SGIQ: SPARSE-GRADIENT-INFORMED GRADIENT ESTIMATION FOR VARIATIONAL QUANTUM CIRCUITS

Anonymous authors
Paper under double-blind review

ABSTRACT

Gradient estimation is a major measurement bottleneck in training complex variational quantum circuits (VQCs) because backpropagation scaling is unattainable for general cases with unknown quantum inputs and no quantum memory. Existing gradient estimators therefore face a cost-accuracy trade-off: parameter-shift rule (PSR) incurs a cost linear in the parameter dimension, while simultaneous perturbation stochastic approximation (SPSA) uses few evaluations but can suffer from dimension-dependent variance. To improve this trade-off, we exploit settings in which loss-gradient energy concentrates on a small subset of parameters. We formalize this property through gradient-energy sparsity and derive sufficient conditions from causal-cone, commutation, and response-decay mechanisms. Building on this structure, we propose SGIQ, a **S**parse-**G**radient-**i**nformed **Q**uantum gradient estimation algorithm that uses averaged SPSA estimates to rank all circuit parameters, replaces the selected coordinate estimates with PSR gradients, and retains the SPSA estimates elsewhere. Its cost scales with the circuit's gradient structure, rather than the parameter dimension. In ideal and finite-shot regression experiments, SGIQ uses fewer gradient circuits than PSR and attains lower normalized test mean absolute error (MAE) than SPSA baselines.

1 INTRODUCTION

Gradient estimation is a major measurement bottleneck in training variational quantum circuits (VQCs), whose measured outputs define trainable models for quantum machine learning and variational optimization (Mitarai et al., 2018; Cerezo et al., 2021a). Unlike classical backpropagation, derivatives through a measured quantum circuit generally require additional quantum evaluations, and backpropagation scaling is unattainable for general quantum models with unknown quantum inputs and no quantum memory (Abbas et al., 2023). The parameter-shift rule (PSR) provides a standard approach to computing VQC gradients (Schuld et al., 2019; Izmaylov et al., 2021), but its circuit cost scales linearly with the number of trainable parameters. For d trainable parameters, a mini-batch of B samples, and Q separately evaluated measurement groups per sample, full PSR requires $2BQd$ gradient-circuit evaluations per optimizer step. In contrast, simultaneous perturbation stochastic approximation (SPSA) (Spall, 1992) estimates all gradient coordinates from m paired perturbations using only $2BQm$ evaluations, independent of d when m is fixed. This reduction in cost, however, comes at the expense of estimation accuracy: each SPSA coordinate estimate mixes contributions from other coordinates, leading to dimension-dependent variance, while finite-shot measurements introduce additional noise. More generally, reducing the per-step directional-gradient cost does not necessarily yield a proportional reduction in total training cost, since parameter dependence can reappear through estimator variance or convergence (Abbas et al., 2023; Coyle et al., 2026). Thus, PSR and SPSA expose a fundamental cost-accuracy trade-off in quantum gradient estimation.

This trade-off raises a natural question: *must all gradient coordinates be estimated with equal accuracy?* For structured VQCs, the answer can be no. Known causal-cone, commutation, and response-locality arguments show that some parameters can have zero or small derivatives (Cerezo et al., 2021b; Uvarov & Biamonte, 2021; Heidari et al., 2025; Else et al., 2020). These structures can cause the loss-gradient energy to concentrate on a small effective coordinate set of size $k \ll d$. In this regime, estimating every coordinate with equal precision is unnecessary, suggesting that gradient-evaluation resources should instead be allocated according to coordinate importance. We

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

formalize this structure through *gradient-energy sparsity* and derive sufficient conditions under which exact or approximate gradient-energy concentration arises.

Building on this structure, we propose SGIQ, a Sparse-Gradient-Informed Quantum gradient estimation algorithm that combines the global coverage of SPSA with the coordinate-wise accuracy of PSR. SGIQ uses averaged SPSA estimates as a low-cost probe over all d circuit parameters and ranks the coordinates by their estimated gradient magnitudes. It then replaces the SPSA estimates on the selected top-k coordinates with PSR gradients, while retaining the SPSA estimates elsewhere. This allocation reduces the per-step gradient-circuit cost to a fraction $(m + k)/d$ of that of full PSR.

Existing measurement-efficient approaches explore different ways of reducing or allocating quantum gradient evaluations. Random coordinate descent evaluates randomly selected gradient coordinates (Ding et al., 2024), while Guided-SPSA combines SPSA and PSR across training samples (Periyasamy et al., 2024). Coyle et al. (Coyle et al., 2026) develop a forward-gradient framework with adaptive measurement allocation. In contrast, SGIQ allocates accurate gradient evaluations across parameter coordinates according to their estimated importance while retaining an SPSA estimate over the full parameter space.

We summarize our contributions as follows. First, we formalize gradient-energy sparsity and derive sufficient conditions from known circuit mechanisms. Second, we introduce SGIQ, which selectively allocates accurate PSR evaluations to important gradient coordinates identified by SPSA. Third, we demonstrate gradient-energy concentration in structured VQC regression models and show, under both ideal and finite-shot simulation, that SGIQ reduces gradient-circuit cost relative to full PSR while outperforming SPSA.

2 PRELIMINARIES

For circuit parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d)$, define $\rho(\boldsymbol{\theta}) = U(\boldsymbol{\theta})\rho_0 U^\dagger(\boldsymbol{\theta})$, where ρ_0 is the input density operator, $U(\boldsymbol{\theta})$ is the parameterized unitary, and $\rho(\boldsymbol{\theta})$ is the output density operator. For a Hermitian observable O, define the measured expectation value $E_O(\boldsymbol{\theta}) = \mathrm{Tr}[O\rho(\boldsymbol{\theta})]$. In supervised learning, measured outputs are combined through a classical readout and loss function to form the scalar mini-batch loss $L(\boldsymbol{\theta})$. The mini-batch and classical readout parameters are held fixed when estimating $\boldsymbol{g} = \nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta})$.

Assume that each parameter appears in a single Pauli-generated rotation $U_j(\theta_j) = e^{-i\theta_j P_j/2}$, with $P_j^2 = I$. For the jth standard basis vector $\boldsymbol{e}_j \in \mathbb{R}^d$, PSR gives

$$\partial_j E_O(\boldsymbol{\theta}) = \frac{1}{2}\left[E_O\Big(\boldsymbol{\theta} + \frac{\pi}{2}\boldsymbol{e}_j\Big) - E_O\Big(\boldsymbol{\theta} - \frac{\pi}{2}\boldsymbol{e}_j\Big)\right].$$

This identity is exact for exact expectation values (Schuld et al., 2019). Finite-shot evaluations add sampling error, and extensions to broader generators are available (Izmaylov et al., 2021). In supervised learning, PSR differentiates the measured outputs, and the chain rule yields the loss gradient. SPSA perturbs all parameters simultaneously (Spall, 1992). For m independent Rademacher directions $\boldsymbol{\Delta}^{(r)} \in \{-1, +1\}^d$ and perturbation radius $c > 0$, define

$$h_j^{(r)} = \frac{E_O(\boldsymbol{\theta} + c\boldsymbol{\Delta}^{(r)}) - E_O(\boldsymbol{\theta} - c\boldsymbol{\Delta}^{(r)})}{2c\,\Delta_j^{(r)}}, \qquad \frac{1}{m}\sum_{r=1}^{m} h_j^{(r)} \approx \partial_j E_O(\boldsymbol{\theta}). \tag{1}$$

Thus one pair of circuit evaluations estimates all d partial derivatives at once, but each directional difference mixes contributions from all coordinates, and finite-shot evaluations add measurement noise.

3 GRADIENT-ENERGY SPARSITY IN VQCS

3.1 GRADIENT-ENERGY SPARSITY

Let $L : \mathbb{R}^d \to \mathbb{R}$ be a differentiable scalar loss and $g(\boldsymbol{\theta}) = \nabla L(\boldsymbol{\theta})$. For $S \subseteq [d] = \{1, \ldots, d\}$, let $\boldsymbol{g}_S \in \mathbb{R}^d$ agree with $\boldsymbol{g}$ on S and vanish elsewhere, so that $\|\boldsymbol{g}_S\|_2^2 = \sum_{j\in S} g_j^2$.

Definition 1 (Gradient-energy sparsity). *For $k \in [d]$ and $\rho \in (0, 1]$, $g(\boldsymbol{\theta})$ is (k, ρ)-sparse if a set $S \subseteq [d]$ with $|S| = k$ satisfies*

$$\|g_S(\boldsymbol{\theta})\|_2^2 \geq \rho \|g(\boldsymbol{\theta})\|_2^2. \tag{2}$$

When $\rho = 1$, $\boldsymbol{g}$ is supported on at most k coordinates; when ρ is close to one and $k \ll d$, its energy is concentrated on k coordinates up to a small tail. Unlike exact sparsity, this criterion is stable under small perturbations of $\boldsymbol{g}$, which suits finite-shot and stochastic optimization.

For a nonzero gradient, let $|g|_{(1)} \geq \cdots \geq |g|_{(d)}$ denote its ordered coordinate magnitudes and define $k_\rho(\boldsymbol{\theta}) := \min\{k : \sum_{j=1}^{k} |g|_{(j)}^2 \geq \rho \|\boldsymbol{g}(\boldsymbol{\theta})\|_2^2\}$.

3.2 Circuit conditions for gradient-energy sparsity

We give circuit conditions under which the observable gradient $\nabla_{\boldsymbol{\theta}} E_O$ satisfies Definition 1. $U(\boldsymbol{\theta}) = U_{>j} U_j(\theta_j) U_{<j}$ with $U_j(\theta_j) = e^{-i\theta_j G_j/2}$ and $\|G_j\| \leq 1$, where $U_{<j}$ and $U_{>j}$ contain the gates before and after U_j. With the observable $O_{>j} = U_{>j}^\dagger O U_{>j}$ and the state $\rho_j = U_j U_{<j} \rho_0 U_{<j}^\dagger U_j^\dagger$, direct differentiation gives

$$\partial_j E_O(\boldsymbol{\theta}) = \frac{i}{2} \operatorname{Tr}(\rho_j [G_j, O_{>j}]), \qquad [A, B] := AB - BA. \tag{3}$$

Throughout, write $O = \sum_\alpha a_\alpha O_\alpha$ with real coefficients a_α and Hermitian terms O_α supported on small qubit sets X_α.

Backward causal cones. The backward causal cone of O_α is obtained by starting from X_α at the output and, moving toward the input, collecting every gate that acts on a qubit already in the set and adding that gate's qubits to the set; gates outside the cone cancel in $U^\dagger O_\alpha U$. Let $\mathcal{C}_O$ contain the parameter indices whose gates lie in the cone of at least one O_α.

Proposition 1 (Causal-cone sparsity). *If $j \notin \mathcal{C}_O$, then $\partial_j E_O(\boldsymbol{\theta}) = 0$ for all $\boldsymbol{\theta}$. Hence, if $|\mathcal{C}_O| \leq k$, $\nabla_{\boldsymbol{\theta}} E_O$ is $(k, 1)$-sparse.*

For shallow, geometrically local circuits with local readouts, the cone can contain far fewer parameterized gates than the full ansatz (Uvarov & Biamonte, 2021; Cerezo et al., 2021b). The proof and a five-qubit example are in Appendix A.1.

Commutation. Inside the cone, a derivative still vanishes when the generator commutes with $O_{>j}$.

Proposition 2 (Commutation-induced sparsity). *Let $\mathcal{A}_O(\boldsymbol{\theta}) = \{j : [G_j, O_{>j}(\boldsymbol{\theta})] \neq 0\}$. If $j \notin \mathcal{A}_O(\boldsymbol{\theta})$, then $\partial_j E_O(\boldsymbol{\theta}) = 0$. Hence, if $|\mathcal{A}_O(\boldsymbol{\theta})| \leq k$, $\nabla_{\boldsymbol{\theta}} E_O(\boldsymbol{\theta})$ is $(k, 1)$-sparse.*

This follows directly from Eq. 3. Related commutator and Lie-algebraic representations appear in analytic quantum-gradient methods (Schuld et al., 2019; Izmaylov et al., 2021; Heidari et al., 2025). The proof and a diagonal Ising example are provided in Appendix A.2.

Decaying response tails. Exact zeros are not necessary for concentration, it suffices that the response $|\partial_j E_{O_\alpha}|$ of a term decays with its distance from the gate. Let $\operatorname{dist}(X_\alpha, \operatorname{supp} G_j)$ be the shortest-path distance between the two qubit sets in the interaction graph, whose edges join qubits coupled by a two-qubit gate.

Assumption 1 (Decaying response envelope). *There exist $L_h > 0$ and a nonincreasing $\chi : [0, \infty) \to [0, 1]$ with $\lim_{r\to\infty} \chi(r) = 0$ such that $|\partial_j E_{O_\alpha}(\boldsymbol{\theta})| \leq L_h \, \chi\big(\operatorname{dist}(X_\alpha, \operatorname{supp} G_j)\big)$ for all $\boldsymbol{\theta}$, inputs, j, and α.*

By Eq. 3, $|\partial_j E_{O_\alpha}| \leq \frac{1}{2} \|[G_j, U_{>j}^\dagger O_\alpha U_{>j}]\|_{\mathrm{op}}$, so Lieb–Robinson-type bounds yield such envelopes, e.g., for power-law interactions (Else et al., 2020). For $\tau > 0$, let $A_j = \sum_\alpha |a_\alpha| \, \chi\big(\operatorname{dist}(X_\alpha, \operatorname{supp} G_j)\big)$ and $\mathcal{T}_\tau(O) = \{j : A_j > \tau\}$ with $k = |\mathcal{T}_\tau(O)| \geq 1$.

