Concentration of Data Encoding in Parameterized Quantum Circuits

Guangxi Li[1,2], **Ruilin Ye**[2,3], **Xuanqiang Zhao**[2], **Xin Wang**[2*]
[1] University of Technology Sydney, NSW, Australia
[2] Institute for Quantum Computing, Baidu Research, Beijing, China
[3] Peking University, Beijing, China

Abstract

Variational quantum algorithms have been acknowledged as the leading strategy to realize near-term quantum advantages in meaningful tasks, including machine learning and optimization. When applied to tasks involving classical data, such algorithms generally begin with data encoding circuits and train quantum neural networks (QNNs) to minimize target functions. Although QNNs have been widely studied to improve these algorithms' performance on practical tasks, there is a gap in systematically understanding the influence of data encoding on the eventual performance. In this paper, we make progress in filling this gap by considering the common data encoding strategies based on parameterized quantum circuits. We prove that, under reasonable assumptions, the distance between the average encoded state and the maximally mixed state could be explicitly upper-bounded with respect to the width and depth of the encoding circuit. This result in particular implies that the average encoded state will concentrate on the maximally mixed state at an exponential speed on depth. Such concentration seriously limits the capabilities of quantum classifiers, and strictly restricts the distinguishability of encoded states from a quantum information perspective. To support our findings, we numerically verify these results on both synthetic and public data sets. Our results highlight the significance of quantum data encoding and may shed light on the future design of quantum encoding strategies.

1 Introduction

Quantum machine learning [1, 2, 3, 4] is an emerging and promising interdisciplinary research direction in the fields of quantum computing and artificial intelligence. In this area, quantum computers are expected to enhance machine learning algorithms through their inherent parallel characteristics, thus demonstrating quantum advantages over classical algorithms [5].

With the increasing enormous efforts from academia and industry, the current quantum devices (usually acknowledged as the *noisy intermediate-scale quantum* (NISQ) devices [6]) already can show quantum advantages on certain carefully designed tasks [7, 8] despite their limitations in quantum circuit width and depth. A timely direction is to explore quantum advantages with near-term quantum devices in practical machine learning tasks, and a leading strategy is the hybrid quantum-classical algorithms, also known as the variational quantum algorithms [9, 10]. Such algorithms usually use a classical optimizer to train *quantum neural networks* (QNNs). They in particular hand over relatively intractable tasks to quantum computers while those rather ordinary tasks to classical computers.

*wangxin73@baidu.com

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

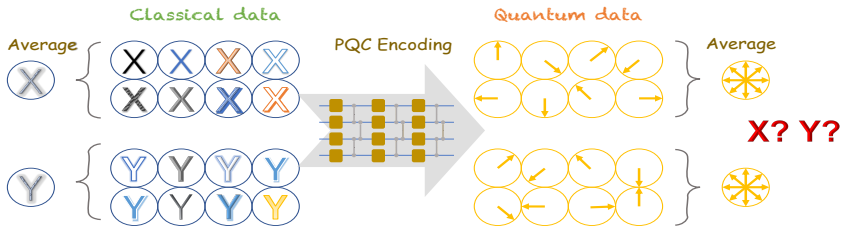


Figure 1: Cartoon illustrating the concentration of PQC-based data encoding. The average encoded quantum states concentrate on the maximally mixed state at an exponential rate on the encoding depth. This concentration implies the theoretical indistinguishability of the encoded quantum data.

For usual quantum machine learning tasks, a quantum circuit used in the variational quantum algorithms is generally divided into two parts: a data encoding circuit and a QNN. On the one hand, developing various QNN architectures is the most popular way to improve these algorithms' ability to deal with practical tasks. Numerous architectures such as strongly entangling circuit architectures [11], quantum convolutional neural networks [12], tree-tensor networks [13], and even automatically searched architectures [14, 15, 16, 17] have been proposed. On the other hand, one has to carefully design the encoding circuit, which could significantly influence the generalization performance of these algorithms [18, 19]. Consider an extreme case. If all classical inputs are encoded into the same quantum state, these algorithms will fail to do any machine learning tasks. In addition, the kernel's perspective [20, 21, 22, 23] also suggests that data encoding strategy plays a vital or even leading role in quantum machine learning algorithms [24, 25, 26]. However, there is much less literature on data encoding strategies, which urgently requires to be studied.

Encoding classical information into quantum data is nontrivial [1], and it is even more difficult on near-term quantum devices. One of the most feasible and popular encoding strategies on NISQ devices is based on *parameterized quantum circuits* (PQCs) [27], such as the empirically designed angle encoding [24, 28], IQP encoding [29], etc. It is natural to ask how to choose these encoding strategies and whether there are theoretical guarantees of using them. More specifically, it is necessary to systematically understand the impact of such PQC-based encoding strategies on the performance of QNNs in practical quantum machine learning tasks.

In this work, we present results from the perspective of average encoded quantum state with respect to the width and depth of the encoding PQCs. A cartoon illustration summarizing our main result is depicted in Fig. 1. We show that for the usual PQC-based encoding strategies with a fixed width, the average encoded state is close to the maximally mixed state at an exponential speed on depth. The contributions of this paper are multi-fold:

- We theoretically give the upper bound of the quantum divergence between the average encoded state and the maximally mixed state, which depends explicitly on the hyper-parameters (e.g., qubit number and encoding depth) of PQCs. From this bound, we find that for a fixed qubit number, the average encoded state concentrates on the maximally mixed state exponentially on the encoding depth.
- We show that the quantum states encoded by deep PQCs will seriously limit the trainability of a quantum classifier and further limit its classification ability.
- We show that the quantum states encoded by deep PQCs are indistinguishable from a quantum information perspective.
- We support the above findings by numerical experiments on both synthetic and public data sets.

Related Work Ref. [18] derived generalization bounds of PQC-based data encoding, which mainly depends on the total number of circuit gates. While we derive quantum divergence bound that depends on the width and depth of PQCs. The works [26, 30] explored the effects of data encoding from the perspective of data re-uploading. [31] studied the robustness of data encoding for quantum classifiers. Data encoding strategies with discrete features were proposed for variational quantum classifiers [32].

1.1 Background

Quantum Basics To better understand this paper, some basic concepts about quantum computing [33] are provided here. In general, quantum information is described by quantum states. An n-qubit quantum state is mathematically represented by a positive semi-definite matrix (a.k.a. a density matrix) $\rho \in \mathbb{C}^{2^n \times 2^n}$ with property $\mathrm{Tr}(\rho) = 1$. If $\mathrm{Rank}(\rho) = 1$, it is called a pure state; otherwise, it is a mixed state. A pure state can also be represented by a unit column vector $|\phi\rangle \in \mathbb{C}^{2^n}$, where $\rho = |\phi\rangle\langle\phi|$ and $\langle\phi| = |\phi\rangle^\dagger$. A mixed state can be regarded as a weighted sum of pure states, i.e., $\rho = \sum_i q_i |\phi_i\rangle\langle\phi_i|$, where $q_i \geq 0$ and $\sum_i q_i = 1$. Specifically, a mixed state whose density matrix is proportional to the identity matrix is called the maximally mixed state $\mathbb{1} \equiv \frac{I}{2^n}$.

A quantum state ρ could be evolved to another state ρ' through a quantum circuit (or gate) mathematically represented by a unitary matrix U, i.e., $\rho' = U\rho U^\dagger$. Typical single-qubit gates include Pauli gates, $X \equiv \left[\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right], Y \equiv \left[\begin{smallmatrix} 0 & -i \\ i & 0 \end{smallmatrix}\right], Z \equiv \left[\begin{smallmatrix} 1 & 0 \\ 0 & -1 \end{smallmatrix}\right]$, and their corresponding rotation gates $R_P(\theta) \equiv \mathrm{e}^{-i\theta P/2}$ with a parameter θ and $P \in \{X, Y, Z\}$. Another commonly used gate $U3$ appeared in this paper is defined as $U3(\theta_1, \theta_2, \theta_3) \equiv R_z(\theta_3)R_y(\theta_2)R_z(\theta_1)$, which can implement an arbitrary single-qubit unitary transformation with appropriate parameters. In this paper, R_z, R_y are equivalent to R_Z, R_Y without specified. A multi-qubit gate can be either an individual gate (e.g., CNOT) or a tensor product of single-qubit gates. To get classical information from quantum state ρ', one needs to perform quantum measurements, e.g., calculating the expectation value $\langle H\rangle = \mathrm{Tr}(H\rho')$ of a Hermitian matrix H, and we often call H an observable.

Quantum Divergence Similar to Kullback-Leibler divergence in machine learning, we use quantum divergence to quantify the difference between quantum states or quantum data. Two widely-used quantum divergences are quantum sandwiched Rényi divergence [34, 35] $\widetilde{D}_\alpha(\rho\|\sigma) \equiv \frac{1}{\alpha-1}\log \mathrm{Tr}\left[\sigma^{\frac{1-\alpha}{2\alpha}}\rho\sigma^{\frac{1-\alpha}{2\alpha}}\right]^\alpha$ and the Petz-Rényi divergence [36] $D_\alpha(\rho\|\sigma) \equiv \frac{1}{\alpha-1}\log \mathrm{Tr}\left[\rho^\alpha\sigma^{1-\alpha}\right]$, where $\alpha \in (0,1) \cup (1,\infty)$ and the latter has an operational significance in quantum hypothesis testing [37, 38, 39]. In this work, for the purposes of analyzing quantum encoding, we focus on the Petz-Rényi divergence with $\alpha = 2$, i.e.,

$$D_2(\rho\|\sigma) = \log \mathrm{Tr}\left[\rho^2\sigma^{-1}\right], \tag{1}$$

which also plays an important role in training quantum neural networks [40] as well as quantum communication [41]. Throughout this paper, when we mention the quantum divergence, we mean the Petz-Rényi divergence D_2 if not specified; $\log$ denotes $\log_2$ if not specified.

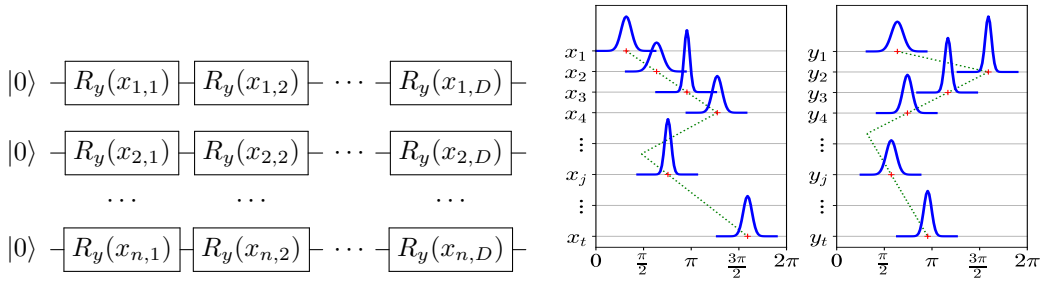
Parameterized Quantum Circuit In general, a parameterized quantum circuit [27] has the form $U(\boldsymbol{\theta}) = \prod_j U_j(\theta_j)V_j$, where $\boldsymbol{\theta}$ is its parameter vector, $U_j(\theta_j) = \mathrm{e}^{-i\theta_j P_j/2}$ with P_j denoting a Pauli gate, and V_j denotes a fixed gate such as Identity, CNOT and so on. In this paper, PQCs are utilized as both data encoding strategies and quantum neural networks. Specifically, when used for data encoding, an n-qubit PQC takes a classical input vector $\boldsymbol{x}$ as its parameters and acts on an initial state $|0\rangle^{\otimes n}$ to obtain the encoded state $|\boldsymbol{x}\rangle$. Here, $|0\rangle^{\otimes n}$ is a 2^n-dimensional vector whose first element is 1 and all other elements are 0.

2 Main Results

2.1 A Warm-up Case

For a quick access, we first consider one of the most straightforward PQC-based data encoding strategies, i.e., consisting of R_y rotations only, cf. Fig. 2(a). It can be viewed as a generalized angle encoding. For a classical input vector $\boldsymbol{x}$ with nD components, the output of this data encoding circuit is a pure state $|\boldsymbol{x}\rangle \in \mathbb{C}^{2^n}$ expanded in a 2^n-dimensional Hilbert space. We denote the density matrix of the output state by $\rho(\boldsymbol{x}) = |\boldsymbol{x}\rangle\langle\boldsymbol{x}|$. If we assume each element of the input vector obeys an *independent Gaussian distribution* (IGD, see Fig. 2(b) for an intuitive illustration), then we have the following theorem.

