

Double Descent in Quantum Extreme Learning Machines

Sabri Meyer¹, Ege Yilmaz^{1,2}, Aurelien Lucchi¹, Francesco Tacchino², Francesco Scala¹

¹*Department of Mathematics and Computer Science University of Basel (Switzerland)*

²*IBM Research – Zurich, 8803 Rüschlikon, Switzerland*

Abstract

Understanding generalization in quantum machine learning, in particular the phenomenon of double descent, is a crucial open problem, especially as these models move into overparameterized regimes. Yet for quantum extreme learning machines (QELMs), the precise relationship between architectural choices, physical parameters, and the resulting shape of the risk curve remains poorly understood. We address this gap using random matrix theory to establish general bounds expressed through the population covariance, which we then study analytically by working in the Fourier representation of QELM feature maps. This approach reveals that stronger input encoding, additional input reuploading, and finite-shot sampling noise all contribute to improved generalization, and it lets us pinpoint the conditions under which a double descent risk curve persists.

Introduction. A central challenge in modern machine learning is providing a precise, quantitative analysis of the risk curve within the double descent paradigm [1]–[6]. Understanding this behaviour is critical, as it directly tells us how to tune model architectures for enhanced generalization. The existence of double descent has recently been established for quantum kernel methods [7], [8]. In QELMs, this phenomenon was observed in studies on shot noise robustness [9], but a general theory explaining what happens to the risk curve when altering key architectural parameters in these models remains unknown. We fill this gap by deriving analytical bounds for linear regression generalization in terms of the population covariance condition number. Leveraging the analysis with QELM feature maps, we bound this condition number as a function of input strength, the number of reuploads, and finite sampling shot noise. Notably, we demonstrate that finite sampling shot noise re-establishes a standard double descent curve with an unshifted interpolation threshold. This work reveals the first theoretically grounded design principles for double descent and generalization with QELM feature maps. The results are supported by numerical experiments.

Double Descent in Models with Correlated Features. We consider supervised learning over a dataset $\mathcal{S} = \{(x_\ell, y_\ell)\}_{\ell=1}^n$ drawn i.i.d. from an unknown distribution $\mathcal{D}$ on $\mathcal{X} \times \mathcal{Y} \subseteq \mathbb{R}^p \times \mathbb{R}$. Given a feature map $\phi : \mathcal{X} \rightarrow \mathbb{R}^d$, we learn a linear predictor $f_w(x) = w^\top \phi(x)$ by minimizing the empirical squared loss $\mathcal{L}_n(w) := \frac{1}{2n} \|\Phi^\top w - Y\|_2^2$ where $\Phi = [\phi(x_1), \dots, \phi(x_n)] \in \mathbb{R}^{d \times n}$ is the feature matrix and $Y \in \mathbb{R}^n$ is the label vector. We assume a generative model such that there exists $w_* \in \mathbb{R}^d$ satisfying $y = \Phi^\top w_* + \mathcal{E}$ for some Gaussian noise $\mathcal{E} \sim \mathcal{N}(0, \sigma^2)$. The generalization performance is measured by the expected risk $\mathcal{R}(\hat{w}) := \mathbb{E}_{(x,y) \sim \mathcal{D}}[(\hat{w}^\top \phi(x) - y)^2]$, where $\hat{w} := \Phi^+ Y$ denotes the min-norm interpolator with the Moore-Penrose pseudo inverse Φ^+ of Φ , the unique minimum ℓ_2 -norm solution to $\min_w \mathcal{L}_n(w)$. The asymptotic risk $\mathcal{R}(\gamma) := \lim_{d,n \rightarrow \infty, d/n \rightarrow \gamma} \mathcal{R}(\hat{w})$ of related models exhibits the *double descent* phenomenon as a function of the overparameterization ratio $\gamma = d/n$ [1]–[4].

