

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

1 Introduction

Quantum many-body physics studies the collective behavior of systems of many quantum particles. One central task in this field is computing the ground state, $|\psi_0\rangle$, of a given Hamiltonian, $\hat{H}$, but solving this analytically is intractable for large systems due to the exponential growth of the Hilbert space, $\mathcal{H}$.

Variational Monte Carlo (VMC) turns this into an optimization problem over a parametric family of states $|\psi_{\boldsymbol{\theta}}\rangle$ via Monte Carlo sampling of spin configurations, $\mathbf{s}$, from the Born distribution, $\mathcal{P}_{\boldsymbol{\theta}}(\mathbf{s}) \propto |\psi_{\boldsymbol{\theta}}(\mathbf{s})|^2$:

$$\boldsymbol{\theta}^* = \arg \min_{\boldsymbol{\theta}} E[\psi_{\boldsymbol{\theta}}] = \arg \min_{\boldsymbol{\theta}} \mathbb{E}_{\mathbf{s} \sim \mathcal{P}_{\boldsymbol{\theta}}} [E_{\boldsymbol{\theta}}^{\text{loc}}(\mathbf{s})], \quad (1)$$

where $E_{\boldsymbol{\theta}}^{\text{loc}}(\mathbf{s}) = \sum_{\mathbf{s}'} \langle \mathbf{s} | \hat{H} | \mathbf{s}' \rangle \psi_{\boldsymbol{\theta}}(\mathbf{s}') / \psi_{\boldsymbol{\theta}}(\mathbf{s})$ is the so-called local energy. This problem can be solved using gradient-based optimizers, where the gradient is estimated via Monte Carlo sampling as

$$\nabla_{\boldsymbol{\theta}} E[\psi_{\boldsymbol{\theta}}] = \mathbb{E}_{\mathbf{s} \sim \mathcal{P}_{\boldsymbol{\theta}}} \left[2 \operatorname{Re} \left\{ \Delta E_{\boldsymbol{\theta}}^{\text{loc}}(\mathbf{s}) (\nabla_{\boldsymbol{\theta}} \log \psi_{\boldsymbol{\theta}}(\mathbf{s}))^* \right\} \right]. \quad (2)$$

where $\Delta E_{\boldsymbol{\theta}}^{\text{loc}}(\mathbf{s}) \equiv E_{\boldsymbol{\theta}}^{\text{loc}}(\mathbf{s}) - \mathbb{E}_{\mathbf{s} \sim \mathcal{P}_{\boldsymbol{\theta}}} [E_{\boldsymbol{\theta}}^{\text{loc}}(\mathbf{s})]$ is the centered local energy.

Regarding the ansatz, Neural Quantum States (NQS) [2, 10, 4] are one of the most popular choices nowadays. They use a neural network $f_{\boldsymbol{\theta}}: \{\pm 1\}^N \rightarrow \mathbb{C}$ to parametrize $\psi_{\boldsymbol{\theta}}(\mathbf{s}) = \langle \mathbf{s} | \psi_{\boldsymbol{\theta}} \rangle$. As for the optimizer, first-order methods do not usually perform well in the complex and noisy energy landscapes of VMC [14, 11]. For this reason, the standard choice in NQS is Stochastic Reconfiguration (SR) [19], a second-order method equivalent to natural gradient descent [1] that solves:

$$\delta \boldsymbol{\theta}^* = \arg \min_{\delta \boldsymbol{\theta}} E[\psi_{\boldsymbol{\theta} + \delta \boldsymbol{\theta}}] - E[\psi_{\boldsymbol{\theta}}] \quad \text{s.t.} \quad \mathcal{I}(\psi_{\boldsymbol{\theta}}, \psi_{\boldsymbol{\theta} + \delta \boldsymbol{\theta}}) \leq \epsilon \in \mathbb{R}, \quad (3)$$

where infidelity, $\mathcal{I}$, is approximated to second order with the Fubini-Study metric:

$$\mathbf{S} = \operatorname{Re} \{ \operatorname{Cov}_{\mathbf{s} \sim \mathcal{P}_{\boldsymbol{\theta}}} [(\nabla_{\boldsymbol{\theta}} \log \psi_{\boldsymbol{\theta}})^*, \nabla_{\boldsymbol{\theta}} \log \psi_{\boldsymbol{\theta}}] \}. \quad (4)$$

