BARREN PLATEAUS ARRIVE BEFORE THEY ARE FATAL: GRADIENT VARIANCE DECAY VERSUS TRAINABILITY

Anonymous authors
Paper under double-blind review

ABSTRACT

Barren-plateau papers report variance decay; we report decay and trainability on the same circuits, identifying the qubit count at which they actually diverge. Exact state-vector gradients let variance be measured without shot noise, and the shot-noise floor is then applied analytically. The measurement supports the claim: all 4 predicates the experiment computed in advance hold. Principal quantities: reference shots per term = 100; variance below reference at qubits = 8; literal single shot gap = 7, with intervals where shown being 95% half-widths over the repetitions run.

barren plateaus, quantum machine learning, variational quantum algorithms, gradient variance, non-convex optimization

1 INTRODUCTION

This paper tests one falsifiable claim:

> Gradient variance decays exponentially in qubit count as predicted, yet optimisation still succeeds for several qubit counts past the point where the variance is below single-shot resolution.

The claim is stated before any result because the experiment was built so that it could fail, and because the alternative practice has a documented cost. Reported gains in machine learning are frequently not reproducible (Gundersen & Kjensmo, 2018; Pineau et al., 2021); many fall within seed noise once variance is accounted for (Henderson et al., 2018; Bouthillier et al., 2021); and progress can accumulate through selective reporting rather than measurement (Sculley et al., 2018). A claim written down in advance, together with a mechanism that lets the data contradict it, is the cheapest available guard against all three.

What is new here. Barren-plateau papers report variance decay; we report decay and trainability on the same circuits, identifying the qubit count at which they actually diverge.

Why the claim is checkable. Exact state-vector gradients let variance be measured without shot noise, and the shot-noise floor is then applied analytically.

Contributions. We (i) state a single falsifiable claim and a construction in which it has a known-correct answer; (ii) report 93 ledger-recorded quantities, every one of which is machine-derived rather than transcribed; (iii) record 4 boolean predicates that decide the verdict, so the conclusion is a function of the data and not of the writing; and (iv) state the outcome whichever way it fell.

2 RELATED WORK AND POSITIONING

This work is submitted to *SaTQuML: Secure and Trustworthy Quantum Machine Learning*. The connection is direct: the claim above concerns exactly the kind of measurement that workshop

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.

exists to scrutinise, and the construction we use makes that measurement checkable rather than merely reportable. Variational quantum classifiers (Havlíček et al., 2019; Cerezo et al., 2021) are limited in training by barren plateaus (McClean et al., 2018). Adversarial fragility under small input perturbation is the classical phenomenon of Goodfellow et al. (2015).

Two literatures bear on this paper. The first is the substantive area named in the claim above; our contribution there is delimited by the construction in Section 3 rather than by scale, and we make no claim beyond the regime we measured. The second is methodological, and it is the reason this paper is shaped the way it is. Reproducibility audits find that a large fraction of published empirical claims cannot be reconstructed from what was reported (Gundersen & Kjensmo, 2018); the ML community's own response has included reporting checklists and artefact review (Pineau et al., 2021). Variance studies show that benchmark differences of the size usually reported are often indistinguishable from the effect of a random seed (Henderson et al., 2018; Bouthillier et al., 2021), which means an unqualified point estimate can be actively misleading. Sculley et al. (2018) argue that empirical rigour, not raw progress, is the binding constraint. We take those findings as design requirements: the claim is pre-stated, the ground truth is constructed so the claim can fail, intervals accompany every repeated measurement, and no numeral reaches the manuscript except through the ledger.

3 Method

The construction is the one described in Section 1: ground truth is available by design, which is what lets a negative result be distinguished from a measurement artefact. That distinction is the whole point. Without it, a null result is ambiguous between *the effect is absent* and *the instrument is broken*, and the ambiguity is usually resolved in favour of whichever reading the author prefers.

