# Stochastic Reconfiguration as Statistical Filtering for Overparameterized Neural Quantum States

**Tak Hur**
Department of Applied Mathematics
University of Waterloo
tak.hur@uwaterloo.ca

## Abstract

Stochastic reconfiguration (SR) is the standard optimizer for neural quantum states (NQS), but modern NQS often have far more parameters than Monte Carlo samples. We show that in this regime the diagonal shift is more than a numerical stabilizer. It acts as a statistical filter for finite-sample generalization. At a fixed wave function, SR is ridge regression from tangent features to the centered local energy. Its residual is the *expressivity gap*, the part of imaginary-time evolution outside the current tangent space. This gap is orthogonal to the tangent space in population, but finite batches make it act as noise that SR can overfit. The shift therefore balances shrinkage of useful update directions against variance from fitting sampled residuals. Exact diagnostics on a $4 \times 4$ Heisenberg graph separate two effects of overparameterization. Larger tangent spaces help when they reduce the expressivity gap, but they can hurt when they overfit a fixed gap. In a $100$-site transverse-field Ising family trained with a foundation NQS, validation risk is U-shaped in the shift while variance decreases, matching the noisy-ridge model. This view leads to multi-shift SR (MS-SR), which averages independent ridge solves at data-adaptive shifts to form a richer, lower-variance spectral filter. Checkpoint-local experiments show that MS-SR lowers validation risk and update variance relative to the fixed-shift SR baseline. We further compare MS-SR and SR in paired online training continuations, with independent endpoint energy evaluations and a separate update-cost benchmark. The code and data for this work is available at `https://github.com/takh04/sr_filter`.

## 1 Introduction

Stochastic Reconfiguration (SR) remains the dominant optimizer for neural quantum states (NQS), but the regime in which it operates has changed fundamentally. Classical variational Monte Carlo with Jastrow, Gutzwiller, and projected mean-field ansatzes, as well as early RBM-based NQS, had modest parameter counts [14, 31, 1, 28, 24, 25, 2]. Kernel-form SR (minSR) removed the need to form the full $P \times P$ quantum geometric tensor (QGT), enabling the shift to transformer-based and foundation-style wave functions with millions of parameters [4, 9, 17, 10, 30, 3, 19, 26, 12, 16, 18]. However, minSR still solves a dense $N_s \times N_s$ system built from the empirical NTK matrix, at cost $O(N_s^2 P)$ to form and $O(N_s^3)$ to solve. Increasing $N_s$ therefore remains the practical bottleneck, pushing modern NQS from the sample-rich regime of classical VMC ($P < N_s$) into a parameter-rich regime where $P$ exceeds $N_s$ by one to three orders of magnitude (Figure 1).

Each SR update solves a linear system with an empirical QGT, which can be ill-conditioned when estimated from finitely many Monte Carlo samples. The standard remedy is to add a diagonal shift before solving [24, 25, 2]. Historically, this shift was therefore understood mainly as a numerical conditioning device. In the modern minSR regime, however, the same operation is applied when the

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

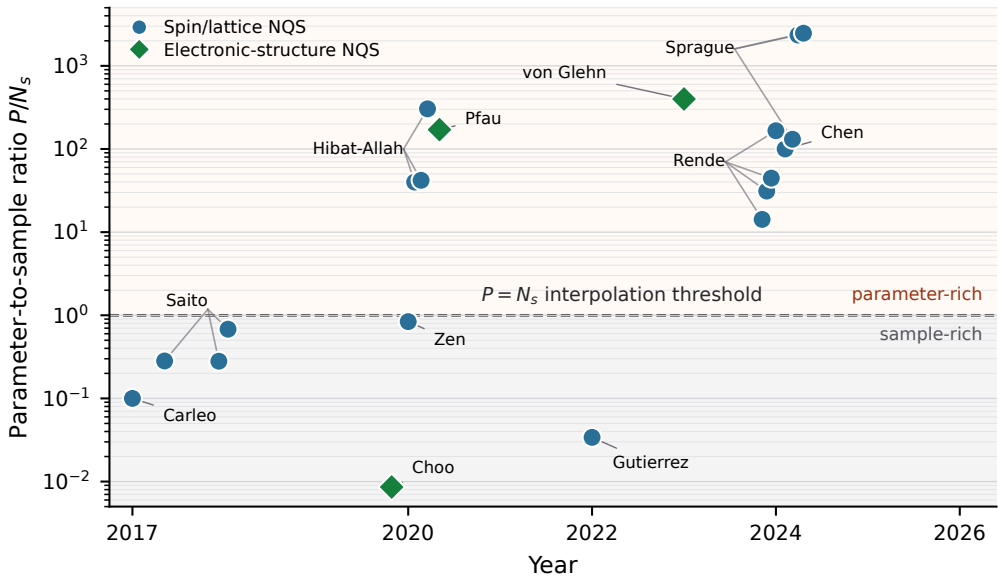



Figure 1: Historical shift in the parameter-to-sample ratio used by VMC/NQS calculations. The dashed line marks $P = N_s$. Marker shape denotes domain: circles are spin/lattice NQS and diamonds are electronic-structure NQS. Sources: [2, 20, 21, 4, 32, 10, 17, 8, 30, 19, 3, 26].

number of variational parameters can far exceed the number of samples, so its role may be statistical as well as numerical. This raises the central question of this paper: is the diagonal shift in SR merely a numerical stabilizer, or is it a statistical filter that makes overparameterized finite-sample updates generalize, and can this view lead to better filters than a single $\lambda$?

This question exposes the central tension of modern NQS optimization. More expressive neural quantum states can reduce the *expressivity gap*, the component of imaginary-time evolution outside the current tangent space. However, finite-sample SR can also overfit the part of this gap that remains. The central contributions of this work are:

- We formulate SR as a misspecified ridge-regression problem. The expressivity gap acts as finite-sample residual noise, yielding an exact held-out error identity and a spectral risk approximation for the bias–variance role of the diagonal shift $\lambda$.
- We test this mechanism in both exactly tractable small-scale and modern large-scale NQS settings. Exact $4 \times 4$ matched-noise and matched-state regression diagnostics separate finite-sample overfitting from expressivity-gap reduction. Large-scale shared-checkpoint diagnostics show the predicted U-shaped validation risk and monotone variance reduction as $\lambda$ increases.
- We use the filter view to design multi-shift SR (MS-SR), which combines independent ridge solves at multiple data-adaptive shifts. Checkpoint-local ablations isolate the roles of multi-shift spectral filtering, independent-batch averaging, and learned stacking. We further compare MS-SR and SR in paired online training continuations, with independent endpoint energy evaluations and measurements of update cost.

Section 2 develops the noisy ridge-regression view of SR. Sections 3 and 4 validate its predictions in exact small-scale and modern large-scale settings. Section 5 shows how this perspective leads to MS-SR, with checkpoint-local ablations and online comparisons against SR.

## 2 SR as Misspecified Ridge Regression

Throughout this section we fix a checkpoint $\theta$ and assume that the wavefunction and Hamiltonian are real in the sampling basis for notational simplicity. The complex case is obtained by replacing transposes with Hermitian conjugates (see Appendix A for detail).

**SR as tangent-space regression.** Let $\psi_\theta$ be a neural quantum state with Born distribution $\pi(x) = |\psi_\theta(x)|^2/\|\psi_\theta\|^2$. Define the centered tangent features and centered local energy by

$$O_c(x) = \nabla_\theta \log \psi_\theta(x) - \mathbb{E}_\pi[\nabla_\theta \log \psi_\theta], \qquad H_{\mathrm{loc},c}(x) = H_{\mathrm{loc}}(x) - E(\theta),$$

where $H_{\mathrm{loc}}(x) = \langle x|H|\psi_\theta\rangle/\langle x|\psi_\theta\rangle$ and $E(\theta) = \mathbb{E}_\pi[H_{\mathrm{loc}}]$. Under normalized imaginary-time evolution, the log amplitude changes as $\Delta \log \psi_\theta(x) = -\eta H_{\mathrm{loc},c}(x) + O(\eta^2)$. A parameter update $\theta^+ = \theta - \eta\delta$ changes the log amplitude as $\log \psi_{\theta^+}(x) - \log \psi_\theta(x) = -\eta O_c(x)^T\delta + O(\eta^2)$, up to an $x$-independent normalization term. Matching these first-order changes shows that fixed-checkpoint SR fits tangent predictions $O_c(x)^T\delta$ to the centered local energy $H_{\mathrm{loc},c}(x)$.

Writing $S^\dagger$ for the Moore–Penrose pseudoinverse of $S = \mathbb{E}_\pi[O_c(x)O_c(x)^T]$, and setting $g = \mathbb{E}_\pi[O_c(x)H_{\mathrm{loc},c}(x)]$, the ideal fixed-checkpoint SR direction is the population least-squares solution

$$\delta^* = S^\dagger g \in \arg\min_\delta \mathbb{E}_\pi\Big[\big(O_c(x)^T\delta - H_{\mathrm{loc},c}(x)\big)^2\Big]. \tag{1}$$

**Empirical SR as misspecified ridge regression.** The centered local energy need not be exactly representable by the current tangent features. We define the tangent-space expressivity gap $\epsilon(x) = H_{\mathrm{loc},c}(x) - O_c(x)^T\delta^*$. The normal equations imply $\mathbb{E}_\pi[\epsilon(x)O_c(x)] = 0$, and the gap variance is $\sigma^2_{\mathrm{gap}} = \mathbb{E}_\pi[\epsilon(x)^2] = \mathrm{Var}_\pi(H_{\mathrm{loc}}) - g^T S^\dagger g$. We call $\epsilon(x)$ the misspecification noise: it is the part of the imaginary-time target that lies outside the current tangent space.

Given a Monte Carlo batch $\mathcal{D} = \{x_j\}_{j=1}^{N_s}$, empirical SR solves the ridge problem

$$\hat\delta_\lambda = (\hat S + \lambda I)^{-1}\hat g, \qquad \hat S = \frac{1}{N_s}\sum_{j=1}^{N_s} O_c(x_j)O_c(x_j)^T, \qquad \hat g = \frac{1}{N_s}\sum_{j=1}^{N_s} O_c(x_j)H_{\mathrm{loc},c}(x_j).$$

Although $\epsilon(x)$ is orthogonal to the tangent features in population, it is still present in the sampled targets $H_{\mathrm{loc},c}(x_j)$. An overparameterized empirical solve can therefore memorize this residual on the batch, even though it does not correspond to a generalizable tangent-space update. In this sense, the expressivity gap behaves as regression noise.

**Excess risk analysis.** Here, we are interested in $S$-norm excess risk $\mathcal{E}_\lambda = \mathbb{E}_\mathcal{D}\|\hat\delta_\lambda - \delta^*\|_S^2$, where the expectation is over the Monte Carlo batch used to form $\hat\delta_\lambda$. This is the out-of-sample error of the checkpoint-local ridge problem with features $O_c(x)$, target $H_{\mathrm{loc},c}(x)$, and residual noise $\epsilon(x)$.

The exact bias–variance identity, together with the spectral proxy (see Appendix A for details), gives

$$\mathcal{E}_\lambda = \underbrace{\Big\|\mathbb{E}_\mathcal{D}[\hat\delta_\lambda] - \delta^*\Big\|_S^2}_{\text{Bias}} + \underbrace{\mathbb{E}_\mathcal{D}\Big\|\hat\delta_\lambda - \mathbb{E}_\mathcal{D}[\hat\delta_\lambda]\Big\|_S^2}_{\text{Variance}} \approx \underbrace{\sum_i s_i\left(\frac{\lambda}{s_i+\lambda}\right)^2(\beta_i^*)^2}_{\mathcal{B}_{\mathrm{shrink}}(\lambda)} + \underbrace{\frac{\sigma^2_{\mathrm{gap}}}{N_s}\sum_i\left(\frac{s_i}{s_i+\lambda}\right)^2}_{\mathcal{V}_{\mathrm{gap}}(\lambda)}, \tag{2}$$

where $S = V\,\mathrm{diag}(s_i)V^T$ and $\beta_i^* = (V^T\delta^*)_i$. The approximation replaces the empirical inverse by its population counterpart and uses a scalar residual-noise covariance. The first term is shrinkage bias: it grows when $\lambda$ damps useful components of $\delta^*$. The second term is finite-sample variance: it decreases because the same filter suppresses fits to sampled gap residuals. Thus $\lambda$ is a statistical regularizer, not only a numerical tolerance. Standard SR, however, is restricted to the one-parameter filter family $s/(s+\lambda)$, motivating the multi-$\lambda$ filters in Section 5.

This $S$-norm excess risk is physically relevant because it is closely related to the infidelity, a standard distance measure between quantum states. For bounded updates,

$$\mathcal{I}_\theta(\delta,\delta') = 1 - |\langle\psi_{\theta-\eta\delta}|\psi_{\theta-\eta\delta'}\rangle|^2 = \eta^2\,\|\delta-\delta'\|_S^2 + O(\eta^3). \tag{3}$$

Therefore $\mathcal{E}_\lambda$ controls the expected local infidelity between the ideal population update and the empirical SR update, up to the factor $\eta^2$ and higher-order terms.

**Large-system diagnostics for excess risk and variance.** For small systems, the population quantities $S$, $\delta^*$, $\sigma^2_{\mathrm{gap}}$, and $\|\hat\delta_\lambda - \delta^*\|_S^2$ can be computed by exact diagonalization. This is impossible

in the large scale regimes where modern NQS are tuned and compared. We therefore use two fixed-checkpoint diagnostics that require only held-out predictions.