While standard double descent theory typically assumes independent structures, isotropic features, or rely on highly non-trivial Stieltjes transform solutions assuming large signal-to-noise ratios [1], [3], [5], [6], [10], a key novelty of our work lies in extending these guarantees by bounding random feature correlations, and by conditioning a non-singular population covariance $\Sigma := \mathbb{E}_x[\phi(x)\phi(x)^\top]$. In the under-parametrized regime ($\gamma < 1$), we show that the classic risk divergence $\mathcal{R}(\gamma) = \sigma^2(1 - \gamma)^{-1}$ is preserved at the interpolation threshold ($\gamma = 1$) if the feature correlations remain limited. If these correlations are too excessive, the exact behaviour at the interpolation threshold becomes unpredictable. Regarding the risk decay in the over-parametrized regime ($\gamma > 1$), we establish the generalization bound,

$$\mathcal{R}(\gamma) = \sigma^2 + \mathcal{O}\left(\|w_*\|^2 \lambda_{\max}(\Sigma) + \frac{\sigma^2 \kappa(\Sigma)}{(\sqrt{\gamma} - C_K)^2}\right) \quad \text{as } \gamma \rightarrow \infty, \quad (1)$$

in terms of the condition number $\kappa(\Sigma) = \lambda_{\max}(\Sigma)/\lambda_{\min}(\Sigma)$ and a constant $C_K > 0$ depending on the sub-Gaussian norm [11], [12] of the whitened features $z := \Sigma^{-\frac{1}{2}}\phi(x)$. The bound holds true if $\|z\| = \sqrt{d}$ almost surely, a geometric restriction causing mild feature correlations, justified when applied to quantum extreme learning machine (QELM) feature maps discussed below.

Application to QELMs. In the context of QELMs, we particularly focus on a quantum feature map $\phi_U(x)$ defined by $\phi_U(x)_k \equiv \langle O_k \rangle_x := \text{Tr}(O_k \rho_U(x))$. Here, $\rho_U(x) := U(x)\rho_0 U(x)^\dagger$ denotes the pre-measurement state for some initial density matrix ρ_0 , where $U(x) := U_{L+1} \prod_{\ell=1}^L \{\mathcal{G}_\ell(x) U_\ell\}$ specifies the quantum architecture defined by alternating $L \in \mathbb{N}$ layers of unitary operators U_ℓ and parametrized quantum circuits $\mathcal{G}_\ell(x)$ encoding the input x . $\{O_k\}_{k \in [d]}$ is a set of Hermitian measurement operators. To apply our previously discussed theoretical framework to such quantum models, we introduce the quantum version of the population covariance, $\Sigma_U := \mathbb{E}_x [\phi_U(x) \phi_U(x)^\top]$. In the special case where $U(x)$ is sampled from a unitary 2-design [13], we have that $\Sigma_U = \Theta(2^{-N})I_d$ is isotropic with an exponentially decaying variance in the number $N \in \mathbb{N}$ of qubits [14]. Due to the absence of feature correlations in this ideal setting, a direct application of the classical Marchenko-Pastur [10] law establishes provable double descent with a clean interpolation threshold at $\gamma = 1$, equivalently shown in the classical literature [1]. Unfortunately, achieving this perfect curve comes at the cost of requiring exponential resources in the number of qubits in order to maintain a feasible signal-to-noise ratio.

For practical applications, one must thus consider non-2-design quantum architectures that generally lead to non-trivial correlations, and thus, non-isotropic Σ_U . As mentioned before, our analytic justifications allow the direct application of Eq. (1) for $\gamma > 1$, and $\mathcal{R}(\gamma) = \sigma^2(1 - \gamma)^{-1}$ for $\gamma < 1$ provided that the underlying correlations remain limited. Focusing on the over-parametrized regime, we provide an in-depth analysis of the condition number $\kappa(\Sigma_U)$ as a function of input strength $s > 0$, defined as the reparametrization $x \mapsto sx$, and as a function of increased reuploads $\Delta L \in \mathbb{N}$ due to $L \mapsto L + \Delta L$. This analysis is performed by expanding the quantum models through a Fourier perspective [15]–[17], $\phi_U(sx) = \sum_{\omega \in \Omega} c_\omega e^{i\omega \cdot (sx)}$, for explicitly determined Fourier coefficients $c_\omega \in \mathbb{C}^d$ which depend exclusively on the layer-wise architecture and are independent of s . The geometry of the Fourier frequency space, including the scaling behavior of its cardinality, depends on the input encoding strategies [16]. We also demonstrate that the maximal Fourier frequency $\max_{\omega \in \Omega} \|\omega\|$ plays a particularly important role in benign over-fitting [5], [6]. Under mild but technical assumptions, the Fourier analysis of the quantum features further allows us to rigorously establish analytic bounds for the condition number, as $s, \Delta L \rightarrow \infty$,