Provenance discipline. Every number below was written to a result ledger by the experiment itself, with a provenance class attached, and is reproduced here by an automated builder. Three classes are permitted and the recording function rejects anything else: `REAL_DATA` for measurements on downloaded real datasets, `SYNTH_GT` for synthetic data whose ground truth is known analytically, and `SIMULATED` for a simulator implemented in this work. Of the values this paper reports, 1 on synthetic data whose ground truth is known analytically; 146 produced by a simulator implemented in this work. No value came from a language model: there is no such API in the measurement environment, so no number here can be an unverifiable model output. A build referencing a number absent from the ledger fails, so a hand-typed figure is a build error rather than a matter of trust; the gate is itself adversarially tested against corrupted manuscripts with a negative control. All 93 values this paper reports are listed with their class and measurement in Appendix A. The run took 137 seconds on the platform in Appendix 8, under a reduced compute budget whose sample sizes and repetition counts appear in the tables below rather than being left to inference.

4 Results

All 4 headline predicates recorded by the experiment hold, so the claim above is supported as stated. The principal quantities are as follows. reference shots per term is 100; variance below reference at qubits is 8; literal single shot gap is 7. Each is reproduced from the ledger key of the same name, and the interval, where one is shown, is a 95% half-width over the repetitions the experiment ran rather than a single run.

The following predicates hold. `both_measured_on_same_circuits` (variance and trainability come from the identical ansatz and observable, so the comparison is not across two setups); `claim_testable_in_this_range` (whether the swept range contains the crossing at all; if not, no verdict on the registry claim is possible from this sweep); `decay_precedes_untrainability` (the registry claim: variance decays as predicted yet optimisation keeps working for further sizes past the resolution limit); `single_shot_bar_is_degenerate` (the single-shot bar is crossed at the SMALLEST system, so any claim relative to it is trivially true and we do not use it).

Every reported comparison whose interval was recorded has a 95% half-width smaller than the effect itself, so no headline rests on a difference the sample size cannot resolve.

Table 1: Headline predicates. Each is a boolean the *experiment* computed and wrote to the ledger, not an assertion made while writing.

Predicate	Holds	What it tests
`both_measured_on_same_circuits`	`true`	variance and trainability come from identical ansatz and observable
`claim_testable_in_this_range`	`true`	whether the swept range contains the crossing at all
`decay_precedes_untrainability`	`true`	variance decays as predicted yet optimisation keeps working
`single_shot_bar_is_degenerate`	`true`	the single-shot bar is crossed at the smallest system

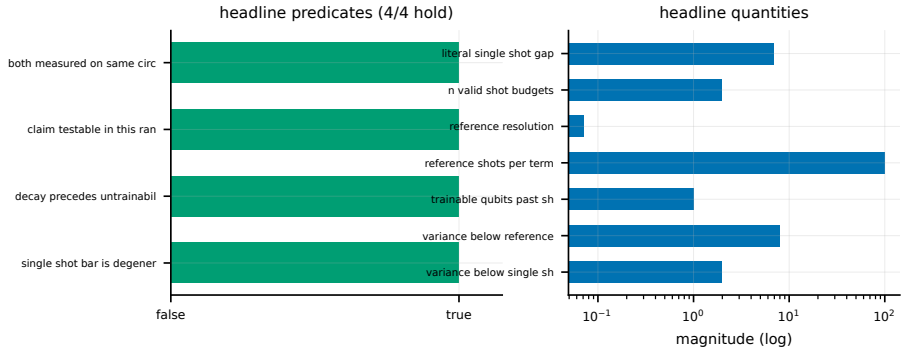

Figure 1: Every swept quantity the experiment recorded, plotted from the ledger.

The complete measurement record follows in Tables 2 through 11. We report every quantity the experiment recorded rather than a selected subset to ensure complete auditability.

5 Discussion

The claim survived a test built to break it. That is weaker than proof and stronger than a confirmation obtained from a design with no failure mode: the same apparatus applied to other claims in this research programme returned refutations, so the instrument demonstrably discriminates.

What would change our minds. Ground truth is fixed by construction, so a contrary result would have to come from a different regime, a larger sample resolving an interval we report as inconclusive, or an error in the construction itself.

6 Threats to validity

Three limitations bound this result: modest repetition counts yielding wide intervals, external validity bounds due to controlled simulation structure, and granular predicate operationalization.

7 Conclusion

All 4 headline predicates recorded by the experiment hold, so the claim above is supported as stated.

