Machine Learning: Science and Technology

TOPICAL REVIEW • OPEN ACCESS

Numerical and geometrical aspects of flow-based variational quantum Monte Carlo

To cite this article: James Stokes *et al* 2023 *Mach. Learn.: Sci. Technol.* **4** 021001

IOP Publishing Mach. Learn.: Sci. Technol. 4 (2023) 021001 https://doi.org/10.1088/2632-2153/acc8b9

MACHINE LEARNING
Science and Technology

OPEN ACCESS

RECEIVED
16 September 2022

REVISED
21 March 2023

ACCEPTED FOR PUBLICATION
29 March 2023

PUBLISHED
9 May 2023

TOPICAL REVIEW

Numerical and geometrical aspects of flow-based variational quantum Monte Carlo

James Stokes[1], **Brian Chen**[2] **and Shravan Veerapaneni**[2,*]

[1] Center for Computational Quantum Physics and Center for Computational Mathematics, Flatiron Institute, New York, NY 10010 United States of America
[2] Department of Mathematics, University of Michigan, Ann Arbor, MI 48109, United States of America
* Author to whom any correspondence should be addressed.

E-mail: shravan@umich.edu

Keywords: variational Monte Carlo, normalizing flows, quantum information

Abstract

This article aims to summarize recent and ongoing efforts to simulate continuous-variable quantum systems using flow-based variational quantum Monte Carlo techniques, focusing for pedagogical purposes on the example of bosons in the field amplitude (quadrature) basis. Particular emphasis is placed on the variational real- and imaginary-time evolution problems, carefully reviewing the stochastic estimation of the time-dependent variational principles and their relationship with information geometry. Some practical instructions are provided to guide the implementation of a PyTorch code. The review is intended to be accessible to researchers interested in machine learning and quantum information science.

1. Introduction

Recent years have witnessed a profitable interchange between the fields of machine learning, quantum many-body physics and quantum information science. This multidisciplinary interaction has been partially facilitated by the discovery that artificial neural networks provide a powerful inductive bias for parametrizing subsets of quantum many-body Hilbert space. Although the description of Hilbert-space vectors via neural networks renders exact linear algebra operations prohibitive for this subset of quantum states, the existence of efficient stochastic approximation algorithms [8, 30] called variational Monte Carlo (VMC) has enabled neural-network-based quantum states (NQS) to accurately reveal properties of the ground state for quantum spin systems, as well as to simulate their time evolution using a time-dependent variant of VMC (the so-called t-VMC) [6, 7]. Since the inception of complex-valued restricted Boltzmann machines [8], the reach of neural-network quantum states has grown to encompass a diversity of quantum systems, made possible by the use of increasingly sophisticated (often multi-layered) architectures. Another driver of interaction is the discovery of close analogies between VMC and variational quantum algorithms (VQAs). Recent work on quantum information geometry in particular, Stokes *et al* [40] has clarified the connection between the natural gradient descent in machine learning [2], stochastic reconfiguration VMC [38] and variational imaginary-time evolution in quantum computing [45].

This tutorial paper is intended to serve as a self-contained review of flow-based VMC and t-VMC for continuous-variable quantum systems. For the purpose of concreteness we frame the discussion around the example of bosonic quantum systems, represented in the field amplitude basis. The field amplitude basis has not been a traditional focal point of the VMC literature[3], which has concentrated on non-relativistic systems that are more interpretable in the Fock basis. The field amplitude basis is natural, however, in systems with relativistic symmetry, where the regulated bosonic Hamiltonian is represented on an L^2-space as a simple Schrödinger operator. The simplicity of the Hamiltonian therefore also offers pedagogical advantages. A possible computational advantage of the field amplitude basis is that it does not require artificially restricting the allowed mode occupation numbers to finite range for its numerical implementation. In an effort to foster

[3] See however [18, 19, 36, 39].

further interactions between machine learning and variational quantum simulation, we chose to advance a geometrical approach based on information geometry. This contrasts with previous work on variational quantum simulation which focuses on Kähler geometry (see [17] for a review).

In a related parallel line of work, flow-based probabilistic neural networks have been applied to investigate ground state properties of lattice-regulated Euclidean quantum field theories [23]. This approach hinges on the existence of a quantum–classical mapping, under which properties of the ground state are related to classical expectation values with respect to a Gibbs distribution with an intractable normalizing factor. The aim of this approach is to accelerate, or eliminate, an otherwise expensive Markov chain Monte Carlo sampler by approximating the Gibbs measure using normalizing flows. The work described in the following aims to expand the applicability of flow-based methods to situations in which the quantum–classical mapping does not produce an efficiently-scaling sampling algorithm, which is understood to be a symptom of a so-called sign/phase problem. It is important to emphasize that sign/phase problems are not the only factors which can impact the quality of expectation values obtained via high-dimensional sampling. Flows have also proven successful in the context of path integral methods for overcoming sampling problems due to critical fluctuations at phase transitions, critical slowing down of Markov chains [1, 11, 14, 23] and frustration effects due to disorder [46]. Although VQMC does not offer an obvious advantage compared to path integrals in the absence of a sign/phase problem, one expects flow-based VQMC to inherit the same sampling advantages of path integrals. Moreover, flows have also been explored for addressing sign/phase problems within the path integral formalism by learning contour deformations [25].

It is illuminating to compare and contrast the approach advocated here with proposals in the quantum computing literature for quantum field simulation. Like [22], we adopt the field amplitude basis, although we remain agnostic about the form of the regulator since it does not alter the general structure of the resulting Schrödinger operator (see [26] for a discussion of possibilities). The choice to focus on normalizing flows is motivated by their exact sampling properties and potential to exploit distributed computing using the embarrassing parallelism of Monte Carlo [33, 46].

In an effort to make the paper accessible to a wide audience we have chosen an exposition emphasizing mathematical and numerical aspects, minimizing where possible, the necessary physics prerequisites. Although a fully rigorous treatment is possible, mathematical rigor is not attempted in this work. Detailed appendices have been included containing proofs of all relevant claims as well as additional physics motivation. Although our target application is bosons, this paper can be read independently as a self-contained review of VMC for continuous-variable quantum systems. The paper is organized as follows. In section 3 we provide a brief refresher on normalizing flows, dynamics of isolated quantum systems and time-dependent variational principles. Section 4 is devoted to explaining the geometry of quantum states and of classical probability densities (information geometry). Time-dependent variational principles are then reviewed in this context in section 5, explaining the role of normalization and holomorphy. The next section specializes to the stochastic approximation of the time-dependent variational principles using VMC and t-VMC, including a discussion of variance reduction. Section 7 comments on prospects and challenges for modeling quantum states with symmetries using equivariant flows. In section 8 we sketch a worked example in PyTorch [32] using normalizing flows to prepare bosonic ground eigenfunctions, showing improved energy compared to optimized Gaussian states. Section 9 concludes with a discussion of future directions.

2. Notation

Let $L^p(\mathbb{R}^d) := L^p(\mathbb{R}^d;\mathbb{C})$ denote the normed space of complex-valued, p-integrable functions with $p \geqslant 1$ and norm $\|\cdot\|_p$. Denote by $\langle\cdot|\cdot\rangle$ the L^2 inner product, with the convention chosen to be linear in the second argument (anti-linear in the first argument). Denote by $\mathrm{Diff}(\mathbb{R}^d)$ the group of smooth bijections from $\mathbb{R}^d$ to itself with smooth inverse and binary operation given by composition. For random vectors X and Y, valued in $\mathbb{C}^m$ and $\mathbb{C}^n$, respectively define the covariance matrix by $\mathrm{cov}(X,Y) := \mathbb{E}[XY^\dagger] - \mathbb{E}[X]\mathbb{E}[Y]^\dagger \in \mathbb{C}^{m\times n}$, where $\dagger$ denotes the conjugate transpose. Denote by $\mathbb{S}^n$, $\mathbb{S}^n_+$ and $\mathbb{S}^n_{++}$, respectively the real symmetric, symmetric positive semi-definite and symmetric positive definite matrices and similarly denote the corresponding complex Hermitian matrices as $\mathbb{H}^n, \mathbb{H}^n_+$ and $\mathbb{H}^n_{++}$. The identity matrix is denoted by $\mathbb{1}$.

3. Preliminaries

Although a formulation of normalizing flows applicable to general Riemannian manifolds is possible [35][4], the exposition is considerably simplified by restricting attention to flows defined on Euclidean spaces. The starting point is to recall that the diffeomorphism group $\mathrm{Diff}(\mathbb{R}^d)$ acts naturally on the normed space

[4] Indeed necessary for the description of nonlinear sigma-model field theories.

$L^p(\mathbb{R}^d)$. In particular, a function $\psi \in L^p(\mathbb{R}^d)$ transforms under a diffeomorphism $f \in \mathrm{Diff}(\mathbb{R}^d)$ to $f \cdot \psi \in L^p(\mathbb{R}^d)$ defined for all $x \in \mathbb{R}^d$ by,

$$(f \cdot \psi)(x) := \left|\det J_{f^{-1}}(x)\right|^{1/p} \psi\big(f^{-1}(x)\big), \tag{1}$$

in such a way that the p-norm is preserved,

$$\|f \cdot \psi\|_p = \|\psi\|_p. \tag{2}$$

The action of $\mathrm{Diff}(\mathbb{R}^d)$ is compatible with the group operation in the sense that for all $f, g \in \mathrm{Diff}(\mathbb{R}^d)$,

$$f \cdot (g \cdot \psi) = (f \circ g) \cdot \psi. \tag{3}$$

Specializing to $p = 1$, a normalized, non-negative (probability) density $p \in L^1(\mathbb{R}^d)$ is carried under the action of f to a normalized probability density $f \cdot p$. The above observation has motivated the investigation of normalizing flows for probabilistic modeling [16, 34]. The aim of this endeavor is approximate a complex, multi-modal target probability density, starting from a simple normalized probability density p (such as a multi-variate Gaussian), by choosing the diffeomorphism amongst a parametrized differentiable family of invertible multi-layer neural networks. If $\theta \in \mathbb{R}^n$ denotes the parameters indexing a family of diffeomorphisms f_θ, then density estimation using normalizing flows is to be performed under the hypothesis $\{p_\theta : \theta \in \mathbb{R}^n\}$ where $p_\theta := f_\theta \cdot p$.

The successful application of normalizing flows to density estimation tasks has motivated their generalization to quantum applications. In particular, setting $p = 2$, normalized densities are replaced by normalizable wavefunctions $\psi \in L^2(\mathbb{R}^d)$ subject to the norm constraint $\|\psi\|_2 = 1$ and the diffeomorphism group acts on the Hilbert space $L^2(\mathbb{R}^d)$ by unitary transformations [9, 44]. In particular, the linear operator $U_f : L^2(\mathbb{R}^d) \to L^2(\mathbb{R}^d)$ defined for each diffeomorphism $f \in \mathrm{Diff}(\mathbb{R}^d)$ by $U_f : \psi \mapsto f \cdot \psi$ is unitary. This fact, which follows immediately from (2) using the polarization identity, can also be seen as a consequence of the following identity for all $\psi, \psi' \in L^2(\mathbb{R}^d)$,

$$\big\langle \psi \,\big|\, U_f(\psi') \big\rangle = \big\langle U_f^{-1}(\psi) \,\big|\, \psi' \big\rangle. \tag{4}$$

A differentiable family of unit-normalized L^2 functions $\psi_\theta := f_\theta \cdot \psi$ is obtained by choosing a base function $\psi \in L^2(\mathbb{R}^d)$ satisfying $\|\psi\|_2 = 1$. Following von Neumann, the space of states for a quantum system described via the Hilbert space $L^2(\mathbb{R}^d)$ is given by set of rank-1 orthogonal projection operators (projectors). The projector P_ψ onto the complex line spanned by the unit vector $\psi \in L^2(\mathbb{R}^d)$ is a linear map defined for all $\psi' \in L^2(\mathbb{R}^d)$ by $P_\psi : \psi' \mapsto \langle \psi | \psi' \rangle \psi$. As we shall review in the subsequent sections, this projector viewpoint of quantum states turns out to be the most natural one for discussing time-dependent variational principles.

The subject of this paper is isolated quantum systems whose dynamics is dictated by a known time-independent Hermitian Hamiltonian operator H. If the system is initialized in the state P_ψ, then the uninterrupted time evolution of the system is described by the sequence of projectors $[0, \infty) \ni t \mapsto P_{\exp(-iHt)\psi}$. It is likewise instructive to consider the sequence of states $[0, \infty) \ni t \mapsto P_{\exp(-Ht)\psi}$ which correspond to unphysical evolution along the imaginary time axis. An important application of imaginary-time evolution is to preparation of a ground eigenfunction, since the component $\psi_\perp$ of ψ lying orthogonal to the ground space experiences exponential damping relative to the parallel component $\psi_\parallel$.

Given initial parameters $\theta_0 \in \mathbb{R}^n$, a time-dependent variational principle is a proposed sequence of parameters $(\theta_t)_{t \geqslant 0}$ such that $(P_{\psi_{\theta_t}})_{t \geqslant 0}$ optimally describes the exact projector evolution $(P_{\exp(-At)\psi_0})_{t \geqslant 0}$ where A denotes either H or iH, for imaginary- or real-time evolution, respectively. In practice, the sequence of parameters is defined implicitly by a system of ordinary differential equations, which must be approximated, leading to accumulation of error with evolution time, in addition to the systematic bias originating in the finite capacity of the variational family. In the case of imaginary time evolution, error accumulation is not of concern if the ultimate goal is preparation of a ground eigenfunction for a Hamiltonian with bounded-below energy, since the Rayleigh-Ritz principle ensures that any trial wavefunction provides an upper bound for the exact energy. Indeed energy optimization using neural-network-based quantum states is an active field of research and a number of proposals have been put forward including stochastic sign gradient descent [28] and Gauss–Newton [42].

