One More Time: Revisiting Neural Quantum States from a Reinforcement Learning Perspective

Juan Agustín Duque[1,2,*] **Sergio García-Heredia**[2,*] **Vinicius Hernandes**[3] **Eliška Greplová**[3]
Thomas Spriggs[3,†] **Aaron Courville**[1,4,†] **Anna Dawid**[2,†]
[1]Mila Quebec AI Institute, Université de Montréal, Montréal, Canada
[2]Applied Quantum Algorithms $\langle \mathrm{aQa}^L \rangle$, LIACS & LION, Leiden University, The Netherlands
[3]QuTech and Kavli Institute of Nanoscience, Delft University of Technology, The Netherlands
[4]CIFAR AI Chair
juanduquevan@gmail.com
[*]Equal contribution. [†]Equal supervision.

Abstract

Neural quantum states (NQS) provide a flexible and scalable framework for approximating quantum many-body wavefunctions. Among NQS parameterizations, autoregressive models are especially attractive because they enable exact, independent sampling from the Born distribution, avoiding the autocorrelation and mixing issues of Markov chain methods. Yet their optimization remains challenging: Adam is a scalable method but ignores function space geometry, while stochastic reconfiguration is principled but costly and numerically fragile in large models. To address this gap, we show that variational energy minimization can be viewed as an advantage policy-gradient problem over the Born distribution, motivating trust-region optimization for NQS training. We introduce *Proximal Wavefunction Optimization* (PWO), a principled trust-region algorithm that clips probability-ratio changes in the amplitude channel and phase increments in the phase channel. PWO avoids explicit matrix inversion, reuses samples across multiple updates, and combines the scalability of first-order optimization with theoretical guarantees. Across Ising and frustrated J_1–J_2 one- and two-dimensional spin systems, PWO improves stability and wall-clock convergence over Adam, minSR, and SPRING. Finally, we fine-tune a 1.5B-parameter RWKV-7 model, demonstrating NQS optimization at a scale over three orders of magnitude beyond prior work.

1 Introduction

Quantum physics seeks to predict and understand the behavior of many interacting quantum particles. However, determining the exact ground (lowest-energy) state of a system of N quantum bits generally requires diagonalizing a $2^N \times 2^N$ Hamiltonian matrix, quickly rendering exact methods intractable at scale. Variational approaches address this exponential complexity by optimizing a parameterized wavefunction ansatz to approximate the ground state. Neural quantum states (NQS) [Carleo and Troyer, 2017, Lange et al., 2024] build upon this by using neural networks as expressive wavefunction representations, capable of capturing complex correlations [Nomura and Imada, 2021] and high entanglement [Gauvin-Ndiaye et al., 2025], as well as scaling to high-dimensional systems [Pescia et al., 2024], where traditional methods struggle.

The practical success of NQS depends not only on expressivity but also on the ability to sample from and optimize the variational distribution. State-of-the-art NQS methods [Wu et al., 2024] primarily rely on Markov Chain Monte Carlo (MCMC) to estimate energies and gradients, which can introduce autocorrelation, slow mixing, and unreliable exploration in difficult regimes [Wolff, 1990, Del Debbio et al., 2004, Krinitsin et al., 2026]. Autoregressive NQS remove this sampling bottleneck by factorizing the Born distribution and enabling exact, independent sampling [Sharir et al.,

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

2020, Hibat-Allah et al., 2020]. As a result, they provide a clean setting for variational optimization: samples can be drawn directly from the current wavefunction, and training dynamics are no longer confounded by Markov chain limitations. Despite these advantages, their adoption remains limited, as stable optimization is a central bottleneck for training accurate autoregressive models.

Existing optimization methods expose a sharp trade-off. First-order optimizers such as Adam [Kingma and Ba, 2014] are computationally efficient and scale naturally to large neural networks, but they ignore the geometry of the variational manifold and can converge unstably or inaccurately in NQS applications [Pfau et al., 2020, Liu et al., 2025]. Stochastic reconfiguration (SR), and scalable variants such as minSR [Chen and Heyl, 2024], are more geometrically principled because they do natural-gradient descent in wavefunction space. However, they require solving large and often ill-conditioned linear systems [Attar et al., 2026], with costs that become prohibitive for large networks or large sample regimes. Thus, autoregressive NQS resolve an important sampling problem, but stable first-order optimization of large autoregressive wavefunction models remains a central obstacle, recently explored with transfer learning [Moss et al., 2025, Merali et al., 2026].

Contributions. This gap points to a missing optimization principle in NQS. In this work, we observe that the variational objective in NQS is mathematically equivalent to a policy-gradient objective in reinforcement learning (RL) under certain assumptions. While this connection has been implicitly present since early work [Carleo and Troyer, 2017, Wu et al., 2019], it has not been formalized or leveraged to design modern optimizers. In this work, we show that the *RL-style trust-region optimization improves NQS training scalability and stability.* Our contributions are as follows:

- We formally connect variational energy minimization and policy-gradient RL, showing that the NQS gradient admits an advantage-weighted form over the Born distribution.
- We introduce *Proximal Wavefunction Optimization (PWO)*, an algorithm for NQS training inspired by Proximal Policy Optimization (PPO) [Schulman et al., 2017], and prove that the PWO surrogate satisfies a trust-region improvement bound, allowing safe sample reuse.
- We show that PWO improves the stability and convergence speed of autoregressive NQS compared to Adam and minSR on standard benchmarks (1D, 2D Ising and J_1–J_2).
- Finally, we demonstrate the scalability of PWO by fine-tuning a 1.5 billion parameter RWKV-7 LLM [Peng et al., 2025] on the 1-D Ising model.

2 Background

2.1 Reinforcement Learning

Reinforcement Learning (RL) is a machine learning paradigm in which an agent learns to make decisions by interacting with an environment. At each time step t, the agent observes a state s_t, selects an action a_t, and then receives a reward r_{t+1} and a new state s_{t+1} from the transition function $P(\cdot|s_t, a_t)$. The goal of the agent is to learn a *policy* that maximizes the expected return over time. Let $\pi_{\boldsymbol{\theta}}$ denote a policy parameterized by $\boldsymbol{\theta}$. Formally, the return is defined over a trajectory $\tau := (s_0, a_0, r_1, s_1, a_1, r_2, \dots)$, which is a sequence of states, actions, and rewards generated by the agent-environment interaction. Following the notation of Agarwal et al. [2021], the probability of a trajectory τ under $\pi_{\boldsymbol{\theta}}$ is given by

$$\mathcal{P}_{\boldsymbol{\theta}}(\tau) = \mu(s_0)\pi_{\boldsymbol{\theta}}(a_0|s_0)P(s_1|s_0, a_0)\pi_{\boldsymbol{\theta}}(a_1|s_1)P(s_2|s_1, a_1)\dots, \tag{1}$$

where $\mu(s_0)$ denotes the initial distribution over states. The objective of RL is to find a policy that maximizes the expected *discounted return*, which is captured by the state-value and action-value functions of a policy $\pi_{\boldsymbol{\theta}}$:

$$V^{\pi_{\boldsymbol{\theta}}}(s) := \mathbb{E}_{\pi_{\boldsymbol{\theta}}}\left[\sum_{t=0}^{\infty}\gamma^t r_t \,\middle|\, s_0 = s\right], \quad Q^{\pi_{\boldsymbol{\theta}}}(s,a) := \mathbb{E}_{\pi_{\boldsymbol{\theta}}}\left[\sum_{t=0}^{\infty}\gamma^t r_t \,\middle|\, s_0 = s,\ a_0 = a\right], \tag{2}$$

respectively. In policy optimization, an agent maximizes its expected return by performing gradient ascent with a *policy gradient* estimator [Williams, 1992] of the form:

$$\boldsymbol{\nabla}_{\boldsymbol{\theta}} V^{\pi_{\boldsymbol{\theta}}}(\mu) = \mathbb{E}_{\tau\sim\mathcal{P}_{\boldsymbol{\theta}}}\left[\sum_{t=0}^{\infty}\gamma^t A^{\pi_{\boldsymbol{\theta}}}(s_t, a_t)\boldsymbol{\nabla}_{\boldsymbol{\theta}}\log\pi_{\boldsymbol{\theta}}(a_t|s_t)\right]. \tag{3}$$

Here $A^{\pi_{\boldsymbol{\theta}}}(s,a) := Q^{\pi_{\boldsymbol{\theta}}}(s,a) - V^{\pi_{\boldsymbol{\theta}}}(s)$ denotes the advantage of taking action a while in state s, *i.e.*, the expected return of taking action a relative to the policy average while in state s.

2.2 Neural Quantum States

One of the key tasks in theoretical quantum many-body physics is to approximate the ground-state wavefunction of a system with many interacting particles. For a system of N spin-$1/2$ particles, a configuration is a binary vector $\mathbf{s} \in \{\pm 1\}^N$, and the wavefunction can be viewed as a complex vector $|\psi\rangle$ in a Hilbert space $\mathcal{H}$ (see Appendix A) indexed by all 2^N configurations. Equivalently, it defines a function $\mathbf{s} \mapsto \psi(\mathbf{s})$, where $\psi(\mathbf{s}) \in \mathbb{C}$ is the amplitude assigned to configuration $\mathbf{s}$:

$$|\psi\rangle = \sum_{\mathbf{s}\in\{\pm 1\}^N} \psi(\mathbf{s})\,|\mathbf{s}\rangle\,, \tag{4}$$

where we use the Dirac, bra-ket, notation (see Appendix A). The exponential size of this vector makes explicit representations intractable for large N. Neural quantum states (NQS) address this by parameterizing the amplitude function with a neural network, $f_{\boldsymbol{\theta}}$, so that $f_{\boldsymbol{\theta}}(\mathbf{s})$ provides a compact approximation to $\psi(\mathbf{s})$ [Carleo and Troyer, 2017, Dawid et al., 2025]. In practice, NQS models output the logarithm of the complex amplitude,

$$f_{\boldsymbol{\theta}}(\mathbf{s}) = \log \psi_{\boldsymbol{\theta}}(\mathbf{s}) = \log|\psi_{\boldsymbol{\theta}}(\mathbf{s})| + i\,\arg\psi_{\boldsymbol{\theta}}(\mathbf{s}), \tag{5}$$

which separates the wavefunction into a log-modulus and a phase, usually handled by separate networks (channels). This representation is numerically convenient because amplitudes can vary over many orders of magnitude. By the Born rule, normalized wavefunctions induce a probability distribution over configurations, $\mathcal{P}_{\boldsymbol{\theta}}(\mathbf{s}) \propto |\psi_{\boldsymbol{\theta}}(\mathbf{s})|^2$. In this work, we focus on autoregressive NQS, which explicitly factorize $\mathcal{P}_{\boldsymbol{\theta}}$ and enable exact independent sampling from the Born distribution.

2.3 Finding the Ground State

The Hamiltonian $\hat{H}$ is the energy operator of a quantum system: it encodes the interactions, external fields, and kinetic terms that determine which wavefunctions have low or high energy. A popular application of the NQS framework is to find the ground state of quantum systems. Given a description of a physics system within a Hamiltonian, $\hat{H}$, the ground state is the eigenstate $|\psi_0\rangle$ corresponding to the lowest eigenvalue E_0. This can be cast as a minimization problem:

$$E_0 \le E[\psi] = \frac{\langle\psi|\hat{H}|\psi\rangle}{\langle\psi|\psi\rangle}, \quad \forall\,|\psi\rangle \in \mathcal{H}, \tag{6}$$

where the equality holds if and only if $|\psi\rangle$ belongs to the ground-state eigenspace – this is known as the *variational principle*. Taking this variational formulation of the ground state as the starting point, so-called variational methods define a parametric family of states $|\psi_{\boldsymbol{\theta}}\rangle$ and approximate the ground state by solving

$$\boldsymbol{\theta}^* = \operatorname*{argmin}_{\boldsymbol{\theta}} E[\psi_{\boldsymbol{\theta}}] = \operatorname*{argmin}_{\boldsymbol{\theta}} \frac{\langle\psi_{\boldsymbol{\theta}}|\hat{H}|\psi_{\boldsymbol{\theta}}\rangle}{\langle\psi_{\boldsymbol{\theta}}|\psi_{\boldsymbol{\theta}}\rangle}. \tag{7}$$

In the case of NQS, this family of functions is given by a neural network, as already discussed.

2.4 Variational Monte Carlo

To efficiently compute a minimizer of Eq. (7), NQS uses Monte Carlo estimates. This approach is known as variational Monte Carlo (VMC). In particular, the energy can be written as

$$E[\psi_{\boldsymbol{\theta}}] = \frac{\langle\psi_{\boldsymbol{\theta}}|\hat{H}|\psi_{\boldsymbol{\theta}}\rangle}{\langle\psi_{\boldsymbol{\theta}}|\psi_{\boldsymbol{\theta}}\rangle} = \sum_{\mathbf{s}} \mathcal{P}_{\boldsymbol{\theta}}(\mathbf{s})\,E^{\mathrm{loc}}_{\boldsymbol{\theta}}(\mathbf{s}) = \mathbb{E}_{\mathbf{s}\sim\mathcal{P}_{\boldsymbol{\theta}}}\left[E^{\mathrm{loc}}_{\boldsymbol{\theta}}(\mathbf{s})\right], \tag{8}$$

where we define

$$\mathcal{P}_{\boldsymbol{\theta}}(\mathbf{s}) = \frac{|\psi_{\boldsymbol{\theta}}(\mathbf{s})|^2}{\sum_{\mathbf{s}'}|\psi_{\boldsymbol{\theta}}(\mathbf{s}')|^2}, \quad E^{\mathrm{loc}}_{\boldsymbol{\theta}}(\mathbf{s}) := \sum_{\mathbf{s}'} \frac{\psi_{\boldsymbol{\theta}}(\mathbf{s}')}{\psi_{\boldsymbol{\theta}}(\mathbf{s})}\,\langle\mathbf{s}|\hat{H}|\mathbf{s}'\rangle\,. \tag{9}$$

With this formulation, NQS estimates the loss function by sampling a set of spin configurations $\{\mathbf{s}_i\}_{i=1}^M$ from the parameterized distribution $\mathcal{P}_{\boldsymbol{\theta}}$ and computing the sample mean of $E^{\mathrm{loc}}_{\boldsymbol{\theta}}(\mathbf{s}_i)$:

$$L(\boldsymbol{\theta}) \approx \bar{E}^{\mathrm{loc}}_{\boldsymbol{\theta}} \equiv \frac{1}{M}\sum_{i=1}^{M} E^{\mathrm{loc}}_{\boldsymbol{\theta}}(\mathbf{s}_i). \tag{10}$$

Differentiating Eq. (8) (see Appendix B.2) yields a gradient expressed as an expected value over $\mathcal{P}_{\boldsymbol{\theta}}$, which can be estimated using MCMC sampling or direct sampling of autoregressive models:

$$\partial_{\theta_i} L(\boldsymbol{\theta}) = \mathbb{E}_{\mathbf{s}\sim\mathcal{P}_{\boldsymbol{\theta}}}\left[2\,\mathrm{Re}\left\{\left(E^{\mathrm{loc}}_{\boldsymbol{\theta}}(\mathbf{s}) - \mathbb{E}_{\mathbf{s}\sim\mathcal{P}_{\boldsymbol{\theta}}}[E^{\mathrm{loc}}_{\boldsymbol{\theta}}(\mathbf{s})]\right)O_i(\mathbf{s})^*\right\}\right], \tag{11}$$

for $i = 1, \ldots, P$, where $O_i(\mathbf{s}) = \partial_{\theta_i}\log\psi_{\boldsymbol{\theta}}$ are the so-called score functions.

Algorithm 1: Proximal Wavefunction Optimization (PWO)

Input: Hamiltonian $\hat{H}$, NQS $\psi_{\boldsymbol{\theta}}$, batch size M, inner epochs K, amplitude clip ϵ, phase clip δ
Initialize $\boldsymbol{\theta}$
while *not converged* **do**
  $\boldsymbol{\theta}_{\text{old}} \leftarrow \boldsymbol{\theta}$
  sample $\{\mathbf{s}_i\}_{i=1}^{M} \sim \mathcal{P}_{\boldsymbol{\theta}_{\text{old}}}$
  Cache reference log-probabilities, $\log \mathcal{P}_{\boldsymbol{\theta}_{\text{old}}}(\mathbf{s}_i)$; phases, $\arg \psi_{\boldsymbol{\theta}_{\text{old}}}(\mathbf{s}_i)$; and normalized real and imaginary advantages, $A^{\mathrm{R}}_{\boldsymbol{\theta}_{\text{old}}}(\mathbf{s}_i)$ and $A^{\mathrm{I}}_{\boldsymbol{\theta}_{\text{old}}}(\mathbf{s}_i)$ respectively, of $\{\mathbf{s}_i\}_{i=1}^{M}$.
  for $k = 1, \ldots, K$ **do**
    Compute current log-probabilities $\log \mathcal{P}_{\boldsymbol{\theta}}(\mathbf{s}_i)$ and phases $\arg \psi_{\boldsymbol{\theta}}(\mathbf{s}_i)$ for $\{\mathbf{s}_i\}_{i=1}^{M}$
    $r_i \leftarrow \exp(\log \mathcal{P}_{\boldsymbol{\theta}}(\mathbf{s}_i) - \log \mathcal{P}_{\boldsymbol{\theta}_{\text{old}}}(\mathbf{s}_i))$
    $\phi_i \leftarrow 2 \cdot \operatorname{atan}_2\big(\sin(\arg \psi_{\boldsymbol{\theta}}(\mathbf{s}_i) - \arg \psi_{\boldsymbol{\theta}_{\text{old}}}(\mathbf{s}_i)), \cos(\arg \psi_{\boldsymbol{\theta}}(\mathbf{s}_i) - \arg \psi_{\boldsymbol{\theta}_{\text{old}}}(\mathbf{s}_i))\big)$
    $\ell_i^R \leftarrow \max\big(r_i A^{\mathrm{R}}_{\boldsymbol{\theta}_{\text{old}}}(\mathbf{s}_i),\ \operatorname{clip}(r_i, 1-\epsilon, 1+\epsilon) A^{\mathrm{R}}_{\boldsymbol{\theta}_{\text{old}}}(\mathbf{s}_i)\big)$
    $\ell_i^I \leftarrow \text{stop_gradient}(r_i) \cdot \max\big(\phi_i A^{\mathrm{I}}_{\boldsymbol{\theta}_{\text{old}}}(\mathbf{s}_i),\ \operatorname{clip}(\phi_i, -\delta, \delta) A^{\mathrm{I}}_{\boldsymbol{\theta}_{\text{old}}}(\mathbf{s}_i)\big)$
    $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \eta \nabla_{\boldsymbol{\theta}} \frac{1}{M} \sum_{i=1}^{M} \left(\ell_i^R + \ell_i^I\right)$.

2.5 Stochastic Reconfiguration

Stochastic Gradient Descent (SGD) [Robbins and Monro, 1951] is a simple and effective optimization method in many scenarios. However, in the context of VMC, it can suffer from slow convergence and instability due to the complex geometry of the parameter space. The key issue is that small changes in the parameters, $\boldsymbol{\theta}$, do not necessarily correspond to small changes in the quantum state $|\psi_{\boldsymbol{\theta}}\rangle$.

Stochastic reconfiguration (SR) [Sorella, 1998], which generalizes natural gradient descent [Amari, 1998] to VMC, preconditions the gradient $\nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta})$ to encourage that the distance between $|\psi_{\boldsymbol{\theta}}\rangle$ at one training iteration and the next is kept small. This distance is given by the infidelity between the two states and approximated to second order by the Fubini-Study metric,

$$S_{ij} = \operatorname{Re}\left\{\operatorname{Cov}_{\mathbf{s}\sim\mathcal{P}_{\boldsymbol{\theta}}}\left[O_i^*(\mathbf{s}), O_j(\mathbf{s})\right]\right\}, \qquad i,j = 1, \ldots, P, \tag{12}$$

which encodes the local geometry of the Hilbert space at the current parameter configuration. In mathematical terms, SR modifies the update rule in the following way:

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \eta \mathbf{S}^{-1} \nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta}), \tag{13}$$

where $\eta \in \mathbb{R}$ is the learning rate. We discuss the high computational costs of SR and its more recent improvements in Appendix C.1, motivating the use of first-order optimization methods.

3 Proximal Wavefunction Optimization

Here, we introduce a new general optimization procedure for training NQS. We first draw an equivalence between the current NQS optimization paradigm and the policy gradient update in RL for fixed-phase wavefunctions. This motivates applying Proximal Policy Optimization to NQS. We then extend PPO to general complex-valued wavefunctions.

Proposition 3.1 (Policy-gradient form of variational energy minimization)**.** *Assume the Hamiltonian is stoquastic, i.e. its matrix representation in a chosen computational basis has non-positive off-diagonal elements. Then we can assume $f_{\boldsymbol{\theta}} = \log \psi_{\boldsymbol{\theta}} = \log |\psi_{\boldsymbol{\theta}}|$, and the gradient of the variational energy can be written in policy-gradient form as*

$$\nabla_{\boldsymbol{\theta}} E[\psi_{\boldsymbol{\theta}}] = \mathbb{E}_{\mathbf{s}\sim\mathcal{P}_{\boldsymbol{\theta}}}\left[\left(E^{\mathrm{loc}}_{\boldsymbol{\theta}}(\mathbf{s}) - \mathbb{E}_{\mathbf{s}\sim\mathcal{P}_{\boldsymbol{\theta}}}[E^{\mathrm{loc}}_{\boldsymbol{\theta}}(\mathbf{s})]\right) \nabla_{\boldsymbol{\theta}} \log \mathcal{P}_{\boldsymbol{\theta}}(\mathbf{s})\right]. \tag{14}$$

In particular, variational energy minimization is equivalent to an advantage policy-gradient update over configurations. For a proof, see Appendix B.1.