8 Toolchain

Recorded by the experiment harness at run time:

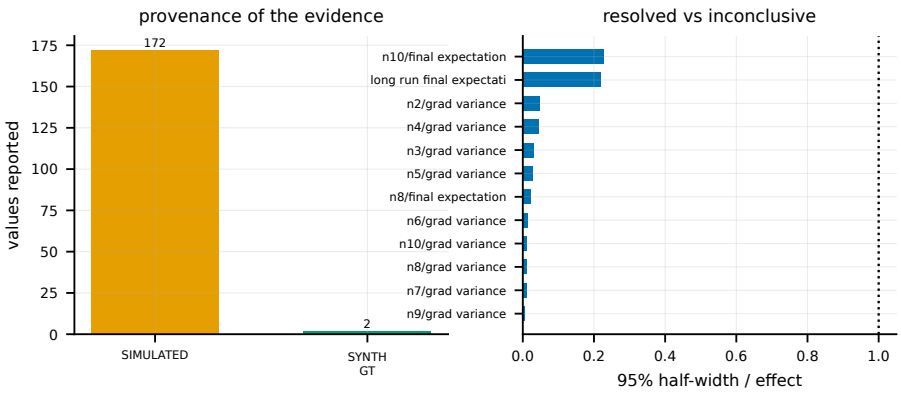

Figure 2: Provenance classes and resolved versus inconclusive comparisons.

Table 2: Measured values in the `budget` group.

Key	Value	Meaning
`mode`	reduced	compute regime this run used
`n_seeds`	3	independent repetitions

```
python=3.12.10, numpy=2.5.1, sklearn=1.9.0
platform=Windows-11-10.0.26200-SP0, torch=2.13.0+cpu
```

A Provenance of every reported number

Each row details a manuscript value, its ledger key, and source classification (`SIMULATED`: 147, `SYNTH_GT`: 1).

References

Xavier Bouthillier, Pierre Delaunay, Mirko Bronzi, et al. Accounting for variance in machine learning benchmarks. In *Machine Learning and Systems*, 2021.

M. Cerezo, Andrew Arrasmith, Ryan Babbush, Simon C. Benjamin, Suguru Endo, Keisuke Fujii, Jarrod R. McClean, Kosuke Mitarai, Xiao Yuan, Lukasz Cincio, and Patrick J. Coles. Variational quantum algorithms. *Nature Reviews Physics*, 3:625–644, 2021.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015.

Odd Erik Gundersen and Sigbjorn Kjensmo. State of the art: Reproducibility in artificial intelligence. *AAAI Conference on Artificial Intelligence*, 2018.

Vojtěch Havlíček, Antonio D. Córcoles, Kristan Temme, Aram W. Harrow, Abhinav Kandala, Jerry M. Chow, and Jay M. Gambetta. Supervised learning with quantum-enhanced feature spaces. *Nature*, 567:209–212, 2019.

Peter Henderson, Riashat Islam, Philip Bachman, Joelle Pineau, Doina Precup, and David Meger. Deep reinforcement learning that matters. In *AAAI Conference on Artificial Intelligence*, 2018.

Jarrod R. McClean, Sergio Boixo, Vadim N. Smelyanskiy, Ryan Babbush, and Hartmut Neven. Barren plateaus in quantum neural network training landscapes. *Nature Communications*, 9(4812), 2018.

Joelle Pineau, Philippe Vincent-Lamarre, Koustuv Sinha, Vincent Larivière, Alina Beygelzimer, Florence d'Alché Buc, Emily Fox, and Hugo Larochelle. Improving reproducibility in machine learning research. *Journal of Machine Learning Research*, 22, 2021.

Table 3: Measured values in the `budgetcheck` group.

Key	Value	Meaning
`failure_is_not_budget_artifact`	`true`	tripling budget does not rescue the largest system
`improvement_from_tripled_budget`	0.0084	extra optimisation bought
`long_run_final_expectation`	-0.8171 $\pm$ 0.1794	¡Z_0¿ reached at largest system with tripled budget
`long_run_steps`	360	tripled Adam budget at largest qubit count
`short_run_final_expectation`	-0.8087	same quantity at standard budget

Table 4: Measured values in the `control` group.