Theorem 1. *Assume each element of an nD-dimensional vector $\boldsymbol{x}$ obeys an IGD, i.e., $x_{j,d} \sim \mathcal{N}(\mu_{j,d}, \sigma_{j,d}^2)$, where $\sigma_{j,d} \geq \sigma$ for some constant σ and $1 \leq j \leq n, 1 \leq d \leq D$. If $\boldsymbol{x}$ is encoded into*



(a) Circuit with R_y rotations only

(b) Assumption of IGD

Figure 2: (a) Circuit for the data encoding strategy with R_y rotations only; (b) An example of a binary data set with two classes of t-dimensional vectors $\boldsymbol{x}$ and $\boldsymbol{y}$. Here, it is assumed that each x_j (or y_j) obeys an independent Gaussian distribution (IGD), i.e., $x_j \sim \mathcal{N}(\mu_{x,j}, \sigma_{x,j}^2)$ (or $y_j \sim \mathcal{N}(\mu_{y,j}, \sigma_{y,j}^2)$), where these mean values (small red cross symbols) range in $[0, 2\pi)$ and form the green dotted lines. Note that the difference between these two lines determines that they belong to different classes.

an n-qubit pure state $\rho(\boldsymbol{x})$ according to the circuit in Fig. 2(a), then the quantum divergence between the average encoded state $\bar{\rho}$ and the maximally mixed state $\mathbb{1}$ is upper-bounded as

$$D_2\left(\bar{\rho} \| \mathbb{1}\right) \leq n \log\left(1 + \mathrm{e}^{-D\sigma^2}\right), \tag{2}$$

where $\bar{\rho}$ is defined as $\bar{\rho} \equiv \mathbb{E}\left[\rho(\boldsymbol{x})\right]$.

This theorem shows that the upper bound of the quantum divergence between $\bar{\rho}$ and $\mathbb{1}$ explicitly depends on the qubit number n and the encoding depth D under certain conditions. Further by approximating Eq. (2) as

$$D_2\left(\bar{\rho} \| \mathbb{1}\right) \leq n \log(1 + \mathrm{e}^{-D\sigma^2}) \approx \begin{cases} n\left(1 - \frac{\sigma^2}{2\ln 2} D\right), & D \in O(1) \\ n\mathrm{e}^{-D\sigma^2}, & D \in \Omega(\mathrm{poly}\log(n)) \end{cases}, \tag{3}$$

we easily find that for a fixed n, the upper bound decays exponentially with D growing in $\Omega(\mathrm{poly}\log(n))$. This means that the average encoded state will quickly approach the maximally mixed state with an arbitrarily small distance under reasonable depths.

Sketch of Proof. For the j-th qubit, let $x_{j,\mathrm{sum}} = \sum_d x_{j,d}$ and $\mu_{j,\mathrm{sum}} = \sum_d \mu_{j,d}$, we get $x_{j,\mathrm{sum}} \sim \mathcal{N}(\mu_{j,\mathrm{sum}}, \sum_d \sigma_{j,d}^2)$. From the fact that $R_y(x_1)R_y(x_2) = R_y(x_1 + x_2)$, we have $\rho\left(\boldsymbol{x}_j\right) = R_y\left(x_{j,\mathrm{sum}}\right)|0\rangle\langle 0| R_y^\dagger\left(x_{j,\mathrm{sum}}\right) = \frac{1}{2}\left[I + \cos\left(x_{j,\mathrm{sum}}\right) Z + \sin\left(x_{j,\mathrm{sum}}\right) X\right]$. Further from the fact that if a variable $x \sim \mathcal{N}(\mu, \sigma^2)$, then $\mathbb{E}\left[\cos(x)\right] = \mathrm{e}^{-\frac{\sigma^2}{2}}\cos(\mu)$ and $\mathbb{E}\left[\sin(x)\right] = \mathrm{e}^{-\frac{\sigma^2}{2}}\sin(\mu)$, together with the condition $\sigma_{j,d} \geq \sigma$, we bound $D_2\left(\mathbb{E}\left[\rho\left(\boldsymbol{x}_j\right)\right] \| \mathbb{1}\right)$ via calculating

$$\mathrm{Tr}\left(\mathbb{E}^2\left[\rho\left(\boldsymbol{x}_j\right)\right]\right) = \frac{1}{2}\left(1 + \mathrm{e}^{-\sum_d \sigma_{j,d}^2}\right) \leq \frac{1}{2}\left(1 + \mathrm{e}^{-D\sigma^2}\right). \tag{4}$$

Finally we generalize Eq. (4) to multi-qubit cases, and complete the proof of Theorem 1. The detailed proof is deferred to Appendix A.

2.2 General Case

Next, we consider the general PQC-based data encoding strategies shown in Fig. 3, where a column of $U3$ gates and a column of entangled gates spread out alternately.

Theorem 2. ***(Data Encoding Concentration)*** *Assume each element of a $3nD$-dimensional vector $\boldsymbol{x}$ obeys an IGD, i.e., $x_{j,d,k} \sim \mathcal{N}(\mu_{j,d,k}, \sigma_{j,d,k}^2)$, where $\sigma_{j,d,k} \geq \sigma$ for some constant σ and $1 \leq j \leq n, 1 \leq d \leq D, 1 \leq k \leq 3$. If $\boldsymbol{x}$ is encoded into an n-qubit pure state $\rho(\boldsymbol{x})$ according to the circuit in Fig. 3, the quantum divergence between the average encoded state $\bar{\rho}$ and the maximally mixed state $\mathbb{1}$ is upper-bounded as*

$$D_2\left(\bar{\rho} \| \mathbb{1}\right) \leq \log(1 + (2^n - 1)\mathrm{e}^{-D\sigma^2}). \tag{5}$$

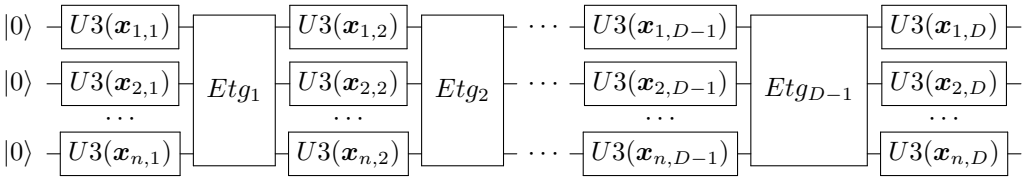


Figure 3: Circuit for the data encoding strategy with D layers of $U3$ gates and $D-1$ layers of entanglements. Here, each $\boldsymbol{x}_{j,d}$ represents three elements $x_{j,d,1}, x_{j,d,2}, x_{j,d,3}$, and each Etg_i denotes an arbitrary group of entangled two-qubit gates, such as CNOT or CZ, where $1 \leq j \leq n, 1 \leq d \leq D, 1 \leq i \leq D-1$.

This theorem shows that, when employing general PQC-based encoding strategies, we could also have an upper bound of the quantum divergence $D_2(\bar{\rho}\|\mathbb{1})$ which explicitly depends on n and D. By approximating the upper bound in Eq. (5) as follows

$$D_2(\bar{\rho}\|\mathbb{1}) \leq \log\left(1 + (2^n - 1)\mathrm{e}^{-D\sigma^2}\right) \approx \begin{cases} n - \frac{\sigma^2}{\ln 2}D, & D \in O(\text{poly}\log(n)) \\ (2^n - 1)\mathrm{e}^{-D\sigma^2}, & D \in \Omega(\text{poly}(n)) \end{cases}, \tag{6}$$

we observe similarly that for some fixed n, the upper bound decays at an exponential speed as D grows in $\Omega(\text{poly}(n))$. In addition, according to our proof analysis, even if each $U3$ gate is replaced by a $U2$ gate containing only two different kinds of Pauli rotations or even a $U1$ gate with only one proper Pauli rotation, we still get a similar bound as Eq. (5). Therefore, we conclude that as long as D grows within a reasonable scope, the average of the quantum states encoded by a wide family of PQCs will quickly concentrates on the maximally mixed state. Unlike the warm-up case, the proof for this theorem is quite non-straightforward due to the tricky entangled gates. For smooth reading, we defer the complete proof in Appendix B.

The average encoded state $\bar{\rho}$ in Theorem 2 is built on infinite data samples, but in practice we do not have infinite ones. Therefore, we provide the following helpful corollary.

Corollary 2.1. *Assume there are M classical vectors $\{\boldsymbol{x}^{(m)}\}_{m=1}^{M}$ sampled from the distributions described in Theorem 2 and define $\bar{\rho}_M \equiv \frac{1}{M}\sum_{m=1}^{M}\rho(\boldsymbol{x}^{(m)})$. Let H be a Hermitian matrix with its eigenvalues ranging in $[-1, 1]$, then given an arbitrary $\epsilon \in (0, 1)$, as long as the encoding depth $D \geq \frac{1}{\sigma^2}\left[(n+4)\ln 2 + 2\ln(1/\epsilon)\right]$, we have*

$$\left|\operatorname{Tr}\left[H\left(\bar{\rho}_M - \mathbb{1}\right)\right]\right| \leq \epsilon \tag{7}$$

with a probability of at least $1 - 2\mathrm{e}^{-M\epsilon^2/8}$.

This corollary implies that for a reasonable encoding depth D and number of samples M, the practical average encoded state $\bar{\rho}_M$ will also be infinitely close to the maximally mixed state with a high probability. The proof is mainly derived from *Hoeffding's inequality* [42] and the relations between quantum divergence and trace norm. The proof details are deferred to Appendix C.

3 Applications in Quantum Supervised Learning

In this section, we will show that the concentrated quantum states encoded by the above PQC-based data encoding strategies will severely limit the performances of quantum supervised learning tasks. Before beginning, let's define the following necessary data set.

Definition 3. ***(Data Set)*** *The K-class data set $\mathcal{D} \equiv \{(\boldsymbol{x}^{(m)}, \boldsymbol{y}^{(m)})\}_{m=1}^{KM} \subset \mathbb{R}^{3nD} \times \mathbb{R}^K$ totally has KM data samples, including M samples in each category. Here, suppose elements in the same entry of all input vectors from the same category are sampled from the same IGD with a variance of at least σ^2 and each $\boldsymbol{x}^{(m)}$ is encoded into the corresponding pure state $\rho(\boldsymbol{x}^{(m)})$ according to the circuit in Fig. 3 with n qubits and D layers of $U3$ gates. The label $\boldsymbol{y}^{(m)}$ is a one-hot vector that indicates which of the K classes $\boldsymbol{x}^{(m)}$ belongs to.*

3.1 In the Scenario of Quantum Classification

Quantum classification, as one of the most significant branches in quantum machine learning, is widely studied nowadays, where quantum-enhanced classification models are expected to achieve quantum advantages against the classical ones in solving classification tasks [29, 43]. Specifically, in the NISQ era, plenty of variational quantum classifiers based on parameterized quantum circuits are developed to make full use of NISQ devices [9, 44, 11, 12, 13, 45, 46]. See [47] for a review on the recent advances of quantum classifiers.

In general, a quantum classifier aims to learn a map from input to label by optimizing a loss function constructed through QNNs to predict the label of an unseen input as accurately as possible. Now, we demonstrate the performance of a quantum classifier on the data set $\mathcal{D}$ defined in Def. 3.

In this paper, the loss function is defined from the cross-entropy loss with softmax function [48]:

$$L\left(\boldsymbol{\theta};\mathcal{D}\right) \equiv \frac{1}{KM}\sum_{m=1}^{KM} L^{(m)} \quad \text{with} \quad L^{(m)}\left(\boldsymbol{\theta};(\boldsymbol{x}^{(m)},\boldsymbol{y}^{(m)})\right) \equiv -\sum_{k=1}^{K} y_k^{(m)} \ln \frac{\mathrm{e}^{h_k}}{\sum_{j=1}^{K}\mathrm{e}^{h_j}}, \tag{8}$$

where $y_k^{(m)}$ denotes the k-th element of the label $\boldsymbol{y}^{(m)}$ and

$$h_k\left(\boldsymbol{x}^{(m)},\boldsymbol{\theta}\right) = \mathrm{Tr}\left[H_k U(\boldsymbol{\theta})\rho(\boldsymbol{x}^{(m)})U^\dagger(\boldsymbol{\theta})\right], \tag{9}$$

which means the Hermitian operator H_k is finally measured after the quantum neural network $U(\boldsymbol{\theta})$. Here, each H_k is chosen from tensor products of various Pauli matrices, such as $Z\otimes I$, $X\otimes Y\otimes Z$ and so on. By minimizing the loss function with a gradient descent method, we could obtain the final trained model $U(\boldsymbol{\theta}^*)$ with the optimal or sub-optimal parameters $\boldsymbol{\theta}^*$. After that, when provided a new input quantum state $\rho(\boldsymbol{x}')$, we compute each h_k' with parameters $\boldsymbol{\theta}^*$ according to Eq. (9), and the index of the largest h_k' is exactly our designated label.

