STATISTICAL INFERENCE FOR LARGE MATRICES

A Thesis
Presented to
The Academic Faculty

by

Dong Xia

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in
Computational Science and Engineering

School of Mathematics
Georgia Institute of Technology
August 2016

STATISTICAL INFERENCE FOR LARGE MATRICES

Approved by:

Professor Vladimir Koltchinskii,
Committee Chair
School of Mathematics
Georgia Institute of Technology

Professor Vladimir Koltchinskii,
Advisor
School of Mathematics
Georgia Institute of Technology

Professor Karim Lounici
School of Mathematics
Georgia Institute of Technology

Professor Justin Romberg
School of Electrical and Computer
Engineering
Georgia Institute of Technology

Professor Le Song
School of Computational Science and
Engineering
Georgia Institute of Technology

Professor Prasad Tetali
School of Mathematics
Georgia Institute of Technology

Date Approved: 18 July 2016

To my family.

PREFACE

Many machine learning tasks can be modeled as problems of estimating high-dimensional low rank matrices, such as building recommender systems and predicting links in social networks. The estimation of low rank density matrices plays the essential role in quantum state tomography. One problem studied in this dissertation is the low rank density matrix estimation based on noisy observations of linear measurements of the unknown density matrix. The minimax lower bounds are established for several statistically relevant distances. Then several estimators are studied, showing that these minimax lower bounds are attained up to logarithmic terms. The main theoretic results have been published in the articles [58] and [101].

While most of the thesis is dedicated to the density matrix estimation, another problem is studied in this dissertation which is related to the spectral perturbation bounds of matrices under Gaussian noise. The eigenvectors and singular vectors of matrices have been widely applied in spectral algorithms for many machine learning problems, such as community detection in social networks and the sub-matrix localization. Sharp upper bounds on the perturbation of linear forms of singular vectors under Gaussian noise are developed. This result has been published in the article [59].

ACKNOWLEDGEMENTS

I met so many great people and friends during my past five years at Georgia Tech. This dissertation would not have been possible without their support.

First and foremost, I express my deepest gratitude to my advisor Prof. Vladimir Koltchinskii for his patience and guidance. Vladimir introduces to me interesting questions in probability, statistics and machine learning, which help me figure out my research direction. My Ph.D. life would not have been smooth or productive without the insightful discussions with him. I appreciate his guidance not only in research but also in life.

I am also grateful to Prof. Karim Lounici for his comments on my research ideas. I appreciate his discussion on the sup norm convergence rate and sign consistency of LASSO estimator. I also want to thank Prof. Ionel Popescu for many motivating discussions about random matrix theory.

I thank my fellow graduate students and friends Dawei He, Fan Zhou, Chenchen Mou, Lei Zhang for years of friendship and many interesting discussions in mathematics and statistics. I will remember the enjoyable time spent with them. Finally, I want to thank Huilin and our parents for their support and encouragement.

My research is partly supported through my advisor's grants from NSF DMS-1207808 and CCF-1523768.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

This dissertation studies two problems related to the statistical inference for large matrices. The first problem is on the estimation of a low rank density matrix based on noisy observations of linear measurements of the unknown density matrix with application in quantum state tomography. The density matrices are positively semi-definite Hermitian matrices of unit trace that describe the state of a quantum system. Most quantum states of physical interest can be accurately described by low rank density matrices. It is therefore important to study the statistical limitations of low rank density matrix estimation based on noisy measurements and to propose computationally friendly estimators achieving the optimal convergence rates. The first goal is to develop minimax lower bounds on the error rates of estimating low rank density matrices in trace regression models used in quantum state tomography (in particular, in the case of Pauli measurements) with explicit dependence of the bounds on the rank and other complexity parameters, such as the dimension and sample size. Such bounds are established for several statistically relevant distances, including quantum versions of Kullback-Leibler divergence (relative entropy distance) and of Hellinger distance (also called Bures distance), and Schatten p-norm distances for all $1 \leq p \leq +\infty$. These bounds are proved in both the trace regression model with bounded response and the trace regression model with Gaussian noise. The second goal is to study several well-known estimators and prove that the optimal convergence rates (with additional logarithmic terms) are attained for these estimators in different distances. These estimators include the least squares estimator (which may be penalized by von Neumann entropy), the simple projection estimator and the Dantzig type estimator, which are popular estimators in problems of low rank matrix estimation.

The second problem studied in this dissertation is on the analysis of the perturbation of linear forms of singular vectors of matrices under Gaussian noise. Let $A \in \mathbb{R}^{m\times n}$ be a matrix of rank r with singular value decomposition (SVD) $A = \sum_{k=1}^{r} \sigma_k (u_k \otimes v_k)$, where $\{\sigma_k, k = 1, \ldots, r\}$ are singular values of A (arranged in a non-increasing order) and $u_k \in \mathbb{R}^m, v_k \in \mathbb{R}^n, k = 1, \ldots, r$ are the corresponding left and right orthonormal singular vectors. Let $\tilde{A} = A + X$ be a noisy observation of A, where $X \in \mathbb{R}^{m\times n}$ is a random matrix with i.i.d. Gaussian entries, $X_{ij} \sim \mathcal{N}(0, \tau^2)$, and consider its SVD $\tilde{A} = \sum_{k=1}^{m\wedge n} \tilde{\sigma}_k (\tilde{u}_k \otimes \tilde{v}_k)$ with singular values $\tilde{\sigma}_1 \geq \ldots \geq \tilde{\sigma}_{m\wedge n}$ and singular vectors $\tilde{u}_k, \tilde{v}_k, k = 1, \ldots, m \wedge n$. The goal is to develop sharp concentration bounds for linear forms $\langle \tilde{u}_k, x\rangle, x \in \mathbb{R}^m$ and $\langle \tilde{v}_k, y\rangle, y \in \mathbb{R}^n$ of the perturbed (empirical) singular vectors in the case when the singular values of A are distinct and, more generally, concentration bounds for bilinear forms of projection operators associated with SVD. In particular, the results imply upper bounds of the order $O\left(\sqrt{\frac{\log(m+n)}{m\vee n}}\right)$ (holding with a high probability) on

$$\max_{1\leq i\leq m} \left|\left\langle \tilde{u}_k - \sqrt{1+b_k} u_k, e_i^m \right\rangle\right| \quad \text{and} \quad \max_{1\leq j\leq n} \left|\left\langle \tilde{v}_k - \sqrt{1+b_k} v_k, e_j^n \right\rangle\right|,$$

where b_k are properly chosen constants characterizing the bias of empirical singular vectors $\tilde{u}_k, \tilde{v}_k$ and $\{e_i^m, i = 1, \ldots, m\}, \{e_j^n, j = 1, \ldots, n\}$ are the canonical bases of $\mathbb{R}^m, \mathbb{R}^n$, respectively.

CHAPTER I

INTRODUCTION TO LOW RANK DENSITY MATRIX ESTIMATION

1.1 Notations and basic definitions

Let $\mathbb{R}$ denote the set of real numbers and $\mathbb{R}_+$ denote the set of nonnegative numbers. We denote the set of complex numbers by $\mathbb{C}$. Then the linear space of m-dimensional vectors is denoted by $\mathbb{R}^m$ when the entries are real numbers. Correspondingly, when the entries are complex numbers, we denote it by $\mathbb{C}^m$. All the vectors in this thesis are column vectors. For $a \in \mathbb{R}$, let $|a|$ denote its absolute value . If $a \in \mathbb{C}$, we denote its modulus by $|a|$. In other words, if $a = x + yi$ for $x, y \in \mathbb{R}$, then

$$|a| = \sqrt{x^2 + y^2}.$$

For a vector $v \in \mathbb{C}^m$, denote its transpose by v' and define its l_p norm by

$$\|v\|_p := \Big(\sum_{i=1}^{m} |v_i|^p\Big)^{1/p}, \quad 1 \leq p \leq +\infty,$$

Then, if $p = +\infty$, we have $\|v\|_\infty = \max_{1 \leq i \leq m} |v_i|$. For a matrix $A \in \mathbb{C}^{m_1 \times m_2}$, denote its transpose by A' and define its Schatten p-norm by

$$\|A\|_p := \Big(\sum_{i=1}^{m_1 \wedge m_2} \sigma_i^p(A)\Big)^{1/p}, \quad 1 \leq p \leq +\infty,$$

where $\sigma_1(A) \geq \sigma_2(A) \geq \ldots \geq \sigma_{m_1 \wedge m_2}(A) \geq 0$ are the singular values of A. Note that $m_1 \wedge m_2 := \min(m_1, m_2)$ and $m_1 \vee m_2 := \max(m_1, m_2)$. The rank of A is defined as $r := \mathrm{rank}(A) := \max\{1 \leq i \leq (m_1 \wedge m_2) : \sigma_i(A) > 0\}$. When $p = 1$, $\|A\|_1$ is usually called the nuclear norm or trace norm. Similarly, $\|A\|_2$ is called the Frobenius norm and $\|A\|_\infty$ is called the operator norm or spectral norm. Note that $\|A\|_\infty = \sigma_1(A)$, the largest singular value of A. Moreover, denote the singular value decomposition (SVD) of $A \in \mathbb{C}^{m_1 \times m_2}$ with rank r by

$$A := \sum_{i=1}^{r} \sigma_i(A) u_i \otimes v_i,$$

where $\{u_1, \dots, u_r\} \subset \mathbb{C}^{m_1}$ and $\{v_1, \dots, v_r\} \subset \mathbb{C}^{m_2}$ are two sets of orthonormal vectors. The notation $\otimes$ stands for the tensor product which means $u \otimes v = uv' \in \mathbb{C}^{m_1 \times m_2}$ for $u \in \mathbb{C}^{m_1}$ and $v \in \mathbb{C}^{m_2}$. Moreover, if $A, B \in \mathbb{C}^{m_1 \times m_2}$, we use $A \otimes B$ to denote the tensor product or kronecker product. For example, if

$$A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \in \mathbb{C}^{2\times 2} \quad \text{and} \quad B = \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{pmatrix} \in \mathbb{C}^{2\times 2}$$

then,

$$A \otimes B = \begin{pmatrix} a_{11}B & a_{12}B \\ a_{21}B & a_{22}B \end{pmatrix} \in \mathbb{C}^{4\times 4}.$$

The set of $m \times m$ Hermitian matrices is denoted by $\mathbb{H}_m$: $\mathbb{H}_m = \{A \in \mathbb{C}^{m\times m} : A = A^*\}$ with A^* denoting the adjoint matrix of A. For $A \in \mathbb{H}_m$, let $\mathrm{tr}(A)$ denote the trace of A and let $A \succcurlyeq 0$ mean that A is positively semi-definite. Let $\mathcal{S}_m := \{S \in \mathbb{H}_m : S \succcurlyeq 0, \mathrm{tr}(S) = 1\}$ be the set of all positively semi-definite Hermitian matrices of unit trace called *density matrices*. The von Neumann entropy of a density matrix $\rho \in \mathcal{S}_m$ is defined as

$$V(\rho) := \mathrm{tr}(\rho \log \rho),$$

which is equivalent to $V(\rho) = \sum_{i=1}^{m} \lambda_i \log(\lambda_i)$ for $\{\lambda_i\}_{i=1}^{m}$ being the eigenvalues of ρ. The von Neumann entropy can be viewed as a quantum version of the classical Shannon entropy.

For a Hermitian matrix $S \in \mathbb{H}_m$, its spectral decomposition is given by the following representation,

$$S := \sum_{k\geq 1}^{m'} \mu_k P_k$$

with μ_k being the distinct eigenvalues of S and P_k being the corresponding spectral projectors for $1 \leq k \leq m'$. Clearly, $m_k := \text{rank}(P_k) \leq m$ is the multiplicity of the eigenvalue μ_k. Then the projector P_k represents the orthogonal projection onto the eigenspace corresponding to the eigenvalue μ_k. For an eigenvalue with multiplicity greater than 1, the eigenvectors are not uniquely defined. However, the corresponding spectral projector is unique. The definition of P_k indicates that $P_k P_{k'} = \mathbf{0}$ for any $k \neq k'$ with $\mathbf{0}$ representing the $m \times m$ zero matrix.

C, C_1, C', c, c', etc will denote constants (that do not depend on parameters of interest such as the dimension m and the sample size n) whose values could change from line to line (or, even, within the same line) without further notice. For nonnegative A and B, $A \lesssim B$ (equivalently, $B \gtrsim A$) means that $A \leq CB$ for some absolute constant $C > 0$, and $A \asymp B$ means that $A \lesssim B$ and $B \lesssim A$. Sometimes, symbols $\lesssim, \gtrsim$ and $\asymp$ could be provided with subscripts (say, $A \lesssim_\gamma B$) to indicate that constant C may depend on a parameter (say, γ).

1.2 Quantum state tomography

1.2.1 Quantum systems, quantum states and density matrices

Quantum systems are the fundamental objects in the study of quantum mechanics. We take the simplest quantum systems, namely the two-state systems, as a basic introduction. A two-state system is a system which can exist in any quantum superposition of two physically distinguishable quantum states (see [75] and [35] for more details). The most famous example of a two-state system is the spin of spin-$\frac{1}{2}$ particles, such as electrons and neutrons, etc. The two distinguishable quantum states of a two-state system can be viewed as quantum analogue of the basic characters 0 and 1 in modern computers, leading to their wide application in quantum computation. As a result, such a two-state quantum system is usually called a qubit, akin to the concept of bit in computer theories. The fundamental difference between a quantum system and a classical system is that a bit in the classical system has to be in one state or the other, while a qubit is allowed to be in a superposition of both states at the same time. This is also why scientists need quantum mechanics to describe the states of quantum systems.

To determine and characterize a quantum system, we need to know its state, which is called quantum state. The quantum states are usually described by state vectors in a Hilbert space over complex numbers. For a two-state quantum system, the two basis states are denoted, following the conventional notations, by $\big|0\big\rangle$ and $\big|1\big\rangle$ known as the basis (state) vectors. Then a pure qubit state is a linear superposition of the basis states, meaning that the pure qubit state can be represented as

$$\big|\psi\big\rangle = \alpha\big|0\big\rangle + \beta\big|1\big\rangle,$$

where the complex number α and β are the probability amplitudes. In other words, if we measure this qubit $\big|\psi\big\rangle$ in the standard basis, then with probability $|\alpha|^2$ the outcome is $\big|0\big\rangle$ and with probability $|\beta|^2$ the outcome is $\big|1\big\rangle$. Therefore, the following constraint is obvious:

$$|\alpha|^2 + |\beta|^2 = 1.$$

A pure state can be represented by a single state vector $\big|\psi\big\rangle$ and $\big|\psi\big\rangle$ is usually normalized such that it has unit norm in the Hilbert space. Given a set of pure states $\big\{\big|\psi_s\big\rangle\big\}_{s=1}^{\infty}$, a non-degenerate statistical ensemble of them is called a mixed state.

It is usually more convenient to characterize the quantum states by a positively semi-definite Hermitian matrix which is called density matrix. The density matrix of a mixed state is defined as

$$\rho := \sum_s p_s \big|\psi_s\big\rangle\big\langle\psi_s\big| \tag{1.2.1}$$

with p_s representing the fraction of each pure states in the statistical ensemble, implying that $p_s \geq 0, \forall s \geq 1$ and $\sum_s p_s = 1$. The notation $\big|\psi_s\big\rangle\big\langle\psi_s\big|$ can be viewed as the outer product of the basis vectors $\left|\psi_s\right\rangle$. By the definition of ρ, it is easy to verify that $\rho \succcurlyeq 0$, $\rho = \rho^\star$ and $\mathrm{Tr}(\rho) = 1$.

1.2.2 Multi-qubit systems and Observables

As discussed in Section 1.2.1, it is easy to see that by taking $\left|0\right\rangle$ and $\left|1\right\rangle$ as the basis state vectors for a one-qubit system, any pure state vector $\left|\psi_s\right\rangle$ can be uniquely determined by a 2-dimensional vector $(\alpha, \beta)'$. As a result, the corresponding density matrix ρ defined as (1.2.1) is equivalent to a 2×2 density matrix which belongs to $\mathcal{S}_2$. The measurement of a quantum system is conducted on the so-called Observables, which can be mathematically represented by certain Hermitian operators. In quantum mechanics, these Observables usually correspond to certain physical properties of the system states, for instance, the superposition of the joint states of spin-$\frac{1}{2}$ particles. For a one-qubit system whose density matrix $\rho \in \mathcal{S}_2$, the corresponding Observables can be viewed as Hermitian matrices in $\mathbb{H}_2$. An important class of Observables for one-qubit systems is called the Pauli matrices. They are defined as the follows:

$$\sigma_0 := \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad \sigma_1 := \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad \sigma_2 := \begin{pmatrix} 0 & i \\ -i & 0 \end{pmatrix}, \quad \sigma_3 := \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix},$$

where the matrices $\sigma_1, \sigma_2, \sigma_3$ are often denoted as $\sigma_x, \sigma_y, \sigma_z$, corresponding to the interaction of the spin of a particle with an external electromagnetic field. They can also be viewed as the spin along the coordinate axes in the three-dimensional Euclidean space $\mathbb{R}^3$.

It is natural to extend the definition of state vectors and density matrices to the multi-qubit systems. For a set of b qubits, any pure state vector can be represented as a linear combination of basis vectors $\left\{\left|i_1 i_2 \dots i_b\right\rangle : i_1, \dots, i_b \in \{0, 1\}\right\}$, resulting in a complex vector with dimension $m = 2^b$. In a similar fashion, we can check that for a b-qubit system, its density matrix $\rho \in \mathcal{S}_m$, implying that the dimension of a density matrix grows exponentially with the number of qubits. As a consequence, a large dimensional density matrix is often needed to characterize the states of many-qubit systems. The Observables (namely, Hermitian operators) can be defined accordingly. For example, the Pauli matrices for a b-qubit system consist of the following $m^2 = 4^b$ matrices:

$$\sigma_{i_1} \otimes \ldots \otimes \sigma_{i_b}, \quad (i_1, \ldots, i_b) \in \{0, 1, 2, 3\}^b.$$

Another fundamental difference between quantum systems and classical systems is that higher correlation is allowed in a set of qubits which is usually called entanglement. For example, consider a set of two qubits. It is obvious that the basis vectors for this system can be expressed as

$$\big|00\big\rangle, \big|01\big\rangle, \big|10\big\rangle \quad \text{and } \big|11\big\rangle.$$

The famous *Bell state* of two entangled qubits has the following state vector

$$\big|\psi\big\rangle = \frac{1}{\sqrt{2}}\big|00\big\rangle + \frac{1}{\sqrt{2}}\big|11\big\rangle,$$

which is usually called the maximally entangled quantum state. The entanglement of multiple particles (qubits) is usually caused by the ways in which the group of particles are generated or interacted such that the quantum state of each qubit can not be described independently. Due to the entanglement of multiple qubits, we have to treat them as a whole system.

1.2.3 Quantum state tomography

An important task in quantum computation and quantum information is to determine the state of given quantum systems, which is equivalent to determine its underlying density matrix. The goal of *quantum state tomography* is to estimate the density matrix for a system prepared in an unknown state based on specially designed measurements. Let $X \in \mathbb{H}_m$ be a Hermitian matrix (*an observable*) with spectral representation $X = \sum_{j=1}^{m'} \lambda_j P_j$, where $m' \leq m$, $\lambda_j \in \mathbb{R}, j = 1, \ldots, m'$ being the distinct eigenvalues of X and $P_j, j = 1, \dots, m'$ being the corresponding eigenprojections. For a system prepared in state $\rho \in \mathcal{S}_m$, possible outcomes of a measurement of observable X are the eigenvalues $\lambda_j, j = 1, \dots, m'$ and they occur with probabilities $p_j := \mathrm{tr}(\rho P_j), j = 1, \dots, m'$. If Y is a random variable representing such an outcome, then

$$\mathbb{E}_\rho Y = \mathrm{tr}(\rho X) = \langle \rho, X \rangle.$$

In a simple model of quantum state tomography considered here, an observable X is sampled at random from some probability distribution Π in $\mathbb{H}_m$, $\mathbb{E}_\rho(Y|X) = \langle \rho, X \rangle$ and $Y = \langle \rho, X \rangle + \xi$ with noise ξ such that $\mathbb{E}_\rho(\xi|X) = 0$. Given a sample $X_1, \dots, X_n$ of n i.i.d. copies of X, n measurements of observables $X_1, \dots, X_n$ are performed for a system identically prepared n times in the same unknown state $\rho \in \mathcal{S}_m$ resulting in outcomes $Y_1, \dots, Y_n$. This leads to the following *trace regression model*

$$Y_j = \langle \rho, X_j \rangle + \xi_j, j = 1, \dots, n \tag{1.2.2}$$

with design variables $X_j, j = 1, \dots, n$, response variables $Y_j, j = 1, \dots, n$ and noise $\xi_j, j = 1, \dots, n$ satisfying the assumption $\mathbb{E}_\rho(\xi_j|X_j) = 0, j = 1, \dots, n$ and $\mathbb{E}_\rho(Y_j|X_j) = \langle \rho, X_j \rangle$. The goal is to estimate the target density matrix ρ based on the data $(X_1, Y_1), \dots, (X_n, Y_n)$, with the estimation error being measured by one of the statistically meaningful distances between density matrices such as the Schatten p-norm distances for $p \in [1, \infty]$ or quantum versions of Hellinger and Kullback-Leibler distances. Remember that the difficulties in estimating density matrices lie in the fact that the dimension m of the underlying density matrix ρ is usually very large. For instance, for a quantum system with only 10 qubits, the dimension $m = 2^{10}$ which results into a density matrix with $m^2 = 2^{20}$ entries.

Let's see what is happening when the measurement X are chosen uniformly from Pauli matrices for b-qubit systems with $m = 2^b$. Note that if we define the matrices $W_i = \frac{1}{\sqrt{2}}\sigma_i, \ i = 0, 1, 2, 3$ (see Section 1.2.2), it is easy to check that

$\{W_0, W_1, W_2, W_3\}$ forms an orthonormal basis (the Pauli basis) of the space $\mathbb{H}_2$. For a system consisting of b qubits, the corresponding observables are $m \times m$ Hermitian matrices with $m = 2^b$. The Pauli basis of $\mathbb{H}_m$ is then defined (as introduced in Section 1.2.2) by tensorizing the Pauli basis of $\mathbb{H}_2$: it consists of $m^2 = 4^b$ tensor products $W_{i_1} \otimes \ldots \otimes W_{i_b}, (i_1, \ldots, i_b) \in \{0, 1, 2, 3\}^b$. Let $E_1 = W_0 \otimes \ldots \otimes W_0$ and let $E_2, \ldots, E_{m^2}$ be the rest of the matrices of the Pauli basis of $\mathbb{H}_m$. Define $\mathcal{E} := \{E_1, \ldots, E_{m^2}\}$. It is easy to verify that $\mathcal{E}$ is an orthonormal basis of $\mathbb{H}_m$. It is straightforward to check that $E_1 = \frac{1}{\sqrt{m}} I_m$, where I_m denotes $m \times m$ identity matrix (thus, $\frac{1}{\sqrt{m}}$ is the only eigenvalue of E_1). Matrices $E_2, \ldots, E_{m^2}$ have eigenvalues $\pm\frac{1}{\sqrt{m}}$. Therefore, $\|E_j\|_\infty = m^{-1/2}$, for all $1 \le j \le m^2$. Matrices E_j have the following spectral representations: $E_j = \frac{1}{\sqrt{m}} P_j^+ - \frac{1}{\sqrt{m}} P_j^-$ with eigenprojections $P_j^+, P_j^-, j = 1, \ldots, m^2$ (for E_1, $P_1^- = 0$). A measurement of E_j for a b qubit system prepared in state ρ results in a random outcome τ_j with two possible values $\pm\frac{1}{\sqrt{m}}$ taken with probabilities $\langle \rho, P_j^\pm \rangle$. For random variable τ_j, $\mathbb{E}_\rho \tau_j = \langle \rho, E_j \rangle$. The density matrix ρ admits the following representation in the Pauli basis:

$$\rho = \sum_{j=1}^{m^2} \frac{\alpha_j}{\sqrt{m}} E_j$$

with $\alpha_1 = 1$ and with some $\alpha_j \in \mathbb{R}, j = 2, \ldots, m^2$. This implies that $\mathbb{E}_\rho \tau_j = \frac{\alpha_j}{\sqrt{m}}$,

$$\mathbb{P}_\rho\Big\{\tau_j = \pm\frac{1}{\sqrt{m}}\Big\} = \frac{1 \pm \alpha_j}{2}$$

and $\mathrm{Var}_\rho(\tau_j) = \frac{1-\alpha_j^2}{m}$. Note that, for $j = 1$, $\alpha_1 = 1$, $\mathbb{P}_\rho\Big\{\tau_1 = \frac{1}{\sqrt{m}}\Big\} = 1$ and $\mathrm{Var}_\rho(\tau_1) = 0$. For $j = 2, \ldots, m^2$, $|\alpha_j| < 1$ and $\mathrm{Var}_\rho(\tau_j) > 0$.

Let ν be picked at random from the set $\{1, \ldots, m^2\}$ (with the uniform distribution) and let $X = E_\nu, Y = \tau_\nu$ (which corresponds to random sampling from the Pauli basis with a subsequent measurement of observable X resulting in the outcome Y). Then $\mathbb{E}_\rho(Y|X) = \langle \rho, X \rangle$ and $Var_\rho(Y|X) = \frac{1-\alpha_\nu^2}{m}$. Moreover, we have

$$\mathbb{P}\Big\{Var_\rho(Y|X) \le \frac{1}{2m}\Big\} = \mathbb{P}\Big\{\alpha_\nu^2 \ge \frac{1}{2}\Big\} \le 2\mathbb{E}\alpha_\nu^2 = \frac{2}{m}\sum_{j=1}^{m^2} \frac{\alpha_j^2}{m} = \frac{2\|\rho\|_2^2}{m}.$$

Since, for $\rho \in \mathcal{S}_m$, $\|\rho\|_2 \leq 1$, this means that, for $m > 2$ with probability at least $1 - \frac{2}{m}$, $\mathrm{V}ar_\rho(Y|X) > \frac{1}{2m}$. In other words, the number of $j = 1, \ldots, m^2$ such that $\mathrm{V}ar_\rho(\tau_j) > \frac{1}{2m}$ is at least $m^2 - 2m$ implying that, for the most of the values of j, $\mathrm{V}ar_\rho(\tau_j) \asymp \frac{1}{m}$.

The variance could be further reduced by repeating the measurement of the observable X K times (for a system identically prepared in state ρ) and averaging the outcomes of the resulting K measurements. In this case, the response variable becomes $Y = \langle \rho, X \rangle + \xi$, where $\mathbb{E}_\rho(\xi|X) = 0$ and $\mathbb{E}_\rho(\xi^2|X) = \mathrm{Var}_\rho(Y|X) = \frac{1-\alpha_\nu^2}{Km}$.

1.3 Low rank (density) matrix estimation

1.3.1 The trace regression model of low rank matrix estimation

Low rank matrix estimation has been studied for several years in the literature, such as [20], [57], [55] and [49] with references therein. In the general settings, we have independent pairs of measurements and outputs, $(X_1, Y_1), \ldots, (X_n, Y_n) \in (\mathbb{R}^{m_1 \times m_2}, \mathbb{R})$ which are related to an unknown matrix $A_0 \in \mathbb{R}^{m_1 \times m_2}$. The dimensions m_1 and m_2 are often very large such that $n \ll m_1 m_2$. It is usually assumed that A_0 has low rank, *i.e.*, $r = \mathrm{rank}(A_0) \ll (m_1 \wedge m_2)$ such that the estimation complexity is significantly reduced. The observations $(X_j, Y_j), j = 1, \ldots, n$ satisfy the trace regression model (also introduced in Section 1.2.3 for quantum state tomography):

$$Y_j = \langle A_0, X_j \rangle + \xi_j, \quad j = 1, \ldots, n \tag{1.3.1}$$

where $\xi_j, j = 1, \ldots, n$ are *i.i.d.* random noises with $\mathbb{E}(\xi|X) = 0$ and $\mathbb{E}(\xi^2|X) \leq \sigma_\xi^2 < +\infty$. Note that when $\sigma_\xi = 0$, it corresponds to the problem of the exact recovery of low rank matrices.

We begin with the clarification of some notations. Let $\langle A, B \rangle$ denote $\mathrm{Tr}(A^T B)$ for any $A, B \in \mathbb{R}^{m_1 \times m_2}$. The measurement X is usually assumed to be sampled randomly from some set (of measurements) $\mathcal{X} \subset \mathbb{R}^{m_1 \times m_2}$. We use Π to denote the distribution of X. The distribution based dot product and L_2-norm are defined as

$$\langle A, B\rangle_{L_2(\Pi)} := \mathbb{E}\langle A, X\rangle \langle B, X\rangle$$

and

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

Given the data $X_1, \ldots, X_n \in \mathbb{R}^{m_1 \times m_2}$, let Π_n denote the empirical distribution constructed from $X_1, \ldots, X_n$. In a similar fashion, we can define the $L_2(\Pi_n)$ norm and inner product as

$$\|A\|^2_{L_2(\Pi_n)} := \frac{1}{n}\sum_{i=1}^n \big\langle A, X_i\big\rangle^2$$

and

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

There are several popular measurements $\mathcal{X}$ and distributions Π considered in the literature. There is an incomplete list of these examples given as follows.

Example 1. Matrix Completion *In this situation, the distribution Π denotes some distribution on the set*

$$\mathcal{X} = \{e_j(m_1) \otimes e_k(m_2), j = 1, \ldots, m_1, k = 1, \ldots, m_2\} \subset \mathbb{R}^{m_1 \times m_2}$$

where $e_j(m)$ denotes the j-th canonical basis vector in $\mathbb{R}^m$. It is usually assumed that Π is a (nearly) uniform distribution on the set $\mathcal{X}$, see [53], [57], [80], [66] and [49]. In other words, the task of matrix completion is to estimate A_0 from randomly observed entries of A_0 which are corrupted with noises. [80] also considered sampling without replacement from $\mathcal{X}$, i.e. $X_1, \ldots, X_n$ must be different from each other. When Π denotes the uniform distribution on $\mathcal{X}$, we have $||A||^2_{L_2(\Pi)} = \frac{1}{m_1 m_2}||A||_2^2$ and $\langle A, B\rangle_{L_2(\Pi)} = \frac{1}{m_1 m_2}\langle A, B\rangle$.

Example 2. Sub-Gaussian Design *In this situation, $X_j, j = 1, \ldots, n$ are i.i.d. random matrices. The entries of every X_j are all i.i.d. sub-Gaussian random variables.*

A real-valued random variable x is said to be sub-Gaussian with parameter $b > 0$ if it has the property that for every $t \in \mathbb{R}$ one has: $\mathbb{E}e^{tx} \leq e^{b^2t^2/2}$. Two important examples of such random variables are Gaussian random variables and Rademacher random variables. A random variable $z \sim \mathcal{N}(0, \sigma^2)$ is a Gaussian random variable with $\mathbb{E}z = 0$, $\mathbb{E}z^2 = \sigma^2$ and a probability density function $f_z(z) = \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{z^2}{2\sigma^2}}$. A random variable ε is called a Rademacher random variable if $\mathbb{P}(z = \pm 1) = \frac{1}{2}$. Note that in the case of Gaussian design and Rademacher design, we have $||A||_{L_2(\Pi)} = ||A||_2$ and $\langle A, B\rangle_{L_2(\Pi)} = \langle A, B\rangle$. It is also studied in [53] the estimation of density matrices in quantum state tomography under sub-Gaussian design. The Gaussian measurements are widely studied in low rank estimation problem for the reason that, with high probability, Gaussian random sampling operator satisfies the Restricted Isometry Property (see [18], [20], [15] and [17]) and restricted strong convexity (see [71], [72]).

Example 3. Rank One Projection *As described in [16], both Example 1 and Example 2 have disadvantages. Under the matrix completion model, in order to get a robust estimation of matrix A_0, as pointed out by [21], [22], [19], [78], [86] and [36], additional structral assumptions are needed. To be more exact, such structural assumptions are called the incoherent conditions. Actually, it is impossible to recover spiked matrices under matrix completion model. On the other hand, by using the sub-Gaussian measurements, every measurement $X_j, j = 1, \ldots, n$ requires $\mathcal{O}(m_1m_2)$ bytes of space for storage, which is huge when m_1 and m_2 are large. Therefore, [16] proposed the rank one projection, $X_j = \alpha_j^T\beta_j$, where $\alpha_j \in \mathbb{R}^{m_1}$ and $\beta_j \in \mathbb{R}^{m_2}$ are i.i.d. sub-Gaussian vectors for $j = 1, \ldots, n$. They proved that under rank one projection, there exists robust procedures to construct a stable estimator without addition structural assumptions. In addition, only $\mathcal{O}(m_1 + m_2)$ bytes of space are needed for storage of every $X_j, j = 1, \ldots, n$.*

It worths to point out that by using the Pauli basis $\mathcal{X} = \mathcal{E} := \{E_1, \ldots, E_{m^2}\}$ (see Section 1.2.3) and by using the uniform distribution over $\mathcal{E}$ as the design of Π, we have $\|A\|_{L_2(\Pi)}^2 = \frac{1}{m^2}\|A\|_2^2$ for all $A \in \mathbb{H}_m$. This model has been studied in [53], [54], [98] and [36]. It will be the standard setting of design in this thesis.

1.3.2 Nuclear norm penalization and computationally feasible approaches

As discussed in Section 1.3.1, the objective of low rank matrix estimation is to recover the underlying matrix $A_0 \in \mathbb{R}^{m_1 \times m_2}$ from the data $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ satisfying the trace regression model (1.3.1) in the case that the dimensions m_1 and m_2 are large such that $n \ll m_1 m_2$. Typically, the assumption that $r := \text{rank}(A_0) \ll (m_1 \wedge m_2)$ is imposed to make it possible to obtain a robust estimation of A_0 when $n = O(m_1 \vee m_2)r$ which can be much smaller than $m_1 m_2$. First, consider the situation that it is known that $\text{rank}(A_0) \leq r$, then an obvious estimator of A_0 is the following one:

$$\mathring{A} := \mathop{\arg\min}_{A \in \mathbb{R}^{m_1 \times m_2}, \text{rank}(A) \leq r} \frac{1}{n} \sum_{i=1}^{n} \big(Y_i - \langle A, X_i \rangle\big)^2. \tag{1.3.2}$$

The estimator (1.3.2) involves a non-convex optimization procedure which is usually computationally infeasible. However, we can write $A = UV'$ such that $U \in \mathbb{R}^{m_1 \times r}$ and $V \in \mathbb{R}^{m_2 \times r}$ when r is known. Then the optimization problem in (1.3.2) can be solved efficiently by the alternating minimization approaches, see [46], [45] and [41]. In the case that the rank of A_0 is unknown, the rank penalized least squares estimator has also studied:

$$\breve{A}_\varepsilon := \mathop{\arg\min}_{A \in \mathbb{R}^{m_1 \times m_2}} \frac{1}{n} \sum_{i=1}^{n} \big(Y_i - \langle A, X_i \rangle\big)^2 + \varepsilon \cdot \text{rank}(A) \tag{1.3.3}$$

with some regularization parameter $\varepsilon > 0$, see [48] and [2]. The estimator (1.3.3) aims at searching for a solution with a balance between the sum of squares (loss) and its rank. Note that in general, the optimization problem in (1.3.3) is difficult to solve, making it computationally infeasible.

The numerical difficulty of the rank penalized approaches for estimating low rank matrices originates from the non-smoothness and non-convexity of the rank function on $\mathbb{R}^{m_1 \times m_2}$. Note that low rank matrix estimation problem has similarities to sparse vector estimation problems which is usually called compressed sensing (CS), where the l_1 norm is used as a convex surrogate of the non-smooth and non-convex l_0 norm, see [89], [104], [103], [10], [18] and a large body of references therein. In view of the similarity between low rank matrix estimation problems and CS, it has been conjectured that the nuclear norm should be a good surrogate for the rank function. This is confirmed in the pioneer work [22], [21] and further developed in a lot of following work, including [79], [45], [57], [6], [88], etc. By considering the nuclear norm as a penalization for the standard least squares estimator, the following so-called matrix LASSO estimator is studied,

$$\hat{A}_\varepsilon := \operatorname*{arg\,min}_{A\in\mathcal{A}} \frac{1}{n}\sum_{i=1}^n \big(Y_i - \langle A, X_i\rangle\big)^2 + \varepsilon\cdot\|A\|_1 \tag{1.3.4}$$

for some regularization parameter $\varepsilon > 0$. The convex set $\mathcal{A}$ can be $\mathbb{R}^{m_1\times m_2}$ in many situations, while in some cases, $\mathcal{A} := \{A \in \mathbb{R}^{m_1\times m_2} : \max_{i,j}|A_{ij}| \le a\}$ meaning that there is a uniform upper bound on the entries of A. This estimator has been well studied in [57], [55], [49], [26], [67] and references therein.

Note that we can rewrite the sum of squares as

$$\frac{1}{n}\sum_{i=1}^n \big(Y_i - \langle A, X_i\rangle\big)^2 = \frac{1}{n}\sum_{i=1}^n \langle A, X_i\rangle^2 - \frac{2}{n}\sum_{i=1}^n Y_i\langle A, X_i\rangle + \frac{1}{n}\sum_{i=1}^n Y_i^2.$$

As a result, it is equivalent to write the matrix LASSO estimator as

$$\hat{A}_\varepsilon := \operatorname*{arg\,min}_{A\in\mathcal{A}} \|A\|^2_{L_2(\Pi_n)} - \frac{2}{n}\sum_{i=1}^n Y_i\langle A, X_i\rangle + \varepsilon\cdot\|A\|_1.$$

Remember that Π_n denotes the empirical version of Π, which is usually known in many problems. After replacing $\|A\|_{L_2(\Pi_n)}$ by $\|A\|_{L_2(\Pi)}$, we get the modified matrix LASSO estimator:

$$\check{A}_\varepsilon := \operatorname*{arg\,min}_{A\in\mathcal{A}} \|A\|^2_{L_2(\Pi)} - \frac{2}{n}\sum_{i=1}^n Y_i\langle A, X_i\rangle + \varepsilon\cdot\|A\|_1, \tag{1.3.5}$$

which was introduced and studied in [57], see also [54], [49]. Remember that in many situations (for example, uniform distribution over an orthonormal basis, see Section 1.3.1), $\|A\|_{L_2(\Pi)} = \frac{1}{\sqrt{m_1 m_2}}\|A\|_2$. In these cases, the estimator (1.3.5) is equivalent to

$$\breve{A}_\varepsilon := \arg\min_{A\in\mathcal{A}} \left\| A - \frac{m_1m_2}{n}\sum_{i=1}^n Y_iX_i \right\|_2^2 + m_1m_2\varepsilon\cdot\|A\|_1.$$

Moreover, if $\mathcal{A} = \mathbb{R}^{m_1\times m_2}$, then the above estimator can be solved by a simple singular value thresholding algorithm applied on the matrix $\frac{m_1m_2}{n}\sum_{i=1}^n Y_iX_i$, see [50], [14], [23], which is usually computationally efficient.

Both the matrix LASSO estimator (1.3.4) and the modified matrix LASSO estimator (1.3.5) are based on (modified) penalized least squares estimator. Another type estimator is called the Dantzig estimator which was first introduced in compressed sensing, see [18], [10], [42], [51] and [102]. The matrix Dantzig estimator is defined as follows

$$\acute{A}_\varepsilon := \arg\min\Big\{\|A\|_1 : \Big\|\frac{1}{n}\sum_{i=1}^n \big(Y_i - \langle A, X_i\rangle\big)X_i\Big\|_\infty \le \varepsilon, A\in\mathcal{A}\Big\}. \tag{1.3.6}$$

Note that the estimator (1.3.6) also involves a convex optimization problem which can be solved efficiently. The Dantzig estimator $\acute{A}_\varepsilon$ aims at searching for a solution A with a minimal nuclear norm over the feasible set containing all the oracles having good fitness in the data. The matrix Dantzig estimator $\acute{A}_\varepsilon$ after setting $\varepsilon = 0$ is the most popular method for exact low rank matrix completion when there is no noise, see [21] and [78].

1.3.3 Low rank density matrix estimation and von Neumann entropy

Remember that in quantum state tomography, the objective is to recover an unknown density matrix $\rho\in\mathcal{S}_m$ from a set of pairs of measurements and outcomes, $\big\{(X_1, Y_1), \ldots, (X_n, Y_n)\big\}$ such that

$$Y_i = \big\langle\rho, X_i\big\rangle + \xi_i, \quad i = 1,\ldots,n.$$

As introduced in Section 1.2.3, the dimension m is usually large and the sample size $n \ll m^2$. However, many important and interesting quantum states have density matrices which are low rank or nearly low rank. Therefore, it is natural to adopt the framework of low rank matrices estimation to the settings of quantum state tomography, see [64], [52], [54] and [30]. Note that the definition of density matrices indicates that $\|S\|_1 = 1$ for all $S \in \mathcal{S}_m$. As a result, the matrix LASSO estimator (1.3.4) in Section 1.3.2 for estimating density matrices is equivalent to

$$\hat{\rho}_\varepsilon := \operatorname*{arg\,min}_{S \in \mathcal{S}_m} \frac{1}{n} \sum_{i=1}^{n} \big(Y_i - \langle S, X_i \rangle\big)^2$$

for any $\varepsilon > 0$, which is actually the standard least squares estimator. Similarly, the modified matrix LASSO estimator (1.3.5) is equivalent to

$$\check{\rho}_\varepsilon := \operatorname*{arg\,min}_{S \in \mathcal{S}_m} \Big\| S - \frac{m^2}{n} \sum_{i=1}^{n} Y_i X_i \Big\|_2^2 \tag{1.3.7}$$

for any $\varepsilon > 0$. It is easy to see that the estimator (1.3.7) is equivalent to projecting the matrix $\frac{m^2}{n} \sum_{i=1}^{n} Y_i X_i$ onto the set of density matrices, meaning that $\check{\rho}_\varepsilon$ is the closest point to $\frac{m^2}{n} \sum_{i=1}^{n} Y_i X_i$ in the Frobenius norm or the Hilbert Schmidt norm.

Recall the definition of von Neumann entropy: $V(S) = -\mathrm{tr}(S \log S)$ for any $S \in \mathcal{S}_m$, which plays an important role in quantum information theory. It is often an important task to produce certain quantum systems with maximum von Neumann entropy, see [13]. In order to obtain an estimation of ρ with maximum entropy, the following estimator was introduced in [53]:

$$\tilde{\rho}^\varepsilon := \operatorname*{arg\,min}_{S \in \mathcal{S}_m} \frac{1}{n} \sum_{i=1}^{n} \Big(Y_i - \langle S, X_i \rangle\Big)^2 + \varepsilon \cdot \mathrm{tr}(S \log S) \tag{1.3.8}$$

with certain penalization parameter $\varepsilon > 0$. As introduced in [53], one advantage of applying the the von Neumann entropy as a penalization for least squares estimator is that the bounds on Kullback-Leibler divergence of $\tilde{\rho}^\varepsilon$ can be attained. The Kullback-Leibler divergence of two density matrices is defined as

$$K(S_1 \| S_2) := \mathrm{tr}\big(S_1(\log S_1 - \log S_2)\big), \quad S_1, S_2 \in \mathcal{S}_m, \tag{1.3.9}$$

which is the quantum version of the canonical Kullback-Leibler divergence between probability measures. It is easy to check that $K(S_1 \| S_2) \geq 0$ for any $S_1, S_2 \in \mathcal{S}_m$.

In the case that $K(S_1\|S_2)$ is undefined (for instance, S_2 is not of full rank), we set $K(S_1\|S_2) = +\infty$. Then, the symmetric Kullback-Leibler divergence is defined as

$$K(S_1, S_2) = K(S_1\|S_2) + K(S_1\|S_1).$$

A version of the (squared) Hellinger distance that will be studied is defined as

$$H^2(S_1, S_2) := 2 - 2\mathrm{tr}\sqrt{S_1^{\frac{1}{2}} S_2 S_1^{\frac{1}{2}}}$$

for $S_1, S_2 \in \mathcal{S}_m$ (see also [75]). Clearly, $0 \le H^2(S_1, S_2) \le 2$. It is usually called Bures distance, but it worth to point out that it does not coincide with $\mathrm{tr}(\sqrt{S_1} - \sqrt{S_2})^2$ (which is another possible non-commutative extension of the classical Hellinger distance). In fact, $H^2(S_1, S_2) \le \mathrm{tr}(\sqrt{S_1} - \sqrt{S_2})^2, S_1, S_2 \in \mathcal{S}_m$, but the opposite inequality does not necessarily hold. The quantity $\mathrm{tr}\sqrt{S_1^{\frac{1}{2}} S_2 S_1^{\frac{1}{2}}}$ in the right hand side of the definition of H^2 is a quantum version of Hellinger affinity.

The following very useful inequality is a noncommutative extension of similar classical inequalities for total variation, Hellinger and Kullback-Leibler distances. It follows from representing the "noncommutative distances" involved in the inequality as suprema of the corresponding classical distances between the distributions of outcomes of measurements for two states S_1, S_2 over all possible measurements represented by positive operator valued measures (see, [75], [47], [53], Section 3 and references therein).

Lemma 1. *For all $S_1, S_2 \in \mathcal{S}_m$, the following inequalities hold:*

$$\frac{1}{4}\|S_1 - S_2\|_1^2 \le H^2(S_1, S_2) \le \big(K(S_1\|S_2) \wedge \|S_1 - S_2\|_1\big). \tag{1.3.10}$$

1.4 The upper bounds of empirical processes

The upper bounds for the supremum of empirical process and Rademacher process are powerful tools in characterizing the excess risk of empirical risk minimization in statistical learning theory. They will be frequently used in proving the low rank oracle inequalities.

1.4.1 Concentration bounds of the supremum of empirical processes

Let $(S, \mathcal{A}, P)$ be a probability space with σ-algebra $\mathcal{A}$. Let $X, X_1, \dots, X_n$ be *i.i.d.* random variables in the measurable space $(S, \mathcal{A})$ with a common distribution P. Then denote the empirical distribution by P_n. Let $\mathcal{F}$ be a class of measurable functions defined on the space $(S, \mathcal{A})$. The empirical process indexed by the function class $\mathcal{F}$ is a stochastic process defined as

$$Z_n(f) := P_n f - Pf, \quad f \in \mathcal{F}.$$

The supremum of this empirical process is denoted by $\|P_n - P\|_{\mathcal{F}} := \sup_{f\in\mathcal{F}} |P_n f - Pf|$, where certain measurability assumptions are required to guarantee the measurability of $\|P_n - P\|_{\mathcal{F}}$, see [52], [28] and [92]. Without further notification, we assume in what follows that $\|P_n - P\|_{\mathcal{F}}$ is a properly measurable random variable on $(S, \mathcal{A})$.

There are several types of concentration inequalities of $\|P_n - P\|_{\mathcal{F}}$, see [52] for a list of these inequalities. We introduce some most useful ones which are easiest to apply in many situations. The first concentration inequality is usually referred as the Bousquet's version of Talagrand's concentration inequality, which assumes that

$$\sup_{f\in\mathcal{F}, x\in S} |f(x)| \le U,$$

namely, the class $\mathcal{F}$ is uniformly upper bounded by a constant $U > 0$.

Theorem 1. *[11] Let the function class $\mathcal{F}$ be uniformly upper bounded by $U > 0$. Then, the following bound holds with probability at least $1 - e^{-t}$ for all $t > 0$,*

$$\|P_n - P\|_{\mathcal{F}} \le \mathbb{E}\|P_n - P\|_{\mathcal{F}} + \sqrt{2\frac{t}{n}\big(\sigma_p^2(\mathcal{F}) + 2\mathbb{E}\|P_n - P\|_{\mathcal{F}}\big)} + \frac{3tU}{n}$$

with $\sigma_p^2(\mathcal{F}) = \sup_{f\in\mathcal{F}} Var(f) = \sup_{f\in\mathcal{F}} \big(Pf^2 - (Pf)^2\big)$.

When the random variable $f(x)$ is unbounded but has an exponential tail for each $f \in \mathcal{F}$, it is also possible to derive a version of Talagrand's concentration inequality, see [1]. This kind of bound is further developed in [60] when the envelop of $\mathcal{F}$ has only

q-th moment for $q > 2$. In order to characterize the tail type of random variables, we introduce the Orlicz norms, see [92] and [52]. For a convex increasing function ψ with $\psi(0) = 0$, define

$$\|f\|_\psi := \inf\Big\{C \geq 0 : \int_S \psi\Big(\frac{|f|}{C}\Big) dP \leq 1\Big\}.$$

Note that if $\psi(x) = x^p, x \geq 0$ for some $p \geq 1$, then the norm $\|\cdot\|_\psi$ is just the L_p norm for $p \geq 1$. Another type of choices are functions $\psi_\alpha(x) = e^{x^\alpha} - 1, x \geq 0, \alpha \geq 1$. If $\|X\|_{\psi_1} < +\infty$, it indicates that X has a sub-exponential tail. Moreover, if $\|X\|_{\psi_2} < +\infty$, then X has a sub-Gaussian tail in which case X is usually called a sub-Gaussian random variable, see Example 2 in Section 1.3.1.

Theorem 2. *Let $F(x), x \in S$ be an envelop function of $\mathcal{F}$ such that $F(x) \geq \sup_{f\in\mathcal{F}}|f(x)|$ for all $x \in S$. Then, the following bound holds with probability at least $1 - e^{-t}$ for all $t > 0$,*

$$\|P_n - P\|_{\mathcal{F}} \leq K\Big[\mathbb{E}\|P_n - P\|_{\mathcal{F}} + \sigma_p(\mathcal{F})\sqrt{\frac{t}{n}} + \Big\|\max_{1\leq i\leq n} F(X_i)\Big\|_{\psi_1}\frac{t}{n}\Big]$$

for some universal constant $K > 0$.

1.4.2 Upper bounds of the expectation of supremum of empirical processes

In order to obtain the upper bounds of the supremum of empirical processes, it is further needed to prove the upper bound of $\mathbb{E}\|P_n - P\|_{\mathcal{F}}$, see Theorem 1 and Theorem 2. It is very helpful to control $\mathbb{E}\|P_n - P\|_{\mathcal{F}}$ by using the expectation of the supremum of the so-called Rademacher process, which is actually a sub-Gaussian process. The Dudley's entropy bound and Talagrand's generic chaining bound are powerful tools to control the expectation of the supremum of sub-Gaussian processes, see [87], [52], [92] and [32].

The Rademacher process indexed by a class $\mathcal{F}$ is defined as

$$R_n(f) := \frac{1}{n}\sum_{i=1}^n \varepsilon_i f(X_i), \quad f \in \mathcal{F},$$

with $\varepsilon_i, i = 1, \ldots, n$ being *i.i.d.* Rademacher random variables independent of $X_i, i = 1, \ldots, n$. The following symmetrization inequality is a useful tool to control $\mathbb{E}\|P_n - P\|_{\mathcal{F}}$, whose proof can be found, for instance, [92] and [52]. Note that the expectation $\mathbb{E}\|R_n\|_{\mathcal{F}}$ is respect both to $X_1, \ldots, X_n$ and $\varepsilon_1, \ldots, \varepsilon_n$.

Lemma 2. *For any class $\mathcal{F}$ of P-integral functions and for any convex function $\phi : \mathbb{R}_+ \mapsto \mathbb{R}_+$,*

$$\mathbb{E}\phi\big(\frac{1}{2}\|R_n\|_{\mathcal{F}_c}\big) \leq \mathbb{E}\phi\big(\|P_n - P\|_{\mathcal{F}}\big) \leq \mathbb{E}\phi\big(2\|R_n\|_{\mathcal{F}}\big),$$

where $\mathcal{F}_c := \{f - Pf, f \in \mathcal{F}\}$. In particular,

$$\frac{1}{2}\mathbb{E}\|R_n\|_{\mathcal{F}_c} \leq \mathbb{E}\|P_n - P\|_{\mathcal{F}} \leq 2\mathbb{E}\|R_n\|_{\mathcal{F}}.$$

If $X_1, \ldots, X_n$ are fixed, then $R_n(f)$ is a sub-Gaussian random variable. Moreover,

$$\mathbb{E}\sup_{f\in\mathcal{F}} R_n(f) = \mathbb{E}_X\mathbb{E}_\varepsilon \sup_{f\in\mathcal{F}} R_n(f).$$

It turns out that the expectation of the supremum of Rademacher process $\mathbb{E}_\varepsilon \sup_{f\in\mathcal{F}} R_n(f)$ plays an essential role in the upper bound of $\mathbb{E}\|P_n - P\|_{\mathcal{F}}$. For any subset $T \subset \mathbb{R}^n$ and *i.i.d.* Rademacher variables $\varepsilon_1, \ldots, \varepsilon_n$, we are interested in the quantity

$$R_n(T) := \mathbb{E}\sup_{t\in T}\big|R_n(t)\big| = \mathbb{E}\sup_{t\in T}\Big|\sum_{i=1}^n \varepsilon_i t_i\Big|.$$

The following inequality is called the contraction inequality for Rademacher processes and was first proved by Talagrand.

