Representation Learning via Quantum Neural Tangent Kernels

Junyu Liu,[1,2,3,*] Francesco Tacchino,[4,†] Jennifer R. Glick,[5,‡] Liang Jiang,[1,2,§]
and Antonio Mezzacapo[5,¶]

[1] *Pritzker School of Molecular Engineering, The University of Chicago, Chicago, Illinois 60637, USA*
[2] *Chicago Quantum Exchange, Chicago, Illinois 60637, USA*
[3] *Kadanoff Center for Theoretical Physics, The University of Chicago, Chicago, Illinois 60637, USA*
[4] *IBM Quantum, IBM Research – Zurich, Rüschlikon 8803, Switzerland*
[5] *IBM Quantum, IBM T. J. Watson Research Center, Yorktown Heights, New York 10598, USA*

(Received 29 April 2022; revised 19 July 2022; accepted 28 July 2022; published 17 August 2022)

Variational quantum circuits are used in quantum machine learning and variational quantum simulation tasks. Designing good variational circuits or predicting how well they perform for given learning or optimization tasks is still unclear. Here we discuss these problems, analyzing variational quantum circuits using the theory of neural tangent kernels. We define quantum neural tangent kernels, and derive dynamical equations for their associated loss function in optimization and learning tasks. We analytically solve the dynamics in the frozen limit, or lazy training regime, where variational angles change slowly and a linear perturbation is good enough. We extend the analysis to a dynamical setting, including quadratic corrections in the variational angles. We then consider a hybrid quantum classical architecture and define a large-width limit for hybrid kernels, showing that a hybrid quantum classical neural network can be approximately Gaussian. The results presented here show limits for which analytical understandings of the training dynamics for variational quantum circuits, used for quantum machine learning and optimization problems, are possible. These analytical results are supported by numerical simulations of quantum machine-learning experiments.

DOI: 10.1103/PRXQuantum.3.030323

I. INTRODUCTION

The idea of using quantum computers for machine learning has recently received attention both in academia and industry [1–13]. While proof-of-principle studies have shown that some problems of mathematical interest quantum computers are useful [13], quantum advantage in machine-learning algorithms for practical applications is still unclear [14]. On classical architectures, a first-principles theory of machine learning, especially the so-called deep learning that uses a large number of layers, is still in development. Early developments of the statistical learning theory provide rigorous guarantees on the learning capability in generic learning algorithms, but theoretical bounds obtained from information theory are sometimes weak in practical settings.

The theory of neural tangent kernel (NTK) has been deemed an important tool to understand deep neural networks [15–21]. In the large-width limit, a generic neural network becomes nearly Gaussian when averaging over the initial weights and biases, and the learning capabilities become predictable. The NTK theory allows an analytical understanding of the neural networks' dynamics to be derived, improving on statistical learning theory and shedding light on the underlying principle of deep learning [22–26]. In quantum machine learning, a similar first-principles theory would help in understanding the training dynamics and selecting appropriate variational quantum circuits to target specific problems. A step in this direction has been considered recently, and originally, for quantum classical neural networks [27]. However, the framework of Ref. [27] is mostly focused on the classical convolutional neural networks combined with quantum circuits, and it does not address the quantum gradient-descent dynamics of variational circuits.

In this paper, we address this problem, focusing on the limit where the learning rate is sufficiently small, inspired

[*] Corresponding author. junyuliu@uchicago.edu
[†] fta@zurich.ibm.com
[‡] jennifer.r.glick@ibm.com
[§] liang.jiang@uchicago.edu
[¶] mezzacapo@ibm.com

by the classical theory of NTK. Following the framework and results from Refs. [24,25,28], we first define a quantum analog of a classical NTK. In the limit where the variational angles do not change much, the so-called *lazy training* [29], the *frozen* quantum neural tangent kernel (QNTK) leads to an exponential decaying of the loss function used on the training set. We furthermore compute the leading-order perturbation above the static limit, where we define a quantum version of the classical *metakernel*. We derive closed-form formulas for the dynamics of the training in terms of parameters of variational quantum circuits, see Fig. 1.

We then move to a hybrid quantum classical neural network framework, and find that it becomes approximately Gaussian, as long as the quantum outputs are sufficiently orthogonal. We present an analytic derivation of the large-width limit where the non-Gaussian contribution to the neuron correlations is suppressed by large width. Interestingly, we observe that now the *width* is defined by the number of independent Hermitian operators in the variational ansatz, which is upper bounded by (a polynomial of) the dimension of the Hilbert space. Thus, a large Hilbert-space size will naturally bring our neural network to the large-width limit. Moreover, the orthogonality assumption in the variational ansatz could be achieved statistically using randomized assumptions. If not, the hybrid quantum classical neural networks could still learn features even at the large width, indicating a significant difference compared to the classical neural networks.

We test the analytical derivations of our theory compared against numerical experiments with the IBM quantum device simulator [30], on a classification problem in the supervised learning setting, finding good agreement with the theory. The structure of this paper and the ideas presented are summarized in Fig. 2.

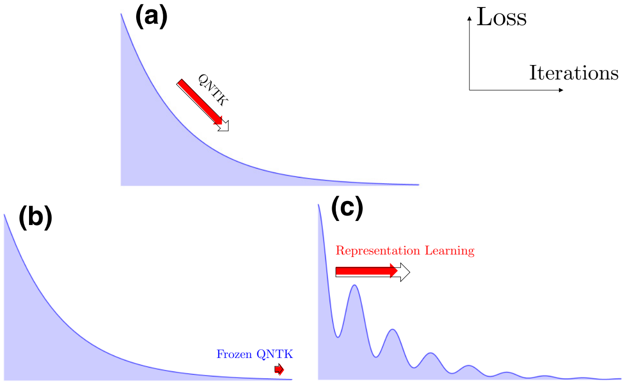

FIG. 1. An illustration of the QNTK theory. (a) The QNTK characterizes the gradient-descent dynamics in the variational quantum circuit. The quantum state modifies according to the QNTK prediction. (b) Around the end of the training, the QNTK is *frozen* and almost a constant. (c) The gradient-descent dynamics could be highly nonlinear, and the QNTK is running during gradient descent, which is a property of representation learning.

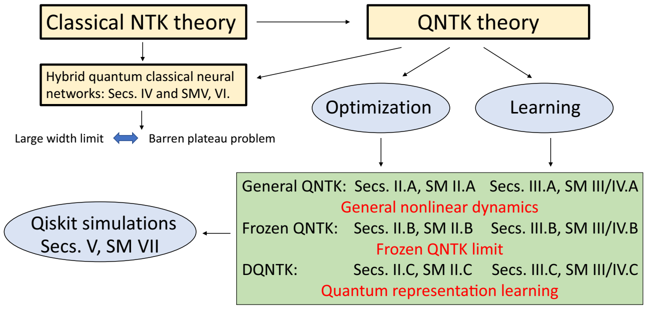
Classical NTK theory
QNTK theory
Hybrid quantum classical neural networks: Secs. IV and SMV, VI.
Large width limit
Barren plateau problem
Optimization
Learning
General QNTK: Secs. II.A, SM II.A Secs. III.A, SM III/IV.A
General nonlinear dynamics
Frozen QNTK: Secs. II.B, SM II.B Secs. III.B, SM III/IV.B
Frozen QNTK limit
DQNTK: Secs. II.C, SM II.C Secs. III.C, SM III/IV.C
Quantum representation learning
Qiskit simulations Secs. V, SM VII

FIG. 2. Structure of our paper. In Sec. II we establish the theory of QNTK in the context of optimization without data for generic variational quantum ansatz, which is the typical task in quantum simulation. In Sec. III, we establish the theory of quantum machine learning with the help of QNTK. In Sec. IV, we define the hybrid quantum classical neural network model, and we prove that in the large-width limit, the model is approximated by the Gaussian process. In Sec. V, we give numerical examples to demonstrate our quantum representation theory. In Sec. VI, we discuss the implication of this work, and outline open problems for future works. In the main text, we mostly highlight our theoretical frameworks and important theorems. Technical details are given in the Supplemental Material (SM) [31].

II. THEORY OF QUANTUM OPTIMIZATION

A. QNTK for optimization

We start from a relatively simple example about the optimization of a quantum cost function, without a model to be learned from some data associated to it. Let a variational quantum wave function [32–37] be given as

$$|\phi(\theta)\rangle = U(\theta)\,|\Psi_0\rangle = \left(\prod_{\ell=1}^{L} W_\ell \exp\left(i\theta_\ell X_\ell\right)\right)|\Psi_0\rangle\,. \quad (1)$$

Here we define L unitary operators of the type $U_\ell(\theta_\ell) = \exp(i\theta_\ell X_\ell)$, with a variational parameter θ_ℓ, and a Hermitian operator X_ℓ associated to them. We denote the vector version of all variational parameters as $\theta = \{\theta_\ell\}$ and the initial state as $|\Psi_0\rangle$. Our ansatz also includes constant gates W_ℓs that do not depend on the variational angles. Here, we write $U_\ell(\theta_\ell)$ as U_ℓ, but U_ℓs are θ dependent.

