The Annals of Statistics
2013, Vol. 41, No. 5, 2462–2504
DOI: 10.1214/13-AOS1156

ASYMPTOTIC EQUIVALENCE OF QUANTUM STATE TOMOGRAPHY AND NOISY MATRIX COMPLETION

BY YAZHEN WANG[1]

University of Wisconsin–Madison

Matrix completion and quantum tomography are two unrelated research areas with great current interest in many modern scientific studies. This paper investigates the statistical relationship between trace regression in matrix completion and quantum state tomography in quantum physics and quantum information science. As quantum state tomography and trace regression share the common goal of recovering an unknown matrix, it is nature to put them in the Le Cam paradigm for statistical comparison. Regarding the two types of matrix inference problems as two statistical experiments, we establish their asymptotic equivalence in terms of deficiency distance. The equivalence study motivates us to introduce a new trace regression model. The asymptotic equivalence provides a sound statistical foundation for applying matrix completion methods to quantum state tomography. We investigate the asymptotic equivalence for sparse density matrices and low rank density matrices and demonstrate that sparsity and low rank are not necessarily helpful for achieving the asymptotic equivalence of quantum state tomography and trace regression. In particular, we show that popular Pauli measurements are bad for establishing the asymptotic equivalence for sparse density matrices and low rank density matrices.

1. Introduction. Compressed sensing and quantum tomography are two disparate scientific fields. The fast developing field of compressed sensing provides innovative data acquisition techniques and supplies efficient accurate reconstruction methods for recovering sparse signals and images from highly undersampled observations [see Donoho (2006)]. Its wide range of applications include signal processing, medical imaging and seismology. The problems to solve in compressed sensing often involve large data sets with complex structures such as data on many variables or features observed over a much smaller number of subjects. As a result, the developed theory of compressed sensing can shed crucial insights on high-dimensional statistics. Matrix completion, a current research focus point in compressed sensing, is to reconstruct a low rank matrix based on

Received June 2012; revised January 2013.

[1]Supported in part by the NSF Grants DMS-10-5635 and DMS-12-65203.

MSC2010 subject classifications. Primary 62B15; secondary 62P35, 62J99, 65F10, 65J20, 81P45, 81P50.

Key words and phrases. Compressed sensing, deficiency distance, density matrix, observable, Pauli matrices, quantum measurement, quantum probability, quantum statistics, trace regression, fine scale trace regression, low rank matrix, sparse matrix.

under-sampled observations. Trace regression is often employed in noisy matrix completion for low rank matrix estimation. Recently several methods were proposed to estimate a low rank matrix by minimizing the squared residual sum plus some penalty. The penalties used include nuclear-norm penalty [Candés and Plan (2009, 2011), Koltchinskii, Lounici and Tsybakov (2011) and Negahban and Wainwright (2011)], rank penalty [Bunea, She and Wegkamp (2011) and Klopp (2011)], the von Neumann entropy penalty [Koltchinskii (2011)], and the Schatten-p quasi-norm penalty [Rohde and Tsybakov (2011)].

Contemporary scientific studies often rely on understanding and manipulating quantum systems. Examples include quantum computation, quantum information and quantum simulation [Nielsen and Chuang (2000) and Wang (2011, 2012)]. The studies particularly frontier research in quantum computation and quantum information stimulate great interest in and urgent demand on quantum tomography. A quantum system is described by its state, and the state is often characterized by a complex matrix on some Hilbert space. The matrix is called density matrix. A density matrix used to characterize a quantum state usually grows exponentially with the size of the quantum system. For the study of a quantum system, it is important but very difficult to know its state. If we do not know in advance the state of the quantum system, we may deduce the quantum state by performing measurements on the quantum system. In statistical terminology, we want to estimate the density matrix based on measurements performed on a large number of quantum systems which are identically prepared in the same quantum state. In the quantum literature, quantum state tomography refers to the reconstruction of the quantum state based on measurements obtained from measuring identically prepared quantum systems.

In this paper, we investigate statistical relationship between quantum state tomography and noisy matrix completion based on trace regression. Trace regression is used to recover an unknown matrix from noisy observations on the trace of the products of the unknown matrix and matrix input variables. Its connection with quantum state tomography is through quantum probability on quantum measurements. Consider a finite-dimensional quantum system with a density matrix. According to the theory of quantum physics, when we measure the quantum system by performing measurements on observables which are Hermitian (or self-adjoint) matrices, the measurement outcomes for each observable are real eigenvalues of the observable, and the probability of observing a particular eigenvalue is equal to the trace of the product of the density matrix and the projection matrix onto the eigen-space corresponding to the eigenvalue, with the expected measurement outcome equal to the trace of the product of the density matrix and the observable. Taking advantage of the connection Gross et al. (2010) has applied matrix completion methods with nuclear norm penalization to quantum state tomography for reconstructing low rank density matrices. As trace regression and quantum state tomography share the common goal of recovering the same matrix parameter, we naturally treat them as two statistical models in the Le Cam paradigm and study

their asymptotic equivalence via Le Cam's deficiency distance. Here equivalence means that each statistical procedure for one model has a corresponding equal-performance statistical procedure for another model. The equivalence study motivates us to introduce a new fine scale trace regression model. We derive bounds on the deficiency distances between trace regression and quantum state tomography with summarized measurement data and between fine scale trace regression and quantum state tomography with individual measurement data, and then under suitable conditions we establish asymptotic equivalence of trace regression and quantum state tomography for both cases. The established asymptotic equivalence provides a sound statistical foundation for applying matrix completion procedures to quantum state tomography under appropriate circumstances. We further analyze the asymptotic equivalence of trace regression and quantum state tomography for sparse matrices and low rank matrices. The detailed analyses indicate that the asymptotic equivalence does not require sparsity nor low rank on matrix parameters, and depending on the density matrix class as well as the set of observables used for performing measurements, sparsity and low rank may or may not make the asymptotic equivalence easier to achieve. In particular, we show that the Pauli matrices as observables are bad for establishing the asymptotic equivalence for sparse matrices and low rank matrices; and for certain class of sparse or low rank density matrices, we can obtain the asymptotic equivalence of quantum state tomography and trace regression in the ultra high dimension setting where the matrix size of the density matrices is comparable to or even exceeds the number of the quantum measurements on the observables.

The rest of paper proceeds as follows. Section 2 reviews trace regression and quantum state tomography and states statistical models and data structures. We consider only finite square matrices, since trace regression handles finite matrices, and density matrices are square matrices. Section 3 frames trace regression and quantum state tomography with summarized measurements as two statistical experiments in Le Cam paradigm and studies their asymptotic equivalence. Section 4 introduces a fine scale trace regression model to match quantum state tomography with individual measurements and investigates their asymptotic equivalence. We illustrate the asymptotic equivalence for sparse density matrix class and low rank density matrix class in Sections 5 and 6, respectively. We collect technical proofs in Section 7, with additional proofs of technical lemmas in the Appendix.

2. Statistical models and data structures.

2.1. *Trace regression in matrix completion.* Suppose that we have n independent random pairs $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$ from the model

$$Y_k = \operatorname{tr}(\mathbf{X}_k^\dagger \boldsymbol{\rho}) + \varepsilon_k, \qquad k = 1, \ldots, n, \tag{1}$$

where tr is matrix trace, $\dagger$ denotes conjugate transpose, $\boldsymbol{\rho}$ is an unknown d by d matrix, ε_k are zero mean random errors, and $\mathbf{X}_k$ are matrix input variables of

size d by d. We consider both fixed and random designs. For the random design case, each $\mathbf{X}_k$ is randomly sampled from a set of matrices. In the fixed design case, $\mathbf{X}_1, \ldots, \mathbf{X}_n$ are fixed matrices. Model (1) is called trace regression and employed in matrix completion. Matrix input variables $\mathbf{X}_k$ are often sparse in a sense that each $\mathbf{X}_k$ has a relatively small number of nonzero entries. Trace regression masks the entries of $\boldsymbol{\rho}$ through $\mathbf{X}_k^\dagger \boldsymbol{\rho}$, and each observation Y_k is the trace of the masked $\boldsymbol{\rho}$ corrupted by noise ε_k. The statistical problem is to estimate all the entries of $\boldsymbol{\rho}$ based on observations $(\mathbf{X}_k, Y_k)$, $k = 1, \ldots, n$, which is often referred to as noisy matrix completion. Model (1) and matrix completion are matrix generalizations of a linear model and sparse signal estimation in compressed sensing. See Candés and Plan (2009, 2011), Candès and Recht (2009), Candès and Tao (2010), Keshavan, Montanari and Oh (2010), Koltchinskii, Lounici and Tsybakov (2011), and Negahban and Wainwright (2011), Koltchinskii (2011) and Rohde and Tsybakov (2011).

Matrix input variables $\mathbf{X}_k$ are selected from a matrix set $\mathcal{B} = \{\mathbf{B}_1, \ldots, \mathbf{B}_p\}$, where $\mathbf{B}_j$ are d by d matrices. Below we list some examples of such matrix sets used in matrix completion.

(i) Let

$$\begin{aligned}\mathcal{B} = \{&\mathbf{B}_j = \mathbf{e}_{\ell_1}\mathbf{e}'_{\ell_2},\ j = (\ell_1 - 1)d + \ell_2,\\ &j = 1, \ldots, p = d^2,\ \ell_1, \ell_2 = 1, \ldots, d\},\end{aligned} \tag{2}$$

where $\mathbf{e}_\ell$ is the canonical basis in Euclid space $\mathbb{R}^d$. In this case, if $\boldsymbol{\rho} = (\rho_{ab})$, then $\operatorname{tr}(\mathbf{B}_j\boldsymbol{\rho}) = \rho_{\ell_1\ell_2}$, and the observation Y_k is equal to some entry of $\boldsymbol{\rho}$ plus noise ε_k. More generally, instead of using single $\mathbf{e}_{\ell_1}\mathbf{e}'_{\ell_2}$, we may define $\mathbf{B}_j$ as the sum of several $\mathbf{e}_{\ell_1}\mathbf{e}'_{\ell_2}$, and then $\operatorname{tr}(\mathbf{B}_j\boldsymbol{\rho})$ is equal to the sum of some entries of $\boldsymbol{\rho}$.

(ii) Set

$$\mathcal{B} = \{\mathbf{B}_j,\ j = 1, \ldots, p = d^2\}, \tag{3}$$

where we identify j with (ℓ_1, ℓ_2), $j = 1, \ldots, p$, $\ell_1, \ell_2 = 1, \ldots, d$, $\mathbf{B}_j = \mathbf{e}_{\ell_1}\mathbf{e}'_{\ell_2}$ for $\ell_1 = \ell_2$,

$$\mathbf{B}_j = \frac{1}{\sqrt{2}}(\mathbf{e}_{\ell_1}\mathbf{e}'_{\ell_2} + \mathbf{e}_{\ell_2}\mathbf{e}'_{\ell_1}) \qquad \text{for } \ell_1 < \ell_2$$

and

$$\mathbf{B}_j = \frac{\sqrt{-1}}{\sqrt{2}}(\mathbf{e}_{\ell_1}\mathbf{e}'_{\ell_2} - \mathbf{e}_{\ell_2}\mathbf{e}'_{\ell_1}) \qquad \text{for } \ell_1 > \ell_2.$$

(iii) For $d = 2$ define

$$\boldsymbol{\sigma}_0 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \qquad \boldsymbol{\sigma}_1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix},$$

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

where $\boldsymbol{\sigma}_1$, $\boldsymbol{\sigma}_2$ and $\boldsymbol{\sigma}_3$ are called the Pauli matrices. For $d = 2^b$ with integer b, we may use b-fold tensor products of $\boldsymbol{\sigma}_0$, $\boldsymbol{\sigma}_1$, $\boldsymbol{\sigma}_2$ and $\boldsymbol{\sigma}_3$ to define general Pauli matrices and obtain the Pauli matrix set

$$\mathcal{B} = \{\boldsymbol{\sigma}_{\ell_1} \otimes \boldsymbol{\sigma}_{\ell_2} \otimes \cdots \otimes \boldsymbol{\sigma}_{\ell_b}, (\ell_1, \ell_2, \ldots, \ell_b) \in \{0, 1, 2, 3\}^b\}, \tag{4}$$

where $\otimes$ denotes tensor product. The Pauli matrices are widely used in quantum physics and quantum information science.

Matrices in (2) are of rank 1 and have eigenvalues 1 and 0. For matrices in (3), the diagonal matrices are of rank 1 and have eigenvalues 1 and 0, and the nondiagonal matrices are of rank 2 and have eigenvalues ± 1 and 0. Pauli matrices in (4) are of full rank, and except for the identity matrix all have eigenvalues ± 1. Denote by $\mathbb{C}^{d\times d}$ the space of all d by d complex matrices and define an inner product $\langle\langle \mathbf{A}_1, \mathbf{A}_2\rangle\rangle = \operatorname{tr}(\mathbf{A}_2^\dagger \mathbf{A}_1)$ for $\mathbf{A}_1, \mathbf{A}_2 \in \mathbb{C}^{d\times d}$. Then both (3) and (4) form orthogonal bases for all complex Hermitian matrices, and the real matrices in (3) or (4) form orthogonal bases for all real symmetric matrices.

For the random design case, with $\mathcal{B} = \{\mathbf{B}_j, j = 1, \ldots, p\}$, we assume that matrix input variables $\mathbf{X}_k$ are independent and sampled from $\mathcal{B}$ according to a distribution $\Pi(j)$ on $\{1, \ldots, p\}$,

$$P(\mathbf{X}_k = \mathbf{B}_{j_k}) = \Pi(j_k), \qquad k = 1, \ldots, n,\ j_k \in \{1, \ldots, p\}. \tag{5}$$

The observations from (1) are $(\mathbf{X}_k, Y_k)$, $k = 1, \ldots, n$, with $\mathbf{X}_k$ sampled from $\mathcal{B}$ according to the distribution $\Pi(\cdot)$. For the fixed design case, matrix input variables $\mathbf{X}_1, \ldots, \mathbf{X}_n$ form a fixed set of matrices, and we assume $n = p$ and $\mathcal{B} = \{\mathbf{X}_1, \ldots, \mathbf{X}_n\} = \{\mathbf{B}_1, \ldots, \mathbf{B}_p\}$. The observations from (1) are $(\mathbf{X}_k, Y_k)$, $k = 1, \ldots, n$, with deterministic $\mathbf{X}_k$.

2.2. *Quantum state and measurements.* For a finite-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) semi-positive definite; (3) unit trace, that is, $\operatorname{tr}(\boldsymbol{\rho}) = 1$.

Experiments are conducted to perform measurements on the quantum system and obtain data for studying the quantum system. Common quantum measurements are on some 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:

$$\mathbf{M} = \sum_{a=1}^{r} \lambda_a \mathbf{Q}_a, \tag{6}$$

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 a 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 $\{\lambda_1, \lambda_2, \ldots, \lambda_r\}$, with probability distribution given by

$$P(R = \lambda_a) = \operatorname{tr}(\mathbf{Q}_a \boldsymbol{\rho}), \qquad a = 1, 2, \ldots, r, \qquad E(R) = \operatorname{tr}(\mathbf{M}\boldsymbol{\rho}). \tag{7}$$

See Holevo (1982), Sakurai and Napolitano (2010), Shankar (1994) and Wang (2012).

Suppose that an experiment is conducted to perform measurements on $\mathbf{M}$ independently for m quantum systems which are identically prepared in the same quantum state $\boldsymbol{\rho}$. From the experiment we obtain individual measurements $R_1, \ldots, R_m$, which are i.i.d. according to distribution (7), and denote their average by $N = (R_1 + \cdots + R_m)/m$.

The following proposition provides a simple multinomial characterization for the distributions of $(R_1, \ldots, R_m)$ and N.

PROPOSITION 2.1. *As random variables $R_1, \ldots, R_m$ take eigenvalues $\lambda_1, \ldots, \lambda_r$, we count the number of $R_1, \ldots, R_m$ taking λ_a and define the counts by $U_a = \sum_{\ell=1}^m 1(R_\ell = \lambda_a)$, $a = 1, \ldots, r$. Then the counts $U_1, \ldots, U_r$ jointly follow the following multinomial distribution*:

$$P(U_1 = u_1, \ldots, U_r = u_r) = \binom{m}{u_1, \ldots, u_r} \big[\operatorname{tr}(\mathbf{Q}_1\boldsymbol{\rho})\big]^{u_1} \cdots \big[\operatorname{tr}(\mathbf{Q}_r\boldsymbol{\rho})\big]^{u_r}, \quad \sum_{a=1}^r u_a = m \tag{8}$$

and

$$N = (R_1 + \cdots + R_m)/m = (\lambda_1 U_1 + \cdots + \lambda_a U_a)/m. \tag{9}$$

We note the difference between the observable $\mathbf{M}$ which is a Hermitian matrix and its measurement result R which is a real-valued random variable. To illustrate the connection between density matrix $\boldsymbol{\rho}$ and the measurements of $\mathbf{M}$, we assume that $\mathbf{M}$ has d different eigenvalues. As in Artiles, Gill and Guţă (2005), we use the normalized eigenvectors of $\mathbf{M}$ to form an orthonormal basis, represent $\boldsymbol{\rho}$ under the basis and denote the resulting matrix by $(\rho_{\ell_1\ell_2})$. Then from (7) we obtain

$$P(R = \lambda_a) = \operatorname{tr}(\mathbf{Q}_a\boldsymbol{\rho}) = \rho_{aa}, \qquad a = 1, 2, \ldots, d.$$

That is, with the representation under the eigen basis of $\mathbf{M}$, measurements on single observable $\mathbf{M}$ contain only information about the diagonal elements of $(\rho_{\ell_1\ell_2})$. No matter how many measurements we perform on $\mathbf{M}$, we cannot draw any inference about the off-diagonal elements of $(\rho_{\ell_1\ell_2})$ based on the measurements on $\mathbf{M}$. We usually need to perform measurements on enough different observables in order to estimate the whole density matrix $(\rho_{\ell_1\ell_2})$. See Artiles, Gill and Guţă (2005), Barndorff-Nielsen, Gill and Jupp (2003) and Butucea, Guţă and Artiles (2007).

2.3. *Quantum state tomography.* In physics literature quantum state tomography refers to the reconstruction of a quantum state based on measurements obtained from quantum systems that are identically prepared under the state. Statistically it is the problem of estimating the density matrix from the measurements. Suppose that quantum systems are identically prepared in a state ρ, $\mathcal{B} = \{\mathbf{B}_1, \ldots, \mathbf{B}_p\}$ is a set of observables available to perform measurements, and each $\mathbf{B}_j$ has a spectral decomposition

$$\mathbf{B}_j = \sum_{a=1}^{r_j} \lambda_{ja} \mathbf{Q}_{ja}, \tag{10}$$

where λ_{ja} are r_j different real eigenvalues of $\mathbf{B}_j$, and $\mathbf{Q}_{ja}$ are projections onto the eigen-spaces corresponding to λ_{ja}. We select an observable, say $\mathbf{B}_j \in \mathcal{B}$, and perform measurements on $\mathbf{B}_j$ for the quantum systems. According to the observable selection we classify the quantum state tomography experiment as either a fixed design or a random design. In a random design, we choose an observable at random from $\mathcal{B}$ to perform measurements for the quantum systems, while a fixed design is to perform measurements on every observable in $\mathcal{B}$ for the quantum systems.

Consider the random design case. We sample an observable $\mathbf{M}_k$ from $\mathcal{B}$ to perform measurements independently for m quantum systems, $k = 1, \ldots, n$, where observables $\mathbf{M}_1, \ldots, \mathbf{M}_n$ are independent and sampled from $\mathcal{B}$ according to a distribution $\Xi(j)$ on $\{1, \ldots, p\}$,

$$P(\mathbf{M}_k = \mathbf{B}_{j_k}) = \Xi(j_k), \qquad k = 1, \ldots, n,\ j_k \in \{1, \ldots, p\}. \tag{11}$$

Specifically we perform measurements on each observable $\mathbf{M}_k$ independently for m quantum systems that are identically prepared under the state $\boldsymbol{\rho}$, and denote by $R_{k1}, \ldots, R_{km}$ the m measurement outcomes and N_k the average of the m measurement outcomes. The resulting individual measurements are the data $(\mathbf{M}_k, R_{k1}, \ldots, R_{km})$, $k = 1, \ldots, n$, and the summarized measurements are the pairs $(\mathbf{M}_k, N_k)$, $k = 1, \ldots, n$, where

$$N_k = (R_{k1} + \cdots + R_{km})/m, \tag{12}$$

$R_{k\ell}$, $k = 1, \ldots, n$, $\ell = 1, \ldots, m$, are independent, and given $\mathbf{M}_k = \mathbf{B}_{j_k}$ for some $j_k \in \{1, \ldots, p\}$, the conditional distributions of $R_{k1}, \ldots, R_{km}$ are given by

$$\begin{aligned} P(R_{k\ell} = \lambda_{j_k a} | \mathbf{M}_k = \mathbf{B}_{j_k}) &= \operatorname{tr}(\mathbf{Q}_{j_k a} \boldsymbol{\rho}), \\ &\quad a = 1, \ldots, r_{j_k},\ \ell = 1, \ldots, m,\ j_k \in \{1, \ldots, p\}, \end{aligned} \tag{13}$$

$$\begin{aligned} E(R_{k\ell} | \mathbf{M}_k = \mathbf{B}_{j_k}) &= \operatorname{tr}(\mathbf{B}_{j_k} \boldsymbol{\rho}), \\ \operatorname{Var}(R_{k\ell} | \mathbf{M}_k = \mathbf{B}_{j_k}) &= \operatorname{tr}(\mathbf{B}_{j_k}^2 \boldsymbol{\rho}) - \big[\operatorname{tr}(\mathbf{B}_{j_k} \boldsymbol{\rho})\big]^2. \end{aligned} \tag{14}$$

The statistical problem is to estimate $\boldsymbol{\rho}$ from the individual measurements $(\mathbf{M}_k, R_{k1}, \ldots, R_{km})$, $k = 1, \ldots, n$, or from the summarized measurements $(\mathbf{M}_1, N_1), \ldots, (\mathbf{M}_n, N_n)$.

