A Locality-Aware Surrogate for Natural-Gradient Descent in Quantum Optimization

Md Mobasshir Arshed Naved
Department of Computer Science
Purdue University
West Lafayette, USA
naved@purdue.edu

Wenbo Xie
Department of Computer Science
Purdue University
West Lafayette, USA
xie401@purdue.edu

Wojciech Szpankowski
Department of Computer Science
Purdue University
West Lafayette, USA
szpan@purdue.edu

Ananth Grama
Department of Computer Science
Purdue University
West Lafayette, USA
ayg@cs.purdue.edu

Abstract

Variational quantum algorithms (VQAs) rely on gradient-based optimization over parameterized circuits whose loss landscape lives on a manifold of quantum states. Quantum natural gradient (QNG) accounts for this geometry through the quantum Fisher information matrix (QFIM). However, estimating the full QFIM is prohibitively expensive within practical quantum circuit shot budgets. Existing tractable approximations either discard inter-block couplings, as in block-diagonal QNG, or replace the QFIM with a measurement-induced classical Fisher surrogate, as in random natural gradient (RNG). We propose a locality-aware surrogate that targets the QFIM through its covariance representation in terms of propagated generators, retaining a structured subset of informative cross-parameter couplings while remaining practical to estimate under finite-shot budgets. Our construction combines three ingredients: weight-k Pauli truncation to control estimation variance, support-overlap masking to exploit sparsity, and single-dataset classical-shadow estimation that reuses the same shadow dataset across all retained metric entries. Truncation level k provides a bias–variance tradeoff: larger k captures more of the propagated-generator structure but increases estimation cost, making it a budget-constrained design choice. Circuit parameters whose truncated generators remain nonzero form an active set receiving surrogate-QNG updates, while the rest follow standard gradient descent. We provide finite-shot concentration bounds for the shadow-estimated surrogate and convergence guarantees to a stationary neighborhood under standard smoothness assumptions. Experiments on variational quantum eigensolver (VQE) instances with local ansätze for Heisenberg and Ising spin chains show improved convergence per cumulative shot relative to SGD, Adam, block-diagonal QNG, and RNG baselines under finite-shot budgets, on circuits with up to 16 qubits.

1 Introduction

Variational quantum algorithms (VQAs) optimize a parameterized quantum circuit $U(\overrightarrow{\theta})$ over $\overrightarrow{\theta} \in \mathbb{R}^p$ to minimize a cost function $C(\overrightarrow{\theta}) = \text{tr}\left(\mathcal{M}\, U(\overrightarrow{\theta})\rho_{\text{in}}U^{\dagger}(\overrightarrow{\theta})\right)$, where $\mathcal{M}$ denotes the Hermitian observable and ρ_{in} the input state [20, 6]. The optimizer operates on a classical objective,

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

but the underlying landscape lives on the manifold of quantum states rather than in a Euclidean parameter space. This mismatch can lead to poor conditioning: equal-sized parameter steps may induce very different changes in the quantum state, slowing convergence and degrading scaling of gradient-based methods.

Quantum natural gradient (QNG) [25] mitigates this mismatch by replacing the Euclidean gradient step with $\overrightarrow{\theta}^{t+1} = \overrightarrow{\theta}^{t} - \eta F(\overrightarrow{\theta}^{t})^{-1} \nabla C(\overrightarrow{\theta}^{t})$, where $F(\overrightarrow{\theta})$ is the quantum Fisher information matrix (QFIM), the metric tensor of the Fubini–Study geometry on pure quantum states [5]. This update follows the steepest-descent direction on the state manifold and often improves conditioning and convergence in practice. The challenge lies in estimating matrix F, since standard approaches require $O(p^2)$ metric entries for a circuit with p parameters and substantial shot budgets, which becomes prohibitive for circuits with large numbers of parameters.

Existing tractable approximations make different trade-offs. Stokes et al. [25] propose a *block-diagonal* approximation that partitions parameters into blocks, typically aligned with circuit layers, reducing the cost to $O(pq)$ for block size q but ignoring inter-block couplings, which can be important in entangling circuits. Kolotouros and Wallden [16] take a different approach with their Random Natural Gradient (RNG) method, which preconditions the gradient with the *classical Fisher information matrix* (CFIM) induced by randomized measurements, requiring only $O(p)$ state preparations per iteration. However, for a fixed measurement model, CFIM is a positive semidefinite (PSD) lower bound on the QFIM and does not generally coincide with it. Consequently, RNG does not generally follow the Fubini–Study geometry, and its preconditioning quality can vary across states and tasks.

Our approach instead exploits the covariance representation of the QFIM in terms of propagated generators with Pauli expansions (Section 3.2), which makes it a natural target for classical-shadow estimation [13]. We construct a locality-aware surrogate using three components: (i) weight-k truncation of propagated generators to control variance in the subsequent metric estimation, (ii) support-overlap masking to exploit sparsity, and (iii) single-dataset classical-shadow estimation that reuses the same shadow dataset across all retained metric entries, rather than separate per-block or per-pair estimation routines. Per iteration, the quantum cost of the metric estimate is N single-shot shadow snapshots, with N controlled by the retained mask structure and truncation level k through Corollary 5.2. In sparse local regimes, the support-overlap mask retains far fewer than p^2 entries. Since a single shadow dataset estimates all retained entries simultaneously, the method avoids separate estimation of the dense QFIM. The truncation also determines the active set used by the hybrid update: parameters whose truncated generators are zero carry no surrogate curvature information and follow standard gradient descent, while the rest receive surrogate-QNG updates. We target VQA regimes with local structure, such as local Hamiltonians and local ansätze, where sparse surrogate geometry is most informative under finite measurement budgets.

Our locality-aware surrogate occupies a distinct point in the QNG design space. Unlike RNG, it targets the QFIM geometry rather than a measurement-induced CFIM. Unlike block-diagonal QNG, it retains a structured subset of cross-parameter couplings instead of discarding couplings across a fixed block partition. Theoretically, we establish finite-shot concentration bounds that decompose the surrogate error into truncation error, masking error, and finite-shot estimation error (Theorem 5.1), together with convergence guarantees to a stationary neighborhood under standard smoothness assumptions (Theorem 5.4). Experimentally, we evaluate the method on VQE instances for Heisenberg and Ising spin chains with up to 16 qubits, where it achieves improved energy convergence per cumulative shot relative to SGD, Adam, block-diagonal QNG, and RNG under finite-shot budgets.

The remainder of the paper is organized as follows: Section 2 presents related work and puts our contributions in the context of these results. Section 3 introduces the circuit model and the covariance form of the QFIM. Section 4 develops the shadow-estimated locality-aware surrogate and the resulting optimization algorithm. Section 5 presents our theoretical analysis, and Section 6 reports experimental results. Discussion and concluding remarks appear in Section 7.

2 Related Work

Beyond block-diagonal approximation [25] and RNG [16] discussed in the introduction, several other approaches have been proposed to make QNG more tractable. Recent work has extended QNG to noisy circuits [15] and analyzed measurement costs of metric-aware variational algorithms [26]. Other methods reduce metric-estimation cost by changing the estimation strategy or the surrogate

itself: Koczor and Benjamin [14] bypass metric estimation through local analytic surrogates of the loss; Gacon et al. [9] use SPSA for p-independent but high-variance QFIM estimation; and Sohail et al. [24] study stochastic pairwise-coordinate updates that estimate one 2×2 block per step. Shi et al. [23] propose WA-QNG, which replaces the full-state QFIM with a Hamiltonian-weighted sum of subsystem Hilbert–Schmidt metric tensors for k-local Hamiltonians. This defines a different locality-weighted geometry rather than a surrogate or sparsification of the QFIM itself. Since our goal is to approximate the Fubini–Study/QFIM geometry, RNG and block-diagonal QNG are the most directly aligned baselines. WA-QNG nevertheless provides a natural locality-aware baseline for the local-Hamiltonian setting considered here, and we include a direct comparison in Appendix G.3.

Classical shadows [13] and the broader randomized-measurement toolbox [8] enable simultaneous estimation of many observables from a single dataset, the property that makes our single-dataset surrogate estimation tractable. Randomized measurements have also been used for scalar QFI estimation [21, 27]; we instead estimate the full QFIM matrix structure on the masked support.

Rudolph et al. [22] provide a Pauli propagation framework, and Angrisani et al. [1] analyze truncated propagation in random circuits. Our setting differs in that parameters are optimization-driven and propagated operators are used to construct a metric surrogate rather than simulate observables. The masking step is also related to masked covariance estimation [7] and banded covariance regularization [3]. For finite-depth local circuits, support-overlap masks remain sparse due to bounded gate lightcones, a locality property related to but distinct from Lieb–Robinson bounds [17].

3 Background and Setup

3.1 Circuit model and propagated generators

We consider an L-layer variational circuit, $U(\overrightarrow{\theta}) = U_L(\theta_L) \cdots U_1(\theta_1)$, where each layer operator is given by:

$$U_j(\theta_j) = \begin{cases} \exp\{i\theta_j G_j\}, & b_j = 1, \\ V_j, & b_j = 0. \end{cases}$$

Here, $\overrightarrow{b} = (b_1, \ldots, b_L) \in \{0,1\}^L$ is a fixed binary mask indicating trainable layers, $G_j \in \{I, X, Y, Z\}^{\otimes n}$ are Pauli-string generators for trainable gates, and V_j are fixed unitaries. For pure input $\rho_{\text{in}} = |\psi_{\text{in}}\rangle\langle\psi_{\text{in}}|$, define:

$$|\psi\rangle = |\psi(\overrightarrow{\theta})\rangle = U(\overrightarrow{\theta})\,|\psi_{\text{in}}\rangle\,, \qquad \rho = \rho(\overrightarrow{\theta}) = |\psi\rangle\langle\psi|\,.$$

For a trainable index j, we write:

$$U(\overrightarrow{\theta}) = U_{>j} U_j U_{<j}, \qquad U_{>j} = U_L \cdots U_{j+1}, \qquad U_{<j} = U_{j-1} \cdots U_1.$$

Then

$$\frac{\partial}{\partial\theta_j}|\psi\rangle = i\widetilde{G}_j\,|\psi\rangle\,, \qquad \widetilde{G}_j = U_{>j} G_j U_{>j}^\dagger, \tag{1}$$

where $\widetilde{G}_j$ is the propagated generator. Although G_j is a single Pauli string, $\widetilde{G}_j$ is generally a Hermitian operator admitting a Pauli expansion [22]. The derivation is deferred to Appendix A.1.

3.2 QFIM as covariance

For pure-state circuits, the QFIM is the pullback of the Fubini–Study metric [5] and admits [25, 18]

$$F_{jl} = 4\Re\left[\left\langle\frac{\partial\psi}{\partial\theta_j}\middle|\frac{\partial\psi}{\partial\theta_l}\right\rangle - \left\langle\frac{\partial\psi}{\partial\theta_j}\middle|\psi\right\rangle\left\langle\psi\middle|\frac{\partial\psi}{\partial\theta_l}\right\rangle\right]. \tag{2}$$

Using Equation (1), the QFIM can be written directly in terms of propagated-generator covariance.

Lemma 3.1 (QFIM via propagated-generator covariance)**.** *For a pure-state variational circuit with single-parameter gates, the QFIM entries satisfy:*

$$F_{jl} = 4\Re\left(Cov_\rho(\widetilde{G}_j, \widetilde{G}_l)\right),$$

where $Cov_\rho(A,B) = \langle AB\rangle_\rho - \langle A\rangle_\rho\langle B\rangle_\rho$ *and* $\langle O\rangle_\rho = tr\,(\rho O)$.

Proof. Substituting Equation (1) into Equation (2) gives the result. Full details are deferred to Appendix A.2. □

This covariance form makes the QFIM a natural target for classical-shadow estimation [13].

4 Locality-Aware Surrogate and Shadow Estimation

Building on Lemma 3.1, we construct a tractable surrogate and estimator. Direct estimation of all p^2 QFIM entries faces two obstacles: (i) under local-Pauli shadows, variance grows exponentially with Pauli weight, and (ii) pairwise estimation is expensive at scale. We address both via truncation followed by masking. Figure 1 gives an overview of the resulting pipeline. The quantities shown in the diagram are defined in the following subsections. A worked example illustrating the full surrogate construction is given in Appendix C.4.

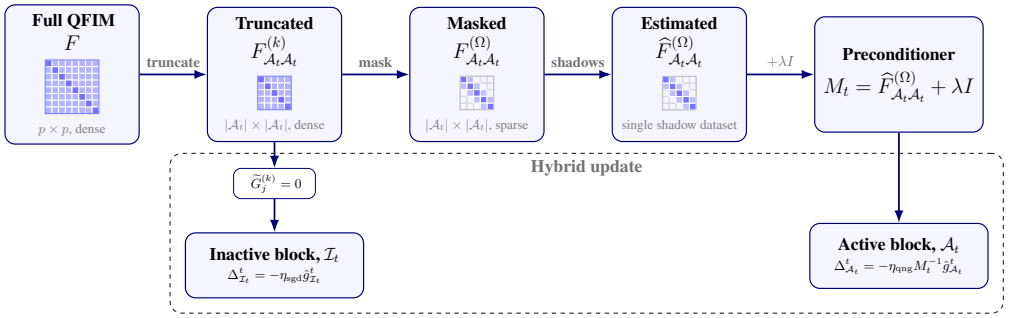


Figure 1: Sparse shadow-based QNG pipeline. Truncation determines the active parameters, support-overlap masking selects the retained metric entries, and shadow estimation constructs the active-block preconditioner. Parameters with zero truncated generators receive an SGD update.

4.1 Truncation and masking

Truncation. Since $\widetilde{G}_j$ generally expands into a sum of Pauli strings of increasing weight, estimating its covariance directly can be costly. Weight-k truncation addresses this by projecting onto a low-weight subspace: for any Hermitian $A = \sum_P c_P P$, where $\mathrm{wt}(P) = |\{i : P_i \neq I\}|$ denotes the Pauli weight of the Pauli string P,

$$T_k(A) = \sum_{\mathrm{wt}(P) \leq k} c_P P, \tag{3}$$

giving $\widetilde{G}^{(k)}_j = T_k(\widetilde{G}_j)$. This is a standard approximation in the Pauli propagation literature [22]. Since each Pauli term appearing in $\frac{1}{2}\left\{\widetilde{G}^{(k)}_j, \widetilde{G}^{(k)}_l\right\}$ has weight at most $2k$, truncation reduces worst-case shadow variance from exponential in n to exponential in k, at the cost of a controlled truncation bias discussed in Section 5.

Masking. Truncation can reduce the effective dimensionality, because some truncated generators may become zero and those parameters carry no surrogate curvature information. However, the block over the remaining parameters can still be dense, so pairwise estimation is still quadratic in the remaining active dimension. Truncation also determines the qubit support of each $\widetilde{G}^{(k)}_j$, which we use to define a sparsity pattern. By Lemma 3.1, each QFIM entry is four times the real part of the covariance between two propagated generators. After truncation, $\mathrm{supp}(\widetilde{G}^{(k)}_j)$ identifies the qubits on which the retained generator acts. If two retained supports overlap, the corresponding generators act on at least one common qubit, making their covariance a natural candidate for a local cross-parameter curvature coupling. We therefore retain only pairs whose truncated generators have overlapping support:

$$\Omega_k = \left\{(j,l) : \mathrm{supp}(\widetilde{G}^{(k)}_j) \cap \mathrm{supp}(\widetilde{G}^{(k)}_l) \neq \emptyset\right\} \cup \{(j,j) : j \in [p]\}, \tag{4}$$

setting all other entries to zero. For local circuits with nearest-neighbor structure, this mask is typically sparse. Retaining all disjoint-support pairs would increase the number of retained metric entries and the worst-case finite-shot estimation burden. This follows the general principle of masked covariance estimation [7], and Ω_k is deterministic given $(\overrightarrow{\theta}, k)$, independent of shadow sampling. We emphasize that support masking encodes a computational prior rather than a physical claim. The off-mask covariances may carry genuine curvature information that is discarded by design. The masking diagnostics in Table 3 confirm that, conditional on truncation, the mask retains above 97.6% of the Frobenius energy of $F^{(k)}$ across all k, providing empirical support for this prior.

4.2 Surrogate metric

With truncated generators in hand, we define the surrogate metric. Since each G_j is a Pauli, T_k preserves Hermiticity (Lemma C.1), so $\widetilde{G}_j^{(k)}$ is Hermitian for all j. This lets us rewrite covariances using the anticommutator $\{A, B\} = AB + BA$ via $\Re\langle AB\rangle_\rho = \frac{1}{2}\langle\{A, B\}\rangle_\rho$ for Hermitian A, B (Appendix C.1), which is the form amenable to shadow estimation.

Definition 4.1 (Truncated covariance metric)**.** Let $\rho = \rho(\overrightarrow{\theta})$. Define $F^{(k)} \in \mathbb{R}^{p\times p}$ by

$$\begin{aligned} F_{jl}^{(k)} &= 4\Re\left(\langle\widetilde{G}_j^{(k)}\widetilde{G}_l^{(k)}\rangle_\rho - \langle\widetilde{G}_j^{(k)}\rangle_\rho\langle\widetilde{G}_l^{(k)}\rangle_\rho\right) \\ &= 4\left(\tfrac{1}{2}\langle\left\{\widetilde{G}_j^{(k)}, \widetilde{G}_l^{(k)}\right\}\rangle_\rho - \langle\widetilde{G}_j^{(k)}\rangle_\rho\langle\widetilde{G}_l^{(k)}\rangle_\rho\right), \quad \langle O\rangle_\rho = \mathrm{tr}\left(\rho O\right). \end{aligned} \tag{5}$$

The anticommutator form on the second line is important. $\frac{1}{2}\left\{\widetilde{G}_j^{(k)}, \widetilde{G}_l^{(k)}\right\}$ and $\widetilde{G}_j^{(k)}$, $\widetilde{G}_l^{(k)}$ are all Hermitian Pauli-weight operators, so their expectations are directly estimable from shadow data. The resulting matrix, $F^{(k)}$ is indeed a valid surrogate metric.

Proposition 4.2 (Variance identity)**.** *For any $v \in \mathbb{R}^p$, let $\widetilde{G}^{(k)}(v) = \sum_{j=1}^p v_j\widetilde{G}_j^{(k)}$. Then*

$$v^\top F^{(k)} v = 4\mathit{var}_\rho\left(\widetilde{G}^{(k)}(v)\right) \geq 0.$$

Hence, $F^{(k)} \succeq 0$.

The proof is given in Appendix C.2.

Applying the support mask entrywise gives the locality-aware surrogate

$$F_{jl}^{(\Omega)} = \begin{cases} F_{jl}^{(k)}, & (j,l) \in \Omega_k, \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

Note that entrywise masking does not preserve positive semidefiniteness; regularization in Section 4.3 addresses this. When $k = n$ and the mask is dense, $F^{(\Omega)}$ recovers the full QFIM covariance form. In the practical regime of finite k and local circuits, $F^{(\Omega)}$ is a sparse, lower-variance approximation that retains useful curvature structure. Since each entry involves expectations of bounded-weight Pauli operators on $\rho(\overrightarrow{\theta})$, the surrogate is directly estimable from classical shadow data [13].

4.3 Shadow estimation of the locality-aware surrogate

We now describe how to estimate $F^{(\Omega)}$ from a single shadow dataset. We adopt local Pauli measurements, where each qubit is independently measured in a uniformly random Pauli basis. From N snapshots $\mathcal{D}_t = \{(U_s, b_s)\}_{s=1}^N$ collected at the current iterate $\overrightarrow{\theta}^t$, the shadow operators $\hat{\rho}_s = U_s^\dagger\left(\bigotimes_{i=1}^n(3\left|b_{s,i}\right\rangle\left\langle b_{s,i}\right| - I)\right)U_s$ satisfy $\mathbb{E}_{\mathcal{D}_t}\left[\frac{1}{N}\sum_{s=1}^N \mathrm{tr}\left(O\hat{\rho}_s\right)\right] = \mathrm{tr}\left(O\rho(\overrightarrow{\theta}^t)\right)$ for any observable O [13]. Setting $X_s(O) = \mathrm{tr}\left(O\hat{\rho}_s\right)$, each entry of $F^{(\Omega)}$ requires two quantities: the anti-commutator expectation $\frac{1}{2}\langle\left\{\widetilde{G}_{j,t}^{(k)}, \widetilde{G}_{l,t}^{(k)}\right\}\rangle_\rho$ and the product of means $\langle\widetilde{G}_{j,t}^{(k)}\rangle_\rho\langle\widetilde{G}_{l,t}^{(k)}\rangle_\rho$. We estimate

these from the same dataset using:

$$\bar{\mu}_{\text{sym}}(j,l) = \frac{1}{N}\sum_{s=1}^{N} X_s\left(\tfrac{1}{2}\left\{\widetilde{G}^{(k)}_{j,t}, \widetilde{G}^{(k)}_{l,t}\right\}\right),$$

$$\bar{\mu}_{\text{prod}}(j,l) = \frac{1}{N(N-1)}\sum_{\substack{s,r=1\\ s\neq r}}^{N} X_s\left(\widetilde{G}^{(k)}_{j,t}\right) X_r\left(\widetilde{G}^{(k)}_{l,t}\right). \tag{7}$$

where $\bar{\mu}_{\text{prod}}$ is the cross-fit U-statistic estimator for the product of means. It is unbiased because distinct shadow snapshots are independent and each shadow feature is unbiased [13]. The masked estimator is then given by:

$$\widehat{F}^{(\Omega)}_{jl} = \begin{cases} 4\big(\bar{\mu}_{\text{sym}}(j,l) - \bar{\mu}_{\text{prod}}(j,l)\big), & (j,l)\in\Omega_{k,t},\\ 0, & \text{otherwise,}\end{cases} \tag{8}$$

and satisfies $\mathbb{E}_{\mathcal{D}_t}\left[\widehat{F}^{(\Omega)}(\overrightarrow{\theta}^t) \mid \overrightarrow{\theta}^t\right] = F^{(\Omega)}(\overrightarrow{\theta}^t)$ (using Lemma C.2). The key practical advantage is that a single dataset $\mathcal{D}_t$ of N snapshots suffices to estimate all on-mask entries simultaneously, without dedicated circuits per pair. Details, including the $O(N^2) \to O(N)$ reduction for $\bar{\mu}_{\text{prod}}$, are presented in Appendix C.3.

