Bayesian learning of parameterised quantum circuits

View the article online for updates and enhancements.

MACHINE LEARNING
Science and Technology

PAPER

CrossMark

Bayesian learning of parameterised quantum circuits

Samuel Duffield[*] ⓘ, Marcello Benedetti ⓘ and Matthias Rosenkranz[*] ⓘ

Quantinuum, Partnership House, Carlisle Place, London SW1P 1BX, United Kingdom
* Authors to whom any correspondence should be addressed.

E-mail: sam.duffield@quantinuum.com and matthias.rosenkranz@quantinuum.com

Abstract

Currently available quantum computers suffer from constraints including hardware noise and a limited number of qubits. As such, variational quantum algorithms that utilise a classical optimiser in order to train a parameterised quantum circuit have drawn significant attention for near-term practical applications of quantum technology. In this work, we take a probabilistic point of view and reformulate the classical optimisation as an approximation of a Bayesian posterior. The posterior is induced by combining the cost function to be minimised with a prior distribution over the parameters of the quantum circuit. We describe a dimension reduction strategy based on a maximum a posteriori point estimate with a Laplace prior. Experiments on the Quantinuum H1-2 computer show that the resulting circuits are faster to execute and less noisy than the circuits trained without the dimension reduction strategy. We subsequently describe a posterior sampling strategy based on stochastic gradient Langevin dynamics. Numerical simulations on three different problems show that the strategy is capable of generating samples from the full posterior and avoiding local optima.

1. Introduction

Variational quantum algorithms (VQAs) [1–4] are the leading paradigm for solving computational problems on current generation quantum computers. A VQA solves the computational problem by turning it into an optimisation problem over the parameters of a quantum circuit. The quantum computer is used to execute the circuit, that is, to prepare a quantum state and perform measurements on it. The classical computer is used to estimate the cost function from measurement outcomes and to update the parameters accordingly. This process is repeated in the hope of finding the parameters yielding minimum cost, effectively encoding a solution to the computational problem.

A number of VQAs have been proposed to attack specific problems in condensed matter physics, quantum chemistry, machine learning and combinatorial optimisation, with demonstrations on existing hardware [1–4]. One of the motivations is that the quantum circuit ansatz can be designed to comply with hardware constraints (e.g. qubit-to-qubit connectivity and coherence time) and to encode domain knowledge about the problem (e.g. symmetries and correlations). This is in contrast to fault-tolerant quantum algorithms which use a large number of error-corrected qubits and deep circuits to solve generic instances of a problem.

Despite the successes of VQAs, it is well known that the optimisation of parameterised quantum circuits (PQCs) does not scale to large systems in general. The optimisation landscape is characterised by features such as barren plateaus [5], narrow gorges [6], and exponentially many local minima [7], most of which have poor quality [8]. Real experiments are further complicated by the fact that hardware is noisy, and execution of quantum gates is slow in some architectures. Any improvement in parameter initialisation (e.g. [9–11]), ansatz design (e.g. [12–14]) and training (e.g. [15–17]) could push the boundaries of VQA applications.

In the context of VQAs there exist a plethora of Bayesian methods which we briefly review here. Bayesian optimisation is a zeroth-order method (i.e. it does not use gradient information) which is popular among VQA practitioners [18–20]. Bayesian optimisation can, however, be used within first-order methods to tune the stepsize [21] or initialisation [22]. Wang *et al* [23] uses Bayesian methods to infer the value of the cost
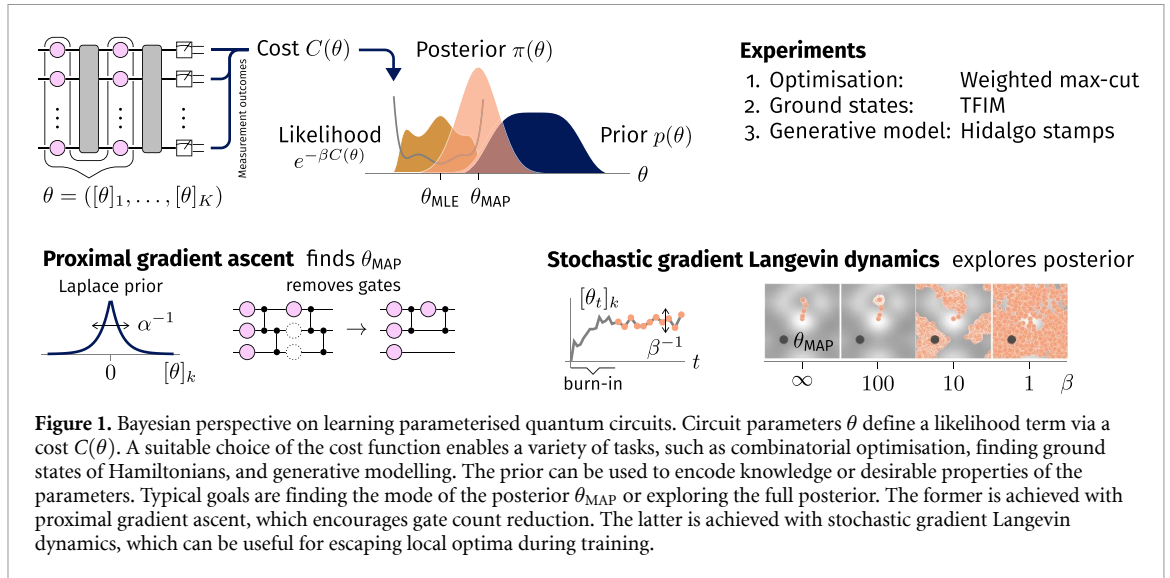
Figure 1. Bayesian perspective on learning parameterised quantum circuits. Circuit parameters θ define a likelihood term via a cost $C(\theta)$. A suitable choice of the cost function enables a variety of tasks, such as combinatorial optimisation, finding ground states of Hamiltonians, and generative modelling. The prior can be used to encode knowledge or desirable properties of the parameters. Typical goals are finding the mode of the posterior θ_{MAP} or exploring the full posterior. The former is achieved with proximal gradient ascent, which encourages gate count reduction. The latter is achieved with stochastic gradient Langevin dynamics, which can be useful for escaping local optima during training.

function from a reduced number of measurements. Benedetti *et al* [24] proposes a VQA for inference of unobserved variables in Bayesian networks. Du *et al* [25] uses entanglement and ancillary qubits to implement a quantum prior distribution over the circuit parameters. Going beyond VQAs we also find a number of quantum algorithms for speeding up Bayesian inference [26, 27] as well as novel Bayesian quantum causal models [28–30].

In this work, we formulate generic VQAs as a Bayesian inference problem over circuit parameters. We propose two algorithms that achieve different goals. One searches for the maximum a posteriori point estimate and automatically removes a given percentage of parameters (quantum gates) to reduce hardware noise and execution speed. The second approximately samples from the posterior distribution over the circuit parameters and reduces sensitivity to initialisation and local optima. Our methods make good use of the gradient of the cost function, which can be estimated from additional measurements on the quantum computer. An overview of the framework and methods is visualised in figure 1.

This article is structured as follows. In section 2 we introduce the Bayesian perspective for VQAs. In section 3 we describe the maximum a posteriori approach for a sparsity-inducing Laplace prior. In section 4 we describe stochastic gradient Langevin dynamics (SGLD) for posterior sampling. In section 5 we numerically investigate the algorithms on instances of weighted max-cut, a transverse field Ising model, and a generative modelling problem. We also show the benefits of using the Laplace prior with an experiment on the Quantinuum H1-2 computer. In section 6 we present our concluding remarks and discuss avenues for future research.

2. A Bayesian perspective

A PQC takes the form $U(\theta) = \prod_{k=1}^{K} W_k U_k([\theta]_k)$, where $\{W_k\}_{k=1}^{K}$ is a set of fixed quantum gates, and $\{U_k([\theta]_k)\}_{k=1}^{K}$ is a set of parameterised gates. The circuit is applied to some initial quantum state. Let $C(\theta)$ be the cost for the problem at hand as a function of the circuit parameters $\theta = ([\theta]_1, \ldots [\theta]_K)$. For example, the cost function could be the expectation value $C(\theta) = \mathrm{Tr}[OU(\theta)|\psi_0\rangle\langle\psi_0|U(\theta)^{\dagger}]$ of an observable O and initial state $|\psi_0\rangle$. In VQAs a classical optimiser is used in the hope of finding

$$\theta_{\mathrm{MLE}} = \mathrm{argmin}_{\theta} C(\theta). \tag{1}$$

Here we use the subscript to indicate that this quantity is akin to a *maximum likelihood* estimator (MLE), as will become apparent. Assuming $C(\theta)$ is differentiable, a single iteration of vanilla (stochastic) gradient descent takes the form

$$\theta_t = \theta_{t-1} - \epsilon_t \widehat{\nabla_\theta C}(\theta_{t-1}), \tag{2}$$

where $\{\epsilon_t\}_{t=1}^{\infty}$ is a schedule of stepsizes and $\widehat{\nabla_\theta C}(\theta)$ is an unbiased estimate of the gradient of the cost function $\nabla_\theta C(\theta)$ [31, 32].

