Expressivity of quantum neural networks

Yadong Wu,[1] Juan Yao,[2] Pengfei Zhang,[3,4,*] and Hui Zhai [1,†]

[1]*Institute for Advanced Study, Tsinghua University, Beijing, 100084, China*

[2]*Guangdong Provincial Key Laboratory of Quantum Science and Engineering, Shenzhen Institute for Quantum Science and Engineering, Southern University of Science and Technology, Shenzhen 518055, Guangdong, China*

[3]*Institute for Quantum Information and Matter, California Institute of Technology, Pasadena, California 91125, USA*

[4]*Walter Burke Institute for Theoretical Physics, California Institute of Technology, Pasadena, California 91125, USA*

In this work, we address the question whether a sufficiently deep quantum neural network can approximate a target function as accurate as possible. We start with typical physical situations that the target functions are physical observables, and then we extend our discussion to situations that the learning targets are not directly physical observables, but can be expressed as physical observables in an enlarged Hilbert space with multiple replicas, such as the Loschmidt echo and the Rényi entropy. The main finding is that an accurate approximation is possible only when all the input wave functions in the dataset do not span the entire Hilbert space that the quantum circuit acts on, and more precisely, the Hilbert space dimension of the former has to be less than half of the Hilbert space dimension of the latter. In some cases, this requirement can be satisfied automatically because of the intrinsic properties of the dataset, for instance, when the input wave function has to be symmetric between different replicas. And if this requirement cannot be satisfied by the dataset, we show that the expressivity capabilities can be restored by adding one ancillary qubit at which the wave function is always fixed at input. Our studies point toward establishing a quantum neural network analogy of the universal approximation theorem that lays the foundation for expressivity of classical neural networks.

I. INTRODUCTION

Neural networks lie at the center of the recent third wave of artificial intelligence. The universal approximation theorem plays an essential role in the development of neural networks, which states that sufficiently wide or sufficiently deep neural networks can approximate a well-behaved function on d-dimensional Euclidean space $\mathbf{R}^d$ with arbitrary accuracy [1]. This theorem lays the foundation of the expressive capability of neural networks and serves as bases for the successes of neural network applications. Quantum neural networks (QNNs) are quantum generalizations of classical feedforward neural networks on future quantum computers [2–5], which lie at the center of the recent development of the quantum machine learning, including quantum unsupervised learning [6–11], quantum generalization of neural networks [12–21], quantum circuit structures developed by classical neural networks [22–29], and information theory in quantum neural networks [30,31]. However, expressivity of QNNs has not been fully explored and there are only few works in this direction [13,32].

Here we consider the quantum generalizations of fully connected neural networks, which contain quantum wave functions of n-qubit states as inputs, parameterized quantum circuits made of local quantum gates, and measurements on readout qubits leading to labels. The parameters in the quantum circuit will be optimized during training that yields the best approximation of the learning target. Since the Hilbert space dimension of an n-qubit state is 2^n, the wave function can encode information of $d = 2^n$ complex numbers, up to a normalization condition and a global phase. In the following, concerning the effects of the depth and the width on the expressive capability of a QNN, we address the question whether a sufficiently deep QNN and a sufficiently wide QNN can express any well-behaved function in the $\mathbf{C}^d$ space.

To be concrete, here we consider a number of typical learning tasks in the quantum physics problem, where the learning targets include (i) physical observables; (ii) the Loschmidt echo, and (iii) the Rényi entropy. We point out that, in contrast to the universal approximation theorem for classical neural networks, a QNN *cannot* express a general well-behaved function with arbitrary accuracy even though the QNN is made sufficiently deep. However, we show in this work that, by enlarging the Hilbert space dimension of the input state, this problem can be solved, and the expressivity can be significantly improved or can even be made as accurate as possible. Enlarging the Hilbert space dimension effectively increases the width of the QNN. This can be achieved either by adding

*pengfeizhang.physics@gmail.com
†hzhai@tsinghua.edu.cn

an ancillary qubit in the input and (or) by duplicating replicas of the input wave functions. These results point toward an analogy of the universal approximation theorem for QNN.

II. RESULTS

We consider a dataset denoted as $\{(|\psi^l\rangle, y^l)\}$, $\{l = 1, \ldots, N_D\}$, where l labels data and N_D is the total number of data in the dataset. Each input quantum state $|\psi^l\rangle$ can be written as $|\psi^l\rangle = \sum_m c_m^l |m\rangle$, where $\{|m\rangle\}$ is a complete set of 2^n bases of the n-qubit Hilbert space, and $\{c_m^l\}$ are 2^n normalized complex numbers with a fixed total phase. Usually the information in the label is much more condensed than the information of the entire input, therefore, here we consider that the label is simply a number $y^l \in [-1, 1]$. To motivate this result, let us first start with a simpler situation that the label is a physical observable $y^l = \langle \psi^l | \hat{O} | \psi^l \rangle$, where $\hat{O}$ is a hermitian operator on n-qubit quantum state [33]. This is equivalent to say that y^l is a quadratic function of these complex numbers as $y^l = \sum_{mm'} O_{mm'} c_m^{l*} c_{m'}^l$, where $O_{mm'} = O_{m'm}^* = \langle m | \hat{O} | m' \rangle$. The QNN we considered is shown in Fig. 1(a).

