

Trainability Beyond Linearity in Variational Quantum Objectives

Gordon Ma¹ and Xiufan Li¹

Centre for Quantum Technologies, National University of Singapore

1 Introduction

Barren-plateau theorems concern gradients of expectation-value objectives [1–3], where the loss has the form

$$L(\theta) = \text{Tr}(H\rho(\theta))$$

for a fixed observable H . This form is load-bearing: each gradient component can then be written as the expectation of a fixed operator, and it is such expectations that concentrate under sufficiently random circuits. Many standard concentration-based proofs rely largely on three ingredients: i) fixed-observable reducibility, ii) sufficient circuit expressivity, and iii) a sufficiently global cost. Most mitigation strategies alter the circuit, cost, or initialization distribution while retaining the first. Such modifications can restore gradients, but tighten a familiar double bind: the same modifications tend to render the model classically tractable [4]—or, at best, relocate the problem to selecting among exponentially many trainable regions [12].

The first ingredient is less often isolated as a structural assumption, in part because $\text{Tr}(H\rho)$ is inherited from the origins of the variational paradigm in ground-state preparation [5] and appears almost intrinsic to what a quantum computer measures. Yet many learning objectives are not of this canonical form: divergences [6], likelihoods [7, 8], and risk functionals such as CVaR [9] are non-affine in the measured statistics. Existing results are conditional or downstream: bounded-sensitivity transfer theorems show that broad classes of non-affine losses inherit barren-plateau scaling [10]; particular unbounded objectives can exhibit distinct gradient behavior [6]; and finite-shot analyses reveal separate operational barriers [11]. These results determine whether particular constructions succeed or fail, but they do not isolate the structural reason that their behavior may differ from expectation-value costs. The upstream question is therefore: for which objectives is the fixed-observable reduction available at all, and what governs gradients when it is not? An exact answer to the first turns the second from a case-by-case question into a three-factor analysis.

2 The structural boundary

Given a set of fixed Hermitian observables $\{O_1, \dots, O_m\}$, we denote their expectations by $F(\rho) = (\text{Tr}(\rho O_1), \dots, \text{Tr}(\rho O_m)) \in \mathbb{R}^m$. We call this chosen family the *measurement interface*, fixed by which expectation values enter the loss rather than by how they are estimated, so the boundary below is independent of the measurement protocol. For an objective $L(\theta) = f(F(\rho(\theta)))$, where the classical function f — the *head* — scores the exposed statistics, our main result characterizes exactly when a fixed-observable representation exists.

Theorem 1. *Let $U \subset \mathbb{R}^m$ be a non-empty open set explored by the ansatz, i.e. $U \subset \{F(\rho(\theta)) : \theta \in \mathbb{R}^P\}$. Then there exist $H \in \text{span}\{O_1, \dots, O_m\}$ and $c \in \mathbb{R}$ such that*

$$L(\theta) = \text{Tr}(H\rho(\theta)) + c \quad \text{for all } \theta \text{ with } F(\rho(\theta)) \in U$$

if and only if f is affine on U .

The proof is elementary (see [13]); its content lies in the premise it makes explicit. Existing transfer results establish, for some families, that particular non-affine losses inherit barren-plateau scaling; the boundary identifies the structural fact upstream of them—exact fixed-observable reducibility holds precisely for affine heads.

Thus non-affineness is not a guarantee of trainability. It marks the point at which the objective is no longer reducible to a single fixed observable, and whether gradients vanish must be analyzed through the interaction between the measurement interface and the classical head.