Now let us consider a probabilistic formulation

$$\pi(\theta) = \frac{p(\theta)\exp(-\beta C(\theta))}{\int p(\theta')\exp(-\beta C(\theta'))\mathrm{d}\theta'}, \tag{3}$$

and generalised optimisation problem

$$\theta_{\mathrm{MAP}} = \mathrm{argmax}_\theta \pi(\theta). \tag{4}$$

This probabilistic viewpoint instead treats the parameter vector θ as a random variable, and the denominator in equation (3) ensures $\pi(\theta)$ is a valid probability distribution, i.e. $\int \pi(\theta)\mathrm{d}\theta = 1$. In this light, we have the following *Bayesian* interpretation:

- $p(\theta)$ represents the *prior distribution* and can be used to encode any pre-experimental knowledge or desirable properties for the parameters θ;
- $\exp(-\beta C(\theta))$ is a generalised *likelihood* term [33, 34] encouraging $C(\theta)$ to be small. The parameter β controls the scaling of the cost function with respect to the prior. For statistical and machine learning tasks, the influence of the data enters through the likelihood term (i.e. $C(\theta) = C(\theta; y)$ for a dataset y);
- $\pi(\theta)$ represents the *posterior distribution* and describes high probability regions of the parameter space with uncertainty quantification.

The maximisation of $\pi(\theta)$ is equivalent to maximisation of $\log \pi(\theta)$ and thus we can take gradients in log-space

$$\nabla_\theta \log \pi(\theta) = \nabla_\theta \log p(\theta) - \beta \nabla_\theta C(\theta). \tag{5}$$

We now observe that when the prior is set to the uniform distribution $p(\theta) \propto 1$, we get $\nabla_\theta \log p(\theta) = 0$ and $\theta_{\mathrm{MAP}} = \theta_{\mathrm{MLE}}$ from equation (1). Subsequent application of a gradient ascent algorithm regains equation (2) (where the β parameter is absorbed by the stepsize). Observe that the paradigm shift (going from minimising a cost function to maximising a posterior distribution) results in a change in terminology, gradient descent to gradient ascent—although in the case of a uniform prior the implementation is identical. This gradient ascent approach searches for a so-called *maximum a posteriori* (MAP) estimator, as denoted in equation (4) and described in generality in algorithm 1 (note the stepsize has been rescaled such that $\beta \to \infty$ regains the maximum likelihood approach, equation (2)).

Algorithm 1. Gradient Ascent.

---

for $t = 1, \ldots$ **do**
$\quad \theta_t = \theta_{t-1} + \epsilon_t \beta^{-1} \nabla_\theta \log p(\theta_{t-1}) - \epsilon_t \widehat{\nabla_\theta C}(\theta_{t-1})$

A major success of the Bayesian paradigm is the ability to analyse uncertainty in the parameter θ, that is to quantify the full posterior $\pi(\theta)$. Unfortunately, aside from trivial cases, the true posterior is intractable and we have to resort finding an approximation $q(\theta) \approx \pi(\theta)$. There then comes a trade-off between quality of approximation and computational cost. The cheapest approximation is that of the already discussed point estimate $q(\theta) = \delta(\theta \mid \theta_{\mathrm{MAP}})$ where δ is the Dirac point measure, however this approach neglects all uncertainty in the parameter θ. A more rigorous approach to approximating the posterior is to construct a Monte Carlo approximation $q(\theta) = \frac{1}{T} \sum_{t=1}^{T} \delta(\theta \mid \theta_t)$. The most popular methods for constructing this Monte Carlo approximation [35, 36] do so in a way that is asymptotically unbiased for the posterior, i.e. $q(\theta) \xrightarrow{T \to \infty} \pi(\theta)$. Naturally, taking $T \to \infty$ is not feasible in practice and instead finite sample sizes are used, a (controllable) bias is therefore induced. A final approach to approximating the posterior is that of variational inference [37], where a parameterised *variational family* of distributions $Q = \{q_\omega(\theta) : \omega \in \Omega\}$ is defined and then the optimal parameters ω^* are sought in order to minimise some tractable measure of the discrepancy between $q_\omega(\theta)$ and $\pi(\theta)$ (most commonly the KL divergence). Variational inference is typically computationally cheaper than the Monte Carlo approach, although induces a bias in the likely case $\pi(\theta) \notin Q$ and this bias can be difficult to assess or control.

The Bayesian paradigm also provides a natural approach to characterising predictions, that is through expectations with respect to the posterior distribution $\mathbb{E}_{\pi(\theta)}[f(\theta)] \approx \mathbb{E}_{q(\theta)}[f(\theta)]$, where $f(\theta)$ is some predictive function. These predictions are trivial to implement in the case of the point estimate $\mathbb{E}_{q(\theta)}[f(\theta)] = f(\theta_{\mathrm{MAP}})$ and Monte Carlo $\mathbb{E}_{q(\theta)}[f(\theta)] = \frac{1}{T} \sum_{t=1}^{T} f(\theta_t)$ approximations. In the case of variational inference, the variational family is often chosen such that $\mathbb{E}_{q_\omega(\theta)}[f(\theta)]$ is analytically tractable for the predictive functions of interest.

3. Laplace prior and proximal gradient ascent

A simple and relevant choice of prior is that of the Laplace distribution (independent across parameters)

$$p(\theta) \propto \exp\left(-\alpha \sum_{k=1}^{K} |[\theta]_k|\right),$$

for $\alpha \in [0, \infty)$. This choice of prior is also known as LASSO or ℓ_1 regularisation [38]. For α large enough it is known that the resulting MAP estimate enforces $[\theta]_k = 0$ for the least influential parameters [38, 39]. In typical PQCs, the parameter $[\theta]_k$ represents an angle of a rotational gate and therefore setting $[\theta]_k = 0$ is equivalent to removing the corresponding gate. Removal of parameterised gates may lead to further gate reductions in a compilation step, e.g. if the removed gate was sandwiched between 2-qubit gates that now evaluate to the identity (see inset 'Proximal gradient ascent' in figure 1).

The non-differentiability of the Laplace prior also means we cannot apply standard gradient techniques. Specifically, algorithm 1 fails to enforce the parameters to be exactly zero. Fortunately we can utilise well-studied proximal gradient methods [39]. A single step of proximal gradient ascent takes the form

$$\theta_t = \text{prox}^{\varphi}_{\alpha \epsilon_t} (\theta_{t-1} - \epsilon_t \nabla_\theta C(\theta_{t-1})),$$

where the proximal operator is defined as $[\text{prox}^{\varphi}_{\upsilon}(x)]_k = \text{argmin}_y \left\{ \varphi(y) + \frac{1}{2\upsilon}\|[x]_k - y\|^2 \right\}$. In general the proximal operator is intractable, however the special case of the Laplace prior $\varphi(y) = \ell_1(y) := |y|$ can be solved analytically giving the *soft-thresholding* function [40]

$$[\text{prox}^{\ell_1}_{\upsilon}(\theta)]_k = \begin{cases} [\theta]_k - \upsilon, & [\theta]_k > \upsilon, \\ 0, & -\upsilon \leqslant [\theta]_k \leqslant \upsilon, \\ [\theta]_k + \upsilon, & [\theta]_k < -\upsilon. \end{cases} \tag{6}$$

Proximal gradient ascent has the benefit of explicitly setting parameters $[\theta]_k = 0$ when they are sufficiently small, it is also known to converge to a (local) MAP estimate of $\pi(\theta) \propto \exp(-\alpha \sum_{k=1}^{K} \varphi([\theta]_k) - \beta C(\theta))$ and at a faster rate than vanilla gradient ascent [39] (which is regained, for the MLE, by setting $\alpha = 0$). We note that the influence of the β parameter is absorbed by rescaling the regularisation parameter α and stepsize ε_t and can therefore be omitted, as described in algorithm 2, where we also allow the parameter $\alpha = \alpha_t$ to adapt over iterations.

Algorithm 2. Proximal Gradient Ascent.

---

for $t = 1, \dots$ **do**
   $\theta_t = \text{prox}^{\varphi}_{\alpha_t \epsilon_t}(\theta_{t-1} - \epsilon_t \widehat{\nabla_\theta C}(\theta_{t-1}))$
     for a Laplace prior, $\text{prox}^{\varphi}_{\upsilon}(\theta) = \text{prox}^{\ell_1}_{\upsilon}(\theta)$ in equation (6)

There are several potential benefits of the Laplace prior and the subsequent reduction in the number of gates. Specifically, the resulting VQA benefits from

- Reduced hardware noise;
- Circuits that are faster to sample.