As a result, SR modifies the update rule in the following way:

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \eta \mathbf{S}^{-1} \nabla_{\boldsymbol{\theta}} E[\psi_{\boldsymbol{\theta}}], \quad \text{with} \quad \eta \in \mathbb{R}. \quad (5)$$

NQS are a promising approach to tackle the exponential complexity of quantum many-body physics. However, there are some issues that are preventing them from solving important real-world problems. First, the standard practice is to discard all the samples after each optimization step, even if the wavefunction change is small, which makes the optimization sample-inefficient. Second, computing $\mathbf{S}^{-1}$ in SR is costly and scales poorly with the number of parameters. Some refinements of this method like minSR [3] and SPRING [6] try to reduce this cost, but at the price of an increased dependence on the number of samples. Finally, Markov Chain Monte Carlo methods struggle to explore the Hilbert space [20, 5], whereas SR is unstable when applied to the alternative approach of autoregressive (AR) models [18, 7, 12].

2 Proximal Wavefunction Optimization

In this work, we show that the advantage policy gradient framework of Reinforcement Learning (RL) maps naturally to VMC optimization. In fact, this correspondence is exact for stoquastic Hamiltonians, for which the wavefunction is real and non-negative. Moreover, Trust Region Policy Optimization (TRPO) [17] is in perfect analogy with SR, replacing the wavefunction with the policy and the Fisher information matrix with $\mathbf{S}$. This motivates the extension to NQS of Proximal Policy Optimization [16]: a first-order optimization method that allows for sample reuse and overcomes the limitations of TRPO, while preserving its trust-region philosophy via simple yet effective heuristics.

The proposed method, Proximal Wavefunction Optimization (PWO), does exactly this. We first show that the VMC gradient can be split into real and imaginary components as

$$\nabla_{\boldsymbol{\theta}} E[\psi_{\boldsymbol{\theta}}] = \mathbb{E}_{\mathbf{s} \sim \mathcal{P}_{\boldsymbol{\theta}}} [2A_{\boldsymbol{\theta}}^{\text{R}}(\mathbf{s}) \nabla_{\boldsymbol{\theta}} \log |\psi_{\boldsymbol{\theta}}|] + \mathbb{E}_{\mathbf{s} \sim \mathcal{P}_{\boldsymbol{\theta}}} [2A_{\boldsymbol{\theta}}^{\text{I}}(\mathbf{s}) \nabla_{\boldsymbol{\theta}} \arg \psi_{\boldsymbol{\theta}}], \quad (6)$$

where $A_{\boldsymbol{\theta}}^{\text{R}}(\mathbf{s}) = \text{Re} \{ \Delta E_{\boldsymbol{\theta}}^{\text{loc}}(\mathbf{s}) \}$ and $A_{\boldsymbol{\theta}}^{\text{I}}(\mathbf{s}) = \text{Im} \{ \Delta E_{\boldsymbol{\theta}}^{\text{loc}}(\mathbf{s}) \}$. Then, we realize that each component has the same structure and is analogous to the advantage policy gradient from RL. For this reason, we propose a clipped surrogate objective for each component analogous to the one that PPO introduces for the policy gradient:

$$L_{\text{PWO}}(\boldsymbol{\theta}; \boldsymbol{\theta}_{\text{old}}) = L_{\text{mod}}^{\text{clip}}(\boldsymbol{\theta}; \boldsymbol{\theta}_{\text{old}}) + L_{\text{arg}}^{\text{clip}}(\boldsymbol{\theta}; \boldsymbol{\theta}_{\text{old}}). \quad (7)$$