We introduce the following mean-squared-error (MSE) loss function when we wish to optimize the expectation value of a Hermitian operator O to its minimal eigenvalue O_0, which is assumed to be known here, over the class of states $|\phi(\theta)\rangle$

$$\mathcal{L}(\theta) = \frac{1}{2}\left(\left\langle\Psi_0\left|U^\dagger(\theta)OU(\theta)\right|\Psi_0\right\rangle - O_0\right)^2 \equiv \frac{1}{2}\varepsilon^2. \quad (2)$$

Here we define the *residual optimization error* $\varepsilon \equiv \left\langle\Psi_0\left|U^\dagger(\theta)OU(\theta)\right|\Psi_0\right\rangle - O_0$. When using gradient descent to optimize Eq. (2), the difference equation for the dynamics of the training parameter is given by

$$\text{đ}\theta_\ell = -\eta \frac{d\mathcal{L}(\theta)}{d\theta_\ell} = -\eta\varepsilon \frac{d\varepsilon}{d\theta_\ell}. \tag{3}$$

We use the notation $\text{đ}o$ to denote the difference between the step $t+1$ and the step t during gradient descent for the quantity o, $\text{đ}o = o(t+1) - o(t)$, associated to a learning rate η. Then we have also, to the linear order in θ,

$$\text{đ}\varepsilon = \sum_\ell \frac{d\varepsilon}{d\theta_\ell}\text{đ}\theta_\ell = -\eta \sum_\ell \frac{d\varepsilon}{d\theta_\ell}\frac{d\varepsilon}{d\theta_\ell}\varepsilon. \tag{4}$$

The object $\sum_\ell d\varepsilon/d\theta_\ell d\varepsilon/d\theta_\ell$ serves to construct a toy version of the NTK in the quantum setup, in the sense that it can be seen as a one-dimensional kernel matrix with *training data* O_0. We can make our definition of a QNTK associated to an optimization problem more precise as follows:

Definition 1 (QNTK for optimization)**.** *The QNTK associated to the optimization problem of Eq. (2) is given by*

$$\begin{aligned} K &= \sum_\ell \frac{d\varepsilon}{d\theta_\ell}\frac{d\varepsilon}{d\theta_\ell} \\ &= -\left\langle \Psi_0 \left| U_{+,\ell}^\dagger \left[X_\ell, U_\ell^\dagger W_\ell^\dagger U_{-,\ell}^\dagger O U_{-,\ell} W_\ell U_\ell\right] U_{+,\ell} \right| \Psi_0 \right\rangle^2, \end{aligned} \tag{5}$$

where

$$U_{-,\ell} \equiv \prod_{\ell'=1}^{\ell-1} W_{\ell'} U_{\ell'}, \quad U_{+,\ell} \equiv \prod_{\ell'=\ell+1}^{L} W_{\ell'} U_{\ell'}. \tag{6}$$

It is easy to show that the quantity squared in Eq (5) is imaginary, hence K is always non-negative, $K \geq 0$. A derivation of Eq. (5) can be found within the Supplemental Material [31].

B. Frozen QNTK limit for optimization

An analytic theory of the NTK is established when the learning rate is sufficiently small. It is defined by solving the coupled difference equations, Eqs. (3), (4), which we report here

$$\begin{aligned} \text{đ}\theta_\ell &= -\eta\varepsilon\frac{d\varepsilon}{d\theta_\ell}, \\ \text{đ}\varepsilon &= -\eta \sum_\ell \frac{d\varepsilon}{d\theta_\ell}\frac{d\varepsilon}{d\theta_\ell}\varepsilon = -\eta K \varepsilon. \end{aligned} \tag{7}$$

In the continuum learning rate limit $\eta \rightarrow 0$, Eqs. (7) become coupled nonlinear ordinary differential equations, which are hard to solve in general. Note that this system of equations stems from a quantum optimization problem and in general it is classically hard to even instantiate.

Nevertheless, in the following we build an analytic model for a quantum version of the *frozen NTK* (frozen QNTK) in the regime of *lazy training*, where variational angles do not change too much. To be more precise, we assume that at a certain value θ^* our variational angles θ change by a small amount, $\theta^* + \delta\varphi$. A typical scenario is to do the Taylor expansion around such values θ^* during the convergence regime for instance. Here δ is a small scaling parameter and we define δ together with φ to denote small perturbations. We call the limit $\delta \rightarrow 0^+$ the *frozen QNTK limit*.

In this limit, one can write $W_\ell U_\ell = W_\ell \exp(i\theta_\ell^* X_\ell) \exp(i\delta\varphi_\ell X_\ell)$, so that the θ^* dependence is absorbed into the nonvariational part of the unitary by defining $W_\ell(\theta_\ell^*) \equiv W_\ell \exp(i\theta_\ell^* X_\ell)$, and we have $W_\ell U_\ell \rightarrow W_\ell(\theta_\ell^*) \exp(i\delta\varphi_\ell X_\ell)$. In what follows, we drop the θ^* notation and understand the variational angles as small parameters that change by δ around a value θ^*. Then, expanding linearly for small δ we can define the following.

Definition 2 (Frozen QNTK for quantum optimization)**.** *In the optimization problem, Eq. (2), the frozen QNTK limit is*

$$\begin{aligned} K = &-\delta^2 \\ &\times \sum_\ell \left\langle \Psi_0 \left| W_{+,\ell}^\dagger \left[X_\ell, W_\ell^\dagger W_{-,\ell}^\dagger O W_{-,\ell} W_\ell\right] W_{+,\ell} \right| \Psi_0 \right\rangle^2, \end{aligned} \tag{8}$$

with

$$W_{-,\ell} \equiv \prod_{\ell'=1}^{\ell-1} W_{\ell'}, \quad W_{+,\ell} \equiv \prod_{\ell'=\ell+1}^{L} W_{\ell'}. \tag{9}$$

In the frozen kernel limit, we can state the following result about the dependency of the residual error ϵ, solving Eq. (7) linearly for small δ.

Theorem 1 (Performance guarantee of optimization within the frozen QNTK approximation)**.** *When using standard gradient descent for the optimization problem, Eq. (2), within the frozen QNTK limit, the residual optimization error* ε *decays exponentially as*

$$\begin{aligned} \varepsilon(t) = (1-\eta K)^t \varepsilon(0) = \varepsilon(0) \times \Bigg(1 + \eta\delta^2 \\ \times \sum_\ell \left\langle \Psi_0 \left| W_{+,\ell}^\dagger \left[X_\ell, W_\ell^\dagger W_{-,\ell}^\dagger O W_{-,\ell} W_\ell\right] W_{+,\ell} \right| \Psi_0 \right\rangle^2 \Bigg)^t, \end{aligned} \tag{10}$$

with a convergence rate defined as

$$\begin{aligned}\tau_c &= -\log(1-\eta K) \approx \eta K \\ &= \eta\delta^2 \sum_{\ell} \left\langle \Psi_0 \left| W_{+,\ell}^{\dagger} \left[X_\ell, W_\ell^{\dagger} W_{-,\ell}^{\dagger} O W_{-,\ell} W_\ell \right] W_{+,\ell} \right| \Psi_0 \right\rangle^2 \leq 2\eta\delta^2 L \|O\|^2 \max_{\ell} \|X_\ell\|^2,\end{aligned} \tag{11}$$

with the $\mathbb{L}_2$ *norm.*

The derivation is given within the Supplemental Material [31]. An immediate consequence is that the residual error will converge to zero,

$$\varepsilon(\infty) = 0. \tag{12}$$

C. Differential of QNTK (DQNTK)

The frozen QNTK limit describes the regime of the linear approximation of nonlinearities. Therefore, the frozen QNTK cannot reflect the nonlinear nature of the variational quantum algorithms. In order to formulate an analytical model of the nonlinearities, we now analyze the leading-order correction in terms of the expansion of the learning rate η and the size of the variational angle δ. We formulate the expansion of $\text{đ}\varepsilon$ to the second order in $\text{đ}\varphi$,

$$\text{đ}\varepsilon = \sum_{\ell} \frac{d\varepsilon}{d\varphi_\ell} \text{đ}\varphi_\ell + \frac{1}{2} \sum_{\ell_1,\ell_2} \frac{d^2\varepsilon}{d\varphi_{\ell_1} d\varphi_{\ell_2}} \text{đ}\varphi_{\ell_1} \text{đ}\varphi_{\ell_2}. \tag{13}$$

This time $\text{đ}\varepsilon$ during gradient descent will follow the equation [25]:

$$\text{đ}\varepsilon = -\eta \sum_{\ell} \frac{d\varepsilon}{d\varphi_\ell} \frac{d\varepsilon}{d\varphi_\ell} \varepsilon + \frac{1}{2}\eta^2\varepsilon^2 \sum_{\ell_1,\ell_2} \frac{d^2\varepsilon}{d\varphi_{\ell_1} d\varphi_{\ell_2}} \frac{d\varepsilon}{d\varphi_{\ell_1}} \frac{d\varepsilon}{d\varphi_{\ell_2}}. \tag{14}$$

With this expansion at second order, we have two contributing terms in Eq. (13). We label the first term of Eq. (13) quantum *effective* kernel, K^E. We use K^E to distinguish it from K, when only a first-order expansion is considered in the description of the dynamics. It is dynamical in the sense that it depends on the value of the training parameter φ during the dynamics regulated by a gradient descent. We label the variable part of the second term in Eq. (14) quantum *metakernel* or DQNTK.

