Laziness, Barren Plateau, and Noise in Machine Learning

Junyu Liu
Pritzker School of Molecular Engineering, The University of Chicago, Chicago, IL 60637, USA
Chicago Quantum Exchange, Chicago, IL 60637, USA
Kadanoff Center for Theoretical Physics, The University of Chicago, Chicago, IL 60637, USA
qBraid Co., Harper Court 5235, Chicago, IL 60615, USA
junyuliu@uchicago.edu

Zexi Lin
Pritzker School of Molecular Engineering, The University of Chicago, Chicago, IL 60637, USA
zexil@uchicago.edu

Liang Jiang
Pritzker School of Molecular Engineering, The University of Chicago, Chicago, IL 60637, USA
Chicago Quantum Exchange, Chicago, IL 60637, USA
liangjiang@uchicago.edu

Abstract

We define *laziness* to describe a large suppression of variational parameter updates for neural networks, classical or quantum. In the quantum case, the suppression is exponential in the number of qubits for randomized variational quantum circuits. We discuss the difference between laziness and *barren plateau* in quantum machine learning created by quantum physicists in [1] for the flatness of the loss function landscape during gradient descent. We address a novel theoretical understanding of those two phenomena in light of the theory of neural tangent kernels. For noiseless quantum circuits, without the measurement noise, the loss function landscape is complicated in the overparametrized regime with a large number of trainable variational angles. Instead, around a random starting point in optimization, there are large numbers of local minima that are good enough and could minimize the mean square loss function, where we still have quantum laziness, but we do not have barren plateaus. However, the complicated landscape is not visible within a limited number of iterations, and low precision in quantum control and quantum sensing. Moreover, we look at the effect of noises during optimization by assuming intuitive noise models, and show that variational quantum algorithms are noise-resilient in the overparametrization regime. Our work precisely reformulates the quantum barren plateau statement towards a precision statement and justifies the statement in certain noise models, injects new hope toward near-term variational quantum algorithms, and provides theoretical connections toward classical machine learning. Our paper provides conceptual perspectives about quantum barren plateaus, together with discussions about the gradient descent dynamics in [2].

1 Barren plateau, laziness and noise

Variational quantum circuits [3–8] can be used to optimize cost function measured on quantum computers. Specifically, these cost functions can be used for machine learning tasks [9–16]. In this case variational quantum circuits are addressed as quantum neural networks.

However, a generically designed variational quantum ansatz may not be applicable to real problems. Specifically, a problem so-called *barren plateau* has been widely discussed in the variational quantum algorithm community, which is believed to be one of the primary problems of quantum machine learning [1]. The argument is given as follows. A typical gradient descent algorithm will look like

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

where θ_μ is the variational angle, and t is referring the time step of gradient descent dynamics. η is the learning rate, and $\mathcal{L}$ is the loss function. The observation [1] is that, if our variational ansatz is highly random, due to the k-design integral formula [17–20], the derivative of the loss function is generically suppressed by the dimension of the Hilbert space N, and we might encounter a situation where the variation of the loss function during gradient descent is very small, namely $\delta\mathcal{L} \equiv \mathcal{L}(t+1) - \mathcal{L}(t) \ll 1$ for the step t. For instance, the second moment formula for Haar ensemble is

$$\int dU U_{ij} U^\dagger_{kl} = \frac{1}{N} \delta_{il}\delta_{jk} \,. \tag{2}$$

Here U is a unitary taken from a 1-design, and δ is the Kronecker delta and i, j, k, l are matrix indexes. For higher moments random integrals [17–21], the factor $\text{poly}(1/N)$ will appear. Thus, the difference between the variational angles during iterations will be suppressed by the dimension of the Hilbert space. The work [1] demonstrates this existence of the *barren plateau* (the statement where $\delta\mathcal{L} \ll 1$) numerically and understands the result as a primary challenge of variational quantum circuits. It is often considered to be quantum analogs to the *vanishing gradient problem*, but the nature is fundamentally different [22, 23]. A further explanation is given in Appendix A.

Although the existence of the barren plateau is verified by numerous works [24–27], the theoretical understanding of the barren plateau problem is unclear. Moreover, the classical machine learning community has been successfully demonstrated its practical usage in science and business for years, and many successful classical neural network algorithms have been run for large scales. For example, `Generative Pre-trained Transformer-3` (`GPT-3`) from `OpenAI` [28] has used 175 billion of training parameters, and it is one of the most successful natural language processing models up to date. Considering the standard LeCun initialization of weights W with the normalization of the variance σ_W^2 [22, 23, 29]

$$\mathbb{E}(W_{ij} W^\dagger_{kl}) = \frac{\sigma_W^2}{\text{width}} \delta_{ik}\delta_{jl} \,, \tag{3}$$

and its formal similarity to Equation 2, we might imagine that similar issues will happen for classical neural networks too: they might be highly overparametrized in the large-width limit. Here, σ_W is a number that is independent of the size of the neural networks, and we set the width of the neural network to be the same in each layer for simplicity. In fact, in Appendix A, we will show that in the classical large-width neural network, the barren plateau will also happen: the trainable weights do not run that much during gradient descent.

So, why classical overparametrized neural networks are supposed to be practical and good, but the barren plateaus of quantum neural networks are crucial challenges? In this paper, we define the primary theoretical argument towards the quantum barren plateau, the large suppression of the right hand side of Equation 1, as *laziness*. In the quantum context, the suppression is from the dimension of the Hilbert space, while in the classical case, the suppression is from the width of the classical neural networks. In a more precise language, laziness is referring to small $\delta\theta_\mu$, and barren plateau is referring to small $\delta\mathcal{L}$.

Moreover, we will show that laziness may not imply the quantum barren plateau, from the perspective of overparametrization theory and representation learning theory through quantum neural tangent kernels (QNTKs) [2, 29]. In this paper, for quantum neural networks *overparametrization* is referring to the fact where $L\text{Tr}(O^2)/N^2 \approx \mathcal{O}(1)$, where O is the operator we are optimizing, L is the number of trainable angles, and η is the learning rate as a constant.

Defining quantum analogs of neural tangent kernels from their classical counterparts [23, 30–40], we show that from the first-principle theoretical derivation, random (noiseless) quantum neural networks are still efficient to learn in the large-L limit without barren plateaus, despite their laziness. In fact, although each trainable angle does not move much due to the small magnitude of the gradient, the combined effect of many of them on the loss function will still be significant. In addition, there

exist good enough achievable local minima that minimize the training error. See Figure 1 for an illustration. The requirements for making this to happen is especially when $L\text{Tr}(O^2)/N^2 \approx \mathcal{O}(1)$, and we have a small learning rate and the mean square loss function. In the case of large Hilbert space dimension without overparametrization, the exponential decay rate during gradient descent might be small, which may not make this phenomenon manifest in the polynomial training iterations. In practice, what we see is a very slow decay of loss functions. Interestingly, in this case quantum noises will not affect us significantly until exponential numbers of iterations. Thus, the averaged QNTK, $\bar{K}$, proportional to $\text{Tr}(O^2)L/N^2$, *explains* the existence of the barren plateau in practice, with or without noises. On the other hand, in the overparametrization regime where $\eta L\text{Tr}(O^2)/N^2 \approx \mathcal{O}(1)$, the exponential decay of gradient descent process is visible.

We note that the large-L expansion is a quantum analog of the classical neural tangent kernel theory at large width. In fact, we will show in Section 2 that we have similar large-width expansion comparing the classical theory, where in our model, *classical width* corresponds to L. The dimension of the Hilbert space plays an important role in the calculation. Moreover, the correspondence between quantum and classical neural networks might be explained by some physical heuristics, from the duality between matrix models and quantum field theories. See Appendix C for a brief discussion.

Moreover, we need to point out that laziness is intrinsically still a precision problem. More precisely, it could be primarily from quantum measurement and quantum control, since the size of classical devices could scale as $\log(1/\epsilon)$ for given precision ϵ, while variational quantum circuits cannot, due to the measurement error and the limitation of quantum control [1]. Thus, it naturally motivates us to think about how to include the effect of noise in the gradient descent calculation. In our work, we introduce a simple and intuitive noise model by adding random variables in the gradient descent dynamics. We show that in the overparametrization regime, our variational quantum algorithms are noise-resilient. More precisely, we find that the residual training error scales as

$$\varepsilon^2(t) \approx (1-\eta K)^{2t}\left(\varepsilon^2(0) - \frac{\sigma_\theta^2}{\eta(2-\eta K)}\right) + \frac{\sigma_\theta^2}{\eta(2-\eta K)}\,, \tag{4}$$

with the neural tangent kernel K and the standard deviation of the noise introduced in the variational angles σ_θ. Thus, in the late time, we get

$$\mathcal{L}(\infty) = \frac{1}{2}\varepsilon^2(\infty) \approx \frac{\sigma_\theta^2}{2\eta(2-\eta K)}\,. \tag{5}$$

In the late time, we have

$$\mathcal{L}(\infty) = \frac{1}{2}\varepsilon^2(\infty) \approx \frac{\sigma_\theta^2}{2\eta(2-\eta K)}\,. \tag{6}$$

Thus, in the overparametrized regime, we could set $\eta K \approx \mathcal{O}(1)$, so schematically,

$$\mathcal{L}(\infty) \approx \mathcal{O}(\frac{\sigma_\theta^2}{\eta})\,, \tag{7}$$

indicating that we could get good predictions at the end as long as we sufficiently control the noises.

