : \|\mathcal{P}_A^\perp(B)\|_1 \leq c_0\|\mathcal{P}_A(B)\|_1\Big\}.$$

Finally, define

$$\mu_{c_0}(A) := \inf\Big\{\mu > 0 : \|\mathcal{P}_A(B)\|_2 \leq \mu\|B\|_{L_2(\Pi)},\ \forall B \in \mathbb{C}_{A,c_0}\Big\}.$$

Note that $\mu_{c_0}(A)$ is a nondecreasing function of c_0. For $c_0 = +\infty$, the quantity $\mu_\infty(A)$ has a simple meaning: it is equal to the norm of the linear transformation $B \mapsto \mathcal{P}_A(B)$ from the space $\mathbb{A}$ equipped with the $L_2(\Pi)$-norm into the space of all matrices equipped with the Frobenius norm. For $c_0 = 0$, $\mu_0(A)$ is the norm of the same linear transformation restricted to the subspace of $\mathbb{A}$ consisting of all matrices $B \in \mathbb{A}$ with $\mathcal{P}_A^\perp(B) = 0$. We are more interested in the intermediate values, $c_0 \in (0, +\infty)$. In this case, $\mu_{c_0}(A)$ is the "norm" of the linear mapping $\mathcal{P}_A$ restricted to the cone of matrices B for which $\mathcal{P}_A(B)$ is the "dominant" part and $\mathcal{P}_A^\perp(B)$ is "small". Note that the rank of $\mathcal{P}_A(B)$ is not larger than 2rank(A), so, when the rank of A is small, the matrices in $\mathbb{C}_{A,c_0}$ are approximately "low-rank". The quantities of the same flavor have been previously used in the literature on Lasso, Dantzig selector and other methods of sparse estimation of vectors. In these problems, they can be expressed in terms of "restricted eigenvalues" of certain Gram matrices, cf. the Restricted Eigenvalue condition in [6] for the fixed design case and similar distribution dependent conditions in [16] for the random design case. Such conditions are also considered in [20] for the matrix case. In what follows, we use the value $c_0 = 5$ and set $\mu(A) := \mu_5(A)$.

THEOREM 2. *Let $\mathbb{A}$ be a linear subspace of $\mathbb{R}^{m_1\times m_2}$. If $\lambda \geq 3\|\mathbf{M}\|_\infty$, then*

$$\|\hat{A}^\lambda - A_0\|^2_{L_2(\Pi)} \leq \inf_{A\in\mathbb{A}}\Big[\|A-A_0\|^2_{L_2(\Pi)} + \lambda^2\mu^2(A)\mathrm{rank}(A)\Big]. \tag{2.19}$$

PROOF. Fix $A\in\mathbb{A}$ with support (S_1,S_2). If $\langle \hat{A}^\lambda - A_0, \hat{A}^\lambda - A\rangle_{L_2(\Pi)} \leq 0$, then we trivially have $\|\hat{A}^\lambda - A_0\|^2_{L_2(\Pi)} \leq \|A-A_0\|^2_{L_2(\Pi)}$ in view of (2.9). Thus, assume that $\langle \hat{A}^\lambda - A_0, \hat{A}^\lambda - A\rangle_{L_2(\Pi)} > 0$. In this case, (2.8) and an obvious modification of (2.10) imply

$$\lambda\|P_{S_1^\perp}\hat{A}^\lambda P_{S_2^\perp}\|_1 \leq \lambda\|\mathcal{P}_A(\hat{A}^\lambda - A)\|_1 + 2\langle \mathbf{M}, \hat{A}^\lambda - A\rangle. \tag{2.20}$$

Now,

$$\begin{aligned}\langle \mathbf{M}, \hat{A}^\lambda - A\rangle &= \langle \mathbf{M}, \mathcal{P}_A(\hat{A}^\lambda - A)\rangle + \langle \mathbf{M}, \mathcal{P}_A^\perp(\hat{A}^\lambda - A)\rangle \\ &\leq \|\mathbf{M}\|_\infty\left(\|\mathcal{P}_A(\hat{A}^\lambda - A)\|_1 + \|\mathcal{P}_A^\perp(\hat{A}^\lambda - A)\|_1\right).\end{aligned} \tag{2.21}$$

By (2.20) and (2.21),

$$(\lambda - 2\Delta)\|\mathcal{P}_A^\perp(\hat{A}^\lambda - A)\|_1 \leq (\lambda+2\Delta)\|\mathcal{P}_A(\hat{A}^\lambda - A)\|_1. \tag{2.22}$$

For $\lambda \geq 3\Delta$, this yields

$$\|\mathcal{P}_A^\perp(\hat{A}^\lambda - A)\|_1 \leq 5\|\mathcal{P}_A(\hat{A}^\lambda - A)\|_1,$$

which implies that $\hat{A}^\lambda - A \in \mathbb{C}_{A,5}$, and thus $\|\mathcal{P}_A(\hat{A}^\lambda - A)\|_2 \leq \mu(A)\|\hat{A}^\lambda - A\|_{L_2(\Pi)}$. Combining this inequality with (2.11), (2.12), (2.14), (2.15) and using that $\lambda \geq 3\Delta$, after some algebra we get

$$\begin{aligned}&\|\hat{A}^\lambda - A_0\|^2_{L_2(\Pi)} + \|\hat{A}^\lambda - A\|^2_{L_2(\Pi)} + (\lambda/3)\|P_{S_1^\perp}\hat{A}^\lambda P_{S_2^\perp}\|_1 \\ &\quad\leq \|A-A_0\|^2_{L_2(\Pi)} + (1+2\sqrt{2}/3)\mu(A)\lambda\sqrt{\mathrm{rank}(A)}\,\|\hat{A}^\lambda - A\|_{L_2(\Pi)} \\ &\quad\leq \|A-A_0\|^2_{L_2(\Pi)} + \|\hat{A}^\lambda - A\|^2_{L_2(\Pi)} + \mu^2(A)\lambda^2\mathrm{rank}(A).\end{aligned}$$

□

As a simple example, consider the case when $m_1 = m_2$, $\mathbb{A}$ is the space of all diagonal matrices, and X_i also belong to $\mathbb{A}$. Then the trace regression model (1.2) becomes the usual linear regression model. The Schatten p-norms are in this case equivalent to the ℓ_p-norms with the operator norm $\|\cdot\|_\infty$ being the ℓ_∞-norm and the rank of matrix A characterizing the sparsity of the corresponding vector. The problem of minimizing the functional $L_n(A)$ over the space $\mathbb{A}$ is a Lasso-type penalized empirical risk minimization. In particular, it coincides with the standard Lasso if all X_i are non-random. Inequalities of Theorem 1 and (2.19) become, in this case, sparsity oracle inequalities for the Lasso-type estimators. It is noteworthy that these inequalities are sharp (i.e., with leading constant 1), which was not achieved in the past work. The random matrix $\mathbf{M}$ is also diagonal and its norm $\|\mathbf{M}\|_\infty$ is just the ℓ_∞-norm of the corresponding random vector, which is the sum of independent random vectors. Hence, it is easy to provide probabilistic bounds on $\|\mathbf{M}\|_\infty$ using, for instance, the classical Bernstein inequality and the union bound. We give an example of such an application of Theorem 2 in Section 5.4.

