Supplemental Material for "Information Scrambling in Quantum Neural Networks"

Huitao Shen,[1] Pengfei Zhang,[2] Yi-Zhuang You,[3] and Hui Zhai[2,*]

[1]*Department of Physics, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA*
[2]*Institute for Advanced Study, Tsinghua University, Beijing, 100084, China*
[3]*Department of Physics, University of California, San Diego, CA 92093, USA*

In this supplemental material, we present more results of magnetization learning, staggered magnetization learning, and winding number learning, along with details of gradient calculation and measurement.

I. MAGNETIZATION LEARNING

In this section, we provide more details of magnetization learning and present an argument on why in magnetization learning, long string operators should exist in $\hat{U}^\dagger \sigma^x_{(n+1)/2} \hat{U}$ when it is expanded under the basis of product of local Pauli matrices.

A. Learning Task Details

Figure 1 shows the distribution of magnetization M_z^α in the training and validation datasets. The magnetization distributions within the training and validation set are similar. There are roughly equal number of wavefuntions that are "ferromagnetic" ($|M_z^\alpha| \geq 0.5$) or "paramagnetic" ($|M_z^\alpha| < 0.5$).

For the AMSGrad algorithm [1], momentum parameters are always $\beta_1 = 0.9$ and $\beta_2 = 0.999$ throughout this work. Becasue the training set is not very big, we use gradient descent instead of stochastic or mini-batch gradient descent. In other words, each epoch involves only one gradient descent step.

We confirm that validation losses also decrease monotonically when the training proceeds (not shown here), indicating that the network can learn to compute the magnetization reasonably well without overfitting.

In Fig. 2, we show the training loss and the tripartite information as functions of the training epoch. We plot both the averaged values over 20 different random initializations and two typical initializations. Both the averaged value and the two training instances show two-stage training dynamics. In particular, Fig. 3(a) and Fig. 4 in the main text use the same initialization as Initialization 2 here.

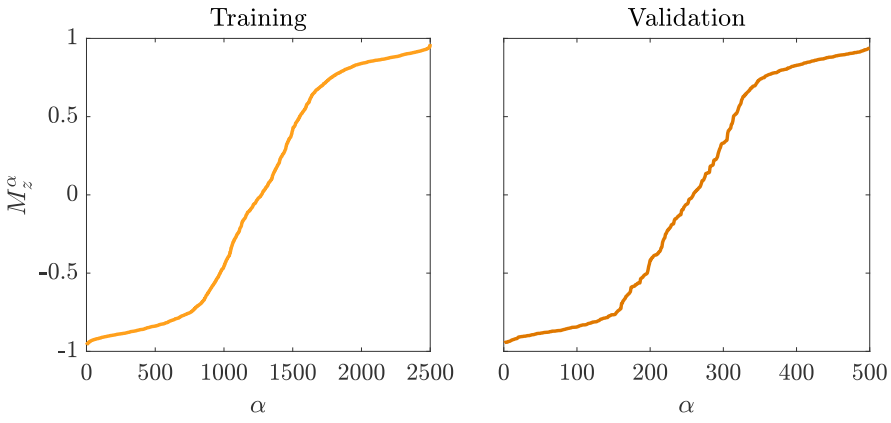

FIG. 1. Distribution of magnetization M_z^α in the training and validation sets. The training and validation dataset contains $N = 2500$ and 500 wavefunction–magnetization pairs respectively, sampled from random Hamiltonian ensemble of system size $n = 9$, where random parameters are distributed uniformly within $J_{ij}/J \in [-1, 0]$, $K_{ij}/J \in [-1, 1]$, $g_i/J \in [-6, 6]$ and $h/J \in [-0.04, 0.04]$. J is the energy unit.

