Analytic Theory for the Dynamics of Wide Quantum Neural Networks

Junyu Liu,[1,2,3,*] Khadijeh Najafi,[4,†] Kunal Sharma,[4,5,‡] Francesco Tacchino,[6,§] Liang Jiang,[1,2,‖] and Antonio Mezzacapo[4,¶]

[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 T. J. Watson Research Center, Yorktown Heights, New York 10598, USA*
[5]*Joint Center for Quantum Information and Computer Science, University of Maryland, College Park, Maryland 20742, USA*
[6]*IBM Quantum, IBM Research, Zurich, 8803 Rüschlikon, Switzerland*

(Received 22 April 2022; revised 11 November 2022; accepted 2 March 2023; published 10 April 2023)

Parametrized quantum circuits can be used as quantum neural networks and have the potential to outperform their classical counterparts when trained for addressing learning problems. To date, much of the results on their performance on practical problems are heuristic in nature. In particular, the convergence rate for the training of quantum neural networks is not fully understood. Here, we analyze the dynamics of gradient descent for the training error of a class of variational quantum machine learning models. We define wide quantum neural networks as parametrized quantum circuits in the limit of a large number of qubits and variational parameters. Then, we find a simple analytic formula that captures the average behavior of their loss function and discuss the consequences of our findings. For example, for random quantum circuits, we predict and characterize an exponential decay of the residual training error as a function of the parameters of the system. Finally, we validate our analytic results with numerical experiments.

DOI: 10.1103/PhysRevLett.130.150601

Machine learning has revolutionized data processing for several practical applications. With the abundance of data and computational resources, heuristic deep learning algorithms have been successfully employed for several applications, including speech recognition, translation, drug discovery, genomics, and self-driving cars [1]. The theory of deep learning owes part of its successes to analytic insights that facilitate the design of learning algorithms [2–8].

Recent experimental progress on quantum hardware and algorithms has generated great excitement in trying to identify applications that can lead to a quantum advantage over classical devices [9–12]. One such application is quantum machine learning (QML) which employs parametrized quantum circuits to analyze either classical or quantum data [13–15]. Contrary to the classical setting, where experiments are routinely performed on large-scale problems, current quantum processors are limited both in number of qubits and by decoherence noise [16], which makes it challenging to test QML algorithms in practice. Analytic tools are currently among the best resources that can help us quantify the performance of QML models and design new algorithms [17].

Similar to classical neural networks, several models for quantum neural networks (QNNs) have been proposed [18–22]. The success of QNNs relies on several factors: trainability, expressivity, generalization, and convergence rate. Although trainability [23–34], expressivity [29,35–38], and generalization capabilities [10,37,39–41] of QNNs have been extensively studied, analytic understanding of the convergence rate of the training error is still lacking.

In this Letter, we present an analytic theory for the dynamics of a wide QNN trained with gradient descent, in the limit of a large number of parameters. Our results are based on the framework of the quantum neural tangent kernel (QNTK), recently developed in Refs. [42,43]. Of particular interest, here, is the lazy training regime where the QNTK becomes constant—or frozen, see Ref. [42].

Using the QNTK framework, we calculate the behavior of the residual training error for random parametrized quantum circuits. In the high-dimensional limit and for a sufficiently large number of variational parameters, we find an analytic solution characterizing the convergence of the residual training error. We denote the rate of convergence as γ and show its dependence on the number of variational parameters L, the dimension of the Hilbert space D, learning rate η, and $\mathrm{Tr}(O^2)$ for an observable O. High values of γ imply that a QNN is in the overparametrized regime, leading to an exponential convergence of the residual error on average. We note that, prior to our analytic results, overparametrization in QNNs was only numerically investigated for some systems in [44,45] and connected to the dimension of the dynamical Lie algebra associated with periodic structure *Ansätze* [46].

The Letter is organized as follows: first, we review the QNTK theory. Using the QNTK, we derive an analytic

solution characterizing the convergence of the residual training error. Then, we derive conditions on the parameters of the system, such that the residual error decays exponentially, and extend our results to general supervised learning problems. Finally, we provide numerics verifying our results. We conclude with a brief summary and discuss the implications of our results. In our Supplemental Material [47], we provide detailed proofs of our results.

Quantum neural tangent kernel.—We begin by reviewing the QNTK theory as described in Ref. [42]. Let D denote the dimension of a Hilbert space $\mathcal{H}$. We consider a general class of parametrized quantum circuits on $\log(D)$ qubits, expressed as follows:

$$U(\vec{\theta}) = \prod_{\ell=1}^{L} W_\ell \exp(i\theta_\ell X_\ell) \equiv \prod_{\ell=1}^{L} W_\ell U_\ell, \tag{1}$$

where $\vec{\theta} = \{\theta_\ell\}_{\ell=1}^{L}$ is a set of continuous parameters, W_ℓ denote unparametrized gates, and X_ℓ are Hermitian operators. Here, $\vec{\theta}$ are optimized to minimize a loss function that can be expressed as the expectation value of an observable O

$$\mathcal{L}(\vec{\theta}) \equiv \frac{1}{2}(\langle\Psi_0|U^\dagger(\vec{\theta})OU(\vec{\theta})|\Psi_0\rangle - O_0)^2 \equiv \frac{1}{2}\varepsilon^2, \tag{2}$$

where $|\psi_0\rangle$ is an input state, O_0 denotes the target value, and ε denotes the residual error [42].

