

Efficient classical computation of the neural tangent kernel of quantum neural networks

Anderson Melchor Hernandez¹, Davide Pastorello², and Giacomo De Palma¹

¹Dipartimento di Matematica, Università di Bologna, Piazza di Porta San Donato 5, 40126, Bologna (Italy)

²Dipartimento di Matematica, Università di Bologna, Piazza di Porta San Donato 5, 40126, Bologna (Italy)
TIFPA-INFN, via Sommarive 14, 38123 Povo (Trento), Italy

We propose an efficient classical algorithm to estimate the Neural Tangent Kernel (NTK) associated with a broad class of quantum neural networks. These networks consist of arbitrary unitary operators belonging to the Clifford group interleaved with parametric gates given by the time evolution generated by an arbitrary Hamiltonian belonging to the Pauli group. The proposed algorithm leverages a key insight: the average over the distribution of initialization parameters in the NTK definition can be exactly replaced by an average over just four discrete values, chosen such that the corresponding parametric gates are Clifford operations. This reduction enables an efficient classical simulation of the circuit. Combined with recent results establishing the equivalence between wide quantum neural networks and Gaussian processes [Girardi *et al.*, Comm. Math. Phys. 406, 92 (2025); Melchor Hernandez *et al.*, Ann. Henri Poincaré (2025)], our method enables efficient computation of the expected output of wide, trained quantum neural networks, and therefore shows that such networks cannot achieve quantum advantage.

Contents

1	Introduction	2
1.1	Our contribution	3
1.2	Outline	5
2	Preliminaries	5
2.1	Training data	5
2.2	The Pauli group and the Clifford group	5
2.3	Assumptions on the architecture	6
2.4	The neural tangent kernel	6
3	Our Results	8
3.1	An efficient classical algorithm to estimate the analytic NTK	8
3.2	Sample complexity	10
4	Proofs	11

Anderson Melchor Hernandez: anderson.melchor@unibo.it

Davide Pastorello: davide.pastorello3@unibo.it

Giacomo De Palma: giacomo.depalma@unibo.it

4.1	Preliminary results	11
4.2	Proof of Theorem 3	13
4.2.1	Estimation of the number of samples	13
4.2.2	Estimation of the number of elementary operations	13
4.3	Proof of Theorem 4	16
4.3.1	Estimation of the number of samples	16
4.3.2	Estimation of the number of elementary operations	22
5	Conclusions	23
A	Bernstein’s upper bound for random matrices	24

1 Introduction

Quantum Machine Learning (QML) is an interdisciplinary field that merges the principles of quantum computing with classical machine learning techniques [9, 21, 25]. One of the core ideas of QML is to harness quantum algorithms and the unique properties of quantum mechanics such as superposition, entanglement, and quantum parallelism to enhance the performance of deep neural models [6]. Quantum neural networks constitute the quantum version of deep neural models. These new models are based on quantum circuits and generate functions given by the expectation values of a quantum observable measured on the output of a quantum circuit made by parametric gates [11, 24]. These parameters are typically optimized by gradient descent, which involves iterative adjustment to minimize a cost function and improve the performance of the quantum circuit in the processing and analysis of data [26]. Several works have focused on the analysis of quantum neural networks, as it is believed that they can combine the computational power of quantum computers with the capabilities of deep learning algorithms [17]. Ref. [16] shows that an exponential quantum speed-up can be obtained via the use of a quantum-enhanced feature space, where each data point is mapped in a non-linear way to a quantum state, and then classified by a linear classifier in a high-dimensional Hilbert space [13]. Nevertheless, a significant disadvantage lies in the need to determine the appropriate parameters to configure the quantum circuit beforehand, and it is not yet clear whether these parameters can be effectively obtained using a variational technique [8].

A rigorous mathematical characterization of the training dynamics of quantum neural networks becomes possible in the limit of infinite width. Ref. [2] proves that in the limit of infinite width with the depth kept fixed, trained quantum neural networks with constant depth operate in the lazy regime (*i.e.*, the maximum amount by which the training can change a parameter tends to zero) and are capable of perfectly fitting the training examples. However, such a regime is not interesting for quantum advantage, since at finite depth, the past light-cone of each measured qubit has a finite size and the expectation value of the measured observable can be efficiently estimated on a classical computer. The recent works [11, 20] consider the joint limit of infinite width and depth. Under the hypothesis that the depth grows at most logarithmically with respect to the number of qubits, they prove that the probability distribution of the trained model function converges in distribution to a Gaussian process. The key element of the proof is showing that the training happens in the lazy regime, and therefore the dependence of the model function on the parameters can be approximated by its linearized version near the initialization values. Consequently, in the limit the model becomes linear and the training has an analytic solution whose probability distribution is Gaussian with analytically computable mean and covariance. Such a linear model is governed by a kernel called the Neural Tangent Kernel (NTK), which has been investigated in several recent studies [15, 23, 25, 27, 32].

Let us recall that, given a general parametric model $f_\theta : \mathcal{X} \rightarrow \mathbb{R}$, with $\theta \in \mathbb{R}^L$, where $\mathcal{X}$ is the feature space, the *empirical* NTK is defined as [14]:

$$\hat{K}_\theta(x, x') = \nabla_\theta f_\theta(x)^T \nabla_\theta f_\theta(x') \quad x, x' \in \mathcal{X}, \quad (1.1)$$

and is the central object to describe the training dynamics of the model generated by the time evolution of the parameters under gradient flow. The *analytic* NTK is defined by the expectation value of the empirical NTK with respect to the random sampling of the parameters at initialization:

$$K(x, x') = \mathbb{E}_\theta \hat{K}_\theta(x, x'), \quad (1.2)$$

and governs the training dynamics of the model in the limit of infinite width.

1.1 Our contribution

In this work, we provide a classical algorithm to efficiently estimate the NTK for a very wide class of quantum neural networks and present quantitative results on both the sample and computational complexity of the proposed algorithm to show the efficiency of the estimation. We then leverage the proposed algorithm to efficiently estimate the average of the model functions of very wide trained quantum neural networks. We consider a finite set $\mathcal{X}$ of possible inputs and a model function defined by

$$f_\theta(x) = \langle 0^n | U_{x,\theta}^\dagger O U_{x,\theta} | 0^n \rangle, \quad (1.3)$$

where n denotes the number of qubits, $|0^n\rangle = |0\rangle^{\otimes n}$, O is the measured n -qubit observable, and $U_{x,\theta}$ is a parametric quantum circuit made by nonparametric unitary operations belonging to the Clifford group, which can depend on the input x in an arbitrary way, interleaved with parametric gates given by the time evolution generated by a Hamiltonian belonging to the Pauli group (see (2.3)), so that each entry of the parameter vector is a rotation angle in $[0, 2\pi)$.

We provide an efficient classical algorithm to estimate the NTK of any such a model function. Our estimator is unbiased, meaning that its expected value coincides with the true kernel. Our first contribution can be stated informally in the following way:

Theorem 1 (Informal statement of Theorem 3). *Let $\mathcal{X}$ be the feature space and consider the model function defined in (1.3). Let each parameter be initialized independently from the uniform distribution on $[0, 2\pi)$, and let us assume that $\mathbb{E}_\theta f_\theta(x) = 0$ for any $x \in \mathcal{X}$. Then, for any $x, x' \in \mathcal{X}$ and any $\epsilon, \delta > 0$ there is a classical probabilistic algorithm that estimates the number $K(x, x')$ with precision ϵ and probability at least $1 - \delta$ with*

$$N = \frac{8L^2 m^2}{3\epsilon^2} \log \frac{2}{\delta} \quad (1.4)$$

iid samples of θ , where L denotes the number of parameters of the model, and where the measured observable O can be expressed as a linear combination of m tensor products of Pauli matrices. The algorithm requires $O(N L^2 m n^2)$ classical operations.

The key idea of the algorithm is that, since $f_\theta(x)$ is a quadratic polynomial in the entries of the unitary matrix associated to each gate and the analytic NTK is quadratic in f , the expectation value with respect to the uniform distribution of the parameters can be replaced by the expectation value with respect to the uniform distribution on the set $\left\{0, \frac{\pi}{2}, \pi, \frac{3\pi}{2}\right\}$. For such values of the parameters, the parametric gates belong to the Clifford group, and since also the nonparametric gates belong to the Clifford group by hypothesis, the whole $U_{x,\theta}$ belongs to

the Clifford group. Therefore, the analytic NTK can be computed via the efficient algorithm for Clifford quantum circuits [1].

Let us now state informally our second contribution. The average of the Gaussian process associated to a trained quantum neural network has a limit $\mu_\infty : \mathcal{X} \rightarrow \mathbb{R}$ for infinite training time. Such a limit depends only on the NTK of the network and on the training data (see subsection 2.4 and [11, 20] for the details). We can now leverage the algorithm of Theorem 1 to estimate μ_∞ :

Theorem 2 (Informal statement of Theorem 4). *Within the same hypotheses of Theorem 1, the following hold true:*

- (a) *For any $x \in \mathcal{X}$ and any $\epsilon, \delta > 0$ there is a classical probabilistic algorithm that, with probability at least $1 - \delta$, outputs an estimate $\tilde{\mu}_\infty(x)$ of the average $\mu_\infty(x)$ of the model function in x in the limit of infinite training time such that*

$$|\tilde{\mu}_\infty(x) - \mu_\infty(x)| < \epsilon \quad (1.5)$$

with

$$N = \frac{24R^2 + 4R\epsilon}{3\epsilon^2} \log \frac{2(1 + d_{\text{train}})}{\delta} + \frac{2(1 + \sqrt{2})^4 L^4 d_{\text{train}}^3 m^8 \|K_{\text{train}}^{-1}\|_{\text{op}}^4 \|Y\|_2^2}{3\epsilon^2} \log \frac{4d_{\text{train}}}{\delta} \quad (1.6)$$

iid samples of θ , where

$$R = 2L\sqrt{d_{\text{train}}m^2} \|K_{\text{train}}^{-1}Y\|_2, \quad (1.7)$$

$$(1.8)$$

d_{train} is the number of training examples, Y is the vector of the training labels and K_{train} is the NTK restricted to the training inputs. The algorithm requires

$$O\left(N L^2 d_{\text{train}} \left(m n^2 + d_{\text{train}}^2\right)\right) \quad (1.9)$$

elementary operations.

- (b) *The output $\tilde{\mu}_\infty$ of the algorithm above satisfies*

$$|\tilde{\mu}_\infty(x) - \mu_\infty(x)| < \epsilon \quad \forall x \in \mathcal{X} \quad (1.10)$$

(i.e., the estimate of μ_∞ has a uniformly bounded error on the whole $\mathcal{X}$) with probability at least $1 - \delta$ with

$$N = \frac{24R^2 + 4R\epsilon}{3\epsilon^2} \log \frac{2|\mathcal{X}|(1 + d_{\text{train}})}{\delta} + \frac{2(1 + \sqrt{2})^4 L^4 d_{\text{train}}^3 m^8 \|K_{\text{train}}^{-1}\|_{\text{op}}^4 \|Y\|_2^2}{3\epsilon^2} \log \frac{4|\mathcal{X}|d_{\text{train}}}{\delta} \quad (1.11)$$

iid samples of θ .

Remark 1. (1.11) coincides with (1.6) upon replacing δ with $\frac{\delta}{|\mathcal{X}|}$. A number of samples growing logarithmically with the size of the feature space is enough to get an estimate of μ_∞ with a uniformly bounded error on the whole $\mathcal{X}$.

Remark 2. [Theorem 2](#) states that sample and time complexities related to the classical algorithm for estimating $\mu_\infty(x)$ are polynomial. However, we do not believe that our exponents are optimal. Our result needs a time complexity scaling as d_{train}^6 (see [\(1.6\)](#) and [\(1.9\)](#)). For realistic values of d_{train} (e.g., the MNIST dataset of handwritten digits includes about 10^5 examples), such a scaling would be prohibitive. We are confident that our upper bound to the time complexity is pessimistic and that numerical experiments, which we plan to address in a future work, may reveal a better scaling.

