

Cautious optimism for quantum kernels [1] and deep parametrized quantum circuits [2]

Marie Kempkes,^{1,2} Aroosa Ijaz,^{3,4} Elies Gil-Fuster,^{5,6} Carlos Bravo-Prieto,⁵ Alissa Wilms,^{5,7} Jakob Spiegelberg,² Jens Eisert,⁵ Evert van Nieuwenburg,¹ and Vedran Dunjko¹

¹*Leiden University, Niels Bohrweg 1, 2333 CA Leiden, Netherlands*

²*Volkswagen Group Innovation, Berliner Ring 2, 38440 Wolfsburg, Germany*

³*Department of Physics and Astronomy, University of Waterloo, ON N2L 3G1, Canada*

⁴*Vector Institute, Toronto, ON M5G 0C6, Canada*

⁵*Dahlem Center for Complex Quantum Systems, Freie Universität Berlin, 14195 Berlin, Germany*

⁶*Fraunhofer Heinrich Hertz Institute, 10587 Berlin, Germany*

⁷*Porsche Digital GmbH, 71636 Ludwigsburg, Germany*

Traditional statistical learning theory suggests that increasing model complexity should eventually harm generalization: models with too many parameters may memorize the training data and perform poorly on unseen samples. Modern machine learning has challenged this view. Large neural networks often display *double descent*: the test error first follows a U-shaped curve, peaks near the interpolation threshold, and then decreases again in the overparameterized regime. This phenomenon has become a central explanation for why classical deep learning models can perform well at scale. Whether analogous behavior occurs in quantum machine learning has remained largely open. This raises a natural question: can double descent appear in QML models? We distinguish between quantum kernels (implicit models) and usual re-uploading PQCs (explicit models). Previous work suggested that PQCs may instead show a traditional U-shaped risk curve [3]. Here, we show the opposite.

Results. In this joint submission, we show that both quantum kernel methods [1] and gradient-trained PQCs [2] can exhibit double descent. In both approaches, we start from a carefully-chosen decomposition for the test error. We then use random matrix theory to show that the expected error explodes with high probability when the number of training data matches the number of parameters of the model, under reasonable assumptions on the data. We perform a range of numerical experiments on both real-life and synthetic learning tasks, which match our theoretical predictions: the test loss as a function of the number of parameters increases in the *underparametrized regime* (fewer parameters than training data), achieves its maximum at interpolation (number of parameters $\approx$ number of training data), and (for gradient-trained PQCs) decreases again as the circuit is further overparameterized (more parameters than training data).

Main ideas. In our first work [1], the notion of error we study is the *excess risk* between a given model and the best possible linear model for the same task, as proposed in Ref. [4]. We first derive the expected behavior of this excess error under the assumption that input data is Gaussian distributed. We conclude by showing that the concentration properties of Gaussian data are preserved under smooth maps, which we combine with prior arguments to justify why this is enough to capture real-life learning tasks. Our numerical experiments match our predictions, and we isolate the several contributing factors via an ablation study. In our second work [2], we study the usual expected risk of a trained re-uploading model, and we show that it peaks at interpolation, under reasonable assumptions on the training phase. The starting point, inspired by the classical milestone work of Ref. [5], relates the expected risk to a variant of the training error called the add-one-in empirical risk, plus a complexity term. Since explicit re-uploading PQCs are less structured than quantum kernels, and in particular they give rise to non-convex optimization landscapes, the proof involves more-elaborate spectral techniques, though we require weaker assumptions. Our results then hold with high probability over a broader class of input data distributions, not only Gaussian.

Potential impact. We adopt a cautiously optimistic perspective on the performance of QML at scale. We are aware that currently-proposed QML models suffer from various problems associated to concentration of measure. What our work says is that there can be good generalization for overparametrized models, including deep re-uploading PQCs. What our work does not say is how to design better QML models that avoid other problems. In this sense, one could think that the promise of deep learning *could* be realized within QML, provided we find advantageous design principles. For quantum kernels, departing from the overparametrized regime may require an exponential-sized training set. For (trainable) re-uploading PQCs, increasing the number of circuit parameters beyond interpolation need not degrade generalization and can be compatible with improved test performance.