4. Information geometry

Although we will ultimately be concerned with the geometry of quantum states, it is illuminating to first consider the geometry of classical probability densities. Indeed, these notions will be seen to coincide for a

wide family of quantum states. There is a natural distance metric on the set of normalized probability densities called the Fisher–Rao distance. Given probability densities $p, q \in L^1(\mathbb{R}^d)$ the Fisher–Rao distance is defined by

$$d_{\mathrm{FR}}(p,q) := \arccos\left(\left\langle p^{1/2} \middle| q^{1/2}\right\rangle\right), \tag{5}$$

where $p^{1/2} \in L^2(\mathbb{R}^d)$ denotes the pointwise square root. The Fisher–Rao distance is manifestly invariant under arbitrary diffeomorphisms $f \in \mathrm{Diff}(\mathbb{R}^d)$,

$$d_{\mathrm{FR}}(p,q) = d_{\mathrm{FR}}(f\cdot p, f\cdot q). \tag{6}$$

In order to expose the Riemannian structure underling d_{FR}, it is useful to consider a parametrized differentiable family of probability densities $\{p_\theta : \theta \in \mathbb{R}^n\}$. Define the symmetric positive semi-definite information matrix $I(\theta) \in \mathbb{S}^n_+$ for all $\theta \in \mathbb{R}^n$,

$$I(\theta) = I[p_\theta] := \mathop{\mathbb{E}}_{x\sim p_\theta}\left[\nabla_\theta \log p_\theta(x)\, \nabla_\theta \log p_\theta(x)^T\right], \tag{7}$$

which is invariant under diffeomorphisms $f \in \mathrm{Diff}(\mathbb{R}^d)$,

$$I[f\cdot p_\theta] = I[p_\theta]. \tag{8}$$

In addition, the information matrix transforms as a covariant tensor under diffeomorphisms of the parameter manifold. In particular, for a diffeomorphism $\phi \in \mathrm{Diff}(\mathbb{R}^n)$, define $\widetilde{p}_\theta := p_{\phi(\theta)}$ and then

$$\widetilde{I}(\theta) := I[\widetilde{p}_\theta] = J_\phi(\theta)^T I(\phi(\theta)) J_\phi(\theta). \tag{9}$$

The above observations concerning the information matrix are in accord with the fact that it provides the coefficient matrix for the infinitesimal line element obtained by restricting the Fisher–Rao distance to the parametric family,

$$d^2_{\mathrm{FR}}(p_\theta, p_{\theta+\mathrm{d}\theta}) = \frac{1}{4}\sum_{\mu,\nu=1}^{n} I_{\mu\nu}(\theta)\mathrm{d}\theta^\mu \mathrm{d}\theta^\nu. \tag{10}$$

Although the information matrix is not a Riemannian metric tensor since it fails the requirement of non-degeneracy, it is however, the pull-back of a Riemannian metric tensor on the infinite-dimensional manifold of strictly positive probability densities[5].

In passing to quantum state space, the natural distance function generalizes to Fubini–Study, which assigns a distance between the projectors onto unit vectors $\psi, \psi' \in L^2(\mathbb{R}^d)$ as follows,

$$d_{\mathrm{FS}}(P_\psi, P_{\psi'}) := \arccos(|\langle\psi|\psi'\rangle|). \tag{11}$$

The Fubini–Study distance inherits the diffeomorphism invariance of Fisher–Rao, and expands it to all unitary transformations $U : L^2(\mathbb{R}^d) \to L^2(\mathbb{R}^d)$,

$$d_{\mathrm{FS}}(P_\psi, P_{\psi'}) = d_{\mathrm{FS}}\left(P_{U(\psi)}, P_{U(\psi')}\right). \tag{12}$$

Paralleling the discussion for probability densities above, now consider a differentiable family of unit-normalized L^2 functions $\{\psi_\theta : \theta \in \mathbb{R}^n\}$ and define the quantum geometric tensor [5] for all $\theta \in \mathbb{R}^n$ as the Hermitian positive semi-definite matrix $G(\theta) \in \mathbb{H}^n_+$ with components,

$$G_{\mu\nu}(\theta) := G_{\mu\nu}[\psi_\theta] := \left\langle \frac{\partial\psi_\theta}{\partial\theta^\mu} \middle| \frac{\partial\psi_\theta}{\partial\theta^\nu}\right\rangle - \left\langle \frac{\partial\psi_\theta}{\partial\theta^\mu} \middle| \psi_\theta\right\rangle \left\langle \psi_\theta \middle| \frac{\partial\psi_\theta}{\partial\theta^\nu}\right\rangle, \qquad 1 \leqslant \mu,\nu \leqslant n \tag{13}$$

whose tensorial property is confirmed by elementary calculus. The quantum geometric tensor is manifestly invariant under unitary transformations,

[5] The so-called the Fisher–Rao metric tensor, which is uniquely characterized by the requirement of diffeomorphism invariance [4]. The Fisher–Rao distance is the geodesic distance function corresponding to the Fisher–Rao metric tensor.

$$G[U(\psi_\theta)] = G[\psi_\theta] \tag{14}$$

and is additionally invariant under a local phase transformation of the parametrized family which has no classical analogue. In particular, for any differentiable function $\omega : \mathbb{R}^n \to \mathbb{R}$,

$$G[\exp(i\omega(\theta))\psi_\theta] = G[\psi_\theta]. \tag{15}$$

The real part of the quantum geometric tensor, $g(\theta) := \mathrm{Re}[G(\theta)]$ is necessarily a real symmetric positive semi-definite matrix[6] $g(\theta) \in \mathbb{S}^n_+$. In direct analogy to the information matrix (7), the matrix $g(\theta)$ is the coefficient matrix for the infinitesimal line element obtained by restricting the Fubini–Study distance to the parametric family

$$d^2_{\mathrm{FS}}(P_{\psi_\theta}, P_{\psi_{\theta+\mathrm{d}\theta}}) = \sum_{\mu,\nu=1}^{n} g_{\mu\nu}(\theta)\mathrm{d}\theta^\mu \mathrm{d}\theta^\nu. \tag{16}$$

By an abuse of terminology, we will refer to $g(\theta)$ as a metric tensor. In the special case when $\psi_\theta(x)$ is real-valued, classical and quantum information geometry coincide in the sense that

$$g[\psi_\theta] = \frac{1}{4} I[\psi^2_\theta]. \tag{17}$$

4.1. Unnormalized wavefunctions and holomorphy

It is often convenient to represent the unit-normalized family of functions ψ_θ via another family of functions $\Psi_\theta \in L^2(\mathbb{R}^d)$ whose normalization is unknown,

$$\psi_\theta(x) = \frac{\Psi_\theta(x)}{\sqrt{\langle \Psi_\theta | \Psi_\theta \rangle}}. \tag{18}$$

If the quantum geometric tensor is expressed in terms of Ψ_θ one obtains the following useful expression,

$$G_{\mu\nu}(\theta) = \frac{1}{\langle \Psi_\theta | \Psi_\theta \rangle} \left\langle \frac{\partial \Psi_\theta}{\partial \theta^\mu} \middle| \frac{\partial \Psi_\theta}{\partial \theta^\nu} \right\rangle - \frac{1}{\langle \Psi_\theta | \Psi_\theta \rangle^2} \left\langle \frac{\partial \Psi_\theta}{\partial \theta^\mu} \middle| \Psi_\theta \right\rangle \left\langle \Psi_\theta \middle| \frac{\partial \Psi_\theta}{\partial \theta^\nu} \right\rangle, \qquad 1 \leqslant \mu, \nu \leqslant n. \tag{19}$$

In many important applications, the unnormalized function Ψ_θ is parametrized in terms of an even number $n = 2m$ of real parameters $\theta = \theta_1 \oplus \theta_2$ satisfying the following differential identities which correspond to the Cauchy–Riemann equations for the components of the complex vector $z := \theta_1 + i\theta_2 \in \mathbb{C}^m$,

$$\nabla_{\theta_2} \Psi_\theta = i \nabla_{\theta_1} \Psi_\theta, \tag{20}$$

which is equivalent to the vanishing of the Wirtinger gradient of Ψ_θ with respect to $\bar{z}$,

$$\nabla_{\bar{z}} \Psi_\theta := \frac{1}{2} \left(\nabla_{\theta_1} + i \nabla_{\theta_2} \right) \Psi_\theta = 0. \tag{21}$$

In this case the quantum geometric tensor and its real part are respectively given by,

$$G(\theta) = \begin{bmatrix} S(z) & iS(z) \\ -iS(z) & S(z) \end{bmatrix}, \qquad g(\theta) = \begin{bmatrix} \mathrm{Re}[S(z)] & -\mathrm{Im}[S(z)] \\ \mathrm{Im}[S(z)] & \mathrm{Re}[S(z)] \end{bmatrix}, \tag{22}$$

where $S(z) \in \mathbb{H}^m_+$ is the Hermitian postive semi-definite sub-block of the quantum geometric tensor corresponding to the θ_1 axis, whose components can be expressed in terms of Wirtinger derivatives as follows,

$$S_{ij}(z) = \frac{1}{\langle \Psi_\theta | \Psi_\theta \rangle} \left\langle \frac{\partial \Psi_\theta}{\partial z^i} \middle| \frac{\partial \Psi_\theta}{\partial z^j} \right\rangle - \frac{1}{\langle \Psi_\theta | \Psi_\theta \rangle^2} \left\langle \frac{\partial \Psi_\theta}{\partial z^i} \middle| \Psi_\theta \right\rangle \left\langle \Psi_\theta \middle| \frac{\partial \Psi_\theta}{\partial z^j} \right\rangle, \qquad 1 \leqslant i, j \leqslant m \tag{23}$$

[6] Which, incidentally, equals four times the quantum Fisher information matrix for pure states.

5. Time-dependent variational principles

With the above information-geometric preliminaries we are prepared to discuss variational principles. Consider the evolution of the unit-normalized function ψ_θ under the operator $\exp(-A\delta t)$ where A denotes either the Hermitian Hamiltonian H or the skew-Hermitian operator iH and $\delta t \geqslant 0$ denotes the evolution time. The unconstrained evolution $\psi_\theta \mapsto \exp(-A\delta t)\psi_\theta$ generically produces a function outside of the set $\{\psi_\theta\}_{\theta\in\mathbb{R}^n}$. Now consider the constrained evolution $\psi_\theta \mapsto \psi_{\theta+\delta\theta}$ induced by a parameter shift $\delta\theta \in \mathbb{R}^n$. The optimal shift should be chosen to minimize the Fubini–Study distance between the associated projectors [8]. Thus, given an initial parameter vector $\theta_0 \in \mathbb{R}^n$ and a step size $\delta t > 0$, define a sequence of parameter vectors $(\theta_k)_{k\in\mathbb{N}}$ by the following iteration

$$\theta_{k+1} = \theta_k + \delta\theta_k \tag{24}$$

$$\delta\theta_k := \underset{\delta\theta\in\mathbb{R}^n}{\arg\min}\, d_{\mathrm{FS}}\left(P_{\exp(-A\delta t)\psi_{\theta_k}}, P_{\psi_{\theta_k+\delta\theta}}\right). \tag{25}$$

In the limit $\delta t \to 0$, the sequence approximates the solution of a system of ordinary differential equations with initial condition $\theta(0) = \theta_0$. In particular, for $A = H$ and $A = iH$ we obtain respectively [3, 40],

$$g(\theta(t))\dot\theta(t) = \begin{cases} -\mathrm{Re}[F(\theta(t))] \\ \mathrm{Im}[F(\theta(t))] \end{cases} \tag{26}$$

where we have defined the following complex vector $F(\theta) \in \mathbb{C}^n$ for all $\theta \in \mathbb{R}^n$,

$$F(\theta) := \langle\nabla_\theta\psi_\theta|H\psi_\theta\rangle - \langle\nabla_\theta\psi_\theta|\psi_\theta\rangle\langle\psi_\theta|H\psi_\theta\rangle. \tag{27}$$

The arguments leading to the above evolution equations are reviewed in appendix D. The same equations are obtained by applying McLachlan's variational principle to the Liouville-von Neumann equation restricted to pure states [45]. McLachlan's variational principle applied to the time-dependent Schrödinger equation, however, yields a different set of evolution equations [29, 45][7]. Using the fact that $\langle\psi_\theta|H\psi_\theta\rangle \in \mathbb{R}$ and $\langle\nabla_\theta\psi_\theta|\psi_\theta\rangle \in i\mathbb{R}$, it follows from (26) that the imaginary-time evolution equation coincides with Riemannian gradient flow in the geometry induced by the Fubini–Study metric [40],

$$g(\theta(t))\dot\theta(t) = -\nabla\mathcal{L}(\theta(t)), \qquad \mathcal{L}(\theta) := \frac{1}{2}\langle\psi_\theta|H\psi_\theta\rangle, \tag{28}$$

which implies, as a trivial consequence of the positive semi-definiteness condition $g(\theta) \in \mathbb{S}^n_+$, that the energy $\mathcal{L}(\theta)$ is non-increasing under imaginary-time evolution,

$$\dot{\mathcal{L}}(t) = -\dot\theta(t)^T g(\theta(t))\,\dot\theta(t) \leqslant 0. \tag{29}$$

Furthermore, in the special case of real-valued ψ_θ, (17) implies that imaginary-time evolution is a special case of natural gradient flow [2]. Finally, since $g(\theta)$ is typically a degenerate metric as discussed in section 4 regularization is typically required in order to obtain a well-posed system of ordinary differential equations.

5.1. Unnormalized wavefunctions and holomorphy

In the case of unnormalized wavefunctions, the expression for $F(\theta)$ becomes,

$$F(\theta) = \frac{\langle\nabla_\theta\Psi_\theta|H\Psi_\theta\rangle}{\langle\Psi_\theta|\Psi_\theta\rangle} - \frac{\langle\nabla_\theta\Psi_\theta|\Psi_\theta\rangle}{\langle\Psi_\theta|\Psi_\theta\rangle}\frac{\langle\Psi_\theta|H\Psi_\theta\rangle}{\langle\Psi_\theta|\Psi_\theta\rangle}. \tag{30}$$

Suppose, in addition, that $n = 2m$ is even and that the holomorphic constraints (20) are satisfied. Let us overload the notation by denoting $F(z) \in \mathbb{C}^m$ the subvector of $F(\theta) \in \mathbb{C}^n$ along the θ_1 axis. Then it follows from (20) and (30) that,

$$F(\theta) = \begin{bmatrix} F(z) \\ -iF(z) \end{bmatrix}. \tag{31}$$

[7] See [45, appendices A.2.2 and B.2.2] and [45, appendices A.1.2 and B.1.2] for the variational forms of the von Neumann and time-dependent Schrödinger equation, respectively. Further clarification about these variational principles is presented in appendix A.

It can then be shown that, under the holomorphic assumption, the evolution equation (26) reduce to the following differential equations for the complex vector $z \in \mathbb{C}^m$,

$$S(z(t))\dot{z}(t) = \begin{cases} -F(z(t)) \\ -iF(z(t)) \end{cases} \tag{32}$$

The instantaneous rate of change of the loss function $\mathcal{L}$ under the evolution equation (32) is given by

$$\dot{\mathcal{L}}(t) = \begin{cases} -\dot{z}(t)^\dagger S(z(t))\dot{z}(t) \\ 0 \end{cases} \tag{33}$$

Recalling the Hermitian positive semi-definiteness of $S(z) \in \mathbb{H}^m_+$, the above equations imply that the energy is non-increasing or conserved under the imaginary- or real-time evolution respectively.

