INSTITUT NATIONAL DE LA STATISTIQUE ET DES ETUDES ECONOMIQUES
Série des Documents de Travail du CREST
(Centre de Recherche en Economie et Statistique)

n° 2010-25

Estimation of High-dimensional Low Rank Matrices

A. ROHDE[1]
A. B. TSYBAKOV[2]

---

1 Université Hamburg.
2 CREST, Laboratoire de Statistique.

ESTIMATION OF HIGH-DIMENSIONAL LOW RANK MATRICES

By Angelika Rohde and Alexandre B. Tsybakov

Universität Hamburg and CREST

Suppose that we observe entries or, more generally, linear combinations of entries of an unknown $m \times T$-matrix A corrupted by noise. We are particularly interested in the high-dimensional setting where the number mT of unknown entries can be much larger than the sample size N. Motivated by several applications, we consider estimation of matrix A under the assumption that it has small rank. This can be viewed as dimension reduction or sparsity assumption. In order to shrink towards a low-rank representation, we investigate penalized least squares estimators with a Schatten-p quasi-norm penalty term, $p \leq 1$. We study these estimators under two possible assumptions – a modified version of the restricted isometry condition and a uniform bound on the ratio "empirical norm induced by the sampling operator/Frobenius norm". The main results are stated as non-asymptotic upper bounds on the prediction risk and on the Schatten-q risk of the estimators, where $q \in [p, 2]$. The rates that we obtain for the prediction risk are of the form rm/N (for $m = T$), up to logarithmic factors, where r is the rank of A. The particular examples of multi-task learning and matrix completion are worked out in detail. The proofs are based on tools from the theory of empirical processes. As a by-product we derive bounds for the kth entropy numbers of the quasi-convex Schatten class embeddings $S_p^M \hookrightarrow S_2^M$, $p < 1$, which are of independent interest.

1. Introduction. Consider the observations (X_i, Y_i) satisfying the model

$$Y_i = \mathrm{tr}(X_i' A^*) \; + \; \xi_i, \quad i = 1, ..., N, \tag{1.1}$$

where $X_i \in \mathbb{R}^{m \times T}$ are given matrices (m rows, T columns), $A^* \in \mathbb{R}^{m \times T}$ is an unknown matrix, $\mathrm{tr}(B)$ denotes the trace of square matrix B and ξ_i are i.i.d. random errors. Our aim is to estimate the matrix A^* and to predict the future Y-values based on the sample $(X_i, Y_i), i = 1, ..., N$.

We will call model (1.1) the *trace regression model.* Clearly, for $T = 1$ it reduces to the standard regression model. The "design" matrices X_i will be called *masks.* This name is motivated by the fact that we focus on the applications of trace regression where X_i are very sparse, i.e., contain only a small percentage of non-zero entries. Therefore, multiplication of A^* by X_i masks most of the entries of A^*. The following two examples are of particular interest.

(i) *Point masks.* For some, typically small, integer d the point masks X_i are defined

*Research partially supported by the grant ANR-06-BLAN-0194, ANR "Parcimonie" and by the PASCAL-2 Network of Excellence

AMS 2000 subject classifications: 62G05, 62F10

Keywords and phrases: High-dimensional low-rank matrices, empirical process, sparse recovery, Schatten norm, penalized least-squares estimator, quasi-convex Schatten class embeddings

as elements of the set

$$\mathcal{X}_d = \left\{ \sum_{i=1}^{d} e_{k_i}(m) e'_{l_i}(T) : 1 \le k_i \le m, 1 \le l_i \le T, \text{ with } (k_i, l_i) \ne (k_{i'}, l_{i'}) \text{ for } i \ne i' \right\},$$

where $e_k(m)$ are the canonical basis vectors of $\mathbb{R}^m$. In particular, for $d = 1$ the point masks X_i are matrices that have only one non-zero entry, which equals to 1. The problem of estimation of A^* in this case becomes the problem of *matrix completion*; the observations Y_i are just some selected entries of A^* corrupted by noise, and the aim is to reconstruct all the entries of A. The problem of matrix completion dates back at least to Srebro et al. (2005a, 2005b). We will analyze the following two special cases of matrix completion:

– *USR (Uniform Sampling at Random) matrix completion.* The masks X_i are independent, uniformly distributed on

$$\mathcal{X}_1 \;=\; \left\{ e_k(m) e'_l(T) : 1 \le k \le m, \; 1 \le l \le T \right\},$$

and independent from $\xi_1, \dots, \xi_N$.

– *Collaborative filtering.* The masks X_i (random or deterministic) belong to $\mathcal{X}_1$, are all distinct, and independent from $\xi_1, \dots, \xi_N$.

An important feature of the real-world matrix completion problems is that the number of observed entries is much smaller than the size of the matrix: $N \ll mT$, whereas mT can be very large. For example, mT is of the order of hundreds of millions for the Netflix problem.

(ii) *Column or row masks.* If X_i has only a small number d of non-zero columns or rows, it is called column or row mask, respectively. We suppose here that d is much smaller than m and T. A remarkable case $d = 1$ is covering the problem known in Statistics and Econometrics as longitudinal (or panel, or cross-section) data analysis and in Machine Learning as multi-task learning. In what follows we will designate this problem as multi-task learning, to avoid ambiguity. In the simplest version of multi-task learning, we have $N = nT$ where T is the number of tasks (for instance, in image detection each task t is associated with a particular type of visual object, e.g., face, car, chair, etc.), and n is the number of observations per task. The tasks are characterized by vectors of parameters $a^*_t \in \mathbb{R}^m$, $t = 1, \dots, T$, which constitute the columns of matrix A^*:

$$A^* = (a^*_1 \cdots a^*_T).$$

The X_i are column masks, each containing only one non-zero column $\mathbf{x}^{(t,s)} \in \mathbb{R}^m$ (with the convention that $\mathbf{x}^{(t,s)}$ is the tth column):

$$X_i \in \{(0 \cdots 0 \underbrace{\mathbf{x}^{(t,s)}}_{t} 0 \cdots 0), \; t = 1, \dots, T, \; s = 1, \dots, n\}.$$

The column $\mathbf{x}^{(t,s)}$ is interpreted as the vector of predictor variables corresponding to sth observation for the tth task. Thus, for each $i = 1, \dots, N$ there exists a pair (t, s) with $t = 1, \dots, T, \; s = 1, \dots, n$, such that

$$\mathrm{tr}(X'_i A^*) = (a^*_t)' \mathbf{x}^{(t,s)}. \tag{1.2}$$

If we denote by $Y^{(t,s)}$ and $\xi^{(t,s)}$ the corresponding values Y_i and ξ_i, then the trace regression model (1.1) can be written as a collection of T standard regression models:

$$Y^{(t,s)} = (a_t^*)'\mathbf{x}^{(t,s)} + \xi^{(t,s)}, \quad t = 1, \ldots, T, \ s = 1, \ldots, n.$$

This is the usual formulation of the multi-task learning model in the literature.

For both examples given above the matrices X_i are sparse in the sense that they have only a small portion of non-zero entries. On the other hand, such a sparsity property is not necessarily granted for the target matrix A^*. Nevertheless, we can always characterize A^* by its rank $r = \text{rank}\,(A^*)$, and say that a matrix is sparse if it has small rank, cf. Recht et al.(2007). For example, the problem of estimation of a square matrix $A^* \in \mathbb{R}^{m\times m}$ is a parametric problem which is formally of dimension m^2 but it has only $(2m-r)r$ free parameters. If r is small as compared to m, then the intrinsic dimension of the problem is of the order rm. In other words, the rank sparsity assumption $r \ll m$ is a dimension reduction assumption. This assumption will be crucial for the interpretation of our results. Another sparsity assumption that we will consider is that Schatten-p norm of A^* (see the definition in Section 2 below) is small for some $0 < p \leq 1$. This is an analog of sparsity expressed in terms of the ℓ_p norm, $0 < p \leq 1$, in vector estimation problems.

Estimation of high-dimensional matrices has been recently studied by several authors in settings different from the ours (cf., e.g., Meinshausen and Bühlmann (2006), Bickel and Levina (2008), Ravikumar et al. (2008), Amini and Wainwright (2009), Cai et al. (2010) and the references cited therein). Most of attention was devoted to estimation of a large covariance matrix or its inverse. In these papers sparsity is characterized by the number of non-zero entries of a matrix.

Candès and Recht (2008), Candès and Tao (2009), Gross (2009), Recht (2009) considered the non-noisy setting ($\xi_i \equiv 0$) of the matrix completion problem under conditions that the singular vectors of A^* are sufficiently spread out on the unit sphere or "incoherent". They focused on exact recovery of A^*. Up to date, the sharpest results are those of Gross (2009) and Recht (2009) who showed that under "incoherence condition" the exact recovery is possible with high probability if $N > C\,r(m+T)\log^2 m$ with some constant $C > 0$ when we observe N entries of a matrix $A^* \in \mathbb{R}^{m\times T}$ with locations uniformly sampled at random. Candès and Plan (2009), Keshavan et al. (2009) explored the same setting in the presence of noise, proposed estimators $\hat{A}$ of A^* and evaluated their Frobenius norm $\|\hat{A} - A^*\|_F$. The better bounds are in Keshavan et al. (2009) who suggest $\hat{A}$ such that for $A^* \in \mathbb{R}^{m\times T}$ and $T = \alpha m$ with $\alpha > 1$ the squared error $\|\hat{A} - A^*\|_F^2$ is of the order $\alpha^{5/2} r m^3 (\log N)/N$ with probability close to 1 when the noise is i.i.d. Gaussian.

In this paper we consider the general noisy setting of the trace regression problem. We study a class of Schatten-p estimators $\hat{A}$, i.e., the penalized least squares estimators with a penalty proportional to Schatten-p norm, cf. (2.5). The special case $p = 1$ corresponds to the "matrix Lasso". We study the convergence properties of their prediction error

$$\hat{d}_{2,N}(\hat{A}, A^*)^2 = N^{-1}\sum_{i=1}^{N} \text{tr}^2(X_i'(\hat{A} - A^*))$$

and of their Schatten-q error. The main contributions of this paper are the following.

(a) For all $0 < p \le 1$, under various assumptions on the masks X_i (no assumption, USR matrix completion, collaborative filtering) we obtain different bounds on the prediction error of Schatten-p estimators involving the Schatten-p norm of A^*.
(b) For p sufficiently close to 0, under a mild assumption on X_i, we show that Schatten-p estimators achieve the prediction error rate of convergence $\frac{r\max(m,T)}{N}$, up to a logarithmic factor. This result is valid for matrices A^* whose eigenvalues are not exponentially large in N. It covers the matrix completion and high-dimensional multi-task learning problems.
(c) For all $0 < p \le 1$, we obtain upper bounds for the prediction error under the matrix Restricted Isometry (RI) condition on the masks X_i, which is a rather strong condition, and under the assumption that $\mathrm{rank}(A^*) \le r$. We also derive the bounds for the Schatten-q error of $\hat{A}$. The rate in the bounds for the prediction error is $r\max(m,T)/N$ when the RI condition is satisfied with scaling factor 1 (i.e., for the case not related to matrix completion and high-dimensional multi-task learning).
(d) We prove the lower bounds showing that the rate $r\max(m,T)/N$ is minimax optimal for the prediction error and Schatten-2 (i.e., Frobenius) norm estimation error under the RI condition on the class of matrices A^* of rank smaller than r. Our result is even more general because we prove our lower bound on the intersection of the Schatten-0 ball with the Schatten-p ball for any $0 < p \le 1$, which allows us to show minimax optimality of the upper bounds of (a) as well. Furthermore, we prove a minimax lower bound for collaborative filtering.

The main message of this paper is to show that the suitably tuned Schatten estimators attain the optimal rate of prediction error up to logarithmic factors. The striking fact is that we can achieve this not only under the very restrictive assumption, such as the RI condition, but also under very mild assumptions on the masks X_i.

Finally, it is useful to compare the results for matrix estimation when the sparsity is expressed by the rank with those for the high-dimensional vector estimation when the sparsity is expressed by the number of non-zero components of the vector. For the vector estimation we have the linear model

$$Y_i \;=\; X_i'\beta \;+\xi_i, \quad i=1,...,N,$$

where $X_i \in \mathbb{R}^p$, $\beta \in \mathbb{R}^p$ and, for example, ξ_i are i.i.d. $\mathcal{N}(0,1)$ random variables. Consider the high-dimensional case, $p \gg N$. (This is analogous to the assumption $m^2 \gg N$ in the matrix problem and means that the nominal dimension is much larger than the sample size.) The sparsity assumption for the vector case has the form $s \ll N$, where s is the number of non-zero components, or the *intrinsic dimension* of β. Let $\hat\beta$ be an estimator of β. Then the optimal rate of convergence of the prediction risk $N^{-1}\sum_{i=1}^N (X_i'(\hat\beta-\beta))^2$ on the class of vectors β with given s is of the order s/N, up to logarithmic factors. This rate is shown to be attained, up to logarithmic factors, for many estimators, such as the BIC, the Lasso, the Dantzig selector, Sparse Exponential Weighting etc., cf., e.g., Bunea et al. (2007), Koltchinskii (2008), Bickel et al. (2009), Dalalyan and Tsybakov (2008). Note that this rate is of the form $\frac{\text{intrinsic dimension}}{\text{sample size}} = \frac{s}{N}$, up to a logarithmic factor. The general interpretation is therefore completely analogous to that of the matrix case: Assume for simplicity that A^* is a square $m \times m$ matrix with rank $(A^*) = r$. As mentioned above, the *intrinsic dimension* (the number of parameters to be estimated to recover A^*) is then $(2m-r)r$, which is of the order $\sim rm$ if $r \ll m$. An interesting difference is that the logarithmic risk inflation factor is inevitable in the vector case (cf. Donoho et al.(1992), Foster and George, (1994)), but not in the matrix problem, as our results reveal.

This paper is organized as follows. In Section 2 we introduce the notation, some definitions, basic facts about the Schatten quasi-norms and define the Schatten-p estimators. Section 3 describes elementary steps in their convergence analysis and presents two general approaches to upper bounds on the estimation and prediction error (cf. Theorems 1 and 2) depending on the efficient noise level τ. Our main results are stated in Sections 4, 5 (matrix completion), 6 (multi-task learning). They are obtained from Theorems 1 and 2 by specifying the effective noise level τ under particular assumptions on the masks X_i. Concentration bounds for certain random matrices leading to the expressions for the effective noise level are gathered in Section 8. Section 7 is devoted to minimax lower bounds. Sections 9 and 10 contain the main proofs. Finally, in Section 11 we establish bounds for the kth entropy numbers of the quasi-convex Schatten class embeddings $S_p^M \hookrightarrow S_2^M$, $p < 1$, which are needed for our proofs and are of independent interest.

2. Preliminaries.

2.1. *Notation, definitions and basic facts.* We will write $|\cdot|_2$ for the Euclidean norm in $\mathbb{R}^d$ for any integer d. For any matrix $A \in \mathbb{R}^{m\times T}$, we denote by $A_{(j,\cdot)}$ for $1 \le j \le m$ its jth row and write $A_{(\cdot,k)}$ for its kth column, $1 \le k \le T$. We denote by $\sigma_1(A) \ge \sigma_2(A) \ge \cdots \ge 0$ the singular values of A. The (quasi-) norm of some (quasi-) Banach space $\mathcal{B}$ is canonically denoted by $\|\cdot\|_{\mathcal{B}}$. In particular, for any matrix $A \in \mathbb{R}^{m\times T}$ and $0 < p < \infty$ we consider the Schatten (quasi-)norms

$$\|A\|_{S_p} = \Big(\sum_{j=1}^{\min(m,T)} \sigma_j(A)^p\Big)^{1/p} \quad \text{and} \quad \|A\|_{S_\infty} = \sigma_1(A).$$

The Schatten spaces S_p are defined as spaces of all matrices $A \in \mathbb{R}^{m\times T}$ equipped with quasi-norm $\|A\|_{S_p}$. In particular, the Schatten-2 norm coincides with the Frobenius norm:

$$\|A\|_{S_2} = \sqrt{\mathrm{tr}(A'A)} = \Big(\sum_{i,j} a_{ij}^2\Big)^{1/2}$$

where a_{ij} denote the elements of matrix $A \in \mathbb{R}^{m\times T}$, A' stands for the transposed of A. Recall that for $0 < p < 1$ the Schatten spaces S_p are not normed but only quasi-normed, and $\|\cdot\|_{S_p}^p$ satisfies the inequality

$$\|A+B\|_{S_p}^p \le \|A\|_{S_p}^p + \|B\|_{S_p}^p \tag{2.1}$$

for any $0 < p \le 1$ and any two matrices $A, B \in \mathbb{R}^{m\times T}$, cf. McCarthy (1967) and Rotfeld (1969). We will use the following well-known *trace duality* property:

$$|\mathrm{tr}(A'B)| \le \|A\|_{S_1}\|B\|_{S_\infty}, \quad \forall\, A, B \in \mathbb{R}^{m\times T}.$$

2.2. *Characteristics of the sampling operator.* Let $\mathcal{L} : \mathbb{R}^{m\times T} \to \mathbb{R}^N$ be the *sampling operator*, i.e., the linear mapping defined by

$$A \mapsto \left(\mathrm{tr}(X_1'A), ..., \mathrm{tr}(X_N'A)\right)/\sqrt{N}.$$

We have

$$|\mathcal{L}(A)|_2^2 = N^{-1}\sum_{i=1}^N \mathrm{tr}^2(X_i'A).$$

Depending on the context we also write $\hat{d}_{2,N}(A,B)$ for $|\mathcal{L}(A-B)|_2$, where A and B are any matrices in $\mathbb{R}^{m\times T}$. Unless the reverse is explicitly stated, it will tacitly be assumed that the X_i are non-random matrices.