For the fixed design case, we take $p = n$ and $\mathcal{B} = \{\mathbf{B}_1, \ldots, \mathbf{B}_n\}$. We perform measurements on every observable $\mathbf{M}_k = \mathbf{B}_k \in \mathcal{B}$ independently for m quantum systems that are identically prepared under the state $\boldsymbol{\rho}$, and denote by $R_{k1}, \ldots, R_{km}$ the m measurement outcomes and N_k the average of the m measurement outcomes. The resulting individual measurements are the data $(\mathbf{M}_k, R_{k1}, \ldots, R_{km})$, $k = 1, \ldots, n$, and the summarized measurements are the pairs $(\mathbf{M}_k, N_k)$, $k = 1, \ldots, n$, where N_k is the same as in (12), $R_{k\ell}$, $k = 1, \ldots, n$, $\ell = 1, \ldots, m$, are independent, and the distributions of $R_{k1}, \ldots, R_{km}$ are given by

$$P(R_{k\ell} = \lambda_{ka}) = \operatorname{tr}(\mathbf{Q}_{ka}\boldsymbol{\rho}), \qquad a = 1, \ldots, r_k, \ell = 1, \ldots, m, \tag{15}$$

$$E(R_{k\ell}) = \operatorname{tr}(\mathbf{M}_k\boldsymbol{\rho}), \qquad \operatorname{Var}(R_{k\ell}) = \operatorname{tr}\big(\mathbf{M}_k^2\boldsymbol{\rho}\big) - \big[\operatorname{tr}(\mathbf{M}_k\boldsymbol{\rho})\big]^2. \tag{16}$$

The statistical problem is to estimate $\boldsymbol{\rho}$ from the individual measurements $(\mathbf{M}_k, R_{k1}, \ldots, R_{km})$, $k = 1, \ldots, n$, or from the summarized measurements $(\mathbf{M}_1, N_1), \ldots, (\mathbf{M}_n, N_n)$.

Because of convenient statistical procedures and fast implementation algorithms, the summarized measurements instead of the individual measurements are often employed in quantum state tomography [Gross et al. (2010), Koltchinskii (2011), Nielsen and Chuang (2000)]. However, in Section 4 we will show that quantum state tomography based on the summary measurements may suffer from substantial loss of information, and we can develop more efficient statistical inference procedures by the individual measurements than by the summary measurements.

In order to estimate all $d^2 - 1$ free entries of $\boldsymbol{\rho}$, we need the quantum state tomography model identifiable. Suppose that all $\mathbf{B}_j$ have exact r distinct eigenvalues. The identifiability may require $n \geq (d^2 - 1)/(r - 1)$ (which is at least $d + 1$) and $m \geq r - 1$ for the individual measurements and $n \geq d^2 - 1$ for the summarized measurements. There is a trade-off between r and m in the individual measurement case. For large r, we need less observables but more measurements on each observable, while for small r, we require more observables but less measurements on each observable. In terms of the total number, mn, of measurement data, the requirement becomes $mn \geq d^2 - 1$.

3. Asymptotic equivalence. Quantum state tomography and trace regression share the common goal of estimating the same unknown matrix $\boldsymbol{\rho}$, and it is nature to put them in the Le Cam paradigm for statistical comparison. We compare trace regression and quantum state tomography in either the fixed design case or the random design case.

First, we consider the fixed design case. Trace regression (1) generates data on dependent variables Y_k with deterministic matrix input variables $\mathbf{X}_k$, and we denote by $\mathbb{P}_{1,n,\boldsymbol{\rho}}$ the joint distribution of Y_k, $k = 1, \ldots, n$. Quantum state tomography performs measurements on a fixed set of observables $\mathbf{M}_k$ and obtains average measurements N_k on $\mathbf{M}_k$ whose distributions are specified by (12) and (15)–(16),

and we denote by $\mathbb{P}_{2,n,\boldsymbol{\rho}}$ the joint distribution of N_k, $k=1,\ldots,n$. Both $\mathbb{P}_{1,n,\boldsymbol{\rho}}$ and $\mathbb{P}_{2,n,\boldsymbol{\rho}}$ are probability distributions on measurable space $(\mathbb{R}^n, \mathcal{F}_{\mathbb{R}}^n)$, where $\mathcal{F}_{\mathbb{R}}$ is the Borel σ-field on $\mathbb{R}$.

Second we consider the random design case. Trace regression (1) generates data on the pairs $(\mathbf{X}_k, Y_k)$, $k=1,\ldots,n$, where matrix input variables $\mathbf{X}_k$ are sampled from $\mathcal{B}$ according to the distribution $\Pi(j)$ given by (5). We denote by $\mathbb{P}_{1,n,\boldsymbol{\rho}}$ the joint distribution of $(\mathbf{X}_k, Y_k)$, $k=1,\ldots,n$, for the trace regression model. Quantum state tomography yields observations in the form of observables $\mathbf{M}_k$ and average measurement results N_k on $\mathbf{M}_k$, $k=1,\ldots,n$, where the distributions of $(\mathbf{M}_k, N_k)$ are specified by (11)–(14). We denote by $\mathbb{P}_{2,n,\boldsymbol{\rho}}$ the joint distribution of $(\mathbf{M}_k, N_k)$, $k=1,\ldots,n$, for the quantum state tomography model. Both $\mathbb{P}_{1,n,\boldsymbol{\rho}}$ and $\mathbb{P}_{2,n,\boldsymbol{\rho}}$ are probability distributions on measurable space $(\mathcal{B}^n \times \mathbb{R}^n, \mathcal{F}_{\mathcal{B}}^n \times \mathcal{F}_{\mathbb{R}}^n)$, where $\mathcal{F}_{\mathcal{B}}$ consists of all subsets of $\mathcal{B}$.

Denote by Θ a class of semi-positive Hermitian matrices with unit trace. For trace regression and quantum state tomography, we define two statistical models

$$\mathcal{P}_{1n} = \bigl\{(\mathcal{X}_1, \mathcal{G}_1, \mathbb{P}_{1,n,\boldsymbol{\rho}}), \boldsymbol{\rho} \in \Theta\bigr\}, \qquad \mathcal{P}_{2n} = \bigl\{(\mathcal{X}_2, \mathcal{G}_2, \mathbb{P}_{2,n,\boldsymbol{\rho}}), \boldsymbol{\rho} \in \Theta\bigr\}, \tag{17}$$

where measurable spaces $(\mathcal{X}_i, \mathcal{G}_i)$, $i=1,2$, are either $(\mathcal{B}^n \times \mathbb{R}^n, \mathcal{F}_{\mathcal{B}}^n \times \mathcal{F}_{\mathbb{R}}^n)$ for the random design case or $(\mathbb{R}^n, \mathcal{F}_{\mathbb{R}}^n)$ for the fixed design case. Models $\mathcal{P}_{1n}$ and $\mathcal{P}_{2n}$ are called statistical experiments in the Le Cam paradigm. We use Le Cam's deficiency distance between $\mathcal{P}_{1n}$ and $\mathcal{P}_{2n}$ to compare the two models. Let $\mathcal{A}$ be a measurable action space, $L\colon \Theta \times \mathcal{A} \to [0,\infty)$ a loss function, and $\|L\| = \sup\{L(\boldsymbol{\rho}, \mathbf{a}) : \boldsymbol{\rho} \in \Theta, \mathbf{a} \in \mathcal{A}\}$. For model $\mathcal{P}_{in}$, $i=1,2$, denote by χ_i a decision procedure and $R_i(\chi_i, L, \boldsymbol{\rho})$ the risk from using procedure χ_i when L is the loss function and $\boldsymbol{\rho}$ is the true value of the parameter. We define deficiency distance $\Delta(\mathcal{P}_{1n}, \mathcal{P}_{2n})$ between $\mathcal{P}_{1n}$ and $\mathcal{P}_{2n}$ as the maximum of $\delta(\mathcal{P}_{1n}, \mathcal{P}_{2n})$ and $\delta(\mathcal{P}_{2n}, \mathcal{P}_{1n})$, where

$$\delta(\mathcal{P}_{1n}, \mathcal{P}_{2n}) = \inf_{\chi_1} \sup_{\chi_2} \sup_{\boldsymbol{\rho}\in\Theta} \sup_{L:\|L\|=1} \bigl|R_1(\chi_1, L, \boldsymbol{\rho}) - R_2(\chi_2, L, \boldsymbol{\rho})\bigr|$$

is referred to as the deficiency of $\mathcal{P}_{1n}$ with respect to $\mathcal{P}_{2n}$. If $\Delta(\mathcal{P}_{1n}, \mathcal{P}_{2n}) \le \epsilon$, then every decision procedure in one of the two experiments $\mathcal{P}_{1n}$ and $\mathcal{P}_{2n}$ has a corresponding procedure in another experiment that comes within ϵ of achieving the same risk for any bounded loss. Two sequences of statistical experiments $\mathcal{P}_{1n}$ and $\mathcal{P}_{2n}$ are called asymptotically equivalent if $\Delta(\mathcal{P}_{1n}, \mathcal{P}_{2n}) \to 0$, as $n \to \infty$. For two asymptotic equivalent experiments $\mathcal{P}_{1n}$ and $\mathcal{P}_{2n}$, any sequence of procedures χ_{1n} in model $\mathcal{P}_{1n}$ has a corresponding sequence of procedures χ_{2n} in model $\mathcal{P}_{2n}$ with risk differences tending to zero uniformly over $\boldsymbol{\rho} \in \Theta$ and all loss L with $\|L\| = 1$, and the procedures χ_{1n} and χ_{2n} are called asymptotically equivalent. See Le Cam (1986), Le Cam and Yang (2000) and Wang (2002).

To establish the asymptotic equivalence of trace regression and quantum state tomography, we need to lay down technical conditions and make some synchronization arrangement between observables in quantum state tomography and matrix input variables in trace regression.

(C1) Assume that $\mathcal{B} = \{\mathbf{B}_1, \ldots, \mathbf{B}_p\}$, and each $\mathbf{B}_j$ is a Hermitian matrix with at most κ distinct eigenvalues, where κ is a fixed integer. Matrix input variables $\mathbf{X}_k$ in trace regression and observables $\mathbf{M}_k$ in quantum state tomography are taken from $\mathcal{B}$. For the fixed design case, we assume $p = n$, and $\mathbf{X}_k = \mathbf{M}_k = \mathbf{B}_k$, $k = 1, \ldots, n$. For the random design case, $\mathbf{X}_k$ and $\mathbf{M}_k$ are independently sampled from $\mathcal{B}$ according to distributions $\Pi(j)$ and $\Xi(j)$, respectively, and assume that as $n, p \to \infty$, $n\gamma_p \to 0$, where

$$\gamma_p = \max_{1 \le j \le p} \left[\left| 1 - \frac{\Pi(j)}{\Xi(j)} \right| + \left| 1 - \frac{\Xi(j)}{\Pi(j)} \right| \right]. \tag{18}$$

(C2) Suppose that two models $\mathcal{P}_{1n}$ and $\mathcal{P}_{2n}$ are identifiable. For trace regression, we assume that $(\mathbf{X}_1, \varepsilon_1), \ldots, (\mathbf{X}_n, \varepsilon_n)$ are independent, and given $\mathbf{X}_k$, ε_k follows a normal distribution with mean zero and variance

$$\operatorname{Var}(\varepsilon_k | \mathbf{X}_k) = \frac{1}{m} \{ \operatorname{tr}(\mathbf{X}_k^2 \boldsymbol{\rho}) - [\operatorname{tr}(\mathbf{X}_k \boldsymbol{\rho})]^2 \}. \tag{19}$$

(C3) For $\mathbf{B}_j \in \mathcal{B}$ with spectral decomposition (10), $j = 1, \ldots, p$, let

$$\mathcal{I}_j(\boldsymbol{\rho}) = \{ a : 0 < \operatorname{tr}(\mathbf{Q}_{ja} \boldsymbol{\rho}) < 1, 1 \le a \le r_j \}. \tag{20}$$

Let c_0 and c_1 be two fixed constants with $0 < c_0 \le c_1 < 1$. Assume for $\boldsymbol{\rho} \in \Theta$,

$$c_0 \le \min_{a \in \mathcal{I}_j(\boldsymbol{\rho})} \operatorname{tr}(\mathbf{Q}_{ja} \boldsymbol{\rho}) \le \max_{a \in \mathcal{I}_j(\boldsymbol{\rho})} \operatorname{tr}(\mathbf{Q}_{ja} \boldsymbol{\rho}) \le c_1, \qquad j = 1, \ldots, p. \tag{21}$$

REMARK 1. Condition (C1) synchronizes matrices used as matrix input variables in trace regression and as observables in quantum state tomography so that we can compare the two models. The synchronization is needed for applying matrix completion methods to quantum state tomography [Gross et al. (2010)]. The finiteness assumption on κ is due to the practical consideration. Observables in quantum state tomography and matrix input variables in trace regression are often of large size. Mathematically the numbers of their distinct eigenvalues could grow with the size, however, in practice matrices with a few distinct eigenvalues are usually chosen as observables to perform measurements in quantum state tomography and as matrix input variables to mask the entries of $\boldsymbol{\rho}$ in matrix completion [Candès and Recht (2009), Gross (2011), Gross et al. (2010), Koltchinskii (2011), Koltchinskii, Lounici and Tsybakov (2011), Nielsen and Chuang (2000), Recht (2011), Rohde and Tsybakov (2011)]. Condition (C2) is to match the variance of N_k in quantum state tomography with the variance of random error ε_k in trace regression in order to obtain the asymptotic equivalence, since N_k and Y_k always have the same mean. Regarding condition (C3), from (8)–(9) and (12)–(16) we may see that each N_k is determined by the counts of random variables $R_{k\ell}$ taking eigenvalues λ_{ja}, and the counts jointly follow a multinomial distribution with parameters of m trials and cell probabilities $\operatorname{tr}(\mathbf{Q}_{ja} \boldsymbol{\rho})$, $a = 1, \ldots, r_j$. Condition (C3) is to ensure that the multinomial distributions (with uniform perturbations) can be

well approximated by multivariate normal distributions so that we can calculate the Hellinger distance between the distributions of N_k (with uniform perturbations) in quantum state tomography and the distributions of ε_k in trace regression and thus establish the asymptotic equivalence of quantum state tomography and trace regression. Index $\mathcal{I}_j(\boldsymbol{\rho})$ in (20) is to exclude all the cases with $\operatorname{tr}(\mathbf{Q}_{ja}\boldsymbol{\rho}) = 0$ or $\operatorname{tr}(\mathbf{Q}_{ja}\boldsymbol{\rho}) = 1$, under which measurement results on $\mathbf{B}_j$ are certain, either never yielding measurement results λ_{ja} or always yielding results λ_{ja}, and their contributions to N_k are deterministic and can be completely separated out from N_k. See further details in Remark 4 below and the proofs of Theorems 1 and 2 in Section 7.

The following theorem provides bounds on deficiency distance $\Delta(\mathcal{P}_{1n}, \mathcal{P}_{2n})$ and establishes the asymptotic equivalence of trace regression and quantum state tomography under the fixed or random designs.

THEOREM 1. *Assume that conditions* (C1)–(C3) *are satisfied.*

(a) *For the random design case, we have*

$$\Delta(\mathcal{P}_{1n}, \mathcal{P}_{2n}) \le n\gamma_p + C\left(\frac{n\zeta_p}{m}\right)^{1/2}, \tag{22}$$

where C is a generic constant depending only on (κ, c_0, c_1), integer κ and constants (c_0, c_1) are, respectively, specified in conditions (C1) *and* (C3), *γ_p is defined in* (18), *and ζ_p is given by*

$$\zeta_p = \max_{\boldsymbol{\rho}\in\Theta}\left\{\sum_{j=1}^{p}\Pi(j)1\big(|\mathcal{I}_j(\boldsymbol{\rho})| \ge 2\big), \sum_{j=1}^{p}\Xi(j)1\big(|\mathcal{I}_j(\boldsymbol{\rho})| \ge 2\big)\right\} \le 1. \tag{23}$$

In particular, if $\Pi(j) = \Xi(j) = 1/p$ for $j = 1, \ldots, p$, then

$$\Delta(\mathcal{P}_{1n}, \mathcal{P}_{2n}) \le C\left(\frac{n\zeta_p}{m}\right)^{1/2}, \tag{24}$$

where now ζ_p can be simplified as

$$\zeta_p = \max_{\boldsymbol{\rho}\in\Theta}\left\{\frac{1}{p}\sum_{j=1}^{p}1\big(|\mathcal{I}_j(\boldsymbol{\rho})| \ge 2\big)\right\} \le 1. \tag{25}$$

(b) *For the fixed design case, we have*

$$\Delta(\mathcal{P}_{1n}, \mathcal{P}_{2n}) \le C\left(\frac{n\zeta_p}{m}\right)^{1/2}, \tag{26}$$

where C is the same as in (a), *and ζ_p is given by* (25).

REMARK 2. Theorem 1 establishes bounds on the deficiency distance between trace regression and quantum state tomography. If the deficiency distance bounds in (22), (24) and (26) go to zero, trace regression and quantum state tomography are asymptotically equivalent under the corresponding cases. ζ_p defined in (23) and (25) has an intuitive interpretation as follows. Proposition 2.1 shows that each observable corresponds to a multinomial distribution in quantum state tomography. Of the p multinomial distributions in quantum state tomography, ζ_p is the maximum of the average fraction of the nondegenerate multinomial distributions (i.e., with at least two cells). As we discussed in Remark 1, the multinomial distributions have cell probabilities $\operatorname{tr}(\mathbf{Q}_{ja}\boldsymbol{\rho})$, $a = 1, \ldots, r_j$. Since for each $\mathbf{B}_j$, $\operatorname{tr}(\mathbf{Q}_{ja}\boldsymbol{\rho})$ is the trace of the density matrix $\boldsymbol{\rho}$ restricted to the corresponding eigenspace, and $\sum_{a=1}^{r_j} \operatorname{tr}(\mathbf{Q}_{ja}\boldsymbol{\rho}) = \operatorname{tr}(\boldsymbol{\rho}) = 1$, thus if $|\mathcal{I}_j(\boldsymbol{\rho})| \geq 2$, $\boldsymbol{\rho}$ cannot live on any single eigen-space corresponding to one eigenvalue of $\mathbf{B}_j$; otherwise measurement results on $\mathbf{B}_j$ are certain, and the corresponding multinomial and normal distributions are reduced to the same degenerate distribution and hence are always equivalent. Therefore, to bound the deficiency distance between quantum state tomography and trace regression we need to consider only the nondegenerate multinomial distributions, and thus ζ_p appears in all the deficiency distance bounds. Since ζ_p is always bounded by 1, from Theorem 1 we have that if $n/m \to 0$, the two models are asymptotically equivalent. As we will see in Sections 5 and 6, depending on density matrix class Θ as well as the matrix set $\mathcal{B}$, ζ_p may or may not go to zero, and we will show that if it approaches to zero, we may have asymptotic equivalence in ultra-high dimensions where d may be comparable to or exceed m.

REMARK 3. The asymptotic equivalence results indicate that we may apply matrix completion methods to quantum state tomography by substituting $(\mathbf{M}_k, N_k)$ from quantum state tomography for $(\mathbf{X}_k, Y_k)$ from trace regression. For example, suppose that $\mathcal{B}$ is an orthonormal basis and $\boldsymbol{\rho}$ has an expansion $\boldsymbol{\rho} = \sum_j \alpha_j \mathbf{B}_j$ with $\alpha_j = \operatorname{tr}(\boldsymbol{\rho}\mathbf{B}_j)$. For trace regression, we may estimate α_j by the average of those Y_k with corresponding $\mathbf{X}_k = \mathbf{B}_j$. Replacing $(\mathbf{X}_k, Y_k)$ from trace regression by $(\mathbf{M}_k, N_k)$ from quantum state tomography we construct an estimator of α_j by taking the average of those N_k with corresponding $\mathbf{M}_k = \mathbf{B}_j$. In fact, the resulting estimator based on N_k can be naturally derived from quantum state tomography. From (7), (14) and (16), we have $\alpha_j = \operatorname{tr}(\boldsymbol{\rho}\mathbf{B}_j) = E(R)$, where R is the outcome of measuring $\mathbf{B}_j$, and hence it is natural to estimate α_j by the average of quantum measurements $R_{k\ell}$ with corresponding $\mathbf{M}_k = \mathbf{B}_j$. As statistical procedures and fast algorithms are available for trace regression, these statistical methods and computational techniques can be easily used to implement quantum state tomography based on the summarized measurements [Gross et al. (2010) and Koltchinskii (2011)].