* hzhai@tsinghua.edu.cn

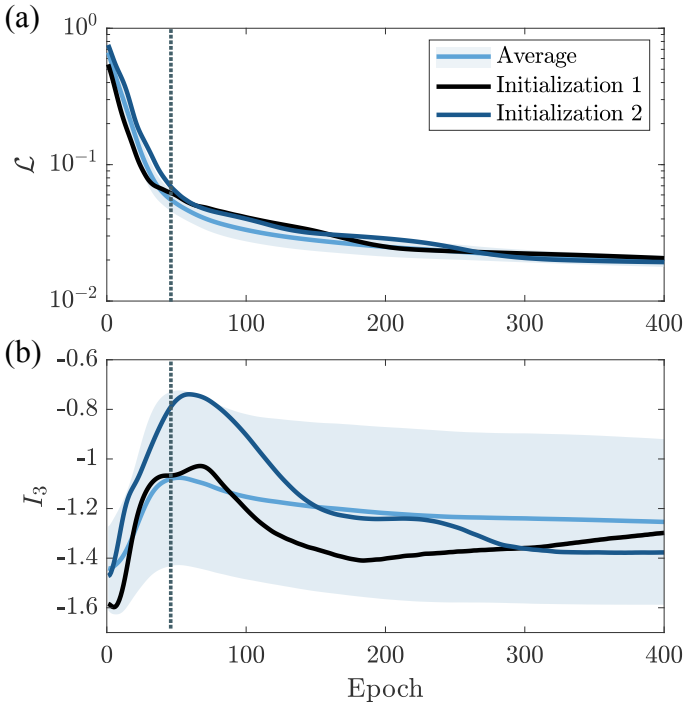

FIG. 2. Magnetization learning. (a) Training loss as functions of the training epoch. Different colors represent the average over 20 different random initializations or typical results from two training instances. The shaded area represents one standard deviation. The network has $n = 9$ qubits and depth $l = 6$. The learning rate is $\lambda = 10^{-2}$. (b) Tripartite information $I_3(A, C, D)$ as a function of the training epoch. Here the input subsystem size $|C| = 5$. The dotted vertical line indicates the boundary between two training stages, which is determined as the local maximum of the averaged I_3.

B. Explicit Construction of Unitary that Learns Magnetization

Generally, it is impossible to find an unitary $\hat{U}$ such that

$$\hat{U}^\dagger \sigma^x_{(n+1)/2} \hat{U} = \hat{M}_z, \tag{1}$$

because the L.H.S. and the R.H.S. of the above equality have different eigenvalues. As a result, we can only expect the above equality to hold at the level of expectation

$$\langle\psi|\hat{U}^\dagger \sigma^x_{(n+1)/2} \hat{U}|\psi\rangle = \langle\psi|\hat{M}_z|\psi\rangle\,, \tag{2}$$

within a subset of states $\{|\psi\rangle\}$ that are of interest [1]. In the following, we present an explicit construction of $\hat{U}^\dagger \sigma^x_{(n+1)/2} \hat{U}$ for the magnetization learning problem when the subset of states are eigenstates of $\hat{M}_z \equiv \sum_{i=1}^n \sigma_i^z/n$. The purpose of this construction is to use an explicit example to demonstrate why it is usually necessary to have string operators in $\hat{U}^\dagger \sigma^x_{(n+1)/2} \hat{U}$ for a quantum neural netowrk (NN) that learns magnetization.

We first elaborate the rationale behind choosing eigenstates of $\hat{M}_z$. For magnetization learning, the dataset consists of ground states of Hamiltonians (Eq. (2) in the main text) with a small pinning field h to trigger spontaneous Z_2 symmetry breaking in the finite-size numerical simulation. The spin-spin interaction is also chosen to be nonlocal to ensure that we have sufficient number of distinct states. To actually probe the physics of Z_2 symmetry breaking in one dimension, we should take the thermodynamics limit $n \to \infty$ while sending $h \to 0$ and fixing the spin-spin interaction range. It is well-known that in such systems, the ordered ferromagnetic ground state is gapped. Consequently, the quantum fluctuation of $\hat{M}_z$ is

$$\left\langle \left(\hat{M}_z - \left\langle \hat{M}_z \right\rangle\right)^2 \right\rangle = \sum_{i,j=1}^n \frac{1}{n^2} \left\langle \delta\sigma_i^z \delta\sigma_j^z \right\rangle = \sum_{i=1}^n \frac{1}{n} \left\langle \delta\sigma_i^z \delta\sigma_1^z \right\rangle \sim \frac{1}{n}, \tag{3}$$

[1] Note that the subset is in general not a subspace as linear combinations in general break the equality Eq. (2).

because $\langle \delta\sigma_i^z \delta\sigma_1^z \rangle$ decays exponentially with i. Therefore, the fluctuation of $\hat{M}_z$ is suppressed, and ground states of our random Hamiltonian can be well approximated by eigenstates of $\hat{M}_z$ in the thermodynamic limit.

We are now ready to present our construction. Denote the eigenstates of $\hat{M}_z$ as $|m, i\rangle$ such that $\hat{M}_z |m, i\rangle = m |m, i\rangle$. Here $m \in [-1, 1]$ is the eigenvalue, which is also the average magnetization. $i = 1, \ldots, d_m$ represents the state in the degenerate eigenspace and d_m is the degeneracy. The states are orthonormal $\langle m, i | m', i' \rangle = \delta_{mm'} \delta_{ii'}$ and complete $\sum_m d_m = 2^n$. Because of the spin-flip symmetry, $d_m = d_{-m}$. In general $d_m > 1$ unless $m = \pm 1$, where all spins are polarized to the same direction. For degenerate subspaces, note that the choice of $|m, i\rangle$ for fixed m but different i is not unique.