Lemma 3. *Let $T \subset \mathbb{R}^n$ and let $\varphi_i : \mathbb{R} \mapsto \mathbb{R}$ be functions such that $\varphi_i(0) = 0$ and*

$$|\varphi_i(u) - \varphi_i(v)| \leq |u - v|, \quad u, v \in \mathbb{R}$$

for all $i = 1, \ldots, n$. Then,

$$\mathbb{E}\sup_{t\in T}\Big|\sum_{i=1}^n \varphi_i(t_i)\varepsilon_i\Big| \leq \mathbb{E}\sup_{t\in T}\Big|\sum_{i=1}^n t_i\varepsilon_i\Big|.$$

Let (T, d) be a pseudo-metric space such that $T \subset \mathbb{R}^n$ is equipped with a pseudo distance d. The diameter of (T, d) is defined as

$$D(T) := \sup_{t_1, t_2 \in T} d(t_1, t_2).$$

A subset $T_\varepsilon \subset T$ is called an ε-covering of T if for any $t \in T$, there exists $t' \in T_\varepsilon$ such that $d(t, t') \leq \varepsilon$. Then let $N(T, \varepsilon, d)$ denote the ε-covering number of (T, d), namely, the smallest cardinality over all possible ε-coverings of T. Also, denote $M(T, \varepsilon, d)$ the ε-packing number of (T, d), namely, the largest number of points in T separated from each other by a distance at least ε. By the definitions, it is easy to check

$$N(T, \varepsilon, d) \leq M(T, \varepsilon, d) \leq N(T, \varepsilon/2, d).$$

Then, the ε-entropy number is defined as

$$H(T, \varepsilon, d) = \log N(T, \varepsilon, d).$$

Theorem 3 (Dudley's entropy bound)**.** *Let $T \subset \mathbb{R}^d$ such that (T, d) is a pseudo-metric space and $\varepsilon_1, \ldots, \varepsilon_n$ be i.i.d. Rademacher variables. Then*

$$\mathbb{E} \sup_{t \in T} R_n(t) \leq C \int_0^{D(T)} H^{1/2}(T, \varepsilon, d) d\varepsilon$$

for some numerical constant $C > 0$. Moreover, for all $t_0 \in T$,

$$\mathbb{E} \sup_{t \in T} |R_n(t) - R_n(t_0)| \leq C \int_0^{D(T)} H^{1/2}(T, \varepsilon, d) d\varepsilon$$

The Dudley's entropy bound holds for all sub-Gaussian processes where the Rademacher process is a special case.

CHAPTER II

OPTIMAL ESTIMATION OF LOW RANK DENSITY MATRICES

In this chapter, we introduce the main results of low rank density matrix estimation. Both the trace regression model with bounded response and the trace regression model with Gaussian noise will be considered. We first prove the minimax lower bounds and these bounds are established in several statistical important distances, including the Schatten p-norms for $1 \leq p \leq +\infty$, the Kullback-Leibler divergence and the Hellinger distance.. Then several estimators (introduced as in Section 1.3.2) are studied, including the least squares estimator (1.3.4), the projection estimator (1.3.5) and the Dantzig-type estimator (1.3.6), showing that these estimators are able to achieve the optimal convergence rates which match the minimax lower bounds except some logarithmic factors.

In general, the trace regression model involves a random couple (X, Y) satisfying the model

$$\mathbb{E}(Y|X) = \big\langle \rho, X \big\rangle$$

for some density matrix $\rho \in \mathcal{S}_m$ with low rank, i.e., $r = \text{rank}(\rho) \ll m$. The measurement (Observable) $X \in \mathbb{H}_m$ is usually assumed to be sampled randomly from some distribution which is called the design distribution. Suppose that

$$\mathcal{D}_n := \{(X_1, Y_1), \dots, (X_n, Y_n)\}$$

contains *i.i.d.* samples from the trace regression model. Then, the task is to develop computationally efficient methods to estimate the unknown density matrix ρ from $\mathcal{D}_n$. We are also interested in the informational theoretic bound of the estimation of

ρ, which is also the so-called minimax lower bound.

2.1 The trace regression model and assumptions

A common choice of design distribution in low rank matrix estimation problems is so called *uniform sampling from an orthonormal basis* described in the following assumptions.

Assumption 1. *Let $\mathcal{E} = \{E_1, \ldots, E_{m^2}\} \subset \mathbb{H}_m$ be an orthonormal basis of $\mathbb{H}_m$ with respect to the Hilbert–Schmidt inner product: $\langle A, B \rangle = tr(AB)$. Moreover, suppose that, for some $U > 0$,*

$$\|E_j\|_\infty \leq U, j = 1, \ldots, n,$$

where $\| \cdot \|_\infty$ denotes the operator norm (the spectral norm).

Since $\|E_j\|_2 = 1$, where $\| \cdot \|_2$ denotes the Hilbert–Schmidt (or Frobenius) norm, we can assume that $U \leq 1$. Moreover, $U \geq m^{-1/2}$ since $1 = \|E_j\|_2 \leq m^{1/2}\|E_j\|_\infty \leq m^{1/2}U$. As introduced in Section 1.2.3, when $\mathcal{E}$ is the Pauli basis, the corresponding $U = \frac{1}{\sqrt{m}}$. The fact that the matrices of this basis have the smallest possible operator norms has been used in quantum compressed sensing (see [37], [36], [64]).

Assumption 2. *Let Π be the uniform distribution in the finite set $\mathcal{E}$ (see Assumption 1), let X be a random variable sampled from Π and let $X_1, \ldots, X_n$ be i.i.d. copies of X.*

It will be assumed in this Chapter that assumptions 1 and 2 hold (unless it is stated otherwise). Under these assumptions, $Y_1, \ldots, Y_n$ could be viewed as noisy observations of a random sample of Fourier coefficients $\langle \rho, X_1 \rangle, \ldots, \langle \rho, X_n \rangle$ of the target density matrix ρ in the basis $\mathcal{E}$. The above model (in which $X_1, \ldots, X_n$ are uniformly sampled from an orthonormal basis and $Y_1, \ldots, Y_n$ are the outcomes of measurements of the observables $X_1, \ldots, X_n$ for the system being identically prepared n times in the same state ρ) will be called in what follows the standard QST (quantum state tomography) model. It is a special case of *trace regression model with bounded response*:

Assumption 3 (Trace regression with bounded response)**.** *Suppose that Assumption 1 holds and let (X, Y) be a random couple such that X is sampled from the uniform distribution Π in an orthonormal basis $\mathcal{E} \subset \mathbb{H}_m$. Suppose also that, for some $\rho \in \mathcal{S}_m$, $\mathbb{E}(Y|X) = \langle \rho, X\rangle$ a.s. and, for some $\bar{U} > 0$, $|Y| \leq \bar{U}$ a.s.. The data $(X_1, Y_1), \ldots (X_n, Y_n)$ consists of n i.i.d. copies of (X, Y).*

We are also interested in the *trace regression model with Gaussian noise*:

Assumption 4 (Trace regression with Gaussian noise)**.** *Suppose Assumption 1 holds and let (X, Y) be a random couple such that X is sampled from the uniform distribution Π in an orthonormal basis $\mathcal{E} \subset \mathbb{H}_m$ and, for some $\rho \in \mathcal{S}_m$, $Y = \langle \rho, X\rangle + \xi$, where ξ is a normal random variable with mean 0 and variance σ_ξ^2, ξ and X being independent. The data $(X_1, Y_1), \ldots (X_n, Y_n)$ consists of n i.i.d. copies of (X, Y).*

Note that this model is not directly applicable to the "standard QST problem" described above, where the response variable Y is discrete. However, if the measurements are repeated multiple times for each observable X_j and the resulting outcomes are averaged to reduce the variance, the noise of such averaged measurements becomes approximately Gaussian and it is of interest to characterize the estimation error in terms of the variance of the noise, see more details in Section 1.2.3.

2.2 Minimax lower bounds

In this section, we provide main results on the minimax lower bounds on the risk of estimation of density matrices with respect to Schatten p-norm for $1 \leq p \leq +\infty$ distances, as well as Hellinger-Bures distance and Kullback-Leibler divergence.

Minimax lower bounds will be derived for the class $\mathcal{S}_{r,m} := \{S \in \mathcal{S}_m : \text{rank}(S) \leq r\}$ consisting of all density matrices of rank at most r (the low rank case). We will start with the case of trace regression with Gaussian noise. Given that the sample $(X_1, Y_1), \dots, (X_n, Y_n)$ satisfies Assumption 4 with the target density matrix $\rho \in \mathcal{S}_m$ and noise variance σ_ξ^2, let $\mathbb{P}_\rho$ denote the corresponding probability distribution of the sample.

Note that [68] developed a method of deriving minimax lower bounds for distances based on unitary invariant norms, including Schatten p-norms in matrix problems, and obtained such lower bounds, in particular, in matrix completion problem. The approach used here is somewhat different and the aim is to develop such bounds under an additional constraint that the target matrix is a density matrix. The resulting bounds are also somewhat different, they involve an additional term that does not depend on the rank, but does depend on p. Essentially, it means that the "complexity" of the problem is controlled by a "truncated rank" $r \wedge \frac{1}{\tau}$, where $\tau = \frac{\sigma_\xi m^{3/2}}{\sqrt{n}}$ rather than by the actual rank r. The upper bounds (for instance, see Section 2.3.4) of several estimators studied in the following sections show that such a structure of the bound is, indeed, necessary. It should be also mentioned that minimax lower bounds on the nuclear norm error of estimation of density matrices have been obtained earlier in [30] (see Remark 1 below).

Theorem 4. *For all* $p \in [1, +\infty]$, *there exist constants* $c, c' > 0$ *such that, the following bounds hold:*

$$\inf_{\hat\rho} \sup_{\rho \in \mathcal{S}_{r,m}} \mathbb{P}_\rho \left\{ \|\hat\rho - \rho\|_p \geq c \left(\frac{\sigma_\xi m^{\frac{3}{2}} r^{1/p}}{\sqrt{n}} \bigwedge \left(\frac{\sigma_\xi m^{3/2}}{\sqrt{n}} \right)^{1-\frac{1}{p}} \bigwedge 1 \right) \right\} \geq c', \tag{2.2.1}$$

$$\inf_{\hat\rho} \sup_{\rho \in \mathcal{S}_{r,m}} \mathbb{P}_\rho \left\{ H^2(\hat\rho, \rho) \geq c \left(\frac{\sigma_\xi m^{\frac{3}{2}} r}{\sqrt{n}} \bigwedge 1 \right) \right\} \geq c', \tag{2.2.2}$$

and

$$\inf_{\hat\rho} \sup_{\rho \in \mathcal{S}_{r,m}} \mathbb{P}_\rho \left\{ K(\rho \| \hat\rho) \geq c \left(\frac{\sigma_\xi m^{\frac{3}{2}} r}{\sqrt{n}} \bigwedge 1 \right) \right\} \geq c', \tag{2.2.3}$$

where $\inf_{\hat\rho}$ *denotes the infimum over all estimators* $\hat\rho$ *in* $\mathcal{S}_m$ *based on the data* $(X_1, Y_1), \dots, (X_n, Y_n)$ *satisfying the trace regression model with noise variance* σ_ξ^2.

Proof. A couple of preliminary facts will be needed in the proof. We start with bounds on the packing numbers of Grassmann manifold $\mathcal{G}_{k,l}$, which is the set of all k-dimensional subspaces L of the l-dimensional space $\mathbb{R}^l$. Given such a subspace $L \subset \mathbb{R}^l$ with $\dim(L) = k$, let P_L be the orthogonal projection onto L and let $\mathfrak{P}_{k,l} := \{P_L : L \in \mathcal{G}_{k,l}\}$. The set of all k-dimensional projectors $\mathfrak{P}_{k,l}$ will be equipped with Schatten p-norm distances for all $p \in [1, +\infty]$ (which also could be viewed as distances on the Grassmannian itself): $d_p(Q_1, Q_2) := \|Q_1 - Q_2\|_p, Q_1, Q_2 \in \mathfrak{P}_{k,l}$. Recall that the ε*-packing number* of a metric space (T, d) is defined as

$$D(T, d, \varepsilon) = \max \Big\{ n : \text{there are } t_1, \dots, t_n \in T, \text{such that } \min_{i \neq j} d(t_i, t_j) > \varepsilon \Big\}.$$

The following lemma (see [77, Proposition 8]) will be used to control the packing numbers of $\mathfrak{P}_{k,l}$ with respect to Schatten distances d_q.

Lemma 4. *For all integer $1 \le k \le l$ such that $k \le l - k$, and all $1 \le p \le \infty$, the following bounds hold*

$$\left(\frac{c}{\varepsilon}\right)^d \le D(\mathfrak{P}_{k,l}, d_p, \varepsilon k^{1/p}) \le \left(\frac{C}{\varepsilon}\right)^d, \ \varepsilon > 0 \tag{2.2.4}$$

with $d = k(l - k)$ and universal positive constants c, C.

In addition to this, we need the following well known information-theoretic bound frequently used in derivation of minimax lower bounds (see [91, Theorem 2.5]). Let $\Theta = \{\theta_0, \theta_1, \dots, \theta_M\}$ be a finite parameter space equipped with a metric d and let $\mathcal{P} := \{\mathbb{P}_\theta : \theta \in \Theta\}$ be a family of probability distributions in some sample space. Given $\mathbb{P}, \mathbb{Q} \in \mathcal{P}$, let $K(\mathbb{P}\|\mathbb{Q}) := \mathbb{E}_{\mathbb{P}} \log \frac{d\mathbb{P}}{d\mathbb{Q}}$ be the Kullback-Leibler divergence between $\mathbb{P}$ and $\mathbb{Q}$.

Proposition 1. *Suppose that the following conditions hold:*

(i) for some $s > 0$, $d(\theta_j, \theta_k) \ge 2s > 0, 0 \le j < k \le M$;

(ii) for some $0 < \alpha < 1/8$, $\frac{1}{M} \sum_{j=1}^{M} K(\mathbb{P}_{\theta_j} \| \mathbb{P}_{\theta_0}) \le \alpha \log M$

Then, for a positive constant c_α,

$$\inf_{\hat\theta}\sup_{\theta\in\Theta}\mathbb{P}_\theta\{d(\hat\theta,\theta)\geq s\}\geq c_\alpha,$$

where the infimum is taken over all estimators $\hat\theta\in\Theta$ *based on an observation sampled from* $\mathbb{P}_\theta$.

We now turn to the actual proof of Theorem 4. Under Assumption 4, the following computation is well known: for $\rho_1,\rho_2\in\mathcal{S}_{r,m}$,

$$\begin{aligned}K(\mathbb{P}_{\rho_1}\|\mathbb{P}_{\rho_2}) &= \mathbb{E}_{\mathbb{P}_{\rho_1}}\log\frac{\mathbb{P}_{\rho_1}}{\mathbb{P}_{\rho_2}}\Big(X_1,Y_1,\dots,X_n,Y_n\Big)\\ &= \mathbb{E}_{\mathbb{P}_{\rho_1}}\sum_{j=1}^{n}\Big[-\frac{(Y_j-\langle\rho_1,X_j\rangle)^2}{2\sigma_\xi^2}+\frac{(Y_j-\langle\rho_2,X_j\rangle)^2}{2\sigma_\xi^2}\Big] \\ &= \mathbb{E}\sum_{j=1}^{n}\frac{\langle\rho_1-\rho_2,X_j\rangle^2}{2\sigma_\xi^2}=\frac{n}{2\sigma_\xi^2}\|\rho_1-\rho_2\|_{L_2(\Pi)}^2.\end{aligned}\tag{2.2.5}$$

It is enough to prove the bounds for $2\leq r\leq m/2$. The proof in the case $r=1$ is simpler and the case $r>m/2$ easily reduces to the case $r\leq m/2$. We will use Lemma 4 to construct a well separated (with respect to d_p) subset of density matrices in $\mathcal{S}_{r,m}$. To this end, first choose a subset $\mathcal{D}_p\subset\mathfrak{P}_{r-1,m-1}$ such that $\mathrm{card}(\mathcal{D}_p)\geq 2^{(r-1)(m-r)}$ and, for some constant c', $\|Q_1-Q_2\|_p\geq c'(r-1)^{1/p}$, $Q_1,Q_2\in\mathfrak{P}_{r-1,m-1},Q_1\neq Q_2$. Such a choice is possible due to the lower bound on the packing numbers of Lemma 4. For $Q\in\mathcal{D}_p$ (note that Q can be viewed as an $(m-1)\times(m-1)$ matrix with real entries) and $\kappa\in(0,1)$, consider the following $m\times m$ matrix

$$S=S_Q=\begin{pmatrix}1-\kappa & \mathbf{0}'\\ \mathbf{0} & \kappa\frac{Q}{r-1}\end{pmatrix}.\tag{2.2.6}$$

Note that S is symmetric positively-semidefinite real matrix of unit trace. It is straightforward to check that it defines a Hermitian positively-semidefinite operator in $\mathbb{C}^m$ of unit trace, and it can be identified with a density matrix $S\in\mathcal{S}_m$. Clearly, S is of rank r, so, $S\in\mathcal{S}_{r,m}$.

We will take $\kappa := c_1 \frac{\sigma_\xi m^{3/2}(r-1)}{\sqrt{n}}$ with a small enough absolute constant $c_1 > 0$ and first assume that $\kappa < 1$ (as it is needed in definition Equation 2.2.6).

Let $\mathcal{S}'_p := \{S_Q : Q \in \mathcal{D}_q\}$ and consider a family of $M+1 = \mathrm{card}(\mathcal{D}_p) \geq 2^{(r-1)(m-r)}$ distributions $\{\mathbb{P}_S : S \in \mathcal{S}'_p\}$. It is immediate that for $S_1 = S_{Q_1}, S_2 = S_{Q_2}$, $Q_1, Q_2 \in \mathcal{D}_p, Q_1 \neq Q_2$, we have

$$\begin{aligned} \|S_1 - S_2\|_p &= \frac{\kappa}{r-1}\|Q_1 - Q_2\|_p \geq c'\kappa(r-1)^{1/p-1} \\ &\geq c'c_1 \frac{\sigma_\xi m^{3/2}(r-1)^{1/p}}{\sqrt{n}} \geq c\frac{\sigma_\xi m^{3/2} r^{1/p}}{\sqrt{n}} \end{aligned} \tag{2.2.7}$$

with some constant $c > 0$, implying condition (i) of Proposition 1 with $s = \frac{c}{2}\frac{\sigma_\xi m^{3/2} r^{1/p}}{\sqrt{n}}$.

We will now check its condition (ii) . In view of (2.2.5), we have, for all $S_1 = S_{Q_1}, S_2 = S_{Q_2} \in \mathcal{S}'_p$,

$$\begin{aligned} K(\mathbb{P}_{S_1}\|\mathbb{P}_{S_2}) &= \frac{n}{2\sigma_\xi^2}\|S_1 - S_2\|^2_{L_2(\Pi)} = \frac{n}{2\sigma_\xi^2 m^2}\|S_1 - S_2\|_2^2 \\ &= \frac{n\kappa^2}{2\sigma_\xi^2 m^2 (r-1)^2}\|Q_1 - Q_2\|_2^2 \leq \frac{4n(r-1)\kappa^2}{2\sigma_\xi^2 m^2 (r-1)^2} = 2c_1^2 m(r-1) \quad (2.2.8) \\ &\leq \alpha m(r-1)/\log(2)/4 \leq \frac{\alpha}{2}(r-1)(m-r)\log(2) \leq \alpha \log M, \end{aligned}$$

provided that constant c_1 is small enough, so, condition (ii) of Proposition 1 is also satisfied. Proposition 1 implies that, under the assumption $\kappa = c_1\frac{\sigma_\xi m^{3/2}(r-1)}{\sqrt{n}} < 1$, the following minimax lower bound holds for some $c, c' > 0$:

$$\inf_{\hat{\rho}} \sup_{\rho \in \mathcal{S}_{r,m}} \mathbb{P}_\rho \left\{ \|\hat{\rho} - \rho\|_p \geq c\frac{\sigma_\xi m^{\frac{3}{2}} r^{1/p}}{\sqrt{n}} \right\} \geq c'. \tag{2.2.9}$$

In the case when

$$c_1 \frac{\sigma_\xi m^{3/2}}{\sqrt{n}} < 1 \leq c_1 \frac{\sigma_\xi m^{3/2}(r-1)}{\sqrt{n}},$$

one can choose $2 \leq r' < r-1$ such that, for some constant $c_2 > 0$,

$$c_2 < c_1 \frac{\sigma_\xi m^{3/2}(r'-1)}{\sqrt{n}} < 1.$$

For such a choice of r', it follows from (2.2.9) that

$$\inf_{\hat{\rho}} \sup_{\rho \in \mathcal{S}_{r',m}} \mathbb{P}_\rho \left\{ \|\hat{\rho} - \rho\|_p \geq c\frac{\sigma_\xi m^{\frac{3}{2}} (r')^{1/p}}{\sqrt{n}} \right\} \geq c'. \tag{2.2.10}$$

The definition of r' implies that

$$r' \asymp r' - 1 \asymp \left(\frac{\sigma_\xi m^{3/2}}{\sqrt{n}}\right)^{-1}.$$

Therefore,

$$\frac{\sigma_\xi m^{\frac{3}{2}}(r')^{1/p}}{\sqrt{n}} \asymp \left(\frac{\sigma_\xi m^{3/2}}{\sqrt{n}}\right)^{1-1/p},$$

and, since $\mathcal{S}_{r',m} \subset \mathcal{S}_{r,m}$, bound (2.2.10) yields

$$\inf_{\hat\rho} \sup_{\rho\in\mathcal{S}_{r,m}} \mathbb{P}_\rho\left\{\|\hat\rho - \rho\|_p \geq c\left(\frac{\sigma_\xi m^{3/2}}{\sqrt{n}}\right)^{1-1/p}\right\} \geq \inf_{\hat\rho} \sup_{\rho\in\mathcal{S}_{r',m}} \mathbb{P}_\rho\left\{\|\hat\rho - \rho\|_p \geq c\left(\frac{\sigma_\xi m^{3/2}}{\sqrt{n}}\right)^{1-1/p}\right\} \geq c' \tag{2.2.11}$$

for some constants $c, c' > 0$. This allows us to recover the second term in the minimum in bound (2.2.1). Finally, in the case when $c_1\frac{\sigma_\xi m^{3/2}}{\sqrt{n}} > 1$, the minimax lower bound becomes a constant (and the proof is based on a simplified version of the above argument that could be done for $r = 1$). This completes the proof of bound (2.2.1) for Schatten p-norms.

The proof of bound (2.2.2) for the Hellinger distance is similar. In the case $r \geq 2$, we will use a "well separated" set of density matrices $\mathcal{S}'_p \subset \mathcal{S}_{r,m}$ for $p = 1$ constructed above. We still use $\kappa := c_1\frac{\sigma_\xi m^{3/2}(r-1)}{\sqrt{n}}$ assuming first that $\kappa \in (0,1)$. For $S_{Q_1}, S_{Q_2} \in \mathcal{S}'_p$ with $Q_1 \neq Q_2$, it follows by a simple computation and using bound (1.3.10) that, for some $c'' > 0$,

$$\begin{aligned} H^2(S_{Q_1}, S_{Q_2}) &= \kappa H^2\left(\frac{Q_1}{r-1}, \frac{Q_2}{r-1}\right) \\ &\geq \frac{1}{4}\frac{\kappa}{(r-1)^2}\|Q_1 - Q_2\|_1^2 \geq \frac{(c')^2}{4}\kappa \geq c''\frac{\sigma_\xi m^{3/2}(r-1)}{\sqrt{n}}. \end{aligned}$$

Repeating the argument based on Proposition 1 yields bound (2.2.2) in the case when $\kappa = c_1\frac{\sigma_\xi m^{3/2}(r-1)}{\sqrt{n}} < 1$, and in the opposite case it is easy to see that the lower bound is a constant.

Finally, bound (2.2.3) for the Kullback–Leibler divergence follows from (2.2.2) and the inequality $K(\rho\|\hat\rho) \geq H^2(\hat\rho, \rho)$ (see inequality 1.3.10).

□

Next we state similar results in the case of trace regression model with bounded response (see Assumption 3). Denote by $\mathcal{P}_{r,m}(\bar{U})$ the class of all distributions P of (X,Y) such that Assumption 3 holds for some $\bar{U}$ and $\mathbb{E}(Y|X) = \langle \rho_P, X\rangle$ for some $\rho_P \in \mathcal{S}_{r,m}$. Given P, $\mathbb{P}_P$ denotes the corresponding probability measure (such that $(X_1,Y_1),\dots,(X_n,Y_n)$ are i.i.d. copies of (X,Y) sampled from P).

Theorem 5. *Suppose $\bar{U} \geq 2U$. For all $p \in [1,+\infty]$, there exist absolute constants $c,c' > 0$ such that the following bounds hold:*

$$\inf_{\hat{\rho}} \sup_{P\in\mathcal{P}_{r,m}(\bar{U})} \mathbb{P}_P\Biggl\{\|\hat{\rho}-\rho_P\|_p \geq c\biggl(\frac{\bar{U}m^{\frac{3}{2}}r^{1/p}}{\sqrt{n}} \bigwedge \biggl(\frac{\bar{U}m^{3/2}}{\sqrt{n}}\biggr)^{1-\frac{1}{p}} \bigwedge 1\biggr)\Biggr\} \geq c', \tag{2.2.12}$$

$$\inf_{\hat{\rho}} \sup_{P\in\mathcal{P}_{r,m}(\bar{U})} \mathbb{P}_P\Biggl\{H^2(\hat{\rho},\rho_P) \geq c\biggl(\frac{\bar{U}m^{\frac{3}{2}}r}{\sqrt{n}} \bigwedge 1\biggr)\Biggr\} \geq c', \tag{2.2.13}$$

and

$$\inf_{\hat{\rho}} \sup_{P\in\mathcal{P}_{r,m}(\bar{U})} \mathbb{P}_P\Biggl\{K(\rho_P\|\hat{\rho}) \geq c\biggl(\frac{\bar{U}m^{\frac{3}{2}}r}{\sqrt{n}} \bigwedge 1\biggr)\Biggr\} \geq c', \tag{2.2.14}$$

where $\inf_{\hat{\rho}}$ denotes the infimum over all estimators $\hat{\rho}$ in $\mathcal{S}_m$ based on the data $(X_1,Y_1),\dots,(X_n,Y_n)$.

Proof. The proof relies on an idea already used in a context of matrix completion by [57] (see their Theorem 7). We need the same family $\mathcal{S}'_p \subset \mathcal{S}_{r,m}$ of "well separated" density matrices of rank r as in the proof of Theorem 4. For a density matrix ρ, let (X,Y) be a random couple such that X is sampled from the uniform distribution Π in $\mathcal{E}$ and, conditionally on X, Y takes value $+\bar{U}$ with probability $p_\rho(X) := \frac{1}{2} + \frac{\langle\rho,X\rangle}{2\bar{U}}$ and value $-\bar{U}$ with probability $q_\rho(X) := \frac{1}{2} - \frac{\langle\rho,X\rangle}{2\bar{U}}$. Since $\bar{U} \geq 2U$ and $|\langle\rho,X\rangle| \leq \|\rho\|_1\|X\|_\infty \leq U$, we have $p_\rho(X), q_\rho(X) \in [1/4, 3/4]$ (so, they are bounded away from 0 and from 1). Clearly, $\mathbb{E}_\rho(Y|X) = \langle\rho,X\rangle$. Let P_ρ denote the distribution of such a couple and $\mathbb{P}_\rho$ denote the corresponding distribution of the data $(X_1,Y_1),\dots,(X_n,Y_n)$. Then, for all $\rho\in\mathcal{S}_{r,m}$, $P_\rho \in \mathcal{P}_{r,m}(\bar{U})$. The only difference with the proof of Theorem 4 is in the bound on Kullback-Leibler divergence $K(\mathbb{P}_{\rho_1}\|\mathbb{P}_{\rho_2})$ (see Equation 2.2.5). It is easy to see that

$$K(\mathbb{P}_{\rho_1}\|\mathbb{P}_{\rho_2}) = n\mathbb{E}\left(p_{\rho_1}(X)\log\frac{p_{\rho_1}(X)}{p_{\rho_2}(X)} + q_{\rho_1}(X)\log\frac{q_{\rho_1}(X)}{q_{\rho_2}(X)}\right). \tag{2.2.15}$$

The following simple inequality will be used: for all $a, b \in [1/4, 3/4]$,

$$a \log \frac{a}{b} + (1-a) \log \frac{1-a}{1-b} \leq 12(a-b)^2.$$

It implies that

$$K(\mathbb{P}_{\rho_1} \| \mathbb{P}_{\rho_2}) \leq 3n\mathbb{E}\frac{\langle \rho_1 - \rho_2, X\rangle^2}{\bar{U}^2} \leq \frac{3n}{\bar{U}^2}\|\rho_1 - \rho_2\|^2_{L_2(\Pi)}.$$

This bound is used instead of identity (2.2.5) from the proof of Theorem 4. The rest of the proof is the same. □

Note that the proof requires the possible range $[-\bar{U}, \bar{U}]$ of response variable Y to be larger than the possible range $[-U, U]$ of Fourier coefficients $\langle \rho, E_j \rangle, j = 1, \ldots, m^2$. This is not the case for standard QST model described in the introduction (see also the example of Pauli measurements) and it is of interest to prove a version of minimax lower bounds without this constraint, including the case when $\bar{U} = U$. The following theorem is a result in this direction.

Theorem 6. *Suppose Assumption 1 is satisfied and, moreover, for some constant $\gamma \in (0, 1)$,*

$$\Big|\mathrm{tr}(E_k)\Big| \leq (1-\gamma)Um, \ \ k = 1, \ldots, m^2. \tag{2.2.16}$$

Then, for all $p \in [1, +\infty]$, there exist constants $c_\gamma, c'_\gamma > 0$ such that the following bounds hold:

$$\inf_{\hat{\rho}} \sup_{P \in \mathcal{P}_{r,m}(U)} \mathbb{P}_P\bigg\{\|\hat{\rho} - \rho_P\|_p \geq c_\gamma \bigg(\frac{Um^{\frac{3}{2}}r^{1/p}}{\sqrt{n}} \bigwedge \Big(\frac{Um^{3/2}}{\sqrt{n}}\Big)^{1-\frac{1}{p}} \bigwedge 1\bigg)\bigg\} \geq c'_\gamma, \tag{2.2.17}$$

$$\inf_{\hat{\rho}} \sup_{P \in \mathcal{P}_{r,m}(U)} \mathbb{P}_P\bigg\{H^2(\hat{\rho}, \rho_P) \geq c_\gamma \bigg(\frac{Um^{\frac{3}{2}}r}{\sqrt{n}} \bigwedge 1\bigg)\bigg\} \geq c'_\gamma, \tag{2.2.18}$$

and

$$\inf_{\hat{\rho}} \sup_{P \in \mathcal{P}_{r,m}(U)} \mathbb{P}_P\bigg\{K(\rho_P \| \hat{\rho}) \geq c_\gamma \bigg(\frac{Um^{\frac{3}{2}}r}{\sqrt{n}} \bigwedge 1\bigg)\bigg\} \geq c'_\gamma, \tag{2.2.19}$$

where $\inf_{\hat{\rho}}$ denotes the infimum over all estimators $\hat{\rho}$ in $\mathcal{S}_m$ based on the data $(X_1, Y_1), \ldots, (X_n, Y_n)$.

Proof. The proof is based on the following lemma:

Lemma 5. *Suppose assumption (2.2.16) holds. Let K be a sufficiently large absolute constant (to be chosen later) and let m satisfy the condition $K\frac{\log m}{\sqrt{m}} \le \frac{\gamma}{2}$ (which means that $m \ge A_\gamma$ for some constant A_γ). Then there exists $v \in \mathbb{C}^m$ with $\|v\| = 1$ such that*

$$\Big|\langle E_k v, v\rangle\Big| \le (1-\gamma/2)U, k = 1, \dots, m^2. \tag{2.2.20}$$

Proof. We will prove this fact by a probabilistic argument. Namely, set

$$v := m^{-1/2}(\varepsilon_1, \dots, \varepsilon_m),$$

where $\varepsilon_j = \pm 1$. We will show that there is a random choice of "signs" ε_j such that (2.2.20) holds. Assume that $\varepsilon_j, j = 1, \dots, m$ are i.i.d. and take values ± 1 with probability 1/2 each. Let $E_k := (a_{ij}^{(k)})_{i,j=1,\dots,m}$. For simplicity, assume that $(a_{ij}^{(k)})_{i,j=1,\dots,m}$ is a symmetric real matrix (in the complex case, the proof can be easily modified). We have

$$\langle E_k v, v\rangle = \frac{1}{m}\sum_{i=1}^m a_{ii}^{(k)}\varepsilon_i^2 + \frac{1}{m}\sum_{i\ne j} a_{ij}^{(k)}\varepsilon_i\varepsilon_j = \frac{\operatorname{tr}(E_k)}{m} + \frac{1}{m}\sum_{i\ne j} a_{ij}^{(k)}\varepsilon_i\varepsilon_j.$$

It is well known that

$$\operatorname{Var}\biggl(\sum_{i\ne j} a_{ij}^{(k)}\varepsilon_i\varepsilon_j\biggr) = \mathbb{E}\biggl(\sum_{i\ne j} a_{ij}^{(k)}\varepsilon_i\varepsilon_j\biggr)^2 = 2\sum_{i\ne j}\Bigl(a_{ij}^{(k)}\Bigr)^2 \le 2\sum_{i,j}\Bigl(a_{ij}^{(k)}\Bigr)^2 = 2\|E_k\|_2^2 = 2.$$

Moreover, it follows from exponential inequalities for Rademacher chaos (see, e.g., Corollary 3.2.6 in [27]) that for some absolute constant $K > 0$ and for all $t > 0$, with probability at least $1 - e^{-t}$

$$\left|\langle E_k v, v\rangle - \frac{\operatorname{tr}(E_k)}{m}\right| = \left|\frac{1}{m}\sum_{i\ne j} a_{ij}^{(k)}\varepsilon_i\varepsilon_j\right| \le \frac{Kt}{m}.$$

Taking $t = 2\log m$ and using the union bound, we conclude that with probability at least $1 - me^{-2\log m} = 1 - \frac{1}{m} > 0$,

$$\max_{1\le k\le m^2}\left|\langle E_k v, v\rangle - \frac{\operatorname{tr}(E_k)}{m}\right| \le \frac{K\log m}{m} \le \frac{K\log m}{\sqrt{m}}U \le \frac{\gamma}{2}U,$$

where we also used the fact that $U \geq m^{-1/2}$. Thus, there exists a choice of signs ε_j such that

$$\max_{1 \leq k \leq m^2} \left| \langle E_k v, v \rangle \right| \leq \max_{1 \leq k \leq m} \left| \frac{\mathrm{tr}(E_k)}{m} \right| + \frac{\gamma}{2} U,$$

which, under condition (2.2.16), implies (2.2.20). □

We set $e_1 := v$ (where v is the unit vector introduced in Lemma 5) and construct an orthonormal basis $e_1, \ldots, e_m$. Assume that matrices S_Q defined by (2.2.6) represent linear transformations in basis $e_1, \ldots, e_m$. Then we have

$$\langle S_Q, E_k \rangle = (1-\kappa) \langle E_k v, v \rangle + \frac{\kappa}{r-1} \langle Q, E_k \rangle.$$

Therefore,

$$\left| \langle S_Q, E_k \rangle \right| \leq (1-\kappa) \left| \langle E_k v, v \rangle \right| + \frac{\kappa}{r-1} \|E_k\|_\infty \|Q\|_1 \leq (1-\kappa)(1-\gamma/2)U + \kappa U = (1-(1-\kappa)(\gamma/2))U.$$

Assuming that $\kappa \leq 1/2$, we get

$$\left| \langle S_Q, E_k \rangle \right| \leq (1-\gamma/4)U, \ k = 1, \ldots, m^2. \tag{2.2.21}$$

The rest of the proof becomes similar to the proof of Theorem 5 (with $\bar{U} = U$). Namely, bound (2.2.21) implies that, for $\rho = S_Q$ and X being sampled from the orthonormal basis $\{E_1, \ldots, E_{m^2}\}$, probabilities $p_\rho(X)$ and $q_\rho(X)$ are bounded away from 0 and from 1 : $p_\rho(X), q_\rho(X) \in [\gamma/8, 1-\gamma/8]$. This allows us to complete the argument of the proof of Theorem 5. □

Theorem 6 does not apply directly to the Pauli basis since condition (2.2.16) fails in this case. Indeed, by the definition of Pauli basis, $U = m^{-1/2}$ and $\mathrm{tr}(E_1) = \sqrt{m} = Um > (1-\gamma)Um$. Note also that $\mathrm{tr}(E_j) = 0, j = 2, \ldots, m^2$. Thus, for Pauli basis, E_1 is the only matrix for which condition (2.2.16) fails. However, for this matrix $\langle \rho, E_1 \rangle = m^{-1/2} \mathrm{tr}(\rho) = m^{-1/2} = U$ for all density matrices $\rho \in \mathcal{S}_m$. This immediately implies that $p_\rho(E_1) = 1$ and $q_\rho(E_1) = 0$ for all $\rho \in \mathcal{S}_m$ and, as a result, the value $X = E_1$ does not have an impact on the computation of Kullback-Leibler divergence in (2.2.15). For the rest of the matrices in the Pauli basis, condition (2.2.16) holds implying also bound (2.2.20). Therefore, if $X \neq E_1$, we still have that, for $\rho = S_Q$, $p_\rho(X), q_\rho(X) \in [\gamma/8, 1-\gamma/8]$, and the proof of Theorem 5 can be completed in this case, too. Note also that, given X sampled from the Pauli basis, the binary random variable Y taking values $\pm U = \pm \frac{1}{\sqrt{m}}$ with probabilities $p_\rho(X)$ and $q_\rho(X)$, respectively (this is exactly the random variable used in the construction of the proof of Theorem 5) coincides with an outcome of a Pauli measurement for the system prepared in state ρ. These considerations yield the following minimax lower bounds for Pauli measurements.

Theorem 7. *Let $\{E_1, \dots, E_{m^2}\}$ be the Pauli basis in the space $\mathbb{H}_m$ of $m \times m$ Hermitian matrices and let $X_1, \dots, X_n$ be i.i.d. random variables sampled from the uniform distribution in $\{E_1, \dots, E_{m^2}\}$. Let $Y_1, \dots, Y_n$ be outcomes of measurements of observables $X_1, \dots, X_n$ for the system being identically prepared n times in state ρ. The corresponding distribution of the data $(X_1, Y_1), \dots, (X_n, Y_n)$ will be denoted by $\mathbb{P}_\rho$. Then, for all $p \in [1, +\infty]$, there exist constants $c, c' > 0$ such that the following bounds hold:*

$$\inf_{\hat\rho} \sup_{\rho \in \mathcal{S}_{r,m}} \mathbb{P}_\rho \left\{ \|\hat\rho - \rho\|_p \geq c \left(\frac{m r^{1/p}}{\sqrt{n}} \bigwedge \left(\frac{m}{\sqrt{n}} \right)^{1-\frac{1}{p}} \bigwedge 1 \right) \right\} \geq c', \tag{2.2.22}$$

$$\inf_{\hat\rho} \sup_{\rho \in \mathcal{S}_{r,m}} \mathbb{P}_\rho \left\{ H^2(\hat\rho, \rho) \geq c \left(\frac{mr}{\sqrt{n}} \bigwedge 1 \right) \right\} \geq c', \tag{2.2.23}$$

and

$$\inf_{\hat\rho} \sup_{\rho \in \mathcal{S}_{r,m}} \mathbb{P}_\rho \left\{ K(\rho \| \hat\rho) \geq c \left(\frac{mr}{\sqrt{n}} \bigwedge 1 \right) \right\} \geq c', \tag{2.2.24}$$

where $\inf_{\hat\rho}$ denotes the infimum over all estimators $\hat\rho$ in $\mathcal{S}_m$ based on the data $(X_1, Y_1), \dots, (X_n, Y_n)$.

Remark 1. *Minimax lower bounds on nuclear norm error of density matrix estimation close to bound (2.4.7) for $p = 1$ (but for a somewhat different "estimation protocol" and stated in a different form) were obtained earlier in [30]. This paper also contains upper bounds on the errors of matrix LASSO and Dantzig selector estimators in the nuclear norm matching the lower bounds up to log-factors.*

Remark 2. *It is easy to see that, if constant $\gamma \in (0,1)$ is small enough (namely, $\gamma < 1 - \frac{1}{\sqrt{2}}$), then, in an arbitrary orthonormal basis $\{E_1, \dots, E_{m^2}\}$, there is at most one matrix E_j such that $|\mathrm{tr}(E_j)| > (1-\gamma)Um$. Indeed, note that $\mathrm{tr}(E_j) = \langle E_j, I_m \rangle$. Since*

$$\sum_{j=1}^{m^2} \langle E_j, I_m \rangle^2 = \|I_m\|_2^2 = m$$

and $U^2 m \geq 1$, we have

$$\mathrm{card}\Big(\Big\{j : |\langle E_j, I_m \rangle| > (1-\gamma)Um\Big\}\Big) \leq \frac{1}{(1-\gamma)^2 U^2 m^2} \sum_{j=1}^{m^2} \langle E_j, I_m \rangle^2$$

$$\leq \frac{m}{(1-\gamma)^2 U^2 m^2} = \frac{1}{(1-\gamma)^2 U^2 m} \leq \frac{1}{(1-\gamma)^2} < 2,$$

provided that $\gamma < 1 - \frac{1}{\sqrt{2}}$.

Remark 3. *It will be shown in Section 2.3.4 that the minimax rates of theorems 4, 5, 6 and 7 are attained up to logarithmic factors for the von Neumann entropy penalized least squares estimator.*

Remark 4. *Similar minimax lower bounds could be proved in certain classes of "nearly low rank" density matrices. Consider, for instance, the following class*

$$B_q(d; m) := \left\{ S \in \mathcal{S}_m : \sum_{j=1}^{m} |\lambda_j(S)|^q \leq d \right\} \tag{2.2.25}$$

for some $d > 0$ and $q \in [0,1]$, where $\lambda_1(S) \geq \dots \geq \lambda_m(S)$ denote the eigenvalues of S. This set consists of density matrices with the eigenvalues decaying at a certain rate (nearly low rank case) and, for $q = 0$, $d = r$ it coincides with $\mathcal{S}_{r,m}$. It turns out that minimax lower bounds of theorems 4 and 5 hold for the class $B_q(d;m)$ (instead of $\mathcal{S}_{r,m}$) with r replaced by

$$\bar{r} := \bar{r}(\tau, d, m, q) = d\tau^{-q} \wedge m,$$

where $\tau := \frac{\sigma_\xi m^{3/2}}{\sqrt{n}}$ in the case of trace regression with Gaussian noise and $\tau := \frac{\bar{U} m^{3/2}}{\sqrt{n}}$ in the case of trace regression with bounded response. These minimax bounds are attained *up to logarithmic factors for a slightly modified von Neumann entropy penalized least squares estimator.*

Note that, for $\rho \in B_q(d,m)$ with eigenvalues $\lambda_1(\rho) \geq \cdots \geq \lambda_m(\rho)$, we have $\lambda_j(\rho) \leq \frac{d^{1/q}}{j^{1/q}}, j = 1, \ldots, m$. Therefore, for $j \geq \bar{r}$, $\lambda_j(\rho) \leq \tau$. Note also that τ characterizes the minimax rate of estimation of $\rho \in \mathcal{S}_{r,m}$ in the operator norm for any value of the rank r (see bound (2.2.1) for $p = +\infty$; the corresponding upper bound also holds for the least squares estimator up to a logarithmic factor, see [101]). Roughly speaking, τ is a threshold below which the estimation of eigenvalues $\lambda_j(\rho)$ becomes impossible and $\bar{r}$ can be viewed as an "effective rank" of nearly low rank density matrices in the class $B_q(d,m)$.

2.3 Least squares estimator: low rank oracle inequalities and its optimality

Recall that the the least squares estimator penalized by the so called von Neumann entropy is defined as

$$\tilde{\rho}^{\varepsilon} := \underset{S \in \mathcal{S}_m}{\arg\min} \Big[\frac{1}{n} \sum_{j=1}^{n} \big(Y_j - \langle S, X_j \rangle \big)^2 + \varepsilon \mathrm{tr}\big(S \log S \big) \Big]. \tag{2.3.1}$$

Note that when $\varepsilon = 0$, it reduces to the standard least squares estimator.

The goal of this section is to study optimality properties of von Neumann entropy penalized least squares estimator $\tilde{\rho}^{\varepsilon}$ defined by (2.3.1). In particular, we establish oracle inequalities for such estimators in the cases of trace regression with bounded response (Subsection 2.3.2) and trace regression with Gaussian noise (Subsection 2.3.3), and prove upper bounds on their estimation errors measured by Schatten p-norm distances for $p \in [1,2]$ and also by Hellinger and Kullback-Leibler distances (Subsection 2.3.4).

2.3.1 Preliminaries and matrix Bernstein inequalities

The following well known *interpolation inequality* for Schatten p-norms will be used to extend the bounds proved for some values of p to the whole range of its values. It easily follows from similar bounds for ℓ_p-spaces.

Lemma 6 (Interpolation inequality). *For $1 \le p < q < r \le \infty$, and let $\mu \in [0,1]$ be such that*

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

Then, for all $A \in \mathbb{H}_m$,

$$\|A\|_q \le \|A\|_p^{\mu} \|A\|_r^{1-\mu}.$$

Given $A \in \mathbb{H}_m$, define a function $f_A : \mathbb{H}_m \mapsto \mathbb{R} : f_A(x) := \langle A, x\rangle, x \in \mathbb{H}_m$. For a given random variable X in $\mathbb{H}_m$ with a distribution Π, we have $\|f_A\|_{L_2(\Pi)}^2 = \mathbb{E} f_A^2(X) = \mathbb{E}\langle A, X\rangle^2$. Sometimes, with a minor abuse of notation (see also Section 1.3.1), we might write $\|A\|_{L_2(\Pi)}^2 = \int_{\mathbb{H}_m} \langle A, x\rangle^2 \Pi(dx) = \|f_A\|_{L_2(\Pi)}^2$. Remember that Π is typically the uniform distribution in an orthonormal basis $\mathcal{E} = \{E_1, \dots, E_{m^2}\} \subset \mathbb{H}_m$, implying that

$$\|f_A\|_{L_2(\Pi)}^2 = \|A\|_{L_2(\Pi)}^2 = m^{-2}\|A\|_2^2,$$

so, the $L_2(\Pi)$-norm is just a rescaled Hilbert–Schmidt norm.

Consider $A \in \mathbb{H}_m$ with spectral representation $A = \sum_{j=1}^{m'} \lambda_j P_j$, $m' \le m$ with distinct non-zero eigenvalues λ_j. Denote by $\operatorname{sign}(A) := \sum_{j=1}^{m'} \operatorname{sign}(\lambda_j) P_j$ and by $\operatorname{supp}(A)$ the linear span of the images of projectors $P_j, j = 1, \dots, m'$ (the subspace $\operatorname{supp}(A) \subset \mathbb{C}^m$ will be called *the support* of A).

Given a subspace $L \subset \mathbb{C}^m$, $L^\perp$ denotes the orthogonal complement of L and P_L denotes the orthogonal projection onto L. Let $\mathcal{P}_L, \mathcal{P}_L^\perp$ be orthogonal projection operators in the space $\mathbb{H}_m$ (equipped with the Hilbert–Schmidt inner product), defined as follows:

$$\mathcal{P}_L^\perp(A) = P_{L^\perp} A P_{L^\perp}, \quad \mathcal{P}_L(A) = A - P_{L^\perp} A P_{L^\perp}.$$

These two operators split any Hermitian matrix A into two orthogonal parts, $\mathcal{P}_L(A)$ and $\mathcal{P}_L^{\perp}(A)$, the first one being of rank at most $2\mathrm{dim}(L)$.

For a convex function $f : \mathbb{H}_m \mapsto \mathbb{R}$, $\partial f(A)$ denotes the subdifferential of f at the point $A \in \mathbb{H}_m$. It is well known that

$$\partial\|A\|_1 = \Big\{\mathrm{sign}(A) + \mathcal{P}_L^{\perp}(M) : M \in \mathbb{H}_m, \|M\|_\infty \leq 1\Big\}, \tag{2.3.2}$$

where $L = \mathrm{supp}(A)$ (see [52], p. 240, [99] and references therein).

Non-commutative (matrix) versions of Bernstein inequality will be used frequently in this Chapter. The most common version is stated (in a convenient form for our applications) in the following lemma.

Lemma 7. *Let $X, X_1, \dots, X_n \in \mathbb{H}_m$ be i.i.d. random matrices with $\mathbb{E}X = 0$, $\sigma_X^2 := \|\mathbb{E}X^2\|_\infty$ and $\|X\|_\infty \leq U$ a.s. for some $U > 0$. Then, for all $t \geq 0$ with probability at least $1 - e^{-t}$,*

$$\left\|\frac{1}{n}\sum_{j=1}^n X_j\right\|_\infty \leq 2\left[\sigma_X\sqrt{\frac{t+\log(2m)}{n}} \bigvee U\frac{t+\log(2m)}{n}\right].$$

The proof of such bounds could be found, e.g., in [90].A simple consequence of the inequality of Lemma 7 is the following expectation bound:

$$\mathbb{E}\left\|\frac{1}{n}\sum_{j=1}^n X_j\right\|_\infty \lesssim \left[\sigma_X\sqrt{\frac{\log(2m)}{n}} \bigvee U\frac{\log(2m)}{n}\right].$$

It follows from the exponential bound by integrating the tail probabilities.

Other versions on matrix Bernstein type inequalities for not necessarily bounded random matrices will be also used in what follows and they could be found in [52], [54] and [53].