We will denote by $\phi_{\max}(1)$ the maximal rank-1 restricted eigenvalue of $\mathcal{L}$:

$$\phi_{\max}(1) = \sup_{A\in\mathbb{R}^{m\times T}:\ \mathrm{rank}(A)=1} \frac{|\mathcal{L}(A)|_2}{\|A\|_{S_2}}. \tag{2.2}$$

Note that

$$\phi_{\max}^2(1) \le \min\left\{ \max_{1\le j\le m} \frac{1}{N}\sum_{i=1}^N |X_{i(j,\cdot)}|_2^2,\ \max_{1\le k\le T} \frac{1}{N}\sum_{i=1}^N |X_{i(\cdot,k)}|_2^2 \right\}. \tag{2.3}$$

We now list basic assumptions on the sampling operator that will be used in the sequel. The sampling operator $\mathcal{L}$ will be called *uniformly bounded* if there exists a constant $c_0 < \infty$ such that

$$\sup_{A\in\mathbb{R}^{m\times T}\setminus\{0\}} \frac{|\mathcal{L}(A)|_2^2}{\|A\|_{S_2}^2} \ \le\ c_0, \quad \text{uniformly in } m,\ T \text{ and } N. \tag{2.4}$$

Clearly, if $\mathcal{L}$ is uniformly bounded, then $\phi_{\max}^2(1) \le c_0$.

The sampling operator $\mathcal{L}$ is said to satisfy the *Restricted Isometry condition* RI (r,ν) for some integer $1 \le r \le \min(m,T)$ and some $0 < \nu < \infty$ if there exists a constant $\delta_r \in (0,1)$ such that

$$(1-\delta_r)\|A\|_{S_2} \ \le\ \nu|\mathcal{L}(A)|_2 \ \le\ (1+\delta_r)\|A\|_{S_2}$$

for all matrices $A \in \mathbb{R}^{m\times T}$ of rank at most r.

A difference of this condition from the Restricted Isometry condition introduced by Candès and Tao (2005) in the vector case or from its analog for the matrix case suggested by Recht et al. (2007), is that we state it with a *scaling factor* ν. This factor is introduced to account for the fact that the masks X_i are typically very sparse, so that they do not induce isometries with coefficient close to one. Indeed, ν will be large in the examples that we consider below.

2.3. *Least squares estimators with Schatten penalty.* In this paper we study the estimators $\hat{A}$ defined as a solution of the minimization problem

$$\min_{A\in\mathbb{R}^{m\times T}} \left(\frac{1}{N}\sum_{i=1}^N \left(Y_i - \mathrm{tr}(X_i'A)\right)^2 \ +\ \lambda\|A\|_{S_p}^p \right) \tag{2.5}$$

with some fixed $0 < p \le 1$ and $\lambda > 0$. The case $p = 1$ (matrix Lasso) is of outstanding interest since the minimization problem is then convex and thus can be efficiently solved in polynomial time. We call $\hat{A}$ the *Schatten-p estimator*. Such estimators have been recently considered by many authors motivated by applications to multi-task learning and collaborative filtering. Probably, the first study is due to Srebro et al. (2005a) who dealt with binary classification and considered the Schatten-1 estimator with the hinge loss rather than squared loss. Argyriou et al. (2007, 2008, 2009), Bach (2008), Abernethy et al.(2009) discussed connections of (2.5) to other related minimization problems, along with characterizations of the solutions and computational issues, mainly focusing on the convex case $p = 1$. Also for the non-convex case ($0 < p < 1$), Argyriou et al. (2007, 2008) suggested an algorithm of approximate computation of Schatten-p estimator or its analogs. However, for $0 < p < 1$ the methods can find only a local minimum in (2.5), so that Schatten estimators with such p remain for the moment mainly of theoretical value. In particular, analyzing these estimators reveals, which rates of convergence can, in principle, be attained.

The statistical properties of Schatten estimators are not yet well understood. To our knowledge, the only previous study is that of Bach (2008) showing that for $p = 1$, under some condition on $X_i's$ (analogous to strong irrepresentability condition in the vector case, cf. Meinshausen and Bühlmann (2006), Zhao and Yu (2006)), $\text{rank}(A^*)$ is consistently recovered by $\text{rank}(\hat{A})$ when m, T are fixed and $N \to \infty$. Our results are of a different kind. They are non-asymptotic and meaningful in the case $mT \gg N > \max(m, T)$. Furthermore, we do not consider the recovery of the rank, but rather the estimation and prediction properties of Schatten-p estimators.

After this paper has been submitted we became aware of interesting contemporaneous and independent works by Candès and Plan (2010), Negahban et al. (2009) and Negahban and Wainwright (2009). Those papers focus on the bounds for the Schatten-2 (i.e., Frobenius) norm error of the matrix Lasso estimator under the matrix RI condition. This is related to the particular instance of our results in item (c) above with $p = 1$ and $q = 2$. Their analysis of this case is complementary to ours in several aspects. Negahban and Wainwright (2009) derive their bound under the assumption that X_i are matrices with i.i.d. standard Gaussian elements and A^* belongs to a Schatten-p' ball with $0 \le p' \le 1$, which leads to rates different from ours if $p' \neq 0$. An assumption used in this context in Negahban and Wainwright (2009) is that $N > CmT$ (in our notation), which excludes the high-dimensional case $mT \gg N$, which we are mainly interested in. Candès and Plan (2010) consider approximately low rank matrices, explore the closely related matrix Dantzig selector and provide lower bounds corresponding to a special case of item (d) above. The results of these papers do not cover the matrix completion and high-dimensional multi-task learning problems, which are in the main focus of our study.

3. Two schemes of analyzing Schatten estimators.

In this section we discuss two schemes of proving upper bounds on the prediction error of $\hat{A}$. The first bound involves only the Schatten-p norm of matrix A^*. The second involves only the rank of A^* but needs the RI condition on the sampling operator.

We start by sketching elementary steps in the convergence analysis of Schatten-p estimators. By definition of $\hat{A}$,

$$\frac{1}{N}\sum_{i=1}^{N}\left(Y_i - \text{tr}(X_i'\hat{A})\right)^2 + \lambda\|\hat{A}\|_{S_p}^p \le \frac{1}{N}\sum_{i=1}^{N}\left(Y_i - \text{tr}(X_i'A^*)\right)^2 + \lambda\|A^*\|_{S_p}^p.$$

Recalling that $Y_i = \mathrm{tr}(X_i' A^*) + \xi_i$, we can transform this by a simple algebra to:

$$\hat{d}_{2,N}(\hat{A}, A^*)^2 \le \frac{2}{N}\sum_{i=1}^{N} \xi_i \mathrm{tr}\Big((\hat{A} - A^*)' X_i\Big) + \lambda\Big(\|A^*\|_{S_p}^p - \|\hat{A}\|_{S_p}^p\Big). \tag{3.1}$$

In the sequel, inequality (3.1) will be referred to as basic inequality and the random variable $N^{-1}\sum_{i=1}^N \xi_i \mathrm{tr}((\hat{A} - A^*)' X_i)$ will be called the *stochastic term.* The core in the analysis of Schatten-p estimators consists in proving tight bounds for the right hand side of the basic inequality (3.1). For this purpose, we first need a control of the stochastic term. Section 8 below demonstrates that such a control strongly depends on the properties of $\mathcal{L}$, i.e., of the problem at hand. In summary, Section 8 establishes that, under suitable conditions, for any $0 < p \le 1$ the stochastic term can be bounded for all $\delta > 0$ with probability close to 1 as follows:

$$\Big|\frac{1}{N}\sum_{i=1}^{N} \xi_i \mathrm{tr}(X_i'(\hat{A} - A^*))\Big| \le \begin{cases} \tau\|\hat{A} - A^*\|_{S_1} & \text{for } p = 1 \\ \frac{\delta}{2}\hat{d}_{2,N}(\hat{A}, A^*)^2 + \tau\delta^{p-1}\|\hat{A} - A^*\|_{S_p}^p & \text{for } 0 < p < 1, \end{cases} \tag{3.2}$$

where $0 < \tau < \infty$ depends on m, T and N. The quantity τ plays a crucial role in this bound. We will call τ the *effective noise level.* Exact expressions for τ under various assumptions on the sampling operator $\mathcal{L}$ and on the noise ξ_i are derived in Section 8. In Table 1 we present the values of τ for three important examples under the assumption that ξ_i are i.i.d. Gaussian $\mathcal{N}(0, \sigma^2)$ random variables. In the cases listed in Table 1, inequality (3.2) holds with probability $1 - \varepsilon$, where $\varepsilon = (1/C)\exp(-C(m+T))$ (first and third example) and $\varepsilon = (1/C')(\max(m, T)+1)^{-C'}$ (second example) with constants $C, C' > 0$ independent of N, m, T.

Assumptions on X_i	Assumptions on N, m, T, p	Value of τ
Uniformly bounded $\mathcal{L}$	$0 < p \le 1$	$c(p)(M/N)^{1-p/2}$
USR matrix completion	$p = 1$, $(m+T)mT > N$	$c\min(M/N, (\log M)/\sqrt{N})$
Collaborative filtering	$p = 1$	$cM^{1/2}/N$

TABLE 1. Effective noise level for uniformly bounded $\mathcal{L}$, USR matrix completion and collaborative filtering. Here $M = \max(m, T)$, and the constants $c > 0, c(p) > 0$ depend only on σ.

The following two points will be important to understand the subsequent results:

- In this paper, we will always choose the regularization parameter λ in the form: $\lambda = 4\tau$.
- With this choice of λ, the smaller is effective noise level τ, the faster is the rate of convergence of the Schatten estimator.

In particular, the first line in Table 1 reveals that when $M = \max(m, T) < N$ the largest τ corresponds to $p = 1$ and it becomes smaller when p decreases to 0. This suggests that choosing Schatten-p estimators with $p < 1$ and especially p close to 0 might be advantageous. Note that the assumption of uniform boundedness of $\mathcal{L}$ is very mild. For example, it is trivially satisfied with $c_0 = 1$ for USR matrix completion and collaborative filtering. Nevertheless, in those cases a specific analysis leads to sharper bounds on the effective noise level listed in the second and third lines of Table 1.

In this section we provide two bounds on the prediction error of $\hat{A}$ with a general effective noise level τ. We then detail them in Sections 4, 5, 6 for particular values of

τ depending on the assumptions on the X_i. The first bound involves the Schatten-p norm of matrix A^*.

THEOREM 1. *Let $A^* \in \mathbb{R}^{m\times T}$, and let $0 < p \le 1$. Assume that (3.2) holds with probability at least $1-\varepsilon$ for some $\varepsilon > 0$ and $0 < \tau < \infty$. Let $\hat{A}$ be the Schatten-p estimator defined as a minimizer of (2.5) with $\lambda = 4\tau$. Then*

$$\hat{d}_{2,N}(\hat{A}, A^*)^2 \le 16\tau \|A^*\|_{S_p}^p. \tag{3.3}$$

holds with probability at least $1-\varepsilon$.

PROOF. From (3.1) and (3.2) with $\delta = 1/2$ and $\lambda = 4\tau$ we get

$$\hat{d}_{2,N}(\hat{A}, A^*)^2 \le 8\tau\Big(\|\hat{A} - A^*\|_{S_p}^p + \|A^*\|_{S_p}^p - \|\hat{A}\|_{S_p}^p\Big).$$

This and the p-norm inequality (2.1) yield (3.3). □

The bound (3.3) depends on the magnitude of the elements of A^* via $\|A^*\|_{S_p}$. The next theorem shows that under the RI condition this dependence can be avoided, and only the rank of A^* affects the rate of convergence.

THEOREM 2. *Let $A^* \in \mathbb{R}^{m\times T}$ with $\mathrm{rank}(A^*) \le r$, and let $0 < p \le 1$. Assume that (3.2) holds with probability at least $1-\varepsilon$ for some $\varepsilon > 0$ and $0 < \tau < \infty$. Assume also that the Restricted Isometry condition RI $((2+a)r,\nu)$ holds with some $0 < \nu < \infty$, with a sufficiently large $a = a(p)$ depending only on p and with $0 < \delta_{(2+a)r} \le \delta_0$ for a sufficiently small $\delta_0 = \delta_0(p)$ depending only on p.*

Let $\hat{A}$ be the Schatten-p estimator defined as a minimizer of (2.5) with $\lambda = 4\tau$. Then with probability at least $1-\varepsilon$ we have

$$\hat{d}_{2,N}(\hat{A}, A^*)^2 \le C_1 r \tau^{\frac{2}{2-p}} \nu^{\frac{2p}{2-p}}, \tag{3.4}$$

$$\|\hat{A} - A^*\|_{S_q}^q \le C_2 r \tau^{\frac{q}{2-p}} \nu^{\frac{2q}{2-p}}, \quad \forall\, q \in [p, 2], \tag{3.5}$$

where C_1 and C_2 are constants, C_1 depends only on p and C_2 depends on p and q.

Proof of Theorem 2 is given in Section 9. The values $a = a(p)$ and $\delta_0(p)$ can be deduced from the proof. In particular, for $p = 1$ it is sufficient to take $a = 19$.

REMARK 1. Note that if $\nu = 1$ the rates in (3.4) and (3.5) do not depend on p if we assume in addition the uniform boundedness of $\mathcal{L}$, which is a very mild condition. Indeed, taking the value of τ from the first line of Table 1 we see that $r\tau^{2/(2-p)}\nu^{2p/(2-p)} \sim rM/N$ for all $0 < p \le 1$. Thus, under the RI condition, using Schatten-p estimators with $p < 1$ does not improve the rate of convergence on the class of matrices A^* of rank at most r.

Discussion about the scaling factor ν. Remark 1 deals with the case $\nu = 1$, which seems to be not always appropriate for trace regression models. To our knowledge, the only available examples of matrices X such that $\mathcal{L}(.) = \mathrm{tr}(X'.)$ satisfies the RI condition with $\nu = 1$ are *complete matrices*, i.e., matrices with all non-zero entries, which are random and have specific distributions (typically, i.i.d. Rademacher or Gaussian entries, cf. Recht et al. (2007)). Except for degenerate cases (such as $N =$

mT, the X_i distinct and of the form $\sqrt{N}e_k(m)e_l(T)'$ for $1 \le k \le m, 1 \le l \le T$) the sampling operator $\mathcal{L}$ defines typically a restricted isometry with $\nu = 1$ only if the matrices X_i contain a considerable number of (uniformly bounded) non-zero entries.

In view of the examples mentioned in the introduction, let us now discuss the form of the RI condition in the context of multi-task learning. Using the analog of (1.2) for a matrix $A = (a_1 \cdots a_T)$ we obtain

$$\begin{aligned} |\mathcal{L}(A)|_2^2 &= N^{-1}\sum_{i=1}^N \mathrm{tr}^2(X_i'A) \\ &= N^{-1}\sum_{t=1}^T\sum_{s=1}^n a_t'\mathbf{x}^{(t,s)}(\mathbf{x}^{(t,s)})'a_t \;=\; T^{-1}\sum_{t=1}^T a_t'\Psi_t a_t \end{aligned}$$

where $\Psi_t = n^{-1}\sum_{s=1}^n \mathbf{x}^{(t,s)}(\mathbf{x}^{(t,s)})'$ is the Gram matrix of predictors for the tth task. These matrices correspond to T separate regression models. The standard assumption is that they are normalized so that all the diagonal elements of each Ψ_t are equal to 1. This suggests that the natural RI scaling factor ν for such model is of the order $\nu \sim \sqrt{T}$. For example, in the simplest case when all the matrices Ψ_t are just equal to the $m \times m$ identity matrix, we find $|\mathcal{L}(A)|_2^2 = T^{-1}\sum_{t=1}^T a_t'\Psi_t a_t = T^{-1}\|A\|_{S_2}^2$. Similarly, we get the RI condition with scaling factor $\nu \sim \sqrt{T}$ when the spectra of all the Gram matrices $\Psi_t,\ t = 1,\dots,T$, are included in a fixed interval $[a,b]$ with $0 < a < b < \infty$. However, this excludes the high-dimensional task regressions, such that the number of parameters m is larger than the sample size, $m > n$. In conclusion, application of the matrix RI techniques in multi-task learning is restricted to low-dimensional regression and the scaling factor is $\nu \sim \sqrt{T}$.

The reason of the failure of RI approach is that the masks X_i are sparse. The sparser are X_i, the larger is ν. The extreme situation corresponds to matrix completion problems. Indeed, if $N < mT$, then there exists a matrix of rank 1 in the null-space of the sampling operator $\mathcal{L}$ and hence the RI condition cannot be satisfied. For $N \ge mT$ we can have the RI condition with scaling factor $\nu \sim \sqrt{mT}$, but $N \ge mT$ means that essentially all the entries are observed, and there is no sense to state the problem of completion.

4. Upper bounds under mild conditions on the sampling operator. The above discussion suggests that Theorem 2 and, in general, the argument based on the restricted isometry or related conditions are not well adapted for several interesting settings. Motivated by this, we propose another approach described in the next theorem, which requires only the comparably mild uniform boundedness condition (2.4). For example, this condition is satisfied in the USR matrix completion problem with $c_0 = 1$. For simplicity we focus on Gaussian errors ξ_i. Denote $M := \max(m,T)$.

THEOREM 3. *Let $\xi_1,\dots,\xi_N$ be i.i.d. $\mathcal{N}(0,\sigma^2)$ random variables. Assume that $1 < M := \max(m,T)$, $N > \mathrm{e}\,M$, and that the uniform boundedness condition (2.4) is satisfied. Let $A^* \in \mathbb{R}^{m\times T}$ with $\mathrm{rank}(A^*) \le r$ and the maximal singular value $\sigma_1(A^*) \le (N/M)^{C^*}$ for some $0 < C^* < \infty$. Set $p = (\log(N/M))^{-1}$, $c_\kappa = (2\kappa - 1)(2\kappa)\kappa^{-1/(2\kappa-1)}$ where $\kappa = (2-p)/(2-2p)$ and*

$$\lambda = 4c_\kappa(\vartheta/p)^{1-p/2}\Big(\frac{M}{N}\Big)^{1-p/2} \tag{4.1}$$

for some $\vartheta \geq C^2$ and C a universal positive constant independent of r, M and N. Then the Schatten-p estimator $\hat{A}$ defined as a minimizer of (2.5) with λ as in (4.1) satisfies:

$$\hat{d}_{2,N}(\hat{A}, A^*)^2 \leq C_3 \vartheta \frac{rM}{N} \log\left(\frac{N}{M}\right) \tag{4.2}$$

with probability at least $1 - C\exp(-\vartheta M/C^2)$ where the positive constant C_3 is independent of r, M and N.

