

Asymptotic Efficiency and Finite Sample Performance of Frequentist Quantum State Estimation

Raj Chakrabarti

Department of Chemistry, Princeton University, Princeton, NJ 08544

Anisha Ghosh

Department of Economics, London School of Economics, London, UK

(Dated: December 27, 2008)

Abstract

We undertake a detailed study of the performance of maximum likelihood estimation (MLE) of the density matrix of finite-dimensional quantum systems in order to interrogate generic properties of frequentist quantum state estimation. Existing literature on frequentist quantum estimation has not rigorously examined the finite sample performance of the estimators and associated methods of statistical inference. While MLE is usually preferred on the basis of its asymptotic properties - it achieves the Cramer-Rao (CR) lower bound - the finite sample properties are often less than optimal. Moreover, in quantum estimation, there are multiple CR-type bounds, with the maximum asymptotic efficiency depending on the choice of measurements. The tightest quantum Cramer-Rao lower bound is not even asymptotically achievable in most circumstances. Here, we compare the asymptotic and finite sample efficiencies of quantum state estimators and test statistics corresponding to various CR bounds for spin-1/2 (one qubit) and spin-1 systems. We show that, in each of these cases, the finite sample properties of MLE differ significantly from the corresponding asymptotic ones for experimentally realistic sample sizes, and that the relative finite sample variances across different measurement strategies are much closer to 1 than the corresponding asymptotic relative efficiencies. These results indicate that in order to fully exploit the information geometry of quantum states and achieve smaller reconstruction errors, the use of Bayesian state reconstruction methods - which, unlike frequentist methods, do not rely on asymptotic properties - is necessary, since the estimation error is typically lower due to the incorporation of prior knowledge.

I. INTRODUCTION

Perhaps the most fundamental problem in quantum statistical inference (QSI) and quantum information theory is the reconstruction of the density matrix of a quantum system on the basis of a limited (finite) number of quantum observations. Due to the rapidly growing interest in quantum computation and quantum control, the ability to retrieve the maximum amount of information about a quantum state based on the smallest number of measurements is a subject of paramount importance. The accuracy of all derivative forms of QSI, including process estimation, are ultimately determined by that of the underlying state estimation.

Methods for quantum state estimation can be formally subdivided into three categories. The first, tomographic inversion [4], is the least computationally expensive and most popular technique. However, tomographic inversion cannot enforce the constraints on the density matrix during estimation, and hence is not amenable to rigorous statistical inference. The second class consists of frequentist techniques of inference based on a likelihood function, the most notable of which is maximum likelihood estimation (MLE) [1]. This class of methods avoids the problems associated with tomography, but nonetheless delivers distributional results for the estimators of the parameters of interest under the assumption of an infinite number of measurements. Hence, for finite sample sizes, the estimated confidence intervals for the parameters may have actual coverages quite different from the corresponding asymptotic ones and, as is well known in the statistics literature, this divergence typically tends to become more pronounced as the number of parameters and nonlinearity of the model increase. All forms of frequentist inference, including tomographic inversion and MLE, require a complete observation level, i.e. $N^2 - 1$ linearly independent observable operators where N is the Hilbert space dimension, in order to estimate all parameters. The third type, Bayesian estimation [8, 9, 11, 13, 16], which is based on updating a prior plausibility distribution about the parameters based on observed data, lends itself readily to both incomplete observation levels and a finite number of measurements. Moreover, the estimation error that arises in Bayesian methods

is typically lower, reflected in shorter lengths of Bayesian confidence intervals compared to their frequentist counterparts, due to the use of such priors [21].

In this paper, we investigate the finite sample properties of maximum likelihood estimators of the density matrix of finite-dimensional, spin-1/2 (one qubit) and spin-1 quantum systems. Among frequentist estimation techniques, MLE is usually preferred on the basis of its asymptotic properties - (i) the MLE estimator is asymptotically efficient in the sense that its asymptotic variance achieves the Cramer-Rao lower bound for consistent estimators, (ii) likelihood based testing approaches are optimal, in the sense of the Neyman-Pearson Fundamental Lemma and the Large Deviation Principle, for a broad class of hypothesis testing problems. However, the finite sample properties of MLE are often less than optimal. Existing literature on MLE has not rigorously examined the finite sample performance of the estimators and associated methods of statistical inference. To our knowledge, there is only one extant study on the efficiency of frequentist quantum state estimation [10], and robust numerical techniques are lacking. Minimizing finite sample estimation errors is essential for making optimal quantum decisions, which underlie emerging quantum feedback control and computation strategies [6]. Lack of rigorous understanding of the small sample estimation errors has inhibited the application of MLE to practical problems in quantum information and control.

In addition, the assessment of the estimation error in QSI is complicated by the existence of multiple measurement strategies due to the noncommutativity of the probability space, and ambiguities regarding the optimal measurement strategy. Optimal quantum measurement theory has been studied from two different perspectives: measurements that minimize estimator variance on *average* across all possible density matrices, and measurements that minimize estimator variance assuming a particular type of density matrix. The relative merits of these strategies in finite samples have not been properly investigated. Moreover, it has recently been shown that the theoretically predicted optimal quantum efficiency bound (the quantum Cramer-Rao bound) is unachievable in most circumstances of practical interest, indicating that average-case optimal measurement strategies - or even suboptimal strategies, which are often easier to implement

experimentally - may be preferred. We compare the efficiencies of optimal, average-case optimal and representative suboptimal measurement strategies.

We shed light on the following issues:

- How are the small sample biases of the MLE affected by the choices of the measurement bases and the sample size?
- How do the confidence regions of parameter estimators constructed using the optimal measurements perform relative to confidence regions based on nonoptimal measurements and how does the coverage of the intervals change with increase in the sample size?
- How does the small sample behavior of the test statistics for physical quantities of interest in quantum decision theory compare to their known asymptotic behavior for different choices of measurement bases and sample size?

We show that the finite sample properties of MLE differ significantly from the corresponding asymptotic ones for experimentally realistic sample sizes, and that the relative finite sample variances across different measurement strategies are much closer to 1 than the corresponding asymptotic relative efficiencies.

The paper is organized as follows. Section II discusses the asymptotic properties of MLE. Section III details the properties of MLE of the quantum density matrix, including the effect of measurement strategy on the Fisher information and the asymptotic efficiency (and the quantum CRB). In Section IV, we describe the Bloch vector representation, the most common approach used to parameterize the density matrix, that is employed in the paper. In Section V, we survey the various types of quantum measurements considered in this work, and their associated Fisher information matrices. Section VI discusses the details of the globally convergent Newton-Raphson and quasi-Newton algorithms used for constrained parameter optimization, along with methods for kernel density estimation of finite sample distributions. Section VII provides the estimation results for various types of spin-1/2 and spin-1 density matrices, a comparison of the finite sample versus

asymptotic properties of estimators, and the role of the choice of measurements. Section VIII compares the finite sample and asymptotic properties of the test statistics for physical quantities of interest in quantum control and quantum decision theory. Finally, in the concluding Section IX, we draw conclusions regarding the efficiency of frequentist quantum state estimation and discuss (Bayesian) extensions.

II. PROPERTIES OF FREQUENTIST ESTIMATORS

A. Maximum Likelihood Estimators

Let $x = (x_1, \dots, x_m)$ be an *i.i.d.* sample of size m from a population with probability density function $p(x|\theta)$ which depends on the unknown parameter vector θ whose true value is θ_0 . The value of the parameter vector that maximizes the likelihood function - the joint density of the sample defined as a function of the unknown parameter vector θ - is called the MLE of θ :

$$\hat{\theta}_{ML}^m = \arg \max_{\theta \in \Theta} L(\theta|x) = \arg \max_{\theta \in \Theta} \left(\prod_{i=1}^m p(x_i|\theta) \right), \quad (1)$$

where Θ denotes the admissible parameter space. Typically, the logarithm of the likelihood function, $\ln L(\theta|x)$, is easier to maximize numerically because of its separability. By maximizing the log likelihood, MLE minimizes the Kullback-Leibler distance between the estimated and true probability distributions.

MLE has several properties that make it an attractive frequentist estimator:

1. *Consistency*: An estimator $\hat{\theta}^m$ is *consistent* for the parameter θ (written as $\text{plim} \hat{\theta}^m = \theta_0$) if for every $\epsilon > 0$,

$$\lim_{m \rightarrow \infty} P_{\theta} \left\{ |\hat{\theta}^m - \theta_0| \geq \epsilon \right\} = 0.$$

The MLE is consistent: $\text{plim} \hat{\theta}_{ML}^m = \theta_0$.

2. *Invariance:* The MLE of $c(\theta)$ is $c(\hat{\theta}_{ML}^m)$.

3. *Asymptotic Normality.* For a sequence of estimators $\hat{\theta}^m$, if $k_m (\hat{\theta}^m - \theta_0) \xrightarrow{d} N(0, \Sigma)$ as $m \rightarrow \infty$, where \xrightarrow{d} denotes convergence in distribution and k_m is any function of m , $\hat{\theta}^m$ is said to be $\sqrt{k_m}$ -consistent for θ and have an asymptotic normal distribution with asymptotic covariance matrix Σ .

The MLE is asymptotically normally distributed:

$$\sqrt{m} [\hat{\theta}_{ML}^m - \theta_0] \rightarrow \mathcal{N}[0, I^{-1}(\theta_0)],$$

$$\text{where } I(\theta_0) = -\text{E} \left[\frac{\partial^2 \ln L(\theta_0|x)}{\partial \theta \partial \theta'} \right].$$

$I(\theta_0)$ is called the expected Fisher information matrix. Note that the asymptotic covariance matrix of the MLE is a function of the unknown parameters. Two approaches exist for consistent estimation of the expected Fisher Information matrix thereby providing feasible versions of the observed Fisher Information matrix. The first estimator replaces the expected second derivatives matrix of the log likelihood function with its sample mean evaluated at the maximum likelihood estimates,

$$\hat{I}_1(\hat{\theta}_{ML}^m) = - \left[\frac{\partial^2 \ln L(\hat{\theta}_{ML}^m|x)}{\partial \theta \partial \theta'} \right] \quad (2)$$

The second estimator is based on the result that the expected second derivatives matrix is the covariance matrix of the first derivatives vector,

$$\hat{I}_2(\hat{\theta}_{ML}^m) = \left[\left(\frac{\partial \ln L(\hat{\theta}_{ML}^m|x)}{\partial \theta} \right) \left(\frac{\partial \ln L(\hat{\theta}_{ML}^m|x)}{\partial \theta} \right)^T \right] \quad (3)$$

4. *Asymptotically efficient.* A sequence of consistent estimators $\hat{\theta}^m$, where m is the sample size, is asymptotically efficient if $\sqrt{m} [\hat{\theta}^m - \theta_0] \rightarrow \mathcal{N}[0, I^{-1}(\theta_0)]$ in distribution where $I(\theta) = -\text{E} \left[\frac{\partial^2 \ln L(\theta|x)}{\partial \theta \partial \theta'} \right]$; $[mI(\theta_0)]^{-1}$ is called the Cramer-Rao lower bound (CRB) for consistent estimators.

Property 4 is the subject of the following classic lemma of frequentist inference.

Lemma 1 *When there are no constraints on the parameters, the eigenvalues of the covariance matrix of parameter estimates of an asymptotically unbiased frequentist estimator are bounded from below by the eigenvalues of $(mI(\theta_0))^{-1} = \left\{ -m\mathbb{E} \left[\frac{\partial^2 \ln L(\theta_0|x)}{\partial\theta\partial\theta'} \right] \right\}^{-1}$. The maximum likelihood estimator asymptotically (i.e., in the limit of an infinite number of measurements) achieves this lower bound.*

We present a brief proof of these standard results in the Appendix since we will later be examining their extension to quantum estimation.

In addition, likelihood based testing approaches are optimal, in the sense of the Neyman-Pearson Fundamental Lemma and the Large Deviation Principle, for a broad class of hypothesis testing problems. A hypothesis test T , based on a test statistic $W(x)$ - a function of the data - is a rule that specifies the subset of the sample space for which values of x (the acceptance region A) the null hypothesis $H_0 : \theta \in \Theta_0$ is accepted, and for which values (the rejection region R) it is rejected (and the alternative hypothesis $H_1 : \theta \in \Theta_0^c$, where Θ_0^c is the complement of Θ_0 , is accepted).

The *size* $s(T)$ of a hypothesis test T is the probability of rejecting the null hypothesis given that it is true. The *power* $p(T)$ of a hypothesis test T is given by the probability of rejecting the null hypothesis given that it is false. Typically, when defining the power of a test, one assigns the test to a *class* based on its size. For $0 \leq \alpha \leq 1$ a test with power function $\beta(\theta)$ is a size α test if $\sup_{\theta \in \Theta_0} \beta(\theta) = \alpha$.

Definition 1 *Given a class of hypothesis tests for testing $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \in \Theta_0^c$, where $\Theta_0 \cup \Theta_0^c = \Omega$, the admissible parameter space, a test in that class with power function $\beta(\theta)$ is uniformly most powerful (UMP) if $\beta(\theta) \geq \beta'(\theta)$ for every $\beta'(\theta)$ in that class.*

An important type of hypothesis test based on MLE is the likelihood ratio test. Likelihood ratio test statistics take the form

$$\lambda(x) = \frac{L(\hat{\theta}_t|x)}{L(\hat{\theta}|x)},$$

where $\hat{\theta}_t$ is the constrained MLE estimator. It can be shown [7, 20] that likelihood ratio tests are UMP in their respective classes. In our simulation analysis, we rely on an

alternative testing procedure, namely the Wald test that inherits the optimality properties of the likelihood ratio test on account of their asymptotic equivalence (Godfrey (1988)). This choice is made primarily on the basis of the ease of computation. The likelihood ratio test requires calculation of both restricted and unrestricted estimators. The Wald test, on the other hand, requires only the unrestricted estimator. Since, most of our hypothesis tests involve nonlinear constraints and estimation of the constrained model is cumbersome, we rely on the Wald testing procedure.

Because of properties 1-4 and the fact that likelihood ratio tests are UMP, the maximum likelihood estimation methodology is considered the most desirable among frequentist estimation techniques.