Proposition 3 (Compressibility from decaying responses). *For every $j \notin \mathcal{T}_\tau(O)$, $|\partial_j E_O(\boldsymbol{\theta})| \leq L_h \tau$, so $\|(\nabla E_O)_{\mathcal{T}_\tau^c}\|_2^2 \leq (d - k) L_h^2 \tau^2$. Consequently, for any $\rho \in (0, 1)$, $\nabla E_O(\boldsymbol{\theta})$ is (k, ρ)-sparse whenever $\|(\nabla E_O)_{\mathcal{T}_\tau}\|_2^2 \geq \frac{\rho}{1-\rho}(d - k) L_h^2 \tau^2$.*

Unlike the first two mechanisms, this one yields approximate rather than exact concentration; the proof is in Appendix A.3.

Algorithm 1 SGIQ for one mini-batch

Require: Mini-batch $\mathcal{B}$, parameters $\boldsymbol{\theta}$, directions m, selected coordinates k, radius c.

1: Forward pass: evaluate $z_{iq} = E_{O_q}(\boldsymbol{\theta}; \boldsymbol{x}_i)$ and compute λ_{iq} by Eq. 4.
2: For each circuit (i, q) and $r = 1, \ldots, m$, draw $\boldsymbol{\Delta}^{(i,q,r)} \sim \mathrm{Unif}(\{-1, +1\}^d)$ and evaluate $E_{O_q}(\boldsymbol{\theta} \pm c\boldsymbol{\Delta}^{(i,q,r)}; \boldsymbol{x}_i)$.
3: Form $\bar{\boldsymbol{g}}$ from Eqs. 1 and 4.
4: $\mathcal{S} \leftarrow \mathrm{TopK}(|\bar{\boldsymbol{g}}|, k)$.
5: For $j \in \mathcal{S}$, evaluate $E_{O_q}(\boldsymbol{\theta} \pm \frac{\pi}{2}\boldsymbol{e}_j; \boldsymbol{x}_i)$ and form g_j^{PSR}.
6: $\widetilde{g}_j \leftarrow g_j^{\mathrm{PSR}}$ for $j \in \mathcal{S}$ and $\widetilde{g}_j \leftarrow \bar{g}_j$ otherwise.
7: **return** $\widetilde{\boldsymbol{g}}$ to the optimizer.

3.3 From circuit structure to loss gradients

The mechanisms above act on circuit outputs, whereas training uses the loss gradient. For sample $\boldsymbol{x}_i$ with target $\widetilde{y}_i$ and observable O_q, let $z_{iq}(\boldsymbol{\theta}) = E_{O_q}(\boldsymbol{\theta}; \boldsymbol{x}_i)$, $\widehat{y}_i = F_{\boldsymbol{\phi}}(z_{i1}, \ldots, z_{iQ})$ with readout parameters $\boldsymbol{\phi}$, and $\mathcal{L}_{\mathcal{B}}(\boldsymbol{\theta}, \boldsymbol{\phi}) = |\mathcal{B}|^{-1} \sum_{i \in \mathcal{B}} \ell(\widehat{y}_i, \widetilde{y}_i)$. The chain rule gives

$$\partial_j \mathcal{L}_{\mathcal{B}} = \sum_{i \in \mathcal{B}} \sum_{q=1}^{Q} \lambda_{iq}\, \partial_j E_{O_q}(\boldsymbol{\theta}; \boldsymbol{x}_i), \qquad \lambda_{iq} := \frac{1}{|\mathcal{B}|} \frac{\partial \ell(\widehat{y}_i, \widetilde{y}_i)}{\partial \widehat{y}_i} \frac{\partial F_{\boldsymbol{\phi}}}{\partial z_{iq}}. \tag{4}$$

Hence a loss-gradient coordinate vanishes if it vanishes for every observable and sample, and is small when these derivatives are small and the weights λ_{iq} are bounded; since the dominant coordinates may vary across samples, observables, and iterates, Section 5 measures concentration directly on the mini-batch loss gradient.

4 SGIQ

4.1 Gradient estimator

At each step, every circuit (sample $\boldsymbol{x}_i$, measurement group q) is differenced along its own m independent Rademacher directions by Eq. 1, and the chain rule Eq. 4, with λ_{iq} from the forward pass, combines the results into the averaged SPSA loss-gradient estimate $\bar{\boldsymbol{g}}$. Each coordinate receives the score $s_j = |\bar{g}_j|$, and $\mathcal{S} = \mathrm{TopK}(\boldsymbol{s}, k)$ collects the indices of the k largest scores, with ties broken by a fixed rule. For $j \in \mathcal{S}$, PSR differentiates the circuit outputs and the chain rule gives g_j^{PSR}. The estimator is

$$\widetilde{g}_j = \begin{cases} g_j^{\mathrm{PSR}}, & j \in \mathcal{S}, \\ \bar{g}_j, & j \notin \mathcal{S}. \end{cases} \tag{5}$$

4.2 Theoretical guarantees

Setting. In the *ideal* setting circuit outputs are exact expectation values, in the *finite-shot* setting each is estimated from N_{sh} shots. With $\boldsymbol{\theta}$ and $\mathcal{B}$ fixed, $\mathbb{E}[\cdot]$ averages over the SPSA directions and, under finite shots, the measurement outcomes. The target is $\boldsymbol{g} := \nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mathcal{B}}(\boldsymbol{\theta})$, and $\nabla \mathcal{J}$ is the gradient of $\mathcal{J} := \mathbb{E}_{\mathcal{B}}[\mathcal{L}_{\mathcal{B}}]$. We write $Q = 1$ and $\lambda_i := \lambda_{i1}$. By Eq. 4, $g_j = \sum_{i \in \mathcal{B}} \Gamma_{ij}$ with $\Gamma_{ij} := \lambda_i\, \partial_j E_O(\boldsymbol{\theta}; \boldsymbol{x}_i)$. Since each circuit has its own directions, the estimator noise is governed by this *per-sample gradient matrix* $\Gamma \in \mathbb{R}^{B \times d}$, whose energy is $\|\Gamma\|_F^2 = \sum_i \lambda_i^2 \|\nabla_{\boldsymbol{\theta}} E_O(\boldsymbol{\theta}; \boldsymbol{x}_i)\|_2^2$. The Taylor expansion of the SPSA difference quotient (Appendix B.1) shows that coordinate j of $\bar{\boldsymbol{g}}$ has bias $c^2 \bar{\beta}_j + O(c^4)$, with

$$\bar{\beta}_j := \sum_{i \in \mathcal{B}} \lambda_i\, b_{ij}, \qquad b_{ij} := \frac{1}{6}\Big[3 \sum_{p=1}^{d} \partial_j \partial_p^2 E_O - 2\, \partial_j^3 E_O\Big](\boldsymbol{\theta}; \boldsymbol{x}_i). \tag{6}$$

Under finite shots, let $\sigma_{\mathrm{P},j}^2$ and σ_{S}^2 be the shot variances of a PSR coordinate and of an SPSA coordinate from one direction, and $\sigma_{\mathrm{P}}^2 := \max_j \sigma_{\mathrm{P},j}^2$. Since a measured output has variance at most

$\|O\|_{\mathrm{op}}^2/N_{\mathrm{sh}}$,

$$\sigma_{\mathrm{P}}^2 \le s/N_{\mathrm{sh}}, \qquad \sigma_{\mathrm{S}}^2 \le s/(N_{\mathrm{sh}}c^2), \qquad s := \tfrac{1}{2}\|O\|_{\mathrm{op}}^2 \textstyle\sum_{i\in\mathcal{B}} \lambda_i^2. \tag{7}$$

In the ideal setting all shot terms vanish and we set $s := 0$. The *realized concentration* of the selected set is $\widehat{\rho} := \frac{\|\boldsymbol{g}_{\mathcal{S}}\|_2^2}{\|\boldsymbol{g}\|_2^2} \in [0,1]$, so that $\boldsymbol{g}$ is $(k,\widehat{\rho})$-sparse in the sense of Definition 1. For a random vector $\boldsymbol{v}$, we write $\mathrm{Var}(\boldsymbol{v}) := \mathbb{E}\|\boldsymbol{v} - \mathbb{E}\boldsymbol{v}\|_2^2$ for its total variance, the sum of its coordinate variances.

Assumption 2 (Predictable selection)**.** *$\mathcal{S}$ is independent of the directions and shots used to form $\bar{\boldsymbol{g}}$ in the current step.*

Algorithm 1 selects $\mathcal{S}$ from the same $\bar{\boldsymbol{g}}$ and violates Assumption 2; Lemma 3 shows that, under a gap condition, it coincides with a predictable estimator except with probability exponentially small in m.

Lemma 1 (Bias)**.** *Under Assumption 2, in both settings, $\mathbb{E}[\widetilde{g}_j] = g_j$ for $j \in \mathcal{S}$ and $\mathbb{E}[\widetilde{g}_j] - g_j = c^2\bar{\beta}_j + \mathcal{O}(c^4)$ for $j \notin \mathcal{S}$, with $\bar{\beta}_j$ from Eq. 6. Hence*

$$F(k) := \big\|\mathbb{E}[\widetilde{\boldsymbol{g}}] - \boldsymbol{g}\big\|_2^2 = \sum_{j\notin\mathcal{S}} \big(c^2\bar{\beta}_j\big)^2 + \mathcal{O}\big((d-k)c^6\big) = (d-k)\,c^4\,\overline{\beta^2}_{\mathcal{S}^{\mathrm{c}}} + \mathcal{O}\big((d-k)c^6\big), \tag{8}$$

where $\overline{\beta^2}_{\mathcal{S}^{\mathrm{c}}} := \frac{1}{d-k}\sum_{j\notin\mathcal{S}} \bar{\beta}_j^2$ is the mean squared bias coefficient over the unselected coordinates.

The bias does not depend on m and vanishes at $k = d$. $F(0)$ is the bias of averaged SPSA. Because the difference quotient is formed on circuit outputs before chaining, shot noise leaves it unchanged (Appendix B.6).

Lemma 2 (Variance)**.** *Under Assumption 2, conditionally on $\mathcal{S}$,*

$$\mathrm{Var}(\widetilde{\boldsymbol{g}}) = \frac{(d-k)\|\Gamma\|_F^2 - \|\Gamma_{\mathcal{S}^{\mathrm{c}}}\|_F^2}{m} + \frac{(d-k)\sigma_{\mathrm{S}}^2}{m} + \sum_{j\in\mathcal{S}} \sigma_{\mathrm{P},j}^2 + \mathcal{O}\Big(\frac{(d-k)c^2}{m}\Big), \tag{9}$$

where $\Gamma_{\mathcal{S}^{\mathrm{c}}}$ keeps the unselected columns of Γ.

Proposition 4 (Gradient estimation error)**.** *(i) For any selection rule,*

$$\mathbb{E}\|\widetilde{\boldsymbol{g}} - \boldsymbol{g}\|_2^2 = \mathbb{E}\|\bar{\boldsymbol{g}} - \boldsymbol{g}\|_2^2 - \mathbb{E}\Big[\sum_{j\in\mathcal{S}} \big((\bar{g}_j - g_j)^2 - \sigma_{\mathrm{P},j}^2\big)\Big]. \tag{10}$$

In the ideal setting $\sigma_{\mathrm{P},j} = 0$, so $\mathbb{E}\|\widetilde{\boldsymbol{g}} - \boldsymbol{g}\|_2^2 \le \mathbb{E}\|\bar{\boldsymbol{g}} - \boldsymbol{g}\|_2^2$; under finite shots and Assumption 2, the same holds when $\sigma_{\mathrm{P}}^2 \le \sigma_{\mathrm{S}}^2/m$. (ii) Under Assumption 2, $\mathbb{E}\|\widetilde{\boldsymbol{g}} - \boldsymbol{g}\|_2^2 = F(k) + \mathrm{Var}(\widetilde{\boldsymbol{g}})$.

Proposition 5 (Descent-direction reliability)**.** *Under Assumption 2, $\mathbb{E}\langle \boldsymbol{g}, \widetilde{\boldsymbol{g}} - \boldsymbol{g}\rangle = c^2\langle \boldsymbol{g}_{\mathcal{S}^{\mathrm{c}}}, \bar{\boldsymbol{\beta}}_{\mathcal{S}^{\mathrm{c}}}\rangle + \mathcal{O}(c^4)$ and*

$$\mathrm{Var}\left(\langle \boldsymbol{g}, \widetilde{\boldsymbol{g}}\rangle\right) \le \frac{(1-\widehat{\rho})\,\|\boldsymbol{g}\|_2^2}{m}\big(2\|\Gamma\|_F^2 + \sigma_{\mathrm{S}}^2\big) + \widehat{\rho}\,\|\boldsymbol{g}\|_2^2\,\sigma_{\mathrm{P}}^2 + \mathcal{O}\Big(\frac{c^2}{m}\Big). \tag{11}$$

In the ideal setting, averaged SPSA is the case $\widehat{\rho} = 0$: the directional-variance bound of SGIQ is that of SPSA multiplied by $1 - \widehat{\rho}$, so the more concentrated the gradient, the more reliable the descent direction.