3 Gradients in the nonlinear regime

With no fixed observable to absorb it, the gradient is separated by the chain rule,

$$\nabla_{\theta} L = J_F^{\top} g_F, \quad g_F = \nabla_F f(F(\theta)),$$

where the feature Jacobian J_F carries the model’s responsivity and g_F the loss-side signal. With $u_{\max}$ a leading left singular vector of J_F , the transmittance $\mathcal{T} = |\langle g_F, u_{\max} \rangle| / \|g_F\| \in [0, 1]$ measures their alignment, and a singular-value bound gives

$$\|\nabla_{\theta} L\| \geq \sigma_{\max}(J_F) \mathcal{T} \|g_F\|.$$

Exponential decay thus requires at least one of the three factors to vanish exponentially; concentration-based arguments constrain only the first. The decomposition connects to the standard diagnostic in one direction: the second moment of $\sigma_{\max}(J_F) \|g_F\|$ upper-bounds the aggregate gradient variance, hence each per-parameter variance [13], so exponential suppression of the factors implies suppression of the variance diagnostic. The converse does not follow: variance suppression does not identify which of the three factors has collapsed.

This separates two loss classes. Bounded-gradient losses—smooth or Lipschitz, including the divergences of adversarial training—cannot compensate a contracting Jacobian and inherit suppression. Losses whose feature-space gradients are unbounded need not: for the negative log-likelihood, $\|g_F\|$ grows as the inverse of small probabilities, of order $2^{n/2}$ near the uniform regime. Amplification is real on the loss side, but only transmitted if $\mathcal{T}$ survives—and at the full probability-vector interface ($m = 2^n$) an isotropic baseline gives $\mathcal{T} = \Theta(2^{-n/2})$, neutralizing the growth. This is the gradient-level face of a distribution-level fact: generative models whose outputs anticoncentrate are untrainable on average [15]. Both classes fail at exponential width: the obstruction is not the circuit alone but its interaction with the interface, whose width is a design choice. Compressing to polynomial width, $m = \text{poly}(n)$, removes the dimensional source of vanishing transmittance and returns the three factors to play. Whether they can be jointly controlled is no longer a question about the loss, but about the representation through which the model meets it.

4 A compressed interface

We test one instantiation of a polynomial-width interface. A domain-wall state $|0\rangle^{\otimes n/2} |1\rangle^{\otimes n/2}$ is evolved under local $U(1)$ -conserving dynamics: total charge is preserved, and the initially sharp density profile relaxes toward a uniform one within the fixed-charge sector. The intermediate regime between these extremes, set by the local dynamics and the evolution time, defines the task: an unconstrained hardware-efficient student is trained to match the evolved teacher’s statistics. Rather than learning the full distribution over 2^n outcomes, the interface exposes only the joint charge of b contiguous blocks—at $n = 8, b = 4$, the domain wall itself reads $|00\,00\,11\,11\rangle \mapsto (0, 0, 2, 2)$. For fixed b this gives width $m(n) = \prod_j (|B_j| + 1) \sim (n/b + 1)^b = \text{poly}(n)$: the same system, read through coarse conserved-charge statistics rather than over individual bitstrings.

Three classical heads are compared on this task—linear (affine), Jensen–Shannon divergence (inheriting), and negative log-likelihood (amplification-capable)—with gradients probed under finite-shot estimation. The chain-rule hierarchy holds both exactly and at finite shots, ordered $\text{NLL} \gg \text{JSD} \gg \text{linear}$, with resolved NLL gradients $\sim 10^4 \times$ the linear baseline at $n = 24$ at matched shot cost (Fig. 1). Classifying the scaling by AICc model selection over equal-complexity models (details in Sec. VI of [13]), the linear and JSD trends decisively favor the exponentially decaying class; the NLL trend separates from it over the tested window ($n = 8\text{--}24$), though the data do not identify a unique sub-exponential trend.