Definition 3 (Quantum metakernel for optimization)**.** *The quantum metakernel associated with the optimization problem in Eq. (2) is defined via*

$$\mu = \sum_{\ell_1,\ell_2} \frac{d^2\varepsilon}{d\varphi_{\ell_1} d\varphi_{\ell_2}} \frac{d\varepsilon}{d\varphi_{\ell_1}} \frac{d\varepsilon}{d\varphi_{\ell_2}}. \tag{15}$$

In the limit of small changes in $\theta = \theta^* + \delta\varphi$, optimization problem, Eq. (2), the quantum metakernel is given at the leading-order perturbation theory in δ as

$$\begin{aligned}\mu = \delta^4 \sum_{\ell_1,\ell_2} &\left\langle \Psi_0 \left| W_{+,\ell_1}^{\dagger} \left[X_{\ell_1}, W_{\ell_1}^{\dagger} W_{-,\ell_1}^{\dagger} O W_{-,\ell_1} W_{\ell_1} \right] W_{+,\ell_1} \right| \Psi_0 \right\rangle \\ &\left\langle \Psi_0 \left| W_{+,\ell_2}^{\dagger} \left[X_{\ell_2}, W_{\ell_2}^{\dagger} W_{-,\ell_2}^{\dagger} O W_{-,\ell_2} W_{\ell_2} \right] W_{+,\ell_2} \right| \Psi_0 \right\rangle \times \\ &\left(\begin{cases} \left\langle \Psi_0 \left| W_{+,\ell_1}^{\dagger} \left[X_{\ell_1}, Q_{\ell_1,\ell_2}^{\dagger} \left[X_{\ell_2}, W_{\ell_2}^{\dagger} W_{-,\ell_2}^{\dagger} O W_{-,\ell_2} W_{\ell_2} \right] Q_{\ell_2,\ell_1} \right] W_{+,\ell_1} \right| \Psi_0 \right\rangle : \ell_1 \geq \ell_2 \\ \left\langle \Psi_0 \left| W_{+,\ell_2}^{\dagger} \left[X_{\ell_2}, Q_{\ell_2,\ell_1}^{\dagger} \left[X_{\ell_1}, W_{\ell_1}^{\dagger} W_{-,\ell_1}^{\dagger} O W_{-,\ell_1} W_{\ell_1} \right] Q_{\ell_1,\ell_2} \right] W_{+,\ell_2} \right| \Psi_0 \right\rangle : \ell_1 < \ell_2 \end{cases} \right).\end{aligned} \tag{16}$$

where

$$\begin{aligned} W_{\ell_1,\ell_2} &\equiv \prod_{\ell=\ell_1+1}^{\ell_2-1} W_\ell \,, \\ Q_{\ell_1,\ell_2} &= \begin{cases} W_{\ell_1,\ell_2} W_{\ell_2} : \ell_1 < \ell_2 \\ 1 : \ell_1 = \ell_2. \end{cases} \end{aligned} \tag{17}$$

The residual error ε in the optimization problem of Eq. (2), can then be computed as

$$\varepsilon = \left\langle \Psi_0 \left| \left(\prod\nolimits_{\ell'=L}^{1} W_{\ell'}^{\dagger}\right) O \left(\prod\nolimits_{\ell=1}^{L} W_{\ell}\right) \right| \Psi_0 \right\rangle - O_0 - i\delta \sum_{\ell} \varphi_\ell \left\langle \Psi_0 \left| W_{+,\ell}^{\dagger} \left[X_\ell, W_\ell^{\dagger} W_{-,\ell}^{\dagger} O W_{-,\ell} W_\ell\right] W_{+,\ell} \right| \Psi_0 \right\rangle$$
$$- \frac{\delta^2}{2} \sum_{\ell_1,\ell_2} \varphi_{\ell_1}\varphi_{\ell_2} \times \begin{cases} \left\langle \Psi_0 \left| W_{+,\ell_1}^{\dagger} \left[X_{\ell_1}, Q_{\ell_1,\ell_2}^{\dagger} \left[X_{\ell_2}, W_{\ell_2}^{\dagger} W_{-,\ell_2}^{\dagger} O W_{-,\ell_2} W_{\ell_2}\right] Q_{\ell_2,\ell_1}\right] W_{+,\ell_1} \right| \Psi_0 \right\rangle : \ell_1 \geq \ell_2 \\ \left\langle \Psi_0 \left| W_{+,\ell_2}^{\dagger} \left[X_{\ell_2}, Q_{\ell_2,\ell_1}^{\dagger} \left[X_{\ell_1}, W_{\ell_1}^{\dagger} W_{-,\ell_1}^{\dagger} O W_{-,\ell_1} W_{\ell_1}\right] Q_{\ell_1,\ell_2}\right] W_{+,\ell_2} \right| \Psi_0 \right\rangle : \ell_1 < \ell_2 \end{cases} . \tag{18}$$

We are now ready to make a statement about the residual error in the limit of the DQNTK

Theorem 2 (Performance guarantee of optimization from DQNTK)**.** *In the optimization problem, Eq. (2), at the DQNTK order, we split the residual optimization error into two pieces, the free part, and the interacting part,*

$$\varepsilon = \varepsilon^F + \varepsilon^I. \tag{19}$$

The free part follows the exponentially decaying dynamics

$$\varepsilon^F = (1 - \eta K)^t \varepsilon(0), \tag{20}$$

and the interacting part is given by

$$\varepsilon^I(t) = -\eta t (1 - \eta K)^{t-1} K^{\Delta} \varepsilon(0). \tag{21}$$

Here we have

$$K^{\Delta} \equiv K^E(0) - K = \left(\sum_{\ell} \frac{d\varepsilon}{d\theta_\ell}\frac{d\varepsilon}{d\theta_\ell}\right)(0) - \sum_{\ell} \frac{d\varepsilon^F}{d\theta_\ell}\frac{d\varepsilon^F}{d\theta_\ell}$$
$$= 2i\delta^3 \sum_{\ell} \left\langle \Psi_0 \left| W_{+,\ell}^{\dagger} \left[X_\ell, W_\ell^{\dagger} W_{-,\ell}^{\dagger} O W_{-,\ell} W_\ell\right] W_{+,\ell} \right| \Psi_0 \right\rangle$$
$$\sum_{\ell'} \left.\begin{array}{l} \left\langle \Psi_0 \left| W_{+,\ell'}^{\dagger} \left[X_{\ell'}, Q_{\ell',\ell}^{\dagger} \left[X_\ell, W_\ell^{\dagger} W_{-,\ell}^{\dagger} O W_{-,\ell} W_\ell\right] W_{\ell,\ell'} W_{\ell'}\right] W_{+,\ell'} \right| \Psi_0 \right\rangle : \ell' \geq \ell \\ \left\langle \Psi_0 \left| W_{+,\ell}^{\dagger} \left[X_\ell, Q_{\ell,\ell'}^{\dagger} \left[X_{\ell'}, W_{\ell'}^{\dagger} W_{-,\ell'}^{\dagger} O W_{-,\ell'} W_{\ell'}\right] W_{\ell',\ell} W_\ell\right] W_{+,\ell} \right| \Psi_0 \right\rangle : \ell' < \ell \end{array}\right. \varphi_{\ell'}(0). \tag{22}$$

Thus, the residual optimization error ε will always finally approach zero,

$$\varepsilon(\infty) = 0. \tag{23}$$

Thus, the leading-order perturbative correction gives the contribution $\mathcal{O}(\delta^3)$.

Moreover, we notice that DQNTK leads to interesting physical consequences. More precisely, the next leading correction above the perturbative limit will cause the so-called *catapult effect*, where there are small bumps appearing before an exponential decay [38]. The reason is rather simple and probably most clearly explained in our draft within the Supplemental Material [31]. We know that the leading order gives schematically the term approximately $\exp(-\eta K t)$ for the residual training error, where η is the small learning rate, K is the quantum neural tangent kernel, and t is the number of iterations. Moreover, we derive the correction towards the residual training error, which scales as approximately $t\exp(-\eta K t)$. In general, in higher-order corrections, we get schematically the correction approximately $t^p \exp(-\eta K t)$ for a more general polynomial t^p in the prefactor of the exponential decay. This type of correction forms a first-principles explanation of the catapult effect, where a similar related model is discussed. A full characterization of the catapult effect in classical and quantum cases is beyond the scope of this paper, and we leave it for future research [39].

III. THEORY OF LEARNING

A. General theory

The results outlined in Sec. II can be extended in the context of supervised learning from a data space $\mathcal{D}$. In particular, we are given a training set contained in the dataspace $\mathcal{A} \subset \mathcal{D}$. The data can be loaded into quantum states through a quantum feature map [9,11]. We define the variational quantum ansatz with a single *layer* by regarding the output of a quantum neural network with the data point $\mathbf{x}_\delta$ as

$$z_{i;\delta} \equiv z_i\left(\theta, \mathbf{x}_\delta\right) = \left\langle \phi\left(\mathbf{x}_\delta\right)\right| U^\dagger O_i U \left|\phi\left(\mathbf{x}_\delta\right)\right\rangle . \qquad (24)$$

Here, we assume that O_i is taken from $\mathcal{O}(\mathcal{H})$, a subset of the space of Hermitian operators of the Hilbert space $\mathcal{H}$, and the index i describes the ith component of the output, associated to the ith operator O_i. The above *Hermitian operator expectation value evaluation* model is a common definition of the quantum neural network. One could also measure the real and imaginary parts directly to define a complexified version of the quantum neural network, useful in the context of amplitude encoding for the $z_{i;\delta}$, as discussed within the Supplemental Material [31]. We are now in the position of introducing the loss function

$$L_{\mathcal{A}}(\theta) = \frac{1}{2}\sum_{\tilde{\alpha},i}\left(y_{i;\tilde{\alpha}} - z_{i;\tilde{\alpha}}\right)^2 = \frac{1}{2}\sum_{\tilde{\alpha},i}\varepsilon^2_{i;\tilde{\alpha}}. \qquad (25)$$