We note that, in Eq. (2), we start with a simpler problem than a general supervised learning task where one has access to a labeled dataset. In general, the loss function for a general supervised learning task is given by

$$\mathcal{L}_{\mathcal{A}}(\vec{\theta}) = \sum_{i,\tilde{\alpha}\in\mathcal{A}} \frac{1}{2}(z_i(\vec{\theta}, \mathbf{x}_{\tilde{\alpha}}) - y_{\tilde{\alpha},i})^2 \equiv \sum_{i,\tilde{\alpha}\in\mathcal{A}} \frac{1}{2}\varepsilon_{\tilde{\alpha},i}^2(\vec{\theta}), \tag{3}$$

where $\tilde{\alpha}$ labels the elements from the training set $\mathcal{A}$, $\mathbf{x}_{\tilde{\alpha}}$ and $y_{\tilde{\alpha},i}$ form the data inputs and outputs, respectively, in the training set, where the output dimension has the index i. $z_i(\mathbf{x}_{\tilde{\alpha}}) = \langle\Psi(\mathbf{x}_{\tilde{\alpha}})|U^\dagger(\vec{\theta})O_iU(\vec{\theta})|\Psi(\mathbf{x}_{\tilde{\alpha}})\rangle$ is the model output with the embedding map $|\Psi(\mathbf{x}_{\tilde{\alpha}})\rangle$, and $\varepsilon_{\tilde{\alpha},i} = z_i(\vec{\theta}, \mathbf{x}_{\tilde{\alpha}}) - y_{\tilde{\alpha},i}$ is the residual training error.

Below, we provide a detailed summary of our results for Eq. (2) and briefly discuss our results for Eq. (3). We provide detailed proofs for both cases in the Supplemental Material [47]. Based on the gradient of the loss function in Eq. (2), the gradient descent algorithm updates the variational parameters as

$$\delta\theta_\ell \equiv \theta_\ell(t+1) - \theta_\ell(t) = -\eta\varepsilon\frac{\partial\varepsilon}{\partial\theta_\ell}, \tag{4}$$

where η is the learning rate and t refers to the time step of the gradient descent dynamics. Similarly, we define the change in the residual training error as $\delta\varepsilon \equiv \varepsilon(t+1) - \varepsilon(t)$. When the learning rate η is small, from the Taylor expansion of $\delta\varepsilon$, we get

$$\delta\varepsilon \approx \sum_\ell \frac{\partial\varepsilon}{\partial\theta_\ell}\delta\theta_\ell = -\eta\sum_\ell \frac{\partial\varepsilon}{\partial\theta_\ell}\frac{\partial\varepsilon}{\partial\theta_\ell}\varepsilon = -\eta K\varepsilon, \tag{5}$$

where the quantity

$$K \equiv \sum_\ell \frac{\partial\varepsilon}{\partial\theta_\ell}\frac{\partial\varepsilon}{\partial\theta_\ell} \tag{6}$$

is called the quantum neural tangent kernel [42], which is a non-negative number. In a general supervised learning setting, as defined in Eq. (3), K is a symmetric positive-semidefinite matrix.

In the regime of lazy training—where variational angles do not change much—QNTK becomes constant (frozen) [42]. For a frozen QNTK, at the gradient descent step t, the residual error decays as follows [42]:

$$\varepsilon(t) \approx (1 - \eta K)^t\varepsilon(0), \tag{7}$$

where $\varepsilon(0)$ denotes the residual error at $t = 0$. Thus, for a small learning rate η and a frozen QNTK, the residual error ε decays exponentially.

First, we analyze Eq. (7) for the case when $K \approx \mathbb{E}(K) \equiv \bar{K}$, where the average of K is over $\vec{\theta}$. Later, we derive conditions under which Eq. (7) is valid. Note that an average of K over $\vec{\theta}$ depends on the choice of the *Ansatz*, as defined in Eq. (1). For such *Ansätze*, $\partial\varepsilon/\partial\theta_l$ can be expressed as

$$\partial\varepsilon/\partial\theta_l = -i\langle\Psi_0|U_{+,\ell}^\dagger[X_\ell, U_{-,\ell}^\dagger OU_{-,\ell}]U_{+,\ell}|\Psi_0\rangle, \tag{8}$$

where $U_{-,\ell} \equiv \prod_{k=1}^{\ell} W_kU_k$ and $U_{+,\ell} \equiv \prod_{k=\ell+1}^{L} W_kU_k$. Let $\mathrm{Tr}(X_l^2) = cN$, where c is a constant.

