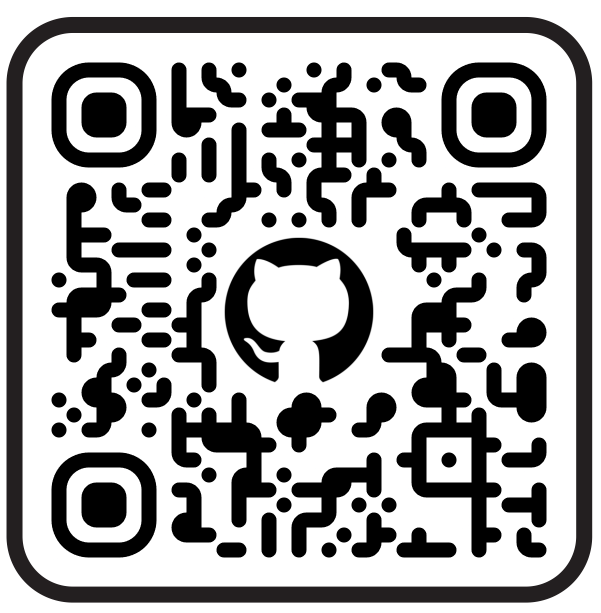




ONE MORE TIME: REVISITING NEURAL QUANTUM STATES FROM A REINFORCEMENT LEARNING PERSPECTIVE

Juan Agustín Duque*, Sergio García Heredia*, Vinicius Hernandes, Eliska Greplova, Thomas Spriggs**, Aaron Courville**, Anna Dawid**

NEURAL QUANTUM STATES

Quantum Many-Body Physics studies the collective behavior of systems of many quantum particles. One fundamental problem is finding the **ground state** of a given Hamiltonian, but solving this analytically is intractable for large systems due to the exponential growth of the Hilbert space.

Variational Monte Carlo (VMC) turns this into an optimization problem via MC sampling

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

This can be solved using gradient-based optimizers, with

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

Stochastic Reconfiguration (SR) is the standard NQS optimizer, which solves:

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

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

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

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

$$\theta \leftarrow \theta - \eta \mathbf{S}^{-1} \nabla_{\theta} E[\psi_{\theta}].$$

Neural Quantum States (NQS) use a neural network as the wavefunction ansatz:

$$\begin{aligned} f_{\theta}: \{\pm 1\}^N &\rightarrow \mathbb{C} \\ s &\mapsto \psi_{\theta}(s) \end{aligned}$$

Problems:

- Computing $\mathbf{S}^{-1}$ is costly and scales poorly with the number of parameters or samples.
- Samples are discarded after each step, despite the small wavefunction change.
- Markov Chain Monte Carlo struggles to explore the Hilbert space and SR is unstable with AR models.

REINFORCEMENT LEARNING

The policy gradient framework of Reinforcement Learning (RL) maps naturally to VMC optimization. This correspondence is exact for stoquastic Hamiltonians, since:

$$\nabla_{\theta} E[\psi_{\theta}] = \mathbb{E}_{s \sim \mathcal{P}_{\theta}} [\Delta E_{\theta}^{\text{loc}}(s) \nabla_{\theta} \log \mathcal{P}_{\theta}(s)].$$

Policy gradient methods learn a policy π_{θ} to maximize the expected return:

$$\theta^* = \arg \max_{\theta} J(\pi_{\theta}) = \arg \max_{\theta} \mathbb{E}_{\tau \sim \mathcal{P}_{\theta}} [R(\tau)].$$

This can be solved using gradient-based optimizers, with

$$\nabla_{\theta} J(\theta) = \mathbb{E}_{(s,a) \sim \mathcal{V}_{\theta}} [A^{\pi_{\theta}}(s,a) \nabla_{\theta} \log \pi_{\theta}(a | s)].$$

