Hedged Maximum Likelihood Quantum State Estimation

Robin Blume-Kohout*

Perimeter Institute for Theoretical Physics, 31 Caroline Street North, Waterloo, Ontario N2L 2Y5, Canada

(Received 17 February 2010; published 10 November 2010)

This Letter proposes and analyzes a new method for quantum state estimation, called hedged maximum likelihood (HMLE). HMLE is a quantum version of Lidstone's law, also known as the "add β" rule. A straightforward modification of maximum likelihood estimation (MLE), it can be used as a plug-in replacement for MLE. The HMLE estimate is a strictly positive density matrix, slightly less likely than the ML estimate, but with much better behavior for predictive tasks. Single-qubit numerics indicate that HMLE beats MLE, according to several metrics, for nearly all "true" states. For nearly pure states, MLE does slightly better, but neither method is optimal.

DOI: 10.1103/PhysRevLett.105.200504 PACS numbers: 03.67.−a, 03.65.Wj, 42.50.Dv

Quantum state estimation is a basic task in quantum information science [1], simple to describe but hard to do right. The estimator gets N independently and identically prepared quantum systems, performs measurements on them, analyzes the data, and reports a single-system density matrix $\hat{\rho}$. The goal is to report the most "accurate" answer possible. What this means is debatable. I adopt three common assumptions: (i) we have N copies of an unknown "true" state ρ; (ii) the goal is to get $\hat{\rho}$ as close as possible to ρ, according to some metric $d(\rho, \hat{\rho})$; and (iii) we are concerned with average (over measurement outcomes) performance. I will not consider how to choose a measurement, seeking instead a protocol that works well for all measurements.

The current standard, maximum likelihood estimation (MLE) [2–5], tends to report rank-deficient estimates with zero eigenvalues [6]. Those eigenvalues represent probabilities. A zero probability indicates extraordinary confidence—confidence that the data do not support, and which can be catastrophic if used for predictive tasks.

This Letter suggests an alternative, hedged maximum likelihood (HMLE), which can be used as a plug-in substitute for MLE. The modification consists, in its entirety, of the following rule. Replace the standard likelihood function $\mathcal{L}(\rho) = \Pr(\text{observed data}|\rho)$ with the product of $\mathcal{L}(\rho)$ and a "hedging function"

$$h(\rho) = \det(\rho)^{\beta}, \tag{1}$$

where $\det(\cdot)$ is the determinant, and $\beta \approx \frac{1}{2}$ is a positive constant chosen by the estimator. The rest of this Letter explains, derives, and analyzes this rule.

Background.—HMLE generalizes a classical rule for probability estimation called "add β," also known as Lidstone's law [7,8]. Suppose we observe N samples from an unknown distribution $\mathbf{p} = \{p_1 \ldots p_K\}$, and see n_k "k"s. What probabilities $\hat{\mathbf{p}}$ should we assign for the next sample? The likelihood, $\mathcal{L}(\mathbf{p}) = \prod_k p_k^{n_k}$, is maximized by the natural and obvious estimate

$$\hat{p}_k = \frac{n_k}{N}. \tag{2}$$

But if k has not yet been observed, $n_k = 0$, and MLE assigns $\hat{p}_k = 0$. This is fine if p_k really is zero, but p_k may well be positive but small. If so, the consequences of assigning $\hat{p}_k = 0$ depend on what the estimate is used for. When it is used for predictive tasks, such as data compression or gambling [9,10], they may be catastrophic.

Compression and gambling define operational interpretations of $\hat{\mathbf{p}}$. Compressors seek to shorten strings reversibly, by replacing each instance of k with a [binary] code word w_k. Code words should (on average) require less space than the original symbols, but their lengths are constrained by the Kraft-McMillan inequality, $\sum_k 2^{-\text{length}(w_k)} \leq 1$. Gamblers seek to grow their bankroll as rapidly as possible (in expectation), by betting it on the possible events $\{k\}$. In a useful and widely studied model based on roulette or horse racing (see [9]), money bet on the event that occurs (e.g., the winning horse) is multiplied by a constant C, while that allocated to other events is lost [11].

