A practical and efficient approach for Bayesian quantum state estimation

Joseph M Lukens, Kody J H Law, Ajay Jasra and Pavel Lougovski

1 Quantum Information Science Group, Oak Ridge National Laboratory, Oak Ridge, Tennessee 37831, United States of America
2 School of Mathematics, University of Manchester, Manchester, M13 9PL, United Kingdom
3 Computer, Electrical and Mathematical Science and Engineering Division, King Abdullah University of Science and Technology, Thuwal, 23955-6900, Saudi Arabia

E-mail: lukensjm@ornl.gov and kodylaw@gmail.com

Keywords: quantum tomography, Bayesian estimation, sampling methods

Abstract

Bayesian inference is a powerful paradigm for quantum state tomography, treating uncertainty in meaningful and informative ways. Yet the numerical challenges associated with sampling from complex probability distributions hampers Bayesian tomography in practical settings. In this article, we introduce an improved, self-contained approach for Bayesian quantum state estimation. Leveraging advances in machine learning and statistics, our formulation relies on highly efficient preconditioned Crank–Nicolson sampling and a pseudo-likelihood. We theoretically analyze the computational cost, and provide explicit examples of inference for both actual and simulated datasets, illustrating improved performance with respect to existing approaches.

1. Introduction

Quantum state tomography (QST) is of fundamental importance in quantum information processing, where realization of computational advantages rests critically on the quality of the underlying quantum resources. In general, QST seeks to estimate the density matrix $\rho$ describing a given state, utilizing the results of measurements on repeated state preparations [1]. As an encapsulation of the quantum state’s properties, the density matrix facilitates quantitative predictions of quantum information protocols, clarifies the effects and sources of noise, and provides the foundation for analyzing entire circuits via quantum process tomography [2, 3].

Yet QST is notoriously challenging for all but the smallest quantum systems. The Hilbert space of a collection of qubits grows exponentially with the number of particles, as does the number of independent quantities needed to fully characterize $\rho$. Indeed, such exponential scaling is the source of the unique computational power inherent in quantum information, and accordingly QST cannot be used for characterizing large-scale QIP systems of the future, at least in their entirety. However, there remains demand for efficient and informative QST techniques that make the most of available resources and push limits on system size. In this vein, Bayesian methods offer exciting promise. Built upon Bayes’ rule for updating a prior probability distribution according to new information (measurements in the context of quantum tomography), Bayesian QST returns a complete probability distribution on $\rho$, quantifying uncertainty in a natural way, utilizing all available information optimally (in terms of minimizing an operational divergence), and avoiding unjustifiably optimistic estimates of low rank [4]. While Bayesian sampling approaches have been applied in several quantum optical experiments [5–7], the numerical challenge of drawing from high-dimensional probability distributions impedes widespread use in the physics community.

In this work, we propose, analyze, and demonstrate a full Bayesian tomography method that is straightforward to implement and numerically efficient. Our stand-alone approach leverages recent developments in multiple fields, including density matrix parameterization [8], PAC-Bayesian machine
learning [9], and Markov chain Monte Carlo (MCMC) algorithms [10]. After introducing the algorithm in detail, we test it on experimental two-qubit data, obtaining a $\sim 3.5\times$ speedup in our custom Metropolis–Hastings method over slice sampling. Additionally, with the aid of simulated data of much higher-dimensional two-qudit measurements, we observe a computational scaling advantage utilizing a pseudo-likelihood in favor of a full multinomial likelihood, demonstrating accurate estimation of states with Hilbert space dimensions up to 49—corresponding to 4851 real parameters. Overall, our method represents an improvement over previous Bayesian QST approaches and should provide a valuable tool for comprehensive, yet numerically efficient, state estimation.

2. Background

In formulating the general problem, consider a system of $n$ qudits—$d$-level quantum information carriers. The Hilbert space dimensionality is then $D = d^n$, and the $D \times D$ density matrix $\rho$ describing a state requires $D^2 - 1$ real numbers for specification. In order to designate a physically realizable state, $\rho$ must be (i) normalized ($\Tr_\rho = 1$), (ii) Hermitian ($\rho^\dagger = \rho$), and (iii) positive semi-definite ($\langle \psi|\rho|\psi \rangle \geq 0$ for all unit-norm $D$-dimensional states $|\psi\rangle$). Historically, three major approaches have been adopted to estimate $\rho$ from measurements.

Linear inversion—the first method considered in quantum information processing, linear inversion tomography relies on the fact that measurement outcome probabilities are linear functions of the individual elements comprising $\rho$ [1]. Thus, if a sufficient number of measurements have been performed to access all $D^2 - 1$ parameters of $\rho$—and the outcome frequencies are equated with these probabilities directly—one can enlist, e.g., least-squares (LS) inversion to obtain an estimate $\rho_{LS}$. While straightforward, LS tends to return nonphysical states: normalization and hermiticity can easily be enforced, but positive semi-definiteness cannot be.

Maximum likelihood—maximum likelihood estimation (MLE) finds the density matrix which is most likely to have produced the observed data $\mathcal{D}$:

$$\rho_{\text{MLE}} = \arg \max_\rho L_\mathcal{D}(\rho),$$

where $L_\mathcal{D} \propto P(\mathcal{D}|\rho)$, the probability of receiving the particular set of outcomes given state $\rho$, as defined by some model [11, 12]. Through appropriate parameterization of $\rho$, this method guarantees a result satisfying all physicality constraints. This advantage has made MLE the dominant approach to QST in recent years. However, as seen in equation (1), $\rho_{\text{MLE}}$ is a point estimate and so does not quantify the level of uncertainty in the result. In practice, error bars have been obtained by modifying the observations according to, e.g., a Poissonian noise model and computing many MLE estimates [13], a procedure which amounts to simulating further experiments and averaging the MLE results obtained from these. While likely to give reasonable estimates, this approach is somewhat ad hoc and conceptually undesirable, as it involves feeding in additional data beyond that obtained experimentally.

