

# Efficient measure for the expressivity of variational quantum algorithms

Yuxuan Du,<sup>1,\*</sup> Zhuozhuo Tu,<sup>2,†</sup> Xiao Yuan,<sup>3,‡</sup> and Dacheng Tao<sup>1,§</sup>

<sup>1</sup>*JD Explore Academy*

<sup>2</sup>*School of Computer Science, The University of Sydney*

<sup>3</sup>*Center on Frontiers of Computing Studies, Department of Computer Science, Peking University, Beijing 100871, China*

The superiority of variational quantum algorithms (VQAs) such as quantum neural networks (QNNs) and variational quantum eigen-solvers (VQEs) heavily depends on the expressivity of the employed ansätze. Namely, a simple ansatz is insufficient to capture the optimal solution, while an intricate ansatz leads to the hardness of trainability. Despite its fundamental importance, an effective strategy of measuring the expressivity of VQAs remains largely unknown. Here, we exploit an advanced tool in statistical learning theory, i.e., covering number, to study the expressivity of VQAs. Particularly, we first exhibit how the expressivity of VQAs with an arbitrary ansatz is upper bounded by the number of quantum gates and the measurement observable. We next explore the expressivity of VQAs on near-term quantum chips, where the system noise is considered. We observe an exponential decay of the expressivity with increasing circuit depth. We also utilize the achieved expressivity to analyze the generalization of QNNs and the accuracy of VQE. We numerically verify our theory employing VQAs with different levels of expressivity. Our work opens the avenue for quantitative understanding of the expressivity of VQAs.

*Introduction.*—A paramount mission in quantum computing is devising learning protocols outperforming classical methods [1–3]. Variational quantum algorithms (VQAs) [4–8] using parameterized quantum circuits — *ansätze* and classical optimizers, serve as promising candidates to achieve this goal, especially in the noisy intermediate-scale quantum (NISQ) era [9]. Theoretical evidence has shown that VQAs may provide runtime speedups and enhanced generalization bounds for quantum information, quantum chemistry, and quantum machine learning (QML) tasks [10–14]. Meanwhile, VQAs are flexible, which can adapt to restrictions imposed by NISQ devices such as qudits connectivity and shallow circuit depth. With this regard, great efforts have been dedicated to designing VQAs with varied ansätze to address different problems. Two important categories of existing VQAs include quantum neural networks (QNNs) [15–17] and variational quantum eigen-solvers (VQEs) [18–20]. Empirical studies have shown VQAs on near-term quantum devices achieving good performance for various tasks [20–23].

In parallel to the algorithm design, another central topic in the context of VQAs is exploring their learnability. A well study of this topic does not only allow us to understand the capabilities and limitations of VQAs with varied ansätze, but can also guide us to devise more powerful quantum protocols. As such, theoretical studies have attempted to exploit learnability of VQAs from distinct views. Refs. [24–27] have exhibited that the optimization of VQA suffers from barren plateaus, where gradients information will be exponentially vanished with respect to the number of qudits and the circuit depth;

Refs. [11, 28] have shown that more measurements, lower noise, and shallower circuit depth contribute to a better convergence of QNNs with gradient descent optimizers; Refs. [11, 12, 29–33] have proven the generalization of QNNs with varied ansätze. Ref. [34] has established quantum no-free lunch theorem of QNN and provided an apparently stronger lower bound than its classical counterpart. Very recently, Refs. [35–37] connect the trainability and expressibility of VQAs, i.e., an ansatz exhibited with higher expressibility implies a flatter loss landscapes and therefore will be harder to train. Hence, to ensure the power of VQAs, it is indispensable to develop an effective tool to measure the expressibility of VQAs. To this end, prior literature uses the unitary  $t$ -design to quantify the expressivity of VQAs [38]. However, such a quantity is hard to calculate for a realistic quantum circuit and VQAs with well-designed ansätze may not obey the assumptions imposed by the unitary  $t$ -design [27, 39]. The above caveats motivate us to rethink: ‘*Is there any effective and generic way to measure the expressivity of VQAs?*’

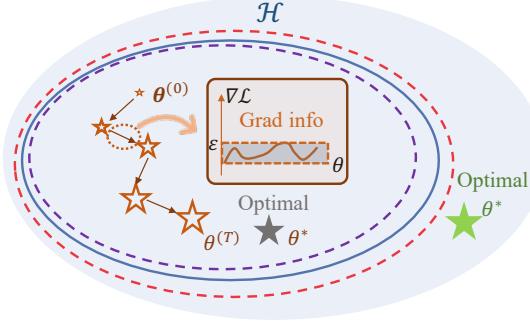
Here, we provide a positive affirmation toward this question. Through connecting the expressivity with the model complexity, we leverage an advanced tool in statistical learning theory — covering number [40], to quantify the expressivity of VQAs. We first exhibit that in the measure of the covering number, the upper bound of the expressivity for a given VQA yields  $\mathcal{O}((N_{gt}\|O\|)^{d^2 k N_{gt}})$ , where  $d$ ,  $N_{gt}$ ,  $k$ , and  $\|O\|$  refer to the dimension of *qudit*, the number of trainable quantum gates, the largest number of qudits operated with a single quantum gate, and the operator norm of the observable  $O$  used in the employed ansätze, respectively. With fixed  $d$  and  $\|O\|$ , the expressivity of VQA can be well controlled by tuning  $N_{gt}$  and  $k$ . Our second contribution is analyzing the expressivity of VQAs under the NISQ setting. When the quantum system noise is simulated by the depolarization channel, the expressivity is upper bounded by

\* [duyuxuan123@gmail.com](mailto:duyuxuan123@gmail.com)

† [zhtu3055@uni.sydney.edu.au](mailto:zhtu3055@uni.sydney.edu.au)

‡ [xiaoyuan@pku.edu.cn](mailto:xiaoyuan@pku.edu.cn)

§ [dacheng.tao@gmail.com](mailto:dacheng.tao@gmail.com)



**FIG. 1: Overview the expressivity of VQAs.** The expressivity of the employed ansätze of VQA rules its hypothesis space  $\mathcal{H}$  (solid blue ellipse). When  $\mathcal{H}$  has the modest size and covers the target concept (grey solid star), VQA can attain a good performance. Conversely, when  $\mathcal{H}$  fails to cover the target concept (green solid star), due to the limited expressivity, VQA achieves a poor performance.

$\mathcal{O}((1-p)^{N_g}(N_{gt}\|O\|)^{d^{2k}N_{gt}})$ , where  $N_g$  is the total number of quantum gates (including both trainable and fixed ones) in the ansätze with  $N_g \geq N_{gt}$  and  $p$  is the depolarization rate. We further harness the derived expressivity to show that the generalization error bound of QNNs scales with  $\tilde{\mathcal{O}}(\sqrt{N_{gt}d^k}/\sqrt{n})$ , where  $n$  is the number of training examples. This means that ansätze constituted by a large number of quantum gates request an increased number of training examples to ensure convergence. We believe that these observations may be of independent interest for the quantum machine learning community.

*Expressivity of VQA.*—We first review the working-flow of VQAs, which contains an  $N$ -qudits quantum circuit and a classical optimizer. In the training stage, VQA follows an iterative manner to proceed optimization, where the optimizer continuously leverages the output of the quantum circuit to update trainable parameters of the adopted ansatz to minimize the predefined objective function  $\mathcal{L}(\cdot)$ . At the  $t$ -th iteration, the updating rule is  $\theta^{(t+1)} = \theta^{(t)} - \eta \frac{\partial \mathcal{L}(h(\theta^{(t)}, O, \rho), c_1)}{\partial \theta}$ , where  $\eta$  is the learning rate,  $c_1 \in \mathbb{R}$  is the target label, and  $h(\theta^{(t)}, O, \rho)$  amounts to the output of the quantum circuit as elaborated below. Define  $\rho \in \mathbb{C}^{d^N \times d^N}$  as the  $N$ -qudit input quantum state,  $O \in \mathbb{C}^{d^N \times d^N}$  as the quantum observable,  $\hat{U}(\theta) = \prod_{l=1}^{N_g} \hat{u}_l(\theta) \in \mathcal{U}(d^N)$  as the applied ansatz, i.e.,  $\theta \in \Theta$  are trainable parameters living in the parameter space  $\Theta$ ,  $\hat{u}_l(\theta) \in \mathcal{U}(d^k)$  refers to the  $l$ -th quantum gate operated with at most  $k$ -qudits with  $k \leq N$ , and  $\mathcal{U}(d^N)$  stands for the unitary group in dimension  $d^N$ . In general,  $\hat{U}(\theta)$  is formed by  $N_{gt}$  trainable gates and  $N_g - N_{gt}$  fixed gates, e.g.,  $\Theta \subseteq [0, 2\pi]^{N_{gt}}$ . Under the above definitions, the explicit form of the output of the quantum circuit under the ideal scenario is

$$h(\theta^{(t)}, O, \rho) := \text{Tr} \left( \hat{U}(\theta^{(t)})^\dagger O \hat{U}(\theta^{(t)}) \rho \right). \quad (1)$$

The gradients information  $\partial \mathcal{L}(h(\theta^{(t)}, O, \rho), c_1)/\partial \theta$  can

be acquired via the parameter shift rule or other methods [16, 41, 42]. The definition of  $h(\theta^{(t)}, O, \rho)$  is generic. Here the unitary  $\hat{U}(\theta)$  covers many representative ansätze in QML and quantum chemistry, e.g., the hardware-efficient ansatz and unitary coupled-cluster ansätze [6], and QNNs and VQEs can be effectively adapted to the form of  $h(\theta^{(t)}, O, \rho)$  (see following sections for details).

We now introduce the relationship between the expressivity and model complexity. In essence, the aim of VQAs is to find a good hypothesis  $h^*(\theta, O, \rho) = \arg \min_{h(\theta, O, \rho) \in \mathcal{H}} \mathcal{L}(h(\theta, O, \rho), c_1)$  that can well approximate the target concept, where  $\mathcal{H}$  refers to the hypothesis space of VQA with

$$\mathcal{H} = \left\{ \text{Tr} \left( \hat{U}(\theta)^\dagger O \hat{U}(\theta) \rho \right) \mid \theta \in \Theta \right\}. \quad (2)$$

An intuition about how the hypothesis space  $\mathcal{H}$  affects the performance of VQA is depicted in Fig. 1. When  $\mathcal{H}$  has *the modest size* and covers the target concepts, the estimated hypothesis could well approximate the target concept. By contrast, when the complexity of  $\mathcal{H}$  is too low, there exists a large gap between the estimated hypothesis and the target concept. Hence, to understand the expressivity of VQAs, it is highly demanded to devise an effective measure to evaluate the complexity of  $\mathcal{H}$ .

Here we employ covering number, an advanced tool broadly used in statistical learning theory [40], to bound the complexity of  $\mathcal{H}$  and measure the expressivity of VQAs.

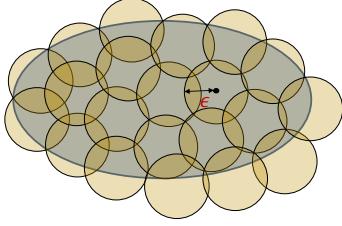
**Definition 1** (Covering number). *The covering number  $\mathcal{N}(\mathcal{U}, \epsilon, \|\cdot\|)$  denotes the least cardinality of any subset  $\mathcal{V} \subset \mathcal{U}$  that covers  $\mathcal{U}$  at scale  $\epsilon$  with a norm  $\|\cdot\|$ , i.e.,  $\sup_{A \in \mathcal{U}} \min_{B \in \mathcal{V}} \|A - B\| \leq \epsilon$ . Here we use this notion to measure the expressivity of VQAs.*

The geometric interpretation of covering number is depicted in Fig. 2, which refers to the minimum number of spherical balls with radius  $\epsilon$  that are required to completely cover a given space with possible overlaps. This notion has been employed to study other crucial topics in quantum physics such as Hamiltonian simulation [43] and entangled states [44]. Note that  $\epsilon$  is a predefined hyper-parameter, i.e., a small constant with  $\epsilon \in (0, 1)$ , and is independent with any factor [40]. This convention has been broadly adopted in the regime of machine learning to evaluate the model capacity of various learning models [40, 45]. The following theorem shows the upper bound of  $\mathcal{N}(\mathcal{H}, \epsilon, |\cdot|)$  whose proof is shown in Appendix A.

**Theorem 1.** *For  $0 < \epsilon < 1/10$ , the covering number of the hypothesis space  $\mathcal{H}$  in Eq. (2) yields*

$$\mathcal{N}(\mathcal{H}, \epsilon, |\cdot|) \leq \left( \frac{7N_{gt}\|O\|}{\epsilon} \right)^{d^{2k}N_{gt}}, \quad (3)$$

where  $\|O\|$  denotes the operator norm of  $O$ .



**FIG. 2: The geometric intuition of covering number.** Covering number concerns the minimum number of spherical balls with radius  $\epsilon$  that occupies the whole space.

It indicates that the most decisive factor, which controls the complexity of  $\mathcal{H}$ , is the employed quantum gates in  $\hat{U}(\boldsymbol{\theta})$ . This claim is ensured by the fact that the term  $d^{2k} N_{gt}$  exponentially scales the complexity  $\mathcal{N}(\mathcal{H}, \epsilon, |\cdot|)$ . Meanwhile, the qudits count  $N$  and the operator norm  $\|O\|$  polynomially scale the complexity of  $\mathcal{N}(\mathcal{H}, \epsilon, |\cdot|)$ . These observations suggest a succinct and direct way to compare the expressivity of VQAs with differed ansätze. Moreover, different from prior works, we first prove that the expressivity of VQAs depends on the type of quantum gates (denoted by the term  $k$ ). Since it is a long standing problem of proving that the expressivity of VQAs depends on the structure information of ansatz such as the location of different quantum gates and the types of the employed quantum gates, our result makes a concrete progress toward this goal. It is noteworthy that our results do not only indicate a general scaling behavior of the model's expressivity, but also provide a practical guidance of designing VQA-based models. In Appendix H, we elaborate how to combine the achieved theoretical results with *structural risk minimization* to enhance the learning performance of VQA-based models [45].

**Remark.** Theorem 1 is *ubiquitous* and *do not rely on the assumption of the unitary  $t$ -design*, which differs from [35]. Moreover, the qubit-based VQAs are a special case of our results with  $d = 2$ . We also study the tightness of the bound In Appendix I.

We next consider how the expressivity, or equivalently covering number, of VQA varies when noise  $\mathcal{E}(\cdot)$  is considered. Under this scenario, the hypothesis space of VQA in Eq. (2) transforms to

$$\tilde{\mathcal{H}} = \left\{ \text{Tr} \left( O \mathcal{E} \left( \hat{U}(\boldsymbol{\theta}) \rho \hat{U}(\boldsymbol{\theta})^\dagger \right) \right) \mid \boldsymbol{\theta} \in \Theta \right\}. \quad (4)$$

The expressivity of noisy VQAs is summarized in Proposition 1, whose proof is provided in Appendix B.

