Demystify Problem-Dependent Power of Quantum Neural Networks on Multi-Class Classification

Yuxuan Du,[1,*] Yibo Yang,[1] Dacheng Tao,[1] and Min-Hsiu Hsieh[2]

[1] *JD Explore Academy, Beijing 10010, China*
[2] *Hon Hai (Foxconn) Research Institute, Taipei, Taiwan*

Quantum neural networks (QNNs) have become an important tool for understanding the physical world, but their advantages and limitations are not fully understood. Some QNNs with specific encoding methods can be efficiently simulated by classical surrogates, while others with quantum memory may perform better than classical classifiers. Here we systematically investigate the problem-dependent power of quantum neural classifiers (QCs) on multi-class classification tasks. Through the analysis of expected risk, a measure that weighs the training loss and the generalization error of a classifier jointly, we identify two key findings: first, the training loss dominates the power rather than the generalization ability; second, QCs undergo a U-shaped risk curve, in contrast to the double-descent risk curve of deep neural classifiers. We also reveal the intrinsic connection between optimal QCs and the Helstrom bound and the equiangular tight frame. Using these findings, we propose a method that uses loss dynamics to probe whether a QC may be more effective than a classical classifier on a particular learning task. Numerical results demonstrate the effectiveness of our approach to explain the superiority of QCs over multilayer Perceptron on parity datasets and their limitations over convolutional neural networks on image datasets. Our work sheds light on the problem-dependent power of QNNs and offers a practical tool for evaluating their potential merit.

arXiv:2301.01597v1 [quant-ph] 29 Dec 2022

I. INTRODUCTION

The advent of hardware fabrication pushes the boundary of quantum computing from verifying its superiority on artificial tasks [1–3] to conquering realistic problems with merits [4–6]. This has led to the emergence of a popular paradigm known as quantum neural networks (QNNs), which combine variational quantum Ansätze with classical optimizers [7, 8]. So far, various QNN-based methods have been proposed to address difficult problems in areas such as quantum physics [9–12], quantum information theory [13–16], combinatorial optimization [17–21], and machine learning [22–26]. Among these applications, QNNs are often deployed as *quantum classifiers* (QCs) to predict correct labels of the input data [27–32], e.g., categorize image objects [33–35], classify phases of quantum matters [36–39], and distinguish entangled states from separable states [40, 41].

To comprehend the full potential of existing quantum classifiers (QCs) and to spur the development of novel QCs, huge efforts have been made to unveil the learnability of QCs [42–44]. Prior literature establishes the foundations of QCs from three primary aspects, i.e., model capacity [45–48], trainability [49–51], and generalization [52–57]. Nevertheless, the advantages and constraints of QCs have rarely been proven [57–62]. Meanwhile, previous results cannot rigorously explain the empirical observations such that QCs generally outperform classical classifiers (CCs) on handcraft or quantum data [44, 63] but are inferior to them on realistic problems [64]. As a result, the need for QCs to address classical issues remains highly questionable.

A principal criteria in characterizing the power of a classifier is the expected risk [65], which weighs the empirical risk (i.e., training loss) and the generalization error (i.e., test loss) jointly. An *optimal* classifier is one which achieves zero expected risk. As shown in Fig. 1(a), the success of deep neural classifiers is attributed to their double-descent risk curves [66, 67]. This means that as the hypothesis space is continually expanded, the expected risk of a trained deep neural classifier initially decreases, increases, and when it overfits the train set, undergoes a second descent. As such, to show the superiority of QCs over CCs, it demands to distill ubiquitous rules that capture the risk curve of diverse QCs in addition to conditions where the expected risk of QCs can be lower than CCs.

In this study, we unify a broad class of QCs in the same framework and understand their problem-dependent ability under the expected risk (see Fig. 1(b)). Our analysis reveals two substantial outcomes: (i) trainability dominates QCs' ability more than generalization ability; (ii) QCs undergo a U-shape risk curve instead of the double-descent curve for CCs. These outcomes consolidate and refine previous observations. Concretely, the first outcome suggests that the deficiency of QCs on classical data stems from their limited ability to fit the train set, resulting in a larger training loss compared to CCs. The second outcome highlights the distinct learning behavior of QCs and CCs. Despite the fact that overparameterization is fundamental to enhance the performance of CCs, it adversely affects the power of QCs. In line with the diverse dynamics of the risk curves for QCs and CCs, we devise an efficient problem-dependent method to recognize potential merits of QCs, as shown in Fig. 1(a). Conceptually, for a given learning task, our method fits the loss (risk) dynamics of QC and CC under the prior (i.e., U-shape versus double descent) and then

* duyuxuan123@gmail.com

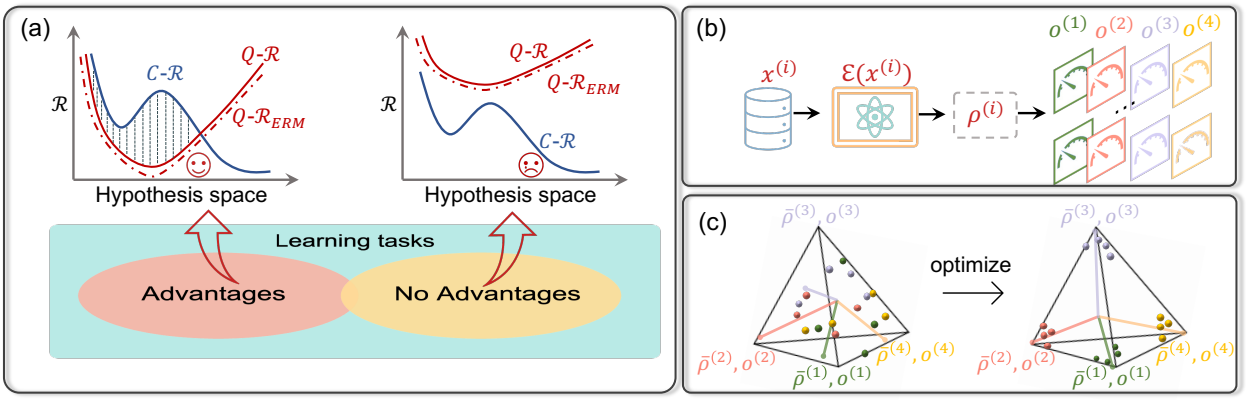

FIG. 1. **Risk curve and geometry of the unified QCs.** (a) The risk curve of QCs and CCs are highlighted by the solid red and blue lines (labeled by 'Q-$\mathcal{R}$' and 'C-$\mathcal{R}$'), respectively. The former yields a 'U' shape while the latter yields a double-descent tendency. Potential advantages of QCs are dominated by the empirical risk, highlighted by the dashed curve. The shaded region refers to the potential merits of QCs. (b) The unified QC consists of two parts, the feature state ρ and the measure operator $\boldsymbol{o}$. This model covers diverse QCs. (c) Geometric relationship between $\{\rho^{(i,k)}\}$ and $\boldsymbol{o}$ of QCs with (near) zero training loss: (i) the feature states associated with train samples belonging to the same class concentrate around their class-feature mean, i.e., $\bar{\rho}^{*(k)} := \rho^{*(1,k)} = ... = \rho^{*(n_c,k)}$ for $\forall k \in [K]$; (ii) the class-feature means are maximally distant with each other, i.e., $\mathrm{Tr}(\bar{\rho}^{*(k)}\bar{\rho}^{*(k')}) \sim \delta_{k,k'}$; (iii) the measure operator should align with class-feature means, i.e., $\mathrm{Tr}(\bar{\rho}^{*(k)}o^{*(k')}) \sim \delta_{k,k'}$.

identify the 'advantage' regime where the risk of QC is lower than CC. Numerical simulations are conducted to support our theoretical results.

On the technical level, we approach the two outcomes by separately quantifying the empirical risk and generalization error of QCs. Specifically, we first prove conditions of QCs that lead to near-zero empirical risk, the geometric interpretation of which is depicted in Fig. 1(c). As a byproduct, we elucidate how such conditions are inherently linked to quantum state discrimination and quantum measurement theory. In addition, we prove that deep QCs can never reach the vanished empirical risk by utilizing the concentration property of quantum observables [68, 69]. We next analyze the generalization error of QCs by exploiting algorithmic robustness [70]. The derived bound surpasses prior results because it is the first non-vacuous bound in the over-parameterized regime. By combining the unreachable zero empirical risk with the manipulatable generalization error, we obtain the first outcome. The second outcome is gained by integrating the fact that deep QCs are unable to reach the vanished empirical risk with the first outcome.

II. MAIN RESULTS

Expected risk.— Let us first introduce a K-class ($K \geq 2$) classification task. Denote the input space as $\mathcal{X}$, the label (class) space as $\mathcal{Y} = \{1, \cdots, K\}$, and the train set as $\mathcal{D} = \bigcup_{k=1}^{K}\{(\boldsymbol{x}^{(i,k)}, y^{(i,k)})\}_{i=1}^{n_k}$ with $|\mathcal{D}|$ samples drawn i.i.d. from an unknown probability distribution $\mathbb{D}$ on $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. In standard scenarios, the number of train samples in each class is the same, i.e., $n_1 = ... = n_k \equiv n_c$ and $|\mathcal{D}| := n = Kn_c$. The purpose of a classification algorithm $\mathcal{A}$ is using $\mathcal{D}$ to infer a hypothesis (a.k.a., a classifier) $h_{\mathcal{A}_\mathcal{D}} : \mathcal{X} \to \mathbb{R}^K$ from the hypothesis space $\mathcal{H}$ to separate train examples from different classes. This is equivalent to identifying an optimal hypothesis in $\mathcal{H}$ minimizing the *expected risk* $\mathsf{R}(h) = \mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim\mathbb{D}}[\ell(h(\boldsymbol{x}), y)]$, where $\ell(\cdot,\cdot)$ is the per-sample loss and for clarity we specify it as the square error with $\ell(\boldsymbol{a}, \boldsymbol{b}) = \frac{1}{2}\|\boldsymbol{a} - \boldsymbol{b}\|_2^2$ [71]. Unfortunately, the inaccessible distribution $\mathbb{D}$ forbids us to assess the expected risk directly. In practice, $\mathcal{A}$ alternatively learns an *empirical classifier* $\hat{h} \in \mathcal{H}$, as the global minimizer of the (regularized) loss function

$$\mathcal{L}(h, \mathcal{D}) = \frac{1}{n}\sum_{i=1}^{n_c}\sum_{k=1}^{K}\ell(h(\boldsymbol{x}^{(i,k)}), y^{(i,k)}) + \mathfrak{E}(h), \qquad (1)$$

where $\mathfrak{E}(h)$ is an optional regularizer.

The foremost role of the risk means that quantum advantages can be ascertained if $\mathsf{R}(\hat{h}_Q) < \mathsf{R}(\hat{h}_C)$, where $\hat{h}_Q$ and $\hat{h}_C$ are the empirical QC and CC on $\mathcal{D}$. Unlike conventions merely focusing on a QC on one specific task, the above criteria orients to unearth *ubiquitous rules* of QCs with computational advantages. To reconcile the intractable issue of $\mathsf{R}(\hat{h})$ and proceed the following analysis, we decomposed it into two measurable terms, i.e.,

$$\mathsf{R}(\hat{h}) = \mathsf{R}_{\mathrm{ERM}}(\hat{h}) + \mathsf{R}_{\mathrm{Gene}}(\hat{h}), \qquad (2)$$

where $\mathsf{R}_{\mathrm{ERM}}(\hat{h}) = \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K}\ell(\hat{h}(\boldsymbol{x}^{(i,k)}), y^{(i,k)})$ is the *empirical risk* and $\mathsf{R}_{\mathrm{Gene}}(\hat{h}) = \mathsf{R}(\hat{h}) - \mathsf{R}_{\mathrm{ERM}}(\hat{h})$ is the *generalization error*. Based on Eq. (2), detecting advances of QCs is translated to deriving under what conditions do QCs commit both lower $\mathsf{R}_{\mathrm{ERM}}$ and $\mathsf{R}_{\mathrm{Gene}}$ than CCs.

To better elucidate our results, let us recall that the general form of QC is $\hat{h}_Q = \arg\min_{h_Q \in \mathcal{H}_Q}\mathcal{L}(h_Q, \mathcal{D})$,

where $\mathcal{L}$ is defined in Eq. (1) and $\mathcal{H}_Q$ is the hypothesis space. For an N-qubit QC, its hypothesis space is

$$\mathcal{H}_Q = \left\{ \left[h_Q(\cdot, U(\boldsymbol{\theta}), O^{(k)}) \right]_{k=1:K} \middle| \boldsymbol{\theta} \in \boldsymbol{\Theta} \right\}, \quad (3)$$

where $[\cdot]_{k=1:K}$ is a K-dimensional vector, its k-th entry $h_Q(\boldsymbol{x}, U(\boldsymbol{\theta}), O^{(k)}) = \mathrm{Tr}(O^{(k)} U(\boldsymbol{\theta}) \sigma(\boldsymbol{x}) U(\boldsymbol{\theta})^\dagger)$ for $\forall k \in [K]$ refers to the output (prediction) of quantum circuits, $\sigma(\boldsymbol{x}) = U_E(\boldsymbol{x})(|0\rangle\langle 0|)^{\otimes N} U_E(\boldsymbol{x})^\dagger$ is the input state of $\boldsymbol{x}$ with the encoding circuit $U_E(\cdot)$, $\boldsymbol{O} = \{O^{(k)}\}_{k=1}^K$ is a set of measure operators, and $U(\boldsymbol{\theta})$ is the adopted Ansatz with trainable parameters $\boldsymbol{\theta}$ living in the parameter space $\boldsymbol{\Theta}$. Without loss of generality, we define $U(\boldsymbol{\theta}) = \prod_{l=1}^{N_t} (u_l(\boldsymbol{\theta}) u_e) \in \mathcal{U}(2^N)$, where $u_l(\boldsymbol{\theta}) \in \mathcal{U}(2^m)$ is the l-th parameterized quantum gate operated with at most m qubits ($m \leq N$) and u_e refers to fixed quantum gates. Similarly, we define $U_E(\boldsymbol{x}) = \prod_{g=1}^{N_g} u_g(\boldsymbol{x}) \in \mathcal{U}(2^N)$, where $u_g(\boldsymbol{x}) \in \mathcal{U}(2^m)$ refers to the g-th quantum gate operated with at most m qubits, and N_g gates contain N_{ge} tunable gates and $N_g - N_{ge}$ fixed gates.

Due to the diverse constructions of $U(\boldsymbol{\theta})$ and $U_E(\cdot)$, it is necessary to unify various QCs into the same framework to obtain the generic results. Notably, the unified QC should be *agnostic to* particular forms of these two terms. A feasible way is rewritten $h_Q(\cdot, U(\boldsymbol{\theta}), O^{(k)})$ as

$$h_Q(\rho^{(i,k)}, o^{(k)}) := \mathrm{Tr}(\rho^{(i,k)} o^{(k)}) \ \forall k \in [K], \quad (4)$$

where $O^{(k)} = \mathbb{I}_{2^{N-D}} \otimes o^{(k)}$ with the nontrivial local operator $o^{(k)} \in \mathbb{C}^{2^D \times 2^D}$, D describes the locality, and $\rho^{(i,k)} = \mathrm{Tr}_D(U(\boldsymbol{\theta}) \sigma(\boldsymbol{x}^{(i,k)}) U(\boldsymbol{\theta})^\dagger)$ corresponds to the state before measurements, named as *feature state*. An intuition of the unified QC is shown in Fig. 1(b).

We are now ready to exploit the unified framework to analyze the expected risk of QCs. Let $\boldsymbol{\rho} = \{\rho^{(i,k)}\}$ and $\boldsymbol{o} = \{o^{(k)}\}$ be two sets collecting all feature states and measure operators. The following theorem exhibits conditions in which QCs allow a low expected risk, where the formal statement and the proof are deferred to SM A.

Theorem 1 (informal)**.** *Following notations in Eqs. (1)-(4), when the train data size is* $nO(KN_{ge} \log \frac{KN_g}{\epsilon\delta})$ *with* ϵ *being the tolerable error, and the optimal sets of* $\boldsymbol{\rho}^*$ *and* $\boldsymbol{o}^*$ *satisfy three conditions: (i) the feature states have the vanished variability in the same class; (ii) all feature states are equal length and are orthogonal in the varied classes; (iii) any feature state is alignment with the measure operator in the same class, with probability* $1-\delta$*, the expected risk of QC tends to be zero, i.e.,* $\mathsf{R}(\hat{h}_Q) \to 0$*.*

Conditions (i)-(iii) visualized in Fig. 1(c) sculpt the geometric interpretations of $\boldsymbol{\rho}^*$ and $\boldsymbol{o}^*$. These properties come across the design philosophy of CCs, e.g., linear discriminant analysis and neural collapse phenomenon appeared in most deep neural classifiers [71–73]. Moreover, these conditions unveil the intrinsic connection between optimal QCs and the quantum state discrimination [74], since $\boldsymbol{\rho}^*$ and $\boldsymbol{o}^*$ should maximize the Helstrom bound [75], which explains the ultimate limit of QCs observed in [76]. However, as will be explained later (see Corollary 1 and Lemma 1), under certain scenarios, it is hard for QCs to meet these conditions. A typical instance is applying QC to learn the image dataset, where the difficulty stems from the limited nonlinearity of QC to fit the train set, thereby inducing a large empirical risk.

Conditions (i)-(iii) also imply how the quantum measurement theory can be used to guide the design of QCs. Namely, the mean feature states of each class $\{\bar{\rho}^{*(k)}\}$ compose the equiangular tight frame (ETF) and Condition (iii) suggests that the optimal measure operators $\{\boldsymbol{o}^*\}$ also satisfy this ETF [77]. Due to the relation between symmetric informationally complete (SIC) measurements and ETF [78, 79], the optimal measure operators could be estimated by various SIC construction strategies [80]. Besides, the locality D of $\{\boldsymbol{o}^*\}$ should be carefully selected in QCs in which a small D precludes the acquisition of the optimal QCs (i.e., the complex ETF does not exist when $2^D = K$ [81, 82]), while an extremely large D may incur the barren plateaus [83, 84]. Furthermore, when K is large, Pauli-based measurements are preferable than computational basis measurements in QCs, since the former allows classical shadow techniques to accelerate the training of QCs [85, 86].