$$\kappa(\Sigma_U^{(L+\Delta L)}(s)) \leq \kappa\left(\sum_{\omega \in \Omega} c_\omega c_\omega^\dagger\right) + \mathcal{O}_s(s^{-K}) = \frac{\mathcal{O}_{\Delta L}(1)}{\lambda_{\min}(\Sigma_U^{(L)}) + R(\Delta L) - \mathcal{O}_s(s^{-K})}, \quad (2)$$

of the population covariance $\Sigma_U^{(L+\Delta L)}(s)$ with ΔL additional reuploads and input strength $s > 0$. Here, $R(\Delta L) > 0$ defines a regularization term, quantifying how much spectral energy the newly added, high Fourier frequencies inject into the near-null space of the base reference covariance $\Sigma_U^{(L)}$ being independent of s . The error term $\mathcal{O}_s(s^{-K})$ defines a polynomial decay with a degree $K \in \mathbb{N}$ depending on the regularity properties of the input probability density. In the limit $s \rightarrow 0$, on the other hand, we have an exploding condition number $\kappa(\Sigma_U) = \Omega(s^{-2})$ regardless of the model architecture, reflecting the collapse of the feature space into a highly redundant, low-rank subspace as the input strength vanishes. In combination with Eq. (1), we thus see that generalization performance is improved when using QELM features with high input strengths and a large number of reuploads, where the specific design of the encoding generators determines the geometry of the Fourier frequency space while the layer-wise ansatz directly influences the structure of the underlying Fourier coefficients.

Finite Shot Noise limits Quantum Feature Correlations. In practice, one relies on finite measurement shots where the true readout $\phi_U(x)$ is approximated with an estimator of the general form $\hat{\phi}_U(x) = \phi_U(x) + \epsilon(x)$ for some centred random vector $\epsilon(x)$ due to $m_0 \in \mathbb{N}$ independent, random measurement outcomes. This introduces yet another, noisy version of the population covariance $\hat{\Sigma}_U := \mathbb{E}_{x, \text{shot}}[\hat{\phi}_U(x) \hat{\phi}_U(x)^\top]$ with respect to the product probability measure $\mathbb{P}_x \times \mathbb{P}_{\text{shot}}$. Calculations

in the underlying product probability space yields the bound

$$\kappa(\hat{\Sigma}_U) \leq \frac{\lambda_{\max}(\Sigma_U) + \sigma_{\max}^2/m_0}{\lambda_{\min}(\Sigma_U) + \sigma_{\min}^2/m_0}, \quad (3)$$

where $\sigma_{\max/\min}^2$ is the maximum/minimum of the input averaged quantum variances $\sigma_k^2 := \mathbb{E}_x[\langle O_k^2 \rangle_x - \langle O_k \rangle_x^2]$ among $k \in [d]$. This demonstrates that finite shot noise acts as a regularizer of an ill conditioned, noiseless population covariance Σ_U . An additional, provable result of finite shot noise is that it further limits correlations among the noiseless features such that an interpolation threshold at $\gamma = 1$ is guaranteed if $m_0 = o(\sqrt{d})$ is bounded as $d \rightarrow \infty$.

