# Hamiltonian Learning and Certification Using Quantum Resources

Nathan Wiebe, <sup>1, 2, 3</sup> Christopher Granade, <sup>4, 3</sup> Christopher Ferrie, <sup>5</sup> and D. G. Cory <sup>6, 3, 7</sup>

<sup>1</sup> Quantum Architectures and Computation Group, Microsoft Research, Redmond, WA 98052, USA

<sup>2</sup> Department of Combinatorics & Optimization, University of Waterloo, Ontario N2L 3G1, Canada

<sup>3</sup> Institute for Quantum Computing, University of Waterloo, Ontario N2L 3G1, Canada

<sup>4</sup> Department of Physics, University of Waterloo, Ontario N2L 3G1, Canada

<sup>5</sup> Center for Quantum Information and Control, University of New Mexico, Albuquerque, New Mexico, 87131-0001

<sup>6</sup> Department of Chemistry, University of Waterloo, Ontario N2L 3G1, Canada

<sup>7</sup> Perimeter Institute, University of Waterloo, Ontario N2L 2Y5, Canada

In recent years quantum simulation has made great strides culminating in experiments that operate in a regime that existing supercomputers cannot easily simulate. Although this raises the possibility that special purpose analog quantum simulators may be able to perform computational tasks that existing computers cannot, it also introduces a major challenge: certifying that the quantum simulator is in fact simulating the correct quantum dynamics. We provide an algorithm that, under relatively weak assumptions, can be used to efficiently infer the Hamiltonian of a large but untrusted quantum simulator using a trusted quantum simulator. We illustrate the power of this approach by showing numerically that it can inexpensively learn the Hamiltonians for large frustrated Ising models, demonstrating that quantum resources can make certifying analog quantum simulators tractable.

Quantum information processing promises to dramatically advance physics and chemistry by providing efficient simulators for the Schrödinger or Dirac equations [1–3]. This is important because conventional methods are inefficient, scaling exponentially in the number of interacting subsystems. Consequently, quantum simulations beyond a few tens of interacting particles are generally believed to be beyond the limitations of conventional supercomputers. This inability to simulate large quantum systems means that important questions in condensed matter, such as the shape of the phase diagram for the Fermi-Hubbard model, remain open. Analog quantum simulation raises the possibility that special purpose analog devices may be able to address such problems using current or near-future hardware [4-6]. A major objection to this avenue of inquiry is that analog simulators are not necessarily trustworthy [7, 8] and certification of them is not known to be efficient. Without such certification, an analog simulator can at best only provide hints about the answer to a given computational question. A resolution to this problem is therefore essential if analog quantum simulators are to compete on an even footing with classical supercomputers.

An important first step towards a resolution is provided in [9], where it is shown that quantum systems with local time—independent Hamiltonians can be efficiently characterized given ensemble readout. However, the method is not generally applicable, can be expensive and is not known to be either error robust or stable in cases where single shot measurements are used. A number of machine learning and statistical inference methods [10–17] have been recently introduced to address similar problems in metrology or Hamiltonian learning. In the context of Hamiltonian learning, such ideas have are known to be error—robust and lead to substantial reductions in the cost of high—precision Hamiltonian inference [15], albeit at the price of sacrificing the efficient

scaling exhibited by [9].

We overcome these challenges by providing a robust method that can be used to characterize unknown Hamiltonians by unifying statistical inference with quantum simulation. The key insight behind this is that Bayesian inference reduces the problem of Hamiltonian estimation to a problem in Hamiltonian simulation that can be efficiently solved using a trusted quantum simulator. Our algorithm achieves this through the following steps. We begin by positing a Hamiltonian model for the system and a probability distribution over the parameters of the Hamiltonian model. We then use a novel guess heuristic for the optimal experiment that adaptively chooses experiments based on the current uncertainty in the Hamiltonian. The experiment is then performed and the trusted quantum simulator is used to efficiently compute the likelihood of the measurement outcome occurring if each hypothetical model were true. These likelihoods are then used by the algorithm to update its knowledge of the Hamiltonian parameter via Bayes rule, resulting in an updated probability distribution, called the posterior distribution. This process is then repeated until the uncertainty in the unknown Hamiltonian parameters (as measured by the posterior variance) becomes sufficiently small. This iterative process is depicted in Figure 1.

To make the problem concrete, we represent each hypothetical Hamiltonian  $H_j$  by a vector of real numbers  $\mathbf{x}_j \in \mathbb{R}^d$  such that  $H_j = H(\mathbf{x}_j)$ . The Hamiltonian model is therefore specified by  $H(\mathbf{x})$ .

We consider three classes of experiments that can be performed to infer the Hamiltonian, H, given an initial state  $|\psi\rangle$  (typically a pseudorandom state [18]): (a) Classical Likelihood Evaluation (CLE), (b) Quantum Likelihood Evaluation (QLE) and (c) Interactive Quantum Likelihood Evaluation (IQLE). CLE is the simplest of these experiments and is discussed in detail in [15]. It involves simply picking an experimental time t, and com-

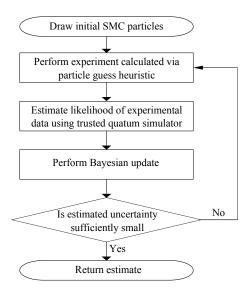


FIG. 1: Flowchart for Hamiltonian learning algorithm.

puting the likelihood  $\Pr(D|\mathbf{x}_i) = |\langle D|e^{-iH(\mathbf{x}_i)t}|\psi\rangle|^2$  using a classical computer, where  $\mathbf{x}_i$  is a given set of Hamiltonian parameters and D is the experimental outcome. This function, known as the likelihood function, will not generally be efficiently computable on a classical computer because it involves quantum simulation.

In QLE experiments, a trusted quantum simulator is used to ameliorate these problems. It does so by estimating  $\Pr(D|\mathbf{x}_i)$  to be the fraction of times outcome D occurs in a sufficiently large set of simulated experiments, which is efficient if  $\Pr(D|\mathbf{x}_i)$  is only polynomially small. This approach allows a complex quantum simulator, such as a fault tolerant quantum computer, to act as a certifier for an analog quantum simulator. A trusted quantum simulator could also be constructed using a bootstrapping protocol wherein a smaller trusted analog simulator is the certifier. This is possible if a compressed simulation scheme [19] for the dynamics exists.

The Loschmidt echo famously shows that, for complex quantum systems, two nearly identical Hamiltonians will typically generate evolutions that diverge exponentially after a short time, before saturating at an exponentially small overlap [20]. This means that QLE will often be restricted to short evolution times to guarantee efficiency (which is undesirable [15]). We resolve this by using IQLE experiments, which are described in Figure 2. These experiments are reminiscent of the Hahn echo experiments commonly used in magnetic resonance and experimental quantum information processing [21]. An IQLE experiment swaps the state of the unknown quantum system with that of a trusted quantum simulator then inverts the evolution based on a guessed Hamiltonian  $H_{-}$ . The measurement in IQLE is always assumed to be in an orthonormal basis that has  $|\psi\rangle$  as an element. This produces  $\Pr(D|\mathbf{x}_i) = |\langle D|e^{iH_-t}e^{-iH(\mathbf{x}_i)t}|\psi\rangle|^2$ .

Although the Loschmidt echo may also seem to be problematic for IQLE experiments, we exploit it in our

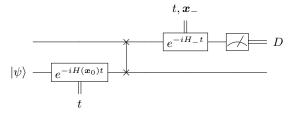


FIG. 2: IQLE. The upper register is the trusted simulator, while the lower register is the system under study. QLE is conceptually similar but with  $H_{-}=0$ .

guess heuristic for  $H_{-}$ . We call this heuristic the "particle guess heuristic" (PGH), which chooses  $H_{-} := H(\boldsymbol{x}_{-})$ by sampling  $x_{-}$  from the prior probability Pr(x), which describes our current knowledge of the Hamiltonian parameters. The set of parameters,  $x_{-}$ , is called a particle because it is described by a Dirac-delta distribution over parameter space. The time t is chosen by drawing a second particle  $x'_{-} \neq x_{-}$  and setting  $t = 1/\|x'_{-} - x_{-}\|_{2}$ . As the uncertainty in the estimated parameter shrinks, the PGH adaptively picks longer times to ensure that informative experiments continue to be chosen as certainty about the unknown parameters increases. The PGH also causes  $e^{-iH(\boldsymbol{x}_i)t}$  to result in substantially different likelihoods for  $x_i$  that are within one standard deviation of the prior's mean, which we show in the appendix is optimal for certain learning problems.

IQLE experiments with two outcomes also ensure that Pr(D|x) will not be exponentially small (with high probability) for H an affine transformation acting on x, since

$$|\langle \psi | e^{iH_{-}t} e^{-iHt} | \psi \rangle| \ge 1 - 2||H - H_{-}||_{2}t$$
  
  $\ge 1 - O(||x - x_{-}||_{2}t).$  (1)

If the prior distribution has converged to a unimodal distribution centered near the correct Hamiltonian (this is typical for Bayesian inference of non-degenerate learning problems [15]) then  $\|\boldsymbol{x}-\boldsymbol{x}_-\|_2 \in \Theta(1/t)$ . This means that if we use a POVM with two elements:  $|\psi\rangle\langle\psi|$  and its orthogonal compliment  $\mathbb{1}-|\psi\rangle\langle\psi|$  then we expect (a) neither probability will be exponentially small if  $\boldsymbol{x}_-$  and  $\boldsymbol{x}_j$  are near the mean and (b)  $\Pr(\psi|\boldsymbol{x}_j)$  will typically be exponentially small for  $H(\boldsymbol{x}_j)$  that differ substantially from the correct Hamiltonian. The PGH therefore leads to IQLE experiments that rapidly eliminate incorrect hypotheses about the correct Hamiltonian.