The first diagnostic is a validation residual. Let $\mathcal{D}_{\mathrm{val}}$ be independent of the batches used to compute the SR updates. Since $H_{\mathrm{loc},c}(x) = O_c(x)^T\delta^* + \epsilon(x)$ with $\mathbb{E}_\pi[\epsilon(x)O_c(x)] = 0$, any fixed update $\delta$ satisfies

$$R_{\mathrm{val}}(\delta) = \frac{1}{|\mathcal{D}_{\mathrm{val}}|}\sum_{x\in\mathcal{D}_{\mathrm{val}}}\big(O_c(x)^T\delta - H_{\mathrm{loc},c}(x)\big)^2, \qquad \mathbb{E}_{\mathcal{D}_{\mathrm{val}}}\left[R_{\mathrm{val}}(\delta)\right] = \|\delta-\delta^*\|_S^2 + \sigma_{\mathrm{gap}}^2. \quad (4)$$

Thus, for independent SR updates $\hat\delta_{\lambda,1},\dots,\hat\delta_{\lambda,m}$ at the same checkpoint, we use $\widehat{\mathcal{E}}_{\mathrm{val}}(\lambda) = m^{-1}\sum_{j=1}^m R_{\mathrm{val}}(\hat\delta_{\lambda,j})$. Its expectation is $\mathcal{E}_\lambda + \sigma_{\mathrm{gap}}^2$, so it tracks excess risk plus the expressivity-gap.

The second diagnostic measures how much the update varies across independent training batches. With $\bar\delta_\lambda = m^{-1}\sum_j \hat\delta_{\lambda,j}$, define

$$\mathcal{V}_{\mathrm{mb}}(\lambda) = \frac{1}{m-1}\sum_{j=1}^m \frac{1}{|\mathcal{D}_{\mathrm{val}}|}\sum_{x\in\mathcal{D}_{\mathrm{val}}}\big(O_c(x)^T(\hat\delta_{\lambda,j}-\bar\delta_\lambda)\big)^2. \quad (5)$$

As $m$ and the validation batch grow, this converges to the variance component of the excess risk. Both diagnostics are practical at large scale because they only require held-out tangent-feature predictions. Since they do not require forming or solving a validation QGT or NTK system, they can be evaluated on validation batches much larger than the SR training batch with modest computational cost.

## 3 Small-scale Dissection through Regression

This section tests the regression mechanism of Section 2 through the real-amplitude component of complex neural quantum states. The experiment uses a $4\times 4$ spin-$1/2$ Heisenberg graph,

$$H = J_1\sum_{(i,j)\in\mathcal{B}_1}\mathbf{S}_i\cdot\mathbf{S}_j + J_2\sum_{(i,j)\in\mathcal{B}_2}\mathbf{S}_i\cdot\mathbf{S}_j, \qquad J_1 = 1, \quad J_2 = 0.5,$$

where $\mathbf{S}_i = \boldsymbol{\sigma}_i/2$. Gap and risk values below concern the real-amplitude channel, where phase contributions are excluded.

We compare an RBM with $P/N_s = 0.27$ and a ViT with $P/N_s = 3.28$, using $N_s = 4096$ and the near-ridgeless shift $\lambda = 10^{-9}$ to expose fitting of sampled residuals. The matched-noise and matched-state comparisons use 100 independent Born-resampled batches. Exact summation over all $2^{16}$ states gives the population reference and gap variance. We report the mean and standard error of the regression excess risk. Figure 2 visualizes the corresponding empirical and population-optimal amplitude tangent predictions.

The experiment separates two effects that are coupled. The matched-noise comparison keeps $\sigma_{\mathrm{gap}}^2$ approximately fixed and tests the finite-sample cost of a larger tangent space. The matched-state comparison keeps the represented wavefunction approximately fixed and tests the expressivity benefit of the larger tangent space. Full model and optimization details are given in Appendix B.

**Matched-noise comparison.** To test the finite-sample cost of overparameterization, we select independently trained RBM and ViT checkpoints with comparable expressivity-gap variance. In Figure 2 panel (a) and (b), both values of $\sigma_{\mathrm{gap}}^2$ are close to $0.8$. At $\lambda = 10^{-9}$, the RBM remains close to the population-optimal amplitude direction, with $\mathcal{E}_\lambda \approx 0.53 \pm 0.08$. The ViT has comparable misspecification noise, but its empirical solve fits the sampled residual much more strongly, increasing the excess-risk estimate to $\mathcal{E}_\lambda \approx 3.39 \pm 0.73$. This illustrates the regression-overfitting mechanism underlying Eq. (2). At comparable amplitude gap variance, the larger tangent space gives the empirical regression more directions in which to fit sample-specific noise.

**Matched-state comparison.** To isolate the benefit of a larger tangent space, we freeze a partially converged RBM checkpoint and fit a ViT to represent the same quantum state by minimizing infidelity [23, 7]. The two represented states have fidelity $F \geq 0.98$. Thus the wavefunction is nearly fixed, while the tangent space changes. Under this matched-state protocol, ViT's larger tangent

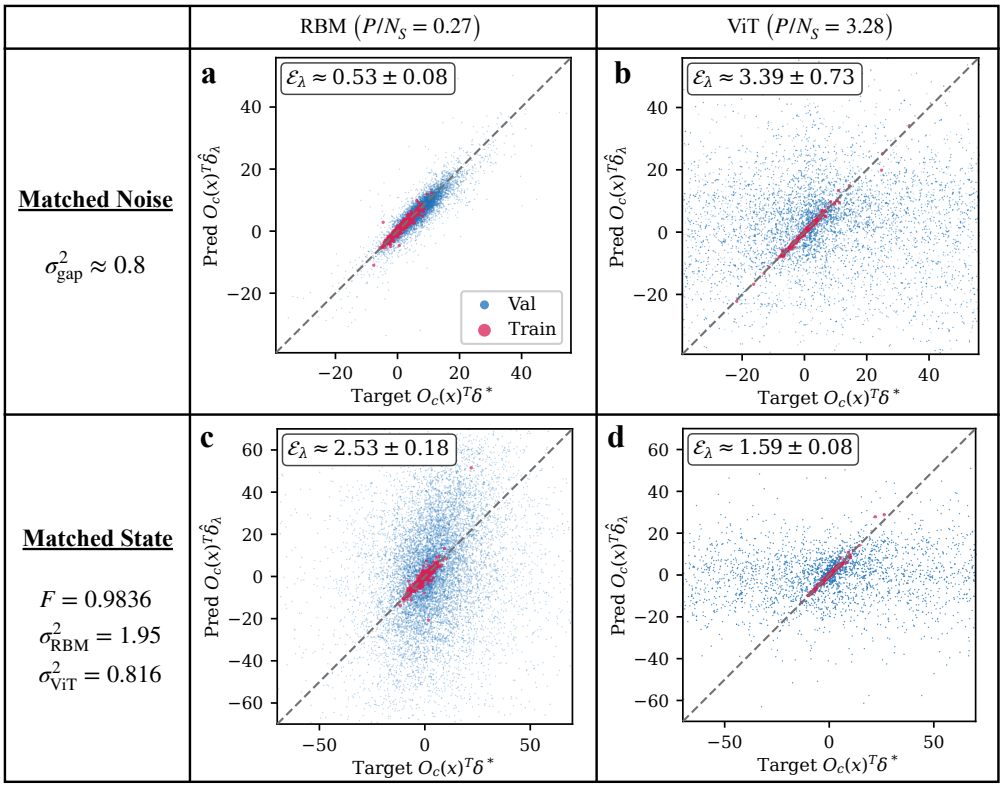



Figure 2: Exact regression diagnostics on the $4 \times 4$ Heisenberg graph specified in Appendix B, at $N_s = 4096$ and $\lambda = 10^{-9}$. Panels compare empirical and population-optimal amplitude tangent predictions. Gap variance and excess risk refer to this regression. Matched noise exposes ViT overfitting at comparable residual variance; matched state shows lower gap variance and excess risk for the ViT.

space lowers the amplitude gap variance from $\sigma^2_{\text{gap}} = 1.95$ to $0.816$. The excess-risk estimate also decreases from $\mathcal{E}_\lambda \approx 2.53 \pm 0.19$ to $\mathcal{E}_\lambda \approx 1.59 \pm 0.08$. Thus overparameterization can improve regression generalization when its additional tangent directions reduce enough misspecification noise.

Together, these comparisons illustrate the corresponding tension in regression: a larger tangent space can help when gap reduction outweighs extra finite-sample error, but can hurt when the solve overfits a comparable gap. This motivates the statistical-filtering interpretation of the diagonal shift $\lambda$.

## 4 Large-Scale Checkpoint Diagnostics of the Noisy-Ridge Mechanism

The exact $4 \times 4$ study isolates the noisy-ridge mechanism using population quantities. We next test whether the same mechanism remains visible in modern overparameterized NQS, where the population QGT, population SR direction, and expressivity-gap variance are unavailable. We therefore test the mechanism using the checkpoint-local diagnostics introduced in Section 2.

We use a foundation neural quantum state (FNQS) on the $L = 100$ periodic one-dimensional transverse-field Ising model (TFIM) family, $H = -h\sum_i X_i - \sum_i Z_i Z_{i+1}$, with $h \in [0.8, 1.2]$. A single conditional wave function $\log \psi_\theta(x; h)$ is trained across many Hamiltonian parameters, so the SR update must aggregate information over a family of Hamiltonians. The model has $P = 198{,}144$ parameters, while each SR step uses a batch of $N_s = 12{,}000$ samples, split over 6000 training $h$ values with two spin configurations per field. We use the same architectural and sampler setting as Rende et al. [18] (see Appendix C for experimental details).

The diagnostic experiment uses a shared-checkpoint protocol. We first take ten late-training checkpoints from the run trained with $\lambda_{\text{train}} = 10^{-4}$. These checkpoints are then held fixed. At each checkpoint, we perform an offline sweep over the solve-time shift $\lambda \in \{10^{-8}, 10^{-7}, \ldots, 10^{-1}\}$,

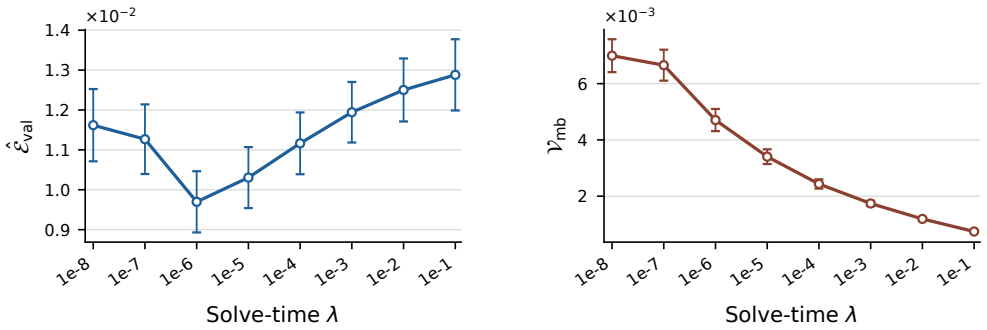



Figure 3: Large-scale shared-checkpoint test of the noisy-ridge mechanism on the periodic $L = 100$ transverse-field Ising model with $h$ ranging from $0.8$ to $1.2$ with foundational NQS. A fixed set of late-training checkpoints is reused while only the solve-time shift $\lambda$ is varied. $\widehat{\mathcal{E}}_{\mathrm{val}}$ is U-shaped, while $\mathcal{V}_{\mathrm{mb}}$ decreases monotonically with $\lambda$, matching the two predictions of Section 2.

recomputing standard SR updates on fresh resampled batches for each value of $\lambda$. This design keeps the wave function, tangent geometry, and expressivity-gap fixed while varying only the ridge penalty. It therefore tests the local bias–variance mechanism directly, not the different training trajectories.

For each checkpoint and each diagnostic shift, we construct $m = 100$ independent standard-SR updates from independently resampled training batches. The validation diagnostic uses $120{,}000$ held-out samples, consisting of $6000$ held-out $h$ values, disjoint from the training $h$ values, and $20$ spin configurations per $h$ value. Predictions are centered separately for each $h$ value, since each Hamiltonian parameter has its own local-energy distribution. The validation pass then estimates both $\widehat{\mathcal{E}}_{\mathrm{val}}$ and $\mathcal{V}_{\mathrm{mb}}$ without forming or solving any validation QGT system.

Figure 3 shows the fixed-checkpoint signature predicted by the noisy-ridge view. The $\hat{\mathcal{E}}_{\mathrm{val}}$ curve is U-shaped in $\lambda$, while $\mathcal{V}_{\mathrm{mb}}$ decreases as $\lambda$ increases. Since all shifts are evaluated on the same checkpoints, this trend reflects the solve-time ridge filter rather than different training trajectories. Small shifts leave useful directions nearly undamped but fit more batch-dependent noise. Large shifts reduce this variance but also shrink useful population directions.

## 5 Multi-Shift SR for Statistical Filtering and Variance Reduction

Motivated by the noisy-ridge view, we introduce multi-shift SR (MS-SR), an optimizer that replaces the single shifted SR solve with a mixture of shifted SRs computed on independent batches. Standard SR uses one ridge scale and therefore applies the same spectral filter $s/(s + \lambda)$ to every QGT eigendirection. MS-SR instead combines several regularization scales, while the independent batches decorrelate finite-sample errors.

At checkpoint $\theta_t$, MS-SR draws $K$ independent batches $\mathcal{D}_1, \ldots, \mathcal{D}_K$, chooses shifts $\lambda_k$ from empirical NTK-spectrum quantiles, and computes candidate updates $\hat{\delta}_k = \mathrm{minSR}(\mathcal{D}_k, \lambda_k)$. It then forms $\hat{\delta}_{\mathrm{MS}} = \sum_{k=1}^{K} w_k \hat{\delta}_k$, with $\mathbf{w} \in \Delta^{K-1}$. The weights are selected by leave-one-batch-out stacking, so each candidate update is evaluated only on batches that were not used for its ridge solve. The full procedure is given in Algorithm 1 in Appendix D.

