Quantum Science and Technology

OPEN ACCESS

RECEIVED
20 June 2022

REVISED
12 October 2022

ACCEPTED FOR PUBLICATION
2 December 2022

PUBLISHED
14 December 2022

PAPER

Markov chain Monte Carlo enhanced variational quantum algorithms

Taylor L Patti[1,2,*], Omar Shehab[2], Khadijeh Najafi[1,2] and Susanne F Yelin[1]

[1] Department of Physics, Harvard University, Cambridge, MA 02138, United States of America
[2] IBM Quantum, IBM T.J. Watson Research Center, Yorktown Heights, NY 10598, United States of America
[*] Author to whom any correspondence should be addressed.

E-mail: tpatti@nvidia.com

Keywords: quantum machine learning, quantum computing, variational quantum algorithm, Markov chain Monte Carlo, quantum optimization, Gibbs state

Abstract

Variational quantum algorithms have the potential for significant impact on high-dimensional optimization, with applications in classical combinatorics, quantum chemistry, and condensed matter. Nevertheless, the optimization landscape of these algorithms is generally nonconvex, leading the algorithms to converge to local, rather than global, minima and the production of suboptimal solutions. In this work, we introduce a variational quantum algorithm that couples classical Markov chain Monte Carlo techniques with variational quantum algorithms, allowing the former to provably converge to global minima and thus assure solution quality. Due to the generality of our approach, it is suitable for a myriad of quantum minimization problems, including optimization and quantum state preparation. Specifically, we devise a Metropolis–Hastings method that is suitable for variational quantum devices and use it, in conjunction with quantum optimization, to construct quantum ensembles that converge to Gibbs states. These performance guarantees are derived from the ergodicity of our algorithm's state space and enable us to place analytic bounds on its time-complexity. We demonstrate both the effectiveness of our technique and the validity of our analysis through quantum circuit simulations for MaxCut instances, solving these problems deterministically and with perfect accuracy, as well as large-scale quantum Ising and transverse field spin models of up to 50 qubits. Our technique stands to broadly enrich the field of variational quantum algorithms, improving and guaranteeing the performance of these promising, yet often heuristic, methods.

1. Introduction

Since the advent of the variational quantum eigensolver (VQE) [1, 2] and quantum approximate optimization algorithm (QAOA) [3], quantum algorithms that function in tandem with classical machine learning have garnered great interest. These variational quantum algorithms (VQAs) typically harness some form of classical gradient descent to tackle a large-scale optimization problem on the exponential state space of quantum hardware [4, 5]. Applications of these methods have included the optimization of NP-hard combinatorial problems [6–10], the identification of eigenstates and energies in quantum chemistry applications [11–13], and the study of condensed matter systems [14–16]. Much like their classical counterparts, the above near-term quantum algorithms can be plagued by nonconvex optimization landscapes, causing them to converge to suboptimal minima [17]. A variety of techniques have been suggested to address this issue in NP-hard combinatorial optimization problems, such as: 'warm starting' procedures [18–20], composition with classical neural networks [21], multibasis encodings with bistable convergence [9], and other techniques [10, 22, 23]. However, these methods offer few provable optimization guarantees of practical utility. While optimization landscapes are known to become more convex with high-depth [17], additional methods of mitigating nonconvexity should be explored as the adverse effect of quantum noise [24, 25] and barren plateaus [26–30] on deep quantum networks is well-documented.

In order to avoid the local minima convergence that plagues VQAs without increasing quantum circuit depth, we introduce MCMC-VQA, a technique that adapts the ergodic exploration of classical Markov chain Monte Carlo (MCMC) to guarantee the global convergence of quantum algorithms. As samples of ergodic systems are representative of their underlying probability distribution, an ergodic VQA necessarily yields a sample that contains states near the global minimum, also known as the ground truth. In particular, our method combines the intrepid sampling of MCMC methods with quantum optimization in order to create and sample from a quantum ensemble that approximates a Gibbs state [31], whose most probable quantum state is that which encodes the global minimum of the optimization problem. Optimization is then finalized by carrying out standard extremization techniques (e.g. gradient descent) that are initialized with the lowest energy state obtained during sampling. Assuming that the quantum circuit is expressive enough (i.e. has a sufficiently intricate ansatz) to produce the distribution's ground state, our method is analytically guaranteed to return solutions close to this ground truth. In this work, the MCMC method that we focus on is the Metropolis–Hastings algorithm, chosen due to its success in high-dimensional spaces and its suitability for unnormalized probability distributions [32]. MCMC-VQA utilizes modified VQAs and their statistics as the Metropolis–Hastings transition kernels (probabilistic transition maps) and quantum state energies as state likelihoods. These quantities are then used to determine the viability of parameter updates. As the quantum components of our algorithm rely on the same expectation value estimations requisite in the original loss function, they do not represent an increase in quantum overhead. Likewise, our algorithm's increase of classical overhead is limited to the evaluation of probability density functions, the calculation of simple functions of expectation value estimates, and the generation of pseudo-random numbers, and thus represents a minimal amount of compute. MCMC-VQA represents a time-discrete, space-continuous Markov chain, as the algorithm progresses in discrete VQA epochs (time steps) while training a continuous-parameter quantum circuit. It can also be classified as a form of Stochastic Gradient Descent MCMC [33, 34]. Although in this work we focus on VQE [2], the crux of our technique is a sampling process that uses Markov chain Monte Carlo comparisons between states that are partially prepared using a combination of gradient descent and stochastic noise. As such, our method is readily applicable to any quantum algorithm that extremizes a loss function using a variational quantum circuit.

While other works have introduced quantum subroutines for classical MCMC methods that offer a quadratic speedup for random walks [35–37] and sampling [38, 39], this manuscript takes the opposite approach by designing a classical MCMC subroutine for quantum algorithms. Likewise, while classical MCMC methods have been used to *simulate* quantum computing routines [40, 41], our work is unique in that it uses classical MCMC to *enhance* the performance of variational quantum algorithms. Similarly, while the preparation of a Gibbs state on a variational quantum computer had been previously proposed via an approximate Fourier series [42] and free energy minimization [43], and has been suggested since the release of this work using time evolution [44] and efficient free energy minimization [45], these methods do not employ MCMC techniques.

We briefly outline VQAs, focusing on VQE (figure 1, gray) for quantum optimization of MaxCut problems, the quantum Ising model, and more general nonlocal transverse spin models. This choice of these applications is motivated by the ample nonconvexity of the corresponding quadratic MaxCut loss functions [9, 17], as well as the relevance of the Ising and other spin models for quantum chemistry and condensed matter physics. VQAs are parameterized by input states $|\psi\rangle$ and quantum circuit unitaries $U_t = U(\hat{\theta}_t)$, where $\hat{\theta}_t$ are the variable parameters learned during epoch $t-1$. Without loss of generality, we choose the n-qubit input state as $|0\rangle = \prod_{i=0}^{n-1}|0\rangle$ such that the output state is entirely defined by $\hat{\theta}$ and assume that the initial parameters $\hat{\theta}_0$ are randomly selected at the start of each new sequence of epochs.

MaxCut is a partitioning problem on undirected graphs Γ (figure 1, black), where edges $\omega_{i,ab}$ connect the ith pair of vertices v_a, v_b, with vertex numbers a and b, respectively [46]. The goal is to optimally assign all vertices $v_a, v_b \in \{-1,1\}$, so as to maximize the objective function

$$\text{maximize} \quad \frac{1}{2}\sum_i w_{i,ab}\left(1 - v_a v_b\right). \tag{1}$$

In this work, we will consider a generalized form of the problem known as *weighted* MaxCut, in which w_i take arbitrary real values.