The second point is also applicable to classical neural networks, where *weight pruning* [41] is used to reduce time and memory costs at test time.

There may also be trainability benefits in some cases (perhaps mitigating the barren-plateau phenomena [5] or simpler optimisation via dimension reduction) or better generalisation (by avoiding overfitting) for machine learning tasks, although this will be very case dependent as is investigated with mixed results in [42] (where they use *weight decay* which is equivalent to a Laplace prior without the explicit removal of gates achieved by proximal gradient ascent).

The downside is that by maximising $\pi(\theta)$ rather than minimising $C(\theta)$ directly, we have $C(\theta_{\text{MAP}}) > C(\theta_{\text{MLE}})$ when the prior is not uniform. We theoretically find a worse solution, although if the circuit is deep and α is small, this difference in cost may be negligible.

Additionally, the proximal operator is not tractable in general (the Laplace distribution is a particular instance where it is [43]). This makes inference difficult for alternative priors such as the spike-and-slab [44] or horseshoe [45] which theoretically achieve dimensionality reduction with a less significant shift in the global optima.

In practice, it is difficult to set the regularisation parameter α. It is more intuitive to set a fixed percentage of the parameters $\theta = ([\theta]_1, \dots [\theta]_K)$ to be zero and have the algorithm adapt α (or rather α_t) accordingly. We can achieve this at each iteration of proximal gradient ascent by choosing

$$\alpha_t \quad \text{such that} \quad \sum_{k=1}^{K} \mathbb{I}\left(\left|\left[\theta_{t-\frac{1}{2}}\right]_k\right| < \alpha_t \epsilon_t\right) = K_0,$$

where $\theta_{t-\frac{1}{2}} = \theta_{t-1} - \epsilon_t \widehat{\nabla_\theta C}(\theta_{t-1})$ and $K_0 \in \{1, \dots, K\}$ is the desired number of parameters to be set to 0 and therefore removed from the circuit. Numerically, this can be done efficiently by setting α_t to be the $\frac{K_0}{K}$th quantile of $\{|[\theta_{t-\frac{1}{2}}]_k|/\epsilon_t\}_{k=1}^{K}$ at each iteration.

4. SGLD

Reducing the entire posterior $\pi(\theta)$ to a single point estimate θ_{MAP} will be a poor description of the true behaviour of the parameter θ unless the posterior is very concentrated. Furthermore, the gradient ascent algorithm may only succeed in finding a local maximum. A more desirable inference procedure characterises the entire distribution $\pi(\theta)$.

The most popular classical approaches to posterior quantification build a Monte Carlo approximation normally either through Metropolis-Hastings based Markov chain Monte Carlo [35] or importance sampling [36]. Unfortunately both of these techniques require access to pointwise evaluations of $\pi(\theta)$ or at least, an unbiased estimate [46]. In our setting, we only have unbiased estimates of $\log \pi(\theta)$ and $\nabla_\theta \log \pi(\theta)$. Fortunately, we can adopt the SGLD method of [47] to generate an asymptotically unbiased Monte Carlo approximation whilst staying entirely in log-space.

Langevin dynamics are described by the following stochastic differential equation

$$\mathrm{d}\theta_t = \nabla_\theta \log \pi(\theta_t)\mathrm{d}t + \sqrt{2}W_t,$$

where W_t is a standard Brownian motion. Langevin dynamics are known to admit $\pi(\theta)$ as a stationary distribution [48]. That is, if we take a sample $\theta_t \sim \pi(\theta)$ and evolve it exactly according to Langevin dynamics (for any time period ε) then the marginal distribution of $\theta_{t+\epsilon}$ will also be $\pi(\theta)$. Therefore, simulating Langevin dynamics exactly and collecting samples along the way will provide a Monte Carlo approximation of the distribution $\pi(\theta)$.

For non-trivial distributions Langevin dynamics cannot be simulated exactly. Instead, an Euler-Maruyama discretisation is commonly applied

$$\theta_t = \theta_{t-1} + \epsilon_t \nabla_\theta \log \pi(\theta_{t-1}) + \sqrt{2\epsilon_t}\xi_t, \qquad \xi_t \sim \mathcal{N}(\xi \mid 0, \mathbb{I}). \tag{7}$$

This discretisation will introduce a bias for practical stepsizes $\epsilon_t > 0$. However, as argued in [47] (and proved in [49]), as long as the stepsize schedule is chosen to decay to zero $\sum_{t=1}^{\infty} \epsilon_t^2 < \infty$ but not too fast $\sum_{t=1}^{\infty} \epsilon_t = \infty$, then the samples will be asymptotically correct for $\pi(\theta)$. It was also noted that this is the case if $\nabla_\theta \log \pi(\theta)$ is replaced with an unbiased estimate, thus obtaining *stochastic gradient* Langevin dynamics. As a result, we can use our unbiased gradient estimate (or even a mini-batched version if applicable) within an SGLD algorithm to obtain a Monte Carlo approximation to the posterior $\pi(\theta)$. The algorithm, described in algorithm 3, represents a modification of gradient ascent with the correct amount of noise added to ensure exploration.

Algorithm 3. Stochastic Gradient Langevin Dynamics.

---

 for $t = 1, \dots$ **do**
  $\xi_t \sim \mathcal{N}(\xi \mid 0, \mathbb{I})$
  $\theta_t = \theta_{t-1} + \epsilon_t \beta^{-1} \nabla_\theta \log p(\theta_{t-1}) - \epsilon_t \widehat{\nabla_\theta C}(\theta_{t-1})$
         $+ \sqrt{2\beta^{-1}\epsilon_t}\xi_t$

There are multiple potential benefits to adding noise to gradient steps in this principled manner. Specifically, the resulting VQA benefits from

- Transitions between local optima and saddle points, for a suitably chosen β;
- Reduced sensitivity to initialisation;

- By replacing point estimate predictions with ergodic averages over the full trajectory (i.e. $\frac{1}{T}\sum_{t=1}^{T}f(\theta_t)$ for a prediction function $f(\theta)$), we quantify uncertainty in the parameter θ. This is both conceptually advantageous and can provide enhanced generalisation and a reduced tendency to overfit for machine learning problems [50].

We also note that, unlike the proximal gradient approach, SGLD is extremely flexible to prior specification.

On the flip side, SGLD brings an additional parameter to tune, β. Setting β too low will result in noisy trajectories that do not successfully find low cost regions of the parameter space. Conversely $\beta \to \infty$ regains vanilla gradient ascent with uniform prior, missing out on the benefits described above. Additionally, quantifying the posterior (via ergodic averages) is a significantly more challenging computational task than a simple point estimate and therefore more iterations may be required.

5. Experiments

We now investigate the benefits, drawbacks and parameter sensitivities of the two aforementioned generalisations of gradient ascent in a selection of VQA experiments. Firstly, we examine an 11 qubit weighted max-cut problem; secondly, we study the problem of sampling the ground state of an 11 qubit transverse-field Ising model (TFIM); before finally exploring the statistical task of using an 8 qubit PQC as a generative model (a so-called quantum *Born machine* [51]) for a real life, integer data set.

One of the simplest circuit parameterisations is via single-qubit gates $U_k([\theta]_k) = e^{-i[\theta]_k V_k/2}$. These are rotations through angles $[\theta]_k \in [0, 2\pi]$, generated by Hermitian operators V_k with eigenvalues ± 1. When the cost function $C(\theta)$ can be expressed as the quantum expectation of an Hermitian observable, the partial derivatives can be evaluated from parameter-shifted circuits (e.g. [52])

$$[\nabla_\theta C(\theta)]_k = \frac{1}{2}\left(C(\theta + \tfrac{\pi}{2}e_k) - C(\theta - \tfrac{\pi}{2}e_k)\right),$$

where e_k is the unit vector in the kth direction.

On a quantum device both $C(\theta)$ and each $[\nabla_\theta C(\theta)]_k$ can be approximated (unbiasedly) by generating n_{shots} samples from the PQC for each expectation, however for our numerical experiments we make use of qujax [53] for exact cost and gradient evaluations (i.e. $n_{\text{shots}} = \infty$).

All simulations are repeated 20 times with new initial parameters sampled from a small perturbation around zero, $[\theta_0]_k \sim \mathcal{U}([\theta]_k \mid -r, r)$ independently for $k = 1, \ldots, K$, with $r = 10^{-3}$. For each experiment we use a decaying stepsize schedule $\epsilon_t = a(t + b)^{-\frac{1}{3}}$ in line with [49], and set $a = 15$, $b = 10$.