Our work conceptually aligns with a recent series of results [\[3, 7, 18, 22, 33\]](#) demonstrating that a wide class of variational quantum circuits can be simulated efficiently by classical computers. In particular, Ref. [\[3\]](#) shows that, if each layer of the circuit contains random iid one-qubit gates sampled from the Haar distribution applied to all the qubits, then, regardless of how the two-qubit gates are placed, the untrained circuit can be simulated efficiently with high probability. Our results imply that wide quantum neural networks with logarithmic depth trained on supervised-learning problems with the input encoded with Clifford gates can be simulated efficiently by classical computers.

1.2 Outline

The organization of the present work is as follows. In [section 2](#), we introduce the structure of the considered parametric quantum circuit, the model function and the NTK. In [section 3](#), we present our algorithm to estimate the NTK and μ_∞ and state our results about the sample and computational complexity of the algorithm. In [section 4](#), we prove our results. Finally, in [section 5](#), we present our conclusions and outline some open questions. In [Appendix A](#), we present Bernstein’s inequality for rectangular matrices, which we will employ to prove concentration bounds for our estimate of the NTK.

2 Preliminaries

Let us start by introducing the notation of the present work.

2.1 Training data

Let $\mathcal{X}$ be the feature space, *i.e.*, the set of all the possible inputs, which we assume to be finite, and we let $\mathbb{R}$ be the output space.

2.2 The Pauli group and the Clifford group

Let $\mathbb{C}^2$ be the Hilbert space of a single qubit. In what follows, we denote by $n \in \mathbb{N}$ the number of qubits of the quantum neural network. Hence, the Hilbert space of the system is $(\mathbb{C}^2)^{\otimes n}$, with dimension 2^n . Let us now briefly recall the definition of the n -qubit Pauli group and of the n -qubit Clifford group. Let

$$\sigma_0 := \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad \sigma_1 := \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \quad \sigma_2 := \begin{bmatrix} 0 & -i \\ i & 0 \end{bmatrix}, \quad \sigma_3 := \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \quad (2.1)$$

be the Pauli matrices.

Definition 1 (*n -qubit Pauli group*). The n -qubit Pauli group $\mathbf{P}_n$ is made by the unitary operators acting on $(\mathbb{C}^2)^{\otimes n}$ of the form $\lambda \sigma_{i_1} \otimes \cdots \otimes \sigma_{i_n}$ where $\lambda \in \{\pm 1, \pm i\}$, and $i_1, \dots, i_n \in \{0, \dots, 3\}$.

Definition 2. The n -qubit Clifford group $\mathbf{C}_n$ is given by all the unitary operators U acting on $(\mathbb{C}^2)^{\otimes n}$ such that

$$U\mathbf{P}_n U^\dagger \subseteq \mathbf{P}_n. \quad (2.2)$$

That is, the Clifford group $\mathbf{C}_n$ is the normalizer of the n -qubit Pauli group.

For details about properties of these groups, see for instance [19].

2.3 Assumptions on the architecture

For any $x \in \mathcal{X}$ and any $\theta \in [0, 2\pi)^L$, we consider quantum circuits of the form

$$U_{x,\theta} = U_x^{(L)} e^{-i\frac{\theta_L}{2} P_L} \dots U_x^{(1)} e^{-i\frac{\theta_1}{2} P_1} U_x^{(0)}, \quad (2.3)$$

where $U_x^{(0)}, \dots, U_x^{(L)} \in \mathbf{C}_n$ can arbitrarily depend on the input x , and $P_1, \dots, P_L \in \mathbf{P}_n$ are self-adjoint. Let us clarify that the requirement of encoding the input via Clifford gates is not a severe restriction. In particular, Clifford gates allow the encoding of bit strings into the corresponding vectors of the computational basis (basis encoding) or, more generally, into any stabilizer state.

We let the measured observable O be a real linear combination of at most m self-adjoint elements of the n -qubit Pauli group:

$$O = \sum_{k=1}^m c_k P_k, \quad (2.4)$$

where $P_k = P_k^\dagger \in \mathbf{P}_n$ and $c_k \in [-1, 1]$.

The model function is then

$$f_\theta(x) = \langle 0^n | U_{x,\theta}^\dagger O U_{x,\theta} | 0^n \rangle, \quad (2.5)$$

where $|0^n\rangle = |0\rangle^{\otimes n}$.

2.4 The neural tangent kernel

Given a training set

$$\mathcal{D} = \left\{ (x^{(i)}, y^{(i)}) \right\}_{i=1, \dots, d_{\text{train}}} \subseteq \mathcal{X} \times \mathbb{R}, \quad (2.6)$$

we will call $d_{\text{train}} = |\mathcal{D}|$ the number of examples and we will represent it in a vectorized form as follows

$$X_{\text{train}} := \begin{pmatrix} x^{(1)} \\ x^{(2)} \\ \vdots \\ x^{(d_{\text{train}})} \end{pmatrix}, \quad Y = \begin{pmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(d_{\text{train}})} \end{pmatrix}. \quad (2.7)$$

In what follows, we assume the following:

Assumption 1. The parameter vector θ is initialized by independently sampling each entry from the uniform distribution on $[0, 2\pi)$.

Here, we note that the parameters are optimized via gradient descent with respect to the quadratic cost function $\mathcal{L}$ as follows:

$$\frac{d\theta_t}{dt} = -\eta \nabla_{\theta} \mathcal{L}(\theta_t), \quad (2.8)$$

where $\mathcal{L}$ is given by

$$\mathcal{L}(\theta) := \frac{1}{2} \sum_{i=1}^{d_{\text{train}}} \left(f_{\theta}(x^{(i)}) - y^{(i)} \right)^2, \quad (2.9)$$

and $\eta > 0$ is the learning rate. Given any function $g : \mathcal{X} \rightarrow \mathbb{R}$, we will often use the following notation:

$$g(X_{\text{train}}) := \begin{pmatrix} g(x^{(1)}) \\ g(x^{(2)}) \\ \vdots \\ g(x^{(d_{\text{train}})}) \end{pmatrix}, \quad g(X_{\text{train}}^T) := \begin{pmatrix} g(x^{(1)}) & g(x^{(2)}) & \cdots & g(x^{(d_{\text{train}})}) \end{pmatrix}. \quad (2.10)$$

Similarly, for any bivariate function $K : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ we will write $K(X_{\text{train}}, X_{\text{train}}^T)$ to indicate the $d_{\text{train}} \times d_{\text{train}}$ matrix with entries $\left(K(X_{\text{train}}, X_{\text{train}}^T) \right)_{ij} := K(x^{(i)}, x^{(j)})$ for $1 \leq i, j \leq d_{\text{train}}$. We set

$$F_{\theta, \text{train}} := \begin{pmatrix} f_{\theta}(x^{(1)}) \\ f_{\theta}(x^{(2)}) \\ \vdots \\ f_{\theta}(x^{(d_{\text{train}})}) \end{pmatrix} = f_{\theta}(X_{\text{train}}). \quad (2.11)$$

Definition 3 (Empirical NTK). The empirical NTK is given by the inner product between the gradients of the model function evaluated for different inputs:

$$\hat{K}_{\theta}(x, x') := (\nabla_{\theta} f_{\theta}(x))^T \nabla_{\theta} f_{\theta}(x'), \quad x, x' \in \mathcal{X}. \quad (2.12)$$

In what follows, we introduce some useful assumptions to treat the behavior of the NTK. Before, let us recall the definition of the analytic NTK.

Definition 4 (Analytic NTK). The analytic NTK is the expectation of the empirical NTK with respect to the probability distribution of the parameters at initialization:

$$K(x, x') := \mathbb{E}_{\theta} \hat{K}_{\theta}(x, x'), \quad x, x' \in \mathcal{X}. \quad (2.13)$$

Assumption 2. We suppose that the analytic NTK restricted to the training inputs

$$K_{\text{train}} = K(X_{\text{train}}, X_{\text{train}}^T) \quad (2.14)$$

is invertible. We denote with $\lambda_{\text{max}}^{K_{\text{train}}}$ and $\lambda_{\text{min}}^{K_{\text{train}}}$ its maximum and minimum eigenvalue, respectively.

From the gradient-flow equation (2.8) and the chain rule, one gets that

$$\begin{cases} \frac{d\theta_t}{dt} = -\eta \nabla_{\theta} f_{\theta_t}(X_{\text{train}}^T) (F_{\theta_t, \text{train}} - Y), \\ \frac{d}{dt} f_{\theta_t}(x) = -\eta (\nabla_{\theta} f_{\theta_t}(x))^T \nabla_{\theta} f_{\theta_t}(X_{\text{train}}^T) (F_{\theta_t, \text{train}} - Y), \end{cases} \quad (2.15)$$

where $\nabla_{\theta} f_{\theta_t}(X_{\text{train}}^T)$ denotes the gradient of $f_{\theta_t}(X_{\text{train}}^T)$ with respect to θ . Recall that T is the transposition operator. Let us notice that (2.15) can be written as

$$\begin{cases} \frac{d\theta_t}{dt} = -\eta \nabla_{\theta} f_{\theta_t}(X_{\text{train}}^T)(F_{\theta_t, \text{train}} - Y), \\ \frac{d}{dt} f_{\theta_t}(x) = -\eta \hat{K}_{\theta_t}(x, X_{\text{train}}^T)(F_{\theta_t, \text{train}} - Y). \end{cases} \quad (2.16)$$

If we assume that the NTK does not change during training, i.e., that $\hat{K}_{\theta_t} = \hat{K}_{\theta_0}$, then the time evolution of the model function becomes linear:

$$\frac{d}{dt} f_{\theta_t}^{\text{lin}}(x) = -\eta \hat{K}_{\theta_0}(x, X_{\text{train}}^T) e^{-\eta \hat{K}_{\theta_0} t} (F_{\theta_0, \text{train}} - Y) \quad (2.17)$$

and has the analytic solution

$$f_{\theta_t}^{\text{lin}}(x) = f_{\theta_0}(x) - \hat{K}_{\theta_0}(x, X_{\text{train}}^T) \hat{K}_{\theta_0}^{-1} \left(\mathbb{1} - e^{-\eta \hat{K}_{\theta_0} t} \right) (F_{\theta_0, \text{train}} - Y). \quad (2.18)$$

Ref. [11] proves that this assumption holds: in the limit of infinite width with the depth growing at most logarithmically with the width, if covariance of f at initialization and the analytic NTK have finite limits $\mathcal{K}_0$ and K , respectively, then the empirical NTK stays close to its value at initialization throughout the whole training. Moreover, the value of the NTK at initialization is with high probability close to its average with respect to the random initialization of the parameters [11, Theorem 4.7]. As a consequence, the probability distribution of the trained model function $\{f_{\theta_t}(x)\}_{x \in \mathcal{X}}$ converges in distribution to the Gaussian process with mean and covariance given by [11, Corollary 4.9]:

$$\begin{aligned} \mathcal{K}_t(x, x') &:= \mathcal{K}_0(x, x') - K(x, X_{\text{train}}^T) K_{\text{train}}^{-1} \left(\mathbb{1} - e^{-t\eta K_{\text{train}}} \right) \mathcal{K}_0(X_{\text{train}}, x') \\ &\quad - K(x', X_{\text{train}}^T) K_{\text{train}}^{-1} \left(\mathbb{1} - e^{-t\eta K_{\text{train}}} \right) \mathcal{K}_0(X_{\text{train}}, x) + \\ &\quad + K(x, X_{\text{train}}^T) K_{\text{train}}^{-1} \left(\mathbb{1} - e^{-t\eta K_{\text{train}}} \right) \mathcal{K}_0(X_{\text{train}}, X_{\text{train}}^T) \left(\mathbb{1} - e^{-t\eta K_{\text{train}}} \right) K_{\text{train}}^{-1} K(X_{\text{train}}, x'), \end{aligned} \quad (2.19)$$

$$\mu_t(x) := K(x, X_{\text{train}}^T) K_{\text{train}}^{-1} \left(\mathbb{1} - e^{-t\eta K_{\text{train}}} \right) Y. \quad (2.20)$$

We stress that both $\mathcal{K}_t(x, x')$ and $\mu_t(x)$ have a limit for $t \rightarrow \infty$. In particular, $\mu_t(x)$ converges to

$$\mu_{\infty}(x) = K(x, X_{\text{train}}^T) K_{\text{train}}^{-1} Y. \quad (2.21)$$

Ref. [20] provides a quantitative version of the convergence that holds at finite width. The results of Ref. [20] hold uniformly with respect to the training time, and therefore prove that the limits of infinite width and infinite training time commute, and therefore the average of the model function trained for infinite time actually converges to (2.21) in the limit of infinite width.