This proposition establishes a direct link between the NQS and RL frameworks (see also Tab. 1) by relating Eqs. (11) and (14), motivating our algorithm, *Proximal Wavefunction Optimization* (PWO). The key observation is that the SR gradient results from solving an approximate constrained optimization problem: minimizing the expected energy while constraining the infidelity between

Reinforcement Learning		Variational Monte Carlo	
Policy	$\pi_{\boldsymbol{\theta}}(a \mid s)$	Born distribution	$\mathcal{P}_{\boldsymbol{\theta}}(\mathbf{s}) \propto \lvert\psi_{\boldsymbol{\theta}}(\mathbf{s})\rvert^2$
Advantage	$A^{\pi_{\boldsymbol{\theta}}}(s_t, a_t)$	Centered local energy	$\Delta E(\mathbf{s}) = E^{\mathrm{loc}}_{\boldsymbol{\theta}}(\mathbf{s}) - E[\psi_{\boldsymbol{\theta}}]$
Policy gradient	$A^{\pi_{\boldsymbol{\theta}}} \boldsymbol{\nabla}_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}$	VMC force	$2\,\mathrm{Re}[\Delta E(\mathbf{s}) \boldsymbol{\nabla}_{\boldsymbol{\theta}} \log \psi^*_{\boldsymbol{\theta}}(\mathbf{s})]$
KL trust region	$D_{\mathrm{KL}}(\pi_{\boldsymbol{\theta}_{\mathrm{old}}} \Vert \pi_{\boldsymbol{\theta}})$	Infidelity	$\mathcal{I}(\psi_{\boldsymbol{\theta}}, \psi_{\boldsymbol{\theta}+\delta\boldsymbol{\theta}})$
Fisher matrix	$\mathbf{F}$	Fubini–Study metric	$\mathbf{S}$

Table 1: Correspondence between RL policy-gradient methods and VMC methods for NQS.

successive wavefunctions. An analogous problem has long been studied in RL as *trust region optimization*, notably in Natural Policy Gradient (NPG) [Kakade, 2001] and Trust Region Policy Optimization (TRPO) [Schulman et al., 2015], which aims to improve expected rewards while limiting changes in the policy distribution. SR, NPG, and TRPO are closely related to natural gradient descent, which uses the geometry of their corresponding functional space to precondition gradients, at a higher computational cost than standard first-order methods. In RL, PPO emerged as a simpler, computationally efficient alternative to TRPO, motivating an analogous approach to NQS optimization.

Proximal Policy Optimization (PPO) [Schulman et al., 2017] introduces a simple heuristic that empirically shows monotonic improvements by constraining updates to keep the updated policy close to the previous one. Applied to NQS, PPO would minimize the following surrogate loss:

$$L^{\mathrm{clip}}_{\mathrm{mod}}(\boldsymbol{\theta}) = \mathbb{E}_{\mathbf{s}\sim\mathcal{P}_{\boldsymbol{\theta}_{\mathrm{old}}}} \left[\max\left(r_{\boldsymbol{\theta}}(\mathbf{s}) A^{\mathrm{R}}_{\boldsymbol{\theta}_{\mathrm{old}}}(\mathbf{s}), \mathrm{clip}(r_{\boldsymbol{\theta}}(\mathbf{s}), 1-\epsilon, 1+\epsilon) A^{\mathrm{R}}_{\boldsymbol{\theta}_{\mathrm{old}}}(\mathbf{s})\right)\right], \tag{15}$$

where $r_{\boldsymbol{\theta}} = \mathcal{P}_{\boldsymbol{\theta}}/\mathcal{P}_{\boldsymbol{\theta}_{\mathrm{old}}}$ is the importance sampling ratio between the updated and reference distributions (see Appendix B.4), ϵ is the amplitude clip, the $\max$ comes from solving a minimization problem, and

$$A^{\mathrm{R}}_{\boldsymbol{\theta}_{\mathrm{old}}}(\mathbf{s}) := \mathrm{Re}\left\{E^{\mathrm{loc}}_{\boldsymbol{\theta}_{\mathrm{old}}}(\mathbf{s})\right\} - \mathbb{E}_{\mathbf{s}\sim\mathcal{P}_{\boldsymbol{\theta}_{\mathrm{old}}}} \left[\mathrm{Re}\left\{E^{\mathrm{loc}}_{\boldsymbol{\theta}_{\mathrm{old}}}(\mathbf{s})\right\}\right], \tag{16}$$

is the equivalent of the advantage estimate, yielding the local energy of configuration $\mathbf{s}$ relative to the expected energy under the current NQS. In the case of RL, clipping ensures that policy updates remain conservative, thereby maintaining stability during optimization.

While PPO is a natural fit for the amplitude channel, it does not by itself account for the full VMC gradient for a general (non-stoquastic) Hamiltonian. As shown in Appendix B.3, the variational gradient also contains an imaginary component, which governs how the phase of the wavefunction should evolve. We, therefore, introduce a second proximal objective that controls phase updates directly by constraining the wrapped phase increment between the current and reference models. Concretely, letting $\phi_{\boldsymbol{\theta}}(\mathbf{s})$ denote twice the wrapped phase difference (see Appendix B.4), we minimize

$$L^{\mathrm{clip}}_{\mathrm{arg}}(\boldsymbol{\theta}) = \mathbb{E}_{\mathbf{s}\sim\mathcal{P}_{\boldsymbol{\theta}_{\mathrm{old}}}} \left[\mathrm{sg}\big(r_{\boldsymbol{\theta}}(\mathbf{s})\big)\ \max\big(\phi_{\boldsymbol{\theta}}(\mathbf{s}) A^{\mathrm{I}}_{\boldsymbol{\theta}_{\mathrm{old}}}(\mathbf{s}), \mathrm{clip}\big(\phi_{\boldsymbol{\theta}}(\mathbf{s}), -\delta, \delta\big) A^{\mathrm{I}}_{\boldsymbol{\theta}_{\mathrm{old}}}(\mathbf{s})\big)\right], \tag{17}$$

where δ is the phase clip, $\mathrm{sg}(\cdot)$ denotes the stop-gradient operator (see Theorem B.5), and

$$A^{\mathrm{I}}_{\boldsymbol{\theta}_{\mathrm{old}}}(\mathbf{s}) := \mathrm{Im}\left\{E^{\mathrm{loc}}_{\boldsymbol{\theta}_{\mathrm{old}}}(\mathbf{s})\right\} - \mathbb{E}_{\mathbf{s}\sim\mathcal{P}_{\boldsymbol{\theta}_{\mathrm{old}}}} \left[\mathrm{Im}\left\{E^{\mathrm{loc}}_{\boldsymbol{\theta}_{\mathrm{old}}}(\mathbf{s})\right\}\right]. \tag{18}$$

The imaginary part of the local energy has zero expectation, so the centering is retained only to mirror the RL advantage. Equation (17) is the phase analog of PPO: instead of clipping a probability ratio, it clips the phase increment itself, preventing abrupt rotations of the wavefunction while still following the imaginary part of the VMC gradient. Using the phase increment instead of the ratio of phases ensures that the gradient of the combined loss (Eqs. (15)+(17)) matches the original one (Eq. (11)) when we are on-policy, *i.e.*, when the importance sampling ratio $r_{\boldsymbol{\theta}}$ is one (no sample reuse).

PWO combines the amplitude and phase losses into an *efficient* (see Appendix C.2) first-order optimization procedure. For the amplitude channel, Proposition 3.1 shows that, when the phase is fixed, variational energy minimization takes the form of an advantage-weighted policy-gradient update over configurations, which naturally motivates a PPO-style clipped surrogate based on importance ratios. For the phase channel, Appendix B.6 shows that the imaginary part of the VMC gradient can be conservatively optimized through a surrogate that constrains the phase increment. In practice, PWO performs K inner updates on a frozen batch of configurations sampled from the reference Born distribution $\mathcal{P}_{\boldsymbol{\theta}_{\mathrm{old}}}$, clipping probability-ratio changes for the amplitude and wrapped phase increments for the phase. Because the same batch is reused across inner steps, the phase objective includes detached importance weights to correct for the mismatch between the reference and current sampling distributions (see Theorem B.5). Algorithm 1 summarizes the resulting procedure.

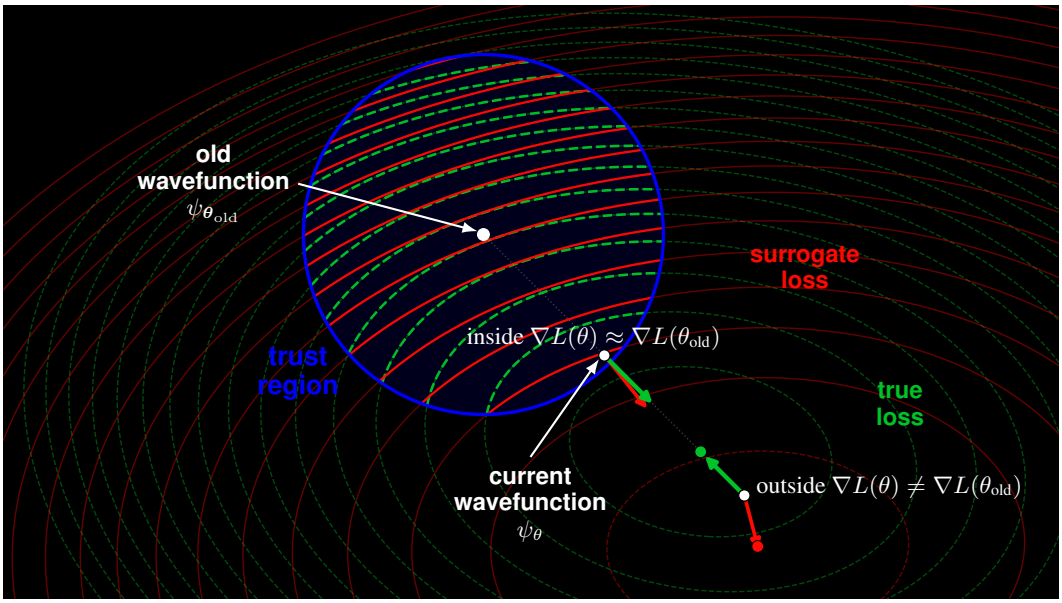


Figure 1: Trust-region intuition for PWO. Inside the trust region, the surrogate landscape gives a reliable local improvement direction; outside the trust region, the surrogate and true landscapes may disagree substantially.

4 Theoretical Analysis

The theoretical justification for PWO parallels monotonic-improvement analyses in policy optimization. In CPI [Kakade and Langford, 2002] and TRPO [Schulman et al., 2015], the true return change is controlled by a surrogate objective under a reference policy plus a penalty for moving too far from it. PWO follows the same principle with the return replaced by negative variational energy, probability ratios controlling the amplitude channel, and phase increments controlling the phase channel. Figure 1 illustrates this principle: trust regions make sample reuse possible by keeping optimization close enough to the reference wavefunction $\psi_{\boldsymbol{\theta}_{\text{old}}}$ so that the surrogate loss function evaluated on old samples remains a controlled approximation to the current energy landscape. Theorem 4.1 shows that this proximal construction preserves the exact local VMC direction.

Theorem 4.1 (First-order consistency of PWO)**.** *Let $|\psi_{\boldsymbol{\theta}}\rangle$ be differentiable at $\boldsymbol{\theta}_{\text{old}}$, and assume common support. For $\epsilon, \delta > 0$,*

$$\boldsymbol{\nabla}_{\boldsymbol{\theta}} \left(L^{\text{clip}}_{\text{mod}}(\boldsymbol{\theta}) + L^{\text{clip}}_{\text{arg}}(\boldsymbol{\theta}) \right) \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_{\text{old}}} = \boldsymbol{\nabla}_{\boldsymbol{\theta}} E[\psi_{\boldsymbol{\theta}}]|_{\boldsymbol{\theta}=\boldsymbol{\theta}_{\text{old}}} . \tag{19}$$

For a proof see Appendix B.6.

Thus, the first PWO inner update exactly matches the VMC gradient; the surrogate only controls how far this direction is followed while reusing samples. For finite updates, local agreement is not enough. Let $r_{\boldsymbol{\theta}} = \mathcal{P}_{\boldsymbol{\theta}}/\mathcal{P}_{\boldsymbol{\theta}_{\text{old}}}$ and $\alpha_{\boldsymbol{\theta}}$ be the wrapped phase increment and define

$$\mathcal{A}_{\boldsymbol{\theta}_{\text{old}}}(\boldsymbol{\theta}) := 2\mathbb{E}_{\mathbf{s}\sim\mathcal{P}_{\boldsymbol{\theta}_{\text{old}}}} \left[\sqrt{r_{\boldsymbol{\theta}}} \left(\cos\alpha_{\boldsymbol{\theta}} A^{\text{R}}_{\boldsymbol{\theta}_{\text{old}}} + \sin\alpha_{\boldsymbol{\theta}} A^{\text{I}}_{\boldsymbol{\theta}_{\text{old}}} \right) \right] , \tag{20}$$

Theorem 4.2 (Infidelity energy bound)**.** *Let $|\psi_{\boldsymbol{\theta}_{\text{old}}}\rangle$ and $|\psi_{\boldsymbol{\theta}}\rangle$ be normalized wavefunctions with common support. Then*

$$E[\psi_{\boldsymbol{\theta}}] - E[\psi_{\boldsymbol{\theta}_{\text{old}}}] \leq \mathcal{A}_{\boldsymbol{\theta}_{\text{old}}}(\boldsymbol{\theta}) + 2 \left\| \hat{H} - E[\psi_{\boldsymbol{\theta}_{\text{old}}}]\mathbb{1} \right\|_{\infty} \left(1 - \sqrt{1 - \mathcal{I}(\psi_{\boldsymbol{\theta}_{\text{old}}}, \psi_{\boldsymbol{\theta}})} \right) . \tag{21}$$

For a proof see Appendix B.4.

This bound is independent of clipping and holds for any finite update. The clipped PWO certificate follows by adding the stronger requirement that the realized update satisfies the amplitude and phase trust regions globally, not only on the sampled batch (which is encouraged by the method but not guaranteed, see Appendix B.7).

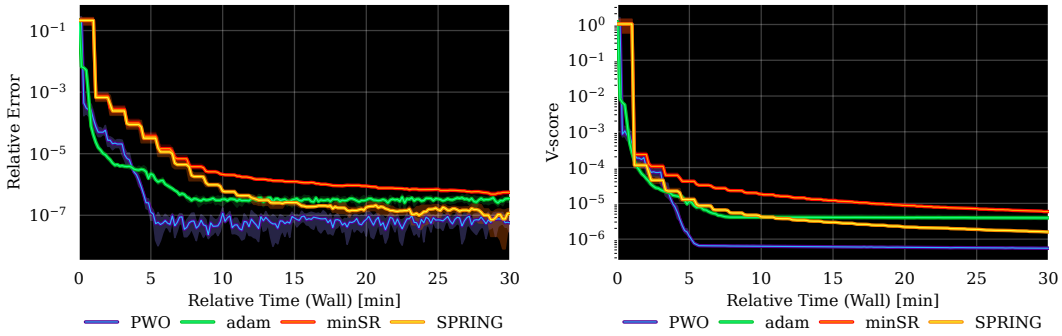


Figure 2: Comparison of PWO, Adam, minSR, and SPRING on the transverse-field Ising model over 10 random seeds. All methods were run with 1024 samples on a single NVIDIA L40S GPU.

Corollary 4.3 (Clipped PWO improvement certificate)**.** *Assume the conditions of Theorem 4.2, bounded centered local energies, and global constraints* $r_{\boldsymbol{\theta}}(\mathbf{s}) \in [1-\epsilon, 1+\epsilon]$ *and* $|\phi_{\boldsymbol{\theta}}(\mathbf{s})| \le \delta$ *for all* $\mathbf{s}$*, with* $0 \le \epsilon \le 1$ *and* $0 \le \delta \le \pi$*. Then there exists* $C_{\boldsymbol{\theta}_{\text{old}}} < \infty$ *such that*

$$E[\psi_{\boldsymbol{\theta}}] - E[\psi_{\boldsymbol{\theta}_{\text{old}}}] \le L^{\text{clip}}_{\text{mod}}(\boldsymbol{\theta}) + L^{\text{clip}}_{\text{arg}}(\boldsymbol{\theta}) + 2C_{\boldsymbol{\theta}_{\text{old}}} \left(1 - \frac{1+\sqrt{1-\epsilon^2}}{2} \cos^2\left(\frac{\delta}{2}\right)\right). \tag{22}$$

For a proof see Appendix B.7.

The final term is the trust-region penalty. It vanishes as $\epsilon, \delta \to 0$ and grows as either clipping range is relaxed. Hence, if the clipped surrogate is minimized until the right-hand side of Eq. (22) is negative, the updated wavefunction is guaranteed to have lower energy. Together, the results show that PWO follows the exact local VMC direction, while finite-update errors are controlled generally by infidelity and, under global clipping, by explicit amplitude and phase trust-regions.

5 Experiments

5.1 Spin-chain benchmarks

Although PWO applies to both autoregressive and non-autoregressive NQS, we focus on autoregressive models, which eliminate MCMC sampling limitations but remain challenging to optimize. We evaluate PWO on two spin systems using 1.5M-parameter autoregressive NQS. Each system is a one-dimensional spin-1/2 chain of $N = 12$ sites with periodic boundary conditions (site indices modulo N). Let $\hat{\sigma}_i^{\alpha}$, $\alpha \in \{x, y, z\}$, denote the Pauli operator on site i, and $\hat{\mathbf{S}}_i = 1/2(\hat{\sigma}_i^x, \hat{\sigma}_i^y, \hat{\sigma}_i^z)$. For each Hamiltonian, we minimize the variational energy $E[\psi_{\boldsymbol{\theta}}]$ and report its relative error with respect to the exact ground-state energy E_0, $\epsilon_{\text{rel}} = (E[\psi_{\boldsymbol{\theta}}] - E_0)/E_0$, computable for these small systems. We also report the V-score [Wu et al., 2024], a scale-invariant convergence metric based on the energy variance that vanishes for exact eigenstates. The chosen Hamiltonians require increasingly complex functions to represent their ground-state physics. We use the same NQS architecture and the number of samples for all optimizers, isolating optimization performance from model capacity. Hyperparameter search and architecture details are provided in Appendices D.1 and D.2.

Transverse Field Ising Model.

$$\hat{H}_{\text{Ising}} = -J \sum_{i=1}^{N} \hat{\sigma}_i^z \hat{\sigma}_{i+1}^z - h \sum_{i=1}^{N} \hat{\sigma}_i^x. \tag{23}$$

Here, $J = 1$ is the ferromagnetic coupling strength and $h = 1$ is the transverse-field strength. This Hamiltonian is a standard sign-problem-free benchmark: the diagonal interaction term favors aligned spin configurations, while the transverse field introduces quantum fluctuations by flipping individual spins. Therefore, Proposition 3.1 holds exactly.

Heisenberg J_1–J_2 Chain.

$$\hat{H}_{J_1\text{–}J_2} = J_1 \sum_{i=1}^{N} \hat{\mathbf{S}}_i \cdot \hat{\mathbf{S}}_{i+1} + J_2 \sum_{i=1}^{N} \hat{\mathbf{S}}_i \cdot \hat{\mathbf{S}}_{i+2}, \qquad J_1 > 0,\ J_2 > 0. \tag{24}$$

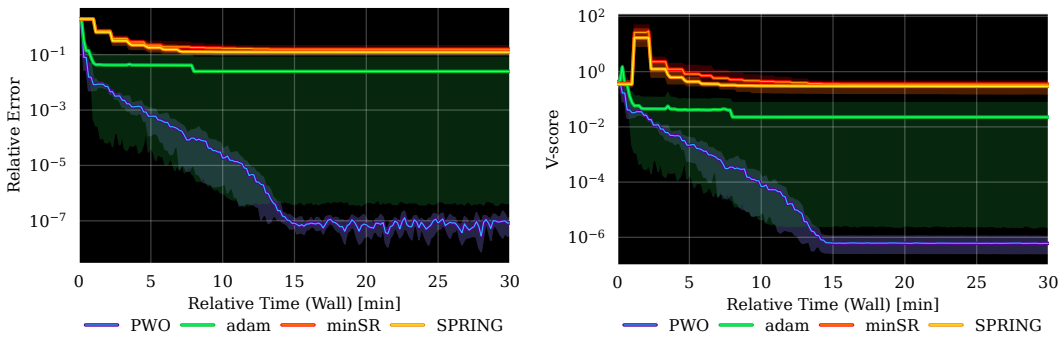


Figure 3: Comparison of PWO, Adam, minSR, and SPRING on the Heisenberg J_1–J_2 chain over 10 random seeds. MinSR is highly unstable on this Hamiltonian, with 6 out of 10 runs producing NaNs. All methods were run with 1024 samples on a single NVIDIA L40S GPU.

We set $J_1 = 1$ and $J_2 = 0.5$, *i.e.*, $J_2/J_1 = 0.5$. At this ratio the next-nearest-neighbor interaction maximally frustrates the nearest-neighbor antiferromagnetic coupling, placing the system at the Majumdar–Ghosh point, where a highly entangled ground state with a complex sign structure makes both variational optimization and Monte Carlo sampling particularly challenging [Sorella, 1998]. Unlike $\hat{H}_{\text{Ising}}$, the Hamiltonian $\hat{H}_{J_1-J_2}$ requires a nontrivial sign structure in the computational basis; we use a complex-valued parameterization so that PWO can learn this structure through its phase channel. This setting evaluates whether the clipped phase surrogate (Eq. (17)) remains a useful practical heuristic. We make an analogous study for $J_1 = 0.25$ and $J_2 = 0$ (the so-called Heisenberg chain) in Appendix E.1.

All results are computed over 10 random seeds. Since several runs reach high-precision errors, numerical instabilities can produce extreme outliers. Mean-and-standard-deviation summaries are therefore poorly suited to this regime, as a single unstable run can distort both the central estimate and the uncertainty band. Following Agarwal et al. [2022], we instead report interquartile statistics: at each evaluation time, curves show the interquartile mean (IQM), computed by averaging the middle 50% of seeds after discarding the lowest and highest quartiles. Shaded regions denote the interquartile range, from the 25th to the 75th percentile. This provides robust central estimates while still showing run-to-run variability. The figures with all individual seeds plotted are shown in Appendices E.1-E.3. We discuss limitations of our approach in Appendix C.4. Ablation studies in Appendix C.5 show that this improvement arises specifically from combining sample reuse with trust-region clipping.

