

# Double Descent in Parameterized Quantum Circuits

Francesco Scala<sup>1</sup>, Sabri Meyer<sup>1</sup>, Francesco Tacchino<sup>2</sup>

<sup>1</sup>*Department of Mathematics and Computer Science University of Basel, Switzerland*

<sup>2</sup>*IBM Research – Zurich, 8803 Rüschlikon, Switzerland*

## Abstract

We present the first analytical and numerical characterization of double descent in parameterized quantum circuits (PQCs) for regression tasks. By exploiting the Fourier structure of PQC models and the lazy training approximation, we establish a precise connection between overparameterized gradient descent on variational parameters, the quantum neural tangent kernel, and the emergence of double descent. Our results show that the effective model complexity in PQCs is determined by both the number of Fourier modes and the number of variational parameters, leading to a richer landscape of interpolation regimes than in classical linear models. Crucially, the position of the second descent depends on correlations encoded in the circuit Jacobian.

**Introduction** While classical statistical learning theory predicts an increasing generalization error with growing model complexity, modern machine learning has profoundly challenged this picture. Belkin et al. [1] proposed a unified *double descent* (DD) risk curve, showing that after an initial U-shaped regime, continuing to increase model complexity beyond an *interpolation threshold*, the point where the model begins to exactly fit the training data, causes test error to decrease again. The mechanism behind this second descent is that in this overparameterized regime the gradient-based optimizer is biased toward the minimum  $\ell_2$ -norm interpolating solution, which tends to be smoother and generalizes better. Despite the rich theoretical understanding of DD in classical settings [2], the phenomenon has remained largely unexplored in PQCs. Indeed, Du et al. [3] analytically showed that quantum neural network classifiers built from 2-design circuits undergo a U-shaped risk curve, explicitly ruling out DD for such models in classification tasks. However, this negative result is specific to classification losses with 2-design models, while non-Haar-random PQC ansätze and other tasks have not been systematically investigated. Recent work has begun to probe DD in quantum kernel methods [4, 5], but DD in the PQC parameter optimization setting, where gradient descent (GD) is used to train variational parameters directly, remains uncharted territory.

**Classical double descent from linear regression** DD can be analytically understood in the tractable setting of linear regression, where the interpolation threshold corresponds to the number of parameters equalling the number of training samples  $n$ . We consider supervised learning over  $\mathcal{S} = \{(\mathbf{x}_\ell, y_\ell)\}_{\ell=1}^n$  drawn i.i.d. from  $\mathcal{D}$  on  $\mathcal{X} \times \mathcal{Y}$ , with labels  $y = f_*(\mathbf{x}) + \varepsilon$ ,  $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ . Given a feature map  $\phi : \mathcal{X} \rightarrow \mathbb{R}^d$ , we learn  $f_w(\mathbf{x}) = w^\top \phi(\mathbf{x})$  by minimizing the empirical mean squared error (MSE)  $\mathcal{L}_n(w) = \frac{1}{2n} \|\Phi^\top w - Y\|_2^2$ , where  $\Phi = [\phi(\mathbf{x}_1), \dots, \phi(\mathbf{x}_n)] \in \mathbb{R}^{d \times n}$  and  $Y \in \mathbb{R}^n$ . Generalization is measured by the expected risk  $\mathcal{R}(w) = \mathbb{E}_{(\mathbf{x}, y) \sim \mathcal{D}}[(w^\top \phi(\mathbf{x}) - y)^2]$ , whose asymptotic value  $\mathcal{R}(\gamma) := \lim_{d, n \rightarrow \infty, d/n \rightarrow \gamma} \mathcal{R}(\hat{w})$  exhibits *double descent* as a function of the overparameterization ratio  $\gamma = d/n$  [1, 2]. The central estimator is the *minimum-norm interpolator*

$$\hat{w} = \Phi^+ Y, \quad \Phi^+ = (\Phi^\top \Phi)^{-1} \Phi \text{ if } \gamma < 1, \quad \Phi^+ = \Phi^\top (\Phi \Phi^\top)^{-1} \text{ if } \gamma > 1, \quad (1)$$

the Moore-Penrose pseudoinverse solution, equivalently obtained by minimizing  $\mathcal{L}_n(w)$  with GD from  $w_0 = 0$ .