Now, we derive our results on the residual training error of random parametrized quantum circuits. Note that our results can be generalized to other *Ansätze*, and later, we discuss the relevance of our results for periodic structure *Ansätze* [30,46]. Suppose that $U(\vec{\theta})$ is sufficiently random, such that for each l, both $U_{-,l}$ and $U_{+,l}$ are independent and match the Haar distribution up to the second moment. Then, we get the following averaged value of K in the large D limit [47]:

$$\bar{K} \approx \frac{L\mathrm{Tr}(O^2)}{D^2}, \tag{9}$$

which implies that, on average, the residual training error decays as

$$\varepsilon(t) \approx e^{-\gamma t}\varepsilon(0), \tag{10}$$

where the decay rate is given by

$$\gamma \equiv \eta \bar{K} = [\eta L \mathrm{Tr}(O^2)]/D^2. \tag{11}$$

Note that Eq. (10) holds for $\eta \ll 1$, as we have discarded higher-order terms in Eq. (5). Later, we show that, in the large-D limit, the second order term can be discarded even for high values of η.

The average result Eq. (9) follows for 2-design random circuits which can be implemented using one-dimensional $\mathcal{O}[\log^2(D)]$ gates [48]. Thus, for such efficiently implementable circuits, Eq. (10) provides an analytic solution to the average behavior of the residual training error. For circuits that approximate a 4-design, fluctuations in K from $\bar{K}$ are also small.

In general, the decay rate γ is small because of the $1/D^2$ dependence on the dimension of the Hilbert space. By setting $L \approx D^2/[\eta \mathrm{Tr}(O^2)]$, the residual training error decays exponentially with the decay with rate $\gamma = \mathcal{O}(1)$. This leads us to define the overparametrized regime for a QNN: A QNN is overparametrized if the number of parameters of the system are sufficiently large such that $\gamma = \mathcal{O}(1)$.

Now, we discuss two cases for overparametrized random quantum circuits. For $\mathrm{Tr}(O^2) \in \mathcal{O}(D)$, which holds for physical Hamiltonians that can be expressed as a linear combination of Pauli operators on $\log D$ qubits, $L \sim D$ is sufficient for making the corresponding QNN overparametrized. Similarly, for low-rank observables, $\mathrm{Tr}(O^2) \in \mathcal{O}[\log(D)]$ which implies that $L \sim D^2$ make the corresponding QNN overparametrized.

Prior to our work, the overparametrization of a QNN was first numerically observed in Ref. [44], where the authors investigated the task of learning Haar random unitaries $U(D)$ using parametrized alternating operator sequences. In particular, they numerically observed that, when the number of parameters in the sequence was greater than or equal to D^2, the gradient descent always finds the target unitary. On the other hand, in Ref. [45], overparametrization phenomena were studied in the context of estimating the ground state energies of transverse-field Ising and XXZ models. In particular, they employed the Hamiltonian variational *Ansatz* from Ref. [49] for these problems and numerically observed that the computational phase transition (or overparametrization) takes place when the number of parameters is much less than D^2.

Furthermore, in Ref. [46], the overparametrization was defined using the rank of the quantum Fisher information matrix associated with a QNN. They particularly focused on periodic structure *Ansätze* (PSAs) and argued that a PSA is overparametrized if the number of parameters scale as the dimension of the dynamical Lie algebra associated with a periodic structure *Ansatz* [46]. The results provided here show a deeper understanding of the training error dynamics, which cannot be obtained with algebraic arguments alone.

Although the overparametrized regime for random quantum circuits provides an analytic understanding of exponential convergence of the training error, it is not amenable to practical implementations as the number of parameters are required to scale as the dimension of the Hilbert space. However, our result should not be interpreted as a no-go theorem. In fact, our model simply cannot make predictions if the number of parameters is not large enough. On the other hand, Eqs. (10) and (11) are, in general, valid for k-design circuits, which can be efficiently implemented, but there are two major issues for such circuits: (1) the decay rate is small due to $1/D^2$ dependence, and (2) these circuits suffer from barren plateaus [23]. Thus, a practical challenge is to identify *Ansätze* that are trainable and have fast exponential convergence of the training error. In this regard, our analytic results could be applied to the case when the dynamical Lie algebra associated with the generators of a periodic structure *Ansatz* share a symmetry [30]. As discussed in Ref. [30], a symmetry can cause the state space to break into invariant subspaces, and the system may become reducible. Let us assume that such a reducible system is controllable on some or all of the invariant subspaces. More concretely, let $\mathcal{H} = \bigoplus_k \mathcal{H}_k$, and let the system be controllable on a subspace $\mathcal{H}_k$ of dimension D_k. If the initial state $\psi \in \mathcal{H}_k$, then under the condition that $D_k \in \mathcal{O}[\mathrm{poly}(\log D)]$, it is possible to get the decay rate $\gamma = \mathcal{O}(1)$ for $L \in \mathcal{O}[\mathrm{poly}(\log D)]$ number of parameters. Thus, for such *Ansätze*, it is possible to get trainability guarantees along with an exponential decay of the residual error.

