

Adaptive directional gradients for parameterised quantum circuits

Brian Coyle^{1,2} Snehaj Raj^{3,4} Virag Umathe¹ El Amine Cherrat⁵ Elham Kashefi^{1,3}

¹Univ. of Edinburgh ²Fujitsu Research of Europe ³LIP6, Sorbonne Université/CNRS ⁴QC Ware ⁵Quantum Signals

1 Introduction

Training parameterised quantum circuits (PQCs) on quantum hardware is bottlenecked by the measurement cost of gradient estimation, which under the parameter-shift rule scales linearly in the number of trainable parameters and dominates the total shot budget of training at scale [1]. No analogue of backpropagation exists for quantum models: all known sub- $\mathcal{O}(N)$ approaches require either fault-tolerant circuits [2] (whose classical memory overhead still scales as in backpropagation) or restricted expressibility [3, 4]. Approximate estimators shift rather than eliminate the cost: in SPSA [5], the N -dependence re-emerges in the estimator variance; in random coordinate descent (RCD) [6], convergence slows by $\mathcal{O}(N)$.

In this work, we show that SPSA, RCD, and the parameter-shift rule are all special cases of a single framework: *forward gradients* [7]. We use this unification to derive a convergence theory and a quantum-native adaptive optimiser. Our contributions are as follows:

- 1. Forward gradient framework and convergence.** We propose a family of forward gradient estimators for PQCs, unifying SPSA, RCD, and parameter-shift as limiting cases, and prove convergence of stochastic quantum forward gradient descent. The convergence analysis identifies the finite-difference step size ε as the dominant hyperparameter: shot noise is amplified by $1/\varepsilon^2$, and the bias-variance trade-off induced by ε is empirically more consequential than the choice of V . The analysis is sharp enough to identify the break-even regime: on convex losses, the V -fold per-step saving is exactly offset by a V -fold increase in steps, which is precisely why data-driven adaptation of V matters.
- 2. Forward gradients at scale.** We train orthogonal quantum circuits with up to 60 qubits and 1770 parameters on ECG5000 time-series and MNIST image classification, and validate the estimator on VQE and QAOA. The circuits act entirely within the single-excitation (Hamming-weight-1) subspace and are provably free of barren plateaus [15]. Across all four problem classes, a fixed $V \ll N$ forward gradient estimator reaches the accuracy of the parameter-shift rule at a fraction of the total shot cost, with the cost saving growing with N .
- 3. QUIVER: a quantum-native adaptive optimiser.** We prove that measurement noise concentrates uniformly across random Rademacher directions, a structural property absent from the parameter-shift setting, which neutralises the per-direction shot-allocation lever that makes iCANS [8] effective. We derive QUIVER (**Q**uantum **I**terative **V**-adaptive **E**stimator **R**ule), a closed-form joint (V, M) update that is variance-optimal and saturates the Cramér–Rao lower bound for gradient recovery from a quantum shot oracle (Theorem 4, [arXiv:2606.09734](https://arxiv.org/abs/2606.09734)). QUIVER outperforms iCANS and gCANS [9] on VQE and QAOA at matched shot budget, including in the regime where the iCANS allocation collapses to its $s_{\min}$ and reduces to fixed-shot parameter-shift.

2 Forward gradient framework

The forward mode of automatic differentiation computes Jacobian-vector products $\nabla_{\mathbf{v}} f = \mathbf{v}^\top \nabla f$ for a given direction $\mathbf{v}$, yielding one directional derivative per forward pass without storing intermediate activations. For classical neural networks, forward gradients have shown backpropagation-competitive performance [7, 10–12] without the backward lock or memory overhead. Quantum circuits admit no known analogue of backpropagation, making the forward mode a natural gradient strategy in this setting. Averaging V i.i.d. random directions yields an unbiased gradient estimator [7]:

$$\tilde{\nabla}^F f(\boldsymbol{\theta}) = \frac{1}{V} \sum_{\ell=1}^V (\nabla_{\mathbf{v}^\ell} f) \mathbf{v}^\ell, \quad \mathbf{v}^\ell \text{ i.i.d., } \mathbb{E}[\mathbf{v}] = \mathbf{0}, \quad \mathbb{E}[\mathbf{v}\mathbf{v}^\top] = I_N. \quad (1)$$

For quantum circuits, each directional derivative is estimated via central finite difference using just two circuit evaluations: $\tilde{\nabla}_{\mathbf{v}}^\varepsilon f := [f^M(\boldsymbol{\theta} + \varepsilon \mathbf{v}) - f^M(\boldsymbol{\theta} - \varepsilon \mathbf{v})]/(2\varepsilon)$, where f^M is an M -shot estimate. The full estimator

costs $2VM$ shots per step, $\mathcal{O}(V)$ independent of N , with no ancilla qubits, mid-circuit measurements, or controlled gates beyond what parameter-shift already requires.