Bayesian—the third and least explored approach, Bayesian QST [4, 5, 14] accounts for experimental uncertainty explicitly through Bayes’ theorem. Suppose $\rho(x)$ is parameterized by some vector $x$, such that any value within $x$’s support returns a physical matrix. Bayes’ theorem states that the posterior probability distribution of $x$, given results $\mathcal{D}$ of some experiment, follows via

$$\pi(x) = \frac{1}{Z} L_\mathcal{D}(x) \pi_0(x),$$

where $L_\mathcal{D}(x)$ is the likelihood (as in MLE), $\pi_0(x)$ is the prior distribution (any beliefs about $\rho$ before the experiment), and $Z$ is a normalizing constant such that $\int \! dx \pi(x) = 1$. With access to $\pi(x)$, the expectation value of any function $\phi$ of $\rho$ can be obtained

$$\langle \phi(\rho) \rangle = \int \! dx \, \pi(x) \phi(\rho(x)),$$

which can be used to compute, e.g., the mean and standard deviation of any quantity of interest.

Nevertheless, evaluating integrals of the form in equation (3) is numerically challenging due to their generally complicated features and high dimensionality, even for moderate-size systems (e.g., two qubits). Accordingly, MCMC methods have been invoked in the literature, such as Metropolis–Hastings [4, 8], sequential Monte Carlo (SMC) [14] and slice sampling [5]. These approaches are designed, in most cases, to
obtain \( R \) samples \( \{x^{(1)}, x^{(2)}, \ldots, x^{(R)}\} \), so that equation (3) can then be approximated as

\[
\langle \phi(\rho) \rangle \approx \frac{1}{R} \sum_{r=1}^{R} \phi \left( \rho(x^{(r)}) \right).
\]  

(4)

Slice sampling in particular is an effective and quite general MCMC method, requiring no proposal distributions and largely insensitive to initial step settings \([15, 16]\). However, the computing time required for convergence can easily make these methods intractable for systems of interest. A major motivation for the current work rests in the realization that a more tailored sampling method—focused on the specific density matrix parameterization and robust to increases in system dimensionality—can attain significant computational speedups.

Finally, before describing in detail the procedure introduced here, we note an alternative view of Bayesian tomography associated with adaptive QST. In this application, Bayes’ theorem is invoked in real-time, with the results from previous measurements used to hone in subsequent measurement choices and reduce the total number of bases required for reconstruction \([17–21]\). While beyond the scope of the present work, where we concentrate on Bayesian state reconstruction post-experiment, we could certainly envision incorporating aspects of the approach into adaptive QST as well.

## 3. Proposed method

### 3.1. Steps

We now outline our proposed Bayesian QST workflow, summarized visually in figure 1. (The steps will be explained in detail in section 3.2.)

(a) Perform a set of measurements on unknown state \( \rho \), amounting to a total of \( N \) individual outcomes (over all measurement settings).

(b) Compute the least-squares estimate \( \rho_{LS} \). If the number of measurements is tomographically incomplete, \( \rho_{LS} \) lives in a subspace spanned by only those directions which were observed, which we can express through the function \( P_M(\cdot) \), i.e., \( \rho_{LS} = [P_M(\rho)]_{LS} \).

(c) Parameterize the \( D \times D \) density matrix by \( D \) nonnegative real numbers, \( y_k \), and \( D \) complex column vectors of length \( D \), \( z_k \). The density matrix for parameter set \( x = \{y_1, \ldots, y_D, z_1, \ldots, z_D\} \) is then

\[
\rho(x) = \sum_{k=1}^{D} \left( \frac{y_k}{\sum_i y_i} \right) \frac{z_k z_k^\dagger}{|z_k|^2}. 
\]  

(5)

This satisfies all physicality conditions.

(d) Take the prior distribution for \( x \) as

\[
\pi_0(x) \propto \prod_{k=1}^{D} y_k^{\alpha-1} e^{-y_k} e^{-\frac{1}{2} \sum_i z_i^\dagger z_i},
\]  

(6)

which amounts to treating the weights as Gamma-distributed random variables \( Y_k \sim \Gamma(\alpha, 1) \) and the vectors as standard-normal complex Gaussians \( Z_k \sim \mathcal{CN}(0, I_D) \).

(e) Define the pseudo-likelihood as

\[
L_D(x) = \exp \left( -\frac{N}{2} \| P_M(\rho(x)) - \rho_{LS} \|_F^2 \right),
\]  

(7)

with \( \| A \|_F \equiv \sqrt{\text{Tr}(A^\dagger A)} \) denoting the Frobenius norm and \( P_M(\cdot) \) the projection introduced in step (b).

(f) Draw \( R \) samples according to the preconditioned Crank–Nicolson (pCN) Metropolis–Hastings procedure of invariant distribution \( \pi(x) \propto L_D(x) \pi_0(x) \), detailed in algorithm 1.

(g) From these samples, estimate any function of \( \rho \) via equation (4).

