Quantum Neural Networks Diagnostics with Neural Tangent Kernels

Francesco Scala[1,2], Dario Gerace[1], Francesco Tacchino[2]
[1] *Dipartimento di Fisica, Università degli Studi di Pavia*
[2] *IBM Quantum, IBM Research, Zurich, 8803 Rüschlikon, Switzerland*

Abstract

In this work, we give evidence that the Quantum Neural Tangent Kernel (QNTK) can be leveraged as a diagnostic tool for Quantum Neural Network (QNN) training and generalization capabilities. We first illustrate that QNTK features can be used to approximate the behaviour of a random QNN model without having to fully train it. We show that the predictive power improves with increasing quantum circuit depth which also corresponds to more parameters. More specifically, the characteristic decay time of the training error for overparametrized QNNs can be estimated with the condition number of the QNTK. Furthermore, the QNN predictions in the infinite time limit are well reproduced by a linear kernel expression. We validate our findings with extensive numerical simulations, using different QNN architectures and datasets. Our results demonstrate that the QNTK yields a satisfying approximation of the QNN behavior for sufficiently deep circuits and motivate empirical investigations for the approximation capabilities of QNTKs for shallow QNNs.

1 Introduction

Quantum Neural Networks (QNNs) [1] consist of parameterized quantum circuits which can be used to train machine learning tasks with near term quantum processors. Various theoretical properties of QNNs have been extensively studied. However, it is still difficult to understand a priori whether a QNN model training will perform well for a given dataset or not. Inspired by classical machine learning research [2]–[4], the Quantum Neural Tangent Kernel (QNTK) [5]–[11] was recently discovered as a convenient tool to study the approximate behaviour of QNNs, in the so-called *lazy training* regime [12]. This corresponds to a setting where the parameters do not change significantly during the training and, hence, remain close to their initial values [6], [9], [12]. Under these conditions, which are realized for example when the model is overparameterized [6], the QNTK can be employed to prove an exponential decay of the residual training error given deep, random [5], [6], or symmetric [7] QNNs.

Given data and labels $\{\mathbf{x}_i, y_i\}_{i=1}^M$, we define the QNN model output as the expectation value of a Hermitian observable O with respect to a parameterized quantum state $|\psi(\mathbf{x}_i, \boldsymbol{\theta})\rangle$:

$$f(\mathbf{x}_i, \boldsymbol{\theta}) = \langle \psi(\mathbf{x}_i, \boldsymbol{\theta}) | O \, | \psi(\mathbf{x}_i, \boldsymbol{\theta}) \rangle \, . \tag{1}$$

For a QNN with a sufficient number of parameters approximating a t-design, the training behaviour of the model becomes amenable to analytical treatment: all the model derivatives with respect to the trainable parameters vanish after the second order and the QNN effectively behaves as a linear model [6]. To train the model, one can define, e.g., a quadratic loss function of the form

$$\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{2} \sum_{i=1}^{M} (f(\mathbf{x}_i, \boldsymbol{\theta}) - y_i)^2 = \frac{1}{2} \sum_i \varepsilon_i^2 \, , \tag{2}$$

where ε denotes the residual training error. Using a gradient descent optimizer with learning rate $\eta \ll 1$, we can compute the change in the residual step-wise training error as

$$\delta \varepsilon_i = -\eta \sum_{i'} K_{ii'} \varepsilon_{i'}, \quad \text{where} \quad K_{ii'} = \sum_l \frac{\partial \varepsilon_i}{\partial \theta_l} \frac{\partial \varepsilon_{i'}}{\partial \theta_l} \tag{3}$$

is the QNTK associated to the quantum model f. The latter is a symmetric positive semidefinite matrix of dimension $M \times M$.