5.2 Convergence and Stability Results

Figure 2 shows that PWO converges the fastest of the four optimizers on the 1D Ising model, reaching relative error 10^{-7} in approximately 5 minutes, while minSR requires approximately 30 minutes to reach the same accuracy. The V-score follows the same trend: PWO rapidly suppresses energy fluctuations, indicating that the proximal objective is not merely improving the energy estimate, but driving the state toward an eigenstate. Adam and SPRING also make steady progress, but require more wall-clock time to reach the same accuracy. minSR is stable on this easier Hamiltonian, but its per-step cost makes it substantially slower in wall-clock time.

On the J_1–J_2 Hamiltonian in Figure 3, PWO reaches relative error 10^{-7} in approximately 15 minutes while maintaining a steadily decreasing V-score, showing that the learned state is converging in both energy and variance. In contrast, Adam plateaus several orders of magnitude above PWO due to outliers. Also, the PWO contrast with minSR and SPRING is sharper than in the Ising case: minSR is numerically unstable, with 6 out of 10 runs producing NaNs, and both remain at a relative error of 10^{-1} after 30 minutes. Despite using curvature-inspired updates, these methods struggle to make progress and exhibit numerical instability. This is the strongest empirical evidence for the central claim of the paper: a PPO-style proximal objective can retain the low wall-clock cost of first-order optimization while providing the stability needed to train expressive NQS on difficult quantum Hamiltonians. Experiments on the Heisenberg chain in Appendix E.1 offer the same conclusion. We also plot all individual seed curves without aggregation or filtering for greater clarity and visibility in Appendix E.1.

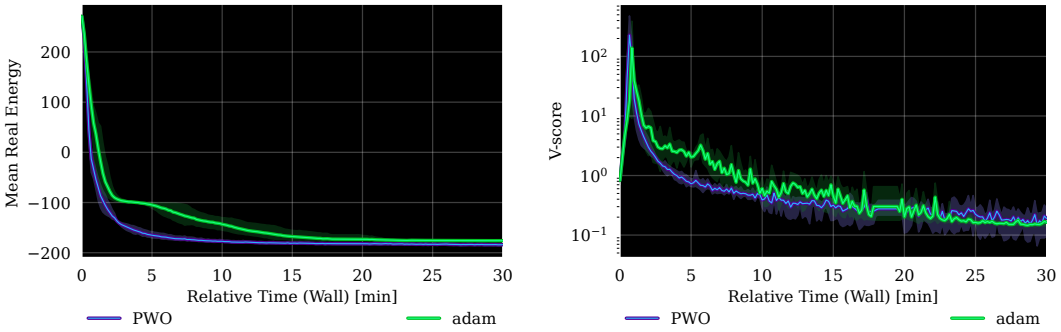


Figure 4: Two-dimensional frustrated J_1–J_2 Heisenberg model on the 10×10 lattice. Left: mean real energy. Right: V-score. PWO reaches lower energies faster than Adam and maintains a lower variance-based error signal over the same wall-clock budget.

5.3 Two-dimensional J_1–J_2 model on the square lattice

We next test whether PWO remains effective beyond one-dimensional chains on a frustrated square-lattice J_1–J_2 Heisenberg model with periodic boundary conditions and system size of 10×10. We use a complex patch-autoregressive transformer with two-dimensional RoPE and enforce the zero-magnetization sector. Since exact diagonalization is intractable at this size, we compare optimizers using the mean real energy and the V-score. As shown in Fig. 4, PWO decreases the energy faster than Adam and reaches a lower-energy regime within the same wall-clock budget. The V-score follows the same trend, indicating that the proximal objective also improves stability on this frustrated two-dimensional benchmark. Individual seed runs are shown in Appendix E.1.

5.4 Scalability Experiments

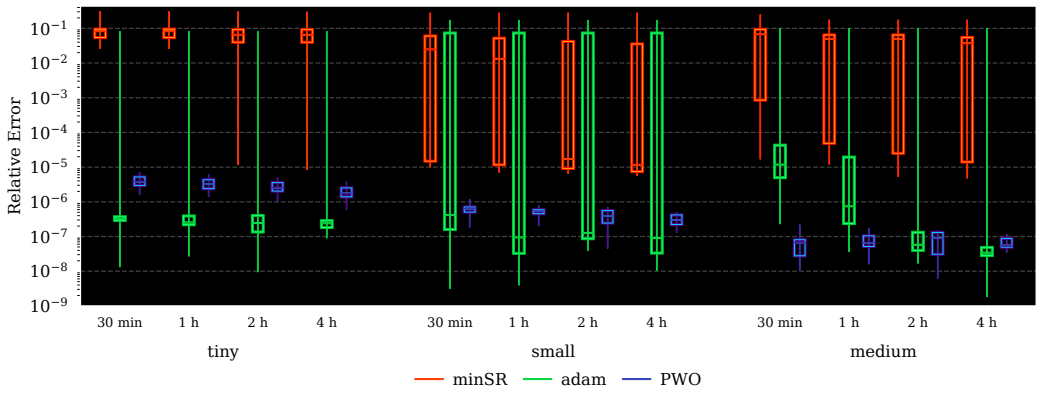


Figure 5: Scaling comparison across model sizes and optimization methods for the Heisenberg J_1–J_2 chain. Boxplots show the relative error distribution over seeds, with boxes indicating the interquartile range and lines, the minimum and maximum. Results are grouped by model size ($\sim 10^4$, $\sim 10^5$ and $\sim 10^6$ parameters) and wall-clock time, and run on a single NVIDIA A100 GPU.

Figure 5 compares the scaling behavior of PWO, Adam, and minSR for increasing NQS model sizes on the Heisenberg J_1–J_2 chain (see Appendix D.7 for details). PWO exhibits consistent stability and monotonic energy improvement as NQS size grows, reaching near-final energies within the first 30 minutes. In contrast, minSR is both unstable and never reaches competitive energy values. Adam performs best for the tiny network, but degrades in stability and efficiency for small and medium NQS, where PWO is both faster and more reliable. In particular, achieving comparable energy scales with the medium NQS requires roughly $4\times$ more time with Adam than with PWO. In Appendix C.2, we show that PWO converges faster than other optimizers because it does more optimization steps per unit of time. Although Adam attains lower energies after extended training of 4 hours, the observed scaling trend suggests that this advantage may diminish for yet larger networks. In Appendix E.2 we perform additional scaling experiments on the number of samples and system size.

5.5 NQS Fine-tuning of Large Language Models

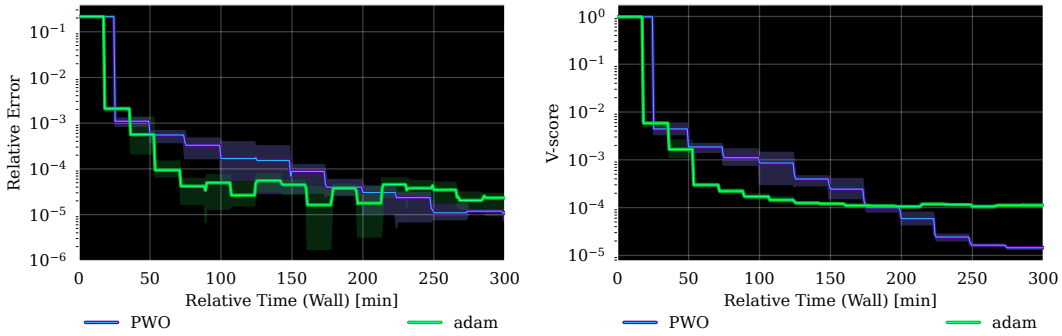


Figure 6: Fine-tuning curves of a 1.5B-parameter RWKV7LLM on the 1-D Ising model.

Figure 6 shows fine-tuning results for a 1.5B-parameter RWKV-7 model on the transverse-field Ising chain. PWO achieves a lower final relative error and V-score than Adam, while maintaining stable training across the full wall-clock budget. This improvement is not meant to establish that LLMs or large networks [Moss et al., 2026] are the right inductive bias for one-dimensional Hamiltonians; rather, it shows that the proximal objective is effective when the ansatz is scaled by more than three orders of magnitude beyond existing NQS in the literature [Rende et al., 2025].

6 Related Work

Reinforcement Learning and Ground State Search. Ground state search has been treated as an optimal control problem via Feynman-Kac representations [Barr et al., 2020], and RL frameworks have been developed for lattice models by treating stoquastic Hamiltonians as reward functions [Gispen and Lamacraft, 2022]. Recent meta-learning approaches [Zhao et al., 2019, Jae et al., 2025] leverage RL to optimize the quantum state learning process itself. However, these works apply RL by either reformulating the problem theoretically or optimizing the learning procedure. We instead use RL as the primary optimization framework applying it directly on the NQS ansatz.

Importance Sampling and Trust Regions. Sample reuse and proximal objectives have been previously explored for non-autoregressive NQS optimization. Yang et al. [2020] introduced importance sampling gradient optimization (ISGO), which reuses Markov chain samples across multiple gradient updates by reweighting local energies with the probability ratio between the current and reference wavefunctions. Building on this, Chen and Xue [2022] introduced a PPO-inspired a cosine penalty on phase variation. Our work formalizes the analogy with RL in Proposition 3.1, uses a different clipping objective directly on the imaginary gradient instead of a penalty, and makes the policy-gradient correspondence exact, allowing stable training of billion-parameter models for the first time.

7 Conclusion

We introduced Proximal Wavefunction Optimization (PWO), a scalable first-order optimizer for neural quantum states (NQS) derived from an explicit connection between variational Monte Carlo and reinforcement learning (RL). In the fixed-phase stoquastic setting, we showed that variational energy minimization can be written as an advantage policy-gradient objective over the Born distribution, where spin configurations play the role of actions and centered local energies act as advantages. This reformulates NQS optimization into a trust-region policy-optimization problem, motivating a PPO-style algorithm that clips probability-ratio changes in the amplitude channel and extends naturally to complex wavefunctions through a clipped phase-increment surrogate. Across spin-chain benchmarks, PWO improves wall-clock convergence and stability over Adam, minSR, and SPRING, with the largest gains in frustrated regimes where optimization is most fragile. Finally, by fine-tuning a 1.5B-parameter RWKV-7 model as an autoregressive NQS, we bring NQS training into the regime of modern large-scale sequence modeling. More broadly, our results bridge the RL and NQS communities, enabling further mutual inspiration. By unlocking large autoregressive NQS, PWO opens the door to studying whether scale alone, in the spirit of the "bitter lesson", can overcome current barriers in quantum many-body simulation.

Reproducibility Statement

Code for reproducing our experiments is available at https://github.com/jduquevan/hyperscalenqs. Full experimental details, including hyperparameter searches, model architectures, optimizer settings, sample budgets, and scaling configurations, are provided in Appendix D.

Acknowledgements

This publication is part of the Quantum Limits project with file number 1.1016 of the research programme SUMMIT which is (partly) financed by the Dutch Research Council (NWO) under the grant https://doi.org/10.61686/EBSOG21615. Also by the Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korean government(MSIT) (No. RS-2024-00457882, National AI Research Lab Project). This publication is also part of the project Optimal Digital-Analog Quantum Circuits with File No. NGF.1582.22.026 of the research programme NGF Quantum Delta NL 2022, which is (partly) financed by the Dutch Research Council (NWO) and the Dutch National Growth Fund initiative Quantum Delta NL. Juan Agustín Duque is supported by the St-Pierre-Larochelle Scholarship at the University of Montreal and by Aaron Courville's CIFAR AI Chair in Representations that Generalize Systematically.

References

Alekh Agarwal, Nan Jiang, Sham Kakade, and Wen Sun. *Reinforcement Learning: Theory and Algorithms*. Draft/Preprint, 2021. URL https://rltheorybook.github.io/.

Rishabh Agarwal, Max Schwarzer, Pablo Samuel Castro, Aaron Courville, and Marc G. Bellemare. Deep reinforcement learning at the edge of the statistical precipice. *arXiv preprint arXiv:2108.13264*, 2022. doi: 10.48550/arXiv.2108.13264. URL https://arxiv.org/abs/2108.13264.

Shun-ichi Amari. Natural gradient works efficiently in learning. *Neural Computation*, 10(2): 251–276, 1998. doi: 10.1162/089976698300017746. URL https://doi.org/10.1162/089976698300017746.

Adil Attar, Amine M. Aboussalah, and Mohamed Hibat-Allah. Quantum geometric tensor preconditioning for stable training of recurrent neural quantum states. *arXiv preprint arXiv:2608.18065*, 2026. doi: 10.48550/arXiv.2608.18065. URL http://arxiv.org/abs/2608.18065.

Ariel Barr, Willem Gispen, and Austen Lamacraft. Quantum ground states from reinforcement learning. In *Proceedings of the First Mathematical and Scientific Machine Learning Conference*, volume 107 of *Proceedings of Machine Learning Research*, pages 635–653, 2020. URL https://proceedings.mlr.press/v107/barr20a.html.

Giuseppe Carleo and Matthias Troyer. Solving the quantum many-body problem with artificial neural networks. *Science*, 355(6325):602–606, 2017. doi: 10.1126/science.aag2302. URL https://doi.org/10.1126/science.aag2302.

Ao Chen and Markus Heyl. Empowering deep neural quantum states through efficient optimization. *Nature Physics*, 20(9):1476–1481, 2024. doi: 10.1038/s41567-024-02566-1. URL https://doi.org/10.1038/s41567-024-02566-1.

Feng Chen and Ming Xue. Neural network quantum state with proximal optimization: A ground-state searching scheme based on variational Monte Carlo. *arXiv preprint arXiv:2210.16493*, 2022. doi: 10.48550/arXiv.2210.16493. URL https://arxiv.org/abs/2210.16493.

Kenny Choo, Antonio Mezzacapo, and Giuseppe Carleo. Fermionic neural-network states for ab-initio electronic structure. *Nature Communications*, 11(1):2368, 2020. doi: 10.1038/s41467-020-15724-9. URL https://doi.org/10.1038/s41467-020-15724-9.

Anna Dawid, Julian Arnold, Borja Requena, Alexander Gresch, Marcin Płodzień, Kaelan Donatella, Kim A. Nicoli, Paolo Stornati, Rouven Koch, Miriam Büttner, Robert Okuła, Gorka Muñoz Gil, Rodrigo A. Vargas-Hernández, Alba Cervera-Lierta, Juan Carrasquilla, Vedran Dunjko, Marylou Gabrié, Patrick Huembeli, Evert van Nieuwenburg, Filippo Vicentini, Lei Wang, Sebastian J. Wetzel, Giuseppe Carleo, Eliška Greplová, Roman Krems, Florian Marquardt, Michał Tomza,

Maciej Lewenstein, and Alexandre Dauphin. *Machine Learning in Quantum Sciences*. Cambridge University Press, 2025. ISBN 9781009504935. doi: 10.1017/9781009504942. URL https://doi.org/10.1017/9781009504942.

Luigi Del Debbio, Gian Mario Manca, and Ettore Vicari. Critical slowing down of topological modes. *Physics Letters B*, 594(3–4):315–323, 2004. doi: 10.1016/j.physletb.2004.05.038. URL https://doi.org/10.1016/j.physletb.2004.05.038.

Mehdi Drissi, James W. T. Keeble, Javier Rozalén Sarmiento, and Arnau Rios. Second-order optimization strategies for neural network quantum states. *Philosophical Transactions of the Royal Society A*, 382(2275):20240057, 2024. doi: 10.1098/rsta.2024.0057. URL https://doi.org/10.1098/rsta.2024.0057.

Chloé Gauvin-Ndiaye, Joseph Tindall, Javier Robledo Moreno, and Antoine Georges. Mott transition and volume law entanglement with neural quantum states. *Physical Review Letters*, 134(7): 076502, 2025. doi: 10.1103/PhysRevLett.134.076502. URL https://doi.org/10.1103/PhysRevLett.134.076502.

Willem Gispen and Austen Lamacraft. Ground states of quantum many body lattice models via reinforcement learning. In *Proceedings of the 2nd Mathematical and Scientific Machine Learning Conference*, volume 145 of *Proceedings of Machine Learning Research*, pages 369–385, 2022. URL https://proceedings.mlr.press/v145/gispen22a.html.

Gil Goldshlager, Nilin Abrahamsen, and Lin Lin. A Kaczmarz-inspired approach to accelerate the optimization of neural network wavefunctions. *Journal of Computational Physics*, 516: 113351, 2024. doi: 10.1016/j.jcp.2024.113351. URL https://doi.org/10.1016/j.jcp.2024.113351.

Jan Hermann, Zeno Schätzle, and Frank Noé. Deep-neural-network solution of the electronic Schrödinger equation. *Nature Chemistry*, 12(10):891–897, 2020. doi: 10.1038/s41557-020-0544-y. URL https://doi.org/10.1038/s41557-020-0544-y.

Mohamed Hibat-Allah, Martin Ganahl, Lauren E. Hayward, Roger G. Melko, and Juan Carrasquilla. Recurrent neural network wave functions. *Physical Review Research*, 2(2):023358, 2020. doi: 10.1103/PhysRevResearch.2.023358. URL https://doi.org/10.1103/PhysRevResearch.2.023358.

Mohamed Hibat-Allah, Estelle M. Inack, Roeland Wiersema, Roger G. Melko, and Juan Carrasquilla. Variational neural annealing. *Nature Machine Intelligence*, 3(11):952–961, 2021. ISSN 2522-5839. doi: 10.1038/s42256-021-00401-3. URL http://dx.doi.org/10.1038/s42256-021-00401-3.

Roger A. Horn and Charles R. Johnson. *Matrix Analysis*. Cambridge University Press, Cambridge, 2 edition, 2012. ISBN 9780521839402. doi: 10.1017/CBO9780511810817. URL https://doi.org/10.1017/CBO9780511810817.

Eduardo Ibarra-García-Padilla, Hannah Lange, Roger G. Melko, Richard T. Scalettar, Juan Carrasquilla, Annabelle Bohrdt, and Ehsan Khatami. Autoregressive neural quantum states of Fermi Hubbard models. *Physical Review Research*, 7:013122, 2025. doi: 10.1103/PhysRevResearch.7.013122. URL https://link.aps.org/doi/10.1103/PhysRevResearch.7.013122.

Jeongwoo Jae, Jeonghoon Hong, Jinho Choo, and Yeong-Dae Kwon. Reinforcement learning to learn quantum states for Heisenberg scaling accuracy. *Advanced Quantum Technologies*, 8(10), 2025. doi: 10.1002/qute.202500206. URL https://doi.org/10.1002/qute.202500206.

Sham Kakade and John Langford. Approximately optimal approximate reinforcement learning. In *Proceedings of the Nineteenth International Conference on Machine Learning*, pages 267–274, 2002. URL https://dl.acm.org/doi/10.5555/645531.656005.

Sham M Kakade. A natural policy gradient. In T. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems*, volume 14. MIT Press, 2001. URL https://proceedings.neurips.cc/paper_files/paper/2001/file/4b86abe48d358ecf194c56c69108433e-Paper.pdf.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. doi: 10.48550/arXiv.1412.6980. URL https://arxiv.org/abs/1412.6980.

Wladislaw Krinitsin, Nikita Alert, Matteo Rizzi, and Markus Schmitt. Comment on "Beyond-classical computation in quantum simulation". *arXiv preprint arXiv:2607.08811*, 2026. doi: 10.48550/arXiv.2607.08811. URL http://arxiv.org/abs/2607.08811.

Hannah Lange, Anka Van de Walle, Atiye Abedinnia, and Annabelle Bohrdt. From architectures to applications: a review of neural quantum states. *Quantum Science and Technology*, 9(4):040501, sep 2024. doi: 10.1088/2058-9565/ad7168. URL https://doi.org/10.1088/2058-9565/ad7168.

Jing Liu, Ying Tang, and Pan Zhang. Efficient optimization of variational autoregressive networks with natural gradient. *Physical Review E*, 111(2):025304, 2025. doi: 10.1103/PhysRevE.111.025304. URL https://doi.org/10.1103/PhysRevE.111.025304.

James Martens and Roger Grosse. Optimizing neural networks with Kronecker-factored approximate curvature. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 2408–2417, 2015. URL https://proceedings.mlr.press/v37/martens15.html.

James Martens, Jimmy Ba, and Matt Johnson. Kronecker-factored curvature approximations for recurrent neural networks. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=HyMTkQZAb.

Ejaaz Merali, Mohamed Hibat-Allah, Mohammad Kohandel, Richard T. Scalettar, and Ehsan Khatami. Parallel scan recurrent neural quantum states for scalable variational Monte Carlo. *arXiv preprint arXiv:2605.13807*, 2026. doi: 10.48550/arXiv.2605.13807. URL https://arxiv.org/abs/2605.13807.

M. Schuyler Moss, Roeland Wiersema, Mohamed Hibat-Allah, Juan Carrasquilla, and Roger G. Melko. Leveraging recurrence in neural network wavefunctions for large-scale simulations of Heisenberg antiferromagnets on the square lattice. *Physical Review B*, 112(13):134450, 2025. ISSN 2469-9950. doi: 10.1103/6ccd-wzhz. URL https://link.aps.org/doi/10.1103/6ccd-wzhz.

M. Schuyler Moss, Alev Orfi, Christopher Roth, Anirvan M. Sengupta, Antoine Georges, Dries Sels, Anna Dawid, and Agnes Valenti. Double descent: When do neural quantum states generalize? *Physical Review E*, 113(4):045303, 2026. ISSN 2470-0045. doi: 10.1103/cwmj-fxr4. URL https://link.aps.org/doi/10.1103/cwmj-fxr4.

Yusuke Nomura and Masatoshi Imada. Dirac-type nodal spin liquid revealed by refined quantum many-body solver using neural-network wave function, correlation ratio, and level spectroscopy. *Physical Review X*, 11(3):031034, 2021. doi: 10.1103/PhysRevX.11.031034. URL https://doi.org/10.1103/PhysRevX.11.031034.

Jannes Nys, Gabriel Pescia, Alessandro Sinibaldi, and Giuseppe Carleo. Ab-initio variational wave functions for the time-dependent many-electron Schrödinger equation. *Nature Communications*, 15(1):9404, 2024. doi: 10.1038/s41467-024-53672-w. URL https://doi.org/10.1038/s41467-024-53672-w.