### 3.2. Further details on specific steps

**Parameterization**—we have opted for the parameterization and prior employed by Mai and Alquier \([8]\), which expresses the density matrix as a superposition of normalized (though non-orthogonal) projectors. Incidentally, this represents an over-parameterization, in that it relies on a total of \( 2D^2 + D \) real numbers,
rather than the minimum of $D^2 - 1$ required for a $D \times D$ density matrix. We have found this parameterization significantly more efficient to sample from and evaluate than the Cholesky approach of references [5, 22] (see appendix for more details). For example, computing the determinant in the integration measure of reference [5]—needed to preserve Haar invariance [23]—requires $O(D^6)$ operations for a given draw. On the other hand, the current over-parameterization utilizes a simple Cartesian differential,

$$dx = \prod_{k=1}^{D} \prod_{l=1}^{D} d(Re z_{k,l})d(Im z_{k,l}),$$

(8)

where $z_{k,m}$ denotes the $m$th component of the complex vector $z_k$. Constructing $\rho$ given $x$ requires only $O(D^3)$ operations (equation (5)), offsetting the small overhead incurred from the additional parameters.

The prior distribution (equation (6)), also from reference [8], is specified by one user-adjustable value, $\alpha$, which can be used to favor low- or high-rank $\rho$, i.e., pure or mixed states, respectively. The collection of $D$ normalized random variables, $Y_k/(\sum_l Y_l)$, with $Y_k \sim \Gamma(\alpha)$, follows a Dirichlet distribution Dir($\alpha$), which guarantees both normalization and nonnegativity; $\alpha = 1$ represents a fully uniform prior, with equal weight given to all physically realizable states, while $\alpha < 1$ favors sparse Dirichlet draws [24] and hence purer states. (It is important to note that Haar invariance of the prior $\pi_0(x)$ obtains for any choice of $\alpha$, due to rotational symmetry of the normal distribution.) Finally, the complex Gaussian vectors, when normalized to $Z_k/|Z_k|$, correspond to uniform draws from the complex unit hypersphere.

Pseudo-likelihood—the particular $L_D(x)$ chosen in equation (7) is ‘pseudo’ in that it does not proceed from an explicit experimental model, but rather merely assigns a loss function between a proposed $\rho(x)$ and experimental data, in this case the least-squares estimate $\rho_{LS}$. Growing in popularity in the context of ‘probably approximately correct’ (PAC) Bayesian machine learning [9], pseudo-likelihoods are useful when a first-principles model is either unknown or too complex to compute efficiently. The downside is the need to separately specify the weight between evidence and prior, as controlled by the constant appearing in the pseudo-likelihood expression $\exp(\text{const} \times \text{loss})$. The larger its value, the more sharply peaked around $\rho_{LS}$ the posterior distribution becomes. As one of the strengths of Bayesian QST lies in its quantification of estimator uncertainty, it is essential that this scale factor reflect confidence levels commensurate with the amount of data gathered. For a quadratic loss function as in figure 1(e), one can associate this constant with $1/(2\sigma^2)$, with $\sigma^2$ the variance. If we take $N$ as the total number of events utilized in the LS estimate, it is reasonable to assume $\sigma^2 \propto 1/N$, although the specific proportionality factor is unclear. Reference [8] conjectures $\sigma^2 = 2/N$ as optimal, but in the absence of more rigorous motivation, we select $\sigma^2 = 1/N$, which in the examples below leads to uncertainties comparable to that of a full likelihood, albeit slightly larger. In general, more thorough methods for selecting the variance represent an important direction for future research.

In steps (b) and (e), we propose treating cases of incomplete measurements by projecting onto only those elements of $\rho$ which are accessed in the experiment, expressed formally through the function $P_M(\cdot)$. Consider the decomposition of a $D \times D$ density matrix in terms of $D^2 - 1$ traceless, Hermitian generators.
\[ \rho = \frac{1}{D} I_D + \frac{1}{2} \sum_{k=1}^{D^2 - 1} \alpha_k \lambda_k. \]  

(9)

In light of orthogonality \((\text{Tr} \lambda_k \lambda_j = 2 \delta_{ij})\), we have \(\alpha_k = \text{Tr} \rho \lambda_k\). Thus, incomplete measurements reflect that only a subset of the \(D^2 - 1\) observables can be estimated through linear inversion. Suppose that \(K_M\) denotes this subset of indices; then we define \(P_M(\cdot)\) as

\[ P_M(\rho) = \frac{1}{D} I_D + \frac{1}{2} \sum_{k \in K_M} \text{Tr}(\rho \lambda_k) \lambda_k. \]  

(10)

For example, in the data utilized in section 4, measurements were sensitive to eight of the fifteen coefficients required to specify a two-qubit state \(|K_M| = 8\). By contrast, a tomographically complete experiment corresponds to \(P_M(\rho) = \rho\). We emphasize that this particular projector definition is merely a convenient choice, and it differs from the prob-estimator in reference [8].

Additionally, even though we have presented a pseudo-likelihood formulation in defining the proposed method, the basic features can be readily applied to a full (model-infused) likelihood as well. Suppose an experiment consists of \(Q\) positive-operator valued measures (POVMs) \(\Lambda^{(q)}\), each with \(S_q\) total outcomes, associated with operators in the set \(\Lambda^{(q)} = \{\Lambda^{(q)}_1, \ldots, \Lambda^{(q)}_{S_q}\}\). Then the full likelihood is the multinomial expression

\[ L_D(x) = \prod_{q=1}^{Q} \prod_{i=1}^{S_q} \left[ \text{Tr}(\rho(x) \Lambda^{(q)}_i) \right]^{N^{(q)}_i}, \]  

