Von Neumann Entropy Penalization and Low Rank Matrix Estimation

Vladimir Koltchinskii *
School of Mathematics
Georgia Institute of Technology
Atlanta, GA 30332-0160
vlad@math.gatech.edu

October 6, 2011

Abstract

We study a problem of estimation of a Hermitian nonnegatively definite matrix ρ of unit trace (for instance, a density matrix of a quantum system) based on n i.i.d. measurements $(X_1, Y_1), \dots, (X_n, Y_n)$, where

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

$\{X_j\}$ being random i.i.d. Hermitian matrices and $\{\xi_j\}$ being i.i.d. random variables with $\mathbb{E}(\xi_j | X_j) = 0$. The estimator

$$\hat{\rho}^{\varepsilon} := \mathrm{argmin}_{S \in \mathcal{S}} \left[n^{-1} \sum_{j=1}^{n} (Y_j - \mathrm{tr}(SX_j))^2 + \varepsilon\ \mathrm{tr}(S \log S) \right]$$

is considered, where $\mathcal{S}$ is the set of all nonnegatively definite Hermitian $m \times m$ matrices of trace 1. The goal is to derive oracle inequalities showing how the estimation error depends on the accuracy of approximation of the unknown state ρ by low-rank matrices.

Keywords and phrases: low rank matrix estimation, von Neumann entropy, matrix regression, empirical processes, noncommutative Bernstein inequality, quantum state tomography, Pauli basis

2010 AMS Subject Classification: 62J99, 62H12, 60B20, 60G15, 81Q99

*Partially supported by NSF Grants DMS-0906880 and CCF-0808863

1 Introduction

Let $\mathbb{M}_m(\mathbb{C})$ be the set of all $m \times m$ matrices with complex entries. In what follows, $\mathrm{tr}(S)$ denotes the trace of $S \in \mathbb{M}_m(\mathbb{C})$ and S^* denotes its adjoint matrix. Let $\mathbb{H}_m(\mathbb{C})$ be the set of all Hermitian $m \times m$ matrices and let $\mathcal{S} := \{S \in \mathbb{H}_m(\mathbb{C}) : S \geq 0, \mathrm{tr}(S) = 1\}$ be the set of all nonnegatively definite Hermitian matrices of trace 1. The matrices from the set $\mathcal{S}$ can be interpreted, for instance, as *density matrices,* describing the states of a quantum system. Let $X \in \mathbb{H}_m(\mathbb{C})$ be a matrix (*an observable*) with spectral representation $X = \sum_{j=1}^m \lambda_j P_j$, where λ_j are the eigenvalues of X and P_j are its spectral projectors. Then, a measurement of X in a state $\rho \in \mathcal{S}$ would result in outcomes λ_j with probabilities $\mathrm{tr}(\rho P_j)$ and its expectation is $\mathbb{E}_\rho X = \mathrm{tr}(\rho X)$. Let $X_1, \ldots, X_n \in \mathbb{H}_m(\mathbb{C})$ be given matrices (observables) and let $\rho \in \mathcal{S}$ be an unknown state of the system. An important problem in *quantum state tomography* is to estimate ρ based on the observations $(X_j, Y_j),\ j = 1, \ldots, n$, where $Y_1, \ldots, Y_n$ are outcomes of measurements of the observables $X_1, \ldots, X_n$ for the system identically prepared n times in the state ρ. In other words, the unknown state ρ of the system is to be learned from a set of linear measurements in a number of "directions" $X_j, j = 1, \ldots, n$ (see Artiles, Gill and Guta (2004) for a general discussion of statistical problems in quantum state tomography). In what follows, it is assumed that the design variables $X_1, \ldots, X_n$ are also random, specifically, they are i.i.d. Hermitian $m \times m$ matrices with distribution Π. In this case, the observations $(X_1, Y_1), \ldots, (X_n, Y_n)$ are i.i.d. and they satisfy the following model

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

$\xi_j,\ j = 1, \ldots, n$ being i.i.d. random variables with $\mathbb{E}(\xi_j | X_j) = 0, j = 1, \ldots, n$.

A typical choice of the design variables already discussed in the literature (see Gross et al (2010), Gross (2011)) can be described as follows. The linear space of matrices $\mathbb{M}_m(\mathbb{C})$ can be equipped with the Hilbert-Schmidt inner product: $\langle A, B \rangle := \mathrm{tr}(AB^*)$. Let $E_i,\ i = 1, \ldots, m^2$ be an orthonormal basis of $\mathbb{M}_m(\mathbb{C})$ consisting of Hermitian matrices E_i. Let $X_j,\ j = 1, \ldots, n$ be i.i.d. random variables sampled from a distribution Π on the set $\{E_1, \ldots, E_{m^2}\}$. We will refer to this model as *sampling from an orthonormal basis.* Most often, the uniform distribution Π that assigns probability m^{-2} to each basis matrix E_i is used. Note that in this case $\mathbb{E}|\langle A, X \rangle|^2 = m^{-2}\|A\|_2^2$, where $\|\cdot\|_2 := \langle \cdot, \cdot \rangle^{1/2}$ is the Hilbert-Schmidt (or the Frobenius) norm.

The following simple example is related to the problems of *matrix completion* extensively discussed in the recent literature (see, e.g., Candes and Recht (2009), Candes

and Tao (2010) and references therein). More precisely, it deals with a version of matrix completion for Hermitian matrices (see Gross (2011)).

Example 1. Matrix completion. Let $\{e_i : i = 1, \dots, m\}$ be the canonical basis of $\mathbb{C}^m$. Then, the set of Hermitian matrices $\Big\{E_{jk} : 1 \le j, k \le m\Big\}$, where

$$E_{jj} := e_j \otimes e_j,\ j = 1, \dots, m,\ E_{jk} := \frac{1}{\sqrt{2}}\Big(e_j \otimes e_k + e_k \otimes e_j\Big),$$
$$E_{kj} := \frac{i}{\sqrt{2}}\Big(e_j \otimes e_k - e_k \otimes e_j\Big),\ j, k = 1, \dots, m, j < k,$$

forms an orthonormal basis of $\mathbb{H}_m(\mathbb{C})$. Here and in what follows $\otimes$ denotes the tensor product of vectors or matrices. For $j < k$, the Fourier coefficients of a Hermitian matrix ρ in this basis are equal to the real and imaginary parts of the entries $\rho_{kj}, j < k$ of matrix ρ multiplied by $\sqrt{2}$; for $j = k$, they are just the diagonal entries of ρ that are real. If now Π is the uniform distribution in this basis, then $\mathbb{E}|\langle A, X\rangle|^2 = m^{-2}\|A\|_2^2$. Sampling from this distribution is equivalent to sampling at random real and imaginary parts of the entries of matrix ρ.

Another example was studied by Gross et al (2010) and by Gross (2011). It is more directly related to the problems of quantum state tomography.

Example 2. Pauli basis. Let $m = 2^k$. Consider the *Pauli basis* in the space of 2×2 matrices $\mathbb{M}_2(\mathbb{C})$: $W_i := \frac{1}{\sqrt{2}}\sigma_i$, where

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

are the *Pauli matrices* (they are both Hermitian and unitary). The Pauli basis in $\mathbb{M}_2(\mathbb{C})$ can be extended to a basis in the space of $m \times m$ matrices $\mathbb{M}_m(\mathbb{C})$. These matrices define linear transformations acting in the linear space $\mathbb{C}^m = \mathbb{C}^{2^k}$ (that is, the k-fold tensor product of spaces $\mathbb{C}^2 : \mathbb{C}^{2^k} = (\mathbb{C}^2)^{\otimes k}$). Then, the Pauli basis in the space of matrices $\mathbb{M}_{2^k}(\mathbb{C})$ consists of all tensor products $W_{i_1} \otimes \cdots \otimes W_{i_k},\ (i_1, \dots, i_k) \in \{1, 2, 3, 4\}^k$. As before, $X_1, \dots, X_n$ are i.i.d. random variables sampled from this basis. Essentially, this is a standard measurement model for a k qubit system frequently used in quantum information, in particular, in quantum state and quantum process tomography (see Nielsen and Chuang (2000), section 8.4.2).

Example 3. Subgaussian design. Another interesting class of examples includes *subgaussian design matrices* X such that $\langle A, X\rangle$ is a subgaussian random variable for each $A \in \mathbb{H}_m(\mathbb{C})$. (Recall that a random variable η is called subgaussian with parameter

σ iff, for all $\lambda \in \mathbb{R}$, $\mathbb{E}e^{\lambda\eta} \le e^{\lambda^2\sigma^2/2}$). These examples are, probably, of less interest in applications to quantum state tomography, but this is an important model, closely related to randomized designs in compressed sensing, for which one can use powerful tools developed in the high-dimensional probability. For instance, one can consider the *Gaussian design,* where X is a symmetric random matrix with real entries such that $\{X_{ij} : 1 \le i \le j \le m\}$ are independent centered normal random variables with $\mathbb{E}X_{ii}^2 = 1,\ i = 1, \dots, m$ and $\mathbb{E}X_{ij}^2 = \frac{1}{2},\ i < j$. Alternatively, one can consider the *Rademacher design* assuming that $X_{ii} = \varepsilon_{ii},\ i = 1, \dots, m$ and $X_{ij} = \frac{1}{\sqrt{2}}\varepsilon_{ij},\ i < j$, where $\{\varepsilon_{ij} : 1 \le i \le j \le m\}$ are i.i.d. Rademacher random variables (that is, random variables taking values $+1$ or -1 with probability $1/2$ each). In both cases, $\mathbb{E}|\langle A, X\rangle|^2 = \|A\|_2^2,\ A \in \mathbb{M}_m(\mathbb{C})$ (such random matrices X will be called *isotropic*) and $\langle A, X\rangle$ is a subgaussian random variable whose subgaussian parameter is equal to $\|A\|_2$ (up to a constant).

The problems of this nature belong to a rapidly growing area of low rank matrix recovery. The most popular methods developed so far are based on nuclear norm regularization. In what follows, the Euclidean norm in the space $\mathbb{C}^m$ will be denoted by $|\cdot|$ and the inner product will be denoted by $\langle\cdot,\cdot\rangle$ (with a little abuse of notation since it has been already used for the Hilbert–Schmidt inner product between matrices). We will denote by $\|\cdot\|_p, p \ge 1$ the *Schatten p-norm* of matrices in $\mathbb{M}_m(\mathbb{C})$ (and, if needed, in other matrix spaces). Specifically, $\|A\|_p := \left(\sum_{j=1}^m \lambda_k^p(|A|)\right)^{1/p}$, where $|A| := (A^*A)^{1/2}$ and, for a Hermitian matrix B, $\lambda_k(B), k = 1, \dots, m$ are the eigenvalues of B (usually arranged in the decreasing order). In particular, $\|\cdot\|_1$ is the usual nuclear norm and $\|\cdot\|_2$ is the Hilbert-Schmidt norm. We will use the notation $\|\cdot\|$ for the operator norm. Given a design distribution Π, we will write

$$\|A\|_{L_2(\Pi)}^2 := \int \langle A, x\rangle^2 \Pi(dx) = \mathbb{E}\langle A, X\rangle^2,\ A \in \mathbb{M}_m(\mathbb{C}),$$

where X is sampled from Π. We will often use the corresponding $L_2(\Pi)$-distance between matrices, that represents the prediction error in statistical problems in question.

In the noiseless case (i.e., when $\xi_j \equiv 0$), the following estimator of ρ has been extensively studied, especially, in the case of matrix completion problems (see Candes and Recht (2009), Candes and Tao (2010), Gross (2011), Recht (2009) and references therein):

$$\hat\rho := \operatorname{argmin}\left\{\|S\|_1 : S \in \mathbb{M}_m(\mathbb{C}), \langle S, X_j\rangle = Y_j, j = 1, \dots, n\right\}.$$

Under so called "low coherence assumptions" on the target matrix ρ, it was shown that, with a high probability, $\hat\rho = \rho$ provided that the number n of observations is sufficiently

large. Namely, up to logarithmic factors and constants, it should be of the order mr, where r is the rank of the target matrix ρ.

In the noisy case, the following penalized least squares estimator, which is akin to the LASSO used in sparse regression, was proposed and studied (see, e.g., Candes and Plan (2009), Rohde and Tsybakov (2011), Koltchinskii (2011) and references therein):

$$\hat{\rho}^{\varepsilon} := \mathrm{argmin}_{S \in \mathbb{M}_m(\mathbb{C})} \left[n^{-1} \sum_{j=1}^{n} (Y_j - \mathrm{tr}(SX_j))^2 + \varepsilon \|S\|_1 \right], \tag{1.1}$$

where ε is a regularization parameter. Candes and Plan (2009) have also studied an estimator based on nuclear norm minimization subject to linear constraints that resembles the Dantzig selector; Rohde and Tsybakov (2011) suggested estimators based on nonconvex penalties involving Schatten "p-norms" for $p < 1$; Koltchinskii, Lounici and Tsybakov (2011) studied a modification of nuclear norm penalized least squares estimator that requires the precise knowledge of the design distribution.

We will study the following estimator of the unknown state ρ defined as a solution of a penalized empirical risk minimization problem

$$\hat{\rho}^{\varepsilon} := \mathrm{argmin}_{S \in \mathcal{S}} \left[n^{-1} \sum_{j=1}^{n} (Y_j - \mathrm{tr}(SX_j))^2 + \varepsilon \ \mathrm{tr}(S \log S) \right], \tag{1.2}$$

where $\varepsilon > 0$ is a regularization parameter. The penalty term is based on the functional $\mathrm{tr}(S \log S) = -\mathcal{E}(S)$, where $\mathcal{E}(S)$ is the *von Neumann entropy* of state S. Thus, the method considered in this paper is based on a trade-off between fitting the model by the least squares in the class of all density matrices and maximizing the entropy of the state. Note that optimization problem (1.2) is convex (this is based on convexity of the penalty term that follows from the concavity of von Neumann entropy, see Nielsen and Chuang (2000)). It is also easy to see that the solution $\hat{\rho}^{\varepsilon}$ of (1.2) is always a full rank matrix (see the proof of Proposition 3). Nevertheless, it will be shown that when the target matrix ρ is nearly low rank, $\hat{\rho}^{\varepsilon}$ is also well approximated by low rank matrices and the error $\|\hat{\rho}^{\varepsilon} - \rho\|^2_{L_2(\Pi)}$ can be controlled in terms of the "approximate rank" of ρ.

One can also consider a version of optimization problem (1.2) that is further constrained to a closed convex subset $\mathbb{D} \subset \mathcal{S}$ of density matrices containing the target matrix ρ. The analysis of such problems is exactly the same as in the case when $\mathbb{D} = \mathcal{S}$, considered in the paper, and the results are also the same (subject to straightforward modifications). In particular, when $\mathbb{D}$ is the set of all diagonal matrices with nonnegative diagonal entries summable to 1 and the design matrices X_j are also diagonal, this allows

one to deduce the results on sparse recovery in convex hulls of finite dictionaries via entropy penalization that are akin to what was obtained earlier by Koltchinskii (2009).

2 An Overview of Main Results

The results of this paper include oracle inequalities for the $L_2(\Pi)$-error of the empirical solution $\hat{\rho}^{\varepsilon}$. They will be stated in a general form in sections 5 and 6. Here we formulate them only in two of the special examples outlined in the Introduction: random sampling from an orthonormal basis and subgaussian isotropic design (such as Gaussian or Rademacher). Assume, for simplicity, that the noise $\{\xi_j\}$ is a sequence of i.i.d. $N(0, \sigma_{\xi}^2)$ random variables independent of $(X_1, \dots, X_n)$ (a Gaussian noise).

In what follows, we write $f(S) := \sum_{j=1}^m f(\lambda_j)(\phi_j \otimes \phi_j)$ for any Hermitian matrix S with spectral representation $S = \sum_{j=1}^m \lambda_j(\phi_j \otimes \phi_j)$ and any function f defined on a set that contains the spectrum of S.

First, we consider the case of sampling from an orthonormal basis $\{E_1, \dots, E_{m^2}\}$ of $\mathbb{M}_m(\mathbb{C})$ (that consists of Hermitian matrices). Let us call the distribution Π in $\{E_1, \dots, E_{m^2}\}$ *nearly uniform* iff there exist constants $c_1, c_2 > 0$ such that $\max_{1 \le j \le m^2} \Pi(\{E_j\}) \le c_1 m^{-2}$ and $\|A\|_{L_2(\Pi)}^2 \ge c_2 m^{-2} \|A\|_2^2, A \in \mathbb{H}_m(\mathbb{C})$. Clearly, both the matrix completion design (Example 1) and sampling from the Pauli basis (Example 2) are special cases of sampling from such nearly uniform distributions, so, the next result does apply to these two examples.

Let $t > 0$ be fixed and denote $t_m := t + \log(2m), \quad \tau_n := t + \log\log_2(2n)$.