6. Stochastic estimation

6.1. VMC and t-VMC

Now we discuss numerical solution of (26) using stochastic estimation, assuming an efficient algorithm to compute $x \mapsto [\psi_\theta(x), \nabla_\theta \psi_\theta(x), (H\psi_\theta)(x)]$ and an efficient algorithm to generate unbiased samples according to the probability density $|\psi_\theta(x)|^2$. Most literature on VMC and time-dependent VMC [6–8, 21] has focused on the assumption of an efficient mapping $x \mapsto [\Psi_\theta(x), \nabla_\theta \Psi_\theta(x), (H\Psi_\theta)(x)]$ and approximate sampling from the density $|\psi_\theta(x)|^2$ using Markov Chain Monte Carlo[8]. Define the Born probability density $\rho_\theta(x) \in [0, \infty)$, the wavefunction score $\sigma_\theta(x) \in \mathbb{C}^n$ and the local energy $l_\theta(x) \in \mathbb{C}$ as follows,

$$\rho_\theta(x) := |\psi_\theta(x)|^2, \qquad \sigma_\theta(x) := \frac{\nabla_\theta \psi_\theta(x)}{\psi_\theta(x)}, \qquad l_\theta(x) := \frac{(H\psi_\theta)(x)}{\psi_\theta(x)}. \tag{34}$$

It is then a simple exercise to confirm that the quantities $g(\theta)$ and $F(\theta)$, $\mathcal{L}(\theta)$ and $\nabla\mathcal{L}(\theta)$ can be expressed as the expectation values of random variables with respect to the Born probability density. In particular,

$$G(\theta) = \mathrm{cov}(\sigma_\theta, \sigma_\theta)^T, \qquad F(\theta) = \mathrm{cov}(l_\theta, \sigma_\theta)^T. \tag{35}$$

The expressions (35) provide the basis for a stochastic approximate solution of (26) called stochastic reconfiguration and t-VMC, respectively, which use Monte Carlo methods to approximate the covariances, in combination with a suitable time-marching scheme (e.g. forward Euler). In the special case of imaginary-time evolution, the algorithm is known as stochastic reconfiguration [38]. If, in addition $\psi_\theta \in \mathbb{R}$ for all $\theta \in \mathbb{R}^n$ then, stochastic reconfiguration becomes stochastic natural gradient descent for the objective $\mathcal{L}(\theta)$.

6.2. Stochastic optimization and variance reduction

Stochastic reconfiguration can be understood as a special case of stochastic gradient-based optimization for the objective $\mathcal{L}(\theta)$. To see this, let $\hat{\mathcal{L}}_\theta(x)$ be any unbiased estimator for the loss function so that

$$\mathcal{L}(\theta) = \mathbb{E}\left[\hat{\mathcal{L}}_\theta(x)\right]. \tag{36}$$

Then by the log-derivative trick,

$$\nabla\mathcal{L}(\theta) = \mathbb{E}\left[\left(\hat{\mathcal{L}}_\theta(x)\mathbb{1} - \frac{B}{2}\right)\nabla_\theta \log \rho_\theta(x)\right] + \mathbb{E}\left[\nabla_\theta \hat{\mathcal{L}}_\theta(x)\right], \tag{37}$$

where $B \in \mathbb{R}^{n\times n}$ is an arbitrary matrix and we have used the fact that $\mathbb{E}[\nabla_\theta \log \rho_\theta(x)] = 0$. The canonical estimator which has been pursued most widely in the literature corresponds to the local energy defined in (34),

$$\hat{\mathcal{L}}^{(\mathrm{can})}_\theta(x) := \frac{1}{2} l_\theta(x). \tag{38}$$

[8] See however [18, 36, 44] for continuum and [20, 37] for discrete quantum systems, respectively.

The canonical estimator has a number of desirable properties including the fact the gradient of the objective (37) becomes independent of the gradient of the estimator. Specifically, plugging (38) into (37) and using Hermiticity of H we obtain,

$$\nabla\mathcal{L}(\theta) = \mathrm{Re}\,\mathbb{E}\big[\big(l_\theta(x)\mathbb{1} - B\big)\overline{\sigma_\theta}(x)\big], \tag{39}$$

which coincides with stochastic reconfiguration for the choice of baseline $B = \mathbb{E}[l_\theta(x)]$. In addition, the variance of the stochastic objective function using the canonical estimator is proportional to the quantum variance of the Hamiltonian in the quantum state P_{ψ_θ},

$$\mathrm{var}(l_\theta) := \mathrm{cov}(l_\theta, l_\theta) = \langle\psi_\theta|H\psi_\theta\rangle - \langle\psi_\theta|H^2\psi_\theta\rangle, \tag{40}$$

which follows from Hermiticity of H. The canonical estimator thus has the desirable property that its variance approaches zero when ψ_θ approaches any eigenfunction of H. This zero-variance principle is an attractive feature of stochastic reconfiguration compared to other stochastic optimization problems, and can be exploited when numerically approaching a ground eigenfunction. It turns out that the canonical estimator is not the only estimator available in continuous-variable VMC, however, and we will discuss one such alternative in the experiments section.

For normalized trial functions, it has been shown empirically [18, 20, 37, 44] that the stochastic reconfiguration choice of baseline $B = \mathbb{E}[l_\theta(x)]$ has reduced variance compared to vanishing baseline. In order to better understand this variance reduction property, let us define the following gradient estimator with arbitrary baseline B,

$$\hat{\nabla}_{\theta,B}(x) := \mathrm{Re}\,\big[\big(l_\theta(x)\mathbb{1} - B\big)\overline{\sigma_\theta}(x)\big]. \tag{41}$$

Anticipating the variance-reduction property of the baseline, introduce a convex objective function for the matrix B; namely, the total variation of the random vector $\hat{\nabla}_{\theta,B}$,

$$V(B) := \mathrm{tr}\,\big[\mathrm{cov}(\hat{\nabla}_{\theta,B})\big]. \tag{42}$$

The stationary points of (42) describe a linear system of equations for the matrix B. In appendix B, it is shown that the system can be solved under the ad-hoc assumption that l_θ and σ_θ are statistically independent under ρ_θ. In particular, one finds that B is an approximate multiple of the identity matrix, $B \approx \mathbb{E}[l_\theta(x)]\mathbb{1}$. This provides a heuristic justification for the variance reduction procedure utilized in [18, 20].

7. Additional comments on quantum flows

7.1. Group equivariant dynamics from flows

An attractive feature of normalizing flows is that they are compatible with group symmetries in the following sense.

Lemma 7.1. *Let $G \leqslant O(d,\mathbb{R})$ be an orthogonal matrix group and $\rho : G \to \mathbb{C}^\times$ a one-dimensional representation of G. Suppose that $0 \neq \psi \in L^p(\mathbb{R}^d)$ and $f \in \mathrm{Diff}(\mathbb{R}^d)$ transform as*

$$\psi(gx) = \rho(g)\,\psi(x), \qquad f(gx) = gf(x), \qquad \text{for all } g \in G \text{ and } x \in \mathbb{R}^d. \tag{43}$$

Then ρ is a unitary representation and $f\cdot\psi \in L^p(\mathbb{R}^d)$ transforms as

$$(f\cdot\psi)(gx) = \rho(g)\,(f\cdot\psi)(x), \qquad \text{for all } g \in G \text{ and } x \in \mathbb{R}^d. \tag{44}$$

The proof is an elementary generalization of [31, lemma 1]. The above lemma has important implications for quantum simulation because it enables to model a flexible family of normalized G-equivariant functions using a simple normalized G-equivariant base function $\psi \in L^2(\mathbb{R}^d)$ and a family of G-equivariant diffeomorphisms. The set of such G-equivariant functions forms a Hilbert subspace of $L^2(\mathbb{R}^d)$, which is stable under real- or imaginary time evolution by a G-equivariant Hamiltonian H satisfying,

$$U_g \circ H = H \circ U_g, \qquad \text{for all } g \in G, \tag{45}$$

where the unitary operator U_g acts on the Hilbert space as $U_g(\psi) = \psi\,(g^{-1}x)$ for all $\psi \in L^2(\mathbb{R}^d)$ (recalling that $|\det g| = 1$). If combined with the time-dependent variational principle, the above construction enables the approximation of real- or imaginary-time dynamics within the G-equivariant subspace of states. In other words, flows can describe dynamically closed superselection sectors of the Hilbert space. This possibility was recently investigated in [44] using the sign representation of a permutation subgroup $G \leqslant O(d,\mathbb{R})$ for the purpose of modeling spinless fermions.

7.2. Universal approximation

Although universal approximation has been proven under the assumption of strict positivity [31], complex-valued flows pose additional subtleties. In particular, as already noted in the context of first-quantized fermions [44], normalizing flows cannot change the topology of the zero level set for the base function and similarly for the level sets of complex phase. More precisely, the level sets are diffeomorphic,

$$L_0(\mathrm{mod} f\cdot\psi) = f\big(L_0(\mathrm{mod}\,\psi)\big), \qquad \forall\theta\in[0,2\pi): L_\theta(\arg f\cdot\psi) = f\big(L_\theta(\arg\psi)\big), \tag{46}$$

where mod denotes the complex modulus and $L_c(\cdot)$ denotes the level set of a function corresponding to the real value $c\in\mathbb{R}$. Although the above identities pose an obstruction to universal approximation, this limitation can be easily overcome in practice by promoting the base to a trainable function, or by multiplying the output of the flow with a learnable complex phase.

8. Experiments

8.1. States and Hamiltonian

Let $p=(1/i)\nabla$ denote the momentum operator canonically conjugate to x and consider the following Hamiltonian operator represented on a suitable subspace of the Hilbert space $L^2(\mathbb{R}^d)$ by

$$H = \frac{1}{2}\begin{bmatrix}x\\p\end{bmatrix}\cdot h\begin{bmatrix}x\\p\end{bmatrix} + \frac{1}{4!}\sum_{ijkl}\lambda_{ijkl}x_ix_jx_kx_l, \tag{47}$$

where $h\in\mathbb{S}^{2d}$ is a symmetric matrix, λ_{ijkl} are the components of a real tensor $\lambda\in\mathbb{R}^{d\times d\times d\times d}$ and $\cdot$ denotes the standard inner product for Euclidean space $\mathbb{R}^{2d}$. The physical significance of the above Hamiltonian is that it describes an indefinite number of bosons occupying d possible modes[9]. Additional assumptions about h and λ are required to ensure the existence of a ground eigenfunction. If, for example $\lambda=0$, then we require $h\in\mathbb{S}^{2d}_{++}$ to be symmetric positive definite. In this case, the exact ground eigenfunction is represented by a family of trial functions $\psi_{\mathrm{G}}\in L^2(\mathbb{R}^d)$ of the following Gaussian form[10],

$$\psi_{\mathrm{G}}(x) = \left(\det\frac{A}{\pi}\right)^{1/4}\exp\left[-\frac{1}{2}(x-\mu)^T(A+iB)(x-\mu)\right], \tag{48}$$

where the variational parameters are constrained to the manifold $(\mu,A,B)\in\mathbb{R}^d\times\mathbb{S}^d_{++}\times\mathbb{S}^d$. Consider the following block decomposition of the symmetric matrix $h\in\mathbb{S}^{2d}$

$$h = \begin{bmatrix}h_{xx} & h_{xp}\\ h_{px} & h_{pp}\end{bmatrix}, \tag{49}$$

where $h_{px}=h_{xp}^T$ and where $h_{xx},h_{pp}\in\mathbb{S}^d$ are symmetric.

The experiments focus on the problem of approximating a ground eigenfunction via the method of natural gradient descent, using a non-canonical estimator of the gradient obtained using an adjoint representation of the Hamiltonian, the details of which are deferred to appendix H. For simplicity, we exemplify the method in the special case of $h_{pp}=\mathbb{1}$ and $h_{xp}=0$, where it should be stressed that this example is only for illustrative purposes since quantum Monte Carlo does not suffer from a sign-problem when $h_{xp}=0$. In addition, since we are targeting the ground eigenfunction rather than the full imaginary-time trajectory, we may employ any step size schedule consistent with decreasing energy. In particular, we combined the natural gradient method with the Adam optimizer. In order to ensure boundedness of the Hamiltonian from below, we chose the interaction tensor $\lambda_{ijkl}=3\delta_{ij}\delta_{kl}u_{ik}$ for some symmetric positive definite matrix $u\in\mathbb{S}^d_{++}$. The resulting Hamiltonian has a unique ground space spanned by a strictly positive eigenfunction [15, section 3.3]. By positivity of the ground function, if the Hamiltonian is G-equivariant for some $G\leqslant O(d,\mathbb{R})$, then it follows that the ground eigenfunction transforms in the trivial representation. Thus, provided that G is a finite group of computationally tractable order, a trial function for the G-invariant ground eigenfunction can be chosen as the square root of a mixture density formed by symmetrizing a classical normalizing flow p_θ as follows,

$$\psi_\theta(x)^2 = \frac{1}{|G|}\sum_{g\in G}p_\theta(gx). \tag{50}$$

[9] A motivating example from field theory is provided in appendix E.

[10] See appendix C for an example in one dimension.