This construction has a simple bias–variance interpretation. For fixed shifts and mixture weights, define $\bar{\delta}_k = \mathbb{E}_{\mathcal{D}_k}\left[\hat{\delta}_k\right]$. Then the excess risk decomposes exactly as

$$\mathbb{E}\left[\left\|\hat{\delta}_{\mathrm{MS}} - \delta^*\right\|_S^2\right] = \left\|\sum_{k=1}^{K} w_k \bar{\delta}_k - \delta^*\right\|_S^2 + \mathbb{E}\left[\left\|\sum_{k=1}^{K} w_k \left(\hat{\delta}_k - \bar{\delta}_k\right)\right\|_S^2\right]. \tag{6}$$

The first term is the bias of the average filter. The second term is the finite-batch variance of the average update. MS-SR targets both terms: multiple shifts enlarge the filter class, while independent batches reduce the variance of the noisy solves.

**Why multiple shifts help.** Consider the population spectral approximation from Section 2. If $S = V\,\mathrm{diag}(s_i)V^T$, then a single ridge solve applies the filter $f_\lambda(s) = s/(s+\lambda)$ to each eigendirection. A convex mixture of ridge solves applies instead $f_{\mathbf{w},\boldsymbol{\lambda}}(s) = \sum_k w_k s/(s+\lambda_k)$. Thus MS-SR is still a spectral filter, but it is no longer restricted to the single-shift family.

The difference is clearest at the two ends of the spectrum. For small $s$, $f_{\mathbf{w},\boldsymbol{\lambda}}(s) = s\sum_k w_k/\lambda_k + O(s^2)$. For large $s$, $f_{\mathbf{w},\boldsymbol{\lambda}}(s) = 1 - (\sum_k w_k\lambda_k)/s + O(s^{-2})$. With a single shift, the same $\lambda$ controls both the suppression of low-eigenvalue directions and the residual shrinkage of high-eigenvalue directions. On the other hand, these two behaviors can be tuned separately with multiple shifts. This can improve the aggregate bias–variance tradeoff when useful update directions and noise-sensitive directions are not controlled well by the same shift.

**Why independent batches help.** The second term in Eq. (6) depends on the correlations between candidate updates. If the candidates are computed on independent batches, the cross terms vanish and $\mathbb{E}\|\sum_k w_k(\hat{\delta}^{(k)}_{\lambda_k} - \bar{\delta}_{\lambda_k})\|_S^2 = \sum_k w_k^2\,\mathbb{E}\|\hat{\delta}^{(k)}_{\lambda_k} - \bar{\delta}_{\lambda_k}\|_S^2$. For uniform weights, this is a factor-$K$ reduction of the average single-batch variance in the ideal independent-batch limit. Within-batch autocorrelation and residual cross-batch covariance are discussed in Appendix A.6.

Independent batches produce a bagging-like effect by decorrelating the sampling noise across candidate updates. If all $K$ shifts are solved on the same batch, the resulting candidates can still enrich the spectral filter, but their sampling errors remain shared, so the independent-batch factor-$K$ reduction no longer applies. A single larger batch of size $KN_s$ would also reduce variance, but it is substantially more expensive computationally, with dense solve cost $O(K^3N_s^3)$. In contrast, $K$ independent solves at batch size $N_s$ cost only $O(KN_s^3)$ and can be run in parallel. This targets variance reduction without the cubic solve cost of one enlarged batch. At $K = 4$, the single-GPU TFIM benchmark gives MS-SR approximately four times the runtime of a single SR update and about $1\%$ overhead over bagged SR (Figure 5(c); Appendix F.1).

**Shift and weight selection.** MS-SR first chooses the shifts and then fits the mixture weights. In our $K = 4$ experiments, the shifts are the empirical NTK-spectrum quantiles $\{0.9, 0.7, 0.4, 0.1\}$. These quantiles place the ridge transition scale at several parts of the empirical spectrum, producing candidate updates with different bias–variance tradeoffs rather than relying on a single regularization.

Given these shifts, Algorithm 1 uses leave-one-batch-out stacking to choose the weights. Each candidate update is evaluated on batches that were not used for its ridge solve. Minimizing the residual between tangent-feature predictions and centered local energies gives a data-dependent cross-validation criterion.

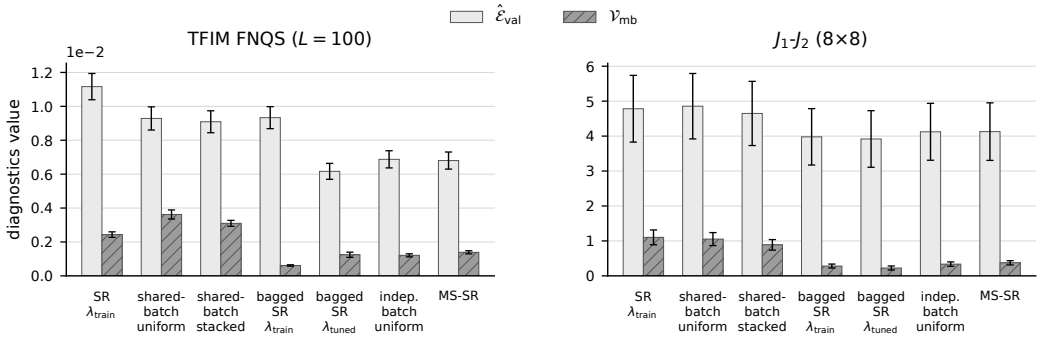



Figure 4: Checkpoint-local ablations for the ingredients of MS-SR at $K = 4$. Paired bars show raw $\hat{\mathcal{E}}_{\mathrm{val}}$ and $\mathcal{V}_{\mathrm{mb}}$ averaged over fixed source checkpoints with one standard error. Standard SR uses $\lambda_{\mathrm{train}} = 10^{-4}$. Both systems show bagged SR at $\lambda_{\mathrm{train}}$ and at $\lambda_{\mathrm{tuned}}$, selected per checkpoint by an eight-point bagged-SR residual sweep using the evaluation data. All multi-shift variants use the same NTK quantile grid.

**Ablation study of MS-SR.** Figure 4 examines three ingredients: independent-batch averaging, shift diversity, and learned weights. We compare validation residuals and update variance at fixed TFIM

FNQS and $8 \times 8$ $J_1$-$J_2$ ViT checkpoints ($J_2/J_1 = 0.5$). Standard SR provides the single-update reference; each ensemble combines four candidates:

| Method | Candidate fitting data | Shifts | Weights |
|---|---|---|---|
| Standard SR | One batch | Training shift | Single update |
| Bagged SR | Four independent batches | One common shift | Uniform |
| Shared-batch multi-shift SR | One shared batch | Four shifts | Uniform |
| Shared-batch stacked SR | One shared batch | Four shifts | Learned |
| Independent-batch multi-shift SR | Four independent batches | Four shifts | Uniform |
| MS-SR | Four independent batches | Four shifts | Learned |

Bagging tests independent averaging at one shift. Shared-batch variants mix shifts with shared sampling noise. Uniform and learned weights use identical candidates within each construction to test stacking. MS-SR combines all three ingredients. All multi-shift variants use the same NTK-quantile grid. Stacking fits weights on samples held out from the corresponding candidate solves.

Figure 4 shows bagging at $\lambda_{\mathrm{train}}$ and a checkpoint-specific $\lambda_{\mathrm{tuned}}$. The latter minimizes the bagged residual on the reporting data, making it an oracle reference. Appendix D gives the fitting and tuning protocols.

Relative to standard SR at the training shift, MS-SR reduces $\widehat{\mathcal{E}}_{\mathrm{val}}$ and $\mathcal{V}_{\mathrm{mb}}$ by $39\%/43\%$ on TFIM and $14\%/66\%$ on $J_1$-$J_2$. Bagging at $\lambda_{\mathrm{tuned}}$ reduces them by $45\%/49\%$ and $18\%/80\%$, respectively, yielding lower mean values than MS-SR in both systems. Shared-batch multi-shift SR reduces the TFIM residual by $17\%$, but increases the $J_1$-$J_2$ residual by $2\%$.

The TFIM comparison also illustrates the role of shift selection. Relative to bagging at $\lambda_{\mathrm{train}}$, MS-SR has $27\%$ lower residual, with the same ordering on all ten checkpoints. Bagging at that larger shift has lower variance but higher total residual, illustrating the cost of shrinkage bias. The fixed-shift sweep in Appendix E shows that bagging achieves a lower residual than MS-SR near $10^{-6}$–$10^{-7}$. Thus a well-chosen single shift can suffice, while MS-SR offers adaptation without a validation-based shift search.

On $J_1$-$J_2$, bagging at $\lambda_{\mathrm{train}}$ has mean residual $3.98$, compared with $3.92$ at $\lambda_{\mathrm{tuned}}$ and $4.13$ for MS-SR. Tuning reduces the mean residual by a further $1.5\%$ relative to training-shift bagging, selecting shifts from $10^{-4}$ to $10^{-1}$.

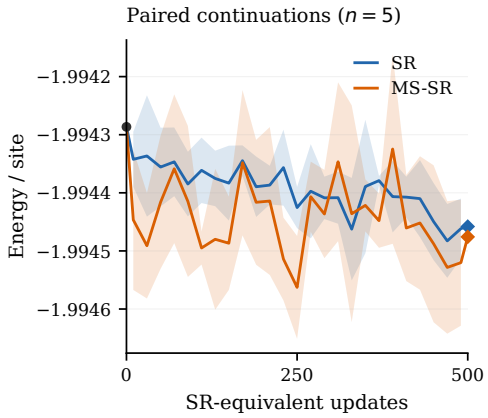
**Robustness and choice of $K$.** Appendix E reports $K \in \{2, 3, 4, 5\}$, two alternative quantile grids, and checkpoints trained at $10^{-3}$ and $10^{-5}$. MS-SR's TFIM residual decreases from $0.00854$ to $0.00638$ as $K$ grows, with diminishing gains and one extra batch and solve per candidate. At every $K$ it beats bagging at the training shift, while bagging at $\lambda_{\mathrm{tuned}}$ remains better. The alternative grids change the residual by at most $3.4\%$, and alternative training shifts give near-parity with that bagged baseline.

**Online training and computational cost.** With $K = 4$, MS-SR combines four candidate SR updates at each optimization step. Figure 5(a) therefore reports progress in nominal SR-equivalent updates: one standard-SR step counts as one unit and one MS-SR step as four. Under this convention, 125 MS-SR steps and 500 SR steps both reach 500 units. This axis counts candidate updates; implementation-dependent computational costs are assessed separately (Appendices C.4 and F.1).

We examine online behavior through a post hoc comparison of five paired continuations from a single shared SR checkpoint at iteration $2500$ on the $8 \times 8$ $J_1$–$J_2$ model, using a ViT ansatz without explicit lattice-symmetry symmetrization. The small expressivity gap early in training motivates using standard SR initially and reserving MS-SR's additional cost for later refinement. Both methods use the same batch size and learning rate. Full model, training, and evaluation details are given in Appendix C.4.

Independent endpoint evaluations give MS-SR lower energy in 3 of 5 pairs (Figure 5(b)), with mean $\Delta E/N = -1.78 \times 10^{-5}$ and paired $95\%$ Student-$t$ interval $[-7.05, 3.50] \times 10^{-5}$. The interval includes zero and summarizes variation across paired differences conditional on this source state. Additional online update-tracking diagnostics against a full-summation population SR reference are reported in Appendix G.

**(a) $J_1$–$J_2$ training loss history**




**(b) $J_1$–$J_2$ endpoint energies/site**

| Pair | SR | MS-SR | $\Delta$ ($10^{-5}$) |
|---|---|---|---|
| 1 | $\mathbf{-1.9944675}$ | $-1.9944307$ | $+3.69 \pm 1.09$ |
| 2 | $-1.9944450$ | $\mathbf{-1.9945160}$ | $-7.10 \pm 1.07$ |
| 3 | $\mathbf{-1.9944733}$ | $-1.9944722$ | $+0.11 \pm 1.11$ |
| 4 | $-1.9944593$ | $\mathbf{-1.9944668}$ | $-0.75 \pm 1.04$ |
| 5 | $-1.9944451$ | $\mathbf{-1.9944934}$ | $-4.83 \pm 1.02$ |
| Mean | $-1.9944580$ | $-1.9944758$ | $-1.78$ |

**(c) TFIM update-construction cost**

| Method | Batches/ solves | Time (s) | vs. SR |
|---|---|---|---|
| SR | 1/1 | 13.38 | 1.00× |
| Bagged SR | 4/4 | 53.55 | 4.00× |
| MS-SR | 4/4 | 54.09 | 4.04× |

Figure 5: Online performance results for MS-SR against SR. **(a)** Five paired continuations from a single shared SR checkpoint. Curves show means of binned traces with $\pm 1$ across-run sample standard deviation. The nominal horizontal axis counts candidate SR updates: one per SR step and four per MS-SR step. Diamonds at 500 mark independent endpoint means. **(b)** Endpoint energies for the same five paired continuations as in (a). Bold marks the lower energy. $\Delta = (E_{\text{MS-SR}} - E_{\text{SR}})/N$. Individual $\pm$ values are combined Monte Carlo standard errors in units of $10^{-5}$. **(c)** A separate TFIM benchmark on one H100 80GB GPU at $K = 4$, showing medians over three paired device/seed settings for the NTK-quantile MS-SR implementation. Protocols and computational accounting are given in Appendices F.1 and C.4.

## 6 Related Work

Neural quantum states have evolved rapidly from compact RBM-like ansätze to deep convolutional, recurrent, transformer-based, fermionic, and foundation-style wave functions [2, 10, 17, 30, 15, 19, 16, 18]. Autoregressive NQS provide a complementary route to scalability because their normalized factorization permits sampling from the model Born distribution rather than Markov-chain sampling [10]. This changes the sampling and autocorrelation structure of VMC, but it does not remove the finite-sample nature of SR, since QGTs and forces are still estimated from finite samples.