3 Our Results

3.1 An efficient classical algorithm to estimate the analytic NTK

In the present work, we propose an efficient classical algorithm to estimate the analytic NTK as well as μ_{∞} to within any specified precision $\epsilon > 0$ with high probability. To that end, we introduce an estimator for the analytic NTK. Let $\mathcal{S} := \{\theta^{(1)}, \dots, \theta^{(N)}\}$ be a collection of parameter samples, where each $\theta^{(i)}$ is drawn uniformly from the set $\{0, \frac{\pi}{2}, \pi, \frac{3\pi}{2}\}^L$. Although

in general the parameters of the circuit that we consider are generic angles in $[0, 2\pi)$, we prove below that due to the structure of the parametric unitary $U_{\theta,x}$, it is sufficient to restrict attention to $\{0, \frac{\pi}{2}, \pi, \frac{3\pi}{2}\}$.

The following lemma will allow us to compute the gradient with respect to the parameters in the definition of the NTK as an exact finite difference:

Lemma 1 (Parameter-shift rule). *Let $f_\theta(x)$ be the model function (2.5) of a parametric quantum circuit as in (2.3). Then, for any $x \in \mathcal{X}$, any $\theta \in [0, 2\pi)^L$ and any $i = 1, \dots, L$ we have*

$$\partial_{\theta_i} f_\theta(x) = \frac{1}{2} \left(f_{\theta + \frac{\pi}{2} e_i}(x) - f_{\theta - \frac{\pi}{2} e_i}(x) \right), \quad (3.1)$$

where e_i is the i -th vector of the canonical basis of $\mathbb{R}^L$.

Proof. This is just a matter of computation, and we refer to [4, 11]. $\square$

With this, we propose the following empirical estimator for the analytic NTK:

$$\begin{aligned} \tilde{K}(x, x') &= \frac{1}{N} \sum_{j=1}^N \sum_{i=1}^L \partial_{\theta_i} f_{\theta^{(j)}}(x) \partial_{\theta_i} f_{\theta^{(j)}}(x') \\ &= \frac{1}{4N} \sum_{j=1}^N \sum_{i=1}^L \left(f_{\theta^{(j)} + \frac{\pi}{2} e_i}(x) - f_{\theta^{(j)} - \frac{\pi}{2} e_i}(x) \right) \left(f_{\theta^{(j)} + \frac{\pi}{2} e_i}(x') - f_{\theta^{(j)} - \frac{\pi}{2} e_i}(x') \right), \end{aligned} \quad (3.2)$$

where we have used the parameter-shift rule to express the derivatives with respect to the parameters as finite differences. From a statistical perspective, the estimator $\tilde{K}$ possesses several desirable properties that make it a reliable tool for approximating the analytic NTK. First, it is an unbiased estimator, meaning that its expected value coincides exactly with the true kernel (see Proposition 1 below). This ensures that, on average, the estimator does not systematically overestimate or underestimate the quantity of interest [5, 12]. Moreover, because the estimator is constructed as an average over N independent and identically distributed random samples, classical concentration inequalities (such as Hoeffding's or Bernstein's inequality) can be applied to provide probabilistic guarantees on the deviation of $\tilde{K}$ from K [29, 30]. We then propose the following classical algorithm to estimate the NTK:

Algorithm 1 (Classical estimation of $K(x, x')$).

- Sample N sets of parameters $\mathcal{S} = \{\theta^{(1)}, \dots, \theta^{(N)}\}$, where each entry of each $\theta^{(i)}$ is sampled from the uniform distribution on $\{0, \frac{\pi}{2}, \pi, \frac{3\pi}{2}\}$.
- Compute $f_{\theta \pm \frac{\pi}{2} e_i}(x)$, $f_{\theta \pm \frac{\pi}{2} e_i}(x')$ for any $\theta \in \mathcal{S}$ and any $i = 1, \dots, L$.
- Return

$$\tilde{K}(x, x') = \frac{1}{4N} \sum_{j=1}^N \sum_{i=1}^L \left(f_{\theta^{(j)} + \frac{\pi}{2} e_i}(x) - f_{\theta^{(j)} - \frac{\pi}{2} e_i}(x) \right) \left(f_{\theta^{(j)} + \frac{\pi}{2} e_i}(x') - f_{\theta^{(j)} - \frac{\pi}{2} e_i}(x') \right).$$

Algorithm 1 can be employed to compute a classical estimate of the average of the trained model function:

Algorithm 2 (Classical estimation of $\mu_\infty(x)$).

- Sample N sets of parameters $\mathcal{S} = \{\theta^{(1)}, \dots, \theta^{(N)}\}$, where each entry of each $\theta^{(i)}$ is sampled from the uniform distribution on $\{0, \frac{\pi}{2}, \pi, \frac{3\pi}{2}\}$.
- Apply [Algorithm 1](#) with the above $\mathcal{S}$ to compute an estimate $\tilde{K}_{\text{train}}$ of the NTK on the training inputs and an estimate $\tilde{K}(x, X_{\text{train}}^T)$ of $K(x, X_{\text{train}}^T)$.
- Compute the inverse matrix $\tilde{K}_{\text{train}}^{-1}$.
- Return

$$\tilde{\mu}_{\infty}(x) = \tilde{K}(x, X_{\text{train}}^T) \tilde{K}_{\text{train}}^{-1} Y.$$

3.2 Sample complexity

The following [Theorem 3](#) determines the sample complexity of estimating the NTK via [Algorithm 1](#):

Theorem 3 (Formal statement of [Theorem 1](#)). *Let $\mathcal{X}$ be the feature space, and for each $x \in \mathcal{X}$ let us consider a parametric quantum circuit as defined in (2.3) with model function as in (2.5). Suppose that [Assumption 1](#) holds true. Let us assume that $\mathbb{E}_{\theta} f_{\theta}(x) = 0$ for any $x \in \mathcal{X}$. Then for any $x, x' \in \mathcal{X}$ and any $\epsilon, \delta > 0$, the output $\tilde{K}(x, x')$ of [Algorithm 1](#) with*

$$N = \frac{8L^2 m^2}{3\epsilon^2} \log \frac{2}{\delta} \quad (3.3)$$

iid samples of θ satisfies

$$\left| \tilde{K}(x, x') - K(x, x') \right| < \epsilon \quad (3.4)$$

with probability at least $1 - \delta$. The algorithm requires $O(N L^2 m n^2)$ elementary operations.

The following [Theorem 4](#) determines the sample and computational complexity of estimating μ_{∞} via [Algorithm 2](#):

Theorem 4 (Formal statement of [Theorem 2](#)). *Under the same hypotheses of [Theorem 3](#), the following hold true.*

(a) *For any $x \in \mathcal{X}$, any*

$$0 < \epsilon < \frac{L}{2} \sqrt{d_{\text{train}}} m^2 \|Y\|_2 \left\| K_{\text{train}}^{-1} \right\|_{\text{op}} \quad (3.5)$$

and any $\delta > 0$, the output $\tilde{\mu}_{\infty}(x)$ of [Algorithm 2](#) with

$$N = \frac{24R^2 + 4R\epsilon}{3\epsilon^2} \log \frac{2(1 + d_{\text{train}})}{\delta} + \frac{2(1 + \sqrt{2})^4 L^4 d_{\text{train}}^3 m^8 \left\| K_{\text{train}}^{-1} \right\|_{\text{op}}^4 \|Y\|_2^2}{3\epsilon^2} \log \frac{4d_{\text{train}}}{\delta} \quad (3.6)$$

iid samples of θ , where

$$R = 2L \sqrt{d_{\text{train}}} m^2 \left\| K_{\text{train}}^{-1} Y \right\|_2, \quad (3.7)$$

satisfies

$$|\tilde{\mu}_{\infty}(x) - \mu_{\infty}(x)| < \epsilon \quad (3.8)$$

with probability at least $1 - \delta$. The algorithm requires

$$O(N L^2 d_{\text{train}} [m n^2 + d_{\text{train}}^2]) \quad (3.9)$$

elementary operations.

(b) For any ϵ satisfying (3.5) and any $\delta > 0$, the output $\tilde{\mu}_\infty$ of *Algorithm 2* applied to each $x \in \mathcal{X}$ with the same $\mathcal{S}$ with

$$N = \frac{24R^2 + 4R\epsilon}{3\epsilon^2} \log \frac{2|\mathcal{X}|(1 + d_{\text{train}})}{\delta} + \frac{2(1 + \sqrt{2})^4 L^4 d_{\text{train}}^3 m^8 \|K_{\text{train}}^{-1}\|_{\text{op}}^4 \|Y\|_2^2}{3\epsilon^2} \log \frac{4|\mathcal{X}|d_{\text{train}}}{\delta} \quad (3.10)$$

iid samples of θ satisfies

$$|\tilde{\mu}_\infty(x) - \mu_\infty(x)| < \epsilon \quad \forall x \in \mathcal{X} \quad (3.11)$$

with probability at least $1 - \delta$.

Remark 3. From part (b) of *Theorem 4*, *Algorithm 2* can estimate μ_∞ with a uniformly bounded error on the whole $\mathcal{X}$ with a number of samples that grows logarithmically with the size of the feature space.

Remark 4. We stress that the algorithm itself does not require any assumptions on the width or depth of the network. Such assumptions are only needed to guarantee that μ_∞ is close to the mean of the trained model function.

4 Proofs

In this section, we provide the proofs of the main results.

4.1 Preliminary results

We now prove that $\tilde{K}$ is an unbiased estimator.

Proposition 1. Suppose that *Assumption 1* holds true. For any $x, x' \in \mathcal{X}$, we have

$$\mathbb{E}_{\theta^{(1)}, \dots, \theta^{(N)} \sim \text{unif}\{0, \frac{\pi}{2}, \pi, \frac{3\pi}{2}\}^L} \tilde{K}(x, x') = K(x, x'). \quad (4.1)$$

Proof. Since each $\theta^{(i)} \in \mathcal{S}$ is independent and identically distributed, then we need to prove that for any $x, x' \in \mathcal{X}$ we have

$$\mathbb{E}_{\theta \sim [0, 2\pi)^L} \left[\nabla_\theta f(\theta, x)^T \nabla_\theta f(\theta, x') \right] = \mathbb{E}_{\theta \sim \{0, \frac{\pi}{2}, \pi, \frac{3\pi}{2}\}^L} \left[\nabla_\theta f(\theta, x)^T \nabla_\theta f(\theta, x') \right]. \quad (4.2)$$

In order to prove (4.2), we need the following:

Proposition 2. Let $d \in \mathbb{N}$, and $P \in \mathbb{C}^{d \times d}$ be hermitian and such that $P^2 = I$. Let μ be a probability distribution on $\mathbb{R}$. Then, $\mathbb{E}_{\theta \sim \mu} \left(e^{i\frac{\theta}{2}P} \otimes e^{-i\frac{\theta}{2}P} \right)^{\otimes 2}$ depends on μ only through the quantities $\mathbb{E}_{\theta \sim \mu} e^{i\theta}$, $\mathbb{E}_{\theta \sim \mu} e^{2i\theta}$. In particular, the uniform distribution on $[0, 2\pi)$ and the uniform distribution on $\{0, \frac{\pi}{2}, \pi, \frac{3\pi}{2}\}$ result in the same expectation value of $\left(e^{i\frac{\theta}{2}P} \otimes e^{-i\frac{\theta}{2}P} \right)^{\otimes 2}$.