Trust Region Policy Optimization (TRPO) ensures monotonic improvement solving:

$$\theta^* = \arg \max_{\theta} J_{\theta}^{\text{CPI}}(\theta) \quad \text{s.t.} \quad \mathbb{E}_{s \sim \mathcal{V}_{\theta_{\text{old}}}} [D_{\text{KL}}(\pi_{\theta} \| \pi_{\theta_{\text{old}}})] \leq \epsilon \in \mathbb{R},$$

where KL divergence, D_{KL} , is approximated to second order with the Fisher Information Matrix:

$$\mathbf{F} = \mathbb{E}_{(s,a) \sim \mathcal{V}_{\theta}} [(\nabla_{\theta} \log \pi_{\theta})(\nabla_{\theta} \log \pi_{\theta})^{\top}].$$

As a result, TRPO modifies policy gradient ascent in the following way:

$$\theta \leftarrow \theta + \eta \mathbf{F}_{\text{old}}^{-1} \nabla_{\theta} J_{\theta}^{\text{CPI}}(\theta).$$

Proximal Policy Optimization (PPO) outperforms TRPO by replacing the costly Fisher inverse with a clipping heuristic:

$$J_{\theta_{\text{old}}}^{\text{clip}}(\theta) = \mathbb{E}_{(s,a) \sim \mathcal{V}_{\theta}} [\min(r_{\theta} A^{\pi_{\theta_{\text{old}}}}, \text{clip}(r_{\theta}, 1 - \epsilon, 1 + \epsilon) A^{\pi_{\theta_{\text{old}}}})].$$

Advantages:

- **First-order** optimizer — no metric inversion.
- **Sample reuse** via importance sampling: multiple gradient steps per batch.
- **Trust-region** constraint that prevents policy collapse.

PPO EXTENSION TO NQS

PROXIMAL WAVEFUNCTION OPTIMIZATION

Proximal Wavefunction Optimization (PWO) splits the gradient into real and imaginary parts,

$$\nabla_{\theta} E[\psi_{\theta}] = \mathbb{E}_{s \sim \mathcal{P}_{\theta}} [2 \text{Re} \{ \Delta E_{\theta}^{\text{loc}}(s) \} \nabla_{\theta} \log |\psi_{\theta}|] + \mathbb{E}_{s \sim \mathcal{P}_{\theta}} [2 \text{Im} \{ \Delta E_{\theta}^{\text{loc}}(s) \} \nabla_{\theta} \arg \psi_{\theta}],$$

and applies a separate clipped surrogate loss to each part:

$$L(\theta; \theta_{\text{old}}) = L_{\text{mod}}^{\text{clip}}(\theta; \theta_{\text{old}}) + L_{\text{arg}}^{\text{clip}}(\theta; \theta_{\text{old}}).$$

- For the real part:

$$L_{\text{mod}}^{\text{clip}}(\theta; \theta_{\text{old}}) = \mathbb{E}_{s \sim \mathcal{P}_{\theta_{\text{old}}}} [\max(r_{\theta}(s) \text{Re} \{ \Delta E_{\theta}^{\text{loc}}(s) \}, \text{clip}(r_{\theta}(s), 1 - \epsilon, 1 + \epsilon) \text{Re} \{ \Delta E_{\theta}^{\text{loc}}(s) \})],$$

where

$$r_{\theta}(s) = \mathcal{P}_{\theta}(s) / \mathcal{P}_{\theta_{\text{old}}}(s).$$

- For the imaginary part:

$$L_{\text{arg}}^{\text{clip}}(\theta; \theta_{\text{old}}) = \mathbb{E}_{s \sim \mathcal{P}_{\theta_{\text{old}}}} [\text{sg}(r_{\theta}(s)) \max(\phi_{\theta}(s) \text{Im} \{ \Delta E_{\theta}^{\text{loc}}(s) \}, \text{clip}(\phi_{\theta}(s), -\delta, \delta) \text{Im} \{ \Delta E_{\theta}^{\text{loc}}(s) \})],$$

where

$$\phi_{\theta}(s) = 2 \text{atan}(\sin(\arg \psi_{\theta} - \arg \psi_{\theta_{\text{old}}}), \cos(\arg \psi_{\theta} - \arg \psi_{\theta_{\text{old}}}).$$