A unitary $\hat{\mathcal{U}}$ is made of a number of two-qubit gates. We use $\hat{u}_{ij}$ to denote a two-qubit gate acting on qubit-i and -j. Each $\hat{u}_{ij}$ is parameterized as

$$\hat{u}_{ij} = e^{\sum_k \alpha_{ij}^k \hat{g}_k}, \tag{1}$$

where $\hat{g}_k$ are SU(4) generators and α_{ij}^k are parameters. In a QNN, these parameters need to be determined by training. We use the brick wall architecture to arrange these $\hat{u}_{ij}$ to form $\hat{\mathcal{U}}$. To make sure that the QNN can realize any kinds of unitary transformations, we set the circuit depth sufficiently large. The entire quantum circuit $\hat{\mathcal{U}}$ acts on the input wave function $|\psi^l\rangle$ and then we perform a measurement, say σ_x, on the readout qubit-r.

The measurement operator is therefore denoted by

$$\hat{M} = \hat{\sigma}_0^1 \otimes \cdots \otimes \hat{\sigma}_x^r \otimes \cdots \otimes \hat{\sigma}_0^N, \tag{2}$$

where the superscription $i = 1, \ldots, N$ denotes the qubits, and σ_0^i denotes the identity matrix. The measurement of the
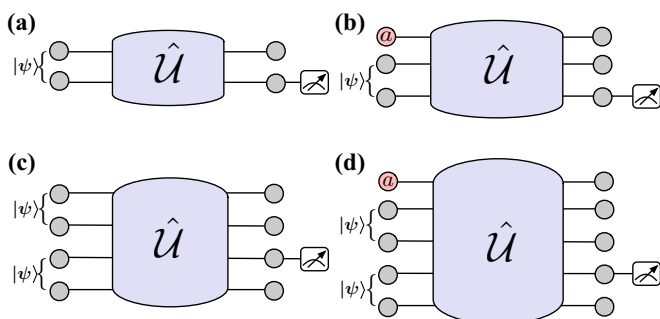

FIG. 1. (a) Conventional QNN structure. (b) QNN with one ancillary qubit. (c) QNN with two duplicated replicas of inputs. (d) QNN with two duplicated replicas of input and one ancillary qubit. Here $|\psi\rangle$ are input wave function in the dataset, $\hat{\mathcal{U}}$ denotes the unitary rotation by quantum circuit, and the detector denotes the readout qubit. The red qubit with label a denotes the ancillary qubit at which the wave function is always fixed.

quantum circuits leads to

$$\tilde{y}^l = \langle \psi^l | \hat{\mathcal{U}}^\dagger \hat{M} \hat{\mathcal{U}} | \psi^l \rangle. \tag{3}$$

The loss function is taken as $\mathcal{L} = \frac{1}{N_D} \sum_l |\tilde{y}^l - y^l|^2$, which enforces $\tilde{y}^l$ to be y^l for all $|\psi^l\rangle$. And we use the Adams method to optimize the coefficient of the generators of each two-qubit gate during training [31].

Here one thing that should be noticed is whether the input wave functions span the entire Hilbert space, which requires $N_D > 2^n$. Here N_D denotes the total number of data in the training set and 2^n is the dimension of the Hilbert space containing n qubits. If $N_D < 2^n$, all the input wave functions in the dataset only occupy a subset of the entire Hilbert space. In some cases, even when $N_D > 2^n$, if the wave functions have certain structures, for instance, if the wave functions are taken as ground states of certain Hamiltonians [30,31], they also do not span the entire Hilbert space. However, if $N_D > 2^n$ and the input wave functions are general enough, they span the entire Hilbert space. In this case, in order for all $\tilde{y}^l$ to faithfully represent y^l, one requires $\hat{O} = \hat{\mathcal{U}}^\dagger \hat{M} \hat{\mathcal{U}}$. However, this is not possible for a general operator $\hat{O}$. Because $\hat{M}$ is a direct product of the $\hat{\sigma}_x$ operator on the read-out qubit-r and the $n - 1$ identity operators on the rest qubits, the eigenvalues of $\hat{M}$ consist 2^{n-1} number of -1 and equal number of $+1$, and any unitary transformation keeps these eigenvalues invariant. That is to say, even though one can make the QNN deep enough to present a generic unitary $\hat{\mathcal{U}}$ in the SU(2^n) group, it always cannot satisfy $\hat{O} = \hat{\mathcal{U}}^\dagger \hat{M} \hat{\mathcal{U}}$. This argument can be easily generalized to situations that measurements are performed in more than one readout qubit.