We will give more details in the following sections.

2 The loss function landscape and the QNTK theory

We begin by considering a variational quantum circuit ansatz, on a Hilbert space of size N with $\log_2 N$ qubits, as follows,

$$U(\theta) = \left(\prod_{\ell=1}^{L} W_\ell \exp\left(i\theta_\ell X_\ell\right)\right) \equiv \left(\prod_{\ell=1}^{L} W_\ell U_\ell\right), \tag{8}$$

with some trainable angles θ_ℓ, constant unitary operators W_ℓ, and Pauli operators X_ℓ. Following [29], we consider the mean square loss function

$$\mathcal{L}(\theta) = \frac{1}{2}\left(\left\langle \Psi_0 \left| U^\dagger(\theta) O U(\theta) \right| \Psi_0 \right\rangle - O_0\right)^2 \equiv \frac{1}{2}\varepsilon^2\,, \tag{9}$$

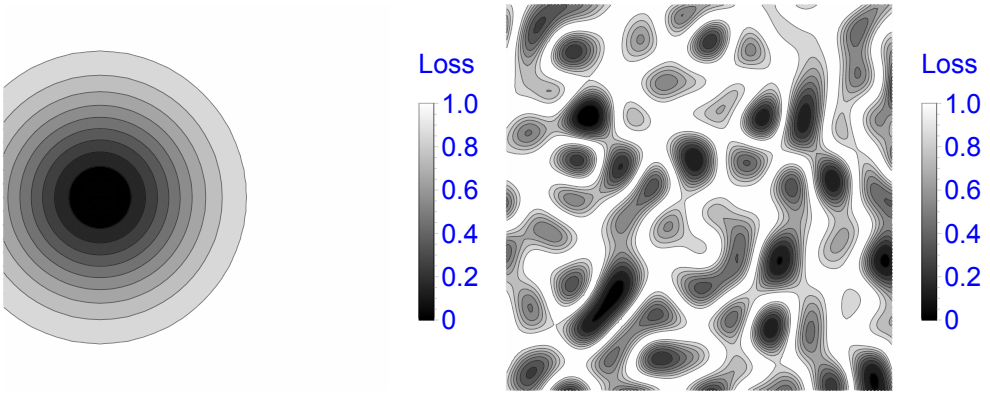

Figure 1: Density plots of the loss function landscape comparing usual and overparametrized variational quantum circuits. We illustrate the landscape by color plots of the loss function for two variational angles. Left: the traditional understanding of *barren plateaus* where we have the a single optimal point. Right: in the overparametrized case, the landscape is not barren, since for a random initial point, we get many good enough local optima that could minimize the loss function. Note that those plots are schematic since it is not possible to directly plot the loss function landscape in very high dimensions. In order to visualize it in $\mathcal{O}(1)$ numbers of iterations, one might have to have the number of trainable angles L comparable to the dimension of the Hilbert space N.

and train the expectation value $\left\langle \Psi_0 \left| U^\dagger(\theta) O U(\theta) \right| \Psi_0 \right\rangle$ on an initial state $|\Psi_0\rangle$ towards a value O_0. We define the residual training error $\varepsilon = \left\langle \Psi_0 \left| U^\dagger(\theta) O U(\theta) \right| \Psi_0 \right\rangle - O_0$. We use the gradient descent algorithm Equation 1 with the learn ing rate η and an initial variational angle $\theta(0)$. We look now at the difference of the residual training error

$$\delta\varepsilon \equiv \varepsilon(t+1) - \varepsilon(t) \, . \tag{10}$$

When the learning rate of Equation 1 η is small, we can perform a Taylor expansion,

$$\delta\varepsilon \approx \sum_\ell \frac{\partial \varepsilon}{\partial \theta_\ell} \delta\theta_\ell = -\eta \sum_\ell \frac{\partial \varepsilon}{\partial \theta_\ell} \frac{\partial \varepsilon}{\partial \theta_\ell} \varepsilon = -\eta K \varepsilon \, . \tag{11}$$

The quantity K here is called the Quantum Neural Tangent Kernel (QNTK) [29], $K = \sum_\ell \frac{\partial \varepsilon}{\partial \theta_\ell} \frac{\partial \varepsilon}{\partial \theta_\ell}$. Note that in a general supervised learning setup where one has a labeled dataset instead of just one expected value O_0, K is a positive-semidefinite and symmetric matrix instead of a non-negative number. Here we focus on the optimization problem Equation 9: this example will demonstrate the validity of our theory, that can be readily generalized to a full supervised quantum machine learning setup.

A *frozen* QNTK will remain constant during a gradient descent flow will lead to gradient flow equations which can be solved exactly [29], showing that the error will decay exponentially at the gradient descent iteration t as

$$\varepsilon(t) = (1 - \eta K)^t \varepsilon(0) \, . \tag{12}$$

For sufficient random variational ansätze, we could compute the value of K based on the same assumption of the barren plateau problem [1]. After computing 2-design random average $\mathbb{E}$ (see [2] for more details)

$$\mathbb{E}(O) = \int_{U \in \text{2-design}} dU O(U) \, , \tag{13}$$

More precisely, we define

$$U_{-,\ell} \equiv \prod_{\ell'=1}^{\ell-1} W_{\ell'} U_{\ell'}, U_{+,\ell} \equiv \prod_{\ell'=\ell+1}^{L} W_{\ell'} U_{\ell'},$$
$$V_{-,\ell} = U_{-,\ell} W_\ell U_\ell, V_{+,\ell} = U_{+,\ell} \,. \tag{14}$$

And we assume that $V_{-,\ell}$ and $V_{+,\ell}$ form 2-designs independently in all ℓs. We get the following expression of the averaged QNTK,

$$\bar{K} = \mathbb{E}(K) = L\left(N\mathrm{Tr}\left(O^2\right) - \mathrm{Tr}^2(O)\right)\frac{2}{N+1}\left(\frac{1}{N^2-1}\right)$$
$$\approx \frac{2L\mathrm{Tr}\left(O^2\right)}{N^2} \,. \tag{15}$$

This simple equation combined with Equation 12 reveals how, on average, the residual training error of a gradient descent dynamics will decay exponentially. Moreover, one should also check the standard deviation ΔK. If $\Delta K \ll \bar{K}$, we get a distribution of K which is concentrated at $\bar{K}$. In fact, one could show that from k-design assumptions,

$$\Delta K \approx \frac{\sqrt{L}}{N^2}\sqrt{\left(8\mathrm{Tr}^2\left(O^2\right) + 12\mathrm{Tr}\left(O^4\right)\right)} \,. \tag{16}$$

Thus, we have $\Delta K/\bar{K} = \mathcal{O}(1/\sqrt{L})$. In the limit where $L \gg 1$, the neural tangent kernel is concentrated around a fixed value $\bar{K}$. A more precise constraint will also include a time-dependent statement including the perturbations of higher-order Taylor expansion of the residual training error, which is characterized by the so-called quantum meta-kernel or dQNTK. See Appendix B for more details.

3 Precision and noise

Now we give some physical interpretations about Equation 15. We see in Section 2 that the theory should work in the regime where $L \gg 1$, and also the overparametrization regime where $\eta K \approx \mathcal{O}(1)$. From Equation 12, we know that $\bar{K}$ would serve as an exponent of exponential decay: the larger $\bar{K}$ is, the faster the algorithm will converge. This qualitative description has been formulated in [29], with numerical evidence in [41] around the same time.

Moreover, a statement about precision could be made by combining Equation 12 and Equation 15. We have

$$\log\frac{1}{\varepsilon_r} \approx -T\log\left(1-\eta\bar{K}\right) \approx \eta\bar{K}T \,. \tag{17}$$

Here, T is the total training steps, and ε_r is the relative residual training error around the end of training $\varepsilon_r = \varepsilon(T)/\varepsilon(0)$. The relative error ε_r could be as small as the precision of the quantum device. Using Equation 15, we get

$$\log\frac{1}{\varepsilon_r} \approx \frac{2\eta L\mathrm{Tr}\left(O^2\right)T}{N^2} \,. \tag{18}$$

Equation 18 makes the barren plateau problem manifestly as a precision problem. If we want to see the convergence within $T \approx \mathcal{O}(1)$, we want $\eta\bar{K} \approx 1$. The smaller $\bar{K}$ is, the smaller decaying exponent we have, and more likely we will experience a barren plateau in practice. Otherwise, there will be good enough local optima around the small random fluctuations of variational angles. The more overparametrized the quantum neural networks are, the faster convergence they could have. In this case, we do not have a barren plateau if we assume that we do not have the measurement noise and the quantum hardware noise, although we have laziness.

Originally, a relation between the barren plateau problem and the precision has also been stated in [1], while we make it more clear by showing that the barren plateau is not algorithmic. In fact, in Appendix A, we show that classical overparametrized neural networks have laziness as well. Many useful, practical machine learning algorithms have to be in this case [23]. Thus, variational quantum

algorithms here have no algorithmic issue, and the origin of the problem comes from measurement and control (see also [42]).