Active and inactive parameters. We define the active and inactive parameters as $\mathcal{A}_t = \left\{j : \widetilde{G}^{(k)}_{j,t} \neq 0\right\}$ and $\mathcal{I}_t = [p] \setminus \mathcal{A}_t$. Parameters in $\mathcal{I}_t$ have no curvature information in $F^{(\Omega)}$ and are updated by standard gradient descent, while parameters in $\mathcal{A}_t$ receive surrogate-QNG updates on the restricted block

$$\Omega^{\mathcal{A}_t}_{k,t} = \Omega_{k,t} \cap (\mathcal{A}_t \times \mathcal{A}_t), \qquad \widehat{F}^{(\Omega)}_{\mathcal{A}_t\mathcal{A}_t} = \big(\widehat{F}^{(\Omega)}_{jl}\big)_{j,l\in\mathcal{A}_t}.$$

Since $\mathcal{A}_t$ and $\Omega_{k,t}$ are deterministic given $(\overrightarrow{\theta}^t, k)$, restricting to $j,l \in \mathcal{A}_t$ preserves unbiasedness entrywise (Lemma C.3).

Although $F^{(k)} \succeq 0$ by Proposition 4.2, entrywise masking and finite-sample shadow noise can introduce small negative eigenvalues in $\widehat{F}^{(\Omega)}_{\mathcal{A}_t\mathcal{A}_t}$ in practice. We therefore regularize with a ridge term, defining the preconditioner as

$$M_t = \widehat{F}^{(\Omega)}_{\mathcal{A}_t\mathcal{A}_t} + \lambda I, \qquad \lambda > 0.$$

By the eigenvalue shift identity, choosing $\lambda > -\lambda_{\min}(\widehat{F}^{(\Omega)}_{\mathcal{A}_t\mathcal{A}_t})$ ensures $M_t \succ 0$. Beyond ensuring invertibility, the ridge term smoothly interpolates back to gradient descent in flat regions where the surrogate carries little reliable curvature information. Algorithm 1 summarizes the full procedure.

Algorithm 1 Locality-Aware Shadow-Hybrid QNG

Require: Circuit $U(\overrightarrow{\theta})$, Cost C, truncation level k, ridge λ, stepsizes $(\eta_{\text{qng}}, \eta_{\text{sgd}})$

1: Initialize $\overrightarrow{\theta}^0$
2: **for** $t = 0, \ldots, T-1$ **do**
3: Estimate the finite-shot gradient $\hat{g}_t$, an estimator of the true gradient $g_t = \nabla C(\overrightarrow{\theta}^t)$ via parameter-shift.
4: Build propagated generators $\left\{\widetilde{G}_{j,t}\right\}_{j=1}^{p}$ and truncate $\widetilde{G}^{(k)}_{j,t} = T_k(\widetilde{G}_{j,t})$
5: Define active/inactive sets $\mathcal{A}_t = \left\{j : \widetilde{G}^{(k)}_{j,t} \neq 0\right\}$, $\mathcal{I}_t = [p] \setminus \mathcal{A}_t$
6: Build support mask $\Omega_{k,t}$ by Equation (4), and restrict $\Omega^{\mathcal{A}_t}_{k,t} = \Omega_{k,t} \cap (\mathcal{A}_t \times \mathcal{A}_t)$
7: Collect N shadow snapshots $\{\hat{\rho}_s\}_{s=1}^{N}$ at $\overrightarrow{\theta}^t$
8: Estimate $\widehat{F}^{(\Omega)}_{\mathcal{A}_t\mathcal{A}_t}$ via Equations (7) and (8)
9: Set preconditioner $M_t = \widehat{F}^{(\Omega)}_{\mathcal{A}_t\mathcal{A}_t} + \lambda I$
10: Update: $\overrightarrow{\theta}^{t+1}_{\mathcal{A}_t} = \overrightarrow{\theta}^{t}_{\mathcal{A}_t} - \eta_{\text{qng}} M_t^{-1}\hat{g}^t_{\mathcal{A}_t}$, $\qquad \overrightarrow{\theta}^{t+1}_{\mathcal{I}_t} = \overrightarrow{\theta}^{t}_{\mathcal{I}_t} - \eta_{\text{sgd}}\hat{g}^t_{\mathcal{I}_t}$
11: **end for**

5 Approximation and Convergence Guarantees

We condition on $\overrightarrow{\theta}^t$ throughout, so $\mathcal{A}_t$ and $\Omega_{k,t}^{\mathcal{A}_t}$ are deterministic functions of $(\overrightarrow{\theta}^t, k)$. Algorithm 1 applies surrogate-QNG updates only to active parameters $\mathcal{A}_t$, while inactive parameters $\mathcal{I}_t$ are updated by standard gradient descent with no metric approximation. The analysis below therefore focuses only on the active block, as the inactive update introduces no additional approximation error beyond finite-shot gradient noise. The construction of the preconditioner can be visualized as follows (see Figure 1 for a schematic overview).

$$F \xrightarrow{\text{truncation}} F^{(k)}_{\mathcal{A}_t\mathcal{A}_t} \xrightarrow{\text{masking}} F^{(\Omega)}_{\mathcal{A}_t\mathcal{A}_t} \xrightarrow{\text{shadow estimation}} \widehat{F}^{(\Omega)}_{\mathcal{A}_t\mathcal{A}_t} \xrightarrow{\text{ridge}} M_t. \tag{9}$$

5.1 Approximation error decomposition

Each step in Equation (9) introduces an error relative to the true QFIM. On the active block, define the truncation and masking approximation errors as:

$$E_{\text{trunc},t} = \|F_{\mathcal{A}_t\mathcal{A}_t} - F^{(k)}_{\mathcal{A}_t\mathcal{A}_t}\|_{\text{op}}, \qquad E_{\text{mask},t} = \|F^{(k)}_{\mathcal{A}_t\mathcal{A}_t} - F^{(\Omega)}_{\mathcal{A}_t\mathcal{A}_t}\|_{\text{op}}.$$

These are deterministic conditional on $\overrightarrow{\theta}^t$, depending on the circuit structure and the design choices $(k, \Omega_{k,t})$, but not on the shadow dataset. With $\widehat{F}^{(\Omega)}_{\mathcal{A}_t\mathcal{A}_t}$ estimated from shadows, let $E_{\text{shot},t} = \|\widehat{F}^{(\Omega)}_{\mathcal{A}_t\mathcal{A}_t} - F^{(\Omega)}_{\mathcal{A}_t\mathcal{A}_t}\|_{\text{op}}$ denote the finite-shot estimation error. Triangle inequality gives:

$$\|\widehat{F}^{(\Omega)}_{\mathcal{A}_t\mathcal{A}_t} - F_{\mathcal{A}_t\mathcal{A}_t}\|_{\text{op}} \le E_{\text{trunc},t} + E_{\text{mask},t} + E_{\text{shot},t}. \tag{10}$$

The finite-shot analysis below controls $E_{\text{shot},t}$. The remaining two terms are deterministic biases arising from the surrogate construction. The truncation error reflects how much of each propagated generator is discarded. It is controlled by the largest residual norm after weight-k projection, vanishing once k exceeds the maximum Pauli weight in the propagated generator; for local nearest-neighbor circuits, this is bounded by the circuit lightcone. The masking error is governed by the magnitude of the off-mask covariances discarded by the support-overlap criterion. These terms are bounded as follows (details in Appendix D):

$$E_{\text{trunc},t} \le 4|\mathcal{A}_t|\tau_{\max}(2+\tau_{\max}), \qquad E_{\text{mask},t} \le 4d^{\text{off}}_{\Omega,t} \max_{\substack{(j,l)\notin\Omega^{\mathcal{A}_t}_{k,t} \\ j\neq l, j,l\in\mathcal{A}_t}} \left|\text{Cov}_\rho\left(\widetilde{G}^{(k)}_j, \widetilde{G}^{(k)}_l\right)\right|, \tag{11}$$

where $\tau_{\max} = \max_{j\in\mathcal{A}_t} \|\widetilde{G}_j - \widetilde{G}^{(k)}_j\|_{\text{op}}$ and $d^{\text{off}}_{\Omega,t}$ is the maximum number of off-mask entries per row. Both bounds are worst-case. In particular, the masking diagnostics in Table 3 indicate that the practical loss from support masking is modest in the regimes we study.

5.2 Finite-shot estimation

We now bound the estimation error $E_{\text{shot},t}$. Conditioned on $\overrightarrow{\theta}^t$, the estimators $\bar{\mu}_{\text{sym}}(j,l)$ and $\bar{\mu}_{\text{prod}}(j,l)$ defined in Equation (7) are respectively a sample mean and a degree-2 cross-fit U-statistic, both unbiased for their population counterparts. Their per-snapshot variances V^{sym}_{jl} and V^{prod}_{jl} are N-independent. Full definitions and the variance decomposition are provided in Appendix E. To state the concentration bound compactly, define

$$V_\star = \max_{(j,l)\in\Omega^{\mathcal{A}_t}_{k,t}} \max\left\{V^{\text{sym}}_{jl}, V^{\text{prod}}_{jl}\right\}, \qquad B_\star = \max\left\{B_{\text{sym}}, B_A^2\right\}, \tag{12}$$

where B_{sym}, B_A bound the shadow features $\left|X_s(\frac{1}{2}\left\{\widetilde{G}^{(k)}_{j,t}, \widetilde{G}^{(k)}_{l,t}\right\})\right|$ and $\left|X_s(\widetilde{G}^{(k)}_{j,t})\right|$ respectively. Let $d_{\Omega,t} = \max_{j\in\mathcal{A}_t}\sum_{l\in\mathcal{A}_t} \mathbf{1}\left\{(j,l)\in\Omega^{\mathcal{A}_t}_{k,t}\right\}$ denote the maximum on-mask row degree.

Theorem 5.1 (Operator-norm concentration)**.** *Assume $N \ge 2$. Conditioned on $\overrightarrow{\theta}^t$, for any $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$\left\|\widehat{F}^{(\Omega)}_{\mathcal{A}_t\mathcal{A}_t} - F^{(\Omega)}_{\mathcal{A}_t\mathcal{A}_t}\right\|_{op} \le d_{\Omega,t}\left[C_1\sqrt{\frac{V_\star\log(4|\Omega^{\mathcal{A}_t}_{k,t}|/\delta)}{N}} + C_2\frac{B_\star\log(4|\Omega^{\mathcal{A}_t}_{k,t}|/\delta)}{N}\right], \tag{13}$$

where $C_1 = 4(\sqrt{2}+\sqrt{3})$, $C_2 = 64/3$.

Proof. Applying entrywise concentration to $\bar{\mu}_{\text{sym}}$ and $\bar{\mu}_{\text{prod}}$, combining via a union bound over $\Omega_{k,t}^{\mathcal{A}_t}$, then converting entrywise control to operator norm using mask sparsity ($d_{\Omega,t}$) gives the desired result. Full details are in Appendix E.6. □

Corollary 5.2 (Per-iteration sample complexity)**.** *To ensure* $\|\widehat{F}^{(\Omega)}_{\mathcal{A}_t\mathcal{A}_t} - F^{(\Omega)}_{\mathcal{A}_t\mathcal{A}_t}\|_{op} \le \varepsilon$ *with probability at least* $1-\delta$*, it suffices that*

$$N = \mathcal{O}\left(\frac{V_\star d_{\Omega,t}^2 \log\left(|\Omega_{k,t}^{\mathcal{A}_t}|/\delta\right)}{\varepsilon^2} + \frac{B_\star d_{\Omega,t} \log\left(|\Omega_{k,t}^{\mathcal{A}_t}|/\delta\right)}{\varepsilon}\right). \tag{14}$$

Explicit constants and the variance-dominated simplification are presented in Appendix E.7.

The k-dependence enters through $V_\star$ and $B_\star$: fixed k avoids exponential scaling with system size, while larger k trades more measurements for richer curvature information.

5.3 Preconditioning and convergence

For each iteration t, let $M_t = \widehat{F}^{(\Omega)}_{\mathcal{A}_t\mathcal{A}_t} + \lambda I$ with $\lambda > 0$. Denote the true gradient by $g^t = \nabla C(\overrightarrow{\theta}^t)$ and its finite-shot parameter-shift estimate by $\hat{g}^t = g^t + \xi^t$, with $\mathbb{E}\left[\xi^t \mid \overrightarrow{\theta}^t\right] = 0$ and $\mathbb{E}\left[\|\xi^t\|_2^2 \mid \overrightarrow{\theta}^t\right] \le \sigma_g^2$. Since gradients and metrics are estimated from disjoint shot batches, $\xi^t \perp M_t \mid \overrightarrow{\theta}^t$. The hybrid updates on the active and inactive blocks are given by:

$$\Delta^t_{\mathcal{A}_t} = -\eta_{\text{qng}} M_t^{-1} \hat{g}^t_{\mathcal{A}_t}, \qquad \Delta^t_{\mathcal{I}_t} = -\eta_{\text{sgd}} \hat{g}^t_{\mathcal{I}_t}.$$

For the preconditioner to remain stable throughout training, we require the estimation error to stay uniformly bounded. Define the good event $\mathcal{E}_T = \bigcap_{t=0}^{T-1} \left\{\|\widehat{F}^{(\Omega)}_{\mathcal{A}_t\mathcal{A}_t} - F^{(\Omega)}_{\mathcal{A}_t\mathcal{A}_t}\|_{\text{op}} \le \varepsilon\right\}$. Applying Corollary 5.2 at each iteration with failure probability δ/T and taking a union bound gives $\Pr(\mathcal{E}_T) \ge 1-\delta$, provided N satisfies the uniform sample condition in Appendix F.1. On $\mathcal{E}_T$, Weyl's inequality provides a sufficient route to verifying bounds of the form $mI \preceq M_t \preceq \Lambda_{\max} I$ for all t, with $m > 0$ and $\Lambda_{\max} < \infty$. See Appendix F.2 for complete expressions.

Theorem 5.3 (Expected one-step descent)**.** *Assume* C *is* β*-smooth. Fix iteration* t*, and suppose* $mI \preceq M_t \preceq \Lambda_{\max} I$*. If* $\eta_{qng} \le m/\beta$ *and* $\eta_{sgd} \le 1/\beta$*, then*

$$\mathbb{E}\left[C(\overrightarrow{\theta}^{t+1}) \mid \overrightarrow{\theta}^t, M_t\right] \le C(\overrightarrow{\theta}^t) - \frac{\eta_{qng}}{2} g^{t\top}_{\mathcal{A}_t} M_t^{-1} g^t_{\mathcal{A}_t} - \frac{\eta_{sgd}}{2}\|g^t_{\mathcal{I}_t}\|_2^2 + \Sigma,$$

where $\Sigma = \frac{\beta}{2}\left(\eta_{qng}^2 \sigma_g^2/m^2 + \eta_{sgd}^2 \sigma_g^2\right)$.

Theorem 5.4 (Convergence to a stationary neighborhood assuming uniform spectral bounds)**.** *Assume additionally that* C *is bounded below by* C_{inf} *and that* $mI \preceq M_t \preceq \Lambda_{\max} I$*,* $\forall t = 0, \dots, T-1$*. Then*

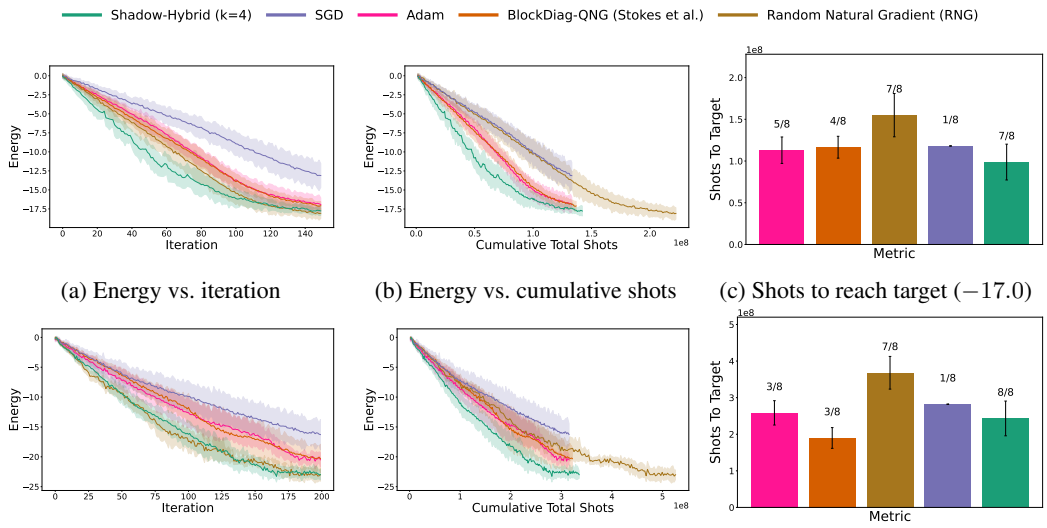
$$\min_{t=0,\dots,T-1} \mathbb{E}\left[\|g^t\|_2^2\right] \le \frac{2(C(\overrightarrow{\theta}^0) - C_{\text{inf}})}{T\eta_{\min}} + \frac{2\Sigma}{\eta_{\min}}, \qquad \eta_{\min} = \min\left\{\frac{\eta_{qng}}{\Lambda_{\max}}, \eta_{sgd}\right\}.$$

Both proofs follow standard smooth nonconvex stochastic optimization arguments; see, e.g., Ghadimi and Lan [10] and Bottou et al. [4]. Full details are given in Appendix F. The concentration result together with the union bound above provides a pathwise sufficient route to the uniform spectral bounds along a trajectory. Theorem 5.4 is conditional on this uniform spectral control and guarantees convergence to a stationary neighborhood whose size depends on Σ, rather than global optimality. This analysis uses the ridge-only preconditioner. In experiments, we apply positive semidefinite projection as a numerical stabilizer, which only tightens the lower spectral bound and ensures $M_t \succeq \lambda I$ (Appendix F.5).

6 Experiments

6.1 Setup

We evaluate our proposed Locality-Aware Shadow-Hybrid QNG on VQE instances for the ground-state energy of the 1D Heisenberg chain $H = \sum_{i=0}^{n-2}(X_iX_{i+1} + Y_iY_{i+1} + Z_iZ_{i+1})$ with open



(a) Energy vs. iteration (b) Energy vs. cumulative shots (c) Shots to reach target (-17.0)

(d) Energy vs. iteration (e) Energy vs. cumulative shots (f) Shots to reach target (-22.0)

Figure 2: Convergence on 1D Heisenberg chains over 8 matched random seeds. Top row: 12-qubit, depth-6. Bottom row: 16-qubit, depth-8. Columns show energy vs. iteration, energy vs. cumulative shots, and shots to reach a preset benchmark target (-17.0 for the top row, -22.0 for the bottom row). Shaded bands in the iteration and cumulative shot panels show pointwise 95% confidence intervals over seeds. The shots-to-target panels report mean $\pm$ SD over successful seeds only: fractions above bars indicate the number of seeds reaching the target. Targets are fixed benchmarking thresholds shared across methods and seeds, not true ground-state energies. Energies at fixed shot budgets are reported in Figure 3 (Appendix G.1).

boundary conditions. In all experiments, the cost observable is $\mathcal{M} = H$. We use a hardware-efficient ansatz with nearest-neighbor CNOT entanglement, where each layer consists of trainable single-qubit rotations followed by linear entanglement (Figure 8); we denote by depth d the number of repeated HEA layers. Shadow-Hybrid uses weight-k truncation with $k = 4$, selected as a fixed operating point from a pilot sweep over $k \in \{2, 3, 4\}$ on the 12-qubit instance (Appendix G.5), reflecting the tradeoff between curvature fidelity and estimation noise. We use a PSD-projected ridge preconditioner with $\lambda = 0.1$ on the active block. We compare against SGD, Adam, BlockDiag-QNG [25], and RNG [16]. We also include a direct comparison with the locality-aware WA-QNG baseline in Appendix G.3. All methods share the same finite-shot energy and gradient estimation pipeline (parameter-shift). The metric-based methods incur additional shots only for preconditioner estimation. Results on the Ising Hamiltonian and additional ablations are in Appendix G. Shot budgets, including shadow snapshots for surrogate estimation, and further implementation details are summarized in Appendix H. Additional small instance diagnostics, including a comparison with the dense exact QFIM, component-wise error decomposition, and an all-to-all connectivity test, are reported in Appendix G.6.

6.2 Results

At 12 qubits (top row of Figure 2), Shadow-Hybrid reaches lower mean energy at matched cumulative shot budgets than all baselines (Figure 2b), while remaining competitive per iteration (Figure 2a). The shots-to-target panel gives a complementary summary of target attainment. Because shots-to-target averages are conditioned on successful runs, we report success counts alongside them. Shadow-Hybrid attains the preset benchmark target in 7 of 8 seeds, matching RNG; Adam succeeds in 5 of 8 seeds, BlockDiag-QNG in 4 of 8, and SGD in 1 of 8. Among the methods with the same success rate, Shadow-Hybrid requires the lowest average cumulative shot cost.

The same trend persists at 16 qubits (bottom row of Figure 2). Shadow-Hybrid again gives the strongest cumulative-shot performance (Figure 2e) and is the only method to reach the target in all 8 seeds. RNG succeeds in 7 of 8 seeds, Adam and BlockDiag-QNG each succeed in 3 of 8, and SGD succeeds in 1 of 8. Thus, the larger instance preserves the main conclusion from the 12-qubit case: Shadow-Hybrid combines strong target-attainment reliability with shot-efficient convergence.

Ablations. We report ablations and diagnostics for the 12-qubit Heisenberg instance in Appendix G.5. A sweep over truncation levels k shows the expected bias–variance tradeoff (Figure 7a): smaller k converges faster early but plateaus sooner, while larger k reaches lower final energies with wider variance bands. As same-ridge controls, we compare against Ridge-SGD and Hybrid-Ridge-SGD, which replace the surrogate by $M_t = \lambda I$ globally or on the active set, respectively (Figures 7b and 7c). Both ridge variants oscillate and stall, whereas Shadow-Hybrid continues to converge, indicating that neither stepsize scaling alone nor the active/inactive split alone reproduce the observed behavior. Spectral diagnostics further confirm that the surrogate carries non-trivial curvature: its dominant eigenvalue exceeds the ridge floor by $\sim 25\times$, and the preconditioned update direction differs substantially from the ridge-only direction (cosine similarity ~ 0.61); see Table 4.

