IMS Collections
From Probability to Statistics and Back: High-Dimensional Models and Processes
Vol. 9 (2013) 213–226

DOI: 10.1214/12-IMSCOLL915

A remark on low rank matrix recovery and noncommutative Bernstein type inequalities

Vladimir Koltchinskii*

School of Mathematics, Georgia Institute of Technology

Abstract: A problem of estimation of a large Hermitian nonnegatively definite matrix of trace 1 (a density matrix of a quantum system) motivated by quantum state tomography is studied. The estimator is based on a modified least squares method suitable in the case of models with random design with known design distributions. The bounds on Hilbert-Schmidt error of the estimator, including low rank oracle inequalities, have been proved. The proofs rely on Bernstein type inequalities for sums of independent random matrices.

1. Introduction

Denote $\mathbb{M}_m(\mathbb{C})$ the set of all $m \times m$ matrices with complex entries and $\mathbb{H}_m$ the set of all $m \times m$ Hermitian matrices. Let

$$\mathcal{S} := \{S : S \in \mathbb{H}_m : S \geq 0, \mathrm{tr}(S) = 1\}$$

be the set of *density matrices*. Here and in what follows $S \geq 0$ means that S is a nonnegatively definite matrix and $\mathrm{tr}(S)$ denotes the trace of S. Density matrices are used in quantum statistics to represent the states of quantum systems. Given a state $\rho \in \mathcal{S}$ and a Hermitian matrix $X \in \mathbb{H}_m$ (*an observable*) with spectral representation $X = \sum_j \lambda_j P_j$ (λ_j being the eigenvalues of X and P_j the corresponding spectral projections), the outcomes of a measurement of X in the state ρ are the numbers λ_j with probabilities $p_j = \mathrm{tr}(\rho P_j)$. In what follows, this probability distribution will be denoted $\mu_{\rho,X}$. Clearly, the mean of $\mu_{\rho,X}$ is equal to

$$\mathbb{E}_\rho X = \int_{\mathbb{R}} u \mu_{\rho,X}(du) = \mathrm{tr}(\rho X).$$

Let $X_1, \dots, X_n$ be independent random Hermitian matrices and $Y_1, \dots, Y_n$ be the outcomes of measurements of $X_1, \dots, X_n$ for the system being identically prepared n times in the state ρ. The goal is to estimate the unknown density matrix ρ based on the measurements $(X_1, Y_1), \dots, (X_n, Y_n)$ (that are independent random couples). Such a problem is very basic in *quantum state tomography* (Nielsen and Chuang [13], Artiles, Gill and Guta [2]). Note that $\mathbb{E}(Y_j | X_j) = \mathrm{tr}(\rho X_j)$, $j = 1, \dots, n$. In

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

*Partially supported by NSF grants DMS-0906880 and CCF-0808863.

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

Keywords and phrases: low rank matrix estimation, matrix regression, noncommutative Bernstein inequality, quantum state tomography

what follows, it will be assumed that the data $(X_1, Y_1), \dots, (X_n, Y_n)$ consists of independent couples satisfying the following linear model

$$Y_j = \mathrm{tr}(\rho X_j) + \xi_j, \qquad j = 1, \dots, n$$

with a random noise $\{\xi_j\}$. Here $\{\xi_j\}$ are independent random variables with $\mathbb{E}(\xi_j | X_j) = 0, j = 1, \dots, n$.

We are interested in this problem in the case when m is large, but the rank of the target matrix ρ is relatively small, so, the problem can be viewed as *low rank matrix recovery*. More generally, ρ can be a full rank matrix, but it should be well approximated by low rank matrices. The problems of this nature have been intensively studied in the recent years, the main approach being based on nuclear norm penalization, see Candes and Recht [4], Candes and Tao [5], Candes and Plan [6], Recht [14], Rohde and Tsybakov [15], Koltchinskii, Lounici and Tsybakov [11], Koltchinskii [9] and references therein. Gross et al [7], Gross [8] studied low rank recovery problems in the context of quantum state tomography and developed a powerful method of their analysis based on noncommutative Bernstein inequalities of Ahlswede and Winter [1] (we follow this approach here). Koltchinskii [10] suggested a method of low rank density matrix recovery based on von Neumann entropy penalization.

2. Low rank recovery of density matrices

The following notations will be used throughout the paper. The operator norm of matrices will be denoted by $\|\cdot\|$ and their Schatten p-norm, $p \geq 1$, will be denoted by $\|\cdot\|_p$. Recall that for Hermitian matrices

$$\|A\|_p^p := \mathrm{tr}(|A|^p), \qquad A \in \mathbb{H}_m,$$

where $|A| := \sqrt{A^2}$, $A \in \mathbb{H}_m$. In particular, $\|\cdot\|_1$ is the nuclear norm and $\|\cdot\|_2$ is the Hilbert–Schmidt norm. The norm $\|\cdot\|_2$ is generated by the Hilbert–Schmidt inner product that will be denoted $\langle \cdot, \cdot \rangle$:

$$\langle A, B \rangle := \mathrm{tr}(AB), \qquad A, B \in \mathbb{H}_m$$

(the same notation is also used for vectors in $\mathbb{C}^m$). We use the notation $\otimes$ for tensor products of vectors in $\mathbb{C}^m$ or matrices in $\mathbb{H}_m$ (viewed as vectors in the Euclidean space $(\mathbb{H}_m, \langle \cdot, \cdot \rangle)$). In the last case, given $A, B \in \mathbb{H}_m$, $A \otimes B$ is the linear transformation of $\mathbb{H}_m$ defined as

$$(A \otimes B)C = A\langle B, C \rangle, \qquad C \in \mathbb{H}_m.$$

Given a random matrix X in $\mathbb{H}_m$ with distribution Π, denote

$$\|A\|_{L_2(\Pi)} := \|\langle A, \cdot \rangle\|_{L_2(\Pi)} = \mathbb{E}^{1/2} \langle A, X \rangle^2.$$

