Submitted to the Annals of Statistics

OPTIMAL LARGE-SCALE QUANTUM STATE TOMOGRAPHY WITH PAULI MEASUREMENTS[*]

By Tony Cai[†], Donggyu Kim[‡], Yazhen Wang[‡], Ming Yuan[‡] and Harrison H. Zhou[§]

University of Pennsylvania[†]*, University of Wisconsin-Madison*[‡] *and Yale University*[§]

Quantum state tomography aims to determine the state of a quantum system as represented by a density matrix. It is a fundamental task in modern scientific studies involving quantum systems. In this paper, we study estimation of high-dimensional density matrices based on a relatively small number of Pauli measurements. In particular, under appropriate notion of sparsity, we establish the minimax optimal rates of convergence for estimation of the density matrix under both the spectral and Frobenius norm losses; and show how these rates can be achieved by a common thresholding approach. Numerical performance of the proposed estimator is also investigated.

1. Introduction. For a range of scientific studies including quantum computation, quantum information and quantum simulation, an important task is to learn and engineer quantum systems (Aspuru-Guzik et. al. (2005), Benenti et. al. (2004, 2007), Brumfiel (2012), Jones (2013), Lanyon et. al. (2010), Nielsen and Chuang (2000), and Wang (2011, 2012)). A quantum system is described by its state characterized by a density matrix, which is a positive semi-defintie Hermitian matrix with unit trace. Determining a quantum state, often referred to as quantum state tomography, is important but difficult (Alquier et. al. (2013), Artiles et. al. (2005), Aubry et. al. (2009), Butucea et. al. (2007), Guţă and Artiles (2007), Häffner et. al. (2005), Wang (2013), and Wang and Xu (2015)). It is often inferred by performing measurements on a large number of identically prepared quantum systems.

[*]The research of Tony Cai was supported in part by NSF Grants DMS-1208982 and DMS-1403708, and NIH Grant R01 CA127334. Yazhen Wang's research was partially supported by NSF grants DMS-105635 and DMS-1265203. The research of Ming Yuan was supported in part by NSF Career Award DMS-1321692 and FRG Grant DMS-1265202. The research of Harrison Zhou is supported in part by NSF Grant DMS-1209191.

MSC 2010 subject classifications: Primary 62H12, 81P50; secondary 62C20, 62P35, 81P45, 81P68

Keywords and phrases: Compressed sensing, density matrix, Pauli matrices, quantum measurement, quantum probability, quantum statistics, sparse representation, spectral norm, minimax estimation

More specifically we describe a quantum spin system by the d-dimensional complex space $\mathbb{C}^d$ and its quantum state by a complex matrix on $\mathbb{C}^d$. When measuring the quantum system by performing measurements on some observables which can be represented by Hermitian matrices, we obtain the measurement outcomes for each observable, where the measurements take values at random from all eigenvalues of the observable, with the probability of observing a particular eigenvalue equal to the trace of the product of the density matrix and the projection matrix onto the eigen-space corresponding to the eigenvalue. To handle the up and down states of particles in a quantum spin system, we usually employ well-known Pauli matrices as observables to perform measurements and obtain the so-called Pauli measurements (Britton et. al. (2012), Johnson et. al. (2011), Liu (2011), Sakurai and Napolitano (2010), Shankar (1994), and Wang (2012, 2013)). Since all Pauli matrices have ± 1 eigen-values, Pauli measurements takes discrete values 1 and -1, and the resulted measurement distributions can be characterized by binomial distributions. Our goal is to estimate the density matrix by the Pauli measurements.

Traditionally quantum tomography employs classical statistical models and methods to deduce quantum states from quantum measurements. These approaches are designed for the setting where the size of a density matrix is greatly exceeded by the number of quantum measurements, which is almost never the case even for moderate quantum systems in practice because the dimension of the density matrix grows exponentially in the size of the quantum system. For example, the density matrix for a one-dimensional quantum spin chain of size b is of size $2^b \times 2^b$. In this paper, we consider specifically how the density matrix could be effectively and efficiently reconstructed for a large-scale quantum system with a relatively limited number of quantum measurements.

Quantum state tomography is fundamentally connected to the problem of recovering a high dimensional matrix based on noisy observations (Wang (2013)). The latter problem arises naturally in many applications in statistics and machine learning and has attracted considerable recent attention. When assuming that the matrix parameter of interest is of (approximately) low-rank, many regularization techniques have been developed. Examples include Candès and Recht (2008), Candès and Tao (2009), Candès and Plan (2009a, b), Keshavan, Montanari, and Oh (2010), Recht, Fazel, and Parrilo (2010), Bunea, She and Wegkamp (2011, 2012), Klopp (2011, 2012), Koltchinskii (2011), Koltchinskii, Lounici and Tsybakov (2011), Negahban and Wainwright (2011), Recht (2011), Rohde and Tsybakov (2011), and Cai and Zhang (2015), among many others. Taking advantage of the low-rank structure of the matrix parameter, these approaches can often be applied to estimate matrix parameters of high dimensions. Yet these methods do not fully account for the specific structure of quantum state tomography. As demonstrated in a pioneering article, Gross et al. (2010) argued that, when considering quantum measurements characterized by the Pauli matrices, the density matrix can often be characterized by the sparsity with respect to the Pauli basis. Built upon this connection, they suggested a compressed sensing (Donoho (2006)) strategy for quantum state tomography (Gross (2011) and Wang (2013)). Although promising, their proposal assumes exact measurements, which is rarely the case in practice, and adopts the constrained nuclear norm minimization method, which may not be an appropriate matrix completion approach for estimating a density matrix with unit trace (or unit nuclear norm). We specifically address such challenges in the present article. In particular, we establish the minimax optimal rates of convergence for the density matrix estimation in terms of both spectral and Frobenius norms when assuming that the true density matrix is approximately sparse under the Pauli basis. Furthermore, we show that these rates could be achieved by carefully thresholding the coefficients with respect to Pauli basis. Because the quantum Pauli measurements are characterized by binomial distributions, the convergence rates and minimax lower bounds are derived by asymptotic analysis with manipulations of binomial distributions instead of the usual normal distribution based calculations.

The rest of paper proceeds as follows. Section 2 gives some background on quantum state tomography and introduces a thresholding based density matrix estimator. Section 3 develops theoretical properties for the density matrix estimation problem. In particular, the convergence rates of the proposed density matrix estimator and its minimax optimality with respect to both the spectral and Frobenius norm losses are established. Section 4 features a simulation study to illustrate finite sample performance of the proposed estimators. All technical proofs are collected in Section 5.

2. Quantum state tomography with Pauli measurements. In this section, we first review the quantum state and density matrix and introduce Pauli matrices and Pauli measurements. We also develop results to describe density matrix representations through Pauli matrices and characterize the distributions of Pauli measurements via binomial distribution before introducing a thresholding based density matrix estimator.

2.1. *Quantum state and measurements.* For a d-dimensional quantum system, we describe its quantum state by a density matrix $\boldsymbol{\rho}$ on d dimensional complex space $\mathbb{C}^d$, where density matrix $\boldsymbol{\rho}$ is a d by d complex matrix satisfying (1) Hermitian, that is, $\boldsymbol{\rho}$ is equal to its conjugate transpose; (2) positive semi-definite; (3) unit trace i.e. $tr(\boldsymbol{\rho}) = 1$.

For a quantum system it is important but difficult to know its quantum state. Experiments are conducted to perform measurements on the quantum system and obtain data for studying the quantum system and estimating its density matrix. In physics literature quantum state tomography refers to reconstruction of a quantum state based on measurements for the quantum systems. Statistically it is the problem of estimating the density matrix from the measurements. Common quantum measurements are on observable $\mathbf{M}$, which is defined as a Hermitian matrix on $\mathbb{C}^d$. Assume that the observable $\mathbf{M}$ has the following spectral decomposition,

$$\mathbf{M} = \sum_{a=1}^{r} \lambda_a \, \mathbf{Q}_a, \tag{2.1}$$

where λ_a are r different real eigenvalues of $\mathbf{M}$, and $\mathbf{Q}_a$ are projections onto the eigen-spaces corresponding to λ_a. For the quantum system prepared in state $\boldsymbol{\rho}$, we need a probability space $(\Omega, \mathcal{F}, P)$ to describe measurement outcomes when performing measurements on the observable $\mathbf{M}$. Denote by R the measurement outcome of $\mathbf{M}$. According to the theory of quantum mechanics, R is a random variable on $(\Omega, \mathcal{F}, P)$ taking values in $\{\lambda_1, \lambda_2, \cdots, \lambda_r\}$, with probability distribution given by

$$P(R = \lambda_a) = tr(\mathbf{Q}_a \, \boldsymbol{\rho}), \; a = 1, 2, \cdots, r, \;\; E(R) = tr(\mathbf{M}\boldsymbol{\rho}). \tag{2.2}$$

We may perform measurements on an observable for a quantum system that is identically prepared under the state and obtain independent and identically distributed observations. See Holevo (1982), Sakurai and Napolitano (2010), and Wang (2012).

2.2. *Pauli measurements and their distributions.* The Pauli matrices as observables are widely used in quantum physics and quantum information science to perform quantum measurements. Let

$$\boldsymbol{\sigma}_0 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \; \boldsymbol{\sigma}_1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \; \boldsymbol{\sigma}_2 = \begin{pmatrix} 0 & -\sqrt{-1} \\ \sqrt{-1} & 0 \end{pmatrix}, \; \boldsymbol{\sigma}_3 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix},$$

where $\boldsymbol{\sigma}_1$, $\boldsymbol{\sigma}_2$ and $\boldsymbol{\sigma}_3$ are called the two dimensional Pauli matrices. Tensor products are used to define high dimensional Pauli matrices. Let $d = 2^b$ for some integer b. We form b-fold tensor products of $\boldsymbol{\sigma}_0$, $\boldsymbol{\sigma}_1$, $\boldsymbol{\sigma}_2$ and $\boldsymbol{\sigma}_3$ to obtain d dimensional Pauli matrices

$$\boldsymbol{\sigma}_{\ell_1} \otimes \boldsymbol{\sigma}_{\ell_2} \otimes \cdots \otimes \boldsymbol{\sigma}_{\ell_b}, \qquad (\ell_1, \ell_2, \cdots, \ell_b) \in \{0, 1, 2, 3\}^b. \tag{2.3}$$

We identify index $j = 1, \cdots, d^2$ with $(\ell_1, \ell_2, \cdots, \ell_b) \in \{0, 1, 2, 3\}^b$. For example, $j = 1$ corresponds to $\ell_1 = \cdots = \ell_b = 0$. With the index identification we denote by $\mathbf{B}_j$ the Pauli matrix $\boldsymbol{\sigma}_{\ell_1} \otimes \boldsymbol{\sigma}_{\ell_2} \otimes \cdots \otimes \boldsymbol{\sigma}_{\ell_b}$, with $\mathbf{B}_1 = \mathbf{I}_d$. We have the following theorem to describe Pauli matrices and represent a density matrix by Pauli matrices.