PROOF. Inequality (3.2) holds with probability at least $1 - C\exp(-\vartheta M/C^2)$ by Lemma 5. We then use (3.3) and note that, under our choice of p, $\tau \leq c\vartheta M/(Np)$ for some constant $c < \infty$, which does not depend on M and N, and

$$\|A^*\|_{S_p}^p \leq r[\sigma_1(A^*)]^p \leq r\left(\frac{N}{M}\right)^{C^*p} = \exp(C^*)\, r.$$

□

Finally, we give the following theorem quantifying the rates of convergence of the prediction risk in terms of the Schatten norms of A^*. Its proof is straightforward in view of Theorem 1 and Lemmas 2, 5.

THEOREM 4. *Let $\xi_1, \ldots, \xi_N$ be i.i.d. $\mathcal{N}(0, \sigma^2)$ random variables and $A^* \in \mathbb{R}^{m\times T}$. Then the Schatten-p estimator $\hat{A}$ has the following properties.*

(i) Let $p = 1$, and $\lambda = 32\sigma\phi_{\max}(1)\sqrt{(m+T)/N}$. Then

$$\hat{d}_{2,N}(\hat{A}, A^*)^2 \leq C\sigma\phi_{\max}(1)\|A^*\|_{S_1}\sqrt{\frac{m+T}{N}} \tag{4.3}$$

with probability at least $1 - 2\exp\{-(2-\log 5)(m+T)\}$ where $C > 0$ is an absolute constant.

(ii) Let $0 < p < 1$ and let the uniform boundedness condition (2.4) hold. Set λ as in (4.1). Then with $M := \max(m, T)$,

$$\hat{d}_{2,N}(\hat{A}, A^*)^2 \leq C\|A^*\|_{S_p}^p \left(\frac{M}{N}\right)^{1-p/2} \tag{4.4}$$

with probability at least $1 - C\exp(-\vartheta M/C^2)$ where the constant $C > 0$ is independent of r, M and N.

In Theorem 5 below we show that these rates are optimal in a minimax sense on the corresponding Schatten-p balls for the sampling operators satisfying the RI condition.

5. Upper bounds for noisy matrix completion. As discussed in Section 3, for matrix completion problems the restricted isometry argument as in Theorem 2 is not applicable. The bound of Theorem 4 is also too coarse. We will therefore use Theorems 1 and 3. Combining them with Lemmas 3 and 4 we get the following two corollaries.

COROLLARY 1 (USR matrix completion). *Let the i.i.d. zero-mean random variables ξ_i satisfy the Bernstein condition (8.2). Assume that $mT(m+T) > N$ and consider the USR matrix completion model. Let τ_2 be given by (8.10) with some $D \geq 2$. Then the Schatten-1 estimator $\hat{A}$ defined with $\lambda = 4\tau_2$ satisfies:*

$$\hat{d}_{2,N}(\hat{A}, A^*)^2 \leq 16\bar{C}\|A^*\|_{S_1}\frac{m+T}{N} \tag{5.1}$$

with probability at least $1 - 4\exp\{-(2-\log 5)(m+T)\}$, where $\bar{C} = 4\sigma\sqrt{10D} + 8HD$.

(ii) Let the i.i.d. zero-mean random variables ξ_i satisfy the light tail condition (8.3), and let τ_3 be given by (8.11) for some $B > 0$. Then the Schatten-1 estimator $\hat{A}$ defined with $\lambda = 4\tau_3$ satisfies:

$$\hat{d}_{2,N}(\hat{A}, A^*)^2 \leq 16\|A^*\|_{S_1}\sqrt{B}\,\frac{\sigma \log\big(\max(m+1, T+1)\big)}{\sqrt{N}} \tag{5.2}$$

with probability at least $1 - (1/C)\max(m+1, T+1)^{-CB}$ for some constant $C > 0$ which does not depend on m, T and N.

COROLLARY 2 (USR matrix completion, non-convex penalty). *Let $\xi_1, \dots, \xi_N$ be i.i.d. $\mathcal{N}(0, \sigma^2)$ random variables. Assume that $M := \max(m, T) > 1$, $N > \mathrm{e}\,M$ and consider the USR matrix completion model. Let $A^* \in \mathbb{R}^{m\times T}$ with $\mathrm{rank}(A^*) \leq r$ and the maximal singular value $\sigma_1(A^*) \leq (N/M)^{C^*}$ for some $0 < C^* < \infty$. Set $p = (\log(N/M))^{-1}$, $c_\kappa = (2\kappa - 1)(2\kappa)\kappa^{-1/(2\kappa-1)}$ where $\kappa = (2-p)/(2-2p)$ and*

$$\lambda = 4c_\kappa(\vartheta/p)^{1-p/2}\Big(\frac{M}{N}\Big)^{1-p/2}$$

for some $\vartheta \geq C^2$ with a universal constant $C > 0$, independent of r, M and N. Then the Schatten-p estimator $\hat{A}$ defined as a minimizer of (2.5) satisfies:

$$\hat{d}_{2,N}(\hat{A}, A^*)^2 \leq C_3\vartheta\,\frac{rM}{N}\log\Big(\frac{N}{M}\Big) \tag{5.3}$$

with probability at least $1 - C\,\exp(-\vartheta M/C^2)$, where the positive constant C_3 is also independent of r, M and N.

Note that the bounds of Corollaries 1 and 2 achieve the rate $r\max(m,T)/N$, up to logarithmic factor, but under different conditions on the maximal singular value of A^*. If $\max(m,T) < N < mT$ then the condition in Corollary 2 does not imply more than a polynomial in $\max(m,T)$ growth condition on $\sigma_1(A^*)$, which is a mild assumption. Note that (5.1) requires uniform boundedness of the maximal singular value of A^* by some constant to achieve the same rate. On the other hand, the estimators of Corollary 2 correspond to non-convex penalty and are computationally hard.

COROLLARY 3 (Collaborative filtering). *Consider the problem of collaborative filtering with random or deterministic $X_1, \dots, X_N$ which are independent from $\xi_1, \dots, \xi_N$.*

(i) Let $\xi_1, ..., \xi_N$ be i.i.d. $\mathcal{N}(0, \sigma^2)$ random variables. Let τ_4 be given by (8.12) with some $D \geq 2$. Then the Schatten-1 estimator $\hat{A}$ defined with $\lambda = 4\tau_4$ satisfies

$$\hat{d}_{2,N}(\hat{A}, A^*)^2 \leq 16\bar{C}\|A^*\|_{S_1}\frac{\sqrt{m+T}}{N} \tag{5.4}$$

with probability at least $1 - 2\exp\{-(D - \log 5)(m + T)\}$, *where* $\bar{C} = 8\sigma\sqrt{D}$.

(ii) Let $\xi_1, ..., \xi_N$ *be i.i.d. zero-mean random variables satisfying the Bernstein condition (8.2). Let* τ_5 *be given by (8.13) with some* $D \geq 2$*. Then the Schatten-*1 *estimator* $\hat{A}$ *defined with* $\lambda = 4\tau_5$ *satisfies:*

$$\hat{d}_{2,N}(\hat{A}, A^*)^2 \leq 64\|A^*\|_{S_1} \frac{\sigma\sqrt{2D(m+T)} + 2HD(m+T)}{N} \tag{5.5}$$

with probability at least $1 - 2\exp\{-(D - \log 5)(m + T)\}$.

REMARK 2. Using the inequality $\|A^*\|_{S_1} \leq \sqrt{r}\|A^*\|_{S_2}$ for matrices A^* of rank at most r we find that that the bound (5.4) in (i) is minimax optimal on the class of matrices

$$\left\{A^* \in \mathbb{R}^{m\times T} : \operatorname{rank}(A^*) \leq r,\ \|A^*\|_{S_2}^2 \leq \sigma^2 \max(m, T) r\right\}$$

as long as the masks $X_1, \ldots, X_N$ fulfill the "dispersion" condition of Theorem 7. In view of the bounds for USR matrix completion, it is further interesting to note that the construction in the proof of Theorem 7 fails if $\|A^*\|_{S_2}^2$ is of lower order than $r \max(m, T)$.

6. Upper bounds for multi-task learning. For multi-task learning we can employ both Theorem 2 and Theorem 3. Theorem 2 imposes a strong assumption on the masks X_i, namely the RI condition. Nevertheless, the advantage is that Theorem 2 covers the computationally easy case $p = 1$.

COROLLARY 4 (Multi-task learning; RI condition). *Let* $\xi_1, \ldots, \xi_N$ *be i.i.d.* $\mathcal{N}(0, \sigma^2)$ *random variables. Consider the multi-task learning problem with* $\operatorname{rank}(A^*) \leq r$*. Assume that the spectra of the Gram matrices* Ψ_t *are uniformly in* t *bounded from above by a constant* $c_1 < \infty$*. Assume also that the Restricted Isometry condition RI* $(21r,\nu)$ *holds with some* $0 < \nu < \infty$ *and with* $0 < \delta_{21r} \leq \delta_0$ *for a sufficiently small* δ_0*. Set*

$$\lambda = 32\sigma\sqrt{\frac{c_1(m+T)}{nT^2}}\,.$$

Let $\hat{A}$ *be the Schatten-*1 *estimator with this parameter* λ*. Then with probability at least* $1 - 2\exp\{-(2 - \log 5)(m + T)\}$ *we have*

$$\begin{aligned}
\hat{d}_{2,N}(\hat{A}, A^*)^2 &\leq \bar{C}_1 c_1 \sigma^2 r \nu^2 \left(\frac{m+T}{nT^2}\right), \\
\|\hat{A} - A^*\|_{S_q}^q &\leq \bar{C}_2 c_1^{q/2} \sigma^q r \nu^{2q} \left(\frac{m+T}{nT^2}\right)^{q/2}, \quad \forall\, q \in [1, 2],
\end{aligned}$$

where $\bar{C}_1$ *is an absolute constant and* $\bar{C}_2$ *depends only on* q.

Proof of Corollary 4 is straightforward in view of Theorem 2, Lemma 2, and the fact that, under the premise of Corollary 4, we have $|\mathcal{L}(A)|_2^2 = T^{-1}\sum_{t=1}^T a_t'\Psi_t a_t \leq (c_1/T)\|A\|_{S_2}^2$ for all matrices $A \in \mathbb{R}^{m\times T}$, so that the sampling operator is uniformly bounded ((2.4) holds with $c_0 = c_1/T$), and thus $\phi_{\max}(1) \leq \sqrt{c_0} \leq \sqrt{c_1/T}$.

Taking in the bounds of Corollary 4 the natural scaling factor $\nu \sim \sqrt{T}$ we obtain the following inequalities

$$\hat{d}_{2,N}(\hat{A}, A^*)^2 \leq \tilde{C}_1 \frac{r(m+T)}{nT}, \tag{6.1}$$

$$\frac{1}{T}\|\hat{A} - A^*\|_{S_2}^2 \leq \tilde{C}_2 \frac{r(m+T)}{nT}, \tag{6.2}$$

where the constants $\tilde{C}_1$ and $\tilde{C}_2$ do not depend on m, T and n.

A remarkable fact is that the rates in Corollary 4 are free of logarithmic inflation factor. This is one of the differences between the matrix estimation problems and vector estimation ones, where the logarithmic risk inflation is inevitable, as first noticed by Donoho et al.(1992), Foster and George (1994). For more details about optimal rates of sparse estimation in the vector case, see Rigollet and Tsybakov (2010).

Corollary 4 and the bounds (6.1), (6.2) can be compared to those obtained for the Group Lasso estimator in multi-task setting by Lounici et al.(2009, 2010). The main difference is that the sparsity index s appearing in Lounici et al.(2009,2010) is now replaced by r. In Lounici et al.(2009, 2010), the columns a_t^* of A^* are supposed to be sparse, with the sets of non-zero elements of cardinality not more than s, whereas here the sparsity is characterized by the rank r of A^*.

Finally, we give the following result based on application of Theorem 3.

Corollary 5 (Multi-task learning; uniformly bounded $\mathcal{L}$). *Let $\xi_1, \ldots, \xi_N$ be i.i.d. $\mathcal{N}(0, \sigma^2)$ random variables, and assume that $n > \mathrm{e}$. Consider the multi-task learning problem with $A^* \in \mathbb{R}^{m\times T}$, $\mathrm{rank}(A^*) \leq r$, such that the maximal singular value $\sigma_1(A^*) \leq n^{C^*}$ for some $0 < C^* < \infty$. Assume that the spectra of the Gram matrices Ψ_t are uniformly in t bounded from above by $c_0 T$ where $c_0 < \infty$ is a constant. Set $p = (\log n)^{-1}$, $c_\kappa = (2\kappa-1)(2\kappa)\kappa^{-1/(2\kappa-1)}$ where $\kappa = (2-p)/(2-2p)$ and*

$$\lambda = 4c_\kappa(\vartheta/p)^{1-p/2}\Big(\frac{1}{n}\Big)^{1-p/2}$$

for some $\vartheta \geq C^2$ and a universal constant $C > 0$, independent of r, m and n. Then the Schatten-p estimator $\hat{A}$ with this parameter λ satisfies

$$\hat{d}_{2,N}(\hat{A}, A^*)^2 \leq C_3 \vartheta \frac{rM}{nT} \log n \tag{6.3}$$

with probability at least $1 - C\exp(-\vartheta M/C^2)$ where $M = \max(m, T)$, and the positive constant C_3 is independent of r, m and n.

Corollary 5 follows from Theorem 3. Indeed, it suffices to remark that, under the premises of Corollary 5, we have $|\mathcal{L}(A)|_2^2 = T^{-1}\sum_{t=1}^T a_t'\Psi_t a_t \leq c_0\|A\|_{S_2}^2$ for all matrices $A \in \mathbb{R}^{m\times T}$, so that the sampling operator is uniformly bounded, cf. (2.4).

For $m = T$, we can write (6.3) in the form

$$\hat{d}_{2,N}(\hat{A}, A^*)^2 \leq C_3' \frac{rm}{nT} \log n \tag{6.4}$$

Clearly, this bound achieves the optimal rate "intrinsic dimension/sample size"$\sim rm/N$, up to logarithms (recall that $N = nT$ in the multi-task learning). The bounds

(6.1) and (6.2) achieve this rate in a more precise sense because they are free of extra logarithmic factors.

Another remark concerns the possible range of m. As discussed in Section 3, the "dimension larger than the sample size" framework is not covered by Corollary 4 since it relies on the RI condition. In contrast to this, the bounds of Corollary 5 make sense when the dimension m is larger than the sample size n of each task; we only need to have $m \ll \exp(n)$ for Corollary 5 to be meaningful. Corollary 5 holds when the RI assumption is violated and under a mild condition on the masks X_i. The price to pay is to assume that the singular values of A^* do not grow exponentially fast. Also, the estimator of Corollary 5 corresponds to $p < 1$, so it is computationally hard.

7. Minimax lower bounds. In this section we derive lower bounds for the prediction error, which show that the upper bounds that we have proved are optimal in a minimax sense for two scenarios: (i) under the RI condition and (ii) for collaborative filtering. Under the RI condition with $\nu = 1$, minimax lower bounds for the Frobenius norm $\|\hat{A} - A^*\|_{S_2}$ on "Schatten-0" balls $\{A^* \in \mathbb{R}^{m\times T} : \operatorname{rank}(A^*) \le r\}$ are derived by Candès and Plan (2010) with a technique different from ours, which does not allow for further boundedness constraints on the subclass of at most rank-r matrices. Specifically, after reduction to a standard linear model their lower bound is obtained by passage to a Bayes risk with an unbounded support prior (Gaussian prior). Our lower bound is obtained on the intersection of Schatten-0 and Schatten-p balls. This is similar in spirit to Rigollet and Tsybakov (2010) establishing minimax lower bounds on the intersection of ℓ_0 and ℓ_1 balls for the vector sparsity scenario. In what follows, $\inf_{\hat{A}}$ denotes the infimum over all estimators based on $(X_1, Y_1), ..., (X_N, Y_N)$.

THEOREM 5 (Lower bound – Restricted Isometry). *Let $\xi_1, ..., \xi_N$ be i.i.d. $\mathcal{N}(0, \sigma^2)$ random variables for some $\sigma^2 > 0$. With $M := \max(m, T)$, let $M \ge 8$, $r \ge 1$, $\min(T, m) \ge r$ and assume that $\mathcal{L}$ satisfies the RI (r, ν)-condition for some $\delta_r \in (0, 1)$. Then for any $p \in (0, 1]$, $\Delta > 0$,*

$$(7.1) \qquad \inf_{\hat{A}} \sup_{\substack{A^* \in \mathbb{R}^{m\times T}: \\ \operatorname{rank}(A^*) \le r,\, \|A^*\|_{S_p} \le \Delta\,\nu\,\sigma}} \mathbb{P}_{A^*}\Big(\hat{d}_{2,N}(\hat{A}, A^*)^2 > C(\alpha)\,\sigma^2\,\psi_{M,N,r,\Delta}\Big) \;\ge\; \beta,$$

and

$$(7.2) \qquad \inf_{\hat{A}} \sup_{\substack{A^* \in \mathbb{R}^{m\times T}: \\ \operatorname{rank}(A^*) \le r,\, \|A^*\|_{S_p} \le \Delta\,\nu\,\sigma}} \mathbb{P}_{A^*}\Big(\|\hat{A} - A^*\|_{S_2}^2 > C(\alpha, \nu)\,\sigma^2\,\psi_{M,N,r,\Delta}\Big) \;\ge\; \beta,$$

for any $0 < \alpha < 1/8$, where $\beta = \beta(M, \alpha) > 0$ is such that $\beta(M, \alpha) \to 1$ as $M \to \infty$, $\alpha \to 0$,

$$\psi_{M,N,r,\Delta} \;=\; \min\Big(\frac{rM}{N},\, \Delta^p\Big(\frac{M}{N}\Big)^{1-p/2},\, \Delta^2\Big),$$

$$C(\alpha) \;=\; \frac{\alpha(1-\delta_r)^2}{(1+\delta_r)^2}\frac{\log 2}{128} \quad \textit{and} \quad C(\alpha, \nu) = \alpha\nu^2(1-\delta_r)^2\frac{\log 2}{128}\,.$$

REMARK 3. In view of the multi-task learning example, it is worth to note that $C(\alpha)$ and $\beta(M, \alpha)$ do not depend on the constant ν of the RI condition.

