

Improved energies and local energies with weighted variational Monte Carlo

Huan Zhang¹, Robert J. Webber², Michael Lindsey³, Timothy C. Berkelbach^{4,5,*} and Jonathan Weare^{1,†}¹*Courant Institute of Mathematical Sciences, New York University, New York 10012, USA*²*Department of Mathematics, University of California San Diego, La Jolla, California 92093, USA*³*Department of Mathematics, University of California Berkeley, California 94720, USA*⁴*Department of Chemistry, Columbia University, New York, New York 10027, USA*⁵*Initiative for Computational Catalysis, Flatiron Institute, New York, New York 10010, USA*

(Received 16 June 2025; accepted 17 December 2025; published 25 February 2026)

Neural network parametrizations have increasingly been used to represent the ground and excited states in variational Monte Carlo (VMC) with promising results. However, traditional VMC methods only optimize the wave function in regions of peak probability. The wave function is uncontrolled in the tails of the probability distribution, which can limit the accuracy of the trained wave function. To improve the approximation accuracy in the probability tails, this paper interprets VMC as a gradient flow in the space of wave functions, followed by a projection step. From this perspective, arbitrary probability distributions can be used in the projection step, allowing the user to prioritize accuracy in different regions of state space. Motivated by this theoretical perspective, the paper tests a weighted VMC method on the antiferromagnetic Heisenberg model for a periodic spin chain. Compared to traditional VMC, weighted VMC reduces the error in the ground state energy by a factor of 2, and it reduces the errors in the local energies away from the mode by large factors of 10^2 – 10^4 .

DOI: [10.1103/6b3v-38jp](https://doi.org/10.1103/6b3v-38jp)

I. INTRODUCTION

The introduction of neural network wave function parametrizations [1–4] has sparked renewed interest in the variational Monte Carlo (VMC) [5–13] approach to approximating quantum many-body ground and excited states. In its simplest version, VMC approximates the ground state of a given system by minimizing an energy functional $E(\psi_\theta)$ over a parametrized class of wave functions ψ_θ . The approach uses Monte Carlo sampling to estimate the parameter gradient $\nabla_\theta E(\psi_\theta)$ needed for energy minimization.

Neural quantum states have shown tremendous promise within VMC, but they also lead to a problem of generalization error [14]. The combination of a nonlinear wave function representation with Monte Carlo approximation can lead to low-accuracy wave function estimates away from regions of high probability, and it can even lead to unstable optimization [15,16].

This paper focuses on computing the ground state of the antiferromagnetic Heisenberg XXX model, which has proved computationally challenging for VMC [16,17] even though the ground state can be derived analytically using the Bethe ansatz [18]. Our previous work [16] documents the

generalization error in the estimated ground state for the XXX model. This error is concentrated in the region of ferromagnetic configurations far away from the mode of the wave function. To help reduce this error, we proposed [16] adding a term to the VMC objective function that prevents the wave function from growing in the ferromagnetic region. However, this approach does not ensure relative accuracy in the approximation of the ground state in the region of low $|\psi_\theta|^2$.

The current paper presents an approach to improve the wave function accuracy based on understanding VMC optimization as a gradient flow followed by a projection step [19,20]. In this theoretical framework, the wave function is first optimized using a gradient flow that reduces the energy. Then, the evolved wave function is projected onto the space of parameterized wave functions. The evolve-and-project approach is at the core of the time-dependent variational principle commonly used to optimize wave function approximations [21,22].

The projection step depends on the choice of distance function, since it identifies the element within the parametric class with the *smallest distance* to the evolved wave function. This paper introduces the idea of a weighted distance that enhances the projection accuracy in low-probability regions. The weighted distance leads to a modified formula for the VMC parameter updates, which can be evaluated using Monte Carlo samples drawn from a weighted wave function density. Conversely, the user is free to generate samples from any convenient distribution, and the parameter update calculated with the samples corresponds to a projection step with an associated set of weights. This “weighted VMC” approach does not require prior knowledge of the problem and should be widely applicable.

*Contact author: t.berkelbach@columbia.edu†Contact author: weare@nyu.edu

Our numerical experiments indicate the success of weighted VMC applied to the Heisenberg XXX model. The experiments identify several sampling methods leading to improved energies and local energies throughout the configuration space. Mixed tempering and well-tempered metadynamics [23,24] are broadly applicable methods that yield a factor of 2 energy improvement compared to traditional VMC. Additionally, uniform collective variable (CV) sampling is based on identifying a physically important collective variable. It yields competitive energy estimates, and it improves the local energies in targeted regions by factors up to 10^4 .

The paper is structured as follows. Section II provides background. Section III introduces weighted VMC. Section IV shows examples of weighted VMC sampling distributions. Section V presents numerical results. Section VI concludes.

II. BACKGROUND

This section first introduces the XXX model and the restricted Boltzmann machine (RBM) wave function ansatz. Then, it describes the standard VMC optimization using stochastic reconfiguration.

A. Model and ansatz

This paper focuses on the antiferromagnetic Heisenberg XXX model for N spin-1/2 particles in a one-dimensional periodic chain. After a stoquastic transformation [25], the Hamiltonian for the XXX model is

$$\hat{H} = \sum_{i=1}^N (-\hat{\sigma}_i^x \hat{\sigma}_{i+1}^x - \hat{\sigma}_i^y \hat{\sigma}_{i+1}^y + \hat{\sigma}_i^z \hat{\sigma}_{i+1}^z), \quad (1)$$

where $\hat{\sigma}_i^x$, $\hat{\sigma}_i^y$, and $\hat{\sigma}_i^z$ are the Pauli operators for the i th spin. Here and throughout, periodic boundary conditions are implied via the identification $\hat{\sigma}_{N+i} = \hat{\sigma}_i$. The Hamiltonian admits a unique, nonnegative-valued ground state wave function due to the Perron-Frobenius theorem.

The Hamiltonian for the XXX model commutes with the translation operator, so each ground and excited state ψ must satisfy a symmetry relationship

$$\psi(T_j \hat{\sigma}) = e^{i2\pi k j/N} \psi(\sigma),$$

where T_j is the translation operator, $(T_j \hat{\sigma})_i = \hat{\sigma}_{i-j}$, and $k \in \{0, 1, \dots, N-1\}$ is the symmetry class. Hereafter, we focus on the $k=0$ subspace, which contains the ground state due to the Perron-Frobenius theorem.

We define the variational wave function using a basis of simultaneous eigenstates of the operators $\{\hat{\sigma}_i\} = \{\hat{\sigma}_i^z\}$, where we drop the z superscript for notational simplicity. We use an RBM [26] ansatz with translational symmetry, which is a one-hidden-layer neural network that has been widely used in VMC and feature learning algorithms [1,15,16,27,28]. See the supplement of Ref. [1] for a detailed derivation. The ansatz can be written as

$$\psi_{\theta}(\sigma) = \prod_{k=1}^{\alpha} \prod_{j=1}^N \cosh \left(b_k + \sum_{i=1}^N W_{ki} \sigma_{i+j} \right), \quad (2)$$

where $\sigma_i \in \{-1, +1\}$ are spin variables and $\theta = \{\mathbf{b}, \mathbf{W}\}$ are variational parameters. We consider both $N=100$ and $N=200$ spin models in our numerical experiments below.