PROPOSITION 1. *(i) Pauli matrices $\mathbf{B}_2, \cdots, \mathbf{B}_{d^2}$ are of full rank and have eigenvalues ± 1. Denote by $\mathbf{Q}_{j\pm}$ the projections onto the eigenspaces of $\mathbf{B}_j$ corresponding to eigenvalues ± 1, respectively. Then for $j, j' = 2, \cdots, d^2$,*

$$tr(\mathbf{Q}_{j\pm}) = \frac{d}{2}, \qquad tr(\mathbf{B}_{j'}\mathbf{Q}_{j\pm}) = \begin{cases} \pm\frac{d}{2} & \text{if } j = j' \\ 0 & \text{if } j \neq j'. \end{cases} \tag{2.4}$$

(ii) Denote by $\mathbb{C}^{d\times d}$ the space of all d by d complex matrices equipped with the Frobenius norm. All Pauli matrices defined by (2.3) form an orthogonal basis for all complex Hermitian matrices. Given a density matrix $\boldsymbol{\rho}$ we can expand it under the Pauli basis as follows,

$$\boldsymbol{\rho} = \frac{\mathbf{I}_d}{d} + \sum_{j=2}^{d^2} \beta_j \frac{\mathbf{B}_j}{d}, \tag{2.5}$$

where β_j are coefficients. For $j = 2, \cdots, d^2$,

$$tr(\boldsymbol{\rho}\mathbf{Q}_{j\pm}) = \frac{1 \pm \beta_j}{2}. \tag{2.6}$$

Suppose that an experiment is conducted to perform measurements on Pauli observable $\mathbf{B}_j$ independently for n quantum systems which are identically prepared in the same quantum state $\boldsymbol{\rho}$. As $\mathbf{B}_j$ has eigenvalues ± 1, the Pauli measurements take values 1 and -1, and thus the average of the n measurements for each $\mathbf{B}_j$ is a sufficient statistics. Denote by N_j the average of the n measurement outcomes obtained from measuring $\mathbf{B}_j$, $j = 2, \cdots, d^2$. Our goal is to estimate $\boldsymbol{\rho}$ based on $N_2, \cdots, N_{d^2}$.

The following proposition provides a simple binomial characterization for the distributions of N_j.

PROPOSITION 2. *Suppose that $\boldsymbol{\rho}$ is given by (2.5). Then $N_2, \cdots, N_{d^2}$ are independent with*

$$E(N_j) = \beta_j, \qquad Var(N_j) = \frac{1 - \beta_j^2}{n},$$

and $n(N_j+1)/2$ follows a binomial distribution with n trials and cell probabilities $tr(\boldsymbol{\rho}\mathbf{Q}_{j+}) = (1+\beta_j)/2$, where $\mathbf{Q}_{j+}$ denotes the projection onto the eigen-space of $\mathbf{B}_j$ corresponding to eigenvalue 1, and β_j is the coefficient of $\mathbf{B}_j$ in the expansion of $\boldsymbol{\rho}$ in (2.5).

2.3. *Density matrix estimation.* Since the dimension of a quantum system grows exponentially with its components such as the number of particles in the system, the matrix size of $\boldsymbol{\rho}$ tends to be very large even for a moderate quantum system. We need to impose some structure such as sparsity on $\boldsymbol{\rho}$ in order to make it consistently estimable. Suppose that $\boldsymbol{\rho}$ has a sparse representation under the Pauli basis, following wavelet shrinkage estimation we construct a density matrix estimator of $\boldsymbol{\rho}$. Assume that representation (2.5) is sparse in a sense that there is only a relatively small number of coefficients β_k with large magnitudes. Formally we specify sparsity by assuming that coefficients $\beta_2, \cdots, \beta_{d^2}$ satisfy

$$\sum_{k=2}^{d^2} |\beta_k|^q \le \pi_n(d), \tag{2.7}$$

where $0 \le q < 1$, and $\pi_n(d)$ is a deterministic function with slow growth in d such as $\log d$.

Since N_k are independent, and $E(N_k) = \beta_k$. We naturally estimate β_k by N_k and threshold N_k to estimate large β_k, ignoring small β_k, and obtain

$$\hat\beta_k = N_k 1(|N_k| \ge \varpi) \text{ or } \hat\beta_k = sign(N_k)\,(|N_k| - \varpi)_+, \quad k = 2, \cdots, d^2, \tag{2.8}$$

and then we use $\hat\beta_k$ to construct the following estimator of $\boldsymbol{\rho}$,

$$\hat{\boldsymbol{\rho}} = \frac{\mathbf{I}_d}{d} + \sum_{k=2}^{p} \hat\beta_k \frac{\mathbf{B}_k}{d}, \tag{2.9}$$

where the two estimation methods in (2.8) are called hard and soft thresholding rules, and ϖ is a threshold value which, we reason below, can be chosen to be $\varpi = \hbar\sqrt{(4/n)\log d}$ for some constant $\hbar > 1$. The threshold value is designed such that for small β_k, N_k must be bounded by threshold ϖ with overwhelming probability, and the hard and soft thresholding rules select only those N_k with large signal components β_k.

As $n(N_k+1)/2 \sim Bin(n, (1+\beta_k)/2)$, an application of Bernstein's inequality leads to that for any $x > 0$,

$$P(|N_k - \beta_k| \ge x) \le 2\exp\left(-\frac{nx^2}{2(1-\beta_k^2 + x/3)}\right) \le 2\exp\left(-\frac{nx^2}{2}\right),$$

and

$$P\left(\max_{2\le k\le d^2}|N_k-\beta_k|\le\varpi\right)=\prod_{k=2}^{d^2}P\left(|N_k-\beta_k|\le\varpi\right)$$
$$\ge\left[1-2\exp\left(-\frac{n\varpi^2}{2}\right)\right]^{d^2-1}=\left[1-2d^{-2\hbar}\right]^{d^2-1}\to 1,$$

as $d\to\infty$, that is, with probability tending to one, $|N_k|\le\varpi$ uniformly for $k=2,\cdots,d^2$. Thus we can select $\varpi=\hbar\sqrt{(4/n)\log d}$ to threshold N_k and obtain $\hat\beta_k$ in (2.8).

3. Asymptotic theory for the density matrix estimator.

3.1. *Convergence rates.* We fix matrix norm notations for our asymptotic analysis. Let $\mathbf{x}=(x_1,\cdots,x_d)^T$ be a d-dimensional vector and $\mathbf{A}=(A_{ij})$ be a d by d matrix, and define their ℓ_α norms

$$\|\mathbf{x}\|_\alpha=\left(\sum_{i=1}^d|x_i|^\alpha\right)^{1/\alpha},\qquad\|\mathbf{A}\|_\alpha=\sup\{\|\mathbf{A}\,\mathbf{x}\|_\alpha,\|\mathbf{x}\|_\alpha=1\},\qquad 1\le\alpha\le\infty.$$

Denote by $\|\mathbf{A}\|_F=\sqrt{tr(\mathbf{A}^\dagger\mathbf{A})}$ the Frobenius norm of $\mathbf{A}$.

For the case of matrix, the ℓ_2 norm is called the matrix spectral norm or operator norm. $\|\mathbf{A}\|_2$ is equal to the square root of the largest eigenvalue of $\mathbf{A}\,\mathbf{A}^\dagger$,

$$\|\mathbf{A}\|_1=\max_{1\le j\le d}\sum_{i=1}^d|A_{ij}|,\qquad\|\mathbf{A}\|_\infty=\max_{1\le i\le d}\sum_{j=1}^d|A_{ij}|,\tag{3.1}$$

and

$$\|\mathbf{A}\|_2^2\le\|\mathbf{A}\|_1\,\|\mathbf{A}\|_\infty.\tag{3.2}$$

For a real symmetric or complex Hermitian matrix $\mathbf{A}$, $\|\mathbf{A}\|_2$ is equal to the largest absolute eigenvalue of $\mathbf{A}$, $\|\mathbf{A}\|_F$ is the square root of the sum of squared eigenvalues, $\|\mathbf{A}\|_F\le\sqrt{d}\,\|\mathbf{A}\|_2$, and (3.1)-(3.2) imply that $\|\mathbf{A}\|_2\le\|\mathbf{A}\|_1=\|\mathbf{A}\|_\infty$.

The following theorem gives the convergence rates for $\hat{\boldsymbol{\rho}}$ under the spectral and Frobenius norms.

THEOREM 1. *Denote by Θ the class of density matrices satisfying the sparsity condition (2.7). Assume $d \geq n^{c_0}$ for some constant $c_0 > 0$. For density matrix estimator $\hat{\boldsymbol{\rho}}$ defined by (2.8)-(2.9) with threshold $\varpi = \hbar\sqrt{(4/n)\log d}$ for some constant $\hbar > 1$, we have*

$$\sup_{\boldsymbol{\rho}\in\Theta} E[\|\hat{\boldsymbol{\rho}} - \boldsymbol{\rho}\|_2^2] \leq c_1\, \pi_n^2(d)\, \frac{1}{d^2} \left(\frac{\log d}{n}\right)^{1-q},$$

$$\sup_{\boldsymbol{\rho}\in\Theta} E[\|\hat{\boldsymbol{\rho}} - \boldsymbol{\rho}\|_F^2] \leq c_2\, \pi_n(d)\, \frac{1}{d} \left(\frac{\log d}{n}\right)^{1-q/2},$$

where c_1 and c_2 are constants free of n and d.

Remark 1. Theorem 1 shows that $\hat{\boldsymbol{\rho}}$ has convergence rate $\pi_n^{1/2}(d)d^{-1/2}$ $(n^{-1/2}\log^{1/2} d)^{1-q/2}$ under the Frobenius norm and convergence rate $\pi_n(d)d^{-1}$ $(n^{-1/2}\log^{1/2} d)^{1-q}$ under the spectral norm, which will be shown to be optimal in next section. Similar to the optimal convergence rates for large covariance and volatility matrix estimation (Cai and Zhou (2012) and Tao, Wang and Zhou (2013)), the optimal convergence rates here have factors involving $\pi_n(d)$ and $\log d/n$. However, unlike the covariance and volatility matrix estimation case, the convergence rates in Theorem 1 have factors $d^{-1/2}$ and d^{-1} for the spectral and Frobenius norms, respectively, and go to zero as d approaches to infinity. In particular the result implies that MSEs of the proposed estimator get smaller for large d. This is quite contrary to large covariance and volatility matrix estimation where the traces are typically diverge, the optimal convergence rates grow with the logarithm of matrix size, and the corresponding MSEs increase in matrix size. The new phenomenon may be due to the unit trace constraint on density matrix and that the density matrix representation (2.5) needs a scaling factor d^{-1} to satisfy the constraint. As a result, the assumption imposed on d and n in Theorem 1 does not include the usual upper bound requirement on d by an exponential growth with sample size in large covariance and volatility matrix estimation. Also for finite sample $\hat{\boldsymbol{\rho}}$ may not be positive semi-definite, we may project $\hat{\boldsymbol{\rho}}$ onto the cone formed by all density matrices and obtain a positive semi-definite density matrix estimator $\tilde{\boldsymbol{\rho}}$. As the underlying true density matrix $\boldsymbol{\rho}$ is positive semi-definite, the distance between $\tilde{\boldsymbol{\rho}}$ and $\boldsymbol{\rho}$ will be bounded by twice the distance between $\hat{\boldsymbol{\rho}}$ and $\boldsymbol{\rho}$, and thus $\tilde{\boldsymbol{\rho}}$ has the same convergence rates as $\hat{\boldsymbol{\rho}}$.

3.2. *Optimality of the density matrix estimator.* The following theorem establishes a minimax lower bound for estimating $\boldsymbol{\rho}$ under the spectral norm.

THEOREM 2. *We assume that $\pi_n(d)$ in the sparsity condition (2.7) satisfies*

$$\pi_n(d) \leq \aleph\, d^v \,(\log d)^{q/2}\,/n^{q/2}, \tag{3.3}$$

for some constant $\aleph > 0$ and $0 < v < 1/2$. Then

$$\inf_{\check{\boldsymbol{\rho}}} \sup_{\boldsymbol{\rho}\in\Theta} E[\|\check{\boldsymbol{\rho}} - \boldsymbol{\rho}\|_2^2] \geq c_3\, \pi_n^2(d)\, \frac{1}{d^2} \left(\frac{\log d}{n}\right)^{1-q},$$

where $\check{\boldsymbol{\rho}}$ denotes any estimator of $\boldsymbol{\rho}$ based on measurement data $N_2, \cdots, N_{d^2}$, and c_3 is a constant free of n and d.