A. Ancillary qubit

Now we show this problem can be solved by adding one ancillary qubit. Instead of $|\psi^l\rangle$, we now add one ancillary qubit and the input wave function is set as $|\alpha\rangle \otimes |\psi^l\rangle$, where the input state at the ancillary qubit is always fixed as $|\alpha\rangle$. The unitary $\hat{\mathcal{U}}$ now acts on the entire 2^{n+1}-dimensional Hilbert space, and the measurement is still performed in the readout qubit and now

$$\hat{\tilde{M}} = \hat{\sigma}_0^a \otimes \hat{\sigma}_0^1 \otimes \cdots \otimes \hat{\sigma}_x^r \otimes \cdots \otimes \hat{\sigma}_0^N, \tag{4}$$

where the superscript a denotes the ancillary qubit. The structure is shown in Fig. 1(b). Now we will show that for any given $|\alpha\rangle$, we can always construct an operator $\hat{\tilde{O}}$, acting on the 2^{n+1} Hilbert space, which satisfies the following two requirements. The first is that the operator $\hat{\tilde{O}}$ can generate the observables as $\langle \alpha | \otimes \langle \psi^l | \hat{\tilde{O}} | \alpha \rangle \otimes | \psi^l \rangle = \langle \psi^l | \hat{O} | \psi^l \rangle = y^l$, and the second is that the eigenvalues of the operator $\hat{\tilde{O}}$ consist of 2^n number of $+1$ and equal number of -1 and are consistent to that of the measurement operator $\hat{\tilde{M}}$. Without loss of generality, we choose $|\alpha\rangle = |\uparrow\rangle$, it can be shown that $\hat{\tilde{O}}$ chosen as $\hat{\sigma}_z^a \otimes \hat{O} + \hat{\sigma}_x^a \otimes \sqrt{I - \hat{O}^2}$ satisfies these two conditions. First,

$$\langle \uparrow | \otimes \langle \psi^l | \hat{\tilde{O}} | \uparrow \rangle \otimes | \psi^l \rangle$$
$$= \langle \uparrow | \hat{\sigma}_z^a | \uparrow \rangle \langle \psi^l | \hat{O} | \psi^l \rangle + \langle \uparrow | \hat{\sigma}_x^a | \uparrow \rangle \langle \psi^l | \sqrt{I - \hat{O}^2} | \psi^l \rangle$$
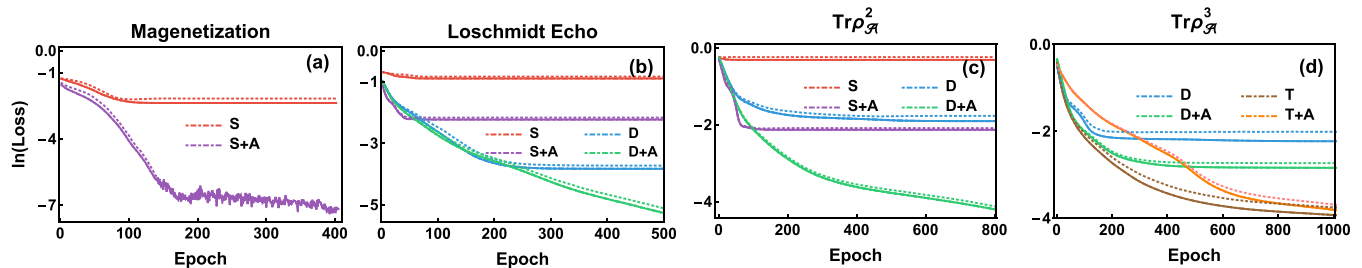$$= \langle \psi^l | \hat{O} | \psi^l \rangle = y^l. \tag{5}$$

FIG. 2. The loss function as a function of the training epoch. The solid lines and the dashed lines are the loss on the training dataset and the validation dataset, respectively. The target functions are, respectively, the magnetization (a), the Loschmidt echo (b), the second Rényi entropy (c) and the third Rényi entropy (d). For the input wave function $|\psi^l\rangle$ in the dataset, the number of qubit $n = 3$ for (a), and $n = 2$ for (b), (c) and (d). [35]. For the Rényi entropy discussed in (c) and (d), one qubit is taken as $\mathcal{A}$ and the other qubits are taken as $\mathcal{B}$. In the legends, S, D and T, respectively, denote single, double, and triple replicas as input for quantum circuit, and +A means that one ancillary qubit is added. Here we use the brick wall architecture made of two-qubit gate for quantum circuit.