Concentration of the QNTK and validity of the analytic regime.—We proved Eq. (10) under two assumptions: (1) For small learning rate η, in Eq. (5), we expanded $\delta\varepsilon$ up to the first order in η, (2) we considered $K \approx \bar{K}$ in Eq. (7). Now, we derive conditions in support of these assumptions. First, we derive conditions under which K does not fluctuate much around $\bar{K}$. In particular, under the assumption that the *Ansatz* in Eq. (1) forms a 4-design and in the large D limit, we get that $\Delta K = \sqrt{\mathbb{E}[(K\text{-}\bar{K})^2]}$ scales as [47]

$$\Delta K \approx \frac{\sqrt{L}}{D^2} \sqrt{[8\mathrm{Tr}^2(O^2) + 12\mathrm{Tr}(O^4)]}. \tag{12}$$

Similarly, we analyze the higher order corrections to the Taylor expansion of $\delta\varepsilon$ in Eq. (5), which was called the quantum metakernel in Ref. [42]. In particular,

$$\delta\varepsilon = -\eta K \varepsilon + (1/2)\eta^2 \varepsilon^2 \mu, \tag{13}$$

where

$$\mu = \sum_{\ell_1,\ell_2} \frac{\partial^2 \varepsilon}{\partial\theta_{\ell_1}\partial\theta_{\ell_2}} \frac{\partial \varepsilon}{\partial\theta_{\ell_1}} \frac{\partial \varepsilon}{\partial\theta_{\ell_2}}. \tag{14}$$

Under the assumption that the *Ansatz* in (1) forms a 6-design, we show that $\mathbb{E}(\mu) = 0$. This condition is similar to the classical counterpart of deep neural networks [50].

Moreover, in the large-D limit, $\Delta\mu \equiv \sqrt{\mathbb{E}(\mu^2)}$ scales as [47]

$$\Delta\mu \approx \frac{\sqrt{32}\eta L}{D^3}\mathrm{Tr}^{3/2}(O^2). \tag{15}$$

Therefore, assumptions made in deriving Eq. (10) are valid as long as

$$\frac{\Delta K}{\bar{K}} \approx \frac{1}{\sqrt{L}} \ll 1, \qquad \frac{\Delta\mu}{\bar{K}} \approx \frac{(\eta\sqrt{\mathrm{Tr}(O^2)})}{D} \ll 1. \tag{16}$$

We refer to these conditions as the concentration conditions. Note that $\mathrm{poly(t)} \cdot \log^2(D)$-depth local random circuits with two qubit nearest-neighbor gates on a one-dimensional lattice are sufficient to realize an approximate k-designs on $\log(D)$ qubits [51]. Thus, for $L \in \mathcal{O}[\log^2(D)]$, in the large-$D$ limit, both conditions in Eq. (16) are satisfied. Interestingly, in the large-D limit, $\Delta\mu/\bar{K} \ll 1$ is satisfied even for high values of η. Thus, our analytic solutions to the dynamics of the training error are valid in the large-D limit, which defines a wide QNN. A wide QNN also has a large number of variational parameters as $L \in \mathcal{O}[\log^2(D)]$ was assumed in deriving Eq. (16). It is an interesting question to determine similar conditions for other parametrized quantum circuits, including problem-inspired *Ansätze* [30]. Furthermore, we remark that $1/\mathrm{width}$ limit is also useful for analytic understanding of classical neural networks [50], where width implies the number of neurons in a single layer.

Supervised learning generalization.—In supervised learning, the QNTK is a symmetric, positive semidefinite matrix [52]. We compute the average behavior of the QNTK [47], in the frozen limit, finding that

$$\bar{K}^{i_1 i_2}_{\delta_1,\delta_2} \approx \frac{2L\mathrm{Tr}(O_{i_1}O_{i_2})}{D^2}\sigma_{\delta_1\delta_2}. \tag{17}$$

Here, $\delta_{1,2}$ correspond to input variables in the training data. The feature dependent analytic result is obtained through the definition of the feature S matrix,

$$S_{\delta_1\delta_2} = |\phi(\mathbf{x}_{\delta_1})\rangle\langle\phi(\mathbf{x}_{\delta_2})|, \tag{18}$$

and feature cross section,

$$\sigma_{\delta_1\delta_2} = |\mathrm{Tr}(S_{\delta_1\delta_2})|^2. \tag{19}$$

see [47] for a full derivation. We use the notation $\mathbf{x}_\delta$ to represent, in general, an element in the data space, i.e., both training and test data. In Eq. (17), we only include the leading order contributions in the large D limit, while the full nonperturbative expressions are given in [47]. Note that Eq. (17) is dependent on the feature maps used, unlike the case of the optimization task considered in Eq. (2). Moreover, using 4-design assumptions (see [47] for more details), we show that $\Delta K/\bar{K} \sim (1/\sqrt{L})$ which is similar to what is observed for classical neural networks.