In the following, we construct matrix elements of $\hat{U}^\dagger \sigma^x_{(n+1)/2} \hat{U}$ under $|m, i\rangle$ basis such that

$$\langle m, i | \hat{U}^\dagger \sigma^x_{(n+1)/2} \hat{U} | m, i \rangle = m, \tag{4}$$

for all m and i. Consider the two-dimensional subspace spanned by $|m, i\rangle$ and $|-m, i\rangle$ for all m and i. Within this subspace, we set

$$\hat{U}^\dagger \sigma^x_{(n+1)/2} \hat{U} = \sin\theta \sigma^x + \cos\theta \sigma^z, \tag{5}$$

where $\theta = \arccos m$. It is straightforward to verify the constraint Eq. (4) is satisfied and the eigenvalues are ± 1. Under this construction, half of $\hat{U}^\dagger \sigma^x_{(n+1)/2} \hat{U}$'s eigenvalues are $+1$ and half are -1. It is then not hard to see that there must exist some $\hat{U}$ such that $\hat{U}^\dagger \sigma^x_{(n+1)/2} \hat{U}$ has the matrix elements under $|m, i\rangle$ basis as constructed.

Although the above matrix is constructed explicitly on a particular choice of basis, it is straightforward to verify that the following basis-independent constraint holds

$$\langle m | \hat{U}^\dagger \sigma^x_{(n+1)/2} \hat{U} | m \rangle = m, \tag{6}$$

where $|m\rangle \equiv \sum_{i=1}^{d_m} c_i |m, i\rangle$ is any linear combination of eigenstates within the same degenerate eigenspace. $\sum_{i=1}^{d_m} |c_i|^2 = 1$.

Because the choice of basis within a degenerate subspace is not unique, our constructions above are not unique either. Nevertheless, generally $|m, i\rangle$ and $|-m, i\rangle$ are related to each other by a string of local Pauli matrices whose length is of order of system size n. A particular choice is that $|-m, i\rangle = \prod_{j=1}^n \sigma_i^x |m, i\rangle$ such that the two states are related by a global spin-flip operator, which is a string operator of length n. Because of Eq. (5), such string operator must exist in $\hat{U}^\dagger \sigma^x_{(n+1)/2} \hat{U}$.

II. STAGGERED MAGNETIZATION LEARNING

In this section, we present results of staggered magnetization learning task, where empirical correlation between the NN performance and the tripartite information is also found.

Dataset. Similar to magnetization learning, the dataset consists of N input-target pairs $\{(|G^\alpha\rangle, \overline{M}_z^\alpha), \alpha = 1, \ldots, N\}$, where the input wavefunction $|G^\alpha\rangle$ is the ground state wavefunction of the parent Hamiltonian with random long-ranged spin-spin interactions:

$$\hat{H} = \sum_{i=1}^{n-1} J_{i,i+1} \sigma_i^z \sigma_{i+1}^z + \sum_{i,j=1}^{n} K_{ij} \sigma_i^x \sigma_j^x + \sum_{i=1}^{n} (g_i \sigma_i^x + h \sigma_i^z), \tag{7}$$

where $J_{i,i+1}$, K_{ij}, g_i and h are all random numbers. Compared with Eq. (2) in the main text, here only nearest-neighbour $\sigma_i^z \sigma_{i+1}^z$ interactions are included to avoid frustration.

The target is the average staggered magnetization computed as $\overline{M}_z^\alpha \equiv \langle G^\alpha | \hat{\overline{M}}_z | G^\alpha \rangle$, where the staggered magnetization operator is $\hat{\overline{M}}_z \equiv \sum_{i=1}^n (-1)^n \sigma_i^z / n$. In sampling the random Hamiltonian, we ensure $J_{ij} \geq 0$ such that the ground state wavefunctions are either "antiferromagnetic" or "paramagnetic" measured under $\hat{\overline{M}}_z$. h is a small pinning field randomly drawn from a distribution with zero mean, which is used to trigger the spontaneous Z_2 symmetry breaking in the antiferromagnetic phase.

Task. The quantum NN takes the input wavefunction $|G^\alpha\rangle$ and applies the unitary transformation $\hat{U}$ on it. The staggered magnetization is readout by measuring σ^x of the central qubit. The loss function to be minimized is the absolute error of the staggered magnetization:

$$\mathcal{L} = \frac{1}{N} \sum_{\alpha=1}^{N} \left| \langle G^\alpha | \hat{U}^\dagger \sigma^x_{(n+1)/2} \hat{U} | G^\alpha \rangle - \overline{M}_z^\alpha \right|. \tag{8}$$