B. Bayesian Estimators

In the alternative paradigm of Bayesian estimation, the estimation error that arises is typically lower than that in frequentist estimation. Bayesian estimation differs fundamentally from frequentist methods in that the parameters θ_i are treated as random variables. The goal is not to estimate a unique probability distribution, which can only truly be known in the limit of an infinite number of measurements, but to update a so-called prior plausibility distribution to a posterior plausibility distribution based (only) on the observed data. The posterior plausibility distribution density is given by

$$p(\theta | x \wedge I) d\theta = \frac{L(x | \theta) p(\theta | I) d\theta}{\int_{\Omega} L(x | \theta) p(\theta | I) d\theta}, \quad (4)$$

where $p(\theta | I)$ denotes the prior plausibility distribution, i.e., the probability of the parameter vector taking on the value θ given our prior information I regarding the parameter space, $L(x | \theta)$ denotes the joint probability density, and Ω denotes the space of admissible parameters θ .

Conditional simulation is required to retrieve quantities of interest, including the parameter estimates, which are given formally by the posterior means

$$\hat{\theta}_i = \frac{\int_{\Omega} \theta_i p(\theta | x \wedge I) d\theta}{\int_{\Omega} p(\theta | x \wedge I) d\theta}.$$

Unlike frequentist estimators, the notion of a confidence or credible interval can be rigorously defined for finite samples only for Bayesian estimators. This allows one to rigorously report finite sample uncertainties. The $100 * c\%$ Bayesian credible interval on parameter estimate $\hat{\theta}_i$ is the interval $[a, b]$ such that

$$\int_{-\infty}^{\infty} \cdots \int_a^b \cdots \int_{-\infty}^{\infty} p(\theta | x \wedge I) d\theta_1, \cdots d\theta_i \cdots d\theta_n = c. \quad (5)$$

III. ESTIMATION OF THE DENSITY MATRIX

A. Quantum estimation and the likelihood function for state reconstruction

In this section we apply and extend the classical estimation framework in Section II to estimation of the quantum density matrix. Quantum statistical inference is based on the notion of a quantum probability space.

Definition 2 Consider a measurable space (χ, \mathcal{A}) , where χ is the set of all possible measurement outcomes and \mathcal{A} is the σ -Algebra of subsets of χ . An operator-valued probability measure (POVM) is a (set) function $M : \mathcal{A} \rightarrow B(\mathcal{H})$, where $B(\mathcal{H})$ is the set of bounded positive semidefinite, Hermitian linear operators on a Hilbert space \mathcal{H} .

Definition 3 A quantum probability space is a measurable space (χ, \mathcal{A}) , together with an operator-valued probability measure M , such that the outcome $x \in \chi$ has probability density function $p(x|\theta) = \text{Tr}(\rho(\theta)F(x))$, where $F(x) \in \mathcal{H}$ and $\rho(\theta)$ is a positive-semidefinite, unit trace, Hermitian matrix (parametrized by a vector θ of parameters) called the density matrix.

Note that x in this context denotes the outcome of a single measurement. The measure M is explicitly defined in terms of the operators $F(x)$, and an associated scalar-valued probability measure μ satisfying $\mu(\chi) = 1$, through the relation $M(A) = \int_A F(x) \mu(dx)$. For N -dimensional (finite) quantum systems, we write $F(x_i) = F_i$, $i = 1, \cdots, N^2 - 1$,

and denote the outcome of the k -th measurement F_{i_k} . F_i is then a $N \times N$ positive-semidefinite Hermitian matrix. The outcomes x are indexed by the set of integers $(1, \dots, N^2 - 1)$. In this case, we have $\mu(\chi) = \frac{1}{N} \text{Tr}(M(\chi)) = 1$. values in

For simplicity of exposition, we collect all the distinct parameters of the density matrix, ρ , into the $(N^2 - 1)$ -dimensional vector, θ . The most convenient parameterization of $\rho(\theta)$ differs based on the state estimation method; various parameterizations are discussed in Section IV. The likelihood function (1) for quantum state estimation is then

$$L(\rho(\theta) | x) = \prod_{k=1}^m \text{Tr}(\rho(\theta) F_{i_k}) \quad (6)$$

which may be interpreted as the probability of obtaining the set of observed outcomes for a given density matrix $\rho(\theta)$. The MLE of the density matrix seeks to identify the admissible parameter vector θ at which this likelihood is maximal.

One frequentist method of quantum state estimation that is not based on a likelihood function is tomographic inversion [14]. In this method, the parameter estimates $\hat{\theta}_j$, $1 \leq j \leq N^2 - 1$ are obtained by inverting a system of equations of the form

$$\text{Tr}(\rho(\theta) F_i) = c_i, \quad 1 \leq i \leq N^2 - 1, \quad (7)$$

where c_i denotes the frequency with which outcome F_i is observed. Introducing the notation $A_{ij} = \frac{\partial \text{Tr}(\rho(\theta) F_i)}{\partial \theta_j}$, we may solve for the estimated parameter vector as $\hat{\theta} = A^{-1} \mathbf{c}$ for any parameterization $\rho(\theta)$ that is linear in θ (see Section IV). However, this method has two major drawbacks. First, since no parametrization $\rho(\theta)$ guarantees satisfaction of each of the positive-semidefiniteness, unit trace, and Hermiticity constraints on ρ (see Section IV), direct inversion can yield unphysical density matrix estimates. Second, in the absence of a likelihood function, there is no way to assign a variance to the estimator on the basis of a single sample. Even if the finite sample variances from multiple estimations are used to approximate the estimator variance, the parameter covariances will not be correct because the $N^2 - 1$ parameters are not estimated jointly in the presence of all constraints. For these reasons, we do not consider tomographic inversion in our assessment of the performance of frequentist quantum estimation.

B. Quantum measurement bases

A *resolution of the identity* on a Hilbert space \mathcal{H} of quantum states is a normalized operator-valued measure. Generally, the resolution of the identity satisfies

$$M(\chi) = \int_{\chi} F(x)\mu(dx) = I.$$

For finite-dimensional systems, to which we restrict our attention,

$$\sum_i F_i \mu(x_i) = I_N,$$

where I_N denotes the $N \times N$ identity matrix.

An important feature of quantum probability is that the operators F_i do not all mutually commute. The subsets $A \in \mathcal{A}$ of the space of possible measurement outcomes χ may be chosen to be pairwise disjoint and associated with subsets $M_A = \{F_i, \dots, F_{i+N-1}\}$ of commuting observables whose members do not commute with those of any other subset. The M_A are then said to constitute distinct measurement “bases”.

Writing each F_i as an $N \times N$ Hermitian matrix, it is convenient to represent each basis $M_{A^{(r)}}$ in terms of an $N \times N$ matrix of common eigenvectors $V^{(r)}$, $1 \leq r \leq N + 1$. Given that the density matrix is a function of $N^2 - 1$ independent parameters, the minimal cardinality resolution of the identity must be composed of $N + 1$ subsets $A \in \mathcal{A}$; $(N + 1)(N - 1) = N^2 - 1$. We note, however, that many resolutions of the identity are redundant in that they are associated with $n > N + 1$ bases M_A .

In the current work, the data x consist of m_i measurement outcomes in each of p measurement bases with $\sum_{i=1}^p m_i = m$. The measurement bases used are discussed further in Section V.

C. Quantum maximum likelihood estimation

Among frequentist estimation techniques, MLE has been employed most extensively for reconstruction of quantum states. In this method, we aim to identify the maximum

of the likelihood function (6) over the set of *admissible* density matrices. Assuming the constraints on the parameter vector θ are of the general form $a_j(\theta) \geq 0$, $j = 1, \dots, N$, this problem can be formulated in terms of the Lagrangian function

$$\mathcal{L}(\theta, \lambda, \gamma|x) = \ln \left[\prod_{k=1}^m \text{Tr}(\rho(\theta) F_i^k) \right] + \sum_{j=1}^N \lambda_j (a_j(\theta) - \gamma_j^2), \quad (8)$$

where the first term is $\ln L(\theta|x)$, the γ_j denote slack variables ($\gamma_j = 0$ in the case of an equality constraint) and the λ_j denote Lagrange multipliers. It is convenient to order the N constraints such that the first constraint enforces the unit trace of ρ , and the following $N - 1$ constraints enforce its positive semidefiniteness. For parameterizations where positive semidefiniteness is implicit in the parametrization (such as the Cholesky parametrization), $\lambda_j = 0$, $j = 2, \dots, N$, and for parameterizations where the unit trace constraint is implicit in the parameterization (such as the Bloch vector parametrization), $\lambda_1 = 0$. We denote the vector of parameters $(\theta, \lambda, \gamma) \equiv \mathbf{t}$. Finding the constrained optimum corresponding to this Lagrangian entails searching for parameters \mathbf{t} θ_i and slack variables γ_j that render the gradient vectors $\nabla L(\theta)$ and a linear combination of $\nabla(a_j(\theta) - \gamma_j)$, $j = 1, \dots, N$ parallel. There are two common approaches to solving this problem: 1) minimization of the “sum of squares” (of the first-order conditions) function $\sum_i \left(\frac{\partial \mathcal{L}}{\partial t_i} \right)^2$; 2) finding the roots of the system of nonlinear equations $\frac{\partial \mathcal{L}}{\partial \mathbf{t}} = 0$ using the Newton-Raphson (NR) method. In fact, methods 1) and 2) may be combined to produce a globally convergent NR algorithm. Further details on solving the constrained optimization problem are provided in Section VI.

Note that the MLE obtained by maximizing the likelihood function in the absence of any constraints is consistent, asymptotically normally distributed, and has an asymptotic covariance matrix equal to the inverse of m times the Fisher information matrix (see Section II for details). However, on constrained parameter spaces such as that found in density matrix estimation, the asymptotic covariance matrix is no longer given by $(mI(\theta_0))^{-1}$. In the present case, the parameters are subject to up to N constraints: $c_1(\theta) = \text{Tr}(\rho(\theta)) - 1 = 0$; $c_{j+1}(\theta) = a_j(\theta) \geq 0$, $j = 1, \dots, N - 1$. Linearizing each constraint with respect to the parameters θ_j of the density matrix, we have $R_{jk}\theta_k = 0$,

where R is a $(N^2 - 1) \times q$ -dimensional Jacobian matrix whose k -th column r_k consists of elements $R_{jk} = \left. \frac{\partial a_j(\theta)}{\partial \theta_k} \right|_{\theta_k=0}$.

Let $\tilde{\theta}$ denote the constrained parameter estimate. Under standard regularity conditions it can be shown (see Appendix for derivation) that the asymptotic variance-covariance matrix for $\hat{\theta}$ is given by

$$\Sigma = \frac{1}{m} \left[I^{-1}(\theta_0) - I^{-1}(\theta_0)R(\theta_0)(R^T(\theta_0)I^{-1}(\theta_0)R(\theta_0))^{-1}R^T(\theta_0)I^{-1}(\theta_0) \right]. \quad (9)$$

We estimate the asymptotic covariance matrix above as follows:

$$\hat{\Sigma} = \frac{1}{m} \left[\hat{I}^{-1}(\tilde{\theta}) - \hat{I}^{-1}(\tilde{\theta})R(\tilde{\theta})(R^T(\tilde{\theta})\hat{I}^{-1}(\tilde{\theta})R(\tilde{\theta}))^{-1}R^T(\tilde{\theta})\hat{I}^{-1}(\tilde{\theta}) \right]. \quad (10)$$

It is possible to also work out the covariances of the slack variables and Lagrange multipliers, but we are not interested in them here.

D. Asymptotic properties of quantum maximum likelihood estimators

1. Quantum Fisher-Bures information and the quantum Cramer-Rao bound

A unique feature of quantum statistics is the quantum Cramer-Rao bound, an extension/generalization of the classical Cramer-Rao bound that originates due to the dependence of the quantum Fisher information on the mode of measurement. Work in quantum probability theory [9] has indicated that $\frac{1}{m}I(\theta_0)^{-1}$ for an arbitrary choice of measurement bases is generally not the tightest asymptotic lower bound achievable in quantum MLE. In quantum statistics, there are multiple Cramer-Rao type inequalities, each with its own associated (quantum) Fisher information. Some of these correspond to particular measurement strategies, whereas others are in fact unachievable. Specifically, instead of using the standard logarithmic derivative of the scalar likelihood (i.e., the classical *score function*), it is necessary to employ the so-called noncommutative logarithmic derivatives. These are defined implicitly in terms of matrix derivatives of ρ with respect to the parameters, $\frac{\partial \ln \rho}{\partial \theta_i}$. Due to the noncommutativity of the space of elementary events, there are several types of these logarithmic derivatives, each associated with its

own Fisher information. The one that is associated with the tightest Cramer-Rao type bound is the *symmetrized logarithmic derivative* (quantum score) l_θ^j , $1 \leq j \leq N^2 - 1$, which is a Hermitian operator defined implicitly by

$$\mathrm{Tr}\left\{\frac{\partial \hat{\rho}}{\partial \theta_j} F\right\} = \frac{1}{2} \mathrm{Tr}\{\hat{\rho}[l_\theta^j F + F l_\theta^j]\} \quad (11)$$

or

$$\frac{\partial \rho}{\partial \theta_j} = \frac{1}{2}(\rho l_\theta^j + l_\theta^j \rho). \quad (12)$$

The associated (expected) Fisher information is given by

$$\mathcal{I}_\theta^{ij} = \frac{1}{2}(l_\theta^i l_\theta^j + l_\theta^j l_\theta^i), \quad (13)$$

which is referred to as the quantum Fisher-Bures information, or simply the quantum Fisher information. Note that l_θ^j , and hence the associated measurement basis (if it exists) depends on the true ρ .