Proof. Let

$$H = \frac{P \otimes I - I \otimes P}{2}. \quad (4.3)$$

We have

$$\left(e^{i\frac{\theta}{2}P} \otimes e^{-i\frac{\theta}{2}P} \right)^{\otimes 2} = e^{i\theta(H_1 + H_2)}, \quad (4.4)$$

where each H_i acts on the i -th subsystem. P has spectrum $\{\pm 1\}$, therefore H has spectrum $\{-1, 0, 1\}$ and $H_1 + H_2$ has spectrum $\{-2, \dots, 1, 2\}$. For any $l = -2, \dots, 2$, let Π_l be the orthogonal projector onto the eigenspace of $H_1 + H_2$ with eigenvalue l . Then,

$$\mathbb{E}_{\theta \sim \mu} \left(e^{i\frac{\theta}{2}P} \otimes e^{-i\frac{\theta}{2}P} \right)^{\otimes 2} = \sum_{l=-2}^2 \left(\mathbb{E}_{\theta \sim \mu} e^{il\theta} \right) \Pi_l. \quad (4.5)$$

The claim follows since for any $l = 1, 2$

$$\mathbb{E}_{\theta \sim \mu} e^{-il\theta} = \left(\mathbb{E}_{\theta \sim \mu} e^{il\theta} \right)^*. \quad (4.6)$$

□

Let us notice that due to [Lemma 1](#), we have that

$$\hat{K}_\theta(x, x') = \frac{1}{4} \sum_{j=1}^L \left(f_{\theta + \frac{\pi}{2}e_j}(x) - f_{\theta - \frac{\pi}{2}e_j}(x) \right) \left(f_{\theta + \frac{\pi}{2}e_j}(x') - f_{\theta - \frac{\pi}{2}e_j}(x') \right). \quad (4.7)$$

Denote by $u_j = e^{-i\frac{\theta_j}{2}P_j}$, for $j = 1, \dots, L$. Notice that $\hat{K}_\theta(x, x')$ can be written as a polynomial of degree 2 in the coefficients of u_j and degree 2 in their complex conjugates. Hence, we have that $\hat{K}_\theta(x, x')$ can be written as

$$\hat{K}_\theta(x, x') = \text{tr} \left(\left(e^{-i\frac{\theta_j}{2}P_j} \otimes e^{i\frac{\theta_j}{2}P_j} \right)^{\otimes 2} J_j(\theta, x, x') \right) \quad (4.8)$$

where $J_j(\theta, x, x')$ is a suitable linear operator that does not depend on θ_j . Then by [Proposition 2](#), (4.8) has the same expected value with respect to the uniform distribution on $[0, 2\pi)^L$ and with respect to the uniform distribution on $\{0, \frac{\pi}{2}, \pi, \frac{3\pi}{2}\}^L$, from which we conclude that (4.2) holds true. Since all the variables in $\mathcal{S}$ are iid, the conclusion of [Proposition 1](#) is proved. □

In what follows, we estimate the deviation of $\tilde{K}_{\text{train}}$ from K_{train} . In the next, let us denote by $\|\cdot\|_{\text{op}}$ the operator norm.

Theorem 5. *Let d_{train} be the number of training examples. We have for any $0 < t \leq \|f\|_\infty^2$*

$$\mathbb{P} \left\{ \left\| \tilde{K}_{\text{train}} - K_{\text{train}} \right\|_{\text{op}} \geq t \right\} \leq 2 d_{\text{train}} \exp \left(- \frac{3 N t^2}{8 L^2 d_{\text{train}}^2 \|f\|_\infty^4} \right). \quad (4.9)$$

Proof. Let us choose in [Corollary 1](#)

$$\mathbf{X}_k = \frac{1}{4} \sum_{i=1}^L \left(F_{\theta^{(k)} + \frac{\pi}{2}e_i, \text{train}} - F_{\theta^{(k)} - \frac{\pi}{2}e_i, \text{train}} \right) \left(F_{\theta^{(k)} + \frac{\pi}{2}e_i, \text{train}} - F_{\theta^{(k)} - \frac{\pi}{2}e_i, \text{train}} \right)^T - K_{\text{train}}. \quad (4.10)$$

We have

$$\begin{aligned} \mathbf{X}_k &\leq \frac{1}{4} \sum_{i=1}^L \left(F_{\theta^{(k)} + \frac{\pi}{2}e_i, \text{train}} - F_{\theta^{(k)} - \frac{\pi}{2}e_i, \text{train}} \right) \left(F_{\theta^{(k)} + \frac{\pi}{2}e_i, \text{train}} - F_{\theta^{(k)} - \frac{\pi}{2}e_i, \text{train}} \right)^T \\ &\leq \frac{1}{4} \sum_{i=1}^L \left\| F_{\theta^{(k)} + \frac{\pi}{2}e_i, \text{train}} - F_{\theta^{(k)} - \frac{\pi}{2}e_i, \text{train}} \right\|^2 \leq L d_{\text{train}} \|f\|_\infty^2. \end{aligned} \quad (4.11)$$

We also have

$$\mathbf{X}_k \geq -K_{\text{train}} \geq -L d_{\text{train}} \|f\|_{\infty}^2, \quad (4.12)$$

such that

$$\|\mathbf{X}_k\| \leq L d_{\text{train}} \|f\|_{\infty}^2. \quad (4.13)$$

We get from [Corollary 1](#)

$$\mathbb{P} \left\{ \left\| \tilde{K}_{\text{train}} - K_{\text{train}} \right\|_{\text{op}} \geq t \right\} \leq 2 d_{\text{train}} \exp \left(- \frac{N t^2 / 2}{L^2 d_{\text{train}}^2 \|f\|_{\infty}^4 + L d_{\text{train}} \|f\|_{\infty}^2 t / 3} \right) \quad (4.14)$$

Since $t \leq \|f\|_{\infty}^2$, we have

$$L^2 d_{\text{train}}^2 \|f\|_{\infty}^4 + L d_{\text{train}} \|f\|_{\infty}^2 t / 3 \leq \frac{4}{3} L^2 d_{\text{train}}^2 \|f\|_{\infty}^4. \quad (4.15)$$

The claim follows. $\square$

4.2 Proof of [Theorem 3](#)

4.2.1 Estimation of the number of samples

Let $\epsilon > 0$. Notice that by [Theorem 5](#) one gets

$$\mathbb{P} \left\{ \left| \tilde{K}(x, x') - K(x, x') \right| \geq \epsilon \right\} \leq 2 \exp \left(- \frac{3 N \epsilon^2}{8 L^2 \|f\|_{\infty}^4} \right). \quad (4.16)$$

Hence, by letting

$$2 \exp \left(- \frac{3 N \epsilon^2}{8 L^2 \|f\|_{\infty}^4} \right) \leq \delta, \quad (4.17)$$

one finds that

$$N \geq \frac{8 L^2 \|f\|_{\infty}^2}{3 \epsilon^2} \log \left(\frac{2}{\delta} \right), \quad (4.18)$$

and where $\|f\|_{\infty} \leq m$.

4.2.2 Estimation of the number of elementary operations

Let us now estimate the required number of operations to compute $\tilde{K}(x, x')$. To this aim, we need the following:

Lemma 2. *Let $P \in \mathbf{P}_n$ be self-adjoint, and let $Q \in \mathbf{P}_n$. Then,*

$$e^{i \frac{\theta}{2} P} Q e^{-i \frac{\theta}{2} P} = \begin{cases} Q & [P, Q] = 0 \\ \cos \theta Q + i \sin \theta P Q & \{P, Q\} = 0 \end{cases}. \quad (4.19)$$

In particular, $e^{-i \frac{\theta}{2} P} \in \mathbf{C}_n$ for any $\theta \in \left\{ 0, \frac{\pi}{2}, \pi, \frac{3\pi}{2} \right\}$.

Proof. If P and Q commute we have

$$e^{i\frac{\theta}{2}P} Q e^{-i\frac{\theta}{2}P} = e^{i\frac{\theta}{2}P} e^{-i\frac{\theta}{2}P} Q = Q \in \mathbf{P}_n. \quad (4.20)$$

If P and Q anticommute, we have

$$e^{i\frac{\theta}{2}P} Q e^{-i\frac{\theta}{2}P} = e^{i\frac{\theta}{2}P} e^{i\frac{\theta}{2}P} Q = e^{i\theta P} Q = (\cos \theta I + i \sin \theta P) Q = \cos \theta Q + i \sin \theta P Q. \quad (4.21)$$

For $\theta = 0, \frac{\pi}{2}, \pi, \frac{3\pi}{2}$, respectively, we have

$$e^{i\frac{\theta}{2}P} Q e^{-i\frac{\theta}{2}P} = Q, i P Q, -Q, -i P Q \in \mathbf{P}_n, \quad (4.22)$$

where we have used that $\mathbf{P}_n$ is closed with respect to multiplication. $\square$

Lemma 3. For any $\theta \in \left\{0, \frac{\pi}{2}, \pi, \frac{3\pi}{2}\right\}^L$ and any $x \in \mathcal{X}$, the complexity of computing $f_\theta(x)$ is $O(L m n^2)$.

Proof. In what follows, we need to introduce some further notation. Let us denote by $\mathbb{F}_2 := \{0, 1\}$ the field with two elements. Let $a, b \in \mathbb{F}_2^{2n}$. Let

$$\sigma_{00} = \tau_{00} = \sigma_0, \quad \sigma_{01} = \tau_{01} = \sigma_1, \quad \sigma_{10} = \tau_{10} = \sigma_3, \quad \sigma_{11} = \sigma_2, \quad \tau_{11} = i\sigma_{11} = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}. \quad (4.23)$$

We use the following notation to indicate tensor products of Pauli matrices. For $a = \begin{pmatrix} v \\ w \end{pmatrix}$ where $v, w \in \mathbb{F}_2^n$, we set

$$\begin{aligned} \sigma_a &:= \sigma_{v_1} \sigma_{w_1} \otimes \cdots \otimes \sigma_{v_n} \sigma_{w_n}, \\ \tau_a &:= \tau_{v_1} \tau_{w_1} \otimes \cdots \otimes \tau_{v_n} \tau_{w_n}, \end{aligned} \quad (4.24)$$

where $\sigma_{v_i}, \sigma_{w_i} \in \mathbf{P}_1$.

Remark 5. An arbitrary element of $\mathbf{P}_n$ can be represented in a unique way by $i^\delta (-1)^\epsilon \tau_a$ where $\delta, \epsilon \in \mathbb{F}_2$ and $a \in \mathbb{F}_2^{2n}$. In the following, we will employ such a parameterization of the Pauli group.

Multiplication of two Pauli group elements can now be translated into binary terms in the following way:

Lemma 4 ([10, Lemma 1]). If $a_1, a_2 \in \mathbb{F}_2^{2n}$, $\delta_1, \delta_2, \epsilon_1, \epsilon_2 \in \mathbb{F}_2$, and τ is defined as in (4.24), then

$$i^{\delta_1} (-1)^{\epsilon_1} \tau_{a_1} i^{\delta_2} (-1)^{\epsilon_2} \tau_{a_2} = i^{\delta_{12}} (-1)^{\epsilon_{12}} \tau_{a_{12}} \quad (4.25)$$

where

$$\begin{aligned} \delta_{12} &= \delta_1 + \delta_2, \\ \epsilon_{12} &= \epsilon_1 + \epsilon_2 + \delta_1 \delta_2 + a_2^T U a_1, \\ a_{12} &= a_1 + a_2, \\ U &= \begin{bmatrix} 0_n & I_n \\ 0_n & 0_n \end{bmatrix}, \end{aligned} \quad (4.26)$$

where the multiplication, and addition of binary variables is modulo 2.