3. Upper bounds for matrix completion. In this section we consider implications of the general oracle inequalities of Theorem 1 for the model of USR matrix completion. Thus, we assume that the matrices X_i are i.i.d. uniformly distributed in the matrix completion basis $\mathcal{X}$, which implies that $\|A\|^2_{L_2(\Pi)} = (m_1m_2)^{-1}\|A\|_2^2$ for all matrices $A \in \mathbb{R}^{m_1\times m_2}$, and we set $\mu = \sqrt{m_1m_2}$. The estimator $\hat{A}^\lambda$ is then defined by (here and further on we set $\mathbb{A} = \mathbb{R}^{m_1\times m_2}$ in the case of matrix completion):

$$\begin{aligned} \hat{A}^\lambda &= \operatorname{argmin}_{A\in\mathbb{R}^{m_1\times m_2}} \left(\frac{1}{m_1m_2}\|A\|_2^2 - \left\langle \frac{2}{n}\sum_{i=1}^n Y_iX_i, A\right\rangle + \lambda\|A\|_1 \right) \\ &= \operatorname{argmin}_{A\in\mathbb{R}^{m_1\times m_2}} \Big(\|A-\mathbf{X}\|_2^2 + \lambda m_1m_2\|A\|_1 \Big), \end{aligned} \tag{3.1}$$

where

$$\mathbf{X} = \frac{m_1m_2}{n}\sum_{i=1}^n Y_iX_i.$$

We can also write $\hat{A}^\lambda$ explicitly:

$$\hat{A}^\lambda = \sum_j \big(\sigma_j(\mathbf{X}) - \lambda m_1m_2/2\big)_+ u_j(\mathbf{X})v_j(\mathbf{X})^\top \tag{3.2}$$

where $x_+ = \max\{x, 0\}$, $\sigma_j(\mathbf{X})$ are the singular values and $u_j(\mathbf{X}), v_j(\mathbf{X})$ are the left and right singular vectors of $\mathbf{X} = \sum_{j=1}^{\mathrm{rank}(\mathbf{X})} \sigma_j(\mathbf{X})u_j(\mathbf{X})v_j(\mathbf{X})^\top$. Thus, the computation of $\hat{A}^\lambda$ is simple; it reduces to soft thresholding of singular values in the SVD of $\mathbf{X}$. To see why (3.2) gives the solution of (3.1), note that, in view of (2.1), the subdifferential of $F(A) = \|A-\mathbf{X}\|_2^2 + \lambda m_1m_2\|A\|_1$ is the set of matrices

$$\partial F(A) = \left\{ 2(A-\mathbf{X}) + \lambda m_1m_2 \left(\sum_{j=1}^r u_jv_j^\top + P_{S_1^\perp}WP_{S_2^\perp} \right) : \ \|W\|_\infty \le 1 \right\},$$

where r, u_j, v_j, S_1, S_2 correspond to the SVD of A. Since $A \mapsto F(A)$ is strictly convex, the minimizer $\hat{A}^\lambda$ is unique, and the condition $\mathbf{0} \in \partial F(\hat{A}^\lambda)$ is necessary and sufficient characterization of the minimum, where $\mathbf{0}$ is the zero $m_1 \times m_2$ matrix. Considering

$$W = \sum_{j:\, \sigma_j(\mathbf{X}) < \lambda m_1 m_2/2} \Big(\frac{2\sigma_j(\mathbf{X})}{\lambda m_1 m_2} - 1\Big) u_j(\mathbf{X}) v_j(\mathbf{X})^\top,$$

it is easy to check that (3.2) satisfies this condition.

In view of Theorem 1, to get the oracle inequalities in a closed form it remains only to specify the value of regularization parameter λ such that $\lambda \geq 2\|\mathbf{M}\|_\infty$ with high probability. This requires some assumptions on the distribution of (X_i, Y_i), and the value of λ will be different under different assumptions. We will consider only the following two cases of particular interest.

- *Sub-exponential noise and matrices with uniformly bounded entries.* There exist constants $\sigma, c_1 > 0$, $\alpha \geq 1$ and $\tilde{c}$ such that

$$\max_{i=1,\dots,n} \mathbb{E} \exp\left(\frac{|\xi_i|^\alpha}{\sigma^\alpha}\right) < \tilde{c}, \quad \mathbb{E}\xi_i^2 \geq c_1 \sigma^2, \ \forall 1 \leq i \leq n, \tag{3.3}$$

  and $\max_{i,j} |a_0(i,j)| \leq a$ for some constant a.
- *Statistical learning setting.* There exists a constant η such that $\max_{i=1,\dots,n} |Y_i| \leq \eta$ almost surely.

In both cases, we obtain the upper bounds for $\|\mathbf{M}\|_\infty$ (that we call the *stochastic error*) using the non-commutative Bernstein inequalities, cf. Section 6. The resulting values of λ and the corresponding oracle inequalities are given in the next two theorems.

Set $m = m_1 + m_2$. In what follows, we will denote by C absolute positive constants, possibly different on different occasions.

THEOREM 3. *Let X_i be i.i.d. uniformly distributed on $\mathcal{X}$, and the pairs (X_i, Y_i) be i.i.d. Assume that $\max_{i,j} |a_0(i,j)| \leq a$ for some constant a, and that condition (3.3) holds. For $t > 0$, consider the regularization parameter λ satisfying*

$$\lambda \geq C(\sigma \vee a) \max\left\{ \sqrt{\frac{t + \log(m)}{(m_1 \wedge m_2) n}}\,,\ \frac{(t + \log(m)) \log^{1/\alpha}(m_1 \wedge m_2)}{n} \right\}, \tag{3.4}$$

where $C > 0$ is a large enough constant that can depend only on $\alpha, c_1, \tilde{c}$. Then with probability at least $1 - 3e^{-t}$ we have

(3.5)
$$\|\hat{A}^\lambda - A_0\|_2^2 \le \|A - A_0\|_2^2 + m_1 m_2 \min\left\{2\lambda\|A\|_1, \left(\frac{1+\sqrt{2}}{2}\right)^2 m_1 m_2 \lambda^2 \operatorname{rank}(A)\right\}$$

for all $A \in \mathbb{R}^{m_1 \times m_2}$.

THEOREM 4. *Let X_i be i.i.d. uniformly distributed on $\mathcal{X}$. Assume that $\max_{i=1,\dots,n} |Y_i| \le \eta$ almost surely for some constant η. For $t > 0$ consider the regularization parameter λ satisfying*

$$\lambda \ge 4\eta \max\left\{\sqrt{\frac{t + \log(m)}{(m_1 \wedge m_2)n}}, \frac{2(t + \log(m))}{n}\right\}. \tag{3.6}$$

Then with probability at least $1 - e^{-t}$ inequality (3.5) holds for all $A \in \mathbb{R}^{m_1 \times m_2}$.

Theorems 3 and 4 follow immediately from Theorem 1 and Lemmas 1, 2 and 3 with $\mu = \sqrt{m_1 m_2}$.

Note that the natural choice of t in Theorems 3 and 4 is of the order $\log(m)$, since a larger t leads to slower rate of convergence and a smaller t does not improve the rate but makes the concentration probability smaller. Note also that, under this choice of t, the second terms under the maxima in (3.4) and (3.6) are negligible for the values of n, m_1, m_2 such that the term containing $\operatorname{rank}(A_0)$ in (3.5) is meaningful. Indeed, if t is of the order $\log(m)$, the condition that $m_1 m_2 \lambda^2 \ll 1$ necessarily implies $n \gg (m_1 \vee m_2)\log(m)$. On the other hand, the negligibility of the second terms under the maxima in (3.4) and (3.6) is approximately equivalent to $n > (m_1 \wedge m_2)\log^{1+2/\alpha}(m)$ and $n > (m_1 \wedge m_2)\log(m)$ respectively. Based on these remarks, we can choose λ in the form

$$\lambda = C_* c_* \sqrt{\frac{\log(m)}{(m_1 \wedge m_2)n}}, \tag{3.7}$$

where c_* equals either $\sigma \vee a$ or η and the constant $C_* > 0$ is large enough, and we can state the following corollary that will be further useful for minimax considerations. Define $\tau > 0$ by

$$\tau^2 = \left(\frac{1+\sqrt{2}}{2}\right)^2 C_*^2 c_*^2 \frac{M \log(m)}{n},$$

where $M = \max(m_1, m_2)$, and $m = m_1 + m_2$.

COROLLARY 2. *Let one of the sets of conditions (i) or (ii) below be satisfied:*

(i) The assumptions of Theorem 3 with λ *as in (3.7),* $n > (m_1 \wedge m_2)\log^{1+2/\alpha}(m)$, $c_* = \sigma \vee a$, *and a large enough constant* $C_* > 0$ *that can depend only on* $\alpha, c_1, \tilde{c}$.

(ii) The assumptions of Theorem 4 with $n > 4(m_1 \wedge m_2)\log(m)$, λ *as in (3.7),* $c_* = \eta$, *and* $C_* = 4$.