Let us take a look at Equation 1 again. To implement variational algorithms, we need to perform measurements to evaluate the loss function or its derivatives (involving quantum measurements), and update the trainable angles through Equation 1 (involving quantum control). On the measurement side, classical computations could handle the precision-ϵ computation with the resource scaling as $\log 1/\epsilon$, while measurement errors will be produced in the quantum setup, making the scaling $1/\epsilon^{\alpha}$ for positive α [1]. There is no known way to date to avoid it because of limitations of metrology [43]. On the control side, it is also challenging to update the variational angles with exponential precision. In a sense, our theory makes the statement from [1] more precise.

The discussion naturally motivates us to introduce the noise model. Heuristically, we will expect that during the gradient descent process, the effective noise term will also be exponentially decaying because of the original recurrence relation and its solution. To verify this, we could add a random fluctuation term $\Delta\theta_\ell$ to model the uncertainty of measuring the expectation value. One could also assume that the random variable $\Delta\theta_\ell$ is Markovian. Namely, it is independent for the time step t. Moreover, we assume that $\Delta\theta_\ell$s are distributed with Gaussian distributions $\mathcal{N}(0,\sigma_\theta^2)$. Note that σ_θ could come from the measurement noise during estimations of quantum observables used for the gradient descent, which scales as $1/\sqrt{n}$, where n is the number of measurements. And the Gaussian assumptions come from the central limit theorem in the large-n limit. Furthermore, σ_θ could also come from the hardware noises. On the other hand, the physical implementation of rotation angle will also have limited precision. One could note that robust quantum control techniques can suppress errors of rotation angles to higher orders, see [44].

Thus, one could show that the residual training error has the recursion relation in the linear order of the Taylor expansion,

$$\delta\varepsilon = -\eta\varepsilon K + \sum_\ell \frac{\partial\varepsilon}{\partial\theta_\ell}\Delta\theta_\ell \ . \tag{19}$$

Now, let us assume that K is still a constant, $K \approx \bar{K}$. Since $\Delta\theta_\ell \sim \mathcal{N}(0,\sigma_\theta^2)$, we get

$$\sum_\ell \frac{\partial\varepsilon}{\partial\theta_\ell}\Delta\theta_\ell \sim \mathcal{N}(0, K\sigma_\theta^2) \ . \tag{20}$$

Including the noise term into the recursion relation, one could show that averaging over the random distribution of the noise, we have

$$\overline{\varepsilon^2}(t) = (1-\eta K)^{2t}\left(\varepsilon^2(0) - \frac{\sigma_\theta^2}{\eta(2-\eta K)}\right) + \frac{\sigma_\theta^2}{\eta(2-\eta K)} \ . \tag{21}$$

Note that the first term is decaying when the time t is increasing. At the late time, we have

$$\overline{\varepsilon^2}(\infty) = \frac{\sigma_\theta^2}{\eta(2-\eta K)} \approx \mathcal{O}(\frac{\sigma_\theta^2}{\eta}) \ , \tag{22}$$

where we assume the overparametrization $\eta K \approx \mathcal{O}(1)$. Thus, at the late time, the loss function will arrive at a constant plateau at $\mathcal{O}(\sigma_\theta^2/\eta)$. One could improve σ_θ to make the constant plateau controllable and do not increase significantly with N, indicating that our algorithm could be noise-resilient. See Appendix D for a more detailed discussion, and see Figure 1 for an illustration. Some numerical results are also obtained in Figure 2 and Figure 3.

4 Conclusion and outlook

In this paper, we point out that for variational circuits with sufficiently large numbers of trainable angles, the gradient descent dynamics could still be efficiently performed, despite the existence of the exponential suppression of the variational angle updates (laziness). We point out that laziness is not uniquely happening in quantum machine learning, but also for overparametrized classical neural networks with large widths. The efficiency of large-width neural networks is justified by the neural tangent kernel theory, so do their quantum counterparts. A solid and simple theory has been established based on the above ideas, and the relation between the number of training steps,

the quantum device error, the trainable depth, the dimension of the Hilbert space, and the norm of operators appearing in the loss function has been explicitly derived. Moreover, we have justified that for simple and natural noise models, we could make the variational quantum circuits noise-resilient in the overparametrized regime, with solid theoretical and numerical evidence.

Our results also indicate a more well-defined path to designing quantum neural networks from the first principle. If we are sampling unitary operators uniformly in the whole unitary group, it is hard to avoid polynomial factors of N, the dimension of the Hilbert space, into the expression of the number of iterations in order to obtain the visible laziness (see parallel efforts in [45, 46]). One idea is to reduce the space of searching, and reduce the space of variational circuits to some subspaces, where people observe some evidence for setups in quantum convolutional neural networks [25, 47] and local loss function [24], and the barren plateau phenomena are less drastic in those cases. However, since the subspace we are searching is reduced, the decreased expressibility will lead to a lower performance for the final convergence of the loss function on the training set [45]: around the end of the training, drastic corrections towards fixed neural tangent kernels will stop the exponential decay, and we get a local minimum which may not be good enough. The design of variational circuits will be a trade-off between barren plateaus and performance [48], which could be manifest in the presence of laziness. Despite generalizations to full learning setups with multiple output dimensions, other interesting directions include detailed discussions about the quantum noise in the real machines during quantum representation learning to understand how the noise will affect laziness and the barren plateau, a justification of our theory with large-scale classical and quantum simulation, and possible theoretical understandings beyond the limit $L \gg 1$. We look forward to further analysis and research along our path.

Note added: When the paper is finished, we notice that another nice independent paper [49] appears in the arxiv, which has very similar conclusion to our results.

Acknowledgments and Disclosure of Funding

We thank Jens Eisert, Keisuke Fujii, Isaac Kim, Risi Kondor, Kenji Kubo, Antonio Mezzacapo, Kosuke Mitarai, Khadijeh Najafi, Sam Pallister, John Preskill, Dan A. Roberts, Norihito Shira, Eva Silverstein, Francesco Tacchino, Shengtao Wang, Xiaodi Wu, Yi-Zhuang You, Han Zheng, and Quntao Zhuang for useful discussions.

We acknowledge support from the ARO (W911NF-18-1-0020, W911NF-18-1-0212), ARO MURI (W911NF-16-1-0349, 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, EEC-1941583, OMA-2137642), NTT Research, and the Packard Foundation (2020-71479).

Appendix

A Comments on the barren plateau in the *classical* machine learning

Now we consider a classical neural network, the MLP model (see [23]). The definition is

$$\begin{aligned}
&z_i^{(1)}(x_\alpha) \equiv b_i^{(1)} + \sum_{j=1}^{n_0} W_{ij}^{(1)} x_{j;\alpha}, \\
&\text{for} \quad i = 1, \ldots, n_1, \\
&z_i^{(\ell+1)}(x_\alpha) \equiv b_i^{(\ell+1)} + \sum_{j=1}^{n_\ell} W_{ij}^{(\ell+1)} \sigma\left(z_j^{(\ell)}(x_\alpha)\right), \\
&\text{for} \quad i = 1, \ldots, n_{\ell+1}; \ell = 1, \ldots, L-1.
\end{aligned} \tag{23}$$

Here, σ is a non-linear activation function, and we have widths $n_{1,2,\cdots,L}$ in layers $\ell = 1, 2, \cdots L$. The input dimension is n_0 and the output dimension is n_L. Weights and biases at layer ℓ are denoted as $W^{(\ell)}$ and $b^{(\ell)}$. $z^{(\ell)}$ is called the *preactivation*. $x_{j,\alpha}$ will denote the data where j is the vector index, and α is the data sample index. At the beginning, we initialize the neural network by

$$\begin{aligned}
&\mathbb{E}\left[b_{i_1}^{(\ell)} b_{i_2}^{(\ell)}\right] = \delta_{i_1 i_2} C_b^{(\ell)}, \\
&\mathbb{E}\left[W_{i_1 j_1}^{(\ell)} W_{i_2 j_2}^{(\ell)}\right] = \delta_{i_1 i_2} \delta_{j_1 j_2} \frac{C_W^{(\ell)}}{n_{\ell-1}}.
\end{aligned} \tag{24}$$

Here, C_b and C_W will set the variance of biases and weights (we use the notation $C_W = \sigma_W^2$ in the main text). And we train the neural networks by gradient descent algorithms. We could consider the simplest version of the gradient descent algorithm,

$$\theta_\mu(t+1) = \theta_\mu(t) - \eta \frac{d\mathcal{L}_\mathcal{A}}{d\theta_\mu}\bigg|_{\theta(t)}. \tag{25}$$

The loss function is

$$\mathcal{L}_\mathcal{A} \equiv \frac{1}{2} \sum_{i,\tilde{\alpha} \in \mathcal{A}} \left(z_i(x_{\tilde{\alpha}}; \theta) - y_{i,\tilde{\alpha}}\right)^2 = \frac{1}{2} \sum_{i,\tilde{\alpha} \in \mathcal{A}} \varepsilon_{i,\tilde{\alpha}}^2, \tag{26}$$

where $\tilde{\alpha} \in \mathcal{A}$ form a training set $\mathcal{A}$, and we have a supervised learning task with the data label y. z_i is the final prediction from the MLP model, $z_i^{(L)}$, η is the training rate. θ_μ is a vector combining all Ws and bs. ε here is the residual training error,

$$\varepsilon_{i,\tilde{\alpha}} = z_i(x_{\tilde{\alpha}}) - y_{i,\tilde{\alpha}}. \tag{27}$$

A.1 The fundamental difference between barren plateau and vanishing gradient

Firstly, we wish to comment on the fact that there is a fundamental difference between the barren plateau problem and the vanishing gradient problem.