Secondly, suppose $\{|m\rangle\}$ is a set of eigenbases in the 2^n-dimensional Hilbert space (without the ancillary qubit) that $\hat{O}|m\rangle = O_m|m\rangle$, and under these bases, $\hat{\bar{O}}$ can be written as $\sum_m (\hat{\sigma}_z^a O_m + \hat{\sigma}_x^a \sqrt{1 - O_m^2}|m\rangle\langle m|$. Therefore, its eigenvalue consists of 2^n number of $+1$ and equal number of -1, which equal the eigenvalues of operator $\hat{M}$. When such an operator $\hat{\bar{O}}$ is found, it is possible to find a unitary $\hat{\mathcal{U}}$ in the 2^{n+1}-dimensional space, such that $\hat{\mathcal{U}}^\dagger \hat{M} \hat{\mathcal{U}} = \hat{\bar{O}}$, and then,

$$\langle\alpha| \otimes \langle\psi^l|\hat{\mathcal{U}}^\dagger \hat{M} \hat{\mathcal{U}}|\alpha\rangle \otimes |\psi^l\rangle = \langle\alpha| \otimes \langle\psi^l|\hat{\bar{O}}|\alpha\rangle \otimes |\psi^l\rangle = y^l. \quad (6)$$

This shows, with the help of one ancillary qubit, the QNN can accurately express the functional mapping $y^l = \langle\psi^l|\hat{O}|\psi^l\rangle$ for all generic quantum states $|\psi^l\rangle$.

In Fig. 2(a), we show that the loss function for learning total magnetization of generic wave functions in a three-qubit quantum state, with $\hat{O}$ chosen as $\sum_i \sigma_z^i/n$. The red lines and the purple lines are results for QNN with structures shown in Figs. 1(a) and 1(b), respectively. The structure shown in Fig. 1(b) has one more ancillary qubit compared with the structure shown in Fig. 1(a). One can see that, if without the ancillary qubit, the loss clearly saturates to a finite value even for sufficient large training epochs. By adding the ancillary qubit, the loss is significantly reduced and approaches to zero.

B. General rule

The lesson from above example is that the learning accuracy can be significantly improved by enlarging the Hilbert space dimension of the input for the quantum circuit. Here we generalize this lesson to a general statement. Suppose $\mathcal{H}$ is the total Hilbert space of input for quantum circuit, $\mathrm{Dim}(\mathcal{H})$ denotes its Hilbert space dimension. Let us consider $\mathcal{H} = \mathcal{H}_0 \bigoplus \mathcal{H}_1$, and suppose all input wave functions in the dataset only reside in $\mathcal{H}_0$. The statement is that, when $\mathrm{Dim}(\mathcal{H}_1) \geqslant \mathrm{Dim}(\mathcal{H}_0)$, we can always find an operator $\hat{\bar{O}}$ acting on entire Hilbert space $\mathcal{H}$, such that, (i) for any wave function $|\psi\rangle$ in $\mathcal{H}_0$, $\langle\psi_\mathcal{H}|\hat{\bar{O}}|\psi_\mathcal{H}\rangle = \langle\psi|\hat{O}|\psi\rangle$, (ii) the eigenvalues of $\hat{\bar{O}}$ consists of an equal number of $+1$ and -1, which are the same as the eigenvalue of the measurement operator. Then, it is possible to find a proper $\hat{\mathcal{U}}$ such that $\hat{\bar{O}} = \hat{\mathcal{U}}^\dagger \hat{M} \hat{\mathcal{U}}$.

The construction of $\hat{\bar{O}}$ is quite similar as the ancillary qubit example. Let us first consider the situation $\mathrm{Dim}(\mathcal{H}_1) = \mathrm{Dim}(\mathcal{H}_0)$, and suppose $\{|m\rangle\}$ is a set of eigenbases in $\mathcal{H}_0$ with $\hat{O}|m\rangle = O_m|m\rangle$. We can then define a set of bases $\{|\bar{m}\rangle\}$ in $\mathcal{H}_1$ that have one-to-one correspondence with the bases in set $\{|m\rangle\}$, say, each $|\bar{m}\rangle$ corresponds to a $|m\rangle$. Then, we can construct $\hat{\bar{O}}$ as

$$\hat{\bar{O}} = \sum_m O_m(|m\rangle\langle m| - |\bar{m}\rangle\langle\bar{m}|)$$
$$+ \sqrt{1 - O_m^2}(|m\rangle\langle\bar{m}| + |\bar{m}\rangle\langle m|). \quad (7)$$