Here, we call $\varepsilon_{i;\tilde{\alpha}}$ the residual training error and we assume $y_{i;\tilde{\alpha}}$ is associated with the encoded data $\phi_i(\mathbf{x}_{\tilde{\alpha}})$. Now, similarly to what is described in Sec. II A, we have the gradient-descent equation

$$\text{đ}z_{i;\delta} = -\eta \sum_{\ell,i',\tilde{\alpha}} \varepsilon_{i';\tilde{\alpha}} \frac{dz_{i;\delta}}{d\theta_\ell}\frac{dz_{i';\tilde{\alpha}}}{d\theta_\ell}, \qquad (26)$$

with an associated kernel

$$K^{ii'}_{\delta,\tilde{\alpha}} = \sum_\ell \frac{dz_{i;\delta}}{d\theta_\ell}\frac{dz_{i';\tilde{\alpha}}}{d\theta_\ell}. \qquad (27)$$

To ease the notation, we define the joint index

$$(\delta, i) = \bar{a}, \quad (\tilde{\alpha}, i') = \hat{b}, \qquad (28)$$

which are running in the space $\mathcal{D}\times\mathcal{O}(\mathcal{H})$ and $\mathcal{A}\times\mathcal{O}(\mathcal{H})$, respectively, (we use $\hat{a}$ to indicate that the corresponding data component is in the sample set $\mathcal{A}$, and if we wish to make a general data point we denote it as $\bar{a}$), and our gradient-descent equations are

$$\text{đ}z_{\bar{a}} = -\eta\sum_{\hat{b}} K_{\bar{a}\hat{b}}\varepsilon_{\hat{b}}. \qquad (29)$$

It is possible to show that this kernel is always positive semidefinite and Hermitian, see Supplementary Material for a proof. Now recalling Eq. (1), we are in the position to give an analytical expression for the QNTK for a supervised learning problem as follows. Details on the derivation can be found within the Supplemental Material [31].

Definition 4 (QNTK for quantum machine learning)**.** *The QNTK for the quantum learning model, Eq. (25), is given by*

$$K^{ii'}_{\delta,\tilde{\alpha}} = \sum_\ell \frac{dz_{i;\delta}}{d\theta_\ell}\frac{dz_{i';\tilde{\alpha}}}{d\theta_\ell}$$
$$= -\sum_\ell \left(\begin{array}{l} \left\langle\phi\left(\mathbf{x}_\delta\right)\middle| U^\dagger_{+,\ell}\left[X_\ell, U^\dagger_\ell W^\dagger_\ell U^\dagger_{-,\ell} O_i U_{-,\ell} W_\ell U_\ell\right] U_{+,\ell}\middle|\phi\left(\mathbf{x}_\delta\right)\right\rangle \times \\ \left\langle\phi\left(\mathbf{x}_{\tilde{\alpha}}\right)\middle| U^\dagger_{+,\ell}\left[X_\ell, U^\dagger_\ell W^\dagger_\ell U^\dagger_{-,\ell} O_{i'} U_{-,\ell} W_\ell U_\ell\right] U_{+,\ell}\middle|\phi\left(\mathbf{x}_{\tilde{\alpha}}\right)\right\rangle \end{array}\right). \qquad (30)$$

B. Absence of representation learning in the frozen limit

In the frozen QNTK case, the kernel is static, and the learning algorithm cannot learn *features* from the data. In the same fashion of Sec. II B, we take *the frozen QNTK limit* where the changes of variational angles θ are small. Using the previous notations we can define the QNTK in for quantum machine learning in the frozen limit, and a performance guarantee for the error on the loss function in this regime as follows.

Definition 5 (Frozen QNTK for quantum machine learning)**.** *In the quantum learning model, Eq. (25), with the frozen QNTK limit,*

$$K^{ii'}_{\delta,\tilde{\alpha}} = -\delta^2\sum_\ell \left(\begin{array}{l} \left\langle\phi\left(\mathbf{x}_\delta\right)\middle| W^\dagger_{+,\ell}\left[X_\ell, W^\dagger_\ell W^\dagger_{-,\ell} O_i W_{-,\ell} W_\ell\right] W_{+,\ell}\middle|\phi\left(\mathbf{x}_\delta\right)\right\rangle \times \\ \left\langle\phi\left(\mathbf{x}_{\tilde{\alpha}}\right)\middle| W^\dagger_{+,\ell}\left[X_\ell, W^\dagger_\ell W^\dagger_{-,\ell} O_{i'} W_{-,\ell} W_\ell\right] W_{+,\ell}\middle|\phi\left(\mathbf{x}_{\tilde{\alpha}}\right)\right\rangle \end{array}\right). \qquad (31)$$

Theorem 3 (Performance guarantee of quantum machine learning in the frozen QNTK limit)**.** *In the quantum learning model, Eq. (25), with the frozen QNTK limit, the residual optimization error decays exponentially during the gradient descent as*

$$\begin{aligned}\varepsilon_{\hat{a}_1}(t) &= \sum_{\hat{a}_2} U_{\hat{a}_1\hat{a}_2}(t)\varepsilon_{\hat{a}_2}(0),\\ U_{\hat{a}_1\hat{a}_2}(t) &= \left[(1-\eta K)^t\right]_{\hat{a}_1\hat{a}_2}.\end{aligned} \tag{32}$$

The convergence rate is defined as

$$\tau_c = \|-\log(1-\eta K)\| \approx \eta \left\|K^{ii'}_{\delta,\tilde{\alpha}}\right\|. \tag{33}$$

Then we obtain for the quantum learning model, Eq. (25), with the frozen QNTK limit, the asymptotic dynamics with the $\mathcal{D}\times\mathcal{O}(\mathcal{H})$ index $\bar{a}$, is given by

$$z_{\bar{a}}(\infty) = z_{\bar{a}}(0) - \sum_{\hat{a}_1,\hat{a}_2} \tilde{K}^{\hat{a}_1\hat{a}_2} K_{\bar{a}\hat{a}_1}\varepsilon_{\hat{a}_2}(0). \tag{34}$$

Here $\tilde{K}$ means that the kernel defined only restricted to the space $\mathcal{A}\times\mathcal{O}(\mathcal{H})$ (note that it is different from the kernel inverse defined for the whole space in general), and we denote the kernel inverse as

$$\sum_{\hat{a}\in\mathcal{A}\times\mathcal{O}(\mathcal{H})} \tilde{K}^{\hat{a}_1\hat{a}_2}\tilde{K}_{\hat{a}_2\hat{a}_3} = \delta^{\hat{a}_1}_{\hat{a}_3}. \tag{35}$$

Specifically, if we assume $\bar{a}$ indicates the data in the space $\mathcal{A}\times\mathcal{O}(\mathcal{H})$, we have $\varepsilon_{\bar{a}}(\infty)=0$. Proofs and details of these results are given within the Supplemental Material [31]. Moreover, the asymptotic value is different from the frozen QNTK case in the optimization problem, because of the existence of the difference between the training set $\mathcal{A}$ and the whole data space.

C. Representation learning in the dynamical setting

In the dynamical case, the kernel is changing during the gradient-descent optimization, due to nonlinearity in the unitary operations. In this case then the variational quantum circuits could naturally serve as architectures of representation learning in the classical sense.

We generalize the leading-order perturbation theory of optimization naturally to the learning case, and we state the main theorems here. First, we have the following.

Theorem 4 (Performance guarantee of quantum machine learning in the DQNTK limit)**.** *In the quantum learning model, Eq. (25), at the DQNTK order, the training error is given by two contributions, a free and interacting part, as follows:*

$$\varepsilon_{\hat{a}}(t) = \varepsilon^F_{\hat{a}}(t) + \varepsilon^I_{\hat{a}}(t), \tag{36}$$

where

$$\begin{aligned}\varepsilon^F_{\hat{a}}(t) &= \sum_{\hat{a}_1} U_{\hat{a}\hat{a}_1}(t)\varepsilon_{\hat{a}_1}(0),\\ U_{\hat{a}_1\hat{a}_2}(t) &= \left[(1-\eta K)^t\right]_{\hat{a}_1\hat{a}_2},\end{aligned} \tag{37}$$

and

$$\varepsilon^I_{\hat{a}}(t) = \left(-\eta\sum_{s=0}^{t-1}(1-\eta K)^{t-1-s}K^{\Delta}(1-\eta K)^s\varepsilon(0)\right)_{\hat{a}}. \tag{38}$$