Bo Peng, Ruichong Zhang, Daniel Goldstein, Eric Alcaide, Xingjian Du, Haowen Hou, Jiaju Lin, Jiaxing Liu, Janna Lu, William Merrill, Guangyu Song, Kaifeng Tan, Saiteja Utpala, Nathan Wilce, Johan S. Wind, Tianyi Wu, Daniel Wuttke, and Christian Zhou-Zheng. RWKV-7 “goose” with expressive dynamic state evolution. *arXiv preprint arXiv:2503.14456*, 2025. doi: 10.48550/arXiv.2503.14456. URL https://arxiv.org/abs/2503.14456.

Gabriel Pescia, Jannes Nys, Jane Kim, Alessandro Lovato, and Giuseppe Carleo. Message-passing neural quantum states for the homogeneous electron gas. *Physical Review B*, 110(3):035108, 2024. doi: 10.1103/PhysRevB.110.035108. URL https://doi.org/10.1103/PhysRevB.110.035108.

David Pfau, James S. Spencer, Alexander G. D. G. Matthews, and W. M. C. Foulkes. Ab initio solution of the many-electron Schrödinger equation with deep neural networks. *Physical Review Research*, 2(3):033429, 2020. doi: 10.1103/PhysRevResearch.2.033429. URL https://doi.org/10.1103/PhysRevResearch.2.033429.

Yi Ren and Donald Goldfarb. Efficient subsampled Gauss-Newton and natural gradient methods for training neural networks. *arXiv preprint arXiv:1906.02353*, 2019. doi: 10.48550/arXiv.1906.02353. URL https://arxiv.org/abs/1906.02353.

Riccardo Rende, Luciano Loris Viteritti, Lorenzo Bardone, Federico Becca, and Sebastian Goldt. A simple linear algebra identity to optimize large-scale neural network quantum states. *Communications Physics*, 7(1):260, 2024. doi: 10.1038/s42005-024-01732-4. URL https://doi.org/10.1038/s42005-024-01732-4.

Riccardo Rende, Luciano Loris Viteritti, Federico Becca, Antonello Scardicchio, Alessandro Laio, and Giuseppe Carleo. Foundation neural-networks quantum states as a unified ansatz for multiple hamiltonians. *Nature Communications*, 16(1):7213, 2025. doi: 10.1038/s41467-025-62098-x. URL https://doi.org/10.1038/s41467-025-62098-x.

Herbert Robbins and Sutton Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3):400–407, 1951. doi: 10.1214/aoms/1177729586. URL https://doi.org/10.1214/aoms/1177729586.

Bidipta Sarkar, Mattie Fellows, Juan Agustin Duque, Alistair Letcher, Antonio León Villares, Anya Sims, Clarisse Wibault, Dmitry Samsonov, Dylan Cope, Jarek Liesen, Kang Li, Lukas Seier, Theo Wolf, Uljad Berdica, Valentin Mohl, Alexander David Goldie, Aaron Courville, Karin Sevegnani, Shimon Whiteson, and Jakob Nicolaus Foerster. Evolution strategies at the hyperscale. *arXiv preprint arXiv:2511.16652*, 2026. doi: 10.48550/arXiv.2511.16652. URL https://arxiv.org/abs/2511.16652.

John Schulman, Sergey Levine, Philipp Moritz, Michael I. Jordan, and Pieter Abbeel. Trust region policy optimization. *arXiv preprint arXiv:1502.05477*, 2015. doi: 10.48550/arXiv.1502.05477. URL https://arxiv.org/abs/1502.05477.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. doi: 10.48550/arXiv.1707.06347. URL https://arxiv.org/abs/1707.06347.

Kristof T. Schütt, Huziel E. Sauceda, Pieter-Jan Kindermans, Alexandre Tkatchenko, and Klaus-Robert Müller. Schnet—a deep learning architecture for molecules and materials. *The Journal of Chemical Physics*, 148(24):241722, 2018. doi: 10.1063/1.5019779. URL https://doi.org/10.1063/1.5019779.

Or Sharir, Yoav Levine, Noam Wies, Giuseppe Carleo, and Amnon Shashua. Deep autoregressive models for the efficient variational simulation of many-body quantum systems. *Physical Review Letters*, 124(2):020503, 2020. doi: 10.1103/PhysRevLett.124.020503. URL https://doi.org/10.1103/PhysRevLett.124.020503.

Ahmedeo Shokry, Alessandro Santini, and Filippo Vicentini. When less is more: Approximating the quantum geometric tensor with block structures. *arXiv preprint arXiv:2510.08430*, 2025. doi: 10.48550/arXiv.2510.08430. URL https://arxiv.org/abs/2510.08430.

Leslie N. Smith and Nicholay Topin. Super-convergence: Very fast training of neural networks using large learning rates. *arXiv preprint arXiv:1708.07120*, 2018. doi: 10.48550/arXiv.1708.07120. URL https://arxiv.org/abs/1708.07120.

Sandro Sorella. Green function Monte Carlo with stochastic reconfiguration. *Physical Review Letters*, 80(20):4558–4561, 1998. doi: 10.1103/PhysRevLett.80.4558. URL https://doi.org/10.1103/PhysRevLett.80.4558.

Ingrid von Glehn, James S. Spencer, and David Pfau. A self-attention ansatz for ab-initio quantum chemistry. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=xveTeHVlF7j.

Jia-Qi Wang, Rong-Qiang He, and Zhong-Yi Lu. Generalized Lanczos method for systematic optimization of neural-network quantum states. *Physical Review B*, 113(8):085120, 2026. doi: 10.1103/PhysRevB.113.085120. URL https://doi.org/10.1103/PhysRevB.113.085120.

Ronald J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8(3–4):229–256, 1992. doi: 10.1007/BF00992696. URL https://doi.org/10.1007/BF00992696.

Ulli Wolff. Critical slowing down. *Nuclear Physics B - Proceedings Supplements*, 17:93–102, 1990. doi: 10.1016/0920-5632(90)90224-I. URL https://doi.org/10.1016/0920-5632(90)90224-I.

Dian Wu, Lei Wang, and Pan Zhang. Solving statistical mechanics using variational autoregressive networks. *Physical Review Letters*, 122:080602, Feb 2019. doi: 10.1103/PhysRevLett.122.080602. URL https://link.aps.org/doi/10.1103/PhysRevLett.122.080602.

Dian Wu, Riccardo Rossi, Filippo Vicentini, Nikita Astrakhantsev, Federico Becca, Xiaodong Cao, Juan Carrasquilla, Francesco Ferrari, Antoine Georges, Mohamed Hibat-Allah, Masatoshi Imada, Andreas M. Läuchli, Guglielmo Mazzola, Antonio Mezzacapo, Andrew Millis, Javier Robledo Moreno, Titus Neupert, Yusuke Nomura, Jannes Nys, Olivier Parcollet, Rico Pohle, Imelda Romero, Michael Schmid, J. Maxwell Silvester, Sandro Sorella, Luca F. Tocchio, Lei Wang, Steven R. White, Alexander Wietek, Qi Yang, Yiqi Yang, Shiwei Zhang, and Giuseppe Carleo. Variational benchmarks for quantum many-body problems. *Science*, 386(6719):296–301, 2024. doi: 10.1126/science.adg9774. URL https://doi.org/10.1126/science.adg9774.

Li Yang, Zhaoqi Leng, Guangyuan Yu, Ankit Patel, Wen-Jun Hu, and Han Pu. Deep learning-enhanced variational Monte Carlo method for quantum many-body physics. *Physical Review Research*, 2(1):012039, 2020. doi: 10.1103/PhysRevResearch.2.012039. URL https://doi.org/10.1103/PhysRevResearch.2.012039.

Kai-Wen Zhao, Wen-Han Kao, Kai-Hsin Wu, and Ying-Jer Kao. Generation of ice states through deep reinforcement learning. *Physical Review E*, 99:062106, Jun 2019. doi: 10.1103/PhysRevE.99.062106. URL https://link.aps.org/doi/10.1103/PhysRevE.99.062106.

Appendix

Table of Contents

A Dirac Notation **17**

B Mathematical Statements **18**
B.1 Proof of Proposition 3.1 . . . 18
B.2 Variational Gradient Derivation . . . 19
B.3 Variational Gradient Decomposition . . . 20
B.4 Exact Complex-Amplitude Decomposition . . . 20
B.5 Amplitude and Phase Trust Regions Control Infidelity . . . 22
B.6 PWO Surrogate Upper Bound and First-Order Consistency . . . 22
B.7 Clipped PWO Improvement Certificate . . . 24

C Additional Sections **26**
C.1 Computational Cost of Stochastic Reconfiguration and Improvements . . . 26
C.2 Relative Speed of PWO per Iteration . . . 26
C.3 Why RWKV? . . . 27
C.4 Limitations of PWO . . . 27
C.5 Ablation Studies . . . 27

D Experimental Details **30**
D.1 Hyperparameter search . . . 30
D.2 Neural Quantum State Architecture . . . 30
D.3 Ising Model . . . 31
D.4 Heisenberg Chain . . . 31
D.5 Heisenberg J_1–J_2 Chain . . . 32
D.6 Two-dimensional J_1–J_2 square-lattice experiment . . . 32
D.7 Scaling Experiment . . . 33
D.8 RWKV-7 on Ising Model . . . 34

E Additional Figures **35**
E.1 Individual Seed Plots for All Hamiltonias . . . 35
E.2 Scaling Samples and System Sizes . . . 36
E.3 Individual Seed Plots for RWKV7 Fine-tuning . . . 37

A Dirac Notation

This appendix provides a self-contained introduction to the Dirac (bra-ket) notation used throughout the paper.

Hilbert space. A quantum state lives in a complex vector space $\mathcal{H}$ called a *Hilbert space*, equipped with an inner product. For a system of N spin-$1/2$ particles, the Hilbert space is $\mathcal{H} = (\mathbb{C}^2)^{\otimes N}$, which has dimension 2^N. A natural orthonormal basis is the *computational* (spin) basis, whose elements $|\mathbf{s}\rangle$ are indexed by binary spin configurations $\mathbf{s} \in \{\pm 1\}^N$.

Ket vectors. A quantum state is written as a *ket* $|\psi\rangle$, which is simply a column vector in $\mathcal{H}$. Any state can be expanded in the computational basis,

$$|\psi\rangle = \sum_{\mathbf{s}\in\{\pm1\}^N} \psi(\mathbf{s})\ |\mathbf{s}\rangle\,, \tag{25}$$

where each coefficient $\psi(\mathbf{s}) \in \mathbb{C}$ is the *amplitude* of configuration $\mathbf{s}$. The vector $|\psi\rangle$ is therefore completely specified by the mapping $\mathbf{s} \mapsto \psi(\mathbf{s})$, which NQS parameterize with a neural network.

Bra vectors and inner product. The *bra* $\langle\psi|$ is the conjugate transpose (Hermitian adjoint) of $|\psi\rangle$,

$$\langle\psi| = |\psi\rangle^\dagger\,. \tag{26}$$

The *inner product* of two states is written $\langle\phi|\psi\rangle$, which evaluates to the complex number $\sum_{\mathbf{s}} \phi(\mathbf{s})^* \psi(\mathbf{s})$. When $\phi = \psi$ this gives the squared norm $\langle\psi|\psi\rangle = \sum_{\mathbf{s}} |\psi(\mathbf{s})|^2 \geq 0$, with equality only for the zero vector. The basis states are orthonormal:

$$\langle\mathbf{s}|\mathbf{s}'\rangle = \delta_{\mathbf{s},\mathbf{s}'}, \tag{27}$$

where $\delta_{\mathbf{s},\mathbf{s}'}$ is the Kronecker delta.

Completeness relation. The basis states form a complete set, meaning the *identity operator* $\mathbb{1}$ can be resolved as

$$\sum_{\mathbf{s}} |\mathbf{s}\rangle\,\langle\mathbf{s}| = \mathbb{1}. \tag{28}$$

This is used repeatedly in derivations (e.g. Appendix B.2) to insert a basis expansion and convert abstract operator equations into sums over spin configurations.

Operators and matrix elements. A quantum *operator* $\hat{A}$ is a linear map $\mathcal{H} \to \mathcal{H}$, represented as a $2^N \times 2^N$ matrix in the computational basis. The entry at row $\mathbf{s}$ and column $\mathbf{s}'$ is the *matrix element*

$$\langle\mathbf{s}|\hat{A}|\mathbf{s}'\rangle = \langle\mathbf{s}|\left(\hat{A}\,|\mathbf{s}'\rangle\right) = (\mathbf{A})_{\mathbf{s},\mathbf{s}'}. \tag{29}$$

An operator is *Hermitian* (or self-adjoint) if $\hat{A} = \hat{A}^\dagger$, which implies real eigenvalues. The Hamiltonian $\hat{H}$ is always Hermitian because energy is a real-valued observable.

Expectation value. The expected value of an observable $\hat{A}$ in state $|\psi\rangle$ is

$$\langle\hat{A}\rangle_\psi := \frac{\langle\psi|\hat{A}|\psi\rangle}{\langle\psi|\psi\rangle}. \tag{30}$$

For the Hamiltonian, this gives the variational energy $E[\psi] = \langle\psi|\hat{H}|\psi\rangle\,/\,\langle\psi|\psi\rangle$ that NQS minimizes. When $|\psi\rangle$ is normalized ($\langle\psi|\psi\rangle = 1$), the denominator can be dropped.

Summary table.

Dirac symbol	Linear-algebra equivalent	Meaning in this paper
$\|\psi\rangle$	column vector in $\mathbb{C}^{2^N}$	quantum state / wavefunction
$\langle\psi\|$	conjugate row vector	dual / adjoint state
$\langle\phi\|\psi\rangle$	$\boldsymbol{\phi}^\dagger\boldsymbol{\psi}$ (dot product)	inner product; overlap of two states
$\langle\psi\|\psi\rangle$	$\|\boldsymbol{\psi}\|^2$	squared norm; $= 1$ for normalized states
$\|\mathbf{s}\rangle$	standard basis vector $e_{\mathbf{s}}$	spin-configuration basis state
$\langle\mathbf{s}\|\hat{H}\|\mathbf{s}'\rangle$	$(\mathbf{H})_{\mathbf{s},\mathbf{s}'}$ matrix entry	Hamiltonian matrix element
$\langle\psi\|\hat{H}\|\psi\rangle$	$\boldsymbol{\psi}^\dagger\mathbf{H}\boldsymbol{\psi}$ (quadratic form)	(unnormalized) variational energy
$\sum_{\mathbf{s}} \|\mathbf{s}\rangle\,\langle\mathbf{s}\|$	$\mathbb{1}_{2^N}$ (identity matrix)	completeness / resolution of identity

B Mathematical Statements

B.1 Proof of Proposition 3.1

Proposition 3.1 (Policy-gradient form of variational energy minimization)**.** *Assume the Hamiltonian is stoquastic, i.e. its matrix representation in a chosen computational basis has non-positive off-diagonal elements. Then we can assume $f_{\boldsymbol{\theta}} = \log \psi_{\boldsymbol{\theta}} = \log |\psi_{\boldsymbol{\theta}}|$, and the gradient of the variational energy can be written in policy-gradient form as*

$$\boldsymbol{\nabla}_{\boldsymbol{\theta}} E[\psi_{\boldsymbol{\theta}}] = \mathbb{E}_{\mathbf{s}\sim\mathcal{P}_{\boldsymbol{\theta}}}\big[\big(E^{\mathrm{loc}}_{\boldsymbol{\theta}}(\mathbf{s}) - \mathbb{E}_{\mathbf{s}\sim\mathcal{P}_{\boldsymbol{\theta}}}[E^{\mathrm{loc}}_{\boldsymbol{\theta}}(\mathbf{s})]\big)\, \boldsymbol{\nabla}_{\boldsymbol{\theta}} \log \mathcal{P}_{\boldsymbol{\theta}}(\mathbf{s})\big]. \tag{14}$$

In particular, variational energy minimization is equivalent to an advantage policy-gradient update over configurations. For a proof, see Appendix B.1.

Proof. We follow the proof strategy of the policy gradient theorem [Agarwal et al., 2021] to connect the standard NQS/VMC gradient estimator with REINFORCE-style updates when the probability distribution is parameterized autoregressively.

Differentiating gives

$$\begin{aligned}
\boldsymbol{\nabla}_{\boldsymbol{\theta}} E[\psi_{\boldsymbol{\theta}}] &= \boldsymbol{\nabla}_{\boldsymbol{\theta}} \sum_{\mathbf{s}} \mathcal{P}_{\boldsymbol{\theta}}(\mathbf{s})\, E^{\mathrm{loc}}_{\boldsymbol{\theta}}(\mathbf{s}) && (31)\\
&= \sum_{\mathbf{s}} \boldsymbol{\nabla}_{\boldsymbol{\theta}} \mathcal{P}_{\boldsymbol{\theta}}(\mathbf{s})\, E^{\mathrm{loc}}_{\boldsymbol{\theta}}(\mathbf{s}) \;+\; \sum_{\mathbf{s}} \mathcal{P}_{\boldsymbol{\theta}}(\mathbf{s})\, \boldsymbol{\nabla}_{\boldsymbol{\theta}} E^{\mathrm{loc}}_{\boldsymbol{\theta}}(\mathbf{s}) && (32)\\
&= \sum_{\mathbf{s}} \mathcal{P}_{\boldsymbol{\theta}}(\mathbf{s})\, E^{\mathrm{loc}}_{\boldsymbol{\theta}}(\mathbf{s}) \boldsymbol{\nabla}_{\boldsymbol{\theta}} \log \mathcal{P}_{\boldsymbol{\theta}}(\mathbf{s}) \;+\; \sum_{\mathbf{s}} \mathcal{P}_{\boldsymbol{\theta}}(\mathbf{s})\, \boldsymbol{\nabla}_{\boldsymbol{\theta}} E^{\mathrm{loc}}_{\boldsymbol{\theta}}(\mathbf{s}) && (33)\\
&= \mathbb{E}_{\mathbf{s}\sim\mathcal{P}_{\boldsymbol{\theta}}}\big[E^{\mathrm{loc}}_{\boldsymbol{\theta}}(\mathbf{s})\, \boldsymbol{\nabla}_{\boldsymbol{\theta}} \log \mathcal{P}_{\boldsymbol{\theta}}(\mathbf{s})\big] \;+\; \mathbb{E}_{\mathbf{s}\sim\mathcal{P}_{\boldsymbol{\theta}}}\big[\boldsymbol{\nabla}_{\boldsymbol{\theta}} E^{\mathrm{loc}}_{\boldsymbol{\theta}}(\mathbf{s})\big]. && (34)
\end{aligned}$$

We now analyze the second term. Define the log-derivative observable

$$\mathbf{O}(\mathbf{s}) := \boldsymbol{\nabla}_{\boldsymbol{\theta}} \log \psi_{\boldsymbol{\theta}}(\mathbf{s}),$$

then

$$\begin{aligned}
\mathbb{E}_{\mathbf{s}\sim\mathcal{P}_{\boldsymbol{\theta}}}\big[\boldsymbol{\nabla}_{\boldsymbol{\theta}} E^{\mathrm{loc}}_{\boldsymbol{\theta}}(\mathbf{s})\big] &= \mathbb{E}_{\mathbf{s}\sim\mathcal{P}_{\boldsymbol{\theta}}}\left[\boldsymbol{\nabla}_{\boldsymbol{\theta}} \sum_{\mathbf{s}'} \frac{\psi_{\boldsymbol{\theta}}(\mathbf{s}')}{\psi_{\boldsymbol{\theta}}(\mathbf{s})} \langle \mathbf{s}|\hat{H}|\mathbf{s}'\rangle\right] && (35)\\
&= \mathbb{E}_{\mathbf{s}\sim\mathcal{P}_{\boldsymbol{\theta}}}\left[\sum_{\mathbf{s}'} \langle \mathbf{s}|\hat{H}|\mathbf{s}'\rangle\, \frac{(\psi_{\boldsymbol{\theta}}(\mathbf{s})\boldsymbol{\nabla}_{\boldsymbol{\theta}}\psi_{\boldsymbol{\theta}}(\mathbf{s}') - \boldsymbol{\nabla}_{\boldsymbol{\theta}}\psi_{\boldsymbol{\theta}}(\mathbf{s})\psi_{\boldsymbol{\theta}}(\mathbf{s}'))}{\psi_{\boldsymbol{\theta}}(\mathbf{s})^2}\right] && (36)\\
&= \mathbb{E}_{\mathbf{s}\sim\mathcal{P}_{\boldsymbol{\theta}}}\left[\sum_{\mathbf{s}'} \langle \mathbf{s}|\hat{H}|\mathbf{s}'\rangle\, \frac{\psi_{\boldsymbol{\theta}}(\mathbf{s}')}{\psi_{\boldsymbol{\theta}}(\mathbf{s})}(\mathbf{O}(\mathbf{s}') - \mathbf{O}(\mathbf{s}))\right] && (37)\\
&= \sum_{\mathbf{s}} \mathcal{P}_{\boldsymbol{\theta}}(\mathbf{s}) \sum_{\mathbf{s}'} \hat{H}_{\mathbf{s},\mathbf{s}'} \frac{\psi_{\boldsymbol{\theta}}(\mathbf{s}')}{\psi_{\boldsymbol{\theta}}(\mathbf{s})}(\mathbf{O}(\mathbf{s}') - \mathbf{O}(\mathbf{s})), && (38)
\end{aligned}$$

where $\hat{H}_{\mathbf{s},\mathbf{s}'} = \langle \mathbf{s}|\hat{H}|\mathbf{s}'\rangle$, which gives the double-sum form

$$\mathbb{E}_{\mathbf{s}\sim\mathcal{P}_{\boldsymbol{\theta}}}\big[\boldsymbol{\nabla}_{\boldsymbol{\theta}} E^{\mathrm{loc}}_{\boldsymbol{\theta}}(\mathbf{s})\big] = \frac{1}{Z_{\boldsymbol{\theta}}} \sum_{\mathbf{s},\mathbf{s}'} \psi_{\boldsymbol{\theta}}(\mathbf{s})\psi_{\boldsymbol{\theta}}(\mathbf{s}')\, \hat{H}_{\mathbf{s},\mathbf{s}'}(\mathbf{O}(\mathbf{s}') - \mathbf{O}(\mathbf{s})). \tag{39}$$

At this point, the only remaining question is whether the second term in (39) vanishes. This happens when the Hamiltonian is stoquastic in the chosen basis. In that case, its matrix elements are real and Hermiticity implies $\hat{H}_{\mathbf{s},\mathbf{s}'} = \hat{H}_{\mathbf{s}',\mathbf{s}}$. Moreover, according to Perron-Frobenius [Horn and Johnson, 2012], all the components of the ground state wavefunction of a stoquastic Hamiltonian are real and strictly positive, so $\psi_{\boldsymbol{\theta}}(\mathbf{s}) = |\psi_{\boldsymbol{\theta}}(\mathbf{s})|$ and the log-derivative observable $\mathbf{O}(\mathbf{s})$ is real-valued.