Lemma 8. *Let $X, X_1, \dots, X_n$ be i.i.d. random matrices in $\mathbb{H}_m$ with $\mathbb{E}X = 0$. Suppose that, for some $\alpha \geq 1$, $U^{(\alpha)} := 2\big\|\|X\|_\infty\big\|_{\psi_\alpha} < +\infty$.* [1] *Let $\sigma_X^2 := \|\mathbb{E}X^2\|_\infty$. Then, for all $t > 0$ with probability at least $1 - e^{-t}$,*

$$\left\|\frac{X_1 + \dots + X_n}{n}\right\|_\infty \leq C\left[\sigma_X\sqrt{\frac{t+\log(2m)}{n}} \bigvee U^{(\alpha)}\log^{1/\alpha}\left(\frac{U^{(\alpha)}}{\sigma}\right)\frac{t+\log(2m)}{n}\right].$$

[1] *Remember that $\|\cdot\|_{\psi_\alpha}$ denotes the ψ_α Orlicz norm in the space of random variables defined as*

2.3.2 Oracle inequalities for trace regression with bounded response

In this subsection, we prove a *sharp low rank oracle inequality* for estimator $\tilde{\rho}^{\varepsilon}$ defined by (2.3.1). It is done in the case of trace regression model with bounded response (that is, under Assumption 3 in Section 2.1). The results of this type show some form of optimality of the estimation method, namely, that the estimator provides an optimal trade-off between the "approximation error" of the target density matrix by a low rank "oracle" and the "estimation error" of the "oracle" that is proportional to its rank. Sharp oracle inequalities (in which the leading constant in front of the "approximation error" is equal to 1, so that the bound mimics precisely the approximation by the oracle) are usually harder to prove. In the case of low rank matrix completion, the first result of this type was proved by [57] for a modified least squares estimator with nuclear norm penalty. A version of such inequality for empirical risk minimization with nuclear norm penalty (that includes matrix LASSO) was first proved by [55]. Low rank oracle inequalities for von Neumann entropy penalized least squares method with the leading constant larger than 1 were proved by [53], see also [31]. The main result of this section refines these previous bounds by proving a sharp oracle inequality, improving the logarithmic factors and removing superfluous assumptions, but also by establishing the inequality in the whole range of values of regularization parameter $\varepsilon \geq 0$ (including the value $\varepsilon = 0$, for which $\tilde{\rho}^{\varepsilon}$ coincides with the least squares estimator $\hat{\rho}$, see Section 1.3.3). In addition to this, for a special choice of regularization parameter ε, the theorem below also provides an upper bound on the Kullback-Leibler error $K(\rho \| \tilde{\rho}^{\varepsilon})$ of $\tilde{\rho}^{\varepsilon}$ that matches the minimax lower bound (2.2.14) up to log-factors (and "second order terms"). It turns out that, for this choice of ε, the

follows (see Section 1.4):

$$\|\eta\|_{\psi_\alpha} := \inf\Big\{c > 0 : \mathbb{E}\exp\Big\{\frac{|\eta|^\alpha}{c^\alpha}\Big\} \leq 2\Big\}.$$

estimator satisfies exactly the same low rank oracle inequality as the best inequalities known for LASSO estimator and minimax optimal error rates are attained for $\tilde{\rho}^{\varepsilon}$ also with respect to Hellinger distance and Schatten p-norm distances for all $p \in [1,2]$ (see Section 2.3.4). For simplicity, it will be assumed that constants U in Assumption 1 and $\bar{U}$ in Assumption 3 coincide (in the upper bounds, one can always replace U and $\bar{U}$ by $U \vee \bar{U}$).

Theorem 8. *Suppose Assumption 3 holds with constant $\bar{U} = U$ and let $\varepsilon \in [0,1]$. Then, there exists a constant $C > 0$ such that for all $t \geq 1$ with probability at least $1 - e^{-t}$*

$$\|f_{\tilde{\rho}^{\varepsilon}} - f_{\rho}\|^2_{L_2(\Pi)} \leq \inf_{S \in \mathcal{S}_m} \biggl[\|f_S - f_\rho\|^2_{L_2(\Pi)} + C \biggl(\mathrm{rank}(S) m^2 \varepsilon^2 \log^2(mn) \\ + U^2 \tfrac{\mathrm{rank}(S) m \log(2m)}{n} + U^2 \tfrac{t + \log\log_2(2n)}{n} \biggr) \biggr]. \tag{2.3.3}$$

In particular, this implies that

$$\|f_{\tilde{\rho}^{\varepsilon}} - f_{\rho}\|^2_{L_2(\Pi)} \leq C \biggl[\mathrm{rank}(\rho) m^2 \varepsilon^2 \log^2(mn) \\ + U^2 \tfrac{\mathrm{rank}(\rho) m \log(2m)}{n} + U^2 \tfrac{t + \log\log_2(2n)}{n} \biggr]. \tag{2.3.4}$$

Moreover, if

$$\varepsilon := \frac{1}{\log(mn)} \biggl[U \sqrt{\frac{\log(2m)}{nm}} \bigvee U^2 \frac{\log(2m)}{n} \biggr],$$

then, with some constant C and with probability at least $1 - e^{-t}$

$$\|f_{\tilde{\rho}^{\varepsilon}} - f_{\rho}\|^2_{L_2(\Pi)} \leq C \biggl[U^2 \tfrac{\mathrm{rank}(\rho) m \log(2m)}{n} \biggl(1 \bigvee U^2 \tfrac{m \log(2m)}{n} \biggr) \\ + U^2 \tfrac{t + \log\log_2(2n)}{n} \biggr] \tag{2.3.5}$$

and

$$K(\rho \| \tilde{\rho}^{\varepsilon}) \leq CU \biggl[\tfrac{\mathrm{rank}(\rho) m^{3/2} \sqrt{\log(2m)} \log(mn)}{\sqrt{n}} \biggl(1 \bigvee U \sqrt{\tfrac{m \log(2m)}{n}} \biggr) \\ + \sqrt{\tfrac{m}{n}} \tfrac{(t + \log\log_2(2n)) \log(mn)}{\sqrt{\log(2m)}} \biggr]. \tag{2.3.6}$$

Proof. The following notations will be used in the proof. Let $\ell(y,u) := (u-y)^2, y, u \in \mathbb{R}$ be the quadratic loss function. For $f : \mathbb{H}_m \mapsto \mathbb{R}$, denote

$$(\ell \bullet f)(x,y) = (f(x)-y)^2, \quad (\ell' \bullet f)(x,y) = 2(f(x)-y)$$

and

$$P(\ell \bullet f) = \mathbb{E}(Y - f(X))^2, \quad P_n(\ell \bullet f) = n^{-1}\sum_{j=1}^{n}(Y_j - f(X_j))^2.$$

For $A \in \mathbb{H}_m$, let $f_A(x) = \langle A, x\rangle, x \in \mathbb{H}_m$. Since for density matrices $S \in \mathcal{S}_m$, $\|S\|_1 = \mathrm{tr}(S) = 1$, the estimator $\tilde{\rho} = \tilde{\rho}^{\varepsilon}$ can be equivalently defined by the following convex optimization problem:

$$\tilde{\rho} = \mathrm{argmin}_{S \in \mathcal{S}_m} L_n(S), \quad L_n(S) := \Big[P_n(\ell \bullet f_S) + \varepsilon \mathrm{tr}(S \log S) + \bar{\varepsilon}\|S\|_1\Big],$$

for an arbitrary $\bar{\varepsilon} > 0$.

The following lemma will be crucial in the proofs of Theorem 8 as well Theorem 9 in the following subsection. Note that it does not rely on Assumption 3, only Assumptions 1 and 2 are needed.

Lemma 9. *Suppose Assumptions 1 and 2 hold. Let $\delta \in (0,1)$ and $S := (1-\delta)S' + \delta\frac{I_m}{m}$, where $S' \in \mathcal{S}_m$, $\mathrm{rank}(S') = r$ and I_m is the $m \times m$ identity matrix. Then the following bound holds:*

$$\begin{aligned}\|f_{\tilde{\rho}} - f_{\rho}\|^2_{L_2(\Pi)} + \tfrac{1}{2}\|f_{\tilde{\rho}} - f_S\|^2_{L_2(\Pi)} + \varepsilon K(\tilde{\rho}; S) + \bar{\varepsilon}\Big\|\mathcal{P}_L^{\perp}(\tilde{\rho})\Big\|_1 \\ \leq \|f_S - f_{\rho}\|^2_{L_2(\Pi)} + rm^2\varepsilon^2\log^2(m/\delta) + rm^2\bar{\varepsilon}^2 \\ +4\bar{\varepsilon}\delta + (P - P_n)(\ell' \bullet f_{\tilde{\rho}})(f_{\tilde{\rho}} - f_S).\end{aligned} \tag{2.3.7}$$

Lemma 9 will be often used together with the following simple bound:

$$\begin{aligned}\|f_S - f_{\rho}\|^2_{L_2(\Pi)} = \tfrac{1}{m^2}\|S - \rho\|_2^2 \leq \\ \tfrac{1}{m^2}\|S' - \rho\|_2^2 + \tfrac{2}{m^2}\|S' - \rho\|_2\|S' - S\|_2 + \tfrac{1}{m^2}\|S' - S\|_2^2 \\ \leq \|f_{S'} - f_{\rho}\|^2_{L_2(\Pi)} + \tfrac{8\delta}{m^2} + \tfrac{4\delta^2}{m^2} \leq \|f_{S'} - f_{\rho}\|^2_{L_2(\Pi)} + \tfrac{12\delta}{m^2}.\end{aligned} \tag{2.3.8}$$

Together, they imply that

$$\begin{aligned}\|f_{\tilde\rho}-f_\rho\|^2_{L_2(\Pi)}+\tfrac{1}{2}\|f_{\tilde\rho}-f_S\|^2_{L_2(\Pi)}+\varepsilon K(\tilde\rho;S)+\bar\varepsilon\left\|\mathcal{P}_L^\perp(\tilde\rho)\right\|_1 \\ \le \|f_{S'}-f_\rho\|^2_{L_2(\Pi)}+rm^2\varepsilon^2\log^2(m/\delta)+rm^2\bar\varepsilon^2 \qquad (2.3.9) \\ +4\bar\varepsilon\delta+\tfrac{12\delta}{m^2}+(P-P_n)(\ell'\bullet f_{\tilde\rho})(f_{\tilde\rho}-f_S).\end{aligned}$$

We will now give the proof of Lemma 9.

Proof. By standard necessary conditions of extremum in convex problems, we get that, for all $S\in\mathcal{S}_m$ and for some $\tilde V\in\partial\|\tilde\rho\|_1$,

$$P_n(\ell'\bullet f_{\tilde\rho})(f_{\tilde\rho}-f_S)+\varepsilon\langle\log\tilde\rho,\tilde\rho-S\rangle+\bar\varepsilon\langle\tilde V,\tilde\rho-S\rangle\le 0$$

(see, e.g., [3], Chapter 2, Corollary 6; see also [52], pp. 198–199; for the computation of derivative of the function $\mathrm{tr}(S\log S)$, see Lemma 1 in [53]). Replacing in the left hand side P by P_n, we get

$$P(\ell'\bullet f_{\tilde\rho})(f_{\tilde\rho}-f_S)+\varepsilon\langle\log\tilde\rho,\tilde\rho-S\rangle+\bar\varepsilon\langle\tilde V,\tilde\rho-S\rangle\le(P-P_n)(\ell'\bullet f_{\tilde\rho})(f_{\tilde\rho}-f_S).$$

It is easy to check that for the quadratic loss

$$P(\ell'\bullet f_{\tilde\rho})(f_{\tilde\rho}-f_S)=P(\ell\bullet f_{\tilde\rho})-P(\ell\bullet f_S)+\|f_{\tilde\rho}-f_S\|^2_{L_2(\Pi)},$$

implying that

$$\begin{aligned}P(\ell\bullet f_{\tilde\rho})-P(\ell\bullet f_S)+\|f_{\tilde\rho}-f_S\|^2_{L_2(\Pi)}+\varepsilon\langle\log\tilde\rho,\tilde\rho-S\rangle+\bar\varepsilon\langle\tilde V,\tilde\rho-S\rangle \\ \le(P-P_n)(\ell'\bullet f_{\tilde\rho})(f_{\tilde\rho}-f_S).\end{aligned}$$

Also, for the quadratic loss,

$$P(\ell\bullet f)-P(\ell\bullet f_\rho)=\|f-f_\rho\|^2_{L_2(\Pi)}.$$

Therefore,

$$\|f_{\tilde\rho}-f_\rho\|^2_{L_2(\Pi)}+\|f_{\tilde\rho}-f_S\|^2_{L_2(\Pi)}+\varepsilon\langle\log\tilde\rho,\tilde\rho-S\rangle+\bar\varepsilon\langle\tilde V,\tilde\rho-S\rangle$$

$$\leq \|f_S - f_\rho\|_{L_2(\Pi)}^2 + (P - P_n)(\ell' \bullet f_{\tilde{\rho}})(f_{\tilde{\rho}} - f_S).$$

Recall that we have set $S = (1-\delta)S' + \delta \frac{I_m}{m}$, where $S' \in \mathcal{S}_m$, $\mathrm{rank}(S') = r$, $\delta \in (0,1)$. Clearly,

$$\left|\langle \tilde{V}, S - S' \rangle\right| \leq \|\tilde{V}\|_\infty \|S - S'\|_1 \leq \|S - S'\|_1 = \delta \left\| S' - \frac{I_m}{m} \right\|_1 \leq 2\delta,$$

where we used the fact that $\|\tilde{V}\|_\infty \leq 1$ for $\tilde{V} \in \partial \|\tilde{\rho}\|_1$. This implies

$$\|f_{\tilde{\rho}} - f_\rho\|_{L_2(\Pi)}^2 + \|f_{\tilde{\rho}} - f_S\|_{L_2(\Pi)}^2 + \varepsilon \langle \log \tilde{\rho}, \tilde{\rho} - S \rangle + \bar{\varepsilon} \langle \tilde{V}, \tilde{\rho} - S' \rangle \tag{2.3.10}$$
$$\leq \|f_S - f_\rho\|_{L_2(\Pi)}^2 + 2\bar{\varepsilon}\delta + (P - P_n)(\ell' \bullet f_{\tilde{\rho}})(f_{\tilde{\rho}} - f_S).$$

Recall formula (2.3.2) for the subdifferential of nuclear norm. Let $L = \mathrm{supp}(S')$. By the duality between the operator and nuclear norms, there exists $M \in \mathbb{H}_m$ with $\|M\|_\infty \leq 1$ such that

$$\langle \mathcal{P}_L^\perp(M), \tilde{\rho} - S' \rangle = \langle M, \mathcal{P}_L^\perp(\tilde{\rho} - S') \rangle = \left\| \mathcal{P}_L^\perp(\tilde{\rho} - S') \right\|_1 = \left\| \mathcal{P}_L^\perp(\tilde{\rho}) \right\|_1.$$

With $V = \mathrm{sign}(S') + \mathcal{P}_L^\perp(M) \in \partial \|S'\|_1$, by monotonicity of subdifferential, we get that

$$\langle \mathrm{sign}(S'), \tilde{\rho} - S' \rangle + \left\| \mathcal{P}_L^\perp(\tilde{\rho}) \right\|_1 = \langle V, \tilde{\rho} - S' \rangle \leq \langle \tilde{V}, \tilde{\rho} - S' \rangle. \tag{2.3.11}$$

In addition to this, we have

$$\langle \log \tilde{\rho}, \tilde{\rho} - S \rangle = \langle \log \tilde{\rho} - \log S, \tilde{\rho} - S \rangle + \langle \log S, \tilde{\rho} - S \rangle = K(\tilde{\rho}; S) + \langle \log S, \tilde{\rho} - S \rangle. \tag{2.3.12}$$

Substituting (2.3.11) and (2.3.12) into (2.3.10), we get

$$\|f_{\tilde{\rho}} - f_\rho\|_{L_2(\Pi)}^2 + \|f_{\tilde{\rho}} - f_S\|_{L_2(\Pi)}^2 + \varepsilon K(\tilde{\rho}; S) + \bar{\varepsilon} \left\| \mathcal{P}_L^\perp(\tilde{\rho}) \right\|_1$$
$$\leq \|f_S - f_\rho\|_{L_2(\Pi)}^2 + \varepsilon \langle \log S, S - \tilde{\rho} \rangle + \bar{\varepsilon} \langle \mathrm{sign}(S'), S' - \tilde{\rho} \rangle \tag{2.3.13}$$
$$+ 2\bar{\varepsilon}\delta + (P - P_n)(\ell' \bullet f_{\tilde{\rho}})(f_{\tilde{\rho}} - f_S).$$

The following bound on $\bar{\varepsilon} \langle \mathrm{sign}(S'), S' - \tilde{\rho} \rangle$ is straightforward:

$$\bar{\varepsilon} \langle \mathrm{sign}(S'), S' - \tilde{\rho} \rangle \leq \bar{\varepsilon} \langle \mathrm{sign}(S'), S - \tilde{\rho} \rangle + \bar{\varepsilon} \|\mathrm{sign}(S')\|_\infty \|S - S'\|_1$$
$$\leq \bar{\varepsilon} \|\mathrm{sign}(S')\|_2 \|S - \tilde{\rho}\|_2 + 2\bar{\varepsilon}\delta \leq \bar{\varepsilon} \sqrt{r} m \|f_S - f_{\tilde{\rho}}\|_{L_2(\Pi)} + 2\bar{\varepsilon}\delta \tag{2.3.14}$$
$$\leq r m^2 \bar{\varepsilon}^2 + \tfrac{1}{4} \|f_S - f_{\tilde{\rho}}\|_{L_2(\Pi)}^2 + 2\bar{\varepsilon}\delta.$$

A similar bound on $\varepsilon\langle \log S, S-\tilde{\rho}\rangle$ is only slightly more complicated. Suppose S' has the following spectral representation: $S' = \sum_{k=1}^{r} \lambda_k P_k$ with eigenvalues $\lambda_k \in (0,1]$ (repeated with their multiplicities) and one-dimensional orthogonal eigenprojectors P_k. We will extend $P_j, j=1,\dots,r$ to the complete orthogonal resolution of the identity $P_j, j=1,\dots,m$. Then

$$\log S = \log\Big((1-\delta)S' + \delta\frac{I_m}{m}\Big) = \sum_{j=1}^{r} \log\Big((1-\delta)\lambda_j + \delta/m\Big)P_j + \sum_{j=r+1}^{m} \log(\delta/m)P_j$$

$$= \sum_{j=1}^{r} \log\Big(1+(1-\delta)m\lambda_j/\delta\Big)P_j + \log(\delta/m)I_m$$

and

$$\langle \log S, S-\tilde{\rho}\rangle = \Big\langle \sum_{j=1}^{r} \log\Big(1+(1-\delta)m\lambda_j/\delta\Big)P_j, S-\tilde{\rho}\Big\rangle + \log(\delta/m)\langle I_m, S-\tilde{\rho}\rangle$$

$$= \Big\langle \sum_{j=1}^{r} \log\Big(1+(1-\delta)m\lambda_j/\delta\Big)P_j, S-\tilde{\rho}\Big\rangle$$

where we used the fact that $\langle I_m, S-\tilde{\rho}\rangle = \mathrm{tr}(S) - \mathrm{tr}(\tilde{\rho}) = 0$. Therefore,

$$\varepsilon\langle \log S, S-\tilde{\rho}\rangle \le \varepsilon\Big\|\textstyle\sum_{j=1}^{r} \log\Big(1+(1-\delta)m\lambda_j/\delta\Big)P_j\Big\|_2 \|S-\tilde{\rho}\|_2 \qquad (2.3.15)$$

$$= \varepsilon m\Big(\textstyle\sum_{j=1}^{r} \log^2\Big(1+(1-\delta)m\lambda_j/\delta\Big)\Big)^{1/2} \|f_S - f_{\tilde{\rho}}\|_{L_2(\Pi)}$$

$$\le \varepsilon\sqrt{r}m\log(m/\delta)\|f_S - f_{\tilde{\rho}}\|_{L_2(\Pi)} \le rm^2\varepsilon^2\log^2(m/\delta) + \tfrac{1}{4}\|f_S - f_{\tilde{\rho}}\|^2_{L_2(\Pi)},$$

where it was used that for $\lambda_j \in [0,1]$

$$\log\Big(1+(1-\delta)m\lambda_j/\delta\Big) \le \log\Big(\frac{\delta+(1-\delta)m}{\delta}\Big) \le \log(m/\delta).$$

Substituting bounds (2.3.14) and (2.3.15) in (2.3.13) we easily get bound (2.3.7), as claimed in the lemma.

□

We will also need the following simple lemma that provides a bound on $K(S'\|\tilde{\rho})$ in terms of $K(S\|\tilde{\rho})$.

Let

$$h(\delta) := \delta \log \frac{1}{\delta} + (1-\delta) \log \frac{1}{1-\delta}.$$

Observe that

$$h(\delta) = \delta \log \frac{1}{\delta} + (1-\delta) \log\left(1 + \frac{\delta}{1-\delta}\right) \leq \delta \log \frac{1}{\delta} + (1-\delta)\frac{\delta}{1-\delta} \leq \delta \log \frac{e}{\delta}$$

(this bound will be used in what follows).

Lemma 10. *Let $\delta \in (0,1)$, $S' \in \mathcal{S}_m$ with $\mathrm{rank}(S') = r$ and $S = (1-\delta)S' + \delta \frac{I_m}{m}$. Then, for any $U \in \mathcal{S}_m$,*

$$K(S' \| U) \leq \frac{K(S\|U) + h(\delta)}{1-\delta}.$$

Proof. The following identities are straightforward:

$$K(S\|U) = \mathrm{tr}(S(\log S - \log U))$$

$$= (1-\delta)\mathrm{tr}(S'(\log S - \log U)) + \delta \mathrm{tr}((I_m/m)(\log S - \log U))$$

$$= (1-\delta)\mathrm{tr}(S'(\log S' - \log U)) + (1-\delta)\mathrm{tr}(S'(\log S - \log S'))$$

$$+\delta \mathrm{tr}((I_m/m)(\log S - \log(I_m/m))) + \delta \mathrm{tr}((I_m/m)(\log(I_m/m) - \log U))$$

$$= (1-\delta)K(S'\|U) - (1-\delta)K(S'\|S) + \delta K(I_m/m\|U) - \delta K(I_m/m\|S).$$

Since $K(I_m/m\|U) \geq 0$, it follows that

$$K(S'\|U) \leq \frac{K(S\|U)}{1-\delta} + K(S'\|S) + \frac{\delta}{1-\delta} K(I_m/m\|S). \tag{2.3.16}$$

Assuming that S' has spectral representation $S' = \sum_{j=1}^r \lambda_j P_j$ with eigenvalues $\lambda_j > 0$ and one-dimensional projectors P_j, we get

$$-K(S'\|S) = \sum_{j=1}^r \lambda_j \log \frac{(1-\delta)\lambda_j + \delta/m}{\lambda_j}$$

$$= \sum_{j=1}^r \lambda_j \log\left(1 - \delta + \frac{\delta}{m\lambda_j}\right) \geq \log(1-\delta) \sum_{j=1}^r \lambda_j = \log(1-\delta),$$

implying that $K(S'\|S) \le \log\frac{1}{1-\delta}$. On the other hand,

$$K(I_m/m\|S) = \frac{1}{m}\sum_{j=1}^{m}\log\frac{1/m}{(1-\delta)\lambda_j + \delta/m} \le \frac{1}{m}\sum_{j=1}^{m}\log\frac{1}{\delta} = \log\frac{1}{\delta}.$$

Substituting these bounds in (2.3.16) yields the result. □

To complete the proof of Theorem 8, we need to control the empirical process $(P-P_n)(\ell'\bullet f_{\tilde\rho})(f_{\tilde\rho}-f_S)$ in the right hand side of bound (2.3.7). Our approach is based on the following empirical processes bound that is a slight modification of Lemma 1 in [55]. As before, we assume that $S=(1-\delta)S'+\delta\frac{I_m}{m}$ with $S'\in\mathcal{S}_m$, $\mathrm{rank}(S')=r$. We will set $\delta := \frac{1}{m^2n^2}$.

Let $\Xi_\varepsilon := n^{-1}\sum_{j=1}^{n}\varepsilon_j X_j$, where ε_j are i.i.d. Rademacher random variables (that is, ε_j takes values $+1$ and -1 with probability $1/2$ each) and $\{\varepsilon_j\},\{X_j\}$ are independent.

Lemma 11. *Given $\delta_1,\delta_2>0$, denote*

$$\alpha_n(\delta_1,\delta_2) := \sup\biggl\{\Bigl|(P_n-P)(\ell'\bullet f_A)(f_A-f_S)\Bigr| : A\in\mathcal{S}_m, \|f_A-f_S\|_{L_2(\Pi)}\le\delta_1, \|\mathcal{P}_L^{\perp}A\|_1\le\delta_2\biggr\}.$$

Let $0<\delta_1^-<\delta_1^+, 0<\delta_2^-<\delta_2^+$. For $t\ge 1$, denote

$$\bar t := t+\log\Bigl([\log_2(\delta_1^+/\delta_1^-)]+2\Bigr)+\log\Bigl([\log_2(\delta_2^+/\delta_2^-)]+2\Bigr)+\log 3.$$

Then, with probability at least $1-e^{-t}$, for all $\delta_1\in[\delta_1^-,\delta_1^+], \delta_2\in[\delta_2^-,\delta_2^+]$,

$$\alpha_n(\delta_1,\delta_2)\le C_1U\mathbb{E}\|\Xi_\varepsilon\|_\infty\Bigl(\sqrt{r}m\delta_1+\delta_2+\delta\Bigr)+C_2U\delta_1\sqrt{\frac{\bar t}{n}}+C_3U^2\frac{\bar t}{n},$$

where $C_1,C_2,C_3>0$ are constants.

We will use this lemma to control the term $(P-P_n)(\ell'\bullet f_{\tilde\rho})(f_{\tilde\rho}-f_S)$ in bound (2.3.7). Let $\delta_1 := \|f_{\tilde\rho}-f_S\|_{L_2(\Pi)}$ and $\delta_2 := \|\mathcal{P}_L^{\perp}\tilde\rho\|_1$. Define also

$$\delta_1^+ := \frac{2}{m},\ \delta_2^+ := 1,\ \delta_1^- = \delta_2^- := \frac{1}{mn},$$

so that $\bar{t} \le t + 2\log(\log_2(mn)+3) + \log 3$. It is easy to see that $\delta_1 \le \delta_1^+$ and $\delta_2 \le \delta_2^+$. If, in addition, $\delta_1 \ge \delta_1^-$, $\delta_2 \ge \delta_2^-$, the bound of Lemma 11 implies that with probability at least $1 - e^{-t}$,

$$
\begin{aligned}
&(P - P_n)(\ell' \bullet f_{\tilde\rho})(f_{\tilde\rho} - f_S) \le \alpha_n(\delta_1, \delta_2) \\
&\le C_1 U \mathbb{E}\|\Xi_\varepsilon\|_\infty \Big(\sqrt{r} m \delta_1 + \delta_2 + \delta\Big) + C_2 U \delta_1 \sqrt{\frac{\bar{t}}{n}} + C_3 U^2 \frac{\bar{t}}{n}.
\end{aligned}
$$

If $\bar\varepsilon \ge C_1 U \mathbb{E}\|\Xi_\varepsilon\|_\infty$, the last bound implies that

$$
\begin{aligned}
&(P - P_n)(\ell' \bullet f_{\tilde\rho})(f_{\tilde\rho} - f_S) \\
&\le \tfrac{1}{4}\|f_{\tilde\rho} - f_S\|_{L_2(\Pi)}^2 + r m^2 \bar\varepsilon^2 + \bar\varepsilon \|\mathcal{P}_L^\perp \tilde\rho\|_1 + \bar\varepsilon \delta \qquad (2.3.17) \\
&+ \tfrac{1}{4}\|f_{\tilde\rho} - f_S\|_{L_2(\Pi)}^2 + (C_2^2 + C_3) U^2 \tfrac{\bar{t}}{n}.
\end{aligned}
$$

Substituting this bound in the right hand side of (2.3.9), we get

$$
\begin{aligned}
&\|f_{\tilde\rho} - f_\rho\|_{L_2(\Pi)}^2 + \varepsilon K(\tilde\rho; S) \\
&\le \|f_{S'} - f_\rho\|_{L_2(\Pi)}^2 + r m^2 \varepsilon^2 \log^2(m/\delta) + 2 r m^2 \bar\varepsilon^2 \qquad (2.3.18) \\
&+ 5\bar\varepsilon\delta + C U^2 \tfrac{\bar{t}}{n} + \tfrac{12\delta}{m^2},
\end{aligned}
$$

where $C := C_2^2 + C_3$.

In the case when $\delta_1 = \|f_{\tilde\rho} - f_S\|_{L_2(\Pi)} \le \delta_1^- = \frac{1}{mn}$ or $\delta_2 = \|\mathcal{P}_L^\perp \tilde\rho\|_1 \le \delta_2^- = \frac{1}{mn}$, we can replace the terms $\frac{1}{4}\|f_{\tilde\rho} - f_S\|_{L_2(\Pi)}^2$ or $\|\mathcal{P}_L^\perp \tilde\rho\|_1$ in bound (2.3.17) by their respective upper bounds ($\frac{1}{4}(\delta_1^-)^2 = \frac{1}{4m^2n^2}$, or $\delta_2^- = \frac{1}{mn}$), which would be smaller than $CU^2\frac{\bar{t}}{n}$ for large enough $C > 0$, so bound (2.3.18) still holds (recall that $U \ge m^{-1/2}$). Note also that $\frac{12\delta}{m^2} = 12\frac{1}{m^4n^2} \le 12U^2\frac{\bar{t}}{n}$. Thus, increasing the value of constant C, one can rewrite (2.3.18) in a simpler form as

$$
\begin{aligned}
&\|f_{\tilde\rho} - f_\rho\|_{L_2(\Pi)}^2 + \varepsilon K(\tilde\rho; S) \\
&\le \|f_{S'} - f_\rho\|_{L_2(\Pi)}^2 + r m^2 \varepsilon^2 \log^2(m/\delta) + 2 r m^2 \bar\varepsilon^2 \qquad (2.3.19) \\
&+ 5\bar\varepsilon\delta + C U^2 \tfrac{\bar{t}}{n}.
\end{aligned}
$$

The following expectation bound is a consequence of a matrix version of Bernstein inequality for $\|\Xi_\varepsilon\|_\infty$ in Lemma 7 in Section 2.3.1 (it follows by integrating out its exponential tails):

$$\mathbb{E}\|\Xi_\varepsilon\|_\infty \leq 4\biggl[\sqrt{\frac{\log(2m)}{nm}} \bigvee U\frac{\log(2m)}{n}\biggr]$$

(it is also used in this computation that, in the case of uniform sampling from an orthonormal basis, $\sigma_{\varepsilon X}^2 = \|\mathbb{E}X^2\|_\infty = \frac{1}{m}$, a simple fact often used in the literature; see, e.g., [53], Section 5). Let

$$\bar{\varepsilon} := D'U\sqrt{\frac{\log(2m)}{nm}}$$

for some constant D'. If D' is sufficiently large and

$$U\frac{\log(2m)}{n} \leq \sqrt{\frac{\log(2m)}{nm}}, \tag{2.3.20}$$

then the condition $\bar{\varepsilon} \geq C_1 U\mathbb{E}\|\Xi_\varepsilon\|_\infty$ is satisfied and bound (2.3.19) holds with probability at least $1-e^{-t}$. Moreover, $\bar{\varepsilon}\delta \lesssim_{D'} \delta \lesssim_{D'} U^2\frac{\bar{t}}{n}$, implying that the term $5\bar{\varepsilon}\delta$ in (2.3.19) can be dropped at a price of further increasing the value of constant C.

If (2.3.20) does not hold, we still have that

$$\|f_{\tilde{\rho}} - f_\rho\|_{L_2(\Pi)}^2 = \frac{\|\tilde{\rho}-\rho\|_2^2}{m^2} \leq \frac{2}{m^2} \leq CU^2\frac{\bar{t}}{n}.$$

Recalling that $\bar{t} \leq t + 2\log(\log_2(mn)+3)$ and $\log(m/\delta) \lesssim \log(mn)$, we deduce from (2.3.19) that with some constant C and with probability at least $1-e^{-t}$

$$\|f_{\tilde{\rho}} - f_\rho\|_{L_2(\Pi)}^2 \leq \|f_{S'} - f_\rho\|_{L_2(\Pi)}^2 + C\biggl[rm^2\varepsilon^2\log^2(mn) + U^2\tfrac{rm\log(2m)}{n} + U^2\tfrac{t+\log(\log_2(mn)+3)}{n}\biggr]. \tag{2.3.21}$$

Note that, for $n \geq 2$,

$$\log(\log_2(mn)+3) = \log\Bigl(\log_2(4m)+\log_2(2n)\Bigr) \leq \log\log_2(4m)+\log\log_2(2n), \tag{2.3.22}$$

since $\log_2(4m)+\log_2(2n) \leq \log_2(4m)\log_2(2n)$. Since also, for $r \geq 1$,

$$U^2\frac{t+\log\log_2(4m)}{n} \lesssim U^2\frac{rm\log(2m)}{n}, \tag{2.3.23}$$

we can replace in bound (2.3.21) the term $U^2\frac{t+\log(\log_2(mn)+3)}{n}$ with the term $U^2\frac{t+\log\log_2(2n)}{n}$ (increasing the value of the constant C accordingly). This yields bound (2.3.3) of the theorem. For $S'=\rho$, it yields bound (2.3.4), and, moreover, for $S'=\rho$ and $S=(1-\delta)\rho+\delta\frac{I_m}{m}$ with $\delta=\frac{1}{m^2n^2}$, bound (2.3.19) also implies that

$$\varepsilon K(\tilde{\rho};S)\leq \mathrm{rank}(\rho)m^2\varepsilon^2\log^2(m/\delta)+2\mathrm{rank}(\rho)m^2\bar{\varepsilon}^2 \tag{2.3.24}$$
$$+5\bar{\varepsilon}\delta+CU^2\tfrac{\bar{t}}{n}.$$

We will now take

$$\bar{\varepsilon}:=D'\biggl[U\sqrt{\frac{\log(2m)}{nm}}\bigvee U^2\frac{\log(2m)}{n}\biggr]$$

for a large enough constant D' so that $\bar{\varepsilon}\geq C_1U\mathbb{E}\|\Xi_\varepsilon\|_\infty$. Assume that

$$\varepsilon:=\frac{1}{\log(mn)}\biggl[U\sqrt{\frac{\log(2m)}{nm}}\bigvee U^2\frac{\log(2m)}{n}\biggr].$$

As before, the term $\bar{\varepsilon}\delta$ in bound (2.3.24) will be absorbed by the term $CU^2\frac{\bar{t}}{n}$ with a larger value of C and also

$$\mathrm{rank}(\rho)m^2\varepsilon^2\log^2(m/\delta)\asymp_{D'}\mathrm{rank}(\rho)m^2\bar{\varepsilon}^2\asymp_{D'}U^2\frac{\mathrm{rank}(\rho)m\log(2m)}{n}\biggl(1\bigvee U^2\frac{m\log(2m)}{n}\biggr).$$

As a result, taking into account (2.3.22), (2.3.23), bound (2.3.24) can be rewritten as follows:

$$\varepsilon K(\tilde{\rho};S)\leq CU^2\biggl[\tfrac{\mathrm{rank}(\rho)m\log(2m)}{n}\biggl(1\bigvee U^2\tfrac{m\log(2m)}{n}\biggr) \tag{2.3.25}$$
$$+\tfrac{t+\log\log_2(2n)}{n}\biggr].$$

Using the bound of Lemma 10 along with the bound

$$h(\delta)\leq\delta\log(e/\delta)=\frac{1}{m^2n^2}\log(em^2n^2)\lesssim U\sqrt{\frac{m}{n}}\frac{(t+\log\log_2(2n))\log(mn)}{\sqrt{\log(2m)}},$$

we easily get that (2.3.6) holds. □

2.3.3 Oracle inequalities for trace regression with Gaussian noise

In this subsection, we establish oracle inequalities for the von Neumann entropy penalized least squares estimator $\tilde{\rho}^{\varepsilon}$ in the case of trace regression model with Gaussian noise (Assumption 4). Unlike in the case of Theorem 8 of the previous section, our aim is not to obtain sharp oracle inequality, but rather to get a clean main term of the random error bound part of the inequality, namely, the term $\sigma_\xi^2 \frac{\mathrm{rank}(S)m(t+\log(2m))}{n}$ in inequality (2.3.27) below. Note that this term depends only on the variance of the noise σ_ξ^2, but not on the constant U from Assumption 1 (the constant U is involved only in the higher order $O(n^{-2})$ terms of the bound). Note also that there are no constraints on the variance σ_ξ^2 that could be arbitrarily small, or even equal to 0 (in which case only higher order terms are present in the bound). This improvement comes at a price of having the leading constant 2 in the oracle inequality and also of imposing assumption (2.3.26) that requires the regularization parameter ε to be bounded away from 0 (again, unlike Theorem 8, where it could be arbitrarily small). As in the previous section, we also obtain a bound on Kullback–Leibler divergence $K(\rho\|\tilde{\rho}^{\varepsilon})$.

Theorem 9. *Let $t \geq 1$. Suppose*

$$\varepsilon \in \left[DU^2 \frac{t+\log^3 m \log^2 n}{n}, \frac{D_1 \sigma_\xi}{\log(mn)} \sqrt{\frac{t+\log(2m)}{nm}} \bigvee DU^2 \frac{t+\log^3 m \log^2 n}{n} \right] \tag{2.3.26}$$

with large enough constants $D, D_1 > 0$. There exists a constant $C > 0$ such that with probability at least $1-e^{-t}$

$$\|f_{\tilde{\rho}^{\varepsilon}} - f_\rho\|_{L_2(\Pi)}^2 \leq \inf_{S \in \mathcal{S}_m} \Bigg[2\|f_S - f_\rho\|_{L_2(\Pi)}^2 + C\bigg(\sigma_\xi^2 \frac{\mathrm{rank}(S)m(t+\log(2m))}{n} + \sigma_\xi^2 U^2 \frac{\mathrm{rank}(S)m^2(t+\log(2m))^2\log(2m)}{n^2} + U^4 \frac{\mathrm{rank}(S)m^2(t+\log^3 m\log^2 n)^2\log^2(mn)}{n^2} \bigg) \Bigg]. \tag{2.3.27}$$

In particular,

$$\|f_{\tilde{\rho}^{\varepsilon}} - f_\rho\|^2_{L_2(\Pi)} \le C\Bigg[\sigma_\xi^2 \tfrac{\mathrm{rank}(\rho) m (t+\log(2m))}{n} \tag{2.3.28}$$

$$+\sigma_\xi^2 U^2 \tfrac{\mathrm{rank}(\rho) m^2 (t+\log(2m))^2 \log(2m)}{n^2} + U^4 \tfrac{\mathrm{rank}(\rho) m^2 (t+\log^3 m \log^2 n)^2 \log^2(mn)}{n^2}\Bigg].$$

Moreover, if

$$\varepsilon := \frac{D_1 \sigma_\xi}{\log(mn)} \sqrt{\frac{t+\log(2m)}{nm}} \bigvee D U^2 \frac{t+\log^3 m \log^2 n}{n}$$

for large enough constants D, D_1, then with some constant C and with the same probability both (2.3.28) and the following bound hold:

$$K(\rho\|\tilde{\rho}^{\varepsilon}) \le C\Bigg[\sigma_\xi \tfrac{\mathrm{rank}(\rho) m^{3/2} (t+\log(2m))^{1/2} \log(mn)}{\sqrt{n}} \tag{2.3.29}$$

$$+\sigma_\xi^2 \tfrac{\mathrm{rank}(\rho) m^2 (t+\log(2m)) \log(2m)}{n} + U^2 \tfrac{\mathrm{rank}(\rho) m^2 (t+\log^3 m \log^2 n) \log^2(mn)}{n}\Bigg].$$

Proof. As in in the proof of Theorem 8, we rely on Lemma 9, but we use a different approach to bounding the empirical process $(P-P_n)(\ell' \bullet f_{\tilde{\rho}})(f_{\tilde{\rho}} - f_S)$. The following identity follows from the definition of quadratic loss ℓ

$$(\ell' \bullet f)(x,y)(f(x) - f_S(x)) = 2(f(x) - f_S(x))^2 + 2(f_S(x) - y)(f(x) - f_S(x))$$

and it implies that

$$(P-P_n)(\ell' \bullet f_{\tilde{\rho}})(f_{\tilde{\rho}} - f_S) = -2(P_n - P)(f_{\tilde{\rho}} - f_S)^2 - 2\langle \Xi, \tilde{\rho} - S\rangle \tag{2.3.30}$$

where

$$\Xi := n^{-1} \sum_{j=1}^{n} (f_S(X_j) - Y_j) X_j - \mathbb{E}(f_S(X) - Y)X.$$

We will bound $(P_n - P)(f_{\tilde{\rho}} - f_S)^2$ in representation (2.3.30) as follows:

$$\Big|(P_n - P)(f_{\tilde{\rho}} - f_S)^2\Big| \le \|\tilde{\rho} - S\|_1^2 \beta_n\bigg(\frac{\|f_{\tilde{\rho}} - f_S\|_{L_2(\Pi)}}{\|\tilde{\rho} - S\|_1}\bigg), \tag{2.3.31}$$

where

$$\beta_n(\Delta) := \sup\bigg\{\Big|(P_n - P)(f_A^2)\Big| : A \in \mathbb{H}_m, \|A\|_1 \le 1, \|f_A\|_{L_2(\Pi)} \le \Delta\bigg\}.$$

The next lemma provides a bound on $\beta_n(\Delta)$. Its proof is somewhat involved and it will be given in Section 2.5. It is based on Rudelson's $L_\infty(P_n)$ generic chaining bound for empirical processes indexed by squares of functions and on the ideas of the paper by [38] combined with Talagrand's concentration inequality (see also [4], [64] and Theorem 3.16, Lemma 9.8 and Proposition 9.2 in [52] for similar arguments).

Lemma 12. *Given $0 < \delta^- < \delta^+$ and $t \geq 1$, let*

$$\bar{t} := t + \log\Big(\log_2(\delta^+/\delta^-) + 3\Big).$$

Then, with some constant C and with probability at least $1 - e^{-t}$, the following bound holds for all $\Delta \in [\delta^-, \delta^+]$:

$$\beta_n(\Delta) \leq C\left[\Delta U \frac{\log^{3/2} m \log n}{\sqrt{n}} + U^2 \frac{\log^3 m \log^2 n}{n} + \Delta U \sqrt{\frac{\bar{t}}{n}} + U^2 \frac{\bar{t}}{n}\right]. \qquad (2.3.32)$$

We will use Lemma 12 to control $\beta_n(\Delta)$ for $\Delta := \frac{\|f_{\tilde{\rho}} - f_S\|_{L_2(\Pi)}}{\|\tilde{\rho} - S\|_1}$. Let $\delta^+ := \frac{1}{m}$ and $\delta^- := \frac{1}{mn}$. With this choice, $\bar{t} \leq t + \log(\log_2 n + 3)$. Note that for $A = \frac{\tilde{\rho} - S}{\|\tilde{\rho} - S\|_1}$, $\|f_A\|_{L_2(\Pi)} = \frac{\|A\|_2}{m} \leq \frac{\|A\|_1}{m} = m^{-1} = \delta^+$. If also $\|f_A\|_{L_2(\Pi)} \geq \delta^-$, then we can substitute bound (2.3.32) on $\beta_n(\Delta)$ into (2.3.31) that yields:

$$\begin{aligned}
\Big|(P_n - P)(f_{\tilde{\rho}} - f_S)^2\Big| &\leq C\Big[\|f_{\tilde{\rho}} - f_S\|_{L_2(\Pi)}\|\tilde{\rho} - S\|_1 U \tfrac{\log^{3/2} m \log n}{\sqrt{n}} \\
&\quad + \|\tilde{\rho} - S\|_1^2 U^2 \tfrac{\log^3 m \log^2 n}{n} + \|f_{\tilde{\rho}} - f_S\|_{L_2(\Pi)}\|\tilde{\rho} - S\|_1 U \sqrt{\tfrac{\bar{t}}{n}} \\
&\quad + \|\tilde{\rho} - S\|_1^2 U^2 \tfrac{\bar{t}}{n}\Big] \\
&\leq \tfrac{1}{32}\|f_{\tilde{\rho}} - f_S\|_{L_2(\Pi)}^2 + 8(C^2 + C/8)U^2 \tfrac{\log^3 m \log^2 n}{n}\|\tilde{\rho} - S\|_1^2 \qquad (2.3.33) \\
&\quad + \tfrac{1}{32}\|f_{\tilde{\rho}} - f_S\|_{L_2(\Pi)}^2 + 8(C^2 + C/8)U^2 \tfrac{\bar{t}}{n}\|\tilde{\rho} - S\|_1^2 \\
&\leq \tfrac{1}{16}\|f_{\tilde{\rho}} - f_S\|_{L_2(\Pi)}^2 + C' U^2 \tfrac{\log^3 m \log^2 n + \bar{t}}{n}\|\tilde{\rho} - S\|_1^2,
\end{aligned}$$

where $C' := 8(C^2 + C/8)$. If, on the other hand, $\|f_A\|_{L_2(\Pi)} \leq \delta^- = \frac{1}{mn}$, then $\|f_{\tilde{\rho}} - f_S\|_{L_2(\Pi)}$ in the above bound can be replaced by $\frac{1}{mn}\|\tilde{\rho} - S\|_1$ and the proof that follows only simplifies since

$$\frac{1}{16}\|f_{\tilde{\rho}} - f_S\|_{L_2(\Pi)}^2 \leq \frac{1}{16}\frac{1}{m^2 n^2}\|\tilde{\rho} - S\|_1^2 \leq \frac{1}{16} U^2 \frac{\log^3 m \log^2 n + \bar{t}}{n}\|\tilde{\rho} - S\|_1^2.$$

Another term in the right hand side of representation (2.3.30) to be controlled is $\langle \Xi, \tilde{\rho} - S\rangle$. Note that $\Xi = \Xi_1 + \Xi_2$, where

$$\Xi_1 := -n^{-1}\sum_{j=1}^{n} \xi_j X_j$$

and

$$\Xi_2 := n^{-1}\sum_{j=1}^{n}(f_S(X_j) - f_\rho(X_j))X_j - \mathbb{E}(f_S(X) - f_\rho(X))X.$$

Recall that $S = (1-\delta)S' + \delta\frac{I_m}{m}$ with $S' \in \mathcal{S}_m$, $\mathrm{rank}(S') = r$, $\mathrm{supp}(S') = L$ and $\delta = \frac{1}{m^2n^2}$.