The measurement outcomes yielded by the experiments are immediately processed using Bayesian inference, as described in Figure 1. This immediate processing allows our algorithm to adaptively choose experiments based on its current knowledge of the correct Hamiltonian. The state of knowledge is represented by a distribution that is called, previous to the next update step, the prior. In the cases we consider, the initial prior distribution before any data is observed is taken to be uniform. This encodes a state of maximum ignorance about the correct  $\boldsymbol{x}$ . The prior distribution is updated as mea-

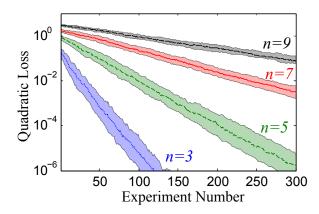


FIG. 3: The quadratic loss plotted as a function of the number of inversion experiments for Ising models on the complete graph. The shaded areas show a 50% confidence interval for the quadratic loss.

surement outcomes are recorded using Bayes' rule, which gives the proper way of computing the probability of each  $x_j$  being correct given the observed data and the prior. It states that if datum D is recorded then

$$\Pr(\boldsymbol{x}_j|D) \propto \Pr(D|\boldsymbol{x}_j) \Pr(\boldsymbol{x}_j),$$
 (2)

up to a normalization factor and  $Pr(x_j|D)$  is called the posterior distribution.

Eq. (2) can be efficiently computed (for a polynomial number of  $H_j$ ) only if the likelihood function  $\Pr(D|\mathbf{x}_j)$  is tractable. QLE and IQLE experiments allows  $\Pr(D|\mathbf{x}_j)$  to be efficiently estimated, which removes the main obstacle to using Bayesian methods to learn the correct  $\mathbf{x}$ .

A secondary problem is that *exact* computations of the update rule are intractable in practice because an infinite number of Hamiltonians could potentially describe the system; hence, a probability distribution over Hamiltonians cannot be exactly represented on either a classical or quantum computer. This problem can be addressed by using the sequential Monte Carlo (SMC) approximation [15, 22, 23], which approximates the probability distribution using a weighted sum of particles (Dirac delta functions). Each particle corresponds to a particlar  $x_i$ and is a hypothesis about the correct Hamiltonian parameters x. SMC assigns a weight  $w_i$  to each particle that represents the probability of that hypothesis. The weights are normalized such that  $\sum_{j} w_{j} = 1$ . The update rule for the probability distribution under the SMC approximation then becomes  $w_i \mapsto \Pr(D|x_i)w_i$ , followed by normalization. If necessary, a resampling step is used after updating to ensure that the inference procedure remains stable, as discussed in [15] and in Appendix B.

The algorithm then iteratively updates the weights  $w_i$  and positions  $x_i$  of the sequential Monte Carlo particles representing the distribution over Hamiltonians  $\Pr(H|D)$ , conditioned on the data recorded at each step. In this way, the full state of knowledge at each step is iteratively carried forward, and is used to heuristically

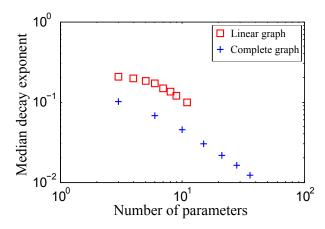


FIG. 4: The median decay exponent for the quadratic loss as a function of the number of parameters in the Ising model, d.

design future experiments according to the PGH. Subsequent updates will then refine this estimate of the unknown Hamiltonian parameter until the uncertainty of the estimated Hamiltonian is sufficiently small, as measured by the trace of the posterior covariance matrix.

Now that we have discussed how our algorithm works, we will proceed to assess its cost. We will show that the cost of Hamiltonian inference on a fixed number of IQLE experiments is exponentially smaller than the cost of using CLE. This is significant because CLE gives the best known methods for some problems [15].

A natural measure of the cost is the number of quantum simulations needed to estimate the Hamiltonian parameters. The total cost is therefore,

$$Cost = N_{steps}(\delta) \times Cost(update; \epsilon).$$
 (3)

Here  $N_{\rm steps}$  is the number of updates needed to make the uncertainty less than  $\delta$  and  $\rm Cost(update;\epsilon)$  is the number of samples from the trusted simulator that are needed to update the particle weights using Eq. (2) within error  $\epsilon$  in the 1–norm. We show in Appendix D that, with high probability,  $\rm Cost(update;\epsilon)$  scales as

$$\frac{|\{\boldsymbol{x}_i\}|}{\epsilon^2} \left( \mathbb{E}_{D|H} \left[ \frac{\max_k \Pr(D|\boldsymbol{x}_k)(1 - \Pr(D|\boldsymbol{x}_k))}{\left(\sum_k \Pr(D|\boldsymbol{x}_k) \Pr(\boldsymbol{x}_k)\right)^2} \right] \right)$$

This implies that the update process will be efficient if the number of particles required is small and the resultant probability distribution is not too flat. That is,  $|\{x_i\}| \in O(\operatorname{poly}(n))$  and  $\sum_k \Pr(D|H(x_k)) \Pr(H(x_k)) \in O(1/\operatorname{poly}(n))$ , where n is the number of interacting systems. It has been shown that SMC algorithms require a number of particles that scales sub–exponentially in d [24], which itself may not be a function of n. This means that in practice, a small number of particles will typically be required. The robustness of the algorithm to sampling errors is discussed in [25] as well as in Appendix D, so relatively large  $\epsilon$  can be tolerated.

If the posterior distribution has converged to a unimodal distribution such that  $\boldsymbol{x}$  is within a fixed distance from the mean, then the PGH and (2) ensure that  $\mathbb{E}_{H_-}[|\langle\psi|\,e^{iH_-t}e^{-iHt}\,|\psi\rangle\,|^2]\in\Theta(1)$  since  $t\in\Theta(|\boldsymbol{x}-\boldsymbol{x}_-|^{-1})$ . If a two outcome measurement is used then Markov's inequality implies that  $\sum_k \Pr(D|H(\boldsymbol{x}_k)) \Pr(H(\boldsymbol{x}_k))\in\Theta(1)$  with high probability. By a trivial generalization of this argument, it is clear that a super–polynomial reduction in the cost of performing (2) relative to CLE is obtained with high probability for IQLE experiments if  $d\in O(\operatorname{poly}(n))$  and the effective number of outcomes,  $\sum_j \Pr(j|\boldsymbol{x}_k)^{-2}$ , is at most  $O(\operatorname{poly}(n))$  for each  $\boldsymbol{x}_k$ .

In contrast, QLE experiments may not lead to a super–polynomial separation in the cost estimates for generic Hamiltonians and large t because  $\sum_k \Pr(D|H(\boldsymbol{x}_k)) \Pr(H(\boldsymbol{x}_k)) \in 2^{-\Theta(n)}$  with high probability for complex quantum systems [20, 26]. This can be rectified by choosing small t as per [9], but such QLE experiments will be much less informative [15].

If a fixed number of updates are required, then the previous discussion and (3) suggest that IQLE will provide an exponential advantage over CLE. If inference within a fixed error tolerance,  $\delta$ , is required then the cost estimate is much more challenging. Each two-outcome measurement yields at most one bit of information about H per measurement hence  $N_{\text{steps}}(\delta) \in \Omega(d\log_2(1/\delta))$ . For most models of interest, d is polynomial (or even constant) in n and hence a small number of updates should typically suffice. It is, however, unclear whether this lower bound is tight; hence, we turn to numerical evidence to show that our algorithm can efficiently learn Hamiltonians in certain cases.

Consider the problem of learning H(x) using IQLE experiments for an Ising model with no transverse field:

$$H(\boldsymbol{x}) = \sum_{(i,j)\in G} x_{i,j} \ \sigma_z^{(i)} \sigma_z^{(j)}, \tag{4}$$

where G is the edge set of an interaction graph on n qubits. Unless otherwise specified, we take  $x_{i,j} \in [-1/2,1/2]$  uniformly at random. We take the initial state for the evolution to be  $|\psi\rangle = |+\rangle^{\otimes n}$ . We choose this Hamiltonian not only because it is physically relevant [27], but also for numerical expediency, since the learning process require the algorithm to perform thousands of simulated evolutions of the initial state. All measurements are performed in the eigenbasis of  $X^{\otimes n}$ . Restricting the measurements to two outcomes is unnecessary for these experiments because IQLE and the PGH concentrates  $\Pr(D|\mathbf{x}_i)$  over a small number of outcomes for this Hamiltonian.

Figure 3 shows that the quadratic loss (a generalization of the mean–squared error for multiple parameters) shrinks exponentially with the number of experiments performed; however, the rate at which the error decreases slows as the number of qubits n increases. This is expected because d = n(n-1)/2 for the case of a complete

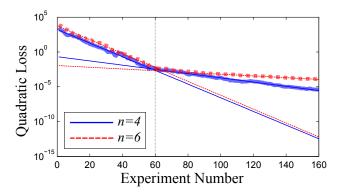


FIG. 5: An approximate 1 parameter Ising model on the complete graph. The thin lines give the best fits to the exponential decays, which scale as  $e^{-0.07N}$  and  $e^{-0.23N}$  for n=4 and as  $e^{-0.029N}$  and  $e^{-0.23N}$  for n=6 qubits.

interaction graph, which implies that the learning problem becomes more difficult as n increases. The data for interactions on the line is similar and is presented in Appendix A 1 and QLE data is given in Appendix A 2.