Then, with probability at least $1 - 3/(m_1 + m_2)$,

$$\frac{1}{m_1 m_2}\|\hat{A}^\lambda - A_0\|_2^2 \le \min_{A\in\mathbb{R}^{m_1\times m_2}} \left(\frac{1}{m_1 m_2}\|A - A_0\|_2^2 + \tau^2 \mathrm{rank}(A)\right), \tag{3.8}$$

and, in particular,

$$\frac{1}{m_1 m_2}\|\hat{A}^\lambda - A_0\|_2^2 \le \left(\frac{1+\sqrt{2}}{2}\right)^2 C_*^2 c_*^2 \log(m) \frac{M\,\mathrm{rank}(A_0)}{n}, \tag{3.9}$$

where $M = \max(m_1, m_2)$, *and* $m = m_1 + m_2$. *Furthermore, with the same probability,*

$$\frac{1}{m_1 m_2}\|\hat{A}^\lambda - A_0\|_2^2 \le \sum_{j=1}^{\mathrm{rank}(A_0)} \min\left\{\tau^2,\ \frac{\sigma_j^2(A_0)}{m_1 m_2}\right\} \le \inf_{0<q\le 2} \frac{\tau^{2-q}\|A_0\|_q^q}{(m_1 m_2)^{q/2}}. \tag{3.10}$$

PROOF. Inequalities (3.8) and (3.9) are straightforward in view of Theorems 3 and 4. To prove (3.10) it suffices to note that, for any $\kappa > 0$, $0 < q \le 2$,

$$\begin{aligned}\min_A \left(\|A - A_0\|_2^2 + \kappa^2\,\mathrm{rank}(A)\right) &= \sum_j \min\{\kappa^2, \sigma_j^2(A_0)\} = \kappa^2 \sum_j \min\left\{1, \left(\frac{\sigma_j(A_0)}{\kappa}\right)^2\right\} \\ &\le \kappa^2 \sum_j \min\left\{1, \left(\frac{\sigma_j(A_0)}{\kappa}\right)^q\right\} \le \kappa^{2-q}\|A_0\|_q^q.\end{aligned}$$

□

Inequality (3.9) guarantees that the normalized Frobenius error $(m_1 m_2)^{-1}\|\hat{A}^\lambda - A_0\|_2^2$ of the estimator $\hat{A}^\lambda$ is small whenever $n > C(m_1 \vee m_2)\log(m)\mathrm{rank}(A_0)$ with a large enough $C > 0$. This quantifies the sample size n necessary for successful matrix completion from noisy data.

Keshavan et al. [15], Theorem 1.1, under a sampling scheme different from ours (sampling without replacement) and sub-gaussian errors, proposed an estimator $\hat{A}$ satisfying, with probability at least $1-(m_1\wedge m_2)^{-3}$,

$$\frac{1}{m_1 m_2}\|\hat{A}-A_0\|_2^2 \leqslant C\sqrt{\beta}\,\log(n)\,\frac{M\,\mathrm{rank}(A_0)}{n}\,, \tag{3.11}$$

where $C>0$ is a constant, and $\beta=(m_1\vee m_2)/(m_1\wedge m_2)$ is the aspect ratio. A drawback is that the construction of $\hat{A}$ in [15] requires the exact knowledge of $\mathrm{rank}(A_0)$. Furthermore, the bound (3.11) is suboptimal for "very rectangular" matrices, i.e., when $\beta \gg 1$.

4. Lower Bounds. In this section, we prove the minimax lower bounds showing that the rates attained by our estimator are optimal up to logarithmic factors. Note that here we cannot apply the lower bounds of Theorem 6 in [23] for USR matrix completion on the Schatten balls because they are achieved on matrices with entries, which are not uniformly bounded for $m_1 m_2 \gg n$.

We will need the following assumption, which is similar in spirit but, in general, substantially weaker than the usual Restricted Isometry condition.

Assumption 2. **(Restricted Isometry in Expectation.)** *For some $1\le r\le \min(m_1,m_2)$ and some $0<\mu<\infty$ that there exists a constant $\delta_r\in[0,1)$ such that*

$$(1-\delta_r)\|A\|_2 \le \mu\|A\|_{L_2(\Pi)} \le (1+\delta_r)\|A\|_2,$$

for all matrices $A\in\mathbb{R}^{m_1\times m_2}$ with rank at most r.

For the particular case of fixed X_i (cf. Example 4 in the Introduction), Assumption 2 coincides with the matrix version of scaled restricted isometry with scaling factor μ [23].

We will denote by $\inf_{\hat{A}}$ the infimum over all estimators $\hat{A}$ with values in $\mathbb{R}^{m_1\times m_2}$. For any integer $r\le\min(m_1,m_2)$ and any $a>0$ we consider the class of matrices

$$\mathcal{A}(r,a)=\left\{A_0\in\mathbb{R}^{m_1\times m_2}:\ \mathrm{rank}(A_0)\le r,\ \max_{i,j}|a_0(i,j)|\ \le\ a\right\}.$$

For any $A\in\mathbb{R}^{m_1\times m_2}$, let $\mathbb{P}_A$ denote the probability distribution of the observations $(X_1,Y_1,\dots,X_n,Y_n)$ with $\mathbb{E}(Y_i|X_i)=\langle A,X_i\rangle$. We set for brevity $M=\max(m_1,m_2)$.

THEOREM 5. *Fix $a > 0$ and an integer $1 \le r \le \min(m_1, m_2)$. Assume that $Mr \le n$, and that conditionally on X_i, the variables ξ_i are Gaussian $\mathcal{N}(0, \sigma^2)$, $\sigma^2 > 0$, for $i = 1, \dots, n$. Let Assumption 2 be satisfied with some $\mu > 0$. Then*

$$\inf_{\hat{A}} \sup_{A_0 \in \mathcal{A}(r,a)} \mathbb{P}_{A_0}\left(\|\hat{A} - A_0\|^2_{L_2(\Pi)} > c(\sigma \wedge a)^2 \frac{Mr}{n} \right) \ \ge \ \beta, \tag{4.1}$$

where $\beta \in (0,1)$ and $c > 0$ are absolute constants.

PROOF. Without loss of generality, assume that $M = \max(m_1, m_2) = m_1 \ge m_2$. For some constant $0 \le \gamma \le 1$ we define

$$\mathcal{C} \ = \left\{ \tilde{A} = (a_{ij}) \in \mathbb{R}^{m_1 \times r} : a_{ij} \in \left\{ 0, \gamma(\sigma \wedge a)\left(\frac{\mu^2 r}{m_2 n} \right)^{1/2} \right\}, \forall 1 \le i \le m_1,\, 1 \le j \le r \right\},$$

and consider the associated set of block matrices

$$\mathcal{B}(\mathcal{C}) \ = \ \left\{ A = (\ \tilde{A} \ | \ \cdots \ | \ \tilde{A} \ | \ O \ ) \in \mathbb{R}^{m_1 \times m_2} : \tilde{A} \in \mathcal{C} \right\},$$

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

By construction, any element of $\mathcal{B}(\mathcal{C})$ as well as the difference of any two elements of $\mathcal{B}(\mathcal{C})$ has rank at most r and the entries of any matrix in $\mathcal{B}(\mathcal{C})$ take values in $[0, a]$. Thus, $\mathcal{B}(\mathcal{C}) \subset \mathcal{A}(r, a)$. Due to the Varshamov-Gilbert bound (cf. Lemma 2.9 in [26]), there exists a subset $\mathcal{A}^0 \subset \mathcal{B}(\mathcal{C})$ with cardinality $\mathrm{Card}(\mathcal{A}^0) \ge 2^{rm_1/8} + 1$ containing the zero $m_1 \times m_2$ matrix $\mathbf{0}$ and such that, for any two distinct elements A_1 and A_2 of $\mathcal{A}^0$,

$$\|A_1 - A_2\|_2^2 \ge \frac{m_1 r}{8} \left(\gamma^2 (\sigma \wedge a)^2 \frac{\mu^2 r}{m_2 n} \right) \left\lfloor \frac{m_2}{r} \right\rfloor \ge \frac{\gamma^2}{16} (\sigma \wedge a)^2 \frac{\mu^2 m_1 r}{n} \, . \tag{4.2}$$

In view of Assumption 2, this implies

$$\|A_1 - A_2\|^2_{L_2(\Pi)} \ge (1 - \delta_r)^2 \frac{\gamma^2}{16} (\sigma \wedge a)^2 \frac{m_1 r}{n}. \tag{4.3}$$

Using that, conditionally on X_i, the distributions of ξ_i are Gaussian, we get that, for any $A \in \mathcal{A}_0$, the Kullback-Leibler divergence $K(\mathbb{P}_{\mathbf{0}}, \mathbb{P}_A)$ between $\mathbb{P}_{\mathbf{0}}$ and $\mathbb{P}_A$ satisfies

$$K(\mathbb{P}_{\mathbf{0}}, \mathbb{P}_A) \ = \ \frac{n}{2\sigma^2} \|A\|^2_{L_2(\Pi)} \le (1 + \delta_r)^2 \frac{\gamma^2}{2} m_1 r. \tag{4.4}$$

From (4.4) we deduce that the condition

$$\frac{1}{\mathrm{Card}(\mathcal{A}^0)-1}\sum_{A\in\mathcal{A}^0} K(\mathbb{P}_{\mathbf{0}},\mathbb{P}_A) \;\le\; \alpha\log\big(\mathrm{Card}(\mathcal{A}^0)-1\big) \tag{4.5}$$

is satisfied for any $\alpha>0$ if $\gamma>0$ is chosen as a sufficiently small numerical constant depending on α. In view of (4.3) and (4.5), the result now follows by application of Theorem 2.5 in [26]. □

In the USR matrix completion problem we have $\|A\|^2_{L_2(\Pi)} = (m_1m_2)^{-1}\|A\|_2^2$ for all matrices $A\in\mathbb{R}^{m_1\times m_2}$. Thus, the corresponding lower bound follows immediately from the previous theorem with $\delta_r=0$ and $\mu=\sqrt{m_1m_2}$.