PROOF. Without loss of generality we assume that $M = m \geq T$. For some real constant γ and natural number $s \in \{1, 2, ..., r\}$ both to be specified later, define

$$\mathcal{A}_{s,\gamma} \;:=\; \Big\{A = (a_{ij}) \in \mathbb{R}^{m\times T} : a_{ij} \in \{0, \gamma\nu/\sqrt{N}\}\,\text{if}\,1 \leq j \leq s;\; a_{ij} = 0 \text{ otherwise}\Big\}.$$

By construction, any element of $\mathcal{A}_{s,\gamma}$ as well as the difference of any two elements of $\mathcal{A}_{s,\gamma}$ has rank at most s. Due to the Varshamov-Gilbert bound, there exists a subset $\mathcal{A}^0_{s,\gamma} \subset \mathcal{A}_{s,\gamma}$ of cardinality $\mathrm{Card}(\mathcal{A}^0_{s,\gamma}) \geq 2^{sm/8} + 1$ containing $A_0 = 0$ such that for any two distinct elements A_1 and A_2 of $\mathcal{A}^0_{s,\gamma}$,

$$\hat{d}_{2,N}(A_1, A_2)^2 \;\geq\; \nu^{-2}(1-\delta_r)^2\|A_1 - A_2\|^2_{S_2} \;\geq\; (1-\delta_r)^2\frac{\gamma^2}{8}\frac{sM}{N}, \tag{7.3}$$

where the first inequality follows from the left hand inequality in the Restricted Isometry condition and is only used to prove (7.1). We will only prove (7.1); the proof of (7.2) is analogous in view of (7.3).

For any $A \in \mathbb{R}^{m\times T}$, let $\mathbb{P}_A$ denote the probability distribution of $(Y_1, \dots, Y_N)$ satisfying (1.1) with $A^* = A$. Then, for any $A_1 \in \mathcal{A}^0_{s,\gamma}$, the Kullback-Leibler divergence $K(\mathbb{P}_{A_0}, \mathbb{P}_{A_1})$ between $\mathbb{P}_{A_0}$ and $\mathbb{P}_{A_1}$ satisfies

$$K(\mathbb{P}_{A_0}, \mathbb{P}_{A_1}) \;=\; \frac{N}{2\sigma^2}\hat{d}_{2,N}(A_0, A_1)^2 \;\leq\; \frac{\gamma^2}{2\sigma^2}(1+\delta_r)^2 sM \tag{7.4}$$

where we used again the RI condition. The condition

$$\frac{1}{\mathrm{Card}(\mathcal{A}^0_{s,\gamma}) - 1}\sum_{A\in\mathcal{A}^0_{s,\gamma}} K(\mathbb{P}_A, \mathbb{P}_{A_0}) \;\leq\; \alpha\log\left(\mathrm{Card}(\mathcal{A}^0_{s,\gamma}) - 1\right) \tag{7.5}$$

for $0 < \alpha < 1/8$ is satisfied in particular when $\gamma^2 \leq 2\alpha\sigma^2(\log 2)/(8(1+\delta_r)^2)$. Define

$$r_\Delta \;=\; \arg\min\Big\{l \in \mathbb{N} : \Delta^p \leq l\Big(\frac{M}{N}\Big)^{p/2}\Big\}.$$

The case $r_\Delta = 1$. Here, $\psi_{M,N,r,\Delta} = \Delta^2$, for every $r \geq 1$, and $\Delta^2 N/M \leq 1$. For $0 < \alpha < 1/8$ define

$$s_1 \;=\; 1 \;\text{ and }\; \gamma_1 \;=\; \Big(\frac{\alpha}{(1+\delta_r)^2}\frac{\log 2}{4}\sigma^2\Delta^2\frac{N}{M}\Big)^{1/2}.$$

Then $\|A\|_{S_p} \leq \|A\|_{S_2} \leq \sqrt{M/N}\nu\gamma \;\leq \sigma\nu\Delta$ for all $A \in \mathcal{A}_{1,\gamma_1}$, i.e. $\mathcal{A}_{1,\gamma_1}$ is contained in the set

$$\Big\{A \in \mathbb{R}^{m\times T} : \mathrm{rank}(A) \leq r,\; \|A\|_{S_p} \leq \Delta\nu\sigma\Big\}.$$

Now, inequality (7.3) shows that $\hat{d}_{2,N}(A_1, A_2)^2 \geq 4C(\alpha)\sigma^2\Delta^2$ for any two different elements $A_1, A_2 \in \mathcal{A}^0_{1,\gamma_1}$, while $\Delta^2 N/M \leq 1$ reveals that $\gamma_1^2 \leq 2\alpha\sigma^2(\log 2)/(8(1+\delta_r)^2)$. Hence condition (7.5) is satisfied.

The case $1 < r_\Delta \leq r$. In this case the rate $\psi_{M,N,r,\Delta}$ is equal to $\Delta^p(M/N)^{1-p/2}$. We consider the set $\mathcal{A}^0_{r_\Delta,\gamma_2}$ with some γ_2 to be specified below. For $A \in \mathcal{A}^0_{r_\Delta,\gamma_2}$, we have $\|A\|_{S_p}^2 \le r_\Delta^{2-p}\|A\|_{S_2}^2 \le r_\Delta^{2-p}\gamma_2^2\nu^2 r_\Delta M/N$, and using that $r_\Delta \le 2\Delta^p(N/M)^{p/2}$, it follows by a simple algebra that $\|A\|_{S_p} \le \sigma\nu\Delta$ whenever

$$2^{1+1/(2-p)}\gamma_2^{2/(2-p)} \;\le\; \left(\Delta\sqrt{N/M}\right)^{1-p}\sigma^{2/(2-p)}. \tag{7.6}$$

But $\Delta\sqrt{N/M} \ge 1$ because of $r_\Delta > 1$. Now define

$$s_2 \;=\; r_\Delta \;\text{ and }\; \gamma_2 \;=\; \left(\frac{\alpha}{(1+\delta_r)^2}\frac{\log 2}{4}\sigma^2\right)^{1/2}.$$

Then γ_2 fulfills condition (7.5) and the constraint (7.6) since $\alpha < 1/8$ and $(\log 2)/4 < 1$. Furthermore, $\mathcal{A}^0_{r_\Delta,\gamma_2}$ is a subset of matrices $A \in \mathbb{R}^{m\times T}$ with $\mathrm{rank}(A) \le r$ and $\|A\|_{S_p} \le \nu\sigma\Delta$. Finally, (7.3) implies that

$$\hat{d}_{2,N}(A_1,A_2)^2 \ge (1-\delta_r)^2\frac{\gamma_2^2}{8}\frac{r_\Delta M}{N} \ge (1-\delta_r)^2\frac{\gamma_2^2}{8}\Delta^p\Big(\frac{M}{N}\Big)^{1-p/2} = 4C(\alpha)\sigma^2\Delta^p\Big(\frac{M}{N}\Big)^{1-p/2}$$

for any two different matrices $A_1, A_2 \in \mathcal{A}_{r_\Delta,\gamma_2}$.

The case $r_\Delta > r$. Here, $\psi_{M,N,\Delta,r} = rM/N$. The required conditions follow immediately as above with $\mathcal{A}^0_{r,\gamma_3}$, where $\gamma_3^2 (= \gamma_2^2) = 2\alpha\sigma^2(\log 2)/(8(1+\delta_r)^2)$ (and $s_3 = r$).

The lower bound in these three cases as stated in the Theorem follows now by an application of Tsybakov (2009), Theorem 2.5. □

REMARK 4. Theorem 5 implies that the rates of convergence in Theorem 4 are optimal in a minimax sense on Schatten-p balls $\{A^* \in \mathbb{R}^{m\times T} : \|A^*\|_{S_p} \le \Delta\}$ under the RI condition. Indeed, using Theorem 5 with no restriction on the rank (i.e., when $r = \min(m,T)$), and putting for simplicity $\Delta = 1$, we find that the rate in the lower bound is of the order $\min\big(\min(m,T)M/N,\ (M/N)^{1-p/2},\ 1\big)$. For $m = T (= M)$ and $m^3 > N > m$ this minimum equals $(M/N)^{1-p/2}$, which coincides with the upper bound of Theorem 4.

The lower bound for the prediction error (7.1) in the above theorem does not apply to matrix completion with $N < mT$ since then the Restricted Isometry condition cannot be satisfied, as discussed in Section 3. However, the conditions for the second bound (7.2) are not prohibitive. Indeed, for (7.2) to hold, we only need the right hand inequality of the RI condition. The next theorem specifies the corresponding lower bound for USR matrix completion.

THEOREM 6 (Lower bound – USR matrix completion). *Let $\xi_1,\dots,\xi_N$ be i.i.d. $\mathcal{N}(0,\sigma^2)$ random variables for some $\sigma^2 > 0$. For $M := \max(m,T)$, let $M \ge 8$, $r \ge 1$, $\min(T,m) \ge r$, and consider the USR matrix completion model. Then for any $p \in (0,1]$, $\Delta > 0$,*

$$\inf_{\hat{A}} \sup_{\substack{A^*\in\mathbb{R}^{m\times T}:\\ \mathrm{rank}(A^*)\le r,\,\|A^*\|_{S_p}\le\Delta\sqrt{mT}\,\sigma}} \mathbb{P}_{A^*}\left(\frac{1}{mT}\|\hat{A}-A^*\|_{S_2}^2 > C'(\alpha)\,\sigma^2\,\psi_{M,N,r,\Delta}\right) \;\ge\; \beta,$$

for any $0 < \alpha < 1/8$ and $C'(\alpha) = \alpha(\log 2)/128$, where $\beta = \beta(M,\alpha)$ and $\psi_{M,N,r,\Delta}$ are as in Theorem 5.

PROOF. We proceed as in Theorem 5 for $\nu = \sqrt{mT}$, with the only difference in the bound on the Kullback-Leibler divergence. Indeed, for USR matrix completion, instead of (7.4) we have

$$K(\mathbb{P}_{A_0}, \mathbb{P}_{A_1}) = \frac{N}{2\sigma^2}\mathbb{E}\left(\hat{d}_{2,N}(A_0, A_1)^2\right) = \frac{N}{2\sigma^2 mT}\|A_0 - A_1\|_{S_2}^2 \leq \frac{\gamma^2 sM}{2\sigma^2}. \tag{7.7}$$

□

In particular, Theorem 6 with $\Delta = \infty$ shows that on the class of matrices of rank smaller than r the lower bound of estimation in the squared Frobenius norm for matrix completion is of the order rM/N. This agrees with the conjecture of optimality derived in Candès and Plan (2009) by a comparison to oracle heuristics. The upper bounds in Keshavan et al. (2009) are also of the same form, up to logarithmic factors. It is worthwhile to note that these papers considered noisy matrix completion not on the class of matrices of rank smaller than r but on some of its subclasses characterized by additional strong restrictions.

Furthermore, for collaborative filtering we can obtain a lower bound for the prediction error without the RI condition, as shows the next theorem. We will need a natural assumption that the observed noisy entries are sufficiently well dispersed, i.e., there there exist r rows or r columns with more that κMr observations for some fixed $\kappa \in (0, 1]$. We state the result with an additional constraint on the Frobenius norm of A^*, in order to be coherent with the upper bound (cf. Remark 2 in Section 5).

THEOREM 7 (Lower bounds - Collaborative filtering). *Let $\xi_1, ..., \xi_N$ be i.i.d. $\mathcal{N}(0, \sigma^2)$ random variables for some $\sigma^2 > 0$ and assume that the masks $X_1 = e_{i_1}(m)e'_{j_1}(T), ..., X_N = e_{i_N}(m)e'_{j_N}(T)$ are pairwise different. With $M = \max(m, T)$, suppose that $\kappa Mr \geq 8$ for some fixed $\kappa \in (0, 1]$, $\min(T, m) \geq r$ and assume furthermore that there exist numbers $1 \leq k_1 < ... < k_r \leq T$ or $1 \leq k'_1 < ... < k'_r \leq m$ such that $\{(i_1, j_1), ..., (i_N, j_N)\} \cap \{(i, k_1), ..., (i, k_r) : i = 1, ..., m\}$ or $\{(i_1, j_1), ..., (i_N, j_N)\} \cap \{(k'_1, j), ..., (k'_r, j) : j = 1, ..., T\}$ has cardinality at least $\kappa Mr + 1$. Define $\mathcal{C}_{\delta,r} = \{A \in \mathbb{R}^{m\times T} : \mathrm{rank}(A) \leq r$ and $\|A\|_{S_2} \leq \delta\}$. Then for any $0 < \alpha < 1/8$ and $\delta^2 \geq \alpha\sigma^2(\log 2)(\kappa Mr + 1)/4$,*

$$\inf_{\hat{A}} \sup_{A^*\in\mathcal{C}_{\delta,r}} \mathbb{P}_{A^*}\left(\hat{d}_{2,N}(\hat{A}, A^*)^2 > C'(\alpha)\,\frac{\sigma^2\kappa rM}{N}\right) \geq \beta\Big(\kappa M, \alpha\Big) > 0,$$

with a function $\beta \to 1$ as $\kappa M \to \infty$, $\alpha \to 0$, and $C'(\alpha) = \alpha(\log 2)/128$.

To prove this theorem, it suffices to repeat the argument in case $r_\Delta > r$ of the proof of Theorem 5 with a minor modification. The "dispersion" condition allows to use a similar construction as in the previous proof to choose now a subset $\mathcal{A}_0$ of matrices with rank at most r and log-cardinality of the order κMr and with prediction loss $\hat{d}_{2,N}(A, B)$ lower bounded by the order $\sqrt{\kappa Mr/N}$ for any two different elements A, B of $\mathcal{A}_0$.

8. Control of the stochastic term.

We consider two approaches for bounding the stochastic term $N^{-1}\sum_{i=1}^N \xi_i \mathrm{tr}\big((\hat{A} - A^*)'X_i\big)$ on the right hand side of the basic inequality (3.1). The first one used for $p = 1$ consists in application of the trace duality

$$\left| \frac{1}{N} \sum_{i=1}^{N} \xi_i \mathrm{tr}\Big((\hat{A} - A^*)' X_i \Big) \right| \;\le\; \| \hat{A} - A^* \|_{S_1} \| \mathbf{M} \|_{S_\infty} \tag{8.1}$$

with $\mathbf{M} = N^{-1} \sum_{i=1}^{N} \xi_i X_i$ and then of suitable exponential bounds for the spectral norm of $\mathbf{M}$ under different conditions on X_i, $i = 1, ..., N$. The second approach used to treat the case $0 < p < 1$ (non-convex penalties) is based on refined empirical process techniques, see Section 8.2. All proofs of this section are deferred to Section 10.

8.1. *Tail bounds for the spectral norm of random matrices.* We say that the random variables ξ_i, $i = 1, \ldots, N$, satisfy the *Bernstein condition* if

$$\max_{1 \le i \le N} \mathbb{E}|\xi_i|^l \;\le\; \frac{1}{2} l!\, \sigma^2 H^{l-2}, \quad l = 2, 3 \ldots, \tag{8.2}$$

with some finite constants σ and H.

The random variables ξ_i, $i = 1, \ldots, N$, are said to satisfy the *light tail condition* if

$$\max_{1 \le i \le N} \mathbb{E}\Big(\exp(\xi_i^2/\sigma^2) \Big) \;\le \exp(1) \tag{8.3}$$

for some positive constant σ^2.

LEMMA 1. *Let the i.i.d. zero-mean random variables ξ_i satisfy the Bernstein condition (8.2). Let also either*

$$\max_{1 \le j \le m} \frac{1}{N} \sum_{i=1}^{N} |X_{i(j,\cdot)}|_2^2 \;\le\; S_{row}^2 \quad \textit{and} \tag{8.4}$$

$$\max_{1 \le j \le m, 1 \le i \le N} |X_{i(j,\cdot)}|_2 \;\le\; H_{row} \tag{8.5}$$

or the conditions

$$\max_{1 \le k \le T} \frac{1}{N} \sum_{i=1}^{N} |X_{i(\cdot,k)}|_2^2 \;\le\; S_{col}^2 \quad \textit{and} \tag{8.6}$$

$$\max_{1 \le k \le T, 1 \le i \le N} |X_{i(\cdot,k)}|_2 \;\le\; H_{col} \tag{8.7}$$

hold true with some constants $S_{row}, H_{row}, S_{col}, H_{col}$. Let $D > 1$. Then, respectively, with probability at least $1 - 2/m^{D-1}$ or at least $1 - 2/T^{D-1}$ we have

$$\| \mathbf{M} \|_{S_\infty} \;\le\; \tau \tag{8.8}$$

where $\tau = \tau_{row} = C_{row}\sqrt{m(\log m)/N}$ if (8.4) and (8.5) are satisfied or $\tau = \tau_{col}$ $= C_{col}\sqrt{T(\log T)/N}\}$ if (8.6) and (8.7) hold. Here

$$C_{row} = \left(\sqrt{2D\sigma^2 S_{row}^2} + 2DH_{row}H\sqrt{\frac{\log m}{N}} \right),\; C_{col} = \left(\sqrt{2D\sigma^2 S_{col}^2} + 2DH_{col}H\sqrt{\frac{\log T}{N}} \right).$$

LEMMA 2. *Let $\xi_1, ..., \xi_N$ be i.i.d. $\mathcal{N}(0, \sigma^2)$ random variables. Then*

$$\|\mathbf{M}\|_{S_\infty} \;\le\; 8\sigma\phi_{\max}(1)\sqrt{\frac{m+T}{N}} \;=:\tau_1 \tag{8.9}$$

with probability at least $1-2\exp\{-(2-\log 5)(m+T)\}$, *where* $\phi_{\max}(1)$ *is the maximal rank-1 eigenvalue of the sampling operator* $\mathcal{L}$.