(11)

where the outcome associated with \(\Lambda^{(q)}_i\) is observed \(N^{(q)}_i\) times. Computationally speaking, considering \(Q\) measurements, each with \(S_q = D\) outcomes (as in standard projective measurement), evaluating this likelihood requires \(\mathcal{O}(QD^3)\) operations. On the other hand, the cost of evaluating the pseudo-likelihood depends on the specifics of the projection operation \(P_M(\cdot)\). For \(Q \sim \mathcal{O}(D)\) tomographically complete measurements—such as mutually unbiased bases (MUBs) [25]—\(P_M(\rho) = \rho\) and no projection is necessary, thus leaving a total evaluation cost of \(\mathcal{O}(D^2)\) for the pseudo-likelihood, compared to \(\mathcal{O}(D^4)\) for the full likelihood. For incomplete measurements, however, \(P_M(\rho)\) must be explicitly computed, thus increasing the pseudo-likelihood’s evaluation cost. For example, in the case of \(Q \sim \mathcal{O}(D)\) but not a tomographically complete set, evaluating the pseudo-likelihood can increase to \(\mathcal{O}(D^4)\) operations. Thus, the pseudo-likelihood in our formulation is expected to impart a computational speedup for complete measurements but not necessarily for the incomplete case, a situation which is consistent with the results of sections 4 and 5.

**Sampling algorithm**—the most challenging feature of Bayesian methods, sampling the posterior distribution faces slow convergence that becomes arbitrarily slow as dimension increases. In 2013, Cotter et al [10] introduced a transformative approach to MCMC sampling which eliminates this ‘curse of dimensionality’ for Gaussian priors, under appropriate assumptions on the likelihood. Titled ‘preconditioned Crank–Nicolson’ (pCN for short), pCN modifies standard random-walk Metropolis sampling by scaling the previous iteration’s position before adding a random shift and generating the proposal \(x'\). In algorithm 1, pCN appears specifically in the factor \(\sqrt{1 - \beta_2^2}\) in step 2. This small modification simplifies the acceptance probability \(A(x', x^{(0)})\) significantly with respect to a standard random walk proposal, by removing terms of the form \(|x'^{(0)}|^2 - |x|^2\) from the exponent. The difference in these terms can be large, which necessitates a smaller stepsize to maintain a given acceptance rate. Alternatively, independence sampling from the prior \((\beta_2 = 1)\) also removes these terms, but the acceptance probability in that case is determined by the ratio of the likelihood at two independent prior samples, which one can expect to be large if the likelihood varies substantially over the support of the prior, i.e., if the posterior differs significantly from the prior (see appendix for more details). Therefore, unlike both standard random-walk Metropolis and independence sampling, the proposal here preserves random walk behavior and provides a simplified acceptance probability.

The specific expression for \(A(x', x^{(0)})\) follows from the standard form for Metropolis–Hastings [16]. Letting \(p(x'|x^{(0)})\) denote the proposal density, we have

\[ A(x', x^{(0)}) = \min \left\{ 1, \frac{p(x'|x^{(0)}) \pi(x^{(0)})}{\pi(x') p(x^{(0)}|x')} \right\}. \]  

(12)

Making use of the densities for the proposal distribution, \(Y^{(j)}_k|x^{(0)} \sim \lognormal(\log Y^{(j)}_k, \beta^2_2)\) and \(Z^{(j)}_k|x^{(0)} \sim \mathcal{CN}(\sqrt{1 - \beta^2_2} Z^{(j)}_k, \beta^2_2 I_D)\), as well as equation (2), (6), and (7), returns the formula in step 3. For
Algorithm 1. pCN sampling procedure

1. Choose the stepsizes $\beta_x, \beta_z \in (0, 1)$. Set $j = 0$, and draw $x^0$ from the prior (equation (6)).
2. For $k \in \{1, \ldots, D\}$, draw $y^{1, k} \sim \mathcal{N}(0,1)$ and independently $\xi_k \sim \mathcal{CN}(0, I_4)$. Then propose the new point $x' = \{y'_1, \ldots, y'_Q, x'_Z\}$, according to
   
   \[ y'_1 = \alpha y^0 + (1 - \alpha) y^{1, k}, \quad y'_2 = \sqrt{1 - \beta_x^2} y^0 + \beta_x \xi_1, \quad \text{and} \quad x'_Z = \sqrt{1 - \beta_z^2} x^0 + \beta_z \xi_k. \]

3. Set $x^{j+1} = x'$ with probability $A(x', x^0)$, where
   
   \[ \log A(x', x^0) = \min \left\{ 0, \log L_p(x') - \log L_p(x^0) + \sum_{i=1}^D \left[ \alpha \log y'_i - y_i^0 - \alpha \log y^{0, i} + y^{0, i}_i \right] \right\}. \]

   Otherwise set $x^{j+1} = x^0$. Increment $j$ by one and return to step 2.

Efficient convergence in the sampling algorithm, we monitor the acceptance rate and increase or decrease the step sizes $\beta_x$ and $\beta_z$ in tandem to maintain an acceptance fraction between 0.1 and 0.3. This range is chosen to enclose 0.234, the optimum acceptance probability, under various assumptions, for random-walk Metropolis–Hastings [26]. Additionally, we note that the adaptation diminishes as the chain evolves, so it does not preclude ergodicity [27].

4. Example with experimental data

To explore the effectiveness of the new method, we perform QST on results from the frequency-bin quantum optics experiment of reference [28], specifically the measurements in figure 4 thereof. We selected this experiment for comparison because: (i) its basis set is tomographically incomplete, thus enabling use of the projector formulation in figure 1(b); and (ii) the results were already analyzed with the method of reference [5], providing an initial reference point. In the following, we perform numerical benchmarking of method performance for a variety of configurations. All tests were completed in 64 bit MATLAB utilizing a single thread on a 2.5 GHz machine with 128 GB of RAM.