Here K is the frozen (linear) part of the QNTK. Using a matrix notation for the compact indices $\hat{a}$, in the space $\mathcal{A}\times\mathcal{O}(\mathcal{H})$, we have

$$\left\|\varepsilon^I(t)\right\| \le \eta t\|1-\eta K\|^{t-1}\left\|K^{\Delta}\right\|\|\varepsilon(0)\|, \tag{39}$$

where K^{Δ} is defined as

$$K^{\Delta,ii'}_{\delta,\tilde{\alpha}} = i\delta^3\sum_{\ell,\ell'}\varphi_{\ell'}(0)G^{\delta,i}_{\ell',\ell}\Theta^{\tilde{\alpha},i'}_{\ell} + i\delta^3\sum_{\ell,\ell'}\varphi_{\ell'}(0)G^{\tilde{\alpha},i'}_{\ell',\ell}\Theta^{\delta,i}_{\ell}, \tag{40}$$

and

$$\begin{aligned}G^{\delta,i}_{\ell_1,\ell_2} &\equiv G_{\ell_1,\ell_2}(\phi(\mathbf{x}_\delta),O_i)\\ &= \left(\begin{cases}\left\langle\phi(\mathbf{x}_\delta)\middle|W^\dagger_{+,\ell_1}\left[X_{\ell_1},Q^\dagger_{\ell_1,\ell_2}\left[X_{\ell_2},W^\dagger_{\ell_2}W^\dagger_{-,\ell_2}O_iW_{-,\ell_2}W_{\ell_2}\right]W_{\ell_2,\ell_1}W_{\ell_1}\right]W_{+,\ell_1}\middle|\phi(\mathbf{x}_\delta)\right\rangle : \ell_1\ge\ell_2\\ \left\langle\phi(\mathbf{x}_\delta)\middle|W^\dagger_{+,\ell_2}\left[X_{\ell_2},Q^\dagger_{\ell_2,\ell_1}\left[X_{\ell_1},W^\dagger_{\ell_1}W^\dagger_{-,\ell_1}O_iW_{-,\ell_1}W_{\ell_1}\right]W_{\ell_1,\ell_2}W_{\ell_2}\right]W_{+,\ell_2}\middle|\phi(\mathbf{x}_\delta)\right\rangle : \ell_1<\ell_2\end{cases}\right),\\ \Theta^{\delta,i}_{\ell} &\equiv \Theta_\ell(\phi(\mathbf{x}_\delta),O_i) = \left\langle\phi(\mathbf{x}_\delta)\middle|W^\dagger_{+,\ell}\left[X_\ell,W^\dagger_\ell W^\dagger_{-,\ell}O_iW_{-,\ell}W_\ell\right]W_{+,\ell}\middle|\phi(\mathbf{x}_\delta)\right\rangle.\end{aligned} \tag{41}$$

For the quantum learning model, Eq. (25), at the DQNTK order, the dynamics given by gradient descent on a general data point is given by

$$z_{\bar{a}}(\infty) = z_{\bar{a}}(0) - \sum_{\hat{a}_1,\hat{a}_2} K_{\bar{a}\hat{a}_1}\tilde{K}^{\hat{a}_1\hat{a}_2}\varepsilon_{\hat{a}_2}(0) + \sum_{\hat{a}_1,\hat{a}_2,\hat{a}_3,\hat{a}_4}\left[\mu_{\hat{a}_1\bar{a}\hat{a}_2} - \sum_{\hat{a}_5,\hat{a}_6} K_{\bar{a}\hat{a}_5}\tilde{K}^{\hat{a}_5\hat{a}_6}\mu_{\hat{a}_1\hat{a}_6\hat{a}_2}\right] Z_A^{\hat{a}_1\hat{a}_2\hat{a}_3\hat{a}_4}\varepsilon_{\hat{a}_3}(0)\varepsilon_{\hat{a}_4}(0)$$
$$+ \sum_{\hat{a}_1,\hat{a}_2,\hat{a}_3,\hat{a}_4}\left[\mu_{\bar{a}\hat{a}_1\hat{a}_2} - \sum_{\hat{a}_5,\hat{a}_6} K_{\bar{a}\hat{a}_5}\tilde{K}^{\hat{a}_5\hat{a}_6}\mu_{\hat{a}_6\hat{a}_1\hat{a}_2}\right] Z_B^{\hat{a}_1\hat{a}_2\hat{a}_3\hat{a}_4}\varepsilon_{\hat{a}_3}(0)\varepsilon_{\hat{a}_4}(0), \quad (42)$$

where $Z_{A,B}$s are called the quantum algorithm projectors (see Refs. [24,28] for their original framework),

$$Z_A^{\hat{a}_1\hat{a}_2\hat{a}_3\hat{a}_4} \equiv \tilde{K}^{\hat{a}_1\hat{a}_3}\tilde{K}^{\hat{a}_2\hat{a}_4} - \sum_{\hat{a}_5}\tilde{K}^{\hat{a}_2\hat{a}_5}X_{\parallel}^{\hat{a}_1\hat{a}_5\hat{a}_3\hat{a}_4},$$
$$Z_B^{\hat{a}_1\hat{a}_2\hat{a}_3\hat{a}_4} \equiv \tilde{K}^{\hat{a}_1\hat{a}_3}\tilde{K}^{\hat{a}_2\hat{a}_4} - \sum_{\hat{a}_5}\tilde{K}^{\hat{a}_2\hat{a}_5}X_{\parallel}^{\hat{a}_1\hat{a}_5\hat{a}_3\hat{a}_4} + \frac{\eta}{2}X_{\parallel}^{\hat{a}_1\hat{a}_2\hat{a}_3\hat{a}_4}, \quad (43)$$

and $X_{\parallel}$ is defined as

$$X_{\parallel}^{\hat{a}_1\hat{a}_2\hat{a}_3\hat{a}_4} = \sum_{s=0}^{\infty}[(1-\eta K)^s]_{\hat{a}_1\hat{a}_3}[(1-\eta K)^s]_{\hat{a}_2\hat{a}_4}, \quad (44)$$

or

$$\delta_{\hat{a}_5}^{\hat{a}_1}\delta_{\hat{a}_6}^{\hat{a}_2} = \sum_{\hat{a}_3,\hat{a}_4} X_{\parallel}^{\hat{a}_1\hat{a}_2\hat{a}_3\hat{a}_4} \times \left(\tilde{K}_{\hat{a}_3\hat{a}_5}\delta_{\hat{a}_4\hat{a}_6} + \delta_{\hat{a}_3\hat{a}_5}\tilde{K}_{\hat{a}_4\hat{a}_6} - \eta\tilde{K}_{\hat{a}_3\hat{a}_5}\tilde{K}_{\hat{a}_4\hat{a}_6}\right). \quad (45)$$

Finally, μ is the quantum metakernel in the quantum machine learning context,

$$\mu_{\delta_0\delta_1\delta_2}^{i_0 i_1 i_2} = \mu_{\bar{a}_0\bar{a}_1\bar{a}_2} = \sum_{\ell_1,\ell_2}\frac{d^2 z_{i_0;\delta_0}}{d\varphi_{\ell_1}d\varphi_{\ell_2}}\left(\frac{dz_{i_1;\delta_1}}{d\varphi_{\ell_1}}\frac{dz_{i_2;\delta_2}}{d\varphi_{\ell_2}}\right)\Bigg|_{\varphi=0}$$
$$= \delta^4\sum_{\ell_1,\ell_2}\Theta_{\ell_1}^{\delta_1,i_1}\Theta_{\ell_2}^{\delta_2,i_2}G_{\ell_1,\ell_2}^{\delta_0,i_0}. \quad (46)$$

Specifically, if we assume that $\bar{a}$ is from $\mathcal{A}\times\mathcal{O}(\mathcal{H})$, we will get $\varepsilon_{\bar{a}}(\infty) = 0$. More details of μ are given within the Supplemental Material [31]. The existence of quantum algorithm projectors shows the *quantum algorithm dependence* of the variational quantum circuits, which indicates powerful representation learning potential because of nonlinearity.

IV. HYBRID QUANTUM CLASSICAL NETWORK AND THE LARGE-WIDTH LIMIT

In this section we define a setting in which one can speak of a quantum analog of the large-width limit for NTKs. In such a limit, we expect that the dynamics linearizes during the whole training process, similar to what happens in the frozen regime of lazy training, and the correlation function of the outputs neurons becomes Gaussian. The classical NTK theory requires a random initialization of weights and bias and takes the large-width limit of neural network architectures. In the quantum setup, the random initialization is a random choice of trainable ansätze.

To see it more clearly, we consider a hybrid quantum classical neural network model [40,41]. Starting from a quantum neural network, we measure the output neurons from the quantum architecture and dress them with a single-layer classical neural network. The output of the classical neural network could be then re-encoded into a quantum register via another quantum feature map. A single quantum to classical step can be called one *hybrid layer*, and then one could construct multiple hybrid layers connected by feature map encoding, see Fig. 3 for an illustration.

