SAND2016-8677J

Behavior of the Maximum Likelihood in Quantum State Tomography

Travis L Scholten and Robin Blume-Kohout
Center for Computing Research (CCR), Sandia National Labs and *University of New Mexico*
(Dated: July 27, 2016)

Quantum state tomography on a d-dimensional system demands resources that grow rapidly with d. Model selection can be used to tailor the number of fit parameters to the data, but some common assumptions that underly canonical model selection techniques are often violated. This is due to the presence of boundaries in the quantum state space. Here, we study the behavior of the maximum likelihood in different Hilbert space dimensions, and derive an expression for a complexity penalty – the expected value of the loglikelihood ratio statistic (roughly, the logarithm of the maximum likelihood) – which can be used to make an appropriate choice for d.

In quantum information science, an experimentalist may wish to determine the quantum state ρ_0 that is produced by a specific initialization procedure. This can be done using quantum state tomography [1]: many copies of ρ_0 are produced; they are measured in diverse ways; and finally the outcomes of those measurements (data) are collated and analyzed to produce an estimate $\hat{\rho}$. This is a straightforward statistical inference process [2, 3], where the data are used to fit the parameters of a statistical model – provided that we know what statistical model to use. But this is not always the case. In state tomography, the parameter is ρ, the model is the set of all possible density matrices on a Hilbert space $\mathcal{H}$ (equipped with the Born rule). It is not always *a priori* obvious what $\mathcal{H}$ or its dimension is; examples include optical modes [4–8] and leakage levels in AMO and superconducting [9, 10] qubits. Choosing an appropriate Hilbert space on the fly is *model selection*, and while model selection is well-studied in classical statistics [11], applying it to quantum tomography leads to some surprising twists. These problems stem from the positivity boundary ($\rho \geq 0$), and the techniques we use to resolve them are broadly applicable to quantum tomography on low-rank states.

Selecting between multiple models often requires fitting each model's parameters using *maximum likelihood estimation* (MLE) [12–14], which reports the parameter values that maximize the likelihood (the probability of the observed data). Classical estimation problems usually satisfy *local asymptotic normality* [15, 16], meaning that: (1) as $N_{\text{samples}} \to \infty$, $\hat{\rho}_{\text{MLE}}$ is normally distributed around the true state ρ_0 with covariance matrix $\mathcal{I}^{-1}$, and (2) the likelihood function in a neighborhood of $\hat{\rho}_{\text{MLE}}$ is locally Gaussian with Hessian $\mathcal{I}$, where $\mathcal{I}$ is the *Fisher information*.

In state tomography, these conditions are violated by the constraint $\hat{\rho}_{\text{MLE}} \geq 0$. This constraint distorts the distribution of estimates, even in the asymptotic limit, precisely when ρ_0 is rank-deficient within $\mathcal{H}$. (That is, when ρ_0 is on the boundary of $\mathcal{H}$. See Figure 1.) While the behavior of MLEs is well-studied classically, its behavior near boundaries is not. How boundaries affect $\hat{\rho}_{\text{MLE}}$ – and its derived properties – is of broad interest in state and process estimation [17, 18], and is also critical for model selection [19–23].

We will focus on the special and simple case where

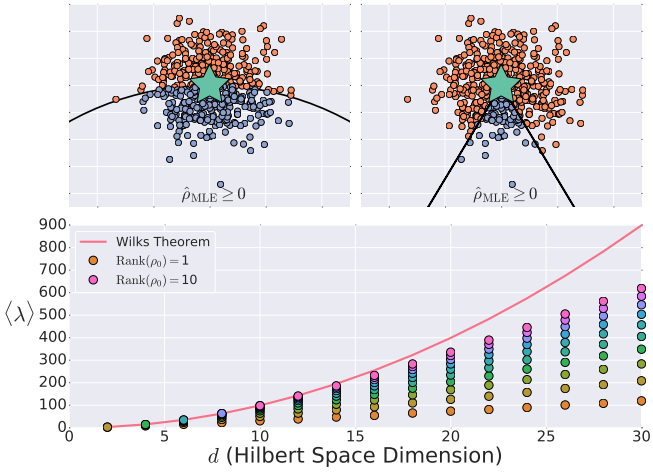

FIG. 1: Impact of the boundary on maximum likelihood tomography. **Top**: Two views through the qutrit state space. Without the positivity constraint, some estimates (orange circles) are negative, and do not represent valid estimates of a quantum state. The distribution of $\hat{\rho}_{\text{MLE}}$ (blue circles) is generally non-normal, and depends on the true state ρ_0 (star). **Bottom**: Comparison of the classical theory (Wilks Theorem) prediction for the loglikelihood ratio $\langle\lambda\rangle$ to numerical data for true states of rank $r = 1 \ldots 10$. The Wilks Theorem fails badly for low-rank states; our main result (Equation 10) fixes this problem (see Figure 3).

$\mathcal{I}$ at ρ_0 is proportional to the Hilbert-Schmidt metric, so the likelihood function (and the distribution of the *unconstrained* estimates $\hat{\rho}$) is given by

$$\mathcal{L}(\rho) = \Pr(\hat{\rho}|\rho) \propto e^{-||\rho-\hat{\rho}||_2^2/2\epsilon^2} \tag{1}$$

for some ϵ that scales as $1/\sqrt{N_{\text{samples}}}$. In practice, $\mathcal{I}$ depends on ρ_0 and the particular tomographic measurement performed, but the interaction of an arbitrary Fisher information with the boundary is complex and intractable. The isotropic assumption greatly simplifies our study of the problem, and permits the derivation of analytic results which capture realistic tomographic scenarios surprisingly well.

In this case, $\hat{\rho}$ is often negative (see Figure 1). For each such $\hat{\rho}$, the actual $\hat{\rho}_{\text{MLE}}$ is the solution to the following

optimization problem:

$$\hat{\rho}_{\text{MLE}} = \underset{\substack{\rho\in\mathcal{B}(\mathcal{H})\\ \text{Tr}(\rho)=1,\rho\geq 0}}{\text{argmin}} \text{Tr}[(\hat{\rho}-\rho)^2], \tag{2}$$

and the distribution of $\hat{\rho}_{\text{MLE}}$ is not generally normal. We do not attempt to derive $\Pr(\hat{\rho}_{\text{MLE}})$ explicitly. Instead, we demonstrate a method for computing the behavior of useful statistics which depend on $\hat{\rho}_{\text{MLE}}$, such as the expected value of the loglikelihood ratio statistic $\langle\lambda\rangle$, a key quantity in model selection.

MODEL SELECTION: Model selection becomes relevant when several candidate models could be fit to the data. A model's "size" is the number of free parameters in it, and as a general rule, the best model is the smallest one that fits the data well. Many model selection techniques use a statistic to quantify goodness of fit; a commonly used statistic is the *loglikelihood ratio* [14, 23],

$$\lambda(\mathcal{M}_1,\mathcal{M}_2) = -2\log\left(\frac{\max\limits_{\rho\in\mathcal{M}_1}\mathcal{L}(\boldsymbol{\theta})}{\max\limits_{\rho\in\mathcal{M}_2}\mathcal{L}(\boldsymbol{\theta})}\right). \tag{3}$$

Intuitively, the model with the higher likelihood is more plausible – except that models with more adjustable parameters will almost always fit the data better! This is very clear in the case of *nested* models (the smaller is a submanifold of the larger) [45]. If two models are equally valid – i.e. they both contain the true state ρ_0 – the larger one will nonetheless usually fit the data better because its extra parameters allow it to fit more of the noise in the data. For the same reason, the larger model's fit will be less accurate. This makes it imperative to correct for overfitting, by handicapping larger models.

For this reason, any model selection method that relies (explicitly or implicitly) on a statistic to quantify "how well Model $\mathcal{M}$ fits the data" also relies on a *null theory* to predict how that statistic will behave *if* $\rho_0\in\mathcal{M}$ [46]. A model selection criterion based on an invalid null theory (or a criterion used in a context where its null theory does not apply) will tend to choose the wrong model.

When local asymptotic normality holds if $\rho_0\in\mathcal{M}_1\subset\mathcal{M}_2$, where $\mathcal{M}_1$ has k free parameters and $\mathcal{M}_2$ has $K+k$ free parameters, then the null theory for λ is given by the *Wilks Theorem* [24]: λ is a χ^2_K random variable, so that $\langle\lambda\rangle = K$.