The natural measure of error for both tasks is relative entropy. If $\mathbf{p}$ are the true probabilities and $\hat{\mathbf{p}}$ are estimates, then $\mathbf{p}$'s entropy, $H(\mathbf{p}) \equiv -\sum_k p_k \log p_k$, is the unavoidable cost of $\mathbf{p}$'s randomness, while the additional cost of error is given by relative entropy,

$$D(\mathbf{p}\|\hat{\mathbf{p}}) = \sum_k p_k(\log p_k - \log \hat{p}_k). \tag{3}$$

The gambler's optimal expected rate of gain is $(n) = (0)e^{n[\log C - H(\mathbf{p})]}$, achieved (uniquely) by betting a fraction p_k of his bankroll on outcome k. Similarly, the compressor's minimum expected output length is $L(n) = nH(\mathbf{p})$, achieved (uniquely) by replacing k with a code word of length $-\log p_k$. But if they act based on $\hat{\mathbf{p}} \neq \mathbf{p}$, then these quantities grow instead as

$$(n) = (0)e^{n[\log C - H(\mathbf{p}) - D(\mathbf{p}\|\hat{\mathbf{p}})]} \tag{4}$$

$$L = n[H(\mathbf{p}) + D(\mathbf{p}\|\hat{\mathbf{p}})]. \tag{5}$$

Setting $\hat{p}_k = 0$ thus implies extreme strategies for gambling (bet everything against k) and data compression (map k to an infinitely long code word). If the next letter is k, the gambler loses his entire bankroll irrevocably, and the compressor's output becomes undecodable (for there is no finite code word to assign).

"Add β" avoids these catastrophes by hedging against as-yet-unseen possibilities. It assigns probabilities

$$\hat{p}_k = \frac{n_k + \beta}{N + K\beta}. \tag{6}$$

The lowest probability that can be assigned is $\frac{\beta}{N+K\beta} \approx \frac{\beta}{N}$. Like Eq. (2), this rule has a statistical derivation. It is the Bayes estimator (i.e., it minimizes expected cost) for a relative entropy cost function and a Dirichlet-β prior

$$P_0(\mathbf{p})d\mathbf{p} \propto \prod_k p_k^{\beta-1} dp_k. \tag{7}$$

Dirichlet priors include the "flat" Lebesgue measure ($\beta = 1$), and Jeffreys' prior ($\beta = \frac{1}{2}$). Given any prior, we can minimize expected relative entropy by (1) updating the prior via Bayes' rule, and (2) reporting its mean value. For the Dirichlet-β prior, this gives the "add β" rule. But the "add β" rule is *not* intrinsically Bayesian. A naive estimator following Eq. (2) can simulate it by adding β dummy observations of each letter k. This yields new frequencies $\{n_k + \beta\}$ and a total of $N + K\beta$ observations. Since $\mathcal{L}(\mathbf{p}) = \Pr(\{n_k\}|\mathbf{p}) = \prod_k p_k^{n_k}$, the dummy observations yield a *hedged* likelihood function

$$\mathcal{L}'(\mathbf{p}) = \prod_k p_k^{n_k+\beta} = \left(\prod_k n_k^{\beta}\right)\mathcal{L}(\mathbf{p}), \tag{8}$$

whose maximum value is achieved by Eq. (6). When β is not an integer, the hedged likelihood [Eq. (8)] remains well defined, and the "add β" rule still maximizes it.

Quantum hedging.—The quantum analogue of a distribution $\mathbf{p}$ is a $d \times d$ density matrix ρ. It cannot be observed directly; observing a sample of ρ requires choosing a particular measurement $\mathcal{M}$ [12]. $\mathcal{M}$ is represented by a positive operator valued measure (POVM), a set of positive operators $\{E_i\}$ summing to $\mathbb{1}$, which determine the probability of outcome "i" as

$$\Pr(i) = \mathrm{Tr}[\rho E_i]. \tag{9}$$

Inferring ρ, from the observed frequencies $\{n_i\}$, is the central problem of quantum state estimation.