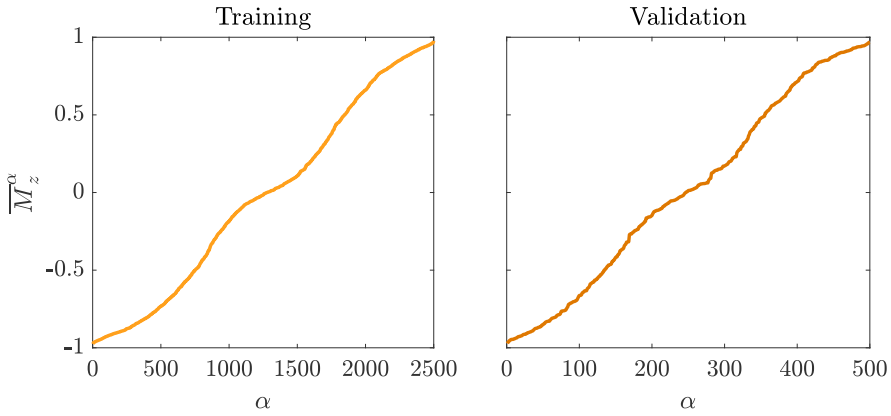

FIG. 3. Distribution of staggered magnetization $\overline{M}_z^\alpha$ in the training and validation sets. The training and validation dataset contains $N = 2500$ and 500 wavefunction–staggered magnetization pairs respectively, sampled from random Hamiltonian ensemble of system size $n = 9$, where random parameters are distributed uniformly within $J_{i,i+1}/J \in [0, 8]$, $K_{ij}/J \in [-1, 1]$, $g_i/J \in [-2, 2]$ and $h/J \in [-0.04, 0.04]$. J is the energy unit.

In Fig. 3, we show the distribution of staggered magnetization $\overline{M}^\alpha$ in the training and validation datasets. The magnetization distributions within the training and validation set are similar. There are roughly equal number of wavefuntions that are "antiferromagnetic" ($|\overline{M}_z^\alpha| \geq 0.5$) or "paramagnetic" ($|\overline{M}_z^\alpha| < 0.5$).

Results. Figure 4 is the training loss and tripartite information during quantum NN training for the staggered magnetization learning task. We confirm the validation loss is similar to that in the training set.

The figure looks almost identical to Fig. 2, despite that we now have a different dataset. The two-stage training dynamics, i.e., an early stage with rapid decrease of loss and increase of tripartite information, followed by a later stage with slow decrease of both loss and tripartite information, can be clearly observed. One can also see the initial rapid linear growth of tripartite information in Fig. 4 with almost identical slopes for both two training instances and the averaged result.

Finally, we have also tried Hamiltonians similar to Eq. (7) but with longer interaction range such that the ground state is frustrated. The quantum NN shows similar performance and training dynamics.

III. WINDING NUMBER LEARNING

In this section we present the results of winding number learning task, which again reinforces the generality of two-stage training dynamics of quantum NNs.

Dataset. The input data consist of N product states of n qubits, where each qubit represents a vector on the xz plane of the Bloch sphere. The target is the winding number of these vectors by treating the n qubits as vectors on an one-dimensional Brillouin zone [2]. Formally, the dataset consists of N input-target pairs $\{(|H^\alpha\rangle, w^\alpha), \alpha = 1, \ldots, N\}$, where the input wavefunction $|H^\alpha\rangle = \prod_{i=1}^n |\psi^\alpha(k_i)\rangle$, $k_i = 2\pi(i-1)/(n-1)$, and $\psi^\alpha(k)$ is the ground state of the following random two-band Hamiltonian in one-dimensional Brillouin zone $k \in [0, 2\pi)$ with chiral symmetry $\sigma_y H(k)\sigma_y = -H(k)$:

$$H(k) = h_x(k)\sigma^x + h_z(k)\sigma^z. \tag{9}$$

Here the coefficient $h_\mu(k)$, $\mu = x, z$ is represented in terms of Fourier components up to p-th harmonic:

$$h_\mu(k) = \sum_{n=0}^{p} \cos(nk)c_n^\mu + \sum_{n=1}^{p} \sin(nk)s_n^\mu, \tag{10}$$

where c_n^μ and s_n^μ are random numbers.

The learning target is the discrete version of winding number:

$$w^\alpha = \frac{1}{2\pi}\sum_{i=1}^{n} \mathrm{Im}\ln\left[e^{i(\phi^\alpha(k_i)-\phi^\alpha(k_{i+1}))}\right], \tag{11}$$

where $\phi^\alpha(k)$ is defined as the argument of the following complex number:

$$e^{i\phi^\alpha(k)} = \frac{h_z^\alpha(k) + ih_x^\alpha(k)}{\sqrt{h_z^\alpha(k)^2 + h_x^\alpha(k)^2}}. \tag{12}$$