Under this assumption, the factor $\psi_{\boldsymbol{\theta}}(\mathbf{s})\psi_{\boldsymbol{\theta}}(\mathbf{s}')\hat{H}_{\mathbf{s},\mathbf{s}'}$ in (39) is symmetric under exchanging $\mathbf{s}$ and $\mathbf{s}'$, whereas the difference $\mathbf{O}(\mathbf{s}') - \mathbf{O}(\mathbf{s})$ is antisymmetric. Therefore, each term in the double sum cancels with the term obtained by swapping $\mathbf{s}$ and $\mathbf{s}'$, and the whole sum is zero. Hence,

$$\mathbb{E}_{\mathbf{s}\sim\mathcal{P}_{\boldsymbol{\theta}}}\big[\boldsymbol{\nabla}_{\boldsymbol{\theta}} E^{\mathrm{loc}}_{\boldsymbol{\theta}}(\mathbf{s})\big] = 0. \tag{40}$$

Substituting back into (34) yields the pure score-function (policy-gradient) estimator:

$$\nabla_{\boldsymbol{\theta}} E[\psi_{\boldsymbol{\theta}}] = \mathbb{E}_{\mathbf{s}\sim\mathcal{P}_{\boldsymbol{\theta}}}\big[E^{\mathrm{loc}}_{\boldsymbol{\theta}}(\mathbf{s})\,\nabla_{\boldsymbol{\theta}} \log \mathcal{P}_{\boldsymbol{\theta}}(\mathbf{s})\big]. \tag{41}$$

Moreover, for any constant c that doesn't depend on $\mathbf{s}$,

$$\begin{aligned}
\mathbb{E}_{\mathbf{s}\sim\mathcal{P}_{\boldsymbol{\theta}}}[c\nabla_{\boldsymbol{\theta}} \log \mathcal{P}_{\boldsymbol{\theta}}(\mathbf{s})] &= c\sum_{\mathbf{s}} \mathcal{P}_{\boldsymbol{\theta}}(\mathbf{s})\nabla_{\boldsymbol{\theta}} \log \mathcal{P}_{\boldsymbol{\theta}}(\mathbf{s}) && (42)\\
&= c\sum_{\mathbf{s}} \nabla_{\boldsymbol{\theta}}\mathcal{P}_{\boldsymbol{\theta}}(\mathbf{s}) && (43)\\
&= c\nabla_{\boldsymbol{\theta}} \sum_{\mathbf{s}} \mathcal{P}_{\boldsymbol{\theta}}(\mathbf{s}) && (44)\\
&= c\nabla_{\boldsymbol{\theta}} 1 = 0. && (45)
\end{aligned}$$

So we may subtract any constant baseline without bias [Agarwal et al., 2021]. Choosing the baseline as the global energy $E[\psi_{\boldsymbol{\theta}}] = \mathbb{E}_{\mathbf{s}\sim\mathcal{P}_{\boldsymbol{\theta}}}[E^{\mathrm{loc}}_{\boldsymbol{\theta}}(\mathbf{s})]$ gives the form

$$\nabla_{\boldsymbol{\theta}} E[\psi_{\boldsymbol{\theta}}] = \mathbb{E}_{\mathbf{s}\sim\mathcal{P}_{\boldsymbol{\theta}}}\big[\big(E^{\mathrm{loc}}_{\boldsymbol{\theta}}(\mathbf{s}) - \mathbb{E}_{\mathbf{s}\sim\mathcal{P}_{\boldsymbol{\theta}}}[E^{\mathrm{loc}}_{\boldsymbol{\theta}}(\mathbf{s})]\big)\,\nabla_{\boldsymbol{\theta}} \log \mathcal{P}_{\boldsymbol{\theta}}(\mathbf{s})\big], \tag{46}$$

which is analogous to the Advantage form of the REINFORCE estimator [Agarwal et al., 2021]. □

B.2 Variational Gradient Derivation

In this section, we show how to derive the well known VMC-gradient in Eq. (11) [Carleo and Troyer, 2017] from Eq. (8). We start from the general expression for the energy, valid for any (not necessarily normalized) state $|\psi_{\boldsymbol{\theta}}\rangle$:

$$E[\psi_{\boldsymbol{\theta}}] = \frac{\langle\psi_{\boldsymbol{\theta}}|\hat{H}|\psi_{\boldsymbol{\theta}}\rangle}{\langle\psi_{\boldsymbol{\theta}}|\psi_{\boldsymbol{\theta}}\rangle}. \tag{47}$$

Differentiating with respect to θ_k using the quotient rule and the Hermiticity of $\hat{H}$ gives:

$$\partial_{\theta_k} E[\psi_{\boldsymbol{\theta}}] = \frac{2\,\mathrm{Re}\left\{(\partial_{\theta_k}\langle\psi_{\boldsymbol{\theta}}|)\hat{H}|\psi_{\boldsymbol{\theta}}\rangle\right\}}{\langle\psi_{\boldsymbol{\theta}}|\psi_{\boldsymbol{\theta}}\rangle} - E[\psi_{\boldsymbol{\theta}}]\frac{2\,\mathrm{Re}\left\{(\partial_{\theta_k}\langle\psi_{\boldsymbol{\theta}}|)|\psi_{\boldsymbol{\theta}}\rangle\right\}}{\langle\psi_{\boldsymbol{\theta}}|\psi_{\boldsymbol{\theta}}\rangle}. \tag{48}$$

Since the final result is norm-invariant, we now choose the convenient normalization $\langle\psi_{\boldsymbol{\theta}}|\psi_{\boldsymbol{\theta}}\rangle = 1$, which simplifies the expression to:

$$\partial_{\theta_k} E[\psi_{\boldsymbol{\theta}}] = 2\,\mathrm{Re}\left\{(\partial_{\theta_k}\langle\psi_{\boldsymbol{\theta}}|)\big(\hat{H} - E[\psi_{\boldsymbol{\theta}}]\big)|\psi_{\boldsymbol{\theta}}\rangle\right\}. \tag{49}$$

We insert the completeness relation $\sum_{\mathbf{s}} |\mathbf{s}\rangle\langle\mathbf{s}| = \mathbb{1}$ twice and use $\mathcal{P}_{\boldsymbol{\theta}}(\mathbf{s}) = |\psi_{\boldsymbol{\theta}}(\mathbf{s})|^2$:

$$\begin{aligned}
(\partial_{\theta_k}\langle\psi_{\boldsymbol{\theta}}|)\big(\hat{H} - E[\psi_{\boldsymbol{\theta}}]\big)|\psi_{\boldsymbol{\theta}}\rangle &= \sum_{\mathbf{s}} (\partial_{\theta_k}\langle\psi_{\boldsymbol{\theta}}|)|\mathbf{s}\rangle\,\langle\mathbf{s}|\big(\hat{H} - E[\psi_{\boldsymbol{\theta}}]\big)|\psi_{\boldsymbol{\theta}}\rangle && (50)\\
&= \sum_{\mathbf{s}} \partial_{\theta_k}\psi^*_{\boldsymbol{\theta}}(\mathbf{s})\left(\sum_{\mathbf{s}'}\langle\mathbf{s}|\hat{H}|\mathbf{s}'\rangle\langle\mathbf{s}'|\psi_{\boldsymbol{\theta}}\rangle - E[\psi_{\boldsymbol{\theta}}]\,\psi_{\boldsymbol{\theta}}(\mathbf{s})\right) && (51)\\
&= \sum_{\mathbf{s}} \psi^*_{\boldsymbol{\theta}}(\mathbf{s})O^*_k(\mathbf{s})\left(\sum_{\mathbf{s}'}\hat{H}_{\mathbf{s},\mathbf{s}'}\psi_{\boldsymbol{\theta}}(\mathbf{s}') - E[\psi_{\boldsymbol{\theta}}]\,\psi_{\boldsymbol{\theta}}(\mathbf{s})\right) && (52)\\
&= \sum_{\mathbf{s}} |\psi_{\boldsymbol{\theta}}(\mathbf{s})|^2\,O^*_k(\mathbf{s})\,\big(E^{\mathrm{loc}}_{\boldsymbol{\theta}}(\mathbf{s}) - E[\psi_{\boldsymbol{\theta}}]\big) && (53)\\
&= \mathbb{E}_{\mathbf{s}\sim\mathcal{P}_{\boldsymbol{\theta}}}\Big[O^*_k(\mathbf{s})\Big(E^{\mathrm{loc}}_{\boldsymbol{\theta}}(\mathbf{s}) - \mathbb{E}_{\mathbf{s}\sim\mathcal{P}_{\boldsymbol{\theta}}}[E^{\mathrm{loc}}_{\boldsymbol{\theta}}(\mathbf{s})]\Big)\Big], && (54)\\
& && (55)
\end{aligned}$$

where in the last step we used that $E[\psi_{\boldsymbol{\theta}}] = \mathbb{E}_{\mathbf{s}\sim\mathcal{P}_{\boldsymbol{\theta}}}[E^{\mathrm{loc}}_{\boldsymbol{\theta}}(\mathbf{s})]$. Therefore,

$$\partial_{\theta_k} E[\psi_{\boldsymbol{\theta}}] = 2\,\mathrm{Re}\left\{\mathbb{E}_{\mathbf{s}\sim\mathcal{P}_{\boldsymbol{\theta}}}\Big[O^*_k(\mathbf{s})\Big(E^{\mathrm{loc}}_{\boldsymbol{\theta}}(\mathbf{s}) - \mathbb{E}_{\mathbf{s}\sim\mathcal{P}_{\boldsymbol{\theta}}}[E^{\mathrm{loc}}_{\boldsymbol{\theta}}(\mathbf{s})]\Big)\Big]\right\}. \tag{56}$$

B.3 Variational Gradient Decomposition

For the development of a practical algorithm we can decompose the VMC gradient into its probability and phase components as follows. Starting from the standard VMC/NQS gradient estimator we can write

$$\nabla_{\boldsymbol{\theta}} E[\psi_{\boldsymbol{\theta}}] = 2\ \mathrm{Re}\left\{\mathbb{E}_{\mathbf{s}\sim\mathcal{P}_{\boldsymbol{\theta}}}\left[\left(E^{\mathrm{loc}}_{\boldsymbol{\theta}}(\mathbf{s}) - \mathbb{E}_{\mathbf{s}\sim\mathcal{P}_{\boldsymbol{\theta}}}[E^{\mathrm{loc}}_{\boldsymbol{\theta}}(\mathbf{s})]\right)\mathbf{O}(\mathbf{s})^*\right]\right\}. \tag{57}$$

Recall that $*$ is the complex conjugate, and

$$E^{\mathrm{loc}}_{\boldsymbol{\theta}}(\mathbf{s}) = \mathrm{Re}\left\{E^{\mathrm{loc}}_{\boldsymbol{\theta}}(\mathbf{s})\right\} + i\cdot\mathrm{Im}\left\{E^{\mathrm{loc}}_{\boldsymbol{\theta}}(\mathbf{s})\right\}, \quad \mathbf{O}(\mathbf{s}) = \mathrm{Re}\left\{\mathbf{O}(\mathbf{s})\right\} + i\cdot\mathrm{Im}\left\{\mathbf{O}(\mathbf{s})\right\}. \tag{58}$$

For convenience, we denote $E_R := \mathrm{Re}\left\{E^{\mathrm{loc}}_{\boldsymbol{\theta}}(\mathbf{s})\right\}$ and $E_I := \mathrm{Im}\left\{E^{\mathrm{loc}}_{\boldsymbol{\theta}}(\mathbf{s})\right\}$ (analogously for $\mathbf{O}(\mathbf{s})$). We can now write

$$\begin{aligned}
\nabla_{\boldsymbol{\theta}} E[\psi_{\boldsymbol{\theta}}] &= 2\ \mathrm{Re}\left\{\mathbb{E}_{\mathbf{s}\sim\mathcal{P}_{\boldsymbol{\theta}}}\left[\left((E_R - \mathbb{E}_{\mathbf{s}\sim\mathcal{P}_{\boldsymbol{\theta}}}[E^{\mathrm{loc}}_{\boldsymbol{\theta}}(\mathbf{s})]) + iE_I\right)(O_R - iO_I)\right]\right\} && (59)\\
&= 2\ \mathrm{Re}\{\mathbb{E}_{\mathbf{s}\sim\mathcal{P}_{\boldsymbol{\theta}}}[(E_R - \mathbb{E}_{\mathbf{s}\sim\mathcal{P}_{\boldsymbol{\theta}}}[E^{\mathrm{loc}}_{\boldsymbol{\theta}}(\mathbf{s})])O_R && (60)\\
&\quad - i(E_R - \mathbb{E}_{\mathbf{s}\sim\mathcal{P}_{\boldsymbol{\theta}}}[E^{\mathrm{loc}}_{\boldsymbol{\theta}}(\mathbf{s})])O_I + iE_IO_R + E_IO_I]\} && (61)\\
&= 2\ \mathrm{Re}\left\{\mathbb{E}_{\mathbf{s}\sim\mathcal{P}_{\boldsymbol{\theta}}}\left[(E_R - \mathbb{E}_{\mathbf{s}\sim\mathcal{P}_{\boldsymbol{\theta}}}[E^{\mathrm{loc}}_{\boldsymbol{\theta}}(\mathbf{s})])O_R\right]\right\} + 2\ \mathrm{Re}\left\{\mathbb{E}_{\mathbf{s}\sim\mathcal{P}_{\boldsymbol{\theta}}}[E_IO_I]\right\}, && (62)
\end{aligned}$$

which shows exactly how to update our model, when we parameterize the real and imaginary parts of the amplitude separately. Moreover, we can subtract a constant baseline, c, from the second term (corresponding to the imaginary parts of our parameterization) to reduce the variance:

$$\begin{aligned}
2\ \mathrm{Re}\left\{\mathbb{E}_{\mathbf{s}\sim\mathcal{P}_{\boldsymbol{\theta}}}[E_I(O_I - c)]\right\} &= 2\ \mathrm{Re}\left\{\mathbb{E}_{\mathbf{s}\sim\mathcal{P}_{\boldsymbol{\theta}}}[E_IO_I]\right\} - 2\ \mathrm{Re}\left\{\mathbb{E}_{\mathbf{s}\sim\mathcal{P}_{\boldsymbol{\theta}}}[E_Ic]\right\} && (63)\\
&= 2\ \mathrm{Re}\left\{\mathbb{E}_{\mathbf{s}\sim\mathcal{P}_{\boldsymbol{\theta}}}[E_IO_I]\right\} - 2c\ \mathrm{Re}\left\{\mathbb{E}_{\mathbf{s}\sim\mathcal{P}_{\boldsymbol{\theta}}}[E_I]\right\} && (64)\\
&= 2\ \mathrm{Re}\left\{\mathbb{E}_{\mathbf{s}\sim\mathcal{P}_{\boldsymbol{\theta}}}[E_IO_I]\right\}, && (65)
\end{aligned}$$

where line 65 comes from the reality of expected values under Hermitian operators. So like before, we may subtract a baseline to reduce variance.

To find the variance-minimizing constant, consider the one-sample estimator for the imaginary contribution to the gradient

$$g_I(\mathbf{s}; c) := 2\,E_I(\mathbf{s})(O_I(\mathbf{s}) - c). \tag{66}$$

Since $\mathbb{E}_{\mathbf{s}\sim\mathcal{P}_{\boldsymbol{\theta}}}[E_I(\mathbf{s})] = 0$, subtracting any constant c leaves the estimator unbiased:

$$\mathbb{E}[g_I(\mathbf{s}; c)] = 2\,\mathbb{E}[E_I(\mathbf{s})\,O_I(\mathbf{s})]\,. \tag{67}$$

Therefore, the optimal constant is the one that minimizes $\mathbb{V}(g_I(\mathbf{s}; c))$, or equivalently its second moment:

$$\underset{c}{\mathrm{argmin}}\ \mathbb{E}\Big[E_I(\mathbf{s})^2(O_I(\mathbf{s}) - c)^2\Big]\,. \tag{68}$$

Differentiating with respect to c and setting the derivative to zero gives

$$\begin{aligned}
0 &= \frac{\partial}{\partial c}\mathbb{E}\left[E_I(\mathbf{s})^2(O_I(\mathbf{s}) - c)^2\right] && (69)\\
&= -2\,\mathbb{E}\left[E_I(\mathbf{s})^2(O_I(\mathbf{s}) - c)\right], && (70)
\end{aligned}$$

hence the variance-minimizing baseline is

$$c^\star = \frac{\mathbb{E}\left[E_I(\mathbf{s})^2\,O_I(\mathbf{s})\right]}{\mathbb{E}\left[E_I(\mathbf{s})^2\right]}. \tag{71}$$

If $O_I(\mathbf{s})$ is vector-valued, this formula is applied coordinate-wise. We tried using this minimizer in practice, but its implementation requires adding an extra forward and backward pass, making it significantly slower than the vanilla gradient expression.

B.4 Exact Complex-Amplitude Decomposition

We now derive an exact decomposition of the energy difference between two normalized complex wavefunctions. Let $|\psi_{\boldsymbol{\theta}}\rangle$ be the reference state and $|\psi_{\boldsymbol{\theta}'}\rangle$ be the candidate state. We choose the global phase of $|\psi_{\boldsymbol{\theta}'}\rangle$ so that

$$\langle\psi_{\boldsymbol{\theta}}|\psi_{\boldsymbol{\theta}'}\rangle = |\langle\psi_{\boldsymbol{\theta}}|\psi_{\boldsymbol{\theta}'}\rangle| = \sqrt{1 - \mathcal{I}(\psi_{\boldsymbol{\theta}}, \psi_{\boldsymbol{\theta}'})}. \tag{72}$$

Define

$$r_{\boldsymbol{\theta}';\boldsymbol{\theta}}(\mathbf{s}) := \frac{\mathcal{P}_{\boldsymbol{\theta}'}(\mathbf{s})}{\mathcal{P}_{\boldsymbol{\theta}}(\mathbf{s})}, \qquad z_{\boldsymbol{\theta}';\boldsymbol{\theta}}(\mathbf{s}) := \frac{\psi_{\boldsymbol{\theta}'}(\mathbf{s})}{\psi_{\boldsymbol{\theta}}(\mathbf{s})}. \tag{73}$$

Note that $r_{\boldsymbol{\theta}';\boldsymbol{\theta}}(\mathbf{s})$ can be exactly computed when using autoregressive models, but has to be approximated by its self-normalized version,

$$r_{\boldsymbol{\theta}';\boldsymbol{\theta}}(\mathbf{s}) \approx \frac{\left|\psi_{\boldsymbol{\theta}'}(\mathbf{s})/\psi_{\boldsymbol{\theta}}(\mathbf{s})\right|^2}{\mathrm{sg}\left(\frac{1}{M}\sum_{j=1}^{M}\left|\psi_{\boldsymbol{\theta}'}(\mathbf{s}_j)/\psi_{\boldsymbol{\theta}}(\mathbf{s}_j)\right|^2\right)}, \tag{74}$$

for unnormalized non-autoregressive wavefunctions, which is a biased but consistent estimator of the true importance sampling ratio.

Writing

$$z_{\boldsymbol{\theta}';\boldsymbol{\theta}}(\mathbf{s}) = \sqrt{r_{\boldsymbol{\theta}';\boldsymbol{\theta}}(\mathbf{s})}e^{i\alpha_{\boldsymbol{\theta}';\boldsymbol{\theta}}(\mathbf{s})},$$

we take the wrapped phase difference

$$\alpha_{\boldsymbol{\theta}';\boldsymbol{\theta}}(\mathbf{s}) = \mathrm{atan}_2\left(\sin\left(\arg\psi_{\boldsymbol{\theta}'}(\mathbf{s}) - \arg\psi_{\boldsymbol{\theta}}(\mathbf{s})\right), \cos\left(\arg\psi_{\boldsymbol{\theta}'}(\mathbf{s}) - \arg\psi_{\boldsymbol{\theta}}(\mathbf{s})\right)\right). \tag{75}$$

For later use, define the PWO phase increment

$$\phi_{\boldsymbol{\theta}';\boldsymbol{\theta}}(\mathbf{s}) := 2\alpha_{\boldsymbol{\theta}';\boldsymbol{\theta}}(\mathbf{s}). \tag{76}$$

The factor of two appears because the VMC phase gradient in Eq. (62) contains $2A^{\mathrm{I}}_{\boldsymbol{\theta}}\boldsymbol{\nabla}\arg\psi_{\boldsymbol{\theta}}$.

