Supplemental Material: Representation Learning via Quantum Neural Tangent Kernels

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

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

(Dated: July 13, 2022)

∗: Corresponding author.

CONTENTS

I. DETAILS ON THE LINEAR AND THE QUADRATIC MODELS

A. Linear model

Before we discuss the quantum model setup, we start by reviewing the linear and quadratic models in the classical setup. The linear and quadratic models are originally solved in the framework presented by [1, 2], and in this section, we will give a brief review.

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

We define the model as

$$z_{i;\delta}(\theta) = \sum_j W_{ij}\phi_j\left(\mathbf{x}_\delta\right) \ . \tag{1}$$

Here $\mathbf{x}_\delta$ is the data point δ in the space $\mathcal{D}$, and W_{ij}s are weights and biases. We use the slack notation such that $W_{0j} = b_j$ includes the bias. The model is called the linear model, which is linear in weights, but we wish to keep the feature map ϕ in general. Moreover, we write θ as a compact notation of the vectorized W.

We wish to optimize the following loss function

$$\mathcal{L}_{\mathcal{A}}(\theta) = \frac{1}{2}\sum_{i,\tilde{\alpha}}\left[y_{i;\tilde{\alpha}} - \sum_j W_{ij}\phi_j\left(\mathbf{x}_{\tilde{\alpha}}\right)\right]^2 \ . \tag{2}$$

For the sample set $\mathcal{A}$. Again, we will use the notation $\tilde{\alpha}$ such that $\tilde{\alpha}$ is in the sample set $\mathcal{A}$. We define the kernel

$$k_{\delta_1\delta_2} \equiv k\left(\mathbf{x}_{\delta_1}, \mathbf{x}_{\delta_2}\right) \equiv \sum_i \phi_i\left(\mathbf{x}_{\delta_1}\right)\phi_i\left(\mathbf{x}_{\delta_2}\right) = \sum_{a,b}\frac{dz_{i;\delta_1}}{dW_{ab}}\frac{dz_{i;\delta_2}}{dW_{ab}} \ . \tag{3}$$

Note that the right hand side does not contain a sum over i. It is equal for all is. Moreover, We call $\delta \in \mathcal{D}$ as the whole data set, while $\mathcal{A} \subset \mathcal{D}$ is the sample set. We define $\widetilde{k}_{\tilde{\alpha}_1\tilde{\alpha}_2}$ with tilde to indicate the submatrix. We also define [3]

$$\sum_{\tilde{\alpha}_2\in\mathcal{A}} \widetilde{k}^{\tilde{\alpha}_1\tilde{\alpha}_2}\widetilde{k}_{\tilde{\alpha}_2\tilde{\alpha}_3} = \delta^{\tilde{\alpha}_1}_{\tilde{\alpha}_3} \ . \tag{4}$$

namely, the upper index means the inverse. Similar to the main text, we consider the gradient descent algorithm,

$$đ\theta_\mu = -\eta\frac{\partial\mathcal{L}_{\mathcal{A}}}{\partial\theta_\mu} \ . \tag{5}$$

The partial derivatives are computed as

$$\begin{aligned}\frac{\partial L_A(W)}{\partial W_{ab}} &= -\sum_{\tilde{\alpha},i,j}\delta_{ia}\delta_{jb}\phi_j\left(\mathbf{x}_{\tilde{\alpha}}\right)\left[y_{i;\tilde{\alpha}} - \sum_j W_{ij}\phi_j\left(\mathbf{x}_{\tilde{\alpha}}\right)\right] \\ &= \sum_{\tilde{\alpha}}\phi_b\left(\mathbf{x}_{\tilde{\alpha}}\right)\left(z_{a;\tilde{\alpha}} - y_{a;\tilde{\alpha}}\right) = \sum_{\tilde{\alpha}}\varepsilon_{a;\tilde{\alpha}}\phi_b\left(\mathbf{x}_{\tilde{\alpha}}\right)\end{aligned} \tag{6}$$

where ε is the residual training error,

$$đW_{ij} = -\eta\sum_{\tilde{\alpha}}\phi_j\left(\mathbf{x}_{\tilde{\alpha}}\right)\varepsilon_{i;\tilde{\alpha}} \ . \tag{7}$$

So we have

$$đz_{i;\delta} = \sum_{a,b}\frac{\partial z_{i;\delta}}{\partial W_{ab}}đW_{ab} = -\eta\sum_{\tilde{\alpha}}k_{\delta\tilde{\alpha}}\varepsilon_{i;\tilde{\alpha}} \ . \tag{8}$$

In the linear model, the kernel $k_{\delta\tilde{\alpha}}$ is static. The solution for the residual training error is

$$\varepsilon_{i;\tilde{\alpha}_1}(t) = \sum_{\tilde{\alpha}_2} U_{\tilde{\alpha}_1\tilde{\alpha}_2}(t)\varepsilon_{i;\tilde{\alpha}_2}(0) \ , \tag{9}$$

where

$$\begin{aligned}U_{\tilde{\alpha}_t\tilde{\alpha}_0}(t) &\equiv \left[(1-\eta k)^t\right]_{\tilde{\alpha}_t\tilde{\alpha}_0} \\ &= \sum_{\tilde{\alpha}_1,\ldots,\tilde{\alpha}_{t-1}}\left(\delta_{\tilde{\alpha}_t\tilde{\alpha}_{t-1}} - \eta k_{\tilde{\alpha}_t\tilde{\alpha}_{t-1}}\right)\cdots\left(\delta_{\tilde{\alpha}_1\tilde{\alpha}_0} - \eta k_{\tilde{\alpha}_1\tilde{\alpha}_0}\right) \ .\end{aligned} \tag{10}$$

Now we could predict the model output for an arbitrary δ. We have

$$\begin{aligned} z_{i;\delta}(\infty) &= z_{i;\delta}(0) - \sum_{\tilde{\alpha}} k_{\delta\tilde{\alpha}} \left\{ \eta \sum_{s=0}^{\infty} \varepsilon_{i;\tilde{\alpha}}(s) \right\} \\ &= z_{i;\delta}(0) - \sum_{\tilde{\alpha}} k_{\delta\tilde{\alpha}} \left\{ \eta \sum_{s=0}^{\infty} \left[\sum_{\tilde{\alpha}_1} U_{\tilde{\alpha}\tilde{\alpha}_1}(s) \varepsilon_{i;\tilde{\alpha}_1}(0) \right] \right\} \\ &= z_{i;\delta}(0) - \sum_{\tilde{\alpha},\tilde{\alpha}_1} k_{\delta\tilde{\alpha}} \left\{ \eta \sum_{s=0}^{\infty} [(1-\eta k)^s]_{\tilde{\alpha}\tilde{\alpha}_1} \right\} \varepsilon_{i;\tilde{\alpha}_1}(0) \\ &= z_{i;\delta}(0) - \sum_{\tilde{\alpha},\tilde{\alpha}_1} k_{\delta\tilde{\alpha}} \left\{ \eta [1-(1-\eta k)]^{-1} \right\}^{\tilde{\alpha}\tilde{\alpha}_1} \varepsilon_{i;\tilde{\alpha}_1}(0) \\ &= z_{i;\delta}(0) - \sum_{\tilde{\alpha},\tilde{\alpha}_1} k_{\delta\tilde{\alpha}} \tilde{k}^{\tilde{\alpha}\tilde{\alpha}_1} \varepsilon_{i;\tilde{\alpha}_1}(0) \, , \end{aligned} \tag{11}$$

where we have made use of geometric sums. It is easy to check that

$$z_{i;\tilde{\delta}}(\infty) = y_{i;\tilde{\delta}} \, , \tag{12}$$

where we take $\tilde{\delta} \in \mathcal{A}$.

B. Quadratic model

Now we start to study perturbative corrections about the linear model. We consider the model definition,

$$z_{i;\delta}(\theta) = \sum_j W_{ij} \phi_j(\mathbf{x}_\delta) + \frac{\sigma}{2} \sum_{j_1,j_2} W_{ij_1} W_{ij_2} \psi_{j_1 j_2}(\mathbf{x}_\delta) \ . \tag{13}$$

Here σ is a small number as a perturbative correction. Thus, the model prediction difference up to the quadratic order will be given by

$$đz_{i;\delta} = \sum_j \left[\phi_j(\mathbf{x}_\delta) + \varepsilon \sum_{j_1=0} W_{ij_1} \psi_{j_1 j}(\mathbf{x}_\delta) \right] đW_{ij} + \frac{\sigma}{2} \sum_{j_1,j_2} \psi_{j_1 j_2}(\mathbf{x}_\delta) \, đW_{ij_1} đW_{ij_2} \, . \tag{14}$$

The first term here is the *effective feature map*,

$$\phi^E_{ij}(\mathbf{x}_\delta) \equiv \frac{dz_{i;\delta}}{dW_{ij}} = \phi_j(\mathbf{x}_\delta) + \sigma \sum_k W_{ik} \psi_{kj}(\mathbf{x}_\delta) \ . \tag{15}$$

Now, we could collect our dynamical equations for z, ϕ^E, W as

$$\begin{aligned} đz_{i;\delta} &= \sum_j đW_{ij} \phi^E_{ij}(\mathbf{x}_\delta) + \frac{\sigma}{2} \sum_{j_1,j_2} đW_{ij_1} đW_{ij_2} \psi_{j_1 j_2}(\mathbf{x}_\delta) \ , \\ đ\phi^E_{ij}(\mathbf{x}_\delta) &= \sigma \sum_k đW_{ik} \psi_{kj}(\mathbf{x}_\delta) \ , \\ đW_{ij} &= -\eta \sum_{\tilde{\alpha}} \phi^E_{ij}(\mathbf{x}_{\tilde{\alpha}}) \varepsilon_{i;\tilde{\alpha}} \, . \end{aligned} \tag{16}$$

Moreover, now we have an MSE loss as

$$\mathcal{L}_{\mathcal{A}} = \frac{1}{2} \sum_{i,\tilde{\alpha}} \left[y_{i;\tilde{\alpha}} - \sum_j W_{ij} \phi_j(\mathbf{x}_{\tilde{\alpha}}) - \frac{\sigma}{2} \sum_{j_1,j_2} W_{ij_1} W_{ij_2} \psi_{j_1 j_2}(\mathbf{x}_{\tilde{\alpha}}) \right]^2 \, . \tag{17}$$

So we could write the dynamics of z more explicitly as

$$\begin{aligned} đz_{i;\delta} = &-\eta \sum_{\tilde{\alpha}} \left[\sum_j \phi_{ij}^E(\mathbf{x}_\delta) \phi_{ij}^E(\mathbf{x}_{\tilde{\alpha}}) \right] \varepsilon_{i;\tilde{\alpha}} \\ &+ \frac{\eta^2}{2} \sum_{\tilde{\alpha}_1,\tilde{\alpha}_2} \left[\sum_{j_1,j_2} \sigma \psi_{j_1 j_2}\left(\mathbf{x}_\delta\right) \phi_{ij_1}^E(\mathbf{x}_{\tilde{\alpha}_1}) \phi_{ij_2}^E(\mathbf{x}_{\tilde{\alpha}_2}) \right] \varepsilon_{i;\tilde{\alpha}_1} \varepsilon_{i;\tilde{\alpha}_2} \ . \end{aligned} \tag{18}$$

We define the *effective kernel*:

$$k_{ii;\delta_1\delta_2}^E \equiv \sum_j \phi_{ij}^E\left(\mathbf{x}_{\delta_1}\right) \phi_{ij}^E\left(\mathbf{x}_{\delta_2}\right) \ , \tag{19}$$

and the *meta kernel*,

$$\begin{aligned} \mu_{\delta_0\delta_1\delta_2} &\equiv \sigma \sum_{j_1,j_2} \psi_{j_1 j_2}\left(\mathbf{x}_{\delta_0}\right) \phi_{j_1}\left(\mathbf{x}_{\delta_1}\right) \phi_{j_2}\left(\mathbf{x}_{\delta_2}\right) \\ &= \sigma \sum_{j_1,j_2} \varepsilon \psi_{j_1 j_2}\left(\mathbf{x}_{\delta_0}\right) \phi_{ij_1}^E\left(\mathbf{x}_{\delta_1}\right) \phi_{ij_2}^E\left(\mathbf{x}_{\delta_2}\right) + \mathcal{O}\left(\sigma^2\right) \ . \end{aligned} \tag{20}$$

So we have

$$đz_{i;\delta} = -\eta \sum_{\tilde{\alpha}} k_{ii;\delta\tilde{\alpha}}^E \varepsilon_{i;\tilde{\alpha}} + \frac{\eta^2}{2} \sum_{\tilde{\alpha}_1,\tilde{\alpha}_2} \mu_{\delta\tilde{\alpha}_1\tilde{\alpha}_2} \varepsilon_{i;\tilde{\alpha}_1} \varepsilon_{i;\tilde{\alpha}_2} + \mathcal{O}\left(\sigma^2\right) \ . \tag{21}$$

The meta kernel is fixed. The effective kernel will satisfy the following dynamics

$$đk_{ii;\delta_1\delta_2}^E = -\eta \sum_{\tilde{\alpha}} \left(\mu_{\delta_1\delta_2\tilde{\alpha}} + \mu_{\delta_2\delta_1\tilde{\alpha}}\right) \varepsilon_{i;\tilde{\alpha}} + O\left(\sigma^2\right) \ . \tag{22}$$

Now we could try solving the output z. Based on perturbation theory, we could divide the whole output by the free term z^F and the interacting term z^I,

$$z_{i;\delta}(t) \equiv z_{i;\delta}^F(t) + z_{i;\delta}^I(t) \ . \tag{23}$$

The free part follows the following exponential dynamics,

$$đz_{i;\delta}^F = -\eta \sum_{\tilde{\alpha}} k_{\delta\tilde{\alpha}} \varepsilon_{i;\tilde{\alpha}}^F \equiv z_{i;\delta}^F - \eta \sum_{\tilde{\alpha}} k_{\delta\tilde{\alpha}} \left[z_{i;\tilde{\alpha}}^F - y_{i;\tilde{\alpha}} \right] \ . \tag{24}$$

Here, k is the old definition of the kernel without quadratic terms, which is different from the effective kernel by $O(\sigma)$. Moreover, we have

$$\begin{aligned} &k_{ii;\delta_1\delta_2}^E(t) \\ =&k_{ii;\delta_1\delta_2}^E(0) - \sum_{\tilde{\alpha}} \left(\mu_{\delta_1\delta_2\tilde{\alpha}} + \mu_{\delta_2\delta_1\tilde{\alpha}}\right) a_{i;\tilde{\alpha}}(t) \ , \end{aligned} \tag{25}$$

where

$$a_{i;\tilde{\alpha}}(t) = \sum_{\tilde{\alpha}_2} \tilde{k}^{\tilde{\alpha}\tilde{\alpha}_2} \left(\varepsilon_{i;\tilde{\alpha}}(t) - \varepsilon_{i;\tilde{\alpha}}^F(t) \right) . \tag{26}$$

where ε^F is the free part of the residual training error $z^F - y$. Then, one could compute the interacting piece. We have

$$đz_{i;\delta}^I = -\sum_{j,\tilde{\alpha}} \eta k_{\delta\tilde{\alpha}} z_{j;\tilde{\alpha}}^I(t) + \eta \mathbb{F}_{i;\delta}(t) \ . \tag{27}$$

Here $\mathbb{F}$ is the damping force

$$\mathbb{F}_{i;\delta}(t) \equiv -\sum_{\tilde{\alpha}} \left[k^E_{\delta\tilde{\alpha}}(t) - k_{\delta\tilde{\alpha}}\right]\varepsilon^F_{i;\tilde{\alpha}}(t) + \frac{\eta}{2}\sum_{\tilde{\alpha}_1,\tilde{\alpha}_2}\mu_{\delta\tilde{\alpha}_1\tilde{\alpha}_2}\varepsilon^F_{i;\tilde{\alpha}_1}(t)\varepsilon^F_{i;\tilde{\alpha}_2}(t)\,, \tag{28}$$

and we have

$$z^I_{i;\delta}(t) = \eta\sum_{s=0}^{t-1}\left[\mathbb{F}_{i;\delta}(s) - \sum_{\tilde{\alpha}_1,\tilde{\alpha}_2} k_{\delta\tilde{\alpha}_1}\tilde{k}^{\tilde{\alpha}_1\tilde{\alpha}_2}\mathbb{F}_{i;\tilde{\alpha}_2}(s)\right] + \sum_{\tilde{\alpha}_1,\tilde{\alpha}_2} k_{\delta\tilde{\alpha}_1}\tilde{k}^{\tilde{\alpha}_1\tilde{\alpha}_2} z^I_{i;\tilde{\alpha}_2}(t)\,. \tag{29}$$

The whole system is a set of non-linear difference equations.

1. The residual training error

Now we could try to solve the residual training error. We note that, for $\tilde{\alpha} \in \mathcal{A}$, we have

$$\varepsilon_{i;\tilde{\alpha}}(t) - \varepsilon^F_{i;\tilde{\alpha}}(t) = z^I_{i;\tilde{\alpha}}(t)\,. \tag{30}$$

So we have

$$k^E_{ii;\delta_1\delta_2}(t) = k^E_{ii;\delta_1\delta_2}(0) - \sum_{\tilde{\alpha},\tilde{\alpha}_2}\left(\mu_{\delta_1\delta_2\tilde{\alpha}} + \mu_{\delta_2\delta_1\tilde{\alpha}}\right)\tilde{k}^{\tilde{\alpha}\tilde{\alpha}_2} z^I_{i;\tilde{\alpha}_2}(t)\,. \tag{31}$$

Moreover, we call

$$k^\Delta_{ii;\delta_1\delta_2} \equiv k^\Delta_{ii;\delta_1\delta_2} \equiv k^E_{ii;\delta_1\delta_2}(0) - k_{\delta_1\delta_2}\,. \tag{32}$$

We note that the effective kernel looks like,

$$\begin{aligned} k^E_{ii;\delta_1\delta_2} &\equiv \sum_j \phi^E_{ij}\left(\mathbf{x}_{\delta_1};0\right)\phi^E_{ij}\left(\mathbf{x}_{\delta_2};0\right) \\ &= \sum_j\left(\phi_j\left(\mathbf{x}_{\delta_1}\right) + \sigma\sum_k W_{ik}(0)\psi_{kj}\left(\mathbf{x}_{\delta_1}\right)\right)\left(\phi_j\left(\mathbf{x}_{\delta_2}\right) + \sigma\sum_{k'} W_{ik}(0)\psi_{kj}\left(\mathbf{x}_{\delta_2}\right)\right) \\ &= k_{\delta_1\delta_2} + \sigma\sum_{j,k} W_{ik}(0)\psi_{kj}\left(\mathbf{x}_{\delta_1}\right)\phi_j\left(\mathbf{x}_{\delta_2}\right) + \sigma\sum_{j,k} W_{ik}(0)\psi_{kj}\left(\mathbf{x}_{\delta_2}\right)\phi_j\left(\mathbf{x}_{\delta_1}\right) + \mathcal{O}(\sigma^2)\,. \end{aligned} \tag{33}$$

Keeping the leading order, we have

$$k^\Delta_{ii,\delta_1\delta_2} = \sigma\sum_{j,k} W_{ik}(0)\psi_{kj}\left(\mathbf{x}_{\delta_1}\right)\phi_j\left(\mathbf{x}_{\delta_2}\right) + \sigma\sum_{j,k} W_{ik}(0)\psi_{kj}\left(\mathbf{x}_{\delta_2}\right)\phi_j\left(\mathbf{x}_{\delta_1}\right)\,, \tag{34}$$

and

$$\begin{aligned} k^E_{ii;\delta_1\delta_2}(t) &= k^E_{ii;\delta_1\delta_2}(0) - \sum_{\tilde{\alpha},\tilde{\alpha}_2}\left(\mu_{\delta_1\delta_2\tilde{\alpha}} + \mu_{\delta_2\delta_1\tilde{\alpha}}\right)\tilde{k}^{\tilde{\alpha}\tilde{\alpha}_2} z^I_{i;\tilde{\alpha}_2}(t) \\ &= k_{\delta_1\delta_2} + k^\Delta_{ii,\delta_1\delta_2} - \sum_{\tilde{\alpha},\tilde{\alpha}_2}\left(\mu_{\delta_1\delta_2\tilde{\alpha}} + \mu_{\delta_2\delta_1\tilde{\alpha}}\right)\tilde{k}^{\tilde{\alpha}\tilde{\alpha}_2} z^I_{i;\tilde{\alpha}_2}(t)\,. \end{aligned} \tag{35}$$

Moreover, let us solve the damping force $\mathbb{F}$:

$$\begin{aligned}
\mathbb{F}_{i;\delta}(t) &= -\sum_{\tilde{\alpha}} \left[k^E_{\delta\tilde{\alpha}}(t) - k_{\delta\tilde{\alpha}}\right]\varepsilon^F_{i;\tilde{\alpha}}(t) + \frac{\eta}{2}\sum_{\tilde{\alpha}_1,\tilde{\alpha}_2}\mu_{\delta\tilde{\alpha}_1\tilde{\alpha}_2}\varepsilon^F_{i;\tilde{\alpha}_1}(t)\varepsilon^F_{i;\tilde{\alpha}_2}(t) \\
&= -\sum_{\tilde{\alpha},\tilde{\alpha}_3}\left[k^E_{ii;\delta\tilde{\alpha}}(0) - \sum_{\tilde{\alpha}_1,\tilde{\alpha}_2}\left(\mu_{\delta\tilde{\alpha}\tilde{\alpha}_1} + \mu_{\tilde{\alpha}\delta\tilde{\alpha}_1}\right)\tilde{k}^{\tilde{\alpha}_1\tilde{\alpha}_2}\varepsilon^I_{i;\tilde{\alpha}_2}(t) - k_{\delta\tilde{\alpha}}\right]U_{\tilde{\alpha}\tilde{\alpha}_3}(t)\varepsilon_{i;\tilde{\alpha}_3}(0) \\
&+ \frac{\eta}{2}\sum_{\tilde{\alpha}_1,\tilde{\alpha}_2,\tilde{\alpha}_3,\tilde{\alpha}_4}\mu_{\delta\tilde{\alpha}_1\tilde{\alpha}_2}U_{\tilde{\alpha}_1\tilde{\alpha}_3}(t)U_{\tilde{\alpha}_2\tilde{\alpha}_4}(t)\varepsilon_{i;\tilde{\alpha}_3}(0)\varepsilon_{i;\tilde{\alpha}_4}(0) \\
&= -\sum_{\tilde{\alpha},\tilde{\alpha}_3}k^\Delta_{ii;\delta\tilde{\alpha}}U_{\tilde{\alpha}\tilde{\alpha}_3}(t)\varepsilon_{i;\tilde{\alpha}_3}(0) + \sum_{\tilde{\alpha},\tilde{\alpha}_1,\tilde{\alpha}_2,\tilde{\alpha}_3}\left(\mu_{\delta\tilde{\alpha}\tilde{\alpha}_1} + \mu_{\tilde{\alpha}\delta\tilde{\alpha}_1}\right)\tilde{k}^{\tilde{\alpha}_1\tilde{\alpha}_2}U_{\tilde{\alpha}\tilde{\alpha}_3}(t)\varepsilon_{i;\tilde{\alpha}_3}(0)z^I_{i;\tilde{\alpha}_2}(t) \\
&+ \frac{\eta}{2}\sum_{\tilde{\alpha}_1,\tilde{\alpha}_2,\tilde{\alpha}_3,\tilde{\alpha}_4}\mu_{\delta\tilde{\alpha}_1\tilde{\alpha}_2}U_{\tilde{\alpha}_1\tilde{\alpha}_3}(t)U_{\tilde{\alpha}_2\tilde{\alpha}_4}(t)\varepsilon_{i;\tilde{\alpha}_3}(0)\varepsilon_{i;\tilde{\alpha}_4}(0)\ .
\end{aligned} \tag{36}$$

The second and the last term in the last equality contributes higher orders. Thus, if the initial k^Δ is not vanishing, we have

$$\mathbb{F}_{i;\delta}(t) = -\sum_{\tilde{\alpha},\tilde{\alpha}_3}k^\Delta_{ii;\delta\tilde{\alpha}}U_{\tilde{\alpha}\tilde{\alpha}_3}(t)\varepsilon_{i;\tilde{\alpha}_3}(0) + \mathcal{O}(\eta\sigma)\ . \tag{37}$$

The contribution in the first term will dominate at least in the early time when k^Δ is not vanishing. In the late time where U decays significantly, we would have some non-perturbative effects.