Let us notice that by definition a Clifford operation Q , maps the Pauli group $\mathbf{P}_n$ to itself under conjugation:

$$Q\tau_a Q^\dagger = i^\delta (-1)^\epsilon \tau_b \quad (4.27)$$

for some parameters δ, ϵ, b functions of a . Since

$$Q\tau_{a_1}\tau_{a_2}Q^\dagger = (Q\tau_{a_1}Q^\dagger)(Q\tau_{a_2}Q^\dagger), \quad (4.28)$$

then we only need to know the image of a generating set of the Pauli group to know the image of all Pauli group elements and define Q up to an overall phase. Hence, it is sufficient to know the image of τ_k , where $\{b_k\}_{k=1}^n$ form a basis of $\mathbb{F}_2^{2n}$. Therefore, it is sufficient to work with Hermitian Pauli group elements only since the image of a Hermitian matrix H under the map $H \mapsto QHQ^\dagger$ will be a Hermitian matrix. Hence, in the notation of the right-hand side of (4.27) Hermitian Pauli group elements are described as

$$i^{a^T U a} (-1)^\epsilon \tau_a. \quad (4.29)$$

Let us notice that $a^T U a$ modulo 2 counts the number of τ_{11} (which is the unique non Hermitian matrix of the four τ matrices) in the tensor product of τ_a . Let us denote by $(e_k)_{k=1}^{2n}$ the standard basis of $\mathbb{F}_2^{2n}$ where e_k is the k -th column of I_{2n} , and consider the hermitian generators of the Pauli group $\{\tau_{e_k}\}_{k=1}^{2n}$. These correspond to single qubit-operations σ_z , and σ_x . So that, we denote their images under $H \mapsto QHQ^\dagger$ by $i^{d_k} (-1)^{h_k} \tau_{c_k}$. We then assemble the matrix C with columns c_k , and the vectors d, h with scalars d_k, h_k , respectively. Since the images are Hermitian, then

$$d_k = c_k^T U c_k \text{ or } d = \text{diag}(C^T U C) \quad (4.30)$$

where $\text{diag}(C^T U C)$ denotes the vector of diagonal elements of $C^T U C$. Hence, given C, d , and h defining the Clifford operation Q , the image $i^{\delta_2} (-1)^{\epsilon_2} \tau_{b_2}$ of $i^{\delta_1} (-1)^{\epsilon_1} \tau_{b_1}$ under $H \mapsto QHQ^\dagger$ can be found by multiplying those operators $i^{d_k} (-1)^{h_k} \tau_{c_k}$ for which $b_{1k} = 1$. Then by Lemma 4

$$\begin{aligned} b_2 &= C b_1, \\ \delta_2 &= \delta_1 + d^T b_1 \\ \epsilon_2 &= \epsilon_1 + h^T b_1 + b_1^T (\text{ lows}(C^T U C + d d^T)) b_1 + \delta_1 d^T b_1, \end{aligned} \quad (4.31)$$

where $\text{ lows}(P)$ denotes the strictly lower triangular part of a matrix P . Let us now consider

$$P_k \mapsto U_x^{(\ell)} P_k U_x^{(\ell)\dagger} \quad \ell = 1, \dots, L, \quad (4.32)$$

where $P_k \in \mathbf{P}_n$, $k = 1, \dots, m$. Hence the images of P_k under (4.32) can be written as in (4.31). Furthermore, by the explicit formulas in (4.31), we have that the number of operations to specify the image is then $O(n^2)$ (see also [1, Section III]). Then for each $\ell = 1, \dots, L$ one gets that by Lemma 2,

$$e^{i\frac{\theta_\ell}{2} P_\ell} U_x^{(\ell)} P_k U_x^{(\ell)\dagger} e^{-i\frac{\theta_\ell}{2} P_\ell} \quad (4.33)$$

belongs to the Clifford group for any $\theta_\ell \in \left\{0, \frac{\pi}{2}, \pi, \frac{3\pi}{2}\right\}$, and applying [Lemma 4](#), its images can be computed as in (4.31) by making $O(n^2)$ elementary operations. Recall that $f_\theta(x)$ can be written as

$$f_\theta(x) = \sum_{k=1}^m c_k \langle 0^n | U_{x,\theta}^\dagger P_k U_{x,\theta} | 0^n \rangle \quad (4.34)$$

where

$$U_{x,\theta} = U_x^{(L)} e^{-i\frac{\theta}{2}P_L} \dots U_x^{(1)} e^{-i\frac{\theta}{2}P_1} U_x^{(0)}, \quad (4.35)$$

so that $U_{x,\theta} P_k U_{x,\theta}^\dagger$ gives us an element of the form $i^{\delta_2} (-1)^{\epsilon_2} \tau_{b_2}$ which can be computed as in (4.31) after $2L + 1$ compositions. Then applying [Lemma 4](#), one gets that the complexity of computing $f_\theta(x)$ is $O(Lmn^2)$. $\square$

Lemma 5. *For any $\theta \in \left\{0, \frac{\pi}{2}, \pi, \frac{3\pi}{2}\right\}^L$, and any $x \in \mathcal{X}$, the complexity of computing $\partial_{\theta_i} f_\theta(x)$ is $O(Lmn^2)$ for each $i = 1, \dots, L$.*

Proof. Notice that by [Lemma 1](#), we have that for each $i = 1, \dots, L$, $\partial_{\theta_i} f_\theta(x)$ is a linear combination of $f_{\theta + \frac{\pi}{2} e_i}(x)$, $f_{\theta - \frac{\pi}{2} e_i}(x)$. Then by [Lemma 3](#), the complexity of computing $\partial_{\theta_i} f_\theta(x)$ is then $O(Lmn^2)$. $\square$

Lemma 6. *For any $\theta \in \left\{0, \frac{\pi}{2}, \pi, \frac{3\pi}{2}\right\}^L$, and any $x, x' \in \mathcal{X}$, the complexity of computing $\hat{K}_\theta(x, x')$, and $\tilde{K}(x, x')$ is $O(L^2 mn^2)$, and $O(NL^2 mn^2)$, respectively.*

Proof. Let us notice that to estimate $\hat{K}_\theta(x, x')$ requires to perform L products of the form $\partial_{\theta_i} f_\theta(x) \partial_{\theta_i} f_\theta(x')$. Then by [Lemma 5](#), to estimate $\hat{K}(x, x')$, we need $O(L^2 mn^2)$ operations. Therefore, the complexity of computing $\tilde{K}(x, x')$ is then $O(NL^2 mn^2)$. $\square$

4.3 Proof of [Theorem 4](#)

We split the proof in two parts. In the first one, we estimate the number of samples N .

4.3.1 Estimation of the number of samples

Notice that

$$\begin{aligned} \mu_\infty(x) - \tilde{\mu}_\infty(x) &= \mu_\infty(x) - \tilde{K}(x, X_{\text{train}}^T) K_{\text{train}}^{-1} Y + \tilde{K}(x, X_{\text{train}}^T) K_{\text{train}}^{-1} Y - \tilde{\mu}(x) \\ &= \left(K(x, X_{\text{train}}^T) - \tilde{K}(x, X_{\text{train}}^T) \right) K_{\text{train}}^{-1} Y + \tilde{K}(x, X_{\text{train}}^T) \left(K_{\text{train}}^{-1} - \tilde{K}_{\text{train}}^{-1} \right) Y. \end{aligned} \quad (4.36)$$

Then for each $\epsilon > 0$ such that

$$\epsilon \leq |\mu_\infty(x) - \tilde{\mu}(x)| \leq \left| \left(K(x, X_{\text{train}}^T) - \tilde{K}(x, X_{\text{train}}^T) \right) K_{\text{train}}^{-1} Y \right| + \quad (4.37)$$

$$+ \left| \tilde{K}(x, X_{\text{train}}^T) \left(K_{\text{train}}^{-1} - \tilde{K}_{\text{train}}^{-1} \right) Y \right| \quad (4.38)$$

one gets

$$\begin{aligned} \mathbb{P}(\epsilon \leq |\mu_\infty(x) - \tilde{\mu}(x)|) &\leq \\ &\mathbb{P}\left(\epsilon/2 \leq \left| \left(K(x, X_{\text{train}}^T) - \tilde{K}(x, X_{\text{train}}^T) \right) K_{\text{train}}^{-1} Y \right| \right) + \\ &+ \mathbb{P}\left(\epsilon/2 \leq \left| \tilde{K}(x, X_{\text{train}}^T) \left(K_{\text{train}}^{-1} - \tilde{K}_{\text{train}}^{-1} \right) Y \right| \right). \end{aligned} \quad (4.39)$$

Let us now consider the first term in the right-hand side of the last identity. Let us set

$$\begin{aligned} \tilde{\mathbf{X}}_k^{x, \text{train}} &:= \frac{1}{4} \sum_{i=1}^L \left(f_{\theta^{(k)} + \frac{\pi}{2} e_i}(x) - f_{\theta^{(k)} - \frac{\pi}{2} e_i}(x) \right) \left(F_{\theta^{(k)} + \frac{\pi}{2} e_i, \text{train}} - F_{\theta^{(k)} - \frac{\pi}{2} e_i, \text{train}} \right)^T \\ &\quad - K(x, X_{\text{train}}^T), \end{aligned} \quad (4.40)$$

and also let us set

$$\mathbf{X}_k^{x, \text{train}} := \tilde{\mathbf{X}}_k^{x, \text{train}} K_{\text{train}}^{-1} Y. \quad (4.41)$$

Notice that

$$\left| \mathbf{X}_k^{x, \text{train}} \right| \leq \left\| \tilde{\mathbf{X}}_k^{x, \text{train}} \right\|_2 \left\| K_{\text{train}}^{-1} Y \right\|_2. \quad (4.42)$$

Notice that

$$\begin{aligned} \left\| \tilde{\mathbf{X}}_k^{x, \text{train}} \right\|_2 &\leq \frac{1}{4} \sum_{i=1}^L \left| f_{\theta^{(k)} + \frac{\pi}{2} e_i}(x) - f_{\theta^{(k)} - \frac{\pi}{2} e_i}(x) \right| \left\| F_{\theta^{(k)} + \frac{\pi}{2} e_i, \text{train}} - F_{\theta^{(k)} - \frac{\pi}{2} e_i, \text{train}} \right\|_2 \\ &\quad + \left\| K(x, X_{\text{train}}^T) \right\|_2, \end{aligned} \quad (4.43)$$

where

$$\left| f_{\theta^{(k)} + \frac{\pi}{2} e_i}(x) - f_{\theta^{(k)} - \frac{\pi}{2} e_i}(x) \right| \leq |f_{\theta^{(k)} + \frac{\pi}{2} e_i}(x)| + |f_{\theta^{(k)} - \frac{\pi}{2} e_i}(x)| \quad (4.44)$$

$$\leq 2m, \quad (4.45)$$

and similarly

$$\left\| F_{\theta^{(k)} + \frac{\pi}{2} e_i, \text{train}} - F_{\theta^{(k)} - \frac{\pi}{2} e_i, \text{train}} \right\|_2 \leq 2\sqrt{d_{\text{train}}}m. \quad (4.46)$$

Hence,

$$\left\| \tilde{\mathbf{X}}_k^{x, \text{train}} \right\|_2 \leq L\sqrt{d_{\text{train}}}m^2 + \left\| K(x, X_{\text{train}}^T) \right\|_2. \quad (4.47)$$

