OPEN ACCESS

# Optimal, reliable estimation of quantum states

New Journal of Physics
The open-access journal for physics

# Optimal, reliable estimation of quantum states

**Robin Blume-Kohout**[1]
Institute for Quantum Information, Caltech 107-81, Pasadena, CA 91125, USA
E-mail: robin@blumekohout.com

*New Journal of Physics* **12** (2010) 043034 (25pp)
Received 7 December 2009
Published 20 April 2010
Online at http://www.njp.org/
doi:10.1088/1367-2630/12/4/043034

**Abstract.** Accurately inferring the state of a quantum device from the results of measurements is a crucial task in building quantum information processing hardware. The predominant state estimation procedure, maximum likelihood estimation (MLE), generally reports an estimate with zero eigenvalues. These cannot be justified. Furthermore, the MLE estimate is incompatible with error bars, so conclusions drawn from it are suspect. I propose an alternative procedure, Bayesian mean estimation (BME). BME never yields zero eigenvalues, its eigenvalues provide a bound on their own uncertainties, and under certain circumstances it is provably the most accurate procedure possible. I show how to implement BME numerically, and how to obtain natural error bars that are compatible with the estimate. Finally, I briefly discuss the differences between Bayesian and frequentist estimation techniques.

[1] Current address: Perimeter Institute for Theoretical Physics, Waterloo, ON N2L2Y5, Canada.

**Contents**



One of the prerequisites for quantum computing is 'the ability to initialize the state of the qubits to a simple fiducial state, such as $|000\ldots\rangle$' [1]. The device that prepares such a state must be tested and characterized, either to confirm that it reliably produces $|000\ldots\rangle$ or to determine what state it *does* produce, so that it can be tuned to emit the desired one. This task, of experimentally finding a density matrix $\hat{\rho}$ to describe the output of a quantum device, is *quantum state estimation*.

State estimation is more generally useful than it may appear. Two of the other quantum computing building blocks listed in DiVincenzo's seminal paper [1] (low-noise universal gates and minimal decoherence) refer to quantum *processes*. Quantum process estimation, used to characterize gates and decoherence, is mathematically equivalent to state estimation [2]. Thanks to quantum error correction and fault tolerant design, states (e.g. of the ancillae used for error correction) and gates for quantum computing need not be perfect, nor does the designer have to characterize them with infinite precision. They must function correctly with probability at least $1-\epsilon$, where $\epsilon$ (the fault tolerance threshold) is thought to be somewhere between $10^{-5}$ [3] and $10^{-2}$ [4]. A procedure for state estimation must accurately estimate probabilities of the order of $\epsilon$ and must provide a reliable bound on the uncertainty in the estimate.

Maximum likelihood estimation (MLE), based on the principle that the best estimate is the state $\hat{\rho}$ that maximizes the probability of the observed data, is the current procedure of choice. Unfortunately, it has serious flaws. It typically yields a rank-deficient estimate, with one or more zero eigenvalues. Such an estimate is implausible, implying that some measurement outcome is literally impossible. No finite amount of data can justify such certainty. More importantly,

it is impossible to bracket a zero probability with consistent error bars[2]. The MLE estimate is at best sub-optimal, and at worst dangerously unreliable (implying, for instance, that certain errors can be ruled out).

Bayesian mean estimation (BME) is an alternative procedure that avoids these pitfalls. Unlike MLE, which seeks a unique maximally plausible state, BME considers other states that are only slightly less plausible. The simple underlying principle is that the best estimate is an average over all states $\rho$ consistent with the data, weighted by their likelihood. The BME estimate is always full-rank, and comes equipped with a natural set of error bars. Moreover, each eigenvalue $\lambda$ of $\hat{\rho}_{\mathrm{BME}}$ yields an upper bound on its own uncertainty ($\Delta\lambda^2 \leqslant \lambda$). Best of all, BME is provably the most accurate scheme possible [5], under certain reasonable assumptions.

The body of this paper is divided into three sections. Section 1 explains the problems with MLE. Section 2 presents and analyzes the BME algorithm, along with one possible implementation. Section 3 discusses some unsolved problems.

## 1. The state of the art

The oldest and simplest estimation procedure is 'quantum state tomography'. In tomography, the estimator repeatedly measures several observables, records the frequencies of the outcomes and identifies the outcomes' frequencies with their probabilities. Inverting Born's rule yields a unique density matrix $\hat{\rho}_{\mathrm{tomo}}$ that predicts these probabilities. The most important problem with tomography is that $\hat{\rho}_{\mathrm{tomo}}$ often has negative eigenvalues, which means that it cannot represent a physical state.

In 1996, Hradil proposed MLE as a more flexible and sophisticated approach [6]. An estimate $\hat{\rho}$ is a theory about the unknown state. Statisticians define the *likelihood* of a theory, $\mathcal{L}(\rho)$, as the probability that the theory ($\rho$) would have predicted for the observed data ($\mathcal{M}$) before the experiment took place:

$$\mathcal{L}(\rho) \equiv p(\mathcal{M}|\rho). \tag{1}$$

Thus, $\hat{\rho}_{\mathrm{MLE}}$ is simply the $\rho$ that maximizes $\mathcal{L}(\cdot)$—i.e. the most 'likely' state. $\mathcal{L}(\rho)$ is *not* a probability distribution over $\rho$, so $\hat{\rho}_{\mathrm{MLE}}$ is not in any well-defined sense the most probable state. Actually finding $\hat{\rho}_{\mathrm{MLE}}$ requires numerics, but several algorithms exist [7]. MLE was successfully applied in 2001 to a quantum optics experiment [8] and has been used extensively since then.

MLE has some critical flaws. The most visible is that $\hat{\rho}_{\mathrm{MLE}}$ can be rank-deficient. If $|\psi\rangle$ is the eigenstate corresponding to a zero eigenvalue, then $\langle\psi|\rho|\psi\rangle = 0$. Such an estimate, although perhaps not unphysical, is *implausible*—i.e. no experimentalist would believe it. It predicts exactly zero probability for every measurement outcome $|\psi\rangle\langle\psi|$ such that $\langle\psi|\rho|\psi\rangle = 0$. This

[2] The savvy reader may be wondering about the error-estimation procedure proposed for MLE in [6] and used (e.g.) in [27]. These error bars are the variance of many MLE estimates, on many datasets obtained by simulating measurements on the original $\hat{\rho}_{\mathrm{MLE}}$. They are simply not compatible with the (rank-deficient) estimate. Are they *good* error bars, i.e. accurate, and compatible with some better estimate? The short answer is *no*, because this procedure computes the Fisher information matrix *at* $\hat{\rho}_{\mathrm{MLE}}$, not the true state. For example, suppose we flip a coin once (getting, w/o l.o.g., 'heads'), and use MLE to estimate $\hat{p}_{\mathrm{heads}} = 1$. All the simulated measurements yield 'heads', leading to the absurd conclusion that $\Delta p = 0$.

implies *absolute* certainty that $|\psi\rangle\langle\psi|$ will not be observed, which cannot be justified by a finite amount of data. If $N$ observations with $d$ possible outcomes are available, then the lowest defensible probability estimate for any event is roughly[3]

$$\hat{p}_{\min} \approx \frac{1}{N+d}. \tag{2}$$

This is a practical concern. One of the seminal papers on MLE, James *et al* [8], estimated the polarization state of two entangled photons, produced by parametric downconversion. The estimated $4 \times 4$ density matrix has two eigenvalues that are exactly zero. More recently, MLE was used to estimate the entangled state of eight ionic qubits in a trap [9]. Of 256 eigenvalues, more than 200 are less than $1/N$ (about $10^6$ measurements were made), and at least 80 are zero (to within machine precision).

Zero eigenvalues are just the most extreme illustration of a more general problem; $\hat{\rho}_{\mathrm{MLE}}$ implies predictions that cannot be justified by the data. After 100 observations, $p = 10^{-8}$ is no more credible than $p = 0$. To put it another way, taking $\hat{\rho}_{\mathrm{MLE}}$ seriously might be exceedingly embarrassing in light of *further* data. Viewed this way, zero eigenvalues are just a symptom of the larger problem. However, they motivate some useful questions that lend structure to this analysis:

1. Why are zero eigenvalues a problem?
2. Why does MLE produce zero eigenvalues?
3. What is the underlying problem with MLE?

### *1.1. Why are zero eigenvalues a problem?*

A quantum state is nothing more or less than a prediction of the future. Like a classical probability distribution, it predicts probabilities for all measurements that could be performed. A state estimate is the estimator's best prediction of what future experimentalists will find when they observe a copy of the estimated system. We should therefore evaluate an estimate on how well it predicts the future.

Quantitative evaluation of an estimate is a matter for debate. Statisticians disagree about how to interpret even a simple statement about a coin flip: 'The probability of observing 'tails' is $p_{\mathrm{tails}} = 3/4$'. However, it is indisputable that '$p_{\mathrm{tails}} = 0$' implies that 'tails' will never be observed, and that '$p_{\mathrm{tails}} = 1$' implies that nothing but 'tails' will ever be observed. Such a statement has the force and status of a mathematical theorem, just like 'There is no largest prime number'.

An estimator should hardly claim '$p_{\mathrm{tails}} = 0$' just because he/she has never observed 'tails'. He/she has a finite number of observations to work from, and $p_{\mathrm{tails}}$ might simply be very small. For example, if the data comprise a single flip, then at least one of the possible outcomes will never have been observed, but this does not justify asserting that it will *never* occur. Even if a dozen trials all yield heads, '$p_{\mathrm{tails}} = 0$' is unjustified. No matter how many data points the estimator has, we can always imagine a much larger dataset in the future, which might (embarrassingly) debunk the prediction '$p = 0$'. Thus, data can never justify reporting $\hat{p} = 0$. Only prior knowledge, such as an impossibility theorem, can do so.

[3] At first glance, $\hat{p}_{\min} = 1/N$ might seem more natural. However, consider rolling a 100-sided die just twice. Assigning $\hat{p} \approx 1/N = \frac{1}{2}$ to 98 unobserved outcomes is impossible, but $\hat{p} \approx \frac{1}{102}$ makes sense.