Selection by the SPSA scores. Let $\mathcal{S}_\star$ index the k largest $|g_j|$, let $\rho_k := \|\boldsymbol{g}_{\mathcal{S}_\star}\|_2^2/\|\boldsymbol{g}\|_2^2$ be the true top-k coverage (reported in Table 1 for $k = k_\rho$), let $b_\infty := \max_j |\mathbb{E}\bar{g}_j - g_j| = O(c^2)$, and let $\gamma := |g|_{(k)} - |g|_{(k+1)}$.

Lemma 3 (Selection consistency and fidelity)**.** *For the in-sample rule $\mathcal{S} = \mathrm{TopK}(|\bar{\boldsymbol{g}}|, k)$ of Algorithm 1 and $\epsilon := \max_j |\bar{g}_j - g_j|$,*

$$\widehat{\rho} \le \rho_k, \qquad \sqrt{\widehat{\rho}} \ge \sqrt{\rho_k} - \frac{2\sqrt{k}\,\epsilon}{\|\boldsymbol{g}\|_2}. \tag{12}$$

If $\gamma > 2b_\infty$, then $\mathcal{S} = \mathcal{S}_\star$, and hence $\widehat{\rho} = \rho_k$, with probability at least $1 - p_{\mathrm{fail}}$, where

$$p_{\mathrm{fail}} \le 2d\exp\Big(-\frac{m\,t_\gamma^2}{2\sigma_X^2 + \frac{2}{3}Rt_\gamma}\Big), \qquad t_\gamma := \frac{\gamma}{2} - b_\infty, \tag{13}$$

so that $(1 - p_{\mathrm{fail}})\,\rho_k \le \mathbb{E}[\widehat{\rho}] \le \rho_k$. Here $\sigma_X^2 = \|\Gamma\|_F^2 + \sigma_{\mathrm{S}}^2 + O(c^2)$ and R bound the variance and the range of the contribution of a single direction to $\bar{g}_j$.

Outside an event of probability p_{fail}, which is small once the SPSA noise $\sigma_X/\sqrt{m}$ is well below the gap, Algorithm 1 returns the estimator built on the fixed set $\mathcal{S}_\star$; this estimator satisfies Assumption 2, so the results above apply with $\widehat{\rho} = \rho_k$. Without a gap, Eq. 12 still ties $\widehat{\rho}$ to ρ_k (Appendix B.9).

We analyze SGD on $\boldsymbol{\theta}$ with the readout fixed, $\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta\,\widetilde{\boldsymbol{g}}_t$, where $\widetilde{\boldsymbol{g}}_t$ is formed on an independently drawn mini-batch $\mathcal{B}_t$ and $\mathcal{S}_t$ is chosen at step t. The experiments use Adam; the estimator results above do not depend on the optimizer.

Assumption 3 (Objective and sampling)**.** *$\mathcal{J}$ is bounded below by $\mathcal{J}_\star$, and $\nabla\mathcal{J}$ is L_g-Lipschitz. Mini-batches satisfy $\mathbb{E}[\boldsymbol{g}_{\mathcal{B}_t} \mid \boldsymbol{\theta}_t] = \nabla\mathcal{J}(\boldsymbol{\theta}_t)$ and $\mathbb{E}[\|\boldsymbol{g}_{\mathcal{B}_t} - \nabla\mathcal{J}(\boldsymbol{\theta}_t)\|_2^2 \mid \boldsymbol{\theta}_t] \le \sigma_{\mathcal{B}}^2$, and $\bar{G} := \sup_t \|\Gamma_t\|_F^2$, $\bar{s} := \sup_t s_t$, and $B_\beta^2 := \sup_t \overline{\beta^2}_{\mathcal{S}_t^c}$ are finite, where Γ_t, s_t (Eq. 7) and $\overline{\beta^2}_{\mathcal{S}_t^c}$ (Lemma 1) are taken at step t.*

Theorem 1 (Convergence of SGIQ)**.** *Under Assumption 3, let $0 < \eta \le 1/(4L_g)$ and $\Delta_{\mathcal{J}} := \mathcal{J}(\boldsymbol{\theta}_0) - \mathcal{J}_\star$. (i) If every $\mathcal{S}_t$ satisfies Assumption 2 given the history,*

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\|\nabla\mathcal{J}(\boldsymbol{\theta}_t)\|_2^2 \le \frac{2\Delta_{\mathcal{J}}}{\eta T} + L_g\,\eta\,v(k) + 3F(k), \tag{14}$$

where $F(k) \le (d-k)\,c^4 B_\beta^2 + \mathcal{O}\big((d-k)c^6\big)$ bounds the bias of Lemma 1 along the run, $v(k) := V(k) + \big(\sigma_{\mathcal{B}} + \sqrt{F(k)}\big)^2$, and, by Lemma 2 and Eq. 7,

$$V(k) = \frac{d-k}{m}\Big(\bar{G} + \frac{\bar{s}}{N_{\mathrm{sh}}c^2}\Big) + \frac{k\,\bar{s}}{N_{\mathrm{sh}}} + \mathcal{O}\Big(\frac{(d-k)c^2}{m}\Big). \tag{15}$$

(ii) For any selection rule, including Algorithm 1, Eq. 14 holds with $F(k)$ and $v(k)$ replaced by e^2 and $(e+\sigma_{\mathcal{B}})^2$, where $e^2 := \sup_t \mathbb{E}_t\|\widetilde{\boldsymbol{g}}_t - \boldsymbol{g}_{\mathcal{B}_t}\|_2^2$ is bounded by Proposition 4(i).

Unlike the variance term, the bias term $3F(k)$ is not controlled by η. If the gap of Lemma 3 holds at every step, Algorithm 1 coincides with the predictable variant using $\mathcal{S}_{\star,t}$ with probability at least $1 - Tp_{\mathrm{fail}}$. Since $G_\infty := \sup_{\boldsymbol{\theta}}\|\nabla\mathcal{J}(\boldsymbol{\theta})\|_2^2 < \infty$ for bounded readout weights, the bounds of part (i) and Corollary 1 then hold for Algorithm 1 itself up to an additive $Tp_{\mathrm{fail}}G_\infty$, which is negligible once m grows logarithmically in dT (Appendix B.9).

Let $A_0 := 8L_g\Delta_{\mathcal{J}} + \sigma_{\mathcal{B}}^2/2$ denote the part shared by all estimators.

Corollary 1 (Explicit rates)**.** *Let $\eta = 1/(4L_g\sqrt{T})$. Under Theorem 1(i), up to terms of order $(d-k)c^2/(m\sqrt{T}) + (d-k)c^6$,*

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\|\nabla\mathcal{J}(\boldsymbol{\theta}_t)\|_2^2 \le \underbrace{\frac{1}{\sqrt{T}}\Big[A_0 + \frac{d-k}{4m}\Big(\bar{G} + \frac{\bar{s}}{N_{\mathrm{sh}}c^2}\Big) + \frac{k\,\bar{s}}{4N_{\mathrm{sh}}}\Big]}_{\textit{convergence rate}} + \underbrace{\frac{7}{2}(d-k)\,c^4 B_\beta^2}_{\textit{accuracy floor}}. \tag{16}$$

(a) Ideal setting *($\bar{s} = 0$). If $c \le (d-k)^{-1/4}T^{-1/8}$ (no condition when $k = d$), then*

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\|\nabla\mathcal{J}(\boldsymbol{\theta}_t)\|_2^2 \le \frac{K(m,k,d)}{\sqrt{T}} + \mathcal{O}\big(T^{-3/4}\big), \tag{17}$$

where $K(m,k,d) := A_0 + \frac{7}{2}B_\beta^2 + (d-k)\,\bar{G}/(4m)$, with $B_\beta := 0$ at $k = d$. (b) Finite-shot setting. *For a fixed c, the bound converges at rate $\mathcal{O}(T^{-1/2})$ to the accuracy floor. With $c = c_\star := \big(\bar{s}/(28\,mN_{\mathrm{sh}}B_\beta^2\sqrt{T})\big)^{1/6}$,*

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\|\nabla\mathcal{J}(\boldsymbol{\theta}_t)\|_2^2 \le \frac{K_{\mathrm{sh}}}{\sqrt{T}} + \frac{C_{\mathrm{sh}}}{T^{1/3}}, \tag{18}$$

where $K_{\mathrm{sh}} := A_0 + (d-k)\bar{G}/(4m) + k\bar{s}/(4N_{\mathrm{sh}})$ and $C_{\mathrm{sh}} := 1.2\,(d-k)\big(B_\beta\bar{s}/(mN_{\mathrm{sh}})\big)^{2/3}$.

Eq. 16 separates the convergence rate, whose algorithm-dependent part scales as $(d-k)/m$, from the accuracy floor, which depends on neither m nor T and vanishes at $k = d$. In the ideal setting, K interpolates between averaged SPSA ($k = 0$) and PSR ($K = A_0$ at $k = d$).

Corollary 2 (Equal-budget comparison)**.** *Let SPSA use $m' = m + k$ directions, so that both methods evaluate $2BQ(m + k)$ gradient circuits per step, and let Assumption 2 hold. (i)* Bias: $F_{\mathrm{SPSA}} - F_{\mathrm{SGiQ}} = c^4\|\tilde{\boldsymbol{\beta}}_{\mathcal{S}}\|_2^2 + \mathcal{O}(c^6) \geq 0$*, independent of m. (ii)* Descent direction: *in the ideal setting and up to $O(c^2)$ terms, the bound of Proposition 5 is smaller than its SPSA counterpart $2\|\boldsymbol{g}\|_2^2\|\Gamma\|_F^2/m'$ if and only if $\widehat{\rho} > k/(m + k)$.*

4.3 Gradient-circuit cost

Let $B = |\mathcal{B}|$ and let Q be the number of separately evaluated measurement groups per sample. PSR evaluates two shifted circuits per parameter, SPSA two per direction, and SGiQ additionally two per selected coordinate, so the per-step costs are $N_{\mathrm{PSR}} = 2BQd$, $N_{\mathrm{SPSA}} = 2BQm$, and $N_{\mathrm{SGiQ}} = 2BQ(m + k)$, with $m + k \leq d$.

Corollary 3 (Total cost)**.** *Over T steps, method X uses $C_X = TN_X$ gradient circuits and $N_{\mathrm{sh}}C_X$ shots, i.e., $C_{\mathrm{PSR}} = 2BQdT$, $C_{\mathrm{SPSA}} = 2BQm'T$ with m' directions, and $C_{\mathrm{SGiQ}} = 2BQ(m + k)T$. In the ideal setting of Corollary 1(a), the average in Eq. 17 is at most ε after $T(\varepsilon) = K(m, k, d)^2/\varepsilon^2$ steps, which requires*

$$C_{\mathrm{SGiQ}}(\varepsilon) = \frac{2BQ(m + k)\,K(m, k, d)^2}{\varepsilon^2}, \qquad C_{\mathrm{PSR}}(\varepsilon) = \frac{2BQd\,A_0^2}{\varepsilon^2} \tag{19}$$

gradient circuits. Under finite shots with $c = c_\star$ (Corollary 1(b)), $2BQ(m + k)\,T(\varepsilon)$ circuits suffice, with $T(\varepsilon) = \max\{(2K_{\mathrm{sh}}/\varepsilon)^2, (2C_{\mathrm{sh}}/\varepsilon)^3\}$.

To reach accuracy ε, the dimension enters $C_{\mathrm{SGiQ}}(\varepsilon)$ explicitly only through $m+k$ and $(d-k)/m$, and the guarantee of SGiQ needs fewer circuits than that of PSR whenever $(m + k)K(m, k, d)^2 < d\,A_0^2$.

5 Experiments

We first measure loss-gradient energy concentration in two structured VQC regression models and then compare SGiQ with full PSR and SPSA under ideal and finite-shot simulation.

5.1 Experimental setup and circuits

Datasets. We use the Friedman1 and Friedman2 regression benchmarks (Friedman, 1991) with Gaussian target noise of standard deviation 0.1; Friedman1 has five input features and Friedman2 has four. Each feature is min-max scaled to $[0, \pi]$ and the target to $[-1, 1]$, and all errors are reported on the normalized target scale. The Friedman2 experiments use 100 training, 20 validation, and 80 test examples, with normalization computed on the full sample pool. The Friedman1 protocol is given in Section 5.3.

Models & Training Objective. The Friedman2 model (Fig. 1) has eight qubits, three variational layers, and $d = 48$ circuit parameters; it measures eight local observables and uses the affine readout $\widehat{y}_i = b + \sum_{q=0}^{7} w_q\langle Z_q\rangle_{\boldsymbol{\theta},\widetilde{\boldsymbol{x}}_i}$. The Friedman1 model (Fig. 2) uses the layered R_y–R_z and CZ-ring pattern of hardware-efficient circuits (Skolik et al., 2021), with five qubits, eight layers, $d = 80$, and the readout $\widehat{y}_i = a\langle Z^{\otimes 5}\rangle_{\boldsymbol{\theta},\widetilde{\boldsymbol{x}}_i}$ with a trainable scale a. Both models minimize the mini-batch mean squared error (MSE) $\mathcal{L}_{\mathcal{B}} = |\mathcal{B}|^{-1}\sum_{i\in\mathcal{B}}(\widehat{y}_i - \widetilde{y}_i)^2$ without regularization. The quantum-gradient estimators act on $\boldsymbol{\theta}$, and the readout parameters are differentiated automatically; every estimator differentiates circuit outputs and applies the chain rule of Eq. 4.

Both circuits instantiate the commutation mechanism of Section 3: the final-layer R_z parameters ($\theta_{40}, \ldots, \theta_{47}$ in Friedman2 and $\theta_{75}, \ldots, \theta_{79}$ in Friedman1) are followed only by CZ gates and Z-type measurements, which commute with their generators, so these parameters have identically zero loss gradients.