To simplify the bounds, assume that $\log\log_2 n \le \log(2m)$ (so, $\tau_n \le t_m$), that $n \ge m t_m \log^2 m$, and, finally, that $\sigma_{\xi} \ge m^{-1/2}$. The last condition just means that the variance of the noise is not "too small" which allows one to suppress "exponential tail terms" in Bernstein type inequalities used in the derivation of the bounds.

Recall that $\rho \in \mathcal{S}$.

Theorem 1 *Suppose that X is sampled at random from a nearly uniform distribution Π. Then, there exists a constant $C > 0$ such that, for all $\varepsilon \in [0,1]$, with probability at least $1 - e^{-t}$,*

$$\|\hat{\rho}^{\varepsilon} - \rho\|_{L_2(\Pi)}^2 \le C\left[\varepsilon\left(\|\log \rho\| \wedge \log\left(\frac{m}{\varepsilon}\right)\right) \bigvee \sigma_{\xi}\sqrt{\frac{t_m}{nm}}\right]. \tag{2.1}$$

In addition, for all sufficiently large $D > 0$, there exists a constant $C > 0$ such that, for

$\varepsilon := D\sigma_\xi\sqrt{\frac{t_m}{mn}}$, *with probability at least* $1-e^{-t}$,

$$\|\hat{\rho}^\varepsilon - \rho\|^2_{L_2(\Pi)} \le \inf_{S\in\mathcal{S}}\biggl[2\|S-\rho\|^2_{L_2(\Pi)} + C\sigma_\xi^2\frac{\mathrm{rank}(S)mt_m\log^2(mn)}{n}\biggr]. \tag{2.2}$$

Theorem 1 follows from the results of Section 5 (see theorems 3 and 4, the Remark after Theorem 4 and Corollary 1). A simple consequence of Theorem 1 is the following bound

$$\|\hat{\rho}^\varepsilon - \rho\|^2_{L_2(\Pi)} \le C\biggl[\sigma_\xi\sqrt{\frac{t_m}{mn}}\log(mn)\bigwedge\sigma_\xi^2\frac{\mathrm{rank}(\rho)mt_m\log^2(mn)}{n}\biggr]$$

that holds with probability at least $1-e^{-t}$ and with some $C>0$ for $\varepsilon = D\sigma_\xi\sqrt{\frac{t_m}{mn}}$. It follows by substituting $S=\rho$ in bound (2.2) and combining it with (2.1).

Next we consider the case of subgaussian isotropic design for which $\|A\|_{L_2(\Pi)} = \|A\|_2,\ A\in\mathbb{M}_m(\mathbb{C})$. To simplify the bounds, we assume again that the noise is Gaussian.

Theorem 2 *Suppose X is a subgaussian isotropic matrix. There exist constants $C>0, c>0$ such that the following holds. Under the assumptions that $\tau_n \le cn$ and $t_m \le n$, for all $\varepsilon\in[0,1]$, with probability at least $1-e^{-t}$*

$$\|\hat{\rho}^\varepsilon - \rho\|^2_{L_2(\Pi)} \le C\biggl(\varepsilon\biggl(\|\log\rho\|\wedge\log\frac{m}{\varepsilon}\biggr)\bigvee\sigma_\xi\sqrt{\frac{mt_m}{n}}\bigvee (\sigma_\xi\vee\sqrt{m})\frac{\sqrt{m}(\tau_n\log n\vee t_m)}{n}\biggr)\biggr]. \tag{2.3}$$

Moreover, there exist a constant $c>0$ and, for all sufficiently large $D>0$, a constant $C>0$ such that, for $\varepsilon := D\sigma_\xi\sqrt{\frac{mt_m}{n}}$, with probability at least $1-e^{-t}$,

$$\|\hat{\rho}^\varepsilon-\rho\|^2_{L_2(\Pi)} \le \inf_{S\in\mathcal{S}}\biggl[2\|S-\rho\|^2_{L_2(\Pi)}+C\biggl(\frac{\sigma_\xi^2\mathrm{rank}(S)mt_m\log^2(mn)}{n}\bigvee\frac{m(\tau_n\log n\vee t_m)}{n}\biggr)\biggr]. \tag{2.4}$$

This theorem follows from the results of Section 6 (see theorems 5, 6 and Corollary 3). As it was the case with Theorem 1, one cas easily derive from Theorem 2 (by substituting $S=\rho$ in (2.4) and combining it with (2.3)) the following inequality that holds, for $\varepsilon := D\sigma_\xi\sqrt{\frac{mt_m}{n}}$, with probability at least $1-e^{-t}$ and with some $C>0$:

$$\|\hat{\rho}^\varepsilon-\rho\|^2_{L_2(\Pi)} \le C\biggl[\biggl(\sigma_\xi\sqrt{\frac{mt_m}{n}}\log\frac{m}{\varepsilon}\bigwedge\frac{\sigma_\xi^2\mathrm{rank}(\rho)mt_m\log^2(mn)}{n}\biggr)\bigvee\frac{m(\tau_n\log n\vee t_m)}{n}\biggr].$$

Note that the first bounds of theorems 1 and 2 (bounds (2.1) and (2.3)) hold for all $\varepsilon \geq 0$, even in the case of unpenalized least squares estimator with $\varepsilon = 0$. The random error parts of these bounds are (up to logarithmic factors) of the order $n^{-1/2}$ as $n \to \infty$. Bounds (2.2) and (2.4) are based on more subtle analysis taking into account the ranks of oracles S approximating the true density matrix ρ. In these bounds, the size of the $L_2(\Pi)$-error $\|\hat{\rho}^{\varepsilon} - \rho\|^2_{L_2(\Pi)}$ is determined by a trade-off between the approximation error $\|S - \rho\|^2_{L_2(\Pi)}$ of an oracle S and the random error. In the case of bounds (2.2) and (2.4), the last error is of the order $\frac{\sigma_{\xi}^2 \mathrm{rank}(S) m}{n}$ (up to logarithmic factors), and it depends on the rank of the oracle S. In particular, taking $S = \rho$, we can conclude that $\|\hat{\rho}^{\varepsilon} - \rho\|^2_{L_2(\Pi)}$ is bounded by $\frac{\sigma_{\xi}^2 \mathrm{rank}(\rho) m}{n}$ (up to constants and logarithmic factors). This means that von Neumann entropy penalization mimics oracles that know precisely which low rank matrices approximate ρ well and can estimate ρ by estimating a "small" number of parameters needed to describe such oracles. This is comparable with recent results for nuclear norm penalization. For instance, Candes and Plan (2009) obtained low rank oracle inequalities for the Frobenius norm under subgaussian type assumptions; Rohde and Tsybakov (2011) proved low rank bounds for the empirical prediction error; Koltchinskii, Lounici and Tsybakov (2011) obtained bounds of the same flavor as in theorems 2, 1, but for a modification of nuclear norm penalized least squares estimator in the case of known design distribution; Negahban and Wainwright (2010) proved similar inequalities for a version of nuclear penalization method with further constraints on the ℓ_{∞}-norm of the matrix. Depending on the values of σ_{ξ}, m, n and other characteristics of the problem more "rough" bounds (2.1) and (2.3) might become even sharper than more "subtle" bounds (2.2) and (2.4) (see Rohde and Tsybakov (2011) for a discussion of a similar phenomenon). Thus, the rate of convergence of the $L_2(\Pi)$-error to zero in a particular asymptotic scenario (when certain characteristics are large) is determined by the bounds of both types.

Theorems 1, 2 and other results of a similar nature will follow as corollaries from more general oracle inequalities that we establish under broader assumptions on the design distributions and on the noise. To prove these results, we need several tools from the empirical processes and random matrices theory, such as noncommutative Bernstein type inequalities and generic chaining bounds for empirical processes. We will discuss these results in Section 3 (as well as some properties of noncommutative Kullback-Leibler, Hellinger and other distances between density matrices). We will then study approximation error bounds for the solution of von Neumann entropy penalized true risk min-

imization problem (Section 4) and, finally, in sections 5 and 6, derive main results of the paper concerning random error bounds for the empirical solution $\hat{\rho}^{\varepsilon}$. More precisely, we bound the squared $L_2(\Pi)$-distance $\|\hat{\rho}^{\varepsilon} - S\|^2_{L_2(\Pi)}$ and symmetrized Kullback-Leibler distance $K(\hat{\rho}^{\varepsilon}; S)$ from $\hat{\rho}^{\varepsilon}$ to an arbitrary "oracle" $S \in \mathcal{S}$ and derive oracle inequalities for the squared $L_2(\Pi)$-error $\|\hat{\rho}^{\varepsilon} - \rho\|^2_{L_2(\Pi)}$ of the empirical solution $\hat{\rho}^{\varepsilon}$. These results are first established for oracles S of full rank and expressed in terms of certain characteristics of the operator $\log S$ (which is, essentially, a subgradient of the von Neumann entropy penalty used in (1.2)). Using simple techniques discussed in Section 4, we then develop the bounds for low rank oracles S (such as the bounds of theorems 1 and 2) and also obtain oracle inequalities for so called "Gibbs oracles".

Recently, several authors obtained minimax lower bounds on the errors of low rank matrix recovery, in particular, in matrix completion problems (see Rohde and Tsybakov (2011), Negahban and Wainwright (2010), Koltchinskii, Lounici and Tsybakov (2011) and references therein). Although it was not our goal in this paper, it would not be hard to extend these results to the framework of low rank density matrix estimation showing the optimality (up to logarithmic factors) of the main terms of our $L_2(\Pi)$-error bounds.

It is worth mentioning that the results of sections 4, 5 provide a way to bound the error of estimator $\hat{\rho}^{\varepsilon}$ not only in the $L_2(\Pi)$-distance, but also in other statistically important distances such as noncommutative Kullback-Leibler, Hellinger and nuclear norm distance.[1] For instance, under the assumptions of Theorem 2, the following bound for the Kullback-Leibler distance holds with probability at least $1 - e^{-t}$

$$K(\rho\|\hat{\rho}^{\varepsilon}) := \mathbb{E}_{\rho}(\log \rho - \log \hat{\rho}^{\varepsilon}) \leq \frac{C}{\varepsilon}\left[\frac{\sigma_{\xi}^2 \mathrm{rank}(\rho) m t_m \log^2(mn)}{n} \bigvee \frac{m(\tau_n \log n \vee t_m)}{n}\right] \quad (2.5)$$

for $\varepsilon := D\sigma_{\xi}\sqrt{\frac{mt_m}{n}}$. In the case of sampling from a nearly uniform distribution in an orthonormal basis (as in Theorem 1), it is easy to derive from Theorem 4 of Section 5 (using also some bounds from the proofs of Proposition 4 and Corollary 1) the following bound on the squared Hellinger distance between $\hat{\rho}^{\varepsilon}$ and ρ :

$$H^2(\hat{\rho}^{\varepsilon}; \rho) \leq C\sigma_{\xi}\frac{\mathrm{rank}(\rho) m^{3/2} t_m^{1/2} \log^2(mn)}{\sqrt{n}}$$

that holds with probability at least $1 - e^{-t}$ for $\varepsilon = D\sigma_{\xi}\sqrt{\frac{t_m}{mn}}$.

[1] A possibility to control Kullback-Leibler and Hellinger distances can be viewed as an advantage of von Neumann entropy penalization method.

3 Preliminaries: Distances in $\mathcal{S}$, Empirical Processes and Exponential Inequalities for Random Matrices

Noncommutative Kullback-Leibler and other distances. We will use noncommutative extensions of classical distances between probability distributions such as Kullback-Leibler and Hellinger distances. These extensions are common in quantum information theory (see Nielsen and Chuang (2000)). In particular, we will use *the symmetrized Kullback-Leibler distance* between two states $S_1, S_2 \in \mathcal{S}$ defined as

$$K(S_1; S_2) := \mathbb{E}_{S_1}(\log S_1 - \log S_2) + \mathbb{E}_{S_2}(\log S_2 - \log S_1) = \mathrm{tr}((S_1 - S_2)(\log S_1 - \log S_2)).$$

We will also use a noncommutative version of *Hellinger distance* defined as follows. For any two states $S_1, S_2 \in \mathcal{S}$, let $F(S_1, S_2) := \mathrm{tr}\sqrt{S_1^{1/2} S_2 S_1^{1/2}}$. This quantity is called the *fidelity* of states S_1, S_2 (see, e.g., Nielsen and Chuang (2000), p. 409). Then, a natural definition of the squared Hellinger distance is $H^2(S_1, S_2) := 2(1 - F(S_1, S_2))$. A remarkable property of this distance is that

$$H^2(S_1, S_2) = \sup H^2(\{p_i\}; \{q_i\}) = \sup \sum_i \left(\sqrt{p_i} - \sqrt{q_i}\right)^2,$$

where the supremum is taken over all POVMs $\{E_i\}$ (positive operator valued measures)[2] and $p_i := \mathrm{tr}(S_1 E_i), q_i := \mathrm{tr}(S_2 E_i)$. Thus, the quantum Hellinger distance is just the largest "classical" Hellinger distance between the probability distributions $\{p_i\}, \{q_i\}$ of a "measurement" $\{E_i\}$ in the states S_1, S_2 (see Nielsen and Chuang (2000), p. 412). The same property also holds for two other important "distances", the trace distance $\|S_1 - S_2\|_1$ and the Kullback-Leibler distance $K(S_1; S_2)$ (see, e.g., Klauck et al (2007)). These properties immediately imply an extension of classical inequalities for these distances:

$$\|S_1 - S_2\|_1^2 \le H^2(S_1, S_2) \le K(S_1; S_2).$$

They also imply the following simple proposition used below. It shows that, if two matrices S_1, S_2 are close in the Hellinger distance and one of them (say, S_2) is "approximately low rank" in the sense that there exists a subspace $L \subset \mathbb{C}^m$ of small dimension such that $\|P_{L^\perp} S_2 P_{L^\perp}\|_1$ is small, then another matrix S_1 is also "approximately low rank" with the same "support" L. [3]

[2] In the discrete case, a positive operator valued measure is a set $\{E_i\}$ of Hermitian nonnegatively definite matrices such that $\sum_i E_i = I$.

[3] Here and in what follows P_L denotes the orthogonal projection onto L and $L^\perp$ denotes the orthogonal complement of L.

Proposition 1 *For all subspaces $L \subset \mathbb{C}^m$ and all $S_1, S_2 \in \mathcal{S}$,*

$$\|P_L S_1 P_L\|_1 \le 2\|P_L S_2 P_L\|_1 + 2H^2(S_1, S_2).$$

Proof. Indeed, take an orthonormal basis $\{e_1, \dots, e_m\}$ in $\mathbb{C}^m$ such that $L = \text{l.s.}(\{e_1, \dots, e_k\})$. Let $p_j := \langle S_1 e_j, e_j\rangle = \text{tr}(S_1(e_j \otimes e_j))$ and $q_j := \langle S_2 e_j, e_j\rangle = \text{tr}(S_2(e_j \otimes e_j))$. Then

$$H^2(S_1, S_2) \ge \sum_{j=1}^m \Bigl(\sqrt{p_j} - \sqrt{q_j}\Bigr)^2 \ge \sum_{j=1}^k \Bigl(\sqrt{p_j} - \sqrt{q_j}\Bigr)^2 = \sum_{j=1}^k p_j + \sum_{j=1}^k q_j - 2\sum_{j=1}^k \sqrt{p_j}\sqrt{q_j},$$

which implies (using that $2\sqrt{ab} \le a/2 + 2b$)

$$\|P_L S_1 P_L\|_1 = \sum_{j=1}^k p_i \le 2\sum_{j=1}^k \sqrt{p_j}\sqrt{q_j} - \sum_{j=1}^k q_j + H^2(S_1, S_2) \le$$

$$\frac{1}{2}\sum_{j=1}^k p_j + \sum_{j=1}^k q_j + H^2(S_1, S_2) = \frac{1}{2}\|P_L S_1 P_L\|_1 + \|P_L S_2 P_L\|_1 + H^2(S_1, S_2),$$

and the result follows. □

Empirical processes bounds. We will use several inequalities for empirical processes indexed by a class of measurable functions $\mathcal{F}$ defined on an arbitrary measurable space $(S, \mathcal{A})$. Let $X, X_1, \dots, X_n$ be i.i.d. random variables in $(S, \mathcal{A})$ with common distribution P. If $\mathcal{F}$ is uniformly bounded by a number U, then Bousquet's version of the famous Talagrand's concentration inequality for empirical processes implies that, for all $t > 0$, with probability at least $1 - e^{-t}$

$$\sup_{f\in\mathcal{F}}\Bigl|n^{-1}\sum_{j=1}^n f(X_j) - \mathbb{E}f(X)\Bigr| \le 2\Bigl[\mathbb{E}\sup_{f\in\mathcal{F}}\Bigl|n^{-1}\sum_{j=1}^n f(X_j) - \mathbb{E}f(X)\Bigr| + \sigma\sqrt{\frac{t}{n}} + U\frac{t}{n}\Bigr],$$

where $\sigma^2 := \sup_{f\in\mathcal{F}} \text{Var}_P(f(X))$. We will also need a version of this bound for function classes that are not necessarily uniformly bounded. Such a bound was recently proved by Adamczak (2008). Recall that, for a convex increasing function ψ with $\psi(0) = 0$,

$$\|f\|_\psi := \inf\Bigl\{C > 0 : \int_S \psi\Bigl(\frac{|f|}{C}\Bigr)dP \le 1\Bigr\}$$