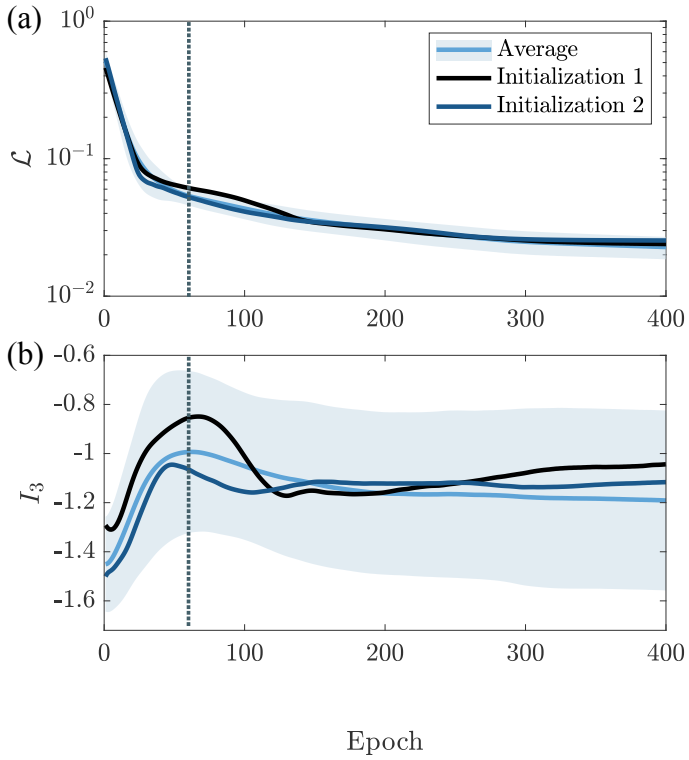

FIG. 4. Staggered magnetization learning. (a) Training loss as functions of the training epoch. Different colors represent the average over 20 different random initializations or typical results from two training instances. The shaded area represents one standard deviation. The network has $n = 9$ qubits and depth $l = 6$. The learning rate is $\lambda = 10^{-2}$. (b) Tripartite information $I_3(A, C, D)$ as a function of the training epoch. Here the input subsystem size $|C| = 5$. The dotted vertical line indicates the boundary between two training stages, which is determined as the local maximum of the averaged I_3.

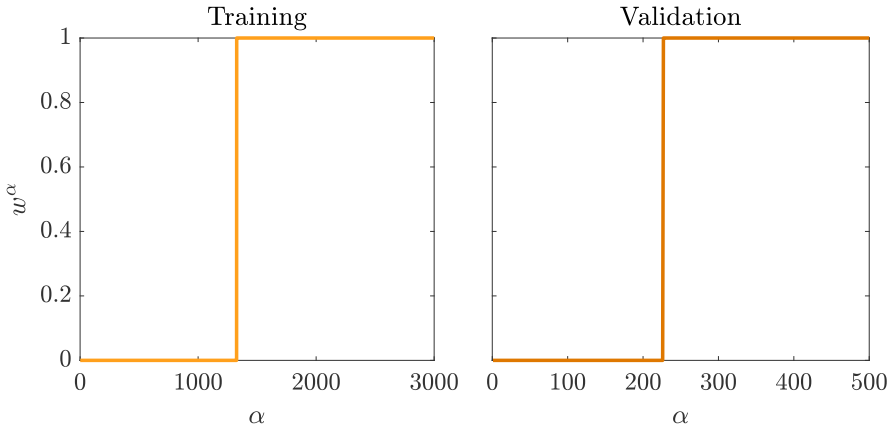

FIG. 5. Distribution of winding number w^α in the training and validation sets.

The branch cut for the logarithm in Eq. (11) is along the negative direction of the x axis such that $\phi(k) - \phi(k') \in [-\pi, \pi)$.

Task. In the following, we set the harmonic cutoff $p = 1$. c_n^μ and s_n^μ are sampled from a uniform distribution between $[-1/3, 1/3]$ for $n = 0$ and $[-1, 1]$ for $n > 0$. We then post-select data with winding number $w = 0, 1$ and discard those with $w = -1$. In this way, the task becomes binary classification. The parameters are chosen such that there are roughly equal number of data with $w = 0$ and 1, as shown in Fig. 5.

The quantum NN takes the input wavefunction $|H^\alpha\rangle$ and applies the unitary transformation $\hat{U}$ on it. The probability that the