5.1. Weighted max-cut

A well-known NP-complete optimisation problem is weighted max-cut. This is the task of taking a graph of nodes and weighted edges then solving for the optimal binary labelling of the nodes. Say we label each node as either '0' or '1' with the first node fixed to be '0', then the optimality of the labelling is defined as maximising the sum of weights on edges between nodes with *differing* labels. The cost function which we look to minimise is therefore defined as

$$C(\theta) = \mathbb{E}_{p(z|\theta)}[-S(z)],$$

where $z \in \{0, 1\}^N$ is a bit string labelling the $N + 1$ nodes (the zeroth node is labelled 0 by default), and $S(z)$ is the sum of weights between nodes with differing labels after labelling the graph according to the bit string z (the jth element of z indicates the label for the jth node).

We map bit strings to measurement operators in the computational basis of N qubits as $z \to |z\rangle\langle z|$. Then, the probability distribution is given by the Born rule $p(z \mid \theta) = |\langle z|U(\theta)|0\rangle^{\otimes N}|^2$. This arises from the inherent randomness of the pure quantum state and depends on the parameter values θ. For this experiment, we use the PQC in figure 2 with $N = 11$. The experiments are repeated across the 20 random seeds where for each seed a 3-regular graph (each node has three connected edges) is randomly generated along with associated weights each sampled uniformly in [0,1].

The top row of figure 3, displays training with a Laplace prior via proximal gradient ascent on shallow and deep circuits as well as training with a uniform prior via SGLD on a shallow circuit.

We vary the regularisation strength of the Laplace prior by changing the number of parameters set to 0 (and therefore removing gates from the circuit) at each iteration within the adaptive proximal gradient ascent, section 3. For the shallower circuit ($L = 1$ layers) we see that the Laplace prior is having a strong impact, forcing the trained cost to be higher. This is somewhat remedied by using a deeper circuit ($L = 7$ layers).
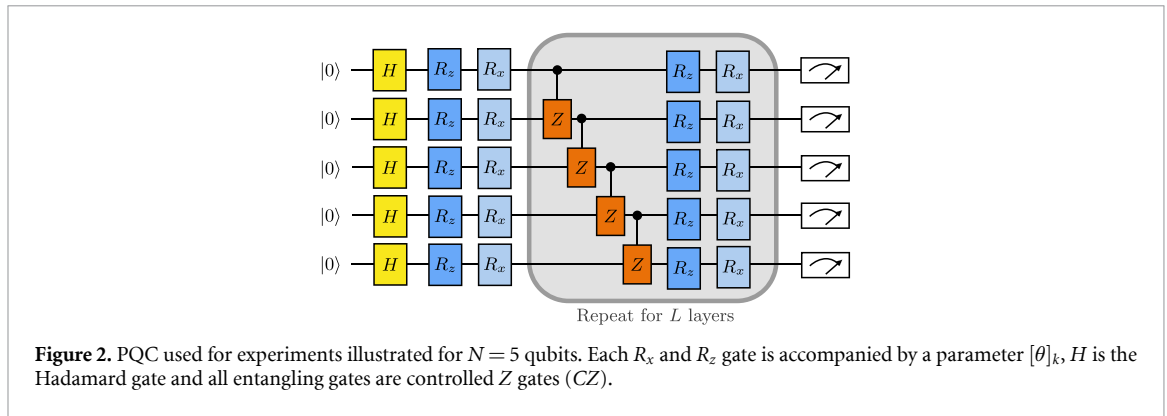
Figure 2. PQC used for experiments illustrated for $N = 5$ qubits. Each R_x and R_z gate is accompanied by a parameter $[\theta]_k$, H is the Hadamard gate and all entangling gates are controlled Z gates (CZ).
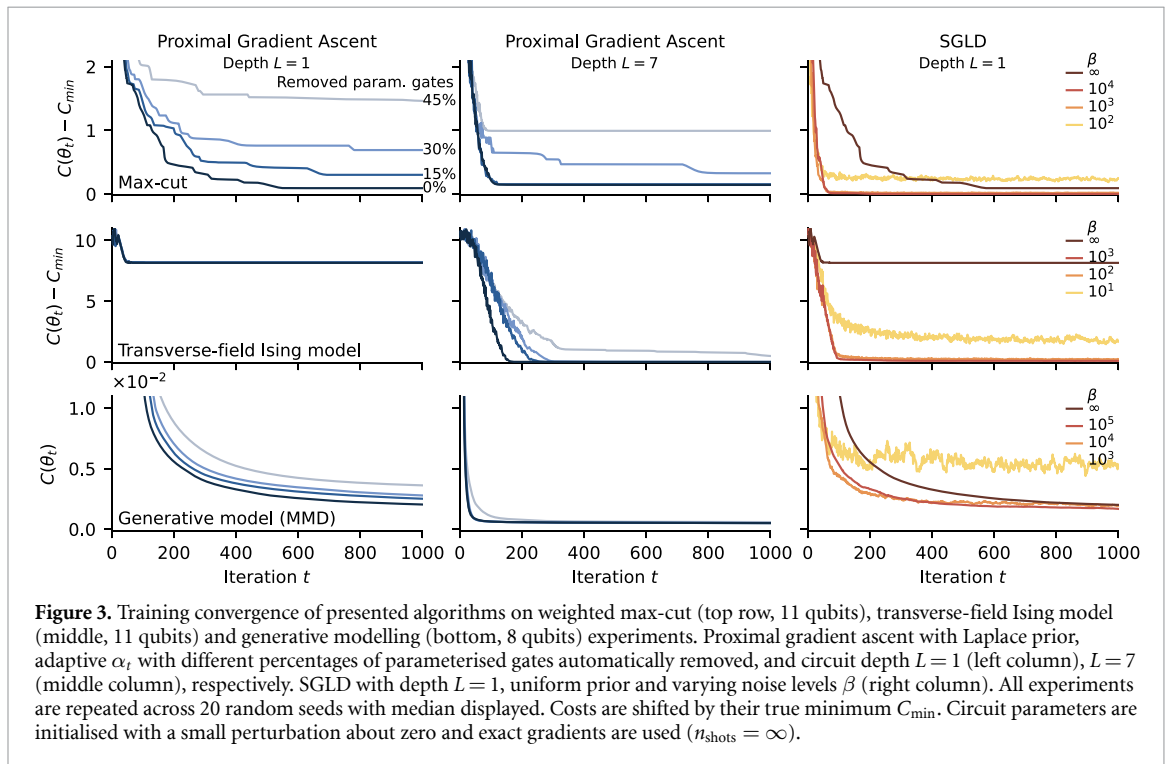

Figure 3. Training convergence of presented algorithms on weighted max-cut (top row, 11 qubits), transverse-field Ising model (middle, 11 qubits) and generative modelling (bottom, 8 qubits) experiments. Proximal gradient ascent with Laplace prior, adaptive α_t with different percentages of parameterised gates automatically removed, and circuit depth $L = 1$ (left column), $L = 7$ (middle column), respectively. SGLD with depth $L = 1$, uniform prior and varying noise levels β (right column). All experiments are repeated across 20 random seeds with median displayed. Costs are shifted by their true minimum $C_{\min}$. Circuit parameters are initialised with a small perturbation about zero and exact gradients are used ($n_{\text{shots}} = \infty$).

In the SGLD plot we use the shallower circuit ($L = 1$ layers). We observe that vanilla gradient ascent ($\beta = \infty$) is getting caught in local optima and that this is avoided by adding a suitable amount of noise in SGLD ($\beta = 10^3$ and $\beta = 10^4$), however adding too much noise ($\beta = 10^2$) prevents convergence to low cost regions.