To solve MaxCut via VQE, a graph Γ is encoded in the Ising model Hamiltonian

$$H = \sum_i \omega_{i,ab}\sigma_a\sigma_b, \tag{2}$$

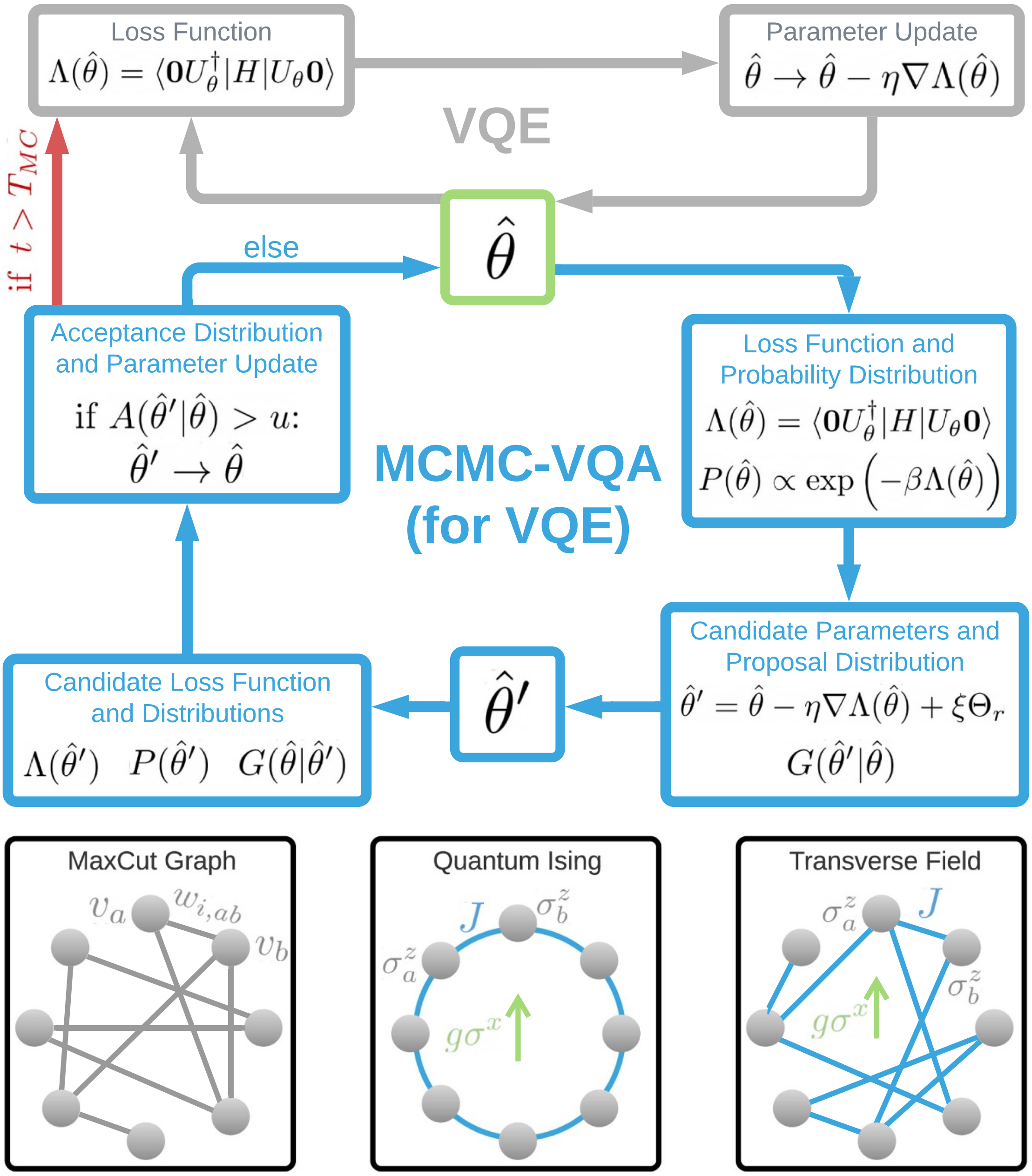
Loss Function
$\Lambda(\hat\theta) = \langle \mathbf{0} U_\theta^\dagger | H | U_\theta \mathbf{0} \rangle$
Parameter Update
$\hat\theta \to \hat\theta - \eta \nabla \Lambda(\hat\theta)$
VQE
if $t > T_{MC}$
else
$\hat\theta$
Acceptance Distribution and Parameter Update
if $A(\hat\theta'|\hat\theta) > u$:
$\hat\theta' \to \hat\theta$
MCMC-VQA (for VQE)
Loss Function and Probability Distribution
$\Lambda(\hat\theta) = \langle \mathbf{0} U_\theta^\dagger | H | U_\theta \mathbf{0} \rangle$
$P(\hat\theta) \propto \exp\left(-\beta \Lambda(\hat\theta)\right)$
Candidate Loss Function and Distributions
$\Lambda(\hat\theta')$ $P(\hat\theta')$ $G(\hat\theta|\hat\theta')$
$\hat\theta'$
Candidate Parameters and Proposal Distribution
$\hat\theta' = \hat\theta - \eta \nabla \Lambda(\hat\theta) + \xi \Theta_r$
$G(\hat\theta'|\hat\theta)$
MaxCut Graph
v_a
$w_{i,ab}$
v_b
Quantum Ising
J
σ_b^z
σ_a^z
$g\sigma^x$
Transverse Field
σ_a^z
J
σ_b^z
$g\sigma^x$

Figure 1. Diagram of the MCMC-VQA algorithm for VQE. VQE (gray, section 1) minimizes the loss function for each $\hat\theta$ by calculating the expectation value $\Lambda(\hat\theta)$ and updating $\hat\theta$ with gradient descent using $\nabla\Lambda(\hat\theta)$. MCMC-VQA for the VQE algorithm (blue, section 2) uses gradient descent with $\nabla\Lambda(\hat\theta)$ and random noise $\xi\Theta_r$ to produce candidate state $\hat\theta\,'$, but also calculates probabilities $P(\hat\theta)$ and $P(\hat\theta\,')$, as well as proposal probabilities $G(\hat\theta\,'|\hat\theta)$ and $G(\hat\theta|\hat\theta\,')$. Using these distribution samples, the acceptance probability $A(\hat\theta\,'|\hat\theta)$ is calculated and compared to random uniform sample $u \sim U(0,1)$. If $A(\hat\theta\,'|\hat\theta) > u$, then $\hat\theta\,' \to \hat\theta$. Otherwise, the MCMC-VQA algorithm restarts with the original $\hat\theta$. (Red) after the maximum number of MCMC-VQA epochs (time steps) T_{MC} have occurred, the sampled parameters with the lowest loss, $\hat\theta_{\min}$, are selected and the optimization completes with a closing sequence of VQE epochs. Hamiltonian models (black insets, sections 1 and 3) MaxCut graphs in this work are generated with normally distributed edge weights $w_{i,ab}$. The objective is to minimize equation (1) by optimally assigning each pair of vertices $v_a, v_b \in \{-1, 1\}$. MaxCut can be solved on a quantum computer by mapping $v_a, v_b \to \sigma_a, \sigma_b$ and minimizing the corresponding H. The quantum Ising model has nearest-neighbor (local) ZZ-coupling and an X-axis transverse field, as defined in equation (4). In this work, we choose $J > 0$ such that the ordered phase is ferromagnetic. The transverse field spin model is a more general (nonlocal) counterpart of the quantum Ising model, as defined in equation (5). In the limit of $g \to 0$, the transverse field reduces to the MaxCut Hamiltonian. See section 3 for Hamiltonian details.

where $\omega_{i,ab}$ remains unchanged from the MaxCut objective function and $v_a, v_b \to \sigma_a, \sigma_b$ for Pauli-Z spin operators σ_a, σ_b. Maximizing the cut of Γ is then equivalent to minimizing the loss function

$$\Lambda_t = \Lambda(\hat\theta_t) = \langle 0 | (U_t^\dagger | H | U_t) | 0 \rangle = \sum_i \omega_{i,ab} \langle \sigma_a \sigma_b \rangle_t = \sum_i \mu_t^i, \qquad (3)$$

where μ_t^i are the expectation values of the quadratic MaxCut terms. VQE circuit training updates parameters $\hat{\theta}$ via gradient descent on Λ_t (figure 1), where the gradient of any $\theta_t^k \in \hat{\theta}_t$ can be calculated as $\nabla_k \Lambda(\hat{\theta}_t) = \left(\Lambda(\hat{\theta}_t + \epsilon \hat{k}) - \Lambda(\hat{\theta}_t - \epsilon \hat{k})\right)/2\epsilon$ by finite difference. As $\nabla \Lambda(\hat{\theta}_t) \rightarrow 0$ in the vicinity of both global *and local* minima, VQE training is prone to stagnation at suboptimal solutions.

In this work, we also explore the preparation of low-energy states of the quantum Ising and more general nonlocal transverse field spin models (see figure 1). The quantum Ising model is defined as [47]

$$H_{\mathrm{QI}} = -J \sum_{i,i+1} \sigma_i^z \sigma_{i+1}^z - g \sum_i \sigma_i^x \tag{4}$$

where the summation over σ^z terms is between nearest-neighbors only. Moreover, we assume a quantum Ising model with periodic boundary conditions, resulting in a ring, rather than a chain, of qubit–qubit interactions. Similarly, more general spin–spin interactions with transverse field can be described with the Hamiltonian

$$H_{\mathrm{TF}} = -J \sum_{i,j} \sigma_i^z \sigma_{i+1}^z - g \sum_i \sigma_i^x \tag{5}$$

where the indices i, j denote some specified set of two-qubit pairs, which are not necessarily nearest-neighbors (i.e. which can be nonlocal interactions).