The scaling behavior of n indicates that it is data-efficient for QCs to attain a low generalization error, where the size of train set only linearly depends on the class number K and the number of encoding gates N_{ge} (see Lemma 3 for the technical elaboration). In other words, the generalization error of QCs can be well controlled by the modest-size train data.

According to Theorem 1, the challenges in satisfying Conditions (i)-(iii) and the well controlled generalization error pinpoint that the risk of a QC is mostly dominated by its empirical loss rather than its generalization error. In this view, the core in devising advanced QCs is tailoring $\mathcal{H}_Q$ in Eq. (3) so that $\hat{h}_Q$ has a (near) zero empirical risk on $\mathcal{D}$, or equivalently examining whether the employed QCs can fulfill Conditions (i)-(iii).

U-shape risk curve.—The risk curve concerns how the expected risk of a classifier behaves with the varied hypothesis space. It is desired that as with deep neural classifiers, QCs undergo a double-descent risk curve in the sense that the over-parameterized QCs consent a low expected risk when the trainable parameters N_t is much greater than the train data n. If so, 'over-parameterization' could serve as a golden law in designing QCs. However, the following corollary refutes the existence of the double-descent risk curve for QCs.

Corollary 1. *Following notations in Theorem 1, when* $\{U_E(\boldsymbol{x}) | \boldsymbol{x} \in \mathcal{X}\}$ *follows the Haar distribution, with probability* $1-\delta$*, the empirical QC follows* $|\mathrm{Tr}\left(\sigma(\boldsymbol{x}^{(i,k)})\sigma(\boldsymbol{x})\right) - \frac{1}{2^N}| \leq \sqrt{\frac{3}{2^{2N}\delta}}$*. When* $\{U(\boldsymbol{\theta}) | \boldsymbol{\theta} \in \Theta\}$ *follows the Haar distribution, with probability* $1-\delta$*, the empirical QC follows* $|\mathrm{Tr}(\rho^{(i,k)} o^{(k')}) - \frac{\mathrm{Tr}(o^{(k')})}{2^D}| <$

$\sqrt{\frac{\mathrm{Tr}(o^{(k')})^2+2\,\mathrm{Tr}((o^{(k')})^2)}{2^{2D}\delta}}$.

The proof is deferred to SM B. The achieved results reveal the caveat of deep QCs. Specifically, when $U_E(\boldsymbol{x})$ is deep, two encoded states $\sigma(\boldsymbol{x}^{(i,k)})$ and $\sigma(\boldsymbol{x}^{(i',k)})$ from the same class tend to be orthogonal, which contradicts with Conditions (i) in Theorem 1. Besides, when $U(\boldsymbol{\theta})$ is deep, the output of QC concentrates to zero, regardless how $o^{(k')}$ and $\rho^{(i,k)}$ are selected. This violates Condition (iii) in Theorem 1. Overall, over-parameterized QCs encounter the high empirical risk and thus the high expected risk, which suggests that QCs experience a *U-shape risk curve*. This observation differs the dynamics of QCs from variational quantum Eigensolvers, since the latter can benefit from over-parameterization, e.g., better trainability and convergence rate [87–90]. Moreover, the rule of thumb in QCs' construction is slimming $\mathcal{H}_Q$ to find the valley region. Intriguingly, tailoring the feature states echoes with quantum metric learning and quantum self-supervised learning [91–95].

Probe power of QCs via loss dynamics.—The distinct tendency of the risk curves between QCs and CCs provides a succinct way to recognize the potential quantum advantages. As shown in Fig. 1(a), given a specific data set, the U-shape risk curve of QCs indicates that its advantages mostly appear in the valley region. Precisely, if the risk values of QC around the basin are lower than those of CC, potential merits may exist; otherwise, QC is inferior to CC. The proved learning behavior of QCs, accompanied with the tight generalization bound, allows us to effectively fit its risk curve according to their loss dynamics. Specifically, our method contains three steps. First, W tuples of $\{n, N_t, T\}$ are initialized based on Theorem 1 so that the collected risk points of QC span the basin area with low generalization error. Second, we execute QC and CC under these W hyper-parameter settings and fit their loss dynamics to attain the risk curve. Last, we compare two risk curves and probe potential advantages. See SM F for the implementation details.

Technical analysis.—Theorem 1 is achieved by analyzing when $\mathsf{R}_{\mathrm{ERM}}(\hat{h}_Q)$ and $\mathsf{R}_{\mathrm{Gene}}(\hat{h}_Q)$ are (near) zero. In the analysis of $\mathsf{R}_{\mathrm{ERM}}(\hat{h}_Q)$, we first consider the most general case in which both $\boldsymbol{\rho}$ and $\boldsymbol{o}$ are tunable, where $\hat{h}_Q \equiv h_Q(\boldsymbol{\rho}^*, \boldsymbol{o}^*)$ with $(\boldsymbol{\rho}^*, \boldsymbol{o}^*) = \min_{\boldsymbol{\rho},\boldsymbol{o}} \mathcal{L}(\boldsymbol{\rho}, \boldsymbol{o})$.

Lemma 1 (Informal)**.** *When the regularizer $\mathfrak{E}$ is considered and $(\boldsymbol{\rho}^*, \boldsymbol{o}^*)$ meets the three conditions in Theorem 1, the global minimizer leads to $\mathsf{R}_{\mathrm{ERM}}(\hat{h}_Q) = C_1^2/2$ with C_1 depending on the hyper-parameters in $\mathfrak{E}$.*

The achieved properties of $\boldsymbol{o}^*$ can be used as a priori to simplify QCs. To this end, the following lemma quantifies $\mathsf{R}_{\mathrm{ERM}}(\hat{h}_Q)$ when $\boldsymbol{o}$ is predefined and $\mathfrak{E} = 0$, where $\hat{h}_Q \equiv h_Q(\boldsymbol{\rho}^*, \boldsymbol{o})$ with $\boldsymbol{\rho}^* = \min_{\boldsymbol{\rho}} \mathcal{L}(\boldsymbol{\rho}, \boldsymbol{o})$.

Lemma 2 (Informal)**.** *When the predefined $\{o^{(k)}\}$ are mutually orthogonal with each other and the conditions in Theorem 1 are satisfied, we have $\mathsf{R}_{\mathrm{ERM}}(\hat{h}_Q) = 0$.*

The proofs of Lemmas 1 and 2 are given in SM C&D.

We next analyze $\mathsf{R}_{\mathrm{Gene}}(\hat{h}_Q)$. Prior results cannot be used to prove Theorem 1, since such bounds polynomially scale with the trainable parameters and become vacuous in the over-parameterized regime. To remedy this issue, we utilize the concept of algorithmic robustness [70].

Definition 1 (Robustness)**.** *A learning algorithm $\mathcal{A}$ is $(R, \nu(\cdot))$-robust with $R \in \mathbb{N}$ and $\nu(\cdot) : \mathcal{Z}^n \to \mathbb{R}$, if $\mathcal{Z}$ can be partitioned into R disjoint sets, denoted by $\{C_r\}_{r=1}^R$, such that the following holds for all $\mathcal{D} \subset \mathcal{Z}^n$: $\forall \boldsymbol{s} = (\boldsymbol{x}^{(i)}, y^{(i)}) \in \mathcal{D}$, $\forall \boldsymbol{z} = (\boldsymbol{x}, y) \in \mathcal{Z}$, $\forall r \in [R]$,*

$$\boldsymbol{s}, \boldsymbol{z} \in \mathcal{C}_r \Rightarrow |l(h_{\mathcal{A}_\mathcal{D}}(\boldsymbol{x}^{(i)}), y^{(i)}) - l(h_{\mathcal{A}_\mathcal{D}}(\boldsymbol{x}), y)| \leq \nu(\mathcal{D}).$$

Concisely, robustness measures how much the loss value can be varied with respect to the input space $\mathcal{Z}$. A higher robustness of a classifier admits lower R, $\nu(\cdot)$, and $\mathsf{R}_{\mathrm{Gene}}$ [70]. The following lemma quantifies the upper bound of $\mathsf{R}_{\mathrm{Gene}}(\hat{h}_Q)$ whose proof is given in SM E.

Lemma 3. *Suppose the measure operator is bounded by C_2 with $\max_{k\in[K]} \|o^{(k)}\| \leq C_2$. Define ϵ as the tolerable error. Following notations in Definition 1, the empirical QC is $(K(28N_{ge}/\epsilon)^{4^m N_{ge}}, 4L_1KC_2\epsilon)$-robust, and with probability $1-\delta$ we have*

$$\mathsf{R}_{\mathrm{Gene}}(\hat{h}_Q) \leq 4L_1KC_2\epsilon + 5\xi(\hat{h}_Q)\sqrt{\frac{|\mathcal{T}_\mathcal{D}|4^m N_{ge} \ln \frac{56KN_{ge}}{\epsilon\delta}}{n}},$$

where L_1 is the Lipschitz constant of ℓ with respect to h_Q, $\mathcal{I}_r^\mathcal{D} = \{i \in [n] : \boldsymbol{z}^{(i)} \in \mathcal{C}_r\}$, $\xi(\hat{h}) := \max_{\boldsymbol{z}\in\mathcal{Z}}(\ell(\hat{h}, \boldsymbol{z}))$, and $\mathcal{T}_\mathcal{D} := \{r \in [R] : |\mathcal{I}_r^\mathcal{D}| \geq 1\}$.

The achieved results convey threefold insights. First, our bound does not explicitly depend on the number of trainable parameters [96]. This unlocks a new way to understand the generalization ability of QCs, especially for the over-parameterized ones. Next, our bound hints that a carefully designed U_E can enhance performance of QCs [53, 97]. Last, $\mathsf{R}_{\mathrm{Gene}}(\hat{h}_Q) \to 0$ requires $n \gg |\mathcal{T}_\mathcal{D}|4^m N_{ge}$. Fortunately, a reasonable value of n is sufficient to warrant this condition, because in general $m \leq 2$, $N_{ge} \propto |\boldsymbol{x}|$, and $|\mathcal{T}_\mathcal{D}|$ is continuously decreased from n to K with respect to the reduced empirical loss.

III. NUMERICAL SIMULATIONS

We conduct numerical simulations to exhibit that the advantages and limitations of QCs on different classification tasks can be interpreted by the derived risk curve and feature states. The omitted construction details and results are deferred to SM G.

We first apply QC to accomplish the binary classification on the parity dataset [98–100]. The number of qubits is $N = 6$ and the hardware-efficient Ansatz is adopted to realize $U(\boldsymbol{\theta})$. The gradient descent method is used as the classical optimizer. Two measure operators are

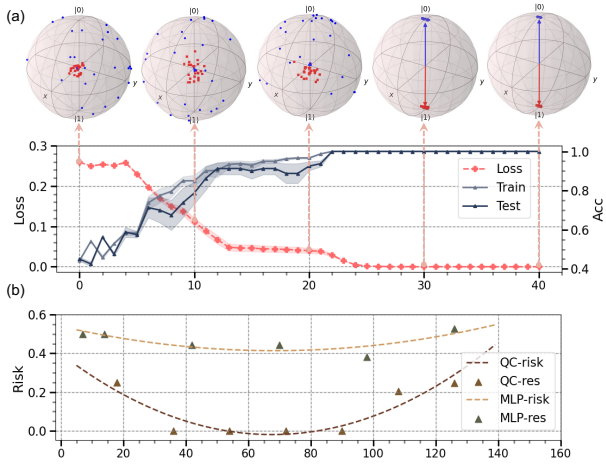

FIG. 2. **Binary classification on the parity dataset.** (a) The learning performance of QC when the layer number is 3. The x-axis denotes the epoch numbers. Shaded region represents variance. The Bloch spheres display the quantum feature states at different epochs. (b) The fitted risk curve of QC and MLP. The x-axis denotes the number of trainable parameters. The label '*QC-risk*' ('*MLP-risk*') refers to the fitted risk curve of QC and MLP. The label '*QC-res*' ('*MLP-res*') refers to the collected results used for fitting the curves.

$o^{(1)} = |0\rangle\langle 0|$ and $o^{(2)} = |1\rangle\langle 1|$. The simulation results of QC with $N_t = 54$ are displayed in Fig. 2(a). Particularly, the averaged train (test) accuracy steadily grows from 44.1% to 100% within 22 epochs, and the corresponding loss decreases from 0.26 to 4×10^{-5}. The dynamics of the feature states $\{\rho^{(i,t)}\}$ with $t \in \{0, 10, 20, 30, 40\}$ visualized by Bloch spheres echo with Lemma 2. Besides, QC becomes more robust when we continue the training. Although the train (test) accuracy reaches the optimum, the loss can be further reduced and suggests a lower risk warranted by Theorem 1. We further compare the risk curve between QC and multilayer Perceptron (MLP) on this dataset. We fit their risk curves following the proposed method to probe potential quantum merits. As shown in Fig. 2(b), QC clearly outperforms MLP when the trainable parameters ranges from 20 to 140 and the valley of the risk curve is around $N_t = 70$ [101].

We then apply QC to learn the Fashion-MNIST image dataset with $K = 9$ [102]. The employed number of qubits is $N = 10$ and the Pauli-based measure operators are employed. Convolutional neural networks (CNNs) are exploited as the reference. For all classifiers, the number of epochs is fixed to be $T = 50$ and the number of trainable parameters N_t ranges from 60 to 9000. Each setting is repeated with 3 times. As shown in Fig. 3, when the layer number is 50 with $N_t = 1500$, both the train and test accuracies of QC are about 50%. This performance is inferior to CNN under the similar setting. To explore whether QC has the potential to outperform CNN on this dataset, we compare their risk curves. As shown in Fig. 3(b), unlike the parity dataset, QC is evidently inferior to CNN on Fashion-MNIST dataset.

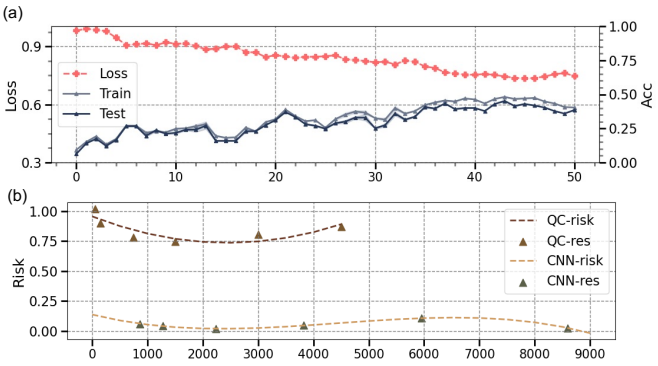

FIG. 3. **Multi-class classification on the image dataset with $K = 9$.** (a) The learning performance of QC when the layer number is 50. (b) The fitted risk curve of QC and CNN. All labels have the same meaning with those used in Fig. 2.

IV. DISCUSSIONS AND OUTLOOK

We understand the potential of diverse QCs in terms of the expected risk. Our theoretical findings demonstrate that the efficacy of QCs is dependent on the problem at hand, which explains the empirical evidence of their superiority on synthetic and quantum datasets, yet inferiority on realistic tasks. With the clear difference between the risk curve of QCs and deep neural classifiers, we present a concise technique to investigate potential quantum benefits by fitting their loss dynamics. Numerical results validate our theoretical results and the effectiveness of our method.

There are several interesting future research directions. The U-shape curve of QCs poses two open questions. First, can contemporary QCs attain quantum benefits on certain classical data when only limited data and restricted computing resources are available? Secondly, is it necessary to redesign QCs such as nonlinear QCs [103, 104] that can also exhibit a double-descent risk curve? Besides, the unearthed connection between the conditions towards optimal empirical risk and quantum state discrimination opens a new research avenue that amplifies the potential of QCs on quantum data aided by quantum information theory. Finally, it is intriguing to extend the developed non-vacuous generalization error bound of QCs to other scenarios, such as out-of-distribution data, in order to identify potential quantum advantages.

ACKNOWLEDGMENTS

The authors thank Xinbiao Wang for valuable input and inspiring discussions.

[1] Frank Arute, Kunal Arya, Ryan Babbush, Dave Bacon, Joseph C Bardin, Rami Barends, Rupak Biswas, Sergio Boixo, Fernando GSL Brandao, David A Buell, et al. Quantum supremacy using a programmable superconducting processor. *Nature*, 574(7779):505–510, 2019.

[2] Han-Sen Zhong, Hui Wang, Yu-Hao Deng, Ming-Cheng Chen, Li-Chao Peng, Yi-Han Luo, Jian Qin, Dian Wu, Xing Ding, Yi Hu, et al. Quantum computational advantage using photons. *Science*, 370(6523):1460–1463, 2020.

[3] Yulin Wu, Wan-Su Bao, Sirui Cao, Fusheng Chen, Ming-Cheng Chen, Xiawei Chen, Tung-Hsun Chung, Hui Deng, Yajie Du, Daojin Fan, et al. Strong quantum computational advantage using a superconducting quantum processor. *Physical review letters*, 127(18):180501, 2021.

[4] Xiao Mi, Pedram Roushan, Chris Quintana, Salvatore Mandra, Jeffrey Marshall, Charles Neill, Frank Arute, Kunal Arya, Juan Atalaya, Ryan Babbush, et al. Information scrambling in quantum circuits. *Science*, 374(6574):1479–1483, 2021.

[5] Yi Xia, Wei Li, Quntao Zhuang, and Zheshen Zhang. Quantum-enhanced data classification with a variational entangled sensor network. *Phys. Rev. X*, 11:021047, Jun 2021.

[6] M Cerezo, Guillaume Verdon, Hsin-Yuan Huang, Lukasz Cincio, and Patrick J Coles. Challenges and opportunities in quantum machine learning. *Nature Computational Science*, 2(9):567–576, 2022.

[7] Marcello Benedetti, Erika Lloyd, Stefan Sack, and Mattia Fiorentini. Parameterized quantum circuits as machine learning models. *Quantum Science and Technology*, 4(4):043001, 2019.

[8] Marco Cerezo, Andrew Arrasmith, Ryan Babbush, Simon C Benjamin, Suguru Endo, Keisuke Fujii, Jarrod R McClean, Kosuke Mitarai, Xiao Yuan, Lukasz Cincio, et al. Variational quantum algorithms. *Nature Reviews Physics*, 3(9):625–644, 2021.