Thus, we can plug the expression back to solve $z^I(t)$. We have

$$\begin{aligned}
z^I_{i;\tilde{\alpha}}(t) &= \eta\sum_{s=0}^{t-1}\sum_{\tilde{\alpha}_1}U_{\tilde{\alpha}\tilde{\alpha}_1}(t-1-s)\mathbb{F}_{i;\tilde{\alpha}_1}(s) \\
&= -\eta\sum_{s=0}^{t-1}\sum_{\tilde{\alpha}_1,\tilde{\alpha}_2,\tilde{\alpha}_3}U_{\tilde{\alpha}\tilde{\alpha}_1}(t-1-s)U_{\tilde{\alpha}_2\tilde{\alpha}_3}(s)k^\Delta_{ii;\tilde{\alpha}_1\tilde{\alpha}_2}\varepsilon_{i;\tilde{\alpha}_3}(0) \\
&= -\eta\sum_{s=0}^{t-1}(1-\eta k)^{t-1-s}k^\Delta_{ii}(\delta-\eta k)^s\varepsilon_i(0)\ ,
\end{aligned} \tag{38}$$

The last formula is given in the following matrix form:

$$\begin{aligned}
&(1-\eta k)_{\tilde{\alpha}_1\tilde{\alpha}_2} = \delta_{\tilde{\alpha}_1\tilde{\alpha}_2} - \eta k_{\tilde{\alpha}_1\tilde{\alpha}_2} = \delta_{\tilde{\alpha}_1\tilde{\alpha}_2} - \eta\sum_j\phi_j\left(\mathbf{x}_{\tilde{\alpha}_1}\right)\phi_j\left(\mathbf{x}_{\tilde{\alpha}_2}\right)\ , \\
&(k^\Delta_{ii})_{\tilde{\alpha}_1\tilde{\alpha}_2} = \sigma\sum_{j,k}W_{ik}(0)\psi_{kj}\left(\mathbf{x}_{\tilde{\alpha}_1}\right)\phi_j\left(\mathbf{x}_{\tilde{\alpha}_2}\right) + \sigma\sum_{j,k}W_{ik}(0)\psi_{kj}\left(\mathbf{x}_{\tilde{\alpha}_2}\right)\phi_j\left(\mathbf{x}_{\tilde{\alpha}_1}\right) \equiv \sigma(\mathbf{M}^i)_{\tilde{\alpha}_1,\tilde{\alpha}_2}\ , \\
&(\varepsilon_i(0))_{\tilde{\alpha}} = \varepsilon_{i;\tilde{\alpha}}(0)\ .
\end{aligned} \tag{39}$$

Moreover, we could indeed get more information by just making the bound. We have

$$\begin{aligned}
\left\|z^I_{i;\tilde{\alpha}}(t)\right\| &= \left\|\eta\sum_{s=0}^{t-1}(1-\eta k)^{t-1-s}k^\Delta_{ii}(1-\eta k)^s\varepsilon_i(0)\right\| \\
&\le \eta\sum_{s=0}^{t-1}\|1-\eta k\|^{t-1-s}\left\|k^\Delta_{ii}\right\|\|1-\eta k\|^s\|\varepsilon_i(0)\| \\
&= \eta\left(\sum_{s=0}^{t-1}\|1-\eta k\|^{t-1}\right)\left\|k^\Delta_{ii}\right\|\|\varepsilon_i(0)\| \\
&= \eta t\|1-\eta k\|^{t-1}\left\|k^\Delta_{ii}\right\|\|\varepsilon_i(0)\| = \sigma\eta t\|1-\eta k\|^{t-1}\left\|\mathbf{M}^i\right\|\|\varepsilon_i(0)\|\ .
\end{aligned} \tag{40}$$

Now we compare the convergence time noticing that

$$\varepsilon^F_{i;\tilde{\alpha}_1}(t) = \sum_{\tilde{\alpha}_2}U_{\tilde{\alpha}_1\tilde{\alpha}_2}(t)\varepsilon_{i;\tilde{\alpha}_2}(0)\ . \tag{41}$$

So

$$\left\|\varepsilon_i^F(t)\right\| \leq \|1-\eta k\|^t \left\|\varepsilon_i(0)\right\| \ . \tag{42}$$

Schematically, we have

$$\frac{\left\|z_i^I(t)\right\|}{\left\|\varepsilon_i^F(t)\right\|} \sim \frac{\sigma\eta t \left\|\mathbf{M}^i\right\| \left\|\varepsilon_i(0)\right\|}{\|1-\eta k\| \left\|\varepsilon_i(0)\right\|} \sim \frac{\sigma\eta t \left\|\mathbf{M}^i\right\|}{\|1-\eta k\|} \ . \tag{43}$$

The relative perturbative error contribution will grow linearly in time.

2. *The asymptotic regime*

Now instead of only looking at the training set, we study the asymptotic regime for general inputs. We start from

$$z_{i;\delta}^I(t) = \eta \sum_{s=0}^{t-1} \left[\mathbb{F}_{i;\delta}(s) - \sum_{\tilde{\alpha}_1,\tilde{\alpha}_2} k_{\delta\tilde{\alpha}_1} \tilde{k}^{\tilde{\alpha}_1\tilde{\alpha}_2} \mathbb{F}_{i;\tilde{\alpha}_2}(s) \right] + \sum_{\tilde{\alpha}_1,\tilde{\alpha}_2} k_{\delta\tilde{\alpha}_1} \tilde{k}^{\tilde{\alpha}_1\tilde{\alpha}_2} z_{i;\tilde{\alpha}_2}^I(t) \ . \tag{44}$$

At the asymptotic convergence, the interacting perturbative correction on the training set will converge to zero, so we have

$$z_{i;\delta}^I(\infty) = \left[\eta \sum_{s=0}^{\infty} \mathbb{F}_{i;\delta}(s) \right] - \sum_{\tilde{\alpha}_1,\tilde{\alpha}_2} k_{\delta\tilde{\alpha}_1} \tilde{k}^{\tilde{\alpha}_1\tilde{\alpha}_2} \left[\eta \sum_{s=0}^{\infty} \mathbb{F}_{i;\tilde{\alpha}_2}(s) \right] \ . \tag{45}$$

So we need to perform the sum

$$\begin{aligned} \eta \sum_{s=0}^{\infty} \mathbb{F}_{i;\delta}(s) &= \sum_{\tilde{\alpha}_0,\tilde{\alpha}} \left(\mu_{\delta\tilde{\alpha}\tilde{\alpha}_0} + \mu_{\tilde{\alpha}\delta\tilde{\alpha}_0}\right) \left\{ \eta \sum_{s=0}^{\infty} a_{i;\tilde{\alpha}_0}(s) \varepsilon_{i;\tilde{\alpha}}^F(s) \right\} \\ &+ \frac{\eta}{2} \sum_{\tilde{\alpha}_1,\tilde{\alpha}_2} \mu_{\delta\tilde{\alpha}_1\tilde{\alpha}_2} \left\{ \eta \sum_{s=0}^{\infty} \varepsilon_{i;\tilde{\alpha}_1}^F(s) \varepsilon_{i;\tilde{\alpha}_2}^F(s) \right\} \ . \end{aligned} \tag{46}$$

We note that

$$\begin{aligned} &\eta \sum_{s=0}^{\infty} \varepsilon_{i;\tilde{\alpha}_1}^F(t) \varepsilon_{i;\tilde{\alpha}_2}^F(t) = \eta \sum_{\tilde{\alpha}_3,\tilde{\alpha}_4} \sum_{s=0}^{\infty} [(1-\eta k)^s]_{\tilde{\alpha}_1\tilde{\alpha}_3} [(1-\eta k)^s]_{\tilde{\alpha}_2\tilde{\alpha}_4} \varepsilon_{i;\tilde{\alpha}_3}(0) \varepsilon_{i;\tilde{\alpha}_4}(0) \\ &= \sum_{\tilde{\alpha}_3,\tilde{\alpha}_4} X_{\parallel}^{\tilde{\alpha}_1\tilde{\alpha}_2\tilde{\alpha}_3\tilde{\alpha}_4} \varepsilon_{i;\tilde{\alpha}_3}(0) \varepsilon_{i;\tilde{\alpha}_4}(0) \ . \end{aligned} \tag{47}$$

Here, we define the following inverting tensor:

$$X_{\parallel}^{\tilde{\alpha}_1\tilde{\alpha}_2\tilde{\alpha}_3\tilde{\alpha}_4} = \sum_{s=0}^{\infty} [(1-\eta k)^s]_{\tilde{\alpha}_1\tilde{\alpha}_3} [(1-\eta k)^s]_{\tilde{\alpha}_2\tilde{\alpha}_4} \ , \tag{48}$$

which is implicitly defined as

$$\begin{aligned} \delta_{\tilde{\alpha}_5}^{\tilde{\alpha}_1} \delta_{\tilde{\alpha}_6}^{\tilde{\alpha}_2} &= \sum_{\tilde{\alpha}_3,\tilde{\alpha}_4} X_{\parallel}^{\tilde{\alpha}_1\tilde{\alpha}_2\tilde{\alpha}_3\tilde{\alpha}_4} \frac{1}{\eta} \left[\delta_{\tilde{\alpha}_3\tilde{\alpha}_5} \delta_{\tilde{\alpha}_4\tilde{\alpha}_6} - \left(\delta_{\tilde{\alpha}_3\tilde{\alpha}_5} - \eta \tilde{k}_{\tilde{\alpha}_3\tilde{\alpha}_5}\right) \left(\delta_{\tilde{\alpha}_4\tilde{\alpha}_6} - \eta \tilde{k}_{\tilde{\alpha}_4\tilde{\alpha}_6}\right) \right] \\ &= \sum_{\tilde{\alpha}_3,\tilde{\alpha}_4} X_{\parallel}^{\tilde{\alpha}_1\tilde{\alpha}_2\tilde{\alpha}_3\tilde{\alpha}_4} \left(\tilde{k}_{\tilde{\alpha}_3\tilde{\alpha}_5} \delta_{\tilde{\alpha}_4\tilde{\alpha}_6} + \delta_{\tilde{\alpha}_3\tilde{\alpha}_5} \tilde{k}_{\tilde{\alpha}_4\tilde{\alpha}_6} - \eta \tilde{k}_{\tilde{\alpha}_3\tilde{\alpha}_5} \tilde{k}_{\tilde{\alpha}_4\tilde{\alpha}_6} \right) \ . \end{aligned} \tag{49}$$

Using the inverting tensor, the final expression is given by

$$\begin{aligned} z_{i;\delta}(\infty) &= z_{i;\delta}(0) - \sum_{\tilde{\alpha}_1,\tilde{\alpha}_2} k_{\delta\tilde{\alpha}_1} \tilde{k}^{\tilde{\alpha}_1\tilde{\alpha}_2} \varepsilon_{i;\tilde{\alpha}_2}(0) \\ &+ \sum_{\tilde{\alpha}_1,\ldots,\tilde{\alpha}_4} \left[\mu_{\tilde{\alpha}_1\delta\tilde{\alpha}_2} - \sum_{\tilde{\alpha}_5,\tilde{\alpha}_6 \in \mathcal{A}} k_{\delta\tilde{\alpha}_5} \tilde{k}^{\tilde{\alpha}_5\tilde{\alpha}_6} \mu_{\tilde{\alpha}_1\tilde{\alpha}_6\tilde{\alpha}_2} \right] Z_A^{\tilde{\alpha}_1\tilde{\alpha}_2\tilde{\alpha}_3\tilde{\alpha}_4} \varepsilon_{i;\tilde{\alpha}_3}(0) \varepsilon_{i;\tilde{\alpha}_3}(0) \\ &+ \sum_{\tilde{\alpha}_1,\ldots,\tilde{\alpha}_4} \left[\mu_{\delta\tilde{\alpha}_1\tilde{\alpha}_2} - \sum_{\tilde{\alpha}_5,\tilde{\alpha}_6 \in \mathcal{A}} k_{\delta\tilde{\alpha}_5} \tilde{k}^{\tilde{\alpha}_5\tilde{\alpha}_6} \mu_{\tilde{\alpha}_6\tilde{\alpha}_1\tilde{\alpha}_2} \right] Z_B^{\tilde{\alpha}_1\tilde{\alpha}_2\tilde{\alpha}_3\tilde{\alpha}_4} \varepsilon_{i;\tilde{\alpha}_3}(0) \varepsilon_{i;\tilde{\alpha}_3}(0) \ , \end{aligned} \tag{50}$$

where Zs are the *algorithm projectors*,

$$Z_A^{\tilde{\alpha}_1\tilde{\alpha}_2\tilde{\alpha}_3\tilde{\alpha}_4} \equiv \tilde{k}^{\tilde{\alpha}_1\tilde{\alpha}_3}\tilde{k}^{\tilde{\alpha}_2\tilde{\alpha}_4} - \sum_{\tilde{\alpha}_5} \tilde{k}^{\tilde{\alpha}_2\tilde{\alpha}_5} X_{\parallel}^{\tilde{\alpha}_1\tilde{\alpha}_5\tilde{\alpha}_3\tilde{\alpha}_4} ,$$
$$Z_B^{\tilde{\alpha}_1\tilde{\alpha}_2\tilde{\alpha}_3\tilde{\alpha}_4} \equiv \tilde{k}^{\tilde{\alpha}_1\tilde{\alpha}_3}\tilde{k}^{\tilde{\alpha}_2\tilde{\alpha}_4} - \sum_{\tilde{\alpha}_5} \tilde{k}^{\tilde{\alpha}_2\tilde{\alpha}_5} X_{\parallel}^{\tilde{\alpha}_1\tilde{\alpha}_5\tilde{\alpha}_3\tilde{\alpha}_4} + \frac{\eta}{2} X_{\parallel}^{\tilde{\alpha}_1\tilde{\alpha}_2\tilde{\alpha}_3\tilde{\alpha}_4} . \tag{51}$$

Finally, in the continuum limit, we could drop out the η terms in the algorithm projectors,

$$Z_A^{\tilde{\alpha}_1\tilde{\alpha}_2\tilde{\alpha}_3\tilde{\alpha}_4} = Z_B^{\tilde{\alpha}_1\tilde{\alpha}_2\tilde{\alpha}_3\tilde{\alpha}_4} \equiv \tilde{k}^{\tilde{\alpha}_1\tilde{\alpha}_3}\tilde{k}^{\tilde{\alpha}_2\tilde{\alpha}_4} - \sum_{\tilde{\alpha}_5} \tilde{k}^{\tilde{\alpha}_2\tilde{\alpha}_5} X_{\parallel}^{\tilde{\alpha}_1\tilde{\alpha}_5\tilde{\alpha}_3\tilde{\alpha}_4} ,$$
$$\sum_{\tilde{\alpha}_3,\tilde{\alpha}_4} X_{\parallel}^{\tilde{\alpha}_1\tilde{\alpha}_2\tilde{\alpha}_3\tilde{\alpha}_4} \left(\tilde{k}_{\tilde{\alpha}_3\tilde{\alpha}_5}\delta_{\tilde{\alpha}_4\tilde{\alpha}_6} + \delta_{\tilde{\alpha}_3\tilde{\alpha}_5}\tilde{k}_{\tilde{\alpha}_4\tilde{\alpha}_6}\right) = \delta_{\tilde{\alpha}_5}^{\tilde{\alpha}_1}\delta_{\tilde{\alpha}_6}^{\tilde{\alpha}_2} . \tag{52}$$

The existence of algorithm projectors shows algorithm dependence in those perturbative corrections. This will typically happen when the model has multiple (local) minimal.

II. DETAILS ON THE QUANTUM OPTIMIZATION

A. General setup

The quantum optimization problem we discuss here has a simpler formulation compared to quantum machine learning. Since the loss function does not contain the training data, we do not need to consider the difference between the training set and the whole space. It could be understood as a limiting case of the quantum machine learning problem.

We use the loss function,

$$\mathcal{L}(\theta) = \frac{1}{2}\left(\left\langle \Psi_0 \left| U^\dagger O U \right| \Psi_0 \right\rangle - O_0\right)^2 \equiv \frac{1}{2}\varepsilon^2 . \tag{53}$$

The trainable ansatz is,

$$U = \prod_{\ell=1}^{L} W_\ell U_\ell = \prod_{\ell=1}^{L} W_\ell \exp(i\theta_\ell X_\ell) . \tag{54}$$

Here, note that since the operator O is Hermitian, the loss function is always real and non-negative. The gradient descent is

$$đ\theta_\ell = -\eta \frac{d\mathcal{L}(\theta)}{d\theta_\ell} = -\eta \left(\left\langle \Psi_0 \left| U^\dagger O U \right| \Psi_0 \right\rangle - O_0\right) \frac{d\left\langle \Psi_0 \left| U^\dagger O U \right| \Psi_0 \right\rangle}{d\theta_\ell} = -\eta\varepsilon \frac{d\varepsilon}{d\theta_\ell} , \tag{55}$$

where

$$\varepsilon = \left\langle \Psi_0 \left| U^\dagger O U \right| \Psi_0 \right\rangle - O_0 . \tag{56}$$

Thus

$$đ\varepsilon = \sum_\ell \frac{d\varepsilon}{d\theta_\ell} đ\theta_\ell = -\eta \sum_\ell \frac{d\varepsilon}{d\theta_\ell}\frac{d\varepsilon}{d\theta_\ell}\varepsilon . \tag{57}$$

The object

$$K = \sum_\ell \frac{d\varepsilon}{d\theta_\ell}\frac{d\varepsilon}{d\theta_\ell} , \tag{58}$$

is the optimization analog of the QNTK. More precisely, we have

$$\frac{d\varepsilon}{d\theta_\ell} = \frac{d\left\langle \Psi_0 \left| U^\dagger O U \right| \Psi_0 \right\rangle}{d\theta_\ell} = -i \left\langle \Psi_0 \left| U_{+,\ell}^\dagger \left[X_\ell, U_\ell^\dagger W_\ell^\dagger U_{-,\ell}^\dagger O U_{-,\ell} W_\ell U_\ell\right] U_{+,\ell} \right| \Psi_0 \right\rangle , \tag{59}$$

with the help of the definition,

$$U_{-,\ell} \equiv \prod_{\ell'=1}^{\ell-1} W_{\ell'} U_{\ell'} \ ,$$
$$U_{+,\ell} \equiv \prod_{\ell'=\ell+1}^{L} W_{\ell'} U_{\ell'} \ . \tag{60}$$

So,

$$đ\varepsilon = -\eta \sum_{\ell} \frac{d\varepsilon}{d\theta_\ell} \frac{d\varepsilon}{d\theta_\ell} \varepsilon = \eta\varepsilon \Big\langle \Psi_0 \Big| U_{+,\ell}^\dagger \left[X_\ell, U_\ell^\dagger W_\ell^\dagger U_{-,\ell}^\dagger O U_{-,\ell} W_\ell U_\ell \right] U_{+,\ell} \Big| \Psi_0 \Big\rangle^2 \ . \tag{61}$$

Can we solve the difference equation? We have

$$đ\theta_\ell = i\eta \left\langle \Psi_0 \left| U_{+,\ell}^\dagger \left[X_\ell, U_\ell^\dagger W_\ell^\dagger U_{-,\ell}^\dagger O U_{-,\ell} W_\ell U_\ell \right] U_{+,\ell} \right| \Psi_0 \right\rangle \ . \tag{62}$$

In the continuum limit, we have

$$\frac{d\theta_\ell}{dt} = i\eta_c \left\langle \Psi_0 \left| U_{+,\ell}^\dagger \left[X_\ell, U_\ell^\dagger W_\ell^\dagger U_{-,\ell}^\dagger O U_{-,\ell} W_\ell U_\ell \right] U_{+,\ell} \right| \Psi_0 \right\rangle \ , \tag{63}$$

where $\eta_c = \eta/dt$ is the continuous version of the learning rate. In principle, this is a coupled nonlinear ODE system, and one could use the ODE theory to study them. Moreover, we have

$$U_{+,\ell}^\dagger \left[X_\ell, U_\ell^\dagger W_\ell^\dagger U_{-,\ell}^\dagger O U_{-,\ell} W_\ell U_\ell \right] U_{+,\ell} = U_{+,\ell}^\dagger U_\ell^\dagger \left[X_\ell, W_\ell^\dagger U_{-,\ell}^\dagger O U_{-,\ell} W_\ell \right] U_\ell U_{+,\ell} \ . \tag{64}$$

Defining

$$A_\ell = \left[X_\ell, W_\ell^\dagger U_{-,\ell}^\dagger O U_{-,\ell} W_\ell \right] \ , \tag{65}$$

and if we are expanding θ_ℓ when it is slightly deviated from 0 (more generally, around a fixed angle, which is equivalent to a redefinition of constant gates), we have,

$$U_{+,\ell}^\dagger U_\ell^\dagger A_\ell U_\ell U_{+,\ell} = U_{+,\ell}^\dagger A_\ell U_{+,\ell} - i\theta_\ell U_{+,\ell}^\dagger \left[X_\ell, A_\ell \right] U_{+,\ell} + \mathcal{O}(\theta_\ell^2) \ . \tag{66}$$

Thus, we note that the second-order expansions of θ_ℓ here means second-order commutators among X_ℓ and some unitary-addressed versions of the operator O, which is the operator we wish to optimize. Our experiences could be easily generalized to high orders.