THEOREM 6. *Let the assumptions of Theorem 5 be satisfied, and let the matrices X_i be i.i.d. uniformly distributed on $\mathcal{X}$. Then*

$$\inf_{\hat A}\sup_{A_0\in\mathcal{A}(r,a)} \mathbb{P}_{A_0}\left(\frac{1}{m_1m_2}\|\hat A-A_0\|_2^2 > c(\sigma\wedge a)^2\frac{Mr}{n}\right) \;\ge\; \beta, \tag{4.6}$$

where $\beta\in(0,1)$ and $c>0$ are absolute constants.

Comparing Theorem 6 with Corollary 2(i) we see that, in the case of Gaussian errors ξ_i, the rate of convergence of our estimator $\hat A^\lambda$ given in (3.9) is optimal (up to a logarithmic factor) in the minimax sense on the class of matrices $\mathcal{A}(r,a)$.

Similar conclusion can be obtained for the statistical learning setting. Indeed, assume that the pairs (X_i,Y_i) are i.i.d. realizations of a random pair (X,Y) with distribution P_{XY} belonging to the class

$$\mathcal{P}_{A_0,\eta} = \{P_{XY}:\ X\sim\Pi_0, |Y|\le\eta\,(\text{a.s.}),\ \mathbb{E}(Y|X)=\langle A_0,X\rangle\,\},$$

where Π_0 is the uniform distribution on $\mathcal{X}$, $1\le r\le\min(m_1,m_2)$ is an integer, and $\eta>0$.

THEOREM 7. *Let n,m_1,m_2,r be as in Theorem 5. Let (X_i,Y_i) be i.i.d. realizations of a random pair (X,Y) with distribution P_{XY}. Then*

$$\inf_{\hat A}\sup_{\mathrm{rank}(A_0)\le r}\sup_{P_{XY}\in\mathcal{P}_{A_0,\eta}} \mathbb{P}\left(\frac{1}{m_1m_2}\|\hat A-A_0\|_2^2 > c\eta^2\frac{Mr}{n}\right) \;\ge\; \beta, \tag{4.7}$$

where $\beta\in(0,1)$ and $c>0$ are absolute constants.

Proof. We act as in the proof of Theorem 5 with some modifications. Assuming that $M = \max(m_1, m_2) = m_1 \geq m_2$ and $0 \leq \gamma \leq 1/2$ we define the class of matrices

$$\mathcal{C}' = \left\{ \tilde{A} = (a_{ij}) \in \mathbb{R}^{m_1 \times r} : a_{ij} \in \left\{ 0, \gamma\eta\left(\frac{\mu^2 r}{m_2 n}\right)^{1/2} \right\}, \forall\, 1 \leq i \leq m_1,\, 1 \leq j \leq r \right\},$$

and take its block extension $\mathcal{B}(\mathcal{C}')$. Consider the joint distributions P_{XY} such that $X \sim \Pi_0$ and, conditionally on X, $Y = \eta$ with probability $p_{A_0}(X) = 1/2 + \langle A_0, X\rangle/(2\eta)$ and $Y = -\eta$ with probability $1 - p_{A_0}(X) = 1/2 - \langle A_0, X\rangle/(2\eta)$, where $A_0 \in \mathcal{B}(\mathcal{C}')$. It is easy to see that such distributions P_{XY} belong to the class $\mathcal{P}_{A_0,\eta}$, and our assumptions guarantee that $1/4 \leq p_{A_0}(X) \leq 3/4$, $\operatorname{rank}(A_0) \leq r$ for all $A_0 \in \mathcal{B}(\mathcal{C}')$. We will denote the corresponding n-product measure by $\mathbb{P}_{A_0}$. For any $A \in \mathcal{B}(\mathcal{C}')$, the Kullback-Leibler divergence between $\mathbb{P}_{\mathbf{0}}$ and $\mathbb{P}_A$ has the form

(4.8)
$$K\big(\mathbb{P}_{\mathbf{0}}, \mathbb{P}_A\big) = n\mathbb{E}\left(p_{\mathbf{0}}(X) \log\frac{p_{\mathbf{0}}(X)}{p_A(X)} + (1 - p_{\mathbf{0}}(X)) \log\frac{1 - p_{\mathbf{0}}(X)}{1 - p_A(X)} \right).$$

Using the inequality $-\log(1+u) \leq -u + u^2/2$, $\forall\, u > -1$, and the fact that $1/4 \leq p_A(X) \leq 3/4$, we find that the expression under the expectation in (4.8) is bounded by $2(p_{\mathbf{0}}(X) - p_A(X))^2$. This implies

$$K\big(\mathbb{P}_{\mathbf{0}}, \mathbb{P}_A\big) \leq \frac{n}{2\eta^2}\|A\|^2_{L_2(\Pi_0)}.$$

The remaining arguments are identical to those in the proof of Theorem 5.
□

5. Further results and examples.

5.1. *Recovery of the rank and specific lower bound.* A notable property of the estimator $\hat{A}^\lambda$ in matrix completion setting is that it has the same rank as the underlying matrix A_0 with probability close to 1. As a consequence we can establish a lower bound for the Frobenius error of $\hat{A}^\lambda$ with the rates matching up to constants the upper bounds of Corollary 2.

Theorem 8. *Let X_i be i.i.d. uniformly distributed on $\mathcal{X}$ and let λ satisfy the inequality $\lambda \geq 2\|\mathbf{M}\|_\infty$ (as in Theorem 1). Consider the estimator $\hat{A}^{\lambda'}$ with $\lambda' = \lambda/(1-\delta)$ for some $0 < \delta < 1$. Set $\hat{r} = \operatorname{rank}(\hat{A}^{\lambda'})$. Then*

$$\hat{r} \leq \operatorname{rank}(A_0). \tag{5.1}$$

If, in addition, $\min_{j:\,\sigma_j(A_0)\neq 0} \sigma_j(A_0) \geq \lambda' m_1 m_2$, *then*

$$\hat{r} \geq \operatorname{rank}(A_0), \tag{5.2}$$

and

$$\|\hat{A}^{\lambda'} - A_0\|_2^2 \geq \frac{\delta^2}{4(1-\delta)^2} \operatorname{rank}(A_0)\,(\lambda m_1 m_2)^2\,. \tag{5.3}$$

PROOF. Note that $\mathbf{X} - A_0 = m_1 m_2 \mathbf{M}$. Using standard matrix perturbation argument (cf. [24], page 203), we get, for all $j = 1, \dots, m_1 \wedge m_2$,

$$|\sigma_j(\mathbf{X}) - \sigma_j(A_0)| \leq \sigma_1(\mathbf{X} - A_0) = m_1 m_2 \|\mathbf{M}\|_\infty \leq \frac{\lambda m_1 m_2}{2} = (1-\delta)\frac{\lambda' m_1 m_2}{2}\,.$$

Since, by (3.2), $\sigma_{\hat{r}}(\mathbf{X}) > \lambda' m_1 m_2/2$, we find that $\sigma_{\hat{r}}(A_0) > \delta\lambda' m_1 m_2/2$. This implies (5.1). Now, if $\sigma_j(A_0) \geq \lambda' m_1 m_2$ we get

$$\sigma_j(\mathbf{X}) \geq \sigma_j(A_0) - |\sigma_j(\mathbf{X}) - \sigma_j(A_0)| \geq \lambda' m_1 m_2 - (1-\delta)\frac{\lambda' m_1 m_2}{2} > \frac{\lambda' m_1 m_2}{2}\,,$$

and thus (5.2) follows.