4. Fine scale trace regression. In Section 3 for quantum state tomography we define $\mathbb{P}_{2,n,\boldsymbol{\rho}}$ and $\mathcal{P}_{2n}$ in (17) based on the average measurements N_k, and the

asymptotic equivalence results show that trace regression matches quantum state tomography with the summarized measurements $(\mathbf{M}_k, N_k)$, $k = 1, \ldots, n$. We may use individual measurements $R_{k1}, \ldots, R_{km}$ instead of their averages N_k [see (12)–(16) for their definitions and relationships], and replace $\mathbb{P}_{2,n,\boldsymbol{\rho}}$ in (17) by the joint distribution, $\mathbb{Q}_{2,n,\boldsymbol{\rho}}$, of $(\mathbf{M}_k, R_{k1}, \ldots, R_{km})$, $k = 1, \ldots, n$, for the random design case [or $(R_{k1}, \ldots, R_{km})$, $k = 1, \ldots, n$, for the fixed design case] to define a new statistical experiment for quantum state tomography with the individual measurements,

$$\mathcal{Q}_{2n} = \big\{(\mathcal{X}_2, \mathcal{G}_2, \mathbb{Q}_{2,n,\boldsymbol{\rho}}), \boldsymbol{\rho} \in \Theta\big\}, \tag{27}$$

where measurable space $(\mathcal{X}_2, \mathcal{G}_2)$ is either $(\mathcal{B}^n \times \mathbb{R}^{mn}, \mathcal{F}_{\mathcal{B}}^n \times \mathcal{F}_{\mathbb{R}}^{mn})$ for the random design case or $(\mathbb{R}^{mn}, \mathcal{F}_{\mathbb{R}}^{mn})$ for the fixed design case.

In general, $\mathcal{P}_{1n}$ and $\mathcal{Q}_{2n}$ may not be asymptotically equivalent. As individual measurements $R_{k1}, \ldots, R_{km}$ may contain more information than their average N_k, $\mathcal{Q}_{2n}$ may be more informative than $\mathcal{P}_{2n}$, and hence $\delta(\mathcal{Q}_{2n}, \mathcal{P}_{2n}) = 0$ but $\delta(\mathcal{P}_{2n}, \mathcal{Q}_{2n})$ may be bounded away from zero. As a consequence, we may have $\delta(\mathcal{Q}_{2n}, \mathcal{P}_{1n})$ goes to zero but $\delta(\mathcal{P}_{1n}, \mathcal{Q}_{2n})$ and $\Delta(\mathcal{P}_{1n}, \mathcal{Q}_{2n})$ are bounded away from zero. For the special case of $\kappa = 2$ where all $\mathbf{B}_j$ have at most two distinct eigenvalues such as Pauli matrices in (4), N_k are sufficient statistics for the distribution of (R_{k1}, R_{k2}), and hence $\mathcal{P}_{2n}$ and $\mathcal{Q}_{2n}$ are equivalent, that is, $\Delta(\mathcal{P}_{2n}, \mathcal{Q}_{2n}) = 0$, $\Delta(\mathcal{P}_{1n}, \mathcal{P}_{2n}) = \Delta(\mathcal{P}_{1n}, \mathcal{Q}_{2n})$, and $\mathcal{P}_{1n}$ and $\mathcal{Q}_{2n}$ can still be asymptotically equivalent. In summary, generally trace regression can be asymptotically equivalent to quantum state tomography with summarized measurements but not with individual measurements. In fact, the individual measurements $(R_{k1}, \ldots, R_{km})$, $k = 1, \ldots, n$, from quantum state tomography contain information about $\operatorname{tr}(\mathbf{Q}_{ja}\boldsymbol{\rho})$, $a = 1, \ldots, r_j$, while observations Y_k, $k = 1, \ldots, n$, from trace regression have information only about $\operatorname{tr}(\mathbf{B}_j\boldsymbol{\rho})$. From (10) we get $\operatorname{tr}(\mathbf{B}_j\boldsymbol{\rho}) = \sum_{a=1}^{r_j} \lambda_{ja} \operatorname{tr}(\mathbf{Q}_{ja}\boldsymbol{\rho})$, so the individual measurements $(R_{k1}, \ldots, R_{km})$ from quantum state tomography may be more informative than observations Y_k from trace regression for statistical inference of $\boldsymbol{\rho}$. To match quantum state tomography with individual measurements, we may introduce a fine scale trace regression model and treat trace regression (1) as a coarse scale model aggregated from the fine scale model as follows. Suppose that matrix input variable $\mathbf{X}_k$ has the following spectral decomposition:

$$\mathbf{X}_k = \sum_{a=1}^{r_k^X} \lambda_{ka}^X \mathbf{Q}_{ka}^X, \tag{28}$$

where λ_{ka}^X are r_k^X real distinct eigenvalues of $\mathbf{X}_k$, and $\mathbf{Q}_{ka}^X$ are the projections onto the eigen-spaces corresponding to λ_{ka}^X. The fine scale trace regression model assumes that observed random pairs $(\mathbf{Q}_{ka}^X, y_{ka})$ obey

$$y_{ka} = \operatorname{tr}\big(\mathbf{Q}_{ka}^X\boldsymbol{\rho}\big) + z_{ka}, \qquad k = 1, \ldots, n, a = 1, \ldots, r_k^X, \tag{29}$$

where z_{ka} are random errors with mean zero.

Models (1) and (29) are trace regression at two different scales and connected through (28) and the following aggregation relations:

$$Y_k = \sum_{a=1}^{r_k^X} \lambda_{ka}^X y_{ka}, \qquad \varepsilon_k = \sum_{a=1}^{r_k^X} \lambda_{ka}^X z_{ka}, \qquad \operatorname{tr}(\mathbf{X}_k \boldsymbol{\rho}) = \sum_{a=1}^{r_k^X} \lambda_{ka}^X \operatorname{tr}(\mathbf{Q}_{ka}^X \boldsymbol{\rho}). \tag{30}$$

The fine scale trace regression model specified by (29) matches quantum state tomography with the individual measurements $(\mathbf{M}_k, R_{k1}, \ldots, R_{km})$, $k = 1, \ldots, n$. Indeed, as (28) indicates a one to one correspondence between $\mathbf{X}_k$ and $\{\lambda_{ka}^X, \mathbf{Q}_{ka}^X, a = 1, \ldots, r_k^X\}$, we replace Y_k by $(y_{k1}, \ldots, y_{kr_k^X})$ and $\mathbb{P}_{1,n,\boldsymbol{\rho}}$ in (17) by the joint distribution, $\mathbb{Q}_{1,n,\boldsymbol{\rho}}$, of $(\mathbf{X}_k, y_{k1}, \ldots, y_{kr_k^X})$, $k = 1, \ldots, n$, for the random design case [or $(y_{k1}, \ldots, y_{kr_k^X})$, $k = 1, \ldots, n$, for the fixed design case], and define the statistical experiment for fine scale trace regression (29) as follows:

$$\mathcal{Q}_{1n} = \{(\mathcal{X}_1, \mathcal{G}_1, \mathbb{Q}_{1,n,\boldsymbol{\rho}}), \boldsymbol{\rho} \in \Theta\}, \tag{31}$$

where measurable space $(\mathcal{X}_1, \mathcal{G}_1)$ is either $(\mathcal{B}^n \times \mathbb{R}^{mn}, \mathcal{F}_{\mathcal{B}}^n \times \mathcal{F}_{\mathbb{R}}^{mn})$ for the random design case or $(\mathbb{R}^{mn}, \mathcal{F}_{\mathbb{R}}^{mn})$ for the fixed design case.

To study the asymptotic equivalence of fine scale trace regression and quantum state tomography with individual measurements, we need to replace condition (C2) by a new condition for fine scale trace regression:

(C2*) Suppose that two models $\mathcal{Q}_{1n}$ and $\mathcal{Q}_{2n}$ are identifiable. For fine scale trace regression (29), random errors $(z_{k1}, \ldots, z_{kr_k^X})$, $k = 1, \ldots, n$, are independent, and given $\mathbf{X}_k$, $(z_{k1}, \ldots, z_{kr_k^X})'$ is a multivariate normal random vector with mean zero and for $a, b = 1, \ldots, r_k^X$, $a \neq b$,

$$\begin{aligned} \operatorname{Var}(z_{ka}|\mathbf{X}_k) &= \frac{1}{m} \operatorname{tr}(\mathbf{Q}_{ka}^X \boldsymbol{\rho})[1 - \operatorname{tr}(\mathbf{Q}_{ka}^X \boldsymbol{\rho})], \\ \operatorname{Cov}(z_{ka}, z_{kb}|\mathbf{X}_k) &= -\frac{1}{m} \operatorname{tr}(\mathbf{Q}_{ka}^X \boldsymbol{\rho}) \operatorname{tr}(\mathbf{Q}_{kb}^X \boldsymbol{\rho}). \end{aligned} \tag{32}$$

We provide bounds on $\Delta(\mathcal{Q}_{1n}, \mathcal{Q}_{2n})$ and establish the asymptotic equivalence of $\mathcal{Q}_{1n}$ and $\mathcal{Q}_{2n}$ in the following theorem.

THEOREM 2. *Assume that conditions* (C1), (C2*) *and* (C3) *are satisfied.*

(a) *For the random design case, we have*

$$\Delta(\mathcal{Q}_{1n}, \mathcal{Q}_{2n}) \leq n\gamma_p + C\left(\frac{n\zeta_p}{m}\right)^{1/2}, \tag{33}$$

where as in Theorem 1, C is a generic constant depending only on (κ, c_0, c_1), integer κ and constants (c_0, c_1) are, respectively, specified in conditions (C1)

and (C3), *and* γ_p *and* ζ_p *are given by* (18) *and* (23), *respectively. In particular, if* $\Pi(j) = \Xi(j) = 1/p$ *for* $j = 1, \ldots, p$, *then*

$$\Delta(\mathcal{Q}_{1n}, \mathcal{Q}_{2n}) \le C\left(\frac{n\zeta_p}{m}\right)^{1/2}, \tag{34}$$

where ζ_p *is given by* (25).

(b) *For the fixed design case, we have*

$$\Delta(\mathcal{Q}_{1n}, \mathcal{Q}_{2n}) \le C\left(\frac{n\zeta_p}{m}\right)^{1/2}, \tag{35}$$

where C *is the same as in* (a), *and* ζ_p *is given by* (25).

REMARK 4. For quantum state tomography we regard summarized measurements and individual measurements as quantum measurements at coarse and fine scales, respectively. Then Theorems 1 and 2 show that quantum state tomography and trace regression are asymptotically equivalent at both coarse and fine scales. Moreover, as measurements at the coarse scale are aggregated from measurements at the fine scale for both quantum state tomography and trace regression, their asymptotic equivalence at the coarse scale is a consequence of their asymptotic equivalence at the fine scale. Specifically, the deficiency distance bounds in (33)–(35) of Theorem 2 are derived essentially from the deficiency distance between n independent multinomial distributions in quantum state tomography and their corresponding multivariate normal distributions in fine scale trace regression, and the deficiency distance bounds in (22), (24) and (26) of Theorem 1 are the consequences of corresponding bounds in Theorem 2. Fine scale trace regression (29) and condition (C2*) indicate that for each k, $(y_{k1}, \ldots, y_{kr_k^x})$ follows a multivariate normal distribution. From (8) and (13)–(16) we see that given $\mathbf{M}_k$, $(R_{k1}, \ldots, R_{km})$ is jointly determined by the counts of $R_{k1}, \ldots, R_{km}$ taking the eigenvalues of $\mathbf{M}_k$, and the counts jointly follow a multinomial distribution, with mean and covariance matching with those of $m(y_{k1}, \ldots, y_{kr_k^x})$. To prove Theorems 1 and 2, we need to derive the Hellinger distances of the multivariate normal distributions and their corresponding multinomial distributions with uniform perturbations. Carter (2002) has established a bound on deficiency distance between a multinomial distribution and its corresponding multivariate normal distribution through the total variation distance between the multivariate normal distribution and the multinomial distribution with uniform perturbation. The main purpose of the multinomial deficiency bound in Carter (2002) is the asymptotic equivalence study for density estimation. Consequently, the multinomial distribution in Carter (2002) is allowed to have a large number of cells, with bounded cell probability ratios, and his proof techniques are geared up for managing such a multinomial distribution under total variation distance. Since quantum state tomography involves many independent multinomial distributions all with a small number of cells, Carter's result is

not directly applicable for proving Theorems 1 and 2, nor his approach suitable for the current model setting. To show Theorems 1 and 2, we deal with n independent multinomial distributions in quantum state tomography by deriving the Hellinger distances between the perturbed multinomial distributions and the corresponding multivariate normal distributions, and then we establish bounds on the deficiency distance between quantum state tomography and trace regression at the fine scale. Moreover, from (9), (12) and (30) we derive N_k from the counts of individual measurements $R_{k1}, \ldots, R_{km}$ for quantum state tomography and Y_k from fine scale observations y_{ka} for trace regression by the same aggregation relationship, and (32) implies (19), so bounds on $\Delta(\mathcal{P}_{1n}, \mathcal{P}_{2n})$ can be obtained from those on $\Delta(\mathcal{Q}_{1n}, \mathcal{Q}_{2n})$. Thus, Theorem 1 may be viewed as a consequence of Theorem 2. For more details see the proofs of Theorems 1 and 2 in Section 7.

5. Sparse density matrices. Since all deficiency distance bounds in Theorems 1 and 2 depend on ζ_p, we further investigate ζ_p for two special classes of density matrices: sparse density matrices in this section and low rank density matrices in Section 6.

COROLLARY 1. *Denote by Θ_s a collection of density matrices with at most s nonzero entries, where s is an integer. Assume that $\mathcal{B}$ is selected as basis* (3), *and* $\Pi(j) = \Xi(j) = 1/p$. *Then*

$$\zeta_p = \max_{\boldsymbol{\rho} \in \Theta_s} \left\{ \frac{1}{p} \sum_{j=1}^{p} 1\big(|\mathcal{I}_j(\boldsymbol{\rho})| \geq 2\big) \right\} \leq \frac{s_d}{d},$$

where s_d is the maximum number of nonzero diagonal entries of $\boldsymbol{\rho}$ over Θ_s. Furthermore, if conditions (C1), (C2), (C2*) *and* (C3) *are satisfied, we have*

$$\Delta(\mathcal{P}_{1n}, \mathcal{P}_{2n}) \leq C\left(\frac{ns_d}{md}\right)^{1/2}, \qquad \Delta(\mathcal{Q}_{1n}, \mathcal{Q}_{2n}) \leq C\left(\frac{ns_d}{md}\right)^{1/2},$$

where C is the same generic constant as in Theorems 1 and 2.

REMARK 5. Since $p = d^2$, $s_d \leq s$, and the deficiency distance bounds in Corollary 1 are of order $[ns_d/(md)]^{1/2}$, if s_d/d goes to zero as $d \to \infty$, we may have that as $m, n, d \to \infty$, $ns_d/(md) \to 0$ and hence the asymptotic equivalence of quantum state tomography and trace regression, while n/m may not necessarily go to zero. Thus, even though sparsity is not required in the asymptotic equivalence of quantum state tomography and trace regression, Corollary 1 shows that with the sparsity the asymptotic equivalence is much easier to achieve. For example, consider the case that s_d is bounded, and n is of order d^2 (suggested by the bounded κ and the identifiability discussion at the end of Section 2.3). In this case the deficiency distance bounds in Corollary 1 are of order $(d/m)^{1/2}$, and we obtain the asymptotic equivalence of quantum state tomography and trace regression, if $d/m \to 0$ with an example $d = O(m/\log m)$.

We illustrate below that the sparse density matrices studied in Corollary 1 have a sparse representation under basis (3). In general, assume that $\mathcal{B}$ is an orthogonal basis for complex Hermitian matrices. Then every density matrix $\boldsymbol{\rho}$ has a representation under the basis $\mathcal{B}$,

$$\boldsymbol{\rho} = \sum_{j=1}^{p} \alpha_j \mathbf{B}_j, \tag{36}$$

where α_j are coefficients. We say a density matrix $\boldsymbol{\rho}$ is s-sparse under the basis $\mathcal{B}$, if the representation (36) of $\boldsymbol{\rho}$ under the basis $\mathcal{B}$ has at most s nonzero coefficients α_j. The sparsity definition via representation (36) is in line with the vector sparsity concept through orthogonal expansion in compressed sensing. It is easy to see that a density matrix $\boldsymbol{\rho}$ with at most s nonzero entries is the same as that $\boldsymbol{\rho}$ is s-sparse under basis (3). However, a s-sparse matrix under the Pauli basis (4) may have more than s nonzero entries. In fact, it may have up to sd nonzero entries. The following corollary exhibits the different behavior of ζ_p for sparse density matrices under the Pauli basis.

COROLLARY 2. *Denote by Θ_s^p the class of all density matrices that are s-sparse under the Pauli basis, where s is an integer. Assume that $\mathcal{B}$ is selected as the Pauli basis* (4), *and $\Pi(j) = \Xi(j) = 1/p$. Then*

$$1 \geq \zeta_p = \max_{\boldsymbol{\rho} \in \Theta_s^p} \left\{ \frac{1}{p} \sum_{j=1}^{p} 1\big(|\mathcal{I}_j(\boldsymbol{\rho})| \geq 2\big) \right\} \geq 1 - \frac{1}{p}.$$

Furthermore, if conditions (C1), (C2), (C2*) *and* (C3) *are satisfied, we have*

$$\Delta(\mathcal{P}_{1n}, \mathcal{P}_{2n}) \leq C\left(\frac{n}{m}\right)^{1/2}, \qquad \Delta(\mathcal{Q}_{1n}, \mathcal{Q}_{2n}) \leq C\left(\frac{n}{m}\right)^{1/2},$$

where C is the same generic constant as in Theorems 1 and 2.

REMARK 6. Corollary 1 shows that for sparse matrices under basis (3), as $d \to \infty$, if $s_p/d \to 0$, ζ_p goes to zero, and hence the sparsity enables us to establish the asymptotic equivalence of quantum state tomography and trace regression under weaker conditions on m and n. However, Corollary 2 demonstrates that ζ_p does not go to zero for sparse matrices under the Pauli basis. Corollary 1 indicates that for a density matrix with s nonzero entries, in order to have small s_p/d, we must make its nonzero diagonal entries as less as possible. The Pauli basis is the worst in a sense that a sparse matrix under the Pauli basis has at least d nonzero entries, and the Pauli basis tends to put many nonzero entries on the diagonal. From Corollaries 1 and 2 we see that ζ_p depends on sparsity of the density matrix class, but more importantly it is determined by how the sparsity is specified by $\mathcal{B}$.

6. Low rank density matrices. Consider the case of low rank density matrices. Assume density matrix $\boldsymbol{\rho}$ has rank at most r, where $r \ll d$. Then $\boldsymbol{\rho}$ has at most r nonzero eigenvalues, and thus its positive eigenvalues are sparse. The following corollary derives the behavior of ζ_p for low rank density matrices and the Pauli basis.

COROLLARY 3. *Denote by Θ_r the collection of all density matrices $\boldsymbol{\rho}$ with rank up to $r \ll d$. Assume that $\mathcal{B}$ is the Pauli basis* (4)*, and $\Pi(j) = \Xi(j) = 1/p$. Then*

$$1 \geq \zeta_p = \max_{\boldsymbol{\rho} \in \Theta_r} \left\{ \frac{1}{p} \sum_{j=1}^{p} 1\big(|\mathcal{I}_j(\boldsymbol{\rho})| \geq 2\big) \right\} \geq 1 - \frac{1}{p}.$$

Furthermore, if conditions (C1), (C2), (C2*) *and* (C3) *are satisfied, we have*

$$\Delta(\mathcal{P}_{1n}, \mathcal{P}_{2n}) \leq C\left(\frac{n}{m}\right)^{1/2}, \qquad \Delta(\mathcal{Q}_{1n}, \mathcal{Q}_{2n}) \leq C\left(\frac{n}{m}\right)^{1/2},$$

where C is the same generic constant as in Theorems 1 *and* 2.

We construct a low rank density matrix class and matrix set for which ζ_p goes to zero in the following corollary.