7 Discussion

We have presented a locality-aware surrogate for quantum natural gradient that combines weight-k truncation, support-overlap masking, and classical-shadow estimation into a single-dataset, shadow-based preconditioner. The method is accompanied by finite-shot concentration bounds and convergence to a stationary neighborhood under standard assumptions. Integrating these components yields a preconditioner that captures meaningful curvature structure while remaining numerically stable and practical under finite-shot budgets. This is reflected in our spectral diagnostics, where the dominant eigenvalue substantially exceeds the ridge floor and the preconditioned update direction differs markedly from the ridge-only baseline. Empirically, the resulting optimizer achieves improved convergence relative to strong baselines, with the clearest gains when additional curvature information justifies its estimation overhead.

Several directions remain open. The truncation level k is fixed in this work; adaptive selection of k is a natural extension suggested by the crossover behavior in Figure 7a. The worst-case truncation bounds (Section 5.1) also appear loose in practice. Our main optimization experiments focus on local ansatzes with nearest-neighbor entanglement. Bounded-degree local circuits at moderate depth provide a favorable regime, where lightcones can limit propagated-generator spreading and keep the retained Pauli terms and support-overlap structure manageable. Our small-instance all-to-all diagnostic (Appendix G.6.3) shows a connectivity-dependent operating-point tradeoff. On this instance, increasing k improves the approximation by retaining more propagated-generator terms. Studying larger and deeper circuits with long-range or dense connectivity remains future work and may require larger k or alternative sparsity structures. Finally, because the covariance-based surrogate remains well-defined for mixed states, extending the framework to noisy circuits is a natural next step, along with corresponding finite-shot concentration bounds.

Acknowledgments and Disclosures

This work is partially supported by the NSF Center for Science of Information (CSoI) Grant CCF0939370, and by NSF Grants CCF-2006440 and CCF-2211423. The authors declare no competing interests.

References

[1] Armando Angrisani, Alexander Schmidhuber, Manuel S Rudolph, Marco Cerezo, Zoë Holmes, and Hsin-Yuan Huang. Classically estimating observables of noiseless quantum circuits. *Physical Review Letters*, 135(17):170602, 2025.

[2] Ville Bergholm, Josh Izaac, Maria Schuld, Christian Gogolin, Shahnawaz Ahmed, Vishnu Ajith, M Sohaib Alam, Guillermo Alonso-Linaje, Bharath Akash Narayanan, Ali Asadi, et al. Pennylane: Automatic differentiation of hybrid quantum-classical computations. *arXiv preprint arXiv:1811.04968*, 2018.

[3] Peter J Bickel and Elizaveta Levina. Regularized estimation of large covariance matrices. 2008.

[4] Léon Bottou, Frank E Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *SIAM Review*, 60(2):223–311, 2018.

[5] Samuel L Braunstein and Carlton M Caves. Statistical distance and the geometry of quantum states. *Physical Review Letters*, 72(22):3439, 1994.

[6] Marco Cerezo, Andrew Arrasmith, Ryan Babbush, Simon C Benjamin, Suguru Endo, Keisuke Fujii, Jarrod R McClean, Kosuke Mitarai, Xiao Yuan, Lukasz Cincio, et al. Variational quantum algorithms. *Nature Reviews Physics*, 3(9):625–644, 2021.

[7] Richard Y Chen, Alex Gittens, and Joel A Tropp. The masked sample covariance estimator: an analysis using matrix concentration inequalities. *Information and Inference: A Journal of the IMA*, 1(1):2–20, 2012.

[8] Andreas Elben, Steven T Flammia, Hsin-Yuan Huang, Richard Kueng, John Preskill, Benoît Vermersch, and Peter Zoller. The randomized measurement toolbox. *Nature Reviews Physics*, 5(1):9–24, 2023.

[9] Julien Gacon, Christa Zoufal, Giuseppe Carleo, and Stefan Woerner. Simultaneous perturbation stochastic approximation of the quantum fisher information. *Quantum*, 5:567, 2021.

[10] Saeed Ghadimi and Guanghui Lan. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM journal on optimization*, 23(4):2341–2368, 2013.

[11] Brian C. Hall. Lie groups, lie algebras, and representations. In *Quantum Theory for Mathematicians*, pages 333–366. Springer, 2013.

[12] Wassily Hoeffding. A class of statistics with asymptotically normal distribution. In *Breakthroughs in statistics: Foundations and basic theory*, pages 308–334. Springer, 1992.

[13] Hsin-Yuan Huang, Richard Kueng, and John Preskill. Predicting many properties of a quantum system from very few measurements. *Nature Physics*, 16(10):1050–1057, 2020.

[14] Bálint Koczor and Simon C Benjamin. Quantum analytic descent. *Physical Review Research*, 4(2):023017, 2022.

[15] Bálint Koczor and Simon C Benjamin. Quantum natural gradient generalized to noisy and nonunitary circuits. *Physical Review A*, 106(6):062416, 2022.

[16] Ioannis Kolotouros and Petros Wallden. Random natural gradient. *Quantum*, 8:1503, 2024.

[17] Elliott H Lieb and Derek W Robinson. The finite group velocity of quantum spin systems. *Communications in mathematical physics*, 28(3):251–257, 1972.

[18] Johannes Jakob Meyer. Fisher information in noisy intermediate-scale quantum applications. *Quantum*, 5:539, 2021.

[19] Thomas Peel, Sandrine Anthoine, and Liva Ralaivola. Empirical Bernstein inequalities for U-statistics. *Advances in Neural Information Processing Systems*, 23, 2010.

[20] Alberto Peruzzo, Jarrod McClean, Peter Shadbolt, Man-Hong Yung, Xiao-Qi Zhou, Peter J Love, Alán Aspuru-Guzik, and Jeremy L O'brien. A variational eigenvalue solver on a photonic quantum processor. *Nature communications*, 5(1):4213, 2014.

[21] Aniket Rath, Cyril Branciard, Anna Minguzzi, and Benoît Vermersch. Quantum Fisher information from randomized measurements. *Physical Review Letters*, 127(26):260501, 2021.

[22] Manuel S Rudolph, Tyson Jones, Yanting Teng, Armando Angrisani, and Zoë Holmes. Pauli propagation: A computational framework for simulating quantum systems. *PRX Quantum*, 7(3):032001, 2026.

[23] Chenyu Shi, Vedran Dunjko, and Hao Wang. Weighted approximate quantum natural gradient for variational quantum eigensolver. *Quantum Science and Technology*, 11(1):015060, 2026.

[24] Mohammad Aamir Sohail, Mohsen Heidari, and S Sandeep Pradhan. Quantum natural stochastic pairwise coordinate descent. *npj Quantum Information*, 11(1):109, 2025.

[25] James Stokes, Josh Izaac, Nathan Killoran, and Giuseppe Carleo. Quantum natural gradient. *Quantum*, 4:269, 2020.

[26] Barnaby van Straaten and Bálint Koczor. Measurement cost of metric-aware variational quantum algorithms. *PRX Quantum*, 2(3):030324, 2021.

[27] Min Yu, Dongxiao Li, Jingcheng Wang, Yaoming Chu, Pengcheng Yang, Musang Gong, Nathan Goldman, and Jianming Cai. Experimental estimation of the quantum Fisher information from randomized measurements. *Physical Review Research*, 3(4):043122, 2021.

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 the introduction point out the major contribution and scope of the paper.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: The limitations of the work are mentioned in Section 7.

   Guidelines:

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

3. **Theory assumptions and proofs**

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

   Answer: [Yes]

Justification: The paper provides complete proof for the theoretical results along with the assumptions where necessary.

Guidelines:

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

4. **Experimental result reproducibility**

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

   Answer: [Yes]

   Justification: Source code with reproduction instructions is provided as supplementary materials.

   Guidelines:

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

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

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

Answer: [Yes]

Justification: The source code is made available along with the submission.

Guidelines:

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

6. **Experimental setting/details**

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

Answer: [Yes]

Justification: The required information is provided along with the source code.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

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

Answer: [Yes]

Justification: The iteration, cumulative-shot, and fixed-budget plots report mean energy with 95% confidence intervals over eight matched random seeds. Shots-to-target error bars report mean $\pm$ SD over successful seeds, with success counts reported alongside them. The diagnostic tables report mean $\pm$ SD.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

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

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: The required information is provided along with the source code as supplementary material.

   Guidelines:

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

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: The authors have made a concerted effort to comply with the NeurIPS Code of Ethics.

   Guidelines:

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

10. **Broader impacts**

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

    Answer: [N/A]

    Justification: This work is not expected to have any direct societal impact, to the best of our understanding.

    Guidelines:

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

11. **Safeguards**

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

    Answer: [N/A]

    Justification: As far as we can tell, the paper does not involve any identifiable risks as mentioned above.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

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

    Answer: [Yes]

    Justification: Pennylane [2] is cited for simulation framework. Baseline methods such as BlockDiag-QNG, RNG are reimplemented following their respective references [25, 16].

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

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

    Answer: [Yes]

    Justification: The source code and relevant information are made available.

    Guidelines:

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

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

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

    Answer: [N/A]

    Justification: This work does not involve research involving human subjects or crowdsourcing experiments.

    Guidelines:

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

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

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

    Answer: [N/A]

    Justification: [N/A]

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

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

    Answer: [N/A]

    Justification: [N/A]

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

A Background

A.1 Derivation of Equation (1)

For each trainable index j, write

$$U(\overrightarrow{\theta}) = U_{>j} U_j U_{<j}, \qquad U_{>j} = U_L \cdots U_{j+1}, \qquad U_{<j} = U_{j-1} \cdots U_1.$$

Then

$$\begin{aligned}
\frac{\partial}{\partial \theta_j} |\psi\rangle &= U_{>j} \left(\frac{\partial}{\partial \theta_j} (\exp\{i\theta_j G_j\}) \right) U_{<j} |\psi_{\text{in}}\rangle \\
&= i U_{>j} G_j U_j U_{<j} |\psi_{\text{in}}\rangle \\
&= i U_{>j} G_j U_{>j}^\dagger U_{>j} U_j U_{<j} |\psi_{\text{in}}\rangle \qquad \left[U_{>j}^\dagger U_{>j} = I\right] \\
&= i U_{>j} G_j U_{>j}^\dagger U(\overrightarrow{\theta}) |\psi_{\text{in}}\rangle \\
&= i U_{>j} G_j U_{>j}^\dagger |\psi\rangle \\
&= i \widetilde{G}_j |\psi\rangle .
\end{aligned}$$

A.2 Proof of Lemma 3.1

Using Equation (1),

$$\left| \frac{\partial \psi}{\partial \theta_j} \right\rangle = i \widetilde{G}_j |\psi\rangle , \qquad \left\langle \frac{\partial \psi}{\partial \theta_j} \right| = -i \langle \psi | \widetilde{G}_j .$$

Hence

$$\begin{aligned}
\left\langle \frac{\partial \psi}{\partial \theta_j} \middle| \frac{\partial \psi}{\partial \theta_l} \right\rangle &= \langle \psi | \widetilde{G}_j \widetilde{G}_l |\psi\rangle = \langle \widetilde{G}_j \widetilde{G}_l \rangle_\rho, \\
\left\langle \frac{\partial \psi}{\partial \theta_j} \middle| \psi \right\rangle &= -i \langle \psi | \widetilde{G}_j |\psi\rangle = -i \langle \widetilde{G}_j \rangle_\rho, \\
\left\langle \psi \middle| \frac{\partial \psi}{\partial \theta_l} \right\rangle &= i \langle \psi | \widetilde{G}_l |\psi\rangle = i \langle \widetilde{G}_l \rangle_\rho,
\end{aligned}$$

where $\langle O \rangle_\rho = \text{tr}(\rho O)$. Therefore, from Equation (2),

$$\begin{aligned}
F_{jl} &= 4 \Re \left(\langle \widetilde{G}_j \widetilde{G}_l \rangle_\rho - \langle \widetilde{G}_j \rangle_\rho \langle \widetilde{G}_l \rangle_\rho \right) \\
&= 4 \Re \left(\text{Cov}_\rho (\widetilde{G}_j, \widetilde{G}_l) \right).
\end{aligned}$$

B Propagated generator calculus

To compute $\widetilde{G}_j = U_{>j} G_j U_{>j}^\dagger$, we push G_j through the gates after j using local conjugation rules.

B.1 Generic Rules

Let P be any Pauli string (possibly multi-qubit), so $P^\dagger = P$ and $P^2 = \mathbb{I}$. Define two linear maps on operators:

$$\text{Ad}_P(A) = PAP^\dagger = PAP, \quad ad_P(A) = [P, A] = PA - AP$$

Ad_P is an involution as:

$$\text{Ad}_P^2(A) = P(PAP)P = (PP)A(PP) = A.$$

So, Ad_P has eigenvalues ± 1 and is diagonalizable. The projectors onto the $\pm$ eigenspaces are:

$$\begin{aligned}
&\Pi_+ = \frac{1}{2}(\mathcal{I} + \text{Ad}_P), \qquad \Pi_- = \frac{1}{2}(\mathcal{I} - \text{Ad}_P), \\
&\Pi_\pm^2 = \Pi_\pm, \ \Pi_+ \Pi_- = 0, \ \Pi_+ + \Pi_- = \mathcal{I}.
\end{aligned}$$

Applying this for any operator A yields the commuting / anticommuting decomposition:

$$A_{\|} = \Pi_{+}(A) = \frac{1}{2}(A + \mathrm{Ad}_P(A)), \qquad A_{\perp} = \Pi_{-}(A) = \frac{1}{2}(A - \mathrm{Ad}_P(A)).$$

By construction we get:

$$\begin{aligned}\left[P, A_{\|}\right] = 0 &\iff P A_{\|} P = A_{\|} \\ \{P, A_{\perp}\} = 0 &\iff P A_{\perp} P = -A_{\perp}.\end{aligned}$$

Lemma B.1 (Hadamard lemma ([11]))**.** *For any operators X,A, the adjoint action satisfies*

$$\exp\{X\} A \exp\{-X\} = \sum_{n=0}^{\infty} \frac{1}{n!} ad_X^n(A) = \exp\{ad_X\}(A), \qquad ad_X(A) = [X, A].$$

Using Lemma B.1 for our unitary $U(\theta) = \exp\{i\theta P\}$ we get,

$$\begin{aligned}U(\theta) A U(\theta)^{\dagger} &= \sum_{n=0}^{\infty} \frac{(i\theta)^n}{n!} \mathrm{ad}_P^n(A) \\ &= \exp\{i\theta \mathrm{ad}_P\}(A) \\ &= \exp\{i\theta \mathrm{ad}_P\}(A_{\|} + A_{\perp}) \\ &= \exp\{i\theta \mathrm{ad}_P\}(A_{\|}) + \exp\{i\theta \mathrm{ad}_P\}(A_{\perp}) \quad \text{[using the linearity of ad}_P\text{]}. \end{aligned} \tag{15}$$

Since $\left[P, A_{\|}\right] = 0$, we have $\mathrm{ad}_P(A_{\|}) = 0$ and hence all higher commutators are zero. So:

$$\exp\{i\theta \mathrm{ad}_P\}(A_{\|}) = A_{\|}. \tag{16}$$

For $A_{\perp}$ with $\{P, A_{\perp}\} = 0$,

$$\mathrm{ad}_P(A_{\perp}) = [P, A_{\perp}] = P A_{\perp} - A_{\perp} P = P A_{\perp} - (-P A_{\perp}) = 2 P A_{\perp}. \tag{17}$$

$$\mathrm{ad}_P^2(A_{\perp}) = \mathrm{ad}_P(2 P A_{\perp}) = 2\left(P P A_{\perp} - P A_{\perp} P\right) = 2\left(A_{\perp} - (-A_{\perp})\right) = 4 A_{\perp}. \tag{18}$$

By induction we have,

$$\mathrm{ad}_P^{2m}(A_{\perp}) = 4^m A_{\perp}, \qquad \mathrm{ad}_P^{2m+1}(A_{\perp}) = 2 \cdot 4^m P A_{\perp}, \quad (m = 0, 1, 2, \dots). \tag{19}$$

Using Equation (16)–Equation (19) on Equation (15),

$$\begin{aligned}U(\theta) A U(\theta)^{\dagger} &= \exp\{i\theta \mathrm{ad}_P\}(A_{\|}) + \exp\{i\theta \mathrm{ad}_P\}(A_{\perp}) \\ &= A_{\|} + \sum_{m=0}^{\infty} \frac{(i\theta)^{2m}}{(2m)!} 4^m A_{\perp} + \sum_{m=0}^{\infty} \frac{(i\theta)^{2m+1}}{(2m+1)!} 2 \cdot 4^m P A_{\perp} \\ &= A_{\|} + \left[\sum_{m=0}^{\infty} (-1)^m \frac{(2\theta)^{2m}}{(2m)!}\right] A_{\perp} + i \left[\sum_{m=0}^{\infty} (-1)^m \frac{(2\theta)^{2m+1}}{(2m+1)!}\right] P A_{\perp} \\ &= A_{\|} + \cos(2\theta) A_{\perp} + i \sin(2\theta) P A_{\perp}.\end{aligned}$$

Pauli specialization: If Q is a Pauli string, then either $[P, Q] = 0$ ($Q = Q_{\|}$) or $\{P, Q\} = 0$ ($Q = Q_{\perp}$). Hence

$$U Q U^{\dagger} = \begin{cases} Q, & [P, Q] = 0, \\ Q\cos(2\theta) + \sin(2\theta) Q', & \{P, Q\} = 0. \end{cases}$$

Here $Q' = iPQ$ is Hermitian in the anticommuting case.

B.2 Pauli evolution through Rotation and CNOT gate

Single-qubit rotations

When $R_P(\phi) = \exp\left\{\pm i \frac{\phi}{2} P\right\}$, then $2\theta \mapsto \phi$ and

$$R_P(\phi) Q R_P(\phi)^{\dagger} = \begin{cases} Q, & [P, Q] = 0, \\ Q\cos\phi \pm i \sin\phi PQ, & \{P, Q\} = 0. \end{cases}$$

In the Table 1 below, X, Y, Z are Pauli operators on the indicated qubit, and ϕ is the rotation angle.

Gate	X	Y	Z
$R_x(\phi)(\exp\{i\phi X/2\})$	X	$Y\cos\phi - Z\sin\phi$	$Z\cos\phi + Y\sin\phi$
$R_y(\phi)(\exp\{i\phi Y/2\})$	$X\cos\phi + Z\sin\phi$	Y	$Z\cos\phi - X\sin\phi$
$R_z(\phi)(\exp\{i\phi Z/2\})$	$X\cos\phi - Y\sin\phi$	$Y\cos\phi + X\sin\phi$	Z

Table 1: Flip the sign in the exponent $\rightarrow$ flip the sign in all $\sin\phi$ terms of the push rules.

CNOT$_{c\to t}$ (control c, target t)

On control c	On target t
$X_c \mapsto X_c X_t$	$X_t \mapsto X_t$
$Y_c \mapsto Y_c X_t$	$Y_t \mapsto Z_c Y_t$
$Z_c \mapsto Z_c$	$Z_t \mapsto Z_c Z_t$

B.3 Example: 3 Qubits, 3 Rotations, 2 CNOTs

This section provides an example illustrating how propagated generators are computed step by step via the conjugation rules of Appendix B.2. The example demonstrates how a single-Pauli generator expands into a multiterm operator after propagation through rotation and entangling gates, which is the mechanism that motivates weight-k truncation in Section 4.1.

Circuit. On qubits (q_0, q_1, q_2), left-to-right:

$$\begin{aligned} U_1 &= R_y^{(0)}(\theta_1), \\ U_2 &= \text{CNOT}_{0\to1}, \\ U_3 &= R_x^{(1)}(\theta_2), \\ U_4 &= \text{CNOT}_{1\to2}, \\ U_5 &= R_z^{(2)}(\theta_3). \end{aligned}$$

Let $\rho_{in} = |000\rangle\langle 000|$ and choose $M = Z_2$. Generators: $G_1 = \frac{1}{2}Y_0$, $G_2 = \frac{1}{2}X_1$, $G_3 = \frac{1}{2}Z_2$.

Propagated generators.

$$\widetilde{G}_3 = \frac{1}{2}Z_2,$$

$$\widetilde{G}_2 = \frac{1}{2}\left(X_1X_2\cos\theta_3 - X_1Y_2\sin\theta_3\right), \quad \text{(through CNOT}_{1\to2}\text{ and } R_z^{(2)})$$

$$\widetilde{G}_1 = \frac{1}{2}\left(Y_0X_1X_2\cos\theta_3 - Y_0X_1Y_2\sin\theta_3\right), \quad \text{(through CNOT}_{0\to1}, R_x^{(1)}, \text{CNOT}_{1\to2}, R_z^{(2)})$$

C Surrogate Proofs and Estimation Details

C.1 Hermiticity of truncated generators and anticommutator identity

Lemma C.1 (Anticommutator identity for Hermitian operators)**.** *Let A, B be Hermitian operators. Then*

$$\Re\langle AB\rangle_\rho = \tfrac{1}{2}\langle\{A, B\}\rangle_\rho.$$

Moreover, if A is Hermitian and T_k is a linear truncation in the Pauli basis that preserves real coefficients, then $T_k(A)$ is also Hermitian.