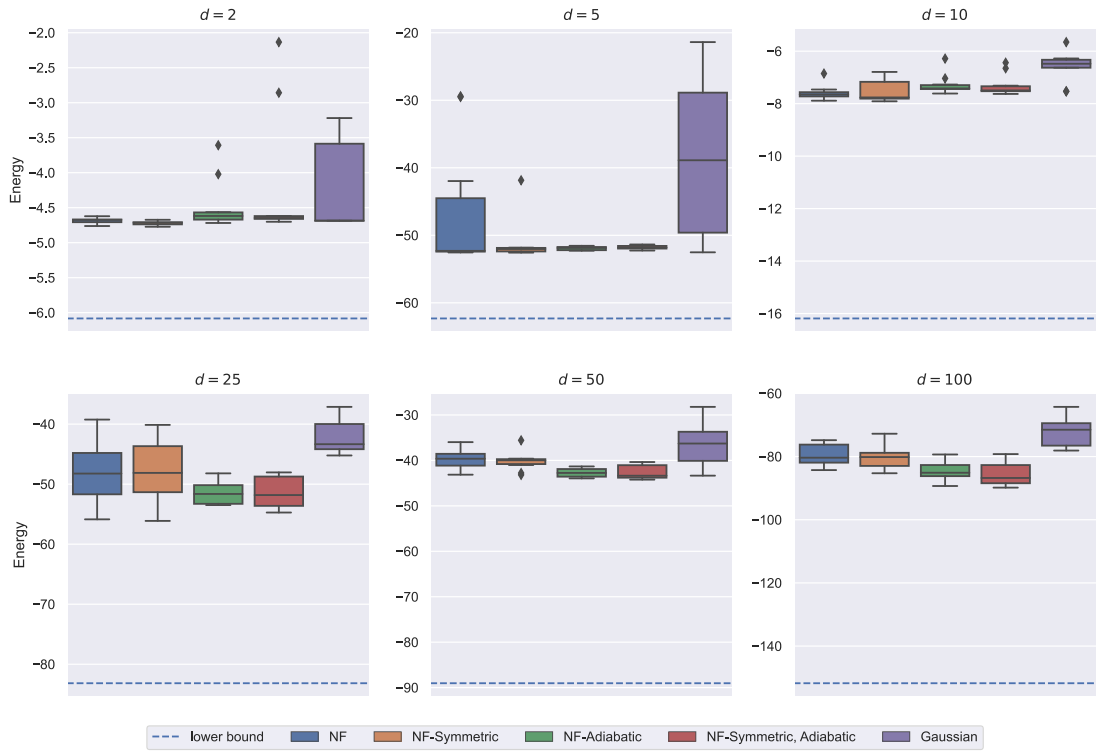

Figure 1. A boxplot showing the energies found via normalizing flows and Gaussian state approximation (results shown for ten random initializations). 'NF' and 'NF-Symmetric' refer to the unsymmetrized and symmetrized normalizing flows, respectively (see (50)). The results labeled with 'adiabatic' are trained using adiabatic retraining, as described in appendix G.3. Although the true ground state energy of these systems is unknown, a lower bound on the energies is plotted for comparison—the calculation of this lower bound is described in appendix I.

In the experiments the matrices u and h_{xx} were chosen randomly to reflect our agnosticism about the nature of the regularization. It follows that the only remaining symmetry of the Hamiltonian is $G = \mathbb{Z}_2$, whose non-trivial element is implemented by field space inversion $x \mapsto -x$. Since $\mathbb{Z}_2$ is a finite group of order 2, the mixture density approach is computationally efficient in this case.

8.2. Experimental setup

Recall that under the above simplifications the Hamiltonian is specified by a symmetric matrix $h_{xx} \in \mathbb{S}^d$ and a symmetric positive-definite matrix $u \in \mathbb{S}^d_{++}$. For each problem dimension $d \in \{2, 5, 10, 25, 50, 100\}$ a random Hamiltonian was selected by choosing the h and λ parameters in (47) subject to the constraints described in the previous subsection, specifically that h is a symmetric block diagonal matrix of the form $h = \mathrm{diag}(h_{xx}, \mathbb{1})$ with h_{xx} symmetric negative definite[11] and that $\lambda_{ijkl} = 3\delta_{ij}\delta_{kl}u_{ik}$ for symmetric positive definite $u \in \mathbb{S}^d_{++}$. The relevant matrix ensembles and sampling procedures are described in appendix G.2.

For each randomly selected Hamiltonian, the ground eigenfunction is approximated by representing it using a $\mathbb{Z}_2$-symmetrical trial function of the form (50) where p_θ was chosen to be a RealNVP-based normalizing flow [12] with variational parameters $\theta \in \mathbb{R}^n$. Starting from random initialization, the parameters θ of the neural network were updated using stochastic natural gradient-based optimization of the loss function (36), employing the adjoint representation of the Hamiltonian. Stochastic estimates of the gradient on each iteration were obtained using PyTorch to back-propagate (36), without the use of baseline adjustment. Additional details on the normalizing flow architecture and training procedure can be found in appendix G.1. Following the recent work of [16], adiabatic retraining was also explored (appendix G.3 for additional detail).

As a baseline, the results are compared to a real-valued Gaussian trial state (48), for which the variational energy can be computed analytically (see appendix G.4). Optimization over the parameter manifold $(\mu, A) \in \mathbb{R}^d \times \mathbb{S}^d_{++}$ was performed using the Pymanopt toolbox [41].

[11] That is, $-h_{xx} \in \mathbb{S}^d_{++}$.

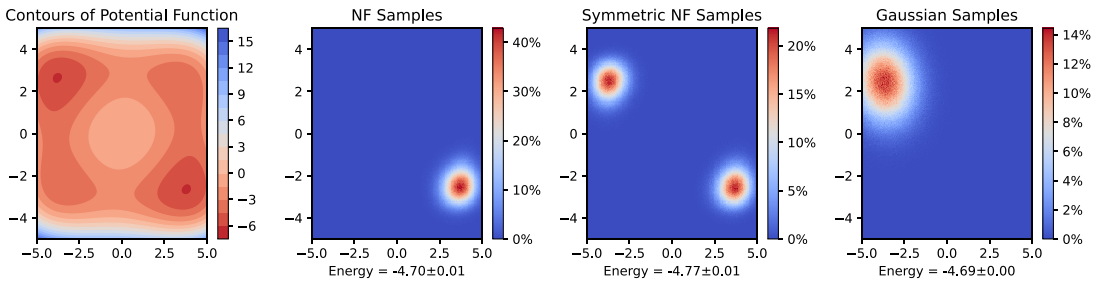

Figure 2. From left to right: contours of the potential function and density plots from the corresponding optimized wave functions for $d = 2$ (unsymmetrized normalizing flows, symmetric normalizing flows, and Gaussian wave function approximation).

8.3. Results

The energies found via the normalizing flow approaches and the Gaussian approximation are shown in figure 1. We can infer that the normalizing flow methods generally lead to lower estimates of the ground state energy than the Gaussian state approximation. In addition, compared to unsymmetrized flows, symmetrized flows with adiabatic retraining yield lower energies in higher dimensions. In figure 2, the potential function and the approximations of the ground-state probability density found via normalizing flows or the Gaussian wave approximation are visualized for problem dimension $d = 2$. Notice that while the symmetrized normalizing flow is able to find the two expected modes for the given the potential function, the unsymmetrized flow and Gaussian approximation collapse around one of the modes.

9. Discussion and future directions

In summary, flow-based parametrizations provide a promising class of trial wavefunctions for use in the continuous-variable VMC, although a number of challenges remain to be solved. Let us conclude by summarizing open problems. An obvious limitation of flow-based parametrizations is their lack of holomorphy, which is required to ensure energy conservation during variational real-time evolution. It would be very interesting to reconcile the holomorphic constraint with the exact sampling property of flows in the continuum. Since VMC methods suffer from finite sampling effects, it will be interesting to undertake finite-sample analysis of the different gradient estimators and to further explore variance reduction strategies. It would also be interesting to better understand the approximation power of G-equivariant flows.

It will be natural to extend the experiments to address the sign-problem, including the determination of ground states for non-stoquastic systems characterized by $h_{xp} \neq 0$, or real-time evolution. In particular, it would also be interesting to numerically investigate the extent to which energy conservation is violated in practice.

An important outstanding challenge is to find physical applications of the scheme which exhibit an advantage compared to the best known algorithm. Aspirational targets include charged scalar fields at finite chemical potential [43] or variational calculation of scattering [26].

Data availability statement

All data that support the findings of this study are included within the article (and any supplementary files).

Acknowledgments

J S thanks Giuseppe Carleo, Damian Hofmann, Di Luo, Matija Medvidović, Jannes Nys, and Gabriel Pescia for helpful discussions. Authors gratefully acknowledge support from NSF under Grant DMS-2038030. This research was supported in part through computational resources and services provided by the Advanced Research Computing (ARC) at the University of Michigan.

Appendix A. McLachlan's variational principles

This section constructs an (admittedly contrived) example demonstrating the failure of McLachlan's variational principle for the time-dependent Schödinger equation. This example complements

[45, section 4.2] which considered real time evolution of a single qubit. Consider the quantum simple harmonic oscillator

$$H = \frac{1}{2}\left(p^2 + x^2\right), \tag{51}$$

initialized in the state given by projection onto the ground eigenfunction,

$$\psi_0(x) = \frac{1}{\pi^{1/4}} \exp\left[-\frac{1}{2}x^2\right] \tag{52}$$

The exact projector dynamics under real-time evolution is clearly given by the trivial constant sequence $(P_{\psi_0})_{t\geqslant 0}$. Now consider the following normalized variational family

$$\psi_\theta(x) = \left(\frac{a}{\pi}\right)^{1/4} \exp\left[-\frac{1}{2}(a+ib)x^2\right], \tag{53}$$

parametrized in terms of $\theta := (\log a, b) \in \mathbb{R}^2$, which ensures the positivity constraint $a > 0$. Recall that variational Liouville-von Neumann equation and the variational time-dependent Schrödigner equation (TDSE) are given, respectively by

$$g(\theta(t))\dot{\theta}(t) = \mathrm{Im}[F(\theta(t))], \qquad \widetilde{g}(\theta(t))\dot{\theta}(t) = \mathrm{Im}[\widetilde{F}(\theta(t))], \tag{54}$$

where

$$F_\mu(\theta) = \left\langle \frac{\partial\psi_\theta}{\partial\theta^\mu} | H\psi_\theta \right\rangle - \left\langle \frac{\partial\psi_\theta}{\partial\theta^\mu} | \psi_\theta \right\rangle \langle \psi_\theta | H\psi_\theta \rangle \qquad \widetilde{F}_\mu(\theta) = \left\langle \frac{\partial\psi_\theta}{\partial\theta^\mu} | H\psi_\theta \right\rangle \tag{55}$$

$$g_{\mu\nu}(\theta) = \mathrm{Re}\left[\left\langle \frac{\partial\psi_\theta}{\partial\theta^\mu} \middle| \frac{\partial\psi_\theta}{\partial\theta^\nu} \right\rangle - \left\langle \frac{\partial\psi_\theta}{\partial\theta^\mu} \middle| \psi_\theta \right\rangle \left\langle \psi_\theta \middle| \frac{\partial\psi_\theta}{\partial\theta^\nu} \right\rangle\right] \qquad \widetilde{g}_{\mu\nu}(\theta) = \mathrm{Re}\left[\left\langle \frac{\partial\psi_\theta}{\partial\theta^\mu} \middle| \frac{\partial\psi_\theta}{\partial\theta^\nu} \right\rangle\right]. \tag{56}$$

In the example above we compute

$$\langle \psi_\theta | H\psi_\theta \rangle = \frac{1 + a^2 + b^2}{4a^2} \tag{57}$$

$$\mathrm{Im}[\widetilde{F}(\theta)] = \begin{bmatrix} \frac{b}{4} \\ -\frac{1}{16} + \frac{3(1+b^2)}{16a^2} \end{bmatrix} \tag{58}$$

$$\mathrm{Im}[F(\theta)] = \begin{bmatrix} \frac{b}{4} \\ -\frac{1}{16} + \frac{3(1+b^2)}{16a^2} \end{bmatrix} - \begin{bmatrix} 0 \\ \frac{1}{16} + \frac{1+b^2}{16a^2} \end{bmatrix} \tag{59}$$

$$= \begin{bmatrix} \frac{b}{4} \\ -\frac{1}{8} + \frac{1+b^2}{8a^2} \end{bmatrix} \tag{60}$$

$$\widetilde{g}(\theta) = \begin{bmatrix} \frac{1}{8} & 0 \\ 0 & \frac{3}{16a^2} \end{bmatrix} \tag{61}$$

$$g(\theta) = \begin{bmatrix} \frac{1}{8} & 0 \\ 0 & \frac{3}{16a^2} \end{bmatrix} - \begin{bmatrix} 0 & 0 \\ 0 & \frac{1}{16a^2} \end{bmatrix} \tag{62}$$

$$= \begin{bmatrix} \frac{1}{8} & 0 \\ 0 & \frac{1}{8a^2} \end{bmatrix}. \tag{63}$$

Putting the above pieces together we obtain the following variational Liouville-von Neumann equation

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{bmatrix} \log a \\ b \end{bmatrix} = \begin{bmatrix} 2b \\ 1 - a^2 + b^2 \end{bmatrix} \tag{64}$$

and the following variational TDSE,

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{bmatrix} \log a \\ b \end{bmatrix} = \begin{bmatrix} 2b \\ 1 - \frac{1}{3}a^2 + b^2 \end{bmatrix}. \tag{65}$$

If the system (64) is initialized at the origin of the $\theta = (\log a, b) \in \mathbb{R}^2$ coordinates, then it will clearly reproduce the exact projector dynamics since the right-hand side of (64) vanishes. In contrast, it can be verified that there is no choice of initialization which produces the exact projector under (65).