B. Frozen QNTK

Now, we rescale the original variational angles by a factor δ where the variational angles are around $\theta^* + \delta\varphi$. The constant term θ^* will produce a constant gate $\exp(i\theta_\ell^* X_\ell)$, such that $W_\ell U_\ell \to W_\ell \exp(i\theta_\ell^* X_\ell) \exp(i\delta\varphi_\ell X_\ell)$. So we could absorb the θ^* dependence to the definition of the constant gate by defining $W_\ell(\theta_\ell^*) \equiv W_\ell \exp(i\theta_\ell^* X_\ell)$, and we have $W_\ell U_\ell \to W_\ell(\theta_\ell^*) \exp(i\delta\varphi_\ell X_\ell)$. Thus, for simplicity, we could understand the trainable gate as $U_\ell \to \exp(i\delta\varphi_\ell X_\ell)$ with a redefined constant gate. In our calculation later, we will drop the θ^* notation and understand the variational angles as small parameters rescaled by δ for our notation convenience.

In the frozen QNTK limit, we have,

$$K = -\delta^2 \sum_{\ell} \left\langle \Psi_0 \left| W_{+,\ell}^\dagger \left[X_\ell, W_\ell^\dagger W_{-,\ell}^\dagger O W_{-,\ell} W_\ell \right] W_{+,\ell} \right| \Psi_0 \right\rangle^2 \ . \tag{67}$$

And we define

$$W_{-,\ell} \equiv \prod_{\ell'=1}^{\ell-1} W_{\ell'} \ , \quad W_{+,\ell} \equiv \prod_{\ell'=\ell+1}^{L} W_{\ell'} \ . \tag{68}$$

Note that the frozen QNTK does not depend on the variational parameters. The gradient descent dynamics, in this case, is very easy to solve. We have

$$\varepsilon(t) = (1-\eta K)^t \varepsilon(0) \equiv \left(1 + \eta\delta^2 \sum_\ell \left\langle \Psi_0 \left| W_{+,\ell}^\dagger \left[X_\ell, W_\ell^\dagger W_{-,\ell}^\dagger O W_{-,\ell} W_\ell\right] W_{+,\ell} \right| \Psi_0 \right\rangle^2\right)^t \varepsilon(0) \ . \tag{69}$$

The convergence rate is given by,

$$\begin{aligned}
&\tau_c = -\log(1-\eta K) \approx \eta K \\
&= \eta\delta^2 \sum_\ell \left\langle \Psi_0 \left| W_{+,\ell}^\dagger \left[X_\ell, W_\ell^\dagger W_{-,\ell}^\dagger O W_{-,\ell} W_\ell\right] W_{+,\ell} \right| \Psi_0 \right\rangle^2 \\
&\leq 2\eta\delta^2 L \|O\|^2 \max_\ell \|X_\ell\|^2 \ .
\end{aligned} \tag{70}$$

C. dQNTK

Now let us focus on the second order to develop an analog of representation learning theory in the quantum optimization example. We have

$$\begin{aligned}
&\varepsilon = \left\langle \Psi_0 \left| \left(\prod\nolimits_{\ell'=L}^{1} W_{\ell'}^\dagger\right) O \left(\prod\nolimits_{\ell=1}^{L} W_\ell\right) \right| \Psi_0 \right\rangle - O_0 \\
&- i\delta \sum_\ell \varphi_\ell \left\langle \Psi_0 \left| W_{+,\ell}^\dagger \left[X_\ell, W_\ell^\dagger W_{-,\ell}^\dagger O W_{-,\ell} W_\ell\right] W_{+,\ell} \right| \Psi_0 \right\rangle \\
&- \frac{\delta^2}{2} \sum_{\ell_1,\ell_2} \varphi_{\ell_1}\varphi_{\ell_2} \times \begin{cases} \left\langle \Psi_0 \left| W_{+,\ell_1}^\dagger \left[X_{\ell_1}, Q_{\ell_1,\ell_2}^\dagger \left[X_{\ell_2}, W_{\ell_2}^\dagger W_{-,\ell_2}^\dagger O W_{-,\ell_2} W_{\ell_2}\right] Q_{\ell_2,\ell_1}\right] W_{+,\ell_1} \right| \Psi_0 \right\rangle : \ell_1 \geq \ell_2 \\ \left\langle \Psi_0 \left| W_{+,\ell_2}^\dagger \left[X_{\ell_2}, Q_{\ell_2,\ell_1}^\dagger \left[X_{\ell_1}, W_{\ell_1}^\dagger W_{-,\ell_1}^\dagger O W_{-,\ell_1} W_{\ell_1}\right] Q_{\ell_1,\ell_2}\right] W_{+,\ell_2} \right| \Psi_0 \right\rangle : \ell_1 < \ell_2 \end{cases} ,
\end{aligned} \tag{71}$$

where

$$\begin{aligned}
&W_{\ell_1,\ell_2} \equiv \prod_{\ell=\ell_1+1}^{\ell_2-1} W_\ell \ , \\
&Q_{\ell_1,\ell_2} = \begin{cases} W_{\ell_1,\ell_2} W_{\ell_2} : \ell_1 < \ell_2 \\ 1 : \ell_1 = \ell_2 \end{cases} \ .
\end{aligned} \tag{72}$$

Moreover, we have

$$đ\varphi_\ell = -\eta\varepsilon \frac{d\varepsilon}{d\varphi_\ell} \ . \tag{73}$$

Thus

$$đ\varepsilon = -\eta \sum_\ell \frac{d\varepsilon}{d\varphi_\ell}\frac{d\varepsilon}{d\varphi_\ell}\varepsilon + \frac{1}{2}\eta^2\varepsilon^2 \sum_{\ell_1,\ell_2} \frac{d^2\varepsilon}{d\varphi_{\ell_1} d\varphi_{\ell_2}} \frac{d\varepsilon}{d\varphi_{\ell_1}} \frac{d\varepsilon}{d\varphi_{\ell_2}} \ . \tag{74}$$

The structure of the gradient descent equation for the residual optimization error is very similar to the case in the quadratic model and representation learning context. We define the free part and the interacting part of ε as

$$\varepsilon = \varepsilon^F + \varepsilon^I \ . \tag{75}$$

The free part is given by,

$$\begin{aligned}
&\varepsilon^F = (1-\eta K)^t \varepsilon(0) \ , \\
&K = -\delta^2 \sum_\ell \left\langle \Psi_0 \left| W_{+,\ell}^\dagger \left[X_\ell, W_\ell^\dagger W_{-,\ell}^\dagger O W_{-,\ell} W_\ell\right] W_{+,\ell} \right| \Psi_0 \right\rangle^2 ,
\end{aligned} \tag{76}$$

and the interacting part is given by

$$\varepsilon^I(t) = -\eta t(1-\eta K)^{t-1} K^\Delta \varepsilon(0) \ , \tag{77}$$

where we define the effective kernel up to the dQNTK order,

$$K^E = \sum_\ell \frac{d\varepsilon}{d\varphi_\ell}\frac{d\varepsilon}{d\varphi_\ell} \ . \tag{78}$$

And we have

$$\begin{aligned} K^\Delta &\equiv K^E(0) - K = \left(\sum_\ell \frac{d\varepsilon}{d\theta_\ell}\frac{d\varepsilon}{d\theta_\ell}\right)(0) - \sum_\ell \frac{d\varepsilon^F}{d\theta_\ell}\frac{d\varepsilon^F}{d\theta_\ell} \\ &= 2i\delta^3 \sum_\ell \left\langle \Psi_0 \left| W_{+,\ell}^\dagger \left[X_\ell, W_\ell^\dagger W_{-,\ell}^\dagger O W_{-,\ell} W_\ell\right] W_{+,\ell} \right| \Psi_0 \right\rangle \\ &\sum_{\ell'} \begin{cases} \left\langle \Psi_0 \left| W_{+,\ell'}^\dagger \left[X_{\ell'}, Q_{\ell',\ell}^\dagger \left[X_\ell, W_\ell^\dagger W_{-,\ell}^\dagger O W_{-,\ell} W_\ell\right] W_{\ell,\ell'} W_{\ell'}\right] W_{+,\ell'} \right| \Psi_0 \right\rangle : \ell' \geq \ell \\ \left\langle \Psi_0 \left| W_{+,\ell}^\dagger \left[X_\ell, Q_{\ell,\ell'}^\dagger \left[X_{\ell'}, W_{\ell'}^\dagger W_{-,\ell'}^\dagger O W_{-,\ell'} W_{\ell'}\right] W_{\ell',\ell} W_\ell\right] W_{+,\ell} \right| \Psi_0 \right\rangle : \ell' < \ell \end{cases} \varphi_{\ell'}(0) \ . \end{aligned} \tag{79}$$

Similarly, one define the quantum meta-kernel (dQNTK) as

$$\begin{aligned} \mu = \delta^4 \sum_{\ell_1,\ell_2} &\left\langle \Psi_0 \left| W_{+,\ell_1}^\dagger \left[X_{\ell_1}, W_{\ell_1}^\dagger W_{-,\ell_1}^\dagger O W_{-,\ell_1} W_{\ell_1}\right] W_{+,\ell_1} \right| \Psi_0 \right\rangle \\ &\left\langle \Psi_0 \left| W_{+,\ell_2}^\dagger \left[X_{\ell_2}, W_{\ell_2}^\dagger W_{-,\ell_2}^\dagger O W_{-,\ell_2} W_{\ell_2}\right] W_{+,\ell_2} \right| \Psi_0 \right\rangle \times \\ \left(\begin{cases} \left\langle \Psi_0 \left| W_{+,\ell_1}^\dagger \left[X_{\ell_1}, Q_{\ell_1,\ell_2}^\dagger \left[X_{\ell_2}, W_{\ell_2}^\dagger W_{-,\ell_2}^\dagger O W_{-,\ell_2} W_{\ell_2}\right] Q_{\ell_2,\ell_1}\right] W_{+,\ell_1} \right| \Psi_0 \right\rangle : \ell_1 \geq \ell_2 \\ \left\langle \Psi_0 \left| W_{+,\ell_2}^\dagger \left[X_{\ell_2}, Q_{\ell_2,\ell_1}^\dagger \left[X_{\ell_1}, W_{\ell_1}^\dagger W_{-,\ell_1}^\dagger O W_{-,\ell_1} W_{\ell_1}\right] Q_{\ell_1,\ell_2}\right] W_{+,\ell_2} \right| \Psi_0 \right\rangle : \ell_1 < \ell_2 \end{cases} \right) \ . \end{aligned} \tag{80}$$

Finally, we wish to mention that in [1], for the classical neural networks they study, the leading order perturbative contribution $\mathcal{O}(1/\text{width})$ is both given by dNTK and ddNTK in dynamics. In our frozen QNTK limit (in the context of lazy training), this does not happen because of power counting in δ.

Finally, we wish to mention that the scaling of ε^I will lead to the so-called *catapult effect*. In general, in higher order corrections, we get schematically the correction $\sim t^p \exp(-\eta K t)$ for a more general polynomial t^p in the prefactor of the exponential decay. This type of correction forms a first-principle explanation of the catapult effect, where a similar related model is discussed. A full characterization of the catapult effect in classical and quantum cases is beyond the scope of this paper, and we refer the detailed research to the future [4].

III. DETAILS ON THE QUANTUM MACHINE LEARNING: HERMITIAN OPERATOR EXPECTATION VALUE EVALUATION

A. General setup

We define our model as

$$z_{i;\delta} \equiv z_i(\theta, \mathbf{x}_\delta) = \langle \phi(\mathbf{x}_\delta)| U^\dagger O_i U |\phi(\mathbf{x}_\delta)\rangle \ . \tag{81}$$

where O_i is the i-th Hermitian observable. We assume that O_i is taken from a subset of Hermitian operators of the Hilbert space $\mathcal{H}$, denoted by $\mathcal{O}(\mathcal{H})$. The dimension of $\mathcal{O}(\mathcal{H})$ is upper bounded by polynomials of the dimension of the Hilbert space, $\dim \mathcal{H}$. The trainable ansatz is,

$$U = \prod_{\ell=1}^L W_\ell U_\ell = \prod_{\ell=1}^L W_\ell \exp(i\theta_\ell X_\ell) \ . \tag{82}$$

One could take the derivative,

$$đ z_{i;\delta} = \sum_\ell \frac{dz_{i;\delta}}{d\theta_\ell} đ\theta_\ell \ , \tag{83}$$

The loss is

$$L_{\mathcal{A}}(\theta) = \frac{1}{2}\sum_{\tilde{\alpha},i}\left(y_{i;\tilde{\alpha}} - z_{i;\tilde{\alpha}}\right)^2 = \frac{1}{2}\sum_{\tilde{\alpha},i}\varepsilon^2_{i;\tilde{\alpha}}\,. \tag{84}$$

So

$$\frac{dL_{\mathcal{A}}(\theta)}{d\theta_\ell} = \sum_{\tilde{\alpha},i}\varepsilon_{i;\tilde{\alpha}}\frac{dz_{i;\tilde{\alpha}}}{d\theta_\ell}\,. \tag{85}$$

The gradient descent rule is

$$đ\theta_\ell = -\eta\frac{d\mathcal{L}_{\mathcal{A}}}{d\theta_\ell} = -\eta\sum_{\tilde{\alpha},i}\varepsilon_{i;\tilde{\alpha}}\frac{dz_{i;\tilde{\alpha}}}{d\theta_\ell}\,, \tag{86}$$

so we have

$$đz_{i;\delta} = -\eta\sum_{\ell,i',\tilde{\alpha}}\varepsilon_{i';\tilde{\alpha}}\frac{dz_{i;\delta}}{d\theta_\ell}\frac{dz_{i';\tilde{\alpha}}}{d\theta_\ell}\,. \tag{87}$$

Since we measure the expectation values of operators, our εs are always real. Defining the kernel,

$$K^{ii'}_{\delta,\tilde{\alpha}} = \sum_{\ell}\frac{dz_{i;\delta}}{d\theta_\ell}\frac{dz_{i';\tilde{\alpha}}}{d\theta_\ell}\,, \tag{88}$$

we have

$$đz_{i;\delta} = -\eta\sum_{\tilde{\alpha},i'}K^{ii'}_{\delta,\tilde{\alpha}}\varepsilon_{i';\tilde{\alpha}}\,. \tag{89}$$

We could also make the joint indices

$$(\delta,i) = \bar{a}\,,\qquad (\tilde{\alpha},i') = \hat{b}\,, \tag{90}$$

which are running in the space $\mathcal{D}\times\mathcal{O}(\mathcal{H})$ and $\mathcal{A}\times\mathcal{O}(\mathcal{H})$ respectively. The notation $\hat{a}$ is indicating that the data point component belongs to the training set $\mathcal{A}$, while the notation $\bar{a}$ means that the data point component is general in $\mathcal{D}$. And we have

$$đz_{\bar{a}} = -\eta\sum_{\hat{b}}K_{\bar{a}\hat{b}}\varepsilon_{\hat{b}}\,. \tag{91}$$

In general, one could prove the following statement.

Theorem 1. *The matrix $K_{\bar{a}\bar{b}}$ is non-negative and symmetric.*

Proof. It is symmetric by definition. Moreover, we consider an arbitrary vector $f_{\bar{a}}$. We have

$$\begin{aligned}\sum_{\bar{a},\bar{b}}f_{\bar{a}}K_{\bar{a}\bar{b}}f_{\bar{b}} &= \sum_{\delta,\delta',i,i'}f_{\delta,i}K^{ii'}_{\delta,\delta'}f_{\delta',i'} = \\ &= \sum_{\delta,\delta',i,i',\ell}f_{\delta,i}f_{\delta',i'}\frac{dz_{i';\delta'}}{d\theta_\ell}\frac{dz_{i;\delta}}{d\theta_\ell} = \sum_{\ell}\left(\sum_{\delta,i}f_{\delta,i}\frac{dz_{i;\delta}}{d\theta_\ell}\right)^2 \geq 0\,.\end{aligned} \tag{92}$$

□

Thus, the matrix K is a proper version of the positive semi-definite symmetric (PDS) kernel [5] in the sense of the classical learning theory.

Now, putting the variational ansatz inside the kernel, we get

$$\begin{aligned}\frac{dz_{i;\delta}}{d\theta_\ell} &= \left\langle\phi\left(\mathbf{x}_\delta\right)\middle|U^\dagger_{+,\ell}\left[X_\ell, U^\dagger_\ell W^\dagger_\ell U^\dagger_{-,\ell}O_i U_{-,\ell}W_\ell U_\ell\right]U_{+,\ell}\middle|\phi\left(\mathbf{x}_\delta\right)\right\rangle\,,\\ \frac{dz_{i';\tilde{\alpha}}}{d\theta_\ell} &= \left\langle\phi\left(\mathbf{x}_{\tilde{\alpha}}\right)\middle|U^\dagger_{+,\ell}\left[X_\ell, U^\dagger_\ell W^\dagger_\ell U^\dagger_{-,\ell}O_{i'} U_{-,\ell}W_\ell U_\ell\right]U_{+,\ell}\middle|\phi\left(\mathbf{x}_{\tilde{\alpha}}\right)\right\rangle\,,\end{aligned} \tag{93}$$

So we have

$$K^{ii'}_{\delta,\tilde{\alpha}} = \sum_{\ell}\frac{dz_{i;\delta}}{d\theta_\ell}\frac{dz_{i';\tilde{\alpha}}}{d\theta_\ell} = -\sum_{\ell}\left(\begin{array}{l}\left\langle\phi\left(\mathbf{x}_\delta\right)\middle|U^\dagger_{+,\ell}\left[X_\ell, U^\dagger_\ell W^\dagger_\ell U^\dagger_{-,\ell}O_i U_{-,\ell}W_\ell U_\ell\right]U_{+,\ell}\middle|\phi\left(\mathbf{x}_\delta\right)\right\rangle\times\\ \left\langle\phi\left(\mathbf{x}_{\tilde{\alpha}}\right)\middle|U^\dagger_{+,\ell}\left[X_\ell, U^\dagger_\ell W^\dagger_\ell U^\dagger_{-,\ell}O_{i'} U_{-,\ell}W_\ell U_\ell\right]U_{+,\ell}\middle|\phi\left(\mathbf{x}_{\tilde{\alpha}}\right)\right\rangle\end{array}\right)\,. \tag{94}$$

B. No representation learning

The statement of *no representation learning* corresponds to the limit where all the change of variational angles are sufficiently close to zero. In this case, the QNTK becomes frozen (static), similar to the optimization problem. With the variational angle redefined, and the frozen QNTK limit, we have,

$$K^{ii'}_{\delta,\tilde{\alpha}} = -\delta^2 \sum_{\ell} \begin{pmatrix} \left\langle \phi\left(\mathbf{x}_{\delta}\right) \middle| W^{\dagger}_{+,\ell} \left[X_{\ell}, W^{\dagger}_{\ell} W^{\dagger}_{-,\ell} O_i W_{-,\ell} W_{\ell}\right] W_{+,\ell} \middle| \phi\left(\mathbf{x}_{\delta}\right) \right\rangle \times \\ \left\langle \phi\left(\mathbf{x}_{\tilde{\alpha}}\right) \middle| W^{\dagger}_{+,\ell} \left[X_{\ell}, W^{\dagger}_{\ell} W^{\dagger}_{-,\ell} O_{i'} W_{-,\ell} W_{\ell}\right] W_{+,\ell} \middle| \phi\left(\mathbf{x}_{\tilde{\alpha}}\right) \right\rangle \end{pmatrix} . \tag{95}$$

Here, δ is the factor we are used to redefine the variational angles θ by $\theta^* + \delta \times \varphi$. Moreover, one could exactly solve the gradient descent dynamics. We start from the equation

$$đ\varepsilon_{\hat{a}} = -\eta \sum_{\hat{b}} K_{\hat{a}\hat{b}} \varepsilon_{\hat{b}} \, . \tag{96}$$

Thus we have

$$\varepsilon_{\hat{a}_1}(t) = \sum_{\hat{a}_2} U_{\hat{a}_1 \hat{a}_2}(t) \varepsilon_{\hat{a}_2}(0) \, , \tag{97}$$

where

$$U_{\hat{a}_1 \hat{a}_2}(t) = \left[(1 - \eta K)^t \right]_{\hat{a}_1 \hat{a}_2} . \tag{98}$$

One can compute the convergence time as

$$\tau_c = \| -\log\left(1 - \eta K\right) \| \approx \eta \left\| K^{ii'}_{\delta,\tilde{\alpha}} \right\| \, . \tag{99}$$

And we could compute the prediction similarly by

$$\begin{aligned} z_{\bar{a}}(\infty) &= z_{\bar{a}}(0) - \eta \sum_{\hat{b}} K_{\bar{a}\hat{b}} \sum_{t=0}^{+\infty} \varepsilon_{\hat{b}}(t) \\ &= z_{\bar{a}}(0) - \eta \sum_{\hat{b}} K_{\bar{a}\hat{b}} \left(\sum_{t=0,\hat{b}_0}^{+\infty} \left[(1 - \eta K)^t \right]_{\hat{b}\hat{b}_0} \right) \varepsilon_{\hat{b}_0}(0) \\ &= z_{\bar{a}}(0) - \eta \sum_{\hat{b}} K_{\bar{a}\hat{b}} \sum_{\hat{b}_0} \left[\left(1 - (1 - \eta K)\right)^{-1} \right]_{\hat{b}\hat{b}_0} \varepsilon_{\hat{b}_0}(0) \\ &= z_{\bar{a}}(0) - \sum_{\hat{b}} K_{\bar{a}\hat{b}} \sum_{\hat{b}_0} \left[K^{i,-1} \right]_{\hat{b}\hat{b}_0} \varepsilon_{\hat{b}_0}(0) \, . \end{aligned} \tag{100}$$