Remark 2. The lower bound in Theorem 2 matches the convergence rate of $\hat{\boldsymbol{\rho}}$ under the spectral norm in Theorem 1, so we conclude that $\hat{\boldsymbol{\rho}}$ achieves the optimal convergence rate under the spectral norm. To establish the minimax lower bound in Theorem 2, we construct a special subclass of density matrices and then apply Le Cam's lemma. Assumption (3.3) is needed to guarantee the positive definiteness of the constructed matrices as density matrix candidates and to ensure the boundedness below from zero for the total variation of related probability distributions in Le Cam's lemma. Assumption (3.3) is reasonable in a sense that if the right hand side of (3.3) is large enough, (3.3) will not impose very restrictive condition on $\pi_n(d)$. We evaluate the dominating factor $n^{-q/2}d^v$ on the right hand side of (3.3) for various scenarios. First consider $q = 0$, the assumption becomes $\pi_n(d) \leq \aleph\, d^v$, $v < 1/2$, and so Assumption (3.3) essentially requires $\pi_n(d)$ grows in d not faster than $d^{1/2}$, which is not restrictive at all as $\pi_n(d)$ usually grows slowly in d. The asymptotic analysis of high dimensional statistics usually allows both d and n go to infinity. Typically we may assume d grows polynomially or exponentially in n. If d grows exponentially in n, that is, $d \sim \exp(b_0\, n)$ for some $b_0 > 0$, then $n^{q/2}$ is negligible in comparison with d^v, and $n^{-q/2}d^v$ behavior like d^v. The assumption in this case is not very restrictive. For the case of polynomial growth, that is, $d \sim n^{b_1}$ for some $b_1 > 0$, then $n^{-q/2}d^v \sim d^{v-q/(2b_1)}$. If $v - q/(2b_1) > 0$, $n^{-q/2}d^v$ grows in d like some positive power of d. Since we may take v arbitrarily close to $1/2$, the positiveness of $v - q/(2b_1)$ essentially requires $b_1 > q$, which can often be quite realistic given that q is usually very small.

The theorem below provides a minimax lower bound for estimating $\boldsymbol{\rho}$ under the Frobenius norm.

THEOREM 3. *We assume that $\pi_n(d)$ in the sparsity condition (2.7) satisfies*

$$\pi_n(d) \leq \aleph'\, d^{v'}/n^q, \tag{3.4}$$

for some constants $\aleph' > 0$ *and* $0 < v' < 2$. *Then*

$$\inf_{\check{\boldsymbol{\rho}}} \sup_{\boldsymbol{\rho} \in \Theta} E[\|\check{\boldsymbol{\rho}} - \boldsymbol{\rho}\|_F^2] \geq c_4 \, \pi_n(d) \frac{1}{d} \left(\frac{\log d}{n} \right)^{1-q/2},$$

where $\check{\boldsymbol{\rho}}$ *denotes any estimator of* $\boldsymbol{\rho}$ *based on measurement data* $N_2, \cdots, N_{d^2}$*, and* c_4 *is a constant free of* n *and* d.

Remark 3. The lower bound in Theorem 3 matches the convergence rate of $\hat{\boldsymbol{\rho}}$ under the Frobenius norm in Theorem 1, so we conclude that $\hat{\boldsymbol{\rho}}$ achieves the optimal convergence rate under the Frobenius norm. Similar to the Remark 2 after Theorem 2, we need to apply Assouad's lemma to establish the minimax lower bound in Theorem 3, and Assumption (3.4) is used to guarantee the positive definiteness of the constructed matrices as density matrix candidates and to ensure the boundedness below from zero for the total variation of related probability distributions in Assouad's lemma. Also the appropriateness of (3.4) is more relaxed than (3.3), as $v' < 2$ and the right hand of (3.4) has main powers more than the square of that of (3.3).

4. A simulation study.

A simulation study was conducted to investigate the performance of the proposed density matrix estimator for the finite sample. We took $d = 32, 64, 128$ and generated a true density matrix $\boldsymbol{\rho}$ for each case as follows. $\boldsymbol{\rho}$ has an expansion over the Pauli basis

$$\boldsymbol{\rho} = d^{-1} \left(\mathbf{I}_d + \sum_{j=2}^{d^2} \beta_j \mathbf{B}_j \right),$$

where $\beta_j = tr(\boldsymbol{\rho} \mathbf{B}_j)$, $j = 2, \cdots, d^2$. From $\beta_2, \cdots, \beta_{d^2}$ we randomly selected $[6 \log d]$ coefficients β_j and set the rest of β_j to be zero. We simulated $[6 \log d]$ values independently from a uniform distribution on $[-0.2, 0.2]$ and assigned the simulated values at random to the selected β_j. We repeated the procedure to generate a positive semi-definite $\boldsymbol{\rho}$ and took it as the true density matrix. The simulation procedure guarantees the obtained $\boldsymbol{\rho}$ is a density matrix and has a sparse representation under the Pauli basis.

For each true density matrix $\boldsymbol{\rho}$, as described in Section 2.2 we simulated data N_j from a binomial distribution with cell probability β_j and the number of cells $n = 100, 200, 500, 1000, 2000$. We constructed coefficient estimators $\hat{\beta}_j$ by (2.8) and obtained density matrix estimator $\hat{\boldsymbol{\rho}}$ using (2.9). The whole estimation procedure is repeated 200 times. The density matrix estimator is measured by the mean squared errors (MSE), $E\|\hat{\boldsymbol{\rho}} - \boldsymbol{\rho}\|_2^2$ and $E\|\hat{\boldsymbol{\rho}} - \boldsymbol{\rho}\|_F^2$, that are evaluated by the average of $\|\hat{\boldsymbol{\rho}} - \boldsymbol{\rho}\|_2^2$ and $\|\hat{\boldsymbol{\rho}} - \boldsymbol{\rho}\|_F^2$ over 200 repetitions, respectively. Three thresholds were used in the simulation study: the universal threshold $1.01\sqrt{4\log d/n}$ for all β_j, the individual threshold $1.01\sqrt{4(1-N_j^2)\log d/n}$ for each β_j, and the optimal threshold for all β_j, which minimizes the computed MSE for each corresponding hard or soft threshold method. The individual threshold takes into account the fact in Theorem 2 that the mean and variance of N_j are β_j and $(1-\beta_j^2)/n$, respectively, and the variance of N_j is estimated by $(1-N_j^2)/n$.

Figures 1 and 2 plot the MSEs of the density matrix estimators with hard and soft threshold rules and its corresponding density matrix estimator without thresholding (i.e. β_j are estimated by N_j in (2.9)) against the sample size n for different matrix size d, and Figures 3 and 4 plot their MSEs against matrix size d for different sample size. The numerical values of the MSEs are reported in Table 1. The figures 1 and 2 show that the MSEs usually decrease in sample size n, and the thresholding density matrix estimators enjoy superior performances than that the density matrix estimator without thresholding even for $n = 2000$; while all threshold rules and threshold values yield thresholding density matrix estimators with very close MSEs, the soft threshold rule with individual and universal threshold values produce larger MSEs than others for larger sample size such as $n = 1000, 2000$, and the soft threshold rule tends to give somewhat better performance than the hard threshold rule for smaller sample size like $n = 100, 200$. Figures 3 and 4 demonstrates that while the MSEs of all thresholding density matrix estimators decrease in the matrix size d, but if we rescale the MSEs by multiplying it with d^2 for the spectral norm case and d for the Frobenius norm case, the rescaled MSEs slowly increase in matrix size d. The simulation results largely confirm the theoretical findings discussed in Remark 1.

5. Proofs. Let $p = d^2$. Denote by C's generic constants whose values are free of n and p and may change from appearance to appearance. Let $u \vee v$ and $u \wedge v$ be the maximum and minimum of u and v, respectively. For two sequences $u_{n,p}$ and $v_{n,p}$, we write $u_{n,p} \sim v_{n,p}$ if $u_{n,p}/v_{n,p} \to 1$ as $n, p \to \infty$, and write $u_{n,p} \asymp v_{n,p}$ if there exist positive constants C_1 and C_2 free of n and p such that $C_1 \le u_{n,p}/v_{n,p} \le C_2$. Let $p = d^2$, and without confusion we may write $\pi_n(d)$ as $\pi_n(p)$.

5.1. *Proofs of Propositions 1 and 2.* **Proof of Proposition 1** In two dimensions, Pauli matrices satisfy $tr(\boldsymbol{\sigma}_0) = 2$, and $tr(\boldsymbol{\sigma}_1) = tr(\boldsymbol{\sigma}_2) = tr(\boldsymbol{\sigma}_3) = 0$, $\boldsymbol{\sigma}_1, \boldsymbol{\sigma}_2, \boldsymbol{\sigma}_3$ have eigenvalues ± 1, the square of a Pauli matrix is equal to the identity matrix, and the multiplications of any two Pauli matrices are equal

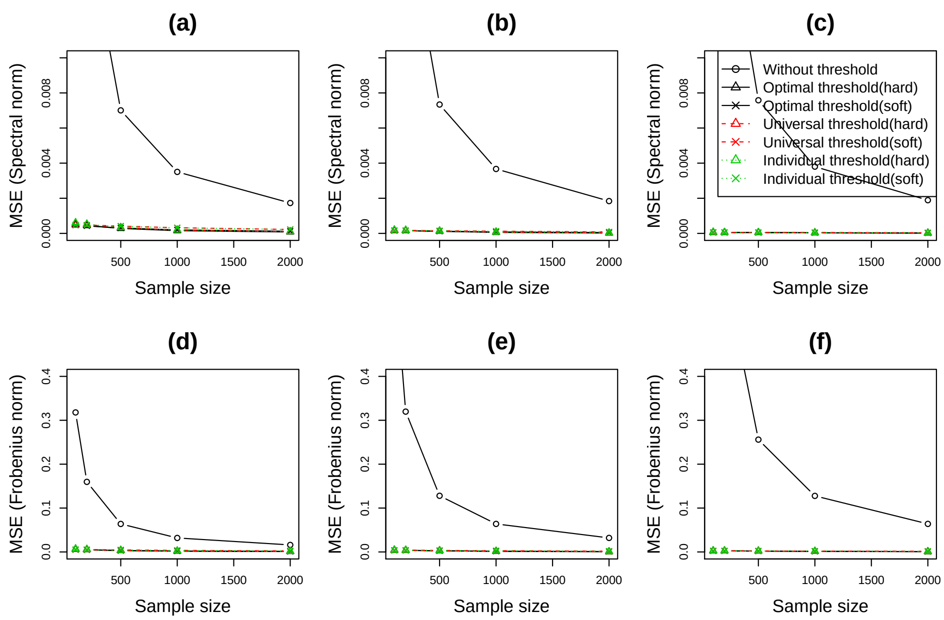

FIG 1. *The MSE plots against sample size for the proposed density estimator with hard and soft threshold rules and its corresponding estimator without thresholding for* $d = 32, 64, 128$. *(a)-(c) are plots of MSEs based on the spectral norm for* $d = 32, 64, 128$, *respectively, and (d)-(f) are plots of MSEs based on the Frobenius norm for* $d = 32, 64, 128$, *respectively.*