(see van der Vaart and Wellner (1996), p. 95). If $\psi(u) = u^p, u \ge 0$, for some $p \ge 1$, the corresponding ψ-norm is just the L_p-norm. Other important choices are functions $\psi_\alpha(t) = e^{t^\alpha} - 1, t \ge 0, \alpha \ge 1$, especially, ψ_2 that is related to subgaussian tails of f and ψ_1 that is related to subexponential tails. Let $F(x) \ge \sup_{f\in\mathcal{F}} |f(x)|, x \in S$, be an

envelope of the class. It follows from Theorem 4 of Adamczak (2008) that there exists a constant $K > 0$ such that for all $t > 0$ with probability at least $1 - e^{-t}$

$$\sup_{f\in\mathcal{F}}\left|n^{-1}\sum_{j=1}^{n} f(X_j)-\mathbb{E}f(X)\right| \le K\left[\mathbb{E}\sup_{f\in\mathcal{F}}\left|n^{-1}\sum_{j=1}^{n} f(X_j)-\mathbb{E}f(X)\right|+\sigma\sqrt{\frac{t}{n}}+\left\|\max_{1\le j\le n}|F(X_j)|\right\|_{\psi_1}\frac{t}{n}\right].$$

In addition to this, we will need bounds on empirical processes indexed by the class of "squares" $\{f^2 : f \in \mathcal{F}\}$ for a given function class $\mathcal{F}$. A usual approach to this problem is based on combining of symmetrization inequality with Talagrand's comparison (contraction) inequality for Rademacher sums (see, e.g., Ledoux and Talagrand (1991), Section 4.5). This, however, would require the class $\mathcal{F}$ to be uniformly bounded by a relatively small constant $U > 0$, which is not sufficient in the case of subgaussian design considered in the last section. A more subtle approach has been developed in the recent years by Klartag and Mendelson (2005), Mendelson (2010) and it is based on generic chaining bounds. Talagrand's *generic chaining complexity* (see Talagrand (2005)) of a metric space (T, d) is defined as follows. An admissible sequence $\{\Delta_n\}_{n\ge 0}$ is an increasing sequence of partitions of T (i.e., each next partition is a refinement of the previous one) such that $\mathrm{card}(\Delta_0) = 1$ and $\mathrm{card}(\Delta_n) \le 2^{2^n},\ n \ge 1$. For $t \in T$, $\Delta_n(t)$ denotes the unique subset in Δ_n that contains t. For a set $A \subset T$, $D(A)$ denotes its diameter. Then, define the generic chaining complexity $\gamma_2(T; d)$ as

$$\gamma_2(T; d) := \inf_{\{\Delta_n\}_{n\ge 0}} \sup_{t\in T} \sum_{n\ge 0} 2^{n/2} D(\Delta_n(t)),$$

where the inf is taken over all admissible sequences of partitions.

The generic chaining complexities were used by Talagrand (2005) to characterize the size of the expected sup-norms of Gaussian processes. Similar quantities can be also used to control the size of empirical processes indexed by a function class $\mathcal{F}$. It is natural to define $\gamma_2(\mathcal{F}; L_2(P))$, that is, $\gamma_2(\mathcal{F}; d)$, where d is the $L_2(P)$-distance. Some other distances are also useful, for instance, the ψ_2-distance associated with the probability space $(S, \mathcal{A}, P)$. The generic chaining complexity that corresponds to the ψ_2-distance will be denoted by $\gamma_2(\mathcal{F}; \psi_2)$. Mendelson (2010) proved the following deep result. Suppose that $\mathcal{F}$ is a symmetric class, that is, $f \in \mathcal{F}$ implies $-f \in \mathcal{F}$, and $Pf = \mathbb{E}f(X) = 0, f \in \mathcal{F}$. Then, for some universal constant $K > 0$,

$$\mathbb{E}\sup_{f\in\mathcal{F}}\left|n^{-1}\sum_{j=1}^{n} f^2(X_j) - \mathbb{E}f^2(X)\right| \le K\left[\sup_{f\in\mathcal{F}}\|f\|_{\psi_1}\frac{\gamma_2(\mathcal{F};\psi_2)}{\sqrt{n}} \bigvee \frac{\gamma_2^2(\mathcal{F};\psi_2)}{n}\right].$$

Noncommutative Bernstein type inequalities. We will need an operator version of Bernstein's inequality which is due to Ahlswede and Winter (2002) (and which has been already successfully used in the low rank recovery problems by Gross et al (2010), Gross (2011), Recht (2009)). Assume that $X, X_1, \dots, X_n$ are i.i.d. random Hermitian $m \times m$ matrices with $\mathbb{E}X = 0$ and $\sigma_X^2 := \|\mathbb{E}X^2\|$. The following bound is an easy consequence of Bernstein type inequality of Ahlswede and Winter (2002): *for all* $t > 0$, *with probability at least* $1 - e^{-t}$

$$\left\| \frac{X_1 + \cdots + X_n}{n} \right\| \le 2\left(\sigma_X \sqrt{\frac{t + \log(2m)}{n}} \bigvee U \frac{t + \log(2m)}{n} \right). \tag{3.1}$$

Moreover, it is possible to replace the L_∞-bound U on $\|X\|$ in the above inequality by bounds on the weaker ψ_α-norms (see also Koltchinskii (2011)). Namely, suppose that, for $\alpha \ge 1$ and for some constant $U_X^{(\alpha)}$, $U_X^{(\alpha)} \ge \max\left(\left\| \|X\| \right\|_{\psi_\alpha}, 2\mathbb{E}^{1/2}\|X\|^2 \right)$.

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

$$\left\| \frac{X_1 + \cdots + X_n}{n} \right\| \le C\left(\sigma_X \sqrt{\frac{t + \log(2m)}{n}} \bigvee U_X^{(\alpha)} \left(\log \frac{U_X^{(\alpha)}}{\sigma_X} \right)^{1/\alpha} \frac{t + \log(2m)}{n} \right). \tag{3.2}$$

Proof. Similarly to the proof of (3.1) discussed in the literature (Ahlswede and Winter (2002), Gross (2011)), we follow the standard derivation of classical Bernstein's inequality and we use the well known *Golden-Thompson inequality* [4] (see, e.g., Simon (1979), p. 94): for arbitrary $A, B \in \mathbb{H}_m(\mathbb{C})$, $\mathrm{tr}(e^{A+B}) \le \mathrm{tr}(e^A e^B)$. Let $Y_n := X_1 + \cdots + X_n$. Note that $\|Y_n\| < t$ if and only if $-tI_m < Y_n < tI_m$ (here and in what follows $A < B$ means that $B - A$ is positively definite). Therefore,

$$\mathbb{P}\{\|Y_n\| \ge t\} \le \mathbb{P}\{Y_n \not< tI_m\} + \mathbb{P}\{Y_n \not> -tI_m\}. \tag{3.3}$$

The following bounds are straightforward by simple matrix algebra:

$$\mathbb{P}\{Y_n \not< tI_m\} = \mathbb{P}\{e^{\lambda Y_n} \not< e^{\lambda t I_m}\} \le \mathbb{P}\Big\{ \mathrm{tr}\Big(e^{\lambda Y_n} \Big) \ge e^{\lambda t} \Big\} \le e^{-\lambda t} \mathbb{E}\mathrm{tr}(e^{\lambda Y_n}). \tag{3.4}$$

To bound the expected value in the right hand side, we use independence of random variables $X_1, \dots, X_n$ and Golden-Thompson inequality:

$$\mathbb{E}\mathrm{tr}(e^{\lambda Y_n}) = \mathbb{E}\mathrm{tr}\Big(e^{\lambda Y_{n-1} + \lambda X_n} \Big) \le \mathbb{E}\mathrm{tr}\Big(e^{\lambda Y_{n-1}} e^{\lambda X_n} \Big) = \mathrm{tr}\Big(\mathbb{E}\Big(e^{\lambda Y_{n-1}} e^{\lambda X_n} \Big) \Big) =$$

[4] See Oliveira (2010), Tropp (2010), Koltchinskii (2011) for other approahes that do not rely on Golden-Thompson inequality.

$$\mathrm{tr}\Big(\mathbb{E}e^{\lambda Y_{n-1}}\mathbb{E}e^{\lambda X_n}\Big) \le \mathbb{E}\mathrm{tr}\Big(e^{\lambda Y_{n-1}}\Big)\Big\|\mathbb{E}e^{\lambda X_n}\Big\|.$$

Since $\mathbb{E}\mathrm{tr}\Big(e^{\lambda X_1}\Big) = \mathrm{tr}\Big(\mathbb{E}e^{\lambda X_1}\Big) \le m\Big\|\mathbb{E}e^{\lambda X}\Big\|$, it is easy to conclude by induction that

$$\mathbb{E}\mathrm{tr}(e^{\lambda Y_n}) \le m\Big\|\mathbb{E}e^{\lambda X}\Big\|^n. \tag{3.5}$$

It remains to bound the norm $\|\mathbb{E}e^{\lambda X}\|$. To this end, we use Taylor expansion and the condition $\mathbb{E}X = 0$ to get

$$\mathbb{E}e^{\lambda X} = I_m + \mathbb{E}\lambda^2 X^2\left[\frac{1}{2!} + \frac{\lambda X}{3!} + \frac{\lambda^2 X^2}{4!} + \dots\right] \le$$

$$I_m + \lambda^2\mathbb{E}X^2\left[\frac{1}{2!} + \frac{\lambda\|X\|}{3!} + \frac{\lambda^2\|X\|^2}{4!} + \dots\right] = I_m + \lambda^2\mathbb{E}X^2\left[\frac{e^{\lambda\|X\|} - 1 - \lambda\|X\|}{\lambda^2\|X\|^2}\right].$$

Therefore, for all $\tau > 0$,

$$\Big\|\mathbb{E}e^{\lambda X}\Big\| \le 1 + \lambda^2\Big\|\mathbb{E}X^2\left[\frac{e^{\lambda\|X\|} - 1 - \lambda\|X\|}{\lambda^2\|X\|^2}\right]\Big\| \le$$

$$1 + \lambda^2\Big\|\mathbb{E}X^2\Big\|\left[\frac{e^{\lambda\tau} - 1 - \lambda\tau}{\lambda^2\tau^2}\right] + \lambda^2\mathbb{E}\|X\|^2\left[\frac{e^{\lambda\|X\|} - 1 - \lambda\|X\|}{\lambda^2\|X\|^2}\right]I(\|X\| \ge \tau).$$

Let $M := 2(\log 2)^{1/\alpha}U_X^{(\alpha)}$ and assume that $\lambda \le 1/M$. Then

$$\mathbb{E}\|X\|^2\left[\frac{e^{\lambda\|X\|} - 1 - \lambda\|X\|}{\lambda^2\|X\|^2}\right]I(\|X\| \ge \tau) \le M^2\mathbb{E}^{1/2}e^{2\|X\|/M}\mathbb{P}^{1/2}\{\|X\| \ge \tau\}.$$

Since, for $\alpha \ge 1$, $M = 2(\log 2)^{1/\alpha}U_X^{(\alpha)} \ge 2\Big\|\|X\|\Big\|_{\psi_1}$ (see van der Vaart and Wellner (1996), p. 95), we have $\mathbb{E}e^{2\|X\|/M} \le 2$ and also $\mathbb{P}\{\|X\| \ge \tau\} \le \exp\Big\{-2^\alpha\log 2\Big(\frac{\tau}{M}\Big)^\alpha\Big\}$. As a result, we get the following bound

$$\Big\|\mathbb{E}e^{\lambda X}\Big\| \le 1 + \lambda^2\sigma_X^2\left[\frac{e^{\lambda\tau} - 1 - \lambda\tau}{\lambda^2\tau^2}\right] + 2^{1/2}\lambda^2M^2\exp\Big\{-2^{\alpha-1}\log 2\Big(\frac{\tau}{M}\Big)^\alpha\Big\}.$$

Let $\tau := M\frac{2^{1/\alpha-1}}{(\log 2)^{1/\alpha}}\log^{1/\alpha}\frac{M^2}{\sigma_X^2}$ and suppose that λ satisfies the condition $\lambda\tau \le 1$. Then, the following bound holds with some constant $C_1 > 0$:

$$\Big\|\mathbb{E}e^{\lambda X}\Big\| \le 1 + C_1\lambda^2\sigma_X^2 \le \exp\{C_1\lambda^2\sigma_X^2\}.$$

Thus, we proved that there exist constants $C_1, C_2 > 0$ such that, for all λ satisfying the condition

$$\lambda\, U_X^{(\alpha)}\left(\log\frac{U_X^{(\alpha)}}{\sigma_X}\right)^{1/\alpha} \le C_2, \tag{3.6}$$

we have $\left\|\mathbb{E}e^{\lambda X}\right\| \le \exp\{C_1\lambda^2\sigma_X^2\}$. This can be combined with (3.3), (3.4) and (3.5) to get

$$\mathbb{P}\{\|Y_n\| \ge t\} \le 2m\exp\Big\{-\lambda t + C_1\lambda^2 n\sigma_X^2\Big\}.$$

It remains now to minimize the last bound with respect to all λ satisfying (3.6) to get that, for some constant $K > 0$,

$$\mathbb{P}\{\|Y_n\| \ge t\} \le 2m\exp\left\{-\frac{1}{K}\frac{t^2}{n\sigma_X^2 + tU_X^{(\alpha)}\log^{1/\alpha}(U_X^{(\alpha)}/\sigma_X)}\right\},$$

which immediately implies (3.2).

□

Note that, in the limit $\alpha \to \infty$, inequality (3.2) coincides with (3.1) (up to a constant).

4 Approximation Error

A natural first step in the analysis of the problem is to study its version with the true risk instead of the empirical risk. The true risk with respect to the quadratic loss is equal to $\mathbb{E}(Y - \langle S, X\rangle)^2 = \mathbb{E}\langle S-\rho, X\rangle^2 + \mathbb{E}\xi^2$, where we used the assumption that $\mathbb{E}(\xi|X) = 0$. Thus, the penalized true risk minimization problem becomes

$$\rho^\varepsilon := \operatorname{argmin}_{S\in\mathcal{S}} L(S), \quad L(S) := \mathbb{E}\langle S-\rho, X\rangle^2 + \varepsilon\,\operatorname{tr}(S\log S) \tag{4.1}$$

and the goal is to study the error of approximation of ρ by ρ^ε depending on the value of regularization parameter $\varepsilon > 0$. The next proposition shows that if there exists an oracle $S \in \mathcal{S}$ that provides a good approximation of the target matrix ρ in a sense that $\|S-\rho\|_{L_2(\Pi)}$ is small, then ρ^ε belongs to an $L_2(\Pi)$-ball around S of small enough radius that can be controlled in terms of the operator norm $\|\log S\|$ or in terms of more subtle characteristics of the oracle S. It also provides upper bounds on the Kullback-Leibler distance $K(\rho^\varepsilon; S)$ to the oracle and on the approximation error $\|\rho^\varepsilon - \rho\|_{L_2(\Pi)}$. We will first obtain such bounds for an arbitrary oracle $S \in \mathcal{S}$ of full rank expressed in terms of the operator norm $\|\log S\|$ of its logarithm. For simplicity, we assume that $\|\log S\| = +\infty$ in the case when $\operatorname{rank}(S) < m$ (and $\log S$ is not defined). Note, however, that $\operatorname{tr}(S\log S)$ is well defined and finite even in the case when $\operatorname{rank}(S) < m$. To obtain more subtle bounds with approximation error of the order $O(\varepsilon^2)$ instead of $O(\varepsilon)$, we introduce and use the following quantity

$$a(W) := a_\Pi(W) := a_X(W) := \sup\Big\{\langle W, U\rangle : U \in \mathbb{H}_m(\mathbb{C}), \operatorname{tr}(U) = 0, \|U\|_{L_2(\Pi)} = 1\Big\},$$

which will be called the *alignment coefficient* of W. This is a straightforward extension of similar quantities in the commutative case (Koltchinskii (2009)). Note that, for all constants c,

$$a(W + cI_m) = a(W) \tag{4.2}$$

(since $\langle I_m, U\rangle = 0$ for all U of zero trace). In addition, we have

$$a_{cX}(W) = \frac{1}{|c|} a_X(W),\ c \neq 0. \tag{4.3}$$

Let $\{E_i : i = 1, \ldots, m^2\}$ be an orthonormal basis of $\mathbb{M}_m(\mathbb{C})$ consisting of Hermitian matrices and let $\mathcal{K} := \Big(\langle E_j, E_k\rangle_{L_2(\Pi)}\Big)_{j,k=1}^{m^2}$ be the Gram matrix of the functions $\{\langle E_j, \cdot\rangle : j = 1, \ldots, m^2\}$ in the space $L_2(\Pi)$. Clearly, the mapping $J : \mathbb{M}_m(\mathbb{C}) \mapsto \ell_2^{m^2}(\mathbb{C})$,

$$JU = \Big(\langle U, E_j\rangle : j = 1, \ldots, m^2\Big),\ U \in \mathbb{M}_m(\mathbb{C}),$$

is an isometry. If now we define $\bar{\mathcal{K}} : \mathbb{M}_m(\mathbb{C}) \mapsto \mathbb{M}_m(\mathbb{C})$ as $\bar{\mathcal{K}} := J^{-1}\mathcal{K}J$, then we also have $\bar{\mathcal{K}}^{1/2} = J^{-1}\mathcal{K}^{1/2}J$, $\bar{\mathcal{K}}^{-1/2} = J^{-1}\mathcal{K}^{-1/2}J$. As a consequence, for any matrix $U = \sum_{j=1}^{m^2} u_j E_j$,

$$\|U\|_{L_2(\Pi)}^2 = \sum_{j,k=1}^{m^2} \langle E_j, E_k\rangle_{L_2(\Pi)} u_j \bar{u}_k = \langle \mathcal{K}u, u\rangle_{\ell_2} = \|\mathcal{K}^{1/2}u\|_{\ell_2}^2 = \|\bar{\mathcal{K}}^{1/2}U\|_2^2,$$

and it is not hard to conclude that $a(W) \leq \|\bar{\mathcal{K}}^{-1/2}W\|_2$. Moreover, in view of (4.2), for an arbitrary scalar c, $a(W) \leq \|\bar{\mathcal{K}}^{-1/2}(W + cI_m)\|_2$. This shows that the size of $a(W)$ depends on how W is "aligned" with the eigenspaces of the Gram matrix $\mathcal{K}$. In a special case when, for all A, $\|A\|_{L_2(\Pi)} = \|A\|_2$, the functions $\{\langle E_j, \cdot\rangle : j = 1, \ldots, m^2\}$ form an orthonormal system in $L_2(\Pi)$ and $\mathcal{K}$ is the identity matrix. In this case, we simply have the bound $a(W) \leq \inf_c \|W + cI_m\|_2$.