Thus,

$$\nabla_{\theta} (L_{\text{mod}}^{\text{clip}}(\theta; \theta_{\text{old}}) + L_{\text{arg}}^{\text{clip}}(\theta; \theta_{\text{old}})) \Big|_{\theta=\theta_{\text{old}}} = \nabla_{\theta} E[\psi_{\theta}] \Big|_{\theta=\theta_{\text{old}}}.$$

Similar to CPI and PPO, PWO provides an upper bound on the energy improvement:

$$E[\psi_{\theta}] - E[\psi_{\theta_{\text{old}}}] \leq L_{\text{mod}}^{\text{clip}}(\theta; \theta_{\text{old}}) + L_{\text{arg}}^{\text{clip}}(\theta; \theta_{\text{old}}) + 2C_{\theta_{\text{old}}} \left(1 - \frac{1 + \sqrt{1 - \epsilon^2}}{2} \cos^2 \frac{\delta}{2} \right).$$

Algorithm 1: Proximal Wavefunction Optimization (PWO)

Input: $\hat{H}$, ψ_{θ} , M , K , ϵ , δ , η .

```

1 Initialize  $\theta$ 
2 while not converged do
3    $\theta_{\text{old}} \leftarrow \theta$ 
4   Sample  $\{s_i\}_{i=1}^M \sim \mathcal{P}_{\theta_{\text{old}}}$ 
5   Cache  $\log \mathcal{P}_{\theta_{\text{old}}}(s_i)$ ,  $\arg \psi_{\theta_{\text{old}}}(s_i)$  and  $\Delta E_{\theta_{\text{old}}}^{\text{loc}}(s)$  for all  $\{s_i\}_{i=1}^M$ 
6   for  $k = 1, \dots, K$  do
7     Compute  $\log \mathcal{P}_{\theta}(s_i)$  and  $\arg \psi_{\theta}(s_i)$  for all  $\{s_i\}_{i=1}^M$ 
8     Compute  $\tilde{L}_{\text{mod}}^{\text{clip}}$  and  $\tilde{L}_{\text{arg}}^{\text{clip}}$ 
9      $\theta \leftarrow \theta - \eta \nabla_{\theta} (\tilde{L}_{\text{mod}}^{\text{clip}} + \tilde{L}_{\text{arg}}^{\text{clip}})$ 

```

EXPERIMENTS

PWO can outperform both Adam and SR in terms of convergence speed, stability and accuracy in standard 1D benchmarks with AR models.