B. Variational Monte Carlo

In this section, we adopt the standard L^2 inner product and norm:

$$\langle \psi, \phi \rangle = \sum_{\sigma} \overline{\psi(\sigma)} \phi(\sigma), \quad \|\psi\|^2 = \sum_{\sigma} |\psi(\sigma)|^2.$$

VMC approximates the ground state by minimizing the Rayleigh quotient

$$E(\psi) := \frac{\langle \psi, H \psi \rangle}{\langle \psi, \psi \rangle} \quad (3)$$

within a parametrized manifold $\psi = \psi_{\theta}$. Stochastic reconfiguration (SR) [6] is a popular method to optimize traditional [29,30] and more recent [1,3] wave functions within VMC. SR is motivated by considering a small parameter update $\theta \leftarrow \theta + \delta$. The parameter update changes the energy in a predictable way, according to the following linearized equation:

$$E_{\text{linear}}(\psi_{\theta+\delta}) = E(\psi_{\theta}) + \delta^* \mathbf{g} + \mathbf{g}^* \delta. \quad (4)$$

Here, $\mathbf{g}$ is the Wirtinger gradient [31] of the energy, defined by

$$\mathbf{g} = \frac{\langle \nabla_{\theta} \psi_{\theta}, [H - E(\psi)] \psi_{\theta} \rangle}{\langle \psi_{\theta}, \psi_{\theta} \rangle}. \quad (5)$$

Each step of SR minimizes the linearized energy (4) subject to quadratic penalty terms. The first quadratic penalty is based on the square angle between wave functions [15]:

$$\delta^* \mathbf{S} \delta \approx (\angle(\psi_{\theta}, \psi_{\theta+\delta}))^2 = \left(\arccos \frac{|\langle \psi_{\theta}, \psi_{\theta+\delta} \rangle|}{\|\psi_{\theta}\| \|\psi_{\theta+\delta}\|} \right)^2,$$

where $\mathbf{S}$ is the overlap matrix defined by

$$\mathbf{S} = \frac{\langle \nabla_{\theta} \psi_{\theta}, \nabla_{\theta} \psi_{\theta}^{\top} \rangle}{\langle \psi_{\theta}, \psi_{\theta} \rangle} - \frac{\langle \nabla_{\theta} \psi_{\theta}, \psi_{\theta} \rangle \langle \psi_{\theta}, \nabla_{\theta} \psi_{\theta}^{\top} \rangle}{\langle \psi_{\theta}, \psi_{\theta} \rangle^2}. \quad (6)$$

The second quadratic penalty is based on the square magnitude of the update, $\eta \|\delta\|^2$. In summary, SR solves the optimization problem

$$\underset{\delta}{\operatorname{argmin}} \left[\frac{1}{\epsilon} \delta^* (\mathbf{S} + \eta \mathbf{I}) \delta + \delta^* \mathbf{g} + \mathbf{g}^* \delta \right], \quad \text{for } \eta \geq 0.$$

The resulting parameter update is the solution to the positive definite linear system

$$(\mathbf{S} + \eta \mathbf{I}) \delta = \epsilon \mathbf{g}. \quad (7)$$

In practice, the exact SR update given in Eq. (7) is hard to calculate, so a Monte Carlo approximation is used. Suppose one can sample from the wave function density

$$\rho_{\theta}(\sigma) = \frac{|\psi_{\theta}(\sigma)|^2}{\sum_{\sigma'} |\psi_{\theta}(\sigma')|^2},$$

using a population of Markov chain Monte Carlo (MCMC) walkers (see Sec. IV). Then, the energy in Eq. (3) can be reformulated and estimated as an expectation

$$E(\psi_{\theta}) = \sum_{\sigma} E_{\text{loc}}(\sigma) \rho_{\theta}(\sigma) = \mathbb{E}_{\rho}[E_{\text{loc}}(\sigma)], \quad (8)$$

where E_{loc} denotes the “local energy,” defined by

$$E_{\text{loc}}(\sigma) = \frac{(H\psi_{\theta})(\sigma)}{\psi_{\theta}(\sigma)}.$$

The gradient vector $\mathbf{g}$ and overlap matrix $\mathbf{S}$ can be reformulated and estimated as covariances with respect to the wave function density $\rho_{\theta}(\sigma)$. The resulting SR algorithm is presented as Algorithm 1.

Algorithm 1 (Stochastic Reconfiguration). Choose the parameter update δ to solve

$$(\mathbf{S} + \eta \mathbf{I})\delta = -\epsilon \mathbf{g}. \quad (9)$$

Here, $\eta \geq 0$ is a non-negative parameter that makes $\mathbf{S} + \eta \mathbf{I}$ positive definite. The energy $E(\psi_{\theta})$, gradient vector $\mathbf{g}$, and overlap matrix $\mathbf{S}$ are defined by

$$E(\psi_{\theta}) = \mathbb{E}_{\rho}[E_{\text{loc}}(\sigma)], \quad (10a)$$

$$\mathbf{g} = \text{Cov}_{\rho}\left[\frac{\nabla_{\theta}\psi_{\theta}(\sigma)}{\psi_{\theta}(\sigma)}, E_{\text{loc}}(\sigma)\right], \quad (10b)$$

$$\mathbf{S} = \text{Cov}_{\rho}\left[\frac{\nabla_{\theta}\psi_{\theta}(\sigma)}{\psi_{\theta}(\sigma)}, \frac{\nabla_{\theta}\psi_{\theta}(\sigma)}{\psi_{\theta}(\sigma)}\right], \quad (10c)$$

where $\mathbb{E}_{\rho}$ and Cov_{ρ} indicate the expectation value and covariance with respect to the wave function density $\rho_{\theta}(\sigma)$.

III. WEIGHTED VARIATIONAL MONTE CARLO

This section derives stochastic reconfiguration from an alternative perspective of gradient flows, building on the approach in Ref. [20]. Then, it proposes a modified gradient flow method to improve generalization error in VMC.

A. Gradient flow perspective

The energy functional $E(\psi)$ has a Wirtinger gradient [31] that is an n -dimensional vector

$$\nabla_{\psi}E(\psi) = \frac{H\psi - E(\psi)\psi}{\langle \psi, \psi \rangle}.$$

Therefore, a theoretical method to find the exact ground state is based on running the continuous-time gradient flow:

$$\partial_t \psi_t = -\nabla_{\psi_t}E(\psi_t) = \frac{E(\psi_t)\psi_t - H\psi_t}{\langle \psi_t, \psi_t \rangle}. \quad (11)$$

Yet, this continuous-time evolution cannot be realized when the wave function lies on a parametrized manifold $\psi_t = \psi_{\theta}$. The gradient flow pushes $\psi_{t+\epsilon}$ off the manifold, so that no parameter vector $\theta_{t+\epsilon}$ exactly corresponds to $\psi_{t+\epsilon}$ defined by Eq. (11).