The term with Ξ_1 is controlled as follows:

$$\begin{aligned}
&\Big|\langle \Xi_1, \tilde{\rho} - S\rangle\Big| \\
&\leq \Big|\langle \mathcal{P}_L(\Xi_1), \tilde{\rho} - S'\rangle\Big| + \Big|\langle \Xi_1, \mathcal{P}_L^\perp(\tilde{\rho} - S')\rangle\Big| + \Big|\langle \mathcal{P}_L^\perp(\Xi_1), S' - S\rangle\Big| \\
&\leq \|\mathcal{P}_L(\Xi_1)\|_2\|\tilde{\rho} - S'\|_2 + \|\Xi_1\|_\infty\|\mathcal{P}_L^\perp(\tilde{\rho})\|_1 + \Big\|\mathcal{P}_L^\perp(\Xi_1)\Big\|_\infty\|S' - S\|_1 \\
&\leq 2\sqrt{2r}m\|\Xi_1\|_\infty\|f_{\tilde{\rho}} - f_S\|_{L_2(\Pi)} + \|\Xi_1\|_\infty\|\mathcal{P}_L^\perp(\tilde{\rho})\|_1 + 4\delta\|\Xi_1\|_\infty \qquad (2.3.34) \\
&\leq 32rm^2\|\Xi_1\|_\infty^2 + \tfrac{1}{16}\|f_{\tilde{\rho}} - f_S\|_{L_2(\Pi)}^2 \\
&\quad + \|\Xi_1\|_\infty\|\mathcal{P}_L^\perp(\tilde{\rho})\|_1 + 4\delta\|\Xi_1\|_\infty.
\end{aligned}$$

We also have

$$\begin{aligned}
\Big|\langle \Xi_2, \tilde{\rho} - S\rangle\Big| &\leq \|\Xi_2\|_\infty\|\tilde{\rho} - S\|_1 \leq \|\Xi_2\|_\infty\|\tilde{\rho} - S'\|_1 + \|\Xi_2\|_\infty\|S' - S\|_1 \\
&\leq \|\Xi_2\|_\infty\|\tilde{\rho} - S'\|_1 + 2\delta\|\Xi_2\|_\infty. \qquad (2.3.35)
\end{aligned}$$

Thus,

$$\begin{aligned}
\Big|\langle \Xi, \tilde{\rho} - S\rangle\Big| &\leq 32rm^2\|\Xi_1\|_\infty^2 + \tfrac{1}{16}\|f_{\tilde{\rho}} - f_S\|_{L_2(\Pi)}^2 \\
&+ \|\Xi_1\|_\infty\|\mathcal{P}_L^\perp(\tilde{\rho})\|_1 + 4\delta\|\Xi_1\|_\infty + \|\Xi_2\|_\infty\|\tilde{\rho} - S'\|_1 + 2\delta\|\Xi_2\|_\infty. \qquad (2.3.36)
\end{aligned}$$

It follows from (2.3.30), (2.3.33) and (2.3.36) that with some constant C'

$$\begin{aligned}(P-P_n)(\ell'\bullet f_{\tilde\rho})(f_{\tilde\rho}-f_S) &\le \\ \tfrac{1}{4}\|f_{\tilde\rho}-f_S\|_{L_2(\Pi)}^2 + C'U^2\tfrac{\log^3 m\log^2 n+\bar t}{n}\|\tilde\rho-S\|_1^2 & \qquad (2.3.37)\\ +64rm^2\|\Xi_1\|_\infty^2 + 2\|\Xi_1\|_\infty\|\mathcal{P}_L^\perp(\tilde\rho)\|_1 + 8\delta\|\Xi_1\|_\infty & \\ +2\|\Xi_2\|_\infty\|\tilde\rho-S'\|_1 + 4\delta\|\Xi_2\|_\infty. &\end{aligned}$$

This bound will be substituted in (2.3.7). Note that, if assumption (2.3.26) on ε holds with a sufficiently large constant D, then we have

$$\varepsilon \ge 8C'U^2\frac{\log^3 m\log^2 n+\bar t}{n}$$

(this follows from the fact that $\bar t \le t+\log(\log_2 n+3) \le t + c\log^3 m\log^2 n$ for some constant $c>0$). Assume also that $\bar\varepsilon \ge 4\|\Xi_1\|_\infty$ and recall that $K(\tilde\rho;S)\ge \frac{1}{4}\|\tilde\rho-S\|_1^2$ (see inequality 1.3.10). Taking all this into account, (2.3.7) implies that

$$\begin{aligned}\|f_{\tilde\rho}-f_\rho\|_{L_2(\Pi)}^2 + \tfrac{1}{4}\|f_{\tilde\rho}-f_S\|_{L_2(\Pi)}^2 + \tfrac{\varepsilon}{2}K(\tilde\rho;S) + \tfrac{\bar\varepsilon}{2}\|\mathcal{P}_L^\perp\tilde\rho\|_1 &\\ \le \|f_S-f_\rho\|_{L_2(\Pi)}^2 + rm^2\varepsilon^2\log^2(m/\delta) + 5rm^2\bar\varepsilon^2 + 6\bar\varepsilon\delta & \qquad (2.3.38)\\ +2\|\Xi_2\|_\infty\|\tilde\rho-S'\|_1 + 4\|\Xi_2\|_\infty\delta. &\end{aligned}$$

It remains to control $\|\Xi_1\|_\infty$ and $\|\Xi_2\|_\infty$. To this end, we use matrix versions of Bernstein inequality. To bound $\|\Xi_2\|_\infty$, we use its standard version as in Lemma 7 which yields that with probability at least $1-e^{-t}$

$$\begin{aligned}\|\Xi_2\|_\infty \le 2\Bigg[\Big\|\mathbb{E}(f_S(X)-f_\rho(X))^2X^2\Big\|_\infty^{1/2}\sqrt{\tfrac{t+\log(2m)}{n}}\\ \bigvee\Big\|(f_S(X)-f_\rho(X))\|X\|_\infty\Big\|_{L_\infty}\tfrac{t+\log(2m)}{n}\Bigg],\end{aligned}$$

where $\|\cdot\|_{L_\infty}$ denotes the essential supremum norm in the space of random variables. Since

$$\Big\|\mathbb{E}(f_S(X)-f_\rho(X))^2X^2\Big\|_\infty \le U^2\|f_S-f_\rho\|_{L_2(\Pi)}^2$$

and

$$\Big\|(f_S(X)-f_\rho(X))\|X\|_\infty\Big\|_{L_\infty}\leq 2U^2,$$

we get

$$\|\Xi_2\|_\infty\leq 4\Bigg[\|f_S-f_\rho\|_{L_2(\Pi)}U\sqrt{\tfrac{t+\log(2m)}{n}}+U^2\tfrac{t+\log(2m)}{n}\Bigg]. \tag{2.3.39}$$

This implies that

$$\begin{aligned}2\|\Xi_2\|_\infty\|\tilde\rho-S'\|_1\leq \|f_S-f_\rho\|^2_{L_2(\Pi)}+16U^2\tfrac{t+\log(2m)}{n}\|\tilde\rho-S'\|_1^2 \\ +8U^2\tfrac{t+\log(2m)}{n}\|\tilde\rho-S'\|_1.\end{aligned} \tag{2.3.40}$$

Note that

$$\begin{aligned}&16U^2\tfrac{t+\log(2m)}{n}\|\tilde\rho-S'\|_1^2\\ &\leq 16U^2\tfrac{t+\log(2m)}{n}\|\tilde\rho-S\|_1^2+16U^2\tfrac{t+\log(2m)}{n}(4\delta+\delta^2)\end{aligned} \tag{2.3.41}$$

and

$$\begin{aligned}&8U^2\tfrac{t+\log(2m)}{n}\|\tilde\rho-S'\|_1\\ &\leq 8U^2\tfrac{t+\log(2m)}{n}\|\mathcal{P}_L^\perp\tilde\rho\|_1+8U^2\tfrac{t+\log(2m)}{n}\|\mathcal{P}_L(\tilde\rho-S')\|_1\\ &\leq 8U^2\tfrac{t+\log(2m)}{n}\|\mathcal{P}_L^\perp\tilde\rho\|_1+8U^2\tfrac{t+\log(2m)}{n}\|\mathcal{P}_L(\tilde\rho-S)\|_1+16U^2\tfrac{t+\log(2m)}{n}\delta.\end{aligned} \tag{2.3.42}$$

Since, for some constant $C''>0$,

$$\begin{aligned}8U^2\tfrac{t+\log(2m)}{n}\|\mathcal{P}_L(\tilde\rho-S)\|_1\leq 8\sqrt{2}U^2\tfrac{t+\log(2m)}{n}\sqrt{r}\|\mathcal{P}_L(\tilde\rho-S)\|_2\\ \leq 8\sqrt{2}U^2\tfrac{t+\log(2m)}{n}\sqrt{r}m\|f_{\tilde\rho}-f_S\|_{L_2(\Pi)}\leq\tfrac14\|f_{\tilde\rho}-f_S\|^2_{L_2(\Pi)}+C''U^4\tfrac{rm^2(t+\log(2m))^2}{n^2},\end{aligned}$$

it follows from (2.3.40), (2.3.41) and (2.3.42) that

$$\begin{aligned}2\|\Xi_2\|_\infty\|\tilde\rho-S'\|_1\leq\|f_S-f_\rho\|^2_{L_2(\Pi)}+\\ +16U^2\tfrac{t+\log(2m)}{n}\|\tilde\rho-S\|_1^2+16U^2\tfrac{t+\log(2m)}{n}(4\delta+\delta^2)\\ +8U^2\tfrac{t+\log(2m)}{n}\|\mathcal{P}_L^\perp\tilde\rho\|_1+16U^2\tfrac{t+\log(2m)}{n}\delta\\ +\tfrac14\|f_{\tilde\rho}-f_S\|^2_{L_2(\Pi)}+C''U^4\tfrac{rm^2(t+\log(2m))^2}{n^2}.\end{aligned} \tag{2.3.43}$$

Note that (2.3.39) also implies that

$$\|\Xi_2\|_\infty \le 4\left[\tfrac{2U}{m}\sqrt{\tfrac{t+\log(2m)}{n}} + U^2\tfrac{t+\log(2m)}{n}\right] \tag{2.3.44}$$

(since $\|f_S - f_\rho\|_{L_2(\Pi)} \le m^{-1}\|S-\rho\|_2 \le 2m^{-1}$). Let us substitute (2.3.43) and (2.3.44) in the last line of (2.3.38). Assume that

$$\bar{\varepsilon} \ge 16U^2\frac{t+\log(2m)}{n}$$

and that constant D in assumption (2.3.26) is large enough so that

$$16U^2\frac{t+\log(2m)}{n}\|\tilde{\rho}-S\|_1^2 \le \frac{\varepsilon}{4}K(\tilde{\rho},S)$$

(recall inequality 1.3.10). It easily follows that with some constants C_1, C_2,

$$\begin{aligned} &\|f_{\tilde{\rho}} - f_\rho\|_{L_2(\Pi)}^2 + \tfrac{\varepsilon}{4}K(\tilde{\rho};S) \\ &\le 2\|f_S - f_\rho\|_{L_2(\Pi)}^2 + C_1 rm^2\varepsilon^2\log^2(m/\delta) + 5rm^2\bar{\varepsilon}^2 \\ &\quad + C_2\bar{\varepsilon}\delta + 32\tfrac{U}{m}\sqrt{\tfrac{t+\log(2m)}{n}}\delta \end{aligned} \tag{2.3.45}$$

(note that the term $C''U^4\frac{rm^2(t+\log(2m))^2}{n^2}$ of bound (2.3.43) is "absorbed" by the term $C_1rm^2\varepsilon^2\log^2(m/\delta)$ of bound (2.3.45) provided that constant C_1 is large enough). Since

$$\delta = \frac{1}{m^2n^2} \le U^2\frac{t+\log(2m)}{n} \le \bar{\varepsilon}$$

(recall that $U^2 \ge m^{-1}$), we have $\bar{\varepsilon}\delta \le \bar{\varepsilon}^2$. Also, since $U \ge m^{-1/2}$,

$$\frac{U}{m}\sqrt{\frac{t+\log(2m)}{n}}\delta = U\sqrt{\frac{t+\log(2m)}{n}}\frac{1}{m^3n^2} \le U^4\left(\frac{t+\log(2m)}{n}\right)^2 \le \bar{\varepsilon}^2.$$

Therefore, (2.3.45) implies that with some constant C

$$\begin{aligned} &\|f_{\tilde{\rho}} - f_\rho\|_{L_2(\Pi)}^2 + \tfrac{\varepsilon}{4}K(\tilde{\rho};S) \\ &\le 2\|f_S - f_\rho\|_{L_2(\Pi)}^2 + C\Big(rm^2\varepsilon^2\log^2(m/\delta) + rm^2\bar{\varepsilon}^2\Big). \end{aligned} \tag{2.3.46}$$

To bound $\|\Xi_1\|_\infty$, we use a version of matrix Bernstein type inequality as in Lemma 8 in Section 2.3.1. Its version for $\alpha = 2$ (with $U^{(\alpha)} \asymp U\sigma_\xi$) implies that for some constant $K > 0$ with probability at least $1 - e^{-t}$

$$\|\Xi_1\|_\infty \leq K\left[\sigma_\xi\sqrt{\frac{t+\log(2m)}{nm}}\bigvee \sigma_\xi U\frac{(t+\log(2m))\log^{1/2}(2Um^{1/2})}{n}\right]. \tag{2.3.47}$$

We choose

$$\bar{\varepsilon} := D_2\left[\sigma_\xi\sqrt{\frac{t+\log(2m)}{nm}}\bigvee(\sigma_\xi \vee U)U\frac{(t+\log(2m))\log^{1/2}(2m)}{n}\right]$$

with a sufficiently large constant D_2 to satisfy the condition $\|\Xi_1\|_\infty \leq 4\bar{\varepsilon}$ with probability at least $1 - e^{-t}$ (the rest of the assumptions we made on $\bar{\varepsilon}$ are also satisfied with this choice).

Bound (2.3.46) then implies that with some constant C and with probability at least $1 - 3e^{-t}$ the following inequality holds:

$$\begin{aligned}\|f_{\tilde{\rho}^\varepsilon} - f_\rho\|^2_{L_2(\Pi)} &\leq 2\|f_S - f_\rho\|^2_{L_2(\Pi)} \\ &+ C\left[\sigma_\xi^2\frac{rm(t+\log(2m))}{n} + \sigma_\xi^2 U^2\frac{rm^2(t+\log(2m))^2\log(2m)}{n^2}\right. \\ &\left.+ U^4\frac{rm^2(t+\log^3 m\log^2 n)^2\log^2(mn)}{n^2}\right].\end{aligned} \tag{2.3.48}$$

Using bound (2.3.8) to replace S in $\|f_S - f_\rho\|^2_{L_2(\Pi)}$ with S' and adjusting the value of constant C to rewrite the probability bound as $1 - e^{-t}$, it is easy to complete the proof of (2.3.27). If $S' = \rho$, this also yields bound (2.3.28). Moreover, with a larger value of regularization parameter

$$\varepsilon := \frac{D_1\sigma_\xi}{\log(mn)}\sqrt{\frac{t+\log(2m)}{nm}}\bigvee DU^2\frac{t+\log^3 m\log^2 n}{n},$$

bound (2.3.46) and Lemma 10 easily imply bound (2.3.29). □

2.3.4 Optimality of von Neumann entropy penalized estimator

We start with upper bounds on the error of estimator $\tilde{\rho}^\varepsilon$ (von Neumann entropy penalized least squares estimator defined by (2.3.1)) in Hellinger, Kullback-Leibler and Schatten p-norm distances for $p \in [1,2]$ for the trace regression model with Gaussian noise (Assumption 4). To avoid the impact of "second order terms" on the upper bounds, we will make the following simplifying assumptions:

$$U\sqrt{\frac{m}{n}}\log m \lesssim 1 \quad \text{and} \quad U^2\sqrt{\frac{m}{n}}\log^{5/2} m \log^2 n \log(mn) \lesssim \sigma_\xi. \tag{2.3.49}$$

Recall that, for the Pauli basis, $U = m^{-1/2}$, so, the above assumptions hold if $n \gtrsim \log^2 m$ and σ_ξ is larger than $\frac{1}{\sqrt{mn}}$ (times a logarithmic factor). We will choose regularization parameter ε as follows:

$$\varepsilon := \frac{D_1\sigma_\xi}{\log(mn)}\sqrt{\frac{\log(2m)}{nm}} \tag{2.3.50}$$

with a sufficiently large constant $D_1 > 0$. The next result shows that minimax rates of Theorem 4 are attained up to logarithmic factors for the estimator $\tilde{\rho}^\varepsilon$.

Theorem 10. *There exists a constant $C > 0$ such that the following bounds hold for all $r = 1, \dots, m$, for all $\rho \in \mathcal{S}_{r,m}$ and for all $p \in [1,2]$ with probability at least $1 - m^{-2}$:*

$$\|\tilde{\rho}^\varepsilon - \rho\|_p \le C\biggl(\frac{\sigma_\xi m^{\frac{3}{2}} r^{1/p}}{\sqrt{n}}\sqrt{\log m}\log^{(2-p)/p}(mn) \bigwedge \biggl(\frac{\sigma_\xi m^{3/2}}{\sqrt{n}}\biggr)^{1-\frac{1}{p}}(\log m)^{\frac{1}{2}-\frac{1}{2p}}\biggr) \bigwedge 2, \tag{2.3.51}$$

$$H^2(\tilde{\rho}^\varepsilon, \rho) \le C\frac{\sigma_\xi m^{\frac{3}{2}} r}{\sqrt{n}}\sqrt{\log m}\log(mn) \bigwedge 2 \tag{2.3.52}$$

and

$$K(\rho\|\tilde{\rho}^\varepsilon) \le C\frac{\sigma_\xi m^{\frac{3}{2}} r}{\sqrt{n}}\sqrt{\log m}\log(mn). \tag{2.3.53}$$

Proof. We will need the following simple lemma.

Lemma 13. *For all $\rho \in \mathcal{S}_m$ and all $l = 1, \dots, m$, there exists $\rho' \in \mathcal{S}_{l,m}$ such that*

$$\|\rho - \rho'\|_2^2 \le \frac{1}{l}.$$

Proof. Suppose that $\rho = \sum_{j=1}^m \lambda_j P_j$, where λ_j are the eigenvalues of ρ repeated with their multiplicities and P_j are orthogonal one-dimensional projectors. Note that

$\{\lambda_j : j = 1, \dots, m\}$ is a probability distribution on the set $\{1, \dots, m\}$. Let ν be a random variable sampled from this distribution and $\nu_1, \dots, \nu_l$ be its i.i.d. copies. Then $\mathbb{E}P_\nu = \rho$ and

$$\mathbb{E}\left\|l^{-1}\sum_{j=1}^{l} P_{\nu_j} - \rho\right\|_2^2 = \frac{\mathbb{E}\|P_\nu - \rho\|_2^2}{l} = \frac{\mathbb{E}\|P_\nu\|_2^2 - \|\rho\|_2^2}{l} = \frac{1 - \|\rho\|_2^2}{l} \leq \frac{1}{l}.$$

Therefore, there exists a realization $\nu_1 = k_1, \dots, \nu_l = k_l$ of r.v. $\nu_1, \dots, \nu_l$ such that

$$\left\|l^{-1}\sum_{j=1}^{l} P_{k_j} - \rho\right\|_2^2 \leq \frac{1}{l}.$$

Denote $\rho' := l^{-1}\sum_{j=1}^{l} P_{k_j}$. Then, $\rho' \in \mathcal{S}_{l,m}$ and $\|\rho - \rho'\|_2^2 \leq \frac{1}{l}$. □

First, we will prove bound (2.3.51) for $p = 2$. To this end, we use oracle inequality (2.3.27) with $t = 2\log m + \log 2$ and with oracle $S = \rho' \in \mathcal{S}_{l,m}$ such that $\|\rho - \rho'\|_2^2 \leq \frac{1}{l}$. Under simplifying assumptions (2.3.49) it yields that with probability at least $1 - \frac{1}{2}m^{-2}$

$$\|\tilde{\rho}^\varepsilon - \rho\|_2^2 = m^2\|f_{\tilde{\rho}^\varepsilon} - f_\rho\|_{L_2(\Pi)}^2 \lesssim \left[\frac{1}{l} + \tau^2 l \log m\right],$$

where $\tau := \frac{\sigma_\xi m^{3/2}}{\sqrt{n}}$. On the other hand, using the same inequality with $S = \rho \in \mathcal{S}_{r,m}$ yields the bound

$$\|\tilde{\rho}^\varepsilon - \rho\|_2^2 \lesssim \tau^2 r \log m$$

that also holds with probability at least $1 - \frac{1}{2}m^{-2}$. Therefore, with probability at least $1 - m^{-2}$

$$\|\tilde{\rho}^\varepsilon - \rho\|_2^2 \lesssim \left(\frac{1}{l} + \tau^2 l \log m\right) \bigwedge \tau^2 r \log m. \tag{2.3.54}$$

Let $\bar{l} = \frac{1}{\tau\sqrt{\log m}}$. If $\bar{l} \in [1, m]$, set $l := [\bar{l}]$. Otherwise, if $\bar{l} > m$, set $l := m$ and, if $\bar{l} < 1$, set $l := 1$. An easy computation shows that with such a choice of l bound (2.3.54) implies (2.3.51) for $p = 2$.

Next we use bound (2.3.29) that, for $t = 2\log m$, implies under assumptions (2.3.49) that with some constant C and with probability at least $1 - m^{-2}$

$$K(\rho\|\tilde{\rho}^\varepsilon) \leq C\sigma_\xi \frac{r m^{3/2}\sqrt{\log m}\log(mn)}{\sqrt{n}}, \tag{2.3.55}$$

which is bound (2.3.53). Bound (2.3.52) also holds in view of inequality (1.3.10).

Now, we prove bound (2.3.51) for $p = 1$ (the bound for $p \in [1, 2]$ will then follow by interpolation). To this end, we will use the following lemma (see Proposition 1 in [53]) that shows that if two density matrices are close in Hellinger distance and one of them is "concentrated around a subspace" L, then another one is also "concentrated around" L.

Lemma 14. *For any $L \subset \mathbb{C}^m$ and all $S_1, S_2 \in \mathcal{S}_m$,*

$$\|\mathcal{P}_L^{\perp} S_1\|_1 \leq 2\|\mathcal{P}_L^{\perp} S_2\|_1 + 2H^2(S_1, S_2).$$

We apply this lemma to $S_1 = \tilde{\rho}^{\varepsilon}$, $S_2 = \rho$ and $L = \mathrm{supp}(\rho)$ so that $\mathcal{P}_L^{\perp}\rho = 0$. It yields that

$$\|\mathcal{P}_L^{\perp}\tilde{\rho}^{\varepsilon}\|_1 \leq 2H^2(\tilde{\rho}^{\varepsilon}, \rho).$$

Therefore,

$$\|\tilde{\rho}^{\varepsilon}-\rho\|_1 \leq \|\mathcal{P}_L(\tilde{\rho}^{\varepsilon}-\rho)\|_1+\|\mathcal{P}_L^{\perp}(\tilde{\rho}^{\varepsilon}-\rho)\|_1 \leq \sqrt{2r}\|\tilde{\rho}^{\varepsilon}-\rho\|_2+\|\mathcal{P}_L^{\perp}\tilde{\rho}^{\varepsilon}\|_1 \leq \sqrt{2r}\|\tilde{\rho}^{\varepsilon}-\rho\|_2+2H^2(\tilde{\rho}^{\varepsilon}, \rho). \tag{2.3.56}$$

Using bounds (2.3.51) for $p = 2$ and (2.3.52), we get from (2.3.56) that

$$\|\tilde{\rho}^{\varepsilon} - \rho\|_1 \leq C\frac{\sigma_{\xi} m^{\frac{3}{2}} r}{\sqrt{n}}\sqrt{\log m}\log(mn) \bigwedge 2, \tag{2.3.57}$$

which is equivalent to (2.3.51) for $p = 1$. Note that by choosing $t = 2\log m + \log 2 + 2$ (which might have an impact only on the constant), we could make probability bounds in (2.3.51) for $p = 2$ and (2.3.52) to be at least $1 - \frac{1}{2}m^{-2}$ implying that (2.3.57) holds with probability at least $1 - m^{-2}$, as it is claimed in the theorem.

To complete the proof, it is enough to use the interpolation inequality of Lemma 6. It follows that, for $p \in (1, 2)$,

$$\|\tilde{\rho}^{\varepsilon} - \rho\|_p \leq \|\tilde{\rho}^{\varepsilon} - \rho\|_1^{\frac{2}{p}-1}\|\tilde{\rho}^{\varepsilon} - \rho\|_2^{2-\frac{2}{p}}.$$

Substituting bound (2.3.51) for $p = 1$ and $p = 2$ into the last inequality yields the result for an arbitrary $p \in (1, 2)$. □

Similarly, in the case of trace regression with bounded response (see Assumption 3), minimax rates of Theorem 5 are also attained for the estimator $\tilde{\rho}^{\varepsilon}$ (up to log factors). In this case, assume that Assumption 3 holds with $\bar{U} = U$ and, in addition, let us make the following simplifying assumptions:

$$U\sqrt{\frac{m\log m}{n}} \lesssim 1 \quad \text{and} \quad \log\log_2 n \lesssim m\log m. \tag{2.3.58}$$

For the Pauli basis ($U = m^{-1/2}$), the first assumption holds if $n \gtrsim \log m$. The second assumption does hold unless n is extremely large ($n \sim 2^{\exp\{m\log m\}}$). Under these assumptions, we will use the following value of regularization parameter ε :

$$\varepsilon := \frac{U}{\log(mn)}\sqrt{\frac{\log(2m)}{nm}}.$$

The following version of Theorem 10 holds in the bounded regression case (with a similar proof).

Theorem 11. *There exists a constant $C > 0$ such that the following bounds hold for all $r = 1, \dots, m$, for all $\rho \in \mathcal{S}_{r,m}$ and for all $p \in [1, 2]$ with probability at least $1 - m^{-2}$:*

$$\|\tilde{\rho}^{\varepsilon} - \rho\|_p \leq C\biggl(\frac{Um^{\frac{3}{2}}r^{1/p}}{\sqrt{n}}\sqrt{\log m}\log^{(2-p)/p}(mn)\bigwedge\Bigl(\frac{Um^{3/2}}{\sqrt{n}}\Bigr)^{1-\frac{1}{p}}(\log m)^{\frac{1}{2}-\frac{1}{2p}}\biggr)\bigwedge 2, \tag{2.3.59}$$

$$H^2(\tilde{\rho}^{\varepsilon}, \rho) \leq C\frac{Um^{\frac{3}{2}}r}{\sqrt{n}}\sqrt{\log m}\log(mn)\bigwedge 2 \tag{2.3.60}$$

and

$$K(\rho\|\tilde{\rho}^{\varepsilon}) \leq C\frac{Um^{\frac{3}{2}}r}{\sqrt{n}}\sqrt{\log m}\log(mn). \tag{2.3.61}$$

Remark 5. *In the case of Pauli basis, the minimax optimal rates (up to constants and logarithmic factors) are: $\frac{mr^{1/p}}{\sqrt{n}} \wedge (\frac{m}{\sqrt{n}})^{1-\frac{1}{p}} \wedge 2$ for Schatten p-norm distances for $p \in [1, 2]$; $\frac{mr}{\sqrt{n}}$ for nuclear norm, squared Hellinger and Kullback-Leibler distances (provided the $mr \lesssim \sqrt{n}$).*

2.4 The projection estimator and Schatten p-norm convergence rates

Our main goal in this section is to study a minimal distance estimator $\check{\rho}$ of ρ (initially proposed in [54]) defined as the projection of a simple unbiased estimator

$$\hat{Z} = \frac{m^2}{n}\sum_{j=1}^{n} Y_j X_j$$

onto the convex set of density matrices $\mathcal{S}_m$, see also (1.3.7) in Section 1.3.3. We show that the minimax error rates established in Section 2.2 for the classes of low rank density matrices are attained for this estimator up to logarithmic factors *in the whole range of Schatten p-norm distances* for $p \in [1, \infty]$ as well as for Bures and relative entropy distance. The proof of these results relies on simple properties of projections of Hermitian matrices onto the convex set $\mathcal{S}_m$ of density matrices (see theorems 15 and 16) that might be of independent interest.

For the model of uniform sampling from an orthonormal basis $\mathcal{E} = \{E_1, \dots, E_{m^2}\}$, the following simple estimator of unknown state $\rho \in \mathcal{S}_m$ is unbiased:

$$\hat{Z} := \frac{m^2}{n}\sum_{j=1}^{n} Y_j X_j.$$

Indeed,

$$\begin{aligned}\mathbb{E}_\rho \hat{Z} &= m^2 \mathbb{E}_\rho (YX) = m^2 \mathbb{E}(\mathbb{E}_\rho(Y|X)X) = m^2 \mathbb{E}\mathrm{tr}(\rho X)X \\ &= m^2 \mathbb{E}\langle \rho, X\rangle X = m^2 \frac{1}{m^2}\sum_{j=1}^{m^2} \langle \rho, E_j\rangle E_j = \rho.\end{aligned}$$

Clearly, $\hat{Z}$ is not necessarily a density matrix.

2.4.1 Schatten p-norm convergence rates of the projection estimator

We will now define the minimal distance estimator $\check{\rho}$ as the projection of $\hat{Z}$ onto the convex set $\mathcal{S}_m$ of all density matrices. More precisely, for an arbitrary $Z \in \mathbb{H}_m$, define

$$\pi_{\mathcal{S}_m}(Z) := \mathrm{argmin}_{S \in \mathcal{S}_m} \|Z - S\|_2^2. \tag{2.4.1}$$

Clearly, $\pi_{\mathcal{S}_m}(Z)$ is the closest density matrix to Z with respect to the Hilbert–Schmidt norm distance (that is, the projection of Z onto $\mathcal{S}_m$; such a closest density matrix exists in view of compactness of $\mathcal{S}_m$ and it is unique in view of strict convexity of $S \mapsto \|Z - S\|_2^2$). Let

$$\check{\rho} := \pi_{\mathcal{S}_m}(\hat{Z}), \tag{2.4.2}$$

which is actually the modified least square estimator $\check{\rho}_\varepsilon$ (1.3.7) for any $\varepsilon > 0$ introduced in Section 1.3.3.

We will show that the upper bounds on the error rates in Schatten p-norm distances for $p \in [1, \infty]$ and in Bures distance that match the minimax lower bounds of Theorems 4, 5 and 7 in Section 2.2 up to logarithmic factors hold for the estimator $\check{\rho}$. We will then introduce a simple modification of this estimator for which a matching upper bound holds also for Kullback-Leibler distance.

First, we consider the case of Gaussian trace regression model (Assumption 4). We need an additional assumption that $\sigma_\xi \geq \frac{U}{m^{1/2}}$ (the variance of the noise is not too small).

Theorem 12. *Suppose Assumption 4 holds and $\sigma_\xi \geq \frac{U}{m^{1/2}}$. For all $p \in [1, +\infty]$, there exists a constant $C > 0$ such that, for all $A \geq 1$ the following bounds hold:*

$$\sup_{\rho \in \mathcal{S}_{r,m}} \mathbb{P}_\rho \left\{ \|\check{\rho} - \rho\|_p \geq C\left(r^{1/p} \frac{\sigma_\xi m^{\frac{3}{2}} \sqrt{A \log(2m)}}{\sqrt{n}} \bigwedge \left(\frac{\sigma_\xi m^{3/2} \sqrt{A \log(2m)}}{\sqrt{n}} \right)^{1 - \frac{1}{p}} \bigwedge 1 \right) \right\} \leq (2m)^{-A} \tag{2.4.3}$$

and

$$\sup_{\rho \in \mathcal{S}_{r,m}} \mathbb{P}_\rho \left\{ H^2(\check{\rho}, \rho) \geq C\left(r \frac{\sigma_\xi m^{\frac{3}{2}} \sqrt{A \log(2m)}}{\sqrt{n}} \bigwedge 1 \right) \right\} \leq (2m)^{-A}. \tag{2.4.4}$$

If $\sigma_\xi < \frac{U}{m^{1/2}}$, the bounds still hold with σ_ξ replaced by $\frac{U}{m^{1/2}}$.

Similarly, in the case of trace regression with a bounded response, the following result holds.

Theorem 13. *Suppose Assumption 3 is satisfied. Then, for all* $p \in [1, +\infty]$, *there exists a constant* $C > 0$ *such that, for all* $A \geq 1$, *the following bounds hold:*

$$\sup_{P \in \mathcal{P}_{r,m}(U)} \mathbb{P}_P \biggl\{ \|\check{\rho} - \rho_P\|_p \geq C \biggl(r^{1/p} \frac{U m^{\frac{3}{2}} \sqrt{A \log(2m)}}{\sqrt{n}} \bigwedge \biggl(\frac{U m^{3/2} \sqrt{A \log(2m)}}{\sqrt{n}} \biggr)^{1 - \frac{1}{p}} \bigwedge 1 \biggr) \biggr\} \leq (2m)^{-A} \tag{2.4.5}$$

and

$$\sup_{P \in \mathcal{P}_{r,m}(U)} \mathbb{P}_P \biggl\{ H^2(\check{\rho}, \rho_P) \geq C \biggl(r \frac{U m^{\frac{3}{2}} \sqrt{A \log(2m)}}{\sqrt{n}} \bigwedge 1 \biggr) \biggr\} \leq (2m)^{-A}. \tag{2.4.6}$$

For completeness, we state also the upper bounds in the case of Pauli measurements (that immediately follow from Theorem 13).

Theorem 14. *Suppose the assumptions of Theorem 7 hold. Then, for all* $p \in [1, +\infty]$, *there exists a constant* C *such that, for all* $A \geq 1$, *the following bounds hold:*

$$\sup_{\rho \in \mathcal{S}_{r,m}} \mathbb{P}_\rho \biggl\{ \|\check{\rho} - \rho\|_p \geq C \biggl(r^{1/p} \frac{m \sqrt{A \log(2m)}}{\sqrt{n}} \bigwedge \biggl(\frac{m \sqrt{A \log(2m)}}{\sqrt{n}} \biggr)^{1 - \frac{1}{p}} \bigwedge 1 \biggr) \biggr\} \leq (2m)^{-A} \tag{2.4.7}$$

and

$$\sup_{\rho \in \mathcal{S}_{r,m}} \mathbb{P}_\rho \biggl\{ H^2(\check{\rho}, \rho) \geq C \biggl(r \frac{m}{\sqrt{n}} \bigwedge 1 \biggr) \biggr\} \leq (2m)^{-A}. \tag{2.4.8}$$

The proof of these results relies on the following fact that might be of independent interest and that essentially shows that $\pi_{\mathcal{S}_m}(Z)$ is the closest density matrix to Z not only in the Hilbert–Schmidt norm distance, but also in the operator norm distance.

Lemma 15. *For all* $Z \in \mathbb{H}_m$,

$$\|Z - \pi_{\mathcal{S}_m}(Z)\|_\infty = \inf_{S \in \mathcal{S}_m} \|Z - S\|_\infty.$$

The proof of this lemma will be given in Section 2.4.2. Here we use it to establish the next result that is the main ingredient of the proofs of theorems 12, 13 and 14.

Lemma 16. *Let* $p \in [1, +\infty]$. *For all* $Z \in \mathbb{H}_m$ *and all* $S \in \mathcal{S}_{r,m}$,

$$\|\pi_{\mathcal{S}_m}(Z) - S\|_p \leq \min\Bigl(2^{3/p+1} r^{1/p} \|Z - S\|_\infty, 2\|Z - S\|_\infty^{1-1/p} \Bigr).$$

The proof relies on Lemma 15 and on a simple lemma stated below.

Lemma 17. *Let $S, S' \in \mathcal{S}_m$ and* $\mathrm{rank}(S) = r$. *Then, for all* $p \in [1, \infty]$,

$$\|S' - S\|_p \le \min\Bigl((8r)^{1/p}\|S' - S\|_\infty, 2^{1/p}\|S' - S\|_\infty^{1-1/p}\Bigr).$$

Proof. Let $S = \sum_{j=1}^{r} \lambda_j(\phi_j \otimes \phi_j)$ be the spectral decomposition of S with eigenvalues λ_j and eigenvectors ϕ_j. Let $L := \mathrm{supp}(S)$ be the linear span of vectors $\phi_1, \dots, \phi_r \in \mathbb{C}^m$. Denote by $P_L, P_{L^\perp}$ the orthogonal projection operators onto subspace L and its orthogonal complement $L^\perp$, respectively. We will need the following projection operators $\mathcal{P}_L, \mathcal{P}_L^\perp : \mathbb{H}_m \mapsto \mathbb{H}_m$:

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

The following bounds are obvious:

$$\|S\|_1 = 1 = \|S'\|_1 = \|S' - S + S\|_1 = \|\mathcal{P}_L(S' - S) + \mathcal{P}_L^\perp(S' - S) + S\|_1$$

$$\ge \|\mathcal{P}_L^\perp(S' - S) + S\|_1 - \|\mathcal{P}_L(S' - S)\|_1.$$

Since $S = P_L S P_L$, we can use the pinching inequality for unitary invariant norm $\|\cdot\|_1$ (see [9], p. 97) to get:

$$\|\mathcal{P}_L^\perp(S' - S) + S\|_1 = \|P_L S P_L + P_{L^\perp}(S' - S)P_{L^\perp}\|_1$$

$$= \|P_L S P_L\|_1 + \|P_{L^\perp}(S' - S)P_{L^\perp}\|_1 = \|S\|_1 + \|\mathcal{P}_L^\perp(S' - S)\|_1.$$

Therefore,

$$\|S\|_1 \ge \|S\|_1 + \|\mathcal{P}_L^\perp(S' - S)\|_1 - \|\mathcal{P}_L(S' - S)\|_1,$$

implying that

$$\|\mathcal{P}_L^\perp(S' - S)\|_1 \le \|\mathcal{P}_L(S' - S)\|_1.$$

It follows from the last bound that

$$\|S' - S\|_1 = \|\mathcal{P}_L(S' - S) + \mathcal{P}_L^\perp(S' - S)\|_1 \le 2\|\mathcal{P}_L(S' - S)\|_1.$$

Since $\dim(L) = r$, the matrix $\mathcal{P}_L(S' - S)$ is of rank at most $2r$. This implies that

$$\|\mathcal{P}_L(S' - S)\|_1 \leq 2r\|\mathcal{P}_L(S' - S)\|_\infty$$

$$\leq 2r(\|(S' - S)P_L\|_\infty + \|P_L(S' - S)P_{L^\perp}\|_\infty) \leq 4r\|S' - S\|_\infty.$$

Therefore, $\|S' - S\|_1 \leq 8r\|S' - S\|_\infty$, and since also $\|S' - S\|_1 \leq 2, S, S' \in \mathcal{S}_m$, we conclude that

$$\|S' - S\|_1 \leq \min(8r\|S' - S\|_\infty, 2).$$

Together with interpolation inequality this yields that for all $p \in [1, \infty]$

$$\|S' - S\|_p \leq \|S' - S\|_1^{1/p}\|S' - S\|_\infty^{1-1/p} \leq \min\Big((8r)^{1/p}\|S' - S\|_\infty, 2^{1/p}\|S' - S\|_\infty^{1-1/p}\Big).$$

□

Proof. We now prove Lemma 16. It immediately follows from Lemma 15 that, for all $S \in \mathcal{S}_m$,

$$\|\pi_{\mathcal{S}_m}(Z) - S\|_\infty \leq \|\pi_{\mathcal{S}_m}(Z) - Z\|_\infty + \|Z - S\|_\infty \leq 2\|Z - S\|_\infty.$$

If $S \in \mathcal{S}_m$ is a density matrix of rank r, the last bound could be combined with the bound of Lemma 17 to get that for all $p \in [1, +\infty]$

$$\|\pi_{\mathcal{S}_m}(Z) - S\|_p \leq \min\Big(2^{3/p+1}r^{1/p}\|Z - S\|_\infty, 2\|Z - S\|_\infty^{1-1/p}\Big).$$

□

Proof. We now turn to the proof of theorems 12, 13 and 14. To this end, we use the bound of Lemma 16 with $Z = \hat{Z}$ and $S = \rho \in \mathcal{S}_{r,m}$ that yields:

$$\|\check{\rho} - \rho\|_p \leq \min\Big(2^{3/p+1}r^{1/p}\|\hat{Z} - \rho\|_\infty, 2\|\hat{Z} - \rho\|_\infty^{1-1/p}\Big). \tag{2.4.9}$$

The control of

$$\|\hat{Z} - \rho\|_\infty = \left\|\frac{m^2}{n}\sum_{j=1}^n Y_j X_j - \rho\right\|_\infty$$

is based on a standard application of matrix Bernstein type inequalities. We give a detailed argument for completeness. Note that $\|\check{\rho}-\rho\|_p$ in the left-hand side of bound (2.4.9) is upper bounded by 2, so, if Bernstein bound on $\|\hat{Z}-\rho\|_\infty$ is larger than 1 (or even $\gtrsim 1$), it could be replaced by the trivial bound equal to 1. In the case of Theorem 13, we use the version of Bernstein inequality for i.i.d. bounded random matrices, see Lemma 7 in Section 2.3.1.

For $V = YX - \mathbb{E}(YX)$, we get, under Assumption 3, that

$$\sigma^2 = \|\mathbb{E}V^2\|_\infty \leq \|\mathbb{E}(Y^2X^2)\|_\infty \leq U^2\|\mathbb{E}X^2\|_\infty.$$

It is also well known that, under the same assumption, $\|\mathbb{E}X^2\|_\infty = m^{-1}$. [Indeed, if $\{e_j, j = 1, \dots, m\}$ is an orthonormal basis of $\mathbb{C}^m$, then

$$\|\mathbb{E}X^2\|_\infty = \sup_{v\in\mathbb{C}^m,|v|\leq 1} \mathbb{E}\langle X^2 v, v\rangle = \sup_{v\in\mathbb{C}^m,|v|\leq 1} \mathbb{E}|Xv|^2 = \sup_{v\in\mathbb{C}^m,|v|\leq 1} \mathbb{E}\sum_{j=1}^m |\langle Xv, e_j\rangle|^2$$

$$= \sup_{v\in\mathbb{C}^m,|v|\leq 1} \mathbb{E}\sum_{j=1}^m |\langle X, v\otimes e_j\rangle|^2 = \sup_{v\in\mathbb{C}^m,|v|\leq 1} \sum_{j=1}^m m^{-2}\sum_{k=1}^{m^2} |\langle E_k, v\otimes e_j\rangle|^2 = \sup_{v\in\mathbb{C}^m,|v|\leq 1} m^{-2}\sum_{j=1}^m \|v\otimes e_j\|_2^2 =$$

$$\sup_{v\in\mathbb{C}^m,|v|\leq 1} m^{-2}\sum_{j=1}^m |v|^2|e_j|^2 = m^{-1}].$$

We use the bound of Lemma 7 with $t = A\log(2m), A \geq 1$ to get that with probability at least $1-(2m)^{-A}$,

$$\left\|\frac{m^2}{n}\sum_{j=1}^n Y_jX_j - \rho\right\|_\infty \leq C\left[Um^{3/2}\sqrt{\frac{A\log(2m)}{n}} \bigvee \frac{U^2m^2A\log(2m)}{n}\right]$$

with some absolute constant $C \geq 1$. If

$$\frac{U^2m^2A\log(2m)}{n} \geq Um^{3/2}\sqrt{\frac{A\log(2m)}{n}},$$

then $Um^{1/2}\sqrt{\frac{A\log(2m)}{n}} \geq 1$ implying that $Um^{3/2}\sqrt{\frac{A\log(2m)}{n}} \geq 1$. Thus, when the bound on $\|\hat{Z}-\rho\|_\infty$ is substituted in bound (2.4.9), it is enough to keep only the first term $Um^{3/2}\sqrt{\frac{A\log(2m)}{n}}$, the second term could be dropped. This implies that with some constant $C' > 0$ (that does not depend on $\rho \in \mathcal{S}_{r,m}$) the inequality

$$\|\check{\rho}-\rho\|_p \leq C'\biggl(r^{1/p}\frac{Um^{\frac{3}{2}}\sqrt{A\log(2m)}}{\sqrt{n}} \bigwedge \biggl(\frac{Um^{3/2}\sqrt{A\log(2m)}}{\sqrt{n}}\biggr)^{1-\frac{1}{p}} \bigwedge 1\biggr)$$

holds with probability at least $1-(2m)^{-A}$, implying the first bound of Theorem 13. The second bound immediately follows from the inequality $H^2(\check{\rho},\rho)\leq \|\check{\rho}-\rho\|_1$ (see Lemma 1). Theorem 14 is an immediate consequence of Theorem 13.

The proof of Theorem 12 is very similar. In this case, Assumption 4 holds and it is natural to split $\hat{Z}-\rho$ into two parts

$$\hat{Z}-\rho = \frac{m^2}{n}\sum_{j=1}^{n}\langle \rho, X_j\rangle X_j - \rho + \frac{m^2}{n}\sum_{j=1}^{n}\xi_j X_j. \tag{2.4.10}$$

and to bound $\|\hat{Z}-\rho\|_\infty$ by triangle inequality. For the first part, an application of matrix Bernstein inequality of Lemma 7 yields the bound

$$\left\|\frac{m^2}{n}\sum_{j=1}^{n}\langle \rho, X_j\rangle X_j - \rho\right\|_\infty \leq C\left[Um\sqrt{\frac{A\log(2m)}{n}}\bigvee \frac{U^2m^2A\log(2m)}{n}\right] \tag{2.4.11}$$

that holds for some absolute constant $C\geq 1$ with probability at least $1-(2m)^{-A}$. Indeed, in this case $V=\langle \rho, X\rangle X - \mathbb{E}\langle \rho, X\rangle X$ and

$$\sigma^2 \leq \|\mathbb{E}\langle \rho, X\rangle^2 X^2\|_\infty \leq U^2\mathbb{E}\langle \rho, X\rangle^2 = \frac{U^2\|\rho\|_2^2}{m^2}\leq \frac{U^2}{m^2},$$

$$\|\langle \rho, X\rangle X\|_\infty \leq \|\rho\|_1\|X\|_\infty^2 \leq \|X\|_\infty^2 \leq U^2,$$

and Lemma 7 implies (2.4.11). As before, if $\frac{U^2m^2A\log(2m)}{n}\geq Um\sqrt{\frac{A\log(2m)}{n}}$, then $Um\sqrt{\frac{A\log(2m)}{n}}\geq 1$. Thus, the second term $\frac{U^2m^2A\log(2m)}{n}$ could be dropped when the bound on $\|\hat{Z}-\rho\|_\infty$ (for which the right hand side of (2.4.11) is a part) is substituted in (2.4.9).

As to the second part of representation (2.4.10) that involves normal random variables ξ_j, it is bounded using another version of matrix Bernstein inequality for not necessarily bounded random variables (see [53], [52], [55]).

We apply the bound of Lemma 8 in the case when $V:=\xi X, \alpha=2$ and $t=A\log(2m)$ for $A\geq 1$. By an easy computation,

$$\sigma^2 = \sigma_\xi^2\|\mathbb{E}X^2\|_\infty = \frac{\sigma_\xi^2}{m}$$

and

$$U^{(2)} = 2\big\|\xi\|X\|_\infty\big\|_{\psi_2} \leq 2U\|\xi\|_{\psi_2} \leq 4\sigma_\xi U.$$

This yields the following bound

$$\left\|\frac{m^2}{n}\sum_{j=1}^n \xi_j X_j\right\|_\infty \leq C\left[\sigma_\xi m^{3/2}\sqrt{\frac{A\log(2m)}{n}}\bigvee \sigma_\xi U\frac{m^2 A\log(2m)\log^{1/2}(4U\sqrt{m})}{n}\right] \tag{2.4.12}$$

that holds with probability at least $1-(2m)^{-A}$ and with some absolute constant $C \geq 1$. If the second term in the maximum in the right hand side of (2.4.12) is dominant, then $Um^{1/2}\sqrt{\frac{A\log(2m)}{n}}\log^{1/2}(4U\sqrt{m}) \geq 1$. Under the condition that $\sigma_\xi \geq Um^{-1/2}$, this implies that also $\sigma_\xi m^{3/2}\sqrt{\frac{A\log(2m)}{n}} \gtrsim 1$. Thus, when the bound in the right hand side of (2.4.12) (used to control $\|\hat{Z}-\rho\|_\infty$) is substituted in (2.4.9), it is enough to keep only the first term in the maximum. Finally, under the assumption $\sigma_\xi \geq Um^{-1/2}$, the first term of bound (2.4.12) dominates the first term of (2.4.11), so, only this term is needed to control $\|\hat{Z}-\rho\|_\infty$ in bound (2.4.9). These considerations imply the bound

$$\|\check{\rho}-\rho\|_p \leq C'\left(r^{1/p}\frac{\sigma_\xi m^{\frac{3}{2}}\sqrt{A\log(2m)}}{\sqrt{n}}\bigwedge\left(\frac{\sigma_\xi m^{3/2}\sqrt{A\log(2m)}}{\sqrt{n}}\right)^{1-\frac{1}{p}}\bigwedge 1\right)$$

that holds with some constant $C' > 0$ (that does not depend on $\rho \in \mathcal{S}_{r,m}$) and with probability at least $1-(2m)^{-A}$. The first bound of Theorem 12 now follows for all $p \in [1,\infty]$ (which also implies the second bound in view of Lemma 1.

□

It turns out that for a slightly modified version of estimator $\check{\rho}$, minimax lower bounds are also attained (up to logarithmic factors) in the case of Kullback-Leibler distance. For $S \in \mathcal{S}_m$ and $\delta \in [0,1]$, define $S_\delta = (1-\delta)S + \delta\frac{I_m}{m}$. Clearly, $S_\delta \in \mathcal{S}_m$. Let $\mathcal{S}_{m,\delta} := \{S_\delta : S \in \mathcal{S}_m\}$. Define $\pi_{\mathcal{S}_{m,\delta}}(Z)$ the projection of $Z \in \mathbb{H}_m$ onto the convex set $\mathcal{S}_{m,\delta}$:

$$\pi_{\mathcal{S}_{m,\delta}}(Z) := \mathrm{argmin}_{S\in\mathcal{S}_{m,\delta}}\|Z-S\|_2^2.$$

Let

$$\check{\rho}_\delta := \pi_{\mathcal{S}_{m,\delta}}(\hat{Z})$$

with $\check{\rho}_0 = \check{\rho}$. We will prove the following versions of theorems 12, 13 and 14 for the estimator $\check{\rho}_\delta$.

Theorem 15. *Suppose Assumption 4 holds,* $\sigma_\xi \geq \frac{U}{m^{1/2}}$ *and*

$$\delta \leq \frac{\sigma_\xi m^{\frac{3}{2}}\sqrt{\log(2m)}}{\sqrt{n}} \bigwedge 1.$$

Then bounds (2.4.3) and (2.4.4) hold for estimator $\check{\rho}_\delta$. *Moreover, for* $A \geq 1$, *define*

$$\lambda := \frac{r\sigma_\xi m^{5/2}\sqrt{\frac{A\log(2m)}{n}} \bigwedge m}{\delta}.$$

Then, for some constant $c > 0$,

$$\sup_{\rho \in \mathcal{S}_{r,m}} \mathbb{P}_\rho \left\{ K(\rho \| \check{\rho}_\delta) \geq c\left(r\frac{\sigma_\xi m^{\frac{3}{2}}\sqrt{A\log(2m)}}{\sqrt{n}} \bigwedge 1\right) \log(1 + c\lambda) \right\} \leq (2m)^{-A}. \quad (2.4.13)$$

If $\sigma_\xi < \frac{U}{m^{1/2}}$, *the bounds still hold with* σ_ξ *replaced by* $\frac{U}{m^{1/2}}$.

Theorem 16. *Suppose Assumption 3 is satisfied and*

$$\delta \leq \frac{Um^{\frac{3}{2}}\sqrt{\log(2m)}}{\sqrt{n}} \bigwedge 1.$$

Then (2.4.5) and (2.4.6) hold for estimator $\check{\rho}_\delta$. *Moreover, for* $A \geq 1$, *define*

$$\lambda := \frac{rUm^{5/2}\sqrt{\frac{A\log(2m)}{n}} \bigwedge m}{\delta}.$$

Then, for some constant $c > 0$,

$$\sup_{\rho \in \mathcal{S}_{r,m}} \mathbb{P}_\rho \left\{ K(\rho \| \check{\rho}_\delta) \geq c\left(r\frac{Um^{\frac{3}{2}}\sqrt{A\log(2m)}}{\sqrt{n}} \bigwedge 1\right) \log(1 + c\lambda) \right\} \leq (2m)^{-A}. \quad (2.4.14)$$

Theorem 17. *Suppose the assumptions of Theorem 7 hold and*

$$\delta \leq \frac{m\sqrt{\log(2m)}}{\sqrt{n}} \bigwedge 1.$$

Then (2.4.7) and (2.4.8) hold for estimator $\check{\rho}_\delta$. *Moreover, for* $A \geq 1$, *define*

$$\lambda := \frac{rm^2\sqrt{\frac{A\log(2m)}{n}} \bigwedge m}{\delta}.$$

Then, for some constant $c > 0$,

$$\sup_{\rho \in \mathcal{S}_{r,m}} \mathbb{P}_\rho \left\{ K(\rho \| \check{\rho}_\delta) \geq c\left(r\frac{m\sqrt{A\log(2m)}}{\sqrt{n}} \bigwedge 1\right) \log(1 + c\lambda) \right\} \leq (2m)^{-A}. \quad (2.4.15)$$

Remark 6. *If, under the assumptions of Theorem 16, we choose*

$$\delta = \frac{Um^{\frac{3}{2}}\sqrt{\log(2m)}}{\sqrt{n}} \bigwedge 1,$$

then the logarithmic factor in bound (2.4.14) satisfies the inequality

$$\log(1 + c\lambda) \leq \log(1 + crm\sqrt{A}),$$

so it is of the order $\log m$. *Under the assumptions of Theorem 15, this would require the choice of* δ

$$\delta = \frac{\sigma_\xi m^{\frac{3}{2}}\sqrt{\log(2m)}}{\sqrt{n}} \bigwedge 1,$$

so δ *would depend on an unknown parameter* σ_ξ. *Replacing* σ_ξ *in the definition of* δ *by the lower bound* $Um^{-1/2}$ *would result in a logarithmic factor* $\lesssim \log\biggl(1 + crm\sqrt{A}\frac{\sigma_\xi}{Um^{-1/2}}\biggr)$.

Proof. We start with the following modification of Theorem 16.

Lemma 18. *Let* $p \in [1, \infty]$. *For all* $Z \in \mathbb{H}_m$ *and all* $S \in \mathcal{S}_{r,m}$, *the following bound holds:*

$$\|\pi_{\mathcal{S}_{m,\delta}}(Z) - S\|_p \leq \min\biggl(2^{3/p+1}r^{1/p}\Bigl(\|Z - S\|_\infty + 2\delta\Bigr), 2(1-\delta)^{1/p}\Bigl(\|Z - S\|_\infty + 2\delta\Bigr)^{1-1/p}\biggr) + 2\delta.$$

Proof. The following formula is straightforward: for $\delta \in [0, 1)$,

$$\pi_{\mathcal{S}_{m,\delta}}(Z) = (1-\delta)\pi_{\mathcal{S}_m}\left(\frac{Z}{1-\delta} - \frac{\delta}{1-\delta}\frac{I_m}{m}\right) + \delta\frac{I_m}{m}.$$

Indeed, $\pi_{\mathcal{S}_{m,\delta}}(Z)$ coincides with $(1-\delta)S' + \delta\frac{I_m}{m}$, where

$$S' := \operatorname{argmin}_{S \in \mathcal{S}_m}\left\|Z - (1-\delta)S - \delta\frac{I_m}{m}\right\|_2^2$$

$$= \operatorname{argmin}_{S \in \mathcal{S}_m}\left\|\frac{Z}{1-\delta} - \frac{\delta}{1-\delta}\frac{I_m}{m} - S\right\|_2^2 = \pi_{\mathcal{S}_m}\left(\frac{Z}{1-\delta} - \frac{\delta}{1-\delta}\frac{I_m}{m}\right),$$

implying the claim.

Let $S \in \mathcal{S}_{r,m}$. Then, for $p \in [1, \infty]$,

$$\|\pi_{\mathcal{S}_{m,\delta}}(Z) - S\|_p \leq \|\pi_{\mathcal{S}_{m,\delta}}(Z) - S_\delta\|_p + \|S_\delta - S\|_p \tag{2.4.16}$$
$$\leq (1-\delta)\left\|\pi_{\mathcal{S}_m}\left(\tfrac{Z}{1-\delta} - \tfrac{\delta}{1-\delta}\tfrac{I_m}{m}\right) - S\right\|_p + 2\delta.$$

To control the first term in the right hand side, we use the bound of Theorem 16, which requires bounding $\left\|\frac{Z}{1-\delta} - \frac{\delta}{1-\delta}\frac{I_m}{m} - S\right\|_\infty$. We have

$$\left\|\tfrac{Z}{1-\delta} - \tfrac{\delta}{1-\delta}\tfrac{I_m}{m} - S\right\|_\infty = \tfrac{1}{1-\delta}\|Z - S_\delta\|_\infty \tag{2.4.17}$$
$$\leq \tfrac{1}{1-\delta}\|Z - S\|_\infty + \tfrac{1}{1-\delta}\|S - S_\delta\|_\infty \leq \tfrac{1}{1-\delta}\|Z - S\|_\infty + \tfrac{2\delta}{1-\delta}.$$

Using bounds (2.4.16), (2.4.17) along with the bound of Theorem 16, we get the bound of the lemma.

□

We will use the bound of Lemma 18 to control $\|\check{\rho}_\delta - \rho\|_p$ for $\rho \in \mathcal{S}_{r,m}$. To this end, we need to bound $\|\hat{Z} - \rho\|_\infty$ using matrix Bernstein inequalities exactly as it was done in the proof of theorems 12, 13 and 14 (under assumptions of these theorems). Denote by $\bar{\Delta}$ such an upper bound on $\|\hat{Z} - \rho\|_\infty$ that holds with probability a least $1 - (2m)^{-A}$. Recall that $\bar{\Delta} \asymp \sigma_\xi m^{3/2}\sqrt{\frac{A\log(2m)}{n}}$ under the conditions of Theorem 12 and $\bar{\Delta} \asymp U m^{3/2}\sqrt{\frac{A\log(2m)}{n}}$ under the conditions of Theorem 13 (it is the same under the conditions of Theorem 14 with $U = m^{-1/2}$). Setting $\Delta = \bar{\Delta} \wedge 1$, we get from the bound of Lemma 18 that

$$\|\check{\rho}_\delta - \rho\|_p \leq \min\left(2^{3/p+1}r^{1/p}\left(\Delta + 2\delta\right), 2(1-\delta)^{1/p}\left(\Delta + 2\delta\right)^{1-1/p}\right) + 2\delta$$

that holds with the same probability at least $1 - (2m)^{-A}$. Recall that we replace $\bar{\Delta}$ by Δ since the left hand side $\|\check{\rho}_\delta - \rho\|_p \leq 2$; for the same reason, we can and do drop the "exponential parts" of matrix Bernstein bounds leaving in the definition of Δ only the "Gaussian parts". For $\delta \lesssim \Delta$, we get

$$\|\check{\rho}_\delta - \rho\|_p \lesssim \min(r^{1/p}\Delta, \Delta^{1-1/p}).$$

Exactly as in the proof of theorems 12, 13 and 14, this implies that bounds (2.4.3), (2.4.4), (2.4.5), (2.4.6), (2.4.7) and (2.4.8) hold for estimator $\check{\rho}_\delta$.