Proposition 3 *For all $S \in \mathcal{S}$,*

$$\|\rho^\varepsilon - \rho\|_{L_2(\Pi)}^2 + \|\rho^\varepsilon - S\|_{L_2(\Pi)}^2 + \varepsilon K(\rho^\varepsilon; S) \leq \|S - \rho\|_{L_2(\Pi)}^2 + 2\varepsilon\|\log S\|.$$

Moreover, for all $S \in \mathcal{S}$,

$$\|\rho^\varepsilon - S\|_{L_2(\Pi)}^2 + 2\varepsilon K(\rho^\varepsilon; S) \leq 2\|S - \rho\|_{L_2(\Pi)}^2 + \varepsilon^2 a^2(\log S)$$

and

$$\|\rho^\varepsilon - \rho\|_{L_2(\Pi)}^2 \leq \inf_{S \in \mathcal{S}} \left[\|S - \rho\|_{L_2(\Pi)}^2 + \frac{\varepsilon^2}{4} a^2(\log S)\right].$$

For a differentiable mapping g from an open subset $G \subset \mathbb{M}_m(\mathbb{C})$ into $\mathbb{M}_m(\mathbb{C})$, denote by $Dg(A;H)$ its differential at a matrix $A \in G$ in the direction $H \in \mathbb{M}_m(\mathbb{C})$, that is, $g(A+H) = g(A) + Dg(A;H) + o(\|H\|)$ as $\|H\| \to 0$ and $Dg(A;H)$ is linear with respect to H. The following lemma is a simple corollary of Theorem V.3.3 in Bhatia (1996):

Lemma 1 *Let f be a function continuously differentiable in an open interval $I \subset \mathbb{R}$. Suppose that A is a Hermitian matrix whose spectrum belongs to I. Then the mapping $B \mapsto g(B) := \mathrm{tr}(f(B))$ is differentiable at A and $Dg(A;H) = \mathrm{tr}(f'(A)H)$.*

Proof of Proposition 3. It is easy to see that the solution ρ^ε of problem (4.1) is a full rank matrix. To prove this, assume that $\mathrm{rank}(\rho^\varepsilon) < m$. Let $\tilde{\rho} := (1-\delta)\rho^\varepsilon + \delta I_m$, where I_m is the $m \times m$ identity matrix. Then, for small enough δ, $\tilde{\rho}$ is a full rank matrix and it is straightforward to show that the penalized risk $L(\tilde{\rho})$ is strictly smaller than $L(\rho^\varepsilon)$ (for some small $\delta > 0$). It is also easy to check that, for any $S \in \mathcal{S}$ of full rank, the differential of the functional L in the direction $\nu \in \mathbb{M}_m(\mathbb{C})$ is equal to

$$DL(S;\nu) = 2\mathbb{E}\langle S-\rho, X\rangle\langle \nu, X\rangle + \varepsilon\ \mathrm{tr}(\nu(\log S + I_m)).$$

This follows from the fact that the first term of the functional L is differentiable since it is quadratic. The differentiability of the penalty term is based on Lemma 1 (it is enough to apply this lemma to the function $f(u) = u\log u$). Since ρ^ε is the minimal point of L in $\mathcal{S}$, we can conclude that, for an arbitrary $S \in \mathcal{S}$, $DL(\rho^\varepsilon; S-\rho^\varepsilon) \geq 0$. This implies that $DL(S; S-\rho^\varepsilon) - DL(\rho^\varepsilon; S-\rho^\varepsilon) \leq DL(S; S-\rho^\varepsilon)$, which, by a simple algebra, becomes

$$2\|S-\rho^\varepsilon\|^2_{L_2(\Pi)} + \varepsilon K(S;\rho^\varepsilon) \leq 2\langle S-\rho, S-\rho^\varepsilon\rangle_{L_2(\Pi)} + \varepsilon\ \langle S-\rho^\varepsilon, \log S\rangle. \tag{4.4}$$

Taking into account that

$$2\langle S-\rho, S-\rho^\varepsilon\rangle_{L_2(\Pi)} = \|\rho^\varepsilon - S\|^2_{L_2(\Pi)} + \|S-\rho\|^2_{L_2(\Pi)} - \|\rho^\varepsilon - \rho\|_{L_2(\Pi)},$$

(4.4) can be rewritten as

$$\|\rho^\varepsilon-\rho\|^2_{L_2(\Pi)} + \|\rho^\varepsilon - S\|^2_{L_2(\Pi)} + \varepsilon K(S;\rho^\varepsilon) \leq \|S-\rho\|^2_{L_2(\Pi)} + \varepsilon\ \langle S-\rho^\varepsilon, \log S\rangle. \tag{4.5}$$

The first inequality of the proposition immediately follows from (4.5) since

$$\Big|\langle S-\rho^\varepsilon, \log S\rangle\Big| \leq \|\log S\|\|S-\rho^\varepsilon\|_1 \leq 2\|\log S\|.$$

To prove the remaining bounds, note that by the definition of alignment coefficient

$$\varepsilon\Big|\langle S-\rho^\varepsilon, \log S\rangle\Big| \leq \varepsilon a(\log S)\|\rho^\varepsilon - S\|_{L_2(\Pi)},$$

and, using an elementary bound

$$\varepsilon a(\log S)\|\rho^\varepsilon - S\|_{L_2(\Pi)} \le \frac{\varepsilon^2 a^2(\log S)}{2\alpha^2} + \frac{\alpha^2}{2}\|\rho^\varepsilon - S\|^2_{L_2(\Pi)}$$

for $\alpha = 1$ and $\alpha = \sqrt{2}$, it is easy to complete the proof. □

A consequence of Proposition 3 is that $\|\rho^\varepsilon - \rho\|^2_{L_2(\Pi)} \le \frac{\varepsilon^2}{4} a^2(\log \rho) \wedge \varepsilon \|\log \rho\|$.

We will now provide versions of approximation error bounds for special types of oracles $S \in \mathcal{S}$.

Low Rank Oracles. First we show how to adapt the bounds of Proposition 3 expressed in terms of the alignment coefficient $a(\log S)$ for a full rank matrix S (for which $\log S$ is well defined) to the case when S is an oracle of a small rank $r < m$. For a subspace L of $\mathbb{C}^m$, denote $\Lambda(L) := \sup_{\|A\|_{L_2(\Pi)} \le 1} \|P_L A P_L\|_2$. Suppose that $S \in \mathcal{S}$ is a matrix of rank r. To be specific, let $S = \sum_{j=1}^r \gamma_j (e_j \otimes e_j)$, where γ_j are positive eigenvalues of S and $\{e_1, \dots, e_m\}$ is an orthonormal basis of $\mathbb{C}^m$. Let L be the linear span of the vectors $e_1, \dots, e_r$.

Proposition 4 *There exists a numerical constant $C > 0$ such that, for all $\varepsilon > 0$,*

$$\|\rho^\varepsilon - \rho\|^2_{L_2(\Pi)} \le \Big(\|S - \rho\|_{L_2(\Pi)} + C\mathbb{E}^{1/2}\|X\|^2\varepsilon\Big)^2 + C\varepsilon^2\Lambda^2(L) r \log^2\left(1 + \frac{m}{\varepsilon \wedge 1}\right).$$

Proof. Note that, for all matrices W of rank r "supported" in the space L in the sense that $W = P_L W P_L$, we have

$$a(W) \le \sup_{\|U\|_{L_2(\Pi)} \le 1} \langle W, U\rangle = \sup_{\|U\|_{L_2(\Pi)} \le 1} \langle W, P_L U P_L\rangle \le \Lambda(L)\|W\|_2.$$

For $\delta \in (0,1)$, consider $S_\delta := (1-\delta)S + \delta \frac{I_m}{m}$. Then, using the fact that $a(W + cI_m) = a(W)$, we get

$$\log S_\delta = \sum_{j=1}^r \Big(\log((1-\delta)\gamma_j + \delta/m) - \log(\delta/m)\Big)(e_j \otimes e_j) + \log(\delta/m) I_m$$

and

$$a(\log S_\delta) = a\left(\sum_{j=1}^r \Big(\log((1-\delta)\gamma_j + \delta/m) - \log(\delta/m)\Big)(e_j \otimes e_j)\right) \le$$

$$\Lambda(L)\left\|\sum_{j=1}^r \Big(\log((1-\delta)\gamma_j + \delta/m) - \log(\delta/m)\Big)(e_j \otimes e_j)\right\|_2 \le$$

$$\Lambda(L)\left(\sum_{j=1}^{r}\log^2\left(1+\frac{m\gamma_j}{\delta}\right)\right)^{1/2}\le\Lambda(L)\sqrt{r}\log\left(1+\frac{m\|S\|}{\delta}\right).$$

Note also that $\|S-S_\delta\|^2_{L_2(\Pi)}=\delta^2\|S-I_m/m\|^2_{L_2(\Pi)}\le 4\delta^2\mathbb{E}\|X\|^2$, since

$$\|S-I_m/m\|^2_{L_2(\Pi)}\le 2(\mathbb{E}\langle S,X\rangle^2+\mathbb{E}\langle I_m/m,X\rangle^2)\le$$

$$2(\|S\|_1^2\mathbb{E}\|X\|^2+\|I_m/m\|_1^2\mathbb{E}\|X\|^2)\le 4\mathbb{E}\|X\|^2.$$

Thus, it easily follows from the last bound of Proposition 3 that

$$\|\rho^\varepsilon-\rho\|^2_{L_2(\Pi)}\le\|S_\delta-\rho\|^2_{L_2(\Pi)}+(\varepsilon^2/4)a^2(\log S_\delta)\le$$

$$\left(\|S-\rho\|_{L_2(\Pi)}+\|S_\delta-S\|_{L_2(\Pi)}\right)^2+(\varepsilon^2/4)\Lambda^2(L)r\log^2\left(1+\frac{m}{\delta}\right).$$

Taking $\delta=\varepsilon\wedge 1$ and using the bound on $\|S-S_\delta\|_{L_2(\Pi)}$, this yields the claim of the proposition. □

Note that if $\{E_i,\ i=1,\dots,m^2\}$ is an orthonormal basis of $\mathbb{M}_m(\mathbb{C})$ consisting of Hermitian matrices and X is uniformly distributed in $\{E_i,\ i=1,\dots,m^2\}$, then, for all Hermitian A, $\|A\|^2_{L_2(\Pi)}=m^{-2}\|A\|_2^2$. Therefore $\Lambda(L)\le\sup_{\|A\|_{L_2(\Pi)}\le 1}\|A\|_2=\sup_{\|A\|_2\le m}\|A\|_2=m$. Also, in this case $\|X\|\le\|X\|_2=1$. Thus, Proposition 4 yields

$$\|\rho^\varepsilon-\rho\|^2_{L_2(\Pi)}\le(\|S-\rho\|_{L_2(\Pi)}+C\varepsilon)^2+Cm^2r\varepsilon^2\log^2\left(1+\frac{m}{\varepsilon\wedge 1}\right).$$

Gibbs Oracles. Let H be a Hermitian matrix ("a Hamiltonian") and let $\beta>0$. Consider the following density matrix (a "Gibbs oracle"): $\rho_{H,\beta}:=\frac{e^{-\beta H}}{\mathrm{tr}(e^{-\beta H})}$. For simplicity, assume in what follows that $\beta=1$ (in fact, one can always replace H by βH) and denote $\rho_H:=\frac{e^{-H}}{\mathrm{tr}(e^{-H})}$. Let $\gamma_1\le\gamma_2\le\cdots\le\gamma_m$ be the eigenvalues of H and $e_1,\dots,e_m$ be the corresponding eigenvectors. Let $L_r=\mathrm{l.s.}(\{e_1,\dots,e_r\})$ and

$$H_{\le r}:=\sum_{j=1}^{r}\gamma_j(e_j\otimes e_j),\quad H_{>r}:=\sum_{j=r+1}^{m}\gamma_j(e_j\otimes e_j).$$

It is easy to see that

$$\|P_{L_r^\perp}\rho_H P_{L_r^\perp}\|_1=\frac{\sum_{k\ge r+1}e^{-\gamma_k}}{\sum_{k\ge 1}e^{-\gamma_k}}=:\delta_r(H).$$

Denote $\tilde\delta_r(H):=\max_{1\le k\le m}\mathbb{E}^{1/2}\langle Xe_k,e_k\rangle^2\delta_r(H)$. Under reasonable conditions on the spectrum of H, the quantity $\tilde\delta_r(H)$ decreases fast enough when r increases. Thus, ρ_H can be well approximated by low rank matrices.

The next statement follows immediately from Proposition 3. Here the unknown density matrix ρ is approximated by a Gibbs model with an arbitrary Hamiltonian. The error is controlled in terms of the $L_2(\Pi)$-distance between ρ and the oracle ρ_H and also in terms of the alignment coefficient $a(H_{\le r})$ for a "low rank part" $H_{\le r}$ of the Hamiltonian H and the quantity $\delta_r(H)$.

Proposition 5 *For all Hermitian nonnegatively definite matrices H and for all $\varepsilon > 0$,*

$$\|\rho^{\varepsilon} - \rho\|^2_{L_2(\Pi)} \le \Bigl(\|\rho_H - \rho\|_{L_2(\Pi)} + 2\tilde{\delta}_r(H)\Bigr)^2 + a^2(H_{\le r})\varepsilon^2.$$

Proof. We will use the last bound of proposition 3 with $S = \rho_{H_{\le r}}$. Note that

$$a(\log \rho_{H_{\le r}}) = a(-H_{\le r} - \log \mathrm{tr}(e^{-H_{\le r}})I_m) = a(H_{\le r}).$$

Therefore, we have $\|\rho^{\varepsilon} - \rho\|^2_{L_2(\Pi)} \le \|\rho_{H_{\le r}} - \rho\|^2_{L_2(\Pi)} + (\varepsilon^2/4)a^2(H_{\le r})$. In addition to this,

$$\|\rho_H - \rho_{H_{\le r}}\|_{L_2(\Pi)} = \left\| \frac{\sum_{k=1}^m e^{-\gamma_k}(e_k \otimes e_k)}{\sum_{k=1}^m e^{-\gamma_k}} - \frac{\sum_{k=1}^r e^{-\gamma_k}(e_k \otimes e_k)}{\sum_{k=1}^r e^{-\gamma_k}} \right\|_{L_2(\Pi)},$$

which can be easily bounded from above by

$$2\delta_r(H) \max_{1\le k\le m} \|e_k \otimes e_k\|_{L_2(\Pi)} = 2\delta_r(H) \max_{1\le k\le m} \mathbb{E}^{1/2}\langle Xe_k, e_k\rangle^2 = 2\tilde{\delta}_r(H).$$

The result follows immediately. □

5 Random Error Bounds and Oracle Inequalities

We now turn to the analysis of random error of the estimator $\hat{\rho}^{\varepsilon}$. We obtain upper bounds on the $L_2(\Pi)$ and Kullback-Leibler distances of this estimator to an arbitrary oracle $S \in \mathcal{S}$ of full rank, and, as a consequence, **oracle inequalities** for the empirical solution $\hat{\rho}^{\varepsilon}$. The size of both errors $\|\hat{\rho}^{\varepsilon} - S\|^2_{L_2(\Pi)}$ and $K(\hat{\rho}^{\varepsilon}; S)$ will be controlled in terms of the squared $L_2(\Pi)$-distance $\|S - \rho\|^2_{L_2(\Pi)}$ from the oracle to the target density matrix ρ and also in terms of such characteristics of the oracle as the norm $\|\log S\|$ or the alignment coefficient $a(\log S)$ that have been already used in the approximation error bounds of the previous section (see Proposition 3). However, in the case of the random error, we also need some additional quantities that describe the properties of the design distribution Π and of the noise ξ. These quantities are explicitly involved in the statements of the results below which makes them somewhat complicated. At the

same time, it is easy to control these quantities in concrete examples and to derive in special cases the bounds that are easier to understand.