Braunstein and Caves [3], following Wootters [19], clarified the relationship between the classical expected Fisher information for the unknown parameters θ of a probability distribution $p(x|\theta)$ and the quantum expected Fisher information for a quantum state $\rho = \rho(\theta)$. For simplicity, consider the case of estimation of a single scalar parameter θ , which is straightforward to generalize to the vector case. We present the inequality derivation for the general case of an arbitrary Hilbert space (finite- or infinite-dimensional quantum systems). First note that for any single measurement, we can write the classical Fisher information in terms of the quantum score:

$$\begin{aligned} I(\theta_0|x) &= \int_x \frac{1}{L(\theta_0|x)} \left(\frac{\partial L(\theta_0|x)}{\partial \theta} \right)^2 \mu(dx) \\ &= \int_x \frac{1}{L} (\mathrm{Tr}(\rho l F(x)))^2 \mu(dx) \\ &= \int_x \frac{1}{L} (\mathrm{Tr}((\rho l + l \rho) F(x)))^2 \mu(dx) \\ &= \int_x \frac{1}{L} (\Re [\mathrm{Tr}(\rho l F(x))])^2 \mu(dx). \end{aligned}$$

where, for simplicity, x again denotes a single measurement (generalization to a set of multiple measurements is straightforward). Starting from this expression, we can establish an upper bound on the classical Fisher information $I(\theta_0|x)$ obtained through any type of measurement:

$$\begin{aligned}
I(\theta_0|x) &= \int_{\chi^+} L(\theta_0|x)^{-1} \Re [\text{Tr}(\rho l F(x))^2] \mu(dx) \\
&\leq \int_{\chi^+} L(\theta_0|x)^{-1} |\text{Tr}(\rho l F(x))|^2 \mu(dx) \\
&= \int_{\chi^+} L(\theta_0|x)^{-1} |\text{Tr}[(F(x)^{1/2} \rho^{1/2})^\dagger (F(x)^{1/2} l \rho^{1/2})]|^2 \cdot \mu(dx) \\
&= \int_{\chi^+} (\text{Tr}(\rho F(x))^{-1} |\text{Tr}[(F(x)^{1/2} \rho^{1/2})^\dagger (F(x)^{1/2} l \rho^{1/2})]|^2 \mu(dx) \\
&= \int_{\chi^+} (\text{Tr}(\rho F))^{-1} \text{Tr}(F^{1/2} \rho^{1/2} \rho^{1/2} F^{1/2}) \text{Tr}(F^{1/2} l \rho^{1/2} \rho^{1/2} l F^{1/2}) \mu(dx) \\
&\leq \int_{\chi^+} \text{Tr}(M(\chi^+) l \rho l) \mu(dx) \\
&\leq \text{Tr}(\rho l^2) \equiv \mathcal{I}(\theta_0),
\end{aligned}$$

where $M(\chi^+)$ denotes the subset of the POVM set that has positive outcome probability given the true ρ . The first inequality follows from $\Re(a + ib)^2 = |a + ib|^2 - \Im(a + ib)^2$. The second inequality is a consequence of the Cauchy-Schwarz inequality, $|\text{Tr}(A^\dagger B)| \leq \text{Tr}(A^\dagger A) \text{Tr}(B^\dagger B)$, with $A = F(x)^{1/2} \rho^{1/2}$, $B = F(x)^{1/2} l \rho^{1/2}$. The last inequality follows from $M(\chi) = M(\chi^+) + M(\chi_0) = I$. The *quantum Cramer-Rao bound* (QCRB), then, is that no consistent frequentist estimator, based on any measurement, has variance smaller than $\frac{1}{m} \mathcal{I}(\theta_0)^{-1}$.

In complex vector spaces, the C-S inequality is an equality iff $A = rB$, $r \in \mathbb{R}$, which implies $\text{Tr}(A^\dagger B)$ is real. So in order for the first two inequalities to be equalities, we must have, respectively [2]:

$$1) F\rho = rF\rho l \quad \text{and} \quad 2) \Im[\text{Tr}(F\rho l)] = 0. \quad (14)$$

The last inequality is an equality iff 3) $M(\chi^+) = M(\chi) = I$, i.e, iff all elements of the POVM set have positive outcome probability. It was originally believed that these three

conditions could always be achieved and that the quantum Fisher-Bures information is nothing else than the maximal classical Fisher information achievable over all possible measurements of quantum system, but recent work has revealed that the QCRB is not reachable under any choice of measurements, for most systems. In Section V, we examine a (special) case where the QCRB is achievable, the corresponding value of the QFI, the associated choice of measurement basis, and the expression for the asymptotic covariance matrix.

2. Average-case optimal Fisher information

The measurements that maximize the Fisher information depend on the true, unknown state of the quantum system. Although the achievability of the QCRB and the choice of measurement bases that can achieve it depends on the true ρ , there exists an approach to optimal measurement that is agnostic to the true value of ρ .

Wootters [19] proposed a construction of measurement bases that maximizes the *average* information (over the set of all possible density matrices) obtained via a set of m measurements. These so-called mutually unbiased measurement bases (MUB) are “maximally noncommutative” in the sense that a measurement in one basis provides no information as to the outcome of a measurement over a basis unbiased with respect to the current one. Let $I(\rho_0)$ denote the Fisher information given true state $\rho_0 = \rho(\theta_0)$. MUB aims to maximize the average Fisher information over all possible ρ_0 ’s:

$$\langle I(\tilde{\theta}) \rangle = \frac{1}{V_0} \int_{B_n} I(\tilde{\theta}, \rho_0) d\rho_0,$$

where B_n denotes the $n = N^2 - 1$ dimensional Bloch vector space of admissible $N \times N$ dimensional density matrices (see Section IV) and V_0 is the volume of B_n , by an appropriate choice of measurement bases. It can be shown that this is equivalent to maximizing the average *Kullback-Leibler (KL) information gain* $\langle D \rangle$ upon updating the flat prior distribution to the asymptotic multivariate normal distribution [22]. The KL

information gain is given by

$$D = \int_{\Omega} f(\tilde{\theta}) \ln \frac{f(\tilde{\theta})}{g(\tilde{\theta})} d\tilde{\theta}, \quad (15)$$

where in the present case $f(\tilde{\theta})$ is the asymptotic multivariate normal distribution over the Bloch vector space after measurements and estimation, and $g(\tilde{\theta})$ is the uniform distribution on the Bloch vector space. The KL gain is defined in terms of the Shannon information (entropy) of a distribution, $E_{\tilde{\theta}}(\ln f(\tilde{\theta})) = \int_{\Omega} f(\tilde{\theta}) \ln f(\tilde{\theta}) d\tilde{\theta}$. A measurement in each basis restricts the variance in $N - 1$ directions, leaving it infinitely broad in the other $N^2 - N$. For estimation of a single parameter, the Shannon entropy of the asymptotic normal distribution of parameter estimates is $-\frac{1}{2} \ln(\pi e) - \ln(\sigma)$.

Maximizing the information gain is equivalent to minimizing the ‘‘uncertainty volume’’ in the parameter space, which is equal to the volume of the Bloch vector space in the absence of measurements. The uncertainty distance for estimation of a single parameter is the standard deviation of the estimator; the uncertainty volume is the product of the standard deviations of the estimators for each of the parameters. In terms of uncertainty volumes, the information gained by measurements is

$$D = -\ln\left(\frac{W}{W_0}\right) - \left(\frac{N^2 - 1}{2}\right) \ln(\pi e),$$

where W is the uncertainty volume after measurements and estimation and W_0 is volume of the Bloch vector space.

Denote by T_r the $(N - 1)$ -dimensional subspace of $su(N)$ (or B_n ?) associated with measurement basis $V^{(r)}$. The total uncertainty volume is diminished by overlaps between the subspaces T_r . It can be shown [19] that this total volume W can be written $W = \frac{W_1 \cdots W_{N+1}}{\text{vol}(T_1, \dots, T_{N+1})}$ where T_r is the $(N - 1)$ -dimensional subspace of $su(N)$ associated with measurement basis $V^{(r)}$. Thus the Kullback-Leibler information gain (15) in updating the flat prior distribution to the asymptotically normal distribution is then $\langle D \rangle = -\sum_{r=1}^{N+1} \langle \ln(W_r) \rangle + \ln[\text{vol}(T_1, \dots, T_{N+1})] + \ln(W_0) - \left(\frac{N^2-1}{2}\right) \ln(\pi e)$.

$\langle \ln(W_r) \rangle$, the log of the average uncertainty volume in subspace T_r , does not depend on the choice of measurements since it is the (log of the) product of standard deviations of

multinomial parameters in a single measurement basis, averaged over all possible multinomial parameters p_1, \dots, p_N . Thus the average Kullback-Leibler or Fisher information is a function of only $\text{vol}(T_1, \dots, T_{N+1})$, which in turn is determined by the relative orientations of the bases (and not the parameters). The total uncertainty volume is minimized when (T_1, \dots, T_{N+1}) is a rectangular solid with the unit vectors defining the edges being orthogonal; this is equivalent to the condition that the subspaces T_1, \dots, T_{N+1} are mutually orthogonal. Wootters showed [19] that this condition is equivalent to requiring that

$$|\langle \mathbf{v}_i, \mathbf{w}_j \rangle| = \frac{1}{\sqrt{N}}, \quad (16)$$

where $|\langle \cdot, \cdot \rangle|$ denotes the modulus of the Hermitian inner product, and where $\mathbf{v}_i, \mathbf{w}_j$ are column vectors in the bases V, W respectively. Whereas mutual nonorthogonality of the edges of the parallelepiped may decrease the asymptotic uncertainty volume in particular subspaces T_r , the total asymptotic uncertainty volume is always increased by such nonorthogonality. Explicit formulas for measurement bases that satisfy (16) are known in the cases where the Hilbert space dimension N is the power of a prime (see Section V), and are discussed in Section VA.

It is important to assess the information loss for finite sample sizes incurred due to not using MUB or the measurements that maximize the QFI. In many experimental setups, it is not convenient to use these specialized bases. We aim to clarify the practical utility of the quantum Fisher information (and other Fisher informations) in MLE, and assess the extent to which frequentist quantum estimation can effectively make use of these optimal efficiencies.

E. Bayesian versus frequentist density matrix estimation

In the alternative, Bayesian approach to quantum state estimation, the posterior distribution $p(\theta | x \wedge I)$ (4) takes the form (use big int):

$$p(\theta | x \wedge I) d\theta = \frac{L(x | \theta) p(\theta | I) d\theta}{\int_{\Omega} L(x | \theta) p(\theta | I) d\theta} = \frac{[\prod_k \text{Tr}(F_{i_k} \rho(\theta))] p(\theta) d\theta}{\int_{\Omega} [\prod_k \text{Tr}(F_{i_k} \rho(\theta))] p(\theta) d\theta} \quad (17)$$

whereas the density matrix can be estimated by the posterior mean of each of its elements (corresponding to a quadratic “loss function”), namely

$$\begin{aligned} \rho_{x \wedge I} &\equiv \int_{\Omega} \rho(\theta) p(\theta | x \wedge I) d\theta \\ &= \frac{\int_{\Omega} \rho(\theta) \left[\prod_k \text{Tr}(F_{i_k}^{(k)} \rho) \right] p(\theta) d\theta}{\int_{\Omega} \left[\prod_k \text{Tr}(F_{i_k}^{(k)} \theta) \right] p(\theta) d\theta} \end{aligned}$$

Alternative loss functions can be used to retrieve other estimable quantiles of interest.

Bayesian credible intervals can be obtained according to expression (5) by sampling from the posterior density (17). These credible intervals do not rely on asymptotic results / Fisher information. We do not compute Bayesian integrals in this work; our goal is rather to determine whether such a need exists given the finite sample performance of the simpler frequentist estimators.

IV. PARAMETERIZATION

In maximum likelihood estimation of the quantum density matrix, a constrained optimization must be carried out, where the constraints correspond to preservation of the unit trace and positive semidefiniteness properties of the density matrix. The dimension of the parameter space increases quadratically with the Hilbert space dimension, necessitating the use of efficient parameterizations of the density matrix. The three most commonly used parameterizations are the Bloch vector [12], Euler angle [17] and the Cholesky [1] parameterizations. Within the last few years, considerable advancements have been made in extending the Bloch and Euler parameterizations to arbitrary N -dimensional Hilbert spaces. The Euler angle parameterization of the density matrix, which is based on the Euler angle parameterization of the special unitary group $SU(N)$, employs the generators of the Lie algebra $su(N)$ to parameterize ρ in terms of the Lie algebra exponential

map. The constraint equations are linear in the parameters, but because the parameters appear as exponents, the likelihood takes on a complicated form. The Cholesky parameterization $\rho = A^\dagger A$ with A upper triangular has real elements on the diagonal. ρ is then automatically positive-semidefinite, but has a nonlinear expression for the likelihood and is not convenient in other applications. Here, we employ the so-called Bloch vector parameterization, where the likelihood of an observable outcome, $\text{Tr}(\rho(\theta)F_i)$, is a simple linear function of the parameter vector θ . Moreover, the Bloch vector parameterization is perhaps the most commonly used in the statistical physics of finite-dimensional quantum systems (especially in quantum information applications).

Bloch vector parameterization

In the Bloch vector parameterization [12], the Hermitian operator ρ is parameterized in terms of an orthogonal basis $\{\lambda_j\}$, $1 \leq j \leq N^2 - 1$ for the vector space of Hermitian operators on an N -dimensional Hilbert space. In two dimensions, these are the familiar Pauli spin matrices, whereas in three dimensions they are the so-called Gell-Mann matrices. ρ can then be written

$$\rho \equiv \rho(\theta) = \frac{1}{N}I_N + \frac{1}{2} \sum_{j=1}^{N^2-1} \theta_j \lambda_j, \quad (\theta_1, \dots, \theta_{N^2-1}) \equiv \theta \in B_{N^2-1} \subset R^{N^2-1}, \quad (18)$$

where the N^2-1 matrices λ_j satisfy the conditions a) $\lambda_j = \lambda_j^\dagger$, b) $\text{Tr}(\lambda_j) = 0$, c) $\text{Tr}(\lambda_i \lambda_j) = 2\delta_{ij}$. These are the defining conditions of the generators of $SU(N)$ that generalize the Pauli spin matrices. The θ_j are given by $\theta_j(\rho) = \text{Tr}(\lambda_j \rho)$ (i.e., are expectation values of the observable generators). The vector $\theta_j \lambda_j$ is called the Bloch vector.