The bound on the Kullback-Leibler divergence $K(\rho\|\check{\rho}_\delta)$ is an immediate consequence of the bound on $\|\check{\rho}_\delta - \rho\|_1$ and the next lemma that follows from Corollary 1 in [5].

Lemma 19. *Let $S_1, S_2 \in \mathcal{S}_m$ be density matrices and let $\beta := \lambda_{\min}(S_2)$ be the smallest eigenvalue of S_2. Suppose that $\beta > 0$. Then*

$$K(S_1\|S_2) \le \|S_1 - S_2\|_1 \log\biggl(1 + \frac{\|S_1 - S_2\|_1}{2\beta}\biggr).$$

We apply Lemma 19 to $S_1 = \rho, S_2 = \check{\rho}_\delta$, observing that $\check{\rho}_\delta \in \mathcal{S}_{m,\delta}$ and $\lambda_{\min}(\check{\rho}_\delta) \ge \delta/m$. We then use the bound on $\|\check{\rho}_\delta - \rho\|_1$ to complete the proof of the bound on $K(\rho\|\check{\rho}_\delta)$.

□

We conclude this section with a simple result concerning the least squares estimator $\hat{\rho}$ defined as

$$\hat{\rho} := \operatorname*{arg\,min}_{S\in\mathcal{S}_m} \frac{1}{n}\sum_{i=1}^{n}\bigl(Y_i - \langle S, X_i\rangle\bigr)^2, \tag{2.4.18}$$

see also (2.3.1) in Section 2.3 when $\varepsilon = 0$. It shows that the estimators $\hat{\rho}$ and $\check{\rho}$ are close in the Hilbert-Schmidt norm. As a result, the bounds of the previous theorems could be applied to estimator $\hat{\rho}$ as well (at least, under some additional assumptions).

Theorem 18. *Under the assumption that i.i.d. design variables $X_1, \dots, X_n$ are sampled from the uniform distribution Π in an orthonormal basis $\mathcal{E} = \{E_1, \dots, E_{m^2}\}$, the following bound holds with some constant $C > 0$ for all $A \ge 1$ with probability at least $1 - (2m^2)^{-A}$:*

$$\|\check{\rho} - \hat{\rho}\|_2 \le Cm\sqrt{\frac{A\log(2m)}{n}}.$$

Proof. Note that the gradient (and subgradient) of convex function $S \mapsto \|S - Z\|_2^2$ is equal to $2(S - Z)$. By a necessary condition of minimum in convex minimization problem (2.4.1), for $\check{\rho} = \pi_{\mathcal{S}_m}(\hat{Z})$, $\hat{Z} - \check{\rho}$ should belong to the normal cone $N_{\mathcal{S}_m}(\check{\rho})$ of the convex set $\mathcal{S}_m$ at point $\check{\rho}$ (see [3], Proposition 5, Chapter 4, Section 1). Since both $\check{\rho}, \hat{\rho} \in \mathcal{S}_m$, this implies that

$$\langle \check{\rho} - Z, \check{\rho} - \hat{\rho} \rangle \le 0. \tag{2.4.19}$$

Similar analysis of convex optimization problem (2.4.18) shows that

$$\left\langle \frac{m^2}{n} \sum_{j=1}^{n} (\langle \hat{\rho}, X_j \rangle - Y_j) X_j, \check{\rho} - \hat{\rho} \right\rangle \ge 0,$$

which could be rewritten as follows:

$$\left\langle \frac{m^2}{n} \sum_{j=1}^{n} \langle \hat{\rho}, X_j \rangle X_j - Z, \check{\rho} - \hat{\rho} \right\rangle \ge 0. \tag{2.4.20}$$

Subtracting (2.4.20) from (2.4.19) yields

$$\left\langle \check{\rho} - \frac{m^2}{n} \sum_{j=1}^{n} \langle \hat{\rho}, X_j \rangle X_j, \check{\rho} - \hat{\rho} \right\rangle \le 0,$$

implying that

$$\|\check{\rho} - \hat{\rho}\|_2^2 = \langle \check{\rho} - \hat{\rho}, \check{\rho} - \hat{\rho} \rangle \le \left\langle \frac{m^2}{n} \sum_{j=1}^{n} \langle \hat{\rho}, X_j \rangle X_j - \hat{\rho}, \check{\rho} - \hat{\rho} \right\rangle. \tag{2.4.21}$$

We will now write [2]

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

$$= m^2 \left[\frac{1}{n} \sum_{j=1}^{n} (X_j \otimes X_j - \mathbb{E}(X \otimes X)) \right] \hat{\rho}.$$

It follows from (2.4.21) that

$$\|\check{\rho} - \hat{\rho}\|_2^2 \le m^2 \left\| \frac{1}{n} \sum_{j=1}^{n} X_j \otimes X_j - \mathbb{E}(X \otimes X) \right\|_{\text{op}} \|\hat{\rho}\|_2 \|\check{\rho} - \hat{\rho}\|_2.$$

[2] Here we view the tensor product $A \otimes B$ of operators $A, B \in \mathbb{M}_m$ as an operator acting from the space $\mathbb{M}_m$ of $m \times m$ matrices equipped with Hilbert-Schmidt inner product $\langle \cdot, \cdot \rangle$ into itself as follows: $(A \otimes B)C = A\langle C, B \rangle$. Let $\| \cdot \|_{\text{op}}$ denote the operator norm of linear operators from $\mathbb{M}_m$ into itself, which corresponds to the $\| \cdot \|_\infty$ in the case of $m \times m$ matrices.

Since $\|\hat{\rho}\|_2 \leq 1$, we get

$$\|\check{\rho} - \hat{\rho}\|_2 \leq m^2 \left\| \frac{1}{n} \sum_{j=1}^{n} X_j \otimes X_j - \mathbb{E}(X \otimes X) \right\|_{\mathrm{op}}. \tag{2.4.22}$$

It remains to control the operator norm in the right hand side for which we can again use matrix Bernstein inequality of Lemma 7 applying it to $V = X \otimes X - \mathbb{E}(X \otimes X)$. In this case,

$$\sigma^2 = \|\mathbb{E}V^2\|_{\mathrm{op}} \leq \|\mathbb{E}(X \otimes X)^2\|_{\mathrm{op}} = \sup_{\|U\|_2 \leq 1} \mathbb{E}\langle (X \otimes X)^2 U, U \rangle = \sup_{\|U\|_2 \leq 1} \mathbb{E}\langle (X \otimes X) U, (X \otimes X) U \rangle$$

$$= \sup_{\|U\|_2 \leq 1} \mathbb{E}|\langle U, X \rangle|^2 \|X\|_2^2 \leq \sup_{\|U\|_2 \leq 1} \mathbb{E}|\langle U, X \rangle|^2 = \sup_{\|U\|_2 \leq 1} \frac{\|U\|_2^2}{m^2} = \frac{1}{m^2}$$

and

$$\|V\|_{\mathrm{op}} \leq \|X \otimes X\|_{\mathrm{op}} + \mathbb{E}\|X \otimes X\|_{\mathrm{op}} = \|X\|_2^2 + \mathbb{E}\|X\|_2^2 \leq 2.$$

Bound (2.4.22) along with the bound of Lemma 7 with $t = A \log(2m^2), A \geq 1$ yield the following inequality

$$\|\check{\rho} - \hat{\rho}\|_2 \lesssim m \sqrt{\frac{A \log(2m)}{n}} \bigvee m^2 \frac{A \log(2m)}{n}$$

that holds with probability at least $1 - (2m^2)^{-A}$. Since $\|\check{\rho} - \hat{\rho}\|_2 \leq 2$, the second term $m^2 \frac{A \log(2m)}{n}$ in the right hand side could be dropped (if this term is dominant, the bound is $\gtrsim 1$). This completes the proof of the theorem. □

Since $\|\check{\rho} - \hat{\rho}\|_\infty \leq \|\check{\rho} - \hat{\rho}\|_2$, the bound of Theorem 18 also holds for $\|\check{\rho} - \hat{\rho}\|_\infty$. Combining this with the bound of Theorem 13 for $p = \infty$, it is easy to conclude that under conditions of this theorem

$$\|\hat{\rho} - \rho\|_\infty \lesssim U m^{3/2} \sqrt{\frac{A \log(2m)}{n}}$$

and that the last bound holds (with a proper choice of constant in relationship $\lesssim$) with probability at least $1 - (2m)^{-A}$. In view of Lemma 17, this immediately implies that all the bounds of Theorem 13 also hold for the least squares estimator $\hat{\rho}$. In a special case of Pauli measurements, this means that Theorem 14 holds for the estimator $\hat{\rho}$. Concerning Theorem 15, the same conclusion is true under the additional assumption that $\sigma_\xi \geq m^{-1/2}$. Moreover, if $\hat{\rho}_\delta$ is the following modification of estimator $\hat{\rho}$

$$\hat{\rho}_\delta := \mathrm{argmin}_{S\in\mathcal{S}_{m,\delta}} \left[n^{-1} \sum_{j=1}^{n} (Y_j - \langle S, X_j\rangle)^2 \right], \tag{2.4.23}$$

then the statements of theorems 15, 16 and 17 hold for the estimator $\hat{\rho}_\delta$ (in the case of Theorem 15, under the additional assumption that $\sigma_\xi \geq m^{-1/2}$).

2.4.2 The minimal distance in spectral norm of the projection estimator

Recall that

$$\pi_{\mathcal{S}_m}(Z) := \mathrm{argmin}_{S\in\mathcal{S}_m} \|Z - S\|_2^2, Z \in \mathbb{H}_m$$

defines the projection of Z onto $\mathcal{S}_m$. The mapping $\mathbb{H}_m \ni Z \mapsto \pi_{\mathcal{S}_m}(Z) \in \mathcal{S}_m$ possesses a couple of simple properties stated in the next proposition. Denote by $\mathcal{S}_m^d$ the set of all diagonal density matrices.

Proposition 2. *1. For all $m \times m$ unitary matrices U,*

$$\pi_{\mathcal{S}_m}(U^{-1}ZU) = U^{-1}\pi_{\mathcal{S}_m}(Z)U, Z \in \mathbb{H}_m.$$

2. If $D \in \mathbb{H}_m$ is a diagonal matrix, then $\pi_{\mathcal{S}_m}(D) \in \mathcal{S}_m^d$.

Proof. To prove the first claim, note that, by the unitary invariance of the Hilbert–Schmidt norm,

$$\|U^{-1}ZU - S\|_2^2 = \|U^{-1}(Z - USU^{-1})U\|_2^2 = \|Z - USU^{-1}\|_2^2.$$

In addition, the mapping $S \mapsto USU^{-1}$ is a bijection from the set $\mathcal{S}_m$ onto itself. This immediately implies that

$$\pi_{\mathcal{S}_m}(U^{-1}ZU) = \mathrm{argmin}_{S\in\mathcal{S}_m} \|Z - USU^{-1}\|_2^2 = U^{-1}\pi_{\mathcal{S}_m}(Z)U.$$

For an $m \times m$ matrix $A = (a_{ij})_{i,j=1}^m \in \mathbb{H}_m$, let A^d be the diagonal matrix with diagonal entries $a_{ii}, i = 1, \ldots, m$. It is easy to see that if A is a density matrix, then A^d is also a density matrix. Moreover, it is also obvious that, for a diagonal matrix D,

$$\|D - A^d\|_2^2 \leq \|D - A\|_2^2, A \in \mathcal{S}_m,$$

with a strict inequality if A is not diagonal. These observations immediately imply the second claim.

□

We will now state and prove a vector version of Theorem 15 in which the role of the set of density matrices $\mathcal{S}_m$ is played by the simplex

$$\Delta_m := \Big\{u = (u_1, \ldots, u_m) \in \mathbb{R}^m : u_j \geq 0, \sum_{j=1}^m u_j = 1\Big\}$$

in $\mathbb{R}^m$ (this is equivalent to considering the set of diagonal density matrices). We will then show that the matrix version of the problem reduces to the vector case.

Define

$$\pi_{\Delta_m}(z) := \operatorname{argmin}_{u \in \Delta_m} \|z - u\|_{\ell_2^m}^2, z \in \mathbb{R}^m.$$

Since the function $u \mapsto \|z - u\|_{\ell_2^m}^2$ is strictly convex and Δ_m is a compact convex set, such a minimizer exists and is unique. In other words, $\pi_{\Delta_m}(z)$ is the projection of the point $z \in \mathbb{R}^m$ onto simplex Δ_m (the closest point to z in the set Δ_m with respect to the Euclidean ℓ_2^m-distance). The next lemma shows that the same point also minimizes the ℓ_∞^m-distance from z to the simplex Δ_m.

Lemma 20. *For all $z \in \mathbb{R}^m$,*

$$\|z - \pi_{\Delta_m}(z)\|_{\ell_\infty^m} = \min_{v \in \Delta_m} \|z - v\|_{\ell_\infty^m}.$$

Proof. Without loss of generality, assume that $z = (z_1, \ldots, z_m) \in \mathbb{R}^m$ is a point with $z_1 \geq \cdots \geq z_m$. Denote

$$\bar{z}_j := \frac{z_1 + \cdots + z_j}{j}, j = 1, \ldots, m.$$

Clearly, $\bar{z}_1 = z_1$ and $\bar{z}_j \geq z_j, j = 1, \ldots, m$. Let

$$k := \max\left\{ j \leq m : \bar{z}_j \leq z_j + \frac{1}{j} \right\}.$$

Note that if $k > 1$, then, for all $j < k$, $\bar{z}_j \leq z_j + \frac{1}{j}$. Indeed,

$$\bar{z}_j = \frac{k\bar{z}_k - \sum_{i=j+1}^{k} z_i}{j} \leq \frac{kz_k + 1 - (k-j)z_k}{j} = \frac{jz_k + 1}{j} = z_k + \frac{1}{j} \leq z_j + \frac{1}{j}.$$

On the other hand, if $k < m$, then $\bar{z}_k > z_{k+1} + \frac{1}{k}$. Indeed, if $\bar{z}_k \leq z_{k+1} + \frac{1}{k}$, then

$$\bar{z}_{k+1} = \frac{k\bar{z}_k + z_{k+1}}{k+1} \leq \frac{kz_{k+1} + 1 + z_{k+1}}{k+1} = z_{k+1} + \frac{1}{k+1},$$

which would contradict the definition of k.

Let $\lambda = (\lambda_1, \ldots, \lambda_m)$, where $\lambda_j = z_j - \bar{z}_k + \frac{1}{k}$ for $j = 1, \ldots, k$ and $\lambda_j = 0$ for $j = k+1, \ldots, m$. Since $\bar{z}_k \leq z_k + \frac{1}{k} \leq z_j + \frac{1}{k}$ for all $j \leq k$, we have $\lambda_j \geq 0, j = 1, \ldots, m$ and

$$\sum_{j=1}^{m} \lambda_j = \sum_{j=1}^{k} \left(z_j - \bar{z}_k + \frac{1}{k} \right) = \sum_{j=1}^{k} z_j - k\bar{z}_k + 1 = 1.$$

Thus, $\lambda \in \Delta_m$. It turns out that $\pi_{\Delta_m}(z) = \lambda$. [3] To prove this it is enough to show that $z - \lambda \in N_{\Delta_m}(\lambda)$, where

$$N_{\Delta_m}(\lambda) := \{u \in \mathbb{R}^m : \langle u, v - \lambda \rangle \leq 0, v \in \Delta_m\}$$

is the normal cone of the convex set Δ_m at point λ (see, e.g., [3], Proposition 5, Chapter 4, Section 1). Let $t := \bar{z}_k - \frac{1}{k}$. Clearly, we have $z_{k+1} < t \leq z_k$ if $k < m$ and $t \leq z_m$ if $k = m$. For $k = m$, $z - \lambda = (t, \ldots, t)$ and

$$\langle z - \lambda, v - \lambda \rangle = \sum_{i=1}^{m} t(v_i - \lambda_i) = t\left(\sum_{i=1}^{m} v_i - \sum_{i=1}^{m} \lambda_i \right) = 0$$

since $v, \lambda \in \Delta_m$. For $k < m$, note that

$$z - \lambda = (t, \ldots t, z_{k+1}, \ldots, z_m)$$

[3] The computation of the projection onto a simplex occurs in many applications and has been studied before: see, e.g. [83] and [70]. See also [25], where an explicit expression for the projection was derived. For completeness, we provide our version of the proof below.

and, for $v \in \Delta_m$,

$$\langle z - \lambda, v - \lambda \rangle = \sum_{i=1}^{k} t(v_i - \lambda_i) + \sum_{i=k+1}^{m} z_i v_i.$$

Using the facts that $\sum_{i=1}^{m} v_i = 1$ and $\sum_{i=1}^{k} \lambda_i = 1$, we get

$$\begin{aligned}\langle z - \lambda, v - \lambda \rangle &= t\left(\sum_{i=1}^{k} v_i - \sum_{i=1}^{k} \lambda_i\right) + \sum_{i=k+1}^{m} z_i v_i \\ &= -t \sum_{i=k+1}^{m} v_i + \sum_{i=k+1}^{m} z_i v_i = \sum_{i=k+1}^{m} (z_i - t) v_i \leq 0,\end{aligned}$$

where we also used that, for all $i = k+1, \ldots, m$, $z_i - t \leq z_{k+1} - t \leq 0$ and $v_i \geq 0$. Thus, $z - \lambda \in N_{\Delta_m}(\lambda)$ and, by the uniqueness of the minimum, $\lambda = \pi_{\Delta_m}(z)$.

Note that

$$\|z - \lambda\|_{\ell_\infty^m} = \max(|t|, |z_{k+1}|, \ldots, |z_m|).$$

For any $v \in \Delta_m$,

$$t = \bar{z}_k - \frac{1}{k} = \frac{1}{k}\sum_{i=1}^{k} z_i - \frac{1}{k}\sum_{i=1}^{m} v_i \leq \frac{1}{k}\sum_{i=1}^{k} z_i - \frac{1}{k}\sum_{i=1}^{k} v_i = \frac{1}{k}\sum_{i=1}^{k}(z_i - v_i) \leq \|z - v\|_{\ell_\infty^m}.$$

On the other hand,

$$z_m \geq z_m - v_m \geq -\|z - v\|_{\ell_\infty^m}.$$

Since

$$t = \bar{z}_k - \frac{1}{k} \geq z_{k+1} \geq \cdots \geq z_m,$$

we conclude that, for all $v \in \Delta_m$,

$$\|z - \lambda\|_{\ell_\infty^m} \leq \|z - v\|_{\ell_\infty^m}.$$

□

We now turn to the proof of Lemma 15.

Proof. Any matrix $Z \in \mathbb{H}_m$ admits spectral representation $Z = U^{-1} D U$, where D is the diagonal matrix with real entries $d_1, \ldots, d_m$ on the diagonal and U is a unitary

$m \times m$ matrix. Let $d = (d_1, \dots, d_m) \in \mathbb{R}^m$. Given $v = (v_1, \dots, v_m) \in \Delta_m$, the diagonal matrix V with entries $v_1, \dots, v_m$ is a density matrix. This defines a bijection $\Delta_m \ni v \mapsto V = J(v)$ between the simplex Δ_m and the set $\mathcal{S}_m^d$ of all diagonal $m \times m$ density matrices. Moreover, J is an isometry of Δ_m and $\mathcal{S}_m^d$: $\|J(v) - J(u)\|_2^2 = \|u - v\|_{\ell_2^m}^2, u, v \in \Delta_m$.

We will now prove the following lemma.

Lemma 21. *Let $Z = U^{-1}DU$ with a unitary $m \times m$ matrix U and diagonal matrix D with $d = (d_1, \dots, d_m) \in \mathbb{R}^m$ being the vector of its diagonal entries. Then*

$$\pi_{\mathcal{S}_m}(Z) = U^{-1}J(\pi_{\Delta_m}(d))U.$$

Proof. This is an immediate consequence of Proposition 2 and the following simple fact:

$$\text{argmin}_{A \in \mathcal{S}_m^d} \|D - A\|_2^2 = J\Bigg(\text{argmin}_{v \in \Delta_m} \|J(d) - J(v)\|_2^2\Bigg)$$

$$J\Bigg(\text{argmin}_{v \in \Delta_m} \|d - v\|_{\ell_2^m}^2\Bigg) = J(\pi_{\Delta_m}(d)).$$

□

To complete the proof of Lemma 15, observe that, In view of lemmas 20, 21,

$$\|Z - \pi_{\mathcal{S}_m}(Z)\|_\infty = \|U^{-1}(J(d) - J(\pi_{\Delta_m}(d)))U\|_\infty$$

$$= \|J(d) - J(\pi_{\Delta_m}(d))\|_\infty = \|d - \pi_{\Delta_m}(d)\|_{\ell_\infty^m} = \inf_{v \in \Delta_m} \|d - v\|_{\ell_\infty^m}.$$

Without loss of generality, assume that $d_1 \geq \cdots \geq d_m$. Let $S \in \mathcal{S}_m$ be a density matrix with eigenvalues $v_1 \geq \cdots \geq v_m$. Clearly, $v = (v_1, \dots, v_m) \in \Delta_m$. Therefore,

$$\|Z - \pi_{\mathcal{S}_m}(Z)\|_\infty \leq \|d - v\|_\infty \leq \|Z - S\|_\infty,$$

where to get the last bound we used Weyl's perturbation inequality (see [9], Corollary III.2.6).

□

2.5 The optimality of Dantzig-type estimator

We define the matrix Dantzig estimator (or selector) $\acute{\rho}^D$ as the solution of the following convex optimization problem:

$$\min \|S\|_1 \quad \text{subject to } S \in \Lambda(\varepsilon), \tag{2.5.1}$$

where

$$\Lambda(\varepsilon) := \Big\{ S \in \mathcal{S}_m, \Big\| \frac{1}{n} \sum_{i=1}^{n} \big(Y_j - \langle S, X_j \rangle\big) X_j \Big\|_\infty \leq \varepsilon \Big\}$$

for some constant $\varepsilon \geq 0$. When $\varepsilon = 0$, it corresponds to the noiseless case(i.e., $\sigma_\xi = 0$) where the exact recovery of ρ is the main interest, see also Section 1.3.2. The original Dantzig estimator was introduced in [20] for low rank matrix estimation and was applied in quantum state tomography for estimating low rank density matrices, see [64], [36] and [30]. Our definition adds another constraint that the solution should be a valid density matrix. They also proved sharp convergence rates in Schatten 1-norm and Schatten 2-norm distances by applying some techniques based on the *restricted isometry property*(RIP) which requires $n \gtrsim mr \log^6 m$ Pauli measurements. Note that RIP is a strong assumption, but there is yet no results related to the convergence rates of $\acute{\rho}^D$ in other Schatten p-norms. It is commonly known that proving the convergence rate in spectral norm (i.e. $p = \infty$) is difficult.

When $S \in \mathcal{S}_m$, the objective function in (2.5.1) is always 1 and has no effect on the optimization problem. Instead, we will study the following estimator:

$$\acute{\rho}_\varepsilon := \arg\min \Big\{ \text{tr}(S \log S) : S \in \Lambda(\varepsilon) \Big\}, \tag{2.5.2}$$

where we replaced the nuclear norm in (2.5.1) with the negative von Neumann entropy. Remember that the von Neumann entropy of a density matrix ρ is defined as

$$V(\rho) := -\text{tr}(\rho \log \rho), \ \forall \rho \in \mathcal{S}_m,$$

which is a concave function on $\mathcal{S}_m$ and then (2.5.2) is actually a convex optimization problem. In this section, we prove the sharp convergence rates of $\acute{\rho}_\varepsilon$ in all the Schatten

p-norms with $p \in [1, +\infty]$. These rates also hold for the standard matrix Dantzig estimator $\acute{\rho}^D$ as the solution in (2.5.1). Moreover, we obtain sharp convergence rate of $\acute{\rho}_\varepsilon$ in Kullback-Leibler divergence. In Section 2.4.1, we proved similar upper bounds of Schatten p-norms for the (modified) least squares estimator, based on a minimal distance projection onto the simplex. It will be shown in Section 2.5.1 that the condition needed for $\acute{\rho}^\varepsilon$ is improved than the projection estimator $\hat{\rho}$ in (2.4.2).

2.5.1 Oracle inequality and the Schatten p-norm convergence rates

Theorem 19 displays the performance of $\acute{\rho}_\varepsilon$ by a *low rank oracle inequality.* The *low rank oracle inequality* has been well studied for (matrix) LASSO estimator, see Section 2.3, also [52] and [55]. When studying Dantzig estimator in compressed sensing, the *sparsity oracle inequality* is considered over all oracles in the feasible set, namely $\Lambda(\varepsilon)$ in (2.5.1), see for example [51]. It is generally impossible to compare the performance of the estimator with sparse oracles (or low rank oracles in the matrix case) when they are not in the feasible set. Surprisingly, we can obtain the following *low rank oracle inequality* for $\acute{\rho}_\varepsilon$ which actually hold for all the oracles in $\mathcal{S}_m$, even when the oracle is infeasible for the optimization problem (2.5.1) and (2.5.2).

Define for any $S \in \mathcal{S}_m$ and $t > 0$,

$$\varphi_1(n, S, t) := \varphi_1(n, S, U, \sigma_\xi, t) := \frac{m\sigma_\xi^2 \mathrm{rank}(S)\big(t + \log(2m)\big)}{n} + \frac{\sigma_\xi^2 m^2 U^2 \mathrm{rank}(S)\big(t + \log(2m)\big)^2}{n^2}$$

and

$$\varphi_2(n, S, t) := \varphi_2(n, S, U, t) := \frac{m^2 U^4 \mathrm{rank}(S)\big(\log^3 m \log^3 n + t\big)^2}{n^2}.$$

These are the main terms in our *low rank oracle inequality* characterizing the convergence rates of $\acute{\rho}_\varepsilon$(and $\acute{\rho}^D$). The first term is directly connected to the noise level σ_ξ. The second term is related with the randomization error and involves the constant U in the higher order term $O(n^{-2})$. These are also the main terms in the oracle inequality for the least squares estimator proved in Section 2.3.3. For the sake of brevity, let denote $\varphi_1(n) := \varphi_1(n, \rho, \log(2m))$ and $\varphi_2(n) := \varphi_2(n, \rho, \log(2m))$. It is possible to improve the exponents of the logarithmic terms which is beyond our main interest and will not be pursued here.

Theorem 19. *Suppose Assumptions 1 and 4 hold and $\rho \in \mathcal{S}_{r,m}$. Let $\acute{\rho}_\varepsilon$ be as defined in (2.5.2) and any $\varepsilon \geq C_1\Big(\sigma_\xi\sqrt{\frac{t+\log(2m)}{nm}} + \sigma_\xi U\frac{t+\log(2m)}{n}\Big)$ for any $t \geq 1$ and some large enough constant $C_1 > 0$. There exists a constant $C > 0$ such that with probability at least $1 - e^{-t}$,*

$$\|\acute{\rho}_\varepsilon - \rho\|_{L_2(\Pi)}^2 \leq \inf_{S \in \mathcal{S}_m}\Big\{2\|S-\rho\|_{L_2(\Pi)}^2 + C\Big(m^2\varepsilon^2 rank(S) + \varphi_1(n, S, t) + \varphi_2(n, S, t)\Big)\Big\}. \tag{2.5.3}$$

Moreover, if $\varepsilon = C_1\Big(\sigma_\xi\sqrt{\frac{\log(2m)}{nm}} + \frac{\sigma_\xi U\log(2m)}{n}\Big)$, then with probability at least $1 - \frac{1}{m}$,

$$\|\acute{\rho}_\varepsilon - \rho\|_{L_2(\Pi)}^2 \leq C\Big(\varphi_1(n) + \varphi_2(n)\Big) \tag{2.5.4}$$

and

$$K(\rho\|\acute{\rho}_\varepsilon) \leq Cm\sqrt{\big(\varphi_1(n) + \varphi_2(n)\big)rank(\rho)}\log\frac{mn}{\sigma_\xi} + \frac{\sigma_\xi}{n}\sqrt{rank(\rho)}\log\frac{mn}{\sigma_\xi}. \tag{2.5.5}$$

Remark 7. *The objective function in optimization problem (2.5.2) is not involved in the proof of (2.5.3). Therefore, the bound (2.5.3) also hold for the standard Dantzig estimator $\acute{\rho}^D$. Moreover, instead of (2.5.3), we actually prove a sharper bound:*

$$\|\acute{\rho}_\varepsilon - \rho\|_{L_2(\Pi)}^2 \leq 2\|S-\rho\|_{L_2(\Pi)}^2 + C\Big(m^2\varepsilon^2 rank(S) + \varphi_1(n, S, t) + \varphi_2(n, S, t)\big(\|\acute{\rho}_\varepsilon - S\|_1^2 + \|S-\rho\|_1^2\big)\Big),$$

for any $S \in \mathcal{S}_m$. It indicates that if Pauli measurements are used($U = \frac{1}{\sqrt{m}}$) and $n \geq C'mr\log^6 m\log^6 n$ for large enough constant $C' > 0$ such that (due to Lemma 17 in Section 2.4.1 with $p = 1$) $\varphi_2(n)\|\acute{\rho}_\varepsilon - \rho\|_1^2 \leq 8\varphi_2(n)r\|\acute{\rho}_\varepsilon - \rho\|_2^2 \leq \frac{1}{2}\|\acute{\rho}_\varepsilon - \rho\|_{L_2(\Pi)}^2$, we *get* $\|\acute{\rho}_\varepsilon - \rho\|^2_{L_2(\Pi)} \leq 2C\big(m^2\varepsilon^2 rank(\rho) + \varphi_1(n)\big)$*, which reduces to the canonical result by applying the restricted isometry property(see [64],[20]). This bound depends linearly on* σ_ξ *(which can be arbitrarily small, even* 0*), see also Remark 8 after Theorem 20.*

Generally, if we assume that (remember that $\sigma_\xi \leq U$)

$$U^2\sqrt{\frac{m}{n}}\log^{5/2} m \log^3 n \leq \sigma_\xi, \tag{2.5.6}$$

the choice of ε can be simplified to

$$\varepsilon = C_1\sigma_\xi\sqrt{\frac{\log(2m)}{mn}}$$

and (2.5.3) and (2.5.4) in Theorem 19 can be simplified into

$$\|\acute{\rho}_\varepsilon - \rho\|_2 \lesssim \sigma_\xi \frac{\sqrt{\text{rank}(\rho)}m^{3/2}\log^{1/2}(2m)}{\sqrt{n}} \tag{2.5.7}$$

and (due to Lemma 17 with $p = 1$)

$$\|\acute{\rho}_\varepsilon - \rho\|_1 \lesssim \sigma_\xi \frac{\text{rank}(\rho)m^{3/2}\log^{1/2}(2m)}{\sqrt{n}} \tag{2.5.8}$$

and

$$K(\rho\|\acute{\rho}_\varepsilon) \lesssim \sigma_\xi \frac{\text{rank}(\rho)m^{3/2}\log^{1/2}(2m)\log(mn/\sigma_\xi)}{\sqrt{n}}.$$

According to the minimax lower bounds established in Section 2.2, these bounds are optimal except the logarithmic terms. Note that by applying the interpolation inequality in Lemma 6 with (2.5.7) and (2.5.8), we can also get the upper bound of $\|\acute{\rho}_\varepsilon - \rho\|_p$ for all $1 \leq p \leq 2$. In the case of Pauli basis where $U = \frac{1}{\sqrt{m}}$, the assumptions (2.5.6) hold if σ_ξ is larger than $\frac{1}{\sqrt{mn}}$(times an additional logarithmic factor), which is also the condition needed for the optimality of the least squares estimator (see Section 2.3.4).

The main technical tool for our proof is the following lemma which gives a probabilistic upper bound of the product empirical processes. For any $\Delta \in [0, 1]$, define the set and quantity

$$\mathcal{A}(\Delta) := \big\{A \in \mathbb{H}_m, \|A\|_1 \leq 1, \|A\|_{L_2(\Pi)} \leq \Delta\big\}$$

and

$$\alpha_n(\Delta_1,\Delta_2) := \sup_{A_1\in\mathcal{A}(\Delta_1)} \sup_{A_2\in\mathcal{A}(\Delta_2)} \Big|\frac{1}{n}\sum_{i=1}^{n}\langle A_1, X_i\rangle\langle A_2, X_i\rangle - \mathbb{E}\langle A_1, X\rangle\langle A_2, X\rangle\Big|.$$

Lemma 22. *Given $0<\delta^-<\delta^+$ and $t\geq 1$, let*

$$\bar{t} := t + \log(\log_2(\delta^+/\delta^-)+3).$$

Then, with some constant C and probability at least $1-e^{-t}$, the following bound holds for all $\frac{\Delta_1+\Delta_2}{2}\in[\delta^-,\delta^+]$:

$$\alpha_n(\Delta_1,\Delta_2) \leq C\Big[(\Delta_1+\Delta_2)U\frac{\log^{3/2} m\log^{3/2} n+\sqrt{\bar{t}}}{\sqrt{n}} + U^2\frac{\log^3 m\log^3 n+\bar{t}}{n}\Big].$$

Generally, tight upper bounds (generic chaining bounds) of product empirical processes are not easy to derive due to the nontrivial geometric structure of the indexing classes of the product empirical processes, see [69] and references therein. Even though we suspect that the bound in Lemma 22 might not be sharp, it is sufficient for us to prove the results we need in this section. Lemma 22 will be used to prove the oracle inequality (2.5.3) and the spectral norm (i.e., $p=+\infty$) convergence rate of $\acute{\rho}_\varepsilon$ in (2.5.15). The proof of Lemma 22 is given in Section 2.5.2.

Proof of Theorem 19. Denote $\Xi_1 = \frac{1}{n}\sum_{i=1}^{n}\xi_i X_i$. By Lemma 8 in Section 2.3.1 with $\alpha=2$, we know that with probability at least $1-e^{-t}$,

$$\big\|\Xi_1\big\|_\infty \leq C\Big(\sigma_\xi\sqrt{\frac{t+\log(2m)}{nm}} + \sigma_\xi U\frac{t+\log(2m)}{n}\Big) \tag{2.5.9}$$

for some constant $C>0$. Note that we used the facts $\|\mathbb{E}\xi^2X^2\|_\infty^{1/2} \leq \sigma_\xi\frac{1}{\sqrt{nm}}$ and $\big\|\|\xi X\|_\infty\big\|_{\psi_2} \leq \|\xi\|_{\psi_2}U \leq 2\sigma_\xi U$ a.s.. The choice of ε in Theorem 19 guarantees the existence of the solution $\acute{\rho}_\varepsilon$ since $\Lambda(\varepsilon)$ is nonempty and $\rho\in\Lambda(\varepsilon)$.

The fact $\acute{\rho}_\varepsilon\in\Lambda(\varepsilon)$ indicates that, for any $S\in\mathcal{S}_m$,

$$\frac{1}{n}\sum_{j=1}^{n}\big(\langle\acute{\rho}_\varepsilon, X_j\rangle - Y_j\big)\langle\acute{\rho}_\varepsilon - S, X_j\rangle \leq \varepsilon\|\acute{\rho}_\varepsilon - S\|_1.$$

Then, by arranging the terms accordingly,

$$\begin{aligned}\big\langle \acute{\rho}_\varepsilon - \rho, \acute{\rho}_\varepsilon - S\big\rangle_{L_2(\Pi)} \leq \varepsilon \|\acute{\rho}_\varepsilon - S\|_1 + \big\langle \Xi_1, \acute{\rho}_\varepsilon - S\big\rangle \\ + \Big| \frac{1}{n}\sum_{i=1}^n \big\langle \acute{\rho}_\varepsilon - \rho, X_i\big\rangle \big\langle \acute{\rho}_\varepsilon - S, X_i\big\rangle - \mathbb{E}\big\langle \acute{\rho}_\varepsilon - \rho, X\big\rangle\big\langle \acute{\rho}_\varepsilon - S, X\big\rangle\Big|.\end{aligned}$$

Observe that

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

Therefore, we get

$$\begin{aligned}\|\acute{\rho}_\varepsilon - \rho\|^2_{L_2(\Pi)} + \|\acute{\rho}_\varepsilon - S\|^2_{L_2(\Pi)} \leq \|S - \rho\|^2_{L_2(\Pi)} + 2(\varepsilon + \|\Xi_1\|_\infty)\|\acute{\rho}_\varepsilon - S\|_1 \\ + 2\Big| \frac{1}{n}\sum_{i=1}^n \big\langle \acute{\rho}_\varepsilon - \rho, X_i\big\rangle \big\langle \acute{\rho}_\varepsilon - S, X_i\big\rangle - \mathbb{E}\big\langle \acute{\rho}_\varepsilon - \rho, X\big\rangle\big\langle \acute{\rho}_\varepsilon - S, X\big\rangle\Big|.\end{aligned} \tag{2.5.10}$$

By definition of $\alpha_n(\Delta_1, \Delta_2)$, we can control the last term in above inequality as follows:

$$\begin{aligned}&\Big| \frac{1}{n}\sum_{i=1}^n \big\langle \acute{\rho}_\varepsilon - \rho, X_i\big\rangle \big\langle \acute{\rho}_\varepsilon - S, X_i\big\rangle - \mathbb{E}\big\langle \acute{\rho}_\varepsilon - \rho, X\big\rangle\big\langle \acute{\rho}_\varepsilon - S, X\big\rangle\Big| \\ &\qquad \leq \|\acute{\rho}_\varepsilon - \rho\|_1 \|\acute{\rho}_\varepsilon - S\|_1 \alpha_n\Big(\frac{\|\acute{\rho}_\varepsilon - \rho\|_{L_2(\Pi)}}{\|\acute{\rho}_\varepsilon - \rho\|_1}, \frac{\|\acute{\rho}_\varepsilon - S\|_{L_2(\Pi)}}{\|\acute{\rho}_\varepsilon - S\|_1}\Big).\end{aligned}$$

We apply Lemma 22 with $\delta^- = \frac{1}{mn}$ and $\delta^+ = \frac{1}{m}$. Clearly, if $\frac{\|\acute{\rho}_\varepsilon - \rho\|_{L_2(\Pi)}}{\|\acute{\rho}_\varepsilon - \rho\|_1} + \frac{\|\acute{\rho}_\varepsilon - S\|_{L_2(\Pi)}}{\|\acute{\rho}_\varepsilon - S\|_1} \geq \delta^-$, Lemma 22 yields that, with probability at least $1 - e^{-t}$,

$$\begin{aligned}&\Big| \tfrac{1}{n}\textstyle\sum_{i=1}^n \big\langle \acute{\rho}_\varepsilon - \rho, X_i\big\rangle \big\langle \acute{\rho}_\varepsilon - S, X_i\big\rangle - \mathbb{E}\big\langle \acute{\rho}_\varepsilon - \rho, X\big\rangle\big\langle \acute{\rho}_\varepsilon - S, X\big\rangle\Big| \\ &\leq \|\acute{\rho}_\varepsilon - \rho\|_1 \|\acute{\rho}_\varepsilon - S\|_1 \Big(\tfrac{\|\acute{\rho}_\varepsilon - \rho\|_{L_2(\Pi)}}{\|\acute{\rho}_\varepsilon - \rho\|_1} + \tfrac{\|\acute{\rho}_\varepsilon - S\|_{L_2(\Pi)}}{\|\acute{\rho}_\varepsilon - S\|_1}\Big) CU\tfrac{\log^{3/2} m \log^{3/2} n + \sqrt{\bar{t}}}{\sqrt{n}} \\ &\quad + \|\acute{\rho}_\varepsilon - \rho\|_1 \|\acute{\rho}_\varepsilon - S\|_1 CU^2 \tfrac{\log^3 m \log^3 n + \bar{t}}{n} \\ &= \|\acute{\rho}_\varepsilon - S\|_1 \|\acute{\rho}_\varepsilon - \rho\|_{L_2(\Pi)} CU \tfrac{\log^{3/2} m \log^{3/2} n + \sqrt{\bar{t}}}{\sqrt{n}} \\ &\quad + \|\acute{\rho}_\varepsilon - \rho\|_1 \|\acute{\rho}_\varepsilon - S\|_{L_2(\Pi)} CU \tfrac{\log^{3/2} m \log^{3/2} n + \sqrt{\bar{t}}}{\sqrt{n}} \\ &\quad + \|\acute{\rho}_\varepsilon - \rho\|_1 \|\acute{\rho}_\varepsilon - S\|_1 CU^2 \tfrac{\log^3 m \log^3 n + \bar{t}}{n},\end{aligned}$$

where $\bar{t} = t + \log(\log_2 n + 3)$. Recall from the proof of Lemma 17 in Section 2.4.1

that $\|\acute{\rho}_\varepsilon - S\|_1 \leq 2\sqrt{2\text{rank}(S)}\|\acute{\rho}_\varepsilon - S\|_2$,

$$\begin{aligned}
\|\acute{\rho}_\varepsilon - S\|_1\|\acute{\rho}_\varepsilon - \rho\|_{L_2(\Pi)}&CU\frac{\log^{3/2} m \log^{3/2} n + \sqrt{t}}{\sqrt{n}} \\
&\leq \frac{1}{4}\|\acute{\rho}_\varepsilon - \rho\|_{L_2(\Pi)}^2 + 2C^2\|\acute{\rho}_\varepsilon - S\|_1^2 U^2\frac{\log^3 m \log^3 n + t}{n} \\
&\leq \frac{1}{4}\|\acute{\rho}_\varepsilon - \rho\|_{L_2(\Pi)}^2 + \frac{1}{4}\|\acute{\rho}_\varepsilon - S\|_{L_2(\Pi)}^2 + c_1\varphi_2(n,S,t)\|\acute{\rho}_\varepsilon - S\|_1^2,
\end{aligned}$$

for some constant $c_1 > 0$, where we applied the inequality $ab \leq \frac{a^2}{4} + b^2$ multiple times. Moreover, since $\|\acute{\rho}_\varepsilon - \rho\|_1 \leq \|\acute{\rho}_\varepsilon - S\|_1 + \|S - \rho\|_1$,

$$\begin{aligned}
\|\acute{\rho}_\varepsilon - \rho\|_1\|\acute{\rho}_\varepsilon - S\|_{L_2(\Pi)}&CU\frac{\log^{3/2} m \log^{3/2} n + \sqrt{t}}{\sqrt{n}} \\
&\leq \frac{1}{8}\|\acute{\rho}_\varepsilon - S\|_{L_2(\Pi)}^2 + 4C^2\|\acute{\rho}_\varepsilon - \rho\|_1^2 U^2\frac{\log^3 m \log^3 n + t}{n} \\
&\leq \frac{1}{8}\|\acute{\rho}_\varepsilon - S\|_{L_2(\Pi)}^2 + 8C^2\|\acute{\rho}_\varepsilon - S\|_1^2 U^2\frac{\log^3 m \log^3 n + t}{n} \\
&\quad + 8C^2\|S - \rho\|_1^2 U^2\frac{\log^3 m \log^3 n + t}{n} \\
&\leq \frac{1}{4}\|\acute{\rho}_\varepsilon - S\|_{L_2(\Pi)}^2 + \frac{1}{4}\|S - \rho\|_{L_2(\Pi)}^2 + c_1\varphi_2(n,S,t)\big(\|\acute{\rho}_\varepsilon - S\|_1^2 + \|S - \rho\|_1^2\big).
\end{aligned}$$

Similarly, we can get

$$\begin{aligned}
\|\acute{\rho}_\varepsilon - \rho\|_1\|\acute{\rho}_\varepsilon - S\|_1 CU^2\frac{\log^3 m \log^3 n + \bar{t}}{n} \leq& \frac{1}{4}\|\acute{\rho}_\varepsilon - S\|_{L_2(\Pi)}^2 \\
&+ c_1\varphi_2(n,S,t)\big(\|\acute{\rho}_\varepsilon - S\|_1^2 + \|S - \rho\|_1^2\big).
\end{aligned}$$

Therefore, we conclude that if $\frac{\|\acute{\rho}_\varepsilon - \rho\|_{L_2(\Pi)}}{\|\acute{\rho}_\varepsilon - \rho\|_1} + \frac{\|\acute{\rho}_\varepsilon - S\|_{L_2(\Pi)}}{\|\acute{\rho}_\varepsilon - S\|_1} \geq \delta^-$, with probability at least $1 - e^{-t}$,

$$\begin{aligned}
\Big|\frac{1}{n}\sum_{i=1}^n \big\langle \acute{\rho}_\varepsilon - \rho, X_i\big\rangle\big\langle \acute{\rho}_\varepsilon - S, X_i\big\rangle - \mathbb{E}\big\langle \acute{\rho}_\varepsilon - \rho, X\big\rangle\big\langle \acute{\rho}_\varepsilon - S, X\big\rangle\Big| \\
\leq \frac{3}{4}\|\acute{\rho}_\varepsilon - S\|_{L_2(\Pi)}^2 + \frac{1}{4}\|\acute{\rho}_\varepsilon - \rho\|_{L_2(\Pi)}^2 + \frac{1}{4}\|S - \rho\|_{L_2(\Pi)}^2 \\
+ c_1\varphi_2(n,S,t)\big(\|\acute{\rho}_\varepsilon - S\|_1^2 + \|S - \rho\|_1^2\big).
\end{aligned}\tag{2.5.11}$$

If, on the other hand, $\frac{\|\acute{\rho}_\varepsilon - \rho\|_{L_2(\Pi)}}{\|\acute{\rho}_\varepsilon - \rho\|_1} + \frac{\|\acute{\rho}_\varepsilon - S\|_{L_2(\Pi)}}{\|\acute{\rho}_\varepsilon - S\|_1} \leq \delta^- = \frac{1}{mn}$, then the proof of (2.5.3) only simplifies since

$$\begin{aligned}
\|\acute{\rho}_\varepsilon - \rho\|_{L_2(\Pi)}^2 \leq \frac{1}{n^2m^2}\|\acute{\rho}_\varepsilon - \rho\|_1^2 &\leq U^4m^2\frac{\log^3 m \log^3 n + t}{n^2}\big(\|\acute{\rho}_\varepsilon - S\|_1^2 + \|S - \rho\|_1^2\big) \\
&\leq \varphi_2(n,S,t)\big(\|\acute{\rho}_\varepsilon - S\|_1^2 + \|S - \rho\|_1^2\big).
\end{aligned}$$

Plugging (2.5.11) into (2.5.10), we get that with probability at least $1-e^{-t}$,

$$\begin{aligned}\frac{3}{4}\|\acute{\rho}_\varepsilon-\rho\|_{L_2(\Pi)}^2+\frac{1}{4}\|\acute{\rho}_\varepsilon-S\|_{L_2(\Pi)}^2\leq\frac{5}{4}\|S-\rho\|_{L_2(\Pi)}^2+2(\varepsilon+\|\Xi_1\|_\infty)\|\acute{\rho}_\varepsilon-S\|_1\\ +c_1\varphi_2(n,S,t)\big(\|\acute{\rho}_\varepsilon-S\|_1^2+\|S-\rho\|_1^2\big).\end{aligned}\tag{2.5.12}$$

By the bound (2.5.9) and the choice of ε, we have

$$\begin{aligned}2(\varepsilon+\|\Xi_1\|_\infty)\|\acute{\rho}_\varepsilon-S\|_1\leq&\frac{1}{4}\|\acute{\rho}_\varepsilon-S\|_{L_2(\Pi)}^2+4m^2\mathrm{rank}(S)(\varepsilon+\|\Xi_1\|_\infty)^2\\ \leq&\frac{1}{4}\|\acute{\rho}_\varepsilon-S\|_{L_2(\Pi)}^2+C\big(m^2\varepsilon^2\mathrm{rank}(S)+\varphi_1(n,S,t)\big).\end{aligned}\tag{2.5.13}$$

By putting the bound (2.5.13) into (2.5.12) and adjusting some constants, we get the bound (2.5.3). Then (2.5.4) is an immediate result from (2.5.3) by setting $S=\rho$.

We are ready to prove (2.5.5). Consider $\rho'=(1-\delta)\rho+\delta\frac{I_m}{m}$ with $\delta=\frac{\sigma_\xi}{n}\leq\frac{U}{n}$, a technique already used in the proof of (2.3.29) in Section 2.3.3. It is easy to check that $\rho'\in\Lambda(\varepsilon)$ as long as the constant C_1 in ε is large enough. By definition of $\acute{\rho}_\varepsilon$ (the subdifferential of function $\mathrm{tr}(S\log S)$ at $\acute{\rho}_\varepsilon$ is $\log(\acute{\rho}_\varepsilon)+I_m$, see [53]), we get

$$\big\langle\log\acute{\rho}_\varepsilon,\acute{\rho}_\varepsilon-\rho'\big\rangle\leq 0,$$

since $\big\langle I_m,\acute{\rho}_\varepsilon-\rho'\big\rangle=0$. Meanwhile, suppose $r=\mathrm{rank}(\rho)$ and $\rho=\sum_{i=1}^r\lambda_jP_j$ with eigenvalues $\lambda_j\in[0,1]$(repeated with their multiplicities) and one dimensional orthogonal eigenprojectors P_j. We extend $P_j, j=1,\dots,r$ to the complete orthogonal resolution of the identity $P_j, j=1,\dots,m$. Then

$$\begin{aligned}\log\rho'=&\log\Big(\big(1-\delta\big)\rho+\delta\frac{I_m}{m}\Big)=\sum_{i=1}^r\log\Big((1-\delta)\lambda_j+\delta/m\Big)+\sum_{j=r+1}^m\log(\delta/m)P_j\\ =&\sum_{j=1}^r\log\Big(1+(1-\delta)m\lambda_j/\delta\Big)P_j+\log(\delta/m)I_m.\end{aligned}$$

Therefore,

$$\begin{aligned}K(\acute{\rho}_\varepsilon,\rho')\leq-\big\langle\log\rho',\acute{\rho}_\varepsilon-\rho'\big\rangle=&\Big\langle\sum_{j=1}^r\log\Big(1+(1-\delta)m\lambda_j/\delta\Big)P_j,\acute{\rho}_\varepsilon-\rho'\Big\rangle\\ \leq&\Big\|\sum_{j=1}^r\log\Big(1+(1-\delta)m\lambda_j/\delta\Big)P_j\Big\|_2\|\acute{\rho}_\varepsilon-\rho'\|_2\end{aligned}$$

Note that $\|\acute{\rho}_\varepsilon - \rho'\|_2 \leq \|\acute{\rho}_\varepsilon - \rho\|_2 + \delta\|\rho - I_m/m\|_2 \leq \|\acute{\rho}_\varepsilon - \rho\|_2 + 2\delta$ and

$$\begin{aligned}\Big\|\sum_{j=1}^{r} \log\Big(1+(1-\delta)m\lambda_j/\delta\Big)P_j\Big\|_2 =& \Big(\sum_{j=1}^{r} \log^2\big(1+(1-\delta)m\lambda_j/\delta\big)\Big)^{1/2} \\ \leq& \sqrt{r}\log(m/\delta).\end{aligned}$$

Then, together with (2.5.4), we get

$$\begin{aligned}K(\acute{\rho}_\varepsilon, \rho') \leq& \sqrt{r}(\|\acute{\rho}_\varepsilon - \rho\|_2 + 2\delta)\log\frac{mn}{\sigma_\xi} \\ \leq& Cm\sqrt{\big(\varphi_1(n)+\varphi_2(n)\big)r}\log\frac{mn}{\sigma_\xi} + 2\frac{\sigma_\xi}{n}\sqrt{r}\log\frac{mn}{\sigma_\xi}\end{aligned} \tag{2.5.14}$$

Recall that $K(\rho'\|\acute{\rho}_\varepsilon) \leq K(\acute{\rho}_\varepsilon, \rho')$ and Lemma 10 in Section 2.3.2, we get $K(\rho\|\acute{\rho}_\varepsilon) \leq 2K(\rho'\|\acute{\rho}_\varepsilon) + 2\frac{\sigma_\xi}{n}\log\frac{en}{\sigma_\xi}$. Replacing $K(\rho'\|\acute{\rho}_\varepsilon)$ with the right hand side of (2.5.14), we obtain (2.5.5). □

Theorem 20. *Suppose Assumption 1 and 4 hold with rank(ρ) $\leq r$. Under the conditions (2.5.6) and the choice of $\varepsilon = C_1\sigma_\xi\sqrt{\frac{\log(2m)}{nm}}$ for some large enough constant $C_1 > 0$, there exists a constant $C > 0$ such that with probability at least $1 - \frac{1}{m}$,*

$$\|\acute{\rho}_\varepsilon - \rho\|_p \leq C\bigg(\frac{\bar{\sigma}m^{\frac{3}{2}}r^{1/p}}{\sqrt{n}}\log^3 m\log^3 n \bigwedge \Big(\frac{\bar{\sigma}m^{3/2}}{\sqrt{n}}\Big)^{1-\frac{1}{p}}\big(\log^3 m\log^3 n\big)^{1-\frac{1}{p}}\bigg)\bigwedge 2, \tag{2.5.15}$$

for all $1 \leq p \leq +\infty$ and $\bar{\sigma} := \big(\sigma_\xi \vee \frac{U}{\sqrt{m}}\big)$.