Suppose we used model selection to decide whether a particular d-dimensional Hilbert space $\mathcal{H}_d$ is suitable by compare $\mathcal{H}_d$ to $\mathcal{H}_{d+1}$ using λ. We define the model $\mathcal{M}_d$ as

$$\mathcal{M}_d = \{\rho \mid \rho\in\mathcal{B}(\mathcal{H}_d),\ \text{Tr}(\rho)=1,\ \rho\geq 0\}, \tag{4}$$

where $\mathcal{B}(\mathcal{H})$ is the space of bounded operators on $\mathcal{H}$. Now, before we can use an empirically observed λ to decide whether $\mathcal{H}_{d+1}$ is significantly better, we need a null theory for its behavior when it *isn't* better (i.e., when ρ_0 is in both $\mathcal{M}_d$ and $\mathcal{M}_{d+1}$). In what follows, it's useful to reduce the problem of computing $\lambda(\mathcal{M}_d,\mathcal{M}_{d+1})$ to that of computing $\lambda(\rho_0,\mathcal{M}_d)$ and $\lambda(\rho_0,\mathcal{M}_{d+1})$ using the identity

$$\lambda(\mathcal{M}_d,\mathcal{M}_{d+1}) = \lambda(\rho_0,\mathcal{M}_{d+1}) - \lambda(\rho_0,\mathcal{M}_d), \tag{5}$$

where

$$\lambda(\rho_0,\mathcal{M}_d) = -2\log\left(\frac{\mathcal{L}(\rho_0)}{\max\limits_{\rho\in\mathcal{M}_d}\mathcal{L}(\rho)}\right). \tag{6}$$

When ρ_0 is full-rank, the Wilks Theorem appplies, and $\lambda\sim\chi^2_{d^2-1}$ [47]. However, the behavior of λ depends crucially on the *rank* of ρ_0. If ρ_0 is rank-deficient, then the boundary looms, local asymptotic normality does not hold, and the Wilks Theorem does not apply! Even if ρ_0 is full-rank in $\mathcal{M}_d$, it will be rank-deficient in $\mathcal{M}_{d+1}$. Thus, to do model selection for Hilbert space dimension in state tomography, we must understand the behavior of $\lambda(\rho_0,\mathcal{M}_d)$ – i.e., derive a replacement Wilks Theorem – for rank-deficient ρ_0. Some of the more technical details of this derivation are deferred to Appendix I in the Supplementary Material.

DERIVING A WILKS REPLACEMENT: In this derivation, we assume that $\rho_0,\hat{\rho}_{\text{MLE}}\in\mathcal{M}_d$, that $r\equiv\text{Rank}(\rho_0)<d$, and that the Fisher information at ρ_0 is $\mathcal{I}=\epsilon^2\mathbb{1}$. The loglikelihood ratio that we're trying to predict is equal to $\lambda(\rho_0,\hat{\rho}_{\text{MLE}}) = \epsilon^{-2}\text{Tr}[(\hat{\rho}_{\text{MLE}}-\rho_0)^2]$, where $\hat{\rho}_{\text{MLE}}$ is the MLE over $\mathcal{M}_d$. The *unconstrained* MLE is distributed as $\Pr(\hat{\rho}) = \mathcal{N}(\rho_0,\epsilon^2\mathbb{1})$.

We need a procedure to compute $\hat{\rho}_{\text{MLE}}$ given $\hat{\rho}$ – i.e., to solve the optimization problem in Eq. (2). Fortunately, for the special case of isotropic Fisher information, this problem was solved in Ref. [25]:

1. Subtract $q\mathbb{1}$ from the unconstrained $\hat{\rho}$, for a particular real scalar q,
2. "Truncate" $\hat{\rho}-q\mathbb{1}$, by replacing each of its negative eigenvalues with zero.

Here, q is defined implicitly such that $\text{Tr}\left[\text{Trunc}(\hat{\rho}-q\mathbb{1})\right]=1$.

Although this was intended as a (very fast) numerical algorithm, we will abuse it (by a series of approximations) to derive a closed-form expression for the average $\langle\lambda\rangle$. We begin by observing that $\lambda(\rho_0,\hat{\rho}_{\text{MLE}})$ can be written as a sum over matrix elements,

$$\begin{aligned}\lambda &= \epsilon^{-2}\text{Tr}[(\hat{\rho}_{\text{MLE}}-\rho_0)^2] = \epsilon^{-2}\sum_{jk}||(\hat{\rho}_{\text{MLE}}-\rho_0)_{jk}||^2\\ &= \sum_{jk}\lambda_{jk} \quad \text{where} \quad \lambda_{jk} = \epsilon^{-2}||(\hat{\rho}_{\text{MLE}}-\rho_0)_{jk}||^2,\end{aligned} \tag{7}$$

and therefore $\langle\lambda\rangle = \sum_{jk}\langle\lambda_{jk}\rangle$. Each term $\langle\lambda_{jk}\rangle$ quantifies the average mean-squared error of a single matrix element of MLE, and while the Wilks Theorem predicts $\langle\lambda_{jk}\rangle = 1$ for all j,k, numerical simulations (see Fig. 2) show that this only holds true for *some* matrix elements. Others contribute much less than 1 unit on average, meaning that the Wilks Theorem predicts too high

a value for the total $\langle\lambda\rangle$. Thus motivated, we divide the parameters of $\hat{\rho}$ into two parts (see Fig. 2),

- The "kite" comprises all diagonal elements *and* all elements on the kernel (null space) of ρ_0,
- The "L" comprises all off-diagonal elements on the support of ρ_0 *and* between the support and the kernel,

and observe that $\langle\lambda\rangle = \langle\lambda\rangle_{\mathrm{L}} + \langle\lambda\rangle_{\mathrm{kite}}$. The rationale for this division is simple: small fluctuations on the "L" do not change the zero eigenvalues of $\hat{\rho}$ to 1st order, whereas those on the "kite" do.

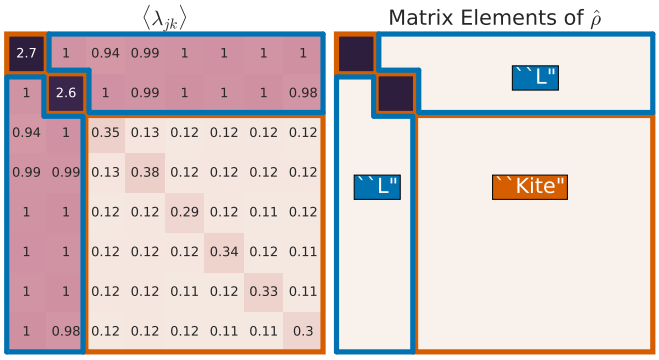

FIG. 2: When a rank-2 state is reconstructed in $d = 8$ dimensions, the total loglikelihood ratio $\lambda(\rho_0, \mathcal{M}_8)$ is the sum of contributions λ_{jk} from errors in each matrix element $(\hat{\rho}_{\mathrm{MLE}})_{jk}$. **Left**: Numerics show a clear division; some matrix elements contribute $\langle\lambda_{jk}\rangle \sim 1$ as predicted by the Wilks Theorem, while others contribute either more or less. **Right**: The numerical results motivate dividing the elements of $\hat{\rho}$ into two parts: the "kite" and the "L".

The error of the unconstrained estimate is $\delta = \hat{\rho} - \rho_0$, a normally-distributed *traceless* matrix. To simplify the analysis, we explicitly drop the $\mathrm{Tr}(\rho) = 1$ constraint and let δ be $\mathcal{N}(0, \epsilon^2 \mathbb{1})$ distributed over the d^2-dimensional space of Hermitian matrices (a good approximation when $d \gg 2$), which makes δ proportional to an element of the Gaussian unitary ensemble (GUE) [26].

Now, to first order in ϵ, elements of δ in the "L" do not affect positivity, so they are unconstrained by the boundary, and behave exactly as expected from classical theory. Each matrix element δ_{jk} in the "L" has Gaussian fluctuations and so for those elements, $\langle\lambda_{jk}\rangle = 1$. As there are $2rd - r(r+1)$ of them in the "L", $\langle\lambda\rangle_{\mathrm{L}} = 2rd - r(r+1)$.