2. Results

In this section, we present our novel method for enhancing the performance of VQAs with classical MCMCs, a technique that we dub MCMC-VQA. We start by briefly reviewing traditional MCMC, focusing on the Metropolis–Hastings algorithm. Then, we introduce MCMC-VQA, derive its behavior, and verify our findings with numerical simulations.

2.1. MCMC-VQA method

MCMC algorithms, such as Metropolis–Hastings, combine the randomized sampling of Monte-Carlo methods with the Markovian dynamics of a Markov chain in order to randomly sample from a distribution that is difficult to characterize deterministically [32]. MCMC is particularly useful for approximations in high-dimensional spaces, where the so-called 'curse of dimensionality' can make techniques such as random sampling prohibitively slow [48]. The core merit of MCMC techniques is their ergodicity, which guarantees that all states of the distribution are eventually sampled in a statistically representative way, regardless of which initial point is chosen. This representative sample is known as the unique stationary distribution π. In particular, any Markov chain that is both irreducible (each state has a non-zero probability of transitioning to any other state) and aperiodic (not partitioned into sets that undergo periodic transitions) will provably converge to its unique stationary distribution π, from which it samples ergodically [49]. The mathematical properties of ergodic Markov chains are well-studied, including analytic bounds for solution quality and mixing time (number of epochs) [50, 51].

In order to obtain π for a distribution of interest, Metropolis–Hastings specifies the transition kernel $P(x'|x)$, which is the probability that state x transitions to state $x^{'}$. Typically, the Markov process is defined such that transitions satisfy the detailed balance condition:

$$P(x)P(x'|x) = P(x')P(x|x'). \tag{6}$$

When equation (6) holds, the chain is said to be reversible and is guaranteed to converge to a stationary distribution. $P(x'|x)$ can be factored into two quantities

$$P(x'|x) = G(x'|x)A(x'|x), \tag{7}$$

where $G(x'|x)$ is the proposal distribution, or the conditional probability of proposing state $x^{'}$ given state x, and $A(x'|x)$ is the acceptance distribution, or the probability of accepting the new state $x^{'}$ given state x. To satisfy equation (6), the acceptance distribution is defined as

$$A(x'|x) = \min\left(1, \frac{P(x')G(x|x')}{P(x)G(x'|x)}\right). \tag{8}$$

Note that as only the ratio $P(x')/P(x)$ is considered, the probability distribution need not be normalized. To determine whether the candidate state $x^{'}$ or the current state x_t should be used as the future state x_{t+1}, a sample u is drawn from the uniform distribution $U(0,1)$. If $A(x'|x_t) \geqslant u$, then $x_{t+1} = x'$ and we say that the candidate state $x^{'}$ is accepted. Otherwise, $x_{t+1} = x_t$ and we say that $x^{'}$ is rejected.

We now present the MCMC-VQA method. Figure 1 contains a diagram of the algorithm (blue). In particular, we focus on an ergodic Metropolis–Hastings algorithm, which is guaranteed to sample states near global minima. We outline the algorithm both idealistically and experimentally, prove its ergodicity and convergence, and verify these findings with numerical simulations.

As we seek the lowest energy eigenstate when solving for the low-lying states of Hamiltonians via VQE. We define $P(\hat{\theta})$ as the Boltzmann distribution

$$P(\hat{\theta}_a) = \exp\left(-\beta\Lambda_a\right)/Z, \qquad Z = \sum_i \exp\left(-\beta\Lambda_i\right), \tag{9}$$

such that a state's probability increases exponentially with decreasing loss function.

To calculate the proposal distribution $G(\hat{\theta}'|\hat{\theta}_t)$, we must consider the sampling statistics of VQAs. Due to quantum uncertainty, a measurement $m_i^r(\hat{\theta}_t)$ of operators $\omega_{i,ab}\sigma_a\sigma_b$ from equation (2) is a sample from a distribution with mean μ_t^i and variance

$$(\Delta_t^i)^2 = \omega_{i.ab}^2[\langle(\sigma_a\sigma_b)^2\rangle_t - \langle\sigma_a\sigma_b\rangle_t^2] = \omega_{i,ab}^2[1-(\mu_t^i)^2]. \tag{10}$$

Similarly, the two qubit terms from equations (4) and (5) result in $(\Delta_t^i)^2 = J^2[1-(\mu_t^i)^2]$, where $\mu_t^i = \langle\sigma_i^z\sigma_{i+1}^z\rangle_t$ and $\mu_t^i = \langle\sigma_i^z\sigma_j^z\rangle_t$, respectively. Likewise, for the single-qubit terms $(\Delta_t^i)^2 = g^2[1-(\mu_t^i)^2]$ where $\mu_t^i = \langle\sigma_i^x\rangle_t$. The Central Limit Theorem asserts that, assuming at least $M \approx 30$ independent and identically distributed measurements (shots) $m_i^r(\hat{\theta}_t)$, an estimate of the loss function Λ_t is the statistic $l_t \sim \mathcal{N}\left(\Lambda_t,\ (\Delta_t^\Lambda)^2\right)$, where $(\Delta_t^\Lambda)^2 = \sum_i(\Delta_t^i)^2/M$ [52, 53]. As precise expectation values usually require far more than 30 measurements (shots), this criterion is easily satisfied. Similarly, $\forall\theta_t^k \in \hat{\theta}_t$ and assuming small parameter shifts ε, the gradient $\nabla_k\Lambda_t = \left(\Lambda(\hat{\theta}_t+\epsilon\hat{k}) - \Lambda(\hat{\theta}_t-\epsilon\hat{k})\right)/2\epsilon$ is the statistic $d_kl_t \sim \mathcal{N}\left(\nabla_k\Lambda_t,\ [\Delta_\Lambda^2(\hat{\theta}_t+\epsilon\hat{k}) + \Delta_\Lambda^2(\hat{\theta}_t-\epsilon\hat{k})]/4\epsilon^2\right)$. The variance of this distribution can be simplified by noting that to first order in ε, the parameter shifted Pauli operators are $\sigma_a^{\pm k} = \sigma_a(\hat{\theta}\pm\epsilon\hat{k}) = \sigma_a \pm \iota_{ak}$, where $\sigma_a = \sigma_a(\hat{\theta})$ and $\iota_{ak} = (\partial\langle\sigma_a\rangle/\partial\theta^k)\epsilon$. We can then simplify the sum $\Delta_i(\hat{\theta}_t+\epsilon\hat{k})^2 + \Delta_i(\hat{\theta}_t-\epsilon\hat{k})^2 = 2\Delta_i(\hat{\theta}_t)^2$ by noting that

$$\Delta_i(\hat{\theta}_t+\epsilon\hat{k})^2 = \langle(\omega_{i,ab}\sigma_a^{\pm k}\sigma_b^{\pm k})^2\rangle - \langle\omega_{i,ab}\sigma_a^{\pm k}\sigma_b^{\pm k}\rangle^2, \tag{11a}$$

$$\langle(\sigma_a^{+k}\sigma_b^{+k})^2\rangle + \langle(\sigma_a^{-k}\sigma_b^{-k})^2\rangle = 2 + \mathcal{O}(\iota^2), \tag{11b}$$

$$\langle\sigma_a^{+k}\sigma_b^{+k}\rangle^2 + \langle\sigma_a^{-k}\sigma_b^{-k}\rangle^2 = 2\langle\sigma_a\sigma_b\rangle + \mathcal{O}(\iota^2). \tag{11c}$$

Now, up to first order in ι, we can derive the gradient's distribution

$$d_kl_t \sim \mathcal{N}\left(\nabla_k\Lambda_t,\ \Delta_\Lambda^2(\hat{\theta}_t)/2\epsilon^2\right). \tag{12}$$

