# 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*
*Center for Quantum Information and Control (CQuIC), University of New Mexico*
(Dated: July 29, 2017)

Quantum state tomography on a $d$-dimensional system demands resources – for example, experimental repetitions, or data post-processing – that grow rapidly with $d$. They may be reduced by using model selection to tailor the number of parameters in the model (i.e., the size of the density matrix). Most model selection methods typically rely on a test statistic and a null theory that describes its behavior when two models are equally good. Here, we use the loglikelihood ratio. Because of the positivity constraint $\rho \geq 0$, quantum state space does not generally satisfy local asymptotic normality, meaning the classical null theory for the loglikelihood ratio (the Wilks theorem) should not be used. Thus, understanding and quantifying how positivity constrains the null theory of this test statistic is vital for enabling model selection in state tomography. Here, we define a new generalization of local asymptotic normality, metric-projected local asymptotic normality, show that quantum state space satisfies it, and derive a replacement for the Wilks theorem. In addition to enabling reliable model selection, our results shed more light on the qualitative effects of the positivity constraint on state tomography.

Determining the quantum state $\rho_0$ produced by a specific preparation procedure for a quantum system is a problem almost as old as quantum mechanics itself [1, 2]. This task, known as *quantum state tomography* [3], is not only useful in its own right (diagnosing and detecting errors in state preparation), but is also used in other characterization protocols including entanglement verification [4–6] and process tomography [7]. A typical state tomography protocol proceeds as follows: many copies of $\rho_0$ 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 [8, 9], where the data are used to fit the parameters of a statistical model. In state tomography, the parameter is $\rho$, and the model is the set of all possible density matrices on a Hilbert space $\mathcal{H}$ (equipped with the Born rule). However, we don't always know what model to use. It is not always *a priori* obvious what $\mathcal{H}$ or its dimension is; examples include optical modes [10–14] and leakage levels in AMO and superconducting [15, 16] qubits. In such situations, we seek to let the data itself determine which of many candidate Hilbert spaces is best suited for reconstructing $\rho_0$.

Choosing an appropriate Hilbert space on the fly is an instance of a general statistical problem called *model selection*. Although model selection has been thoroughly explored in classical statistics [17], its application to quantum tomography encounters some obstacles. They stem from the fact that quantum states (and therefore, estimates of them) must satisfy a *positivity constraint* ($\rho \geq 0$). A similar constraint (complete positivity) applies to process tomography. The impact of positivity constraints on state and process tomography is an active area of research [18–21], and its implications for model selection have also been considered [22–28]. In this paper, we address a specific question at the heart of this matter: *How does the loglikelihood ratio statistic used in many model selection protocols, including (but not limited to) information criteria such as Akaike's AIC [29], behave in the presence of the positivity constraint $\rho \geq 0$?*

We begin in Section I by introducing the loglikelihood ratio statistic $\lambda$, and outline how it can be used to choose a Hilbert space. In Section II, we show how and why the classical null theory for its behavior, the Wilks theorem, falls apart in the presence of the positivity constraint, because quantum state space does not generally satisfy *local asymptotic normality* (LAN). We define a new generalization of LAN, *metric-projected local asymptotic normality*, in Section III; this generalization explicitly accounts for the positivity constraint, and is satisfied by quantum state space. Using this generalization, we derive a closed-form approximation for $\lambda$'s expected value in Section IV, thereby providing a replacement for the Wilks theorem that *is* applicable in state tomography. Finally, we test the validity of our approximate theory under the assumptions used in its derivation (Section V A), and under harsh conditions by comparing it to numerical results for the realistic problem of optical heterodyne tomography (Section V B).

## I. INTRODUCTION - STATISTICAL MODEL SELECTION

Discussing model selection for state tomography requires introducing some basic statistical notions. A *model* is a parameterized family of probability distributions over some data $D$, usually denoted as $\Pr_{\boldsymbol{\theta}}(D)$, where $\boldsymbol{\theta}$ are the *parameters* of the model. In state tomography, the data are the observed outcomes of a measurement of a positive operator-valued measure (POVM) $\{E_j\}$, the parameters are a quantum state $\rho$, and the probability of observing outcome "$j$" is given by the Born rule: $p_j = \mathrm{Tr}(\rho E_j)$ [30]. In this paper, a model is a set of density matrices, and a state $\rho$ is a particular choice

of the model's parameters.

Suppose we have data $D$ obtained from an unknown state $\rho_0$, and two candidate models $\mathcal{M}_1, \mathcal{M}_2$ that could be used to reconstruct it. Many of the known methods for choosing between them (i.e., model selection) involve quantifying how well each model fits the data by its *likelihood*. The likelihood of a simple hypothesis $\rho$ is defined as $\mathcal{L}(\rho) = \Pr(\mathrm{D}|\rho)$. Models, however, are *composite* hypotheses, comprising many possible values of $\rho$. A canonical way to define model $\mathcal{M}$'s likelihood is via the general method of *maximum likelihood* (ML), by maximizing $\mathcal{L}(\rho)$ over $\rho \in \mathcal{M}$. In practice, the maximization is usually done explicitly to find an ML estimate $\hat{\rho}_{\mathrm{ML},\mathcal{M}}$ [31–33] of $\mathcal{M}$'s parameters, and then $\mathcal{L}(\mathcal{M}) = \mathcal{L}(\hat{\rho}_{\mathrm{ML},\mathcal{M}})$. (Although it is common to refer to $\hat{\rho}_{\mathrm{ML}}$ without specifying the model over which $\mathcal{L}$ was maximized, we list the model explicitly in the subscript because in this paper we are frequently using many different models!)

Then, the models can be compared using the *loglikelihood ratio statistic* [27, 33, 34]:

$$\lambda(\mathcal{M}_1, \mathcal{M}_2) \equiv -2 \log \left( \frac{\mathcal{L}(\mathcal{M}_1)}{\mathcal{L}(\mathcal{M}_2)} \right) \tag{1}$$

$$= -2 \log \left( \frac{\mathcal{L}(\hat{\rho}_{\mathrm{ML},\mathcal{M}_1})}{\mathcal{L}(\hat{\rho}_{\mathrm{ML},\mathcal{M}_2})} \right) \tag{2}$$

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

All else being equal, a positive $\lambda$ favors $\mathcal{M}_2$ – i.e., the model with the higher likelihood is more plausible, because it fits the data better. However, all else is rarely equal. If both models are equally valid – e.g., they both contain $\rho_0$ – but $\mathcal{M}_2$ has more parameters, then $\mathcal{M}_2$ will very probably fit the data better. Models with more adjustable parameters do a better job of fitting *noise* (e.g., finite sample fluctuations) in the data. This becomes strictly true when the models are *nested*, so that $\mathcal{M}_1 \subset \mathcal{M}_2$. In this case, the likelihood of $\mathcal{M}_2$ is at least as high as that of $\mathcal{M}_1$; not only is $\lambda \geq 0$, but almost surely $\lambda > 0$.

Remarkably, the same effect also makes $\mathcal{M}_2$'s fit less accurate (almost surely), because the fit incorporates more of the noise in the data. These two effects constitute *overfitting*, which can be summed up as "Extra parameters make the fit look better, but perform worse." An overfitted model would fit *current* data extremely well, but would fail to accurately predict *future* data well. This provides strong motivation to correct for overfitting by penalizing or handicapping larger models, to prevent them from being chosen over smaller models that are no less valid, and may even yield better estimates in practice [29].

For this reason, any model selection method/criterion 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 $\rho_0 \in \mathcal{M}$. 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.

The null theory can be used to formulate a *decision rule* for choosing between models. If we know how the statistic behaves when both models are equally valid, then we can evaluate the observed value of the statistic under the null theory. If the observed value is very improbable under the null theory, then that constitutes evidence against the smaller model, and justifies rejecting it. On the other hand, if the observed value is *consistent* with the null theory, there is no reason to reject the smaller model.

The standard null theory for $\lambda$ is the *Wilks theorem* [35]. It relies on *local asymptotic normality* (LAN) [36, 37]. LAN is a property of a statistical model $\mathcal{M}$. If $\mathcal{M}$ satisfies LAN, then as $N_{\mathrm{samples}} \to \infty$:

- The ML estimate $\hat{\rho}_{\mathrm{ML},\mathcal{M}}$ is normally distributed around $\rho_0$ with covariance matrix $\mathcal{I}^{-1}$:

$$\Pr(\hat{\rho}_{\mathrm{ML},\mathcal{M}}) \propto \exp\left[-\mathrm{Tr}[(\rho_0 - \hat{\rho}_{\mathrm{ML},\mathcal{M}})\mathcal{I}(\rho_0 - \hat{\rho}_{\mathrm{ML},\mathcal{M}})]/2\right]. \tag{4}$$

- The likelihood function in a neighborhood of $\hat{\rho}_{\mathrm{ML},\mathcal{M}}$ is locally Gaussian with Hessian $\mathcal{I}$:

$$\mathcal{L}(\rho) \propto \exp\left[-\mathrm{Tr}[(\rho - \hat{\rho}_{\mathrm{ML},\mathcal{M}})\mathcal{I}(\rho - \hat{\rho}_{\mathrm{ML},\mathcal{M}})]/2\right]. \tag{5}$$

Here, $\mathcal{I}$ is the (classical) *Fisher information matrix* associated with the POVM. It quantifies how much information the data carry about a parameter in the model. (Note that in expressions involving $\mathcal{I}$, states $\rho$ are treated as vectors in state space, and $\mathcal{I}$ is a matrix or 2-index tensor acting on that state space.) Generally speaking, the Fisher information depends strongly on $\rho_0$ and the POVM being measured (see Figure 5). However, as the next section explains, it's common in classical statistics to assume *regularity conditions* that make it possible to change to *Fisher-adjusted* coordinates in which $\mathcal{I}$ is isotropic (i.e., $\mathcal{I} \propto \mathbb{1}$).

Most statistical models satisfy LAN. When LAN holds *and* $N_{\mathrm{samples}}$ is large enough to reach the "asymptotic" regime, we can invoke the Wilks theorem to determine the behavior of $\lambda$. This theorem says that under suitable regularity conditions, if $\rho_0 \in \mathcal{M}_1 \subset \mathcal{M}_2$, where $\mathcal{M}_2$ has $K$ more parameters than $\mathcal{M}_1$, then $\lambda$ is a $\chi^2_K$ random variable. This is a complete null theory for $\lambda$ (under the specified conditions), and implies that $\langle\lambda\rangle = K$ and $(\Delta\lambda)^2 = 2K$.

Therefore, in the "Wilks regime", a simple criterion for model selection would be to compare the observed value of $\lambda$ to $\lambda_{\mathrm{thresh}} = \langle\lambda\rangle + n\Delta\lambda$, for some $n \approx 1$, and reject the smaller model if $\lambda > \lambda_{\mathrm{thresh}}$. While model selection rules can be more subtle and complex than this [29, 38–40], they usually take the general form of a threshold in which $\langle\lambda\rangle$ plays a key role. Rather than attempting to define a specific rule, our purpose in this paper is to understand the behavior of $\langle\lambda\rangle$ and derive an approximate expression for it in the context of state tomography.

The first step in doing so is to explain how and why the Wilks theorem breaks down in that context.