One might object that an estimate carries with it an implied uncertainty. For instance, $\hat{p} = 0.5$ is clearly a decent estimate of $p = 0.51$; why is $\hat{p} = 0$ not an equally good estimate of $p = 0.01$? The reason is that zero probabilities are not compatible with *any* error bars[4]. The estimate $\hat{p} = 0.5$ could mean $\hat{p} = 0.5 \pm 0.01$, meaning '$p$ is probably between 0.49 and 0.51'. To report $\hat{p} = 0 \pm 0.01$, however, is nonsensical. This would mean '$p$ is probably between $-0.01$ and 0.01', but because $p$ must be non-negative, an unconditionally better description is '$p$ is probably between 0 and 0.01', or $\hat{p} = 0.05 \pm 0.05$.

This is not the only way of representing '$p$ is probably between 0 and 0.01'. If the estimator's confidence is skewed toward one side of the interval, then the best $\hat{p}$ might not be at its center. However, it should necessarily be within the interval, not on its boundary. An estimate on the boundary cannot be optimal, because moving the estimate inside the boundary by some tiny $\epsilon$ improves it (even if the optimal $\epsilon$ is unknown). Since $p = 0$ is on the boundary of any interval, $\hat{p} = 0$ is only optimal when the confidence interval has zero width. Taken seriously, a zero probability thus implies both: (i) absolute certainty about the outcomes of future measurements, and (ii) absolute certainty about the probability itself.

This has practical consequences. If we accept that zero probabilities are implausible, then each zero eigenvalue in $\hat{\rho}$ should be replaced by a small, but finite, $\epsilon$. This poses two substantial problems. Firstly, what is $\epsilon$? It clearly declines with $N$, but whether it should scale as $1/N$ or $1/\sqrt{N}$ is unclear. Moreover, when statistics from many distinct observables are collated, it is not clear what $N$ is. Secondly, how does 'fixing' $\hat{\rho}$'s small eigenvalues affect its large eigenvalues? Since $\mathrm{Tr}\,\hat{\rho} = 1$ is fixed, increasing many small eigenvalues will require decreasing the largest ones. These large eigenvalues are critical to most of the quantities of interest—entanglement, gate fidelity, etc. The only way to resolve this messy situation is to avoid zero eigenvalues in the first place.

### *1.2. Why does MLE produce zero eigenvalues?*

The zero eigenvalues in $\hat{\rho}_{\mathrm{MLE}}$ are connected to the negativity of tomographic estimates. What I will show in this section is that, for a given dataset, if $\hat{\rho}_{\mathrm{tomo}}$ is not positive, then $\hat{\rho}_{\mathrm{MLE}}$ is rank-deficient. On the other hand, if the tomographic estimate is positive, then $\hat{\rho}_{\mathrm{MLE}} = \hat{\rho}_{\mathrm{tomo}}$. MLE is thus a sort of 'corrected tomography'[5].

The valid *state-set*, comprising all positive density matrices, is a convex subset of Hilbert–Schmidt space, the $(d^2 - 1)$-dimensional vector space of Hermitian, trace-1 matrices.

[4] Here is an alternative explanation of why an error of $\Delta p = 0.01$ is acceptable for $\hat{p} = 0.5$, but not for $\hat{p} = 0$. The canonical application of probabilities is in making bets; if event $E$ has probability $p$, then a canny bettor is justified in accepting odds of $(p^{-1} - 1) : 1$ or better against its occurrence. When $p = 0.51$, a bettor who accepts $1 : 1$ odds will slowly lose his money— but when $p = 0.01$, the bettor who believes $\hat{p} = 0$ and accepts $\infty : 1$ odds is truly courting disaster! The operational penalty for believing an (even slightly) erroneous estimate of $p = 0$ is, indeed, severe. Of course, this argument is moot if the estimator's purpose in reporting a probability is not to predict the future in any sense (in which case, generalized gambling is not a useful paradigm)— but this seems to obviate the very meaning of 'probability'.

[5] This discussion is rigorously correct only when the estimator measures a complete—rather than overcomplete—set of observables. For an overcomplete set, it is not entirely obvious what 'tomography' means. The most common answer would be a least-squares fit, in which case $\hat{\rho}_{\mathrm{tomo}} \geqslant 0$ does not imply $\hat{\rho}_{\mathrm{tomo}} = \hat{\rho}_{\mathrm{MLE}}$. The general conclusions still hold, however, and rigor can be retained by defining 'tomography' to mean unconstrained MLE, for overcomplete data.

Its boundary comprises the rank-deficient states. Whenever $\hat{\rho}_{\text{tomo}}$ lies outside this boundary, MLE squashes it down onto the boundary, producing a rank-deficient estimate.

*1.2.1. How tomography works.* Quantum state tomography is based on inverting Born's rule: If a positive operator valued measure (POVM) measurement $\mathcal{P} = \{E_1 \ldots E_N\}$ is performed on a system in state $\rho$, then the probability of observing $E_i$ is $p_i = \text{Tr}(E_i \rho)$. The probabilities for $d^2$ linearly independent outcomes single out a unique $\hat{\rho}_{\text{tomo}}$ consistent with those probabilities. Several projective measurements (at least $d+1$) can, in aggregate, form a *quorum*—i.e. provide sufficient information to identify $\hat{\rho}_{\text{tomo}}$.

Note, however, that no measurement can reveal the probability of an event. Repeated measurements yield frequencies, from which the tomographic estimator infers probabilities. The measurement is repeated $N$ times, and if outcome $E_i$ appears $n_i$ times, we estimate $\hat{p}_i = n_i/N$. If the measurements form a quorum, then the equations

$$\text{Tr}\left(\hat{\rho}_{\text{tomo}} E_i\right) = \frac{n_i}{N} \tag{3}$$

can be solved to yield a unique $\hat{\rho}_{\text{tomo}}$.

Tomography thus seeks a density matrix whose predictions agree exactly with the observed frequencies. Unfortunately, this matrix is not always a state. Suppose that an experimentalist, estimating the state of a single qubit, measures $\sigma_x$, $\sigma_y$ and $\sigma_z$—but only one time each! Having observed the +1 result in each case, he/she seeks a $\hat{\rho}_{\text{tomo}}$ satisfying $\langle\sigma_x\rangle = \langle\sigma_y\rangle = \langle\sigma_z\rangle = 1$. Such a matrix exists,

$$\hat{\rho}_{\text{tomo}} = \begin{pmatrix} 1 & \dfrac{1+i}{2} \\ \dfrac{1-i}{2} & 0 \end{pmatrix}, \tag{4}$$

but it has a negative eigenvalue $\lambda = 1 - \sqrt{3}/2 \approx -0.366$. Moreover, this 'state' implies that all three spin measurements would be perfectly predictable, which is impossible.

Estimating the state from a single measurement of each basis is a rather extreme example. However, it illustrates a point. Tomography, in a single-minded quest to match Born's rule to observed frequencies, pays no attention to positivity. As the number of measurements ($N$) increases, the possible tomographic estimates form an $N \times N \times N$ grid. They fill a 'Bloch cube', defined by $\langle\sigma_{x,y,z}\rangle \in [-1 \ldots 1]$, which circumscribes the Bloch sphere and contains a lot of negative states (see figure 1). If the true state is sufficiently pure, then the probability of obtaining a negative estimate can remain as high as 50% for arbitrarily large $N$, since the true state is very close to the boundary between physical and unphysical states.

In larger Hilbert spaces, the problem gets worse for two reasons. Firstly, the state-set's dimensionality (and therefore the number of independent parameters in $\rho$) grows as $d^2 - 1$. In order to keep the RMS error ($\Delta_2 = \sqrt{\text{Tr}\left[(\hat{\rho}_{\text{tomo}} - \rho)^2\right]}$) fixed, $N$ must grow proportional to $d$. Secondly, $\hat{\rho}_{\text{tomo}}$ has more eigenvalues, so the probability of at least one negative eigenvalue grows with $d$ (for fixed $\Delta_2$). Together, these scalings ensure that tomographic estimates of large systems are rarely non-negative.

The problems with tomography are well known—negative eigenvalues were precisely the embarrassing feature that motivated MLE. As we shall see, however, MLE's implausible zero eigenvalues are closely related to tomography's negative ones.

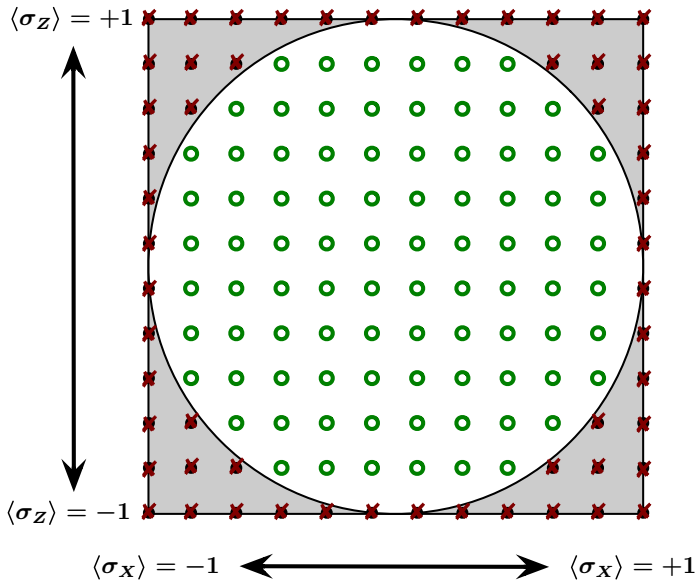


**Figure 1.** A cross section of the 'Bloch cube', which contains all possible tomographic estimates, and circumscribes the Bloch sphere containing all positive estimates. The points shown are possible tomographic estimates for $N = 11$ measurements each of $\sigma_x$ and $\sigma_z$, with $\langle\sigma_y\rangle$ set to zero for the sake of simplicity. Of the 144 $\hat{\rho}_{\text{tomo}}$ shown, 54 are non-positive (it should be kept in mind that the $\sigma_y$ dimension is ignored). Depending on the state, some $\hat{\rho}_{\text{tomo}}$ will of course be more likely than others; this figure merely illustrates the array of possible non-positive estimates.

*1.2.2. How MLE works.* MLE, although sometimes complex in implementation, is very simple in theory. Given a measurement record $\mathcal{M} = \{M_1, M_2, M_3 \ldots M_N\}$ (where $M_i$ is a positive operator representing the $i$th observation), the estimator seeks the maximum of the likelihood function,

$$\mathcal{L}(\rho) = p(\mathcal{M}|\rho) = \prod_i (\text{Tr}[M_i\rho]). \tag{5}$$

$\mathcal{M}$ can be compactly represented as a list of frequencies. Define a set $\mathcal{P} = \{E_1 \ldots E_m\}$ containing all possible outcomes, and let $n_i$ be the number of times that $E_i$ appears in $\mathcal{M}$. Then $\mathcal{M} \sim \{n_1 \ldots n_m\}$. As $N$ increases, the frequency representation of $\mathcal{M}$ remains short.