Standard gradient descent would propose the candidate state $\hat{\theta}' = \hat{\theta} - \eta\nabla\Lambda_t$, however MCMC-VQA adds a normally distributed random noise term $\Theta_r \sim \mathcal{N}(0,1)$ with scale parameter ξ in order to expand the support of the proposal distribution $G(\hat{\theta}'|\hat{\theta}_t)$. This specifies

$$G\left(\hat{\theta}'|\hat{\theta}_t\right) = \prod_k G\left(\hat{\theta}'|\hat{\theta}_t\right)_k, \quad G\left(\hat{\theta}'|\hat{\theta}_t\right)_k = \mathrm{pdf}\left[\mathcal{N}\left(\eta\nabla_k\Lambda(\hat{\theta}_t),\ \xi^2 + \eta^2\frac{(\Delta_t^\Lambda)^2}{2\epsilon^2}\right)\right]\left(\hat{\theta}_t - \hat{\theta}'\right), \tag{13}$$

where the notation $\mathrm{pdf}\left[\mathcal{N}\left(\mu,\sigma^2\right)\right](x)$ denotes the probability density function at point x of a normal distribution with mean μ and variance σ^2. It follows that the acceptance distribution is given by

$$A\left(\hat{\theta}'|\hat{\theta}_t\right) = \min\left(1, \frac{P\left(\hat{\theta}'\right)G\left(\hat{\theta}_t|\hat{\theta}'\right)}{P\left(\hat{\theta}_t\right)G\left(\hat{\theta}'|\hat{\theta}_t\right)}\right). \tag{14}$$

We note that $G(\hat{\theta}_t|\hat{\theta}')$ is obtained by simply exchanging $\hat{\theta}_t$ and $\hat{\theta}'$ in equation (13). A random uniform sample $u \sim U(0,1)$ is then drawn for comparison, such that $\hat{\theta}_{t+1} = \hat{\theta}'$ if $A(\hat{\theta}'|\hat{\theta}_t) > u$ and $\hat{\theta}_{t+1} = \hat{\theta}_t$ otherwise.

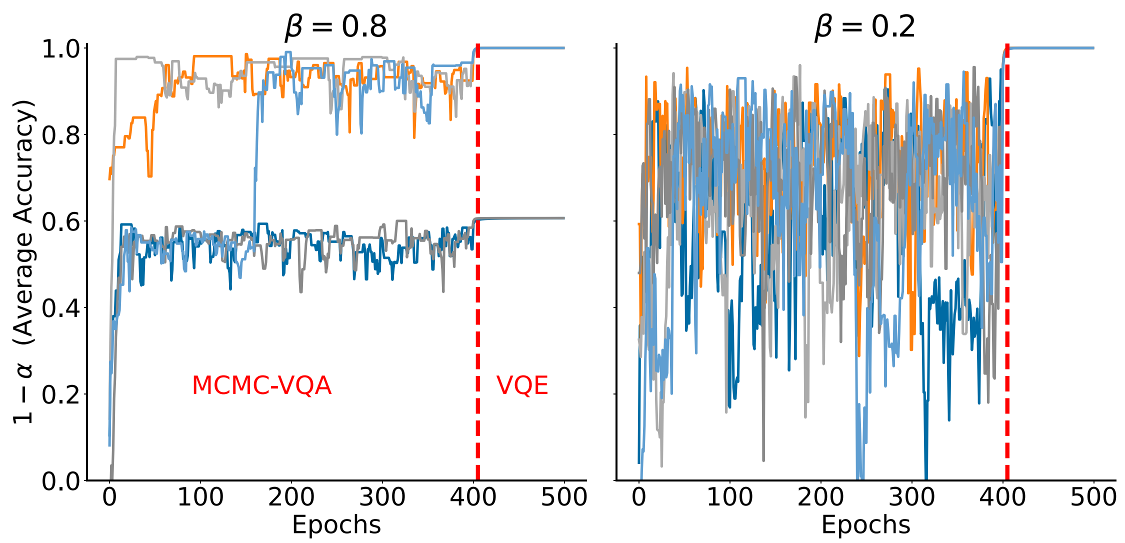

Figure 2. Example MaxCut trajectories with inverse thermodynamic temperature $\beta = 0.8$ (left) and $\beta = 0.2$ (right). Each color is a distinct trajectory representing a unique random initialization. The first four-hundred time steps are MCMC-VQA epochs (Markovian epochs), which are then followed by a closing sequence of VQE epochs (beginning at the vertical red dashed line). As discussed in section 2.1 and illustrated in figure 1, at the beginning of the closing VQE epochs, each trajectory is initialized with the best parameters $\hat{\theta}_{\min}$ found during the Markovian epochs. At lower temperature ($\beta = 0.8$), trajectories become trapped in local minima and such that reaching ergodicity is a lengthy process. Conversely, the high-temperature ($\beta = 0.2$) trajectories rapidly reach burn-in, generating $\hat{\theta}_{\min}$ that lead to near perfect convergence during the VQE closing sequence. As a deterministic algorithm initiated near an extremum, VQE converges rapidly, leveling-off at the bottom of the minimum within a few epochs (time steps). See section 3 for simulation details.

After T_{MC} epochs of the above Markovian process, MCMC-VQA implements a short series of traditional VQA epochs for rapid convergence to the nearest minimum. In particular, these closing VQA epochs are initialized with $\hat{\theta}_{\min}$, the parameter set of lowest expectation value $\Lambda_{\min}$ found during the Metropolis–Hastings phase. In this manner, MCMC-VQA can be considered a 'warm starting' procedure [18–20], but with ergodic guarantees.

Example MCMC-VQA trajectories are shown in figure 2 with inverse thermodynamic temperatures $\beta = 0.8$ and $\beta = 0.2$. The details of all simulations are given in section 3. Our algorithm combines the gradient descent-based optimization of VQE with a Markovian process that escapes local minima. Such exploration is significantly greater at the higher-temperature $\beta = 0.2$, where rather than settling into distinct loss function basins from which escape is relatively rare, the trajectories display the trademark 'burn-in' behavior of ergodic Markov chains. By the time that the closing VQE epochs are applied, the ergodic $\beta = 0.2$ MCMC-VQA chains have sampled states sufficiently near the global minimum and converge to the groundtruth nearly uniformly.

Figure 3(left) displays the average accuracy $1 - \alpha$ (where α is the average error, blue), and standard deviation (gray) of MaxCut solutions with MCMC-VQA as a function of β. Dashed lines represent the performance of traditional VQE on the same set of graphs and circuit ansatz. We note that all simulated β values outperform traditional VQE. Until $\beta \sim 0.2$, higher temperature MCMC-VQA chains have higher accuracy and better convergence, as their more permissive temperature parameter biases the acceptance distribution towards accepting the candidate states. However, performance decreases at very high temperatures, for which the MCMC-VQA chains are no longer appreciably biased towards energy minimization and the algorithm becomes more like random sampling than intrepid gradient descent. Likewise, the optimal amount of parameter update noise ξ is inversely proportional to β (figure 3, right), as higher temperatures permit more radical deviations from standard gradient descent.

Figure 4 demonstrates the effectiveness of MCMC-VQA on large-scale (50-qubit) quantum Ising (local, left blue) and transverse spin model (nonlocal, left dark gray) interactions. In both cases, MCMC-VQA outperforms VQE considerably, providing greater benefit in the ordered phases $|g| < |J|$, as well as local models, the later of which is likely due to the local ansatze of the quantum circuits in this manuscript. In all cases, the benefits of MCMC-VQA for larger systems ($n = 50$) are considerably greater than those for smaller systems ($n = 10$, left black), even when the same number of Markovian (sampling) epochs and sampled parameters are used (see section 3 for simulation details). The exploration of local minima by MCMC-VQA leads to the population of lower energy states of diverse Hamiltonians with higher probability than traditional initialization, including for large models and relatively few Markovian epochs (figure 4, right).