The simplest procedure is linear inversion tomography [13], which assumes Eq. (2) and inverts Born's rule [Eq. (9)] to get an estimate $\hat{\rho}_{\mathrm{tomo}}$ satisfying [14]

$$\mathrm{Tr}[\hat{\rho}_{\mathrm{tomo}} E_i] = \frac{n_i}{N} \quad \text{for } i = 1, \ldots, m. \tag{10}$$

Often, $\hat{\rho}_{\mathrm{tomo}}$ has negative eigenvalues—which is awkward, for they represent probabilities. Linear inversion ignores the shape of state space: to fit data from a single POVM $\mathcal{M}$, it happily assigns negative probabilities for unperformed measurements.

MLE [2] remedies this problem, assigning the $\hat{\rho}$ that maximizes the likelihood,

$$\mathcal{L}(\rho) = \Pr(\{n_i\}|\rho) = \prod_i \mathrm{Tr}[\rho E_i]^{n_i}. \tag{11}$$

Maximizing over all trace-1 Hermitian matrices yields $\hat{\rho}_{\mathrm{tomo}}$, but restricting to $\rho \geq 0$ yields a non-negative $\hat{\rho}_{\mathrm{MLE}}$.

However, $\hat{\rho}_{\mathrm{MLE}}$ can still assign zero probabilities—just as in Eq. (2). If $\hat{\rho}_{\mathrm{tomo}}$ is negative, $\hat{\rho}_{\mathrm{MLE}}$ will have a zero eigenvalue [6]. Moreover, quantum MLE usually assigns zero probability to a measurement outcome $|\psi\rangle\langle\psi|$ that is not in $\mathcal{M}$, and could never have been observed, whereas classically $p_k = 0$ only when k has been given N chances to appear and (so far) has not. So although $\hat{\rho}_{\mathrm{MLE}}$ may be the right estimator for some task, its zero eigenvalues are implausibly and (for predictive tasks like gambling and compression—see [15]) catastrophically overconfident. Prediction demands a hedged estimator.

Bayesian mean estimation (BME) is hedged, and with suitable priors has extremely good predictive behavior [6]. But quantum BME is computationally formidable, with no known closed-form solutions. This is unfortunate, for Bayes estimation of classical probabilities works very well. They yield "add β" rules when applied to Dirichlet-β priors, which are well motivated. Jeffreys' prior ($\beta = \frac{1}{2}$) yields estimators that are asymptotically optimal (by the minimax criterion) for data compression [19], Krichevskiy showed that "add 0.50922..." outperforms all other rules for predicting the next event [20], and Braess *et al.* [21] pointed out that $\beta \approx 1$ works well because large-N behavior depends only weakly on β.

This suggests adapting "add β" to quantum state estimation. Obvious methods like dummy counts do not work. If we estimate a qubit source by measuring σ_x, σ_y, and σ_z 10 times each, and (by unlikely chance) all the outcomes are $+1$, then $\hat{\rho}_{\mathrm{tomo}}$ is quite negative. $\hat{\rho}_{\mathrm{MLE}}$ is the projector onto its largest eigenvector. Adding $\beta = 1$ dummy counts has no effect: $\hat{\rho}_{\mathrm{tomo}}$ remains negative, and $\hat{\rho}_{\mathrm{MLE}}$ is unchanged.

Dummy data work classically because only K different events exist; dummy observations rule out $p_k = 0$ for any event. A quantum state assigns probabilities to infinitely many different measurement outcomes, and a finite set of dummy observations cannot bound all of them away from zero.