Computing $\langle\lambda\rangle_{\mathrm{kite}}$ is a bit harder, because the boundary *does* constrain elements in the "kite". Here, we turn to the "truncation" algorithm given above for finding $\hat{\rho}_{\mathrm{MLE}}$, which is most naturally performed in the eigenbasis of $\hat{\rho}$. Exact diagonalization of $\hat{\rho}$ is not feasible analytically, but only the *small* eigenvalues of $\hat{\rho}$ are critical in truncation. As long as all the nonzero eigenvalues of ρ_0 are much larger than ϵ, the eigenbasis of $\hat{\rho}$ is accurately approximated by: (1) the eigenvectors of ρ_0 on its support; and (2) the eigenvectors of $\delta_{\mathrm{ker}} = \Pi_{\mathrm{ker}} \delta \Pi_{\mathrm{ker}}$, where Π_{ker} is the projector onto the kernel of ρ_0.

Changing to this basis diagonalizes the "kite" portion of δ, and leaves all elements of the "L" unchanged (to 1st order in ϵ). The diagonal elements of $\hat{\rho}$ now fall into two categories:

1. r elements corresponding to the eigenvalues of ρ_0 and given by $p_j = \rho_{jj} + \delta_{jj}$ where $\delta_{jj} \sim \mathcal{N}(0, \epsilon^2)$.
2. $N \equiv d - r$ elements that are eigenvalues of δ_{ker}, which we denote $\boldsymbol{\kappa} = \{\kappa_j : \ j = 1 \ldots N\}$.

The κ_j are random variables, but are not normally distributed. Instead, δ_{ker} is proportional to a GUE(N) matrix. For $N \gg 1$, the distribution of any eigenvalue converges to a Wigner semicircle distribution [27] given by $\Pr(\kappa) = \frac{2}{\pi R^2}\sqrt{R^2 - \kappa^2}$ for $|\kappa| \leq R$, with $R = 2\epsilon\sqrt{N}$. The eigenvalues are not independent; they tend to avoid collisions ("level avoidance" [28]), and typically form a surprisingly regular array over the support of the Wigner semicircle. Since our goal is to compute $\langle\lambda\rangle$, we can capitalize on this behavior by replacing each random sample of $\boldsymbol{\kappa}$ with a *typical sample* $\bar{\boldsymbol{\kappa}}$ given by its order statistics. These are the average values of the *sorted* $\boldsymbol{\kappa}$, so $\overline{\kappa}_j$ is the average value of the jth largest value of $\boldsymbol{\kappa}$. Large random samples are usually well approximated (for many purposes) by their order statistics even when the elements of the sample are independent, and level avoidance makes the approximation even better. We make one further approximation, by assuming (as an ansatz) that $N \gg 1$, and thus that the distribution of the $\overline{\kappa}_j$ is effectively continuous and identical to $\Pr(\kappa)$. (See Appendix I for a more detailed discussion of this series of approximations.)

To proceed with truncation, we observe that the $\overline{\kappa}_j$ are symmetrically distributed around $\kappa = 0$, so half of them are negative. Therefore, with high probability, $\mathrm{Tr}\left[\mathrm{Trunc}(\hat{\rho})\right] > 1$, and so we will need to subtract $q\mathbb{1}$ from $\hat{\rho}$ before truncating. The appropriate q solves $\mathrm{Tr}\left[\mathrm{Trunc}(\hat{\rho} - q\mathbb{1})\right] = 1$. This equation can be solved using the ansatz established so far, and some series expansions (see Appendix I), and yields the solution:

$$q \approx 2\sqrt{N} - \frac{(240r\pi)^{2/5}}{4} N^{1/10} + \frac{(240r\pi)^{4/5}}{80} N^{-3/10}. \quad (8)$$

Now that we know how much to subtract off in the truncation process, we can compute $\langle\lambda\rangle_{\mathrm{kite}}$. Defining

$(x)^+ = \max(x, 0)$:

$$\begin{aligned}\langle\lambda\rangle_{\text{kite}} &= \sum_{j=1}^{r}(\rho_{jj} - (p_j - q))^2 + \sum_{j=1}^{N}\left[(\overline{\kappa}_j - q)^+\right]^2 \\ &\approx r + rq^2 + N\int_{\kappa=q}^{2\epsilon\sqrt{N}} Pr(\kappa)(\kappa - q)^2 d\kappa \\ &= r + rq^2 + \frac{N(N+q^2)}{\pi}\left(\frac{\pi}{2} - \sin^{-1}\left(\frac{q}{2\sqrt{N}}\right)\right) \\ &\quad - \frac{q(q^2+26N)}{24\pi}\sqrt{4N-q^2}. \end{aligned} \tag{9}$$

Thus, the total expected value, $\langle\lambda\rangle = \langle\lambda\rangle_{\text{L}} + \langle\lambda\rangle_{\text{kite}}$, is

$$\begin{aligned}\langle\lambda(\rho_0, \mathcal{M}_d)\rangle &= 2rd - r^2 + rq^2 \\ &\quad + \frac{N(N+q^2)}{\pi}\left(\frac{\pi}{2} - \sin^{-1}\left(\frac{q}{2\sqrt{N}}\right)\right) \\ &\quad - \frac{q(q^2+26N)}{24\pi}\sqrt{4N-q^2} \end{aligned} \tag{10}$$

where q is given in Equation (8), $N = d - r$, and $r = \text{Rank}(\rho_0)$.

Equation (10) is our main result. To test its validity, we compare it to numerical simulations for $d = 2, \ldots, 30$ and $r = 1, \ldots, 10$, in Figure 3. The prediction of Wilks theorem is wildly incorrect for $r \ll d$. In contrast, Equation 10 is almost perfectly accurate when $r \ll d$, but it does begin to break down (albeit fairly gracefully) as r becomes comparable to d. We conclude that our analysis (and Equation (10)) correctly models tomography *if* the Fisher information is isotropic ($\mathcal{I} \propto \mathbb{1}$).

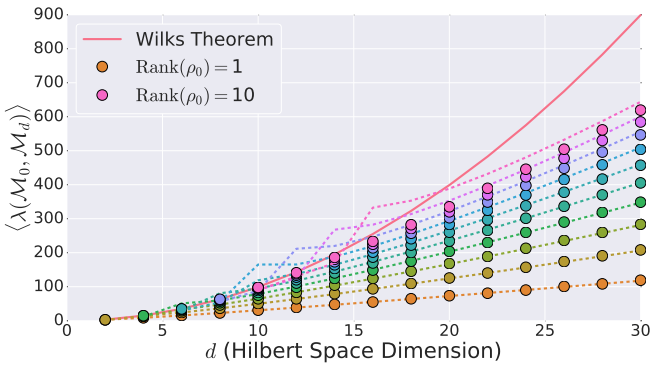

FIG. 3: Numerical results for $\langle\lambda\rangle$ (dots) compared to the prediction of the Wilks Theorem (solid red line) and our replacement theory as given in Eq. 10, (dashed lines). Our formula depends on the rank r of ρ_0 (unlike the Wilks prediction), and is nearly perfect for $r \ll d$. It becomes less accurate as r approaches $d/2$, and is invalid when $r \approx d$.

HETERODYNE TOMOGRAPHY: In practice, the Fisher information is rarely isotropic. So we tested our idealized result by applying it to a realistic, challenging, and experimentally relevant problem: quantum heterodyne state tomography [5, 6, 8, 29] of a single optical mode. States of this continuous-variable system are described by density operators on the infinite-dimensional Hilbert space $L^2(\mathbb{R})$. Fitting these infinitely many parameters to finitely much data demands simpler models. We consider a family of nested models motivated by a low-energy (few-photon) ansatz, and choose the Hilbert space $\mathcal{H}_d$ to be that spanned by the photon number states $\{|0\rangle \ldots |d-1\rangle\}$. Heterodyne (a.k.a. double homodyne) tomography reconstructs ρ_0 using data from repeated measurements of the coherent-state POVM, $\{|\alpha\rangle\langle\alpha|/\pi,\ \alpha = x + ip \in \mathbb{C}\}$, which corresponds to sampling directly from the state's Husimi Q-function [30].