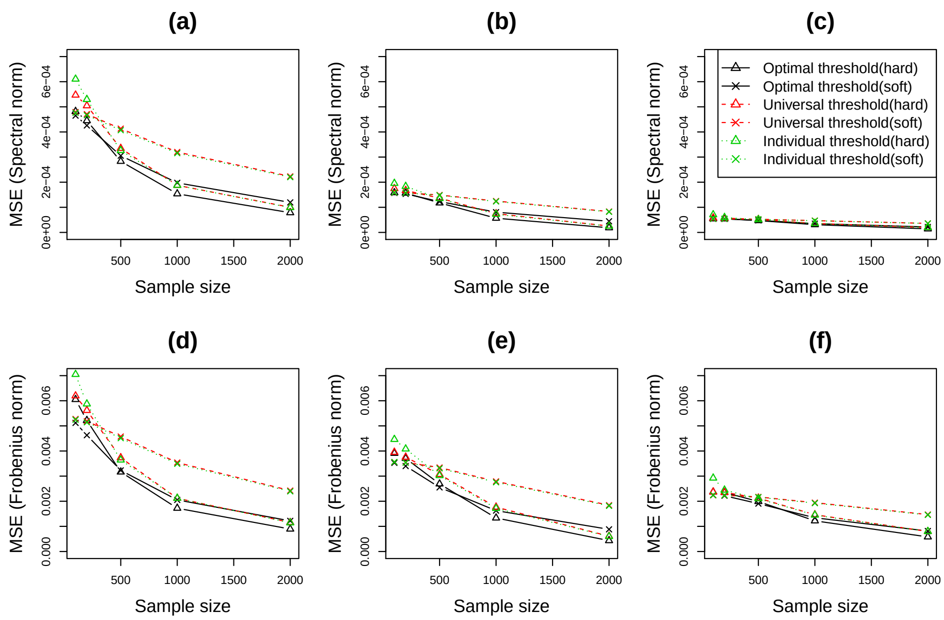

FIG 2. *The MSE plots against sample size for the proposed density estimator with hard and soft threshold rules for* $d = 32, 64, 128$. *(a)-(c) are plots of MSEs based on the spectral norm for* $d = 32, 64, 128$*, respectively, and (d)-(f) are plots of MSEs based on the Frobenius norm for* $d = 32, 64, 128$*, respectively.*

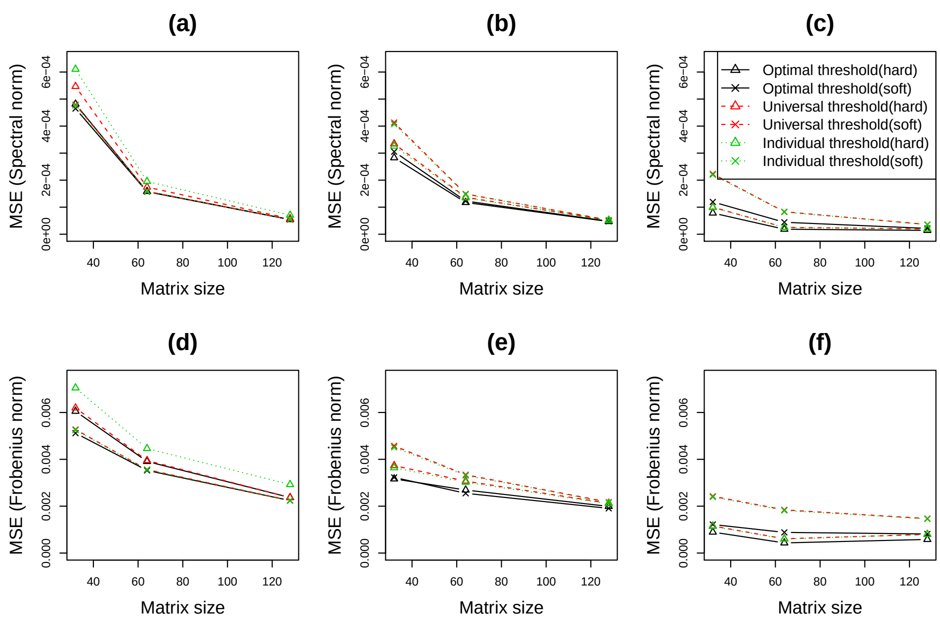

Fig 3. *The MSE plots against matrix size for the proposed density estimator with hard and soft threshold rules for $n = 100, 500, 2000$. (a)-(c) are plots of MSEs based on the spectral norm for $n = 100, 500, 2000$, respectively, and (d)-(f) are plots of MSEs based on the Frobenius norm for $n = 100, 500, 2000$, respectively.*

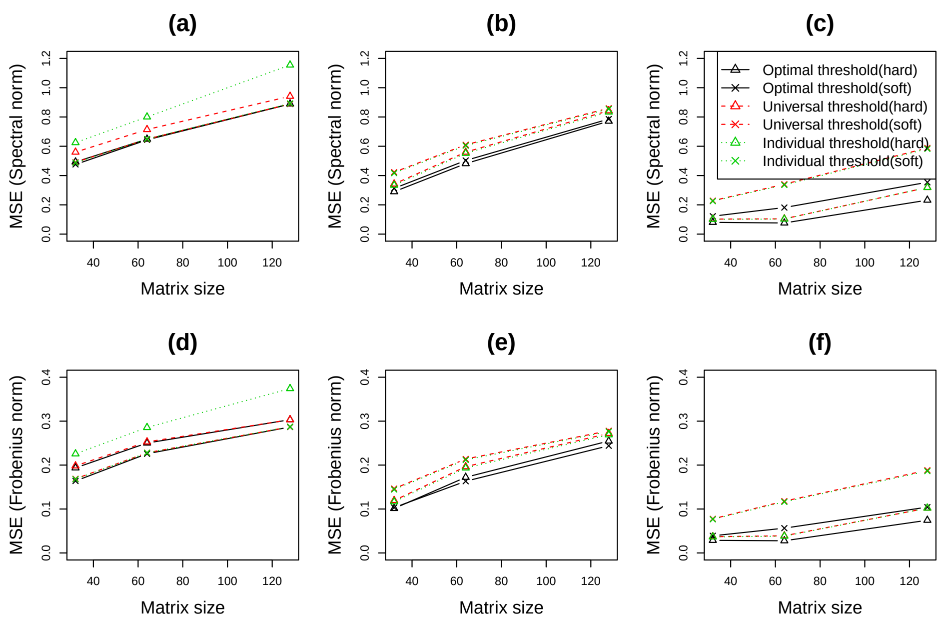

FIG 4. *The plots of MSEs multiplying by d or d^2 against matrix size d for the proposed density estimator with hard and soft threshold rules for $n = 100, 500, 2000$. (a)-(c) are plots of d^2 times of MSEs based on the spectral norm for $n = 100, 500, 2000$, respectively, and (d)-(f) are plots of d times of MSEs based on the Frobenius norm for $n = 100, 500, 2000$, respectively.*

TABLE 1

MSEs based on spectral and Frobenius norms of the density estimator defined by (2.8) and (2.9) and its corresponding density matrix estimator without thresholding, and threshold values used for $d = 32, 64, 128,$ and $n = 100, 200, 500, 1000, 2000$.

		MSE (Spectral norm) $\times 10^4$							Threshold value (ϖ) $\times 10^2$		
		Without threshold	Optimal threshold		Universal threshold		Individual threshold		Universal	Optimal	
d	n	Density estimator	Hard	Soft	Hard	Soft	Hard	Soft	Universal	Hard	Soft
32	100	348.544	4.816	4.648	5.468	4.790	6.104	4.762	24.782	15.180	0.619
	200	175.034	4.449	4.257	5.043	4.708	5.293	4.667	17.524	7.739	0.562
	500	70.069	2.831	3.054	3.344	4.130	3.260	4.071	11.083	2.397	0.373
	1000	35.028	1.537	1.974	1.875	3.201	1.875	3.155	7.837	1.099	0.212
	2000	17.307	0.785	1.195	1.001	2.230	0.989	2.200	5.541	0.551	0.116
64	100	368.842	1.583	1.572	1.744	1.583	1.954	1.586	27.148	16.660	0.395
	200	183.050	1.565	1.534	1.669	1.575	1.833	1.571	19.196	9.252	0.376
	500	73.399	1.175	1.228	1.367	1.490	1.347	1.476	12.141	2.900	0.307
	1000	36.692	0.566	0.807	0.747	1.249	0.722	1.233	8.585	1.308	0.177
	2000	18.402	0.186	0.443	0.255	0.832	0.251	0.820	6.070	0.657	0.061
128	100	381.032	0.543	0.542	0.574	0.543	0.705	0.545	29.323	17.500	0.237
	200	190.113	0.541	0.539	0.570	0.542	0.594	0.542	20.734	10.246	0.235
	500	75.824	0.471	0.480	0.514	0.525	0.509	0.522	13.114	3.547	0.213
	1000	38.010	0.309	0.350	0.355	0.470	0.354	0.466	9.273	1.613	0.146
	2000	18.907	0.142	0.216	0.194	0.359	0.194	0.356	6.557	0.725	0.080
		MSE (Frobenius norm) $\times 10^3$							Threshold value (ϖ) $\times 10^2$		
		Without threshold	Optimal threshold		Universal threshold		Individual threshold		Universal	Optimal	
d	n	Density estimator	Hard	Soft	Hard	Soft	Hard	Soft	Universal	Hard	Soft
32	100	317.873	6.052	5.119	6.195	5.274	7.050	5.246	24.782	11.004	9.936
	200	159.679	5.217	4.629	5.616	5.187	5.874	5.143	17.524	5.681	3.771
	500	63.823	3.165	3.229	3.732	4.575	3.642	4.512	11.083	2.286	0.954
	1000	31.856	1.722	2.053	2.119	3.540	2.119	3.492	7.837	1.100	0.401
	2000	15.967	0.894	1.219	1.155	2.424	1.141	2.394	5.541	0.546	0.182
64	100	641.437	3.909	3.528	3.951	3.563	4.463	3.562	27.148	13.719	13.234
	200	319.720	3.706	3.401	3.755	3.548	4.082	3.536	19.196	7.042	5.515
	500	127.958	2.691	2.551	3.069	3.342	3.023	3.309	12.141	2.800	1.275
	1000	63.845	1.335	1.628	1.765	2.791	1.717	2.756	8.585	1.277	0.548
	2000	31.952	0.433	0.882	0.610	1.842	0.596	1.817	6.070	0.647	0.258
128	100	1283.182	2.370	2.240	2.370	2.242	2.924	2.245	29.323	15.989	16.128
	200	639.556	2.349	2.219	2.354	2.238	2.444	2.238	20.734	8.218	7.799
	500	255.954	1.990	1.906	2.125	2.172	2.102	2.160	13.114	3.355	1.773
	1000	127.714	1.221	1.341	1.463	1.943	1.448	1.924	9.273	1.546	0.729
	2000	63.921	0.581	0.815	0.798	1.471	0.798	1.456	6.557	0.719	0.327

to the third Pauli matrix multiplying by $\sqrt{-1}$, for example, $\boldsymbol{\sigma}_1\boldsymbol{\sigma}_2 = \sqrt{-1}\boldsymbol{\sigma}_3$, $\boldsymbol{\sigma}_2\boldsymbol{\sigma}_3 = \sqrt{-1}\boldsymbol{\sigma}_1$, and $\boldsymbol{\sigma}_3\boldsymbol{\sigma}_1 = \sqrt{-1}\boldsymbol{\sigma}_2$.

For $j = 2, \cdots, p$, consider $\mathbf{B}_j = \boldsymbol{\sigma}_{\ell_1} \otimes \boldsymbol{\sigma}_{\ell_2} \otimes \cdots \otimes \boldsymbol{\sigma}_{\ell_b}$. $tr(\mathbf{B}_j) = tr(\boldsymbol{\sigma}_{\ell_1}) tr(\boldsymbol{\sigma}_{\ell_2}) \cdots tr(\boldsymbol{\sigma}_{\ell_b}) = 0$, and $\mathbf{B}_j$ has eigenvalues ± 1, $\mathbf{B}_j^2 = \mathbf{I}_d$.