To apply gradient flow within a parametrized manifold, it is necessary to combine evolution with projection. We first evolve the wave function from ψ to $\psi_{t+\epsilon}$. We then identify a parameter vector $\theta_{t+\epsilon}$ that projects $\psi_{t+\epsilon}$ onto the manifold:

$$\theta_{t+\epsilon} = \underset{\theta}{\text{argmin}} d(\psi_{t+\epsilon}, \psi_{\theta}). \quad (12)$$

The projection step can be carried out in any metric $d(\cdot, \cdot)$, and one natural choice is the Fubini-Study metric [19]

$$d(\phi, \psi) = \angle(\phi, \psi) = \arccos \frac{|\langle \phi, \psi \rangle|}{\|\phi\| \|\psi\|},$$

which measures the angle between two unnormalized wave functions. In the limit of small time steps $\epsilon \rightarrow 0$, the evolve-and-project method leads to a parameter update,

$$\theta_{t+\epsilon} = \theta_t - \frac{\epsilon}{\|\psi_{\theta}\|^2} \mathbf{S}^{-1} \mathbf{g} + \mathcal{O}(\epsilon^2).$$

This update corresponds to the SR method (9) with $\eta = 0$ and a modified step size $\epsilon/\|\psi_{\theta}\|^2$.

B. Weighted variational Monte Carlo

We now exploit the flexibility afforded by the evolve-and-project method. Given any positive weight function $\mu(\sigma)$, we can introduce a weighted Fubini-Study metric

$$d_{\mu}(\phi, \psi) = \arccos \frac{|\langle \phi, \psi \rangle_{\mu}|}{\|\phi\|_{\mu} \|\psi\|_{\mu}},$$

with an inner product and norm defined by

$$\langle \phi, \psi \rangle_{\mu} = \sum_{\sigma} \mu(\sigma) \overline{\phi(\sigma)} \psi(\sigma),$$

$$\|\psi\|_{\mu}^2 = \sum_{\sigma} \mu(\sigma) |\psi(\sigma)|^2.$$

Like the standard Fubini-Study metric, d_{μ} is non-negative and $d_{\mu}(\phi, \psi) = 0$ only when $\phi = c\psi$, where c is a complex number. However, the weight function $\mu(\sigma)$ can be optimized in this approach.

The change in metric leads to a change in the projection step. Now, as $\epsilon \rightarrow 0$, the evolve-and-project method leads to

$$\theta_{t+\epsilon} = \theta_t - \frac{\epsilon}{\|\psi_{\theta}\|_{\mu}^2} \tilde{\mathbf{S}}^{-1} \tilde{\mathbf{g}} + \mathcal{O}(\epsilon^2).$$

Here, the gradient vector $\mathbf{g}$ in Eq. (5) and overlap matrix $\mathbf{S}$ in Eq. (6) are modified to become

$$\tilde{\mathbf{g}} = \frac{\langle \nabla_{\theta}\psi_{\theta}, (H - E(\psi))\psi_{\theta} \rangle_{\mu}}{\langle \psi_{\theta}, \psi_{\theta} \rangle_{\mu}},$$

$$\tilde{\mathbf{S}} = \frac{\langle \nabla_{\theta}\psi_{\theta}, \nabla_{\theta}\psi_{\theta}^{\top} \rangle_{\mu}}{\langle \psi_{\theta}, \psi_{\theta} \rangle_{\mu}} - \frac{\langle \nabla_{\theta}\psi_{\theta}, \psi_{\theta} \rangle_{\mu}}{\langle \psi_{\theta}, \psi_{\theta} \rangle_{\mu}} \frac{\langle \psi_{\theta}, \nabla_{\theta}\psi_{\theta}^{\top} \rangle_{\mu}}{\langle \psi_{\theta}, \psi_{\theta} \rangle_{\mu}}.$$

The major difference is that the L^2 inner products have been replaced by μ -weighted inner products. This evolve-and-project approach can be understood as an example of the time-dependent variational principle [21,22,32,33].

In weighted VMC, the parameter update $\tilde{\mathbf{S}}^{-1} \tilde{\mathbf{g}}$ is hard to calculate exactly. Instead, $\tilde{\mathbf{g}}$ and $\tilde{\mathbf{S}}$ can be estimated using covariances with respect to a weighted wave function density:

$$\tilde{\rho}_{\theta}(\sigma) = \frac{\mu(\sigma) |\psi_{\theta}(\sigma)|^2}{\sum_{\sigma'} \mu(\sigma') |\psi_{\theta}(\sigma')|^2}. \quad (13)$$

The “weighted VMC” method based on Eq. (13) is presented as Algorithm 2. As usual, the practical algorithm includes a small factor $\eta \mathbf{I}$, which helps stabilize the updates.

Algorithm 2 (Weighted VMC). Choose the parameter update δ to solve

$$(\tilde{\mathbf{S}} + \eta \mathbf{I})\delta = -\epsilon \tilde{\mathbf{g}}. \quad (14)$$

Here, $\eta \geq 0$ is a non-negative parameter that makes $\tilde{S} + \eta I$ positive definite. The energy $E(\psi_\theta)$, gradient vector $\tilde{g}$, and overlap matrix $\tilde{S}$ are defined by

$$E(\psi_\theta) = \mathbb{E}_\rho[E_{\text{loc}}(\sigma)], \quad (15a)$$

$$\tilde{g} = \text{Cov}_{\tilde{\rho}} \left[\frac{\nabla_\theta \psi_\theta(\sigma)}{\psi_\theta(\sigma)}, E_{\text{loc}}(\sigma) \right], \quad (15b)$$

$$\tilde{S} = \text{Cov}_{\tilde{\rho}} \left[\frac{\nabla_\theta \psi_\theta(\sigma)}{\psi_\theta(\sigma)}, \frac{\nabla_\theta \psi_\theta(\sigma)}{\psi_\theta(\sigma)} \right], \quad (15c)$$

where $\mathbb{E}_\rho$ indicates the expectation value with respect to the wave function density $\rho_\theta(\sigma)$ and $\text{Cov}_{\tilde{\rho}}$ indicates the covariance with respect to the weighted wave function density $\tilde{\rho}_\theta(\sigma)$.

Weighted VMC uses a sequence of sampling distributions that are defined iteratively. At each optimization step, we generate samples from a weighted distribution $\tilde{\rho}_\theta(\sigma)$ to optimize the parameter vector θ . Then, after all the optimization steps are complete, we generate samples from $\rho_\theta(\sigma)$ to evaluate the energy $E(\psi_\theta)$. This approach has the same computational cost as standard VMC (Algorithm 1).

What opportunities does weighted VMC provide us? In standard VMC, the projection step barely emphasizes the regions of small wave function density $\rho_\theta(\sigma)$. With only a finite number of samples from ρ_θ , few, if any, samples will lie in low-probability regions, leaving the wave function unconstrained there. In contrast, weighted VMC controls the relative importance of different regions using a weighted sampling distribution $\tilde{\rho}_\theta(\sigma)$. Increasing the relative density of $\tilde{\rho}_\theta(\sigma)$ at a location σ results in greater agreement between $\psi_{t+\epsilon}(\sigma)$ and the parametrized update $\psi_{t+\delta}(\sigma)$, which can lead to higher quality at the selected location. By adjusting the sampling distribution, we can selectively refine the wave function where refinement is needed most.

IV. WEIGHTED DISTRIBUTIONS

The key to weighted VMC is the freedom to choose a weighted distribution that emphasizes important regions of configuration space. This section describes several distributions that can be used for sampling from the XXX Heisenberg model. The section first describes the standard MCMC approach to sample from the unweighted wave function density. Then, it proposes three modifications that emphasize low-probability regions.