For the modulus component,

$$L_{\text{mod}}^{\text{clip}}(\boldsymbol{\theta}; \boldsymbol{\theta}_{\text{old}}) = \mathbb{E}_{\mathbf{s} \sim \mathcal{P}_{\boldsymbol{\theta}_{\text{old}}}} [\max(r_{\boldsymbol{\theta}}(\mathbf{s}) A_{\boldsymbol{\theta}_{\text{old}}}^{\text{R}}(\mathbf{s}), \text{clip}(r_{\boldsymbol{\theta}}(\mathbf{s}), 1 - \epsilon, 1 + \epsilon) A_{\boldsymbol{\theta}_{\text{old}}}^{\text{R}}(\mathbf{s}))], \quad (8)$$

where

$$r_{\boldsymbol{\theta}}(\mathbf{s}) = \mathcal{P}_{\boldsymbol{\theta}}(\mathbf{s}) / \mathcal{P}_{\boldsymbol{\theta}_{\text{old}}}(\mathbf{s}) \quad (9)$$

is the importance sampling ratio and $\epsilon \in \mathbb{R}$ is the clipping threshold. For the phase component,

$$L_{\text{arg}}^{\text{clip}}(\boldsymbol{\theta}; \boldsymbol{\theta}_{\text{old}}) = \mathbb{E}_{\mathbf{s} \sim \mathcal{P}_{\boldsymbol{\theta}_{\text{old}}}} [\text{sg}(r_{\boldsymbol{\theta}}(\mathbf{s})) \max(\phi_{\boldsymbol{\theta}}(\mathbf{s}) A_{\boldsymbol{\theta}_{\text{old}}}^{\text{I}}(\mathbf{s}), \text{clip}(\phi_{\boldsymbol{\theta}}(\mathbf{s}), -\delta, \delta) A_{\boldsymbol{\theta}_{\text{old}}}^{\text{I}}(\mathbf{s}))], \quad (10)$$

where

$$\phi_{\boldsymbol{\theta}}(\mathbf{s}) = 2 \text{atan}_2(\sin(\arg \psi_{\boldsymbol{\theta}} - \arg \psi_{\boldsymbol{\theta}_{\text{old}}}), \cos(\arg \psi_{\boldsymbol{\theta}} - \arg \psi_{\boldsymbol{\theta}_{\text{old}}})) \quad (11)$$

is twice the phase difference wrapped modulo 2π ; $\text{sg}(\cdot)$ denotes the stop-gradient operator, which detaches its input from the gradient computation; and $\delta \in [0, \pi]$ is the phase-clipping threshold.

The gradient of this objective is consistent with the original one every time we sample new spin configurations from the current parametrization of the wavefunction:

$$\nabla_{\boldsymbol{\theta}} L_{\text{PWO}}(\boldsymbol{\theta}; \boldsymbol{\theta}_{\text{old}}) \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_{\text{old}}} = \nabla_{\boldsymbol{\theta}} E[\psi_{\boldsymbol{\theta}}] \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_{\text{old}}}. \quad (12)$$

Moreover, similar to the monotonic improvement guarantees offered by trust-region methods like Conservative Policy Iteration (CPI) [8], TRPO [17] or PPO [16] from RL, PWO is supported by the following upper bound on the energy difference between two parametrizations of the wavefunction:

$$E[\psi_{\boldsymbol{\theta}}] - E[\psi_{\boldsymbol{\theta}_{\text{old}}}] \leq L_{\text{PWO}}(\boldsymbol{\theta}; \boldsymbol{\theta}_{\text{old}}) + 2C_{\boldsymbol{\theta}_{\text{old}}} \mathcal{I}(\psi_{\boldsymbol{\theta}}, \psi_{\boldsymbol{\theta}_{\text{old}}}), \quad (13)$$

with

$$C_{\boldsymbol{\theta}_{\text{old}}} = \left\| \hat{H} - E[\psi_{\boldsymbol{\theta}_{\text{old}}}] \mathbb{1} \right\|_{\infty} + \|A_{\boldsymbol{\theta}_{\text{old}}}^{\text{R}}\|_{\infty} + 4\pi^2 \|A_{\boldsymbol{\theta}_{\text{old}}}^{\text{I}}\|_{\infty}. \quad (14)$$