**Proposition 1.** *Following notations and conditions in Theorem 1, the covering number of  $\tilde{\mathcal{H}}$  in Eq. (4) satisfies  $\mathcal{N}(\tilde{\mathcal{H}}, \epsilon, |\cdot|) \leq 2\|O\| \left( \frac{7N_{gt}}{\epsilon} \right)^{d^{2k} N_{gt}}$ . If  $\mathcal{E}(\cdot)$  further refers to the depolarization channel  $\mathcal{E}_p(\rho) = (1-p)\rho + p\mathbb{I}/d^N$  that is applied to each quantum gate, the covering number of  $\tilde{\mathcal{H}}$  satisfies  $\mathcal{N}(\tilde{\mathcal{H}}, \epsilon, |\cdot|) \leq (1-p)^{N_g} \left( \frac{7N_{gt}\|O\|}{\epsilon} \right)^{d^{2k} N_{gt}}$ .*

Proposition 1 indicates the following insights. First, the expressivity of VQAs under general system noise setting can not be better than their ideal cases, since for both cases, the term  $N_{gt} d^{2k}$  exponentially scales the expressivity of  $\tilde{\mathcal{H}}$ . Second, the upper bounds about the expressivity given in Eq. (3) and Proposition 1 suggest that quantum noise cannot increase the expressivity of VQA compared with its ideal case [46]. Additionally, in the worst scenario where the depolarization noise is considered, the factor  $(1-p)^{N_g}$  shrinks the expressivity of  $\mathcal{H}$ . These insights enables us to compare the expressivity of different VQAs in the NISQ scenario. Meanwhile, the system noise may *forbid us* to devise high-expressive VQAs, due to the term  $(1-p)^{N_g}$ . Hence, integrating error mitigation techniques with VQAs is desired [47–52].

To better elucidate how Theorem 1 and Proposition 1 contribute to concrete quantum learning tasks, in the following, we separately explore the expressivity of QNNs and VQEs, as two crucial subclasses of VQAs.

*Expressivity of quantum neural networks.*—The aim of machine learning is devising an algorithm  $\mathcal{A}$  so that given a training dataset  $S = \{(\mathbf{x}^{(i)}, \mathbf{y}^{(i)})\}_{i=1}^n$  sampled from the domain  $\mathcal{X} \times \mathcal{Y}$ ,  $\mathcal{A}$  can use  $S$  to infer a hypothesis  $h_{\mathcal{A}(S)}^*(\cdot)$  from its hypothesis space to minimize the expected risk  $\mathcal{R}(\mathcal{A}(S)) = \mathbb{E}_{\mathbf{x}, \mathbf{y}}(\ell(h_{\mathcal{A}(S)}(\mathbf{x}), \mathbf{y}))$  [53], where the randomness is taken over  $\mathcal{A}$  and  $S$ , and  $\ell$  refers to the designated loss function. Since the probability distribution behind data space  $\mathcal{X} \times \mathcal{Y}$  is generally inaccessible, the minimization of  $\mathcal{R}(\mathcal{A}(S))$  becomes intractable. To tackle this issue, an alternative way of inferring  $h^*(\cdot)$  is minimizing the empirical risk  $\hat{\mathcal{R}}_S(\mathcal{A}(S)) = \frac{1}{n} \sum_{i=1}^n \ell(h_{\mathcal{A}(S)}(\mathbf{x}^{(i)}), \mathbf{y}^{(i)})$ .

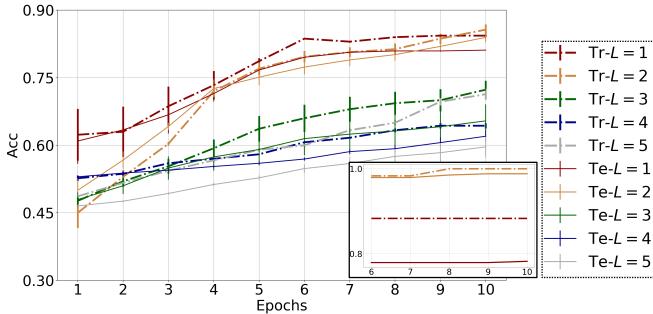
When QNN is employed to implement  $\mathcal{A}$  (as denoted by  $\mathcal{A}_{\text{QNN}}$ ) to minimize  $\hat{\mathcal{R}}_S(\mathcal{A}(S))$ , its paradigm can be cast into Eq. (1). Given the classical example  $\mathbf{x}^{(i)}$ , QNN first prepares an input quantum state  $\rho_{\mathbf{x}^{(i)}} \in \mathbb{C}^{2^N \times 2^N}$  that loads  $\mathbf{x}^{(i)}$  adopting various encoding methods [4, 54]. Once the state  $\rho_{\mathbf{x}^{(i)}}$  is prepared, the ansatz  $\hat{U}(\boldsymbol{\theta}^{(t)})$  is applied to this state, followed by a predefined quantum measurement  $O$ . To this end, the explicit form of a hypothesis for QNN is

$$h_{\mathcal{A}_{\text{QNN}}(S)}(\mathbf{x}^{(i)}) = \text{Tr} \left( \hat{U}(\boldsymbol{\theta})^\dagger O \hat{U}(\boldsymbol{\theta}) \rho_{\mathbf{x}^{(i)}} \right), \quad (5)$$

where  $\mathcal{A}_{\text{QNN}}(S) = \boldsymbol{\theta} \in \Theta$  represents the updated parameters. Since the parameter space  $\Theta$  is bounded, the hypothesis space of QNN follows

$$\mathcal{H}_{\text{QNN}} = \left\{ h_{\mathcal{A}_{\text{QNN}}(S)}(\cdot) \mid \boldsymbol{\theta} \in \Theta \right\}. \quad (6)$$

The explicit form of  $\mathcal{H}_{\text{QNN}}$  allows us to directly make use of Theorem 1 and Proposition 1 to analyze the expressivity of various QNNs. To facilitate understanding, in Appendix D, we analyze the expressivity of QNNs with typical ansätze such as hardware-efficient and tensor-network based ansätze.



**FIG. 3: Simulation results of QNN with the varied layer number  $L$ .** The label  $\text{Tr-}L = a$  (or  $\text{Te-}L = a$ ) refers to the train (or test) accuracy of QNN with layer number  $L = a$ . The outer (or inner) plot shows the statistical (top-1) results of QNNs with  $L = \{1, 2, \dots, 5\}$  ( $L = \{1, 2\}$ ). Vertical bars refer to the variance of obtained results.

Here we also explore the *generalization error* of QNNs, an important concept in quantum learning theory, which explains that when and how minimizing  $\hat{\mathcal{R}}_S(\mathcal{A}(S))$  is a sensible approach to minimizing  $\mathcal{R}(\mathcal{A}(S))$  by analyzing the upper bound of  $\mathcal{R}(\mathcal{A}(S)) - \hat{\mathcal{R}}_S(\mathcal{A}(S))$ . The generalization bound can be effectively derived when the complexity of hypothesis space is accessible [55]. Hence, we use Theorem 1 to obtain the following claim whose proof is given in Appendix C.

**Theorem 2.** *Assume the loss  $\ell$  is  $L_1$ -Lipschitz and upper bounded by  $C_1$ . Following notations in Eq. (6), for  $0 < \epsilon < 1/10$ , with probability at least  $1 - \delta$  with  $\delta \in (0, 1)$ ,*

$$\mathcal{R}(\mathcal{A}(S)) - \hat{\mathcal{R}}_S(\mathcal{A}(S)) \leq \tilde{\mathcal{O}}\left(\frac{8L_1 + C_1 + 24L_1d^k\sqrt{N_{gt}}}{\sqrt{n}}\right).$$

The employed assumption is very mild, since the loss functions adopted in QNNs are generally Lipschitz continuous and can be upper bounded by a constant  $C_1$ . This property has been broadly employed to understand the capability of QNNs [11, 12, 25, 29, 56, 57]. The achieved results provide three-fold implications. First, the generalization bound has an exponential dependence with the term  $k$  and the *sublinear* dependence with the number of trainable quantum gates  $N_{gt}$ . This observation reveals an Occam's razor principle in the quantum version [58], where parsimony of the output hypothesis implies predictive power. Second, increasing the size of training examples  $n$  contributes to an improved generalization bound. This outcome requests us to involve more training data to optimize intricate ansätze. Last, the sublinear dependence of  $N_{gt}$  may limit our result to accurately assess the generalization ability for the over-parameterized QNNs [59]. A future work is to integrate our results with deep learning theory that focuses on over-parameterized model to derive a tighten bound [60]. All of these implications can be employed as guidance to design powerful QNNs.

We conduct numerical simulations to validate our theoretical results. Specifically, we apply QNNs to accomplish

the binary classification task on the synthetic dataset  $S$ . The construction of  $S$  follows [17], where the dataset consists of 400 examples, and the feature dimension of  $\mathbf{x}^{(i)}$  is 7 and the corresponding label  $y^{(i)} \in \{0, 1\}$  is binary  $\forall i \in [400]$ . At the data preprocessing stage,  $S$  is divided into the training set and the test set with size 60 and 340, respectively. The implementation of QNN is as follows. The qubit encoding method is used to load  $\rho_{\mathbf{x}^{(i)}}$ . The the layer number  $L$  of  $\hat{U}(\boldsymbol{\theta}) = \prod_{l=1}^L U(\boldsymbol{\theta}^l)$  is varied from 1 to 5. Notably, when  $L \geq 2$ , the target concept is contained in  $\mathcal{H}_{\text{QNN}}$ . We repeat each setting with 5 times to gain statistical results. See Appendix E for construction details.

The simulation results are exhibited in Fig. 3. Although  $\mathcal{H}_{\text{QNN}}$  with the layer number  $L \in \{2, 3, 4, 5\}$  covers the target concept, the trainability, as reflected by the training accuracy in the outer plot, becomes deteriorating with respect to increased  $L$ . This result echoes with Theorem 1 in the sense that high expressivity implies poor trainability. Moreover, the discrepancy between train and test accuracy of QNN becomes large, especially for  $L = 5$ . This result accords with Proposition 1 such that higher expressivity results in larger generalization error. Eventually, in conjunction with the inner and outer plots with  $L = 1$ , we conclude that when the expressivity of  $\mathcal{H}_{\text{QNN}}$  is too small, which excludes the target concept, the training of QNN is stable but with a high empirical risk. The performance of QNNs in the NISQ case is deferred to Appendix E.

*Expressivity of variational quantum eigen-solvers.*—A central task in quantum chemistry is designing an efficient algorithm to estimate low-lying eigenstates and corresponding eigenvalues of an input Hamiltonian [61]. Variational quantum eigen-solvers (VQEs), denoted by  $\mathcal{A}_{\text{VQE}}$ , are the most popular protocols to reach this goal in the NISQ era [19], owing to their capability and flexibility. The training of VQE also adopts the iterative manner and each iteration includes two steps. Initially, VQE applies an ansatz  $U(\boldsymbol{\theta}) = \prod_{l=1}^L U_l(\boldsymbol{\theta})$  to a fixed  $N$ -qubit quantum state  $\rho_0 = (|0\rangle\langle 0|)^{\otimes N}$ , followed by measuring the Hamiltonian  $H$  to collect the classical outputs. Then, the classical optimizer utilizes the output information to update  $\boldsymbol{\theta}$  via gradient descent method to minimize  $\text{Tr}(HU(\boldsymbol{\theta})\rho_0U(\boldsymbol{\theta})^\dagger)$ . The hypothesis space of VQE can be exactly formulated by Eq. (2), i.e.,

$$\mathcal{H}_{\text{VQE}} = \{h_{\mathcal{A}_{\text{VQE}}(H)}(\rho_0) := \text{Tr}(HU(\boldsymbol{\theta})\rho_0U(\boldsymbol{\theta})^\dagger) | \boldsymbol{\theta} \in \Theta\}.$$

The form of  $\mathcal{H}_{\text{VQE}}$  enables us to efficiently measure the expressivity of an arbitrary ansatze used in VQEs by using Theorem 1 and Proposition 1. For concreteness, we quantify the expressivity of unitary coupled-cluster ansatze truncated up to single and double excitations (UCCSD) [62] whose proof is given in Appendix F.

**Corollary 1.** *Under the ideal setting, the covering number of VQE with UCCSD is upper bounded by  $\mathcal{N}(\mathcal{H}_{\text{VQE}}, \epsilon, |\cdot|) \leq \mathcal{O}(\frac{7N^5\|H\|}{\epsilon})^{d^{2k}N^5}$ . When the system noise is considered and simulated by the depolariza-*

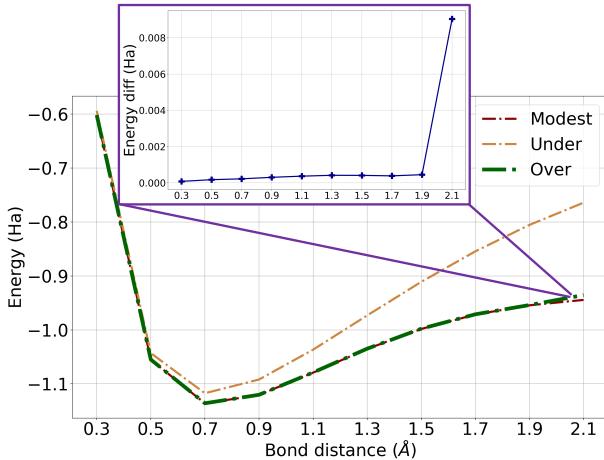


FIG. 4: **Simulation results of VQE.** The labels ‘Under’, ‘Modest’, and ‘Over’ refer to the estimated energy of VQEs when the employed ansätze have the restricted, modest, and overwhelming expressivity, respectively. ‘Ha’ (Hartrees) and ‘Å’ (Angstroms) refer to the units for energy and the bond lengths. The inner plot shows the energy gap of VQEs in ‘Over’ and ‘Modest’ cases.

tion channel, the corresponding covering number is upper bounded by  $\mathcal{N}(\widehat{\mathcal{H}}_{\text{VQE}}, \epsilon, |\cdot|) \leq \mathcal{O}((1-p)^{N^5} (\frac{7N^5\|H\|}{\epsilon})^{d^{2k}N^5})$ .

We conduct numerical simulations to explore how the expressivity of ansätze affects the performance of VQEs. In particular, we apply VQEs using three different ansätze with insufficient, modest, and overwhelming expressivity [63], and estimate the ground state energy of Hydrogen molecule with varied bond length ranging from  $0.3\text{\AA}$  to  $2.1\text{\AA}$ . As shown in Fig. 4, VQE with the restricted ansätze performs worse than the modest and overwhelming ansätze, where there exists an apparent energy gap when the bond length is larger than  $0.7\text{\AA}$ . Furthermore, although VQEs with the modest and over-

whelming ansätze demonstrate similar behavior in all bond lengths, the former always outperforms the latter as shown in the inner plot of Fig. 4. The collected results indicate that too limited or too redundant expressivity of the employed ansätze may prohibit the trainability of VQE (in Appendix G, we numerically evidence that such a claim still holds when the learning rate is allowed to be adaptive).