COROLLARY 4. *Suppose that $\mathbf{g}_1, \ldots, \mathbf{g}_d$ form an orthonormal basis in $\mathbb{R}^d$, and*

$$\mathcal{B} = \Big\{ \mathbf{g}_\ell \mathbf{g}_\ell', \frac{1}{\sqrt{2}}(\mathbf{g}_{\ell_1}\mathbf{g}_{\ell_2}' + \mathbf{g}_{\ell_2}\mathbf{g}_{\ell_1}'), \frac{\sqrt{-1}}{\sqrt{2}}(\mathbf{g}_{\ell_2}\mathbf{g}_{\ell_1}' - \mathbf{g}_{\ell_1}\mathbf{g}_{\ell_2}'), \\ \ell, \ell_1, \ell_2 = 1, \ldots, d, \ell_1 < \ell_2 \Big\}.$$

Assume that $\gamma \ll d$ and $r \ll d$ are integers. Denote by $\Theta_{r\gamma}$ a collection of density matrices $\boldsymbol{\rho}$ with the form

$$\boldsymbol{\rho} = \sum_{j=1}^{r} \xi_j U_j U_j^\dagger, \tag{37}$$

where $\xi_j \geq 0$, $\xi_1 + \cdots + \xi_r = 1$, and U_j are unit vectors in $\mathbb{C}^d$ whose real and imaginary parts are linear combinations of $\mathbf{g}_{\ell_1}, \ldots, \mathbf{g}_{\ell_k}$, $1 \leq \ell_1, \ldots, \ell_k \leq d$ and $1 \leq k \leq \gamma$. Assume $\Pi(j) = \Xi(j) = 1/p$. Then

$$\zeta_p = \max_{\boldsymbol{\rho} \in \Theta_{r\gamma}} \left\{ \frac{1}{p} \sum_{j=1}^{p} 1\big(|\mathcal{I}_j(\boldsymbol{\rho})| \geq 2\big) \right\} \leq \frac{2r\gamma(4\gamma + 1)}{p}.$$

Furthermore, if conditions (C1), (C2), (C2*) *and* (C3) *are satisfied, we have*

$$\Delta(\mathcal{P}_{1n}, \mathcal{P}_{2n}) \leq C\left(\frac{nr\gamma^2}{mp}\right)^{1/2}, \qquad \Delta(\mathcal{Q}_{1n}, \mathcal{Q}_{2n}) \leq C\left(\frac{nr\gamma^2}{mp}\right)^{1/2},$$

where C is the same generic constant as in Theorems 1 and 2.

REMARK 7. It is known that a density matrix of rank up to r has representation (37), and matrix $\boldsymbol{\rho}$ with representation (37) has rank at most r. Corollary 3 shows that for the class of density matrices with rank at most r, ζ_p does not go to zero under the Pauli basis. Corollary 4 constructs a basis $\mathcal{B}$ and a subclass of low rank density matrices, for which ζ_p can go to zero, and the deficiency distance bounds are of order $[nr\gamma^2/(mp)]^{1/2}$. Since $r, \gamma \ll d$ and $p = d^2$, $r\gamma^2/p$ may go to zero very fast as $d \to \infty$. As $m, n, d \to \infty$, if $nr\gamma^2/(mp) \to 0$, we obtain the asymptotic equivalence of quantum state tomography and trace regression. For example, consider the case that r and γ are bounded, and n is of order d^2 (suggested by the bounded κ and the identifiability discussion at the end of Section 2.3). In this case the deficiency distance bounds in Corollary 4 are of order $m^{-1/2}$, and we conclude that if $m \to \infty$, the two models are asymptotically equivalent for any (n, d) compatible with the model identifiability condition. A particular example is that $n = d^2$ and d grows exponentially faster than m.

REMARK 8. The low rank condition $r \ll d$ on a density matrix indicates that it has a relatively small number of positive eigenvalues, that is, its positive eigenvalues are sparse. We may also explain the condition on the eigenvectors U_j in (37) via sparsity as follows. Since $\{\mathbf{g}_1, \ldots, \mathbf{g}_d\}$ is an orthonormal basis in $\mathbb{R}^d$, the real part, $\mathrm{Re}(U_j)$, and imaginary part, $\mathrm{Im}(U_j)$, of U_j have the following expansions under the basis:

$$\mathrm{Re}(U_j) = \sum_{\ell=1}^{d} \alpha_{1\ell}^j \mathbf{g}_\ell, \qquad \mathrm{Im}(U_j) = \sum_{\ell=1}^{d} \alpha_{2\ell}^j \mathbf{g}_\ell, \tag{38}$$

where $\alpha_{1\ell}^j$ and $\alpha_{2\ell}^j$ are coefficients. Then a low rank density matrix with representation (37) belongs to $\Theta_{r\gamma}$, if for $j = 1, \ldots, r$, $\{\ell, \alpha_{1\ell}^j \neq 0\}$ and $\{\ell, \alpha_{2\ell}^j \neq 0\}$ have cardinality at most γ, that is, there are at most γ nonzero coefficients in the expansions (38). As $\gamma \ll d$, the eigenvectors U_j have sparse representations. Thus, the subclass $\Theta_{r\gamma}$ of density matrices imposes some sparsity conditions on not only the eigenvalues but also the eigenvectors of its members. In fact, Witten, Tibshirani and Hastie (2009) indicates that we need some sparsity on both eigenvalues and eigenvectors for estimating large matrices. An important class of quantum states are pure states, which correspond to density matrices of rank one. In order to have a pure state in $\Theta_{r\gamma}$, its eigenvector U_1 corresponding to eigenvalue 1 must be a liner combination of at most γ basis vectors $\mathbf{g}_\ell$. Such a requirement can be met for a large class of pure states through the selection of proper γ and suitable bases in $\mathbb{R}^d$. It is interesting to see that matrices themselves in $\Theta_{r\gamma}$ of Corollary 4 may not be sparse. For example, taking $\mathbf{g}_1, \ldots, \mathbf{g}_d$ as the Haar basis in $\mathbb{R}^d$ [see Vidakovic (1999)], we obtain that rank one matrix $\boldsymbol{\rho} = (1, 1, \ldots, 1)'(1, 1, \ldots, 1)/d$ and rank two matrix $\boldsymbol{\rho} = 3(1, 1, \ldots, 1)'(1, 1, \ldots, 1)/$

$(4d) + (1, \ldots, 1, -1, \ldots, -1)'(1, \ldots, 1, -1, \ldots, -1)/(4d)$, which are inside $\Theta_{r\gamma}$ for $(r, \gamma) = (1, 1)$ and $(r, \gamma) = (2, 2)$, respectively, but not sparse.

REMARK 9. From Corollaries 1–4, we see that whether ζ_p goes to zero or not is largely dictated by $\mathcal{B}$ used in the two models. As we discussed in Remarks 5 and 7, for certain classes of sparse or low rank density matrices, ζ_p goes to zero, and we can achieve the asymptotic equivalence of quantum state tomography and trace regression when d is comparable to or exceeds m. In particular for a special subclass of low rank density matrices we can obtain the asymptotic equivalence even when d grows exponentially faster than m. We should emphasize that the claimed asymptotic equivalences in the ultra high dimension setting are under some sparse circumstances for which ζ_p goes to zero, that is, of the p multinomial distributions in the quantum state tomography model, a relatively small number of multinomial distributions are nondegenerate, and similarly, the trace regression model as the approximating normal experiment consists of the same small number of corresponding nondegenerate normal distributions. In other words, the asymptotic equivalence in ultra high dimensions may be interpreted as the approximation of a sparse quantum state tomography model by a sparse Gaussian trace regression model. This is the first asymptotic equivalence result in ultra high dimensions. It leads us to speculate that sparse Gaussian experiments may play an important role in the study of asymptotic equivalence in the ultra high dimension setting.

7. Proofs.

7.1. *Basic facts and technical lemmas.* We need some basic results about the Markov kernel method which are often used to bound $\delta(\mathcal{P}_{2n}, \mathcal{P}_{1n})$ and prove asymptotic equivalence of $\mathcal{P}_{1n}$ and $\mathcal{P}_{2n}$ [see Le Cam (1986) and Le Cam and Yang (2000)]. A Markov kernel $K(\omega, A)$ is defined for $\omega \in \mathcal{X}_2$ and $A \in \mathcal{G}_1$ such that for a given $\omega \in \mathcal{X}_2$, $K(\omega, \cdot)$ is a probability measure on the σ-field $\mathcal{G}_1$, and for a fixed $A \in \mathcal{G}_1$, $K(\cdot, A)$ is a measurable function on $\mathcal{X}_2$. The Markov kernel maps any $\mathbb{P}_{2,n,\boldsymbol{\rho}} \in \mathcal{P}_{2n}$ into another probability measure $[K(\mathbb{P}_{2,n,\boldsymbol{\rho}})](A) = \int K(\omega, A)\mathbb{P}_{2,n,\boldsymbol{\rho}}(d\omega) \in \mathcal{P}_{1n}$. We have the following result:

$$\delta(\mathcal{P}_{2n}, \mathcal{P}_{1n}) \le \inf_K \sup_{\boldsymbol{\rho}\in\Theta} \|\mathbb{P}_{1,n,\boldsymbol{\rho}} - K(\mathbb{P}_{2,n,\boldsymbol{\rho}})\|_{\mathrm{TV}}, \tag{39}$$

where the infimum is over all Markov kernels, and $\|\cdot\|_{\mathrm{TV}}$ is the total variation norm.

We often use the Hellinger distance to bound total variation norm and handle product probability measures. For two probability measures P and Q on a common measurable space, we define the Hellinger distance

$$H^2(P, Q) = \int \left|\sqrt{\frac{dP}{d\mu}} - \sqrt{\frac{dQ}{d\mu}}\right|^2 d\mu, \tag{40}$$

where μ is any measure that dominates P and Q, and if P and Q are equivalent,

$$H^2(P, Q) = 2 - 2E_P\left[\sqrt{\frac{dQ}{dP}}\right], \tag{41}$$

where E_P denotes expectation under P. We have

$$\|P - Q\|_{\mathrm{TV}} \le H(P, Q), \tag{42}$$

and for any event A,

$$\begin{aligned} H^2(P, Q) &\le 2 - 2E_P\left[1_A\sqrt{\frac{dQ}{dP}}\right] = 2P(A^c) + 2E_P\left[1_A\left(1 - \sqrt{\frac{dQ}{dP}}\right)\right] \\ &\le 2P(A^c) + E_P\left[1_A \log\frac{dP}{dQ}\right], \end{aligned} \tag{43}$$

where the last inequality is from the fact that $x - 1 \ge \log x$ for any $x > 0$.

Carter (2002) has established an asymptotic equivalence of a multinomial distribution and its corresponding multivariate normal distribution through bounding the total variation distance between the multivariate normal distribution and the multinomial distribution with uniform perturbation. The approach in Carter (2002) is to break dependence in the multinomial distribution and create independence by successively conditioning on pairs and thus establish a bound on the total variation distance of the perturbed multinomial distribution and the multivariate normal distribution. Carter (2002) works for the multinomial distribution with a large number of cells, while quantum state tomography involves many independent multinomial distributions all with a small number of cells. To handle the many small independent multinomial distributions for quantum state tomography and prove Theorems 1 and 2, we need to derive the Hellinger distances between the perturbed multinomial distributions and multivariate normal distributions instead of total variation distance. Carter's approach is geared up for total variation distance and the result cannot be directly used to prove Theorems 1 and 2. Our approach to proving Lemma 2 below is to directly decompose a multinomial distribution as products of conditional distributions and then establish a bound on the Hellinger distance between the perturbed multinomial distribution and its corresponding multivariate normal distribution.

Denote by C a generic constant whose value may change from appearance to appearance. The value of C may depends on fixed constants (κ, c_0, c_1) given by conditions (C1) and (C3) but is free of (m, n, d, p) and individual $\boldsymbol{\rho}$.

First, we describe a known result between binomial and normal distributions [see Carter (2002), B2 of the Appendix].

LEMMA 1. *Suppose that P is a binomial distribution* $\mathrm{Bin}(m, \theta)$ *with* $\theta \in (0, 1)$, *and Q is a normal distribution with mean $m\theta$ and variance $m\theta(1 - \theta)$.*

Let P^* be the convolution distribution of P and an independent uniform distribution on $(-1/2, 1/2)$. Then

$$P^*(A^c) \le \exp(-Cm^{1/3}), \qquad E_{P^*}\left[1(A)\log\frac{dP^*}{dQ}\right] \le \frac{C}{m\theta(1-\theta)},$$

where $A = \{|U - m\theta| \le m[\theta(1-\theta)]^{2/3}\}$, and random variable U has the distribution P.

We give bounds on the Hellinger distances between the perturbed multinomial distributions and their corresponding multivariate normal distributions in next two lemmas whose proofs are collected in the Appendix.

LEMMA 2. *Suppose that P is a multinomial distribution $\mathcal{M}(m, \theta_1, \dots, \theta_r)$, where $r \ge 2$ is a fixed integer,*

$$\theta_1 + \cdots + \theta_r = 1, \qquad c_0 \le \min(\theta_1, \dots, \theta_r) \le \max(\theta_1, \dots, \theta_r) \le c_1$$

and $0 < c_0 \le c_1 < 1$ are two fixed constants. Denote by Q the multivariate normal distribution whose mean and covariance are the same as P. Let P^* be the convolution of the distribution P and the distribution of $(\psi_1, \dots, \psi_r)$, where $\psi_1, \dots, \psi_{r-1}$ are independent and follow a uniform distribution on $(-1/2, 1/2)$, and $\psi_r = -\psi_1 - \cdots - \psi_{r-1}$. Then

$$H(P^*, Q) \le r^2 \exp(-Cm^{1/3}) + \frac{Cr}{\sqrt{m}}.$$

LEMMA 3. *Suppose that for $k = 1, \dots, n$, P_k is a multinomial distribution $\mathcal{M}(m, \theta_{k1}, \dots, \theta_{k\nu_k})$, where $\nu_k \le \kappa$, κ is a fixed integer, $\theta_{k1} + \cdots + \theta_{k\nu_k} = 1$, and for constants c_0 and c_1,*

$$0 < c_0 \le \min(\theta_{k1}, \dots, \theta_{k\nu_k}) \le \max(\theta_{k1}, \dots, \theta_{k\nu_k}) \le c_1 < 1.$$

Denote by Q_k the multivariate normal distribution whose mean and covariance are the same as P_k. If $\nu_k \ge 2$, following the same way as in Lemma 2 we define P_k^* as the convolution of P_k and an independent uniform distribution on $(-1/2, 1/2)$, and if $\nu_k \le 1$ let $P_k^* = P_k$. Assume that P_k, P_k^*, Q_k for different k are independent, and define product probability measures

$$P = \prod_{k=1}^n P_k, \qquad P^* = \prod_{k=1}^n P_k^*, \qquad Q = \prod_{k=1}^n Q_k.$$

Then we have

$$H^2(P^*, Q) \le \frac{C\kappa^2}{m}\sum_{k=1}^n 1(\nu_k \ge 2).$$

We need the following lemma on total variation distance of two joint distributions whose proof is in the Appendix.

LEMMA 4. *Suppose that U_1 and V_1 are discrete random variables, and random variables (U_1, U_2) and (V_1, V_2) have joint distributions F and G, respectively. Let $F(u_1, u_2) = F_1(u_1) \times F_{2|1}(u_2|u_1)$ and $G(v_1, v_2) = G_1(v_1) \times G_{2|1}(v_2|v_1)$, where F_1 and G_1 are the respective marginal distributions of U_1 and V_1, and $F_{2|1}$ and $G_{2|1}$ are the conditional distributions of U_2 given U_1 and V_2 given V_1, respectively. Then*

$$\begin{aligned}\|F - G\|_{\mathrm{TV}} \le \max_x \left|1 - \frac{P(U_1 = x)}{P(V_1 = x)}\right| \\ + E_{F_1}\big[\|F_{2|1}(\cdot|U_1) - G_{2|1}(\cdot|V_1)\|_{\mathrm{TV}}|U_1 = V_1\big],\end{aligned} \tag{44}$$

where E_{F_1} denotes expectation under F_1, $\|F_{2|1}(\cdot|U_1) - G_{2|1}(\cdot|V_1)\|_{\mathrm{TV}}$ denotes the total variation norm of the difference of the two conditional distributions $F_{2|1}$ and $G_{2|1}$, and the value of the second term on the right-hand side of (44) is clearly specified as follows:

$$\begin{aligned}&E_{F_1}\big[\|F_{2|1}(\cdot|U_1) - G_{2|1}(\cdot|V_1)\|_{\mathrm{TV}}|U_1 = V_1\big] \\ &\quad = \sum_x \|F_{2|1}(\cdot|x) - G_{2|1}(\cdot|x)\|_{\mathrm{TV}} P(U_1 = x).\end{aligned}$$

7.2. *Proofs of Theorems 1 and 2.*

PROOF OF THEOREM 1. Denote by $\mathbb{P}^k_{1,n,\boldsymbol{\rho}}$ the distribution of $(\mathbf{X}_k, Y_k)$ and $\mathbb{P}^k_{2,n,\boldsymbol{\rho}}$ the distribution of $(\mathbf{M}_k, N_k)$, $k = 1, \ldots, n$. For different k, $(\mathbf{X}_k, Y_k)$ from trace regression are independent, and $(\mathbf{M}_k, N_k)$ from quantum state tomography are independent, so $\mathbb{P}^k_{1,n,\boldsymbol{\rho}}$ and $\mathbb{P}^k_{2,n,\boldsymbol{\rho}}$ for different k are independent, and

$$\mathbb{P}_{1,n,\boldsymbol{\rho}} = \prod_{k=1}^n \mathbb{P}^k_{1,n,\boldsymbol{\rho}}, \qquad \mathbb{P}_{2,n,\boldsymbol{\rho}} = \prod_{k=1}^n \mathbb{P}^k_{2,n,\boldsymbol{\rho}}, \tag{45}$$

where $\mathbb{P}_{1,n,\boldsymbol{\rho}}$ and $\mathbb{P}_{2,n,\boldsymbol{\rho}}$ are given in (17).

Suppose that $\mathbf{M}_k$ has ν_k different eigenvalues, and let $U_{ka} = \sum_{\ell=1}^m 1(R_{k\ell} = \lambda_{ka})$, $a = 1, \ldots, \nu_k$, and $\mathbf{U}_k = (U_{k1}, \ldots, U_{k\nu_k})'$. Denote by $\mathbb{Q}^k_{2,n,\boldsymbol{\rho}}$ the distribution of $(\mathbf{M}_k, \mathbf{U}_k)$. If $\nu_k \ge 2$, we let $\mathbb{Q}^{k*}_{2,n,\boldsymbol{\rho}}$ be the distribution of $(\mathbf{M}_k, \mathbf{U}^*_k)$, where $\mathbf{U}^*_k = (U^*_{k1}, \ldots, U^*_{k\nu_k})'$, U^*_{ka} is equal to U_{ka} plus an independent uniform random variable on $(-1/2, 1/2)$, $a = 1, \ldots, \nu_k - 1$ and $U^*_{k\nu_k} = m - U^*_{k1} - \cdots - U^*_{k,\nu_k-1}$. Note that $\mathbb{P}^k_{2,n,\boldsymbol{\rho}}$ is the distribution of $(\mathbf{M}_k, N_k)$, and

$$N_k = (R_{k1} + \cdots + R_{km})/m = (\lambda_{k1}U_{k1} + \cdots + \lambda_{k\nu_k}U_{k\nu_k})/m. \tag{46}$$

Analog to the expression (46) of N_k in terms of $\mathbf{U}_k = (U_{k1}, \ldots, U_{km})'$, we define

$$N_k^* = \big(\lambda_{k1} U_{k1}^* + \cdots + \lambda_{k\nu_k} U_{k\nu_k}^*\big)/m, \tag{47}$$

and denote by $\mathbb{P}_{2,n,\boldsymbol{\rho}}^{k*}$ the distribution of $(\mathbf{M}_k, N_k^*)$. If $\nu_k \le 1$, let $\mathbb{Q}_{2,n,\boldsymbol{\rho}}^{k*} = \mathbb{Q}_{2,n,\boldsymbol{\rho}}^{k}$ and $\mathbb{P}_{2,n,\boldsymbol{\rho}}^{k*} = \mathbb{P}_{2,n,\boldsymbol{\rho}}^{k}$. As $\mathbb{Q}_{2,n,\boldsymbol{\rho}}^{k}$, $\mathbb{Q}_{2,n,\boldsymbol{\rho}}^{k*}$, and $\mathbb{P}_{2,n,\boldsymbol{\rho}}^{k*}$ for different k are independent, define their product probability measures

$$\mathbb{Q}_{2,n,\boldsymbol{\rho}} = \prod_{k=1}^{n} \mathbb{Q}_{2,n,\boldsymbol{\rho}}^{k}, \qquad \mathbb{Q}_{2,n,\boldsymbol{\rho}}^{*} = \prod_{k=1}^{n} \mathbb{Q}_{2,n,\boldsymbol{\rho}}^{k*}, \qquad \mathbb{P}_{2,n,\boldsymbol{\rho}}^{*} = \prod_{k=1}^{n} \mathbb{P}_{2,n,\boldsymbol{\rho}}^{k*}. \tag{48}$$

Note that, since $\mathbf{U}_k$ and $(R_{k1}, \ldots, R_{km})$ have a one to one correspondence, and the two statistical experiments formed by the distribution of $(\mathbf{M}_k, \mathbf{U}_k)$ and the distribution of $(\mathbf{M}_k, R_{k1}, \ldots, R_{km})$ have zero deficiency distance, without confusion we abuse the notation $\mathbb{Q}_{2,n,\boldsymbol{\rho}}$ by using it here for the joint distribution of $(\mathbf{M}_k, \mathbf{U}_k)$, $k = 1, \ldots, n$, as well as in (27) for the joint distribution of $(\mathbf{M}_k, R_{k1}, \ldots, R_{km})$, $k = 1, \ldots, n$.