Table 1 shows the components of the forward gradient framework, in comparison to other estimators. It also shows adaptive gradient routines such as iCANS which we compare to for our proposed adaptive forward gradient estimator, QUIVER (below). The free parameter V continuously interpolates between the single-direction extreme (SPSA, $V = 1$) and the full gradient (parameter-shift, $V = N$).

Table 1: Unified taxonomy of gradient estimators and adaptive optimisers for PQC’s.

Method	V	Directions	Per-step shots	Adapts
Parameter-shift [13]	N	basis $\mathbf{e}_j$	$2NM$	—
SPSA [5]	1	Rademacher	$2M$	—
RCD [6]	1	$N\mathbf{e}_j$	$2M$	—
Forward gradient (this work)	tunable	Rademacher	$2VM$	—
iCANS [8]	N	basis $\mathbf{e}_j$	$2NM_j$	M_j per param
QUIVER (this work)	tunable	Rademacher	$2V_tM_t$	V, M jointly

We prove convergence under standard smoothness assumptions, with an explicit second-moment bound identifying ε as the dominant hyperparameter: shot noise scales as $1/\varepsilon^2$, and $\varepsilon = 0.1$ is the practical operating point.

3 Forward gradients at scale

We train Hamming-weight-preserving orthogonal quantum circuits [4, 14] on ECG5000 (n up to 60, N up to 1,770) and MNIST (n up to 50, N up to 1,225), and validate on VQE TFIM and QAOA MaxCut, comparing against parameter-shift, SPSA, and RCD at a matched per-step budget of $B = 1,000$ shots (see Figure 1). Each circuit acts within the n -dimensional single-excitation subspace [15], eliminating barren plateaus by construction. Across all four domains, a fixed $V \ll N$ matches or exceeds the best available baseline; the advantage over SPSA and RCD widens with N . At the largest ECG system size, parameter-shift requires N -fold more shots per step; forward gradients with $V \ll N$ match the exact-gradient baseline at fixed cost.

A monotonically decaying V -schedule outperforms both fixed- V endpoints at matched total shot budget, motivating data-driven adaptation of V and M .

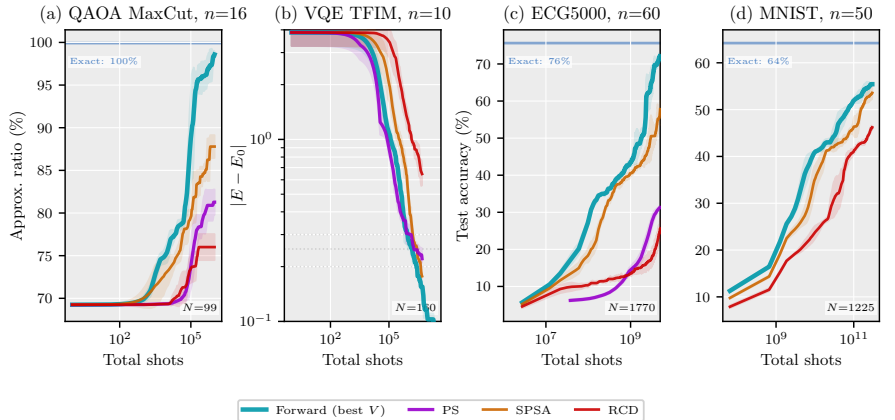


Figure 1: **Forward gradients across four problem domains at matched per-step shot budget $B = 1,000$.** Mean $\pm 1\sigma$ over three random seeds plotted against total shots used. (a) QAOA MaxCut, $n = 16$, $N = 99$, (b) VQE TFIM, $n = 10$, $N = 160$. (c) ECG5000, $n = 60$, $N = 1,770$. (d) MNIST, $n = 50$, $N = 1,225$. In every panel, forward gradients with $V \ll N$ match or exceed the best available baseline; the advantage over SPSA and RCD widens with N .