So HMLE modifies $\mathcal{L}$ directly, multiplying it by a unitarily invariant hedging function [Eq. (1)] independent of $\mathcal{M}$. This is directly analogous to the effect of dummy counts in Eq. (8), because $\det(\rho)$ is the product of ρ's eigenvalues. Hedging penalizes small probabilities, steering the maximum of $\mathcal{L}'(\cdot)$ away from boundaries. When a single basis is measured, HMLE reproduces "add β" exactly: the HMLE estimate is

$$\hat{\rho}_H = \sum_k \frac{n_k + \beta}{N + K\beta} |k\rangle\langle k|. \tag{12}$$

Equation (1) is the only measurement-independent smooth modification of $\mathcal{L}(\rho)$ that yields "add β" for every basis (see Appendix B of [22]).

Performance.—The point of HMLE is to give more accurate estimates than MLE. Predictive tasks (e.g., compression or gambling) suggest quantum relative entropy, $D(\rho|\hat{\rho}) \equiv \mathrm{Tr}\rho\log\rho - \mathrm{Tr}\rho\log\hat{\rho}$, as a measure of inaccuracy.

Evaluating MLE this way is difficult, for if $\hat{\rho}$ is rank deficient on ρ's support, then $D(\rho\|\hat{\rho}) = \infty$. The expected value of $D(\rho|\hat{\rho}_{\mathrm{MLE}})$ is always infinite, for measurement results yielding a rank-deficient $\hat{\rho}_{\mathrm{MLE}}$ can always occur. We can compare different amounts of hedging. Figure 1 shows relative-entropy error for $\beta = 10^{-2}, 10^{-1}, \frac{1}{2}$, applied to a single qubit measured $N = 10, 10^2, 10^3$ times in each of the Pauli bases. The error depends on ρ, most strongly on its radial coordinate $r = \sqrt{(1 + \mathrm{Tr}\rho^2)/2}$. For highly mixed states ($1 - r^2 \gg \sqrt{3/N}$), MLE rarely yields rank-deficient estimates, and accuracy increases slowly with β. For slightly mixed states ($1 - r^2 \approx \sqrt{3/N}$), $\hat{\rho}_{\mathrm{MLE}}$ is often rank deficient, and accuracy improves dramatically with β, up to $\beta \approx 1/2$. Nearly pure states ($1 - r^2 \ll \sqrt{3/N}$) display unexpected and complex behavior. A very small and N-dependent amount of hedging ($\beta_{\mathrm{optimal}} \approx 1/2\sqrt{N}$) is optimal; further hedging decreases accuracy for nearly pure states.

Other error metrics include Euclidean distance $[\sqrt{\mathrm{Tr}[(\rho - \hat{\rho})^2]}]$, infidelity $[1 - (\mathrm{Tr}\sqrt{\sqrt{\rho}\sigma\sqrt{\rho}})^2]$, and trace distance $(\mathrm{Tr}|\rho - \sigma|)$ [1]. Though not particularly appropriate for comparing $\hat{\rho}$ to ρ, they are widely used, so Fig. 2 illustrates their behavior for MLE and HMLE, applied to a single qubit measured in the Pauli bases. They all show the same basic behavior. For nearly pure states, MLE is more accurate. For highly mixed states, HMLE improves accuracy slightly. The biggest improvement comes in the intermediate regime where $O(1/N) < 1 - r^2 < O(1/\sqrt{N})$. These states are not quite pure, but close enough that MLE yields rank-deficient estimates a substantial fraction of the time, and hedging provides substantial improvement. So, even though HMLE is not designed to maximize fidelity or trace distance, it improves on MLE for all but the purest states.

Discussion.—There are other ways to avoid zero eigenvalues. Bayesian mean estimation [6] is well motivated and accurate, but is computationally formidable (requiring integration over $d^2 - 1$ dimensions), and philosophically objectionable to frequentists. On the other hand, simply mixing the MLE result with a small amount of the maximally mixed state [23] is easy to do, but *ad hoc* and ill motivated. Because HMLE is based on a [slightly] modified likelihood function, it is easier to analyze and justify than *ad hoc* schemes, while avoiding controversial Bayesian reasoning.