Appendix B. Justification for choice of baseline

In this appendix argue that the stationary points of (42) are approximately solved by the stochastic reconfiguration baseline. Starting from the definition of the loss function (42) and recalling that $\hat{\nabla}_{\theta,B}(x)$ is real-valued,

$$V(B) = \mathrm{tr}\left[\mathrm{cov}\left(\hat{\nabla}_{\theta,B}\right)\right] \tag{66}$$

$$= \mathrm{tr}\left\{\mathbb{E}\left[\hat{\nabla}_{\theta,B}(x)\otimes\hat{\nabla}_{\theta,B}(x)\right] - \mathbb{E}\left[\hat{\nabla}_{\theta,B}(x)\right]\otimes\mathbb{E}\left[\hat{\nabla}_{\theta,B}(x)\right]\right\} \tag{67}$$

$$= \mathrm{tr}\left\{\mathbb{E}\left[\hat{\nabla}_{\theta,B}(x)\otimes\hat{\nabla}_{\theta,B}(x)\right] - \nabla\mathcal{L}(\theta)\otimes\nabla\mathcal{L}(\theta)\right\} \tag{68}$$

$$= \mathbb{E}\left[\hat{\nabla}_{\theta,B}(x)^T\hat{\nabla}_{\theta,B}(x)\right] - \nabla\mathcal{L}(\theta)^T\nabla\mathcal{L}(\theta). \tag{69}$$

Now using the fact that the second term above is manifestly independent of B we obtain,

$$\frac{\partial V(B)}{\partial B} = \frac{\partial}{\partial B}\mathbb{E}\left[\hat{\nabla}_{\theta,B}(x)^T\hat{\nabla}_{\theta,B}(x)\right] \tag{70}$$

$$= -2\mathbb{E}\left[\hat{\nabla}_{\theta,B}(x)\otimes\mathrm{Re}\left[\overline{\sigma_\theta}(x)\right]\right] \tag{71}$$

$$= -2\mathbb{E}\left[\mathrm{Re}\left[\left(l_\theta(x)-B\right)\overline{\sigma_\theta}(x)\right]\otimes\mathrm{Re}\left[\overline{\sigma_\theta}(x)\right]\right] \tag{72}$$

$$= -2\mathbb{E}\left[\mathrm{Re}\left[\left(\overline{l_\theta}(x)-B\right)\sigma_\theta(x)\right]\otimes\mathrm{Re}\left[\sigma_\theta(x)\right]\right]. \tag{73}$$

Setting the gradient to zero we obtain,

$$B\,\mathbb{E}\left[\mathrm{Re}[\sigma_\theta(x)]\otimes\mathrm{Re}[\sigma_\theta(x)]\right] = \mathbb{E}\left[\mathrm{Re}[\overline{l_\theta}(x)\sigma_\theta(x)]\otimes\mathrm{Re}[\sigma_\theta(x)]\right]. \tag{74}$$

Now the right-hand side is

$$\mathbb{E}\left[\mathrm{Re}[\overline{l_\theta}(x)\sigma_\theta(x)]\otimes\mathrm{Re}[\sigma_\theta(x)]\right] = \frac{1}{2}\mathbb{E}\left[\left(\overline{l_\theta}(x)\sigma_\theta(x)+l_\theta(x)\overline{\sigma_\theta}(x)\right)\otimes\mathrm{Re}[\sigma_\theta(x)]\right] \tag{75}$$

$$= \frac{1}{2}\mathbb{E}\left[\overline{l_\theta}(x)\sigma_\theta(x)\otimes\mathrm{Re}[\sigma_\theta(x)] + l_\theta(x)\overline{\sigma_\theta}(x)\otimes\mathrm{Re}[\sigma_\theta(x)]\right] \tag{76}$$

$$\approx \frac{1}{2}\mathbb{E}\left[\overline{l_\theta}(x)\right]\mathbb{E}\left[\sigma_\theta(x)\otimes\mathrm{Re}[\sigma_\theta(x)]\right] + \frac{1}{2}\mathbb{E}\left[l_\theta(x)\right]\mathbb{E}\left[\overline{\sigma_\theta}(x)\otimes\mathrm{Re}[\sigma_\theta(x)]\right] \tag{77}$$

$$= \frac{1}{2}\mathbb{E}\left[l_\theta(x)\right]\mathbb{E}\left[\left(\sigma_\theta(x)+\overline{\sigma_\theta}(x)\right)\otimes\mathrm{Re}[\sigma_\theta(x)]\right] \tag{78}$$

$$= \mathbb{E}\left[l_\theta(x)\right]\mathbb{E}\left[\mathrm{Re}[\sigma_\theta(x)]\otimes\mathrm{Re}[\sigma_\theta(x)]\right], \tag{79}$$

where in the third equality we have assumed that l_θ and σ_θ are approximately independent under ρ_θ. Plugging back into the right-hand side of (74) we find the approximate solution $B \approx \mathbb{E}[l_\theta(x)]\mathbb{1}$.

Appendix C. Quadratic Hamiltonian in one dimension

In one problem dimension ($d = 1$) and in the limit of vanishing interaction potential,

$$H = \frac{1}{2}\begin{bmatrix}x\\p\end{bmatrix}\cdot\begin{bmatrix}h_{xx} & h_{xp}\\ h_{xp} & h_{pp}\end{bmatrix}\begin{bmatrix}x\\p\end{bmatrix} \tag{80}$$

where $h_{xx}h_{pp} - (h_{xp})^2 > 0$ and the ground eigenfunction is of the form (53) with

$$a = \frac{\sqrt{h_{xx}h_{pp}-(h_{xp})^2}}{h_{pp}} \tag{81}$$

$$b = \frac{h_{xp}}{h_{pp}}. \tag{82}$$

Appendix D. Derivation of time-dependent variational principles

In this section we review the derivation of (26), synthesizing the results of [3, 40]. In the following summation over repeated indices is implied. Denote $\widetilde{\psi}_\theta := e^{-A\delta t}\psi_\theta$. Then

$$\arg\min_{\delta\theta\in\mathbb{R}^n} d_{\mathrm{FS}}(P_{\widetilde{\psi}_\theta}, P_{\psi_{\theta+\delta\theta}}) = \arg\max_{\delta\theta\in\mathbb{R}^n} \frac{\left|\left\langle\widetilde{\psi}_\theta|\psi_{\theta+\delta\theta}\right\rangle\right|}{|\langle\widetilde{\psi}_\theta|\widetilde{\psi}_\theta\rangle\langle\psi_{\theta+\delta\theta}|\psi_{\theta+\delta\theta}\rangle|} = \arg\max_{\delta\theta\in\mathbb{R}^n}\left|\left\langle\widetilde{\psi}_\theta|\psi_{\theta+\delta\theta}\right\rangle\right|^2, \tag{83}$$

where we used the monotonicity of elementary functions and the normalization of $\psi_{\theta+\delta\theta}$. We have

$$\langle\widetilde{\psi}_\theta|\psi_{\theta+\delta\theta}\rangle = \langle\widetilde{\psi}_\theta|\psi_\theta\rangle + \left\langle\widetilde{\psi}_\theta\middle|\frac{\partial\psi_\theta}{\partial\theta^\mu}\right\rangle\delta\theta^\mu + \frac{1}{2}\left\langle\widetilde{\psi}_\theta\middle|\frac{\partial^2\psi_\theta}{\partial\theta^\mu\partial\theta^\nu}\right\rangle\delta\theta^\mu\delta\theta^\nu + \cdots. \tag{84}$$

So Taylor expanding $|\langle\widetilde{\psi}_\theta|\psi_{\theta+\delta\theta}\rangle|^2$ to quadratic order in the displacement gives,

$$\begin{aligned}|\langle\widetilde{\psi}_\theta|\psi_{\theta+\delta\theta}\rangle|^2 = |\langle\widetilde{\psi}_\theta|\psi_\theta\rangle|^2 &+ \left[\langle\psi_\theta|\widetilde{\psi}_\theta\rangle\left\langle\widetilde{\psi}_\theta\middle|\frac{\partial\psi_\theta}{\partial\theta^\mu}\right\rangle + \left\langle\frac{\partial\psi_\theta}{\partial\theta^\mu}\middle|\widetilde{\psi}_\theta\right\rangle\langle\widetilde{\psi}_\theta|\psi_\theta\rangle\right]\delta\theta^\mu \\ &+ \left[\left\langle\frac{\partial\psi_\theta}{\partial\theta^\mu}\middle|\widetilde{\psi}_\theta\right\rangle\left\langle\widetilde{\psi}_\theta\middle|\frac{\partial\psi_\theta}{\partial\theta^\nu}\right\rangle + \frac{1}{2}\langle\psi_\theta|\widetilde{\psi}_\theta\rangle\left\langle\widetilde{\psi}_\theta\middle|\frac{\partial^2\psi_\theta}{\partial\theta^\mu\partial\theta^\nu}\right\rangle + \frac{1}{2}\left\langle\frac{\partial^2\psi_\theta}{\partial\theta^\mu\theta^\nu}\middle|\widetilde{\psi}_\theta\right\rangle\langle\widetilde{\psi}_\theta|\psi_\theta\rangle\right] \\ &\times\delta\theta^\mu\delta\theta^\nu + \cdots\end{aligned} \tag{85}$$

Expanding the exponential $e^{-A\delta t}$ in δt and neglecting cubic-order terms in the multi-variable Taylor expansion in $\delta\theta$ and δt,

$$|\langle\widetilde{\psi}_\theta|\psi_{\theta+\delta\theta}\rangle|^2 = |\langle\widetilde{\psi}_\theta|\psi_\theta\rangle|^2 - 2\,\mathrm{Re}\left[\left\langle\frac{\partial\psi_\theta}{\partial\theta^\mu}\middle|A\psi_\theta\right\rangle + \langle A\psi_\theta|\psi_\theta\rangle\left\langle\frac{\partial\psi_\theta}{\partial\theta^\mu}\middle|\psi_\theta\right\rangle\right]\delta\theta^\mu\delta t - \mathrm{Re}[G_{\mu\nu}(\theta)]\delta\theta^\mu\delta\theta^\nu + \cdots, \tag{86}$$

The first-order optimality condition $0 = \frac{\partial}{\partial\delta\theta^\mu}|\langle\widetilde{\psi}_\theta|\psi_{\theta+\delta\theta}\rangle|^2$, at lowest order in $\delta\theta$ and δt, thus gives

$$0 = -\,\mathrm{Re}\left[\left\langle\frac{\partial\psi_\theta}{\partial\theta^\mu}\middle|A\psi_\theta\right\rangle + \langle A\psi_\theta|\psi_\theta\rangle\left\langle\frac{\partial\psi_\theta}{\partial\theta^\mu}\middle|\psi_\theta\right\rangle\right]\delta t - \mathrm{Re}[G_{\mu\nu}(\theta)]\delta\theta^\nu + \cdots. \tag{87}$$

In the limit $\delta t\to 0$, neglecting higher order terms gives (26).

Appendix E. Field theory motivation

In this section we use the heuristic arguments to motivate a Hamiltonian of the form (47). Consider the Hamiltonian for a neutral scalar field with quartic self-coupling defined, for simplicity, on the compact interval[12] $I = [0, L]$ with periodic boundary conditions

$$H = \int_I \mathrm{d}x\, h(x), \qquad h(x) = \frac{1}{2}\left[\pi(x)^2 + \phi'(x)^2 + \mu\phi(x)^2\right] + \frac{\lambda}{4!}\phi(x)^4, \tag{88}$$

where the field amplitude and the momentum density satisfy $[\phi(x), \pi(y)] = i\delta(x-y)$ for $x, y\in I$. If we now expand the field operators ϕ and π using an orthonormal system of periodic real-valued functions $(f_i)_{i\in\mathbb{N}}$,

$$\phi(x) = \sum_{i\in\mathbb{N}}\hat{\phi}_i f_i(x), \qquad \pi(x) = \sum_{i\in\mathbb{N}}\hat{\pi}_i f_i(x), \tag{89}$$

then we find that the operator-valued coefficients $\hat{\phi}_i$ and $\hat{\pi}_i$ satisfy the standard canonical commutation relations $[\hat{\phi}_i, \hat{\pi}_j] = i\delta_{ij}$. Substituting back into the Hamiltonian, using orthonormality and truncating the infinite series using the dth partial sums, we obtain a regulated Hamiltonian which is represented on $L^2(\mathbb{R}^d)$ by the Schrödinger operator (47) with the following identifications

$$h = \begin{bmatrix}\mu\mathbb{1} + \alpha & 0\\ 0 & \mathbb{1}\end{bmatrix}, \qquad \alpha_{ij} = \int_I \mathrm{d}x f_i'(x) f_j'(x), \qquad \lambda_{ijkl} = \lambda\int_I \mathrm{d}x f_i(x) f_j(x) f_k(x) f_l(x) \tag{90}$$

Repeating the above analysis for a charged scalar field with nonzero chemical potential we obtain a similiar Hamiltonian in which the off-diagonal blocks of h are nonzero.

[12] Compactness is convenient because it avoids the possibility of spontaneous symmetry breakdown.

Appendix F. Technical proofs

Proof of (3). Let $f, g \in \mathrm{Diff}(\mathbb{R}^d)$ and $\psi \in L^p(\mathbb{R}^d)$. Then

$$(f \circ g) \cdot \psi(x) = \left|\det J_{g^{-1} \circ f^{-1}}(x)\right|^{1/p} \psi\left(g^{-1} \circ f^{-1}(x)\right) \tag{91}$$

$$= \left|\det J_{g^{-1}}(f^{-1}(x)) \det J_{f^{-1}}(x)\right|^{1/p} \psi\left(g^{-1} \circ f^{-1}(x)\right) \tag{92}$$

$$= \left|\det J_{f^{-1}}(x)\right|^{1/p} \left|\det J_{g^{-1}}(f^{-1}(x))\right|^{1/p} \psi\left(g^{-1} \circ f^{-1}(x)\right) \tag{93}$$

$$= f \cdot (g \cdot \psi)(x) \tag{94}$$

□

Proof of (4). Suppose that $f \in \mathrm{Diff}(\mathbb{R}^d)$ and let $y = f(x)$ denote the image of $x \in \mathbb{R}^d$. By the inverse function theorem, $J_{f^{-1}}(y) = J_f(x)^{-1}$, so $\left|\det J_{f^{-1}}(y)\right| = \left|\det J_f(x)\right|^{-1}$. Now ($p = 2$ here)

$$\langle \psi | f \cdot \psi' \rangle = \int_{\mathbb{R}^d} \mathrm{d}y\, \overline{\psi(y)} \left|\det J_{f^{-1}}(y)\right|^{1/2} \psi'\left(f^{-1}(y)\right) \tag{95}$$

$$= \int_{\mathbb{R}^d} \mathrm{d}x \left|\det J_f(x)\right| \overline{\psi(f(x))} \left|\det J_f(x)\right|^{-1/2} \psi'(x) \tag{96}$$

$$= \int_{\mathbb{R}^d} \mathrm{d}y\, \overline{\left|\det J_f(x)\right|^{1/2} \psi(f(x))} \psi'(x) \tag{97}$$

$$= \left\langle f^{-1} \cdot \psi | \psi' \right\rangle. \tag{98}$$

Now by (3),

$$U_f(U_{f^{-1}}(\psi)) = f \cdot (f^{-1} \cdot \psi) = (f \circ f^{-1})(\psi) = \psi \tag{99}$$

and likewise $U_{f^{-1}}(U_f(\psi)) = \psi$ so $U_{f^{-1}} = U_f^{-1}$.