I. DOUBLE DESCENT IN QUANTUM KERNEL METHODS [1]

In Quantum Kernel Methods (QKMs), we consider a quantum kernel function $\kappa_\rho(x, x') = \text{Tr}\{\rho(x)\rho(x')\}$, given by a quantum feature map $x \mapsto \rho(x)$. Given a training set $S = \{(x_i, y_i)\}_{i=1}^N$, we can solve binary classification and regression tasks using functions of the form $f_\alpha(x) = \sum_{i=1}^N \alpha_i \kappa_\rho(x_i, x)$. The main step in the training of such a model is to find optimal choices of α given the training data. The mathematical formalism of training kernel methods is well understood, and it follows the usual language of linear regression on feature spaces [4]. After training is finish, we consider a previously-unseen new input x , for which our trained model assigns output y_α . We denote by y_* the output that the best-possible linear model would produce (think of $\alpha^* = \arg \min_\alpha \text{MSE}(\alpha)$ for the mean-squared error loss).

We first show that the difference between y_α and y_* admits a closed-form formula: $y_\alpha - y_* = \text{Tr}\{\rho(x)V\Sigma^+U^\dagger E\} + A_{\text{over}}$, where V, Σ , and U arise from a singular value decomposition; E is related to the mistakes made by the best linear fit α_* , and A_{over} is an additive term related to the overparametrized regime. We re-write the first summand as a sum $\text{Tr}\{\rho(x)V\Sigma^+U^\dagger E\} = \sum_{r=1}^R \frac{1}{\sigma_r} \text{Tr}\{V_r \rho(x)\} \langle u_r, E \rangle$, where R is the rank of Σ^+ . Written this way, it is clear that this term will explode if the smallest non-zero singular value σ_R becomes very small (hence $1/\sigma_R$ becomes very large). The remaining ingredient is then to show that the smallest non-zero singular value indeed becomes very small. (To be fully precise, we also need to confirm that the other terms remain under control. We test this numerically in the full version of the text [1]).

Theorem 1 (Double descent in QKMs – informal) Consider a Lipschitz continuous quantum feature map ρ with p linearly independent dimensions, and consider a linear regression problem in the corresponding feature space. Let $\{(\rho(x_i), y_i)\}_{i=1}^N$ be a training set, where $x_i \sim \mathcal{N}(0, \mathbb{I}_d)$ are d -dimensional i.i.d. Gaussian normal random samples. Then, the test error of the linear model based on $\rho(x)$ peaks with high probability at $N = p$, for large enough N and p . Furthermore, for a quantum feature map on n qubits, the training set size N at which the peak in test error occurs fulfills $N \in \mathcal{O}(\exp(n))$ with high probability.

Proof (informal): For the reciprocal singular values to be large, the corresponding non-zero singular values must be small. The generalized Marčenko-Pastur law from Ref. [6] implies that for a Lipschitz-continuous feature map applied to Gaussian normal random vectors, the spectral behavior of the sample covariance matrix is the same as that of the sample covariance matrix constructed from a data matrix of Gaussian random entries. As $N, p \rightarrow \infty$ at some constant rate, the MP law predicts that the smallest non-zero eigenvalues are smallest when $N = p$. Consequently, the test error is maximized with high probability at $N = p$, as $N, p \rightarrow \infty$. Additionally, the number of linearly independent dimensions of any n -qubit quantum feature map cannot exceed the linear dimension of the set of 2^n -dimensional Hermitian matrices, which is 4^n . $\square$

We validate our theory by providing empirical evidence of the double descent phenomenon in QML. Our experiments employ usual quantum kernel methods, which are formally equivalent to the linear-regression formalism presented above. We use three different datasets, including the standard Fashion MNIST and California Housing. Implementation details can be found in the public code repository [7]. Fig. 1 shows the test error in different scenarios, all of which exhibit a clear double descent peak at interpolation $N/p = 1$. These findings provide compelling empirical evidence for the presence of double descent in QKMs.