The risk’s bias-variance decomposition shapes  $\mathcal{R}(\gamma)$  across three regimes. In the *underparameterized regime* ( $\gamma < 1$ ), bias decreases and variance rises with  $\gamma$ , giving a classical U-shaped curve. At the *interpolation threshold* ( $\gamma \rightarrow 1$ ),  $\Phi$  becomes asymptotically square, its minimum singular value vanishes, the feature covariance  $\Sigma = \mathbb{E}_{\mathbf{x}}[\phi(\mathbf{x})\phi(\mathbf{x})^\top]$  becomes ill-conditioned, and variance diverges. In the *overparameterized regime* ( $\gamma > 1$ ), the system is underdetermined and the minimum-norm solution acts as an implicit regularizer, suppressing noise amplification along small-singular-value directions and driving the second descent. This confirms that DD stems from the inductive bias of the learning algorithm toward the minimum-norm solution, not merely from model expressivity.

**PQCs as linear regression models** Parameterized quantum circuits (PQCs) have emerged as a leading paradigm for quantum machine learning (QML). A standard PQC model takes the form  $f_\theta(\mathbf{x}) = \langle 0|U^\dagger(\mathbf{x}, \theta) O U(\mathbf{x}, \theta)|0\rangle$ , and Schuld et al. [6] showed that such models are naturally expressed as truncated Fourier series in the input data,

$$f(\mathbf{x}, \theta) = \sum_{\omega \in \Omega} c_{\omega}(\theta) e^{i\omega \cdot \mathbf{x}} = \phi(\mathbf{x})^{\top} c(\theta), \quad (2)$$

where the accessible frequency set  $\Omega$  is fixed by the data-encoding gates and the Fourier coefficients  $c_{\omega}(\theta)$  are non-trivially determined by the variational parameters  $\theta$ . This Fourier picture provides a natural basis for studying the complexity of PQC models: the number of Fourier modes  $|\Omega|$  sets the effective degrees of freedom accessible to the model which can be seen as a linear regression model with parameters  $w = c(\theta)$ .

**Fourier Analysis and Lazy Training** We investigate DD in PQCs trained by GD on a MSE loss. Our entry point is the Fourier representation Eq. (2), mapping the learning problem to a form amenable to linear regression analysis, where the relevant complexity is the number of frequencies  $|\Omega|$  and the interpolation threshold is  $|\Omega| = n$ . A key subtlety distinguishes PQCs from classical linear models: gradient descent updates the variational parameters  $\theta$ , not the Fourier coefficients  $c_{\omega}(\theta)$  directly, which depend on  $\theta$  in a highly non-trivial, correlated manner. Here, we initialize the model such that  $c(\theta_0) = 0$  and invoke the *lazy training* approximation [7]: parameters remain close to initialization throughout training, so the model output linearizes via a first-order Taylor expansion around  $\theta_0$ :

$$f(\mathbf{x}, \theta) \approx f(\mathbf{x}, \theta_0) + \phi(\mathbf{x})^{\top} J_0 \delta\theta = \tilde{\phi}(\mathbf{x})^{\top} \delta\theta, \quad \tilde{\phi}(\mathbf{x}) = J_0^{\top} \phi(\mathbf{x}) \in \mathbb{C}^p, \quad (3)$$

where  $J_0 = \partial c(\theta)/\partial\theta|_{\theta_0} \in \mathbb{C}^{|\Omega| \times p}$  is the Jacobian at initialization,  $\delta\theta = \theta - \theta_0$ , and  $f(\mathbf{x}, \theta_0) = 0$  since  $c(\theta_0) = 0$ . In this regime, GD on  $\theta$  is equivalent to linear regression via Eq. (1), with effective feature map  $\tilde{\phi}(\mathbf{x})$  governed by  $J_0$ , connecting PQC training to overparameterized linear regression.