□

Proof of (8). Recalling the definition (1) of $f \cdot p_\theta$ (for $p = 1$) and using the fact that $f \in \mathrm{Diff}(\mathbb{R}^d)$ is independent of θ,

$$I[f \cdot p_\theta] = \int_{\mathbb{R}^d} \mathrm{d}x\, (f \cdot p_\theta)(x) \left[\nabla_\theta \log (f \cdot p_\theta)(x)\, \nabla_\theta \log (f \cdot p_\theta)(x)^T\right] \tag{100}$$

$$= \int_{\mathbb{R}^d} \mathrm{d}x \left|\det J_{f^{-1}}(x)\right| p_\theta(f^{-1}(x)) \left[\nabla_\theta \log p_\theta(f^{-1}(x))\, \nabla_\theta \log p_\theta(f^{-1}(x))^T\right] \tag{101}$$

$$= I[p_\theta] \tag{102}$$

□

Proof of (9). By the chain rule

$$\frac{\partial}{\partial \theta^\mu} \log \widetilde{p}_\theta(x) = \sum_{\nu=1}^n \frac{\partial \phi^\nu(\theta)}{\partial \theta^\mu} \left.\frac{\partial}{\partial \theta^\nu}\right|_{\phi(\theta)} \log p_\theta(x) \tag{103}$$

$$= \sum_{\nu=1}^n (J_\phi)^\nu{}_\mu(\theta) \left.\frac{\partial}{\partial \theta^\nu}\right|_{\phi(\theta)} \log p_\theta(x) \tag{104}$$

and thus

$$\widetilde{I}[\theta] := \mathop{\mathbb{E}}_{x \sim \widetilde{p}_\theta} \left[\nabla_\theta \log \widetilde{p}_\theta(x)\, \nabla_\theta \log \widetilde{p}_\theta(x)^T\right] \tag{105}$$

$$= J_\phi(\theta)^T I(\phi(\theta)) J_\phi(\theta), \tag{106}$$

□

Proof of (15). Let $\widetilde{\psi}_\theta := e^{i\omega(\theta)}\psi_\theta$. Then

$$\left\langle \frac{\partial\widetilde{\psi}_\theta}{\partial\theta^\mu}\middle|\frac{\partial\widetilde{\psi}_\theta}{\partial\theta^\nu}\right\rangle = \left\langle \frac{\partial\psi_\theta}{\partial\theta^\mu}+i\frac{\partial\omega}{\partial\theta^\mu}\psi_\theta\middle|\frac{\partial\psi_\theta}{\partial\theta^\nu}+i\frac{\partial\omega}{\partial\theta^\nu}\psi_\theta\right\rangle \tag{107}$$

$$= \left\langle \frac{\partial\psi_\theta}{\partial\theta^\mu}\middle|\frac{\partial\psi_\theta}{\partial\theta^\nu}\right\rangle + \frac{\partial\omega}{\partial\theta^\mu}\frac{\partial\omega}{\partial\theta^\nu} - i\frac{\partial\omega}{\partial\theta^\mu}\left\langle \psi_\theta\middle|\frac{\partial\psi_\theta}{\partial\theta^\nu}\right\rangle + i\frac{\partial\omega}{\partial\theta^\nu}\left\langle \frac{\partial\psi_\theta}{\partial\theta^\mu}\middle|\psi_\theta\right\rangle, \tag{108}$$

and

$$\left\langle \frac{\partial\widetilde{\psi}_\theta}{\partial\theta^\mu}\middle|\widetilde{\psi}_\theta\right\rangle = \left\langle \frac{\partial\psi_\theta}{\partial\theta^\mu}+i\frac{\partial\omega}{\partial\theta^\mu}\psi_\theta\middle|\psi_\theta\right\rangle \tag{109}$$

$$= \left\langle \frac{\partial\psi_\theta}{\partial\theta^\mu}\middle|\psi_\theta\right\rangle - i\frac{\partial\omega}{\partial\theta^\mu}, \tag{110}$$

so

$$\left\langle \frac{\partial\widetilde{\psi}_\theta}{\partial\theta^\mu}\middle|\widetilde{\psi}_\theta\right\rangle\left\langle \widetilde{\psi}_\theta\middle|\frac{\partial\widetilde{\psi}_\theta}{\partial\theta^\nu}\right\rangle = \left\langle \frac{\partial\psi_\theta}{\partial\theta^\mu}\middle|\psi_\theta\right\rangle\left\langle \psi_\theta\middle|\frac{\partial\psi_\theta}{\partial\theta^\nu}\right\rangle + \frac{\partial\omega}{\partial\theta^\mu}\frac{\partial\omega}{\partial\theta^\nu} - i\frac{\partial\omega}{\partial\theta^\mu}\left\langle \psi_\theta\middle|\frac{\partial\psi_\theta}{\partial\theta^\nu}\right\rangle + i\frac{\partial\omega}{\partial\theta^\nu}\left\langle \frac{\partial\psi_\theta}{\partial\theta^\mu}\middle|\psi_\theta\right\rangle. \tag{111}$$

So

$$\left\langle \frac{\partial\widetilde{\psi}_\theta}{\partial\theta^\mu}\middle|\frac{\partial\widetilde{\psi}_\theta}{\partial\theta^\nu}\right\rangle - \left\langle \frac{\partial\widetilde{\psi}_\theta}{\partial\theta^\mu}\middle|\widetilde{\psi}_\theta\right\rangle\left\langle \widetilde{\psi}_\theta\middle|\frac{\partial\widetilde{\psi}_\theta}{\partial\theta^\nu}\right\rangle = \left\langle \frac{\partial\psi_\theta}{\partial\theta^\mu}\middle|\frac{\partial\psi_\theta}{\partial\theta^\nu}\right\rangle - \left\langle \frac{\partial\psi_\theta}{\partial\theta^\mu}\middle|\psi_\theta\right\rangle\left\langle \psi_\theta\middle|\frac{\partial\psi_\theta}{\partial\theta^\nu}\right\rangle. \tag{112}$$

□

Proof of (19) and (30). Details are provided only for (19) since (30) follows similarly. By the chain rule

$$\frac{\partial\psi_\theta}{\partial\theta^\mu} = \frac{1}{\langle\Psi_\theta|\Psi_\theta\rangle^{1/2}}\left[\frac{\partial\Psi_\theta}{\partial\theta^\mu} - \frac{1}{2}\frac{\Psi_\theta}{\langle\Psi_\theta|\Psi_\theta\rangle}\frac{\partial\langle\Psi_\theta|\Psi_\theta\rangle}{\partial\theta^\mu}\right]. \tag{113}$$

So

$$\left\langle \frac{\partial\psi_\theta}{\partial\theta^\mu}\middle|\frac{\partial\psi_\theta}{\partial\theta^\nu}\right\rangle = \frac{1}{\langle\Psi_\theta|\Psi_\theta\rangle}\left\langle \frac{\partial\Psi_\theta}{\partial\theta^\mu} - \frac{1}{2}\frac{\Psi_\theta}{\langle\Psi_\theta|\Psi_\theta\rangle}\frac{\partial\langle\Psi_\theta|\Psi_\theta\rangle}{\partial\theta^\mu}\middle|\frac{\partial\Psi_\theta}{\partial\theta^\nu} - \frac{1}{2}\frac{\Psi_\theta}{\langle\Psi_\theta|\Psi_\theta\rangle}\frac{\partial\langle\Psi_\theta|\Psi_\theta\rangle}{\partial\theta^\nu}\right\rangle \tag{114}$$

$$= \frac{1}{\langle\Psi_\theta|\Psi_\theta\rangle}\left[\left\langle \frac{\partial\Psi_\theta}{\partial\theta^\mu}\middle|\frac{\partial\Psi_\theta}{\partial\theta^\nu}\right\rangle + \frac{1}{4}\frac{1}{\langle\Psi_\theta|\Psi_\theta\rangle}\frac{\partial\langle\Psi_\theta|\Psi_\theta\rangle}{\partial\theta^\mu}\frac{\partial\langle\Psi_\theta|\Psi_\theta\rangle}{\partial\theta^\nu} + \right. \tag{115}$$

$$\left. -\frac{1}{2}\frac{1}{\langle\Psi_\theta|\Psi_\theta\rangle}\left\langle \frac{\partial\Psi_\theta}{\partial\theta^\mu}\middle|\Psi_\theta\right\rangle\frac{\partial\langle\Psi_\theta|\Psi_\theta\rangle}{\partial\theta^\nu} - \frac{1}{2}\frac{1}{\langle\Psi_\theta|\Psi_\theta\rangle}\left\langle \Psi_\theta\middle|\frac{\partial\Psi_\theta}{\partial\theta^\nu}\right\rangle\frac{\partial\langle\Psi_\theta|\Psi_\theta\rangle}{\partial\theta^\mu},\right] \tag{116}$$

and

$$\left\langle \frac{\partial\psi_\theta}{\partial\theta^\mu}\middle|\psi_\theta\right\rangle\left\langle \psi_\theta\middle|\frac{\partial\psi_\theta}{\partial\theta^\nu}\right\rangle = \frac{1}{\langle\Psi_\theta|\Psi_\theta\rangle^2}\left[\left\langle \frac{\partial\Psi_\theta}{\partial\theta^\mu}\middle|\Psi_\theta\right\rangle - \frac{1}{2}\frac{\partial\langle\Psi_\theta|\Psi_\theta\rangle}{\partial\theta^\mu}\right]\left[\left\langle \Psi_\theta\middle|\frac{\partial\Psi_\theta}{\partial\theta^\nu}\right\rangle - \frac{1}{2}\frac{\partial\langle\Psi_\theta|\Psi_\theta\rangle}{\partial\theta^\nu}\right]. \tag{117}$$

Expanding out one finds that the offending terms cancel and we obtain:

$$\left\langle \frac{\partial\psi_\theta}{\partial\theta^\mu}\middle|\frac{\partial\psi_\theta}{\partial\theta^\nu}\right\rangle - \left\langle \frac{\partial\psi_\theta}{\partial\theta^\mu}\middle|\psi_\theta\right\rangle\left\langle \psi_\theta\middle|\frac{\partial\psi_\theta}{\partial\theta^\nu}\right\rangle = \frac{1}{\langle\Psi_\theta|\Psi_\theta\rangle}\left\langle \frac{\partial\Psi_\theta}{\partial\theta^\mu}\middle|\frac{\partial\Psi_\theta}{\partial\theta^\nu}\right\rangle - \frac{1}{\langle\Psi_\theta|\Psi_\theta\rangle^2}\left\langle \frac{\partial\Psi_\theta}{\partial\theta^\mu}\middle|\Psi_\theta\right\rangle\left\langle \Psi_\theta\middle|\frac{\partial\Psi_\theta}{\partial\theta^\nu}\right\rangle \tag{118}$$

□

Proof of (32). Using (22) and (31) we obtain the following $\theta = \theta_1 \oplus \theta_2$ decompositions,

$$g(\theta)\dot{\theta} = \begin{bmatrix} \mathrm{Re}[S(z)]\dot{\theta}_1 - \mathrm{Im}[S(z)]\dot{\theta}_2 \\ \mathrm{Im}[S(z)]\dot{\theta}_1 + \mathrm{Re}[S(z)]\dot{\theta}_2 \end{bmatrix} \tag{119}$$

$$-\mathrm{Re}[F(\theta)] = \begin{bmatrix} -\mathrm{Re}[F(z)] \\ -\mathrm{Im}[F(z)] \end{bmatrix} \tag{120}$$

$$\mathrm{Im}[F(\theta)] = \begin{bmatrix} \mathrm{Im}[F(z)] \\ -\mathrm{Re}[F(z)] \end{bmatrix} \tag{121}$$

Plug these expressions into the evolution equation (26) and consider a linear combination consisting of the θ_1 rows superposed with an imaginary unit multiplying the θ_2 rows to obtain (32). □

Proof of (33). Taking the time derivative of the loss function $\mathcal{L}$ defined in (28) and assuming that the holomorphic constraints (20) are satisfied we obtain,

$$\dot{\mathcal{L}}(\theta) = \dot{\theta}_1^T \nabla_{\theta_1} \mathcal{L}(\theta) + \dot{\theta}_2^T \nabla_{\theta_2} \mathcal{L}(\theta) \tag{122}$$

$$= \dot{\theta}_1^T \mathrm{Re}[F(z)] + \dot{\theta}_2^T \mathrm{Im}[F(z)] \tag{123}$$

$$= \mathrm{Re}[\dot{z}^\dagger F(z)] \tag{124}$$

$$= \begin{cases} -\mathrm{Re}[\dot{z}^\dagger S(z)\dot{z}] \\ \mathrm{Re}[i\dot{z}^\dagger S(z)\dot{z}] \end{cases} \tag{125}$$

$$= \begin{cases} -\dot{z}^\dagger S(z)\dot{z} \\ 0 \end{cases} . \tag{126}$$

The first equality is the chain rule. The second equality used $\nabla \mathcal{L}(\theta) = \mathrm{Re}[F(\theta)]$ together with (31). The third equality used $\dot{z} = \dot{\theta}_1 + i\dot{\theta}_2$. The fourth equality used (32) and the final equality used the fact that $S(z) \in \mathbb{H}_+^m$ is Hermitian positive semi-definite.

□

Proof of (35). Recalling our conventions for complex covariance matrices in section 2,

$$\mathrm{cov}(\sigma_\theta, \sigma_\theta) = \mathbb{E}[\sigma_\theta(x)\sigma_\theta(x)^\dagger] - \mathbb{E}[\sigma_\theta(x)]\mathbb{E}[\sigma_\theta(x)]^\dagger \tag{127}$$

$$\mathrm{cov}(\sigma_\theta, \sigma_\theta)^T = \mathbb{E}[\overline{\sigma_\theta}(x)\sigma_\theta(x)^T] - \mathbb{E}[\overline{\sigma_\theta}(x)]\mathbb{E}[\sigma_\theta(x)]^T \tag{128}$$

$$= G(\theta). \tag{129}$$

Similarly,

$$\mathrm{cov}(l_\theta, \sigma_\theta) = \mathbb{E}[l_\theta(x)\sigma_\theta(x)^\dagger] - \mathbb{E}[l_\theta(x)]\mathbb{E}[\sigma_\theta(x)]^\dagger \tag{130}$$

$$\mathrm{cov}(l_\theta, \sigma_\theta)^T = \mathbb{E}[\overline{\sigma_\theta}(x) l_\theta(x)] - \mathbb{E}[\overline{\sigma_\theta}(x)]\mathbb{E}[l_\theta(x)] \tag{131}$$