Assumptions on the design distribution. In this section, it will be assumed that X is a random Hermitian $m\times m$ matrix and that, for some constants $0<U\le U_2$, $\|X\|\le U$ and $\|X\|_2\le U_2$. We will denote $\sigma_X^2:=\|\mathbb{E}X^2\|$, $\sigma^2_{X\otimes X}:=\|\mathbb{E}(X\otimes X-\mathbb{E}(X\otimes X))^2\|$. [5]

Let $L\subset\mathbb{C}^m$ be a subspace of dimension $r\le m$ and let $\mathcal{P}_L:\mathbb{M}_m(\mathbb{C})\mapsto\mathbb{M}_m(\mathbb{C})$, $\mathcal{P}_Lx:=x-P_{L^\perp}xP_{L^\perp}$. We will use the following quantity:

$$\beta(L):=\sup_{A\in\mathbb{H}_m(\mathbb{C}),\|A\|_{L_2(\Pi)}\le 1}\|\mathcal{P}_LA\|_{L_2(\Pi)}.$$

Note that $\|\mathcal{P}_LA\|_2\le\|A\|_2$ (for a proof, choose a basis $\{e_1,\dots,e_m\}$ of $\mathbb{C}^m$ such that $L=\text{l.s.}(e_1,\dots,e_r)$ and represent $A,\mathcal{P}_LA$ in this basis). If, for all A, $K_1\|A\|_2\le\|A\|_{L_2(\Pi)}\le K_2\|A\|_2$, then $\beta(L)\le K_2/K_1$. In particular, if $K_1=K_2$, then $\beta(L)=1$ (which is the case, for instance, when X is sampled at random from an orthonormal basis).

Assumptions on the noise. Let $(X_1,Y_1),\dots,(X_n,Y_n)$ be i.i.d. copies of (X,Y). Denote $\xi:=Y-\text{tr}(\rho X)$. Then $\xi_1,\dots,\xi_n$ are i.i.d. copies of ξ. Recall also that $\mathbb{E}(\xi|X)=0$ and assume that $\mathbb{P}$ a.s. $\mathbb{E}(\xi^2|X)\le\sigma_\xi^2$, where $\sigma_\xi^2\ge 0$ is a constant. We will further assume that the noise is uniformly bounded by a constant $c_\xi>0:|\xi|\le c_\xi$.

Given $t>0$, denote $t_m:=t+\log(2m)$, $\tau_n:=t+\log\log_2(2n)$ and

$$\varepsilon_{n,m}:=(\sigma_\xi\sigma_X\vee\sigma_{X\otimes X})\sqrt{\frac{t_m}{n}}\bigvee c_\xi U\frac{t_m}{n}.$$

We will start with a simple result akin to the first bound of Proposition 3.

Theorem 3 *There exists a constant $C>0$ such that, for all $S\in\mathcal{S}$ and for all $\varepsilon\ge 0$, with probability at least $1-e^{-t}$*

$$\|\hat\rho^\varepsilon-S\|^2_{L_2(\Pi)}\le\|S-\rho\|^2_{L_2(\Pi)}+C\Bigg[\varepsilon(\|\log S\|\bigwedge\log\Gamma)\bigvee\|S-\rho\|_{L_2(\Pi)}U\sqrt{\frac{t_m}{n}}\bigvee$$
$$(\sigma_\xi\sigma_X\vee\sigma_{X\otimes X})\sqrt{\frac{t_m}{n}}\bigvee(c_\xi U\vee U_2^2)\frac{t_m}{n}\Bigg] \tag{5.1}$$

[5] In this section, the notation $A\otimes B$ means the tensor product of the matrices A,B viewed as vectors of the Euclidean space $(\mathbb{M}_m(\mathbb{C}),\langle\cdot,\cdot\rangle):(A\otimes B)V=A\langle B,V\rangle, V\in\mathbb{M}_m(\mathbb{C})$.

and

$$\|\hat{\rho}^{\varepsilon} - \rho\|^2_{L_2(\Pi)} \le \|S - \rho\|^2_{L_2(\Pi)} + C\Bigg[\varepsilon(\|\log S\| \wedge \log \Gamma) \bigvee \|S - \rho\|_{L_2(\Pi)} U \sqrt{\frac{t_m}{n}} \bigvee$$
$$(\sigma_\xi \sigma_X \vee \sigma_{X\otimes X}) \sqrt{\frac{t_m}{n}} \bigvee (c_\xi U \vee U_2^2) \frac{t_m}{n}\Bigg], \tag{5.2}$$

where $\Gamma := \frac{m\mathbb{E}^{1/2}\|X\|^2}{\sqrt{\varepsilon}} \vee m$. *In particular,*

$$\|\hat{\rho}^{\varepsilon} - \rho\|^2_{L_2(\Pi)} \le C\Bigg[\varepsilon(\|\log \rho\| \wedge \log \Gamma) \bigvee (\sigma_\xi \sigma_X \vee \sigma_{X\otimes X}) \sqrt{\frac{t_m}{n}} \bigvee (c_\xi U \vee U_2^2) \frac{t_m}{n}\Bigg]. \tag{5.3}$$

Note that this result holds for all $\varepsilon \ge 0$, including the case of $\varepsilon = 0$ that corresponds to the least squares estimator over the set $\mathcal{S}$ of all density matrices. The approximation error term $\|\log S\|\varepsilon$ in the bounds of Theorem 3 is of the order $O(\varepsilon)$ (as in the first bound of Proposition 3) and the random error terms are, up to logarithmic factors, of the order $O(\frac{1}{\sqrt{n}})$ with respect to the sample size n.

The next result provides a more subtle oracle inequality in spirit of the second and third bounds of Proposition 3. In this oracle inequality, the approximation error term due to von Neumann entropy penalization is $a^2(\log S)\varepsilon^2$ (as in Proposition 3), so, it is of the order $O(\varepsilon^2)$. Note that it is assumed implicitly that $a^2(\log S) < +\infty$, i.e., that S is of full rank and the matrix $\log S$ is well defined. The random error terms are of the order $O(n^{-1})$ as $n \to \infty$ (up to logarithmic factors) with an exception of the term $\sigma_\xi \sigma_X \|P_{L^\perp} S P_{L^\perp}\|_1 \sqrt{\frac{t_m}{n}}$, which depends on how well the oracle S is approximated by low rank matrices. If $\|P_{L^\perp} S P_{L^\perp}\|_1$ is small, say of the order $n^{-1/2}$ for a subspace L of a small dimension r, this term becomes comparable to other terms in the bound, or even smaller. The inequalities hold only for the values of regularization parameter ε above certain threshold. The first bound shows that if there is an oracle $S \in \mathcal{S}$ such that: (a) it is "well aligned", that is, $a(\log S)$ is small; (b) there exists a subspace L of small dimension r such that the oracle matrix S is "almost supported" in L, that is, $\|P_{L^\perp} S P_{L^\perp}\|_1$ is small; and (c) S provides a good approximation of the density matrix ρ, that is, $\|S - \rho\|^2_{L_2(\Pi)}$ is small, then the empirical solution $\hat{\rho}^{\varepsilon}$ will be in the intersection of the $L_2(\Pi)$-ball and the Kullback-Leibler "ball" of small enough radii around the oracle S. The second bound is an oracle inequality showing how the $L_2(\Pi)$-error $\|\hat{\rho}^{\varepsilon} - \rho\|^2_{L_2(\Pi)}$ depends on the properties of the oracle S.

Theorem 4 *There exist numerical constants* $C > 0, D > 0$ *such that the following holds. For all* $t > 0$, *for all* $\varepsilon \ge D\varepsilon_{n,m}$, *for all subspaces* $L \subset \mathbb{C}^m$ *with* $\dim(L) := r$, *and for all*

$S \in \mathcal{S}$, *with probability at least* $1 - e^{-t}$,

$$\|\hat{\rho}^{\varepsilon} - S\|^2_{L_2(\Pi)} + \frac{\varepsilon}{4} K(\hat{\rho}^{\varepsilon}; S) \le 2\|S - \rho\|^2_{L_2(\Pi)} + C\Bigg[a^2(\log S)\varepsilon^2 \bigvee \tag{5.4}$$

$$\sigma_{\xi}^2 \beta^2(L) \frac{mr + \tau_n}{n} \bigvee \sigma_{\xi}\sigma_X \|P_{L^{\perp}} S P_{L^{\perp}}\|_1 \sqrt{\frac{t_m}{n}} \bigvee c_{\xi} U \frac{\tau_n \vee t_m}{n} \bigvee U_2^2 \frac{t_m}{n}\Bigg]$$

and

$$\|\hat{\rho}^{\varepsilon} - \rho\|^2_{L_2(\Pi)} \le \|S - \rho\|^2_{L_2(\Pi)} + C\Bigg[a^2(\log S)\varepsilon^2 \bigvee \|S - \rho\|_{L_2(\Pi)} U \sqrt{\frac{t_m}{n}} \bigvee$$

$$\sigma_{\xi}^2 \beta^2(L) \frac{mr + \tau_n}{n} \bigvee \sigma_{\xi}\sigma_X \|P_{L^{\perp}} S P_{L^{\perp}}\|_1 \sqrt{\frac{t_m}{n}} \bigvee c_{\xi} U \frac{\tau_n \vee t_m}{n} \bigvee U_2^2 \frac{t_m}{n}\Bigg]. \tag{5.5}$$

Remark. In the case when the noise is not necessarily bounded, but $\|\xi\|_{\psi_1} < +\infty$ (for instance, Gaussian noise), the results still hold with the following simple modifications. In bounds (5.1), (5.2), (5.3) and in the definition of $\varepsilon_{n,m}$, the term $c_{\xi} U \frac{t_m}{n}$ is to be replaced by $\|\xi\|_{\psi_1} U \log\Big(\frac{\|\xi\|_{\psi_1}}{\sigma_{\xi}} \frac{U}{\sigma_X}\Big) \frac{t_m}{n}$. In the bounds of Theorem 4, the term $c_{\xi} U \frac{\tau_n \vee t_m}{n}$ is to be replaced by $\|\xi\|_{\psi_1} U \frac{\tau_n \log n}{n} \bigvee \|\xi\|_{\psi_1} U \log\Big(\frac{\|\xi\|_{\psi_1}}{\sigma_{\xi}} \frac{U}{\sigma_X}\Big) \frac{t_m}{n}$. For such an unbounded noise, one should replace in the proofs of theorems 3 and 4 the noncommutative Bernstein inequality of Ahlswede and Winter by the bound of Proposition 2. One should also use a version of concentration inequality for empirical processes by Adamczak (2008) instead of the usual version of Talagrand for bounded function classes (see Section 3).

We will provide a detailed proof of Theorem 4. The proof of Theorem 3 is its simplified version and it will be skipped. Throughout the proofs below, $C, C_1, \dots$ are numerical constants whose values might be different in different places.

Proof of Theorem 4. Denote

$$L_n(S) := n^{-1} \sum_{j=1}^{n} (Y_j - \operatorname{tr}(S X_j))^2 + \varepsilon \operatorname{tr}(S \log S).$$

For any $S \in \mathcal{S}$ of full rank and any direction $\nu \in \mathbb{M}_m(\mathbb{C})$, we have

$$DL_n(S; \nu) = 2n^{-1} \sum_{j=1}^{n} (\langle S, X_j \rangle - Y_j) \langle \nu, X_j \rangle + \varepsilon \operatorname{tr}(\nu(\log S + I_m)).$$

By necessary conditions of extrema in the convex optimization problem (1.2), $DL_n(\hat{\rho}^{\varepsilon}; \hat{\rho}^{\varepsilon} - S) \le 0$, which implies

$$DL(\hat{\rho}^{\varepsilon}; \hat{\rho}^{\varepsilon} - S) - DL(S; \hat{\rho}^{\varepsilon} - S) \le -DL(S; \hat{\rho}^{\varepsilon} - S) + DL(\hat{\rho}^{\varepsilon}; \hat{\rho}^{\varepsilon} - S) - DL_n(\hat{\rho}^{\varepsilon}; \hat{\rho}^{\varepsilon} - S). \tag{5.6}$$

By a simple algebra similar to what has been already used in the proof of Proposition 3 (see the derivation of (4.4), (4.5)), we get from (5.6) the following bound:

$$2\|\hat{\rho}^{\varepsilon}-S\|^2_{L_2(\Pi)}+2\langle S-\rho,\hat{\rho}^{\varepsilon}-S\rangle_{L_2(\Pi)}+\varepsilon K(\hat{\rho}^{\varepsilon};S)= \tag{5.7}$$

$$\|\hat{\rho}^{\varepsilon}-S\|^2_{L_2(\Pi)}+\|\hat{\rho}^{\varepsilon}-\rho\|^2_{L_2(\Pi)}-\|S-\rho\|^2_{L_2(\Pi)}+\varepsilon K(\hat{\rho}^{\varepsilon};S)\le$$

$$\varepsilon a(\log S)\|\hat{\rho}^{\varepsilon}-S\|_{L_2(\Pi)}-\frac{2}{n}\sum_{j=1}^n\Big(\langle\hat{\rho}^{\varepsilon}-S,X_j\rangle^2-\mathbb{E}\langle\hat{\rho}^{\varepsilon}-S,X\rangle^2\Big)+$$

$$\frac{2}{n}\sum_{j=1}^n\Big(\langle S-\rho,X_j\rangle\langle\hat{\rho}^{\varepsilon}-S,X_j\rangle-\mathbb{E}\langle S-\rho,X\rangle\langle\hat{\rho}^{\varepsilon}-S,X\rangle\Big)-\frac{2}{n}\sum_{j=1}^n\xi_j\langle\hat{\rho}^{\varepsilon}-S,X_j\rangle,$$

where we also used that $\varepsilon|\mathrm{tr}((\hat{\rho}^{\varepsilon}-S)\log S)|\le\varepsilon a(\log S)\|\hat{\rho}^{\varepsilon}-S\|_{L_2(\Pi)}$.

We need to bound the empirical processes in the right hand side of bound (5.7). We will do it in three steps by bounding each term separately (which leads to different ingredients in bounds (5.4) and (5.5)). The first two steps are based on simple applications of noncommutative Bernstein's inequality (3.1); the third step relies in addition on Talagrand's concentration inequality and empirical processes bounds.

Step 1. To bound the first term note that

$$\frac{1}{n}\sum_{j=1}^n\Big(\langle\hat{\rho}^{\varepsilon}-S,X_j\rangle^2-\mathbb{E}\langle\hat{\rho}^{\varepsilon}-S,X\rangle^2\Big)=\Big\langle(\hat{\rho}^{\varepsilon}-S)\otimes(\hat{\rho}^{\varepsilon}-S),\frac{1}{n}\sum_{j=1}^n((X_j\otimes X_j)-\mathbb{E}(X\otimes X))\Big\rangle.$$

Applying (3.1) to the sum of independent random matrices $X_j\otimes X_j-\mathbb{E}(X\otimes X)$, we can claim that with probability at least $1-e^{-t}$

$$\Big|\frac{1}{n}\sum_{j=1}^n\Big(\langle\hat{\rho}^{\varepsilon}-S,X_j\rangle^2-\mathbb{E}\langle\hat{\rho}^{\varepsilon}-S,X\rangle^2\Big)\Big|\le\|\hat{\rho}^{\varepsilon}-S\|_1^2\Big\|\frac{1}{n}\sum_{j=1}^n((X_j\otimes X_j)-\mathbb{E}(X\otimes X))\Big\|\le$$

$$4\Big(\sigma_{X\otimes X}\sqrt{\frac{t+\log(2m^2)}{n}}\bigvee U_2^2\frac{t+\log(2m^2)}{n}\Big)\|\hat{\rho}^{\varepsilon}-S\|_1^2\le$$

$$4\sigma_{X\otimes X}\sqrt{\frac{t+\log(2m^2)}{n}}\|\hat{\rho}^{\varepsilon}-S\|_1^2\bigvee 16U_2^2\frac{t+\log(2m^2)}{n}.$$

We also used the fact that $\|X\otimes X\|=\|X\|_2^2\le U_2^2$, $\|X\otimes X-\mathbb{E}(X\otimes X)\|\le 2U_2^2$ as well as the bounds $\|(\hat{\rho}^{\varepsilon}-S)\otimes(\hat{\rho}^{\varepsilon}-S)\|_1=\|\hat{\rho}^{\varepsilon}-S\|_2^2\le\|\hat{\rho}^{\varepsilon}-S\|_1^2$ and $\|\hat{\rho}^{\varepsilon}-S\|_1\le 2$.