The assumptions of Lemma 1 imply that $\phi_{\max}(1)$ is bounded by a constant independent of m, T and N, cf. (2.3). If m and T have the same order of magnitude, the bound of Lemma 2 is better, since it does not contain extra logarithmic factors. On the other hand, if m and T differ dramatically, for example, $m \gg T$, then Lemma 1 can provide a significant improvement. Indeed, the "column" version of Lemma 1 guarantees the rate $\tau \sim \sqrt{T\log T}/\sqrt{N}$ which in this case is much smaller than $\sqrt{m/N}$. In all the cases, the concentration rate in Lemma 2 is exponential and thus faster than in Lemma 1.

We now give bounds on the stochastic term for USR matrix completion problem and collaborative filtering. Candès and Plan (2009) assume the locations of the observed noisy entries in the matrix completion setting selected at random *without* replacement. The following lemma allows to treat matrix completion with random noise under different conditions than in Candès and Plan (2009), and shows that there are some unusual effects.

LEMMA 3 (USR matrix completion). *(i) Let the i.i.d. zero-mean random variables ξ_i satisfy the Bernstein condition (8.2). Consider the USR matrix completion problem and assume that $mT(m+T) > N$. Then, for any $D \ge 2$ and*

$$\|\mathbf{M}\|_{S_\infty} \;\le\; (4\sigma\sqrt{10D} + 8HD)\frac{m+T}{N} \;=:\; \tau_2 \tag{8.10}$$

with probability at least $1 - 4\exp\{-(2-\log 5)(m+T)\}$.

(ii) Assume that the i.i.d. zero-mean random variables ξ_i satisfy the light tail condition (8.3) for some $\sigma^2 > 0$. Then for any $B > 0$,

$$\|\mathbf{M}\|_{S_\infty} \;\le\; \sqrt{B}\,\frac{\sigma\log\big(\max(m+1, T+1)\big)}{\sqrt{N}} \;=:\; \tau_3 \tag{8.11}$$

with probability at least $1 - (1/C)\max(m+1, T+1)^{-CB}$ *for some constant* $C > 0$ *which does not depend on m, T and N.*

The proof of part (i) is based on a refinement of a technique in Vershynin (2007), whereas that of part (ii) follows immediately from the large deviations inequality of Nemirovski (2004). For example, if $\xi_i \sim \mathcal{N}(0, \sigma^2)$, in which case both results apply, the bound (ii) is tighter than (i) for sample sizes $N \ll (m+T)^2$ which is the most interesting case for matrix completion.

Slightly surprising, much tighter bounds are available when the X_i are forced to be pairwise different. Besides it is noteworthy that the rates in (8.12) and (8.13) below are different for Gaussian and Bernstein errors.

LEMMA 4 (Collaborative filtering). *Consider the problem of collaborative filtering with random or deterministic $X_1, \ldots, X_N$ which are independent from $\xi_1, \ldots, \xi_N$.*

(i) Let $\xi_1, ..., \xi_N$ be i.i.d. $\mathcal{N}(0, \sigma^2)$ random variables. Then, for any $D \geq 2$,

$$\|\mathbf{M}\|_{S_\infty} \leq 8\sigma\sqrt{D}\,\frac{\sqrt{m+T}}{N} =: \tau_4 \tag{8.12}$$

with probability at least $1 - 2\exp\{-(D - \log 5)(m+T)\}$.

(ii) Let $\xi_1, ..., \xi_N$ be i.i.d. zero-mean random variables satisfying the Bernstein condition (8.2). Then, for any $D \geq 2$ and

$$\|\mathbf{M}\|_{S_\infty} \leq \frac{4\sigma\sqrt{2D(m+T)} + 8HD(m+T)}{N} =: \tau_5 \tag{8.13}$$

with probability at least $1 - 2\exp\{-(D - \log 5)(m+T)\}$.

(iii) Let $\xi_1, ..., \xi_N$ be i.i.d. $\mathcal{N}(0, \sigma^2)$ random variables. Then for any $A > 1$,

$$\|\mathbf{M}\|_{S_\infty} \leq \frac{\sqrt{2A}}{N}\max\left\{\Big\|\sum_{i=1}^N X_i' X_i\Big\|_{S_\infty}^{1/2}, \Big\|\sum_{i=1}^N X_i X_i'\Big\|_{S_\infty}^{1/2}\right\}\sigma\sqrt{\log(m+T)} =: \tau_6$$

with probability at least $1 - 2(m+T)^{1-A}$.

Since the masks X_i are distinct, the maximum appearing in (iii) is bounded by $\sqrt{\max(m,T)}$; in case it is attained, the bound (8.12) is slightly stronger since it is free from the logarithmic factor. For $N \ll mT$ the tightness of the bound in (iii) depends strongly on the geometry of the X_i's and the maximum can be significantly smaller than $\sqrt{\max(m,T)}$. Note also that the concentration in (8.12) is exponential, while it is only polynomial in (iii).

8.2. *Concentration bounds for the stochastic term under non-convex penalties.* The last bound in this section applies in the case $0 < p < 1$. It is given in the following lemma.

LEMMA 5. *Let $\xi_1, ..., \xi_N$ be i.i.d. $\mathcal{N}(0, \sigma^2)$ random variables, $0 < p < 1$ and $M = \max(m,T)$. Assume that the sampling operator $\mathcal{L}$ is uniformly bounded, cf. (2.4). Set $c_\kappa = (2\kappa - 1)(2\kappa)\kappa^{-1/(2\kappa-1)}$ where $\kappa = (2-p)/(2-2p)$. Then for any fixed $\delta > 0$, $\vartheta \geq C^2$ and $\tau_7 = c_\kappa(\vartheta/p)^{1-p/2}(M/N)^{1-p/2}$ we have*

$$\Big|\frac{1}{N}\sum_{i=1}^N \xi_i \mathrm{tr}(X_i'(\hat{A} - A^*))\Big| \leq \frac{\delta}{2}\hat{d}_{2,N}(\hat{A}, A^*)^2 + \tau_7\,\delta^{p-1}\,\|\hat{A} - A^*\|_{S_p}^p \tag{8.14}$$

with probability at least $1 - C\,\exp(-\vartheta M/C^2)$ for some constant $C = C(p, c_0, \sigma^2) > 0$ which is independent of M and N and satisfies $\sup_{0<p\leq q} C(p, c_0, \sigma) < \infty$ for all $q < 1$.

Note at this point that we cannot rely the proof of Lemma 5 directly on the trace duality and norm interpolation (cf. Lemma 11), i.e., on the inequalities

$$\begin{aligned}\Big|\frac{1}{N}\sum_{i=1}^N \xi_i \mathrm{tr}(X_i'(\hat{A} - A^*))\Big| &\leq \|\hat{A} - A^*\|_{S_1}\|\mathbf{M}\|_{S_\infty} \\ &\leq \|\hat{A} - A^*\|_{S_2}^{1-\frac{p}{2-p}}\|\hat{A} - A^*\|_{S_p}^{\frac{p}{2-p}}\|\mathbf{M}\|_{S_\infty}. \end{aligned} \tag{8.15}$$

Indeed, one may think that we could have bounded here the S_∞-norm of $\mathbf{M}$ in the same way as in Section 8.1, and then the proof would be complete after suitable decoupling if we were able to bound from above $\|\hat{A} - A^*\|_{S_2}^2$ by $\hat{d}_{2,N}(\hat{A}, A^*)^2$ times a constant factor. However, this is not possible. Even the Restricted Isometry condition cannot help here because $\hat{A} - A^*$ is not necessarily of small rank. Nevertheless, we will show that by other techniques it is possible to derive an inequality similar to (8.15) with $\hat{d}_{2,N}(\hat{A}, A^*)$ instead of $\|\hat{A} - A^*\|_{S_2}$. Further details are given in Sections 10 and 11.

9. Proof of Theorem 2.

Preliminaries. We first give two lemmas on matrix decomposition needed in our proof, which are essentially provided by Recht, Fazel and Parrilo (2008) (subsequently, RFP(08) for short).

LEMMA 6. *Let A and B be matrices of the same dimension. If $AB' = 0$, $A'B = 0$, then*

$$\|A + B\|_{S_p}^p \;=\; \|A\|_{S_p}^p \;+\; \|B\|_{S_p}^p, \quad \forall\, p > 0.$$

PROOF. For $p = 1$ the result is Lemma 2.3 in RFP(08). The argument obviously extends to any $p > 0$ since RFP(08) show that the singular values of $A+B$ are equal to the union (with repetition) of the singular values of A and B. □

LEMMA 7. *Let $A \in \mathbb{R}^{m\times T}$ with $\mathrm{rank}(A) = r$ and singular value decomposition $A = U\Lambda V'$. Let $B \in \mathbb{R}^{m\times T}$ be arbitrary. Then there exists a decompositon $B = B_1 + B_2$ with the following properties:*

(i) $\mathrm{rank}(B_1) \le 2\mathrm{rank}(A) = 2r$,

(ii) $AB_2' = 0$, $A'B_2 = 0$,

(iii) $\mathrm{tr}(B_1'B_2) = 0$.

(iv) B_1 and B_2 are of the form

$$B_1 \;=\; U\begin{pmatrix} \tilde{B}_{11} & \tilde{B}_{12} \\ \tilde{B}_{21} & 0 \end{pmatrix}V' \quad \textit{and} \quad B_2 \;=\; U\begin{pmatrix} 0 & 0 \\ 0 & \tilde{B}_{22} \end{pmatrix}V', \quad \textit{with } \tilde{B}_{11} \in \mathbb{R}^{r\times r}.$$

The points (i)-(iii) are the statement of Lemma 3.4 in RFP(08), the representation (iv) is provided in its proof.

Proof of Theorem 2. First note that there exists a decomposition $\hat{A} = \hat{A}^{(1)} + \hat{A}^{(2)}$ with the following properties:

(i) $\mathrm{rank}(\hat{A}^{(1)} - A^*) \le 2\,\mathrm{rank}(A^*) = 2r$,

(ii) $A^*(\hat{A}^{(2)})' = 0$, $(A^*)'\hat{A}^{(2)} = 0$,

(iii) $\mathrm{tr}\big((\hat{A}^{(1)} - A^*)'\hat{A}^{(2)}\big) = 0$.

This follows from Lemma 7 with $A = A^*$ and $B = \hat{A} - A^*$. In the notation of Lemma 7 we have $B_1 = \hat{A}^{(1)} - A^*$ and $B_2 = \hat{A}^{(2)}$.

From the basic inequality (3.1) and (3.2) with $\delta = 1/2$ we find

$$(1 - I_{\{0<p<1\}}/2)\hat{d}_{2,N}(\hat{A}, A^*)^2 \leq 2^{2-p}\tau\|\hat{A} - A^*\|_{S_p}^p + 4\tau\Big(\|A^*\|_{S_p}^p - \|\hat{A}\|_{S_p}^p\Big). \tag{9.1}$$

In particular, for the case $p = 1$

$$\hat{d}_{2,N}(\hat{A}, A^*)^2 \leq 2\tau\|\hat{A} - A^*\|_{S_p}^p + 4\tau\Big(\|A^*\|_{S_p}^p - \|\hat{A}\|_{S_p}^p\Big). \tag{9.2}$$

For brevity, we will conduct the proof with the numerical constants given in (9.2), i.e., with those for $p = 1$. The proof for general p differs only in the values of the constants, but their expressions become cumbersome.

Using (2.1), we get

$$\hat{d}_{2,N}(\hat{A}, A^*)^2 \leq 2\tau\|\hat{A}^{(1)} - A^*\|_{S_p}^p + 4\tau\|A^*\|_{S_p}^p + 2\tau\|\hat{A}^{(2)}\|_{S_p}^p - 4\tau\|\hat{A}\|_{S_p}^p. \tag{9.3}$$

By (2.1) again and by Lemma 6,

$$\begin{aligned}\|\hat{A}\|_{S_p}^p &\geq \|A^* + \hat{A}^{(2)}\|_{S_p}^p - \|\hat{A}^{(1)} - A^*\|_{S_p}^p \\ &= \|A^*\|_{S_p}^p + \|\hat{A}^{(2)}\|_{S_p}^p - \|\hat{A}^{(1)} - A^*\|_{S_p}^p,\end{aligned}$$

since $(A^*)'\hat{A}^{(2)} = 0$ and $A^*(\hat{A}^{(2)})' = 0$ by construction. Together with (9.3) this yields

$$\hat{d}_{2,N}(\hat{A}, A^*)^2 \leq 2\tau\|\hat{A}^{(1)} - A^*\|_{S_p}^p - 2\tau\|\hat{A}^{(2)}\|_{S_p}^p + 4\tau\|\hat{A}^{(1)} - A^*\|_{S_p}^p, \tag{9.4}$$

from which one may deduce

$$\hat{d}_{2,N}(\hat{A}, A^*)^2 \leq 6\tau\|\hat{A}^{(1)} - A^*\|_{S_p}^p \quad \text{and} \tag{9.5}$$

$$\|\hat{A}^{(2)}\|_{S_p}^p \leq 3\|\hat{A}^{(1)} - A^*\|_{S_p}^p. \tag{9.6}$$

Consider now the following decomposition of the matrix $\hat{A}^{(2)}$. First recall that $\hat{A}^{(2)}$ is of the form

$$\hat{A}^{(2)} = U\begin{pmatrix} 0 & 0 \\ 0 & \tilde{B}_{22}\end{pmatrix}V'.$$

Write $\tilde{B}_{22} = W_1\Lambda(\tilde{B}_{22})W_2'$ with diagonal matrix $\Lambda(\tilde{B}_{22})$ of dimension r' and $W_1'W_1 = W_2'W_2 = I_{r'\times r'}$ for some $r' \leq \min(m, T)$. In the next step, W_1 and W_2 are complemented to orthogonal matrices $\bar{W}_1$ and $\bar{W}_2$ of dimension $\min(m,T) \times \min(m,T)$. For instance, set

$$\bar{W}_2' = \begin{pmatrix} & & 0 \\ * & & \\ & & W_2'\end{pmatrix} \in \mathbb{R}^{\min(m,T)\times\min(m,T)},$$

where $*$ complements the columns of the matrix $\binom{0}{W_2'}$ to an orthonormal basis in $\mathbb{R}^{m\times T}$, and proceed analogously with W_1. In particular, $\bar{W}_1'\bar{W}_1 = \bar{W}_2'\bar{W}_2 = I_{\min(m,T)\times\min(m,T)}$. Also

$$\hat{A}^{(2)} = U\begin{pmatrix} 0 & 0 \\ 0 & W_1\Lambda(\tilde{B}_{22})W_2'\end{pmatrix}V' = U\bar{W}_1\begin{pmatrix} 0 & 0 \\ 0 & \Lambda(\tilde{B}_{22})\end{pmatrix}\bar{W}_2'V' =: U\bar{W}_1 D\bar{W}_2'V'.$$

We now represent $\hat{A}^{(2)}$ as a finite sum of matrices $\hat{A}^{(2)} = \sum_{j=1}^{R'} \hat{A}_j^{(2)}$ with

$$\hat{A}_i^{(2)} \;=\; U\bar{W}_1 D_i \bar{W}_2' V'$$

and

$$D_i \;=\; \begin{pmatrix} 0 & 0 \\ 0 & \Lambda_i \end{pmatrix}$$

where the $r' \times r'$ diagonal matrix Λ_i has the form $\Lambda_i = \mathrm{diag}(\lambda_j I_{\{j \in I_i\}})$, $i \geq 1$. We denote here by I_1 the set of ar indices from $\{1, \ldots, \min(m, T)\}$ corresponding to the ar largest in absolute value diagonal entries of Λ, by I_2 the set of indices corresponding to the next ar largest in absolute value diagonal entries λ_j, etc. Clearly, the matrices $\hat{A}_k^{(2)}$ are mutually orthogonal: $\mathrm{tr}\big((\hat{A}_j^{(2)})'\hat{A}_k^{(2)}\big) = 0$ for $j \neq k$ and $\mathrm{rank}(\hat{A}_j^{(2)}) \leq ar$. Moreover, $\hat{A}_i^{(2)}$ is orthogonal to $\hat{A}^{(1)} - A^*$.