However, all these graceful expectations can only be established on gradients with relatively large absolute values. On the contrary, gradients with significantly small absolute values will cause a severe training problem, for example the barren plateau issue [49]. Therefore, in the following we investigate the partial gradient of the cost defined in Eq. (8) with regard to its parameters. The results are exhibited in Proposition 4.

Proposition 4. *Consider a K-classification task with the data set $\mathcal{D}$ defined in Def. 3. If the encoding depth $D \geq \frac{1}{\sigma^2}\left[(n+4)\ln 2 + 2\ln(1/\epsilon)\right]$ for some $\epsilon \in (0,1)$, then the partial gradient of the loss function defined in Eq. (8) with respect to each parameter θ_i of the employed QNN is bounded as*

$$\left|\frac{\partial L\left(\boldsymbol{\theta};\mathcal{D}\right)}{\partial \theta_i}\right| \leq K\epsilon \tag{10}$$

with a probability of at least $1-2\mathrm{e}^{-M\epsilon^2/8}$.

From this proposition, we observe that no matter what QNN structures are selected, the absolute gradient value can be arbitrarily small with a very high probability for the above data set $\mathcal{D}$, provided that the encoding depth D and the number of data samples M are sufficiently large. This vanishing of the gradients will severely restrict the trainability of QNNs. Moreover, if before training $U(\boldsymbol{\theta})$ is initialized to satisfy a certain randomness, such as unitary 2-design [50], then each h_k in Eq. (9) will concentrate on 0 with a high probability, thus the loss in Eq. (8) will concentrate on $\ln K$. This concentration of loss is also verified through numerical simulations, as presented in Sec. 4. This phenomenon, together with Proposition 4, implies that large encoding depth will significantly hinder the training of a quantum classifier and probably lead to poor classification accuracy. Please refer to Appendix D for the proof of Proposition 4.

3.2 In the Scenario of Quantum State Discrimination

Quantum state discrimination [51] is a central information-theoretic task and finds applications in various topics such as quantum cryptography [52], quantum error mitigation [53], and quantum data hiding [54]. It aims to distinguish quantum states using a positive operator-valued measure (POVM), a set of positive semi-definite operators that sum to the identity operator. Here, we have to seriously note that in quantum state discrimination, we can only measure each quantum state once, instead of measuring repeatedly and calculating the expectations as shown in quantum classification.

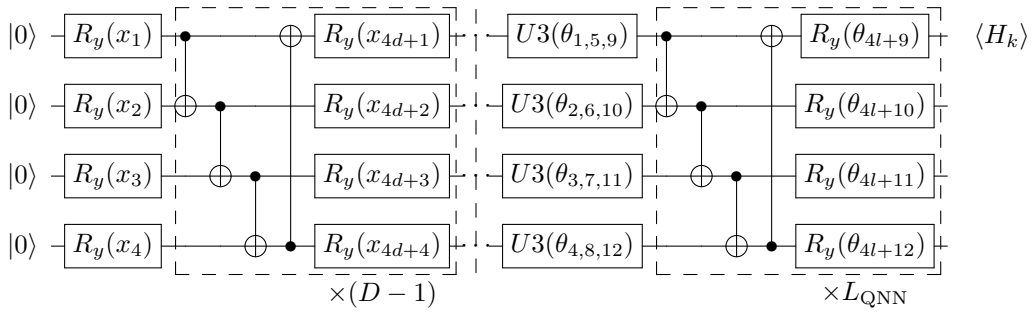


Figure 4: Circuits for data encoding (before the barrier line) and quantum neural network (after the barrier line) in the 4-qubit case. Here the input $\boldsymbol{x} \in \mathbb{R}^{4D}$ and $d \in [1, D-1]$. The QNN totally has $L_{\mathrm{QNN}} + 1$ layers with parameters $\boldsymbol{\theta} \in \mathbb{R}^{4L_{\mathrm{QNN}}+12}$, where the first layer $U3$ gates consists of 12 parameters. After QNN, there are K expectations $\{\langle H_k \rangle\}_{k=1}^{K}$ for K-classification tasks.

In general, a perfect discrimination (i.e., a perfect POVM) can not be achieved if quantum states are non-orthogonal. A natural alternative option is by adopting some metrics such as the success probability so that the optimal POVM could be obtained via various kinds of optimization ways, e.g., Helstrom bound [55] and semi-definite programming (SDP) [56]. Recently, researchers also try to train QNNs as substitutions for optimal POVMs [57, 58].

Next, we demonstrate the impact of the encoded quantum states from the data set $\mathcal{D}$ defined in Def. 3 on quantum state discrimination. Our goal is to obtain the maximum success probability p_{succ} by maximizing the success probability over all POVMs with K operators:

$$p_{\mathrm{succ}} \equiv \max_{\{\Pi_k\}_k} \frac{1}{K} \sum_{k=1}^{K} \mathrm{Tr}\left[\Pi_k \bar{\rho}_{k,M}\right] \quad \text{with} \quad \bar{\rho}_{k,M} \equiv \frac{1}{M} \sum_{m=1}^{KM} y_k^{(m)} \rho(\boldsymbol{x}^{(m)}), \tag{11}$$

where $y_k^{(m)}$ denotes the k-th element of the label $\boldsymbol{y}^{(m)}$ and $\{\Pi_k\}_{k=1}^{K}$ denotes a POVM, which satisfies $\sum_{k=1}^{K} \Pi_k = I$.

Proposition 5. *Consider a K-class discrimination task with the data set $\mathcal{D}$ defined in Def. 3. If the encoding depth $D \geq \frac{1}{\sigma^2}\left[(n+4)\ln 2 + 2\ln(1/\epsilon)\right]$ for a given $\epsilon \in (0,1)$, then with a probability of at least $1 - 2\mathrm{e}^{-M\epsilon^2/8}$, the maximum success probability p_{succ} is bounded as*

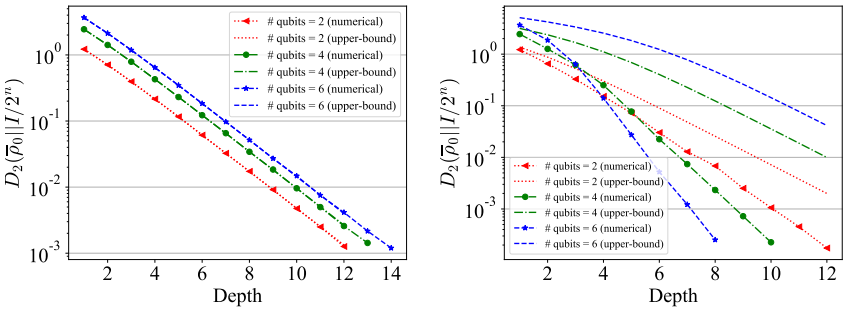
$$p_{\mathrm{succ}} \leq 1/K + \epsilon. \tag{12}$$

This proposition implies that as long as the encoding depth D and the data numbers M are large enough, the optimal success probability p_{succ} could be arbitrarily close to $\frac{1}{K}$ with a remarkably high probability for the data set $\mathcal{D}$. This nearly blind-guessing success probability shows that the concentration of the encoded quantum states in the data set $\mathcal{D}$ will lead to the failure of state discrimination via POVM. As POVMs are the most general kind of measurements one can implement to extract classical information from quantum systems [33], we conclude that the above different classes of encoded states are indistinguishable from the perspective of quantum information. The proof of Proposition 5 could be derived straightforwardly by combining Eq. (11) with Corollary 2.1.

4 Numerical Experiments

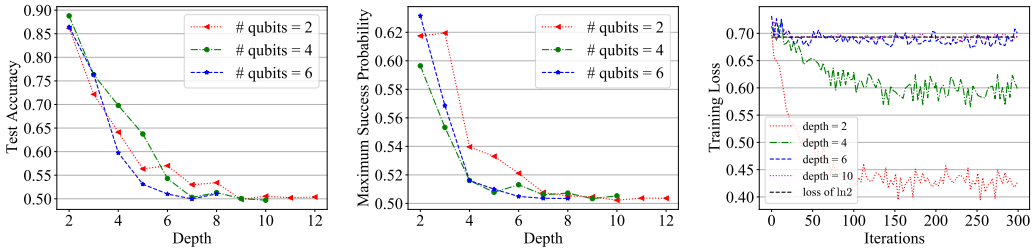
Previous sections demonstrate that the average encoded state will concentrate on the maximally mixed state under PQC-based data encoding strategies with large depth. These encoded states theoretically cannot be utilized to train QNNs or distinguished by POVMs. In this section, we verify these results on both synthetic and public data sets by choosing a commonly employed strongly entangling circuit [11], which helps to understand the concentration rate intuitively for realistic encoding circuits. All the simulations and optimization loop are implemented via Paddle Quantum[2] on the PaddlePaddle Deep Learning Platform [59].

[2]https://github.com/paddlepaddle/Quantum



(a) Encoding Strategy in Fig. 2(a) (b) Encoding Strategy in Fig. 4

Figure 5: Exponential decay of quantum divergence $D_2(\bar{\rho}_0||\mathbb{1})$ vs. encoding depth under different qubit cases for synthetic data set. Here, there are one million data samples for calculating average encoded state $\bar{\rho}_0$ for class 0 at each point in numerical lines. And the upper-bounds come from (a) Theorem 1, (b) Theorem 2, respectively.



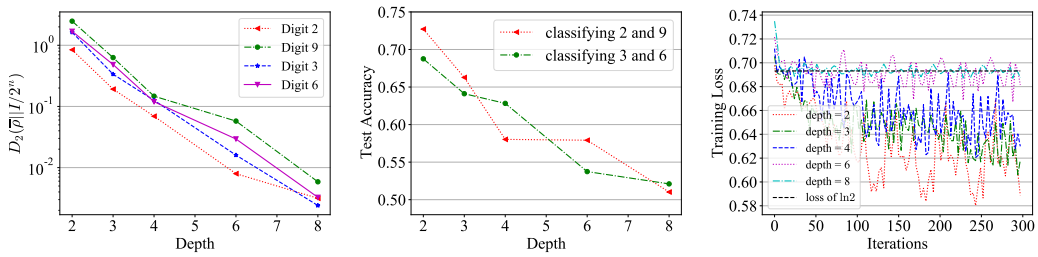
(a) (b) (c)

Figure 6: Numerical results for synthetic data sets under the encoding strategy in Fig. 4. In all qubit cases, (a) the test accuracy of QNN (or (b) the maximum success probability of POVM) will eventually decay to 50% or so with the depth growing; (c) In the 4 qubit case, for instance, the training losses of QNN do not decrease and stay at about $\ln 2$ in the training process when the depth becomes large enough.

4.1 On Synthetic Data Set

Data Set The synthetic two-class data set $\{(\boldsymbol{x}^{(m)}, \boldsymbol{y}^{(m)})\}_{m=1}^{M}$ is generated following the distributions depicted in Fig. 2(b), where each $x_j^{(m)} \sim \mathcal{N}(\mu_j, \sigma_j^2)$ for $1 \leq j \leq t$ and $\boldsymbol{y}^{(m)}$ denotes a one-hot label. Here, we assume all means come from two lines, i.e., $\mu_j = \frac{2\pi}{16}(j-1) \bmod 2\pi$ for class 0 and $\mu_j = \frac{2\pi}{16}(16-j) \bmod 2\pi$ for class 1, and all σ_j's are set as 0.8. Note that the same variance is selected for both classes to facilitate the demonstration of the experiment. Other choices of σ_j's would have similar effects.

We first verify our two main upper bounds given in Theorems 1 and 2 by encoding the nD-dimensional inputs that belong to the same class into n-qubit quantum states with D encoding depths under the encoding strategies illustrated in Figs. 2(a) and 4, respectively. Here, n is set as $2, 4, 6$ and $D \in [1, 14]$. The results are displayed in Fig. 5, from which we can intuitively see that the divergences decrease exponentially on depth. Specifically, from Fig. 5(a), we know the upper bound in Theorem 2(a) is tight, which is also easily verified from our proof. From Fig. 5(b), we learn that the upper bound in Theorem 2 is quite loose, which suggests that the real situation is much worse than our theoretical analysis. We also notice in Fig. 5(b) that for this strongly entangling encoding strategy, the larger the qubit number is, the faster the divergence decreases. This unexpected phenomenon reveals the possibility that specific structures of encoding circuits may lead to more severe concentrations for larger numbers of qubits and is worthy of further studies.



(a) (b) (c)

Figure 7: Numerical results of QNN for MNIST data set under the encoding strategy in Fig. 4. (a) The curves for the quantum divergence between the averaged encoded state $\bar{\rho}$ of each handwritten digit and the maximally mixed state $\mathbb{1}$ decrease exponentially on depth. (b) The test accuracy reduce rapidly with a larger encoding depth; (c) In the case of classifying digits 3 and 6, when the depth is large (e.g., 8), it is difficult to keep the training loss away from $\ln 2$ in the training process.