We examined the behavior of λ for 13 distinct true states ρ_0, both pure and mixed, supported on $\mathcal{H}_2, \mathcal{H}_3, \mathcal{H}_4$, and $\mathcal{H}_5$. We used rejection sampling to simulate 100 heterodyne datasets with up to $N_{samples} = 10^5$, and found MLEs over each of the 9 models $\mathcal{M}_2, \ldots, M_{10}$ using numerical optimization [48]. For each true state and each d, we averaged $\lambda(\rho_0, \mathcal{M}_d)$ over all 100 datasets to obtain an empirical average loglikelihood ratio $\bar{\lambda}$ for each (ρ_0, d) pair.

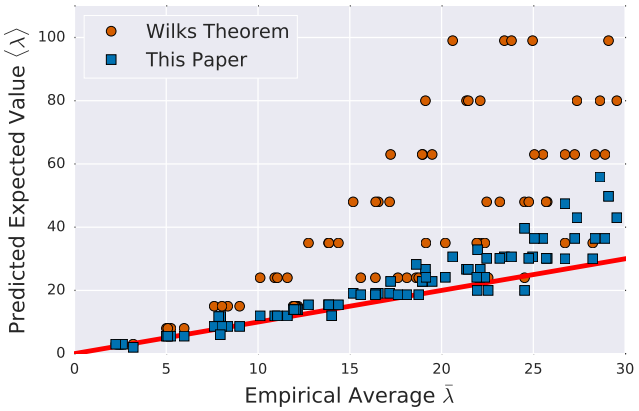

FIG. 4: The Wilks Theorem (orange dots) dramatically overestimates $\langle\lambda(\rho_0, \mathcal{M}_d)\rangle$ in optical heterodyne tomography. Our formula, Equation 10 (blue squares), is far more accurate. Residual discrepancies, discussed in Appendix III, appear to stem largely from non-asymptoticity (N_{samples} is not yet "asymptotically large"). The solid red line of equality corresponds to perfect correlation between theory ($\langle\lambda\rangle$) and practice ($\bar{\lambda}$).

Results of this test are shown in Figure 4, where we plot (a) the Wilks prediction, and (b) Equation (10), against the empirical $\bar{\lambda}$, for a variety of true states and reconstruction dimensions. Our formula correlates very well with the empirical average, while the Wilks Theorem (unsurprisingly) overestimates λ dramatically for low-rank states. Whereas a model selection procedure based on Wilks Theorem would tend to falsely reject larger Hilbert spaces (by setting the threshold for acceptance too high), our formula provides a reliable null theory.

Interestingly, as d grows, Eq. (10) also begins to overpredict. Further investigation (see Appendix III) indicates that a more accurate description is that the numer-

ical experiments are *underachieving*, because N_{samples} is not large enough – $\bar{\lambda}$ is still growing with N_{samples}. Our analysis is explicitly asymptotic, and does not cover this case. Failing to reach asymptotic behavior at $N_{\text{samples}} = 10^5$ is surprising, and suggests that heterodyne tomography is a particular exceptional and challenging case to model statistically.

CONCLUSIONS: Wilks Theorem is not generally reliable in quantum state tomography, but our Eq. (10) provides a much more broadly applicable replacement that can be used in model selection methods. This includes protocols like the AIC and BIC [11, 31, 32] that do not explicitly use the Wilks Theorem, but rely on the same assumptions (asymptotic normality, etc). Null theories of the loglikelihood ratios have many other applications, including hypothesis testing [14, 23] and confidence regions [33], and our result is directly applicable to them. Refs. [23, 33] both point out explicitly that their methods are unreliable near boundaries and therefore cannot be applied to rank-deficient states; our result fixes this outstanding problem. However, our numerical experiments with heterodyne tomography show (perhaps surprisingly) that quantum tomography can still surprise, and may violate *all* asymptotic statistics results. In such cases, bootstrapping [34] may be the only reliable way to construct null theories for λ. Finally, the *methods* presented here have application beyond the analysis of loglikelihoods. They shed light on the behavior of $\hat{\rho}_{\text{MLE}}$ for rank-deficient states, and can be used to predict other derived properties such as the average rank of the estimate, which is independently interesting for (e.g.) quantum compressed sensing [18, 35, 36].

ACKNOWLEDGEMENTS: The authors are grateful for those who provide support for the following software packages: iPython/Jupyter [37], matplotlib [38], mpi4py [39], NumPy [40], pandas [41], Python 2.7 [42], seaborn [43], and SciPy [44]. TLS thanks Jonathan A Gross for helpful discussions on software design, coding, and statistics, as well as John King Gamble for useful insights on parallelized computation and feedback on earlier drafts of this paper.

Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

[1] M. G. A. Paris and J. Rehacek, eds., *Quantum State Estimation* (Springer, Berlin-Heidelberg, 2004).
[2] N. Reid and D. R. Cox, International Statistical Review **83**, 293 (2015).
[3] L. Wasserman, *All of Statistics: A Concise Course in Statistical Inference* (Springer New York, 2004).
[4] J. B. Altepeter, E. R. Jefferey, and P. G. Kwiat, in *Advances in Atomic, Molecular, and Optical Physics* (Elsevier B.V., 2005), vol. 52, chap. 2, pp. 1–54.
[5] J. I. Bertrand and P. Bertrand, Foundations of Physics **17**, 397 (1987).
[6] A. I. Lvovsky, Reviews of Modern Physics **81**, 299 (2009).
[7] G. Breitenbach, S. Schiller, and J. Mlynek, Nature **387**, 471 (1997).
[8] U. Leonhardt and H. Paul, Progress in Quantum Electronics **19**, 89 (1995).
[9] F. Motzoi, J. M. Gambetta, P. Rebentrost, and F. K. Wilhelm, Physical Review Letters **103**, 1 (2009).
[10] R. Fazio, G. Palma, and J. Siewert, Physical Review Letters **83**, 5385 (1999).
[11] K. P. Burnham, Sociological Methods & Research **33**, 261 (2004).
[12] Z. Hradil, Physical Review A **55**, 1561 (1997).
[13] D. F. V. James, P. G. Kwiat, W. J. Munro, and A. G. White, Physical Review A **64**, 1 (2001).
[14] R. Blume-Kohout, J. O. S. Yin, and S. J. Van Enk, Physical Review Letters **105**, 1 (2010).
[15] L. Le Cam, Annals of Mathematical Statistics **41**, 802 (1970).
[16] L. Le Cam, Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability pp. 129–156 (1956).
[17] E. J. Candes and T. Tao, IEEE Transactions on Information Theory **52**, 5406 (2006).
[18] S. Flammia, D. Gross, Y. Liu, and J. Eisert, New Journal of Physics **14** (2012).
[19] L. Schwarz and S. van Enk, Physical Review A **88** (2013).
[20] M. Guta, T. Kypraios, and I. Dryden, New Journal of Physics **14** (2012).
[21] S. J. Van Enk and R. Blume-Kohout, New Journal of Physics **15** (2013).
[22] J. O. S. Yin and S. J. van Enk, Physical Review A **83**, 062110 (2011).
[23] T. Moroder, M. Kleinmann, P. Schindler, T. Monz, O. Guhne, and R. Blatt, Physical Review Letters **110**, 1 (2013).
[24] S. S. Wilks, The Annals of Mathematical Statistics **9**, 60 (1938).
[25] J. A. Smolin, J. M. Gambetta, and G. Smith, Physical Review Letters **108**, 1 (2012).
[26] Y. V. Fyodorov, *Introduction to the Random Matrix Theory: Gaussian Unitary Ensemble and Beyond* (Cambridge University Press, 2005).
[27] E. P. Wigner, Annals of Mathematics **67**, 325 (1958).
[28] T. Tao and V. Vu, Random Matrices: Theory and Applications **2** (2013).
[29] A. I. Lvovsky, H. Hansen, T. Aichele, O. Benson, J. Mlynek, and S. Schiller, Physical Review Letters **87**, 050402 (2001).
[30] K. Husimi, in *Proceedings of the Physico-Mathematical Society of Japan* (1940), vol. 22, pp. 264 – 314.
[31] H. Akaike, IEEE Transactions on Automatic Control **19**, 716 (1974).
[32] G. Schwarz, The Annals of Statistics **6**, 461 (1978).
[33] S. Glancy, E. Knill, and M. Girard, New Journal of Physics **14** (2012).