A. Markov chain Monte Carlo

Here, we explain the basic approach to sample from the wave function density $\rho_\theta(\sigma)$ using MCMC. All the sampling methods for weighted distributions $\tilde{\rho}_\theta(\sigma)$ are modified from this basic MCMC sampling scheme.

Our approach uses a population of $M = 100$ independent MCMC walkers, which are labeled σ_t^m for $m = 1, \dots, M$ and $t = 0, 1, \dots$. We evolve each walker by proposing a new configuration σ_{t+1}^m where a randomly chosen positive spin $(\sigma_t^m)_i = +1$ and a randomly chosen negative spin $(\sigma_t^m)_j = -1$

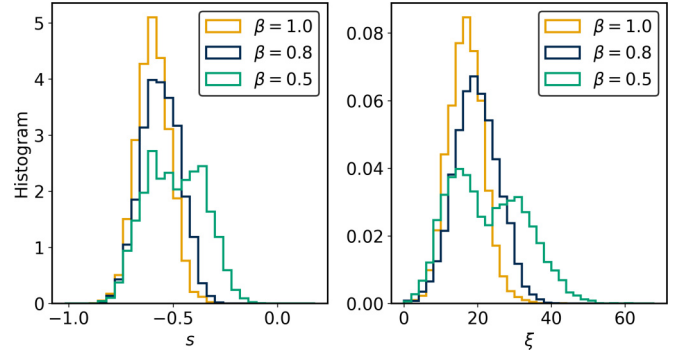


FIG. 1. Histograms of the collective variables $s(\sigma)$ defined in Eq. (17) (left) and $\xi(\sigma)$ defined in Eq. (18) (right), when 4000 MCMC samples are generated from the MT distribution with $\beta = 1.0, 0.8$, or 0.5 .

are simultaneously flipped,

$$(\sigma_{t+1}^m)_\ell = \begin{cases} -(\sigma_t^m)_\ell, & \ell \in \{i, j\}, \\ (\sigma_t^m)_\ell, & \text{otherwise.} \end{cases}$$

We accept the “swap move” with probability

$$\min \left\{ 1, \frac{|\psi_\theta(\sigma_{t+1}^m)|^2}{|\psi_\theta(\sigma_t^m)|^2} \right\}.$$

Otherwise, we reject and set $\sigma_{t+1}^m = \sigma_t^m$ instead. Unless otherwise stated, we carry out 20N Metropolis steps in this way, subsampling the σ values once per N steps. The resulting total number of samples used to estimate the covariances in Eqs. (10) and (15) is 2000. This MCMC sampling scheme preserves the total magnetization, $\sum_{\ell=1}^n (\sigma_t^m)_\ell = 0$, which is appropriate since the exact ground state of the XXX model (1) is supported on states of total magnetization zero whenever there is a bipartite lattice [25].

B. Mixed tempering

The first version of weighted VMC uses a “mixed tempering” (MT) distribution, where one half of the MCMC walkers are sampled from $\rho_\theta(\sigma)$ and the other half are sampled from a tempered distribution:

$$\frac{\rho_\theta(\sigma)^\beta}{\sum_\sigma \rho_\theta(\sigma)^\beta},$$

determined by an “inverse temperature” parameter $\beta \in [0, 1]$. Therefore, the MT distribution can be written as

$$\tilde{\rho}_\theta(\sigma) = \frac{\rho_\theta(\sigma)}{2} + \frac{\rho_\theta(\sigma)^\beta}{2 \sum_\sigma \rho_\theta(\sigma)^\beta}. \quad (16)$$

Figure 1 visualizes the MT distribution with $\beta = 1.0, 0.8$, and 0.5 along two choices of “collective variables.” The first CV is the nearest-neighbor spin correlation function

$$s(\sigma) \equiv \frac{1}{N} \sum_{i=1}^N \sigma_i \sigma_{i+1}. \quad (17)$$

The second CV is the negative logarithm of the unnormalized wave function density

$$\xi(\sigma) \equiv -\log |\psi_\theta(\sigma)|^2. \quad (18)$$

When $\beta = 1$ in Eq. (16), the MT distribution reduces to the wave function density $\rho_\theta(\sigma)$. However, as $\beta \rightarrow 0$, the second group of walkers spreads more uniformly over the state space $\{\sigma\}$. This spreading out is reflected in the histograms in Fig. 1.

C. Well-tempered metadynamics

Another version of weighted VMC is based on well-tempered metadynamics (WTMD) [23,24]. WTMD is a Monte Carlo method that estimates the marginal distribution of a high-dimensional probability along a user-specified CV. WTMD drives the walkers to explore regions where the estimated marginal density is small, and it uses reweighting to recover statistics.

In this study, we apply a sampling method inspired by WTMD where the CV is the negative logarithm of the unnormalized wave function density, $\xi(\sigma) = -\log |\psi_\theta(\sigma)|^2$ as in Eq. (18). Our specific sampling procedure follows:

(1) *Initialization.* Set w and ΔT parameters for the marginal density estimation. Set the lower bound L_ξ and bin interval $\Delta\xi$ for the histogram of ξ . Set the initial histogram values $V = (0, \dots, 0) \in \mathbb{R}^{n+1}$, and choose an initial configuration of walkers $\sigma_0^1, \dots, \sigma_0^M$.

(2) *Sampling.* For each $t = 0, \dots, T$ and $m = 1, \dots, M$, perform the following steps.

(a) Propose an evolution from σ_t^m to σ_{t+1}^m using a random swap move. Accept σ_{t+1}^m with probability

$$\min \left\{ 1, \frac{e^{V(\sigma_t^m)} |\psi_\theta(\sigma_{t+1}^m)|^2}{e^{V(\sigma_{t+1}^m)} |\psi_\theta(\sigma_t^m)|^2} \right\}. \quad (19)$$

Otherwise, reject and set $\sigma_{t+1}^m = \sigma_t^m$.

(b) Compute a bin index $i \in \mathbb{Z}$ such that

$$\xi(\sigma_{t+1}^m) \in (L_\xi + (i-1)\Delta\xi, L_\xi + i\Delta\xi].$$

Round up to $i = 0$ if i is negative. Round down to $i = n+1$ if i exceeds $n+1$. Then, update the associated histogram value

$$V(i) = V(i) + w e^{-V(i)/\Delta T}.$$

In the WTMD sampling procedure, the walkers are driven to explore new regions because of the term $e^{V(\sigma_t^m) - V(\sigma_{t+1}^m)}$ in the acceptance probability (19). After T steps, the walkers are done exploring, and we define the sampling distribution as the average of delta functions on the walkers' locations:

$$\tilde{\rho}_\theta(\sigma) = \frac{1}{MT} \sum_{m=1}^M \sum_{t=1}^T \delta_{\sigma_t^m}(\sigma).$$

We start fresh by reinitializing the marginal density approximation and the sampling distribution after updating the VMC parameter weights.