For the first test, we focus on the speed of the pCN sampling method, using slice sampling with MATLAB’s built-in algorithm for comparison. In this example, we take $\alpha = 1$ for a uniform prior and invoke the full likelihood of equation (11). We increase the number of points in the Markov chain and monitor the mean ($F$) and standard deviation $\Delta F$ of the fidelity $F(x)$ and standard deviation $\Delta F$ of the fidelity $F(x) = \langle \Psi | \rho(x) | \Psi \rangle$, where $|\Psi\rangle = \frac{1}{\sqrt{2}} (|0\rangle + |1\rangle)$ is the ideal entangled state; $R = 2^{10}$ samples are kept from a total of $RT$, with $T$ the thinning parameter used to reduce serial correlation in the chain. For each value of $T$, we run 100 independent samplers (i.e., with random initial points), returning 100 separate estimates of $F$ and $\Delta F$. The results for slice sampling appear in figure 2(a), plotted against total time logged by each sampler. Thinning increases by factors of two from $2^0$ to $2^7$ on this plot, and we use a box plot format to summarize the statistics at each $T$: the center mark denotes the median, upper and lower lines enclose the 25th–75th percentiles, and the whiskers extend to the smaller of the farthest point or 1.5× the length of the box. Fidelity converges to $F = 0.93 \pm 0.01$, slightly higher than the mean of 0.92 found in reference [28]. This difference is unsurprising, though, since here we do not consider single most (i.e., events where only one of the two photons is detected), so our likelihood model differs.

Figure 2(b) furnishes results for the identical test performed with our custom pCN algorithm, for thinning from $2^0$ to $2^7$. The total times are lower by approximately $18 \times$ for the same value of $T$. Examining the codes in detail indicates this difference is caused by the fact the slice sampler evaluates $\pi(x)$ many more times than pCN. Nevertheless, figure 2 reveals pCN’s need for larger thinness values to reach the same level of convergence as slice, so it is not clear a priori what, if any, quantitative advantage is obtained.

Accordingly, we next plot $\Delta F$ for both slice and pCN on the same logarithmic scale in figure 3. Initially, both approaches obtain a reduction in $\Delta F$ with log–log slope of $-1/2$ (i.e., $\Delta F \propto \text{time}^{-1/2}$), until converging to final values. Linear least-squares fits to the first five and seven points of the slice and pCN curves, respectively, give a $\sim 3.5 \times$ temporal speedup for pCN over slice at the same convergence level. Such an improvement—even for this comparatively small system of two qubits—is significant for practical QST, where computational time represents a precious commodity.

In the second test, we shift focus away from the sampling procedure and concentrate on the likelihood, comparing the full (equation (11)) and pseudo (equation (7)) versions directly. Since the experiment in question measured in combinations of the Pauli-X and Z bases for the two qubits, but not Pauli-Y, the experimental LS estimate consists of only eight of the fifteen total Pauli basis components, thus requiring the projector formalism in figure 1(b). From a computational perspective, this projection can be efficiently implemented as a linear transformation, by writing the density matrix elements as a length-$D^2$ column vector $\rho_{vec}$ and finding the matrix $\mathcal{V}$ such that $[P_{\mathcal{M}}(\rho)]_{vec} = \mathcal{V} \rho_{vec}$, which can be precomputed according to the relationship between Pauli and computational basis representations [29]. Likewise, the probabilities
Figure 2. State fidelity estimates for the example experimental data for independent samplers as the Markov chain length (and hence total time) increases, for (a) slice sampling and (b) the proposed pCN algorithm.

Figure 3. Convergence of $\Delta F$ in the results of figure 2. The fits have log–log slopes of $-1/2$, with the pCN curve shifted to the left by a factor of $\sim 3.5 \times$.

appearing in the full likelihood (equation (11)) can be vectorized so that $\log L_D = N^T \log W \rho$, where $N$ denotes the vector of counts and $W$ is the linear transformation mapping matrix elements to probabilities. Matrix-vector multiplication reduces function evaluation time and is essential in providing a fair comparison between the likelihood approaches.

Figure 4(a) plots $\langle F \rangle$ and $\Delta F$ for 100 samplers utilizing the pCN algorithm, as the number of points increases, for both the full (left) and psuedo (right) likelihood models; both consider $\alpha = 1$ in the prior. (The full likelihood results are the same as figure 2(b), reproduced here for comparison.) Both likelihoods converge to similar values, though the pseudo case returns slightly lower mean and higher uncertainty. The general congruity between the two cases offers evidence in favor of our choice for the variance ($\sigma^2 = 1/N$) and the projector-based approach to limited measurements. Even the slightly lower fidelity for the pseudo-likelihood is a positive feature, in that it does not overestimate the state’s fidelity beyond that predicted by a complete model.

One of the advantages of the prior formulation is its ability to impart sparsity to the density matrix parameterization, through $\alpha$, and thus favor pure states. In figure 4(b), we consider $\alpha = 1/4$ and repeat the convergence tests with all other settings the same. In both the full and pseudo cases, the fidelity is slightly higher compared to $\alpha = 1$, which makes sense in light of the extra weight given to pure states. Interestingly, the full-likelihood case shows additional outliers at this $\alpha$ value (note the much wider $y$-axis scale). Evidently, the sparser prior increases the tendency for trapping of the Markov chain around local maxima. By contrast, the pseudo-likelihood results remain much more consistent throughout the convergence plot. While it would be unwise to infer too much from acknowledged outliers, the pseudo-likelihood approach nonetheless appears slightly more robust to fluctuations in the sampling algorithm, a valuable feature for Bayesian QST.
Yet the pseudo-likelihood does not lead to any observable speedup in sampler time. In this particular example ($D = 4$ and 16 total measurement outcomes), such a situation obtains, first, because computational cost is dominated not by calculating $L_D(x)$ but rather constructing $\rho(x)$ (equation (5)) and, second, because of the additional cost of computing the projection $P_M(\rho)$ in this limited measurement case. As dimension $D$ increases, though, the pseudo-likelihood’s improved efficiency should ultimately surface, a question we address with simulated data in the next section.