[34] B. Efron, The Annals of Statistics **7**, 1 (1979).
[35] A. Kalev and C. H. Baldwin, arxiv:arXiv:1511.01433v1 pp. 1–5 (2015).
[36] A. Kalev, R. L. Kosut, and I. H. Deutsch, npj Quantum Information **1**, 15018 (2015).
[37] F. Pérez and B. E. Granger, Computing in Science and Engineering **9**, 21 (2007).
[38] J. D. Hunter, Computing in Science and Engineering **9**, 90 (2007).
[39] L. Dalcin, R. Paz, and M. Storti, Journal of Parallel and Distributed Computing **65**, 1108 (2005).
[40] S. Van Der Walt, S. C. Colbert, and G. Varoquaux, Computing in Science and Engineering **13**, 22 (2011).
[41] W. McKinney, in *Proceedings of the 9th Python in Science Conference*, edited by S. van der Walt and J. Millman (2010), pp. 51–56.
[42] G. van Rossum, *Python Language Reference* (1995).
[43] M. Waskom (2016).
[44] T. E. Oliphant, Computing in Science and Engineering **9**, 10 (2007).
[45] If $\mathcal{M}_1 \subset \mathcal{M}_2$, then the maximum likelihood of $\mathcal{M}_2$ is at least as high as that of $\mathcal{M}_1$.
[46] Knowing the behavior of the statistic when $\rho_0 \in \mathcal{M}$ allows us, once we've taken some data, to compute the observed value of the statistic, and thereby decide whether ρ_0 is in fact in $\mathcal{M}$.
[47] Because ρ_0 is *fixed*, the number of free parameters is 0.
[48] The model $\mathcal{M}_1$ is trivial, in the sense $\mathcal{M}_1 = |0\rangle\langle 0|$. This model will almost always be wrong, unless one is actually trying to prepare the vacuum state.

SUPPLEMENTAL MATERIAL

I. DERIVATION OF THE WILKS REPLACEMENT: TECHNICAL DETAILS

In our derivation of the loglikelihood ratio statistic's expected value $\langle\lambda\rangle$, we used several approximations and ansätze. Perhaps the most important was the replacement of a random sample of GUE eigenvalues, $\boldsymbol{\kappa}$, by a "typical" sample $\overline{\kappa}$ (and the subsequent replacement of this discrete typical sample with a continuous density of values. Let us discuss this in a bit more detail.

FIG. 5: Typical samples of GUE(N) eigenvalues are accurately approximated by order statistics of the distribution (average values of sorted sample). **Top**: The sorted eigenvalues of one randomly chosen GUE(100) matrix. **Bottom**: Approximate values of the order statistics of the GUE(100) distribution, computed as the average of 100 the sorted eigenvalues of 100 randomly chosen GUE(100) matrices.

First, we illustrate the claim made in the main text that the eigenvalues of a GUE(N) matrix are usually well-approximated by their order statistics. Suppose that $\boldsymbol{\kappa}$ are the eigenvalues of a GUE(N) matrix, sorted from

highest to lowest. Figure 5 illustrates such a sample for $N = 100$. It also shows the *average* values of 100 such samples (all sorted). These are the *order statistics* $\overline{\kappa}$ of the distribution (more precisely, what is shown is a good *estimate* of the order statistics; the actual order statistics would be given by the average over infinitely many samples). The point of the figure is to show that, while the order statistics *are* slightly more smoothly and predictably distributed than a single (sorted) sample... the two are remarkably similar. A single sample $\boldsymbol{\kappa}$ will fluctuate around the order statistics, but these fluctuations are relatively small, partly because the sample is large, and partly because the GUE eigenvalues experience level repulsion. Thus, the "typical" behavior of a sample – by which we mean the mean value of a statistic of the sample – is well captured by the order statistics (which have no fluctuations at all).

We now turn to the problem of modeling $\boldsymbol{\kappa}$ quantitatively. We note up front that we are only going to be interested in certain properties of $\boldsymbol{\kappa}$: specifically, partial sums of all κ_i greater or less than some threshold (this threshold is q in the main text), or partial sums of functions of the κ_i (e.g. $(\kappa_i - q)^2$). We require only that an ansatz be accurate for such quantities. We do not use this fact explicitly, but it motivates our approach – and we do not claim that our ansatz is accurate for *all* conceivable statistics.

In general, if a sample $\boldsymbol{\kappa}$ of size N is drawn so that each κ has the same probability distribution function $\Pr(\kappa)$, then a good approximation for the jth order statistic is given by the inverse *cumulative* distribution function (c.d.f.):

$$\overline{\kappa}_j \approx \mathrm{CDF}^{-1}\left(\frac{j-1/2}{N}\right). \tag{11}$$

This is closely related to the observation that the histogram of a sample tends to look similar to the underlying probability distribution function. More precisely, it is equivalent to the observation that the empirical distribution function (the c.d.f. of the histogram) tends to be (even more) similar to the underlying c.d.f. (For i.i.d. samples, this is the content of the GlivenkoCantelli theorem). Figure 6 compares the order statistics of GUE(100) and GUE(10) eigenvalues (computed as numerical averages over 100 random samples) to the inverse c.d.f. for the Wigner semicircle distribution. Even though the Wigner semicircle model of GUE eigenvalues is only exact as $N \to \infty$, it provides a nearly-perfect model for $\overline{\kappa}$ even at $N = 10$ (and remains surprisingly good all the way down to $N = 2$).

In the main text, we make the further simplifying approximation that the distribution of the $\overline{\kappa}_j$ is effectively continuous. For the quantities that we compute, this is equivalent to replacing the empirical distribution function (which is a step function) by the c.d.f. of the Wigner semicircle distribution. So, whereas for any given sample the partial sum of all $\kappa_i > q$ jumps discontinuously when $q = \kappa_i$ for any i, in this approximation it changes smoothly. This accurately models the *average* behavior of partial sums.

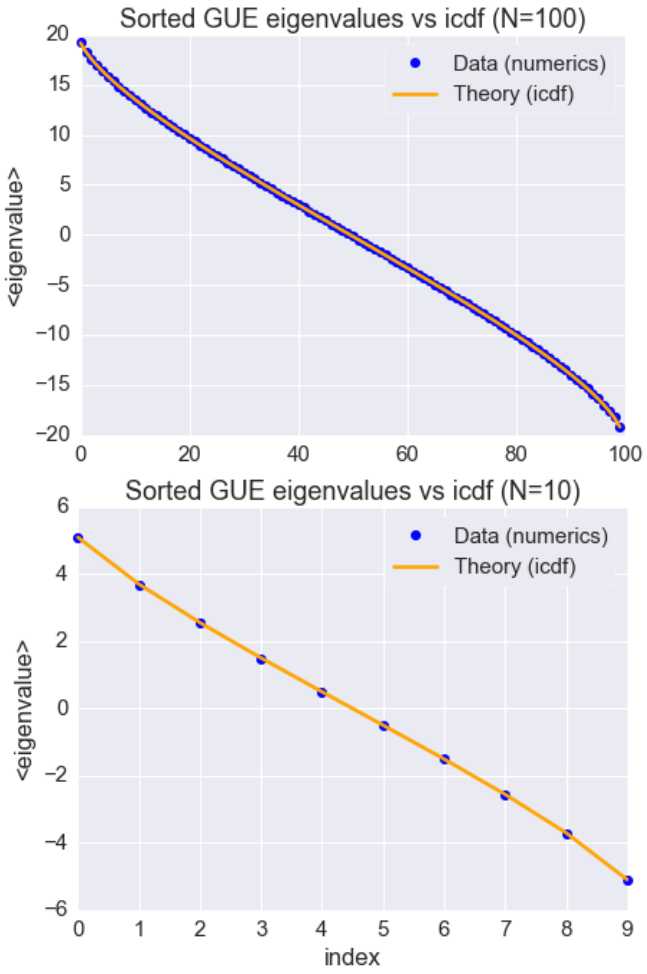

FIG. 6: Order statistics of the GUE(N) eigenvalue distribution are very well approximated by the inverse c.d.f. of the Wigner semicircle distribution. **Top**: Comparison of the order statistics of the GUE(100) distribution (as shown in Fig. 5) to the inverse cdf of the Wigner semicircle distribution for N=100. **Bottom**: Comparison of the order statistics of the GUE(10) distribution to the inverse cdf of the Wigner semicircle distribution for N=10. Agreement in both cases is essentially perfect.