This means that, as long as the infidelity between consecutive updates of the wavefunction in the optimization process is kept small enough, minimizing the PWO objective yields a monotonic decrease in energy. Moreover, if $r_{\boldsymbol{\theta}}(\mathbf{s}) \in [1 - \epsilon, 1 + \epsilon]$ and $\phi_{\boldsymbol{\theta}}(\mathbf{s}) \in [-\delta, \delta]$ for all $\mathbf{s}$, which PWO encourages but does not guarantee, then infidelity presents the following upper bound:

$$\mathcal{I}(\psi_{\boldsymbol{\theta}}, \psi_{\boldsymbol{\theta}_{\text{old}}}) \leq \left(1 - \frac{1 + \sqrt{1 - \epsilon^2}}{2} \cos^2 \frac{\delta}{2} \right), \quad (15)$$

which behaves as $(\epsilon^2 + \delta^2)/4 + \mathcal{O}(\epsilon^4 + \epsilon^2\delta^2 + \delta^4)$ for small ϵ and δ .

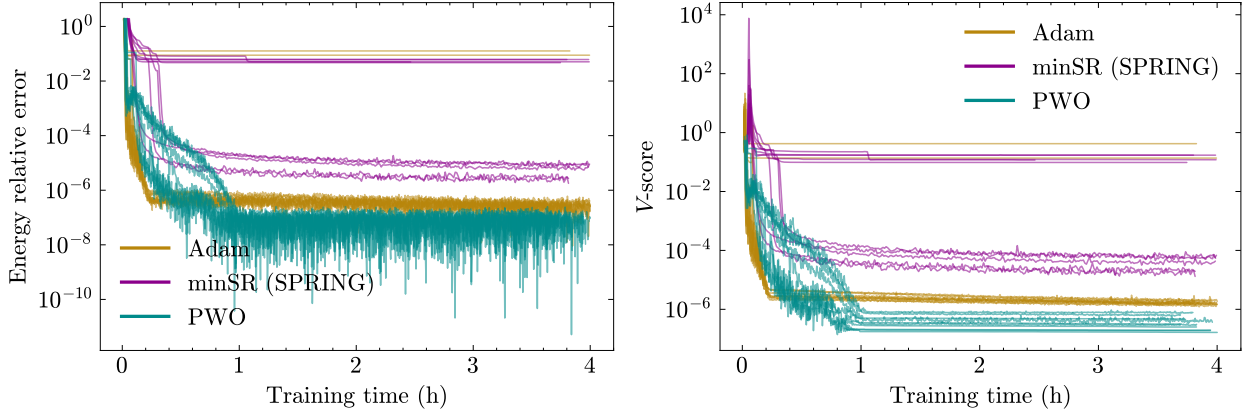


Figure 1: Training history for the J_1 - J_2 Heisenberg model in a periodic chain of 12 spins at $J_2/J_1 = 0.5$. PWO outperforms Adam and SR, which struggle to converge in this setting.