Finding $\hat{\rho}_{\text{MLE}}$ is feasible because $\mathcal{L}(\rho)$ has two convenient properties. Firstly, it is non-negative, so we can maximize $\log(\mathcal{L}(\rho))$. Secondly, $\log(\mathcal{L}(\rho))$ is convex. The proof is quite simple: we observe that $\log(\mathcal{L}(\rho)) = \sum_i \log \text{Tr}[M_i\rho]$; that $\text{Tr}[M_i\rho]$ is a non-negative, linear function of $\rho$; that the logarithm of a linear function is convex; and that the sum of convex functions is convex. Among other things, this means that $\mathcal{L}(\rho)$ has a unique local maximum.

*1.2.3. The relationship between tomography and MLE.* The likelihood function has another elegant property: If there is a state $\hat{\rho}_{\text{tomo}}$, such that the probability predicted for every outcome is equal to its observed frequency, then $\hat{\rho}_{\text{tomo}}$ is the maximum of $\mathcal{L}(\rho)$. To prove this, let us write

log $\mathcal{L}(\rho)$ in terms of (a) the observed frequencies ($f_j = n_j/N$) and (b) the predicted probabilities ($p_j = \mathrm{Tr}[E_j\rho]$) for all $E_j$:

$$\mathcal{L}(\rho) = \prod_i (\mathrm{Tr}[M_i\rho]) = \prod_j \mathrm{Tr}[E_j\rho]^{n_j}, \tag{6}$$

$$\log(\mathcal{L}(\rho)) = \sum_j n_j \log\big(\mathrm{Tr}[E_j\rho]\big) \tag{7}$$

$$= N\sum_j f_j \log p_j \tag{8}$$

$$= N\sum_j \left[f_j \log f_j - (f_j \log f_j - f_j \log p_j)\right]$$

$$= -N\left[H(f) + D(f||p)\right]. \tag{9}$$

The last line invokes two information-theoretic quantities, entropy $H(\cdot)$ and relative entropy $D(\cdot||\cdot)$. $H(f)$ does not depend on $p$, so it is irrelevant for maximization. The relevant quantity is $D(f||p)$, which is always non-negative, and uniquely zero when $p = f$. Thus, $\log(\mathcal{L}(\rho))$ is uniquely maximized when $p = f$. □

So, if $\hat{\rho}_{\mathrm{tomo}}$ is a valid state, then $\hat{\rho}_{\mathrm{MLE}} = \hat{\rho}_{\mathrm{tomo}}$. What if $\hat{\rho}_{\mathrm{tomo}}$ exists, but is not a valid state? It must still be Hermitian and have unit trace. Furthermore, it predicts non-negative probability for each $M_i$ observed, so $\mathrm{Tr}[E_i\hat{\rho}_{\mathrm{tomo}}] \geqslant 0$ for all $i$. The hyperplanes $\mathrm{Tr}[E_i\rho] = 0$ define a polytope in Hilbert–Schmidt space—a simple example is the 'Bloch cube' in figure 1—which contains $\hat{\rho}_{\mathrm{tomo}}$.

If we extend the domain of $\mathcal{L}(\rho)$ to this polytope and its interior, then its maximum must coincide with $\hat{\rho}_{\mathrm{tomo}}$, since $\hat{\rho}_{\mathrm{tomo}}$ predicts the correct frequencies. Tomography, in other words, is essentially unconstrained MLE.

Because $\mathcal{L}(\rho)$ has a unique local maximum at $\hat{\rho}_{\mathrm{tomo}}$, its maximum over a closed region that does not contain $\hat{\rho}_{\mathrm{tomo}}$ must lie on the boundary of that region (see figure 2). The set of non-negative density matrices is precisely such a closed region, so whenever $\hat{\rho}_{\mathrm{tomo}}$ is not a valid state, $\hat{\rho}_{\mathrm{MLE}}$ must lie on the boundary of the state-set. That is, it will be rank-deficient.

MLE and tomography are thus variants of the same procedure, distinguished only by the positivity constraint[6]. MLE is a sort of minimal fix for tomography, returning the non-negative state that is in some sense 'closest' to $\hat{\rho}_{\mathrm{tomo}}$. Actually computing the number of zero eigenvalues in $\hat{\rho}_{\mathrm{MLE}}$ seems difficult, but numerical exploration for 1, 2, 3 and 4 qubit problems suggests that $\hat{\rho}_{\mathrm{MLE}}$ usually has at least as many zero eigenvalues as $\hat{\rho}_{\mathrm{tomo}}$ has negative ones. In conjunction with the observation that large-system tomography tends to yield many negative eigenvalues, this explains the many zero eigenvalues in experimental applications of MLE.

[6] In fact, tomography does have a *de facto* positivity constraint; all the probabilities for observed events must be non-negative. Quantum mechanics, on the other hand, demands that all the probabilities for any event that could ever be observed must be non-negative. These distinct constraints lead, respectively, to the 'Bloch polytope' and to the Bloch sphere, as the set of valid states. This distinction between observed and observable events is what undermines frequentism in quantum estimation.

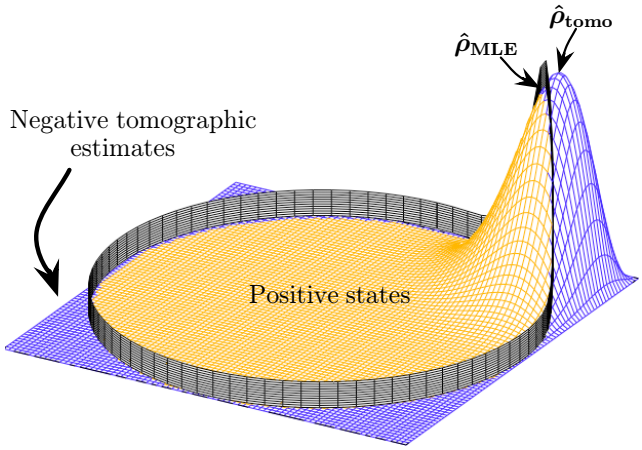


**Figure 2.** An example of a likelihood function (for a single qubit) whose unconstrained maximum lies outside the state-set and whose constrained maximum therefore lies on its boundary. The domain shown here is a cross section of the Bloch sphere, with $\langle\sigma_y\rangle = 0$. This particular likelihood function is obtained from 16 measurements each of $\sigma_x$ and $\sigma_z$, comprising 14 $|1\rangle$ and $|+\rangle$ results and 2 $|0\rangle$ and $|-\rangle$ results. The unconstrained maximum of $\mathcal{L}(\rho)$ is at $\hat{\rho}_{\mathrm{tomo}} = \frac{1}{2}\left(\mathbb{1} + \frac{3}{4}\sigma_x + \frac{3}{4}\sigma_z\right)$, which has a negative eigenvalue. The constrained maximum is at $\hat{\rho}_{\mathrm{MLE}} = |\psi\rangle\langle\psi|$, where $|\psi\rangle = (2+\sqrt{2})^{-1/2}$.

### *1.3. What is the underlying flaw?*

Tomography and MLE maximize $\mathcal{L}(\rho)$ over different domains. They display the same pathology, implying unjustifiable (zero or negative) probabilities. The underlying problem is simple: maximum likelihood methods are frequentistic; they interpret observed frequencies as probabilities. By maximizing $\mathcal{L}(\rho)$, they seek to fit the observed frequencies as precisely as possible. If there exists a $\hat{\rho}$ that fits the data exactly, then that is always the best estimate.

The point of state estimation, however, is not solely to explain the data. Rather, it is to find a state that will predict the future. It should concisely describe what the estimator knows about the system being estimated. Mindless data fitting accomplishes only retrodiction, of the past. The best description of the past (i.e. data) probably does not describe the estimator's knowledge, especially his/her uncertainty.

For example, consider estimating the bias of a coin after flipping it just once. The best fit to the data is to assume that the coin always comes up the same way. This clearly does not describe the estimator's knowledge—an honest description would perhaps include the word 'scant'. Ironically, it is the high entropy of the estimator's knowledge that causes a spuriously low-entropy estimate.

The problem with MLE is that it matches probabilities to observed frequencies, consistent with frequentist statistics. This is actually unfair to frequentism, which begins by defining probability as the infinite-sample-size limit of frequency. A true frequentist avoids making statements about probabilities in the absence of an infinite ensemble, so applying a frequentist method to relatively small amounts of data is inherently disaster-prone. Nonetheless, this is precisely what is happening in MLE. For further discussion, see section 2.4.

## 2. BME

Bayesian methods provide a different perspective on statistics (see, e.g., [10] as a general reference). Bayesian mean estimation (BME) avoids the pitfalls of MLE. Here are three basic tenets, each of which independently motivates BME:

1. *Consider all the possibilities*. MLE identifies the best fit to observed data, but many nearby states are almost equally likely. An honest estimate should incorporate these alternatives.
2. *Demand error bars*. The estimate should be compatible with error bars, e.g. $\hat{\rho} \pm \Delta\rho$. This implies a region containing most of the plausible states, of size $\Delta\rho$, with $\hat{\rho}$ somewhere around the center. If $\hat{\rho}$ is rank-deficient, no such region exists. Thus, $\hat{\rho}$ should lie far enough from the state-set's boundary to be compatible with well-motivated error bars.
3. *Optimize accuracy*. Obviously, the estimate $\hat{\rho}$ should be close to the true $\rho$. How do we evaluate this? Quantum strictly proper scoring rules [5] yield a class of metrics designed to measure this closeness, called 'operational divergences'. BME uniquely minimizes the expected value of every operational divergence.

Each of these motivations illustrates one of BME's major advantages. The estimate predicts reliable probabilities for all measurement outcomes, it comes with a free set of error bars, and it is (on average) the most accurate estimate that can be made from the data.

Bayesian approaches have been previously discussed in various contexts. Helstrom [11] applied Bayesian methodology extensively to estimation. He considered a variety of utility functions, especially the rather pathological $\delta$-function utility that motivates MLE, without paying particular attention to the posterior mean. Jones [12] applied Bayesian inference with Haar measure, focusing on information-theoretic bounds. Derka *et al* [13] examined Bayesian estimation in some detail, primarily in its connections to tomography and maximum entropy. Schack *et al* [14] formalized a deep connection to exchangeable (deFinetti) states. More recently, Neri [15] considered Bayesian estimation of phase difference in coherent light. Tanaka and Komaki [16] proved the optimality of Bayesian estimation with respect to relative entropy.