For $j, j' = 2, \cdots, p$, $j \neq j'$, $\mathbf{B}_j = \boldsymbol{\sigma}_{\ell_1} \otimes \boldsymbol{\sigma}_{\ell_2} \otimes \cdots \otimes \boldsymbol{\sigma}_{\ell_b}$ and $\mathbf{B}_{j'} = \boldsymbol{\sigma}_{\ell'_1} \otimes \boldsymbol{\sigma}_{\ell'_2} \otimes \cdots \otimes \boldsymbol{\sigma}_{\ell'_b}$,

$$\mathbf{B}_j \mathbf{B}_{j'} = [\boldsymbol{\sigma}_{\ell_1}\boldsymbol{\sigma}_{\ell'_1}] \otimes [\boldsymbol{\sigma}_{\ell_2}\boldsymbol{\sigma}_{\ell'_2}] \otimes \cdots \otimes [\boldsymbol{\sigma}_{\ell_b}\boldsymbol{\sigma}_{\ell'_b}],$$

is equal to a d dimensional Pauli matrix multiplying by $(\sqrt{-1})^b$, which has zero trace. Thus, $tr(\mathbf{B}_j\mathbf{B}_{j'}) = 0$, that is , $\mathbf{B}_j$ and $\mathbf{B}_{j'}$ are orthogonal, and $\mathbf{B}_1, \cdots, \mathbf{B}_p$ form an orthogonal basis. $tr(\boldsymbol{\rho}\mathbf{B}_j/d) = \beta_k tr(\mathbf{B}_j^2)/d = \beta_k$. In particular $\mathbf{B}_1 = \mathbf{I}_d$, and $\beta_1 = tr(\boldsymbol{\rho}\mathbf{B}_1) = tr(\boldsymbol{\rho}) = 1$.

Denote by $\mathbf{Q}_{j\pm}$ the projections onto the eigen-spaces corresponding to eigenvalues ± 1, respectively. Then for $j = 2, \cdots, p$,

$$\mathbf{B}_j = \mathbf{Q}_{j+} - \mathbf{Q}_{j-}, \quad \mathbf{B}_j^2 = \mathbf{Q}_{j+} + \mathbf{Q}_{j-} = \mathbf{I}_d, \quad \mathbf{B}_j \mathbf{Q}_{j\pm} = \pm \mathbf{Q}_{j\pm}^2 = \pm \mathbf{Q}_{j\pm},$$
$$0 = tr(\mathbf{B}_j) = tr(\mathbf{Q}_{j+}) - tr(\mathbf{Q}_{j-}), \qquad d = tr(\mathbf{I}_d) = tr(\mathbf{Q}_{j+}) + tr(\mathbf{Q}_{j-}),$$

and solving the equations we get

$$tr(\mathbf{Q}_{j\pm}) = d/2, \qquad tr(\mathbf{B}_j \mathbf{Q}_{j\pm}) = \pm tr(\mathbf{Q}_{j\pm}) = \pm d/2. \tag{5.1}$$

For $j \neq j'$, $j, j' = 2, \cdots, p$, $\mathbf{B}_j$ and $\mathbf{B}_{j'}$ are orthogonal,

$$0 = tr(\mathbf{B}_{j'} \mathbf{B}_j) = tr(\mathbf{B}_{j'} \mathbf{Q}_{j+}) - tr(\mathbf{B}_{j'} \mathbf{Q}_{j-}),$$

and

$$\mathbf{B}_{j'} \mathbf{Q}_{j+} + \mathbf{B}_{j'} \mathbf{Q}_{j-} = \mathbf{B}_{j'} (\mathbf{Q}_{j+} + \mathbf{Q}_{j-}) = \mathbf{B}_{j'},$$
$$tr(\mathbf{B}_{j'} \mathbf{Q}_{j+}) + tr(\mathbf{B}_{j'} \mathbf{Q}_{j-}) = tr(\mathbf{B}_{j'}) = 0,$$

which imply

$$tr(\mathbf{B}_{j'} \mathbf{Q}_{j\pm}) = 0, \qquad j \neq j', \; j, j' = 2, \cdots, p. \tag{5.2}$$

For a density matrix $\boldsymbol{\rho}$ with representation (2.5) under the Pauli basis (2.3), from (5.1) we have $tr(\mathbf{Q}_{k\pm}) = d/2$ and $tr(\mathbf{B}_k \mathbf{Q}_{k\pm}) = \pm d/2$, and thus (5.3)

$$tr(\boldsymbol{\rho} \mathbf{Q}_{k\pm}) = \frac{1}{d} tr(\mathbf{Q}_{k\pm}) + \sum_{j=2}^{p} \frac{\beta_j}{d} tr(\mathbf{B}_j \mathbf{Q}_{k\pm}) = \frac{1}{2} + \frac{\beta_k}{d} tr(\mathbf{B}_k \mathbf{Q}_{k\pm}) = \frac{1 \pm \beta_k}{2}.$$

Proof of Proposition 2

We perform measurements on each Pauli observable $\mathbf{B}_k$ independently for n quantum systems that are identically prepared under state $\boldsymbol{\rho}$. Denote by $R_{k1} \cdots, R_{kn}$ the n measurement outcomes for measuring $\mathbf{B}_k$, $k = 2, \cdots, p$.

$$N_k = (R_{k1} + \cdots + R_{kn})/n, \tag{5.4}$$

$R_{k\ell}$, $k = 2, \cdots, p$, $\ell = 1, \cdots, n$, are independent, and take values ± 1, with distributions given by

$$P(R_{k\ell} = \pm 1) = tr(\boldsymbol{\rho} \mathbf{Q}_{k\pm}), \;\; k = 2, \cdots, p, \;\; \ell = 1, \cdots, n. \tag{5.5}$$

As random variables $R_{k1}, \cdots, R_{kn}$ are i.i.d. and take eigenvalues ± 1, $n(N_k + 1)/2 = \sum_{\ell=1}^{n} (R_{k\ell} + 1)/2$ is equal to the total number of random variables $R_{k1}, \cdots, R_{kn}$ taking eigenvalue 1, and thus $n(N_k + 1)/2$ follows a binomial distribution with n trials and cell probability $P(R_{k1} = 1) = tr(\boldsymbol{\rho} \mathbf{Q}_{k+})$. From (5.4)-(5.5) and Proposition 1 we have for $k = 2, \cdots, p$,

$$tr(\boldsymbol{\rho} \mathbf{Q}_{k+}) = \frac{1 + \beta_k}{2}, \qquad E(N_k) = E(R_{k1}) = tr(\boldsymbol{\rho} \mathbf{B}_k) = \beta_k tr(\mathbf{B}_k^2)/d = \beta_k,$$
$$Var(N_k) = \frac{1 - \beta_k^2}{n}.$$

5.2. *Proof of Theorem 1: Upper bound.*

LEMMA 1. *If β_j satisfy sparsity condition (2.7), then for any a,*

$$\sum_{j=2}^{p} |\beta_j| 1(|\beta_j| \le a\varpi) \le a^{1-q}\pi_n(p)\varpi^{1-q},$$

$$\sum_{j=2}^{p} 1(|\beta_j| \ge a\varpi) \le a^{-q}\pi_n(p)\varpi^{-q}.$$

Proof. Simple algebraic manipulation shows

$$\sum_{j=2}^{p} |\beta_j| 1(|\beta_j| \le a\varpi) \le (a\varpi)^{1-q} \sum_{j=2}^{p} |\beta_j|^q 1(|\beta_j| \le a\varpi)$$
$$\le a^{1-q}\pi_n(p)\varpi^{1-q},$$

and

$$\sum_{j=2}^{p} 1(|\beta_j| \ge a\varpi) \le \sum_{j=2}^{p} [|\beta_j|/(a\varpi)]^q 1(|\beta_j| \ge a\varpi)$$
$$\le (a\varpi)^{-q} \sum_{j=2}^{p} |\beta_j|^q \le a^{-q}\pi_n(p)\varpi^{-q}.$$

LEMMA 2. *With $\varpi = \hbar n^{-1/2}\sqrt{2\log p}$ for some positive constant $\hbar$, we have for any $a \ne 1$,*

$$P(N_j - \beta_j \le -|a-1|\varpi) \le 2\,p^{-\hbar^2|a-1|^2}, \qquad P(N_j - \beta_j \ge |a-1|\varpi) \le 2\,p^{-\hbar^2|a-1|^2}.$$

Proof. From Proposition 2 and (5.4)-(5.5) we have that N_j is the average of $R_{j1}, \cdots, R_{jn}$, which are i.i.d. random variables taking values ± 1, $P(R_{j1} = \pm 1) = (1 \pm \beta_j)/2$, $E(R_{j1}) = \beta_j$ and $Var(R_{j1}) = 1 - \beta_j^2$. Applying Bernstein's inequality we obtain for any $x > 0$,

$$P(|N_j - \beta_j| \ge x) \le 2\exp\left(-\frac{nx^2}{2(1 - \beta_j^2 + x/3)}\right) \le 2\exp\left(-\frac{nx^2}{2}\right).$$

Both $P(N_j - \beta_j \le -|a-1|\varpi)$ and $P(N_j - \beta_j \ge |a-1|\varpi)$ are less than $P(|N_j - \beta_j| \ge |a-1|\varpi)$, which is bounded by

$$2\exp\left(-\frac{n|a-1|^2\varpi^2}{2}\right) = 2\exp\left(-\hbar^2|a-1|^2\log p\right) = 2\,p^{-\hbar^2|a-1|^2}.$$

LEMMA 3.

$$E\|\hat{\boldsymbol{\rho}} - \boldsymbol{\rho}\|_F^2 = p^{-1/2} \sum_{j=2}^{p} E|\hat{\beta}_j - \beta_j|^2, \tag{5.6}$$

$$p^{1/2} E\|\hat{\boldsymbol{\rho}} - \boldsymbol{\rho}\|_2 \le \sum_{j=2}^{p} E|\hat{\beta}_j - \beta_j|, \tag{5.7}$$

$$\begin{aligned} pE\|\hat{\boldsymbol{\rho}} - \boldsymbol{\rho}\|_2^2 \le \sum_{j=2}^{p} E[|\hat{\beta}_j - \beta_j|^2] + \left\{ \sum_{j=2}^{p} E[|\hat{\beta}_j - \beta_j|] \right\}^2 \\ - \sum_{j=2}^{p} \{E(|\hat{\beta}_j - \beta_j|)\}^2. \end{aligned} \tag{5.8}$$

Proof. Since Pauli matrices $\mathbf{B}_j$ are orthogonal under the Frobenius norm, with $\|\mathbf{B}_j\|_F = d^{1/2}$, and $\|\mathbf{B}_j\|_2 = 1$, we have

$$\begin{aligned} \|\hat{\boldsymbol{\rho}} - \boldsymbol{\rho}\|_F^2 &= \| \sum_{j=2}^{p} (\hat{\beta}_j - \beta_j)\mathbf{B}_j\|_F^2 / d^2 = \sum_{j=2}^{p} |\hat{\beta}_j - \beta_j|^2 \|\mathbf{B}_j\|_F^2 / d^2 \\ &= \sum_{j=2}^{p} |\hat{\beta}_j - \beta_j|^2 / d, \end{aligned} \tag{5.9}$$

$$\begin{aligned} p^{1/2}\|\hat{\boldsymbol{\rho}} - \boldsymbol{\rho}\|_2 &= \| \sum_{j=2}^{p} (\hat{\beta}_j - \beta_j)\mathbf{B}_j\|_2 \le \sum_{j=2}^{p} |\hat{\beta}_j - \beta_j| \|\mathbf{B}_j\|_2 \\ &= \sum_{j=2}^{p} |\hat{\beta}_j - \beta_j|, \end{aligned} \tag{5.10}$$

$$\begin{aligned} p\|\hat{\boldsymbol{\rho}} - \boldsymbol{\rho}\|_2^2 &= \| \sum_{j=2}^{p} (\hat{\beta}_j - \beta_j)\mathbf{B}_j\|_2^2 \\ &\le \sum_{j=2}^{p} |\hat{\beta}_j - \beta_j|^2 \|\mathbf{B}_j\|_2^2 + 2\sum_{i<j}^{p} |(\hat{\beta}_i - \beta_i)(\hat{\beta}_j - \beta_j)| \|\mathbf{B}_i \mathbf{B}_j\|_2 \\ &\le \sum_{j=2}^{p} |\hat{\beta}_j - \beta_j|^2 \|\mathbf{B}_j\|^2 + 2\sum_{i<j}^{p} |(\hat{\beta}_i - \beta_i)(\hat{\beta}_j - \beta_j)| \|\mathbf{B}_i\|_2 \|\mathbf{B}_j\|_2 \\ &= \sum_{j=2}^{p} |\hat{\beta}_j - \beta_j|^2 + 2\sum_{i<j}^{p} |(\hat{\beta}_i - \beta_i)(\hat{\beta}_j - \beta_j)|. \end{aligned} \tag{5.11}$$