Optimization & Model Selection. Circuit parameters use learning rate 0.03 and readout parameters 0.1. SPSA and SGiQ use $c = 0.1$; SPSA-m denotes SPSA with m directions, and SGiQ (m, k) uses m directions and k selected coordinates. After each epoch we evaluate the validation mean absolute error, $\mathrm{MAE} = n^{-1}\sum_{i=1}^{n}|\widehat{y}_i - \widetilde{y}_i|$ over the n examples of a split, and report the test MAE and MSE of the checkpoint with the lowest validation MAE. The Friedman2 runs use Adam with $(\beta_1, \beta_2) = (0.5, 0.9)$.

Simulation. The ideal experiments use statevector expectation values and have no shot noise. The finite-shot experiment uses a local Aer simulator with 1024 shots per circuit evaluation and no gate, readout, depolarizing, or device-calibrated noise model.

For Friedman1, $Q = 1$.

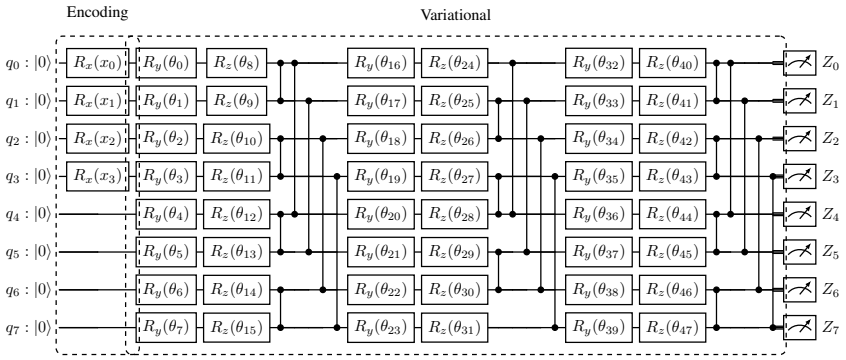

Figure 1: Friedman2 circuit ($d = 48$). q_0–q_3 encode the features with $R_x(x_j)$; each of three R_y–R_z layers is followed by nearest-neighbor and feature-pair CZ gates $(0, 4), \ldots, (3, 7)$. The readout measures $Z_0, \ldots, Z_7$.

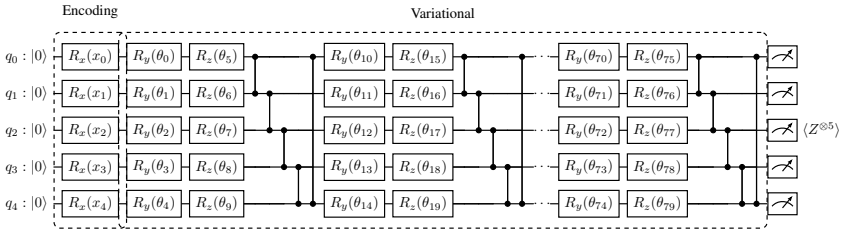

Figure 2: Friedman1 circuit ($d = 80$). Five features are encoded with $R_x(x_j)$; each of eight R_y–R_z layers is followed by the CZ ring $(0, 1), \ldots, (4, 0)$, and layers 3–7 are elided. The readout is $\langle Z^{\otimes 5} \rangle$.

5.2 Gradient concentration in structured VQCs

We measure the gradient of the complete mini-batch MSE with respect to the quantum parameters. At the beginning of each run, one training mini-batch of size 20 is fixed as a probe batch. At epochs 5, 10, 15, and 20, full PSR computes $\boldsymbol{g} = \nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mathcal{B}_{\text{probe}}}(\boldsymbol{\theta}, \boldsymbol{\phi})$.

Table 1 reports, for this $\boldsymbol{g}$, the count k_ρ defined in Section 3.1 and its coverage $\sum_{j=1}^{k_\rho} |g|^2_{(j)} / \|\boldsymbol{g}\|_2^2$; k_ρ is a post-hoc diagnostic and is distinct from the manually chosen algorithmic budget k.

Table 1: PSR loss-gradient energy concentration on a fixed training probe batch. k_ρ is the minimum number of coordinates capturing ρ of the total squared-gradient magnitude.

	Friedman2 ($d = 48$, $\rho = 0.90$)			**Friedman1** ($d = 80$, $\rho = 0.80$)		
Epoch	k_ρ	k_ρ/d	**Coverage**	k_ρ	k_ρ/d	**Coverage**
5	9	0.1875	0.9219	25	0.3125	0.8078
10	12	0.2500	0.9139	29	0.3625	0.8069
15	5	0.1042	0.9067	22	0.2750	0.8030
20	6	0.1250	0.9133	26	0.3250	0.8082

Table 1 shows strong concentration in the structured Friedman2 circuit: only 5–12 of 48 parameters capture at least 90% of the loss-gradient energy. Friedman1 is less concentrated but still nonuniform,

Table 2: Normalized test errors and gradient-circuit evaluations per training run.

Setting	Method	Test MAE	Test MSE	Cost
Ideal	PSR	0.08489 ± 0.01192	0.01304 ± 0.00261	960,000
	SPSA-30	0.10924 ± 0.01759	0.02122 ± 0.00370	360,000
	SPSA-60	0.10454 ± 0.01799	0.01888 ± 0.00601	720,000
	SGiQ $(40, 10)$	0.09504 ± 0.00342	0.01647 ± 0.00197	600,000
Finite-shot (1024 shots)	PSR	0.10865 ± 0.02196	0.02094 ± 0.01101	960,000
	SPSA-20	0.11579 ± 0.01090	0.02265 ± 0.00449	240,000
	SPSA-60	0.12070 ± 0.00264	0.02559 ± 0.00317	720,000
	SGiQ $(40, 20)$	0.10714 ± 0.01119	0.02075 ± 0.00518	720,000

with 22–29 of 80 parameters capturing at least 80%. Since the two blocks use different target ratios, they represent two operating regimes rather than directly ranking the circuit architectures.

On Friedman2, full PSR reaches a test MAE of 0.3684 and SGiQ with $(m, k) = (10, 7)$ reaches 0.3686 while using $(m + k)/d = 17/48 \approx 35\%$ of the PSR gradient circuits per step.

5.3 Friedman1 comparison

We use a fixed Friedman1 split of 200 training, 40 validation, and 80 test examples, with normalization fitted on the training data, and the circuit of Fig. 2. Both environments train for 30 epochs ($B = 20$, $T = 300$ updates) with Adam, $\beta = (0.9, 0.999)$, and the learning rates above. In the ideal environment, $c = 0.1$ satisfies the condition $c \leq (d - k)^{-1/4} T^{-1/8}$ of Corollary 1(a) for every k, since $80^{-1/4} 300^{-1/8} \approx 0.16$. For each of three seeds, all methods share the data split, initialization, and mini-batch order. Configurations are selected by mean best validation MAE, and test errors are reported as mean $\pm$ sample standard deviation over the three runs. Costs are the gradient circuits of Section 4.3, $2BTd$, $2BTm$, and $2BT(m + k)$, excluding forward, evaluation, and search circuits.

We screen SPSA with $m \in \{10, 20, \ldots, 80\}$ and the SGiQ configurations listed in Appendix C. Validation selects SPSA-30 and SGiQ $(40, 10)$ in the ideal environment, and SPSA-60 and SGiQ $(40, 20)$ in the finite-shot environment. Table 2 also includes PSR, ideal SPSA-60 as a higher-cost reference, and finite-shot SPSA-20 as a low-cost reference.

In the ideal environment, SGiQ $(40, 10)$ reduces mean test MAE by 9.1% relative to SPSA-60 using 16.7% fewer gradient circuits. It improves on SPSA-30 at higher cost and uses 37.5% fewer circuits than PSR at higher error. In the finite-shot environment, $(40, 20)$ reduces mean test MAE and MSE by 11.2% and 18.9% relative to equal-budget SPSA-60, respectively, and achieves comparable mean errors to PSR with 25% fewer gradient circuits. These results show that the selected configurations provide empirical cost–accuracy trade-offs in both environments (Appendix C).

6 Conclusion

Efficient quantum gradient estimation is essential for backpropagation in quantum neural networks. Motivated by concentrated loss-gradient energy, we introduced SGiQ to allocate measurement resources according to coordinate importance. Averaged SPSA provides a low-cost global probe, and PSR replaces selected coordinate estimates while SPSA estimates are retained elsewhere. Experiments show lower gradient-circuit cost than PSR and better predictive accuracy than SPSA. These results support sparsity-informed gradient estimation as a practical approach to reducing the cost of quantum backpropagation.

AI use statement

We used LLMs to assist with polishing and improving the clarity of the manuscript's writing. We also used LLMs for literature retrieval and discovery, such as identifying potentially relevant prior work. All AI-generated content and outputs were carefully reviewed and independently verified by the authors to ensure their accuracy and appropriateness. The authors take full responsibility for the final content of the paper.

REFERENCES

Amira Abbas, Robbie King, Hsin-Yuan Huang, William J. Huggins, Ramis Movassagh, Dar Gilboa, and Jarrod Ryan McClean. On quantum backpropagation, information reuse, and cheating measurement collapse. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=HF6bnhfSqH.

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, 2021a.

Marco Cerezo, Akira Sone, Tyler Volkoff, Lukasz Cincio, and Patrick J Coles. Cost function dependent barren plateaus in shallow parametrized quantum circuits. *Nature Communications*, 12 (1):1791, 2021b.

Brian Coyle, Snehal Raj, Virag Umathe, El Amine Cherrat, and Elham Kashefi. Adaptive directional gradients for parameterised quantum circuits. *arXiv preprint arXiv:2606.09734*, 2026.

Zhiyan Ding, Taehee Ko, Jiahao Yao, Lin Lin, and Xiantao Li. Random coordinate descent: a simple alternative for optimizing parameterized quantum circuits. *Physical Review Research*, 6(3):033029, 2024.

Dominic V Else, Francisco Machado, Chetan Nayak, and Norman Y Yao. Improved lieb-robinson bound for many-body hamiltonians with power-law interactions. *Physical Review A*, 101(2): 022333, 2020.

Jerome H Friedman. Multivariate adaptive regression splines. *The Annals of Statistics*, 19(1):1–67, 1991.

Mohsen Heidari, Masih Mozakka, and Wojciech Szpankowski. Hadamard test is sufficient for efficient quantum gradient estimation with lie algebraic symmetries. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=PdUdzvYzoH.

Artur F Izmaylov, Robert A Lang, and Tzu-Ching Yen. Analytic gradients in variational quantum algorithms: Algebraic extensions of the parameter-shift rule to general unitary transformations. *Physical Review A*, 104(6):062443, 2021.

Kosuke Mitarai, Makoto Negoro, Masahiro Kitagawa, and Keisuke Fujii. Quantum circuit learning. *Physical Review A*, 98(3):032309, 2018.

Maniraman Periyasamy, Axel Plinge, Christopher Mutschler, Daniel D Scherer, and Wolfgang Mauerer. Guided-spsa: Simultaneous perturbation stochastic approximation assisted by the parameter shift rule. In *2024 IEEE International Conference on Quantum Computing and Engineering (QCE)*, volume 1, pp. 1504–1515. IEEE, 2024.

Maria Schuld, Ville Bergholm, Christian Gogolin, Josh Izaac, and Nathan Killoran. Evaluating analytic gradients on quantum hardware. *Physical Review A*, 99(3):032331, 2019.

Andrea Skolik, Jarrod R McClean, Masoud Mohseni, Patrick Van Der Smagt, and Martin Leib. Layerwise learning for quantum neural networks. *Quantum Machine Intelligence*, 3(1):5, 2021.

James C Spall. Multivariate stochastic approximation using a simultaneous perturbation gradient approximation. *IEEE Transactions on Automatic Control*, 37(3):332–341, 1992.

AV Uvarov and Jacob D Biamonte. On barren plateaus and cost function locality in variational quantum algorithms. *Journal of Physics A: Mathematical and Theoretical*, 54(24):245301, 2021.

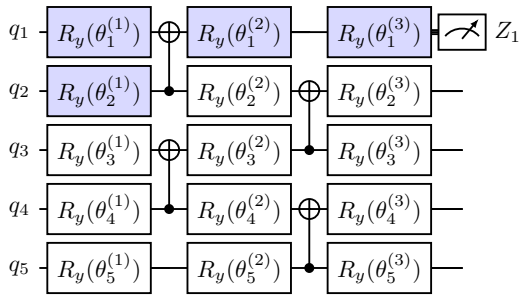

Figure 3: Five-qubit causal-cone example. The shaded rotations are the four parameters in the backward causal cone of Z_1 and are the only rotations that can affect $E_{Z_1}(\boldsymbol{\theta})$.

A Structural proofs and examples

A.1 Causal-cone proof and five-qubit example

Proof of Proposition 1. For $j \notin \mathcal{C}_O$, backward propagation of each local term O_α through the gates following U_j produces an effective observable $O_{\alpha,>j}$ whose support does not intersect that of G_j. Therefore,

$$[G_j, O_{\alpha,>j}] = 0 \qquad \text{for every } \alpha.$$

Using linearity and Eq. 3,

$$\partial_j E_O = \sum_\alpha a_\alpha \frac{i}{2} \operatorname{Tr}(\rho_j [G_j, O_{\alpha,>j}]) = 0.$$

Thus, $\nabla_{\boldsymbol{\theta}} E_O$ is supported on $\mathcal{C}_O$. If $|\mathcal{C}_O| \le k$, the support can be extended to a set of size k, and Definition 1 holds with $\rho = 1$. □

Example 1 (Five-qubit local readout)**.** *Consider the five-qubit circuit*

$$U(\boldsymbol{\theta}) = U_R^{(3)} U_E^{(2)} U_R^{(2)} U_E^{(1)} U_R^{(1)}, \qquad U_R^{(\ell)} = \prod_{i=1}^{5} R_y(\theta_i^{(\ell)}),$$

where

$$U_E^{(1)} = \mathrm{CNOT}_{2\to1}\,\mathrm{CNOT}_{4\to3}, \qquad U_E^{(2)} = \mathrm{CNOT}_{3\to2}\,\mathrm{CNOT}_{5\to4},$$

and the measured observable is $O = Z_1$.