A random matrix X is called *isotropic* iff

$$\mathbb{E}(X \otimes X) = \mathrm{Id}_{\mathbb{H}_m},$$

where $\mathrm{Id}_{\mathbb{H}_m}$ denotes the identity transformation of $\mathbb{H}_m$. Equivalently, the distribution Π of X is isotropic iff $\int_{\mathbb{H}_m} (x \otimes x) \Pi(dx) = \mathrm{Id}_{\mathbb{H}_m}$, or

$$\|A\|_{L_2(\Pi)} = \|A\|_2, \qquad A \in \mathbb{H}_m.$$

Suppose that the independent "design matrices" $X_1, \dots, X_n$ have been sampled from distributions $\Pi_1, \dots, \Pi_n$ in $\mathbb{H}_m$. Let

$$\bar{\Pi} := \bar{\Pi}_n = n^{-1} \sum_{j=1}^{n} \Pi_j.$$

In what follows, it will be assumed that the distribution $\bar{\Pi}$ is isotropic (in particular, this is the case if all the distributions Π_j are isotropic). This implies that

$$n^{-1} \sum_{j=1}^{n} \mathbb{E}(X_j \otimes X_j) = n^{-1} \sum_{j=1}^{n} \int_{\mathbb{H}_m} (x \otimes x) \Pi_j(dx) = \int_{\mathbb{H}_m} (x \otimes x) \bar{\Pi}(dx) = \mathrm{Id}_{\mathbb{H}_m}$$

Note that, for all $j = 1, \dots, n$

$$\mathbb{E}(Y_j X_j) = \mathbb{E}\mathbb{E}(Y_j | X_j) X_j = \mathbb{E}\langle \rho, X_j \rangle X_j = \mathbb{E}(X_j \otimes X_j)\rho.$$

Thus,

$$\mathbb{E} n^{-1} \sum_{j=1}^{n} Y_j X_j = \rho.$$

Suppose that $\mathbb{D} \subset \mathbb{H}_m$ is a given closed convex subset of Hermitian matrices that is known to include the target state ρ. We will study the following estimator of ρ :

$$\hat{\rho} := \mathrm{argmin}_{S \in \mathbb{D}} \left[\|S\|_2^2 - 2 \left\langle n^{-1} \sum_{j=1}^{n} Y_j X_j, S \right\rangle \right]. \tag{2.1}$$

Replacing in (2.1) the matrix $n^{-1} \sum_{j=1}^{n} Y_j X_j$ by its expectation results in the following minimization problem

$$\|S\|_2^2 - 2\langle \rho, S \rangle = \|S - \rho\|_2^2 - \|\rho\|_2^2 \longrightarrow \min, \qquad S \in \mathbb{D},$$

whose unique solution is ρ. The estimator $\hat{\rho}$ can be viewed as a modification of the standard least squares estimator defined as a solution of the following minimization problem:

$$n^{-1} \sum_{j=1}^{n} \big(Y_j - \langle S, X_j \rangle \big)^2 \longrightarrow \min, \qquad S \in \mathbb{D}. \tag{2.2}$$

Indeed, (2.2) is equivalent to

$$\|S\|^2_{L_2(\hat{\Pi}_n)} - 2 \left\langle n^{-1} \sum_{j=1}^{n} Y_j X_j, S \right\rangle \longrightarrow \min, \qquad S \in \mathbb{D}, \tag{2.3}$$

where $\hat{\Pi}_n$ is the empirical measure based on $(X_1, \dots, X_n)$. In (2.1), $\|S\|^2_{L_2(\hat{\Pi}_n)}$ is replaced by its expectation:

$$\mathbb{E}\|S\|^2_{L_2(\hat{\Pi}_n)} = \mathbb{E} n^{-1} \sum_{j=1}^{n} \langle S, X_j \rangle^2 = n^{-1} \sum_{j=1}^{n} \|S\|^2_{L_2(\Pi_j)} = \|S\|^2_{L_2(\bar{\Pi})} = \|S\|_2^2$$

since $\bar{\Pi}$ is isotropic.

Similar approach was used in Koltchinskii, Lounici and Tsybakov [11] where it was assumed that the design distributions $\Pi_j, j = 1, \dots, n$ are known (not necessarily isotropic) and the empirical norm $\|S\|^2_{L_2(\hat{\Pi}_n)}$ in a version of problem (2.3) with nuclear norm penalization was replaced by its expectation $n^{-1}\sum_{j=1}^n \|S\|^2_{L_2(\Pi_j)}$ (in this paper, the domain $\mathbb{D}$ was a linear space of matrices rather than a subset of density matrices). Koltchinskii [10] studied a penalized version of problem (2.2) with the complexity penalty $\varepsilon \mathrm{tr}(S\log S) = -\varepsilon \mathcal{E}(S)$, where $\mathcal{E}(S)$ is the von Neumann entropy of S and $\varepsilon > 0$ is a regularization parameter.

Our main observation in this note is that, even without any regularization, the error of the estimator $\hat{\rho}$ defined by (2.1) can be controlled in terms of the rank of the target matrix ρ (or in terms of low rank matrices approximating ρ). The same is true for the least squares estimator (2.2) with somewhat different error bounds and with a little bit more involved proofs (see Koltchinskii [9], Chapter 9).

In what follows, we denote

$$\Xi := n^{-1}\sum_{j=1}^n Y_j X_j - \rho \quad \text{and} \quad \Delta := \|\Xi\|.$$

In the next section, we will describe noncommutative Bernstein type inequalities that give upper bounds on Δ. The following statement provides a way to control the Hilbert–Schmidt error of $\hat{\rho}$ in terms of quantity Δ and the rank of the target matrix ρ.

Theorem 1. *Suppose that $\mathbb{D} \subset \mathcal{S}$ is a closed convex set, $\rho \in \mathbb{D}$ and $\hat{\rho}$ is defined by (2.1). Then, the following bound holds:*

$$\|\hat{\rho} - \rho\|_2^2 \le \min\bigl(4\Delta, (1+\sqrt{2})^2\Delta^2 \mathrm{rank}(\rho)\bigr). \tag{2.4}$$

Proof. The argument is very similar to the proofs of some of the results by Koltchinskii, Lounici and Tsybakov [11]. We give the proof for completeness. It is well known (see, e.g., Watson [17]) that the subdifferential of the nuclear norm of Hermitian matrices is given by the following formula:

$$\partial\|S\|_1 = \bigl\{\mathrm{sign}(S) + P_{L^\perp} W P_{L^\perp} : W \in \mathbb{H}_m, \|W\| \le 1\bigr\}. \tag{2.5}$$

Here S is a Hermitian matrix with spectral representation $S = \sum_{j=1}^r \lambda_j(\phi_j \otimes \phi_j)$, r being the rank of S, $\lambda_j, j = 1, \dots, r$ being its nonzero eigenvalues and $\phi_j, j = 1, \dots, r$ being the corresponding eigenvectors;

$$\mathrm{sign}(S) := \sum_{j=1}^r \mathrm{sign}(\lambda_j)(\phi_j \otimes \phi_j);$$

$L := \mathrm{l.s.}(\phi_1, \dots, \phi_r)$; $L^\perp$ is the orthogonal complement of L and $P_{L^\perp}$ the corresponding orthogonal projection. In particular, for all $V \in \partial\|S\|_1$, we have $\|V\| \le 1$.