The vanishing gradient problem is claimed to be a challenge of machine learning algorithms, where the gradient is vanishing for some neural network constructions, and it will be challenging to train the network [50, 51]. A standard and traditional explanation of the vanishing gradient problem is due to multiplicatively large number of layers in a deep neural network. The loss will have exponential behavior against some multiplicative factors during gradient descent, which will cause either exploding or vanishing of the loss function if there is no fine tuning. A resolution of the vanishing gradient problem is associated with the idea of *He initialization* or *Kaiming initialization*, which fine-tunes the neural network towards its critical point [52] (see also [23]).

The *barren plateau problem* is a term invented from the quantum community since [1]. As far as we know, there is no such term in classical machine learning instead of geography. The theoretical

argument from the barren plateau problem is the following, where we define the argument as *laziness*. If we consider the gradient descent process of the variational angles,

$$\theta_\mu(t+1) = \theta_\mu(t) - \eta \frac{d\mathcal{L}_\mathcal{A}}{d\theta_\mu}\bigg|_{\theta(t)} . \tag{28}$$

and if we make a sufficiently random variational ansatz, the factor $\mathrm{poly}(\dim \mathcal{H})$ where $\dim \mathcal{H}$ is the dimension of the Hilbert space, will appear in the formula of $d\mathcal{L}_A/d\theta_\mu$. Thus, the change of the variational angle will always suppressed by the dimension of the Hilbert space. A simple example of the Haar random factor $\mathrm{poly}(\dim \mathcal{H})$ will be the integration formula over a 2-design,

$$\int dU U_{ij} U^\dagger_{kl} = \frac{\delta_{il}\delta_{jk}}{\dim \mathcal{H}} , \tag{29}$$

where the matrix U forms a 2-design. The higher k is in a k-design, the higher factor of $\dim \mathcal{H}$ will appear if we consider higher moments of U. Thus, one claim that the variational angles almost cannot run in the randomized variational quantum architectures.

We could notice that the argument of the barren plateau problem using laziness is fundamentally different from the vanishing gradient problem: the vanishing gradient problem is *dynamical* when going to deeper and deeper neural networks, while the laziness is *static* and appears everywhere. Thus they are two intrinsically different problems. Moreover, from the similarity between the 2-design integral formula 29 and the LeCun parametrization 24, we could expect that the large-width neural networks will have similar behaviors: their weights and biases will also almost not run. Considering that classical overparametrized neural networks are proven to be practically useful (see, for instance, a comparison [53]), and large-scale neural networks could be implemented commonly nowadays, laziness may not always be bad in the actual machine learning tasks.

A.2 Classical large-width neural network has laziness as well

Now we prove that in the above setup, the large-width classical neural network will also have laziness. We have

$$\begin{aligned}
\frac{d\mathcal{L}_\mathcal{A}}{d\theta_\mu} &= \sum_{i,\tilde{\alpha}} \varepsilon_{i,\tilde{\alpha}} \frac{d\varepsilon_{i,\tilde{\alpha}}}{d\theta_\mu} = \sum_{i,\tilde{\alpha}} \varepsilon_{i,\tilde{\alpha}} \frac{dz_{i,\tilde{\alpha}}}{d\theta_\mu} \\
&= \sum_{i,\tilde{\alpha}} y_{i,\tilde{\alpha}} \frac{dz_{i,\tilde{\alpha}}}{d\theta_\mu} + \sum_{i,\tilde{\alpha}} z_{i,\tilde{\alpha}} \frac{dz_{i,\tilde{\alpha}}}{d\theta_\mu} .
\end{aligned} \tag{30}$$

We wish to represent the derivatives over W and b by the derivatives of early-layer preactivation $z^{(\ell)}$,

$$\begin{aligned}
&\frac{dz^{(L)}_{i;\alpha}}{db^{(\ell)}_j} = \frac{dz^{(L)}_{i;\alpha}}{dz^{(\ell)}_{j;\alpha}} , \\
&\frac{dz^{(L)}_{i;\alpha}}{dW^{(\ell)}_{jk}} = \sum_m \frac{dz^{(L)}_{i;\alpha}}{dz^{(\ell)}_{m;\alpha}} \frac{dz^{(\ell)}_{m;\alpha}}{dW^{(\ell)}_{jk}} = \frac{dz^{(L)}_{i;\alpha}}{dz^{(\ell)}_{j;\alpha}} \sigma^{(\ell-1)}_{k;\alpha} .
\end{aligned} \tag{31}$$

Here, $\sigma^{(\ell)}$ is a short-hand notation of $\sigma(z^{(\ell)})$, and we introduce $\sigma^{(\ell)}_{j;\alpha}$ as $\sigma(z^{(\ell)}_{j;\alpha})$. Finally, we have,

$$\begin{aligned}
&\frac{dz^{(L)}_{i;\alpha}}{dz^{(\ell)}_{j;\alpha}} = \sum_{k=1}^{n_{\ell+1}} \frac{dz^{(L)}_{i;\alpha}}{dz^{(\ell+1)}_{k;\alpha}} \frac{dz^{(\ell+1)}_{k;\alpha}}{dz^{(\ell)}_{j;\alpha}} = \sum_{k=1}^{n_{\ell+1}} \frac{dz^{(L)}_{i;\alpha}}{dz^{(\ell+1)}_{k;\alpha}} W^{(\ell+1)}_{kj} \sigma'^{(\ell)}_{j;\alpha} \\
&\text{for} \quad \ell < L , \\
&\frac{dz^{(L)}_{i;\alpha}}{dz^{(L)}_{j;\alpha}} = \delta_{ij} .
\end{aligned} \tag{32}$$

This is a back-propagation iterative formula, giving the recurrence relation from the end of the neural networks to the beginning. Moreover, we use σ' to denote derivatives of σ. So we get

$$\begin{aligned}
\frac{dz_{i;\alpha}^{(L)}}{dz_{j;\alpha}^{(\ell)}} &= \sum_{k=1}^{n_{\ell+1}} \frac{dz_{i;\alpha}^{(L)}}{dz_{k;\alpha}^{(\ell+1)}} \frac{dz_{k;\alpha}^{(\ell+1)}}{dz_{j;\alpha}^{(\ell)}} = \sum_{k=1}^{n_{\ell+1}} \frac{dz_{i;\alpha}^{(L)}}{dz_{k;\alpha}^{(\ell+1)}} W_{kj}^{(\ell+1)} \sigma_{j;\alpha}^{(\ell)\prime} \\
&= \sum_{k_{\ell+1},k_{\ell+2}}^{n_{\ell+1},n_{\ell+2}} \frac{dz_{i;\alpha}^{(L)}}{dz_{k_{\ell+2};\alpha}^{(\ell+2)}} W_{k_{\ell+2}j}^{(\ell+2)} W_{k_{\ell+1}j}^{(\ell+1)} \sigma_{j;\alpha}^{(\ell+1)\prime} \sigma_{j;\alpha}^{(L-2)\prime} \\
&= \sum_{k_{\ell+1},k_{\ell+2},\ldots,k_L}^{n_{\ell+1},n_{\ell+2},\ldots,n_L} \frac{dz_{i;\alpha}^{(L)}}{dz_{k_L;\alpha}^{(L)}} W_{k_Lj}^{(L)} W_{k_{L-1}j}^{(L-1)} \ldots W_{k_{\ell+2}j}^{(\ell+2)} W_{k_{\ell+1}j}^{(\ell+1)} \\
&\times \sigma_{j;\alpha}^{(L-1)\prime} \sigma_{j;\alpha}^{(L-2)\prime} \ldots \sigma_{j;\alpha}^{(\ell+1)\prime} \sigma_{j;\alpha}^{(L-2)\prime} \\
&= \sum_{k_{\ell+1},k_{\ell+2},\ldots,k_{L-1}}^{n_{\ell+1},n_{\ell+2},\ldots,n_{L-1}} W_{i,j}^{(L)} W_{k_{L-1}j}^{(L-1)} \ldots W_{k_{\ell+2}j}^{(\ell+2)} W_{k_{\ell+1}j}^{(\ell+1)} \\
&\sigma_{j;\alpha}^{(L-1)\prime} \sigma_{j;\alpha}^{(L-2)\prime} \ldots \sigma_{j;\alpha}^{(\ell+1)\prime} \sigma_{j;\alpha}^{(L-2)\prime} \,. \qquad (33)
\end{aligned}$$