Importantly, a difference between the supervised learning case and the optimization case is the dependence of the QNTK on the size of the training data. For $\sigma_{\delta_1\delta_2} \approx \delta_{\delta_1\delta_2}$ and $O_i = O$ for all i, we are able to model the eigenvalues of the QNTK [47]. We find that, for a training set size $|\mathcal{A}|$, there are $|\mathcal{A}| - 1$ eigenvalues which do not depend on the size of the data set. On the other hand, there is a one-dimensional eigenspace with the kernel eigenvalue, $[2L(D - |\mathcal{A}|)/(D^2 - 1)^2][D\mathrm{Tr}(O^2) - \mathrm{Tr}^2(O)]$, for all $D \geq |A|$. Thus, for high values of $|A|$, the behavior of the eigenvalues suggests slower decay rates for larger datasets.

Numerical experiments.—In what follows, we present numerical results verifying our analytic results presented above. First, for the optimization problem, we consider a variational *Ansatz*, as defined in Eq. (1) with

$$U_\ell = \exp(iP_\ell\theta_\ell), \qquad W_\ell = \mathrm{Haar} \in \mathrm{U}(D), \tag{20}$$

where we sample P_ℓ uniformly from the D-qubit Pauli group. Moreover, W_l is sampled with respect to a Haar measure on $U(D)$ and then kept fixed during the optimization. Let $O = \sum_{j=1}^{10} c_j\tilde{P}_j$, where $\tilde{P}_j$ are also sampled from the D-qubit Pauli group, and $c_j \in (0, 1)$. After sampling once, we keep P_ℓ, $\tilde{P}_j$, and c_j fixed during the optimization.

In Fig. 1, we study the scaling of $\bar{K}$ and $\Delta K/\bar{K}$ with respect to L for four qubits. In the inset of Fig. 1, we verify the linear scaling of $\bar{K}$ with L, as derived in Eq. (9). Similarly, Fig. 1 also follows the analytic scaling of $\Delta K/\bar{K} \approx 1/\sqrt{L}$, as derived in Eq. (12). In Fig. 2, we plot the residual training error ε versus the gradient descent optimization steps (time) for two qubits and $L = 64$, for 50 independent random initializations. Figure 2 verifies that,

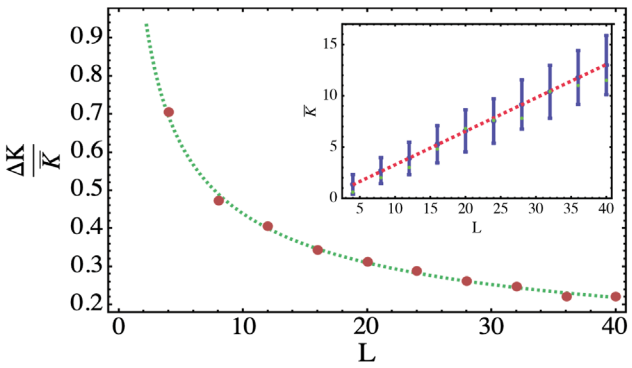

FIG. 1. Concentration of the QNTK on four qubits as a function of the circuit depth L. We pick different values of L (up to 40) in the randomized *Ansatz*, defined in Eq. (20). We sample over 1000 variational angles from Eq. (20) independently and uniformly over $[0, 2\pi]$. We plot $\Delta K/\bar{K}$ versus L, which verifies the analytic scaling of $L^{-1/2}$, as derived in Eq. (15). In the inset, we plot $\bar{K}$ versus L, where the dashed line represents the theoretical values of $\bar{K}$ from Eq. (9), green dots represent the numerical values of $\bar{K}$, and the blue error bars represent ΔK.

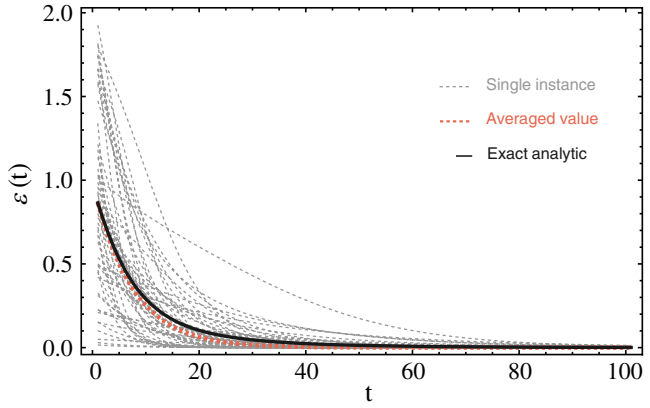

FIG. 2. Residual training error ε versus the gradient descent steps t for two qubits and $L = 64$ for the random *Ansatz* defined in Eq. (20). We use random initial angles in $[0, 2\pi]$ and perform the gradient descent experiment with a learning rate $\eta = 10^{-4}$ with 1000 steps. For 50 different initializations, we plot the dynamics of $\varepsilon(t)$, the theoretical prediction for the average dynamics of $\varepsilon(t)$, and the numerical values for the averaged $\varepsilon(t)$.

in the large-L limit, the residual error ε decays exponentially, as derived in Eqs. (10) and (11).