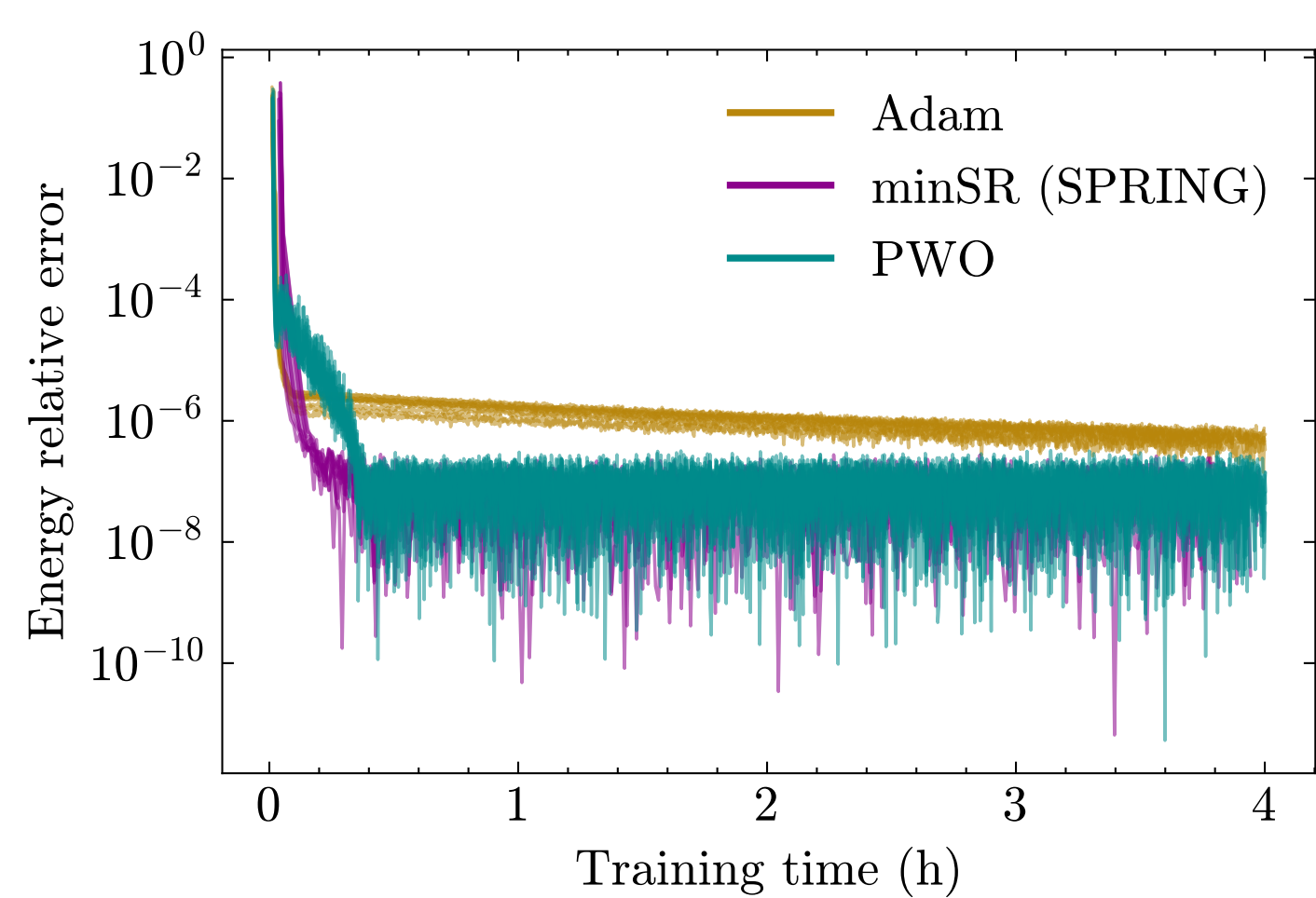