We are deliberate about what this does and does not show. The shot budget required to resolve these gradients grows exponentially for all three heads (Fig. 1B,D); by the estimation criterion the task remains within the barren-plateau phenomenon, and a larger resolved gradient need not translate into faster optimization. What the numerics establish is different: the amplification mechanism

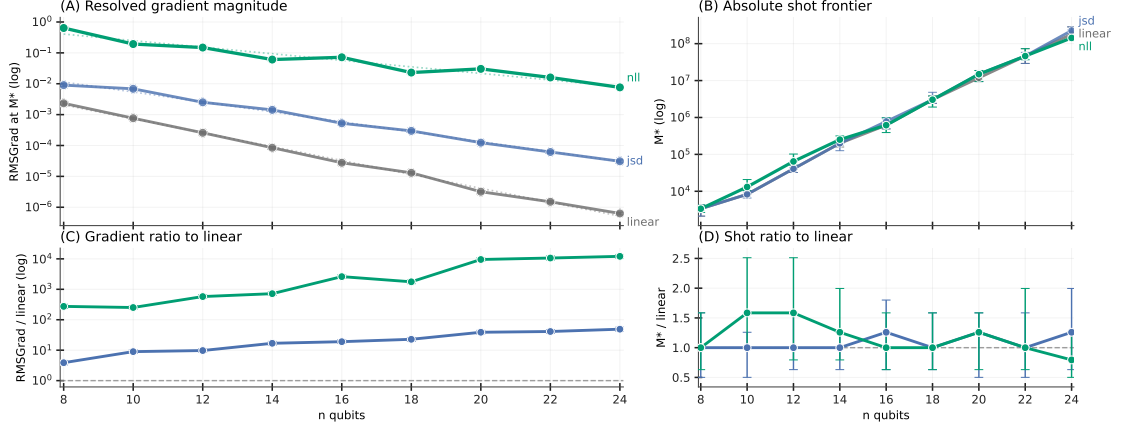


Fig. 1. Single-parameter finite-shot probe on the charge-conserving block interface ($b = 4$). The accepted frontier M^* is the smallest shot budget with median signal-to-noise ratio at least 2. NLL resolves gradients $\sim 10^4 \times$ larger than the affine baseline at $n = 24$ at comparable shot cost, while the observed M^* grows exponential-like for all three heads.

of Section 3 is real and measurable, and the remaining failure is localized to a single factor—the responsivity $\sigma_{\max}(J_F)$, the sensitivity of the exposed statistics to the circuit parameters. On this interface it decays because the statistics are still extensive: each block aggregates $\Theta(n/b)$ qubits, so the exposed variables grow with system size even though $m(n)$ stays polynomial. The head is thus a genuine design axis—it demonstrably changes the gradient scaling class—but one axis is not enough; the polynomial-width condition needs a finer-grained refinement toward statistics that remain intensive.

5 Outlook

Beyond the affine boundary, exponential suppression is not structurally enforced by the fixed-observable template alone: gradients are governed by three factors, set jointly by the exposed statistics and the head that scores them—non-affinity is not a loophole in the concentration proofs but a design space. The contribution here sharpens two programs. For barren-plateau theory, the non-affine regime is a ripe target: concentration guarantees that do not route through a fixed observable, in particular for unbounded heads, would be strictly more general than the affine template. For constructive work, the target is a positive demonstration under realistic resources: polynomial shots [11] and optimization that survives a rugged landscape [14]. We leave the constructive side as a hypothesis, Polynomially-Barren & Just-Right (PB&J): some physically natural tasks admit a polynomial-width interface and an initialization region of inverse-polynomial measure on which all three chain-rule factors, and hence the gradients, remain inverse-polynomial [13]. The diagnosis of Section 4 suggests a design principle: exposed statistics should remain intensive as the system grows—plausibly through transport-adapted or band-limited interfaces exposing slow modes associated with conservation laws, complementing recent spectral perspectives on quantum machine learning [16]. Neither program presumes quantum advantage: the head is classical, so the loss is classically simulable whenever the interface statistics are, and the double bind relocates to the interface rather than dissolving [4]. The boundary is what these arguments trace; the question that follows is which representations admit learning, and which do not.