In Fig. 3, we focus on the data-dependent lowest kernel eigenvalue,

$$K_{\text{eigen}} = \frac{2L(D - |\mathcal{A}|)}{(D^2 - 1)^2}[D\text{Tr}(O^2) - \text{Tr}^2(O)]. \quad (21)$$

We consider a 4-qubit example and set $L = 64$. Other parameters are the same as in Fig. 1. We consider different values of $|\mathcal{A}|$ ranging from 2 to 10, and plot, both numerically and analytically, the value of the smallest eigenvalue, observing fair agreement. We generate the input data vectors such that they are orthogonal to each other. We provide further numerical results on supervised learning problems in [47].

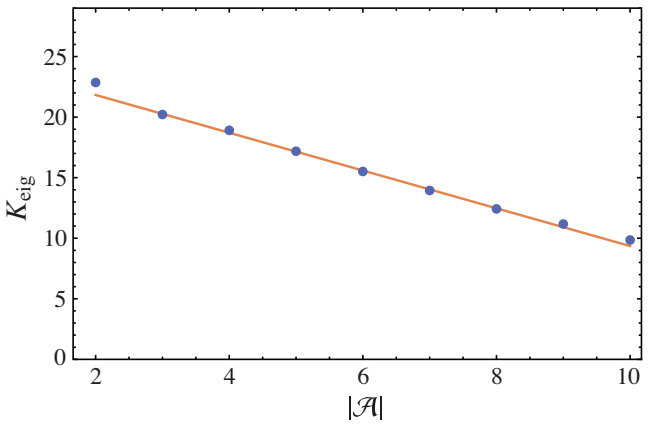

FIG. 3. Lowest eigenvalue for the kernel of a supervised learning task defined on four qubits and $L = 64$, as a function of the training data size $|\mathcal{A}|$. The predicted linear relation $[2L(D - |\mathcal{A}|)/(D^2 - 1)^2][D\text{Tr}(O^2) - \text{Tr}^2(O)]$ (orange solid line) agrees with numerical estimations with 50 independent instances of the *Ansatz*.

Discussion.—Using the quantum neural tangent kernel theory, we analytically solved the dynamics of the residual training error corresponding to variational quantum cost functions. Using these analytic solutions, we characterized an exponential decay of the residual training error as a function of the parameters of random quantum circuits. We derived conditions for which the second-order effects to the residual error and the fluctuations in the QNTK are negligible for wide QNNs.

One application of the theory developed in our work is to analyze the overparametrization of symmetric QNNs. As discussed previously, for the case when the subspace dimension of symmetric QNNs grows polynomially in the number of qubits, the number of parameters needed to observe a nonvanishing decay of the training error is also polynomial in number of qubits. Therefore, extending our results to symmetric QNNs [30] will be an important direction for future research. Another open question is to establish connections between the QNTK and the generalization error for quantum machine learning models [39,53,54], as well as with the quantum information theory bottleneck, as developed in [54]. Finally, note that the symmetric quantum neural networks already lead to desirable features in variational quantum algorithms, like the absence of barren plateaus [25].

We thank Jens Eisert, Keisuke Fujii, Dan A. Roberts, and Xiaodi Wu, for useful discussions. 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 the ARO (W911NF-23-1-0077), ARO MURI (W911NF-21-1-0325), AFOSR MURI (FA9550-19-1-0399, FA9550-21-1-0209), AFRL (FA8649-21-P-0781), DoE Q-NEXT, NSF (OMA-1936118, ERC-1941583, OMA-2137642), NTT Research, and the Packard Foundation (2020-71479). This research used resources of the Oak Ridge Leadership Computing Facility, which is a DOE Office of Science User Facility supported under Contract No. DE-AC05-00OR22725.

*junyuliu@uchicago.edu
†knajafi@ibm.com
‡kunals@ibm.com
§fta@zurich.ibm.com
‖liang.jiang@uchicago.edu
¶mezzacapo@ibm.com

[1] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton, Deep learning, Nature (London) **521**, 436 (2015).