B_{N^2-1} is a compact convex subset of \mathbb{R}^{N^2-1} . Let $a_i(\lambda)$ denote the coefficients of the characteristic polynomial of $\rho, \det(yI_N - \rho)$, where ρ takes the form above. The unit trace constraint is automatically satisfied in the Bloch vector parameterization [23]. It can be shown that the conditions of Hermiticity and positive-semidefiniteness of ρ correspond to

the following definition of the ‘‘Bloch vector set’’ B_{N^2-1} of admissible values of θ

$$B_{N^2-1} \equiv \{\theta \in \mathbb{R}^{N^2-1} \mid a_i(\theta) \geq 0, \quad i = 1, \dots, N\}. \quad (19)$$

The a_i in the above definition of B are polynomials whose coefficients are expressed in terms of the structure constants of the Lie algebra $su(N)$, and will be written explicitly below for $N = 2, 3, 4$. The structure constants, which characterize the generators of $su(N)$, are the elements of the completely antisymmetric and completely symmetric tensors f and g , respectively defined by the relations:

$$\begin{aligned} [\lambda_i, \lambda_j] &= 2\iota f_{ijk} \lambda_k \\ [\lambda_i, \lambda_j]_+ &= \frac{4}{N} \delta_{ij} I_N + 2g_{ijk} \lambda_k, \end{aligned}$$

which can be solved ([17]) for f_{ijk} , g_{ijk} :

$$f_{ijk} = \frac{1}{4\iota} \text{Tr}\{[\lambda_i, \lambda_j] \lambda_k\} \quad (20)$$

$$g_{ijk} = \frac{1}{4} \text{Tr}\{([\lambda_i, \lambda_j]_+ - \frac{4}{N} \delta_{ij}) \lambda_k\}. \quad (21)$$

where $[\cdot, \cdot]$ denotes the antisymmetric commutator and $[\cdot, \cdot]_+$ denotes the symmetric commutator, and where we have used the Einstein (implicit) summation convention for repeated indices. It can be shown [12] that the generators λ_i that satisfy these conditions can be expressed:

$$\{\lambda_i\}_{i=1}^{N^2-1} = \{\{u_{jk}\}, \{v_{jk}\}, \{w_l\}\}$$

where

$$\{u_{jk}\} = \{ |j\rangle\langle k| + |k\rangle\langle j| \mid 1 \leq j < k \leq N \} \quad (22)$$

$$\{v_{jk}\} = \{ \iota(|j\rangle\langle k| - |k\rangle\langle j|) \mid 1 \leq j < k \leq N \}; \quad (23)$$

$$\{w_l\} = \left\{ \frac{\sqrt{2}}{l(l+1)} \sum_{j=1}^l (|j\rangle\langle j| - l|l+1\rangle\langle l+1|) \mid 1 \leq l \leq N-1 \right\}. \quad (24)$$

The generators λ_i of $SU(3)$ are presented explicitly in Appendix A.

1. Spin-1/2 systems

When $N = 2$, the conditions $a_i \geq 0$ (equation (19)) correspond to $1!a_1 = 1$ and

$$2!a_2 = \frac{N-1}{N} - \frac{1}{2}|\theta|^2 \geq 0. \quad (25)$$

The latter condition defines the familiar Bloch sphere for spin- $\frac{1}{2}$ systems. The Lagrangian (8) in this case becomes

$$\mathcal{L}(\theta, \lambda) = \ln \left[\prod_{k=1}^m \text{Tr}(\rho(\theta)F_{i_k}) \right] + \lambda_2 \left(\frac{N-1}{N} - \frac{1}{2}|\theta|^2 - \gamma_2^2 \right), \quad (26)$$

where we have omitted the constant term originating from a_1 since it is independent of θ .

2. Spin-1 systems

For $N = 3$, in addition to the constraint (25), we have

$$3!a_3 = \frac{(N-1)(N-2)}{N^2} - \frac{3(N-2)}{2N}|\theta|^2 + \frac{1}{2}g_{ijk}\theta_i\theta_j\theta_k \geq 0, \quad (27)$$

where the structure constants g_{ijk} are components of the completely symmetric tensor of the Lie algebra $su(N)$ (Eq. 20). The Lagrangian (8) for spin-1 systems then becomes

$$\begin{aligned} \mathcal{L}(\theta, \lambda) = \ln \left[\prod_{k=1}^m \text{Tr}(\rho(\theta)F_{i_k}) \right] + \lambda_2 \left(\frac{N-1}{N} - \frac{1}{2}|\theta|^2 - \gamma_2^2 \right) + \\ \lambda_3 \left[\frac{(N-1)(N-2)}{N^2} - \frac{3(N-2)}{2N}|\theta|^2 + \frac{1}{2}g_{ijk}\theta_i\theta_j\theta_k - \gamma_3^2 \right], \quad (28) \end{aligned}$$

where we have (again) used the Einstein summation convention for repeated indices, i.e., $\frac{1}{2}g_{ijk}\theta_i\theta_j\theta_k = \sum_{ijk} g_{ijk}\theta_i\theta_j\theta_k$.

Note that while for $N = 2$, the Bloch vector space is exactly a ball, the additional constraints starting with $a_3 \geq 0$ restrict the Bloch vector space for $N = 3$ and higher dimensions to a proper subset of a ball. Since the structure constants of $su(N)$ for $N \geq 3$ have no rotational invariance, neither do these conditions. The Bloch vector space has an asymmetric structure in \mathbb{R}^{N^2-1} for $N \geq 3$.

3. Spin-3/2 systems

Spin $\frac{3}{2}$ systems ($N = 4$) correspond to the important case of coupled qubits, which arise frequently in quantum information processing. Here, in addition to conditions (25) and (27), we have

$$4!a_4 = \frac{(N-1)(N-2)(N-3)}{N^3} - \frac{3(N-2)(N-3)}{N^2}|\theta|^2 + \frac{3(N-2)}{4N}|\theta|^4 + \frac{2(N-2)}{N}g_{ijk}\theta_i\theta_j\theta_k - \frac{3}{4}g_{ijk}g_{klm}\theta_i\theta_j\theta_l\theta_m. \quad (29)$$

In the present work, we report state estimation results for spin-1/2 and spin-1 systems; extensions to spin-3/2 systems will be considered in a future work.

In the Bloch vector parameterization, the Fisher information takes on a particularly simple analytical form. For $N = 2$, we have for the score vectors

$$\begin{aligned} \frac{\partial \ln L(\theta)}{\partial \theta_1} &= \frac{1}{2} \sum_{k=1}^m \frac{1}{\text{Tr}(\rho F_{i_k})} [F_{i_k}(2, 1) + F_{i_k}(1, 2)] \\ \frac{\partial \ln L(\theta)}{\partial \theta_2} &= \frac{1}{2} \sum_{k=1}^m \frac{k}{\text{Tr}(\rho F_{i_k})} [F_{i_k}(1, 2) + F_{i_k}(2, 1)] \\ \frac{\partial \ln L(\theta)}{\partial \theta_3} &= \frac{1}{2} \sum_{k=1}^m \frac{1}{\text{Tr}(\rho F_{i_k})} [F_{i_k}(1, 1) + F_{i_k}(2, 2)] \end{aligned}$$

with $\left(\hat{I}(\tilde{\theta})\right)_{ij} = \frac{\partial \ln L(\tilde{\theta})}{\partial \theta_i} \frac{\partial \ln L(\tilde{\theta})}{\partial \theta_j}$. Also, for $N = 2$, the constraint matrix is simply a row vector, $R_{1j} = \frac{\partial a_2(\theta)}{\partial \theta_j}$.

The Fisher information decomposes similarly to $N = 2$ for $N = 3$ due to the linearity of the likelihood in θ .

V. CHOICE OF MEASUREMENT BASES

In this section we discuss the different measurement strategies used in the paper. Note that each measurement strategy yields a different asymptotic variance for the MLE, i.e., a different Fisher information matrix, as discussed in Section III C.

A. Mutually unbiased (average case optimal) measurements

For 1-qubit systems ($N = 2$), the MUB bases $V^{(r)}$, $0 \leq r \leq N$ can be written

$$V^{(0)} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad V^{(1)} = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}, \quad V^{(2)} = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ i & -i \end{pmatrix}. \quad (30)$$

The observables are then simply the standard Pauli spin operators.

For $N = 3$, or more generally when N is the power of an odd prime, these bases $V^{(r)}$ are given by

$$V_{pq}^{(r)} = \begin{cases} \delta_{pq}, & r = 0 \\ \frac{1}{\sqrt{N}} \exp\left[\frac{2\pi i}{N}(rp^2 + pq)\right], & 1 \leq r \leq N. \end{cases}$$

The orthonormal observables can then be taken to be

$$F_{r(N-1)+i} = V^{(r)} \tilde{F}_i (V^{(r)})^\dagger, \quad \tilde{F}_i = |i\rangle\langle i| = \text{diag}(0, \dots, 1, \dots, 0), \quad 1 \leq i \leq N-1. \quad (31)$$

Recall that the MUB measurements maximize the average Fisher information over the set of all true density matrices ρ_0 , and hence are ‘‘average-case optimal’’ as discussed in Section III C.

B. Complete, average case suboptimal measurements

In order to interrogate the asymptotic and finite sample losses induced by using biased measurement bases, MUB bases are rotated, causing the associated paralleliped (T_1, \dots, T_{N+1}) in Section IIID.2 to no longer be a rectangular solid and the average Fisher information to decrease. After rotation, the new bases can be written

$$\tilde{V}^{(r)} = U(s) V^{(r)} U^\dagger(s) \quad (32)$$

where $U(s) = e^{iA^{(r)}s}$, A being a random Hermitian matrix specifying a random axis of rotation in the N -dimensional Hilbert space; s is a scalar parameter specifying the extent

of rotation (magnitude of the solid angle). To generate sets of measurement bases that differ systematically in asymptotic efficiency from the MUBs, s is incrementally increased until $|\langle \mathbf{v}_i^{(r)}, \mathbf{w}_j^{(r)} \rangle|$ in equation (16) for at least one r, i, j exceeded a fixed multiple of $\frac{1}{\sqrt{N}}$.

VI. NUMERICAL IMPLEMENTATION

In the Bloch vector parametrization, the constrained maximum of the likelihood (6) corresponding to Lagrangian function (8) can be found by solving for roots of the nonlinear system of $N^2 + 2N - 3$ equations $\frac{\partial \mathcal{L}}{\partial \mathbf{t}}$ in $N^2 + 2N - 3$ unknowns $\theta_i, \gamma_j, \lambda_k$. The number of constraints and hence unknowns will differ in other parametrizations; for example, for parameterizations where positive semidefiniteness is implicit in the parametrization (such as the Cholesky parametrization), $\lambda_j = 0, j = 1, \dots, N - 1$ in equation (8). The Newton-Raphson algorithm can be used to find the roots of this nonlinear system. Writing $\frac{\partial \mathcal{L}}{\partial \mathbf{t}} = \mathbf{H}(\mathbf{t})$, the Newton step for

$$\mathbf{H}(\mathbf{t}) = 0$$

is

$$\mathbf{t}_{\text{new}} = \mathbf{t}_{\text{old}} + \delta \mathbf{t},$$

with $\delta \mathbf{t} = -\mathbf{J}^{-1}\mathbf{H}$, where $J_{ij} = \frac{\partial H_i}{\partial t_j}$ is the Jacobian matrix. Denoting the rows of \mathbf{H} by H_i , we have:

$$\begin{aligned} H_i(\theta) &= \frac{\partial \mathcal{L}(\theta, \lambda, \gamma|x)}{\partial \theta_i} = \frac{\partial \ln L(\theta|x)}{\partial \theta_i} = 0, & 1 \leq i \leq N^2 - 1, \\ H_{N^2+j-1}(\theta) &= \frac{\partial \mathcal{L}(\theta, \lambda, \gamma|x)}{\partial \lambda_j} = a_j(\theta) = 0, & 1 < j \leq N - 1, \\ H_{N^2+N+j-2}(\lambda, \gamma) &= \frac{\partial \mathcal{L}(\theta, \lambda, \gamma|x)}{\partial \gamma_j} = 2\lambda_j\gamma_j = 0 & 1 < j \leq N - 1. \end{aligned}$$

In order to facilitate global convergence of the Newton-Raphson algorithm, the “sum-of-squares” function $h = \mathbf{H} \cdot \mathbf{H}$ is evaluated after each iteration, and the step length progressively shortened until the value of this function is found to decrease (the existence of such a step length is guaranteed) [15].

Alternatively, the “sum-of-squares” function $h(\mathbf{t})$ may be minimized directly to locate the constrained maximum of the likelihood. In general, direct minimization of this function may be prone to encountering local traps. In the present case, minimization using an optimization algorithm capable of escaping from traps was employed. The quasi-Newton algorithm, in which the algorithmic step $k + 1$ is given by $\mathbf{t}^{(k+1)} - \mathbf{t}^{(k)} = -\mathbf{A}^{-1} \nabla h(\mathbf{t}^{(k)})$, where \mathbf{A}^{-1} denotes the approximate inverse Hessian computed with the Broyden-Fletcher-Goldfarb-Shanno (BFGS) update, was first used to search for a zero of $\nabla h(\mathbf{t})$ until convergence slowed “below a specified stepwise tolerance”, again using an adaptive line search strategy to identify the optimal step size. Traps were often encountered that could not be escaped from using the above technique. To surmount them, a stochastic (simulated annealing) step... When computing asymptotic covariances of the parameters at the global optimum according to equation (9), the perturbation $0.001 * I_{N^2-1}$ was added to the Fisher information in order to avoid numerical singularities in inversion.

In order to have a scalar measure of estimation accuracy of the entire density matrix, the Josza fidelity (generalized overlap) $\mathcal{F} = \text{Tr}^2\{\sqrt{\sqrt{\hat{\rho}}\rho\sqrt{\hat{\rho}}}\}$, which is related to the statistical distance on the space of density matrices, was used. The convergence tolerance (i.e., objective function value below which the optimized parameter estimate was accepted as an estimator) for each ρ was chosen by running a set of MLE optimizations with a very large sample size ($m = 10000$ observations), and determining the objective function value below which $\mathcal{F} \leq$ (a fixed value). In order to make the convergence tolerance compatible across difference sample sizes, the log of the likelihood function (6) was scaled as $\ln L(x | \rho(\theta)) \rightarrow \frac{1}{m} \ln L(x | \rho(\theta))$.

Kernel density estimators (KDEs) were used to estimate finite sample probability density functions. KDEs are nonparametric density estimators that avoid some of the deficiencies of histograms. Unlike histograms, they are smooth. KDEs center a kernel function at each data point; the contribution of data point $x(i)$ to the estimate at x^* depends on $x^* - x(i)$. The estimated density takes the form

$$\hat{f}(x) = \frac{1}{n} \sum_{i=1}^n K\left(\frac{x - x(i)}{h}\right)$$

with $\int K(t)dt = 1$. Bandwidth (h) optimization, which avoids values of h that lead to spiky estimates (undersmoothing) or oversmoothing, was based on minimization of the asymptotic mean integrated squared error (AMISE) [18], i.e., $h_{opt} = \arg \min \text{AMISE}$. A Gaussian kernel

$$\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}u^2\right),$$

was used.