Proof. For the first claim, $AB = \frac{1}{2}\{A, B\} + \frac{1}{2}[A, B]$. Since A, B are Hermitian, $\{A, B\}$ is Hermitian and $[A, B]$ is skew-Hermitian, so $\langle\{A, B\}\rangle_\rho$ is real and $\langle[A, B]\rangle_\rho$ is imaginary. Taking the real part yields the identity. For the second claim, T_k retains only real-coefficient Pauli terms, so $T_k(A)$ is a real linear combination of Hermitian Pauli operators and $T_k(A)^\dagger = T_k(A)$. □

C.2 Proof of Proposition 4.2

Proof. Let $v \in \mathbb{R}^p$ and define

$$\widetilde{G}^{(k)}(v) = \sum_{j=1}^{p} v_j \widetilde{G}_j^{(k)}.$$

Using Equation (5) and bilinearity,

$$\begin{aligned} v^\top F^{(k)} v &= 4\Re\left(\left\langle \sum_{j,l} v_j v_l \widetilde{G}_j^{(k)} \widetilde{G}_l^{(k)} \right\rangle_\rho - \left\langle \sum_j v_j \widetilde{G}_j^{(k)} \right\rangle_\rho^2\right) \\ &= 4\left(\left\langle \left(\widetilde{G}^{(k)}(v)\right)^2 \right\rangle_\rho - \left\langle \widetilde{G}^{(k)}(v) \right\rangle_\rho^2\right) \\ &= 4\mathrm{var}_\rho\left(\widetilde{G}^{(k)}(v)\right). \end{aligned}$$

Since each $\widetilde{G}_j^{(k)}$ is Hermitian and $v_j \in \mathbb{R}$, $\widetilde{G}^{(k)}(v)$ is Hermitian. The variance of a Hermitian operator is nonnegative, hence

$$v^\top F^{(k)} v \geq 0 \quad \forall v \in \mathbb{R}^p,$$

so $F^{(k)} \succeq 0$. □

C.3 Shadow estimation: unbiasedness and efficient computation

Lemma C.2 (Unbiasedness of shadow covariance)**.** *Let A, B be fixed Hermitian observables independent of $\mathcal{D}$. Then*

$$\mathbb{E}_{\mathcal{D}}\left[\widehat{Cov}_{\mathcal{D}}(A,B)\right] = Cov_\rho(A,B).$$

Proof. Since A, B are fixed and independent of $\mathcal{D}$, the unbiasedness of the shadow mean gives $\mathbb{E}_{\mathcal{D}}\left[\widehat{\langle AB\rangle}_{\mathcal{D}}\right] = \langle AB\rangle$, and the cross-fit estimator satisfies $\mathbb{E}_{\mathcal{D}}\left[\widehat{\langle A\rangle\langle B\rangle}_{\mathcal{D}}\right] = \langle A\rangle\langle B\rangle$. Subtracting yields the claim. □

Lemma C.3 (Unbiasedness on the active block)**.** *If $\mathcal{A}_t$ and $\Omega_{k,t}$ are deterministic given $(\overrightarrow{\theta}^t, k)$ and independent of the shadow randomness, then*

$$\mathbb{E}_{\mathcal{D}_t}\left[\widehat{F}^{(\Omega)}_{\mathcal{A}_t\mathcal{A}_t} \mid \overrightarrow{\theta}^t\right] = F^{(\Omega)}_{\mathcal{A}_t\mathcal{A}_t}(\overrightarrow{\theta}^t).$$

Proof. Since $\mathcal{A}_t$ and $\Omega_{k,t}$ are fixed given $\overrightarrow{\theta}^t$, Lemma C.2 gives conditional unbiasedness for each on-mask entry. Restricting to $j, l \in \mathcal{A}_t$ preserves this entrywise. □

C.3.1 Efficient computation of $\bar{\mu}_{\text{prod}}$.

Computing $\bar{\mu}_{\text{prod}}$ naively costs $O(N^2)$ per pair, but noting

$$\sum_{\substack{s,r=1 \\ s\neq r}}^{N} x_s(j)x_r(l) = \left(\sum_{s=1}^{N} x_s(j)\right)\left(\sum_{s=1}^{N} x_s(l)\right) - \sum_{s=1}^{N} x_s(j)x_s(l), \tag{20}$$

where $x_s(j) = X_s(\widetilde{G}_{j,t}^{(k)})$, the three accumulators $\sum_s x_s(j)$, $\sum_s x_s(l)$, and $\sum_s x_s(j)x_s(l)$ can be computed in a single pass over the data, reducing cost to $O(N)$.

C.4 Worked low-qubit surrogate construction

We continue the three-qubit propagated-generator example from Appendix B.3 through truncation and masking. To make the masking step nontrivial, we extend the circuit by appending the local rotation

$$U_6 = R_y^{(0)}(\theta_4), \qquad G_4 = \frac{1}{2}Y_0.$$

Since this is the final gate, its propagated generator is

$$\widetilde{G}_4 = \frac{1}{2}Y_0.$$

Because $R_y^{(0)}(\theta_4)$ commutes with the Y_0 factor in $\widetilde{G}_1$ and acts on a qubit outside the supports of $\widetilde{G}_2$ and $\widetilde{G}_3$, the first three propagated-generator expressions remain unchanged.

Truncation. Consider weight-2 truncation. Every Pauli term in $\widetilde{G}_1$ has weight three, while the terms in the other propagated generators have weight at most two. Therefore,

$$\begin{aligned}
\widetilde{G}_1^{(2)} &= 0, \\
\widetilde{G}_2^{(2)} &= \frac{1}{2}\left(X_1X_2\cos\theta_3 - X_1Y_2\sin\theta_3\right), \\
\widetilde{G}_3^{(2)} &= \frac{1}{2}Z_2, \\
\widetilde{G}_4^{(2)} &= \frac{1}{2}Y_0.
\end{aligned}$$

The resulting active set is

$$\mathcal{A} = \{2, 3, 4\}.$$

Retained supports and overlap mask. The retained generator supports are

$$\mathrm{supp}(\widetilde{G}_2^{(2)}) = \{1, 2\}, \qquad \mathrm{supp}(\widetilde{G}_3^{(2)}) = \{2\}, \qquad \mathrm{supp}(\widetilde{G}_4^{(2)}) = \{0\}.$$

The supports of parameters 2 and 3 overlap, while the support of parameter 4 is disjoint from both. So, the mask restricted to the active set is

$$\begin{aligned}
\Omega_2^{\mathcal{A}} &= \Omega_2 \cap (\mathcal{A} \times \mathcal{A}) \\
&= \{(2,2), (2,3), (3,2), (3,3), (4,4)\}.
\end{aligned}$$

Masked metric. On the active set $\mathcal{A} = \{2, 3, 4\}$, the masked metric is

$$F_{\mathcal{A}\mathcal{A}}^{(\Omega)} = \begin{pmatrix} F_{22}^{(2)} & F_{23}^{(2)} & 0 \\ F_{32}^{(2)} & F_{33}^{(2)} & 0 \\ 0 & 0 & F_{44}^{(2)} \end{pmatrix}.$$

Thus, truncation removes parameter 1 from the active block. The mask retains the coupling between parameters 2 and 3 and discards the off-diagonal couplings involving parameter 4. The mask imposes these zeros. The corresponding entries of $F^{(2)}$ and the exact QFIM may still be nonzero. We then estimate the retained entries from a shared shadow dataset and regularize them to form the active-block preconditioner, completing the pipeline in Figure 1.

D Approximation Error Analysis

Lemma D.1 (Covariance Cauchy–Schwarz)**.** *For Hermitian X, Y and any state ρ,*

$$|\mathrm{Cov}_\rho(X, Y)| \le \sqrt{var_\rho(X)\, var_\rho(Y)} \le \|X\|_{op}\|Y\|_{op}.$$

Proof.

$$\mathrm{Cov}_\rho(X,Y) = \langle (X - \langle X\rangle_\rho I)\,(Y - \langle Y\rangle_\rho I)\rangle_\rho\,.$$

Applying Cauchy–Schwarz to the centered operators gives

$$\begin{aligned}|\mathrm{Cov}_\rho(X,Y)| &\le \sqrt{\left\langle (X - \langle X\rangle_\rho I)^2 \right\rangle_\rho}\sqrt{\left\langle (Y - \langle Y\rangle_\rho I)^2 \right\rangle_\rho} \\ &= \sqrt{\mathrm{var}_\rho\,(X)\,\mathrm{var}_\rho\,(Y)}.\end{aligned}$$

Also,

$$\mathrm{var}_\rho\,(X) = \langle X^2\rangle_\rho - \langle X\rangle_\rho^2 \le \langle X^2\rangle_\rho \le \|X\|_{\mathrm{op}}^2,$$

and similarly for Y. Hence

$$|\mathrm{Cov}_\rho(X,Y)| \le \|X\|_{\mathrm{op}}\|Y\|_{\mathrm{op}}.$$

□

D.1 Proof of truncation bound in Equation (11)

For $j, l \in \mathcal{A}_t$, let

$$R_j = \widetilde{G}_j - \widetilde{G}_j^{(k)}, \qquad R_l = \widetilde{G}_l - \widetilde{G}_l^{(k)}.$$

Here R_j and R_l are the operators formed by Pauli terms discarded by weight-k truncation. Since each propagated generator $\widetilde{G}_u$ is Hermitian and T_k preserves Hermiticity by Lemma C.1, both $\widetilde{G}_u^{(k)}$ and R_u are Hermitian for $u \in \{j, l\}$.

Using bilinearity of covariance,

$$\begin{aligned}\mathrm{Cov}_\rho(\widetilde{G}_j, \widetilde{G}_l) - \mathrm{Cov}_\rho(\widetilde{G}_j^{(k)}, \widetilde{G}_l^{(k)}) &= \mathrm{Cov}_\rho(\widetilde{G}_j, \widetilde{G}_l - \widetilde{G}_l^{(k)}) + \mathrm{Cov}_\rho(\widetilde{G}_j - \widetilde{G}_j^{(k)}, \widetilde{G}_l^{(k)}) \\ &= \mathrm{Cov}_\rho(\widetilde{G}_j, R_l) + \mathrm{Cov}_\rho(R_j, \widetilde{G}_l^{(k)}).\end{aligned}$$

Since $F_{jl} = 4\Re[\mathrm{Cov}_\rho(\widetilde{G}_j, \widetilde{G}_l)]$ and $F_{jl}^{(k)} = 4\Re[\mathrm{Cov}_\rho(\widetilde{G}_j^{(k)}, \widetilde{G}_l^{(k)})]$,

$$|F_{jl} - F_{jl}^{(k)}| \le 4\left(|\mathrm{Cov}_\rho(\widetilde{G}_j, R_l)| + |\mathrm{Cov}_\rho(R_j, \widetilde{G}_l^{(k)})|\right).$$

Using Lemma D.1, $\|\widetilde{G}_j\|_{\mathrm{op}} \le 1$ (unitary conjugate of a Pauli), and

$$\|R_j\|_{\mathrm{op}}, \|R_l\|_{\mathrm{op}} \le \tau_{\max}, \qquad \|\widetilde{G}_l^{(k)}\|_{\mathrm{op}} \le \|\widetilde{G}_l\|_{\mathrm{op}} + \|\widetilde{G}_l - \widetilde{G}_l^{(k)}\|_{\mathrm{op}} \le 1 + \tau_{\max},$$

we obtain,

$$\begin{aligned}\left|\mathrm{Cov}_\rho(\widetilde{G}_j, R_l)\right| &\le \|\widetilde{G}_j\|_{\mathrm{op}}\|R_l\|_{\mathrm{op}} \le \tau_{\max}, \\ \left|\mathrm{Cov}_\rho(R_j, \widetilde{G}_l^{(k)})\right| &\le \|R_j\|_{\mathrm{op}}\|\widetilde{G}_l^{(k)}\|_{\mathrm{op}} \le \tau_{\max}(1 + \tau_{\max}).\end{aligned}$$

Therefore,

$$|F_{jl} - F_{jl}^{(k)}| \le 4\tau_{\max}(2 + \tau_{\max}). \tag{21}$$

By applying Equation (21) entrywise on the active block we get,

$$\begin{aligned}\|F_{\mathcal{A}_t\mathcal{A}_t} - F_{\mathcal{A}_t\mathcal{A}_t}^{(k)}\|_\infty &= \max_{j\in\mathcal{A}_t} \sum_{l\in\mathcal{A}_t} \left|F_{jl} - F_{jl}^{(k)}\right| \\ &\le |\mathcal{A}_t| 4\tau_{\max}(2 + \tau_{\max}).\end{aligned}$$

Since $F_{\mathcal{A}_t\mathcal{A}_t} - F_{\mathcal{A}_t\mathcal{A}_t}^{(k)}$ is symmetric,

$$\|F_{\mathcal{A}_t\mathcal{A}_t} - F_{\mathcal{A}_t\mathcal{A}_t}^{(k)}\|_{\mathrm{op}} \le \|F_{\mathcal{A}_t\mathcal{A}_t} - F_{\mathcal{A}_t\mathcal{A}_t}^{(k)}\|_\infty.$$

So,

$$E_{\mathrm{trunc},t} = \|F_{\mathcal{A}_t\mathcal{A}_t} - F_{\mathcal{A}_t\mathcal{A}_t}^{(k)}\|_{\mathrm{op}} \le 4|\mathcal{A}_t|\tau_{\max}(2 + \tau_{\max}).$$

We note that this is a conservative worst-case bound, obtained by entrywise covariance control followed by a row-sum norm estimate, and it may therefore be loose when many parameters are active.

D.2 Proof of masking bound in Equation (11)

Define $d_{\Omega,t}^{\text{off}} = \max_{j\in\mathcal{A}_t} \sum_{l\in\mathcal{A}_t} \mathbf{1}\left\{(j,l) \notin \Omega_{k,t}^{\mathcal{A}_t}, j \neq l\right\}$, the maximum number of off-mask entries in any row. By construction of the mask, for $j, l \in \mathcal{A}_t$,

$$F_{jl}^{(k)} - F_{jl}^{(\Omega)} = 0 \quad \text{if } (j,l) \in \Omega_{k,t}^{\mathcal{A}_t},$$

and in particular on the diagonal. For off-mask pairs $(j,l) \notin \Omega_{k,t}^{\mathcal{A}_t}$ with $j \neq l$, we have $F_{jl}^{(\Omega)} = 0$ by definition, so

$$\begin{aligned}\left|F_{jl}^{(k)} - F_{jl}^{(\Omega)}\right| &= \left|F_{jl}^{(k)}\right| \\ &= 4\left|\Re\left[\text{Cov}_\rho\left(\widetilde{G}_j^{(k)}, \widetilde{G}_l^{(k)}\right)\right]\right| \\ &\leq 4\left|\text{Cov}_\rho\left(\widetilde{G}_j^{(k)}, \widetilde{G}_l^{(k)}\right)\right|.\end{aligned}$$

Applying the off-mask entry bound row-wise over $\mathcal{A}_t$,

$$\begin{aligned}\|F_{\mathcal{A}_t\mathcal{A}_t}^{(k)} - F_{\mathcal{A}_t\mathcal{A}_t}^{(\Omega)}\|_\infty &= \max_{j\in\mathcal{A}_t} \sum_{l\in\mathcal{A}_t} \left|F_{jl}^{(k)} - F_{jl}^{(\Omega)}\right| \\ &\leq 4d_{\Omega,t}^{\text{off}} \max_{\substack{(j,l)\notin\Omega_{k,t}^{\mathcal{A}_t} \\ j\neq l, j,l\in\mathcal{A}_t}} \left|\text{Cov}_\rho\left(\widetilde{G}_j^{(k)}, \widetilde{G}_l^{(k)}\right)\right|.\end{aligned}$$

Since $F_{\mathcal{A}_t\mathcal{A}_t}^{(k)} - F_{\mathcal{A}_t\mathcal{A}_t}^{(\Omega)}$ is symmetric,

$$\|F_{\mathcal{A}_t\mathcal{A}_t}^{(k)} - F_{\mathcal{A}_t\mathcal{A}_t}^{(\Omega)}\|_{\text{op}} \leq \|F_{\mathcal{A}_t\mathcal{A}_t}^{(k)} - F_{\mathcal{A}_t\mathcal{A}_t}^{(\Omega)}\|_\infty.$$

Hence,

$$\begin{aligned}E_{\text{mask},t} &= \|F_{\mathcal{A}_t\mathcal{A}_t}^{(k)} - F_{\mathcal{A}_t\mathcal{A}_t}^{(\Omega)}\|_{\text{op}} \\ &\leq 4d_{\Omega,t}^{\text{off}} \max_{\substack{(j,l)\notin\Omega_{k,t}^{\mathcal{A}_t} \\ j\neq l, j,l\in\mathcal{A}_t}} \left|\text{Cov}_\rho\left(\widetilde{G}_j^{(k)}, \widetilde{G}_l^{(k)}\right)\right|,\end{aligned}$$

with the fact that this maximum is 0 when the index set is empty.

E Finite shot Analysis

Throughout this section, we condition on $\overrightarrow{\theta}^t$ so that the truncated generators, the mask, and the active set are all deterministic. The subscript $\mathcal{D}_t$ denotes randomness of the shadow dataset and angle brackets $\langle\cdot\rangle$ denote quantum expectations under $\rho(\overrightarrow{\theta}^t)$.

E.1 Estimator definitions and variance structure

For each on-mask pair $(j,l) \in \Omega_{k,t}^{\mathcal{A}_t}$, the two population quantities to be estimated are

$$\begin{aligned}\mu_{jl}^{\text{sym}} &= \left\langle \tfrac{1}{2}\left\{\widetilde{G}_{j,t}^{(k)}, \widetilde{G}_{l,t}^{(k)}\right\}\right\rangle, \\ \mu_{jl}^{\text{prod}} &= \langle\widetilde{G}_{j,t}^{(k)}\rangle\langle\widetilde{G}_{l,t}^{(k)}\rangle.\end{aligned}$$

These quantities are estimated from shadow data by $\bar{\mu}_{\text{sym}}(j,l)$ and $\bar{\mu}_{\text{prod}}(j,l)$, as defined in Equation (7). The equality for the product term uses the cross-fit structure of $\bar{\mu}_{\text{prod}}$, which averages a product kernel over off-diagonal pairs $s \neq r$. The off-diagonal restriction is important because distinct shadow snapshots are independent. Thus, for fixed observables A, B,

$$\mathbb{E}\left[X_s(A)X_r(B) \mid \overrightarrow{\theta}^t\right] = \mathbb{E}\left[X_s(A) \mid \overrightarrow{\theta}^t\right]\mathbb{E}\left[X_r(B) \mid \overrightarrow{\theta}^t\right], \qquad s \neq r.$$

Also, by shadow unbiasedness, for any observable O,

$$\mathbb{E}\left[X_s(O) \mid \overrightarrow{\theta}^t\right] = \langle O \rangle.$$

Combining the two identities gives

$$\mathbb{E}\left[X_s(A)X_r(B) \mid \overrightarrow{\theta}^t\right] = \langle A \rangle \langle B \rangle, \qquad s \neq r.$$

Taking $A = \widetilde{G}^{(k)}_{j,t}$ and $B = \widetilde{G}^{(k)}_{l,t}$, averaging over all off-diagonal pairs gives

$$\mathbb{E}_{\mathcal{D}_t}\left[\bar{\mu}_{\text{prod}}(j,l) \mid \overrightarrow{\theta}^t\right] = \langle \widetilde{G}^{(k)}_{j,t} \rangle \langle \widetilde{G}^{(k)}_{l,t} \rangle.$$

Similarly, $\bar{\mu}_{\text{sym}}(j,l)$ is unbiased by applying shadow unbiasedness to the observable $\frac{1}{2}\left\{\widetilde{G}^{(k)}_{j,t}, \widetilde{G}^{(k)}_{l,t}\right\}$. Thus,

$$\mu^{\text{sym}}_{jl} = \mathbb{E}_{\mathcal{D}_t}\left[\bar{\mu}_{\text{sym}}(j,l)\middle|\overrightarrow{\theta}^t\right],$$
$$\mu^{\text{prod}}_{jl} = \mathbb{E}_{\mathcal{D}_t}\left[\bar{\mu}_{\text{prod}}(j,l)\middle|\overrightarrow{\theta}^t\right].$$

The **linear estimator** $\bar{\mu}_{\text{sym}}(j,l)$ is a sample mean of N i.i.d. shadow features $X_s(\frac{1}{2}\left\{\widetilde{G}^{(k)}_{j,t}, \widetilde{G}^{(k)}_{l,t}\right\})$, so its variance scales as $1/N$:

$$\text{var}_{\mathcal{D}_t}\left(\bar{\mu}_{\text{sym}}(j,l) \mid \overrightarrow{\theta}^t\right) = \frac{1}{N}V^{\text{sym}}_{jl}, \qquad V^{\text{sym}}_{jl} = \text{var}_{\mathcal{D}_t}\left(X_s\left(\tfrac{1}{2}\left\{\widetilde{G}^{(k)}_{j,t}, \widetilde{G}^{(k)}_{l,t}\right\}\right) \mid \overrightarrow{\theta}^t\right).$$

The **quadratic estimator** $\bar{\mu}_{\text{prod}}(j,l)$ is a degree-2 U-statistic. A degree-2 U-statistic averages a kernel over all distinct pairs of samples and provides an unbiased estimator of the expectation of that kernel [12]. Here, the kernel is $h(s,r) = X_s(\widetilde{G}^{(k)}_{j,t})X_r(\widetilde{G}^{(k)}_{l,t})$, averaged over all distinct snapshot pairs. This structure admits a Hoeffding variance decomposition [13, Supplement C, Lemma 2], which separates the variance into first-order terms $V_{1,jl}, V_{2,jl}$ and a second-order term $V_{3,jl}$:

$$V_{1,jl} = \left\langle \widetilde{G}^{(k)}_{l,t} \right\rangle^2 \text{var}_{\mathcal{D}_t}\left(X_s(\widetilde{G}^{(k)}_{j,t})\middle|\overrightarrow{\theta}^t\right), \tag{22}$$

$$V_{2,jl} = \left\langle \widetilde{G}^{(k)}_{j,t} \right\rangle^2 \text{var}_{\mathcal{D}_t}\left(X_s(\widetilde{G}^{(k)}_{l,t})\middle|\overrightarrow{\theta}^t\right), \tag{23}$$

$$V_{3,jl} = \text{var}_{\mathcal{D}_t}\left(X_s(\widetilde{G}^{(k)}_{j,t})X_r(\widetilde{G}^{(k)}_{l,t})\middle|\overrightarrow{\theta}^t\right), \quad s \neq r, \tag{24}$$

where X_s, X_r are evaluated on independent shadow snapshots. $V_{1,jl}$ and $V_{2,jl}$ capture the variance from estimating each marginal mean, and $V_{3,jl}$ captures the joint fluctuation of two independent snapshots. The overall U-statistic variance satisfies

$$\text{var}_{\mathcal{D}_t}\left(\bar{\mu}_{\text{prod}}(j,l) \mid \overrightarrow{\theta}^t\right) \leq \frac{2}{N}\left(V_{1,jl} + V_{2,jl} + \frac{1}{N}V_{3,jl}\right) \leq \frac{1}{N}V^{\text{prod}}_{jl}, \tag{25}$$

with

$$V^{\text{prod}}_{jl} = 2\left(V_{1,jl} + V_{2,jl} + V_{3,jl}\right).$$

Hence V^{prod}_{jl} is N-independent. The worst-case quantities $V_\star$ and $B_\star$ defined in Equation (12) are the maxima of these variances and feature bounds over all on-mask pairs.