SR is the natural-gradient method of VMC, with the QGT playing the role of the Fisher metric on the variational manifold [24, 25, 2]. Recent sample-space formulations, including minSR, avoid forming the full $P \times P$ QGT and instead solve an $N_s \times N_s$ kernel system, making SR practical for highly overparameterized NQS [3, 19]. SPRING builds on this view with a Kaczmarz-inspired projected-increment update that reuses information across minibatches to accelerate convergence [5, 6]. These methods improve the cost and convergence of SR, while our focus is the statistical filtering role of the diagonal shift.

A parallel line of work approximates the QGT or Fisher metric structurally. KFAC uses Kronecker-factored curvature approximations and has been important in FermiNet-style electronic-structure VMC [13, 17]. Block-diagonal QGT methods partition the NQS metric by network layers, preserving intra-layer curvature while suppressing noisy cross-layer correlations [22]. These approaches change the metric approximation, but MS-SR keeps the sample-space SR geometry but replaces the single-shift ridge filter with a data-adaptive mixture of shifted solves.

Our analysis connects SR to random-design ridge regression [11]. In this view, the local-energy target is generally misspecified by the current tangent space, and the expressivity gap acts as the irreducible residual noise. Sketch-and-project and Kaczmarz interpretations clarify the structure of the overparameterized linear solve [27, 5, 6]. The present work identifies the statistical source of the noise being fitted and uses that view to design richer spectral filters for SR.

## 7 Conclusion

Modern NQS optimization has entered a finite-sample, parameter-rich regime. In this regime the diagonal shift in SR is not just a safeguard against an ill-conditioned empirical QGT. It is also the spectral regularizer that determines how much of the sampled tangent-space target is trusted. By recasting a single SR step as misspecified ridge regression, we identified the tangent-space expressivity gap as the effective noise source: it is orthogonal to the tangent features in population, but it is present in every finite batch and can be fitted by overparameterized empirical solves.

This viewpoint explains the apparent tension observed throughout the paper. Larger models can improve SR when gap reduction outweighs additional finite-sample error, but they can also amplify overfitting when a comparable gap remains. The exact small-system experiments and the large fixed-checkpoint SR diagnostics reveal this bias–variance structure across scales.

The filter perspective also leads to MS-SR, which mixes independent ridge solves at data-adaptive shifts. The checkpoint ablations separate the roles of these ingredients. Independent batches provide robust variance reduction, multiple shifts enlarge the filter class, and stacking adapts the mixture to the local checkpoint. Paired online continuations and independent endpoint energy evaluations extend the study to training trajectories, while a separate timing benchmark quantifies the cost of constructing the updates.

A natural next step is to investigate MS-SR in difficult systems and ask whether it allows models to reach states that standard optimizers miss. It is also important to quantify when the extra $K$-fold optimizer cost, from solving $K$ shifted SR, is worth paying. Such a study requires substantial compute and system-level engineering, and is a scope of its own.

Overall, overparameterization should not be avoided in NQS, but it should be treated with the right statistical regularization. Treating SR as a statistical filtering problem provides both a diagnostic language for understanding existing optimizers and a design principle for building more reliable ones.

## Acknowledgments and Disclosure of Funding

This research was supported by the education and training program of the Quantum Information Research Support Center, funded through the National Research Foundation of Korea (NRF) by the Ministry of Science and ICT (MSIT) of the Korean government (No. RS-2023-NR057243). The author thanks all members of the NQX Lab at École Polytechnique for their guidance and support in learning about neural quantum states during the visit. The experimental tools were developed on top of NetKet [29], an open-source neural-quantum-state and variational Monte Carlo library distributed under the Apache License 2.0. The author declares no competing interests.

## References

[1] Manuela Capello, Federico Becca, Michele Fabrizio, Sandro Sorella, and Erio Tosatti. Variational description of Mott insulators. *Physical Review Letters*, 94:026406, 2005. doi: 10.1103/PhysRevLett.94.026406.

[2] Giuseppe Carleo and Matthias Troyer. Solving the quantum many-body problem with artificial neural networks. *Science*, 355(6325):602–606, 2017.

[3] Ao Chen and Markus Heyl. Empowering deep neural quantum states through efficient optimization. *Nature Physics*, 20:1476–1481, 2024.

[4] Kenny Choo, Antonio Mezzacapo, and Giuseppe Carleo. Fermionic neural-network states for ab-initio electronic structure. *Nature Communications*, 11:2368, 2020.

[5] Gil Goldshlager, Nilin Abrahamsen, and Lin Lin. A kaczmarz-inspired approach to accelerate the optimization of neural network wavefunctions. *Journal of Computational Physics*, 511: 113351, 2024.

[6] Gil Goldshlager, Jiang Hu, and Lin Lin. A sketch-and-project analysis of subsampled natural gradient algorithms. *arXiv preprint arXiv:2508.21022*, 2025.

[7] Luca Gravina, Vincenzo Savona, and Filippo Vicentini. Neural projected quantum dynamics: a systematic study. *Quantum*, 9:1803, 2025. doi: 10.22331/q-2025-07-22-1803.

[8] Iván L. Gutiérrez and Christian B. Mendl. Real time evolution with neural-network quantum states. *Quantum*, 6:627, 2022. doi: 10.22331/q-2022-01-20-627.

[9] Jan Hermann, Zeno Schätzle, and Frank Noé. Deep-neural-network solution of the electronic Schrödinger equation. *Nature Chemistry*, 12:891–897, 2020.

[10] Mohamed Hibat-Allah, Martin Ganahl, Lauren E. Hayward, Roger G. Melko, and Juan Carrasquilla. Recurrent neural network wave functions. *Physical Review Research*, 2:023358, 2020.

[11] Daniel Hsu, Sham M Kakade, and Tong Zhang. Random design analysis of ridge regression. *Foundations of Computational Mathematics*, 14:569–600, 2014.

[12] Jane Kim, Gabriel Pescia, Bryce Fore, Jannes Nys, Giuseppe Carleo, Stefano Gandolfi, Morten Hjorth-Jensen, Alessandro Lovato, et al. Neural-network quantum states for ultra-cold Fermi gases. *Communications Physics*, 7:148, 2024.

[13] James Martens and Roger Grosse. Optimizing neural networks with kronecker-factored approximate curvature. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 2408–2417. PMLR, 2015. URL `https://proceedings.mlr.press/v37/martens15.html`.

[14] W. L. McMillan. Ground state of liquid $He^4$. *Physical Review*, 138:A442–A451, 1965. doi: 10.1103/PhysRev.138.A442.

[15] 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.

[16] Rajah P Nutakki, Shokry Ahmedov, and Filippo Vicentini. Design principles of deep translationally-symmetric neural quantum states for frustrated magnets. *Physical Review Research*, 7:043099, 2025.

[17] 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*, 2:033429, 2020.

[18] 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*, 16:7213, 2025.

[19] Riccardo Loris Rende, Lorenzo Luciano Viteritti, Lorenzo Bardone, Federico Becca, and Sebastian Goldt. A simple linear algebra identity to optimize large-scale neural network quantum states. *Communications Physics*, 7:260, 2024.

[20] Hiroki Saito. Solving the Bose-Hubbard model with machine learning. *Journal of the Physical Society of Japan*, 86:093001, 2017. doi: 10.7566/JPSJ.86.093001.

[21] Hiroki Saito and Masaya Kato. Machine learning technique to find quantum many-body ground states of bosons on a lattice. *Journal of the Physical Society of Japan*, 87:014001, 2018. doi: 10.7566/JPSJ.87.014001.

[22] 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.

[23] Alessandro Sinibaldi, Clemens Giuliani, Giuseppe Carleo, and Filippo Vicentini. Unbiasing time-dependent variational monte carlo by projected quantum evolution. *Quantum*, 7:1131, 2023.

[24] Sandro Sorella. Green function Monte Carlo with stochastic reconfiguration. *Physical Review Letters*, 80(20):4558, 1998.

[25] Sandro Sorella. Generalized Lanczos algorithm for variational quantum Monte Carlo. *Physical Review B*, 64(2):024512, 2001.

[26] Kyle Sprague and Stefanie Czischek. Variational Monte Carlo with large patched transformers. *Communications Physics*, 7:90, 2024.

[27] Thomas Strohmer and Roman Vershynin. A randomized Kaczmarz algorithm with exponential convergence. *Journal of Fourier Analysis and Applications*, 15:262–278, 2009.

[28] Daisuke Tahara and Masatoshi Imada. Variational Monte Carlo method combined with quantum-number projection and multi-variable optimization. *Journal of the Physical Society of Japan*, 77:114701, 2008. doi: 10.1143/JPSJ.77.114701.

[29] Filippo Vicentini, Damian Hofmann, Attila Szabó, Dian Wu, Christopher Roth, Clemens Giuliani, Gabriel Pescia, Jannes Nys, Vladimir Vargas-Calderón, Nikita Astrakhantsev, and Giuseppe Carleo. NetKet 3: Machine learning toolbox for many-body quantum systems. *SciPost Physics Codebases*, page 7, 2022. doi: 10.21468/SciPostPhysCodeb.7.

[30] Ingrid von Glehn, James S. Spencer, and David Pfau. A self-attention ansatz for ab-initio quantum chemistry. *arXiv preprint arXiv:2211.13672*, 2023.

[31] Hisatoshi Yokoyama and Hiroyuki Shiba. Variational Monte-Carlo studies of Hubbard model. I. *Journal of the Physical Society of Japan*, 56(4):1490–1506, 1987. doi: 10.1143/JPSJ.56.1490.

[32] Remmy Zen, Long My, Ryan Tan, Frédéric Hébert, Mario Gattobigio, Christian Miniatura, Dario Poletti, and Stéphane Bressan. Finding quantum critical points with neural-network quantum states. In *ECAI 2020: 24th European Conference on Artificial Intelligence*, volume 325, pages 1962–1969, 2020. doi: 10.3233/FAIA200315.

# A Fixed-Checkpoint SR Regression Identities

This appendix gives the derivations behind Section 2. All statements are checkpoint-local: the wavefunction $\psi_\theta$, sampling distribution $\pi$, QGT, and tangent features are held fixed while only the Monte Carlo batch used to estimate the SR direction varies. We write $\|u\|_S^2 = u^T S u$. Since $S$ may be singular, null QGT directions do not change tangent predictions and have zero $S$-seminorm. Throughout, $^\dagger$ denotes the Moore–Penrose pseudoinverse; Hermitian adjoints are written explicitly as $\overline{u}^T$.

## A.1 Real and complex conventions

The main text assumes real wavefunctions and real tangent features in the sampling basis. Then

$$S = \mathbb{E}_\pi[O_c(x) O_c(x)^T], \qquad g = \mathbb{E}_\pi[O_c(x) H_{\mathrm{loc},c}(x)].$$

The fixed-checkpoint population problem is the ordinary least-squares problem

$$\min_\delta \mathbb{E}_\pi\left[\left(O_c(x)^T\delta - H_{\mathrm{loc},c}(x)\right)^2\right],$$

and its minimum-norm solution is $\delta^* = S^\dagger g$.

For complex wavefunctions, let

$$O_c(x) = \nabla_\theta \log\psi_\theta(x) - \mathbb{E}_\pi[\nabla_\theta \log\psi_\theta], \qquad H_{\mathrm{loc},c}(x) = H_{\mathrm{loc}}(x) - \mathbb{E}_\pi H_{\mathrm{loc}}.$$

The Hermitian QGT and force are

$$Q = \mathbb{E}_\pi[\overline{O_c(x)}\, O_c(x)^T], \qquad f = \mathbb{E}_\pi[\overline{O_c(x)}\, H_{\mathrm{loc},c}(x)].$$

For complex parameter increments $\delta \in \mathbb{C}^p$, the regression loss is $\mathbb{E}_\pi|O_c(x)^T\delta - H_{\mathrm{loc},c}(x)|^2$, the normal equations are $Q\delta = f$, and the corresponding seminorm is $\overline{\delta}^T Q\delta = \mathbb{E}_\pi|O_c(x)^T\delta|^2$. For real parameter increments in a complex wavefunction, minimizing the same modulus-squared loss over $\delta \in \mathbb{R}^p$ gives $\mathrm{Re}[Q\,\delta] = \mathrm{Re}[f]$. Equivalently, one stacks the real and imaginary residuals and solves an ordinary real least-squares problem. Thus the real-parameter SR equation uses $S = \mathrm{Re}[Q]$ and $g = \mathrm{Re}[f]$. The derivations below are written in the real convention. In the complex convention, replace scalar squares by modulus squares and replace the $S$-seminorm by the corresponding real or Hermitian QGT seminorm.

### A.2 Validation and variance identities

Let $\delta^*$ be the population least-squares SR direction and define

$$\epsilon(x) = H_{\mathrm{loc},c}(x) - O_c(x)^T \delta^*.$$

The population normal equations give $\mathbb{E}_\pi[O_c(x)\epsilon(x)] = 0$. Hence, for any fixed direction $\delta$,

$$\begin{aligned}\mathbb{E}_\pi\left[\left(O_c(x)^T\delta - H_{\mathrm{loc},c}(x)\right)^2\right] &= \mathbb{E}_\pi\left[\left(O_c(x)^T(\delta-\delta^*) - \epsilon(x)\right)^2\right] \\ &= \|\delta-\delta^*\|_S^2 + \sigma_{\mathrm{gap}}^2,\end{aligned}$$

where $\sigma_{\mathrm{gap}}^2 = \mathbb{E}_\pi[\epsilon(x)^2]$. Thus an independent held-out residual estimates the $S$-excess risk up to the constant expressivity-gap term.