Remark 8. *In the case of Pauli measurements, we consider that $\sigma_\xi \asymp \frac{U}{\sqrt{m}}$, i.e. every Y_i is taken as the average of m outcomes from independent measurements (this is also the experimental scheme proposed in [30, section II.A]). Note that the condition $\sigma_\xi \geq \frac{U}{\sqrt{m}}$ is also needed in the proof of the Schatten p-norm convergence rates of the projection estimator, see Section 2.4.1 The bound (2.5.15) is equivalent(up to logarithmic terms) to*

$$\|\acute{\rho}_\varepsilon - \rho\|_p \lesssim \sqrt{\frac{m}{n}}r^{1/p}\bigwedge\Big(\sqrt{\frac{m}{n}}\Big)^{1-\frac{1}{p}}\bigwedge 1, \tag{2.5.16}$$

for all $1 \leq p \leq +\infty$. It matches the minimax lower bounds shown in Theorem 4 in Section 2.2 by setting $\sigma_\xi = \frac{1}{m}$ there. Essentially, it means that the "complexity" of *the problem is controlled by a "truncated rank"* $r \wedge \sqrt{\frac{n}{m}}$*. Basically, bound (2.5.16) indicates that whenever* $\sigma_\xi \geq \frac{1}{m}$*, estimator* $\acute{\rho}_\varepsilon$ *can achieve optimal convergence rates (up to logarithmic factors) in all the Schatten p-norms for* $p \in [1, +\infty]$*. Note that even though our proof does not require* $n \gtrsim mr$*(with logarithmic terms), the bound (2.5.16) is nontrivial only when* $n \gtrsim m$ *for* $1 \leq p \leq +\infty$*.*

Moreover, it worths to point out that actually

$$\bar{\sigma} = \big(\sigma_\xi \vee \frac{U}{\sqrt{m}} \|\acute{\rho}_\varepsilon - \rho\|_1 \vee U^2 \sqrt{\frac{m}{n}} \|\acute{\rho}_\varepsilon - \rho\|_1\big)$$

in (2.5.15). Then, by the bound (2.5.8) and $U = \frac{1}{\sqrt{m}}$*, we get from (2.5.15) that (up to logarithmic factors)*

$$\|\acute{\rho}_\varepsilon - \rho\|_p \lesssim \Big(\frac{\sigma_\xi m^{3/2} r^{1/p}}{\sqrt{n}} \vee \frac{\sigma_\xi m^2 r^{1+\frac{1}{p}}}{n}\Big) \bigwedge \Big(\frac{\sigma_\xi m^{3/2}}{\sqrt{n}} \vee \frac{\sigma_\xi m^2 r}{n}\Big)^{1-\frac{1}{p}} \bigwedge 1$$

for all $1 \leq p \leq +\infty$*. This bound is important because of its linear dependence on the noise level* σ_ξ*(see also [86]), which can be significantly small in certain situations. If* $n \gtrsim mr^2$*(with logarithmic terms), we get a cleaner bound(recall also Remark 7),*

$$\|\acute{\rho}_\varepsilon - \rho\|_p \lesssim \Big(\frac{\sigma_\xi m^{3/2} r^{1/p}}{\sqrt{n}}\Big) \bigwedge 1, \quad 1 \leq p \leq +\infty, \tag{2.5.17}$$

which holds even when σ_ξ *is significantly smal (it even holds when* $\sigma_\xi = 0$*). It is interesting to notice that the condition* $n \gtrsim mr^2$*(with logarithmic factors) is also needed in proving the optimal convergence rates in Schatten p-norms of Dantzig estimator for estimating general low rank matrices with Gaussian measurements, see [100]. It is still an open problem that whether this condition is necessary.*

Proof of Theorem 20. We begin with the proof of the spectral norm $\|\acute{\rho}_\varepsilon - \rho\|_\infty$. Note that

$$\frac{\|\acute{\rho}_\varepsilon - \rho\|_\infty}{m^2} \leq \Big\|\frac{1}{n}\sum_{i=1}^n \langle \acute{\rho}_\varepsilon - \rho, X_i\rangle X_i\Big\|_\infty + \Big\|\frac{1}{n}\sum_{i=1}^n \langle \acute{\rho}_\varepsilon - \rho, X_i\rangle X_i - \mathbb{E}\langle \acute{\rho}_\varepsilon - \rho, X\rangle X\Big\|_\infty.$$

The first term is upper bounded by $2\varepsilon = C_1\sigma_\xi\sqrt{\frac{\log(2m)}{nm}}$ with probability at least $1-\frac{1}{2m}$, since $\acute{\rho}_\varepsilon \in \Lambda(\varepsilon)$ and,

$$\Big\|\frac{1}{n}\sum_{i=1}^{n}\big\langle \acute{\rho}_\varepsilon-\rho, X_i\big\rangle X_i\Big\|_\infty \le \varepsilon + \|\Xi_1\|_\infty.$$

By the definition of spectral norm, the second term can be written as follows (recall the definition of $\mathcal{A}(\Delta)$ in Lemma 22):

$$\begin{aligned}\Big\|\frac{1}{n}\sum_{i=1}^{n}&\big\langle \acute{\rho}_\varepsilon-\rho, X_i\big\rangle X_i - \mathbb{E}\big\langle \acute{\rho}_\varepsilon-\rho, X\big\rangle X\Big\|_\infty \\ &= \sup_{V\in\mathcal{A}(\frac{1}{m})}\Big|\frac{1}{n}\sum_{i=1}^{n}\big\langle \acute{\rho}_\varepsilon-\rho, X_i\big\rangle\big\langle V, X_i\big\rangle - \mathbb{E}\big\langle \acute{\rho}_\varepsilon-\rho, X\big\rangle\big\langle V, X\big\rangle\Big| \qquad (2.5.18)\\ &\le \|\acute{\rho}_\varepsilon-\rho\|_1\alpha_n\Big(\frac{\|\acute{\rho}_\varepsilon-\rho\|_{L_2(\Pi)}}{\|\acute{\rho}_\varepsilon-\rho\|_1}, \frac{1}{m}\Big).\end{aligned}$$

To this end, we can apply Lemma 22 with $\delta^- = \frac{1}{2m}$ and $\delta^+ = \frac{1}{m}$ and get

$$\alpha_n\Big(\frac{\|\acute{\rho}_\varepsilon-\rho\|_{L_2(\Pi)}}{\|\acute{\rho}_\varepsilon-\rho\|_1}, \frac{1}{m}\Big) \le C\Big(\frac{U}{\sqrt{m}}\frac{\log^{3/2} m\log^{3/2} n}{\sqrt{nm}} + U^2\frac{\log^3 m\log^3 n}{n}\Big),$$

which holds with probability at least $1-\frac{1}{2m}$. We can simply replace $\|\acute{\rho}_\varepsilon-\rho\|_1$ with 2 in (2.5.18) and get that, with probability at least $1-\frac{1}{m}$,

$$\begin{aligned}\frac{\|\acute{\rho}_\varepsilon-\rho\|_\infty}{m^2} \le& C\Big(\big(\sigma_\xi\vee\frac{U}{\sqrt{m}}\big)\frac{\log^{3/2} m\log^{3/2} n}{\sqrt{nm}} + U^2\frac{\log^3 m\log^3 n}{n}\Big)\\ \le& C\Big(\sigma_\xi\vee\frac{U}{\sqrt{m}}\vee U^2\sqrt{\frac{m}{n}}\Big)\frac{\log^3 m\log^3 n}{\sqrt{nm}}.\end{aligned}$$

Observe that if $U^2\sqrt{\frac{m}{n}} \ge \frac{U}{\sqrt{m}}$, then $U\frac{m}{\sqrt{n}} \ge 1$. Since $\|\acute{\rho}_\varepsilon-\rho\|_p$ has a trivial upper bound 2, we conclude that the term $\sigma_\xi\vee\frac{U}{\sqrt{m}}$ will be sufficient in the above bounds. Therefore,

$$\|\acute{\rho}_\varepsilon-\rho\|_\infty \le C\bar{\sigma}\sqrt{\frac{m^3}{n}}\log^3 m\log^3 n\bigwedge 2.$$

By the proof of Lemma 17 with $p=1$, we get that with the same probability,

$$\|\acute{\rho}_\varepsilon-\rho\|_1 \le 2\|\mathcal{P}_L(\acute{\rho}_\varepsilon-\rho)\|_1 \le 2r\|\acute{\rho}_\varepsilon-\rho\|_\infty \le C\bar{\sigma}r\sqrt{\frac{m^3}{n}}\log^3 m\log^3 n\bigwedge 2,$$

where L denotes the support of ρ. Applying the *interpolation inequality* from Lemma 6,

$$\|\acute{\rho}_\varepsilon-\rho\|_p \le \|\acute{\rho}_\varepsilon-\rho\|_1^{1/p}\|\acute{\rho}_\varepsilon-\rho\|_\infty^{1-1/p}$$

for all $1\le p\le+\infty$, we will get bound (2.5.15). □

2.5.2 Upper bounds of the product empirical processes

Proof of Lemma 22. For any $\Delta \in [0,1]$, define the following quantity

$$\beta_n(\Delta) := \sup_{A \in \mathcal{A}(\Delta)} \Big| \frac{1}{n} \sum_{i=1}^n \big\langle A, X_i \big\rangle^2 - \mathbb{E}\big\langle A, X \big\rangle^2 \Big|.$$

For all $A_1 \in \mathcal{A}(\Delta_1)$ and $A_2 \in \mathcal{A}(\Delta_2)$, the following fact is clear,

$$\begin{aligned}
\Big| \frac{1}{n} \sum_{i=1}^n \big\langle A_1, X_i \big\rangle \big\langle A_2, X_i \big\rangle - \mathbb{E}\big\langle A_1, X \big\rangle \big\langle A_2, X \big\rangle \Big| \\
\leq & \frac{1}{4} \Big| \frac{1}{n} \sum_{i=1}^n \big\langle A_1 + A_2, X_i \big\rangle^2 - \mathbb{E}\big\langle A_1 + A_2, X \big\rangle^2 \Big| \\
+ & \frac{1}{4} \Big| \frac{1}{n} \sum_{i=1}^n \big\langle A_1 - A_2, X_i \big\rangle^2 - \mathbb{E}\big\langle A_1 - A_2, X \big\rangle^2 \Big| \\
\leq & \beta_n\big(\|A_1 + A_2\|_{L_2(\Pi)}/2 \big) + \beta_n\big(\|A_1 - A_2\|_{L_2(\Pi)}/2 \big),
\end{aligned}$$

where the last inequality holds because $\frac{A_1 \pm A_2}{2} \in \mathcal{A}\big(\|A_1 \pm A_2\|_{L_2(\Pi)}/2 \big)$. Observe that $\frac{\|A_1 \pm A_2\|_{L_2(\Pi)}}{2} \leq \frac{\Delta_1 + \Delta_2}{2}$ for all $A_1 \in \mathcal{A}(\Delta_1)$ and $A_2 \in \mathcal{A}(\Delta_2)$. Therefore,

$$\alpha_n(\Delta_1, \Delta_2) \leq 2\beta_n\Big(\frac{\Delta_1 + \Delta_2}{2} \Big).$$

It suffices to prove an upper bound for $\beta_n(\Delta)$ for $\Delta \in [\delta^-, \delta^+]$. Remember that the upper bound for $\beta_n(\Delta)$ has been claimed in Section 2.3.3 without proof. We give the proof based on Dudley's entropy bound and the $L_\infty(\Pi_n)$ complexity of unit ball in $\mathbb{H}_m$ equipped with Schatten 1-norm.

Assume that $\Delta \in [\delta^-, \delta^+]$, the main strategy is that we derive the upper bound of $\beta_n(\Delta)$ for $\Delta \in [\delta_j, \delta_{j+1}]$ with $\delta_j = 2^j \delta^-$ for $j = 0, 1, 2, \ldots, \lfloor \log_2 \frac{\delta^+}{\delta^-} \rfloor$. Then, we take the bounds uniformly over the whole range $[\delta^-, \delta^+]$, which is a standard argument.

For a fixed j such that $\Delta \in [\delta_j, \delta_{j+1}]$, we apply Bousquet's version (see [52, Chapter 2]) of Talagrand's inequality for empirical processes and get that with probability at least $1 - e^{-t}$,

$$\beta_n(\Delta) \leq 2\mathbb{E}\beta_n(\Delta) + 2U\Delta\sqrt{\frac{t}{n}} + 2U^2\frac{t}{n}$$

for any $t \geq 1$. We used the facts

$$\sup_{A \in \mathcal{A}(\Delta)} \mathbb{E}\langle A, X\rangle^4 \leq U^2 \sup_{A \in \mathcal{A}(\Delta)} \mathbb{E}\langle A, X\rangle^2 \leq U^2 \Delta^2$$

and $\langle A, X\rangle^2 \leq U^2$. To control $\mathbb{E}\beta_n(\Delta)$, by the symmetrization inequality, we get

$$\mathbb{E}\beta_n(\Delta) \leq 2\mathbb{E}_X \mathbb{E}_\varepsilon \sup_{A \in \mathcal{A}(\Delta)} \Big| \frac{1}{n} \sum_{i=1}^n \varepsilon_i \langle A, X_i\rangle^2 \Big|$$

where $\varepsilon_1, \ldots, \varepsilon_n$ are *i.i.d.* Rademacher random variables.

Now, we fix $X_1, X_2, \ldots, X_n$ and consider the sub-Gaussian process indexed by $A \in \mathcal{A}(\Delta)$ defined as

$$G_A := \frac{1}{\sqrt{n}} \sum_{i=1}^n \varepsilon_i \langle A, X_i\rangle^2.$$

This is a sub-Gaussian process with respect to the pseudo-distance

$$d(A_1, A_2) := \mathbb{E}^{1/2}(G_{A_1} - G_{A_2})^2 = \Big(\frac{1}{n} \sum_{i=1}^n \langle A_1 - A_2, X_i\rangle^2 \langle A_1 + A_2, X_i\rangle^2\Big)^{1/2}$$
$$\leq 2\sigma_n \|A_1 - A_2\|_{L_\infty(\Pi_n)},$$

where $\sigma_n^2 := \sup_{A \in \mathcal{A}(\Delta)} \frac{1}{n} \sum_{i=1}^n \langle A, X_i\rangle^2$. According to Dudley's entropy bound (see Theorem 3 in Section 1.4),

$$\mathbb{E}_\varepsilon \sup_{A \in \mathcal{A}(\Delta)} |G_A| \lesssim \int_0^{4U\sigma_n} H^{1/2}(\mathcal{A}(\Delta), d, u) du,$$

where the entropy number $H(\mathcal{A}(\Delta), d, u) = \log N(\mathcal{A}(\Delta), d, u)$, the logarithmic of u-covering number of $\mathcal{A}(\Delta)$ with respect to the pseudo-metric d.

Since $d(A_1, A_2) \leq 2\sigma_n \|A_1 - A_2\|_{L_\infty(\Pi_n)}$, we have

$$H^{1/2}(\mathcal{A}(\Delta), d, u) \leq H^{1/2}(\mathcal{A}(\Delta), L_\infty(\Pi_n), \frac{u}{2\sigma_n}).$$

As a consequence,

$$\mathbb{E}_\varepsilon \sup_{A \in \mathcal{A}(\Delta)} G_A \lesssim \int_0^{4U\sigma_n} H^{1/2}(\mathcal{A}(\Delta), L_\infty(\Pi_n), \frac{u}{2\sigma_n}) du$$
$$\leq 2\sigma_n \int_0^{2U} H^{1/2}(\mathcal{A}(\Delta), L_\infty(\Pi_n), u) du.$$

The $L_\infty(\Pi_n)$-complexity of unit balls in $\mathbb{H}_m$ equipped with nuclear norm distance has been thoroughly studied. When $X_1,\dots,X_n$ are fixed, the vector $\big(\langle A,X_1\rangle,\dots,\langle A,X_n\rangle\big)'$ belongs to the cube $[-U,U]^n$. The l_∞-covering number is upper bounded by

$$N(\mathcal{A}(\Delta),L_\infty(\Pi_n),u)\leq\Big(1+\frac{2U}{u}\Big)^n.$$

This bound will be used when u is small. When u is large, we apply the following bound, see [64, (21)], [38],[4, Lemma A5],

$$N(\mathcal{A}(\Delta),L_\infty(\Pi_n),u)\leq\exp\Big\{C\frac{U^2\log^3 m\log n}{u^2}\Big\}$$

for some constant $C>0$. Then, by setting $K=\frac{U}{\sqrt{n}}$,

$$\begin{aligned}\mathbb{E}_\varepsilon\sup_{A\in\mathcal{A}(\Delta)}G_A\lesssim&\sigma_n\int_0^K\sqrt{n}\log^{\frac12}\Big(1+\frac{2U}{u}\Big)du+\sigma_n\int_K^{2U}\frac{U\log^{3/2}m\log^{1/2}n}{u}du\\ \lesssim&\sigma_n\sqrt{n}K\log(1+\frac{2U}{K})+U\sigma_n\log^{3/2}m\log^{1/2}n\log\frac{U}{K}\\ \lesssim&U\sigma_n\log^{3/2}m\log^{3/2}n.\end{aligned}$$

Therefore, we conclude that

$$\mathbb{E}\beta_n(\Delta)=\frac{1}{\sqrt{n}}\mathbb{E}_X\mathbb{E}_\varepsilon\sup_{A\in\mathcal{A}(\Delta)}G_A\lesssim\frac{1}{\sqrt{n}}\mathbb{E}_XU\sigma_n\log^{3/2}m\log^{3/2}n.$$

Note that

$$\begin{aligned}\mathbb{E}_X\sigma_n=\mathbb{E}_X\sqrt{\sup_{A\in\mathcal{A}(\Delta)}\frac{1}{n}\sum_{i=1}^n\big\langle A,X_i\big\rangle^2}\leq&\sqrt{\mathbb{E}_X\sup_{A\in\mathcal{A}(\Delta)}\frac{1}{n}\sum_{i=1}^n\big\langle A,X_i\big\rangle^2}\\ \leq&\sqrt{\mathbb{E}\beta_n(\Delta)+\Delta^2}.\end{aligned}$$

Therefore, we get

$$\mathbb{E}\beta_n(\Delta)\lesssim\sqrt{\mathbb{E}\beta_n(\Delta)+\Delta^2}\frac{U\log^{3/2}m\log^{3/2}n}{\sqrt{n}},$$

which can be simplified into

$$\mathbb{E}\beta_n(\Delta)\lesssim\Delta U\frac{\log^{3/2}m\log^{3/2}n}{\sqrt{n}}+\frac{U^2\log^3 m\log^3 n}{n}.$$

Therefore, for $\Delta \in [\delta_j, \delta_{j+1}]$, with probability at least $1 - e^{-t}$,

$$\beta_n(\Delta) \leq C\Delta U \frac{\log^{3/2} m \log^{3/2} n}{\sqrt{n}} + CU^2 \frac{\log^3 m \log^3 n}{n} + 2U\Delta\sqrt{\frac{t}{n}} + 2U^2\frac{t}{n}.$$

for some $C > 0$. By making it uniform over all $j = 0, 1, \ldots, \lfloor \log_2 \frac{\delta^+}{\delta^-} \rfloor$ and adjusting t to $t + \log(\log_2 \frac{\delta^+}{\delta^-} + 2)$, we get the uniform upper bound of $\beta_n(\Delta)$ for $\Delta \in [\delta^-, \delta^+]$. $\square$

CHAPTER III

SIMULATION RESULTS OF LOW RANK DENSITY MATRICES ESTIMATION

The main purpose of this chapter is to discuss the numerical algorithms for solving the estimators studied in Chapter 2, including the least squares estimator, the projection estimator and the Dantzig-type estimator. There are many algorithms available for solving the optimization problems involved in these estimators. For instance, the proximal gradient method is a popular class of algorithms for solving the constrained convex optimization problems, see [65] and [74], etc. In principle, the proximal gradient method is equivalent to the projected gradient descent method. In this chapter, we focus on the alternating minimization method for the least squares estimator. The Dantzig-type estimator is usually formulated as a semi-definite programming.

3.1 Algorithms

3.1.1 The ADMM algorithm for the least squares estimator

The alternating direction method of multipliers (ADMM) is a popular algorithm in convex optimization problems, see a comprehensive introduction in [12], the application of ADMM in matrix estimation problems in [63], [24] and refernces therein. The ADMM algorithm has shown great success in many problems, which (empirically) converges much faster than many famous algorithms, such as Nesterov's accelerated gradient algorithms, see [73].

Recall that the least squares estimator with penalization is defined as

$$\tilde{\rho}_\varepsilon := \operatorname*{arg\,min}_{S \in \mathcal{S}_m} L_\varepsilon(S), \tag{3.1.1}$$

where $L_\varepsilon(S) := \frac{1}{n}\sum_{j=1}^n \big(Y_j - \langle S, X_j\rangle\big)^2 + \varepsilon \cdot \mathrm{tr}\big(S \log S\big)$. The optimization problem in (3.1.1) belongs to the standard form of optimization problems considered in the ADMM algorithms, see [12]. It usually involves the sum of two functions of the underlying parameter, i.e., a loss function and a penalization function. Then ADMM algorithm solves the optimization problem by introducing a new variable such that the following function is considered in stead of $L_\varepsilon(S)$:

$$L_\varepsilon(S_1, S_2) := \frac{1}{n}\sum_{j=1}^{n}\big(Y_j - \langle S_1, X_j\rangle\big)^2 + \varepsilon \cdot \mathrm{tr}\big(S_2 \log S_2\big).$$

Then, it is easy to check that we can define $\tilde{\rho}_\varepsilon$ equivalently as

$$(\tilde{\rho}_\varepsilon, \tilde{\rho}_\varepsilon) := \operatorname*{arg\,min}_{S_1 \in \mathcal{S}_m, S_1 = S_2} L_\varepsilon(S_1, S_2).$$

The augmented Lagrangian multipliers of function $L_\varepsilon(S_1, S_2)$ is defined as

$$L_\varepsilon(S_1, S_2, Z, \lambda) := L_\varepsilon(S_1, S_2) + \big\langle S_1 - S_2, Z\big\rangle + \frac{\lambda}{2}\|S_1 - S_2\|_2^2,$$

for some $\lambda \geq 0$ and $Z \in \mathbb{H}_m$. The parameter λ can be fixed and can also be predetermined a sequence of sizes $\{\lambda_k\}_{k\geq 1}$. Then ADMM algorithm updates S_1 and S_2 aternatively and the multiplier Z is updated by the difference between S_1 and S_2. The algorithm is listed as in Algorithm 1, with tolerance $\varepsilon_{tol} > 0$ being the stopping criterion and $\max\_Iteration$ being the maximum number of iterations.

In order to update S_2, we need to solve the following optimization problem:

$$S_2^{(k+1)} := \operatorname*{arg\,min}_{S \in \mathcal{S}_m} \frac{\lambda}{2}\Big\|S_1^{(k+1)} - S + \frac{Z^{(k)}}{\lambda}\Big\|_2^2 + \varepsilon \mathrm{tr}\big(S \log S\big).$$

In the case $\varepsilon = 0$, $S_2^{(k+1)}$ is equivalent to the projection of the matrix $S_1^{(k+1)} + \frac{Z^{(k)}}{\lambda}$ onto the compact and convex set $\mathcal{S}_m$ whose explicit solution is given as in Lemma 20 in Section 2.4.1. It is clear that when $\varepsilon > 0$, there is no explicit solution for $S_2^{(k+1)}$, in which cases, it is usually solved by iterative algorithms such as [8], which is often referred as the projected gradient descent algorithm.

Algorithm 1 ADMM Algorithm

```
Set up value of max_Iteration and tolerance $\varepsilon_{\text{tol}} > 0$
Initiate $S_1^{(0)} \in \mathbb{H}_m$, $S_2^{(0)} \in \mathcal{S}_m$ and $Z^{(0)} = \mathbf{0} \in \mathbb{H}_m$, k=0
while k<max_Iteration do
    $S_1^{(k+1)} = \underset{S \in \mathbb{H}_m}{\arg\min} \quad \frac{1}{n} \sum_{j=1}^{n} (Y_j - \text{Tr}(SX_j))^2 + \left\langle S - S_2^{(k)}, Z^{(k)} \right\rangle + \frac{\lambda}{2} ||S - S_2^{(k)}||_2^2$
    $S_2^{(k+1)} = \underset{S \in \mathcal{S}_m}{\arg\min} \quad \varepsilon \cdot \text{Tr}(S \log(S)) + \left\langle S_1^{(k+1)} - S, Z^{(k)} \right\rangle + \frac{\lambda}{2} ||S_1^{(k+1)} - S||_2^2$
      $\triangleright$ $S_2^{(k+1)}$ is also the minimizer of $\frac{\lambda}{2} ||S_1^{(k+1)} - S + Z^{(k)}/\lambda||_2^2 + \varepsilon \cdot \text{Tr}(S \log(S))$
    $Z^{(k+1)} = Z^{(k)} + \lambda(S_1^{(k+1)} - S_2^{(k+1)})$
    if $||S_2^{(k+1)} - S_2^{(k)}||_2^2 \leq \varepsilon_{\text{tol}}$ or $||Z^{(k+1)} - Z^{(k)}||_2^2 \leq \varepsilon_{\text{tol}} \lambda^2$ then
        Reaching the tolerance. Return $S_2^{(k+1)}$.
    end if
    k=k+1
end while
Return $S_2^{(k+1)}$.
```

3.1.2 The computational advantages of the projection estimator

An advantage of the minimal distance estimator $\check{\rho} = \pi_{\mathcal{S}_m}(\hat{Z})$ is the simplicity of its computational implementation. The computation of the matrix $\hat{Z} = \frac{m^2}{n} \sum_{i=1}^{n} Y_i X_i$ requires $O(nm^2)$ operations. It is followed by an eigen-decomposition of Z that requires $O(m^3)$ operations(see [33]); there exist efficient software packages designed for this kind of tasks, for instance, LINPACK and PROPACK, etc.). As it is shown in the previous section, the problem of computing $\pi_{\mathcal{S}_m}(\hat{Z})$ then reduces to projecting of the vector of eigenvalues of Z arranged in a non-increasing order onto the simplex Δ_m. The last problem has been studied in the literature (see [70], [83], [25]) and it has an explicit solution of computational complexity proportional to m (see the proof of Lemma 20). Thus, the computational implementation of the minimal distance estimator $\check{\rho}$ requires $O((n+m)m^2)$ operations.

Recall that the matrix LASSO estimator for estimating density matrices is defined as

$$\hat{\rho} := \underset{S \in \mathcal{S}_m}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} \Big(Y_i - \langle S, X_i \rangle \Big)^2 \tag{3.1.2}$$

which is actually the least squares estimator. Clearly, there is no explicit solution for this optimization problem and it is usually solved by iterative algorithms, such as the ADMM algorithm introduced in Section 3.1.1. In addition to the ADMM algorithm explained in Section 3.1.1, a well know iterative singular value thresholding (SVT) algorithm was proposed in [14], and was also implemented in quantum compressed sensing in [30]. The main idea is that (3.1.2) is equivalent to the following optimization problem: for any $\tau > 0$,

$$\hat{\rho} := \underset{S \in \mathcal{S}_m, Z \in \mathbb{H}_m, S = Z}{\arg\min} \frac{m^2}{n} \sum_{i=1}^{n} \Big(Y_i - \big\langle Z, X_i \big\rangle \Big)^2 + \tau \|S - Z\|_2^2.$$

The proposed algorithm updates Z and S alternatively, with the only constraint for S being that $S \in \mathcal{S}_m$. Therefore, the main ingredient of SVT is the following iterative updating rule (with initial $Z_0 = 0$): for $k = 1, 2, \ldots$,

$$\begin{cases} S_k = \pi_{\mathcal{S}_m}(Z_{k-1}) \\ Z_k = S_k + \delta_k \big(\hat{Z} - \frac{m^2}{n} \sum_{i=1}^{n} \big\langle S_k, X_i \big\rangle X_i \big) \end{cases} \tag{3.1.3}$$

with certain pre-determined step sizes $\delta_k > 0$. The algorithm terminates at some step $k = N$ and outputs $S_N \in \mathcal{S}_m$ when $\|S_N - S_{N-1}\|_2 \leq \varepsilon$ for some numerical threshold $\varepsilon > 0$. It is clear that the minimal distance estimator $\check{\rho}$ can be produced by the above algorithm with one iteration and the initialization $Z_0 = \hat{Z}, \delta_1 = 0$. When the number of qubits k is not small (for instance, about 20) and the dimension m is very large, the iterative algorithm (3.1.3) is much more computationally expensive than the algorithm for the minimal distance estimator (since every iteration requires the eigen-decomposition of a high dimensional matrix).

3.1.3 A semidefinite program for the Dantzig-type estimator

Given the data $\mathcal{D}_n := \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$, define a linear map as follows:

$$\mathcal{T} : \mathbb{H}_m \mapsto \mathbb{C}^n. \quad \mathcal{T}(S) := \big(\big\langle S, X_1 \big\rangle, \ldots, \big\langle S, X_n \big\rangle \big)' \in \mathbb{C}^n.$$

Its adjoint operator is easily defined as

$$\mathcal{T}^\star : \mathbb{C}^n \mapsto \mathbb{H}_m. \quad \mathcal{T}^\star(r) := \sum_{i=1}^{n} r_i X_i \in \mathbb{H}_m.$$

Then, the Dantzig-type estimator can be equivalently written as (by written $\mathcal{Y} = (Y_1, \ldots, Y_n) \in \mathbb{R}^n$)

$$\acute{\rho}_\varepsilon^D := \arg\min \ \big\{ \|S\|_1, \big\| \mathcal{T}^\star\big(\mathcal{Y} - \mathcal{T}(S)\big) \big\|_\infty \leq \varepsilon, S \in \mathcal{S}_m \big\}. \tag{3.1.4}$$

It is also possible to replace the nuclear norm $\|S\|_1$ with the negative von Neumann entropy $-V(S) = \operatorname{tr}(S \log S)$. Note that the feasible set in (3.1.4) involves an upper bound on the spectral norm of $\mathcal{T}^\star\big(\mathcal{Y} - \mathcal{T}(S)\big) \in \mathbb{H}_m$ which can be written as

$$\begin{pmatrix} \varepsilon I_m & \mathcal{T}^\star\big(\mathcal{Y} - \mathcal{T}(S)\big) \\ \mathcal{T}^\star\big(\mathcal{Y} - \mathcal{T}(S)\big) & \varepsilon I_m \end{pmatrix} \succcurlyeq 0.$$

As a result, we can write the Dantzig-type estimator $\acute{\rho}_\varepsilon^D$ as the following semidefinite program:

$$\begin{aligned} \text{minimize} \quad & \|S\|_1 \equiv 1 \text{ or } \operatorname{tr}(S \log S) \\ \text{subject to} \quad & \begin{pmatrix} S & 0 & 0 \\ 0 & \varepsilon I_m & \mathcal{T}^\star\big(\mathcal{Y} - \mathcal{T}(S)\big) \\ 0 & \mathcal{T}^\star\big(\mathcal{Y} - \mathcal{T}(S)\big) & \varepsilon I_m \end{pmatrix} \succcurlyeq 0 \\ & S^\star = S, \operatorname{tr}(S) = 1. \end{aligned} \tag{3.1.5}$$

There are many efficient algorithms (for instance, interior point algorithm, see [39] and [93]) and softwares available for solving semidefinite programs, see [85] and [34].

3.2 Numerical results

In this section, we display the numerical simulation results of several estimators. The data is generated according to the trace regression model with Gaussian noise. The bounded response model is not considered here since it usually requires a larger sample size, and its result can be reproduced by the Gaussian noise model with large noise variance. The numerical results of Dantzig estimator will not be presented since they are pretty close to the least squares estimator.

Table 1: The Schatten p-norms for $n = 600$ and different ranks

Rank	p=1	p=2	p=3	p=4	p=5	p=6	p=7	p=8	p=9	p=10
2	0.733	0.297	0.240	0.222	0.213	0.208	0.205	0.203	0.201	0.200
7	1.068	0.261	0.183	0.161	0.154	0.151	0.150	0.150	0.149	0.149
12	1.057	0.233	0.168	0.154	0.150	0.149	0.148	0.148	0.148	0.148
17	1.057	0.233	0.177	0.166	0.163	0.162	0.162	0.162	0.162	0.162
22	0.954	0.196	0.146	0.136	0.134	0.133	0.133	0.132	0.132	0.132

3.2.1 Low rank estimation with small noise level

The example considered in this section is related to the trace regression model with small noise. The density matrix ρ is considered with 6-qubits such that $m = 2^6 = 64$. The Pauli measurements are sampled uniformly with n being the sample size. The data is generated with Gaussian noise such that $\sigma_\xi = \frac{0.1}{m}$ which we call small noise level. If $\sigma_\xi > \frac{1}{m}$, we refer it to large noise level. The least squares estimator is considered here and the singular value thresholding algorithm is applied as introduced in Section 3.1. We also considered different cases such that ρ has an increasing rank. Note that we are interested in the Schatten p-norm convergence rates for different values of p, which decreases as p increases. We consider that $p = 1, 2, 3, \ldots, 10$. Actually when $p \geq 5$, the Schatten p-norms are nearly equal to the operator norm in many cases.

The low rank density matrix $\rho \in \mathcal{S}_m$ with $m = 64$ is constructed as follows. The eigenvectors are generated randomly from a Gaussian random matrix and the eigenvalues have the following form: if rank$(\rho) = 1$, then we set $\lambda_1(\rho) = 1$; if rank$(\rho) > 1$, then we set $\lambda_1(\rho) = \frac{1}{2}$ and $\lambda_2(\rho) = \ldots = \lambda_r = \frac{1}{2r}$. The reason that we set $\lambda_1(\rho) = \Omega(1)$ is that we don't want $\|\rho\|_\infty$ to decrease as the rank increases in order to emphasize the convergence rates in spectral norm. The ranks considered in this example are $r = 2, 7, 12, 17, 22$. The sample sizes considered are $n = 200, 300, 400, 500, 600, 700, 800, 900, 1000$.

Table 2: The Schatten p-norms for $n = 1000$ and different ranks

Rank	p=1	p=2	p=3	p=4	p=5	p=6	p=7	p=8	p=9	p=10
2	0.331	0.133	0.108	0.100	0.097	0.095	0.093	0.092	0.092	0.091
7	0.873	0.215	0.147	0.127	0.119	0.115	0.113	0.113	0.112	0.112
12	0.876	0.181	0.117	0.099	0.092	0.089	0.087	0.086	0.086	0.086
17	0.896	0.175	0.117	0.103	0.098	0.096	0.096	0.095	0.095	0.095
22	0.835	0.153	0.101	0.089	0.085	0.083	0.082	0.082	0.082	0.082

Table 1 and Table 2 show the average values of $\|\hat{\rho} - \rho\|_p$ for $p = 1, \ldots, 10$ and $\text{rank}(\rho) = 2, 7, 12, 17, 22$ when $n = 600$ and $n = 1000$. For every r and n, the experiments are repeated for 5 times and the average values of $\|\hat{\rho} - \rho\|_p$ are shown in these tables. It is interesting to notice that there is a big gap between the trace norm distance and the Frobenius norm distance. Moreover, when the rank r increases, there is no clear increase in the Schatten p-norm distances. This might be due to the special structure of ρ we constructed.

In Figure 1, we showed the convergence rates of Schatten norms when the rank is 3 and the sample size n is increasing. The Schatten 1-norm, 2-norm and 5-norm are considered when the noise level $\sigma_\xi = \frac{0.1}{m}$. It is clear that when the sample size n increases, the Schatten norm distances decrease which is within our expectation.

Next, we are also interested in the dependence of the Schatten norms on the noise level, namely σ_ξ. In the following, we consider an example with $n = 1000, m = 64, r = 3$ and different levels of noise $m\sigma_\xi = 0.01, 0.05, 0.1, 0.15, 0.20$. The result is shown in Table 3. It is clear that the error rates have a positive relation with the noise level σ_ξ. However, it seems that the error rate is not increasing proportionally with respect to the noise variance.

Figure 1: The convergence rates of Schatten p-norm distances

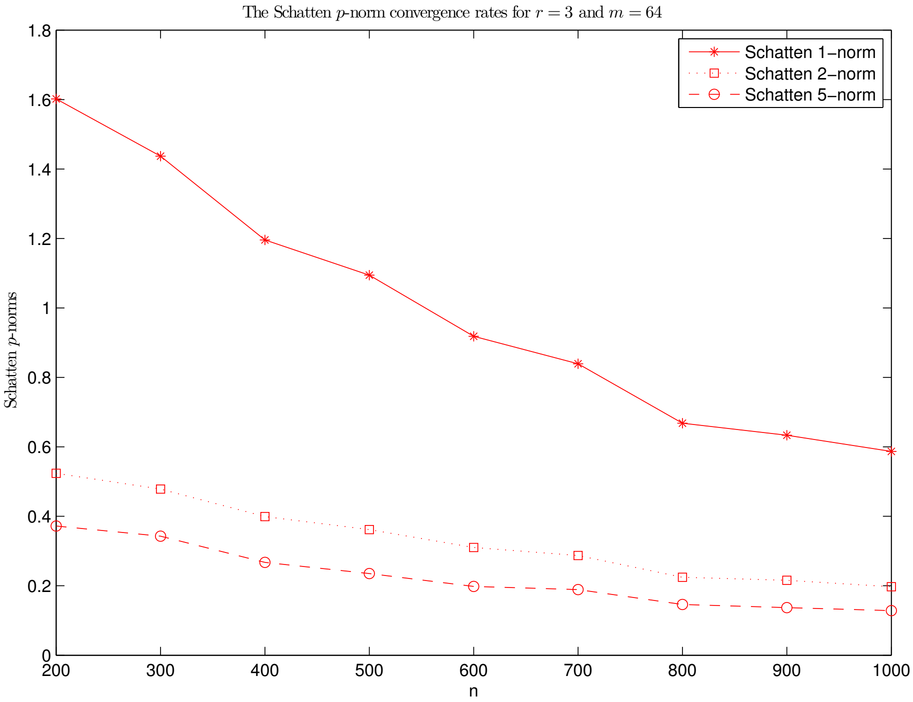


Table 3: The Schatten p-norms for $n = 1000, m = 64, r = 3$ and different noise levels

$m\sigma_\xi$	$p = 1$	$p = 2$	$p = 3$	$p = 4$	$p = 5$	$p = 6$	$p = 7$	$p = 8$	$p = 9$	$p = 10$
0.01	0.377	0.126	0.095	0.085	0.081	0.078	0.076	0.075	0.074	0.074
0.05	0.406	0.136	0.103	0.092	0.086	0.083	0.081	0.080	0.079	0.078
0.10	0.428	0.145	0.109	0.098	0.092	0.089	0.088	0.086	0.086	0.085
0.15	0.533	0.182	0.137	0.123	0.116	0.112	0.110	0.109	0.108	0.107
0.20	0.675	0.230	0.173	0.154	0.145	0.140	0.137	0.135	0.134	0.133

Table 4: The Schatten p-norms, KL-divergence and Hellinger distance for $n = 1000, m = 64, r = 3$ and different choices of regularization parameter ε

ε	$p=1$	$p=2$	$p=3$	$p=4$	$p=5$	$p=6$	$p=7$	$p=8$	$K(\rho\|\tilde{\rho}_\varepsilon)$	$H(\rho, \tilde{\rho}_\varepsilon)$
10^{-4}	0.621	0.207	0.157	0.141	0.133	0.129	0.126	0.125	0.675	0.527
2×10^{-4}	0.599	0.197	0.151	0.136	0.129	0.125	0.123	0.121	0.623	0.515
5×10^{-4}	0.698	0.220	0.173	0.158	0.150	0.146	0.143	0.142	0.529	0.590
10^{-3}	1.079	0.335	0.271	0.251	0.241	0.236	0.233	0.231	0.917	0.788
1.5×10^{-3}	1.274	0.394	0.323	0.300	0.290	0.285	0.282	0.281	1.168	0.884

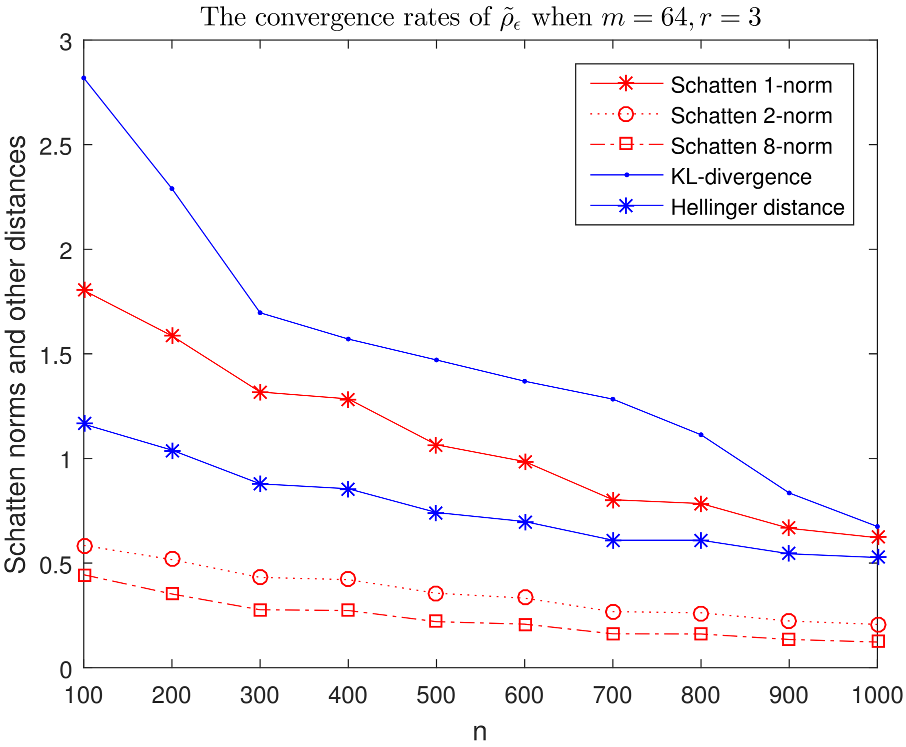
3.2.2 Estimation with von Neumann entropy penalization and small noise level

In this section, we considered the least squares estimator with von Neumann entropy as the penalization. The optimization problem can be solved by the ADMM algorithm introduced in Section 3.1 and also the singular value thresholding algorithm. Actually, both two algorithms produce similar results. In addition to the Schatten p-norms considered as in previous section, we will also include the Kullback-Leibler divergence and Hellinger distance. It worths to point out that due to the unboundedness of the derivative of the entropy function, both the singular value thresholding algorithm and the ADMM algorithm converge slowly, especially when $\varepsilon > 0$ is larger than the optimal choice based on theoretic analysis (see Section 2.3).

In Table 4, we provide the Schatten p-norm distances, Kullback-Leibler divergence and the Hellinger distance of the estimator $\tilde{\rho}_\varepsilon$ according to different choices of ε. In this example, we set $m = 6, r = 3, n = 1000$ and noise level $\sigma_\xi = \frac{0.1}{m}$. For every choice of $\varepsilon = 10^{-4}, 5 \times 10^{-4}, 10 \times 10^{-4}$ and 15×10^{-4}, the experiment is repeated for 5 times and the average errors are taken.

There are several important observations from Table 4. Even though we know that when $\varepsilon > 0$, the estimator $\tilde{\rho}_\varepsilon$ has full rank. The simulation result indicates that when ε is small $\big($for instance, $O(10^{-4})\big)$, $\tilde{\rho}_\varepsilon$ has actually nearly low rank, many

Figure 2: The convergence rates of von Neumann entropy penalized least squares estimator with noise level $\sigma_\xi = \frac{0.1}{m}$ and $\varepsilon = 10^{-4}$.



of its eigenvalues are extremely close to 0. In this cases, the von Neumann entropy penalized estimator $\tilde{\rho}_\varepsilon$ is close to the standard least squares estimator $\hat{\rho}$. Moreover, when $\varepsilon > 0$ is small, the Kullback-Leibler divergence $K(\rho\|\tilde{\rho}_\varepsilon)$ is larger than the Schatten 1-norm distance $\|\tilde{\rho}_\varepsilon - \rho\|_1$. When ε slightly increases, the KL divergence $K(\rho\|\tilde{\rho}_\varepsilon)$ decreases and $\|\tilde{\rho}_\varepsilon - \rho\|_1$ increases such that $K(\rho\|\tilde{\rho}_\varepsilon)$ is becoming smaller than $\|\tilde{\rho}_\varepsilon - \rho\|_1$. If ε becomes even larger, then the KL-divergence $K(\rho\|\tilde{\rho}_\varepsilon)$ is also increasing and is still smaller than $\|\tilde{\rho}_\varepsilon - \rho\|_1$. Note that by the theoretical analysis, we know that the optimal choice of ε is of the order $O\big(\sigma_\xi\sqrt{\frac{\log m}{mn}}\big)$ which should be $O(10^{-5})$ in this example.

Next, we are also interested in the dependence of the convergence rates of $\tilde{\rho}_\varepsilon$ on the sample size n. We still focus on the same example, namely, $m = 64, r = 3$ and $\sigma_\xi = \frac{0.1}{m}$. Moreover, the regularization parameter ε is set to be $\varepsilon = 10^{-4}$. The result is shown as in Figure 2.

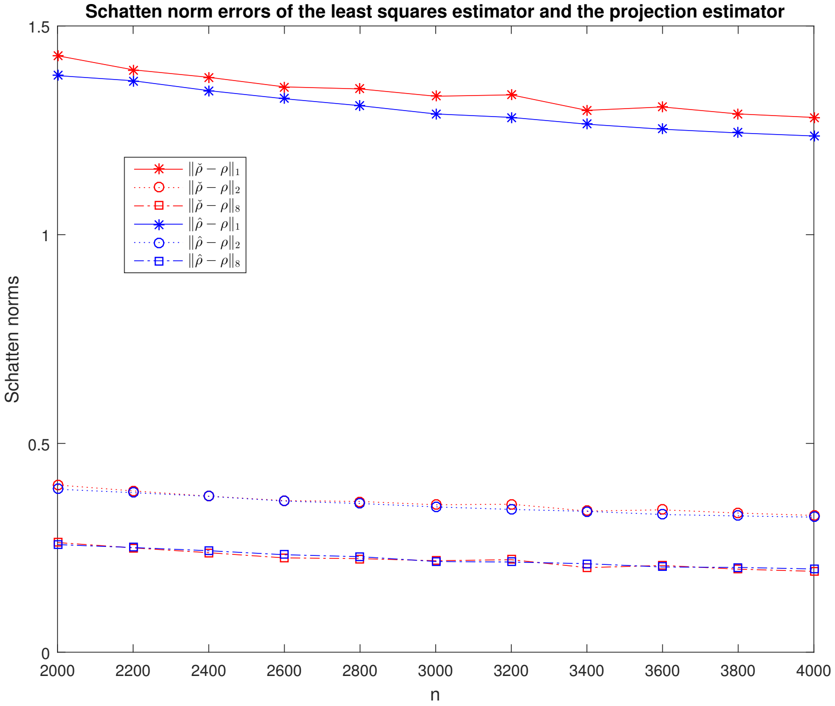
It shows that when $\varepsilon = 10^{-4}$, the Kullback-Leibler divergence is always larger than the trace distance and the Hellinger distance is always between the KL-divergence and trace distance.

3.2.3 The projection estimator when noise level is large

In this section, we consider that the noise level σ_ξ is large such that $\sigma_\xi \geq \frac{1}{m}$. As proved in Section 2.4.1, the simple projection estimator is able to achieve the optimal convergence rates in all the Schatten p-norms for $1 \leq p \leq +\infty$ when $\sigma_\xi \geq \frac{1}{m}$. The goal of this section is to numerically verify this claim. The example considered in this example is $m = 2^6, r = 10$ and the density matrix ρ is constructed as in Section 3.2.1. The noise variance is chosen as $\sigma_\xi = \frac{1}{m}$.

Both the standard least squares estimator and the projection estimator are implemented for this example under the same settings to verify that the simple projection estimator is able to achieve the optimal convergence rates. Moreover, the projection estimator is much more computationally friendly. The results are shown in Figure 3. The least squares estimator is slightly better than the projection estimator in trace norm distance. There is no clear difference in other Schatten norm convergence rates.

Figure 3: Comparison of the convergence rates between the least squares estimator($\hat{\rho}$) and the projection estimator ($\check{\rho}$) when $\sigma_\xi = \frac{1}{m}$.



CHAPTER IV

PERTURBATION OF LINEAR FORMS OF SINGULAR VECTORS UNDER GAUSSIAN NOISE

4.1 Introduction and notations

Analysis of perturbations of singular vectors of matrices under a random noise is of importance in a variety of areas including, for instance, digital signal processing, numerical linear algebra and spectral based methods of community detection in large networks (see [44], [84], [29], [62], [81], [61], [43], [40] and references therein). Recently, random perturbations of singular vectors have been studied in [95], [97], [76], [7]. However, up to our best knowledge, our work proposes first sharp results concerning concentration of *the components* of singular vectors of randomly perturbed matrices. At the same time, there has been interest in the recent literature in so called "delocalization" properties of eigenvectors of random matrices, see [82], [96] and references therein. In this case, the "information matrix" A is equal to zero, $\tilde{A} = X$ and, under certain regularity conditions, it is proved that the magnitudes of the components for the eigenvectors of X (in the case of symmetric square matrix) are of the order $O\big(\frac{\log(n)}{\sqrt{n}}\big)$ with a high probability. This is somewhat similar to the results on "componentwise concentration" of singular vectors of $\tilde{A} = A + X$ proved in this chapter, but the analysis in the case when $A \neq 0$ is quite different (it relies on perturbation theory and on the condition that the gaps between the singular values are sufficiently large).

Later in this section, we provide a formal description of the problem studied in this chapter. Before this, we introduce the notations that will be used (some of them have been introduced in Section 1.1). In what follows, $\langle \cdot, \cdot \rangle$ denotes the inner product of finite-dimensional Euclidean spaces. For $N \geq 1$, $e_j^N, j = 1, \ldots, N$ denotes the canonical basis of the space $\mathbb{R}^N$. If P is the orthogonal projector onto a subspace $L \subset \mathbb{R}^N$, then $P^{\perp}$ denotes the projector onto the orthogonal complement $L^{\perp}$. The Schatten p-norms of matrices and the l_p norms of vectors will be denoted, following the notations used in Section 1.1, by $\|\cdot\|_p$ for $1 \leq p \leq +\infty$. With a minor abuse of notation, $\|\cdot\|$ denotes both the l_2-norm of vectors in finite-dimensional spaces and the operator norm of matrices (i.e., their largest singular value). In other words, for any matrix $A \in \mathbb{R}^{m\times n}$, $\|A\|$ is equivalent to $\|A\|_\infty$.