5.2. TFIM

In the second experiment we are tasked with finding the ground state of a TFIM with nearest neighbour interactions

$$H = -\sum_{i=1}^{N-1} Z_i Z_{i+1} - g \sum_{i=1}^{N} X_i. \tag{8}$$

This is a model of quantum magnetism, where g corresponds to the applied transverse field in units of the Ising coupling strength [54]. For the experiments, we choose $N = 11$ and generate random TFIM instances by sampling $g \sim \mathcal{N}(\cdot \mid 0, \frac{1}{4})$. In the classical limit ($g = 0$) the ground state is either all spins up or down, and the system remains in a ferromagnetic phase for $|g| < 1$. It undergoes a phase transition at a critical point $|g| = 1$ and remains in a disordered phase for $|g| > 1$ [54]. The natural cost function for finding the ground state of equation (8) with a PQC is

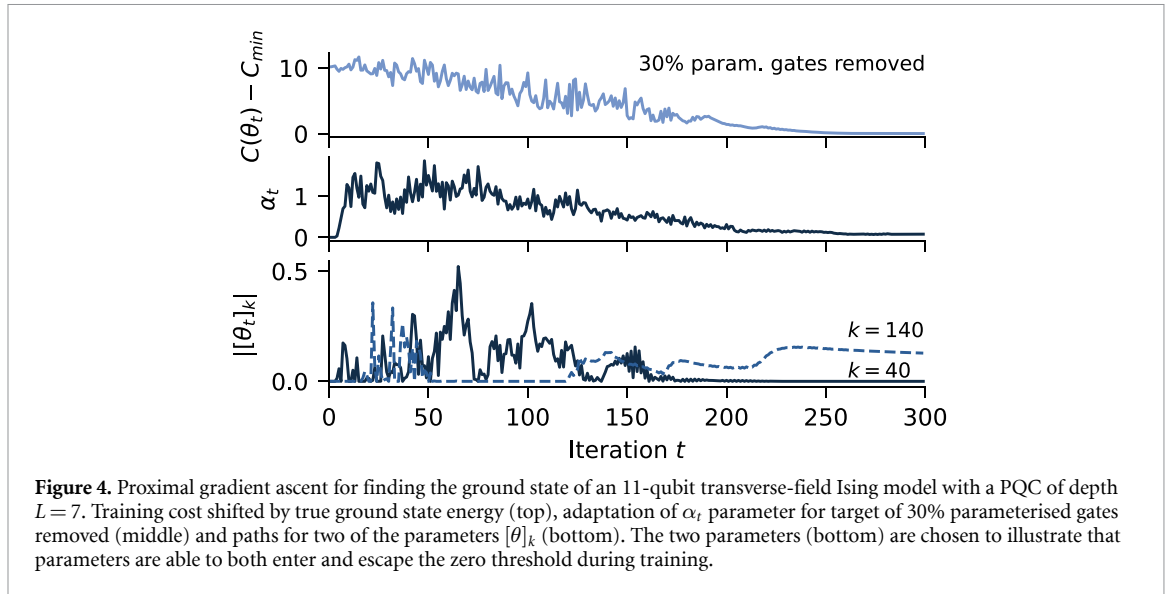$$C(\theta) = \langle 0|^{\otimes N} U(\theta)^\dagger H U(\theta) |0\rangle^{\otimes N}.$$

Figure 4. Proximal gradient ascent for finding the ground state of an 11-qubit transverse-field Ising model with a PQC of depth $L = 7$. Training cost shifted by true ground state energy (top), adaptation of α_t parameter for target of 30% parameterised gates removed (middle) and paths for two of the parameters $[\theta]_k$ (bottom). The two parameters (bottom) are chosen to illustrate that parameters are able to both enter and escape the zero threshold during training.

We run the same setup as in the weighted max-cut experiment and display the training results in the middle row of figure 3. We again observe severe local optima behaviour in the shallow circuit but this time the deep circuit successfully trains and finds the ground state even with 30% of the gates removed from the circuit. We also notice that the addition of noise in SGLD is very effective at escaping the local optima, although it does not find the exact ground energy, most likely due to the reduced expressivity of the shallow circuit versus the deep circuit.

For this experiment, we also visualise the adaptation of the Laplace prior regularisation parameter as described in section 3. In figure 4, for a specific instance of the quantum Hamiltonian we observe that the regularisation parameter α_t quickly becomes large and then decreases as the algorithm converges to low cost regions of the optimisation landscape, where at each iteration 30% of the parameters are set to zero. We additionally see that the set of zero parameters changes during training—as some parameters are dragged within the zero threshold this allows others to leave.

5.3. Generative modelling

In our final experiment, we use the PQC in figure 2 as a generative model (or Born machine) for the 1872 Hidalgo stamp dataset [55]. This dataset, y, represents measurements of the thickness of 485 stamps, on the μm scale these measurements are integers ranging from 60 μm to 131 μm and thus an 8-qubit generative model is sufficient to model in the binary expansion. This stamp data is displayed in the histograms on the top row of figure 5.

The goal of this generative modelling experiment is to drive the samples from the Born machine to be as close as possible to the true data set. The *ideal* posterior is $p(\theta \mid y) \propto p(\theta) \prod_{j=1}^{485} p([y]_j \mid \theta)$ where $[y]_j$ is a single datum of the dataset $y = \{[y]_j\}_{j=1}^{485}$ and $p([y]_j \mid \theta)$ is the probability of the Born machine generating said datum for given parameters θ. Pointwise evaluations of this likelihood $p([y]_j \mid \theta)$ are inherently intractable by the nature of quantum computation, and thus we cannot use it within our cost function. Instead, we utilise a generalised Bayesian inference framework [56], replacing the true loglikelihood with a two-sample test or scoring rule [34] that provides a measure of distance between a sample generated by the Born machine and the true data. In particular, we use the *maximum mean discrepancy* (MMD) [57, 58]

$$C(\theta) = \mathbb{E}_{p(z|\theta)p(z'|\theta)}[k(z,z')]$$
$$- 2\mathbb{E}_{p(z|\theta)\nu(z')}[k(z,z')] + \mathbb{E}_{\nu(z)\nu(z')}[k(z,z')],$$

where $\nu(z) = \frac{1}{485} \sum_{j=1}^{485} \delta(z \mid y_j)$ is the empirical distribution representing the dataset y. Here $k(z,z')$ is a kernel measuring the distance between the integers z and z', in this experiment we use a Gaussian kernel $k(z,z') = \exp(-(z-z')^2/(2\sigma^2))$ and set the bandwidth, σ, using the median heuristic [57] applied to the data y.