Strict frequentist methodology suggests choosing the most plausible state—i.e., $\hat{\rho}_{\mathrm{MLE}}$. Among choices with identical properties, we may as well pick the more likely (plausible) one. But if we choose some other estimate (e.g., $\hat{\rho}_H$) on its merits, then we ought to confirm that it is almost as likely as $\hat{\rho}_{\mathrm{MLE}}$.

Consider the classical case. If $n_k = 0$, then $\hat{\mathbf{p}}_{\mathrm{MLE}}$ assigns $\hat{p}_k = 0$. But it is equally plausible that $p_k \approx 1/N$, in which case k probably would not appear in the first N samples. The likelihood function bears this out: the most likely state sets $\hat{p}_k = 0$, but nearby states with nonzero p_k have almost the same likelihood. If $\hat{\mathbf{p}}_H$ assigns $\hat{p}_k = \frac{\beta}{N}$ and $\hat{p}_j = (1 - \frac{\beta}{N})\frac{n_j}{N}$ for $j \neq k$, then

$$\frac{\mathcal{L}(\hat{\mathbf{p}}_H)}{\mathcal{L}(\hat{\mathbf{p}}_{\mathrm{MLE}})} = \left(1 - \frac{\beta}{N}\right)^N \approx e^{-\beta}. \tag{13}$$

Likelihood ratios between e^{-1} and e are "barely worth mentioning" [24], so if $\beta < 1$, then $\hat{\mathbf{p}}_H$ is essentially as

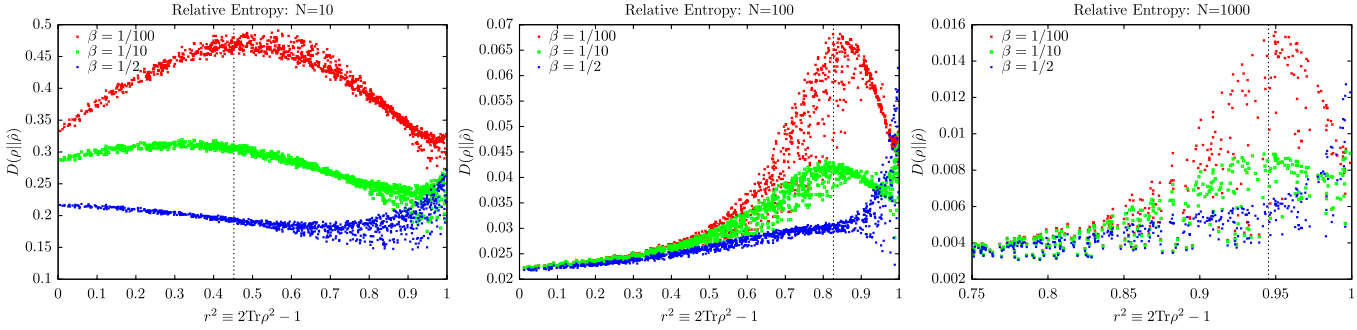

FIG. 1 (color online). Methodology: 10^3 single-qubit states ρ_{true} were selected at random from the Hilbert-Schmidt ("flat") measure on the Bloch sphere. For each state, 10^3 separate data sets were generated, each consisting of $3N$ ($N = 10, 100, 1000$) measurements divided among the three Pauli operators. HMLE estimates (with several β values) were calculated. For each ρ_{true}, relative entropy error was averaged over all 10^3 data sets. Results: Error is strongly correlated with $r^2 = \frac{1}{2}(1 + \mathrm{Tr}\rho^2)$. There are three regimes, separated by $1 - r^2 \approx \sqrt{3/N}$ (dotted line). For mixed states with $1 - r^2 \gg \sqrt{3/N}$, accuracy increases slightly with the amount of hedging (quantified by β). For slightly mixed states with $1 - r^2 \approx \sqrt{3/N}$, accuracy improves substantially with hedging, but only up to $\beta \approx \frac{1}{2}$. For nearly pure states with $1 - r^2 \ll \sqrt{3/N}$, a small amount of hedging improves accuracy, but higher β increases error, and the optimal β decreases with N.