We find the expectation value will vanish directly (which is exactly similar to the quantum case). Thus, we could estimate the norm by computing the variance of the gradients from,

$$\begin{aligned}
&\mathbb{E}\left(\left(\frac{dz_{i;\alpha}^{(L)}}{dz_{j;\alpha}^{(\ell)}}\right)^2\right) \\
&= \sum_{k_{\ell+1},k_{\ell+2},\ldots,k_{L-1},\bar{k}_{\ell+1},\bar{k}_{\ell+2},\ldots,\bar{k}_{L-1}}^{n_{\ell+1},n_{\ell+2},\ldots,n_{L-1},n_{\ell+1},n_{\ell+2},\ldots,n_{L-1}} \mathbb{E}\left(\begin{array}{l} W_{i,j}^{(L)} W_{i,j}^{(L)} W_{k_{L-1}j}^{(L-1)} W_{\bar{k}_{L-1}j}^{(L-1)} \cdots \\ W_{k_{\ell+2}j}^{(\ell+2)} W_{\bar{k}_{\ell+2}j}^{(\ell+2)} W_{k_{\ell+1}j}^{(\ell+1)} W_{\bar{k}_{\ell+1}j}^{(\ell+1)} \end{array} \right) \mathbb{E}\left(\left(\Sigma_{j;\alpha}^{(\ell);(L-1)}\right)^2\right) \\
&= \sum_{k_{\ell+1},k_{\ell+2},\ldots,k_{L-1},\bar{k}_{\ell+1},\bar{k}_{\ell+2},\ldots,\bar{k}_{L-1}}^{n_{\ell+1},n_{\ell+2},\ldots,n_{L-1},n_{\ell+1},n_{\ell+2},\ldots,n_{L-1}} \mathbb{E}\left(\begin{array}{l} W_{i,j}^{(L)} W_{i,j}^{(L)} W_{k_{L-1}j}^{(L-1)} W_{\bar{k}_{L-1}j}^{(L-1)} \cdots \\ W_{k_{\ell+2}j}^{(\ell+2)} W_{\bar{k}_{\ell+2}j}^{(\ell+2)} W_{k_{\ell+1}j}^{(\ell+1)} W_{\bar{k}_{\ell+1}j}^{(\ell+1)} \end{array} \right) \mathbb{E}\left(\left(\Sigma_{j;\alpha}^{(\ell);(L-1)}\right)^2\right) \\
&= \frac{1}{n_L} C_W^{(L)} C_W^{(L-1)} \ldots C_W^{(\ell+1)} \mathbb{E}\left(\left(\Sigma_{j;\alpha}^{(\ell);(L-1)}\right)^2\right) \,, \qquad (34)
\end{aligned}$$

where

$$\Sigma_{j;\alpha}^{(\ell);(L-1)} = \sigma_{j;\alpha}^{(L-1)\prime} \sigma_{j;\alpha}^{(L-2)\prime} \ldots \sigma_{j;\alpha}^{(\ell+2)\prime} \sigma_{j;\alpha}^{(\ell+1)\prime} \,. \qquad (35)$$

We have used the Wick contraction rule and the LeCun parametrization 24 according to [23]. Plug Equation 34 back to Equation 31, we see that this $1/n_L$ factor appears. This is the classical barren plateau in the large-width classical neural networks.

A.3 Classical large-width neural network could still learn efficiently

Here we show that the classical neural tangent kernel (NTK) will not vanish in classical MLPs, despite its laziness. This indicates that there are many good enough local minima around the point of initialization, so even the variational angles run slowly (the barren plateau problem), it will not matter for our practical purpose. On the other hand, more variational parameters will make us converge faster.

This part is a review of existing results, presented in the language of [23]. In classical MLPs, similar to the quantum cases we have discussed in the whole paper, the residual training error ε will decay exponentially at large width. We define the NTK as

$$H_{i_1 i_2;\alpha_1\alpha_2} \equiv \sum_{\mu} \frac{dz_{i_1;\alpha_1}}{d\theta_\mu} \frac{dz_{i_2;\alpha_2}}{d\theta_\mu} \,. \qquad (36)$$

The gradient descent rule will imply,

$$\delta\varepsilon_{i;\delta} = -\eta \sum_{i_1,\tilde{\alpha}\in\mathcal{A}} H_{ii_1;\delta\tilde{\alpha}} \varepsilon_{i_1,\tilde{\alpha}} \,. \qquad (37)$$

One could compute the average of the NTK. One could define the frozen NTK and the fluctuating NTK as

$$H_{i_1 i_2;\alpha_1\alpha_2} = \bar{H}_{i_1 i_2;\alpha_1\alpha_2} + \Delta H_{i_1 i_2;\alpha_1\alpha_2} \,, \tag{38}$$

and we have

$$\mathbb{E}\left(\Delta H_{i_1 i_2;\alpha_1\alpha_2}\Delta H_{i_3 i_4;\alpha_3\alpha_4}\right) = \frac{1}{n_{L-1}}\left[\delta_{i_1 i_2}\delta_{i_3 i_4}A_{(\alpha_1\alpha_2)(\alpha_3\alpha_4)} + \delta_{i_1 i_3}\delta_{i_2 i_4}B_{\alpha_1\alpha_3\alpha_2\alpha_4} + \delta_{i_1 i_4}\delta_{i_2 i_3}B_{\alpha_1\alpha_4\alpha_2\alpha_3}\right] \,. \tag{39}$$

The full expressions of A, B are given in Chapter 8 of [23]. Similarly, in the statistics language, one could check [31]. The suppression of ΔH in the large width indicates that the large-width neural networks will learn efficiently through non-trivial $\bar{H}_{i_1 i_2;\alpha_1\alpha_2}$, which is guaranteed to converge exponentially. In the large-width limit, the gradient descent algorithm is theoretically equivalent to the kernel method, where the kernel is defined effectively by NTKs. In Chapter 11 of [23], it is shown that dNTK, the higher-order corrections to the exponential decay, will vanish on its own, averaging over the Gaussian distribution of weights and bias. Moreover, the correlations between dNTK and other operators, which cause even numbers of Ws in total, will be suppressed by the large width polynomially. Those theoretical results are classical analogs of random unitary calculations done in our work.

B Some further details about concentration conditions

For concentration conditions including the quantum meta-kernel, one could see [2] for further details. Here we provide a simple review.

Now, we would like to ask when the QNTK approximation is valid. When the learning rate is small, the error of the prediction in Equation 15 could possibly come from two sources: the fluctuation of K about $\bar{K}$ during the gradient descent, and the higher-order corrections comparing the leading order Taylor expansion in Equation 11. The fluctuation ΔK could come from higher-order statistical calculations over the k-design assumption, similar to the analysis of higher-order effects in the barren plateau setup [26],

$$\Delta K = \sqrt{\mathbb{E}\left(\left(K-\bar{K}\right)^2\right)} \approx \frac{\sqrt{L}}{N^2}\sqrt{\left(8\mathrm{Tr}^2\left(O^2\right) + 12\mathrm{Tr}\left(O^4\right)\right)} \,, \tag{40}$$

in the large-N limit, and we present a detailed calculation in [2] with formulas up to 4-design. Moreover, we could look at higher order corrections to the Taylor expansion by the quantum meta-kernel (dQNTK) [29],

$$\begin{aligned}\delta\varepsilon &= -\eta\sum_{\ell}\frac{d\varepsilon}{d\theta_\ell}\frac{d\varepsilon}{d\theta_\ell}\varepsilon + \frac{1}{2}\eta^2\varepsilon^2\sum_{\ell_1,\ell_2}\frac{d^2\varepsilon}{d\theta_{\ell_1}d\theta_{\ell_2}}\frac{d\varepsilon}{d\theta_{\ell_1}}\frac{d\varepsilon}{d\theta_{\ell_2}} \\ &\equiv -\eta K\varepsilon + \frac{1}{2}\eta^2\varepsilon^2\mu \,. \end{aligned} \tag{41}$$

Here $\mu = \sum_{\ell_1,\ell_2}\frac{d^2\varepsilon}{d\theta_{\ell_1}d\theta_{\ell_2}}\frac{d\varepsilon}{d\theta_{\ell_1}}\frac{d\varepsilon}{d\theta_{\ell_2}}$ could be computed statistically using k-design formulas again. One can show that $\mathbb{E}(\mu) = 0$ (which is the same as its classical counterpart [23]), and we have

$$\Delta\mu = \sqrt{\mathbb{E}\left(\mu^2\right)} \approx \frac{\sqrt{32}L}{N^3}\mathrm{Tr}^{3/2}\left(O^2\right) \,, \tag{42}$$

in the large-N limit. The condition where the QNTK estimation in Equation 15 is valid when

$$\Delta K \ll K \Leftrightarrow L \gg 1 \,, \tag{43}$$

$$\begin{aligned}&\frac{1}{2}\eta^2\varepsilon^2\Delta\mu \ll \eta\bar{K}\varepsilon \Leftrightarrow \eta\varepsilon(0)\frac{L}{N^3}\mathrm{Tr}^{3/2}\left(O^2\right) \ll \frac{L\mathrm{Tr}\left(O^2\right)}{N^2} \\ &\Leftrightarrow \frac{\eta\Omega_O}{N}\varepsilon(0) \ll 1 \,. \end{aligned} \tag{44}$$

We call the conditions 43 and 44 as the *concentration conditions*. Here, we denote $\varepsilon(0) = \varepsilon(t = 0)$, and we assume that $\text{Tr}(O^2) \equiv \Omega_O^2 > \text{Tr}^2(O)$. This is correct, for instance, if O is a Pauli operator, where we have $\text{Tr}(O^2) = N$ but $\text{Tr}^2(O) = 0$.