Next, we examine the performance of QNNs and POVMs by generating 20k data samples for training and 4k for testing under the encoding strategy in Fig. 4, where half of the data belong to class 0, and the others belong to class 1. The QNNs are designed according to the right hand side of Fig. 4, where the number of layers L_{QNN} is set as $n+2$, the finally measured Hermitian operators are set as $H_1 = Z$ and $H_2 = X$ on the first qubit, and all parameters $\boldsymbol{\theta}$ are initialized randomly in $[0, 2\pi]$. During the optimization, we adopt the Adam optimizer [60] with a batch size of 200 and a learning rate of 0.02. In the POVM setting, we directly employ semi-definite programming [56] to obtain the maximum success probability P_{succ} on the training data samples. The results are illustrated in Fig. 6. We observe that both the test accuracy of the QNNs and the maximum success probability P_{succ} of the POVMs eventually decay to about 0.5 as the encoding depth grows, indicating that the classification abilities of both the QNNs and the POVMs are no better than random guessing. In addition, the training losses of QNNs in Fig. 6(c) gradually approach $\ln 2$ as the depth grows and finally do not go down anymore during the whole training process, which implies that the concentration of this data set on the maximally mixed state would limit the trainability of QNNs as we predicted in Sec. 3.1. All these results are in line with our theoretical expectations.

4.2 On Public Data Set

Data Set and Preprocessing The handwritten digit data set MNIST [61] consists of 70k images labeled from '0' to '9', each of which contains 28×28 gray scale pixels valued in $[0, 255]$. In order to facilitate encoding, these images are first resized to 4×4 and then normalized to values between 0 and π. Finally, we select all images corresponding to two pairs of labels, i.e., $(2, 9)$ and $(3, 6)$, for two binary classification tasks. For each task, there are about 12k training samples and 2k testing samples, and each category accounts for half or so.

Here we mainly consider the performance of QNN on this data set because POVMs are generally not suitable for prediction. These 16-dimensional preprocessed images are first encoded into n-qubit quantum states with encoding depth D and then fed into a QNN (cf. Fig. 4 again). We set n as 2,3,4,6,8 and D as 8,6,4,3,2 accordingly. The settings of QNN are almost the same as those used in the synthetic case, except for a new learning rate of 0.05. From Fig. 7(a), we see that the average state of each digit class concentrates on the maximally mixed state at an approximately exponential speed on depth, which is consistent with our main result. Furthermore, the outcomes in Figs. 7(b) and 7(c) also confirm the incapability of training of QNNs, provided that the classical inputs are encoded by a higher depth PQC.

5 Discussion

We have witnessed both theoretically and numerically that for usual PQC-based data encoding strategies with higher depth, the average encoded state concentrates on the maximally mixed state. We further show that such concentration severely limits the capabilities of quantum classifiers for

practical tasks. Such limitation indicates that we should pay more attention to methods encoding classical data into PQCs in quantum supervised learning.

Our work suggests that the distance between the average encoded state and the maximally mixed state may be a reasonable metric to quantify how well the quantum encoding preserves the features in quantum supervised learning. The result on the encoding concentration also motivates us to consider how to design PQC-based encoding strategies better to avoid the exponentially decayed distance. An obvious way this paper implies might be to keep the depth shallow while accompanied by a higher width. Still, it will render poor generalization performance [19] as well as the notorious barren plateau issue [49]. Therefore, it will be desirable to develop nontrivial quantum encoding strategies to guarantee the effectiveness and efficiency of quantum supervised learning as well as quantum kernel methods [24, 20, 28]. Amplitude encoding [24] may serve as a good candidate, but due to its unavailability on NISQ devices, it is rarely considered or implemented at current stage. Recent works on data re-uploading [26, 30, 62] and pooling [19, 25] of quantum neural networks may also provide potential methods for improving quantum encoding efficiency.

Acknowledgements

We would like to thank Sanjiang Li, Yuan Feng, Hongshun Yao, and Zhan Yu for helpful discussions. G. L. acknowledges the support from the Baidu-UTS AI Meets Quantum project, the China Scholarship Council (No. 201806070139), and the Australian Research Council project (Grant No: DP180100691). Part of this work was done when R. Y. and X. Z. were research interns at Baidu Research.

References

[1] Jacob Biamonte, Peter Wittek, Nicola Pancotti, Patrick Rebentrost, Nathan Wiebe, and Seth Lloyd. Quantum machine learning. *Nature*, 549(7671):195–202, sep 2017.

[2] Maria Schuld, Ilya Sinayskiy, and Francesco Petruccione. An introduction to quantum machine learning. *Contemporary Physics*, 56(2):172–185, apr 2015.

[3] Carlo Ciliberto, Mark Herbster, Alessandro Davide Ialongo, Massimiliano Pontil, Andrea Rocchetto, Simone Severini, and Leonard Wossnig. Quantum machine learning: a classical perspective. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 474(2209):20170551, jan 2018.

[4] Seth Lloyd, Masoud Mohseni, and Patrick Rebentrost. Quantum algorithms for supervised and unsupervised machine learning. *arXiv:1307.0411*, jul 2013.

[5] Aram W Harrow and Ashley Montanaro. Quantum computational supremacy. *Nature*, 549(7671):203–209, 2017.

[6] John Preskill. Quantum Computing in the NISQ era and beyond. *Quantum*, 2:79, aug 2018.

[7] 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.

[8] 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.

[9] Kishor Bharti, Alba Cervera-Lierta, Thi Ha Kyaw, Tobias Haug, Sumner Alperin-Lea, Abhinav Anand, Matthias Degroote, Hermanni Heimonen, Jakob S Kottmann, Tim Menke, et al. Noisy intermediate-scale quantum algorithms. *Reviews of Modern Physics*, 94(1):015004, 2022.

[10] 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.

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

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

[13] 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):65, dec 2018.

[14] Mateusz Ostaszewski, Lea M Trenkwalder, Wojciech Masarczyk, Eleanor Scerri, and Vedran Dunjko. Reinforcement learning for optimization of variational quantum circuit architectures. *Advances in Neural Information Processing Systems*, 34:18182–18194, 2021.

[15] Mateusz Ostaszewski, Edward Grant, and Marcello Benedetti. Structure optimization for parameterized quantum circuits. *Quantum*, 5:391, 2021.

[16] Shi-Xin Zhang, Chang-Yu Hsieh, Shengyu Zhang, and Hong Yao. Differentiable Quantum Architecture Search. *arXiv:2010.08561*, oct 2020.

[17] Yuxuan Du, Tao Huang, Shan You, Min-Hsiu Hsieh, and Dacheng Tao. Quantum circuit architecture search: error mitigation and trainability enhancement for variational quantum solvers. *arXiv preprint arXiv:2010.10217*, 2020.

[18] 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.

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

[20] 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):2631, dec 2021.

[21] 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.

[22] Tongyang Li, Shouvanik Chakrabarti, and Xiaodi Wu. Sublinear quantum algorithms for training linear and kernel-based classifiers. In *International Conference on Machine Learning*, pages 3815–3824. PMLR, 2019.

[23] Jonas Kübler, Simon Buchholz, and Bernhard Schölkopf. The inductive bias of quantum kernels. *Advances in Neural Information Processing Systems*, 34, 2021.

[24] Maria Schuld. Quantum machine learning models are kernel methods. *arXiv:2101.11020*, pages 1–26, jan 2021.

[25] Seth Lloyd, Maria Schuld, Aroosa Ijaz, Josh Izaac, and Nathan Killoran. Quantum embeddings for machine learning. *arXiv:2002.08953*, pages 1–11, jan 2020.

[26] 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.

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

[28] 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. *arXiv:2101.09581*, pages 1–20, jan 2021.

[29] 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, mar 2019.

[30] Maria Schuld, Ryan Sweke, and Johannes Jakob Meyer. Effect of data encoding on the expressive power of variational quantum-machine-learning models. *Physical Review A*, 103(3):032430, 2021.

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

[32] Hiroshi Yano, Yudai Suzuki, Rudy Raymond, and Naoki Yamamoto. Efficient discrete feature encoding for variational quantum classifier. In *2020 IEEE International Conference on Quantum Computing and Engineering (QCE)*, pages 11–21. IEEE, 2020.

[33] Michael A. Nielsen and Isaac Chuang. Quantum Computation and Quantum Information. *American Journal of Physics*, 70(5):558–559, may 2002.

[34] Martin Müller-Lennert, Frédéric Dupuis, Oleg Szehr, Serge Fehr, and Marco Tomamichel. On quantum Rényi entropies: A new generalization and some properties. *Journal of Mathematical Physics*, 54(12):122203, dec 2013.

[35] Mark M Wilde, Andreas Winter, and Dong Yang. Strong Converse for the Classical Capacity of Entanglement-Breaking and Hadamard Channels via a Sandwiched Rényi Relative Entropy. *Communications in Mathematical Physics*, 331(2):593–622, oct 2014.

[36] Dénes Petz. Quasi-entropies for finite quantum systems. *Reports on Mathematical Physics*, 23(1):57–65, feb 1986.

[37] Koenraad M R Audenaert. A sharp continuity estimate for the von Neumann entropy. *Journal of Physics A: Mathematical and Theoretical*, 40(28):8127–8136, jul 2007.

[38] Michael Nussbaum and Arleta Szkoła. The Chernoff lower bound for symmetric quantum hypothesis testing. *The Annals of Statistics*, 37(2):1040–1057, 2009.

[39] Robert Salzmann, Nilanjana Datta, Gilad Gour, Xin Wang, and Mark M. Wilde. Symmetric distinguishability as a quantum resource. *New Journal of Physics*, 23(8):083016, aug 2021.

[40] Maria Kieferova, Ortiz Marrero Carlos, and Nathan Wiebe. Quantum Generative Training Using Rényi Divergences. *arXiv:2106.09567*, jun 2021.

[41] Kun Fang and Hamza Fawzi. Geometric Rényi Divergence and its Applications in Quantum Channel Capacities. *Communications in Mathematical Physics*, 384(3):1615–1677, jun 2021.

[42] Wassily Hoeffding. Probability inequalities for sums of bounded random variables. In *The collected works of Wassily Hoeffding*, pages 409–426. Springer, 1994.

[43] Maria Schuld and Nathan Killoran. Quantum Machine Learning in Feature Hilbert Spaces. *Physical Review Letters*, 122(4):040504, feb 2019.

[44] Edward Farhi and Hartmut Neven. Classification with Quantum Neural Networks on Near Term Processors. *arXiv:1802.06002*, pages 1–21, feb 2018.

[45] K. Mitarai, M. Negoro, M. Kitagawa, and K. Fujii. Quantum circuit learning. *Physical Review A*, 98(3):032309, sep 2018.

[46] 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.

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

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

[49] Jarrod R. McClean, Sergio Boixo, Vadim N. Smelyanskiy, Ryan Babbush, and Hartmut Neven. Barren plateaus in quantum neural network training landscapes. *Nature Communications*, 9(1):4812, dec 2018.

[50] Christoph Dankert, Richard Cleve, Joseph Emerson, and Etera Livine. Exact and approximate unitary 2-designs and their application to fidelity estimation. *Physical Review A*, 80(1):012304, jul 2009.

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

[52] Ran Canetti. Universally composable security: A new paradigm for cryptographic protocols. In *Proceedings 42nd IEEE Symposium on Foundations of Computer Science*, pages 136–145. IEEE, 2001.

[53] Ryuji Takagi, Suguru Endo, Shintaro Minagawa, and Mile Gu. Fundamental limits of quantum error mitigation. *arXiv preprint arXiv:2109.04457*, 2021.

[54] DP DiVincenzo, DW Leung, and BM Terhal. Quantum data hiding. *IEEE Transactions on Information Theory*, 48(3):580–598, 2002.

[55] Carl W Helstrom. Quantum detection and estimation theory. *Journal of Statistical Physics*, 1(2):231–252, 1969.

[56] Joonwoo Bae and Won-Young Hwang. Minimum-error discrimination of qubit states: Methods, solutions, and properties. *Phys. Rev. A*, 87:012334, Jan 2013.

[57] Hongxiang Chen, Leonard Wossnig, Simone Severini, Hartmut Neven, and Masoud Mohseni. Universal discriminative quantum neural networks. *Quantum Machine Intelligence*, 3(1):1–11, 2021.

[58] Andrew Patterson, Hongxiang Chen, Leonard Wossnig, Simone Severini, Dan Browne, and Ivan Rungger. Quantum state discrimination using noisy quantum neural networks. *Physical Review Research*, 3(1):013063, 2021.

[59] Yanjun Ma, Dianhai Yu, Tian Wu, and Haifeng Wang. PaddlePaddle: An Open-Source Deep Learning Platform from Industrial Practice. *Frontiers of Data and Domputing*, 1(1):105–115, 2019.