*Discussion and conclusions.*—We devise an efficient measure to quantify the expressivity of VQAs, including QNNs and VQEs, controlled by the qudits count, the involved quantum gates in ansätze, the operator norm of the observable, and the system noise. Compared with the prior study [35], our results allow a succinct and direct way to compare the expressivity of different ansätze and devise novel ansätze. Our work mainly concentrates on the upper bounds of expressivity, whereas a promising research direction is to derive lower bounds and tighten the expressivity quantity. The developed tool here can be extended to analyze generalization ability of other advanced QNNs such as quantum convolutional neural networks. Besides, considering that generalization bounds can be used to design an ansatz with good learning performance via the framework of structural risk minimization, it is intrigued to use our results as a theoretical guidance to devise advanced QNNs. Another crucial research direction is exploring explicit quantification of the ‘modest expressivity’ of VQAs. A deep understanding of this issue contributes to integrate various NISQ-oriented techniques such as error mitigation and quantum circuit architecture design techniques to boost the VQAs performance.

## ACKNOWLEDGMENTS

Y. Du thanks V. Schmiesing for providing valuable discussions. X. Yuan is supported by the National Natural Science Foundation of China Grant No. 12175003.

- 
- [1] Jacob Biamonte, Peter Wittek, Nicola Pancotti, Patrick Rebentrost, Nathan Wiebe, and Seth Lloyd. Quantum machine learning. *Nature*, 549(7671):195, 2017.
  - [2] Vedran Dunjko and Hans J Briegel. Machine learning & artificial intelligence in the quantum domain: a review of recent progress. *Reports on Progress in Physics*, 81(7):074001, 2018.
  - [3] Aram W Harrow and Ashley Montanaro. Quantum computational supremacy. *Nature*, 549(7671):203, 2017.
  - [4] Marcello Benedetti, Erika Lloyd, Stefan Sack, and Mattia Fiorentini. Parameterized quantum circuits as machine learning models. *Quantum Science and Technology*, 4(4):043001, 2019.
  - [5] Kishor Bharti, Alba Cervera-Lierta, Thi Ha Kyaw, Tobias Haug, Sumner Alperin-Lea, Abhinav Anand, Matthias Degroote, Hermanni Heimonen, Jakob S Kottmann, Tim Menke, et al. Noisy intermediate-scale quantum (nisq) algorithms. *arXiv preprint arXiv:2101.08448*, 2021.
  - [6] M Cerezo, Andrew Arrasmith, Ryan Babbush, Simon C Benjamin, Suguru Endo, Keisuke Fujii, Jarrod R McClean, Kosuke Mitarai, Xiao Yuan, Lukasz Cincio, et al. Variational quantum algorithms. *arXiv preprint arXiv:2012.09265*, 2020.
  - [7] Yuxuan Du, Min-Hsiu Hsieh, Tongliang Liu, and Dacheng Tao. Expressive power of parametrized quantum circuits. *Phys. Rev. Research*, 2:033125, Jul 2020.
  - [8] Suguru Endo, Zhenyu Cai, Simon C Benjamin, and Xiao Yuan. Hybrid quantum-classical algorithms and quantum error mitigation. *Journal of the Physical Society of Japan*, 90(3):032001, 2021.
  - [9] John Preskill. Quantum computing in the nisq era and

- beyond. *Quantum*, 2:79, 2018.
- [10] Yuxuan Du, Min-Hsiu Hsieh, Tongliang Liu, and Dacheng Tao. A grover-search based quantum learning scheme for classification. *New Journal of Physics*, 23(2):023020, feb 2021.
- [11] Yuxuan Du, Min-Hsiu Hsieh, Tongliang Liu, Shan You, and Dacheng Tao. On the learnability of quantum neural networks. *arXiv preprint arXiv:2007.12369*, 2020.
- [12] Hsin-Yuan Huang, Richard Kueng, and John Preskill. Information-theoretic bounds on quantum advantage in machine learning. *Physical Review Letters*, 126(19):190505, 2021.
- [13] Huitao Shen, Pengfei Zhang, Yi-Zhuang You, and Hui Zhai. Information scrambling in quantum neural networks. *Physical Review Letters*, 124(20):200504, 2020.
- [14] Yadong Wu, Juan Yao, Pengfei Zhang, and Hui Zhai. Expressivity of quantum neural networks. *arXiv preprint arXiv:2101.04273*, 2021.
- [15] Kerstin Beer, Dmytro Bondarenko, Terry Farrelly, Tobias J Osborne, Robert Salzmann, Daniel Scheiermann, and Ramona Wolf. Training deep quantum neural networks. *Nature Communications*, 11(1):1–6, 2020.
- [16] Kosuke Mitarai, Makoto Negoro, Masahiro Kitagawa, and Keisuke Fujii. Quantum circuit learning. *Physical Review A*, 98(3):032309, 2018.
- [17] Vojtěch Havlíček, Antonio D Cárcoles, Kristan Temme, Aram W Harrow, Abhinav Kandala, Jerry M Chow, and Jay M Gambetta. Supervised learning with quantum-enhanced feature spaces. *Nature*, 567(7747):209, 2019.
- [18] Kunkun Wang, Lei Xiao, Wei Yi, Shi-Ju Ran, and Peng Xue. Quantum image classifier with single photons. *arXiv preprint arXiv:2003.08551*, 2020.
- [19] Alberto Peruzzo, Jarrod McClean, Peter Shadbolt, Man-Hong Yung, Xiao-Qi Zhou, Peter J Love, Alán Aspuru-Guzik, and Jeremy L O’Brien. A variational eigenvalue solver on a photonic quantum processor. *Nature communications*, 5:4213, 2014.
- [20] Abhinav Kandala, Antonio Mezzacapo, Kristan Temme, Maika Takita, Markus Brink, Jerry M Chow, and Jay M Gambetta. Hardware-efficient variational quantum eigensolver for small molecules and quantum magnets. *Nature*, 549(7671):242–246, 2017.
- [21] Google AI Quantum et al. Hartree-fock on a superconducting qubit quantum computer. *Science*, 369(6507):1084–1089, 2020.
- [22] He-Liang Huang, Yuxuan Du, Ming Gong, Youwei Zhao, Yulin Wu, Chaoyue Wang, Shaowei Li, Futian Liang, Jin Lin, Yu Xu, et al. Experimental quantum generative adversarial networks for image generation. *arXiv preprint arXiv:2010.06201*, 2020.
- [23] Daiwei Zhu, Norbert M Linke, Marcello Benedetti, Kevin A Landsman, Nhung H Nguyen, C Huerta Alderete, Alejandro Perdomo-Ortiz, Nathan Korda, A Garfoot, Charles Brecque, et al. Training of quantum circuits on a hybrid quantum computer. *Science advances*, 5(10):eaaw9918, 2019.
- [24] M Cerezo, Akira Sone, Tyler Volkoff, Lukasz Cincio, and Patrick J Coles. Cost-function-dependent barren plateaus in shallow quantum neural networks. *arXiv preprint arXiv:2001.00550*, 2020.
- [25] Jarrod R McClean, Sergio Boixo, Vadim N Smelyanskiy, Ryan Babbush, and Hartmut Neven. Barren plateaus in quantum neural network training landscapes. *Nature communications*, 9(1):1–6, 2018.
- [26] Samson Wang, Enrico Fontana, Marco Cerezo, Kunal Sharma, Akira Sone, Lukasz Cincio, and Patrick J Coles. Noise-induced barren plateaus in variational quantum algorithms. *arXiv preprint arXiv:2007.14384*, 2020.
- [27] Kaining Zhang, Min-Hsiu Hsieh, Liu Liu, and Dacheng Tao. Toward trainability of quantum neural networks. *arXiv preprint arXiv:2011.06258*, 2020.
- [28] Ryan Sweke, Frederik Wilde, Johannes Meyer, Maria Schuld, Paul K Fährmann, Barthélémy Meynard-Piganeau, and Jens Eisert. Stochastic gradient descent for hybrid quantum-classical optimization. *Quantum*, 4:314, 2020.
- [29] Amira Abbas, David Sutter, Christa Zoufal, Aurélien Lucchi, Alessio Figalli, and Stefan Woerner. The power of quantum neural networks. *arXiv preprint arXiv:2011.00027*, 2020.
- [30] Leonardo Banchi, Jason Pereira, and Stefano Pirandola. Generalization in quantum machine learning: a quantum information perspective. *arXiv preprint arXiv:2102.08991*, 2021.
- [31] Kaifeng Bu, Dax Enshan Koh, Lu Li, Qingxian Luo, and Yaobo Zhang. On the statistical complexity of quantum circuits. *arXiv preprint arXiv:2101.06154*, 2021.
- [32] Matthias C Caro and Ishaun Datta. Pseudo-dimension of quantum circuits. *Quantum Machine Intelligence*, 2(2):1–14, 2020.
- [33] Lena Funcke, Tobias Hartung, Karl Jansen, Stefan Kühn, and Paolo Stornati. Dimensional expressivity analysis of parametric quantum circuits. *Quantum*, 5:422, 2021.
- [34] Kyle Poland, Kerstin Beer, and Tobias J Osborne. No free lunch for quantum machine learning. *arXiv preprint arXiv:2003.14103*, 2020.
- [35] Zoë Holmes, Kunal Sharma, M Cerezo, and Patrick J Coles. Connecting ansatz expressibility to gradient magnitudes and barren plateaus. *arXiv preprint arXiv:2101.02138*, 2021.
- [36] Sukin Sim, Peter D Johnson, and Alán Aspuru-Guzik. Expressibility and entangling capability of parameterized quantum circuits for hybrid quantum-classical algorithms. *Advanced Quantum Technologies*, 2(12):1900070, 2019.
- [37] Kouhei Nakaji and Naoki Yamamoto. Expressibility of the alternating layered ansatz for quantum computation. *Quantum*, 5:434, 2021.
- [38] Aram W Harrow and Richard A Low. Random quantum circuits are approximate 2-designs. *Communications in Mathematical Physics*, 291(1):257–302, 2009.
- [39] William Huggins, Piyush Patil, Bradley Mitchell, K Birgitta Whaley, and E Miles Stoudenmire. Towards quantum machine learning with tensor networks. *Quantum Science and technology*, 4(2):024001, 2019.
- [40] Vladimir Vapnik. *The nature of statistical learning theory*. Springer science & business media, 2013.
- [41] Maria Schuld, Ville Bergholm, Christian Gogolin, Josh Izaac, and Nathan Killoran. Evaluating analytic gradients on quantum hardware. *Physical Review A*, 99(3):032331, 2019.
- [42] James Stokes, Josh Izaac, Nathan Killoran, and Giuseppe Carleo. Quantum natural gradient. *Quantum*, 4:269, 2020.
- [43] David Poulin, Angie Qarry, Rolando Somma, and Frank Verstraete. Quantum simulation of time-dependent hamiltonians and the convenient illusion of hilbert space. *Physical review letters*, 106(17):170501, 2011.

- [44] Stanisław J Szarek, Elisabeth Werner, and Karol Życzkowski. How often is a random quantum state k-entangled? *Journal of Physics A: Mathematical and Theoretical*, 44(4):045303, 2010.
- [45] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2018.
- [46] We remark that to obtain a general result that covers any type of noise, a relatively loose relaxation technique is used to infer  $\mathcal{N}(\hat{\mathcal{H}}, \epsilon, |\cdot|)$ . This leads to a different scaling behavior in term of  $\|O\|$  comparing with the ideal and depolarizing cases.
- [47] Zhenyu Cai, Xiaosi Xu, and Simon C Benjamin. Mitigating coherent noise using pauli conjugation. *npj Quantum Information*, 6(1):1–9, 2020.
- [48] Yuxuan Du, Tao Huang, Shan You, Min-Hsiu Hsieh, and Dacheng Tao. Quantum circuit architecture search: error mitigation and trainability enhancement for variational quantum solvers. *arXiv preprint arXiv:2010.10217*, 2020.
- [49] Sam McArdle, Xiao Yuan, and Simon Benjamin. Error-mitigated digital quantum simulation. *Physical review letters*, 122(18):180501, 2019.
- [50] Jarrod R McClean, Zhang Jiang, Nicholas C Rubin, Ryan Babbush, and Hartmut Neven. Decoding quantum errors with subspace expansions. *Nature communications*, 11(1):1–9, 2020.
- [51] Armands Strikis, Dayue Qin, Yanzhu Chen, Simon C Benjamin, and Ying Li. Learning-based quantum error mitigation. *arXiv preprint arXiv:2005.07601*, 2020.
- [52] Jinzhao Sun, Xiao Yuan, Takahiro Tsunoda, Vlatko Vedral, Simon C Benjamin, and Suguru Endo. Mitigating realistic noise in practical noisy intermediate-scale quantum devices. *Physical Review Applied*, 15(3):034026, 2021.
- [53] Kenji Kawaguchi, Leslie Pack Kaelbling, and Yoshua Bengio. Generalization in deep learning. *arXiv preprint arXiv:1710.05468*, 2017.
- [54] Ryan LaRose and Brian Coyle. Robust data encodings for quantum classifiers. *Physical Review A*, 102(3):032420, 2020.
- [55] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. Foundations of machine learning, 2012.
- [56] Carlos Ortiz Marrero, Mária Kieferová, and Nathan Wiebe. Entanglement induced barren plateaus. *arXiv preprint arXiv:2010.15968*, 2020.
- [57] Ryan Sweke, Frederik Wilde, Johannes Meyer, Maria Schuld, Paul K Fährmann, Barthélémy Meynard-Piganeau, and Jens Eisert. Stochastic gradient descent for hybrid quantum-classical optimization. *Quantum*, 4:314, 2020.
- [58] Anselm Blumer, Andrzej Ehrenfeucht, David Haussler, and Manfred K Warmuth. Occam’s razor. *Information processing letters*, 24(6):377–380, 1987.
- [59] Martin Larocca, Nathan Ju, Diego García-Martín, Patrick J. Coles, and M. Cerezo. Theory of over-parametrization in quantum neural networks, 2021.
- [60] Abdulkadir Canatar, Blake Bordelon, and Cengiz Pehlevan. Spectral bias and task-model alignment explain generalization in kernel regression and infinitely wide neural networks. *Nature communications*, 12(1):1–12, 2021.
- [61] Sam McArdle, Suguru Endo, Alan Aspuru-Guzik, Simon C Benjamin, and Xiao Yuan. Quantum computational chemistry. *Reviews of Modern Physics*, 92(1):015003, 2020.
- [62] Yudong Cao, Jonathan Romero, Jonathan P Olson, Matthias Degroote, Peter D Johnson, Mária Kieferová, Ian D Kivlichan, Tim Menke, Borja Peropadre, Nicolas PD Sawaya, et al. Quantum chemistry in the age of quantum computing. *Chemical reviews*, 119(19):10856–10915, 2019.
- [63] The separated expressivity of different ansatze is completed by controlling the involved number of quantum gates, supported by Theorem 1. See Appendix G for construction details.
- [64] Thomas Barthel and Jianfeng Lu. Fundamental limitations for measurements in quantum many-body systems. *Phys. Rev. Lett.*, 121:080406, Aug 2018.
- [65] Michael A Nielsen and Isaac L Chuang. *Quantum computation and quantum information*. Cambridge University Press, 2010.
- [66] Sham M. Kakade, K. Sridharan, and Ambuj Tewari. On the complexity of linear prediction: Risk bounds, margin bounds, and regularization. In *NIPS*, 2008.
- [67] Richard M Dudley. The sizes of compact subsets of hilbert space and continuity of gaussian processes. *Journal of Functional Analysis*, 1(3):290–330, 1967.
- [68] Jonathan Romero, Ryan Babbush, Jarrod R McClean, Cornelius Hempel, Peter J Love, and Alán Aspuru-Guzik. Strategies for quantum computing molecular energies using the unitary coupled cluster ansatz. *Quantum Science and Technology*, 4(1):014008, 2018.
- [69] Sergey B Bravyi and Alexei Yu Kitaev. Fermionic quantum computation. *Annals of Physics*, 298(1):210–226, 2002.
- [70] Jarrod R McClean, Nicholas C Rubin, Kevin J Sung, Ian D Kivlichan, Xavier Bonet-Monroig, Yudong Cao, Chengyu Dai, E Schuyler Fried, Craig Gidney, Brendan Gimby, et al. Openfermion: the electronic structure package for quantum computers. *Quantum Science and Technology*, 5(3):034014, 2020.
- [71] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.
- [72] Matthias C Caro, Elies Gil-Fuster, Johannes Jakob Meyer, Jens Eisert, and Ryan Sweke. Encoding-dependent generalization bounds for parametrized quantum circuits. *arXiv preprint arXiv:2106.03880*, 2021.
- [73] Casper Gyurik, Dyon van Vreumingen, and Vedran Dunjko. Structural risk minimization for quantum linear classifiers. *arXiv preprint arXiv:2105.05566*, 2021.
- [74] M Bilkis, M Cerezo, Guillaume Verdon, Patrick J Coles, and Lukasz Cincio. A semi-agnostic ansatz with variable structure for quantum machine learning. *arXiv preprint arXiv:2103.06712*, 2021.
- [75] Harper R Grimsley, Sophia E Economou, Edwin Barnes, and Nicholas J Mayhall. An adaptive variational algorithm for exact molecular simulations on a quantum computer. *Nature communications*, 10(1):1–9, 2019.
- [76] En-Jui Kuo, Yao-Lung L Fang, and Samuel Yen-Chi Chen. Quantum architecture search via deep reinforcement learning. *arXiv preprint arXiv:2104.07715*, 2021.
- [77] Mateusz Ostaszewski, Edward Grant, and Marcello Benedetti. Structure optimization for parameterized quantum circuits. *Quantum*, 5:391, 2021.
- [78] Ho Lun Tang, VO Shkolnikov, George S Barron,