For a random diagnostic update $Z = \hat{\delta}_\lambda$ computed by resampling an SR batch at the same checkpoint, let $\mu = \mathbb{E}_{\mathcal{D}} Z$. Adding and subtracting $\mu$ gives

$$\mathcal{E}_\lambda = \mathbb{E}_{\mathcal{D}} \|Z - \delta^*\|_S^2 = \|\mu - \delta^*\|_S^2 + \mathbb{E}_{\mathcal{D}} \|Z - \mu\|_S^2 .$$

The cross term vanishes because $\mathbb{E}_{\mathcal{D}}[Z - \mu] = 0$. This is the usual squared-bias plus variance decomposition, with distances measured in the checkpoint QGT seminorm.

Finally, let $Z_1, \dots, Z_m$ be independent copies of $Z$ and $\bar{Z} = m^{-1}\sum_j Z_j$. The standard sample-variance identity holds in the same seminorm:

$$\mathbb{E}\left[\frac{1}{m-1}\sum_{j=1}^m \left\|Z_j - \bar{Z}\right\|_S^2\right] = \mathbb{E}\,\|Z-\mu\|_S^2 .$$

Equivalently,

$$\frac{1}{m-1}\sum_{j=1}^m \left\|Z_j - \bar{Z}\right\|_S^2 = \frac{1}{\binom{m}{2}}\sum_{a<b}\frac{1}{2}\|Z_a - Z_b\|_S^2 .$$

Replacing each $S$-seminorm prediction norm by its empirical held-out average $|\mathcal{D}_{\mathrm{val}}|^{-1}\sum_x (O_c(x)^T\cdot)^2$ gives the multi-batch diagnostic used in the main text.

### A.3 Spectral filter proxy

On the non-null QGT subspace, write $S = V\,\mathrm{diag}(s_i)V^T$ with $s_i > 0$, and let $\beta_i^* = (V^T\delta^*)_i$. The population ridge direction is

$$\delta_\lambda = (S+\lambda I)^{-1} g = (S+\lambda I)^{-1} S\delta^*.$$

Therefore

$$\delta_\lambda - \delta^* = -\lambda(S+\lambda I)^{-1}\delta^*,$$

and the exact population shrinkage bias is

$$\|\delta_\lambda - \delta^*\|_S^2 = \sum_i s_i \left(\frac{\lambda}{s_i+\lambda}\right)^2 (\beta_i^*)^2.$$

The scalar variance term in (2) is a reference calculation, not an assumption needed by the diagnostics. To isolate the residual-noise contribution, write

$$\hat{g} = \hat{S}\delta^* + \hat{\xi}, \qquad \hat{\xi} = \frac{1}{N_s}\sum_{j=1}^{N_s} O_c(x_j)\epsilon(x_j), \qquad \mathbb{E}[\hat{\xi}] = 0.$$

If one ignores the covariance-estimation fluctuation in $\hat{S}$ and replaces the empirical inverse by $A_\lambda = (S+\lambda I)^{-1}$, then

$$\hat{\delta}_\lambda - \delta_\lambda \approx A_\lambda \hat{\xi}.$$

Let $\Sigma_\epsilon = \mathbb{E}_\pi[O_c(x)O_c(x)^T\epsilon(x)^2]$. The force-noise proxy is then

$$\mathbb{E}_{\mathcal{D}}\left\|\hat{\delta}_\lambda - \delta_\lambda\right\|_S^2 \approx \frac{1}{N_s}\,\mathrm{tr}[SA_\lambda\Sigma_\epsilon A_\lambda] .$$

Under the scalar, homoscedastic approximation $\Sigma_\epsilon \approx \sigma_{\mathrm{gap}}^2 S$, this becomes

$$\mathbb{E}_{\mathcal{D}}\left\|\hat{\delta}_\lambda - \delta_\lambda\right\|_S^2 \approx \frac{\sigma_{\mathrm{gap}}^2}{N_s}\,\mathrm{tr}[A_\lambda S A_\lambda S] = \frac{\sigma_{\mathrm{gap}}^2}{N_s}\sum_i \left(\frac{s_i}{s_i+\lambda}\right)^2 .$$

### A.4 Local Fubini–Study metric

Let $|\Psi(\theta)\rangle$ denote the normalized variational state and let $P_\perp = I - |\Psi(\theta)\rangle\langle\Psi(\theta)|$. The real QGT is the pullback of the Fubini–Study metric,

$$S_{ij} = \mathrm{Re}\,\langle P_\perp \partial_i \Psi(\theta), P_\perp \partial_j \Psi(\theta)\rangle\,.$$

In the real sampling-basis convention, this equals $\mathbb{E}_\pi[O_{c,i}(x)O_{c,j}(x)]$. A Taylor expansion of the normalized state therefore gives, for nearby displacements $a$ and $b$,

$$1 - |\langle\Psi(\theta+a) \mid \Psi(\theta+b)\rangle|^2 = (a-b)^T S (a-b) + O\big((\|a\| + \|b\|)^3\big)\,.$$

Taking $a = -\eta\delta$ and $b = -\eta\delta'$ gives

$$\mathcal{I}_\theta(\delta, \delta') = \eta^2(\delta - \delta')^T S (\delta - \delta') + O(\eta^3)$$

for bounded update directions. Hence the $S$-norm excess risk controls the expected local infidelity between the population SR update and the empirical SR update, up to the factor $\eta^2$ and higher-order terms.

### A.5 Sample centering

Practical SR uses sample-centered log-derivatives and sample-centered local energies. This is exactly the ridge-regression problem with an unpenalized intercept,

$$\min_{a,\delta} \frac{1}{n}\sum_{j=1}^{n} \big(a + O(x_j)^T\delta - H_{\mathrm{loc}}(x_j)\big)^2 + \lambda\|\delta\|_2^2.$$

For fixed $\delta$, the optimal intercept is $a = \bar{H}_{\mathrm{loc}} - \bar{O}^T\delta$. Substitution gives

$$\frac{1}{n}\sum_{j=1}^{n} \big((O(x_j) - \bar{O})^T\delta - (H_{\mathrm{loc}}(x_j) - \bar{H}_{\mathrm{loc}})\big)^2 + \lambda\|\delta\|_2^2.$$

Thus sample centering is simply the intercept form of the same regression. In population, the intercept removes only the constant energy/normalization component. It does not remove the expressivity gap, because

$$\epsilon(x) = H_{\mathrm{loc},c}(x) - O_c(x)^T\delta^*$$

has $\mathbb{E}_\pi[\epsilon(x)] = 0$ and $\mathbb{E}_\pi[O_c(x)\epsilon(x)] = 0$. A finite-sample intercept also removes the sample mean of the residual, replacing $\epsilon_j$ by $\epsilon_j - \bar{\epsilon}$, but it cannot fit any $x$-dependent part of the gap. These centering terms are lower-order finite-sample fluctuations and do not change the checkpoint-local bias–variance identities.

### A.6 Correlated Monte Carlo samples

The population regression identities do not require independent samples within an SR batch. Autocorrelation does affect the variance of the empirical force. For a stationary chain, define $\zeta_t = O_c(x_t)\epsilon(x_t)$ and $\Gamma_\ell = \mathrm{Cov}(\zeta_0, \zeta_\ell)$. If these covariances are summable, the long-run force-noise covariance is

$$\Omega_\epsilon = \Gamma_0 + \sum_{\ell=1}^{\infty}(\Gamma_\ell + \Gamma_\ell^T), \qquad \mathrm{Cov}(\hat{\xi}) = \frac{\Omega_\epsilon}{N_s} + o(N_s^{-1}).$$

Under the same fixed-inverse approximation as the spectral proxy, its variance term therefore becomes $N_s^{-1}\,\mathrm{tr}[SA_\lambda\Omega_\epsilon A_\lambda]$. Replacing $N_s$ by an effective sample size is a useful scalar approximation when autocorrelation inflates the relevant directions similarly; a single scalar generally cannot represent the full matrix $\Omega_\epsilon$. The large-scale diagnostics use MCMC samples and thus already include the within-batch correlation present in those runs.

Cross-batch dependence is a separate issue. Independent sampler streams are used for the candidate batches and held-out evaluation. For independent, identically distributed updates at a fixed shift, uniform averaging reduces the QGT-norm variance by exactly $K$ even if samples within each batch are correlated. Correlated candidate batches introduce cross-covariance terms. Moreover, shifts estimated from one candidate batch and weights learned from the candidates are data dependent; the exact factor-$K$ statement does not automatically apply to the complete adaptive MS-SR procedure. Its variance reduction is assessed by the multi-batch diagnostic.

## B Experimental Details: Small-Scale Experiments

**System and bond convention.** The experiment in Figure 2 uses 16 spin-$1/2$ sites labeled $i = 4x+y$, with $x, y \in \{0, 1, 2, 3\}$. Nearest-neighbor bonds are periodic in both directions. Diagonal bonds join $(x, y)$ to $(x + 1, y \pm 1 \bmod 4)$ only for $x = 0, 1, 2$. There are 32 nearest-neighbor and 24 diagonal bonds, with couplings $J_1 = 1$ and $J_2 = 0.5$ multiplying $\mathbf{S}_i \cdot \mathbf{S}_j$, where $\mathbf{S}_i = \boldsymbol{\sigma}_i/2$. The historical bond constructor omitted the eight diagonal bonds across the $x = 3$ to $x = 0$ seam. The reported measurements retain that graph rather than the fully periodic $J_1$-$J_2$ lattice. No magnetization sector is imposed: all $2^{16} = 65{,}536$ states enter the exact Born distribution and population sums. The archived graph is exposed explicitly as `legacy_nnn_open_x` in the code; a separate `periodic` option restores the missing bonds for new experiments.

**Regression convention.** The complex wavefunction is retained when computing $H_{\mathrm{loc}}(x) = (H\psi_\theta)(x)/\psi_\theta(x)$ and the Born distribution. The offline diagnostic then uses only $A(x) = \nabla_\theta \operatorname{Re} \log \psi_\theta(x)$ and $y(x) = \operatorname{Re} H_{\mathrm{loc}}(x)$, with real parameter coordinates. Writing $A_c$ and $y_c$ for their Born-centered versions, it computes

$$S_{\mathrm{A}} = \mathbb{E}_\pi[A_c A_c^T], \qquad g_{\mathrm{A}} = \mathbb{E}_\pi[A_c y_c], \qquad \delta^*_{\mathrm{A}} = S^\dagger_{\mathrm{A}} g_{\mathrm{A}}, \qquad \sigma^2_{\mathrm{gap,A}} = \mathbb{E}_\pi[(y_c - A_c^T \delta^*_{\mathrm{A}})^2].$$

For complex states, the full real-parameter QGT additionally contains $\mathrm{Cov}_\pi(\nabla_\theta \operatorname{Im} \log \psi_\theta)$. That contribution and the imaginary local-energy target are omitted here. Thus these measurements are regression excess risks, not full complex-SR errors or the complete local infidelity in Eq. (3). The parameter directions may still change the wavefunction phase; the diagnostic simply does not score that output. The gap and risk labels in Figure 2 refer to this restricted regression, for which the least-squares decomposition is exact.

**Models.** The underparameterized model is a complex-valued RBM with hidden-to-visible ratio $\alpha = 2$, giving $P = 1120$ real parameters and $P/N_s = 0.273$ at the diagnostic sample size $N_s = 4096$. The overparameterized model is a complex-valued ViT with depth 2, embedding dimension 24, 4 attention heads, MLP expansion factor 4, and $2 \times 2$ patches, giving $P = 13{,}448$ real parameters and $P/N_s = 3.283$. Parameter counts are real counts, obtained by unrolling each complex parameter into real and imaginary parts.

**Training and checkpoint diagnostics.** The source checkpoints are produced by VMC with ordinary complex SR, 4096 Metropolis-local samples per optimization step, learning rate $10^{-3}$, and training diagonal shift $10^{-3}$. Exact regression summaries are saved every 10 optimizer steps. For each diagnostic checkpoint we store the amplitude gap and target variances, the population-optimal amplitude direction, and the full-Hilbert-space arrays needed for the offline regression.

**Matched-noise protocol.** For the matched-noise comparison, the RBM and ViT are trained independently and checkpoint pairs are selected by matching the amplitude gap variance $\sigma^2_{\mathrm{gap,A}}$. The pair used in Figure 2 is RBM step 490 and ViT step 10, with $\sigma^2_{\mathrm{gap}} = 0.8088$ and $0.7957$ respectively. At these fixed checkpoints we draw 100 independent training batches of size $N_s = 4096$ from the exact Born distribution and solve the ridge regression at $\lambda = 10^{-9}$.

**Matched-state protocol.** For the matched-state comparison, we first choose a partially converged RBM checkpoint and then fit a ViT to the same wavefunction by minimizing the exact Fubini–Study infidelity

$$\mathcal{I}(\theta_{\mathrm{ViT}}) = 1 - \frac{\left|\sum_x \psi^*_{\mathrm{RBM}}(x)\psi_{\mathrm{ViT}}(x;\theta_{\mathrm{ViT}})\right|^2}{\left(\sum_x |\psi_{\mathrm{RBM}}(x)|^2\right)\left(\sum_x |\psi_{\mathrm{ViT}}(x;\theta_{\mathrm{ViT}})|^2\right)}. \tag{7}$$

All sums run over the $2^{16}$ basis states, so the objective and diagnostics do not use stochastic overlap estimates. The pair used in Figure 2 has fidelity $F = 0.983626$ between the two represented complex states. Its exact amplitude gap variances are $\sigma^2_{\mathrm{gap}} = 1.9469$ for the RBM and $0.8161$ for the ViT. We use 10 independent Born-resampled training batches of size 4096 for this matched-state regression diagnostic.

