Beyond MNIST: Limitations of Amplitude Encoding on Quantum Classification

Xin Wang
Department of Automation
Tsinghua University
Beijing, 100084, China

Yabo Wang*
School of Mathematics
Southwest Jiaotong University
Chengdu, 611756, China

Rebing Wu*
Department of Automation
Tsinghua University
Beijing, 100084, China

Abstract

It remains unclear whether quantum machine learning (QML) truly holds an advantage when tackling practical and meaningful tasks. In exploring this question, the importance of classical data encoding, often overlooked, becomes increasingly evident. Amplitude encoding, which can embed 2^n classical data into n qubits, is widely used due to its apparent efficiency. However, its potential limitations for QML have yet to be fully explored. In this paper, we investigate a standard linear non-separability limitation on the normalized quadratic features induced by amplitude encoding. The quantum model output is a linear function of these features. We establish a criterion based on the proximity of class-conditional expected encoded states that bounds the expected absolute prediction error independently of the training process. Our findings shed new light on a long-standing puzzle in the field of QML: why some QML models perform well on simple datasets like MNIST but fail to generalize to more complex practical tasks. By highlighting the pivotal role of encoding design in QML, our work clearly indicates that future research needs to focus more on the design of classical data encoding to advance the effectiveness of QML.

1 Introduction

QML has long been anticipated as a promising paradigm that could potentially surpass classical machine learning, thanks to its inherent advantages in state representation, quantum parallelism, and entanglement [1–4]. Classification is a fundamental task in machine learning [5–8] and quantum classifiers [9–13] have demonstrated advantages on toy classical datasets [9, 14, 15, 11, 10, 16, 17, 13] and carefully designed datasets [18, 12]. However, on practical classification tasks, QML performance remains far from competitive. For instance, while state-of-the-art classical models achieve 99.5% accuracy on CIFAR-10 [19, 20], quantum classifiers without classical assistance have not exceeded 80%, with many results below 60% [21–24]. These gaps raise critical questions about the practical effectiveness of quantum classifiers and underscore the need for deeper investigation into their underlying limitations.

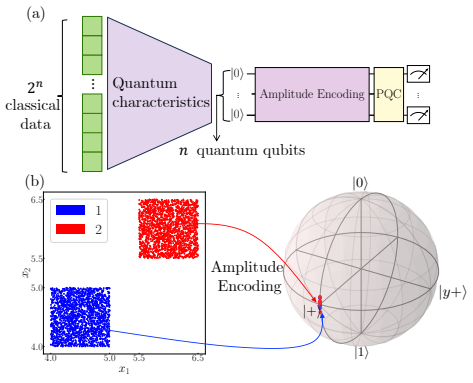

Figure 1: (a) Amplitude encoding embeds 2^n classical values into n qubits. (b) This can cause classically distinct data to become hard to distinguish as quantum states.

*Corresponding authors.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

Previous studies have attributed the limitations of QML to the trainability of parameterized quantum circuits (PQCs) [25], focusing on the barren-plateau phenomenon [26] and spurious local minima [27]. Despite extensive research into their origins [28–31] and proposed remedies, such as over-parameterization [32], and barren-plateau-free ansätze [33–35], QML limitations persist. We argue that a more severe bottleneck lies in the quantum encoding itself, an indispensable step for applying QML to classical tasks [36–38] that remains comparatively underexplored.

The two predominant encoding schemes are angle encoding [39, 14] and amplitude encoding [40]. Angle encoding has been shown to suffer from intrinsic limitations [41–43]; without data re-uploading it can scarcely learn MNIST [41]. In contrast, amplitude encoding exploits quantum superposition to embed 2^n classical amplitudes into n qubits, and is widely regarded as the most "quantum" encoding without reported limitations. In this work, however, we investigate a standard linear non-separability limitation on the normalized quadratic features induced by amplitude encoding.