Propagating Z_1 backward first encounters $R_y(\theta_1^{(3)})$. Neither CNOT in $U_E^{(2)}$ expands its support, so the middle rotation layer contributes only $R_y(\theta_1^{(2)})$. Under $\mathrm{CNOT}_{2\to1}$,

$$Z_1 \longmapsto Z_2 Z_1,$$

and hence the first rotation layer contributes $R_y(\theta_1^{(1)})$ and $R_y(\theta_2^{(1)})$.

The backward-cone parameter set is therefore

$$\mathcal{C}_O = \{(1,1),(1,2),(2,1),(3,1)\},$$

where (ℓ, i) indexes $\theta_i^{(\ell)}$. Only four of the fifteen rotation parameters can have nonzero derivatives. Consequently,

$$\nabla_{\boldsymbol{\theta}} E_{Z_1}$$

is supported on at most four coordinates and is therefore $(4,1)$-sparse.

A.2 Commutation proof and diagonal Ising example

Proof of Proposition 2. If $j \notin \mathcal{A}_O(\boldsymbol{\theta})$, then $[G_j, O_{>j}] = 0$ at this $\boldsymbol{\theta}$, and Eq. 3 gives

$$\partial_j E_O(\boldsymbol{\theta}) = \frac{i}{2} \operatorname{Tr}(\rho_j [G_j, O_{>j}]) = 0.$$

Hence $\nabla_{\boldsymbol{\theta}} E_O(\boldsymbol{\theta})$ is supported on $\mathcal{A}_O(\boldsymbol{\theta})$. If $|\mathcal{A}_O(\boldsymbol{\theta})| \le k$, the support can be extended to a set of size k, and Definition 1 holds with $\rho = 1$. □

Example 2 (Diagonal Ising block)**.** *Let $\mathcal{G} = (V, E)$ be a weighted graph and define*

$$O_C = \sum_{(u,v)\in E} w_{uv} Z_u Z_v, \qquad U_D(\boldsymbol{\gamma}) = \prod_{(u,v)\in E} e^{-i\gamma_{uv} Z_u Z_v/2}.$$

Here, V and E are the vertex and edge sets, w_{uv} is the weight of edge (u, v), and γ_{uv} parameterizes the corresponding diagonal gate. Consider

$$U(\boldsymbol{\omega}, \boldsymbol{\gamma}) = U_D(\boldsymbol{\gamma}) U_M(\boldsymbol{\omega}),$$

where $U_M(\boldsymbol{\omega})$ is a mixer block with d_M trainable parameters.

Every generator in U_D commutes with every term in O_C. Therefore,

$$U_D^\dagger(\boldsymbol{\gamma}) O_C U_D(\boldsymbol{\gamma}) = O_C$$

for all $\boldsymbol{\gamma}$, and

$$\frac{\partial E_{O_C}}{\partial \gamma_{uv}} = 0 \qquad \textit{for every } (u, v) \in E.$$

The mixer coordinates may remain active; for example,

$$[X_u, Z_u Z_v] \neq 0.$$

The circuit contains

$$d = d_M + |E|$$

parameters but has at most d_M potentially nonzero gradient coordinates. Thus,

$$\nabla E_{O_C}$$

is $(d_M, 1)$-sparse, and these zeros are structural. For a dense n-vertex graph with $|E| = \Theta(n^2)$ and $d_M = O(n)$, the fraction of potentially active coordinates satisfies

$$\frac{d_M}{d_M + |E|} = O\left(\frac{1}{n}\right).$$

This example demonstrates algebraic sparsity: a parameter can have zero derivative even when its gate is geometrically connected to the measured qubits.

A.3 Proof of the decaying-response bound

Proof of Proposition 3. By linearity, the triangle inequality, and Assumption 1,

$$|\partial_j E_O| \le \sum_\alpha |a_\alpha| \, |\partial_j E_{O_\alpha}| \le L_h \sum_\alpha |a_\alpha| \, \chi\big(\operatorname{dist}(X_\alpha, \operatorname{supp} G_j)\big) = L_h A_j.$$

For $j \notin \mathcal{T}_\tau$, $A_j \le \tau$, so $|\partial_j E_O| \le L_h \tau$. Summing the squares over the $d - k$ coordinates outside $\mathcal{T}_\tau$ gives $t := \|(\nabla E_O)_{\mathcal{T}_\tau^c}\|_2^2 \le (d - k) L_h^2 \tau^2$. Let $a := \|(\nabla E_O)_{\mathcal{T}_\tau}\|_2^2$. If $a \ge \frac{\rho}{1-\rho}(d - k) L_h^2 \tau^2$, then $a \ge \frac{\rho}{1-\rho} t$, that is, $a \ge \rho(a + t) = \rho \|\nabla E_O\|_2^2$. This is Eq. 2 with $S = \mathcal{T}_\tau$ and $|S| = k$. □

B Proofs of the bias, variance, error, descent, convergence, and cost results

B.1 Estimator, notation, and the common decomposition

Fix $(\boldsymbol{\theta}, \mathcal{B})$ and suppress the conditioning. We write $Q = 1$; for $Q > 1$ every pair (sample, measurement group) is a separate circuit with its own directions and becomes a separate row of Γ, and all statements hold verbatim.

Estimator. For sample i and direction r, let $\boldsymbol{\Delta}^{(i,r)} \in \{-1,+1\}^d$ be Rademacher vectors, independent across i and r. Let $\widehat{E}_O$ denote a measured circuit output, with $\widehat{E}_O = E_O$ in the ideal setting. SGIQ forms the difference quotient on the circuit output of each sample and chains afterwards,

$$\bar{g}_j = \frac{1}{m}\sum_{r=1}^{m}\sum_{i\in\mathcal{B}} \lambda_i \frac{\widehat{E}_O(\boldsymbol{\theta}+c\boldsymbol{\Delta}^{(i,r)};\boldsymbol{x}_i) - \widehat{E}_O(\boldsymbol{\theta}-c\boldsymbol{\Delta}^{(i,r)};\boldsymbol{x}_i)}{2c\,\Delta_j^{(i,r)}}, \tag{20}$$
$$g_j^{\mathrm{PSR}} = \sum_{i\in\mathcal{B}} \lambda_i \frac{\widehat{E}_O(\boldsymbol{\theta}+\frac{\pi}{2}\boldsymbol{e}_j;\boldsymbol{x}_i) - \widehat{E}_O(\boldsymbol{\theta}-\frac{\pi}{2}\boldsymbol{e}_j;\boldsymbol{x}_i)}{2},$$

and $\widetilde{g}_j = g_j^{\mathrm{PSR}}$ for $j \in \mathcal{S}$, $\widetilde{g}_j = \bar{g}_j$ otherwise.

Shot noise. Write $\varepsilon^{(i,r)}$ for the shot error of the SPSA difference of sample i and direction r, and $\zeta_j := g_j^{\mathrm{PSR}} - g_j$. Both have zero mean, are independent of the directions, and are independent of each other and across (i,r) and j. The PSR circuits are run after $\mathcal{S}$ is chosen, with fresh shots. We set

$$\sigma_{\mathrm{S}}^2 := \frac{1}{4c^2}\sum_{i\in\mathcal{B}} \lambda_i^2 \operatorname{Var}\left(\varepsilon^{(i,r)}\right), \qquad \sigma_{\mathrm{P},j}^2 := \operatorname{Var}(\zeta_j).$$

A measured output is the mean of N_{sh} independent single-shot outcomes with values in $[-\|O\|_{\mathrm{op}}, \|O\|_{\mathrm{op}}]$, so its variance is at most $\|O\|_{\mathrm{op}}^2/N_{\mathrm{sh}}$. A difference of two independently measured outputs therefore has shot variance at most $2\|O\|_{\mathrm{op}}^2/N_{\mathrm{sh}}$, and ζ_j is $\sum_i \lambda_i/2$ times such differences. With $s = \frac{1}{2}\|O\|_{\mathrm{op}}^2 \sum_i \lambda_i^2$,

$$\sigma_{\mathrm{S}}^2 \le \frac{1}{4c^2}\sum_i \lambda_i^2 \frac{2\|O\|_{\mathrm{op}}^2}{N_{\mathrm{sh}}} = \frac{s}{N_{\mathrm{sh}}c^2}, \qquad \sigma_{\mathrm{P},j}^2 \le \frac{1}{4}\sum_i \lambda_i^2 \frac{2\|O\|_{\mathrm{op}}^2}{N_{\mathrm{sh}}} = \frac{s}{N_{\mathrm{sh}}},$$

which is Eq. 7. For Pauli observables $\|O\|_{\mathrm{op}} = 1$ and the single-shot variance is $1 - E_O^2$. In the finite-shot setting the factors λ_i are computed from the forward pass, which is independent of the gradient circuits and shared by all three estimators. For the MSE loss with a linear readout, $\partial\ell/\partial\widehat{y}_i$ is affine in the measured output, so this estimate $\widehat{\lambda}_i$ is unbiased. Hence $\sum_i \widehat{\lambda}_i \nabla E_O(\boldsymbol{\theta};\boldsymbol{x}_i)$ has mean $\boldsymbol{g}_{\mathcal{B}}$. We condition on $\widehat{\lambda}$ throughout Sections B.1–B.6 and write λ for it. Its fluctuation is counted in $\sigma_{\mathcal{B}}^2$ in Section B.7.

Regularity. Each θ_j enters a single Pauli rotation, so E_O is a trigonometric polynomial of degree one in every θ_j. Repeated use of the shift rule then gives $|\partial^\alpha E_O| \le \|O\|_{\mathrm{op}}$ for every multi-index α. All $O(c^p)$ terms below are therefore uniform, with constants depending only on d, $\|O\|_{\mathrm{op}}$ and $\max_i|\lambda_i|$.

Common decomposition. Set $\boldsymbol{x} = c\boldsymbol{\Delta}^{(i,r)}$. In the Taylor expansion of $E_O(\boldsymbol{\theta}+\boldsymbol{x}) - E_O(\boldsymbol{\theta}-\boldsymbol{x})$ the even orders cancel:

$$E_O(\boldsymbol{\theta}+\boldsymbol{x}) - E_O(\boldsymbol{\theta}-\boldsymbol{x}) = 2\boldsymbol{x}^{\mathsf{T}}\nabla E_O + \tfrac{1}{3}\sum_{a,b,e} \partial_{abe}E_O\, x_a x_b x_e + O(c^5).$$

Dividing by $2c\Delta_j^{(i,r)}$ and using $(\Delta_j)^{-1} = \Delta_j$,

$$\widehat{u}_j^{(i,r)} = \partial_j E_O + \underbrace{\sum_{p\neq j} \partial_p E_O\, \Delta_p^{(i,r)}\Delta_j^{(i,r)}}_{=:\xi_{ij}^{(i,r)}} + \underbrace{\frac{c^2}{6}\sum_{a,b,e} \partial_{abe}E_O\, \Delta_j^{(i,r)}\Delta_a^{(i,r)}\Delta_b^{(i,r)}\Delta_e^{(i,r)} + O(c^4)}_{=:r_{ij}^{(i,r)}}. \tag{21}$$

No bound has been used: ξ is the exact coordinate-mixing term and r is the Taylor remainder. Multiplying by λ_i and adding the shot error,

$$\bar{g}_j - g_j = \frac{1}{m}\sum_{r=1}^{m}\sum_{i\in\mathcal{B}} \lambda_i\Big(\xi_{ij}^{(i,r)} + r_{ij}^{(i,r)} + \frac{\varepsilon^{(i,r)}\Delta_j^{(i,r)}}{2c}\Big). \tag{22}$$

Fourth moments and the bias coefficient. For Rademacher entries, $\mathbb{E}[\Delta_j\Delta_a\Delta_b\Delta_e] = 1$ if the four indices pair up and 0 otherwise. The three pairings $(a{=}j, b{=}e)$, $(b{=}j, a{=}e)$ and $(e{=}j, a{=}b)$ each contribute $\sum_p \partial_j\partial_p^2 E_O$, and the case $a{=}b{=}e{=}j$ is counted three times instead of once. Hence

$$\mathbb{E}\big[r_{ij}^{(i,r)}\big] = c^2 b_{ij} + O(c^4), \qquad b_{ij} := \tfrac{1}{6}\Big[3\sum_p \partial_j\partial_p^2 E_O - 2\partial_j^3 E_O\Big](\boldsymbol{\theta}; \boldsymbol{x}_i), \qquad \beta_{ij} := \lambda_i b_{ij}, \tag{23}$$

so that $\bar{\beta}_j = \sum_i \beta_{ij}$ is the coefficient of Eq. 6.