As for classical models, DD depends on the feature covariance matrix, here  $\tilde{\Sigma} = \mathbb{E}_{\mathbf{x}}[\tilde{\phi}(\mathbf{x}) \tilde{\phi}(\mathbf{x})^{\top}] = J_0^{\top} J_0 \in \mathbb{R}^{p \times p}$ , which is set by the Jacobian, generally correlated, possibly rank-deficient, and input-independent as a structural consequence of Fourier feature orthogonality under uniform inputs. Setting  $\theta_0 = 0$ , the overparameterized solution  $\hat{\theta} = \tilde{\Phi}^{\top} (\tilde{\Phi} \tilde{\Phi}^{\top})^{-1} Y$  (with  $\tilde{\Phi} \in \mathbb{C}^{n \times p}$  the effective feature map for the training set) gives the prediction formula

$$f(\mathbf{x}_{test}, \theta) \approx \tilde{\phi}(\mathbf{x}_{test})^{\top} \tilde{\Phi}^{\top} (\tilde{\Phi} \tilde{\Phi}^{\top})^{-1} Y = K(\mathbf{x}_{test}, X) K(X, X)^{-1} Y, \quad (4)$$

with  $K_{ij} = K(\mathbf{x}_i, \mathbf{x}_j) := \tilde{\phi}(\mathbf{x}_i)^{\top} \tilde{\phi}(\mathbf{x}_j)$  the Quantum Neural Tangent Kernel (QNTK) [8], linking the DD mechanism to existing QNTK diagnostics and motivating the analysis of double descent under lazy training [9].

**Two complexity regimes** The lazy training picture reveals two regimes controlled by the interplay between the number of variational parameters  $p$ , the number of Fourier modes  $|\Omega|$ , and the number of training samples  $n$ . The effective model complexity is set by the smaller of  $p$  and  $|\Omega|$  (the rank of  $J_0$ ), unlike classical models where these coincide since  $w$  is trained directly. This gives two regimes:

- *Frequency-limited regime* ( $p > |\Omega|$ ): the Fourier spectrum saturates before the parameter count does, so the model behaves analogously to classical linear regression with  $d = |\Omega|$  degrees of freedom.
- *Parameter-limited regime* ( $p < |\Omega|$ ): parameters are exhausted before all Fourier modes can be independently controlled, giving  $d = p$  degrees of freedom. This is the regime typically expected in practice, since  $|\Omega|$  can scale exponentially in the number of qubits.

Under the lazy training assumption, we derive an analytic bias-variance decomposition of the test error that tracks the DD curve through both regimes, showing the role played by  $J_0$ . The initialization protocol is essential: the Fourier coefficients must vanish at initialization,  $c_{\omega}(\theta_0) = 0 \forall \omega \in \Omega$ , requiring  $\theta_0 = 0$  along with other design choices. This mimics the classical zero-initialization protocol selecting the minimum-norm interpolant in the overparameterized regime. A different initialization can instead inflate the bias term, potentially masking the variance behaviour.

**Observable dependence of lazy training** Our framework critically depends on the validity of lazy training, which requires the Jacobian  $J$  to be approximately Lipschitz, i.e., stable throughout optimization. Ref. [7] already noted that this validity is strongly sensitive to the choice of observable: for global observables acting on all qubits,  $J$  varies substantially during training, breaking the assumption, while for local observables acting on single qubits, lazy training is expected to hold. This allows the DD curve to be forecast before training, and its features characterized via the effective covariance matrix  $\tilde{\Sigma}$ . In particular, the interpolation threshold can be estimated empirically by monitoring its rank: when  $\text{rank}(\tilde{\Sigma}) = n$ , the model is at threshold.