In what follows, $A' \in \mathbb{R}^{n\times m}$ denotes the transpose of a matrix $A \in \mathbb{R}^{m\times n}$. The following mapping $\Lambda : \mathbb{R}^{m\times n} \mapsto \mathbb{R}^{(m+n)\times(m+n)}$ will be frequently used:

$$\Lambda(A) := \begin{pmatrix} 0 & A \\ A' & 0 \end{pmatrix}, A \in \mathbb{R}^{m\times n}.$$

Note that the image $\Lambda(A)$ is a symmetric $(m+n)\times(m+n)$ matrix.

Vectors $u \in \mathbb{R}^m, v \in \mathbb{R}^n$, etc. will be viewed as column vectors (or $m \times 1, n \times 1$, etc matrices). For $u \in \mathbb{R}^m, v \in \mathbb{R}^n$, denote by $u \otimes v$ the matrix $uv' \in \mathbb{R}^{m\times n}$. In other words, $u \otimes v$ can be viewed as a linear transformation from $\mathbb{R}^n$ into $\mathbb{R}^m$ defined as follows: $(u \otimes v)x = u\langle v, x\rangle, x \in \mathbb{R}^n$.

Let $A \in \mathbb{R}^{m\times n}$ be an $m \times n$ matrix and let

$$A = \sum_{i=1}^{m\wedge n} \sigma_i (u_i \otimes v_i)$$

be its singular value decomposition (SVD) with singular values $\sigma_1 \geq \ldots \geq \sigma_{m\wedge n} \geq 0$, orthonormal left singular vectors $u_1, \ldots, u_{m\wedge n} \in \mathbb{R}^m$ and orthonormal right singular vectors $v_1, \ldots, v_{m\wedge n} \in \mathbb{R}^n$. If A is of rank $\mathrm{rank}(A) = r \leq m \wedge n$, then $\sigma_i = 0, i > r$ and the SVD can be written as $A = \sum_{i=1}^r \sigma_i(u_i \otimes v_i)$. Note that in the case when there are repeated singular values σ_i, the singular vectors are not unique. In this case, let $\mu_1 > \ldots \mu_d > 0$ with $d \leq r$ be distinct singular values of A arranged in decreasing order and denote $\Delta_k := \{i : \sigma_i = \mu_k\}, k = 1, \ldots, d$. Let $\nu_k := \mathrm{card}(\Delta_k)$ be the multiplicity of $\mu_k, k = 1, \dots, d$. Denote

$$P_k^{uv} := \sum_{i \in \Delta_k} (u_i \otimes v_i),\ P_k^{vu} := \sum_{i \in \Delta_k} (v_i \otimes u_i),$$
$$P_k^{uu} := \sum_{i \in \Delta_k} (u_i \otimes u_i),\ P_k^{vv} := \sum_{i \in \Delta_k} (v_i \otimes v_i).$$

It is straightforward to check that the following relationships hold:

$$(P_k^{uu})' = P_k^{uu},\ (P_k^{uu})^2 = P_k^{uu},\ P_k^{vu} = (P_k^{uv})',\ P_k^{uv} P_k^{vu} = P_k^{uu}. \tag{4.1.1}$$

This implies, in particular, that the operators P_k^{uu}, P_k^{vv} are orthogonal projectors (in the spaces $\mathbb{R}^m, \mathbb{R}^n$, respectively). It is also easy to check that

$$P_k^{uu} P_{k'}^{uu} = 0,\ P_k^{vv} P_{k'}^{vv} = 0,\ P_k^{vu} P_{k'}^{uv} = 0,\ P_k^{uv} P_{k'}^{vu} = 0,\ k \neq k'. \tag{4.1.2}$$

The SVD of matrix A can be rewritten as $A = \sum_{k=1}^d \mu_k P_k^{uv}$ and it can be shown that the operators $P_k^{uv}, k = 1, \dots, d$ are uniquely defined. Let

$$B = \Lambda(A) = \begin{pmatrix} 0 & A \\ A' & 0 \end{pmatrix} = \sum_{k=1}^d \mu_k \begin{pmatrix} 0 & P_k^{uv} \\ P_k^{vu} & 0 \end{pmatrix}.$$

For $k = 1, \dots, d$, denote

$$P_k := \frac{1}{2} \begin{pmatrix} P_k^{uu} & P_k^{uv} \\ P_k^{vu} & P_k^{vv} \end{pmatrix},\ P_{-k} := \frac{1}{2} \begin{pmatrix} P_k^{uu} & -P_k^{uv} \\ -P_k^{vu} & P_k^{vv} \end{pmatrix},$$

and also

$$\mu_{-k} := -\mu_k.$$

Using relationships (4.1.1), (4.1.2), it is easy to show that $P_k P_{k'} = P_{k'} P_k = \mathbb{1}(k = k') P_k$ for all $k, k', 1 \leq |k| \leq d, 1 \leq |k'| \leq d$. Since the operators $P_k : \mathbb{R}^{m+n} \mapsto \mathbb{R}^{m+n}, 1 \leq |k| \leq d$ are also symmetric, they are orthogonal projectors onto mutually orthogonal subspaces of $\mathbb{R}^{m+n}$. Note that, by a simple algebra, $B = \sum_{1 \leq |k| \leq d} \mu_k P_k$, implying that μ_k are distinct eigenvalues of B and P_k are the corresponding eigenprojectors. Note also that if $2 \sum_{k=1}^d \nu_k < m + n$, then zero is also an eigenvalue of B (that will be denoted by μ_0) of multiplicity $\nu_0 := n + m - 2\sum_{k=1}^d \nu_k$. Representation $A \mapsto B = \Lambda(A) = \begin{pmatrix} 0 & A \\ A' & 0 \end{pmatrix}$ will play a crucial role in what follows since it allows to reduce the analysis of SVD for matrix A to the spectral representation $B = \sum_{1\leq|k|\leq d} \mu_k P_k$. In particular, the operators P_k^{uv} involved in the SVD $A = \sum_{k=1}^d \mu_k P_k^{uv}$ can be recovered from the eigenprojectors P_k of matrix B (hence, they are uniquely defined). Define also $\theta_i := \frac{1}{\sqrt{2}}\begin{pmatrix} u_i \\ v_i \end{pmatrix}$ and $\theta_{-i} := \frac{1}{\sqrt{2}}\begin{pmatrix} u_i \\ -v_i \end{pmatrix}$ for $i = 1, \dots, r$ and let $\Delta_{-k} := \{-i : i \in \Delta_k\}, k = 1, \dots, d$. Then, $\theta_i, 1 \leq |i| \leq r$ are orthonormal eigenvectors of B (not necessarily uniquely defined) corresponding to its non-zero eigenvalues $\sigma_1 \geq \cdots \geq \sigma_r > 0 > \sigma_{-r} \geq \cdots \geq \sigma_{-1}$ with $\sigma_{-i} = -\sigma_i$ and

$$P_k = \sum_{i\in\Delta_k} (\theta_i \otimes \theta_i), 1 \leq |k| \leq d.$$

It will be assumed in what follows that A is perturbed by a random matrix $X \in \mathbb{R}^{m\times n}$ with i.i.d. entries $X_{ij} \sim \mathcal{N}(0, \tau^2)$ for some $\tau > 0$. Given the SVD of the perturbed matrix

$$\tilde{A} = A + X = \sum_{j=1}^{m\wedge n} \tilde{\sigma}_i (\tilde{u}_i \otimes \tilde{v}_i),$$

our main interest lies in estimating singular vectors u_i and v_i of the matrix A in the case when its singular values σ_i are distinct, or, more generally, in estimating the operators $P_k^{uu}, P_k^{uv}, P_k^{vu}, P_k^{vv}$. To this end, we will use the estimators

$$\tilde{P}_k^{uu} := \sum_{i\in\Delta_k} (\tilde{u}_i \otimes \tilde{u}_i),\ \tilde{P}_k^{uv} := \sum_{i\in\Delta_k} (\tilde{u}_i \otimes \tilde{v}_i),$$
$$\tilde{P}_k^{vu} := \sum_{i\in\Delta_k} (\tilde{v}_i \otimes \tilde{u}_i),\ \tilde{P}_k^{vv} := \sum_{i\in\Delta_k} (\tilde{v}_i \otimes \tilde{v}_i),$$

and our main goal will be to study the fluctuations of the bilinear forms of these random operators around the bilinear forms of operators P_k^{uu}, P_k^{uv}, P_k^{vu}, P_k^{vv}. In the case when the singular values of A are distinct, this would allow us to study the fluctuations of linear forms of singular vectors $\tilde{u}_i, \tilde{v}_i$ around the corresponding linear forms of u_i, v_i which would provide a way to control the fluctuations of components of "empirical" singular vectors in a given basis around their true counterparts. Clearly, the problem can be and will be reduced to the analysis of spectral representation of a symmetric random matrix

$$\tilde{B} = \Lambda(\tilde{A}) = \begin{pmatrix} 0 & \tilde{A} \\ \tilde{A}' & 0 \end{pmatrix} = B + \Gamma, \quad \text{where } \Gamma = \Lambda(X) = \begin{pmatrix} 0 & X \\ X' & 0 \end{pmatrix}, \tag{4.1.3}$$

that can be viewed as a random perturbation of the symmetric matrix B. The spectral representation of this matrix can be written in the form

$$\tilde{B} = \sum_{1 \le |i| \le (m \wedge n)} \tilde{\sigma}_i (\tilde{\theta}_i \otimes \tilde{\theta}_i),$$

where

$$\tilde{\sigma}_{-i} = -\tilde{\sigma}_i, \ \tilde{\theta}_i := \frac{1}{\sqrt{2}} \begin{pmatrix} \tilde{u}_i \\ \tilde{v}_i \end{pmatrix}, \ \tilde{\theta}_{-i} := \frac{1}{\sqrt{2}} \begin{pmatrix} \tilde{u}_i \\ -\tilde{v}_i \end{pmatrix}, \ i = 1, \dots, (m \wedge n).$$

If the operator norm $\|\Gamma\|$ of the "noise" matrix Γ is small enough comparing with the "spectral gap" of the k-th eigenvalue μ_k of B (for some $k = 1, \dots, d$), then it is easy to see that $\tilde{P}_k := \sum_{i \in \Delta_k} (\tilde{\theta}_i \otimes \tilde{\theta}_i)$ is the orthogonal projector on the direct sum of eigenspaces of $\tilde{B}$ corresponding to the "cluster" $\{\tilde{\sigma}_i : i \in \Delta_k\}$ of its eigenvalues localized in a neighborhood of μ_k. Moreover, $\tilde{P}_k = \frac{1}{2} \begin{pmatrix} \tilde{P}_k^{uu} & \tilde{P}_k^{uv} \\ \tilde{P}_k^{vu} & \tilde{P}_k^{vv} \end{pmatrix}$. Thus, it is enough to study the fluctuations of bilinear forms of random orthogonal projectors $\tilde{P}_k$ around the corresponding bilinear form of the spectral projectors P_k to derive similar properties of operators $\tilde{P}_k^{uu}, \tilde{P}_k^{uv}, \tilde{P}_k^{vu}, \tilde{P}_k^{vv}$.

We will be interested in bounding the bilinear forms of operators $\tilde{P}_k - P_k$ for $k = 1, \dots, d$. To this end, we will provide separate bounds on the random error $\tilde{P}_k - \mathbb{E}\tilde{P}_k$ and on the bias $\mathbb{E}\tilde{P}_k - P_k$. For $k = 1, \dots, d$, $\bar{g}_k$ denotes the distance from the eigenvalue μ_k to the rest of the spectrum of A (the eigengap of μ_k). More specifically, for $2 \le k \le d-1$, $\bar{g}_k = \min(\mu_k - \mu_{k+1}, \mu_{k-1} - \mu_k)$, $\bar{g}_1 = \mu_1 - \mu_2$ and $\bar{g}_d = \min(\mu_{d-1} - \mu_d, \mu_d)$.

The main assumption in the results that follow is that $\mathbb{E}\|X\| < \frac{\bar{g}_k}{2}$ (more precisely, $\mathbb{E}\|X\| \le (1-\gamma)\frac{\bar{g}_k}{2}$ for a positive γ). In view of the concentration inequality of

Lemma 23 in the next section, this essentially means that the operator norm of the random perturbation matrix $\|\Gamma\| = \|X\|$ is strictly smaller than one half of the spectral gap $\bar{g}_k$ of singular value μ_k. Since, again by Lemma 23, $\mathbb{E}\|X\| \asymp \tau\sqrt{m \vee n}$, this assumption also means that $\bar{g}_k \gtrsim \tau\sqrt{m \vee n}$ (so, the spectral gap $\bar{g}_k$ is sufficiently large). Our goal is to prove that, under this assumption, the values of bilinear form $\langle \tilde{P}_k x, y\rangle$ of random spectral projector $\tilde{P}_k$ have tight concentration around their means (with the magnitude of deviations of the order $\sqrt{\frac{1}{m\vee n}}$). We will also show that the bias $\mathbb{E}\tilde{P}_k - P_k$ of the spectral projector $\tilde{P}_k$ is "aligned" with the spectral projector P_k (up to an error of the order $\sqrt{\frac{1}{m\vee n}}$ in the operator norm).

4.2 Preliminary lemmas

The proofs follow the approach of [56] who did a similar analysis in the problem of estimation of spectral projectors of sample covariance. We start with discussing several preliminary facts used in what follows. Lemma 23 and Lemma 24 below provide moment bounds and a concentration inequality for $\|\Gamma\| = \|X\|$. The bound on $\mathbb{E}\|X\|$ of Lemma 23 is available in many references (see, e.g., [94]). The concentration bound for $\|X\|$ is a straightforward consequence of the Gaussian concentration inequality. The moment bounds of Lemma 24 can be easily proved by integrating out the tails of the exponential bound that follows from the concentration inequality of Lemma 23.

Lemma 23. *There exist absolute constants $c_0, c_1, c_2 > 0$ such that*

$$c_0\tau\sqrt{m \vee n} \le \mathbb{E}\|X\| \le c_1\tau\sqrt{m \vee n}$$

and for all $t > 0$,

$$\mathbb{P}\big\{\big|\|X\| - \mathbb{E}\|X\|\big| \ge c_2\tau\sqrt{t}\big\} \le e^{-t}.$$

Lemma 24. *For all $p \ge 1$, it holds that*

$$\mathbb{E}^{1/p}\|X\|^p \asymp \tau\sqrt{m \vee n}$$

According to a well-known result that goes back to Weyl, for symmetric (or Hermitian) $N \times N$ matrices C, D

$$\max_{1 \le j \le N} \left| \lambda_j^{\downarrow}(C) - \lambda_j^{\downarrow}(D) \right| \le \|C - D\|,$$

where $\lambda^{\downarrow}(C), \lambda^{\downarrow}(D)$ denote the vectors consisting of the eigenvalues of matrices C, D, respectively, arranged in a non-increasing order. This immediately implies that, for all $k = 1, \ldots, d$,

$$\max_{j \in \Delta_k} |\tilde{\sigma}_j - \mu_k| \le \|\Gamma\|$$

and

$$\min_{j \in \cup_{k' \neq k} \Delta_{k'}} |\tilde{\sigma}_j - \mu_k| \ge \bar{g}_k - \|\Gamma\|.$$

Assuming that $\|\Gamma\| < \frac{\bar{g}_k}{2}$, we get that $\{\tilde{\sigma}_j : j \in \Delta_k\} \subset (\mu_k - \bar{g}_k/2, \mu_k + \bar{g}_k/2)$ and the rest of the eigenvalues of $\tilde{B}$ are outside of this interval. Moreover, if $\|\Gamma\| < \frac{\bar{g}_k}{4}$, then the cluster of eigenvalues $\{\tilde{\sigma}_j : j \in \Delta_k\}$ is localized inside a shorter interval $(\mu_k - \bar{g}_k/4, \mu_k + \bar{g}_k/4)$ of radius $\bar{g}_k/4$ and its distance from the rest of the spectrum of $\tilde{B}$ is $> \frac{3}{4}\bar{g}_k$. These simple considerations allow us to view the projection operator $\tilde{P}_k = \sum_{j \in \Delta_k} (\tilde{\theta}_j \otimes \tilde{\theta}_j)$ as a projector on the direct sum of eigenspaces of $\tilde{B}$ corresponding to its eigenvalues located in a "small" neighborhood of the eigenvalue μ_k of B, which makes $\tilde{P}_k$ a natural estimator of P_k.

Define operators C_k as follows:

$$C_k = \sum_{s \neq k} \frac{1}{\mu_s - \mu_k} P_s.$$

In the case when $2 \sum_{k=1}^{d} \nu_k < m + n$ and, hence, $\mu_0 = 0$ is also an eigenvalue of B, it will be assumed that the above sum includes $s = 0$ with P_0 being the corresponding spectral projector.

The next simple lemma can be found, for instance, in [56]. Its proof is based on a standard perturbation analysis utilizing Riesz formula for spectral projectors.

Lemma 25. *The following bound holds:*

$$\|\tilde{P}_k - P_k\| \leq 4\frac{\|\Gamma\|}{\bar{g}_k}.$$

Moreover,

$$\tilde{P}_k - P_k = L_k(\Gamma) + S_k(\Gamma),$$

where $L_k(\Gamma) := C_k \Gamma P_k + P_k \Gamma C_k$ and

$$\|S_k(\Gamma)\| \leq 14\left(\frac{\|\Gamma\|}{\bar{g}_k}\right)^2.$$

4.3 Main results and proofs

Theorem 21. *Suppose that for some $\gamma \in (0,1)$, $\mathbb{E}\|X\| \leq (1-\gamma)\frac{\bar{g}_k}{2}$. There exists a constant $C_\gamma > 0$ such that, for all $x, y \in \mathbb{R}^{m+n}$ and for all $t \geq 1$, the following inequality holds with probability at least $1 - e^{-t}$:*

$$\left|\left\langle(\tilde{P}_k - \mathbb{E}\tilde{P}_k)x, y\right\rangle\right| \leq C_\gamma \frac{\tau\sqrt{t}}{\bar{g}_k}\Big(\frac{\tau\sqrt{m \vee n} + \tau\sqrt{t}}{\bar{g}_k} + 1\Big)\|x\|\|y\|. \tag{4.3.1}$$

Assuming that $t \lesssim m \vee n$ and taking into account that $\tau\sqrt{m \vee n} \asymp \mathbb{E}\|X\| \leq \bar{g}_k$, we easily get from the bound of Theorem 21 that

$$\left|\left\langle(\tilde{P}_k - \mathbb{E}\tilde{P}_k)x, y\right\rangle\right| \lesssim_\gamma \frac{\tau\sqrt{t}}{\bar{g}_k}\|x\|\|y\| \lesssim_\gamma \sqrt{\frac{t}{m \vee n}}\|x\|\|y\|,$$

so, the fluctuations of $\langle \tilde{P}_k x, y\rangle$ around its expectation are indeed of the order $\sqrt{\frac{1}{m \vee n}}$.

Proof of Theorem 21. Since $\mathbb{E}L_k(\Gamma) = 0$, it is easy to check that

$$\tilde{P}_k - \mathbb{E}\tilde{P}_k = L_k(\Gamma) + S_k(\Gamma) - \mathbb{E}S_k(\Gamma) =: L_k(\Gamma) + R_k(\Gamma). \tag{4.3.2}$$

We will first provide a bound on the bilinear form of the remainder $\big\langle R_k(\Gamma)x, y\big\rangle$. Note that

$$\langle R_k(\Gamma)x, y\rangle = \langle S_k(\Gamma)x, y\rangle - \langle \mathbb{E}S_k(\Gamma)x, y\rangle$$

is a function of the random matrix $X \in \mathbb{R}^{m\times n}$ since $\Gamma = \Lambda(X)$ (see (4.1.3)). When we need to emphasize this dependence, we will write Γ_X instead of Γ. With some abuse of notation, we will view X as a point in $\mathbb{R}^{m\times n}$ rather than a random variable.

Let $0 < \gamma < 1$ and define a function $h_{x,y,\delta}(\cdot) : \mathbb{R}^{m\times n} \to \mathbb{R}$ as follows:

$$h_{x,y,\delta}(X) := \langle S_k(\Gamma_X)x, y\rangle\, \phi\biggl(\frac{\|\Gamma_X\|}{\delta}\biggr),$$

where ϕ is a Lipschitz function with constant $\frac{1}{\gamma}$ on $\mathbb{R}_+$ and $0 \le \phi(s) \le 1$. More precisely, assume that $\phi(s) = 1, s \le 1$, $\phi(s) = 0, s \ge (1+\gamma)$ and ϕ is linear in between. We will prove that the function $X \mapsto h_{x,y,\delta}(X)$ satisfy the Lipschitz condition. Note that

$$|\langle (S_k(\Gamma_{X_1}) - S_k(\Gamma_{X_2}))\, x, y\rangle| \le \|S_k(\Gamma_{X_1}) - S_k(\Gamma_{X_2})\|\|x\|\|y\|.$$

To control the norm $\|S_k(\Gamma_{X_1}) - S_k(\Gamma_{X_2})\|$, we need to apply Lemma 4 from [56]. It is stated below without the proof.

Lemma 26. *Let $\gamma \in (0,1)$ and suppose that $\delta \le \frac{1-\gamma}{1+\gamma}\frac{\bar{g}_k}{2}$. There exists a constant $C_\gamma > 0$ such that, for all symmetric $\Gamma_1, \Gamma_2 \in \mathbb{R}^{(m+n)\times(m+n)}$ satisfying the conditions $\|\Gamma_1\| \le (1+\gamma)\delta$ and $\|\Gamma_2\| \le (1+\gamma)\delta$,*

$$\|S_k(\Gamma_1) - S_k(\Gamma_2)\| \le C_\gamma \frac{\delta}{\bar{g}_k^2}\|\Gamma_1 - \Gamma_2\|.$$

We now derive the Lipschitz condition for the function $X \mapsto h_{x,y,\delta}(X)$.

Lemma 27. *Under the assumption that $\delta \le \frac{1-\gamma}{1+\gamma}\frac{\bar{g}_k}{2}$, there exists a constant $C_\gamma > 0$,*

$$|h_{x,y,\delta}(X_1) - h_{x,y,\delta}(X_2)| \le C_\gamma \frac{\delta\|X_1 - X_2\|_2}{\bar{g}_k^2}\|x\|\|y\|. \tag{4.3.3}$$

Proof. Suppose first that $\max(\|\Gamma_{X_1}\|, \|\Gamma_{X_2}\|) \le (1+\gamma)\delta$. Using Lemma 26 and Lipschitz properties of function ϕ, we get

$$\begin{aligned}
|h_{x,y,\delta}(X_1) - h_{x,y,\delta}(X_2)| &= \left| \langle S_k(\Gamma_{X_1})x, y\rangle \phi\Big(\frac{\|\Gamma_{X_1}\|}{\delta}\Big) - \langle S_k(\Gamma_{X_2})x, y\rangle \phi\Big(\frac{\|\Gamma_{X_2}\|}{\delta}\Big) \right| \\
&\le \|S_k(\Gamma_{X_1}) - S_k(\Gamma_{X_2})\| \|x\| \|y\| \phi\Big(\frac{\|\Gamma_{X_1}\|}{\delta}\Big) \\
&+ \|S_k(\Gamma_{X_2})\| \left| \phi\Big(\frac{\|\Gamma_{X_1}\|}{\delta}\Big) - \phi\Big(\frac{\|\Gamma_{X_2}\|}{\delta}\Big) \right| \|x\| \|y\| \\
&\le C_\gamma \frac{\delta \|\Gamma_{X_1} - \Gamma_{X_2}\|}{\bar{g}_k^2} \|x\| \|y\| + \frac{14(1+\gamma)^2\delta^2}{\bar{g}_k^2} \frac{\|\Gamma_{X_1} - \Gamma_{X_2}\|}{\gamma\delta} \|x\| \|y\| \\
&\lesssim_\gamma \frac{\delta \|\Gamma_{X_1} - \Gamma_{X_2}\|}{\bar{g}_k^2} \|x\| \|y\| \lesssim_\gamma \frac{\delta \|X_1 - X_2\|_2}{\bar{g}_k^2} \|x\| \|y\|.
\end{aligned}$$

In the case when $\min(\|\Gamma_{X_1}\|, \|\Gamma_{X_2}\|) \ge (1+\gamma)\delta$, we have $h_{x,y,\delta}(X_1) = h_{x,y,\delta}(X_2) = 0$, and (4.3.3) trivially holds. Finally, in the case when $\|\Gamma_{X_1}\| \le (1+\gamma)\delta \le \|\Gamma_{X_2}\|$, we have

$$\begin{aligned}
|h_{x,y,\delta}(X_1) - h_{x,y,\delta}(X_2)| &= \left| \langle S_k(\Gamma_{X_1})x, y\rangle \phi\Big(\frac{\|\Gamma_{X_1}\|}{\delta}\Big) \right| \\
&= \left| \langle S_k(\Gamma_{X_1})x, y\rangle \phi\Big(\frac{\|\Gamma_{X_1}\|}{\delta}\Big) - \langle S_k(\Gamma_{X_1})x, y\rangle \phi\Big(\frac{\|\Gamma_{X_2}\|}{\delta}\Big) \right| \\
&\le \|S_k(\Gamma_{X_1})\| \left| \phi\Big(\frac{\|\Gamma_{X_1}\|}{\delta}\Big) - \phi\Big(\frac{\|\Gamma_{X_2}\|}{\delta}\Big) \right| \|x\| \|y\| \\
&\le 14 \left(\frac{(1+\gamma)\delta}{\bar{g}_k}\right)^2 \frac{\|\Gamma_{X_1} - \Gamma_{X_2}\|}{\gamma\delta} \|x\| \|y\| \\
&\lesssim_\gamma \frac{\delta \|X_1 - X_2\|_2}{\bar{g}_k^2} \|x\| \|y\|.
\end{aligned}$$

The case $\|\Gamma_{X_2}\| \le (1+\gamma)\delta \le \|\Gamma_{X_1}\|$ is similar. □

Our next step is to apply the following concentration bound that easily follows from the Gaussian isoperimetric inequality.

Lemma 28. *Let $f : \mathbb{R}^{m\times n} \mapsto \mathbb{R}$ be a function satisfying the following Lipschitz condition with some constant $L > 0$:*

$$|f(A_1) - f(A_2)| \le L\|A_1 - A_2\|_2, A_1, A_2 \in \mathbb{R}^{m\times n}$$

Suppose X is a random $m \times n$ matrix with i.i.d. entries $X_{ij} \sim \mathcal{N}(0, \tau^2)$. Let M be a real number such that

$$\mathbb{P}\big\{f(X) \geq M\big\} \geq \frac{1}{4} \text{ and } \mathbb{P}\big\{f(X) \leq M\big\} \geq \frac{1}{4}.$$

Then there exists some constant $D_1 > 0$ such that for all $t \geq 1$,

$$\mathbb{P}\Big\{\big|f(X) - M\big| \geq D_1 L\tau\sqrt{t}\Big\} \leq e^{-t}.$$

The next lemma is the main ingredient in the proof of Theorem 21. It provides a Bernstein type bound on the bilinear form $\langle R_k(\Gamma)x, y\rangle$ of the remainder R_k in the representation (4.3.2).

Lemma 29. *Suppose that, for some $\gamma \in (0,1)$, $\mathbb{E}\|\Gamma\| \leq (1-\gamma)\frac{\bar{g}_k}{2}$. Then, there exists a constant $C_\gamma > 0$ such that for all $x, y \in \mathbb{R}^{m+n}$ and all $t \geq \log(4)$, the following inequality holds with probability at least $1 - e^{-t}$*

$$|\langle R_k(\Gamma)x, y\rangle| \leq C_\gamma \frac{\tau\sqrt{t}}{\bar{g}_k}\left(\frac{\tau\sqrt{m \vee n} + \tau\sqrt{t}}{\bar{g}_k}\right)\|x\|\|y\|.$$

Proof. Define $\delta_{n,m}(t) := \mathbb{E}\|\Gamma\| + c_2\tau\sqrt{t}$. By the second bound of Lemma 23, with a proper choice of constant $c_2 > 0$, $\mathbb{P}\{\|\Gamma\| \geq \delta_{n,m}(t)\} \leq e^{-t}$. We first consider the case when $c_2\tau\sqrt{t} \leq \frac{\gamma}{2}\frac{\bar{g}_k}{2}$, which implies that

$$\delta_{n,m}(t) \leq (1-\gamma/2)\frac{\bar{g}_k}{2} = \frac{1-\gamma'}{1+\gamma'}\frac{\bar{g}_k}{2}$$

for some $\gamma' \in (0,1)$ depending only on γ. Therefore, it enables us to use Lemma 27 with $\delta := \delta_{n,m}(t)$. Recall that $h_{x,y,\delta}(X) = \big\langle S_k(\Gamma)x, y\big\rangle\phi\left(\frac{\|\Gamma\|}{\delta}\right)$ and let $M := \mathrm{Med}\big(\big\langle S_k(\Gamma)x, y\big\rangle\big)$. Observe that, for $t \geq \log(4)$,

$$\begin{aligned}&\mathbb{P}\{h_{x,y,\delta}(X) \geq M\} \geq \mathbb{P}\{h_{x,y,\delta}(X) \geq M, \|\Gamma\| \leq \delta_{n,m}(t)\}\\ \geq&\mathbb{P}\{\big\langle S_k(\Gamma)x, y\big\rangle \geq M\big\} - \mathbb{P}\{\|\Gamma\| > \delta_{n,m}(t)\} \geq \frac{1}{2} - e^{-t} \geq \frac{1}{4}\end{aligned}$$

and, similarly. $\mathbb{P}(h_{x,y,\delta}(X) \leq M) \geq \frac{1}{4}$. Therefore, by applying lemmas 27,28, we conclude that with probability at least $1 - e^{-t}$,

$$\big|h_{x,y,\delta}(X) - M\big| \lesssim_\gamma \frac{\delta_{n,m}(t)\tau\sqrt{t}}{\bar{g}_k^2}\|x\|\|y\|$$

Since, by the first bound of Lemma 23, $\delta_{n,m}(t) \lesssim \tau(\sqrt{m \vee n} + \sqrt{t})$, we get that with the same probability

$$\big|h_{x,y,\delta}(X) - M\big| \lesssim_{\gamma} \frac{\tau\sqrt{t}}{\bar{g}_k}\frac{\tau\sqrt{m \vee n} + \tau\sqrt{t}}{\bar{g}_k}\|x\|\|y\|.$$

Moreover, on the event $\{\|\Gamma\| \le \delta_{n,m}(t)\}$ that holds with probability at least $1 - e^{-t}$, $h_{x,y,\delta}(X) = \big\langle S_k(\Gamma)x, y\big\rangle$. Therefore, the following inequality holds with probability at least $1 - 2e^{-t}$:

$$\big|\big\langle S_k(\Gamma)x, y\big\rangle - M\big| \lesssim_{\gamma} \frac{\tau\sqrt{t}}{\bar{g}_k}\frac{\tau\sqrt{m \vee n} + \tau\sqrt{t}}{\bar{g}_k}\|x\|\|y\|. \tag{4.3.4}$$

We still need to prove a similar inequality in the case $c_2\tau\sqrt{t} \ge \frac{\gamma}{2}\frac{\bar{g}_k}{2}$. In this case,

$$\mathbb{E}\|\Gamma\| \le (1-\gamma)\frac{\bar{g}_k}{2} \le \frac{2c_2(1-\gamma)}{\gamma}\tau\sqrt{t},$$

implying that $\delta_{n,m}(t) \lesssim_{\gamma} \tau\sqrt{t}$. It follows from Lemma 25 that

$$\big|\big\langle S_k(\Gamma)x, y\big\rangle\big| \le \|S_k(\Gamma)\|\|x\|\|y\| \lesssim \frac{\|\Gamma\|^2}{\bar{g}_k^2}\|x\|\|y\|$$

This implies that with probability at least $1 - e^{-t}$,

$$\big|\big\langle S_k(\Gamma)x, y\big\rangle\big| \lesssim \frac{\delta_{n,m}^2(t)}{\bar{g}_k^2}\|x\|\|y\| \lesssim_{\gamma} \frac{\tau^2 t}{\bar{g}_k^2}\|x\|\|y\|.$$

Since $t \ge \log(4)$ and $e^{-t} \le 1/4$, we can bound the median M of $\big\langle S_k(\Gamma)x, y\big\rangle$ as follows:

$$M \lesssim_{\gamma} \frac{\tau^2 t}{\bar{g}_k^2}\|x\|\|y\|,$$

which immediately implies that bound (4.3.4) holds under assumption $c_2\tau\sqrt{t} \ge \frac{\gamma}{2}\frac{\bar{g}_k}{2}$ as well. By integrating out the tails of exponential bound (4.3.4), we obtain that

$$\big|\mathbb{E}\big\langle S_k(\Gamma)x, y\big\rangle - M\big| \le \mathbb{E}\big|\big\langle S_k(\Gamma)x, y\big\rangle - M\big| \lesssim_{\gamma} \frac{\tau^2\sqrt{m \vee n}}{\bar{g}_k^2}\|x\|\|y\|,$$

which allows us to replace the median by the mean in concentration inequality (4.3.4). To complete the proof, it remains to rewrite the probability bound $1 - 2e^{-t}$ as $1 - e^{-t}$ by adjusting the value of the constant C_{γ}. □

Recalling that $\tilde{P}_k - \mathbb{E}\tilde{P}_k = L_k(\Gamma) + R_k(\Gamma)$, it remains to study the concentration of $\big\langle L_k(\Gamma)x, y\big\rangle$.

Lemma 30. *For all* $x, y \in \mathbb{R}^{m+n}$ *and* $t > 0$,

$$\mathbb{P}\left(\big|\big\langle L_k(\Gamma)x, y\big\rangle\big| \geq 4\frac{\tau\|x\|\|y\|\sqrt{t}}{\bar{g}_k}\right) \leq e^{-t}.$$

Proof. Recall that $L_k(\Gamma) = P_k\Gamma C_k + C_k\Gamma P_k$ implying that

$$\langle L_k(\Gamma)x, y\rangle = \langle \Gamma P_k x, C_k y\rangle + \langle \Gamma C_k x, P_k y\rangle.$$

If $x = \begin{pmatrix} x_1 \\ \\ x_2 \end{pmatrix}, y = \begin{pmatrix} y_1 \\ \\ y_2 \end{pmatrix}$, where $x_1, y_1 \in \mathbb{R}^m, x_2, y_2 \in \mathbb{R}^n$, then it is easy to check that

$$\langle \Gamma x, y\rangle = \langle X x_2, y_1\rangle + \langle X y_2, x_1\rangle.$$

Clearly, the random variable $\langle \Gamma x, y\rangle$ is normal with mean zero and variance

$$\mathbb{E}\langle \Gamma x, y\rangle^2 \leq 2\Big[\mathbb{E}\langle X x_2, y_1\rangle^2 + \mathbb{E}\langle X y_2, x_1\rangle^2\Big].$$

Since X is an $m \times n$ matrix with i.i.d. $\mathcal{N}(0, \tau^2)$ entries, we easily get that

$$\mathbb{E}\langle X x_2, y_1\rangle^2 = \mathbb{E}\langle X, y_1 \otimes x_2\rangle^2 = \tau^2\|y_1 \otimes x_2\|_2^2 = \tau^2\|x_2\|^2\|y_1\|^2$$

and, similarly,

$$\mathbb{E}\langle X y_2, x_1\rangle^2 = \tau^2\|x_1\|^2\|y_2\|^2.$$

Therefore,

$$\begin{aligned}\mathbb{E}\langle \Gamma x, y\rangle^2 \leq & 2\tau^2\Big[\|x_2\|^2\|y_1\|^2 + \|x_1\|^2\|y_2\|^2\Big] \\ & \leq 2\tau^2\Big[(\|x_1\|^2 + \|x_2\|^2)(\|y_1\|^2 + \|y_2\|^2)\Big] = 2\tau^2\|x\|^2\|y\|^2.\end{aligned}$$

As a consequence, the random variable $\langle L_k(\Gamma)x, y\rangle$ is also normal with mean zero and its variance is bounded from above as follows:

$$\begin{aligned}\mathbb{E}\langle L_k(\Gamma)x, y\rangle^2 \leq & 2\Big[\mathbb{E}\langle \Gamma P_k x, C_k y\rangle^2 + \mathbb{E}\langle \Gamma C_k x, P_k y\rangle^2\Big] \\ & \leq 4\tau^2\Big[\|P_k x\|^2\|C_k y\|^2 + \|C_k x\|^2\|P_k y\|^2\Big].\end{aligned}$$

Since $\|P_k\| \leq 1$ and $\|C_k\| \leq \frac{1}{\bar{g}_k}$, we get that

$$\mathbb{E}\langle L_k(\Gamma)x, y\rangle^2 \leq \frac{8\tau^2}{\bar{g}_k^2}\|x\|^2\|y\|^2.$$

The bound of the lemma easily follows from standard tail bounds for normal random variables. □

The upper bound on $|\langle(\tilde{P}_k - \mathbb{E}\tilde{P}_k)x, y\rangle|$ claimed in Theorem 21 follows by combining Lemma 29 and Lemma 30. □

The next result shows that the bias $\mathbb{E}\tilde{P}_k - P_k$ of $\tilde{P}_k$ can be represented as a sum of a "low rank part" $P_k(\mathbb{E}\tilde{P}_k - P_k)P_k$ and a small remainder.

Theorem 22. *The following bound holds with some constant $C > 0$:*

$$\left\|\mathbb{E}\tilde{P}_k - P_k\right\| \leq C\frac{\tau^2(m \vee n)}{\bar{g}_k^2}. \tag{4.3.5}$$

Moreover, suppose that for some $\gamma \in (0,1)$, $\mathbb{E}\|X\| \leq (1-\gamma)\frac{\bar{g}_k}{2}$. Then, there exists a constant $C_\gamma > 0$ such that

$$\big\|\mathbb{E}\tilde{P}_k - P_k - P_k(\mathbb{E}\tilde{P}_k - P_k)P_k\big\| \leq C_\gamma\frac{\nu_k\tau^2\sqrt{m \vee n}}{\bar{g}_k^2}. \tag{4.3.6}$$

Since, under the assumption $\mathbb{E}\|X\| \leq (1-\gamma)\frac{\bar{g}_k}{2}$, we have $\bar{g}_k \gtrsim \tau\sqrt{m \vee n}$, bound (4.3.6) implies that the following representation holds

$$\mathbb{E}\tilde{P}_k - P_k = P_k(\mathbb{E}\tilde{P}_k - P_k)P_k + T_k$$

with the remainder T_k satisfying the bound

$$\|T_k\| \lesssim_\gamma \frac{\tau^2\sqrt{m \vee n}}{\bar{g}_k^2} \lesssim_\gamma \frac{\nu_k}{\sqrt{m \vee n}}.$$

Proof of Theorem 22. Note that, since $\tilde{P}_k - P_k = L_k(\Gamma) + S_k(\Gamma)$ and $\mathbb{E}L_k(\Gamma) = 0$, we have

$$\mathbb{E}\tilde{P}_k - P_k = \mathbb{E}S_k(\Gamma).$$

It follows from the bound on $\|S_k(\Gamma)\|$ of Lemma 25 that

$$\left\|\mathbb{E}\tilde{P}_k - P_k\right\| \leq \mathbb{E}\|S_k(\Gamma)\| \leq 14\frac{\mathbb{E}\|\Gamma\|^2}{\bar{g}_k^2} \tag{4.3.7}$$

and the bound of Lemma 24 implies that

$$\left\|\mathbb{E}\tilde{P}_k - P_k\right\| \lesssim \frac{\tau^2(m \vee n)}{\bar{g}_k^2},$$

which proves (4.3.5).

Let

$$\delta_{n,m} := \mathbb{E}\|\Gamma\| + c_2\tau\sqrt{\log(m+n)}.$$

It follows from Lemma 23 that, with a proper choice of constant $c_2 > 0$,

$$\mathbb{P}\left(\|\Gamma\| \geq \delta_{n,m}\right) \leq \frac{1}{m+n}.$$

In the case when $c_2\tau\sqrt{\log(m+n)} > \frac{\gamma}{2}\frac{\bar{g}_k}{2}$, the proof of bound (4.3.6) is trivial. Indeed, in this case

$$\left\|\mathbb{E}\tilde{P}_k - P_k\right\| \ \leq \mathbb{E}\|\tilde{P}_k\| + \|P_k\| \leq 2 \lesssim_\gamma \frac{\tau^2\log(m+n)}{\bar{g}_k^2} \lesssim \frac{\nu_k\tau^2\sqrt{m \vee n}}{\bar{g}_k^2}.$$

Since $\left\|P_k(\mathbb{E}\tilde{P}_k - P_k)P_k\right\| \leq \left\|\mathbb{E}\tilde{P}_k - P_k\right\|$, bound (4.3.6) of the theorem follows when $c_2\tau\sqrt{\log(m+n)} > \frac{\gamma}{2}\frac{\bar{g}_k}{2}$.

In the rest of the proof, it will be assumed that $c_2\tau\sqrt{\log(m+n)} \leq \frac{\gamma}{2}\frac{\bar{g}_k}{2}$ which, together with the condition $\mathbb{E}\|\Gamma\| = \mathbb{E}\|X\| \leq (1-\gamma)\frac{\bar{g}_k}{2}$, implies that $\delta_{n,m} \leq (1-\gamma/2)\frac{\bar{g}_k}{2}$. On the other hand, $\delta_{n,m} \lesssim \tau\sqrt{m \vee n}$. The following decomposition of the bias $\mathbb{E}\tilde{P}_k - P_k$ is obvious:

$$\begin{aligned}\mathbb{E}\tilde{P}_k - P_k &= \mathbb{E}S_k(\Gamma) = \mathbb{E}P_kS_k(\Gamma)P_k \\ &\quad+\mathbb{E}\left(P_k^\perp S_k(\Gamma)P_k + P_kS_k(\Gamma)P_k^\perp + P_k^\perp S_k(\Gamma)P_k^\perp\right)\mathbb{1}(\|\Gamma\| \leq \delta_{n,m}) \\ &\quad+\mathbb{E}\left(P_k^\perp S_k(\Gamma)P_k + P_kS_k(\Gamma)P_k^\perp + P_k^\perp S_k(\Gamma)P_k^\perp\right)\mathbb{1}(\|\Gamma\| > \delta_{n,m})\end{aligned} \tag{4.3.8}$$

We start with bounding the part of the expectation in the right hand side of (4.3.8) that corresponds to the event $\{\|\Gamma\| \leq \delta_{n,m}\}$ on which we also have $\|\Gamma\| < \frac{\bar{g}_k}{2}$. Under this assumption, the eigenvalues μ_k of B and $\sigma_j(\tilde{B}), j \in \Delta_k$ of $\tilde{B}$ are inside the circle γ_k in $\mathbb{C}$ with center μ_k and radius $\frac{\bar{g}_k}{2}$. The rest of the eigenvalues of $B, \tilde{B}$ are outside of γ_k. According to the Riesz formula for spectral projectors,

$$\tilde{P}_k = -\frac{1}{2\pi i} \oint_{\gamma_k} R_{\tilde{B}}(\eta) d\eta,$$

where $R_T(\eta) = (T - \eta I)^{-1}, \eta \in \mathbb{C} \setminus \sigma(T)$ denotes the resolvent of operator T ($\sigma(T)$ being its spectrum). It is also assumed that the contour γ_k has a counterclockwise orientation. Note that the resolvents will be viewed as operators from $\mathbb{C}^{m+n}$ into itself. The following power series expansion is standard:

$$\begin{aligned} R_{\tilde{B}}(\eta) =& R_{B+\Gamma}(\eta) = (B + \Gamma - \eta I)^{-1} \\ =& [(B - \eta I)(I + (B - \eta I)^{-1}\Gamma)]^{-1} \\ =& (I + R_B(\eta)\Gamma)^{-1} R_B(\eta) = \sum_{r \geq 0} (-1)^r [R_B(\eta)\Gamma]^r R_B(\eta), \end{aligned}$$

where the series in the last line converges because $\|R_B(\eta)\Gamma\| \leq \|R_B(\eta)\| \|\Gamma\| < \frac{2}{\bar{g}_k} \frac{\bar{g}_k}{2} = 1$. The inequality $\|R_B(\eta)\| \leq \frac{2}{\bar{g}_k}$ holds for all $\eta \in \gamma_k$. One can easily verify that

$$\begin{aligned} P_k =& -\frac{1}{2\pi i} \oint_{\gamma_k} R_B(\eta) d\eta, \\ L_k(\Gamma) =& \frac{1}{2\pi i} \oint_{\gamma_k} R_B(\eta)\Gamma R_B(\eta) d\eta, \\ S_k(\Gamma) =& -\frac{1}{2\pi i} \oint_{\gamma_k} \sum_{r \geq 2} (-1)^r [R_B(\eta)\Gamma]^r R_B(\eta) d\eta. \end{aligned}$$

The following spectral representation of the resolvent will be used

$$R_B(\eta) = \sum_s \frac{1}{\mu_s - \eta} P_s,$$

where the sum in the right hand side includes $s = 0$ in the case when $\mu_0 = 0$ is an eigenvalue of B (equivalently, in the case when $2\sum_{k=1}^d \nu_k < m + n$). Define

$$\tilde{R}_B(\eta) := R_B(\eta) - \frac{1}{\mu_k - \eta} P_k = \sum_{s \neq k} \frac{1}{\mu_s - \eta} P_s.$$

Then, for $r \geq 2$,

$$
\begin{aligned}
P_k^{\perp}[R_B(\eta)\Gamma]^r R_B(\eta)P_k &= \frac{1}{\mu_k - \eta}P_k^{\perp}[R_B(\eta)\Gamma]^r P_k \\
=&\frac{1}{(\mu_k-\eta)^2}\sum_{s=2}^{r}(\tilde{R}_B(\eta)\Gamma)^{s-1}P_k\Gamma(R_B(\eta)\Gamma)^{r-s}P_k + \frac{1}{\mu_k-\eta}(\tilde{R}_B(\eta)\Gamma)^r P_k.
\end{aligned}
$$

The above representation easily follows from the following simple observation: let $a := \frac{P_k}{\mu_k-\eta}\Gamma$ and $b := \tilde{R}_B(\eta)\Gamma$. Then

$$
\begin{aligned}
(a+b)^r =& a(a+b)^{r-1} + b(a+b)^{r-1} \\
=& a(a+b)^{r-1} + ba(a+b)^{r-2} + b^2(a+b)^{r-2} \\
=& a(a+b)^{r-1} + ba(a+b)^{r-2} + b^2a(a+b)^{r-3} + b^3(a+b)^{r-3} \\
=& \ldots = \sum_{s=1}^{r} b^{s-1}a(a+b)^{r-s} + b^r.
\end{aligned}
$$

As a result,

$$
\begin{aligned}
P_k^{\perp}S_k(\Gamma)P_k = -\sum_{r\geq 2}(-1)^r\frac{1}{2\pi i}\oint_{\gamma_k}\Bigg[&\frac{1}{(\mu_k-\eta)^2}\sum_{s=2}^{r}(\tilde{R}_B(\eta)\Gamma)^{s-1}P_k\Gamma(R_B(\eta)\Gamma)^{r-s}P_k \\
+&\frac{1}{\mu_k-\eta}(\tilde{R}_B(\eta)\Gamma)^r P_k\Bigg]d\eta
\end{aligned}
\tag{4.3.9}
$$

Let $P_k = \sum_{l\in\Delta_k}\theta_l \otimes \theta_l$, where $\{\theta_l, l \in \Delta_k\}$ are orthonormal eigenvectors corresponding to the eigenvalue μ_k. Therefore, for any $y \in \mathbb{R}^{m+n}$,

$$
\begin{aligned}
(\tilde{R}_B(\eta)\Gamma)^{s-1}P_k\Gamma(R_B(\eta)\Gamma)^{r-s}P_k y &= \sum_{l\in\Delta_k}(\tilde{R}_B(\eta)\Gamma)^{s-1}\theta_l\otimes\theta_l\Gamma(R_B(\eta)\Gamma)^{r-s}P_k y \\
=\sum_{l\in\Delta_k}&\left\langle\Gamma(R_B(\eta)\Gamma)^{r-s}P_k y, \theta_l\right\rangle(\tilde{R}_B(\eta)\Gamma)^{s-2}\tilde{R}_B(\eta)\Gamma\theta_l
\end{aligned}
\tag{4.3.10}
$$

Since $|\left\langle\Gamma(R_B(\eta)\Gamma)^{r-s}P_k y, \theta_l\right\rangle| \leq \|\Gamma\|^{r-s+1}\|R_B(\eta)\|^{r-s}\|y\|$, we get

$$
\mathbb{E}|\left\langle\Gamma(R_B(\eta)\Gamma)^{r-s}P_k y, \theta_l\right\rangle|^2\mathbb{1}(\|\Gamma\|\leq\delta_{n,m}) \leq \delta_{n,m}^{2(r-s+1)}\left(\frac{2}{\bar{g}_k}\right)^{2(r-s)}\|y\|^2.
$$

Also, for any $x \in \mathbb{R}^{m+n}$, we have to bound

$$\mathbb{E}\left|\left\langle (\tilde{R}_B(\eta)\Gamma)^{s-2}\tilde{R}_B(\eta)\Gamma\theta_l, x\right\rangle\right|^2 \mathbb{1}(\|\Gamma\| \le \delta_{n,m}). \tag{4.3.11}$$

In what follows, we need some additional notations. Let $X_1^c, \dots, X_n^c \sim \mathcal{N}(0, \tau^2 I_m)$ be the i.i.d. columns of X and $(X_1^r)', \dots, (X_n^r)' \sim \mathcal{N}(0, \tau^2 I_n)$ be its i.i.d. rows (here I_m and I_n are $m \times m$ and $n \times n$ identity matrices). For $j = 1, \dots, n$, define the vector $\check{X}_j^c = ((X_j^c)', 0)' \in \mathbb{R}^{m+n}$, representing the $(m+j)$-th column of matrix Γ. Similarly, for $i = 1, \dots, m$, $\check{X}_i^r = (0, (X_i^r)')' \in \mathbb{R}^{m+n}$ represents the i-th row of Γ. With these notations, the following representations of Γ holds

$$\Gamma = \sum_{j=1}^n e_{m+j}^{m+n} \otimes \check{X}_j^c + \sum_{j=1}^n \check{X}_j^c \otimes e_{m+j}^{m+n},$$

$$\Gamma = \sum_{i=1}^m \check{X}_i^r \otimes e_i^{m+n} + \sum_{i=1}^m e_i^{m+n} \otimes \check{X}_i^r,$$

and, moreover,

$$\sum_{j=1}^n e_{m+j}^{m+n} \otimes \check{X}_j^c = \sum_{i=1}^m \check{X}_i^r \otimes e_i^{m+n}, \quad \sum_{j=1}^n \check{X}_j^c \otimes e_{m+j}^{m+n} = \sum_{i=1}^m e_i^{m+n} \otimes \check{X}_i^r.$$

Therefore,

$$\begin{aligned}\left\langle (\tilde{R}_B(\eta)\Gamma)^{s-2}\tilde{R}_B(\eta)\Gamma\theta_l, x\right\rangle &= \sum_{j=1}^n \left\langle \check{X}_j^c, \theta_l\right\rangle \left\langle (\tilde{R}_B(\eta)\Gamma)^{s-2}\tilde{R}_B(\eta)e_{m+j}^{m+n}, x\right\rangle \\ &+ \sum_{j=1}^n \left\langle e_{m+j}^{m+n}, \theta_l\right\rangle \left\langle (\tilde{R}_B(\eta)\Gamma)^{s-2}\tilde{R}_B(\eta)\check{X}_j^c, x\right\rangle =: I_1(x) + I_2(x),\end{aligned}$$

and we get

$$\begin{aligned}&\mathbb{E}\left|\left\langle (\tilde{R}_B(\eta)\Gamma)^{s-2}\tilde{R}_B(\eta)\Gamma\theta_l, x\right\rangle\right|^2 \mathbb{1}(\|\Gamma\| \le \delta_{n,m}) \\ &\qquad \le 2\mathbb{E}(|I_1(x)|^2 + |I_2(x)|^2)\mathbb{1}(\|\Gamma\| \le \delta_{n,m}).\end{aligned} \tag{4.3.12}$$

Observe that the random variable $(\tilde{R}_B(\eta)\Gamma)^{s-2}\tilde{R}_B(\eta)$ is a function of $\{P_t\check{X}_j^c, t \neq k, j = 1, \dots, n\}$. Indeed, since $\tilde{R}_B(\eta)$ is a linear combination of operators $P_t, t \neq k$, it is easy to see that $(\tilde{R}_B(\eta)\Gamma)^{s-2}\tilde{R}_B(\eta)$ can be represented as a linear combination of operators

$$(P_{t_1}\Gamma P_{t_2})(P_{t_2}\Gamma P_{t_3}) \dots (P_{t_{s-2}}\Gamma P_{t_{s-1}})$$

with $t_j \neq k$ and with non-random complex coefficients. On the other hand,

$$P_{t_k}\Gamma P_{t_{k+1}} = \sum_{j=1}^{n} P_{t_k} e_{m+j}^{m+n} \otimes P_{t_{k+1}} \check{X}_j^c + \sum_{j=1}^{n} P_{t_k} \check{X}_j^c \otimes P_{t_{k+1}} e_{m+j}^{m+n}.$$

These two facts imply that $(\tilde{R}_B(\eta)\Gamma)^{s-2}\tilde{R}_B(\eta)$ is a function of $\{P_t\check{X}_j^c, t \neq k, j = 1, \ldots, n\}$. Similarly, it is also a function of $\{P_t\check{X}_i^r, t \neq k, i = 1, \ldots, m\}$.