The rate at which the learning process slows as n increases is investigated in Figure 4. We examine the slowing of the learning problem by fitting the quadratic loss,  $\delta$ , in each experiment to  $Ae^{-\gamma N_{\rm steps}}$ . The median decay exponent, which is the median of the values of  $\gamma$  attained for a set of experiments with constant n, measures how rapidly the algorithm learns the unknown parameters. Figure 4 shows that these decay constants scale as O(1/d) for the complete graph, and provides weaker evidence for the line. This implies that  $N_{\rm steps}(\delta) = O(d\log(1/\delta))$  for this Hamiltonian, which implies that the inference is efficient. Similarly, the PGH implies that the total simulation time needed (for fixed  $|\{x_i\}|$ ) scales as  $N_{\rm steps}\delta^{-1} \approx \delta^{-3}$ , which is relevant in cases where the cost of a simulation is dominated by the evolution time.

Although d = n(n-1)/2 or d = n-1 in the examples considered above, d can be approximately independent of n in some cases. An example of this behavior is given in Figure 5, where we consider the case where each of the  $x_{i,j}$  is approximately the same value chosen uniformly on [0, 100], but with small normally distributed fluctuations with mean 0 and variance  $10^{-4}$ . This causes the learning problem to be effectively one-dimensional initially, and then transition to d = n when the small fluctuations need to be identified to learn the Hamiltonian parameters within a fixed accuracy. The transition from a singleto a multi-parameter learning problem happens at  $\delta \approx$  $d \times 10^{-4} \approx 10^{-3}$ , which coincides with the point when the slope in Figure 5 changes. This emphasizes that the cost of Hamiltonian estimation using our method only implicitly depends on n through d. In fact, the difference in the observed scaling of  $\gamma$  is approximately a factor of 2.5, which is what would be expected if  $\gamma \propto 1/d$ .

In conclusion, we have shown that Bayesian inference combined with the SMC approximation provides an ideal way to leverage a (potentially non–universal) quantum

simulator to characterize an unknown or unreliable quantum system. We provide theoretical evidence that shows that the update rule, which is at the heart of the learning algorithm, can be performed efficiently using quantum resources. We then illustrate the practicality of the algorithm and show that it is capable of learning unknown Ising couplings with surprisingly few experiments even in the presence of sampling errors. We will show elsewhere that the algorithm is highly resilient to depolarizing noise and other forms of noise that can be introduced via a noisy swap gate.

#### ACKNOWLEDGMENTS

We acknowledge Troy Borneman for suggesting bootstrapping and also Krysta Svore and Allan Geller for useful feedback and discussion. The numerical experiments performed here used SciPy, F2Py and QInfer [28–30]. This work was supported by funding from USARO-DTO, NSERC, CERC and CIFAR. CF was supported in part by NSF Grant Nos. PHY-1212445 and PHY-1005540.

- S. Lloyd et al., Universal quantum simulators, SCIENCE-NEW YORK THEN WASHINGTON- pp. 1073–1077 (1996).
- [2] A. Aspuru-Guzik, A. D. Dutoi, P. J. Love, and M. Head-Gordon, Simulated quantum computation of molecular energies, Science 309, 1704 (2005), http://www.sciencemag.org/content/309/5741/1704.full.pdf.
- [3] R. Gerritsma, G. Kirchmair, F. Zähringer, E. Solano, R. Blatt, and C. F. Roos, Quantum simulation of the Dirac equation, Nature (London) 463, 68 (2010).
- [4] J. Simon, W. S. Bakr, R. Ma, M. E. Tai, P. M. Preiss, and M. Greiner, Quantum simulation of antiferromagnetic spin chains in an optical lattice, Nature (London) 472, 307 (2011), 1103.1372.
- [5] J. W. Britton, B. C. Sawyer, A. C. Keith, C.-C. J. Wang, J. K. Freericks, H. Uys, M. J. Biercuk, and J. J. Bollinger, Engineered two-dimensional Ising interactions in a trapped-ion quantum simulator with hundreds of spins, Nature (London) 484, 489 (2012), 1204.5789.
- [6] K. Kim, M.-S. Chang, S. Korenblit, R. Islam, E. E. Edwards, J. K. Freericks, G.-D. Lin, L.-M. Duan, and C. Monroe, Quantum simulation of frustrated Ising spins with trapped ions, Nature (London) 465, 590 (2010).
- [7] P. Hauke, F. M. Cucchietti, L. Tagliacozzo, I. Deutsch, and M. Lewenstein, *Can one trust quantum simulators?*, Reports on Progress in Physics **75**, 082401 (2012).
- [8] C. Gogolin, M. Kliesch, L. Aolita, and J. Eisert, Bosonsampling in the light of sample complexity, arXiv preprint arXiv:1306.3995 (2013).
- [9] M. P. da Silva, O. Landon-Cardinal, and D. Poulin, Practical characterization of quantum devices without tomography, Phys. Rev. Lett. 107, 210404 (2011).
- [10] A. Hentschel and B. C. Sanders, Machine learning for precise quantum measurement, Physical Review Letters 104, 063603 (2010).
- [11] A. Hentschel and B. C. Sanders, Efficient algorithm for optimizing adaptive quantum metrology processes, Physical Review Letters 107, 233601 (2011).
- [12] A. Sergeevich, A. Chandran, J. Combes, S. D. Bartlett, and H. M. Wiseman, Characterization of a qubit Hamiltonian using adaptive measurements in a fixed basis, 1102.3700 (2011).
- [13] C. Ferrie, C. Granade, and D. Cory, How to best sample a periodic probability distribution, or on the accuracy of Hamiltonian finding strategies, Quantum Information Processing pp. 1–13 (2012), ISSN 1570-0755.
- [14] A. Sergeevich and S. D. Bartlett, Optimizing qubit

- hamiltonian parameter estimation algorithms using PSO, arXiv:1206.3830 (2012), proceedings of 2012 IEEE Conference on Evolutionary Computation (CEC), 10-15 June 2012.
- [15] C. E. Granade, C. Ferrie, N. Wiebe, and D. G. Cory, Robust online Hamiltonian learning, New Journal of Physics
   14, 103013 (2012), ISSN 1367-2630.
- [16] N. B. Lovett, C. Crosnier, M. Perarnau-Llobet, and B. C. Sanders, Differential evolution for many-particle adaptive quantum metrology, arXiv:1304.2246 (2013).
- [17] K. M. Svore, M. B. Hastings, and M. Freedman, Faster phase estimation, arXiv:1304.0741 (2013).
- [18] J. Emerson, Y. S. Weinstein, M. Saraceno, S. Lloyd, and D. G. Cory, Pseudo-random unitary operators for quantum information processing, Science 302, 2098 (2003), ISSN 0036-8075, 1095-9203, PMID: 14684815.
- [19] B. Kraus, Compressed Quantum Simulation of the Ising Model, Physical Review Letters 107, 250503 (2011), 1109.2455.
- [20] F. Haake, Quantum Signatures of Chaos 2<sup>nd</sup> Edition (Springer-Verlag New York, 2004).
- [21] E. L. Hahn, *Spin echoes*, Physical Review **80**, 580 (1950).
- [22] F. Huszr and N. M. T. Houlsby, Adaptive bayesian quantum tomography, 1107.0895 (2011).
- [23] A. Doucet and A. M. Johansen, A tutorial on particle filtering and smoothing: fifteen years later (2011).
- [24] A. Beskos, D. Crisan, and A. Jasra, arXiv e-print 1103.3965 (2011), URL http://arxiv.org/abs/1103. 3965.
- [25] C. Ferrie and C. E. Granade, Likelihood-free quantum inference: tomography without the Born rule, arXiv:1304.5828 (2013).
- [26] C. Ududec, N. Wiebe, and J. Emerson, Equilibration of Measurement Statistics Under Complex Dynamics, ArXiv e-prints (2012), 1208.3419.
- [27] P. Richerme, C. Senko, S. Korenblit, J. Smith, A. Lee, W. C. Campbell, and C. Monroe, Trapped-ion quantum simulation of an Ising model with transverse and longitudinal fields, arXiv:1303.6983 (2013).
- [28] E. Jones, T. Oliphant, P. Peterson, et al., SciPy: Open source scientific tools for Python (2001-), URL http://www.scipy.org/.
- [29] P. Peterson, F2PY: a tool for connecting Fortran and Python programs, International Journal of Computational Science and Engineering 4, 296 (2009).
- [30] C. Ferrie, C. Granade, et al., QInfer: Library for statistical inference in quantum information (2012–), URL

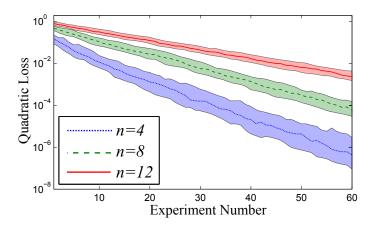


FIG. 6: This figure shows the quadratic loss plotted as a function of the number of IQLE experiments for 4, 8, 12 qubits (from bottom to top) interacting on the line. The dashed lines show a 50% confidence interval for the quadratic loss. 10 000, 10 000 and 20 000 particles were used in the n = 4, n = 8 and n = 12 cases respectively.

https://github.com/csferrie/python-qinfer.