Note that, since $\|S\|_1 = \mathrm{tr}(S) = 1$, $S \in \mathbb{D} \subset \mathcal{S}$, we have

$$\hat{\rho} = \mathrm{argmin}_{S\in\mathbb{D}} L_n(S),$$

where

$$L_n(S) := \|S\|_2^2 - 2\left\langle n^{-1}\sum_{j=1}^n Y_j X_j, S\right\rangle + 2\Delta\|S\|_1.$$

We will use a well known necessary condition of minimum of a convex function (see Aubin and Ekeland [3], Chapter 4, Section 2, Corollary 6): since L_n is a convex function in $\mathbb{H}_m$, $\mathbb{D} \subset \mathbb{H}_m$ is closed and convex and $\hat{\rho}$ is a minimal point of L_n in $\mathbb{D}$, we have

$$0 \in \partial L_n(\hat{\rho}) + N_{\mathbb{D}}(\hat{\rho}),$$

where $\partial L_n(\hat{\rho})$ is the subdifferential of L_n at the point $\hat{\rho}$ and $N_{\mathbb{D}}(\hat{\rho})$ is the normal cone of the convex set $\mathbb{D}$ at $\hat{\rho}$ (see [3] for precise definitions). Thus, there exists a point $C \in \partial L_n(\hat{\rho}) \cap (-N_{\mathbb{D}}(\hat{\rho}))$. It follows from the definition of the normal cone, that $\langle C, \hat{\rho} - S\rangle \le 0$ for all $S \in \mathbb{D}$. Note also that

$$C = 2\hat{\rho} - \frac{2}{n}\sum_{j=1}^{n} Y_j X_j + 2\Delta\hat{V} = 2\hat{\rho} - 2\rho - 2\Xi + 2\Delta\hat{V},$$

for some $\hat{V} \in \partial\|\hat{\rho}\|_1$. Therefore, we have

$$2\langle\hat{\rho} - \rho, \hat{\rho} - \rho\rangle - 2\langle\Xi, \hat{\rho} - \rho\rangle + 2\Delta\langle\hat{V}, \hat{\rho} - \rho\rangle \le 0.$$

This implies that, for any $V \in \partial\|\rho\|_1$,

$$\|\hat{\rho} - \rho\|_2^2 + \Delta\langle\hat{V} - V, \hat{\rho} - \rho\rangle \le \langle\Xi, \hat{\rho} - \rho\rangle - \Delta\langle V, \hat{\rho} - \rho\rangle.$$

Using monotonicity of subdifferentials of convex functions, we conclude that

$$\langle\hat{V} - V, \hat{\rho} - \rho\rangle \ge 0$$

and

$$\|\hat{\rho} - \rho\|_2^2 \le \langle\Xi, \hat{\rho} - \rho\rangle - \Delta\langle V, \hat{\rho} - \rho\rangle. \tag{2.6}$$

Observe that

$$|\langle\Xi, \hat{\rho} - \rho\rangle| \le \|\Xi\|\|\hat{\rho} - \rho\|_1 \le \Delta(\|\hat{\rho}\|_1 + \|\rho\|_1) = 2\Delta,$$

since $\hat{\rho}, \rho \in \mathbb{D} \subset \mathcal{S}$. Also,

$$|\langle V, \hat{\rho} - \rho\rangle| \le \|V\|\|\hat{\rho} - \rho\|_1 \le 2\|V\| \le 2,$$

since, for $V \in \partial\|\rho\|_1$, $\|V\| \le 1$. Thus, (2.6) implies that $\|\hat{\rho} - \rho\|_2^2 \le 4\Delta$.

We will use (2.5) for $S = \rho$ (assuming that $\rho = \sum_{j=1}^{r} \lambda_j(\phi_j \otimes \phi_j)$ and $L = \text{l.s.}(\phi_1, \dots, \phi_r)$). Substitute in (2.6) $V = \text{sign}(\rho) + P_{L^\perp} W P_{L^\perp}$, where $W \in \mathbb{H}_m$, $\|W\| \le 1$ and

$$\langle P_{L^\perp} W P_{L^\perp}, \hat{\rho} - \rho\rangle = \langle P_{L^\perp} W P_{L^\perp}, \hat{\rho}\rangle = \langle W, P_{L^\perp}\hat{\rho}P_{L^\perp}\rangle = \|P_{L^\perp}\hat{\rho}P_{L^\perp}\|_1$$

(the existence of such a W easily follows from the duality between nuclear and operator norms). For such a $V \in \partial\|\rho\|_1$, (2.6) yields

$$\|\hat{\rho} - \rho\|_2^2 + \Delta\|P_{L^\perp}\hat{\rho}P_{L^\perp}\|_1 \le \langle\Xi, \hat{\rho} - \rho\rangle - \Delta\big\langle\text{sign}(\rho), \hat{\rho} - \rho\big\rangle. \tag{2.7}$$

It remains to bound the right hand side from above. To this end, note that

$$\big|\big\langle\text{sign}(\rho), \hat{\rho} - \rho\big\rangle\big| \le \|\text{sign}(\rho)\|_2\|\hat{\rho} - \rho\|_2 = \sqrt{\text{rank}(\rho)}\|\hat{\rho} - \rho\|_2. \tag{2.8}$$

Let $\mathcal{P}_L, \mathcal{P}_L^{\perp} : \mathbb{H}_m \mapsto \mathbb{H}_m$ be the orthogonal projectors defined as follows:

$$\mathcal{P}_L A := A - P_{L^{\perp}} A P_{L^{\perp}}, \qquad \mathcal{P}_L^{\perp} A := P_{L^{\perp}} A P_{L^{\perp}}, \quad A \in \mathbb{H}_m.$$