These approximations provide the ansatz that we use in the main text, for the eigenvalues of $\hat{\rho}$, as $\{p_j\} \cup \{\overline{\kappa}_j\}$, where the p_j are $\mathcal{N}(\rho_{jj}, \epsilon^2)$ random variables, and the $\overline{\kappa}_j$ are the (fixed, smoothed) order statistics of a Wigner semicircle distribution.

Another rather technical portion of our derivation is the finding of an (approximate) solution to the equation

$$f(q) = 1 \tag{12}$$

where

$$f(q) = \mathrm{Tr}\left[\mathrm{Trunc}(\hat{\rho} - q\mathbb{1})\right], \tag{13}$$

Given the assumptions and approximations above, and

letting $(x)^+$ denote $\max(x, 0)$,

$$\begin{aligned} f(q) &= \sum_{j=1}^{r}(p_j - q) + \sum_{j=1}^{N}(\overline{\kappa}_j - q)^+ \\ &\approx 1 - rq + \Delta + N\int_{\kappa=q}^{2\epsilon\sqrt{N}}(\kappa - q)\Pr(\kappa)\mathrm{d}\kappa \\ &= 1 - rq + \Delta + \frac{\epsilon}{12\pi}\left[\begin{array}{l}(q^2+8N)\sqrt{-q^2+4N} \\ -12qN\left(\frac{\pi}{2} - \sin^{-1}\left(\frac{q}{2\sqrt{N}}\right)\right)\end{array}\right], \end{aligned} \tag{14}$$

where $\Delta = \sum_{j=1}^{r}\delta_{jj}$ is a $\mathcal{N}(0, r\epsilon^2)$ random variable. We also chose to replace a discrete sum (line 1) with an integral (line 2). This approximation is valid when $N \gg 1$, where (as noted in the main text) we can accurately approximate a discrete collection of closely spaced real numbers by a smooth density or distribution over the real numbers that has approximately the same continuous distribution function (cdf). It is also remarkably accurate in practice.

In yet another approximation, we replace Δ with its average value, which is zero. We could obtain an even more accurate expression by treating the fluctuations in Δ more carefully, but this crude approximation turns out to be quite accurate already.

To solve for $f(q) = 1$, it is necessary to further simplify the complicated expression resulting from the integral. To do so, we assume that ρ_0 is relatively low-rank, so $r \ll N$. In this case, the sum of the positive $\overline{\kappa}_j$ is large compared with r, almost all of them need to be subtracted away, and therefore q is close to $2\epsilon\sqrt{N}$. We therefore replace the complicated expression with its leading order Taylor expansion around $q = 2\epsilon\sqrt{N}$, substitute into $f(q) = 1$, and obtain the equation

$$rq = \frac{4}{15\pi}n^{1/4}\left(2\sqrt{N} - q\right)^{5/2}. \tag{15}$$

This equation is a quintic polynomial, so it has no closed-form solution. However, its roots have a well-defined asymptotic ($n \to \infty$) expansion that becomes accurate quite rapidly (e.g., for $N > 4$):

$$q \approx 2\sqrt{N} - \frac{(240r\pi)^{2/5}}{4}N^{1/10} + \frac{(240r\pi)^{4/5}}{80}N^{-3/10}. \tag{16}$$

This completes our exposition of two significant technical details in the derivation.

II. THE "L"

In the main text, we asserted several properties of matrix elements of δ which lie in the "L". Here, we show how those properties arise. To do so, we observe that that the δ_{jk} in the "L" may be seen as errors which arise due to small unitary perturbations of the true state ρ_0. Writing $\hat{\rho} = U^\dagger\rho_0 U$, where $U = e^{i\epsilon H}$, we have

$$\hat{\rho} \approx \rho_0 + i\epsilon[\rho_0, H] + \mathcal{O}(\epsilon^2)$$

Taking matrix elements, we have

$$\langle j|\hat{\rho}|k\rangle = \langle j|\rho_0|k\rangle + i\epsilon\langle j|[\rho_0, H]|k\rangle$$

In the event $j = k$, then $\langle j|[\rho_0, H]|j\rangle = 0$. If $\langle j|\rho_0|j\rangle = 0$ as well, then $\langle j|\hat{\rho}|j\rangle = 0$. In turn, this implies $\langle j|\delta|j\rangle = 0$. Thus, small unitaries cannot create errors in the diagonal matrix elements unless ρ_0 had support there as well. If $j \neq k$, then $\langle j|\hat{\rho}|k\rangle = i\epsilon\langle j|[\rho_0, H]|k\rangle \neq 0$, in general. This implies $\delta_{jk} \neq 0$, in general. Thus, small unitaries can introduce errors only on off-diagonal elements, or diagonal elements on which ρ_0 has support.

These off-diagonal elements are precisely the "L". They are described by the set $\{\delta_{jk}\ s.t.\ \langle j|\rho_0|j\rangle \neq 0, j \neq k,\ 0 \leq j, k \leq d-1\}$.

Now, because the δ_{jk} in this set arise due to small unitary perturbations of ρ_0, it follows that they cannot change the spectrum of $\hat{\rho}$, at least to first order in ϵ. As such, they may fluctuate in an unconstrained manner, and because those fluctuations are Gaussian in nature, $\langle\lambda_{jk}\rangle = 1$. There are $2rd - r(r+1)$ such elements, giving a total expected value $\langle\lambda\rangle_L = 2rd - r(r+1)$.

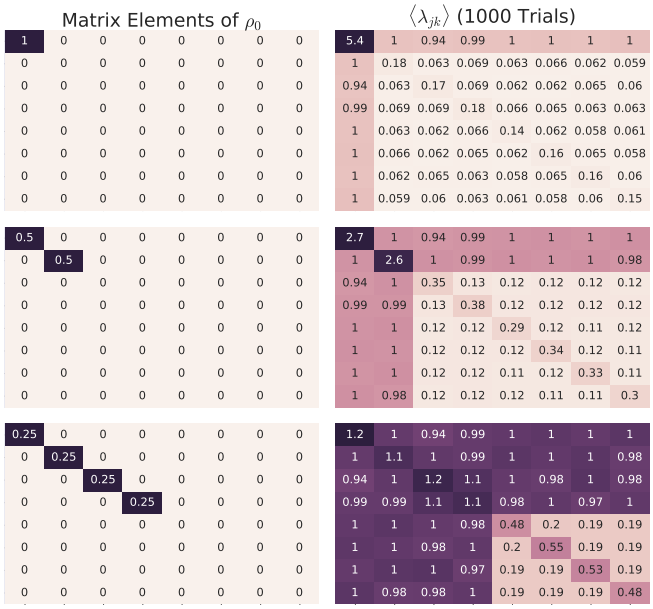

FIG. 7: Contributions of each matrix element of $\hat{\rho}_{\mathrm{MLE}}$ to $\langle\lambda\rangle$, for three different ρ_0, as computed in numerical simulations for $d = 8$. **Left**: The matrix elements of ρ_0. **Right**: The average scaled mean-squared-errors of individual matrix elements, $\langle\lambda_{jk}\rangle$, as defined in Eq. 6. **Discussion**: While the Wilks Theorem predicts that each matrix element should contribute ≈ 1, only those in the "L" (see Fig. 3) do so; those in the "kite" behave quite differently and (generally) contribute less to $\langle\lambda\rangle$.

III. BREAKDOWN OF OUR MODEL IN THE CASE OF HETERODYNE TOMOGRAPHY

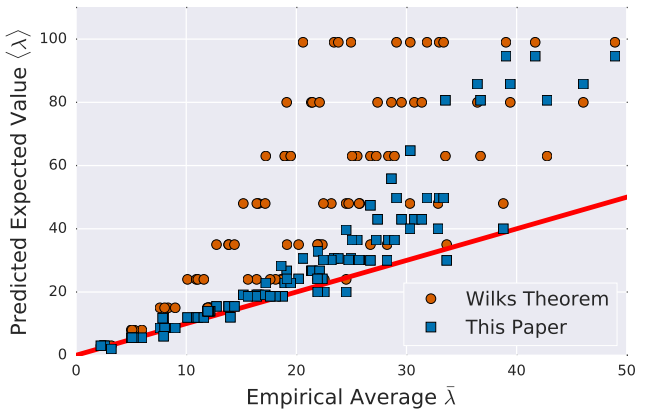

FIG. 8: Preliminary results comparing our theory (Eq. 10) and the Wilks theorem to the empirical expectation value of λ for heterodyne tomography in larger dimension (d) than shown in Fig. 4. For these larger Hilbert space dimensions, our theory begins to fail, for several reasons that we detail in this section. Most importantly, even at $N = 10^5$ samples, we observe (see Fig. 9) that the empirical λ has not yet asymptoted – it is still increasing (very slowly) with N. Thus, simulations of heterodyne tomography at the N values considered so far *underperform* our theory. We conjecture that our theory will become more and more accurate as $N \to \infty$, but that (because certain important events are very, very rare), this convergence will only occur at very large $N = e^{O(d)}$. A secondary effect is the anisotropy of the Fisher information (see Fig. 10), which causes smaller but significant deviations from our theory.