Now, by applying the previous reasoning, we have $\left\| K(x, X_{\text{train}}^T) \right\|_2 \leq L\sqrt{d_{\text{train}}}m^2$, and thus

$$\left\| \tilde{\mathbf{X}}_k^{x, \text{train}} \right\|_2 \leq 2L\sqrt{d_{\text{train}}}m^2 =: \tilde{R}, \quad (4.48)$$

and so that

$$\nu(\tilde{\mathbf{X}}^{x, \text{train}}) \leq N\tilde{R}^2, \quad (4.49)$$

where

$$\tilde{\mathbf{X}}^{x, \text{train}} := \sum_{k=1}^N \tilde{\mathbf{X}}_k^{x, \text{train}}, \quad (4.50)$$

and $\nu(\cdot)$ as defined in (A.3). Therefore, by [Corollary 1](#)

$$\begin{aligned}
\mathbb{P} \left(\left\| \frac{1}{N} \sum_{k=1}^N \mathbf{X}_k^{x, \text{train}} \right\|_2 \geq \epsilon/2 \right) &\leq \mathbb{P} \left(\left\| K_{\text{train}}^{-1} Y \right\|_2 \left\| \frac{1}{N} \sum_{k=1}^N \tilde{\mathbf{X}}_k^{x, \text{train}} \right\|_2 \geq \epsilon/2 \right) \\
&= \mathbb{P} \left(\left\| \frac{1}{N} \sum_{k=1}^N \tilde{\mathbf{X}}_k^{x, \text{train}} \right\|_2 \geq \left\| K_{\text{train}}^{-1} Y \right\|_2^{-1} \epsilon/2 \right) \\
&\leq (1 + d_{\text{train}}) \cdot \exp \left(- \frac{N \epsilon^2/8}{\left\| K_{\text{train}}^{-1} Y \right\|_2^2 \tilde{R}^2 + \left\| K_{\text{train}}^{-1} Y \right\|_2 \tilde{R} \epsilon/6} \right).
\end{aligned} \tag{4.51}$$

It remains to bound $\tilde{K}(x, X_{\text{train}}^T) (K_{\text{train}}^{-1} - \tilde{K}_{\text{train}}^{-1}) Y$. Let us define for $0 \leq t \leq 1$

$$K_t := (1 - t)K_{\text{train}} + t\tilde{K}_{\text{train}}. \tag{4.52}$$

Notice that

$$\frac{dK_t^{-1}}{dt} = K_t^{-1}(\tilde{K}_{\text{train}} - K_{\text{train}})K_t^{-1}. \tag{4.53}$$

Then we have that

$$\frac{d \left\| K_t^{-1} \right\|_{\text{op}}}{dt} \leq \left\| \frac{dK_t^{-1}}{dt} \right\|_{\text{op}} \leq \left\| K_t^{-1} \right\|_{\text{op}}^2 \left\| \tilde{K}_{\text{train}} - K_{\text{train}} \right\|_{\text{op}}. \tag{4.54}$$

To solve the above differential equation, we let

$$\phi(t) = \exp \left(- \frac{1}{\left\| K_{\text{train}} - \tilde{K}_{\text{train}} \right\|_{\text{op}} y_t} \right), \quad y_t = \left\| K_t^{-1} \right\|_{\text{op}}. \tag{4.55}$$

Then we obtain that

$$\frac{d\phi(t)}{dt} \leq \phi(t), \tag{4.56}$$

and so

$$\phi(t) \leq \phi(0) \exp(t). \tag{4.57}$$

Therefore one has that

$$y_t \leq \frac{y_0}{1 - t \left\| K_{\text{train}} - \tilde{K}_{\text{train}} \right\|_{\text{op}} y_0}, \quad y_0 = \left\| K_{\text{train}}^{-1} \right\|_{\text{op}} \tag{4.58}$$

as soon as $1 - t \left\| K_{\text{train}} - \tilde{K}_{\text{train}} \right\|_{\text{op}} y_0 > 0$. In what follows, we are interested to consider y_t with $t = 1$, and the case where (4.58) holds true. Notice that by (4.54), for $\epsilon > 0$

$$\begin{aligned}
\epsilon/2 &\leq \left| \tilde{K}(x, X_{\text{train}}^T) \left(K_{\text{train}}^{-1} - \tilde{K}_{\text{train}}^{-1} \right) Y \right| \\
&\leq \left\| \tilde{K}(x, X_{\text{train}}^T) \right\|_2 \left\| K_{\text{train}}^{-1} - \tilde{K}_{\text{train}}^{-1} \right\|_{\text{op}} \|Y\|_2 \\
&\leq \left\| \tilde{K}(x, X_{\text{train}}^T) \right\|_2 \|Y\|_2 \int_0^1 \frac{d}{dt} \left\| K_t^{-1} \right\|_{\text{op}} dt \\
&\leq \left\| \tilde{K}(x, X_{\text{train}}^T) \right\|_2 \|Y\|_2 \int_0^1 \left\| \frac{d}{dt} K_t^{-1} \right\|_{\text{op}} dt \\
&\leq \left\| \tilde{K}(x, X_{\text{train}}^T) \right\|_2 \|Y\|_2 \int_0^1 \left\| K_t^{-1} \right\|_{\text{op}}^2 \left\| \tilde{K} - K \right\|_{\text{op}} dt \\
&\leq \left\| \tilde{K}(x, X_{\text{train}}^T) \right\|_2 \left(\frac{\left\| K_{\text{train}} - \tilde{K}_{\text{train}} \right\|_{\text{op}} \left\| K_{\text{train}}^{-1} \right\|_{\text{op}}^2}{\left(1 - \left\| K_{\text{train}} - \tilde{K}_{\text{train}} \right\|_{\text{op}} \left\| K_{\text{train}}^{-1} \right\|_{\text{op}} \right)^2} \right) \|Y\|_2
\end{aligned} \tag{4.59}$$

where in the last inequality we have used (4.58). Let us define

$$a := \left\| K_{\text{train}} - \tilde{K}_{\text{train}} \right\|_{\text{op}} \left\| K_{\text{train}}^{-1} \right\|_{\text{op}}. \tag{4.60}$$

We have

$$\begin{aligned}
&\mathbb{P} \left(\epsilon/2 \leq \left| \tilde{K}(x, X_{\text{train}}^T) \left(K_{\text{train}}^{-1} - \tilde{K}_{\text{train}}^{-1} \right) Y \right| \right) \leq \\
&\leq \mathbb{P} \left(\epsilon/2 \leq \left| \tilde{K}(x, X_{\text{train}}^T) \left(K_{\text{train}}^{-1} - \tilde{K}_{\text{train}}^{-1} \right) Y \right|, 1 - a \leq 0 \right) + \\
&\quad + \mathbb{P} \left(\epsilon/2 \leq \left| \tilde{K}(x, X_{\text{train}}^T) \left(K_{\text{train}}^{-1} - \tilde{K}_{\text{train}}^{-1} \right) Y \right|, 1 - a > 0 \right) \\
&\leq \mathbb{P} (1 - a \leq 0) + \\
&\quad + \mathbb{P} \left(\epsilon/2 \leq \left\| \tilde{K}(x, X_{\text{train}}^T) \right\|_2 \left(\frac{\left\| K_{\text{train}} - \tilde{K}_{\text{train}} \right\|_{\text{op}} \left\| K_{\text{train}}^{-1} \right\|_{\text{op}}^2}{\left(1 - \left\| K_{\text{train}} - \tilde{K}_{\text{train}} \right\|_{\text{op}} \left\| K_{\text{train}}^{-1} \right\|_{\text{op}} \right)^2} \right) \|Y\|_2, 1 - a > 0 \right) \\
&= \mathbb{P} (1 - a \leq 0) + \mathbb{P} \left(1 - a > 0, \epsilon/2 \leq \frac{b}{(1 - a)^2} \right),
\end{aligned} \tag{4.61}$$

where a is defined in (4.60), and b is defined

$$b := \left\| \tilde{K}(x, X_{\text{train}}^T) \right\|_2 \left\| K_{\text{train}} - \tilde{K}_{\text{train}} \right\|_{\text{op}} \left\| K_{\text{train}}^{-1} \right\|_{\text{op}}^2 \|Y\|_2. \tag{4.62}$$

Notice that since $\left\| \tilde{K}(x, X_{\text{train}}^T) \right\|_2 \leq L\sqrt{d_{\text{train}}}m^2$, then

$$b \leq L\sqrt{d_{\text{train}}}m^2 \left\| K_{\text{train}} - \tilde{K}_{\text{train}} \right\|_{\text{op}} \left\| K_{\text{train}}^{-1} \right\|_{\text{op}}^2 \|Y\|_2 =: \tilde{b}, \tag{4.63}$$

so that

$$\begin{aligned}
& \mathbb{P}(1 - a \leq 0) + \mathbb{P}\left(1 - a > 0, \epsilon/2 \leq \frac{b}{(1 - a)^2}\right) \\
& \leq \mathbb{P}(1 - a \leq 0) + \mathbb{P}\left(1 - a > 0, \epsilon/2 \leq \frac{\tilde{b}}{(1 - a)^2}\right) \\
& = \mathbb{P}\left(1 - a \leq \sqrt{\frac{\tilde{b}}{\epsilon/2}}\right).
\end{aligned} \tag{4.64}$$

Notice that

$$\begin{aligned}
a + \frac{\tilde{b}^{\frac{1}{2}}}{(\epsilon/2)^{\frac{1}{2}}} &= \|K_{\text{train}} - \tilde{K}_{\text{train}}\|_{\text{op}} \|K_{\text{train}}^{-1}\|_{\text{op}} + \\
&+ (\epsilon/2)^{-1/2} \left(L \sqrt{d_{\text{train}}} m^2 \|K_{\text{train}} - \tilde{K}_{\text{train}}\|_{\text{op}} \|K_{\text{train}}^{-1}\|_{\text{op}}^2 \|Y\|_2 \right)^{\frac{1}{2}}
\end{aligned} \tag{4.65}$$

Let us consider the equation

$$1 = a + \frac{\tilde{b}^{\frac{1}{2}}}{(\epsilon/2)^{\frac{1}{2}}}, \tag{4.66}$$

and we solve it for $\|K_{\text{train}} - \tilde{K}_{\text{train}}\|_{\text{op}}^{\frac{1}{2}} = x$. Then we obtain

$$\begin{aligned}
1 &= x^2 \|K_{\text{train}}^{-1}\|_{\text{op}} + x(\epsilon/2)^{-\frac{1}{2}} \sqrt{L \sqrt{d_{\text{train}}} m^2 \|K_{\text{train}}^{-1}\|_{\text{op}}^2 \|Y\|_2} \\
&=: x^2 \alpha + x(\epsilon/2)^{-\frac{1}{2}} \beta
\end{aligned} \tag{4.67}$$

whose solutions for x are given by

$$x = \frac{-\beta(\epsilon/2)^{-\frac{1}{2}} \pm \sqrt{(\epsilon/2)^{-1} \beta^2 + 4\alpha}}{2\alpha}. \tag{4.68}$$