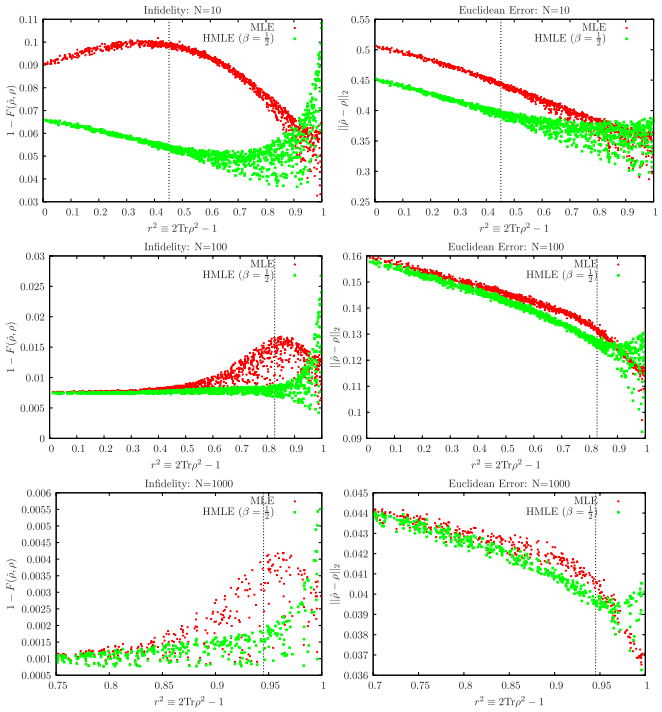

FIG. 2 (color online). Methodology: See Fig. 1. MLE and HMLE (with several β values) estimates were calculated, and for each ρ, Euclidean distance and infidelity were averaged over all data sets. (Trace and Euclidean distances are equivalent for qubits: $\|\hat{\rho} - \rho\| = \sqrt{2}\|\hat{\rho} - \rho\|_2$.) Results: As in Fig. 1, there are three regimes. Hedging provides a small but consistent improvement for highly mixed states, substantial improvement for slightly mixed states, but decreases accuracy on pure states. Because baseline inaccuracy is lower for near-pure states, hedging achieves better overall (i.e., worst-case) accuracy than MLE. Choosing β between 0.25 and 1 seems optimal.

plausible as $\hat{\mathbf{p}}_{\mathrm{MLE}}$. Actually, $\hat{\mathbf{p}}_{\mathrm{MLE}}$ comprises $K - 1$ independent parameters, and in this case likelihood ratios between e^{-K} and e^{K} are insignificant. [Typically, $\mathcal{L}(\mathbf{p}_{\mathrm{true}}) \approx e^{-K}\mathcal{L}(\hat{\mathbf{p}}_{\mathrm{MLE}})$, so tighter significance criteria would reject the true state.] If $\hat{\mathbf{p}}_{\mathrm{MLE}}$ assigns zero probability to $M < K$ different events, and $\hat{\mathbf{p}}_H$ hedges all M of them, then the argument leading to Eq. (13) gives a likelihood ratio of $e^{-M\beta}$, which is not significant. Quantum HMLE satisfies a very similar condition (see Appendix A [22]),

$$\frac{\mathcal{L}(\hat{\rho}_H)}{\mathcal{L}(\hat{\rho}_{\mathrm{MLE}})} \geq e^{-d\beta}, \tag{14}$$

so we do not pay a large price for the benefits of hedging—$\hat{\rho}_H$ is not significantly less plausible than $\hat{\rho}_{\mathrm{MLE}}$.