My goal in this section is to propose BME as a practical procedure for state estimation, and to describe its operational advantages. I begin by concisely presenting the BME algorithm, then discuss in section 2.2 how it can be implemented. Section 2.3 analyzes the properties of $\hat{\rho}_{\mathrm{BME}}$, focusing on the three advantages asserted above. Finally, section 2.4 contrasts the Bayesian and frequentist approaches.

### 2.1. The BME algorithm

BME is conceptually simple.

1. Use the data to generate a likelihood function, $\mathcal{L}(\rho) = p(\mathcal{M}|\rho)$. $\mathcal{L}$ is not a probability distribution; it quantifies the relative plausibility of different state assignments.
2. Choose a prior distribution over states, $\pi_0(\rho)\mathrm{d}\rho$. It represents the estimator's ignorance, and should generally be chosen to be as 'uniform', or uninformative, as possible.
3. Multiply the prior by the likelihood and normalize to obtain a posterior distribution

$$\pi_f(\rho)\mathrm{d}\rho \propto \mathcal{L}(\rho)\pi_0(\rho)\mathrm{d}\rho, \tag{10}$$

which represents the estimator's knowledge. The proportionality constant is set by normalization.

4. Report the mean of this posterior,

$$\hat{\rho}_{\mathrm{BME}} = \int \rho \pi_f(\rho) \mathrm{d}\rho. \tag{11}$$

This is the best concise description of the estimator's knowledge.

### 2.2. Implementation

In practice, BME comes down to computing an integral. The best way of doing this remains uncertain, as does the existence of an exact solution. The numerical algorithm presented below has been demonstrated to work well in a small variety of cases. However, it could be improved in many ways, and has some glaring deficiencies. This algorithm should thus be taken as a proof of principle (i.e. it is possible to do Bayesian estimation) rather than an optimal approach.

An important observation for any integration procedure is that the likelihood is easy to compute. $\mathcal{L}(\rho)$ is the probability of observing a sequence of outcomes $\mathcal{M} = \{M_1 \ldots M_N\}$, given $\rho$. This is the product of the probabilities for the individual $M_i$, each of which is given by Born's rule:

$$\mathcal{L}(\rho) = \mathrm{Tr}(M_1\rho)\,\mathrm{Tr}(M_2\rho)\,\mathrm{Tr}(M_3\rho)\ldots\mathrm{Tr}(M_N\rho). \tag{12}$$

When $\mathcal{M}$ is represented using frequencies ($E_i$ was observed $n_i$ times, for $i \in [1 \ldots m]$), this can be evaluated in $O(m)$ time:

$$\mathcal{L}(\rho) = \mathrm{Tr}(E_1\rho)^{n_1}\,\mathrm{Tr}(E_2\rho)^{n_2}\ldots\mathrm{Tr}(E_m\rho)^{n_m}. \tag{13}$$

*2.2.1. The Metropolis–Hastings algorithm.* In the absence of an analytic solution to the integral, we fall back to numerical Monte Carlo methods. Because $\mathcal{L}(\rho)$ is usually a sharply peaked function over a high-dimensional space, brute-force random sampling will converge extremely slowly. Metropolis algorithms [17] were conceived for precisely such situations. A variant known as Metropolis–Hastings [18, 19] is commonly used for Bayesian estimation, and can be adapted straightforwardly to quantum states.

The Metropolis–Hastings algorithm computes the average value of a function (in this case, $\rho$) over an integration measure (in this case, $\mathcal{L}(\rho)\pi_0(\rho)\mathrm{d}\rho$ ). It leverages the fact that $\mathcal{L}(\rho)\pi_0(\rho)\mathrm{d}\rho$ is typically dominated by a small region of high likelihood. Whereas basic Monte Carlo methods jump randomly around the integration measure, Metropolis–Hastings makes local, biased jumps. This samples the most relevant parts of the sample space preferentially. After each jump, the current value of $\rho$ is added to a running tally. This tally, divided by the total number of jumps, becomes the final average.

To implement Metropolis–Hastings, we begin with a rule $J$ for jumping from any $\rho$ to a nearby $\rho' = J(\rho)$. The precise form of the rule is unimportant; it is usually stochastic, although a deterministic rule (traversing a quasi-random set) is conceivable. What is important is that $J$ should generate the underlying measure $\mathrm{d}\rho$: for any $\rho_0$, the set $\{J^n(\rho_0) : n \in [0 \ldots N]\}$ should sample uniformly from $\mathrm{d}\rho$ as $N \to \infty$. For example, we can sample from Lebesgue measure over the interval $[0 \ldots 1]$ using the rule $J(x) = (x + y) \bmod 1$, where $y$ is selected from a Gaussian distribution with zero mean and fixed variance.

Such a rule, unmodified, would compute $\int_\rho f(\rho)\mathrm{d}\rho$. To average instead over $\mathcal{L}(\rho)\pi_0(\rho)\mathrm{d}\rho$, we modify the rule as follows. After choosing $\rho'$, but before jumping to it, we compute the likelihood ratio

$$r = \frac{\mathcal{L}(\rho')\pi_0(\rho')}{\mathcal{L}(\rho)\pi_0(\rho)}. \tag{14}$$

If $r > 1$ ($\rho'$ is more likely than $\rho$), then we jump as before. If not, we jump to $\rho'$ with probability $r$, and stay at $\rho$ (adding it, once again, to the running total) with probability $1 - r$.

This biasing ensures that the algorithm spends more time at more likely spots, and tends to lurch uphill into regions of high probability. Unlike a gradient algorithm (as might be used for MLE), it does not actively seek the point of highest probability; jumping to a region of lower probability is both possible and necessary. Detailed discussion and explanation of why this works can be found in [19].

*2.2.2. Applying Metropolis–Hastings to quantum states.* The heart of the algorithm is the rule $J$. It determines $\mathrm{d}\rho$, and its form is critical to the algorithm's performance. Different underlying measures will require different rules. Measures with some claim to 'uniformity' are usually invariant under a symmetry group. The natural group for quantum states on a $d$-dimensional Hilbert space is SU($d$), and the measure that this group induces over pure states is called Haar measure. A sensible prior should extend over all possible states, so we need measures extending to mixed states. However, there is no uniquely suitable measure over mixed states, because their spectral degrees of freedom (eigenvalues) have no obvious symmetry. One appealing class of measures, proposed by Zyzkowski and Sommers [20], is the set of induced measures, denoted here by $\mathrm{d}_k\rho$. They are obtained by beginning with Haar measure on a $d \times k$ dimensional system, then tracing out the ancillary factor. Thus, $\mathrm{d}_1\rho$ is simply the Haar measure on pure states; while $\mathrm{d}_d\rho$ is the Hilbert–Schmidt measure (Lebesgue measure on the vector space of Hermitian $d \times d$ matrices).

These induced measures are easy to implement. Instead of keeping track of $\rho$ itself, we generate and track a pure state $|\psi_{d\times k}\rangle$ in $d \times k$ dimensions. At each step, $\rho$ is obtained by tracing out part of $|\psi_{d\times k}\rangle\langle\psi_{d\times k}|$. The ancillary degree of freedom acts as a sort of hidden variable, internal to the algorithm. We need only a rule $J$ to implement Haar measure over the larger Hilbert space.

This could be done in many ways—for instance, at each step, we could generate a random unitary from Haar measure. This has two huge drawbacks. Firstly, the jumps are non-local, which negates the key advantage of the Metropolis–Hastings algorithm. Secondly, generating and applying a random unitary is computationally expensive. Instead, we need a relatively small set of efficiently constructable unitaries that generate the entire group.

Here is an efficient local random walk rule that generates Haar measure on a $d$-dimensional Hilbert space:

1. Choose a direction, by generating two random integers $i, j \in [0 \ldots d-1]$. Select a Hermitian operator $H_{ij}$ that acts only on the $\{|i\rangle, |j\rangle\}$ subspace. Define $H_{ij} = \sigma_z$ if $i = j$, $H = \sigma_x$ if $i < j$ and $H = \sigma_y$ if $i > j$.
2. Choose a distance, $\delta$, from a distribution (e.g. Gaussian) with $\langle\delta\rangle = 0$ and $\langle\delta^2\rangle = \Delta^2$. We will discuss the choice of $\Delta$ below.

3. Let $J(|\psi\rangle) = e^{i\delta H_{ij}}|\psi\rangle$. Since $U$ acts nontrivially only on the $|i\rangle$, $|j\rangle$ subspace, this can be done very easily and quickly.

Each step's distance is chosen randomly to ensure uniform sampling—with a fixed step size, it is just barely conceivable that this algorithm might trace out a discrete lattice of states. The average step size $\Delta$ is important: if $\Delta$ is too large, the algorithm will not find small regions of high probability efficiently; if $\Delta$ is too small, it will explore the space very slowly. The optimal $\Delta$ will depend on $\mathcal{L}(\rho)$, and there is no way to identify it *a priori*.

The algorithm must therefore vary $\Delta$ dynamically, with feedback. If $\Delta$ is very small, then almost every jump will be accepted, whereas if $\Delta$ is large, very few will be accepted. A good heuristic is that the acceptance ratio should be around 60% (other values are also suggested [19]). The algorithm should track the acceptance ratio over the last $\sim$1000 jumps, gradually changing $\Delta$ as appropriate to maintain it around 60%.

Dynamically adjusting the step size like this can, in theory, break the convergence properties of the algorithm. This occurs if the distribution over states is multimodal; the step size is reduced in order to explore one narrow peak in detail, and a far-off peak becomes inaccessible. Fortunately, the likelihood function itself is guaranteed to be log-convex and therefore unimodal. For well-behaved priors with convex support (e.g. the Hilbert–Schmidt prior, $\mathrm{d}_d\rho$), this means that $\mathcal{L}(\rho)\pi_0(\rho)$ can safely be sampled this way.

Other priors—in particular, the Haar prior, which is interesting as a limiting case—do not have convex support. These priors will yield multimodal posterior distributions. How to effectively and reliably sample from such distributions is an open problem. Repeating the sampling many times, with randomly distributed starting points, is not reliable. It fails badly if two similar peaks in the distribution have unequally sized regions of convergence; the peak with the larger convergence region will be relatively oversampled.