Note that the condition Equation 44 is a weak condition. It only tells that how small η is needed to make sure the nearly expansion is valid. In practice, we often assume that $\eta < \mathcal{O}(1)$ and $\Omega_O \geq \mathcal{O}(N)$, so Equation 44 is automatically satisfied. The condition that usually matters is Equation 43, which is the definition of overparametrization here $L \gg 1$. Thus, if L is large, the prediction will be correct, no matter how large N is. But if N is large, the decay rate itself $\bar{K}$ will be small. So this is exactly the definition of the barren plateau!

Furthermore, we wish to mention that if we only count for powers of N and L, we have

$$\frac{\Delta K}{\bar{K}} = \mathcal{O}\left(\frac{1}{\sqrt{L}}\right) , \frac{\Delta \mu}{\bar{K}} = \mathcal{O}\left(\frac{1}{N}\right) . \tag{45}$$

If we demand $\bar{K} = \mathcal{O}(1)$ and ignore η, we get $L = \mathcal{O}(N)$, so we get $\frac{\Delta K}{\bar{K}} = \mathcal{O}\left(\frac{1}{N}\right)$ as well. The $1/N$ or $1/\text{width}$ expansion is exactly observed in the classical neural networks [23]. The origin of this equivalence comes from the similarity between Equation 2 and Equation 46, while a higher level (but heuristic) understanding comes from a connection between quantum field theory and the large-width expansion [23, 37, 38] and a similarity between Feynman rules in quantum field theory and matrix models [54], which we will briefly explain in Appendix C for readers who are interested in how observations about this paper might be discovered from another perspective.

C A physical interpretation

Here we make some comments about possible, heuristic, physical interpretations of the agreement between classical and quantum neural networks. There is a duality, pointed out in [23, 37–39] where the large-width classical neural networks could be understood in the quantum field theory language. In the large-width limit, the output of neural networks will follow a Gaussian process, averaging with respect to Gaussian distribution over weights and bias according to the LeCun parametrization,

$$\mathbb{E}\left(W_{ij}W_{kl}\right) = \frac{\sigma_W^2}{\text{width}}\delta_{ik}\delta_{jl} , \tag{46}$$

or more generally,

$$\mathbb{E}\left(W_{i_1j_1}W_{i_2j_2}\dots W_{i_{2k-1}j_{2k-1}}W_{i_{2k}j_{2k}}\right) = \mathcal{O}(\frac{1}{\text{poly}(\text{width})}) , \tag{47}$$

for all positive integer k. Here, we are considering the multilayer perceptron (MLP) model with weights W, and the width is defined as the number of neurons in each layer. The limit is mathematically similar to the large-N limit of gauge theories, which becomes almost generalized free theories. We could understand the ratio between the depth, the number of layers, and the width, the number of neurons, as perturbative corrections against the Gaussian process, which is similar to what we have done in the large-N expansion of gauge theories.

This physical interpretation will be helpful also when we consider its quantum generalization. If classical MLPs are similar to quantum field theories, quantum neural networks will be similar to matrix models [55, 56]. Matrix models have been studied for a long time, around and after the second string theory revolution [54], and they have deep connections to the holographic principle [57] and the AdS/CFT correspondence [58, 59]. Haar ensembles are toy versions of matrix models, which have been widely studied as toy models of chaotic quantum black holes [17, 60]. The similarity between the LeCun parametrization 46 and the 1-design Haar integral formula

$$\mathbb{E}(U_{ij}U_{kl}^\dagger) = \frac{1}{\dim \mathcal{H}}\delta_{il}\delta_{jk} , \tag{48}$$

or more generally,

$$\mathbb{E}\left(U_{i_1j_1}U_{i_2j_2}^\dagger\dots U_{i_{2k-1}j_{2k-1}}U_{i_{2k}j_{2k}}^\dagger\right) = \mathcal{O}(\frac{1}{\text{poly}(\dim \mathcal{H})}) , \tag{49}$$

where $\dim \mathcal{H}$ is the dimension of the Hilbert space, might be potentially related to the similarity of Feynman rules between matrix models and quantum field theories. Thus, the similarity between quantum and classical neural networks might have a physical interpretation between matrix models and their effective field theory descriptions.

The above analogy is heuristic. We should point out that machine learning and physical systems are very different. Some mathematical similarities could provide guidance towards new discoveries and better insights, but we have to be careful that they are intrinsically different phenomena.

D Noises

Now let us add the affection of the noise. From the original gradient descent equation,

$$\theta_\ell(t+1) - \theta_\ell(t) \equiv \delta\theta_\mu = -\eta \frac{\partial \mathcal{L}}{\partial \theta_\ell} = i\eta \left\langle \Psi_0 \left| V_{+,\ell}^\dagger \left[X_\ell, V_{-,\ell}^\dagger O V_{-,\ell} \right] V_{+,\ell} \right| \Psi_0 \right\rangle \ , \tag{50}$$

we add a random fluctuation term $\Delta\theta_\ell$ to model the uncertainty of measuring the expectation value. We assume that the random variable $\Delta\theta_\ell$ is Markovian. Namely, it is independent for the time step t. Moreover, we assume that $\Delta\theta_\ell$s are distributed with Gaussian distributions $\mathcal{N}(0, \sigma_\theta^2)$.

Thus, the residual training error has the recursion relation in the linear order of the Taylor expansion,

$$\delta\varepsilon = -\eta\varepsilon K + \sum_\ell \frac{\partial \varepsilon}{\partial \theta_\ell} \Delta\theta_\ell \ . \tag{51}$$

Now, let us assume that K is still a constant. Since $\Delta\theta_\ell \sim \mathcal{N}(0, \sigma_\theta^2)$, we get

$$\sum_\ell \frac{\partial \varepsilon}{\partial \theta_\ell} \Delta\theta_\ell \sim \mathcal{N}(0, K\sigma_\theta^2) \ . \tag{52}$$

Thus, we could write the recursion relation as

$$\delta\varepsilon = -\eta\varepsilon K + \sqrt{K}\Delta\theta \ . \tag{53}$$

Here, $\Delta\theta \approx \mathcal{N}(0, \sigma_\theta^2)$. One can solve the difference equation iteratively. The answer is

$$\varepsilon(t) = (1-\eta K)^t \varepsilon(0) + \sqrt{K} \sum_{i=0}^{t-1} (1-\eta K)^i \Delta\theta(t-1-i) \ . \tag{54}$$

Now, we have

$$\begin{aligned} &\sqrt{K} \sum_{i=0}^{t-1} (1-\eta K)^i \Delta\theta(t-1-i) \sim \mathcal{N}(0, K\sigma_\theta^2 \sum_{i=0}^{t-1} (1-\eta K)^{2i}) \\ &= \mathcal{N}(0, \sigma_\theta^2 \frac{1-(1-\eta K)^{2t}}{\eta(2-\eta K)}) \ . \end{aligned} \tag{55}$$

At the initial time $t = 0$, there is no effect of noise. The relative size of the error will grow during time compared to the exponential decay term without noises. Based on the distribution, we could compute the average ε^2 against the noises, $\overline{\varepsilon^2}$, as

$$\overline{\varepsilon^2}(t) = (1-\eta K)^{2t} \left(\varepsilon^2(0) - \frac{\sigma_\theta^2}{\eta(2-\eta K)} \right) + \frac{\sigma_\theta^2}{\eta(2-\eta K)} \ . \tag{56}$$

Note that the first term is decaying when the time t is increasing. At the late time, we have

$$\overline{\varepsilon^2}(\infty) = \frac{\sigma_\theta^2}{\eta(2-\eta K)} \approx \mathcal{O}(\frac{\sigma_\theta^2}{\eta}) \ , \tag{57}$$

where we assume the overparametrization $\eta K \approx \mathcal{O}(1)$. Thus, at the late time, the loss function will arrive at a constant plateau at $\mathcal{O}(\sigma_\theta^2/\eta)$. One could improve σ_θ to make the constant plateau controllable and do not increase significantly with N, indicating that our algorithm could be noise-resilient.

One could also estimate the time scale where the contribution of the noise could emerge. We could define the time scale, T_{noise}, as,

$$(1-\eta K)^{T_{\text{noise}}}\varepsilon(0) \approx \sigma_\theta \sqrt{\frac{1-(1-\eta K)^{2T_{\text{noise}}}}{\eta(2-\eta K)}} \ . \tag{58}$$

It means that at T_{noise}, the noise contribution is comparable to the noiseless part in the residual training error. We have,

$$\begin{aligned} T_{\text{noise}} &\approx \frac{\log\left(\frac{\sigma_\theta}{\sqrt{2\varepsilon^2(0)\eta-\varepsilon^2(0)\eta^2 K+\sigma_\theta^2}}\right)}{\log(1-\eta K)} \ , \\ \varepsilon(T_{\text{noise}}) &= 2(1-\eta K)^{T_{\text{noise}}}\varepsilon(0) = \frac{2\sigma_\theta^2}{\sqrt{\varepsilon(0)^2(2\eta-\eta^2 K)+\sigma_\theta^2}}\varepsilon(0) \ . \end{aligned} \tag{59}$$

We find that choosing $\eta \approx \mathcal{O}(1/K)$ will minimize $\varepsilon(T_{\text{noise}})$. It is exactly the overparametrization condition we use in this paper.