Given $\mathbf{M}_k = \mathbf{B}_{j_k}$, let $\nu_k = r_{j_k}$, and $\mathbf{U}_k = (U_{k1}, \ldots, U_{kr_{j_k}})'$ follows a multinomial distribution $\mathcal{M}(m, \operatorname{tr}(\mathbf{Q}_{j_k1}\boldsymbol{\rho}), \ldots, \operatorname{tr}(\mathbf{Q}_{j_kr_{j_k}}\boldsymbol{\rho}))$, where r_j and $\mathbf{Q}_{ja}$ are defined in (10), and

$$\begin{aligned}
E(U_{ka}|\mathbf{M}_k = \mathbf{B}_{j_k}) &= m \operatorname{tr}(\mathbf{Q}_{j_ka}\boldsymbol{\rho}),\\
\operatorname{Var}(U_{ka}|\mathbf{M}_k = \mathbf{B}_{j_k}) &= m \operatorname{tr}(\mathbf{Q}_{j_ka}\boldsymbol{\rho})\big[1 - \operatorname{tr}(\mathbf{Q}_{j_ka}\boldsymbol{\rho})\big],\\
\operatorname{Cov}(U_{ka}, U_{kb}|\mathbf{M}_k = \mathbf{B}_{j_k}) &= -m \operatorname{tr}(\mathbf{Q}_{j_ka}\boldsymbol{\rho}) \operatorname{tr}(\mathbf{Q}_{j_kb}\boldsymbol{\rho}),\\
&\qquad a \ne b, a, b = 1, \ldots, r_{j_k}.
\end{aligned}$$

Then

$$\begin{aligned}
E(N_k|\mathbf{M}_k = \mathbf{B}_{j_k}) &= \sum_{a=1}^{r_{j_k}} \lambda_{j_ka} \operatorname{tr}(\mathbf{Q}_{j_ka}\boldsymbol{\rho}) = \operatorname{tr}(\mathbf{B}_{j_k}\boldsymbol{\rho}) = \operatorname{tr}(\mathbf{M}_k\boldsymbol{\rho}),\\
\operatorname{Var}(N_k|\mathbf{M}_k = \mathbf{B}_{j_k}) &= \frac{1}{m}\sum_{a=1}^{r_{j_k}} \lambda_{j_ka}^2 \operatorname{tr}(\mathbf{Q}_{j_ka}\boldsymbol{\rho})\big[1 - \operatorname{tr}(\mathbf{Q}_{j_ka}\boldsymbol{\rho})\big]\\
&\quad - \frac{2}{m}\sum_{a=1}^{r_{j_k}}\sum_{b=a+1}^{r_{j_k}} \lambda_{j_ka}\lambda_{j_kb} \operatorname{tr}(\mathbf{Q}_{j_ka}\boldsymbol{\rho}) \operatorname{tr}(\mathbf{Q}_{j_kb}\boldsymbol{\rho})\\
&= \frac{1}{m}\big\{\operatorname{tr}\big(\mathbf{B}_{j_k}^2\boldsymbol{\rho}\big) - \big[\operatorname{tr}(\mathbf{B}_{j_k}\boldsymbol{\rho})\big]^2\big\}\\
&= \frac{1}{m}\big\{\operatorname{tr}\big(\mathbf{M}_k^2\boldsymbol{\rho}\big) - \big[\operatorname{tr}(\mathbf{M}_k\boldsymbol{\rho})\big]^2\big\}.
\end{aligned}$$

From (28) and (29), we have that given $\mathbf{X}_k = \mathbf{B}_{j_k}$, $r_k^X = r_{j_k}$, and multivariate normal random vector $\mathbf{V}_k = (V_{k1}, \ldots, V_{kr_{j_k}})' = m(y_{k1}, \ldots, y_{kr_{j_k}})'$ has conditional

mean and conditional covariance matching those of $\mathbf{U}_k = (U_{k1}, \ldots, U_{kr_{j_k}})'$. With $\mathbf{X}_k = \mathbf{B}_{j_k}$ we may rewrite (29) and (30) as follows:

$$
\begin{aligned}
V_{ka} &= m \operatorname{tr}(\mathbf{Q}_{j_k a} \boldsymbol{\rho}) + m z_{ka}, \qquad a = 1, \ldots, r_{j_k}, \\
Y_k &= \frac{1}{m} \sum_{a=1}^{r_{j_k}} \lambda_{ka} V_{ka}, \qquad \varepsilon_k = \sum_{a=1}^{r_{j_k}} \lambda_{ka} z_{ka}.
\end{aligned} \tag{49}
$$

Denote by $\mathbb{Q}^k_{1,n,\boldsymbol{\rho}}$ the distribution of $(\mathbf{X}_k, \mathbf{V}_k)$. Then $\mathbb{Q}^k_{1,n,\boldsymbol{\rho}}$ for different k are independent, and

$$
\mathbb{Q}_{1,n,\boldsymbol{\rho}} = \prod_{k=1}^{n} \mathbb{Q}^k_{1,n,\boldsymbol{\rho}}, \tag{50}
$$

where $\mathbb{Q}_{1,n,\boldsymbol{\rho}}$ is the joint distribution of $(\mathbf{X}_k, V_{k1}, \ldots, V_{kr_k^X})$, $k = 1, \ldots, n$. Note that, since $\mathbf{V}_k = (V_{k1}, \ldots, V_{kr_{j_k}})' = m(y_{k1}, \ldots, y_{kr_{j_k}})'$, and the two statistical experiments formed by the distribution of $(\mathbf{X}_k, V_{k1}, \ldots, V_{kr_{j_k}})$ and the distribution of $(\mathbf{X}_k, y_{k1}, \ldots, y_{kr_{j_k}})$ have zero deficiency distance, without confusion we abuse the notation $\mathbb{Q}_{1,n,\boldsymbol{\rho}}$ by using it here for the joint distribution of $(\mathbf{X}_k, V_{k1}, \ldots, V_{kr_k^X})$, $k = 1, \ldots, n$, as well as in (31) for the joint distribution of $(\mathbf{X}_k, y_{k1}, \ldots, y_{kr_k^X})$, $k = 1, \ldots, n$.

Conditional on $\mathbf{M}_k = \mathbf{B}_{j_k}$, for $k = 1, \ldots, n$, if $|\mathcal{I}_{j_k}(\boldsymbol{\rho})| \le 1$, $\mathbb{Q}^k_{1,n,\boldsymbol{\rho}}$ and $\mathbb{Q}^k_{2,n,\boldsymbol{\rho}}$ are the same degenerate distribution; if $|\mathcal{I}_{j_k}(\boldsymbol{\rho})| \ge 2$, $\mathbb{Q}^k_{2,n,\boldsymbol{\rho}}$ is a multinomial distribution with $\mathbb{Q}^{k*}_{2,n,\boldsymbol{\rho}}$ its uniform perturbation, and $\mathbb{Q}^k_{1,n,\boldsymbol{\rho}}$ is a multivariate normal distribution with mean and covariance matching those of $\mathbb{Q}^k_{2,n,\boldsymbol{\rho}}$. Thus applying Lemma 3, we obtain that given $(\mathbf{X}_1, \ldots, \mathbf{X}_n) = (\mathbf{M}_1, \ldots, \mathbf{M}_n) = (\mathbf{B}_{j_1}, \ldots, \mathbf{B}_{j_n})$,

$$
\|\mathbb{Q}_{1,n,\boldsymbol{\rho}} - \mathbb{Q}^*_{2,n,\boldsymbol{\rho}}\|^2_{\mathrm{TV}} \le H^2(\mathbb{Q}_{1,n,\boldsymbol{\rho}}, \mathbb{Q}^*_{2,n,\boldsymbol{\rho}}) \le \frac{C\kappa^2}{m} \sum_{k=1}^{n} 1(|\mathcal{I}_{j_k}(\boldsymbol{\rho})| \ge 2), \tag{51}
$$

where the first inequality is due to (42). As (47) and (49) imply that N^*_k and Y_k are the same weighted averages of components of $\mathbf{U}^*_k$ and $\mathbf{V}_k$, respectively, $\mathbb{P}_{1,n,\boldsymbol{\rho}}$ and $\mathbb{P}^*_{2,n,\boldsymbol{\rho}}$ are the same respective marginal probability measures of $\mathbb{Q}_{1,n,\boldsymbol{\rho}}$ and $\mathbb{Q}^*_{2,n,\boldsymbol{\rho}}$. Hence, conditional on $(\mathbf{X}_1, \ldots, \mathbf{X}_n) = (\mathbf{M}_1, \ldots, \mathbf{M}_n)$,

$$
\|\mathbb{P}_{1,n,\boldsymbol{\rho}} - \mathbb{P}^*_{2,n,\boldsymbol{\rho}}\|_{\mathrm{TV}} \le \|\mathbb{Q}_{1,n,\boldsymbol{\rho}} - \mathbb{Q}^*_{2,n,\boldsymbol{\rho}}\|_{\mathrm{TV}}. \tag{52}
$$

With $\mathbf{X}_k$ and $\mathbf{M}_k$ are sampled from $\mathcal{B}$ according to distributions Π and Ξ, respectively, we have

$$
\begin{aligned}
&\|\mathbb{P}_{1,n,\boldsymbol{\rho}} - \mathbb{P}^*_{2,n,\boldsymbol{\rho}}\|_{\mathrm{TV}} \\
&\quad \le \max_{1 \le j \le p} \left| 1 - \frac{\Pi^n(j)}{\Xi^n(j)} \right| \\
&\qquad + E_{\Pi}\big(E_{\Pi}\big[\|\mathbb{P}_{1,n,\boldsymbol{\rho}} - \mathbb{P}^*_{2,n,\boldsymbol{\rho}}\|_{\mathrm{TV}} | \mathbf{X}_1 = \mathbf{M}_1, \ldots, \mathbf{X}_n = \mathbf{M}_n\big]\big)
\end{aligned} \tag{53}
$$

$$
\begin{aligned}
&\le n \max_{1\le j\le p}\left|1-\frac{\Pi(j)}{\Xi(j)}\right| \\
&\quad + E_\Pi\big(E_\Pi\big[\big\|\mathbb{Q}_{1,n,\boldsymbol{\rho}} - \mathbb{Q}^*_{2,n,\boldsymbol{\rho}}\big\|_{\mathrm{TV}} \big| \mathbf{X}_1 = \mathbf{M}_1, \ldots, \mathbf{X}_n = \mathbf{M}_n\big]\big) \\
&\le n\gamma_p + \frac{C\kappa}{\sqrt{m}} E_\Pi\left(\left[\sum_{k=1}^n 1\big(|\mathcal{I}_{j_k}(\boldsymbol{\rho})| \ge 2\big)\right]^{1/2}\right) \\
&\le n\gamma_p + \frac{C\kappa}{\sqrt{m}}\left(\sum_{k=1}^n E_\Pi\big[1\big(|\mathcal{I}_{j_k}(\boldsymbol{\rho})| \ge 2\big)\big]\right)^{1/2} \\
&\le n\gamma_p + \frac{C\kappa}{\sqrt{m}}\left(\sum_{k=1}^n \sum_{j=1}^p \Pi(j) 1\big(|\mathcal{I}_j(\boldsymbol{\rho})| \ge 2\big)\right)^{1/2} \\
&= n\gamma_p + \frac{C\kappa}{\sqrt{m}}\left(n\sum_{j=1}^p \Pi(j) 1\big(|\mathcal{I}_j(\boldsymbol{\rho})| \ge 2\big)\right)^{1/2} \\
&\le n\gamma_p + C\kappa\left(\frac{n\zeta_p}{m}\right)^{1/2},
\end{aligned}
$$

where the first three inequalities are, respectively, from Lemma 4, (52) and (51), the fourth inequality is applying Hölder's inequality, and the fifth inequality is due the fact that $\mathbf{X}_k$ and $\mathbf{M}_k$ are the i.i.d. sample from $\mathcal{B}$. Combining (39) and (53), we obtain

$$
\begin{aligned}
\delta(\mathcal{P}_{2n}, \mathcal{P}_{1n}) &\le \inf_K \sup_{\boldsymbol{\rho}\in\Theta} \big\|\mathbb{P}_{1,n,\boldsymbol{\rho}} - K(\mathbb{P}_{2,n,\boldsymbol{\rho}})\big\|_{\mathrm{TV}} \\
&\le \sup_{\boldsymbol{\rho}\in\Theta} \big\|\mathbb{P}_{1,n,\boldsymbol{\rho}} - \mathbb{P}^*_{2,n,\boldsymbol{\rho}}\big\|_{\mathrm{TV}} \qquad (54) \\
&\le n\gamma_p + C\kappa\left(\frac{n\zeta_p}{m}\right)^{1/2}.
\end{aligned}
$$

To bound $\delta(\mathcal{P}_{1n}, \mathcal{P}_{2n})$, we employ a round-off procedure to invert the uniform perturbation used to obtain $\mathbb{Q}^*_{2,n,\boldsymbol{\rho}}$ and $\mathbb{P}^*_{2,n,\boldsymbol{\rho}}$ in (48) [also see Carter (2002), Section 5]. Specifically let $\mathbf{V}^*_k = (V^*_{k1}, \ldots, V^*_{k\nu_k})'$, where V^*_{ka} is a random vector obtained by rounding V_{ka} off to the nearest integer, $a = 1, \ldots, \nu_k - 1$, and $V^*_{k\nu_k} = m - V^*_{k1} - \cdots - V^*_{k,\nu_k-1}$. Denote by $\mathbb{Q}^{k*}_{1,n,\boldsymbol{\rho}}$ the distribution of $(\mathbf{X}_k, \mathbf{V}^*_k)$ and $\mathbb{P}^{k*}_{1,n,\boldsymbol{\rho}}$ the distribution of $(\mathbf{X}_k, (\lambda_{k1}V^*_{k1} + \cdots + \lambda_{k\nu_k}V^*_{k\nu_k})/m)$, and let

$$
\mathbb{Q}^*_{1,n,\boldsymbol{\rho}} = \prod_{k=1}^n \mathbb{Q}^{k*}_{1,n,\boldsymbol{\rho}}, \qquad \mathbb{P}^*_{1,n,\boldsymbol{\rho}} = \prod_{k=1}^n \mathbb{P}^{k*}_{1,n,\boldsymbol{\rho}}. \tag{55}
$$

It is easy to see that for any integer-valued random variable W,

$$
\text{round-off of } \big[W + \text{uniform}(-1/2, 1/2)\big] = W,
$$

and thus the round-off procedure inverts the uniform perturbation procedure. Denote by K_0 and K_1 the uniform perturbation and the round-off procedure, respectively. Then from (48), (50) and (55) we have

$$\begin{aligned} K_1(\mathbb{Q}_{1,n,\boldsymbol{\rho}}) &= \mathbb{Q}^*_{1,n,\boldsymbol{\rho}}, \qquad K_0(\mathbb{Q}_{2,n,\boldsymbol{\rho}}) = \mathbb{Q}^*_{2,n,\boldsymbol{\rho}}, \\ K_1\big[K_0(\mathbb{Q}_{2,n,\boldsymbol{\rho}})\big] &= K_1\big[\mathbb{Q}^*_{2,n,\boldsymbol{\rho}}\big] = \mathbb{Q}_{2,n,\boldsymbol{\rho}}. \end{aligned} \tag{56}$$

From (56), we show that conditional on $(\mathbf{X}_1, \ldots, \mathbf{X}_n) = (\mathbf{M}_1, \ldots, \mathbf{M}_n)$,

$$\begin{aligned} \big\|\mathbb{Q}^*_{1,n,\boldsymbol{\rho}} - \mathbb{Q}_{2,n,\boldsymbol{\rho}}\big\|_{\mathrm{TV}} &= \big\|K_1(\mathbb{Q}_{1,n,\boldsymbol{\rho}}) - K_1\big[K_0(\mathbb{Q}_{2,n,\boldsymbol{\rho}})\big]\big\|_{\mathrm{TV}} \\ &= \big\|K_1\big[\mathbb{Q}_{1,n,\boldsymbol{\rho}} - K_0(\mathbb{Q}_{2,n,\boldsymbol{\rho}})\big]\big\|_{\mathrm{TV}} \\ &\le \big\|\mathbb{Q}_{1,n,\boldsymbol{\rho}} - K_0(\mathbb{Q}_{2,n,\boldsymbol{\rho}})\big\|_{\mathrm{TV}} \\ &= \big\|\mathbb{Q}_{1,n,\boldsymbol{\rho}} - \mathbb{Q}^*_{2,n,\boldsymbol{\rho}}\big\|_{\mathrm{TV}}, \end{aligned} \tag{57}$$

which is bounded by (51). Using the same arguments for showing (52) and (53) we derive from (51) and (57) the following result:

$$\begin{aligned} &\big\|\mathbb{P}^*_{1,n,\boldsymbol{\rho}} - \mathbb{P}_{2,n,\boldsymbol{\rho}}\big\|_{\mathrm{TV}} \\ &\quad\le n \max_{1\le j\le p}\left|1 - \frac{\Xi(j)}{\Pi(j)}\right| + \frac{C\kappa}{\sqrt{m}}\left(n\sum_{j=1}^{p}\Xi(j)1\big(|\mathcal{I}_j(\boldsymbol{\rho})|\ge 2\big)\right)^{1/2} \\ &\quad\le n\delta_p + C\kappa\left(\frac{n\zeta_p}{m}\right)^{1/2}, \end{aligned} \tag{58}$$

and applying (39) we conclude

$$\begin{aligned} \delta(\mathcal{P}_{1n}, \mathcal{P}_{2n}) &\le \inf_K \sup_{\boldsymbol{\rho}\in\Theta}\big\|K(\mathbb{P}_{1,n,\boldsymbol{\rho}}) - \mathbb{P}_{2,n,\boldsymbol{\rho}}\big\|_{\mathrm{TV}} \\ &\le \sup_{\boldsymbol{\rho}\in\Theta}\big\|\mathbb{P}^*_{1,n,\boldsymbol{\rho}} - \mathbb{P}_{2,n,\boldsymbol{\rho}}\big\|_{\mathrm{TV}} \\ &\le n\delta_p + C\kappa\left(\frac{n\zeta_p}{m}\right)^{1/2}. \end{aligned} \tag{59}$$

Collecting together the deficiency bounds in (54) and (59) we establish (22) to bound the deficiency distance $\Delta(\mathcal{P}_{1n}, \mathcal{P}_{2n})$ for the random design case. For the special case of $\Pi(j) = \Xi(j) = 1/p$, $\gamma_p = 0$ and

$$\begin{aligned} \zeta_p &= \max\left\{\sum_{j=1}^{p}\Pi(j)1\big(|\mathcal{I}_j(\boldsymbol{\rho})|\ge 2\big), \sum_{j=1}^{p}\Xi(j)1\big(|\mathcal{I}_j(\boldsymbol{\rho})|\ge 2\big)\right\} \\ &= \frac{1}{p}\sum_{j=1}^{p}1\big(|\mathcal{I}_j(\boldsymbol{\rho})|\ge 2\big). \end{aligned}$$

The result (24) follows. □

For the fixed design case, the arguments for proving (26) are the same except for now we simply combine (51), (52) and (57) but no need for (53) and (58).