It is easy to see that $\hat{\bar{O}}$ constructed as Eq. (7) satisfies the above two requirements. This can be extended to situations $\mathrm{Dim}(\mathcal{H}_1) > \mathrm{Dim}(\mathcal{H}_0)$. In this case, $\mathcal{H}_1$ is larger than the space spanned by $\{|\bar{m}\rangle\}$, and we choose $\hat{\bar{O}}$ to be diagonal with equal number of $+1$ and -1 eigenvalues in the residual Hilbert space. The ancillary qubit is a specific example of this general statement, where $\mathcal{H}_0$ consists states $|\uparrow\rangle \otimes |\psi\rangle$ and $\mathcal{H}_1$ consists states $|\downarrow\rangle \otimes |\psi\rangle$, where $|\psi\rangle$ denotes the input wave functions in the dataset. We have proved that if the condition $\mathrm{Dim}(\mathcal{H})/2 \geqslant \mathrm{Dim}(\mathcal{H}_0)$ is fulfilled, then a learnable observable can be constructed. However, we shall also note that this condition is sufficient but not necessary. For certain special cases, if a specific target observable $\hat{O}$ happens to share the same set of eigenvalues as the measurement operator $\hat{M}$, the measurement operator can be rotated to the target observable, such that $\hat{U}^\dagger \hat{M} \hat{U} = \hat{O}$ even without requiring $\mathrm{Dim}(\mathcal{H})/2 \geqslant \mathrm{Dim}(\mathcal{H}_0)$ condition.

C. Replica

Now we move to consider the learning tasks such as the Loschmidt echo and the Rényi entropy. The Loschmidt echo is an interference between two wave functions, starting from the same input wave function $|\psi^l\rangle$ and evolved by two different Hamiltonians $\hat{H}_a$ and $\hat{H}_b$ for time duration t, that is, $y^l = |\langle\psi^l|e^{i\hat{H}_a t}e^{-i\hat{H}_b t}|\psi^l\rangle|^2$. Here we denote $\hat{W} = e^{i\hat{H}_a t}e^{-i\hat{H}_b t}$, and for most Hamiltonians, $\hat{W}$ is a sufficiently chaotic operator for long enough t [34]. In Fig. 2(b), adopting the QNN in Figs. 1(a) and 1(b) as before, we show the loss function for learning the Loschmidt echo. We can see that even with

an ancillary qubit, the loss still saturates to a finite nonzero value even with sufficient long training epochs. The reason is also quite obvious. It is because for the Loschmidt echo, the label y is a quartic function of $\{c_m\}$, while $\tilde{y}$ given the QNN through Eq. (3) is only a quadratic function of $\{c_m\}$. Thus, to accurately capture the learning target such as the Loschmidt echo, nonlinearity is necessary.

There are also various discussions on adding nonlinearity in QNN. Here we show that duplicating replica of the input states is another way to incorporate the nonlinearity. In fact, it is a quite efficient way in this case, which can be easily seen from

$$
\begin{aligned}
y^l &= |\langle\psi^l|\hat{W}|\psi^l\rangle|^2 = \langle\psi^l|\otimes\langle\psi^l|\hat{W}\otimes\hat{W}^\dagger|\psi^l\rangle\otimes|\psi^l\rangle \\
&= \langle\psi^l|\otimes\langle\psi^l|\hat{W}^\dagger\otimes\hat{W}|\psi^l\rangle\otimes|\psi^l\rangle \\
&= \langle\psi^l|\otimes\langle\psi^l|\tfrac{1}{2}(\hat{W}^\dagger\otimes\hat{W}+\hat{W}\otimes\hat{W}^\dagger)|\psi^l\rangle\otimes|\psi^l\rangle. \quad (8)
\end{aligned}
$$

Suppose the input wave function is a n-qubit state, and when we double the input to a $2n$-qubit state, the Loschmidt echo returns to a quadratic function in the enlarged Hilbert space. In the doubled space, the Loschmidt echo becomes a physical observable with $\hat{O} = (\hat{W}^\dagger\otimes\hat{W}+\hat{W}\otimes\hat{W}^\dagger)/2$ being a Hermitian operator. The fact that a mth-degree polynomial function of the elements of the density matrix ρ can be written as the expectation value of an observable on m copies of ρ has also been discussed in Ref. [36].