To be self-consistent, we need to check if the choice $\eta \approx \mathcal{O}(1/K)$ is consistent with the concentration condition about dQNTK. In fact, we find that $\eta \approx \mathcal{O}(1/K)$ will naturally satisfy the dQNTK concentration condition if $\varepsilon(0) < \mathcal{O}(L\sqrt{N})$. This is naturally satisfied in generic situations in variational quantum algorithms since we will usually not have an exponential amount of residual training error initially.

E Numerical results

In this part, we show some simple numerical evidences based on the analysis done in [2]. We will use the randomized version of the hardware-efficient variational ansatz defined in [2]. In Figure 2, for each σ_θ value, we run 10 experiments of 100 steps using the same setup of the ansatz $U(\theta)$, the operator O and the input state θ_0 as in [2]. After that, we get the residual error of the last step and take the average value over 10 experiments to get the mean ε value, shown with black dots in the figure. The red line in the figure is the theoretical prediction. In these experiments, $L = 64$, and we have 4 qubits. We can further get the analytic result of the mean value of $\overline{\varepsilon}$ after a long time as

$$\overline{\varepsilon} = \sqrt{\frac{2}{\pi}} \cdot \frac{\sigma_\theta}{\sqrt{2\eta-\eta^2 K}} \ , \tag{60}$$

where the K value is taken from the value of the last step, as it fluctuates a lot in the early time.

We run multiple experiments to approach the theoretical value as much as possible, where 10 experiments are done for each σ_θ value. To verify that the numerical result lies in a reasonable regime, we calculated the 90% confidence interval of ε theoretically.

To compensate for the effect of large K on our numerical simulations, since in every experiment setup, due to randomness, the training will lead the parameters to different regimes of different Ks, we choose those experiments which fulfill our theoretical restrictions for small K. The numerical results above are with $K \approx \mathcal{O}(10)$, which still shows great agreement with our theoretical formalism.

More precisely, in Figure 2, we get the relationship between residual error fluctuation and noise. For each σ_θ value, we calculated the standard deviation with final residual error data from 10 experiments, shown as black dots. The final residual error that we get from the numerical experiments is taken absolute value for the benefit of the log scale. We find the numerical results follow the theoretical prediction in a reasonable confidence interval. Moreover, we verify the extent of our final residual error that can achieve as a function of noise σ_θ with numerical evidence.

In Figure 3, we verify the prediction of standard deviation of $\varepsilon(\infty)$, σ_ε, in the small η regime. In these numerical experiments, the inaccuracy comes mainly from a limited number of experiments and a limited time scale ($t = 100$). Especially for experiments with a small learning rate η with random initial states, T_{noise} may be large for 100 steps to cover.

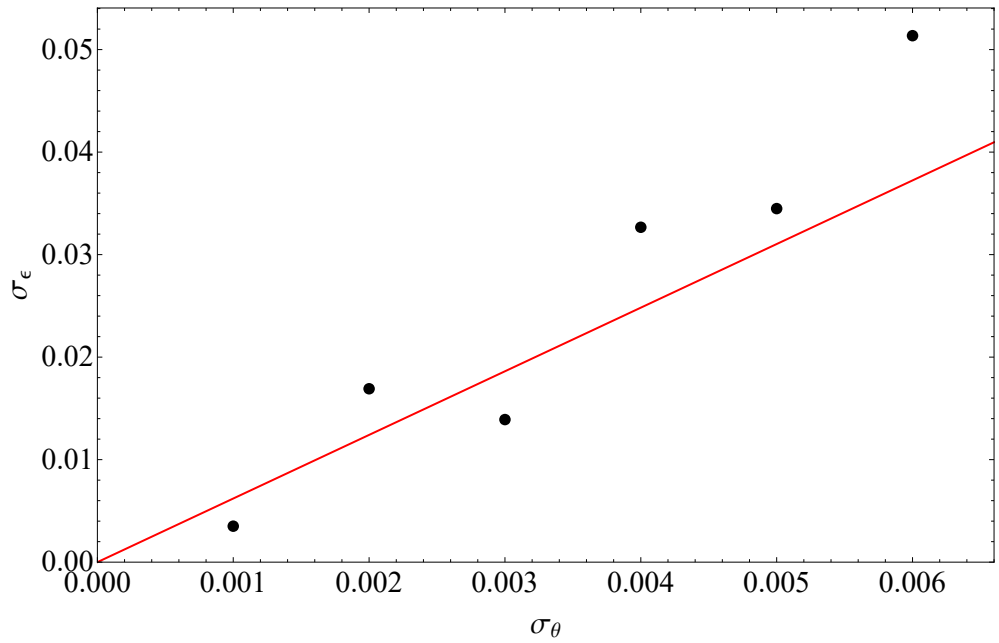

Figure 2: Noise standard deviation σ_θ as a function of standard deviation of final residual error σ_ε after training long enough time, with both numerical result (black dots) and theoretical prediction (red line). In this figure, $\eta = 0.005$, $K \approx 25$, $\varepsilon(0) \approx 1$.

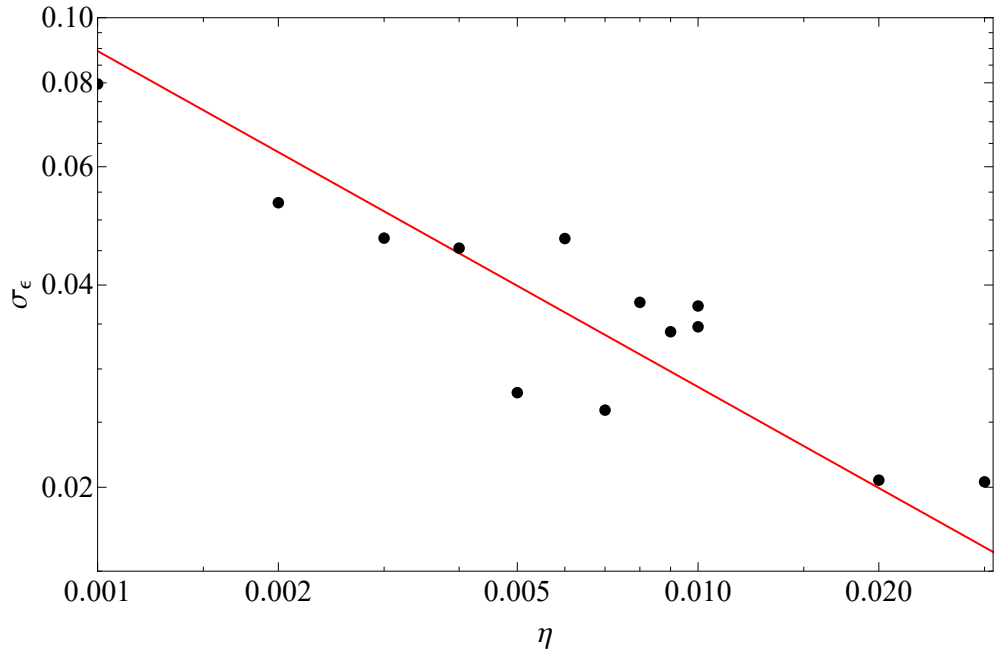

Figure 3: Standard deviation of final residual error σ_ε as a function of learning rate η after training long enough time, with both numerical result (black dots) and theoretical prediction (red line). In this figure, $\sigma_\theta = 0.005$, $K \approx 35$, $\varepsilon(0) \approx 1$, $t = 100$.

References

[1] Jarrod R McClean, Sergio Boixo, Vadim N Smelyanskiy, Ryan Babbush, and Hartmut Neven. Barren plateaus in quantum neural network training landscapes. *Nature communications*, 9(1): 1–6, 2018.

[2] Junyu Liu, Khadijeh Najafi, Kunal Sharma, Francesco Tacchino, Liang Jiang, and Antonio Mezzacapo. An analytic theory for the dynamics of wide quantum neural networks. 3 2022.

[3] Alberto Peruzzo, Jarrod McClean, Peter Shadbolt, Man-Hong Yung, Xiao-Qi Zhou, Peter J Love, Alán Aspuru-Guzik, and Jeremy L O?brien. A variational eigenvalue solver on a photonic quantum processor. *Nature communications*, 5(1):1–7, 2014.

[4] M-H Yung, Jorge Casanova, Antonio Mezzacapo, Jarrod Mcclean, Lucas Lamata, Alan Aspuru-Guzik, and Enrique Solano. From transistor to trapped-ion computers for quantum chemistry. *Scientific reports*, 4(1):1–7, 2014.

[5] Jarrod R McClean, Jonathan Romero, Ryan Babbush, and Alán Aspuru-Guzik. The theory of variational hybrid quantum-classical algorithms. *New Journal of Physics*, 18(2):023023, 2016.

[6] Abhinav Kandala, Antonio Mezzacapo, Kristan Temme, Maika Takita, Markus Brink, Jerry M Chow, and Jay M Gambetta. Hardware-efficient variational quantum eigensolver for small molecules and quantum magnets. *Nature*, 549(7671):242–246, 2017.

[7] Marco Cerezo, Andrew Arrasmith, Ryan Babbush, Simon C Benjamin, Suguru Endo, Keisuke Fujii, Jarrod R McClean, Kosuke Mitarai, Xiao Yuan, Lukasz Cincio, et al. Variational quantum algorithms. *Nature Reviews Physics*, pages 1–20, 2021.