PROOF OF THEOREM 2. The proof of Theorem 1 has essentially established Theorem 2. All we need is to modify the arguments as follows. As in the derivation of (53) we apply Lemma 4 directly to $\mathbb{Q}_{1,n,\boldsymbol{\rho}}$ and $\mathbb{Q}^*_{2,n,\boldsymbol{\rho}}$ and use (51) to get

$$\begin{aligned}
&\big\|\mathbb{Q}_{1,n,\boldsymbol{\rho}} - \mathbb{Q}^*_{2,n,\boldsymbol{\rho}}\big\|_{\mathrm{TV}} \\
&\quad\le \max_{1\le j\le p}\left|1 - \frac{\Pi^n(j)}{\Xi^n(j)}\right| \\
&\qquad + E_\Pi\big(E_\Pi\big[\big\|\mathbb{Q}_{1,n,\boldsymbol{\rho}} - \mathbb{Q}^*_{2,n,\boldsymbol{\rho}}\big\|_{\mathrm{TV}}|\mathbf{X}_1 = \mathbf{M}_1, \ldots, \mathbf{X}_n = \mathbf{M}_n\big]\big) \\
&\quad\le n\gamma_p + C\kappa\left(\frac{n\zeta_p}{m}\right)^{1/2},
\end{aligned}$$

and then we obtain, instead of (54), the following result:

$$\begin{aligned}
\delta(\mathcal{Q}_{2n}, \mathcal{Q}_{1n}) &\le \inf_K \sup_{\boldsymbol{\rho}\in\Theta}\big\|\mathbb{Q}_{1,n,\boldsymbol{\rho}} - K(\mathbb{Q}_{2,n,\boldsymbol{\rho}})\big\|_{\mathrm{TV}} \\
&\le \sup_{\boldsymbol{\rho}\in\Theta}\big\|\mathbb{Q}_{1,n,\boldsymbol{\rho}} - \mathbb{Q}^*_{2,n,\boldsymbol{\rho}}\big\|_{\mathrm{TV}} \\
&\le n\gamma_p + C\kappa\left(\frac{n\zeta_p}{m}\right)^{1/2}.
\end{aligned} \tag{60}$$

As in the derivation of (58), we apply Lemma 4 to $\mathbb{Q}^*_{1,n,\boldsymbol{\rho}}$ and $\mathbb{Q}_{2,n,\boldsymbol{\rho}}$ and use (51) and (57) to get

$$\begin{aligned}
\big\|\mathbb{Q}^*_{1,n,\boldsymbol{\rho}} - \mathbb{Q}_{2,n,\boldsymbol{\rho}}\big\|_{\mathrm{TV}} &\le n\max_{1\le j\le p}\left|1 - \frac{\Xi(j)}{\Pi(j)}\right| \\
&\quad + \frac{C\kappa}{\sqrt{m}}\left(n\sum_{j=1}^p \Xi(j)1\big(\big|\mathcal{I}_j(\boldsymbol{\rho})\big| \ge 2\big)\right)^{1/2} \\
&\le n\delta_p + C\kappa\left(\frac{n\zeta_p}{m}\right)^{1/2},
\end{aligned}$$

and then we obtain, instead of (59), the following result:

$$\begin{aligned}
\delta(\mathcal{Q}_{1n}, \mathcal{Q}_{2n}) &\le \inf_K \sup_{\boldsymbol{\rho}\in\Theta}\big\|K(\mathbb{Q}_{1,n,\boldsymbol{\rho}}) - \mathbb{Q}_{2,n,\boldsymbol{\rho}}\big\|_{\mathrm{TV}} \\
&\le \sup_{\boldsymbol{\rho}\in\Theta}\big\|\mathbb{Q}^*_{1,n,\boldsymbol{\rho}} - \mathbb{Q}_{2,n,\boldsymbol{\rho}}\big\|_{\mathrm{TV}} \\
&\le n\delta_p + C\kappa\left(\frac{n\zeta_p}{m}\right)^{1/2}.
\end{aligned} \tag{61}$$

Putting together the deficiency bounds in (60) and (61) we establish (33) to bound the deficiency distance $\Delta(\mathcal{Q}_{1n}, \mathcal{Q}_{2n})$ for the random design case. □

7.3. *Proofs of corollaries.* To prove corollaries, from Theorems 1 and 2 we need to show the given bounds on ζ_p and then substitute them into (24) and (34). Below we will derive ζ_p for each case.

PROOF OF COROLLARY 1. We first analyze the eigen-structures of basis matrices given by (3). For diagonal basis matrix $\mathbf{B}_j$ with 1 on (ℓ, ℓ) entry and 0 elsewhere, its eigenvalues are 1 and 0. Corresponding to eigenvalue 1, the eigenvector is $\mathbf{e}_\ell$, and corresponding to eigenvalue 0, the eigen-space is the orthogonal complement of $\operatorname{span}\{\mathbf{e}_\ell\}$. Denote by $\mathbf{Q}_{j0}$ and $\mathbf{Q}_{j1}$ the projections on the eigen-spaces corresponding to eigenvalues 0 and 1, respectively.

For real symmetric nondiagonal $\mathbf{B}_j$ with $1/\sqrt{2}$ on (ℓ_1, ℓ_2) and (ℓ_2, ℓ_1) entries and 0 elsewhere, the eigenvalues are 1, -1 and 0. Corresponding to eigenvalues ± 1, the eigenvectors are $(\mathbf{e}_{\ell_1} \pm \mathbf{e}_{\ell_2})/\sqrt{2}$, respectively, and corresponding to eigenvalue 0, the eigen-space is the orthogonal complement of $\operatorname{span}\{\mathbf{e}_{\ell_1} \pm \mathbf{e}_{\ell_2}\}$. Denote by $\mathbf{Q}_{j0}$, $\mathbf{Q}_{j1}$ and $\mathbf{Q}_{j,-1}$ the projections on the eigen-spaces corresponding to eigenvalues 0, 1 and -1, respectively.

For imaginary Hermitian $\mathbf{B}_j$ with $-\sqrt{-1}/\sqrt{2}$ on (ℓ_1, ℓ_2) entry, $\sqrt{-1}/\sqrt{2}$ on (ℓ_2, ℓ_1) entry and 0 elsewhere, the eigenvalues are 1, -1 and 0. Corresponding to eigenvalues ± 1, the eigenvector are $(\mathbf{e}_{\ell_1} \pm \sqrt{-1}\mathbf{e}_{\ell_2})/\sqrt{2}$, respectively, and corresponding to eigenvalue 0, the eigen-space is the orthogonal complement of $\operatorname{span}\{\mathbf{e}_{\ell_1} \pm \sqrt{-1}\mathbf{e}_{\ell_2}\}$. Denote by $\mathbf{Q}_{j0}$, $\mathbf{Q}_{j1}$ and $\mathbf{Q}_{j,-1}$ the projections on the eigen-spaces corresponding to eigenvalues 0, 1 and -1, respectively.

For diagonal $\mathbf{B}_j$ with 1 on (ℓ, ℓ) entry, it is a binomial case,

$$\operatorname{tr}(\boldsymbol{\rho}\mathbf{Q}_{j0}) = 1 - \operatorname{tr}(\boldsymbol{\rho}\mathbf{Q}_{j1}), \qquad \operatorname{tr}(\boldsymbol{\rho}\mathbf{Q}_{j1}) = \mathbf{e}_\ell' \boldsymbol{\rho} \mathbf{e}_\ell = \rho_{\ell\ell}$$

and

$$\bigl|\mathcal{I}_j(\boldsymbol{\rho})\bigr| = 2 \cdot 1\bigl(0 < \operatorname{tr}(\boldsymbol{\rho}\mathbf{Q}_{j1}) < 1\bigr) + 1\bigl(\operatorname{tr}(\boldsymbol{\rho}\mathbf{Q}_{j1}) = 1\bigr) + 1\bigl(\operatorname{tr}(\boldsymbol{\rho}\mathbf{Q}_{j1}) = 0\bigr).$$

In order to have $|\mathcal{I}_j(\boldsymbol{\rho})| \geq 2$, we need $\operatorname{tr}(\boldsymbol{\rho}\mathbf{Q}_{j1}) = \rho_{\ell\ell} \in (0, 1)$. Since $\boldsymbol{\rho}$ has at most s_d nonzero diagonal entries, among all the d diagonal matrices $\mathbf{B}_j$ there are at most s_d of diagonal matrices $\mathbf{B}_j$ for which it is possible to have $\operatorname{tr}(\boldsymbol{\rho}\mathbf{Q}_{j1}) \in (0, 1)$ and thus $|\mathcal{I}_j(\boldsymbol{\rho})| \geq 2$.

For nondiagonal $\mathbf{B}_j$, it is a trinomial case,

$$\operatorname{tr}(\boldsymbol{\rho}\mathbf{Q}_{j0}) = 1 - \operatorname{tr}(\boldsymbol{\rho}\mathbf{Q}_{j1}) - \operatorname{tr}(\boldsymbol{\rho}\mathbf{Q}_{j,-1}),$$

and $\operatorname{tr}(\boldsymbol{\rho}\mathbf{Q}_{j\pm1})$ depend on whether $\mathbf{B}_j$ is real or complex.

For real symmetric nondiagonal $\mathbf{B}_j$ with $1/\sqrt{2}$ on (ℓ_1, ℓ_2) and (ℓ_2, ℓ_1) entries,

$$\begin{aligned}\operatorname{tr}(\boldsymbol{\rho}\mathbf{Q}_{j\pm1}) &= (\mathbf{e}_{\ell_1} \pm \mathbf{e}_{\ell_2})' \boldsymbol{\rho} (\mathbf{e}_{\ell_1} \pm \mathbf{e}_{\ell_2})/2 \\ &= (\rho_{\ell_1\ell_1} + \rho_{\ell_2\ell_2} \pm \rho_{\ell_1\ell_2} \pm \rho_{\ell_2\ell_1})/2 \\ &= \frac{1}{2}(1, \pm 1) \begin{pmatrix} \rho_{\ell_1\ell_1} & \rho_{\ell_1\ell_2} \\ \rho_{\ell_2\ell_1} & \rho_{\ell_2\ell_2} \end{pmatrix} \begin{pmatrix} 1 \\ \pm 1 \end{pmatrix};\end{aligned}$$

and for imaginary Hermitian nondiagonal $\mathbf{B}_j$ with $-\sqrt{-1}/\sqrt{2}$ on (ℓ_1, ℓ_2) entry and $\sqrt{-1}/\sqrt{2}$ on (ℓ_2, ℓ_1) entry,

$$\begin{aligned}\operatorname{tr}(\boldsymbol{\rho}\mathbf{Q}_{j\pm1}) &= (\mathbf{e}_{\ell_1} \pm \sqrt{-1}\mathbf{e}_{\ell_2})^{\dagger}\boldsymbol{\rho}(\mathbf{e}_{\ell_1} \pm \sqrt{-1}\mathbf{e}_{\ell_2})/2\\ &= (\rho_{\ell_1\ell_1} + \rho_{\ell_2\ell_2} \pm \sqrt{-1}\rho_{\ell_1\ell_2} \mp \sqrt{-1}\rho_{\ell_2\ell_1})/2\\ &= \frac{1}{2}(1, \mp\sqrt{-1})\begin{pmatrix}\rho_{\ell_1\ell_1} & \rho_{\ell_1\ell_2}\\ \rho_{\ell_2\ell_1} & \rho_{\ell_2\ell_2}\end{pmatrix}\begin{pmatrix}1\\ \pm\sqrt{-1}\end{pmatrix}.\end{aligned}$$

As $\boldsymbol{\rho}$ is semi-positive with trace 1, matrix

$$\begin{pmatrix}\rho_{\ell_1\ell_1} & \rho_{\ell_1\ell_2}\\ \rho_{\ell_2\ell_1} & \rho_{\ell_2\ell_2}\end{pmatrix}$$

must be semi-positive with trace no more than 1. Of $\rho_{\ell_1\ell_1}$ and $\rho_{\ell_2\ell_2}$, if one of them is zero, the semi-positiveness implies $\rho_{\ell_1\ell_2} = \rho_{\ell_2\ell_1} = 0$. Thus, the 2 by 2 matrix has four scenarios:

$$\begin{pmatrix}\rho_{\ell_1\ell_1} & \rho_{\ell_1\ell_2}\\ \rho_{\ell_2\ell_1} & \rho_{\ell_2\ell_2}\end{pmatrix} \quad \text{or} \quad \begin{pmatrix}\rho_{\ell_1\ell_1} & 0\\ 0 & 0\end{pmatrix} \quad \text{or} \quad \begin{pmatrix}0 & 0\\ 0 & \rho_{\ell_2\ell_2}\end{pmatrix} \quad \text{or} \quad \begin{pmatrix}0 & 0\\ 0 & 0\end{pmatrix}.$$

For the last three scenarios under both real symmetric and imaginary Hermitian cases, we obtain

$$\operatorname{tr}(\boldsymbol{\rho}\mathbf{Q}_{j1}) = \operatorname{tr}(\boldsymbol{\rho}\mathbf{Q}_{j,-1}) = \rho_{\ell_1\ell_1}/2 \quad \text{or} \quad \rho_{\ell_2\ell_2}/2 \quad \text{or} \quad 0.$$

For both real symmetric and imaginary Hermitian cases, in order to have $|\mathcal{I}_j(\boldsymbol{\rho})| \geq 2$ possible, at lease one of $\rho_{\ell_1\ell_1}$ and $\rho_{\ell_2\ell_2}$ needs to be nonzero. Since $\boldsymbol{\rho}$ has at most s_d nonzero diagonal entries, among $(d^2 - d)/2$ real symmetric nondiagonal matrices $\mathbf{B}_j$ [or $(d^2 - d)/2$ imaginary Hermitian nondiagonal matrices $\mathbf{B}_j$], there are at most $ds_d - s_d(s_d + 1)/2$ of real symmetric nondiagonal $\mathbf{B}_j$ (or imaginary Hermitian nondiagonal matrices $\mathbf{B}_j$) for which it is possible to have $\operatorname{tr}(\boldsymbol{\rho}\mathbf{Q}_{j1}) \in (0, 1)$ or $\operatorname{tr}(\boldsymbol{\rho}\mathbf{Q}_{j,-1}) \in (0, 1)$ and thus $|\mathcal{I}_j(\boldsymbol{\rho})| \geq 2$.

Finally, for $\boldsymbol{\rho} \in \Theta_s$, putting together the results on the number of $\mathbf{B}_j$ for which it is possible to have $|\mathcal{I}_j(\boldsymbol{\rho})| \geq 2$ in the diagonal, real symmetric and imaginary Hermitian cases, we conclude

$$\sum_{j=1}^{p} 1\big(\big|\mathcal{I}_j(\boldsymbol{\rho})\big| \geq 2\big) \leq ds_d - s_d(s_d + 1) + s_d \leq ds_d$$

and

$$\zeta_p = \max_{\boldsymbol{\rho}\in\Theta_s}\left\{\frac{1}{p}\sum_{j=1}^{p} 1\big(\big|\mathcal{I}_j(\boldsymbol{\rho})\big| \geq 2\big)\right\} \leq \frac{s_d}{d}.$$

□

PROOF OF COROLLARY 2. The Pauli basis (4) has $p = d^2$ matrices with $d = 2^b$. We identify index $j = 1, \ldots, p$ with $(\ell_1, \ell_2, \ldots, \ell_b) \in \{0, 1, 2, 3\}^b$, $j = 1$

corresponds to $\ell_1 = \cdots = \ell_b = 0$, and $\mathbf{B}_1 = \mathbf{I}_d$. In two dimensions, Pauli matrices satisfy $\operatorname{tr}(\sigma_0) = 2$, and $\operatorname{tr}(\sigma_1) = \operatorname{tr}(\sigma_2) = \operatorname{tr}(\sigma_3) = 0$. Consider $\mathbf{B}_j = \boldsymbol{\sigma}_{\ell_1} \otimes \boldsymbol{\sigma}_{\ell_2} \otimes \cdots \otimes \boldsymbol{\sigma}_{\ell_b}$. $\operatorname{tr}(\mathbf{B}_j) = \operatorname{tr}(\boldsymbol{\sigma}_{\ell_1}) \operatorname{tr}(\boldsymbol{\sigma}_{\ell_2}) \cdots \operatorname{tr}(\boldsymbol{\sigma}_{\ell_b})$; $\operatorname{tr}(\mathbf{B}_1) = d$; for $j \neq 1$ [or $(\ell_1, \ldots, \ell_b) \neq (0, \ldots, 0)$], $\operatorname{tr}(\mathbf{B}_j) = 0$ and $\mathbf{B}_j$ has eigenvalues ± 1. Denote by $\mathbf{Q}_{j\pm}$ the projections onto the eigen-spaces corresponding to eigenvalues ± 1, respectively. Then for $j \neq 1$,

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

and solving the equations we get

$$\operatorname{tr}(\mathbf{Q}_{j\pm}) = d/2, \qquad \operatorname{tr}(\mathbf{B}_j \mathbf{Q}_{j\pm}) = \pm \operatorname{tr}(\mathbf{Q}_{j\pm}) = \pm d/2, \qquad j \neq 1. \tag{62}$$

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

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

and further if $j, j' \neq 1$,

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

which imply

$$\operatorname{tr}(\mathbf{B}_{j'} \mathbf{Q}_{j\pm}) = 0, \qquad j \neq j', j, j' \neq 1. \tag{63}$$

For any density matrix $\boldsymbol{\rho}$ with representation (36) under the Pauli basis (4), we have $1 = \operatorname{tr}(\boldsymbol{\rho}) = \alpha_1 \operatorname{tr}(\mathbf{B}_1) = d\alpha_1$ and hence $\alpha_1 = 1/d$. Consider special density matrices $\boldsymbol{\rho} \in \Theta_s$ with expression

$$\boldsymbol{\rho} = \frac{1}{d} \mathbf{I}_d + \frac{\beta}{d} \mathbf{B}_{j^*}, \tag{64}$$

where β is a real number with $|\beta| < 1$, and index $j^* \neq 1$.

To check if $|\mathcal{I}_j(\boldsymbol{\rho})| \geq 2$, we need to evaluate $\operatorname{tr}(\boldsymbol{\rho} \mathbf{Q}_{j\pm})$ for $\boldsymbol{\rho}$ given by (64), $j = 1, \ldots, p$.

For $j = 1$, $\mathbf{B}_1 = \mathbf{Q}_{1+} = \mathbf{I}_d$, and since $\operatorname{tr}(\mathbf{B}_{j^*}) = 0$, we have

$$\operatorname{tr}(\boldsymbol{\rho} \mathbf{Q}_{1+}) = \frac{1}{d} \operatorname{tr}(\mathbf{I}_d) + \frac{\beta}{d} \operatorname{tr}(\mathbf{B}_{j^*}) = 1. \tag{65}$$

For $j = j^*$, from (62) we have $\operatorname{tr}(\mathbf{Q}_{j^*\pm}) = d/2$ and $\operatorname{tr}(\mathbf{B}_{j^*} \mathbf{Q}_{j^*\pm}) = \pm d/2$, and thus

$$\operatorname{tr}(\boldsymbol{\rho} \mathbf{Q}_{j^*\pm}) = \frac{1}{d} \operatorname{tr}(\mathbf{Q}_{j^*\pm}) + \frac{\beta}{d} \operatorname{tr}(\mathbf{B}_{j^*} \mathbf{Q}_{j^*\pm}) = \frac{1 \pm \beta}{2} \in (0, 1). \tag{66}$$

For $j \neq j^*$ or 1 [i.e., $(\ell_1, \ldots, \ell_b) \neq (\ell_1^*, \ldots, \ell_b^*)$ or $(0, \ldots, 0)$], from (63) we have $\operatorname{tr}(\mathbf{B}_{j^*} \mathbf{Q}_{j\pm}) = 0$, and thus

$$\operatorname{tr}(\boldsymbol{\rho} \mathbf{Q}_{j\pm}) = \frac{1}{d} \operatorname{tr}(\mathbf{Q}_{j\pm}) + \frac{\beta}{d} \operatorname{tr}(\mathbf{B}_{j^*} \mathbf{Q}_{j\pm}) = \frac{1}{d} \operatorname{tr}(\mathbf{Q}_{j\pm}) = \frac{1}{2}. \tag{67}$$

Equations (65)–(67) immediately show that for $\boldsymbol{\rho}$ given by (64) and $j \neq 1$, $\operatorname{tr}(\boldsymbol{\rho}\mathbf{Q}_{j\pm}) \in [(1-|\beta|)/2, (1+|\beta|)/2]$, $|\mathcal{I}_j(\boldsymbol{\rho})| = 2$, and

$$\sum_{j=1}^{p} 1\big(|\mathcal{I}_j(\boldsymbol{\rho})| \geq 2\big) = p-1,$$

which implies

$$\max_{\boldsymbol{\rho}\in\Theta_s^p} \left\{ \frac{1}{p} \sum_{j=1}^{p} 1\big(|\mathcal{I}_j(\boldsymbol{\rho})| \geq 2\big) \right\} \geq 1 - \frac{1}{p}. \qquad \square$$

PROOF OF COROLLARY 3. We use the notation and facts about the Pauli basis (4) in the proof of Corollary 2: $p = d^2$, $d = 2^b$, and we identify index $j = 1, \ldots, p$ with $(\ell_1, \ell_2, \ldots, \ell_b) \in \{0,1,2,3\}^b$. Consider $\mathbf{B}_j = \boldsymbol{\sigma}_{\ell_1} \otimes \boldsymbol{\sigma}_{\ell_2} \otimes \cdots \otimes \boldsymbol{\sigma}_{\ell_b}$. For $j = 1$ [or $\ell_1 = \cdots = \ell_b = 0$], $\mathbf{B}_1 = \mathbf{I}_d$, and for $j \neq 1$ [or $(\ell_1, \ldots, \ell_b) \neq (0, \ldots, 0)$], $\mathbf{B}_j$ has eigenvalues ± 1, $\mathbf{Q}_{j\pm}$ are the projections onto the eigen-spaces corresponding to eigenvalues ± 1, respectively, $\mathbf{B}_j = \mathbf{Q}_{j+} - \mathbf{Q}_{j-}$, and $\mathbf{I}_d = \mathbf{Q}_{j+} + \mathbf{Q}_{j-}$.