For simulating quantum observations from a given basis $V^{(r)}$, the multinomial distribution probabilities (p_1, \dots, p_N) were computed from $p_i = \text{Tr}(\rho(\theta)F_i)$ for each F_i belonging to basis r , and draws were sampled from this multinomial distribution. If multinomial outcome i was associated with draw k , observable F_i was assigned as simulated quantum observation k .

Codes implementing the above algorithms, including both optimization and hypothesis testing, are available upon request from the author. Future updates will include additional parameterizations including the Euler and Cholesky parameterizations.

VII. ESTIMATION AND HYPOTHESIS TESTING RESULTS

For both spin-1/2 and spin-1 systems, MLE estimations were carried out for 3 different ρ 's: two random and one pure state. For each ρ , samples of size m were drawn in simulations of i.i.d quantum observations according to the method described in Section VI, and the parameters of ρ were estimated by MLE optimization for each sample. Of the types of measurement strategies considered in Section V, mutually unbiased measurement bases (MUB) are the most suitable for minimizing (asymptotic) estimation errors across the widest range of Hilbert space dimensions and ρ 's. MUB's were therefore used for the majority of estimations, with the performance of non-MUB bases compared later.

According to Lemmas 1 and 4 (see Appendix), asymptotically for large m , the density matrix parameter estimates should be distributed with covariance matrix given by equation (9). A primary goal was to determine the rate at which asymptotically predicted

standard errors corresponding to the CRB are attained. In order for this to be possible, it is necessary to obtain an accurate estimate of the variance of the estimator for each sample size m . For each m , the distribution of parameter estimates approaches a standard normal in the limit of an infinite number of samples of this size that are estimated. To assess the normality of these distributions of $\sqrt{m}(\tilde{\theta}_i - \theta_{i,0})$, the MSE, bias, and skewness of the distributions were compared for successively larger numbers of simulations. The results are presented in Table I and Figure 1.

Evidently, the normality of the distributions for each m does not substantially improve beyond 1000 simulations. Thus, the corresponding variance was taken as an estimate of the finite sample variance of $\sqrt{m}(\tilde{\theta}_i - \theta_{i,0})$ for each sample size. These distributions were then plotted for successively increasing sample sizes $m = 100, 400, 1000$, for selected parameters (Figs. 2 and 3). Asymptotic standard errors are superimposed for each parameter. These were estimated using $\hat{I}(\tilde{\theta})$ to approximate the asymptotic covariance matrix of the constrained parameter estimates according to (9), based on a single sample of each size chosen randomly from the set of 1000 simulations. The distributions of $\tilde{\theta}_i - \theta_{i,0}$ are also plotted for comparison (Figs. 4 and 5).

If the asymptotic distributions are to provide a good approximation in finite samples, the finite sample distributions $\sqrt{m}(\tilde{\theta}_i - \theta_{i,0})$ must approach the asymptotic distributions for large sample sizes m . Clearly, this is not the case for experimentally realistic sample sizes; the finite sample distribution of $\sqrt{m}(\tilde{\theta}_i - \theta_{i,0})$ actually diverges further from the asymptotic distribution. On the other hand, $\tilde{\theta}_i - \theta_{i,0}$ does narrow with m (Fig. 3). This indicates that for common experimental sample sizes, the asymptotic predictions of frequentist MLE are not practically useful.

Figures 6 and 7 depict the corresponding distributions for selected density matrix elements $\rho_{ij}(\theta)$. The asymptotic variances in the estimates of the elements of ρ can be calculated given the covariance matrix Σ according to the expression:

$$\text{var } f(\theta) = \left(\frac{\partial f(\theta)}{\partial \theta_k} \right)^T \Sigma \frac{f(\theta)}{\partial \theta_k} \quad (33)$$

where $f(\theta) = \rho_{ij}(\theta)$.

FIG. 1: Finite sample distributions of parameter estimates, spin-1/2 systems, MUB bases, for various numbers of simulations. All ρ_1 , θ_1 A) Sample size 20; B) Sample size 40; C) Sample size 60. In each panel, the finite sample distributions for 500, 1000, and 2000 simulations are shown.

FIG. 2: Finite sample distributions of $\sqrt{m}(\tilde{\theta}_i - \mu_{\tilde{\theta}_i})$, spin-1/2 systems, MUB bases. All ρ_1 . Panels A-C, θ_1 . A) θ_1 , sample size 100; B) θ_1 , sample size 400; C) θ_1 , sample size 1000. Panels D-F, θ_2 . D) θ_2 , sample size 100; E) θ_2 , sample size 400; F) θ_3 , sample size 1000. In each panel, the finite sample distributions (1000 simulations) are shown alongside the corresponding asymptotic distributions ($\mathcal{N}(\tilde{\theta}_i, \hat{\Sigma}_{ii})$ with $\hat{\Sigma}$ computed according to equation (10)). Asymptotic variances are estimated from one of the 1000 repeated samples (chosen at random).

Tables 1 and 2 present the bias, finite sample standard error, mean absolute error, and asymptotic 95% confidence interval for each parameter and the diagonal elements of a random spin-1/2 mixed state (ρ_1) and a pure spin-1/2 state (ρ_3). Tables 3 and 4 present the corresponding statistics for a random spin-1 mixed state and a pure spin-1 state. The asymptotic 95% confidence interval for $\tilde{\theta}_i$ is $[\theta_{i,L}, \theta_{i,R}]$ such that $\int_{\theta_{i,L}}^{\theta_{i,R}} \mathcal{N}(\tilde{\theta}_i, v(\tilde{\theta})) = 0.95$. The mean standard error (MSE) is also reported in each case.

FIG. 3: Finite sample distributions of $\sqrt{m}(\tilde{\theta}_i - \mu_{\tilde{\theta}_i})$, spin-1 systems, MUB bases. All ρ_1 . Panels A-C, θ_1 . A) θ_1 , sample size 100; B) θ_1 , sample size 400; C) θ_1 , sample size 1000. Panels D-F, θ_2 . D) θ_2 , sample size 100; E) θ_2 , sample size 400; F) θ_2 , sample size 1000. In each panel, the finite sample distributions (1000 simulations) are shown alongside the corresponding asymptotic distributions ($\mathcal{N}(\mu_{\tilde{\theta}_i}, \hat{\Sigma}_{ii})$ with $\hat{\Sigma}$ computed according to equation (9)). Asymptotic variances are estimated from one of the 1000 repeated samples (chosen at random).

FIG. 4: Finite sample distributions of parameter estimates, spin-1/2 systems, MUB bases. All ρ_1 . Panels A-C, θ_1 . A) θ_1 , sample size 100; B) θ_1 , sample size 400; C) θ_1 , sample size 1000. Panels D-F, θ_2 . D) θ_2 , sample size 100; E) θ_2 , sample size 400; F) θ_3 , sample size 1000. The finite sample distributions were computed from 1000 simulations.

FIG. 5: Finite sample distributions of parameter estimates, spin-1 systems, MUB bases. All ρ_1 . Panels A-C, θ_1 . A) θ_1 , sample size 100; B) θ_1 , sample size 400; C) θ_1 , sample size 1000. Panels D-F, θ_2 . D) θ_2 , sample size 100; E) θ_2 , sample size 400; F) θ_2 , sample size 1000. The finite sample distributions were computed from 1000 simulations.

A. Effect of measurement bases

The above results were all obtained for sample draws/estimations employing mutually unbiased measurement bases. In order to interrogate the effect of the choice of measurement bases on the asymptotic efficiencies and finite sample efficiencies, the three alternative measurement scenarios described in Section V were compared. A primary goal was to compare the asymptotic and finite sample relative efficiencies across these measurement scenarios/strategies.

In order to properly compare the asymptotic variances corresponding to different measurement strategies, the expected Fisher information was approximated in each case by

FIG. 6: Finite sample distributions of ρ matrix elements, spin-1/2 systems, MUB bases. All ρ_1 . Panels A-C, ρ_{11} . A) ρ_{11} , sample size 100; B) ρ_{11} , sample size 400; C) ρ_{11} , sample size 1000. Panels D-F, ρ_{22} . D) ρ_{22} , sample size 100; E) ρ_{22} , sample size 400; F) ρ_{11} , sample size 1000. In each panel, finite sample distributions (1000 simulations) are shown alongside the corresponding asymptotic distribution. $\sqrt{m}(\text{Re}, \text{Im}(\tilde{\rho}_{ij}) - \text{Re}, \text{Im}(\rho_{ij}^0))$, which asymptotically has a nondegenerate distribution, is plotted in each case.

FIG. 7: Finite sample distributions of ρ matrix elements, spin-1/2 systems, MUB bases. All ρ_1 . Panels A-C, ρ_{11} . A) ρ_{11} , sample size 100; B) ρ_{11} , sample size 400; C) ρ_{11} , sample size 1000. Panels D-F, ρ_{22} . D) ρ_{22} , sample size 100; E) ρ_{22} , sample size 400; F) ρ_{22} , sample size 1000. In each panel, finite sample distributions (1000 simulations) are shown alongside the corresponding asymptotic distribution. $\sqrt{m}(\text{Re}, \text{Im}(\tilde{\rho}_{ij}) - \text{Re}, \text{Im}(\rho_{ij}^0))$, which asymptotically has a nondegenerate distribution, is plotted in each case.

the observed Fisher information for a large sample size ($m = 10000$), which makes expression (10) a good approximation of expression (9).

1. Complete non-optimal

Given the difficulty of implementing MUB or QCR-optimal measurements in the laboratory, it is essential to determine whether nonoptimal measurement bases achieve similar estimation accuracies in finite samples. In order for a density matrix to be *identifiable* by frequentist inference, measurements in at least $N + 1$ bases are required. (Most laboratory setups require/use $> N + 1$ bases and are (redundant) not optimal from either the standpoint of QFI or average-case optimal FI.) In order to compare nonoptimal and average-case optimal measurement strategies on an equal footing, two approaches were used. In both cases, the $|\langle \mathbf{u}_i, \mathbf{w}_j \rangle|$'s were recorded. In the first, the bases were randomly generated orthogonal sets of N unit vectors, while in the second, MUB bases were randomly perturbed according to the method described in Section V, i.e., equation (32). (Figs. 8, 10)

The asymptotic relative efficiencies of these bases are:

2. QCR optimal

(Fig. 9)

FIG. 8: Finite sample distributions of $\sqrt{m}(\tilde{\theta}_i - \mu_{\tilde{\theta}_i})$, spin-1/2 systems: comparison of MUB and non-MUB bases. All ρ_1, θ_1 . A) Sample size 100; B) Sample size 400; C) Sample size 1000. In each panel, the finite sample distributions for MUB and non-MUB bases are shown alongside the respective asymptotic distributions. Asymptotic variances were obtained from an estimation employing a sample size $m = 10000$ in order to obtain accurate approximations of the expected Fisher informations for MUB and non-MUB bases.

FIG. 9: Comparison of $\sqrt{m}(\tilde{\theta}_i - \mu_{\tilde{\theta}_i})$ distributions (sample size 1000) of parameter estimates, spin-1/2 systems, in optimal (red), average case optimal (MUB, blue), and random (green) measurement bases with asymptotic predictions. All pure state $\rho_3, \eta =$. A) Finite sample distributions; B) Asymptotic distributions. The asymptotic confidence intervals for the optimal measurement basis corresponds approximately to the quantum Cramer-Rao bound associated with the quantum Fisher information (purple). Asymptotic variances were obtained from an estimation employing a sample size $m = 10000$ in order to obtain accurate approximations of the expected Fisher informations.

The asymptotic relative efficiencies of the QCR optimal and MUB bases are:

FIG. 10: Finite sample distributions of parameter estimates, spin-1 systems: comparison of MUB and non-MUB bases. All ρ_1, θ_1 . A) Sample size 100; B) Sample size 400; C) Sample size 1000. In each panel, the finite sample distributions for MUB and non-MUB bases are shown alongside the respective asymptotic distributions. $\sqrt{m}(\tilde{\theta}_i - \theta_{i,0})$, which asymptotically has a nondegenerate distribution, is plotted in each case. Asymptotic variances were obtained from an estimation employing a sample size $m = 10000$ in order to obtain accurate approximations of the expected Fisher informations for MUB and non-MUB bases.

FIG. 11: Q-Q plots of finite sample estimate quantiles of $\sqrt{m}(\tilde{\rho}_{ij} - \mu_{\tilde{\rho}_{ij}})$ versus asymptotic quantiles: spin-1/2 systems. $\rho =; \rho_{11}$ A) Sample size 100; B) Sample size 400; C) Sample size 1000.

B. Comparison of finite sample performance and asymptotic efficiency

In order to more rigorously assess the correspondence/similarity between the finite sample and asymptotic MLE distributions, quantile-quantile or “Q-Q” plots may be used. A quantile-quantile plot is a 2-d plot consisting of points corresponding to successive quantiles of two datasets, or one dataset and a parametric distribution. The closer the points lie to a line of unit slope, the more similar are the two distributions. Figs. 11 and 12 display Q-Q plots for finite sample estimate distributions for selected elements of ρ , determined using MUB and non-MUB bases. For (an order of magnitude) larger sample sizes, the quantiles of the distributions are closer to the asymptotic ones, but only marginally so.

FIG. 12: Q-Q plots of finite sample estimate quantiles of $\sqrt{m}(\tilde{\rho}_{ij} - \mu_{\tilde{\rho}_{ij}})$ versus asymptotic quantiles: spin-1/2 systems, non-MUB. $\rho =; \rho_{11}$. A) Sample size 100; B) Sample size 400; C) Sample size 1000.

C. Hypothesis Testing

An important application of quantum state estimation is optimal decision making and control, and by extension, quantum information processing. Control logic is often Boolean, in that decisions are made based on whether a hypothesis is true or false, rather than the precise state of the system. An important conjecture to test regarding the density matrix, which we examine here, is whether it is a pure or mixed state, or more generally that an eigenvalue of ρ has a given value. The appropriate hypothesis test for (purity) is the T test. As an example, we test here the hypothesis H_0 that an eigenvalue of ρ equals its true (known) value. When constructing asymptotic standard errors for a bounded function of the parameters, it is necessary to take a transformation of that function whose range is unbounded on the real line.