Let $\sigma_1 \geq \sigma_2 \geq \ldots$ be the singular values of $\hat{A}^{(2)}$, then $\sigma_1 \geq \ldots \geq \sigma_{ar}$ are the singular values of $\hat{A}_1^{(2)}$, $\sigma_{ar+1} \geq \ldots \geq \sigma_{2ar}$ those of $\hat{A}_2^{(2)}$, etc. By construction, we have $\mathrm{Card}(I_i) = ar$ for all i, and for all $k \in I_{i+1}$:

$$\sigma_k \;\leq\; \min_{j \in I_i} \sigma_j \;\leq\; \Big(\frac{1}{ar} \sum_{j \in I_i} \sigma_j^p\Big)^{1/p}.$$

Thus,

$$\sum_{k \in I_{i+1}} \sigma_k^2 \;\leq\; ar\Big(\frac{1}{ar} \sum_{j \in I_i} \sigma_j^p\Big)^{2/p}$$

from which one can deduce for all $j \geq 2$:

$$\|\hat{A}_j^{(2)}\|_{S_2} \;=\; \Big(\sum_{k \in I_j} \sigma_k^2\Big)^{1/2} \;\leq\; (ar)^{\frac{1}{2}-\frac{1}{p}}\Big(\sum_{k \in I_{j-1}} \sigma_k^p\Big)^{1/p} \;=\; (ar)^{\frac{1}{2}-\frac{1}{p}} \|\hat{A}_{j-1}^{(2)}\|_{S_p}$$

and consequently

$$\sum_{j \geq 2} \|\hat{A}_j^{(2)}\|_{S_2} \;\leq\; (ar)^{\frac{1}{2}-\frac{1}{p}} \sum_{j \geq 1} \|\hat{A}_j^{(2)}\|_{S_p}$$

Because of the elementary inequality $x^{1/p} + y^{1/p} \leq (x+y)^{1/p}$ for any non-negative x, y and $0 < p \leq 1$,

$$\sum_{j \geq 2} \|\hat{A}_j^{(2)}\|_{S_p} \;=\; \sum_{j \geq 2} \Big(\sum_{k \in I_j} \sigma_k^p\Big)^{1/p} \;\leq\; \Big(\sum_{j \geq 2} \sum_{k \in I_j} \sigma_k^p\Big)^{1/p} \;\leq\; \Big(\sum_k \sigma_k^p\Big)^{1/p} \;=\; \|\hat{A}^{(2)}\|_{S_p}.$$

Therefore,

$$\begin{aligned}
\sum_{j \geq 2} \|\hat{A}_j^{(2)}\|_{S_2} &\leq (ar)^{\frac{1}{2}-\frac{1}{p}} \|\hat{A}^{(2)}\|_{S_p} \\
&\leq 3^{1/p}(ar)^{\frac{1}{2}-\frac{1}{p}} \|\hat{A}^{(1)} - A^*\|_{S_p} \quad \text{(using inequality (9.6))} \\
&\leq 3^{1/p}(ar)^{\frac{1}{2}-\frac{1}{p}}(2r)^{\frac{1}{p}-\frac{1}{2}} \|\hat{A}^{(1)} - A^*\|_{S_2},
\end{aligned}$$

whereby the last inequality results from $\mathrm{rank}(\hat{A}^{(1)} - A^*) \leq 2r$ and

$$\Big(\frac{1}{2r} \sum_{k \leq 2r} \sigma_k^p\Big)^{1/p} \;\leq\; \Big(\frac{1}{2r} \sum_{k \leq 2r} \sigma_k^2\Big)^{1/2}.$$

Finally,

$$\sum_{j\geq 2}\|\hat A_j^{(2)}\|_{S_2} \;\leq\; 3^{1/p}\Big(\frac{a}{2}\Big)^{\frac12-\frac1p}\|\hat A^{(1)}-A^*\|_{S_2}. \tag{9.7}$$

We now proceed with the final argument. First note that $\mathrm{rank}\big((\hat A^{(1)}-A^*)+\hat A_1^{(2)}\big)\leq (2+a)r$. Next, by the triangular inequality, the restricted isometry condition and the orthogonality of $\hat A_j^{(2)}$ and $\hat A^{(1)}-A^*$ we obtain

$$\begin{aligned}
\nu\hat d_{2,N}(\hat A,A^*) \;&=\; \nu\,|\,\mathcal{L}(\hat A-A^*)\,|_2\\
&\geq\; \nu\,|\,\mathcal{L}(\hat A^{(1)}-A^*+\hat A_1^{(2)})\,|_2-\nu\sum_{j\geq 2}|\,\mathcal{L}(\hat A_j^{(2)})\,|_2\\
&\geq\; (1-\delta_{(2+a)r})\,\|\hat A^{(1)}-A^*+\hat A_1^{(2)}\|_{S_2} \;-\; (1+\delta_{ar})\sum_{j\geq 2}\|\hat A_j^{(2)}\|_{S_2}\\
&\geq\; \|\hat A^{(1)}-A^*\|_{S_2}\Big((1-\delta_{(2+a)r})-(1+\delta_{ar})3^{1/p}\Big(\frac{a}{2}\Big)^{\frac12-\frac1p}\Big). &(9.8)
\end{aligned}$$

Define

$$a \;=\; a(p) \;=\; \min\Big\{k\in\mathbb{N}:\, k>(6^{1/p}/\sqrt2)^{\frac{2p}{2-p}}\Big\}.$$

Then $1-3^{1/p}(a/2)^{\frac12-\frac1p}>0$. Now, $\delta_{(2+a)r}\geq\delta_{ar}$, and thus

$$(1-\delta_{(2+a)r})-(1+\delta_{ar})3^{1/p}\Big(\frac{a}{2}\Big)^{\frac12-\frac1p} \;\geq\; \Big(1-3^{1/p}\Big(\frac{a}{2}\Big)^{\frac12-\frac1p}\Big)-2\delta_{(2+a)r} \;>\; 0$$

whenever

$$\delta_{(2+a)r} \;<\; \frac12\Big(1-3^{1/p}\Big(\frac{a}{2}\Big)^{\frac12-\frac1p}\Big). \tag{9.9}$$

In case of (9.9), there exists a universal constant $\kappa=\kappa(p)$ such that

$$\nu^2\hat d_{2,N}(\hat A,A^*)^2 \;\geq\; \kappa\|\hat A^{(1)}-A^*\|_{S_2}^2. \tag{9.10}$$

Now, the inequalities (9.5) and (9.10) yield

$$\kappa\|\hat A^{(1)}-A^*\|_{S_2}^2 \;\leq\; 6\tau\nu^2\|\hat A^{(1)}-A^*\|_{S_p}^p \;\leq\; 6\tau\nu^2(2r)^{1-p/2}\|\hat A^{(1)}-A^*\|_{S_2}^p, \tag{9.11}$$

where the latter inequality results from the fact that we have $\mathrm{rank}(\hat A^{(1)}-A^*)\leq 2r$, which implies

$$\|\hat A^{(1)}-A^*\|_{S_p} \;\leq\; (2r)^{1-p/2}\|\hat A^{(1)}-A^*\|_{S_2}. \tag{9.12}$$

From (9.10) and (9.11) we obtain

$$\kappa\|\hat A^{(1)}-A^*\|_{S_2}^{2-p} \;\leq\; 6\tau\nu^2(2r)^{1-p/2}. \tag{9.13}$$

Furthermore, from (9.5), (9.12) and (9.13) we find

$$\hat d_{2,N}(\hat A,A^*)^2 \;\leq\; 6\tau(2r)^{1-p/2}\|\hat A^{(1)}-A^*\|_{S_2}^p\leq 2r(6\tau)^{\frac{2}{2-p}}\kappa^{-\frac{p}{2-p}}\nu^{\frac{2p}{2-p}}. \tag{9.14}$$

This proves (3.4). It remains to prove (3.5). We first demonstrate (3.5) for $q=2$, then for $q=p$, and finally obtain (3.5) for all $q\in[p,2]$ by Schatten norm interpolation.

Using (9.7), (9.8), (9.14), we find

$$
\begin{aligned}
(1-\delta_{(2+a)r})\,\|\hat{A}^{(1)}-A^*+\hat{A}^{(2)}_1\|_{S_2} &\le \nu\hat{d}_{2,N}(\hat{A},A^*)+(1+\delta_{ar})\sum_{j\ge 2}\|\hat{A}^{(2)}_j\|_{S_2}\\
&\le C\sqrt{r}\,\tau^{\frac{1}{2-p}}\nu^{\frac{2}{2-p}}
\end{aligned}
$$

for some constant $C=C(p)>0$. This and again (9.7) yield

$$
\|\hat{A}-A^*\|_{S_2}\le\|\hat{A}^{(1)}-A^*+\hat{A}^{(2)}_1\|_{S_2}+\sum_{j\ge 2}\|\hat{A}^{(2)}_j\|_{S_2}\le C'\sqrt{r}\,\tau^{\frac{1}{2-p}}\nu^{\frac{2}{2-p}}
$$

for some constant $C'=C'(p)>0$. Thus, we have proved (3.5) for $q=2$. Next, using inequalities (2.1) and (9.6) we obtain

$$
\|\hat{A}-A^*\|^p_{S_p}\le\|\hat{A}^{(1)}-A^*\|^p_{S_p}+\|\hat{A}^{(2)}\|^p_{S_p}\le 4\|\hat{A}^{(1)}-A^*\|^p_{S_p}.
$$

Combining this with (9.12) and (9.13) we get (3.5) for $q=p$. Finally, (3.5) for arbitrary $q\in[p,2]$ follows from the norm interpolation formula

$$
\|A\|^q_{S_q}\le\|A\|_{S_p}^{\frac{p(2-q)}{2-p}}\,\|A\|_{S_2}^{\frac{2(q-p)}{2-p}},
$$

cf. Lemma 11 of Section 11 with $\theta=\frac{p(2-p)}{q(2-q)}$. □

10. Proofs of the Lemmas.

PROOF OF LEMMA 1. First observe that

$$
\|\mathbf{M}\|_{S_\infty}\;=\;\sup_{\substack{u\in\mathbb{R}^T:\\|u|_2=1}}|\mathbf{M}u|_2\;\le\;\sqrt{m}\max_{1\le j\le m}\sup_{\substack{u\in\mathbb{R}^T:\\|u|_2=1}}|u'\bar{\eta}_j|,
$$

with vectors $\bar{\eta}_j=N^{-1}\sum_{i=1}^N\xi_iX_{i(j,\cdot)}$. Consequently, for any $t>0$,

$$
\begin{aligned}
\mathbb{P}\Big(\|\mathbf{M}\|_{S_\infty}\;\ge\;t\sqrt{\frac{m\log m}{N}}\Big)\;&\le\;\mathbb{P}\Big(\sqrt{m}\max_{1\le j\le m}|\bar{\eta}_j|_2\ge t\sqrt{\frac{m\log m}{N}}\Big)\\
&\le\;m\max_{1\le j\le m}\mathbb{P}\Big(|\bar{\eta}_j|_2\ge t\sqrt{\frac{\log m}{N}}\Big).
\end{aligned}
$$

To proceed with the evaluation of the latter probability we use the following concentration bound (Pinelis and Sakhanenko, 1985).

LEMMA 8. *Let $\zeta_1,\dots,\zeta_N$ be independent zero mean random variables in a separable Hilbert space $\mathcal{H}$ such that*

$$
\sum_{i=1}^N\mathbb{E}\|\zeta_i\|^l_{\mathcal{H}}\;\le\;\frac{1}{2}l!\,B^2L^{l-2},\quad l=2,3\dots, \tag{10.1}
$$

with some finite constants $B,L>0$. Then

$$
\mathbb{P}\Big(\Big\|\sum_{i=1}^N\zeta_i\Big\|_{\mathcal{H}}\;\ge\;x\Big)\;\le\;2\exp\Big(-\frac{x^2}{2B^2+2xL}\Big),\quad\forall\,x>0.
$$

Setting $\zeta_i = \xi_i X_{i(j,\cdot)}$, $\mathcal{H} = \mathbb{R}^T$, note first that, by the Bernstein condition (8.2),

$$\begin{aligned}\sum_{i=1}^N \mathbb{E}\|\zeta_i\|_{\mathcal{H}}^l &= \mathbb{E}|\xi_i|^l \sum_{i=1}^N |X_{i(j,\cdot)}|_2^l \\ &\le \frac{1}{2} l!\, \sigma^2 H^{l-2}\Big(\max_j \sum_{i=1}^N |X_{i(j,\cdot)}|_2^2\Big) \max_{i,j} |X_{i(j,\cdot)}|_2^{l-2} \\ &\le \frac{1}{2} l!\, B^2 L^{l-2},\end{aligned}$$

where $B^2 = \sigma^2 S_{row}^2 N$ and $L = H_{row} H$, i.e., condition (10.1) is satisfied. Now an application of Lemma 8 yields for any $t > 0$

$$\begin{aligned}\mathbb{P}\Big(|\bar{\eta}_j|_2 \ \ge\ t\sqrt{\frac{\log m}{N}}\Big) &= \mathbb{P}\Big(\Big|\frac{1}{\sqrt{N}}\sum_{i=1}^N \xi_i\Big|_2 \ >\ t\sqrt{\log m}\Big) \\ &\le 2\exp\Big(-\frac{N(\log m)t^2}{2B^2 + 2tL\sqrt{N\log m}}\Big) \\ &= 2\exp\Big(-\frac{N(\log m)t^2}{2\sigma^2 S_{row}^2 N + 2tL\sqrt{N\log m}}\Big).\end{aligned}$$

Define $t = \sqrt{2D\sigma^2 S_{row}^2} + 2DL\sqrt{\frac{\log m}{N}}$ for some $D > 1$. Then

$$\frac{t^2}{\bar{B} + \bar{L}t} \ \ge\ D, \quad \text{where } \bar{B} = 2\sigma^2 S_{row}^2,\ \bar{L} = 2L\sqrt{\frac{\log m}{N}}.$$

With this choice of t,

$$\mathbb{P}\Big(|\bar{\eta}_j|_2 \ \ge\ t\sqrt{\frac{\log m}{N}}\Big) \ \le\ 2\exp\big(-D\log m\big) \ =\ 2m^{-D}$$

and therefore $\mathbb{P}(\|\mathbf{M}\|_{S_\infty} \ge \tau_{row}) \le 2m^{1-D}$, where

$$\tau_{row} \ =\ \Big(\sqrt{2D\sigma^2 S_{row}^2} + 2DH_{row}H\sqrt{\frac{\log m}{N}}\Big)\sqrt{\frac{m\log m}{N}}\,.$$

Similarly, using $\|\mathbf{M}\|_{S_\infty} = \sup_{|v|_2=1} |v'\mathbf{M}|_2$, and assuming (8.6) and (8.7), we get $\mathbb{P}(\|\mathbf{M}\|_{S_\infty} \ge \tau_{col}) \le 2T^{1-D}$, where

$$\tau_{col} \ =\ \Big(\sqrt{2D\sigma^2 S_{col}^2} + 2DH_{col}H\sqrt{\frac{\log T}{N}}\Big)\sqrt{\frac{T\log T}{N}}\,.$$

□

Proof of Lemma 3. The matrix $\mathbf{M} = \frac{1}{N}\sum_{i=1}^N \xi_i X_i$ is a sum of i.i.d. random matrices. Therefore, part (ii) of the lemma follows by direct application of the large deviations inequality of Nemirovski (2004).

To prove part (i) of the lemma, we use bounds on maximal eigenvalues of subgaussian matrices due to Mendelson et al. (2007), see also Vershynin (2007). However, direct application of these bounds (based on the overall subgaussianity) does not lead to rates that are accurate enough for our purposes. We therefore need to refine the argument using the specific structure of the matrices. Note first that

$$\|\mathbf{M}\|_{S_\infty} = \max_{v\in\mathcal{S}^{T-1}} |\mathbf{M}v|_2 = \max_{u\in\mathcal{S}^{m-1},v\in\mathcal{S}^{T-1}} u'\mathbf{M}v,$$

where $\mathcal{S}^{m-1}$ is the unit sphere in $\mathbb{R}^m$. Therefore, denoting by $\mathcal{M}_m$ and $\mathcal{M}_T$ the minimal 1/2-nets in Euclidean metric on $\mathcal{S}^{m-1}$ and $\mathcal{S}^{T-1}$ respectively, we easily get

$$\|\mathbf{M}\|_{S_\infty} \le 2\max_{v\in\mathcal{M}_T} |\mathbf{M}v|_2 \le 4 \max_{u\in\mathcal{M}_m,v\in\mathcal{M}_T} |u'\mathbf{M}v|.$$

Now, $\mathrm{Card}(\mathcal{M}_m)\le 5^m$, cf. Kolmogorov and Tikhomirov (1959), so that by the union bound, for any $\tau>0$,

$$\mathbb{P}\Big(\|\mathbf{M}\|_{S_\infty} \ge \tau\Big) \le 5^{m+T} \max_{u\in\mathcal{M}_m,v\in\mathcal{M}_T} \mathbb{P}\Big(|u'\mathbf{M}v|\ge \tau/4\Big). \tag{10.2}$$

It remains to bound the last probability in (10.2) for fixed u,v. Let us fix some $u\in\mathcal{S}^{m-1}, v\in\mathcal{S}^{T-1}$ and introduce the random event

$$\mathcal{A}=\left\{\frac{1}{N}\sum_{i=1}^N (u'X_iv)^2 \le \frac{5(m+T)}{N}\right\}.$$

Note that $\mathbb{E}(u'X_iv)^2 = \sum_{k=1}^m\sum_{l=1}^T u_k^2v_l^2\mathbb{P}(X_1=e_k(m)e_l'(T)) = (mT)^{-1}|u|_2^2|v|_2^2 = (mT)^{-1}$, and consider the zero-mean random variables $\eta_i=(u'X_iv)^2-\mathbb{E}(u'X_iv)^2 = (u'X_iv)^2-(mT)^{-1}$. We have $|\eta_i|\le 2\max_i(u'X_iv)^2\le 2|u|_2^2|v|_2^2=2$ (a.s.). Furthermore,

$$\begin{aligned}\mathbb{E}(\eta_i^2) \le \mathbb{E}(u'X_iv)^4 &\le \sum_{k=1}^m\sum_{l=1}^T u_k^4v_l^4\mathbb{P}(X_1=e_k(m)e_l'(T))\\ &= (mT)^{-1}\sum_{k=1}^m u_k^4\sum_{l=1}^T v_l^4 \le (mT)^{-1}.\end{aligned}$$

Therefore, using Bernstein's inequality and the condition $(m+T)/N > (mT)^{-1}$ we get

$$\begin{aligned}\mathbb{P}(\mathcal{A}^c) &\le 2\exp\left(-\frac{N(4(m+T)/N)^2}{2(mT)^{-1}+(4/3)(4(m+T)/N)}\right)\\ &\le 2\exp(-2(m+T)),\end{aligned} \tag{10.3}$$

where $\mathcal{A}^c$ is the complement of $\mathcal{A}$. We now bound the conditional probability

$$\mathbb{P}\Big(|u'\mathbf{M}v|\ge\tau/4\Big|X_1,\dots,X_N\Big) = \mathbb{P}\Big(\Big|\frac{1}{N}\sum_{i=1}^N\xi_i(u'X_iv)\Big|\ge\tau/4\Big|X_1,\dots,X_N\Big).$$