E.2 Bernstein inversion

Let $Y_1, \ldots, Y_N$ be i.i.d. mean-zero random variables with $\text{var}(Y_s) \leq \sigma^2$ and $|Y_s| \leq b$ almost surely. Then

$$\Pr\left(\left|\frac{1}{N}\sum_{s=1}^{N} Y_s\right| > u\right) \leq 2\exp\left(-\frac{Nu^2}{2\sigma^2 + \frac{2}{3}bu}\right). \tag{26}$$

Now setting, the right-hand side equal to $\alpha \in (0,1)$:

$$2\exp\left(-\frac{Nu^2}{2\sigma^2+\frac{2}{3}bu}\right)=\alpha \quad \Longleftrightarrow \quad \frac{Nu^2}{2\sigma^2+\frac{2}{3}bu}=\log(2/\alpha)$$

Hence u must satisfy

$$Nu^2-\frac{2}{3}b\log(2/\alpha)u-2\sigma^2\log(2/\alpha)=0.$$

The positive root is

$$u^\star=\frac{\frac{2}{3}b\log(2/\alpha)+\sqrt{\frac{4}{9}b^2(\log(2/\alpha))^2+8N\sigma^2\log(2/\alpha)}}{2N}.$$

Using $\sqrt{x+y}\le\sqrt{x}+\sqrt{y}$ we get,

$$\begin{aligned}u^\star &\le \frac{\frac{2}{3}b\log(2/\alpha)+\frac{2}{3}b\log(2/\alpha)+\sqrt{8N\sigma^2\log(2/\alpha)}}{2N}\\ &= \frac{\frac{4}{3}b\log(2/\alpha)+2\sqrt{2N\sigma^2\log(2/\alpha)}}{2N}\\ &= \sqrt{\frac{2\sigma^2\log(2/\alpha)}{N}}+\frac{2b\log(2/\alpha)}{3N}. \end{aligned} \tag{27}$$

Therefore, with probability at least $1-\alpha$,

$$\left|\frac{1}{N}\sum_{s=1}^{N}Y_s\right|\le\sqrt{\frac{2\sigma^2\log(2/\alpha)}{N}}+\frac{2b\log(2/\alpha)}{3N}. \tag{28}$$

E.3 Linear Term Concentration Detail

Recall from Equation (12) that B_{sym} and B_A denote uniform bounds on the shadow features, satisfying

$$\left|X_s(\tfrac{1}{2}\left\{\widetilde{G}^{(k)}_{j,t},\widetilde{G}^{(k)}_{l,t}\right\})\right|\le B_{\text{sym}},\qquad \left|X_s(\widetilde{G}^{(k)}_{j,t})\right|\le B_A.$$

Lemma E.1 (Linear-term concentration)**.** *For any $(j,l)\in\Omega^{\mathcal{A}_t}_{k,t}$ and $\delta_0\in(0,1)$, conditioned on $\overrightarrow{\theta}^t$, with probability at least $1-\delta_0/2$,*

$$\left|\bar{\mu}_{sym}(j,l)-\mu^{sym}_{jl}\right|\le\sqrt{\frac{2V^{sym}_{jl}\log(4/\delta_0)}{N}}+\frac{4B_{sym}\log(4/\delta_0)}{3N}. \tag{29}$$

Proof. Let,

$$Y_s=X_s\left(\tfrac{1}{2}\left\{\widetilde{G}^{(k)}_{j,t},\widetilde{G}^{(k)}_{l,t}\right\}\right)-\mu^{\text{sym}}_{jl}.$$

Then Y_s are i.i.d., and $\text{var}_{\mathcal{D}_t}\left(Y_s\mid\overrightarrow{\theta}^t\right)=V^{\text{sym}}_{jl}$. By the assumed bound,

$$\left|X_s\left(\tfrac{1}{2}\left\{\widetilde{G}^{(k)}_{j,t},\widetilde{G}^{(k)}_{l,t}\right\}\right)\right|\le B_{\text{sym}},$$

hence,

$$|Y_s|\le\left|X_s\left(\tfrac{1}{2}\left\{\widetilde{G}^{(k)}_{j,t},\widetilde{G}^{(k)}_{l,t}\right\}\right)\right|+\left|\mu^{\text{sym}}_{jl}\right|\le 2B_{\text{sym}}.$$

Bernstein's inequality with range bound $2B_{\text{sym}}$ gives

$$\Pr_{\mathcal{D}_t}\left(\left|\frac{1}{N}\sum_{s=1}^{N}Y_s\right|>u\middle|\overrightarrow{\theta}^t\right)\le 2\exp\left(-\frac{Nu^2}{2V^{\text{sym}}_{jl}+\frac{4}{3}B_{\text{sym}}u}\right).$$

Setting the RHS to $\delta_0/2$ and inverting via Equation (28) with $\sigma^2=V^{\text{sym}}_{jl}$, $b=2B_{\text{sym}}$, and $\alpha=\delta_0/2$ (so $\log(2/\alpha)=\log(4/\delta_0)$) yields Equation (29). □

E.4 Quadratic term concentration Details

From now on we assume $N \geq 2$ for the remaining proof as we are working degree-2 U-statistic quadratic term.

Lemma E.2 (Variance bridge for the symmetrized kernel)**.** *Fix $(j,l) \in \Omega_{k,t}^{\mathcal{A}_t}$ and condition on $\overrightarrow{\theta}^t$. Define*

$$h(s,r) = X_s(\widetilde{G}_{j,t}^{(k)})X_r(\widetilde{G}_{l,t}^{(k)}), \qquad h_{sym}(s,r) = \tfrac{1}{2}\left(h(s,r) + h(r,s)\right),$$

where $s \neq r$ index two independent shadow snapshots drawn from the shadow sampling process at $\overrightarrow{\theta}^t$. Then

$$var_{\mathcal{D}_t}\left(h_{sym}(s,r) \mid \overrightarrow{\theta}^t\right) \leq V_{3,jl} \leq \tfrac{1}{2}V_{jl}^{prod}.$$

Proof. By definition,

$$\begin{aligned}
\mathrm{var}_{\mathcal{D}_t}\left(h_{\mathrm{sym}}(s,r) \mid \overrightarrow{\theta}^t\right) &= \frac{1}{4}\mathrm{var}_{\mathcal{D}_t}\left(h(s,r) + h(r,s) \mid \overrightarrow{\theta}^t\right) \\
&\leq \frac{1}{4}\Big[\mathrm{var}_{\mathcal{D}_t}\left(h(s,r) \mid \overrightarrow{\theta}^t\right) + \mathrm{var}_{\mathcal{D}_t}\left(h(r,s) \mid \overrightarrow{\theta}^t\right) \\
&\quad + 2\left|\mathrm{Cov}_{\mathcal{D}_t}\left(h(s,r), h(r,s) \mid \overrightarrow{\theta}^t\right)\right|\Big] \\
&\leq \frac{1}{4}\Big[\mathrm{var}_{\mathcal{D}_t}\left(h(s,r) \mid \overrightarrow{\theta}^t\right) + \mathrm{var}_{\mathcal{D}_t}\left(h(r,s) \mid \overrightarrow{\theta}^t\right) \\
&\quad + 2\sqrt{\mathrm{var}_{\mathcal{D}_t}\left(h(s,r) \mid \overrightarrow{\theta}^t\right)\mathrm{var}_{\mathcal{D}_t}\left(h(r,s) \mid \overrightarrow{\theta}^t\right)}\Big],
\end{aligned}$$

where the last inequality uses $|\mathrm{Cov}\,(X,Y)| \leq \sqrt{\mathrm{var}\,(X)\,\mathrm{var}\,(Y)}$ by Cauchy–Schwarz with $X = h(s,r)$ and $Y = h(r,s)$.

Since s, r are independent and exchangeable conditioned on $\overrightarrow{\theta}^t$, using Equation (24),

$$\mathrm{var}_{\mathcal{D}_t}\left(h(s,r) \mid \overrightarrow{\theta}^t\right) = \mathrm{var}_{\mathcal{D}_t}\left(h(r,s) \mid \overrightarrow{\theta}^t\right) = V_{3,jl}.$$

Hence

$$\mathrm{var}_{\mathcal{D}_t}\left(h_{\mathrm{sym}}(s,r) \mid \overrightarrow{\theta}^t\right) \leq \frac{1}{4}(V_{3,jl} + V_{3,jl} + 2V_{3,jl}) = V_{3,jl}.$$

Also,

$$V_{jl}^{\mathrm{prod}} = 2(V_{1,jl} + V_{2,jl} + V_{3,jl}) \geq 2V_{3,jl},$$

so $V_{3,jl} \leq \frac{1}{2}V_{jl}^{\mathrm{prod}}$. □

Lemma E.3 (Quadratic-term concentration via Peel et al. [19])**.** *Fix $(j,l) \in \Omega_{k,t}^{\mathcal{A}_t}$, condition on $\overrightarrow{\theta}^t$. Let $n_{eff} = \lfloor N/2 \rfloor$. Then with probability at least $1 - \delta_0/2$,*

$$\left|\bar{\mu}_{prod}(j,l) - \mu_{jl}^{prod}\right| \leq \sqrt{\frac{V_{jl}^{prod}\log(4/\delta_0)}{n_{eff}}} + \frac{4B_A^2\log(4/\delta_0)}{3n_{eff}}. \tag{30}$$

Proof. Let

$$h(s,r) = X_s(\widetilde{G}_{j,t}^{(k)})X_r(\widetilde{G}_{l,t}^{(k)}), \qquad h_{\mathrm{sym}}(s,r) = \tfrac{1}{2}\left(h(s,r) + h(r,s)\right).$$

The ordered-pair estimator is equivalently written in symmetric U-statistic form as

$$\bar{\mu}_{\mathrm{prod}}(j,l) = \frac{1}{\binom{N}{2}}\sum_{s<r} h_{\mathrm{sym}}(s,r).$$

By the assumption of bounded features and triangle inequality,

$$|h_{\text{sym}}(s,r)| \le \tfrac{1}{2}\left(|X_s(\widetilde{G}_{j,t}^{(k)})||X_r(\widetilde{G}_{l,t}^{(k)})| + |X_r(\widetilde{G}_{j,t}^{(k)})||X_s(\widetilde{G}_{l,t}^{(k)})|\right) \le B_A^2,$$

so the rescaled kernel

$$\tilde{h}(s,r) = \frac{h_{\text{sym}}(s,r) + B_A^2}{2B_A^2} \in [0,1]$$

satisfies the kernel assumption required by [19, Theorem 2, Eq. (6)]. Let $\bar{\mu}_{\tilde{h}} = \frac{1}{\binom{N}{2}} \sum_{s<r} \tilde{h}(s,r)$ denote the corresponding U-statistic. Applying the result cited at the confidence level $\delta_0/2$ gives

$$\left|\bar{\mu}_{\tilde{h}} - \mathbb{E}_{\mathcal{D}_t}\left[\bar{\mu}_{\tilde{h}} \mid \overrightarrow{\theta}^t\right]\right| \le \sqrt{\frac{2\text{var}_{\mathcal{D}_t}\left(\tilde{h}(s,r) \mid \overrightarrow{\theta}^t\right)\log(4/\delta_0)}{n_{\text{eff}}}} + \frac{2\log(4/\delta_0)}{3n_{\text{eff}}}. \tag{31}$$

Since $h_{\text{sym}} = 2B_A^2\tilde{h} - B_A^2$, averaging gives $\bar{\mu}_{\text{prod}}(j,l) = 2B_A^2\bar{\mu}_{\tilde{h}} - B_A^2$. Now we get,

$$\bar{\mu}_{\text{prod}}(j,l) - \mu_{jl}^{\text{prod}} = 2B_A^2\left(\bar{\mu}_{\tilde{h}} - \mathbb{E}_{\mathcal{D}_t}\left[\bar{\mu}_{\tilde{h}} \mid \overrightarrow{\theta}^t\right]\right).$$

Taking absolute values and substituting Equation (31) results in

$$\left|\bar{\mu}_{\text{prod}}(j,l) - \mu_{jl}^{\text{prod}}\right| \le 2B_A^2\left(\sqrt{\frac{2\text{var}_{\mathcal{D}_t}\left(\tilde{h}(s,r) \mid \overrightarrow{\theta}^t\right)\log(4/\delta_0)}{n_{\text{eff}}}} + \frac{2\log(4/\delta_0)}{3n_{\text{eff}}}\right). \tag{32}$$

By using variance property we get,

$$\text{var}_{\mathcal{D}_t}\left(\tilde{h}(s,r) \mid \overrightarrow{\theta}^t\right) = \frac{1}{4B_A^4}\text{var}_{\mathcal{D}_t}\left(h_{\text{sym}}(s,r) \mid \overrightarrow{\theta}^t\right).$$

Substituting into Equation (32),

$$\left|\bar{\mu}_{\text{prod}}(j,l) - \mu_{jl}^{\text{prod}}\right| \le \sqrt{\frac{2\text{var}_{\mathcal{D}_t}\left(h_{\text{sym}}(s,r) \mid \overrightarrow{\theta}^t\right)\log(4/\delta_0)}{n_{\text{eff}}}} + \frac{4B_A^2\log(4/\delta_0)}{3n_{\text{eff}}}. \tag{33}$$

By Lemma E.2, $\text{var}_{\mathcal{D}_t}\left(h_{\text{sym}}(s,r) \mid \overrightarrow{\theta}^t\right) \le V_{3,jl} \le \frac{1}{2}V_{jl}^{\text{prod}}$. Substituting into Equation (33):

$$\sqrt{\frac{2\cdot\frac{1}{2}V_{jl}^{\text{prod}}\log(4/\delta_0)}{n_{\text{eff}}}} = \sqrt{\frac{V_{jl}^{\text{prod}}\log(4/\delta_0)}{n_{\text{eff}}}},$$

which yields Equation (30). □

Lemma E.4 (Quadratic-term concentration)**.** *Conditioned on $\overrightarrow{\theta}^t$, for any $(j,l) \in \Omega_{k,t}^{\mathcal{A}_t}$ and $\delta_0 \in (0,1)$, with probability at least $1-\delta_0/2$,*

$$\left|\bar{\mu}_{prod}(j,l) - \mu_{jl}^{prod}\right| \le \sqrt{3}\sqrt{\frac{V_{jl}^{prod}\log(4/\delta_0)}{N}} + \frac{4B_A^2\log(4/\delta_0)}{N}.$$

Proof. Apply Lemma E.3:

$$\left|\bar{\mu}_{\text{prod}}(j,l) - \mu_{jl}^{\text{prod}}\right| \le \sqrt{\frac{V_{jl}^{\text{prod}}\log(4/\delta_0)}{n_{\text{eff}}}} + \frac{4B_A^2\log(4/\delta_0)}{3n_{\text{eff}}}, \quad n_{\text{eff}} = \lfloor N/2 \rfloor.$$

For $N \ge 2$, using $n_{\text{eff}} \ge N/3$ yields the looser but simpler bound. □

E.5 Entrywise concentration

Lemma E.5 (Entrywise concentration)**.** *For any $(j,l) \in \Omega_{k,t}^{\mathcal{A}_t}$ and $\delta_0 \in (0,1)$, with probability at least $1-\delta_0$,*

$$\left|\widehat{F}_{jl}^{(\Omega)} - F_{jl}^{(\Omega)}\right| \le C_1\sqrt{\frac{V_\star \log(4/\delta_0)}{N}} + C_2\frac{B_\star \log(4/\delta_0)}{N},$$

where

$$C_1 = 4(\sqrt{2}+\sqrt{3}), \qquad C_2 = \frac{64}{3}.$$

Proof. From Equation (8),

$$\widehat{F}_{jl}^{(\Omega)} - F_{jl}^{(\Omega)} = 4\left[\left(\bar{\mu}_{\text{sym}}(j,l) - \mu_{jl}^{\text{sym}}\right) - \left(\bar{\mu}_{\text{prod}}(j,l) - \mu_{jl}^{\text{prod}}\right)\right].$$

Hence,

$$\left|\widehat{F}_{jl}^{(\Omega)} - F_{jl}^{(\Omega)}\right| \le 4\left(\left|\bar{\mu}_{\text{sym}}(j,l) - \mu_{jl}^{\text{sym}}\right| + \left|\bar{\mu}_{\text{prod}}(j,l) - \mu_{jl}^{\text{prod}}\right|\right). \tag{34}$$

By Lemma E.1,

$$\left|\bar{\mu}_{\text{sym}}(j,l) - \mu_{jl}^{\text{sym}}\right| \le \sqrt{\frac{2V_{jl}^{\text{sym}}\log(4/\delta_0)}{N}} + \frac{4B_{\text{sym}}\log(4/\delta_0)}{3N}$$

with probability at least $1-\delta_0/2$. By Lemma E.4,

$$\left|\bar{\mu}_{\text{prod}}(j,l) - \mu_{jl}^{\text{prod}}\right| \le \sqrt{3}\sqrt{\frac{V_{jl}^{\text{prod}}\log(4/\delta_0)}{N}} + 4\frac{B_A^2\log(4/\delta_0)}{N}$$

with probability at least $1-\delta_0/2$. Applying a union bound, both inequalities hold simultaneously with probability at least $1-\delta_0$. Using

$$V_{jl}^{\text{sym}}, V_{jl}^{\text{prod}} \le V_\star, \qquad B_{\text{sym}}, B_A^2 \le B_\star,$$

and Equation (34) we get,

$$\left|\widehat{F}_{jl}^{(\Omega)} - F_{jl}^{(\Omega)}\right| \le 4(\sqrt{2}+\sqrt{3})\sqrt{\frac{V_\star\log(4/\delta_0)}{N}} + \frac{64}{3}\frac{B_\star\log(4/\delta_0)}{N}.$$

□

E.6 Proof of Theorem 5.1

Proof. For each $(j,l) \in \Omega_{k,t}^{\mathcal{A}_t}$, define the failure event

$$\mathcal{E}_{jl} = \left\{\left|\widehat{F}_{jl}^{(\Omega)} - F_{jl}^{(\Omega)}\right| > C_1\sqrt{\frac{V_\star\log\left(4|\Omega_{k,t}^{\mathcal{A}_t}|/\delta\right)}{N}} + C_2\frac{B_\star\log\left(4|\Omega_{k,t}^{\mathcal{A}_t}|/\delta\right)}{N}\right\}.$$

By Lemma E.5 with $\delta_0 = \delta/|\Omega_{k,t}^{\mathcal{A}_t}|$,

$$\Pr_{\mathcal{D}_t}(\mathcal{E}_{jl} \mid \overrightarrow{\theta}^t) \le \frac{\delta}{|\Omega_{k,t}^{\mathcal{A}_t}|}.$$

Now, we apply a union bound over all on-mask pairs:

$$\Pr_{\mathcal{D}_t}\left(\bigcup_{(j,l)\in\Omega_{k,t}^{\mathcal{A}_t}}\mathcal{E}_{jl}\,\middle|\,\overrightarrow{\theta}^t\right) \le \sum_{(j,l)\in\Omega_{k,t}^{\mathcal{A}_t}}\Pr_{\mathcal{D}_t}(\mathcal{E}_{jl} \mid \overrightarrow{\theta}^t) \le |\Omega_{k,t}^{\mathcal{A}_t}|\cdot\frac{\delta}{|\Omega_{k,t}^{\mathcal{A}_t}|} = \delta.$$

Hence, with probability at least $1-\delta$, conditioned on $\overrightarrow{\theta}$, none of the $\mathcal{E}_{jl}$ occurs,

$$\left|\widehat{F}^{(\Omega)}_{jl} - F^{(\Omega)}_{jl}\right| \le C_1\sqrt{\frac{V_\star \log\left(4|\Omega^{\mathcal{A}_t}_{k,t}|/\delta\right)}{N}} + C_2\frac{B_\star \log\left(4|\Omega^{\mathcal{A}_t}_{k,t}|/\delta\right)}{N}, \qquad \forall (j,l) \in \Omega^{\mathcal{A}_t}_{k,t},$$

Now, since $\widehat{F}^{(\Omega)}_{\mathcal{A}_t\mathcal{A}_t}$ and $F^{(\Omega)}_{\mathcal{A}_t\mathcal{A}_t}$ are symmetric masked matrices, their difference $\widehat{F}^{(\Omega)}_{\mathcal{A}_t\mathcal{A}_t} - F^{(\Omega)}_{\mathcal{A}_t\mathcal{A}_t}$ is symmetric and supported on $\Omega^{\mathcal{A}_t}_{k,t}$. Therefore, each row has at most $d_{\Omega,t}$ nonzero entries. Then every nonzero entry satisfies the following:

$$\left|\widehat{F}^{(\Omega)}_{jl} - F^{(\Omega)}_{jl}\right| \le C_1\sqrt{\frac{V_\star \log\left(4|\Omega^{\mathcal{A}_t}_{k,t}|/\delta\right)}{N}} + C_2\frac{B_\star \log\left(4|\Omega^{\mathcal{A}_t}_{k,t}|/\delta\right)}{N}.$$

Hence,

$$\begin{aligned}\left\|\widehat{F}^{(\Omega)}_{\mathcal{A}_t\mathcal{A}_t} - F^{(\Omega)}_{\mathcal{A}_t\mathcal{A}_t}\right\|_\infty &= \max_j \sum_l \left|\widehat{F}^{(\Omega)}_{jl} - F^{(\Omega)}_{jl}\right| \\ &\le d_{\Omega,t}\left[C_1\sqrt{\frac{V_\star \log\left(4|\Omega^{\mathcal{A}_t}_{k,t}|/\delta\right)}{N}} + C_2\frac{B_\star \log\left(4|\Omega^{\mathcal{A}_t}_{k,t}|/\delta\right)}{N}\right].\end{aligned}$$