## II. QUANTUM STATE TOMOGRAPHY AND THE BREAKDOWN OF THE WILKS THEOREM

Quantum state tomography typically begins with $N_{\text{samples}}$ independently and identically prepared quantum systems – i.e., $N_{\text{samples}}$ copies of an unknown state $\rho_0$. Each copy is measured, and without loss of generality we can assume that each measurement is described by the same positive operator-valued measure (POVM). A POVM is a collection of positive operators $\{E_j\}$ summing to $\mathbb{1}$, and the probability of outcome "$j$" is given by $\text{Tr}(\rho_0 E_j)$. The results of all $N_{\text{samples}}$ measurements constitute data, represented as a record of the frequencies of the possible outcomes $\{n_j\}$, where $n_j$ is the number of times "$j$" was observed, and $\sum_j n_j = N_{\text{samples}}$. Finally, this data is processed through some *estimator* to yield an estimate of $\rho_0$, denoted $\hat{\rho}$ . A variety of estimators have been proposed [31–33, 41–44], but they are not our concern here. However, since we *are* concerned with computing the likelihood of a model $\mathcal{M}$, which is defined as the likelihood of the most likely $\rho \in \mathcal{M}$, we will make extensive use of the *maximum likelihood* (ML) estimator. This should not be taken as advocacy for the ML estimator; it is only a convenient way to find the maximum of $\mathcal{L}$ over $\mathcal{M}$, and once a model is chosen, a different estimator could be used.

The likelihood $\mathcal{L}(\rho)$ is

$$\mathcal{L}(\rho) = \prod_j \text{Tr}(\rho E_j)^{n_j}, \tag{6}$$

and $\hat{\rho}_{\text{ML},\mathcal{M}}$ is the solution to the optimization problem

$$\hat{\rho}_{\text{ML},\mathcal{M}} = \underset{\rho\in\mathcal{M}}{\text{argmax}}\ \mathcal{L}(\rho) = \underset{\rho\in\mathcal{M}}{\text{argmin}}\ \left(-\log[\mathcal{L}(\rho)]\right). \tag{7}$$

In state tomography, $\mathcal{M}$ is almost always the set of all density matrices over a Hilbert space $\mathcal{H}$:

$$\mathcal{M}_{\mathcal{H}} = \{\rho \mid \rho \in \mathcal{B}(\mathcal{H}),\ \text{Tr}(\rho) = 1,\ \rho \geq 0\}, \tag{8}$$

where $\mathcal{B}(\mathcal{H})$ is the space of bounded linear operators on $\mathcal{H}$. To determine $\hat{\rho}_{\text{ML},\mathcal{M}}$, we can use the following facts: (a) $\mathcal{M}_{\mathcal{H}}$ is a convex set, and (b) $\hat{\rho}_{\text{ML},\mathcal{M}}$ minimizes the value of $-\log[\mathcal{L}(\rho)]$, which is a convex function. Because $\hat{\rho}_{\text{ML},\mathcal{M}}$ is the solution to minimizing a convex function over a convex set, it can be found efficiently via any of several algorithms for convex optimization [45].

Usually, $\mathcal{H}$ is taken for granted or chosen by fiat. In this paper, we will consider a nested family of different Hilbert spaces, indexed by their dimension $d$: $\mathcal{H}_1 \subset \cdots \subset \mathcal{H}_d \subset \mathcal{H}_{d+1} \subset \cdots$. The models we consider are therefore given by:

$$\mathcal{M}_d \equiv \mathcal{M}_{\mathcal{H}_d} = \{\rho \mid \rho \in \mathcal{B}(\mathcal{H}_d),\ \text{Tr}(\rho) = 1,\ \rho \geq 0\}. \tag{9}$$

For notational brevity, we will use $\hat{\rho}_{\text{ML},d}$ to denote the ML estimate over $\mathcal{M}_d$. To select between these models, we need to determine whether one model (say, $\mathcal{M}_{d+1}$) is "better" than another (say, $\mathcal{M}_d$). To evaluate which is better, we typically compute the likelihood of each model, and then use $\lambda(\mathcal{M}_d, \mathcal{M}_{d+1})$ to choose between them. As mentioned in the previous section, this requires having a *null theory* for $\lambda$ that describes its behavior when $\rho_0 \in \mathcal{M}_d, \mathcal{M}_{d+1}$.

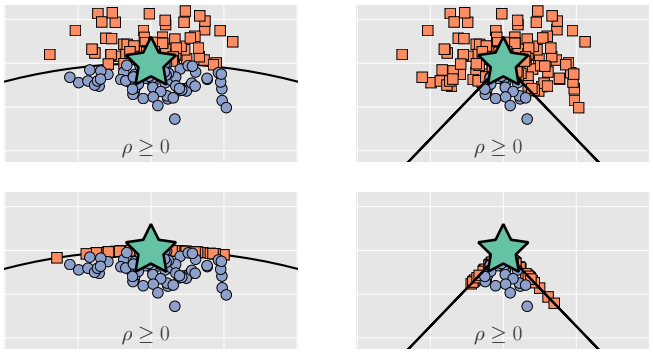


FIG. 1. Two views of the impact of the boundary on maximum likelihood (ML) tomography of a qutrit state $\rho_0$ (star). **Top**: Without the positivity constraint, some ML estimates (orange squares) are not valid estimates of a quantum state, because are not positive semidefinite. However, some ML estimates (blue circles) are. **Bottom**: Imposing the positivity constraint forces the (previously negative) ML estimates to "pile up" on the boundary of state space; $\Pr(\hat{\rho}_{\text{ML},\mathcal{M}})$ is not Gaussian, and LAN is not satisfied. In turn, the assumptions necessary to invoke the Wilks theorem are not satisfied either.

The Wilks theorem, which is the classical null theory for $\lambda$, relies on local asymptotic normality (LAN). In classical statistics, it is common to assume that boundaries are not relevant, either because the models of interest have none, or because the true parameter values $\rho_0$ lie far away from them. In the absence of boundaries, and in the asymptotic limit where the curvature of the Fisher metric is also negligible, many calculations can be simplified by changing to *Fisher-adjusted* coordinates in which the Fisher information is isotropic (i.e., $\mathcal{I} \propto \mathbb{1}$). Under these assumptions and simplifications, the Wilks theorem can be derived.

However, the models used in state tomography *do* have boundaries. Given a model $\mathcal{M}_d$, its boundar is the set of rank-deficient states within it. When $\rho_0 \in \mathcal{M}_d$ and is full rank, LAN will hold – which is to say that, asymptotically, the boundary can indeed be ignored. But when $\rho_0$ is rank-deficient, it lies *on* the boundary of the model. LAN is not satisfied, because positivity constrains $\hat{\rho}_{\text{ML},d}$, and so $\Pr(\hat{\rho}_{\text{ML},d})$ is not Gaussian (see Figure 1). The Wilks theorem does not apply in this case. (See Figure 2.) Moreover, this is the relevant situation for our analysis, because *even if $\rho_0$ is full-rank in $\mathcal{M}_d$, it must be rank-deficient in $\mathcal{M}_{d+1}$*. So we require a replacement for the Wilks theorem, a null theory for $\lambda$ when $\rho_0$ is rank-deficient.

The next section of this paper is devoted to developing a framework for deriving such a replacement for the Wilks theorem. This framework is a new generalization of LAN, called metric-projected local asymptotic normality (MP-

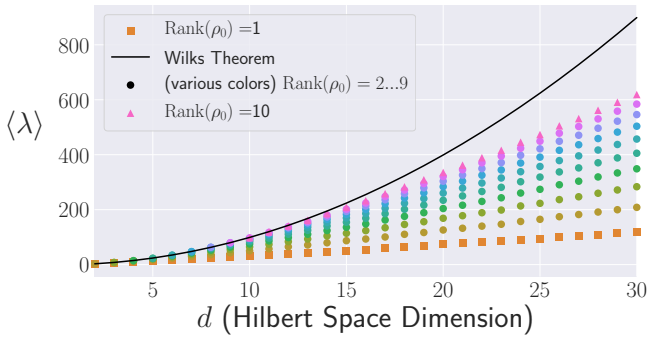


FIG. 2. Comparison of the Wilks theorem prediction for the expected value of the loglikelihood ratio statistic $\langle\lambda\rangle$ to numerical data for states $\rho_0$ with ranks $r = 1, \ldots, 10$. The Wilks theorem fails badly for low-rank states; our main result (Equation 26) gives a replacement that works correctly (see Figure 9).

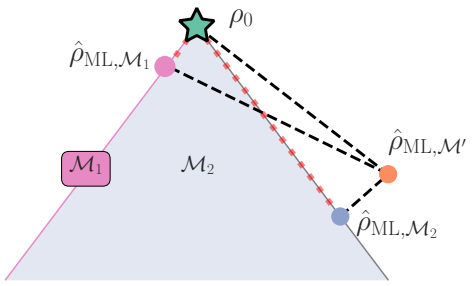


FIG. 3. For any model $\mathcal{M}_k$, $\lambda(\rho_0, \mathcal{M}_k)$ is the difference between the distance from $\rho_0$ to $\hat{\rho}_{\mathrm{ML},\mathcal{M}'}$ and that from $\hat{\rho}_{\mathrm{ML},\mathcal{M}_k}$ to $\hat{\rho}_{\mathrm{ML},\mathcal{M}'}$ (black lines). When MP-LAN holds, $\lambda$ is equal, in the asymptotic limit, to the distance from $\hat{\rho}_{\mathrm{ML},\mathcal{M}_k}$ to $\rho_0$ (red lines), because $\rho_0, \hat{\rho}_{\mathrm{ML},\mathcal{M}'}$, and $\hat{\rho}_{\mathrm{ML},\mathcal{M}_k}$ form a right triangle. This is also true for state spaces with *curved* boundaries because asymptotically, the local state space converges to the solid tangent cone, whose boundaries are always *flat*.

LAN). (For other generalizations of LAN, see [46, 47].) Like LAN, MP-LAN is a property that a statistical model may satisfy. Unlike LAN, MP-LAN is satisfied by quantum state space. For any model which satisfies MP-LAN (quantum or classical), we compute an exact expression for $\lambda$, a necessary building block in our replacement for the Wilks theorem. In Section IV, we show that the models $\mathcal{M}_d$ satisfy MP-LAN, and derive an approximation for $\langle\lambda\rangle$ (Equation (26), on page 11). Section V A compares our theory to numerical results, and Section V B applies our theory to the problem of heterodyne tomography.

## III. GENERALIZING LAN TO DEAL WITH CONSTRAINTS

In this section, we develop a framework that will allow us to derive a replacement for the Wilks theorem valid for rank-deficient $\rho_0$. To do so, we define a generalization of LAN in the presence of boundaries, which we call *metric-projected local asymptotic normality*.

### A. Definitions; Overview of Results

The main definitions and results required for the remainder of the paper are presented in this subsection. Technical details and proofs are presented in the next subsection.

**Definition 1** (Metric-Projected Local Asymptotic Normality (MP-LAN))**.** *A model $\mathcal{M}$ satisfies MP-LAN if, and only if, $\mathcal{M}$ is a convex subset of a model $\mathcal{M}'$ that satisfies LAN.*

Although the definition of MP-LAN is rather short, properties of models which satisfy it are of importance for model selection in state tomography. These properties follow from the fact that, as $N_{\mathrm{samples}} \to \infty$, to understand the behavior of $\hat{\rho}_{\mathrm{ML},\mathcal{M}}$ and $\lambda$, it's sufficient to consider their behavior in an arbitrarily small region of $\mathcal{M}$ around $\rho_0$.

Models which satisfy MP-LAN have the following *asymptotic* properties.