D. Uniform CV sampling

The MT and WTMD sampling distributions can be applied to any quantum system since they only depend on the

unnormalized wave function density $|\psi_\theta(\sigma)|^2$. However, a different approach leverages prior knowledge of the physical system in the form of a well-chosen CV. For example, the ground state of the XXX model is primarily supported on configurations in the antiferromagnetic region, where

$$s(\sigma) = \frac{1}{N} \sum_{i=1}^N \sigma_i \sigma_{i+1} < 0.$$

In contrast, the wave function density puts little mass on the ferromagnetic region, $\{\sigma | s(\sigma) > 0\}$. We can help address this imbalance by applying weighted VMC with a target distribution

$$\tilde{\rho}_\theta(\sigma) = 0.05 p(\sigma) + 0.95 \rho_\theta(\sigma), \quad (20)$$

where $p(\sigma)$ has a nearly uniform marginal distribution along the CV $s(\sigma)$. In practice, we can pregenerate samples from $p(\sigma)$ and randomly draw from the pool in each iteration of weighted VMC. See Appendix B for details.

V. NUMERICAL EXPERIMENTS

This section describes numerical experiments comparing traditional and weighted VMC methods (Algorithms 1 and 2). These computations are implemented using the JAX library for Python [34], and complete scripts for the algorithms and output analysis are available on GitHub [35]. We tested both VMC approaches on the XXX Heisenberg model for a periodic chain of $N = 100$ spins. We choose $\alpha = 30$, resulting in $\alpha(N+1) = 3030$ trainable parameters, which is greater than the number of Monte Carlo samples, 2000, used to estimate the covariances in Eqs. (10) and (15). The regularization parameter η in Eqs. (9) and (14) is set to 0.001. For comparison with other choices of the width parameter α , see Appendix A. We evaluated performance based on two metrics: energy error and local energy error.

The energy error is defined as

$$[E(\psi_\theta) - E_0]/N. \quad (21)$$

Here, $E(\psi_\theta)$ is the energy associated with the VMC wave function ψ_θ , and E_0 is the exact ground state energy. Throughout this section, energies are normalized by the number of spins N .

The local energy error is defined as

$$|E_{\text{loc}}(\sigma) - E_0|/N. \quad (22)$$

Here, $E_{\text{loc}}(\sigma) = (H\psi)(\sigma)/\psi(\sigma)$ is the “local energy,” which satisfies $E_{\text{loc}}(\sigma) = E_0$ when ψ is the exact ground state. Therefore, the local energy error quantifies local inaccuracies in the VMC wave function approximation.

The energy and local energy errors were numerically evaluated as follows. First, we calculated the ground state energy E_0 exactly using the Bethe ansatz [18]. Next, to evaluate the energy $E(\psi_\theta)$ via Eq. (8), we extracted the parameter vector θ after 5000 iterations of traditional or weighted VMC. We ran $N = 100$ MCMC walkers for 20 000 Metropolis steps targeting the wave function density ρ_θ . We subsampled every 100 steps and removed the first half of the time series to account for initialization bias. Finally, we evaluated the

TABLE I. Energy errors of traditional VMC versus weighted VMC with the MT ($\beta = 0.5$) or WTMD distribution, computed from ten independent trials.

Method	Mean	Median
VMC	2.66×10^{-6}	2.34×10^{-6}
MT	1.46×10^{-6}	1.27×10^{-6}
WTMD	1.45×10^{-6}	1.48×10^{-6}

local energy error for traditional and weighted VMC using the diverse sample set described in Appendix B.

The numerical experiments lead to three main results:

(1) Weighted VMC methods reduce the energy error by a factor of 2 compared to traditional VMC. See Sec. IV.

(2) Weighted VMC methods reduce the local energy errors in the tails of the probability distribution by factors of 10^2 – 10^4 compared to traditional VMC. See Sec. VB.

(3) Weighted VMC methods extend to higher-dimensional systems with $N = 200$ spins, yielding the same pattern of improved energies and local energies. See Sec. VC.

The rest of the section describes these results in detail.

A. Improved energies

As the first main result, Table I demonstrates a factor of 2 energy improvement when switching from traditional to weighted VMC using either the MT ($\beta = 0.5$) or WTMD sampling distribution (Secs. IVB and IV C). The same data are visualized in a different way using boxplots in Fig. 2. The comparison between traditional and weighted VMC is reasonable since these methods have the same computational cost.

To investigate the source of the energy improvement, we select one run of each VMC, MT, and WTMD for further analysis. In Fig. 3, we plot histograms and local energy errors across two collective variables, $s(\sigma)$ defined in Eq. (17) and $\xi(\sigma)$ defined in Eq. (18). From the histograms, we see that weighted VMC increases the number of samples in regions far away from the mode of ρ_θ . From the local energies, we see that weighted VMC leads to improved wave function accuracy

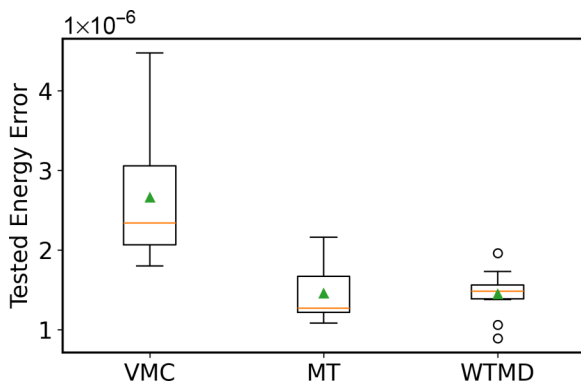


FIG. 2. Boxplots of energy errors in traditional VMC vs weighted VMC (MT, WTMD). The orange lines are median values and the green triangles are mean values. The dots are outliers beyond $4 \times$ the interquartile range.

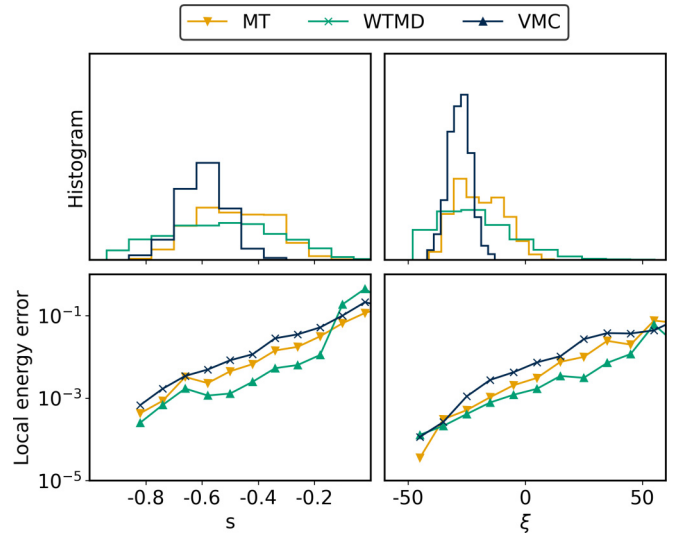


FIG. 3. Histograms and magnitudes of local energy errors evaluated across two collective variables: $s(\sigma)$ (left) and $\xi(\sigma)$ (right). Curves show the average local energy error for each bin.

for all the CV values where the histograms are populated. This robust improvement in the wave function quality implies a surprising conclusion: Improved sampling in the region *surrounding* the mode can enhance local accuracy *at* the mode.

We observe minor differences between MT and WTMD sampling distributions. The variance of the WTMD energy estimate is slightly lower than the variance of the MT energy estimate, as seen in Fig. 2. Nevertheless, MT and WTMD lead to similar mean and median energy errors as reported in Table I. From this comparison, weighted VMC can perform well with multiple sampling distributions, and neither MT nor WTMD uniformly improves on the other.