With these notations, we get

$$\begin{aligned} |\langle \Xi, \hat{\rho} - \rho \rangle| &\le |\langle \Xi, \mathcal{P}_L(\hat{\rho} - \rho) \rangle| + |\langle \Xi, \mathcal{P}_L^{\perp}(\hat{\rho} - \rho) \rangle| \\ &\le \|\mathcal{P}_L \Xi\|_2 \|\hat{\rho} - \rho\|_2 + \|\Xi\| \|P_{L^{\perp}} \hat{\rho} P_{L^{\perp}}\|_1 \\ &\le \Delta \sqrt{2\mathrm{rank}(\rho)} \|\hat{\rho} - \rho\|_2 + \Delta \|P_{L^{\perp}} \hat{\rho} P_{L^{\perp}}\|_1, \end{aligned} \tag{2.9}$$

where we used the bound

$$\|\mathcal{P}_L \Xi\|_2^2 = \|P_L \Xi\|_2^2 + \|P_{L^{\perp}} \Xi P_L\|_2^2 \le 2\mathrm{rank}(\rho) \|\Xi\|^2 = 2\mathrm{rank}(\rho) \Delta^2$$

that holds because $\|P_L \Xi\| \le \|\Xi\|, \|P_{L^{\perp}} \Xi P_L\| \le \|\Xi\|$ and

$$\mathrm{rank}(P_L \Xi) \le \dim(L) = \mathrm{rank}(\rho), \qquad \mathrm{rank}(P_{L^{\perp}} \Xi P_L) \le \dim(L) = \mathrm{rank}(\rho).$$

Substituting (2.8) and (2.9) in (2.7), we get

$$\|\hat{\rho} - \rho\|_2^2 + \Delta \|P_{L^{\perp}} \hat{\rho} P_{L^{\perp}}\|_1 \le (1 + \sqrt{2}) \Delta \sqrt{\mathrm{rank}(\rho)} \|\hat{\rho} - \rho\|_2 + \Delta \|P_{L^{\perp}} \hat{\rho} P_{L^{\perp}}\|_1,$$

implying

$$\|\hat{\rho} - \rho\|_2^2 \le (1 + \sqrt{2})^2 \Delta^2 \mathrm{rank}(\rho),$$

which completes the proof. □

With a minor modification of the proof, it is easy to obtain the following more general *low rank oracle inequality* that provides a way to control the estimation error $\|\hat{\rho} - \rho\|_2^2$ in terms of low rank oracles $S \in \mathbb{D}$ with a small approximation error $\|S - \rho\|_2^2$.

Theorem 2. *Under the assumptions of Theorem 1, the following bound holds:*

$$\|\hat{\rho} - \rho\|_2^2 \le \inf_{S \in \mathbb{D}} [\|S - \rho\|_2^2 + (1 + \sqrt{2})^2 \Delta^2 \mathrm{rank}(S)].$$

Thus, the problem of bounding the Hilbert–Schmidt error $\|\hat{\rho} - \rho\|_2^2$ is reduced to bounding the operator norm Δ of a sum of independent Hermitian random matrices. This will be done using noncommutative Bernstein type inequalities discussed in the following section.

3. Bernstein type inequalities for sums of independent random matrices

Let $X_1, \dots, X_n$ be independent Hermitian $m \times m$ random matrices. Suppose that, for some constant $U > 0$, $\|X_j\| \le U$, $j = 1, \dots, n$ and that $\mathbb{E} X_j = 0$, $j = 1, \dots, n$. Denote

$$S_n := X_1 + \cdots + X_n \quad \text{and} \quad B_n := \|\mathbb{E}(X_1^2 + \cdots + X_n^2)\|.$$

The following remarkable extension of the classical Bernstein inequality goes back to Ahlswede and Winter [1]. The precise form of the inequality below, in particular, of variance B_n is due to Tropp [16].

Theorem 3. *For all* $t > 0$,

$$\mathbb{P}\{\|S_n\| \geq t\} \leq 2m \exp\left\{-\frac{t^2}{2B_n + 2Ut/3}\right\}. \tag{3.1}$$

We will also discuss a version of Bernstein inequality in the cases when $\|X_j\|$ are not necessarily uniformly bounded, but rather have subexponential tails. Recall the definition of Orlicz norms (see, e.g., van der Vaart and Wellner [18], p. 95). Namely, let ψ be a convex nondecreasing function from $\mathbb{R}_+$ into $\mathbb{R}_+$ with $\psi(0) = 0$. For a random variable ξ on a probability space $(\Omega, \Sigma, \mathbb{P})$, define

$$\|\xi\|_\psi := \inf\left\{C > 0 : \mathbb{E}\psi\left(\frac{|\xi|}{C}\right) \leq 1\right\}.$$

The most common choices of ψ are $\psi(u) = u^p$, $p \geq 1$ (in this case, the ψ-norm is just the L_p-norm), $\psi(u) = \psi_1(u) := e^u - 1$ (subexponential tails) and $\psi(u) = \psi_2(u) = e^{u^2} - 1$ (subgaussian tails). In what follows, we assume that

$$\psi(u) \geq e^u - 1 - u, \qquad u \geq 1 \quad \text{and} \quad \psi(u) \geq u^p, \qquad u \geq 0 \text{ for some } p \geq 1. \tag{3.2}$$

Denote $\sigma^2 := n^{-1}B_n$.

Theorem 4. *Suppose that, for some* $M > 0$,

$$\max_{1\leq j\leq n} \big\|\|X_j\|\big\|_{\psi^2} \leq M.$$

Let $\delta \in (0, \frac{2}{\psi(1)})$ *and*

$$\bar{U} := M\psi^{-1}\left(\frac{2}{\delta}\frac{M^2}{\sigma^2}\right).$$

Then, for $t\bar{U} \leq (e-1)(1+\delta)B_n$,

$$\mathbb{P}\{\|S_n\| \geq t\} \leq 2m \exp\left\{-\frac{t^2}{2(1+\delta)B_n + 2\bar{U}t/3}\right\} \tag{3.3}$$

and, for $t\bar{U} > (e-1)(1+\delta)B_n$,

$$\mathbb{P}\{\|S_n\| \geq t\} \leq 2m \exp\left\{-\frac{t}{(e-1)\bar{U}}\right\}. \tag{3.4}$$

As a standard example, consider the case when $\psi(u) = \psi_\alpha(u) = e^{u^\alpha} - 1, u \geq 0, \alpha \geq 1$. Clearly, conditions (3.2) are satisfied with $p = \alpha$. Also, $\psi_\alpha^2(u) \leq e^{2u^\alpha} - 1$. This implies that

$$\|\xi\|_{\psi_\alpha^2} \leq 2^{1/\alpha}\|\xi\|_{\psi_\alpha}.$$

Finally, we have $\psi^{-1}(u) = \log^{1/\alpha}(1+u)$.