- The local state space converges to the *solid tangent cone* at $\rho_0$, denoted $T(\rho_0)$.
- The ML estimate $\hat{\rho}_{\mathrm{ML},\mathcal{M}}$ is given by the *metric projection* of $\hat{\rho}_{\mathrm{ML},\mathcal{M}'}$ onto $T(\rho_0)$:

$$\hat{\rho}_{\mathrm{ML},\mathcal{M}} = \operatorname*{argmin}_{\sigma \in T(\rho_0)} (\sigma - \hat{\rho}_{\mathrm{ML},\mathcal{M}'}) \mathcal{I} (\sigma - \hat{\rho}_{\mathrm{ML},\mathcal{M}'}). \tag{10}$$

  (We first encountered the term "metric projection" in the convex optimization literature [48, 49], and inspires our choice for the acronym "MP-LAN". However, it should be noted that in the problem setting considered in those references, $\mathcal{I} = \mathbb{1}$.)
- The loglikelihood ratio $\lambda(\rho_0, \mathcal{M})$, defined as

$$\lambda(\rho_0, \mathcal{M}) = -2 \log \left( \frac{\mathcal{L}(\rho_0)}{\max_{\rho \in \mathcal{M}} \mathcal{L}(\rho)} \right), \tag{11}$$

  takes the following simple form

$$\lambda(\rho_0, \mathcal{M}) = \mathrm{Tr}[(\rho_0 - \hat{\rho}_{\mathrm{ML},\mathcal{M}}) \mathcal{I} (\rho_0 - \hat{\rho}_{\mathrm{ML},\mathcal{M}})]. \tag{12}$$

  (This property is non-trivial; see Figure 3.)

Just because $\mathcal{M}$ satisfies MP-LAN, it does not necessarily follow that these properties hold when $N_{\mathrm{samples}}$ is finite. (They hold, after all, only in the asymptotic limit.) However, when $N_{\mathrm{samples}}$ is sufficiently large, we can (and will!) use results derived in the asymptotic limit.

The following subsection presents the technical details and definitions necessary to show the above results. The reader may skip it without loss of continuity, and proceed to Section IV.

## B. Technical Details

Consider a statistical model $\mathcal{M}$ that satisfies MP-LAN. In the next sub-subsections, we prove the properties of $\mathcal{M}$ previously described.

### 1. Convergence of the Local State Space to the Solid Tangent Cone

Because $\mathcal{M}$ satisfies MP-LAN, there exists a model $\mathcal{M}'$ of dimension $d'$ that satisfies LAN. This means that as $N_{\text{samples}} \to \infty$, the distribution of $\hat{\rho}_{\text{ML},\mathcal{M}'}$ converges to a Gaussian:

$$\Pr(\hat{\rho}_{\text{ML},\mathcal{M}'}) \xrightarrow{\text{d}} \mathcal{N}(\rho_0, \Sigma/N_{\text{samples}}),$$

where $\xrightarrow{\text{d}}$ means "converges in distribution to", and $\Sigma = \mathcal{I}^{-1}$. The shape of the distribution is entirely determined by $\mathcal{I}$. As $N_{\text{samples}} \to \infty$, this Gaussian distribution becomes more and more tightly concentrated around $\rho_0$. However, there is still a non-zero probability that $\hat{\rho}_{\text{ML},\mathcal{M}'}$ will be arbitrarily far away from $\rho_0$. Thankfully, it is possible to find a ball $B$ of sufficient size such that it contains every $\hat{\rho}_{\text{ML},\mathcal{M}'}$ with probability 1 as $N_{\text{samples}} \to \infty$.

Before showing $B$ exists, we switch coordinates by sending $\rho \to \rho - \rho_0$, thereby establishing $\rho_0$ as the origin of our coordinate system. In these coordinates, $\hat{\rho}_{\text{ML},\mathcal{M}'} \sim \mathcal{N}(0, \Sigma/N_{\text{samples}})$, and we may easily prove the following lemma showing $B$ exists.

**Lemma 1.** *Let $\hat{\rho}_{\text{ML},\mathcal{M}'} \sim \mathcal{N}(0, \Sigma/N_{\text{samples}})$, and let $\lambda_{\max}(\Sigma)$ denote the largest eigenvalue of $\Sigma$. Define a ball $B = \{\rho \in \mathcal{M}' \mid \text{Tr}(\rho^2) \leq r^2\}$, where $r = \sqrt{\lambda_{\max}(\Sigma)}/N_{\text{samples}}^{1/4}$. Then, $\lim_{N_{\text{samples}}\to\infty} \Pr(\hat{\rho}_{\text{ML},\mathcal{M}'} \in B) = 1$.*

*Proof.* Let $B_0$ be an ellipsoidal ball defined by $\{\rho \in \mathcal{M}' \mid \text{Tr}(\rho\Sigma^{-1}\rho) \leq 1/N_{\text{samples}}^{1/2}\}$. Change coordinates by defining $\sigma = N_{\text{samples}}^{1/2}\Sigma^{-1/2}\rho$. In these new coordinates $\hat{\sigma}_{\text{ML},\mathcal{M}'} \sim \mathcal{N}(0, \mathbb{1}_{d'})$, and $B_0 = \{\sigma \in \mathcal{M}' \mid \text{Tr}(\sigma^2) \leq N_{\text{samples}}^{1/2}\}$. Therefore,

$$\begin{aligned}\Pr(\hat{\sigma}_{\text{ML},\mathcal{M}'} \in B_0) &= \Pr(\text{Tr}(\hat{\sigma}^2_{\text{ML},\mathcal{M}'}) \leq N_{\text{samples}}^{1/2}) \\ &= \int_0^{N_{\text{samples}}^{1/2}} \chi^2_{d'}(z)\, dz,\end{aligned}$$

because $\text{Tr}(\hat{\sigma}^2_{\text{ML},\mathcal{M}'})$ is a $\chi^2_{d'}$ random variable. It follows that

$$\lim_{N_{\text{samples}}\to\infty} \Pr(\hat{\sigma}_{\text{ML},\mathcal{M}'} \in B_0) = \int_0^{\infty} \chi^2_{d'}(z)\, dz = 1.$$

Switching back to the original coordinates, we have

$$B_0 = \{\rho \in \mathcal{M}' \mid \text{Tr}(\rho\Sigma^{-1}\rho) \leq 1/N_{\text{samples}}^{1/2}\},$$

and $\lim_{N_{\text{samples}}\to\infty} \Pr(\hat{\rho}_{\text{ML},\mathcal{M}'} \in B_0) = 1$.

Now that we know $B_0$ contains all the $\hat{\rho}_{\text{ML},\mathcal{M}'}$ asymptotically, we can now show the same holds true for $B$. We do so by showing $B_0 \subset B$. To see this, we write the equation for $B_0$ in the standard quadratic form for an ellipsoid:

$$B_0 = \{\rho \in \mathcal{M}' \mid \text{Tr}(\rho(N_{\text{samples}}^{1/2}\Sigma^{-1})\rho) \leq 1\}.$$

The standard ellipsoid $\{\mathbf{x} \mid \mathbf{x}^T A\mathbf{x} \leq 1\}$ has semi-major axes whose lengths $s_j$ are related to the eigenvalues $a_j$ of $A$: $s_j = 1/\sqrt{a_j}$. The matrix $A = N_{\text{samples}}^{1/2}\Sigma^{-1}$ has eigenvalues $N_{\text{samples}}^{1/2}/\lambda_j$, where $\lambda_j$ are the eigenvalues of $\Sigma$. Thus, the lengths of the semi-major axes of $B_0$ are given by $s_j = 1/\sqrt{N_{\text{samples}}^{1/2}/\lambda_j} = \sqrt{\lambda_j}/N_{\text{samples}}^{1/4}$. Letting $\lambda_{\max}(\Sigma)$ denote the largest eigenvalue of $\Sigma$, the longest semi-major axis of $B_0$ has length $\sqrt{\lambda_{\max}(\Sigma)}/N_{\text{samples}}^{1/4}$. Because $B$ is a ball whose radius is equal to this length, $B$ circumscribes $B_0$, and $B_0 \subset B$.

As $B_0 \subset B$, it follows that $\Pr(\hat{\rho}_{\text{ML},\mathcal{M}'} \in B_0) \leq \Pr(\hat{\rho}_{\text{ML},\mathcal{M}'} \in B)$. As the asymptotic limit of $\Pr(\hat{\rho}_{\text{ML},\mathcal{M}'} \in B_0)$ is 1, and $\Pr(\hat{\rho}_{\text{ML},\mathcal{M}'} \in B)$ itself is bounded above by 1, it follows from the squeeze theorem that $\lim_{N_{\text{samples}}\to\infty} \Pr(\hat{\rho}_{\text{ML},\mathcal{M}'} \in B) = 1$. □

Informally speaking, this lemma implies that "all the action" about the ML estimates $\hat{\rho}_{\text{ML},\mathcal{M}'}$ takes place inside $B$. Accordingly, to understand the behavior of quantities which depend on $\hat{\rho}_{\text{ML},\mathcal{M}'}$ (such as $\hat{\rho}_{\text{ML},\mathcal{M}}$ and $\lambda$), it is sufficient to consider their behavior within $B$.

Things kind of fall apart here, and stop flowing well. Help?

As defined in Lemma 1, $B$ shrinks as $N_{\text{samples}}$ increases. We now change coordinates so that $B, \mathcal{M}$, and $\mathcal{M}'$ all grow as $N_{\text{samples}}$ increases, by performing a *homothetic transformation* of each. In turn, we will be able to show that the local state space (the region contained within $B \cap \mathcal{M}$) converges to the *solid tangent cone*.

**Definition 2** (Homothetic Transformation)**.** *Given a convex set $C$, the homothetic transformation of $C$ with respect to any point $X \in C$, with homothety coefficient $h$, is the set $C_h$ defined by*

$$C_h = \{X + hY \mid \forall\ Y \in C, Y \neq X\}.$$

**Definition 3** (Solid Tangent Cone)**.** *For each point $X$ in a convex set $C$, let $C_h$ be the homoethetic transformation of $C$ with respect to $X$, with homoethety coefficient $h$. Then, the solid tangent cone $T(X)$ is defined as the following limit:*

$$T(X) = \lim_{h\to\infty} C_h.$$

Tangent cones are a general feature of convex sets; see [50], Chapter 6, Section A for more information about them and their properties.

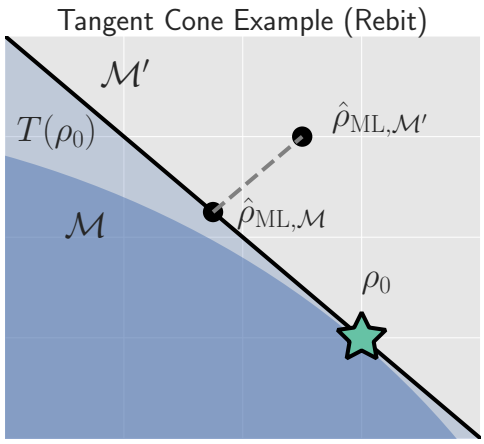


FIG. 4. Example of the solid tangent cone for a rebit. As $N_{\mathrm{samples}} \to \infty$, the local state space around $\rho_0$ converges to $T(\rho_0)$. In Fisher-adjusted coordinates, it's easy to show that (a) $\hat{\rho}_{\mathrm{ML},\mathcal{M}}$ is the metric projection of $\hat{\rho}_{\mathrm{ML},\mathcal{M}'}$ onto $T(\rho_0)$, and (b) $\lambda(\rho_0, \mathcal{M}) = \mathrm{Tr}[(\hat{\rho}_{\mathrm{ML},\mathcal{M}} - \rho_0)^2]$.

Now, we show that the local state space converges to the solid tangent cone. Let $C = \mathcal{M} \cap B$ in Hilbert-Schmidt coordinates. In these coordinates, $B$ shrinks as $N_{\mathrm{samples}}$ increases. Consider the homothetic transformation of $C$ with respect to the origin with homothety coefficient $N_{\mathrm{samples}}$:

$$C_{N_{\mathrm{samples}}} = \{N_{\mathrm{samples}}\rho \;\forall\; \rho \in C\}$$

Under this transformation, $B$ and $\mathcal{M}$ both grow. Taking the limit $N_{\mathrm{samples}} \to \infty$, $C_{N_{\mathrm{samples}}}$ converges to the solid tangent cone at $\rho_0$ (the origin), denoted $T(\rho_0)$.