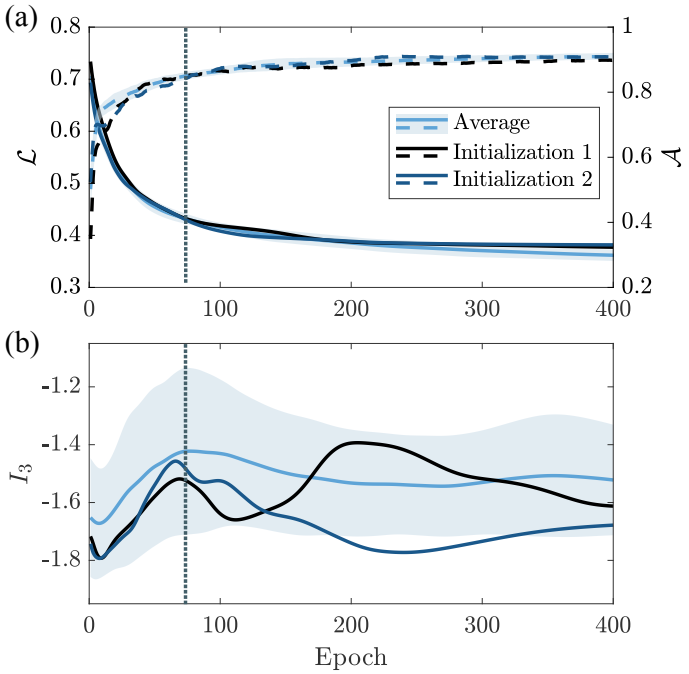

FIG. 6. Winding number learning. (a) Training loss (solid, left) and accuracy (dashed, right) as functions of the training epoch. Different colors represent the average over 20 different random initializations or typical results from two training instances. The shaded area represents one standard deviation. The network has $n = 9$ qubits and depth $l = 8$. The training and validation dataset contains $N = 3000$ and 500 wavefunction-winding number pairs respectively, sampled from random wavefunctions defined in the main text. The learning rate is $\lambda = 10^{-2}$. (b) Tripartite information $I_3(A, C, D)$ as a function of the training epoch for different initializations. Here the input subsystem size $|C| = 5$. The dotted vertical line indicates the boundary between two training stages, which is determined as the local maximum of the averaged I_3.

$w^\alpha = 1$ is readout by measuring σ^x of the central qubit:

$$p^\alpha = \frac{1 + \langle H^\alpha | \hat{U}^\dagger \sigma^x_{(n+1)/2} \hat{U} | H^\alpha \rangle}{2}. \tag{13}$$

Therefore, the loss function to be minimized is the negative binary cross-entropy:

$$\mathcal{L} = \frac{1}{N} \sum_{\alpha=1}^{N} \left[-w^\alpha \ln p^\alpha - (1 - w^\alpha) \ln(1 - p^\alpha) \right]. \tag{14}$$

A more sensible metric is the prediction accuracy. Let the prediction of the winding number be $o^\alpha \equiv (1 + \mathrm{sgn}(p_a - 1/2))/2$. The prediction accuracy is then

$$\mathcal{A} \equiv 1 - \frac{1}{N} \sum_{\alpha=1}^{N} |o^\alpha - w^\alpha|. \tag{15}$$

Results. In Fig. 6, we present the training loss and accuracy for the winding number learning task, along with the tripartite information. We confirm the validation loss and accuracy is similar to that in the training set. The network depth l is larger than that in the magnetization learning as we suspect the winding learning task is more difficult. However, using a shallower network will not affect the performance significantly. Because of the difficulty of this task, not all initializations can lead to high accuracies after 400 epochs. In computing the average, we post-select 20 different initializations with smallest training losses out of 50 initializations.

First, the quantum NN manages to learn distinguish wavefunctions with winding number $w = 0$ and 1, as the final accuracy is more than 90%. Second, the trend of the loss function and the tripartite information is similar to that in (staggered) magnetization learning: At the early stage of the training, the loss decreases rapidly and the tripartite information increases. In the later stage, the tripartite information decreases again. The trend is robust when different initializations are averaged. However, we note the tripartite information is slightly more volatile in the later stage than that in the (staggered) magnetization learning, which is

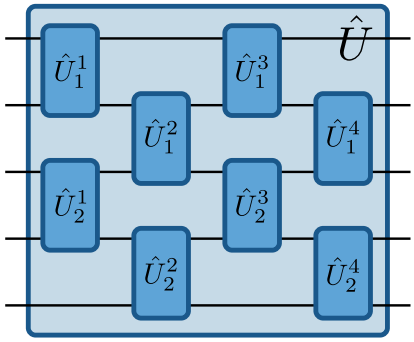

FIG. 7. Schematic of a quantum circuit with brick-wall geometry. Here the network has $n = 5$ qubits and depth $l = 4$. All these two-qubit gates form a giant unitary transformation $\hat{U}$. The i-th two-qubit gate in the d-th layer is denoted as $\hat{U}_i^d$

reflected by a second local maximum of the averaged I_3 around 350 epochs in Fig. 6. Because this behavior does not appear in other tasks, we believe it is not as universal and leave the in-depth understanding of this behavior for future research.

Compared with the (staggered) magnetization task, the input wavefunction here is a product state and is essentially classical, and the target is now a binary label instead of a real number. Despite the very different nature of this task, the empirical correlation between the NN performance and the tripartite information still holds. This suggests the generality of the two-stage training dynamics of quantum NNs.

IV. GRADIENTS IN QUANTUM NNS

In this section, we report the method of computing gradients of quantum NNs in this work.

A. In Classical Simulations

A schematic of the quantum NN with $n = 5$ qubits and depth $l = 4$ is shown in Fig. 7. The i-th two-qubit gate in the d-th layer is denoted as $\hat{U}_i^d$. Assuming n is odd, here $i = 1, 2 \dots (n-1)/2$. It follows the giant unitary $\hat{U}$ is the composition of $\hat{U}_i^d$:

$$\hat{U} = \left(\prod_{i=1}^{(n-1)/2} \hat{U}_i^l\right) \dots \left(\prod_{i=1}^{(n-1)/2} \hat{U}_i^2\right) \left(\prod_{i=1}^{(n-1)/2} \hat{U}_i^1\right) \equiv \prod_{d=1}^{l} \left(\prod_{i=1}^{(n-1)/2} \hat{U}_i^d\right). \tag{16}$$

The order of unitaries within a layer does not matter because these unitaries are applied on non-overlapping qubits.