B.2 Proof of Lemma 1

For $j \in \mathcal{S}$, $\mathbb{E}[g_j^{\mathrm{PSR}}] = g_j$ because the shift rule is exact and ζ_j has zero mean. For $j \notin \mathcal{S}$, take expectations in Eq. 22. For $p \neq j$ the entries Δ_p and Δ_j are independent with zero mean, so $\mathbb{E}[\xi] = 0$. The shot term has zero mean, and Eq. 23 gives the remainder. Under Assumption 2, conditioning on $\mathcal{S}$ does not change these expectations, so $\mathbb{E}[\widetilde{g}_j] - g_j = c^2\sum_i \beta_{ij} + O(c^4) = c^2\bar{\beta}_j + O(c^4)$.

Squaring and summing over $j \notin \mathcal{S}$ gives $F(k) = \sum_{j\notin\mathcal{S}} \big(c^2\bar{\beta}_j + O(c^4)\big)^2 = c^4\sum_{j\notin\mathcal{S}} \bar{\beta}_j^2 + O\big((d-k)c^6\big)$, which is Eq. 8. Exactly, $F(k)$ is the sum over $\mathcal{S}^{\mathrm{c}}$ of the squared coordinate biases of $\bar{\boldsymbol{g}}$. These biases depend on neither m nor $\mathcal{S}$, so enlarging a nested $\mathcal{S}$ can only remove nonnegative terms, and $F(d) = 0$. □

B.3 Proof of Lemma 2

For $j \in \mathcal{S}$, $\mathrm{Var}(\widetilde{g}_j) = \sigma_{\mathrm{P},j}^2$. For $j \notin \mathcal{S}$, the summands of Eq. 22 are independent across (i, r). The shot term has zero mean and is independent of the directions, so it is uncorrelated with $\xi + r$. Hence

$$\mathrm{Var}(\bar{g}_j) = \frac{1}{m}\sum_{i\in\mathcal{B}} \lambda_i^2\Big[\mathrm{Var}(\xi_{ij}) + 2\,\mathrm{Cov}(\xi_{ij}, r_{ij}) + \mathrm{Var}(r_{ij})\Big] + \frac{\sigma_{\mathrm{S}}^2}{m}.$$

Mixing. Since $\Delta_j^2 = 1$, $\mathbb{E}[\Delta_p\Delta_j\Delta_{p'}\Delta_j] = \delta_{pp'}$, and $\mathrm{Var}(\xi_{ij}) = \sum_{p\neq j}(\partial_p E_O)^2$. *Cross term.* Again using $\Delta_j^2 = 1$ and the fourth-moment count,

$$\mathrm{Cov}(\xi_{ij}, r_{ij}) = \frac{c^2}{6}\sum_{p\neq j}\partial_p E_O \sum_{a,b,e}\partial_{abe}E_O\,\mathbb{E}[\Delta_p\Delta_a\Delta_b\Delta_e] + O(c^4) = c^2\sum_{p\neq j}\partial_p E_O\, b_{ip} + O(c^4).$$

Remainder. $\mathrm{Var}(r_{ij}) = O(c^4)$.

Multiplying by λ_i^2 turns $\partial_p E_O$ and b_{ip} into Γ_{ip} and β_{ip}, and summing over i gives

$$\mathrm{Var}(\bar{g}_j) = \frac{1}{m}\Big[\|\Gamma\|_F^2 - \|\Gamma_{:,j}\|_2^2\Big] + \frac{2c^2}{m}\Big[\langle\Gamma,\beta\rangle_F - \sum_i \Gamma_{ij}\beta_{ij}\Big] + \frac{\sigma_{\mathrm{S}}^2}{m} + O\Big(\frac{c^4}{m}\Big).$$

Under Assumption 2, $\mathrm{Var}(\widetilde{\boldsymbol{g}})$ is the sum of these coordinate variances. Summing over $j \notin \mathcal{S}$ with $\sum_{j\notin\mathcal{S}}\|\Gamma_{:,j}\|_2^2 = \|\Gamma_{\mathcal{S}^{\mathrm{c}}}\|_F^2$ and $\sum_{j\notin\mathcal{S}}\sum_i \Gamma_{ij}\beta_{ij} = \langle\Gamma_{\mathcal{S}^{\mathrm{c}}}, \beta_{\mathcal{S}^{\mathrm{c}}}\rangle_F$, and adding $\sum_{j\in\mathcal{S}}\sigma_{\mathrm{P},j}^2$, gives Eq. 9, whose $O((d-k)c^2/m)$ term is $\mathcal{C}_c := \frac{2c^2}{m}\big[(d-k)\langle\Gamma,\beta\rangle_F - \langle\Gamma_{\mathcal{S}^{\mathrm{c}}}, \beta_{\mathcal{S}^{\mathrm{c}}}\rangle_F\big] + O\big(\frac{(d-k)c^4}{m}\big)$. Finally, $0 \le \|\Gamma_{\mathcal{S}^{\mathrm{c}}}\|_F^2 \le \|\Gamma\|_F^2$ places the mixing term between $(d-k-1)\|\Gamma\|_F^2/m$ and $(d-k)\|\Gamma\|_F^2/m$ in the ideal setting. □

B.4 Proof of Proposition 4

(i) On $\mathcal{S}$, $\widetilde{g}_j - g_j = \zeta_j$; off $\mathcal{S}$, $\widetilde{g}_j - g_j = \bar{g}_j - g_j$. Hence, pathwise and for any rule,

$$\|\widetilde{\boldsymbol{g}} - \boldsymbol{g}\|_2^2 = \|\bar{\boldsymbol{g}} - \boldsymbol{g}\|_2^2 - \sum_{j\in\mathcal{S}}(\bar{g}_j - g_j)^2 + \sum_{j\in\mathcal{S}}\zeta_j^2.$$

The PSR shots are fresh, so $\mathbb{E}[\zeta_j^2 \mid \mathcal{S}, \bar{\boldsymbol{g}}] = \sigma_{\mathrm{P},j}^2$. Taking expectations gives Eq. 10. In the ideal setting $\zeta_j = 0$ and the subtracted term is nonnegative. Under Assumption 2, the proof of Lemma 2 gives $\mathbb{E}(\bar{g}_j - g_j)^2 \ge \mathrm{Var}(\bar{g}_j) \ge \sigma_{\mathrm{S}}^2/m$, so each term of the subtracted sum is nonnegative when $\sigma_{\mathrm{P}}^2 \le \sigma_{\mathrm{S}}^2/m$.

(ii) This is the bias–variance identity $\mathbb{E}\|\widetilde{\boldsymbol{g}}-\boldsymbol{g}\|_2^2 = \|\mathbb{E}\widetilde{\boldsymbol{g}}-\boldsymbol{g}\|_2^2 + \mathrm{Var}(\widetilde{\boldsymbol{g}})$ combined with Lemmas 1–2. At $k=0$, $\|\Gamma_{\mathcal{S}^c}\|_F^2 = \|\Gamma\|_F^2$, and the mixing term is $(d-1)\|\Gamma\|_F^2/m$. Hence, in the ideal setting, the mixing term of SGIQ lies between $(d-k-1)\|\Gamma\|_F^2/m$ and $(d-k)\|\Gamma\|_F^2/m$ (proof of Lemma 2), versus $(d-1)\|\Gamma\|_F^2/m$ for averaged SPSA. □

B.5 Proof of Proposition 5

Write $D := \langle \boldsymbol{g}, \widetilde{\boldsymbol{g}} - \boldsymbol{g}\rangle = \sum_{j\notin\mathcal{S}} g_j(\bar{g}_j - g_j) + \sum_{j\in\mathcal{S}} g_j\zeta_j$. Since $\langle \boldsymbol{g}, \boldsymbol{g}\rangle$ is deterministic, $\mathrm{Var}(\langle \boldsymbol{g}, \widetilde{\boldsymbol{g}}\rangle) = \mathrm{Var}(D)$. The mean follows from Lemma 1. By Eq. 22, $D = M + R + N_{\mathrm{S}} + N_{\mathrm{P}}$ with mixing part M, remainder part R, SPSA shot part N_{S} and PSR shot part N_{P}. The shot parts have zero mean, are independent of each other, and are independent of the directions, so they are uncorrelated with $M+R$. Moreover $\mathrm{Var}(R) = O(c^4/m)$ and $\mathrm{Cov}(M,R) = O(c^2/m)$.

Mixing part. $M = \frac{1}{m}\sum_r \sum_i X_i^{(r)}$ with $X_i = \sum_{j\notin\mathcal{S}}\sum_{p\neq j} g_j\Gamma_{ip}\Delta_j\Delta_p$. The summands are independent across (i,r), so $\mathrm{Var}(M) = m^{-1}\sum_i V_i$ with $V_i := \mathrm{Var}(X_i)$. Group the terms of X_i by unordered pairs $\{j,p\}$. A pair with both indices in $\mathcal{S}^{\mathrm{c}}$ appears twice, with coefficient $g_j\Gamma_{ip} + g_p\Gamma_{ij}$. A pair with $j\notin\mathcal{S}$ and $p\in\mathcal{S}$ appears once, with coefficient $g_j\Gamma_{ip}$. A pair inside $\mathcal{S}$ does not appear. The products $\Delta_j\Delta_p$ over distinct pairs have zero mean, unit variance, and are pairwise uncorrelated, so

$$V_i = \sum_{\{j,p\}\subseteq\mathcal{S}^{\mathrm{c}}} \big(g_j\Gamma_{ip} + g_p\Gamma_{ij}\big)^2 + \sum_{j\notin\mathcal{S},\, p\in\mathcal{S}} g_j^2\Gamma_{ip}^2.$$

By $(a+b)^2 \le 2a^2+2b^2$, the first sum is at most $2\|\boldsymbol{g}_{\mathcal{S}^{\mathrm{c}}}\|_2^2\|\Gamma_{i,\mathcal{S}^{\mathrm{c}}}\|_2^2$. The second equals $\|\boldsymbol{g}_{\mathcal{S}^{\mathrm{c}}}\|_2^2\|\Gamma_{i,\mathcal{S}}\|_2^2$. Hence $V_i \le 2\|\boldsymbol{g}_{\mathcal{S}^{\mathrm{c}}}\|_2^2\|\Gamma_i\|_2^2$, and, with $\|\boldsymbol{g}_{\mathcal{S}^{\mathrm{c}}}\|_2^2 = (1-\widehat{\rho})\|\boldsymbol{g}\|_2^2$, $\mathrm{Var}(M) \le 2(1-\widehat{\rho})\|\boldsymbol{g}\|_2^2\|\Gamma\|_F^2/m$.

SPSA shot part. $N_{\mathrm{S}} = \frac{1}{m}\sum_{r,i}\lambda_i\frac{\varepsilon^{(i,r)}}{2c}\sum_{j\notin\mathcal{S}} g_j\Delta_j^{(i,r)}$. Since $\mathbb{E}\big(\sum_{j\notin\mathcal{S}} g_j\Delta_j\big)^2 = \|\boldsymbol{g}_{\mathcal{S}^{\mathrm{c}}}\|_2^2$, $\mathrm{Var}(N_{\mathrm{S}}) = (1-\widehat{\rho})\|\boldsymbol{g}\|_2^2\sigma_{\mathrm{S}}^2/m$.

PSR shot part. $\mathrm{Var}(N_{\mathrm{P}}) = \sum_{j\in\mathcal{S}} g_j^2\sigma_{\mathrm{P},j}^2 \le \widehat{\rho}\,\|\boldsymbol{g}\|_2^2\sigma_{\mathrm{P}}^2$.

Adding the four parts proves Eq. 11. □

B.6 Finite shots: the perturbation radius and the difference quotient

Choice of c. By Proposition 4(ii), the c-dependent part of the error is $f(c) = \kappa_{\mathrm{cr}}c^2 + \kappa_{\mathrm{bs}}c^4 + \kappa_{\mathrm{sh}}c^{-2}$, with

$$\kappa_{\mathrm{cr}} = \tfrac{2}{m}\big[(d-k)\langle\Gamma,\beta\rangle_F - \langle\Gamma_{\mathcal{S}^{\mathrm{c}}},\beta_{\mathcal{S}^{\mathrm{c}}}\rangle_F\big], \qquad \kappa_{\mathrm{bs}} = \|\bar{\boldsymbol{\beta}}_{\mathcal{S}^{\mathrm{c}}}\|_2^2 \ge 0,$$
$$\kappa_{\mathrm{sh}} = \tfrac{(d-k)c^2\sigma_{\mathrm{S}}^2}{m} = \Theta\Big(\tfrac{d-k}{mN_{\mathrm{sh}}}\Big).$$

If $\kappa_{\mathrm{bs}} > 0$, then $f\to\infty$ at both ends, so a finite minimizer exists, and it satisfies $2\kappa_{\mathrm{bs}}c^6 + \kappa_{\mathrm{cr}}c^4 = \kappa_{\mathrm{sh}}$. When the cross term is negligible, $c_\star = (\kappa_{\mathrm{sh}}/2\kappa_{\mathrm{bs}})^{1/6} = \Theta(N_{\mathrm{sh}}^{-1/6})$. If $\kappa_{\mathrm{cr}} > 0$ dominates at small c, then $c_\star = \Theta(N_{\mathrm{sh}}^{-1/4})$. If $\kappa_{\mathrm{cr}} < 0$, then $c_\star \to \sqrt{|\kappa_{\mathrm{cr}}|/(2\kappa_{\mathrm{bs}})}$ as $N_{\mathrm{sh}}\to\infty$. The sign of κ_{cr} depends on the circuit, so we treat c as a tuned hyperparameter.