Moreover, we could compute the prediction by firstly defining the kernel inverse. We define

$$\sum_{\hat{a} \in \mathcal{A} \times \mathcal{O}(\mathcal{H})} \tilde{K}^{\hat{a}_1 \hat{a}_2} \tilde{K}_{\hat{a}_2 \hat{a}_3} = \delta^{\hat{a}_1}_{\hat{a}_3} \, , \tag{101}$$

so we get

$$z_{\bar{a}}(\infty) = z_{\bar{a}}(0) - \sum_{\hat{a}_1, \hat{a}_2} \tilde{K}^{\hat{a}_1 \hat{a}_2} K_{\bar{a}\hat{a}_1} \varepsilon_{\hat{a}_2}(0) \, . \tag{102}$$

C. Representation learning

Now we start to develop our quantum representation learning theory at the dQNTK order. We make a quadratic expansion:

$$
\begin{aligned}
&\varepsilon_{i;\tilde{\alpha}} = \left\langle \phi(\mathbf{x}_{\tilde{\alpha}}) \right| U^\dagger O_i U \left| \phi(\mathbf{x}_{\tilde{\alpha}}) \right\rangle - y_{i;\tilde{\alpha}} \\
&\approx \left\langle \phi(\mathbf{x}_{\tilde{\alpha}}) \right| \left(\prod_{\ell'=L}^{1} W_{\ell'}^\dagger \right) O_i \left(\prod_{\ell=1}^{L} W_\ell \right) \left| \phi(\mathbf{x}_{\tilde{\alpha}}) \right\rangle - y_{i;\tilde{\alpha}} \\
&- i\delta \sum_\ell \varphi_\ell \left\langle \phi(\mathbf{x}_{\tilde{\alpha}}) \right| W_{+,\ell}^\dagger \left[X_\ell, W_\ell^\dagger W_{-,\ell}^\dagger O_i W_{-,\ell} W_\ell \right] W_{+,\ell} \left| \phi(\mathbf{x}_{\tilde{\alpha}}) \right\rangle - \frac{\delta^2}{2} \sum_{\ell_1,\ell_2} \varphi_{\ell_1} \varphi_{\ell_2} \times \\
&\begin{cases} \left\langle \psi(\mathbf{x}_{\tilde{\alpha}}) \right| W_{+,\ell_1}^\dagger \left[X_{\ell_1}, Q_{\ell_1,\ell_2}^\dagger \left[X_{\ell_2}, W_{\ell_2}^\dagger W_{-,\ell_2}^\dagger O_i W_{-,\ell_2} W_{\ell_2} \right] W_{\ell_2,\ell_1} W_{\ell_1} \right] W_{+,\ell_1} \left| \psi(\mathbf{x}_{\tilde{\alpha}}) \right\rangle : \ell_1 \geq \ell_2 \\ \left\langle \psi(\mathbf{x}_{\tilde{\alpha}}) \right| W_{+,\ell_2}^\dagger \left[X_{\ell_2}, Q_{\ell_2,\ell_1}^\dagger \left[X_{\ell_1}, W_{\ell_1}^\dagger W_{-,\ell_1}^\dagger O_i W_{-,\ell_1} W_{\ell_1} \right] W_{\ell_1,\ell_2} W_{\ell_2} \right] W_{+,\ell_2} \left| \psi(\mathbf{x}_{\tilde{\alpha}}) \right\rangle : \ell_1 < \ell_2 \end{cases} .
\end{aligned} \tag{103}
$$

At the dQNTK order, we have,

$$
đz_{i;\delta} = \sum_\ell \frac{dz_{i;\delta}}{d\varphi_\ell} đ\varphi_\ell - \frac{1}{2}\delta^2 \sum_{\ell_1,\ell_2} đ\varphi_{\ell_1} đ\varphi_{\ell_2} G^{\delta,i}_{\ell_1,\ell_2} , \tag{104}
$$

where we define

$$
\begin{aligned}
&G^{\delta,i}_{\ell_1,\ell_2} \equiv G_{\ell_1,\ell_2}(\phi(\mathbf{x}_\delta), O_i) \\
&= \left(\begin{cases} \left\langle \phi(\mathbf{x}_\delta) \right| W_{+,\ell_1}^\dagger \left[X_{\ell_1}, Q_{\ell_1,\ell_2}^\dagger \left[X_{\ell_2}, W_{\ell_2}^\dagger W_{-,\ell_2}^\dagger O_i W_{-,\ell_2} W_{\ell_2} \right] W_{\ell_2,\ell_1} W_{\ell_1} \right] W_{+,\ell_1} \left| \phi(\mathbf{x}_\delta) \right\rangle : \ell_1 \geq \ell_2 \\ \left\langle \phi(\mathbf{x}_\delta) \right| W_{+,\ell_2}^\dagger \left[X_{\ell_2}, Q_{\ell_2,\ell_1}^\dagger \left[X_{\ell_1}, W_{\ell_1}^\dagger W_{-,\ell_1}^\dagger O_i W_{-,\ell_1} W_{\ell_1} \right] W_{\ell_1,\ell_2} W_{\ell_2} \right] W_{+,\ell_2} \left| \phi(\mathbf{x}_\delta) \right\rangle : \ell_1 < \ell_2 \end{cases} \right) , \\
&\Theta^{\delta,i}_\ell \equiv \Theta_\ell(\phi(\mathbf{x}_\delta), O_i) = \left\langle \phi(\mathbf{x}_\delta) \right| W_{+,\ell}^\dagger \left[X_\ell, W_\ell^\dagger W_{-,\ell}^\dagger O_i W_{-,\ell} W_\ell \right] W_{+,\ell} \left| \phi(\mathbf{x}_\delta) \right\rangle .
\end{aligned} \tag{105}
$$

Moreover, we want to define

$$
\Theta^{\delta,i}_\ell \equiv \Theta_\ell(\phi(\mathbf{x}_\delta), O_i) = \left\langle \phi(\mathbf{x}_\delta) \right| W_{+,\ell}^\dagger \left[X_\ell, W_\ell^\dagger W_{-,\ell}^\dagger O_i W_{-,\ell} W_\ell \right] W_{+,\ell} \left| \phi(\mathbf{x}_\delta) \right\rangle . \tag{106}
$$

We will now compute the effective kernel. We have,

$$
\begin{aligned}
\frac{dz_{i;\delta}}{d\varphi_\ell} &= -i\delta \Theta^{\delta,i}_\ell - \delta^2 \sum_{\ell'} \varphi_{\ell'} G^{\delta,i}_{\ell',\ell} , \\
\frac{dz_{i';\tilde{\alpha}}}{d\varphi_\ell} &= -i\delta \Theta^{\tilde{\alpha},i'}_\ell - \delta^2 \sum_{\ell'} \varphi_{\ell'} G^{\tilde{\alpha},i'}_{\ell',\ell} .
\end{aligned} \tag{107}
$$

Thus, we define

$$
K^{E,ii'}_{\delta,\tilde{\alpha}} = \sum_\ell \frac{dz_{i;\delta}}{d\varphi_\ell} \frac{dz_{i';\tilde{\alpha}}}{d\varphi_\ell} , \tag{108}
$$

and

$$
K^{E,ii'}_{\delta,\tilde{\alpha}}(0) = K^{ii'}_{\delta,\tilde{\alpha}} + K^{\Delta,ii'}_{\delta,\tilde{\alpha}} . \tag{109}
$$

We find

$$
K^{\Delta,ii'}_{\delta,\tilde{\alpha}} = i\delta^3 \sum_{\ell,\ell'} \varphi_{\ell'}(0) G^{\delta,i}_{\ell',\ell} \Theta^{\tilde{\alpha},i'}_\ell + i\delta^3 \sum_{\ell,\ell'} \varphi_{\ell'}(0) G^{\tilde{\alpha},i'}_{\ell',\ell} \Theta^{\delta,i}_\ell . \tag{110}
$$

Now we could write down the prediction on the training set. We have

$$
\varepsilon_{\hat{a}}(t) = \varepsilon^F_{\hat{a}}(t) + \varepsilon^I_{\hat{a}}(t) , \tag{111}
$$

where

$$\varepsilon_{\hat{a}}^{F}(t)=\sum_{\hat{a}_1}U_{\hat{a}\hat{a}_1}(t)\varepsilon_{\hat{a}_1}(0)\ ,$$
$$U_{\hat{a}_1\hat{a}_2}(t)=\left[(1-\eta K)^{t}\right]_{\hat{a}_1\hat{a}_2}\ , \tag{112}$$

and

$$\varepsilon_{\hat{a}}^{I}(t)=\left(-\eta\sum_{s=0}^{t-1}(1-\eta K)^{t-1-s}K^{\Delta}(1-\eta K)^{s}\varepsilon(0)\right)_{\hat{a}}\ . \tag{113}$$

It is the compact matrix product form in the space $\mathcal{A}\times\mathcal{O}(\mathcal{H})$. Moreover, we have

$$\left\|\varepsilon^{I}(t)\right\|\leq\eta t\|1-\eta K\|^{t-1}\left\|K^{\Delta}\right\|\|\varepsilon(0)\|\ . \tag{114}$$

Finally, we discuss the asymptotic convergence regime. We could compute the quantum meta-kernel. We notice that

$$đz_{i;\delta}=-\eta\sum_{\ell,i',\tilde{\alpha}}\frac{dz_{i;\delta}}{d\varphi_{\ell}}\frac{dz_{i';\tilde{\alpha}}}{d\varphi_{\ell}}\varepsilon_{i;\tilde{\alpha}}+\eta^{2}\sum_{\ell_1,\ell_2,\tilde{\alpha}_1,i_1,\tilde{\alpha}_2,i_2}\frac{d^{2}z_{i;\delta}}{d\varphi_{\ell_1}d\varphi_{\ell_2}}\frac{dz_{i_1;\tilde{\alpha}_1}}{d\varphi_{\ell_1}}\frac{dz_{i_2;\tilde{\alpha}_2}}{d\varphi_{\ell_2}}\varepsilon_{i_1;\tilde{\alpha}_1}\varepsilon_{i_2;\tilde{\alpha}_2}\ . \tag{115}$$

So we define

$$\mu_{\delta\tilde{\alpha}_1\tilde{\alpha}_2}^{ii_1i_2}=\mu_{\bar{a}\hat{a}_1\hat{a}_2}=\sum_{\ell_1,\ell_2}\frac{d^{2}z_{i;\delta}}{d\varphi_{\ell_1}d\varphi_{\ell_2}}\left(\frac{dz_{i_1;\tilde{\alpha}_1}}{d\varphi_{\ell_1}}\frac{dz_{i_2;\tilde{\alpha}_2}}{d\varphi_{\ell_2}}\right)\Big|_{\varphi=0}\ , \tag{116}$$

in the leading order. Note that it is natural to extend the definition to more general inputs

$$\mu_{\delta_0\delta_1\delta_2}^{i_0i_1i_2}=\mu_{\bar{a}_0\bar{a}_1\bar{a}_2}=\sum_{\ell_1,\ell_2}\frac{d^{2}z_{i_0;\delta_0}}{d\varphi_{\ell_1}d\varphi_{\ell_2}}\left(\frac{dz_{i_1;\delta_1}}{d\varphi_{\ell_1}}\frac{dz_{i_2;\delta_2}}{d\varphi_{\ell_2}}\right)\Big|_{\varphi=0}\ . \tag{117}$$

So the asymptotic convergence is given by

$$\begin{aligned}z_{\bar{a}}(\infty)&=z_{\bar{a}}(0)-\sum_{\hat{a}_1,\hat{a}_2}K_{\bar{a}\hat{a}_1}\tilde{K}^{\hat{a}_1\hat{a}_2}\varepsilon_{\hat{a}_2}(0)\\&+\sum_{\hat{a}_1,\hat{a}_2,\hat{a}_3,\hat{a}_4}\left[\mu_{\hat{a}_1\bar{a}\hat{a}_2}-\sum_{\hat{a}_5,\hat{a}_6}K_{\bar{a}\hat{a}_5}\tilde{K}^{\hat{a}_5\hat{a}_6}\mu_{\hat{a}_1\hat{a}_6\hat{a}_2}\right]Z_{A}^{\hat{a}_1\hat{a}_2\hat{a}_3\hat{a}_4}\varepsilon_{\hat{a}_3}(0)\varepsilon_{\hat{a}_4}(0)\\&+\sum_{\hat{a}_1,\hat{a}_2,\hat{a}_3,\hat{a}_4}\left[\mu_{\bar{a}\hat{a}_1\hat{a}_2}-\sum_{\hat{a}_5,\hat{a}_6}K_{\bar{a}\hat{a}_5}\tilde{K}^{\hat{a}_5\hat{a}_6}\mu_{\hat{a}_6\hat{a}_1\hat{a}_2}\right]Z_{B}^{\hat{a}_1\hat{a}_2\hat{a}_3\hat{a}_4}\varepsilon_{\hat{a}_3}(0)\varepsilon_{\hat{a}_4}(0)\ ,\end{aligned} \tag{118}$$

where the algorithm projector $Z_{A,B}$s are

$$\begin{aligned}Z_{A}^{\hat{a}_1\hat{a}_2\hat{a}_3\hat{a}_4}&\equiv\tilde{K}^{\hat{a}_1\hat{a}_3}\tilde{K}^{\hat{a}_2\hat{a}_4}-\sum_{\hat{a}_5}\tilde{K}^{\hat{a}_2\hat{a}_5}X_{\|}^{\hat{a}_1\hat{a}_5\hat{a}_3\hat{a}_4}\ ,\\Z_{B}^{\hat{a}_1\hat{a}_2\hat{a}_3\hat{a}_4}&\equiv\tilde{K}^{\hat{a}_1\hat{a}_3}\tilde{K}^{\hat{a}_2\hat{a}_4}-\sum_{\hat{a}_5}\tilde{K}^{\hat{a}_2\hat{a}_5}X_{\|}^{\hat{a}_1\hat{a}_5\hat{a}_3\hat{a}_4}+\frac{\eta}{2}X_{\|}^{\hat{a}_1\hat{a}_2\hat{a}_3\hat{a}_4}\ ,\end{aligned} \tag{119}$$

and

$$X_{\|}^{\hat{a}_1\hat{a}_2\hat{a}_3\hat{a}_4}=\sum_{s=0}^{\infty}[(1-\eta K)^{s}]_{\hat{a}_1\hat{a}_3}[(1-\eta K)^{s}]_{\hat{a}_2\hat{a}_4}\ , \tag{120}$$

which is defined implicitly as

$$\delta_{\hat{a}_5}^{\hat{a}_1}\delta_{\hat{a}_6}^{\hat{a}_2}=\sum_{\hat{a}_3,\hat{a}_4}X_{\|}^{\hat{a}_1\hat{a}_2\hat{a}_3\hat{a}_4}\left(\tilde{K}_{\hat{a}_3\hat{a}_5}\delta_{\hat{a}_4\hat{a}_6}+\delta_{\hat{a}_3\hat{a}_5}\tilde{K}_{\hat{a}_4\hat{a}_6}-\eta\tilde{K}_{\hat{a}_3\hat{a}_5}\tilde{K}_{\hat{a}_4\hat{a}_6}\right)\ . \tag{121}$$

Now we write down all components of μ. We have

$$\begin{aligned}
&\frac{dz_{i_1;\delta_1}}{d\varphi_{\ell_1}}(\varphi=0) = -i\delta\Theta^{\delta_1,i_1}_{\ell_1} \ , \\
&\frac{dz_{i_2;\delta_2}}{d\varphi_{\ell_2}}(\varphi=0) = -i\delta\Theta^{\delta_2,i_2}_{\ell_2} \ , \\
&\frac{d^2 z_{i_0;\delta_0}}{d\varphi_{\ell_1} d\varphi_{\ell_2}} = -\delta^2 G^{\delta_0,i_0}_{\ell_1,\ell_2} \ .
\end{aligned} \tag{122}$$

So

$$\mu^{i_0 i_1 i_2}_{\delta_0\delta_1\delta_2} = \sum_{\ell_1,\ell_2} \frac{d^2 z_{i_0;\delta_0}}{d\varphi_{\ell_1} d\varphi_{\ell_2}} \left(\frac{dz_{i_1;\delta_1}}{d\varphi_{\ell_1}} \frac{dz_{i_2;\delta_2}}{d\varphi_{\ell_2}} \right)\Bigg|_{\varphi=0} = \delta^4 \sum_{\ell_1,\ell_2} \Theta^{\delta_1,i_1}_{\ell_1} \Theta^{\delta_2,i_2}_{\ell_2} G^{\delta_0,i_0}_{\ell_1,\ell_2} \ . \tag{123}$$

IV. DETAILS ON THE QUANTUM MACHINE LEARNING: AMPLITUDE ENCODING

A. General setup

Here, we consider alternative quantum machine learning models with the amplitude encoding. In this case, we naturally extend the NTK formalism to the complex version. We consider the variational setup where

$$z_{i;\delta} \equiv z_i(\theta, \mathbf{x}_\delta) = \left\langle i \left| \prod_{\ell=1}^{L} W_\ell U_\ell \right| \phi(\mathbf{x}_\delta) \right\rangle \ . \tag{124}$$

One could take the derivative,

$$đz_{i;\delta} = \sum_\ell \frac{dz_{i;\delta}}{d\theta_\ell} đ\theta_\ell \ , \tag{125}$$

The loss is

$$L_{\mathcal{A}}(\theta) = \frac{1}{2}\sum_{\tilde{\alpha},i} |y_{i;\tilde{\alpha}} - z_{i;\tilde{\alpha}}|^2 = \frac{1}{2}\sum_{\tilde{\alpha},i} (y_{i;\tilde{\alpha}} - z_{i;\tilde{\alpha}})(y_{i;\tilde{\alpha}} - z_{i;\tilde{\alpha}})^* \ . \tag{126}$$

So

$$\frac{dL_{\mathcal{A}}(\theta)}{d\theta_\ell} = \frac{1}{2}\sum_{\tilde{\alpha},i} \varepsilon_{i;\tilde{\alpha}} \frac{dz^*_{i;\tilde{\alpha}}}{d\theta_\ell} + \frac{1}{2}\sum_{\tilde{\alpha},i} \varepsilon^*_{i;\tilde{\alpha}} \frac{dz_{i;\tilde{\alpha}}}{d\theta_\ell} = \mathrm{Re}\sum_{\tilde{\alpha},i} \varepsilon_{i;\tilde{\alpha}} \frac{dz^*_{i;\tilde{\alpha}}}{d\theta_\ell} \ . \tag{127}$$

Note that if we count the number of times that U appears in the loss function, the amplitude encoding model is a *squareroot* of the operator expectation value model. The gradient descent rule is

$$đ\theta_\ell = -\eta \frac{d\mathcal{L}_{\mathcal{A}}}{d\theta_\ell} = -\frac{\eta}{2}\sum_{\tilde{\alpha},i} \varepsilon_{i;\tilde{\alpha}} \frac{dz^*_{i;\tilde{\alpha}}}{d\theta_\ell} - \frac{\eta}{2}\sum_{\tilde{\alpha},i} \varepsilon^*_{i;\tilde{\alpha}} \frac{dz_{i;\tilde{\alpha}}}{d\theta_\ell} \ , \tag{128}$$

so we have

$$\begin{aligned}
đz_{i;\delta} &= -\eta \sum_{\ell,i',\tilde{\alpha}} \frac{dz_{i;\delta}}{d\theta_\ell} \mathrm{Re}\left(\varepsilon_{i;\tilde{\alpha}} \frac{dz^*_{i';\tilde{\alpha}}}{d\theta_\ell} \right) \\
&= -\frac{\eta}{2}\sum_{\ell,i',\tilde{\alpha}} \varepsilon_{i';\tilde{\alpha}} \frac{dz_{i;\delta}}{d\theta_\ell} \frac{dz^*_{i';\tilde{\alpha}}}{d\theta_\ell} - \frac{\eta}{2}\sum_{\ell,i',\tilde{\alpha}} \varepsilon^*_{i';\tilde{\alpha}} \frac{dz_{i;\delta}}{d\theta_\ell} \frac{dz_{i';\tilde{\alpha}}}{d\theta_\ell} \ .
\end{aligned} \tag{129}$$

We notice that the variable z is complex in general. So we could write,

$$\begin{aligned}
đz_{i;\delta} &= -\frac{\eta}{2}\sum_{\ell,i',\tilde{\alpha}} \varepsilon_{i';\tilde{\alpha}} \frac{dz_{i;\delta}}{d\theta_\ell} \frac{dz^*_{i';\tilde{\alpha}}}{d\theta_\ell} - \frac{\eta}{2}\sum_{\ell,i',\tilde{\alpha}} \varepsilon^*_{i';\tilde{\alpha}} \frac{dz_{i;\delta}}{d\theta_\ell} \frac{dz_{i';\tilde{\alpha}}}{d\theta_\ell} \ , \\
đz^*_{i;\delta} &= -\frac{\eta}{2}\sum_{\ell,i',\tilde{\alpha}} \varepsilon^*_{i';\tilde{\alpha}} \frac{dz^*_{i;\delta}}{d\theta_\ell} \frac{dz_{i';\tilde{\alpha}}}{d\theta_\ell} - \frac{\eta}{2}\sum_{\ell,i',\tilde{\alpha}} \varepsilon_{i';\tilde{\alpha}} \frac{dz^*_{i;\delta}}{d\theta_\ell} \frac{dz^*_{i';\tilde{\alpha}}}{d\theta_\ell} \ .
\end{aligned} \tag{130}$$

Defining the kernel

$$\begin{pmatrix} K^{+,ii'}_{\delta,\tilde{\alpha}} & K^{-,ii'}_{\delta,\tilde{\alpha}} \\ K^{*,-,ii'}_{\delta,\tilde{\alpha}} & K^{*,+,ii'}_{\delta,\tilde{\alpha}} \end{pmatrix} = \begin{pmatrix} \sum\limits_{\ell} \frac{dz_{i;\delta}}{d\theta_\ell} \frac{dz^*_{i';\tilde{\alpha}}}{d\theta_\ell} & \sum\limits_{\ell} \frac{dz_{i;\delta}}{d\theta_\ell} \frac{dz_{i';\tilde{\alpha}}}{d\theta_\ell} \\ \sum\limits_{\ell} \frac{dz^*_{i;\delta}}{d\theta_\ell} \frac{dz^*_{i';\tilde{\alpha}}}{d\theta_\ell} & \sum\limits_{\ell} \frac{dz^*_{i;\delta}}{d\theta_\ell} \frac{dz_{i';\tilde{\alpha}}}{d\theta_\ell} \end{pmatrix} , \tag{131}$$

we have

$$\begin{aligned} đz_{i;\delta} &= -\frac{\eta}{2} \sum_{\tilde{\alpha},i'} K^{+,ii'}_{\delta,\tilde{\alpha}} \varepsilon_{i';\tilde{\alpha}} - \frac{\eta}{2} \sum_{\tilde{\alpha},i'} K^{-,ii'}_{\delta,\tilde{\alpha}} \varepsilon^*_{i';\tilde{\alpha}} , \\ đz^*_{i;\delta} &= -\frac{\eta}{2} \sum_{\tilde{\alpha},i'} K^{*,-,ii'}_{\delta,\tilde{\alpha}} \varepsilon_{i';\tilde{\alpha}} - \frac{\eta}{2} \sum_{\tilde{\alpha},i'} K^{*,+,ii'}_{\delta,\tilde{\alpha}} \varepsilon^*_{i';\tilde{\alpha}} , \end{aligned} \tag{132}$$

Here we also use K^* to denote the complex conjugate. We introduce the worldsheet index $\dot{\alpha}, \dot{\beta}$ to denote the two-component of the complex variable $(đz, đz^*)$. So we have [6]

$$đz^{\dot{\alpha}}_{i;\delta} = -\frac{\eta}{2} \sum_{\tilde{\alpha},\dot{\beta},i'} K^{\dot{\alpha}\dot{\beta},ii'}_{\delta,\tilde{\alpha}} \varepsilon^{\dot{\beta}}_{i';\tilde{\alpha}} , \tag{133}$$

where

$$\varepsilon^{\dot{\alpha}}_{i;\delta} = \begin{pmatrix} \varepsilon_{i;\delta} \\ \varepsilon^*_{i;\delta} \end{pmatrix} , \quad K^{\dot{\alpha}\dot{\beta},ii'}_{\delta,\tilde{\alpha}} = \begin{pmatrix} K^{+,ii'}_{\delta,\tilde{\alpha}} & K^{-,ii'}_{\delta,\tilde{\alpha}} \\ K^{*,-,ii'}_{\delta,\tilde{\alpha}} & K^{*,+,ii'}_{\delta,\tilde{\alpha}} \end{pmatrix} . \tag{134}$$

We could also make the joint index

$$(\delta, \dot{\alpha}, i) = \bar{\mu} , \quad (\tilde{\alpha}, \dot{\beta}, i') = \hat{\nu} , \tag{135}$$

which is running in the space $\mathcal{D} \times \mathbb{Z}_2 \times \mathcal{H}$ and $\mathcal{A} \times \mathbb{Z}_2 \times \mathcal{H}$ respectively. The notation $\hat{\mu}$ is indicating that the data point component belongs to the training set $\mathcal{A}$, while the notation $\bar{\mu}$ means that the data point component is general in $\mathcal{D}$. And we have

$$đz_{\bar{\mu}} = -\frac{\eta}{2} \sum_{\hat{\nu}} K_{\bar{\mu}\hat{\nu}} \varepsilon_{\hat{\nu}} . \tag{136}$$

In general, one could prove the following statement.