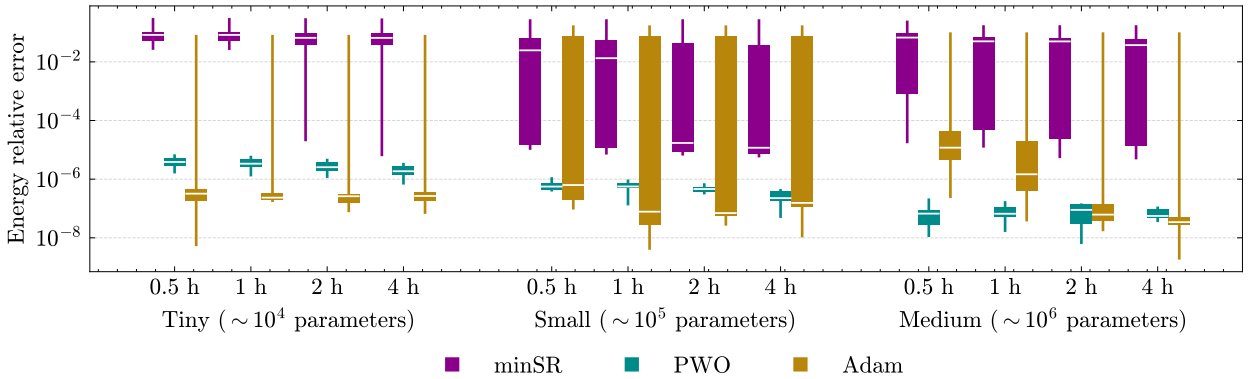


Figure 2: Scaling of PWO with the number of parameters (depth and width) for the same set up as in Figure 1. The speed and accuracy of PWO improve as the model size increases, while stability is maintained, which shows that it can take advantage of the increased expressivity of larger models.

3 Experiments

Across frustrated J_1 - J_2 one- and two-dimensional spin systems [21], PWO improves stability and wall-clock convergence over Adam [9], minSR, and SPRING. Figure 1 shows the results for the J_1 - J_2 chain, while Figure 2 presents the results of the scalability experiment with the model size. Additionally, we fine-tune a 1.5B-parameter RWKV-7 model [13], demonstrating NQS optimization at a scale three orders of magnitude beyond prior work [15].

4 Conclusions

We introduced Proximal Wavefunction Optimization (PWO), a scalable first-order optimization method for NQS, derived from an explicit connection between VMC and RL, that resembles PPO. This method is computationally cheaper than Adam and minSR, allows for sample reuse across optimization steps and provides the convergence stability of higher order methods via a trust region in wavefunction space. The proposed algorithm is supported by both theoretical and empirical evidence. More broadly, our results bridge the RL and NQS communities, enabling further mutual inspiration.

References

- [1] Shun-ichi Amari. “Natural Gradient Works Efficiently in Learning”. In: *Neural Computation* 10.2 (1998), pp. 251–276. DOI: 10.1162/089976698300017746. URL: <https://doi.org/10.1162/089976698300017746>.
- [2] Giuseppe Carleo and Matthias Troyer. “Solving the Quantum Many-Body Problem with Artificial Neural Networks”. In: *Science* 355.6325 (2017), pp. 602–606. DOI: 10.1126/science.aag2302. URL: <https://doi.org/10.1126/science.aag2302>.
- [3] Ao Chen and Markus Heyl. “Empowering Deep Neural Quantum States Through Efficient Optimization”. In: *Nature Physics* 20.9 (2024), pp. 1476–1481. DOI: 10.1038/s41567-024-02566-1. URL: <https://doi.org/10.1038/s41567-024-02566-1>.
- [4] Anna Dawid et al. *Machine Learning in Quantum Sciences*. Cambridge University Press, 2025. ISBN: 9781009504935. DOI: 10.1017/9781009504942. URL: <https://doi.org/10.1017/9781009504942>.
- [5] Luigi Del Debbio, Gian Mario Manca, and Ettore Vicari. “Critical Slowing Down of Topological Modes”. In: *Physics Letters B* 594.3–4 (2004), pp. 315–323. DOI: 10.1016/j.physletb.2004.05.038. URL: <https://doi.org/10.1016/j.physletb.2004.05.038>.
- [6] Gil Goldshlager, Nilin Abrahamsen, and Lin Lin. “A Kaczmarz-Inspired Approach to Accelerate the Optimization of Neural Network Wavefunctions”. In: *Journal of Computational Physics* 516 (2024), p. 113351. DOI: 10.1016/j.jcp.2024.113351. URL: <https://doi.org/10.1016/j.jcp.2024.113351>.
- [7] Mohamed Hibat-Allah et al. “Recurrent Neural Network Wave Functions”. In: *Physical Review Research* 2.2 (2020), p. 023358. DOI: 10.1103/PhysRevResearch.2.023358. URL: <https://doi.org/10.1103/PhysRevResearch.2.023358>.
- [8] Sham Kakade and John Langford. “Approximately Optimal Approximate Reinforcement Learning”. In: *Proceedings of the Nineteenth International Conference on Machine Learning*. 2002, pp. 267–274. URL: <https://dl.acm.org/doi/10.5555/645531.656005>.
- [9] Diederik P. Kingma and Jimmy Ba. “Adam: A Method for Stochastic Optimization”. In: *arXiv preprint arXiv:1412.6980* (2014). DOI: 10.48550/arXiv.1412.6980. URL: <https://arxiv.org/abs/1412.6980>.
- [10] Hannah Lange et al. “From Architectures to Applications: A Review of Neural Quantum States”. In: *arXiv preprint arXiv:2402.09402* (2024). DOI: 10.48550/arXiv.2402.09402. URL: <https://arxiv.org/abs/2402.09402>.
- [11] Jing Liu, Ying Tang, and Pan Zhang. “Efficient Optimization of Variational Autoregressive Networks with Natural Gradient”. In: *Physical Review E* 111.2 (2025), p. 025304. DOI: 10.1103/PhysRevE.111.025304. URL: <https://doi.org/10.1103/PhysRevE.111.025304>.
- [12] M. Schuyler Moss et al. “Leveraging recurrence in neural network wavefunctions for large-scale simulations of Heisenberg antiferromagnets on the square lattice”. In: *Physical Review B* 112.13 (2025), p. 134450. ISSN: 2469-9950. DOI: 10.1103/6ccd-wzhz.
- [13] Bo Peng et al. “RWKV-7 “Goose” with Expressive Dynamic State Evolution”. In: *arXiv preprint arXiv:2503.14456* (2025). DOI: 10.48550/arXiv.2503.14456. URL: <https://arxiv.org/abs/2503.14456>.
- [14] David Pfau et al. “Ab Initio Solution of the Many-Electron Schrödinger Equation with Deep Neural Networks”. In: *Physical Review Research* 2.3 (2020), p. 033429. DOI: 10.1103/PhysRevResearch.2.033429. URL: <https://doi.org/10.1103/PhysRevResearch.2.033429>.