[9] Xiao Yuan, Suguru Endo, Qi Zhao, Ying Li, and Simon C Benjamin. Theory of variational quantum simulation. *Quantum*, 3:191, 2019.

[10] Sam McArdle, Suguru Endo, Alán Aspuru-Guzik, Simon C Benjamin, and Xiao Yuan. Quantum computational chemistry. *Reviews of Modern Physics*, 92(1):015003, 2020.

[11] Cristina Cirstoiu, Zoe Holmes, Joseph Iosue, Lukasz Cincio, Patrick J Coles, and Andrew Sornborger. Variational fast forwarding for quantum simulation beyond the coherence time. *npj Quantum Information*, 6(1):1–10, 2020.

[12] Google AI Quantum, Collaborators*†, Frank Arute, Kunal Arya, Ryan Babbush, Dave Bacon, Joseph C Bardin, Rami Barends, Sergio Boixo, Michael Broughton, Bob B Buckley, et al. Hartree-fock on a superconducting qubit quantum computer. *Science*, 369(6507):1084–1089, 2020.

[13] Jonathan Romero, Jonathan P Olson, and Alan Aspuru-Guzik. Quantum autoencoders for efficient compression of quantum data. *Quantum Science and Technology*, 2(4):045001, 2017.

[14] Yuxuan Du and Dacheng Tao. On exploring practical potentials of quantum auto-encoder with advantages. *arXiv preprint arXiv:2106.15432*, 2021.

[15] Marco Cerezo, Alexander Poremba, Lukasz Cincio, and Patrick J Coles. Variational quantum fidelity estimation. *Quantum*, 4:248, 2020.

[16] Dmytro Bondarenko and Polina Feldmann. Quantum autoencoders to denoise quantum data. *Physical review letters*, 124(13):130502, 2020.

[17] Edward Farhi and Aram W Harrow. Quantum supremacy through the quantum approximate optimization algorithm. *arXiv preprint arXiv:1602.07674*, 2016.

[18] Leo Zhou, Sheng-Tao Wang, Soonwon Choi, Hannes Pichler, and Mikhail D Lukin. Quantum approximate optimization algorithm: Performance, mechanism, and implementation on near-term devices. *Physical Review X*, 10(2):021067, 2020.

[19] Matthew P Harrigan, Kevin J Sung, Matthew Neeley, Kevin J Satzinger, Frank Arute, Kunal Arya, Juan Atalaya, Joseph C Bardin, Rami Barends, Sergio Boixo, et al. Quantum approximate optimization of non-planar graph problems on a planar superconducting processor. *Nature Physics*, 17(3):332–336, 2021.

[20] Zeqiao Zhou, Yuxuan Du, Xinmei Tian, and Dacheng Tao. Qaoa-in-qaoa: solving large-scale maxcut problems on small quantum machines. *arXiv preprint arXiv:2205.11762*, 2022.

[21] Guido Pagano, Aniruddha Bapat, Patrick Becker, Katherine S Collins, Arinjoy De, Paul W Hess, Harvey B Kaplan, Antonis Kyprianidis, Wen Lin Tan, Christopher Baldwin, et al. Quantum approximate optimization of the long-range ising model with a trapped-ion quantum simulator. *Proceedings of the National Academy of Sciences*, 117(41):25396–25401, 2020.

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

[23] 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. *Physical Review Applied*, 16(2):024051, 2021.

[24] Jinkai Tian, Xiaoyu Sun, Yuxuan Du, Shanshan Zhao, Qing Liu, Kaining Zhang, Wei Yi, Wanrong Huang, Chaoyue Wang, Xingyao Wu, et al. Recent advances for quantum neural networks in generative learning. *arXiv preprint arXiv:2206.03066*, 2022.

[25] Xinbiao Wang, Yuxuan Du, Yong Luo, and Dacheng Tao. Towards understanding the power of quantum kernels in the nisq era. *Quantum*, 5:531, 2021.

[26] Yuxuan Du, Zhuozhuo Tu, Bujiao Wu, Xiao Yuan, and Dacheng Tao. Theory of quantum generative learning models with maximum mean discrepancy. *arXiv preprint arXiv:2205.04730*, 2022.

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

[28] Kosuke Mitarai, Makoto Negoro, Masahiro Kitagawa, and Keisuke Fujii. Quantum circuit learning. *arXiv preprint arXiv:1803.00745*, 2018.

[29] Maria Schuld, Alex Bocharov, Krysta M Svore, and Nathan Wiebe. Circuit-centric quantum classifiers. *Physical Review A*, 101(3):032308, 2020.

[30] Guangxi Li, Zhixin Song, and Xin Wang. Vsql: Variational shadow quantum learning for classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 8357–8365, 2021.

[31] Adrián Pérez-Salinas, Alba Cervera-Lierta, Elies Gil-Fuster, and José I Latorre. Data re-uploading for a universal quantum classifier. *Quantum*, 4:226, 2020.

[32] Weikang Li and Dong-Ling Deng. Recent advances for quantum classifiers. *Science China Physics, Mechanics & Astronomy*, 65(2):1–23, 2022.

[33] 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, 2021.

[34] Samuel Yen-Chi Chen, Chih-Min Huang, Chia-Wei Hsing, and Ying-Jer Kao. An end-to-end trainable hybrid classical-quantum classifier. *Machine Learning: Science and Technology*, 2(4):045021, 2021.

[35] Evan Peters, João Caldeira, Alan Ho, Stefan Leichenauer, Masoud Mohseni, Hartmut Neven, Panagiotis Spentzouris, Doug Strain, and Gabriel N Perdue. Machine learning of high dimensional data on a noisy quantum processor. *npj Quantum Information*, 7(1):1–5, 2021.

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

[37] Ming Gong, He-Liang Huang, Shiyu Wang, Chu Guo, Shaowei Li, Yulin Wu, Qingling Zhu, Youwei Zhao, Shaojun Guo, Haoran Qian, et al. Quantum neuronal sensing of quantum many-body states on a 61-qubit programmable superconducting processor. *arXiv preprint arXiv:2201.05957*, 2022.

[38] Johannes Herrmann, Sergi Masot Llima, Ants Remm, Petr Zapletal, Nathan A McMahon, Colin Scarato, François Swiadek, Christian Kraglund Andersen, Christoph Hellings, Sebastian Krinner, et al. Realizing quantum convolutional neural networks on a superconducting quantum processor to recognize quantum phases. *Nature Communications*, 13(1):1–7, 2022.

[39] Huili Zhang, Si Jiang, Xin Wang, Wengang Zhang, Xianzhi Huang, Xiaolong Ouyang, Yefei Yu, Yanqing Liu, Dong-Ling Deng, and L-M Duan. Experimental demonstration of adversarial examples in learning topological phases. *Nature communications*, 13(1):1–8, 2022.

[40] Edward Grant, Marcello Benedetti, Shuxiang Cao, Andrew Hallam, Joshua Lockhart, Vid Stojevic, Andrew G Green, and Simone Severini. Hierarchical quantum classifiers. *npj Quantum Information*, 4(1):1–8, 2018.

[41] Xu-Fei Yin, Yuxuan Du, Yue-Yang Fei, Rui Zhang, Li-Zheng Liu, Yingqiu Mao, Tongliang Liu, Min-Hsiu Hsieh, Li Li, Nai-Le Liu, et al. Efficient bipartite entanglement detection scheme with a quantum adversarial solver. *Physical Review Letters*, 128(11):110501, 2022.

[42] Amira Abbas, David Sutter, Christa Zoufal, Aurélien Lucchi, Alessio Figalli, and Stefan Woerner. The power of quantum neural networks. *Nature Computational Science*, 1(6):403–409, 2021.

[43] Yuxuan Du, Min-Hsiu Hsieh, Tongliang Liu, Shan You, and Dacheng Tao. Learnability of quantum neural networks. *PRX Quantum*, 2(4):040337, 2021.

[44] Hsin-Yuan Huang, Michael Broughton, Masoud Mohseni, Ryan Babbush, Sergio Boixo, Hartmut Neven, and Jarrod R McClean. Power of data in quantum machine learning. *Nature communications*, 12(1):1–9, 2021.

[45] Yuxuan Du, Min-Hsiu Hsieh, Tongliang Liu, and Dacheng Tao. Expressive power of parametrized quantum circuits. *Phys. Rev. Research*, 2:033125, Jul 2020.

[46] Tobias Haug, Kishor Bharti, and MS Kim. Capacity and quantum geometry of parametrized quantum circuits. *PRX Quantum*, 2(4):040309, 2021.

[47] Huitao Shen, Pengfei Zhang, Yi-Zhuang You, and Hui Zhai. Information scrambling in quantum neural networks. *Physical Review Letters*, 124(20):200504, 2020.

[48] Yadong Wu, Juan Yao, Pengfei Zhang, and Hui Zhai. Expressivity of quantum neural networks. *Physical Review Research*, 3(3):L032049, 2021.

[49] Eric R Anschuetz and Bobak T Kiani. Quantum variational algorithms are swamped with traps. *Nature Communications*, 13(1):1–10, 2022.

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

[51] Zoë Holmes, Kunal Sharma, Marco Cerezo, and Patrick J Coles. Connecting ansatz expressibility to gradient magnitudes and barren plateaus. *PRX Quantum*, 3(1):010313, 2022.

[52] Leonardo Banchi, Jason Pereira, and Stefano Pirandola. Generalization in quantum machine learning: A quantum information standpoint. *PRX Quantum*, 2(4):040321, 2021.

[53] Matthias C Caro, Elies Gil-Fuster, Johannes Jakob Meyer, Jens Eisert, and Ryan Sweke. Encoding-dependent generalization bounds for parametrized quantum circuits. *Quantum*, 5:582, 2021.

[54] Matthias C Caro, Hsin-Yuan Huang, M Cerezo, Kunal Sharma, Andrew Sornborger, Lukasz Cincio, and Patrick J Coles. Generalization in quantum machine learning from few training data. *arXiv preprint arXiv:2111.05292*, 2021.

[55] Yuxuan Du, Zhuozhuo Tu, Xiao Yuan, and Dacheng Tao. Efficient measure for the expressivity of variational quantum algorithms. *Physical Review Letters*, 128(8):080506, 2022.

[56] Casper Gyurik, Dyon van Vreumingen, and Vedran Dunjko. Structural risk minimization for quantum linear classifiers. *arXiv preprint arXiv:2105.05566*, 2021.

[57] Hsin-Yuan Huang, Richard Kueng, and John Preskill. Information-theoretic bounds on quantum advantage in machine learning. *Physical Review Letters*, 126(19):190505, 2021.

[58] Hsin-Yuan Huang, Michael Broughton, Jordan Cotler, Sitan Chen, Jerry Li, Masoud Mohseni, Hartmut Neven, Ryan Babbush, Richard Kueng, John Preskill, et al. Quantum advantage in learning from experiments. *Science*, 376(6598):1182–1186, 2022.

[59] Carlo Ciliberto, Andrea Rocchetto, Alessandro Rudi, and Leonard Wossnig. Statistical limits of supervised quantum learning. *Physical Review A*, 102(4):042414, 2020.

[60] Jonas Landman, Slimane Thabet, Constantin Dalyac, Hela Mhiri, and Elham Kashefi. Classically Approximating Variational Quantum Machine Learning with Random Fourier Features, 2022. arXiv:2210.13200v1.

[61] Giacomo De Palma, Milad Marvian, Cambyse Rouzé, and Daniel Stilck França. Limitations of variational quantum algorithms: a quantum optimal transport approach. *arXiv preprint arXiv:2204.03455*, 2022.

[62] Franz J Schreiber, Jens Eisert, and Johannes Jakob Meyer. Classical surrogates for quantum learning models. *arXiv preprint arXiv:2206.11740*, 2022.

[63] Yunchao Liu, Srinivasan Arunachalam, and Kristan Temme. A rigorous and robust quantum speed-up in supervised machine learning. *Nature Physics*, 17(9):1013–1017, 2021.

[64] Yang Qian, Xinbiao Wang, Yuxuan Du, Xingyao Wu, and Dacheng Tao. The dilemma of quantum neural networks. *arXiv preprint arXiv:2106.04975*, 2021.

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

[66] Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, and Ilya Sutskever. Deep double descent: Where bigger models and more data hurt. In *International Conference on Learning Representations*, 2020.

[67] Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019.

[68] Michael J Bremner, Caterina Mora, and Andreas Winter. Are random pure states useful for quantum computation? *Physical review letters*, 102(19):190502, 2009.

[69] David Gross, Steve T Flammia, and Jens Eisert. Most quantum states are too entangled to be useful as computational resources. *Physical review letters*, 102(19):190501, 2009.

[70] Huan Xu and Shie Mannor. Robustness and generalization. In Adam Tauman Kalai and Mehryar Mohri, editors, *COLT 2010 - The 23rd Conference on Learning Theory, Haifa, Israel, June 27-29, 2010*, pages 503–515. Omnipress, 2010.

[71] Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*, volume 4. Springer, 2006.

[72] Vardan Papyan, XY Han, and David L Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40):24652–24663, 2020.

[73] Yibo Yang, Shixiang Chen, Xiangtai Li, Liang Xie, Zhouchen Lin, and Dacheng Tao. Inducing neural collapse in imbalanced learning: Do we really need a learnable classifier at the end of deep neural network? In *NeurIPS*, 2022.

[74] Joonwoo Bae and Leong-Chuan Kwek. Quantum state discrimination and its applications. *Journal of Physics A: Mathematical and Theoretical*, 48(8):083001, 2015.

[75] for any two varied classes, $o^{(k)}$ and $o^{(k')}$ classify $\bar{\rho}^{*(k)}$ and $\bar{\rho}^{*(k')}$ with probability 1.

[76] Bingzhi Zhang and Quntao Zhuang. Fast decay of classification error in variational quantum circuits. *Quantum Science and Technology*, 2022.

[77] Note that Conditions (i)&(ii) imply that $\{\bar{\rho}^{*(k)}\}$ forms an orthogonal frame. Since any orthogonal frame can trivially be turned into a simplex ETF by reducing its global mean, we argue that $\{\bar{\rho}^{*(k)}\}$ composes ETF.

[78] Joseph M Renes, Robin Blume-Kohout, Andrew J Scott, and Carlton M Caves. Symmetric informationally complete quantum measurements. *Journal of Mathematical Physics*, 45(6):2171–2180, 2004.

[79] Andrew J Scott. Tight informationally complete quantum measurements. *Journal of Physics A: Mathematical and General*, 39(43):13507, 2006.

[80] Guillermo García-Pérez, Matteo AC Rossi, Boris Sokolov, Francesco Tacchino, Panagiotis Kl Barkoutsos, Guglielmo Mazzola, Ivano Tavernelli, and Sabrina Maniscalco. Learning to measure: Adaptive informationally complete generalized measurements for quantum algorithms. *Prx quantum*, 2(4):040342, 2021.

[81] Joel A Tropp. Complex equiangular tight frames. In *Wavelets XI*, volume 5914, page 591401. SPIE, 2005.

[82] Mátyás A Sustik, Joel A Tropp, Inderjit S Dhillon, and Robert W Heath Jr. On the existence of equiangular tight frames. *Linear Algebra and its applications*, 426(2-3):619–635, 2007.

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

[84] Stefan H Sack, Raimel A Medina, Alexios A Michailidis, Richard Kueng, and Maksym Serbyn. Avoiding barren plateaus using classical shadows. *PRX Quantum*, 3(2):020365, 2022.

[85] Hsin-Yuan Huang, Richard Kueng, and John Preskill. Predicting many properties of a quantum system from very few measurements. *Nature Physics*, 16(10):1050–1057, 2020.

[86] Hsin-Yuan Huang. Learning quantum states from their classical shadows. *Nature Reviews Physics*, 4(2):81–81, 2022.

[87] Junyu Liu, Zexi Lin, and Liang Jiang. Laziness, barren plateau, and noise in machine learning. *arXiv preprint arXiv:2206.09313*, 2022.

[88] Junyu Liu, Khadijeh Najafi, Kunal Sharma, Francesco Tacchino, Liang Jiang, and Antonio Mezzacapo. An analytic theory for the dynamics of wide quantum neural networks. *arXiv preprint arXiv:2203.16711*, 2022.

[89] Xinbiao Wang, Junyu Liu, Tongliang Liu, Yong Luo, Yuxuan Du, and Dacheng Tao. Symmetric pruning in quantum neural networks. *arXiv preprint arXiv:2208.14057*, 2022.

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

[91] Seth Lloyd, Maria Schuld, Aroosa Ijaz, Josh Izaac, and Nathan Killoran. Quantum embeddings for machine learning. *arXiv preprint arXiv:2001.03622*, 2020.

[92] Nhat A Nghiem, Samuel Yen-Chi Chen, and Tzu-Chieh Wei. Unified framework for quantum classification. *Physical Review Research*, 3(3):033056, 2021.

[93] Ryan LaRose and Brian Coyle. Robust data encodings for quantum classifiers. *Physical Review A*, 102(3):032420, 2020.

[94] Ben Jaderberg, Lewis W Anderson, Weidi Xie, Samuel Albanie, Martin Kiffner, and Dieter Jaksch. Quantum self-supervised learning. *Quantum Science and Technology*, 7(3):035005, 2022.

[95] Rui Yang, Samuel Bosch, Bobak Kiani, Seth Lloyd, and Adrian Lupascu. An analog quantum variational em-

bedding classifier, 2022. arXiv:2211.02748v1.

[96] The exact form of the first term in the generalization bound should be $4L_1KC_2f(U(\boldsymbol{\theta}))\epsilon$ with $f(U(\boldsymbol{\theta})) \leq 1$ for any Ansatz. Therefore, for simplicity, we discard the term $f(U(\boldsymbol{\theta}))$.

[97] Masahito Hayashi and Yuxiang Yang. Efficient algorithms for quantum information bottleneck. *arXiv preprint arXiv:2208.10342*, 2022.

[98] Andrew W Cross, Graeme Smith, and John A Smolin. Quantum learning robust against noise. *Physical Review A*, 92(1):012327, 2015.

[99] Diego Ristè, Marcus P Da Silva, Colm A Ryan, Andrew W Cross, Antonio D Córcoles, John A Smolin, Jay M Gambetta, Jerry M Chow, and Blake R Johnson. Demonstration of quantum advantage in machine learning. *npj Quantum Information*, 3(1):1–5, 2017.