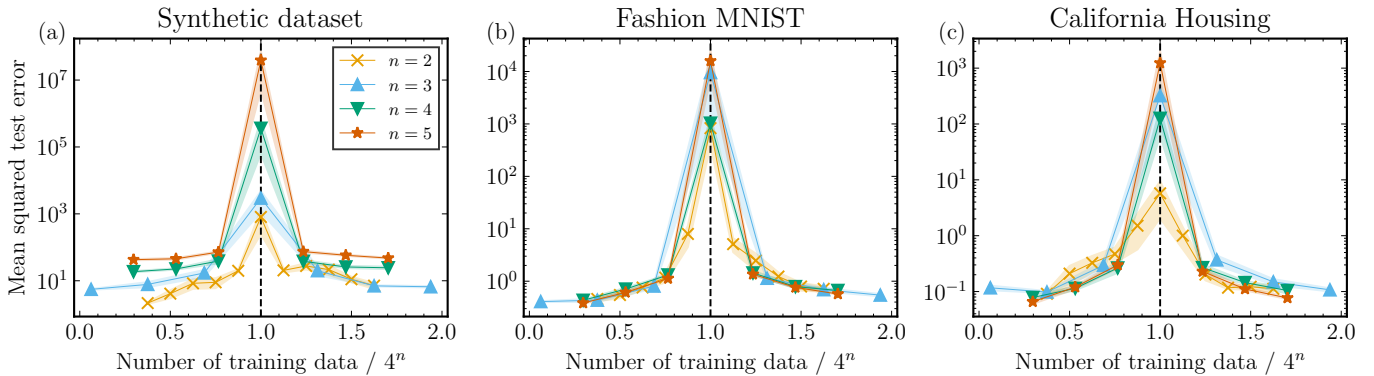


Figure 1. Empirical evidence of double descent in QKMs. Mean squared test error as a function of the normalized number of training data points ($N/4^n$) for the (a) Synthetic, (b) Fashion MNIST, and (c) California Housing datasets. The shaded area corresponds to the standard deviation for five independent experiment repetitions, each using independently sampled training data. The dotted black line indicates the interpolation threshold. The region to the left of the threshold corresponds to the overparameterized regime, while the region to the right corresponds to the underparameterized regime. The consistent peak across all datasets and system sizes n confirms the presence of double descent.

II. DOUBLE DESCENT IN DEEP PQCS PQCS [2]

In data re-uploading, we consider a parametrized quantum circuit $x, \theta \mapsto \rho(x, \theta)$ with parameters $\theta \in [0, 2\pi]^p$. We consider functions of the form $f_\theta(x) = \text{Tr}\{\rho(x, \theta)\mathcal{M}\}$, where $\mathcal{M}$ is a fixed observable. Higher-dimensional inputs can be reached for instance by having different observables $\mathcal{M}_k$ for each output dimension $k \in [K]$. Given a training set $S = \{(x_i, y_i)\}_{i=1}^N$, we use the mean-squared loss $\ell((x, y), \theta) = \|f_\theta(x) - y\|^2$ and assume that gradient-based training converges to a local minimum of the empirical risk (training error). For a learning task with K -dimensional output (like multi-class classification), we distinguish between *underparametrization* $p < NK$ and *overparametrization* $p > NK$. This recovers the usual cases of $p < N$, $p > N$ for $K = 1$. This assumption separates the question of *generalization after successful training* from the separate question of whether optimization succeeds. Our starting point is an add-one-in decomposition for the expected risk at a stationary point [5]: $L(\hat{\theta}_S) = L'(S) + (N+1)^{-1}\text{Tr}[\hat{H}_S(\hat{\theta}_S)^{-1}C(\hat{\theta}_S)] + O(N^{-2})$, where $L'(S)$ is the add-one-in loss (which is closely related to the usual empirical risk [5]), $\hat{H}_S$ is the empirical Hessian, and C is the uncentered covariance of loss gradients. We refer to the trace term as the *complexity term*. It measures how strongly the learned hypothesis reacts to perturbations of the training set.

In our analysis, we separate the Hessian of the training error into several contributions, using the chain rule. To be able to make meaningful statements about the scaling of the complexity term, we introduce 4 reasonable assumptions which carefully capture the expected behavior of all the terms arising from the chain rule of the Hessian. Our assumptions follow in style those of Ref. [5] though the ones we give are less restrictive. *Assumption 1*: training converges to a stationary point (a local minimum of the empirical risk). *Assumption 2*: in the underparameterized regime $p < NK$, the largest eigenvalue of the *functional* Hessian at a stationary point decreases with the number of parameters p near interpolation. *Assumption 3*: in the overparameterized regime $p > NK$, the *matrix of function Jacobians* has rank NK . *Assumption 4*: the minimum non-zero eigenvalue of the *covariance of function Jacobians* after training is upper- and lower-bounded by the corresponding one at initialization.

Theorem 2 (Double descent in deep PQCs – informal) Consider an n -qubit quantum re-uploading model $f_\theta(x) = \text{Tr}\{\rho(x, \theta)\mathcal{M}\}$, with $x \in \mathbb{R}^d$, $y \in \mathbb{R}$, and $\theta \in [0, 2\pi]^p$. Consider a learning task with empirical mean-squared loss $\ell((x, y), \theta) = (f_\theta(x) - y)^2$. Then, under Assumptions 1 – 4, the expected risk achieves a maximum at $p = NK$ with high probability in the limit $N, p \rightarrow \infty$, with fixed ratio $N/p = \gamma$.