We can also consider another example of the Rényi entropy. For an input wave function $|\psi^l\rangle$, by partially tracing out a subsystem $\mathcal{B}$, the reduced density matrix for remaining subsystem $\mathcal{A}$ is given by $\rho_{\mathcal{A}} = \mathrm{Tr}_{\mathcal{B}}|\psi^l\rangle\langle\psi^l|$. The output of the QNN is in the range $[-1, 1]$, while the mth order Rényi entropy is defined as $S_A^{(m)} = (\log\mathrm{Tr}\rho_{\mathcal{A}}^m)/(1-m)$. Consequently, we choose the label y^l as $y^l = \exp((1-m)S_A^{(m)}) = \mathrm{Tr}\rho_{\mathcal{A}}^m$. We show in Fig. 2(c) the loss function for learning the second Rényi entropy. Here we consider the two-qubit system, one qubit is taken as $\mathcal{A}$ and the other qubits are taken as $\mathcal{B}$. Similar as the Loschmidt echo case, without replica the loss still saturates to a finite nonzero value at sufficient long training epochs even with an ancillary qubit. Similarly, for considering the second Rényi entropy, let us first study

$$
y^l = \mathrm{Tr}\rho_{\mathcal{A}}^2 = \langle\psi^l|\otimes\langle\psi^l|\hat{\mathcal{X}}_{\mathcal{A}}\otimes\hat{\mathcal{I}}_{\mathcal{B}}|\psi^l\rangle|\otimes|\psi^l\rangle, \quad (9)
$$

where $\hat{\mathcal{X}}_{\mathcal{A}}$ is the swap operator acting on subsystem $\mathcal{A}$ between two replicas, and $\hat{\mathcal{I}}_{\mathcal{B}}$ is the identity operator acting on subsystem $\mathcal{B}$ between two replicas. Then, the second Rényi entropy, defined as $-\log y^l$, can be related to physical observable in the doubled Hilbert space. This can also be generalized to the higher order Rényi entropy, and in general, for considering the mth order Rényi entropy, we study

$$
y^l = \mathrm{Tr}\rho_{\mathcal{A}}^m = \langle\psi^l|^{\otimes m}\hat{\mathcal{X}}_{\mathcal{A}}\otimes\hat{\mathcal{I}}_{\mathcal{B}}|\psi^l\rangle^{\otimes m}, \quad (10)
$$

for which we need m replicas, and the mth order Rényi entropy is defined as $(\log y^l)/(1-m)$.

Hence, we double the size of the input for the quantum circuit and duplicate two replicas of the input wave functions as the input. The unitary $\hat{\mathcal{U}}$ then acts on the total Hilbert space with $\mathrm{Dim}(\mathcal{H}) = 2^{2n}$, with the same measurement on the readout qubit as discussed above. The structure is shown in Fig. 1(c). Now the question is that, with the enlarged Hilbert

space, whether we still need an ancillary qubit, as shown in Fig. 1(d). Note that the input wave functions are all subjected to a constraint that they have to be symmetric between two replicas, therefore, these wave functions do not span all 2^{2n} dimensional Hilbert space. Let us denote such symmetric Hilbert space as $\mathcal{H}_0$, and with the general statement we discussed above, it is important to analyze whether $\mathrm{Dim}(\mathcal{H}_0)$ is larger than the half of $\mathrm{Dim}(\mathcal{H})$. It can be shown that

$$
\mathrm{Dim}(\mathcal{H}_0) = 2^n + \frac{2^n(2^n-1)}{2} = 2^{2n-1} + 2^{n-1}, \quad (11)
$$

where the first term in the middle counts the dimension of the sub-Hilbert space spanned by $|m\rangle\otimes|m\rangle$, and the second term counts the dimension of the sub-Hilbert space spanned by $(|m\rangle\otimes|m'\rangle+|m'\rangle\otimes|m\rangle)/\sqrt{2}$. It is obvious that $\mathrm{Dim}(\mathcal{H}_0)$ is larger than $\mathrm{Dim}(\mathcal{H})/2$. In other word, because $\mathrm{Dim}(\mathcal{H}_1) < \mathrm{Dim}(\mathcal{H}_0)$, one still needs the ancillary qubit in order to construct a proper $\hat{\hat{O}}$. This can be seen from Figs. 2(b) and 2(c) for the cases of learning the Loschmidt echo and the second Rényi entropy, respectively. Especially, for learning the second Rényi entropy, one can see that the loss can still be reduced by adding an ancillary qubit even with doubled input.

The situation becomes different when one considers the tripled Hilbert space, for instance, when considering the third Rényi entropy. For the tripled Hilbert space, if we still require the wave functions to be symmetric between three replicas, similar as the analyzation that leads to Eq. (11), we obtain that the Hilbert space dimension $\mathrm{Dim}(\mathcal{H}_0)$ is given by

$$
\begin{aligned}
\mathrm{Dim}(\mathcal{H}_0) &= 2^n + 2^n(2^n-1) + \frac{2^n(2^n-1)(2^n-2)}{6} \quad (12) \\
&= \frac{1}{3}(2^{3n-1} + 3\times 2^{2n-1} + 2^n) \leqslant \frac{1}{2}2^{3n} \\
&= \frac{1}{2}\mathrm{Dim}(\mathcal{H}). \quad (13)
\end{aligned}
$$

Therefore, in this case, the requirement for finding a proper $\hat{\hat{O}}$ can be satisfied without adding an ancila qubit. This can be seen in Fig. 2(d). One can see with two replicas, the loss cannot be reduced to a sufficient small value, for both cases without and with the ancila qubit. However, when there are three replicas, even without an ancila qubit, the loss can already drop to be sufficiently close to zero. And for a sufficiently long training epoch, the losses for QNN with or without ancila qubit approach the same value. This shows that the ancila qubit is not necessary in this case when there are three replicas. And the same conclusion can be generalized to situations with more than three replicas.