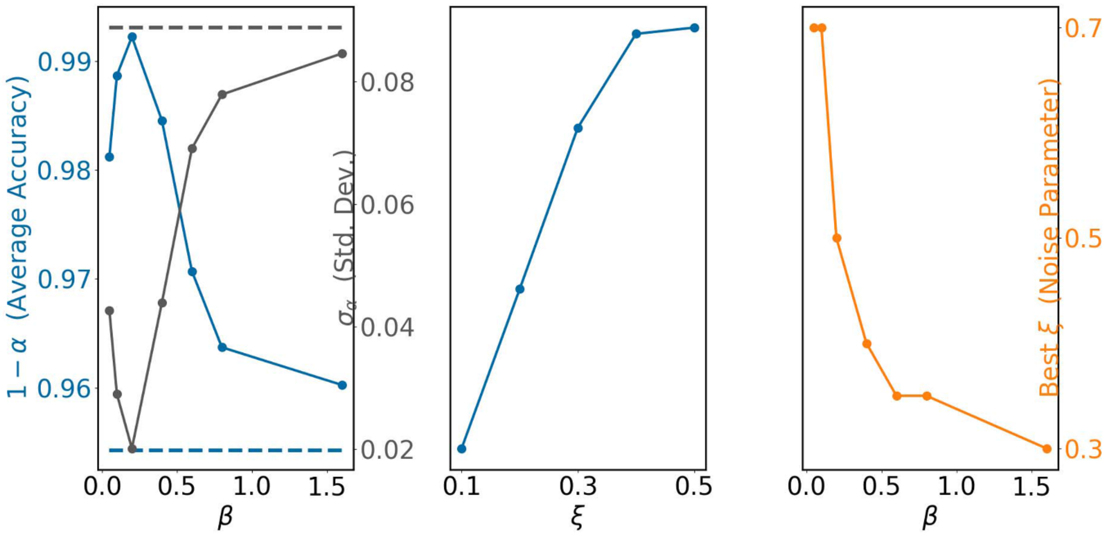

Figure 3. (Left, blue) Average MCMC-VQA accuracy ($1-\alpha$, for average error α) for MaxCut vs inverse thermodynamic temperature β. Nearly perfect average accuracy is obtained for properly tuned hyperparameter β (here, $\beta \approx 0.2$). At low temperature (large β), the algorithm mixes slowly, only partially approximating ergodicity in $T_{\mathrm{MC}}=400$ Markovian epochs. This partial convergence results in lower accuracy, which approaches that of traditional VQE (blue dashed line) in the limit of large β. Conversely, for high temperature (small β), the algorithm is insufficiently biased towards low-energy solutions, which renders its gradient descent inefficient and reduces its accuracy. (Center, blue) Average MCMC accuracy vs noise parameter ξ for inverse temperature $\beta=0.2$, where ξ is the gradient descent noise parameter ($\hat{\theta}'=\hat{\theta}-\eta\nabla\Lambda_t+\xi\Theta_r$) and each trajectory undergoes $T_{\mathrm{MC}}=400$ Markovian epochs. Larger noise parameters ξ generate more intrepid sampling and thus more rapid mixing, resulting in higher accuracies until the optimal value of ξ is surpassed (see right panel). (Left, gray) The standard deviation of MCMC-VQE accuracy vs β. Higher standard deviation directly corresponds with lower accuracy. As discussed above, at high β, this is due to runs trapped in local minima (see figure 2), while at low β, this stems from the lack of energy-preferred convergence. (Right) Optimal value of ξ vs β. As higher temperatures generate more permissive acceptance distributions $A(\hat{\theta}'|\hat{\theta})$, the optimal values of ξ also increase with temperature as they lead to more efficient mixing in the low-β limit. We note that for any finite β, too large of noise parameter ξ will result in the proposition of unlikely candidate states and thus inefficient state updates. See section 3 for simulation details.

For a sufficient number of Markovian epochs, the trajectory becomes ergodic, and convergence to the minimum becomes guaranteed (see section 2.3 and figure 5).

MCMC-VQA's ability to navigate local minima without increasing circuit depth makes it a useful alternative to deep-circuit ansatze, which are known to cause barren plateaus and noise-induced barren plateaus. In the case of barren plateaus, increasing circuit depth increases the concentration of measure, causing the gradient $\nabla_k\Lambda(\hat{\theta})$ and gradient variance $\mathrm{var}(\nabla_k\Lambda(\hat{\theta}))$ of any parameter θ_k to approach 0 and 2^{-n}, respectively [26]. Likewise, in the case of noise-induced barren plateaus, quantum noise effects obscure the gradient, causing $\nabla_k\Lambda(\hat{\theta})$ to shrink proportional to 2^{-L} for circuit depth L [25]. In addition to serving as an alternative to deeper circuits for local minima mitigation, MCMC-VQA provides parameter-update Langevin noise that has been proven to combat barren plateaus and demonstrated to increase circuit trainability [27]. Specifically, when training a circuit with barren plateaus using MCMC-VQA, the parameter update increments $(\hat{\theta}'-\hat{\theta})_k$ between the initial state $\hat{\theta}_k$ candidate state $\hat{\theta}'_k$ remain normally distributed with an unbiased mean $\nabla_k\Lambda(\hat{\theta})=0$, however they maintain a finite variance $\mathrm{var}(\nabla_k\Lambda(\hat{\theta}))=\xi^2$ rather than approaching 2^{-n}.

2.2. Implementation of MCMC-VQA on quantum hardware

As discussed above, the loss function Λ_t is not precisely determined on actual quantum hardware, but rather estimated as a statistic $l_t=\sum_i q_t^i$, where $q_t^i=\frac{1}{M}\sum_{r=1}^{M} m_i^r(\hat{\theta}_t)$. As a result, the variance of a single observable measurement $(\Delta_t^i)^2$ is estimated by $(\delta_t^i)^2=\omega_{i,ab}^2[1-(q_t^i)^2]$, while that of the total loss function $(\Delta_t^\Lambda)^2$ is estimated by $(\delta_t^\Lambda)^2=\sum_i(\delta_t^i)^2/M=\sum_i\omega_{i,ab}^2[1-(q_t^i)^2]/M$, for M-measurements per observable. Alternatively, the variances could be directly estimated from the standard deviations of expectation value statistics. We then define $a(\hat{\theta}'|\hat{\theta}_t)$, the acceptance distribution on quantum hardware, as

$$a(\hat{\theta}'|\hat{\theta}_t)=\min\left(1,\frac{p(\hat{\theta}')g(\hat{\theta}_t|\hat{\theta}')}{p(\hat{\theta}_t)g(\hat{\theta}'|\hat{\theta}_t)}\right), \tag{15a}$$

$$p(\hat{\theta})\propto\exp(-\beta l_t), \tag{15b}$$

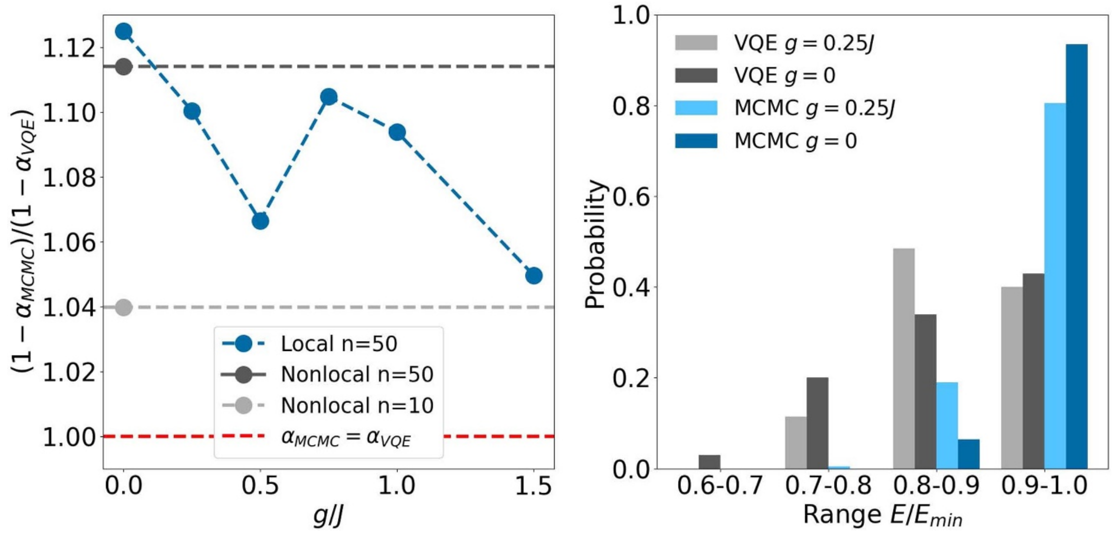

Figure 4. (Left, blue) The relative average accuracy between MCMC-VQA and VQE energy minimization for large-scale ($n=50$) local, ferromagnetic Quantum Ising models vs the magnitude of transverse field g. At $g=0$, the quantum Ising model reduces to its classical counterpart, which has identical Hamiltonian structure the implementation of MaxCut on a variational quantum computer. While MCMC-VQA provides marked improvement for both ordered ($|g|<|J|$) and disordered ($|g|>|J|$) phases, it is particularly advantageous for navigating the single-axis dominated minima of the former. The $g=0$ (MaxCut) energy minimization for nonlocal spin models with both $n=10$ (light gray) and $n=50$ (dark gray) qubits illustrates the feasibility of MCMC-VQA for large system sizes, with $n=50$ performing significantly better than $n=10$. All systems benefit from MCMC-VQA, such that all performance ratios are greater than the red dashed at line at $\alpha_{MCMC}=\alpha_{VQE}$. (Right) Energy histogram for the performance of VQE and MCMC-VQA on nonlocal spin models of $n=50$ qubits with $g=0$ and $g=0.25J$. MCMC-VQA's exploration of local minima reduces the proportion of trajectories that settle in low-lying states, shifting the ensemble of trajectories towards lower energy states. See section 3 for simulation details.