□

Proof of (39). Starting from (37) we obtain,

$$\nabla\mathcal{L}(\theta) = \mathbb{E}\left[\left(\hat{\mathcal{L}}_\theta(x)\mathbb{1} - \frac{B}{2}\right)\nabla_\theta \log\rho_\theta(x)\right] + \mathbb{E}\left[\nabla_\theta\hat{\mathcal{L}}_\theta(x)\right] \tag{132}$$

$$= \frac{1}{2}\mathbb{E}\left[(l_\theta(x)\mathbb{1} - B)\left(\overline{\sigma_\theta}(x) + \sigma_\theta(x)\right)\right] + \frac{1}{2}\mathbb{E}\left[\nabla_\theta l_\theta(x)\right], \tag{133}$$

where we have used

$$\nabla_\theta \log\rho_\theta(x) = \frac{\nabla_\theta\rho_\theta(x)}{\rho_\theta(x)} \tag{134}$$

$$= \frac{\nabla_\theta\overline{\psi_\theta}(x)\,\psi_\theta(x) + \overline{\psi_\theta}(x)\nabla_\theta\psi_\theta(x)}{|\psi_\theta(x)|^2} \tag{135}$$

$$= \overline{\sigma_\theta}(x) + \sigma_\theta(x). \tag{136}$$

Now

$$\mathbb{E}\left[\nabla_\theta l_\theta(x)\right] = \int_{\mathbb{R}^d} \mathrm{d}x|\psi_\theta(x)|^2\left[\frac{\nabla_\theta(H\psi_\theta)(x)}{\psi_\theta(x)} - \frac{(H\psi_\theta)(x)}{\psi_\theta(x)}\frac{\nabla_\theta\psi_\theta(x)}{\psi_\theta(x)}\right] \tag{137}$$

$$= \int_{\mathbb{R}^d} \mathrm{d}x\overline{\psi_\theta}(x)\nabla_\theta(H\psi_\theta)(x) - \mathbb{E}\left[l_\theta(x)\sigma_\theta(x)\right] \tag{138}$$

$$= \int_{\mathbb{R}^d} \mathrm{d}x\overline{(H\psi_\theta)}(x)\nabla_\theta\psi_\theta(x) - \mathbb{E}\left[l_\theta(x)\sigma_\theta(x)\right] \tag{139}$$

$$= \int_{\mathbb{R}^d} \mathrm{d}x|\psi_\theta(x)|^2\overline{\left[\frac{(H\psi_\theta)(x)}{\psi_\theta(x)}\right]}\frac{\nabla_\theta\psi_\theta(x)}{\psi_\theta(x)} - \mathbb{E}\left[l_\theta(x)\sigma_\theta(x)\right] \tag{140}$$

$$= \mathbb{E}\left[\overline{l_\theta}(x)\sigma_\theta(x) - l_\theta(x)\sigma_\theta(x)\right], \tag{141}$$

where we have interchanged the order of operations of the gradient ∇_θ with the Hamiltonian H and also used the fact that H is Hermitian. Thus,

$$\nabla\mathcal{L}(\theta) = \frac{1}{2}\mathbb{E}\left[l_\theta(x)\overline{\sigma_\theta}(x) + \overline{l_\theta}(x)\sigma_\theta(x)\right] - \frac{1}{2}B\mathbb{E}\left[\sigma_\theta(x) + \overline{\sigma_\theta}(x)\right] \tag{142}$$

$$= \mathrm{Re}\,\mathbb{E}\left[\left(l_\theta(x) - B\right)\overline{\sigma_\theta}(x)\right], \tag{143}$$

Alternatively, starting from the definition of the loss function (28),

$$\nabla\mathcal{L}(\theta) = \frac{\langle\nabla_\theta\psi_\theta|H\psi_\theta\rangle + \langle\psi_\theta|H\nabla_\theta\psi_\theta\rangle}{2} \tag{144}$$

$$= \frac{\langle\nabla_\theta\psi_\theta|H\psi_\theta\rangle + \langle H\psi_\theta|\nabla_\theta\psi_\theta\rangle}{2} \tag{145}$$

$$= \frac{1}{2}\langle\nabla_\theta\psi_\theta|H\psi_\theta\rangle + \text{c.c.} \tag{146}$$

$$= \frac{1}{2}\mathbb{E}\left[l_\theta(x)\overline{\sigma_\theta}(x)\right] + \text{c.c.} \tag{147}$$

$$= \mathbb{E}\,\mathrm{Re}\left[l_\theta(x)\overline{\sigma_\theta}(x)\right], \tag{148}$$

where we have used the product rule in the first equality, Hermiticity of H in the second equality and conjugate symmetry of $\langle\cdot|\cdot\rangle$ in the third equality. □

Proof of (46). By definition of the zero level set,

$$L_0(\mathrm{mod} f\cdot\psi) := \{x\in\mathbb{R}^d : |\det J_{f^{-1}}(x)|^{1/p}\,\mathrm{mod}\,\psi\big(f^{-1}(x)\big)=0\} \tag{149}$$

$$= \{x\in\mathbb{R}^d : \mathrm{mod}\,\psi\big(f^{-1}(x)\big)=0\} \tag{150}$$

$$=: (\mathrm{mod}\,\psi\circ f^{-1})^{-1}[\{0\}] \tag{151}$$

$$= f\big((\mathrm{mod}\,\psi)^{-1}[\{0\}]\big) \tag{152}$$

$$= f\big(L_0(\mathrm{mod}\,\psi)\big), \tag{153}$$

where the second equality used the fact that f is a diffeomorphism to divide out the everywhere nonzero Jacobian factor $|\det J_{f^{-1}}(x)|>0$, the third equality is by definition of the pre-image and the fourth equality follows from the property of pre-images under a composition of maps. By the same reasoning,

$$L_\theta(\arg f\cdot\psi) = \{x\in\mathbb{R}^d : \arg\psi\big(f^{-1}(x)\big)=\theta\} \tag{154}$$

$$=: (\arg\psi\circ f^{-1})^{-1}[\{\theta\}] \tag{155}$$

$$= f\big((\arg\psi)^{-1}[\{\theta\}]\big) \tag{156}$$

$$= f\big(L_\theta(\arg\psi)\big) \tag{157}$$

□

Proof of lemma 7.1. Starting with the expression $\psi(gx)=\rho(g)\psi(x)$, taking the complex modulus, raising to the pth power and integrating we obtain

$$\int_{\mathbb{R}^d}\mathrm{d}x|\psi(gx)|^p = |\rho(g)|^p\int_{\mathbb{R}^d}\mathrm{d}x|\psi(x)|^p \tag{158}$$

Now changing integration variables on the left-hand side,

$$|\det(g^{-1})|\int_{\mathbb{R}^d}\mathrm{d}x|\psi(x)|^p = |\rho(g)|^p\int_{\mathbb{R}^d}\mathrm{d}x|\psi(x)|^p. \tag{159}$$

Recalling that $\|\psi\|_p\neq 0$ and that G is an orthogonal group we conclude $|\rho(g)|=1$, as required for a one-dimensional unitary representation. Recalling the definition (1) we obtain

$$(f\cdot\psi)(gx) = \big|\det J_{f^{-1}}(gx)\big|^{1/p}\,\psi\big(f^{-1}(gx)\big). \tag{160}$$

Now $\psi\big(f^{-1}(gx)\big)=\psi\big(gf^{-1}(x)\big)=\rho(g)\psi\big(f^{-1}(x)\big)$. In addition, as shown in [31], $\big|\det J_{f^{-1}}(gx)\big| = \big|\det J_{f^{-1}}(x)\big|$ and therefore $(f\cdot\psi)(gx)=\rho(g)\,(f\cdot\psi)(x)$.

□

Appendix G. Architecture and training details

In this section, additional details on the architecture and training procedure used for the experiments are provided.

Table 1. Ablation study on some of the different options for using the natural gradient (with or without the Adam optimizer and changing the preconditioning terms γ for the Fisher information matrix). Using the standard gradient with the Adam optimizer is also considered. This ablation study uses the standard normalizing flow approach (no symmetrization or adiabatic retraining). The results shown indicate the mean ending energy across ten runs using different random initializations, with the error bounds equal to two times the standard error.

Dimension	2	5	10
Natural Gradient, $\gamma = 1.0$	-3.84 ± 0.53	-48.76 ± 4.77	-3.24 ± 0.42
Natural Gradient, $\gamma = 1.0$, Adam	$\mathbf{-4.69 \pm 0.02}$	-46.75 ± 6.12	$\mathbf{-7.58 \pm 0.18}$
Natural Gradient, $\gamma = 0.1$	-4.38 ± 0.29	-47.41 ± 5.15	-6.78 ± 0.59
Natural Gradient, $\gamma = 0.1$, Adam	-4.69 ± 0.02	$\mathbf{-50.12 \pm 3.06}$	-7.52 ± 0.26
Standard Gradient, Adam	-4.68 ± 0.02	-48.99 ± 4.82	-7.46 ± 0.26
Gaussian	-4.39 ± 0.39	-40.22 ± 7.64	-6.92 ± 0.43
Dimension	25	50	100
Natural Gradient, $\gamma = 1.0$	-35.78 ± 7.49	-28.38 ± 3.76	-75.85 ± 2.95
Natural Gradient, $\gamma = 1.0$, Adam	-47.87 ± 3.43	-39.79 ± 1.39	-79.43 ± 2.17
Natural Gradient, $\gamma = 0.1$	-46.63 ± 2.93	$\mathbf{-41.51 \pm 1.20}$	$\mathbf{-80.11 \pm 1.77}$
Natural Gradient, $\gamma = 0.1$, Adam	$\mathbf{-48.29 \pm 3.46}$	-39.91 ± 1.75	-79.66 ± 2.54
Standard Gradient, Adam	-45.98 ± 3.82	-40.97 ± 1.08	-76.51 ± 3.18
Gaussian	-42.53 ± 2.77	-35.95 ± 1.43	-72.83 ± 4.00

Note: The bold indicates the best performance amongst the set of parameters varied.

G.1. Architecture and training procedure

The training procedure requires both sampling as well as calculating the log probabilities for a given sample. As such, efficiency in both the forward and backward pass of a normalizing flow is desired, which motivates the choice of RealNVP [12] as the architecture for the flow. The architecture is modeled off the nflows package [13] and built in PyTorch [32].

The normalizing flows are trained using natural gradient descent, using the Adam optimizer [24] applied to natural gradient estimates. In the calculation of the Fisher information matrix, a preconditioning term γ of 0.1 is added to the diagonal in order to stabilize training. We start with an initial learning rate of 0.01—this initial learning rate is decayed using a cosine decay schedule with no warm restarts [27]. The per-sample gradients necessary for calculating the Fisher information matrix are calculated using the Backpack package [10]. An ablation study comparing different optimization methods is shown in table 1.

G.2. Randomly generating Hamiltonians

In this subsection, we describe the process for selecting the matrices h_{xx} and u which define the Hamiltonian. The procedure for selecting u for dimension d is as follows:

- Select d eigenvalues uniformly randomly from the interval $[0.1, 2]$. Let Σ be a $d \times d$ diagonal matrix with these eigenvalues along the diagonal.
- Sample a random matrix U from the Haar distribution over the orthogonal group in dimension d.
- Let u equal $U\Sigma U^T$.

The procedure for selecting h_{xx} is identical, except in the final step, h_{xx} is set to be equal to $-U\Sigma U^T$. This ensures that u is positive definite and h_{xx} is negative definite, while maintaining bounds on the condition number of each matrix. For each dimension d, we run each approach (normalizing flows, symmetric normalizing flows, the normalizing flows approaches with adiabatic retraining, and Gaussian state approximation) ten times using different random initializations.

G.3. Adiabatic retraining and flow-distance regularization

Following the recent work of [16], we investigate two methods for improving the final energy, namely adiabatic retraining and flow-distance regularization. Adiabatic retraining involves varying the parameters of the target objective in such a way that it interpolates between a relatively simple problem to a more complicated problem. To implement this, the quadratic term in (47) is multiplied by a term α, where α ranges from 0 to 1 during training. Following [16], α is exponentially decayed—in other words, we let

$$\alpha = \frac{e^{-kt} - e^{-k}}{1 - e^{-k}}, \tag{161}$$

where k is a hyperparameter. Unlike [16], adiabatic retraining is used during the entire training procedure, rather than for a interval in the middle of training.

Hackett *et al* [16] also introduce flow-distance regularization, which imposes a penalty on the flow for transforming samples z from the base distribution to significantly different outputs. This penalty is enforced using a l_2 norm between samples from the base distribution and the outputs. The penalty is annealed to zero by the end of training. Unlike Hackett *et al* we did not find a significant difference when using flow distance regularization.

G.4. Gaussian state approximation

As a point of comparison for the normalizing flows approach, the energy of an optimal Gaussian wave function (48) was estimated by optimizing over the variational parameters using Riemannian gradient descent. Setting $\lambda_{ijkl} = 3\delta_{ij}\delta_{kl}u_{ik}$ and $h = \mathrm{diag}(h_{xx}, \mathbb{1})$ in the (47) and setting $B = 0$ in (48) we obtain,

$$\begin{aligned}\langle\psi_G|H\psi_G\rangle =& \frac{1}{4}\,\mathrm{tr}(A) + \frac{1}{4}\mathrm{sum}\left(h_{xx}\odot A^{-1}\right) + \frac{1}{32}\left(\mathrm{diag}\left(A^{-1}\right) + 2\mu^{\odot 2}\right)^T u\left(\mathrm{diag}\left(A^{-1}\right) + 2\mu^{\odot 2}\right)\\ &+ \frac{1}{16}\mathrm{sum}\left(u\odot\left(A^{-1}\right)^{\odot 2}\right) + \frac{1}{2}\mu^T\left(h_{xx} + \frac{1}{2}uA^{-1}\right)\mu \end{aligned} \tag{162}$$

Appendix H. Loss function estimator

In this section we illustrate the adjoint loss function estimator $\hat{\mathcal{L}}_\theta^{(\mathrm{adj})}(x)$ as an alternative to the canonical estimator $\hat{\mathcal{L}}_\theta^{(\mathrm{can})}(x)$ in the simple example of the harmonic oscillator Hamiltonian (51),

$$\hat{\mathcal{L}}_\theta^{(\mathrm{can})}(x) := \frac{1}{2}l_\theta(x) \tag{163}$$

$$= \frac{1}{4}\left[-\frac{\psi_\theta''(x)}{\psi_\theta(x)} + x^2\right] \tag{164}$$

$$\hat{\mathcal{L}}_\theta^{(\mathrm{adj})}(x) = \frac{1}{4}\left[\left|\frac{\psi_\theta'(x)}{\psi_\theta(x)}\right|^2 + x^2\right]. \tag{165}$$

It should be noted that unlike the canonical estimator, the adjoint estimator does not exhibit the zero-variance property, as described in section 6.2. For the example of the simple harmonic oscillator (51), consider the variational class given by the family of wave functions (53), parametrized by $a > 0$ with b fixed to zero ($a = 1$ is the ground eigenfunction (52)). Figure 3 shows the variance of both estimators as a function of the variational parameter a, showing clearly the zero-variance principle of the canonical estimator. Although the adjoint estimator is subject to an irreducible quantum uncertainty, the standard error of the estimate can be reduced to zero at the Monte Carlo rate of $1/\sqrt{N}$ simply by increasing the number of samples N.