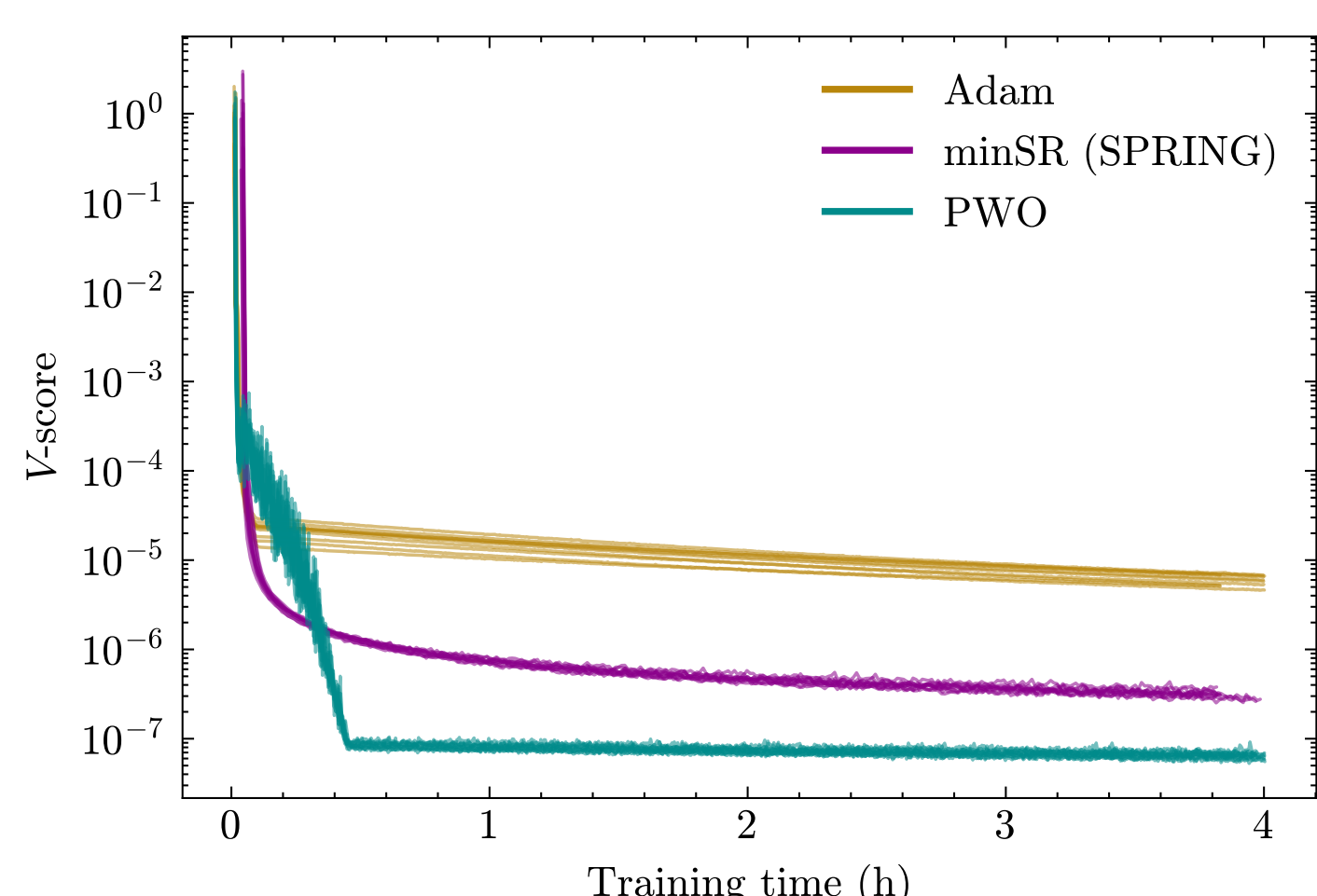