### 2.3. *[Good] properties of the BME estimate*

Why should an experimentalist use BME? After all, BME (via Monte Carlo) is more computationally intensive than MLE. The answer, of course, is that BME provides a better estimate than MLE. Specifically: (i) $\hat{\rho}_{\mathrm{BME}}$'s eigenvalues are never unjustifiably small (or zero); (ii) the procedure can easily be made to yield well-motivated error bars that are compatible with $\hat{\rho}_{\mathrm{BME}}$; (iii) BME is, in a particular sense, the most accurate possible estimate—not just asymptotically, but for finite $N$.

*2.3.1. The estimate is plausible.* The first objection to MLE is that $\hat{\rho}_{\mathrm{MLE}}$ is implausible; it can (and often does) have zero eigenvalues, which imply an unjustified certainty. Any alternative procedure should yield a strictly positive estimate. BME yields just such an estimate, subject to a very weak restriction on the prior.

Consider a simple and illustrative example in classical estimation. We estimate the bias $b$ of a coin, which comes up 'heads' with probability $b$, and 'tails' with probability $1-b$.

Flipping the coin $N$ times yields a measurement record consisting of $n$ heads and $N-n$ tails. The likelihood function is

$$\mathcal{L}(b) = b^n(1-b)^{N-n}, \tag{15}$$

and so the MLE estimate is

$$\hat{b}_{\mathrm{MLE}} = \frac{n}{N}. \tag{16}$$

If $n = 0$ or $n = N$, then $\hat{b}_{\mathrm{MLE}}$ will assign zero probability to observing either 'heads' or 'tails', respectively.

If we adopt a Bayesian approach, then we must choose a prior—e.g. the uniform prior with respect to Lebesgue measure, $\pi_0(b)\mathrm{d}b = \mathrm{d}b$. The mean of the posterior is an integral of the likelihood, and we obtain

$$\hat{b}_{\mathrm{BME}} = \frac{n+1}{N+2}. \tag{17}$$

Since $0 \leqslant n \leqslant N$, the Bayesian estimator never assigns zero probability to anything. The lowest possible probability assignment for either heads or tails is $p_{\min} = 1/(N+2)$. With no data at all, the Bayesian assigns $p = \frac{1}{2}$ to both outcomes; after a single flip, he/she assigns $p = \frac{2}{3}$ to the outcome that was observed and $p = \frac{1}{3}$ to the other.

This is the property that we want in a quantum estimation procedure. The probabilities assigned to unobserved events are not only nonzero, but also sensible—after $N$ trials, it is reasonable to assume that the probability of an as-yet-unobserved outcome is at most $1/N$, and to assign $p_{\min} \approx 1/N$.

However, this property depends on the prior. Consider the prior $\pi_0(b) = \frac{1}{2}(\delta(b) + \delta(1-b))$. After one observation of 'tails', our Bayesian estimate would be $\hat{b}_{\mathrm{BME}} = 0$, which is implausible. The problem is that a finite number of observations (one) ruled out every $b$ in support of $\pi_0$ that ascribed nonzero probability to 'heads'.

The situation gets even worse if the next observation is 'heads'. The data now rule out every hypothesis, the posterior $\pi_f(\rho)\mathrm{d}\rho$ vanishes entirely, and the Bayesian procedure simply fails. This stems from a contradiction. A prior over *states* implies a probability distribution over *observations* as well. $\pi_0$ assigned exactly zero probability to $\mathcal{M} = \{\text{'heads'}, \text{'tails'}\}$—which was then observed, causing a contradiction.

The following statements about a prior $\pi_0$ are logically equivalent:

(a) $\pi_0$ assigns zero probability to some (finite-length) measurement record.

(b) Bayesian estimation using $\pi_0$ will, for some measurement record, yield an estimate with zero probability.

(c) There exists a measurement record that will annihilate $\pi_0$, so that Bayesian estimation fails completely.

Let us define a *fragile* prior as one for which these statements hold (and which can therefore yield a rank-deficient estimate). An estimator should choose a *robust* (i.e. not fragile) prior, which in turn guarantees a full-rank estimate.

In classical probability estimation, avoiding fragility is simple: a prior is robust if and only if it has support in the interior of the probability simplex. States in the interior do not predict zero probability for any observation. They can never be ruled out, so a prior supported on one can never be annihilated by the data. Conversely, every prior supported only on the boundary will be annihilated by a measurement record that includes every possible outcome.

Intriguingly, this condition does not extend to the quantum problem. Support in the interior (i.e. on the full-rank states) is sufficient, but not necessary, for robustness. Consider estimation of a single qubit using the Haar prior, which is restricted to (and uniform over) the pure states. Each observation rules out, at most, a single pure state—if $|0\rangle\langle 0|$ is observed, then the true state cannot be $|1\rangle\langle 1|$. There are uncountably many distinct candidate pure states, which means that no (finite-length) measurement record can annihilate the prior. The Haar prior is robust.

As a general rule, just about every prior that a halfway-sane estimator would pick is, in fact, robust. Not only the Haar prior (which implies absolute certainty that $\rho$ is pure), but much more extreme priors, such as an equatorial distribution on the Bloch sphere—or, for that matter, any continuous curve on the Bloch sphere's surface—are robust. The appendix demonstrates a necessary and sufficient condition.

*2.3.2. The estimate comes with natural error bars.* Another objection to the MLE procedure is that $\hat{\rho}_{\mathrm{MLE}}$ is not, in general, compatible with any error bars. This is an obvious consequence of zero eigenvalues; error bars imply a region of plausibility *surrounding* the point estimate. When the estimate lies on the state-set's boundary, no such region can exist—in order that $\hat{\rho}_{\mathrm{MLE}}$ be in its interior, the region would have to contain negative matrices.

The BME estimate is always full-rank, which is encouraging. This in itself does not guarantee compatibility with sensible error bars. The estimate $\hat{\rho} = \frac{99}{100}\hat{\rho}_{\mathrm{MLE}} + \frac{1}{100d}\mathbb{1}$ is full-rank, but the estimator's honest uncertainty about $\rho$ might well be greater than $\pm 1\%$. Happily, the BME estimation procedure can easily be adapted to yield natural error bars, which are compatible with the point estimate.

First, let us consider what form these error bars should take. Intuitively, the qualified estimate should look like

$$\rho = \hat{\rho} \pm \Delta\rho, \tag{18}$$

but what, precisely, is '$\Delta\rho$'? As $\hat{\rho}$ is a $d \times d$ matrix, we might suppose that $\Delta\rho$ is also a $d \times d$ matrix, so $\rho_{ij} = \hat{\rho}_{ij} \pm \Delta\rho_{ij}$. This fails to account for covariance between distinct elements of $\rho$. For example, the diagonal elements of $\hat{\rho}$ must vary together to maintain $\mathrm{Tr}(\rho) = 1$.

The correct way to think about the estimator's uncertainty begins by representing the estimate, $\hat{\rho}_{\mathrm{BME}}$, as a $d^2 - 1$ dimensional vector in Hilbert–Schmidt space. For a single qubit:

$$\rho \sim \begin{pmatrix} x \\ y \\ z \end{pmatrix} \equiv \begin{pmatrix} \mathrm{Tr}(\sigma_x \rho) \\ \mathrm{Tr}(\sigma_y \rho) \\ \mathrm{Tr}(\sigma_z \rho) \end{pmatrix}. \tag{19}$$

The estimator's uncertainty is represented as a symmetric covariance matrix on the same space:

$$\Delta\rho \sim \begin{pmatrix} \Delta x^2 & \Delta xy & \Delta xz \\ \Delta xy & \Delta y^2 & \Delta yz \\ \Delta xz & \Delta yz & \Delta z^2 \end{pmatrix}. \tag{20}$$

The elements of $\Delta\rho$ involve two different expectation values: one with respect to the state, denoted $\langle X \rangle_\rho \equiv \mathrm{Tr}(X\rho)$; and one with respect to the posterior probability, denoted $\overline{f} \equiv \int f(\rho)\pi(\rho)\mathrm{d}\rho$. Using this notation,

$$\Delta x^2 = \overline{\langle x \rangle^2} - \left(\overline{\langle x \rangle}\right)^2, \tag{21}$$

with the other elements given by the obvious generalization.

Represented as a covariance matrix, $\Delta\rho$ quantifies the second cumulant of the estimator's probability distribution $\pi_f(\rho)\mathrm{d}\rho$. It defines an ellipsoid in Hilbert–Schmidt space, which is a credible interval (the Bayesian version of a confidence interval). The eigenvectors of $\Delta\rho$ are

operators that define the principal axes of this ellipsoid, and the corresponding eigenvalues are their lengths.

As a matrix that acts on density matrices, $\Delta\rho$ is a superoperator. It is symmetric and non-negative, but not completely positive or trace-preserving, so it cannot be interpreted as a quantum process. However, the superoperator interpretation gives a formula for the estimator's uncertainty about the expectation value of a particular operator $X$. Defining $\Delta\rho[X]$ to be the superoperator's action on $X$,

$$\Delta \langle X\rangle^2 = \mathrm{Tr}\left(X^\dagger\ \Delta\rho[X]\right), \tag{22}$$

quantifies the estimator's expected error in $\langle X\rangle$.