In general, each two-qubit gate $\hat{U}_i^d$ is a 4×4 matrix in the $\mathrm{SU}(4)$ group and can be parametrized by 15 parameters. However, as explained in the main text, in this work we restrict $\hat{U}_i^d$ to $\mathrm{SO}(4)$ with 6 Euler angles: Generally, a matrix in $\mathrm{SO}(4)$ can be parametrized by a vector $\boldsymbol{\theta}$ with 6 components [3]:

$$\hat{U}_{\mathrm{SO}(4)} = O_{34}(\theta_1) O_{23}(\theta_2) O_{12}(\theta_3) O_{34}(\theta_4) O_{23}(\theta_5) O_{34}(\theta_6). \tag{17}$$

Here $O_{ij}(\theta) \equiv \exp(\theta J_{ij})$ is a rotation in the ij plane: J_{ij} an antisymmetric matrix with ij (ji) element equal to 1 (-1) and all other elements zero. As a result there are $l(n-1)/2$ independent vectors $\boldsymbol{\theta}_i^d$ and thus $6l(n-1)/2$ independent parameters in total to fully describe the quantum NN.

To be concrete, in the following, we use magnetization learning as the example. The staggered magnetization learning and winding number learning are similar. The loss function in magnetization learning is

$$\mathcal{L} = \frac{1}{N} \sum_{\alpha=1}^{N} \left| \langle G^\alpha | \hat{U}^\dagger \sigma^x_{(n+1)/2} \hat{U} | G^\alpha \rangle - M_z^\alpha \right|. \tag{18}$$

The gradient of $\mathcal{L}$ with respect to $\theta_{j,a}^d$, $a = 1, \dots, 6$ is

$$\frac{\partial \mathcal{L}}{\partial \theta_{j,a}^d} = \frac{1}{N} \sum_{\alpha=1}^{N} \mathrm{sgn}\left(\langle G^\alpha | \hat{U}^\dagger \sigma^x_{(n+1)/2} \hat{U} | G^\alpha \rangle - M_z^\alpha \right) \frac{\partial \langle G^\alpha | \hat{U}^\dagger \sigma^x_{(n+1)/2} \hat{U} | G^\alpha \rangle}{\partial \theta_{j,a}^d}. \tag{19}$$

The gradient of the network output can be further simplified as

$$\begin{aligned}
&\frac{\partial}{\partial\theta_{j,a}^{d}}\left\langle G^{\alpha}\right|\hat{U}^{\dagger}\sigma_{(n+1)/2}^{x}\hat{U}\left|G^{\alpha}\right\rangle \\
&=\left\langle G^{\alpha}\right|\hat{U}^{\dagger}\sigma_{(n+1)/2}^{x}\frac{\partial\hat{U}}{\partial\theta_{j,a}^{d}}\left|G^{\alpha}\right\rangle+\text{h.c.} \\
&=\left\langle G^{\alpha}\right|\hat{U}^{\dagger}\sigma_{(n+1)/2}^{x}\frac{\partial}{\partial\theta_{j,a}^{d}}\left[\prod_{d'=1}^{l}\left(\prod_{i=1}^{(n-1)/2}\hat{U}_{i}^{d'}\right)\right]\left|G^{\alpha}\right\rangle+\text{h.c.} \\
&=\left\langle G^{\alpha}\right|\hat{U}^{\dagger}\sigma_{(n+1)/2}^{x}\left(\prod_{i=1}^{(n-1)/2}\hat{U}_{i}^{l}\right)\ldots\left(\hat{U}_{1}^{d}\hat{U}_{2}^{d}\ldots\frac{\partial\hat{U}_{j}^{d}}{\partial\theta_{j,a}^{d}}\ldots\hat{U}_{\frac{n-1}{2}}^{d}\right)\ldots\left(\prod_{i=1}^{(n-1)/2}\hat{U}_{i}^{1}\right)\left|G^{\alpha}\right\rangle+\text{h.c.},
\end{aligned}\tag{20}$$

where, $\partial\hat{U}_{j}^{d}/\partial\theta_{j,a}^{d}$ can be further simplified using Eq. (17). For example,

$$\frac{\partial\hat{U}_{j}^{d}}{\partial\theta_{j,4}^{d}}=O_{34}(\theta_{j,1}^{d})O_{23}(\theta_{j,2}^{d})O_{12}(\theta_{j,3}^{d})J_{34}O_{34}(\theta_{j,4}^{d})O_{23}(\theta_{j,5}^{d})O_{34}(\theta_{j,6}^{d}).\tag{21}$$

Gradients with respect to other components a can be computed in the similar way by adding an additional corresponding J matrices.

In this work, we directly compute the gradient according to Eqs. (19), (20) and (21) in the classical simulation.