The geometry of $T(\rho_0)$ depends strongly on $\rho_0$. If $\rho_0$ is rank-deficient within $\mathcal{M}$, then $T(\rho_0)$ is the cone whose faces touch $\mathcal{M}$ at $\rho_0$. (See Figure 4 for a rebit example.) However, when $\rho_0$ is full-rank, $T(\rho_0)$ is $\mathbb{R}^{d^2-1}$.

### 2. *MLE as Metric Projection*

Asymptotically, all the $\hat{\rho}_{\mathrm{ML},\mathcal{M}'}$ are contained within the ball $B$, and the local state space is the solid tangent cone. Because $\mathcal{M}'$ satisfies LAN, the likelihood function around each $\hat{\rho}_{\mathrm{ML},\mathcal{M}'}$ is Gaussian, meaning the optimization problem defining $\hat{\rho}_{\mathrm{ML},\mathcal{M}}$ is given by

$$\hat{\rho}_{\mathrm{ML},\mathcal{M}} = \underset{\sigma \in T(\rho_0)}{\mathrm{argmin}}\ \mathrm{Tr}[(\sigma - \hat{\rho}_{\mathrm{ML},\mathcal{M}'})\mathcal{I}(\sigma - \hat{\rho}_{\mathrm{ML},\mathcal{M}'})]. \quad (13)$$

That is, $\hat{\rho}_{\mathrm{ML},\mathcal{M}}$ is the *metric projection* of $\hat{\rho}_{\mathrm{ML},\mathcal{M}'}$ onto the tangent cone. See Figure 4 for a rebit example. (Notice that if $\hat{\rho}_{\mathrm{ML},\mathcal{M}'} \in T(\rho_0)$, then $\hat{\rho}_{\mathrm{ML},\mathcal{M}} = \hat{\rho}_{\mathrm{ML},\mathcal{M}'}$.)

### 3. *Expression for $\lambda(\rho_0, \mathcal{M})$*

In comparing any two models using the loglikelihood ratio statistic $\lambda(\mathcal{M}_1, \mathcal{M}_2)$, it's possible to do so by comparing each to any *reference model* $\mathcal{R}$:

$$\lambda(\mathcal{M}_1, \mathcal{M}_2) = \lambda(\mathcal{R}, \mathcal{M}_2) - \lambda(\mathcal{R}, \mathcal{M}_1),$$

where the quantity $\lambda(\mathcal{R}, \mathcal{M})$ is given by

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

Taking $\mathcal{R} = \rho_0$ itself, and using the fact that asymptotically $\mathcal{L}(\rho)$ is Gaussian around $\hat{\rho}_{\mathrm{ML},\mathcal{M}'}$, we have

$$\begin{aligned}\lambda(\rho_0, \mathcal{M}) &= -2\log\left(\frac{\mathcal{L}(\rho_0)}{\max\limits_{\rho\in\mathcal{M}} \mathcal{L}(\rho)}\right) \\ &= \mathrm{Tr}[(\rho_0 - \hat{\rho}_{\mathrm{ML},\mathcal{M}'})\mathcal{I}(\rho_0 - \hat{\rho}_{\mathrm{ML},\mathcal{M}'})] \\ &\quad - \mathrm{Tr}[(\hat{\rho}_{\mathrm{ML},\mathcal{M}} - \hat{\rho}_{\mathrm{ML},\mathcal{M}'})\mathcal{I}(\hat{\rho}_{\mathrm{ML},\mathcal{M}} - \hat{\rho}_{\mathrm{ML},\mathcal{M}'})]. \end{aligned} \quad (14)$$

It turns out $\lambda(\rho_0, \mathcal{M})$ has a simple form.

**Lemma 2.** $\lambda(\rho_0, \mathcal{M}) = \mathrm{Tr}[(\rho_0 - \hat{\rho}_{\mathrm{ML},\mathcal{M}})\mathcal{I}(\rho_0 - \hat{\rho}_{\mathrm{ML},\mathcal{M}})]$.

*Proof.* We switch to Fisher-adjusted coordinates, through the mapping $\rho \to \mathcal{I}^{1/2}\rho$. Then, we may replace $\mathcal{I}$ by $\mathbb{1}$:

$$\lambda(\rho_0, \mathcal{M}) = \mathrm{Tr}[(\rho_0 - \hat{\rho}_{\mathrm{ML},\mathcal{M}'})^2] - \mathrm{Tr}[(\hat{\rho}_{\mathrm{ML},\mathcal{M}} - \hat{\rho}_{\mathrm{ML},\mathcal{M}'})^2]. \quad (15)$$

To demonstrate the lemma statement, we must consider two cases:

*Case 1*: Assume $\hat{\rho}_{\mathrm{ML},\mathcal{M}'} \notin T(\rho_0)$. Because $\hat{\rho}_{\mathrm{ML},\mathcal{M}}$ is the metric projection of $\hat{\rho}_{\mathrm{ML},\mathcal{M}'}$ onto $T(\rho_0)$ (Equation (13)), the line joining $\hat{\rho}_{\mathrm{ML},\mathcal{M}'}$ and $\hat{\rho}_{\mathrm{ML},\mathcal{M}}$ is normal to $T(\rho_0)$ at $\hat{\rho}_{\mathrm{ML},\mathcal{M}}$. Because $T(\rho_0)$ contains $\rho_0$ (as its origin), it follows that the lines joining $\rho_0$ to $\hat{\rho}_{\mathrm{ML},\mathcal{M}}$, and $\hat{\rho}_{\mathrm{ML},\mathcal{M}}$ to $\hat{\rho}_{\mathrm{ML},\mathcal{M}'}$, are at right angles to one another. (See Figure 4.)

By the Pythagorean theorem, we have

$$\mathrm{Tr}[(\rho_0 - \hat{\rho}_{\mathrm{ML},\mathcal{M}'})^2] = \mathrm{Tr}[(\rho_0 - \hat{\rho}_{\mathrm{ML},\mathcal{M}})^2] + \mathrm{Tr}[(\hat{\rho}_{\mathrm{ML},\mathcal{M}} - \hat{\rho}_{\mathrm{ML},\mathcal{M}'})^2]$$

Subtracting $\mathrm{Tr}[(\hat{\rho}_{\mathrm{ML},\mathcal{M}} - \hat{\rho}_{\mathrm{ML},\mathcal{M}'})^2]$ from both sides, and comparing to Equation (15), yields the lemma statement in Fisher-adjusted coordinates.

*Case 2*: Assume $\hat{\rho}_{\mathrm{ML},\mathcal{M}'} \in T(\rho_0)$. Then, $\hat{\rho}_{\mathrm{ML},\mathcal{M}} = \hat{\rho}_{\mathrm{ML},\mathcal{M}'}$, and Equation (15) simplifies to the lemma statement in Fisher-adjusted coordinates.

Switching back from Fisher-adjusted coordinates, we have $\lambda(\rho_0, \mathcal{M}) = \mathrm{Tr}[(\rho_0 - \hat{\rho}_{\mathrm{ML},\mathcal{M}})\mathcal{I}(\rho_0 - \hat{\rho}_{\mathrm{ML},\mathcal{M}})]$. □

Thus, when MP-LAN is satisfied, asymptotically the loglikelihood ratio statistic is related to *squared error/loss* (as measured by the Fisher metric.) This result may be of independent interest in, for example, considering/defining new information criteria, which attempt to balance goodness of fit (as measured by $\lambda$) against error/loss (generally, as measured by the squared error).

With these technical results in hand, we may proceed to compute $\langle\lambda(\mathcal{M}_d, \mathcal{M}_{d+1})\rangle$. Doing so is the subject of the next section.

## IV. A WILKS THEOREM FOR QUANTUM STATE SPACE

With the framework of MP-LAN available to us, we are ready to derive a replacement for the Wilks theorem. We first need to show that the models $\mathcal{M}_d$ satisfy MP-LAN. Thankfully, it is straightforward to do so.

**Lemma 3.** *The models $\mathcal{M}_d$, defined in Equation (9), satisfy MP-LAN.*

*Proof.* Let $\mathcal{M}'_d = \{\sigma \mid \dim(\sigma) = d, \sigma = \sigma^\dagger\}$. (That is, $\mathcal{M}'_d$ is the set of all $d \times d$ Hermitian matrices, but we do not require them to be non-negative, nor trace-1.) It is clear $\mathcal{M}_d \subset \mathcal{M}'_d$. Next, $\forall\ \sigma \in \mathcal{M}'_d$, the likelihood $\mathcal{L}(\sigma)$ is twice continuously differentiable, meaning $\mathcal{M}'_d$ satisfies LAN. Thus, $\mathcal{M}_d$ satisfies MP-LAN. □

It's useful to reduce the problem of computing $\lambda(\mathcal{M}_d, \mathcal{M}_{d+1})$ to that of computing $\lambda(\rho_0, \mathcal{M}_k)$ for $k = d, d+1$ using the identity

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

where $\lambda(\rho_0, \mathcal{M}_k)$ is given in Equation (11). Because each model satisfies MP-LAN, it is possible to write $\lambda(\rho_0, \mathcal{M}_k)$ in a very simple form, via Equation (12):

$$\lambda(\rho_0, \mathcal{M}_k) = \mathrm{Tr}[(\rho_0 - \hat{\rho}_{\mathrm{ML},\mathcal{M}_k})\mathcal{I}_k(\rho_0 - \hat{\rho}_{\mathrm{ML},\mathcal{M}_k})].$$

The Fisher information $\mathcal{I}_k$ is generally anisotropic, depending on $\rho_0$, the POVM being measured, and the model $\mathcal{M}_k$. (see Figure 5). And while the $\rho \geq 0$ constraint that invalidated LAN in the first place is at least somewhat tractable in standard (Hilbert-Schmidt) coordinates, it becomes completely intractable in Fisher-adjusted coordinates. So, to obtain a semi-analytic null theory for $\lambda$, we will simplify to the case where $\mathcal{I}_k = \mathbb{1}_k/\epsilon$ for some $\epsilon$ that scales as $1/\sqrt{N_{\mathrm{samples}}}$. (That is, $\mathcal{I}_k$ is proportional to the Hilbert-Schmidt metric for model $\mathcal{M}_k$.) This simplification permits the derivation of analytic results which capture realistic tomographic scenarios surprisingly well [51].

With this simplification, $\lambda(\mathcal{M}_d, \mathcal{M}_{d+1})$ is given by

$$\lambda = \frac{1}{\epsilon^2}\left(\mathrm{Tr}[(\rho_0 - \hat{\rho}_{\mathrm{ML},d+1})^2] - \mathrm{Tr}[(\rho_0 - \hat{\rho}_{\mathrm{ML},d})^2]\right). \quad (16)$$

That is, $\lambda$ is a *difference* in Hilbert-Schmidt distances. This expression makes it clear why a null theory for $\lambda$ is necessary: even if $\rho_0 \in \mathcal{M}_d, \mathcal{M}_{d+1}$, the error of $\hat{\rho}_{\mathrm{ML},d+1}$ will be greater than that of $\hat{\rho}_{\mathrm{ML},d}$ (because there are more parameters that can fit noise in the data). The null theory for $\lambda$ tells us how much extra error we will have in using $\mathcal{M}_{d+1}$ to reconstruct $\rho_0$ when $\mathcal{M}_d$ is just as good.

In computing $\Pr(\lambda)$, we face the obstacle that the distributions of $\hat{\rho}_{\mathrm{ML},d}, \hat{\rho}_{\mathrm{ML},d+1}$ are complicated, highly non-Gaussian, and singular (because estimates "pile up" on the various faces of the boundary; recall Figure 1). For this reason, we will not attempt to compute $\Pr(\lambda)$ directly. Instead, we focus on deriving a good approximation for $\langle\lambda\rangle$.