Note that the term $\sigma_{X\otimes X}\sqrt{\frac{t_m}{n}}$ in the threshold $\varepsilon_{n,m}$ originates in this step.

Step 2. The second term can be written as

$$\frac{1}{n}\sum_{j=1}^n\Big(\langle S-\rho,X_j\rangle\langle\hat{\rho}^{\varepsilon}-S,X_j\rangle-\mathbb{E}\langle S-\rho,X\rangle\langle\hat{\rho}^{\varepsilon}-S,X\rangle\Big)=$$

$$\left\langle \hat{\rho}^{\varepsilon} - S, \frac{1}{n}\sum_{j=1}^{n}\Big(\langle S-\rho, X_j\rangle X_j - \mathbb{E}\langle S-\rho, X\rangle X\Big)\right\rangle$$

and bounded as follows: with probability at least $1-e^{-t}$,

$$\left|\frac{1}{n}\sum_{j=1}^{n}\Big(\langle S-\rho, X_j\rangle\langle \hat{\rho}^{\varepsilon}-S, X_j\rangle - \mathbb{E}\langle S-\rho, X\rangle\langle \hat{\rho}^{\varepsilon}-S, X\rangle\Big)\right| \le$$

$$\|\hat{\rho}^{\varepsilon}-S\|_1\left\|\frac{1}{n}\sum_{j=1}^{n}\Big(\langle S-\rho, X_j\rangle X_j - \mathbb{E}\langle S-\rho, X\rangle X\Big)\right\| \le$$

$$2\left\|\frac{1}{n}\sum_{j=1}^{n}\Big(\langle S-\rho, X_j\rangle X_j - \mathbb{E}\langle S-\rho, X\rangle X\Big)\right\| \le$$

$$8U\|S-\rho\|_{L_2(\Pi)}\sqrt{\frac{t+\log(2m)}{n}} \bigvee 8U^2\|S-\rho\|_1\frac{t+\log(2m)}{n}.$$

Here we applied bound (3.1) to sums of independent random matrices $Y_j - \mathbb{E}Y_j$, where $Y_j = \langle S-\rho, X_j\rangle\langle \hat{\rho}^{\varepsilon}-S, X_j\rangle$ and also used simple bounds

$$\|\hat{\rho}^{\varepsilon}-S\|_1 \le 2, \ \ \|\mathbb{E}\langle S-\rho, X\rangle^2 X^2\| \le U^2\|S-\rho\|^2_{L_2(\Pi)} \ \text{ and } \ \|\langle S-\rho, X\rangle X\| \le U^2\|S-\rho\|_1.$$

The bound of this step is the origin of the terms $\|S-\rho\|_{L_2(\Pi)}\sqrt{\frac{t_m}{n}}$, $U^2\frac{t_m}{n}$ in the inequalities of the Theorem.

Step 3. We turn now to bounding the third term in the right hand side of (5.7). It is easy to decompose it as follows:

$$\frac{1}{n}\sum_{j=1}^{n}\xi_j\langle \hat{\rho}^{\varepsilon}-S, X_j\rangle = \left\langle P_{L^{\perp}}(\hat{\rho}^{\varepsilon}-S)P_{L^{\perp}}, \frac{1}{n}\sum_{j=1}^{n}\xi_j P_{L^{\perp}}X_j P_{L^{\perp}}\right\rangle +$$
$$\frac{1}{n}\sum_{j=1}^{n}\xi_j\langle \hat{\rho}^{\varepsilon}-S, \mathcal{P}_L X_j\rangle. \tag{5.8}$$

Note that

$$\left|\left\langle P_{L^{\perp}}(\hat{\rho}^{\varepsilon}-S)P_{L^{\perp}}, \frac{1}{n}\sum_{j=1}^{n}\xi_j P_{L^{\perp}}X_j P_{L^{\perp}}\right\rangle\right| \le \|P_{L^{\perp}}(\hat{\rho}^{\varepsilon}-S)P_{L^{\perp}}\|_1\left\|\frac{1}{n}\sum_{j=1}^{n}\xi_j P_{L^{\perp}}X_j P_{L^{\perp}}\right\|.$$

Applying bound (3.1) one more time, we have that with probability at least $1-e^{-t}$

$$\left|\left\langle P_{L^{\perp}}(\hat{\rho}^{\varepsilon}-S)P_{L^{\perp}}, \frac{1}{n}\sum_{j=1}^{n}\xi_j P_{L^{\perp}}X_j P_{L^{\perp}}\right\rangle\right| \le$$

$$2\|P_{L^\perp}(\hat\rho^\varepsilon - S)P_{L^\perp}\|_1\left[\sigma_\xi\sigma_X\sqrt{\frac{t+\log(2m)}{n}}\bigvee 2c_\xi U\frac{t+\log(2m)}{n}\right],$$

where we also used a simple bound $\|\mathbb{E}\xi^2(P_{L^\perp}XP_{L^\perp})^2\| \le \sigma_\xi^2\|\mathbb{E}X^2\| = \sigma_\xi^2\sigma_X^2$.

To bound the second term in the right hand side of (5.8), denote

$$\alpha_n(\delta) := \sup_{\rho_1,\rho_2\in\mathcal{S},\|\rho_1-\rho_2\|_{L_2(\Pi)}\le\delta}\left|\frac{1}{n}\sum_{j=1}^n \xi_j\langle\rho_1-\rho_2,\mathcal{P}_L X_j\rangle\right|.$$

Clearly, $\left|\frac{1}{n}\sum_{j=1}^n \xi_j\langle\hat\rho^\varepsilon - S,\mathcal{P}_L X_j\rangle\right| \le \alpha_n(\|\hat\rho^\varepsilon - S\|_{L_2(\Pi)})$. To control $\alpha_n(\delta)$, we use Talagrand's concentration inequality for empirical processes. It implies that, for all $\delta>0$, with probability at least $1-e^{-s}$,

$$\alpha_n(\delta) \le 2\left[\mathbb{E}\alpha_n(\delta) + \sigma_\xi\beta(L)\delta\sqrt{\frac{s}{n}} + 4c_\xi U\frac{s}{n}\right]. \tag{5.9}$$

Here we used the facts that $\mathbb{E}\xi^2\langle\rho_1-\rho_2,\mathcal{P}_L X\rangle^2 \le \sigma_\xi^2\beta^2(L)\|\rho_1-\rho_2\|^2_{L_2(\Pi)}$ and

$$\left|\xi\langle\rho_1-\rho_2,\mathcal{P}_L X\rangle\right| \le c_\xi\|\rho_1-\rho_2\|_1\|\mathcal{P}_L X\| \le 2c_\xi(\|X\| + \|P_{L^\perp}XP_{L^\perp}\|) \le 4c_\xi\|X\| \le 4c_\xi U.$$

We will make the bound on $\alpha_n(\delta)$ uniform in $\delta\in[Un^{-1},2U]$. To this end, we apply bound (5.9) for $\delta=\delta_j = 2^{-j+1}U$, $j=0,1,\dots$ and with $s=\tau_n := t+\log\log_2(2n)$. The union bound and the monotonicity of $\alpha_n(\delta)$ with respect to δ implies that with probability at least $1-e^{-t}$ for all $\delta\in[Un^{-1},2U]$

$$\alpha_n(\delta) \le C\left[\mathbb{E}\alpha_n(\delta) + \sigma_\xi\beta(L)\delta\sqrt{\frac{\tau_n}{n}} + c_\xi U\frac{\tau_n}{n}\right], \tag{5.10}$$

where $C>0$ is a numerical constant. Now it remains to bound the expected value $\mathbb{E}\alpha_n(\delta)$. Let $e_1,\dots,e_m$ be the orthonormal basis of $\mathbb{C}^m$ such that $L = \text{l.s.}\{e_1,\dots,e_r\}$. Denote $E_{ij}(x)$ the entries of the linear transformation $x\in\mathbb{M}_m(\mathbb{C})$ in this basis. Clearly, the function $\langle\rho_1-\rho_2,\mathcal{P}_L x\rangle$ belongs to the space $\mathcal{L} := \text{l.s.}\{E_{ij} : i\le r \text{ or } j\le r\}$ of dimension $m^2-(m-r)^2 = 2mr-r^2$. Therefore,

$$\mathbb{E}\alpha_n(\delta) \le \mathbb{E}\sup_{f\in\mathcal{L},\|f\|_{L_2(\Pi)}\le\beta(L)\delta}\left|\frac{2}{n}\sum_{j=1}^n \xi_j f(X_j)\right|.$$

Using standard bounds for empirical processes indexed by finite dimensional function classes, we get $\mathbb{E}\alpha_n(\delta) \le 2\sqrt{2}\sigma_\xi\beta(L)\delta\sqrt{\frac{mr}{n}}$. We can conclude that the following bound on $\alpha_n(\delta)$ holds with probability at least $1-e^{-t}$ for all $\delta\in[Un^{-1},2U]$:

$$\alpha_n(\delta) \le C\left[\sigma_\xi\beta(L)\delta\sqrt{\frac{mr}{n}} + \sigma_\xi\beta(L)\delta\sqrt{\frac{\tau_n}{n}} + c_\xi U\frac{\tau_n}{n}\right]. \tag{5.11}$$

Note that since $\|\hat\rho^\varepsilon - S\|_1 \le 2$ and $\|X\| \le U$, we have $\|\hat\rho^\varepsilon - S\|^2_{L_2(\Pi)} = \mathbb{E}\langle \hat\rho^\varepsilon - S, X\rangle^2 \le 4U^2$, so, $\|\hat\rho^\varepsilon - S\|_{L_2(\Pi)} \le 2U$. As a result, with probability at least $1 - e^{-t}$, we either have $\|\hat\rho^\varepsilon - S\|_{L_2(\Pi)} < Un^{-1}$, or

$$\left|\frac{1}{n}\sum_{j=1}^n \xi_j \langle \hat\rho^\varepsilon - S, \mathcal{P}_L X_j\rangle\right| \le C\left[\sigma_\xi \beta(L)\|\hat\rho^\varepsilon - S\|_{L_2(\Pi)}\sqrt{\frac{mr}{n}} + \sigma_\xi\beta(L)\|\hat\rho^\varepsilon - S\|_{L_2(\Pi)}\sqrt{\frac{\tau_n}{n}} + c_\xi U\frac{\tau_n}{n}\right].$$

In the first case, we still have

$$\left|\frac{1}{n}\sum_{j=1}^n \xi_j \langle \hat\rho^\varepsilon - S, \mathcal{P}_L X_j\rangle\right| \le C\left[\sigma_\xi \beta(L)\frac{U}{n}\sqrt{\frac{mr}{n}} + \sigma_\xi\beta(L)\frac{U}{n}\sqrt{\frac{\tau_n}{n}} + c_\xi U\frac{\tau_n}{n}\right].$$

Let us assume in what follows that $\|\hat\rho^\varepsilon - S\|_{L_2(\Pi)} \ge Un^{-1}$ since another case is even easier to handle.

The terms $\sigma_\xi^2\beta^2(L)\frac{mr+\tau_n}{n}$, $\sigma_\xi\sigma_X\|P_{L^\perp}SP_{L^\perp}\|_1\sqrt{\frac{t_m}{n}}$ in the inequalities of the Theorem have their origin in this step.

We now substitute the bounds of steps 1–3 in the right hand side of (5.7) to get the following inequality that holds with some constant $C > 0$ and with probability at least $1 - 4e^{-t}$:

$$\begin{aligned}
&\|\hat\rho^\varepsilon - S\|^2_{L_2(\Pi)} + \|\hat\rho^\varepsilon - \rho\|^2_{L_2(\Pi)} + \varepsilon K(\hat\rho^\varepsilon; S) \le \\
&\|S - \rho\|^2_{L_2(\Pi)} + \varepsilon a(\log S)\|\hat\rho^\varepsilon - S\|_{L_2(\Pi)} + \\
&16\sigma_{X\otimes X}\sqrt{\frac{t_m}{n}}\|\hat\rho^\varepsilon - S\|_1^2 + 64U_2^2\frac{t_m}{n} + 16U\|S-\rho\|_{L_2(\Pi)}\sqrt{\frac{t_m}{n}}\bigvee 16U^2\frac{t_m}{n} + \\
&4\|P_{L^\perp}(\hat\rho^\varepsilon - S)P_{L^\perp}\|_1\left[\sigma_\xi\sigma_X\sqrt{\frac{t_m}{n}}\bigvee 2c_\xi U\frac{t_m}{n}\right] + \\
&C\left[\sigma_\xi\beta(L)\|\hat\rho^\varepsilon - S\|_{L_2(\Pi)}\sqrt{\frac{mr+\tau_n}{n}}\bigvee c_\xi U\frac{\tau_n}{n}\right].
\end{aligned} \tag{5.12}$$

Under the assumption $\varepsilon \ge D\varepsilon_{n,m}$ with a sufficiently large constant $D > 0$, it is easy to get that

$$16\sigma_{X\otimes X}\sqrt{\frac{t_m}{n}}\|\hat\rho^\varepsilon - S\|_1^2 \le \frac{\varepsilon}{2}\|\hat\rho^\varepsilon - S\|_1^2 \le \frac{\varepsilon}{2}K(\hat\rho^\varepsilon; S). \tag{5.13}$$

Also, by Proposition 1,

$$\|P_{L^\perp}(\hat\rho^\varepsilon - S)P_{L^\perp}\|_1 \le \|P_{L^\perp}\hat\rho^\varepsilon P_{L^\perp}\|_1 + \|P_{L^\perp}SP_{L^\perp}\|_1 \le 3\|P_{L^\perp}SP_{L^\perp}\|_1 + 2K(\hat\rho^\varepsilon; S),$$

and, under the same assumption that $\varepsilon \geq D\varepsilon_{n,m}$ with a sufficiently large constant $D > 0$,

$$4\|P_{L^\perp}(\hat{\rho}^\varepsilon - S)P_{L^\perp}\|_1 \left[\sigma_\xi \sigma_X \sqrt{\frac{t_m}{n}} \bigvee 2c_\xi U \frac{t_m}{n}\right] \leq \tag{5.14}$$
$$C\|P_{L^\perp} S P_{L^\perp}\|_1 \left[\sigma_\xi \sigma_X \sqrt{\frac{t_m}{n}} \bigvee c_\xi U \frac{t_m}{n}\right] + \frac{\varepsilon}{4} K(\hat{\rho}^\varepsilon; S).$$

Combining bounds (5.13) and (5.14) with (5.12) yields

$$\|\hat{\rho}^\varepsilon - S\|^2_{L_2(\Pi)} + \|\hat{\rho}^\varepsilon - \rho\|^2_{L_2(\Pi)} + \frac{\varepsilon}{4} K(\hat{\rho}^\varepsilon; S) \leq \tag{5.15}$$
$$\|S - \rho\|^2_{L_2(\Pi)} + \varepsilon a(\log S)\|\hat{\rho}^\varepsilon - S\|_{L_2(\Pi)} +$$
$$C\left[\|\hat{\rho}^\varepsilon - S\|_{L_2(\Pi)} \sigma_\xi \beta(L) \sqrt{\frac{mr + \tau_n}{n}} \bigvee U\|S - \rho\|_{L_2(\Pi)} \sqrt{\frac{t_m}{n}} \bigvee \right.$$
$$\left. \|P_{L^\perp} S P_{L^\perp}\|_1 \sigma_\xi \sigma_X \sqrt{\frac{t_m}{n}} \bigvee c_\xi U \frac{\tau_n \vee t_m}{n} \bigvee U_2^2 \frac{t_m}{n}\right]$$

with some constant $C > 0$. It follows from the last inequality that

$$\|\hat{\rho}^\varepsilon - S\|^2_{L_2(\Pi)} \leq A\|\hat{\rho}^\varepsilon - S\|_{L_2(\Pi)} + B - \frac{\varepsilon}{4} K(\hat{\rho}^\varepsilon; S), \tag{5.16}$$

where $A := \frac{\varepsilon}{2} a(\log S) + C\sigma_\xi \beta(L)\sqrt{\frac{mr+\tau_n}{n}}$ and

$$B := \|S - \rho\|^2_{L_2(\Pi)} - \|\hat{\rho}^\varepsilon - \rho\|^2_{L_2(\Pi)} +$$
$$C\left[\|S - \rho\|_{L_2(\Pi)} U \sqrt{\frac{t_m}{n}} \bigvee \|P_{L^\perp} S P_{L^\perp}\|_1 \sigma_\xi \sigma_X \sqrt{\frac{t_m}{n}} \bigvee c_\xi U \frac{\tau_n \vee t_m}{n} \bigvee U_2^2 \frac{t_m}{n}\right].$$