Theorem B.1 (Exact complex-amplitude decomposition)**.** *Let*

$$\hat{K}_{\boldsymbol{\theta}} := \hat{H} - E[\psi_{\boldsymbol{\theta}}]\mathbb{1}, \qquad |\eta\rangle := |\psi_{\boldsymbol{\theta}'}\rangle - |\psi_{\boldsymbol{\theta}}\rangle,$$

and let

$$\Delta E^{\mathrm{loc}}_{\boldsymbol{\theta}} := E^{\mathrm{loc}}_{\boldsymbol{\theta}} - E[\psi_{\boldsymbol{\theta}}] = A^{\mathrm{R}}_{\boldsymbol{\theta}} + iA^{\mathrm{I}}_{\boldsymbol{\theta}}.$$

Then

$$E[\psi_{\boldsymbol{\theta}'}] - E[\psi_{\boldsymbol{\theta}}] = 2\,\mathrm{Re}\,\mathbb{E}_{\mathbf{s}\sim\mathcal{P}_{\boldsymbol{\theta}}}\left[z_{\boldsymbol{\theta}';\boldsymbol{\theta}}(\mathbf{s})^*\Delta E^{\mathrm{loc}}_{\boldsymbol{\theta}}(\mathbf{s})\right] + \langle\eta|\hat{K}_{\boldsymbol{\theta}}|\eta\rangle. \tag{77}$$

Consequently,

$$E[\psi_{\boldsymbol{\theta}'}] - E[\psi_{\boldsymbol{\theta}}] \le 2\mathbb{E}_{\mathbf{s}\sim\mathcal{P}_{\boldsymbol{\theta}}}\left[\sqrt{r}\cos\alpha A^{\mathrm{R}}_{\boldsymbol{\theta}}\right] + 2\mathbb{E}_{\mathbf{s}\sim\mathcal{P}_{\boldsymbol{\theta}}}\left[\sqrt{r}\sin\alpha A^{\mathrm{I}}_{\boldsymbol{\theta}}\right] \tag{78}$$

$$+ 2\left\|\hat{H} - E[\psi_{\boldsymbol{\theta}}]\mathbb{1}\right\|_\infty\left(1 - \sqrt{1 - \mathcal{I}(\psi_{\boldsymbol{\theta}}, \psi_{\boldsymbol{\theta}'})}\right), \tag{79}$$

where, for readability, we have suppressed the $(\boldsymbol{\theta}';\boldsymbol{\theta})$ *subscripts on* r *and* α*.*

Proof. Since $\langle\psi_{\boldsymbol{\theta}}|\hat{K}_{\boldsymbol{\theta}}|\psi_{\boldsymbol{\theta}}\rangle = 0$, expanding around the reference state gives

$$E[\psi_{\boldsymbol{\theta}'}] - E[\psi_{\boldsymbol{\theta}}] = \langle\psi_{\boldsymbol{\theta}'}|\hat{K}_{\boldsymbol{\theta}}|\psi_{\boldsymbol{\theta}'}\rangle \tag{80}$$

$$= 2\,\mathrm{Re}\,\langle\eta|\hat{K}_{\boldsymbol{\theta}}|\psi_{\boldsymbol{\theta}}\rangle + \langle\eta|\hat{K}_{\boldsymbol{\theta}}|\eta\rangle. \tag{81}$$

Moreover, $\eta(\mathbf{s}) = (z_{\boldsymbol{\theta}';\boldsymbol{\theta}}(\mathbf{s}) - 1)\psi_{\boldsymbol{\theta}}(\mathbf{s})$, and

$$\langle\mathbf{s}|\hat{K}_{\boldsymbol{\theta}}|\psi_{\boldsymbol{\theta}}\rangle = \psi_{\boldsymbol{\theta}}(\mathbf{s})\Delta E^{\mathrm{loc}}_{\boldsymbol{\theta}}(\mathbf{s}).$$

Therefore,

$$\langle\eta|\hat{K}_{\boldsymbol{\theta}}|\psi_{\boldsymbol{\theta}}\rangle = \mathbb{E}_{\mathbf{s}\sim\mathcal{P}_{\boldsymbol{\theta}}}\left[\left(z_{\boldsymbol{\theta}';\boldsymbol{\theta}}(\mathbf{s})^* - 1\right)\Delta E^{\mathrm{loc}}_{\boldsymbol{\theta}}(\mathbf{s})\right] \tag{82}$$

$$= \mathbb{E}_{\mathbf{s}\sim\mathcal{P}_{\boldsymbol{\theta}}}\left[z_{\boldsymbol{\theta}';\boldsymbol{\theta}}(\mathbf{s})^*\Delta E^{\mathrm{loc}}_{\boldsymbol{\theta}}(\mathbf{s})\right], \tag{83}$$

because $\mathbb{E}_{\mathbf{s}\sim\mathcal{P}_{\boldsymbol{\theta}}}[\Delta E^{\mathrm{loc}}_{\boldsymbol{\theta}}(\mathbf{s})] = 0$. The residual satisfies

$$\langle\eta|\hat{K}_{\boldsymbol{\theta}}|\eta\rangle \le \left|\langle\eta|\hat{K}_{\boldsymbol{\theta}}|\eta\rangle\right| \le \left\|\hat{K}_{\boldsymbol{\theta}}\right\|_\infty\|\eta\|^2. \tag{84}$$

By the phase convention in Eq. (72),

$$\|\eta\|^2 = 2 - 2\,\mathrm{Re}\,\langle\psi_{\boldsymbol{\theta}}|\psi_{\boldsymbol{\theta}'}\rangle = 2\left(1 - \sqrt{1 - \mathcal{I}(\psi_{\boldsymbol{\theta}}, \psi_{\boldsymbol{\theta}'})}\right). \tag{85}$$

Finally,

$$\mathrm{Re}\left[z^*\Delta E^{\mathrm{loc}}_{\boldsymbol{\theta}}\right] = \sqrt{r}\left(\cos\alpha\, A^{\mathrm{R}}_{\boldsymbol{\theta}} + \sin\alpha\, A^{\mathrm{I}}_{\boldsymbol{\theta}}\right), \tag{86}$$

which proves the result. □

B.5 Amplitude and Phase Trust Regions Control Infidelity

The bound above, in the Conservative Policy Iteration (CPI) style, depends on the infidelity between the current and candidate wavefunctions. We now show that separate trust regions on the Born distribution and the phase imply such an infidelity trust region.

Lemma B.2 (Amplitude and phase trust regions imply an infidelity trust region)**.** *Let $0 \le \epsilon < 1$ and $0 \le \delta \le \pi$. Assume that we have no measure zero events and that for every configuration with $\mathcal{P}_{\boldsymbol{\theta}}(\mathbf{s}) > 0$,*

$$1 - \epsilon \le r_{\boldsymbol{\theta}';\boldsymbol{\theta}}(\mathbf{s}) \le 1 + \epsilon, \tag{87}$$

and that the phases satisfy, up to a global phase,

$$\left|\alpha_{\boldsymbol{\theta}';\boldsymbol{\theta}}(\mathbf{s})\right| \le \delta/2. \tag{88}$$

Then

$$\mathcal{I}(\psi_{\boldsymbol{\theta}}, \psi_{\boldsymbol{\theta}'}) \le 1 - \cos^2\left(\frac{\delta}{2}\right)\frac{1 + \sqrt{1-\epsilon^2}}{2}. \tag{89}$$

Proof. The overlap satisfies

$$\langle \psi_{\boldsymbol{\theta}'} | \psi_{\boldsymbol{\theta}} \rangle = \sum_{\mathbf{s}} \sqrt{\mathcal{P}_{\boldsymbol{\theta}}(\mathbf{s})\mathcal{P}_{\boldsymbol{\theta}'}(\mathbf{s})} e^{i\alpha_{\boldsymbol{\theta}';\boldsymbol{\theta}}(\mathbf{s})} = \mathbb{E}_{\mathbf{s}\sim\mathcal{P}_{\boldsymbol{\theta}}}\left[\sqrt{r_{\boldsymbol{\theta}';\boldsymbol{\theta}}(\mathbf{s})} e^{i\alpha_{\boldsymbol{\theta}';\boldsymbol{\theta}}(\mathbf{s})}\right]. \tag{90}$$

Since $|\alpha_{\boldsymbol{\theta}';\boldsymbol{\theta}}(\mathbf{s})| \le \delta/2 \le \pi/2$,

$$|\langle \psi_{\boldsymbol{\theta}'} | \psi_{\boldsymbol{\theta}} \rangle| \ge \mathrm{Re}\,\langle \psi_{\boldsymbol{\theta}'} | \psi_{\boldsymbol{\theta}} \rangle \tag{91}$$

$$= \mathbb{E}_{\mathbf{s}\sim\mathcal{P}_{\boldsymbol{\theta}}}\left[\sqrt{r_{\boldsymbol{\theta}';\boldsymbol{\theta}}(\mathbf{s})}\cos\alpha_{\boldsymbol{\theta}';\boldsymbol{\theta}}(\mathbf{s})\right] \tag{92}$$

$$\ge \cos\left(\frac{\delta}{2}\right)\mathbb{E}_{\mathbf{s}\sim\mathcal{P}_{\boldsymbol{\theta}}}\left[\sqrt{r_{\boldsymbol{\theta}';\boldsymbol{\theta}}(\mathbf{s})}\right]. \tag{93}$$

It remains to lower bound the amplitude factor. Set $a_0 = 1-\epsilon$ and $b_0 = 1+\epsilon$. Since $r_{\boldsymbol{\theta}';\boldsymbol{\theta}}(\mathbf{s}) \in [a_0, b_0]$ and $\sqrt{x}$ is concave, it lies above its chord on $[a_0, b_0]$:

$$\sqrt{r_{\boldsymbol{\theta}';\boldsymbol{\theta}}(\mathbf{s})} \ge \frac{b_0 - r_{\boldsymbol{\theta}';\boldsymbol{\theta}}(\mathbf{s})}{b_0 - a_0}\sqrt{a_0} + \frac{r_{\boldsymbol{\theta}';\boldsymbol{\theta}}(\mathbf{s}) - a_0}{b_0 - a_0}\sqrt{b_0}. \tag{94}$$

Taking expectations, using linearity of expectation and

$$\mathbb{E}_{\mathbf{s}\sim\mathcal{P}_{\boldsymbol{\theta}}}[r_{\boldsymbol{\theta}';\boldsymbol{\theta}}(\mathbf{s})] = \sum_{\mathbf{s}} \mathcal{P}_{\boldsymbol{\theta}'}(\mathbf{s}) = 1, \tag{95}$$

gives

$$\mathbb{E}_{\mathbf{s}\sim\mathcal{P}_{\boldsymbol{\theta}}}\left[\sqrt{r_{\boldsymbol{\theta}';\boldsymbol{\theta}}(\mathbf{s})}\right] \ge \frac{b_0 - 1}{b_0 - a_0}\sqrt{a_0} + \frac{1 - a_0}{b_0 - a_0}\sqrt{b_0} = \frac{\sqrt{1-\epsilon} + \sqrt{1+\epsilon}}{2}. \tag{96}$$

Combining Eqs. (93) and (96), then squaring, gives Eq. (89). The compact form follows from

$$\left(\frac{\sqrt{1-\epsilon} + \sqrt{1+\epsilon}}{2}\right)^2 = \frac{1 + \sqrt{1-\epsilon^2}}{2}.$$

The small-trust-region expansion follows from $\cos^2\delta = 1 - \delta^2 + O(\delta^4)$ and $(1 + \sqrt{1-\epsilon^2})/2 = 1 - \epsilon^2/4 + O(\epsilon^4)$. □

B.6 PWO Surrogate Upper Bound and First-Order Consistency

We now relate the exact complex-amplitude decomposition to the practical PWO surrogate. The exact decomposition contains the coefficients $2\sqrt{r}\cos\alpha$ and $2\sqrt{r}\sin\alpha$, whereas PWO uses the simpler first-order surrogates r and $r\phi = 2r\alpha$. The following lemma controls the phase approximation error.

Lemma B.3 (Quadratic error of the PWO phase coefficient)**.** *For every $t \ge 0$ and $\alpha \in [-\pi, \pi]$,*

$$\left|2t\sin\alpha - 2t^2\alpha\right| \le 4\pi^2\left|te^{i\alpha} - 1\right|^2. \tag{97}$$

Proof. Set

$$d(t,\alpha) := |te^{i\alpha} - 1|^2 = (t-1)^2 + 2t(1-\cos\alpha). \tag{98}$$

We consider three ranges of t. If $0 \le t \le 1/2$, then $d(t,\alpha) \ge (1-t)^2 \ge 1/4$, while

$$|2t\sin\alpha - 2t^2\alpha| \le 2t + 2\pi t^2 \le 1 + \frac{\pi}{2} \le 4\pi^2 d(t,\alpha). \tag{99}$$

If $t \ge 2$, then $d(t,\alpha) \ge (t-1)^2 \ge t^2/4$, and

$$|2t\sin\alpha - 2t^2\alpha| \le 2t + 2\pi t^2 \le 4\pi^2 d(t,\alpha). \tag{100}$$

It remains to consider $1/2 \le t \le 2$. We write

$$2t\sin\alpha - 2t^2\alpha = 2t(\sin\alpha - \alpha) + 2t\alpha(1-t), \tag{101}$$

and use the bounds, derived from the Taylor expansions, $|\sin\alpha - \alpha| \le \alpha^2/2$ and $1-\cos\alpha \ge 2\alpha^2/\pi^2$ on $[-\pi,\pi]$. Thus

$$d(t,\alpha) \ge (t-1)^2 + \frac{4t\alpha^2}{\pi^2}. \tag{102}$$

The term $t\alpha^2$ is bounded by $(\pi^2/4)d(t,\alpha)$. For the mixed term, let $x = |1-t|$ and $y = 2\sqrt{t}|\alpha|/\pi$. Then $d(t,\alpha) \ge x^2 + y^2$ and

$$2t|\alpha||1-t| = \pi\sqrt{t}\,xy \le \frac{\pi\sqrt{t}}{2}(x^2+y^2) \le \frac{\pi\sqrt{2}}{2}d(t,\alpha). \tag{103}$$

Combining these bounds gives

$$|2t\sin\alpha - 2t^2\alpha| \le \left(\frac{\pi^2}{4} + \frac{\pi\sqrt{2}}{2}\right) d(t,\alpha) \le 4\pi^2 d(t,\alpha). \tag{104}$$

□

Define the unclipped PWO surrogates

$$S_{\mathrm{mod}}(\boldsymbol{\theta}';\boldsymbol{\theta}) := \mathbb{E}_{\mathbf{s}\sim\mathcal{P}_{\boldsymbol{\theta}}}\left[r_{\boldsymbol{\theta}';\boldsymbol{\theta}}(\mathbf{s})A^{\mathrm{R}}_{\boldsymbol{\theta}}(\mathbf{s})\right], \tag{105}$$

$$S_{\arg}(\boldsymbol{\theta}';\boldsymbol{\theta}) := \mathbb{E}_{\mathbf{s}\sim\mathcal{P}_{\boldsymbol{\theta}}}\left[r_{\boldsymbol{\theta}';\boldsymbol{\theta}}(\mathbf{s})\phi_{\boldsymbol{\theta}';\boldsymbol{\theta}}(\mathbf{s})A^{\mathrm{I}}_{\boldsymbol{\theta}}(\mathbf{s})\right]. \tag{106}$$

Theorem B.4 (PWO modulus–phase surrogate upper bound)**.** *Assume $A^{\mathrm{R}}_{\boldsymbol{\theta}}$ and $A^{\mathrm{I}}_{\boldsymbol{\theta}}$ are bounded. Then*

$$E[\psi_{\boldsymbol{\theta}'}] - E[\psi_{\boldsymbol{\theta}}] \le S_{\mathrm{mod}}(\boldsymbol{\theta}';\boldsymbol{\theta}) + S_{\arg}(\boldsymbol{\theta}';\boldsymbol{\theta}) + 2C_{\boldsymbol{\theta}}\left(1 - \sqrt{1-\mathcal{I}(\psi_{\boldsymbol{\theta}},\psi_{\boldsymbol{\theta}'})}\right), \tag{107}$$

where one may take

$$C_{\boldsymbol{\theta}} = \left\|\hat{H} - E[\psi_{\boldsymbol{\theta}}]\mathbb{1}\right\|_\infty + \left\|A^{\mathrm{R}}_{\boldsymbol{\theta}}\right\|_\infty + 4\pi^2\left\|A^{\mathrm{I}}_{\boldsymbol{\theta}}\right\|_\infty. \tag{108}$$

In particular,

$$E[\psi_{\boldsymbol{\theta}'}] - E[\psi_{\boldsymbol{\theta}}] \le S_{\mathrm{mod}}(\boldsymbol{\theta}';\boldsymbol{\theta}) + S_{\arg}(\boldsymbol{\theta}';\boldsymbol{\theta}) + 2C_{\boldsymbol{\theta}}\mathcal{I}(\psi_{\boldsymbol{\theta}},\psi_{\boldsymbol{\theta}'}). \tag{109}$$

Proof. For readability, write $r = r_{\boldsymbol{\theta}';\boldsymbol{\theta}}$, $\alpha = \alpha_{\boldsymbol{\theta}';\boldsymbol{\theta}}$, $\phi = 2\alpha$, and $z = \sqrt{r}e^{i\alpha}$. Since

$$|z-1|^2 = 1 + r - 2\sqrt{r}\cos\alpha,$$

and $\mathbb{E}_{\mathbf{s}\sim\mathcal{P}_{\boldsymbol{\theta}}}[A^{\mathrm{R}}_{\boldsymbol{\theta}}] = 0$, we have

$$2\mathbb{E}_{\mathcal{P}_{\boldsymbol{\theta}}}[\sqrt{r}\cos\alpha\, A^{\mathrm{R}}_{\boldsymbol{\theta}}] = \mathbb{E}_{\mathcal{P}_{\boldsymbol{\theta}}}[rA^{\mathrm{R}}_{\boldsymbol{\theta}}] - \mathbb{E}_{\mathcal{P}_{\boldsymbol{\theta}}}[|z-1|^2 A^{\mathrm{R}}_{\boldsymbol{\theta}}] \tag{110}$$

$$\le S_{\mathrm{mod}} + \left\|A^{\mathrm{R}}_{\boldsymbol{\theta}}\right\|_\infty \mathbb{E}_{\mathcal{P}_{\boldsymbol{\theta}}}[|z-1|^2]. \tag{111}$$

For the phase term, apply Lemma B.3 with $t = \sqrt{r}$:

$$2\mathbb{E}_{\mathcal{P}_{\boldsymbol{\theta}}}[\sqrt{r}\sin\alpha\, A^{\mathrm{I}}_{\boldsymbol{\theta}}] = \mathbb{E}_{\mathcal{P}_{\boldsymbol{\theta}}}[r\phi A^{\mathrm{I}}_{\boldsymbol{\theta}}] + \mathbb{E}_{\mathcal{P}_{\boldsymbol{\theta}}}[(2\sqrt{r}\sin\alpha - 2r\alpha)A^{\mathrm{I}}_{\boldsymbol{\theta}}] \tag{112}$$

$$\le S_{\arg} + 4\pi^2\left\|A^{\mathrm{I}}_{\boldsymbol{\theta}}\right\|_\infty \mathbb{E}_{\mathcal{P}_{\boldsymbol{\theta}}}[|z-1|^2]. \tag{113}$$

Finally,

$$\mathbb{E}_{\mathcal{P}_{\boldsymbol{\theta}}}[|z-1|^2] = \|\psi_{\boldsymbol{\theta}'} - \psi_{\boldsymbol{\theta}}\|^2 \tag{114}$$

$$= 2\left(1 - \sqrt{1 - \mathcal{I}(\psi_{\boldsymbol{\theta}}, \psi_{\boldsymbol{\theta}'})}\right), \tag{115}$$

where we used the global-phase convention Eq. (72). Combining Eqs. (79), (111), and (113) proves Eq. (107). The linear-infidelity form follows from $1 - \sqrt{1-x} \leq x$. □

Theorem B.5 (First-order consistency of the PWO surrogate)**.** *Assume that $\psi_{\boldsymbol{\theta}}$ is differentiable in a neighborhood of the reference parameters. For the phase surrogate, detach the importance ratio from the gradient computation:*

$$S_{\arg}(\boldsymbol{\theta}'; \boldsymbol{\theta}) = \mathbb{E}_{\mathbf{s}\sim\mathcal{P}_{\boldsymbol{\theta}}}\left[\mathrm{sg}\left(r_{\boldsymbol{\theta}';\boldsymbol{\theta}}(\mathbf{s})\right)\phi_{\boldsymbol{\theta}';\boldsymbol{\theta}}(\mathbf{s})A^{\mathrm{I}}_{\boldsymbol{\theta}}(\mathbf{s})\right].$$

Then

$$\boldsymbol{\nabla}_{\boldsymbol{\theta}'}\left[S_{\mathrm{mod}}(\boldsymbol{\theta}'; \boldsymbol{\theta}) + S_{\arg}(\boldsymbol{\theta}'; \boldsymbol{\theta})\right]\big|_{\boldsymbol{\theta}'=\boldsymbol{\theta}} = \boldsymbol{\nabla}_{\boldsymbol{\theta}} E[\psi_{\boldsymbol{\theta}}]. \tag{116}$$

The same derivative is obtained from the clipped objectives at the reference point, since $r = 1$ and $\phi = 0$ lie in the interior of the clipping intervals.

Proof. At $\boldsymbol{\theta}' = \boldsymbol{\theta}$,

$$\boldsymbol{\nabla}_{\boldsymbol{\theta}'} r_{\boldsymbol{\theta}';\boldsymbol{\theta}}(\mathbf{s})\big|_{\boldsymbol{\theta}'=\boldsymbol{\theta}} = \boldsymbol{\nabla}_{\boldsymbol{\theta}} \log \mathcal{P}_{\boldsymbol{\theta}}(\mathbf{s}) = 2\boldsymbol{\nabla}_{\boldsymbol{\theta}} \log|\psi_{\boldsymbol{\theta}}(\mathbf{s})|, \tag{117}$$

$$\boldsymbol{\nabla}_{\boldsymbol{\theta}'} \phi_{\boldsymbol{\theta}';\boldsymbol{\theta}}(\mathbf{s})\big|_{\boldsymbol{\theta}'=\boldsymbol{\theta}} = 2\boldsymbol{\nabla}_{\boldsymbol{\theta}} \arg \psi_{\boldsymbol{\theta}}(\mathbf{s}). \tag{118}$$

Therefore the gradient of S_{mod} recovers the $2A^{\mathrm{R}}_{\boldsymbol{\theta}}\boldsymbol{\nabla}\log|\psi_{\boldsymbol{\theta}}|$ term in Eq. (62), while the detached phase surrogate recovers the $2A^{\mathrm{I}}_{\boldsymbol{\theta}}\boldsymbol{\nabla}\arg\psi_{\boldsymbol{\theta}}$ term. Thus the sum of the two surrogate gradients equals the VMC gradient. □

Theorem 4.1 (First-order consistency of PWO)**.** *Let $|\psi_{\boldsymbol{\theta}}\rangle$ be differentiable at $\boldsymbol{\theta}_{\mathrm{old}}$, and assume common support. For $\epsilon, \delta > 0$,*

$$\boldsymbol{\nabla}_{\boldsymbol{\theta}}\left(L^{\mathrm{clip}}_{\mathrm{mod}}(\boldsymbol{\theta}) + L^{\mathrm{clip}}_{\arg}(\boldsymbol{\theta})\right)\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_{\mathrm{old}}} = \boldsymbol{\nabla}_{\boldsymbol{\theta}} E[\psi_{\boldsymbol{\theta}}]|_{\boldsymbol{\theta}=\boldsymbol{\theta}_{\mathrm{old}}}. \tag{19}$$

For a proof see Appendix B.6.