2 QNNs diagnostics with the QNTK

Theory. In analogy to classical machine learning literature [13], we use the QNTK, evaluated at the initialization stage, to predict both training and generalization performances of a QNN before actually training it. It is possible to show that in the lazy training regime, the expected model output at the training step t for a *training* or *test* data point $\mathbf{x}_j$ can be approximated via the QNTK (at initialization) as

$$\mathbb{E}_\theta[f(\mathbf{x}_j, \boldsymbol{\theta}_t)] = \sum_{i,i',\gamma,\gamma'} K_{j,i}[K^{-1}]_{i,i'} \left[\left[\mathbb{1} - e^{-\eta K t} \right]_{\gamma\gamma'} y_{\gamma'} \right], \tag{4}$$

where the indices i, i', γ, γ' denote data points from the *training* dataset. From this equation, we can derive two results. Firstly, the training dynamics considering all training data points $\mathbf{x}_j$ is approximated via $\frac{1}{M}\sum_j \mathbb{E}_\theta[f(\mathbf{x}_j, \boldsymbol{\theta}_t)]$. In this average K and K^{-1} cancel each other out, due to the summation over j. One can, now, diagonalize $e^{-\eta K t}$ and impose a critical learning rate via $\eta_{crit}\lambda_{max} \sim \mathcal{O}(1)$ for the maximal QNTK eigenvalue λ_{max}. In this setting, the residual training error decays exponentially with a decay time $\tau \sim \kappa$ given by the QNTK condition number $\kappa = {}^{\lambda_{min}}/_{\lambda_{max}}$, where λ_{min} denotes the smallest eigenvalue of the QNTK. Furthermore, the QNTK inference formula can be obtained for the infinite time limit as

$$\mathbb{E}_\theta[f(\mathbf{x}_j, \boldsymbol{\theta}_{t\to\infty})] = \sum_{i,i'} K_{j,i}[K^{-1}]_{i,i'} y_{i'}. \tag{5}$$

Numerical results. We verify our predictions by numerically showing that κ can be employed to predict how fast an overparameterized QNN loss function converges towards a global minimum. We can also demonstrate that the QNTK is powerful in approximating the inference behaviour on unseen data. Our study is conducted on multiple datasets and different QNN layouts to showcase that the effect is neither ansatz nor dataset dependent. Specifically, the QNN architectures vary in both the data-encoding part – for which we are able to engineer both low and high data-encoding frequency spectra ω [14] – and the trainable component, for which we choose hardware efficient (HEA) and Hamiltonian Variational (HVA) ansatzes. The choice of different encoding strategies is particularly important, since it allows us to show that the inference approximation is accurate even if the QNN is not able to generalize properly. We analyze both regression and classification tasks on synthetic datasets employing limited-memory BFGS as optimizer [15] and Qiskit [16] statevector simulator. As a representative example, we report the results obtained for a regression tasks on an approximated transverse magnetization $\langle \sigma_X \rangle$ as a function of the transverse field for a 1D quantum Ising model (TFIM) with periodic boundary conditions ($N = 6$) $H = -J_Z \sum_{i=1}^{N} \sigma_Z^i \sigma_Z^{i+1} - h_X \sum_{i=1}^{N} \sigma_X^i$ in Fig. 1. The groundstates of H are obtained with a variational quantum eigensolver (VQE). The TFIM dataset comprises of 20 points with $-5 \leq h_X/J_Z \leq 4.5$, randomly split in training and test set with a 50% ratio.

The results are collected in Fig. 1, where the top panels illustrate the QNN training behaviour and the bottom ones represent the QNTK predictions. Fig. 1a shows the final mean squared error (MSE) loss of the training process and we notice that all models achieve residual errors lower than 10^{-4} starting from 10 layers on. This optimal performance is reached with decreasing number of training epochs for increasing number of layers, see Fig. 1b. In particular, high frequency models are the fastest to reach convergence with lowest error. Then, in Fig. 1c, we verify that the relative change in the parameters drops as we increase the number of layers. For the QNTK, we first plot the condition number κ, which is representative ot the typical decay time, for our different models. In Fig. 1d, we see that κ decreases as we increase the depth of the model and this is reflected in the trend of training loss value / number of required training epochs. In particular, the QNNs with high frequency encoding have a smaller condition number, which enables the predicti prediction of fast convergence and suggests that overfitting may even be taking place. An early stopping approach – informed by our QNTK-based diagnostics – could therefore be beneficial when dealing with models of this kind.