Alternatively, $\Delta\rho$ can be represented as an unnormalized symmetric bipartite state,

$$\Delta\rho = \overline{\rho\otimes\rho} - \overline{\rho}\otimes\overline{\rho} \tag{23}$$

$$= \int \rho\otimes\rho\ \pi_f(\rho)\mathrm{d}\rho - \hat{\rho}_{\mathrm{BME}}\otimes\hat{\rho}_{\mathrm{BME}}, \tag{24}$$

and in this representation, the estimator's expected error in $\langle X\rangle$ is

$$\Delta \langle X\rangle^2 = \mathrm{Tr}\left(X\otimes X\Delta\rho\right). \tag{25}$$

This $\Delta\rho$ is a consistent representation of the estimator's uncertainty; for any $X$, it yields the same $\Delta\langle X\rangle^2$ that an independent estimate of $\langle X\rangle$ would. To see this, let $X$ be an arbitrary observable with eigenvalues between $x_{\min}$ and $x_{\max}$. The variance computed via BME is

$$\begin{aligned}\Delta \langle X\rangle^2 &= \mathrm{Tr}(X\otimes X\Delta\rho)\\ &= \mathrm{Tr}\left[X\otimes X\int\rho\otimes\rho\ \pi_f(\rho)\mathrm{d}\rho\right] - \mathrm{Tr}\left[X\otimes X\hat{\rho}_{\mathrm{BME}}\otimes\hat{\rho}_{\mathrm{BME}}\right]\\ &= \int \mathrm{Tr}[X\rho]\cdot\mathrm{Tr}[X\rho]\pi_f(\rho)\mathrm{d}\rho - \mathrm{Tr}[X\hat{\rho}_{\mathrm{BME}}]^2\\ &= \int \langle X\rangle_\rho^2\,\pi_f(\rho)\mathrm{d}\rho - \left[\int\langle X\rangle_\rho\,\pi_f(\rho)\mathrm{d}\rho\right]^2.\end{aligned} \tag{26}$$

Because $\langle X\rangle$ parameterizes exactly one of the dimensions of Hilbert–Schmidt space, we can compute a marginal distribution over $\langle X\rangle$ by integrating $\pi_f(\rho)\mathrm{d}\rho$ over its other $d^2-2$ dimensions, which we denote by $\sigma$. Then $\mathrm{d}\rho = \mathrm{d}\langle X\rangle\mathrm{d}\sigma$, and

$$\pi_f(\langle X\rangle)\mathrm{d}\langle X\rangle \equiv \int_\sigma \pi_f(\rho)\mathrm{d}\rho, \tag{27}$$

in terms of which,

$$\Delta \langle X\rangle^2 = \int \langle X\rangle^2\,\pi_f(\langle X\rangle)\mathrm{d}\langle X\rangle - \left[\int\langle X\rangle\,\pi_f(\langle X\rangle)\mathrm{d}\langle X\rangle\right]^2, \tag{28}$$

which is the familiar formula for the variance of the univariate distribution $\pi_f(\langle X\rangle)$.

In particular, if $|\psi\rangle$ is an eigenvector of $\hat{\rho}_{\mathrm{BME}}$, let $X = |\psi\rangle\langle\psi|$. Then $\lambda = \langle X\rangle$ is the corresponding eigenvalue, and $\Delta\lambda^2 = \Delta\langle X\rangle^2$ is the reported uncertainty about it. Since $\langle X\rangle_\rho$ is between 0 and 1 for all $\rho$, $\pi_f(\lambda)\mathrm{d}\lambda$ is a distribution over the interval $[0\ldots1]$. For any such

distribution, $\Delta\lambda^2 \leqslant \lambda(1-\lambda)$, so every eigenvalue yields an upper bound for its own uncertainty. Note, too, that this bound is uniquely saturated by $\pi(\lambda) = (1-p)\delta(\lambda) + p\delta(1-\lambda)$, which is maximally bimodal. In practice, well-behaved priors will produce convex posteriors, for which $\Delta\lambda^2 \lesssim \lambda^2$ (i.e. $\Delta\lambda$ is no greater than $\lambda$ itself) can reasonably be expected.

*2.3.3. BME optimizes accuracy.* Above all else, an estimation procedure should yield an *accurate* estimate—one as close to the 'true' state as possible. While the concept of a 'true' state is problematic in actual experiments, it makes perfect sense in the context of a simple game. An impartial judge selects a state $\rho$, and provides $N$ copies of it to the estimator, who measures them and reports an estimate $\hat{\rho}$. The best procedure is the one that consistently makes $\hat{\rho}$ as close as possible to the unknown [to the estimator] $\rho$.

Which procedure is 'best' depends on the situation. For instance, if the judge always picks a particular $\rho_0$, and the estimator knows this, then the obvious best procedure is 'Report $\hat{\rho} = \rho_0$ no matter what!' This is a trivial case. This section considers a situation where the judge selects $\rho$ at random from a distribution $\pi_0(\rho)\mathrm{d}\rho$. For a frequentist, this scenario is well defined, but not completely general (but see comments at the end of the section). For a Bayesian, this description is so general as to be tautological—any uncertainty about the preparation can (and should) be represented this way.

I will show that BME with the prior $\pi_0(\rho)\mathrm{d}\rho$ is unconditionally the most accurate scheme possible. It minimizes the expected error between $\hat{\rho}$ and $\rho$. This optimality holds for every finite $N$, not just asymptotically. It depends of course on the measure of 'error' between $\rho$ and $\hat{\rho}$ adopted. The error measures optimized by BME, operational divergences, are arguably the best-motivated such measures. The argument presented here is brief; for more details, see [5].

Operational divergences, denoted $\Delta(\rho : \hat{\rho})$, measure how well the density matrix $\hat{\rho}$ describes (or estimates) the quantum state $\rho$. A certain subtlety should be noted here: whereas $\rho$ represents the state of a quantum system, $\hat{\rho}$ is a classical description of a state—e.g. a density matrix written down on paper. Two natural requirements constrain operational divergences. Firstly, $\Delta$ must represent the outcome of some physically implementable process. Secondly, the best description of $\rho$ had better be $\rho$ itself.

To satisfy operationality, we imagine trying to motivate the estimator to do a good job. A third-party verifier, equipped with the estimate $\hat{\rho}$, will perform a measurement on $\rho$. This measurement, $\mathcal{P}(\hat{\rho}) = \{E_1 \ldots E_m\}$, is an arbitrary POVM that may depend on $\hat{\rho}$. Depending on the outcome $(i)$, the verifier pays the estimator an amount $r_i(\hat{\rho})$.

The estimator's reward is represented by an operator

$$\mathcal{R}(\hat{\rho}) = \sum_i r_i(\hat{\rho}) E_i(\hat{\rho}), \tag{29}$$

and his/her expected reward (which he/she hopes to maximize) is

$$r(\rho : \hat{\rho}) = \mathrm{Tr}(\rho \mathcal{R}(\hat{\rho})). \tag{30}$$

The amount that he/she loses by inaccurately describing the state,

$$\Delta(\rho : \hat{\rho}) \equiv r(\rho : \hat{\rho}) - r(\rho : \rho) = \mathrm{Tr}\left[\rho\left(\mathcal{R}(\hat{\rho}) - \mathcal{R}(\rho)\right)\right], \tag{31}$$

is an operational divergence. Note that (i) it is operationally significant; and (ii) the best description of $\rho$ is $\rho$ itself.

Of course, not every reward scheme is strictly proper, satisfying the condition that $\rho$ be its own best estimate,

$$r(\rho:\rho) > r(\rho:\hat{\rho})\ \forall\ \hat{\rho} \neq \rho. \tag{32}$$

Equation (32) is a constraint on $\mathcal{R}(\hat{\rho})$. If we define $G(\rho) \equiv r(\rho:\rho)$ as the expected reward for a perfect estimate, then a bit of algebra yields

$$G(\rho) > G(\hat{\rho}) + \mathrm{Tr}\left[\left(\rho - \hat{\rho}\right)\mathcal{R}(\hat{\rho})\right]. \tag{33}$$

Equation (33) holds if and only if (i) $r(\rho:\hat{\rho})$ (as a function of $\rho$) is tangent to $G(\rho)$; and (ii) $G(\cdot)$ is strictly concave. Thus, for every strictly concave function $G(\cdot)$ on density operators, there is a unique operational divergence[7]:

$$\Delta(\rho:\hat{\rho}) = G(\rho) - G(\hat{\rho}) - \mathrm{Tr}\left[(\rho-\hat{\rho})\nabla G(\hat{\rho})\right], \tag{34}$$

where $\nabla G(\cdot)$ is the gradient of $G(\cdot)$.

Operational divergences include widely used measures such as the squared Hilbert–Schmidt or $L_2$ distance,

$$\Delta_2(\rho:\hat{\rho}) = \mathrm{Tr}\left[(\rho-\hat{\rho})^2\right], \tag{35}$$

associated with $G(\rho) = \mathrm{Tr}(\rho^2)$; and the relative entropy or Kullback–Leibler divergence,

$$\Delta_{KL}(\rho:\hat{\rho}) = \mathrm{Tr}\left[\rho\left(\log\rho - \log\hat{\rho}\right)\right], \tag{36}$$

associated with $G(\rho) = -H(\rho) = \mathrm{Tr}(\rho\log\rho)$.

Now that we have determined how to measure accuracy, let us try to optimize it. This is an easy task for an omniscient estimator, because the best estimate of $\rho$ is $\rho$ itself. If the estimator actually knows $\rho$, then her best plan is to report $\hat{\rho} = \rho$. The interesting case is an uncertain estimator. She must consider all the possible $\rho$, in order to choose the best $\hat{\rho}$. A risk-neutral estimator seeks to maximize his/her expected reward, averaged over all possible $\rho$.

Consider any estimation procedure, such as a map from measurement records $\mathcal{M}$ to estimates $\hat{\rho}(\mathcal{M})$. Which procedure should the estimator choose? Suppose that the unknown state $\rho$ to be estimated will be drawn from an ensemble described by $\pi_0(\rho)\mathrm{d}\rho$. The expected reward yielded by the procedure $\hat{\rho}(\mathcal{M})$ is an average over (i) possible $\rho$ and (ii) the ensuing $\mathcal{M}$.

$$\overline{r} = \int_\rho \pi_0(\rho)\mathrm{d}\rho \sum_{\mathcal{M}} p(\mathcal{M}|\rho) r\left(\rho:\hat{\rho}(\mathcal{M})\right). \tag{37}$$

Inserting $r(\rho:\hat{\rho}) = \mathrm{Tr}\left[\rho\mathcal{R}(\hat{\rho})\right]$ (equation (30)),

$$\overline{r} = \int_\rho \pi_0(\rho)\mathrm{d}\rho \sum_{\mathcal{M}} p(\mathcal{M}|\rho)\mathrm{Tr}\left[\rho\mathcal{R}\left(\hat{\rho}(\mathcal{M})\right)\right]. \tag{38}$$

The trace, sum and integral are all linear, so we can rearrange them as

$$\overline{r} = \sum_{\mathcal{M}} \mathrm{Tr}\left[\left(\int_\rho \rho p(\mathcal{M}|\rho)\pi_0(\rho)\mathrm{d}\rho\right)\mathcal{R}\left(\hat{\rho}(\mathcal{M})\right)\right]. \tag{39}$$

[7] Actually, if $G(\cdot)$ is not differentiable at a point (i.e. it has a cusp), then a family of operational divergences exist, indexed by the possible sub-gradients $\nabla G(\cdot)$. This seems to be a purely technical point, with no real significance in practice.