[100] Pinaki Sen, Amandeep Singh Bhatia, Kamalpreet Singh Bhangu, and Ahmed Elbeltagi. Variational quantum classifiers through the lens of the hessian. *Plos one*, 17(1):e0262346, 2022.

[101] The disappeared double-descent curve of MLP is caused by the limited train data. In other words, overparameterization and sufficient train data are two necessary conditions to induce the double-descent curve, while parity dataset can only provide limited train data.

[102] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms, 2017.

[103] Maria Schuld and Nathan Killoran. Is quantum advantage the right goal for quantum machine learning? *PRX Quantum*, 3:030101, Jul 2022.

[104] Zoë Holmes, Nolan Coble, Andrew T Sornborger, and Yiğit Subaşı. On nonlinear transformations in quantum computation. *arXiv preprint arXiv:2112.12307*, 2021.

[105] Kenji Kawaguchi, Zhun Deng, Kyle Luh, and Jiaoyang Huang. Robustness implies generalization via data-dependent generalization bounds. In *International Conference on Machine Learning*, pages 10866–10894. PMLR, 2022.

[106] Thomas Barthel and Jianfeng Lu. Fundamental limitations for measurements in quantum many-body systems. *Phys. Rev. Lett.*, 121:080406, Aug 2018.

[107] Amit Daniely and Eran Malach. Learning parities with neural networks. *Advances in Neural Information Processing Systems*, 33:20356–20365, 2020.

[108] Boaz Barak, Benjamin L Edelman, Surbhi Goel, Sham Kakade, Eran Malach, and Cyril Zhang. Hidden progress in deep learning: Sgd learns parities near the computational limit. *arXiv preprint arXiv:2207.08799*, 2022.

[109] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

[110] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(7), 2011.

The organization of the supplementary material (SM) is as follows. In SM A, we present the proof of Theorem 1. Then, we provide the proof of Corollary 1 in SM B. Subsequently, we demonstrate the proof of Lemma 1 and Lemma 2 in SM C and SM D, respectively. Next, in SM E, we exhibit the proof of Lemma 3. In the end, we elucidate the details of numerical simulations in SM G.

SM A: Proof of Theorem 1

For convenience, let us first recall the settings and notations introduced in the main text. When QCs are applied to accomplish the multi-class classification task, the training dataset $\mathcal{D}$ contains n examples and the number of examples in each class is the same with $n = n_c K$. Moreover, the per-sample loss is specified as the mean square error.

We next introduce the formal description of Theorem 1. In particular, Theorem 1 is established on Lemma 2, where the regularization term is set as zero (i.e., $\mathfrak{E} = 0$) and the set of measure operator is predefined, i.e., $\boldsymbol{o}$ spans the space $\mathbb{C}^{2^D \times 2^D}$ and satisfies $\mathrm{Tr}(o^{(k)} o^{(k')}) = B\delta_{k,k'}$ where $B \geq 1$ is a constant. The requirements of $\boldsymbol{o}$ aims to preserve Condition (iii) in Lemma 1. Note that the focus on these specific settings adopted in Lemma 1 instead of the most general settings (i.e., $\boldsymbol{o}$ is tunable and $\mathfrak{E}$ is nonzero) is motivated by Lemma 1, which promises a lower expected risk. Following the above elaboration, the loss function of QC to be minimized can be explicitly written as

$$\mathcal{L}(\boldsymbol{\rho}) = \frac{1}{2n} \sum_{i=1}^{n_c} \sum_{k=1}^{K} \left([\mathrm{Tr}(\rho^{(i,k)} o^{(k)})]_{k=1:K} - \boldsymbol{y}^{(i,k)} \right)^2, \tag{A1}$$

where $\boldsymbol{y}^{(i,k)}$ is the unit basis whose k-th entry is 1 for $\forall i \in [n_c]$, $\forall k \in [K]$. Denote $\boldsymbol{\rho}^* = \min_\rho \mathcal{L}(\boldsymbol{\rho})$ and the empirical risk of QC as $\mathsf{R}_{\mathrm{ERM}}(\hat{h}_Q)$ with $\hat{h}_Q \equiv \hat{h}_Q(\boldsymbol{\rho}^*)$. The formal statement of Theorem 1 is as follows.

Theorem (Formal statement of Theorem 1)**.** *Following notations in Lemmas 2 and 3, with probability $1-\delta$, the expected risk of QC tends to be zero, i.e., $\mathsf{R}_{\mathrm{ERM}}(\hat{h}_Q) = 0$, when the size of train dataset satisfies $n \gg O(KN_{ge} \log \frac{KN_g}{\epsilon\delta})$ and the global minimizer $\boldsymbol{\rho}^*$ in Eq. (A1) satisfies*

$$(i)\bar{\rho}^{*(k)} := \rho^{*(1,k)} = ... = \rho^{*(n_c,k)};\ (ii)\,\mathrm{Tr}(\bar{\rho}^{*(k)} \bar{\rho}^{*(k')}) = B\delta_{k,k'};\ (iii)\,\mathrm{Tr}(\bar{\rho}^{*(k)} o^{(k')}) = \delta_{k,k'}. \tag{A2}$$

Proof of Theorem 1. Following Eq. (2) and the results in Lemma 3, with probability $1-\delta$, the expected risk of an optimal empirical QC is upper bounded by

$$\mathsf{R}(\hat{h}_Q) \leq \mathsf{R}_{\mathrm{ERM}}(\hat{h}_Q) + 4L_1 K C_2 \epsilon + 3\xi(\hat{h}) \sqrt{\frac{|\mathcal{T}_{\mathcal{D}}| 4^m N_{ge} \ln(56KN_{ge}/(\epsilon\delta))}{n}} + \xi(\hat{h}) \frac{2|\mathcal{T}_{\mathcal{D}}| 4^m N_{ge} \ln(56KN_{ge}/(\epsilon\delta))}{n}. \tag{A3}$$

Then, when $\boldsymbol{\rho}^*$ satisfies Eq. (A2), Lemma 2 warrants $\mathsf{R}_{\mathrm{ERM}}(\hat{h}_Q) = 0$, which gives

$$\mathsf{R}(\hat{h}_Q) \leq 4L_1 K C_2 \epsilon + 3\xi(\hat{h}) \sqrt{\frac{|\mathcal{T}_{\mathcal{D}}| 4^m N_{ge} \ln(56KN_{ge}/(\epsilon\delta))}{n}} + \xi(\hat{h}) \frac{2|\mathcal{T}_{\mathcal{D}}| 4^m N_{ge} \ln(56KN_{ge}/(\epsilon\delta))}{n}. \tag{A4}$$

This bound can be further simplified when the training of QC is perfect. Note that Condition (i) implies $|\mathcal{T}_{\mathcal{D}}| = K$, since all feature states from the same class collapse to the same point. Meanwhile, since $\xi(\hat{h})$ and C_2 are bounded, and m and ϵ are small constant, we can conclude that when $n \gg O(KN_{ge} \log(KN_g/(\epsilon\delta)))$, the expected risk can approach to zero. □

SM B: Proof of Corollary 1

The proof leverages the following two lemmas related to the Haar measure and the unitary t-design.

Lemma 4. *Let $\{W_y\}_{y \in Y} \subset U(d)$ form a unitary t-design with $t > 1$, and let $A, B : \mathcal{H}_d \to \mathcal{H}_d$ be arbitrary linear operators. Then*

$$\frac{1}{|Y|} \sum_{y \in Y} \mathrm{Tr}[W_y A W_y^\dagger B] = \int_{Haar} d\mu(W)\, \mathrm{Tr}[W_y A W_y^\dagger B] = \frac{\mathrm{Tr}[A]\,\mathrm{Tr}[B]}{d}. \tag{B1}$$

Lemma 5. *Let $\{W_y\}_{y\in Y}\subset U(d)$ form a unitary t-design with $t>1$, and let $A,B,C,D:\mathcal{H}_d\to\mathcal{H}_d$ be arbitrary linear operators. Then*

$$\begin{aligned}\frac{1}{|Y|}\sum_{y\in Y}\operatorname{Tr}[W_yAW_y^\dagger B]\operatorname{Tr}[W_yCW_y^\dagger D]=&\int_{Haar}d\mu(W)\operatorname{Tr}[W_yAW_y^\dagger B]\operatorname{Tr}[W_yCW_y^\dagger D]\\ =&\frac{1}{d^2-1}\left(\operatorname{Tr}[A]\operatorname{Tr}[B]\operatorname{Tr}[C]\operatorname{Tr}[D]+\operatorname{Tr}[AC]\operatorname{Tr}[BD]\right)\\ &-\frac{1}{d(d^2-1)}\left(\operatorname{Tr}[AC]\operatorname{Tr}[B]\operatorname{Tr}[D]+\operatorname{Tr}[A]\operatorname{Tr}[C]\operatorname{Tr}[BD]\right).\end{aligned}\tag{B2}$$

Corollary (Restatement of Corollary 1)**.** *Following notations in Lemmas 2 and 3, when the encoding unitary $\{U_E(\boldsymbol{x})|\boldsymbol{x}\in\mathcal{X}\}$ follows the Haar distribution, with probability $1-\delta$, the empirical QC follows $|\operatorname{Tr}\left(\sigma(\boldsymbol{x}^{(i,k)})\sigma(\boldsymbol{x})\right)-\frac{1}{2^N}|\le\sqrt{\frac{3}{2^{2N}\delta}}$. When the adopted Ansatz $\{U(\boldsymbol{\theta})|\boldsymbol{\theta}\in\Theta\}$ follows the Haar distribution, with probability $1-\delta$, the empirical QC follows $|\operatorname{Tr}(\rho^{(i,k)}o^{(k')})-\frac{\operatorname{Tr}(o^{(k')})}{2^D}|<\sqrt{\frac{\operatorname{Tr}(o^{(k')})^2+2\operatorname{Tr}((o^{(k')})^2)}{2^{2D}\delta}}$.*

Proof of Corollary 1. We complete the proof by separately analyzing the concentration behavior of the encoding unitary and the Ansätze.

Concentration of the encoding unitary. Recall that Condition (iii) in Lemma 2 concerns the distance between two feature states $\rho^{(i,k)}$ and $\rho^{(i',k')}$ for $\forall i,i\in[n_c]$ and $\forall k,k'\in[K]$. In this regard, we quantify the distance between the encoded state $\sigma(\boldsymbol{x}^{(i,k)})$ and $\sigma(\boldsymbol{x})$ with $\boldsymbol{x}\sim\mathcal{X}$ when the deep encoding Ansatz U_E is employed. In particular, we have

$$\begin{aligned}&\mathbb{E}_{\boldsymbol{x}\sim\mathcal{X}}\left(\operatorname{Tr}\left(\sigma(\boldsymbol{x}^{(i,k)})\sigma(\boldsymbol{x})\right)\right)\\ =&\mathbb{E}_{\boldsymbol{x}\sim\mathcal{X}}\left(\operatorname{Tr}\left(\sigma(\boldsymbol{x}^{(i,k)})U_E(\boldsymbol{x})(|0\rangle\langle 0|)^{\otimes N}U_E(\boldsymbol{x})^\dagger\right)\right)\\ =&\int_{\text{Haar}}d\mu(U)\operatorname{Tr}\left(\sigma(\boldsymbol{x}^{(i,k)})U(|0\rangle\langle 0|)^{\otimes N}U\right)\\ =&\frac{\operatorname{Tr}(\sigma(\boldsymbol{x}^{(i,k)}))\operatorname{Tr}(|0\rangle\langle 0|)^{\otimes N})}{2^N}\\ =&\frac{1}{2^N},\end{aligned}\tag{B3}$$

where the third equality uses Lemma 4. Moreover, the variance of the term $\operatorname{Tr}(\sigma(\boldsymbol{x}^{(i,k)})\sigma(\boldsymbol{x}))$ yields

$$\begin{aligned}&\operatorname{Var}_{\boldsymbol{x}\sim\mathcal{X}}\left(\operatorname{Tr}\left(\sigma(\boldsymbol{x}^{(i,k)})\sigma(\boldsymbol{x})\right)\right)\\ =&\mathbb{E}_{\boldsymbol{x}\sim\mathcal{X}}\left(\operatorname{Tr}\left(\sigma(\boldsymbol{x}^{(i,k)})\sigma(\boldsymbol{x})\right)^2\right)-\mathbb{E}_{\boldsymbol{x}\sim\mathcal{X}}\left(\operatorname{Tr}\left(\sigma(\boldsymbol{x}^{(i,k)})\sigma(\boldsymbol{x})\right)\right)^2\\ =&\int_{\text{Haar}}d\mu(U)\operatorname{Tr}\left(\sigma(\boldsymbol{x}^{(i,k)})U(|0\rangle\langle 0|)^{\otimes N}U\right)\operatorname{Tr}\left(\sigma(\boldsymbol{x}^{(i,k)})U(|0\rangle\langle 0|)^{\otimes N}U\right)-\frac{1}{2^{2N}}\\ =&\frac{1}{2^{2N}-1}\left(1+\operatorname{Tr}(\sigma(\boldsymbol{x}^{(i,k)})^2)\right)-\frac{1}{2^{2N}(2^{2N}-1)}\left(\operatorname{Tr}(\sigma(\boldsymbol{x}^{(i,k)})^2)+1\right)-\frac{1}{2^{2N}}\\ \le&\frac{1}{2^{2N-2}}-\frac{1}{2^{2N}}\\ =&\frac{3}{2^{2N}},\end{aligned}\tag{B4}$$

where the second equality uses the property that the deep encoding unitary follows the Haar distribution and the result in Eq. (B3), the third equality comes from Lemma 4, the inequality adopts $\operatorname{Tr}(\sigma^2)\le 1$ and $2^{2N}-1>2^{2N-1}$, and the last equality is obtained via simplification.

Supported by the Chebyshev's inequality $\Pr(|X-\mathbb{E}[X]|\ge a)\le\operatorname{Var}[X]/a^2$, Eqs. (B3) and (B4) indicate

$$\Pr\left(\left|\operatorname{Tr}\left(\sigma(\boldsymbol{x}^{(i,k)})\sigma(\boldsymbol{x})\right)-\frac{1}{2^N}\right|\ge\tau\right)\le\frac{3}{2^{2N}\tau^2}.$$

Equivalently, with probability $1-\delta$, we have

$$\left|\operatorname{Tr}\left(\sigma(\boldsymbol{x}^{(i,k)})\sigma(\boldsymbol{x})\right)-\frac{1}{2^N}\right|\le\sqrt{\frac{3}{2^{2N}\delta}}.\tag{B5}$$

Concentration of the deep Ansatze. Recall Condition (ii) in Lemma 2. Given a feature state $\rho^{(i,k)}$, for $\forall i \in [n_c]$ and $\forall k \in [K]$ and a measure operator $o^{(k)}$, the optimal feature state should satisfy

$$\mathrm{Tr}(\rho^{*(i,k)} o^{(k')}) = \delta_{k,k'}.$$

In other words, we should examine the value of $\mathrm{Tr}(\rho^{(i,k)} o^{(k')})$ when $\rho^{(i,k)}$ is prepared by a deep Ansatze $U(\boldsymbol{\theta})$. Specifically, we have

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{\theta}\sim\Theta}\left(\mathrm{Tr}(\rho^{(i,k)} o^{(k')})\right)\\
=&\mathbb{E}_{\boldsymbol{\theta}\sim\Theta}\left(\mathrm{Tr}(U(\boldsymbol{\theta})\sigma(\boldsymbol{x}^{(i,k)})U(\boldsymbol{\theta})^\dagger (o^{(k')}\otimes \mathbb{I}_{2^{N-D}})\right)\\
=&\int_{\mathrm{Haar}} d\mu(U)\,\mathrm{Tr}\left(U\sigma(\boldsymbol{x}^{(i,k)})U^\dagger (o^{(k')}\otimes \mathbb{I}_{2^{N-D}})\right)\\
=&\frac{\mathrm{Tr}(o^{(k')})(2^{N-D})}{2^N}\\
=&\frac{\mathrm{Tr}(o^{(k')})}{2^D},
\end{aligned} \tag{B6}$$

where the first equality comes from the explicit form of QC in Eq. (4), the second equality uses the fact that U follows the Haar distribution, and the last second equality comes from Lemma 4.

We then quantify the variance of $\mathrm{Tr}(\rho^{(i,k)} o^{(k')})$, i.e.,

$$\begin{aligned}
&\mathrm{Var}_{\boldsymbol{\theta}\sim\Theta}\left(\mathrm{Tr}(\rho^{(i,k)} o^{(k')})\right)\\
=&\mathbb{E}_{\boldsymbol{\theta}\sim\Theta}\left(\mathrm{Tr}(\rho^{(i,k)} o^{(k')})^2\right) - \left(\mathbb{E}_{\boldsymbol{\theta}\sim\Theta}\left(\mathrm{Tr}(\rho^{(i,k)} o^{(k')})\right)\right)^2\\
=&\int_{\mathrm{Haar}} d\mu(U)\,\mathrm{Tr}\left(U\sigma(\boldsymbol{x}^{(i,k)})U^\dagger (o^{(k')}\otimes \mathbb{I}_{2^{N-D}})\right)^2 - \frac{\mathrm{Tr}(o^{(k')})^2}{2^{2D}}\\
=&\frac{1}{2^{2N}-1}\left(\mathrm{Tr}(\sigma(\boldsymbol{x}^{(i,k)}))\,\mathrm{Tr}(o^{(k')}\otimes \mathbb{I}_{2^{N-D}})\,\mathrm{Tr}(\sigma(\boldsymbol{x}^{(i,k)}))\,\mathrm{Tr}(o^{(k')}\otimes \mathbb{I}_{2^{N-D}}) + \mathrm{Tr}(\sigma(\boldsymbol{x}^{(i,k)})^2)\,\mathrm{Tr}((o^{(k')}\otimes \mathbb{I}_{2^{N-D}})^2)\right)\\
&-\frac{1}{2^N(2^{2N}-1)}\left(\mathrm{Tr}(\sigma(\boldsymbol{x}^{(i,k)})^2)\,\mathrm{Tr}(o^{(k')}\otimes \mathbb{I}_{2^{N-D}})^2 + \mathrm{Tr}(\sigma(\boldsymbol{x}^{(i,k)}))^2\,\mathrm{Tr}((o^{(k')}\otimes \mathbb{I}_{2^{N-D}})^2)\right) - \frac{\mathrm{Tr}(o^{(k')})^2}{2^{2D}}\\
\le&\frac{1}{2^{2N}-1}\left(\mathrm{Tr}(o^{(k')}\otimes \mathbb{I}_{2^{N-D}})^2 + \mathrm{Tr}((o^{(k')}\otimes \mathbb{I}_{2^{N-D}})^2)\right) - \frac{\mathrm{Tr}(o^{(k')})^2}{2^{2D}}\\
=&\frac{1}{2^{2N}-1}\left(\mathrm{Tr}(o^{(k')})^2 2^{2N-2D} + \mathrm{Tr}((o^{(k')})^2) 2^{2N-2D}\right) - \frac{\mathrm{Tr}(o^{(k')})^2}{2^{2D}}\\
\le&\frac{\mathrm{Tr}(o^{(k')})^2 + \mathrm{Tr}((o^{(k')})^2)}{2^{2D-1}} - \frac{\mathrm{Tr}(o^{(k')})^2}{2^{2D}}\\
=&\frac{\mathrm{Tr}(o^{(k')})^2 + 2\,\mathrm{Tr}((o^{(k')})^2)}{2^{2D}}.
\end{aligned} \tag{B7}$$

where the second equality uses the fact that U follows the Haar distribution and Eq. (B6), the the third equality comes from Lemma 5, the first inequality arises from dropping some positive terms, the last second equality employs $\mathrm{Tr}(A\otimes B) = \mathrm{Tr}(A)\,\mathrm{Tr}(B)$ and $(A\otimes B)(C\otimes D) = (AC)\otimes(BD)$, and the last inequality exploits $(2^{2N}-1)^{-1} > (2^{N-1})^{-1}$, and the last equalities is obtained via simplification.