To prove (5.3), denote by $\mathcal{P} : \mathbb{R}^{m_1 \times m_2} \to \mathbb{R}^{m_1 \times m_2}$ the projector on the linear span of matrices $\big(u_j(\mathbf{X}) v_j(\mathbf{X})^\top, j = 1, \dots, r\big)$, where $r = \operatorname{rank}(A_0)$. We have $\|\hat{A}^{\lambda'} - A_0\|_2 \geq \|\mathcal{P}(\hat{A}^{\lambda'} - A_0)\|_2 \geq \|\mathcal{P}(\hat{A}^{\lambda'} - \mathbf{X})\|_2 - \|\mathcal{P}(\mathbf{X} - A_0)\|_2$. Here $\|\mathcal{P}(\hat{A}^{\lambda'} - \mathbf{X})\|_2 = \sqrt{r}\lambda' m_1 m_2/2$ in view of (3.2) and the fact that $\hat{r} = r$, cf. (5.1) and (5.2). On the other hand, $\|\mathcal{P}(\mathbf{X} - A_0)\|_2 \leq \sqrt{r}\|\mathbf{M}\|_\infty m_1 m_2 \leq \sqrt{r}\lambda m_1 m_2/2$. This implies

$$\|\hat{A}^{\lambda'} - A_0\|_2 \geq \sqrt{\hat{r}}\left(\frac{\lambda' m_1 m_2}{2} - (1-\delta)\frac{\lambda' m_1 m_2}{2}\right) = \delta\sqrt{\hat{r}}\frac{\lambda' m_1 m_2}{2}.$$

□

COROLLARY 3. *Let the assumptions of Corollary 2 be satisfied. Consider the estimator* $\hat{A}^{\lambda'}$ *with*

$$\lambda' = \frac{C_* c_*}{1-\delta}\sqrt{\frac{\log(m)}{(m_1 \wedge m_2) n}}$$

for some $0 < \delta < 1$. *Set* $\hat{r} = \operatorname{rank}(\hat{A}^{\lambda'})$. *Then* $\hat{r} \leq \operatorname{rank}(A_0)$ *with probability at least* $1 - 3/(m_1 + m_2)$. *If, in addition,*

$$\min_{j:\,\sigma_j(A_0)\neq 0} \sigma_j(A_0) \geq \frac{C_* c_*}{1-\delta}\sqrt{m_1 m_2}\sqrt{\frac{\log(m)(m_1 \vee m_2)}{n}}, \tag{5.4}$$

then $\hat{r} \geq \mathrm{rank}(A_0)$ *and*

$$\frac{1}{m_1 m_2}\|\hat{A}^{\lambda'} - A_0\|_2^2 \geq \frac{\delta^2 C_*^2 c_*^2}{4(1-\delta)^2}\mathrm{rank}(A_0)\frac{\log(m)(m_1 \vee m_2)}{n}, \tag{5.5}$$

with the same probability.

We note that the lower bound for $\sigma_j(A_0)$ in (5.4) is not excessively high, since $\sqrt{m_1 m_2}$ is a "typical" order of the largest singular value $\sigma_1(A_0)$ for non-lacunary matrices A_0. For example, if all the entries of A_0 are equal to some constant a, the left hand side of (5.4) is equal to $\sigma_1(A_0) = a\sqrt{m_1 m_2}$.

5.2. *Risk bounds in statistical learning.* The results of the previous sections can be also extended to the traditional statistical learning setting where (X_i, Y_i) is a sequence of i.i.d. replications of a random pair (X, Y) with $X \in \mathbb{R}^{m_1 \times m_2}$ and $Y \in \mathbb{R}$, and there is no underlying model determined by matrix A_0, i.e., we do not assume that $\mathbb{E}(Y|X) = \langle A_0, X\rangle$. Then the above oracle inequalities can be reformulated in terms of the prediction risk

$$R(A) = \mathbb{E}\big[(Y - \langle A, X\rangle)^2\big], \quad \forall\ A \in \mathbb{R}^{m_1 \times m_2}.$$

We illustrate this by an example dealing with USR matrix completion. Specifically, Theorem 4 is reformulated in the following way.

THEOREM 9. *Let X_i be i.i.d. uniformly distributed on $\mathcal{X}$. Assume that $|Y| \leq \eta$ almost surely for some constant η. For $t > 0$ consider the regularization parameter λ satisfying (3.6). Then with probability at least $1 - e^{-t}$ we have*

$$R(\hat{A}^\lambda) \leq R(A) + \min\left\{2\lambda\|A\|_1,\ \left(\frac{1+\sqrt{2}}{2}\right)^2 m_1 m_2 \lambda^2\,\mathrm{rank}(A)\right\} \tag{5.6}$$

for all $A \in \mathbb{R}^{m_1 \times m_2}$. In particular, under the assumptions of Corollary 2(ii),

$$R(\hat{A}^\lambda) \leq \min_{A \in \mathbb{R}^{m_1 \times m_2}}\left(R(A) + 4\left(1+\sqrt{2}\right)^2 \eta^2 \log(m)\frac{M\,\mathrm{rank}(A)}{n}\right). \tag{5.7}$$

This theorem can be also viewed as a result about the approximate sparsity. We do not know whether the true underlying model is described by some matrix A_0, but we can guarantee that our estimator is not far from the best approximation provided by matrices A with small rank or small nuclear norm.

Note that the results of Theorem 9 are uniform over the class of distributions

$$\mathcal{P}_\eta = \{P_{XY} : \ X \sim \Pi_0, \ |Y| \le \eta \,(\text{a.s.})\,\},$$

where Π_0 is the uniform distribution on $\mathcal{X}$, and $\eta > 0$ is a constant. The corresponding lower bound is given in the next theorem.

THEOREM 10. *Let n, m_1, m_2, r be as in Theorem 5. Let (X_i, Y_i) be i.i.d. realizations of a random pair (X, Y) with distribution P_{XY}. Then*

$$\inf_{\hat{A}} \sup_{\text{rank}(A) \le r} \sup_{P_{XY} \in \mathcal{P}_\eta} \mathbb{P}\left(R(\hat{A}) \ge R(A) + c\eta^2 \frac{Mr}{n} \right) \ \ge \ \beta, \tag{5.8}$$

where $\beta \in (0,1)$ and $c > 0$ are absolute constants.

PROOF. For $\mathbb{E}(Y|X) = \langle A_0, X \rangle$ we have $R(A) = \|A - A_0\|_{L_2(\Pi)}^2 + \sigma^2 = (m_1 m_2)^{-1} \|A - A_0\|_2^2 + \sigma^2$, where $\sigma^2 = \mathbb{E}\big[(Y - \mathbb{E}(Y|X))^2\big]$. Thus, using Theorem 7 we get

$$\begin{aligned} &\sup_{\text{rank}(A) \le r} \sup_{P_{XY} \in \mathcal{P}_\eta} \mathbb{P}\left(R(\hat{A}) \ge R(A) + c\eta^2 \frac{Mr}{n} \right) \\ \ge \quad & \sup_{\text{rank}(A) \le r} \sup_{P_{XY} \in \mathcal{P}_{A,\eta}} \mathbb{P}\left(\frac{1}{m_1 m_2} \|\hat{A} - A\|_2^2 > c\eta^2 \frac{Mr}{n} \right) > \beta. \end{aligned}$$

□

Inequalities (5.7) and (5.8) imply minimax rate optimality of $\hat{A}^\lambda$ up to a logarithmic factor in the statistical learning setting.

5.3. *Risks bounds in spectral norm.* All the results of the previous sections on the Frobenius norm can be extended to the spectral norm. We consider in this subsection the USR matrix completion problem, i.e, we assume that the matrices X_i are i.i.d. uniformly distributed in the matrix completion basis $\mathcal{X}$, which implies that $\|A\|_2^2 = (m_1 m_2)^{-1} \|A\|_2^2$ for all matrices $A \in \mathbb{R}^{m_1 \times m_2}$.