B. In Real Quantum NNs

In a real quantum NN, this gradient could instead be determined through the measurement of the following Hermitian operator:

$$\hat{g}_{j,a}^{d}=\sigma_{(n+1)/2}^{x}\frac{\partial\hat{U}}{\partial\theta_{j,a}^{d}}\hat{U}^{\dagger}+\text{h.c.}.\tag{22}$$

It is straightforward to see that

$$\left\langle G^{\alpha}\right|\hat{U}^{\dagger}\hat{g}_{j,a}^{d}U\left|G^{\alpha}\right\rangle=\frac{\partial}{\partial\theta_{j,a}^{d}}\left\langle G^{\alpha}\right|\hat{U}^{\dagger}\sigma_{(n+1)/2}^{x}\hat{U}\left|G^{\alpha}\right\rangle.$$

However, this operator is generally non-local and is hard to measure.

Alternatively, one could perform the following three measurements [4, 5]:

1. Measure the output of the quantum NN normally with the original parameter $\boldsymbol{\theta}_{i}^{d}$. The result is denoted as $o_1\equiv\left\langle G^{\alpha}\right|\hat{U}^{\dagger}\sigma_{(n+1)/2}^{x}\hat{U}\left|G^{\alpha}\right\rangle$;
2. Measure the output of the quantum NN with $\theta_{j,a}^{d}$ replaced by $\theta_{j,a}^{d}+\pi/4$. The result is denoted as o_2;
3. Measure the output of the quantum NN with $\theta_{j,a}^{d}$ replaced by $\theta_{j,a}^{d}+\pi/2$. The result is denoted as o_3.

It follows the desired gradient is

$$\frac{\partial}{\partial\theta_{j,a}^{d}}\left\langle G^{\alpha}\right|\hat{U}^{\dagger}\sigma_{(n+1)/2}^{x}\hat{U}\left|G^{\alpha}\right\rangle=2o_2-o_1-o_3.\tag{23}$$

The reason is that if we focus on some specific $\theta_{j,a}^{d}$, we have

$$o_1=\left\langle\ldots O_{p,p+1}^{\dagger}(\theta_{j,a}^{d})\ldots O_{p,p+1}(\theta_{j,a}^{d})\ldots\right\rangle,\tag{24}$$

$$\begin{aligned}
o_2&=\left\langle\ldots O_{p,p+1}^{\dagger}(\theta_{j,a}^{d}+\pi/4)\ldots O_{p,p+1}(\theta_{j,a}^{d}+\pi/4)\ldots\right\rangle \\
&=\left\langle\ldots\left[(1+J_{p,p+1})O_{p,p+1}(\theta_{j,a}^{d})\right]^{\dagger}\ldots(1+J_{p,p+1})O_{p,p+1}(\theta_{j,a}^{d})\ldots\right\rangle/2,
\end{aligned}\tag{25}$$

$$\begin{aligned}
o_3&=\left\langle\ldots O_{p,p+1}^{\dagger}(\theta_{j,a}^{d}+\pi/2)\ldots O_{p,p+1}(\theta_{j,a}^{d}+\pi/2)\ldots\right\rangle \\
&=\left\langle\ldots\left[J_{p,p+1}O_{p,p+1}(\theta_{j,a}^{d})\right]^{\dagger}\ldots J_{p,p+1}O_{p,p+1}(\theta_{j,a}^{d})\ldots\right\rangle.
\end{aligned}\tag{26}$$

Here $p(p+1)$ is the rotation plane associated with a. As a result:

$$2o_2 - o_1 - o_3 = \left\langle \dots O^{\dagger}_{p,p+1}(\theta^d_{j,a}) \dots J_{p,p+1} O_{p,p+1}(\theta^d_{j,a}) \dots \right\rangle + \text{h.c.} = \frac{\partial}{\partial \theta^d_{j,a}} \langle G^{\alpha}| \hat{U}^{\dagger} \sigma^x_{(n+1)/2} \hat{U} |G^{\alpha}\rangle \,. \tag{27}$$

The above method can be easily generalized to $\mathrm{SU}(4)$ as well.

---

[1] Sashank J. Reddi, Satyen Kale, and Sanjiv Kumar, "On the Convergence of Adam and Beyond," in *International Conference on Learning Representations* (2018).

[2] Pengfei Zhang, Huitao Shen, and Hui Zhai, "Machine Learning Topological Invariants with Neural Networks," Phys. Rev. Lett. **120**, 066401 (2018).

[3] P Dita, "Factorization of unitary matrices," J. Phys. A **36**, 2781–2789 (2003).

[4] K. Mitarai, M. Negoro, M. Kitagawa, and K. Fujii, "Quantum circuit learning," Phys. Rev. A **98**, 032309 (2018).

[5] Maria Schuld, Alex Bocharov, Krysta M. Svore, and Nathan Wiebe, "Circuit-centric quantum classifiers," Phys. Rev. A **101**, 032308 (2020).