[2] Jaehoon Lee, Yasaman Bahri, Roman Novak, Samuel S. Schoenholz, Jeffrey Pennington, and Jascha Sohl-Dickstein, Deep neural networks as gaussian processes, in International Conference on Learning Representations (2018), https://openreview.net/forum?id=B1EA-M-0Z.
[3] Arthur Jacot, Franck Gabriel, and Clément Hongler, Neural tangent kernel: Convergence and generalization in neural networks, arXiv:1806.07572.
[4] Jaehoon Lee, Lechao Xiao, Samuel Schoenholz, Yasaman Bahri, Roman Novak, Jascha Sohl-Dickstein, and Jeffrey Pennington, Wide neural networks of any depth evolve as linear models under gradient descent, Adv. Neural Inf. Process. Syst. **32**, 8572 (2019).
[5] Sanjeev Arora, Simon S Du, Wei Hu, Zhiyuan Li, Ruslan Salakhutdinov, and Ruosong Wang, On exact computation with an infinitely wide neural net, in Proceedings of the 33rd International Conference on Neural Information Processing Systems (2019).
[6] Jascha Sohl-Dickstein, Roman Novak, Samuel S Schoenholz, and Jaehoon Lee, On the infinite width limit of neural networks with a standard parameterization, arXiv:2001.07301.
[7] Greg Yang and Edward J. Hu, Feature learning in infinite-width neural networks, arXiv:2011.14522.
[8] Sho Yaida, Non-gaussian processes and neural networks at finite widths, in *Mathematical and Scientific Machine Learning* (PMLR, 2020), pp. 165–192.
[9] Hsin-Yuan Huang, Michael Broughton, Masoud Mohseni, Ryan Babbush, Sergio Boixo, Hartmut Neven, and Jarrod R McClean, Power of data in quantum machine learning, Nat. Commun. **12**, 2631 (2021).
[10] Amira Abbas, David Sutter, Christa Zoufal, Aurélien Lucchi, Alessio Figalli, and Stefan Woerner, The power of quantum neural networks, NATO ASI series Series F, Computer and system sciences **1**, 403 (2021).
[11] Yunchao Liu, Srinivasan Arunachalam, and Kristan Temme, A rigorous and robust quantum speed-up in supervised machine learning, Nat. Phys. **17**, 1013 (2021).
[12] Dorit Aharonov, Jordan Cotler, and Xiao-Liang Qi, Quantum algorithmic measurement, Nat. Commun. **13**, 887 (2022).
[13] Maria Schuld, Ilya Sinayskiy, and Francesco Petruccione, An introduction to quantum machine learning, Contemp. Phys. **56**, 172 (2015).
[14] Jacob Biamonte, Peter Wittek, Nicola Pancotti, Patrick Rebentrost, Nathan Wiebe, and Seth Lloyd, Quantum machine learning, Nature (London) **549**, 195 (2017).
[15] Vedran Dunjko and Hans J. Briegel, Machine learning and artificial intelligence in the quantum domain: A review of recent progress, Rep. Prog. Phys. **81**, 074001 (2018).
[16] https://research.ibm.com/blog/127-qubit-quantum-processor-eagle.
[17] Maria Schuld and Nathan Killoran, Is quantum advantage the right goal for quantum machine learning?, PRX Quantum **3**, 030101 (2022).
[18] Edward Farhi and Hartmut Neven, Classification with quantum neural networks on near term processors, arXiv:1802.06002.
[19] Iris Cong, Soonwon Choi, and Mikhail D. Lukin, Quantum convolutional neural networks, Nat. Phys. **15**, 1273 (2019).
[20] Johannes Bausch, Recurrent quantum neural networks, Adv. Neural Inf. Process. Syst. **33**, 1368 (2020).
[21] Kerstin Beer, Dmytro Bondarenko, Terry Farrelly, Tobias J. Osborne, Robert Salzmann, Daniel Scheiermann, and Ramona Wolf, Training deep quantum neural networks, Nat. Commun. **11**, 1 (2020).
[22] S. Mangini, F. Tacchino, D. Gerace, D. Bajoni, and C. Macchiavello, Quantum computing models for artificial neural networks, Europhys. Lett. **134**, 10002 (2021).
[23] Jarrod R. McClean, Sergio Boixo, Vadim N. Smelyanskiy, Ryan Babbush, and Hartmut Neven, Barren plateaus in quantum neural network training landscapes, Nat. Commun. **9**, 1 (2018).
[24] Marco Cerezo, Akira Sone, Tyler Volkoff, Lukasz Cincio, and Patrick J. Coles, Cost function dependent barren plateaus in shallow parametrized quantum circuits, Nat. Commun. **12**, 1 (2021).
[25] Arthur Pesah, M. Cerezo, Samson Wang, Tyler Volkoff, Andrew T. Sornborger, and Patrick J. Coles, Absence of Barren Plateaus in Quantum Convolutional Neural Networks, Phys. Rev. X **11**, 041011 (2021).
[26] Kunal Sharma, Marco Cerezo, Lukasz Cincio, and Patrick J. Coles, Trainability of Dissipative Perceptron-Based Quantum Neural Networks, Phys. Rev. Lett. **128**, 180505 (2022).
[27] Zidu Liu, Li-Wei Yu, L.-M. Duan, and Dong-Ling Deng, The Presence and Absence of Barren Plateaus in Tensor-Network Based Machine Learning, Phys. Rev. Lett. **129**, 270501 (2022).
[28] Maria Kieferova, Ortiz Marrero Carlos, and Nathan Wiebe, Quantum generative training using Renyi divergences, arXiv:2106.09567.
[29] Z. Holmes, K. Sharma, M. Cerezo, and P. J. Coles, Connecting ansatz expressibility to gradient magnitudes and barren plateaus, PRX Quantum **3**, 010313 (2022).
[30] Martin Larocca, Piotr Czarnik, Kunal Sharma, Gopikrishnan Muraleedharan, Patrick J. Coles, and M. Cerezo, Diagnosing barren plateaus with tools from quantum optimal control, Quantum **6**, 824 (2022).
[31] Taylor L. Patti, Khadijeh Najafi, Xun Gao, and Susanne F. Yelin, Entanglement devised barren plateau mitigation, Phys. Rev. Res. **3**, 033090 (2021).
[32] Chen Zhao and Xiao-Shan Gao, Analyzing the barren plateau phenomenon in training quantum neural networks with the ZX-calculus, Quantum **5**, 466 (2021).
[33] Samson Wang, Enrico Fontana, Marco Cerezo, Kunal Sharma, Akira Sone, Lukasz Cincio, and Patrick J. Coles, Noise-induced barren plateaus in variational quantum algorithms, Nat. Commun. **12**, 1 (2021).
[34] Supanut Thanasilp, Samson Wang, Nhat A. Nghiem, Patrick J. Coles, and Marco Cerezo, Subtleties in the trainability of quantum machine learning models, arXiv:2110.14753.
[35] Sukin Sim, Peter D. Johnson, and Alán Aspuru-Guzik, Expressibility and entangling capability of parameterized quantum circuits for hybrid quantum-classical algorithms, Adv. Quantum Technol. **2**, 1900070 (2019).
[36] Kouhei Nakaji and Naoki Yamamoto, Expressibility of the alternating layered ansatz for quantum computation, Quantum **5**, 434 (2021).