The training performance of proximal gradient ascent and SGLD is again plotted on the bottom row of figure 3. We observe that the deep circuit fits to the data very well and quickly even with 45% of the gates removed. The shallow circuit takes longer to fit although this is mitigated by the noise in SGLD—perhaps helping the parameters to more quickly escape a difficult region in the initialisation around zero.
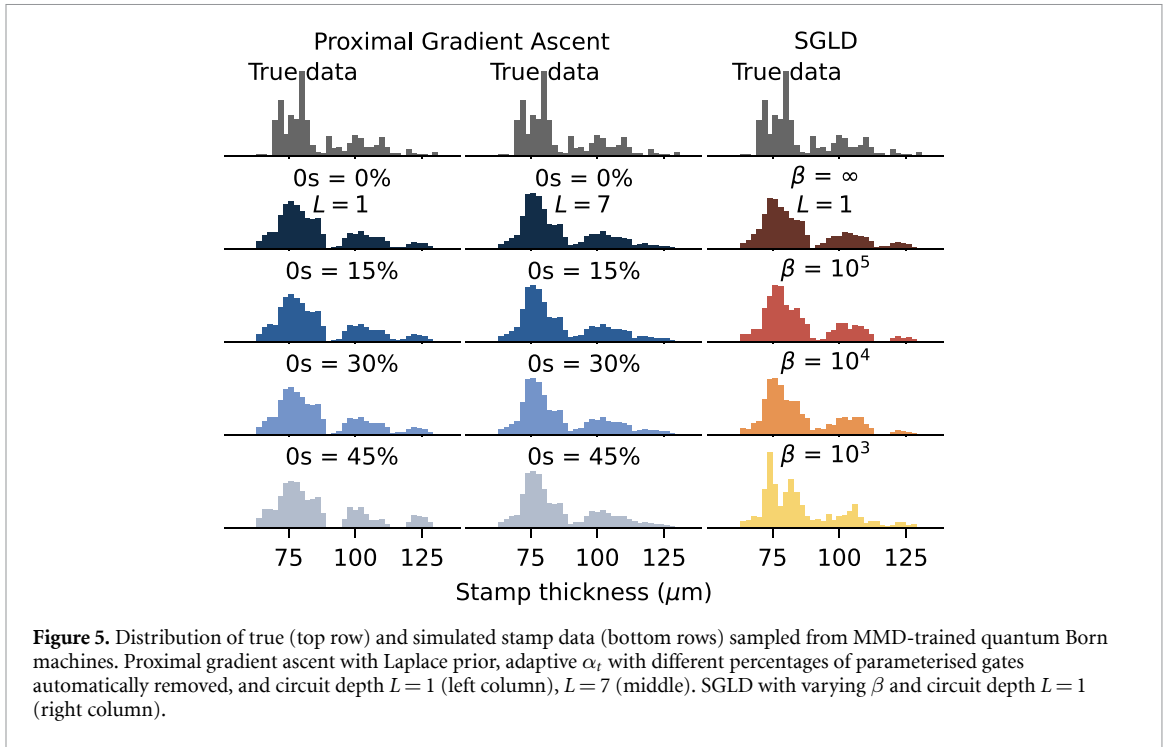
Figure 5. Distribution of true (top row) and simulated stamp data (bottom rows) sampled from MMD-trained quantum Born machines. Proximal gradient ascent with Laplace prior, adaptive α_t with different percentages of parameterised gates automatically removed, and circuit depth $L = 1$ (left column), $L = 7$ (middle). SGLD with varying β and circuit depth $L = 1$ (right column).
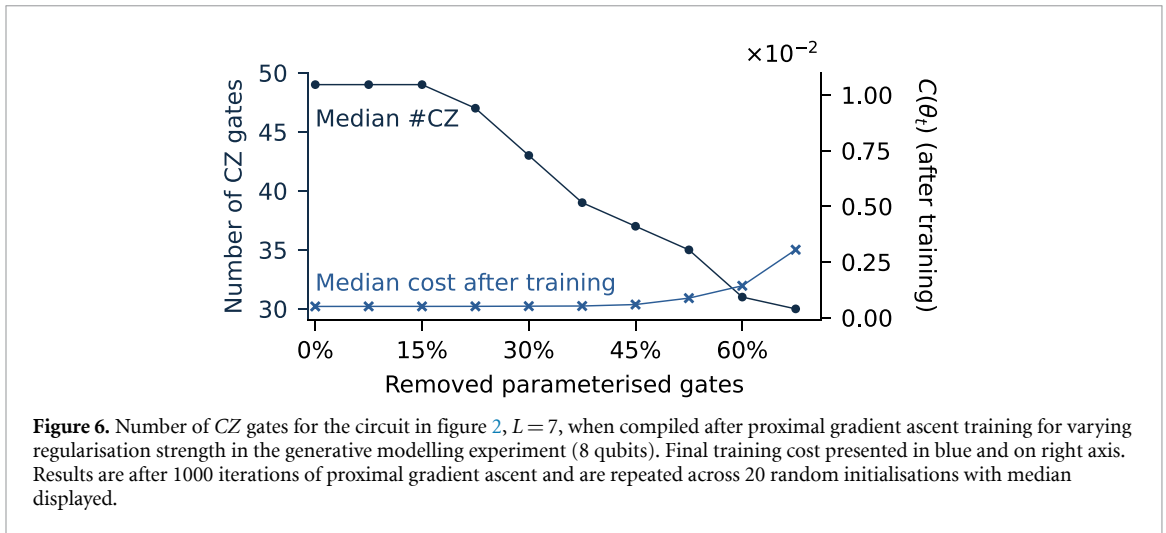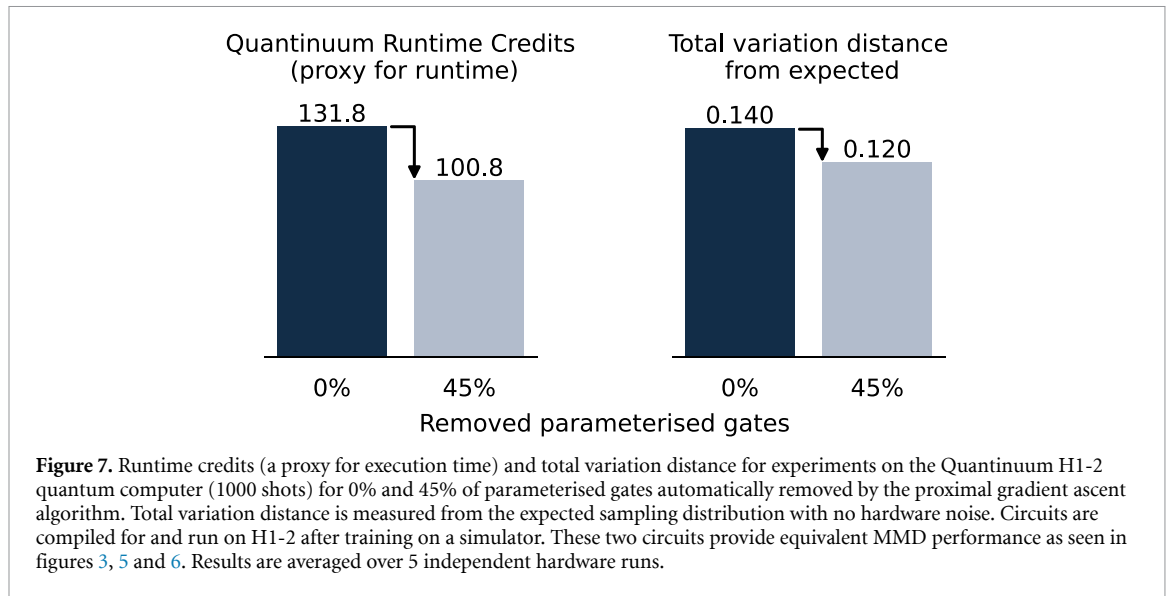

Figure 6. Number of CZ gates for the circuit in figure 2, $L = 7$, when compiled after proximal gradient ascent training for varying regularisation strength in the generative modelling experiment (8 qubits). Final training cost presented in blue and on right axis. Results are after 1000 iterations of proximal gradient ascent and are repeated across 20 random initialisations with median displayed.

In figure 5, we generate a simulated dataset of 10^3 samples using converged parameters and visually compare with the true data. We observe that the Laplace prior approach has fitted the data well with up to 30% of the gates removed in the shallow circuit, and up to 45% for the deep circuit. For the SGLD parameters, we take a different approach where we simulate our dataset by taking 100 samples from each parameter along the training trajectory (after discarding a burn-in of 400 samples), so-called ergodic averages. We observe that this approach provides an implicit regularisation and produces consistent simulated data even in the large noise setting $\beta = 10^3$.

In figure 6, we depict the number of CZ gates that are cancelled out during compilation of the circuit in figure 2 due to the regularisation of the Laplace prior and subsequent removal of R_x and R_z gates. We see that a significant proportion of CZ gates are removed when more than 40% of the number of the rotational parameters are removed although when more than 50% are removed we start to take a hit and suffer poorer performance on the trained cost.

Finally, in figure 7, we take two instances of the trained circuits with depth $L = 7$ and compare the sampling cost and hardware noise on the Quantinuum H1-2 trapped-ion quantum computer [59] after compiling via tket [60]. Observe that the circuit with 45% of parmterised gates removed is significantly cheaper and less noisy than its full parameter equivalent. We also remark that these benefits are multiplied by a factor of $2K$ in each gradient calculation if we employ the parameter shift rule.

Figure 7. Runtime credits (a proxy for execution time) and total variation distance for experiments on the Quantinuum H1-2 quantum computer (1000 shots) for 0% and 45% of parameterised gates automatically removed by the proximal gradient ascent algorithm. Total variation distance is measured from the expected sampling distribution with no hardware noise. Circuits are compiled for and run on H1-2 after training on a simulator. These two circuits provide equivalent MMD performance as seen in figures 3, 5 and 6. Results are averaged over 5 independent hardware runs.

6. Discussion

In this article we have described a very general Bayesian framework for the probabilistic treatment of variational quantum circuits. By considering the probability distribution $\pi(\theta) \propto p(\theta) \exp(-\beta C(\theta))$ we generalise the existing optimisation framework as a special case where a uniform prior $p(\theta) \propto 1$ is used implicitly and gradient ascent techniques are applied to find a maximum a posteriori estimate $\theta_{\text{MAP}} = \text{argmax}_\theta \pi(\theta)$. We move beyond the uniform prior and show how a Laplace distribution can be used to enforce customisable levels of sparsity in the parameter θ_{MAP}. This dimension reduction has benefits including faster sampling and reduced hardware noise, as well as potential trainability benefits for large problems. We additionally detail how to generate a Monte Carlo approximation that is asymptotically unbiased for the posterior $\pi(\theta)$ via an application of SGLD. A characterisation of the posterior beyond a point estimate is highly desirable for landscapes exhibiting complex contours and local optima, as demonstrated by the weighted max-cut and transverse field Ising model experiments.

This Bayesian perspective leaves many questions for future research, of which we will describe a few here.

Both of the described inference algorithms bring with them an additional tuning parameter, the regularisation strength α in the case of proximal gradient ascent and the level of noise added β in the case of SGLD. We described an adaptive method for the regularisation strength α, however this approach simply helps via the intuitive nature of deciding a priori how many parameters to remove, leaving an alternative tuning parameter. Another approach would be to put a prior (such as a gamma or log-normal distribution) on α or β and include them in the inference procedure.

An extremely successful modification of gradient descent in the case of stochastic gradients is the addition of momenta [61, 62]. Indeed, a compelling future direction is the extension to proximal gradient ascent with momenta [63] and the sampling analogue of momenta, underdamped Langevin dynamics [64, 65]. Whilst there is also the opportunity to incorporate second-order information into the Bayesian inference regimes via the use of a preconditioner [39, 48].

Additionally, we have only investigated point estimate and Monte Carlo approximations to the posterior. A natural next step would be to consider a variational inference approach [37], although care would need to be taken when constructing a variational family of distributions that are well-defined for angular parameters.