It is easy to check that

$$\|\hat{\rho}^\varepsilon - S\|^2_{L_2(\Pi)} \leq \left(\frac{A + \sqrt{A^2 + 4(B - (\varepsilon/4)K(\hat{\rho}^\varepsilon; S))}}{2}\right)^2 \leq \left(A + \sqrt{\left(B - \frac{\varepsilon}{4} K(\hat{\rho}^\varepsilon; S)\right)_+}\right)^2.$$

If $\frac{\varepsilon}{4} K(\hat{\rho}^\varepsilon; S) \geq B$, then $\|\hat{\rho}^\varepsilon - S\|^2_{L_2(\Pi)} \leq A^2$, which, in view of (5.16), implies

$$\|\hat{\rho}^\varepsilon - S\|^2_{L_2(\Pi)} + \frac{\varepsilon}{4} K(\hat{\rho}^\varepsilon; S) \leq A^2 + B.$$

Otherwise, we have $\|\hat{\rho}^\varepsilon - S\|^2_{L_2(\Pi)} \leq A^2 + 2A\sqrt{B} + B - \frac{\varepsilon}{4} K(\hat{\rho}^\varepsilon; S)$, which implies that

$$\|\hat{\rho}^\varepsilon - S\|^2_{L_2(\Pi)} + \frac{\varepsilon}{4} K(\hat{\rho}^\varepsilon; S) \leq 3A^2 + \frac{3}{2} B.$$

Thus, the last bound holds in both cases, and, by the definitions of A and B and elementary algebra, one can easily get that

$$\|\hat\rho^\varepsilon - S\|^2_{L_2(\Pi)} + \frac{3}{2}\|\hat\rho^\varepsilon - \rho\|^2_{L_2(\Pi)} + \frac{\varepsilon}{4}K(\hat\rho^\varepsilon; S) \le$$
$$\frac{3}{2}\|S-\rho\|^2_{L_2(\Pi)} + C\biggl[a^2(\log S)\varepsilon^2 \bigvee \|S-\rho\|_{L_2(\Pi)}U\sqrt{\frac{t_m}{n}} \bigvee \sigma_\xi^2\beta^2(L)\frac{mr+\tau_n}{n} \bigvee$$
$$\sigma_\xi\sigma_X\|P_{L^\perp}SP_{L^\perp}\|_1\sqrt{\frac{t_m}{n}} \bigvee c_\xi U\frac{\tau_n \vee t_m}{n} \bigvee U_2^2\frac{t_m}{n}\biggr], \tag{5.17}$$

which holds with probability at least $1-4e^{-t}$ and with a sufficiently large constant C. To replace the probability $1-4e^{-t}$ by $1-e^{-t}$, it is enough to replace t by $t+\log 4$ and to adjust the values of constants C, D accordingly. Then, (5.17) easily imply the bounds of the theorem. □

Remark. Note that replacing in Step 1 of the proof rather simple bounds based on Ahlswede-Winter inequality by a more sophisticated argument based on Talagrand's generic chaining, one can obtain another version of the bounds of Theorem 4 that might be stronger in certain applications. For instance, one can use Theorem 3 in [4] (that relies on the results of [11]) to obtain the following version of (5.5) that holds for $\varepsilon \ge D\varepsilon_{n,m}$ with $\varepsilon_{n,m} = \sigma_\xi\sigma_X\sqrt{\frac{t_m}{n}} \vee c_\xi U\frac{t_m}{n}$:

$$\|\hat\rho^\varepsilon - \rho\|^2_{L_2(\Pi)} \le \|S-\rho\|^2_{L_2(\Pi)} + C\biggl[a^2(\log S)\varepsilon^2 \bigvee \|S-\rho\|_{L_2(\Pi)}U\sqrt{\frac{t_m}{n}} \bigvee$$
$$\sigma_\xi^2\beta^2(L)\frac{mr+\tau_n}{n} \bigvee \sigma_\xi\sigma_X\|P_{L^\perp}SP_{L^\perp}\|_1\sqrt{\frac{t_m}{n}} \bigvee c_\xi U\frac{\tau_n \vee t_m}{n} \bigvee U^2\frac{t+\log^5 m\log n}{n}\biggr].$$

This could be better than (5.5) since there is no term $\sigma_{X\otimes X}\sqrt{\frac{t_m}{n}}$ in the new definition of $\varepsilon_{n,m}$ and also because $U^2\frac{t+\log^5 m\log n}{n}$ could be smaller than $U_2^2\frac{t_m}{n}$ when U is much smaller than U_2 (for instance, in the case of sampling from the Pauli basis, $U_2 = 1$ and $U = m^{-1/2}$).

Example. Sampling from an orthonormal basis. Recall that in this case Π is the distribution in an orthonormal basis $E_1, \dots, E_{m^2}$ that consists of Hermitian matrices. Since $\|X\|_2 = 1$, one can always assume that $U_2 = 1$ and $U \le 1$. Denote $\pi_j := \Pi(\{E_j\})$ and $\bar\pi_m := \max_{1\le j\le m^2}\pi_j$. Then, it is easy to check that $\sigma_X^2 \le m\bar\pi_m, \quad \sigma^2_{X\otimes X} \le \bar\pi_m$.

Indeed, for an orthonormal basis $e_1, \dots, e_m$ of $\mathbb{C}^m$,

$$\sigma_X^2 = \|\mathbb{E}X^2\| = \sup_{v\in\mathbb{C}^m, |v|=1} \mathbb{E}\langle X^2 v, v\rangle = \sup_{v\in\mathbb{C}^m, |v|=1} \mathbb{E}\langle Xv, Xv\rangle = \sup_{v\in\mathbb{C}^m, |v|=1} \mathbb{E}|Xv|^2 =$$

$$\sup_{v\in\mathbb{C}^m, |v|=1} \sum_{j=1}^m \mathbb{E}|\langle Xv, e_j\rangle|^2 = \sup_{v\in\mathbb{C}^m, |v|=1} \sum_{j=1}^m \sum_{k=1}^{m^2} \pi_k |\langle E_k, v\otimes e_j\rangle|^2 \le$$

$$\bar{\pi}_m \sup_{v\in\mathbb{C}^m, |v|=1} \sum_{j=1}^m \|v\otimes e_j\|_2^2 \le m\bar{\pi}_m,$$

where we used Bessel's inequality for the basis $\{E_1, \dots, E_{m^2}\}$. Similarly,

$$\sigma_{X\otimes X}^2 \le \|\mathbb{E}(X\otimes X)^2\| = \sup_{\|V\|_2=1} \mathbb{E}\|(X\otimes X)V\|_2^2 = \sup_{\|V\|_2=1} \mathbb{E}|\langle X, V\rangle|_2^2 \|X\|_2^2 \le$$

$$\sup_{\|V\|_2=1} \sum_{k=1}^{m^2} \pi_k |\langle E_k, V\rangle|_2^2 \le \bar{\pi}_m \sup_{\|V\|_2=1} \|V\|_2^2 = \bar{\pi}_m,$$

where we used the fact that $\|X\|_2 = 1$ and, again, Bessel's inequality. Note also that $\|A\|_{L_2(\Pi)}^2 \le \bar{\pi}_m \|A\|_2^2,\ A \in \mathbb{M}_m(\mathbb{C})$.

In the case of **a nearly uniform design** already defined in Section 2, $\sigma_X^2 \le c_1 m^{-1}$, $\sigma_{X\otimes X}^2 \le c_1 m^{-2}$, and $\|A\|_{L_2(\Pi)}^2 \le c_1 m^{-2}\|A\|_2^2$. We also have that $\|A\|_{L_2(\Pi)}^2 \ge c_2 m^{-2}\|A\|_2^2, A \in \mathbb{H}_m(\mathbb{C})$, which implies that the quantity $\beta(L)$ involved in Theorem 4 is bounded by $\sqrt{\frac{c_1}{c_2}}$.

We can derive the following corollary of Theorem 4. To simplify its statement, we will assume that, for some $\lambda > 0$,

$$\log\log_2(2n) \le \log(2m), \quad \sigma_\xi \ge m^{-1/2}, \quad c_\xi U \le \lambda\left(\sigma_\xi \sqrt{\frac{n}{mt_m}} \bigwedge \sigma_\xi^2 m \log^2(mn)\right). \quad (5.18)$$

Essentially, it means that the variance σ_ξ^2 of the noise is not too small [6] and the constant c_ξ is not too large comparing with the variance, which makes it possible to suppress the exponential tails in Bernstein type inequalities. In this case, we can take $\varepsilon_{n,m} := \sigma_\xi \sqrt{\frac{t_m}{mn}}$ and let $\varepsilon = D\varepsilon_{n,m}$ for a sufficiently large constant $D > 0$.

Corollary 1 *Suppose that* Π *is a nearly uniform distribution in a basis* $\{E_1, \dots, E_{m^2}\}$ *that consists of Hermitian matrices. There exists a numerical constant* $C > 0$ *such that*

[6] Using the remark after the proof of Theorem 4, one can drop the condition that $\sigma_\xi \ge m^{-1/2}$; however, some additional terms will be needed in the bound of Corollary 1.

the following holds. For all $t > 0$, for all sufficiently large D and for $\varepsilon = D\varepsilon_{n,m}$, with probability at least $1 - e^{-t}$,

$$\|\hat{\rho}^{\varepsilon} - \rho\|^2_{L_2(\Pi)} \le \inf_{S \in \mathcal{S}} \left[2\|S - \rho\|^2_{L_2(\Pi)} + CD^2\sigma_\xi^2 \frac{\mathrm{rank}(S) m t_m \log^2(mn)}{n} \right]. \quad (5.19)$$

Proof (sketch). We will use the second bound of Theorem 4. Note that in the case under consideration $\Lambda(L) \le \frac{m}{\sqrt{c_2}}$.[7] Suppose now that $S \in \mathcal{S}$ is an arbitrary oracle of rank r. Then there exists a subspace L of dimension r such that $P_{L^\perp} S P_{L^\perp} = 0$. We will use bound (5.5) for $S_\delta := (1-\delta)S + \delta \frac{I_m}{m}$, where $\delta = \varepsilon \wedge 1$, as we did in the proof of Proposition 4. As in this proof, we have, for some constant $C_1 > 0$,

$$a(\log S_\delta) \le m\sqrt{r} \log\left(1 + \frac{m}{\delta}\right) \le C_1 m\sqrt{r}\log(mn)$$

and $\|S - S_\delta\|^2_{L_2(\Pi)} \le 4\delta^2 \mathbb{E}\|X\|^2 \le 4\delta^2 \le 4\varepsilon^2$. Finally, note that

$$\|P_{L^\perp} S_\delta P_{L^\perp}\|_1 \le (1-\delta)\|P_{L^\perp} S P_{L^\perp}\|_1 + \delta \|P_{L^\perp}(I_m/m)P_{L^\perp}\|_1 \le \delta \le \varepsilon.$$

Substituting these inequalities in (5.5) (with S replaced by S_δ), taking into account the bounds on σ_X, $\sigma_{X\otimes X}$ and $\beta(L)$ that hold in the case of nearly uniform design and bounding $\|S_\delta - \rho\|^2_{L_2(\Pi)}$ in terms of $\|S - \rho\|^2_{L_2(\Pi)}$ and $\|S_\delta - S\|^2_{L_2(\Pi)}$ (similarly to what was done in the proof of Proposition 4), it is easy to derive (5.19) from (5.5). □

Similarly, it is easy to obtain another corollary where the $L_2(\Pi)$-error of estimator $\hat{\rho}^{\varepsilon}$ is controlled in terms of Gibbs oracles. Recall the notations at the end of Section 4 and also denote $\Gamma_r := \|H_{\le r}\|_2^2 = \sum_{k=1}^r \gamma_k^2$. and assume that $\Gamma_1 \ge 1$ and also that (5.18) holds.

Corollary 2 *There exists a numerical constant $C > 0$ such that the following holds. For all $t > 0$, for all sufficiently large D and for $\varepsilon = D\varepsilon_{n,m}$, for all Hermitian matrices H and for all $r \le m$, with probability at least $1 - e^{-t}$,*

$$\|\hat{\rho}^{\varepsilon} - \rho\|^2_{L_2(\Pi)} \le 2\|\rho_H - \rho\|^2_{L_2(\Pi)} + C\left[\frac{\sigma_\xi^2 (D^2 t_m \Gamma_r + r) m}{n} \bigvee m^{-2}\delta_r^2(H) \right]. \quad (5.20)$$

Remarks. Note that both **the matrix completion design** of Example 1 in the Introduction and **sampling from the Pauli basis** (Example 2) are special cases of nearly uniform design. In the case of matrix completion $c_1 = 2$, $c_2 = 1$ and $U = 1$. In the

[7] recall the definition of $\Lambda(L)$ given before Proposition 4

case of sampling from the Pauli basis, $c_1 = c_2 = 1$ and it is easy to see that $U = m^{-1/2}$. Thus, in these two examples the statements of corollaries 1 and 2 hold under assumption (5.18) with proper values of U.

Note also that the bounds of theorems 3, 4 and corollaries 1,2 can be proved in the case when the noise is unbounded, in particular, Gaussian (see the remark after Theorem 4). This immediately leads to Theorem 1 stated in the Introduction. To this end, it is enough to modify slightly conditions (5.18) by replacing c_ξ by another quantity defined in terms of $\|\xi\|_{\psi_1}$, which, in the case of Gaussian noise, is of the same order as σ_ξ (again, see the remark after Theorem 4). Then, the bound of Corollary 1 becomes the second bound of Theorem 1; the first bound follows from Theorem 3.

6 Oracle Inequalities: Subgaussian Design Case

In this section, we turn to the case of *subgaussian design matrices.* More precisely, we assume that X is a Hermitian random matrix with distribution Π such that, for some constant $b_0 > 0$ and for all Hermitian matrices $A \in \mathbb{M}_m(\mathbb{C})$, $\langle A, X\rangle$ is a subgaussian random variable with parameter $b_0\|A\|_{L_2(\Pi)}$. This implies that $\mathbb{E}X = 0$ and, for some constant $b_1 > 0$,

$$\Big\|\langle A, X\rangle\Big\|_{\psi_2} \le b_1\|A\|_{L_2(\Pi)},\ A \in \mathbb{M}_m(\mathbb{C}). \tag{6.1}$$

In addition to this, assume that, for some constant $b_2 > 0$,

$$\|A\|_{L_2(\Pi)} = \Big\|\langle A, X\rangle\Big\|_{L_2(\Pi)} \le b_2\|A\|_2,\ A \in \mathbb{M}_m(\mathbb{C}). \tag{6.2}$$

A Hermitian random matrix X satisfying the above conditions will be called a *subgaussian* matrix. Moreover, if X also satisfies the condition

$$\|A\|^2_{L_2(\Pi)} = \mathbb{E}|\langle A, X\rangle|^2 = \|A\|_2^2,\ A \in \mathbb{M}_m(\mathbb{C}), \tag{6.3}$$

then it will be called an *isotropic subgaussian* matrix. As it was already mentioned in the introduction, the last class of matrices includes such examples as Gaussian and Rademacher design matrices. It easily follows from the basic properties of Orlicz norms (see, e.g., van der Vaart and Wellner (1996), p. 95) that for subgaussian matrices $\|A\|_{L_p(\Pi)} = \mathbb{E}^{1/p}\Big|\langle A, X\rangle\Big|^p \le c_p b_1 b_2\|A\|_2^2$ and $\|A\|_{\psi_1} := \Big\|\langle A, X\rangle\Big\|_{\psi_1} \le cb_1b_2\|A\|_2, A \in \mathbb{M}_m(\mathbb{C}), p \ge 1$, with some numerical constants $c_p > 0$ and $c > 0$. The following fact is well known (see, e.g., Rudelson and Vershynin (2010), Proposition 2.4).