For the quantum part of the circuit, we use the same structure of quantum neural networks with Hermitian operator expectation values. Mathematically, the model is

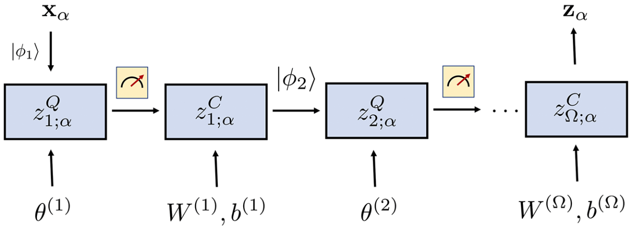

FIG. 3. The hybrid quantum classical neural network considered here. We repetitively apply quantum and classical neural networks in our architecture, with feature map encoding $|\psi\rangle$ and quantum measurements, mapping the data point $\mathbf{x}_\alpha$ to the prediction $\mathbf{z}_\alpha$.

defined as

$$z^{Q}_{1;\alpha;j_1}=\left\langle \phi_1\left(\mathbf{x}_\alpha\right)\left|U^{\dagger,1}\left(\theta^1\right)O^1_{j_1}U^1\left(\theta^1\right)\right|\phi_1\left(\mathbf{x}_\alpha\right)\right\rangle, \tag{47}$$

$$z^{Q}_{\omega;\alpha;j_\omega}=\left\langle \phi_\omega\left(\mathbf{w}_{\omega-1;\alpha}\right)\right| U^{\dagger,\omega}(\theta^\omega)O^\omega_{j_\omega}U^\omega(\theta^\omega)\left|\phi_\omega\left(\mathbf{w}_{\omega-1;\alpha}\right)\right\rangle, \tag{48}$$

$$w_{\omega;\alpha;j^C_\omega}=\sigma^\omega_{j^C_\omega}\left(\sum_{j_\omega=1}^{\dim\mathcal{O}^\omega(\mathcal{H}^\omega)}W^\omega_{j^C_\omega,j_\omega}z^Q_{\omega;\alpha;j_\omega}+b^\omega_{j^C_\omega}\right)\equiv \sigma^\omega_{j^C_\omega}\left(z^C_{\omega;\alpha;j^C_\omega}\right). \tag{49}$$

Here, Eq. (47) initializes the quantum neural network, mapping the data $\mathbf{x}_\alpha$ to components j_1, labeling the index in the space of Hermitian operator we use $\mathcal{O}^1(\mathcal{H}^1)$. The variational ansatz is similar to what we have discussed before, but they might be different in different layers. We use the label ω to denote the order of hybrid layers, ranging from 1 to the total number of hybrid layers. We introduce the quantum ansatz $U^\omega(\theta^\omega)=\prod_{\ell_\omega=1}^{L_\omega}W^\omega_{\ell_\omega}\exp\left(i\theta^\omega_{\ell_\omega}X^\omega_{\ell_\omega}\right)$, the feature map ϕ_ω, and the operator space $\mathcal{O}^\omega(\mathcal{H}^\omega)$ index j_ω. Equation (48) introduces the recursive encoding from the classical neural network data $\mathbf{w}_{\omega-1;\alpha}=(w_{\omega-1;\alpha})_{j^C_{\omega-1}}$ to the space $\mathcal{O}^\omega(\mathcal{H}^\omega)$, where the classical data vector $w_{\omega;\alpha;j^C_\omega}$ is obtained through a single-layer classical neural network with the nonlinear activation $\sigma^\omega_{j^C_\omega}$, weight matrix $W^\omega_{j^C_\omega,j_\omega}$, and bias vector $b^\omega_{j^C_\omega}$ with the classical index j^C_ω, and the preactivation $z^C_{\omega;\alpha;j^C_\omega}$. When we intialize the hybrid network, the classical weights and biases are statistically Gaussian following the LeCun parametrization [42],

$$\mathbb{E}\left(W^\omega_{j^C_{1,\omega},j_{1,\omega}}W^\omega_{j^C_{2,\omega},j_{2,\omega}}\right)=\delta_{j^C_{1,\omega},j^C_{2,\omega}}\delta_{j_{1,\omega},j_{2,\omega}}\frac{C^\omega_W}{\dim\mathcal{O}^\omega(\mathcal{H}^\omega)},$$
$$\mathbb{E}\left(b^\omega_{j^C_{1,\omega}}b^\omega_{j^C_{2,\omega}}\right)=\delta_{j^C_{1,\omega},j^C_{2,\omega}}C^\omega_b. \tag{50}$$

Note that in this case, the role of *width* in the large-width theory is replaced with the dimension of the operator space, $\dim(\mathcal{O}^\omega(\mathcal{H}^\omega))$. The value of the dimension (width) could be arbitrary in principle, but it is upper bounded by the square of the dimension of the Hilbert space, $\dim(\mathcal{O}^\omega(\mathcal{H}^\omega))\le\dim(\mathcal{H}^\omega)^2$, in the qubit system.

If we now assume that our quantum training parameters θ^ω are chosen from ensembles (or the variational ansätze themselves are from some ensembles), similar to the classical assumption. Denoting the expectation value from quantum ensembles as $\mathbb{E}$, we show the following statement.

Theorem 5 (Non-Gaussianity from large width)**.** *The four-point function of classical preactivations is nearly Gaussian if* $\dim(\mathcal{O}^\omega(\mathcal{H}^\omega))$ *is large,*

$$\mathbb{E}_{\text{conn}}\left(z^C_{\omega;\alpha_1;j^C_{1,\omega}}z^C_{\omega;\alpha_2;j^C_{2,\omega}}z^C_{\omega;\alpha_3;j^C_{3,\omega}}z^C_{\omega;\alpha_4;j^C_{4,\omega}}\right)=\mathcal{O}\left(\frac{1}{\dim(\mathcal{O}^\omega(\mathcal{H}^\omega))}\right), \tag{51}$$

as long as,

$$\mathbb{E}_{\text{conn}}\left(z^Q_{\omega;\alpha_1;j_{1,\omega}}z^Q_{\omega;\alpha_2;j_{1,\omega}}z^Q_{\omega;\alpha_3;j_{2,\omega}}z^Q_{\omega;\alpha_4;j_{2,\omega}}\right)=\mathcal{O}(1)\times\delta_{j_{1,\omega},j_{2,\omega}}, \tag{52}$$

and their permutations for all ωs. *Here the notation* $\mathbb{E}_{\text{conn}}$ *means the connected Gaussian correlators subtracting Wick contractions.*

More details are given within the Supplemental Material [31]. The orthogonal condition Eq. (52) can be naturally achieved by randomized architectures, for instance, Haar randomness and k designs. We interpret the result as follows:

(a) In this hybrid case, the role of *width* in the neural network is upper bounded by the square dimension of the wth Hilbert space. Thus, if we scale up the number of qubits, we are naturally in the large-width limit. However, if our variational ansatz is sparse enough such that the operator space dimension $\dim\mathcal{O}(\mathcal{H})$ is small, then we will have significant finite width effects.

(b) The condition, Eq. (52), for quantum outputs is naturally satisfied by random architectures. If we assume that our variational ansatz is highly random, we are expected to have similar Gaussian process behaviors as the large-width limit of classical neural networks. However, the same assumption will generically lead to the barren plateau problem [8], where the derivatives of the loss function will move slowly when we scale up our operator space dimension. Our result shows a possible connection between the large-width limit and the barren plateau problem.

(c) Moreover, the orthogonal condition in Eq. (52) we impose does not mean that we have to set the ansatz to be highly random. It could also be potentially satisfied by fixed ansätze, for instance, with some error-correction types of orthogonal conditions. If the condition is generally not satisfied, the variational architecture we study could have highly non-Gaussian and representation learning features, although it might be theoretically hard to understand [43].

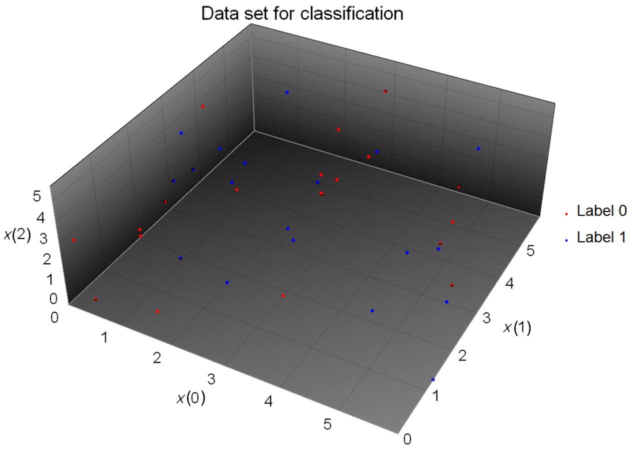

FIG. 4. Three-dimensional illustration of our data set for training. Here we use three inputs $x[0, 1, 2]$ and label them with 0 or 1, colored by red or blue, respectively. The data set is generated using `ad_hoc_data`.

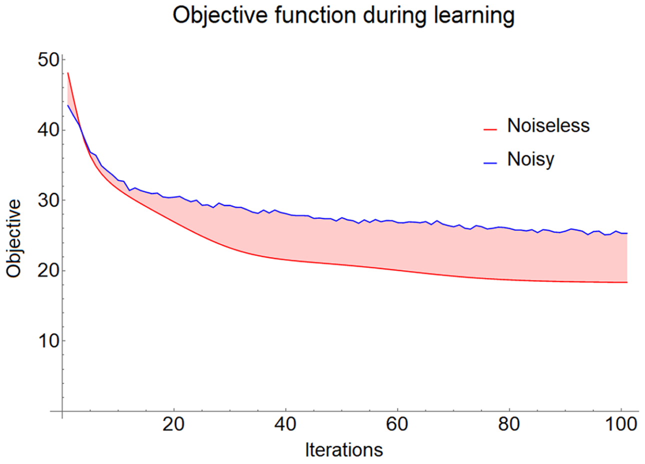

FIG. 5. Convergence of the objective function during gradient descent. Here we compare the ideal and noisy cases, labeled by red or blue, respectively.

V. NUMERICAL RESULTS

In this section, we test our QNTK theory in practice, using the Qiskit software library [30] to simulate the implementation of a paradigmatic quantum machine-learning task on quantum processors, both in noiseless and noisy cases. We consider a variational classification problem in supervised learning with three qubits. The data set is generated with the `ad_hoc_data` functionality as provided in `qiskit.ml.datasets` within the Qiskit Machine Learning module [44], see Fig. 4 for an illustration.