We also explored how the inverse temperature β affects the energy estimate using MT sampling. Figure 4 shows boxplots of the MT energy errors with parameters $\beta = 1.0, 0.8, 0.5$, and 0.2 . The choice $\beta = 1.0$ corresponds to traditional VMC sampling. As β decreases, the energy error initially improves, reaching its lowest value at $\beta = 0.5$. However, further reduction in β leads to a worse energy estimate, as the samples become too dispersed to effectively refine the wave function

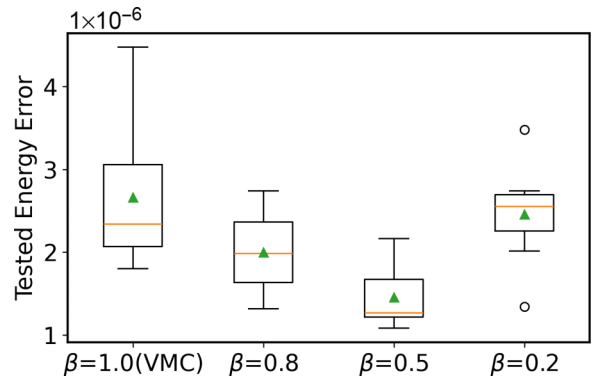


FIG. 4. Boxplots of energy errors using the MT sampling distribution with parameters $\beta = 1.0, 0.8, 0.5$, and 0.2 .

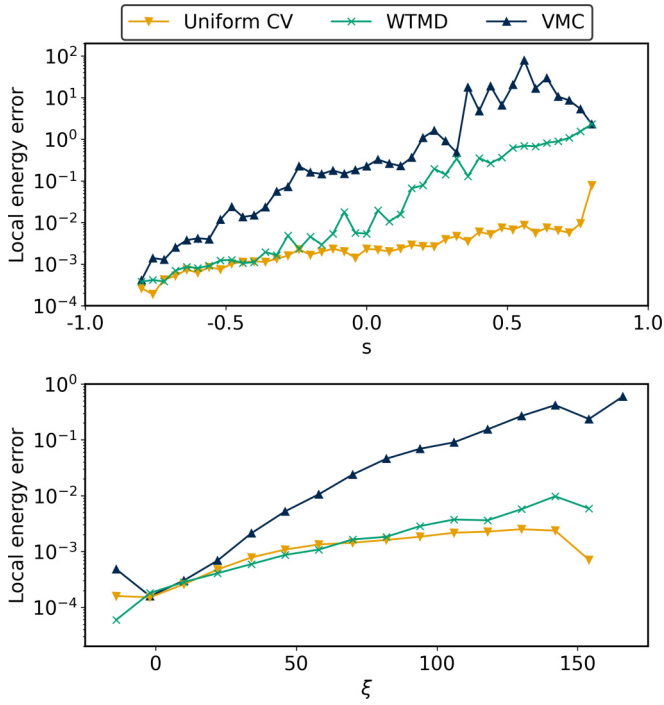


FIG. 5. Local energy errors for uniform CV sampling (uniform CV, yellow), WTMD (green), and traditional VMC (VMC, dark blue). The spike in the traditional VMC error at the lowest level of $\xi(\sigma)$ represents a spurious mode (see Ref. [16]).

estimate near the mode of ρ_θ . See Fig. 1 for an illustration of the MT distribution with several choices of β .

B. Improved local energies in the tails

Weighted VMC makes it possible to improve the local energy accuracy in low-probability regions while still maintaining an advantage over VMC for energy estimation. These benefits are already evident when the user applies a general-purpose sampling distribution such as MT or WTMD. The benefits are further amplified when weighted VMC is applied with a physically informed sampling distribution that drives walkers toward important tail regions identified by a user-chosen collective variable (Sec. IV D).

Figure 5 compares the local energy errors for traditional VMC, uniform CV sampling, and WTMD sampling. In the uniform CV approach, sampling from $\tilde{\rho}_\theta$ in Eq. (20) is carried out by drawing from ρ_θ as described in Sec. IV A and independently drawing the same number of samples from p using the pregenerated pool. To ensure a fair comparison, we adjusted the number of Monte Carlo samples in traditional VMC and WTMD to 3000, so that all methods require the same number of wave function queries. In the figure, uniform CV sampling outperforms. WTMD for $s(\sigma) > 0$: WTMD beats traditional VMC by a factor of 10–100 while uniform CV sampling beats traditional VMC by a larger factor of 10^3 – 10^4 .

In summary, the most powerful sampling distribution leverages physical insight to target important regions in the low-probability tails. This physically inspired VMC sampling approach also leads to competitive energy estimates, as shown in Table II.

TABLE II. Energy errors of traditional VMC versus weighted VMC with uniform CV sampling or WTMD, computed from ten independent trials.

Method	Mean	Median
VMC	2.10×10^{-6}	1.95×10^{-6}
Uniform CV	1.71×10^{-6}	1.64×10^{-6}
WTMD	1.59×10^{-6}	1.52×10^{-6}

C. Higher-dimensional results

Finally, we tested energy and local energy errors on a higher-dimensional XXX system with $N = 200$ spins. We choose $\alpha = 50$ for these tests, resulting in 10 050 trainable parameters. Sampling strategies remain unchanged. The results are qualitatively similar to those with $N = 100$ spins, and they indicate an improvement of weighted VMC over traditional VMC. Figure 6 shows that MT and WTMD still outperform VMC on the higher-dimensional problem, although the WTMD results show a bit more spread than at $N = 100$. Similar to the results in Fig. 4, the energy errors from MT show a decrease-then-increase trend as the inverse temperature β decreases. In the $N = 200$ case, the optimized parameter setting is slightly higher, $\beta = 0.8$.

VI. CONCLUSION

While neural networks are highly expressive models for quantum states, they lead to a well-known problem of

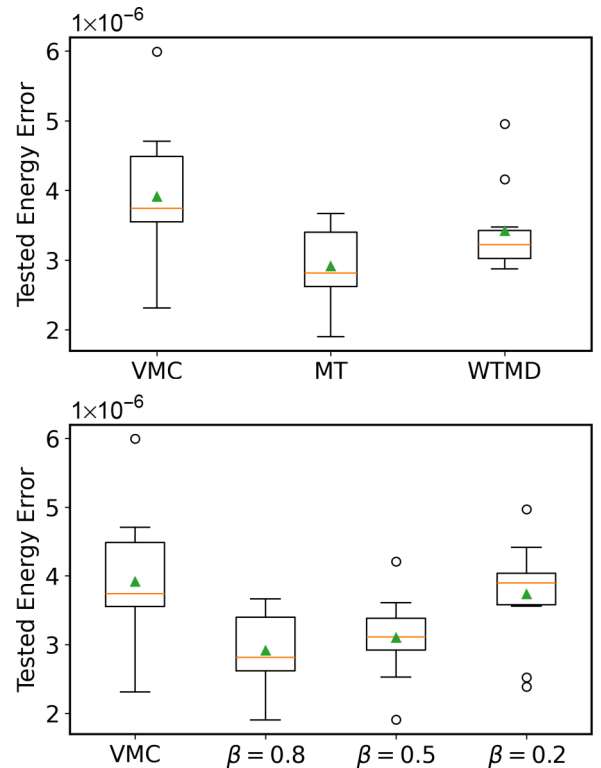


FIG. 6. Boxplots of energy errors for $N = 200$ spins using traditional VMC vs weighted VMC (MT and WTMD). Compare to Figs. 2 and 4, which use $N = 100$ spins.