$$g(\hat{\theta}'|\hat{\theta}_t)=\prod_k g(\hat{\theta}'|\hat{\theta}_t)_k, \tag{15c}$$

$$g(\hat{\theta}'|\hat{\theta}_t)_k=\mathrm{pdf}\left[\mathcal{N}\left(\eta d_k l_t,\ \xi^2+\eta^2\frac{(\delta_t^\Lambda)^2}{2\epsilon^2}\right)\right]\left(\hat{\theta}_t-\hat{\theta}'\right). \tag{15d}$$

MCMC-VQA does not increase the quantum complexity of VQAs (number of operations carried out on quantum hardware), as the expectation values that comprise $\Lambda(\hat{\theta})$ are calculated in the same fashion as they would be for the unmodified quantum variational algorithm and the analysis of MCMC-VQA is designed to as to be general for limited precision (e.g. shot noise). Moreover, the additional classical overhead of MCMC-VQA is minimal, as the acceptance probability and its components are computed classically with simple arithmetic, probability density function calculations, and pseudo-random number sampling.

2.3. Proof of ergodicity

If a Metropolis–Hastings algorithm is *irreducible* and *aperiodic*, then the resulting Markov chain is provably ergodic [49]. That is, it will explore all areas of the probability distribution, converging on average to the Markov process' unique stationary distribution, which includes the global minimum of the solution space. Moreover, as we have chosen to sample from the Boltzmann distribution of the loss function, we sample from states near optimal solutions with exponentially higher probability. In what follows, we demonstrate the ergodicity of our method. We note that the resultant Markov chain is ergodic with respect to the *accessible* distribution, such that the true ground state can only be obtained if the quantum circuit ansatz is sufficiently expressive (i.e. capable of preparing the ground state).

2.3.1. Irreducibility

The VQA Metropolis–Hastings Markov chain is irreducible if $\forall\hat{\theta}_a,\hat{\theta}_b\in[0,2\pi]$, $\exists T,\{\hat{\theta}_1,\hat{\theta}_2,\ldots,\hat{\theta}_T\}$ such that

$$p(\hat{\theta}_1|\hat{\theta}_a)p(\hat{\theta}_b|\hat{\theta}_T)\prod_{i=1}^{T-1}p(\hat{\theta}_{i+1}|\hat{\theta}_i)>0. \tag{16}$$

That is, the Markov chain is irreducible if, for any two points in parameter space $\hat{\theta}_a,\hat{\theta}_b$, there exists a series of transitions of any length T such that $\hat{\theta}_a\rightarrow\hat{\theta}_b$ with non-zero probability [54]. While this definition of irreducibility is sufficient, we will instead focus on the yet more powerful condition of *strong* irreducibility. A Markov chain is strongly irreducible if

$$g(\hat{\theta}_a|\hat{\theta}_b) > 0, \forall\hat{\theta}_a, \hat{\theta}_b, \tag{17}$$

meaning that all points in parameter space have a non-zero probability of transitioning to all other points [55]. This condition is then equivalent to

$$g(\hat{\theta}_b|\hat{\theta}_a)_k = \frac{(2\pi)^{-1/2}}{\sqrt{\xi^2+\eta^2(\delta_a^\Lambda)^2/2\epsilon^2}} \exp\left[\frac{-\left(\theta_a^k-\theta_b^k-\eta d_k l_a\right)^2}{2\left(\xi^2+\eta^2(\delta_a^\Lambda)^2/2\epsilon^2\right)}\right] > 0,\ \forall k, \tag{18}$$

where we note that $\delta_\Lambda^2(\hat{\theta}_t) \propto 1/M$.

Equation (17) is satisfied, at least technically to some tolerance, $\forall\hat{\theta}_a, \hat{\theta}_b$. Although $g(\hat{\theta}_b|\hat{\theta}_a)_k$ may become very small, it will generally retain a non-zero probability for virtually all transitions, and the chain will be strongly irreducible, albeit perhaps slow to convergence. More precise arguments can be made in the limit of large ξ, where to first order in small $1/\xi$, $g(\hat{\theta}_b|\hat{\theta}_a)_k \to 1/\sqrt{2\pi}\xi$ and all transitions become equally likely. While this extreme ξ limit is too random to result in efficient gradient descent, it illustrates a concrete transition to irreducibility with increasing ξ. Moreover, due to the uncertainty introduced by finite statistics $d_k l_a$ and $(\delta_a^\Lambda)^2$, sampling of the proposition kernel $g(\hat{\theta}_b|\hat{\theta}_a)_k$ can allow for otherwise unlikely transitions.

2.3.2. Aperiodicity

In the case of strong irreducibility argued above (equation (17)), aperiodicity is automatically satisfied. Assuming only the weaker irreducibility of equation (16), it is sufficient to show that [54]

$$a(\hat{\theta}_a|\hat{\theta}_a)g(\hat{\theta}_a|\hat{\theta}_a) = g(\hat{\theta}_a|\hat{\theta}_a) = \frac{(2\pi)^{-1/2}}{\sqrt{\xi^2+\eta^2(\delta_a^\Lambda)^2/2\epsilon^2}} \exp\left[\frac{-\left(\eta d_k l_a\right)^2}{2\left(\xi^2+\eta^2(\delta_a^\Lambda)^2/2\epsilon^2\right)}\right] > 0. \tag{19}$$

As long as $\eta \ggg \xi$, equation (19) holds for all but singular points $\hat{\theta}_a$.

2.4. Mixing time

The mixing time τ of a Markov chain is the number of epochs required to reach a certain threshold of convergence. For an ergodic, discrete-time Markov chain, τ is analytically bounded by

$$\tau \leqslant \frac{2}{\Phi^2}\ln\left(\frac{1}{\alpha_{\mathrm{MC}}\sqrt{\pi^*}}\right), \tag{20}$$

where $\alpha_{\mathrm{MC}} = |S-\pi|$ is the distance between the Markov chain's sampled distribution S and the true stationary distribution π, π^* is the probability of the least likely (maximum energy) state of π, and Φ is the conductance or 'Cheeger constant' of the Markov process [50]. The conductance can be understood as the minimum of normalized ergodic flows between all possible partitions of the state space.

Figure 5 demonstrates that the performance of MCMC-VQA is consistent with the theoretical predictions of ergodic Markov chains (equation (20)). That is, the time dependence of MCMC-VQA optimization error α follows the same $\ln(1/\alpha)$ scaling as the distribution distance α_{MC} in equation (20). Moreover, least-squares analysis of figure 5 data reveals a β-dependent scale factor that is proportional to $\ln(1/\sqrt{\pi^*})$, which is consistent with the Boltzmann distribution $p(\hat{\theta}_a) \propto \exp(-\beta\Lambda_a)$ from which our method samples. This temperature-dependent time-complexity further verifies that MCMC-VQA is an ergodic Markov process that successfully samples from the target distribution.

3. Numerical simulations

The simulations in this work are done using a modified version of TensorLy-Quantum, an open-source software package for quantum circuit simulation using factorized tensors [56, 57]. TensorLy-Quantum specializes in exact tensor contraction, such that the simulations are carried out without truncation or approximation.