[60] Diederik P. Kingma and Jimmy Lei Ba. Adam: A method for stochastic optimization. *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, dec 2015.

[61] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[62] Sofiene Jerbi, Lukas J. Fiderer, Hendrik Poulsen Nautrup, Jonas M. Kübler, Hans J. Briegel, and Vedran Dunjko. Quantum machine learning beyond kernel methods. *arXiv:2110.13162*, pages 1–13, 2021.

[63] Christopher A Fuchs and Jeroen Van De Graaf. Cryptographic distinguishability measures for quantum-mechanical states. *IEEE Transactions on Information Theory*, 45(4):1216–1227, 1999.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See Sec. 2.2.
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...

(a) Did you state the full set of assumptions of all theoretical results? [Yes] See Fig. 2(b) and the descriptions in every theorem.

(b) Did you include complete proofs of all theoretical results? [Yes] All proofs are included in Appendix.

3. If you ran experiments...

(a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [No] Our experimental results could be easily reproduced.

(b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Sec. 4.

(c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [No] All our numerical experiments could be run on a personal laptop.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...

(a) If your work uses existing assets, did you cite the creators? [Yes] See Sec. 4.2

(b) Did you mention the license of the assets? [No] It is publicly available.

(c) Did you include any new assets either in the supplemental material or as a URL? [No]

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]

5. If you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

Supplementary Material for "Concentration of Data Encoding in Parameterized Quantum Circuits"

A Proof of Theorem 1

Theorem 1. *Assume each element of an nD-dimensional vector $\boldsymbol{x}$ obeys an IGD, i.e., $x_{j,d} \sim \mathcal{N}(\mu_{j,d}, \sigma_{j,d}^2)$, where $\sigma_{j,d} \geq \sigma$ for some constant σ and $1 \leq j \leq n, 1 \leq d \leq D$. If $\boldsymbol{x}$ is encoded into an n-qubit pure state $\rho(\boldsymbol{x})$ according to the circuit in Fig. S1, then the quantum divergence between the average encoded state $\bar{\rho} = \mathbb{E}\left[\rho(\boldsymbol{x})\right]$ and the maximally mixed state $\mathbb{1} = \frac{I}{2^n}$ is upper bounded as*

$$D_2\left(\bar{\rho} \| \mathbb{1}\right) \equiv \log \operatorname{Tr}\left(\bar{\rho}^2 \cdot \mathbb{1}^{-1}\right) \leq n \log\left(1 + \mathrm{e}^{-D\sigma^2}\right). \tag{S1}$$

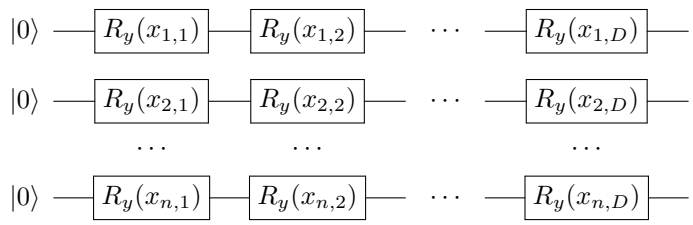


Figure S1: Circuit for the data encoding strategy with R_y rotations only.

Proof. Let $\rho(\boldsymbol{x}_j) \equiv R_y(x_{j,1} + \cdots + x_{j,D}) \left|0\right\rangle\left\langle 0\right| R_y^\dagger(x_{j,1} + \cdots + x_{j,D})$, then

$$\rho(\boldsymbol{x}) = \rho(\boldsymbol{x}_1) \otimes \rho(\boldsymbol{x}_2) \otimes \cdots \otimes \rho(\boldsymbol{x}_n). \tag{S2}$$

Due to the independence of each $\rho(\boldsymbol{x}_j)$, we have

$$\bar{\rho} = \mathbb{E}\left[\rho(\boldsymbol{x})\right] = \mathbb{E}\left[\rho(\boldsymbol{x}_1)\right] \otimes \mathbb{E}\left[\rho(\boldsymbol{x}_2)\right] \otimes \cdots \otimes \mathbb{E}\left[\rho(\boldsymbol{x}_n)\right]. \tag{S3}$$

What's more, for $j = 1, \ldots, n$,

$$\mathbb{E}\left[\rho(\boldsymbol{x}_j)\right] = \frac{1}{2}\mathbb{E}\begin{bmatrix} 1 + \cos\left(\sum_d x_{j,d}\right) & \sin\left(\sum_d x_{j,d}\right) \\ \sin\left(\sum_d x_{j,d}\right) & 1 - \cos\left(\sum_d x_{j,d}\right) \end{bmatrix} \tag{S4}$$

$$= \frac{1}{2}\begin{bmatrix} 1 + \mathbb{E}\left[\cos\left(\sum_d x_{j,d}\right)\right] & \mathbb{E}\left[\sin\left(\sum_d x_{j,d}\right)\right] \\ \mathbb{E}\left[\sin\left(\sum_d x_{j,d}\right)\right] & 1 - \mathbb{E}\left[\cos\left(\sum_d x_{j,d}\right)\right] \end{bmatrix}. \tag{S5}$$

Lemma 1. *Assume a variable $x \sim \mathcal{N}(\mu, \sigma^2)$, then*

$$\mathbb{E}\left[\cos(x)\right] = \mathrm{e}^{-\frac{\sigma^2}{2}} \cos(\mu); \qquad \mathbb{E}\left[\sin(x)\right] = \mathrm{e}^{-\frac{\sigma^2}{2}} \sin(\mu). \tag{S6}$$

We know $\sum_d x_{j,d} \sim \mathcal{N}(\sum_d \mu_{j,d}, \sqrt{\sum_d \sigma_{j,d}^2})$, and combining Eq. (S5) with Lemma 1, we have

$$\left|\left|\mathbb{E}\left[\rho(\boldsymbol{x}_j)\right]\right|\right|_F^2 = \frac{1}{2} + \frac{1}{2}\mathbb{E}^2\left[\cos\left(\sum_d x_{j,d}\right)\right] + \frac{1}{2}\mathbb{E}^2\left[\sin\left(\sum_d x_{j,d}\right)\right] \tag{S7}$$

$$= \frac{1}{2} + \frac{1}{2}\left(\mathrm{e}^{-\frac{\sum_d \sigma_{j,d}^2}{2}} \cos\left(\sum_d \mu_{j,d}\right)\right)^2 + \frac{1}{2}\left(\mathrm{e}^{-\frac{\sum_d \sigma_{j,d}^2}{2}} \sin\left(\sum_d \mu_{j,d}\right)\right)^2 \tag{S8}$$

$$= \frac{1}{2} + \frac{1}{2}\mathrm{e}^{-\sum_d \sigma_{j,d}^2} \tag{S9}$$

$$\leq \frac{1}{2} + \frac{1}{2}\mathrm{e}^{-D\sigma^2}. \tag{S10}$$

Finally, from Eq. (S3), we have

$$\log \operatorname{Tr}\left(\bar{\rho}^2 \cdot \left(\frac{I}{2^n}\right)^{-1}\right) = \log\left(2^n \|\bar{\rho}\|_F^2\right) \tag{S11}$$

$$= \log\left(2^n \prod_{j=1}^{n} \left|\left|\mathbb{E}\left[\rho(\boldsymbol{x}_j)\right]\right|\right|_F^2\right) \tag{S12}$$

$$\leq \log\left(2^n \left(\frac{1+\mathrm{e}^{-D\sigma^2}}{2}\right)^n\right) \tag{S13}$$

$$= n\log(1+\mathrm{e}^{-D\sigma^2}). \tag{S14}$$

This completes the proof. □

B Proof of Theorem 2

Theorem 2. *Assume each element of a $3nD$-dimensional vector $\boldsymbol{x}$ obeys an IGD, i.e., $x_{j,d,k} \sim \mathcal{N}(\mu_{j,d,k}, \sigma^2_{j,d,k})$, where $\sigma_{j,d,k} \geq \sigma$ for some constant σ and $1 \leq j \leq n, 1 \leq d \leq D, 1 \leq k \leq 3$. If $\boldsymbol{x}$ is encoded into an n-qubit pure state $\rho(\boldsymbol{x})$ according to the circuit in Fig. S2, the quantum divergence between the average encoded state $\bar{\rho}$ and the maximally mixed state $\mathbb{1}$ is upper bounded as*

$$D_2\left(\bar{\rho}\|\mathbb{1}\right) \equiv \log \operatorname{Tr}\left(\bar{\rho}^2 \cdot \mathbb{1}^{-1}\right) \leq \log\left(1+(2^n-1)\mathrm{e}^{-D\sigma^2}\right). \tag{S15}$$

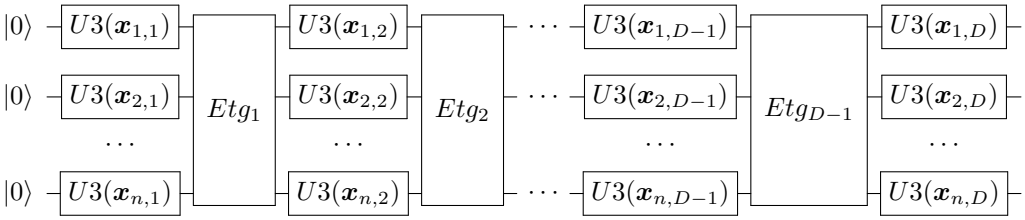


Figure S2: Circuit for the data encoding strategy with D layers of $U3$ gates and $D-1$ layers of entanglements. Here, each $\boldsymbol{x}_{j,d}$ represents three elements $x_{j,d,1}, x_{j,d,2}, x_{j,d,3}$, and each Etg_i denotes an arbitrary group of entangled two-qubit gates, such as CNOT or CZ, where $1 \leq j \leq n, 1 \leq d \leq D, 1 \leq i \leq D-1$.

Proof. In this proof, we consider the $U3(x_{j,d,1}, x_{j,d,2}, x_{j,d,3})$ gate as $R_z(x_{j,d,3}) \cdot R_y(x_{j,d,2}) \cdot R_z(x_{j,d,1})$, which is one of the most commonly used ones. Of course, other forms of U3 gate are similar.

Outline of Proof. **1) Decomposing initial state.** Firstly, we decompose the initial state according to Pauli bases; **2) Vectors transition.** Then by taking the corresponding coefficients as a row vector, we state that each action of a group of entangled gates Etg_i or a column of $U3$ gates is equivalent to multiplying the previous coefficient vector by a transition matrix; **3) Bound by singular value.** Finally, we get the upper bound by investigating the singular values of these transition matrices.

1) Decomposing initial state. The state after the first column of $U3$ gates becomes

$$\rho_1 = \frac{1}{2}\begin{bmatrix} 1+\cos(x_{1,1,2}) & \mathrm{e}^{-ix_{1,1,3}}\sin(x_{1,1,2}) \\ \mathrm{e}^{ix_{1,1,3}}\sin(x_{1,1,2}) & 1-\cos(x_{1,1,2}) \end{bmatrix} \otimes \frac{1}{2}\begin{bmatrix} 1+\cos(x_{2,1,2}) & \mathrm{e}^{-ix_{2,1,3}}\sin(x_{2,1,2}) \\ \mathrm{e}^{ix_{2,1,3}}\sin(x_{2,1,2}) & 1-\cos(x_{2,1,2}) \end{bmatrix}$$
$$\otimes \cdots\cdots \otimes \frac{1}{2}\begin{bmatrix} 1+\cos(x_{n,1,2}) & \mathrm{e}^{-ix_{n,1,3}}\sin(x_{n,1,2}) \\ \mathrm{e}^{ix_{n,1,3}}\sin(x_{n,1,2}) & 1-\cos(x_{n,1,2}) \end{bmatrix}. \tag{S16}$$

Table 1: The transition table for tensor products of Pauli bases when applying CNOT or CZ gates.