THEOREM 11. *Let X_i be i.i.d. uniformly distributed on $\mathcal{X}$. Consider the estimator $\hat{A}^\lambda$ defined in (3.1). If $\lambda \ge \|\mathbf{M}\|$, then we have*

$$\|\hat{A}^\lambda - A_0\| \le \frac{3}{2} m_1 m_2 \lambda.$$

PROOF. We have

$$\|\hat{A}^\lambda - A_0\| \le \|\hat{A}^\lambda - \mathbf{X}\| + m_1 m_2 \|\mathbf{M}\|,$$

where we recall that $\mathbf{X} = \frac{m_1 m_2}{n} \sum_{i=1}^n Y_i X_i$, $\mathbb{E}(\mathbf{X}) = A_0$ and $\mathbf{M}$ is defined in (2.2). In view of (3.2), we clearly have $\|\hat{A}^\lambda - \mathbf{X}\| \le \lambda m_1 m_2/2$. The result follows immediately on the event $\|\mathbf{M}\| \le \lambda$. □

As a consequence of the above theorem, we can derive the optimal rate (up a to logarithmic factor) of USR matrix completion for the spectral norm when the noise is sub-exponential or in the statistical learning setting.

THEOREM 12. *Let one of the sets of conditions (i) or (ii) in Corollary 2 be satisfied. Then, with probability at least* $1 - 3/(m_1 + m_2)$, *we have*

$$\|\hat{A}^\lambda - A_0\| \le C C_* c_* \sqrt{m_1 m_2} \sqrt{\frac{(m_1 \vee m_2)\log m}{n}},$$

where $C > 0$ *is an absolute constant.*

PROOF. The proof of this result is immediate by combining Theorem 11 and Lemmas 1, 2 and 3. □

THEOREM 13. *(i) Let the conditions of Theorem 6 be satisfied. Then*

$$\inf_{\hat{A}} \sup_{A_0 \in \mathcal{A}(r,a)} \mathbb{P}_{A_0}\left(\|\hat{A} - A_0\| > c(\sigma \wedge a)\sqrt{m_1 m_2} \sqrt{\frac{m_1 \vee m_2}{n}} \right) \ \ge\ \beta, \tag{5.9}$$

where $\beta \in (0,1)$ *and* $c > 0$ *are absolute constants.*

(ii) Let the conditions of Theorem 7 be satisfied. Then

$$\inf_{\hat{A}} \sup_{\operatorname{rank}(A_0) \le r} \sup_{P_{XY} \in \mathcal{P}_{A_0,\eta}} \mathbb{P}\left(\|\hat{A} - A_0\| > c\eta \sqrt{m_1 m_2} \sqrt{\frac{m_1 \vee m_2}{n}} \right) \ \ge\ \beta, \tag{5.10}$$

where $\beta \in (0,1)$ *and* $c > 0$ *are absolute constants.*

PROOF. Note first that Assumption 2 is satisfied with $\delta_r = 0$ and $\mu = \sqrt{m_1 m_2}$ in the USR matrix completion problem.

We prove the case (i). We consider the set of matrices $\mathcal{A}_0$ introduced in the proof of Theorem 5. For any two distinct matrices A_1, A_2 of $\mathcal{A}_0$, we have

$$\|A_1 - A_2\| \ge \sqrt{\frac{\gamma}{16}} (\sigma \wedge a) \sqrt{m_1 m_2} \sqrt{\frac{m_1 \vee m_2}{n}}. \tag{5.11}$$

Indeed, in the opposite case, we get

$$\|A_1 - A_2\|_2^2 \le \operatorname{rank}(A_1 - A_2)\|A_1 - A_2\|^2 < \frac{\gamma}{16}(\sigma \wedge a)^2 m_1 m_2 \frac{(m_1 \vee m_2)r}{n},$$

since $\operatorname{rank}(A_1 - A_2) \le r$ by construction of $\mathcal{A}_0$, which contradicts (4.3).

Next, (4.5) is satisfied for any $\alpha > 0$ if $\gamma > 0$ is chosen as a sufficiently small numerical constant depending on α.

Combining (5.11) with (4.5) and Theorem 2.5 in [26] gives the result.

The proof of (ii) follows the same arguments. □

5.4. *Sharp oracle inequality for the Lasso.* As we already mentioned in Example 4 and in the remark after Theorem 2, one can exploit (2.19) to derive a sharp (i.e., with the leading constant 1) sparsity oracle inequality for the Lasso. Indeed, if $m_1 = m_2 = p$ and A and X_i are diagonal matrices, then the trace regression model (1.2) becomes

$$Y_i = x_i^\top \beta^* + \xi_i, \quad i = 1, \dots, n,$$

where $x_i, \beta^* \in \mathbb{R}^p$ denote the vectors of diagonal elements of X_i, A_0, respectively. Set $\mathbb{X} = (x_1, \dots, x_n)^\top \in \mathbb{R}^{n\times p}$ to be the design matrix of this linear regression model. For a vector $z = (z^{(1)}, \dots, z^{(d)}) \in \mathbb{R}^d$, define $|z|_q = \left(\sum_{j=1}^d |z^{(j)}|^q\right)^{1/q}$ for $1 \le q < \infty$ and $|z|_\infty = \max_{1\le j\le d} |z^{(j)}|$.

Assume in what follows that x_i are fixed. Then for $A = \operatorname{diag}(\beta)$ we have $\|A\|_{L_2(\Pi)}^2 = n^{-1}|\mathbb{X}\beta|_2^2$, where $\operatorname{diag}(\beta)$ denotes the diagonal $p \times p$ matrix with the components of β on the diagonal. We will assume without loss of generality that the diagonal elements of the Gram matrix $\frac{1}{n}\mathbb{X}^\top\mathbb{X}$ are not larger than 1 (the general case is obtained from this by simple rescaling).

The estimator $\hat{A}^\lambda$ defined in (1.7) becomes the usual Lasso estimator

$$\hat{\beta}^\lambda = \arg\min_{\beta\in\mathbb{R}^p} \left\{ \frac{1}{n}\sum_{i=1}^n (Y_i - x_i^\top\beta)^2 + \lambda|\beta|_1 \right\}.$$

For a vector $\beta \in \mathbb{R}^p$, we set, with a little abuse of notation, $\mu_{c_0}(\beta) = \mu_{c_0}(\operatorname{diag}(\beta))$, $\mu(\beta) = \mu_5(\beta)$. Let $M(\beta)$ denote the number of nonzero components of β.

For simplicity, the result is stated only in the case of Gaussian noise.

THEOREM 14. *Let ξ_i be i.i.d. Gaussian $\mathcal{N}(0, \sigma^2)$ and let the diagonal elements of matrix $\frac{1}{n}\mathbb{X}^\top\mathbb{X}$ be not larger than* 1. *Take*

$$\lambda = C\sigma\sqrt{\frac{\log p}{n}},$$

where $C = 3a\sqrt{2}, a \geq 1$*. Then, with probability at least* $1 - \frac{1}{p^{a^2-1}\sqrt{\pi \log p}}$*, we have*

$$\frac{1}{n}|\mathbb{X}(\hat{\beta}^{\lambda} - \beta^*)|_2^2 \leq \inf_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{n}|\mathbb{X}(\beta - \beta^*)|_2^2 + C^2\sigma^2 \frac{\mu^2(\beta)M(\beta)\log p}{n} \right\}. \tag{5.12}$$

Proof. Combine Theorem 2 and a standard bound on the tail of the Gaussian distribution, which assures that with probability at least $1 - \frac{1}{p^{a^2-1}\sqrt{\pi \log p}}$,

$$\|\mathbf{M}\|_\infty = \left| \frac{1}{n}\sum_{i=1}^n \xi_i x_i \right|_\infty \leq a\sigma\sqrt{\frac{2\log p}{n}}.$$

□

Given $\beta \in \mathbb{R}^p$ and $J \subset \{1, \dots, p\}$, denote by β_J the vector in $\mathbb{R}^p$ which has the same coordinates as β on J and zero coordinates on the complement J^c of J.

We recall the Restricted Eigenvalue condition of [6]:

Condition RE(s, c_0)**.** *For some integer* s *such that* $1 \leq s \leq p$*, and a positive number* c_0 *the following condition holds:*

$$\kappa(s, c_0) \triangleq \min_{\substack{J \subseteq \{1,\dots,p\},\\ |J| \leq s}} \min_{\substack{u \in \mathbb{R}^p,\, u \neq 0,\\ |u_{J^c}|_1 \leq c_0 |u_J|_1}} \frac{|\mathbb{X}u|_2}{\sqrt{n}|u_J|_2} > 0.$$

We have the following corollary.