It is easy to see that random variables $\{P_k\check{X}_j^c, j = 1, \ldots, n\}$ and $\{P_t\check{X}_j^c, j = 1, \ldots, n, t \neq k\}$ are independent. Since they are mean zero normal random variables and $\check{X}_j^c, j = 1, \ldots, n$ are independent, it is enough to check that, for all $j = 1, \ldots, n$, $t \neq k$, $P_k\check{X}_j^c$ and $P_t\check{X}_j^c$ are uncorrelated. To this end, observe that

$$\begin{aligned}
\mathbb{E}(P_k\check{X}_j^c \otimes P_t\check{X}_j^c) =& P_k\mathbb{E}(\check{X}_j^c \otimes \check{X}_j^c)P_t \\
=& \frac{1}{4}\begin{pmatrix} P_k^{uu} & P_k^{uv} \\ P_k^{vu} & P_k^{vv} \end{pmatrix}\begin{pmatrix} I_m & 0 \\ 0 & 0 \end{pmatrix}\begin{pmatrix} P_t^{uu} & P_t^{uv} \\ P_t^{vu} & P_t^{vv} \end{pmatrix} \\
=& \frac{1}{4}\begin{pmatrix} P_k^{uu}P_t^{uu} & P_k^{uu}P_t^{uv} \\ P_k^{vu}P_t^{uu} & P_k^{vu}P_t^{uv} \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix},
\end{aligned}$$

where we used orthogonality relationships (4.1.2). Quite similarly, one can prove independence of $\{P_k\check{X}_i^r, i = 1, \ldots, m\}$ and $\{P_t\check{X}_i^r, i = 1, \ldots, m, t \neq k\}$.

We will now provide an upper bound on $\mathbb{E}|I_1(x)|^2\mathbb{1}(\|\Gamma\| \leq \delta_{n,m})$. To this end, define

$$\begin{aligned}
\omega_j(x) =& \left\langle (\tilde{R}_B(\eta)\Gamma)^{s-2}\tilde{R}_B(\eta)e_{m+j}^{m+n}, x \right\rangle, \quad j = 1, \ldots, n \\
=& \omega_j^{(1)}(x) + i\omega_j^{(2)}(x) \in \mathbb{C}.
\end{aligned}$$

Let $I_1(x) = \kappa^{(1)}(x) + i\kappa^{(2)}(x) \in \mathbb{C}$. Then, conditionally on $\{P_t\check{X}_j^c : t \neq k, j = 1, \ldots, n\}$, the random vector $(\kappa^{(1)}(x), \kappa^{(2)}(x))$ has the same distribution as mean zero Gaussian random vector in $\mathbb{R}^2$ with covariance,

$$\left(\sum_{j=1}^{n} \frac{\tau^2}{2}\omega_j^{k_1}(x)\omega_j^{k_2}(x)\right), k_1, k_2 = 1, 2$$

(to check the last claim, it is enough to compute conditional covariance of $(\kappa^{(1)}(x), \kappa^{(2)}(x))$ given $\{P_t\check{X}_j^c : t \neq k, j = 1, \dots, n\}$ using the fact that $(\tilde{R}_B(\eta)\Gamma)^{s-2}\tilde{R}_B(\eta)$ is a function of $\{P_t\check{X}_j^c, t \neq k, j = 1, \dots, n\}$). Therefore,

$$\begin{aligned}
&\mathbb{E}\left(|I_1(x)|^2 \Big| P_t\check{X}_j^c : t \neq k, j = 1, \dots, n\right) \\
=&\mathbb{E}\left((\kappa^{(1)}(x))^2 + (\kappa^{(2)}(x))^2 \Big| P_t\check{X}_j^c : t \neq k, j = 1, \dots, n\right) \\
=&\frac{\tau^2}{2}\sum_{j=1}^n \Big((\omega_j^{(1)}(x))^2 + (\omega_j^{(2)}(x))^2\Big) = \frac{\tau^2}{2}\sum_{j=1}^n |\omega_j(x)|^2.
\end{aligned}$$

Furthermore,

$$\begin{aligned}
\sum_{j=1}^n \tau^2|\omega_j(x)|^2 &= \tau^2\sum_{j=1}^n |\omega_j(x)|^2 \\
&=\tau^2 \sum_{j=1}^n \left|\left\langle \tilde{R}_B(\eta)(\Gamma\tilde{R}_B(\eta))^{s-2}x, e_{m+j}^{m+n}\right\rangle\right|^2 \\
&=\tau^2 \left\langle \tilde{R}_B(\eta)(\Gamma\tilde{R}_B(\eta))^{s-2}x, \tilde{R}_B(\eta)(\Gamma\tilde{R}_B(\eta)\Gamma)^{s-2}x\right\rangle \\
&\leq \tau^2\|\tilde{R}_B(\eta)\|^{2(s-1)}\|\Gamma\|^{2(s-2)}\|x\|^2.
\end{aligned}$$

Under the assumption $\delta_{n,m} < \frac{\bar{g}_k}{2}$, the following inclusion holds:

$$\{\|\Gamma\| \leq \delta_{n,m}\} \subset \left\{\sum_{j=1}^n \tau^2|\omega_j(x)|^2 \leq \tau^2\left(\frac{2}{\bar{g}_k}\right)^{2(s-1)}\delta_{n,m}^{2(s-2)}\|x\|^2\right\} =: G$$

Therefore,

$$\begin{aligned}
&\mathbb{E}|I_1(x)|^2\mathbb{1}(\|\Gamma\| \leq \delta_{n,m}) \leq \mathbb{E}|I_1(x)|^2\mathbb{1}_G = \mathbb{E}\mathbb{E}\left(|I_1(x)|^2\Big| P_t\check{X}_j^c, t \neq k, j = 1, \dots, n\right)\mathbb{1}_G \\
=&\mathbb{E}\mathbb{E}\left(\sum_{j=1}^n \tau^2|\omega_j(x)|^2\Big| P_t\check{X}_j^c, t \neq k, j = 1, \dots, n\right)\mathbb{1}_G \leq \tau^2\left(\frac{2}{\bar{g}_k}\right)^{2(s-1)}\delta_{n,m}^{2(s-2)}\|x\|^2.
\end{aligned} \tag{4.3.13}$$

A similar bound holds also for $\mathbb{E}|I_2(x)|^2\mathbb{1}(\|\Gamma\| \leq \delta_{n,m})$:

$$\mathbb{E}|I_2(x)|^2\mathbb{1}(\|\Gamma\| \leq \delta_{n,m}) \leq \tau^2\left(\frac{2}{\bar{g}_k}\right)^{2(s-1)}\delta_{n,m}^{2(s-2)}\|x\|^2. \tag{4.3.14}$$

For the proof, it is enough to observe that

$$\begin{aligned}
I_2(x) &= \sum_{j=1}^{n} \left\langle e^{m+n}_{m+j}, \theta_l \right\rangle \left\langle (\tilde{R}_B(\eta)\Gamma)^{s-2}\tilde{R}_B(\eta)\check{X}^c_j, x \right\rangle \\
&= \left\langle (\tilde{R}_B(\eta)\Gamma)^{s-2}\tilde{R}_B(\eta)\Big(\textstyle\sum_{j=1}^{n} \check{X}^c_j \otimes e^{m+n}_{m+j}\Big)\theta_l, x \right\rangle \\
&= \left\langle (\tilde{R}_B(\eta)\Gamma)^{s-2}\tilde{R}_B(\eta)\Big(\textstyle\sum_{i=1}^{m} e^{m+n}_{i} \otimes \check{X}^r_i\Big)\theta_l, x \right\rangle \\
&= \sum_{i=1}^{m} \left\langle \check{X}^r_i, \theta_l \right\rangle \left\langle (\tilde{R}_B(\eta)\Gamma)^{s-2}\tilde{R}_B(\eta)e^{m+n}_i, x \right\rangle
\end{aligned}$$

and to repeat the previous conditioning argument (this time, given $\{P_t\check{X}^r_i : t \neq k, i = 1, \dots, m\}$).

Combining bounds (4.3.13), (4.3.14) and (4.3.12), we get

$$\mathbb{E}\left|\left\langle (\tilde{R}_B(\eta)\Gamma)^{s-2}\tilde{R}_B(\eta)\Gamma\theta_l, x \right\rangle\right|^2 \mathbb{1}(\|\Gamma\| \leq \delta_{n,m}) \leq 2\tau^2 \left(\frac{2}{\bar{g}_k}\right)^{2(s-1)} \delta^{2(s-2)}_{n,m}\|x\|^2.$$

Then, it follows that

$$\begin{aligned}
&\left|\mathbb{E}\left\langle \Gamma(R_B(\eta)\Gamma)^{r-s}P_k y, \theta_l \right\rangle \left\langle (\tilde{R}_B(\eta)\Gamma)^{s-2}\tilde{R}_B(\eta)\Gamma\theta_l, x \right\rangle \mathbb{1}(\|\Gamma\| \leq \delta_{n,m})\right| \\
&\leq \left(\mathbb{E}\left|\left\langle \Gamma(R_B(\eta)\Gamma)^{r-s}P_k y, \theta_l \right\rangle\right|^2 \mathbb{1}(\|\Gamma\| \leq \delta_{n,m})\right)^{1/2} \\
&\times \left(\mathbb{E}\left|\left\langle (\tilde{R}_B(\eta)\Gamma)^{s-2}\tilde{R}_B(\eta)\Gamma\theta_l, x \right\rangle\right|^2 \mathbb{1}(\|\Gamma\| \leq \delta_{n,m})\right)^{1/2} \\
&\leq \sqrt{2}\tau \left(\frac{2\delta_{n,m}}{\bar{g}_k}\right)^{r-1} \|x\|\|y\|,
\end{aligned}$$

which, taking into account (4.3.10), implies that

$$\begin{aligned}
&\left|\mathbb{E}\left\langle (\tilde{R}_B(\eta)\Gamma)^{s-1}P_k\Gamma(R_B(\eta)\Gamma)^{r-s}P_k y, x \right\rangle \mathbb{1}(\|\Gamma\| \leq \delta_{n,m})\right| \\
&\qquad\qquad\qquad\qquad \leq \sqrt{2}\nu_k\tau \left(\frac{2\delta_{n,m}}{\bar{g}_k}\right)^{r-1} \|x\|\|y\|
\end{aligned}$$

Since $(\tilde{R}_B(\eta)\Gamma)^r P_k = (\tilde{R}_B(\eta)\Gamma)^{r-1}\tilde{R}_B(\eta)\Gamma P_k$, it can be proved by a similar argument that

$$\left|\mathbb{E}\left\langle (\tilde{R}_B(\eta)\Gamma)^r P_k y, x \right\rangle \mathbb{1}(\|\Gamma\| \leq \delta_{n,m})\right| \leq \sqrt{2}\nu_k\tau\frac{2}{\bar{g}_k} \left(\frac{2\delta_{n,m}}{\bar{g}_k}\right)^{r-1} \|x\|\|y\|.$$

Therefore, substituting the above bounds in (4.3.9) and taking into account that $|\mu_k - \eta| = \frac{\bar{g}_k}{2}, \eta \in \gamma_k$ and that the length of the contour of integration γ_k is equal to $2\pi \frac{\bar{g}_k}{2}$, we get

$$\left| \mathbb{E} \left\langle P_k^{\perp} S_k(\Gamma) P_k y, x \right\rangle \mathbb{1}(\|\Gamma\| \leq \delta_{n,m}) \right| \leq \sum_{r \geq 2} \frac{r \bar{g}_k}{2} \left(\frac{2}{\bar{g}_k} \right)^2 \sqrt{2} \nu_k \tau \left(\frac{2\delta_{n,m}}{\bar{g}_k} \right)^{r-1} \|x\| \|y\|$$

$$= \frac{2}{\bar{g}_k} \sqrt{2} \nu_k \tau \sum_{r \geq 2} r \left(\frac{2\delta_{n,m}}{\bar{g}_k} \right)^{r-1} \|x\| \|y\| \lesssim_{\gamma} \nu_k \tau \frac{\delta_{n,m}}{\bar{g}_k^2} \|x\| \|y\|,$$

where we also used the condition $\delta_{n,m} \leq (1 - \gamma/2) \frac{\bar{g}_k}{2}$ implying that $\frac{2\delta_{n,m}}{\bar{g}_k} \leq 1 - \gamma/2$. Clearly, this implies that

$$\left\| \mathbb{E} P_k^{\perp} S_k(\Gamma) P_k \right\| \mathbb{1}(\|\Gamma\| \leq \delta_{n,m}) \lesssim_{\gamma} \nu_k \tau \frac{\delta_{n,m}}{\bar{g}_k^2} \lesssim_{\gamma} \frac{\nu_k \tau \sqrt{m \vee n}}{\bar{g}_k^2}.$$

Furthermore, the same bound, obviously, holds for

$$\left\| \mathbb{E} \big\langle P_k S_k(\Gamma) P_k^{\perp} y, x \big\rangle \mathbb{1}(\|\Gamma\| \leq \delta_{n,m}) \right\| = \left\| \mathbb{E} \big\langle P_k^{\perp} S_k(\Gamma) P_k x, y \big\rangle \mathbb{1}(\|\Gamma\| \leq \delta_{n,m}) \right\|$$

and, by similar arguments, it can be demonstrated that it also holds for

$$\left\| \mathbb{E} P_k^{\perp} S_k(\Gamma) P_k^{\perp} \right\| \mathbb{1}(\|\Gamma\| \leq \delta_{n,m})$$

(the only different term in this case is $(\tilde{R}_B(\eta)\Gamma)^r \tilde{R}_B(\eta)$, but, since $\{\mu_t, t \neq k\}$ are outside of the circle γ_k, it simply leads to $\oint_{\gamma_k} (\tilde{R}_B(\eta)\Gamma)^r \tilde{R}_B(\eta) d\eta = 0$).

It remains to observe that

$$\begin{aligned} &\left\| \mathbb{E} \left(P_k^{\perp} S_k(\Gamma) P_k + P_k S_k(\Gamma) P_k^{\perp} + P_k^{\perp} S_k(\Gamma) P_k^{\perp} \right) \mathbb{1}(\|\Gamma\| > \delta_{n,m}) \right\| \\ \leq &\mathbb{E} \left\| P_k^{\perp} S_k(\Gamma) P_k + P_k S_k(\Gamma) P_k^{\perp} + P_k^{\perp} S_k(\Gamma) P_k^{\perp} \right\| \mathbb{1}(\|\Gamma\| > \delta_{n,m}) \\ \leq &\mathbb{E} \|S_k(\Gamma)\| \mathbb{1}(\|\Gamma\| > \delta_{n,m}) \\ \leq &(\mathbb{E} \|S_k(\Gamma)\|^2)^{1/2} \mathbb{P}^{1/2}(\|\Gamma\| > \delta_{n,m}) \\ \lesssim &\mathbb{E}^{1/2} \left(\frac{\|\Gamma\|}{\bar{g}_k} \right)^4 \mathbb{P}^{1/2}(\|\Gamma\| > \delta_{n,m}) \lesssim \frac{1}{\sqrt{m \vee n}} \frac{\tau^2 (m \vee n)}{\bar{g}_k^2} \lesssim \frac{\tau^2 \sqrt{m \vee n}}{\bar{g}_k^2} \end{aligned}$$

and to substitute the above bounds to identity (4.3.8) to get that

$$\left\| \mathbb{E} \tilde{P}_k - P_k - P_k \mathbb{E} S_k(\Gamma) P_k \right\| \lesssim_{\gamma} \frac{\nu_k \tau^2 \sqrt{m \vee n}}{\bar{g}_k^2},$$

which implies the claim of the theorem. □

We will now consider a special case when μ_k has multiplicity 1 ($\nu_k = 1$). In this case, $\Delta_k = \{i_k\}$ for some $i_k \in \{1, \dots, (m \wedge n)\}$ and $P_k = \theta_{i_k} \otimes \theta_{i_k}$. Let $\tilde{P}_k := \tilde{\theta}_{i_k} \otimes \tilde{\theta}_{i_k}$. Note that on the event $\|\Gamma\| = \|X\| < \frac{\bar{g}_k}{2}$ that is assumed to hold with a high probability, the multiplicity of $\tilde{\sigma}_{i_k}$ is also 1 (see the discussion in the next section after Lemma 24). Note also that the unit eigenvectors $\theta_{i_k}, \tilde{\theta}_{i_k}$ are defined only up to their signs. Due to this, we will assume without loss of generality that $\langle \tilde{\theta}_{i_k}, \theta_{i_k} \rangle \geq 0$.

Since $P_k = \theta_{i_k} \otimes \theta_{i_k}$ is an operator of rank 1, we have

$$P_k(\mathbb{E}\tilde{P}_k - P_k)P_k = b_k P_k,$$

where

$$b_k := \left\langle (\mathbb{E}\tilde{P}_k - P_k)\theta_{i_k}, \theta_{i_k} \right\rangle = \mathbb{E}\langle \tilde{\theta}_{i_k}, \theta_{i_k} \rangle^2 - 1.$$

Therefore,

$$\mathbb{E}\tilde{P}_k = (1 + b_k)P_k + T_k$$

and b_k turns out to be the main parameter characterizing the bias of $\tilde{P}_k$. Clearly, $b_k \in [-1, 0]$ (note that $b_k = 0$ is equivalent to $\tilde{\theta}_{i_k} = \theta_{i_k}$ a.s. and $b_k = -1$ is equivalent to $\tilde{\theta}_{i_k} \perp \theta_{i_k}$ a.s.). On the other hand, by bound (4.3.5) of Theorem 22,

$$|b_k| \leq \left\| \mathbb{E}\tilde{P}_k - P_k \right\| \lesssim \frac{\tau^2(m \vee n)}{\bar{g}_k^2}. \tag{4.3.15}$$

In the next theorem, it will be assumed that the bias is not too large in the sense that b_k is bounded away by a constant $\gamma > 0$ from -1.

Theorem 23. *Suppose that, for some $\gamma \in (0, 1)$, $\mathbb{E}\|X\| \leq (1 - \gamma)\frac{\bar{g}_k}{2}$ and $1 + b_k \geq \gamma$. Then, for all $x \in \mathbb{R}^{m+n}$ and for all $t \geq 1$ with probability at least $1 - e^{-t}$,*

$$\left|\langle \tilde{\theta}_{i_k} - \sqrt{1 + b_k}\theta_{i_k}, x \rangle\right| \lesssim_\gamma \frac{\tau\sqrt{t}}{\bar{g}_k}\Big(\frac{\tau\sqrt{m \vee n} + \tau\sqrt{t}}{\bar{g}_k} + 1\Big)\|x\|.$$

Assuming that $t \lesssim m \vee n$, the bound of Theorem 23 implies that

$$\left|\langle \tilde{\theta}_{i_k} - \sqrt{1 + b_k}\theta_{i_k}, x \rangle\right| \lesssim_\gamma \frac{\tau\sqrt{t}}{\bar{g}_k}\|x\| \lesssim_\gamma \sqrt{\frac{t}{m \vee n}}\|x\|.$$

Therefore, the fluctuations of $\langle \tilde{\theta}_{i_k}, x \rangle$ around $\sqrt{1 + b_k}\langle \theta_{i_k}, x \rangle$ are of the order $\sqrt{\frac{1}{m \vee n}}$.

Proof of Theorem 23. By a simple computation (see Lemma 8 and the derivation of (6.6) in [56]), the following identity holds

$$\begin{aligned}\left\langle \tilde{\theta}_{i_k} - \sqrt{1+b_k}\theta_{i_k}, x\right\rangle &= \frac{\rho_k(x)}{\sqrt{1+b_k+\rho_k(x)}} \\ &- \frac{\sqrt{1+b_k}}{\sqrt{1+b_k+\rho_k(x)}\big(\sqrt{1+b_k+\rho_k(x)}+\sqrt{1+b_k}\big)}\rho_k(\theta_{i_k})\left\langle \theta_{i_k}, x\right\rangle,\end{aligned} \tag{4.3.16}$$

where $\rho_k(x) := \big\langle (\tilde{P}_k - (1+b_k)P_k)\theta_{i_k}, x\big\rangle, x \in \mathbb{R}^{m+n}$. In what follows, assume that $\|x\| = 1$. By the bounds of theorems 21 and 22, with probability at least $1 - e^{-t}$:

$$|\rho_k(x)| \le D_\gamma \frac{\tau\sqrt{t}}{\bar{g}_k}\Big(\frac{\tau\sqrt{m\vee n}+\tau\sqrt{t}}{\bar{g}_k}+1\Big).$$

The assumption $\mathbb{E}\|X\| \le (1-\gamma)\frac{\bar{g}_k}{2}$ implies that $\tau\sqrt{m\vee n} \lesssim \bar{g}_k$. Therefore, if t satisfies the assumption $\frac{\tau\sqrt{t}}{\bar{g}_k} \le c_\gamma$ for a sufficiently small constant $c_\gamma > 0$, then we have $|\rho_k(x)| \le \gamma/2$. By the assumption that $1+b_k \ge \gamma$, this implies that $1+b_k+\rho_k(x) \ge \gamma/2$. Thus, it easily follows from identity (4.3.16) that with probability at least $1-2e^{-t}$

$$\left|\left\langle \tilde{\theta}_{i_k} - \sqrt{1+b_k}\theta_{i_k}, x\right\rangle\right| \lesssim_\gamma \frac{\tau\sqrt{t}}{\bar{g}_k}\Big(\frac{\tau\sqrt{m\vee n}+\tau\sqrt{t}}{\bar{g}_k}+1\Big).$$

It remains to show that the same bound holds when $\frac{\tau\sqrt{t}}{\bar{g}_k} > c_\gamma$. In this case, we simply have that

$$\left|\left\langle \tilde{\theta}_{i_k} - \sqrt{1+b_k}\theta_{i_k}, x\right\rangle\right| \le \|\tilde{\theta}_{i_k}\| + (1+b_k)\|\theta_{i_k}\| \le 2 \lesssim_\gamma \frac{\tau^2 t}{\bar{g}_k^2},$$

which implies the bound of the theorem. □

Recall that $\theta_{i_k} := \frac{1}{\sqrt{2}}\begin{pmatrix} u_{i_k} \\ v_{i_k}\end{pmatrix}$, where u_{i_k}, v_{i_k} are left and right singular vectors of A corresponding to its singular value μ_k. Theorem 23 easily implies the following corollary.

Corollary 1. *Under the conditions of Theorem 23, with probability at least* $1-\frac{1}{m+n}$,

$$\max\Big\{\big\|\tilde{u}_{i_k} - \sqrt{1+b_k}u_{i_k}\big\|_\infty, \big\|\tilde{v}_{i_k} - \sqrt{1+b_k}v_{i_k}\big\|_\infty\Big\} \lesssim \sqrt{\frac{\log(m+n)}{m\vee n}}.$$

For the proof, it is enough to take $t = 2\log(m+n)$, $x = e_i^{m+n}, i = 1, \dots, (m+n)$ and to use the bound of Theorem 23 along with the union bound. Then recalling that $\theta_{i_k} = \frac{1}{\sqrt{2}}(u'_{i_k}, v'_{i_k})'$, Theorem 23 easily implies the claim.

Theorem 23 shows that the "naive estimator" $\langle \tilde{\theta}_{i_k}, x\rangle$ of linear form $\langle \theta_{i_k}, x\rangle$ could be improved by reducing its bias that, in principle, could be done by its simple rescaling $\langle \tilde{\theta}_{i_k}, x\rangle \mapsto \langle (1+b_k)^{-1/2}\tilde{\theta}_{i_k}, x\rangle$. Of course, the difficulty with this approach is related to the fact that the bias parameter b_k is unknown. We will outline below a simple approach based on repeated observations of matrix A. More specifically, let $\tilde{A}^1 = A + X^1$ and $\tilde{A}^2 = A + X^2$ be two independent copies of $\tilde{A}$ and denote $\tilde{B}^1 = \Lambda(\tilde{A}^1), \tilde{B}^2 = \Lambda(\tilde{A}^2)$. Let $\tilde{\theta}^1_{i_k}$ and $\tilde{\theta}^2_{i_k}$ be the eigenvectors of $\tilde{B}^1$ and $\tilde{B}^2$ corresponding to their eigenvalues $\tilde{\sigma}^1_{i_k}, \tilde{\sigma}^2_{i_k}$. The signs of $\tilde{\theta}^1_{i_k}$ and $\tilde{\theta}^2_{i_k}$ are chosen so that $\left\langle \tilde{\theta}^1_{i_k}, \tilde{\theta}^2_{i_k}\right\rangle \geq 0$. Let

$$\tilde{b}_k := \left\langle \tilde{\theta}^1_{i_k}, \tilde{\theta}^2_{i_k}\right\rangle - 1. \tag{4.3.17}$$

Given $\gamma > 0$, define

$$\hat{\theta}^{(\gamma)}_{i_k} := \frac{\tilde{\theta}^1_{i_k}}{\sqrt{1 + \tilde{b}_k \vee \frac{\sqrt{\gamma}}{2}}}.$$

Corollary 2. *Under the assumptions of Theorem 23, there exists a constant $C_\gamma > 0$ such that for all $x \in \mathbb{R}^{m+n}$ and all $t \geq 1$ with probability at least $1 - e^{-t}$,*

$$|\hat{b}_k - b_k| \leq C_\gamma \frac{\tau\sqrt{t}}{\bar{g}_k}\Big[\frac{\tau\sqrt{m \vee n} + \tau\sqrt{t}}{\bar{g}_k} + 1\Big] \tag{4.3.18}$$

and

$$|\langle \hat{\theta}^{(\gamma)}_{i_k} - \theta_{i_k}, x\rangle| \leq C_\gamma \frac{\tau\sqrt{t}}{\bar{g}_k}\Big[\frac{\tau\sqrt{m \vee n} + \tau\sqrt{t}}{\bar{g}_k} + 1\Big]\|x\|. \tag{4.3.19}$$

Note that $\hat{\theta}^{(\gamma)}_{i_k}$ is not necessarily a unit vector. However, its linear form provides a better approximation of the linear forms of θ_{i_k} than in the case of vector $\tilde{\theta}^1_{i_k}$ that is properly normalized. Clearly, the result implies similar bounds for the singular vectors $\hat{u}^{(\gamma)}_{i_k}$ and $\hat{v}^{(\gamma)}_{i_k}$.

Proof of Corollary 2. By a simple algebra,

$$\begin{aligned}|\tilde{b}_k - b_k| =& \Bigl|\bigl\langle \tilde{\theta}^1_{i_k}, \tilde{\theta}^2_{i_k}\bigr\rangle - (1+b_k)\Bigr| \le \Bigl|\sqrt{1+b_k}\bigl\langle \tilde{\theta}^1_{i_k} - \sqrt{1+b_k}\theta_{i_k}, \theta_{i_k}\bigr\rangle\Bigr| \\ &+\Bigl|\sqrt{1+b_k}\bigl\langle \tilde{\theta}^2_{i_k} - \sqrt{1+b_k}\theta_{i_k}, \theta_{i_k}\bigr\rangle\Bigr| + \Bigl|\bigl\langle \tilde{\theta}^1_{i_k} - \sqrt{1+b_k}\theta_{i_k}, \tilde{\theta}^2_{i_k} - \sqrt{1+b_k}\theta_{i_k}\bigr\rangle\Bigr|.\end{aligned}$$

Corollary 2 implies that with probability at least $1-e^{-t}$

$$\Bigl|\sqrt{1+b_k}\bigl\langle \tilde{\theta}^1_{i_k} - \sqrt{1+b_k}\theta_{i_k}, \theta_{i_k}\bigr\rangle\Bigr| \lesssim_\gamma \frac{\tau\sqrt{t}}{\bar{g}_k}\Bigl[\frac{\tau\sqrt{m\vee n}+\tau\sqrt{t}}{\bar{g}_k}+1\Bigr],$$

where we also used the fact that $1+b_k \in [0,1]$. A similar bound holds with the same probability for

$$\Bigl|\sqrt{1+b_k}\bigl\langle \tilde{\theta}^2_{i_k} - \sqrt{1+b_k}\theta_{i_k}, \theta_{i_k}\bigr\rangle\Bigr|.$$

To control the remaining term

$$\Bigl|\bigl\langle \tilde{\theta}^1_{i_k} - \sqrt{1+b_k}\theta_{i_k}, \tilde{\theta}^2_{i_k} - \sqrt{1+b_k}\theta_{i_k}\bigr\rangle\Bigr|,$$

note that $\tilde{\theta}^1_{i_k}$ and $\tilde{\theta}^2_{i_k}$ are independent. Thus, applying the bound of Theorem 23 conditionally on $\tilde{\theta}^2_{i_k}$, we get that with probability at least $1-e^{-t}$

$$\Bigl|\bigl\langle \tilde{\theta}^1_{i_k} - \sqrt{1+b_k}\theta_{i_k}, \tilde{\theta}^2_{i_k} - \sqrt{1+b_k}\theta_{i_k}\bigr\rangle\Bigr| \lesssim_\gamma \frac{\tau\sqrt{t}}{\bar{g}_k}\Bigl[\frac{\tau\sqrt{m\vee n}+\tau\sqrt{t}}{\bar{g}_k}+1\Bigr]\|\tilde{\theta}^2_{i_k} - \sqrt{1+b_k}\theta_{i_k}\|.$$

It remains to observe that

$$\|\tilde{\theta}^2_{i_k} - \sqrt{1+b_k}\theta_{i_k}\| \le 2$$

to complete the proof of bound (4.3.18).

Assume that $\|x\| \le 1$. Recall that under the assumptions of the corollary, $\tau\sqrt{m\vee n} \lesssim_\gamma \bar{g}_k$ and, if $\frac{\tau\sqrt{t}}{\bar{g}_k} \le c_\gamma$ for a sufficiently small constant c_γ, then bound (4.3.18) implies that $|\tilde{b}_k - b_k| \le \gamma/4$ (on the event of probability at least $1-e^{-t}$). Since $1+b_k \ge \gamma/2$, on the same event we also have $1+\tilde{b}_k \ge \gamma/4$ implying that $\hat{\theta}^{(\gamma)}_{i_k} = \frac{\tilde{\theta}^1_{i_k}}{\sqrt{1+\tilde{b}_k}}$. Therefore,

$$\begin{aligned}\Bigl|\bigl\langle \hat{\theta}^{(\gamma)}_{i_k} - \theta_{i_k}, x\bigr\rangle\Bigr| &= \tfrac{1}{\sqrt{1+\tilde{b}_k}}\Bigl|\bigl\langle \tilde{\theta}^1_{i_k} - \sqrt{1+\tilde{b}_k}\theta_{i_k}, x\bigr\rangle\Bigr| \\ &\lesssim_\gamma \Bigl|\bigl\langle \tilde{\theta}^1_{i_k} - \sqrt{1+b_k}\theta_{i_k}, x\bigr\rangle\Bigr| + \Bigl|\sqrt{1+b_k} - \sqrt{1+\tilde{b}_k}\Bigr|.\end{aligned} \tag{4.3.20}$$

The first term in the right hand side can be bounded using Theorem 23 and, for the second term,

$$\left|\sqrt{1+b_k} - \sqrt{1+\tilde{b}_k}\right| = \frac{|\tilde{b}_k - b_k|}{\sqrt{1+b_k} + \sqrt{1+\tilde{b}_k}} \lesssim_\gamma |\tilde{b}_k - b_k|,$$

so bound (4.3.18) can be used. Substituting these bounds in (4.3.20), we derive (4.3.19) in the case when $\frac{\tau\sqrt{t}}{\bar{g}_k} \le c_\gamma$.

In the opposite case, when $\frac{\tau\sqrt{t}}{\bar{g}_k} > c_\gamma$, we have

$$\left|\langle \hat{\theta}_{i_k}^{(\gamma)} - \theta_{i_k}, x\rangle\right| \le \|\hat{\theta}_{i_k}^{(\gamma)}\| + \|\theta_{i_k}\| \le \frac{1}{\sqrt{1+\tilde{b}_k \vee \frac{\sqrt{\gamma}}{2}}} + 1 \le \frac{2}{\sqrt{\gamma}} + 1.$$

Therefore,

$$\left|\langle \hat{\theta}_{i_k}^{(\gamma)} - \theta_{i_k}, x\rangle\right| \lesssim_\gamma \frac{\tau\sqrt{t}}{\bar{g}_k},$$

which implies (4.3.19) in this case.

□

REFERENCES

[1] Adamczak, R., "A tail inequality for suprema of unbounded empirical processes with applications to Markov chains," *Electronic Journal of Probability*, vol. 13, pp. 1000–1034, 2007.

[2] Alquier, P., Butucea, C., Hebiri, M., Meziani, K., and Morimae, T., "Rank-penalized estimation of a quantum system," *Physical Review A*, vol. 88, no. 3, p. 032113, 2013.

[3] Aubin, J.-P. and Ekeland, I., *Applied Nonlinear Analysis.* Courier Corporation, 2006.

[4] Aubrun, G., "On almost randomizing channels with a short Kraus decomposition," *Communications in Mathematical Physics*, vol. 288, no. 3, pp. 1103–1116, 2009.

[5] Audenaert, K. M. and Eisert, J., "Continuity bounds on the quantum relative entropy-ii," *Journal of Mathematical Physics*, vol. 52, no. 11, p. 112201, 2011.

[6] Bach, F. R., "Consistency of trace norm minimization," *The Journal of Machine Learning Research*, vol. 9, pp. 1019–1048, 2008.

[7] Benaych-Georges, F. and Nadakuditi, R. R., "The singular values and vectors of low rank perturbations of large rectangular random matrices," *Journal of Multivariate Analysis*, vol. 111, pp. 120–135, 2012.

[8] Bertsekas, D. P., "Projected newton methods for optimization problems with simple constraints," *SIAM Journal on control and Optimization*, vol. 20, no. 2, pp. 221–246, 1982.

[9] Bhatia, R., *Matrix analysis*, vol. 169. Springer Science & Business Media, 2013.

[10] Bickel, P. J., Ritov, Y., and Tsybakov, A. B., "Simultaneous analysis of LASSO and Dantzig selector," *The Annals of Statistics*, pp. 1705–1732, 2009.

[11] Bousquet, O., "A Bennett concentration inequality and its application to suprema of empirical processes," *Comptes Rendus Mathematique*, vol. 334, no. 6, pp. 495–500, 2002.

[12] Boyd, S., Parikh, N., Chu, E., Peleato, B., and Eckstein, J., "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends® in Machine Learning*, vol. 3, no. 1, pp. 1–122, 2011.

[13] Bužek, V., "Quantum tomography from incomplete Data via MaxEnt principle," in *Quantum State Estimation*, pp. 189–234, Springer, 2004.

[14] Cai, J.-F., Candès, E. J., and Shen, Z., "A singular value thresholding algorithm for matrix completion," *SIAM Journal on Optimization*, vol. 20, no. 4, pp. 1956–1982, 2010.

[15] Cai, T. T. and Zhang, A., "Sharp RIP bound for sparse signal and low-rank matrix recovery," *Applied and Computational Harmonic Analysis*, vol. 35, no. 1, pp. 74–93, 2013.

[16] Cai, T. T. and Zhang, A., "ROP: Matrix recovery via rank-one projections," *The Annals of Statistics*, vol. 43, no. 1, pp. 102–138, 2015.

[17] Cai, T. T. and Zhang, A., "Compressed sensing and affine rank minimization under restricted isometry," *EEE Transactions on Signal Processing*, vol. 61, no. 13, pp. 3279–3290, 2013.

[18] Candes, E. and Tao, T., "The Dantzig selector: statistical estimation when p is much larger than n," *The Annals of Statistics*, pp. 2313–2351, 2007.

[19] Candes, E. J. and Plan, Y., "Matrix completion with noise," *Proceedings of the IEEE*, vol. 98, no. 6, pp. 925–936, 2010.

[20] Candés, E. J. and Plan, Y., "Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements," *IEEE Transactions on Information Theory*, vol. 57, no. 4, pp. 2342–2359, 2011.

[21] Candès, E. J. and Recht, B., "Exact matrix completion via convex optimization," *Foundations of Computational mathematics*, vol. 9, no. 6, pp. 717–772, 2009.

[22] Candès, E. J. and Tao, T., "The power of convex relaxation: Near-optimal matrix completion," *IEEE Transactions on Information Theory*, vol. 56, no. 5, pp. 2053–2080, 2010.

[23] Chatterjee, S., "Matrix estimation by universal singular value thresholding," *The Annals of Statistics*, vol. 43, no. 1, pp. 177–214, 2015.

[24] Chen, C., He, B., and Yuan, X., "Matrix completion via an alternating direction method," *IMA Journal of Numerical Analysis*, vol. 32, no. 1, pp. 227–245, 2012.

[25] Chen, Y. and Ye, X., "Projection onto a simplex," *arXiv preprint arXiv:1101.6081*, 2011.

[26] Davenport, M. A., Plan, Y., van den Berg, E., and Wootters, M., "1-bit matrix completion," *Information and Inference*, vol. 3, no. 3, pp. 189–223, 2014.

[27] de la Peña, V. H. and Giné, E., *Decoupling. From Dependence to Independence.* Springer, 1999.

[28] Dudley, R. M., *Uniform central limit theorems*, vol. 23. Cambridge Univ Press, 1999.

[29] Eisenstat, S. C. and Ipsen, I. C., "Relative perturbation bounds for eigenspaces and singular vector subspaces," in *Proceedings of the Fifth SIAM Conference on Applied Linear Algebra*, pp. 62–66, 1994.

[30] Flammia, S. T., Gross, D., Liu, Y.-K., and Eisert, J., "Quantum tomography via compressed sensing: error bounds, sample complexity and efficient estimators," *New Journal of Physics*, vol. 14, no. 9, p. 095022, 2012.

[31] Gaiffas, S. and Lecué, G., "Sharp oracle inequalities for high-dimensional matrix prediction," *IEEE Transactions on Information Theory*, vol. 57, no. 10, pp. 6942–6957, 2011.

[32] Giné, E. and Nickl, R., "Mathematical foundations of infinite-dimensional statistical models," *Cambridge Series in Statistical and Probabilistic Mathematics*, 2015.

[33] Golub, G. H. and Van Loan, C. F., *Matrix Computations*, vol. 3. JHU Press, 2012.

[34] Grant, M., Boyd, S., and Ye, Y., "CVX: Matlab software for disciplined convex programming," 2008.

[35] Griffiths, D. J., *Introduction to quantum mechanics.* Pearson Education India, 2005.

[36] Gross, D., "Recovering low-rank matrices from few coefficients in any basis," *IEEE Transactions on Information Theory*, vol. 57, no. 3, pp. 1548–1566, 2011.

[37] Gross, D., Liu, Y.-K., Flammia, S. T., Becker, S., and Eisert, J., "Quantum state tomography via compressed sensing," *Physical Review Letters*, vol. 105, no. 15, p. 150401, 2010.

[38] Guédon, O., Mendelson, S., Pajor, A., and Tomczak-Jaegermann, N., "Majorizing measures and proportional subsets of bounded orthonormal systems," *Revista Matemática Iberoamericana*, vol. 24, no. 3, pp. 1075–1095, 2008.

[39] Helmberg, C., Rendl, F., Vanderbei, R. J., and Wolkowicz, H., "An interior-point method for semidefinite programming," *SIAM Journal on Optimization*, vol. 6, no. 2, pp. 342–361, 1996.

[40] Huang, L., Yan, D., Taft, N., and Jordan, M. I., "Spectral clustering with perturbed data," in *Advances in Neural Information Processing Systems*, pp. 705–712, 2009.

[41] Jain, P., Netrapalli, P., and Sanghavi, S., "Low-rank matrix completion using alternating minimization," in *Proceedings of the forty-fifth annual ACM symposium on Theory of computing*, pp. 665–674, ACM, 2013.

[42] James, G. M., Radchenko, P., and Lv, J., "DASSO: connections between the Dantzig selector and LASSO," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 71, no. 1, pp. 127–142, 2009.

[43] Jin, J., "Fast community detection by SCORE," *The Annals of Statistics*, vol. 43, no. 1, pp. 57–89, 2015.

[44] Kannan, R. and Vempala, S., *Spectral algorithms.* Now Publishers Inc, 2009.

[45] Keshavan, R., Montanari, A., and Oh, S., "Matrix completion from noisy entries," in *Advances in Neural Information Processing Systems*, pp. 952–960, 2009.

[46] Keshavan, R. H., Oh, S., and Montanari, A., "Matrix completion from a few entries," in *Information Theory, 2009. ISIT 2009. IEEE International Symposium on*, pp. 324–328, IEEE, 2009.

[47] Klauck, H., Nayak, A., Ta-Shma, A., and Zuckerman, D., "Interaction in quantum communication," *IEEE Transactions on Information Theory*, vol. 53, no. 6, pp. 1970–1982, 2007.

[48] Klopp, O., "Rank penalized estimators for high-dimensional matrices," *Electronic Journal of Statistics*, vol. 5, pp. 1161–1183, 2011.

[49] Klopp, O., "Noisy low-rank matrix completion with general sampling distribution," *Bernoulli*, vol. 20, no. 1, pp. 282–303, 2014.

[50] Klopp, O., "Matrix completion by singular value thresholding: sharp bounds," *Electronic Journal of Statistics*, vol. 9, no. 2, pp. 2348–2369, 2015.

[51] Koltchinskii, V., "The Dantzig selector and sparsity oracle inequalities," *Bernoulli*, vol. 15, no. 3, pp. 799–828, 2009.

[52] Koltchinskii, V., *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems: École d'Été de Probabilités de Saint-Flour XXXVIII-2008.* Springer, 2011.

[53] Koltchinskii, V., "von Neumann entropy penalization and low-rank matrix estimation," *The Annals of Statistics*, vol. 39, no. 6, pp. 2936–2973, 2011.

[54] Koltchinskii, V., "A remark on low rank matrix recovery and noncommutative Bernstein type inequalities," in *From Probability to Statistics and Back: High-Dimensional Models and Processes–A Festschrift in Honor of Jon A. Wellner*, pp. 213–226, Institute of Mathematical Statistics, 2013.

[55] Koltchinskii, V., "Sharp oracle inequalities in low rank estimation," in *Empirical Inference*, pp. 217–230, Springer, 2013.

[56] Koltchinskii, V. and Lounici, K., "Asymptotics and concentration bounds for spectral projectors of sample covariance," *arXiv preprint arXiv:1408.4643*, 2014.

[57] Koltchinskii, V., Lounici, K., and Tsybakov, A. B., "Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion," *The Annals of Statistics*, vol. 39, no. 5, pp. 2302–2329, 2011.

[58] Koltchinskii, V. and Xia, D., "Optimal estimation of low rank density matrices," *Journal of Machine Learning Research*, vol. 16, pp. 1757–1792, 2015.

[59] Koltchinskii, V. and Xia, D., "Perturbation of linear forms of singular vectors under gaussian noise," *arXiv preprint arXiv:1506.02764*, 2015.

[60] Lederer, J. and Van De Geer, S., "New concentration inequalities for suprema of empirical processes," *Bernoulli*, vol. 20, no. 4, pp. 2020–2038, 2014.

[61] Lei, J. and Rinaldo, A., "Consistency of spectral clustering in stochastic block models," *The Annals of Statistics*, vol. 43, no. 1, pp. 215–237, 2014.

[62] Li, R.-C., "Relative perturbation theory: Eigenspace and singular subspace variations," *SIAM Journal on Matrix Analysis and Applications*, vol. 20, no. 2, pp. 471–492, 1998.

[63] Lin, Z., Chen, M., and Ma, Y., "The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices," *arXiv preprint arXiv:1009.5055*, 2010.

[64] Liu, Y.-K., "Universal low-rank matrix recovery from Pauli measurements," in *Advances in Neural Information Processing Systems*, pp. 1638–1646, 2011.

[65] Liu, Y.-J., Sun, D., and Toh, K.-C., "An implementable proximal point algorithmic framework for nuclear norm minimization," *Mathematical Programming*, vol. 133, no. 1-2, pp. 399–436, 2012.

[66] Lounici, K., "Optimal spectral norm rates for noisy low-rank matrix completion." arxiv:1110.5346, 2011.

[67] Lounici, K., "High-dimensional covariance matrix estimation with missing observations," *Bernoulli*, vol. 20, no. 3, pp. 1029–1058, 2014.

[68] Ma, Z. and Wu, Y., "Volume ratio, sparsity, and minimaxity under unitarily invariant norms," in *Information Theory Proceedings (ISIT), 2013 IEEE International Symposium*, pp. 1027–1031, IEEE, 2013.

[69] Mendelson, S., "Upper bounds on product and multiplier empirical processes," *arXiv preprint arXiv:1410.8003*, 2014.

[70] Michelot, C., "A finite algorithm for finding the projection of a point onto the canonical simplex of $\mathbb{R}^n$," *Journal of Optimization Theory and Applications*, vol. 50, no. 1, pp. 195–200, 1986.

[71] Negahban, S. and Wainwright, M. J., “Restricted strong convexity and weighted matrix completion: Optimal bounds with noise,” *The Journal of Machine Learning Research*, vol. 13, no. 1, pp. 1665–1697, 2012.

[72] Negahban, S., Yu, B., Wainwright, M. J., and Ravikumar, P. K., “A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers,” in *Advances in Neural Information Processing Systems*, pp. 1348–1356, 2009.

[73] Nesterov, Y., “A method of solving a convex programming problem with convergence rate $O(1/k^2)$,”

[74] Nesterov, Y. and Nemirovski, A., “On first-order algorithms for l 1/nuclear norm minimization,” *Acta Numerica*, vol. 22, pp. 509–575, 2013.

[75] Nielsen, M. and Chuang, I., *Quantum Computation and Quantum Information.* Cambridge University Press, 2000.

[76] O’Rourke, S., Vu, V., and Wang, K., “Random perturbation of low rank matrices: Improving classical bounds,” *arXiv preprint arXiv:1311.2657*, 2013.

[77] Pajor, A., “Metric entropy of the Grassmann manifold,” *Convex Geometric Analysis*, vol. 34, pp. 181–188, 1998.

[78] Recht, B., “A simpler approach to matrix completion,” *The Journal of Machine Learning Research*, vol. 12, pp. 3413–3430, 2011.

[79] Recht, B., Fazel, M., and Parrilo, P. A., “Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization,” *SIAM Review*, vol. 52, no. 3, pp. 471–501, 2010.

[80] Rohde, A. and Tsybakov, A. B., “Estimation of high-dimensional low-rank matrices,” *The Annals of Statistics*, vol. 39, no. 2, pp. 887–930, 2011.

[81] Rohe, K., Chatterjee, S., and Yu, B., "Spectral clustering and the high-dimensional stochastic block model," *The Annals of Statistics*, vol. 39, no. 4, pp. 1878–1915, 2011.

[82] Rudelson, M. and Vershynin, R., "Delocalization of eigenvectors of random matrices with independent entries," *arXiv preprint arXiv:1306.2887*, 2013.

[83] Shalev-Shwartz, S. and Singer, Y., "Efficient learning of label ranking by soft projections onto polyhedra," *The Journal of Machine Learning Research*, vol. 7, pp. 1567–1599, 2006.

[84] Stewart, G. W., "Perturbation theory for the singular value decomposition," 1998.

[85] Sturm, J. F., "Using SeDuMi 1.02, a MATLAB toolbox for optimization over symmetric cones," *Optimization Methods and Software*, vol. 11, no. 1-4, pp. 625–653, 1999.

[86] Sun, T. and Zhang, C.-H., "Calibrated elastic regularization in matrix completion," in *Advances in Neural Information Processing Systems*, pp. 863–871, 2012.

[87] Talagrand, M., *The generic chaining: upper and lower bounds of stochastic processes.* Springer Science & Business Media, 2006.

[88] Tao, M. and Yuan, X., "Recovering low-rank and sparse components of matrices from incomplete and noisy observations," *SIAM Journal on Optimization*, vol. 21, no. 1, pp. 57–81, 2011.

[89] Tibshirani, R., "Regression shrinkage and selection via the LASSO," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288, 1996.

[90] Tropp, J. A., "User-friendly tail bounds for sums of random matrices," *Foundations of Computational Mathematics*, vol. 12, no. 4, pp. 389–434, 2012.

[91] Tsybakov, A. B., *Introduction to Nonparametric Estimation.* Springer, 2008.

[92] Van Der Vaart, A. W. and Wellner, J. A., *Weak Convergence.* Springer, 1996.

[93] Vandenberghe, L. and Boyd, S., "Semidefinite programming," *SIAM Review*, vol. 38, no. 1, pp. 49–95, 1996.

[94] Vershynin, R., "Introduction to the non-asymptotic analysis of random matrices," *arXiv preprint arXiv:1011.3027*, 2010.

[95] Vu, V., "Singular vectors under random perturbation," *Random Structures & Algorithms*, vol. 39, no. 4, pp. 526–538, 2011.

[96] Vu, V. and Wang, K., "Random weighted projections, random quadratic forms and random eigenvectors," *Random Structures & Algorithms*, 2014.

[97] Wang, R., "Singular vector perturbation under Gaussian noise," *SIAM Journal on Matrix Analysis and Applications*, vol. 36, no. 1, pp. 158–177, 2015.

[98] Wang, Y., "Asymptotic equivalence of quantum state tomography and noisy matrix completion," *The Annals of Statistics*, vol. 41, no. 5, pp. 2462–2504, 2013.

[99] Watson, G. A., "Characterization of the subdifferential of some matrix norms," *Linear algebra and its applications*, vol. 170, pp. 33–45, 1992.

[100] Xia, D., "Optimal Schatten-q and Ky-Fan-k norm rate of low rank matrix estimation," *arXiv preprint arXiv:1403.6499*, 2014.

[101] Xia, D. and Koltchinskii, V., "Estimation of low rank density matrices: bounds in Schatten norms and other distances," *arXiv preprint arXiv:1604.04600*, 2016.

[102] Ye, F. and Zhang, C.-H., "Rate minimaxity of the LASSO and Dantzig selector for the l_q loss in l_r balls," *The Journal of Machine Learning Research*, vol. 11, pp. 3519–3540, 2010.

[103] Zhao, P. and Yu, B., "On model selection consistency of LASSO," *The Journal of Machine Learning Research*, vol. 7, pp. 2541–2563, 2006.

[104] Zou, H., "The adaptive LASSO and its oracle properties," *Journal of the American Statistical Association*, vol. 101, no. 476, pp. 1418–1429, 2006.

VITA

Dong Xia was born in Huanggang, Hubei, China in 1989. He obtained his B.S. in Information and Computing Science from School of Mathematics at University of Science and Technology of China in 2007. He has been a Ph.D. student in Computational Science and Engineering at Georgia Institute of Technology since the Fall of 2011. He started working with Prof. Vladimir Koltchinskii since the spring of 2013 and his research is on mathematical statistics and machine learning. He works as a visiting assistant professor in the Department of Statistics at University of Wisconsin - Madison after receiving his Ph.D. degree in Summer 2016.