[8] Edward Farhi, Jeffrey Goldstone, and Sam Gutmann. A quantum approximate optimization algorithm. *arXiv preprint arXiv:1411.4028*, 2014.

[9] Peter Wittek. *Quantum machine learning: what quantum computing means to data mining*. Academic Press, 2014.

[10] Nathan Wiebe, Ashish Kapoor, and Krysta M Svore. Quantum deep learning. *arXiv preprint arXiv:1412.3489*, 2014.

[11] Jacob Biamonte, Peter Wittek, Nicola Pancotti, Patrick Rebentrost, Nathan Wiebe, and Seth Lloyd. Quantum machine learning. *Nature*, 549(7671):195–202, 2017.

[12] Maria Schuld and Nathan Killoran. Quantum machine learning in feature hilbert spaces. *Physical review letters*, 122(4):040504, 2019.

[13] Vojtěch Havlíček, Antonio D Córcoles, Kristan Temme, Aram W Harrow, Abhinav Kandala, Jerry M Chow, and Jay M Gambetta. Supervised learning with quantum-enhanced feature spaces. *Nature*, 567(7747):209–212, 2019.

[14] Yunchao Liu, Srinivasan Arunachalam, and Kristan Temme. A rigorous and robust quantum speed-up in supervised machine learning. *Nature Physics*, pages 1–5, 2021.

[15] Junyu Liu. *Does Richard Feynman Dream of Electric Sheep? Topics on Quantum Field Theory, Quantum Computing, and Computer Science*. PhD thesis, Caltech, 2021.

[16] Edward Farhi and Hartmut Neven. Classification with quantum neural networks on near term processors. *arXiv preprint arXiv:1802.06002*, 2018.

[17] Daniel A. Roberts and Beni Yoshida. Chaos and complexity by design. *JHEP*, 04:121, 2017. doi: 10.1007/JHEP04(2017)121.

[18] Jordan Cotler, Nicholas Hunter-Jones, Junyu Liu, and Beni Yoshida. Chaos, Complexity, and Random Matrices. *JHEP*, 11:048, 2017. doi: 10.1007/JHEP11(2017)048.

[19] Junyu Liu. Spectral form factors and late time quantum chaos. *Phys. Rev. D*, 98(8):086026, 2018. doi: 10.1103/PhysRevD.98.086026.

[20] Junyu Liu. Scrambling and decoding the charged quantum information. *Phys. Rev. Res.*, 2: 043164, 2020. doi: 10.1103/PhysRevResearch.2.043164.

[21] Motohisa Fukuda, Robert König, and Ion Nechita. Rtni: A symbolic integrator for haar-random tensor networks. *Journal of Physics A: Mathematical and Theoretical*, 52(42):425303, 2019.

[22] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2018.

[23] Daniel A Roberts, Sho Yaida, and Boris Hanin. The principles of deep learning theory. *arXiv preprint arXiv:2106.10165*, 2021.

[24] Marco Cerezo, Akira Sone, Tyler Volkoff, Lukasz Cincio, and Patrick J Coles. Cost function dependent barren plateaus in shallow parametrized quantum circuits. *Nature communications*, 12(1):1–12, 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. *Physical Review X*, 11(4):041011, 2021.

[26] Marco Cerezo and Patrick J Coles. Higher order derivatives of quantum neural networks with barren plateaus. *Quantum Science and Technology*, 6(3):035006, 2021.

[27] Andrew Arrasmith, M Cerezo, Piotr Czarnik, Lukasz Cincio, and Patrick J Coles. Effect of barren plateaus on gradient-free optimization. *Quantum*, 5:558, 2021.

[28] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

[29] Junyu Liu, Francesco Tacchino, Jennifer R. Glick, Liang Jiang, and Antonio Mezzacapo. Representation Learning via Quantum Neural Tangent Kernels. 11 2021.

[30] Jaehoon Lee, Yasaman Bahri, Roman Novak, Samuel S Schoenholz, Jeffrey Pennington, and Jascha Sohl-Dickstein. Deep neural networks as gaussian processes. *arXiv preprint arXiv:1711.00165*, 2017.

[31] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *arXiv preprint arXiv:1806.07572*, 2018.

[32] 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. *Advances in neural information processing systems*, 32:8572–8583, 2019.

[33] Jascha Sohl-Dickstein, Roman Novak, Samuel S Schoenholz, and Jaehoon Lee. On the infinite width limit of neural networks with a standard parameterization. *arXiv preprint arXiv:2001.07301*, 2020.

[34] Greg Yang and Edward J Hu. Feature learning in infinite-width neural networks. *arXiv preprint arXiv:2011.14522*, 2020.

[35] Sho Yaida. Non-gaussian processes and neural networks at finite widths. In *Mathematical and Scientific Machine Learning*, pages 165–192. PMLR, 2020.

[36] Sanjeev Arora, Simon S Du, Wei Hu, Zhiyuan Li, Ruslan Salakhutdinov, and Ruosong Wang. On exact computation with an infinitely wide neural net. *arXiv preprint arXiv:1904.11955*, 2019.

[37] Ethan Dyer and Guy Gur-Ari. Asymptotics of wide networks from feynman diagrams. *arXiv preprint arXiv:1909.11304*, 2019.

[38] James Halverson, Anindita Maiti, and Keegan Stoner. Neural networks and quantum field theory. *Machine Learning: Science and Technology*, 2(3):035002, 2021.

[39] Daniel A Roberts. Why is ai hard and physics simple? *arXiv preprint arXiv:2104.00008*, 2021.

[40] Daniel A Roberts and Sho Yaida. Effective theory of deep learning: Beyond the infinite-width limit. *Deep Learning Theory Summer School at Princeton*, 2021.

[41] Norihito Shirai, Kenji Kubo, Kosuke Mitarai, and Keisuke Fujii. Quantum tangent kernel. *arXiv preprint arXiv:2111.02951*, 2021.

[42] Samson Wang, Enrico Fontana, Marco Cerezo, Kunal Sharma, Akira Sone, Lukasz Cincio, and Patrick J Coles. Noise-induced barren plateaus in variational quantum algorithms. *Nature communications*, 12(1):1–11, 2021.

[43] Emanuel Knill, Gerardo Ortiz, and Rolando D Somma. Optimal quantum measurements of expectation values of observables. *Physical Review A*, 75(1):012328, 2007.

[44] Lieven MK Vandersypen and Isaac L Chuang. Nmr techniques for quantum control and computation. *Reviews of modern physics*, 76(4):1037, 2005.

[45] Erfan Abedi, Salman Beigi, and Leila Taghavi. Quantum lazy training. *arXiv preprint arXiv:2202.08232*, 2022.

[46] Xuchen You, Shouvanik Chakrabarti, and Xiaodi Wu. A convergence theory for over-parameterized variational quantum eigensolvers. *arXiv preprint arXiv:2205.12481*, 2022.

[47] Iris Cong, Soonwon Choi, and Mikhail D Lukin. Quantum convolutional neural networks. *Nature Physics*, 15(12):1273–1278, 2019.

[48] Martin Larocca, Nathan Ju, Diego García-Martín, Patrick J Coles, and M Cerezo. Theory of overparametrization in quantum neural networks. *arXiv preprint arXiv:2109.11676*, 2021.

[49] Eric R. Anschuetz and Bobak T. Kiani. Beyond Barren Plateaus: Quantum Variational Algorithms Are Swamped With Traps. 5 2022.

[50] Sepp Hochreiter. The vanishing gradient problem during learning recurrent neural nets and problem solutions. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 6(02):107–116, 1998.

[51] Sepp Hochreiter, Yoshua Bengio, Paolo Frasconi, Jürgen Schmidhuber, et al. Gradient flow in recurrent nets: the difficulty of learning long-term dependencies, 2001.

[52] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015.

[53] Anna Golubeva, Behnam Neyshabur, and Guy Gur-Ari. Are wider nets better given the same number of parameters? *arXiv preprint arXiv:2010.14495*, 2020.

[54] Edward Witten. String theory dynamics in various dimensions. *Nucl. Phys. B*, 443:85–126, 1995. doi: 10.1016/0550-3213(95)00158-O.

[55] Tom Banks, W. Fischler, S. H. Shenker, and Leonard Susskind. M theory as a matrix model: A Conjecture. *Phys. Rev. D*, 55:5112–5128, 1997. doi: 10.1103/PhysRevD.55.5112.

[56] David Eliecer Berenstein, Juan Martin Maldacena, and Horatiu Stefan Nastase. Strings in flat space and pp waves from N=4 superYang-Mills. *JHEP*, 04:013, 2002. doi: 10.1088/1126-6708/2002/04/013.

[57] Leonard Susskind. The World as a hologram. *J. Math. Phys.*, 36:6377–6396, 1995. doi: 10.1063/1.531249.

[58] Juan Martin Maldacena. The Large N limit of superconformal field theories and supergravity. *Adv. Theor. Math. Phys.*, 2:231–252, 1998. doi: 10.1023/A:1026654312961.

[59] Edward Witten. Anti-de Sitter space and holography. *Adv. Theor. Math. Phys.*, 2:253–291, 1998. doi: 10.4310/ATMP.1998.v2.n2.a2.

[60] Patrick Hayden and John Preskill. Black holes as mirrors: Quantum information in random subsystems. *JHEP*, 09:120, 2007. doi: 10.1088/1126-6708/2007/09/120.