[37] Yuxuan Du, Zhuozhuo Tu, Xiao Yuan, and Dacheng Tao, An Efficient Measure for the Expressivity of Variational Quantum Algorithms, Phys. Rev. Lett. **128**, 080506 (2022).
[38] Kunal Sharma, M Cerezo, Zoë Holmes, Lukasz Cincio, Andrew Sornborger, and Patrick J. Coles, Reformulation of the No-Free-Lunch Theorem for Entangled Datasets, Phys. Rev. Lett. **128**, 070501 (2022).
[39] Matthias C. Caro, Hsin-Yuan Huang, M. Cerezo, Kunal Sharma, Andrew Sornborger, Lukasz Cincio, and Patrick J. Coles, Generalization in quantum machine learning from few training data, Nat. Commun. **13**, 4919 (2022).
[40] Hsin-Yuan Huang, Richard Kueng, and John Preskill, Information-Theoretic Bounds on Quantum Advantage in Machine Learning, Phys. Rev. Lett. **126**, 190505 (2021).
[41] 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**, 2631 (2021).
[42] Junyu Liu, Francesco Tacchino, Jennifer R. Glick, Liang Jiang, and Antonio Mezzacapo, Representation Learning via Quantum Neural Tangent Kernels, PRX Quantum **3**, 030323 (2021).
[43] Norihito Shirai, Kenji Kubo, Kosuke Mitarai, and Keisuke Fujii, Quantum tangent kernel, arXiv:2111.02951.
[44] Bobak Toussi Kiani, Seth Lloyd, and Reevu Maity, Learning unitaries by gradient descent, arXiv:2001.11897.
[45] Roeland Wiersema, Cunlu Zhou, Yvette de Sereville, Juan Felipe Carrasquilla, Yong Baek Kim, and Henry Yuen, Exploring entanglement and optimization within the hamiltonian variational ansatz, PRX Quantum **1**, 020319 (2020).
[46] Martin Larocca, Nathan Ju, Diego García-Martín, Patrick J Coles, and M Cerezo, Theory of overparametrization in quantum neural networks, arXiv:2109.11676.
[47] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevLett.130.150601 for additional calculations and experiments.
[48] Christoph Dankert, Richard Cleve, Joseph Emerson, and Etera Livine, Exact and approximate unitary 2-designs and their application to fidelity estimation, Phys. Rev. A **80**, 012304 (2009).
[49] Dave Wecker, Matthew B. Hastings, and Matthias Troyer, Progress towards practical quantum variational algorithms, Phys. Rev. A **92**, 042303 (2015).
[50] Daniel A. Roberts, Sho Yaida, and Boris Hanin, *The Principles of Deep Learning Theory* (Cambridge University Press, Cambridge, England, 2022).
[51] Aram Harrow and Saeed Mehraban, Approximate unitary *t*-designs by short random quantum circuits using nearest-neighbor and long-range gates, arXiv:1809.06957.
[52] Junyu Liu, Francesco Tacchino, Jennifer R Glick, Liang Jiang, and Antonio Mezzacapo, Representation learning via quantum neural tangent kernels, PRX Quantum **3**, 030323 (2022).
[53] James B. Simon, Madeline Dickens, and Michael R. DeWeese, Neural tangent kernel eigenvalues accurately predict generalization, arXiv:2110.03922.
[54] Leonardo Banchi, Jason Pereira, and Stefano Pirandola, Generalization in quantum machine learning: A quantum information standpoint, PRX Quantum **2**, 040321 (2021).