Supported by the Chebyshev's inequality $\Pr(|X - \mathbb{E}[X]| \ge a) \le \mathrm{Var}[X]/a^2$, Eqs. (B6) and (B7) indicate

$$\Pr\left(\left|\,\mathrm{Tr}(\rho^{(i,k)} o^{(k')}) - \mathbb{E}\left(\mathrm{Tr}(\rho^{(i,k)} o^{(k')})\right)\,\right| \ge \tau\right) \le \frac{\mathrm{Tr}(o^{(k')})^2 + 2\,\mathrm{Tr}((o^{(k')})^2)}{2^{2D}\tau^2}.$$

Equivalently, with probability $1-\delta$, we have

$$\left|\,\mathrm{Tr}(\rho^{(i,k)} o^{(k')}) - \frac{\mathrm{Tr}(o^{(k')})}{2^D}\right| < \sqrt{\frac{\mathrm{Tr}(o^{(k')})^2 + 2\,\mathrm{Tr}((o^{(k')})^2)}{2^{2D}\delta}}. \tag{B8}$$

□

SM C: Proof of Lemma 1

In this section, we derive the geometric properties of the global optimizer under the unconstraint loss function $\mathcal{L}(\boldsymbol{\rho}, \boldsymbol{o})$ in which both $\boldsymbol{\rho}$ and $\boldsymbol{o}$ are tunable and the regularization term is considered. Mathematically, the regularizer in Eq. (1) is defined as $\mathfrak{E} = \frac{\lambda_\rho}{2}\sum_{i=1}^{n_c}\sum_{k=1}^{K}\|\rho^{(i,k)}\|_F^2 + \frac{\lambda_o}{2}\sum_{k=1}^{K}\|o^{(k)}\|_F^2$ with λ_ρ and λ_o being hyper-parameters. The explicit form of the loss function is

$$\mathcal{L}(\boldsymbol{\rho}, \boldsymbol{o}) = \frac{1}{2n}\sum_{i=1}^{n_c}\sum_{k=1}^{K}\left(\left[\mathrm{Tr}(\rho^{(i,k)}o^{(k)})\right]_{k=1:K} - \boldsymbol{y}^{(i,k)}\right)^2 + \frac{\lambda_\rho}{2}\sum_{i=1}^{n_c}\sum_{k=1}^{K}\|\rho^{(i,k)}\|_F^2 + \frac{\lambda_o}{2}\sum_{j=1}^{K}\|o^{(j)}\|_F^2. \tag{C1}$$

Denote the global optima as $(\boldsymbol{\rho}^*, \boldsymbol{o}^*) = \min_{\boldsymbol{\rho},\boldsymbol{o}}\mathcal{L}(\boldsymbol{\rho}, \boldsymbol{o})$ and the empirical QC as $\hat{h}_Q \equiv h_Q(\boldsymbol{\rho}^*, \boldsymbol{o}^*)$. The restatement of Lemma 1 is as follows.

Lemma (Formal statement of Lemma 1)**.** *Define $C_1 := K\sqrt{n_c\lambda_o\lambda_\rho}$. If $2^D \geq K$, $C_1 \leq 1$, and $\lambda_o \leq n_c\lambda_\rho$, the global minimizer $(\boldsymbol{\rho}^*, \boldsymbol{o}^*)$ of $\mathcal{L}(\boldsymbol{\rho}, \boldsymbol{o})$ in Eq. (C1) satisfies for $\forall k, k' \in [K]$:*

$$\begin{aligned}
&(i)\bar{\rho}^{*(k)} := \rho^{*(1,k)} = \cdots = \rho^{*(n_c,k)};\\
&(ii)\,\mathrm{Tr}(\bar{\rho}^{*(k)}\bar{\rho}^{*(k')}) = (1 - C_1)\sqrt{\frac{\lambda_o}{n\lambda_\rho}}\delta_{k,k'};\\
&(iii)o^{*(k)} = \sqrt{\frac{n\lambda_\rho}{\lambda_o}}\bar{\rho}^{*(k)}.
\end{aligned} \tag{C2}$$

The corresponding empirical risk is $\mathsf{R}_{\mathrm{ERM}}(\hat{h}_Q) = C_1$.

Proof of Lemma 1. Conceptually, the global optimizer can be identified by lower bounding $\mathcal{L}(\boldsymbol{\rho}, \boldsymbol{o})$, where the equality conditions of $\boldsymbol{\rho}$ amount to the properties of global minimizer. In particular, the lower bound of $\mathcal{L}(\boldsymbol{\rho}, \boldsymbol{o})$ yields

$$\begin{aligned}
&\frac{1}{2Kn_c}\sum_{i=1}^{n_c}\sum_{k=1}^{K}\left([\mathrm{Tr}(\rho^{(i,k)}o^{(j)})]_{j=1:K} - \boldsymbol{y}^{(i,k)}\right)^2 + \frac{\lambda_\rho}{2}\sum_{i=1}^{n_c}\sum_{k=1}^{K}\|\rho^{(i,k)}\|_F^2 + \frac{\lambda_o}{2}\sum_{j=1}^{K}\|o^{(j)}\|_F^2\\
\geq&\frac{1}{2Kn_c}\sum_{i=1}^{n_c}\sum_{k=1}^{K}\left(\mathrm{Tr}(\rho^{(i,k)}o^{(k)}) - 1\right)^2 + \frac{\lambda_\rho}{2}\sum_{i=1}^{n_c}\sum_{k=1}^{K}\|\rho^{(i,k)}\|_F^2 + \frac{\lambda_o}{2}\sum_{j=1}^{K}\|o^{(j)}\|_F^2\\
=&\frac{1}{2Kn_c}\sum_{k=1}^{K}\sum_{i=1}^{n_c}n_c\frac{1}{n_c}\left(\mathrm{Tr}(\rho^{(i,k)}o^{(k)}) - 1\right)^2 + \frac{\lambda_\rho}{2}\sum_{k=1}^{K}\sum_{i=1}^{n_c}n_c\frac{1}{n_c}\|\rho^{(i,k)}\|_F^2 + \frac{\lambda_o}{2}\sum_{j=1}^{K}\|o^{(j)}\|_F^2\\
\geq&\frac{1}{2K}\sum_{k=1}^{K}\left(\mathrm{Tr}\left(\sum_{i=1}^{n_c}\frac{1}{n_c}\rho^{(i,k)}o^{(k)}\right) - 1\right)^2 + \frac{\lambda_\rho}{2}\sum_{k=1}^{K}n_c\left\|\sum_{i=1}^{n_c}\frac{1}{n_c}\rho^{(i,k)}\right\|_F^2 + \frac{\lambda_o}{2}\sum_{j=1}^{K}\|o^{(j)}\|_F^2,
\end{aligned} \tag{C3}$$

where the first inequality uses the fact $\|\boldsymbol{a} - \boldsymbol{b}\|^2 = \sum_i(\boldsymbol{a}^{(i)} - \boldsymbol{b}^{(i)})^2 \geq (\boldsymbol{a}^{(k)} - \boldsymbol{b}^{(k)})^2$ and the k-th entry of $\boldsymbol{y}^{(i,k)}$ equals to 1, and the second inequality comes from the Jensen's inequality $f(\mathbb{E}(x)) \leq \mathbb{E}(f(x))$. The equality condition of the first inequality holds if and only if

$$\mathrm{Tr}\left(\rho^{(i,k)}o^{(j)}\right) = 0,\ \ (\forall j \in [K] \setminus \{k\}) \wedge (\forall i \in [n_c])\,;$$

and the equality condition of the second inequality holds if and only if

$$\rho^{(1,k)} = \cdots = \rho^{(i,k)} = \cdots = \rho^{(n_c,k)},\ \forall k \in [K].$$

Denote the mean of the feature state for the k-th class as $\bar{\rho}^{(k)} = \sum_{i=1}^{n_c}\frac{1}{n_c}\rho^{(i,k)}$ for $\forall k \in [K]$. The above two equality conditions suggest that the global minimizer $(\boldsymbol{\rho}^*, \boldsymbol{o}^*)$ satisfies

$$\begin{aligned}
&\bar{\rho}^{*(k)} \equiv \rho^{*(1,k)} = \cdots = \rho^{*(n_c,k)},\ \forall k \in [K]\\
&\mathrm{Tr}(\bar{\rho}^{*(k)}o^{*(j)}) = 0,\ \forall j \in [K] \setminus \{k\}.
\end{aligned} \tag{C4}$$

To thins end, we obtain Conditions (i) in Lemma 1, which describe the geometric properties of $\boldsymbol{\rho}^*$, i.e.,

$$(i)\bar{\rho}^{*(k)} := \rho^{*(1,k)} = \cdots = \rho^{*(n_c,k)}. \tag{C5}$$

The next part of the proof is showing that the global minimizer satisfies Condition (iii). Combining Eqs. (C3) and (C4), the lower bound of the loss function in Eq. (C3) follows

$$\begin{aligned}
&\mathcal{L}(\boldsymbol{\rho},\boldsymbol{o})\\
\geq&\frac{1}{2K}\sum_{k=1}^{K}\left(\mathrm{Tr}\left(\bar{\rho}^{(k)}o^{(k)}\right)-1\right)^2+\frac{\lambda_\rho}{2}\sum_{k=1}^{K}n_c\left\|\bar{\rho}^{(k)}\right\|_F^2+\frac{\lambda_o}{2}\sum_{j=1}^{K}\|o^{(j)}\|_F^2\\
=&\frac{1}{2K}\sum_{k=1}^{K}\left(\mathrm{Tr}\left(\bar{\rho}^{(k)}o^{(k)}\right)-1\right)^2+\frac{\lambda_\rho}{2}K\sum_{k=1}^{K}\frac{1}{K}n_c\left\|\bar{\rho}^{(k)}\right\|_F^2+\frac{\lambda_o}{2}K\sum_{j=1}^{K}\frac{1}{K}\|o^{(j)}\|_F^2\\
\geq&\frac{1}{2}\left(\sum_{k=1}^{K}\frac{1}{K}\mathrm{Tr}\left(\bar{\rho}^{(k)}o^{(k)}\right)-1\right)^2+\frac{\lambda_\rho}{2}Kn_c\left\|\sum_{k=1}^{K}\frac{1}{K}\bar{\rho}^{(k)}\right\|_F^2+\frac{\lambda_o}{2}K\|\sum_{j=1}^{K}\frac{1}{K}o^{(j)}\|_F^2,
\end{aligned}\tag{C6}$$

where the second inequality comes from the Jensen's inequality and the equality condition holds if and only if for $\forall k,k'\in[K]$,

$$\mathrm{Tr}\left(\bar{\rho}^{(k)}o^{(k)}\right)=\mathrm{Tr}\left(\bar{\rho}^{(k')}o^{(k')}\right),\ \|\bar{\rho}^{(k)}\|_F=\|\bar{\rho}^{(k')}\|_F,\ \|o^{(k)}\|_F=\|o^{(k')}\|_F.\tag{C7}$$

Then, supported by the inequality $a+b\geq 2\sqrt{ab}$, the loss $\mathcal{L}(\boldsymbol{\rho},\boldsymbol{o})$ can be further lower bounded by

$$\begin{aligned}
&\frac{1}{2}\left(\mathrm{Tr}\left(\bar{\rho}^{(k)}o^{(k)}\right)-1\right)^2+\frac{\lambda_\rho}{2}Kn_c\left\|\bar{\rho}^{(k)}\right\|_F^2+\frac{\lambda_o}{2}K\|o^{(j)}\|_F^2\\
\geq&\frac{1}{2}\left(\mathrm{Tr}\left(\bar{\rho}^{(k)}o^{(k)}\right)-1\right)^2+K\sqrt{n_c\lambda_o\lambda_\rho}\left\|\bar{\rho}^{(k)}\right\|_F\|o^{(j)}\|_F,
\end{aligned}\tag{C8}$$

where the equality condition holds if and only if

$$\lambda_o\|o^{(j)}\|_F^2=n_c\lambda_\rho\left\|\bar{\rho}^{(k)}\right\|_F^2,\forall k\in[K].\tag{C9}$$

Note that the requirements $C_1\leq 1$ and $\lambda_o\leq n_c\lambda_\rho$ in Lemma 1 imply $\|\bar{\rho}^{*(k)}\|\leq 1$ and hence ensure that $\bar{\rho}^{*(k)}$ is a meaningful quantum state for $\forall k\in[K]$.

Since $\mathrm{Tr}\left(\bar{\rho}^{(k)}o^{(k)}\right)=\|\bar{\rho}^{(k)}\|\|o^{(k)}\|\cos(\angle(\rho^{(k)},o^{(k)}))$, the lower bound of $\mathcal{L}(\boldsymbol{\rho},\boldsymbol{o})$ in Eq. (C8) is equivalent to

$$\frac{1}{2}\left(\|\bar{\rho}^{(k)}\|\|o^{(k)}\|\cos(\angle(\rho^{(k)},o^{(k)}))-1\right)^2+C_1\left\|\bar{\rho}^{(k)}\right\|_F\|o^{(j)}\|_F.$$

Define $\|\bar{\rho}^{(k)}\|\|o^{(k)}\|=a$ and $\angle(\rho^{(k)},o^{(k)})=\alpha$. The above equation is described by the function $f(a,\alpha)=(a\cos\alpha-1)^2/2+C_1a$ and its minimum is $C_1-C_1^2/2$ when $\alpha^*=0$ and $a^*=1-C_1$. The derivation is as follows. Since $a>0$ and its maxima is unbounded, we first consider the case $0<a<1$. In this case, the minimum of $f(a,\alpha)$ is $C_1-C_1^2/2$ with $\alpha^*=0$ and $a^*=1-C_1$. Otherwise, when $a\geq 1$, the minimum of $f(a,\alpha)$ is C_1 with $\alpha^*=\arccos(1/a)$ and $a^*=1$. Note that the minimum value of $f(a,\alpha)$ in the second case is always larger than that of the first case. Therefore, the minimum of $f(a,\alpha)$ is $C_1-C_1^2/2$ with $\alpha^*=0$ and $a^*=1-C_1$. Combining the observation that $\bar{\rho}^{*(k)}$ and $o^{(k)}$ are in the same direction with Eq. (C9), we achieve Condition (iii), i.e.,

$$o^{*(k)}=\sqrt{\frac{n_c\lambda_\rho}{\lambda_o}}\bar{\rho}^{*(k)}.$$

The last part is proving Condition (ii). Combining the result $\|\bar{\rho}^{*(k)}\|\|o^{*(k)}\|=1-C_1$ for $\forall k\in[K]$ with Eq. (C4) and Condition (iii), we immediately obtain condition (ii), i.e.,

$$(ii)\sqrt{\frac{n_c\lambda_\rho}{\lambda_o}}\|\rho^{*(k)}\|\|\rho^{*(k')}\|=(1-C_1)\delta_{k,k'}\Rightarrow\mathrm{Tr}(\bar{\rho}^{*(k)}\bar{\rho}^{*(k')})=(1-C_1)\sqrt{\frac{\lambda_o}{n_c\lambda_\rho}}\delta_{k,k'}.\tag{C10}$$

To summarize, given the global optima satisfying the above three conditions, the corresponding empirical risk is

$$\mathsf{R}_{\mathrm{ERM}}(\hat{h}_Q)=\frac{1}{2n}\sum_{i=1}^{n_c}\sum_{k=1}^{K}\left([\mathrm{Tr}(\rho^{*(i,k)}o^{*(k)})]_{k=1:K}-\boldsymbol{y}^{(i,k)}\right)^2=\frac{C_1^2}{2}\tag{C11}$$

□

SM D: Results related to Lemma 2

This section is composed of two parts. In SM D 1, we present the proof of Lemma 2. In SM D 2, we explain that the requirements in Lemma 2 are mild.

1. Proof of Lemma 2

Different from Lemma 1, here we focus the setting such that the regularization term is set as $\mathfrak{E} = 0$ and the operator $\boldsymbol{o}$ is predefined. The explicit form of the loss function $\mathcal{L}$ is defined in Eq. (A1). Denote the optimal feature states $\boldsymbol{\rho}^* = \min_{\boldsymbol{\rho}} \mathcal{L}(\boldsymbol{\rho})$, we quantify the value of $\mathsf{R}_{\mathrm{ERM}}(\hat{h}_Q)$ with $\hat{h}_Q \equiv h_Q(\boldsymbol{\rho}^*)$.