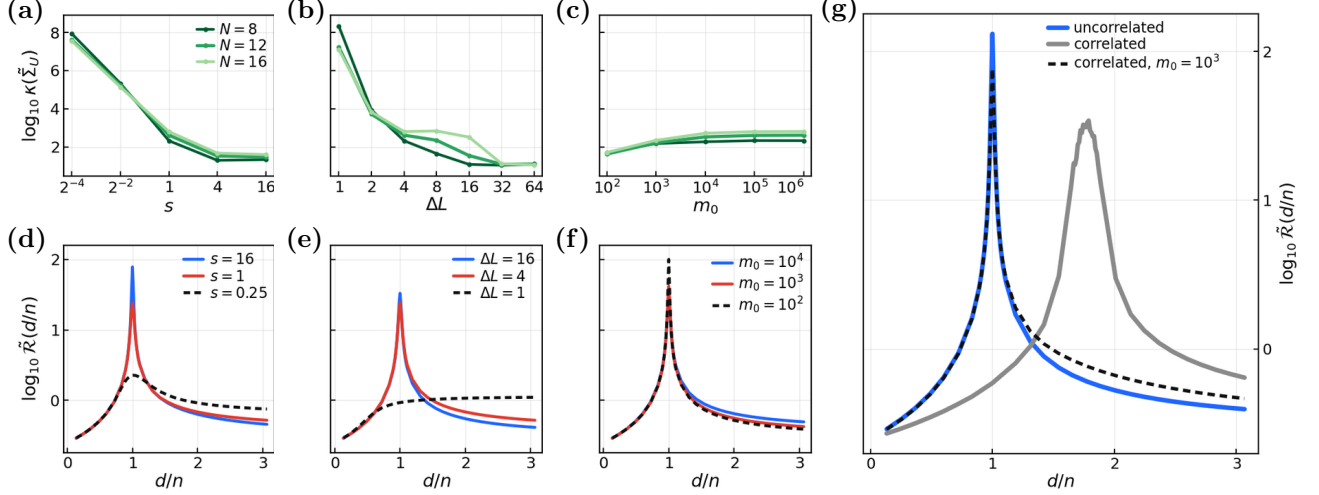


Figure 1: The effect of increasing input strength s , reuploads ΔL , and total shot count m_0 on the condition number $\kappa(\tilde{\Sigma}_U)$, varying in the number N of qubits (a)–(c), and the risk curve $\tilde{\mathcal{R}}(d/n)$ (d)–(f). Panel (g) compares the $\tilde{\mathcal{R}}(d/n)$ between correlating and non-correlating quantum architectures, and how finite shot noise re-establishes the interpolation threshold at $d/n = 1$.

Numerical Experiments. Fig. 1 confirms the theoretical results experimentally for the empirical risk $\tilde{\mathcal{R}}(d/n) := \frac{1}{n_{\text{test}}} \sum_{j=1}^{n_{\text{test}}} (\hat{w}^\top \phi_U(x_j) - y)^2$ and sample covariance $\tilde{\Sigma}_U := \frac{1}{n_{\text{test}}} \sum_{j=1}^{n_{\text{test}}} \phi_U(x_j) \phi_U(x_j)^\top$, respectively. We use the non-2-design QELM model in Ref. [18] to create features with limited correlations. As predicted by Eq. (2), we see in (a)–(b) how $\kappa(\tilde{\Sigma}_U)$ converges to lower values as $s \rightarrow \infty$ and ΔL increases. A similar effect is observed with finite shot noise in panel (c) when limiting the total shot count m_0 in Eq. (3). Panels (d)–(f) show the effect on the empirical risk curve $\tilde{\mathcal{R}}(d/n)$, respectively. The emergence of double descent with an interpolation threshold at $d/n = 1$ indicates that random feature correlations remain bounded. In contrast, when using the chaotic correlator model from Ref. [19], we observe an unpredictable shift of the interpolation threshold in (g), which is re-established at $d/n = 1$ when bounding m_0 . The default setup is $s = 1$, $\Delta L = 4$, $m_0 = \infty$ and $N = 8$, unless specified otherwise in the plot. We use $n = 90$ training samples and $n_{\text{test}} = 1000$ test samples with scalar inputs $x \sim \text{Unif}([-1, 1])$, and labels generated by a random linear teacher applied to the QELM features with additive Gaussian noise $\mathcal{E} \sim \mathcal{N}(0, 0.5^2)$. For each plotted feature count, we sample different subsets from $d = 276$ weight-1 and weight-2 Pauli observables and averaged their test errors into one curve. Specifically, each plotted point uses 3200 random feature subsets and for each subset we average over 800 independent noisy training-label realizations. Finally, we average those 3200 subset means into the plotted value.