We now observe that $\int p(\mathcal{M}|\rho)\pi_0(\rho)\mathrm{d}\rho = p(\mathcal{M})$, the unconditional probability of observing $\mathcal{M}$. Furthermore, $\int \rho p(\mathcal{M}|\rho)\pi_0(\rho)\mathrm{d}\rho = \hat{\rho}_{\mathrm{BME}}(\mathcal{M})$, the BME estimate given $\pi_0$. Using these identities, the estimator's expected reward is

$$\bar{r} = \sum_{\mathcal{M}} p(\mathcal{M})\mathrm{Tr}\left[\hat{\rho}_{\mathrm{BME}}(\mathcal{M})\mathcal{R}\left(\hat{\rho}(\mathcal{M})\right)\right], \tag{40}$$

$$= \sum_{\mathcal{M}} p(\mathcal{M}) r\left(\hat{\rho}_{\mathrm{BME}}(\mathcal{M}) : \hat{\rho}(\mathcal{M})\right), \tag{41}$$

and each term in the sum can be independently maximized. For each $\mathcal{M}$, the optimal $\hat{\rho}(\mathcal{M})$ is $\hat{\rho}_{\mathrm{BME}}$—which means that BME is unconditionally the optimal estimation procedure.

This result is remarkable because it makes no appeal to asymptotics; the optimality holds for 100, 10, or even just 1 observation. Of course, when the estimator has insufficient data, the resulting estimate will not be very accurate—but neither will any other estimate. Crucially, his/her uncertainty will be reflected in a highly mixed estimate, with large error bars. Unlike MLE, BME fails gracefully, making the best use of the available data without over-reaching.

There are a few caveats that should be kept in mind. Firstly, BME need not optimize measures that are not operational divergences—e.g. trace distance or fidelity. Measuring estimation performance using these measures is generally unwise, but they (especially fidelity) are commonly misused. Secondly, optimality requires the estimator's prior to coincide with the ensemble from which the unknown states were selected. A sufficiently 'wrong' prior will lead to horrendous results. The BME estimate is still the most honest and accurate representation of the estimator's knowledge, but that knowledge may have been predicated on dangerous prior assumptions.

This appears, at first glance, like an unbridgeable gap between Bayesians and frequentists. Happily, it is not, and there is a clear road forward. The problem of estimating classical probability distributions (in identical circumstances to our problem), has been largely solved using the *minimax* framework [21–23]. The basic idea is to consider the set of all possible estimation protocols $\mathcal{M} \to \hat{\rho}(\mathcal{M})$, and for each protocol to identify the state $\rho$ for which it works the *worst*. Then the protocols are ranked by their worst-case performance, and the one with the best worst-case performance is chosen. This 'minimax solution' is guaranteed to work pretty well for every state, so its optimality can be demonstrated even to a die-hard frequentist. However, the minimax solution is (provably) *always* a BME protocol, with a particular 'non-informative' prior. Hence, in probability estimation, open-minded Bayesians and frequentists can agree on a single unconditionally reliable protocol. This brief summary does not do justice to the subtle and surprising details of minimax analysis; suffice it to say that similar results for quantum estimation would be very useful, but deriving them is not a trivial problem.

### *2.4. Bayesian and frequentist approaches*

Having examined both frequentist and Bayesian approaches, I have focused on the concrete details—(How does MLE fail? Why does BME do better? How is BME done?)—because estimation is an operational task. Certain readers may, however, ask 'what is wrong with frequentism, anyway?' Others may be wondering what really distinguishes Bayesian and frequentist methods, since $\mathcal{L}(\rho)$ is crucial to both. I attempt to address these questions below.

*2.4.1. Why frequentism fails.* The frequentist approach has dominated statistics for most of the 20th century, so its failure in quantum state estimation requires some explanation. To see why frequentism fails, we might first ask why it should succeed.

MLE attempts to fit the observed data, and so the MLE estimate is the best 'predictor' of the past. Since the goal of a state estimate is to predict the future, frequentist estimation techniques can be justified by the following axiom: *the future will look [statistically] identical to the past*. If this axiom is true, then $\hat{\rho}_{\mathrm{MLE}}$ is the best possible estimate. The law of large numbers implies its validity as $N \to \infty$, and the central limit theorem quantifies this convergence.

For classical systems, it is always *possible* that the frequentist axiom will hold. If the coin comes up 'heads' the first time, it is entirely possible that it will always come up heads. Moreover, the rules are not going to change—the possible outcomes in the past were 'heads' and 'tails', and they will remain the only possible outcomes in the future.

This does not hold for quantum systems. The past, represented by the estimator's data, comprises a finite set of observations extracted from a finite variety of measurements. For instance, the estimator might have measured $\sigma_x$, $\sigma_y$ and $\sigma_z$ on a qubit. Future experimenters, however, might choose to measure *any* observable—and there are infinitely many. A quantum state, by definition, predicts the probabilities for every possible measurement. The frequentist axiom cannot possibly hold; any future observer could violate it at will, simply by making a novel measurement.

Frequentist methods for classical probabilities yield zero probabilities only when

(a) event '$i$' has never been observed,

(b) in every trial, something in the complement of event '$i$' *was* observed.

That is, event '$i$' *could* have happened, but it did not. When MLE is used on quantum systems, the $|\phi_i\rangle\langle\phi_i|$ that ends up getting assigned zero probability is almost never something that could have been observed. The Achilles' heel of frequentist quantum estimation is that it happily assigns zero probability to events that were never observed *not* to happen. To avoid this problem, we need a method that does not begin by assuming 'the future will look like the past', because for a quantum system, that cannot be true.

*2.4.2. How the Bayesian and frequentist approaches differ.* $\mathcal{L}(\rho)$ is the key ingredient in Bayesian methods, just as in frequentist ones. It represents everything relevant about the data. In frequentist methods, $\mathcal{L}(\rho)$ is the sole ingredient, and so the only natural thing to do is to find the $\hat{\rho}$ that maximizes it. Bayesian methods, in contrast, transform the likelihood into a probability distribution,

$$\mathcal{L}(\rho) \longrightarrow \pi(\rho)\,\mathrm{d}\rho \propto \mathcal{L}(\rho)\pi_0(\rho)\,\mathrm{d}\rho, \tag{42}$$

by multiplying it by a prior distribution $\pi_0(\rho)\mathrm{d}\rho$.

A common misconception is that this transformation is trivial when $\pi_0(\rho)\mathrm{d}\rho$ is 'flat' (e.g. coincides with a Lebesgue or Haar measure). On the contrary, it transforms a *function* into a *distribution* (or measure), which is an entirely different mathematical object. Functions, like $\mathcal{L}(\rho)$, have values. Distributions have integrals—they assign values not to points, but to regions.

For example, if $\mathcal{L}(x)$ is defined for real-valued $x$, then $\mathcal{L}(0)$ and $\mathcal{L}(1)$ are well defined, but $\int_0^1 \mathcal{L}(x)$ is purely meaningless. To integrate, we must multiply by $\mathrm{d}x$ (a measure), obtaining a distribution $\mathcal{L}(x)\mathrm{d}x$. This can be integrated over the interval $[0, 1]$—but evaluating $\mathcal{L}(x)\mathrm{d}x$ at $x = 0$ is ill defined (and infinitesimal in any case).

This difference between functions and distributions enforces a difference in approach between frequentist and Bayesian methods. Frequentists, abjuring priors, can only work with the function $\mathcal{L}(\rho)$. The corresponding estimate, $\hat{\rho}_0$, will be distinguished by the *value* of $\mathcal{L}(\hat{\rho}_0)$. The Bayesian approach begins and ends with a distribution that has no values. Everything must be phrased in terms of measurable subsets (e.g. intervals), and integration over them.

Estimation algorithms transform data (observed in the past) into an estimate (which predicts the future). In order to select the best estimate, we must logically connect the past and the future. Frequentist methods implicitly use the frequentist axiom, while Bayesian approaches take a weighted average over all possible theories. This averaging is particularly apropos for quantum state estimation, because density matrices have a natural convex structure. Suppose a physicist knows that a qubit's state is $|0\rangle$ with probability $\frac{1}{3}$ and $|1\rangle$ with probability $\frac{2}{3}$. He will describe it by the *average* state, $\rho = \frac{1}{3}(|0\rangle\langle 0| + 2\,|1\rangle\langle 1|)$—not the most likely state, $\rho = |1\rangle\langle 1|$.

Viewed this way, the prior replaces the frequentist axiom as a connection between the past and the future. This can be an advantage, for a Bayesian is capable of gracefully acknowledging that the data are not descriptive of the true state—that they are unlikely, or simply insufficient. However, the price paid for this flexibility is the need to choose a prior, often without any good justification.

## 3. Where do we go from here?

The Bayesian approach to state estimation has undeniable advantages. It is accurate, it honestly represents the estimator's knowledge, and it conforms to quantum states' role as predictors. Purely frequentist approaches—e.g. maximum likelihood as it is currently used—cannot match these qualities.

Nonetheless, BME comes with an array of concomitant challenges. These range from the purely practical (integration is hard) to the fundamental (How do we choose a prior?). While some are specific to Bayesian methods, others cast doubt on the scalability of *any* state estimation procedure.

### *3.1. The Prior's Tale*

Of all the problems and caveats raised by BME, none is more pressing or obvious than 'How do we choose a prior?'. BME's optimality depends on the estimator's prior matching the 'true' distribution of unknown states. This is fine in the rather artificial context of a state-estimating game that might be played many times, but physics experiments are not drawn from an ensemble. Each experiment is, as a rule, unique.

The prior is therefore a necessary fiction. As a convenient way of representing the estimator's ignorance (either genuine, or assumed for the sake of scientific impartiality), it ought to be as uninformative as possible. Unitary invariance is a good first guideline (see section 3.3 below, however). Over the spectrum of $\rho$, however, no uniquely suitable measure exists. Identifying particularly useful and non-informative priors remains an open and urgent question.

A related open question is 'What is the penalty for choosing the wrong prior?'. If accuracy is measured by an operational divergence, then BME must outperform MLE and all other methods—if the estimator's prior matches the distribution of unknown states. Its robustness

to a poor prior is unknown. The optimality proof given previously is elegant in its simplicity, but precisely because of that elegance, it provides few clues to this problem.

### *3.2. Practical matters*

Every calculus student learns that integration is harder than differentiation. Numerical integration is an active and challenging field of numerical analysis, whereas differentiation involves little more than function evaluation. BME consists almost entirely of integration, whereas MLE is a maximization problem. Unsurprisingly, the implementation of BME described above is roughly an order of magnitude slower than MLE. Experimentalists, already frustrated by MLE analyses that run for a week or more [28], may be nonplussed.