We emphasize that the modifications of $\mathfrak{E}$ and $\boldsymbol{o}$ allow a lower optimal empirical risk. Recall the results of Lemma 1. In the most general case, the optimal empirical risk depends on the regularization term, i.e., $\mathsf{R}_{\mathrm{ERM}}(\hat{h}_Q) \to C_1^2/2$. The dependance on C_1 motivates us to explore the empirical risk of QC when $\mathfrak{E} = 0$. Furthermore, Condition (iii) in Lemma 1 delivers the crucial properties of the optimal measure operator, i.e., the optimal measure operators are orthogonal with each other. Such properties contribute to construct a more effective QCs. Instead of optimizing, the measure operator $\boldsymbol{o}$ can be predefined by inheriting the properties proved in Lemma 1, that is, $\boldsymbol{o}$ are required to span the space $\mathbb{C}^{2^D \times 2^D}$ and satisfy $\mathrm{Tr}(o^{(k)} o^{(k')}) = B\delta_{k,k'}$ with $B \geq 1$ being a constant. Notably, these requirement are mild, covering frequently used measures such as computational basis and Pauli-based measures, as explained in SM D 2.

Lemma (Formal statement of Lemma 2)**.** *Suppose that the adopted measure operator $\boldsymbol{o}$ spans the space $\mathbb{C}^{2^D \times 2^D}$ and satisfies $\mathrm{Tr}(o^{(k)} o^{(k')}) = B\delta_{k,k'}$ where $B \geq 1$ is a constant. The empirical risk of $\hat{h}_Q$ is $\mathsf{R}_{\mathrm{ERM}}(\hat{h}_Q) = 0$ when the global minimizer $\boldsymbol{\rho}^*$ satisfies*

$$(i)\bar{\rho}^{*(k)} := \rho^{*(1,k)} = ... = \rho^{*(n_c,k)};\ (ii)\,\mathrm{Tr}(\bar{\rho}^{*(k)}\bar{\rho}^{*(k')}) = B\delta_{k,k'};\ (iii)\,\mathrm{Tr}(\bar{\rho}^{*(k)} o^{(k')}) = \delta_{k,k'}. \tag{D1}$$

Proof of Lemma 2. The concept of the proof is analogous to Lemma 1, i.e., the global optimizer is identified by lower bounding the loss $\mathcal{L}(\boldsymbol{\rho})$. To this end, the lower bound of $\mathcal{L}(\boldsymbol{\rho})$ yields

$$\begin{aligned}
&\frac{1}{2Kn_c}\sum_{i=1}^{n_c}\sum_{k=1}^{K}\left([\mathrm{Tr}(\rho^{(i,k)}o^{(j)})]_{j=1:K} - \boldsymbol{y}^{(i,k)}\right)^2 \\
\geq &\frac{1}{2Kn_c}\sum_{i=1}^{n_c}\sum_{k=1}^{K}\left(\mathrm{Tr}(\rho^{(i,k)}o^{(k)}) - 1\right)^2 \\
= &\frac{1}{2Kn_c}\sum_{k=1}^{K}\sum_{i=1}^{n_c} n_c \frac{1}{n_c}\left(\mathrm{Tr}(\rho^{(i,k)}o^{(k)}) - 1\right)^2 \\
\geq &\frac{1}{2K}\sum_{k=1}^{K}\left(\mathrm{Tr}\left(\sum_{i=1}^{n_c}\frac{1}{n_c}\rho^{(i,k)}o^{(k)}\right) - 1\right)^2,
\end{aligned} \tag{D2}$$

where the first inequality uses the facts $n = Kn_c$, $\|\boldsymbol{a} - \boldsymbol{b}\|^2 = \sum_i (\boldsymbol{a}^{(i)} - \boldsymbol{b}^{(i)})^2 \geq (\boldsymbol{a}^{(k)} - \boldsymbol{b}^{(k)})^2$, and only the k-th entry of $\boldsymbol{y}^{(i,k)}$ equals to 1, and the second inequality comes from the Jensen's inequality $\mathbb{E}(f(x)) \geq f(\mathbb{E}(x))$ when $f(\cdot)$ is convex. Note that the equality condition of the first inequality holds if and only if

$$\mathrm{Tr}(\rho^{(i,k)}o^{(j)}) = 0, (\forall j \in [K] \setminus \{k\}) \wedge (\forall i \in [n_c])\,;$$

And the equality condition of the second inequality holds if and only if

$$\rho^{(1,k)} = \cdots = \rho^{(i,k)} = \cdots = \rho^{(n_c,k)}, \forall k \in [K].$$

Denote the mean of the feature state for the k-th class as $\bar{\rho}^{(k)} = \sum_{i=1}^{n_c} \frac{1}{n_c}\rho^{(i,k)}$ for $\forall k \in [K]$. The above two equality conditions suggest that the global minimizer yields

$$\bar{\rho}^{*(k)} \equiv \rho^{*(1,k)} = \cdots = \rho^{*(n_c,k)},\ \forall k \in [K] \tag{D3}$$

$$\mathrm{Tr}(\bar{\rho}^{*(k)}o^{(j)}) = 0, \forall j \in [K] \setminus \{k\}. \tag{D4}$$

Combining Eqs. (D2)-(D4), the lower bound of the loss function $\mathcal{L}(\boldsymbol{\rho})$ satisfies

$$\frac{1}{2K}\sum_{k=1}^{K}\left(\mathrm{Tr}\left(\bar{\rho}^{(k)}o^{(k)}\right)-1\right)^2 \geq \frac{1}{2}\left(\sum_{k=1}^{K}\frac{1}{K}\mathrm{Tr}\left(\bar{\rho}^{(k)}o^{(k)}\right)-1\right)^2, \tag{D5}$$

where the inequality comes from the Jensen's inequality and the equality condition holds if and only if $\forall k, k' \in [K]$,

$$\mathrm{Tr}\left(\bar{\rho}^{(k)}o^{(k)}\right) = \mathrm{Tr}\left(\bar{\rho}^{(k')}o^{(k')}\right). \tag{D6}$$

Supported by Eq. (D6), we can further lower bound $\mathcal{L}(\boldsymbol{\rho})$ with

$$\frac{1}{2}\left(\mathrm{Tr}\left(\bar{\rho}^{(k)}o^{(k)}\right)-1\right)^2 \geq 0, \tag{D7}$$

where the equality condition is achieved when $\mathrm{Tr}(\bar{\rho}^{(k)}o^{(k)}) = 1$ for $\forall k \in [K]$.

Taken together, the global optimizer $\boldsymbol{\rho}^*$ should satisfy Condition (i)&(iii) in Lemma 2, where

$$\begin{aligned}&(i)\bar{\rho}^{*(k)} := \rho^{*(1,k)} = ... = \rho^{*(n_c,k)};\\ &(iii)\,\mathrm{Tr}(\bar{\rho}^{*(k)}o^{(k')}) = \delta_{k,k'}.\end{aligned} \tag{D8}$$

We last prove that Condition (iii) and the requirements of $\boldsymbol{o}$ lead to Condition (ii). In particular, denote the vectorization of $\rho^{*(k)}$ and $o^{(k)}$ as $|\rho^{*(k)}\rangle\rangle$ and $|o^{(k)}\rangle\rangle$, respectively. Condition (iii) can be rewritten as

$$\left\langle\left\langle \bar{\rho}^{*(k)}, o^{(k')}\right\rangle\right\rangle = \delta_{k,k'}. \tag{D9}$$

Moreover, since the set of measure operators $\{o^{(k)}\}$ is required to be complete in the space of $\mathbb{C}^{2^D}$ and $\mathrm{Tr}(o^{(k)}o^{(k')}) = B\delta_{k,k'}$ with $B \geq 1$ for $\forall k, k' \in [K]$, we have

$$\sum_k \left|o^{(k)}\right\rangle\!\!\right\rangle\left\langle\!\!\left\langle o^{(k)}\right|\right. = B\mathbb{I}_{2^D}.$$

Then, Condition (ii) can be derived as follows, i.e.,

$$\begin{aligned}&\mathrm{Tr}(\rho^{*(k)}\rho^{*(k')})\\ =&\langle\langle\bar{\rho}^{*(k)}|\mathbb{I}_{2^D}|\rho^{*(k')}\rangle\rangle\\ =&\frac{1}{B}\left\langle\left\langle \bar{\rho}^{*(k)}\right|\sum_{k''}|o^{(k'')}\rangle\rangle\langle\langle o^{(k'')}|\left|\rho^{*(k')}\right\rangle\right\rangle\\ =&\frac{1}{B}\left\langle\left\langle \bar{\rho}^{*(k)}\right||o^{(k)}\rangle\rangle\langle\langle o^{(k)}|\left|\rho^{*(k')}\right\rangle\right\rangle + \left\langle\left\langle \bar{\rho}^{*(k)}\right|\sum_{k''\neq k}|o^{(k'')}\rangle\langle o^{(k'')}|\left|\rho^{*(k')}\right\rangle\right\rangle\\ =&\frac{1}{B}\delta_{k,k'}.\end{aligned} \tag{D10}$$

□

2. Requirement of $\boldsymbol{o}$ used in Lemma 2

Here we elucidate that the requirements adopted in Lemma 2, i.e., $\boldsymbol{o}$ spans the complex space $2^D \times 2^D$ and satisfies $\mathrm{Tr}(o^{(k)}o^{(k')}) = B\delta_{k,k'}$ with $B \geq 1$, are mild. Specifically, the employed measurements in most QNN-based classifiers satisfy these requirements, including the computational basis measurements and Pauli measurements.

Computational basis measurements. In this setting, the local measurement $o^{(k)}$ is set as $|k\rangle\langle k|$ with $|k\rangle$ being the k-th computational basis for $\forall k \in [K]$. When $2^D = K$, $\{|k\rangle\}$ spans the whole space of $\mathbb{C}^{2^D\times 2^D}$ and we have $\mathrm{Tr}(o^{(k)}o^{(k')}) = (\langle k|k'\rangle)^2 = \delta_{k,k'}$ with $B = 1$. The assumptions are satisfied.

Pauli measurements. Denote the Pauli operation applied to the i-th qubit as $P_a^{(i)}$ with $a \in \{X, Y, Z, I\}$ for $\forall i \in [D]$. Then, there are in total 4^D Pauli strings $P = \otimes_{i=1}^{D} P_a^{(i)}$ that form a orthogonal basis for the space $\mathbb{C}^{2^D\times 2^D}$. With setting $2^D = K$, each $o^{(k)}$ corresponds to one Pauli string with $\mathrm{Tr}(o^{(k)}o^{(k')}) = K\delta_{k,k'}$ with $B = K$.

SM E: Proof of Lemma 3

For elucidating, let us restate Lemma 3 below and introduce the proof sketch before moving on to present the proof details.

Lemma (Formal statement of Lemma 3)**.** *Denote L_1 as the Lipschitz constant of ℓ in Eq. (1) with respect to h. Given a QC defined in Eq. (3), let $\mathcal{E}$ be a quantum channel with*

$$h_Q(\boldsymbol{x}, U(\boldsymbol{\theta}), O^{(k)}) \equiv \mathrm{Tr}(o^{(k)}\mathcal{E}(\sigma(\boldsymbol{x}))),\ \forall k \in [K]. \tag{E1}$$

Suppose the measure operator follows $\max_{k\in[K]} \|o^{(k)}\| \leq C_2$. The explicit form of the encoding unitary follows $U_E(\boldsymbol{x}) = \prod_{g=1}^{N_g} u_g(\boldsymbol{x}) \in \mathcal{U}(2^N)$ with the g-th quantum gate $u_g(\boldsymbol{x}) \in \mathcal{U}(2^m)$ operating with at most m qubits with $m \leq N$ and N_g gates consisting of N_{ge} variational gates and $N_g - N_{ge}$ fixed gates,

Following above notations and Definition 1, the empirical QC is $(K(\frac{28N_{ge}}{\epsilon})^{4^m N_{ge}}, 4L_1KC_2\epsilon)$-robust and with probability $1-\delta$, its generalization error yields

$$\mathsf{R}_{\mathrm{Gene}}(\hat{h}) \leq 4L_1KC_2\epsilon + 3\xi(\hat{h})\sqrt{\frac{|\mathcal{T}_{\mathcal{D}}|4^m N_{ge}\ln(56KN_{ge}/(\epsilon\delta))}{n}} + \xi(\hat{h})\frac{2|\mathcal{T}_{\mathcal{D}}|4^m N_{ge}\ln(56KN_{ge}/(\epsilon\delta))}{n},$$

where L_1 is the Lipschitz constant of ℓ with respect to h, $\mathcal{I}_r^{\mathcal{D}} = \{i \in [n] : \boldsymbol{z}^{(i)} \in \mathcal{C}_r\}$, $\xi(\hat{h}) := \max_{\boldsymbol{z}\in\mathcal{Z}} \ell(\hat{h}, \boldsymbol{z})$, and $\mathcal{T}_{\mathcal{D}} := \{r \in [R] : |\mathcal{I}_r^{\mathcal{D}}| \geq 1\}$.

The proof of Lemma 3 is established on the following lemma, which leverages the algorithmic robustness to quantify the upper bound of the generalization error.

Lemma 6 (Theorem 1, [105])**.** *If the learning algorithm $\mathcal{A}$ is $(R, \nu(\cdot))$-robust with $\{\mathcal{C}_r\}_{r=1}^R$, then for any $\delta > 0$, with probability at least $1-\delta$ over an i.i.d drawn of n samples $\mathcal{D} = \{\boldsymbol{z}^{(i)}\}_{i=1}^n$ with $\boldsymbol{z}^{(i)} = (\boldsymbol{x}^{(i)}, y^{(i)})$, the returned hypothesis $\hat{h}$ by $\mathcal{A}$ on $\mathcal{D}$ satisfies*

$$\mathsf{R}_{\mathrm{Gene}}(\hat{h}) \leq \nu(\mathcal{D}) + \xi(\hat{h})\left((\sqrt{2}+1)\sqrt{\frac{|\mathcal{T}_{\mathcal{D}}|\ln(2R/\delta)}{n}} + \frac{2|\mathcal{T}_{\mathcal{D}}|\ln(2R/\delta)}{n}\right), \tag{E2}$$

where $\mathcal{I}_r^{\mathcal{D}} = \{i \in [n] : \boldsymbol{z}^{(i)} \in \mathcal{C}_r\}$, $\xi(\hat{h}) := \max_{\boldsymbol{z}\in\mathcal{Z}}(\ell(\hat{h}, \boldsymbol{z}))$, and $\mathcal{T}_{\mathcal{D}} := \{r \in [R] : |\mathcal{I}_r^{\mathcal{D}}| \geq 1\}$.

The above result hints that given a hypothesis $\hat{h}$, its generalization error is upper bounded by the disjoint sets $\{\mathcal{C}_r\}_{r=1}^R$, where a lower cardinality R allows a lower generalization error. A natural approach to realize these disjoint partitions is covering number [70].

Definition 2 (Covering number, [65])**.** *Given a metric space $(\mathcal{U}, \|\cdot\|)$, 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 ϵ with a norm $\|\cdot\|$, i.e., $\sup_{A\in\mathcal{U}} \min_{B\in\mathcal{V}} \|A - B\| \leq \epsilon$.*

In conjunction with Lemma 6 and Definition 2, the analysis of $\mathsf{R}_{\mathrm{Gene}}(\hat{h})$ of an N-qubit QC amounts to quantifying the covering number of the space of the input quantum states, i.e.,

$$\mathcal{X}_Q = \left\{U_E(\boldsymbol{x})(|0\rangle\langle 0|)^{\otimes N}U_E(\boldsymbol{x})^\dagger \middle| \boldsymbol{x} \in \mathcal{X}\right\}. \tag{E3}$$

The following lemma connects the robustness and covering number of $\mathcal{X}_Q$ of QCs whose proof is provided in Sec. E 1.

Lemma 7. *Following the settings in Eqs. (E1)-(E3), the corresponding QC is $(K(\frac{28N_{ge}}{\epsilon})^{4^m N_{ge}}, 4L_1KC_2\|\mathcal{E}\|_\diamond\epsilon)$-robust.*

We are now ready to prove Lemma 3.

Proof of Lemma 3. The generalization error bound can be acquired by combining Lemmas 6 and 7, i.e.,

$$\begin{aligned}
\mathsf{R}_{\mathrm{Gene}}(\hat{h}) \leq& 4L_1KC_2\|\mathcal{E}\|_\diamond\epsilon + \xi(\hat{h})\left((\sqrt{2}+1)\sqrt{\frac{|\mathcal{T}_{\mathcal{D}}|\ln(2K(\frac{28N_{ge}}{\epsilon})^{4^m N_{ge}}/\delta)}{n}} + \frac{2|\mathcal{T}_{\mathcal{D}}|\ln(2K(\frac{28N_{ge}}{\epsilon})^{4^m N_{ge}}/\delta)}{n}\right)\\
\leq& 4L_1KC_2\|\mathcal{E}\|_\diamond\epsilon + \xi(\hat{h})\left(3\sqrt{\frac{|\mathcal{T}_{\mathcal{D}}|4^m N_{ge}\ln(56KN_{ge}/(\epsilon\delta))}{n}} + \frac{2|\mathcal{T}_{\mathcal{D}}|4^m N_{ge}\ln(56KN_{ge}/(\epsilon\delta))}{n}\right)\\
\leq& 4L_1KC_2\epsilon + \xi(\hat{h})\left(3\sqrt{\frac{|\mathcal{T}_{\mathcal{D}}|4^m N_{ge}\ln(56KN_{ge}/(\epsilon\delta))}{n}} + \frac{2|\mathcal{T}_{\mathcal{D}}|4^m N_{ge}\ln(56KN_{ge}/(\epsilon\delta))}{n}\right),
\end{aligned} \tag{E4}$$

where $\mathcal{I}_r^{\mathcal{D}} = \{i \in [n] : \boldsymbol{z}^{(i)} \in \mathcal{C}_r\}$, $\xi(\hat{h}) := \max_{\boldsymbol{z}\in\mathcal{Z}}(\ell(\hat{h}, \boldsymbol{z}))$, and $\mathcal{T}_{\mathcal{D}} := \{r \in [R] : |\mathcal{I}_r^{\mathcal{D}}| \geq 1\}$. □

1. Proof of Lemma 7

The proof uses the following lemma to quantify the covering number of $\mathcal{X}_Q$ whose proof is given in SM E 2.