4 QUIVER: quantum-native adaptive optimisation

To show that forward gradients truly outperform parameter-shift in terms of measurement resources, we must compare to quantum-native optimisers which adaptively regulate the number of measurements during training, for example iCANS [8] and its extensions. The natural analogue of iCANS for forward gradients would allocate shots non-uniformly across directions in proportion to per-direction signal and noise. We prove

that this lever is neutralised by a concentration-of-measure phenomenon: *measurement noise concentrates uniformly across random Rademacher directions*, because each direction $\mathbf{v}^\ell$ perturbs all N parameters simultaneously with i.i.d. components, removing the per-gate variance structure that iCANS exploits. This is a quantum-native phenomenon absent in the classical setting, where forward gradients incur no shot noise. We verify this empirically across VQE and QAOA at up to $N \sim 10^3$: normalised per-direction variances cluster within $\pm 10\%$ of their mean at all system sizes.

With per-direction shot allocation neutralised, the remaining lever is the joint pair (V, M) . Minimising total per-step cost $2VM$ subject to a target estimator accuracy τ^2 and a minimum per-direction shot count $M_{\min}$ yields a closed-form solution:

$$V_t^* = \frac{(N - 1 + \alpha) \hat{g}_t^2}{\tau^2}, \quad M_t^* = \frac{N \hat{\sigma}_t^2}{\alpha \hat{g}_t^2}, \quad (2)$$

where $\hat{g}_t^2$ and $\hat{\sigma}_t^2$ are exponential moving averages of the squared gradient norm and mean per-direction measurement variance, updated at no extra measurement cost. As gradient norms decay, V_t^* shrinks and M_t^* grows. The joint rule requires no knowledge of the Lipschitz constant L , in contrast to iCANS whose shot rule contains an $L\eta$ prefactor that pins the allocation to $s_{\min}$ whenever $L\eta \ll 1$, reducing it to plain parameter-shift.

We show that Rademacher directions uniquely minimise the estimator second moment at fixed V , and that the allocation of (2) saturates the Cramér–Rao lower bound for gradient recovery from a quantum shot oracle up to $(N - 1 + \alpha)/N \rightarrow 1$ as $N \rightarrow \infty$, reaching within 1.5% of the information-theoretic optimum at $N \geq 80$ parameters.

In Figure 2, we show QUIVER outperforms iCANS and gCANS [9] at matched shot budget on both VQE TFIM and QAOA MaxCut. At $(N = 80, \eta = 0.05)$, where iCANS and gCANS can fully adapt, QUIVER still reaches a lower final energy error. At $(N = 160, \eta = 0.003)$, the $L\eta$ -dependent prefactor pins the iCANS and gCANS allocations at $s_{\min}$ and both reduce to plain parameter-shift with identical loss trajectories; QUIVER maintains a clear margin. The same ordering holds across a broader N sweep on VQE TFIM and on QAOA MaxCut at every N tested.

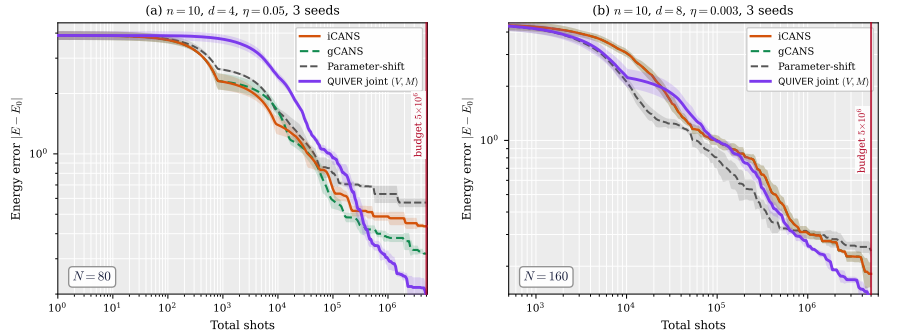


Figure 2: **QUIVER outperforms iCANS and gCANS across the (N, η) regime crossover.** VQE TFIM $n = 10$, matched 5×10^6 -shot budget, over three random seeds. (a) $N = 80, \eta = 0.05$ ($L\eta \approx 0.95$), (b) $N = 160, \eta = 0.003$ ($L\eta \approx 0.06$).