1.1 Related Works

Limitations of Quantum Encoding. All previously reported limitations of quantum encoding concern angle encoding. When no parametrized gates are inserted between the encoding gates and each element of the encoded data is independent, the expected encoded state converges exponentially fast to the maximally mixed state as the encoding depth grows, causing the gradients of the PQC's cross-entropy loss with respect to its parameters to become small [41]. This conclusion was later extended to the data re-uploading scenario: even with interleaved parametrized layers, the expected encoded state still approaches the maximally mixed state, leaving predictions close to random guessing regardless of training effort [43]. For kernel methods that rely on angle encoding, once the encoding circuit approaches the Haar measure the quantum kernel collapses to a fixed, data-independent form, crippling generalization to unseen data [42].

Preparation of Amplitude Encoding. Due to the importance of amplitude encoding, unlike angle encoding, preparing amplitude-encoded states is nontrivial, prompting extensive research on both exact preparation [44–50] and PQC-based approximation [51, 17]. Notably, although the PQCs for approximating amplitude encoding [17] resemble those used to study angle encoding limitations [41], one cannot infer that amplitude encoding shares similar limitations with angle encoding. This is because the trained PQC parameters are not independent, and correlated data can avoid angle encoding limitations [43]. Moreover, amplitude encoding limitations stem from its intrinsic structure, independent of implementation.

Applications of Amplitude Encoding. QML based on amplitude encoding has demonstrated good performance on MNIST [52, 53, 17, 54, 13, 55–57] (a dataset that cannot be well classified by angle encoding [41]), FashionMNIST [58], and other toy datasets [9], making the potential limitations of amplitude encoding on QML performance more likely to be overlooked.

1.2 Contributions

We show, as illustrated in Fig. 1, that although amplitude encoding represents 2^n classical data elements using n qubits, originally classically separable data need not remain linearly separable in the induced normalized quadratic feature space.

We clarify two common misconceptions regarding the normalized quadratic-feature limitation of amplitude encoding: (1) This limitation does not require the concentration assumptions used in prior angle-encoding analyses [41, 43, 42]. The relevant condition is that class-conditional expected encoded states are identical or close; their common reference state need not be maximally mixed. (2) It is not a quantum rephrasing of the fact that ℓ_2 normalization discards magnitude information. After ℓ_2 normalization, overlapping points do not render classes completely inseparable, and accuracy remains far above random guessing. Yet linear separability may be lost in normalized quadratic feature space, even on datasets that remain linearly separable after ℓ_2 normalization.

We present a criterion for limitations of linear classification in the encoded feature space. Regarding the criteria, prior works [41, 43, 42] attribute the limitation to concentration around the maximally mixed state, a state without information. We show, instead, that the limitation arises whenever class-conditional expected encoded states are sufficiently close, without requiring a particular common reference state. Regarding the limitation itself, our result is independent of qubit number, encoding

scheme, or PQC choice. This fundamentally differs from prior training-difficulty results tied to specific loss functions [41, 42], and from barren plateaus and local minima that can be mitigated by ansatz design. For angle encoding, new encoding circuits have been proposed as a remedy [59]. In contrast, the limitation of amplitude encoding is unrelated to circuit design, and no effective mitigation has been found to date.

Our findings shed new light on a long-standing puzzle in the QML field: why some QML models perform well on simple datasets like MNIST, where data is sparse and targets are clearly separated from backgrounds without interference, but fail to meet expectations on complex real-world datasets where data is dense and targets cannot be easily distinguished from backgrounds. This suggests that to effectively apply QML to real-world problems, the encoding of classical data must be given greater attention and the encoding scheme needs to be carefully designed.