This state of affairs may owe a great deal to the fact that MLE algorithms, unlike BME, have been developed and used for 5–10 years. Substantial speedups are likely in the future—precisely because numerical integration remains something of an art. The Metropolis–Hastings algorithm already provides a tremendous advantage over naïve Monte Carlo, so a few more orders of magnitude may be feasible.

One reason for optimism is that the BME integral appears, in principle, to have a rather simple analytic form. The likelihood function is a polynomial, the product of many linear functions, of the form $\mathrm{Tr}(\rho M_i)$. For certain priors (e.g. Hilbert–Schmidt), the resulting posterior looks a lot like a beta distribution of the form $\beta(x) = x^n(1-x)^{N-n}$. This appears in classical estimation and is easy to integrate. What makes the quantum case hard is the boundary conditions. Unlike the classical probability simplex, the quantum state-set has curved edges that are awkward to integrate over. However, analytic solutions can be obtained for small $N$, and a general solution might be possible.

### *3.3. Scalability*

Quantum devices exist that provide coherent control over 8–12 qubits [9, 24]. Twenty or 30 qubits will probably be controlled within the next five years (if only for a short time, and with limited fidelity). The Hilbert space of a 30-qubit quantum register is enormous—to merely store one density matrix for such a device would require just under 1 million terabytes of memory. State estimation, as we know it, is impossible in this context.

Nonetheless, characterizing quantum hardware will remain important. A quantum computer will not need state estimation; its results will appear as a computational basis state, determined by a projective measurement. Development and testing of components, however, will depend crucially on state estimation. It is not sufficient to know whether or not the desired state is produced; the designer will want to know the nature of the errors, so as to correct them. Eventually, these errors need to be reduced below a fault-tolerance threshold that is probably less than $10^{-3}$.

As the states that are estimated grow larger and the uses to which they are put become more demanding, utterly new techniques will be needed. Unbiased estimation—i.e. guessing the system's state without any pre-existing assumptions—becomes exceedingly data intensive for large Hilbert spaces. Making use of the estimator's prior knowledge will be essential. The Bayesian approach presented here provides a natural framework for doing so. However, a framework for reliably representing that prior knowledge (without falling prey to self-fulfilling

prophecies) will be necessary. This reason alone would justify further study of Bayesian state estimation.

Another approach to the same problem is to focus on certain properties of the state. This underlies Gross *et al*'s compressed tomography [25] and Aaronson's PAC learning approach [26]. Compressed tomography is targeted at low-rank states: if the data were in fact generated by a nearly pure state, then techniques from compressed sensing (see references cited in [25]) can be used to reconstruct that state from data that would otherwise be incomplete. There is a price to be paid, in that higher precision is required, and so compressed tomography appears to require the same total number of measurement 'clicks'.

Aaronson's unique approach, in [26], is based on Valiant's notion of PAC (probably approximately correct) learning. He shows that any fixed set of measurements on an $N$-qubit system can be PAC learned using only poly($N$) copies. This seems impossible on the face of it, since conventional state estimation requires $O(2^N)$ different measurement settings. What enables Aaronson's intriguing result is a crucial assumption about the estimator's goals: future measurements will be drawn from an ensemble that is known in advance. PAC learning requires accurately estimating the probabilities for most future measurements—but it is okay to be wildly wrong about a tiny (exponentially small in $N$) fraction of them. PAC learning is possible because one of the following must hold:

1. There are at most poly($N$) measurements in the set.
2. Most of them are highly correlated with each other, so there are at most poly($N$) linearly independent measurements in the set.
3. For every possible $\rho$, most of the measurements in the set have nearly uniform probability distributions.

If the measurements to be learned are tomographically complete (e.g. a full set of mutually unbiased bases), then the third case holds. The estimator can just write down $\hat{\rho} = \frac{\mathbb{1}}{d}$—without making any measurements at all.

Compressed tomography and PAC learning are not alternatives to MLE, BME or linear inversion. They are not procedures for state estimation. They are new definitions of the problem! Both are potentially amenable to frequentist, Bayesian or *ad hoc* statistical methods. Adapting the Bayesian principle and methods discussed here to them is a challenge (for instance, compressed tomography explicitly seeks a low-rank $\hat{\rho}$, while BME explicitly avoids rank-deficient estimates), but a worthy one. For one thing, these approaches lie squarely in the regime of incomplete and undercomplete data—precisely where Bayesian methods excel. These estimates will have large and unpredictable error bars. Characterizing the estimator's uncertainty will be important. These problems demand a reasoned and careful approach to statistical inference—which is exactly what Bayesian inference can provide.

## Acknowledgments

This paper is the result of more than two years of thinking about state estimation. Much of that thinking has been done out loud, and the author is exceptionally grateful to his conversational partners, in particular, Stephen Bartlett, Carlton Caves, Hartmut Häffner, Patrick Hayden, Richard Gill, Daniel James, Dominik Janzing, Jan Korsbakken, Karan Malhotra, Serge Massar, Colin McCormick, John Preskill, Andrew Silberfarb, Rob Spekkens and Steven Van Enk.

## Appendix. Necessary and sufficient condition for a prior's robustness

**Theorem 1.** *A prior $\pi_0(\rho)\mathrm{d}\rho$ over $d \times d$ density operators is robust (and therefore guaranteed to generate full-rank estimates for any finite measurement record) if and only if its support in Hilbert–Schmidt space is not a subset of a finite intersection of $((d-1)^2-1)$-dimensional hyperplanes that are tangent to the state-set.*

**Proof.** A prior is fragile if and only if it can be annihilated by a some finite-length measurement record: that is, there exists $\mathcal{M} = \{M_1, M_2, \ldots M_N\}$ so that $\mathcal{L}(\rho) = \prod_{i=1}^{N} \mathrm{Tr}(M_i\rho)$ is zero on the prior's support. $\mathcal{L}(\rho)$ is zero if and only if, for some $i$, $\mathrm{Tr}(M_i\rho) = 0$. Each $M_i$ thus eliminates every $\rho$ supported on $M_i$'s null space, which has at most $(d-1)$ dimensions. The Hermitian trace-1 matrices supported on $M_i$'s null space form a $(d-1)^2-1$-dimensional hyperplane in Hilbert–Schmidt space. This hyperplane contains non-negative states, which are necessarily orthogonal to $M_i$ and therefore lie on the boundary of the state-set. Thus, $M_i$ eliminates all density matrices lying within a hyperplane that includes states, but does not include full-rank states—i.e. a hyperplane that is tangent to the state-set. The states eliminated by $\mathcal{M}$ are, therefore, merely the intersection of $N$ such hyperplanes, and if the prior's support does not lie within such an intersection, it cannot be eliminated.

Conversely, suppose that the prior's support does lie within an intersection of $N$ such tangent hyperplanes. Each hyperplane is closed under convex combination, so we can define a convex combination of every non-negative element, $\rho_0$, which is itself an element of the hyperplane. Since the hyperplane is tangent to the state-set, no element can lie in the interior, and so $\rho_0$ is not full-rank—i.e. it is orthogonal to some $|\psi\rangle\langle\psi|$. Since $\rho_0$ is a convex combination of every state in the hyperplane, the entire hyperplane is orthogonal to $|\psi\rangle\langle\psi|$ and is therefore eliminated by observing $|\psi\rangle\langle\psi|$. A measurement record consisting of the annihilating projectors for each of the $N$ hyperplanes will therefore annihilate the prior, so it is fragile. □

**Corollary 2.** *Any prior with support on a smooth curve in at least $(d-1)^2$ dimensions is robust.*

**Proof.** Since the curve occupies at least $(d-1)^2$ dimensions of Hilbert–Schmidt space, it cannot be contained in a $((d-1)^2-1)$-dimensional hyperplane. If it could be contained in a finite union of such planes, then it would not be smooth. □

## References


[1] DiVincenzo D P 2000 *Fortschr. Phys.* **48** 771
[2] Altepeter J B *et al* 2003 *Phys. Rev. Lett.* **90** 193601
[3] Aliferis P, Gottesman D and Preskill J 2006 *Quantum Inf. Comput.* **6** 97
[4] Knill E 2005 *Nature* **439** 39
[5] Blume-Kohout R and Hayden P 2006 arXiv:quant-ph/0603116v1
[6] Hradil Z 1997 *Phys. Rev.* A **55** R1561
[7] Hradil Z, Rehácek J, Fiurasek J and Jezek M 2004 *Lect. Notes Phys.* **649** 59
[8] James D F V, Kwiat P G, Munro W J and White A G 2001 *Phys. Rev.* A **64** 052312
[9] Haeffner H *et al* 2005 *Nature* **438** 643
[10] Carlin B P and Louis T A 2009 *Bayesian Methods for Data Analysis (Texts in Statistical Science)* (Boca Raton, FL: CRC Press)

[11] Helstrom C W 1976 *Quantum Detection and Estimation Theory* (New York: Academic)
[12] Jones K R W 1991 *Ann. Phys.* **207** 140
[13] Derka R, Buzek V, Adam G and Knight P L 1996 *J. Fine Mech. Opt.* **11–12** 341
[14] Schack R, Brun T A and Caves C M 2001 *Phys. Rev.* A **64** 014305
[15] Neri F 2005 *Phys. Rev.* A **72** 062306
[16] Tanaka F and Komaki F 2005 *Phys. Rev.* A **71** 052323
[17] Metropolis N, Rosenbluth A W, Rosenbluth M N, Teller A H and Teller E 1953 *J. Chem. Phys.* **21** 1087
[18] Hastings W K 1970 *Biometrika* **57** 97
[19] Chib S and Greenberg E 1995 *Am. Stat.* **49** 327
[20] Zyczkowski K and Sommers H-J 2001 *J. Phys. A: Math. Gen.* **34** 7111
[21] Clarke B and Barron A 1994 *J. Stat. Plan. Inference* **41** 37–60
[22] Merhav N and Feder M 1998 *IEEE Trans. Inf. Theory* **44** 2124
[23] Xie Q and Barron A 2000 *IEEE Trans. Inf. Theory* **46** 431
[24] Negrevergne C *et al* 2006 *Phys. Rev. Lett.* **96** 170501
[25] Gross D, Liu Y-K, Flammia S T, Becker S and Eisert J 2009 arXiv:0909.3304v2
[26] Aaronson S 2007 *Proc. R. Soc.* A **463** 3089–114
[27] Altepeter J B, Jeffrey E R and Kwiat P G 2005 *Adv. At. Mol. Opt. Phys.* **52** 105
[28] Häffner H 2006 private communication