**One-step regression metrics.** For each resampled training batch $\mathcal{D}_b$, the excess-risk estimate uses the exact amplitude covariance:

$$\widehat{\mathcal{E}}_{\lambda,\mathrm{A}} = \frac{1}{B}\sum_{b=1}^{B} \|\hat{\delta}_{\mathrm{A},\lambda}(\mathcal{D}_b) - \delta^*_{\mathrm{A}}\|^2_{S_{\mathrm{A}}}.$$

We also compute the training-batch residual and the population real-local-energy regression residual, which adds the constant amplitude gap variance $\sigma^2_{\mathrm{gap,A}}$. Figure 2 suppresses the A subscripts.

# C Experimental Details: Large-Scale Experiments

This appendix gives the reproducibility details for the two large-scale systems used in Sections 4 and 5: the $L = 100$ TFIM foundation NQS and the $8 \times 8$ $J_1$-$J_2$ ViT.

## C.1 FNQS on TFIM

**Hamiltonian.** The periodic one-dimensional transverse-field Ising model is

$$H(h) = -h\sum_{i=1}^{L} X_i - \sum_{i=1}^{L} Z_i Z_{i+1}, \qquad L = 100,$$

with transverse field $h \in [0.8, 1.2]$. This interval straddles the critical point $h = 1$.

**Foundation NQS objective.** Following Rende et al. [18], we train a conditional wavefunction $\log \psi_\theta(x; h)$ rather than a separate model for each field value. The field $h$ is supplied as a coupling token to the ViT. Training minimizes the field-averaged variational energy

$$\mathcal{L}(\theta) = \mathbb{E}_{h\sim p(h)}\left[\frac{\langle\psi_\theta(\cdot;h)|H(h)|\psi_\theta(\cdot;h)\rangle}{\langle\psi_\theta(\cdot;h)|\psi_\theta(\cdot;h)\rangle}\right], \tag{8}$$

where $p(h)$ is uniform over $R = 6000$ training fields in $[0.8, 1.2]$.

**Model and optimizer.** The model is a real-valued ViTFNQS with 6 layers, $d_{\mathrm{model}} = 72$, 12 attention heads, patch size $b = 4$ (effective sequence length 25), one coupling input (`n_coups=1`), and translational invariance. It has $P = 198{,}144$ trainable parameters. Each SR step uses $N_s = 12{,}000$ samples, organized as 6000 field replicas with two spin configurations per field. Spin configurations are drawn with a local Metropolis rule and sweep size $L$. The optimizer is SGD with fixed learning rate $\eta = 0.005$ and minSR/NTK preconditioning using the `pinv` linear solver.

**Training and checkpoints.** The source run used for the shared-checkpoint diagnostic protocol is trained with $\lambda_{\mathrm{train}} = 10^{-4}$ for 2000 steps. Checkpoints are saved every 100 steps. The ten diagnostic checkpoints are $\{1100, 1200, \ldots, 2000\}$.

**Held-out field bank.** Validation uses a fixed held-out bank of 6000 transverse fields drawn uniformly from $[0.8, 1.2]$ with seed 0, disjoint from the training fields. Each validation pass uses 20 spin configurations per held-out field, giving $|\mathcal{D}_{\mathrm{val}}| = 120{,}000$.

## C.2 $J_1$-$J_2$ ViT with symmetry ramp

**Hamiltonian.** The two-dimensional system is the $8 \times 8$ periodic $J_1$-$J_2$ Heisenberg model

$$H = J_1 \sum_{\langle i,j\rangle} \boldsymbol{\sigma}_i \cdot \boldsymbol{\sigma}_j + J_2 \sum_{\langle\langle i,j\rangle\rangle} \boldsymbol{\sigma}_i \cdot \boldsymbol{\sigma}_j, \qquad J_1 = 1, \quad J_2 = 0.5.$$

Here $\boldsymbol{\sigma}_i$ is the Pauli vector. These raw energy units are four times those of the same graph with spin operators $\mathbf{S}_i = \boldsymbol{\sigma}_i/2$ and the same numerical couplings. All energies and energy-squared diagnostics retain these Pauli units. The graph is fully periodic, with 128 bonds of each coupling type, and sampling is restricted to $\sum_i S^z_i = 0$. No Marshall sign rule is imposed.

**Model and optimizer.** The ansatz is the ViT architecture of Rende et al. [19], with depth 6, $d_{\text{model}} = 48$, 8 attention heads, MLP expansion factor 4, and $2 \times 2$ patches. The real parameter count is $P = 147{,}216$. SR uses the minSR formulation with Cholesky solves, peak learning rate $\eta = 0.01$, a 100-step linear warmup, and a cosine decay schedule.

**Symmetry ramp.** Training uses three phases with increasing symmetry:

- identity symmetry for 200 steps with $N_s = 8192$,
- translations for 3500 steps with $N_s = 8192$,
- translations plus $C_4$ rotations for 3500 steps with $N_s = 6000$.

The total training length is 7200 steps. The checkpoint source used for the MS-SR ablations is the $\lambda_{\text{train}} = 10^{-4}$ run. The ten diagnostic checkpoints are steps $\{900, 1600, 2300, 3000, 3700, 4400, 5100, 5800, 6500, 7200\}$.

### C.3 Checkpoint-Local Diagnostic Protocol

**Solve-time shift sweep.** The large-scale diagnostics hold the source checkpoint fixed and vary only the solve-time shift. For the TFIM U-curve in Figure 3, the diagnostic grid is

$$\lambda \in \{10^{-8}, 10^{-7}, 10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}.$$

Each value recomputes SR updates from fresh training batches at the same stored parameters.

**Delta banks.** For each checkpoint and solve-time shift, we build a bank of independent SR updates by repeatedly restoring the checkpoint state, drawing a fresh training batch, computing the SR update, and discarding the parameter update. This makes the stored directions independent draws from the checkpoint-local law of $\hat{\delta}_\lambda$. The TFIM U-curve in Figure 3 uses $m = 100$ deltas per shift. The Figure 4 baseline and ablation comparisons use 20 resampling repeats per checkpoint.

**Validation pass.** The validation diagnostics are computed by a streamed pass over a fixed validation block. Local energies and log-derivative Jacobian-vector products are computed once per validation sample and reused for all stored deltas. The $J_1$-$J_2$ validation block has 100,000 samples. The TFIM validation block has 120,000 samples, organized as 6000 held-out fields times 20 spin configurations per field.

**Centering and metrics.** The large-scale implementation uses the energy-gradient convention with regression target $2H_{\text{loc},c}$ and correspondingly doubled SR directions. Figures 3–4 and the large-scale tables retain these raw residual and variance values: both diagnostics are four times those under the $H_{\text{loc},c}$ normalization in Section 2. Relative comparisons, shift selection, and method ordering are unchanged by this common normalization. The Pauli-to-spin Hamiltonian conversion for $J_1$-$J_2$ is separate: at fixed checkpoints, consistently dividing the Hamiltonian and directions by four divides squared diagnostics by 16. The figures apply neither conversion. For $J_1$-$J_2$, local energies and predictions are centered globally over the validation block. For TFIM, they are centered separately within each held-out field, matching the replica structure of the training objective. The validation residual $\widehat{\mathcal{E}}_{\text{val}}$ is the mean squared held-out regression residual for the stored updates. The multi-batch variance $\mathcal{V}_{\text{mb}}$ is the centered prediction-space disagreement of the stored updates around their bank mean; it is computed from $O_c(x)^T \hat{\delta}_j$ rather than from raw parameter differences, so flat parameter directions do not contribute.

### C.4 Historical $J_1$-$J_2$ Online Continuations

**Source state and paired continuations.** Figure 5(a,b) compares five paired continuations from one shared SR checkpoint at iteration 2500 of the $8 \times 8$ $J_1$-$J_2$ model, using the same Pauli Hamiltonian and zero-magnetization sector specified above. The ViT has depth 8, embedding dimension 72, 12 attention heads, and MLP expansion factor 4. It uses only the identity symmetry operation: the wavefunction is not explicitly symmetrized over lattice translations or rotations. This is a separate setup from the symmetry-ramp experiment above. The continuation seeds are 104–108, corresponding in order to pairs 1–5 in Figure 5(b). Thus the five pairs sample continuation variability conditional on one source state, rather than variability across independently trained source states.

**Algorithm 1** Multi-Shift SR (MS-SR): one optimization step

**Require:** parameters $\theta_t$; number of shifts $K \geq 2$; batch size $N_s$; NTK-spectrum quantiles $\{q_j\}_{j=1}^{K}$.

1: Draw independent batches $\mathcal{D}_1, \ldots, \mathcal{D}_K \sim \pi_{\theta_t}$, each of size $N_s$. On each batch compute centered features $O_{c,j}$ and centered local energies $H_{\mathrm{loc},c,j}$.

2: Estimate one empirical NTK spectrum $\{s_i\} = \mathrm{spec}_+(O_{c,1}O_{c,1}^T/N_s)$, and set

$$\lambda_j \leftarrow \mathrm{Quantile}(\{s_i\}, q_j), \qquad j = 1, \ldots, K.$$

3: In parallel, compute $\hat{\delta}_j \leftarrow \mathrm{minSR}(\mathcal{D}_j, \lambda_j)$ for $j = 1, \ldots, K$.

4: Fit simplex weights $\mathbf{w}$ by numerically minimizing over $\mathbf{v} \in \Delta^{K-1}$ the leave-one-batch-out objective

$$\sum_{j=1}^{K} \left\| \sum_{i \neq j} \frac{v_i}{1 - v_j} O_{c,j} \hat{\delta}_i - H_{\mathrm{loc},c,j} \right\|^2 .$$

5: Apply the mixture update:

$$\theta_{t+1} \leftarrow \theta_t - \eta \sum_{j=1}^{K} w_j \hat{\delta}_j .$$

6: **return** $\theta_{t+1}, \boldsymbol{\lambda}, \mathbf{w}$.

**Update construction and budgets.** Both methods use Cholesky solves, $N_s = 8192$ samples per batch, and constant learning rate $\eta = 0.01$. SR uses $\lambda = 10^{-4}$. The historical MS-SR implementation constructs four candidate updates on independent batches at the fixed shifts $\{10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}\}$, then fits stacking weights on an additional independent batch. Its callback also executes a base-driver SR solve on another batch. Each MS-SR update therefore uses five solves and six sampled batches; 125 updates use 625 solves and 750 batches, compared with 500 of each for the 500-update SR control. This fixed-shift, separate-weight-batch protocol differs from the NTK-quantile, leave-one-batch-out method in Algorithm 1. The historical runs do not compute NTK quantiles; their additional solve comes from the base driver. For the NTK-quantile implementation, spectrum estimation and its reuse within the solver are described separately in Appendix F.1.

**Endpoint evaluation and uncertainty.** Each final state is evaluated independently using two fresh-chain replicas, with 8,388,608 retained samples in total. For pair $i$ we report $\Delta_i = (E_{\text{MS-SR},i} - E_{\mathrm{SR},i})/N$, where $N = 64$; energies, differences, and their uncertainties are all in Pauli units. The individual error bars combine the two methods' Monte Carlo standard errors in quadrature. The 95% interval for the mean paired difference is $\overline{\Delta} \pm t_{0.975,4} s_\Delta/\sqrt{5}$, where $s_\Delta$ is the sample standard deviation across the five pairs. This interval summarizes variation across continuations conditional on the shared source state.

**Training-curve display.** The horizontal axis in Figure 5(a) counts one nominal SR-equivalent update per SR step and four per MS-SR step; both curves end at 500. It excludes the historical callback's extra solve and sampled batches. The shared point at zero is the mean paired initial observation. Subsequent training observations are averaged within each continuation in half-open bins of width 20 on this nominal axis, excluding the initial observation; the curves and bands show the mean and $\pm 1$ sample standard deviation across the five continuations. Diamonds at 500 show the independent endpoint means.

## D Offline Diagnostics and MS-SR Protocols

**MS-SR shifts.** All MS-SR ablations in Figure 4 use $K = 4$, with NTK-spectrum quantiles $\{0.9, 0.7, 0.4, 0.1\}$. For these offline diagnostics, the empirical spectrum is computed on a separately seeded reference batch and cached once per checkpoint, then reused across all 20 resampling repeats. This holds the shift grid fixed while comparing candidate updates and does not fit the shifts on the final validation data. Algorithm 1 shows the online construction, where the spectrum can instead

be obtained from the first candidate batch at the current step, as in the optimized benchmark of Appendix F.1.

Here $O_{c,j}$ has the unnormalized centered features as its rows, so the factor $1/N_s$ makes the NTK eigenvalues agree with the nonzero eigenvalues of the empirical QGT. Quantiles use only strictly positive eigenvalues. In the offline diagnostics, a nonfinite or nonpositive selected quantile falls back to $\max(s_{\min,+}, 10^{-12} s_{\max})$; the optimized benchmark additionally floors every selected shift at this value. For real parameters and complex features, the corresponding real-stacked feature matrix is used.

**Shared-batch and independent-batch constructions.** The shared-batch variants solve all four shifts on one shared fitting batch. This preserves the exact finite-sample interpretation as a convex mixture of ridge filters on a single empirical QGT. The independent-batch variants solve one shift per independent fitting batch. They no longer correspond to a single empirical spectral filter, but their sampling errors are decorrelated, giving the variance-reduction mechanism analyzed in Section 5.

**Stacking weights.** For MS-SR, weights are selected by leave-one-batch-out stacking. Candidate update $\hat{\delta}_i$ is evaluated on batches that did not produce it, and the weights are constrained to the simplex. The shared-batch stacked ablation uses the same candidate deltas as the shared-batch uniform method. TFIM fits the simplex weights on an independent weight-fit batch. The $J_1$-$J_2$ shared-batch variants instead split one sampled batch into a fitting part and a held-out weight-fit part: $6144/2048$ samples through step $3700$, and $4496/1504$ thereafter. Independent-batch methods use the full $8192$ or $6000$ samples per candidate. Uniform variants set $w_i = 1/K$. The leave-one-batch-out objective is defined for $w_j < 1$: a simplex vertex would leave no candidate after removing its sole nonzero weight. The implementation rejects trial weights with $1 - w_j \leq 10^{-12}$ rather than evaluating this undefined denominator. The weights are obtained with SLSQP initialized at the uniform mixture. The objective contains the ratios $w_i/(1 - w_j)$ and is not generally convex; numerical fitting does not guarantee a global minimum. At $K = 2$, the leave-one-batch-out objective is constant for interior simplex weights, since each held-out prediction uses only one remaining candidate. The implementation consequently retains the uniform initialization, explaining the identical uniform and stacked results in that case.