Theorem 2. *The matrix $K_{\bar{\mu}\bar{\nu}}$ is non-negative and Hermitian.*

Proof. One could easily check that the matrix form

$$\begin{aligned} K^\dagger_{\bar{\mu}\bar{\nu}} &= \begin{pmatrix} K^{+,ii'}_{\delta,\tilde{\alpha}} & K^{-,ii'}_{\delta,\tilde{\alpha}} \\ K^{*,-,ii'}_{\delta,\tilde{\alpha}} & K^{*,+,ii'}_{\delta,\tilde{\alpha}} \end{pmatrix}^{*,T} = \begin{pmatrix} K^{*,+,ii'}_{\delta,\tilde{\alpha}} & K^{*-,ii'}_{\delta,\tilde{\alpha}} \\ K^{-,ii'}_{\delta,\tilde{\alpha}} & K^{+,ii'}_{\delta,\tilde{\alpha}} \end{pmatrix}^T \\ &= \begin{pmatrix} (K^{*,+,ii'}_{\delta,\tilde{\alpha}})^T & (K^{-,ii'}_{\delta,\tilde{\alpha}})^T \\ (K^{*,-,ii'}_{\delta,\tilde{\alpha}})^T & (K^{+,ii'}_{\delta,\tilde{\alpha}})^T \end{pmatrix} = \begin{pmatrix} K^{+,ii'}_{\delta,\tilde{\alpha}} & K^{-,i'i}_{\delta,\tilde{\alpha}} \\ K^{*,-,ii'}_{\delta,\tilde{\alpha}} & K^{*,+,ii'}_{\delta,\tilde{\alpha}} \end{pmatrix} = K_{\bar{\mu}\bar{\nu}} , \end{aligned} \tag{137}$$

is Hermitian. Moreover, consider an arbitrary vector $f_{\bar{\mu}}$, we have

$$\begin{aligned} &\sum_{\bar{\mu},\bar{\nu}} f^*_{\bar{\mu}} K_{\bar{\mu}\bar{\nu}} f_{\bar{\nu}} = \sum_{\delta,\delta',i,i'} \begin{pmatrix} f^*_{\delta,i} & f_{\delta,i} \end{pmatrix} \begin{pmatrix} K^{+,ii'}_{\delta,\delta'} & K^{-,ii'}_{\delta,\delta'} \\ K^{*,-,ii'}_{\delta,\delta'} & K^{*,+,ii'}_{\delta,\delta'} \end{pmatrix} \begin{pmatrix} f_{\delta',i'} \\ f^*_{\delta',i'} \end{pmatrix} \\ &= \sum_{\delta,\delta',i,i'} \left(f^*_{\delta,i} f_{\delta',i'} K^{+,ii'}_{\delta,\delta'} + f_{\delta,i} f_{\delta',i'} K^{*,-,ii'}_{\delta,\delta'} + f^*_{\delta,i} f^*_{\delta',i'} K^{-,ii'}_{\delta,\delta'} + f_{\delta,i} f^*_{\delta',i'} K^{*,+,ii'}_{\delta,\delta'} \right) \\ &= \sum_{\delta,\delta',i,i',\ell} \begin{pmatrix} f^*_{\delta,i} f_{\delta',i'} \frac{dz_{i;\delta}}{d\theta_\ell} \frac{dz^*_{i';\delta'}}{d\theta_\ell} + f_{\delta,i} f_{\delta',i'} \frac{dz^*_{i;\delta}}{d\theta_\ell} \frac{dz^*_{i';\delta'}}{d\theta_\ell} \\ + f^*_{\delta,i} f^*_{\delta',i'} \frac{dz_{i;\delta}}{d\theta_\ell} \frac{dz_{i';\delta'}}{d\theta_\ell} + f_{\delta,i} f^*_{\delta',i'} \frac{dz^*_{i;\delta}}{d\theta_\ell} \frac{dz_{i';\delta'}}{d\theta_\ell} \end{pmatrix} \\ &= \sum_{\delta,\delta',i,i',\ell} \left(f_{\delta,i} \frac{dz^*_{i;\delta}}{d\theta_\ell} + f^*_{\delta,i} \frac{dz_{i;\delta}}{d\theta_\ell} \right) \left(f_{\delta',i'} \frac{dz^*_{i';\delta'}}{d\theta_\ell} + f^*_{\delta',i'} \frac{dz_{i';\delta'}}{d\theta_\ell} \right) \\ &= \sum_{\ell} \left(\sum_{\delta,i} f_{\delta,i} \frac{dz^*_{i;\delta}}{d\theta_\ell} + f^*_{\delta,i} \frac{dz_{i;\delta}}{d\theta_\ell} \right)^2 = 4 \sum_{\ell} \left(\mathrm{Re} \sum_{\delta,i} f_{\delta,i} \frac{dz^*_{i;\delta}}{d\theta_\ell} \right)^2 \geq 0 . \end{aligned} \tag{138}$$

□

Thus, the matrix K is a complexified version of the positive semi-definite symmetric (PDS) kernel in the sense of the kernel method of statistical learning theory (see, for instance, [5]), but running during the training dynamics. Moreover, our work provides a complexified version of the NTK theory that could be useful for machine learning itself.

Now, putting the variational ansatz inside the kernel, we get

$$\frac{dz_{i;\alpha}}{d\theta_\ell} = \frac{d}{d\theta_\ell}\left\langle i\left|\prod_{\ell'=1}^{L} W_{\ell'}U_{\ell'}\right|\phi\left(\mathbf{x}_\alpha\right)\right\rangle = i\left\langle i\left|\prod_{\ell'=1}^{\ell-1} W_{\ell'}U_{\ell'}X_\ell\prod_{\ell''=\ell}^{L} W_{\ell''}U_{\ell''}\right|\phi\left(\mathbf{x}_\alpha\right)\right\rangle \ , \tag{139}$$

$$\frac{dz^*_{i;\alpha}}{d\theta_\ell} = \frac{d}{d\theta_\ell}\left\langle \phi\left(\mathbf{x}_\alpha\right)\left|\prod_{\ell'=L}^{1} U^\dagger_{\ell'}W^\dagger_{\ell'}\right|i\right\rangle = -i\left\langle \phi\left(\mathbf{x}_\alpha\right)\left|\prod_{\ell'=L}^{\ell+1} U^\dagger_{\ell'}W^\dagger_{\ell'}X_\ell\prod_{\ell''=1}^{\ell} U^\dagger_{\ell''}W^\dagger_{\ell''}\right|i\right\rangle \ . \tag{140}$$

And we have

$$\begin{aligned}\frac{dz_{i;\alpha}}{d\theta_\ell} &= i\left\langle i\left|U_{-,\ell}X_\ell W_\ell U_\ell U_{+,\ell}\right|\phi\left(\mathbf{x}_\alpha\right)\right\rangle \ ,\\ \frac{dz^*_{i;\alpha}}{d\theta_\ell} &= -i\left\langle \phi\left(\mathbf{x}_\alpha\right)\left|U^\dagger_{+,\ell}U^\dagger_\ell W^\dagger_\ell X_\ell U^\dagger_{-,\ell}\right|i\right\rangle \ .\end{aligned} \tag{141}$$

So we have

$$\begin{aligned}\sum_\ell \frac{dz_{i;\delta}}{d\theta_\ell}\frac{dz^*_{i';\tilde\alpha}}{d\theta_\ell} &= \sum_\ell \left\langle i\left|U_{-,\ell}X_\ell W_\ell U_\ell U_{+,\ell}\right|\phi\left(\mathbf{x}_\delta\right)\right\rangle\left\langle \phi\left(\mathbf{x}_{\tilde\alpha}\right)\left|U^\dagger_{+,\ell}U^\dagger_\ell W^\dagger_\ell X_\ell U^\dagger_{-,\ell}\right|i'\right\rangle \ ,\\ \sum_\ell \frac{dz_{i;\delta}}{d\theta_\ell}\frac{dz_{i';\tilde\alpha}}{d\theta_\ell} &= -\sum_\ell \left\langle i\left|U_{-,\ell}X_\ell W_\ell U_\ell U_{+,\ell}\right|\phi\left(\mathbf{x}_\delta\right)\right\rangle\left\langle i'\left|U_{-,\ell}X_\ell W_\ell U_\ell U_{+,\ell}\right|\phi\left(\mathbf{x}_{\tilde\alpha}\right)\right\rangle \ ,\end{aligned} \tag{142}$$

and their conjugates. Now, we define some notations. We define the *feature S-matrix*

$$\rho_{\delta\tilde\alpha} = \left|\phi\left(\mathbf{x}_\delta\right)\right\rangle\left\langle\phi\left(\mathbf{x}_{\tilde\alpha}\right)\right| \ , \tag{143}$$

and the *feature projector*

$$\Lambda_{i';\delta} = \left|\phi\left(\mathbf{x}_\delta\right)\right\rangle\left\langle i'\right| \ . \tag{144}$$

We have

$$\begin{aligned}K^{+,ii'}_{\delta,\tilde\alpha} &= \sum_\ell \left\langle i\left|U_{-,\ell}X_\ell W_\ell U_\ell U_{+,\ell}\rho_{\delta\tilde\alpha}U^\dagger_{+,\ell}U^\dagger_\ell W^\dagger_\ell X_\ell U^\dagger_{-,\ell}\right|i'\right\rangle \ ,\\ K^{-,ii'}_{\delta,\tilde\alpha} &= -\sum_\ell \left\langle i'\left|U_{-,\ell}X_\ell W_\ell U_\ell U_{+,\ell}\Lambda_{i;\tilde\alpha}U_{-,\ell}X_\ell W_\ell U_\ell U_{+,\ell}\right|\phi\left(\mathbf{x}_\delta\right)\right\rangle \ .\end{aligned} \tag{145}$$

B. No representation learning

In the frozen QNTK limit with the variational angle redefined, one could compute the expressions of the kernel as

$$\begin{aligned}K^{+,ii'}_{\delta,\tilde\alpha} &= \delta^2\sum_\ell \left\langle i\left|W_{-,\ell}X_\ell W_\ell W_{+,\ell}\rho_{\delta\tilde\alpha}W^\dagger_{+,\ell}W^\dagger_\ell X_\ell W^\dagger_{-,\ell}\right|i'\right\rangle \ ,\\ K^{-,ii'}_{\delta,\tilde\alpha} &= -\delta^2\sum_\ell \left\langle i'\left|W_{-,\ell}X_\ell W_\ell U_{+,\ell}\Lambda_{i;\tilde\alpha}W_{-,\ell}X_\ell W_\ell W_{+,\ell}\right|\phi\left(\mathbf{x}_\delta\right)\right\rangle \ .\end{aligned} \tag{146}$$

Here, δ is the factor we are used to redefine the variational angles θ by $\theta^* + \delta\times\varphi$, and we define

$$W_{-,\ell} \equiv \prod_{\ell'=1}^{\ell-1} W_{\ell'} \ , \quad W_{+,\ell} \equiv \prod_{\ell'=\ell+1}^{L} W_{\ell'} \ . \tag{147}$$

Moreover, one could exactly solve the gradient descent dynamics. We start from the equation

$$đ\varepsilon_{\hat\mu} = -\frac{\eta}{2}\sum_{\hat\nu} K_{\hat\mu\hat\nu}\varepsilon_{\hat\nu} \ . \tag{148}$$

Thus we have

$$\varepsilon_{\hat{\mu}_1}(t) = \sum_{\hat{\mu}_2} U_{\hat{\mu}_1\hat{\mu}_2}(t)\varepsilon_{\hat{\mu}_2}(0) \ , \tag{149}$$

where

$$U_{\hat{\mu}_1\hat{\mu}_2}(t) = \left[\left(1-\frac{\eta}{2}K\right)^t\right]_{\hat{\mu}_1\hat{\mu}_2} \ . \tag{150}$$

One can compute the convergence time as

$$\tau_c = \left\| -\log\left(1-\frac{\eta}{2}K\right)\right\| \approx \frac{\eta}{2}\left\| \begin{pmatrix} K^{+,ii'}_{\delta,\tilde{\alpha}} & K^{-,ii'}_{\delta,\tilde{\alpha}} \\ K^{*,-,ii'}_{\delta,\tilde{\alpha}} & K^{*,+,ii'}_{\delta,\tilde{\alpha}} \end{pmatrix}\right\| \ . \tag{151}$$

And we could compute the prediction similarly by

$$\begin{aligned} z_{\bar{\mu}}(\infty) &= z_{\bar{\mu}}(0) - \frac{\eta}{2}\sum_{\hat{\nu}} K_{\bar{\mu}\hat{\nu}} \sum_{t=0}^{+\infty} \varepsilon_{\hat{\nu}}(t) \\ &= z_{\bar{\mu}}(0) - \frac{\eta}{2}\sum_{\hat{\nu}} K_{\bar{\mu}\hat{\nu}} \left(\sum_{t=0,\hat{\nu}_0}^{+\infty} \left[\left(1-\frac{\eta}{2}K\right)^t\right]_{\hat{\nu}\hat{\nu}_0}\right) \varepsilon_{\hat{\nu}_0}(0) \\ &= z_{\bar{\mu}}(0) - \frac{\eta}{2}\sum_{\hat{\nu}} K_{\bar{\mu}\hat{\nu}} \sum_{\hat{\nu}_0} \left[\left(1-\left(1-\frac{\eta}{2}K\right)\right)^{-1}\right]_{\hat{\nu}\hat{\nu}_0} \varepsilon_{\hat{\nu}_0}(0) \\ &= z_{\bar{\mu}}(0) - \sum_{\hat{\nu}} K_{\bar{\mu}\hat{\nu}} \sum_{\hat{\nu}_0} \left[K^{i,-1}\right]_{\hat{\nu}\hat{\nu}_0} \varepsilon_{\hat{\nu}_0}(0) \ . \end{aligned} \tag{152}$$

Moreover, we could compute the prediction by firstly defining the kernel inverse. We define

$$\sum_{\hat{\mu}\in\mathcal{A}\times\mathbb{Z}_2\times\mathcal{H}} \tilde{K}^{\hat{\mu}_1\hat{\mu}_2}\tilde{K}_{\hat{\mu}_2\hat{\mu}_3} = \delta^{\hat{\mu}_1}_{\hat{\mu}_3} \ , \tag{153}$$

so we get

$$z_{\bar{\mu}}(\infty) = z_{\bar{\mu}}(0) - \sum_{\hat{\mu}_1,\hat{\mu}_2} \tilde{K}^{\hat{\mu}_1\hat{\mu}_2} K_{\bar{\mu}\hat{\mu}_1}\varepsilon_{\hat{\mu}_2}(0) \ . \tag{154}$$

C. Representation learning

Now we start to develop our quantum representation learning theory at the dQNTK order. We make a quadratic expansion:

$$\begin{aligned} U(\theta) \to \prod_{\ell} W_\ell \exp\left(i\delta\varphi_\ell X_\ell\right) &= \prod_{\ell} W_\ell + i\delta\sum_{\ell}\varphi_\ell W_{\ell,-}W_\ell X_\ell W_{\ell,+} \\ -\frac{1}{2}\delta^2\sum_{\ell_1,\ell_2}\varphi_{\ell_1}\varphi_{\ell_2} &\left\{ \begin{array}{l} W_{\ell_1,-}W_{\ell_1}X_{\ell_1}W_{\ell_1,\ell_2}W_{\ell_2}X_{\ell_2}W_{\ell_2,+} : \ell_1 \le \ell_2 \\ W_{\ell_2,-}W_{\ell_2}X_{\ell_2}W_{\ell_2,\ell_1}W_{\ell_1}X_{\ell_1}W_{\ell_1,+} : \ell_1 > \ell_2 \end{array}\right. \ . \end{aligned} \tag{155}$$

Now, we could call

$$\begin{aligned} G^{(1)}_\ell &\equiv W_\ell X_\ell \ , \\ G^{(2)}_{\ell_1\ell_2} &\equiv \left\{ \begin{array}{l} W_{\ell_1,-}W_{\ell_1}X_{\ell_1}W_{\ell_1,\ell_2}W_{\ell_2}X_{\ell_2}W_{\ell_2,+} : \ell_1 \le \ell_2 \\ W_{\ell_2,-}W_{\ell_2}X_{\ell_2}W_{\ell_2,\ell_1}W_{\ell_1}X_{\ell_1}W_{\ell_1,+} : \ell_1 > \ell_2 \end{array}\right. \ . \end{aligned} \tag{156}$$

So the model will look like

$$z_{j;\delta} = \sum_i (\prod_\ell W_\ell)_{ji}\phi_i\left(\mathbf{x}_\delta\right) + i\delta\sum_{i,\ell}\varphi_\ell G^{(1)}_{\ell,ji}\phi_i\left(\mathbf{x}_\delta\right) - \frac{1}{2}\delta^2\sum_{i,\ell_1,\ell_2}\varphi_{\ell_1}\varphi_{\ell_2}G^{(2)}_{\ell_1\ell_2,ji}\phi_i\left(\mathbf{x}_\delta\right) \ . \tag{157}$$

So we could write down the derivatives

$$\frac{dz_{j;\delta}}{d\varphi_\ell} = i\delta \sum_i G^{(1)}_{\ell,ji}\phi_i\left(\mathbf{x}_\delta\right) - \frac{1}{2}\delta^2 \sum_{i,\ell'} \varphi_{\ell'}(G^{(2)}_{\ell\ell',ji} + G^{(2)}_{\ell'\ell,ji})\phi_i\left(\mathbf{x}_\delta\right) . \tag{158}$$

Moreover, we have

$$\begin{aligned} đz_{j;\delta} = & i\delta \sum_i \sum_\ell đ\varphi_\ell G^{(1)}_{\ell,ji}\phi_i\left(\mathbf{x}_\delta\right) - \frac{1}{2}\delta^2 \sum_i \sum_{\ell_1,\ell_2} đ\varphi_{\ell_1}\varphi_{\ell_2} G^{(2)}_{\ell_1\ell_2,ji}\phi_i\left(\mathbf{x}_\delta\right) \\ & - \frac{1}{2}\delta^2 \sum_i \sum_{\ell_1,\ell_2} \varphi_{\ell_1} đ\varphi_{\ell_2} G^{(2)}_{\ell_1\ell_2,ji}\phi_i\left(\mathbf{x}_\delta\right) - \frac{1}{2}\delta^2 \sum_i \sum_{\ell_1,\ell_2} đ\varphi_{\ell_1} đ\varphi_{\ell_2} G^{(2)}_{\ell_1,\ell_2,ji}\phi_i\left(\mathbf{x}_\delta\right) . \end{aligned} \tag{159}$$

The leading order piece is exactly the effective feature map. A more compact version is,

$$đz_{j;\delta} = \sum_\ell \frac{dz_{j;\delta}}{d\varphi_\ell} đ\varphi_\ell - \frac{1}{2}\delta^2 \sum_i \sum_{\ell_1,\ell_2} đ\varphi_{\ell_1} đ\varphi_{\ell_2} G^{(2)}_{\ell_1,\ell_2,ji}\phi_i\left(\mathbf{x}_\delta\right) . \tag{160}$$

Firstly we need to compute the effective kernel, we have

$$\begin{aligned} \frac{dz_{j;\delta}}{d\varphi_\ell} &= i\delta \sum_i G^{(1)}_{\ell,ji}\phi_i\left(\mathbf{x}_\delta\right) - \frac{1}{2}\delta^2 \sum_{i,\ell'} \varphi_{\ell'}\left(G^{(2)}_{\ell\ell',ji} + G^{(2)}_{\ell'\ell,ji}\right)\phi_i\left(\mathbf{x}_\delta\right) , \\ \frac{dz_{i';\tilde{\alpha}}}{d\varphi_\ell} &= i\delta \sum_i G^{(1)}_{\ell,i'i}\phi_i\left(\mathbf{x}_{\tilde{\alpha}}\right) - \frac{1}{2}\delta^2 \sum_{i,\ell'} \varphi_{\ell'}\left(G^{(2)}_{\ell\ell',i'i} + G^{(2)}_{\ell'\ell,i'i}\right)\phi_i\left(\mathbf{x}_{\tilde{\alpha}}\right) , \\ \frac{dz^*_{i';\tilde{\alpha}}}{d\varphi_\ell} &= -i\delta \sum_i \phi^*_i\left(\mathbf{x}_{\tilde{\alpha}}\right) G^{\dagger,(1)}_{\ell,ii'} - \frac{1}{2}\delta^2 \sum_{i,\ell'} \varphi_{\ell'}\phi^*_i\left(\mathbf{x}_{\tilde{\alpha}}\right)\left(G^{\dagger,(2)}_{\ell\ell',ii'} + G^{\dagger,(2)}_{\ell'\ell,ii'}\right) . \end{aligned} \tag{161}$$