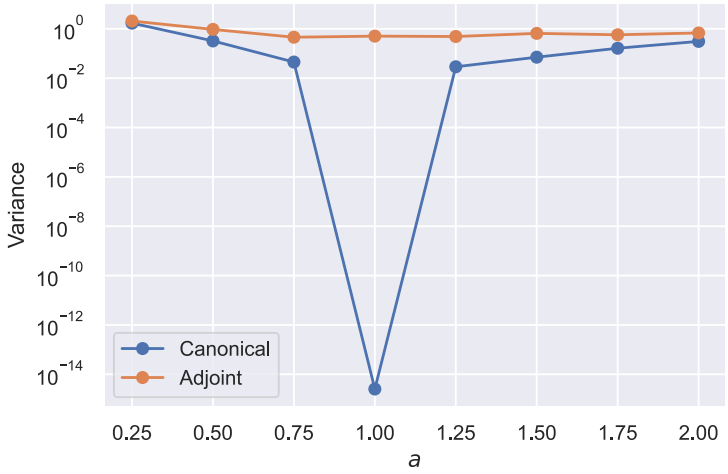

Figure 3. A plot of the variance of the canonical and adjoint energy estimators $\hat{\mathcal{L}}_\theta^{(\mathrm{can})}(x)$ and $\hat{\mathcal{L}}_\theta^{(\mathrm{adj})}(x)$ for the simple harmonic oscillator Hamiltonian (51) as a function of the variational parameter a in (53) with $b=0$ fixed.

Appendix I. Lower bounds on the energies

The lower bound on energies displayed in figure 1 is obtained using the following inequalities, which are valid for a Hamiltonian operator of the form $H=-\frac{1}{2}\nabla^2+V(x)$ acting on $L^2(\mathbb{R}^d)$,

$$\langle\psi|H\psi\rangle=\int_{\mathbb{R}^d}\mathrm{d}^dx\,\overline{\psi}(x)(H\psi)(x) \tag{166}$$

$$=\int_{\mathbb{R}^d}\mathrm{d}^dx\left[\frac{1}{2}\|\nabla\psi\|^2+|\psi(x)|^2V(x)\right] \tag{167}$$

$$\geqslant\int_{\mathbb{R}^d}\mathrm{d}^dx\,|\psi(x)|^2V(x), \tag{168}$$

$$\geqslant V(x_{\min}), \tag{169}$$

where in the second equality we used integration by parts and where $x_{\min}$ denotes any classical minimizer of V. Setting ψ equal to the unique ground eigenfunction we obtain the lower bound

$$\lambda_{\min}\geqslant V(x_{\min}). \tag{170}$$

This is a non-trivial bound for our problem since V is a quartic polynomial whose optimization is generically NP-hard. However, $V(x_{\min})$ can be estimated by performing numerical optimization with multiple random restarts using the Adam optimizer with cosine decay learning rate; the loss curves from these optimizations are shown in figure 4. Given the above lower bound, the relative error can be bounded as follows,

$$\text{relative error}=\frac{E_{\mathrm{Gauss}}-\lambda_{\min}}{-\lambda_{\min}}=\frac{E_{\mathrm{Gauss}}}{-\lambda_{\min}}+1\leqslant\frac{E_{\mathrm{Gauss}}}{-V(x_{\min})}+1=\frac{E_{\mathrm{Gauss}}-V(x_{\min})}{-V(x_{\min})}, \tag{171}$$

where the above inequality used (170) and the fact that $\lambda_{\min}$, $V(x_{\min})$ and E_{Gauss} are all negative. The results are as follows:

d	Max relative error
2	0.30
5	0.38
10	0.59
25	0.49
50	0.59
100	0.52

Granted, this is only an upper bound on the relative error, and thus the true error could be considerably lower.

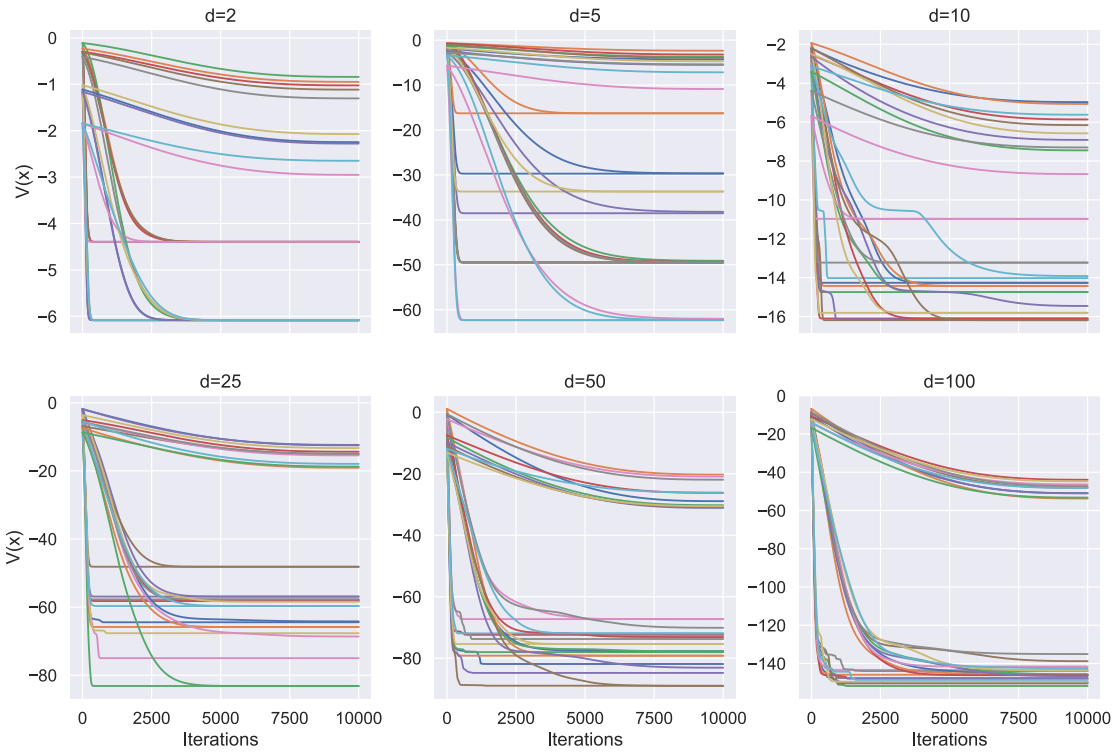

Figure 4. Loss curves from the numerical optimization of $V(x_{\text{min}})$ with multiple random restarts.

ORCID iD

Shravan Veerapaneni https://orcid.org/0000-0002-2294-7233

References

[1] Albergo M S, Kanwar G and Shanahan P E 2019 Flow-based generative models for Markov chain Monte Carlo in lattice field theory *Phys. Rev.* D **100** 034515
[2] Amari S-I 1998 Natural gradient works efficiently in learning *Neural Comput.* **10** 251–76
[3] Barison S, Vicentini F and Carleo G 2021 An efficient quantum algorithm for the time evolution of parameterized circuits *Quantum* **5** 512
[4] Bauer M, Bruveris M and Michor P W 2016 Uniqueness of the Fisher–Rao metric on the space of smooth densities *Bull. London Math. Soc.* **48** 499–506
[5] Berry M V 1989 The quantum phase, five years after *Geometric Phases in Physics* (Singapore: World Scientific) pp 7–28
[6] Carleo G, Becca F, Sanchez-Palencia L, Sorella S and Fabrizio M 2014 Light-cone effect and supersonic correlations in one-and two-dimensional bosonic superfluids *Phys. Rev.* A **89** 031602
[7] Carleo G, Becca F, Schiró M and Fabrizio M 2012 Localization and glassy dynamics of many-body quantum systems *Sci. Rep.* **2** 1–6
[8] Carleo G and Troyer M 2017 Solving the quantum many-body problem with artificial neural networks *Science* **355** 602–6
[9] Cranmer K, Golkar S and Pappadopulo D 2019 Inferring the quantum density matrix with machine learning (arXiv:1904.05903)
[10] Dangel F, Kunstner F and Hennig P 2020 Backpack: packing more into backprop *Int. Conf. on Learning Representations*
[11] Del Debbio L, Marsh Rossney J and Wilson M 2021 Efficient modeling of trivializing maps for lattice ϕ 4 theory using normalizing flows: a first look at scalability *Phys. Rev.* D **104** 094507
[12] Dinh L, Sohl-Dickstein J and Bengio S 2017 Density estimation using real NVP *5th Int. Conf. on Learning Representations (ICLR 2017) (Toulon, France, 24–26 April 2017)* (*Conf. Track Proc.*) (OpenReview.net)
[13] Durkan C, Bekasov A, Murray I and Papamakarios G 2020 nflows: normalizing flows in PyTorch (Zenodo) (https://doi.org/10.5281/zenodo.4296287)
[14] Finkenrath J 2022 Tackling critical slowing down using global correction steps with equivariant flows: the case of the schwinger model (arXiv:2201.02216)
[15] Glimm J and Jaffe A 2012 *Quantum Physics: A Functional Integral Point of View* (Berlin: Springer)
[16] Hackett D C, Hsieh C-C, Albergo M S, Boyda D, Chen J-W, Chen K-F, Cranmer K, Kanwar G and Shanahan P E 2021 Flow-based sampling for multimodal distributions in lattice field theory (arXiv:2107.00734)
[17] Hackl L, Guaita T, Shi T, Haegeman J, Demler E and Cirac I 2020 Geometry of variational methods: dynamics of closed quantum systems *SciPost Phys.* **9** 048
[18] Han X and Hartnoll S A 2020 Deep quantum geometry of matrices *Phys. Rev.* X **10** 011069
[19] Han X and Rinaldi E 2021 Neural quantum states for supersymmetric quantum gauge theories (arXiv:2112.05333)
[20] Hibat-Allah M, Ganahl M, Hayward L E, Melko R G and Carrasquilla J 2020 Recurrent neural network wave functions *Phys. Rev. Res.* **2** 023358

[21] Ido K, Ohgoe T and Imada M 2015 Time-dependent many-variable variational Monte Carlo method for nonequilibrium strongly correlated electron systems *Phys. Rev.* B **92** 245106
[22] Jordan S P, Lee K S M and Preskill J 2014 Quantum computation of scattering in scalar quantum field theories *Quantum Inf. Comput.* **14** 1014–80
[23] Kanwar G, Albergo M S, Boyda D, Cranmer K, Hackett D C, Racanière S, Jimenez Rezende D and Shanahan P E 2020 Equivariant flow-based sampling for lattice gauge theory *Phys. Rev. Lett.* **125** 121601
[24] Kingma D P and Ba J 2015 Adam: a method for stochastic optimization *3rd Int. Conf. on Learning Representations (ICLR 2015) (San Diego, CA, USA, 7–9 May 2015)* (*Conf. Track Proc.*) ed Y Bengio and Y LeCun
[25] Lawrence S and Yamauchi Y 2021 Normalizing flows and the real-time sign problem *Phys. Rev.* D **103** 114509
[26] Liu J, Sun J and Yuan X 2021 Towards a variational Jordan–Lee–Preskill quantum algorithm (arXiv:2109.05547)
[27] Loshchilov I and Hutter F 2017 SGDR: stochastic gradient descent with warm restarts *5th Int. Conf. on Learning Representations (ICLR 2017) (Toulon, France, 24–26 April 2017)* (*Conf. Track Proc.*) (OpenReview.net)
[28] Luo D and Clark B K 2019 Backflow transformations via neural networks for quantum many-body wave functions *Phys. Rev. Lett.* **122** 226401
[29] McLachlan A D 1964 A variational solution of the time-dependent Schrodinger equation *Mol. Phys.* **8** 39–44
[30] Lauchlin McMillan W 1965 Ground state of liquid He^4 *Phys. Rev.* **138** A442
[31] Papamakarios G, Nalisnick E, Jimenez Rezende D, Mohamed S and Lakshminarayanan B 2021 Normalizing flows for probabilistic modeling and inference *J. Mach. Learn. Res.* **22** 1–64
[32] Paszke A *et al* 2019 Pytorch: an imperative style, high-performance deep learning library *Advances in Neural Information Processing Systems* vol 32, ed H Wallach, H Larochelle, A Beygelzimer, F d'Alché-Buc, E Fox and R Garnett (Curran Associates, Inc.) pp 8024–35
[33] Pfau D and Rezende D 2020 Integrable nonparametric flows (arXiv:2012.02035)
[34] Jimenez Rezende D and Mohamed S 2015 Variational inference with normalizing flows (arXiv:1505.05770)
[35] Jimenez Rezende D, Papamakarios G, Racanière S, Albergo M, Kanwar G, Shanahan P and Cranmer K 2020 Normalizing flows on tori and spheres *Int. Conf. on Machine Learning* (PMLR) pp 8083–92
[36] Rinaldi E, Han X, Hassan M, Feng Y, Nori F, McGuigan M and Hanada M 2022 Matrix-model simulations using quantum computing, deep learning and lattice Monte Carlo *PRX Quantum* **3** 010324
[37] Sharir O, Levine Y, Wies N, Carleo G and Shashua A 2020 Deep autoregressive models for the efficient variational simulation of many-body quantum systems *Phys. Rev. Lett.* **124** 020503
[38] Sorella S, Casula M and Rocca D 2007 Weak binding between two aromatic rings: feeling the van der Waals attraction by quantum Monte Carlo methods *J. Chem. Phys.* **127** 014105
[39] Stokes J, De S, Veerapaneni S and Carleo G 2023 Continuous-variable neural network quantum states and the quantum rotor model *Quantum Mach. Intell.* **5** 12
[40] Stokes J, Izaac J, Killoran N and Carleo G 2020 Quantum natural gradient *Quantum* **4** 269
[41] Townsend J, Koep N and Weichwald S 2016 Pymanopt: a Python toolbox for optimization on manifolds using automatic differentiation *J. Mach. Learn. Res.* **17** 1–5
[42] Webber R J and Lindsey M 2021 Rayleigh-Gauss–Newton optimization with enhanced sampling for variational Monte Carlo (arXiv:2106.10558)
[43] Weir D J 2010 Studying a relativistic field theory at finite chemical potential with the density matrix renormalization group *Phys. Rev.* D **82** 025003
[44] Xie H, Zhang L and Wang L 2021 *Ab-initio* study of interacting fermions at finite temperature with neural canonical transformation (arXiv:2105.08644)
[45] Yuan X, Endo S, Zhao Q, Li Y and Benjamin S C 2019 Theory of variational quantum simulation *Quantum* **3** 191
[46] Zhao T, De S, Chen B, Stokes J and Veerapaneni S 2021 Overcoming barriers to scalability in variational quantum Monte Carlo *Proc. Int. Conf. for High Performance Computing, Networking, Storage and Analysis* pp 1–13