III. DISCUSSION

In this work, we consider the expressivity of QNN for learning targets that are observables (i.e., expectations of a hermitian operator) of input wave functions. These also include the situations that the learning targets are not observable of input wave functions, but can be expressed as observables in the enlarged Hilbert space with multiple replicas of input wave functions, such as the Loschmidt echo and the Rényi entropy. The main finding of this work is that such target can be expressed accurately only when the input wave functions in

all dataset only occupy a subset $\mathcal{H}_0$ of the entire Hilbert space $\mathcal{H}$ that the quantum circuit acts on, especially, we require the condition $\text{Dim}(\mathcal{H}_0) < \text{Dim}(\mathcal{H})/2$. An accurate approximation of the learning target is possible for a sufficiently deep QNN either when this condition is satisfied naturally by the dataset, or when the condition is enforced by artificially adding an ancillary qubit.

Our discussions also provide a general recipe for improving the learning accuracy, provided that no prior knowledge of such learning task is known. First, one can first try to add an ancillary quibit. If not satisfactory, then one can duplicate two replicas with an ancillary qubit. Finally, if still not satisfied, one can add more replicas, and when the number of replica equals or is greater than three, the ancillary qubit is no longer needed. However, we shall also note that increasing the number of replicas cause serious computational resources and this method cannot be extended to a large number of replicase. This limits this way of adding nonlinearity and for situations that the nonlinearity is too strong, this method should be combined with other ways of adding nonlinearity.

In the future, we can consider a number of generalizations of such studies. First, here we focus on learning targets that are observables or generalized observables, and we can consider more sophisticated learning targets. Secondly, here we focus on regression tasks, and we can consider classification tasks. Thirdly, here we focus on the fully connected architectures, and we can consider other architectures of QNN, such as the convolutional QNN [25–27] and the recurrent QNN [28,29]. We hope such studies can lead to analogy of the universal approximation theorem for QNN and lay the foundation of the expressive power of QNN.

[1] I. Goodfellow, Y. Bengio, and A. Couville, *Deep Learning* (MIT Press, Cambridge, MA, 2016).

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

[3] J. Preskill, Quantum computing in the NISQ era and beyond, Quantum **2**, 79 (2018).

[4] M. Broughton, G. Verdon, T. McCourt, Antonio J. Martinez, J. H. Yoo, S. V. Isakov, P. Massey, M. Y. Niu, R. Halavati, E. Peters, M. Leib, A. Skolik, M. Streif, D. Von Dollen, Jarrod R. McClean, S. Boixo, D. Bacon, A. K. Ho, H. Neven, M. Mohseni, TensorFlow quantum: A software framework for quantum machine learning, arXiv:2003.02989.

[5] L. Lamata, Quantum machine learning and quantum biomimetics: A perspective, Mach. Learn.: Sci. Technol. **1**, 033002 (2020).

[6] E. Farhi, and H. Neven, Classification with quantum neural networks on near term processors, arXiv:1802.06002.

[7] E. Grant, M. Benedetti, S. Cao, A. Hallam, J. Lockhart, V. Stojevic, A. G. Green, and S. Severini, Hierarchical quantum classifiers, npj Quantum Inf. **4**, 65 (2018).

[8] J.-G. Liu and L. Wang, Differentiable learning of quantum circuit Born machines, Phys. Rev. A **98**, 062324 (2018).

[9] J. Zeng, Y. Wu, J.-G. Liu, L. Wang, and J. Hu, Learning and inference on generative adversarial quantum circuits, Phys. Rev. A **99**, 052306 (2019).

[10] M. Schuld, A. Bocharov, K. M. Svore, and N. Wiebe, Circuit-centric quantum classifiers, Phys. Rev. A **101**, 032308 (2020).

[11] X. Liang, Y. Wu, and H. Zhai, The quantum cocktail party problem, Science China Phys. Mech. Astronomy **63**, 250362 (2020).

[12] M. Benedetti, J. Realpe-Gómez, R. Biswas, and A. Perdomo-Ortiz, Quantum-Assisted Learning of Hardware-Embedded Probabilistic Graphical Models, Phys. Rev. X **7**, 041052 (2017).