Thus, we define

$$K^{E,\dot{\alpha}\dot{\beta},ii'}_{\delta,\tilde{\alpha}} = \begin{pmatrix} K^{E,+,,ii'}_{\delta,\tilde{\alpha}} & K^{E,-,,ii'}_{\delta,\tilde{\alpha}} \\ K^{*,E,-,,ii'}_{\delta,\tilde{\alpha}} & K^{*,E,+,,ii'}_{\delta,\tilde{\alpha}} \end{pmatrix}(\varphi) = \begin{pmatrix} \sum\limits_\ell \frac{dz_{i;\delta}}{d\varphi_\ell}\frac{dz^*_{i';\tilde{\alpha}}}{d\varphi_\ell} & \sum\limits_\ell \frac{dz_{i;\delta}}{d\varphi_\ell}\frac{dz_{i';\tilde{\alpha}}}{d\varphi_\ell} \\ \sum\limits_\ell \frac{dz^*_{i;\delta}}{d\varphi_\ell}\frac{dz^*_{i';\tilde{\alpha}}}{d\varphi_\ell} & \sum\limits_\ell \frac{dz^*_{i;\delta}}{d\varphi_\ell}\frac{dz_{i';\tilde{\alpha}}}{d\varphi_\ell} \end{pmatrix} , \tag{162}$$

and

$$K^{E,\dot{\alpha}\dot{\beta},ii'}_{\delta,\tilde{\alpha}}(0) = K^{\dot{\alpha}\dot{\beta},ii'}_{\delta,\tilde{\alpha}} + K^{\Delta,\dot{\alpha}\dot{\beta},ii'}_{\delta,\tilde{\alpha}} = \begin{pmatrix} K^{+,ii'}_{\delta,\tilde{\alpha}} & K^{-,ii'}_{\delta,\tilde{\alpha}} \\ K^{*,-,ii'}_{\delta,\tilde{\alpha}} & K^{*,+,j}_{\delta,\tilde{\alpha}} \end{pmatrix} + \begin{pmatrix} K^{\Delta,+,ii'}_{\delta,\tilde{\alpha}} & K^{\Delta,-,ii'}_{\delta,\tilde{\alpha}} \\ K^{*,\Delta,-,ii'}_{\delta,\tilde{\alpha}} & K^{*,\Delta,+,ii'}_{\delta,\tilde{\alpha}} \end{pmatrix} . \tag{163}$$

We find

$$\begin{aligned} K^{\Delta,+,ii'}_{\delta,\tilde{\alpha}} &= \frac{i}{2}\delta^3 \sum_{j,\ell',i''} \varphi_{\ell'}(0)\left(G^{(2)}_{\ell\ell',ij} + G^{(2)}_{\ell'\ell,ij}\right) G^{\dagger,(1)}_{\ell,i''i'}\phi_j\left(\mathbf{x}_\delta\right)\phi^*_{i''}\left(\mathbf{x}_{\tilde{\alpha}}\right) \\ &- \frac{i}{2}\delta^3 \sum_{j,\ell',i''} \varphi_{\ell'}(0)\left(G^{\dagger,(2)}_{\ell\ell',i''i'} + G^{\dagger,(2)}_{\ell'\ell,i''i'}\right) G^{(1)}_{\ell,ij}\phi_j\left(\mathbf{x}_\delta\right)\phi^*_{i''}\left(\mathbf{x}_{\tilde{\alpha}}\right) , \\ K^{\Delta,-,ii'}_{\delta,\tilde{\alpha}} &= -\frac{i}{2}\delta^3 \sum_{j,\ell',i''} \varphi_{\ell'}(0)\left(G^{(2)}_{\ell\ell',ij} + G^{(2)}_{\ell'\ell,ij}\right) G^{(1)}_{\ell,i'i''}\phi_j\left(\mathbf{x}_\delta\right)\phi_{i''}\left(\mathbf{x}_{\tilde{\alpha}}\right) \\ &- \frac{i}{2}\delta^3 \sum_{j,\ell',i''} \varphi_{\ell'}(0)\left(G^{(2)}_{\ell\ell',i'i''} + G^{(2)}_{\ell'\ell,i'i''}\right) G^{(1)}_{\ell,ij}\phi_j\left(\mathbf{x}_\delta\right)\phi_{i''}\left(\mathbf{x}_{\tilde{\alpha}}\right) . \end{aligned} \tag{164}$$

We could also write it in the bracket fashion

$$\begin{aligned} K^{\Delta,+,ii'}_{\delta,\tilde{\alpha}} &= \frac{i}{2}\delta^3 \sum_{\ell'} \varphi_{\ell'}(0) \begin{pmatrix} \langle i|\left(G^{(2)}_{\ell\ell'} + G^{(2)}_{\ell'\ell}\right)|\phi\left(\mathbf{x}_\delta\right)\rangle\langle\phi\left(\mathbf{x}_{\tilde{\alpha}}\right)| G^{\dagger,(1)}_\ell |i'\rangle \\ -\langle\phi\left(\mathbf{x}_{\tilde{\alpha}}\right)|\left(G^{\dagger,(2)}_{\ell\ell'} + G^{\dagger,(2)}_{\ell'\ell}\right)|i'\rangle\langle i| G^{(1)}_\ell |\phi\left(\mathbf{x}_\delta\right)\rangle \end{pmatrix} , \\ K^{\Delta,-,ii'}_{\delta,\tilde{\alpha}} &= -\frac{i}{2}\delta^3 \sum_{\ell'} \varphi_{\ell'}(0) \begin{pmatrix} \langle i|\left(G^{(2)}_{\ell\ell'} + G^{(2)}_{\ell'\ell}\right)|\phi\left(\mathbf{x}_\delta\right)\rangle\langle i'| G^{(1)}_\ell |\phi\left(\mathbf{x}_{\tilde{\alpha}}\right)\rangle \\ +\langle i'|\left(G^{(2)}_{\ell\ell'} + G^{(2)}_{\ell'\ell}\right)|\phi_{i''}\left(\mathbf{x}_{\tilde{\alpha}}\right)\rangle\langle i| G^{(1)}_\ell |\phi\left(\mathbf{x}_\delta\right)\rangle \end{pmatrix} . \end{aligned} \tag{165}$$

Now we could write down the prediction on the training set. We have

$$\varepsilon_{\hat{\mu}}(t) = \varepsilon^F_{\hat{\mu}}(t) + \varepsilon^I_{\hat{\mu}}(t) \ , \tag{166}$$

where

$$\begin{aligned} \varepsilon^F_{\hat{\mu}}(t) &= \sum_{\hat{\mu}_1} U_{\hat{\mu}\hat{\mu}_1}(t)\varepsilon_{\hat{\mu}_1}(0) \ , \\ U_{\hat{\mu}_1\hat{\mu}_2}(t) &= \left[\left(1-\frac{\eta}{2}K\right)^t\right]_{\hat{\mu}_1\hat{\mu}_2} \ , \end{aligned} \tag{167}$$

and

$$\varepsilon^I_{\hat{\mu}}(t) = \left(-\frac{\eta}{2}\sum_{s=0}^{t-1}\left(1-\frac{\eta}{2}K\right)^{t-1-s}K^{\Delta}\left(1-\frac{\eta}{2}K\right)^{s}\varepsilon(0)\right)_{\hat{\mu}} \ . \tag{168}$$

It is the compact matrix product form in the space $\mathcal{A}\times\mathbb{Z}_2\times\mathcal{H}$. Moreover, we have

$$\left\|\varepsilon^I(t)\right\| \leq \frac{\eta}{2}t\left\|1-\frac{\eta}{2}K\right\|^{t-1}\left\|K^{\Delta}\right\|\left\|\varepsilon(0)\right\| \ . \tag{169}$$

Finally, we could discuss the asymptotic convergence. We could start by computing the meta-kernel. We notice that

$$\begin{aligned} đz_{i;\delta} &= -\frac{\eta}{2}\sum_{\ell,i',\tilde{\alpha}}\varepsilon_{i;\tilde{\alpha}}\frac{dz_{i;\delta}}{d\varphi_\ell}\frac{dz^*_{i';\tilde{\alpha}}}{d\varphi_\ell} - \frac{\eta}{2}\sum_{\ell,i',\tilde{\alpha}}\varepsilon^*_{i;\tilde{\alpha}}\frac{dz_{i;\delta}}{d\varphi_\ell}\frac{dz_{i';\tilde{\alpha}}}{d\varphi_\ell} \\ &+ \left(\frac{\eta}{2}\right)^2\sum_{\ell_1,\ell_2}\frac{d^2z_{i;\delta}}{d\varphi_{\ell_1}d\varphi_{\ell_2}}\left(\sum_{\tilde{\alpha}_1,i_1}\varepsilon_{i_1;\tilde{\alpha}_1}\frac{dz^*_{i_1;\tilde{\alpha}_1}}{d\varphi_{\ell_1}} + \sum_{\tilde{\alpha}_1,i_1}\varepsilon^*_{i_1;\tilde{\alpha}_1}\frac{dz_{i_1;\tilde{\alpha}_1}}{d\varphi_{\ell_2}}\right)\left(\sum_{\tilde{\alpha}_2,i_2}\varepsilon_{i_2;\tilde{\alpha}_2}\frac{dz^*_{i_2;\tilde{\alpha}_2}}{d\varphi_{\ell_2}} + \sum_{\tilde{\alpha}_2,i_2}\varepsilon^*_{i_2;\tilde{\alpha}_2}\frac{dz_{i_2;\tilde{\alpha}_2}}{d\varphi_{\ell_2}}\right) \ , \end{aligned} \tag{170}$$

We have

$$\begin{aligned} &\sum_{\ell_1,\ell_2}\frac{d^2z_{i;\delta}}{d\varphi_{\ell_1}d\varphi_{\ell_2}}\left(\sum_{\tilde{\alpha}_1,i_1}\varepsilon_{i_1;\tilde{\alpha}_1}\frac{dz^*_{i_1;\tilde{\alpha}_1}}{d\varphi_{\ell_1}} + \sum_{\tilde{\alpha}_1,i_1}\varepsilon^*_{i_1;\tilde{\alpha}_1}\frac{dz_{i_1;\tilde{\alpha}_1}}{d\varphi_{\ell_2}}\right)\left(\sum_{\tilde{\alpha}_2,i_2}\varepsilon_{i_2;\tilde{\alpha}_2}\frac{dz^*_{i_2;\tilde{\alpha}_2}}{d\varphi_{\ell_2}} + \sum_{\tilde{\alpha}_2,i_2}\varepsilon^*_{i_2;\tilde{\alpha}_2}\frac{dz_{i_2;\tilde{\alpha}_2}}{d\varphi_{\ell_2}}\right) \\ &= \sum_{\ell_1,\ell_2,\tilde{\alpha}_1,i_1,\tilde{\alpha}_2,i_2,\dot{\alpha}_1,\dot{\alpha}_2}\frac{d^2z_{i;\delta}}{d\varphi_{\ell_1}d\varphi_{\ell_2}}\frac{dz^{1-\dot{\alpha}_1}_{i_1;\tilde{\alpha}_1}}{d\varphi_{\ell_1}}\frac{dz^{1-\dot{\alpha}_2}_{i_2;\tilde{\alpha}_2}}{d\varphi_{\ell_2}}\varepsilon^{\dot{\alpha}_1}_{i_1;\tilde{\alpha}_1}\varepsilon^{\dot{\alpha}_2}_{i_2;\tilde{\alpha}_2} \ . \end{aligned} \tag{171}$$

So we define

$$\mu^{ii_1i_2;\dot{\beta}\dot{\alpha}_1\dot{\alpha}_2}_{\delta\tilde{\alpha}_1\tilde{\alpha}_2} = \mu_{\bar{\mu}\hat{\mu}_1\hat{\mu}_2} = \sum_{\ell_1,\ell_2}\frac{d^2z^{\dot{\beta}}_{i;\delta}}{d\varphi_{\ell_1}d\varphi_{\ell_2}}\left(\frac{dz^{1-\dot{\alpha}_1}_{i_1;\tilde{\alpha}_1}}{d\varphi_{\ell_1}}\frac{dz^{1-\dot{\alpha}_2}_{i_2;\tilde{\alpha}_2}}{d\varphi_{\ell_2}}\right)\Bigg|_{\varphi=0} \ , \tag{172}$$

in the leading order. Moreover, in general we have,

$$\mu^{i_0i_1i_2;\dot{\alpha}_0\dot{\alpha}_1\dot{\alpha}_2}_{\delta_0\delta_1\delta_2} = \mu_{\bar{\mu}_0\bar{\mu}_1\bar{\mu}_2} = \sum_{\ell_1,\ell_2}\frac{d^2z^{\dot{\alpha}_0}_{i_0;\delta_0}}{d\varphi_{\ell_1}d\varphi_{\ell_2}}\left(\frac{dz^{1-\dot{\alpha}_1}_{i_1;\delta_1}}{d\varphi_{\ell_1}}\frac{dz^{1-\dot{\alpha}_2}_{i_2;\delta_2}}{d\varphi_{\ell_2}}\right)\Bigg|_{\varphi=0} \ . \tag{173}$$

So the asymptotic convergence is given by

$$\begin{aligned} z_{\bar{\mu}}(\infty) &= z_{\bar{\mu}}(0) - \sum_{\hat{\mu}_1,\hat{\mu}_2}K_{\bar{\mu}\hat{\mu}_1}\tilde{K}^{\hat{\mu}_1\hat{\mu}_2}\varepsilon_{\hat{\mu}_2}(0) \\ &+ \sum_{\hat{\mu}_1,\hat{\mu}_2,\hat{\mu}_3,\hat{\mu}_4}\left[\mu_{\hat{\mu}_1\bar{\mu}\hat{\mu}_2} - \sum_{\hat{\mu}_5,\hat{\mu}_6}K_{\bar{\mu}\hat{\mu}_5}\tilde{K}^{\hat{\mu}_5\hat{\mu}_6}\mu_{\hat{\mu}_1\hat{\mu}_6\hat{\mu}_2}\right]Z_A^{\hat{\mu}_1\hat{\mu}_2\hat{\mu}_3\hat{\mu}_4}\varepsilon_{\hat{\mu}_3}(0)\varepsilon_{\hat{\mu}_4}(0) \\ &+ \sum_{\hat{\mu}_1,\hat{\mu}_2,\hat{\mu}_3,\hat{\mu}_4}\left[\mu_{\bar{\mu}\hat{\mu}_1\hat{\mu}_2} - \sum_{\hat{\mu}_5,\hat{\mu}_6}K_{\bar{\mu}\hat{\mu}_5}\tilde{K}^{\hat{\mu}_5\hat{\mu}_6}\mu_{\hat{\mu}_6\hat{\mu}_1\hat{\mu}_2}\right]Z_B^{\hat{\mu}_1\hat{\mu}_2\hat{\mu}_3\hat{\mu}_4}\varepsilon_{\hat{\mu}_3}(0)\varepsilon_{\hat{\mu}_4}(0) \ , \end{aligned} \tag{174}$$

where the algorithm projector $Z_{A,B}$s are

$$
\begin{aligned}
Z_A^{\hat{\mu}_1\hat{\mu}_2\hat{\mu}_3\hat{\mu}_4} &\equiv \tilde{K}^{\hat{\mu}_1\hat{\mu}_3}\tilde{K}^{\hat{\mu}_2\hat{\mu}_4} - \sum_{\hat{\mu}_5}\tilde{K}^{\hat{\mu}_2\hat{\mu}_5}X_{\parallel}^{\hat{\mu}_1\hat{\mu}_5\hat{\mu}_3\hat{\mu}_4}\ , \\
Z_B^{\hat{\mu}_1\hat{\mu}_2\hat{\mu}_3\hat{\mu}_4} &\equiv \tilde{K}^{\hat{\mu}_1\hat{\mu}_3}\tilde{K}^{\hat{\mu}_2\hat{\mu}_4} - \sum_{\hat{\mu}_5}\tilde{K}^{\hat{\mu}_2\hat{\mu}_5}X_{\parallel}^{\hat{\mu}_1\hat{\mu}_5\hat{\mu}_3\hat{\mu}_4} + \frac{\eta}{4}X_{\parallel}^{\hat{\mu}_1\hat{\mu}_2\hat{\mu}_3\hat{\mu}_4}\ ,
\end{aligned}
\tag{175}
$$

and

$$
X_{\parallel}^{\hat{\mu}_1\hat{\mu}_2\hat{\mu}_3\hat{\mu}_4} = \sum_{s=0}^{\infty}\left[(1-\frac{\eta}{2}K)^s\right]_{\hat{\mu}_1\hat{\mu}_3}\left[(1-\frac{\eta}{2}K)^s\right]_{\hat{\mu}_2\hat{\mu}_4}\ ,
\tag{176}
$$

which is defined implicitly as

$$
\delta_{\hat{\mu}_5}^{\hat{\mu}_1}\delta_{\hat{\mu}_6}^{\hat{\mu}_2} = \sum_{\hat{\mu}_3,\hat{\mu}_4} X_{\parallel}^{\hat{\mu}_1\hat{\mu}_2\hat{\mu}_3\hat{\mu}_4}\left(\tilde{K}_{\hat{\mu}_3\hat{\mu}_5}\delta_{\hat{\mu}_4\hat{\mu}_6} + \delta_{\hat{\mu}_3\hat{\mu}_5}\tilde{K}_{\hat{\mu}_4\hat{\mu}_6} - \frac{\eta}{2}\tilde{K}_{\hat{\mu}_3\hat{\mu}_5}\tilde{K}_{\hat{\mu}_4\hat{\mu}_6}\right)\ .
\tag{177}
$$

Now we write down all components of μ. We have

$$
\begin{aligned}
&\frac{d^2 z_{i_0;\delta_0}}{d\varphi_{\ell_1}d\varphi_{\ell_2}} = -\frac{1}{2}\delta^2\sum_i\left(G^{(2)}_{\ell_1\ell_2,i_0 i} + G^{(2)}_{\ell_2\ell_1,i_0 i}\right)\phi_i\left(\mathbf{x}_{\delta_0}\right), \\
&\frac{d^2 z^*_{i_0;\delta_0}}{d\varphi_{\ell_1}d\varphi_{\ell_2}} = -\frac{1}{2}\delta^2\sum_i\phi_i^*\left(\mathbf{x}_{\delta_0}\right)\left(G^{\dagger,(2)}_{\ell_1\ell_2,i i_0} + G^{\dagger,(2)}_{\ell_2\ell_1,i i_0}\right), \\
&\frac{dz^*_{i_1;\delta_1}}{d\varphi_{\ell_1}} = -i\delta\sum_{i'}\phi^*_{i'}\left(\mathbf{x}_{\delta_1}\right)G^{\dagger,(1)}_{\ell_1,i'i_1}\ , \qquad \frac{dz_{i_1;\delta_1}}{d\varphi_{\ell_1}} = i\delta\sum_{i'}G^{(1)}_{\ell_1,i_1 i'}\phi_{i'}\left(\mathbf{x}_{\delta_1}\right)\ , \\
&\frac{dz^*_{i_2;\delta_2}}{d\varphi_{\ell_2}} = -i\delta\sum_{i''}\phi^*_{i''}\left(\mathbf{x}_{\delta_2}\right)G^{\dagger,(1)}_{\ell_2,i''i_2}\ , \qquad \frac{dz_{i_2;\delta_2}}{d\varphi_{\ell_2}} = i\delta\sum_{i''}G^{(1)}_{\ell_2,i_2 i''}\phi_{i''}\left(\mathbf{x}_{\delta_2}\right)\ .
\end{aligned}
\tag{178}
$$