Pauli bases	Apply CNOT	Apply CZ
$I \otimes I$	$I \otimes I$	$I \otimes I$
$I \otimes Z$	$Z \otimes Z$	$I \otimes Z$
$I \otimes X$	$I \otimes X$	$Z \otimes X$
$I \otimes Y$	$Z \otimes Y$	$Z \otimes Y$
$Z \otimes I$	$Z \otimes I$	$Z \otimes I$
$Z \otimes Z$	$I \otimes Z$	$Z \otimes Z$
$Z \otimes X$	$Z \otimes X$	$I \otimes X$
$Z \otimes Y$	$I \otimes Y$	$I \otimes Y$
$X \otimes I$	$X \otimes X$	$X \otimes Z$
$X \otimes Z$	-$Y \otimes Y$	$X \otimes I$
$X \otimes X$	$X \otimes I$	$Y \otimes Y$
$X \otimes Y$	$Y \otimes Z$	-$Y \otimes X$
$Y \otimes I$	$Y \otimes X$	$Y \otimes Z$
$Y \otimes Z$	$X \otimes Y$	$Y \otimes I$
$Y \otimes X$	$Y \otimes I$	-$X \otimes Y$
$Y \otimes Y$	-$X \otimes Z$	$X \otimes X$

Now we define $\rho_1 \equiv \rho_{1,1} \otimes \rho_{1,2} \otimes \cdots \otimes \rho_{1,n}$, where

$$\rho_{1,j} \equiv \frac{1}{2}\begin{bmatrix} 1+\cos(x_{j,1,2}) & \mathrm{e}^{-ix_{j,1,3}}\sin(x_{j,1,2}) \\ \mathrm{e}^{ix_{j,1,3}}\sin(x_{j,1,2}) & 1-\cos(x_{j,1,2}) \end{bmatrix}. \tag{S17}$$

And from Lemma 1, we have

$$\mathbb{E}\left[\rho_{1,j}\right] = \frac{1}{2}\begin{bmatrix} 1+A_{j,1,2}\cos(\mu_{j,1,2}) & A_{j,1,3}\mathrm{e}^{-i\mu_{j,1,3}}A_{j,1,2}\sin(\mu_{j,1,2}) \\ A_{j,1,3}\mathrm{e}^{i\mu_{j,1,3}}A_{j,1,2}\sin(\mu_{j,1,2}) & 1-A_{j,1,2}\cos(\mu_{j,1,2}) \end{bmatrix}, \tag{S18}$$

where we define $A_{j,d,k} = \mathrm{e}^{-\frac{\sigma^2_{j,d,k}}{2}}$ for writing convenience. Here we note that due to all $x_{j,d,k}$'s being independent of each other, calculating the expectation of $\rho_{1,j}$ in advance does not affect the following computations. Next we decompose $\mathbb{E}\left[\rho_{1,j}\right]$ according to the Pauli bases, i.e., I, Z, X, Y,

$$\begin{aligned}\mathbb{E}\left[\rho_{1,j}\right] = \frac{1}{2}I + \frac{A_{j,1,2}\cos(\mu_{j,1,2})}{2}Z + \frac{A_{j,1,3}\cos(\mu_{j,1,3})A_{j,1,2}\sin(\mu_{j,1,2})}{2}X \\ + \frac{A_{j,1,3}\sin(\mu_{j,1,3})A_{j,1,2}\sin(\mu_{j,1,2})}{2}Y.\end{aligned} \tag{S19}$$

Then from Eq. (S16), we could derive that $\mathbb{E}\left[\rho_1\right] = \bigotimes_{j=1}^{n}\mathbb{E}\left[\rho_{1,j}\right]$ can also be decomposed in accordance with various tensor products of Pauli bases. Therefore, studying the state after the gate Etg_1 could be transferred to what performance it will be when entangled two-qubit gates act on the tensor products of Pauli bases.

2) Vectors transition. Here, we focus on the two widely employed two-qubit entangled gates CNOT and CZ, and the calculations are concluded in Table 1. The results in the table show that the transitions are closed for tensor products of Pauli bases. Here we note that other entangled two-qubit gates will have a similar effect.

Next, we consider the effects of applying the gate $U3(x_1, x_2, x_3) = R_z(x_3)R_y(x_2)R_z(x_1)$ to four Pauli matrices. And the results of the calculations are as follows:

$$\mathbb{E}\left[U3 \cdot I \cdot U3^\dagger\right] = I \tag{S20}$$

$$\mathbb{E}\left[U3 \cdot Z \cdot U3^\dagger\right] = p_{zz}Z + p_{zx}X + p_{zy}Y \tag{S21}$$

$$\mathbb{E}\left[U3 \cdot X \cdot U3^\dagger\right] = p_{xz}Z + p_{xx}X + p_{xy}Y \tag{S22}$$

$$\mathbb{E}\left[U3 \cdot Y \cdot U3^\dagger\right] = p_{yz}Z + p_{yx}X + p_{yy}Y, \tag{S23}$$

where

$$p_{zz} = A_2 \cos(\mu_2) \tag{S24}$$
$$p_{zx} = A_2 \sin(\mu_2) A_3 \cos(\mu_3) \tag{S25}$$
$$p_{zy} = A_2 \sin(\mu_2) A_3 \sin(\mu_3) \tag{S26}$$
$$p_{xz} = -A_2 \sin(\mu_2) A_1 \cos(\mu_1) \tag{S27}$$
$$p_{xx} = A_2 \cos(\mu_2) A_1 \cos(\mu_1) A_3 \cos(\mu_3) - A_1 \sin(\mu_1) A_3 \sin(\mu_3) \tag{S28}$$
$$p_{xy} = A_2 \cos(\mu_2) A_1 \cos(\mu_1) A_3 \sin(\mu_3) + A_1 \sin(\mu_1) A_3 \cos(\mu_3) \tag{S29}$$
$$p_{yz} = A_2 \sin(\mu_2) A_1 \sin(\mu_1) \tag{S30}$$
$$p_{yx} = -A_2 \cos(\mu_2) A_1 \sin(\mu_1) A_3 \cos(\mu_3) - A_1 \cos(\mu_1) A_3 \sin(\mu_3) \tag{S31}$$
$$p_{yy} = -A_2 \cos(\mu_2) A_1 \sin(\mu_1) A_3 \sin(\mu_3) + A_1 \cos(\mu_1) A_3 \cos(\mu_3). \tag{S32}$$

Here, x_k, A_k, μ_k are the abbreviations for $x_{j,d,k}, A_{j,d,k}, \mu_{j,d,k}$, respectively.

Now we record Eqs. (S20)-(S23) as a matrix T, which we call *transition matrix*, i.e.,

$$T \equiv \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & p_{zz} & p_{zx} & p_{zy} \\ 0 & p_{xz} & p_{xx} & p_{xy} \\ 0 & p_{yz} & p_{yx} & p_{yy} \end{bmatrix}. \tag{S33}$$

By carefully calculating Eqs. (S20)-(S33) again, we also have

$$T = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & A_1 \cos(\mu_1) & A_1 \sin(\mu_1) \\ 0 & 0 & -A_1 \sin(\mu_1) & A_1 \cos(\mu_1) \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & A_2 \cos(\mu_2) & A_2 \sin(\mu_2) & 0 \\ 0 & -A_2 \sin(\mu_2) & A_2 \cos(\mu_2) & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & A_3 \cos(\mu_3) & A_3 \sin(\mu_3) \\ 0 & 0 & -A_3 \sin(\mu_3) & A_3 \cos(\mu_3) \end{bmatrix}, \tag{S34}$$

where the three matrices correspond to the effects of applying $R_z(x_1)$, $R_y(x_2)$ and $R_z(x_3)$, respectively.

If we further record an arbitrary input $\rho_{in} \equiv \alpha_1 I + \alpha_2 Z + \alpha_3 X + \alpha_4 Y$ as a row vector $\pi_{in} = [\alpha_1 \quad \alpha_2 \quad \alpha_3 \quad \alpha_4]$, then applying the gate $U3(x_1, x_2, x_3)$ to ρ_{in} will result in the output $\rho_{out} \equiv \beta_1 I + \beta_2 Z + \beta_3 X + \beta_4 Y$, where

$$\pi_{out} \equiv [\beta_1 \quad \beta_2 \quad \beta_3 \quad \beta_4] = [\alpha_1 \quad \alpha_2 \quad \alpha_3 \quad \alpha_4] \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & p_{zz} & p_{zx} & p_{zy} \\ 0 & p_{xz} & p_{xx} & p_{xy} \\ 0 & p_{yz} & p_{yx} & p_{yy} \end{bmatrix} = \pi_{in} T. \tag{S35}$$

This is a fundamental relationship in this proof, which can be easily verified in multi-qubit and multi-depth cases. Hence, we could rewrite each $\mathbb{E}[\rho_d]$, $0 \le d \le D$, as follows

$$\mathbb{E}[\rho_0] \quad \longleftrightarrow \quad \pi_0 = \otimes_{j=1}^n \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 & 0 \end{bmatrix} \tag{S36}$$
$$\mathbb{E}[\rho_1] \quad \longleftrightarrow \quad \pi_1 = \pi_0 \cdot \otimes_{j=1}^n T_{j,1} \cdot \widetilde{Etg_1} \tag{S37}$$
$$\cdots$$
$$\mathbb{E}[\rho_d] \quad \longleftrightarrow \quad \pi_d = \pi_{d-1} \cdot \otimes_{j=1}^n T_{j,d} \cdot \widetilde{Etg_d} \tag{S38}$$
$$\cdots$$
$$\mathbb{E}[\rho_{D-1}] \quad \longleftrightarrow \quad \pi_{D-1} = \pi_{D-2} \cdot \otimes_{j=1}^n T_{j,D-1} \cdot \widetilde{Etg_{D-1}} \tag{S39}$$
$$\mathbb{E}[\rho_D] \quad \longleftrightarrow \quad \pi_D = \pi_{D-1} \cdot \otimes_{j=1}^n T_{j,D}, \tag{S40}$$

where each $T_{j,d}$ represents that this transition matrix is constructed based on the gate $U3(x_{j,d,1}, x_{j,d,2}, x_{j,d,3})$ and each $\widetilde{Etg_i}$ means rearranging the elements of the previously multiplied row vector, which is equivalent to the effect after applying Etg_i, $1 \le i \le D-1$. Here, please note that we omit the possible negative sign described in Table 1, because in the following proof, it has no influence.

From the fact that $\mathrm{Tr}\left(P_i^2\right) = 2$, $\mathrm{Tr}\left(P_i P_j\right) = 0$, where P_i, P_j denote different Pauli matrices, and combining the relationship in Eq. (S40), we have

$$\mathrm{Tr}\left(\mathbb{E}[\rho_D]\right)^2 = 2^n \cdot \pi_D \left(\pi_D\right)^\top. \tag{S41}$$

What's more, we also find that every $\otimes_{j=1}^{n} T_{j,d}$ and $\widetilde{Etg_i}$ always have an element 1 in the top left corner, i.e.,

$$\otimes_{j=1}^{n} T_{j,d} \equiv \begin{bmatrix} 1 & \\ & \mathcal{T}_d \end{bmatrix}, \qquad \widetilde{Etg_i} \equiv \begin{bmatrix} 1 & \\ & \mathcal{E}_i \end{bmatrix}, \tag{S42}$$

where $\mathcal{T}_d, \mathcal{E}_i \in \mathbb{R}^{(4^n-1)\times(4^n-1)}$ and $1 \le d \le D, 1 \le i \le D-1$. Therefore,

$$\begin{aligned}
\mathrm{Tr}\left(\mathbb{E}\left[\rho_D\right]\right)^2 &= 2^n \cdot \pi_D \left(\pi_D\right)^\top && \text{(S43)}\\
&= 2^n \cdot \pi_0 \begin{bmatrix} 1 & \\ & \mathcal{T}_1\mathcal{E}_1\mathcal{T}_2\mathcal{E}_2\cdots\mathcal{T}_D \end{bmatrix} \begin{bmatrix} 1 & \\ & \mathcal{T}_D^\top\cdots\mathcal{E}_2^\top\mathcal{T}_2^\top\mathcal{E}_1^\top\mathcal{T}_1^\top \end{bmatrix} \left(\pi_0\right)^\top && \text{(S44)}\\
&= 2^n \cdot \begin{bmatrix} \frac{1}{2^n} & \mathring{\pi}_0 \end{bmatrix} \begin{bmatrix} 1 & \\ & \mathcal{T}_1\mathcal{E}_1\mathcal{T}_2\mathcal{E}_2\cdots\mathcal{T}_D\mathcal{T}_D^\top\cdots\mathcal{E}_2^\top\mathcal{T}_2^\top\mathcal{E}_1^\top\mathcal{T}_1^\top \end{bmatrix} \begin{bmatrix} \frac{1}{2^n} \\ \mathring{\pi}_0^\top \end{bmatrix} && \text{(S45)}\\
&= \frac{1}{2^n} + 2^n \cdot \mathring{\pi}_0\mathcal{T}_1\mathcal{E}_1\cdots\mathcal{T}_{D-1}\mathcal{E}_{D-1}\mathcal{T}_D\mathcal{T}_D^\top\mathcal{E}_{D-1}^\top\mathcal{T}_{D-1}^\top\cdots\mathcal{E}_1^\top\mathcal{T}_1^\top\mathring{\pi}_0^\top, && \text{(S46)}
\end{aligned}$$

where $\mathring{\pi}_0$ means that the row vector π_0 removes the first element.

3) Bound by singular value. Next, in order to further calculate it, we need to prove first the following two Lemmas.