5. Examples with simulated data

In order to explore dimensionalities beyond that of the experimental data available to us, we next generate simulated tomographic data for entangled two-qudit states with the ground truth density matrix

$$\rho = \lambda |\Psi\rangle\langle\Psi| + \frac{1 - \lambda}{D} I_D,$$

(13)

where $|\Psi\rangle = \sum_{k=1}^{d} |k\rangle_M |\lambda\rangle_B$ is a high-dimensional Bell state, $D = d^2$, and $I_D$ is the $D \times D$ identity matrix. $\lambda \in [0, 1]$ controls the fidelity with respect to the ideal $|\Psi\rangle$; i.e., fidelity $F$ and $\lambda$ are related via $F = (\lambda - 1)^2 + 1$. Count data are obtained by cycling through all pairwise combinations of $(d + 1)$ MUBs \[25\], computing the $d^2$ outcome probabilities associated with the state in equation (13), and drawing from a multinomial to emulate an experimental coincidence distribution. This procedure amounts to $Q = (d + 1)^2 = (D + 2\sqrt{D} + 1)$ total measurement settings, each with $S_d = D$ outcomes, which are then either used to compute $q_{LS}$ and perform pseudo-likelihood–based QST or inserted directly as exponents in the full likelihood.

Explicitly, we consider qudit dimensions $d \in \{2, 3, 5, 7\}$ (Hilbert space dimensions $D \in \{4, 9, 25, 49\}$). Prime $d$ are chosen for convenience, for a complete $(d + 1)$ set of MUBs can be generated easily in these cases utilizing Weyl operators \[30\]. We then set $\lambda = 0.95$ and acquire 100 $D$ coincidences for each pair of bases in the simulated experiments. Running the pCN algorithm on these observations and recording the time per sample for $2^{14}$ points (following a burn-in period of $2^{10}$ points), we find the trends in figure 5. While comparable at low $D$, the evaluation times for the full and pseudo-likelihood approaches become increasingly disparate as $D$ grows, reaching $\sim 10 \times$ for $D = 49$. Due to limits on the size of the datasets we could generate, we were not able to reach the $O(D^3)$ asymptotic scaling improvement of section 3. Nonetheless, these time tests confirm that the pseudo-likelihood offers computational speedups under appropriate conditions.

The practical value of this observed speedup becomes evident in continuing with full tomography of these high-dimensional states, utilizing our pCN sampling algorithm and the pseudo-likelihood formalism—the full workflow of figure 1. Specifically, we consider a prior with $\alpha = 1$ and individual MCMC chains of length $RT$, doubling the thinning value $T$ until convergence is reached for each input dataset; we define convergence as the condition beyond which larger thinning returns the same $\langle F \rangle \pm \Delta F$ (with $\Delta F$ rounded to its most significant digit). In order to confirm comparable performance for noise levels other than $\lambda = 0.95$, we generate datasets for $\lambda = 0.85$ and 0.75 as well, and also for dimensions $D = \{4, 9, 25, 49\}$ and 100 $D$ events per basis pair. The converged fidelities for all cases are provided in figure 6. Each point shows the mean $\langle F \rangle$ and standard deviation $\Delta F$ of $R = 2^{10}$ samples, labeled by the time

![Figure 4](image-url). Convergence plots for the priors (a) $\alpha = 1$ and (b) $\alpha = 1/4$, utilizing 100 independent pCN samplers for increasing Markov chain lengths.
Figure 5. Computational cost of our algorithm for simulated data with increasing dimension. Circles indicate the mean, and error bars the standard deviation, for $2^{14}$ points.

Figure 6. Tomographic reconstruction of simulated datasets. Each fidelity and error bar is the mean and standard deviation, respectively, of $2^{10}$ samples obtained with our pCN method and pseudo-likelihood; labels show the total sampler time. The curves give the ground truth fidelities of the states used to generate the datasets for each $\lambda$.

 elapsed to complete the sampling loop (rounded to 1 s); the curves trace out the ground truth fidelities as set by the specific $\lambda$ and $D$. Overall, the estimates are in good agreement with the underlying states. The sampler time increases with $D$ as expected, due to the additional computational cost per point in the chain (figure 5), as well as the need for longer chains to reach convergence as $D$ grows. Importantly, though, the performance variation with $\lambda$ appears to be small. States with lower fidelities do seem to converge slightly faster, yet reconstruction accuracy is comparable for all $\lambda$, highlighting the generality of our method.

Successful realization of Bayesian QST on such large systems—4851 unknowns for $D = 49$ ($2D^2 + D$ total real parameters)—reveals the strengths of our proposed method in expanding the dimensionalities that can be addressed through Bayesian quantum state estimation.

6. Conclusion

Continued research should enable even further improvements to the sampling algorithm. While here we have applied the pCN approach to parameters with Gaussian prior distributions, pCN can be extended to non-Gaussian priors as well [31], provided one can select a proposal distribution which preserves reversibility with respect to the prior. It is worth noting also that there exist additional enhancements, for example utilizing derivative information, which are out of scope of the present work [32–34]. The method can also be embedded within SMC samplers [35, 36], and as $D \to \infty$ one can leverage finite approximations to further improve complexity [36]. These directions are under investigation and will be reported in future work.