Proof. Since clipping is inactive to first order at $r = 1$ and $\phi = 0$, the clipped objectives have the same derivative at the reference point and the proof of Theorem B.5 still applies. □

B.7 Clipped PWO Improvement Certificate

We finally combine the PWO surrogate bound with the amplitude–phase infidelity control. This gives the deterministic population-level certificate corresponding to the clipped PWO objective.

Define the clipped losses

$$L^{\mathrm{clip}}_{\mathrm{mod}}(\boldsymbol{\theta}'; \boldsymbol{\theta}) := \mathbb{E}_{\mathcal{P}_{\boldsymbol{\theta}}}\left[\max\left(rA^{\mathrm{R}}_{\boldsymbol{\theta}}, \mathrm{clip}(r, 1-\epsilon, 1+\epsilon)A^{\mathrm{R}}_{\boldsymbol{\theta}}\right)\right], \tag{119}$$

$$L^{\mathrm{clip}}_{\arg}(\boldsymbol{\theta}'; \boldsymbol{\theta}) := \mathbb{E}_{\mathcal{P}_{\boldsymbol{\theta}}}\left[\mathrm{sg}(r)\max\left(\phi A^{\mathrm{I}}_{\boldsymbol{\theta}}, \mathrm{clip}(\phi, -\delta, \delta)A^{\mathrm{I}}_{\boldsymbol{\theta}}\right)\right], \tag{120}$$

where, as above, $r = r_{\boldsymbol{\theta}';\boldsymbol{\theta}}$ and $\phi = \phi_{\boldsymbol{\theta}';\boldsymbol{\theta}} = 2\alpha$.

Corollary 4.3 (Clipped PWO improvement certificate)**.** *Assume the conditions of Theorem 4.2, bounded centered local energies, and global constraints $r_{\boldsymbol{\theta}}(\mathbf{s}) \in [1-\epsilon, 1+\epsilon]$ and $|\phi_{\boldsymbol{\theta}}(\mathbf{s})| \leq \delta$ for all $\mathbf{s}$, with $0 \leq \epsilon \leq 1$ and $0 \leq \delta \leq \pi$. Then there exists $C_{\boldsymbol{\theta}_{\mathrm{old}}} < \infty$ such that*

$$E[\psi_{\boldsymbol{\theta}}] - E[\psi_{\boldsymbol{\theta}_{\mathrm{old}}}] \leq L^{\mathrm{clip}}_{\mathrm{mod}}(\boldsymbol{\theta}) + L^{\mathrm{clip}}_{\arg}(\boldsymbol{\theta}) + 2C_{\boldsymbol{\theta}_{\mathrm{old}}}\left(1 - \frac{1+\sqrt{1-\epsilon^2}}{2}\cos^2\left(\frac{\delta}{2}\right)\right). \tag{22}$$

For a proof see Appendix B.7.

Proof of Corollary 4.3. The clipped energy bound first follows from the linear-infidelity bound (109) and the pointwise inequalities

$$L^{\mathrm{clip}}_{\mathrm{mod}}(\boldsymbol{\theta}) \geq S_{\mathrm{mod}}(\boldsymbol{\theta}; \boldsymbol{\theta}_{\mathrm{old}}), \qquad L^{\mathrm{clip}}_{\mathrm{arg}}(\boldsymbol{\theta}) \geq S_{\mathrm{arg}}(\boldsymbol{\theta}; \boldsymbol{\theta}_{\mathrm{old}}), \tag{121}$$

which hold because each clipped objective takes the maximum of the unclipped term and its clipped counterpart, and $\mathrm{sg}(r)$ has the same forward value as r. Thus

$$E[\psi_{\boldsymbol{\theta}}] - E[\psi_{\boldsymbol{\theta}_{\mathrm{old}}}] \leq L^{\mathrm{clip}}_{\mathrm{mod}}(\boldsymbol{\theta}) + L^{\mathrm{clip}}_{\mathrm{arg}}(\boldsymbol{\theta}) + 2C_{\boldsymbol{\theta}_{\mathrm{old}}}\mathcal{I}(\psi_{\boldsymbol{\theta}_{\mathrm{old}}}, \psi_{\boldsymbol{\theta}}). \tag{122}$$

By Lemma B.2, the assumptions $r_{\boldsymbol{\theta}} \in [1-\epsilon, 1+\epsilon]$ and $|\phi_{\boldsymbol{\theta}}| \leq \delta$ imply

$$\mathcal{I}(\psi_{\boldsymbol{\theta}_{\mathrm{old}}}, \psi_{\boldsymbol{\theta}}) \leq 1 - \frac{1+\sqrt{1-\epsilon^2}}{2} \cos^2\left(\frac{\delta}{2}\right), \tag{123}$$

because the wrapped phase itself is $\phi_{\boldsymbol{\theta}}/2$. Substituting this estimate into the previous display gives Eq. (22). □

PWO only encourages, but does not guarantee, the global trust-region conditions required by the certificate. First, the initial step is always taken from the unclipped objective, since clipping is inactive at $(\boldsymbol{\theta} = \boldsymbol{\theta}_{\mathrm{old}})$. Second, even after clipping activates on a sampled configuration, an update driven by other unclipped configurations can still change its amplitude or phase because all configurations share parameters. Thus, clipping controls the empirical surrogate on the sampled batch, but does not impose pointwise bounds globally. The certificate should therefore be read as a conditional guarantee for realized updates that satisfy the amplitude and phase trust regions over the full support.

C Additional Sections

C.1 Computational Cost of Stochastic Reconfiguration and Improvements

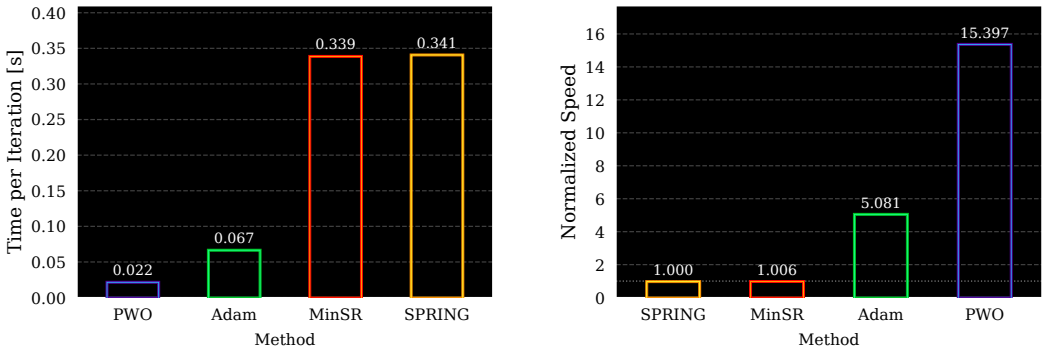
A significant bottleneck arises when using the original SR formulation for optimization. The vanilla implementation involves inverting a $P \times P$ matrix, where P denotes the number of parameters. Alternatively, Chen and Heyl [2024] showed that this inversion problem can be recast into one that inverts a matrix of size $M \times M$, where M is the number of samples used per iteration in the Monte Carlo estimation – this goes by the name of minSR and builds upon ideas from the machine learning optimization literature [Ren and Goldfarb, 2019]. The required operations for minSR are $O(M^2P) + O(M^3)$ instead of $O(P^3)$, and the memory usage is only $O(MP)$ instead of $O(P^2)$. It was used to train a Vision Transformer with around 3×10^5 parameters, giving the lowest energy so far for the difficult benchmark of the Majumdar–Ghosh point of the 10×10 J1-J2 Heisenberg system [Rende et al., 2024].

Despite the improvements brought by minSR, several implementations of NQS optimization, especially for electronic structure problems, employ KFAC [Martens and Grosse, 2015] instead of SR. The $O(M^3)$ cost can still be prohibitive when training networks for problems that require a large number of samples; KFAC provides a practical compromise: it avoids explicit matrix inversion through block-diagonal approximation, offering scalability at the potential cost of missing information about the geometry of the Hilbert space. KFAC is commonly used for applications to fermionic systems [Choo et al., 2020], where NQS achieved remarkable results, including architectures like FermiNet [Pfau et al., 2020], SchNet [Schütt et al., 2018], PauliNet [Hermann et al., 2020], and more recent transformer-based approaches [von Glehn et al., 2023].

Recent advances have aimed to improve upon both approaches. These include SPRING, which combines minSR and KFAC to avoid explicit inversion [Goldshlager et al., 2024], methods that exploit block-diagonal structure in the quantum geometric tensor to improve conditioning and scalability [Nys et al., 2024, Shokry et al., 2025], and variational Lanczos techniques that accelerate convergence by extracting information across multiple eigenstates [Wang et al., 2026]. For a more detailed analysis of improvements in second-order optimization of NQS, the reader can refer to Drissi et al. [2024].

Complementary strategies have been developed specifically for autoregressive NQS, particularly recurrent neural networks (RNNs). Variational neural annealing and gradual ramping of Hamiltonian parameters help convergence [Hibat-Allah et al., 2021, Ibarra-García-Padilla et al., 2025]. Second-order methods have also been adapted to recurrent architectures, including KFAC [Martens et al., 2018], while recent work shows that regularization can stabilize minSR for RNN wavefunctions [Attar et al., 2026].

C.2 Relative Speed of PWO per Iteration



(a) Wall-clock time per optimizer iteration.

(b) Normalized iteration speed relative to SPRING.

Figure 7: Per-iteration computational cost and normalized iteration speed for PWO and the baseline optimizers on the Heisenberg J_1–J_2 chain with the default architecture D.2. Both measurements are obtained on a single NVIDIA L40S GPU. Normalized speed is reported as the number of optimizer iterations per unit time relative to SPRING, so higher is faster.

Beyond convergence in wall-clock time, it is useful to isolate the computational cost of each optimization update. Figure 7b compares the normalized iteration speed (optimization steps per unit of time) of PWO against the baselines. The speed advantage over minSR is expected: PWO is a first-order method and does not require constructing or inverting the stochastic-reconfiguration matrix. In contrast, minSR replaces the original $P \times P$ SR solve by an $M \times M$ solve, where M is the number of samples, but it still requires expensive Jacobian contractions and a matrix inversion or linear solve. PWO avoids this curvature computation entirely.

The comparison with Adam is more subtle. A single PWO update has additional bookkeeping relative to Adam, including probability ratios, clipping terms, and wrapped phase increments. However, in VMC the dominant cost is often not the optimizer algebra itself, but sampling configurations and evaluating the corresponding local energies. PWO computes local energies and advantages once for a sampled batch and then reuses that batch across multiple proximal inner epochs. Adam, in contrast, performs a single update per sampled batch, so the local-energy overhead is paid again for each parameter update. Thus, PWO can amortize the expensive energy computation across several optimization steps, making its effective per-update cost competitive with Adam despite the more structured surrogate objective.

C.3 Why RWKV?

RWKV is a particularly natural architecture for large-scale autoregressive NQS because it combines the expressivity of sequence models with recurrent inference. Unlike Transformer-based autoregressive models, RWKV does not require a growing key-value cache, so its per-token inference cost and memory footprint remain constant in the sequence length [Peng et al., 2025]. This is well aligned with autoregressive VMC, where spin configurations are generated sequentially and model evaluations are repeatedly invoked during Monte Carlo optimization. Recurrent architectures have also been successful in autoregressive NQS more broadly, where they provide exact independent sampling from the Born distribution while avoiding Markov-chain mixing issues [Hibat-Allah et al., 2020, Merali et al., 2026]. Finally, recent hyperscale RWKV fine-tuning work provides a practical implementation foundation for adapting billion-parameter recurrent models to nonstandard optimization objectives [Sarkar et al., 2026]. For these reasons, RWKV is a useful stress test for whether PWO can scale beyond conventional NQS architectures.

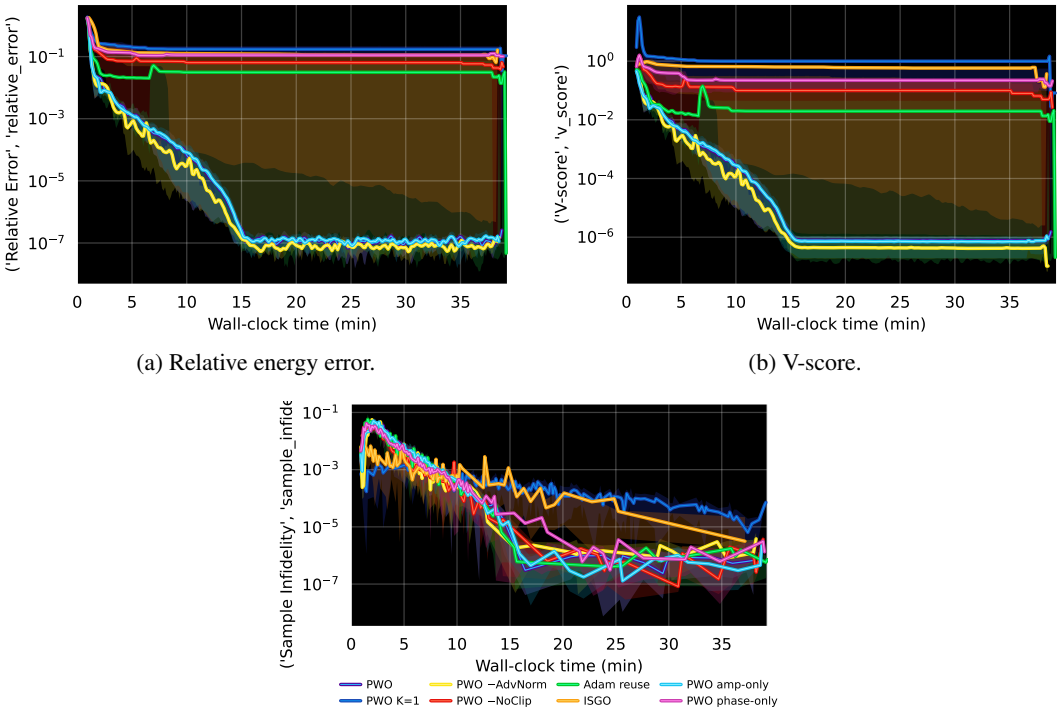
C.4 Limitations of PWO

Although PWO is supported by a first-order consistency result and a clipped improvement bound, the theory remains local and conservative: it relies on common-support and boundedness assumptions and does not imply global convergence for neural-network training. Empirically, our results include one-dimensional chains, a frustrated two-dimensional square lattice, and a large-scale RWKV fine-tuning experiment, but they are still not a comprehensive NQS benchmark across Hamiltonian families. In particular, broader validation on larger two-dimensional systems, fermionic models, and electronic-structure problems remains open. Finally, PWO introduces additional optimization hyperparameters, such as amplitude and phase clipping thresholds and the number of inner epochs, whose optimal values may depend on the Hamiltonian, ansatz, system size, and sample budget.

C.5 Ablation Studies

To isolate the contributions of the different components of PWO, we perform ablation studies on the frustrated Heisenberg J_1–J_2 chain with $N = 12$, periodic boundary conditions, $J_1 = 1$, and $J_2 = 0.5$. We use the same neural quantum state architecture, sampling budget, learning-rate schedule, and remaining optimization hyperparameters as in the main experiment (Appendix D.5). Unless otherwise specified, PWO uses $K = 4$ inner optimization steps per sampled batch, amplitude clipping with $\epsilon = 10^{-3}$, phase clipping with $\delta = 0.3$, and advantage normalization.

We compare the full PWO algorithm against several variants designed to isolate sample reuse and the different components of the proximal objective. First, *PWO* $K = 1$ performs only a single update per sampled batch, thereby removing sample reuse. Since the current and reference wavefunctions coincide at the beginning of this update, clipping is initially inactive and this variant reduces to an on-policy first-order VMC update (up to advantage normalization). Second, *PWO without clipping* retains $K = 4$ reused-batch updates but removes both the amplitude and phase trust-region constraints. We additionally remove advantage normalization, or retain only one of the two clipping mechanisms, yielding *amplitude-only* and *phase-only* PWO respectively. In the latter two ablations,



(a) Relative energy error.

(b) V-score.

(c) Sample infidelity.

Figure 8: Ablation study of PWO on the frustrated Heisenberg J_1–J_2 chain ($N = 12$, $J_1 = 1$, $J_2 = 0.5$, periodic boundary conditions). We compare full PWO against variants removing sample reuse ($K = 1$), clipping, or advantage normalization; variants retaining only amplitude or phase clipping; Adam with matched reused-batch updates; ISGO; and standard Adam. Curves report the interquartile mean over 10 random seeds and shaded regions denote the interquartile range.

both amplitude and phase gradient terms remain present; only the corresponding clipping operation is removed.

To distinguish the effect of sample reuse from that of the PWO objective, we also evaluate *Adam reuse*, which performs the same $K = 4$ optimization steps on each frozen batch using Adam but without the PWO proximal objective. We further include ISGO [Yang et al., 2020], which also reuses samples through importance weighting, as well as the standard Adam baseline from the main experiment.

Figure 8 reveals a clear separation between the components responsible for PWO's improvement. Full PWO, PWO without advantage normalization, and amplitude-only PWO behave nearly identically, reaching a relative error of approximately 10^{-7} within the wall-clock budget while simultaneously reducing the V-score. In contrast, PWO with $K = 1$, PWO without clipping, phase-only PWO, Adam with matched sample reuse, and ISGO remain in the same substantially worse performance regime as standard Adam. The $K = 1$ ablation shows that the improvement is not explained simply by the choice of first-order gradient estimator: removing the repeated updates on each batch eliminates most of the advantage of PWO. Sample reuse by itself, however, is also insufficient. Performing the same number of reused-batch updates with Adam does not reproduce the improvement, and ISGO similarly remains far above the accuracy reached by full PWO. Thus, repeatedly optimizing a frozen batch without controlling the resulting distribution shift does not provide the same benefit.

The clipping ablations identify trust-region control as the second essential ingredient. Removing clipping while retaining sample reuse eliminates the gain, whereas retaining amplitude clipping alone recovers essentially the full performance of PWO. Conversely, phase-only clipping performs substantially worse. On this benchmark, the dominant source of instability during sample reuse therefore appears to be drift in the Born distribution rather than the phase update. This should not

be interpreted as showing that phase clipping is unnecessary in general, but rather that amplitude clipping accounts for most of the empirical improvement on this particular Hamiltonian.

Finally, removing advantage normalization has negligible effect, indicating that it is not responsible for the observed performance difference. Taken together, these ablations support the interpretation of PWO as a *trust-region sample-reuse* method: sample reuse amortizes each expensive local-energy computation across several optimization steps, while clipping keeps those increasingly off-policy updates close to the reference wavefunction for the reused surrogate to remain useful.

D Experimental Details

D.1 Hyperparameter search

For each method, we perform a small hyperparameter search over the learning rate using a grid sweep and select the configuration with the best validation performance. For PWO, we additionally tune the clipping parameters that control the amplitude and phase surrogate objectives. Specifically, after selecting an initial learning rate, we independently grid search each clipping parameter while holding the best values found so far fixed. This sequential procedure provides a simple and computationally tractable way to tune the additional PWO-specific hyperparameters without requiring an exhaustive joint sweep over all combinations.

For Adam-based methods, including PWO, we use a cosine one-cycle learning-rate schedule [Smith and Topin, 2018]. The learning rate is first increased to a peak value during a short warmup phase and is then annealed smoothly to a small final value by cosine decay. The peak learning rate is selected by the grid search above, while the other schedule parameters are fixed across runs. For PWO, which performs multiple optimization epochs per sampled batch, we scale the transition horizon by the number of PPO epochs so that the schedule evolves on a comparable outer-iteration timescale. For minSR and SPRING, a constant learning rate performed better than scheduled variants.

Because the Heisenberg and J1–J2 chains are more challenging optimization problems than the transverse-field Ising model, we use longer learning-rate schedules for these settings (compare Appendices D.3-D.5). In particular, we allocate more transition steps to the Heisenberg chain than to Ising, and more transition steps to J1–J2 than to Heisenberg. This gives the optimizer a longer annealing horizon on the harder Hamiltonians, while keeping the schedule structure fixed across problems. To improve readability, we report the method-specific hyper-parameters for each Hamiltonian separately. The Ising configuration is shown first, followed by the Heisenberg and J1–J2 settings in the subsequent subsections.

D.2 Neural Quantum State Architecture

Across experiments, we use the same autoregressive recurrent neural network for PWO and minSR, ensuring that performance differences arise from the optimizer rather than the parametrization. Given a spin configuration $\mathbf{s} \in \{\pm 1\}^N$, we map spins to binary tokens, prepend a beginning-of-sequence token, and embed the resulting sequence with learned token embeddings of dimension 32 and learned positional embeddings. The embedded inputs are projected to dimension 256, followed by a $\tanh$ nonlinearity and layer normalization. The backbone consists of GRU layers with residual connections between recurrent blocks. From the shared recurrent representation, we use two separate two-layer MLP heads and GELU activations: an amplitude head, which outputs conditional log-probabilities via a log-softmax, and a phase head, which outputs bounded phases through a $\tanh$ nonlinearity scaled by π. The model has 1.4M parameters. Unless otherwise stated, all methods and Hamiltonians use the same architecture. The shared hyperparameters are summarized in Table 2.