Why the quotient is formed on circuit outputs. Suppose instead the quotient were formed on a shot estimate $\widehat{\mathcal{L}}_{\mathcal{B}}$ of the scalar loss. Write $\widehat{y}_i = y_i^\star + \delta_i$ with $\mathbb{E}[\delta_i] = 0$ and $\mathrm{Var}(\delta_i) = \nu_i/N_{\mathrm{sh}}$. For the MSE loss this gives the exact identity

$$\mathbb{E}\big[\widehat{\mathcal{L}}_{\mathcal{B}}(\boldsymbol{\theta})\big] = \mathcal{L}_{\mathcal{B}}(\boldsymbol{\theta}) + \varsigma^2(\boldsymbol{\theta}), \qquad \varsigma^2(\boldsymbol{\theta}) := \frac{1}{|\mathcal{B}|N_{\mathrm{sh}}}\sum_{i\in\mathcal{B}}\nu_i(\boldsymbol{\theta}). \tag{24}$$

This holds at every evaluation point, so such a scheme performs SPSA on $\mathcal{L}_{\mathcal{B}} + \varsigma^2$. Its bias contains $\partial_j\varsigma^2 = O(1/N_{\mathrm{sh}})$, which depends on neither c nor m, and it pushes the optimizer towards states of small readout variance. Eq. 20 uses the same circuit evaluations but never squares a measured quantity before differencing, so its bias remains that of Lemma 1.

B.7 Proof of Theorem 1

Let $\mathbb{E}_t$ condition on the history before step t, and let $\boldsymbol{G}_t := \nabla\mathcal{J}(\boldsymbol{\theta}_t)$, $\boldsymbol{b}_t := \mathbb{E}_t[\widetilde{\boldsymbol{g}}_t] - \boldsymbol{G}_t$ and $v_t := \mathrm{Var}_t(\widetilde{\boldsymbol{g}}_t)$.

Step 1: descent inequality for any estimator. Assume $\|\boldsymbol{b}_t\|_2^2 \le b^2$ and $v_t \le v$ for all t. By L_g-smoothness, $\mathcal{J}(\boldsymbol{\theta}_{t+1}) \le \mathcal{J}(\boldsymbol{\theta}_t) - \eta\langle \boldsymbol{G}_t, \widetilde{\boldsymbol{g}}_t\rangle + \frac{L_g\eta^2}{2}\|\widetilde{\boldsymbol{g}}_t\|_2^2$. Here $\mathbb{E}_t\langle \boldsymbol{G}_t, \widetilde{\boldsymbol{g}}_t\rangle = \|\boldsymbol{G}_t\|_2^2 + \langle \boldsymbol{G}_t, \boldsymbol{b}_t\rangle$ and $\mathbb{E}_t\|\widetilde{\boldsymbol{g}}_t\|_2^2 = \|\boldsymbol{G}_t + \boldsymbol{b}_t\|_2^2 + v_t \le 2\|\boldsymbol{G}_t\|_2^2 + 2b^2 + v$. Young's inequality gives $-\langle \boldsymbol{G}_t, \boldsymbol{b}_t\rangle \le \frac{1}{4}\|\boldsymbol{G}_t\|_2^2 + b^2$. With $L_g\eta \le \frac{1}{4}$,

$$\mathbb{E}_t\mathcal{J}(\boldsymbol{\theta}_{t+1}) \le \mathcal{J}(\boldsymbol{\theta}_t) - \frac{\eta}{2}\|\boldsymbol{G}_t\|_2^2 + \frac{5\eta}{4}b^2 + \frac{L_g\eta^2}{2}v.$$

Taking total expectations, summing over $t = 0, \dots, T-1$, using $\mathcal{J}(\boldsymbol{\theta}_T) \ge \mathcal{J}_\star$ and dividing by $\eta T/2$,

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\boldsymbol{G}_t\|_2^2 \le \frac{2\Delta_{\mathcal{J}}}{\eta T} + L_g\eta v + \frac{5}{2}b^2 \le \frac{2\Delta_{\mathcal{J}}}{\eta T} + L_g\eta v + 3b^2. \tag{25}$$

Step 2: part (i). Condition on $\mathcal{B}_t$. By Lemma 1, $\mathbb{E}[\widetilde{\boldsymbol{g}}_t \mid \mathcal{B}_t] = \boldsymbol{g}_{\mathcal{B}_t} + \boldsymbol{\beta}_t$ with $\|\boldsymbol{\beta}_t\|_2^2 \le F(k)$. Since $\mathbb{E}_t\boldsymbol{g}_{\mathcal{B}_t} = \boldsymbol{G}_t$, Jensen's inequality gives $\|\boldsymbol{b}_t\|_2^2 = \|\mathbb{E}_t\boldsymbol{\beta}_t\|_2^2 \le F(k)$. By the law of total variance,

$$v_t = \mathbb{E}_t\big[\operatorname{Var}(\widetilde{\boldsymbol{g}}_t \mid \mathcal{B}_t)\big] + \operatorname{Var}_t\big(\boldsymbol{g}_{\mathcal{B}_t} + \boldsymbol{\beta}_t\big).$$

The first term is at most $V(k)$ by Lemma 2, after dropping $-\|\Gamma_{\mathcal{S}^c}\|_F^2/m$ and using $\|\Gamma_t\|_F^2 \le \bar{G}$, $\sigma_{\mathrm{P},j}^2 \le \sigma_{\mathrm{P}}^2 \le \bar{s}/N_{\mathrm{sh}}$, $\sigma_{\mathrm{S}}^2 \le \bar{s}/(N_{\mathrm{sh}}c^2)$ (Eq. 7) and $|\mathcal{C}_c| = O((d-k)c^2/m)$; likewise $\|\boldsymbol{\beta}_t\|_2^2 \le (d-k)c^4B_\beta^2 + O((d-k)c^6)$ by Lemma 1 and the definition of B_β^2. For the second term, Minkowski's inequality for random vectors gives $\sqrt{\operatorname{Var}(X+Y)} \le \sqrt{\operatorname{Var}(X)} + \sqrt{\operatorname{Var}(Y)}$. Here $\operatorname{Var}_t(\boldsymbol{g}_{\mathcal{B}_t}) \le \sigma_{\mathcal{B}}^2$ and $\operatorname{Var}_t(\boldsymbol{\beta}_t) \le \mathbb{E}_t\|\boldsymbol{\beta}_t\|_2^2 \le F(k)$. Hence $v_t \le V(k) + (\sigma_{\mathcal{B}} + \sqrt{F(k)})^2 = v(k)$, and Eq. 25 gives Eq. 14.

Step 3: part (ii). No property of $\mathcal{S}_t$ is used. Since $\mathbb{E}_t\boldsymbol{g}_{\mathcal{B}_t} = \boldsymbol{G}_t$, $\boldsymbol{b}_t = \mathbb{E}_t[\widetilde{\boldsymbol{g}}_t - \boldsymbol{g}_{\mathcal{B}_t}]$, and Jensen's inequality gives $\|\boldsymbol{b}_t\|_2^2 \le \mathbb{E}_t\|\widetilde{\boldsymbol{g}}_t - \boldsymbol{g}_{\mathcal{B}_t}\|_2^2 \le e^2$. By Minkowski's inequality in L^2, $v_t \le \mathbb{E}_t\|\widetilde{\boldsymbol{g}}_t - \boldsymbol{G}_t\|_2^2 \le (e + \sigma_{\mathcal{B}})^2$. Eq. 25 completes the proof. □

B.8 Proofs of Corollaries 1–3

Corollary 1. Since $T \ge 1$, $\eta = 1/(4L_g\sqrt{T}) \le 1/(4L_g)$ and Theorem 1 applies, with $2\Delta_{\mathcal{J}}/(\eta T) = 8L_g\Delta_{\mathcal{J}}/\sqrt{T}$ and $L_g\eta v(k) = v(k)/(4\sqrt{T})$. By $(a+b)^2 \le 2a^2 + 2b^2$, $v(k) \le V(k) + 2\sigma_{\mathcal{B}}^2 + 2F(k)$, so the bound of part (i) is at most

$$\frac{1}{\sqrt{T}}\Big[8L_g\Delta_{\mathcal{J}} + \frac{\sigma_{\mathcal{B}}^2}{2} + \frac{V(k)}{4}\Big] + \Big(3 + \frac{1}{2\sqrt{T}}\Big)F(k), \qquad 3 + \frac{1}{2\sqrt{T}} \le \frac{7}{2}.$$

Inserting $V(k)$ from Eq. 15 and $F(k) \le (d-k)c^4B_\beta^2 + O((d-k)c^6)$ gives Eq. 16; the $O((d-k)c^2/m)$ term of $V(k)$ contributes $O((d-k)c^2/(m\sqrt{T}))$.

(a) With $\bar{s} = 0$ and $c \le (d-k)^{-1/4}T^{-1/8}$, we have $(d-k)c^4 \le T^{-1/2}$, so the floor is at most $\frac{7}{2}B_\beta^2/\sqrt{T}$. The neglected terms satisfy $(d-k)c^2/(m\sqrt{T}) \le (d-k)^{1/2}T^{-3/4}/m$ and $(d-k)c^6 \le (d-k)^{-1/2}T^{-3/4}$, both $O(T^{-3/4})$. This gives Eq. 17. For $k = d$ every coordinate is estimated by PSR, $F = V = 0$, and the bound is $A_0/\sqrt{T}$ for every c.

(b) For fixed c the bracket of Eq. 16 is a constant, which gives the first claim. The two terms that depend on c are $a\,c^{-2} + b\,c^4$ with $a = (d-k)\bar{s}/(4mN_{\mathrm{sh}}\sqrt{T})$ and $b = \frac{7}{2}(d-k)B_\beta^2$. Setting the derivative $-2ac^{-3} + 4bc^3$ to zero gives $c_\star^6 = a/(2b) = \bar{s}/(28\,mN_{\mathrm{sh}}B_\beta^2\sqrt{T})$, and the minimum value is

$$\tfrac{3}{2}\,(2b)^{1/3}a^{2/3} = \tfrac{3}{2}\cdot 7^{1/3}4^{-2/3}\,(d-k)\Big(\frac{B_\beta\bar{s}}{mN_{\mathrm{sh}}}\Big)^{2/3}T^{-1/3} \le C_{\mathrm{sh}}\,T^{-1/3},$$

since $\frac{3}{2}\cdot 7^{1/3}4^{-2/3} \approx 1.14 < 1.2$. The remaining terms of the bracket give $K_{\mathrm{sh}}/\sqrt{T}$. At $c = c_\star$ the neglected terms are $O((d-k)T^{-1/2})$ and $O(T^{-2/3})$, of lower order than $T^{-1/3}$.

For part (ii) of Theorem 1, the same substitution gives $\frac{1}{\sqrt{T}}\big[8L_g\Delta_{\mathcal{J}} + \frac{1}{4}(e+\sigma_{\mathcal{B}})^2\big] + 3e^2$. □

Corollary 2. *(i)* The coordinate biases of $\bar{\boldsymbol{g}}$ do not depend on m (Lemma 1). Hence $F_{\mathrm{SPSA}} - F_{\mathrm{SGiQ}}$ is the sum over $\mathcal{S}$ of these squared biases, which is nonnegative and equals $c^4\|\bar{\boldsymbol{\beta}}_{\mathcal{S}}\|_2^2 + O(c^6)$. *(ii)* Proposition 5 with $\widehat{\rho} = 0$ and m' directions gives the SPSA bound $2\|\boldsymbol{g}\|_2^2\|\Gamma\|_F^2/m'$, and $2(1-\widehat{\rho})/m < 2/(m+k)$ if and only if $\widehat{\rho} > k/(m+k)$. □

Corollary 3. Each step evaluates, per sample and measurement group, two circuits per PSR coordinate and two per SPSA direction (Section 4.3); multiplying by T and by N_{sh} gives the totals. In the ideal setting, Eq. 17 is at most ε, up to its $O(T^{-3/4})$ term, once $K(m,k,d)/\sqrt{T} \le \varepsilon$, that is, $T \ge K(m,k,d)^2/\varepsilon^2$. Multiplying by the per-step cost $2BQ(m+k)$ gives $C_{\mathrm{SGiQ}}(\varepsilon)$. For PSR ($k = d$), $K = A_0$ and the per-step cost is $2BQd$, which gives $C_{\mathrm{PSR}}(\varepsilon)$. Under finite shots with $c = c_\star$, $T \ge (2K_{\mathrm{sh}}/\varepsilon)^2$ gives $K_{\mathrm{sh}}/\sqrt{T} \le \varepsilon/2$, and $T \ge (2C_{\mathrm{sh}}/\varepsilon)^3$ gives $C_{\mathrm{sh}}T^{-1/3} \le \varepsilon/2$. Hence the bound is at most ε. □

B.9 Proof of Lemma 3

Fix $\boldsymbol{\theta}$ and $\mathcal{B}$, and write $\boldsymbol{e} := \bar{\boldsymbol{g}} - \boldsymbol{g}$ and $\epsilon := \max_j |e_j|$.

Relation between $\widehat{\rho}$ and ρ_k. Since $\mathcal{S}_\star$ maximizes $\|\boldsymbol{g}_S\|_2$ over $|S| = k$, $\widehat{\rho} \le \rho_k$. Since $\mathcal{S}$ maximizes $\|\bar{\boldsymbol{g}}_S\|_2$ over $|S| = k$, and $\|\boldsymbol{e}_S\|_2 \le \sqrt{k}\,\epsilon$ for every $|S| = k$,

$$\|\boldsymbol{g}_{\mathcal{S}}\|_2 \ge \|\bar{\boldsymbol{g}}_{\mathcal{S}}\|_2 - \sqrt{k}\,\epsilon \ge \|\bar{\boldsymbol{g}}_{\mathcal{S}_\star}\|_2 - \sqrt{k}\,\epsilon \ge \|\boldsymbol{g}_{\mathcal{S}_\star}\|_2 - 2\sqrt{k}\,\epsilon.$$

Dividing by $\|\boldsymbol{g}\|_2$ gives Eq. 12.