In view of current classical computer architectures in the beyond Moore’s law era, parallel implementation is an increasingly important consideration moving forward. For very large systems, the current methods are amenable to parallelization within each serial step of the MCMC using standard batching. As an example, the proposals for the $D$ variables $y_k$, and the real and imaginary parts of each of the $D$ $D$-dimensional complex variables $z_k$ can be simulated in parallel. Then for the construction of $\rho$ each scalar multiple of an outer product can be computed in parallel, with each row of the outer product also
computed in parallel, and the sum of these rank one matrices can then be executed logarithmically by computing pairwise sums in parallel at each step (first $D/2$ pairs, then the resulting $D/4$ pairs, etc). Parallel chains can also be leveraged, as in reference [37] (see also references therein), but this is complicated and controversial, given the intrinsically serial nature of MCMC methods. In reference [37] the authors propose to utilize a convergence diagnostic based on comparing between and within chain statistics of independent chains, and they argue that once the diagnostic indicates that all chains have converged, then the samples can be pooled together to construct estimators. However, parallelization is cleanest when using either population-type methods [38] or else debiasing techniques [39, 40]. Investigations are also ongoing regarding the implementation of these strategies.

As it stands in its current form, the key features of the present method should find application in a myriad of quantum inference problems. With only minor modifications, models previously tackled by slice sampling—such as photon loss [5] or linear-optic transformations with dark counts [6]—can be transformed into a pCN formalism for algorithmic speedup. On the other hand, the impact of the pseudo-likelihood in improving practical QST is less clear in our view. The pseudo-likelihood certainly appears more robust to initial conditions of the sampler (figure 4(b)), with computational improvements for sufficiently large Hilbert spaces, thus enabling Bayesian QST at previously unreached system sizes (figure 6). Yet at the dimensionalities where the pseudo-likelihood provides critical speedups (e.g., $D \sim 50$ here), it is possible that the experimental challenge of acquiring the needed QST data will so outweigh the computational challenge of evaluating the full likelihood as to render the pseudo-likelihood superfluous. That being said, the pseudo-likelihood’s general, model-independent form could provide advantages which may not be evident in the specific QST problem of interest here, so that the full potential of the pseudo-likelihood remains an unanswered question.

In conclusion, we have introduced a Bayesian inference method for efficient quantum state tomography. Compatible with any number of observations, our approach enjoys all the standard advantages of Bayesian QST but with significantly improved computational efficiency, through a combination of well-chosen parameterization, likelihood, and MCMC sampling algorithm. Our numerical investigations on both real and simulated data confirm the promise of our approach, particularly the power of advanced statistical techniques such as pCN in practical quantum tomography.

**Acknowledgments**

We thank R S Bennink and B P Williams for discussions. This work was funded by the U.S. Department of Energy, Office of Advanced Scientific Computing Research, through the Quantum Algorithm Teams and Early Career Research Programs. This work was performed in part at Oak Ridge National Laboratory, operated by UT-Battelle for the U.S. Department of Energy under contract no. DE-AC05-00OR22725. MATLAB source code for the results in this article are available at https://github.com/jmlukens/BayesQuanTom.

**Appendix A. Comparing with alternative methods**

In this article, we have focused on benchmarking our proposed pCN sampling method against slice sampling, and the pseudo-likelihood against a full multinomial likelihood, since these are the closest contenders in terms of efficiency that we have found for the problem of interest here. In this appendix, we provide some additional results validating our choices made, specifically (i) the use of a projector-based parameterization over Cholesky and (ii) use of a modified random walk rather than independence sampling.

Previous Bayesian QST papers [5, 22] have parameterized a $D$-dimensional density matrix through the Cholesky decomposition $\rho = LL^\dagger$, where $L$ is a lower triangular matrix with positive real diagonal elements, defined in terms of $D^2 - 1$ hyperspherical coordinates. (See appendix A of reference [5] for a complete definition.) While an informationally efficient parameterization in that it relies on the smallest possible number of independent real numbers, it is computationally demanding. For if we let $\tau_j$ ($j \in \{1, 2, \ldots, D^2 - 1\}$) denote each hyperspherical parameter, maintaining Haar invariance requires computation of the integration measure $\sqrt{g}$, where $g_{jk} = \text{Tr} \left( \frac{\partial L}{\partial \tau_j} \cdot \frac{\partial L}{\partial \tau_k} \right)$. This is a highly involved calculation requiring $O(D^6)$ operations and is the reason behind our selection of the parameterization in equation (5) instead. In figure A1(a) we provide an example of just how wide this speed gap is, even for the $D = 4$ case of section 4. This plot gives the fidelity and uncertainty recovered from slice sampling on single chains, as a function of total sampler time (as in the main text, we keep $2^{10}$ samples and increase thinning to
reach convergence). We focus on slice sampling here since the hyperspherical parameterization is not readily expressed in terms of normal or gamma random variables. The total times for the points at which convergence is visually achieved are circled: the Cholesky parameterization is over $17 \times$ slower than the projector version in this case. While it is certainly possible that research into alternative formulations of the integration measure could enable improvements in the Cholesky approach, it seems clear for now that the projector version is far superior for efficient state estimation.