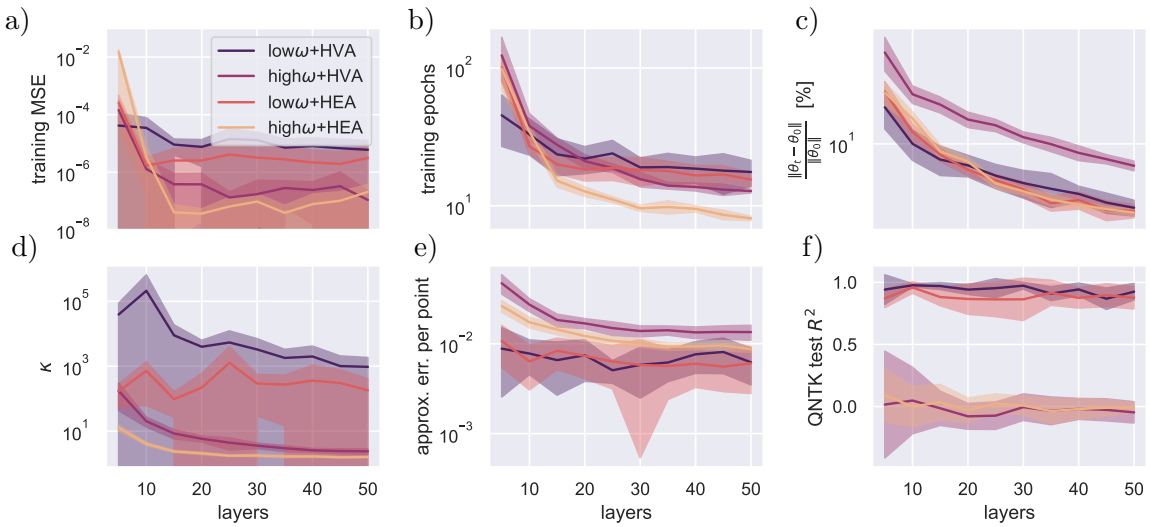

Figure 1: Summary of QNN and QNTK performances on the magnetization in a 1D Ising model with external transverse field (6 qubit chain). **a)** R^2 training score **b)** Training epochs **c)** Relative parameter change during training **d)** Condition number. **e)** Approximation error per point **f)** R^2 test score of QNTK. Each plot displays the average over 10 different initial points as a solid line, while shades represent 1 standard deviation.

This is in line with suggestions presented in the pioneering work of Jacot et al. [2]. The overfitting conjecture is further confirmed in Fig. 1f which represents the fitting performance of the QNTK on the test set by means of the R^2 score [17]. Indeed, low frequency QNNs are able to fit the test data almost perfectly, while high frequency models perform worse. This analysis on the test set can be performed by using either the fully trained QNN model or the QNTK prediction via Eq. (5) since the approximation error per point is negligible, see Fig. 1e. Interestingly enough, the approximation error is almost constant for different depths in the case of low frequency models. This may suggest that these models are correctly approximated even in the case of shallow QNNs with few qubits.

It is worth mentioning that we also investigated datasets such as noisy sinusoidal functions for regression and the moon dataset from `Scikit-learn` [18] for classificationn. The respective results are in line with the TFIM experiments presented above.

3 Conclusions

Inspired by the use of NTKs in classical ML, we showed that the QNTK can be a useful diagnostic tool for gaining insights into the behaviour of QNN training dynamics and generalization capabilities. Our approach may be of particular value for models of increasing size, both in terms of number of qubits, circuit depth and number of data points. Indeed, computing the QNTK at initialization only requires one gradient calculation per training and test point, while training a QNN would require all gradients at each training epoch. Moreover, our work can pave the way for investigating the use of QNTKs as a diagnostic tool for shallow QNNs with few qubits and low frequency data encoding spectrum. We aim to investigate these aspects in near-term future research.