- Harper R Grimsley, Nicholas J Mayhall, Edwin Barnes, and Sophia E Economou. qubit-adapt-vqe: An adaptive algorithm for constructing hardware-efficient ansätze on a quantum processor. *PRX Quantum*, 2(2):020310, 2021.
- [79] Shi-Xin Zhang, Chang-Yu Hsieh, Shengyu Zhang, and Hong Yao. Neural predictor based quantum architecture search. *arXiv preprint arXiv:2103.06524*, 2021.
- [80] Samira Pouyanfar, Saad Sadiq, Yilin Yan, Haiman Tian, Yudong Tao, Maria Presa Reyes, Mei-Ling Shyu, Shu-Ching Chen, and Sundaraja S Iyengar. A survey on deep learning: Algorithms, techniques, and applications. *ACM Computing Surveys (CSUR)*, 51(5):1–36, 2018.
- [81] Ruoyu Sun. Optimization for deep learning: theory and algorithms. *arXiv preprint arXiv:1912.08957*, 2019.
- [82] Peter Bartlett, Dylan Foster, and Matus Telgarsky. Spectrally-normalized margin bounds for neural networks. *Advances in Neural Information Processing Systems*, 30:6241–6250, 2017.
- [83] Nico JD Nagelkerke et al. A note on a general definition of the coefficient of determination. *Biometrika*, 78(3):691–692, 1991.

## Appendix A: Proof of Theorem 1

The proof of Theorem 1 employs the definition of the operator norm.

**Definition 2** (Operator norm). *Suppose  $A$  is an  $n \times n$  matrix. The operator norm  $\|A\|$  is defined as*

$$\|A\| = \sup_{\|\mathbf{x}\|_2=1, \mathbf{x} \in \mathbb{C}^n} \|A\mathbf{x}\|. \quad (\text{A1})$$

Alternatively,  $\|A\| = \sqrt{\lambda_1(AA^\dagger)}$ , where  $\lambda_i(AA^\dagger)$  is the  $i$ -th largest eigenvalue of the matrix  $AA^\dagger$ .

Besides the above definition, the proof of Theorem 1 leverages the the following two lemmas. In particular, The first lemma enables us to employ the covering number of one metric space  $(\mathcal{H}_1, d_1)$  to bound the covering number of an another metric space  $(\mathcal{H}_2, d_2)$ .

**Lemma 1** (Lemma 5, [64]). *Let  $(\mathcal{H}_1, d_1)$  and  $(\mathcal{H}_2, d_2)$  be metric spaces and  $f : \mathcal{H}_1 \rightarrow \mathcal{H}_2$  be bi-Lipschitz such that*

$$d_2(f(\mathbf{x}), f(\mathbf{y})) \leq Kd_1(\mathbf{x}, \mathbf{y}), \quad \forall \mathbf{x}, \mathbf{y} \in \mathcal{H}_1, \quad (\text{A2})$$

and

$$d_2(f(\mathbf{x}), f(\mathbf{y})) \geq kd_1(\mathbf{x}, \mathbf{y}), \quad \forall \mathbf{x}, \mathbf{y} \in \mathcal{H}_1 \text{ with } d_1(\mathbf{x}, \mathbf{y}) \leq r. \quad (\text{A3})$$

Then their covering numbers obey

$$\mathcal{N}(\mathcal{H}_1, 2\epsilon/k, d_1) \leq \mathcal{N}(\mathcal{H}_2, \epsilon, d_2) \leq \mathcal{N}(\mathcal{H}_1, \epsilon/K, d_1), \quad (\text{A4})$$

where the left inequality requires  $\epsilon \leq kr/2$ .

The second lemma presents the covering number of the operator group

$$\mathcal{H}_{circ} := \left\{ \hat{U}(\boldsymbol{\theta})^\dagger O \hat{U}(\boldsymbol{\theta}) \mid \boldsymbol{\theta} \in \Theta \right\}, \quad (\text{A5})$$

where  $\hat{U}(\boldsymbol{\theta}) = \prod_{i=1}^{N_g} \hat{u}_i(\boldsymbol{\theta}_i)$  and only  $N_{gt} \leq N_g$  gates in  $U(\boldsymbol{\theta})$  are trainable. The detailed proof is deferred to Appendix A 1.

**Lemma 2.** *Following notations in Theorem 1, suppose that the employed  $N$ -qubit Ansätze containing in total  $N_g$  gates with  $N_g > N$ , each gate  $\hat{u}_i(\boldsymbol{\theta})$  acting on at most  $k$  qudits, and  $N_{gt} \leq N_g$  gates in  $U(\boldsymbol{\theta})$  are trainable. The  $\epsilon$ -covering number for the operator group  $\mathcal{H}_{circ}$  in Eq. (A5) with respect to the operator-norm distance obeys*

$$\mathcal{N}(\mathcal{H}_{circ}, \epsilon, \|\cdot\|) \leq \left( \frac{7N_{gt}\|O\|}{\epsilon} \right)^{d^{2k}N_{gt}}, \quad (\text{A6})$$

where  $\|O\|$  denotes the operator norm of  $O$ .

We are now ready to present the proof of Theorem 1.

*Proof of Theorem 1.* The intuition of the proof is as follows. Recall the definition of the hypothesis space  $\mathcal{H}$  in Eq. (2) and Lemma 1. When  $\mathcal{H}_1$  refers to the hypothesis space  $\mathcal{H}$  and  $\mathcal{H}_2$  refers to the unitary group  $\mathcal{U}(d^N)$ , the upper bound of the covering number of  $\mathcal{H}$ , i.e.,  $\mathcal{N}(\mathcal{H}_1, d_1, \epsilon)$ , can be derived by first quantifying  $K$  Eq. (A2), and then interacting with  $\mathcal{N}(\mathcal{H}_{circ}, \epsilon, \|\cdot\|)$  in Lemma 2. Under the above observations, in the following, we analyze the upper bound of the covering number  $\mathcal{N}(\mathcal{H}, \epsilon, |\cdot|)$ .

We now derive the Lipschitz constant  $K$  in Eq. (A2), as the precondition to achieve the upper bound of  $\mathcal{N}(\mathcal{H}, \epsilon, |\cdot|)$ . Define  $\hat{U} \in \mathcal{U}(d^N)$  as the employed Ansätze composed of  $N_g$  gates, i.e.,  $\hat{U} = \prod_{i=1}^{N_g} \hat{u}_i$ . Let  $\hat{U}_\epsilon$  be the quantum circuit where each of the  $N_g$  gates is replaced by the nearest element in the covering set. The relation between the distance  $d_2(\text{Tr}(\hat{U}_\epsilon^\dagger O \hat{U}_\epsilon \rho), \text{Tr}(\hat{U}^\dagger O \hat{U} \rho))$  and the distance  $d_1(\hat{U}_\epsilon, \hat{U})$  yields

$$\begin{aligned} & d_2(\text{Tr}(\hat{U}_\epsilon^\dagger O \hat{U}_\epsilon \rho), \text{Tr}(\hat{U}^\dagger O \hat{U} \rho)) \\ &= |\text{Tr}(\hat{U}_\epsilon^\dagger O \hat{U}_\epsilon \rho) - \text{Tr}(\hat{U}^\dagger O \hat{U} \rho)| \\ &= |\text{Tr}((\hat{U}_\epsilon^\dagger O \hat{U}_\epsilon - \hat{U}^\dagger O \hat{U})\rho)| \\ &\leq \|\hat{U}_\epsilon^\dagger O \hat{U}_\epsilon - \hat{U}^\dagger O \hat{U}\| \text{Tr}(\rho) \\ &= d_1(\hat{U}_\epsilon^\dagger O \hat{U}_\epsilon, \hat{U}^\dagger O \hat{U}), \end{aligned} \quad (\text{A7})$$

where the first equality comes from the explicit form of hypothesis, the first inequality uses the Cauchy-Schwartz inequality, and the last inequality employs  $\text{Tr}(\rho) = 1$  and

$$\left\| \hat{U}_\epsilon^\dagger O \hat{U}_\epsilon - \hat{U}^\dagger O \hat{U} \right\| = d_1(\hat{U}_\epsilon^\dagger O \hat{U}_\epsilon, \hat{U}^\dagger O \hat{U}). \quad (\text{A8})$$

The above equation indicates  $K = 1$ . Combining the above result with Lemma 1 (i.e., Eq. (A2)) and Lemma 2, we obtain

$$\mathcal{N}(\mathcal{H}, \epsilon, |\cdot|) \leq \mathcal{N}(\mathcal{H}_{circ}, \epsilon, \|\cdot\|) \leq \left( \frac{7N_{gt}\|O\|}{\epsilon} \right)^{d^{2k}N_{gt}}. \quad (\text{A9})$$

This relation ensures

$$\mathcal{N}(\mathcal{H}, \epsilon, |\cdot|) \leq \left( \frac{7N_{gt}\|O\|}{\epsilon} \right)^{d^{2k}N_{gt}}. \quad (\text{A10})$$

□

### 1. Proof of Lemma 2

The proof of Lemma 2 exploits the following result.

**Lemma 3** (Lemma 1, [64]). *For  $0 < \epsilon < 1/10$ , the  $\epsilon$ -covering number for the unitary group  $U(d^k)$  with respect to the operator-norm distance in Definition 2 obeys*

$$\left( \frac{3}{4\epsilon} \right)^{d^{2k}} \leq \mathcal{N}(U(d^k), \epsilon, \|\cdot\|) \leq \left( \frac{7}{\epsilon} \right)^{d^{2k}}. \quad (\text{A11})$$

*Proof of Lemma 2.* The goal of Lemma 2 is to measure the covering number of the operator group  $\mathcal{H}_{circ} = \{\hat{U}(\boldsymbol{\theta})^\dagger O \hat{U}(\boldsymbol{\theta}) | \boldsymbol{\theta} \in \Theta\}$  in Eq. (A5), where the trainable unitary  $\hat{U}(\boldsymbol{\theta}) = \prod_{i=1}^{N_g} \hat{u}_i(\boldsymbol{\theta}_i)$  consists of  $N_{gt}$  trainable gates and  $N_g - N_{gt}$  fixed gates. To achieve this goal, we consider a fixed  $\epsilon$ -covering  $\tilde{\mathcal{S}}$  for the set  $\mathcal{N}(U(d^k), \epsilon, \|\cdot\|)$  of all possible gates and define the set

$$\tilde{\mathcal{S}} := \left\{ \prod_{i \in \{N_{gt}\}} \hat{u}_i(\boldsymbol{\theta}_i) \mid \prod_{j \in \{N_g - N_{gt}\}} \hat{u}_j \mid \hat{u}_i(\boldsymbol{\theta}_i) \in \mathcal{S} \right\}, \quad (\text{A12})$$

where  $\hat{u}_i(\boldsymbol{\theta}_i)$  and  $\hat{u}_j$  specify to the trainable and fixed quantum gates in the employed Ansätze, respectively. Note that for any circuit  $\hat{U}(\boldsymbol{\theta}) = \prod_{i=1}^{N_g} \hat{u}_i(\boldsymbol{\theta}_i)$ , we can always find a  $\hat{U}_\epsilon(\boldsymbol{\theta}) \in \tilde{\mathcal{S}}$  where each  $\hat{u}_i(\boldsymbol{\theta}_i)$  of trainable gates is replaced with the nearest element in the covering set  $\mathcal{S}$ , and the discrepancy  $\|\hat{U}(\boldsymbol{\theta})^\dagger O \hat{U}(\boldsymbol{\theta}) - \hat{U}_\epsilon(\boldsymbol{\theta})^\dagger O \hat{U}_\epsilon(\boldsymbol{\theta})\|$  satisfies

$$\begin{aligned} & \|\hat{U}(\boldsymbol{\theta})^\dagger O \hat{U}(\boldsymbol{\theta}) - \hat{U}_\epsilon(\boldsymbol{\theta})^\dagger O \hat{U}_\epsilon(\boldsymbol{\theta})\| \\ & \leq \|\hat{U} - \hat{U}_\epsilon\| \|O\| \\ & \leq N_{gt} \|O\| \epsilon, \end{aligned} \quad (\text{A13})$$

where the first inequality uses the triangle inequality, and the second inequality follows from  $\|\hat{U} - \hat{U}_\epsilon\| \leq N_{gt}\epsilon$ .