Since we have imposed the relation $\|K_{\text{train}} - \tilde{K}_{\text{train}}\|_{\text{op}}^{\frac{1}{2}} = x$, it implies that $x \geq 0$, and also that the negative solution in (4.68) is excluded. So that, from here the only solution that we need is

$$\begin{aligned}
x &= \frac{-\beta(\epsilon/2)^{-\frac{1}{2}} + \sqrt{(\epsilon/2)^{-1} \beta^2 + 4\alpha}}{2\alpha} = \frac{4\alpha}{2\alpha(\beta(\epsilon/2)^{-\frac{1}{2}} + \sqrt{(\epsilon/2)^{-1} \beta^2 + 4\alpha})} \\
&= \frac{2}{\beta(\epsilon/2)^{-\frac{1}{2}} + \sqrt{(\epsilon/2)^{-1} \beta^2 + 4\alpha}},
\end{aligned} \tag{4.69}$$

and where we are interested in the case where $\|K_{\text{train}} - \tilde{K}_{\text{train}}\|_{\text{op}}^{\frac{1}{2}} \geq x$. Now by [item 3.5](#), one has that

$$4\alpha \leq (\epsilon/2)^{-1}\beta^2, \quad (4.70)$$

and thus

$$\left\| K_{\text{train}} - \tilde{K}_{\text{train}} \right\|_{\text{op}}^{\frac{1}{2}} \geq \frac{2}{\beta(\epsilon/2)^{-\frac{1}{2}}(1 + \sqrt{2})} \quad (4.71)$$

that is,

$$\left\| K_{\text{train}} - \tilde{K}_{\text{train}} \right\|_{\text{op}} \geq \frac{4}{\beta^2(\epsilon/2)^{-1}(1 + \sqrt{2})^2} \quad (4.72)$$

Therefore,

$$\mathbb{P} \left(1 - a \leq \sqrt{\frac{\tilde{b}}{\epsilon/2}} \right) \leq \mathbb{P} \left(\left\| K_{\text{train}} - \tilde{K}_{\text{train}} \right\|_{\text{op}} \geq \frac{2\epsilon}{\beta^2(1 + \sqrt{2})^2} \right) \quad (4.73)$$

where

$$\beta^2 = L\sqrt{d_{\text{train}}}m^2 \left\| K_{\text{train}}^{-1} \right\|_{\text{op}}^2 \|Y\|_2. \quad (4.74)$$

Therefore, by [Theorem 5](#)

$$\mathbb{P} \left(\epsilon/2 \leq \left| \tilde{K}(x, X_{\text{train}}^T) \left(K_{\text{train}}^{-1} - \tilde{K}_{\text{train}}^{-1} \right) Y \right| \right) \leq \quad (4.75)$$

$$\leq 2 d_{\text{train}} \exp \left(- \frac{6 N \epsilon^2}{4(1 + \sqrt{2})^4 L^4 d_{\text{train}}^3 m^8 \left\| K_{\text{train}}^{-1} \right\|_{\text{op}}^4 \|Y\|_2^2} \right). \quad (4.76)$$

Therefore,

$$\begin{aligned} \mathbb{P}(\epsilon \leq |\mu_{\infty}(x) - \tilde{\mu}(x)|) &\leq (1 + d_{\text{train}}) \cdot \exp \left(- \frac{N \epsilon^2/8}{R^2 + R \epsilon/6} \right) + \\ &+ 2 d_{\text{train}} \exp \left(- \frac{6 N \epsilon^2}{4(1 + \sqrt{2})^4 L^4 d_{\text{train}}^3 m^8 \left\| K_{\text{train}}^{-1} \right\|_{\text{op}}^4 \|Y\|_2^2} \right), \end{aligned} \quad (4.77)$$

where

$$R = 2L\sqrt{d_{\text{train}}}m^2 \left\| K_{\text{train}}^{-1} Y \right\|_2. \quad (4.78)$$

Let us now impose that

$$(1 + d_{\text{train}}) \cdot \exp\left(-\frac{N \epsilon^2/8}{R^2 + R \epsilon/6}\right) \leq \frac{\delta}{2}, \quad (4.79)$$

$$2 d_{\text{train}} \exp\left(-\frac{6 N \epsilon^2}{4(1 + \sqrt{2})^4 L^4 d_{\text{train}}^3 m^8 \|K_{\text{train}}^{-1}\|_{\text{op}}^4 \|Y\|_2^2}\right) \leq \frac{\delta}{2}. \quad (4.80)$$

By solving for N we obtain the following lower bounds:

$$N \geq \frac{48R^2 + 8R\epsilon}{6\epsilon^2} \log\left(\frac{2(1 + d_{\text{train}})}{\delta}\right), \quad (4.81)$$

$$N \geq \frac{4(1 + \sqrt{2})^4 L^4 d_{\text{train}}^3 m^8 \|K_{\text{train}}^{-1}\|_{\text{op}}^4 \|Y\|_2^2}{6 \epsilon^2} \log\left(\frac{4d_{\text{train}}}{\delta}\right). \quad (4.82)$$

Then, we need that

$$\begin{aligned} N \geq & \frac{24R^2 + 4R\epsilon}{3\epsilon^2} \log\left(\frac{2(1 + d_{\text{train}})}{\delta}\right) \\ & + \frac{2(1 + \sqrt{2})^4 L^4 d_{\text{train}}^3 m^8 \|K_{\text{train}}^{-1}\|_{\text{op}}^4 \|Y\|_2^2}{3 \epsilon^2} \log\left(\frac{4d_{\text{train}}}{\delta}\right). \end{aligned} \quad (4.83)$$

In what follows, we determine the number of iid samples of θ needed to estimate μ_∞ with a uniformly bounded error on the whole $\mathcal{X}$. That is, we want to determine N such that $\max_{x \in \mathcal{X}} |\mu_\infty(x) - \tilde{\mu}_\infty(x)| < \epsilon$. Notice that by item (a), for any $\epsilon > 0$, and any $\delta' = \frac{\delta}{|\mathcal{X}|}$, $0 < \delta < 1$, we have

$$\begin{aligned} \mathbb{P}(\epsilon \leq \|\mu_\infty - \tilde{\mu}_\infty\|_\infty) & \leq \mathbb{P}\left(\bigcup_{x \in \mathcal{X}} \{\epsilon \leq |\mu_\infty(x) - \tilde{\mu}_\infty(x)|\}\right) \\ & \leq \sum_{x \in \mathcal{X}} \mathbb{P}(\epsilon \leq |\mu_\infty(x) - \tilde{\mu}_\infty(x)|) \\ & \leq |\mathcal{X}| \delta' \\ & = \delta. \end{aligned} \quad (4.84)$$

By the previous arguments, we conclude that μ_∞ can be estimated with a uniformly bounded error on the whole $\mathcal{X}$ with a number of samples equal to (3.6) with δ replaced by $\frac{\delta}{|\mathcal{X}|}$.

4.3.2 Estimation of the number of elementary operations

Let us now to estimate the number of elementary operations required to compute $\mu_\infty(x)$. Recall that by Lemma 3 and Lemma 5, the number of operations required to calculate $f_\theta(x)$ and $\nabla_\theta f_\theta(x)$ is $O(Lmn^2)$, and $O(L^2mn^2)$ respectively. Furthermore, the complexity of computing $\nabla_\theta f_\theta(X_{\text{train}})$ is $O(d_{\text{train}} L^2 mn^2)$. On the other hand, to compute the product $\nabla_\theta f_\theta(x) \nabla_\theta f_\theta(X_{\text{train}}^T)$ we need to perform $L d_{\text{train}}$ products. Hence, to compute $K(x, X_{\text{train}}^T)$, we need $O(NL[d_{\text{train}} Lmn^2 + d_{\text{train}}])$ operations. On the other hand, by [28], the complexity of inverting K_{train} is $O(d_{\text{train}}^3)$. Hence, the complexity of computing K_{train}^{-1} is $O(NL[Lmn^2 d_{\text{train}} + d_{\text{train}}^2] + d_{\text{train}}^3)$. Therefore, the complexity of computing $\mu_\infty(x)$ is then

$$O(NL[Lmn^2d_{\text{train}} + d_{\text{train}} + d_{\text{train}}^2] + d_{\text{train}}^3 + d_{\text{train}}). \quad (4.85)$$

In conclusion, the algorithm requires

$$O(NL^2 d_{\text{train}} [m n^2 + d_{\text{train}}^2]) \quad (4.86)$$

elementary operations.

5 Conclusions

We have proposed an efficient classical algorithm to estimate the NTK arising from a broad family of parametric quantum circuits constructed using Clifford unitaries that depend arbitrarily on the input interleaved by parametric Pauli rotations. Despite the parametric Pauli rotations belong to the Clifford group only for few special values of the parameters, we demonstrated that the proposed [Algorithm 1](#) can always efficiently estimate the NTK of the model with provable accuracy. Moreover, we have showed that [Algorithm 1](#) can be combined with the recently proved equivalence between very wide trained quantum neural networks and Gaussian processes [[11](#), [20](#)] to efficiently estimate the mean of the trained model function in the limit of infinite width ([Algorithm 2](#)). Therefore, our results prove that wide quantum neural networks with logarithmic depth trained on supervised-learning problems with the input encoded with Clifford gates can be simulated efficiently by classical computers and cannot achieve quantum advantage.

Our work aligns with a recent series of results [[3](#), [7](#), [18](#), [22](#), [33](#)] showing that the output of several parametric quantum circuits can be estimated efficiently using classical resources. These findings suggest a need for a more refined understanding of where true quantum advantage lies, and under what architectural constraints an efficient classical simulation is feasible.

Our results open the way to several promising research directions. For instance, it would be interesting to investigate whether similar classical simulability results hold for broader classes of quantum circuits beyond the ansatz considered here. In particular, it would be interesting to consider the scenarios where the input is not encoded via Clifford gates, or when at initialization the parameters are sampled from a distribution that is different from the uniform one. Such research directions could shed light on which architectures for quantum neural networks can have hope of quantum advantages.

Finally, the proposed algorithm to efficiently compute the NTK of complex parametric quantum circuits can be viewed as a method to synthesize expressive kernels for classical learning tasks, independently of whether the kernel originates from a physically implemented quantum circuit. In this view, a parametric quantum circuit serves as a generative mechanism for feature maps and the associate kernel, computed by our algorithm, can be directly used in a classical kernel method. This opens the possibility of leveraging quantum structural priors without relying on quantum hardware.

Acknowledgements

GDP has been supported by the HPC Italian National Centre for HPC, Big Data and Quantum Computing – Proposal code CN00000013 – CUP J33C22001170001 and by the Italian Extended Partnership PE01 – FAIR Future Artificial Intelligence Research – Proposal code PE00000013 – CUP J33C22002830006 under the MUR National Recovery and Resilience Plan funded by the European Union – NextGenerationEU. Funded by the European Union – NextGenerationEU

under the National Recovery and Resilience Plan (PNRR) – Mission 4 Education and research – Component 2 From research to business – Investment 1.1 Notice Prin 2022 – DD N. 104 del 2/2/2022, from title “understanding the LEarning process of QUantum Neural networks (LeQun)”, proposal code 2022WHZ5XH – CUP J53D23003890006. DP has been supported by project SERICS (PE00000014) under the MUR National Recovery and Resilience Plan funded by the European Union – NextGenerationEU. GDP and DP are members of the “Gruppo Nazionale per la Fisica Matematica (GNFM)” of the “Istituto Nazionale di Alta Matematica “Francesco Severi” (INdAM)”. AMH has been supported by project PRIN 2022 “understanding the LEarning process of QUantum Neural networks (LeQun)”, proposal code 2022WHZ5XH – CUP J53D23003890006. The author AMH is a member of the “Gruppo Nazionale per l’Analisi Matematica, la Probabilità e le loro Applicazioni (GNAMPA)” of the “Istituto Nazionale di Alta Matematica “Francesco Severi” (INdAM)”.

Declarations

Conflict of interest The authors confirm that there is no Conflict of interest.

A Bernstein’s upper bound for random matrices

In what follows, we aim to recall the so-called Bernstein’s inequality. In the next, we denote by $\|\cdot\|_{\text{op}}$ the operator norm.