- [31] J. Liu and M. West, Combined parameter and state estimation in simulation-based filtering (Springer-Verlag, 2000).
- [32] S. L. Braunstein and C. M. Caves, Statistical distance and the geometry of quantum states, Physical Review Letters 72, 3439 (1994).

## Appendix A: Supplemental Data

# 1. Error Scaling for Linear Interaction Graph

In the main body of the text, we showed that our algorithm learns information about the Hamiltonian at a rate that scales exponentially with the number of experiments taken for both the complete graph and the line, but only presented an example of the raw data for the case where the interaction graph is complete. For completeness, we provide here analogous data for the case where the interaction graph is a line.

This data clearly shows that IQLE experiments are similarly effective in the case of linear interaction graphs in that the data follows an exponential scaling. The fits of the median quadratic loss to an exponential of the form  $Ae^{-\gamma N}$ , where N is the experiment number, is given in Figure 4.

# 2. Error Scaling for QLE

A major problem facing the use of QLE experiments is efficiently estimating the likelihood function using quantum simulation. Despite this problem, if we grant the algorithm the ability to do perfect likelihood evaluations at unit cost then QLE experiments can be highly informative. For example, the typical variation of the likelihood function for QLE experiments with large t on random Hamiltonians acting on n qubits drawn from the Gaussian Unitary Ensemble (such random Hamiltonians model complex quantum systems with time-reversal symmetry [20]) is on the order of  $2^{-n}$  [26] which is on the same order as the typical values of the likelihood. This means that, if we do not consider sampling errors, then late time QLE experiments will allow learning to occur even for complex Hamiltonians.

In spite of this, Hamiltonian learning using QLE experiments is expected to be much less stable in this regime. This is because  $\Pr(D|\mathbf{x}_j)$  can be approximately the same as (or larger than)  $\Pr(D|\mathbf{x})$  even if  $||\mathbf{x}_k - \mathbf{x}||_2$  is large (here  $\mathbf{x}$  is the correct Hamiltonian parameter). This can cause the learning algorithm to get confused and move particles to near  $\mathbf{x}_k$  during the resampling step. This makes it harder for the algorithm to recover from the bad inference and continue to learn. Thus even if we grant QLE experiments the ability to perform exact likelihood evaluations at unit cost, then we still do not expect such experiments to be as robust to bad inferences as IQLE.

Figure 7 confirms these expectations. It shows that the  $25^{\text{th}}$  percentile of the quadratic loss for QLE experiments for the Ising model on the line is similar to that of IQLE experiments. The most notable difference between the data sets is that the 50% confidence intervals overlap. This is because, in each case, the learning algorithm is more likely to get confused in IQLE experiments versus QLE experiments. Additionally, the  $75^{\text{th}}$  percentile of the quadratic loss is  $much\ worse$  for n=12 and suggests that the learning algorithm eventually fails in such cases. Similar problems are

observed in the median and  $75^{\text{th}}$  percentile of the n=4 data. These problems are not fatal: they can be addressed by repeating the learning algorithm several times and using a majority vote scheme to reduce the impact of instances where the learning algorithm becomes confused. As mentioned previously, we expect real problems to emerge for QLE experiments when inexact likelihood calls are considered.

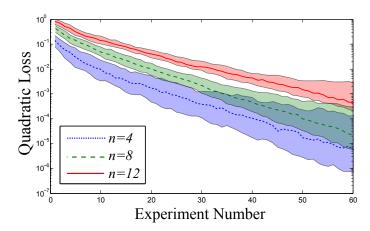


FIG. 7: The quadratic loss plotted as a function of the number of QLE experiments for 4, 8, 12 qubits (from bottom to top) interacting on the line. The dashed lines show a 50% confidence interval for the quadratic loss. 10 000, 10 000 and 20 000 particles were used in the n = 4, n = 8 and n = 12 cases respectively.

#### 3. Errors in Likelihood Evaluations

The numerical examples provided in the main body of the text assumed that the error in the inference process due to using a finite number of samples to compute the likelihoods  $\Pr(D|\mathbf{x}_i)$  is negligible. Here we provide evidence showing that the learning process is robust to such errors for IQLE experiments on the Ising models considered previously.

In particular, let us define  $\mathcal{P}$  to be the uncertainty in the estimated probability that results from estimating the likelihood in using the trusted quantum simulator. Here we simulate the use of MLE (Maximum Likelihood Estimation) or ALE (Adaptive Likelihood Estimation [25]) methods by adding normally distributed noise with zero mean and standard deviation  $\mathcal{P}$ , and then clip the likelihood to the interval [0, 1]. This is chosen in preference to MLE or ALE because it is expedient to compute and it models the results of either method closely. We find that even if  $\mathcal{P}$  is a large constant then our algorithm continues to reduce the quadratic loss at a rate that scales as  $e^{-\gamma N}$ ; albeit at a reduced value of  $\gamma$ . This clearly indicates that we do not need to take  $\epsilon$  to small if we need small error.

The robustness of Bayesian inference using the SMC approximation is illustrated in Figure 8, where we show that our algorithm is robust to sampling errors for 9 qubits interacting on the line for  $\mathcal{P} = 0.1$  and  $\mathcal{P} = 0$ . We see that the data for QLE experiments agrees with that of IQLE experiments for short times, which is expected because the probability distribution has not had time to reach its maximum support. At later times, QLE experiments with  $\mathcal{P} = 0.1$  fare much worse than IQLE experiments. Nonetheless, IQLE experiments (and QLE for this value of  $\mathcal{P}$ ) still exhibit exponential scaling of the error with the number of experiments. This may be surprising because the errors in  $\Pr(D|\mathbf{x}_i)$  can be as large as 0.1, which one may assume would be catastrophic given that many of the outcomes will have probability less than 0.1 in such models. We note that in particular, IQLE experiments are more robust to such noise than QLE experiments. This is because the inversion employed by IQLE concentrates the probability over a smaller number of outcomes; leading to smaller relative errors in the likelihood evaluations in such cases.

Figure 9 gives a more clear picture of the effects of sampling error on the resultant distribution for IQLE experiments. We observe that the presence of such noise does not qualitatively change the scaling of  $\gamma$ , where  $\gamma$  is the decay exponent that describes how the quadratic loss shrinks as more experimental data is provided. Specifically, we find that for  $\mathcal{P}=0, \ \gamma \propto d^{-1}$  whereas for  $\mathcal{P}=0.4/n$  (which corresponds to  $\epsilon \approx \sum_{i=1}^d 0.4/n \approx 0.4$ ) we find that  $\gamma \propto d^{-3/2}$ . This shows that large sampling errors do not necessarily prevent our algorithm from learning the Hamiltonian parameters at a rate that scales as  $Ae^{-\gamma N}$  and further suggests that this learning process is efficient for the problem of learning unknown Ising couplings. It also suggests that a constant value of  $\epsilon$  may suffice for certain experiments.

The surprising robustness of our method comes in part from the fact that the likelihood function must be approximated for each particle. This means that if the algorithm errs in the update of a particular particle due to inexact evaluation of the likelihood function then it may not err substantially in evolving the total probability in the region

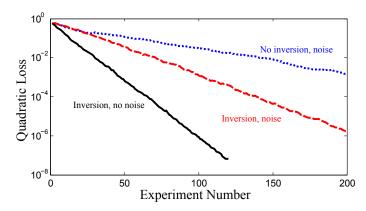


FIG. 8: This plot shows the median quadratic loss for a 9 qubit Ising model on the line for the cases where inversion is used and when inversion is not used for cases where the sample standard deviation is  $\mathcal{P}$  is the uncertainty in the estimated likelihood. 10 000 particles were used for the calculation.

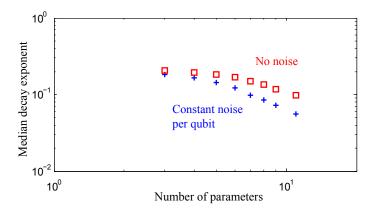


FIG. 9: This plot shows the median value of  $\gamma$  computed for the case of IQLE experiments where the interaction graph is a line and n ranges from 1 to 12 with different levels of noise. 20 000 particles were used for these numerical experiments.

that many such particles are in. For example, consider a region R that has 10 000 particles in it. The probability density in that region is then  $\sum_{x_i \in R} w_i/V(R)$ , where V(R) is the volume of the region. We then see that if errors are independent and identically distributed over each particle, then the total error in the update of the probability density will be roughly 1/100 the error that would be expected if all the errors were in fact correlated. Thus the robustness of the algorithm may be understood in part as a consequence of the fact that the errors are (approximately) unbiased about the true likelihood and that the particle number will typically be large for high precision inferences.

## Appendix B: Bayesian Inference of Hamiltonians

Sequential Monte Carlo (SMC) has before been considered in the context of quantum information [22], and in particular for its utility in estimating Hamiltonian dynamics [15]. Here, we summarize and review the sequential Monte Carlo algorithm and approximation, as SMC is an important tool for the practical implementation of statistical inference according to Bayes' rule, and in particular for our proposed methods. The SMC approximation is of central importance here because Hamiltonian models are typically parameterized by a vector of real numbers rather than discrete numbers. This means that there are an infinite number of hypothetical Hamiltonians that could represent the system, which makes the update of the prior distribution intractable. The SMC approximation is used to model the continuous distribution over model parameters (which is computationally difficult to sample from) by a discrete distribution that preserves the low–order moments of the distribution and thereby making estimation of the unknown Hamiltonian parameters tractable.