Minimal proof sketch: We re-write the complexity term in terms of the eigenvalues of $\hat{H}_S(\hat{\theta}_S)$ and $C(\hat{\theta}_S)$. One main contributor to this term is the smallest non-zero eigenvalue of the Hessian, which in turn becomes the largest eigenvalue of the inverse of the Hessian. Depending on whether we give the upper or lower bound, the other main contributor is either the largest or smallest eigenvalue of the covariance $C(\hat{\theta}_S)$, though we argue these are always under control. Then, following random matrix theory, we argue that the Hessian follows a generalized MP law, and hence its smallest non-zero eigenvalue is smallest in the regime specified by the theorem.

We test the prediction using finite-size PQCs trained by gradient-based optimization. The circuits use $n = 8$ qubits and a data re-uploading architecture: each input is encoded through angle embeddings, followed by trainable single-qubit rotations and nearest-neighbor controlled- Z gates. The output is obtained by measuring Pauli- X expectation values on $K = 8$ qubits.

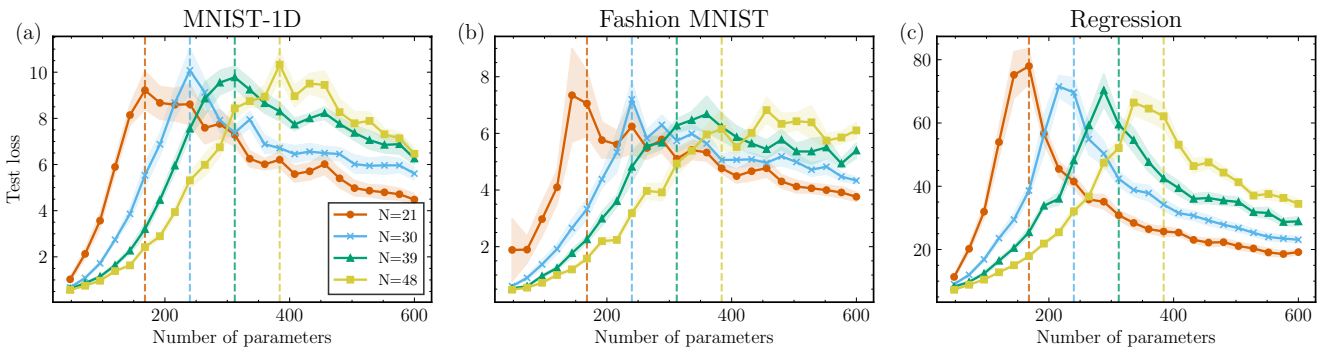


Figure 2. **Empirical evidence of double descent in PQCs.** Test loss as a function of the number of parameters p for (a) MNIST-1D and (b) Fashion MNIST classification, and (c) a multidimensional synthetic regression task, for different training set sizes N , and for $K = 8$ classes. Vertical dashed lines indicate the predicted interpolation thresholds $p = NK$. In all three tasks, the test loss peaks close to interpolation and decreases again in the overparameterized regime, displaying a characteristic double descent profile. The shaded areas correspond to the standard deviation for ten independent experiment repetitions, each using independently sampled training data.

-
- [1] M. Kempkes, A. Ijaz, E. Gil-Fuster, C. Bravo-Prieto, J. Spiegelberg, E. van Nieuwenburg, and V. Dunjko, Double descent in quantum kernel methods, [PRX Quantum](#) **7**, 010312 (2026).
 - [2] M. Kempkes, A. Ijaz, E. Gil-Fuster, C. Bravo-Prieto, A. Wilms, J. Eisert, E. van Nieuwenburg, and V. Dunjko, Cautious optimism for deep parametrized quantum circuits, In preparation (2026).
 - [3] Y. Du, Y. Yang, D. Tao, and M.-H. Hsieh, Problem-dependent power of quantum neural networks on multiclass classification, [Physical Review Letters](#) **131**, 140601 (2023).
 - [4] R. Schaeffer, M. Khona, Z. Robertson, A. Boopathy, K. Pistunova, J. W. Rocks, I. R. Fiete, and O. Koyejo, Double descent demystified: Identifying, interpreting & ablating the sources of a deep learning puzzle, [arXiv:2303.14151](#) (2023).
 - [5] S. P. Singh, A. Lucchi, T. Hofmann, and B. Schölkopf, Phenomenology of double descent in finite-width neural networks, in [International Conference on Learning Representations \(ICLR 2022\)](#) (2022).
 - [6] C. Louart and R. Couillet, Concentration of measure and large random matrices with an application to sample covariance matrices, [arXiv:1805.08295](#) (2018).
 - [7] Double descent in quantum kernel methods, [GitHub repository](#) (2025).