Since the difference matrix is symmetric,

$$\left\|\widehat{F}^{(\Omega)}_{\mathcal{A}_t\mathcal{A}_t} - F^{(\Omega)}_{\mathcal{A}_t\mathcal{A}_t}\right\|_{\text{op}} \le \left\|\widehat{F}^{(\Omega)}_{\mathcal{A}_t\mathcal{A}_t} - F^{(\Omega)}_{\mathcal{A}_t\mathcal{A}_t}\right\|_\infty.$$

Combining the above inequalities gives Equation (13). □

E.7 Proof of Corollary 5.2

Proof. Setting each of the two terms in Equation (13) to at most $\varepsilon/2$ gives the sufficient conditions

$$d_{\Omega,t}C_1\sqrt{\frac{V_\star \log(4\left|\Omega^{\mathcal{A}_t}_{k,t}\right|/\delta)}{N}} \le \frac{\varepsilon}{2}, \qquad d_{\Omega,t}C_2\frac{B_\star \log(4\left|\Omega^{\mathcal{A}_t}_{k,t}\right|/\delta)}{N} \le \frac{\varepsilon}{2}.$$

These are implied by

$$N \ge \frac{4C_1^2 V_\star d^2_{\Omega,t}\log(4\left|\Omega^{\mathcal{A}_t}_{k,t}\right|/\delta)}{\varepsilon^2}, \qquad N \ge \frac{2C_2 B_\star d_{\Omega,t}\log(4\left|\Omega^{\mathcal{A}_t}_{k,t}\right|/\delta)}{\varepsilon}.$$

Taking the maximum yields the following:

$$N \ge \max\left\{\frac{4C_1^2 V_\star d^2_{\Omega,t}\log\left(4|\Omega^{\mathcal{A}_t}_{k,t}|/\delta\right)}{\varepsilon^2}, \frac{2C_2 B_\star d_{\Omega,t}\log\left(4|\Omega^{\mathcal{A}_t}_{k,t}|/\delta\right)}{\varepsilon}\right\}, \tag{35}$$

with $C_1 = 4(\sqrt{2}+\sqrt{3})$ and $C_2 = 64/3$. The $\mathcal{O}(\cdot)$ form in Equation (14) follows by absorbing numerical constants. In the variance-dominated regime, the ε^{-1} term is negligible and the complexity simplifies to

$$N = \mathcal{O}\left(\frac{V_\star d^2_{\Omega,t}\log(\left|\Omega^{\mathcal{A}_t}_{k,t}\right|/\delta)}{\varepsilon^2}\right). \tag{36}$$

□

Remark E.6 (Pauli level scaling of $B_\star$ and $V_\star$). For local Pauli shadows, a Pauli observable Q of weight w satisfies

$$\mathbb{E}\left[X_s(Q)^2\right] \le 3^w$$

[13]. In our construction, $X_s(\widetilde{G}^{(k)}_{j,t})$ only involves Pauli strings of weight at most k, while $X_s\left(\frac{1}{2}\left\{\widetilde{G}^{(k)}_{j,t}, \widetilde{G}^{(k)}_{l,t}\right\}\right)$ involves weight at most $2k$. Therefore, up to coefficient size factors (iterate and potentially k-dependent, but independent of N),

$$B_A = \mathcal{O}(3^k), \qquad B_{\text{sym}} = \mathcal{O}(3^{2k}), \qquad V^{\text{sym}}_{jl} = \mathcal{O}(3^{2k}), \qquad V^{\text{prod}}_{jl} = \mathcal{O}(3^{2k}).$$

Hence

$$B_\star = \max\left\{B_{\text{sym}}, B_A^2\right\} = \mathcal{O}(3^{2k}), \qquad V_\star = \mathcal{O}(3^{2k}),$$

and both are N-independent. This makes k a budget-constrained design knob. Larger k retains more curvature but demands exponentially more shots.

Remark E.7 (Interpretation of the sample complexity). The bound exhibits three key features. (i) The number of shadow snapshots scales logarithmically in the number of on mask entries $|\Omega^{\mathcal{A}_t}_{k,t}|$, reflecting the classical shadow advantage. A single measurement dataset supports simultaneous estimation of all on-mask entries without dedicated circuits for each pair. (ii) The exponential dependence on the truncation level k (through $V_\star$, $B_\star$) makes k a budget-constrained knob. Larger k retains more of the true QFIM structure but demands proportionally more measurements, so under a fixed shot budget one may choose the largest k for which the estimation error remains controlled. (iii) The sparsity of the mask enters through $d^2_{\Omega,t}$, so sparser masks (smaller $d_{\Omega,t}$) reduce the required sample size.

This bound is a worst case sufficient condition and may substantially overestimate the shot budget needed in practice.

F Convergence Guarantees

This section provides the proofs of Theorems 5.3 and 5.4. Throughout this section we use the notation of Section 5.3.

$$\begin{aligned} M_t &= \widehat{F}^{(\Omega)}_{\mathcal{A}_t\mathcal{A}_t} + \lambda I \\ g^t &= \nabla C(\overrightarrow{\theta}^t) \\ \hat{g}^t &= g^t + \xi^t, \end{aligned}$$

with $\mathbb{E}\left[\xi^t \mid \overrightarrow{\theta}^t\right] = 0$ and $\mathbb{E}\left[\|\xi^t\|_2^2 \mid \overrightarrow{\theta}^t\right] \le \sigma_g^2$ and expectations involving ξ^t are taken over the gradient estimation randomness. Since gradient and metric estimators use disjoint shot batches, $\xi^t \perp M_t \mid \overrightarrow{\theta}^t$. Therefore,

$$\mathbb{E}\left[\xi^t \mid \overrightarrow{\theta}^t, M_t\right] = 0, \qquad \mathbb{E}\left[\|\xi^t\|_2^2 \mid \overrightarrow{\theta}^t, M_t\right] \le \sigma_g^2.$$

F.1 Uniform sample-size condition

Let the good event be $\mathcal{E}_T$ as defined in Section 5.3. This is the event that the finite-shot estimation error ($E_{\text{shot},t}$) is at most ε simultaneously for all t. To obtain a single sample-size requirement uniform over t, let us define

$$\begin{aligned} V_\star^{\max} &= \max_{0\le t<T} V_{\star,t}, \qquad B_\star^{\max} = \max_{0\le t<T} B_{\star,t}, \\ d_\Omega^{\max} &= \max_{0\le t<T} d_{\Omega,t}, \qquad |\Omega|_{\max} = \max_{0\le t<T} |\Omega^{\mathcal{A}_t}_{k,t}|. \end{aligned} \tag{37}$$

By Corollary 5.2, for each $t = 0, \dots, T-1$, $\Pr_{\mathcal{D}_t}\left(E_{\text{shot},t} \le \varepsilon \mid \overrightarrow{\theta}^t\right) \ge 1 - \frac{\delta}{T}$. Recalling the event $\mathcal{E}_T = \bigcap_{t=0}^{T-1}\{E_{\text{shot},t} \le \varepsilon\}$, an iterated conditioning argument together with a union bound

yields $\Pr(\mathcal{E}_T) \geq 1-\delta$, where the probability is over the joint shadow randomness across iterations, provided

$$N \geq \max\left\{ \frac{4C_1^2 V_\star^{\max}(d_\Omega^{\max})^2 \log(4T|\Omega|_{\max}/\delta)}{\varepsilon^2}, \frac{2C_2 B_\star^{\max} d_\Omega^{\max} \log(4T|\Omega|_{\max}/\delta)}{\varepsilon} \right\}. \tag{38}$$

Equation (38) should be interpreted pathwise where the maxima are taken along the realized trajectory. The $\log T$ factor relative to the per-iteration bound (Corollary 5.2) reflects the cost of simultaneous control across all iterations.

F.2 Spectral bounds for the preconditioner

Assume there exist constants $m > 0$ and $\Lambda_{\max} < \infty$ such that on $\mathcal{E}_T$,

$$mI \preceq M_t \preceq \Lambda_{\max} I, \qquad \forall t = 0, \ldots, T-1. \tag{39}$$

These spectral bounds are an assumption of the convergence analysis. The finite-shot concentration of Corollary 5.2 provides one concrete route to verifying them. In the shadow-estimation setting, Weyl's inequality yields the following sufficient spectral bounds. On $\mathcal{E}_T$, for each t,

$$\begin{aligned}
\lambda_{\min}(M_t) &= \lambda_{\min}(\widehat{F}^{(\Omega)}_{\mathcal{A}_t\mathcal{A}_t}) + \lambda \\
&\geq \lambda_{\min}(F^{(\Omega)}_{\mathcal{A}_t\mathcal{A}_t}) - \|\widehat{F}^{(\Omega)}_{\mathcal{A}_t\mathcal{A}_t} - F^{(\Omega)}_{\mathcal{A}_t\mathcal{A}_t}\|_{\text{op}} + \lambda \\
&\geq \lambda_{\min}(F^{(\Omega)}_{\mathcal{A}_t\mathcal{A}_t}) - \varepsilon + \lambda.
\end{aligned}$$

Since entrywise masking need not preserve positive semidefiniteness, $F^{(\Omega)}_{\mathcal{A}_t\mathcal{A}_t}$ may have negative eigenvalues. Define $\kappa_{\max} = \max_{0\leq t<T} \max\left\{0, -\lambda_{\min}(F^{(\Omega)}_{\mathcal{A}_t\mathcal{A}_t})\right\} < \infty$, so that $\lambda_{\min}(F^{(\Omega)}_{\mathcal{A}_t\mathcal{A}_t}) \geq -\kappa_{\max}$. This gives

$$m = \lambda - \varepsilon - \kappa_{\max}. \tag{40}$$

provided $\lambda > \varepsilon + \kappa_{\max}$ so that $m > 0$. For the upper bound, again on $\mathcal{E}_T$,

$$\begin{aligned}
\|M_t\|_{\text{op}} &= \|\widehat{F}^{(\Omega)}_{\mathcal{A}_t\mathcal{A}_t} + \lambda I\|_{\text{op}} \\
&= \|F^{(\Omega)}_{\mathcal{A}_t\mathcal{A}_t} + (\widehat{F}^{(\Omega)}_{\mathcal{A}_t\mathcal{A}_t} - F^{(\Omega)}_{\mathcal{A}_t\mathcal{A}_t}) + \lambda I\|_{\text{op}} \\
&\leq \|F^{(\Omega)}_{\mathcal{A}_t\mathcal{A}_t}\|_{\text{op}} + \|\widehat{F}^{(\Omega)}_{\mathcal{A}_t\mathcal{A}_t} - F^{(\Omega)}_{\mathcal{A}_t\mathcal{A}_t}\|_{\text{op}} + \|\lambda I\|_{\text{op}} \\
&\leq \|F^{(\Omega)}_{\mathcal{A}_t\mathcal{A}_t}\|_{\text{op}} + \varepsilon + \lambda \\
&\leq F_{\max} + \varepsilon + \lambda,
\end{aligned}$$

where $F_{\max} = \max_{0\leq t<T} \|F^{(\Omega)}_{\mathcal{A}_t\mathcal{A}_t}\|_{\text{op}} < \infty$. Thus a sufficient upper spectral bound is

$$\Lambda_{\max} = F_{\max} + \varepsilon + \lambda.$$

Therefore, under Equation (38) and provided $\lambda > \varepsilon + \kappa_{\max}$ (so that $m > 0$ and $M_t \succ 0$), the preconditioner satisfies Equation (39) along the realized trajectory with $m = \lambda - \varepsilon - \kappa_{\max}$ and $\Lambda_{\max} = F_{\max} + \varepsilon + \lambda$. This provides a pathwise sufficient route to verifying the spectral assumption used in the convergence analysis.

F.3 Proof of Theorem 5.3

Proof. Condition on $(\overrightarrow{\theta}^t, M_t)$, under which g^t and M_t are deterministic and the remaining expectation is over the gradient estimation noise ξ^t.

By β-smoothness of C, for any step $\Delta^t = \overrightarrow{\theta}^{t+1} - \overrightarrow{\theta}^t$,

$$C(\overrightarrow{\theta}^t + \Delta^t) \leq C(\overrightarrow{\theta}^t) + \langle g^t, \Delta^t \rangle + \frac{\beta}{2}\|\Delta^t\|_2^2.$$

Taking expectation over gradient noise gives:

$$\mathbb{E}\left[C(\overrightarrow{\theta}^{t+1}) \mid \overrightarrow{\theta}^t, M_t\right] \leq C(\overrightarrow{\theta}^t) + \mathbb{E}\left[\langle g^t, \Delta^t\rangle \mid \overrightarrow{\theta}^t, M_t\right] + \frac{\beta}{2}\mathbb{E}\left[\|\Delta^t\|_2^2 \mid \overrightarrow{\theta}^t, M_t\right]. \tag{41}$$

The update $\Delta^t = (\Delta^t_{\mathcal{A}_t}, \Delta^t_{\mathcal{I}_t})$ is block-diagonal, so both the inner product and the squared norm split with no cross terms:

$$\langle g^t, \Delta^t\rangle = \langle g^t_{\mathcal{A}_t}, \Delta^t_{\mathcal{A}_t}\rangle + \langle g^t_{\mathcal{I}_t}, \Delta^t_{\mathcal{I}_t}\rangle, \qquad \|\Delta^t\|_2^2 = \|\Delta^t_{\mathcal{A}_t}\|_2^2 + \|\Delta^t_{\mathcal{I}_t}\|_2^2.$$

Active block. The update is $\Delta^t_{\mathcal{A}_t} = -\eta_{\text{qng}} M_t^{-1} \hat{g}^t_{\mathcal{A}_t}$.

Linear term. Since M_t is fixed under the conditioning and $\mathbb{E}\left[\hat{g}^t_{\mathcal{A}_t} \mid \overrightarrow{\theta}^t, M_t\right] = g^t_{\mathcal{A}_t}$ by unbiasedness,

$$\begin{aligned}\mathbb{E}\left[\langle g^t_{\mathcal{A}_t}, \Delta^t_{\mathcal{A}_t}\rangle \mid \overrightarrow{\theta}^t, M_t\right] &= -\eta_{\text{qng}} g^{t\top}_{\mathcal{A}_t} M_t^{-1} \mathbb{E}\left[\hat{g}^t_{\mathcal{A}_t} \mid \overrightarrow{\theta}^t, M_t\right] \\ &= -\eta_{\text{qng}} g^{t\top}_{\mathcal{A}_t} M_t^{-1} g^t_{\mathcal{A}_t}. \end{aligned} \tag{42}$$

Quadratic term. Substituting the update rule gives $\|\Delta^t_{\mathcal{A}_t}\|_2^2 = \eta_{\text{qng}}^2 \|M_t^{-1}\hat{g}^t_{\mathcal{A}_t}\|_2^2$. Writing $\hat{g}^t_{\mathcal{A}_t} = g^t_{\mathcal{A}_t} + \xi^t_{\mathcal{A}_t}$ and expanding:

$$\begin{aligned}\|M_t^{-1}(g^t_{\mathcal{A}_t} + \xi^t_{\mathcal{A}_t})\|_2^2 &= (g^t_{\mathcal{A}_t} + \xi^t_{\mathcal{A}_t})^\top M_t^{-2}(g^t_{\mathcal{A}_t} + \xi^t_{\mathcal{A}_t}) \\ &= g^{t\top}_{\mathcal{A}_t} M_t^{-2} g^t_{\mathcal{A}_t} + 2 g^{t\top}_{\mathcal{A}_t} M_t^{-2} \xi^t_{\mathcal{A}_t} + \xi^{t\top}_{\mathcal{A}_t} M_t^{-2} \xi^t_{\mathcal{A}_t}, \end{aligned} \tag{43}$$

where $M_t^{-\top} = M_t^{-1}$ by symmetry of M_t, so $M_t^{-\top} M_t^{-1} = M_t^{-2}$. Taking expectation, the cross term vanishes because $g^t_{\mathcal{A}_t}$ and M_t^{-2} are fixed and $\mathbb{E}\left[\xi^t_{\mathcal{A}_t} \mid \overrightarrow{\theta}^t, M_t\right] = 0$:

$$\mathbb{E}\left[\|M_t^{-1}\hat{g}^t_{\mathcal{A}_t}\|_2^2 \mid \overrightarrow{\theta}^t, M_t\right] = \underbrace{g^{t\top}_{\mathcal{A}_t} M_t^{-2} g^t_{\mathcal{A}_t}}_{\text{signal}} + \underbrace{\mathbb{E}\left[\xi^{t\top}_{\mathcal{A}_t} M_t^{-2} \xi^t_{\mathcal{A}_t} \mid \overrightarrow{\theta}^t, M_t\right]}_{\text{noise}}. \tag{44}$$

We bound each piece using $M_t \succeq mI$.

Signal bound. Since

$$M_t \succeq mI \Longrightarrow \lambda_{\min}(M_t) \geq m \Longrightarrow M_t^{-1} \preceq m^{-1} I,$$

Applying congruence monotonicity ($A \preceq B \Rightarrow C^\top A C \preceq C^\top B C$) with $C = M_t^{-1/2}$ gives

$$M_t^{-2} = M_t^{-1/2} M_t^{-1} M_t^{-1/2} \preceq m^{-1} M_t^{-1}.$$

Therefore,

$$g^{t\top}_{\mathcal{A}_t} M_t^{-2} g^t_{\mathcal{A}_t} \leq m^{-1} g^{t\top}_{\mathcal{A}_t} M_t^{-1} g^t_{\mathcal{A}_t}. \tag{45}$$

Noise bound. For any vector v, $\|M_t^{-1}v\|_2 \leq \|M_t^{-1}\|_{\text{op}} \|v\|_2$, so $\|M_t^{-1}v\|_2^2 \leq \|M_t^{-1}\|_{\text{op}}^2 \|v\|_2^2$. From $M_t \succeq mI$, $\|M_t^{-1}\|_{\text{op}} \leq m^{-1}$. By taking expectation:

$$\mathbb{E}\left[\|M_t^{-1}\xi^t_{\mathcal{A}_t}\|_2^2 \mid \overrightarrow{\theta}^t, M_t\right] \leq m^{-2}\mathbb{E}\left[\|\xi^t_{\mathcal{A}_t}\|_2^2 \mid \overrightarrow{\theta}^t, M_t\right] \leq m^{-2}\sigma_g^2. \tag{46}$$

Substituting Equation (45) and Equation (46) into Equation (44), multiplying by $\beta\eta_{\text{qng}}^2/2$, and adding Equation (42):

$$\begin{aligned}&\mathbb{E}\left[\langle g^t_{\mathcal{A}_t}, \Delta^t_{\mathcal{A}_t}\rangle + \frac{\beta}{2}\|\Delta^t_{\mathcal{A}_t}\|_2^2 \mid \overrightarrow{\theta}^t, M_t\right] \\ &\leq -\eta_{\text{qng}} g^{t\top}_{\mathcal{A}_t} M_t^{-1} g^t_{\mathcal{A}_t} + \frac{\beta\eta_{\text{qng}}^2}{2}\left(m^{-1} g^{t\top}_{\mathcal{A}_t} M_t^{-1} g^t_{\mathcal{A}_t} + m^{-2}\sigma_g^2\right) \\ &= -\eta_{\text{qng}}\left(1 - \frac{\beta\eta_{\text{qng}}}{2m}\right) g^{t\top}_{\mathcal{A}_t} M_t^{-1} g^t_{\mathcal{A}_t} + \frac{\beta\eta_{\text{qng}}^2\sigma_g^2}{2m^2}. \end{aligned} \tag{47}$$

The stepsize condition $\eta_{\text{qng}} \leq m/\beta$ ensures $(1 - \beta\eta_{\text{qng}}/2m) \geq 1/2$, so

$$\mathbb{E}\left[\langle g^t_{\mathcal{A}_t}, \Delta^t_{\mathcal{A}_t}\rangle + \frac{\beta}{2}\|\Delta^t_{\mathcal{A}_t}\|_2^2 \mid \overrightarrow{\theta}^t, M_t\right] \leq -\frac{\eta_{\text{qng}}}{2} g^{t\top}_{\mathcal{A}_t} M_t^{-1} g^t_{\mathcal{A}_t} + \frac{\beta\eta_{\text{qng}}^2\sigma_g^2}{2m^2}. \tag{48}$$

Inactive block. The update is $\Delta^t_{\mathcal{I}_t} = -\eta_{\text{sgd}}\hat{g}^t_{\mathcal{I}_t}$, which is standard SGD. Repeating the same smoothness and noise argument above with identity preconditioner yields

$$\mathbb{E}\left[\langle g^t_{\mathcal{I}_t}, \Delta^t_{\mathcal{I}_t}\rangle + \frac{\beta}{2}\|\Delta^t_{\mathcal{I}_t}\|_2^2 \mid \overrightarrow{\theta}^t, M_t\right] \leq -\frac{\eta_{\text{sgd}}}{2}\|g^t_{\mathcal{I}_t}\|_2^2 + \frac{\beta\eta_{\text{sgd}}^2\sigma_g^2}{2}, \tag{49}$$

under $\eta_{\text{sgd}} \leq 1/\beta$. Substituting Equation (48) and Equation (49) into Equation (41):

$$\mathbb{E}\left[C(\overrightarrow{\theta}^{t+1}) \mid \overrightarrow{\theta}^t, M_t\right] \leq C(\overrightarrow{\theta}^t) - \frac{\eta_{\text{qng}}}{2} g^{t\top}_{\mathcal{A}_t} M_t^{-1} g^t_{\mathcal{A}_t} - \frac{\eta_{\text{sgd}}}{2}\|g^t_{\mathcal{I}_t}\|_2^2 + \underbrace{\frac{\beta}{2}\left(\frac{\eta_{\text{qng}}^2\sigma_g^2}{m^2} + \eta_{\text{sgd}}^2\sigma_g^2\right)}_{\Sigma}.$$

□

F.4 Proof of Theorem 5.4

Proof. Rearranging Theorem 5.3:

$$\mathbb{E}\left[C(\overrightarrow{\theta}^t) - C(\overrightarrow{\theta}^{t+1}) \mid \overrightarrow{\theta}^t, M_t\right] \geq \frac{\eta_{\text{qng}}}{2} g_{\mathcal{A}_t}^{t\top} M_t^{-1} g_{\mathcal{A}_t}^t + \frac{\eta_{\text{sgd}}}{2} \|g_{\mathcal{I}_t}^t\|_2^2 - \Sigma. \tag{50}$$