Figure 1. Training history for the Transverse Field Ising Model (TFIM) in a periodic chain of 12 spins at the critical point. This is an example of a stoquastic Hamiltonian, where the policy gradient correspondence is exact.

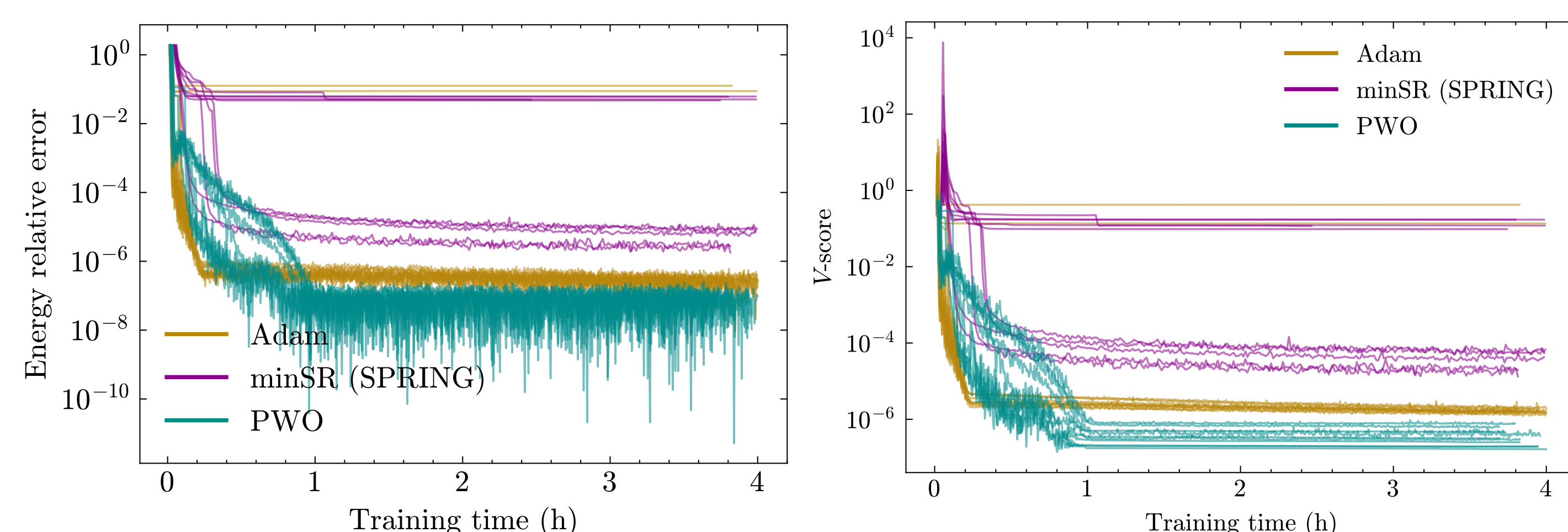


Figure 2. Training history for the J_1 - J_2 Heisenberg model in a periodic chain of 12 spins at $J_2/J_1 = 0.5$. This is an example of a non-stoquastic Hamiltonian, where the policy gradient correspondence is not exact, but PWO still outperforms Adam and SR, which struggle to converge in this setting.

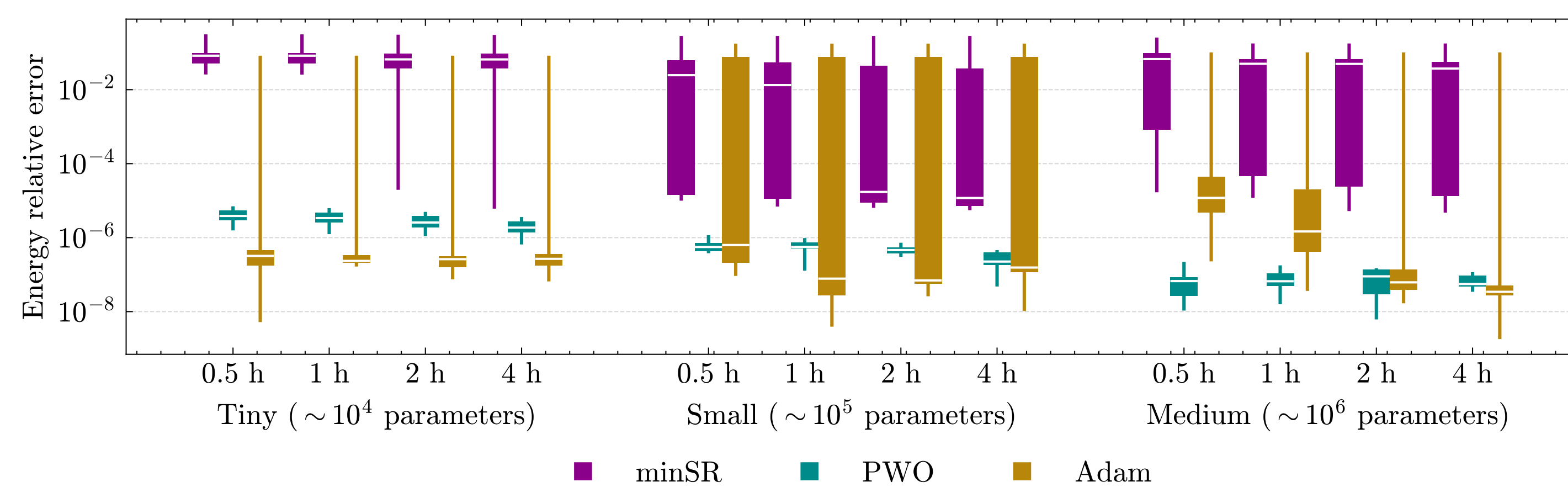


Figure 3. Scaling of PWO with the number of parameters (depth and width) for the same set up as in Figure 2. The speed and accuracy of PWO improve as the model size increases, while stability is maintained, which shows that it can take advantage of the increased expressivity of larger models.