The T test statistic for the null hypothesis that the estimated value of a parameter is equal to the true value is

$$T = \frac{1}{\sigma(\tilde{\theta}_i)}(\tilde{\theta}_i - \theta_{i,0}), \quad (34)$$

where $\sigma(\tilde{\theta}_i)$ is the asymptotic standard error of the estimate. The T test is conventionally defined with a rejection region R corresponding to the 5% of the cumulative probability density region of the tails of the test statistic's asymptotic distribution. Asymptotically, the T statistic distribution converges in the limit to the standard normal distribution; thus, the hypothesis is rejected if $T < -1.96$ or $T > 1.96$. When ρ is a mixed state, a T test of H_0 is generally a two-sided test, whereas when ρ is a pure state, the test is one-sided, since the eigenvalues fall on the boundary of the admissible interval.

We denote the i -th eigenvalue of the estimated density matrix $\tilde{\rho}(\theta)$ by $\delta_i(\theta)$. The Hellmann-Feynman theorem can be used to compute $\frac{\partial f(\theta)}{\partial \theta_k}$ where $f(\theta) = \delta_i(\theta)$:

$$\begin{aligned} \frac{\partial \delta_i}{\partial \theta} &= \frac{\partial \mathbf{v}_i^\dagger}{\partial \theta} \tilde{\rho} \mathbf{v}_i + \mathbf{v}_i^\dagger \tilde{\rho} \frac{\partial \mathbf{v}_i}{\partial \theta} + \mathbf{v}_i^\dagger \frac{\partial \tilde{\rho}}{\partial \theta} \mathbf{v}_i \\ &= \delta_i \frac{\partial \mathbf{v}_i^\dagger}{\partial \theta} \mathbf{v}_i + \delta_i \mathbf{v}_i^\dagger \frac{\partial \mathbf{v}_i}{\partial \theta} + \mathbf{v}_i^\dagger \frac{\partial \tilde{\rho}}{\partial \theta} \mathbf{v}_i \\ &= \mathbf{v}_i^\dagger \frac{\partial \tilde{\rho}}{\partial \theta} \mathbf{v}_i, \end{aligned}$$

since $\langle \mathbf{v}_i | \mathbf{v}_i \rangle = 1$, $\frac{\partial}{\partial \theta} \langle \mathbf{v}_i | \mathbf{v}_i \rangle = \frac{\partial}{\partial \theta} (1) = 0$. However, when computing asymptotic covariances of the eigenvalues of the density matrix, it is important to note that they are constrained to lie in the range $0 \leq \delta_{i,0} \leq 1$. Thus, the asymptotic distributions of the estimates $\tilde{\delta}_i$ cannot be normally distributed in correspondence with the CRB. A typical workaround in such cases is to compute finite sample and asymptotic standard errors of some function of the quantity of interest that can take values over the entire real line. In the present case, a suitable choice is the function $f(\delta_i) = \tanh^{-1}(2(\delta_i - 1))$, since the inverse hyperbolic tangent has range $-\infty \leq \tanh^{-1}(x) \leq \infty$, with $\tanh^{-1}(-1) = -\infty$ and $\tanh^{-1}(1) = \infty$.

It is difficult to analytically compute asymptotic standard errors for functions of the eigenvalues by using the Hellmann-Feynman theorem. For spin-1/2 systems, however, the exact analytical solution for the eigenvalue derivatives is available directly since the characteristic polynomial is a quadratic:

$$\begin{aligned}\delta_1 &= r_{22} + r_{11} - \frac{1}{2} \sqrt{r_{22}^2 - 2r_{11}r_{22} + 4r_{12}^2 + r_{11}^2 + i_{12}^2} \\ \delta_2 &= r_{22} + r_{11} + \frac{1}{2} \sqrt{r_{22}^2 - 2r_{11}r_{22} + 4r_{12}^2 + r_{11}^2 + i_{12}^2}\end{aligned}$$

where $r_{ij} \equiv \text{Re}(\rho_{ij})$, $i_{ij} \equiv \text{Im}(\rho_{ij})$. Since we are interested in the asymptotic distribution of the function $\tanh^{-1}(2(\tilde{\delta}_i - 1))$, we will need the derivatives

$$\begin{aligned}\frac{\partial \tanh^{-1}(2(\delta_i - 1))}{\partial r_{11}} &= \left(\frac{1}{2\delta_i - 1} + \frac{1}{3 - 2\tilde{\delta}_i} \right) \frac{\delta_1}{r_{11}} \\ &= \left(\frac{1}{2\delta_i - 1} + \frac{1}{3 - 2\tilde{\delta}_i} \right) \left[-\frac{1}{2} f^{-\frac{1}{2}}(-2r_{22} + 2r_{11}) + 1 \right].\end{aligned}$$

where $f \equiv r_{22}^2 - 2r_{11}r_{22} + 4r_{12}^2 + r_{11}^2 + i_{12}^2$.

In certain applications, such as quantum information processing, it is important to simultaneously test whether several elements or parameters of the density matrix have prescribed values. The Wald test is ideal for this purpose. As an example, we test the hypothesis that each of the parameters θ equal their true known values.

The Wald statistic for the null hypothesis that the estimated values of a set of parameters, say $(\tilde{\theta}_1, \dots, \tilde{\theta}_{N^2-1})$, are equal to their true values $(\theta_1, \dots, \theta_{N^2-1})$, is given by

$$W = \mathbf{v}^T \Sigma^{-1} \mathbf{v},$$

where $\mathbf{v} = (\tilde{\theta}_1 - \theta_{1,0}, \dots, \tilde{\theta}_{N^2-1} - \theta_{N^2-1,0})$ and Σ is the asymptotic covariance matrix of the parameter estimates. Asymptotically the Wald statistic assumes a chi squared distribution χ_k^2 , with the number of degrees of freedom k equal to the number of parameters (here $N^2 - 1$). The Wald test is conventionally defined with a rejection region R corresponding to the tail of the χ_k^2 distribution beyond which the cumulative probability density is 0.05.

The size and power (as well as test stat distributions) of the T and Wald tests provide another means of interrogating the finite sample vs. asymptotic behavior/efficiency of quantum state MLE. The finite sample size $s(T)$ of the T test is given by the fraction of times T is greater than 1.96 or less than -1.96, $(f_{T>1.96} + f_{T<-1.96})$; as conventionally defined, the T test has an asymptotic size of .05.

MAY DO T TEST FOR ONLY MIXED STATE;

The finite sample power $p(T)$ is evaluated using the test statistic

$$T' = \frac{1}{\sigma(\tilde{\theta}_i)}(\tilde{\theta}_i - \theta_i^n),$$

where $\theta_i^n \neq \theta_0$; the power is equal to the fraction of times T' is greater than 1.96 or less than -1.96. Asymptotically, the power of the T test is 1. It can be shown [5] that the T test based on MLE estimates is equivalent to a likelihood ratio test, and hence as described in Section II, it is UMP. Thus no other frequentist test could do better at testing whether the estimated value of a single parameter is the true value. The Wald test as conventionally defined asymptotically has a size of 1. Defining $W' = \mathbf{v}'^T \Sigma \mathbf{v}'$ with $\mathbf{v}' = (\tilde{\theta}_1 - \theta_1^n, \dots, \tilde{\theta}_{N^2-1} - \theta_{N^2-1}^n)$, if we denote by c the critical value under which 95% of the χ_k^2 distribution lies, $f_{W'<c}$ gives the finite sample power of the test. Asymptotically, the power of the Wald test is also 1. Like the T test, the Wald test based on MLE estimates is equivalent to a likelihood ratio test, and hence is UMP.

For a selected spin-1/2 system, Fig. 13A displays the finite sample T statistic distribution for the null hypothesis $\tanh^{-1}(2(\tilde{\delta}_i - 1)) = \tanh^{-1}(2(\delta_{i,0} - 1))$ (or $\tilde{\delta}_i = \delta_{i,0}$), whereas Fig. 13B displays the Wald statistic distribution for the null hypothesis that $\tilde{\theta}_i = \theta_{i,0}$, $i = 1, \dots, N^2 - 1$. Figs. 14A and 14B depict these distributions for a selected

FIG. 13: Finite sample test statistic distributions, spin-1/2 quantum systems. $\rho =$ (specify from text) A) T statistic distribution for null hypothesis $\tanh^{-1} \tilde{\delta}_1 = \tanh^{-1} \delta_{1,0}$, where $\delta_{1,0}$ denotes the first eigenvalue of the true density matrix ρ_0 . The asymptotic T stat distribution should be standard normal (superimposed). The inverse hyperbolic tangent of δ_1 was used because its range is the real line. B) Wald test? Asymptotic Wald stat distribution should be chi squared with degrees of freedom equal to number of restrictions (superimposed)

FIG. 14: Finite sample test statistic distributions, spin-1 quantum systems. ρ (specify from text) A) T statistic distribution for null hypothesis $\tanh^{-1} \tilde{\delta}_1 = \tanh^{-1} \delta_{1,0}$. Asymptotic T stat distribution should be standard normal (superimposed). The inverse hyperbolic tangent of δ_1 was used because its range is the real line. B) Wald test?

spin-1 ρ . (may eliminate) Tables 7 and 8 report the finite sample size and power of the T test for the null hypothesis that the first eigenvalue of $\tilde{\rho}$ is equal to its true value, for a spin-1/2 and spin-1 mixed state, respectively. (May choose eval number so it is near 0.5). For the same systems, the finite sample size and power of the Wald test for the null hypothesis that all the parameters $\tilde{\theta}_i$ are equal to their true values $\theta_{0,1}$ are also displayed in Tables 7 and 8. The salient point is that the size and power of these tests fall substantially far from the asymptotically predicted optimal values, and that the finite sample test statistic distributions are not normal and chi-squared (clear from inspection), as would be expected from asymptotic theory.

VIII. DISCUSSION AND EXTENSIONS

In this paper we have examined the performance of frequentist estimators of the density matrix of a quantum system using quantum observations simulated under appropriate parameterizations of the density matrix. We provided robust numerical techniques for

likelihood optimization under multiple constraints that are robust across arbitrary spin-1/2 and spin-1 system density matrices. In addition, we have presented methodologies and prescriptions for hypothesis testing in the frequentist quantum framework, an essential requirement for optimal decision making and control.

Performing inference in the frequentist framework, we find that the finite sample variances are significantly larger than the asymptotic bounds (predicted by the Cramer-Rao theorem) for typical experimental sample sizes, and that these bounds are not approached at the rate predicted by asymptotic theory for large m . A prior study showed close correspondence between asymptotic and finite sample performance in a single example, but did not consider the rate of convergence to the CRB, or higher-dimensional systems. This finding is typically robust to the choice of the density matrix and is more pronounced for small datasets. Recent work [14] has aimed to assess the resource requirements of various quantum tomography implementations, but these assessments are only valid in the asymptotic limit of an infinite number of measurements. A proper/complete resource analysis must take into account finite sample losses and the rate at which asymptotic predictions are approached.

Given the dependence of the optimal measurement strategy/Fisher information on the state, adaptive measurement schemes have been proposed as alternatives to MUB. However, this method cannot be used to achieve the QCR bound for more than one unknown parameter. Moreover, even average case optimal (MUB) measurements are difficult to apply for systems with Hilbert space dimensions $N > 2$. Given that these asymptotic (quantum) Fisher informations are unachievable and given that the relative finite sample variances fall short of the asymptotic relative efficiencies, we conclude that in order for this information to improve parameter estimates, it must be incorporated into the estimation procedure in a manner independent of the specific measurement strategy used; i.e., it is important to find other means of exploiting the information geometry of quantum states [24]

An ideal approach in this regard is Bayesian estimation. Bayesian estimation, which is based on updating a prior plausibility distribution about the parameters based on observed

data, is considerably more general than standard frequentist methods, lending itself readily to cases with incomplete observation levels and a finite number of measurements. The prior plausibility distribution permits the introduction of auxiliary information about the parameter space that is not contained within the likelihood. By contrast, such information is impossible to incorporate in frequentist estimation techniques, which as we have seen have no rigorous asymptotic finite sample variances. Moreover, it is well-known that the credible intervals produced by Bayesian inference are generally more reliable than asymptotic frequentist confidence intervals.

The differences between the density matrix estimates, and especially the associated statistical uncertainties, obtained through Bayesian and MLE methods have very important implications for the representation of the true degree of knowledge regarding a quantum system contained in a finite number of measurements. However, in quantum mechanics, the subjective quality of plausibility distributions in Bayesian statistics has aroused some criticism of its place within the foundations of the subject. As discussed above, in the Bayesian approach, a flat prior distribution will produce the same parameter estimates as MLE. However, note that even though the prior is flat or noninformative with respect to the parameters of the density matrix, it would not be flat with respect to a transformation of the same. The subjectivity of the prior distribution is one of the main differences between the classical and Bayesian schools of statistics. The formal postulates of quantum mechanics (QM) assume that the state of every quantum system is described by a density matrix, independent of observations on that system. As such, they do not explicitly provide for the existence of a plausibility distribution, based on physical prior knowledge regarding the system, over the set of possible density matrices. As the MLE approach to state estimation does not make explicit use of such plausibility distributions, it appears that it does not require additional assumptions beyond those of the postulates of QM. However, as mentioned, frequentist QM does implicitly assume a generic prior plausibility distribution that does not follow from any of the standard postulates of QM. By contrast, so-called invariant priors exist, which follow naturally from the Riemannian metric on the density matrix parameter space and which may be able to

effectively exploit the geometric information implicit in the quantum Fisher information without relying on the often unachievable QCRB. As these priors follow purely from considerations of symmetry, Bayesian QM, in fact, requires the adoption of smallest number of additional assumptions beyond the standard postulates of QM among all methods of statistical estimation. Owing to their theoretical underpinnings, more precise and reliable results are likely to be obtained for these prior specifications. Also, this analysis would provide a framework for gauging the loss from relying on subjective prior choices. In a companion paper, we are addressing this class of issues.