Let

$$\mathbf{e} = \sqrt{2/7}\big[(\sqrt{3}/2, 1/2)' + (\sqrt{3}/2, \sqrt{-1}/2)'\big] = \big(\sqrt{6/7}, \sqrt{1/14} + \sqrt{-1/14}\big)'.$$

Then for $\ell = 0, 1, 2, 3$, $\varpi_\ell = \mathbf{e}^\dagger \boldsymbol{\sigma}_\ell \mathbf{e}$ is equal to 1, $2\sqrt{3}/7$, $2\sqrt{3}/7$ and $5/7$, respectively. Let $U = \mathbf{e}^{\otimes b}$ and $\boldsymbol{\rho} = UU^\dagger$. Then $\boldsymbol{\rho}$ is a rank one density matrix, and

$$\begin{aligned}
&\operatorname{tr}(\boldsymbol{\rho}\mathbf{Q}_{j+}) + \operatorname{tr}(\boldsymbol{\rho}\mathbf{Q}_{j-}) = \operatorname{tr}(\boldsymbol{\rho}) = 1,\\
&\operatorname{tr}(\boldsymbol{\rho}\mathbf{Q}_{j+}) - \operatorname{tr}(\boldsymbol{\rho}\mathbf{Q}_{j-}) = \operatorname{tr}(\boldsymbol{\rho}\mathbf{B}_j) = U^\dagger \mathbf{B}_j U = \big(\mathbf{e}^\dagger \boldsymbol{\sigma}_{\ell_1} \mathbf{e}\big) \times \cdots \times \big(\mathbf{e}^\dagger \boldsymbol{\sigma}_{\ell_b} \mathbf{e}\big)\\
&\qquad\qquad = \varpi_{\ell_1} \cdots \varpi_{\ell_b}.
\end{aligned}$$

Solving the two equations we obtain $\operatorname{tr}(\boldsymbol{\rho}\mathbf{Q}_{j\pm}) = (1 \pm \varpi_{\ell_1} \cdots \varpi_{\ell_b})/2$.

For $j \neq 1$ [or $(\ell_1, \ldots, \ell_b) \neq (0, \ldots, 0)$], $(\varpi_{\ell_1}, \ldots, \varpi_{\ell_b}) \neq (1, \ldots, 1)$, and $0 \leq \varpi_{\ell_1} \cdots \varpi_{\ell_b} \leq 5/7$, and thus $\operatorname{tr}(\boldsymbol{\rho}\mathbf{Q}_{j+}) \geq 1/2$ and $\operatorname{tr}(\boldsymbol{\rho}\mathbf{Q}_{j-}) \geq 1/7$, which immediately shows that for the given rank one density matrix $\boldsymbol{\rho}$ and $j \neq 1$, $|\mathcal{I}_j(\boldsymbol{\rho})| = 2$, and

$$\sum_{j=1}^{p} 1\big(|\mathcal{I}_j(\boldsymbol{\rho})| \geq 2\big) = p-1,$$

which implies

$$\max_{\boldsymbol{\rho}\in\Theta_r} \left\{ \frac{1}{p} \sum_{j=1}^{p} 1\big(|\mathcal{I}_j(\boldsymbol{\rho})| \geq 2\big) \right\} \geq 1 - \frac{1}{p}. \qquad \square$$

PROOF OF COROLLARY 4. Since under $\mathbf{g}_1, \ldots, \mathbf{g}_d$, basis matrices $\mathbf{B}_j$ defined in the corollary have the same behavior as matrix basis (3) under $\mathbf{e}_1, \ldots, \mathbf{e}_d$, from

the proof of Corollary 1 on the eigen-structures of matrix basis (3) we see that under $\mathbf{g}_1, \ldots, \mathbf{g}_d$, $\mathbf{B}_j$ has possible eigenvalues 0 and 1 for diagonal $\mathbf{B}_j$ and eigenvalues 0, 1 and -1 for nondiagonal $\mathbf{B}_j$. For the diagonal case, corresponding to eigenvalue 1, the eigenvector is $\mathbf{g}_\ell$; for the real symmetric nondiagonal case, corresponding to eigenvalues ± 1, the eigenvectors are $(\mathbf{g}_{\ell_1} \pm \mathbf{g}_{\ell_2})/\sqrt{2}$, respectively; and for the complex Hermitian nondiagonal case, corresponding to eigenvalue ± 1, the eigenvectors are $(\mathbf{g}_{\ell_1} \pm \sqrt{-1}\mathbf{g}_{\ell_2})/\sqrt{2}$, respectively. Denote by $\mathbf{Q}_{j0}$, $\mathbf{Q}_{j1}$ and $\mathbf{Q}_{j,-1}$ the projections on the eigen-spaces corresponding to eigenvalues 0, 1 and -1, respectively.

For diagonal $\mathbf{B}_j$ with j corresponding to (ℓ, ℓ), it is a binomial case,

$$\operatorname{tr}(\boldsymbol{\rho}\mathbf{Q}_{j0}) = 1 - \operatorname{tr}(\boldsymbol{\rho}\mathbf{Q}_{j1}), \qquad \operatorname{tr}(\boldsymbol{\rho}\mathbf{Q}_{j1}) = \mathbf{g}_\ell' \boldsymbol{\rho} \mathbf{g}_\ell = \sum_{a=1}^{r} \xi_a \big| U_a^\dagger \mathbf{g}_\ell \big|^2$$

and

$$\begin{aligned} \big|\mathcal{I}_j(\boldsymbol{\rho})\big| = {} & 2 \cdot 1\big(0 < \operatorname{tr}(\boldsymbol{\rho}\mathbf{Q}_{j1}) < 1\big) + 1\big(\operatorname{tr}(\boldsymbol{\rho}\mathbf{Q}_{j1}) = 1\big) \\ & + 1\big(\operatorname{tr}(\boldsymbol{\rho}\mathbf{Q}_{j1}) = 0\big). \end{aligned}$$

In order to have $|\mathcal{I}_j(\boldsymbol{\rho})| \geq 2$ possible, we need $\operatorname{tr}(\boldsymbol{\rho}\mathbf{Q}_{j1}) \in (0, 1)$. Since $\boldsymbol{\rho}$ is generated by at most r vectors U_a, and for each U_a there are at most 2γ of $\mathbf{g}_\ell$ with $U_a^\dagger \mathbf{g}_\ell \neq 0$, among all the d diagonal matrices $\mathbf{B}_j$ there are at most $2r\gamma$ of diagonal matrices $\mathbf{B}_j$ for which it is possible to have $\operatorname{tr}(\boldsymbol{\rho}\mathbf{Q}_{j1}) \in (0, 1)$ and thus $|\mathcal{I}_j(\boldsymbol{\rho})| \geq 2$.

For nondiagonal $\mathbf{B}_j$, it is a trinomial case,

$$\operatorname{tr}(\boldsymbol{\rho}\mathbf{Q}_{j0}) = 1 - \operatorname{tr}(\boldsymbol{\rho}\mathbf{Q}_{j1}) - \operatorname{tr}(\boldsymbol{\rho}\mathbf{Q}_{j,-1}),$$

and $\operatorname{tr}(\boldsymbol{\rho}\mathbf{Q}_{j\pm1})$ depend on whether $\mathbf{B}_j$ is real or complex.

For real symmetric nondiagonal $\mathbf{B}_j$ with j corresponding to (ℓ_1, ℓ_2),

$$\operatorname{tr}(\boldsymbol{\rho}\mathbf{Q}_{j\pm1}) = (\mathbf{g}_{\ell_1} \pm \mathbf{g}_{\ell_2})' \boldsymbol{\rho} (\mathbf{g}_{\ell_1} \pm \mathbf{g}_{\ell_2})/2 = \sum_{a=1}^{r} \xi_a \big| U_a^\dagger (\mathbf{g}_{\ell_1} \pm \mathbf{g}_{\ell_2}) \big|^2 /2;$$

and for imaginary Hermitian nondiagonal $\mathbf{B}_j$ with j corresponding to (ℓ_1, ℓ_2),

$$\begin{aligned} \operatorname{tr}(\boldsymbol{\rho}\mathbf{Q}_{j\pm1}) &= (\mathbf{g}_{\ell_1} \pm \sqrt{-1}\mathbf{g}_{\ell_2})^\dagger \boldsymbol{\rho} (\mathbf{g}_{\ell_1} \pm \sqrt{-1}\mathbf{g}_{\ell_2})/2 \\ &= \sum_{a=1}^{r} \xi_a \big| U_a^\dagger (\mathbf{g}_{\ell_1} \pm \sqrt{-1}\mathbf{g}_{\ell_2}) \big|^2 /2. \end{aligned}$$

In order to have $|\mathcal{I}_j(\boldsymbol{\rho})| \geq 2$ possible, we need $\operatorname{tr}(\boldsymbol{\rho}\mathbf{Q}_{j1}) \in (0, 1)$ or $\operatorname{tr}(\boldsymbol{\rho}\mathbf{Q}_{j-1}) \in (0, 1)$. Since $\boldsymbol{\rho}$ is generated by at most r vectors U_a, and for each U_a there are at most 2γ number of $\mathbf{g}_\ell$ with $U_a^\dagger \mathbf{g}_\ell \neq 0$, among $(d^2 - d)/2$ real symmetric nondiagonal matrices $\mathbf{B}_j$ [or $(d^2 - d)/2$ imaginary Hermitian nondiagonal matrices $\mathbf{B}_j$], there are at most $4r\gamma^2$ of real symmetric nondiagonal $\mathbf{B}_j$ (or imaginary Hermi-

tian nondiagonal matrices $\mathbf{B}_j$) for which it is possible to have $\operatorname{tr}(\boldsymbol{\rho}\mathbf{Q}_{j1}) \in (0, 1)$ or $\operatorname{tr}(\boldsymbol{\rho}\mathbf{Q}_{j,-1}) \in (0, 1)$ and thus $|\mathcal{I}_j(\boldsymbol{\rho})| \geq 2$.

Finally, for $\boldsymbol{\rho} \in \Theta_{r\gamma}$, combining the results on the number of $\mathbf{B}_j$ for which it is possible to have $|\mathcal{I}_j(\boldsymbol{\rho})| \geq 2$ in the diagonal, real symmetric and imaginary Hermitian cases, we conclude

$$\sum_{j=1}^{p} 1\big(|\mathcal{I}_j(\boldsymbol{\rho})| \geq 2\big) \leq 8r\gamma^2 + 2r\gamma,$$

and

$$\zeta_p = \max_{\boldsymbol{\rho}\in\Theta_{r\gamma}} \left\{ \frac{1}{p} \sum_{j=1}^{p} 1\big(|\mathcal{I}_j(\boldsymbol{\rho})| \geq 2\big) \right\} \leq \frac{2r\gamma(4\gamma+1)}{p}. \qquad \square$$

APPENDIX: PROOFS OF LEMMAS 2–4

PROOF OF LEMMA 2. For $r = 2$, it is the binomial case, and the lemma is a consequence of (43) and Lemma 1.

For $r = 3$, write $(U_1, U_2, U_3) \sim P$ and $(V_1, V_2, V_3) \sim Q$. Add independent uniforms on $(-1/2, 1/2)$ to U_1 and U_2, denote the resulting random variables by U_1^* and U_2^*, respectively, and let $U_3^* = m - U_1^* - U_2^*$. Then $(U_1^*, U_2^*, U_3^*) \sim P^*$. Note that $U_1 + U_2 + U_3 = U_1^* + U_2^* + U_3^* = V_1 + V_2 + V_3 = m$, and U_1 and U_2 are equal to the round-offs, $[U_1^*]$ and $[U_2^*]$, of U_1^* and U_2^*, respectively, here round-off $[x]$ means rounding x off to the nearest integer.

For trinomial random variable $(U_1, U_2, U_3) \sim \mathcal{M}(m, \theta_1, \theta_2, \theta_3)$, we have $U_1 \sim \operatorname{Bin}(m, \beta_1) = P_1$, the conditional distribution of U_2 given U_1: $U_2|U_1 \sim \operatorname{Bin}(m - U_1, \beta_2) = P_2$, and $U_3 = m - U_1 - U_2$, where $\beta_1 = \theta_1$, $\beta_2 = \theta_2/(\theta_2 + \theta_3)$, $\beta_3 = \theta_3/(\theta_2 + \theta_3)$. Since θ_j are between c_0 and c_1, β_2 and β_3 are between $c_0/(c_0 + c_1)$ and $c_1/(c_0 + c_1)$. We have decomposition $P = P_1 P_2$.

Denote by P_1^* the distribution of U_1^* and P_2^* the conditional distribution of U_2^* given U_1^*. Then P_1^* is the convolution of P_1 and an independent uniform distribution on $(-1/2, 1/2)$. Since the added uniforms are independent of U_j, and U_j is the round-off of U_j^*, the conditional distribution of U_2^* given U_1^* is equal to the conditional distribution of U_2^* given $U_1 = [U_1^*]$, which in turn is equal to the convolution of P_2 and an independent uniform distribution on $(-1/2, 1/2)$. We have decomposition $P^* = P_1^* P_2^*$.

For trivariate normal random variable $(V_1, V_2, V_3) \sim Q$, we have $V_1 \sim N(m\beta_1, m\beta_1(1 - \beta_1)) = Q_1$, the conditional distribution of V_2 given V_1: $V_2|V_1 \sim N((m - V_1)\beta_2, m(1 - \beta_1)\beta_2\beta_3) = Q_2$, and $V_3 = m - V_1 - V_2$. We have decomposition $Q = Q_1 Q_2$.

As there is a difference in conditional variance between P_2 and Q_2, we define $V_2' \sim Q_2' = N((m-V_1)\beta_2, (m-V_1)\beta_2\beta_3)$ to match the conditional variance of P_2, and $V_3' = m - V_1 - V_2'$. Simple direct calculations show that given V_1,

$$H^2(Q_2, Q_2') \le \frac{3}{2}\left(1 - \frac{m - V_1}{m(1-\beta_1)}\right)^2. \tag{68}$$

Note that $P^* = P_1^* P_2^*$ and $Q = Q_1 Q_2$ are probability measures on $\{(x_1, x_2, x_3) : x_1 + x_2 + x_3 = m\}$. Define probability measures $Q_1 Q_2'$ and $P_1^* Q_2'$ on $\{(x_1, x_2, x_3) : x_1 + x_2 + x_3 = m\}$, where Q_1 and P_1^* are their respective marginal distributions of the first component, and Q_2' is their conditional distribution of the second component given the first component. We use $Q_1 Q_2'$ and $P_1^* Q_2'$ to bridge between $P^* = P_1^* P_2^*$ and $Q = Q_1 Q_2$. Applying triangle inequality we obtain

$$\begin{aligned} H(P^*, Q) &\le H(P^*, Q_1 Q_2') + H(Q_1 Q_2', Q) \\ &\le H(P_1^* P_2^*, P_1^* Q_2') + H(P_1^* Q_2', Q_1 Q_2') \\ &\quad + H(Q_1 Q_2', Q_1 Q_2). \end{aligned} \tag{69}$$

Using (40), (43), Lemma 1 and (68) we evaluate the Hellinger distances on the right-hand side of (69) as follows:

$$\begin{aligned} H^2(Q_1 Q_2', Q_1 Q_2) &= \int \left| \sqrt{\frac{dQ_1}{dx_1}\frac{dQ_2}{dx_2}} - \sqrt{\frac{dQ_1}{dx_1}\frac{dQ_2'}{dx_2}} \right|^2 dx_1\, dx_2 \\ &= \int dQ_1 \int \left| \sqrt{\frac{dQ_2}{dx_2}} - \sqrt{\frac{dQ_2'}{dx_2}} \right|^2 dx_2 \\ &= E_{Q_1}[H^2(Q_2, Q_2')] \\ &\le E_{Q_1}\left[\frac{3}{2}\left(1 - \frac{m - V_1}{m(1-\beta_1)}\right)^2\right] \\ &= \frac{3\beta_1}{2m(1-\beta_1)} \le \frac{3\theta_1}{2m(\theta_2+\theta_3)} \le \frac{C}{m}, \end{aligned} \tag{70}$$

where (68) is used to bound $H^2(Q_2, Q_2')$ and obtain the first inequality

$$\begin{aligned} H^2(P_1^* Q_2', Q_1 Q_2') &= \int \left| \sqrt{\frac{dP_1^*}{dx_1}} - \sqrt{\frac{dQ_1}{dx_1}} \right|^2 dx_1 \int dQ_2' \\ &= \int \left| \sqrt{\frac{dP_1^*}{dx_1}} - \sqrt{\frac{dQ_1}{dx_1}} \right|^2 dx_1 = H^2(P_1^*, Q_1) \\ &\le \exp(-Cm^{1/3}) + \frac{C}{m\theta_1(1-\theta_1)} \le \frac{C}{m}, \end{aligned} \tag{71}$$

where Lemma 1 and (43) are used to bound $H^2(P_1^*, Q_1)$ and obtain the first inequality

$$
\begin{aligned}
H^2(P_1^*P_2^*, P_1^*Q_2') &= \int dP_1^* \int \left|\sqrt{\frac{dP_2^*}{dx_2}} - \sqrt{\frac{dQ_2'}{dx_2}}\right|^2 dx_2 \\
&= E_{P_1^*}\left[H^2(P_2^*, Q_2')\right] \\
&\le 2 - 2E_{P_1^*}\left\{E_{P_2^*}\left[1_A\sqrt{\frac{dP_2^*}{dQ_2'}}\middle|U_1\right]\right\} \\
&\le 2P^*(A^c) + E_{P_1^*}\left\{1_{A_1}E_{P_2^*}\left[1_{A_2}\log\frac{dP_2^*}{dQ_2'}\middle|U_1\right]\right\},
\end{aligned} \tag{72}
$$

where we use (43) to bound $H^2(P_2^*, Q_2')$ and obtain the last two inequalities, $A = A_1 \cap A_2$, and

$$
\begin{aligned}
A_1 &= \left\{|U_1 - m\beta_1| \le \left[m\beta_1(1-\beta_1)\right]^{2/3}\right\}, \\
A_2 &= \left\{\left|U_2 - (m-U_1)\beta_2\right| \le \left[(m-U_1)\beta_2(1-\beta_2)\right]^{2/3}\right\}.
\end{aligned}
$$

We evaluate $P^*(A^c)$ as follows:

$$
\begin{aligned}
P^*(A^c) &= P\left(A_1^c \cup \left[A_2^c \cap A_1\right]\right) = P\left(A_1^c\right) + P\left(A_2^c \cap A_1\right) \\
&= P_1\left(A_1^c\right) + E_P\left[1_{A_1}P\left(A_2^c|U_1\right)\right] \\
&\le \exp\left(-Cm^{1/3}\right) + E_P\left[1_{A_1}\exp\left(-C\{m-U_1\}^{1/3}\right)\right] \\
&\le \exp\left(-Cm^{1/3}\right) + \exp\left(-C\left\{m - m\beta_1 - \left[m\beta_1(1-\beta_1)\right]^{2/3}\right\}^{1/3}\right) \\
&\le 2\exp\left(-Cm^{1/3}\right),
\end{aligned} \tag{73}
$$

where we utilize Lemma 1 to derive $P_1(A_1^c)$ and $P(A_2^c|U_1)$, and bound $m - U_1$ by using the fact that on A_1, $U_1 \le m\beta_1 + [m\beta_1(1-\beta_1)]^{2/3}$. Again we apply Lemma 1 to bound $E_{P_2^*}[1_{A_2}\log\frac{dP_2^*}{dQ_2'}|U_1]$ and obtain

$$
\begin{aligned}
&E_{P_1^*}\left\{1_{A_1}E_{P_2^*}\left[1_{A_2}\log\frac{dP_2^*}{dQ_2'}\middle|U_1\right]\right\} \\
&\quad\le E_{P_1^*}\left\{1_{A_1}\frac{C}{(m-U_1)\beta_2(1-\beta_2)}\right\} \\
&\quad\le \frac{C}{(m - m\beta_1 - [m\beta_1(1-\beta_1)]^{2/3})\beta_2(1-\beta_2)} \le \frac{C}{m},
\end{aligned} \tag{74}
$$

where to bound $1/(m-U_1)$ we use the fact that on A_1, $U_1 \le m\beta_1 + [m\beta_1(1-\beta_1)]^{2/3}$.

Substituting (73) and (74) into (72) and then combining it with (69)–(71) we prove that the lemma is true for $r = 3$.

Consider the $r+1$ case. Write $(U_1,\dots,U_r,U_{r+1})\sim P$, $U_1+\cdots+U_{r+1}=m$, and decompose $P=P_1P_2\cdots P_{r-1}P_r$, where $U_1\sim P_1=\mathrm{Bin}(m,\beta_1)$, $P_j=\mathrm{Bin}(m-T_{j-1},\beta_j)$ is the conditional distribution of U_j given $U_1,\dots,U_{j-1}$, $T_j=U_1+\cdots+U_j$, $\beta_1=\theta_1$, $\beta_j=\theta_j/(1-\theta_1-\cdots-\theta_{j-1})$. Since θ_j are between c_0 and c_1, all β_j are between $c_0/(c_0+rc_1)$ and $c_1/(c_0+c_1)$ that are bounded away from 0 and 1.