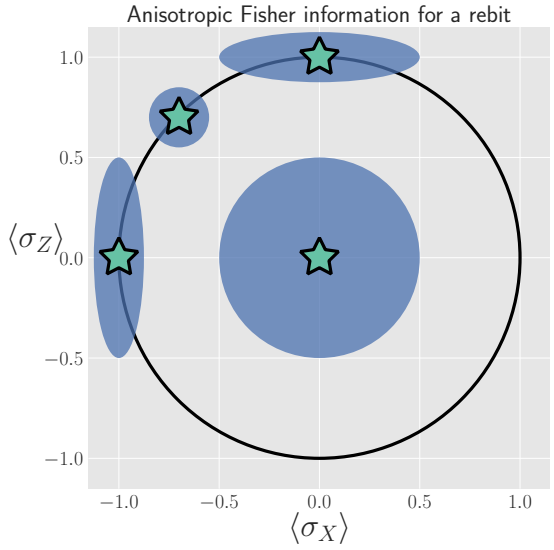


FIG. 5. Anisotropy of the Fisher information for a rebit measured using the POVM $\frac{1}{2}\{|0\rangle\langle 0|, |1\rangle\langle 1|, |+\rangle\langle +|, |-\rangle\langle -|\}$. Depending on $\rho_0$ (star), the distribution of the unconstrained estimates $\hat{\rho}_{\mathrm{ML}}$ (ellipses) may be anisotropic. Imposing the positivity constraint $\rho \geq 0$ is difficult in Fisher-adjusted coordinates; in this paper, we simplify these complexities to the case where $\mathcal{I} \propto \mathbb{1}$, *and is independent of* $\rho_0$.

Because $\langle\lambda\rangle$ itself depends on two terms, each of which has a similar form, we can further simplify our task to that of computing $\epsilon^2\langle\lambda(\rho_0, \mathcal{M}_d)\rangle = \langle\mathrm{Tr}[(\hat{\rho}_{\mathrm{ML},d} - \rho_0)^2]\rangle$ for arbitrary $d$. Doing so involves two main steps:

(1) Identifying degrees of freedom in $\hat{\rho}_{\mathrm{ML},\mathcal{M}'_d}$ that are, and are not, affected by projection onto the tangent cone $T(\rho_0)$.

(2) For both types of degrees of freedom, evaluate the contributions of each to the value of $\langle\lambda\rangle$.

The remainder of this section is devoted to both steps. In Section IV A, we identify two types of degrees of freedom in $\hat{\rho}_{\mathrm{ML},\mathcal{M}'}$, which we call the "L" and the "kite". Section IV B computes the contribution of degrees of freedom in the "L", and Section IV C computes the contribution from the "kite". The total expected value is given in Equation (26) in Section IV D, on page 11.

### A. Separating out Degrees of Freedom in $\hat{\rho}_{\mathrm{ML},\mathcal{M}'_d}$

We begin by observing that $\lambda(\rho_0, \mathcal{M}_d)$ can be written as a sum over matrix elements,

$$\lambda = \epsilon^{-2}\mathrm{Tr}[(\hat{\rho}_{\mathrm{ML},d} - \rho_0)^2] = \epsilon^{-2}\sum_{jk}|(\hat{\rho}_{\mathrm{ML},d} - \rho_0)_{jk}|^2$$
$$= \sum_{jk}\lambda_{jk} \quad \text{where} \quad \lambda_{jk} = \epsilon^{-2}|(\hat{\rho}_{\mathrm{ML},d} - \rho_0)_{jk}|^2, \quad (17)$$

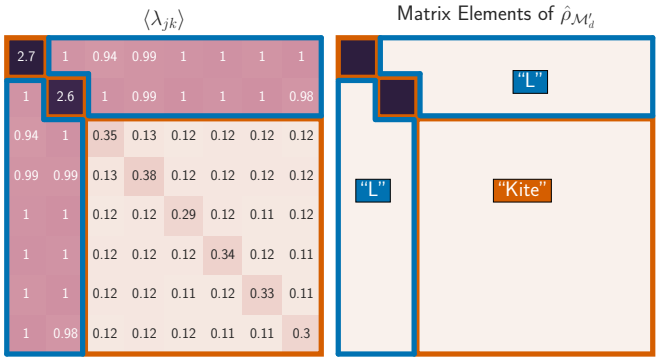


FIG. 6. When a rank-2 state is reconstructed in $d = 8$ dimensions, the total loglikelihood ratio $\lambda(\rho_0, \mathcal{M}_8)$ is the sum of terms $\lambda_{jk}$ from errors in each matrix element $(\hat{\rho}_{\mathrm{ML},d})_{jk}$. **Left**: Numerics show a clear division; some matrix elements have $\langle\lambda_{jk}\rangle \sim 1$ as predicted by the Wilks theorem, while others are either more or less. **Right**: The numerical results support our theoretical reasoning for dividing the matrix elements of $\hat{\rho}_{\mathrm{ML},\mathcal{M}'_d}$ into two parts: the "kite" and the "L".

and therefore $\langle\lambda\rangle = \sum_{jk}\langle\lambda_{jk}\rangle$. Each term $\langle\lambda_{jk}\rangle$ quantifies the mean-squared error of a single matrix element of $\hat{\rho}_{\mathrm{ML},d}$, and while the Wilks theorem predicts $\langle\lambda_{jk}\rangle = 1$ for all $j, k$, due to positivity constraints, this no longer holds. In particular, the matrix elements of $\hat{\rho}_{\mathrm{ML},d}$ now fall into two parts:

1. Those for which the corresponding matrix elements of $\hat{\rho}_{\mathrm{ML},\mathcal{M}'_d}$ *are constrained*.
2. Those for which the corresponding matrix elements of $\hat{\rho}_{\mathrm{ML},\mathcal{M}'_d}$ *are not constrained*, as they correspond to directions on the surface of the tangent cone $T(\rho_0)$. (Recall Figure 4 - as a component of $\hat{\rho}_{\mathrm{ML},\mathcal{M}'}$ along $T(\rho_0)$ changes, the component of $\hat{\rho}_{\mathrm{ML},\mathcal{M}}$ changes by the same amount. These elements are unconstrained.)

The former, which lie in what we call the "L", comprise all off-diagonal elements on the support of $\rho_0$ *and* between the support and the kernel, while the latter, which lie in what we call the "kite", are all diagonal elements *and* all elements on the kernel (null space) of $\rho_0$.

Performing this division is also supported by numerical simulations (see Figure 6). Matrix elements in the "L" appear to contribute $\langle\lambda_{jk}\rangle = 1$, consistent with the Wilks theorem, while those in the "kite" contribute more (if they are within the support of $\rho_0$) or less (if they are in the kernel). Having performed the division of the matrix elements of $\hat{\rho}_{\mathrm{ML},\mathcal{M}'_d}$, we observe that $\langle\lambda\rangle = \langle\lambda_{\mathrm{L}}\rangle + \langle\lambda_{\mathrm{kite}}\rangle$. Because each $\langle\lambda_{jk}\rangle$ is not necessarily equal to one (as in the Wilks theorem), and because many of them are less than 1, it's clear that their total $\langle\lambda\rangle$ is dramatically lower than that predicted by the Wilks theorem. (Recall Figure 2.)

In the following subsections, we develop a theory to explain the behavior of $\langle\lambda_{\mathrm{L}}\rangle$ and $\langle\lambda_{\mathrm{kite}}\rangle$. In doing so, it is helpful to think about the matrix $\delta \equiv \hat{\rho}_{\mathrm{ML},\mathcal{M}'_d} - \rho_0$, a normally-distributed *traceless* matrix. To simplify the analysis, we explicitly drop the $\mathrm{Tr}(\delta) = 0$ constraint and let $\delta$ 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 $\delta$ proportional to an element of the Gaussian Unitary Ensemble (GUE) [52].

### B. Computing $\langle\lambda_{\mathrm{L}}\rangle$

The value of each $\delta_{jk}$ in the "L" is invariant under projection onto the boundary (the surface of the tangent cone $T(\rho_0)$), meaning that it is also equal to the error $(\hat{\rho}_{\mathrm{ML},d} - \rho_0)_{jk}$. Therefore, $\langle\lambda_{jk}\rangle = \langle\delta_{jk}^2\rangle/\epsilon^2$. Because $\mathcal{M}'$ satisfies LAN, it follows that each $\delta_{jk}$ is an i.i.d. Gaussian random variable with mean zero and variance $\epsilon^2$. Thus, $\langle\lambda_{jk}\rangle = 1\ \forall\ (j,k)$ in the "L". The dimension of the surface of the tangent cone is equal to the dimension of the manifold of rank-$r$ states in a $d$-dimensional space. A direct calculation of that quantity yields $2rd - r(r+1)$, so $\langle\lambda_{\mathrm{L}}\rangle = 2rd - r(r+1)$.

Another way of viewing this result is to observe that the $\delta_{jk}$ in the "L" can be seen as errors that arise due to small unitary perturbations of $\rho_0$. Writing $\hat{\rho}_{\mathrm{ML},\mathcal{M}'_d} = U^\dagger\rho_0 U$, where $U = e^{i\epsilon H}$, we have

$$\hat{\rho}_{\mathrm{ML},\mathcal{M}'_d} \approx \rho_0 + i\epsilon[\rho_0, H] + \mathcal{O}(\epsilon^2),$$

and $\delta \approx i\epsilon[\rho_0, H]$. If $j = k$, then $\delta_{jj} = 0$. Thus, small unitaries cannot create errors in the diagonal matrix elements, at $\mathcal{O}(\epsilon)$. If $j \neq k$, then $\delta_{jk} \neq 0$, in general. (Small unitaries *can* introduce errors on off-diagonal elements.)

However, if either $j$ or $k$ (or both) lie within the *kernel* of $\rho_0$ (i.e., $\langle k|\rho_0|k\rangle$ or $\langle j|\rho_0|j\rangle$ is 0), then the corresponding $\delta_{jk}$ are zero. The only off-diagonal elements where small unitaries can introduce errors are those which are coherent between the kernel of $\rho_0$ and its support. These off-diagonal elements are precisely the "L", and are the set $\{\delta_{jk} \mid \langle j|\rho_0|j\rangle \neq 0, j \neq k,\ 0 \leq j,k \leq d-1\}$. This set contains $2rd - r(r+1)$ elements, each of which has $\langle\lambda_{jk}\rangle = 1$, so we again arrive at $\langle\lambda_{\mathrm{L}}\rangle = 2rd - r(r+1)$.

### C. Computing $\langle\lambda_{\mathrm{kite}}\rangle$

Computing $\langle\lambda_{\mathrm{L}}\rangle$ was made easy by the fact that the matrix elements of $\delta$ in the "L" are invariant under the projection of $\hat{\rho}_{\mathrm{ML},\mathcal{M}'_d}$ onto $T(\rho_0)$. Computing $\langle\lambda_{\mathrm{kite}}\rangle$ is a bit harder, because the boundary *does* constrain its elements. Thus, we may naively think we need to explicitly compute $\hat{\rho}_{\mathrm{ML},d}$ given $\hat{\rho}_{\mathrm{ML},\mathcal{M}'_d}$, and only then determine $\langle\lambda_{\mathrm{kite}}\rangle$. However, this is not exactly the case.

To solve the optimization problem in Equation (10), one can use an algorithm developed in [51]:

1. Subtract $q\mathbb{1}$ from $\hat{\rho}_{\mathrm{ML},\mathcal{M}'_d}$, for a particular $q \in \mathbb{R}$.
2. "Truncate" $\hat{\rho}_{\mathrm{ML},\mathcal{M}'_d} - q\mathbb{1}$, by replacing each of its negative eigenvalues with zero.

Here, $q$ is defined implicitly such that $\mathrm{Tr}\left[\mathrm{Trunc}(\hat{\rho}_{\mathrm{ML},\mathcal{M}'_d} - q\mathbb{1})\right] = 1$.