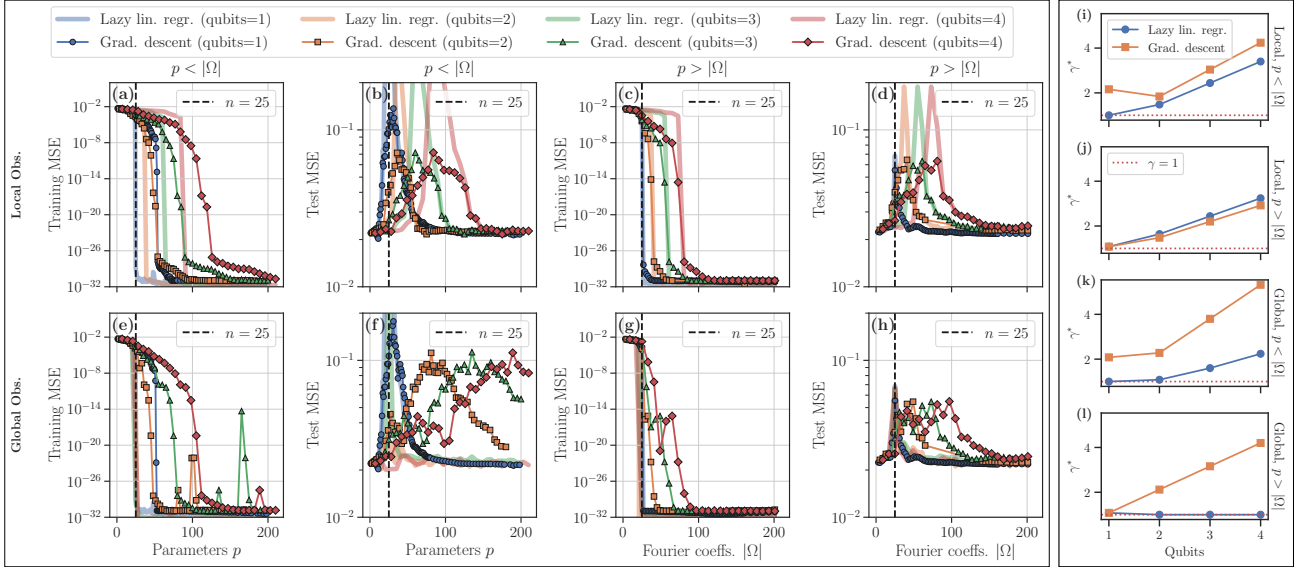


Figure 1: **Comparison of PQCs trained with GD and its lazy approximation** (Left panel) Training and test MSE as a function of degrees of freedom  $\min(p, |\Omega|)$  for  $n = 25$  training samples. The curves are averaged over 15 independent noisy datasets. Panels (a) to (d) show the local observable case, while panels (e) to (h) show the global observable case. The first two columns show the  $p < |\Omega|$  case, while third and fourth the  $p > |\Omega|$  case. (Right panel) Position of the interpolation threshold  $\gamma^*$  for different numbers of qubits. Both comparison panels include gradient descent and the lazy linearized prediction for qubit numbers 1 to 4.

Correlations in  $J_0$ , and hence in the effective feature map, can slow the growth of  $\text{rank}(\tilde{\Sigma})$ , effectively shifting the threshold's position.