APPENDIX A: GENERATORS FOR $SU(3)$

$$\lambda_1 = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad \lambda_2 = \begin{pmatrix} 0 & -i & 0 \\ i & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad \lambda_3 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

$$\lambda_4 = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}, \quad \lambda_5 = \begin{pmatrix} 0 & 0 & -i \\ 0 & 0 & 0 \\ i & 0 & 0 \end{pmatrix}, \quad \lambda_6 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}$$

$$\lambda_7 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & -i \\ 0 & i & 0 \end{pmatrix}, \quad \lambda_8 = \frac{1}{\sqrt{3}} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -2 \end{pmatrix}.$$

APPENDIX B: CLASSICAL CRAMER-RAO BOUND AND UNCONSTRAINED MLE VARIANCE-COVARIANCE

Lemma 2 *When there are no constraints on the parameters, the eigenvalues of the covariance matrix of the estimates from any consistent frequentist estimator are bounded from below by the eigenvalues of $(mI(\theta_0))^{-1} = \left\{ mE \left[\frac{\partial \ln L(\theta_0|x)}{\partial \theta} \left(\frac{\partial \ln L(\theta_0|x)}{\partial \theta} \right)^T \right] \right\}^{-1}$.*

Proof. Let $\mathbf{w} = \hat{\theta} - \theta_0$, $\mathbf{v} = \frac{\partial \ln L(\hat{\theta}|x)}{\partial \theta} - \frac{\partial \ln L(\theta_0|x)}{\partial \theta}$. The covariance matrix of parameter estimates may be written $E(\mathbf{w}\mathbf{w}^T)$, and the Fisher information $E(\mathbf{v}\mathbf{v}^T)$. According to the Cauchy-Schwarz (C-S) inequality, $|\langle \alpha, \beta \rangle|^2 \leq |\alpha|^2 |\beta|^2$, where α, β denote any two (possibly infinite-dimensional) vectors. Here, we use the expectation value inner product $E(f(\hat{\theta})) = \int_{\mathcal{X}} L(\theta_0|x) f(\hat{\theta}) dx$. Let $f = \mathbf{w}^T \mathbf{b}$, $g = \mathbf{v}^T \mathbf{a}$ for any two vectors \mathbf{a}, \mathbf{b} (assumed unit vectors without loss of generality). Then, the C-S inequality implies

$$E[\mathbf{a}^T \mathbf{w}\mathbf{w}^T \mathbf{b}]^2 \leq E[\mathbf{a}^T \mathbf{w}\mathbf{w}^T \mathbf{a}] E[\mathbf{b}^T \mathbf{v}\mathbf{v}^T \mathbf{b}]. \quad (\text{B1})$$

The left side of this inequality can be simplified by noting that the elements of the matrix $E(\mathbf{w}\mathbf{w}^T)$ are

$$E \left[(\hat{\theta} - \theta_0) \left(\frac{\partial \ln L(\hat{\theta}|x)}{\partial \theta} - \frac{\partial \ln L(\theta_0|x)}{\partial \theta} \right)^T \right]_{ij} = \delta_{ij},$$

since we have a) $E(\hat{\theta}_i \frac{\partial}{\partial \theta_i} \ln L(\theta_0|x)) = \frac{\partial}{\partial \theta_i} \int_{\mathcal{X}} \hat{\theta}_i L(\theta_0|x) dx = \frac{\partial}{\partial \theta_i} E(\hat{\theta}_i) = 1$, and b) $E(\hat{\theta}_i \frac{\partial}{\partial \theta_j} \ln L(\theta_0|x)) = \frac{\partial}{\partial \theta_j} \int_{\mathcal{X}} \hat{\theta}_i L(\theta_0|x) dx = \frac{\partial}{\partial \theta_j} E(\hat{\theta}_i) = 0$. Moreover, we have $E(\mathbf{v}\mathbf{v}^T) = mI(\theta_0)$, since $E(\frac{\partial}{\partial \theta} \ln L(\theta_0|x)) = \frac{\partial}{\partial \theta} \int_{\mathcal{X}} \ln L(\theta_0|x) dx = 0$. Thus (B1) becomes

$$(\mathbf{a}^T I \mathbf{b})^2 \leq \mathbf{a}^T \Sigma \mathbf{a} \mathbf{b}^T mI(\theta_0) \mathbf{b}.$$

Now letting $\mathbf{b} = \frac{1}{m} I^{-1}(\theta_0) \mathbf{a}$, since \mathbf{b} can be any vector, we obtain

$$\frac{1}{m^2} (\mathbf{a}^T I^{-1}(\theta_0) \mathbf{a})^2 \leq \frac{1}{m} \mathbf{a}^T \Sigma \mathbf{a} \mathbf{a}^T I^{-1}(\theta_0) \mathbf{a}$$

or

$$\mathbf{a}^T I^{-1}(\theta_0) \mathbf{a} \leq m \mathbf{a}^T \Sigma \mathbf{a},$$

so $\mathbf{a}^T(m\Sigma - I^{-1}(\theta_0))\mathbf{a} \geq 0$ for any \mathbf{a} . This implies $\Sigma - (mI)^{-1}$ is positive-semidefinite. ■

Lemma 3 (*MLE achieving CRB*)

The covariance matrix of MLE estimates asymptotically achieves the Cramer-Rao limiting covariance matrix.

Proof. For simplicity we adopt the shorthand $L(\theta) \equiv L(\theta|x)$. Expand the derivative of the log likelihood function around the true value of the parameter vector θ_0 to first order:

$$\frac{\partial \ln L(\hat{\theta})}{\partial \theta} = \frac{\partial \ln L(\theta_0)}{\partial \theta} + \frac{\partial^2 \ln L(\theta_0)}{\partial \theta \partial \theta'} (\hat{\theta} - \theta_0).$$

Regularity conditions justify the omission of terms above first order (second derivatives of $\ln L$). By definition, $\frac{\partial \ln L(\hat{\theta})}{\partial \theta} = 0$, so we have

$$\begin{aligned} \sqrt{m}(\hat{\theta} - \theta_0) &= -\sqrt{m} \left(\frac{\partial^2 \ln L(\theta_0)}{\partial \theta \partial \theta'} \right)^{-1} \frac{\partial \ln L(\theta_0)}{\partial \theta} \\ &= - \left(\frac{1}{m} \frac{\partial^2 \ln L(\theta_0)}{\partial \theta \partial \theta'} \right)^{-1} \frac{1}{\sqrt{m}} \frac{\partial \ln L(\theta_0)}{\partial \theta}. \end{aligned}$$

We are interested in the distribution of $\sqrt{m}(\hat{\theta} - \theta_0)$ in the limit $m \rightarrow \infty$. We proceed by determining a) the limiting distribution to which $-\frac{1}{\sqrt{m}} \frac{\partial \ln L(\theta_0)}{\partial \theta}$ converges, and b) the limiting mean of $-\left(\frac{1}{m} \frac{\partial^2 \ln L(\theta_0)}{\partial \theta \partial \theta'}\right)^{-1}$. For a), we need both the limiting mean and limiting variance of the quantity. As shown in the proof for Lemma 2, $-\frac{1}{\sqrt{m}} \mathbb{E} \left[\frac{\partial \ln L(\theta_0)}{\partial \theta} \right] = 0$, and the limiting variance-covariance matrix of $-\frac{1}{\sqrt{m}} \frac{\partial \ln L(\theta_0)}{\partial \theta}$ is $m \mathbb{E} \left[\frac{\partial \ln L(\theta_0)}{\partial \theta} \left(\frac{\partial \ln L(\theta_0)}{\partial \theta} \right)^T \right]$. Thus, by the Lindeberg central limit theorem, as $m \rightarrow \infty$, $\frac{\partial \ln L(\theta_0)}{\partial \theta}$ converges to $\mathcal{N}[0, I(\theta_0)]$ in distribution. For b), we have $-\mathbb{E} \left[\frac{1}{m} \frac{\partial^2 \ln L(\theta_0)}{\partial \theta \partial \theta'} \right] = -\frac{1}{m} \int_{\mathcal{X}} L^{-1}(\theta_0|x) \frac{\partial^2}{\partial \theta \partial \theta'} L(\theta_0|x) dx + \frac{1}{m} \int_{\mathcal{X}} L^{-2}(\theta_0|x) \frac{\partial}{\partial \theta} L(\theta_0|x) \left(\frac{\partial}{\partial \theta} L(\theta_0|x) \right)^T dx$. The first term is 0, so we have $\mathbb{E} \left[\frac{1}{m} \frac{\partial^2 \ln L(\theta_0)}{\partial \theta \partial \theta'} \right] = I(\theta_0)$; hence $\frac{1}{m} \frac{\partial^2 \ln L(\theta_0)}{\partial \theta \partial \theta'}$ converges to $I(\theta_0)$ in probability according to Khinchin's weak law of large numbers. Substituting, we get

$$\sqrt{m}(\hat{\theta} - \theta_0) \rightarrow I^{-1}(\theta_0) \mathcal{N}[0, I(\theta_0)] = \mathcal{N}[0, I^{-1}(\theta_0)]. \quad \blacksquare$$

APPENDIX C: DERIVATION OF THE CONSTRAINED ASYMPTOTIC CO-VARIANCE MATRIX

Lemma 4 *The asymptotic variance-covariance matrix Σ for maximum likelihood estimation of a n -component parameter vector θ subject to l nonlinear constraints $a_j(\theta) = 0$, $1 \leq j \leq l$, whose first order Taylor expansions about $\theta = 0$ are $R_{jk}\theta_k = c_j$, where R is a $l \times n$ matrix, is $\Sigma(\tilde{\theta}) = I^{-1}(\tilde{\theta}) - I^{-1}(\tilde{\theta})R^T \left(RI^{-1}(\tilde{\theta})R^T \right)^{-1} RI^{-1}(\tilde{\theta})$.*

Proof. Let us denote the parameter vector and Lagrange multiplier vector at the constrained optimum of the likelihood as $\tilde{\theta}$, $\tilde{\lambda}$, respectively. Taylor expanding the constraints around $\theta = 0$, such that $\frac{\partial}{\partial \theta_k}(a_j(\theta))\big|_{\theta=0}\theta - \gamma_j^2 := R_{jk}\theta_k - c_j$, we may rewrite equation (8) as

$$\mathcal{L}(\theta, \lambda) = \ln L(\theta) - \lambda^T(R\theta - \mathbf{c}),$$

where λ and \mathbf{c} are n -dimensional vectors and $L(\theta) = \prod_{k=1}^m \text{Tr}(\rho(\theta)F_{i_k})$ as in equation (6). The first-order conditions for this Lagrangian are:

$$\frac{\partial \mathcal{L}(\theta, \lambda)}{\partial \theta}\bigg|_{\tilde{\theta}, \tilde{\lambda}} = \frac{\partial \ln L(\tilde{\theta})}{\partial \theta} - R^T \tilde{\lambda} = 0 \tag{C1}$$

$$\frac{\partial \mathcal{L}(\theta, \lambda)}{\partial \lambda}\bigg|_{\tilde{\theta}, \tilde{\lambda}} = R\tilde{\theta} - \mathbf{c} = 0. \tag{C2}$$

where Taylor expanding $\frac{\partial \ln L(\tilde{\theta})}{\partial \theta}$ to first order around $\hat{\theta}$, the unrestricted parameter estimate, we have:

$$\frac{\partial \ln L(\tilde{\theta})}{\partial \theta} = \frac{\partial \ln L(\hat{\theta})}{\partial \theta} + \frac{\partial^2 \ln L(\hat{\theta})}{\partial \theta \partial \theta'}(\tilde{\theta} - \hat{\theta}).$$

Standard regularity condition on the likelihood function allow us to ignore higher order terms. Now based on the definition of $\hat{\theta}$, $\frac{\partial \ln L(\hat{\theta})}{\partial \theta} = 0$, such that

$$\frac{\partial \ln L(\tilde{\theta})}{\partial \theta} = \frac{\partial^2 \ln L(\hat{\theta})}{\partial \theta \partial \theta'}(\tilde{\theta} - \hat{\theta}), \tag{C3}$$

and according to (the first order condition) (C1), $\frac{\partial \ln L(\hat{\theta})}{\partial \theta} = R^T \tilde{\lambda} = 0$, allowing us to write

$$\begin{aligned} R(\tilde{\theta} - \hat{\theta}) &= R \left(\frac{\partial^2 \ln L(\hat{\theta})}{\partial \theta \partial \theta'} \right)^{-1} \frac{\partial \ln L(\tilde{\theta})}{\partial \theta} \\ &= R \left(\frac{\partial \ln L(\hat{\theta})}{\partial \theta \partial \theta'} \right)^{-1} R^T \tilde{\lambda}. \end{aligned}$$

We can now solve for $\tilde{\lambda}$ as follows:

$$\begin{aligned} \tilde{\lambda} &= \left[R \left(\frac{\partial^2 \ln L(\hat{\theta})}{\partial \theta \partial \theta'} \right)^{-1} R^T \right]^{-1} R(\tilde{\theta} - \hat{\theta}) \\ &= \left[R \left(\frac{\partial^2 \ln L(\hat{\theta})}{\partial \theta \partial \theta'} \right)^{-1} R^T \right]^{-1} (c - R\hat{\theta}). \end{aligned}$$

Substituting this expression in equation (B1), and using (C3), we get

$$\begin{aligned} \frac{\partial \ln L(\tilde{\theta})}{\partial \theta} &= R^T \left[R \left(\frac{\partial^2 \ln L(\hat{\theta})}{\partial \theta \partial \theta'} \right)^{-1} R^T \right]^{-1} (c - R\hat{\theta}) \\ \frac{\partial^2 \ln L(\hat{\theta})}{\partial \theta \partial \theta'} (\tilde{\theta} - \hat{\theta}) &= R^T \left[R^T \left(\frac{\partial^2 \ln L(\hat{\theta})}{\partial \theta \partial \theta'} \right)^{-1} R \right]^{-1} (c - R\hat{\theta}). \end{aligned}$$

Solving for $\tilde{\theta}$, we have

$$\tilde{\theta} = \hat{\theta} + \left(\frac{\partial^2 \ln L(\hat{\theta})}{\partial \theta \partial \theta'} \right)^{-1} R^T \left[R \left(\frac{\partial^2 \ln L(\hat{\theta})}{\partial \theta \partial \theta'} \right)^{-1} R^T \right]^{-1} (c - R\hat{\theta}).$$

This relation allows us to express the asymptotic variance-covariance matrix for the constrained parameter vector $\tilde{\theta}$, $\text{Var}(\tilde{\theta}) \equiv \Sigma$, in terms of the variance-covariance matrix for the unconstrained parameter vector $\hat{\theta}$. We first assign \sqrt{m} , m coefficients:

$$\sqrt{m}(\tilde{\theta} - \theta_0) = \sqrt{m}(\hat{\theta} - \theta_0) + \left(\frac{1}{m} \frac{\partial^2 \ln L(\hat{\theta})}{\partial \theta \partial \theta'} \right)^{-1} R^T \left[R \left(\frac{1}{m} \frac{\partial^2 \ln L(\hat{\theta})}{\partial \theta \partial \theta'} \right)^{-1} R^T \right]^{-1} \sqrt{m} [c - R(\hat{\theta} - \theta_0)].$$

As in the proof of Lemma 3, to compute the limiting distribution of $\sqrt{m}(\tilde{\theta} - \theta_0)$, we need to compute the limiting mean of $\frac{1}{m} \frac{\partial^2 \ln L(\hat{\theta})}{\partial \theta \partial \theta'}$ and the limiting distribution of $\sqrt{m}(\hat{\theta} - \theta_0)$.