The truncation algorithm given in [51] for finding $\hat{\rho}_{\mathrm{ML},d}$ is most naturally performed in the eigenbasis of $\hat{\rho}_{\mathrm{ML},\mathcal{M}'_d}$. Exact diagonalization of $\hat{\rho}_{\mathrm{ML},\mathcal{M}'_d}$ is not feasible analytically, but only its *small* eigenvalues are critical in truncation. Further, only knowledge of the *typical* eigenvalues of $\hat{\rho}_{\mathrm{ML},d}$ is necessary for computing $\langle\lambda_{\mathrm{kite}}\rangle$. Therefore, we do not need to determine $\hat{\rho}_{\mathrm{ML},d}$ exactly, which would require explicitly solving Equation (10); instead, we need a procedure for determining its typical eigenvalues.

We assume that $N_{\mathrm{samples}}$ is sufficiently large so that all the nonzero eigenvalues of $\rho_0$ are much larger than $\epsilon$. This means the eigenbasis of $\hat{\rho}_{\mathrm{ML},\mathcal{M}'_d}$ is accurately approximated by: (1) the eigenvectors of $\rho_0$ on its support; and (2) the eigenvectors of $\delta_{\mathrm{ker}} = \Pi_{\mathrm{ker}}\delta\Pi_{\mathrm{ker}} = \Pi_{\mathrm{ker}}\hat{\rho}_{\mathrm{ML},\mathcal{M}'_d}\Pi_{\mathrm{ker}}$, where $\Pi_{\mathrm{ker}}$ is the projector onto the kernel of $\rho_0$.

Changing to this basis diagonalizes the "kite" portion of $\delta$, and leaves all elements of the "L" unchanged (at $\mathcal{O}(\epsilon)$). The diagonal elements fall into two categories:

1. $r$ elements corresponding to the eigenvalues of $\rho_0$, which are given by $p_j = \rho_{jj} + \delta_{jj}$ where $\rho_{jj}$ is the $j^{\mathrm{th}}$ eigenvalue of $\rho_0$, and $\delta_{jj} \sim \mathcal{N}(0, \epsilon^2)$.
2. $N \equiv d - r$ elements that are eigenvalues of $\delta_{\mathrm{ker}}$, which we denote by $\boldsymbol{\kappa} = \{\kappa_j : \ j = 1 \ldots N\}$.

In turn, $q$ is the solution to

$$\sum_{j=1}^{r}(p_j - q)^+ + \sum_{j=1}^{N}(\kappa_j - q)^+ = 1, \tag{18}$$

where $(x)^+ = \max(x, 0)$, and $\lambda_{\mathrm{kite}}$ is

$$\epsilon^2\lambda_{\mathrm{kite}} = \sum_{j=1}^{r}[\rho_{jj} - (p_j - q)^+]^2 + \sum_{j=1}^{N}\left[(\kappa_j - q)^+\right]^2. \tag{19}$$

To solve Equation (18), and derive an approximation for (19), we use the fact that we are interested in computing the *average value* of $\lambda_{\mathrm{kite}}$, which justifies approximating the random variable $q$ by a closed-form, deterministic value. To do so, we need to understand the behavior of $\boldsymbol{\kappa}$. Developing such an understanding, and a theory of its *typical value*, is the subject of the next section.

### 1. *Approximating the eigenvalues of a GUE(N) matrix*

We first observe that while the $\kappa_j$ are random variables, they are *not* normally distributed. Instead, because $\delta_{\mathrm{ker}}$ is proportional to a GUE($N$) matrix, for $N \gg 1$, the distribution of any eigenvalue $\kappa_j$ converges to a Wigner semicircle distribution [53], 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" [54]), and typically form a surprisingly regular array over the support of the Wigner semicircle.

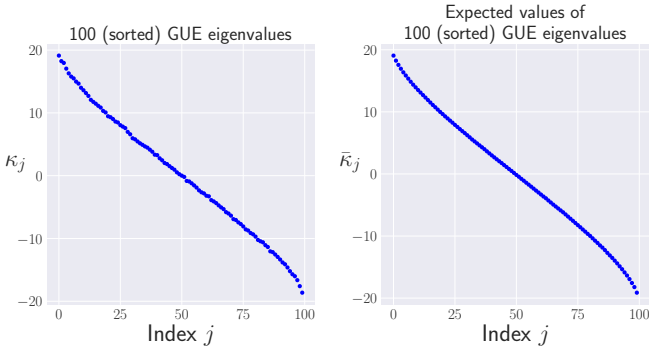


FIG. 7. Typical samples of GUE($N$) eigenvalues are accurately approximated by order statistics of the distribution (average values of a sorted sample). **Left**: The sorted eigenvalues (i.e., order statistics $\kappa_j$) of one randomly chosen GUE(100) matrix. **Right**: Approximate expected values of the order statistics, $\bar{\kappa}_j$, of the GUE(100) distribution, computed as the average of the sorted eigenvalues of 100 randomly chosen GUE(100) matrices.

Since our goal is to compute $\langle\lambda_{\mathrm{kite}}\rangle$, we can capitalize on this behavior by replacing each random sample of $\boldsymbol{\kappa}$ with a *typical sample* given by its order statistics $\bar{\boldsymbol{\kappa}}$. These are the average values of the *sorted* $\boldsymbol{\kappa}$, so $\overline{\kappa}_j$ is the average value of the $j^{\mathrm{th}}$ 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.

Suppose that $\boldsymbol{\kappa}$ are the eigenvalues of a GUE($N$) matrix, sorted from highest to lowest. Figure 7 illustrates such a sample for $N = 100$. It also shows the *average* values of 100 such samples (all sorted). These are the *order statistics* $\overline{\boldsymbol{\kappa}}$ 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). As the figure shows, 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 $\kappa_j$ greater or less than the threshold $q$, or partial sums of functions of the $\kappa_j$ (e.g., $(\kappa_j - q)^2$). 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 functions.

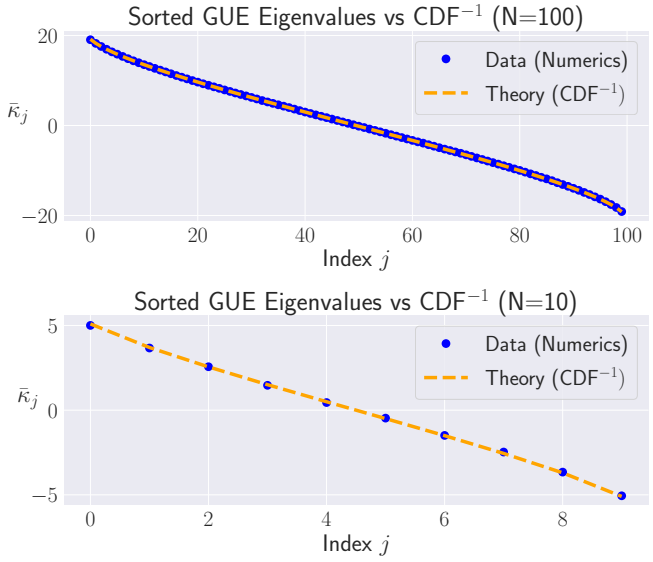


FIG. 8. Order statistics of the GUE($N$) eigenvalue distribution are very well approximated by the inverse CDF of the Wigner semicircle distribution. In both figures, we compare the order statistics of a GUE($N$) distribution to the inverse CDF of the Wigner semicircle distribution. **Top**: $N = 100$. **Bottom**: $N = 10$. Agreement in both cases is essentially perfect.

In general, if a sample $\boldsymbol{\kappa}$ of size $N$ is drawn so that each $\kappa$ has the same probability density function $\Pr(\kappa)$, then a good approximation for the $j^{\text{th}}$ order statistic is given by the inverse *cumulative* distribution function (CDF):

$$\overline{\kappa}_j \approx \text{CDF}^{-1}\left(\frac{j-1/2}{N}\right). \tag{20}$$

This is closely related to the observation that the histogram of a sample tends to look similar to the underlying probability density function. (More precisely, it is equivalent to the observation that the empirical distribution function (the CDF of the histogram) tends to be (even more) similar to the underlying CDF. For i.i.d. samples, this is the content of the Glivenko-Cantelli theorem [55]). Figure 8 compares the order statistics of GUE(100) and GUE(10) eigenvalues (computed as numerical averages over 100 random samples) to the inverse CDF 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 $\overline{\boldsymbol{\kappa}}$ even at $N = 10$ (and remains surprisingly good all the way down to $N = 2$).

We make one further approximation, by assuming that $N \gg 1$, so the distribution of the $\overline{\kappa}_j$ is effectively continuous and identical to $\Pr(\kappa)$. For the quantities that we compute, this is equivalent to replacing the empirical distribution function (which is a step function) by the CDF of the Wigner semicircle distribution. So, whereas for any given sample the partial sum of all $\kappa_j > q$ jumps discontinuously when $q = \kappa_j$ for any $j$, in this approximation it changes smoothly. This accurately models the *average* behavior of partial sums.

### 2. *Deriving an approximation for q*

The approximations of the previous section allow us to use the following ansatz for the eigenvalues of $\hat{\rho}_{\text{ML},\mathcal{M}'_d}$; namely, $\{p_j\} \cup \{\overline{\kappa}_j\}$, 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. In turn, the defining equation for $q$ (Equation (18)) is well approximated as

$$\sum_{j=1}^{r}(p_j - q)^+ + \sum_{j=1}^{N}(\overline{\kappa}_j - q)^+ \approx 1. \tag{21}$$

To solve this equation, we observe that the $\overline{\kappa}_j$ are symmetrically distributed around $\kappa = 0$, so half of them are negative. Therefore, with high probability, $\text{Tr}\left[\text{Trunc}(\hat{\rho}_{\text{ML},\mathcal{M}'_d})\right] > 1$, and so we will need to subtract $q\mathbb{1}$ from $\hat{\rho}_{\text{ML},\mathcal{M}'_d}$ before truncating.

Because we have assumed $N_{\text{samples}}$ is sufficiently large ($N_{\text{samples}} >> \min_j 1/\rho_{jj}^2$), it follows that the eigenvalues of $\rho_0$ are large compared to the perturbations $\delta_{jj}$ and $q$. This implies $(p_j - q)^+ = p_j - q$. Under this assumption, $q$ is the solution to

$$\begin{aligned} 1 &\approx \sum_{j=1}^{r}(p_j - q)^+ + \sum_{j=1}^{N}(\overline{\kappa}_j - q)^+ \\ &\approx 1 - rq + \Delta + N\int_{\kappa=q}^{2\epsilon\sqrt{N}}(\kappa - q)\Pr(\kappa)\mathrm{d}\kappa \\ \implies 0 &= -rq + \Delta + \frac{\epsilon}{12\pi}\left[\begin{array}{l}(q^2+8N)\sqrt{-q^2+4N} \\ -12qN\left(\frac{\pi}{2} - \sin^{-1}\left(\frac{q}{2\sqrt{N}}\right)\right)\end{array}\right], \end{aligned} \tag{22}$$

where $\Delta = \sum_{j=1}^{r}\delta_{jj}$ is a $\mathcal{N}(0, r\epsilon^2)$ random variable. We choose to replace a discrete sum (line 1) with an integral (line 2). This approximation is valid when $N \gg 1$, as 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 CDF. It is also remarkably accurate in practice.