Deterministic recovery. If $\epsilon < \gamma/2$, then for $j \in \mathcal{S}_\star$ and $l \notin \mathcal{S}_\star$ we have $|g_j| \ge |g_l| + \gamma$ and

$$|\bar{g}_j| > |g_j| - \frac{\gamma}{2} \ge |g_l| + \frac{\gamma}{2} > |\bar{g}_l|,$$

so every score in $\mathcal{S}_\star$ strictly exceeds every score outside it, and $\mathrm{TopK}(|\bar{\boldsymbol{g}}|, k) = \mathcal{S}_\star$.

Probability of recovery. Write $\bar{g}_j = \frac{1}{m}\sum_{r=1}^{m} X_j^{(r)}$, where $X_j^{(r)} := \sum_{i\in\mathcal{B}} \lambda_i\, \widehat{u}_{ij}^{(r)}$ collects the contribution of the r-th direction over the batch. The $X_j^{(r)}$ are i.i.d. in r, with mean μ_j satisfying $|\mu_j - g_j| \le b_\infty$ (Lemma 1) and variance at most $\sigma_X^2 := \|\Gamma\|_F^2 + \sigma_{\mathrm{S}}^2 + O(c^2)$ (the single-direction case of Lemma 2). They are bounded: measured outputs satisfy $|\widehat{E}_O| \le \|O\|_{\mathrm{op}}$, so each difference quotient is at most $\|O\|_{\mathrm{op}}/c$ in magnitude, and in the ideal setting also at most $\|\nabla E_O\|_1 \le d\,\|O\|_{\mathrm{op}}$ by the mean-value theorem. Hence $|X_j^{(r)} - \mu_j| \le R := 2\Lambda\|O\|_{\mathrm{op}}\, r_c$ with $\Lambda := \sum_i |\lambda_i|$, where $r_c = \min\{d, 1/c\}$ in the ideal setting and $r_c = 1/c$ under finite shots. Bernstein's inequality gives, for every j,

$$\Pr\left(|\bar{g}_j - \mu_j| \ge t_\gamma\right) \le 2\exp\Big(-\frac{m\,t_\gamma^2}{2\sigma_X^2 + \frac{2}{3}Rt_\gamma}\Big).$$

Since $|e_j| \le |\bar{g}_j - \mu_j| + b_\infty$, the event $\epsilon \ge \gamma/2$ implies $|\bar{g}_j - \mu_j| \ge t_\gamma$ for some j, and a union bound over the d coordinates gives Eq. 13.

Consequences. On the recovery event, $\widehat{\rho} = \rho_k$, and $\widetilde{\boldsymbol{g}}$ coincides with the estimator that uses the fixed set $\mathcal{S}_\star$ with the same directions and shots; this estimator satisfies Assumption 2. Since $\widehat{\rho} \ge 0$, $\mathbb{E}[\widehat{\rho}] \ge \rho_k \Pr(\mathcal{S} = \mathcal{S}_\star) \ge (1 - p_{\mathrm{fail}})\rho_k$. Along a run, if the gap holds at every step, an induction over t shows that the iterates of Algorithm 1 and of the predictable variant using $\mathcal{S}_{\star,t}$ coincide on the event A that every step recovers $\mathcal{S}_{\star,t}$, and a union bound over the T steps gives $\Pr(A^{\mathrm{c}}) \le T p_{\mathrm{fail}}$. Let $X := \frac{1}{T}\sum_{t<T}\|\nabla\mathcal{J}(\boldsymbol{\theta}_t)\|_2^2 \in [0, G_\infty]$ for Algorithm 1 and $X_\star$ the same quantity for the variant. Since $X = X_\star$ on A, $\mathbb{E}[X] \le \mathbb{E}[X_\star] + G_\infty \Pr(A^{\mathrm{c}}) \le \mathbb{E}[X_\star] + T p_{\mathrm{fail}} G_\infty$. By Eq. 13, $T p_{\mathrm{fail}} G_\infty \le \varepsilon$ once $m \ge (2\sigma_X^2 + \frac{2}{3}Rt_\gamma)\, t_\gamma^{-2} \log(2dTG_\infty/\varepsilon)$. □

C Parameter screening and additional results

We evaluate SPSA with $m = 10, 20, \ldots, 80$, covering gradient budgets from 12.5% to 100% of PSR. The ideal SGiQ candidates are $(20, 10), (22, 8), (30, 10), (40, 10)$; finite-shot candidates are

$(10, 10), (12, 8), (20, 10), (30, 10), (40, 10), (40, 20)$. The latter span 25%–75% of the PSR budget. All configurations use the same three seeds and data split.

Within each run, minimum validation MAE selects the checkpoint; configurations are ranked by the mean of the three run-specific minima. These minima may occur at different epochs. The reported test scores use these checkpoints without further training. The tables below report the full screening together with the test errors of all screened configurations. SPSA-60 is the validation-selected finite-shot baseline; SPSA-50 has the lowest SPSA test MAE (0.11148). The experiments are exploratory: test results were inspected during development, so these are not independent confirmation results.

Table 3: Ideal screening and available tests under Adam $(0.9, 0.999)$. Values are means $\pm$ sample standard deviations over three seeds. A dash denotes an unavailable test summary.

Method	Best val. MAE	Test MAE	Test MSE	Cost
PSR	0.08635 ± 0.00807	0.08489 ± 0.01192	0.01304 ± 0.00261	960,000
SPSA-10	0.10794 ± 0.00371	0.12177 ± 0.01907	0.02688 ± 0.00813	120,000
SPSA-20	0.10077 ± 0.00453	0.10984 ± 0.00880	0.02151 ± 0.00341	240,000
SPSA-30	0.09019 ± 0.00347	0.10924 ± 0.01759	0.02122 ± 0.00370	360,000
SPSA-40	0.09861 ± 0.01143	0.10496 ± 0.00338	0.01991 ± 0.00308	480,000
SPSA-50	0.09624 ± 0.01312	0.10456 ± 0.01092	0.02087 ± 0.00504	600,000
SPSA-60	0.09247 ± 0.01132	0.10454 ± 0.01799	0.01888 ± 0.00601	720,000
SPSA-70	0.09226 ± 0.00771	0.10488 ± 0.00999	0.02033 ± 0.00490	840,000
SPSA-80	0.09512 ± 0.02361	0.09418 ± 0.01344	0.01741 ± 0.00351	960,000
SGiQ (20,10)	0.08943 ± 0.00782	0.11976 ± 0.01136	0.02519 ± 0.00723	360,000
SGiQ (22,8)	0.10406 ± 0.01308	–	–	360,000
SGiQ (30,10)	0.10768 ± 0.01552	–	–	480,000
SGiQ (40,10)	0.08343 ± 0.01582	0.09504 ± 0.00342	0.01647 ± 0.00197	600,000

Table 4: Finite-shot screening and available tests under Adam $(0.9, 0.999)$. Values are means $\pm$ sample standard deviations over three seeds. A dash denotes an unavailable test summary.

Method	Best val. MAE	Test MAE	Test MSE	Cost
PSR	0.09034 ± 0.01967	0.10865 ± 0.02196	0.02094 ± 0.01101	960,000
SPSA-10	0.11791 ± 0.00631	0.14602 ± 0.00621	0.03733 ± 0.00558	120,000
SPSA-20	0.09219 ± 0.01149	0.11579 ± 0.01090	0.02265 ± 0.00449	240,000
SPSA-30	0.10714 ± 0.00203	0.12727 ± 0.00227	0.02630 ± 0.00248	360,000
SPSA-40	0.09514 ± 0.01471	0.11657 ± 0.00938	0.02339 ± 0.00600	480,000
SPSA-50	0.09503 ± 0.00792	0.11148 ± 0.01654	0.02288 ± 0.00643	600,000
SPSA-60	0.08897 ± 0.00865	0.12070 ± 0.00264	0.02559 ± 0.00317	720,000
SPSA-70	0.09691 ± 0.00737	0.11249 ± 0.00733	0.02076 ± 0.00104	840,000
SPSA-80	0.11538 ± 0.02127	0.12000 ± 0.01074	0.02541 ± 0.00090	960,000
SGiQ (10,10)	0.13059 ± 0.00499	–	–	240,000
SGiQ (12,8)	0.11261 ± 0.01516	–	–	240,000
SGiQ (20,10)	0.10388 ± 0.00988	–	–	360,000
SGiQ (30,10)	0.10481 ± 0.00699	–	–	480,000
SGiQ (40,10)	0.09637 ± 0.01606	–	–	600,000
SGiQ (40,20)	0.08644 ± 0.00781	0.10714 ± 0.01119	0.02075 ± 0.00518	720,000

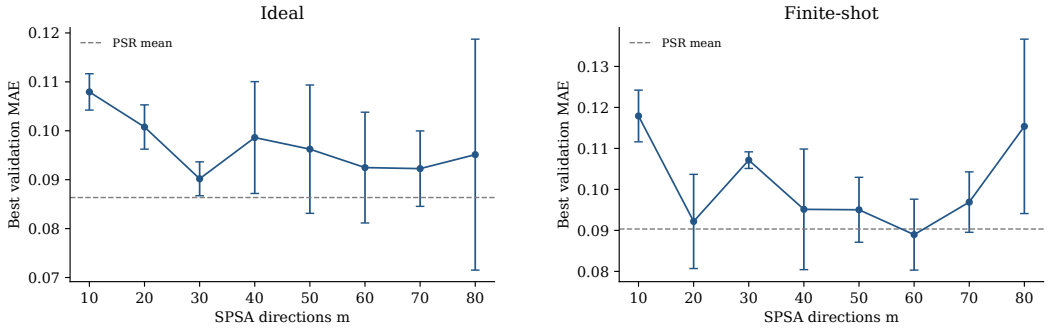

Figure 4: SPSA validation screening. Points and error bars show means and sample standard deviations across three seeds. PSR means are shown as references. Lines connect evaluated configurations only.

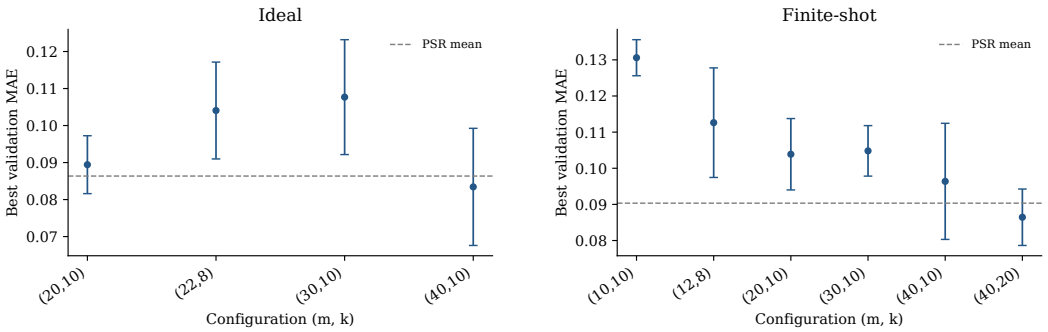

Figure 5: SGIQ validation screening. Labels indicate (m, k). At fixed $k = 10$, the ideal results are non-monotonic in m. In the finite-shot experiment, increasing k from 10 to 20 at $m = 40$ reduces mean best validation MAE from 0.09637 to 0.08644, while increasing the gradient budget by 20%.

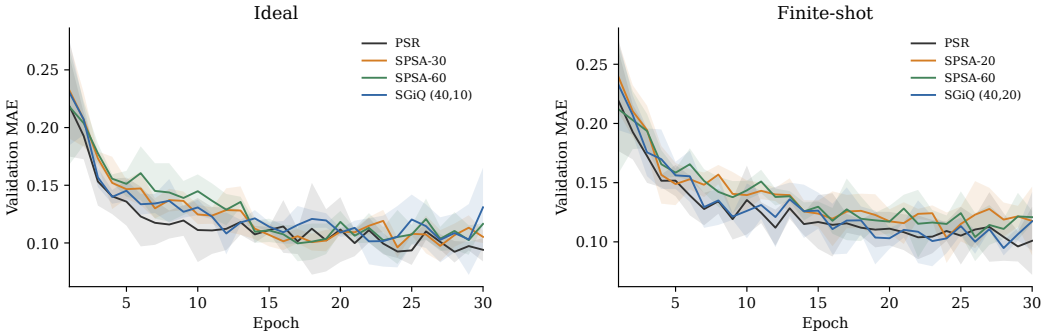

Figure 6: Per-epoch validation MAE of the main-text configurations. Curves show means and shaded bands show sample standard deviations across three seeds. These are end-of-epoch scores, not running minima. Test errors are evaluated only at validation-selected checkpoints.

Budget comparisons. Under ideal simulation, $(40, 10)$ also improves mean test MAE over equal-budget SPSA-50 (0.09504 versus 0.10456), whereas $(20, 10)$ does not improve over equal-budget SPSA-30 (0.11976 versus 0.10924). Under finite shots, $(40, 20)$ improves mean test MAE over SPSA-50 (0.10714 versus 0.11148) at 20% higher cost. Thus the benefit depends on the allocation and environment. The three-run standard deviations describe variability on one fixed split; they do not establish statistical significance or variation across data splits.