Architecture hyperparameter	**Value**
Token embedding dimension	32
Site embeddings	Learned
Backbone input dimension	256
Backbone nonlinearity	$\tanh$
Backbone normalization	Layer normalization
Recurrent backbone	GRU
Number of GRU layers	3
GRU hidden dimension	256
Amplitude head	2-layer MLP
Phase head	2-layer MLP
Head hidden dimension	256
Head activation	GELU
Phase output	$\pi \tanh(\cdot)$
Phase scale	π

Table 2: Shared NQS architecture used across all Hamiltonians and optimization methods. The same autoregressive GRU-based parameterization is used for Adam, PWO, minSR, and SPRING.

D.3 Ising Model

Hyperparameter	Adam	PWO	minSR	SPRING
Optimizer / method	Adam	Adam + PWO	minSR	SPRING
Learning rate	10^{-5}	10^{-5}	10^{-2}	10^{-2}
Peak learning rate	10^{-4}	10^{-4}	–	–
Transition steps	5,000	20,000	–	–
PPO epochs	–	4	–	–
PPO clip ϵ	–	10^{-3}	–	–
Advantage normalization	–	Yes	–	–
Phase loss	–	$\Delta\phi$ clip	–	–
Phase coefficient	–	1.0	–	–
Phase clip	–	0.3	–	–
Center imaginary advantage	–	Yes	–	–
Normalize imaginary advantage	–	Yes	–	–
Phase Jacobian baseline	–	Yes	–	–
SR diagonal shift	–	–	10^{-2}	10^{-2}
NTK / minSR mode	–	–	Yes	Yes
On-the-fly SR	–	–	Yes	Yes
SPRING momentum	–	–	–	0.8

Table 3: Hyperparameters used for the Ising experiments. All methods use $N = 12$, periodic boundary conditions, $J = 1$, $h = 1$, complex-valued neural quantum states, 1024 training samples, exact diagonalization for evaluation, and evaluation every 200 iterations. Adam and PWO use the cosine one-cycle schedule.

D.4 Heisenberg Chain

Hyperparameter	Adam	PWO	minSR	SPRING
Optimizer / method	Adam	Adam + PWO	minSR	SPRING
LR parameter / constant LR	10^{-5}	10^{-5}	10^{-3}	10^{-3}
Peak learning rate	3×10^{-4}	10^{-4}	–	–
Transition steps	10,000	40,000	–	–
PPO epochs	–	4	–	–
PPO clip ϵ	–	10^{-3}	–	–
Advantage normalization	–	Yes	–	–
Phase loss	–	$\Delta\phi$ clip	–	–
Phase coefficient	–	1.0	–	–
Phase clip	–	0.3	–	–
Center imaginary advantage	–	Yes	–	–
Normalize imaginary advantage	–	Yes	–	–
Phase Jacobian baseline	–	Yes	–	–
SR diagonal shift	–	–	10^{-2}	10^{-2}
NTK / minSR mode	–	–	Yes	Yes
On-the-fly SR	–	–	Yes	Yes
SPRING momentum	–	–	–	0.8

Table 4: Hyperparameters for the Heisenberg-chain experiments. All methods use $N = 12$, periodic boundary conditions, coupling $J = 0.25$, no sign rule, complex-valued neural quantum states, 1024 training samples, exact evaluation, and evaluation every 200 iterations. Adam and PWO use a cosine one-cycle learning-rate schedule; minSR and SPRING use a constant learning rate.

D.5 Heisenberg J_1–J_2 Chain

Hyperparameter	Adam	PWO	minSR	SPRING
Optimizer / method	Adam	Adam + PWO	minSR	SPRING
LR parameter / constant LR	10^{-5}	10^{-5}	10^{-3}	10^{-3}
Peak learning rate	3×10^{-4}	10^{-4}	–	–
Transition steps	10,000	40,000	–	–
PPO epochs	–	4	–	–
PPO clip ϵ	–	10^{-3}	–	–
Advantage normalization	–	Yes	–	–
Phase loss	–	$\Delta\phi$ clip	–	–
Phase coefficient	–	1.0	–	–
Phase clip	–	0.3	–	–
Center imaginary advantage	–	Yes	–	–
Normalize imaginary advantage	–	Yes	–	–
Phase Jacobian baseline	–	Yes	–	–
SR diagonal shift	–	–	10^{-2}	10^{-2}
NTK / minSR mode	–	–	Yes	Yes
On-the-fly SR	–	–	Yes	Yes
SPRING momentum	–	–	–	0.8

Table 5: Hyperparameters for the frustrated Heisenberg J_1–J_2 experiments. All methods use $N = 12$, periodic boundary conditions, couplings $J_1 = 1$ and $J_2 = 0.5$, no sign rule, complex-valued neural quantum states, 1024 training samples, exact evaluation, and evaluation every 200 iterations. Adam and PWO use a cosine one-cycle learning-rate schedule; minSR and SPRING use a constant learning rate.

D.6 Two-dimensional J_1–J_2 square-lattice experiment

Using the same method hyper-parameters as above, we evaluate PWO on the frustrated spin-$1/2$ J_1–J_2 Heisenberg model on a two-dimensional square lattice,

$$\hat{H}^{\mathrm{2D}}_{J_1\text{–}J_2} = J_1 \sum_{\langle i,j \rangle} \hat{\mathbf{S}}_i \cdot \hat{\mathbf{S}}_j + J_2 \sum_{\langle\langle i,j \rangle\rangle} \hat{\mathbf{S}}_i \cdot \hat{\mathbf{S}}_j, \tag{124}$$

where $\langle i,j \rangle$ and $\langle\langle i,j \rangle\rangle$ denote nearest- and next-nearest-neighbor pairs, respectively. We use an $L \times L$ square lattice with $L = 10$, periodic boundary conditions, $J_1 = 1$, and $J_2 = 0.5$. All runs are restricted to the zero-magnetization sector, $S^z_{\mathrm{tot}} = 0$, by masking infeasible autoregressive choices during sampling and evaluation.

For this experiment we use a complex-valued patch-autoregressive transformer designed for two-dimensional lattices. Instead of generating spins one at a time, the model partitions the 10×10 lattice into non-overlapping 2×2 patches. Each patch is represented as a categorical token with vocabulary size $2^4 = 16$, so the full configuration is generated as a sequence of

$$T = \left(\frac{L}{2}\right)^2 = 25 \tag{125}$$

autoregressive tokens. The wavefunction is factorized over patch tokens as

$$\log \psi_\theta(\mathbf{s}) = \sum_{t=1}^{T} \left[\frac{1}{2} \log p_\theta(a_t \mid a_{<t}) + i\, \phi_\theta(a_t \mid a_{<t}) \right], \tag{126}$$

where the factor $1/2$ corresponds to the Born-rule convention $P_\theta(\mathbf{s}) = |\psi_\theta(\mathbf{s})|^2$.

The model prepends a beginning-of-sequence token and embeds patch tokens with learned token embeddings of dimension 64. We also use learned site embeddings and prefix-count features that encode the partial magnetization constraint. These embeddings are projected to width 96, normalized with layer normalization, and passed through an 8-layer causal transformer backbone. Each transformer block uses 6 attention heads, two-dimensional axial RoPE with base 100, residual connections, and a feedforward width of $4 \times 96 = 384$.

From the final transformer representation, the model uses separate amplitude and phase heads. Both heads are two-layer MLPs with hidden dimension 192 and GELU activations. The amplitude head outputs masked conditional log-probabilities through a log-softmax, ensuring exact normalization over feasible patch choices. The phase head outputs centered phase increments using a $\pi \tanh(\cdot)$ parameterization, with small initialization scale 10^{-3}. This gives an exactly sampleable complex autoregressive neural quantum state adapted to the two-dimensional square lattice.

While the aim of this experiment is to compare optimizers and not to achieve the state-of-the-art variational energy (around -199.0536 [Rende et al., 2024]), we note that the energy of the PWO-trained autoregressive transformer of 1.5M parameters is -185 after 30 mins, reaches -195.6 after 24 hours, and keeps decreasing. The run is on a single GPU and imposes no symmetries, except for zero-magnetization sampling.

Hyperparameter	**Value**
Lattice size	10×10
Number of spins	100
Boundary conditions	Periodic
Hamiltonian couplings	$J_1 = 1,\ J_2 = 0.5$
Magnetization sector	$S^z_{\text{tot}} = 0$
Patch size	2×2
Patch vocabulary size	16
Autoregressive tokens	25
Token embedding dimension	64
Transformer width	96
Transformer depth	8
Attention heads	6
Transformer MLP hidden dimension	384
RoPE type	2D axial RoPE
RoPE base	100
Amplitude head	2-layer MLP
Phase head	2-layer MLP
Head hidden dimension	192
Phase parameterization	$\pi \tanh(\cdot)$
Phase initialization std.	10^{-3}
Prefix-count features	Yes
Learned site embeddings	Yes

Table 6: Architecture used for the two-dimensional frustrated J_1–J_2 square-lattice experiment. The model is a complex patch-autoregressive transformer over 2×2 spin patches, with two-dimensional axial RoPE and an exact zero-magnetization constraint enforced through autoregressive masking.

D.7 Scaling Experiment

To study how performance scales with model capacity, we train three model sizes on the frustrated Heisenberg J_1–J_2 chain ($N = 12$, $J_1 = 1$, $J_2 = 0.5$, periodic boundary conditions) using PWO, Adam, and minSR. All three sizes share the same autoregressive GRU architecture described in Appendix D.2; they differ only in the number of recurrent layers and hidden dimensions. Table 7 summarises the sizes and the corresponding total parameter counts.

Size	**GRU layers**	**RNN hidden**	**Head hidden**	**Parameters**
Tiny	1	64	64	44,452
Small	2	128	128	269,156
Medium	3	256	256	1,456,356

Table 7: Model sizes used in the scaling experiment. All models use an embedding dimension of 32 and are evaluated on the frustrated Heisenberg J_1–J_2 chain with $N = 12$ sites. Parameter counts include all weights and biases of the full autoregressive network (embedding, recurrent backbone, amplitude head, and phase head).

The hyperparameters used for each optimization method in the scaling experiment are listed in Table 8. The schedule and clipping parameters are identical to those of the main J_1–J_2 experiments (see Tables 5 and 2); only the number of transition steps differs to account for the change in model size.

Hyperparameter	**Adam**	**PWO**	**minSR**
Optimizer / method	Adam	Adam + PWO	minSR
Learning rate	10^{-5}	10^{-5}	10^{-3}
Peak learning rate	3×10^{-4}	10^{-4}	–
Transition steps	40,000	40,000	–
PPO epochs	–	4	–
PPO clip ϵ	–	10^{-3}	–
Advantage normalization	–	Yes	–
Phase loss	–	$\Delta\phi$ clip	–
Phase coefficient	–	1.0	–
Phase clip	–	0.3	–
Center imaginary advantage	–	Yes	–
Normalize imaginary advantage	–	Yes	–
Phase Jacobian baseline	–	Yes	–
SR diagonal shift	–	–	10^{-2}
NTK / minSR mode	–	–	Yes
On-the-fly SR	–	–	Yes
Training samples		2,048	
Seeds		10	

Table 8: Hyperparameters for the scaling experiment on the frustrated Heisenberg J_1–J_2 chain ($N = 12$, $J_1 = 1$, $J_2 = 0.5$, periodic boundary conditions). All methods and sizes use the same optimization hyperparameters; only the model architecture varies across runs (see Table 7). Adam and PWO use a cosine one-cycle learning-rate schedule; minSR uses a constant learning rate.

D.8 RWKV-7 on Ising Model

Hyperparameter	**Adam**	**PWO**
Optimizer / method	Adam	Adam + PWO
Model	RWKV-7	RWKV-7
Model size	1.5B	1.5B
Learning rate	10^{-5}	10^{-5}
Transition steps	1,200	4,800
Decay rate	0.5	0.5
PPO epochs	–	4
PPO clip ϵ	–	10^{-3}
Advantage normalization	–	Yes
Batch size	150	150
Machine power	2	2
Evaluation samples	4,096	4,096
Evaluation batch size	128	128
Exact diagonalization	Yes	Yes

Table 9: Hyperparameters for the RWKV-7 fine-tuning experiments on the transverse-field Ising model. Both methods use an autoregressive RWKV-7 model with 1.5B parameters, $N = 12$, periodic boundary conditions, $J = 1$, $h = 1$, exact diagonalization for evaluation, and samples drawn exactly from the autoregressive Born distribution. Adam corresponds to the single-epoch first-order baseline, while PWO performs four proximal inner epochs per sampled batch.

E Additional Figures

E.1 Individual Seed Plots for All Hamiltonias

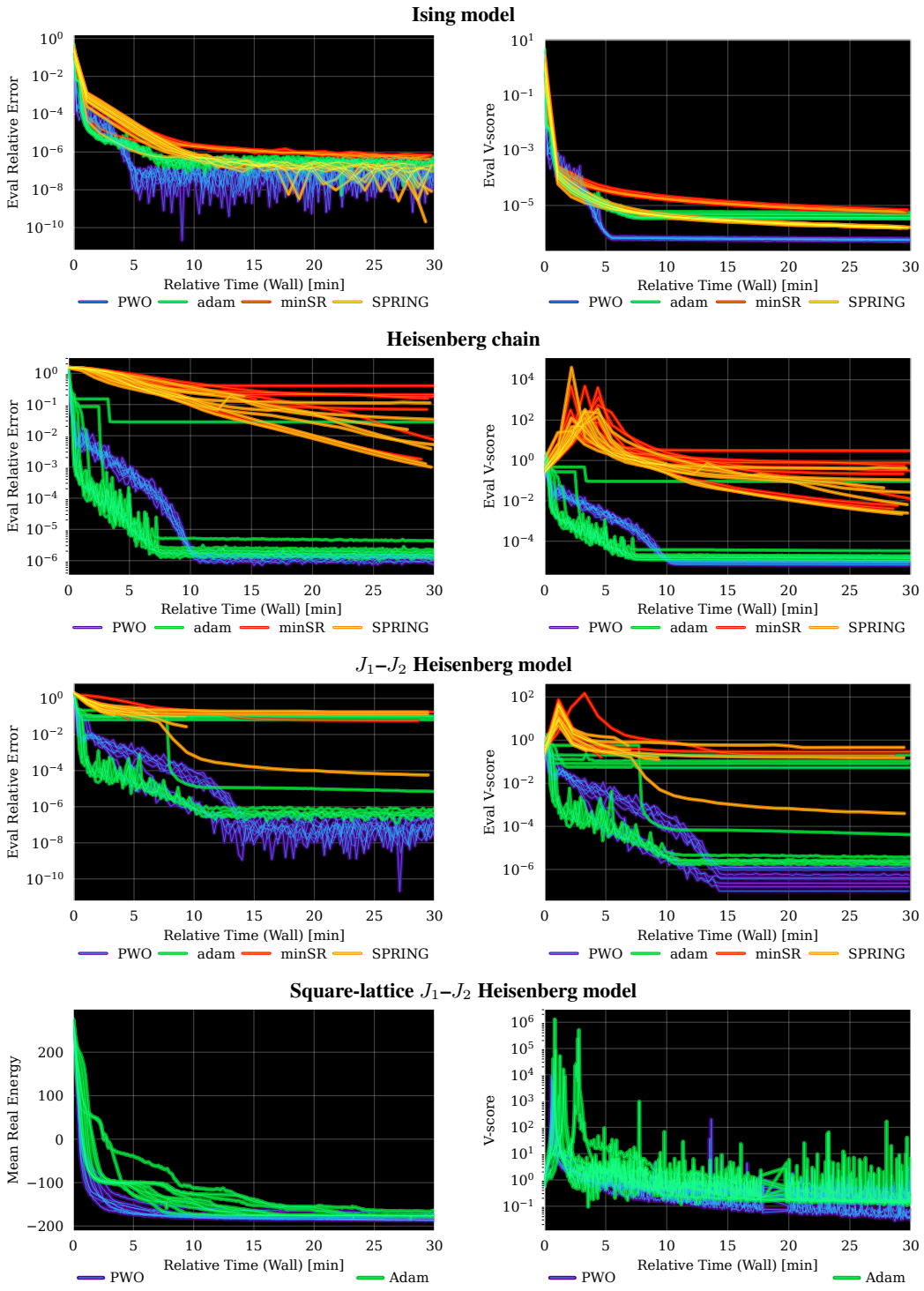


Figure 9: Individual-seed learning curves for all Hamiltonians. Each row corresponds to one Hamiltonian, and the right column reports the V-score.

E.2 Scaling Samples and System Sizes

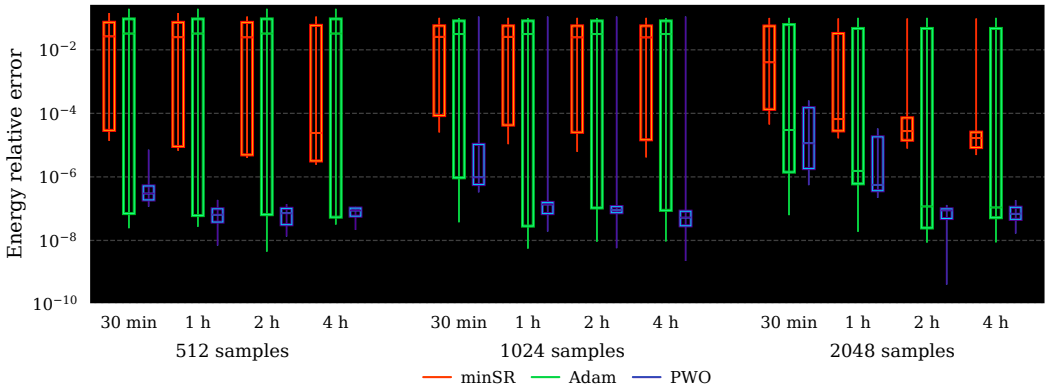


Figure 10: Wall-clock scaling comparison across number of samples and optimization methods for the Heisenberg J_1–J_2 chain ($N = 12$, $J_1 = 1$, $J_2 = 0.5$, periodic boundary conditions). The model architecture is the one corresponding to medium size in Figure 5. Boxplots show the relative error distribution over seeds, with boxes indicating the interquartile range and lines, the minimum and maximum. Results are grouped by number of samples and wall-clock time, and run on a single NVIDIA A100 GPU.

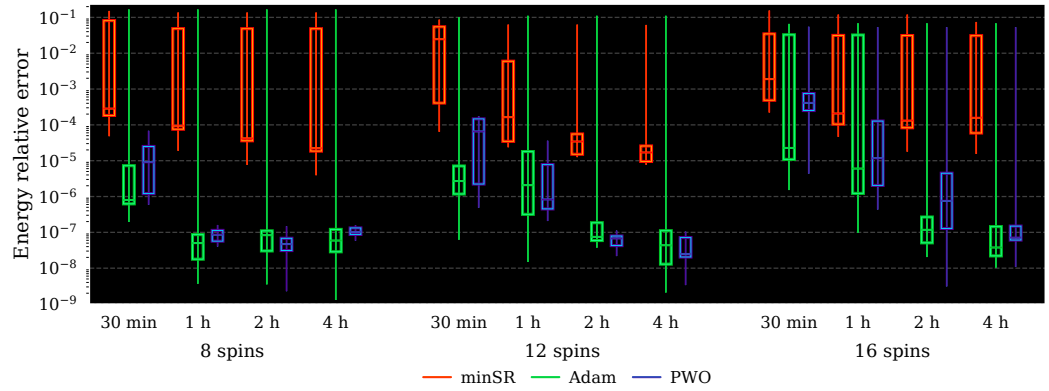


Figure 11: Wall-clock scaling comparison across system size and optimization methods for the Heisenberg J_1–J_2 chain ($J_1 = 1$, $J_2 = 0.5$, periodic boundary conditions and 2048 samples). The model architecture is the one corresponding to medium size in Figure 5. Boxplots show the relative error distribution over seeds, with boxes indicating the interquartile range and lines, the minimum and maximum. Results are grouped by system size and wall-clock time, and run on a single NVIDIA A100 GPU.

E.3 Individual Seed Plots for RWKV7 Fine-tuning

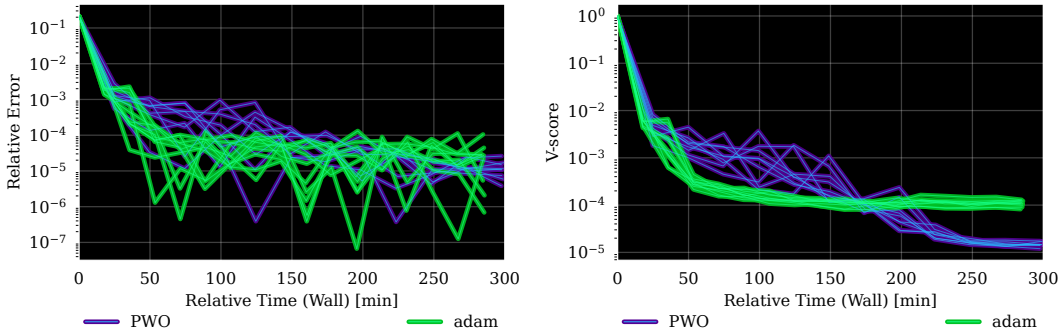


Figure 12: Individual-seed fine-tuning curves for the 1.5B-parameter RWKV-7 neural quantum state on the transverse-field Ising model. Each curve corresponds to one random seed, with relative error shown on the left and V-score on the right. PWO consistently remains stable across seeds and reaches lower final error and variance than the Adam baseline, indicating that the proximal objective improves robustness even in the billion-parameter regime.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction summarize the contributions of the paper supported in the other sections.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We added a Limitations section in Appendix C.4.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: We have clearly stated the assumptions of the only theoretical result, namely proposition 3.1. The proof is in Appendix B.1.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: All the details to reproduce the experiments are provided in the method description and the Appendix (Including hyper-parameters).

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.

(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: Due to anonymity concerns we will not share links to the code until the paper is accepted. All the details to reproduce the experiments are provided in the method description and the Appendix.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Experimental details not provided in the main text can be found in Appendix D.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.

- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We report Inter-Quartile (IQ) results for the main experiments computed over 10 random seeds.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: All experiments report the wall clock time and compute type used to perform them.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

Answer: [Yes]

Justification: This paper conforms with the code of ethics.

Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [N/A]

Justification: We are not aware of any potential societal impacts of our work.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: N.A.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.

- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We cite the owners of any assets used.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: N.A.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: N.A.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: N.A.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: N.A.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.