Corollary 1. *Let* $\delta \in (0, \frac{2}{e-1})$,

$$M = M^{(\alpha)} = 2^{1/\alpha} \max_{1\leq j\leq n} \big\|\|X_j\|\big\|_{\psi_\alpha}$$

and

$$\bar{U} = \bar{U}^{(\alpha)} = M \log^{1/\alpha}\left(\frac{2}{\delta}\frac{M^2}{\sigma^2} + 1\right).$$

Then bounds (3.3) and (3.4) hold.

Note that, if for some constant $U > 0$, $\|X_j\| \le U, j = 1, \dots, n$, then, for all $\delta > 0$,

$$\limsup_{\alpha\to\infty} M^{(\alpha)} \le U, \qquad \limsup_{\alpha\to\infty} \bar{U}^{(\alpha)} \le U.$$

Passing in (3.3) and (3.4) to the limit when $\alpha \to \infty$ and then when $\delta \to 0$, yields the bounds

$$\mathbb{P}\{\|S_n\| \ge t\} \le 2m \exp\left\{-\frac{t^2}{2B_n + 2Ut/3}\right\}$$

for $tU < (e-1)B_n$ and

$$\mathbb{P}\{\|S_n\| \ge t\} \le 2m \exp\left\{-\frac{t}{(e-1)U}\right\}$$

for $tU > (e-1)B_n$, almost recovering the result of Theorem 3.

Proof. For completeness, we prove both Theorems 3 and 4. Denote $\lambda^+(A), \lambda^-(A)$ the largest and the smallest eigenvalues of $A \in \mathbb{H}_m$. Clearly, $\|S_n\| \ge t$ if and only if $\lambda^+(S_n) \ge t$, or $\lambda^-(S_n) \le -t$, implying that

$$\mathbb{P}\{\|S_n\| \ge t\} \le \mathbb{P}\{\lambda^+(S_n) \ge t\} + \mathbb{P}\{\lambda^-(S_n) \le -t\}. \tag{3.5}$$

It is enough to control only one of the probabilities in the right hand side (another one is controlled similarly). For all $\lambda > 0$, we have

$$\mathbb{P}\{\lambda^+(S_n) \ge t\} \le \mathbb{P}\{\mathrm{tr}(e^{\lambda S_n}) \ge e^{\lambda t}\} \le e^{-\lambda t}\mathbb{E}\mathrm{tr}(e^{\lambda S_n}). \tag{3.6}$$

To bound $\mathbb{E}\mathrm{tr}(e^{\lambda S_n})$, we will use an approach by Tropp [16] that relies on the following lemma due to Lieb [12]. The original approach by Ahlswede and Winter [1] was based on Golden–Thompson inequality $\mathrm{tr}(e^{A+B}) \le \mathrm{tr}(e^A e^B), A, B \in \mathbb{H}_m$; it gives the same bound in the i.i.d. case and a slightly weaker bound in the general case.

Lemma 1. *For all $B \in \mathbb{H}_m$, the function*

$$F_B(S) := \mathrm{tr}\exp\{B + \log S\}$$

is concave on the cone $\{S : S \in \mathbb{H}_m, S \ge 0\}$.

Tropp's approach is based on the following bound:

$$\mathbb{E}\mathrm{tr}(e^{\lambda S_n}) \le \mathrm{tr}\exp\{\log \mathbb{E}e^{\lambda X_1} + \log \mathbb{E}e^{\lambda X_2} + \cdots + \log \mathbb{E}e^{\lambda X_n}\}. \tag{3.7}$$

To prove (3.7), denote by $\mathbb{E}_n$ the conditional expectation given $X_1, \dots, X_{n-1}$ and use iteratively Lieb's lemma combined with Jensen's inequality:

$$\begin{aligned}
\mathbb{E}\mathrm{tr}(e^{\lambda S_n}) &= \mathbb{E}\mathbb{E}_n \mathrm{tr}\exp\{\lambda S_{n-1} + \log e^{\lambda X_n}\} = \mathbb{E}\mathbb{E}_n F_{\lambda S_{n-1}}(e^{\lambda X_n}) \\
&\le \mathbb{E}F_{\lambda S_{n-1}}(\mathbb{E}e^{\lambda X_n}) = \mathbb{E}\mathrm{tr}\exp\{\lambda S_{n-1} + \log \mathbb{E}e^{\lambda X_n}\} \\
&= \mathbb{E}\mathbb{E}_{n-1}\mathrm{tr}\exp\{\lambda S_{n-2} + \log \mathbb{E}e^{\lambda X_n} + \log e^{\lambda X_{n-1}}\} \\
&= \mathbb{E}\mathbb{E}_{n-1}F_{\lambda S_{n-2} + \log \mathbb{E}e^{\lambda X_n}}(e^{\lambda X_{n-1}}) \\
&\le \mathbb{E}\mathrm{tr}\exp\{\lambda S_{n-2} + \log \mathbb{E}e^{\lambda X_{n-1}} + \log \mathbb{E}e^{\lambda X_n}\} \le \dots \\
&\le \mathrm{tr}\exp\{\log \mathbb{E}e^{\lambda X_1} + \log \mathbb{E}e^{\lambda X_2} + \cdots + \log \mathbb{E}e^{\lambda X_n}\}.
\end{aligned}$$

In view of (3.7), it remains to bound $\mathbb{E}e^{\lambda X}$ for $X = X_1, \dots, X_n$. Denote

$$\phi(u) := \frac{e^u - 1 - u}{u^2}.$$

Under the assumptions $\mathbb{E}X = 0$ and $\|X\| \le U$, the Taylor expansion yields:

$$\begin{aligned}\mathbb{E}e^{\lambda X} &= I_m + \mathbb{E}\lambda^2 X^2\left[\frac{1}{2!} + \frac{\lambda X}{3!} + \frac{\lambda^2 X^2}{4!} + \dots\right]\\ &\le I_m + \lambda^2 \mathbb{E}X^2\left[\frac{1}{2!} + \frac{\lambda\|X\|}{3!} + \frac{\lambda^2\|X\|^2}{4!} + \dots\right]\\ &= I_m + \lambda^2 \mathbb{E}X^2\phi(\lambda\|X\|) \le I_m + \lambda^2\phi(\lambda U)\mathbb{E}X^2,\end{aligned}$$

which implies

$$\log \mathbb{E}e^{\lambda X} \le \lambda^2\phi(\lambda U)\mathbb{E}X^2.$$

It remains to substitute this bound in (3.7) (for each of random matrices X_j):

$$\begin{aligned}\mathbb{E}\mathrm{tr}\big(e^{\lambda S_n}\big) &\le \mathrm{tr}\exp\big\{\lambda^2\phi(\lambda U)\mathbb{E}\big(X_1^2 + \dots + X_n^2\big)\big\}\\ &\le m\exp\big\{\lambda^2\phi(\lambda U)\|\mathbb{E}\big(X_1^2 + \dots + X_n^2\big)\|\big\}.\end{aligned} \tag{3.8}$$

In view of (3.5), (3.6) and (3.8), the proof of Theorem 3 can be now completed exactly as in the case of classical Bernstein inequality.