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

To solve Equation (22), it is necessary to further simplify the complicated expression resulting from the integral (line 3). To do so, we assume $\rho_0$ 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

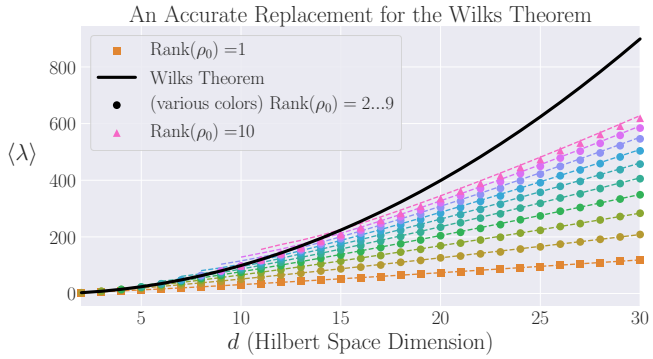


FIG. 9. Numerical results for $\langle\lambda\rangle$ compared to the prediction of the Wilks theorem (solid line) and our replacement theory as given in Equation (26) (dashed lines). Our formula depends on the rank $r$ of $\rho_0$ (unlike the Wilks prediction), and is nearly perfect for $r \ll d$. It becomes less accurate as $r$ approaches $d/2$, and is invalid when $r \approx d$.

leading order Taylor expansion around $q = 2\epsilon\sqrt{N}$, substitute into Equation (22), and obtain the equation

$$\frac{rq}{\epsilon} = \frac{4}{15\pi} N^{1/4} \left(2\sqrt{N} - \frac{q}{\epsilon}\right)^{5/2}. \tag{23}$$

This equation is a quintic polynomial in $q/\epsilon$, so by the Abel-Ruffini theorem, it has no algebraic solution. However, as $N \to \infty$, its roots have a well-defined algebraic *approximation* that becomes accurate quite rapidly (e.g., for $N > 4$):

$$z \equiv q/\epsilon \approx 2\sqrt{N} - \left(\frac{15\pi r}{2}\right)^{2/5} N^{1/10} + \frac{1}{5}\left(\frac{15\pi r}{2}\right)^{4/5} N^{-3/10} - \frac{1}{100}\left(\frac{15\pi r}{2}\right)^{6/5} N^{-7/10}. \tag{24}$$

#### 3. *Expression for* $\langle\lambda_{\text{kite}}\rangle$

Now that we know how much to subtract off in the truncation process, we can approximate $\langle\lambda_{\text{kite}}\rangle$, originally given in Equation (19):

$$\begin{aligned}\langle\lambda_{\text{kite}}\rangle &\approx \frac{1}{\epsilon^2}\left\langle \sum_{j=1}^{r} [\rho_{jj} - (p_j - q)^+]^2 + \sum_{j=1}^{N} \left[(\bar{\kappa}_j - q)^+\right]^2 \right\rangle \\ &\approx \frac{1}{\epsilon^2}\left\langle \sum_{j=1}^{r} [-\delta_{jj} + q]^2 + \sum_{j=1}^{N} \left[(\bar{\kappa}_j - q)^+\right]^2 \right\rangle \\ &\approx r + rz^2 + \frac{N}{\epsilon^2}\int_{\kappa=q}^{2\epsilon\sqrt{N}} \Pr(\kappa)(\kappa - q)^2 d\kappa \\ &= r + rz^2 + \frac{N(N+z^2)}{\pi}\left(\frac{\pi}{2} - \sin^{-1}\left(\frac{z}{2\sqrt{N}}\right)\right) \\ &\quad - \frac{z(z^2+26N)}{24\pi}\sqrt{4N - z^2}. \end{aligned} \tag{25}$$

### D. Complete Expression for $\langle\lambda\rangle$

The total expected value, $\langle\lambda\rangle = \langle\lambda_{\text{L}}\rangle + \langle\lambda_{\text{kite}}\rangle$, is thus

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

where $z$ is given in Equation (24), $N = d - r$, and $r = \text{Rank}(\rho_0)$.

## V. COMPARISON TO NUMERICAL EXPERIMENTS

### A. Isotropic Fisher Information

Equation (26) is our main result. To test its validity, we compare it to numerical simulations for the case of an isotropic Fisher information with $d = 2, \ldots, 30$ and $r = 1, \ldots, 10$ in Figure 9. The prediction of the Wilks theorem is wildly incorrect for $r \ll d$. In contrast, Equation (26) 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 (26)) correctly models tomography *if* the Fisher information is isotropic ($\mathcal{I} \propto \mathbb{1}$).

### B. Comparison to 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 (equivalent to double homodyne) state tomography [11, 12, 14, 56] of a single optical mode. (See Figure 10 for a plot of the *condition number* – the ratio of the largest eigenvalue to the smallest – of the estimated Fisher information. It is clear that $\mathcal{I} \not\propto \mathbb{1}$.) 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 tomography reconstructs $\rho_0$ 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 Husimi $Q$-function of $\rho_0$ [57].

We examined the behavior of $\lambda$ for 13 distinct states $\rho_0$, 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 $N_{\text{samples}} = 10^5$, and found MLEs

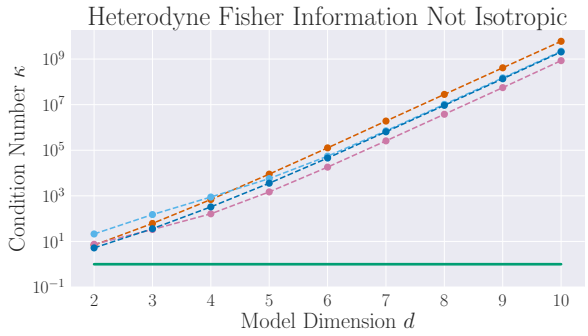


FIG. 10. The condition number $\kappa$ – the ratio of the largest eigenvalue to the smallest – of the estimated heterodyne Fisher information. (Estimates are the average over 100 Hessians of the loglikelihood function.) $\kappa$ grows with model dimension, meaning anisotropy is increasing. The dashed lines indicate different states $\rho_0$, and the solid line is $\kappa = 1$ (i.e., $\mathcal{I} \propto \mathbb{1}$.).

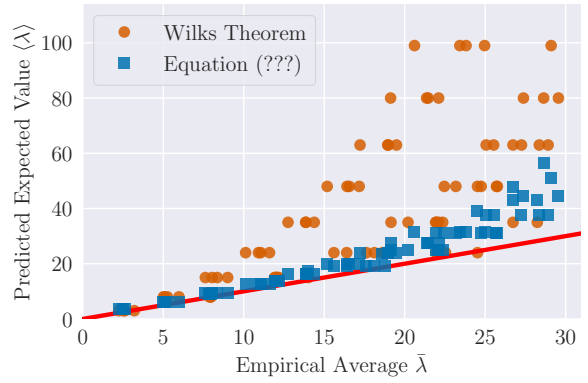


FIG. 11. The Wilks theorem (orange dots) dramatically overestimates $\langle\lambda(\rho_0, \mathcal{M}_d)\rangle$ in optical heterodyne tomography. Our formula, Equation 26 (blue squares), is far more accurate. Residual discrepancies occur in large part because $N_{\text{samples}}$ is not yet "asymptotically large". The solid red line corresponds to perfect correlation between theory ($\langle\lambda\rangle$) and practice ($\bar{\lambda}$).

Put the right equation number in here

over each of the 9 models $\mathcal{M}_2, \ldots, M_{10}$ using numerical optimization [58]. For each $\rho_0$ and each $d$, we averaged $\lambda(\rho_0, \mathcal{M}_d)$ over all 100 datasets to obtain an empirical average loglikelihood ratio $\bar{\lambda}(\rho_0, d)$.

Results of this test are shown in Figure 11, where we plot the predictions for $\langle\lambda\rangle$ given by the Wilks theorem and Equation (26), against the empirical average $\bar{\lambda}$, for a variety of $\rho_0$ and $d$. Our formula correlates very well with the empirical average, while the Wilks theorem (unsurprisingly) overestimates $\lambda$ dramatically for low-rank states. Whereas a model selection procedure based on the 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, Equation (26) also begins to overpredict. As Figure 12 indicates, a more accurate description is that the numerical experiments are *underachieving*, because $\bar{\lambda}$ is still growing with $N_{\text{samples}}$. Both the Wilks theorem and our analysis are derived in the asymptotic limit $N_{\text{samples}} \to \infty$; for finite but large $N_{\text{samples}}$, both may be invalid. Figure 12 shows that, even at $N_{\text{samples}} \sim 10^5$, the behavior of $\bar{\lambda}$ has failed to become asymptotic. This is surprising, and suggests heterodyne tomography is a particularly exceptional and challenging case to model statistically.

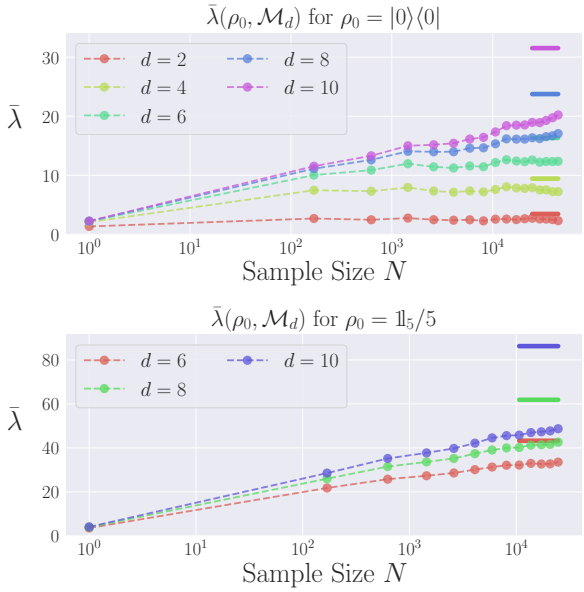


FIG. 12. The empirical average $\bar{\lambda}$ may have achieved its asymptotic value, or is still growing, depending on the true state $\rho_0$ and the model dimension $d$. Solid lines indicate the value of our formula for the asymptotic expected value, given in Equation (26).

However, our model *does* get some of the qualitative features correct. In Figure 13, we look at $\langle\lambda_{jk}\rangle$ for an isotropic Fisher information and for simulated heterodyne tomography. While the values of $\langle\lambda_{jk}\rangle$ do not agree exactly, they still decompose into two types, the "L" and the "kite". (See Figure 14 for an analysis of the discrepancies.)

## VI. CONCLUSIONS AND DISCUSSION

Local asymptotic normality, a key property satisfied by classical statistical models, is not satisfied by quantum state space. Through the introduction of metric-projected local asymptotic normality, we have provided a new framework for reasoning about classical statistical results for models which violate their assumptions.

We explicitly investigated one such result, the Wilks theorem, found it is not generally reliable in quantum state tomography, and provided a much more broadly

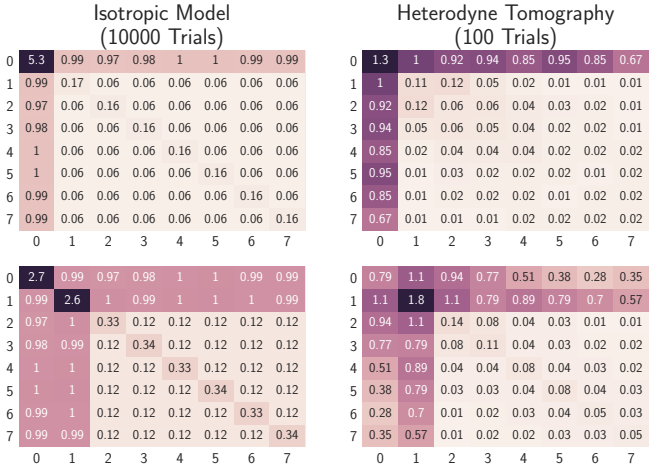


FIG. 13. The values of $\langle \lambda_{jk} \rangle$ for an isotropic Fisher information (left), and for heterodyne tomography (right). **Top**: $\rho_0 = |0\rangle\langle 0|$. **Bottom**: $\rho_0 = \mathcal{I}_2/2$. **Discussion**: Qualitatively, the behavior is the same, though there are quantitative differences, particularly within the kite.