In terms of sampling methodologies, both our version of pCN and slice sampling are built on Markov chains where each proposed point depends on the current one. However, it is possible to consider variations of independence sampling, where each proposed sample—or a subset of the parameters comprising it—is drawn independently of the previous. Such an approach is equivalent to importance sampling and is not able to find regions of high probability efficiently, unless the importance distribution is very close to the target. Therefore it tends to suffer from much slower convergence than the random walk type algorithms we consider here. To give an idea of this difference, we compare our pCN sampling procedure (algorithm 1) against a basic independence sampler where each proposal $x' \sim$ is drawn directly from the prior; i.e., the proposal density is $p(x' | x) = \pi_0(x')$. In this case, the Metropolis acceptance probability (step 3 of algorithm 1) reduces to

$$
\log A(x',x) = \min \{0, \log L_D(x') - \log L_D(x)\}. 
$$

(A.1)

Running a convergence test in the same manner as figure A1(a) (full likelihood, uniform prior, experimental two-qubit dataset, and increasing thin), we find the results in figure A1(b). While pCN converges with a chain taking only 5 s to complete, a massive $1 \times 10^5$ s ($\sim 33$ h) is not sufficient to confirm convergence in the independence sampler. This example underlines the vast efficiency improvement in well-designed MCMC methods, and reinforces our focus on pCN and slice sampling in the main text.

ORCID iDs

Joseph M Lukens https://orcid.org/0000-0001-9650-4462
Kody J H Law https://orcid.org/0000-0003-3133-2537
Pavel Lougovski https://orcid.org/0000-0002-9229-1444

References

[1] Nielsen M A and Chuang I L 2000 *Quantum Computation and Quantum Information* (Cambridge: Cambridge University Press)
[2] Chuang I L and Nielsen M A 1997 *J. Mod. Opt.* 44 2455
[3] Poyatos J F, Cirac J I and Zoller P 1997 *Phys. Rev. Lett.* 78 390
[4] Blume-Kohout R 2010 *New J. Phys.* 12 043034
[5] Williams B P and Lougovski P 2017 *New J. Phys.* 19 043003
[6] Lu H-H, Lukens J M, Williams B P, Imany P, Peters N A, Weiner A M and Lougovski P 2019 *npj Quantum Inf.* 5 24
[7] Williams B P, Lukens J M, Peters N A, Qi B and Grice W P 2019 *Phys. Rev. A* 99 062311
[8] Mai T T and Alquier P 2017 *J. Stat. Plan. Inference* 184 62
[9] Guedj B 2019 arXiv:1901.05353
[10] Cotter S L, Roberts G O, Stuart A M and White D 2013 *Stat. Sci.* 28 424
[11] Hradil Z 1997 *Phys. Rev. A* 55 R1561
[12] James D F V, Kwiat P G, Munro W J and White A G 2001 *Phys. Rev. A* 64 052312
[13] Altepeter J, Jeffrey E and Kwiat P 2005 *Adv. At. Mol. Opt. Phys.* 52 105
[14] Granade C, Combes J and Cory D G 2016 New J. Phys. 18 033024
[15] Neal R M 2003 Ann. Stat. 31 705
[16] MacKay D J C 2003 Information Theory, Inference, and Learning Algorithms (Cambridge: Cambridge University Press)
[17] Huszár F and Houdsby N M T 2012 Phys. Rev. A 85 052120
[18] Kravtsov K S, Straupe S S, Radchenko I V, Houdsby N M T, Huszár F and Kulik S P 2013 Phys. Rev. A 87 062122
[19] Struchalin G I, Pogorelov I A, Straupe S S, Kravtsov K S, Radchenko I V and Kulik S P 2016 Phys. Rev. A 93 012103
[20] Pogorelov I A, Struchalin G I, Straupe S S, Radchenko I V, Kravtsov K S and Kulik S P 2017 Phys. Rev. A 95 012302
[21] Granade C, Ferrie C and Flammini S T 2017 New J. Phys. 19 113017
[22] Seah Y-L, Shang J, Ng H K, Nott D J and Englert B-G 2015 New J. Phys. 17 043018
[23] Fyodorov Y V 2004 arXiv:math-ph/0412017
[24] Telgarsky M 2013 arXiv:1301.4917
[25] Wootters W K and Fields B D 1989 Ann. Phys. 191 363
[26] Roberts G O, Gelman A and Gilks W R 1997 Ann. Appl. Probab. 7 110
[27] Roberts G O and Rosenthal J S 2007 J. Appl. Probab. 44 458
[28] Lu H-H, Lukens J M, Peters N A, Williams B P, Weiner A M and Lougovski P 2018 Optica 5 1455
[29] Gamel O 2016 Phys. Rev. A 93 062320
[30] Sheridan L and Scarani V 2010 Phys. Rev. A 82 030301
[31] Vollmer S J 2015 SIAM/ASA J. Uncertain. Quantif. 3 535
[32] Law K J 2014 J. Comput. Appl. Math. 262 127
[33] Cui T, Law K J and Marzouk Y M 2016 J. Comput. Phys. 304 109
[34] Beskos A, Girolami M, Lan S, Farrell P E and Stuart A M 2017 J. Comput. Phys. 335 327
[35] Kantas N, Beskos A and Jasra A 2014 SIAM/ASA J. Uncertain. Quantif. 2 464
[36] Beskos A, Jasra A, Law K, Marzouk Y and Zhou Y 2018 SIAM/ASA J. Uncertain. Quantif. 6 762
[37] Chen Y, Keyes D, Law K J H and Ltaief H 2016 SIAM J. Sci. Comput. 38 5539
[38] Del Moral P 2004 Feynman–Kac Formulae (Berlin: Springer)
[39] Glynn P W and Rhee C-H 2014 J. Appl. Probab. 51A 377
[40] Jacob P E, O’Leary J and Atchadé Y F 2017 arXiv:1708.03625