The posterior $\pi(\theta)$ represents an instance of generalised Bayesian inference [33, 56], there are indeed alternative posterior formulations providing a probabilistic interpretation of uncertainty over θ. A particularly compelling alternative approach corresponds to *approximate Bayesian computation* (e.g. [66]) which has desirable asymptotics and permits a Metropolis-Hastings accept-reject step. This approach is expanded on in appendix and represents a significant reformulation of the cost function and inference procedure although remains an intriguing future direction nonetheless.

One of the major concerns for the utility of VQAs is that of trainability in large circuits and the so-called barren plateau phenomenon [5] where the gradients of randomly initialised circuits vanish exponentially as the number of qubits increases. It is a natural question to consider whether the choice of prior can mitigate the barren plateau phenomenon either via automatic dimension reduction of the Laplace prior or by introducing correlations amongst parameters [67] via e.g. a (correlated) von Mises prior distribution [68].

A significant motivation for regularisation in classical statistics and machine learning is that of generalisation. ℓ_1 regularisation (without the proximal approach) is investigated for quantum supervised learning with preliminary mixed results in [42], it would be intriguing to investigate whether alternative priors could help quantum circuits avoid overfitting. In the same vein, classical Bayesian deep learning [69] (where the posterior samples are preferred over a point estimate) represents a compelling approach to improving generalisation, strongly motivating the extension of the quantum Bayesian learning framework described here to supervised learning.

Data availability statement

No new data were created or analysed in this study.

Acknowledgments

Appendix. A note on approximate Bayesian computation

A suitably decaying stepsize schedule is one approach to correct for the discretisation error in the Langevin proposal (7). Another possibility is to keep the stepsize constant and apply a Metropolis-Hastings accept-reject step, where a sample from a proposal distribution $q(\theta' \mid \theta_{t-1})$ is accepted $\theta_t = \theta'$ with probability

$$\alpha_t = \min(1, r_t), \qquad \text{where} \qquad r_t = \frac{\pi(\theta')q(\theta_{t-1} \mid \theta')}{\pi(\theta_{t-1})q(\theta' \mid \theta_{t-1})},$$

otherwise the previous is duplicated $\theta_t = \theta_{t-1}$.

The $\pi(\theta)$ evaluations within r_t only need to be up to normalising constant and can even be replaced with an unbiased estimate [46]. However, in the present formulation (3) we only have access to unbiased estimates of the (unnormalised) log density $\log p(\theta) - \beta C(\theta)$, which cannot be easily translated into an unbiased estimate of the required $p(\theta)\exp(-\beta C(\theta))$.

An alternative formulation instead works directly in the density space. Denote the quantum circuit as a conditional distribution $p(z \mid \theta)$ and a weighting function $k(z)$ that is large when the output z is accurate/desirable and small when z is inaccurate/undesirable. Note that we cannot evaluate $p(z \mid \theta)$ but can extract unbiased estimates for quantities of the form $\mathbb{E}_{p(z|\theta)}[f(z)]$ and $\nabla_\theta \mathbb{E}_{p(z|\theta)}[f(z)]$. This formulation falls within the field of approximate Bayesian computation (ABC) [66] where we target the extended distribution

$$\pi_{\mathrm{ABC}}(\theta, z) \propto p(\theta)p(z \mid \theta)k(z).$$

Discarding the simulated output z amounts to marginalisation

$$\pi_{\mathrm{ABC}}(\theta) \propto \int p(\theta)p(z \mid \theta)k(z)\mathrm{d}z,$$
$$\propto p(\theta)\mathbb{E}_{p(z|\theta)}[k(z)].$$

The extended distribution permits a Metropolis-Hastings step [46] and a Langevin proposal, although the Langevin proposal requires the gradient

$$\nabla_\theta \log \pi_{\mathrm{ABC}}(\theta) = \nabla_\theta \log p(\theta) + \nabla_\theta \log \mathbb{E}_{p(z|\theta)}[k(z)],$$
$$= \nabla_\theta \log p(\theta) + \frac{\nabla_\theta \mathbb{E}_{p(z|\theta)}[k(z)]}{\mathbb{E}_{p(z|\theta)}[k(z)]},$$

which differs from (5).

In statistics or machine learning settings, e.g. section 5.3, we have $k(z) = k(z, y)$ encouraging the output z to be similar to a given dataset y. Here $\pi_{\mathrm{ABC}}(\theta)$ has the desirable property that as $k(z, y) \to \delta(z \mid y)$ we get $\pi_{\mathrm{ABC}}(\theta) \to p(\theta \mid z) \propto p(\theta)p(z \mid \theta)$ which is in some sense the *ideal* posterior. ABC targets an extended distribution and this has largely limited the approach to low dimensional settings, however this could be mitigated by the use of the gradients above (ABC is usually gradient-free) or by accepting a bias [70]. However, this alternative ABC formulation loses the seamless transition from existing optimisation-based VQAs, algorithm 1; a numerical investigation into $\pi_{\mathrm{ABC}}(\theta)$ is therefore left for future work.

ORCID iDs

Samuel Duffield ⬤ https://orcid.org/0000-0002-8656-8734
Marcello Benedetti ⬤ https://orcid.org/0000-0003-0231-1729
Matthias Rosenkranz ⬤ https://orcid.org/0000-0002-1605-9141