For the former we have $-I^{-1}(\hat{\theta})$ while for the latter we have $\mathcal{N}(0, I(\theta_0))$. We therefore have

$$\text{Var}(\sqrt{m}(\tilde{\theta} - \theta_0)) \equiv \frac{1}{m}\Sigma = I^{-1}(\theta_0) - I^{-1}(\hat{\theta})R^T \left[RI^{-1}(\hat{\theta})R^T \right]^{-1} RI^{-1}(\theta_0).$$

Finally, we can substitute $\hat{I}(\tilde{\theta})$, the observed Fisher information at the constrained parameter estimates, for $I(\hat{\theta})$, as a consistent estimator of the observed Fisher information at the maximum of the likelihood, giving

$$\Sigma \approx \hat{\Sigma} = \frac{1}{m} \left[\hat{I}^{-1}(\tilde{\theta}) - \hat{I}^{-1}(\tilde{\theta})R^T \left[R\hat{I}^{-1}(\tilde{\theta})R^T \right]^{-1} R\hat{I}^{-1}(\tilde{\theta}) \right]. \quad \blacksquare \quad (\text{C4})$$

As expected, the constraints decrease the uncertainties in the parameter estimates.

-
- [1] K. Banaszek, G.M. D'Ariano, M.G.A. Paris, and M.F. Sacchi. Maximum likelihood estimation of the density matrix. *Phys. Rev. A*, 61:010304, 1999.
 - [2] O. E. Barndorff-Nielsen and R. D. Gill. Fisher information in quantum statistics. *J. Phys. A*, 33:4481–4490, 2000.
 - [3] S. L. Braunstein and C. M. Caves. Statistical distance and the geometry of quantum states. *Phys. Rev. Lett.*, 72:3439, 1994.
 - [4] V. Buzek. Reconstruction of quantum states of spin systems: from quantum bayesian inference to quantum tomography. *Ann. of Phys.*, 266:454–496, 1998.
 - [5] G. Casella and R. L. Berger. *Statistical Inference*. Duxbury Press, Pacific Grove, CA, 2002.
 - [6] R. Chakrabarti and H. Rabitz. Quantum control landscapes. *Int. Rev. Phys. Chem.*, 26:671–735, 2007.
 - [7] W. Hoeffding. Asymptotically optimal tests for multinomial distributions. *Ann. of Math. Stat.*, 36:369–408, 1963.
 - [8] C. W. Holevo. *Probabilistic and statistical aspects of quantum theory*. North-Holland, Amsterdam, 1982.
 - [9] C. W. Holevo. *Statistical structure of quantum theory*. Springer, Berlin, 2002.

- [10] Z. Hradil and J. Rehacek. Efficiency of maximum-likelihood reconstruction of quantum states. *Fortschr. Phys.*, 49:1083–1088, 2001.
- [11] K. R. W. Jones. Principles of quantum inference. *Ann. of Phys.*, 207:140–170, 1991.
- [12] G. Kimura. Bloch vector for n-level systems. 5:121, 2003.
- [13] A. Mansson, P. G. L. Mana, and G. Bjork. Numerical bayesian state assignment for a three-level quantum system. i. absolute frequency data; constant and gaussian priors. 5:121, 2007.
- [14] M. Mohseni, A. T. Rezakhani, and D. A. Lidar. Quantum process tomography: resource analysis of different strategies. pages 1–13, 2007.
- [15] W. Press, S. Teukolsky, W. Vetterling, and B. Flannery. *Numerical recipes in C⁺⁺*. Cambridge Univ. Press, Cambridge, 2002.
- [16] P. B. Slater. Bayesian quantum mechanics. *Nature*, 367:328, 1994.
- [17] T. Tilma and E. C. G. Sudarshan. Generalized euler angle parameterization of $su(n)$. *J. Phys. A*, 35:10467–10501, 2002.
- [18] C. van Eeden. Mean integrated squared error of kernel estimators when the density and its derivative are not necessarily continuous. *Ann. Inst. Statist. Math.*, 37:461, 1985.
- [19] W. K. Wootters and B. D. Fields. Optimal state determination by mutually unbiased measurements. *Ann. of Phys.*, 191:363–381, 1989.
- [20] O. Zeitouni and M. Gutman. On universal hypothesis testing via large deviations. *IEEE Transactions on Information Theory*, 37:285–290, 1991.
- [21] The principle of entropy maximization (PEM), is an estimation methodology that can consistently estimate all parameters with an incomplete observation level, since it implicitly assumes a prior plausibility distribution over the parameter space. However, it has been shown that the von Neumann entropy employed in PEM is not the appropriate measure of information-theoretic entropy; hence, we ignore it here
- [22] In frequentist statistics, the relative entropy is always defined in terms of the passage from the flat plausibility distribution to an asymptotically (multivariate) normal distribution.
- [23] In the alternative Euler angle parameterization, neither the unit trace nor the positive

semidefiniteness constraints are automatically satisfied.

- [24] An alternative approach to achieving greater asymptotic Fisher information, not considered here, is to execute simultaneous joint measurements on multiple particles.

<i>Panel A: Sample size 100; 100 repeated samples</i>					
	Bias	Median	Finite σ	MSE	95% conf intrv
<i>Parameter</i>					
θ_1					
θ_2					
θ_3					
ρ_{11}					
ρ_{22}					
<i>Panel B: Sample size 100; 1000 repeated samples</i>					
	Bias	Median	Finite σ	MSE	95% conf intrv
<i>Parameter</i>					
θ_1					
θ_2					
θ_3					
ρ_{11}					
ρ_{22}					
<i>Panel C: Sample size 1000; 1000 repeated samples</i>					
	Bias	Median	Finite σ	MSE	95% conf intrv
<i>Parameter</i>					
θ_1					
θ_2					
θ_3					
ρ_{11}					
ρ_{22}					

TABLE I: **Finite sample distribution statistics for state estimation of spin-1/2 quantum systems: MUB measurement bases. ρ_1**

<i>Panel A: Sample size 100</i>					
	Bias	Median	Finite σ	MSE	95% conf intrv
<i>Parameter</i>					
θ_1					
θ_2					
θ_3					
ρ_{11}					
ρ_{22}					
<i>Panel B: Sample size 400</i>					
	Bias	Median	Finite σ	MSE	95% conf intrv
<i>Parameter</i>					
θ_1					
θ_2					
θ_3					
ρ_{11}					
ρ_{22}					
<i>Panel C: Sample size 1000</i>					
	Bias	Median	Finite σ	MSE	95% conf intrv
<i>Parameter</i>					
θ_1					
θ_2					
θ_3					
ρ_{11}					
ρ_{22}					

TABLE II: **Finite sample distribution statistics (1000 repeated samples) for state estimation of spin-1/2 quantum systems: MUB measurement bases.** ρ_3 (pure)

<i>Panel A: Sample size 100</i>					
	Bias	Median	Finite σ	MSE	95% conf intrv
<i>Parameter</i>					
θ_1					
θ_2					
θ_3					
ρ_{11}					
ρ_{22}					

<i>Panel B: Sample size 400</i>					
	Bias	Median	Finite σ	MSE	95% conf intrv
<i>Parameter</i>					
θ_1					
θ_2					
θ_3					
ρ_{11}					
ρ_{22}					

<i>Panel C: Sample size 1000</i>					
	Bias	Median	Finite σ	MSE	95% conf intrv
<i>Parameter</i>					
θ_1					
θ_2					
θ_3					
ρ_{11}					
ρ_{22}					

TABLE III: **Finite sample distribution statistics (1000 repeated samples) for state estimation of spin-1 quantum systems: MUB measurement bases.** ρ_1

<i>Panel A: Sample size 100</i>					
	Bias	Median	Finite σ	MSE	95% conf intrv
<i>Parameter</i>					
θ_1					
θ_2					
θ_3					
ρ_{11}					
ρ_{22}					
<i>Panel B: Sample size 400</i>					
	Bias	Median	Finite σ	MSE	95% conf intrv
<i>Parameter</i>					
θ_1					
θ_2					
θ_3					
ρ_{11}					
ρ_{22}					
<i>Panel C: Sample size 1000</i>					
	Bias	Median	Finite σ	MSE	95% conf intrv
<i>Parameter</i>					
θ_1					
θ_2					
θ_3					
ρ_{11}					
ρ_{22}					

TABLE IV: **Finite sample distribution statistics (1000 repeated samples) for state estimation of spin-1 quantum systems: MUB measurement bases. ρ_3 (pure)**

<i>Panel A: Sample size 100</i>					
	Bias	Median	Finite σ	MSE	95% conf intrv
<i>Parameter</i>					
θ_1					
θ_2					
θ_3					
ρ_{11}					
ρ_{22}					

<i>Panel B: Sample size 400</i>					
	Bias	Median	Finite σ	MSE	95% conf intrv
<i>Parameter</i>					
θ_1					
θ_2					
θ_3					
ρ_{11}					
ρ_{22}					

<i>Panel C: Sample size 1000</i>					
	Bias	Median	Finite σ	MSE	95% conf intrv
<i>Parameter</i>					
θ_1					
θ_2					
θ_3					
ρ_{11}					
ρ_{22}					

TABLE V: **Finite sample distribution statistics (1000 repeated samples) for state estimation of spin-1/2 quantum systems: random non-MUB measurement bases.**

<i>Panel A: Sample size 100</i>					
	Bias	Median	Finite σ	MSE	95% conf intrv
<i>Parameter</i>					
θ_1					
θ_2					
θ_3					
ρ_{11}					
ρ_{22}					
<i>Panel B: Sample size 400</i>					
	Bias	Median	Finite σ	MSE	95% conf intrv
<i>Parameter</i>					
θ_1					
θ_2					
θ_3					
ρ_{11}					
ρ_{22}					
<i>Panel C: Sample size 1000</i>					
	Bias	Median	Finite σ	MSE	95% conf intrv
<i>Parameter</i>					
θ_1					
θ_2					
θ_3					
ρ_{11}					
ρ_{22}					

TABLE VI: **Finite sample distribution statistics (1000 repeated samples) for state estimation of spin-1 quantum systems: random non-MUB measurement bases.** ρ_1

<i>Panel A: Sample size 100</i>			
<i>T stat</i>		<i>Wald stat</i>	
T stat size	T stat power	Wald stat size	Wald stat power
<i>Hypothesis</i>			
$\tilde{\delta}_1 = \delta_{1,0}$			
$\tilde{\delta}_2 = \delta_{2,0}$			
$\tilde{\delta}_3 = \delta_{3,0}$			
$\tilde{\theta}_i = \theta_{i,0}, \forall i$			
<i>Panel B: Sample size 400</i>			
<i>T stat</i>		<i>Wald stat</i>	
T stat size	T stat power	Wald stat size	Wald stat power
<i>Hypothesis</i>			
$\tilde{\delta}_1 = \delta_{1,0}$			
$\tilde{\delta}_2 = \delta_{2,0}$			
$\tilde{\delta}_3 = \delta_{3,0}$			
$\tilde{\theta}_i = \theta_{i,0}, \forall i$			
<i>Panel C: Sample size 1000</i>			
<i>T stat</i>		<i>Wald stat</i>	
T stat size	T stat power	Wald stat size	Wald stat power
<i>Hypothesis</i>			
$\tilde{\delta}_1 = \delta_{1,0}$			
$\tilde{\delta}_2 = \delta_{2,0}$			
$\tilde{\delta}_3 = \delta_{3,0}$			
$\tilde{\theta}_i = \theta_{i,0}, \forall i$			

TABLE VII: Finite sample test statistic size, power for state estimation of spin-1/2 quantum systems (1000 repeated samples). ρ_1

<i>Panel A: Sample size 100</i>			
<i>T stat</i>		<i>Wald stat</i>	
T stat size	T stat power	Wald stat size	Wald stat power
<i>Hypothesis</i>			
$\tilde{\delta}_1 = \delta_{1,0}$			
$\tilde{\delta}_2 = \delta_{2,0}$			
$\tilde{\delta}_3 = \delta_{3,0}$			
$\tilde{\theta}_i = \theta_{i,0}, \forall i$			

<i>Panel B: Sample size 400</i>			
<i>T stat</i>		<i>Wald stat</i>	
T stat size	T stat power	Wald stat size	Wald stat power
<i>Hypothesis</i>			
$\tilde{\delta}_1 = \delta_{1,0}$			
$\tilde{\delta}_2 = \delta_{2,0}$			
$\tilde{\delta}_3 = \delta_{3,0}$			
$\tilde{\theta}_i = \theta_{i,0}, \forall i$			

<i>Panel C: Sample size 1000</i>			
<i>T stat</i>		<i>Wald stat</i>	
T stat size	T stat power	Wald stat size	Wald stat power
<i>Hypothesis</i>			
$\tilde{\delta}_1 = \delta_{1,0}$			
$\tilde{\delta}_2 = \delta_{2,0}$			
$\tilde{\delta}_3 = \delta_{3,0}$			
$\tilde{\theta}_i = \theta_{i,0}, \forall i$			

TABLE VIII: Finite sample test statistic size, power for state estimation of spin-1 quantum systems (1000 repeated samples). ρ_1