Since $M_t \preceq \Lambda_{\max} I$ and $M_t \succ 0$, inverting the ordering gives $M_t^{-1} \succeq \Lambda_{\max}^{-1} I$. Therefore:

$$g_{\mathcal{A}_t}^{t\top} M_t^{-1} g_{\mathcal{A}_t}^t \geq \frac{1}{\Lambda_{\max}} g_{\mathcal{A}_t}^{t\top} g_{\mathcal{A}_t}^t = \frac{1}{\Lambda_{\max}} \|g_{\mathcal{A}_t}^t\|_2^2.$$

Substituting into Equation (50):

$$\mathbb{E}\left[C(\overrightarrow{\theta}^t) - C(\overrightarrow{\theta}^{t+1}) \mid \overrightarrow{\theta}^t, M_t\right] \geq \frac{\eta_{\text{qng}}}{2\Lambda_{\max}} \|g_{\mathcal{A}_t}^t\|_2^2 + \frac{\eta_{\text{sgd}}}{2} \|g_{\mathcal{I}_t}^t\|_2^2 - \Sigma. \tag{51}$$

Define the minimum effective stepsize $\eta_{\min} = \min\{\eta_{\text{qng}}/\Lambda_{\max}, \eta_{\text{sgd}}\}$. Since $\mathcal{A}_t$ and $\mathcal{I}_t$ partition $[p]$, we have $\|g^t\|_2^2 = \|g_{\mathcal{A}_t}^t\|_2^2 + \|g_{\mathcal{I}_t}^t\|_2^2$, so:

$$\frac{\eta_{\text{qng}}}{2\Lambda_{\max}} \|g_{\mathcal{A}_t}^t\|_2^2 + \frac{\eta_{\text{sgd}}}{2} \|g_{\mathcal{I}_t}^t\|_2^2 \geq \frac{\eta_{\min}}{2} \left(\|g_{\mathcal{A}_t}^t\|_2^2 + \|g_{\mathcal{I}_t}^t\|_2^2\right) = \frac{\eta_{\min}}{2} \|g^t\|_2^2. \tag{52}$$

Combining Equation (51) and Equation (52):

$$\mathbb{E}\left[C(\overrightarrow{\theta}^t) - C(\overrightarrow{\theta}^{t+1}) \mid \overrightarrow{\theta}^t, M_t\right] \geq \frac{\eta_{\min}}{2} \|g^t\|_2^2 - \Sigma. \tag{53}$$

Taking expectation and applying the tower property yields

$$\mathbb{E}\left[C(\overrightarrow{\theta}^t) - C(\overrightarrow{\theta}^{t+1})\right] \geq \frac{\eta_{\min}}{2} \mathbb{E}\left[\|g^t\|_2^2\right] - \Sigma. \tag{54}$$

Summing Equation (54) from $t = 0$ to $T - 1$:

$$\sum_{t=0}^{T-1} \mathbb{E}\left[C(\overrightarrow{\theta}^t) - C(\overrightarrow{\theta}^{t+1})\right] \geq \frac{\eta_{\min}}{2} \sum_{t=0}^{T-1} \mathbb{E}\left[\|g^t\|_2^2\right] - T\Sigma. \tag{55}$$

The left side telescopes by linearity of expectation:

$$\begin{aligned}
\sum_{t=0}^{T-1} \mathbb{E}\left[C(\overrightarrow{\theta}^t) - C(\overrightarrow{\theta}^{t+1})\right] &= \mathbb{E}\left[C(\overrightarrow{\theta}^0) - C(\overrightarrow{\theta}^T)\right] \\
&= C(\overrightarrow{\theta}^0) - \mathbb{E}\left[C(\overrightarrow{\theta}^T)\right] \\
&\leq C(\overrightarrow{\theta}^0) - C_{\inf},
\end{aligned} \tag{56}$$

where the last step uses $C(\overrightarrow{\theta}^T) \geq C_{\inf}$ (bounded below) and $C(\overrightarrow{\theta}^0)$ is deterministic (fixed initialization). Combining Equation (55) and Equation (56):

$$\frac{\eta_{\min}}{2} \sum_{t=0}^{T-1} \mathbb{E}\left[\|g^t\|_2^2\right] \leq C(\overrightarrow{\theta}^0) - C_{\inf} + T\Sigma.$$

Dividing both sides by $T\eta_{\min}/2$:

$$\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\left[\|g^t\|_2^2\right] \leq \frac{2(C(\overrightarrow{\theta}^0) - C_{\inf})}{T\eta_{\min}} + \frac{2\Sigma}{\eta_{\min}}.$$

Since the minimum of a sequence is at most its average, $\min_{0 \leq t < T} \mathbb{E}\left[\|g^t\|_2^2\right] \leq \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\left[\|g^t\|_2^2\right]$, which completes the proof. □

F.5 Effect of PSD projection

In experiments, we apply PSD projection before ridge regularization, using

$$M_t = \Pi_{\succeq 0}\left(\widehat{F}^{(\Omega)}_{\mathcal{A}_t\mathcal{A}_t}\right) + \lambda I.$$

Accordingly, for the experimental algorithm the approximation chain in Equation (9) acquires one additional step:

$$F \xrightarrow{\text{truncation}} F^{(k)}_{\mathcal{A}_t\mathcal{A}_t} \xrightarrow{\text{masking}} F^{(\Omega)}_{\mathcal{A}_t\mathcal{A}_t} \xrightarrow{\text{shadow estimation}} \widehat{F}^{(\Omega)}_{\mathcal{A}_t\mathcal{A}_t} \xrightarrow{\text{PSD proj.}} \Pi_{\succeq 0}\left(\widehat{F}^{(\Omega)}_{\mathcal{A}_t\mathcal{A}_t}\right) \xrightarrow{\text{ridge}} M_t.$$

This adds the stabilization term

$$E_{\mathrm{psd},t} = \left\|\Pi_{\succeq 0}\left(\widehat{F}^{(\Omega)}_{\mathcal{A}_t\mathcal{A}_t}\right) - \widehat{F}^{(\Omega)}_{\mathcal{A}_t\mathcal{A}_t}\right\|_{\mathrm{op}},$$

which, for a symmetric matrix, is controlled by the negative spectrum and in particular by the magnitude of the most negative eigenvalue. Empirically, Table 4 indicates that this correction is modest: although a nontrivial fraction of eigenvalues are negative before projection, the negative spectral mass is small, suggesting that PSD projection mainly removes a numerically noisy tail rather than substantial curvature information.

For convergence, PSD projection simplifies the lower spectral bound. Since

$$\Pi_{\succeq 0}\left(\widehat{F}^{(\Omega)}_{\mathcal{A}_t\mathcal{A}_t}\right) \succeq 0,$$

we have

$$M_t = \Pi_{\succeq 0}\left(\widehat{F}^{(\Omega)}_{\mathcal{A}_t\mathcal{A}_t}\right) + \lambda I \succeq \lambda I.$$

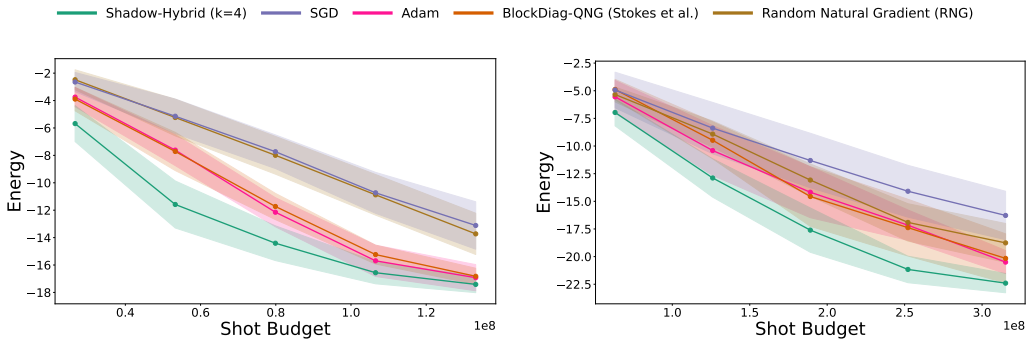
Moreover,

$$\begin{aligned}\|M_t\|_{\mathrm{op}} &\le \left\|\Pi_{\succeq 0}\left\{\widehat{F}^{(\Omega)}_{\mathcal{A}_t\mathcal{A}_t}\right\}\right\|_{\mathrm{op}} + \lambda \\ &\le \left\|\widehat{F}^{(\Omega)}_{\mathcal{A}_t\mathcal{A}_t}\right\|_{\mathrm{op}} + \lambda,\end{aligned}$$

since PSD projection does not increase the operator norm. Thus, the same spectral-bound argument used in Theorems 5.3 and 5.4 extends to the PSD-projected preconditioner used in experiments.

G Additional Experiments

G.1 Energy at fixed shot budgets



(a) Energy at fixed shot budgets on the 12-qubit, depth-6 Heisenberg model.

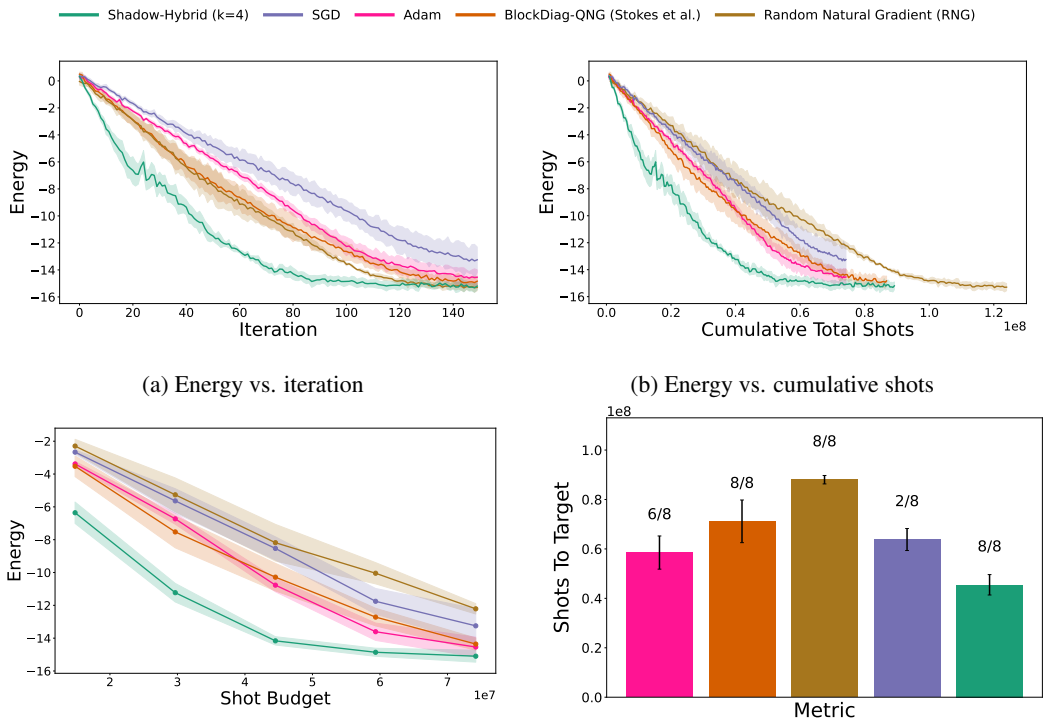
(b) Energy at fixed shot budgets on the 16-qubit, depth-8 Heisenberg model.

Figure 3: Energy at fixed cumulative shot budgets.

Figures 3a and 3b show the energy achieved at fixed cumulative shot budgets for the 12-qubit and 16-qubit Heisenberg instances. The plotted values report the mean over 8 matched random seeds,

with shaded bands showing 95% confidence intervals. Shadow-Hybrid achieves the lowest energy at moderate budgets, where the curvature information provides the largest marginal benefit. At very low budgets, metric-estimation overhead is not yet amortized; at very high budgets, all methods eventually converge.

G.2 10-qubit Heisenberg



(a) Energy vs. iteration

(b) Energy vs. cumulative shots

(c) Energy at fixed shot budgets

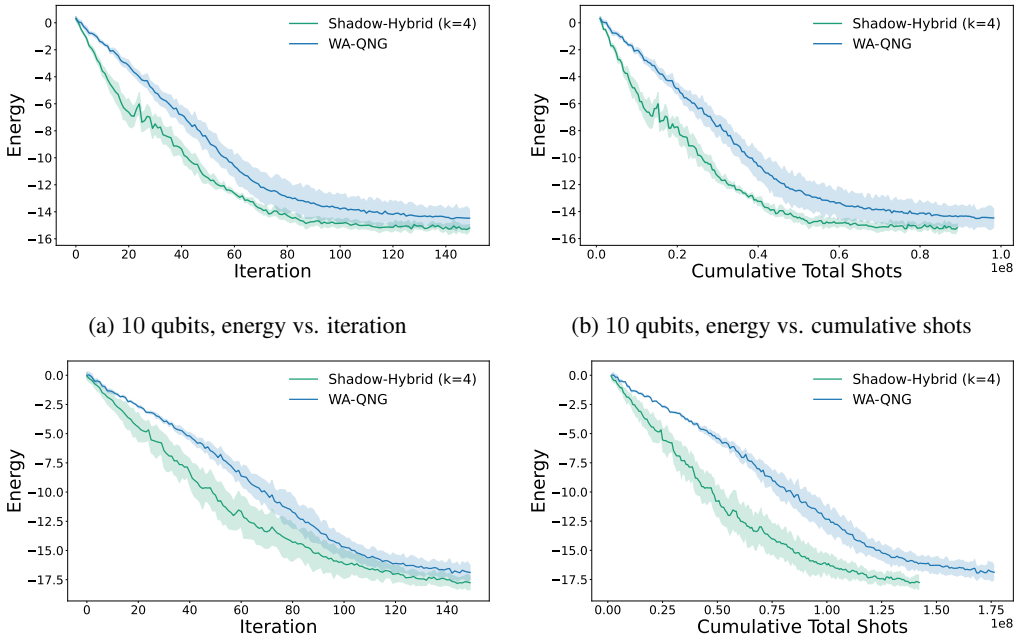
(d) Shots to reach target (-14.0)

Figure 4: Convergence on the 10-qubit, depth-4 Heisenberg model over 8 matched random seeds. (a) Energy vs. iteration. (b) Energy vs. cumulative shots. (c) Energy at fixed shot budgets. (d) Shots to reach preset benchmark target (-14.0); fractions indicate seeds (out of 8) reaching the target. The target is a fixed threshold, not the true ground-state energy. Shaded bands in (a)–(c) show 95% confidence intervals over seeds. Error bars in (d) show mean $\pm$ SD over successful seeds only.

Figure 4 shows convergence on the 10-qubit, depth-4 Heisenberg instance, the smallest Heisenberg setting in our experiments. Shadow-Hybrid achieves lower energy at matched cumulative shot budgets than all baselines, with faster per-iteration convergence early in optimization. At this scale, several methods reach the preset benchmark target reliably: Shadow-Hybrid, RNG, and BlockDiag-QNG each succeed in 8 of 8 seeds, Adam succeeds in 6 of 8, and SGD succeeds in 2 of 8. Among successful runs, Shadow-Hybrid reaches the target with the lowest average cumulative shot cost. The per-iteration differences between Shadow-Hybrid, Adam, and the QNG baselines narrow in the later stages of this smaller instance, while the methods remain more clearly separated in the 16-qubit (Figure 2) case, where curvature information becomes more valuable.

G.3 Comparison with WA-QNG

WA-QNG [23] is a locality-aware baseline developed for local Hamiltonians, although it uses a different metric construction from our QFIM surrogate. We compare the two methods using the same ansatz, matched random seeds, parameter-shift gradients, energy-estimation protocol, and cumulative-shot accounting.



(a) 10 qubits, energy vs. iteration

(b) 10 qubits, energy vs. cumulative shots

(c) 12 qubits, energy vs. iteration

(d) 12 qubits, energy vs. cumulative shots

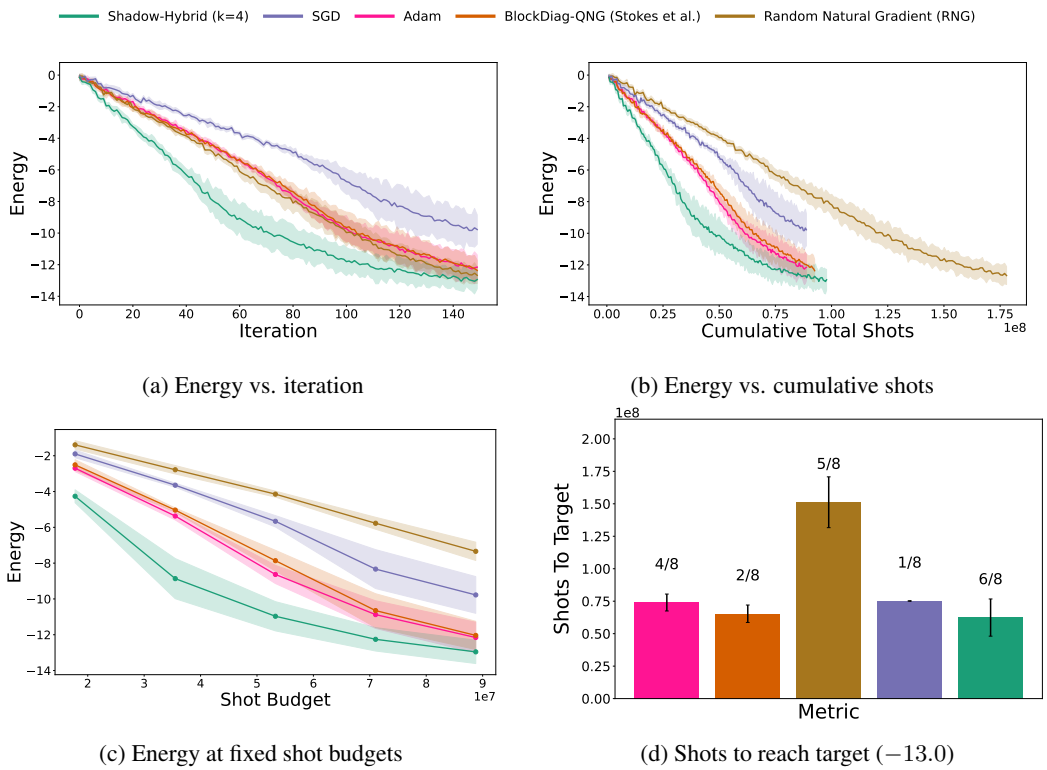
Figure 5: Comparison of Shadow-Hybrid and WA-QNG on the 1D Heisenberg model over 8 matched random seeds. The 10-qubit instance has depth 4, and the 12-qubit instance has depth 6. Shaded bands show 95% confidence intervals over seeds.

On both instances, Shadow-Hybrid reaches lower mean energy earlier in the optimization. The cumulative-shot curves extend to different total budgets because the methods have different per-iteration measurement costs. Over the budget range shared by both methods, Shadow-Hybrid reaches lower mean energy for nearly the entire range.

For the preset benchmark target of -14.0 on the 10-qubit instance, Shadow-Hybrid succeeds in 8 of 8 seeds, compared with 6 of 8 for WA-QNG. Among successful runs, the corresponding mean shots-to-target are 4.57×10^7 and 5.78×10^7. For the target of -17.0 on the 12-qubit instance, Shadow-Hybrid succeeds in 7 of 8 seeds, compared with 4 of 8 for WA-QNG, with mean shots-to-target of 9.91×10^7 and 1.42×10^8. Thus, among successful runs, Shadow-Hybrid has about 21% lower mean shots-to-target at 10 qubits and about 30% lower mean shots-to-target at 12 qubits, while also reaching the targets on more seeds.

G.4 12-qubit Ising Hamiltonian

To test generalization beyond the Heisenberg model, we evaluate on the 12-qubit 1D transverse-field Ising Hamiltonian, $H = \sum_{i=0}^{n-2} JZ_iZ_{i+1} + h\sum_{i=0}^{n-1} X_i$, with $J = h = 1.0$, using the same ansatz and hyperparameters. Figure 6 shows the convergence results. The qualitative pattern is consistent with the Heisenberg experiments. Shadow-Hybrid achieves lower energy at a given cumulative shot budget than all baselines, showing the same per-shot advantage observed in the 10 and 12-qubit Heisenberg cases, with the lowest energy at moderate shot budgets. In the per-iteration view, it remains broadly competitive with RNG, with both ahead of the remaining baselines. It reaches the preset benchmark target energy in 6 of 8 seeds, while RNG reaches it in 5 of 8, Adam in 4 of 8, BlockDiag-QNG in 2 of 8 and SGD in 1 of 8 seeds. Among the methods that reach the target, Shadow-Hybrid does so at the lowest average cumulative shot cost. These results indicate that the locality-aware preconditioning generalizes beyond a single Hamiltonian family.



(a) Energy vs. iteration

(b) Energy vs. cumulative shots

(c) Energy at fixed shot budgets

(d) Shots to reach target (-13.0)

Figure 6: Convergence on the 12-qubit, depth-6 Ising model over 8 matched random seeds. (a) Energy vs. iteration. (b) Energy vs. cumulative shots. (c) Energy at fixed shot budgets. (d) Shots to reach preset benchmark target (-13.0); fractions indicate seeds (out of 8) reaching the target. The target is a fixed threshold, not the true ground-state energy. Shaded bands in (a)–(c) show 95% confidence intervals over seeds. Error bars in (d) show mean $\pm$ SD over successful seeds only.

G.5 Ablations

G.5.1 Weight-k truncation

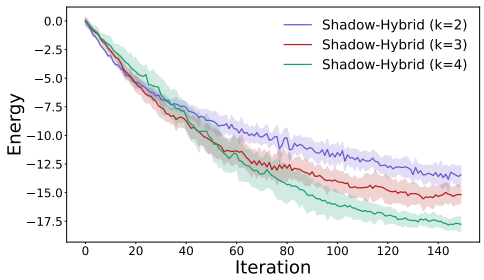
The truncation level k acts as a bias–variance knob in the surrogate metric. Figure 7a shows convergence trajectories for $k \in \{2, 3, 4\}$ on the 12-qubit Heisenberg instance. Lower k converges faster in early iterations due to reduced estimation noise, but plateaus at a higher energy because the truncated metric captures less curvature structure. Higher k reaches lower final energies at the cost of wider variance across seeds, reflecting the increased shot noise from estimating higher-weight Pauli observables.