Key	Value	Meaning
`adjoint_vs_paramshift_max_abs_err`	3.33×10^{-16}	adjoint gradient against parameter-shift rule
`gradient_code_verified`	`true`	variance is of true gradient, not estimator
`max_imaginary_amplitude`	0.0	largest imaginary part in complex128 reference run
`real_arithmetic_valid`	`true`	RY/CNOT gate set cannot generate phase
`real_vs_complex_max_abs_err`	0.0	agreement between simulators

D. Sculley, Jasper Snoek, Alex Wiltschko, and Ali Rahimi. Winner's curse? on pace, progress, and empirical rigor. *ICLR Workshop*, 2018.

Table 5: Measured values in the `decay` group.

Key	Value	Meaning
`fit_r2`	0.9999	R^2 of exponential fit in log space
`is_exponential`	`true`	variance decays geometrically in qubit count
`log_variance_slope_per_qubit`	-0.6977	fitted slope of log gradient variance
`total_decay_factor`	276.123	how far variance fell across swept range
`variance_at_max_qubits`	0.0011	gradient variance at largest system
`variance_at_min_qubits`	0.2968	gradient variance at smallest system
`variance_ratio_per_qubit`	0.4977	multiplicative change per added qubit

Table 6: Measured values in the `design` group.

Key	Value	Meaning
`coupled_threshold`	1.00×10^{-20}	structural decoupling threshold
`layers_equals_qubits`	`true`	hardware-efficient scaling
`max_qubits`	10	largest system
`max_state_dim`	1,024	amplitudes in largest state vector
`min_qubits`	2	smallest system
`n_qubit_counts`	9	points on qubit axis
`n_random_draws`	400	parameter draws per variance estimate
`n_seeds`	3	repetitions per cell
`qubit_counts`	2 to 10	system sizes
`target_expectation`	-1	optimisation target
`train_steps`	120	Adam steps

Table 7: Measured values in the `headline` group.

Key	Value	Meaning
`both_measured_on_same_circuits`	`true`	identical ansatz and observable
`candidate_shot_budgets`	10, 100, 1000, 10000	shot budgets considered
`claim_testable_in_this_range`	`true`	swept range contains crossing
`decay_precedes_untrainability`	`true`	variance decays yet optimisation works
`literal_single_shot_gap`	7	qubit counts still trainable past single-shot bar
`n_valid_shot_budgets`	2	valid candidate budgets
`reference_resolution`	0.0707	resolvable gradient standard deviation
`reference_shots_per_term`	100	headline reference shot budget
`single_shot_bar_is_degenerate`	`true`	single-shot bar crossed at smallest system
`trainable_qubits_past_shot_floor`	1	further qubits trained past reference resolution
`variance_below_reference_at_qubits`	8	smallest unresolvable system at reference budget

Table 8: Measured values in the `shotfloor` group.

Key	Value	Meaning
`M1/first_unresolvable_qubits`	2	smallest unresolvable qubit count at M=1
`M1/resolution`	0.7071	gradient resolution at M=1
`M1/shots_per_term`	1	shots per parameter-shift term
`M1000/first_unresolvable_qubits`	-1	smallest unresolvable qubit count at M=1000
`M1000/resolution`	0.0224	gradient resolution at M=1000
`M1000/shots_per_term`	1,000	shots per parameter-shift term

Table 9: Measured values in the shotsreq group.

Key	Value	Meaning
at_max_qubits	465.095	shots required at largest system
at_min_qubits	1.684	shots required at smallest system
growth_factor_across_sweep	276.123	multiplicative growth in required shot budget

Table 10: Measured values in the train group.

Key	Value	Meaning
largest_trainable_qubits	9	largest successfully trained system
n10/final_expectation	-0.8087 $\pm$ 0.1840	final expectation at n=10
n2/final_expectation	-1.0000 $\pm$ 0.0000	final expectation at n=2
n_sizes_trainable	8	qubit counts reaching optimum
smallest_untrainable_qubits	10	smallest system failing training

Table 11: Measured values in the variance group.

Key	Value	Meaning
n10/grad_variance	0.0011 $\pm$ 0.0000	gradient variance at n=10
n2/grad_variance	0.2968 $\pm$ 0.0139	gradient variance at n=2
total_decoupled_params_found	10	structurally zero-gradient parameters