To clarify, suppose we have fixed an input state  $|\psi\rangle$  and measurement basis  $\{|D\rangle\}$ , but that the Hamiltonian under which the state evolves is unknown. Had we known that the Hamiltonian was H, we apply Born's rule to obtain the

probability distribution for the outcomes of the experiment:

$$\Pr(D|H) = |\langle D|e^{-iHt}|\psi\rangle|^2. \tag{B1}$$

This is called the *likelihood function*. When we write a probability distribution Pr(x|y), we are stating how likely the proposition x is true *given* y is known to be true. In a Hamiltonian learning problem, H is unknown and the measurement result is given. Bayes' rule provides a way to invert the conditioning to provide the probability that H is the true Hamiltonian given that datum D is recorded:

$$\Pr(H|D) = \frac{\Pr(D|H)\Pr(H)}{\Pr(D)} = \frac{\Pr(D|H)\Pr(H)}{\int \Pr(D|H)\Pr(H) dH}.$$
(B2)

Here, Pr(H) is called the *prior* and formally encodes any *a priori* knowledge of the unknown Hamiltonian. The probability of interest, Pr(H|D) is called the *posterior* since it encodes our *a posteriori* knowledge. The final term Pr(D) can simply be thought as a normalization factor that can be found implicitly by integrating over the unnormalized distribution. Since each measurement is statistically independent given H, the processing of the data can be done on- or off-line; Bayesian updating (or Bayesian learning or Bayesian inference) allows us to sequentially update our knowledge of the Hamiltonian through a sequence of probability distributions  $Pr(H|\{D_1, D_2, \ldots\})$ .

In practice, the Bayesian update rule and the expectations listed above are analytically and computationally intractable since they involve integrals over multidimensional parameter spaces. However, if we drop the requirement of a deterministic algorithm, we can efficient compute them using Monte Carlo techniques. Our numerical algorithm fits within the subclass of Monte Carlo methods called *sequential Monte Carlo* or SMC [23].

The first step in the approximation method is to think of H as a function that maps a parameterization x of a Hamiltonian to a Hermitian operator H(x). Doing so allows us to reduce the dimension of the random variable that we are reasoning about, called the model dimension, by using knowledge about the class of Hamiltonians which are plausible given the physics of the system. We then approximate the probability distribution by a weighted sum of Dirac delta-functions,

$$\Pr(H(\boldsymbol{x})) \approx \sum_{j=1}^{|\{\boldsymbol{x}_i\}|} w_j \delta(\boldsymbol{x} - \boldsymbol{x}_j),$$
(B3)

where the weights at each step are iteratively calculated from the previous step via

$$w_i \mapsto \Pr(D|\mathbf{x}_i)w_i,$$
 (B4)

followed by a normalization step. The elements of the set  $\{x_j\}_{j=0}^{|\{x_i\}|}$  are called *particles*. Here,  $|\{x_i\}|$  is the number of particles and controls the accuracy of the approximation. Like all Monte Carlo algorithms, the SMC algorithm approximates expectation values, such that

$$\mathbb{E}_{\boldsymbol{x}}[f(H(\boldsymbol{x}))] \approx \sum_{j=1}^{|\{\boldsymbol{x}_i\}|} w_j f(H(\boldsymbol{x}_j)). \tag{B5}$$

In other words, sequential Monte Carlo allows us to efficiently compute multidimensional integrals with respect to the measure defined by the probability distribution.

An iterative numerical algorithm such as SMC requires care to ensure stability. In the next section, we derive the conditions for stability of the algorithm. But first we describe one additional, and important, step in the iteration. The step is called *resampling* and is required to ensure that the SMC particles explore the space of Hamiltonians rather than staying fixed at the  $|\{x_i\}|$  initially chosen hypotheses. This is necessary both intuitively and, as we will see next, computationally.

The idea is simple: if the weight associated to a particle is too is small, move the particles to a region where the weight is large. We follow the methodology of Liu and West [31]. First, to determine when to resample, we compare the effective sample size  $N_{\rm ess} = 1/\sum_j w_j^2$  to a threshold (typically  $|\{x_i\}|/2$ ). If the threshold is not met, we randomly select  $|\{x_i\}|$  new particles according to the distribution of the current weights. Additionally, we incorporate randomness to search larger volumes of the parameter space. This randomness is inserted by applying a random perturbation to the location of each new particle. Thus, the new particles are randomly spread around the previous locations of the old. After drawing  $|\{x_i\}|$  new particles, we set the weight of each new particle to  $1/|\{x_i\}|$  so that  $N_{\rm ess} = |\{x_i\}|$ . To clarify, the Liu and West resampler algorithm updates the position of a particle  $x_i$ , which is sampled from the posterior distribution, by drawing a particle from a Gaussian distribution with mean

$$\boldsymbol{\mu}_i = a\boldsymbol{x}_i + (1-a)\boldsymbol{\mu},\tag{B6}$$

where  $\mu = \mathbb{E}_{x}[x]$  is the posterior mean of the particle location and  $a \in [0, 1]$  is a constant. The covariance matrix for the Gaussian distribution is given by

$$\Sigma = (1 - a^2) \operatorname{Cov}(\boldsymbol{x}), \tag{B7}$$

where Cov(x) is the covariance matrix for the particle positions. The resampler therefore introduces randomness into the problem that depends on the current level of uncertainty in the unknown Hamiltonian parameters. We find for the learning problems that we consider a = 0.9 performs well, whereas for simpler learning problems a = 0.98 was found to be superior [15, 31]. Full algorithmic details of SMC, including resampling, are given in [15].

The resultant posterior probability provides a full specification of our knowledge. However, in most applications, it is sufficient—and certainly more efficient—to summarize this distribution. In our context, the optimal single Hamiltonian to report is the mean of the posterior distribution (here, we have omitted for brevity the fact that the posterior is conditioned on the data)

$$\mu_H := \mathbb{E}_{\boldsymbol{x}}[H(\boldsymbol{x})] = \sum_{j=1}^{|\{\boldsymbol{x}_i\}|} w_j H(\boldsymbol{x}_j). \tag{B8}$$

However, a single point is the space of unknown Hamiltonians does not provide information of the uncertainties in this estimate. For that we turn to regions. In particular, the set of Hamiltonians X is an  $\alpha$ -credible region if

$$\Pr(H \in X) \ge 1 - \alpha. \tag{B9}$$

That is, a set is an  $\alpha$ -credible region if no more than  $\alpha$  probability mass exists outside the region or, equivalently, at least  $1-\alpha$  probability mass is contained within the region.

After a sufficient number of experiments, we assume the posterior distribution will be approximately Gaussian in terms of our chosen parameterization, so that  $\mathbf{x} \sim \mathcal{N}(\hat{\mathbf{x}}, \text{Cov}[\mathbf{x}])$ , where  $\hat{\mathbf{x}} = \mathbb{E}[\mathbf{x}]$ . Then, an  $\alpha$ -credible region estimate is given by the covariance ellipse

$$X = \{ H(\boldsymbol{x}) : (\boldsymbol{x} - \mathbb{E}_{\boldsymbol{x}}[\boldsymbol{x}])^{\mathrm{T}} \operatorname{Cov}[\boldsymbol{x}]^{-1} (\boldsymbol{x} - \mathbb{E}_{\boldsymbol{x}}[\boldsymbol{x}]) \le Z_{\alpha}^{2} \},$$
(B10)

where  $Z_{\alpha}^2$  is the  $\alpha$ -quantile of the  $\chi_d^2$  distribution. Such estimates are important because they allow SMC methods to characterize the uncertainty in an estimate of the unknown Hamiltonian [15]. We do not emphasize the ability of our learning algorithm to perform region estimation in the main body of the text, but the algorithm's capability of specifying the uncertainty in the unknown Hamiltonian through the form of a region estimation provides a powerful advantage over tomographic methods wherein such a characterization of the uncertainty is much less natural.

## Appendix C: Solution in tractable cases

The mathematical tools we use to solve the Hamiltonian identification problem are those of decision theory and statistical learning. We have used a combination of methods from computation statistics to approximate the optimal solution to the problem of learning a Hamiltonian. However, for the case of a single unknown parameter, the equations can be solved analytically. These solutions provide much of the insight into designing the numerical algorithm to solve the general problem as well as serve to explain, in a broader context, why our method succeeds.

### 1. Statistical decision theory of learning

To evaluate the performance of any algorithm, we compare the estimated Hamiltonian parameters  $\hat{x}$  to the true parameters  $x_0$  by using the quadratic loss  $L(\hat{x}, x) = \|\hat{x} - x\|^2$ . This loss function generalizes the mean squared error to multiple parameters, and quantifies the error we incur due to the estimation procedure.