Therefore, by Definition 1, we know that  $\tilde{\mathcal{S}}$  is a  $N_{gt}\|O\|\epsilon$ -covering set for  $\mathcal{H}_{circ}$ . Recall that the upper bound in Lemma 3 gives  $|\mathcal{S}| \leq \left(\frac{7}{\epsilon}\right)^{d^{2k}}$ . Since there are  $|\mathcal{S}|^{N_{gt}}$  combinations for the gates in  $\tilde{\mathcal{S}}$ , we have  $|\tilde{\mathcal{S}}| \leq \left(\frac{7}{\epsilon}\right)^{d^{2k}N_{gt}}$  and the covering number for  $\mathcal{H}_{circ}$  satisfies

$$\mathcal{N}(\mathcal{H}_{circ}, N_{gt}\|O\|\epsilon, \|\cdot\|) \leq \left( \frac{7}{\epsilon} \right)^{d^{2k}N_{gt}}. \quad (\text{A14})$$

An equivalent representation of the above inequality is

$$\mathcal{N}(\mathcal{H}_{circ}, \epsilon, \|\cdot\|) \leq \left( \frac{7N_{gt}\|O\|}{\epsilon} \right)^{d^{2k}N_{gt}}. \quad (\text{A15})$$

□

## Appendix B: Proof of Proposition 1

*Proof of Proposition 1.* In this proof, we first derive the covering number of VQA for the general noisy quantum channel  $\mathcal{E}(\cdot)$ , and then analyze the covering number of VQA when  $\mathcal{E}(\cdot)$  specifies to the depolarization noise.

The general quantum channel  $\mathcal{E}(\cdot)$ . We follow the same routine as the proof of Theorem 1 to acquire the upper bound of  $\mathcal{N}(\tilde{\mathcal{H}}, \epsilon, |\cdot|)$ . Namely, supported by Lemma 1, once we establish the relation between  $d_2(\text{Tr}(O\mathcal{E}(\hat{U}\rho\hat{U}^\dagger)), \text{Tr}(O\mathcal{E}(\hat{U}_\epsilon\rho\hat{U}_\epsilon^\dagger)))$ , i.e.,

$$d_2\left(\text{Tr}\left(O\mathcal{E}\left(\hat{U}\rho\hat{U}^\dagger\right)\right), \text{Tr}\left(O\mathcal{E}\left(\hat{U}_\epsilon\rho\hat{U}_\epsilon^\dagger\right)\right)\right) = \left|\text{Tr}\left(O\mathcal{E}\left(\hat{U}_\epsilon\rho\hat{U}_\epsilon^\dagger\right) - O\mathcal{E}\left(\hat{U}\rho\hat{U}^\dagger\right)\right)\right|, \quad (\text{B1})$$

and  $d_1(\hat{U}_\epsilon\rho\hat{U}_\epsilon^\dagger, \hat{U}\rho\hat{U}^\dagger)$ , the covering number  $\mathcal{N}(\mathcal{H}_{circ}, \epsilon, \|\cdot\|)$  in Lemma 2 can be utilized to infer the upper bound of  $\mathcal{N}(\tilde{\mathcal{H}}, \epsilon, |\cdot|)$ .

Under the above observation, we now derive the term  $K$  such that  $d_2(\text{Tr}(O\mathcal{E}(\hat{U}\rho\hat{U}^\dagger)), \text{Tr}(O\mathcal{E}(\hat{U}_\epsilon\rho\hat{U}_\epsilon^\dagger))) \leq Kd_1(\hat{U}_\epsilon\rho\hat{U}_\epsilon^\dagger, \hat{U}\rho\hat{U}^\dagger)$ . In particular, we have

$$\begin{aligned} & d_2\left(\text{Tr}\left(O\mathcal{E}\left(\hat{U}\rho\hat{U}^\dagger\right)\right), \text{Tr}\left(O\mathcal{E}\left(\hat{U}_\epsilon\rho\hat{U}_\epsilon^\dagger\right)\right)\right) \\ &= \left|\text{Tr}\left(O\left(\mathcal{E}\left(\hat{U}_\epsilon\rho\hat{U}_\epsilon^\dagger\right) - \mathcal{E}\left(\hat{U}\rho\hat{U}^\dagger\right)\right)\right)\right| \\ &\leq \|O\| \left|\text{Tr}\left(\mathcal{E}\left(\hat{U}_\epsilon\rho\hat{U}_\epsilon^\dagger\right) - \mathcal{E}\left(\hat{U}\rho\hat{U}^\dagger\right)\right)\right| \\ &\leq \|O\| \left|\text{Tr}\left(\hat{U}_\epsilon\rho\hat{U}_\epsilon^\dagger - \hat{U}\rho\hat{U}^\dagger\right)\right| \\ &\leq 2\|O\| \left\|\hat{U}_\epsilon\rho\hat{U}_\epsilon^\dagger - \hat{U}\rho\hat{U}^\dagger\right\|, \end{aligned} \quad (\text{B2})$$

where the first inequality uses the Cauchy-Schwartz inequality, the second inequality employs the contractive property of quantum channels (Theorem 9.2, [65]), the last inequality comes from the fact that  $\hat{U}_\epsilon\rho\hat{U}_\epsilon^\dagger$  and  $\hat{U}\rho\hat{U}^\dagger$  are two rank-1 states (i.e., this implies that the rank of  $\hat{U}_\epsilon\rho\hat{U}_\epsilon^\dagger - \hat{U}\rho\hat{U}^\dagger$  is at most 2) and  $\text{Tr}(\cdot) \leq \text{rank}(\cdot)\|\cdot\|$ .

With setting the operator  $O$  in  $d_1$  as  $\rho$ , we obtain

$$d_2\left(\text{Tr}\left(O\mathcal{E}\left(\hat{U}\rho\hat{U}^\dagger\right)\right), \text{Tr}\left(O\mathcal{E}\left(\hat{U}_\epsilon\rho\hat{U}_\epsilon^\dagger\right)\right)\right) \leq \|O\| 2d_1(\hat{U}_\epsilon\rho\hat{U}_\epsilon^\dagger, U\rho U^\dagger), \quad (\text{B3})$$

which indicates that the term  $K$  in Eq. (A3) is

$$K = 2\|O\|. \quad (\text{B4})$$

Supporting by Lemma 2, the covering number of VQA under the noisy setting is upper bounded by

$$\mathcal{N}(\tilde{\mathcal{H}}, \epsilon, |\cdot|) \leq 2\|O\| \left(\frac{7N_{gt}\|\rho\|}{\epsilon}\right)^{d^{2k}N_{gt}} = 2\|O\| \left(\frac{7N_{gt}}{\epsilon}\right)^{d^{2k}N_{gt}}, \quad (\text{B5})$$

where the equality exploits the spectral property of the quantum state.

The local depolarization channel  $\mathcal{E}_p(\cdot)$ . We next consider the covering number of VQA when the noisy quantum channel is simulated by the local depolarization noise, i.e., the depolarization channel  $\mathcal{E}_p(\cdot)$  is applied to each quantum gate in  $\hat{U}(\boldsymbol{\theta})$ . Following the explicit form of the depolarization channel, the distance  $d_2(\text{Tr}(O\mathcal{E}_p(\hat{U}\rho\hat{U}^\dagger)), \text{Tr}(O\mathcal{E}_p(\hat{U}_\epsilon\rho\hat{U}_\epsilon^\dagger)))$  and distance  $d_1(\hat{U}_\epsilon\rho\hat{U}_\epsilon^\dagger, \hat{U}\rho\hat{U}^\dagger)$  satisfies

$$\begin{aligned} & d_2\left(\text{Tr}\left(O\mathcal{E}_p\left(\hat{U}\rho\hat{U}^\dagger\right)\right), \text{Tr}\left(O\mathcal{E}_p\left(\hat{U}_\epsilon\rho\hat{U}_\epsilon^\dagger\right)\right)\right) \\ &= \left|\text{Tr}\left(O\mathcal{E}_p\left(\hat{U}_\epsilon\rho\hat{U}_\epsilon^\dagger\right) - O\mathcal{E}_p\left(\hat{U}\rho\hat{U}^\dagger\right)\right)\right| \\ &= (1-p)^{N_g} \left|\text{Tr}\left(O\left(\hat{U}_\epsilon\rho\hat{U}_\epsilon^\dagger\right) - O\left(\hat{U}\rho\hat{U}^\dagger\right)\right)\right| \\ &\leq (1-p)^{N_g} \left\|\hat{U}_\epsilon^\dagger O \hat{U}_\epsilon - \hat{U}^\dagger O \hat{U}\right\| \text{Tr}(\rho) \\ &= (1-p)^{N_g} \left\|\hat{U}_\epsilon^\dagger O \hat{U}_\epsilon - \hat{U}^\dagger O \hat{U}\right\| \\ &= (1-p)^{N_g} d_1(\hat{U}_\epsilon^\dagger O \hat{U}_\epsilon, \hat{U}^\dagger O \hat{U}), \end{aligned} \quad (\text{B6})$$

where the second equality comes from the property of the local depolarization noise given in [11, Lemma 5], i.e.,

$$\mathcal{E}_p\left(\hat{U}\rho\hat{U}^\dagger\right) = \mathcal{E}_p(u_{N_g}(\boldsymbol{\theta})\dots u_2(\boldsymbol{\theta})\mathcal{E}_p(u_1(\boldsymbol{\theta})\rho u_1(\boldsymbol{\theta})^\dagger)u_2(\boldsymbol{\theta})^\dagger\dots u_{N_g}(\boldsymbol{\theta})^\dagger) = (1-p)^{N_g}(\hat{U}(\boldsymbol{\theta})\rho\hat{U}(\boldsymbol{\theta})^\dagger) + (1-(1-p)^{N_g})\frac{\mathbb{I}}{N^d}. \quad (\text{B7})$$

This result indicates that the term  $K$  in Eq. (A2) is

$$K = (1-p)^{N_g}. \quad (\text{B8})$$

Supporting by Lemma 2, the covering number of VQA under the depolarization noise is upper bounded by

$$\mathcal{N}(\tilde{\mathcal{H}}, \epsilon, |\cdot|) \leq (1-p)^{N_g} \left( \frac{7N_{gt}\|O\|}{\epsilon} \right)^{d^{2k}N_{gt}}. \quad (\text{B9})$$

□

### Appendix C: Proof of Theorem 2

**Lemma 4** (Theorem 1, [66]). *Assume the loss  $\ell$  is  $L_1$ -Lipschitz and upper bounded by  $C_1$ . With probability at least  $1 - \delta$  over a sample  $\mathcal{S}$  of size  $n$ , every  $h \in \mathcal{H}_{\text{QNN}}$  satisfies*

$$\mathcal{R}(\mathcal{A}(S)) \leq \hat{\mathcal{R}}_S(\mathcal{A}(S)) + 2L_1\mathfrak{R}(\mathcal{H}_{\text{QNN}}) + 3C_1\sqrt{\frac{\ln(2/\delta)}{2n}}, \quad (\text{C1})$$

where  $\mathfrak{R}(\mathcal{H}_{\text{QNN}})$  is the empirical Rademacher complexity of the hypothesis space  $\mathcal{H}_{\text{QNN}}$  and  $n$  is the sample size of  $\mathcal{S}$ .

*Proof of Theorem 2.* The result of Lemma 4 indicates that the precondition to infer the generalization error is deriving the upper bound of the Rademacher complexity  $\mathfrak{R}(\mathcal{H}_{\text{QNN}})$ . The mathematical expression of the Rademacher complexity is

$$\mathfrak{R}(\mathcal{H}_{\text{QNN}}) := n^{-1}\mathbb{E}\left(\sup_{h \in \mathcal{H}} \sum_{i=1}^n \epsilon_i h(x_i, y_i)\right), \quad (\text{C2})$$

where the expectation is over the Rademacher random variables  $(\epsilon_1, \dots, \epsilon_n)$ , which are i.i.d with  $\Pr[\epsilon_1 = 1] = \Pr[\epsilon_1 = -1] = 1/2$ . To achieve this goal, we employ the Dudley entropy integral bound [67] to connect Rademacher complexity with covering number, i.e.,

$$\mathfrak{R}(\mathcal{H}_{\text{QNN}}) \leq \inf_{\alpha > 0} \left( 4\alpha + \frac{12}{\sqrt{n}} \int_{\alpha}^1 \sqrt{\ln \mathcal{N}((\mathcal{H}_{\text{QNN}})_{|S}, \epsilon, \|\cdot\|_2)} d\epsilon \right), \quad (\text{C3})$$

where  $(\mathcal{H}_{\text{QNN}})_{|S}$  denotes the set of vectors formed by the hypothesis with  $n$  examples, i.e.,  $\{[h_{\mathcal{A}_{\text{QNN}}(S)}(\mathbf{x}^{(i)})]_{i=1:n} \mid \boldsymbol{\theta} \in \Theta\}$ .

We first establish the relation between the covering number of  $(\mathcal{H}_{\text{QNN}})_{|S}$  and  $(\mathcal{H}_{\text{QNN}})_{|\mathbf{x}^{(i)}}$  to derive the upper bound of  $\ln \mathcal{N}((\mathcal{H}_{\text{QNN}})_{|S}, \epsilon, \|\cdot\|_2)$ . As with Lemma 2, denote a fixed  $(\epsilon/\sqrt{n})$ -covering  $\mathcal{S}$  for the set  $(\mathcal{H}_{\text{QNN}})_{|\mathbf{x}^{(i)}}$ . Then for any function  $h_{\mathcal{A}_{\text{QNN}}(S)}(\cdot) \in \mathcal{H}_{\text{QNN}}$  in Eq. (6), we can always find a  $h'_{\mathcal{A}_{\text{QNN}}(S)}(\mathbf{x}^{(i)}) \in \mathcal{S}$  such that  $\forall i \in [n]$ ,  $|h_{\mathcal{A}_{\text{QNN}}(S)}(\mathbf{x}^{(i)}) - h'_{\mathcal{A}_{\text{QNN}}(S)}(\mathbf{x}^{(i)})| \leq \epsilon/\sqrt{n}$ , and the discrepancy  $\|[h_{\mathcal{A}_{\text{QNN}}(S)}(\mathbf{x}^{(i)})]_{i=1:n} - [h'_{\mathcal{A}_{\text{QNN}}(S)}(\mathbf{x}^{(i)})]_{i=1:n}\|_2$  satisfies

$$\begin{aligned} & \left\| [h_{\mathcal{A}_{\text{QNN}}(S)}(\mathbf{x}^{(i)})]_{i=1:n} - [h'_{\mathcal{A}_{\text{QNN}}(S)}(\mathbf{x}^{(i)})]_{i=1:n} \right\|_2 \\ &= \sqrt{\sum_{i=1}^n |h_{\mathcal{A}_{\text{QNN}}(S)}(\mathbf{x}^{(i)}) - h'_{\mathcal{A}_{\text{QNN}}(S)}(\mathbf{x}^{(i)})|^2} \\ &\leq \epsilon. \end{aligned} \quad (\text{C4})$$

Therefore, by Definition 1, we know that  $\mathcal{S}$  is a  $\epsilon$ -covering set for  $(\mathcal{H}_{\text{QNN}})_{|S}$ . This result gives

$$\ln(\mathcal{N}((\mathcal{H}_{\text{QNN}})_{|S}, \epsilon, \|\cdot\|_2)) \leq \ln\left(\mathcal{N}\left((\mathcal{H}_{\text{QNN}})_{|\mathbf{x}^{(i)}}, \frac{\epsilon}{\sqrt{n}}, |\cdot|\right)\right). \quad (\text{C5})$$

The right hand-side in Eq. (C5) can be further upper bounded as

$$\begin{aligned}
& \ln \left( \mathcal{N} \left( (\mathcal{H}_{\text{QNN}})_{|\boldsymbol{x}^{(i)}}, \frac{\epsilon}{\sqrt{n}}, |\cdot| \right) \right) \\
& \leq \ln \left( \left( \frac{7\sqrt{n}N_{gt}\|O\|}{\epsilon} \right)^{d^{2k}N_{gt}} \right) \\
& = d^{2k}N_{gt} \ln \left( \frac{7\sqrt{n}N_{gt}\|O\|}{\epsilon} \right), \tag{C6}
\end{aligned}$$

where the first inequality can be easily derived based on the proof of Lemma 2 and the second inequality uses the result of Theorem 1. To this end, the integration term in Eq. (C3) follows

$$\begin{aligned}
& \frac{12}{\sqrt{n}} \int_{\alpha}^1 \sqrt{\ln \mathcal{N}((\mathcal{H}_{\text{QNN}})_{|S}, \epsilon, \|\cdot\|_2)} d\epsilon \\
& \leq \frac{12}{\sqrt{n}} \int_{\alpha}^1 \sqrt{d^{2k}N_{gt} \ln \left( \frac{7\sqrt{n}N_{gt}\|O\|}{\epsilon} \right)} d\epsilon \\
& \leq \frac{12}{\sqrt{n}} \int_{\alpha}^1 d^k \sqrt{N_{gt}} \ln \left( \frac{7\sqrt{n}N_{gt}\|O\|}{\epsilon} \right) d\epsilon \\
& = \frac{12}{\sqrt{n}} d^k \sqrt{N_{gt}} \epsilon \left( \ln \left( \frac{7\sqrt{n}N_{gt}\|O\|}{\epsilon} \right) + 1 \right) \Big|_{\epsilon=\alpha}^1 \\
& = \frac{12}{\sqrt{n}} d^k \sqrt{N_{gt}} \left( \ln (7\sqrt{n}N_{gt}\|O\|) + 1 \right) - \frac{12}{\sqrt{n}} d^k \sqrt{N_{gt}} \alpha \left( \ln \left( \frac{7\sqrt{n}N_{gt}\|O\|}{\alpha} \right) + 1 \right) \tag{C7}
\end{aligned}$$

where the first inequality employs the upper bound of the covering number of  $\mathcal{H}_{\text{QNN}}$  in Theorem 2, and the second inequality uses the monotony of integral.