Bibliography

- [1] McClean, Jarrod R., Sergio Boixo, Vadim N. Smelyanskiy, Ryan Babbush, and Hartmut Neven. Barren plateaus in quantum neural network training landscapes. *Nature Communications* 9.1 (2018): 4812.
- [2] Cerezo, Marco, Akira Sone, Tyler Volkoff, Lukasz Cincio, and Patrick J. Coles. Cost function dependent barren plateaus in shallow parametrized quantum circuits. *Nature Communications* 12.1 (2021): 1791.
- [3] Larocca, Martin, Supanut Thanasilp, Samson Wang, Kunal Sharma, Jacob Biamonte, Patrick J. Coles, Lukasz Cincio, Jarrod R. McClean, Zoë Holmes, and Marco Cerezo. Barren Plateaus in Variational Quantum Computing. *Nature Reviews Physics* 7.4 (2025): 174-189.
- [4] Cerezo, Marco, Martin Larocca, Diego García-Martín, Nelson L. Diaz, Paolo Braccia, Enrico Fontana, Manuel S. Rudolph et al. Does provable absence of barren plateaus imply classical simulability? *Nature Communications* 16, no. 1 (2025): 7907.
- [5] Peruzzo, Alberto, Jarrod McClean, Peter Shadbolt, Man-Hong Yung, Xiao-Qi Zhou, Peter J. Love, Alán Aspuru-Guzik, and Jeremy L. O'Brien. A variational eigenvalue solver on a photonic quantum processor. *Nature Communications* 5, no. 1 (2014): 4213.
- [6] Kieferova, Maria, Ortiz Marrero Carlos, and Nathan Wiebe. Quantum Generative Training Using Rényi Divergences. *arXiv* 2106.09567 (2021).
- [7] Coopmans, Luuk, and Marcello Benedetti. On the sample complexity of quantum Boltzmann machine learning. *Communications Physics* 7, no. 1 (2024): 274.
- [8] Verdon, Guillaume, Jacob Marks, Sasha Nanda, Stefan Leichenauer, and Jack Hidary. Quantum Hamiltonian-based models and the variational quantum thermalizer algorithm. *arXiv* 1910.02071 (2019).
- [9] Barkoutsos, Panagiotis Kl, Giacomo Nannicini, Anton Robert, Ivano Tavernelli, and Stefan Woerner. Improving variational quantum optimization using CVaR. *Quantum* 4 (2020): 256.
- [10] Thanasilp, Supanut, Samson Wang, Nhat Anh Nghiem, Patrick Coles, and Marco Cerezo. Subtleties in the trainability of quantum machine learning models. *Quantum Machine Intelligence* 5, no. 1 (2023): 21.
- [11] Rudolph, Manuel S., Sacha Lerch, Supanut Thanasilp, Oriel Kiss, Oxana Shaya, Sofia Vallecorsa, Michele Grossi, and Zoë Holmes. Trainability Barriers and Opportunities in Quantum Generative Modeling. *npj Quantum Information* 10, no. 1 (2024): 116.
- [12] Kulshrestha, Ankit, Ricard Puig, Diego García-Martín, et al. Exponentially Many Initializations to Avoid Barren Plateaus. *arXiv* 2606.18515 (2026).
- [13] Ma, Gordon, and Xiufan Li. Trainability Beyond Linearity in Variational Quantum Objectives. *arXiv* 2604.18846 (2026).
- [14] Anschuetz, Eric R., and Bobak T. Kiani. Quantum variational algorithms are swamped with traps. *Nature Communications* 13 (2022): 7760.
- [15] Herbst, Sabrina, Ivona Brandić, and Adrián Pérez-Salinas. Limits of quantum generative models with classical sampling hardness. *arXiv* 2512.24801 (2025).
- [16] Belis, Vasilis, Joseph Bowles, Rishabh Gupta, Evan Peters, and Maria Schuld. Spectral methods: crucial for machine learning, natural for quantum computers? *arXiv* 2603.24654 (2026).