**Numerical results** Fig. 1 confirms our theoretical predictions on a synthetic noisy dataset  $y = f_*(x) + \varepsilon$  with 25 training and 200 test samples. The circuits alternate data encoding and trainable layers: each layer applies RY encodings on all qubits, then single qubit RX/RZ rotations and nearest neighbor CRX entanglers, with readout via Pauli Y (local or global). For the global observable, a pre-layer of  $\text{RX}(-\pi/2)$  gates on all but one qubit ensures  $c(\theta_0) = 0$  and  $J_0 \neq 0$ . Varying the number of layers and gates in the encoding and variational ansatz controls both  $p$  and  $|\Omega|$ . We compare PQCs trained with GD against the lazy approximation obtained via inversion (Eq. (1)). For local observables, the lazy theory well approximates the full DD curve of test MSE (panels (b), (d)) and the shifted interpolation point across qubit numbers, indicating that Jacobian statistics near initialization already capture most of the relevant learning geometry; past the interpolation threshold the two curves overlap, confirming Eq. (4). For global observables (panels (f), (h)), this picture breaks down: the lazy approximation predicts a DD peak at  $\gamma^* = d/n = 1$ , while GD shows a shifted threshold, indicating that correlations change during training and the linearized approximation fails. This is detailed further in the right panels: for local observables the threshold shifts away from  $\gamma = 1$  and is well captured by the lazy approximation, while for global observables the lazy model stays pinned at  $\gamma = 1$  and GD exhibits a clear qubit dependent shift. Here  $\gamma^*$  is defined as the point where  $\text{rank}(\tilde{\Sigma}) \geq n$  for the linearized model, and as the point where the training MSE falls below  $10^{-13}$  for GD. The training MSE plots (panels (a), (c), (e), (g)) show the same pattern: agreement between linearized and GD values for local observables, while for global observables the linearized train MSE drops to 0 at  $\gamma = 1$ . This gap partly reflects our convergence criterion: to limit computational cost, GD training was stopped at  $2.5 \cdot 10^6$  epochs, gradient norm below  $10^{-20}$ , or training loss below  $1 \cdot 10^{-31}$ . While our approach generalizes to any number of qubits, we limited experiments to a few qubits to keep runtimes reasonable across settings.

**Conclusions** While the zero-initialization protocol combined with the lazy training assumption yields classically simulable models [10], this framework opens a path toward understanding double descent in parameterized quantum circuits. Future work may eventually combine these findings with encodings that offer computational advantages. The introduced analytical approach to double descent in PQCs provides a useful testbed for understanding overparameterized learning beyond classical settings.

## References

- [1] Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019.
- [2] Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, and Ilya Sutskever. Deep double descent: Where bigger models and more data hurt. In *International Conference on Learning Representations (ICLR)*, 2020.
- [3] Yuxuan Du, Yuhan Yang, Dacheng Tao, and Min-Hsiu Hsieh. Problem-dependent power of quantum neural networks on multiclass classification. *Physical Review Letters*, 131:140601, 2023.
- [4] Marie Kempkes, Aroosa Ijaz, Elies Gil-Fuster, Carlos Bravo-Prieto, Jakob Spiegelberg, Evert van Nieuwenburg, and Vedran Dunjko. Double descent in quantum kernel methods. *PRX Quantum*, 2026.
- [5] Kensuke Kamisoyama et al. Double descent in quantum kernel ridge regression. *arXiv preprint*, 2026.
- [6] Maria Schuld, Ryan Sweke, and Johannes Jakob Meyer. Effect of data encoding on the expressive power of variational quantum-machine-learning models. *Physical Review A*, 103:032430, 2021.
- [7] Erfan Abedi, Salman Beigi, and Leila Taghavi. Quantum lazy training. *Quantum*, 7:989, 2023.
- [8] Junyu Liu, Khadijeh Najafi, Kunal Sharma, Francesco Tacchino, Liang Jiang, and Antonio Mezzacapo. Representation learning via quantum neural tangent kernels. *PRX Quantum*, 3:030323, 2022.
- [9] Francesco Scala, Christa Zoufal, Dario Gerace, and Francesco Tacchino. Towards practical quantum neural network diagnostics with neural tangent kernels. *npj Quantum Information*, 2026.
- [10] Marco Cerezo, Martin Larocca, Diego García-Martín, Nelson L Diaz, Paolo Braccia, Enrico Fontana, Manuel S Rudolph, Pablo Bermejo, Aroosa Ijaz, Supanut Thanasilp, et al. Does provable absence of barren plateaus imply classical simulability? *Nature Communications*, 16(1):7907, 2025.