Although Figure 4 shows reasonable agreement between our expression for $\langle\lambda\rangle$ and the numerically computed average $\bar{\lambda}$, the discrepancies are sufficiently large that it behooves us to examine why this may be the case. In this section, we show how finite-sample effects, anisotropy of the Fisher information, and the behavior of the loglikelihood ratio statistic for Poisson-distributed data could explain some of the discrepancy we see in preliminary results for reconstructions in higher dimension d (see Fig. 8).

We start by plotting $\bar{\lambda}$ as a function of the number of heterodyne counts N. Both the Wilks Theorem and our result are derived in an asymptotic limit $N \to \infty$; for finite but large N, our results may be invalid. As seen in Figure 9, $\bar{\lambda}$ may be growing, even for $N \sim 10^5$, depending on ρ_0 and d.

Thus, part of the discrepancy between our expression and the numerical results is simply a finite-sample effect. Next, we check whether the Fisher information is actually isotropic - a key ingredient in our derivation. To estimate the Fisher information, we use a numerical average of 100 observed informations (Hessians of the loglikelihood function), $\bar{H}$. In Figure 10, we plot, for

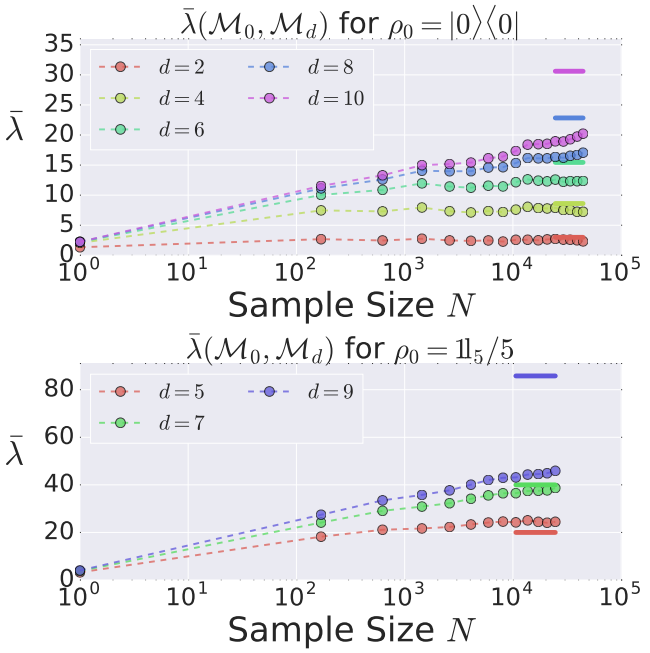

FIG. 9: The empirical average $\bar{\lambda}$ may have achieved its asymptotic value, or is still growing, depending on the true state ρ_0 and the model dimension d. Solid lines indicate the value of our formula for the asymptotic expected value, given in Equation (10).

several $\bar{H}$, their condition number κ, which is the ratio of the largest eigenvalue to the smallest. Importantly, $\kappa = 1$ if, and only if, all the eigenvalues are the same (i.e., the Fisher information is isotropic). The condition number appears to grow with d, meaning the anisotropy is becoming more severe.

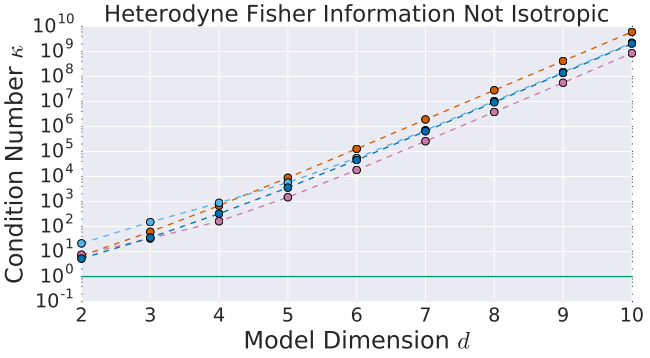

FIG. 10: The condition number κ of the estimated Fisher Information matrix grows with the model dimension, indicating an increase in anisotropy. The dashed lines indicate different true states ρ_0, and the solid line is $\kappa = 1$.

If the anisotropy is so severe, why is our model reasonably accurate? To answer this question, we compute the scaled mean squared errors $\langle\lambda_{jk}\rangle$ (defined in Equation (7)) for the heterodyne estimates, and compare them to those of our model. As shown in Figure 11, we see

qualitative agreement - matrix elements in the coherent "L" contribute about 1 unit of expected value, while matrix elements in the kernel of ρ_0 contribute less than our model predicts.

Looking at each $\langle \lambda_{jk} \rangle$ as a function of the sample size N and (j, k), we see in Figure 12 that each contribution grows starting from zero, and some elements (notably those in the "L") rapidly achieve their asymptotic contribution of 1. Those within the "kite", however, grow much more slowly. Further, the anisotropic nature of the Fisher information may be such that the "kite" elements will never attain the values predicted by our theory.

Finally, we conjecture, based on strong theoretical support from an an analysis of the loglikelihood ratio statistic for classical Poisson distributions (Appendix IV) that certain matrix elements — namely, those in the coherent "L" with support on higher-energy photon states $|n\rangle$ — do not reach their asymptotic contribution level until *extremely* large sample sizes N. If this conjecture is true, then at least some of our model's failure to match the observed values occurs simply because the empirical datasets are not yet "asymptotically" large, meaning individual $\langle \lambda_{jk} \rangle$ are not yet asymptotic, causing $\langle \lambda \rangle$ to not have achieved its asymptotic value either.

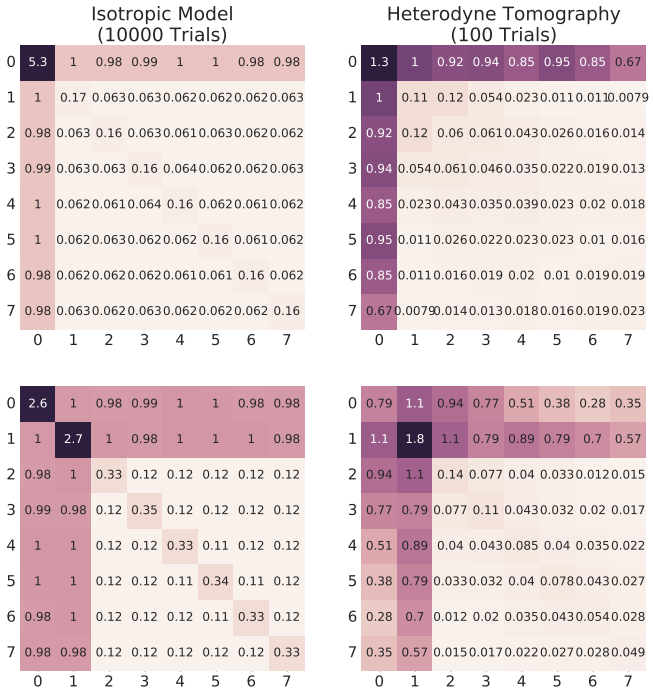

FIG. 11: The scaled mean squared errors $\langle \lambda_{jk} \rangle$ assuming an isotropic Fisher information (left), and for heterodyne tomography (right). Top: $\rho_0 = |0\rangle\langle 0|$, Bottom: $\rho_0 = \mathcal{I}_2/2$. Qualitatively, the behavior of the contributions is the same, though there are quantitative differences, particularly within the kite.