So

$$
\begin{aligned}
\mu^{i_0 i_1 i_2;000}_{\delta_0\delta_1\delta_2} &= \frac{1}{2}\delta^4\sum_{i,i',i'',\ell_1,\ell_2}\left(G^{(2)}_{\ell_1\ell_2,i_0 i} + G^{(2)}_{\ell_2\ell_1,i_0 i}\right)G^{\dagger,(1)}_{\ell_1,i'i_1}G^{\dagger,(2)}_{\ell_2,i''i_2}\phi_i\left(\mathbf{x}_{\delta_0}\right)\phi^*_{i'}\left(\mathbf{x}_{\delta_1}\right)\phi^*_{i''}\left(\mathbf{x}_{\delta_2}\right), \\
\mu^{i_0 i_1 i_2;001}_{\delta_0\delta_1\delta_2} &= -\frac{1}{2}\delta^4\sum_{i,i',i'',\ell_1,\ell_2}\left(G^{(2)}_{\ell_1\ell_2,i_0 i} + G^{(2)}_{\ell_2\ell_1,i_0 i}\right)G^{\dagger,(1)}_{\ell_1,i'i_1}G^{(2)}_{\ell_2,i_2 i''}\phi_i\left(\mathbf{x}_{\delta_0}\right)\phi^*_{i'}\left(\mathbf{x}_{\delta_1}\right)\phi_{i''}\left(\mathbf{x}_{\delta_2}\right), \\
\mu^{i_0 i_1 i_2;010}_{\delta_0\delta_1\delta_2} &= -\frac{1}{2}\delta^4\sum_{i,i',i'',\ell_1,\ell_2}\left(G^{(2)}_{\ell_1\ell_2,i_0 i} + G^{(2)}_{\ell_2\ell_1,i_0 i}\right)G^{(1)}_{\ell_1,i_1 i'}G^{\dagger,(2)}_{\ell_2,i''i_2}\phi_i\left(\mathbf{x}_{\delta_0}\right)\phi_{i'}\left(\mathbf{x}_{\delta_1}\right)\phi^*_{i''}\left(\mathbf{x}_{\delta_2}\right), \\
\mu^{i_0 i_1 i_2;011}_{\delta_0\delta_1\delta_2} &= \frac{1}{2}\delta^4\sum_{i,i',i'',\ell_1,\ell_2}\left(G^{(2)}_{\ell_1\ell_2,i_0 i} + G^{(2)}_{\ell_2\ell_1,i_0 i}\right)G^{(1)}_{\ell_1,i_1 i'}G^{(2)}_{\ell_2,i_2 i''}\phi_i\left(\mathbf{x}_{\delta_0}\right)\phi_{i'}\left(\mathbf{x}_{\delta_1}\right)\phi_{i''}\left(\mathbf{x}_{\delta_2}\right), \\
\mu^{i_0 i_1 i_2;100}_{\delta_0\delta_1\delta_2} &= \frac{1}{2}\delta^4\sum_{i,i',i'',\ell_1,\ell_2}\left(G^{\dagger,(2)}_{\ell_1\ell_2,i i_0} + G^{\dagger,(2)}_{\ell_2\ell_1,i i_0}\right)G^{\dagger,(1)}_{\ell_1,i'i_1}G^{\dagger,(2)}_{\ell_2,i''i_2}\phi^*_i\left(\mathbf{x}_{\delta_0}\right)\phi^*_{i'}\left(\mathbf{x}_{\delta_1}\right)\phi^*_{i''}\left(\mathbf{x}_{\delta_2}\right), \\
\mu^{i_0 i_1 i_2;101}_{\delta_0\delta_1\delta_2} &= -\frac{1}{2}\delta^4\sum_{i,i',i'',\ell_1,\ell_2}\left(G^{\dagger,(2)}_{\ell_1\ell_2,i i_0} + G^{\dagger,(2)}_{\ell_2\ell_1,i i_0}\right)G^{\dagger,(1)}_{\ell_1,i'i_1}G^{(2)}_{\ell_2,i_2 i''}\phi^*_i\left(\mathbf{x}_{\delta_0}\right)\phi^*_{i'}\left(\mathbf{x}_{\delta_1}\right)\phi_{i''}\left(\mathbf{x}_{\delta_2}\right), \\
\mu^{i_0 i_1 i_2;110}_{\delta_0\delta_1\delta_2} &= -\frac{1}{2}\delta^4\sum_{i,i',i'',\ell_1,\ell_2}\left(G^{\dagger,(2)}_{\ell_1\ell_2,i i_0} + G^{\dagger,(2)}_{\ell_2\ell_1,i i_0}\right)G^{(1)}_{\ell_1,i_1 i'}G^{\dagger,(2)}_{\ell_2,i''i_2}\phi^*_i\left(\mathbf{x}_{\delta_0}\right)\phi_{i'}\left(\mathbf{x}_{\delta_1}\right)\phi^*_{i''}\left(\mathbf{x}_{\delta_2}\right), \\
\mu^{i_0 i_1 i_2;111}_{\delta_0\delta_1\delta_2} &= \frac{1}{2}\delta^4\sum_{i,i',i'',\ell_1,\ell_2}\left(G^{\dagger,(2)}_{\ell_1\ell_2,i i_0} + G^{\dagger,(2)}_{\ell_2\ell_1,i i_0}\right)G^{(1)}_{\ell_1,i_1 i'}G^{(2)}_{\ell_2,i_2 i''}\phi^*_i\left(\mathbf{x}_{\delta_0}\right)\phi_{i'}\left(\mathbf{x}_{\delta_1}\right)\phi_{i''}\left(\mathbf{x}_{\delta_2}\right).
\end{aligned}
\tag{179}
$$

One could also write them in the bracket notations.

$$\mu^{i_0i_1i_2;000}_{\delta_0\delta_1\delta_2} = \frac{1}{2}\delta^4 \sum_{\ell_1,\ell_2} \langle i_0| \left(G^{(2)}_{\ell_1\ell_2} + G^{(2)}_{\ell_2\ell_1}\right) |\phi\left(\mathbf{x}_{\delta_0}\right)\rangle \langle\phi\left(\mathbf{x}_{\delta_1}\right)| G^{\dagger,(1)}_{\ell_1} |i_1\rangle \langle\phi\left(\mathbf{x}_{\delta_2}\right)| G^{\dagger,(1)}_{\ell_2} |i_2\rangle \ ,$$

$$\mu^{i_0i_1i_2;001}_{\delta_0\delta_1\delta_2} = -\frac{1}{2}\delta^4 \sum_{\ell_1,\ell_2} \langle i_0| \left(G^{(2)}_{\ell_1\ell_2} + G^{(2)}_{\ell_2\ell_1}\right) |\phi\left(\mathbf{x}_{\delta_0}\right)\rangle \langle\phi\left(\mathbf{x}_{\delta_1}\right)| G^{\dagger,(1)}_{\ell_1} |i_1\rangle \langle i_2| G^{(1)}_{\ell_2} |\phi\left(\mathbf{x}_{\delta_2}\right)\rangle \ ,$$

$$\mu^{i_0i_1i_2;010}_{\delta_0\delta_1\delta_2} = -\frac{1}{2}\delta^4 \sum_{\ell_1,\ell_2} \langle i_0| \left(G^{(2)}_{\ell_1\ell_2} + G^{(2)}_{\ell_2\ell_1}\right) |\phi\left(\mathbf{x}_{\delta_0}\right)\rangle \langle i_1| G^{(1)}_{\ell_1} |\phi\left(\mathbf{x}_{\delta_1}\right)\rangle \langle\phi\left(\mathbf{x}_{\delta_2}\right)| G^{\dagger,(1)}_{\ell_2} |i_2\rangle \ ,$$

$$\mu^{i_0i_1i_2;011}_{\delta_0\delta_1\delta_2} = \frac{1}{2}\delta^4 \sum_{\ell_1,\ell_2} \langle i_0| \left(G^{(2)}_{\ell_1\ell_2} + G^{(2)}_{\ell_2\ell_1}\right) |\phi\left(\mathbf{x}_{\delta_0}\right)\rangle \langle i_1| G^{(1)}_{\ell_1} |\phi\left(\mathbf{x}_{\delta_1}\right)\rangle \langle i_2| G^{(1)}_{\ell_2} |\phi\left(\mathbf{x}_{\delta_2}\right)\rangle \ ,$$

$$\mu^{i_0i_1i_2;100}_{\delta_0\delta_1\delta_2} = \frac{1}{2}\delta^4 \sum_{\ell_1,\ell_2} \langle\phi\left(\mathbf{x}_{\delta_0}\right)| \left(G^{\dagger,(2)}_{\ell_1\ell_2} + G^{\dagger,(2)}_{\ell_2\ell_1}\right) |i_0\rangle \langle\phi\left(\mathbf{x}_{\delta_1}\right)| G^{\dagger,(1)}_{\ell_1} |i_1\rangle \langle\phi\left(\mathbf{x}_{\delta_2}\right)| G^{\dagger,(1)}_{\ell_2} |i_2\rangle \ ,$$

$$\mu^{i_0i_1i_2;101}_{\delta_0\delta_1\delta_2} = -\frac{1}{2}\delta^4 \sum_{\ell_1,\ell_2} \langle\phi\left(\mathbf{x}_{\delta_0}\right)| \left(G^{\dagger,(2)}_{\ell_1\ell_2} + G^{\dagger,(2)}_{\ell_2\ell_1}\right) |i_0\rangle \langle\phi\left(\mathbf{x}_{\delta_1}\right)| G^{\dagger,(1)}_{\ell_1} |i_1\rangle \langle i_2| G^{(1)}_{\ell_2} |\phi\left(\mathbf{x}_{\delta_2}\right)\rangle \ ,$$

$$\mu^{i_0i_1i_2;110}_{\delta_0\delta_1\delta_2} = -\frac{1}{2}\delta^4 \sum_{\ell_1,\ell_2} \langle\phi\left(\mathbf{x}_{\delta_0}\right)| \left(G^{\dagger,(2)}_{\ell_1\ell_2} + G^{\dagger,(2)}_{\ell_2\ell_1}\right) |i_0\rangle \langle i_1| G^{(1)}_{\ell_1} |\phi\left(\mathbf{x}_{\delta_1}\right)\rangle \langle\phi\left(\mathbf{x}_{\delta_2}\right)| G^{\dagger,(1)}_{\ell_2} |i_2\rangle \ ,$$

$$\mu^{i_0i_1i_2;111}_{\delta_0\delta_1\delta_2} = \frac{1}{2}\delta^4 \sum_{\ell_1,\ell_2} \langle\phi\left(\mathbf{x}_{\delta_0}\right)| \left(G^{\dagger,(2)}_{\ell_1\ell_2} + G^{\dagger,(2)}_{\ell_2\ell_1}\right) |i_0\rangle \langle i_1| G^{(1)}_{\ell_1} |\phi\left(\mathbf{x}_{\delta_1}\right)\rangle \langle i_2| G^{(1)}_{\ell_2} |\phi\left(\mathbf{x}_{\delta_2}\right)\rangle \ . \tag{180}$$

D. Reading the amplitude

Here, we review the amplitude protocol for realizing the evaluation of the inner product of two states: $\langle x|y\rangle$ [7]. The protocol is very similar to the celebrated Hadamard test. For a given pair of quantum states $|x\rangle$ and $|y\rangle$, we need to get access to the state

$$|\varphi\rangle = \frac{1}{\sqrt{2}}(|0\rangle|x\rangle + |1\rangle|y\rangle) \ . \tag{181}$$

The first qubit is serving as an ancillary qubit. Applying the Hadamard gate to the first qubit, we get

$$|\varphi\rangle \to \frac{1}{2}(|0\rangle(|x\rangle + |y\rangle) + |1\rangle(|x\rangle - |y\rangle)) \ . \tag{182}$$

Now we could measure the probability to obtain $|0\rangle$ in the ancillary system. We have

$$p = \frac{1}{2}(1 + \mathrm{Re}(\langle x|y\rangle) \ . \tag{183}$$

Thus, we could use the probability to estimate the real part of the inner product. Moreover, we could add a phase rotation to $|\varphi\rangle$ to get

$$|\varphi\rangle = \frac{1}{\sqrt{2}}(|0\rangle|x\rangle - i|1\rangle|y\rangle) \ . \tag{184}$$

Then, we apply the same Hadamard gate and measure the first qubit in the Pauli-Z basis. We get the probability,

$$p = \frac{1}{2}(1 + \mathrm{Im}(\langle x|y\rangle) \ , \tag{185}$$

to obtain $|0\rangle$. Similar to the Hadamard test where we need to get access to a controlled unitary acting on an arbitrary state, here we need to get access to the state $|\varphi\rangle$. The inner product evaluation operation has a statistical error coming from the measurement. If we measure N times, the error scales as $1/\sqrt{N}$.

V. SUPPRESSION OF NON-GAUSSIANITY IN THE LARGE-WIDTH LIMIT

In this section, we visit the statistics of hybrid quantum-classical neural networks. The model is defined as the following. First, we initialize the neural network by a quantum model,

$$z^Q_{1;\alpha;j_1} = \left\langle \phi_1\left(\mathbf{x}_\alpha\right) \left| U^{\dagger,1}\left(\theta^1\right) O^1_{j_1} U^1\left(\theta^1\right) \right| \phi_1\left(\mathbf{x}_\alpha\right)\right\rangle \ . \tag{186}$$

Here we use the notation j_ω to denote the index of the operator space $\mathcal{O}^\omega(\mathcal{H}^\omega)$, where $\omega \in [1, 2, \cdots, \Omega]$ is denoting the layer of the hybrid quantum-classical neural network. Here we are starting our first layer, so $\omega = 1$. We use $z^Q_{\omega;\alpha;j_\omega}$ to denote our quantum model output. We also use

$$U^\omega(\theta^\omega) = \prod_{\ell_\omega=1}^{L_\omega} W^\omega_{\ell_\omega} \exp\left(i\theta^\omega_{\ell_\omega} X^\omega_{\ell_\omega}\right) \ , \tag{187}$$

to denote our ωth quantum ansatz, and ϕ_ω is used to denote the ωth feature map. In each layer, after the quantum network, we connect it with a classical neural network by

$$w^C_{\omega;\alpha;j^C_\omega} = \sigma^\omega_{j^C_\omega}\left(\sum_{j_\omega=1}^{\dim \mathcal{O}^\omega(\mathcal{H}^\omega)} W^\omega_{j^C_\omega,j_\omega} z^Q_{\omega;\alpha;j_\omega} + b^\omega_{j^C_\omega}\right) \equiv \sigma^\omega_{j^C_\omega}\left(z^C_{\omega;\alpha;j^C_\omega}\right) \ . \tag{188}$$

Here, we call

$$z^C_{\omega;\alpha;j^C_\omega} = \sum_{j_\omega=1}^{\dim \mathcal{O}^\omega(\mathcal{H}^\omega)} W^\omega_{j^C_\omega,j_\omega} z^Q_{\omega;\alpha;j_\omega} + b^\omega_{j^C_\omega} \ , \tag{189}$$

as classical preactivation in each layer, and we use the non-linear activation σ. At initialization, we will set all Ws and bs distributed randomly from the following Gaussian statistics:

$$\begin{aligned}
&\mathbb{E}\left(W^\omega_{j^C_{1,\omega},j_{1,\omega}} W^\omega_{j^C_{2,\omega},j_{2,\omega}}\right) = \delta_{j^C_{1,\omega},j^C_{2,\omega}} \delta_{j_{1,\omega},j_{2,\omega}} \frac{C^\omega_W}{\dim \mathcal{O}^\omega(\mathcal{H}^\omega)} \ , \\
&\mathbb{E}\left(b^\omega_{j^C_{1,\omega}} b^\omega_{j^C_{2,\omega}}\right) = \delta_{j^C_{1,\omega},j^C_{2,\omega}} C^\omega_b \ .
\end{aligned} \tag{190}$$

Moreover, after the classical network in each layer, we can move to the next quantum layer by doing the following encoding,

$$z^Q_{\omega;\alpha;j_\omega} = \left\langle \phi_\omega\left(\mathbf{w}_{\omega-1;\alpha}\right)\right| U^{\dagger,\omega}(\theta^\omega) O^\omega_{j_\omega} U^\omega(\theta^\omega) \left|\phi_\omega\left(\mathbf{w}_{\omega-1;\alpha}\right)\right\rangle \ , \tag{191}$$

where we use the vector notation $\mathbf{w}_{\omega;\alpha} = \left(w_{\omega;\alpha}\right)_{j^C_\omega}$. Moreover, we will assume that the initial distribution of quantum variational angles is given by a statistical ensemble. We denote all those ensemble averages as $\mathbb{E}$. One could, for instance, compute the two-point function as

$$\begin{aligned}
&\mathbb{E}\left(z^C_{\omega;\alpha;j^C_{1,\omega}} z^C_{\omega;\beta;j^C_{2,\omega}}\right) = \mathbb{E}\left(\left(\sum_{j_{1,\omega}=1}^{\dim \mathcal{O}^\omega(\mathcal{H}^\omega)} W^\omega_{j^C_{1,\omega} j_{1,\omega}} z^Q_{\omega;\alpha;j_{1,\omega}} + b^\omega_{j^C_{1,\omega}}\right)\left(\sum_{j_{2,\omega}=1}^{\dim \mathcal{O}^\omega(\mathcal{H}^\omega)} W^\omega_{j^C_{2,\omega},j_{2,\omega}} z^Q_{\omega;\beta;j_{2,\omega}} + b^\omega_{j^C_{2,\omega}}\right)\right) \\
&= \sum_{j_{1,\omega},j_{2,\omega}=1}^{\dim \mathcal{O}^\omega(\mathcal{H}^\omega)} \mathbb{E}\left(W^\omega_{j^C_{1,\omega} j_{1,\omega}} W^\omega_{j^C_{2,\omega},j_{2,\omega}}\right) \mathbb{E}\left(z^Q_{\omega;\alpha;j_{1,\omega}} z^Q_{\omega;\beta;j_{2,\omega}}\right) + \mathbb{E}\left(b^\omega_{j^C_{1,\omega}} b^\omega_{j^C_{2,\omega}}\right) \\
&= \frac{C^\omega_W}{\dim \mathcal{O}^\omega(\mathcal{H}^\omega)} \sum_{j_{1,\omega},j_{2,\omega}=1}^{\dim \mathcal{O}^\omega(\mathcal{H}^\omega)} \delta_{j^C_{1,\omega},j^C_{2,\omega}} \delta_{j_{1,\omega},j_{2,\omega}} \mathbb{E}(z^Q_{\omega;\alpha;j_{1,\omega}} z^Q_{\omega;\beta;j_{2,\omega}}) + \delta_{j^C_{1,\omega},j^C_{2,\omega}} C^\omega_b \\
&= \delta_{j^C_{1,\omega},j^C_{2,\omega}} \left(\frac{C^\omega_W}{\dim \mathcal{O}^\omega(\mathcal{H}^\omega)} \sum_{j_\omega=1}^{\dim \mathcal{O}^\omega(\mathcal{H}^\omega)} \mathbb{E}(z^Q_{\omega;\alpha;j_\omega} z^Q_{\omega;\beta;j_\omega}) + C^\omega_b\right) \ .
\end{aligned} \tag{192}$$

In order to compute non-Gaussianities of preactivations, we start from the connected part of the two-point function, which is given by

$$\begin{aligned}
&\mathbb{E}_{\text{conn}}\left(z^C_{\omega;\alpha_1;j^C_{1,\omega}} z^C_{\omega;\alpha_2;j^C_{2,\omega}} z^C_{\omega;\alpha_3;j^C_{3,\omega}} z^C_{\omega;\alpha_4;j^C_{4,\omega}}\right) \\
&\equiv \mathbb{E}\left(z^C_{\omega;\alpha_1;j^C_{1,\omega}} z^C_{\omega;\alpha_2;j^C_{2,\omega}} z^C_{\omega;\alpha_3;j^C_{3,\omega}} z^C_{\omega;\alpha_4;j^C_{4,\omega}}\right) - \mathbb{E}\left(z^C_{\omega;\alpha_1;j^C_{1,\omega}} z^C_{\omega;\alpha_2;j^C_{2,\omega}}\right)\mathbb{E}\left(z^C_{\omega;\alpha_3;j^C_{3,\omega}} z^C_{\omega;\alpha_4;j^C_{4,\omega}}\right) \\
&- \mathbb{E}\left(z^C_{\omega;\alpha_1;j^C_{1,\omega}} z^C_{\omega;\alpha_3;j^C_{3,\omega}}\right)\mathbb{E}\left(z^C_{\omega;\alpha_2;j^C_{2,\omega}} z^C_{\omega;\alpha_4;j^C_{4,\omega}}\right) - \mathbb{E}\left(z^C_{\omega;\alpha_1;j^C_{1,\omega}} z^C_{\omega;\alpha_4;j^C_{4,\omega}}\right)\mathbb{E}\left(z^C_{\omega;\alpha_2;j^C_{2,\omega}} z^C_{\omega;\alpha_3;j^C_{3,\omega}}\right) . \qquad (193)
\end{aligned}$$

We proceed by direct computation. We have

$$\begin{aligned}
&\mathbb{E}\left(z^C_{\omega;\alpha_1;j^C_{1,\omega}} z^C_{\omega;\alpha_2;j^C_{2,\omega}} z^C_{\omega;\alpha_3;j^C_{3,\omega}} z^C_{\omega;\alpha_4;j^C_{4,\omega}}\right) \\
&= \mathbb{E}\left(\begin{pmatrix}\sum_{j_{1,\omega}=1}^{\dim \mathcal{O}^\omega(\mathcal{H}^\omega)} W^\omega_{j^C_{1,\omega},j_{1,\omega}} z^Q_{\omega;\alpha_1;j_{1,\omega}} + b^\omega_{j^C_{1,\omega}} \\ \sum_{j_{3,\omega}=1}^{\dim \mathcal{O}^\omega(\mathcal{H}^\omega)} W^\omega_{j^C_{3,\omega},j_{3,\omega}} z^Q_{\omega;\alpha_3;j_{3,\omega}} + b^\omega_{j^C_{3,\omega}}\end{pmatrix}\begin{pmatrix}\sum_{j_{2,\omega}=1}^{\dim \mathcal{O}^\omega(\mathcal{H}^\omega)} W^\omega_{j^C_{2,\omega},j_{2,\omega}} z^Q_{\omega;\alpha_2;j_{2,\omega}} + b^\omega_{j^C_{2,\omega}} \\ \sum_{j_{4,\omega}=1}^{\dim \mathcal{O}^\omega(\mathcal{H}^\omega)} W^\omega_{j^C_{4,\omega},j_{4,\omega}} z^Q_{\omega;\alpha_4;j_{4,\omega}} + b^\omega_{j^C_{4,\omega}}\end{pmatrix}\right) \\
&= \mathbb{E}\left(\sum_{j_{1,\omega},j_{2,\omega},j_{3,\omega},j_{4,\omega}=1}^{\dim \mathcal{O}^\omega(\mathcal{H}^\omega)} W^\omega_{j^C_{1,\omega},j_{1,\omega}} W^\omega_{j^C_{2,\omega},j_{2,\omega}} W^\omega_{j^C_{3,\omega},j_{3,\omega}} W^\omega_{j^C_{4,\omega},j_{4,\omega}} z^Q_{\omega;\alpha_1;j_{1,\omega}} z^Q_{\omega;\alpha_2;j_{2,\omega}} z^Q_{\omega;\alpha_3;j_{3,\omega}} z^Q_{\omega;\alpha_4;j_{4,\omega}}\right) \\
&+ \mathbb{E}\left(b^\omega_{j^C_{3,\omega}} b^\omega_{j^C_{4,\omega}} \sum_{j_{1,\omega},j_{2,\omega}=1}^{\dim \mathcal{O}^\omega(\mathcal{H}^\omega)} W^\omega_{j^C_{1,\omega},j_{1,\omega}} W^\omega_{j^C_{2,\omega},j_{2,\omega}} z^Q_{\omega;\alpha_1;j_{1,\omega}} z^Q_{\omega;\alpha_2;j_{2,\omega}}\right) + (5 \text{ perms.}) \\
&+ \mathbb{E}\left(b^\omega_{j^C_{1,\omega}} b^\omega_{j^C_{2,\omega}} b^\omega_{j^C_{3,\omega}} b^\omega_{j^C_{4,\omega}}\right) . \qquad (194)
\end{aligned}$$