**Ablations.** Figure 4 compares the following checkpoint-local updates:

- **Standard SR:** one SR solve at $\lambda_{\mathrm{train}} = 10^{-4}$.
- **Bagged SR (two variants):** four independent SR solves, averaged uniformly at the common shift $\lambda_{\mathrm{train}}$ or the checkpoint-specific validation-selected $\lambda_{\mathrm{tuned}}$.
- **Shared-batch multi-shift, uniform:** four NTK-quantile shifts on one batch, averaged uniformly.
- **Shared-batch multi-shift, stacked:** the same four shared-batch candidates with simplex weights fit on held-out samples as described above.
- **Independent-batch multi-shift, uniform:** four independent batches, one NTK-quantile shift per batch, averaged uniformly.
- **MS-SR:** four independent batches, NTK-quantile shifts, and leave-one-batch-out stacking weights.

At each checkpoint, the tuned shift $\lambda_{\mathrm{tuned}}$ minimizes the mean held-out residual of the four-update bagged average across 20 resampling repeats, over $\{10^{-8}, 10^{-7}, \ldots, 10^{-1}\}$. The same four sampled batches are reused across shifts within each repeat. This requires additional shifted solves and evaluations using the reporting data, making it an oracle reference. Since the grid includes $\lambda_{\mathrm{train}}$, the selected mean residual cannot exceed that of training-shift bagging on these data. The variance is reported at the same selected shift; it is not optimized separately. Bagging at $\lambda_{\mathrm{train}}$ needs no such search. For TFIM, $\lambda_{\mathrm{tuned}} = 10^{-7}$ at step $1100$ and $10^{-6}$ at the other nine checkpoints. The $J_1$-$J_2$ choices range from $10^{-4}$ to $10^{-1}$. We use $\lambda_{\mathrm{single}}$ for the shift minimizing the single-SR residual on the same eight-point grid. This separate selection is retained for the single-SR diagnostic in Table 1 and the $K$ and source-training-shift robustness experiments below. The ablation separates the effect of richer spectral filters from the effect of independent-batch averaging. Bars in the main figure are

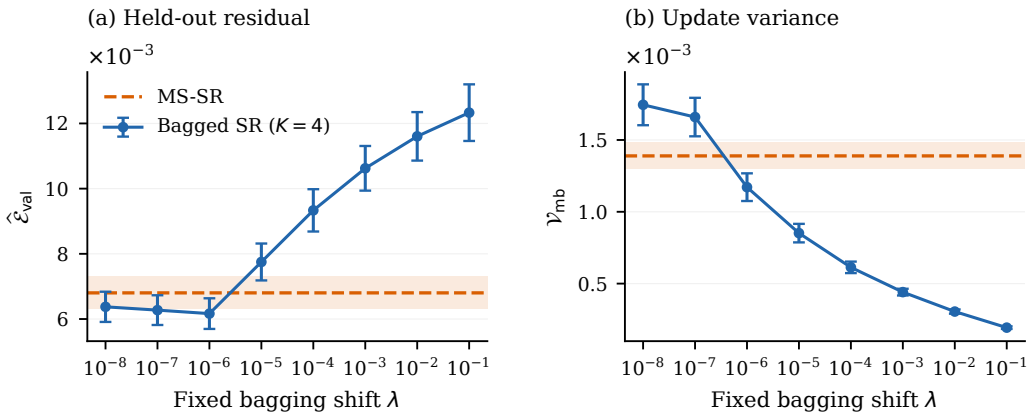



Figure 6: Fixed-shift bagged SR on TFIM ($K = 4$). Points are means over ten checkpoint-level diagnostics; error bars are standard errors over those checkpoints. The dashed line and shaded band show the MS-SR mean and standard error on the same checkpoints. The training shift is $10^{-4}$. Minimizing the bagged residual selects $10^{-7}$ at step 1100 and $10^{-6}$ at the other nine checkpoints. The complete sweep and checkpoint standard deviations are retained in the released summary data.

computed from the same fixed source checkpoints for each method; error bars use the checkpoint standard deviation (divisor 10) divided by $\sqrt{10}$.

# E Robustness of the Checkpoint-Local Comparisons

## E.1 Bagging at fixed shifts

Table 1 summarizes the TFIM baselines at the training shift and at shifts selected for the single-SR or bagged-SR residual.

| TFIM update rule | $10^3\widehat{\mathcal{E}}_{\text{val}}$ | $10^3\mathcal{V}_{\text{mb}}$ |
|---|---|---|
| Single SR, $\lambda_{\text{single}}$ | $9.70 \pm 2.43$ | $4.71 \pm 1.24$ |
| Bagged SR, $\lambda_{\text{train}}$ | $9.33 \pm 2.05$ | $0.61 \pm 0.13$ |
| Bagged SR, $\lambda_{\text{tuned}}$ | $6.17 \pm 1.49$ | $1.24 \pm 0.48$ |
| MS-SR, NTK quantiles | $6.80 \pm 1.59$ | $1.39 \pm 0.29$ |

Table 1: TFIM baseline comparison: mean $\pm$ standard deviation across ten checkpoints (20 resampling repeats per checkpoint). Bagged SR and MS-SR use four independent batches and four solves per candidate shift. Here $\lambda_{\text{train}} = 10^{-4}$ and $\lambda_{\text{single}} = 10^{-6}$ throughout. Bagged-residual tuning selects $10^{-7}$ at step 1100 and $10^{-6}$ elsewhere. Both tuned references use the evaluation data for selection; the additional shift-search cost is excluded from the four-solve count.

Figure 6 sweeps all eight bagging shifts on the ten TFIM checkpoints, using the same Monte Carlo batches as the $K = 4$ ablation. Increasing the shift reduces $\mathcal{V}_{\text{mb}}$, but the held-out residual also includes shrinkage bias. MS-SR has mean held-out residual 0.00680, compared with 0.00933 for bagging at the training shift $10^{-4}$, a reduction of approximately 27%. Selecting the bagging shift per checkpoint gives 0.00617, which is lower than MS-SR. The fixed shift $10^{-6}$ is optimal on nine of the ten checkpoints; step 1100 selects $10^{-7}$. Thus the comparison depends on the shift available to the single-shift method; lower variance alone does not imply a better update.

## E.2 Number of candidates and quantile grid

For this robustness study, $\lambda_{\text{tuned}} = 10^{-6}$ denotes a common reference shift across the ten TFIM checkpoints. It is selected by the single-SR validation sweep and also minimizes the checkpoint-averaged bagged residual over the eight tested shifts in the $K = 4$ sweep. We retain this reference as

| Method | $K = 2$ | $K = 3$ | $K = 4$ | $K = 5$ |
|---|---|---|---|---|
| MS-SR | $8.54/2.28$ | $7.16/1.74$ | $6.80/1.39$ | $6.38/1.27$ |
| Bagged SR, $\lambda_{\text{train}}$ | $9.93/1.22$ | $9.53/0.82$ | $9.33/0.61$ | $9.21/0.49$ |
| Bagged SR, $\lambda_{\text{tuned}}$ | $7.33/2.34$ | $6.56/1.56$ | $6.17/1.17$ | $5.93/0.94$ |

Table 2: Sensitivity to $K$ on TFIM. Each entry is $\widehat{\mathcal{E}}_{\text{val}}/\mathcal{V}_{\text{mb}}$, with both values in units of $10^{-3}$, averaged over ten checkpoint-level diagnostics. Here $\lambda_{\text{train}} = 10^{-4}$ and $\lambda_{\text{tuned}} = 10^{-6}$. The sampling and solve budget grows linearly with $K$.

| Quantile grid | $\widehat{\mathcal{E}}_{\text{val}}$ | $\mathcal{V}_{\text{mb}}$ |
|---|---|---|
| $\{0.9, 0.7, 0.4, 0.1\}$ (default) | $6.802 \pm 1.595$ | $1.389 \pm 0.291$ |
| $\{0.95, 0.8, 0.5, 0.2\}$ | $6.938 \pm 1.625$ | $1.469 \pm 0.283$ |
| $\{0.8, 0.6, 0.3, 0.05\}$ | $6.574 \pm 1.568$ | $1.369 \pm 0.321$ |

Table 3: Sensitivity to the $K = 4$ NTK quantile grid on TFIM. Values are in units of $10^{-3}$ and show means $\pm$ standard deviations across ten checkpoint-level diagnostics, using 20 resampling repeats per checkpoint.

$K$ varies. It also gives the lowest observed mean residual among $\{10^{-6}, 10^{-5}, 10^{-4}\}$ in the saved $K = 2, 3, 5$ comparisons. Figure 4 instead uses the checkpoint-specific bagged optimum, which differs from this common reference only at step 1100.

For a fixed shift and checkpoint, independent uniform averaging leaves bias unchanged and divides variance by $K$. Hence its predicted residual is $\widehat{\mathcal{E}}_{\text{val,single}} - (1 - 1/K)\mathcal{V}_{\text{mb,single}}$. At $\lambda_{\text{tuned}} = 10^{-6}$ and $K = 4$, this gives $0.00970 - 0.75(0.00471) \approx 0.00617$, matching the observed bagged residual. Using the paired single-SR diagnostics for $K = 2$, 3, and 5 gives agreement within $0.3\%$. This checks the benefit of averaging at fixed parameters; it does not compare the resulting update to $K$ sequential optimization steps.

We repeat the TFIM diagnostic with $K \in \{2, 3, 4, 5\}$, using 20 resampling repeats per checkpoint and seed-paired comparisons. The quantile grids are $\{0.9, 0.1\}$, $\{0.9, 0.5, 0.1\}$, $\{0.9, 0.7, 0.4, 0.1\}$, and $\{0.9, 0.7, 0.4, 0.2, 0.1\}$, respectively. Table 2 reports the checkpoint means. The held-out residual improves as $K$ increases for both MS-SR and bagging; this also increases the number of sampled batches and SR solves. MS-SR improves on bagging at the training shift at every tested $K$, while bagging at $\lambda_{\text{tuned}}$ remains stronger on this TFIM window.

At $K = 4$, we also shift the quantile grid upward or downward (Table 3). The mean held-out residual ranges from $0.00657$ to $0.00694$, within approximately $3.4\%$ of the default result. The default quantile grid is fixed across systems; its numerical shifts adapt to the empirical spectrum, and stacking adapts the mixture weights without a held-out validation sweep.

### E.3 Source training shift

To test whether the comparison depends on the source trajectory's diagonal shift, we repeat the diagnostic at steps 1100, 1500, and 2000 of TFIM runs trained with $\lambda_{\text{train}} = 10^{-3}$ and $10^{-5}$. The single-SR validation sweep is repeated for each source trajectory and checkpoint, with $\lambda_{\text{single}} \in \{10^{-5}, 10^{-6}\}$. Table 4 shows similar mean held-out residuals for MS-SR and bagging at $\lambda_{\text{single}}$ on both trajectories. These three-checkpoint comparisons support robustness to the source training shift within this TFIM setup.

## F Compute Resources

### F.1 Steady-state update-cost benchmark

Figure 5(c) measures the direct update constructors at the TFIM step-2000 checkpoint, with $L = 100$ and $N_s = 12{,}000$ samples per candidate batch. Each run uses one NVIDIA H100 80GB HBM3 GPU on a host with two Intel Xeon Platinum 8481C processors. All three methods use the same Hermitian eigendecomposition-based pseudoinverse solver with relative cutoff $10^{-12}$. SR uses one

| $\lambda_{\mathrm{train}}$ | Method | $\widehat{\mathcal{E}}_{\mathrm{val}}$ | $\mathcal{V}_{\mathrm{mb}}$ |
|---|---|---|---|
| $10^{-3}$ | Single SR, $\lambda_{\mathrm{single}}$ | $36.832 \pm 6.731$ | $16.348 \pm 1.470$ |
| | Bagged SR, $\lambda_{\mathrm{single}}$ | $24.590 \pm 6.990$ | $4.063 \pm 0.369$ |
| | MS-SR | $24.726 \pm 4.623$ | $5.384 \pm 0.759$ |
| $10^{-5}$ | Single SR, $\lambda_{\mathrm{single}}$ | $2.107 \pm 0.621$ | $0.700 \pm 0.179$ |
| | Bagged SR, $\lambda_{\mathrm{single}}$ | $1.580 \pm 0.482$ | $0.174 \pm 0.044$ |
| | MS-SR | $1.575 \pm 0.470$ | $0.256 \pm 0.060$ |

Table 4: TFIM diagnostics on trajectories trained with alternative diagonal shifts. Values are in units of $10^{-3}$ and show means $\pm$ standard deviations across the three checkpoint-level diagnostics. Bagged SR and MS-SR use $K = 4$.

batch and one solve at $\lambda = 10^{-4}$; bagged SR uses four batches and four solves at that shift, followed by uniform averaging; MS-SR uses four batches, four NTK-quantile solves, and exact leave-one-batch-out stacking. The bagged-SR timing corresponds to $\lambda_{\mathrm{train}}$; it excludes the additional sweep needed to select $\lambda_{\mathrm{tuned}}$.