Lemma 8. *Following the settings in Eq. (E1), the covering number of $\mathcal{X}_Q$ in Eq. (E3) is*

$$\mathcal{N}(\mathcal{X}_Q, \epsilon, \|\cdot\|_F) \leq \left(\frac{28N_{ge}}{\epsilon}\right)^{4^m N_{ge}}. \tag{E5}$$

Proof of Lemma 7. When QC is applied to accomplish the K-class classification task, the sample space is $\mathcal{Z} = \mathcal{X}_Q \times \mathcal{Y}$ with $\mathcal{Y} = \{1, 2, ..., K\}$. Denote $\tilde{\mathcal{X}}_Q$ as the ϵ-cover set of $\mathcal{X}_Q$ with the covering number $\mathcal{N}(\mathcal{X}_Q, \epsilon, \|\cdot\|_F)$ in Definition 2. Supported by the ϵ-cover set $\tilde{\mathcal{X}}_Q$, the space $\mathcal{X}_Q \times \{i\}$ can be divided into $\mathcal{N}(\mathcal{X}_Q, \epsilon, \|\cdot\|_F)$ sets for $\forall i \in [K]$. In other words, we can divide $\mathcal{Z}$ into $K\mathcal{N}(\mathcal{X}_Q, \epsilon, \|\cdot\|_F)$ sets denoted by $\{\mathcal{Z}_i\}_{i=1}^{K\mathcal{N}(\mathcal{X}_Q,\epsilon,\|\cdot\|_F)}$.

We then utilize the divided sets of $\mathcal{Z}$ to connect the robustness with covering number according to Definition 1. Given a training example $(\boldsymbol{x}^{(i)}, y^{(i)})$ and a test example $(\boldsymbol{x}, y)$, suppose that the corresponding quantum examples $(\sigma(\boldsymbol{x}^{(i)}), y^{(i)})$ and $(\sigma(\boldsymbol{x}), y)$ are in the same set of $\{\mathcal{Z}_i\}_{i=1}^{K\mathcal{N}(\mathcal{X}_Q,\epsilon,\|\cdot\|_F)}$. For convenience, we abbreviate $\sigma(\boldsymbol{x}^{(i)})$ and $\sigma(\boldsymbol{x})$ as $\sigma^{(i)}$ and σ, respectively. Following the definition of covering number, we have

$$y^{(i)} = y \text{ and } \|\sigma^{(i)} - \sigma\|_F \leq 2\epsilon. \tag{E6}$$

Since the encoded state takes the form $\sigma = U_E(\boldsymbol{x})(|0\rangle\langle 0|)^{\otimes N} U_E(\boldsymbol{x})^\dagger$, we have

$$rank(\sigma^{(i)} - \sigma) \leq 2. \tag{E7}$$

Then, in accordance with the definition of robustness, we bound the discrepancy of the loss values for $\sigma^{(i)}$ and σ, i.e.,

$$\begin{aligned}
&\left| l(h_Q(\sigma^{(i)}), y^{(i)}) - l(h_Q(\sigma), y) \right| \\
\leq& L_1 \left\| [\mathrm{Tr}(\mathcal{E}(\sigma^{(i)})o^{(k)})]_{k=1:K} - [\mathrm{Tr}(\mathcal{E}(\sigma))o^{(k)})]_{k=1:K} \right\|_2 \\
\leq& L_1 K \max_{k\in K} |\,\mathrm{Tr}(\mathcal{E}(\sigma^{(i)}))o^{(k)}) - \mathrm{Tr}(\mathcal{E}(\sigma)o^{(k)})| \\
\leq& L_1 K \max_k \left\| o^{(k)} \right\|_2 \mathrm{Tr}(|\mathcal{E}(\sigma^{(i)} - \sigma)|) \\
\leq& 2L_1 K C_2 \|\mathcal{E}\|_\diamond \|\sigma^{(i)} - \sigma\|_F \\
\leq& 4L_1 K C_2 \|\mathcal{E}\|_\diamond \epsilon,
\end{aligned} \tag{E8}$$

where the first inequality uses the Lipschitz property of the loss function with $\ell(\boldsymbol{a}, \boldsymbol{b}) - \ell(\boldsymbol{c}, \boldsymbol{d}) \leq L_1 \|\boldsymbol{a} - \boldsymbol{c}\|_2$ and the form of $\mathcal{E}$ in Lemma 7, the second inequality comes from the definition of l_2 norm, the third inequality exploits von Neumann's trace inequality $|\,\mathrm{Tr}(AB)| \leq \|A\|_p \|B\|_q$ with $1/p + 1/q = 1$ and the linear property of CPTP map with $\mathcal{E}(\rho) - \mathcal{E}(\sigma) = \mathcal{E}(\rho - \sigma)$, the last second inequality employs $\max_k \left\|o^{(k)}\right\|_2 \leq C_2$, the relation $\|\mathcal{E}(\rho - \sigma)\|_1 \leq \|\mathcal{E}\|_\diamond \|\rho - \sigma\|_1$ and $\|A\|_1 \leq rank(A)\|A\|_F$, and the last inequality adopts the result in Eq. (E6).

The above result exhibits that the learned QC is $(K\mathcal{N}(\mathcal{X}_Q, \epsilon, \|\cdot\|), 4L_1 K C_2 \|\mathcal{E}\|_\diamond \epsilon)$-robust. In this regard, the proof can be completed when the upper bound of the covering number $\mathcal{N}(\mathcal{X}_Q, \epsilon, \|\cdot\|_F)$ is known. Supported by Lemma 8, we obtain $\mathcal{N}(\mathcal{X}_Q, \epsilon, \|\cdot\|_F) \leq (\frac{28N_{ge}}{\epsilon})^{4^m N_{ge}}$. Taken together, the learned QC is

$$\left(K \left(\frac{28N_{ge}}{\epsilon}\right)^{4^m N_{ge}}, 4L_1 K C_2 \|\mathcal{E}\|_\diamond \epsilon \right) - robust.$$

□

2. Proof of Lemma 8

The derivation of the covering number of $\mathcal{X}_Q$ in Eq. (E3) uses the following lemma.

Lemma 9 (Lemma 1, [106])**.** *For $0 < \epsilon < 1/10$, the ϵ-covering number for the unitary group $U(2^m)$ with respect to the Frobenius-norm distance in Definition 2 obeys*

$$\left(\frac{3}{4\epsilon}\right)^{4^m} \leq \mathcal{N}(U(2^m), \epsilon, \|\cdot\|_F) \leq \left(\frac{7}{\epsilon}\right)^{4^m}. \tag{E9}$$

Proof of of Lemma 8. Recall the input state space is $\mathcal{X}_Q = \{U_E(\boldsymbol{x})(|0\rangle\langle 0|)^{\otimes N}U_E(\boldsymbol{x})^\dagger | \boldsymbol{x} \in \mathcal{X}\}$, where the encoding unitary $U_E(\boldsymbol{x}) = \prod_{g=1}^{N_g} u_g(\boldsymbol{x}) \in \mathcal{U}(2^N)$ consists of N_{ge} variational gates and $N_g - N_{ge}$ fixed gates. To quantify the covering number $\mathcal{N}(\mathcal{X}_Q, \epsilon, \|\cdot\|_F)$, we define $\tilde{S}$ as the ϵ-covering set for the unitary group $U(2^m)$, $\tilde{\mathcal{X}}_Q$ as the ϵ'-covering set of $\mathcal{X}_Q$, and define a set

$$\tilde{\mathcal{U}}_E := \left\{ \prod_{i\in\{N_{ge}\}} u_i(\boldsymbol{x}) \prod_{j\in\{N_g - N_{ge}\}} u_j(\boldsymbol{x}) \Big| u_i(\boldsymbol{x}) \in \tilde{S} \right\}, \tag{E10}$$

where $u_i(\boldsymbol{\theta}_i)$ and u_j specify to the variational and fixed quantum gates, respectively. Note that for any encoding circuit $U_E(\boldsymbol{x})$, we can always find a unitary $U_{E,\epsilon}(\boldsymbol{x}) \in \tilde{\mathcal{U}}_E$ where each $u_g(\boldsymbol{x})$ is replaced by the nearest element in the covering set $\tilde{S}$. To this end, following the definition of covering number, the discrepancy between $U_E(\boldsymbol{x})(|0\rangle\langle 0|)^{\otimes N}U_E(\boldsymbol{x})^\dagger \in \mathcal{X}_Q$ and $U_{E,\epsilon}(\boldsymbol{x})(|0\rangle\langle 0|)^{\otimes N}U_{E,\epsilon}(\boldsymbol{x})^\dagger \in \tilde{\mathcal{X}}_Q$ under the Frobenius norm satisfies

$$\begin{aligned}
&\left\|U_E(\boldsymbol{x})(|0\rangle\langle 0|)^{\otimes N}U_E(\boldsymbol{x})^\dagger - U_{E,\epsilon}(\boldsymbol{x})(|0\rangle\langle 0|)^{\otimes N}U_{E,\epsilon}(\boldsymbol{x})^\dagger\right\|_F \\
\leq& 2\left\|U_E(\boldsymbol{x})(|0\rangle\langle 0|)^{\otimes N}U_E(\boldsymbol{x})^\dagger - U_{E,\epsilon}(\boldsymbol{x})(|0\rangle\langle 0|)^{\otimes N}U_{E,\epsilon}(\boldsymbol{x})^\dagger\right\| \\
\leq& 2\|U_E(\boldsymbol{x}) - U_{E_\epsilon}(\boldsymbol{x})\|\|(|0\rangle\langle 0|)^{\otimes N}\| \\
\leq& 4N_{ge}\epsilon,
\end{aligned} \tag{E11}$$

where the first inequality uses $\|X\|_F \leq rank(X)\|X\|$ and the relation in Eq. (E7), the second inequality comes from the Cauchy–Schwarz inequality, and the last inequality follows $\|U_E(\boldsymbol{x}) - U_{E,\epsilon}(\boldsymbol{x})\| \leq N_{ge}\epsilon$ and $\|(|0\rangle\langle 0|)^{\otimes N}\| = 1$. In other words, $\epsilon' = 2N_{ge}\epsilon$ and $\tilde{\mathcal{X}}_Q$ is a $(4N_{ge}\epsilon)$-covering set for $\mathcal{X}_Q$. In conjunction with the observation that there are $|\tilde{\mathcal{S}}|^{N_{ge}}$ combinations for the gates in $\tilde{\mathcal{U}}_E$ and the results in Lemma 9, we obtain the cardinality of the set $\tilde{\mathcal{U}}_E$ is upper bounded by $|\tilde{\mathcal{U}}_E| \leq \left(\frac{7}{\epsilon}\right)^{4^m N_{ge}}$. Accordingly, supported by Eq. (E11), the covering number of $\mathcal{X}_Q$ satisfies

$$\mathcal{N}(\mathcal{X}_Q, 4N_{ge}\epsilon, \|\cdot\|_F) \leq \left(\frac{7}{\epsilon}\right)^{4^m N_{ge}}. \tag{E12}$$

After simplification, we have

$$\mathcal{N}(\mathcal{X}_Q, \epsilon, \|\cdot\|_F) \leq \left(\frac{28N_{ge}}{\epsilon}\right)^{4^m N_{ge}}. \tag{E13}$$

□

SM F: Implementation of the algorithm to probe potential advantages of QCs

The expected risk is the most principal criteria to quantify the power of a classifier. As a result, to probe whether a QC holds potential advantages over a CC on a specific learning task, the simplest way is comparing their risk curves. Nevertheless, capturing these two risk curves are difficult, because of many flexible hyper-parameter settings to initiate a classifier.

The developed theories in Theorem 1 and Lemmas 1-3 deliver concrete rules to set up these hyper-parameters and thus allow an efficient way to estimate these risk curves. In particular, the derived U-shape curve of QCs indicates that the minimum risk of QC locates at the modest size of the hypothesis space $\mathcal{H}_Q$. In other words, the number of trainable parameters N_T should be lower than $O(poly(N))$, with N being the number of qubits in QC. Moreover, Lemma 3 hints that the generalization error of QC can be well suppressed by using the modest number of train examples. As such, if the available number of training examples in $\mathcal{D}$ is tremendous, we can distill a subset from $\mathcal{D}$ to better recognize quantum advantages.

The Pseudo code of the proposed method is presented in Alg. 1. To make a fair comparison, the hyper-parameter settings applied to QC and CC, especially for those relating to the computational resources, are required to keep to be the same. Specifically, in each comparison, the employed loss function, the train examples n, the number of trainable parameters N_t, and the number of epochs T applied to QC and CC should be identical. Note that the learning rate, the adopted optimizer, and the batch size can be varied of different classifiers to better estimate the empirical hypothesis. To ensure that the collected results of QC span its basin of the risk curve, the employed W settings of N_t can be acquired by uniformly interpolating from $O(1)$ to $O(poly(N))$. The iteration T should ensure the convergence of QC. Once the loss values of QC and CC under $\{n^{(w)}, N_t^{(w)}, T^{(w)}\}_{w=1}^W$ are obtained, we can apply certain fitting algorithms to attain their risk curves.

Algorithm 1: Estimate risk curves of quantum and classical classifiers

Data: The train dataset $\mathcal{D}$, the test dataset $\mathcal{D}_{Test}$, QC h_Q associated with the hypothesis space $\mathcal{H}_Q$, CC h_C associated with the hypothesis space $\mathcal{H}_Q$, the loss function $\mathcal{L}(\cdot,\cdot)$.

Result: The estimated risk curves of QC and CC.

Initialization: W tuples of hyper-parameter settings $\{n^{(w)}, N_t^{(w)}, T^{(w)}\}_{w=1}^{W}$ with n being train examples, N_t being the number of trainable parameters, and T being the number of epochs;

for $w = 1$, $w \leq W$, $w++$ **do**

Initialize train data as $\mathcal{D}^{(w)}$ by distilling $n^{(w)}$ examples from $\mathcal{D}$;

Collect loss dynamics of QC ;

Minimize the loss function $\mathcal{L}(\cdot,\cdot)$ via gradient descent methods to obtain the empirical quantum classifier $\bar{h}_Q^{(w)} \in \mathcal{H}_Q$ using $\mathcal{D}^{(w)}$ within $T^{(w)}$ epochs and N_T trainable parameters;

Record the loss value $\mathcal{L}(\bar{h}_Q^{(w)}, \mathcal{D}_{Test})$;

Collect loss dynamics of CC ;

Minimize the loss function $\mathcal{L}(\cdot,\cdot)$ via gradient descent methods to obtain the empirical classical classifier $\bar{h}_C^{(w)} \in \mathcal{H}_C$ using $\mathcal{D}^{(w)}$ within $T^{(w)}$ epochs and N_T trainable parameters;

Record the loss value $\mathcal{L}(\bar{h}_C^{(w)}, \mathcal{D}_{Test})$;

end

Fitting the loss dynamics of $\{\mathcal{L}(\bar{h}_Q^{(w)}, \mathcal{D}_{Test})\}_{w=1}^{W}$ to obtain the estimated risk curve of QC ;

Fitting the loss dynamics of $\{\mathcal{L}(\bar{h}_C^{(w)}, \mathcal{D}_{Test})\}_{w=1}^{W}$ to obtain the estimated risk curve of CC.

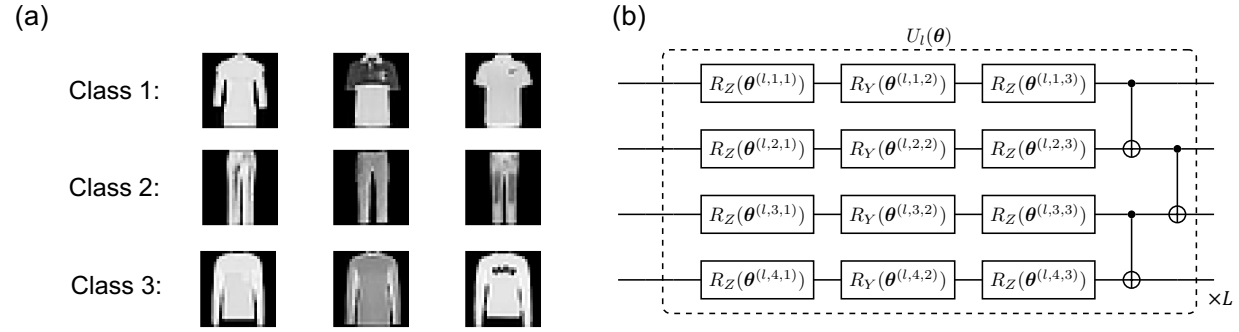

FIG. G.4. **Visualization of image dataset and hardware-efficient Ansatz.** (a) Image instances sampled from the Fashion-MNIST dataset. (b) The circuit architecture of the employed Hardware-efficient Ansatz. The label '$\times L$' denotes the layer number, which means repeating the gates in the dashed box with L times.

SM G: Numerical simulation details

Dataset. The construction of the parity dataset mainly follows from Ref. [98]. Note that this task has also been broadly studied in the field of deep learning to show the limits of deep neural classifiers [107, 108]. The constructed dataset contains in total 64 examples. Each example corresponds to a bit-string with the length 6, i.e., $\boldsymbol{x} \in \{0,1\}^6$. The label of $\boldsymbol{x}$ is assigned to be 1 if the number of '0' in $\boldsymbol{x}$ is even; otherwise, the label is 0. We split it into train dataset and test dataset with the train-test-split ratio being 0.75. The number of train examples in each class is controlled to be the same. For each example, its feature dimension is 10. The image dataset is adapted from Ref. [102]. Specifically, the data from the first nine classes are preserved and the total number of examples is 180. The train-test-split ratio is set as 0.5 to construct the train and test dataset. Each example corresponds to an image with 28×28 pixels. In the preprocessing stage, we flatten all examples followed by padding and normalization. The processed example yields an 10-qubit state with $\boldsymbol{x} \in \mathbb{R}^{2^{10}}$ and $\|\boldsymbol{x}\|_2^2 = 1$. Some examples after preprocessing are illustrated in Fig. G.4(a).