The notation 5 perms means that $\begin{pmatrix}4\\2\end{pmatrix} - 1 = 5$ permutations of indices 1,2,3,4. Now, combining with the disconnected four-point function result,

$$\begin{aligned}
&\mathbb{E}\left(z^C_{\omega;\alpha_1;j^C_{1,\omega}} z^C_{\omega;\alpha_2;j^C_{2,\omega}}\right)\mathbb{E}\left(z^C_{\omega;\alpha_3;j^C_{3,\omega}} z^C_{\omega;\alpha_4;j^C_{4,\omega}}\right) \\
&= \delta_{j^C_{1,\omega},j^C_{2,\omega}}\delta_{j^C_{3,\omega},j^C_{4,\omega}}\begin{pmatrix}\left(\frac{C^\omega_W}{\dim \mathcal{O}^\omega(\mathcal{H}^\omega)} \sum_{j_{1,\omega}=1}^{\dim \mathcal{O}^\omega(\mathcal{H}^\omega)} \mathbb{E}\left(z^Q_{\omega;\alpha_1;j_{1,\omega}} z^Q_{\omega;\alpha_2;j_{1,\omega}}\right) + C^\omega_b\right) \\ \left(\frac{C^\omega_W}{\dim \mathcal{O}^\omega(\mathcal{H}^\omega)} \sum_{j_{2,\omega}=1}^{\dim \mathcal{O}^\omega(\mathcal{H}^\omega)} \mathbb{E}\left(z^Q_{\omega;\alpha_3;j_{2,\omega}} z^Q_{\omega;\alpha_4;j_{2,\omega}}\right) + C^\omega_b\right)\end{pmatrix} \\
&= \delta_{j^C_{1,\omega},j^C_{2,\omega}}\delta_{j^C_{3,\omega},j^C_{4,\omega}}\begin{pmatrix}\left(\frac{C^\omega_W}{\dim \mathcal{O}^\omega(\mathcal{H}^\omega)}\right)^2 \left(\sum_{j_{1,\omega}=1}^{\dim \mathcal{O}^\omega(\mathcal{H}^\omega)} \mathbb{E}\left(z^Q_{\omega;\alpha_1;j_{1,\omega}} z^Q_{\omega;\alpha_2;j_{1,\omega}}\right)\right)\left(\sum_{j_{2,\omega}=1}^{\dim \mathcal{O}^\omega(\mathcal{H}^\omega)} \mathbb{E}\left(z^Q_{\omega;\alpha_3;j_{2,\omega}} z^Q_{\omega;\alpha_4;j_{2,\omega}}\right)\right) \\ + \frac{C^\omega_W C^\omega_b}{\dim \mathcal{O}^\omega(\mathcal{H}^\omega)}\left(\sum_{j_{1,\omega}=1}^{\dim \mathcal{O}^\omega(\mathcal{H}^\omega)} \mathbb{E}\left(z^Q_{\omega;\alpha_1;j_{1,\omega}} z^Q_{\omega;\alpha_2;j_{1,\omega}}\right) + \sum_{j_{2,\omega}=1}^{\dim \mathcal{O}^\omega(\mathcal{H}^\omega)} \mathbb{E}\left(z^Q_{\omega;\alpha_3;j_{2,\omega}} z^Q_{\omega;\alpha_4;j_{2,\omega}}\right)\right) \\ +(C^\omega_b)^2\end{pmatrix}, \qquad (195)
\end{aligned}$$

and its two other t (14-23) and u-channel (13-24) permutations, we have

$$\begin{aligned}
&\mathbb{E}_{\text{conn}}\left(z^C_{\omega;\alpha_1;j^C_{1,\omega}} z^C_{\omega;\alpha_2;j^C_{2,\omega}} z^C_{\omega;\alpha_3;j^C_{3,\omega}} z^C_{\omega;\alpha_4;j^C_{4,\omega}}\right) \\
&= \left(\frac{C^\omega_W}{\dim \mathcal{O}^\omega(\mathcal{H}^\omega)}\right)^2 \begin{pmatrix}\delta_{j^C_{1,\omega},j^C_{2,\omega}}\delta_{j^C_{3,\omega},j^C_{4,\omega}} \sum_{j_{1,\omega},j_{2,\omega}=1}^{\dim \mathcal{O}^\omega(\mathcal{H}^\omega)} \mathbb{E}_{\text{conn}}\left(z^Q_{\omega;\alpha_1;j_{1,\omega}} z^Q_{\omega;\alpha_2;j_{1,\omega}} z^Q_{\omega;\alpha_3;j_{2,\omega}} z^Q_{\omega;\alpha_4;j_{2,\omega}}\right) \\ +\delta_{j^C_{1,\omega},j^C_{3,\omega}}\delta_{j^C_{2,\omega},j^C_{4,\omega}} \sum_{j_{1,\omega},j_{2,\omega}=1}^{\dim \mathcal{O}^\omega(\mathcal{H}^\omega)} \mathbb{E}_{\text{conn}}\left(z^Q_{\omega;\alpha_1;j_{1,\omega}} z^Q_{\omega;\alpha_3;j_{1,\omega}} z^Q_{\omega;\alpha_2;j_{2,\omega}} z^Q_{\omega;\alpha_4;j_{2,\omega}}\right) \\ +\delta_{j^C_{1,\omega},j^C_{4,\omega}}\delta_{j^C_{2,\omega},j^C_{3,\omega}} \sum_{j_{1,\omega},j_{2,\omega}=1}^{\dim \mathcal{O}^\omega(\mathcal{H}^\omega)} \mathbb{E}_{\text{conn}}\left(z^Q_{\omega;\alpha_1;j_{1,\omega}} z^Q_{\omega;\alpha_4;j_{1,\omega}} z^Q_{\omega;\alpha_2;j_{2,\omega}} z^Q_{\omega;\alpha_3;j_{2,\omega}}\right)\end{pmatrix} . \qquad (196)
\end{aligned}$$

Here, the connected piece made by quantum circuits is given by

$$\begin{aligned}
&\mathbb{E}_{\text{conn}}\left(z^Q_{\omega;\alpha_1;j_{1,\omega}} z^Q_{\omega;\alpha_2;j_{1,\omega}} z^Q_{\omega;\alpha_3;j_{2,\omega}} z^Q_{\omega;\alpha_4;j_{2,\omega}}\right)\\
&\equiv \mathbb{E}\left(z^Q_{\omega;\alpha_1;j_{1,\omega}} z^Q_{\omega;\alpha_2;j_{1,\omega}} z^Q_{\omega;\alpha_3;j_{2,\omega}} z^Q_{\omega;\alpha_4;j_{2,\omega}}\right)\\
&- \mathbb{E}\left(z^Q_{\omega;\alpha_1;j_{1,\omega}} z^Q_{\omega;\alpha_2;j_{1,\omega}}\right)\mathbb{E}\left(z^Q_{\omega;\alpha_3;j_{2,\omega}} z^Q_{\omega;\alpha_4;j_{2,\omega}}\right) \ .
\end{aligned} \tag{197}$$

Now, we note that since we have $\mathcal{O}\left(\left(\dim \mathcal{O}^\omega(\mathcal{H}^\omega)\right)^2\right)$ terms in the sum, the connected part will at most scale as $\mathcal{O}(1)$ since the quantum outputs are made by normalized states whose norms are bounded by 1. However, if we assume

$$\mathbb{E}_{\text{conn}}\left(z^Q_{\omega;\alpha_1;j_{1,\omega}} z^Q_{\omega;\alpha_2;j_{1,\omega}} z^Q_{\omega;\alpha_3;j_{2,\omega}} z^Q_{\omega;\alpha_4;j_{2,\omega}}\right) = \mathcal{O}(1)\times\delta_{j_{1,\omega},j_{2,\omega}} \ , \tag{198}$$

and their permutations for all ωs, we have

$$\begin{aligned}
&\mathbb{E}_{\text{conn}}\left(z^Q_{\omega;\alpha_1;j_{1,\omega}} z^Q_{\omega;\alpha_2;j_{1,\omega}} z^Q_{\omega;\alpha_3;j_{2,\omega}} z^Q_{\omega;\alpha_4;j_{2,\omega}}\right)\\
&= \left(\frac{C_W^\omega}{\dim \mathcal{O}^\omega(\mathcal{H}^\omega)}\right)^2\left(\delta_{j^C_{1,\omega},j^C_{2,\omega}}\delta_{j^C_{3,\omega},j^C_{4,\omega}} + \delta_{j^C_{1,\omega},j^C_{3,\omega}}\delta_{j^C_{2,\omega},j^C_{4,\omega}} + \delta_{j^C_{1,\omega},j^C_{4,\omega}}\delta_{j^C_{2,\omega},j^C_{3,\omega}}\right)\times\sum_{j_{1,\omega},j_{2,\omega}=1}^{\dim \mathcal{O}^\omega(\mathcal{H}^\omega)}\delta_{j_{1,\omega},j_{2,\omega}}\times\mathcal{O}(1)\\
&= \left(\frac{C_W^\omega}{\dim \mathcal{O}^\omega(\mathcal{H}^\omega)}\right)^2\times\mathcal{O}(\dim \mathcal{O}^\omega(\mathcal{H}^\omega))\times\left(\delta_{j^C_{1,\omega},j^C_{2,\omega}}\delta_{j^C_{3,\omega},j^C_{4,\omega}} + \delta_{j^C_{1,\omega},j^C_{3,\omega}}\delta_{j^C_{2,\omega},j^C_{4,\omega}} + \delta_{j^C_{1,\omega},j^C_{4,\omega}}\delta_{j^C_{2,\omega},j^C_{3,\omega}}\right)\\
&= \frac{\left(C_W^\omega\right)^2}{\dim \mathcal{O}^\omega(\mathcal{H}^\omega)}\times\mathcal{O}(1)\times\left(\delta_{j^C_{1,\omega},j^C_{2,\omega}}\delta_{j^C_{3,\omega},j^C_{4,\omega}} + \delta_{j^C_{1,\omega},j^C_{3,\omega}}\delta_{j^C_{2,\omega},j^C_{4,\omega}} + \delta_{j^C_{1,\omega},j^C_{4,\omega}}\delta_{j^C_{2,\omega},j^C_{3,\omega}}\right) \ .
\end{aligned} \tag{199}$$

Thus, the connected part is suppressed by the large width $1/\dim \mathcal{O}^\omega(\mathcal{H}^\omega)$.

The orthogonal condition we impose for quantum neural networks might be satisfied by random assumptions of choices of operators O or random variational ansätze (say, averaging over the Pauli group or the 1-design). Similar arguments could be made for the dynamical NTK. It is beyond our scope in this work to discuss how to connect the suppressed non-Gaussian correlations randomized over initialization, and the suppression of dNTK (dQNTK) in dynamics in the large width. Thus, we hereby make more general arguments and leave the detailed work in the future. Consider the schematic formula of dQNTK in the hybrid network

$$\mu^{i_0 i_1 i_2}_{\delta_0\delta_1\delta_2} = \sum_{\ell_1,\ell_2}\frac{d^2 z_{i_0;\delta_0}}{d\theta_{\ell_1}d\theta_{\ell_2}}\left(\frac{dz_{i_1;\delta_1}}{d\theta_{\ell_1}}\frac{dz_{i_2;\delta_2}}{d\theta_{\ell_2}}\right) \ , \tag{200}$$

We schematically use notation θ_ℓ to denote both classical and quantum training variables. The sum will lead to a combination of the following three cases: classical contribution, quantum contribution, and the classical-quantum mixed contribution. For pure classical contribution, the formula from Gaussian correlations will lead to $1/$width suppression. For quantum contribution and its mixture with classical ones, the meta-kernel will be naturally suppressed by $1/\dim \mathcal{O}(\mathcal{H})$ for its corresponding operator space dimension $\dim \mathcal{O}(\mathcal{H})$. Otherwise, there are non-negligible terms that give $\mathcal{O}(1)$ modification to NTK during the gradient descent, leading to a non-linear regime of training dynamics and representation learning. Otherwise, the training dynamics will linearize, and we get an exponential convergence of the residual training error. The observations indicate a possible connection between barren plateau, random unitary, and large-width limit in classical neural networks, where we will leave details to our future works.

VI. ON THE LECUN PARAMETRIZATION AND THE NTK PARAMETRIZATION

In this section, we will clarify the issue of the initialization (parametrization) convention we use comparing to different literature. In several original papers about classical NTK [8], there is a convention which is called the NTK parametrization, which defines the Gaussian correlation of weights as

$$\mathbb{E}\left(W_{ij}W_{kl}\right) = \delta_{ik}\delta_{jl}C_W \ . \tag{201}$$

The initialization will lead to rescaling of parameters as the following table,

In our work, we use the standard (LeCun) scaling, where we will still naturally obtain the natural NTK, and the dNTK suppression results [1]. However, there are also studies pointing out alternative parametrizations [9].

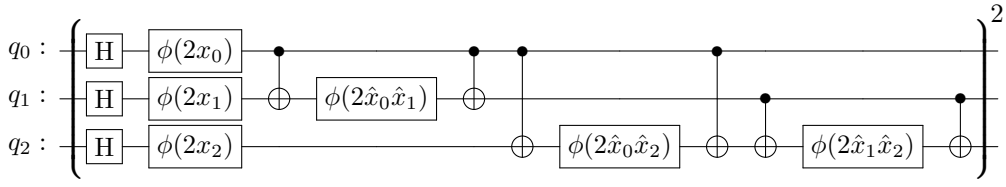
Type	Weight	Single-layer	Learning rate
LeCun	$\mathbb{E}\left(W_{ij}W_{kl}\right) = \delta_{ik}\delta_{jl}C_W/\text{width}$	$Wx+b$	$\eta = \mathcal{O}(1)/\text{width}$
NTK	$\mathbb{E}\left(W_{ij}W_{kl}\right) = \delta_{ik}\delta_{jl}C_W$	$(Wx+b)/\sqrt{\text{width}}$	$\eta = \mathcal{O}(1)$

TABLE I. Comparison between the LeCun (standard) parametrization and the NTK parametrization.

VII. SIMULATION DETAILS AND ADDITIONAL DATA

We simulate the NTK dynamics in a concrete example using the Qiskit python library. In the main text, we consider a binary classification task carried out with a 3-qubit quantum neural network (QNN) in a supervised learning setting, using an ad-hoc data set provided in Qiskit Machine Learning. The input data points have 3-dimensional features, and are partitioned in a training set (20 elements) and a test set (5 elements). Each input $\mathbf{x} = (\texttt{x[0]}, \texttt{x[1]}, \texttt{x[2]})$ is encoded through the `ZZFeatureMap`

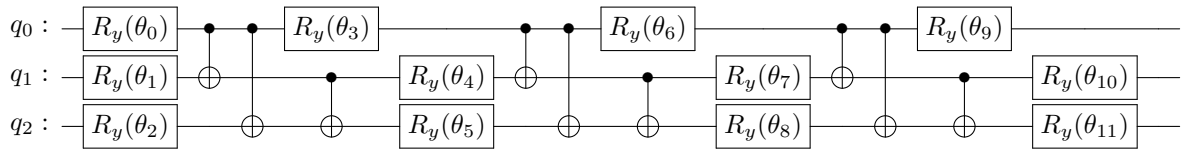
$\texttt{ZZFeatureMap} =$



(202)

where $\phi(\theta)$ is the single-qubit phase gate and $\hat{x}_i = \pi - x_i$. The trainable part of the quantum neural network, which appended to the quantum circuit after the feature map, is represented by a 3-layer `RealAmplitudes` variational ansatz, combining parametrized R_y rotations and entangling CNOT operations:

$\texttt{RealAmplitudes} =$



(203)

The QNN output is obtained as the expectation value of the 3-qubit observable $Z_0Z_1Z_2$, where Z_i is the single qubit Pauli Z operator. Our model is

$$z_\delta = \langle \phi(\mathbf{x}_\delta)| U^\dagger O U |\phi(\mathbf{x}_\delta)\rangle \ , \tag{204}$$

where O is given by

$$O = \text{diag}(1,-1,-1,1,-1,1,1,-1) \ . \tag{205}$$

Note that we only have one observable O, so our quantum neural network only has a scalar as the output. In this setup, the QNTK is given by

$$K_{\delta,\tilde{\alpha}} = \sum_\ell \frac{dz_\delta}{d\theta_\ell}\frac{dz_{\tilde{\alpha}}}{d\theta_\ell} = -\sum_\ell \begin{pmatrix} \left\langle \phi\left(\mathbf{x}_\delta\right)\right| U_{+,\ell}^\dagger \left[X_\ell, U_\ell^\dagger W_\ell^\dagger U_{-,\ell}^\dagger O U_{-,\ell} W_\ell U_\ell\right] U_{+,\ell} \left|\phi\left(\mathbf{x}_\delta\right)\right\rangle \times \\ \left\langle \phi\left(\mathbf{x}_{\tilde{\alpha}}\right)\right| U_{+,\ell}^\dagger \left[X_\ell, U_\ell^\dagger W_\ell^\dagger U_{-,\ell}^\dagger O U_{-,\ell} W_\ell U_\ell\right] U_{+,\ell} \left|\phi\left(\mathbf{x}_{\tilde{\alpha}}\right)\right\rangle \end{pmatrix} , \tag{206}$$

and the frozen QNTK is given by

$$K_{\delta,\tilde{\alpha}} = -\delta^2 \sum_\ell \begin{pmatrix} \left\langle \phi\left(\mathbf{x}_\delta\right)\right| W_{+,\ell}^\dagger \left[X_\ell, W_\ell^\dagger W_{-,\ell}^\dagger O W_{-,\ell} W_\ell\right] W_{+,\ell} \left|\phi\left(\mathbf{x}_\delta\right)\right\rangle \times \\ \left\langle \phi\left(\mathbf{x}_{\tilde{\alpha}}\right)\right| W_{+,\ell}^\dagger \left[X_\ell, W_\ell^\dagger W_{-,\ell}^\dagger O W_{-,\ell} W_\ell\right] W_{+,\ell} \left|\phi\left(\mathbf{x}_{\tilde{\alpha}}\right)\right\rangle \end{pmatrix} , \tag{207}$$

with the help of the initial angle redefinition. All the theories we have discussed could be applied in a straightforward way in our simulation case. For the theoretical prediction, we take the variational angle θ^* at the last step of our training.

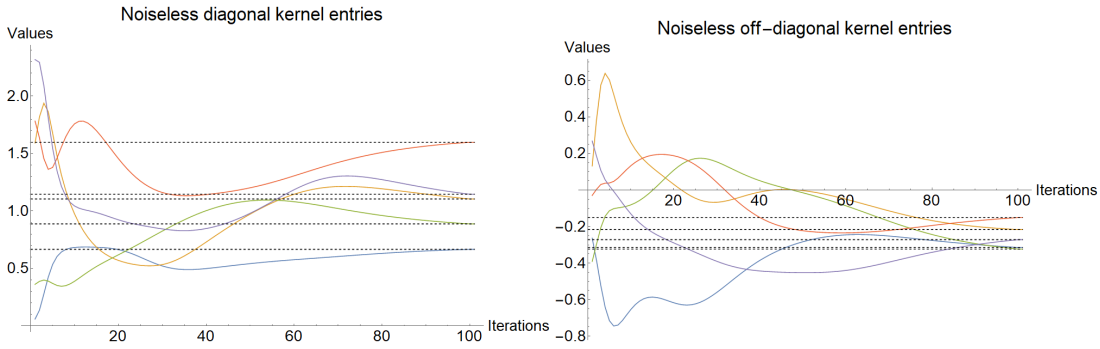


FIG. 1. Noiseless diagonal and off-diagonal entries of QNTK during the noiseless gradient descent dynamics. We compute the gradient descent evolution of five random diagonal and off-diagonal elements in the QNTK. The solid line is the actual value of the QNTK entries during the experiment, and the dashed line is the theoretical prediction of the frozen QNTK.

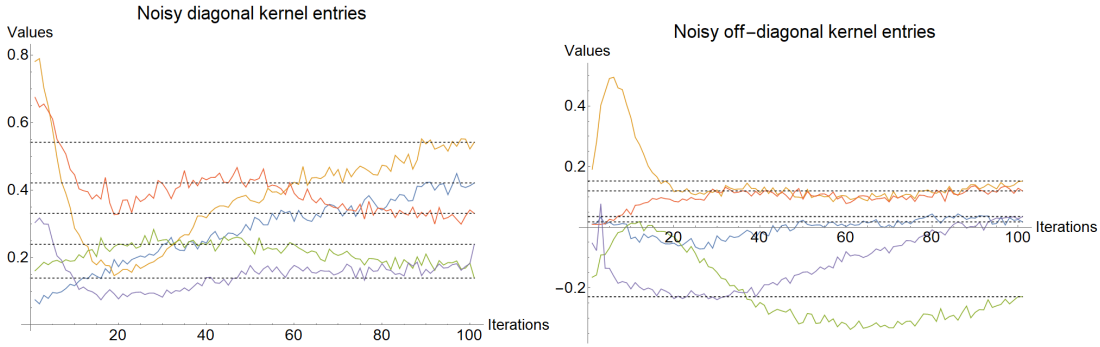


FIG. 2. Noisy diagonal and off-diagonal entries of QNTK during the gradient descent dynamics. We compute the gradient descent evolution of five random diagonal and off-diagonal elements in the QNTK. The solid line is the actual value of the QNTK entries during the simulation of a noisy quantum processor, and the dashed line is the theoretical prediction of the frozen QNTK.

In addition to the plots reported in the main text, here we show more data about the properties of QNTK during the gradient descent. See Fig. 1 and Fig. 2. We could see that, even including a model of device noise, the simulation is successful and the kernel is stable at the late time including the effect of the error mitigation. This will indicate an exponential decay of the residual training error at the late time for noisy quantum circuits.

[1] D. A. Roberts, S. Yaida, and B. Hanin, arXiv preprint arXiv:2106.10165 (2021).
[2] D. A. Roberts and S. Yaida, Deep Learning Theory Summer School at Princeton (2021).
[3] Note that the inverse has to be taken in the over-parametrized regime where the number of variational angles are larger than the training samples to make the formalism simple, which is not the case in our numerical simulation in this paper.
[4] D. Meltzer, H. Zheng, D. Yi-Hsien, D. R. Roberts, and J. Liu, .
[5] M. Mohri, A. Rostamizadeh, and A. Talwalkar, *Foundations of machine learning* (MIT press, 2018).
[6] The convention factor $1/2$ appearing in η is from the worldsheet index defined for holomorphic and anti-holomorphic variables. Similar conventions appear in the two-dimensional conformal field theories, see [10].
[7] L. Zhao, Z. Zhao, P. Rebentrost, and J. Fitzsimons, Quantum Machine Intelligence **3**, 1 (2021).
[8] A. Jacot, F. Gabriel, and C. Hongler, arXiv preprint arXiv:1806.07572 (2018).
[9] J. Sohl-Dickstein, R. Novak, S. S. Schoenholz, and J. Lee, arXiv preprint arXiv:2001.07301 (2020).
[10] J. Polchinski, *String theory: Volume 1, an introduction to the bosonic string* (Cambridge university press, 1998).