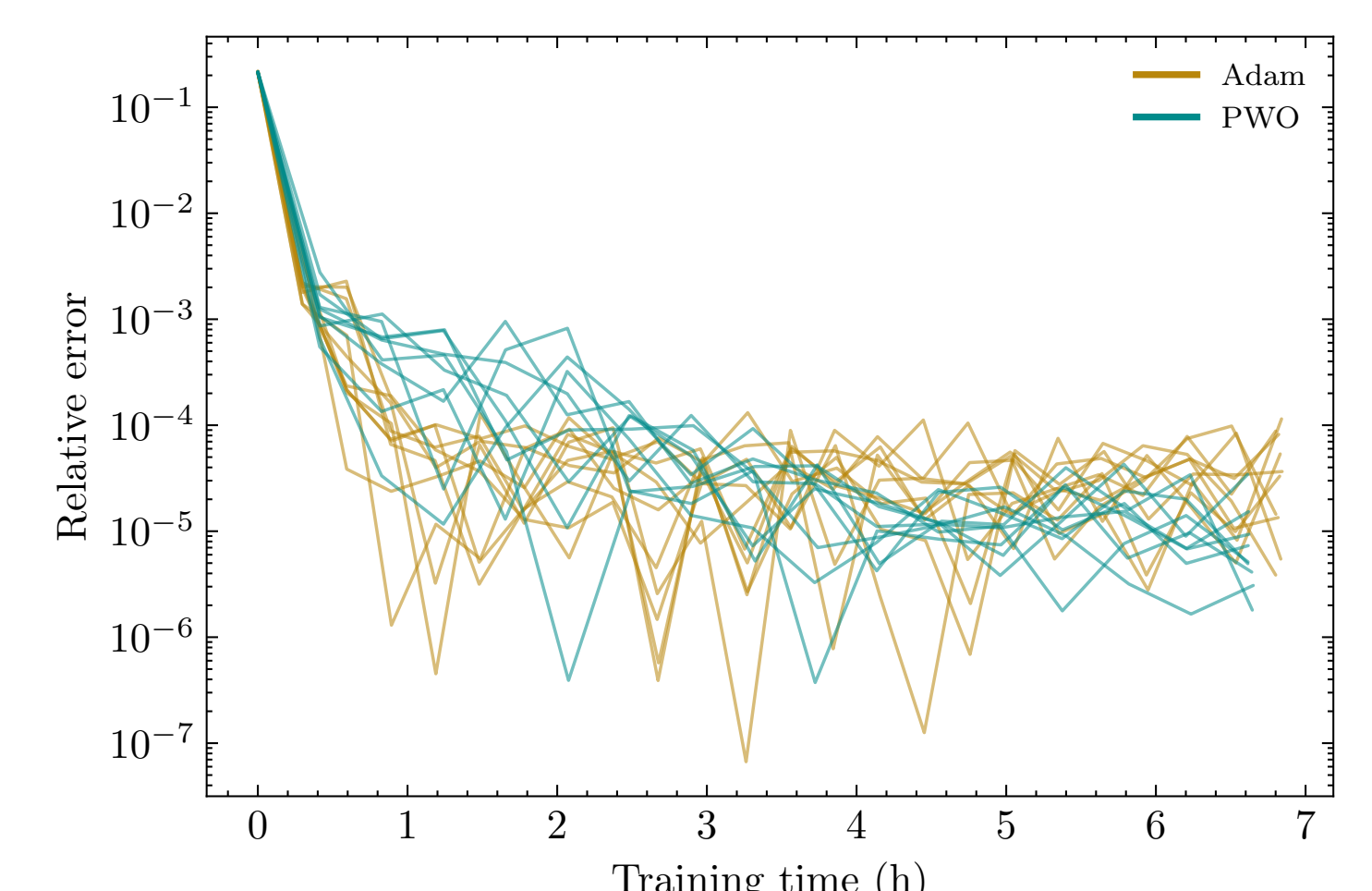
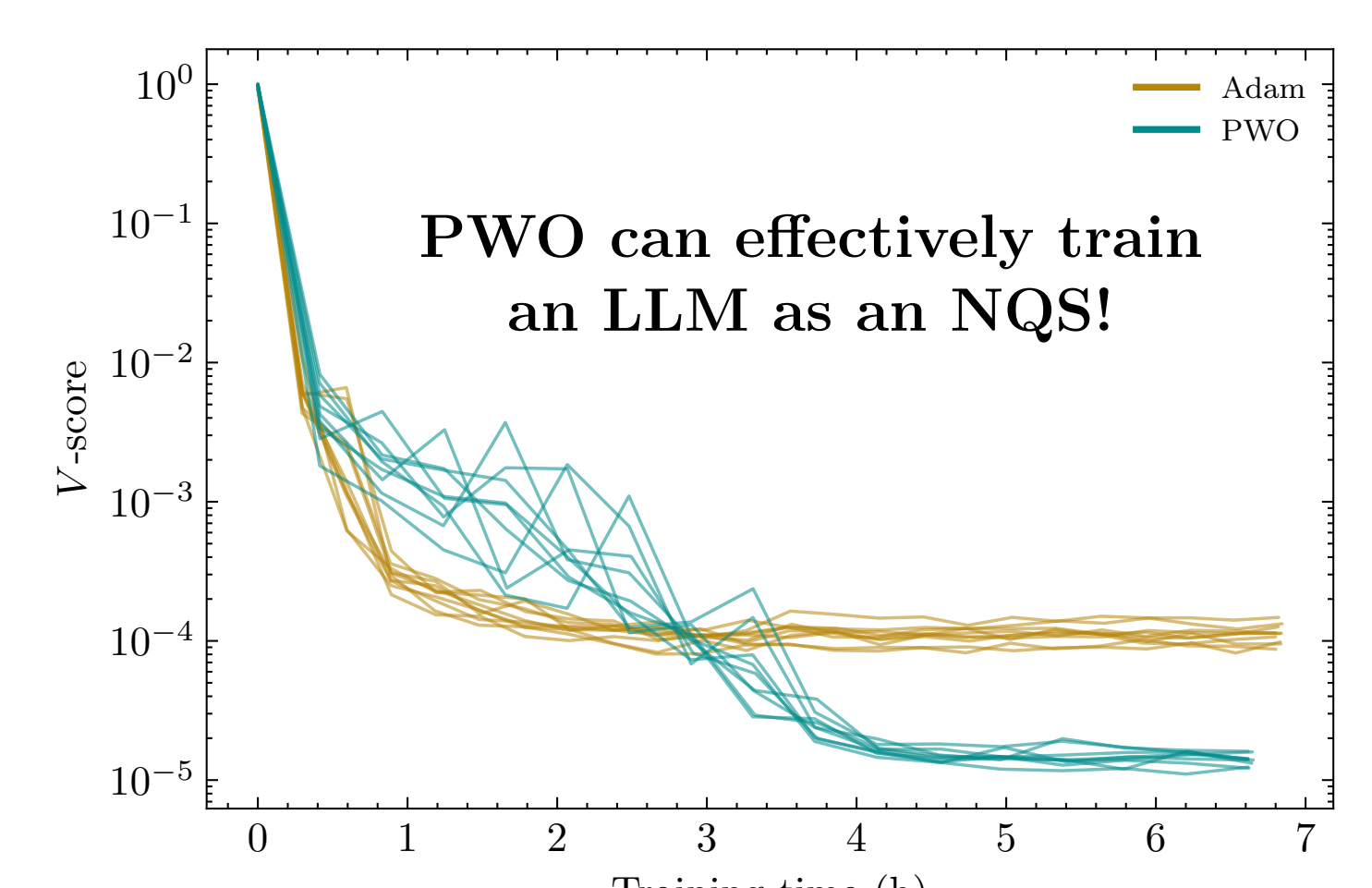


Figure 4. Training history for the 1.5 B parameter RWKV-7 LLM for the TFIM in a periodic chain of 12 spins at the critical point. This is three orders of magnitude bigger than existing NQS in the literature. Only 150 samples and one NVIDIA L40S were used.