Our task is to choose an estimator  $\hat{x}(D)$ , a function from the possible data sets to valid parameters. This problem is most naturally cast in the language of decision theory. There is ostensibly one general approach: minimize—in some sense—the expected loss, or *risk*. That is we choose the estimator which satisfies

$$\hat{\boldsymbol{x}}_{\text{opt}} := \operatorname{argmin}_{\hat{\boldsymbol{x}}} \mathbb{E}_{\boldsymbol{x},D}[\|\boldsymbol{x} - \hat{\boldsymbol{x}}(D)\|^2], \tag{C1}$$

where the expectation is with respect to the distribution of x and D for the given experiment. This objective function is denoted r(e) for the experiment e, and called the *Bayes risk*. Under some regularity conditions, the unique best strategy is the Bayesian one, selecting as the estimator the mean of the posterior distribution

$$\hat{\boldsymbol{x}}_{\text{opt}} = \mathbb{E}_{\boldsymbol{x}|D}[x]. \tag{C2}$$

An important and useful consequence of using the quadratic loss is that the Bayes risk is equal to the expected trace of the covariance matrix of the posterior distribution (in the case of a single parameter it is simply the variance).

### 2. Single parameter problem

For the single parameter problem, the Hamiltonian reads

$$H(x) = x\sigma_z^{(1)}\sigma_z^{(2)},\tag{C3}$$

and the initial state is  $|+\rangle$  and final measurement is in the basis  $\{|+\rangle, |-\rangle\}$  (labeling the outcomes  $\{0,1\}$ ). If we evolve for a time t and allow an IQLE experiment with inversion Hamiltonian  $H_- := H(x_-)$ , the output probability distribution (the likelihood function) is

$$Pr(d|x; \mathbf{x}_{-}, t) = \frac{1}{2} (1 + (1 - 2d) \cos[2(x - \mathbf{x}_{-})t])$$
 (C4)

This model, for CLE experiments, was studied in detail in [13]. To obtain asymptotic expressions, we assume that probability distribution of x is approximately Gaussian and remains so after a subsequent measurement. The risk incurred between these two measurements then provides an asymptotic approximation to Bayes risk of the algorithm. Formally, we assume

$$\Pr(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{\sigma^2}\right),\tag{C5}$$

with mean  $\mu$  and variance  $\sigma^2$ . The posterior distribution

$$\Pr(x|d,\mu,\sigma;\boldsymbol{x}_{-},t) = \frac{\Pr(d|x,\mu,\sigma;\boldsymbol{x}_{-},t)\Pr(x|\mu,\sigma)}{\Pr(d|\mu,\sigma;\boldsymbol{x}_{-},t)}$$
(C6)

gives a mean of

$$\hat{x}_{\text{opt}} = \mu + \frac{2i(1 - 2d)\sigma^2 t(e^{4i\mu t} - e^{4i\mathbf{x}_- t})}{2e^{2t(i(\mu + \mathbf{x}_-) + \sigma^2 t)} + (1 - 2d)(e^{4i\mu t} - e^{4i\mathbf{x}_- t})},$$
(C7)

which is the final estimator. The risk incurred by this estimator (which, recall, is the optimal one) is explored next.

## 3. Asymptotic risk and the particle guess heuristic

These calculations rapidly become too cumbersome to display. The behavior of these more complex systems, however, can be described concisely with a few graphs, as shown in Figure 10.

Figure 10 shows the mean squared error for different choices of  $(t, \mathbf{x}_{-})$ . Notice the envelope

$$1 - 4\sigma^2 t^2 e^{-4\sigma^2 t^2} \le \frac{r(\mathbf{x}_-, t)}{\sigma^2} \le 1.$$
 (C8)

This tells us that the posterior variance cannot increase on average. In other words, there is no such thing as a strictly bad experiment. It also gives us a theoretical lower bound; the "risk envelope" has a minimum at

$$t_{\rm opt} = \frac{1}{2\sigma}. (C9)$$

This leads to

$$\min_{\mathbf{x}_{-},t} r(\mathbf{x}_{-},t) = (1 - e^{-1})\sigma^{2}.$$
 (C10)

That is, per measurement, the risk is reduced by a factor of about 0.63, which leads to the exponential scaling observed in [13].

Notice, however, in Figure 10 that the risk rapidly oscillates within the risk envelope. This shows, in particular, that the minimum corresponds to the solution of challenging global optimization problem. This is where we see the

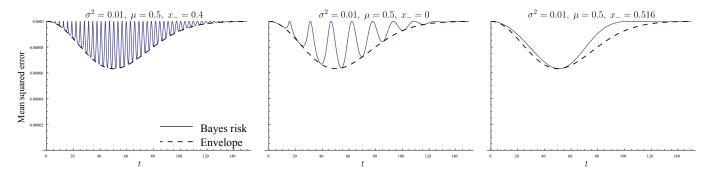


FIG. 10: The mean squared error of the optimal estimator as a function of evolution time. The initial distribution of x is normal  $\mathcal{N}(0.5, 0.01)$ . The left figure shows the mean squared error with no inversion. The remaining figures show the mean squared error with an inversion Hamiltonian. Note that the right figure uses  $x_{-} \approx \mu + \sigma$ . The black dotted line lower bounds the mean squared error and its minimum value is the ultimate limit on the performance of any strategy.

advantage of inversion; the effect of inversion is to "wash out" these oscillations. Thus, errors in approximations misplacing the optimal evolution time are less severe. Based on numerical testing, the optimal experimental inversion parameter is

$$x_{-} \approx \mu \pm \sigma.$$
 (C11)

These results and conclusions are only valid for the 1-dimensional parameter estimation problem. However, from these results we can gain intuition for what ought to happen in the multi-dimensional case. First, since the role of time is identical, we would expect that the optimal algorithm achieves exponential scaling, as we indeed observe. Second, we should expect that the optimal time for each experiment be proportional to some function of the inverse covariance matrix of the current distribution. Computing and inverting the covariance matrix is computationally inconvenient and so we use the heuristic

$$t = \frac{1}{\|\boldsymbol{x}' - \boldsymbol{x}\|},\tag{C12}$$

where  $x \neq x'$  are two particles drawn at random from the distribution of particles weights. This is a proxy for the inverse of the standard deviation. Finally, a computationally efficient analog of the experiment design in equation (C11) is to simply select  $x_{-}$  at random from the distribution of particle weights. As we will see next, this added randomization has a positive effect when additional errors are present.

One final note before we move on is that the above analysis assumes the distribution is approximately Gaussian. This will eventually be true but in practice we require a "warm-up" phase of experiment designs before we employ the the heuristics motivated by the asymptotic analysis. Fortunately, the randomization included in the particle guess heuristic provides a way to adaptively warm-up the learning algorithm without including an ad-hoc warmup heuristic, as was done in previous studies [15].

#### 4. Robustness of inversion to sampling error

For a two-outcome model, the only possible errors (regardless of origin—physical, modeling, sampling, etc.) manifest as bit-flips. If we assume the process is symmetric, we have a noisy version of the likelihood function,

$$\Pr(d|x; \boldsymbol{x}_{-}, t, \alpha) = \alpha + (1 - 2\alpha)\Pr(d|x; \boldsymbol{x}_{-}, t), \tag{C13}$$

where  $\alpha$  is the probability of a bit-flip. Now, since we assume the the algorithm is blind to this added noise, the posterior does not change. Thus, the estimator (the posterior mean) and the variance do not change either. If this seems odd, one must think of the posterior as a logical construct which is updated with assumed model—not the true model. To evaluate the Bayes risk however, we must take the average with respect to data of true model:

$$r(\boldsymbol{x}_{-},t,\alpha) = \mathbb{E}_{d|\boldsymbol{x}_{-},t,\alpha}[\operatorname{Var}_{x|d;\boldsymbol{x}_{-},t}(x)]. \tag{C14}$$

This quantity is shown in Figure 11 for various values of  $\alpha$ . The important thing to note is that the strategy with no inversion possesses a risk which can now *increase*. Now "bad" experiments are just as likely as good ones near

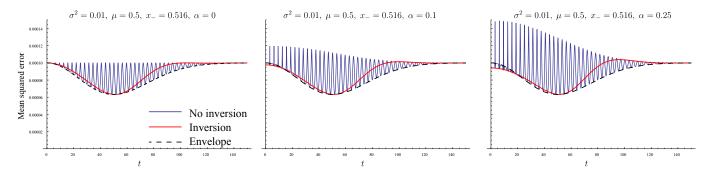


FIG. 11: The mean squared error of the optimal estimator as a function of evolution time. This is identical to the far right plot in Figure 10 except that we had added an additional bit-flip noise source of varying strength.

the optimal evolution time. Remarkably, the inversion model is complete insensitive to any strength of noise near the optimal evolution time. Moreover, the "particle guess heuristic" achieves the same performance independent of noise, which implies that the experimenter need not change their strategy depending on whether noise is present or not.

### 5. Consistency in multiple dimensions

The above analysis considered the case of a single unknown parameter. While this makes the statistical lessons learned equally valid when moving to more unknown parameters plausible, it would be more comforting to have similar results for more than a single parameter. Unfortunately, the integrals required appear to be analytically intractable. We can, however, perform simulations to obtain an approximate function form for the Bayes risk. To this end, we consider the 3-qubit problem:

$$H(x_1, x_2) = x_1 \,\sigma_z^{(1)} \sigma_z^{(2)} + x_2 \,\sigma_z^{(2)} \sigma_z^{(3)}. \tag{C15}$$

The results of the simulations, analogous to those presented in Figures 10 and 11, are shown in Figure 12. The conclusions drawn from the 2-qubit case remain; inversion enhances the performance of the estimation algorithm by smoothing out the Bayes risk and leaving the improvement unchanged near the optimal evolution time.