We now turn to the proof of Theorem 4. Let $\tau > 0$. Again, using the Taylor expansion we easily get that

$$\begin{aligned}\log \mathbb{E}e^{\lambda X} &\le \lambda^2\mathbb{E}X^2\phi(\lambda\|X\|)\\ &\le \lambda^2\phi(\lambda\tau)\mathbb{E}X^2 + I_m\lambda^2\mathbb{E}\|X\|^2\phi(\lambda\|X\|)I(\|X\| \ge \tau).\end{aligned} \tag{3.9}$$

For $\lambda \le 1/M$, the second term in the right hand side can be bounded as follows:

$$\begin{aligned}&\lambda^2\mathbb{E}\|X\|^2\phi(\lambda\|X\|)I(\|X\| \ge \tau)\\ &\quad\le \lambda^2 M^2\mathbb{E}\frac{\|X\|^2}{M^2}\phi\left(\frac{\|X\|}{M}\right)I\left(\frac{\|X\|}{M} \ge \frac{\tau}{M}\right)\\ &\quad\le \lambda^2 M^2\mathbb{E}\left(\exp\left\{\frac{\|X\|}{M}\right\} - 1 - \frac{\|X\|}{M}\right)I\left(\frac{\|X\|}{M} \ge \frac{\tau}{M}\right)\\ &\quad\le \lambda^2 M^2\mathbb{E}\psi^2\left(\frac{\|X\|}{M}\right)\left(\psi\left(\frac{\tau}{M}\right)\right)^{-1} \le \lambda^2 M^2\left(\psi\left(\frac{\tau}{M}\right)\right)^{-1}.\end{aligned}$$

For $\tau = \bar{U}$, we have

$$M^2\left(\psi\left(\frac{\bar{U}}{M}\right)\right)^{-1} = \frac{\delta\sigma^2}{2}.$$

Thus, we get

$$\lambda^2\mathbb{E}\|X\|^2\phi(\lambda\|X\|)I(\|X\| \ge \bar{U}) \le \delta\sigma^2\frac{\lambda^2}{2}, \tag{3.10}$$

which, together with (3.9), yields

$$\log \mathbb{E}e^{\lambda X} \le \lambda^2\phi(\lambda\bar{U})\mathbb{E}X^2 + I_m\delta\sigma^2\frac{\lambda^2}{2}.$$

Substituting this bound for $X = X_j, j = 1, \dots, n$ into (3.7), we get that, for all $\lambda \le 1/M$,

$$\begin{aligned} \mathbb{E}\mathrm{tr}\big(e^{\lambda S_n}\big) &\le m \exp\left\{\lambda^2 \phi(\lambda \bar{U}) \|\mathbb{E}\big(X_1^2 + \cdots + X_n^2\big)\| + \delta n \sigma^2 \frac{\lambda^2}{2}\right\} \\ &\le m \exp\left\{\lambda^2 \phi(\lambda \bar{U}) B_n + \delta B_n \frac{\lambda^2}{2}\right\} \le m \exp\big\{\lambda^2 \phi(\lambda \bar{U})(1+\delta) B_n\big\}, \end{aligned} \tag{3.11}$$

where we used the fact $\lambda^2 \phi(\lambda \bar{U}) \ge \phi(0)\lambda^2 = \frac{\lambda^2}{2}$. It remains to use bounds (3.5), (3.6), (3.11) and to repeat a standard argument of the classical proof of Bernstein inequality to derive the bounds on tail probabilities from the bounds on the exponential moments and to complete the proof. □

4. Applications of Bernstein type inequalities

In this final section, we will use noncommutative Bernstein type inequalities to control the size of random variable Δ involved in the error bounds of Section 2. This would lead to error bounds with an explicit dependence on the important parameters of the problem such as the sample size n, the size and rank of the target matrix and the variance of the noise.

Recall that

$$\Delta = \|\Xi\| = \left\| n^{-1} \sum_{j=1}^n Y_j X_j - \rho \right\| = \left\| \frac{1}{n} \sum_{j=1}^n (\langle \rho, X_j \rangle X_j - \rho) + \frac{1}{n} \sum_{j=1}^n \xi_j X_j \right\|.$$

We will bound Δ by applying the bounds of Section 3 separately to

$$\left\| \frac{1}{n} \sum_{j=1}^n (\langle \rho, X_j \rangle X_j - \rho) \right\| \quad \text{and} \quad \left\| \frac{1}{n} \sum_{j=1}^n \xi_j X_j \right\|.$$

Assume that, for some constant $U_X \ge 0$,

$$\|X_j\| \le U_X, \qquad j = 1, \dots, n.$$

Let

$$\sigma_X^2 := \max_{1 \le j \le n} \|\mathbb{E} X_j^2\|.$$

Recall that $\mathbb{E}(\xi_j | X_j) = 0$ and suppose that for some constant $\sigma_\xi^2 \ge 0$

$$\mathbb{E}(\xi_j^2 | X_j) \le \sigma_\xi^2, \qquad j = 1, \dots, n \ \text{ a.s.}$$

Finally, we will assume that either, for some constant $U_\xi \ge 0$,

$$|\xi_j| \le U_\xi, \qquad j = 1, \dots, n \text{ a.s.},$$

or, for some $\alpha \ge 1$ and for some constant $M_\xi^{(\alpha)} \ge 0$,

$$2^{1/\alpha} \|\xi_j\|_{\psi_\alpha} \le M_\xi^{(\alpha)}, \qquad j = 1, \dots, n.$$

In the first case, set

$$U_{X,\xi} := U_X U_\xi;$$

in the second case, set

$$U_{X,\xi} := U_X M_\xi^{(\alpha)} \log^{1/\alpha}\left(2\frac{U_X M_\xi^{(\alpha)}}{\sigma_X \sigma_\xi} + 1\right).$$

Note also that σ_X, σ_ξ in the above definitions can be replaced by upper bounds not exceeding $U_X, M_\xi^{(\alpha)}$, respectively.