References

[1] McClean J R, Romero J, Babbush R and Aspuru-Guzik A 2016 The theory of variational hybrid quantum-classical algorithms *New J. Phys.* **18** 023023
[2] Benedetti M, Lloyd E, Sack S H and Fiorentini M 2019 Parameterized quantum circuits as machine learning models *Quantum Sci. Technol.* **4** 043001
[3] Cerezo M *et al* 2021 Variational quantum algorithms *Nat. Rev. Phys.* **3** 625
[4] Bharti K *et al* 2022 Noisy intermediate-scale quantum algorithms *Rev. Mod. Phys.* **94** 015004
[5] McClean J R, Boixo S, Smelyanskiy V N, Babbush R and Neven H 2018 Barren plateaus in quantum neural network training landscapes *Nat. Commun.* **9** 4812
[6] Cerezo M, Sone A, Volkoff T, Cincio L and Coles P J 2021 Cost function dependent barren plateaus in shallow parametrized quantum circuits *Nat. Commun.* **12** 1791
[7] You X and Wu X 2021 Exponentially many local minima in quantum neural networks *Proc. 38th Int. Conf. on Machine Learning* (*Proc. Machine Learning Research* vol 139) ed M Meila and T Zhang (PMLR) pp 12144–55
[8] Anschuetz E R and Kiani B T 2022 Beyond barren plateaus:quantum variational algorithms are swamped with traps (arXiv:2205.05786)
[9] Egger D J, Marecek J and Woerner S 2021 Warm-starting quantum optimization *Quantum* **5** 479
[10] Jain N, Coyle B, Kashefi E and Kumar N 2021 Graph neural network initialisation of quantum approximate optimisation (arXiv:2111.03016)
[11] Zhang K, Hsieh M-H, Liu L and Tao D 2022 Gaussian initializations help deep variational quantum circuits escape from the barren plateau (arXiv:2203.09376)
[12] Gard B T, Zhu L, Barron G S, Mayhall N J, Economou S E and Barnes E 2020 Efficient symmetry-preserving state preparation circuits for the variational quantum eigensolver algorithm *npj Quantum Inf.* **6** 1
[13] Cong I, Choi S and Lukin M D 2019 Quantum convolutional neural networks *Nat. Phys.* **15** 1273
[14] Meyer J J, Mularski M, Gil-Fuster E, Mele A A, Arzani F, Wilms A and Eisert J 2022 Exploiting symmetry in variational quantum machine learning (arXiv:2205.06217)
[15] Skolik A, McClean J R, Mohseni M, van der Smagt P and Leib M 2021 Layerwise learning for quantum neural networks *Quantum Mach. Intell.* **3** 5
[16] Zhang K, Hsieh M-H, Liu L and Tao D 2021 Toward trainability of deep quantum neural networks (arXiv:2112.15002)
[17] Sack S H, Medina R A, Michailidis A A, Kueng R and Serbyn M 2022 Avoiding barren plateaus using classical shadows *PRX Quantum* **3** 020365
[18] Otterbach J S *et al* 2017 Unsupervised machine learning on a hybrid quantum computer (arXiv:1712.05771)
[19] Zhu D *et al* 2019 Training of quantum circuits on a hybrid quantum computer *Sci. Adv.* **5** eaaw9918
[20] Self C N, Khosla K E, Smith A W R, Sauvage F, Haynes P D, Knolle J, Mintert F and Kim M S 2021 Variational quantum algorithm with information sharing *npj Quantum Inf.* **7** 116
[21] Tamiya S and Yamasaki H 2021 Stochastic gradient line Bayesian optimization: reducing measurement shots in optimizing parameterized quantum circuits (arXiv:2111.07952)
[22] Rad A, Seif A and Linke N M 2022 Surviving the barren plateau in variational quantum circuits with Bayesian learning initialization (arXiv:2203.02464)
[23] Wang G, Koh D E, Johnson P D and Cao Y 2021 Minimizing estimation runtime on noisy quantum computers *PRX Quantum* **2** 010346
[24] Benedetti M, Coyle B, Fiorentini M, Lubasch M and Rosenkranz M 2021 Variational inference with a quantum computer *Phys. Rev. Appl.* **16** 044057
[25] Du Y, Hsieh M-H, Liu T and Tao D 2020 Expressive power of parametrized quantum circuits *Phys. Rev. Res.* **2** 033125
[26] Low G H, Yoder T J and Chuang I L 2014 Quantum inference on Bayesian networks *Phys. Rev. A* **89** 062315
[27] Harrow A W and Wei A Y 2020 Adaptive quantum simulated annealing for Bayesian inference and estimating partition functions *Proc. 2020 ACM-SIAM Symp. on Discrete Algorithms* (*SODA*) pp 193–212
[28] Tucci R R 1995 Quantum Bayesian nets *Int. J. Mod. Phys.* B **09** 295
[29] Leifer M and Poulin D 2008 Quantum graphical models and belief propagation *Ann. Phys., NY* **323** 1899
[30] Allen J-M A, Barrett J, Horsman D C, Lee C M and Spekkens R W 2017 Quantum common causes and quantum causal models *Phys. Rev. X* **7** 031021
[31] Harrow A and Napp J 2021 Low-depth gradient measurements can improve convergence in variational hybrid quantum-classical algorithms *Phys. Rev. Lett.* **126** 140502
[32] Sweke R, Wilde F, Meyer J, Schuld M, Faehrmann P K, Meynard-Piganeau B and Eisert J 2020 Stochastic gradient descent for hybrid quantum-classical optimization *Quantum* **4** 314
[33] Bissiri P G, Holmes C C and Walker S G 2016 A general framework for updating belief distributions *J. R. Stat. Soc.* B **78** 1103
[34] Pacchiardi L and Dutta R 2021 Generalized Bayesian likelihood-free inference using scoring rules estimators (arXiv:2104.03889)
[35] Brooks S, Gelman A, Jones G and Meng X-L 2011 *Handbook of Markov Chain Monte Carlo* (Boca Raton, FL: CRC Press)
[36] Chopin N and Papaspiliopoulos O 2020 *An Introduction to Sequential Monte Carlo* (*Springer Series in Statistics*) (New York: Springer) (https://books.google.co.uk/books)
[37] Blei D M, Kucukelbir A and McAuliffe J D 2017 Variational inference: a review for statisticians *J. Am. Stat. Assoc.* **112** 859
[38] Tibshirani R 1996 Regression shrinkage and selection via the lasso *J. R. Stat. Soc.* B **58** 267
[39] Parikh N and Boyd S 2014 Proximal algorithms *Found. Trends Optim.* **1** 127–239
[40] Combettes P L and Wajs V R 2005 Signal recovery by proximal forward-backward splitting *Multiscale Model. Simul.* **4** 1168
[41] Williams P M 1995 Bayesian regularization and pruning using a Laplace prior *Neural Comput.* **7** 117
[42] Qian Y, Wang X, Du Y, Wu X and Tao D 2021 The dilemma of quantum neural networks (arXiv:2106.04975)

[43] Polson N G, Scott J G and Willard B T 2015 Proximal algorithms in statistics and machine learning *Stat. Sci.* **30** 559

[44] Ishwaran H and Rao J S 2005 Spike and slab variable selection: frequentist and Bayesian strategies *Ann. Stat.* **33** 730

[45] Carvalho C M, Polson N G and Scott J G 2009 Handling sparsity via the horseshoe *Proc. 12th Int. Conf. on Artificial Intelligence and Statistics* (*Proc. Machine Learning Research* vol 5) (*Hilton Clearwater Beach Resort, Clearwater Beach, FL, USA*) ed D van Dyk and M Welling (PMLR) pp 73–80

[46] Andrieu C and Roberts G O 2009 The pseudo-marginal approach for efficient Monte Carlo computations *Ann. Stat.* **37** 697

[47] Welling M and Teh Y W 2011 Bayesian learning via stochastic gradient Langevin dynamics *Proc. 28th Int. Conf. on Machine Learning (ICML-11)* pp 681–8

[48] Ma Y-A, Chen T and Fox E 2015 A Complete recipe for stochastic gradient MCMC *Advances in Neural Information Processing Systems* vol 28, ed C Cortes, N Lawrence, D Lee, M Sugiyama and R Garnett (Curran Associates, Inc.)

[49] Teh Y W, Thiery A H and Vollmer S J 2016 Consistency and fluctuations for stochastic gradient Langevin dynamics *J. Mach. Learn. Res.* **17** 193–225

[50] Izmailov P, Podoprikhin D, Garipov T, Vetrov D, and Wilson A G, 2018 Averaging weights leads to wider optima and better generalization (arXiv:1803.05407)

[51] Cheng S, Chen J and Wang L 2018 Information perspective to probabilistic modeling: Boltzmann machines versus born machines *Entropy* **20** 583

[52] Li J, Yang X, Peng X and Sun C-P 2017 Hybrid quantum-classical approach to quantum optimal control *Phys. Rev. Lett.* **118** 150503

[53] Duffield S, Matos G and Johannsen M 2022 Qujax: simulating quantum circuits with JAX (available at: https://github.com/CQCL/qujax)

[54] Sachdev S 2011 *Quantum Phase Transitions* 2nd edn (Cambridge: Cambridge University Press)

[55] Izenman A J and Sommer C J 1988 Philatelic mixtures and multimodal densities *J. Am. Stat. Assoc.* **83** 941

[56] Knoblauch J, Jewson J, and Damoulas T, 2019 Generalized variational inference: three arguments for deriving new posteriors (arXiv:1904.02063)

[57] Gretton A, Borgwardt K M, Rasch M J, Schölkopf B and Smola A 2012 A kernel two-sample test *J. Mach. Learn. Res.* **13** 723

[58] Liu J-G and Wang L 2018 Differentiable learning of quantum circuit born machines *Phys. Rev.* A **98** 062324

[59] Pino J M *et al* 2021 Demonstration of the trapped-ion quantum CCD computer architecture *Nature* **592** 209

[60] Sivarajah S, Dilkes S, Cowtan A, Simmons W, Edgington A and Duncan R 2020 t|ket⟩: a retargetable compiler for NISQ devices *Quantum Sci. Technol.* **6** 014003

[61] Nesterov Y 1983 A method for solving the convex programming problem with convergence rate $o(1/k^2)$ *Proc. USSR Acad. Sci.* **269** 543

[62] Kingma D P and Ba J 2015 Adam: a method for stochastic optimization *3rd Int. Conf. on Learning Representations (ICLR 2015)* (*San Diego, CA, USA, 7–9 May 2015*) (*Conf. Track Proc.*) ed Y Bengio and Y LeCun

[63] Li Q, Zhou Y, Liang Y and Varshney P K 2017 Convergence analysis of proximal gradient with momentum for nonconvex optimization *Proc. 34th Int. Conf. on Machine Learning* (*Proc. Machine Learning Research* vol 70) ed D Precup and Y W Teh (PMLR) pp 2111–9

[64] Chen T, Fox E and Guestrin C 2014 Stochastic gradient Hamiltonian Monte Carlo *Proc. 31st Int. Conf. on Machine Learning* (*Proc. Machine Learning Research* vol 32) ed E P Xing and T Jebara (Bejing: PMLR) pp 1683–91

[65] Leimkuhler B and Shang X 2016 Adaptive thermostats for noisy gradient systems *SIAM J. Sci. Comput.* **38** A712

[66] Beaumont M A 2019 Approximate Bayesian computation *Annu. Rev. Stat. Appl.* **6** 379

[67] Volkoff T and Coles P J 2021 Large gradients via correlation in random parameterized quantum circuits *Quantum Sci. Technol.* **6** 025008

[68] Mardia K V, Jupp P E and Mardia K 2000 *Directional Statistics* vol 2 (New York: Wiley)

[69] Wilson A G and Izmailov P 2020 Bayesian deep learning and a probabilistic perspective of generalization *Proc. 34th Int. Conf. on Neural Information Processing Systems* (*NIPS'20*) (Red Hook, NY: Curran Associates Inc.)

[70] Duffield S and Singh S S 2022 Ensemble Kalman inversion for general likelihoods *Stat. Probab. Lett.* **187** 109523