### Appendix D: Conditions for Asymptotic Stability of Bayesian Inference

We have already discussed the need for resampling as a means of maintaining the stability of performing Bayesian inference using the SMC approximation. Here we discuss why these instabilities arise, and whether there are other sources of instabilities that can arise in quantum Hamiltonian estimation. We show that the errors in the updating procedure will, on average, be small given that experiments are chosen that do not yield small likelihoods for probable events and given that the particle weights used in the SMC approximation do not become too small.

We consider, for now, only one step in the updating procedure. There are two sources of errors that can arise in the update procedure: (1) errors in the prior that have arisen due to previous approximate updates or numerical errors in the initial prior (2) errors in the likelihood evaluation. Let us assume that datum D is obtained and let the error–free prior probability of Hamiltonian  $H(x_j)$ , for any j, be denoted  $\Pr(x_j)$  and similarly the actual likelihood is  $\Pr(D|x_j)$ . We then denote the approximate analogs of these distributions  $\Pr(x_j)$  and  $\Pr(D|x_j)$ . The error in the posterior probability of  $x_j$  is

$$\epsilon_{j} = \left| \frac{\Pr(D|\boldsymbol{x}_{j}) \Pr(\boldsymbol{x}_{j})}{\sum_{j} \Pr(D|\boldsymbol{x}_{j}) \Pr(\boldsymbol{x}_{j})} - \frac{\tilde{\Pr}(D|\boldsymbol{x}_{j}) \tilde{\Pr}(\boldsymbol{x}_{j})}{\sum_{j} \tilde{\Pr}(D|\boldsymbol{x}_{j}) \tilde{\Pr}(\boldsymbol{x}_{j})} \right|$$
(D1)

For simplicity, we will now introduce variables that describe the variation of the approximate probabilities from the precise probabilities. These deviations can, in many circumstances, be thought of as random variables since the majority of the error in this protocol will arise from using sampling to estimate the likelihood function.

$$\tilde{\Pr}(\boldsymbol{x}_j) := \eta_j' + \Pr(\boldsymbol{x}_j) 
\tilde{\Pr}(D|\boldsymbol{x}_j) := \eta_j + \Pr(D|\boldsymbol{x}_j)$$
(D2)

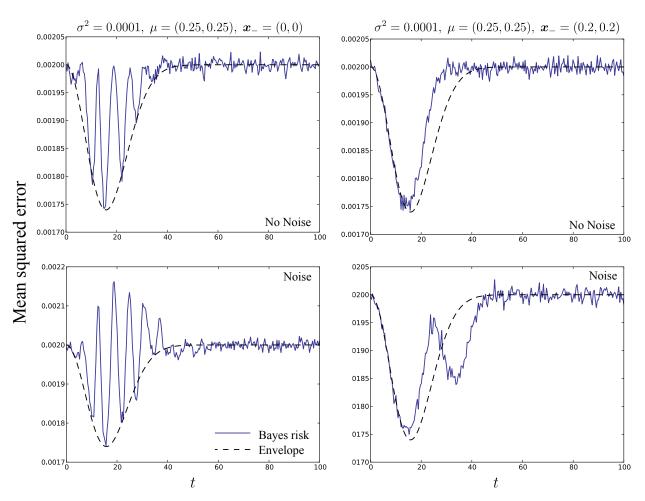


FIG. 12: The mean squared error of the optimal estimator as a function of evolution time. The initial distribution of  $(x_1, x_2)$  is chosen to be normally distributed with mean  $\mu = (0.25, 0.25)$  and diagonal covariance matrix with equal variances for both coordinates:  $\sigma^2 = 0.001$ . The top row displays the mean squared error when there is no noise with the left plot showing the case of no inversion and right showing the choice of an inversion Hamiltonian with parameters roughly a distance from the mean that is given by the square root of the trace of the covariance matrix  $(2\sigma)$  in this case. The lower two plots show the mean squared error for the same two inversion strategies when there is 10% noise present.

We make use of the fact that

$$\eta := \sum_{j} \Pr(D|\boldsymbol{x}_{j}) |\eta'_{j}| + \Pr(\boldsymbol{x}_{j}) |\eta_{j}| + |\eta_{j}\eta'_{j}| \le \frac{1}{2} \sum_{j} \Pr(D|\boldsymbol{x}_{j}) \Pr(\boldsymbol{x}_{j}). \tag{D3}$$

Then assuming that  $\max\{|\eta_i'|, |\eta_j|\} \in O(\eta)$  we find by using Taylor's theorem and the triangle inequality that

$$\epsilon_{j} = \left| \frac{[\Pr(D|\boldsymbol{x}_{j}) + \eta_{j}][\Pr(\boldsymbol{x}_{j}) + \eta'_{j}]}{\sum_{k} [\Pr(D|H(\boldsymbol{x}_{k})) + \eta_{k}][\Pr(H(\boldsymbol{x}_{k})) + \eta'_{k}]} - \frac{\Pr(D|\boldsymbol{x}_{j}) \Pr(\boldsymbol{x}_{j})}{\sum_{k} \Pr(D|H(\boldsymbol{x}_{k})) \Pr(H(\boldsymbol{x}_{k}))} \right| \\
\leq \left| (\Pr(\boldsymbol{x}_{j}) + \eta'_{j})(\Pr(D|\boldsymbol{x}_{j}) + \eta_{j}) \left( \frac{1}{\sum_{k} \Pr(D|H(\boldsymbol{x}_{k})) \Pr(H(\boldsymbol{x}_{k}))} + \frac{2\eta}{(\sum_{k} \Pr(D|H(\boldsymbol{x}_{k})) \Pr(H(\boldsymbol{x}_{k})))^{2}} \right) - \frac{\Pr(D|\boldsymbol{x}_{j}) \Pr(\boldsymbol{x}_{j})}{\sum_{k} \Pr(D|H(\boldsymbol{x}_{k})) \Pr(H(\boldsymbol{x}_{k}))} \right| \\
\leq \left( \frac{\Pr(D|\boldsymbol{x}_{j}) |\eta'_{j}| + \Pr(\boldsymbol{x}_{j}) |\eta_{j}|}{\sum_{k} \Pr(D|H(\boldsymbol{x}_{k})) \Pr(H(\boldsymbol{x}_{k}))} + \frac{2\eta \Pr(D|\boldsymbol{x}_{j}) \Pr(\boldsymbol{x}_{j})}{(\sum_{k} \Pr(D|H(\boldsymbol{x}_{k})))^{2}} \right) + O(\eta^{2}) \tag{D4}$$

The overall error as measured by the 1–norm is  $\epsilon = \sum_{j} \epsilon_{j}$  and hence (D4) gives

$$\epsilon \leq \frac{3\sum_{k} \Pr(D|H(\boldsymbol{x}_{k}))|\eta'_{k}| + \Pr(H(\boldsymbol{x}_{k}))|\eta_{k}|}{\sum_{k} \Pr(D|H(\boldsymbol{x}_{k})) \Pr(H(\boldsymbol{x}_{k}))} + O(\eta^{2})$$

$$\leq \frac{3\left(\sqrt{\sum_{k} \Pr^{2}(D|H(\boldsymbol{x}_{k}))}\sqrt{\sum_{k}|\eta'_{k}|^{2}} + \sqrt{\sum_{k} \Pr^{2}(H(\boldsymbol{x}_{k}))}\sqrt{\sum_{k}|\eta_{k}|^{2}}\right)}{\sum_{k} \Pr(D|H(\boldsymbol{x}_{k})) \Pr(H(\boldsymbol{x}_{k}))} + O(\eta^{2}). \tag{D5}$$

Equation (D5) provides an upper bound for the error in the Bayesian update for a fixed measured datum D. In practice, surprising outcomes can destabilize the update according to (D5). The contribution of such surprising results to the overall error is small if

$$\sqrt{\sum_{j} \eta_{j}^{\prime 2}} \ll \frac{\sum_{k} \Pr(D|H(\boldsymbol{x}_{k})) \Pr(H(\boldsymbol{x}_{k}))}{\sqrt{\sum_{k} \Pr^{2}(D|H(\boldsymbol{x}_{k}))}},$$
(D6)

$$\sqrt{\sum_{j} \eta_{j}^{2}} \ll \frac{\sum_{k} \Pr(D|H(\boldsymbol{x}_{k})) \Pr(H(\boldsymbol{x}_{k}))}{\sqrt{\sum_{k} \Pr^{2}(H(\boldsymbol{x}_{k}))}} = \sqrt{N_{\text{ess}}} \left(\sum_{k} \Pr(D|H(\boldsymbol{x}_{k})) \Pr(H(\boldsymbol{x}_{k}))\right), \tag{D7}$$

where  $N_{\rm ess}$  is the effective sample size.

These equations give two different criteria for the stability of the Bayesian update. Equation (D6) states that if the weights of the particles are too small then an unreasonably small value of  $\eta'_j$  may be required to ensure that the error in the update is small. This justifies the need for using resampling in SMC methods, and further justifies the criteria used for resampling in our algorithm:  $N_{\text{ess}} = (\sum_k \Pr^2(H(\boldsymbol{x}_k)))^{-1} \leq |\{\boldsymbol{x}_i\}|/2$ . Equation (D7) makes a more interesting claim. It states that the update rule can become unstable if the expectation value of  $\Pr(D|H(\boldsymbol{x}_k))$  over the prior  $\Pr(H(\boldsymbol{x}_k))$  is small for typical values of D.