Similarly write $(V_1,\dots,V_r,V_{r+1})\sim Q$, $V_1+\cdots+V_{r+1}=m$, and decompose $Q=Q_1Q_2\cdots Q_{r-1}Q_r$, where $V_1\sim Q_1=N(m\beta_1,m\beta_1(1-\beta_1))$, and $Q_j=N((m-S_{j-1})\beta_j,m(\theta_j+\cdots+\theta_{r+1})\beta_j(1-\beta_j))$ is the conditional distribution of V_j given $V_1,\dots,V_{j-1}$, where $S_j=V_1+\cdots+V_j$.

As there are differences in conditional variance between P_j and Q_j, we handle the differences by introducing $Q'_j\cdots Q'_r$ as follows. Given $V_1,\dots,V_{j-1}$ we define $(V'_j,\dots,V'_r,V'_{r+1})\sim Q'_j\cdots Q'_r$, where the conditional distribution of V'_ℓ given $V_1,\dots,V_{j-1},V'_j,\dots,V'_{\ell-1}$ is $Q'_\ell=N((m-S'_{\ell-1})\beta_\ell,(m-S'_{\ell-1})\beta_\ell(1-\beta_\ell))$ for $\ell=j,\dots,r$, $V'_{r+1}=m-V_1-\cdots-V_{j-1}-V'_j-\cdots-V'_r$, and $S'_\ell=V_1+\cdots+V_{j-1}+V'_j+\cdots+V'_\ell$. Then given $V_1,\dots,V_{j-1}$,

$$H^2(Q_j,Q'_j)\le\frac{3}{2}\biggl(1-\frac{m-S_{j-1}}{m(\theta_j+\cdots+\theta_{r+1})}\biggr)^2. \tag{75}$$

Add independent uniforms on $(-1/2,1/2)$ to $U_1,\dots,U_r$, denote the resulting corresponding random variables by U_j^*, and let $U_{r+1}^*=m-U_1^*-\cdots-U_r^*$. Then $(U_1^*,\dots,U_{r+1}^*)\sim P^*$. Note that $U_1+\cdots+U_{r+1}=U_1^*+\cdots+U_{r+1}^*=V_1+\cdots+V_{r+1}=m$, and U_j is equal to the round-off of U_j^*. Let $P^*=P_1^*P_2^*\cdots P_{r-1}^*P_r^*$, where we denote by P_1^* the distribution of U_1^* and P_j^* the conditional distribution of U_j^* given $U_1^*,\dots,U_{j-1}^*$. Then P_1^* is the convolution of P_1 and an independent uniform distribution on $(-1/2,1/2)$. Since the added uniforms are independent of U_j, and U_j is the round-off of U_j^*, the conditional distribution of U_j^* given $U_1^*,\dots,U_{j-1}^*$ is equal to the conditional distribution of U_j^* given $U_1=[U_1^*],\dots,U_{j-1}=[U_{j-1}^*]$, which in turn is equal to the convolution of P_j and an independent uniform distribution on $(-1/2,1/2)$.

Note that $P^*=P_1^*\cdots P_r^*$ and $Q=Q_1\cdots Q_r$ are probability measures on $\{(x_1,\dots,x_r,x_{r+1}):x_1+\cdots+x_{r+1}=m\}$. We define probability measures $Q_1\cdots Q_jQ'_{j+1}\cdots Q'_r$ and $P_1^*\cdots P_{j-1}^*Q'_j\cdots Q'_r$ on $\{(x_1,\dots,x_r,x_{r+1}):x_1+\cdots+x_{r+1}=m\}$, $j=2,\dots,r$, and use them to bridge between P^* and Q. Applying triangle inequality, we have

$$\begin{aligned}H(P^*,Q)&\le H(P^*,Q_1\cdots Q_{r-1}Q'_r)+H(Q_1\cdots Q_{r-1}Q'_r,Q)\\&\le H(P^*,Q_1\cdots Q_{r-2}Q'_{r-1}Q'_r)\\&\quad+H(Q_1\cdots Q_{r-2}Q'_{r-1}Q'_r,Q_1\cdots Q_{r-1}Q'_r)\\&\quad+H(Q_1\cdots Q_{r-1}Q'_r,Q)\le\cdots\end{aligned} \tag{76}$$

$$\le H(P^*, Q_1 Q_2' \cdots Q_r') + \sum_{j=2}^{r} H(Q_1 \cdots Q_{j-1} Q_j' \cdots Q_r', Q_1 \cdots Q_j Q_{j+1}' \cdots Q_r')$$

and

$$\begin{aligned} &H(P^*, Q_1 Q_2' \cdots Q_r') \\ &\quad \le H(P^*, P_1^* \cdots P_{r-1}^* Q_r') + H(P_1^* \cdots P_{r-1}^* Q_r', Q_1 Q_2' \cdots Q_r') \\ &\quad \le H(P^*, P_1^* \cdots P_{r-1}^* Q_r') + H(P_1^* \cdots P_{r-1}^* Q_r', P_1^* \cdots P_{r-2}^* Q_{r-1}' Q_r') \\ &\qquad + H(P_1^* \cdots P_{r-2}^* Q_{r-1}' Q_r', Q_1 Q_2' \cdots Q_r') \\ &\quad \le \cdots \le \sum_{j=1}^{r} H(P_1^* \cdots P_j^* Q_{j+1}' \cdots Q_r', P_1^* \cdots P_{j-1}^* Q_j' \cdots Q_r'). \end{aligned} \tag{77}$$

Substitute (77) into (76) to get

$$\begin{aligned} H(P^*, Q) \le &\sum_{j=1}^{r} H(P_1^* \cdots P_j^* Q_{j+1}' \cdots Q_r', P_1^* \cdots P_{j-1}^* Q_j' \cdots Q_r') \\ &+ \sum_{j=2}^{r} H(Q_1 \cdots Q_{j-1} Q_j' \cdots Q_r', Q_1 \cdots Q_j Q_{j+1}' \cdots Q_r'). \end{aligned} \tag{78}$$

Using (40), (43), Lemma 1 and (75) we evaluate the Hellinger distances on the right-hand side of (78) as follows:

$$\begin{aligned} &H^2(Q_1 \cdots Q_{j-1} Q_j' \cdots Q_r', Q_1 \cdots Q_j Q_{j+1}' \cdots Q_r') \\ &\quad = \int dQ_1 \cdots dQ_{j-1} \int \left| \sqrt{\frac{dQ_j}{dx_j}} - \sqrt{\frac{dQ_j'}{dx_j}} \right|^2 dx_j \int dQ_{j+1}' \cdots dQ_r' \\ &\quad = \int dQ_1 \cdots dQ_{j-1} \int \left| \sqrt{\frac{dQ_j}{dx_j}} - \sqrt{\frac{dQ_j'}{dx_j}} \right|^2 dx_j \\ &\quad = E_{Q_1 \cdots Q_{j-1}}[H^2(Q_j, Q_j')] \\ &\quad \le E_{Q_1 \cdots Q_{j-1}}\left[\frac{3}{2} \left(1 - \frac{m - S_{j-1}}{m(\theta_j + \cdots + \theta_{r+1})} \right)^2 \right] \\ &\quad = \frac{3(1 - \theta_j - \cdots - \theta_{r+1})}{2m(\theta_j + \cdots + \theta_{r+1})} \le \frac{3\theta_1}{2m(\theta_r + \theta_{r+1})} \\ &\quad \le \frac{C}{m}, \end{aligned} \tag{79}$$

where we use (75) to bound the Hellinger distance $H^2(Q_j, Q'_j)$ and obtain the first inequality

$$
\begin{aligned}
&H^2\big(P_1^*\cdots P_j^* Q'_{j+1}\cdots Q'_r, P_1^*\cdots P_{j-1}^* Q'_j\cdots Q'_r\big)\\
&\quad= \int dP_1^*\cdots dP_{j-1}^* \int \Big|\sqrt{\frac{dP_j^*}{dx_j}} - \sqrt{\frac{dQ'_j}{dx_j}}\Big|^2 dx_j \int dQ'_{j+1}\cdots dQ'_r\\
&\quad= \int dP_1^*\cdots dP_{j-1}^* \int \Big|\sqrt{\frac{dP_j^*}{dx_j}} - \sqrt{\frac{dQ'_j}{dx_j}}\Big|^2 dx_j\\
&\quad= E_{P_1^*\cdots P_{j-1}^*}\big[H^2(P_j^*, Q'_j)\big]\\
&\quad\le E_{P_1^*\cdots P_{j-1}^*}\Big(2P_j^*\big(A_1^c\cup\cdots\cup A_j^c|U_1,\ldots,U_{j-1}\big)\\
&\qquad\qquad + 1_{A_1\cdots A_{j-1}} E_{P_j^*}\Big[1_{A_j}\log\frac{P_j^*}{Q'_j}\Big|U_1,\ldots,U_{j-1}\Big]\Big) \qquad (80)\\
&\quad= 2P^*\big(A_1^c\cup\cdots\cup A_j^c\big)\\
&\qquad + E_{P_1^*\cdots P_{j-1}^*}\Big(1_{A_1\cdots A_{j-1}} E_{P_j^*}\Big[1_{A_j}\log\frac{P_j^*}{Q'_j}\Big|U_1,\ldots,U_{j-1}\Big]\Big)\\
&\quad\le 2P^*\big(A_1^c\cup\cdots\cup A_j^c\big) + E_{P_1^*\cdots P_{j-1}^*}\Big(1_{A_1\cdots A_{j-1}}\frac{C}{(m-T_{j-1})\beta_j(1-\beta_j)}\Big),
\end{aligned}
$$

where we use (43) to bound the Hellinger distance $H^2(P_j^*, Q'_j)$ and obtain the first inequality, we employ Lemma 1 to bound $E_{P_j^*}[1_{A_j}\log\frac{dP_j^*}{dQ'_j}|U_1,\ldots,U_{j-1}]$ and get the last inequality, and for $\ell = 1,\ldots,j$,

$$
A_\ell = \big\{\big|U_\ell - (m-U_1-\cdots-U_{\ell-1})\beta_\ell\big| \le \big[(m-U_1-\cdots-U_{\ell-1})\beta_\ell(1-\beta_\ell)\big]^{2/3}\big\}.
$$

Note that on A_{j-1}, $U_{j-1} \le (m-T_{j-2})\beta_{j-1} + [m\beta_{j-1}(1-\beta_{j-1})]^{2/3}$. Then for $j = 1,\ldots,r$ we have on $A_1\cdots A_{j-1}$,

$$
\begin{aligned}
&m - T_{j-1}\\
&\quad= m - T_{j-2} - U_{j-1}\\
&\quad\ge (m-T_{j-2})(1-\beta_{j-1}) - \big[m\beta_{j-1}(1-\beta_{j-1})\big]^{2/3}\\
&\quad\ge (m-T_{j-3})(1-\beta_{j-2})(1-\beta_{j-1})\\
&\qquad - (1-\beta_{j-1})\big[m\beta_{j-2}(1-\beta_{j-2})\big]^{2/3} - \big[m\beta_{j-1}(1-\beta_{j-1})\big]^{2/3} \ge\cdots\\
&\quad\ge m(1-\beta_1)\cdots(1-\beta_{j-1}) \qquad (81)
\end{aligned}
$$

$$
\begin{aligned}
&- m^{2/3} \sum_{\ell=1}^{j-1} [\beta_\ell(1-\beta_\ell)]^{2/3}(1-\beta_\ell)\cdots(1-\beta_{j-1}) \\
&\geq Cm
\end{aligned}
$$

and thus

$$
E_{P_1^*\cdots P_{j-1}^*}\left(1_{A_1\cdots A_{j-1}} \frac{C}{(m-T_{j-1})\beta_j(1-\beta_j)}\right) \leq \frac{C}{m}. \tag{82}
$$

We evaluate $P^*(A_1^c \cup \cdots \cup A_j^c)$ as follows:

$$
\begin{aligned}
\bigcup_{\ell=1}^{j} A_\ell^c &= \bigcup_{\ell=1}^{j} (A_\ell^c A_{\ell-1}\cdots A_1), \\
P^*\left(\bigcup_{\ell=1}^{j} A_\ell^c\right) &= \sum_{\ell=1}^{j} P^*(A_\ell^c A_{\ell-1}\cdots A_1) \\
&= P^*(A_1^c) + \sum_{\ell=2}^{j} E_{P^*}\left[1_{A_1\cdots A_{\ell-1}} P^*(A_\ell^c | U_1, \ldots, U_{\ell-1})\right] \qquad (83) \\
&\leq \exp[-Cm^{1/3}] + \sum_{\ell=2}^{j} E_{P^*}\left(1_{A_1\cdots A_{\ell-1}} \exp[-C(m-T_{\ell-1})^{1/3}]\right) \\
&\leq \sum_{\ell=1}^{j} \exp[-Cm^{1/3}] \leq j\exp[-Cm^{1/3}],
\end{aligned}
$$

where Lemma 1 is employed to bound $P^*(A_1^c)$ and $P^*(A_\ell^c | U_1, \ldots, U_{\ell-1})$, and we use (81) to bound $m - T_{\ell-1}$.

Plugging (82) and (83) into (80) and combining it together with (78) and (79), we obtain

$$
\begin{aligned}
H(P^*, Q) &\leq \frac{C(r-1)}{\sqrt{m}} + \sum_{j=1}^{r} \left\{2j\exp[-Cm^{1/3}] + \frac{C}{m}\right\}^{1/2} \\
&\leq \frac{Cr}{\sqrt{m}} + r^2 \exp[-Cm^{1/3}],
\end{aligned}
$$

which proves the lemma for the $r+1$ case. □

PROOF OF LEMMA 3. Since P_k, P_k^*, Q_k for different k are independent, an application of the Hellinger distance property for product probability measures [Le Cam and Yang (2000)] leads to

$$
H^2(P^*, Q) \leq \sum_{k=1}^{n} H^2(P_k^*, Q_k).
$$

We note that if $v_k \le 1$, both P_k and Q_k are point mass at m and thus $H(P_k, Q_k) = 0$. Hence,

$$H^2(P^*, Q) \le \sum_{k=1}^{n} H^2(P_k^*, Q_k) 1(v_k \ge 2).$$

Applying Lemma 2, we obtain

$$H^2(P^*, Q) \le \sum_{k=1}^{n} \left[\kappa^4 \exp(-Cm^{1/3}) + \frac{C\kappa^2}{m} \right] 1(v_k \ge 2).$$

For m exceeding certain integer m_0,

$$\frac{C\kappa^2}{m} \ge \kappa^4 \exp(-Cm^{1/3})$$

and hence for $m > m_0$,

$$H^2(P^*, Q) \le \frac{C\kappa^4}{m} \sum_{k=1}^{n} 1(v_k \ge 2).$$

For $m \le m_0$, we may adjust constant C to make the above inequality still holds for $m \le m_0$. □

PROOF OF LEMMA 4.

$$\begin{aligned}
\|F - G\|_{\mathrm{TV}} &= \|F_1(x) \times F_{2|1}(y|x) - G_1(x) \times G_{2|1}(y|x)\|_{\mathrm{TV}} \\
&\le \|F_1(x) \times F_{2|1}(y|x) - F_1(x) \times G_{2|1}(y|x)\|_{\mathrm{TV}} \\
&\quad + \|F_1(x) \times G_{2|1}(y|x) - G_1(x) \times G_{2|1}(y|x)\|_{\mathrm{TV}} \\
&= \|F_1(x)[F_{2|1}(y|x) - G_{2|1}(y|x)]\|_{\mathrm{TV}} \\
&\quad + \|F_1(x)G(x, y)/G_2(x) - G(x, y)\|_{\mathrm{TV}},
\end{aligned}$$

where

$$\begin{aligned}
&\|F_1(x)[F_{2|1}(y|x) - G_{2|1}(y|x)]\|_{\mathrm{TV}} \\
&\quad = E_{F_1}[\|F_{2|1}(\cdot|U_1) - G_{2|1}(\cdot|V_1)\|_{\mathrm{TV}}|U_1 = V_1], \\
&\|F_1(x)G(x, y)/G_2(x) - G(x, y)\|_{\mathrm{TV}} \\
&\quad = \|[F_1(x)/G_1(x) - 1]G(x, y)\|_{\mathrm{TV}} \\
&\quad \le \max_x \left\{ \left| \frac{P(U_1 = x)}{P(V_1 = x)} - 1 \right| \|G(x, y)\|_{\mathrm{TV}} \right\} = \max_x \left| \frac{P(U_1 = x)}{P(V_1 = x)} - 1 \right|.
\end{aligned}$$

□

REFERENCES

ARTILES, L. M., GILL, R. D. and GUŢĂ, M. I. (2005). An invitation to quantum tomography. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **67** 109–134. MR2136642

BARNDORFF-NIELSEN, O. E., GILL, R. D. and JUPP, P. E. (2003). On quantum statistical inference (with discussion). *J. R. Stat. Soc. Ser. B Stat. Methodol.* **65** 775–816. MR2017871

BUNEA, F., SHE, Y. and WEGKAMP, M. H. (2011). Optimal selection of reduced rank estimators of high-dimensional matrices. *Ann. Statist.* **39** 1282–1309. MR2816355

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. MR2336856

CANDÈS, E. J. and PLAN, Y. (2009). Matrix completion with noise. *Proceedings of the IEEE* **98** 925–936.

CANDÈS, E. J. and PLAN, Y. (2011). Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements. *IEEE Trans. Inform. Theory* **57** 2342–2359. MR2809094

CANDÈS, E. J. and RECHT, B. (2009). Exact matrix completion via convex optimization. *Found. Comput. Math.* **9** 717–772. MR2565240

CANDÈS, E. J. and TAO, T. (2010). The power of convex relaxation: Near-optimal matrix completion. *IEEE Trans. Inform. Theory* **56** 2053–2080. MR2723472

CARTER, A. V. (2002). Deficiency distance between multinomial and multivariate normal experiments. *Ann. Statist.* **30** 708–730. MR1922539

DONOHO, D. L. (2006). Compressed sensing. *IEEE Trans. Inform. Theory* **52** 1289–1306. MR2241189

GROSS, D. (2011). Recovering low-rank matrices from few coefficients in any basis. *IEEE Trans. Inform. Theory* **57** 1548–1566. MR2815834

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.

HOLEVO, A. S. (1982). *Probabilistic and Statistical Aspects of Quantum Theory*. *North-Holland Series in Statistics and Probability* **1**. North-Holland, Amsterdam. MR0681693

KESHAVAN, R. H., MONTANARI, A. and OH, S. (2010). Matrix completion from noisy entries. *J. Mach. Learn. Res.* **11** 2057–2078. MR2678022

KLOPP, O. (2011). Rank penalized estimators for high-dimensional matrices. *Electron. J. Stat.* **5** 1161–1183. MR2842903

KOLTCHINSKII, V. (2011). Von Neumann entropy penalization and low-rank matrix estimation. *Ann. Statist.* **39** 2936–2973. MR3012397

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. MR2906869

LE CAM, L. (1986). *Asymptotic Methods in Statistical Decision Theory*. Springer, New York. MR0856411

LE CAM, L. and YANG, G. L. (2000). *Asymptotics in Statistics*: *Some Basic Concepts*, 2nd ed. Springer, New York. MR1784901

NEGAHBAN, S. and WAINWRIGHT, M. J. (2011). Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *Ann. Statist.* **39** 1069–1097. MR2816348

NIELSEN, M. A. and CHUANG, I. L. (2000). *Quantum Computation and Quantum Information*. Cambridge Univ. Press, Cambridge. MR1796805

RECHT, B. (2011). A simpler approach to matrix completion. *J. Mach. Learn. Res.* **12** 3413–3430. MR2877360

ROHDE, A. and TSYBAKOV, A. B. (2011). Estimation of high-dimensional low-rank matrices. *Ann. Statist.* **39** 887–930. MR2816342

SAKURAI, J. J. and NAPOLITANO, J. (2010). *Modern Quantum Mechanics*, 2nd ed. Addison-Wesley, Reading, MA.

SHANKAR, R. (1994). *Principles of Quantum Mechanics*, 2nd ed. Plenum, New York. MR1343488

VIDAKOVIC, B. (1999). *Statistical Modeling by Wavelets*. Wiley, New York. MR1681904

WANG, Y. (2002). Asymptotic nonequivalence of Garch models and diffusions. *Ann. Statist.* **30** 754–783. MR1922541

WANG, Y. (2011). Quantum Monte Carlo simulation. *Ann. Appl. Stat.* **5** 669–683. MR2840170

WANG, Y. (2012). Quantum computation and quantum information. *Statist. Sci.* **27** 373–394. MR3012432

WITTEN, D. M., TIBSHIRANI, R. and HASTIE, T. (2009). A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics* **10** 515–534.

DEPARTMENT OF STATISTICS
UNIVERSITY OF WISCONSIN–MADISON
1300 UNIVERSITY AVENUE
MADISON, WISCONSIN 53706
USA
E-MAIL: yzwang@stat.wisc.edu