References

[1] S. Mangini, F. Tacchino, D. Gerace, *et al.*, "Quantum computing models for artificial neural networks," *Europhysics Letters*, vol. 134, no. 1, p. 10 002, 2021.

[2] A. Jacot, F. Gabriel, and C. Hongler, "Neural tangent kernel: Convergence and generalization in neural networks," in *Advances in Neural Information Processing Systems*, S. Bengio, H. Wallach, H. Larochelle, *et al.*, Eds., vol. 31, Curran Associates, Inc., 2018.

[3] J. Lee, L. Xiao, S. Schoenholz, *et al.*, "Wide neural networks of any depth evolve as linear models under gradient descent," in *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, *et al.*, Eds., vol. 32, Curran Associates, Inc., 2019.

[4] D. A. Roberts, S. Yaida, and B. Hanin, *The Principles of Deep Learning Theory: An Effective Theory Approach to Understanding Neural Networks*. Cambridge University Press, May 2022.

[5] J. Liu, F. Tacchino, J. R. Glick, *et al.*, "Representation learning via quantum neural tangent kernels," *PRX Quantum*, vol. 3, p. 030 323, 3 2022.

[6] J. Liu, K. Najafi, K. Sharma, *et al.*, "Analytic theory for the dynamics of wide quantum neural networks," *Phys. Rev. Lett.*, vol. 130, p. 150 601, 15 2023.

[7] X. Wang, J. Liu, T. Liu, *et al.*, *Symmetric pruning in quantum neural networks*, 2023. arXiv: 2208.14057 [quant-ph].

[8] B. Zhang, J. Liu, X.-C. Wu, *et al.*, *Dynamical phase transition in quantum neural networks with large depth*, 2023. arXiv: 2311.18144 [quant-ph].

[9] N. Shirai, K. Kubo, K. Mitarai, *et al.*, *Quantum tangent kernel*, 2022. arXiv: 2111.02951 [quant-ph].

[10] M. Incudini, M. Grossi, A. Mandarino, *et al.*, "The quantum path kernel: A generalized neural tangent kernel for deep quantum machine learning," *IEEE Transactions on Quantum Engineering*, vol. 4, pp. 1–16, 2023.

[11] Y. Wang and J. Liu, *A comprehensive review of quantum machine learning: From nisq to fault tolerance*, 2024. arXiv: 2401.11351 [quant-ph].

[12] E. Abedi, S. Beigi, and L. Taghavi, "Quantum Lazy Training," *Quantum*, vol. 7, p. 989, Apr. 2023.

[13] L. Xiao, J. Pennington, and S. Schoenholz, "Disentangling trainability and generalization in deep neural networks," in *Proceedings of the 37th International Conference on Machine Learning*, H. D. III and A. Singh, Eds., ser. Proceedings of Machine Learning Research, vol. 119, PMLR, 2020, pp. 10 462–10 472.

[14] M. Schuld, R. Sweke, and J. J. Meyer, "Effect of data encoding on the expressive power of variational quantum-machine-learning models," *Phys. Rev. A*, vol. 103, p. 032 430, 3 2021.

[15] D. C. Liu and J. Nocedal, "On the limited memory bfgs method for large scale optimization.," *Math. Program.*, vol. 45, no. 1-3, pp. 503–528, 1989.

[16] A. Javadi-Abhari, M. Treinish, K. Krsulich, *et al.*, *Quantum computing with Qiskit*, 2024. arXiv: 2405.08810 [quant-ph].

[17] S. A. Glantz and B. K. Slinker, *Primer of applied regression and analysis of variance*, en, 2nd ed. New York, NY: McGraw-Hill Professional, Jan. 1991.

[18] F. Pedregosa, G. Varoquaux, A. Gramfort, *et al.*, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.