TABLE III. Mean energy errors of traditional VMC versus uniform CV sampling, computed from ten independent trials.

α	Uniform CV	VMC
6	3.63×10^{-5}	3.86×10^{-3}
10	1.49×10^{-5}	1.57×10^{-4}
20	4.61×10^{-6}	4.16×10^{-6}
30	2.35×10^{-6}	3.83×10^{-6}
50	2.30×10^{-6}	4.03×10^{-6}

generalization error. The learned wave functions can be inaccurate in low-probability regions, which makes them risky to use for downstream applications.

In this study, we proposed a weighted VMC method that improves the accuracy of neural network approximations in low-probability regions. The core idea is to decouple the sampling distribution $\tilde{\rho}_\theta(\sigma)$ from the wave function density $\rho_\theta(\sigma)$.

We found that the weighted VMC approach has several advantages over traditional VMC:

(1) *Improved energy estimation.* Weighted VMC improves the accuracy of the estimated energies and local energies without additional computational cost.

(2) *Flexible sampling.* Weighted VMC can use an arbitrary sampling distribution targeting regions where wave function refinement is needed.

(3) *Flexible implementation.* The method supports advanced sampling techniques such as tempering, metadynamics, and importance sampling, as well as problem-specific modifications, without any reweighting. It is broadly applicable to any quantum system with or without detailed prior knowledge of the system.

(4) *Extensions.* Our general approach can be extended to other optimization problems consisting of gradient flow and projection steps. The metric used in the projection step can be designed to improve accuracy in specific regions.

While our analysis focused on the energy and the local energy accuracy in low-probability regions, we anticipate that weighted VMC will also positively impact other observables beyond energy, such as spin correlation functions. An interesting question, which we do not explore here, is how to choose the weighted sample density $\tilde{\rho}_\theta$ to improve the accuracy for any particular observable. Finally, we have chosen to present our weighted VMC approach within the familiar SR optimization framework. We expect that weighted VMC will also improve training when employed with popular optimization strategies for large, state-of-the-art parametrizations [36–38], but we leave that question for future work.

ACKNOWLEDGMENTS

This work was supported in part through the NYU IT High Performance Computing resources, services, and staff expertise. H.Z. and J.W. acknowledge support from the Army Research Office under Grant No. W911NF-22-2-0124. M.L. acknowledges support from the U.S. Department of Energy, Office of Science, Accelerated Research in Quantum Computing Centers, Quantum Utility through Advanced Computational Quantum Algorithms, Grant No. DE-SC0025572 and from the Applied Mathematics

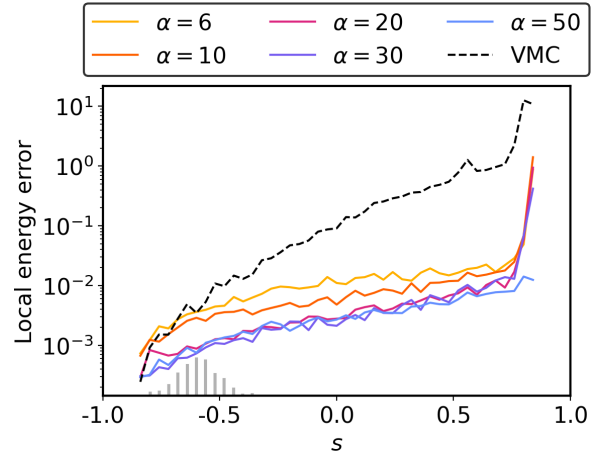


FIG. 7. Local energy errors for traditional VMC ($\alpha = 50$) and uniform CV sampling ($\alpha = 6, 10, 20, 30$, and 50). Histogram shows 1000 Monte Carlo samples from ρ_θ .

Program of the U.S. Department of Energy (DOE) Office of Advanced Scientific Computing Research under Contract No. DE-AC02-05CH11231.

DATA AVAILABILITY

All code and parameters required to reproduce the results in the article, along with scripts to generate the figures are openly available at [39].

APPENDIX A: THE ROLE OF WIDTH

This Appendix explores the impact of the RBM width parameter α , which controls the total number of free parameters via $\alpha(N + 1)$. When α is very low ($\alpha \leq 10$), traditional VMC can lead to spurious modes that compromise the accuracy of the energy estimates [16]. Additionally, when α is low ($\alpha \leq 20$), the energy estimates from weighted VMC with uniform CV sampling (20) exhibit large errors. See Table III, which reports the energy errors for traditional and weighted VMC with $\alpha = 6, 10, 20, 30$, and 50 for $N = 100$ spins.

For further insight, Fig. 7 shows the local energy errors for traditional VMC with $\alpha = 50$ versus weighted VMC with $\alpha = 6, 10, 20, 30$, and 50 . When $\alpha \geq 30$, the local energy errors are uniformly smaller for weighted VMC than traditional VMC, both in the high-probability and low-probability regions. This makes sense because the width parameter contributes flexibility to the projection step in Eq. (12). Combining a flexible projection step with a weighted projection metric leads to a close match between the parametrized wave function and the gradient flow direction.

APPENDIX B: DETAILS OF UNIFORM CV SAMPLING

This Appendix describes the sampling distribution $p(\sigma)$, which is used for visualization in Figs. 3, 5, and 7 and used for sampling in the uniform CV method (Sec. IV D). To generate samples, we introduce $L = 52$ unnormalized Gaussian density functions,

$$p_i(\sigma) = \exp\left(-\frac{(s(\sigma) - m_i)^2}{2\kappa^2}\right),$$

which have the same standard deviation $\kappa = 0.04$ but different means,

$$m_i = -1.06 + 0.04i, \text{ for } i = 1, \dots, 52.$$

To sample each density p_i , we initialize ten MCMC samplers. Then, we run 2×10^6 Metropolis steps and collect one sample every 2×10^3 steps. In this way, we build a pool consisting of 5.2×10^5 samples from p , from which we randomly pick 1000 samples for visualization.