Corollary 4. *Let the assumptions of Theorem 14 hold, and let condition* **RE**$(s, 5)$ *be satisfied for some* $1 \leq s \leq p$*. Then, with probability at least* $1 - \frac{1}{p^{a^2-1}\sqrt{\pi \log p}}$,

$$\frac{1}{n}|\mathbb{X}(\hat{\beta}^{\lambda} - \beta^*)|_2^2 \leq \inf_{\beta \in \mathbb{R}^p \,:\, M(\beta) \leq s} \left\{ \frac{1}{n}|\mathbb{X}(\beta - \beta^*)|_2^2 + \frac{C^2\sigma^2}{\kappa^2(s,5)} \frac{M(\beta)\log p}{n} \right\}. \tag{5.13}$$

Proof. Recall that $e_j(p)$ denote the canonical basis vectors of $\mathbb{R}^p$. For any $p \times p$ diagonal matrix A with support (S_1, S_2), $S_1 = S_2 = \{e_j(p), j \in J\}$, where $J \subset \{1, \dots, p\}$ has cardinality $|J| \leq s$, and an arbitrary $p \times p$ diagonal matrix $B = \mathrm{diag}(u)$, where $u \in \mathbb{R}^p$, we have

$$\|\mathcal{P}_A(B)\|_1 = |u_J|_1, \quad \|\mathcal{P}_A^{\perp}(B)\|_1 = |u_{J^c}|_1,$$

and

$$\mathbb{C}_{A,c_0} = \{ \operatorname{diag}(u) : u \in \mathbb{R}^p, |u_{J^c}|_1 \le c_0 |u_J|_1 \} .$$

Thus

$$\kappa(s, c_0) = \inf_{u\in\mathbb{R}^p :\, u\neq 0,\, M(u)\le s} \frac{1}{\mu_{c_0}(\operatorname{diag}(u))} .$$

Since Condition $\mathbf{RE}(s,5)$ is satisfied, Theorem 14 yields the result. □

REMARK. Oracle inequalities (5.12) and (5.13) can be straightforwardly extended to the model $Y_i = f_i + \xi_i, i = 1,\dots,n$, where f_i are arbitrary fixed values and not necessarily $f_i = x_i^\top \beta^*$. This setting is interesting in the context of aggregation. Then $x_1,\dots,x_n$ are vectors of values of some given dictionary of p functions at n given points and f_i are the values of an unknown regression function at the same points. Under this model, inequalities (5.12) and (5.13) hold true with the only difference that $\mathbb{X}\beta^*$ should be replaced by the vector $f = (f_1,\dots,f_n)^\top$. With such a modification, (5.13) improves upon Theorem 6.1 of [6] where the leading constant is larger than 1.

6. Control of the stochastic error. In this section, we obtain the probability inequalities for the stochastic error $\|\mathbf{M}\|_\infty$. For brevity, we will write throughout $\|\cdot\|_\infty = \|\cdot\|$. The following proposition is an immediate consequence of the matrix version of Bernstein's inequality (Corollary 9.1 in [25]).

PROPOSITION 1. *Let $Z_1,\dots,Z_n$ be independent random matrices with dimensions $m_1 \times m_2$ that satisfy $\mathbb{E}(Z_i) = 0$ and $\|Z_i\| \le U$ almost surely for some constant U and all $i = 1,\dots,n$. Define*

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

Then, for all $t > 0$, with probability at least $1 - e^{-t}$ we have

$$\left\| \frac{Z_1 + \cdots + Z_n}{n} \right\| \le 2 \max\left\{ \sigma_Z \sqrt{\frac{t + \log(m)}{n}}, \ U \frac{t + \log(m)}{n} \right\} ,$$

where $m = m_1 + m_2$.

Furthermore, it is possible to replace the L_∞-bound U on $\|Z\|$ in the above inequality by bounds on the weaker ψ_α-norms of $\|Z\|$ defined by

$$U_Z^{(\alpha)} = \inf\Big\{ u > 0 : \mathbb{E}\exp(\|Z\|^\alpha / u^\alpha) \le 2 \Big\}, \quad \alpha \ge 1.$$

PROPOSITION 2. *Let $Z, Z_1, \dots, Z_n$ be i.i.d. random matrices with dimensions $m_1 \times m_2$ that satisfy $\mathbb{E}(Z) = 0$. Suppose that $U_Z^{(\alpha)} < \infty$ for some $\alpha \geq 1$. Then there exists a constant $C > 0$ such that, for all $t > 0$, with probability at least $1 - e^{-t}$*

$$\left\| \frac{Z_1 + \cdots + Z_n}{n} \right\| \leq C \max \left\{ \sigma_Z \sqrt{\frac{t + \log(m)}{n}}, \quad U_Z^{(\alpha)} \left(\log \frac{U_Z^{(\alpha)}}{\sigma_Z} \right)^{1/\alpha} \frac{t + \log(m)}{n} \right\},$$

where $m = m_1 + m_2$.

This is an easy consequence of Proposition 2 in [17], which provides an analogous result for Hermitian matrices Z. Its extension to rectangular matrices stated in Proposition 2 is straightforward via the self-adjoint dilation, cf., for example, the proof of Corollary 9.1 in [25].

The next lemma gives a control of the stochastic error for USR matrix completion in the statistical learning setting.

LEMMA 1. *Let X_i be i.i.d. uniformly distributed on $\mathcal{X}$. Assume that $\max_{i=1,\dots,n} |Y_i| \leq \eta$ almost surely for some constant η. Then for any $t > 0$ with probability at least $1 - e^{-t}$ we have*

$$\|\mathbf{M}\| \leq 2\eta \max \left\{ \sqrt{\frac{t + \log(m)}{(m_1 \wedge m_2)n}}, \quad \frac{2(t + \log(m))}{n} \right\}. \tag{6.1}$$

PROOF. We apply Proposition 1 with $Z_i = Y_i X_i - \mathbb{E}(Y_i X_i)$. Recall that here X_i are i.i.d. with the same distribution as X and Y_i are not necessarily i.i.d. Observe that

$$\|X\| = 1, \quad \|\mathbb{E}(X)\| = \sqrt{\frac{1}{m_1 m_2}}, \quad \sigma_X^2 = \frac{1}{m_1 \wedge m_2}. \tag{6.2}$$

Therefore, $\|Z_i\| \leq 2\eta$, $\sigma_Z \leq \eta \sigma_X$, and the result follows from Proposition 1. □

We now consider the USR matrix completion with sub-exponential errors. Recall that in this case we assume that the pairs (X_i, Y_i) are i.i.d. We have

$$\begin{aligned} \|\mathbf{M}\| &= \left\| \frac{1}{n} \sum_{i=1}^{n} (Y_i X_i - \mathbb{E}(Y_i X_i)) \right\| \\ &\leq \left\| \frac{1}{n} \sum_{i=1}^{n} \xi_i X_i \right\| + \left\| \frac{1}{n} \sum_{i=1}^{n} \Big(\operatorname{tr}(A_0^\top X_i) X_i - \mathbb{E}(\operatorname{tr}(A_0^\top X) X) \Big) \right\| \\ &= \Delta_1 + \Delta_2. \end{aligned}$$

We treat the terms Δ_1 and Δ_2 separately in the two lemmas below.