The following proposition is a direct and easy consequence of Theorem 3 and Corollary 1.

Proposition 1. *There exists a numerical constant $C > 0$ such that for all $t > 0$ with probability at least $1 - e^{-t}$*

$$\left\| \frac{1}{n}\sum_{j=1}^n \xi_j X_j \right\| \le C\left[\sigma_X \sigma_\xi \sqrt{\frac{t+\log(2m)}{n}} \bigvee U_{X,\xi}\frac{t+\log(2m)}{n}\right].$$

Suppose that there exists a constant $U_{\rho,X} \ge 0$ such that

$$|\langle \rho, X_j\rangle| \le U_{\rho,X}, \qquad j = 1, \dots, n \text{ a.s.}$$

Then the following proposition is an immediate consequence of Theorem 3.

Proposition 2. *There exists a numerical constant $C > 0$ such that for all $t > 0$ with probability at least $1 - e^{-t}$*

$$\left\| \frac{1}{n}\sum_{j=1}^n (\langle \rho, X_j\rangle X_j - \rho) \right\| \le C\left[\sigma_X U_{\rho,X} \sqrt{\frac{t+\log(2m)}{n}} \bigvee U_X U_{\rho,X}\frac{t+\log(2m)}{n}\right].$$

Let $t > 0$ and define

$$\varepsilon_{n,m} := \sigma_X(\sigma_\xi \vee U_{\rho,X})\sqrt{\frac{t+\log(2m)}{n}} \bigvee (U_{X,\xi} \vee U_X U_{\rho,X})\frac{t+\log(2m)}{n}.$$

The next statement follows from Theorems 1, 2 and Propositions 1, 2.

Corollary 2. *Suppose that $\mathbb{D} \subset \mathcal{S}$ is a closed convex set, $\rho \in \mathbb{D}$ and $\hat\rho$ is defined by (2.1). Then, there exists a numerical constant $C > 0$ such that the following bound holds with probability at least $1 - e^{-t}$:*

$$\|\hat\rho - \rho\|_2^2 \le C \min(\varepsilon_{n,m}, \varepsilon_{n,m}^2 \mathrm{rank}(\rho)).$$

Moreover, with the same probability,

$$\|\hat\rho - \rho\|_2^2 \le \inf_{S\in\mathbb{D}} [\|S - \rho\|_2^2 + C\varepsilon_{n,m}^2 \mathrm{rank}(S)].$$

We now turn to a popular model of *sampling from an orthonormal basis.* Namely, let $\{E_1, \dots, E_{m^2}\}$ be an orthonormal basis of Hermitian matrices and let Π be the uniform distribution in $\{E_1, \dots, E_{m^2}\}$. Let X be a random matrix sampled from Π. Conditionally on the random matrix X with spectral representation $\sum_j \lambda_j P_j$, where $\{\lambda_j\}$ are the eigenvalues and $\{P_j\}$ are the spectral projectors of X, let the random variable Y take values λ_j with probabilities $\mathrm{tr}(\rho P_j)$ (for the system prepared in the state ρ). We will assume that $(X_1, Y_1) \dots, (X_n, Y_n)$ are i.i.d. copies of (X, Y). Denote

$$U := \max_{1\le j\le m^2} \|E_j\|.$$

Clearly, $U \le 1$ since $\|E_j\| \le \|E_j\|_2 = 1$. Under this definition, we have $U_X = U$ and

$$|\langle \rho, X_j\rangle| \le \|\rho\|_1 \|X_j\| \le \|X_j\| \le U, \qquad j = 1, \dots, n$$

since $\|\rho\|_1 = \mathrm{tr}(\rho) = 1$. Thus, we can set $U_{\rho,X} = U$. We also have that $|Y_j| \le U$ (since the eigenvalues of X_j are in $[-U, U]$) and $|\xi_j| \le 2U$. Thus, we can take $U_\xi = 2U$ and replace σ_ξ by its upper bound $2U$. As a result, $U_{X,\xi} = 2U^2$. Finally, for an arbitrary orthonormal basis $\{e_1, \dots, e_m\}$ of $\mathbb{C}^m$, the following holds:

$$\begin{aligned}
\|\mathbb{E}X^2\| &= \sup_{v\in\mathbb{C}^m} \mathbb{E}\langle X^2 v, v\rangle = \sup_{v\in\mathbb{C}^m} \mathbb{E}|Xv|^2 = \sup_{v\in\mathbb{C}^m} \mathbb{E}\sum_{k=1}^m |\langle Xv, e_k\rangle|^2 \\
&= \sup_{v\in\mathbb{C}^m} \mathbb{E}\sum_{k=1}^m |\langle X, v\otimes e_k\rangle|^2 = \sup_{v\in\mathbb{C}^m} m^{-2}\sum_{k=1}^m\sum_{j=1}^{m^2} |\langle E_j, v\otimes e_k\rangle|^2 \\
&= \sup_{v\in\mathbb{C}^m} m^{-2}\sum_{k=1}^m \|v\otimes e_k\|_2^2 = m^{-2} \sup_{v\in\mathbb{C}^m}\sum_{k=1}^m |v|^2|e_k|^2 = m^{-1}.
\end{aligned}$$

Therefore, $\sigma_X = m^{-1/2}$.

Note that, in the case of sampling from an orthonormal basis, we have

$$\|A\|_{L_2(\Pi)}^2 = m^{-2}\|A\|_2^2, \qquad A \in \mathbb{H}_m$$

and $\mathbb{E}(X\otimes X) = m^{-2}\mathrm{Id}_{\mathbb{H}_m}$. Thus, the problem has to be rescaled in order to apply Corollary 2. To this end, let $X_j' = mX_j$ and $Y_j' = mY_j$. The estimator $\hat\rho$ defined by (2.1) should be now based on the data $(X_1', Y_1'), \dots (X_n', Y_n')$. In terms of $(X_1, Y_1), \dots, (X_n, Y_n)$, it can be written as

$$\hat\rho := \mathrm{argmin}_{S\in\mathbb{D}}\left[m^{-2}\|S\|_2^2 - 2\left\langle n^{-1}\sum_{j=1}^n Y_j X_j, S\right\rangle\right]. \tag{4.1}$$