Lemma 2. *Given a Hermitian matrix $H \in \mathbb{C}^{n\times n}$ with all its eigenvalues no larger than λ, and an n-dimensional vector $\boldsymbol{x}$, then*

$$\boldsymbol{x}^\dagger H \boldsymbol{x} \le \|\boldsymbol{x}\|_2^2 \lambda, \tag{S47}$$

where $\|\cdot\|_2$ denotes the l_2-norm.

Proof. Assume H has the spectral decomposition

$$H = \sum_{i=1}^{n} \lambda_i \boldsymbol{u}_i \boldsymbol{u}_i^\dagger, \tag{S48}$$

then $\boldsymbol{x}$ can be uniquely decomposed as $\boldsymbol{x} = \sum_{i=1}^{n} \alpha_i \boldsymbol{u}_i$ with $\sum_{i=1}^{n} |\alpha_i|^2 = \|\boldsymbol{x}\|_2^2$. Finally we have

$$\boldsymbol{x}^\dagger H \boldsymbol{x} = \sum_{i=1}^{n} |\alpha_i|^2 \lambda_i \le \sum_{i}^{n} |\alpha_i|^2 \lambda = \|\boldsymbol{x}\|_2^2 \lambda. \tag{S49}$$

□

Lemma 2. *Given a Hermitian matrix $H \in \mathbb{C}^{n\times n}$ with all its eigenvalues no larger than λ, and an arbitrary matrix $Q \in \mathbb{C}^{n\times n}$ with all its singular values no larger than s, then the largest eigenvalue of $QHQ^\dagger$ is no larger than $s^2\lambda$.*

Proof. The largest eigenvalue of $QHQ^\dagger$ can be computed as $\lambda_{max} \equiv \max_{\boldsymbol{x}} \boldsymbol{x}^\dagger QHQ^\dagger \boldsymbol{x}$, where $\boldsymbol{x}$ denotes a unit vector. Assume Q has the singular value decomposition

$$Q = USV^\dagger = \sum_{i=1}^{n} s_i \boldsymbol{u}_i \boldsymbol{v}_i^\dagger = \begin{bmatrix} \boldsymbol{u}_1 & \boldsymbol{u}_2 & \cdots & \boldsymbol{u}_n \end{bmatrix} \begin{bmatrix} s_1 & 0 & 0 & 0 \\ 0 & s_2 & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & s_n \end{bmatrix} \begin{bmatrix} \boldsymbol{v}_1^\dagger \\ \boldsymbol{v}_2^\dagger \\ \vdots \\ \boldsymbol{v}_n^\dagger \end{bmatrix} \tag{S50}$$

and $\boldsymbol{x} = \sum_{i=1}^{n} \alpha_i \boldsymbol{u}_i$ with $\sum_{i=1}^{n} |\alpha_i|^2 = 1$, then

$$\boldsymbol{x}^\dagger QHQ^\dagger \boldsymbol{x} = \boldsymbol{x}^\dagger USV^\dagger HVSU^\dagger \boldsymbol{x} = \begin{bmatrix} \alpha_1^\dagger s_1 & \alpha_2^\dagger s_2 & \cdots & \alpha_n^\dagger s_n \end{bmatrix} V^\dagger HV \begin{bmatrix} \alpha_1 s_1 \\ \alpha_2 s_2 \\ \cdots \\ \alpha_n s_n \end{bmatrix}. \tag{S51}$$

Consider $V^\dagger HV$ as a new Hermitian matrix and $SU^\dagger \boldsymbol{x}$ as a new vector $\tilde{\boldsymbol{x}}$, then all the eigenvalues of $V^\dagger HV$ are still no larger than λ and the square of the l_2-norm of $\tilde{\boldsymbol{x}}$ is computed as

$$\|\tilde{\boldsymbol{x}}\|_2^2 = \sum_{i=1}^{n} |\alpha_i|^2 s_i^2 \leq \sum_{i=1}^{n} |\alpha_i|^2 s^2 = s^2. \tag{S52}$$

From Lemma 2, we have

$$\boldsymbol{x}^\dagger QHQ^\dagger \boldsymbol{x} \leq \|\tilde{\boldsymbol{x}}\|_2^2 \lambda \leq s^2 \lambda. \tag{S53}$$

As $\boldsymbol{x}$ is arbitrary, we can obtain that $\lambda_{max} \equiv \max_{\boldsymbol{x}} \boldsymbol{x}^\dagger QHQ^\dagger \boldsymbol{x}$ is no larger than $s^2\lambda$ as well. □

Now, let us investigate the singular values of $T_{j,d}$. From Eq. (S33), we know it always has the trivial biggest singular value 1. The second-biggest singular value s_m can be derived from

$$s_m^2 = \max_{\boldsymbol{u}} \boldsymbol{u}^\dagger \begin{bmatrix} p_{zz} & p_{zx} & p_{zy} \\ p_{xz} & p_{xx} & p_{xy} \\ p_{yz} & p_{yx} & p_{yy} \end{bmatrix} \begin{bmatrix} p_{zz} & p_{xz} & p_{yz} \\ p_{zx} & p_{xx} & p_{yx} \\ p_{zy} & p_{xy} & p_{yy} \end{bmatrix} \boldsymbol{u}, \tag{S54}$$

where $\boldsymbol{u} \in \mathbb{C}^3$ denotes a unit column vector. From Eq. (S34), we derive that

$$\begin{bmatrix} p_{zz} & p_{zx} & p_{zy} \\ p_{xz} & p_{xx} & p_{xy} \\ p_{yz} & p_{yx} & p_{yy} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & A_1\cos(\mu_1) & A_1\sin(\mu_1) \\ 0 & -A_1\sin(\mu_1) & A_1\cos(\mu_1) \end{bmatrix} \begin{bmatrix} A_2\cos(\mu_2) & A_2\sin(\mu_2) & 0 \\ -A_2\sin(\mu_2) & A_2\cos(\mu_2) & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & A_3\cos(\mu_3) & A_3\sin(\mu_3) \\ 0 & -A_3\sin(\mu_3) & A_3\cos(\mu_3) \end{bmatrix} \tag{S55}$$

$$= \begin{bmatrix} 1 & 0 & 0 \\ 0 & A_1\cos(\mu_1) & A_1\sin(\mu_1) \\ 0 & -A_1\sin(\mu_1) & A_1\cos(\mu_1) \end{bmatrix} \begin{bmatrix} \cos(\mu_2) & \sin(\mu_2) & 0 \\ -\sin(\mu_2) & \cos(\mu_2) & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} A_2 & & \\ & A_2A_3 & 0 \\ 0 & 0 & A_3 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos(\mu_3) & \sin(\mu_3) \\ 0 & -\sin(\mu_3) & \cos(\mu_3) \end{bmatrix}, \tag{S56}$$

hence,

$$\begin{bmatrix} p_{zz} & p_{zx} & p_{zy} \\ p_{xz} & p_{xx} & p_{xy} \\ p_{yz} & p_{yx} & p_{yy} \end{bmatrix} \begin{bmatrix} p_{zz} & p_{xz} & p_{yz} \\ p_{zx} & p_{xx} & p_{yx} \\ p_{zy} & p_{xy} & p_{yy} \end{bmatrix} = Q \begin{bmatrix} \cos(\mu_2) & \sin(\mu_2) & 0 \\ -\sin(\mu_2) & \cos(\mu_2) & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} A_2^2 & & \\ & (A_2A_3)^2 & 0 \\ 0 & 0 & A_3^2 \end{bmatrix} \begin{bmatrix} \cos(\mu_2) & -\sin(\mu_2) & 0 \\ \sin(\mu_2) & \cos(\mu_2) & 0 \\ 0 & 0 & 1 \end{bmatrix} Q^\top, \tag{S57}$$

where $Q \equiv \begin{bmatrix} 1 & 0 & 0 \\ 0 & A_1\cos(\mu_1) & A_1\sin(\mu_1) \\ 0 & -A_1\sin(\mu_1) & A_1\cos(\mu_1) \end{bmatrix}$ has the largest singular value 1.

From Lemma 2, we deduce that the largest eigenvalue of the matrix in Eq. (S57) is $\max\left\{A_2^2, A_3^2\right\}$, which is no larger than $\mathrm{e}^{-\sigma^2}$. Further combining Eq. (S54), we infer that s_m is no larger than $\mathrm{e}^{-\frac{\sigma^2}{2}}$, i.e., the second-biggest singular value of each $T_{j,d}$ is no larger than $\mathrm{e}^{-\frac{\sigma^2}{2}}$. What's more, we could derive that their tensor product $\otimes_{j=1}^n T_{j,d}$ also has the trivial largest singular value 1 and the second-largest singular value which is no larger than $\mathrm{e}^{-\frac{\sigma^2}{2}}$.

From the definition of $\mathcal{T}_d$ in Eq. (S42), we declare that the largest singular value of each $\mathcal{T}_d$ is no larger than $\mathrm{e}^{-\frac{\sigma^2}{2}}$. Let's go back to the following formula in Eq. (S46) to continue estimating $\mathrm{Tr}\left(\mathbb{E}\left[\rho_D\right]\right)^2$, i.e.,

$$\mathring{\pi}_0 \mathcal{T}_1 \mathcal{E}_1 \cdots \mathcal{T}_{D-1} \mathcal{E}_{D-1} \mathcal{T}_D \mathcal{T}_D^\top \mathcal{E}_{D-1}^\top \mathcal{T}_{D-1}^\top \cdots \mathcal{E}_1^\top \mathcal{T}_1^\top \mathring{\pi}_0^\top. \tag{S58}$$

Since the largest eigenvalue of $\mathcal{T}_D\mathcal{T}_D^\top$ is no larger than $\mathrm{e}^{-\sigma^2}$, and each $\mathcal{E}_i$, defined in Eq. (S42), is a unitary matrix, by repeatedly applying Lemma 2, we obtain that largest eigenvalue of $\mathcal{T}_1\mathcal{E}_1 \cdots \mathcal{T}_D\mathcal{T}_D^\top \cdots \mathcal{E}_1^\top\mathcal{T}_1^\top$ is no larger than $\mathrm{e}^{-D\sigma^2}$. Furthermore, from Eq. (S36) and the definition of $\mathring{\pi}_0$, we know $\mathring{\pi}_0$ has $4^n - 1$ dimensions, where $2^n - 1$ elements are $\frac{1}{2^n}$ and the others are 0. Hence, $\|\mathring{\pi}_0\|_2^2 = \frac{2^n-1}{2^{2n}}$. Combining these with Lemma 2, we have

$$\mathring{\pi}_0 \mathcal{T}_1 \mathcal{E}_1 \cdots \mathcal{T}_{D-1} \mathcal{E}_{D-1} \mathcal{T}_D \mathcal{T}_D^\top \mathcal{E}_{D-1}^\top \mathcal{T}_{D-1}^\top \cdots \mathcal{E}_1^\top \mathcal{T}_1^\top \mathring{\pi}_0^\top \leq \|\mathring{\pi}_0\|_2^2 \mathrm{e}^{-D\sigma^2} = \frac{2^n-1}{2^{2n}} \mathrm{e}^{-D\sigma^2}. \tag{S59}$$

Go further, and we have, together with Eq. (S46),

$$\mathrm{Tr}\left(\mathbb{E}\left[\rho_D\right]\right)^2 = \frac{1}{2^n} + 2^n \cdot \mathring{\pi}_0 \mathcal{T}_1 \mathcal{E}_1 \cdots \mathcal{T}_{D-1} \mathcal{E}_{D-1} \mathcal{T}_D \mathcal{T}_D^\top \mathcal{E}_{D-1}^\top \mathcal{T}_{D-1}^\top \cdots \mathcal{E}_1^\top \mathcal{T}_1^\top \mathring{\pi}_0^\top \tag{S60}$$

$$\leq \frac{1}{2^n} + 2^n \cdot \frac{2^n-1}{2^{2n}} \mathrm{e}^{-D\sigma^2} \tag{S61}$$

$$= \frac{1 + (2^n - 1)\,\mathrm{e}^{-D\sigma^2}}{2^n}. \tag{S62}$$

Finally, we have

$$\log \operatorname{Tr}\left(\bar{\rho}^2 \cdot \left(\frac{I}{2^n}\right)^{-1}\right) = \log\left(2^n \cdot \operatorname{Tr}\left(\mathbb{E}\left[\rho_D\right]\right)^2\right) \tag{S63}$$

$$\leq \log\left(1 + (2^n - 1)\mathrm{e}^{-D\sigma^2}\right). \tag{S64}$$

This completes the proof of Theorem 2.

Without stopping here, we also analyse some generalizations of Theorem 2.