LEMMA 2. *Let X_i be i.i.d. uniformly distributed on $\mathcal{X}$, and the pairs (X_i, Y_i) be i.i.d. Assume that condition (3.3) holds. Then there exists an absolute constant $C > 0$ that can depend only on $\alpha, c_1, \tilde{c}$ and such that, for all $t > 0$, with probability at least $1 - 2e^{-t}$ we have*

$$\Delta_1 \leq C\sigma \max\left\{\sqrt{\frac{t + \log(m)}{(m_1 \wedge m_2)n}}\,,\ \frac{(t + \log(m))\log^{1/\alpha}(m_1 \wedge m_2)}{n}\right\}. \tag{6.3}$$

PROOF. Observe first that for $\tilde{X} = X - \mathbb{E}(X)$ we have

$$\sigma^2_{\tilde{X}} = \frac{1}{m_1 \wedge m_2}\,. \tag{6.4}$$

Now,

$$\begin{aligned}
\left\|\frac{1}{n}\sum_{i=1}^n \xi_i X_i\right\| &\leq \left\|\frac{1}{n}\sum_{i=1}^n \xi_i (X_i - \mathbb{E}X_i)\right\| + \left\|\frac{1}{n}\sum_{i=1}^n \xi_i \mathbb{E}(X_i)\right\| \\
&\leq \left\|\frac{1}{n}\sum_{i=1}^n \xi_i\left(X_i - \mathbb{E}X\right)\right\| + \sqrt{\frac{1}{m_1 m_2}}\left|\frac{1}{n}\sum_{i=1}^n \xi_i\right|. \qquad (6.5)
\end{aligned}$$

Set $Z_i = \xi_i\left(X_i - \mathbb{E}X\right)$. These are i.i.d. random matrices having the same distribution as a random matrix Z. It follows from (6.2) that $\|Z_i\| \leq 2|\xi_i|$, and thus condition (3.3) implies that $U_Z^{(\alpha)} \leq c\sigma$ for some constant $c > 0$. Furthermore, in view of (6.4), we have $\sigma_Z \leq c'\sigma\sigma_{\tilde{X}} = c'\sigma/(m_1 \wedge m_2)^{1/2}$ for some constant $c' > 0$ and $\sigma_Z \geq c_1^{1/2}\sigma/(2(m_1 \wedge m_2))^{1/2}$. Using these remarks we can deduce from Proposition 2 that there exists an absolute constant $\tilde{C} > 0$ such that for any $t > 0$ with probability at least $1 - e^{-t}$ we have

$$\begin{aligned}
&\left\|\frac{1}{n}\sum_{i=1}^n \xi_i\left(X_i - \mathbb{E}X\right)\right\| \\
&\quad \leq \tilde{C}\max\left\{\sigma_Z\sqrt{\frac{t + \log(m)}{n}}\,,\ U_Z^{(\alpha)}\left(\log\frac{U_Z^{(\alpha)}}{\sigma_Z}\right)^{1/\alpha}\frac{t + \log(m)}{n}\right\} \\
&\quad \leq C\sigma\max\left\{\sqrt{\frac{t + \log(m)}{(m_1 \wedge m_2)n}}\,,\ \frac{(t + \log(m))\log^{1/\alpha}(m_1 \wedge m_2)}{n}\right\}.
\end{aligned}$$

Finally, in view of Condition (3.3) and Bernstein's inequality for sub-exponential noise, we have for any $t > 0$, with probability at least $1 - e^{-t}$,

$$\left| \frac{1}{n} \sum_{i=1}^{n} \xi_i \right| \quad \leq \quad C\sigma \max \left\{ \sqrt{\frac{t}{n}}, \frac{t}{n} \right\},$$

where $C > 0$ depends only on $\tilde{c}$. We complete the proof by using the union bound. □

Define now

$$|A_0|_* = \max \left\{ \sqrt{\max_{1\leq i \leq m_1} \sum_{j=1}^{m_2} a_0^2(i,j)},\ \sqrt{\max_{1\leq j \leq m_2} \sum_{i=1}^{m_1} a_0^2(i,j)} \right\} \leq \max_{i,j} |a_0(i,j)| \sqrt{m_1 \vee m_2}\,.$$

LEMMA 3. *Let X_i be i.i.d. random variables uniformly distributed in $\mathcal{X}$. Then, for all $t > 0$, with probability at least $1 - e^{-t}$ we have*

$$\Delta_2 \leq 2 \max \left\{ |A_0|_* \sqrt{\frac{t + \log(m)}{m_1 m_2 n}}\ ,\ \ 2 \max_{i,j} |a_0(i,j)| \frac{t + \log(m)}{n} \right\}. \tag{6.6}$$

If $\max_{i,j} |a_0(i,j)| \leq a$ *for some $a > 0$, then with the same probability*

$$\Delta_2 \leq\ 2a\ \max \left\{ \sqrt{\frac{t + \log(m)}{(m_1 \wedge m_2) n}},\ \ \frac{2(t + \log(m))}{n} \right\}.$$

PROOF. We apply Proposition 1 for the random variables $Z_i = \mathrm{tr}(A_0^\top X_i) X_i - \mathbb{E}(\mathrm{tr}(A_0^\top X) X)$. Using (6.2) we get $\|Z_i\| \leq 2 \max_{i,j} |a_0(i,j)|$ and

$$\sigma_Z^2 \leq \max \left\{ \|\mathbb{E}\big(\langle A_0, X\rangle^2 X X^\top\big)\|,\ \|\mathbb{E}\big(\langle A_0, X\rangle^2 X^\top X\ \big)\| \right\} \leq \frac{|A_0|_*^2}{m_1 m_2}\,.$$

Thus, (6.6) follows from Proposition 1.

□

References.

[1] AHLSWEDE, R. AND WINTER, A. (2002) Strong converse for identification via quantum channels. *IEEE Transactions on Information Theory,* 48, 3, pp. 569–679.

[2] ARGYRIOU, A., EVGENIOU, T. AND PONTIL, M. (2008) Convex multi-task feature learning. *Machine Learning,* ***73***, 243–272.

[3] Argyriou, A., Micchelli, C.A., and Pontil, M. (2010) On spectral learning. *Journal of Machine Learning Research*, 11, 935-953.

[4] Argyriou, A., Micchelli, C.A., Pontil, M., and Ying, Y. (2007) A Spectral Regularization Framework for Multi-Task Structure Learning. *Proceedings of NIPS-2007.*

[5] Bach, F.R. (2008). Consistency of trace norm minimization. *Journal of Machine Learning Research,* ***9***, 1019–1048.

[6] Bickel, P., Ritov, Y. and Tsybakov, A. B. (2009). Simultaneous analysis of Lasso and Dantzig selector. *Ann. Statist.* **37** 1705–1732.

[7] Bunea, F., She, Y. and Wegkamp, M.H. (2010) Optimal selection of reduced rank estimators of high-dimensional matrices. `arXiv:1004.2995`. April 2010.

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

[9] Candès, E. J. and Plan, Y. (2010) Tight oracle bounds for low-rank matrix recovery from a mininal number of noisy random measurements. `arXiv:1001.0339`. January, 2010.

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

[11] Candès, E. and Tao, T. (2009) The power of convex relaxation: Near-optimal matrix completion. `arXiv:0903.1476`

[12] Giraud, C. (2010) Low rank multivariate regression. `arXiv:1009.5165`

[13] Gaiffas, S. and Lecué, G. (2010) Sharp oracle inequalities for the prediction of a high-dimensional matrix. `arXiv:1008.4886`

[14] Gross, D. (2009) Recovering low-rank matrices from few coefficients in any basis. `arXiv:0910.1879`.

[15] Keshavan, R.H., Montanari, A. and Oh, S. (2009) Matrix completion from noisy entries. `arXiv:0906.2027`

[16] Koltchinskii, V. (2009) The Dantzig Selector and Sparsity Oracle Inequalities. *Bernoulli,* **15**, 799-828.

[17] Koltchinskii, V. (2010) von Neumann entropy penalization and low rank matrix approximation. `arXiv:1009.2439`

[18] Negahban, S. and Wainwright, M.J. (2009) Estimation of (near) low rank matrices with noise and high-dimensional scaling. `arXiv:0912.5100`, December 2009.

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

[20] Negahban, S., Ravikumar, P., Wainwright, M.J., and Yu, B. (2010) A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. `arXiv:1010.2731`

[21] Recht, B. (2009) A simpler approach to matrix completion. `arXiv:0910.0651`

[22] Recht, B., Fazel, M. and Parrilo, P.A. (2007) Guaranteed Minimum-Rank Solutions of Linear Matrix Equations via Nuclear Norm Minimization. `arXiv:0706.4138`

[23] Rohde, A. and Tsybakov, A. (2009) Estimation of high-dimensional low rank matrices. `arXiv:0912.5338` December 2009.

[24] Stewart, G. W., Sun, J. (1990), Matrix Perturbation Theory. New York, Academic Press

[25] Tropp, J. A. (2010) User-friendly tail bounds for sums of random matrices. `arXiv:1004.4389`, April 2010.

[26] Tsybakov, A. (2009) *Introduction to Nonparametric Estimation.* Springer.

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

School of Mathematics
Georgia Institute of Technology
Atlanta, GA 30332-0160
E-mail: vlad@math.gatech.edu
klounici@math.gatech.edu

CREST
3, Av. Pierre Larousse
92240 Malakoff, France
E-mail: alexandre.tsybakov@ensae.fr