Timing includes sampling, local energies, candidate solves, and construction of the final direction. Parameters are held fixed during timing; the common final parameter write and optional residual diagnostics are excluded. For MS-SR, the first unshifted NTK eigendecomposition supplies the positive spectrum quantiles and is reused for the first shifted solve. Stacking reuses the four candidate batches and their local energies. Its twelve off-batch candidate predictions are evaluated in four batched Jacobian-vector-product calls; sufficient statistics are reduced on device before transferring the small Gram matrices and target projections to the simplex fitter. The median prediction and fitting times are $0.477$ s and $0.006$ s, respectively. Reusing the spectrum in this way requires a solver that exposes its eigendecomposition; an implementation based only on Cholesky or iterative solves must account separately for spectrum estimation.

We use three paired device/seed configurations, with all methods on the same physical GPU within each pair. Each method receives two warm-ups followed by five device-synchronized timed repetitions. Method-independent sampler names pair the Monte Carlo draws: SR's batch matches the first batch of both $K = 4$ methods, and all four batch fingerprints match between bagged SR and MS-SR. We take medians over repetitions within each pair, then report the median and range over the three pair-level summaries. Differences and ratios are computed on paired repetitions before taking these medians. The pair-level time ranges are $13.15$–$13.45$ s for SR, $52.71$–$53.60$ s for bagged SR, and $53.14$–$54.18$ s for MS-SR. MS-SR adds a paired median of $0.550$ s ($1.027\%$) over bagged SR, with a three-pair range of $0.437$–$0.579$ s ($0.829$–$1.080\%$). Thus the methods share the dominant four-batch/four-solve cost, with a small measured stacking increment in this implementation.

The recorded environment uses Python 3.12.13, JAX 0.7.2, NetKet 3.21.0, and SciPy 1.18.0. The benchmark implementation is available through `sr-filter benchmark`; released timing summaries are under `artifacts/paper/figure5/` in the repository. The recorded device-memory high-water marks are $58.0$ GB for SR and $58.1$ GB for each $K = 4$ method (decimal GB), identical across the three pairs at this precision. These are process-lifetime/JIT high-water marks and must not be interpreted as isolated per-update peaks.

### F.2 Historical experiment accounting

Table 5 records the original experiment campaigns. Its 762 H100 GPU-hours are a subtotal for those listed campaigns; they do not include the additional robustness diagnostics, the optimized cost benchmark, or the final online comparison.

| Experiment | Manuscript use | Worker | Wall time | GPU-hours |
|---|---|---|---|---|
| 4 by 4 small experiments | Figures 2 and 7 | $1\times$ H100 80GB GPU | $\approx 30$ h | $\approx 30$ |
| TFIM-FNQS training | Figure 3 and 4 checkpoints | $8\times$ H100 80GB GPUs | $\approx 1.9$ h | $\approx 15$ |
| TFIM single-shift sweep | Figure 3 | $8\times$ H100 80GB GPUs | $\approx 25.5$ h | $\approx 204$ |
| $J_1$–$J_2$ ViT training | Figure 4 checkpoints | $8\times$ H100 80GB GPUs | $\approx 15.1$ h | $\approx 121$ |
| $J_1$–$J_2$ MS-SR ablations | Figure 4 | $8\times$ H100 80GB GPUs | $\approx 46$ h | $\approx 368$ |
| TFIM MS-SR ablations | Figure 4 | $8\times$ H100 80GB GPUs | $\approx 3.0$ h | $\approx 24$ |

Table 5: Compute-resource accounting for the original experiment campaigns. The listed entries sum to approximately 762 H100 GPU-hours; this is a historical subtotal rather than the total for all camera-ready experiments.

## G Online Tracking of a Population SR Reference

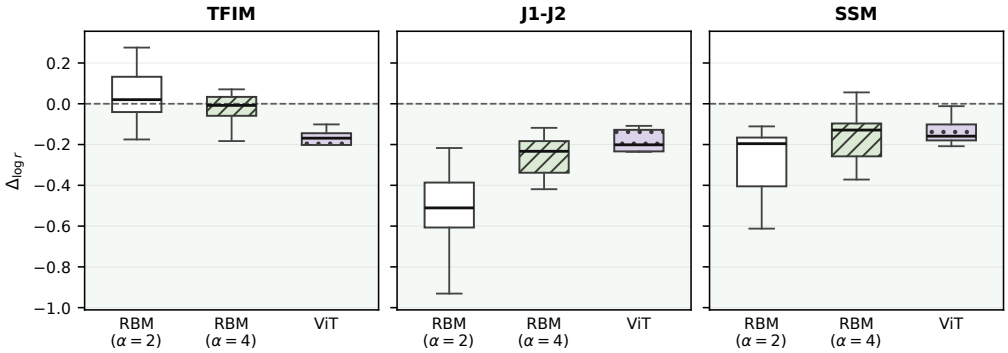



Figure 7: Online update tracking in exact $4 \times 4$ systems. SR, SPRING, and MS-SR are trained separately from the same initialization. At each step we compare the stochastic update with a full-summation population SR reference at shift $10^{-4}$. We plot the full-trajectory log-ratio $\Delta_{\log r} := \log_{10}(\overline{r}_{\text{MS-SR}} / \min\{\overline{r}_{\text{SR}}, \overline{r}_{\text{SPRING}}\})$. Negative values favor MS-SR. Boxes aggregate random seeds, sample sizes, and learning rates for each Hamiltonian–ansatz pair. Iteration counts are matched; MS-SR uses four candidate solves plus spectrum and weight-fit overhead. This variant smooths its stacking weights across iterations.

Figure 7 compares stochastic updates to a population SR reference along separate online trajectories of SR, SPRING [5, 6], and MS-SR. At each optimizer's parameters $\theta_{a,t}$, exact summation over the configured Hilbert space gives $S_{a,t}$ and $g_{a,t}$, including both real and imaginary tangent channels for complex wavefunctions. The reference $\delta^{\text{ref}}_{a,t}$ is computed with shift $10^{-4}$ and the smoothed pseudoinverse solver; it is a numerically regularized population update, rather than the unregularized $\delta^*$ of Section 2. We measure $r_{a,t} = \|\hat{\delta}_{a,t} - \delta^{\text{ref}}_{a,t}\|^2_{S_{a,t}} / \|\delta^{\text{ref}}_{a,t}\|^2_{S_{a,t}}$. Eq. (3) relates the numerator to local update infidelity with respect to this reference. The MS-SR variant uses NTK-quantile shifts, an independent weight-fit batch, and an exponential moving average of its stacking weights; it differs from the leave-one-batch-out construction in Section 5.

The plotted data comprise 96 completed matched configurations across three $4 \times 4$ Hamiltonians (TFIM, $J_1$-$J_2$, and Shastry–Sutherland), three ansätze (RBM $\alpha = 2$, RBM $\alpha = 4$, and a small ViT), and sample sizes $N_s \in \{512, 1024, 2048\}$. The available runs use learning rate $0.01$ with seeds 0 and 1, and $0.005$ with seeds 0, 1, and 2; the plot retains only configurations completed by all three methods for $T = 100$ iterations. For each optimizer we summarize $\overline{r}_a = T^{-1} \sum_t r_{a,t}$. Subsection G.1 gives the per-panel counts and update variant.

The paired log-ratio in Figure 7 compares MS-SR with the smaller of the SR and SPRING trajectory-averaged errors. Negative values favor MS-SR over both baselines. Most completed settings favor MS-SR in eight of nine Hamiltonian–ansatz pairs; the exception is TFIM with RBM $\alpha = 2$, where $4/10$ settings favor it. These are local errors along distinct trajectories, not distances to one common ideal trajectory or an equal-budget final-energy comparison.

**Hamiltonians and sectors.** These online experiments use separate constructors from the amplitude diagnostic in Appendix B. All three models have $4 \times 4$ sites and periodic boundaries, and use Pauli operators. The $J_1$-$J_2$ Hamiltonian has 32 nearest-neighbor bonds with $J_1 = 1$ and all 32 diagonal bonds with $J_2 = 0.5$. The Shastry–Sutherland Hamiltonian has 32 square-lattice bonds with coupling 1 and 8 orthogonal dimer bonds with coupling $0.8$. Both Heisenberg models use $\sum_i S^z_i = 0$, with $\binom{16}{8} = 12{,}870$ basis states, and no Marshall sign transformation. The TFIM uses $H = +\sum_{\langle i,j \rangle} Z_i Z_j - \sum_i X_i$ and the unrestricted $2^{16} = 65{,}536$-state space. Its positive Ising coupling differs from the one-dimensional TFIM convention in Appendix C; on this even bipartite square torus, conjugation by $X$ on one sublattice reverses the bond sign while leaving the field term unchanged. The relative QGT errors use the full configured geometry, rather than the amplitude-only covariance of Figure 2.

| System | RBM, $\alpha = 2$ | RBM, $\alpha = 4$ | ViT |
|---|---|---|---|
| TFIM | $4/10$ | $8/11$ | $12/12$ |
| $J_1$-$J_2$ | $11/11$ | $10/10$ | $8/9$ |
| SSM | $12/12$ | $10/11$ | $9/10$ |

Table 6: Per-panel counts for Figure 7. Each entry is the number of configurations with lower MS-SR relative QGT-norm error than both SR and SPRING, divided by the number of completed matched configurations in that panel. The denominators sum to 96.

### G.1 Protocol and per-panel counts

Figure 7 uses the completed matched subset of five 100-iteration sweeps: learning rate $0.01$ with seeds $0, 1$, and learning rate $0.005$ with seeds $0, 1, 2$. The sample sizes are $512$, $1024$, and $2048$. The plotted asset contains $96$ matched configurations across Hamiltonians, ansatzes, sample sizes, learning rates, and seeds. A configuration contributes only when the selected MS-SR variant, SR, and SPRING have complete tracking records at the target iteration count. Consequently, the panel counts differ (Table 6); the figure does not represent a completed Cartesian sweep of all configurations. These counts and plotted values are retained from the archived figure-source table. The presently available raw archive has nine missing method configuration files and four additional trajectories ending before step $100$; consequently, those entries cannot currently be reconstructed independently from the available raw trajectories.

The selected variant is `mssr_ema_ntk_quantile`. It recomputes the four NTK-quantile shifts at each step, fits simplex weights on an independently sampled weight-fit batch, and smooths them as $w_t = 0.9w_{t-1} + 0.1w_t^{\mathrm{raw}}$ after initialization. These runs use the research callback with an additional baseline driver solve and spectrum probe, so the four candidate solves do not represent the complete per-step cost. This is a comparison at matched iteration counts using this online variant, with a separate weight-fit batch and smoothed weights; the optimized exact leave-one-batch-out construction is documented in Appendix F.1.

At every tracked state, the reference direction is computed by exact summation over the configured Hilbert space with diagonal shift $10^{-4}$ and the `pinv_smooth` solver. It is therefore a population, regularized SR reference $\delta_{\mathrm{ref}}$, rather than the unregularized $\delta^* = S^\dagger g$ used in the regression identities. The QGT metric is also computed by full summation. Each plotted point compares the trajectory-mean relative QGT-norm error of the selected MS-SR variant with the smaller of the corresponding SR and SPRING errors. These measurements describe local direction tracking along the optimizers' own trajectories, rather than a comparison of final energies at equal sampling or wall-time budgets.

## 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 state the paper's main claims: the fixed-checkpoint ridge-regression view of SR, the role of the expressivity gap, the exact and large-scale diagnostics, and the MS-SR ablations.

2. **Limitations**

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

   Answer: [Yes]

   Justification: Section 7 discusses testing MS-SR in more difficult systems and determining when its additional optimizer cost is worthwhile. Section 3 specifies the scope of the exact amplitude-regression experiments, and Section 2 and Appendix A state the assumptions of the checkpoint-local analysis.

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: Section 2 states the fixed-checkpoint regression setup, and Appendix A gives the supporting identities and spectral-risk derivation.

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: Sections 3–5 and Appendices B–F describe the systems, sample budgets, diagnostic construction, and MS-SR ablations needed to reproduce the reported comparisons.

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: [Yes]

   Justification: The Code and data availability statement links a public repository containing the code and data for this work: `https://github.com/takh04/sr_filter`. The manuscript and appendices document the experimental protocols and run settings, and the repository includes quickstart and figure-regeneration commands.

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: The main experimental sections and appendices specify the Hamiltonians, architectures, sample sizes, checkpoint selection, validation batches, shift sweeps, and MS-SR weighting rules.

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: The exact experiments use full-population quantities where available, and the large-scale/MS-SR diagnostics report checkpoint-local averages and standard-error summaries for the figure ablations. Tables define standard deviations across checkpoints, and the timing benchmark reports medians and ranges over paired device/seed settings.

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: Appendix F reports the compute workers and the original experiments' GPU-hour subtotal; Appendix F.1 details the paired benchmark hardware, timing protocol, and process-lifetime memory accounting. Additional robustness runs are not included in the original compute subtotal.

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: The work is fundamental research on optimization and diagnostics for variational quantum many-body simulation, and it does not involve human subjects, private data, or deployed decision systems.

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: The work studies optimizer regularization for neural quantum states and does not introduce a deployed system or a direct application with identifiable societal impact.

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: The paper does not release a model, dataset, or system with high misuse risk that would require special safeguards.

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: The experimental tools were developed on top of NetKet, an open-source neural quantum state and variational Monte Carlo library distributed under the Apache License 2.0, which is credited in the acknowledgments. The paper does not use third-party datasets or pretrained models.

13. **New assets**

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

    Answer: [Yes]

    Justification: The paper releases a public repository containing the code and data for this work: `https://github.com/takh04/sr_filter`. The repository is linked in the code and data availability statement and includes quickstart and figure-regeneration commands; the paper documents the experimental protocols and compute resources in the appendix.

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: The work does not involve crowdsourcing or human subjects research.

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: The work does not involve human subjects, so IRB or equivalent approval is not applicable.

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: LLMs are not part of the scientific method, algorithm, experiments, or evaluation described in the paper.