For simplicity, we set  $\alpha = 1/\sqrt{n}$  in Eq. (C3) and then the Rademacher complexity  $\mathfrak{R}(\mathcal{H}_{\text{QNN}})$  is upper bounded by

$$\mathfrak{R}(\mathcal{H}_{\text{QNN}}) \leq \frac{4}{\sqrt{n}} + \frac{12}{\sqrt{n}} d^k \sqrt{N_{gt}} \left( \ln (7\sqrt{n}N_{gt}\|O\|) + 1 \right). \tag{C8}$$

In conjunction Lemma 4 with Eq. (C8), with probability  $1 - \delta$ , the generalization bound of QNN yields

$$\mathcal{R}(\mathcal{A}(S)) - \hat{\mathcal{R}}_S(\mathcal{A}(S)) \leq \frac{8L_1}{\sqrt{n}} + \frac{24L_1}{\sqrt{n}} d^k \sqrt{N_{gt}} \left( \ln (7\sqrt{n}N_{gt}\|O\|) + 1 \right) + 3C_1 \sqrt{\frac{\ln(1/\delta)}{2n}}. \tag{C9}$$

□

#### Appendix D: Expressivity of other advanced quantum neural networks

To better understand how the covering number effects the expressivity of VQAs, in this section, we explicitly quantify the covering number of QNNs with several representative Ansätze, i.e., the hardware-efficient Ansätze, the tensor-network based Ansätze with the matrix product state structure, and the tensor-network based Ansätze with the tree structure.

**Hardware-efficient Ansätze.** We first quantify the expressivity of QNN proposed by [17], where  $\hat{U}(\boldsymbol{\theta})$  is implemented by the hardware-efficient Ansätze, under the both the ideal and NISQ settings. An  $N$ -qubits hardware-efficient Ansatz is composed of  $L$  layers, i.e.,  $U(\boldsymbol{\theta}) = \prod_{l=1}^L U(\boldsymbol{\theta}^l)$  with  $L \sim \text{poly}(N)$ . For all layers, the arrangement of quantum gates in  $U(\boldsymbol{\theta}^l)$  is identical, which generally consists of parameterized single-qubit gates and fixed two-qubit gates. Moreover, each qubit is operated with at least one parameterized single-qubit gate, and two qubits gates within the layer can adaptively connect two qubits depending on the qubits connectivity of the employed quantum hardware. An example of the 7-qubits hardware-efficient Ansatz is illustrated in the left panel of Fig. E.6. The parameterized single-qubit gate  $U$  can be realized by the rotational qubit gates, e.g.,  $U \in \{R_X(\theta), R_Y(\theta), R_Z(\theta)\}$  or  $U = R_Z(\beta)R_Y(\gamma)R_Z(\nu)$  with  $\theta, \gamma, \beta, \nu \in [0, 2\pi]$ . The topology of two-qubit gates, i.e., CNOT gates, aims to adapt to the chain-like connectivity restriction.

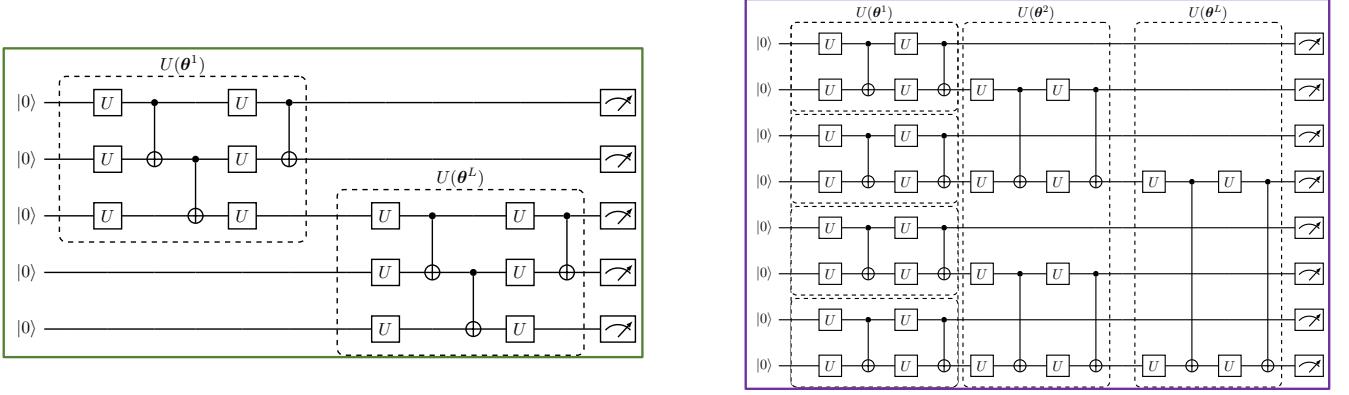


FIG. D.5: **Illustration of two tensor-network based Ansätze used in QNNs.** The left panel presents the tensor network based Ansätze with the matrix product state structure (highlighted by the green box), and the right panel refers to the tensor network based Ansätze with the tree structure (highlighted by the purple box).

The hardware-efficient Ansatz considered here is the most general case. Specifically, the single-qubit gate  $U$  contains three trainable parameters and the number of two-qubit gates in each layer is set as  $N$ . Under this setting, the total number of quantum gates in  $U(\boldsymbol{\theta}) = \prod_{l=1}^L U(\boldsymbol{\theta}^l)$  is

$$N_g = L(3N + N) = 4LN. \quad (\text{D1})$$

Based on the above settings, we achieve the expressivity of QNN with the hardware-efficient Ansätze, supported by Theorem 1 and Proposition 1.

**Corollary 2.** *Under the ideal setting, the covering number of QNN with the hardware-efficient Ansatz is upper bounded by  $\mathcal{N}(\mathcal{H}_{\text{QNN}}, \epsilon, |\cdot|) \leq \left(\frac{21NL\|O\|}{\epsilon}\right)^{6NL}$ . When the system noise is considered and simulated by the depolarization channel, the corresponding covering number is upper bounded by  $\widetilde{\mathcal{N}}(\mathcal{H}_{\text{QNN}}, \epsilon, |\cdot|) \leq (1-p)^{4NL} \left(\frac{21NL\|O\|}{\epsilon}\right)^{6NL}$ .*

**Tensor-network based Ansätze with the matrix product state structure.** Another type of Ansätze inherits the tensor-network structures, i.e., matrix product states and tree tensor network [39]. The left panel in Fig. D.5 illustrates the tensor-network based Ansätze with the matrix product state structure. Mathematically, for an  $N$ -qubit quantum circuit, the corresponding Ansätze yields

$$\hat{U}(\boldsymbol{\theta}) = \prod_{l=1}^L \left( \mathbb{I}_{2^{(M_1-1)*(l-1)}} \otimes U(\boldsymbol{\theta}^l) \otimes \mathbb{I}_{2^{N-1-(M_1-1)*l}} \right), \quad (\text{D2})$$

where  $U(\boldsymbol{\theta}^l)$  is applied to  $M_1$  qubits for  $\forall l \in [L]$  with  $2 \leq M_1 < N$ . The topology as shown in Fig. D.5 indicates that the maximum circuit depth of the tensor-network based Ansätze with the matrix product state structure is  $L = \lceil N/(M_1 - 1) \rceil$ . Suppose that the total number of single-qubit and two-qubit quantum gates in  $U(\boldsymbol{\theta}^l)$  is  $3M_1$  and  $M_1$  respectively, we have

$$N_g = 4M_1 \lceil N/(M_1 - 1) \rceil \leq 4(N + M_1 + N/M_1) \leq 4(N + 2\sqrt{N}). \quad (\text{D3})$$

Based on the above settings, we achieve the expressivity of QNN with tensor-network based Ansätze with the matrix product state structure, supported by Theorem 1 and Proposition 1.

**Corollary 3.** *Under the ideal setting, the covering number of QNN with tensor-network based Ansätze with the matrix product state structure is upper bounded by  $\mathcal{N}(\mathcal{H}_{\text{QNN}}, \epsilon, |\cdot|) \leq \left(\frac{21(N+2\sqrt{N})\|O\|}{\epsilon}\right)^{6(N+2\sqrt{N})}$ . When the system noise is considered and simulated by the depolarization channel, the corresponding covering number is upper bounded by  $\widetilde{\mathcal{N}}(\mathcal{H}_{\text{QNN}}, \epsilon, |\cdot|) \leq (1-p)^{4(N+2\sqrt{N})} \left(\frac{21(N+2\sqrt{N})\|O\|}{\epsilon}\right)^{6(N+2\sqrt{N})}$ .*

**Tensor-network based Ansätze with the tree structure.** The right panel in Fig. D.5 illustrates the tensor-network based Ansätze with tree structure. Intuitively, the involved number of quantum gates is exponentially decreased in terms of  $l \in [L]$ . Suppose that the local unitary, as highlighted by the dotted box in the right panel

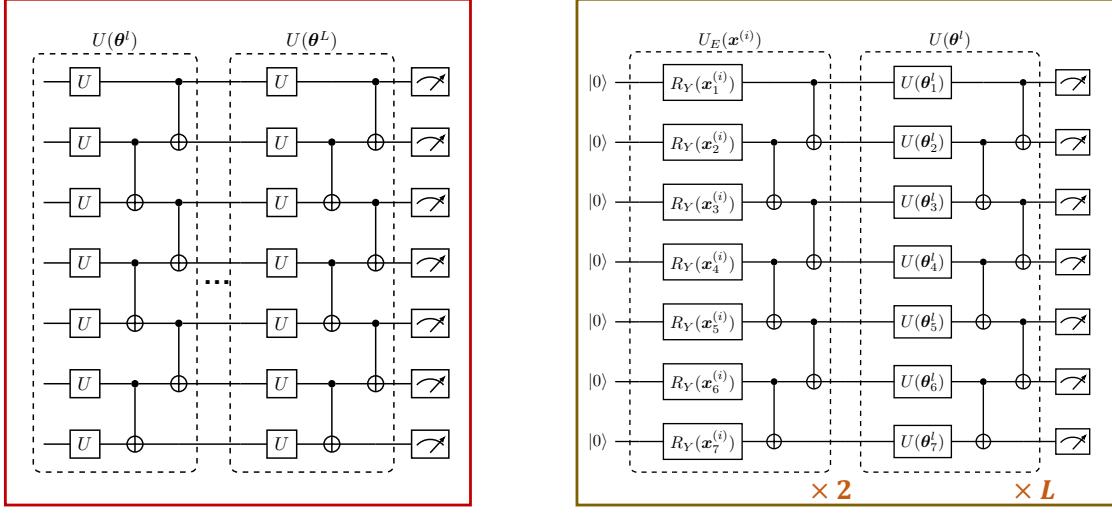


FIG. E.6: **QNN with the hardware-efficient Ansätze.** The left panel depicts the 7-qubit hardware-efficient Ansätze (highlighted by the red box). The right panel illustrates the implementation of QNN used in the numerical simulation (highlighted by the yellow box).

of Fig. D.5 with  $l = 1$ , contains six single-qubit gates (i.e., each qubit is operated with  $R_Z(\beta)R_Y(\gamma)R_Z(\nu)$ ) and one two-qubit gates. Then for an  $N$ -qubit quantum circuit, the total number of quantum gates in  $\hat{U}$  is

$$N_g = 7\lceil N/2 \rceil + 7\lceil N/4 \rceil + \dots + 7 * 2 \leq 7N. \quad (\text{D4})$$

Based on the above settings, we achieve the expressivity of QNN with tensor-network based Ansätze with the matrix product state structure, supported by Theorem 1 and Proposition 1.

**Corollary 4.** *Under the ideal setting, the covering number of QNN with tensor-network based Ansätze with the matrix product state structure is upper bounded by  $\mathcal{N}(\mathcal{H}_{\text{QNN}}, \epsilon, |\cdot|) \leq \left(\frac{73.5N\|O\|}{\epsilon}\right)^{10.5N}$ . When the system noise is considered and simulated by the depolarization channel, the corresponding covering number is upper bounded by  $\widetilde{\mathcal{N}}(\mathcal{H}_{\text{QNN}}, \epsilon, |\cdot|) \leq (1-p)^{7N} \left(\frac{73.5N\|O\|}{\epsilon}\right)^{10.5N}$ .*

#### Appendix E: Numerical simulation details of QNN

**Implementation.** The implementation of QNN employed in the numerical simulations is shown in the right panel of Fig. E.6. In particular, the qubit encoding method [54] is exploited to load classical data into quantum forms. The explicit form of the encoding circuit is

$$U_E(\mathbf{x}^{(i)}) = U_{\text{Eng}} \left( \bigotimes_{j=1}^7 R_Y(\mathbf{x}_j^{(i)}) \right) U_{\text{Eng}} \left( \bigotimes_{j=1}^7 R_Y(\mathbf{x}_j^{(i)}) \right), \quad (\text{E1})$$

where the unitary  $U_{\text{Eng}}$  is formed by CNOT gates as shown in Fig. E.6. The parameterized single-qubit qubit used in the Ansätze yields  $U(\theta_j^l) = R_Z(\beta)R_Y(\gamma)R_Z(\nu)$  for  $\forall j \in [N]$  and  $\forall l \in [L]$ , where  $\beta, \gamma, \nu \in [0, 2\pi)$  are independent trainable parameters.