Note that conditionally on $X_1,\dots,X_N$, the $\xi_i(u'X_iv)$ are independent zero-mean random variables with

$$\sum_{i=1}^N\mathbb{E}\Big(|\xi_i(u'X_iv)|^l\Big|X_1,\dots,X_N\Big)\le\mathbb{E}|\xi_1|^l\sum_{i=1}^N|u'X_iv|^2,\quad\forall\, l\ge 2,$$

where we used the fact that $|u'X_iv|^{l-2} \le (|u|_2|v|_2)^{l-2} = 1$ (a.s.) for $l \ge 2$. This and the Bernstein condition (8.2) yield that, for $(X_1, \dots, X_N) \in \mathcal{A}$,

$$\sum_{i=1}^{N} \mathbb{E}\Big(|\xi_i(u'X_iv)|^l \Big| X_1, \dots, X_N\Big) \le \frac{l!}{2} B^2 H^{l-2}.$$

with $B^2 = 5(m+T)\sigma^2$. Therefore, by Lemma 8, for $(X_1, \dots, X_N) \in \mathcal{A}$ we have

$$\mathbb{P}\Big(|u'\mathbf{M}v| \ge \tau/4 \Big| X_1, \dots, X_N\Big) \quad \le \quad 2\exp\left(-\frac{N^2\tau^2/16}{10\sigma^2(m+T) + N\tau H/2}\right). \tag{10.4}$$

For τ defined in (8.10) the last expression does not exceed $2\exp(-D(m+T))$. Together with (10.2) and (10.3), this proves the lemma. ☐

Proof of Lemma 2. We act as in the proof of Lemma 3 but since the matrices X_i are now deterministic, we do not need to introduce the event $\mathcal{A}$. By the definition of $\phi_{\max}(1)$,

$$\frac{1}{N}\sum_{i=1}^{N}(u'X_iv)^2 = |\mathcal{L}(uv')|_2^2 \le \phi_{\max}^2(1)\|uv'\|_{S_2}^2 = \phi_{\max}^2(1)$$

for all $u \in \mathcal{S}^{m-1}, v \in \mathcal{S}^{T-1}$. Hence, $\frac{1}{N}\sum_{i=1}^{N}\xi_i(u'X_iv)$ is a zero-mean Gaussian random variable with variance not larger than $\phi_{\max}^2(1)\sigma^2/N$. Therefore,

$$\mathbb{P}\Big(|u'\mathbf{M}v| \ge \tau/4\Big) \quad \le \quad 2\exp\left(-\frac{N\tau^2}{32\phi_{\max}^2(1)\sigma^2}\right).$$

For τ as in (8.9) the last expression does not exceed $2\exp(-2(m+T))$. Combining this with (10.2) we get the lemma. ☐

Proof of Lemma 4. We act again as in the proof of Lemmas 3 and 2. Denote by Ω the set of pairs (k,l) such that $\{X_1, \dots, X_N\} = \{e_k(m)e_l'(T), (k,l) \in \Omega\}$ (recall that all X_i are distinct by assumption). Then,

$$\sum_{i=1}^{N}(u'X_iv)^2 = \sum_{(k,l)\in\Omega} u_k^2 v_l^2 \le |u|_2^2|v|_2^2 = 1 \tag{10.5}$$

for any $u \in \mathcal{S}^{m-1}, v \in \mathcal{S}^{T-1}$. Hence, under the assumptions of part (i) of the lemma,

$$\mathbb{P}\Big(|u'\mathbf{M}v| \ge \tau/4\Big) \quad \le \quad 2\exp\left(-\frac{N^2\tau^2}{32\sigma^2}\right)$$

which does not exceed $2\exp(-D(m+T))$ for τ defined in (8.12). Combining this with (10.2) we get part (i) of the lemma. To prove part (ii) we note that, as in the proof of Lemma 3, $|u'X_iv|^{l-2} \le 1$ (a.s.) for $l \ge 2$. This and (10.5) yield

$$\sum_{i=1}^{N}\mathbb{E}(|\xi_i(u'X_iv)|^l) \le \frac{l!}{2}B^2H^{l-2}, \quad \forall\, l \ge 2,$$

with $B^2 = \sigma^2$. Therefore, by Lemma 8, we have

$$\mathbb{P}\Big(|u'\mathbf{M}v| \ge \tau/4\Big) \quad \le \quad 2\exp\left(-\frac{N^2\tau^2/16}{2\sigma^2 + N\tau H/2}\right),$$

and we complete the proof of (ii) in the same way as in Lemmas 3 and 2.

Part (iii) follows by an application of Theorem 2.1, Tropp (2010), after replacing every X_i by its self-adjoint dilation (see Paulsen 1986). □

For the proof of Lemma 5 we will need some notation. The pth Schatten class of $M \times M$-matrices is denoted by S_p^M, and we write

$$\mathcal{B}(S_p^M) \;=\; \left\{A \in \mathbb{R}^{M\times M} : \|A\|_{S_p} \le 1\right\}$$

for the corresponding closed Schatten-p unit ball in $\mathbb{R}^{M\times M}$. For any pseudo-metric space $(\mathcal{T}, d)$ and any $\varepsilon > 0$ we define the covering number

$$\mathcal{N}(\mathcal{T}, d, \varepsilon) \;=\; \min\Big\{\mathrm{Card}(\mathcal{T}_0) : \mathcal{T}_0 \subset \mathcal{T} \text{ and } \inf_{s\in\mathcal{T}_0} d(t,s) \le \varepsilon \text{ for all } t \in \mathcal{T}\Big\}.$$

In other words, $\mathcal{N}(\mathcal{T}, d, \varepsilon)$ is the smallest number of closed balls of radius ε in the metric d needed to cover the set $\mathcal{T}$. We will sometimes write $\mathcal{N}(\mathcal{T}, \|\cdot\|, \varepsilon)$ instead of $\mathcal{N}(\mathcal{T}, d, \varepsilon)$ if the metric d is associated with the norm $\|\cdot\|$. The empirical norm $\|\cdot\|_{2,N}$ corresponds to $\hat{d}_{2,N}$, i.e., for all $A \in \mathbb{R}^{M\times M}$,

$$\|A\|_{2,N}^2 \;=\; \frac{1}{N}\sum_{j=1}^{N} \mathrm{tr}(A'X_j)^2.$$

Proof of Lemma 5. Let us first assume that $m = T \equiv M$. Since

$$\sup_{B\in\mathbb{R}^{M\times M}} \left| \frac{\frac{1}{\sqrt{N}}\sum_{i=1}^N \xi_i \mathrm{tr}(B'X_i)}{\|B\|_{2,N}^{1-\frac{p}{2-p}} \|B\|_{S_p}^{\frac{p}{2-p}}} \right| \;=\; \sup_{B\in\mathcal{B}(S_p^M)} \left| \frac{\frac{1}{\sqrt{N}}\sum_{i=1}^N \xi_i \mathrm{tr}(B'X_i)}{\|B\|_{2,N}^{1-\frac{p}{2-p}}} \right|,$$

the expression on the LHS of (8.14) is not greater than

$$\frac{\sqrt{M}}{\sqrt{p}\sqrt{N}}\, \hat{d}_{2,N}(\hat{A}, A^*)^{1-\frac{p}{2-p}} \,\|\hat{A} - A^*\|_{S_p}^{\frac{p}{2-p}} \sup_{B\in\mathcal{B}(S_p^M)} \left| \frac{(M/p)^{\frac{p-2}{2p}} N^{-1/2} \sum_{i=1}^N \xi_i \mathrm{tr}(B'X_i)}{\left((M/p)^{\frac{p-2}{2p}} \|B\|_{2,N}\right)^{1-\frac{p}{2-p}}} \right|.$$

Due to the linear dependence in M of the ε-entropies of the quasi-convex Schatten class embeddings $S_p^M \hookrightarrow S_2^M$ (cf. Corollary 6) and the fact that the required bound should be uniform in M and in p for $p \searrow 0$, we introduced an additional weighting by $(M/p)^{\frac{p-2}{2p}}$. Now define

$$\mathcal{G}_{M,p} \;=\; \Big\{A \in \mathbb{R}^{M\times M} : (M/p)^{\frac{2-p}{2p}} A \in \mathcal{B}(S_p^M)\Big\}.$$

By the entropy bound of Corollary 6 and the uniform boundedness condition (2.4),

$$\log \mathcal{N}(\mathcal{G}_{M,p}, \hat{d}_{2,N}, \varepsilon) \;\le\; \log \mathcal{N}(\mathcal{G}_{M,p}, \sqrt{c_0}\,\|\cdot\|_{S_2}, \varepsilon) \;\le\; p\,\alpha_0(p)(\varepsilon/\sqrt{c_0})^{-\frac{2p}{2-p}},$$

whence

$$\int_0^\delta \sqrt{\log \mathcal{N}(\mathcal{G}_{M,p}, \hat{d}_{2,N}, \varepsilon)}\, d\varepsilon \;\le\; c_0^{\frac{p}{2(2-p)}}\, p\,\alpha_0(p) \frac{2-p}{2-2p}\, \delta^{1-\frac{p}{2-p}}. \tag{10.6}$$

We remark that due to the order specification of α_0 in Corollary 6, the expression

$$c_0^{\frac{p}{2(2-p)}}\, p\,\alpha_0(p) \frac{2-p}{2-2p} \tag{10.7}$$

is uniformly bounded as long as p stays uniformly bounded away from 1. Note that for $p = 1$ the entropy integral on the LHS in (10.6) does not converge.

Claim 1. *For any $q \in (0,1)$, there exist constants $c(q)$ and $c'(q)$, such that for all $0 < p \le q$, all $0 < \delta \le \sqrt{c_0}$ and uniformly in M and N:*

$$\mathbb{P}\Big(\sup_{\substack{B \in \mathcal{G}_{M,p}:\\ \|B\|_{2,N} \le \delta}} \Big| \frac{1}{\sqrt{N}} \sum_{j=1}^N \xi_j \operatorname{tr}(X_j' B) \Big| \ge T \Big) \le c(q) \exp\Big(- \frac{T^2}{c(q)^2 \delta^2} \Big) \tag{10.8}$$

for all $T \ge c'(q) \delta^{1 - \frac{p}{2-p}}$.

Proof of Claim 1. The bound is essentially stated in van de Geer (2000) as Lemma 3.2 (further referred to as VG(00)). The constant in VG(00) depends neither on the $\|\cdot\|_{2,N}$-diameter of the function class nor on the function class itself and is valid, in particular, for $\varepsilon = 0$, in the notation of VG(00). The uniformity in $0 < p \le q$ follows from the uniform boundedness of (10.7) over $p \in (0, q]$. The required case corresponds to $K = \infty$ in the notation of VG(00). Its proof follows by taking $\varepsilon = 0$ and applying the theorem of monotone convergence as $K \to \infty$, since the RHS of the inequality is independent of K.

Claim 2. *For any $q \in (0,1)$, there exists a constant $C(q)$ such that for any $0 < p \le q$*

$$\mathbb{P}\Big(\sup_{B \in \mathcal{G}_{M,p}} \Big| \frac{\frac{1}{\sqrt{N}} \sum_{j=1}^N \xi_j \operatorname{tr}(B' X_j)}{\|B\|_{2,N}^{1 - \frac{p}{2-p}}} \Big| \ge T \Big) \le C(q) \exp\Big(- T^2 M / C(q)^2 \Big) \tag{10.9}$$

for all $T \ge C(q)$.

Proof of Claim 2. First observe that

$$\sup_{A \in \mathcal{G}_{M,p}} \|A\|_{2,N} \le \sqrt{c_0} \sup_{A \in \mathcal{G}_{M,p}} \|A\|_{S_2} \le \sqrt{c_0} (M/p)^{\frac{p-2}{2p}} \sup_{A \in \mathcal{B}(S_2^M)} \|A\|_{S_2} = \sqrt{c_0} (M/p)^{\frac{p-2}{2p}},$$

where the last inequality follows from $\mathcal{B}(S_p^M) \subset \mathcal{B}(S_2^M)$. Define the decomposition of $\mathcal{G}_{M,p}$

$$\mathcal{G}_{M,p}^{(k)} = \Big\{ A \in \mathcal{G}_{M,p} : 2^{k \frac{p-2}{2p}} \sqrt{c_0} (M/p)^{\frac{p-2}{2p}} \le \|A\|_{2,N} \le 2^{(k-1)\frac{p-2}{2p}} \sqrt{c_0} (M/p)^{\frac{p-2}{2p}} \Big\}, \ k \in \mathbb{N}.$$

Then by straightforward peeling-off the class $\mathcal{G}_{M,p}$, we obtain for all $T \ge c'(q)$

$$\begin{aligned}
&\mathbb{P}\Big(\sup_{B \in \mathcal{G}_{M,p}} \Big| \frac{\frac{1}{\sqrt{N}} \sum_{j=1}^N \xi_j \operatorname{tr}(B' X_j)}{\|B\|_{2,N}^{1 - \frac{p}{2-p}}} \Big| \ge T \Big) \\
&\le \sum_{k=1}^\infty \mathbb{P}\Big(\sup_{B \in \mathcal{G}_{M,p}^{(k)}} \Big| \frac{1}{\sqrt{N}} \sum_{j=1}^N \xi_j \operatorname{tr}(B' X_j) \Big| \ge T \big(2^{k \frac{p-2}{2p}} \sqrt{c_0} (M/p)^{\frac{p-2}{2p}} \big)^{1 - \frac{p}{2-p}} \Big) \\
&\le \sum_{k=1}^\infty c(q) \exp\Big(- \frac{T^2 \big(2^{k \frac{p-2}{2p}} \sqrt{c_0} (M/p)^{\frac{p-2}{2p}} \big)^{-\frac{2p}{2-p}}}{c(q)^2} \Big) \\
&\le \sum_{k=1}^\infty c(q) \exp\Big(- \frac{T^2 M 2^k C_0(q)}{q c(q)^2} \Big) \qquad (10.10)
\end{aligned}$$

with the definition

$$C_0(q) \;=\; \inf_{0<p\le q} c_0^{-\frac{p}{2-p}}.$$

It remains to note that the last sum in (10.10) is bounded by $C(q)\exp\left(-T^2M/C(q)^2\right)$ whenever $T \ge C(q)$ for some suitable constant $C(q)$.

In particular, the result reveals that the LHS of (8.14) is bounded by

$$\hat{d}_{2,N}(\hat{A}, A^*)^{1-\frac{p}{2-p}} \,\|\,\hat{A} - A^*\,\|_{S_p}^{\frac{p}{2-p}} \sqrt{\vartheta/p}\Big(\frac{M}{N}\Big)^{1/2} \tag{10.11}$$

with probability at least $1 - C\,\exp(-\vartheta M/C^2)$ for any $\sqrt{\vartheta} \ge C(q)$.

We now use the following simple consequence of the concavity of the logarithm which is stated, for instance, in Tsybakov and van de Geer (2005) (Lemma 5).

LEMMA 9. *For any positive v, t and any $\kappa \ge 1$, $\delta > 0$ we have*

$$vt^{1/(2\kappa)} \;\le\; (\delta/2)t + c_\kappa \delta^{-1/(2\kappa-1)} v^{2\kappa/(2\kappa-1)},$$

where $c_\kappa = (2\kappa-1)(2\kappa)\kappa^{-1/(2\kappa-1)}$.

Taking in Lemma 9

$$t = \hat{d}_{2,N}(\hat{A}, A^*)^2, \quad v = \,\|\,\hat{A} - A^*\,\|_{S_p}^{\frac{p}{2-p}} \sqrt{\vartheta/p}\Big(\frac{M}{N}\Big)^{1/2},$$

and $\kappa = (2-p)/(2-2p)$ shows that for any $\delta > 0$

$$(10.11) \;\le\; (\delta/2)\,\hat{d}_{2,N}(\hat{A}, A^*)^2 \;+\; \tau_7 \delta^{p-1} \,\|\,\hat{A} - A^*\,\|_{S_p}^p$$

with probability at least $1 - C\,\exp(-\vartheta M/C^2)$.

The case $m \ne T$ can be deduced from the above result by the following observation. For any matrix $B = (b_{ij}) \in \mathbb{R}^{m\times T}$, define the extension $\tilde{B} = (\tilde{b}_{ij}) \in \mathbb{R}^{M\times M}$ with $M = \max(m, T)$ as follows: $\tilde{b}_{ij} = b_{ij}$ for $1 \le i \le m$, $1 \le j \le T$, and $\tilde{b}_{ij} = 0$ otherwise. Then one easily checks that $\|\tilde{B}\|_{S_p} = \|B\|_{S_p}$ for all $p \in [0,\infty]$. Furthermore, $\mathrm{tr}(B'X_i) = \mathrm{tr}(\tilde{B}'\tilde{X}_i)$ and

$$\sup_{A\in\mathbb{R}^{M\times M}\setminus\{0\}} \frac{N^{-1}\sum_{k=1}^N \mathrm{tr}\big(\tilde{X}_i' A\big)^2}{\|A\|_{S_2}^2} \;=\; \sup_{A\in\mathbb{R}^{m\times T}\setminus\{0\}} \frac{\|A\|_{2,N}^2}{\|A\|_{S_2}^2} \;\le\; c_0.$$

Consequently, the result follows now from the already established proof for the case $m = T$. □

11. Entropy numbers for quasi-convex Schatten class embeddings. Here we derive bounds for the kth entropy numbers of the embeddings $S_p^M \hookrightarrow S_2^M$ for $0 < p < 1$, where S_p^M denotes the pth Schatten class of real $M \times M$-matrices. Corresponding results for the $l_p^M \hookrightarrow l_2^M$-embeddings are given first by Edmunds and Triebel (1989) but their proof does not carry over to the Schatten spaces. Pajor (1998) provides bounds for the $S_p^M \hookrightarrow S_2^M$ embeddings in the convex case, $p \ge 1$.