As $N_2, \cdots, N_p$ are independent, $\hat{\beta}_2, \cdots, \hat{\beta}_p$ are independent. Thus, from

(5.9)-(5.11) we obtain (5.6)-(5.7), and

$$
\begin{aligned}
pE\|\hat{\boldsymbol{\rho}}-\boldsymbol{\rho}\|_2^2 &\le \sum_{j=2}^{p} E|\hat{\beta}_j-\beta_j|^2 + 2\sum_{i<j}^{p} E|(\hat{\beta}_i-\beta_i)\,(\hat{\beta}_j-\beta_j)| \\
&= \sum_{j=2}^{p} E|\hat{\beta}_j-\beta_j|^2 + 2\sum_{i<j}^{p} E|\hat{\beta}_i-\beta_i|\,E|\hat{\beta}_j-\beta_j| \\
&= \sum_{j=2}^{p} E[|\hat{\beta}_j-\beta_j|^2] + \left\{\sum_{j=2}^{p} E[|\hat{\beta}_j-\beta_j|]\right\}^2 - \sum_{j=2}^{p}\{E(|\hat{\beta}_j-\beta_j|)\}^2.
\end{aligned}
$$

LEMMA 4.

$$
\sum_{j=2}^{p} E|\hat{\beta}_j-\beta_j| \le C_1\pi_n(d)\varpi^{1-q}, \tag{5.12}
$$

$$
\sum_{j=2}^{p} [E|\hat{\beta}_j-\beta_j|]^2 \le \sum_{j=2}^{p} E[|\hat{\beta}_j-\beta_j|^2] \le C_2\pi_n(d)\varpi^{2-q}. \tag{5.13}
$$

Proof. Using (2.8) we have

$$
\begin{aligned}
&E|\hat{\beta}_j-\beta_j| \le E\left[(|N_j-\beta_j|+\varpi)1(|N_j|\ge\varpi)\right] + |\beta_j|P(|N_j|\le\varpi) \\
&\le \left[E|N_j-\beta_j|^2 P(|N_j|\ge\varpi)\right]^{1/2} + \varpi P(|N_j|\ge\varpi) + |\beta_j|P(|N_j|\le\varpi) \\
&\le \left[n^{-1}(1-\beta_j^2)P(|N_j|\ge\varpi)\right]^{1/2} + \varpi P(|N_j|\ge\varpi) + |\beta_j|P(|N_j|\le\varpi) \\
&\le 2\varpi\left[P(|N_j|\ge\varpi)\right]^{1/2} + |\beta_j|P(|N_j|\le\varpi) \\
&= 2\varpi\left[P(|N_j|\ge\varpi)\right]^{1/2}\{1(|\beta_j|>a_1\varpi)+1(|\beta_j|\le a_1\varpi)\} \\
&\qquad + |\beta_j|P(|N_j|\le\varpi)\{1(|\beta_j|>a_2\varpi)+1(|\beta_j|\le a_2\varpi)\} \\
&\le 2\varpi 1(|\beta_j|>a_1\varpi) + 2\varpi\left[P(|N_j|\ge\varpi)\right]^{1/2}1(|\beta_j|\le a_1\varpi) \\
&\qquad + P(|N_j|\le\varpi)1(|\beta_j|>a_2\varpi) + |\beta_j|1(|\beta_j|\le a_2\varpi),
\end{aligned}
$$

where a_1 and a_2 are two constants satisfying $a_1 < 1 < a_2$ whose values will be chosen later, and

$$
\begin{aligned}
\sum_{j=2}^{p} E|\hat{\beta}_j-\beta_j| &\le 2\varpi\sum_{j=2}^{p} 1(|\beta_j|>a_1\varpi) \\
&\quad + 2\varpi\sum_{j=2}^{p}\left[P(|N_j|\ge\varpi)\right]^{1/2}1(|\beta_j|\le a_1\varpi) \\
&\quad + \sum_{j=2}^{p} P(|N_j|\le\varpi)1(|\beta_j|>a_2\varpi) + \sum_{j=2}^{p}|\beta_j|1(|\beta_j|\le a_2\varpi).
\end{aligned} \tag{5.14}
$$

Similarly,

$$
\begin{aligned}
&[E(|\hat{\beta}_j - \beta_j|)]^2 \le E[|\hat{\beta}_j - \beta_j|^2] \\
&\le E[2(|N_j - \beta_j|^2 + \varpi^2)1(|N_j| \ge \varpi)] + |\beta_j|^2 P(|N_j| \le \varpi) \\
&\le 2[E|N_j - \beta_j|^4 P(|N_j| \ge \varpi)]^{1/2} + 2\varpi^2 P(|N_j| \ge \varpi) + |\beta_j|^2 P(|N_j| \le \varpi) \\
&\le c\varpi^2 [P(|N_j| \ge \varpi)]^{1/2} + |\beta_j|^2 P(|N_j| \le \varpi)\} \\
&= c\varpi^2 [P(|N_j| \ge \varpi)]^{1/2} \{1(|\beta_j| > a_1\varpi) + 1(|\beta_j| \le a_1\varpi)\} \\
&\qquad + |\beta_j|^2 P(|N_j| \le \varpi)\}[1(|\beta_j| > a_2\varpi) + 1(|\beta_j| \le a_2\varpi)] \\
&\le c\varpi^2 1(|\beta_j| > a_1\varpi) + c\varpi^2 \left[P(|N_j| \ge \varpi)\right]^{1/2} 1(|\beta_j| \le a_1\varpi) \\
&+ P(|N_j| \le \varpi) 1(|\beta_j| > a_2\varpi) + |\beta_j|^2 1(|\beta_j| \le a_2\varpi),
\end{aligned}
$$

and

$$
\begin{aligned}
\sum_{j=2}^{p} E[|\hat{\beta}_j - \beta_j|^2] \le c\varpi^2 \sum_{j=2}^{p} 1(|\beta_j| > a_1\varpi) \\
+ c\varpi^2 \sum_{j=2}^{p} \left[P(|N_j| \ge \varpi)\right]^{1/2} 1(|\beta_j| \le a_1\varpi) \\
+ \sum_{j=2}^{p} P(|N_j| \le \varpi) 1(|\beta_j| > a_2\varpi) + \sum_{j=2}^{p} |\beta_j|^2 1(|\beta_j| \le a_2\varpi).
\end{aligned}
\tag{5.15}
$$

By Lemma 1, we have

$$
\sum_{j=2} |\beta_j| 1(|\beta_j| \le a_2\varpi) \le a_2^{1-q} \pi_n(d) \varpi^{1-q}, \tag{5.16}
$$

$$
\begin{aligned}
&\sum_{j=2} |\beta_j|^2 1(|\beta_j| \le a_2\varpi) \\
&\le (a_2\varpi)^{2-q} \sum_{j=2} |\beta_j|^q 1(|\beta_j| \le a_2\varpi) \le a_2^{2-q} \pi_n(d) \varpi^{2-q},
\end{aligned}
\tag{5.17}
$$

$$
\varpi \sum_{j=2}^{p} 1(|\beta_j| \ge a_1 t) \le \pi_n(d) \varpi^{1-q}. \tag{5.18}
$$

On the other hand,

$$
\begin{aligned}
(5.19)\sum_{j=2}^{p} & P(|N_j| \le \varpi) 1(|\beta_j| > a_2\varpi) \\
& \le \sum_{j} P(-\varpi - \beta_j \le N_j - \beta_j \le \varpi - \beta_j) 1(|\beta_j| > a_2\varpi) \\
& \le \sum_{j=2}^{p} [P(N_j - \beta_j \le -|a_2 - 1|\varpi) + P(N_j - \beta_j \ge |a_2 - 1|\varpi)] \\
& \le 4\, p^{1-\hbar^2|a_2-1|^2} = 4\, p^{-1-(2-q)/(2c_0)} \le 4p^{-1} n^{-(q-2)/2} = o(\pi_n(d)\varpi^{2-q}),
\end{aligned}
$$

where the third inequality is from Lemma 2, the first equality is due the fact that we take $a_2 = 1 + \{2 + (2-q)/(2c_0)\}^{1/2}/\hbar$ so that $\hbar^2(1-a_2)^2 = 2 + (2-q)/(2c_0)$, and c_0 is the constant in Assumption $p \ge n^{c_0}$. Finally we can show

$$
\begin{aligned}
(5.20) \quad \varpi & \sum_{j=2}^{p} [P(|N_j| \ge \varpi)]^{1/2} 1(|\beta_j| \le a_1\varpi) \\
& \le \varpi \sum_{j=2}^{p} [P(N_j - \beta_j \le -\varpi - \beta_j) \\
& \qquad + P(N_j - \beta_j \ge \varpi - \beta_j)]^{1/2} 1(|\beta_j| \le a_1\varpi) \\
& \le \varpi \sum_{j=2}^{p} [P(N_j - \beta_j \le -|1 - a_1|\varpi) + P(N_j - \beta_j \ge |1 - a_1|\varpi)]^{1/2} \\
& \le 2\varpi p^{1-\hbar^2(1-a_1)^2/2} = 2\varpi p^{-1} = o(\pi_n(d)\varpi^{1-q}),
\end{aligned}
$$

where the third inequality is from Lemma 2, and the first equality is due to the fact that we take $a_1 = 1 - 2/\hbar$ so that $\hbar^2(1-a_1)^2 = 4$. Plugging (5.16)-(5.20) into (5.15) and (5.15) we prove the lemma.

Proof of Theorem 1. Combining Lemma 4 and (5.6)-(5.7) in Lemma 3 we easily obtain

$$
E[\|\hat{\boldsymbol{\rho}} - \boldsymbol{\rho}\|_2] \le C_1 \frac{\pi_n(d)}{p^{1/2}} \left(\frac{\log p}{n}\right)^{\frac{1-q}{2}},
$$

$$
E[\|\hat{\boldsymbol{\rho}} - \boldsymbol{\rho}\|_F^2] \le C_0 \pi_n(d) \frac{1}{d} \left(\frac{\log p}{n}\right)^{1-q/2}.
$$

Using Lemma 4 and (5.9) in Lemma 3 we conclude

$$E[\|\hat{\boldsymbol{\rho}}-\boldsymbol{\rho}\|_2^2] \le C_2\left[\pi_n^2(d)\,\frac{1}{p}\left(\frac{\log p}{n}\right)^{1-q} + \pi_n(d)\left(\frac{\log p}{n}\right)^{1-q/2}\right] \le C\frac{\pi_n^2(d)}{d^2}\left(\frac{\log p}{n}\right)^{1-q}, \tag{5.21}$$

where the last inequality is due to the fact that the first term on the right hand side of (5.21) dominates its second term.

5.3. *Proofs of Theorems 2 and 3: Lower bound.* **Proof of Theorem 2** for the lower bound under the spectral norm.