Theorem 6 (Matrix Bernstein [31, Theorem 6.1.1]). *Consider a finite sequence $\{\mathbf{X}_k\}$ of independent, random matrices with dimension $d_1 \times d_2$. Assume that each random matrix satisfies*

$$\mathbb{E} \mathbf{X}_k = \mathbf{0} \quad \text{and} \quad \|\mathbf{X}_k\|_{\text{op}} \leq R \quad \text{almost surely.} \quad (\text{A.1})$$

Introduce the matrix

$$\mathbf{X} := \sum_k \mathbf{X}_k. \quad (\text{A.2})$$

Let $\nu(\mathbf{X})$ be the matrix variance statistic of the sum:

$$\nu(\mathbf{X}) := \max \left\{ \left\| \sum_k \mathbb{E}(\mathbf{X}_k \mathbf{X}_k^T) \right\|_{\text{op}}, \left\| \sum_k \mathbb{E}(\mathbf{X}_k^T \mathbf{X}_k) \right\|_{\text{op}} \right\}. \quad (\text{A.3})$$

Then, for all $t \geq 0$,

$$\mathbb{P} \left\{ \|\mathbf{X}\|_{\text{op}} \geq t \right\} \leq (d_1 + d_2) \cdot \exp \left(-\frac{t^2/2}{\nu(\mathbf{X}) + Rt/3} \right). \quad (\text{A.4})$$

Corollary 1. *Consider a finite sequence $\{\mathbf{X}_1, \dots, \mathbf{X}_N\}$ of independent and identically distributed random matrices with dimension $d_1 \times d_2$. Assume that each random matrix satisfies*

$$\mathbb{E} \mathbf{X}_k = \mathbf{0} \quad \text{and} \quad \|\mathbf{X}_k\|_{\text{op}} \leq R \quad \text{almost surely.} \quad (\text{A.5})$$

Then, for all $t \geq 0$,

$$\mathbb{P} \left\{ \left\| \frac{1}{N} \sum_{k=1}^N \mathbf{X}_k \right\|_{\text{op}} \geq t \right\} \leq (d_1 + d_2) \cdot \exp \left(-\frac{N t^2/2}{\nu(\mathbf{X}) + Rt/3} \right). \quad (\text{A.6})$$

References

- [1] Scott Aaronson and Daniel Gottesman. Improved simulation of stabilizer circuits. *Physical Review A*, 70(5), November 2004. ISSN 1094-1622. DOI: [10.1103/PhysRevA.70.052328](https://doi.org/10.1103/PhysRevA.70.052328). URL <http://dx.doi.org/10.1103/PhysRevA.70.052328>.
- [2] Erfan Abedi, Salman Beigi, and Leila Taghavi. Quantum Lazy Training. *Quantum*, 7:989, April 2023. ISSN 2521-327X. DOI: [10.22331/q-2023-04-27-989](https://doi.org/10.22331/q-2023-04-27-989). URL <https://doi.org/10.22331/q-2023-04-27-989>.
- [3] Armando Angrisani, Alexander Schmidhuber, Manuel S. Rudolph, M. Cerezo, Zoë Holmes, and Hsin-Yuan Huang. Classically estimating observables of noiseless quantum circuits. *arXiv preprint arXiv:2409.01706*, 2024. DOI: <https://doi.org/10.1103/lh6x-7rc3>.
- [4] Leonardo Banchi and Gavin E Crooks. Measuring analytic gradients of general quantum evolution with the stochastic parameter shift rule. *Quantum*, 5:386, 2021. DOI: <https://doi.org/10.22331/q-2021-01-25-386>.
- [5] E. W. Barankin. Locally best unbiased estimates. *The Annals of Mathematical Statistics*, 20(4):477–501, 1949. DOI: <https://doi.org/10.1214/aoms/1177729943>.
- [6] Jacob Biamonte, Peter Wittek, Nicola Pancotti, Patrick Rebentrost, Nathan Wiebe, and Seth Lloyd. Quantum machine learning. *Nature*, 549(7671):195–202, 2017. DOI: <https://doi.org/10.1038/nature23474>.
- [7] M. Cerezo, Martin Larocca, Diego García-Martín, N. L. Diaz, Paolo Braccia, Enrico Fontana, Manuel S. Rudolph, Pablo Bermejo, Aroosa Ijaz, Supanut Thanasilp, Eric R. Anschuetz, and Zoë Holmes. Does provable absence of barren plateaus imply classical simulability? *Nature Communications*, 16(1), August 2025. ISSN 2041-1723. DOI: [10.1038/s41467-025-63099-6](https://doi.org/10.1038/s41467-025-63099-6). URL <http://dx.doi.org/10.1038/s41467-025-63099-6>.
- [8] Lucas Pinheiro Cinelli, Matheus Araújo Marins, Eduardo Antonio Barros Da Silva, and Sérgio Lima Netto. *Variational methods for machine learning with applications to deep networks*, volume 15. Springer, 2021. DOI: <https://doi.org/10.1007/978-3-030-70679-1>.
- [9] Franklin De Lima Marquezino, Renato Portugal, and Carlile Lavor. *A primer on quantum computing*. Springer, 2019. DOI: <https://doi.org/10.1007/978-3-030-19066-8>.
- [10] Jeroen Dehaene and Bart De Moor. Clifford group, stabilizer states, and linear and quadratic operations over $gf(2)$. *Physical Review A*, 68(4), October 2003. ISSN 1094-1622. DOI: [10.1103/PhysRevA.68.042318](https://doi.org/10.1103/PhysRevA.68.042318). URL <http://dx.doi.org/10.1103/PhysRevA.68.042318>.
- [11] Filippo Girardi and Giacomo De Palma. Trained quantum neural networks are gaussian processes. *Communications in Mathematical Physics*, 406(4), April 2025. ISSN 1432-0916. DOI: [10.1007/s00220-025-05238-0](https://doi.org/10.1007/s00220-025-05238-0). URL <http://dx.doi.org/10.1007/s00220-025-05238-0>.
- [12] Paul R Halmos. The theory of unbiased estimation. *The Annals of Mathematical Statistics*, 17(1):34–43, 1946. DOI: <https://doi.org/10.1214/aoms/1177731020>.
- [13] Vojtěch Havlíček, Antonio D Córcoles, Kristan Temme, Aram W Harrow, Abhinav Kandala, Jerry M Chow, and Jay M Gambetta. Supervised learning with quantum-enhanced feature spaces. *Nature*, 567(7747):209–212, 2019. DOI: <https://doi.org/10.1038/s41586-019-0980-2>.
- [14] Arthur Jacot, Franck Gabriel, and Clement Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. DOI: <https://doi.org/10.48550/arXiv.1806.07572>. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/5a4be1fa34e62bb8a6ec6b91d2462f5a-Paper.pdf.

- [15] Junyu Liu, Francesco Tacchino, Jennifer R. Glick, Liang Jiang, and Antonio Mezzacapo. Representation learning via quantum neural tangent kernels. *PRX Quantum*, 3:030323, Aug 2022. DOI: [10.1103/PRXQuantum.3.030323](https://doi.org/10.1103/PRXQuantum.3.030323). URL <https://link.aps.org/doi/10.1103/PRXQuantum.3.030323>.
- [16] Yunchao Liu, Srinivasan Arunachalam, and Kristan Temme. A rigorous and robust quantum speed-up in supervised machine learning. *Nature Physics*, 17(9):1013–1017, 2021. DOI: <https://doi.org/10.1038/s41567-021-01287-z>.
- [17] Seth Lloyd, Maria Schuld, Aroosa Ijaz, Josh Izaac, and Nathan Killoran. Quantum embeddings for machine learning. *arXiv preprint arXiv:2001.03622*, 2020. DOI: <https://doi.org/10.48550/arXiv.2001.03622>.
- [18] Victor Martinez, Armando Angrisani, Ekaterina Pankovets, Omar Fawzi, and Daniel Stilck França. Efficient simulation of parametrized quantum circuits under nonunitary noise through pauli backpropagation. *Phys. Rev. Lett.*, 134:250602, Jun 2025. DOI: [10.1103/jlgg-s6zb](https://doi.org/10.1103/jlgg-s6zb). URL <https://link.aps.org/doi/10.1103/jlgg-s6zb>.
- [19] Kieran Mastel. The clifford theory of the n -qubit clifford group. *Journal of Mathematical Physics*, 2026. DOI: <https://doi.org/10.1063/5.0311547>.
- [20] Anderson Melchor Hernandez, Filippo Girardi, Davide Pastorello, and Giacomo De Palma. Quantitative convergence of trained quantum neural networks to a gaussian process: A. melchor hernandez et al. *Annales Henri Poincaré*, pages 1–57, 2025. DOI: <https://doi.org/10.1007/s00023-025-01631-6>.
- [21] Davide Pastorello. *Concise guide to quantum machine learning*. Springer, 2023. DOI: <https://doi.org/10.1007/978-981-19-6897-6>.
- [22] Oliver Reardon-Smith, Michał Oszmaniec, and Kamil Korzekwa. Improved simulation of quantum circuits dominated by free fermionic operations. *Quantum*, 8:1549, December 2024. ISSN 2521-327X. DOI: [10.22331/q-2024-12-04-1549](https://doi.org/10.22331/q-2024-12-04-1549). URL <https://doi.org/10.22331/q-2024-12-04-1549>.
- [23] Francesco Scala, Christa Zoufal, Dario Gerace, and Francesco Tacchino. Towards practical quantum neural network diagnostics with neural tangent kernels. *arXiv preprint arXiv:2503.01966*, 2025. DOI: <https://doi.org/10.48550/arXiv.2503.01966>.
- [24] Maria Schuld and Francesco Petruccione. *Supervised learning with quantum computers*, volume 17. Springer, 2018. DOI: <https://doi.org/10.1007/978-3-319-96424-9>.
- [25] Maria Schuld, Ilya Sinayskiy, and Francesco Petruccione. An introduction to quantum machine learning. *Contemporary Physics*, 56(2):172–185, 2015. DOI: <https://doi.org/10.1080/00107514.2014.964942>.
- [26] 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(3):032430, 2021. DOI: <https://doi.org/10.1103/PhysRevA.103.032430>.
- [27] Norihito Shirai, Kenji Kubo, Kosuke Mitarai, and Keisuke Fujii. Quantum tangent kernel. *Phys. Rev. Res.*, 6(3):033179, 2024. DOI: [10.1103/PhysRevResearch.6.033179](https://doi.org/10.1103/PhysRevResearch.6.033179).
- [28] Volker Strassen. Gaussian elimination is not optimal. *Numerische mathematik*, 13(4):354–356, 1969. DOI: <https://doi.org/10.1007/BF02165411>.
- [29] Michel Talagrand. The missing factor in hoeffding’s inequalities. *Annales de l’IHP Probabilités et statistiques*, 31(4):689–702, 1995. URL https://www.numdam.org/item/AIHPB_1995__31_4_689_0/.
- [30] Joel A. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434, August 2011. ISSN 1615-3383. DOI: [10.1007/s10208-011-9099-z](https://doi.org/10.1007/s10208-011-9099-z). URL <http://dx.doi.org/10.1007/s10208-011-9099-z>.
- [31] Joel A. Tropp. An introduction to matrix concentration inequalities. 2015. DOI: <https://doi.org/10.48550/arXiv.1501.01571>. URL <https://arxiv.org/abs/1501.01571>.

- [32] Li-Wei Yu, Weikang Li, Qi Ye, Zhide Lu, Zizhao Han, and Dong-Ling Deng. Expressibility-induced concentration of quantum neural tangent kernels. *Reports on Progress in Physics*, 87(11):110501, oct 2024. DOI: [10.1088/1361-6633/ad82cf](https://doi.org/10.1088/1361-6633/ad82cf). URL <https://dx.doi.org/10.1088/1361-6633/ad82cf>.
- [33] Yifan Zhang and Yuxuan Zhang. Classical simulability of quantum circuits with shallow magic depth. *PRX Quantum*, 6:010337, Feb 2025. DOI: [10.1103/PRXQuantum.6.010337](https://doi.org/10.1103/PRXQuantum.6.010337). URL <https://link.aps.org/doi/10.1103/PRXQuantum.6.010337>.