[13] Y. Du, M.-H. Hsieh, T. Liu, and D. Tao, The expressive power of parameterized quantum circuits, Phys. Rev. Research **2**, 033125 (2020).

[14] J. R. McClean, S. Boixo, V. N. Smelyanskiy, R. Babbush, and H. Neven, Barren plateaus in quantum neural network training landscapes, Nat. Commun. **9**, 4812 (2018).

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

[16] M. Benedetti, D. Garcia-Pintos, O. Perdomo, V. Leyton-Ortega, Y. Nam, and A. PerdomoOrtiz, A generative modeling approach for benchmarking and training shallow quantum circuits, npj Quantum Inf. **5**, 45 (2019).

[17] M. J. S. Beach, Roger G. Melko, T. Grover, and T. H. Hsieh, Making trotters sprint: A variational imaginary time ansatz for quantum many-body systems, Phys. Rev. B **100**, 094434 (2019).

[18] K. Beer, D. Bondarenko, T. Farrelly, T. J. Osborne, R. Salzmann, and R. Wolf, Efficient learning for deep quantum neural networks, Nat. Commun. **11**, 808 (2020).

[19] W. Huggins, P. Patil, B. Mitchell, K. B. Whaley, and E. M. Stoudenmire, Towards quantum machine learning with tensor networks, Quantum Sci. Technol. **4**, 024001 (2019).

[20] E. Torrontegui, and J. J. García-Ripoll, Unitary quantum perceptron as efficient universal approximator, EPL Europhys. Lett. **125**, 30004 (2019).

[21] S. Yan, H. Qi, and W. Cui, Nonlinear quantum neuron: A fundamental building block for quantum neural networks, Phys. Rev. A **102**, 052421 (2020).

[22] J. Romero, J. P. Olson, and A. Aspuru-Guzik, Quantum autoencoders for efficient compression of quantum data, Quantum Sci. Technol. **2**, 045001 (2017).

[23] D. Bondarenko, and P. Feldmann, Quantum Autoencoders to Denoise Quantum Data, Phys. Rev. Lett. **124**, 130502 (2020).

[24] A. Pepper, N. Tischler, and Geoff J. Pryde, Experimental Realization of a Quantum Autoencoder: The Compression of Qutrits via Machine Learning, Phys. Rev. Lett. **122**, 060501 (2019).

[25] I. Cong, S. Choi, and Mikhail D. Lukin, Quantum convolutional neural networks, Nat. Phys. **15**, 1273 (2019).

[26] M. Henderson, S. Shakya, S. Pradhan, and T. Cook, Quanvolutional neural networks: powering image recognition with quantum circuits, arXiv:1904.04767.

[27] S. Oh, J. Choi, and J. Kim, A tutorial on quantum convolutional neural networks (QCNN), arXiv:2009.09423.

[28] J. Bausch, Recurrent quantum neural networks, in *Advances in Neural Information Processing Systems 33* (NeurIPS 2020).

[29] L. Behera, I. Kar, and A. C. Elitzur, Recurrent quantum neural network and its applications, *The Emerging Physics of Consciousness* (The Frontiers Collection, Springer, Berlin, Heidelberg, 2006).

[30] H. Shen, P. Zhang, Y.-Z. You, and H. Zhai, Information Scrambling in Quantum Neural Networks, Phys. Rev. Lett. **124**, 200504 (2020).

[31] Y. Wu, P. Zhang, and H. Zhai, Scrambling ability of quantum neural networks architectures, arXiv:2011.07698.

[32] L. G. Wright and P. L. McMahon, The capacity of quantum neural networks, arXiv:1908.01364.

[33] We can normalize $\hat{O}$ operator such that all its eigenvalues are between -1 and $+1$.

[34] Two different Hamiltonians $\hat{H}_a$ and $\hat{H}_b$ are random generated

$$\hat{H} = \sum_{\alpha=x,y,z} \sum_{i=1}^{n} h_\alpha^i \hat{\sigma}_\alpha^i + \sum_{\alpha,\beta=x,y,z} \sum_{i,j=1}^{n} J_{\alpha\beta}^{ij} \hat{\sigma}_\alpha^i \hat{\sigma}_\beta^j,$$

where $\{h\}$, $\{J\}$ are all randomly sampled with the uniform distribution in the range $(-2, 2)$. We take the evaluation time $t = 8\pi$.

[35] The computational resources needed in the simulation grow exponentially in terms of the quit number n. For m replicas with an ancillary qubit, the dimension is 2^{mn+1}. The large number of parameters to be optimized limits the size of our simulation, and therefore, here we consider a small system size with $n = 2, 3$.

[36] T. A. Brun, Measuring polynomial functions of states, Quantum Inf. Comput. **4**, 401 (2004).