His approach is based on the trace duality (Hölder inequality for $p^{-1}+q^{-1}=1$) and the geometric formulation of Sudakov's minoration

$$\varepsilon\sqrt{\log\mathcal{N}(A,|\cdot|_2,\varepsilon)} \;\le\; c\,\mathbb{E}\sup_{t\in A}\langle G,t\rangle$$

for some positive constant c, with a d-dimensional standard Gaussian vector G and an arbitrary subset A of $\mathbb{R}^d$. Here $|\cdot|_2$ is the Euclidean norm in $\mathbb{R}^d$ and $\langle\cdot,\cdot\rangle$ is the corresponding scalar product. Guédon and Litvak (2000) derive a slightly sharper bound for the $l_p \hookrightarrow l_q$-embeddings than Edmunds and Triebel (1989) with a different technique. In addition, they prove lower bounds. We adjust their ideas concerning finite ℓ_p spaces to the non-convex Schatten spaces.

We denote by $e_k(id^M_{p,r})$ the kth entropy number of the embedding $S^M_p \hookrightarrow S^M_r$ for $0<p<r\le\infty$, i.e., the infimum of all $\varepsilon>0$ such that there exist 2^{k-1} balls in S^M_r of radius ε that cover $\mathcal{B}(S^M_p)$. For the general definition of kth entropy numbers $e_k(T:F\to E)$ of bounded linear operators T between quasi-Banach spaces F and E we refer to Edmunds and Triebel (1996).

Recall that a homogeneous non-negative functional $\|\cdot\|$ is called C-quasi-norm, if it satisfies for all x,y the inequality $\|x+y\|\le C\max\left(\|x\|,\|y\|\right)$. Finally, any p-norm is a C-quasi-norm with $C=2^{1/p}$ (cf., e.g., Edmunds and Triebel 1996, page 2). We will use the following lemma.

LEMMA 10 (Guédon and Litvak, 2000). *Assume that $\|\cdot\|_i$ are symmetric C_i-quasi-norms on $\mathbb{R}^n$ for $i=0,1$, and for some $\theta\in(0,1)$, $\|\cdot\|_\theta$ is a quasi-norm on $\mathbb{R}^n$ such that $\|x\|_\theta\le\|x\|_0^\theta\|x\|_1^{1-\theta}$ for all $x\in\mathbb{R}^n$. Then for any quasi-normed space F, any linear operator $T:F\to\mathbb{R}^n$, and all integers k and m, we have*

$$e_{m+k-1}(T:F\to E_\theta) \;\le\; \Big(C_0 e_m(T:F\to E_0)\Big)^\theta\Big(C_1 e_k(T:F\to E_1)\Big)^{1-\theta}.$$

where E_t stands for $\mathbb{R}^n$ equipped with quasi-norm $\|\cdot\|_t$, $t\in\{0,\theta,1\}$.

Guédon and Litvak did not specify the notion of symmetry they used. So we have to clarify that here, a (quasi-)norm $\|.\|$ is said to be symmetric if $(\mathbb{R}^n,\|.\|)$ is isometrically isomorphic to a symmetrically (quasi-)normed operator ideal. This includes the diagonal operator spaces (finite ℓ_p) as well as the Schatten spaces. The proof of Lemma 10 follows the lines of Pietsch (1980), Prop. 12.1.12, replacing the triangle inequality by the quasi-triangle inequality. Note that the Schatten classes S_p form interpolation couples like their commutative analogs ℓ_p.

LEMMA 11 (Interpolation inequality). *For $0<p<q<r<\infty$ let $\theta\in[0,1]$ be such that*

$$\frac{\theta}{p}+\frac{1-\theta}{r} = \frac{1}{q}.$$

Then, for all $A\in\mathbb{R}^{m\times T}$,

$$\|A\|_{S_q} \;\le\; \|A\|^\theta_{S_p}\,\|A\|^{1-\theta}_{S_r}.$$

PROOF is immediate in view of the inequalities

$$\sum_j a_j^q=\sum_j a_j^{\theta q}a_j^{(1-\theta)q}\le\Big(\sum_j a_j^p\Big)^{\frac{\theta q}{p}}\Big(\sum_j a_j^r\Big)^{\frac{(1-\theta)q}{r}}$$

valid for any non-negative a_j's.

PROPOSITION 1 (Entropy numbers). *Let $0 < p < 1$, $p < r \le \infty$. Then there exists an absolute constant β independent of p and r, such that for all integers k and M we have*

$$e_k(id^M_{p,r}) \ \le \ \min\Big\{1,\ \alpha(\beta,p,r)\Big(\frac{M}{k}\Big)^{1/p-1/r}\Big\}$$

with

$$\alpha(\beta,p,r) \ \le \ 2^{1+1/r}\Big(\frac{\beta}{p}\Big)^{1/p-1/r}\Big(\frac{1}{1-p}\Big)^{(1/p-1)(1/p-1/r)}.$$

PROOF. The fact that $e_k(id^M_{p,r})$ is bounded by 1 is obvious, since $\mathcal{B}(S^M_p) \subset \mathcal{B}(S^M_r)$. Consider the other case. We start with $r = \infty$ and then extend the result to $r < \infty$ by interpolation. Fix some number $L > M$ and let $D = D(M,L,p)$ be the smallest constant which satisfies, for all $1 \le k \le L$,

$$e_k(id^M_{p,\infty}) \ \le \ D\Big(\frac{M}{k}\Big)^{1/p}. \tag{11.1}$$

Let us show that $\alpha = \sup_{M,L} D(M,L,p)$ is finite. Since $\|\cdot\|_{S_p}$, $p < 1$, can be viewed as a quasi-norm on $\mathbb{R}^{M^2}$ (isomorphic to $\mathbb{R}^{M\times M}$), Lemma 10 applies with $F = E_0 = S^M_p$, $E_1 = S^M_\infty$, $\theta = p$, $E_\theta = S^M_1$ and $m = 1$. This gives

$$e_k(id^M_{p,1}) \ \le \ 4\Big(e_k(id^M_{p,\infty})\Big)^{1-p}. \tag{11.2}$$

Here the factor 4 follows from the relations $C_1 = 2$ and $C_0^p \le 2$. Now, (11.2) and the factorization theorem for entropy numbers of bounded linear operators between quasi-Banach spaces (see, e.g., Edmunds and Triebel 1996, page 8), with factorization via S^M_1, leads to the bound

$$\begin{aligned} e_k(id^M_{p,\infty}) \ &\le \ e_{[(1-p)k]}(id^M_{p,1})e_{[pk]}(id^M_{1,\infty}) \\ &\le \ 4\Big(e_{[(1-p)k]}(id^M_{p,\infty})\Big)^{1-p} e_{[pk]}(id^M_{1,\infty}), \end{aligned} \tag{11.3}$$

where for any $x \in (0,\infty)$, $[x]$ denotes the smallest integer which is larger or equal to x. Proposition 5 of Pajor (1998) entails $\log \mathcal{N}(\mathcal{B}(S^M_1), \|\cdot\|_{S_\infty}, \varepsilon) \ \le \ c\,M/\varepsilon$, $\forall\, \varepsilon > 0$, and hence

$$e_k(id^M_{1,\infty}) \ \le \ c'\,M/k \tag{11.4}$$

with constants c and c' independent of M, ε and k. Note that, in contrast to the $l^M_1 \hookrightarrow l^M_\infty$-embedding, for which the k'th entropy numbers are bounded by $c''k^{-1}\log(1 + M/k)$ with some $c'' > 0$ and $\log_2 M \le k \le M$ (see, e.g., Edmunds and Triebel 1996, page 98), we have in (11.4) not a logarithmic but linear dependence of M in the upper bound. Plugging (11.1) and (11.4) into (11.3) yields

$$\begin{aligned} e_k(id^M_{p,\infty}) \ &\le \ 4\Big(D\Big(\frac{M}{(1-p)k}\Big)^{1/p}\Big)^{1-p}\frac{c'M}{pk} \\ &= \ \frac{4\,c'}{p}\Big(\frac{1}{1-p}\Big)^{(1-p)/p} D^{1-p}\Big(\frac{M}{k}\Big)^{1/p}. \end{aligned}$$

Thus, by definition of D,

$$D^p \;\le\; \frac{4\,c'}{p}\Big(\frac{1}{1-p}\Big)^{(1-p)/p},$$

which shows that D is uniformly bounded in M and L. This proves the proposition for $r=\infty$.

Consider now the case $r<\infty$. In view of Lemma 11 with $\theta=p/r$, we can apply Lemma 10 with $F=E_0=S_p^M$, $E_1=S_\infty^M$, $\theta=p/r$, $E_\theta=S_r^M$ and $m=1$. This yields

$$\begin{aligned} e_k\big(id^M_{p,r}\big) \;&\le\; 2^{1+1/r}\Big(e_k\big(id^M_{p,\infty}\big)\Big)^{1-p/r} \\ &\le\; 2^{1+1/r}D^{1-p/r}\Big(\frac{M}{k}\Big)^{1/p-1/r}. \end{aligned}$$

□

COROLLARY 6. *For any $p\in(0,1)$, there exists a positive constant $\alpha_0(p)$ such that for all integers $M\ge 1$ and any $\varepsilon\in(0,1]$,*

$$\log\mathcal{N}(\mathcal{B}(S_p^M),\|\cdot\|_{S_2},\varepsilon) \;\le\; \alpha_0(p)\,M\,\varepsilon^{-\frac{2p}{2-p}}.$$

Moreover, $\alpha_0(p)=O(1/p)$ for $p\searrow 0$.

PROOF. The result follows by transforming the entropy number bound of Proposition 1 into an entropy bound. Specification of the constant in Proposition 1 yields

$$\alpha_0(p) \;=\; O\bigg(\frac{\beta}{p}\Big(1+\frac{p}{1-p}\Big)^{(1-p)/p}\bigg) = O(1/p)$$

as $p\searrow 0$. □

Acknowledgment. We are grateful to Alain Pajor for pointing out reference [22].

References.

[1] Abernethy,J., Bach,F., Evgeniou,T. and Vert, J.-P. (2009). A New Approach to Collaborative Filtering: Operator Estimation with Spectral Regularization. *Journal of Machine Learning Research,* ***10*** 803–826.

[2] Amini, A. and Wainwright, M. (2009) High-dimensional analysis of semidefinite relaxations for sparse principal components. *Annals of Statistics,* ***37***, 2877–2921.

[3] Argyriou, A., Evgeniou, T. and Pontil, M. (2008) Convex multi-task feature learning. *Machine Learning,* ***73***, 243–272.

[4] Argyriou, A., Micchelli, C.A., and Pontil, M. (2008) On spectral learning. *Journal of Machine Learning Research*, to appear.

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

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

[7] Bickel, P.J., Levina, E. (2008) Regularized estimation of large covariance matrices. *Annals of Statistics,* ***36***, 199–227.

[8] Bickel, P.J., Ritov, Y. and Tsybakov, A.B. (2009) Simultaneous analysis of Lasso and Dantzig selector. *Annals of Statistics,* ***37***, 1705–1732.

[9] Bunea, F., Tsybakov, A.B. and Wegkamp, M.H. (2007) Aggregation for Gaussian regression. *Annals of Statistics* ***35***, 1674–1697.

[10] Cai, T., Zhang, C.-H. and Zhou, H.H. (2010) Optimal Rates of Convergence for Covariance Matrix Estimation. *Annals of Statistics* ***38***, 2118–2144.

[11] Candès, E. J. and Plan, Y. (2009) Matrix completion with noise. *Proceedings of IEEE*, 2009.
[12] 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.
[13] Candès, E. J. and Recht, B. (2008) Exact matrix completion via convex optimization. *Found. of Comput. Math.*, to appear.
[14] Candès, E. J. and Tao,T. (2005) Decoding by linear programming. *IEEE Transactions on Information Theory* ***51*** 4203–4215.
[15] Candès, E. J. and Tao,T. (2009) The power of convex relaxation: Near-optimal matrix completion. Submitted.
[16] Mc Carthy, C.A. (1967) C_p. *Israel J. Math.* **5** 249– 272.
[17] Dalalyan, A. and Tsybakov, A. (2008). Aggregation by exponential weighting, sharp oracle inequalities and sparsity. *Machine Learning* ***72*** 39–61.
[18] Donoho, D. L., Johnstone, I. M., Hoch, J. C. and Stern, A. S. (1992). Maximum entropy and the nearly black object. *J. Roy. Statist. Soc. Ser. B*, **54** 41–81. With discussion and a reply by the authors.
[19] Edmunds, D.E. and Triebel, H. (1996) *Function spaces, Entropy numbers, Differential operators.* Cambridge University Press, Cambridge.
[20] Edmunds, D.E. and Triebel, H. (1989) Entropy numbers and approximation numbers in function spaces. *Proc. London Math. Soc.* ***58***, 137–152.
[21] Foster, D. P. and George, E. I. (1994) The risk inflation criterion for multiple regression *Annals of Statistics* ***22*** 1947–1975.
[22] Guédon, O. and Litvak, A. E. (2000). Euclidean projections of a p-convex body. Geometric aspects of functional analysis, Israel Seminar (GAFA) 1996–2000, V.D. Milman and G. Schechtman, eds., *Lecutre Notes in Mathematics* ***1745***, 95–108, Springer 2000.
[23] Gross, D. (2009) Recovering low-rank matrices from few coefficients in any basis. arXiv:0910.1879. 2009.
[24] Keshavan, R.H., Montanari, A. and Oh, S. (2009) Matrix completion from noisy entries. arXiv:0906.2027
[25] Kolmogorov, A.N. and Tikhomirov,V.M. (1959) The ϵ-entropy and ϵ-capacity of sets in function spaces *Uspekhi Matem. Nauk,* ***14***, 3–86.
[26] Koltchinskii, V. (2008) Oracle inequalities in empirical risk minimization and sparse recovery problems. Saint-Flour Lectures Notes.
[27] Lounici, K., Pontil, M., Tsybakov, A.B. and van de Geer, S. (2009) Taking advantage of sparsity in multi-task learning. *Proceedings of COLT-2009.*
[28] Lounici, K., Pontil, M., Tsybakov, A.B. and van de Geer, S. (2010) Oracle Inequalities and Optimal Inference under Group Sparsity. arXiv:1007.1771
[29] Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the Lasso. *Annals of Statistics* **34** 1436–1462.
[30] Mendelson, S., Pajor, A. and Tomczak-Jaegermann, N. (2007) Reconstruction and subgaussian operators in asymptotic geometric analysis. *Geometric and Functional Analysis* ***17***, 1248–1282.
[31] Negahban, S., Ravikumar, P., Wainwright, M.J. and Yu, B. (2009) A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Advances in Neural Information Processing Systems, NIPS-2009.*
[32] Negahban, S. and Wainwright, M.J. (2009) Estimation of (near) low rank matrices with noise and high-dimensional scaling. arXiv:0912.5100, December 2009.
[33] Nemirovski, A. (2004). Regular Banach spaces and large deviations of random sums. *Unpublished manuscript.*
[34] Pajor, A. (1998). Metric entropy of the Grassmann manifold. *Convex Geometric Analysis* ***34***, 181–188.
[35] Paulsen, V. I. (1986) *Completely bounded maps and dilations.* Number 146 in Pitman Research Notes in Mathematics. Longman Scientific & Technical, New York, NY, 1986.
[36] Pietsch, A. (1980) Operator ideals. Elsevier North-Holland/Amsterdam.
[37] Pinelis, I.F. and Sakhanenko, A.I. (1985) Remarks on inequalities for the probabilities of large deviations. *Theory of Prob. Appl.* ***30***, 143–148.
[38] Ravikumar, P., Wainwright, M., Raskutti, G. and Yu, B. (2008) High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. Submitted.
[39] Recht, B. (2009) A simpler approach to matrix completion. arXiv:0910.0651

[40] Recht, B., Fazel, M. and Parrilo, P.A. (2007) Guaranteed Minimum-Rank Solutions of Linear Matrix Equations via Nuclear Norm Minimization. arXiv:0706.4138
[41] Rigollet, Ph. and Tsybakov, A.B. (2010) Exponential Screening and optimal rates of sparse estimation. arXiv:1003.2654
[42] Rotfeld, S. Yu. (1969) The singular numbers of the sum of completely continuous operators. *Topics in Mathematical Physics*, vol.3. Spectral Theory (Birman, M.Sh., ed.), N.Y.-London, 73-78.
[43] Srebro, N., Rennie, J. and Jaakkola, T. (2005) Maximum margin matrix factorization. *Advances in Neural Information Processing Systems, NIPS-2005*, 1329-1336.
[44] Srebro, N. and Shraibman, A. (2005) Rank, trace-norm and max-norm. *Learning Theory, Proceedings of COLT-2005*, 545-560.
[45] Tropp, J. A. (2010) User-friendly tail bounds for sums of random matrices. arXiv:1004.4389, April 2010.
[46] Tsybakov, A. (2009) *Introduction to nonparametric estimation.* Springer.
[47] Tsybakov, A. and van de Geer, S. (2005). Square root penalty: adaptation to the margin in classification and in edge estimation *Ann. Statist.* ***33***, 1203–1224.
[48] van de Geer, S. (2000) *Empirical Processes in M-estimation.* Cambridge University Press, Cambridge.
[49] Vershynin, R. (2007) Some problems in asymptotic convex geometry and random matrices motivated by numerical algorithms. *Banach spaces and their applications in analysis*, 209–218, Walter de Gruyter, Berlin.
[50] Zhao, P. and Yu, B. (2006). On model selection consistency of Lasso. *Journal of Machine Learning Research* **7** 2541–2563.

Universität Hamburg
Department Mathematik
Bundesstrasse 55
D-20146 Hamburg
Germany
E-mail: angelika.rohde@math.uni-hamburg.de

Laboratoire de Statistique, CREST
3, avenue Pierre Larousse
F-92240 Malakoff Cedex
France
E-mail: alexandre.tsybakov@upmc.fr