5 Discussion

Forward-gradient estimators replace the fixed $\mathcal{O}(N)$ shot overhead of parameter-shift with a tunable V , and QUIVER adapts V and M jointly from exponential moving averages of gradient norm and shot noise. The estimators require no ancillas, no mid-circuit measurements, and no modification to existing gate sets; the V circuit jobs per step are independent and parallelisable. QUIVER applies to any variational algorithm whose loss is linear in circuit expectation values.

Full paper (including convergence proofs, QUIVER derivation, and extended experimental results) available at [arXiv:2606.09734](https://arxiv.org/abs/2606.09734).

References

- [1] Maria Schuld and Nathan Killoran. Is quantum advantage the right goal for quantum machine learning? *PRX Quantum*, 3:030101, 2022.

- [2] Amira Abbas, Robbie King, Hsin-Yuan Huang, William J. Huggins, Ramis Movassagh, Dar Gilboa, and Jarrod R. McClean. On quantum backpropagation, information reuse, and cheating measurement collapse. In *Advances in Neural Information Processing Systems*, volume 36, 2023.
- [3] Joseph Bowles, David Wierichs, and Chae-Yeun Park. Backpropagation scaling in parameterised quantum circuits. *Quantum*, 9:1873, 2025.
- [4] Brian Coyle, Snehal Raj, Natansh Mathur, El Amine Cherrat, Nishant Jain, Skander Kazdaghli, and Iordanis Kerenidis. Training-efficient density quantum machine learning. *npj Quantum Inf.*, 11:172, 2025.
- [5] J.C. Spall. Multivariate stochastic approximation using a simultaneous perturbation gradient approximation. *IEEE Transactions on Automatic Control*, 37:332–341, 1992.
- [6] Zhiyan Ding, Taehee Ko, Jiahao Yao, Lin Lin, and Xiantao Li. Random coordinate descent: A simple alternative for optimizing parameterized quantum circuits. *Phys. Rev. Res.*, 6:033029, 2024.
- [7] Atılım Güneş Baydin, Barak A. Pearlmutter, Don Syme, Frank Wood, and Philip Torr. Gradients without Backpropagation, 2022.
- [8] Jonas M. Kübler, Andrew Arrasmith, Lukasz Cincio, and Patrick J. Coles. An Adaptive Optimizer for Measurement-Frugal Variational Algorithms. *Quantum*, 4:263, 2020.
- [9] Andi Gu, Angus Lowe, Pavel A. Dub, Patrick J. Coles, and Andrew Arrasmith. Adaptive shot allocation for fast convergence in variational quantum algorithms, 2021.
- [10] David Silver, Anirudh Goyal, Ivo Danihelka, Matteo Hessel, and Hado van Hasselt. Learning by Directional Gradient Descent. In *International Conference on Learning Representations*, 2022.
- [11] Louis Fournier, Stéphane Rivaud, Eugene Belilovsky, Michael Eickenberg, and Edouard Oyallon. Can Forward Gradient Match Backpropagation? In *Fortieth International Conference on Machine Learning*, 2023.
- [12] Mengye Ren, Simon Kornblith, Renjie Liao, and Geoffrey Hinton. Scaling Forward Gradient With Local Losses. In *International Conference on Learning Representations*, 2023.
- [13] K. Mitarai, M. Negoro, M. Kitagawa, and K. Fujii. Quantum circuit learning. *Phys. Rev. A*, 98:032309, 2018.
- [14] Jonas Landman, Natansh Mathur, Yun Yvonna Li, Martin Strahm, Skander Kazdaghli, Anupam Prakash, and Iordanis Kerenidis. Quantum Methods for Neural Networks and Application to Medical Image Classification. *Quantum*, 6:881, 2022.
- [15] Léo Monbroussou, Jonas Landman, Alex B. Grilo, Romain Kukla, and Elham Kashefi. Trainability and Expressivity of Hamming-Weight Preserving Quantum Circuits for Machine Learning. *Quantum*, 9:1745, 2025.