Our numerical experiments are performed first using the noiseless `statevector_simulator` backend, then including both statistical ($n_{\text{shots}} = 8192$) and simulated hardware noise with the Qiskit `qasm_simulator`. A simplified model of device noise, featuring the qubit relaxation and dephasing, single-qubit and two-qubit gate errors and readout inaccuracies, is constructed with the `NoiseModel.from_backend()` Qiskit method and parametrized using our calibration data from the `ibmq_bogota` superconducting processor (accessed on October, 15 2021).

We implement supervised learning using a Qiskit Machine Learning `NeuralNetworkClassifier` with a squared error loss, obtaining reasonable convergence with gradient-descent algorithms (see Fig. 5). The underlying variational quantum classifier is based on the `TwoLayerQNN` design, with a three input `ZZFeatureMap` and a `RealAmplitudes` trainable ansatz with three repetitions and 12 parameters. Further details on numerical simulations are given within the Supplemental Material [31]. Note that we do not demand a perfect convergence around the global minimum, since the QNTK theory only cares about the derivatives of the residual learning errors, which is invariant by shifting a constant or changing the initial condition when solving the training dynamics. In the classical theory of NTK, in the infinite width case, for instance, the multilayer perceptron (MLP) model is both overparametrized and generalized, and the answer would give the global minimum. Including finite width corrections, there might be multiple local minima, and it is a feature of representation learning. Moreover, we use the error-mitigation protocol by applying `CompleteMeasFitter` from `qiskit.ignis.mitigation.measurement` to mitigate readout noise.

In Fig. 6, we compute the QNTK eigenvalues for both the noiseless and noisy simulations, comparing them with theoretical predictions. Since we are in the underparametrized regime, the number of nonzero eigenvalues of the QNTK is the same as the number of variational angles, which is 12 in our experiments. We find agreement between those two in the late time, which shows the power of predictability using the QNTK theory.

Here we give an additional summary based on our numerical analysis. Physically, there are two differences between quantum and classical in terms of DNTK. First, non-Gaussianity could be generated if the quantum part of the neural networks is not orthogonal enough, and the orthogonality is likely caused by Haar randomness or quantum error-correction conditions, see Refs. [45–47]. Secondly, the effect of quantum noises is significant in the near-term quantum devices, and we can clearly see the correction between quantum and classical in Fig. 6 as numerical examples.

Finally, we wish to address the limitation of our theories. In fact, our theory is limited by the nature of the perturbative method, and many of the quantum machine-learning dynamics could be highly nonlinear. Furthermore,

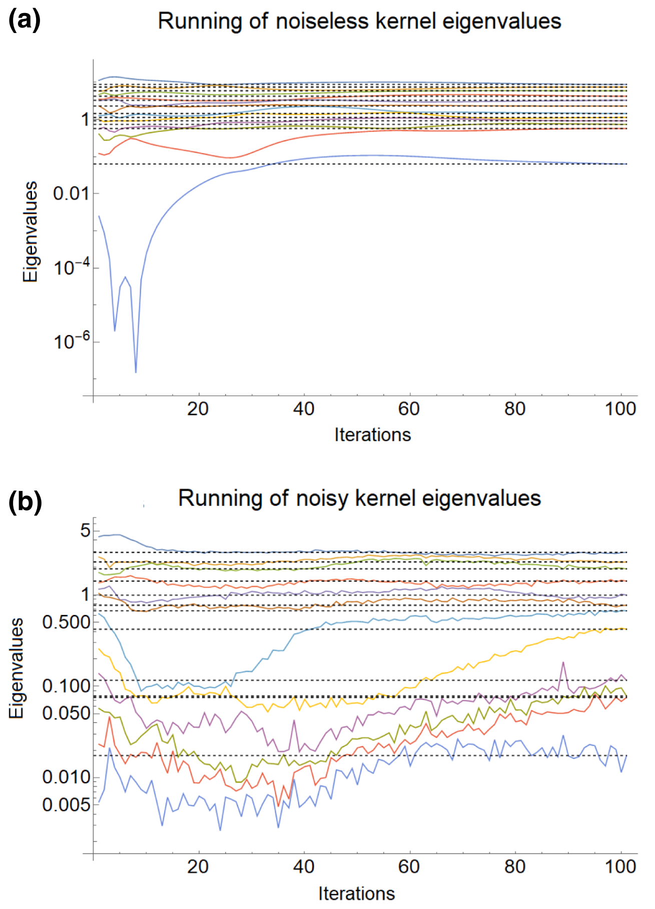

FIG. 6. Kernel eigenvalues during the gradient-descent dynamics. Up: noiseless simulation. Down: noisy simulation including a model of device errors. The solid curves are 12 nonzero eigenvalues of the QNTK, while the dashed lines are theoretical predictions of the frozen QNTK at the late time.

the presence of noise on current quantum processors will also limit the scope of the theory, and it is worth trying to explore further in experiments and theories about how useful the theory is in practice.

VI. CONCLUSIONS AND OPEN PROBLEMS

The results presented here establish a general framework of a QNTK theory, deriving analytical treatment of the optimization and learning dynamics in that regime. We outline the following open problems for future study.

(a) Our research gives practical guidance to the design variational quantum algorithms. One could compute the QNTK, or the kernel itself in the quantum kernel method [11]. It will be interesting to compare those results with other theoretical criteria about the quality of the variational quantum algorithms [48,49]. In fact, higher eigenvalues of neural tangent kernels lead to faster convergences and fewer generalization errors with good alignment with the target function we want to fit [50–55]. It will be interesting to explore in practice if the analytical assumptions made here on the large-width limit and small nonlinearity hold for practical use cases with a number of parameters scaling polynomially with the system size (see classical analogs [56]).

(b) It would be interesting if one could investigate when the frozen QNTK limit is useful in other contexts. In Trotter product formulas, for instance,

$$\lim_{n\to\infty}\left(e^{iaX/n}e^{ibZ/n}\right)^n, \tag{53}$$

to implement the gate $U = e^{i(aX+bZ)}$. Thus, small variational angles might widely appear in real cases of quantum architectures, even beyond the regime of lazy training around convergence.

(c) Connection to the barren plateau problem (see Refs. [46,47]). Our work suggests a possible connection between the barren plateau problem in variational quantum algorithms and the large-width limit in classical neural networks, by observing the following similarity between the LeCun parametrization above

$$\mathbb{E}\left(W_{j_1^C j_1} W_{j_2^C j_2}\right) = \delta_{j_1^C j_2^C}\delta_{j_1,j_2}\frac{C_W}{\text{width}}, \tag{54}$$

and the 1-design random formula [57]

$$\mathbb{E}(U_{ij}U^{\dagger}_{kl}) = \frac{\delta_{il}\delta_{jk}}{\dim\mathcal{H}}. \tag{55}$$

(d) It will be interesting to explore the robustness of the QNTK theory against noise. Specifically, we have obtained an exponential convergence when the QNTK is frozen.

(e) Finally, it will be interesting to explore the nonlinear regime where the perturbative analysis fails. One could draw phase diagrams of quantum machine learning about some *order parameters*, for instance, the learning rate [38]. Those studies will deepen our theoretical understanding of quantum machine learning. More directions include comparisons between the classical and quantum cases, exploring non-Gaussianity in and out of the large-width limit, and exploring the consequences when the orthogonality condition is not met.

ACKNOWLEDGMENTS

We thank Bryan Clark, Cliff Cheung, Jens Eisert, Thomas Faulkner, Guy Gur-Ari, Yoni Kahn, Risi Kondor, David Meltzer, Ryan Mishmash, John Preskill, Dan

A. Roberts, David Simmons-Duffin, Kristan Temme, Lian-Tao Wang, Rebecca Willett, and Changchun Zhong for useful discussions. Specifically, J.L. would also acknowledge conversations with Dan A. Roberts on clarifying his book and lecture [25,28]. J.L. is supported in part by International Business Machines (IBM) Quantum through the Chicago Quantum Exchange, and the Pritzker School of Molecular Engineering at the University of Chicago through AFOSR MURI (FA9550-21-1-0209). L.J. acknowledges support from the ARO (W911NF-18-1-0020, W911NF-18-1-0212), ARO MURI (W911NF-16-1-0349), AFOSR MURI (FA9550-19-1-0399, FA9550-21 -1-0209), DoE Q-NEXT, NSF (EFMA-1640959, OMA-1936118, EEC-1941583), NTT Research, and the Packard Foundation (2013-39273).

Note added.—A recent independent paper [58] that addresses the same topic has been posted publicly on arXiv.

[1] A. W. Harrow, A. Hassidim, and S. Lloyd, Quantum Algorithm for Linear Systems of Equations, Phys. Rev. Lett. **103**, 150502 (2009).

[2] N. Wiebe, D. Braun, and S. Lloyd, Quantum Algorithm for Data Fitting, Phys. Rev. Lett. **109**, 050505 (2012).

[3] S. Lloyd, M. Mohseni, and P. Rebentrost, Quantum principal component analysis, Nat. Phys. **10**, 631 (2014).

[4] P. Wittek, *Quantum Machine Learning: What Quantum Computing Means to Data Mining* (Academic Press, Cambridge, USA, 2014), https://www.amazon.com/Quantum-Machine-Learning-Computing-Mining/dp/0128100400.

[5] N. Wiebe, A. Kapoor, and K. M. Svore, Quantum deep learning, arXiv preprint arXiv:1412.3489 (2014).

[6] P. Rebentrost, M. Mohseni, and S. Lloyd, Quantum Support Vector Machine for Big Data Classification, Phys. Rev. Lett. **113**, 130503 (2014).

[7] J. Biamonte, P. Wittek, N. Pancotti, P. Rebentrost, N. Wiebe, and S. Lloyd, Quantum machine learning, Nature **549**, 195 (2017).