The MaxCut instances optimized in this work are generated from ten graphs. Each graph has ten vertices and an equal number of randomly selected edges, which are randomly generated from the unit normal distribution. Such graphs are equivalent to the Gilbert model of random graphs [58]. The number of edges was chosen to be equal to that of vertices as this ratio is observed to pose high difficulty for random MaxCut problems of this model [59, 60]. Likewise, the number of two-qubit interactions in the quantum Ising and transverse field spin models was chosen to be equal to the number of qubits ($n = 50$), in accordance with the

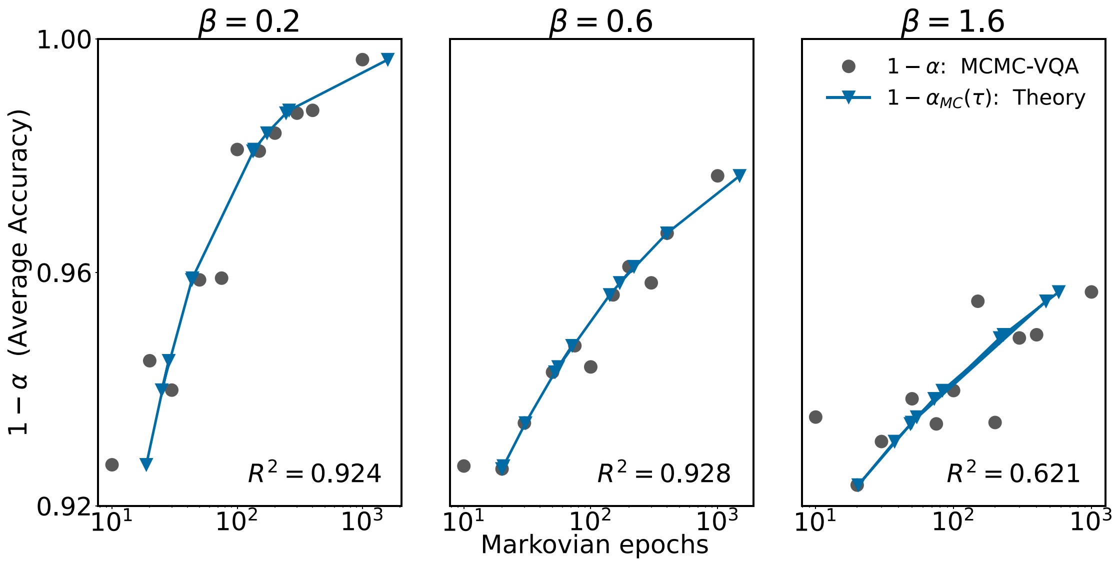

Figure 5. Average accuracy vs Markovian epochs for three different β values. Gray dots are the average MCMC-VQA accuracy $1-\alpha$, and blue curves are a least squares fit of this data to the analytical accuracy of an ergodic Markov chain $1-\alpha_{\text{MC}}(\tau)$, with theoretical mixing time τ (see equation (20)). The analytical time-dependence of α_{MC} matches the observed scaling of α ($R^2 \sim 1$) for low values of inverse temperature $\beta = 0.2, 0.6$, affirming that MCMC-VQA is an ergodic Markov chain at sufficiently high temperatures, and thus guaranteeing convergence to the global minimum. Furthermore, the ratio of observed scale parameters between MCMC-VQA simulations with different β values is consistent with the analytic dependence $\tau \propto \ln(1/\sqrt{\pi^*})$ (equation (20)) on the least likely state $\pi^* \propto \exp(-\beta\Lambda_{\max})$ (equation (9)). This functional dependence on temperature further supports our claims of ergodically sampling from $P(\hat{\theta})$ and thus deterministically converging to the global minimum.

definition of the former and in order to provide a more consistent comparison between the models. While the quantum Ising model topology is uniquely defined, five random transverse field models were studied.

All numerical simulations in this work are done using the Hamiltonians described above, with twenty randomly initialized runs completed for each graph. The quantum circuits for MaxCut use one parameterized rotation per vertex. The quantum circuits for the quantum Ising and transverse field models use two parameterized rotations per qubit, however only 10 parameters undergo Markovian update per Markovian epoch such that the same amount of computational overhead is used in the MCMC optimization of the $n = 50$ and the $n = 10$ graphs alike and the favorable scaling of MCMC-VQA is demonstrated. We illustrate our work using circuits with relatively few parameters, because their optimization landscape is especially nonconvex and thus prone to convergence in local minima [17], however MCMC-VQA can be used with arbitrary parameterization. The circuit gates are alternated between a layer of single-qubit parameterized rotations (angles $\hat{\theta}$) about the y-axis and a layer of two-qubit control-Z gates. For each method (VQE or MCMC-VQA) and set of hyperparameters, a variety of learning rates are scanned so that numerical comparisons could be drawn against the optimal performance of each algorithm. All VQE sequences for MaxCut consisted of 100 epochs of vanilla gradient descent. This quantum Ising and transverse field spin models were optimized over 400 Markovian epochs with a 200 epoch VQE closing sequence in the MCMC-VQA case, or 600 VQE epochs for the VQE comparison. MCMC-VQA epochs (time steps) with rejected samples were still counted as a completed epoch and did not result in an extra step. Figure 2 shows an ensemble of trajectories whereas figures 3–5 are the average over the optimal learning rate for all MaxCut graphs with 20 random initializations each and all spin model graphs with 40 initializations each. For simplicity, we take the large M limit, assuming many measurements and precise expectation values.

4. Discussion

In this work, we have introduced MCMC-VQA: a novel variational quantum algorithm that harnesses classical Makov chains to obtain analytic convergence guarantees for parameterized quantum circuits. As ergodic Markov chains representatively sample a target probability distribution, they identify regions near the global minimum with high probability. We present MCMC-VQA, both from a theoretical and practical perspective, prove its ergodicity, and derive its time-complexity (mixing time) as a function of both accuracy and inverse thermodynamic temperature. Focusing on MaxCut optimization within the VQE framework due to its plentiful local minima and on the formation of low-energy states for the quantum Ising and transverse field spin model Hamiltonians due to their relevance to quantum chemistry [11–13] and condensed matter physics [14–16], we employ a reversible Metropolis–Hastings Markov process suitable for variational quantum circuits. We demonstrate the ergodicity of our method and the validity of our analytical findings, ultimately observing the capacity of MCMC-VQA to not only outperform traditional VQAs, but to do so with up to perfect and deterministic convergence.

In future research, MCMC-VQA should be studied for an even wider variety of different applications, quantum algorithms, and Markov processes. This manuscript's study of canonical quantum models could be furthered by study of more intricate quantum systems, such as the identification of molecular groundstates [11–13]. In such applications, exploration of the loss function landscape is of the upmost importance, as even simple quantum Hamiltonians, such as the transverse field Ising model, are known to acutely struggle with premature convergence to local, rather than global, minima. Similarly, our technique could be extended to QAOA [3] or any of the numerous VQAs that have been proposed in recent years. Finally, tens of MCMCs have been devised over the past 70 years, each with their own advantages, with variations featuring Gibbs sampling [61], parallel tempering [62], and independence sampling [63]. These methods could be substituted for Metropolis–Hastings in order to produce algorithms with lower computational overhead and faster mixing times. In short, varieties of MCMC-VQA can be developed for a broad spectrum of variational quantum algorithms to both improve and guarantee performance.

Data availability statement

The data that support the findings of this study are available upon reasonable request from the authors.

Acknowledgments

O S likes to thank Katie Pizzolato for accommodating HPC resource requests on IBM Cloud. This work was done during T L P's internship at IBM Quantum, for which T L P thanks Katie Pizzolato and the entire IBM Quantum team. T L P also thanks NVIDIA for the use of GPU computational resources. S F Y would like to acknowledge funding by NSF and AFOSR.

ORCID iDs

Taylor L Patti https://orcid.org/0000-0002-4242-6072
Susanne F Yelin https://orcid.org/0000-0003-1655-9151

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] Peruzzo A, McClean J, Shadbolt P, Yung M-H, Zhou X-Q, Love P J, Aspuru-Guzik A and O'brien J L 2014 A variational eigenvalue solver on a photonic quantum processor *Nat. Commun.* **5** 4213

[3] Farhi E, Goldstone J and Gutmann S 2014 A quantum approximate optimization algorithm (arXiv:1411.4028)

[4] Cerezo M *et al* 2021 Variational quantum algorithms *Nat. Rev. Phys.* **3** 625–44

[5] Lavrijsen W, Tudor A, Müller J, Iancu C and de Jong W 2020 Classical optimizers for noisy intermediate-scale quantum devices *2020 IEEE Int. Conf. on Quantum Computing and Engineering (QCE)* (IEEE) pp 267–77

[6] Garey M R and Johnson D S 2002 *Computers and Intractability* vol 29 (New York: W H Freeman)

[7] Nannicini G 2019 Performance of hybrid quantum-classical variational heuristics for combinatorial optimization *Phys. Rev.* E **99** 013304

[8] Braine L, Egger D J, Glick J and Woerner S 2021 Quantum algorithms for mixed binary optimization applied to transaction settlement *IEEE Trans. Quantum Eng.* **2** 1–8

[9] Patti T L, Kossaifi J, Anandkumar A and Yelin S F 2021 Variational quantum optimization with multi-basis encodings (arXiv:2106.13304)