Conclusions.—Hedging is a simple, easy-to-implement solution to the zero eigenvalue problem. HMLE can be implemented by a near-trivial change to any MLE routine, and may even be easier than MLE. The hedged likelihood goes smoothly to zero near the boundary, so no explicit positivity constraint is needed, and simple gradient-crawling methods should work. $\hat{\rho}_H$ is always full rank, so it can be used for predictive tasks like gambling and data compression. For qubits, HMLE provides improved accuracy by almost all metrics. Very small values of β are best for nearly pure states, and in general the optimal β is not clear. This contrast with the classical case, where $\beta \approx \frac{1}{2}$ is known to be asymptotically optimal, suggests that alternative hedging functions may work better for quantum estimation.

I thank Alexei Gilchrist, Daniel James, Jordan LaPointe, and Rob Spekkens for discussions, and the Government of Canada (through Industry Canada) and the Province of Ontario (through the Ministry of Research & Innovation) for support.

*robin@blumekohout.com

[1] M. A. Nielsen and I. L. Chuang, *Quantum Computation and Quantum Information* (Cambridge University Press, Cambridge, England, 2000).
[2] Z. Hradil, Phys. Rev. A **55**, R1561 (1997).
[3] D. F. V. James *et al.*, Phys. Rev. A **64**, 052312 (2001).
[4] Z. Hradil *et al.*, Lect. Notes Phys. **649**, 59 (2004).
[5] H. Haeffner *et al.*, Nature (London) **438**, 643 (2005).
[6] R. Blume-Kohout, New J. Phys. **12**, 043034 (2010).
[7] G. J. Lidstone, Trans. Fac. Actuaries **8**, 182 (1920).
[8] E. Ristad, arXiv:cmp-lg/9508012.
[9] T. H. Cover and J. A. Thomas, *Elements of Information Theory* (Wiley-InterScience, New York, 1991).
[10] Q. Xie and A. Barron, IEEE Trans. Inf. Theory **46**, 431 (2000).
[11] There are many different ways to gamble; many (including investment) can be transformed into this form. For lack of space in this Letter, the reader is directed to the extensive literature.
[12] Several distinct POVMs $\mathcal{M}_j$, performed on disjoint subsamples of size N_j, can be treated as a single POVM $\mathcal{M} = \bigcup_j w_j \mathcal{M}_j$, where $w_j = N_j/N$.
[13] K. Vogel and H. Risken, Phys. Rev. A **40**, 2847 (1989).
[14] If these equations are overcomplete, $\hat{\rho}_{\mathrm{tomo}}$ is chosen by least-squares fitting.
[15] Quantum data compression is well studied, beginning with [16]. Typical-subspace algorithms (e.g., Ref. [16]) fail catastrophically whenever $\hat{\rho} \neq \rho$, but quantum variable-length compression algorithms exist [17], and suffer relative-entropy penalties for error. Quantum gambling has not been thoroughly studied, but a simple model presented in Ref. [18] illustrates a relative entropy penalty.
[16] B. Schumacher, Phys. Rev. A **51**, 2738 (1995).
[17] S. L. Braunstein *et al.*, IEEE Trans. Inf. Theory **46**, 1644 (2000).
[18] R. Blume-Kohout and P. Hayden, arXiv:quant-ph/0603116.
[19] B. Clarke and A. Barron, J. Stat. Plann. Infer. **41**, 37 (1994).
[20] R. Krichevskiy, IEEE Trans. Inf. Theory **44**, 296 (1998).
[21] D. Braess *et al.*, Lect. Notes Comput. Sci. **2533**, 153 (2002).
[22] R. Blume-Kohout, arXiv:1001.2029.
[23] C. H. Bennett, A. W. Harrow, and S. Lloyd, Phys. Rev. A **73**, 032336 (2006).
[24] H. Jeffreys, *Theory of Probability* (Oxford University, New York, 1998).