[8] J. R. McClean, S. Boixo, V. N. Smelyanskiy, R. Babbush, and H. Neven, Barren plateaus in quantum neural network training landscapes, Nat. Commun. **9**, 1 (2018).

[9] M. Schuld and N. Killoran, Quantum Machine Learning in Feature Hilbert Spaces, Phys. Rev. Lett. **122**, 040504 (2019).

[10] E. Tang, in *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing* (2019), p. 217, https://dl.acm.org/doi/10.1145/3313276.3316310.

[11] V. Havlíček, A. D. Córcoles, K. Temme, A. W. Harrow, A. Kandala, J. M. Chow, and J. M. Gambetta, Supervised learning with quantum-enhanced feature spaces, Nature **567**, 209 (2019).

[12] H.-Y. Huang, M. Broughton, M. Mohseni, R. Babbush, S. Boixo, H. Neven, and J. R. McClean, Power of data in quantum machine learning, Nat. Commun. **12**, 1 (2021).

[13] Y. Liu, S. Arunachalam, and K. Temme, A rigorous and robust quantum speed-up in supervised machine learning, Nat. Phys. **17**, 1013 (2021).

[14] H.-Y. Huang, R. Kueng, and J. Preskill, Information-Theoretic Bounds on Quantum Advantage in Machine Learning, Phys. Rev. Lett. **126**, 190505 (2021).

[15] J. Lee, Y. Bahri, R. Novak, S. S. Schoenholz, J. Pennington, and J. Sohl-Dickstein, Deep neural networks as Gaussian processes, arXiv preprint arXiv:1711.00165 (2017).

[16] A. Jacot, F. Gabriel, and C. Hongler, Neural tangent kernel: Convergence and generalization in neural networks, arXiv preprint arXiv:1806.07572 (2018).

[17] J. Lee, L. Xiao, S. Schoenholz, Y. Bahri, R. Novak, J. Sohl-Dickstein, and J. Pennington, Wide neural networks of any depth evolve as linear models under gradient descent, Adv. Neural Inf. Process. Syst. **32**, 8572 (2019).

[18] S. Arora, S. S. Du, W. Hu, Z. Li, R. Salakhutdinov, and R. Wang, On exact computation with an infinitely wide neural net, arXiv preprint arXiv:1904.11955 (2019).

[19] J. Sohl-Dickstein, R. Novak, S. S. Schoenholz, and J. Lee, On the infinite width limit of neural networks with a standard parameterization, arXiv preprint arXiv:2001.07301 (2020).

[20] G. Yang and E. J. Hu, Feature learning in infinite-width neural networks, arXiv preprint arXiv:2011.14522 (2020).

[21] S. Yaida, in *Mathematical and Scientific Machine Learning* (PMLR, 2020), p. 165, https://proceedings.mlr.press/v107/yaida20a.html.

[22] E. Dyer and G. Gur-Ari, Asymptotics of wide networks from Feynman diagrams, arXiv preprint arXiv:1909.11304 (2019).

[23] J. Halverson, A. Maiti, and K. Stoner, Neural networks and quantum field theory, Machine Learning: Sci. Technol. **2**, 035002 (2021).

[24] D. A. Roberts, Why is AI hard and physics simple?, arXiv preprint arXiv:2104.00008 (2021).

[25] D. A. Roberts, S. Yaida, and B. Hanin, The principles of deep learning theory, arXiv preprint arXiv:2106.10165 (2021).

[26] J. Liu, Ph.D. thesis, Caltech, 2021.

[27] K. Nakaji, H. Tezuka, and N. Yamamoto, Quantum-enhanced neural networks in the neural tangent kernel framework, arXiv:2109.03786 [quant-ph] (2021).

[28] D. A. Roberts and S. Yaida, Effective theory of deep learning: Beyond the infinite-width limit, Deep Learning Theory Summer School at Princeton (2021).

[29] L. Chizat, E. Oyallon, and F. Bach, On lazy training in differentiable programming, arXiv preprint arXiv:1812.07956 (2018).

[30] G. Aleksandrowicz *et al.*, Qiskit: An open-source framework for quantum computing (2019).

[31] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PRXQuantum.3.030323 for we provide technical details, proofs, and extensions of the quantum neural tangent kernel theory, which includes Ref. [7].

[32] A. Peruzzo, J. McClean, P. Shadbolt, M.-H. Yung, X.-Q. Zhou, P. J. Love, A. Aspuru-Guzik, and J. L. O'brien, A variational eigenvalue solver on a photonic quantum processor, Nat. Commun. **5**, 1 (2014).

[33] E. Farhi, J. Goldstone, and S. Gutmann, A quantum approximate optimization algorithm, arXiv preprint arXiv:1411.4028 (2014).

[34] J. R. McClean, J. Romero, R. Babbush, and A. Aspuru-Guzik, The theory of variational hybrid quantum-classical algorithms, New J. Phys. **18**, 023023 (2016).

[35] A. Kandala, A. Mezzacapo, K. Temme, M. Takita, M. Brink, J. M. Chow, and J. M. Gambetta, Hardware-efficient variational quantum eigensolver for small molecules and quantum magnets, Nature **549**, 242 (2017).

[36] S. McArdle, S. Endo, A. Aspuru-Guzik, S. C. Benjamin, and X. Yuan, Quantum computational chemistry, Rev. Mod. Phys. **92**, 015003 (2020).

[37] M. Cerezo, A. Arrasmith, R. Babbush, S. C. Benjamin, S. Endo, K. Fujii, J. R. McClean, K. Mitarai, X. Yuan, and L. Cincio *et al.*, Variational quantum algorithms, Nat. Rev. Phys., 1 (2021).

[38] A. Lewkowycz, Y. Bahri, E. Dyer, J. Sohl-Dickstein, and G. Gur-Ari, The large learning rate phase of deep learning: the catapult mechanism, arXiv preprint arXiv:2003.02218 (2020).

[39] D. Meltzer, H. Zheng, Y.-H. Du, D. R. Roberts, and J. Liu, To appear.

[40] J. Otterbach, R. Manenti, N. Alidoust, A. Bestwick, M. Block, B. Bloom, S. Caldwell, N. Didier, E. S. Fried, and S. Hong *et al.*, Unsupervised machine learning on a hybrid quantum computer, arXiv preprint arXiv:1712.05771 (2017).

[41] E. Farhi and H. Neven, Classification with quantum neural networks on near term processors, arXiv preprint arXiv:1802.06002 (2018).

[42] This is somewhat different from the so-called *NTK parametrization* in some literature. See details within the Supplemental Material [31]. Moreover, here we assume the weights and biases are real.

[43] Similar analysis could be done on dynamics, see the Supplemental Material [31] for further comments.

[44] Qiskit machine learning, https://qiskit.org/documentation/machine-learning/ (2021).

[45] J. Liu, Scrambling and decoding the charged quantum information, Phys. Rev. Res. **2**, 043164 (2020).

[46] J. Liu, K. Najafi, K. Sharma, F. Tacchino, L. Jiang, and A. Mezzacapo, An analytic theory for the dynamics of wide quantum neural networks, arXiv:2203.16711 [quant-ph] (2022).

[47] J. Liu, Z. Lin, and L. Jiang, Laziness, barren plateau, and noise in machine learning, arXiv:2206.09313 [cs.LG] (2022).

[48] A. Abbas, D. Sutter, C. Zoufal, A. Lucchi, A. Figalli, and S. Woerner, The power of quantum neural networks, Nat. Comput. Sci. **1**, 403 (2021).

[49] T. Haug, K. Bharti, and M. Kim, Capacity and quantum geometry of parametrized quantum circuits, arXiv preprint arXiv:2102.01659 (2021).

[50] J. Liu, C. Zhong, M. Otten, C. L. Cortes, C. Ti, S. K. Gray, and X. Han, Quantum Kerr learning, arXiv preprint arXiv:2205.12004 (2022).

[51] B. Bordelon, A. Canatar, and C. Pehlevan, in *International Conference on Machine Learning* (PMLR, 2020), p. 1024.

[52] A. Canatar, B. Bordelon, and C. Pehlevan, Spectral bias and task-model alignment explain generalization in kernel regression and infinitely wide neural networks, Nat. Commun. **12**, 1 (2021).

[53] J. B. Simon, M. Dickens, and M. R. DeWeese, Neural tangent kernel eigenvalues accurately predict generalization, arXiv preprint arXiv:2110.03922 (2021).

[54] Y. Bahri, E. Dyer, J. Kaplan, J. Lee, and U. Sharma, Explaining neural scaling laws, arXiv preprint arXiv:2102.06701 (2021).

[55] A. Atanasov, B. Bordelon, and C. Pehlevan, Neural networks as kernel learners: The silent alignment effect, arXiv preprint arXiv:2111.00034 (2021).

[56] G. Yang, E. J. Hu, I. Babuschkin, S. Sidor, X. Liu, D. Farhi, N. Ryder, J. Pachocki, W. Chen, and J. Gao, Tensor programs v: Tuning large neural networks via zero-shot hyperparameter transfer, arXiv preprint arXiv:2203.03466 (2022).

[57] D. A. Roberts and B. Yoshida, Chaos and complexity by design, JHEP **04**, 121 (2017).

[58] N. Shirai, K. Kubo, K. Mitarai, and K. Fujii, Quantum tangent kernel, arXiv:2111.02951 [quant-ph] (2021).