[10] Fuller B, Hadfield C, Glick J R, Imamichi T, Itoko T, Thompson R J, Jiao Y, Kagele M M, Blom-Schieber A W, Raymond R and Mezzacapo A 2021 Approximate solutions of combinatorial problems via quantum relaxations (arXiv:2111.03167)

[11] McArdle S, Endo S, Aspuru-Guzik A, Benjamin S C and Yuan X 2020 Quantum computational chemistry *Rev. Mod. Phys.* **92** 015003

[12] Kandala A, Mezzacapo A, Temme K, Takita M, Brink M, Chow J M and Gambetta J M 2017 Hardware-efficient variational quantum eigensolver for small molecules and quantum magnets *Nature* **549** 242

[13] Grimsley H R, Economou S E, Barnes E and Mayhall N J 2019 An adaptive variational algorithm for exact molecular simulations on a quantum computer *Nat. Commun.* **10** 3007

[14] Ritter M B 2019 Near-term quantum algorithms for quantum many-body systems *J. Phys.: Conf. Ser.* **1290** 012003

[15] Vogt N, Zanker S, Reiner J-M, Eckl T, Marusczyk A and Marthaler M 2020 Preparing symmetry broken ground states with variational quantum algorithms (arXiv:2007.01582)

[16] Zhang F, Gomes N, Yao Y, Orth P P and Iadecola T 2021 Adaptive variational quantum eigensolvers for highly excited states *Phys. Rev.* B **104** 075159

[17] Lee J, Magann A B, Rabitz H A and Arenz C 2021 Progress toward favorable landscapes in quantum combinatorial optimization *Phys. Rev.* A **104** 032401

[18] Beaulieu D and Pham A 2021 Max-cut clustering utilizing warm-start QAOA and IBM runtime (arXiv:2108.13464)
[19] Egger D J, Mareček J and Woerner S 2021 Warm-starting quantum optimization *Quantum* **5** 479
[20] van Dam W, Eldefrawy K, Genise N and Parham N 2021 Quantum optimization heuristics with an application to knapsack problems (arXiv:2108.08805)
[21] Rivera-Dean J, Huembeli P, Acín A and Bowles J 2021 Avoiding local minima in variational quantum algorithms with neural networks (arXiv:2104.02955)
[22] Harwood S M, Trenev D, Stober S T, Barkoutsos P, Gujarati T P, Mostame S and Greenberg D 2021 Improving the variational quantum eigensolver using variational adiabatic quantum computing (arXiv:2102.02875)
[23] Shehab O, Kim I H, Nguyen N H, Landsman K, Alderete C H, Zhu D, Monroe C and Linke N M 2019 Noise reduction using past causal cones in variational quantum algorithms (arXiv:1906.00476)
[24] Bravyi S, Gosset D and König R 2018 Quantum advantage with shallow circuits *Science* **362** 308–11
[25] Wang S, Fontana E, Cerezo M, Sharma K, Sone A, Cincio L and Coles P J 2021 Noise-induced barren plateaus in variational quantum algorithms *Nat. Commun.* **12** 1–11
[26] 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** 1–6
[27] Patti T L, Najafi K, Gao X and Yelin S F 2021 Entanglement devised barren plateau mitigation *Phys. Rev. Res.* **3** 033090
[28] Ortiz Marrero C, Kieferová M and Wiebe N 2021 Entanglement-induced barren plateaus *PRX Quantum* **2** 040316
[29] Holmes Z, Sharma K, Cerezo M and Coles P J 2021 Connecting ansatz expressibility to gradient magnitudes and barren plateaus (arXiv:2101.02138)
[30] 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
[31] Gibbs J W 1902 *Elementary Principles in Statistical Mechanics, Developed with Especial Reference to the Rational Foundation of Thermodynamics* (New York: Charles Scribner's Sons)
[32] Metropolis N, Rosenbluth A W, Rosenbluth M N, Teller A H and Teller E 1998 Equation of state calculations by fast computing machines *J. Chem. Phys.* **21** 1087
[33] Robbins H and Monro S 1951 A stochastic approximation method *Ann. Math. Stat.* **22** 400–7
[34] Nemeth C and Fearnhead P 2021 Stochastic gradient Markov chain Monte Carlo *J. Am. Stat. Assoc.* **116** 433–50
[35] Szegedy M 2004 Quantum speed-up of Markov chain based algorithms *Proc. 45th Annual IEEE Symp. on Foundations of Computer Science (FOCS '04)* (IEEE Computer Society) pp 32–41
[36] Temme K, Osborne T J, Vollbrecht K G, Poulin D and Verstraete F 2011 Quantum metropolis sampling *Nature* **471** 87–90
[37] Lemieux J, Heim B, Poulin D, Svore K and Troyer M 2020 Efficient quantum walk circuits for Metropolis–Hastings algorithm *Quantum* **4** 287
[38] Montanaro A 2016 Quantum speedup of Monte Carlo methods *Proc. R. Soc.* A **471** 20150301
[39] Cornelissen A and Jerbi S 2021 Quantum algorithms for multivariate Monte Carlo estimation (arXiv:2107.03410)
[40] Wang Y, Wu S and Zou J 2016 Quantum annealing with Markov chain Monte Carlo simulations and D-wave quantum computers *Stat. Sci.* **31** 362–98
[41] Medvidovic M and Carleo G 2021 Classical variational simulation of the quantum approximate optimization algorithm *npj Quantum Inf.* **7** 101
[42] Chowdhury A N, Low G H and Wiebe N 2020 A variational quantum algorithm for preparing quantum Gibbs states (arXiv:2002.00055)
[43] Wang Y, Li G and Wang X 2021 Variational quantum Gibbs state preparation with a truncated taylor series *Phys. Rev. Appl.* **16** 054035
[44] Shtanko O and Movassagh R 2021 Algorithms for Gibbs state preparation on noiseless and noisy random quantum circuits (arXiv:2112.14688)
[45] Warren A, Zhu L, Mayhall N J, Barnes E and Economou S E 2022 Adaptive variational algorithms for quantum Gibbs state preparation (arXiv:2203.12757)
[46] Commander C W 2009 Maximum cut problem, MAX-CUT *Encyclopedia of Optimization* (Boston, MA: Springer) pp 1991–9
[47] Strecka J and Jascur M 2015 A brief account of the Ising and Ising-like models: mean-field, effective-field and exact results (arXiv:1511.03031)
[48] Geyer C J 1992 Practical Markov chain Monte Carlo *Stat. Sci.* **7** 473–83
[49] Brooks S P 1998 Markov chain Monte Carlo method and its application *J. R. Stat. Soc.* D **47** 69–100
[50] Montenegro R and Tetali P 2006 Mathematical aspects of mixing times in Markov chains *Found. Trends Theor. Comput. Sci.* **1** 237–354
[51] March N M 2011 The eigenvalue gap and mixing time (available at: https://math.dartmouth.edu/~pw/M100W11/nathan.pdf)
[52] Kim T K 2015 T test as a parametric statistic *Korean J. Anesthesiol.* **68** 540
[53] Kwak S G and Kim J H 2017 Central limit theorem: the cornerstone of modern statistics *Korean J. Anesthesiol.* **70** 144
[54] Daskalakis C 2011 6.896: probability and computation (available at: https://people.csail.mit.edu/costis/6896sp11/)
[55] Whiteley N 2008 The metropolis-hastings algorithm (available at: www.webpages.uidaho.edu/~stevel/565/U.%20Bristol/folien6.pdf)
[56] Patti T L, Kossaifi J, Yelin S F and Anandkumar A 2021 Tensorly-quantum: quantum machine learning with tensor methods (arXiv: 2112.10239)
[57] Tensorly-Quantum 2021 Tensor-based quantum machine learning (available at: http://tensorly.org/quantum/dev/)
[58] Gilbert E N 1959 Random graphs *Ann. Math. Stat.* **30** 1141–4
[59] Coppersmith D, Gamarnik D, Hajiaghayi M and Sorkin G B 2004 Random max sat, random max cut and their phase transitions *Random Struct. Algorithms* **24** 502–45
[60] Luczak T 1990 On the equivalence of two basic models of random graph *Proc. Random Graphs* vol 87 pp 151–9
[61] Gelfand A E 2000 Gibbs sampling *J. Am. Stat. Assoc.* **95** 1300–4
[62] Earl D J and Deem M W 2005 Parallel tempering: theory, applications and new perspectives *Phys. Chem. Chem. Phys.* **7** 3910–6
[63] Hastings W K 1970 Monte Carlo sampling methods using Markov chains and their applications *Biometrika* **57** 97–109