**Data construction.** The construction of the synthetic dataset  $S = \{\mathbf{x}^{(i)}, y^{(i)}\}_{i=1}^n$  imitates the studies [10, 17]. Specifically, for each example, the feature dimension of  $\mathbf{x}^{(i)}$  is set as 7, i.e.,  $\mathbf{x}^{(i)} = [\omega_1^{(i)}, \omega_2^{(i)}, \omega_3^{(i)}, \omega_4^{(i)}, \omega_5^{(i)}, \omega_6^{(i)}, \omega_7^{(i)}]^T \in [0, 2\pi)^7$ , and the label  $y^{(i)} \in \{0, 1\}$  is binary. The assignment of the label  $y^{(i)}$  is accomplished as follows. Define  $V \in \mathcal{SU}(2^7)$  as a fixed unitary operator,  $O = \mathbb{I}_{2^6} \otimes |0\rangle\langle 0|$  as the measurement operator, and the gap threshold  $\Delta$  is set as 0.2. The label of  $\mathbf{x}^{(i)}$  is assigned as ‘1’ if

$$\langle 0^{\otimes 7} | U_E(\mathbf{x}^{(i)})^\dagger V^\dagger O V U_E(\mathbf{x}^{(i)}) | 0^{\otimes 7} \rangle \geq 0.5 + \Delta; \quad (\text{E2})$$

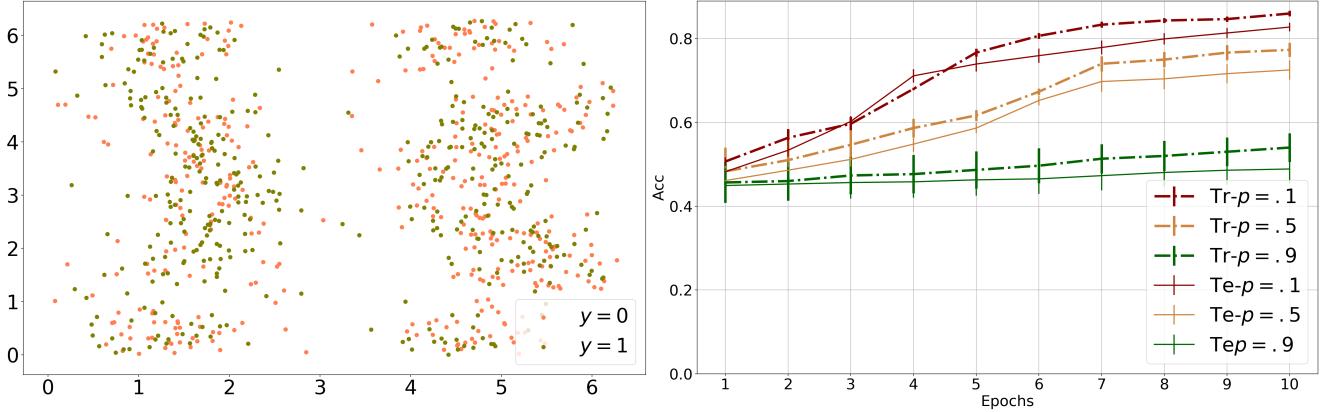


FIG. E.7: **The synthetic dataset and simulation results of noisy QNN.** The left plot illustrates the first two dimensions of the data points, where the green dots (pink dots) correspond to the data with label ‘1’ (‘0’). The right plot exhibits the learning performance of QNN when the depolarization noise is considered. The label ‘Tr- $p = a$ ’ refers to the training accuracy of QNN when the depolarization rate is set as  $p = a$ . Similarly, the label ‘Te- $p = a$ ’ refers to the test accuracy of QNN when the depolarization rate is set as  $p = a$ .

The label of  $\mathbf{x}^{(i)}$  is assigned as ‘0’ if

$$\langle 0^{\otimes 7} | U_E(\mathbf{x}^{(i)})^\dagger V^\dagger O V U_E(\mathbf{x}^{(i)}) | 0^{\otimes 7} \rangle \leq 0.5 - \Delta. \quad (\text{E3})$$

We note that  $V$  is realized by the Ansätze  $U(\boldsymbol{\theta}^*) = \prod_{l=1}^2 U(\boldsymbol{\theta}^{*l})$  shown in Fig. E.6, where the corresponding parameters  $\boldsymbol{\theta}^*$  are sampled with the random seed ‘1’. This setting ensures the target concept  $V$  is always covered by the hypothesis space  $\mathcal{H}_{\text{QNN}}$  once  $L \geq 2$ .

Based on the construction rule in the above two equations, we collect the dataset  $S$  with  $n = 400$ , where the positive and negative examples are equally distributed. We illustrate some examples of  $S$  in the left panel of Fig. E.7. Given access to  $S$ , we split the dataset into the training datasets with size 60 and the test dataset with 340.

**Hyper-parameters setting.** The hyper-parameters setting used in our experiment is as follows. At each epoch, we shuffle the training set in  $S$ . An epoch means that an entire dataset is passed forward through the quantum learning model, e.g., when the dataset contains 1000 training examples, and only two examples are fed into the quantum learning model each time, then it will take 500 iterations to complete 1 epoch. The learning rate is set as  $\eta = 0.2$ . The batch gradient descent method is adopted to be the optimizer with batch size equal to 4.

**The performance of noisy QNNs.** Here we apply noisy QNN to learn the synthetic data  $S$  introduced above to validate the correctness of Proposition 1. In particular, all settings, i.e., the employed Ansätze, the optimizer, and the hyper-parameters, are identical to the noiseless case, except that the employed quantum circuit is interacted with the depolarization noise. With the aim of understanding how the depolarization rate  $p$  shrinks the expressivity of  $\mathcal{H}_{\text{QNN}}$ , we set the layer number of the hardware-efficient Ansatz as  $L = 2$  and the depolarization rate as  $p \in \{0.1, 0.5, 0.9\}$ . We repeat each setting with 5 times to collect the statistical results.

The simulation results are presented in the right panel of Fig. E.7. Recall that the training performance of the noiseless QNN with  $L = 2$  is above 85% at the 10-th epoch, as shown in Fig. 3. Meanwhile, the construction rule of  $S$  indicates that the target concept is contained in  $U(\boldsymbol{\theta}) = \prod_{l=1}^2 U(\boldsymbol{\theta}^l)$ . However, the results in Fig. E.7 reflect that both the training and test accuracies continuously degrade in terms of the increased  $p$ . When  $p = 0.9$ , the learning performance is around 50%, which is no better than the random guess. These observations accord with Proposition 1 such that an increased depolarization rate suppresses the expressivity of  $\mathcal{H}_{\text{QNN}}$  and excludes the target concept out of the hypothesis space, which leads to a poor learning performance.

## Appendix F: Proof of Corollary 1

For completeness, let us first briefly introduce the unitary coupled-cluster Ansätze truncated up to single and double excitations (UCCSD) before presenting the proof of Corollary 1. Please refer to Refs. [62, 68] for comprehensive explanations. UCCSD belongs to a special type of unitary coupled-cluster (UCC) operator, which takes the form  $e^{T-T^\dagger}$ , where  $T$  corresponds to excitation operators defined for the configuration interaction. Since the unitary  $e^{T-T^\dagger}$  is difficult to implement on quantum computers, an alternative Ansätze is truncating UCC up to single and double

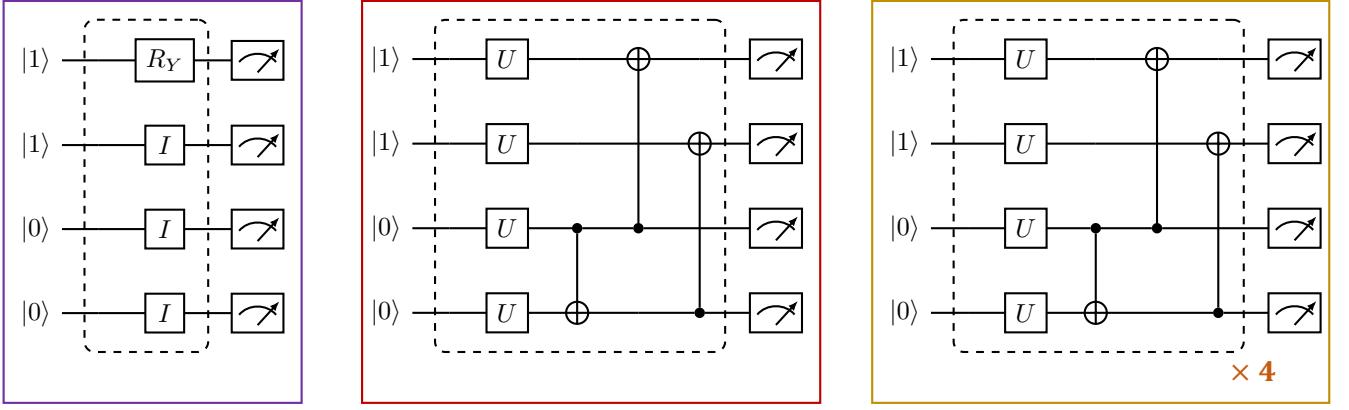


FIG. G.8: **Implementation of VQEs with different Ansätze.** The left, middle, and right panels depict the construction of VQE with restricted, modest, and overwhelming expressivity. The subscript ‘ $\times 4$ ’ in the right panel means repeating the circuit architecture in the dotted box with four times.

excitations, as so-called UCCSD, which can be used to accurately describe many molecular systems and is exact for systems with two electrons. Mathematically, UCCSD estimates  $T$  by  $T_1 + T_2$ . The study [62] has indicated that for both the Bravyi-Kitaev and the Jordan-Wigner transformations, the required number of quantum gates to implement UCCSD is upper bounded by  $N_g \sim O(N^5)$ .

*Proof of Corollary 1.* The results of Corollary 1 can be immediately achieved by substituting  $N_g \sim O(N^5)$  as explained above with Theorem 2 and Proposition 1.  $\square$

#### Appendix G: Numerical simulation details of VQE

**Implementation.** The implementation of VQEs employed in numerical simulations is shown in Fig. G.8. Based on the results in Theorem 1, we control the involved number of quantum gates to separate the expressivity of different Ansätze. In particular, the Ansätze as shown in left panel has a restricted expressivity, which only contains a single trainable quantum gate. The Ansätze as shown in middle panel has a modest expressivity, where  $U(\theta_j^l) = R_Z(\beta)R_Y(\gamma)R_Z(\nu)$  for  $\forall j \in [N]$  and  $\forall l \in [L]$ . In other words, the total number of trainable quantum gates is 12. Note that an Ansatz is sufficient to locate the minimum energy of  $H$ . The Ansatz as shown in middle panel has an overwhelming expressivity. Compared with the Ansatz with the modest expressivity, the number of trainable quantum gates scales by four times. Such an over-parameterized model may suffer from the training hardness, caused by the barren plateaus phenomenon.

**The qubit Hamiltonians of the hydrogen molecule.** The Bravyi-Kitaev transformation [69] is used to attain the qubit Hamiltonian of the hydrogen molecule at each bond length. The mathematical form of the obtained qubit Hamiltonian yields

$$\begin{aligned} H = & f_0 \mathbb{I} + f_1 Z_0 + f_2 Z_1 + f_3 Z_2 + f_4 Z_0 Z_1 + f_4 Z_0 Z_2 + f_5 Z_1 Z_3 + f_6 X_0 Z_1 X_2 + f_6 Y_0 Z_1 Y_2 + f_7 Z_0 Z_1 Z_2 \\ & + f_4 Z_0 Z_2 Z_3 + f_3 Z_1 Z_2 Z_3 + f_6 X_0 Z_1 X_2 Z_3 + f_6 Y_0 Z_1 Y_2 Z_3 + f_7 Z_0 Z_1 Z_2 Z_3, \end{aligned} \quad (\text{G1})$$

where  $\{X_i, Y_i, Z_i\}$  stands for applying the Pauli operators on the  $i$ -th qubit and the coefficients  $\{f_j\}_{j=1}^7$  are determined by the bond length. In the numerical simulations, we use OpenFermion Library [70] to load these coefficients.

**Hyper-parameters setting.** The hyper-parameters setting related to the optimization of VQEs is as follows. The total number of iteration is set as 300. The tolerant error is set as  $10^{-6}$ . The gradient descent optimizer is adopted and the learning rate is set as  $\eta = 0.4$ . The random seed used to initialize trainable parameters is set as 0.

**The performance of VQEs with the Adam optimizer.** We conduct additional numerical simulations to explore how the expressivity of Ansätze affects the performance of VQEs when the adaptive optimizer is adopted. More precisely, we aim to investigate whether VQE with the over-parameterized Ansatz can outperform VQE with the modest Ansatz when the adaptive optimizer is adopted. The appended numerical simulations mainly follow the setup introduced in the main text. In particular, the hardware-efficient VQEs with the different layer number  $L$  are employed to estimate the ground state energy of the Hydrogen molecule with varied bond length ranging from  $0.3\text{\AA}$  to  $2.1\text{\AA}$ . Notably, we substitute the SGD optimizer with the Adam optimizer [71] to update trainable parameters  $\theta$

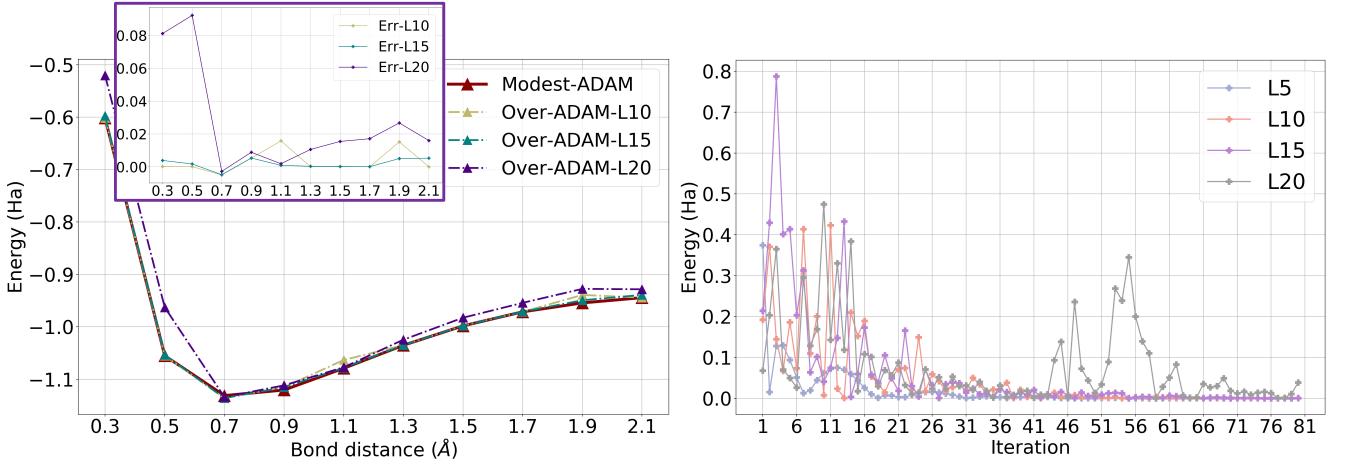


FIG. G.9: **Simulation results of VQE with the Adam optimizer.** The outer plot in the left panel illustrates the estimated energy of the hardware-efficient VQE with different number of layers  $L$ , i.e.,  $U(\boldsymbol{\theta}) = \prod_{l=1}^L U_l(\boldsymbol{\theta})$ . The label ‘Modest-Adam’ refers to the estimated energy of VQEs with the modest Ansatz introduced in the main text and the Adam optimizer. The labels ‘over-Adam-L- $b$ ’ refer to the estimated energy of VQEs when the employed Ansätze has the layer number  $L = b$  and the Adam optimizer is employed. The inner plot of the left panel shows the energy gap of VQEs in ‘Over-L- $b$ ’ and ‘Modest’ cases. ‘Ha’ (Hartrees) and ‘Å’ (Angstroms) refer to the units for energy and the bond lengths. The right panel depicts the energy difference of VQE between the neighboring two iterations with the setting  $L = 5, 10, 15, 20$ .