-
- [15] Riccardo Rende et al. “Foundation neural-networks quantum states as a unified Ansatz for multiple hamiltonians”. In: *Nature Communications* 16.1 (2025), p. 7213. DOI: 10.1038/s41467-025-62098-x. eprint: 2502.09488.
- [16] John Schulman et al. “Proximal Policy Optimization Algorithms”. In: *arXiv preprint arXiv:1707.06347* (2017). DOI: 10.48550/arXiv.1707.06347. URL: <https://arxiv.org/abs/1707.06347>.
- [17] John Schulman et al. “Trust Region Policy Optimization”. In: *arXiv preprint arXiv:1502.05477* (2015). DOI: 10.48550/arXiv.1502.05477. URL: <https://arxiv.org/abs/1502.05477>.
- [18] Or Sharir et al. “Deep Autoregressive Models for the Efficient Variational Simulation of Many-Body Quantum Systems”. In: *Physical Review Letters* 124.2 (2020), p. 020503. DOI: 10.1103/PhysRevLett.124.020503. URL: <https://doi.org/10.1103/PhysRevLett.124.020503>.
- [19] Sandro Sorella. “Green Function Monte Carlo with Stochastic Reconfiguration”. In: *Physical Review Letters* 80.20 (1998), pp. 4558–4561. DOI: 10.1103/PhysRevLett.80.4558. URL: <https://doi.org/10.1103/PhysRevLett.80.4558>.
- [20] Ulli Wolff. “Critical Slowing Down”. In: *Nuclear Physics B - Proceedings Supplements* 17 (1990), pp. 93–102. DOI: 10.1016/0920-5632(90)90224-I. URL: [https://doi.org/10.1016/0920-5632\(90\)90224-I](https://doi.org/10.1016/0920-5632(90)90224-I).
- [21] Dian Wu et al. “Variational Benchmarks for Quantum Many-Body Problems”. In: *Science* 386.6719 (2024), pp. 296–301. DOI: 10.1126/science.adg9774. URL: <https://doi.org/10.1126/science.adg9774>.