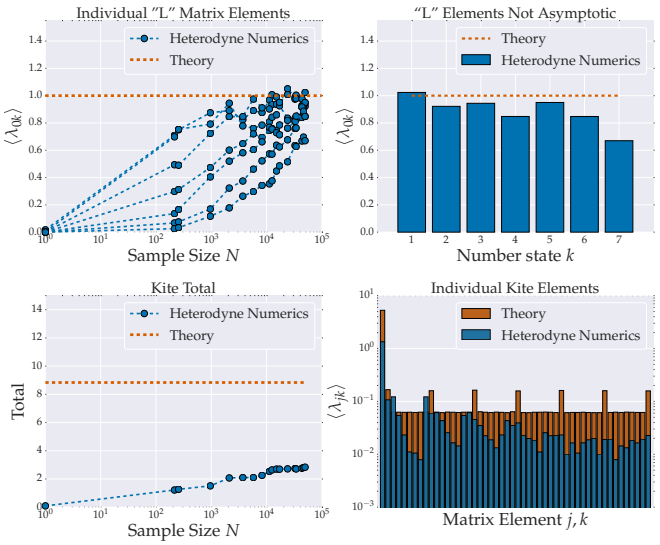


FIG. 14. Examining how our prediction for $\langle \lambda_{jk} \rangle$ disagrees with simulated heterodyne experiments. We take $\rho_0 = |0\rangle\langle 0|$ and $d = 8$. **Top Left**: The values of $\langle \lambda_{0k} \rangle$ in the "L" as a function of $N_{\text{samples}}$. **Top Right**: At the largest $N_{\text{samples}}$ studied, $\langle \lambda_{0k} \rangle$ is less than 1, especially for the higher number states. **Bottom Left**: The total from the "kite" versus $N_{\text{samples}}$. It is clear the total is still growing. **Bottom Right**: The individual "kite" elements $\langle \lambda_{jk} \rangle$ at the largest $N_{\text{samples}}$ studied; most are small compared to their values in the isotropic case.

applicable replacement that can be used in model selection methods (Equation (26)). This includes information criteria such as the AIC and BIC [17, 29, 38, 39] that do not explicitly use the Wilks theorem, but rely on the same assumptions (local asymptotic normality, equivalence between loglikelihood and squared error, etc.). Null theories of loglikelihood ratios have many other applications, including hypothesis testing [27, 33] and confidence regions [59], and our result is directly applicable to them. Refs. [27, 59] 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 unexpected behavior, indicating that quantum tomography can still surprise, and may violate *all* asymptotic statistics results. In such cases, bootstrapping [60, 61] may be the only reliable way to construct null theories for $\lambda$.

Finally, the *methods* presented here have application beyond the analysis of loglikelihoods. Metric-projected local asymptotic normality provides a new and rigorous framework for describing the behavior of the maximum likelihood estimator in the presence of constraints. This framework helps shed light on the behavior of $\hat{\rho}_{\text{ML},d}$ 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 [19, 62–64].

## VII. ACKNOWLEDGEMENTS:

TLS thanks Jonathan A Gross for helpful discussions on software design, coding, and statistics, John King Gamble for useful insights on parallelized computation and feedback on earlier drafts of this paper, and Anupam Mitra, and Kenneth M Rudinger, and Daniel Suess for proofreading edits. The authors are grateful for those who provide support for the following software packages: iPython/Jupyter [65], matplotlib [66], mpi4py [67], NumPy [68], pandas [69], Python 2.7 [70], SciPy [71], seaborn [72], and SymPy [73].

Sandia National Laboratories is a multi-mission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC., a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA-0003525.

[1] J. V. Corbett, Reports on Mathematical Physics **57**, 53 (2006).
[2] W. Pauli, in *Handbuch der Physik*, Vol. XXIV, edited by H. Geiger and K. Scheel (Springer Verlag, 1933).
[3] M. G. A. Paris and J. Rehacek, eds., *Quantum State Estimation* (Springer, Berlin-Heidelberg, 2004).
[4] M. Steffen, M. Ansmann, R. C. Bialczak, N. Katz, E. Lucero, R. McDermott, M. Neeley, E. M. Weig, A. N.

Cleland, and J. M. Martinis, Science **313**, 1423 (2006).
[5] R. Blume-Kohout, J. O. S. Yin, and S. J. van Enk, Physical Review Letters **105**, 170501 (2010).
[6] S. J. van Enk, N. Lutkenhaus, and H. J. Kimble, Physical Review A **75**, 052318 (2007).
[7] A. Anis and A. I. Lvovsky, Arxiv:arXiv:1204.5936v2 , 1 (2012).
[8] N. Reid and D. R. Cox, International Statistical Review **83**, 293 (2015).
[9] L. Wasserman, *All of Statistics: A Concise Course in Statistical Inference* (Springer New York, 2004).
[10] J. B. Altepeter, E. R. Jeffrey, and P. G. Kwiat, in *Advances in Atomic, Molecular and Optical Physics*, Vol. 52 (Elsevier, 2005) pp. 105–159.
[11] J. I. Bertrand and P. Bertrand, Foundations of Physics **17**, 397 (1987).
[12] A. I. Lvovsky and M. G. Raymer, Reviews of Modern Physics **81**, 299 (2009).
[13] G. Breitenbach, S. Schiller, and J. Mlynek, Nature **387**, 471 (1997).
[14] U. Leonhardt and H. Paul, Progress in Quantum Electronics **19**, 89 (1995).
[15] F. Motzoi, J. M. Gambetta, P. Rebentrost, and F. K. Wilhelm, Physical Review Letters **103**, 110501 (2009).
[16] R. Fazio, G. Palma, and J. Siewert, Physical Review Letters **83**, 5385 (1999).
[17] K. P. Burnham, Sociological Methods & Research **33**, 261 (2004).
[18] E. J. Candes and T. Tao, IEEE Transactions on Information Theory **52**, 5406 (2006).
[19] S. T. Flammia, D. Gross, Y.-K. Liu, and J. Eisert, New Journal of Physics **14** (2012), 10.1088/1367-2630/14/9/095022.
[20] D. Suess, Ł. Rudnicki, and D. Gross, arXiv:1608.00374 (2016).
[21] A. Carpentier, J. Eisert, D. Gross, and R. Nickl, arXiv:1504.03234v2 (2015).
[22] L. Schwarz and S. J. van Enk, Physical Review A **88**, 032318 (2013).
[23] M. Guta, T. Kypraios, and I. Dryden, New Journal of Physics **14** (2012), 10.1088/1367-2630/14/10/105002.
[24] S. J. van Enk and R. Blume-Kohout, New Journal of Physics **15**, 025024 (2013).
[25] N. K. Langford, New Journal of Physics **15** (2013), 10.1088/1367-2630/15/3/035003.
[26] J. O. S. Yin and S. J. van Enk, Physical Review A **83**, 062110 (2011).
[27] T. Moroder, M. Kleinmann, P. Schindler, T. Monz, O. Guhne, and R. Blatt, Physical Review Letters **110**, 180401 (2013).
[28] L. Knips, C. Schwemmer, N. Klein, J. Reuter, G. Tóth, and H. Weinfurter, arXiv:1512.06866 (2015).
[29] H. Akaike, IEEE Transactions on Automatic Control **19**, 716 (1974).
[30] The index $j$ may be continuous or discrete.
[31] Z. Hradil, Physical Review A **55**, R1561 (1997).
[32] D. F. V. James, P. G. Kwiat, W. J. Munro, and A. G. White, Physical Review A **64**, 052312 (2001).
[33] R. J. Blume-Kohout, Physical Review Letters **105**, 200504 (2010).
[34] J. Neyman and E. S. Pearson, Philosophical Transactions of the Royal Society of London **231**, 289 (1933).
[35] S. S. Wilks, The Annals of Mathematical Statistics **9**, 60 (1938).
[36] L. Le Cam, Annals of Mathematical Statistics **41**, 802 (1970).
[37] L. Le Cam, in *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, edited by J. Neyman (1956) pp. 129–156.
[38] G. Schwarz, The Annals of Statistics **6**, 461 (1978).
[39] R. E. Kass and A. E. Raftery, Journal of the American Statistical Association **90**, 773 (1995).
[40] D. J. Spiegelhalter, N. G. Best, B. P. Carlin, and A. Van Der Linde, Journal of the Royal Statistical Society: Series B (Statistical Methodology) **64** (2002), 10.1111/1467-9868.00353.
[41] K. Vogel and H. Risken, Physical Review A **40**, 2847 (1989).
[42] R. J. Blume-Kohout, New Journal of Physics **12**, 043034 (2010).
[43] H. Zhu, Physical Review A **90**, 012115 (2014).
[44] C. Ferrie and R. J. Blume-Kohout, Physical Review Letters **116**, 090407 (2016).
[45] S. Boyd, *Convex Optimization*, 7th ed. (Cambridge University Press, New York, New York, USA, 2009).
[46] G. G. Roussas and D. Bhattacharya, in *Advances in Directional and Linear Statistics* (Springer Physica-Verlag HD, 2010) pp. 253–280.
[47] P. Jeganathan, The Indian Journal of Statistics, Series A **44** (1982).
[48] M. B. McCoy and J. A. Tropp, Discrete and Computational Geometry **51**, 926 (2014).
[49] D. Amelunxen and M. Lotz, Information and Inference **1**, 1 (2014).
[50] R. T. Rockafellar and R. J.-B. Wets, *Variational Analysis*, 1st ed. (Springer-Verlag Berlin Heidelberg, 1998).
[51] J. A. Smolin, J. M. Gambetta, and G. Smith, Physical Review Letters **108**, 070502 (2012).
[52] Y. V. Fyodorov, Arxiv:math-ph/0412017 (2004).
[53] E. P. Wigner, Annals of Mathematics **67**, 325 (1958).
[54] T. Tao and V. Vu, Random Matrices: Theory and Applications **2** (2013), 10.1142/S201032631350007X.
[55] A. W. van der Vaart, *Asymptotic Statistics* (Cambridge University Press, 2000).
[56] A. I. Lvovsky, H. Hansen, T. Aichele, O. Benson, J. Mlynek, and S. Schiller, Physical Review Letters **87**, 050402 (2001).
[57] K. Husimi, in *Proceedings of the Physico-Mathematical Society of Japan*, Vol. 22 (1940) pp. 264 – 314.
[58] The model $\mathcal{M}_1$ is trivial, as $\mathcal{M}_1 = \{|0\rangle\langle 0|\}$. This model will almost always be wrong, in general.
[59] S. Glancy, E. Knill, and M. Girard, New Journal of Physics **14** (2012), 10.1088/1367-2630/14/9/095017.
[60] B. Efron, The Annals of Statistics **7**, 1 (1979).
[61] J. Higgins, *An Introduction to Modern Nonparametric Statistics* (Brooks/Cole, 2004).
[62] A. Steffens, C. Riofrio, W. McCutcheon, I. Roth, B. A. Bell, A. McMillan, M. S. Tame, J. G. Rarity, and J. Eisert, arXiv:1611.01189 (2016).
[63] A. Kalev and C. H. Baldwin, arXiv:1511.01433v1 (2015).
[64] A. Kalev, R. L. Kosut, and I. H. Deutsch, npj Quantum Information **1**, 15018 (2015).
[65] F. Pérez and B. E. Granger, Computing in Science and Engineering **9**, 21 (2007).
[66] J. D. Hunter, Computing in Science and Engineering **9**, 90 (2007).
[67] L. Dalcin, R. Paz, and M. Storti, Journal of Parallel and Distributed Computing **65**, 1108 (2005).

[68] S. Van Der Walt, S. C. Colbert, and G. Varoquaux, Computing in Science and Engineering **13**, 22 (2011).
[69] W. McKinney, in *Proceedings of the 9th Python in Science Conference*, edited by S. van der Walt and J. Millman (2010) pp. 51–56.
[70] G. van Rossum, "Python Language Reference," (1995).
[71] T. E. Oliphant, Computing in Science and Engineering **9**, 10 (2007).
[72] M. Waskom, "seaborn," (2016).
[73] A. Meurer, PeerJ Computer Science **3**, e103 (2017).