Choosing k under a fixed shot budget. The appropriate value of k depends on the circuit and the available shot budget. Table 2 summarizes a practical fixed-budget guideline. This is a screening heuristic rather than an automatic selection rule.

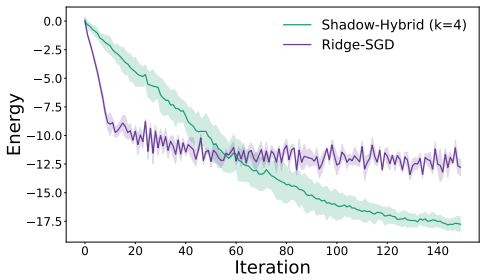
G.5.2 Ridge baselines

We compare against two ridge baselines to isolate the contribution of curvature information from the effect of stepsize scaling via the ridge parameter λ.

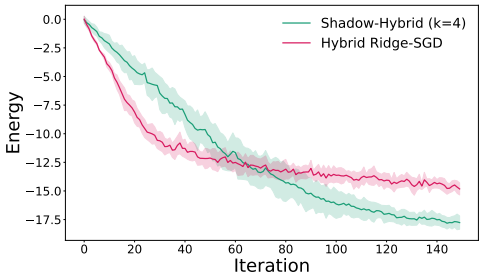
Ridge-SGD (Figure 7b) applies $M_t = \lambda I$ to all parameters, using the same stepsizes and ridge value as Shadow-Hybrid. This is equivalent to rescaled gradient descent with no curvature information. Ridge-SGD oscillates and stalls around -10, while Shadow-Hybrid continues to converge past -17, confirming that the improvement is not due to stepsize scaling alone.



(a) Weight-k truncation Ablation



(b) Ridge Ablation



(c) Hybrid Ridge Ablation

Figure 7: Ablation convergence on 12-qubit, depth-6 1D Heisenberg model over 8 matched random seeds. Shaded bands show 95% confidence intervals over seeds.

Stage	Diagnostic	Decision
Screen	Retained-term count, active-set size, number of retained masked entries, and maximum support-overlap degree	Filter out candidate values that are too costly for the available computational or measurement budget.
Stability test	Compare preconditioned update directions from independent shadow batches at the same parameter point	Estimate finite-shot variability for each feasible k.
Choose	Compare the change between adjacent k values with the observed shadow variability	Prefer the smaller k when the change is no larger than the observed shadow variability. Consider the larger k only when the change exceeds that variability and its estimate remains stable and affordable.

Table 2: Practical fixed-budget guideline for selecting the truncation level k. This is a screening heuristic, not a universal or automatic selection rule.

Hybrid-Ridge-SGD (Figure 7c) uses the same active/inactive splitting as Shadow-Hybrid but replaces the shadow-estimated metric with $M_t = \lambda I$ on the active block. Active parameters receive $\Delta_{\mathcal{A}_t} = -\eta_{\text{qng}}(\lambda I)^{-1}\hat{g}_{\mathcal{A}_t}$ and inactive parameters receive $\Delta_{\mathcal{I}_t} = -\eta_{\text{sgd}}\hat{g}_{\mathcal{I}_t}$. This ablation controls for the hybrid update structure. Hybrid-Ridge-SGD also stalls, indicating that under matched hyperparameters, the active/inactive splitting alone does not reproduce Shadow-Hybrid's convergence — the shadow-estimated curvature on the active block contributes materially to the gains observed.

G.5.3 Masking diagnostics

k	$\lvert\mathcal{A}_t\rvert/p$	Mask density on $[p]\times[p]$		Mask density on $\mathcal{A}_t\times\mathcal{A}_t$		Relative op.-norm loss	Frobenius retention
		diag+off-diag	off-diag only	diag+off-diag	off-diag only	$\frac{\lVert F^{(k)}-F^{(\Omega)}\rVert_{\mathrm{op}}}{\lVert F^{(k)}\rVert_{\mathrm{op}}}$	$\frac{\lVert F^{(\Omega)}\rVert_F^2}{\lVert F^{(k)}\rVert_F^2}$
2	0.5833	0.1233	0.1171	0.3537	0.3460	0.1613 ± 0.0536	0.9786 ± 0.0076
3	0.6250	0.2005	0.1949	0.5067	0.5011	0.1771 ± 0.0512	0.9767 ± 0.0069
4	0.6667	0.2394	0.2341	0.5334	0.5285	0.1848 ± 0.0530	0.9781 ± 0.0066

Table 3: Masking diagnostics versus truncation level k on the 12-qubit, depth-6 1D Heisenberg model, averaged over iterations and seeds. For each index domain ($[p]\times[p]$ and $\mathcal{A}_t\times\mathcal{A}_t$), "diag+off-diag" denotes mask density over all entries, while "off-diag only" denotes mask density restricted to off-diagonal entries. The final columns report the relative operator-norm masking loss and Frobenius retention ratio induced by applying the support-overlap mask to the already truncated surrogate $F^{(k)}$. They do not measure truncation error relative to the full QFIM.

Table 3 reports mask density and Frobenius retention as a function of k. The active fraction $|\mathcal{A}_t|/p$ increases with k (more parameters have nonzero truncated generators), and the mask becomes denser as higher-weight terms create more support overlaps. Despite this, the mask remains well below full density ($\sim$53% on $\mathcal{A}_t\times\mathcal{A}_t$ at $k=4$), and the Frobenius retention exceeds 97.6% at all k. The relative operator-norm masking loss is stable at $\sim$17%, indicating that the discarded off-mask entries carry modest spectral weight.

These diagnostics are computed from the exact (simulation-accessible) truncated metric $F^{(k)}$, not from shadow estimates, and are averaged over iterations and seeds.

G.5.4 Curvature diagnostics

Table 4 reports a set of spectral diagnostics for our proposed Shadow-Hybrid method at $k=4$ on the 12-qubit depth-6 Heisenberg instance. The goal is to verify that the shadow-estimated surrogate metric carries meaningful curvature beyond the ridge regularization term, rather than acting as a scaled identity or stepsize rescaling. These diagnostics complement the ridge-baseline ablations by showing that the improvement is driven by curvature information in the surrogate.

Metric	**Definition**	**Across seeds (mean $\pm$ sd)**	**Interpretation**
Curvature vs. ridge scale	$\lambda_{\max}/\lambda$	25.2197 ± 2.3004	Curvature dominates ridge.
Strong eigenmode fraction	$\#\{\lambda_i > \lambda\}/\lvert\mathcal{A}_t\rvert$	0.5670 ± 0.0310	Majority of modes carry curvature above ridge floor.
Raw indefiniteness	$\#\{\lambda_i < 0\}/\lvert\mathcal{A}_t\rvert$	0.1276 ± 0.0116	The raw estimate is often indefinite, so PSD projection may help in stabilization.
Negative spectral mass ratio	$\frac{\sum_{\lambda_i<0}\lvert\lambda_i\rvert}{\sum_i\lvert\lambda_i\rvert}$	2.2978×10^{-2} $\pm\, 4.2778\times10^{-3}$	Small but nonzero; noise creates some negative eigenvalues but they carry little total weight.
Directional alignment signal	$\frac{\langle\Delta\theta,\Delta\theta_{\mathrm{ridge}}\rangle}{\lVert\Delta\theta\rVert\lVert\Delta\theta_{\mathrm{ridge}}\rVert}$	0.6110 ± 0.0141	Update differs from plain ridge direction; curvature rotates the step.
Energy drop	$\Delta E = E_{150} - E_1$	-17.1074 ± 1.1454	Consistent descent across seeds.

Table 4: Curvature diagnostics for our Shadow-Hybrid method with $k=4$ on the 12-qubit, depth-6 1D Heisenberg model across multiple seeds. Reported values are mean $\pm$ sd across runs. The table summarizes curvature scale, indefiniteness, directional deviation from a ridge-only step, and observed descent.

Table 4 confirms that the surrogate carries meaningful and non-trivial curvature information:

Curvature scale: The largest eigenvalue exceeds the ridge floor by $\sim$25$\times$, indicating that the surrogate captures substantial curvature beyond the regularization baseline.

Strong eigenmodes: 57% of eigenmodes lie above the ridge threshold, so the majority of the active parameter space receives genuine curvature-informed preconditioning.

Indefiniteness: $\sim$13% of eigenvalues are negative before PSD projection, validating the need for this stabilization step. The negative spectral mass is small ($\sim$2.3%), so PSD projection corrects a numerically noisy tail without discarding significant curvature.

Directional alignment: The cosine similarity between the preconditioned update and the ridge-only update is $\sim$0.61, confirming that the curvature rotates the step direction substantially away from the plain gradient.

Descent: All seeds achieve consistent energy drops of $\sim$17 units over 150 iterations.

G.5.5 Scaling diagnostics

We examine how the support-overlap mask behaves at fixed $k = 4$ as system size grows, evaluating Heisenberg instances with $(n, d) \in \{(10, 4), (12, 6), (14, 7), (16, 8)\}$. Table 5 reports the active-set size, mask density, masking-induced operator-norm loss, Frobenius retention, and the wall-clock time needed to build the truncated propagated generators.

n	depth	p	$\lvert\mathcal{A}_t\rvert$	$\lvert\mathcal{A}_t\rvert/p$	mask density	rel. op.-loss	Frob. ret.	T_{prop} (s)
10	4	80	56	0.7000	0.2581	0.1920 ± 0.1017	0.9685 ± 0.0293	0.136 ± 0.025
12	6	144	96	0.6667	0.2394	0.1826 ± 0.0545	0.9787 ± 0.0070	1.152 ± 0.023
14	7	196	126	0.6429	0.2124	0.1192 ± 0.0541	0.9873 ± 0.0058	2.863 ± 0.134
16	8	256	160	0.6250	0.1895	0.1008 ± 0.0252	0.9929 ± 0.0029	5.168 ± 0.034

Table 5: Scaling diagnostics for Heisenberg chains at fixed truncation $k = 4$, varying (n, d). Mask density reports $|\Omega_k|/p^2$ on $[p] \times [p]$. Means $\pm$ std are over seeds $\{5, 7, 9, 11\}$ and snapshots $t \in \{1, 50, 100, 150\}$. The masking-loss and Frobenius-retention columns measure the effect of applying the support-overlap mask to the already truncated surrogate $F^{(k)}$, not truncation error relative to the full QFIM. T_{prop} is the wall-clock time for one truncated propagated-generator construction.

The mask becomes sparser as the system grows: its density on $[p] \times [p]$ decreases from 0.2581 at $(10, 4)$ to 0.1895 at $(16, 8)$. Frobenius retention moves the other way, from 0.9685 to 0.9929. So at fixed $k = 4$, the mask discards more entries at larger scales while keeping nearly all of the Frobenius mass of $F^{(k)}$. This matches the locality of the truncated surrogate: each $\widetilde{G}_j^{(k)}$ has bounded Pauli weight, so support-overlap masking retains only the parameter pairs whose truncated supports actually intersect.

Classical preprocessing time T_{prop} grows from 0.136 s at $(10, 4)$ to 5.168 s at $(16, 8)$. This is only the cost of building the truncated generators; it does not include quantum shots or full optimizer runtime.

Two caveats. First, n and depth increase together in this sweep, so the table is a fixed-k operating-point study, not an isolated scaling law in n. Second, what we measure is the mask's quality after truncation, not the truncation error relative to the full QFIM. Thus, while the fixed choice $k = 4$ works well across the 10-, 12-, and 16-qubit Heisenberg studies reported here, we do not claim that it remains sufficient for arbitrarily larger or deeper circuits.

These diagnostics are computed from the exact, simulation-accessible truncated metric $F^{(k)}$, rather than from finite-shot shadow estimates.

G.6 Small instance QFIM diagnostics

We use a 6-qubit, depth-4 Heisenberg instance, where exact statevector evaluation of the dense QFIM and intermediate surrogate metrics is tractable.

G.6.1 Comparison with the dense exact QFIM

Estimator	Rel. op. err.	Rel. Fro. err.	Rel. Eig. spectrum err.
Shadow surrogate, $k = 2$	0.783 ± 0.041	0.802 ± 0.023	0.520 ± 0.021
Shadow surrogate, $k = 3$	0.638 ± 0.034	0.657 ± 0.030	0.365 ± 0.012
Shadow surrogate, $k = 4$	0.566 ± 0.019	0.623 ± 0.017	0.209 ± 0.016
BlockDiag-QNG (exact)	0.743 ± 0.037	0.682 ± 0.029	0.537 ± 0.031
RNG (exact)	0.730 ± 0.067	0.651 ± 0.027	0.464 ± 0.029

Table 6: Direct comparison with the dense exact QFIM F on the 6-qubit, depth-4 Heisenberg instance. Each shadow-surrogate row uses 10^4 snapshots, while the reference QFIM and baseline metrics are evaluated exactly by statevector simulation. Here, $\widetilde{F}$ denotes the metric reported in the corresponding row. The relative operator and Frobenius errors are $\|\widetilde{F} - F\|_{\text{op}}/\|F\|_{\text{op}}$ and $\|\widetilde{F} - F\|_F/\|F\|_F$. The relative spectrum error is $\|\boldsymbol{\lambda}(\widetilde{F} + 0.1I) - \boldsymbol{\lambda}(F + 0.1I)\|_2/\|\boldsymbol{\lambda}(F + 0.1I)\|_2$, where $\boldsymbol{\lambda}(\cdot)$ denotes the sorted eigenvalue vector. Values are mean $\pm$ SD over 5 circuit seeds.

Table 6 shows that the shadow surrogate becomes more accurate as k increases. At $k = 4$, it has the lowest mean error across all three metrics. At $k = 3$, it has lower mean errors than BlockDiag-QNG across all three metrics and lower operator-norm and spectrum errors than RNG/CFIM, with comparable Frobenius error. This indicates that the retained cross-parameter couplings preserve useful structure from the dense QFIM on this instance.

G.6.2 Component-wise error decomposition

We next separate the truncation, masking, and finite-shot components of the shadow surrogate error. The total error column in Table 7 corresponds to the operator-norm error of the shadow surrogate rows in Table 6.

Estimator	Rel. trunc. op. err.	Rel. mask op. err.	Rel. shot op. err.	Rel. total op. err.
Shadow surrogate, $k = 2$	0.761 ± 0.044	0.150 ± 0.036	0.078 ± 0.015	0.783 ± 0.041
Shadow surrogate, $k = 3$	0.613 ± 0.041	0.154 ± 0.029	0.149 ± 0.030	0.638 ± 0.034
Shadow surrogate, $k = 4$	0.540 ± 0.023	0.159 ± 0.031	0.249 ± 0.026	0.566 ± 0.019

Table 7: Component-wise operator-norm diagnostic for the 6-qubit, depth-4 Heisenberg instance at a fixed 10^4-snapshot budget. The columns report $\|F - F^{(k)}\|_{\text{op}}/\|F\|_{\text{op}}$, $\|F^{(k)} - F^{(\Omega)}\|_{\text{op}}/\|F\|_{\text{op}}$, $\|\widehat{F}^{(\Omega)} - F^{(\Omega)}\|_{\text{op}}/\|F\|_{\text{op}}$, and $\|\widehat{F}^{(\Omega)} - F\|_{\text{op}}/\|F\|_{\text{op}}$, respectively. Values are mean $\pm$ SD over 5 seeds.

Increasing k from 2 to 4 reduces the truncation error, while the finite shot error increases at the same snapshot budget. The masking error changes only modestly, and the total error decreases over the tested values. The three component errors are not additive parts of the total error. Their sum gives a triangle-inequality upper bound, while the measured total can be smaller because the residual matrices need not align.

The exact decomposition requires the dense QFIM and is therefore used only as a small-instance diagnostic. For larger instances, Table 2 gives a practical guideline based on quantities available after propagation and truncation and recommends checking stability across independent shadow batches.

G.6.3 Dense connectivity diagnostic

Finally, we compare the nearest-neighbor circuit with an all-to-all ansatz of the same size and depth. For the all-to-all variant, we retain the same single-qubit R_y and R_z rotation layers and replace each nearest-neighbor CNOT entangling layer with an all-to-all layer containing one CNOT for every pair of qubits, using the lower qubit index as control and the higher qubit index as target. Table 8 uses the same error definitions as Table 6.

This small instance experiment shows a connectivity-dependent operating-point tradeoff rather than an abrupt failure outside 1D. On this instance, the all-to-all architecture gives a less accurate $k = 4$ surrogate in operator norm, while increasing the truncation level to $k = 5$ recovers much of the

Connectivity	Rel. op. err.	Rel. Fro. err.	Rel. eig. spectrum err.
1D, $k = 4$	0.566 ± 0.019	0.623 ± 0.017	0.209 ± 0.016
All-to-all, $k = 4$	0.649 ± 0.085	0.634 ± 0.051	0.217 ± 0.031
All-to-all, $k = 5$	0.574 ± 0.085	0.614 ± 0.040	0.195 ± 0.027

Table 8: Comparison with the dense exact QFIM for 6-qubit, depth-4 nearest-neighbor and all-to-all ansatzes. Each shadow-surrogate estimate uses 10^4 snapshots. Values are mean $\pm$ SD over 5 circuit seeds.

lost approximation fidelity by retaining more propagated-generator terms. We therefore do not claim uniform scaling across arbitrary dense architectures. The advantage depends on whether the retained-term count and support-overlap degree remain computationally manageable.

H Implementation details

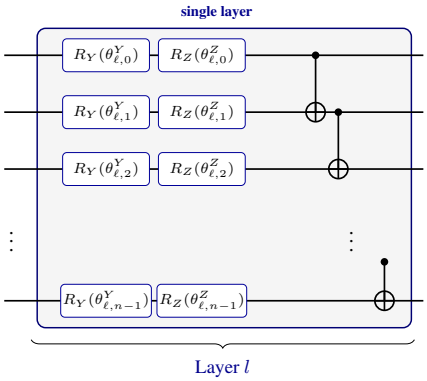


Figure 8: A single layer of the hardware-efficient ansatz used in the experiments. Each layer applies trainable R_Y and R_Z rotations on every qubit, followed by a nearest-neighbor CNOT ladder. The full circuit is obtained by repeating this layer for $l = 1, \ldots, d$.

All experiments were implemented in Python using PennyLane [2] (v0.44.0) with finite shot simulation and were run on a multi-core CPU computing cluster. We used the same HEA (Figure 8), initialization, and finite-shot parameter-shift gradient pipeline across all methods. Metric-based methods (BlockDiag-QNG, RNG, and ours) additionally re-estimate a preconditioner at each iteration. SGD and Adam use only gradient shots. Learning rates were tuned independently per method over $[0.01,\ 0.1]$, and we report the best setting for each method. For our method, the inactive-block learning rate η_{sgd} was fixed to the best tuned learning rate of the SGD baseline, and only the active-block learning rate η_{qng} was tuned. We use weight-k truncation together with PSD projection and ridge regularization on the active block with ridge parameter $\lambda = 0.1$ unless otherwise stated. For baseline metric methods, we use their corresponding preconditioner constructions with ridge regularization for stable inversion.

For shot accounting, we use $S_{\text{grad}} = 1024$ shots per qubit-wise commuting (QWC) measurement group per shifted parameter-shift circuit for all methods. Metric-based baselines use $S_{\text{met}} = 2048$ shots per metric-estimation circuit. Corollary 5.2 provides a conservative worst-case sufficient condition for per-iteration operator-norm accuracy, so we do not use it to set shadow snapshot counts numerically. Instead, after fixing k, we choose the shadow snapshot count N empirically as a stable operating point for surrogate-QNG optimization. Our method uses $N = 50000$ single-shot classical-shadow snapshots for the 10-qubit experiments, $N = 60000$ snapshots for the 12-qubit experiments and $N = 100000$ snapshots for the 16-qubit experiments. Each snapshot is a single-shot measurement under a random single-qubit Pauli basis, so N snapshots correspond to N total shots.

In Table 9, B_{BD} denotes the number of blocks used by BlockDiag-QNG. For uniform block size q, $B_{\text{BD}} = p/q$. In the 12-qubit setting with $p = 144$, we use $B_{\text{BD}} = 12$ uniform blocks of size $q = 12$, corresponding to the 12 layer-and-rotation-type subgroups: each of the 6 circuit layers contributes

Method	Gradient	Metric estimation	Total per iteration
SGD	$2pS_{\text{grad}}$	—	$2pS_{\text{grad}}$
Adam	$2pS_{\text{grad}}$	—	$2pS_{\text{grad}}$
BlockDiag-QNG	$2pS_{\text{grad}}$	$B_{\text{BD}}\, S_{\text{met}}$	$2pS_{\text{grad}} + B_{\text{BD}}\, S_{\text{met}}$
RNG	$2pS_{\text{grad}}$	$(2p+1)\, S_{\text{met}}$	$2pS_{\text{grad}} + (2p+1)\, S_{\text{met}}$
Ours Shadow-Hybrid	$2pS_{\text{grad}}$	N shadow snapshots	$2pS_{\text{grad}} + N$

Table 9: Per-iteration shot budget

one block of R_Y rotations and one block of R_Z rotations. The quantity S_{grad} is defined per QWC group, so the absolute gradient cost includes an additional shared factor equal to the number of QWC groups, namely 3 for the Heisenberg Hamiltonian and 2 for the transverse-field Ising Hamiltonian. This factor is common to all methods and is omitted from Table 9 for readability.

At 12-qubits ($p = 144$), our metric-estimation cost is $N = 60000$ shadow snapshots per iteration. For comparison, RNG requires $(2p+1)S_{\text{met}} = 289 \times 2048 \approx 5.92 \times 10^5$ metric-estimation shots per iteration, whereas BlockDiag-QNG requires $B_{\text{BD}}S_{\text{met}} = 12 \times 2048 \approx 2.46 \times 10^4$ shots. Thus, our method sits between BlockDiag-QNG and RNG in per-iteration metric-estimation cost. It is about $2.4\times$ more expensive than BlockDiag-QNG, but about $9.9\times$ cheaper than RNG, while still capturing inter-block curvature that block-diagonal estimation discards. RNG's metric-estimation cost scales explicitly with p, whereas our method's cost depends on N and the truncation level k. The practical payoff is captured by the energy reached per cumulative shot. Consistent with this cost comparison, our method reaches lower energies per cumulative shot on both tested Heisenberg and Ising instances, see Figures 2 and 6.