Proposition 6 *Let X be a subgaussian $m \times m$ matrix. Then, there exists a constant $B > 0$ such that $\Big\| \|X\| \Big\|_{\psi_2} \leq B\sqrt{m}$.*

Below, we give oracle inequalities and random error bounds in the subgaussian design case. We will use the following notations. Given $t > 0$, let

$$t_m := t + \log(2m), \quad \tau_n := t + \log\log_2(2n), \quad \text{and} \quad t_{n,m} := \tau_n \log n \vee t_m.$$

In what follows, the noise satisfies the assumptions of the previous section except the boundedness assumption. Instead, it is supposed that $\|\xi\|_{\psi_2} < +\infty$. Denote $c_\xi := \|\xi\|_{\psi_2} \log \frac{\|\xi\|_{\psi_2}}{\sigma_\xi}$ and let

$$\varepsilon_{n,m} := \sigma_\xi \sqrt{\frac{mt_m}{n}} \bigvee c_\xi \frac{\sqrt{m} t_m}{n}.$$

Theorem 5 *There exist constants $C > 0, c > 0$ such that the following holds. For all $t > 0$ such that $\tau_n \leq cn$, for all $S \in \mathcal{S}$ and for all $\varepsilon \in [0,1]$, with probability at least $1 - e^{-t}$*

$$\|\hat{\rho}^\varepsilon - S\|_{L_2(\Pi)}^2 \leq 2\|S - \rho\|_{L_2(\Pi)}^2 + C\Bigg[\varepsilon\bigg(\|\log S\| \bigwedge \log\frac{m}{\varepsilon}\bigg) \bigvee \sigma_\xi \sqrt{\frac{mt_m}{n}} \bigvee \frac{mt_m}{n} \bigvee (c_\xi \vee \sqrt{m}) \frac{\sqrt{m} t_{n,m}}{n}\Bigg] \tag{6.4}$$

and

$$\|\hat{\rho}^\varepsilon - \rho\|_{L_2(\Pi)}^2 \leq \|S - \rho\|_{L_2(\Pi)}^2 + C\Bigg[\varepsilon\bigg(\|\log S\| \wedge \log\frac{m}{\varepsilon}\bigg) \bigvee \|S - \rho\|_{L_2(\Pi)} \sqrt{\frac{mt_m}{n}} \bigvee \sigma_\xi \sqrt{\frac{mt_m}{n}} \bigvee \frac{mt_m}{n} \bigvee (c_\xi \vee \sqrt{m}) \frac{\sqrt{m} t_{n,m}}{n}\Bigg]. \tag{6.5}$$

In particular,

$$\|\hat{\rho}^\varepsilon - \rho\|_{L_2(\Pi)}^2 \leq C\Bigg[\varepsilon\bigg(\|\log \rho\| \wedge \log\frac{m}{\varepsilon}\bigg) \bigvee \sigma_\xi \sqrt{\frac{mt_m}{n}} \bigvee (c_\xi \vee \sqrt{m}) \frac{\sqrt{m} t_{n,m}}{n}\Bigg].$$

We now turn to more subtle oracle inequalities that take into account low rank properties of oracles $S \in \mathcal{S}$.

Theorem 6 *There exist numerical constants $C > 0, D > 0, c > 0$ such that the following holds. For all $t > 0$ such that $\tau_n \leq cn$, for all $\varepsilon \geq D\varepsilon_{n,m}$, for all subspaces $L \subset \mathbb{C}^m$ with*

$\dim(L) := r$ *and for all* $S \in \mathcal{S}$, *with probability at least* $1 - e^{-t}$,

$$\|\hat{\rho}^{\varepsilon} - S\|^2_{L_2(\Pi)} + \frac{\varepsilon}{4} K(\hat{\rho}^{\varepsilon}; S) \le 2\|S - \rho\|^2_{L_2(\Pi)} + \tag{6.6}$$

$$C\left[a^2(\log S)\varepsilon^2 \bigvee \sigma_{\xi}^2 \beta^2(L) \frac{mr + \tau_n}{n} \bigvee \sigma_{\xi} \|P_{L^{\perp}} S P_{L^{\perp}}\|_1 \sqrt{\frac{mt_m}{n}} \bigvee (c_{\xi} \vee \sqrt{m}) \frac{\sqrt{m} t_{n,m}}{n}\right]$$

and

$$\|\hat{\rho}^{\varepsilon} - \rho\|^2_{L_2(\Pi)} \le \|S - \rho\|^2_{L_2(\Pi)} + C\left[a^2(\log S)\varepsilon^2 \bigvee \|S - \rho\|_{L_2(\Pi)} \sqrt{\frac{mt_m}{n}} \bigvee \right.$$

$$\left. \sigma_{\xi}^2 \frac{mr + \tau_n}{n} \bigvee \sigma_{\xi} \|P_{L^{\perp}} S P_{L^{\perp}}\|_1 \sqrt{\frac{mt_m}{n}} \bigvee (c_{\xi} \vee \sqrt{m}) \frac{\sqrt{m} t_{n,m}}{n}\right]. \tag{6.7}$$

Proof of Theorem 6. It follows the lines of the proof of Theorem 4 very closely with only minor modifications in steps 2,3 and with more substantial changes in Step 1, where one has to control $\frac{1}{n}\sum_{j=1}^n (\langle \hat{\rho}^{\varepsilon} - S, X_j\rangle^2 - \mathbb{E}\langle \hat{\rho}^{\varepsilon} - S, X\rangle^2)$. To this end, we will study the empirical process

$$\Delta_n(\delta) := \sup_{f \in \mathcal{F}_{\delta}} \left| n^{-1} \sum_{j=1}^n (f^2(X_j) - Pf^2) \right|,$$

where $\mathcal{F}_{\delta} := \{\langle S_1 - S_2, \cdot\rangle : S_1, S_2 \in \mathcal{S}, \|S_1 - S_2\|_{L_2(\Pi)} \le \delta\}$. Clearly,

$$\left| \frac{1}{n} \sum_{j=1}^n \left(\langle \hat{\rho}^{\varepsilon} - S, X_j\rangle^2 - \mathbb{E}\langle \hat{\rho}^{\varepsilon} - S, X\rangle^2 \right) \right| \le \Delta_n(\|\hat{\rho}^{\varepsilon} - S\|_{L_2(\Pi)}).$$

Our goal is to obtain an upper bound on $\Delta_n(\delta)$ uniformly in $\delta \in [(m/n)^{1/2}, 2b_2]$. First we use a version of Talagrand's concentration inequality for empirical processes indexed by unbounded functions due to Adamczak (see Section 3). It implies that with some constant $C > 0$ and with probability at least $1 - e^{-t}$

$$\Delta_n(\delta) \le 2\mathbb{E}\Delta_n(\delta) + C\delta^2 \sqrt{\frac{t}{n}} + C\frac{mt \log n}{n}. \tag{6.8}$$

Here we used the following bounds on the uniform variance and on the envelope of the function class $\mathcal{F}_{\delta}^2$: for the uniform variance, with some constant $c > 0$,

$$\sup_{f \in \mathcal{F}_{\delta}} (Pf^4)^{1/2} = \sup_{S_1, S_2 \in \mathcal{S}, \|S_1 - S_2\|_{L_2(\Pi)} \le \delta} \|S_1 - S_2\|^2_{L_4(\Pi)} \le c\delta^2,$$

by the equivalence properties of the norms in Orlicz spaces. For the envelope,

$$\sup_{f \in \mathcal{F}_{\delta}} f^2(X) = \sup_{S_1, S_2 \in \mathcal{S}, \|S_1 - S_2\|_{L_2(\Pi)} \le \delta} \langle S_1 - S_2, X\rangle^2 \le 4\|X\|^2$$

and

$$\Big\|\max_{1\le i\le n}\sup_{f\in\mathcal{F}_\delta} f^2(X_i)\Big\|_{\psi_1} \le c_1\Big\|\|X\|^2\Big\|_{\psi_1}\log n \le c_2\Big\|\|X\|\Big\|^2_{\psi_2}\log n \le c_3 m\log n,$$

for some constants $c_1, c_2, c_3 > 0$, where we used well known inequalities for maxima of random variables in Orlicz spaces (see, e.g., Lemma 2.2.2 in van der Vaart and Wellner (1996)).

To bound the expectation $\mathbb{E}\Delta_n(\delta)$ we use a recent result by Mendelson (2010) (see Section 3).[8] It gives

$$\mathbb{E}\Delta_n(\delta) \le c\left[\sup_{f\in\mathcal{F}_\delta}\|f\|_{\psi_1}\frac{\gamma_2(\mathcal{F}_\delta;\psi_2)}{\sqrt{n}}\bigvee\frac{\gamma_2^2(\mathcal{F}_\delta;\psi_2)}{n}\right] \tag{6.9}$$

with some constant $c > 0$. It follows from (6.1) that the ψ_1 and ψ_2-norms of functions from the class $\mathcal{F}_\delta$ can be bounded from above by a constant times the $L_2(P)$-norm. As a result,

$$\sup_{f\in\mathcal{F}_\delta}\|f\|_{\psi_1} \le c\delta \tag{6.10}$$

and the following bound holds for Talagrand's generic chaining complexities:

$$\gamma_2(\mathcal{F}_\delta;\psi_2) \le \gamma_2(\mathcal{F}_\delta; c\|\cdot\|_{L_2(\Pi)}), \tag{6.11}$$

where c is a constant. Let G be a symmetric real valued random matrix with independent centered Gaussian entries $\{g_{ij}\}$ on the diagonal and above, where $\mathbb{E}g_{ii}^2 = 1$ and $\mathbb{E}g_{ij}^2 = \frac{1}{2}, i \ne j$. Under condition (6.2), $\mathbb{E}|\langle S_1, G\rangle - \langle S_2, G\rangle|^2 = \|S_1 - S_2\|_2^2 \ge c_1\|S_1 - S_2\|_{L_2(\Pi)}^2$ for some constant c_1, and it easily follows from Talagrand's generic chaining bound that, for some constant $C > 0$,

$$\gamma_2(\mathcal{F}_\delta; c\|\cdot\|_{L_2(\Pi)}) \le C\mathbb{E}\sup_{S_1,S_2\in\mathcal{S},\|S_1-S_2\|_{L_2(\Pi)}\le\delta}|\langle S_1 - S_2, G\rangle| =: C\omega(G;\delta). \tag{6.12}$$

It follows from (6.9), (6.10), (6.11) and (6.12) that

$$\mathbb{E}\Delta_n(\delta) \le C\left[\delta\frac{\omega(G;\delta)}{\sqrt{n}}\bigvee\frac{\omega^2(G;\delta)}{n}\right]. \tag{6.13}$$

By Proposition 6, we get

$$\omega(G;\delta) = \mathbb{E}\sup_{S_1,S_2\in\mathcal{S},\|S_1-S_2\|_{L_2(\Pi)}\le\delta}\Big|\langle S_1-S_2, G\rangle\Big| \le \mathbb{E}\|G\|\sup_{S_1,S_2\in\mathcal{S}}\|S_1-S_2\|_1 \le 2\mathbb{E}\|G\| \le c\sqrt{m}.$$

[8] In fact, even earlier result by Klartag and Mendelson (2005) with the ψ_2-diameter instead of ψ_1-diameter would suffice for our purposes.

Substituting this bound in (6.13) yields that, for some constant $C > 0$,

$$\mathbb{E}\Delta_n(\delta) \le C\left[\delta\sqrt{\frac{m}{n}} \bigvee \frac{m}{n}\right] \tag{6.14}$$

and combining (6.14) with (6.8) gives that with probability at least $1 - e^{-t}$

$$\Delta_n(\delta) \le C\left[\delta\sqrt{\frac{m}{n}} \bigvee \frac{m}{n} \bigvee \delta^2\sqrt{\frac{t}{n}} \bigvee \frac{mt\log n}{n}\right]. \tag{6.15}$$

It is easy to make bound (6.15) uniform in $\delta \in [(m/n)^{1/2}, 2b_2]$ by a simple discretization argument (as we did in Step 3 of the proof of Theorem 4). This leads to the following result: with probability at least $1 - e^{-t}$, for all $\delta \in [(m/n)^{1/2}, 2b_2]$,

$$\Delta_n(\delta) \le C\left[\delta\sqrt{\frac{m}{n}} \bigvee \frac{m}{n} \bigvee \delta^2\sqrt{\frac{\tau_n}{n}} \bigvee \frac{m\tau_n\log n}{n}\right], \tag{6.16}$$

where $\tau_n = t + \log\log_2(2n)$. Thus, with the same probability and with a proper choice of constant $C > 0$

$$\left|\frac{1}{n}\sum_{j=1}^n \Big(\langle\hat\rho^\varepsilon - S, X_j\rangle^2 - \mathbb{E}\langle\hat\rho^\varepsilon - S, X\rangle^2\Big)\right| \le$$

$$C\left[\|\hat\rho^\varepsilon - S\|_{L_2(\Pi)}\sqrt{\frac{m}{n}} \bigvee \frac{m}{n} \bigvee \|\hat\rho^\varepsilon - S\|^2_{L_2(\Pi)}\sqrt{\frac{\tau_n}{n}} \bigvee \frac{m\tau_n\log n}{n}\right]$$

provided that $\|\hat\rho^\varepsilon - S\|_{L_2(\Pi)} \in [(m/n)^{1/2}, 2b_2]$.

The rest is a straightforward modification of the proof of Theorem 4. □

For simplicity, we state the next corollary (similar to corollary 1) only in the case of *subgaussian isotropic design*. Recall that in this case $\|\cdot\|_{L_2(\Pi)} = \|\cdot\|_2$ and $\beta(L) = 1$.

Corollary 3 *There exist numerical constants $C > 0, c > 0$ such that the following holds. For all $t > 0$ such that $\tau_n \le cn$, for all sufficiently large $D > 0$ and for $\varepsilon = D\varepsilon_{n,m}$, for all matrices $S \in \mathcal{S}$ of rank r, with probability at least $1 - e^{-t}$,*

$$\|\hat\rho^\varepsilon - \rho\|^2_{L_2(\Pi)} \le 2\|S - \rho\|^2_{L_2(\Pi)} + C\left[D^2\left(\sigma_\xi^2\frac{rmt_m}{n} \bigvee c_\xi^2\frac{rmt_m^2}{n^2}\right)\log^2(mn) \bigvee\right.$$

$$\left.\sigma_\xi^2\frac{\tau_n}{n} \bigvee (c_\xi \vee \sqrt{m})\frac{\sqrt{m}t_{n,m}}{n}\right]. \tag{6.17}$$

In a special case of Gaussian noise, the bounds of the above corollary can be simplified since in this case $c_\xi \le c\sigma_\xi$ for some numerical constant c. In particular, Theorem 5 and Corollary 3 immediately imply the bounds of Theorem 2 in the Introduction (to this end, one just has to drop the terms in the bounds of Theorem 5 and Corollary 3 that are dominated by the main terms under the assumption that the noise is Gaussian and other assumptions of Theorem 2).

References

[1] Adamczak, R. (2008) A tail inequality for suprema of unbounded empirical processes with applications to Markov chains. *Electronic Journal of Probability,* 13, 34, 1000–1034.

[2] Artiles, L.M., Gill, R. and Guta, M.I.(2004) An invitation to quantum tomography. *J. Royal Statistical Society,* Ser. B, v. 67, 1, 109–134.

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

[4] Aubrun, G. (2009) On almost randomizing channels with a short Kraus decomposition. *Commun. Math. Physics*, 288, 1103–1116.

[5] Bhatia, R. (1997) Matrix Analysis. Springer, New York.

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

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

[8] Candes, E. and Plan, Y. (2009) Tight Oracle Bounds for Low-Rank Matrix Recovery from a Minimal Number of Random Measurements. *IEEE Transactions on Information Theory*, to appear.

[9] 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.

[10] Gross, D. (2011) Recovering Low-Rank Matrices From Few Coefficients in Any Basis. *IEEE Transactions on Information Theory*, 57, 3, 1548–1566.

[11] Guédon, O., Mendelson, S., Pajor, A. and Tomczak-Jaegermann, N. (2008) Majorizing measures and proportional subsets of bounded orthonormal systems. *Rev. Mat. Iberoamericana*, 24, 3, 1075–1095.

[12] Klartag, B. and Mendelson, S. (2005) Empirical Processes and Random Projections. *Journal of Functional Analysis,* 225(1), 229–245.

[13] Klauck, H., Nayak, A., Ta-Shma, A. and Zuckerman, D. (2007) Interactions in Quantum Communication. *IEEE Transactions on Information Theory,* 53, 6, 1970–1982.

[14] Koltchinskii, V. (2009) Sparse recovery in convex hulls via entropy penalization. *Annals of Statistics,* 37(3), 1332–1359.

[15] 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.

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

[17] Ledoux, M. and Talagrand, M. (1991) Probability in Banach Spaces. Springer.

[18] Mendelson, S. (2010) Empirical processes with a bounded ψ_1 diameter. *Geometric and Functional Analysis*, 20, 4, 988–1027.

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

[20] Nielsen, M.A. and Chuang, I.L. (2000) Quantum Computation and Quantum Information, Cambridge University Press.

[21] Oliveira, R.I. (2010) Sums of random Hermitian matrices and an inequality by Rudelson. *Electronic Communications in Probability*, 15, 203-212.

[22] Recht, B. (2009) A Simpler Approach to Matrix Completion. *Journal of Machine Learning Research*. to appear.

[23] Rudelson, M. and Vershynin, R. (2010) Non-asymptotic theory of random matrices: extreme singular values. *Proceedings of the International Congress of Mathematicians,* Hyderabad, India.

[24] Rohde, A. and Tsybakov, A. (2011) Estimation of high-dimensional low rank matrices. *Annals of Statistics*, 39, 2, 887–930.

[25] Simon, B. (1979) Trace Ideals and their Applications. Cambridge University Press.

[26] Talagrand, M. (2005) The Generic Chaining. Springer.

[27] Tropp, J.A. (2010) User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, to appear.

[28] van der Vaart, A. and Wellner, J. (1996) Weak Convergence and Empirical Processes. With Applications to Statistics. Springer.