-
- [1] G. Carleo and M. Troyer, Solving the quantum many-body problem with artificial neural networks, *Science* **355**, 602 (2017).
 - [2] D. Luo and B. K. Clark, Backflow transformations via neural networks for quantum many-body wave functions, *Phys. Rev. Lett.* **122**, 226401 (2019).
 - [3] D. Pfau, J. S. Spencer, A. G. D. G. Matthews, and W. M. C. Foulkes, *Ab initio* solution of the many-electron Schrödinger equation with deep neural networks, *Phys. Rev. Res.* **2**, 033429 (2020).
 - [4] J. Hermann, Z. Schätzle, and F. Noé, Deep-neural-network solution of the electronic Schrödinger equation, *Nat. Chem.* **12**, 891 (2020).
 - [5] J. Gubernatis, N. Kawashima, and P. Werner, *Quantum Monte Carlo Methods: Algorithms for Lattice Models* (Cambridge University Press, Cambridge, UK, 2016).
 - [6] F. Becca and S. Sorella, *Quantum Monte Carlo Approaches for Correlated Systems* (Cambridge University Press, Cambridge, UK, 2017).
 - [7] K. Choo, G. Carleo, N. Regnault, and T. Neupert, Symmetries and many-body excitations with neural-network quantum states, *Phys. Rev. Lett.* **121**, 167204 (2018).
 - [8] Y. Nomura, Helping restricted Boltzmann machines with quantum-state representation by restoring symmetry, *J. Phys.: Condens. Matter* **33**, 174003 (2021).
 - [9] S. Pathak, B. Busemeyer, J. N. B. Rodrigues, and L. K. Wagner, Excited states in variational Monte Carlo using a penalty method, *J. Chem. Phys.* **154**, 034101 (2021).
 - [10] M. T. Entwistle, Z. Schätzle, P. A. Erdman, J. Hermann, and F. Noé, Electronic excited states in deep variational Monte Carlo, *Nat. Commun.* **14**, 274 (2023).
 - [11] C. Filippi, M. Zaccheddu, and F. Buda, Absorption spectrum of the green fluorescent protein chromophore: A difficult case for *ab initio* methods? *J. Chem. Theory Comput.* **5**, 2074 (2009).
 - [12] P. M. Zimmerman, J. Toulouse, Z. Zhang, C. B. Musgrave, and C. J. Umrigar, Excited states of methylene from quantum Monte Carlo, *J. Chem. Phys.* **131**, 124103 (2009).
 - [13] L. Zhao and E. Neuscamman, An efficient variational principle for the direct optimization of excited states, *J. Chem. Theory Comput.* **12**, 3436 (2016).
 - [14] T. Westerhout, N. Astrakhantsev, K. S. Tikhonov, M. I. Katsnelson, and A. A. Bagrov, Generalization properties of neural network approximations to frustrated magnet ground states, *Nat. Commun.* **11**, 1593 (2020).
 - [15] R. J. Webber and M. Lindsey, Rayleigh-Gauss-Newton optimization with enhanced sampling for variational Monte Carlo, *Phys. Rev. Res.* **4**, 033099 (2022).
 - [16] H. Zhang, R. J. Webber, M. Lindsey, T. C. Berkelbach, and J. Weare, Understanding and eliminating spurious modes in variational Monte Carlo using collective variables, *Phys. Rev. Res.* **5**, 023101 (2023).
 - [17] C.-Y. Park and M. J. Kastoryano, Geometry of learning neural quantum states, *Phys. Rev. Res.* **2**, 023232 (2020).
 - [18] M. Karbach, K. Hu, and G. Müller, Introduction to the Bethe ansatz, II, *Comput. Phys.* **12**, 565 (1998).
 - [19] J. Stokes, J. Izaac, N. Killoran, and G. Carleo, Quantum natural gradient, *Quantum* **4**, 269 (2020).
 - [20] K. Neklyudov, J. Nys, L. Thiede, J. Carrasquilla, Q. Liu, M. Welling, and A. Makhzani, Wasserstein quantum Monte Carlo: A novel approach for solving the quantum many-body Schrödinger equation, in *Advances in Neural Information Processing Systems* (Curran Associates, Inc., New York, 2023), Vol. 36, pp. 63461–63482.
 - [21] P. A. M. Dirac, Note on exchange phenomena in the Thomas atom, *Math. Proc. Cambridge Philos. Soc.* **26**, 376 (1930).
 - [22] J. Haegeman, C. Lubich, I. Oseledets, B. Vandereycken, and F. Verstraete, Unifying time evolution and optimization with matrix product states, *Phys. Rev. B* **94**, 165116 (2016).
 - [23] A. Laio and M. Parrinello, Escaping free-energy minima, *Proc. Natl. Acad. Sci. USA* **99**, 12562 (2002).
 - [24] A. Barducci, G. Bussi, and M. Parrinello, Well-tempered metadynamics: A smoothly converging and tunable free-energy method, *Phys. Rev. Lett.* **100**, 020603 (2008).
 - [25] R. F. Bishop and D. J. J. Farnell, Marshall-Peierls sign rules, the quantum Monte Carlo method, and frustration, in *Recent Progress in Many-Body Theories* (World Scientific, Singapore, 2000), pp. 457–460.
 - [26] T. Tieleman, Training restricted Boltzmann machines using approximations to the likelihood gradient, in *Proceedings of the 25th International Conference on Machine Learning, ICML'08* (Association for Computing Machinery, New York, NY, USA, 2008), pp. 1064–1071.
 - [27] K. Sohn and H. Lee, Learning invariant representations with local transformations, in *Proceedings of the 29th International Conference on Machine Learning, ICML'12* (Omnipress, Madison, WI, USA, 2012), pp. 1339–1346.
 - [28] M. Norouzi, M. Ranjbar, and G. Mori, Stacks of convolutional restricted Boltzmann machines for shift-invariant feature learning, in *2009 IEEE Conference on Computer Vision and Pattern Recognition* (IEEE, Miami, FL, USA, 2009), pp. 2735–2742.
 - [29] S. Sorella, Generalized Lanczos algorithm for variational quantum Monte Carlo, *Phys. Rev. B* **64**, 024512 (2001).
 - [30] S. Sorella, M. Casula, and D. Rocca, Weak binding between two aromatic rings: Feeling the van der Waals attraction by quantum Monte Carlo methods, *J. Chem. Phys.* **127**, 014105 (2007).
 - [31] W. Wirtinger, Zur formalen Theorie der Funktionen von mehr komplexen Veränderlichen, *Math. Ann.* **97**, 357 (1927).
 - [32] J. Bruna, B. Peherstorfer, and E. Vanden-Eijnden, Neural Galerkin schemes with active learning for high-dimensional evolution equations, *J. Comput. Phys.* **496**, 112588 (2024).

- [33] H. Zhang, Y. Chen, E. Vanden-Eijnden, and B. Peherstorfer, Sequential-in-time training of nonlinear parametrizations for solving time-dependent partial differential equations, SIAM review (to be published).
- [34] J. Bradbury, R. Frostig, P. Hawkins, M. J. Johnson, C. Leary, D. Maclaurin, G. Necula, A. Paszke, J. VanderPlas, S. Wanderman-Milne, and Q. Zhang, JAX composable transformations of Python+NumPy programs (2018), <http://github.com/google/jax>.
- [35] Huan Zhang, Weighted VMC, GitHub, 2025, <https://github.com/hzhang-math/WeightedVMC>.
- [36] J. Martens and R. Grosse, Optimizing neural networks with Kronecker-factored approximate curvature, in *Proceedings of the 32nd International Conference on Machine Learning*, edited by F. Bach and D. Blei, in *Proceedings of Machine Learning Research* (PMLR, Lille, France, 2015), Vol. 37, pp. 2408–2417.
- [37] A. Chen and M. Heyl, Empowering deep neural quantum states through efficient optimization, *Nat. Phys.* **20**, 1476 (2024).
- [38] G. Goldshlager, N. Abrahamsen, and L. Lin, A Kaczmarz-inspired approach to accelerate the optimization of neural network wavefunctions, *J. Comput. Phys.* **516**, 113351 (2024).
- [39] <https://github.com/hzhang-math/WeightedVMC>.