We first define a subset of the parameter space Θ. It will be shown later that the risk upper bound under the spectral norm is sharp up a constant factor, when the parameter space is sufficiently sparse. Consider a subset of the Pauli basis, $\{\sigma_{l_1}\otimes\sigma_{l_2}\otimes\cdots\otimes\sigma_{l_b}\}$, where $\sigma_{l_1}=\sigma_0$ or σ_3. Its cardinality is $d=2^b=p^{1/2}$. Denote each element of the subset by $\mathbf{B}_j$, $j=1,2,\dots,d$, and let $\mathbf{B}_1=\mathbf{I}_d$. We will define each element of Θ as a linear combination of $\mathbf{B}_j$. Let $\gamma_j\in\{0,1\}$, $j\in\{1,2,\dots,d\}$, and denote $\eta=\sum_j\gamma_j=\|\gamma\|_0$. The value of η is either 0 or K, where K is the largest integer less than or equal to $\pi_n(d)/\left(\frac{\log p}{n}\right)^{q/2}$. By Assumption (3.3) we have

$$1\le K = O\left(d^v\right), \text{ with } v<1/2. \tag{5.22}$$

Let $\varepsilon^2=(1-2v)/4$ and set $a=\varepsilon\sqrt{\frac{\log p}{n}}$. Now we are ready to define Θ,

$$\Theta=\left\{\rho(\gamma):\rho(\gamma)=\frac{\mathbf{I}_d}{d}+a\sum_{j=2}^{d}\gamma_j\frac{\mathbf{B}_j}{d},\ \text{and } \eta=0 \text{ or } K\right\}. \tag{5.23}$$

Note that Θ is a subset of the parameter space, since

$$\sum_{j=2}^{d}(a\gamma_j)^q\le Ka^q\le\varepsilon^q\pi_n(d)\le\pi_n(d),$$

and its cardinality is $1+\binom{d-1}{K}$.

We need to show that

$$\inf_{\hat{\boldsymbol{\rho}}}\sup_{\Theta}E\|\hat{\boldsymbol{\rho}}-\boldsymbol{\rho}\|_2^2\gtrsim\pi_n^2(d)\,\frac{1}{p}\left(\frac{\log p}{n}\right)^{1-q}.$$

Note that for each element in Θ, its first entry ρ_{11} may take the form $1/d + a\sum_{j=2}^{d}\gamma_j/d = 1/d + (a/d)\eta$. It can be shown that

$$\inf_{\hat{\boldsymbol{\rho}}}\sup_{\Theta} E\left\|\hat{\boldsymbol{\rho}} - \boldsymbol{\rho}\right\|_2^2 \geq \inf_{\hat{\rho}_{11}}\sup_{\Theta} E\left(\hat{\rho}_{11}-\rho_{11}\right)^2 \geq \frac{a^2}{d^2}\inf_{\hat{\eta}}\sup_{\Theta} E\left(\hat{\eta}-\eta\right)^2.$$

It is then enough to show that

$$\inf_{\hat{\eta}}\sup_{\Theta} E\left(\hat{\eta}-\eta\right)^2 \gtrsim K^2, \tag{5.24}$$

which immediately implies

$$\inf_{\hat{\boldsymbol{\rho}}}\sup_{\Theta} E\left\|\hat{\boldsymbol{\rho}} - \boldsymbol{\rho}\right\|_2^2 \gtrsim K^2\frac{a^2}{d^2} \gtrsim \pi_n^2\left(d\right)\frac{1}{p}\left(\frac{\log p}{n}\right)^{1-q}.$$

We prove Equation (5.24) by applying Le Cam's lemma. From observations N_j, $j = 2, \ldots, d$, we define $\tilde{N}_j = n\left(N_j + 1\right)/2$, which is *Binomial* $\left(n, \frac{1+a\gamma_j}{2}\right)$. Let $\mathbb{P}_\gamma$ be the joint distribution of independent random variables $\tilde{N}_2, \tilde{N}_3, \ldots, \tilde{N}_d$. The cardinality of $\{\mathbb{P}_\gamma\}$ is $1 + \binom{d-1}{K}$. For two probability measures $\mathbb{P}$ and $\mathbb{Q}$ with density f and g with respect to any common dominating measure μ, write the total variation affinity $\|\mathbb{P} \wedge \mathbb{Q}\| = \int f \wedge g d\mu$, and the Chi-Square distance $\chi^2\left(\mathbb{P}, \mathbb{Q}\right) = \int \frac{g^2}{f} - 1$. Define

$$\bar{\mathbb{P}} = \binom{d-1}{K}^{-1}\sum_{\|\gamma\|_0 = K}\mathbb{P}_\gamma.$$

The following lemma is a direct consequence of Le Cam's lemma (cf. Le Cam (1973) and Yu (1997)).

Lemma 5. *Let $\hat{\eta}$ be any estimator of η based on an observation from a distribution in the collection $\{\mathbb{P}_\gamma\}$, then*

$$\inf_{\hat{k}}\sup_{\Theta} E\left(\hat{\eta}-\eta\right)^2 \geq \frac{1}{4}\left\|\mathbb{P}_{\mathbf{0}} \wedge \bar{\mathbb{P}}\right\|^2 \cdot K^2.$$

We will show that there is a constant $c > 0$ such that

$$\left\|\mathbb{P}_{\mathbf{0}} \wedge \bar{\mathbb{P}}\right\| \geq C, \tag{5.25}$$

which, together with Lemma 5, immediately imply Equation (5.24).

LEMMA 6. *Under conditions (5.22) and (5.23), we have*

$$\inf_{\hat{\boldsymbol{\rho}}} \sup_{\Theta} E\left(\hat{\eta}-\eta\right)^2 \gtrsim K^2,$$

which implies

$$\inf_{\hat{\boldsymbol{\rho}}} \sup_{\Theta} E\left\|\hat{\boldsymbol{\rho}}-\boldsymbol{\rho}\right\|_2^2 \gtrsim \pi_n^2\left(d\right)\frac{1}{p}\left(\frac{\log p}{n}\right)^{1-q}.$$

Proof. It is enough to show that

$$\chi^2\left(\mathbb{P}_{\mathbf{0}},\bar{\mathbb{P}}\right) \to 0,$$

which implies $\left\|\mathbb{P}_{\mathbf{0}}-\bar{\mathbb{P}}\right\|_{TV} \to 0$, then we have $\left\|\mathbb{P}_{\mathbf{0}}\wedge\bar{\mathbb{P}}\right\| \to 1$. Let $J\left(\gamma,\gamma^{'}\right)$ denote the number of overlapping nonzero coordinates between γ and $\gamma^{'}$. Note that

$$\begin{aligned}
\chi^2\left(\mathbb{P}_{\mathbf{0}},\bar{\mathbb{P}}\right) &= \int\frac{\left(d\bar{\mathbb{P}}\right)^2}{d\mathbb{P}_{\mathbf{0}}}-1\\
&= \binom{d-1}{K}^{-2}\sum_{0\le j\le K}\sum_{J\left(\gamma,\gamma^{'}\right)=j}\left(\int\frac{d\mathbb{P}_{\gamma}\cdot d\mathbb{P}_{\gamma^{'}}}{d\mathbb{P}_{\mathbf{0}}}-1\right).
\end{aligned}$$

When $J\left(\gamma,\gamma^{'}\right)=j$, we have

$$\begin{aligned}
&\int\frac{d\mathbb{P}_{\gamma}\cdot d\mathbb{P}_{\gamma^{'}}}{d\mathbb{P}_{\mathbf{0}}}\\
&= \left(\sum_{l=0}^{n}\left[\binom{n}{l}\frac{1}{2^l}\frac{1}{2^{n-l}}\cdot\left(1+a\right)^{2l}\left(1-a\right)^{2n-2l}\right]\right)^j\\
&= \left(\sum_{l=0}^{n}\left[\binom{n}{l}\left(\frac{\left(1+a\right)^2}{2}\right)^l\left(\frac{\left(1-a\right)^2}{2}\right)^{n-l}\right]\right)^j\\
&= \left(\frac{\left(1+a\right)^2}{2}+\frac{\left(1-a\right)^2}{2}\right)^{nj}\\
&= \left(1+a^2\right)^{nj} \le \exp\left(na^2 j\right),
\end{aligned}$$

which implies

$$\begin{aligned}
\chi^2\left(\mathbb{P}_{\mathbf{0}}, \bar{\mathbb{P}}\right) &\leq \binom{d-1}{K}^{-2} \sum_{0 \leq j \leq K} \sum_{J(\gamma,\gamma')=j} \left(\exp\left(na^2 j\right) - 1\right) \\
&\leq \binom{d-1}{K}^{-2} \sum_{1 \leq j \leq K} \sum_{J(\gamma,\gamma')=j} \exp\left(na^2 j\right) \\
&= \sum_{1 \leq j \leq K} \frac{\binom{K}{j}\binom{d-1-K}{K-j}}{\binom{d-1}{K}} d^{2\varepsilon^2 j}.
\end{aligned}$$

Since

$$\begin{aligned}
\frac{\binom{K}{j}\binom{d-1-K}{K-j}}{\binom{d-1}{K}} &= \frac{\left[K \cdot \ldots \cdot (K-j+1)\right]^2 \cdot (d-1-K) \cdot \ldots \cdot (d-2K+j)}{j! \cdot (d-1) \cdot \ldots \cdot (d-K)} \\
&\leq \frac{K^{2j}\left(d-1-K\right)^{K-j}}{\left(d-K\right)^K} \leq \left(\frac{K^2}{d-K}\right)^j,
\end{aligned}$$

and $\varepsilon^2 = (1-2v)/4$, we then have

$$\begin{aligned}
\chi^2\left(\mathbb{P}_{\mathbf{0}}, \bar{\mathbb{P}}\right) &\leq \sum_{1 \leq j \leq K} \left[\frac{K^2}{d-K} d^{2\varepsilon^2}\right]^j \\
&\leq \sum_{1 \leq j \leq K} \left[\frac{d^{2v+(1-2v)/2}}{d-K}\right]^j \to 0.
\end{aligned}$$

Proof of Theorem 3 for the lower bound under the Frobenius norm. Recall that Θ is the collection of density matrices such that

$$\boldsymbol{\rho} = \frac{1}{d}\left(\mathbf{I}_d + \sum_{j=2}^{p} \beta_j \mathbf{B}_j\right),$$

where

$$\sum_{j=2}^{p} |\beta_j|^q \leq \pi_n(p).$$

Apply Assouad's lemma we show below that

$$\inf_{\check{\boldsymbol{\rho}}} \sup_{\boldsymbol{\rho} \in \Theta} E[\|\check{\boldsymbol{\rho}} - \boldsymbol{\rho}\|_F^2] \geq C\, \pi_n(p) \left(\frac{\log p}{n}\right)^{1-q/2},$$

where $\check{\boldsymbol{\rho}}$ denotes any estimator of $\boldsymbol{\rho}$ based on measurement data $N_2, \cdots, N_p$, and C is a constant free of n and p.

To this end, it suffices to construct a collection of $M+1$ density matrices $\{\boldsymbol{\rho}_0 = \mathbf{I}_d/d, \boldsymbol{\rho}_1, \cdots, \boldsymbol{\rho}_M\} \subset \Theta$ such that (i) for any distinct k and k_0,

$$\|\boldsymbol{\rho}_k - \boldsymbol{\rho}_{k_0}\|_F^2 \geq C_1\, \pi_n(p) \frac{1}{d} \left(\frac{\log p}{n}\right)^{1-q/2},$$

where C_1 is a constant; (ii) there exists a constant $0 < C_2 < 1/8$ such that

$$\frac{1}{M}\sum_{k=1}^{M} D_{KL}(P_{\boldsymbol{\rho}_k}, P_{\boldsymbol{\rho}_0}) \leq C_2 \log M,$$

where D_{KL} denotes the Kullback-Leibler divergence.