(I) From Eqs. (S54)-(S57), we find that removing the matrix Q will have no influence on the final result. Hence, we directly generalize that if there is only one column of $R_z R_y$ or $R_y R_z$ gates in each layer, we will get the same upper bound. In fact, according to our proof method, we infer that as long as there are two different kinds of rotation gates in each encoding layer, this upper bound is valid.

(II) What is the result for the case with only R_y rotation gates in each encoding layer? Since each $\mathcal{T}_d$ has the largest singular value 1, it is not suitable for the above proof. However, through analyzing the transition rule in Table 1, we find that the largest singular value of $\mathcal{T}_{d-1}\mathcal{E}_{d-1}\mathcal{T}_d$ is still no larger than $\mathrm{e}^{-\frac{\sigma^2}{2}}$, which means every two encoding layers have the same effect as above with one layer. Therefore, the final upper bound can be changed to $\log\left(1 + (2^n - 1)\mathrm{e}^{-\lfloor \frac{D}{2} \rfloor \sigma^2}\right)$. Since it has the same trend as the original bound, it has no impact on our final analysis. □

C Proof of Corollary 2.1

Corollary 2.1. *Assume there are M classical vectors $\{\boldsymbol{x}^{(m)}\}_{m=1}^M$ sampled from the distributions described in Theorem 2 and define $\bar{\rho}_M \equiv \frac{1}{M}\sum_{m=1}^M \rho(\boldsymbol{x}^{(m)})$. Let H be a Hermitian matrix with its eigenvalues ranging in $[-1, 1]$, then given an arbitrary $\epsilon \in (0, 1)$, as long as the depth $D \geq \frac{1}{\sigma^2}\left[(n+4)\ln 2 + 2\ln(1/\epsilon)\right]$, we have*

$$\left|\operatorname{Tr}\left[H\left(\bar{\rho}_M - \mathbb{1}\right)\right]\right| \leq \epsilon \tag{S65}$$

with a probability of at least $1 - 2\mathrm{e}^{-M\epsilon^2/8}$.

Proof. Let $\bar{\rho} \equiv \mathbb{E}\left[\rho\left(\boldsymbol{x}^{(m)}\right)\right]$, then we have

$$\left|\operatorname{Tr}\left[H\left(\bar{\rho}_M - \frac{I}{2^n}\right)\right]\right| = \left|\operatorname{Tr}\left[H\left(\bar{\rho}_M - \bar{\rho} + \bar{\rho} - \frac{I}{2^n}\right)\right]\right| \tag{S66}$$

$$\leq \left|\operatorname{Tr}\left[H\left(\bar{\rho}_M - \bar{\rho}\right)\right]\right| + \left|\operatorname{Tr}\left[H\left(\bar{\rho} - \frac{I}{2^n}\right)\right]\right|, \tag{S67}$$

where the inequality is due to triangle inequality.

Now we first consider the first term in Eq. (S67). Since $\frac{1}{M}\operatorname{Tr}\left(H\rho(\boldsymbol{x}^{(1)})\right), \ldots, \frac{1}{M}\operatorname{Tr}\left(H\rho(\boldsymbol{x}^{(M)})\right)$ are i.i.d. and $\frac{-1}{M} \leq \frac{1}{M}\operatorname{Tr}\left(H\rho(\boldsymbol{x}^{(m)})\right) \leq \frac{1}{M}$, through *Hoeffding's inequality* [42], we have

$$P\left(\left|\sum_{m=1}^M \frac{1}{M}\operatorname{Tr}\left(H\rho(\boldsymbol{x}^{(m)})\right) - \mathbb{E}\left[\operatorname{Tr}\left(H\rho(\boldsymbol{x}^{(m)})\right)\right]\right| \leq t\right) \geq 1 - 2\mathrm{e}^{-\frac{Mt^2}{2}}. \tag{S68}$$

From the fact that $\sum_{m=1}^M \frac{1}{M}\operatorname{Tr}\left(H\rho(\boldsymbol{x}^{(m)})\right) = \operatorname{Tr}\left(H\bar{\rho}_M\right)$ and $\mathbb{E}\left[\operatorname{Tr}\left(H\rho(\boldsymbol{x}^{(m)})\right)\right] = \operatorname{Tr}\left(H\bar{\rho}\right)$, we obtain

$$\left|\operatorname{Tr}\left(H\bar{\rho}_M\right) - \operatorname{Tr}\left(H\bar{\rho}\right)\right| \leq \frac{\epsilon}{2} \tag{S69}$$

with a probability of at least $1 - 2\mathrm{e}^{-\frac{M\epsilon^2}{8}}$.

Next we consider the second term in Eq. (S67). Since the eigenvalues of H range in $[-1, 1]$, we obtain

$$\left|\operatorname{Tr}\left(H\left(\bar{\rho} - \frac{I}{2^n}\right)\right)\right| \leq \left\|\bar{\rho} - \frac{I}{2^n}\right\|_{\mathrm{tr}} \leq 2\sqrt{1 - F\left(\bar{\rho}, \frac{I}{2^n}\right)}, \tag{S70}$$

where $\|\cdot\|_{\text{tr}}$ denotes the trace norm and the second inequality is from the *Fuchs–van de Graaf inequalities* [63], i.e., $1-\sqrt{F(\rho,\rho')} \leq \frac{1}{2}\|\rho-\rho'\|_{\text{tr}} \leq \sqrt{1-F(\rho,\rho')}$. By combining the upper bound in Theorem 2 with the fact that $-\log F(\rho,\rho') \leq D_2(\rho,\rho')$ [36], we have

$$F\left(\bar{\rho}, \frac{I}{2^n}\right) \geq \frac{1}{2^{D_2\left(\bar{\rho}\|\frac{I}{2^n}\right)}} \geq \frac{1}{1+(2^n-1)\mathrm{e}^{-D\sigma^2}} \tag{S71}$$

$$\geq \frac{1}{1+\frac{(2^n-1)\epsilon^2}{2^{n+4}}} \tag{S72}$$

$$\geq \frac{1}{1+\frac{2^n\epsilon^2}{16\cdot 2^n}} \tag{S73}$$

$$= \frac{16}{16+\epsilon^2}, \tag{S74}$$

where in Eq. (S72) we use the condition $D \geq \frac{1}{\sigma^2}\left[(n+4)\ln 2 + 2\ln(1/\epsilon)\right]$. By inserting Eq. (S74) into Eq. (S70), we can get

$$\left|\operatorname{Tr}\left(H\left(\bar{\rho}-\frac{I}{2^n}\right)\right)\right| \leq 2\sqrt{1-\frac{16}{16+\epsilon^2}} = \frac{2\epsilon}{\sqrt{16+\epsilon^2}} \leq \frac{\epsilon}{2}. \tag{S75}$$

Bringing Eqs. (S69) and (S75) into Eq. (S67), we complete the proof of Corollary 2.1. □

D Proof of Proposition 4

Proposition 4. *Consider a K-classification task with the data set $\mathcal{D}$ defined in Def. 3. If the encoding depth $D \geq \frac{1}{\sigma^2}\left[(n+4)\ln 2 + 2\ln(1/\epsilon)\right]$ for some $\epsilon \in (0,1)$, then the partial gradient of the loss function defined in Eq. (8) with respect to each parameter θ_i of the employed QNN is bounded as*

$$\left|\frac{\partial L(\boldsymbol{\theta};\mathcal{D})}{\partial \theta_i}\right| \leq K\epsilon \tag{S76}$$

with a probability of at least $1-2\mathrm{e}^{-M\epsilon^2/8}$.

Proof. From the chain rule, we know

$$\frac{\partial L(\boldsymbol{\theta};\mathcal{D})}{\partial \theta_i} = \frac{1}{KM}\sum_{m=1}^{KM}\frac{\partial L^{(m)}}{\partial \theta_i} = \frac{1}{KM}\sum_{m=1}^{KM}\sum_{l=1}^{K}\frac{\partial L^{(m)}}{\partial h_l}\frac{\partial h_l}{\partial \theta_i}. \tag{S77}$$

We first calculate $\frac{\partial L^{(m)}}{\partial h_l}$ as follows

$$\frac{\partial L^{(m)}}{\partial h_l} = \frac{\partial \sum_{k=1}^{K} y_k^{(m)}\left(\ln\sum_{j=1}^{K}\mathrm{e}^{h_j} - h_k\right)}{\partial h_l} = \begin{cases} \sum_{k=1}^{K} y_k^{(m)}\frac{\mathrm{e}^{h_l}}{\sum_{j=1}^{K}\mathrm{e}^{h_j}}, & l \neq k \\ \sum_{k=1}^{K} y_k^{(m)}\left(\frac{\mathrm{e}^{h_l}}{\sum_{j=1}^{K}\mathrm{e}^{h_j}}-1\right), & l = k \end{cases} \tag{S78}$$

Since $\mathrm{e}^{h_l}/\sum_{j=1}^{K}\mathrm{e}^{h_j} \in (0,1)$ and $\boldsymbol{y}^{(m)}$ is one-hot, we can get its upper bound as $|\frac{\partial L^{(m)}}{\partial h_l}| \leq 1$. Next, from the parameter-shift rule [45], we calculate $\frac{\partial h_l}{\partial \theta_i}$ as follows

$$\frac{\partial h_l}{\partial \theta_i} = \frac{1}{2}\left[\operatorname{Tr}\left(H_l U(\boldsymbol{\theta}_{+\frac{\pi}{2}})\rho(\boldsymbol{x}^{(m)})U^\dagger(\boldsymbol{\theta}_{+\frac{\pi}{2}})\right) - \operatorname{Tr}\left(H_l U(\boldsymbol{\theta}_{-\frac{\pi}{2}})\rho(\boldsymbol{x}^{(m)})U^\dagger(\boldsymbol{\theta}_{-\frac{\pi}{2}})\right)\right], \tag{S79}$$

where $\boldsymbol{\theta}_{+\frac{\pi}{2}}$ means adding $\frac{\pi}{2}$ to θ_i and keeping the others unchanged, and $\boldsymbol{\theta}_{-\frac{\pi}{2}}$ is similarly defined. If we define

$$\tilde{H}_l \equiv \frac{1}{2}\left[U^\dagger(\boldsymbol{\theta}_{+\frac{\pi}{2}})H_l U(\boldsymbol{\theta}_{+\frac{\pi}{2}}) - U^\dagger(\boldsymbol{\theta}_{-\frac{\pi}{2}})H_l U(\boldsymbol{\theta}_{-\frac{\pi}{2}})\right], \tag{S80}$$

then together with Eqs. (S77)-(S79), we could bound the gradient as

$$\left|\frac{\partial L\left(\boldsymbol{\theta};\mathcal{D}\right)}{\partial\theta_i}\right| \leq \left|\frac{1}{KM}\sum_{m=1}^{KM}\sum_{l=1}^{K}\frac{\partial h_l}{\partial\theta_i}\right| = \left|\frac{1}{KM}\sum_{m=1}^{KM}\sum_{l=1}^{K}\mathrm{Tr}\left(\tilde{H}_l\rho(\boldsymbol{x}^{(m)})\right)\right| \tag{S81}$$

$$\leq \sum_{l=1}^{K}\left|\frac{1}{KM}\sum_{m=1}^{KM}\mathrm{Tr}\left(\tilde{H}_l\rho(\boldsymbol{x}^{(m)})\right)\right| \tag{S82}$$

$$= \sum_{l=1}^{K}\left|\frac{1}{KM}\sum_{m=1}^{KM}\sum_{k=1}^{K}y_k^{(m)}\,\mathrm{Tr}\left(\tilde{H}_l\rho(\boldsymbol{x}^{(m)})\right)\right| \tag{S83}$$

$$\leq \sum_{l=1}^{K}\frac{1}{K}\sum_{k=1}^{K}\left|\frac{1}{M}\sum_{m=1}^{KM}y_k^{(m)}\,\mathrm{Tr}\left(\tilde{H}_l\rho(\boldsymbol{x}^{(m)})\right)\right| \tag{S84}$$

$$\leq \sum_{l=1}^{K}\frac{1}{K}\sum_{k=1}^{K}\epsilon. \tag{S85}$$

Here, it could be easily verified that the eigenvalues of $\tilde{H}_l$ (defined in Eq. (S80)) range in $[-1,1]$ and $\mathrm{Tr}(\tilde{H}_l)=0$. Then from Corollary 2.1, we could bound Eq. (S84) as Eq. (S85), i.e., for any $\epsilon\in(0,1)$, provided that the encoding depth $D\geq\frac{1}{\sigma^2}\left[(n+4)\ln 2+2\ln\left(1/\epsilon\right)\right]$, we have $|\frac{\partial L(\boldsymbol{\theta};\mathcal{D})}{\partial\theta_i}|\leq K\epsilon$ with a probability of at least $1-2\mathrm{e}^{-M\epsilon^2/8}$. This completes the proof of Proposition 4. □