of the Ansatz with the adaptive learning rate. Mathematically, the updating rule of the Adam optimizer yields

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} - \eta^{(t+1)} \frac{a^{(t+1)}}{\sqrt{b^{(t+1)}} + \epsilon}, \quad (\text{G2})$$

where  $a^{(t+1)} = [\beta_1 a^{(t)} + (1 - \beta_1) \nabla \mathcal{L}(\boldsymbol{\theta}^{(t)})]/(1 - \beta_1)$ ,  $b^{(t+1)} = [\beta_2 b^{(t)} + (1 - \beta_2) \nabla \mathcal{L}(\boldsymbol{\theta}^{(t)})^{\odot 2}]/(1 - \beta_2)$ , and  $\eta^{(t+1)} = \eta^{(t)} \frac{\sqrt{(1-\beta_2)}}{(1-\beta_1)}$ . The hyper-parameters settings are as follows. The maximum number of iterations are fixed to be 81. The optimization is criticized to be converged and the updating is stopped when the energy difference between two iterations is lower than  $10^{-6}$ , i.e.,  $|\mathcal{L}(\boldsymbol{\theta}^{(t)}) - \mathcal{L}(\boldsymbol{\theta}^{(t-1)})| \leq 10^{-6}$ . The learning rate at the initial step is set as  $\eta = 0.4$ . The momentum parameters are set as  $\beta_1 = 0.9$  and  $\beta_2 = 0.99$ . The tolerance parameter is set as  $\epsilon = 10^{-6}$ . The random seed used to initialize trainable parameters is set as 0. The number of layers of the hardware-efficient Ansatz, i.e.,  $U(\boldsymbol{\theta}) = \prod_{l=1}^L U_l(\boldsymbol{\theta})$  exhibited in Fig. G.8, varies from  $L = 5$  to  $L = 20$ . Each setting is repeated 5 times to collect the statistical results.

The achieved simulation results are depicted in Fig. G.9. Specifically, when the Adam optimizer is utilized to adaptively adjust the learning rate at each iteration, VQE with the modest Ansatz still attains a better performance than VQE with the overwhelming-expressivity Ansatz. Moreover, as shown in the inner plot of the left panel, the performance of VQE continuously degrades with respect to the increased number of layer number  $L$ . The right panel in Fig. G.9 further exhibits the energy difference of VQE between the neighboring two iterations, i.e.,  $|\mathcal{L}(\boldsymbol{\theta}^{(t)}) - \mathcal{L}(\boldsymbol{\theta}^{(t-1)})|$ , when the bond length is 0.3 Å. For the setting  $L = 5$ , the optimization is converged when  $t = 79$  with  $|\mathcal{L}(\boldsymbol{\theta}^{(79)}) - \mathcal{L}(\boldsymbol{\theta}^{(78)})| = 9.4 \times 10^{-7}$ . For the setting  $L = 10, 15, 20$ , the energy difference of VQE between the last two iterations yields  $|\mathcal{L}(\boldsymbol{\theta}^{(79)}) - \mathcal{L}(\boldsymbol{\theta}^{(78)})| = 1.9 \times 10^{-4}, 8.9 \times 10^{-4}, 3.9 \times 10^{-2}$ , respectively. These observations indicate that our results still hold when the adaptive learning rate is considered. That is, too limited or too redundant expressivity of the employed Ansätze may prohibit the trainability of VQE.

#### Appendix H: Implications of Theorem 1 and Theorem 2 from the practical perspective

In this section, we elucidate how our theoretical results, i.e., Theorems 1 and 2, contribute to improve the learning performance of VQA-based models in practice. Concretely, the established theoretical results can be integrated with structural risk minimization [45] to enhance the learning performance of VQA-based models. Interestingly, the similar topic, i.e., the employment the expressivity of VQAs as guidance to enhance learning performance, has also been discussed in two very recent studies [72, 73].

Before moving on to explain how our results contribute to the structural risk minimization of VQA-based learning models in practice, let us first recap the theory of structural risk minimization. As shown in Fig. H.10, the learning

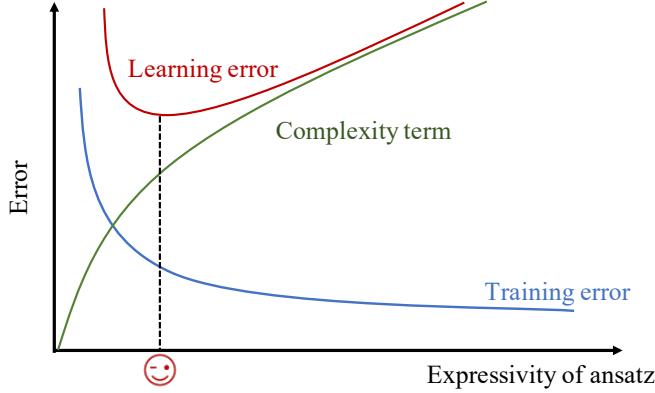


FIG. H.10: Illustration of the structural risk minimization adapted from [45].

performance of VQAs is determined by both the training error and the model's complexity term. Although the training error can be continuously suppressed by increasing the model's expressivity, the price to pay is increasing the complexity term, which may deteriorate its learning performance (e.g., the test accuracy for the unseen data). Overall, increasing the expressivity of VQAs beyond a certain threshold is no longer contributing to the improvement of learning performance or could even lead to a degraded learning performance. With this regard, it is of great importance to develop an efficient method that seeks an Ansatz with a ‘modest’ level (i.e., the smile face in Fig. H.10), which can well balance the tradeoff between the expressivity of the hypothesis space  $\mathcal{H}$  and the performance of a learning model. In other words, the core of the VQA-based model design is controlling its expressivity at a ‘modest’ level, envisioned by the statistical learning theory. With this aim, structural risk minimization is proposed as a concrete method that balances the trade-off between the expressivity and training error to attain the best possible learning performance. The mathematical expression of structural risk minimization can be formulated as

$$\min_{\theta \in \Theta, \Xi} \hat{\mathcal{R}}_S + g(S, \theta, \Xi) , \quad (\text{H1})$$

where  $\hat{\mathcal{R}}_S = \frac{1}{n} \sum_{i=1}^n \ell(h_{\mathcal{A}(S)}(\mathbf{x}^{(i)}), \mathbf{y}^{(i)})$  refers to the empirical risk (i.e., the training error term) defined in the main text and  $g(\cdot)$  refers to the complexity term, which is controlled by the input problem  $S$ , the trainable parameters  $\theta$  and the architecture of learning models.

We now detail three approaches that harness our theoretical results to engineer the complexity term  $g(\cdot)$  and hence improve the learning performance of VQA-based learning models.

1. The first approach is setting  $g(\cdot)$  as a regularizer with respect to the trainable parameters  $\theta$ , e.g.,  $g(\cdot) = \lambda \|\theta\|_2$  or  $g(\cdot) = \lambda \|\theta\|_0$ , where  $\lambda$  refers to a hyper-parameter. In this way, the optimal solution of the structural risk minimization in Eq. (H1) implicitly controls the expressivity of the learning model by sparsifying the trainable parameters, ensured by our theoretical results that the expressivity of VQAs depends on the number of trainable parameters.
2. The second approach is tailoring the spectral norm of the observable  $O$  when it is trainable, supported by our theoretical results that the expressivity of VQAs depends on  $\|O\|$ . For instance, the complexity term can be set as  $g(\cdot) = \lambda \|\theta\|_2 + \|O(\gamma)\|$ , where  $\gamma$  denotes the parameters of the trainable observable.
3. The last approach is carefully designing the complexity term of  $g(\cdot)$  that is determined by both the trainable parameters  $\theta$  and the quantum circuit architecture  $\Xi$ . The key motivation of updating the architecture of the quantum circuits is warranted by our theoretical results in Theorem 1, since the expressivity of VQAs depends on the adopted types of quantum gates (denoted by the term  $k$ ). Following this routine, several studies have proposed different variable structure methods to build Ansätze [48, 74–79]. Conceptually, these proposals developed a set of heuristic rules that during the optimization, the quantum gates in quantum circuits can either be added or deleted to find the optimal solution of the structural risk minimization in Eq. (H1).

We remark that the first two approaches presented above have also been discussed in Refs. [72, 73], where the analyzed expressivity of VQAs can be applied to structural risk minimization to improve the performance of VQAs. However, the derived bounds in their results are relatively loose and therefore fail to unveil the expressivity of VQAs is controlled by the types of quantum gates. With this regard, the achieved results in our study provide more concrete

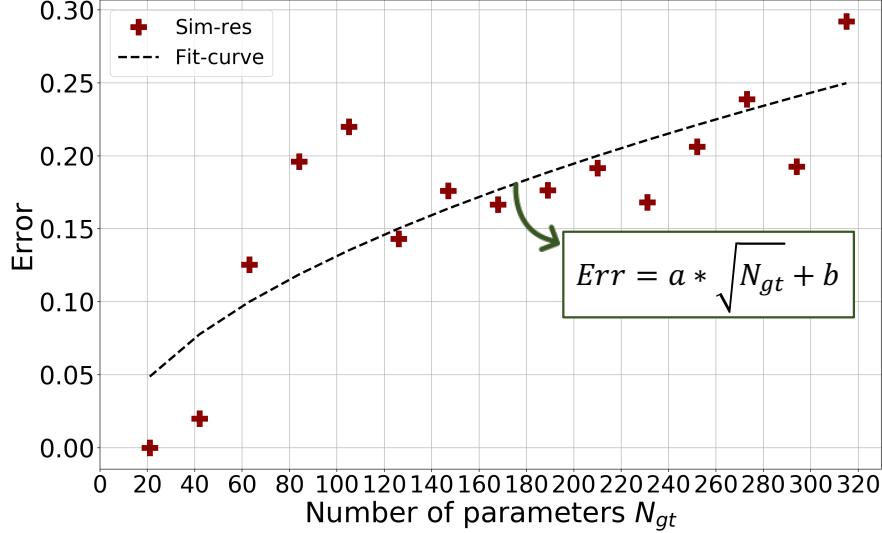


FIG. I.11: **The relationship between the learning error and the number of trainable parameters  $N_{gt}$ .** The label ‘Sim-res’ refers to the averaged simulation results. The label ‘Fit-curve’ refers to the fitting curve, i.e., the mathematical form is  $a\sqrt{N_{gt}} + b$  with  $a, b \in \mathbb{R}$ , with respect to the collected simulation results with the varied settings.

guidance (i.e., especially for the third approach) to implement structural risk minimization of VQA-based models in practice.

#### Appendix I: The tightness of the derived upper bound in Theorem 1

The derivation of the upper and lower bounds with respect to the model’s expressivity are at the heart of both classical and quantum machine learning. In the regime of deep learning, numerous studies [80, 81] devote to analyze the expressivity of deep neural networks (DNNs). Notably, most results focus on achieving a tighter upper bound for the expressivity of DNNs, while only few studies attempt to analyze the corresponding lower bound. This phenomenon is caused by the fact that the derivation of the lower bound is more difficult than the upper bound case. For instance, the seminal paper [82] only proved that the expressivity of DNNs is only lower bounded by the spectral norm of the input data and is independent with other parameters. The development of quantum learning theory also encounters the similar scenario. To our best knowledge, the lower bound of the expressivity for the VQA-based model has not been explored. With the aim of narrowing this knowledge gap, in the following, we conduct numerical simulations to empirically understand the tightness of the derived bound in Theorem 1.

Here we empirically explore whether the derived upper bound in Theorem 1 is tight with respect to the number of parameters  $N_{gt}$ . Note that directly calculating the covering number  $\mathcal{N}((\mathcal{H}_{\text{QNN}})_{|S}, \epsilon, \|\cdot\|_2)$  is very challenged in general. To this end, we devise an alternative method to examine the tightness of our bounds. Recall that the main conclusion achieved in Theorem 2 is

$$\begin{aligned}
& \mathcal{R}(\mathcal{A}(S)) - \hat{\mathcal{R}}_S(\mathcal{A}(S)) \\
& \leq 2L_1 \inf_{\alpha > 0} \left( 4\alpha + \frac{12}{\sqrt{n}} \int_{\alpha}^1 \sqrt{\ln \mathcal{N}((\mathcal{H}_{\text{QNN}})_{|S}, \epsilon, \|\cdot\|_2)} d\epsilon \right) + 3C_1 \sqrt{\frac{\ln(2/\delta)}{2n}} \\
& \leq \frac{8L_1}{\sqrt{n}} + \frac{24L_1}{\sqrt{n}} d^k \sqrt{N_{gt}} (\ln(7\sqrt{n}N_{gt}\|O\|) + 1) + 3C_1 \sqrt{\frac{\ln(1/\delta)}{2n}}.
\end{aligned}$$

This result connects the generalization error  $\mathcal{R}(\mathcal{A}(S)) - \hat{\mathcal{R}}_S(\mathcal{A}(S))$  with the expressivity of QNN, i.e.,  $\mathcal{N}((\mathcal{H}_{\text{QNN}})_{|S}, \epsilon, \|\cdot\|_2)$ . With this regard, the quantification of the tightness of the derived upper bound amounts to examining whether the generalization error of QNNs is linearly scaled with  $\sqrt{N_{gt}}$ . Mathematically, we aim to observe the relation  $\mathcal{R}(\mathcal{A}(S)) - \hat{\mathcal{R}}_S(\mathcal{A}(S)) \sim O(\sqrt{N_{gt}})$  to validate the tightness of our bounds in terms of  $N_{gt}$ .

We now employ the numerical simulations related to QNNs as introduced in the main text and Appendix E to complete the above examination. Specifically, all hyper-parameters settings are identical to those introduced in the

main text. The only modification is varying the layer number from  $L = 1$  to  $L = 15$ . The simulations results are depicted in Fig. I.11. Through fitting the simulation results, we observe that the learning error linearly scales with  $\sqrt{N_{gt}}$ , which accords with our theoretical results. We further use the coefficient of determination [83], denoted as  $R^2 \in [0, 1]$ , to measure the error of fitting. Intuitively, a higher  $R^2$  reflects a good fitting curve, where  $R^2 = 1$  indicates that the model explains all the variability of the response data around its mean. The coefficient of determination for the fitting curve shown in Fig. I.11 yields  $R^2 = 0.6687$ . These observations provide concrete evidence that the derived bound is tight with respect to the number of parameters.