Eqns. (D5) and (D7) imply that the error due to estimating the likelihood via sampling is asymptotically bounded above by  $\epsilon$  if

$$\sum_{j} \eta_{j}^{2} \in O\left(\epsilon^{2} N_{\text{ess}}\left(\sum_{k} \Pr(D|H(\boldsymbol{x}_{k})) \Pr(H(\boldsymbol{x}_{k}))\right)^{2}\right). \tag{D8}$$

Since  $\sum_{i} \eta_{i}^{2} \leq |\{x_{i}\}| \eta_{i \max}^{2}$ , where  $\eta_{i \max} = \max_{j} \eta_{j}$ , we have that (D8) is satisfied if

$$\eta_{j \max} \in O\left(\epsilon \sqrt{\frac{N_{\text{ess}}}{|\{\boldsymbol{x_i}\}|}} \left(\sum_{k} \Pr(D|H(\boldsymbol{x}_k)) \Pr(H(\boldsymbol{x}_k))\right)\right).$$
(D9)

Our criteria for resampling is that  $N_{\text{ess}} \leq |\{x_i\}|/2$ . So, we have that  $N_{\text{ess}} \in \Theta(|\{x_i\}|)$  and hence (D8) is implied by

$$\eta_{j \max} \in O\left(\epsilon \left(\sum_{k} \Pr(D|H(\boldsymbol{x}_{k})) \Pr(H(\boldsymbol{x}_{k}))\right)\right).$$
(D10)

We use as our estimate of  $\Pr(D|H(\boldsymbol{x}_k))$  the fraction of samples drawn from the simulator that yield outcome D. The resultant distribution for the number of samples that yield D is a binomial distribution with mean  $N_{\text{samp}} \Pr(D|H(\boldsymbol{x}_k))$  and variance  $N_{\text{samp}} \Pr(D|H(\boldsymbol{x}_k))(1 - \Pr(D|H(\boldsymbol{x}_k)))$ . Hence, if  $N_{\text{samp}}$  samples are drawn from the simulator then the uncertainty in our estimate of  $\Pr(D|H(\boldsymbol{x}_k))$  obeys

$$\eta_{j \max} \in O\left(\sqrt{\frac{\max_k \Pr(D|H(\boldsymbol{x}_k))(1 - \Pr(D|H(\boldsymbol{x}_k)))}{N_{\text{samp}}}}\right).$$
(D11)

Therefore we have from (D10) that (D8) is satisfied if

$$\sqrt{\frac{\max_{k} \Pr(D|H(\boldsymbol{x}_{k}))(1 - \Pr(D|H(\boldsymbol{x}_{k})))}{N_{\text{samp}}}} \in O\left(\epsilon \sum_{k} \Pr(D|H(\boldsymbol{x}_{k})) \Pr(H(\boldsymbol{x}_{k}))\right), \tag{D12}$$

which is equivalent to saying that

$$N_{\text{samp}} \in \Omega \left( \frac{\max_{k} \Pr(D|H(\boldsymbol{x}_{k}))(1 - \Pr(D|H(\boldsymbol{x}_{k})))}{\left(\epsilon \sum_{k} \Pr(D|H(\boldsymbol{x}_{k})) \Pr(H(\boldsymbol{x}_{k}))\right)^{2}} \right), \tag{D13}$$

We require that  $N_{\text{samp}}$  samples are drawn for each particle in  $\{x_i\}$  and hence it is sufficient to take a number of simulations that scales as

$$N_{\text{sim}} \in \Theta\left(\frac{|\{\boldsymbol{x}_i\}| \max_k \Pr(D|H(\boldsymbol{x}_k))(1 - \Pr(D|H(\boldsymbol{x}_k)))}{(\epsilon \sum_k \Pr(D|H(\boldsymbol{x}_k)) \Pr(H(\boldsymbol{x}_k)))^2}\right). \tag{D14}$$

Our method uses the mean of the posterior distribution as an estimator for the true Hamiltonian,  $H(\boldsymbol{x})$ , which means that more work is needed to determine how an error of  $\epsilon$  in the update procedure propagates to errors in the mean and the variance of the posterior distribution. Let  $\mu_H := \sum_i \Pr(H(\boldsymbol{x}_i)|D)H(\boldsymbol{x}_i)$  be the posterior mean and  $\tilde{\mu}_H := \sum_i \tilde{\Pr}(H(\boldsymbol{x}_i)|D)H(\boldsymbol{x}_i)$  be the posterior mean calculated by approximate likelihood evaluation. The error in the estimated Hamiltonian, as measured by the 2–norm is then

$$\|\mu_H - \tilde{\mu}_H\| \le \max_i \|H(\boldsymbol{x}_i)\| \sum_i |\Pr(H(\boldsymbol{x}_i)|D) - \Pr(H(\boldsymbol{x}_i)|D)| = \max_i \|H(\boldsymbol{x}_i)\|\epsilon.$$
 (D15)

Similarly, it is straight forward to see that

$$\left| \sum_{i} \Pr(H(\boldsymbol{x}_i)|D) \|H(\boldsymbol{x}_i) - \mu_H\|^2 - \tilde{\Pr}(H(\boldsymbol{x}_i)|D) \|H(\boldsymbol{x}_i) - \tilde{\mu}_H\|^2 \right| \in O(\max_{i} \|H(\boldsymbol{x}_i)\|^2 \epsilon), \tag{D16}$$

where  $\sum_{i} \Pr(H(\boldsymbol{x}_i)|D) \|H(\boldsymbol{x}_i) - \mu_H\|^2$  is the posterior variance.

It may be tempting to conclude that after N steps, the error in the estimate is  $N \max_i ||H(x_i)|| \epsilon$ , but because the Bayesian update rule is non-linear it is difficult to prove such a bound. Instead, note that Bayesian inference is robust to the choice of prior [15] and thus the inference process will remain stable under such errors. We therefore can consider beginning the inference process using the erroneous posterior as the prior and expect convergence if the relative errors in the variance are small. In particular, we expect stability if

$$\max_{i} ||H(\boldsymbol{x}_i)||^2 \epsilon \in O(\delta), \tag{D17}$$

where  $\delta \leq Ae^{-\gamma N}$  is defined to be the error in the estimate of the unknown Hamiltonian and  $\gamma$  is approximately a constant function in N for the test cases considered in the main body of the paper. We therefore expect the algorithm to be stable if  $\epsilon$  is chosen as above and hence, it will suffice to use a number of simulations in an update that approximately scales as

$$N_{\text{sim}} \in \Theta\left(\frac{\max_{i} \|H(\boldsymbol{x}_{i})\|^{4} |\{\boldsymbol{x}_{i}\}|}{\delta^{2}} \frac{\max_{k} \Pr(D|H(\boldsymbol{x}_{k}))(1 - \Pr(D|H(\boldsymbol{x}_{k})))}{\left(\sum_{k} \Pr(D|H(\boldsymbol{x}_{k})) \Pr(H(\boldsymbol{x}_{k}))\right)^{2}}\right). \tag{D18}$$

It is then easy to see from Markov's inequality that with high probability over the experiments the cost of any given update will be at most a constant multiple of the cost of the expectation value over all prior distributions  $Pr(x_i)$  that appear in the learning process and all outcomes D observed. Therefore, our approximation to the total number of simulations required to learn the parameters within loss  $\delta$  scales, with high probability, as

$$N_{\text{total}} \in \Theta\left(\frac{N \max_{i} \|H(\boldsymbol{x}_{i})\|^{4} |\{\boldsymbol{x}_{i}\}|}{\delta^{2}} \mathbb{E}\left(\frac{\max_{k} \Pr(D|H(\boldsymbol{x}_{k}))(1 - \Pr(D|H(\boldsymbol{x}_{k})))}{\left(\sum_{k} \Pr(D|H(\boldsymbol{x}_{k})) \Pr(H(\boldsymbol{x}_{k}))\right)^{2}}\right)\right)$$

$$\in \Theta\left(\frac{\log(1/\delta) \max_{i} \|H(\boldsymbol{x}_{i})\|^{4} |\{\boldsymbol{x}_{i}\}|}{\gamma \delta^{2}} \mathbb{E}\left(\frac{\max_{k} \Pr(D|H(\boldsymbol{x}_{k}))(1 - \Pr(D|H(\boldsymbol{x}_{k})))}{\left(\sum_{k} \Pr(D|H(\boldsymbol{x}_{k})) \Pr(H(\boldsymbol{x}_{k}))\right)^{2}}\right)\right). \tag{D19}$$

This suggests that QLE and IQLE may be efficient, given that  $\gamma \in \Omega(\text{poly}(1/n))$ ,  $\max_j \|H_j\| \in O(\text{poly}(n))$  and experiments that yield, with high probability,  $\Pr(D|H(\boldsymbol{x}_k)) \in O(1/\text{poly}(n))$  are avoided. We observed that these scalings are obeyed for the examples considered in the main body if IQLE and the particle guess heuristic are employed. More complex examples may require local optimization of the guesses in order to avoid multi-modal prior distributions, which can be problematic for the PGH; however, we saw no benefit to local optimization for the Ising models considered previously. Finally, the scaling predicted for  $N_{\text{total}}$  as a function of  $\delta$  does not seem to be tight in the prior examples since  $N_{\text{total}}$  does not appear to strongly depend on  $\delta$  in those examples. A more careful analysis of the uncertainty is likely to reveal that our conditions for stability are unnecessarily pessimistic.