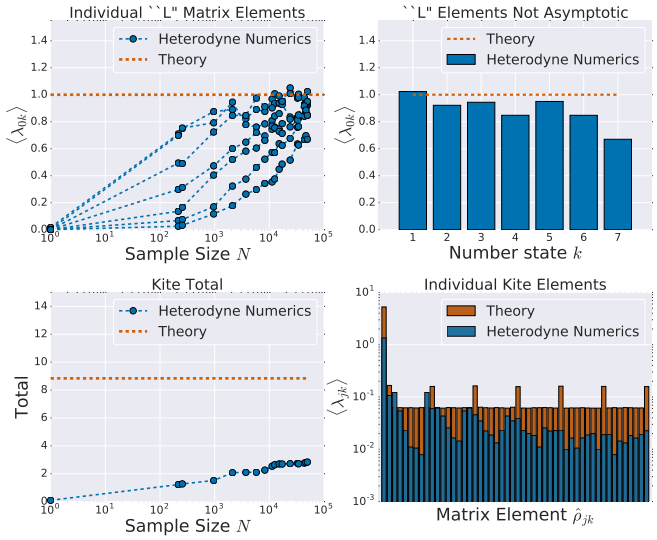

FIG. 12: Examining how our model for $\langle \lambda \rangle$ disagrees with simulated heterodyne experiments. We take $\rho_0 = |0\rangle\langle 0|$ and $d = 8$. Top Left: The fluctuations in the "L" $\langle \lambda_{0k} \rangle$ as a function of sample size N. Top Right: At the largest N studied, the fluctuations in the "L" from coherences with different number states. For the higher number states, $\langle \lambda_{0k} \rangle$ is substantially less than 1. Bottom Left: The total from the "kite" versus N. It is clear the total is still growing. Bottom Right: The individual "kite" elements $\langle \lambda_{jk} \rangle$; most are small compared to the model.

IV. THE POISSON DIP/BUMP PHENOMENON

Here we show how, $\langle \lambda \rangle$ depends strongly on the expected number of counts for Poisson-distributed data. Suppose we are given count data which is Poisson distributed, so that the probability of observing K counts is

$$\Pr(K) = \frac{e^{-\theta_0} \theta_0^K}{K!} \tag{17}$$

where θ_0 is the rate parameter, and $\langle K \rangle = \theta_0$. Suppose further we observed C counts. The likelihood function for the rate parameter θ is simply

$$\mathcal{L}(\theta) = \Pr(C|\theta) = \frac{e^{-\theta} \theta^C}{C!} \tag{18}$$

We will compare the models

$$\mathcal{M}_0 : \theta = \theta_0 \qquad \mathcal{M}_1 : \theta \in [0, \infty)$$

using the loglikelihood ratio statistic. By definition, we have

$$\lambda(\theta_0, \hat{\theta}) = -2 \log \left(\frac{\mathcal{L}(\theta_0)}{\mathcal{L}(\hat{\theta})} \right) \tag{19}$$

with $\hat{\theta}$ as the MLE of the likelihood function in Equation (18). Doing the maximization, we find $\hat{\theta} = C$. Plugging

in this MLE and writing out the likelihood function yields

$$\lambda(\theta_0, \hat{\theta}) = -2 \log \left(e^{C-\theta_0} \left(\frac{\theta_0}{C} \right)^C \right)$$
$$= -2 \left(C - \theta_0 + C \log(\theta_0) - C \log(C) \right) \quad (20)$$

According to the Wilks Theorem, $\lambda \sim \chi^2_1$, since $\mathcal{M}_1$ is a one-dimensional model. Accordingly, $\langle \lambda \rangle = 1$. Let's compute $\langle \lambda \rangle$ using the expression above:

$$\langle \lambda \rangle = -2(\langle C \rangle - \theta_0 + \langle C \rangle \log(\theta_0) - \langle C \log(C) \rangle) \quad (21)$$

Using the fact $\langle C \rangle = \theta_0$, the above equation simplifies some:

$$\langle \lambda \rangle = -2(\theta_0 \log(\theta_0) - \langle C \log(C) \rangle) \quad (22)$$

We compute the expected value of $C \log(C)$ as

$$\langle C \log(C) \rangle = \sum_{C=0}^{\infty} C \log(C) \Pr(C) = \sum_{C=0}^{\infty} C \log(C) \frac{e^{-\theta_0} \theta_0^C}{C!}$$
$$= \theta_0 e^{-\theta_0} \sum_{C=0}^{\infty} \frac{\log(C+1) \theta_0^C}{C!}$$

(In general, for a random variable X, $\langle f(X) \rangle \neq f(\langle X \rangle)$, unless f is *linear*.) Putting this expression into equation (22) gives

$$\langle \lambda \rangle = -2 \left(\theta_0 \log(\theta_0) - \theta_0 e^{-\theta_0} \sum_{C=0}^{\infty} \frac{\log(C+1) \theta_0^C}{C!} \right) \quad (23)$$

In Figure 13, we plot $\langle \lambda \rangle$ as a function of θ_0, where we truncate the sum at 150 terms. It is clear $\langle \lambda \rangle$ depends strongly on θ_0. Curiously, we see that as $\theta_0 \to 0^+$, $\langle \lambda \rangle \to 0$. Why is this the case? When the number of counts is 0, the parameter θ is not a "parameter" at all, since it *cannot fluctuate*. It has to be set to 0! Phrased another way, the Wilks Theorem establishes a conversion factor between fluctuating parameters and $\langle \lambda \rangle$. When parameters do not fluctuate very much, they do not contribute much to the statistic. Taken to the extreme, if a parameter cannot fluctuate at all, and the value it is pinned to happens to be the true value, then its scaled mean squared error is 0, so $\langle \lambda \rangle = 0$.

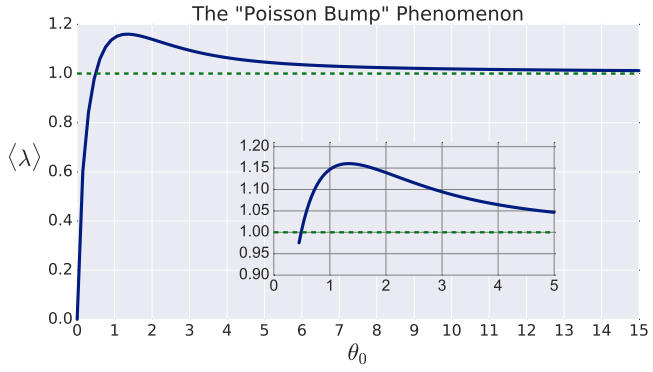

FIG. 13: For Poisson-distributed data, the behavior of the loglikelihood ratio statistic depends strongly on the expected number of counts θ_0. The single parameter in the Poisson model will be fully contributing – i.e., $\langle \lambda \rangle = 1$– only when the expected number of counts is large. When $\theta_0 << 1$, the parameter is "under-fluctuating", as $\langle \lambda \rangle << 1$. As shown in the inset, there exists a regime ($1 \lesssim \theta_0 \lesssim 5$) where the parameter is "super-fluctuating" - i.e., $1 \lesssim \langle \lambda \rangle$.

In heterodyne tomography, if we build a model with support in a D-dimensional Hilbert space $\mathcal{M}_D$, but the true state which generated the data has support on a $d < D$-dimensional Hilbert space, then the probability of observing any α values with radius $r > \sqrt{d}$ drops off as e^{-d}. (Recall $\Pr(\alpha) = \langle \alpha | \rho | \alpha \rangle \propto e^{-|\alpha|^2}$.) With very few "counts" out in this region in phase space, the additional $\mathcal{O}(D^2 - d^2)$ parameters we added do not fluctuate very much, and these parameters are "under-fluctuating". Conversely, in order for a parameter in $\mathcal{M}_D$ which has support on a high-energy Fock state $|n\rangle$, with $n > d$, to contribute to the LLRS, we may have to draw $\mathcal{O}(e^n)$ samples before its contribution is 1. Thus, in order for the parameters with support in $|D\rangle$ to contribute, we may require $\mathcal{O}(e^D)$ samples. In looking at Figure 13, we see that, if $\theta_0 \sim 10$, we have $\langle \lambda \rangle \sim 1$. Thus, when building an estimate in $\mathcal{M}_{10}$, we may need approximately $10e^{10} \sim 2.2 \times 10^5$ coherent state outcomes before the scaled mean squared error of the matrix element $(10, 10)$ in the estimate is 1, which is consistent with the plots shown in Figure 12.