By the Gilbert-Varshamov bound (cf. Nielsen and Chuang (2000)) we have that for any $h < p/8$, there exist M binary vectors $\boldsymbol{\gamma}_k = (\gamma_{k2}, \cdots, \gamma_{kp})' \in \{0,1\}^{p-1}$, $k = 1, \cdots, M$, such that (i) $\|\boldsymbol{\gamma}_k\|_1 = \sum_{j=2}^p |\gamma_{kj}| = h$, (ii) $\|\boldsymbol{\gamma}_k - \boldsymbol{\gamma}_{k_0}\|_1 = \sum_{j=2}^p |\gamma_{kj} - \gamma_{k_0 j}| \geq h/2$, and (iii) $\log M > 0.233\, h\, \log(p/h)$. Let

$$\boldsymbol{\rho}_k = \frac{1}{d}\left(\mathbf{I}_d + \epsilon \sum_{j=2}^{p} \gamma_{kj} \mathbf{B}_j\right),$$

where

$$\epsilon = C_3 \left(\frac{\pi_n(p)}{h}\right)^{1/q}.$$

Since $\sum_{j=2}^p |\epsilon \gamma_{kj}|^q = \epsilon^q h = C_3 \pi_n(p)$, $\boldsymbol{\rho}_k \in \Theta$ whenever $C_3 \leq 1$. Moreover,

$$\|\boldsymbol{\rho}_k - \boldsymbol{\rho}_{k_0}\|_F^2 = \epsilon^2 \|\boldsymbol{\gamma}_k - \boldsymbol{\gamma}_{k_0}\|_1 \geq \frac{\epsilon^2\, h}{4}.$$

On the other hand,

$$\begin{aligned} D_{KL}(P_{\boldsymbol{\rho}_k}, P_{\boldsymbol{\rho}_0}) &= h D_{KL}\left(Bin\left(n, \frac{1+\epsilon}{2}\right), Bin\left(n, \frac{1}{2}\right)\right) \\ &= h\, n\, \frac{\epsilon}{2} \log \frac{1/2+\epsilon}{1/2-\epsilon} \leq C_4\, h\, n\, \epsilon^2. \end{aligned}$$

Now the lower bound can be established by taking

$$h = \pi_n(p) \left(\frac{\log p}{n}\right)^{-q/2},$$

and then

$$\epsilon = C_3 \left(\frac{\log p}{n}\right)^{1/2}, \qquad \frac{\epsilon^2 h}{4} = C_3 \, \pi_n(p) \left(\frac{\log p}{n}\right)^{1-q/2},$$

$$\begin{aligned} &C_4 \, h \, n \, \epsilon^2 = C_4 \, h \, \log p, \qquad h \, \log(p/h) = h \, \log p - h \, \log h, \\ &\log h \sim \log \pi_n(p) + \frac{q}{2} \log n - \frac{q}{2} \log\log p, \end{aligned}$$

which are allowed by the assumption $\log \pi_n(p) + \frac{q}{2} \log n < v' \log p$ for $v' < 1$.

References.

[1] Alquier, P., Butucea, C., Hebiri, M. and Meziani, K. (2013). Rank penalized estimation of a quantum system. Phys. Rev. A. 88 032133.

[2] Artiles, L., Gill, R., and Guţă, M. (2005). An invitation to quantum tomography. J. Roy. Statist. Soc. **67**, 109-134.

[3] Aspuru-Guzik, A., Dutoi, A. D., Love, P. J. and Head-Gordon, M. (2005). Simulated quantum computation of molecular energies. Science 309, 1704-1707.

[4] Aubry, J. M., Butucea, C. and Meziani, K. (2009). State estimation in quantum homodyne tomography with noisy data. Inverse Problem 25, 015003(22pp).

[5] Barndorff-Nielsen, O. E., Gill, R. and Jupp, P. E. (2003). On quantum statistical inference (with discussion). J. R. Statist. Soc. B **65**, 775-816.

[6] Benenti, G., Casati, G. and Strini, G. (2004). Principles of Quantum Computation and Information Volume I: Basic Concepts. World Scientific Publishing Company, Incorporated. Singapore.

[7] Benenti, G., Casati, G. and Strini, G. (2007). Principles of Quantum Computation And Information Volume II: Basic Tools And Special Topics. World Scientific Publishing Company, Incorporated. Singapore.

[8] Britton, J. W., Sawyer, B.C., Keith, A., Wang, C.-C.J., Freericks, J. K., Uys, H., Biercuk, M. J. and Bollinger, J. J. (2012). Engineered 2D Ising interactions on a trapped-ion quantum simulator with hundreds of spins. Nature 484, 489-492.

[9] Brumfiel, G. (2012). Simulation: Quantum leaps. Nature 491, 322-324.

[10] Bunea, F., She, Y. and Wegkamp, M. (2011). Optimal selection of reduced rank estimators of high-dimensional matrices. Ann. Statist. **39**, 1282-1309.

[11] Bunea, F., She, Y. and Wegkamp, M. (2012). Joint variable and rank selection for parsimonious estimation of high dimensional matrices. Ann. Statist. 40, 2359-2388.

[12] Butucea, C., Guţă, M. and Artiles, L. (2007). Minimax and adaptive estimation of the Wigner function in quantum homodyne tomography with noisy data. Ann. Statist. **35**, 465-494.

[13] Cai, T. T. and Zhang, A. (2015). ROP: Matrix recovery via rank-one projections. Ann. Statist. 43, 102-138.

[14] Cai, T. and Zhou, H. (2012). Optimal rates of convergence for sparse covariance matrix estimation. Ann. Statist. 40, 2389-2420.

[15] Candès, E. J. and Plan, Y. (2009a). Matrix completion with noise. Proceedings of the IEEE 98(6), 925-936.

[16] Candès, E. J. and Plan, Y. (2009b). Tight oracle bounds for low-rank matrix recovery from a minimal number of random measurements. IEEE Transactions on Information Theory 57(4), 2342-2359.

[17] Candès, E. J. and Tao, T. (2009). The power of convex relaxation: Near-optimal matrix completion. IEEE Trans. Inform. Theory 56(5), 2053-2080.

[18] Candès, E. J. and Recht, B. (2008). Exact matrix completion via convex optimization. Found. of Comput. Math. 9, 717-772.

[19] Donoho, D. L. (2006). Compressed sensing. IEEE Transactions on Information Theory 52, 1289-1306.

[20] Gross, D. (2011). Recovering low-rank matrices from few coefficients in any basis. IEEE Transactions on Information Theory 57, 1548-1566.

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

[22] Guţă, M. and Artiles, L. (2007). Minimax estimation of the Wigner function in quantum homodyne tomography with ideal detectors. Mathematical Methods of Statistics 16, 1-15.

[23] Häffner, H., Hänsel, W., Roos, C. F., Benhelm, J., Chek-al-Kar, D., Chwalla, M., Körber, T., Rapol, U.D., Riebe, M., Schmidt, P. O., Becher, C., Gühne, O., Dür, W. and Blatt, R. (2005). Scalable multiparticle entanglement of trapped ions. Nature 438, 643-646.

[24] Holevo, A. S. (1982). Probabilistic and Statistical Aspects of Quantum Theory. North-Holland, Amsterdam.

[25] Johnson, M. W., M. H. S. Amin, S. Gildert, T. Lanting, F. Hamze, N. Dickson, R. Harris, A. J. Berkley, J. Johansson, P. Bunyk, E. M. Chapple, C. Enderud, J. P. Hilton, K. Karimi, E. Ladizinsky, N. Ladizinsky, T. Oh, I. Perminov, C. Rich1, M. C. Thom, E. Tolkacheva, C. J. S. Truncik, S. Uchaikin, J. Wang, B. Wilson and G. Rose (2011). Quantum annealing with manufactured spins. Nature 473, 194-198.

[26] Jones, N. (2013). Computing: The quantum company. Nature 498, 286-288.

[27] Keshavan, R. H., Montanari, A. and Oh, S. (2010). Matrix completion from noisy entries. The Journal of Machine Learning Research 11, 2057-2078.

[28] Klopp, O. (2011). Rank penalized estimators for high-dimensional matrices. Electronic Journal of Statistics 5, 1161-1183.

[29] Klopp, O. (2012). Noisy low-rank matrix completion with general sampling distribution. Manuscript.

[30] Koltchinskii, V. (2011). Von Neumann entropy penalization and low rank matrix estimation. Ann. Statist. 39, 2936-2973.

[31] Koltchinskii, V., Lounici, K. and Tsybakov, A. B. (2011). Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. Ann. Statist. 39, 2302-2329.

[32] Lanyon, B. P., Whitfield, J. D., Gillett, G.G., Goggin, M., E., Almeida, M.P., Kassal, I. and Biamonte, J., D. (2010). Towards quantum chemistry on a quantum computer. Nature Chemistry 2, 106-111.

[33] Le Cam, L. (1973). Convergence of estimates under dimensionality restrictions. Ann. Statist. 1, 38-53.

[34] Liu, Y. K. (2011). Universal low-rank matrix recovery from Pauli measurements. Unpublished manuscript.

[35] Negahban, S. and Wainwright, M. J. (2011). Estimation of (near) low-rank matrices with noise and high-dimensional scaling. Ann. Statist. 39, 1069-1097.

[36] Nielsen, M. and Chuang, I. (2000). Quantum Computation and Quantum Information. Cambridge: Cambridge University Press.

[37] Recht, B., Fazel, M. and Parrilo, P. A. (2010). Guaranteed minimum rank solutions to linear matrix equations via nuclear norm minimization. SIAM Review.

Vol 52, no 3, 471-501.
[38] RECHT, B. (2011). A simpler approach to matrix completion. Journal of Machine Learning Research. Vol 12. pp. 3413–3430.
[39] ROHDE, A. and TSYBAKOV, A. B. (2011). Estimation of high-dimensional low-rank matrices. Ann. Statist. 39, 887-930.
[40] SAKURAI, J. J. and NAPOLITANO, J. (2010). Modern Quantum Mechanics. Addison-Wesley, Reading, Massachusetts. Second edition.
[41] SHANKAR, R. (1994). Principles of Quantum Mechanics. Springer. Second edition.
[42] TAO, M., WANG, Y. and ZHOU, H. H. (2013). Optimal sparse volatility matrix estimation for high Dimensional Itô processes with measurement errors. Ann. Statist. 41, 1816-1864.
[43] WANG, Y. (2011). Quantum Monte Carlo simulation. Ann. Appl. Statist. 5, 669-683.
[44] WANG, Y. (2012). Quantum computation and quantum information. Statistical Science 27, 373-394.
[45] WANG, Y. (2013). Asymptotic equivalence of quantum state tomography and noisy matrix completion. Ann. Statist. 41, 2462-2504.
[46] WANG, Y. and XU, C. (2015). Density matrix estimation in quantum homodyne tomography. To appear in Statistica Sinica.
[47] YU, B. (1997). Assouad, Fano, and Le Cam. In: Pollard, D., Torgersen, E., Yang, G. (Eds.), Festschrift for Lucien Le Cam Research Papers in Probability and Statistics, Springer, New York. pp. 423-435.

ADDRESS OF T. CAI
DEPARTMENT OF STATISTICS
THE WHARTON SCHOOL
UNIVERSITY OF PENNSYLVANIA
PHILADELPHIA, PENNSYLVANIA 19104
USA
E-MAIL: tcai@wharton.upenn.edu

ADDRESS OF D. KIM, Y. WANG AND M. YUAN
DEPARTMENT OF STATISTICS
UNIVERSITY OF WISCONSIN-MADISON
MADISON, WISCONSIN 53706
USA
E-MAIL: kimd@stat.wisc.edu
yzwang@stat.wisc.edu
myuan@stat.wisc.edu

ADDRESS OF H. H. ZHOU
DEPARTMENT OF STATISTICS
YALE UNIVERSITY
NEW HAVEN, CONNECTICUT 06511
USA
E-MAIL: huibin.zhou@yale.edu