Construction of QCs. The quantum subroutine of QC consists of the encoding circuit U_E and the Ansatz $U(\boldsymbol{\theta})$. For all learning tasks, the hardware-efficient Ansatz is employed whose mathematical expression is $U(\boldsymbol{\theta}) = \prod_l^L U_l(\boldsymbol{\theta})$. The layout of the hardware-efficient Ansatz follows the layer-wise structure and the gate arrangement at each layer is the same. For $\forall l \in [L]$, $U_l(\boldsymbol{\theta}) = \bigotimes_{i=1}^{N}(\mathrm{RZ}(\boldsymbol{\theta}^{(l,i,1)})\,\mathrm{RY}(\boldsymbol{\theta}^{(l,i,2)})\,\mathrm{RZ}(\boldsymbol{\theta}^{(l,i,3)}))U_{ent}$ with U_{ent} being the entanglement layer formed by CNOT gates. Fig. G.4(b) depicts the adopted hardware-efficient Ansatz with L layers.

The encoding methods for the parity dataset classification and the digit images classification are different. The former uses the basis encoding method. Specifically, for a classical example $\boldsymbol{x} \in \mathbb{R}^d$, the employed encoding unitary

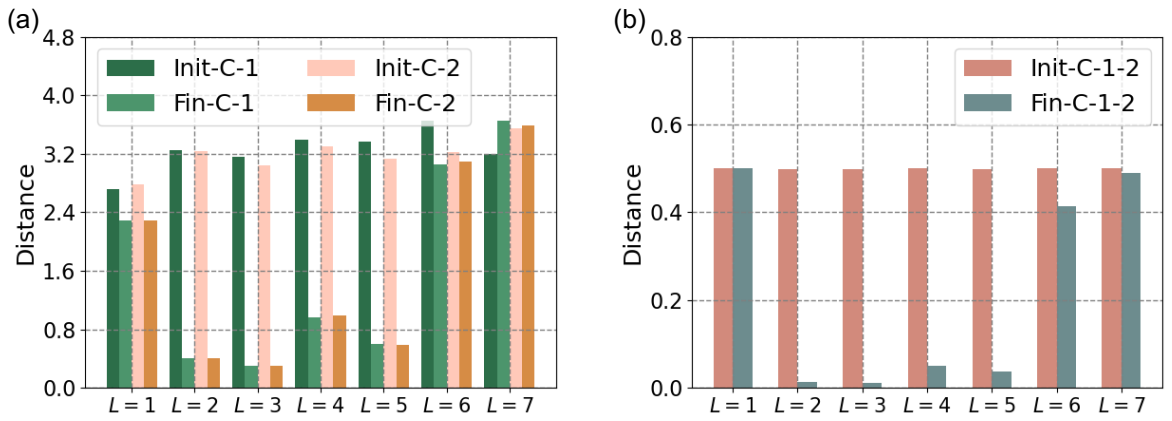

FIG. G.5. **Geometric properties of the quantum feature states on parity dataset.** (a) The averaged performance of QC evaluated by $\mathcal{M}_1$ defined in Eq. (G1). The label 'Init-C-k' with $k = 1, 2$ refers that the value of $\mathcal{M}_1^{(k)}$ at the initialization. Similarly, the label 'Final-C-k' with $k = 1, 2$ refers that the value of $\mathcal{M}_1^{(k)}$ when the training of QC is completed. (b) The averaged performance of QC evaluated by $\mathcal{M}_2$ defined in Eq. (G2). The label 'Init-C-1-2' ('Final-C-1-2') refers that the value of $\mathcal{M}_2$ before and after training of QC. The label '$L = a$' in the x-axis stands for that the layer number of hardware-efficient Ansatz is a.

is $U_E(\boldsymbol{x})\left|0\right\rangle^{\otimes d} = \left|\boldsymbol{x}\right\rangle$, which maps $\boldsymbol{x}$ to a 2^d dimensional quantum state $U_E(\boldsymbol{x})\left|0\right\rangle^{\otimes d}$. The latter uses the amplitude encoding method. Given a normalized image $\boldsymbol{x} \in \mathbb{R}^{64}$ with $\|\boldsymbol{x}\|_2^2 = 1$, the corresponding unitary encodes it into a 6-qubit state with $U_E(\boldsymbol{x})\left|0\right\rangle^{\otimes 6} = \sum_{j=1}^{64} \boldsymbol{x}_j \left|j\right\rangle$.

The Pauli-based measure operators are used in learning Fashion-MNIST dataset. Since the preprocessed dataset contains 9 classes, there are in total 9 measure operators, i.e., $o^{(1)} = X\otimes X\otimes \mathbb{I}^{\otimes 8}$, $o^{(2)} = X\otimes Y\otimes \mathbb{I}^{\otimes 8}$, $o^{(3)} = X\otimes Z\otimes \mathbb{I}^{\otimes 8}$, $o^{(4)} = Y\otimes X\otimes \mathbb{I}^{\otimes 8}$, $o^{(5)} = Y\otimes Y\otimes \mathbb{I}^{\otimes 8}$, $o^{(6)} = Y\otimes Z\otimes \mathbb{I}^{\otimes 8}$, $o^{(7)} = Z\otimes X\otimes \mathbb{I}^{\otimes 8}$, $o^{(8)} = Z\otimes Y\otimes \mathbb{I}^{\otimes 8}$, $o^{(9)} = Z\otimes Z\otimes \mathbb{I}^{\otimes 8}$.

Multilayer Perceptron. To better justify the capability and performance of QCs, we apply the multilayer perceptron (MLP) as the reference [109]. MLP is composed of an input layer, L hidden layers with $L \geq 1$, and an output layer. The dimension of the input layer is equivalent to the feature dimension of the input. ReLU activations are added in the hidden layer to perform nonlinear transformation. In the output layer, the activation function, Softmax, is employed. The number of layers L depends on the assigned tuples $\{n, N_t, T\}$.

Convolutional neural network. In the task of image classification, convolutional neural networks (CNNs) is employed as the reference [109]. The employed CNN is formed by two convolutional layers and one fully-connected layer. ReLU activations and the pooling operation are added in the hidden layer to perform nonlinear transformation. The number of channels for the first convolutional layer is fixed to be 8 and the corresponding kernel size is 9×9. The kernel size of the pooling operation applied to the two convolutional layers is 2×2. The kernel size for the second convolutional layer is fixed to be 5×5 but the number of output channels is varied depending on the settings in Alg. 1. For the sake of fair comparison, the number of output channels is set as $2, 6, 15, 30, 50, 75$, where the corresponding number of parameters is 860, 1284, 2238, 3828, 5948, and 8598, respectively.

Optimizer and other hyper-parameters. The adaptive gradient descent method, named AdaGrad optimizer [110], is used to optimize QCs and MLPs. Compared to the vanilla gradient descent method, AdaGrad permits better performance, since it adapts the learning rate for each feature depending on the estimated geometry of the problem. In the task of parity learning, the initial learning rate is set as $\eta = 0.5$ for QC and $\eta = 0.01$ for MLP, respectively. For both classifiers, the batch size is fixed to be 4. In the task of image classification, the initial learning rate is set as $\eta = 0.05$ for QC and $\eta = 0.01$ for CNN, respectively. The batch size for both classifiers is set as 1.

Curve fitting method. To capture the risk curve, Alg. 1 requests a curve fitting method. For all experiments, we adopt the polynomial fitting to derive the risk curve by using the collected results. The least squares method in determining the best fitting functions.

Source code. The source code used in numerical simulations will be available at Github repository `https://github.com/yuxuan-du/Problem-dependent-power-of-QNNs`.

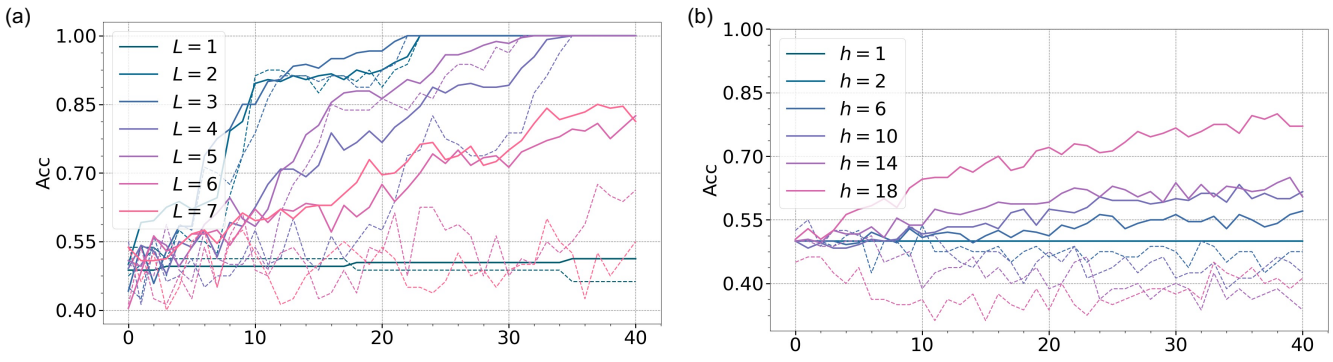

FIG. G.6. **Train (test) accuracy versus epoch on parity dataset.** (a) Train accuracy and test accuracy of QC with the varied layer number. The label '$L = a$' refers that the layer number used in hardware-efficient Ansatz is a. The solid line and the dashed line separately correspond to the train and test accuracies of QC. (b) Train accuracy and test accuracy of MLP with the varied number of hidden neurons. The label '$h = a$' refers that the number of neurons is a. The solid and dashed lines have the same meaning with those in QC.

1. Simulation results of the binary classification for the parity dataset

The feature states before and after training. We explore the geometric properties of feature states when the layer number of hardware-efficient Ansatz varies from $L = 1$ to $L = 7$. Other settings are identical to those introduced in the main text. Condition (i) in Lemma 2 is evaluated by the metric

$$\mathcal{M}_1^{(k)} = \sum_{i=1}^{n_c} \|\rho^{(i,k)} - \bar{\rho}^{(k)}\|, \tag{G1}$$

where the number of train examples $\{\rho^{(i,k)}\}_{i=1}^{n_c}$ belonging to the k-th class is n_c and $\bar{\rho}^{(k)}$ refers to their class-feature mean. Since parity learning is a binary classification task, Condition (ii) in Lemma 2 is evaluated by

$$\mathcal{M}_2 = \mathrm{Tr}(\bar{\rho}^{(0)}\bar{\rho}^{(1)}). \tag{G2}$$

The geometric properties of the feature states in the measure of $\mathcal{M}_1^{(k)}$ and $\mathcal{M}_2$ are visualized in Fig. G.5. The left panel shows that when $L \in \{2, 3, 4, 5\}$, both the value of $\mathcal{M}_1^{(1)}$ (highlighted by the green color) and $\mathcal{M}_1^{(2)}$ (highlighted by the pink color) decrease from ~ 3.2 (epoch $t = 0$) to ~ 0.5 (epoch $t = 40$). These results comply with Condition (i) in the sense that the feature states in the same class concentrates to the class-feature mean and leads to the low empirical risk. By contrast, when L is too small or too large, the value of $\mathcal{M}_1^{(1)}$ changes subtly before and after optimization, which is above 3.2. The large deviation of feature states incurs the degrade performance of QC. The right panel depicts that when $L \in \{2, 3, 4, 5\}$, the value of $\mathcal{M}_1^{(2)}$ decreases from 0.5 (epoch $t = 0$) to 0.05 (epoch $t = 40$). This reduction means that the class-feature means are maximally separated and thus ensure a good learning performance. On the contrary, when $L \in \{1, 6, 7\}$, the the value of $\mathcal{M}_1^{(2)}$ oscillates around 0.5, which implies that the class-feature means $\bar{\rho}^{(1)}$ and $\bar{\rho}^{(2)}$ are highly overlapped.

The learning dynamics of QC and MLP. Fig. G.6 visualizes the learning dynamics of QC and MLP with respect to the varied trainable parameters. The left panel indicates that when the layer number is $L = 2, 3, 4$, both train and test accuracies of QC fast converge to 100% with 25 epochs. When $L = 1$, both train and test accuracies oscillate to 50%. When $L = 7$, the number of train data becomes insufficient and the overfitting phenomenon appears. These results accord with the U-shape risk curve of QCs. The right panel shows that when the number of hidden neurons ranges from $h = 1$ to $h = 18$, the test accuracy of MLP is no higher that 55%. These results reflect the incapability of MLP in learning parity dataset compared with QCs.

2. Simulation results of multi-class classification for the Fashion-MNIST images dataset

The feature states before and after training. Here we discuss the geometric properties of feature states when the layer number of hardware-efficient Ansatz varies from $L = 2$ to $L = 150$. The metrics $\mathcal{M}_1^{(k)}$ and $\mathcal{M}_2$ defined in

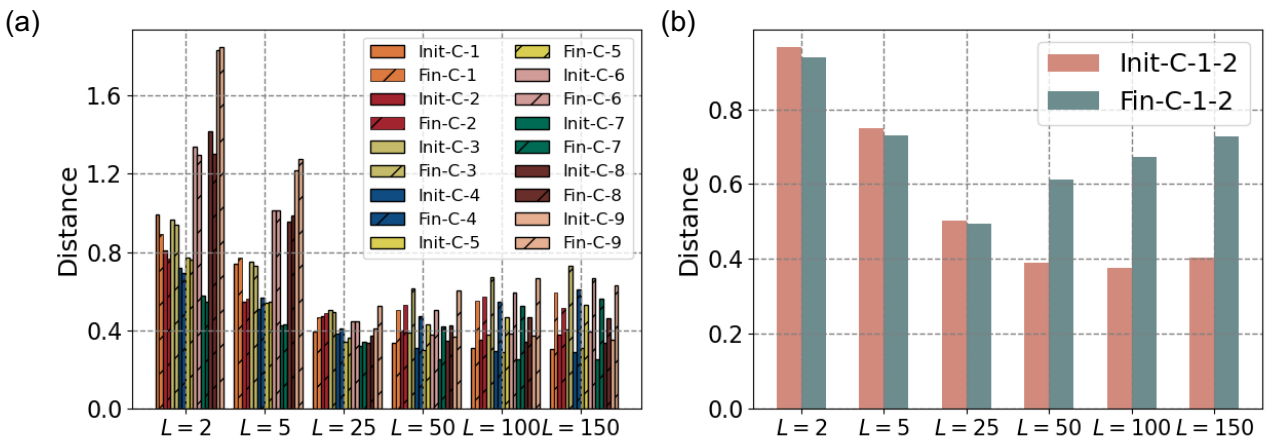

FIG. G.7. **Geometric properties of the quantum feature states on Fashion-MNIST dataset.** (a) The averaged performance of QC evaluated by $\mathcal{M}_1$ defined in Eq. (G1). (b) The averaged performance of QC evaluated by $\mathcal{M}_2$ defined in Eq. (G2). All labels have the same meaning with those introduced in Fig. G.5.

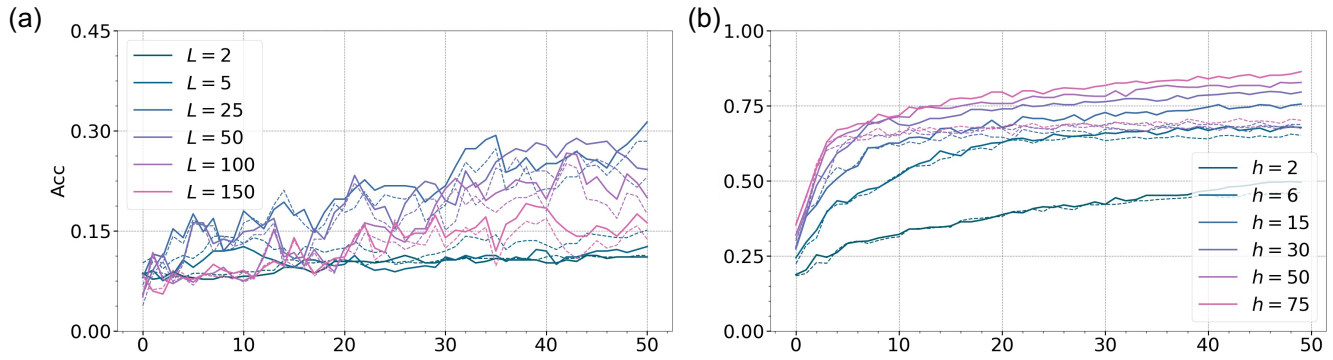

FIG. G.8. **Train (test) accuracy versus epoch on Fashion-MNIST dataset.** (a) Train accuracy and test accuracy of QC with the varied layer number. The labels have the same meaning with those presented in Fig. G.6. (b) Train accuracy and test accuracy of CNN with the varied number of trainable parameters. The label '$h = a$' refers that the number of output channels at the second layer is a. The solid and dashed lines have the same meaning with those in QC.

Eqs. (G1) and (G2) are employed. In the measure of $\mathcal{M}_2$, since the performance of QC for any two classes is similar, we only study the first two classes for ease of visualization.

Fig. G.7 depicts the geometric properties of the feature states in the measure of $\mathcal{M}_1^{(k)}$ and $\mathcal{M}_2$. The left panel shows that for all settings with $L \in \{2, 5, 25, 50, 100, 150\}$, the value $\mathcal{M}_1^{(k)}$ at the initial step and the final step is very similar and $\mathcal{M}_1^{(k)}$ is larger than 0.2 for $\forall k \in \{1, 2, ..., 9\}$. These results indicate that QC cannot satisfy Condition (i) when learning Fashion-MNIST dataset, where the feature states from the same class cannot collapse to a unique point. Moreover, when we examine the performance of intra-class, the right panel implies that after training, the class-feature means of QC are still highly overlapping. The distance for all settings of L is above 0.3. The inability to achieve the optimal training loss shows the the limited power of QC on learning Fashion-MNIST dataset.

The learning dynamics of QC and CNN. Fig. G.8 depicts the learning dynamics of QC and CNN with the varied number of trainable parameters. The left panel indicates that QC achieves the best performance when the layer number is $L \in [25, 100]$, where the corresponding number of parameters ranges from 750 to 3000. In these settings, both train and test accuracies of QC are around 30% after 50 epochs. When $L < 25$ or $L > 100$, both train and test accuracies oscillate at 15%. These results accord with the U-shape risk curve of QCs. The right panel shows that the train and test accuracies of CNN are steadily growing with the increased number of channels. That is, when the number of channels at the second layer is not less than 6, both the train and test accuracies are higher than 60%. These results indicate that the employed QC does not have potential advantages in learning image dataset compared with CNN.