A natural measure of its error is

$$\|\hat\rho - \rho\|_{L_2(\Pi)}^2 = m^{-2}\|\hat\rho - \rho\|_2^2.$$

Note that $\sigma_{X'} = m\sigma_X = m^{1/2}$, $U_{X'} = mU_X = mU$ and $U_{\rho,X'} = mU_{\rho,X} = mU$. Denoting $\xi_j' = Y_j' - \mathrm{tr}(\rho X_j')$, we also have $U_{\xi'} = mU_\xi = 2mU$, $\sigma_{\xi'} = m\sigma_\xi \le 2mU$ and $U_{X',\xi'} = 2m^2U^2$. This yields the following formula for $\varepsilon_{n,m}$:

$$\varepsilon_{n,m} = m^{3/2}U\sqrt{\frac{t+\log(2m)}{n}} \bigvee m^2U^2\frac{t+\log(2m)}{n}.$$

Under the assumption that

$$U^2\frac{m(t+\log(2m))}{n} \le 1,$$

we have

$$\varepsilon_{n,m} = m^{3/2}U\sqrt{\frac{t+\log(2m)}{n}},$$

and, it follows from Corollary 2 that with probability at least $1 - e^{-t}$,

$$\|\hat\rho - \rho\|_{L_2(\Pi)}^2 \le C\left(U\sqrt{\frac{t+\log(2m)}{mn}} \bigwedge \frac{mU^2\mathrm{rank}(\rho)(t+\log(2m))}{n}\right). \tag{4.2}$$

and

$$\|\hat{\rho}-\rho\|^2_{L_2(\Pi)} \le \inf_{S\in\mathbb{D}}\left[\|S-\rho\|^2_{L_2(\Pi)} + C\frac{mU^2\mathrm{rank}(S)(t+\log(2m))}{n}\right]. \tag{4.3}$$

As an interesting special example, consider the case of *Pauli basis*. Recall that *Pauli matrices* are defined as

$$\sigma_1 := \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad \sigma_2 := \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}, \quad \sigma_3 := \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \quad \text{and} \quad \sigma_4 := \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

The Pauli basis in the space $\mathbb{H}_2$ consists of the matrices $W_j := \frac{1}{\sqrt{2}}\sigma_j, j = 1,2,3,4$. If now $m = 2^k, k \ge 1$, then the Pauli basis in $\mathbb{H}_m$ consists of all tensor products (here $\otimes$ means the tensor product of linear transformations) $W_{i_1}\otimes\cdots\otimes W_{i_k}$, $(i_1,\dots,i_k) \in \{1,2,3,4\}^k$. This provides a measurement model for a k qubit system in quantum state and quantum process tomography (see Nielsen and Chuang [13], Section 8.4.2).

In the case of Pauli basis, we have that $\|W_i\| \le \frac{1}{\sqrt{2}}$ and, as a consequence, $\|W_{i_1}\otimes\cdots\otimes W_{i_k}\| \le 2^{-k/2} \le m^{-1/2}$, $(i_1,\dots,i_k) \in \{1,2,3,4\}^k$. Thus, $U = m^{-1/2}$ and bounds (4.2), (4.3) imply that with probability at least $1-e^{-t}$

$$\|\hat{\rho}-\rho\|^2_{L_2(\Pi)} \le C\left(\frac{1}{m}\sqrt{\frac{t+\log(2m)}{n}} \bigwedge \frac{\mathrm{rank}(\rho)(t+\log(2m))}{n}\right). \tag{4.4}$$

and

$$\|\hat{\rho}-\rho\|^2_{L_2(\Pi)} \le \inf_{S\in\mathbb{D}}\left[\|S-\rho\|^2_{L_2(\Pi)} + C\frac{\mathrm{rank}(S)(t+\log(2m))}{n}\right]. \tag{4.5}$$

References

[1] Ahlswede, R. and Winter, A. (2002). Strong converse for identification via quantum channels. *IEEE Trans. Inform. Theory* **48**(3) 569–679.

[2] Artiles, L. M., Gill, R. and Guta, M. I. (2005). An invitation to quantum tomography. *J. Royal Stat. Soc. Ser. B* **67**(1) 109–134.

[3] Aubin, J.-P. and Ekeland, I. (1984). *Applied Nonlinear Analysis*. J. Wiley & Sons, New York.

[4] Candes, E. and Recht, B. (2009). Exact matrix completion via convex optimization. *Foundat. Comput. Math.* **9**(6) 717–772.

[5] Candes, E. and Tao, T. (2010). The power of convex relaxation: Near-optimal matrix completion. *IEEE Trans. Inform. Theory* **56** 2053–2080.

[6] Candes, E. and Plan, Y. (2011). Tight oracle bounds for low-rank matrix recovery from a minimal number of random measurements. *IEEE Trans. Inform. Theory* **57**(4) 2342–2359.

[7] Gross, D., Liu, Y.-K., Flammia, S. T., Becker, S. and Eisert, J. (2010). Quantum state tomography via compressed sensing. *Phys. Rev. Lett.* **105**(15) 150401, October 2010.

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

[9] Koltchinskii, V. (2011). Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems. Ecole d'Eté de Probabilités de Saint-Flour XXXVIII-2008. *Lecture Notes in Mathematics*, Springer.

[10] Koltchinskii, V. (2011). Von Neumann entropy penalization and low rank matrix estimation. *Ann. Statist.* **39**(6) 2936–2973.
[11] Koltchinskii, V., Lounici, K. and Tsybakov, A. (2011). Nuclear norm penalization and optimal rates for noisy low rank matrix completion. *Ann. Statist.* **39**(5) 2302–2329.
[12] Lieb, E. H. (1973). Convex trace functions and the Wigner-Yanase-Dyson conjecture. *Adv. Math.* **11** 267–288.
[13] Nielsen, M. A. and Chuang, I. L. (2000). *Quantum Computation and Quantum Information*. Cambridge University Press.
[14] Recht, B. (2011). A simpler approach to matrix completion. *J. Mach. Learn. Res.* **12** 3413–3430.
[15] Rohde, A. and Tsybakov, A. (2011). Estimation of high-dimensional low rank matrices. *Ann. Statist.* **39**(2) 887–930.
[16] Tropp, J.-A. (2012). User-friendly tail bounds for sums of random matrices. *Found. Comput. Math.* **12**(4) 389–434.
[17] Watson, G. A. (1992). Characterization of the subdifferential of some matrix norms. *Linear Algebra Appl.* **170** 33–45.
[18] van der Vaart, A. and Wellner, J. (1996). *Weak Convergence and Empirical Processes. With Applications to Statistics.* Springer.