Conclusion. We have demonstrated analytically and empirically how changing input strength, input reuploads and finite shot noise affect the condition number of the covariance and directly impact the generalization of linear regression models with QELM feature maps. We also showed how finite shot noise helps maintaining a perfect double descent curve. Our results establish a theoretically grounded bridge between classical double descent literature and quantum extreme learning machines, paving the way for practical applications of quantum feature maps in machine learning models.

References

- [1] E. Dobriban and S. Wager, “High-dimensional asymptotics of prediction: Ridge regression and classification,” *The Annals of Statistics*, vol. 46, no. 1, pp. 247–279, 2018.
- [2] M. Belkin, D. Hsu, S. Ma, et al., “Reconciling modern machine-learning practice and the classical bias–variance trade-off,” *Proceedings of the National Academy of Sciences*, vol. 116, no. 32, pp. 15 849–15 854, 2019.
- [3] S. Mei and A. Montanari, “The generalization error of random features regression: Precise asymptotics and the double descent curve,” *Communications on Pure and Applied Mathematics*, vol. 75, no. 4, pp. 667–766, 2022.
- [4] T. Hastie, A. Montanari, S. Rosset, et al., “Surprises in high-dimensional ridgeless least squares interpolation,” *Annals of statistics*, vol. 50, no. 2, p. 949, 2022.
- [5] P. L. Bartlett, P. M. Long, G. Lugosi, et al., “Benign overfitting in linear regression,” *Proceedings of the National Academy of Sciences*, vol. 117, no. 48, pp. 30 063–30 070, 2020.
- [6] A. Tsigler and P. L. Bartlett, “Benign overfitting in ridge regression,” *Journal of Machine Learning Research*, vol. 24, no. 123, pp. 1–76, 2023.
- [7] M. Kempkes, A. Ijaz, E. Gil-Fuster, et al., “Double descent in quantum kernel methods,” *PRX Quantum*, vol. 7, no. 1, p. 010 312, 2026.
- [8] K. Kamisoyama, L. Nagano, and K. Terashi, “Double descent in quantum kernel ridge regression,” *arXiv preprint arXiv:2604.17202*, 2026.
- [9] F. Hu and X. Gao, “Generalization error in quantum machine learning in the presence of sampling noise,” *arXiv preprint arXiv:2410.14654*, 2024.
- [10] Z. Bai, J. W. Silverstein, et al., *Spectral analysis of large dimensional random matrices*. Springer, 2010, vol. 20.
- [11] R. Vershynin, “High-dimensional probability,” *University of California, Irvine*, vol. 10, no. 11, p. 31, 2020.
- [12] R. Vershynin, *Introduction to the non-asymptotic analysis of random matrices*. 2012.
- [13] A. A. Mele, “Introduction to haar measure tools in quantum information: A beginner’s tutorial,” *Quantum*, vol. 8, p. 1340, 2024.
- [14] D. García-Martín, M. Larocca, and M. Cerezo, “Quantum neural networks form gaussian processes,” *Nature Physics*, vol. 21, no. 7, pp. 1153–1159, 2025.
- [15] M. Schuld, R. Sweke, and J. J. Meyer, “Effect of data encoding on the expressive power of variational quantum-machine-learning models,” *Physical Review A*, vol. 103, no. 3, p. 032 430, 2021.
- [16] H. Mhiri, L. Monbroussou, M. Herrero-Gonzalez, et al., “Constrained and vanishing expressivity of quantum fourier models,” *Quantum*, vol. 9, p. 1847, 2025.
- [17] W. Xiong, G. Facelli, M. Sahebi, et al., “On fundamental aspects of quantum extreme learning machines,” *Quantum Machine Intelligence*, vol. 7, no. 1, p. 20, 2025.
- [18] T. Dao, E. Yilmaz, I. Shehzad, et al., “Breaking concentration barriers for quantum extreme learning on digital quantum processors,” *arXiv preprint arXiv:2603.13005*, 2026.
- [19] L. E. Fischer, M. Leahy, A. Eddins, et al., “Dynamical simulations of many-body quantum chaos on a quantum computer,” *Nature Physics*, pp. 1–6, 2026.