2 Preliminaries and Framework

2.1 Amplitude Encoding

Without loss of generality, suppose that the classical data feature $\boldsymbol{x} \in \mathbb{R}^{2^n}$. By amplitude encoding, we can embed the feature $\boldsymbol{x} = [x_0, x_1, \ldots, x_{2^n-1}]^\top$ into a quantum state vector $|\phi(\boldsymbol{x})\rangle$ as

$$|\phi(\boldsymbol{x})\rangle = \frac{1}{\|\boldsymbol{x}\|_2} \sum_{i=0}^{2^n-1} x_i |i\rangle , \tag{1}$$

where $\|\boldsymbol{x}\|_2 = \sqrt{x_0^2 + x_1^2 + \cdots + x_{2^n-1}^2}$ is the ℓ_2 norm of $\boldsymbol{x}$. An equivalent representation in terms of the density matrix $\rho(\boldsymbol{x})$ reads

$$\rho(\boldsymbol{x}) = |\phi(\boldsymbol{x})\rangle \langle\phi(\boldsymbol{x})| = \frac{1}{\|\boldsymbol{x}\|_2^2} \sum_{i=0}^{2^n-1} \sum_{j=0}^{2^n-1} x_i x_j |i\rangle \langle j| . \tag{2}$$

2.2 Quantum Classifier

To illustrate its limitations, we consider the simplest case of binary classification. Denote the feature space by $\mathcal{X} = \mathbb{R}^{2^n}$ and the label space by $\mathcal{Y} = \{0, 1\}$. Given a dataset $S = \left\{ \left(\boldsymbol{x}^{(m)}, y^{(m)}\right) \right\}_{m=1}^{M}$ containing M independent and identically distributed (i.i.d.) samples, where each sample $\left(\boldsymbol{x}^{(m)}, y^{(m)}\right) \in \mathcal{X} \times \mathcal{Y}$ obeys an unknown joint distribution $\mathcal{D}$, a quantum classifier learns from the dataset S to obtain a hypothesis h_S. Our aim is to enable the hypothesis h_S to accurately classify new, unseen samples that follow the same distribution $\mathcal{D}$.

To be specific, we first embed the feature $\boldsymbol{x}^{(m)}$ into a feature quantum state $\rho\left(\boldsymbol{x}^{(m)}\right)$, and then train a quantum classifier using a PQC $U(\boldsymbol{\theta})$. For binary classification, we select $|0\rangle\langle 0|$ and $|1\rangle\langle 1|$ in the first qubit as the observables for class 0 and 1, respectively, where both observables have eigenvalues in $\{0, 1\}$. The output associated with the class k ($k \in \{0, 1\}$) is defined as

$$h_k\left(\boldsymbol{x}^{(m)}, \boldsymbol{\theta}\right) = \mathrm{Tr}\left[|k\rangle\langle k| U(\boldsymbol{\theta}) \rho(\boldsymbol{x}^{(m)}) U^\dagger(\boldsymbol{\theta}) \right] .$$

We train the parameters $\boldsymbol{\theta}$ of the quantum classifier on dataset $S = \{(\boldsymbol{x}^{(m)}, y^{(m)})\}_{m=1}^{M}$ using the cross-entropy loss:

$$\mathcal{L}(\boldsymbol{\theta}; S) = -\frac{1}{M} \sum_{m=1}^{M} \left[y^{(m)} \log(h_1(\boldsymbol{x}^{(m)}, \boldsymbol{\theta})) + (1 - y^{(m)}) \log(h_0(\boldsymbol{x}^{(m)}, \boldsymbol{\theta})) \right]. \tag{3}$$

Since we choose the observables $|0\rangle\langle 0|$ and $|1\rangle\langle 1|$, for any $\boldsymbol{x}^{(m)}$ it holds that $h_0(\boldsymbol{x}^{(m)}, \boldsymbol{\theta}) + h_1(\boldsymbol{x}^{(m)}, \boldsymbol{\theta}) = 1$, and $h_k(\boldsymbol{x}^{(m)}, \boldsymbol{\theta}) \in [0, 1]$ for $k \in \{0, 1\}$. Specifically, considering only $h_1(\boldsymbol{x}^{(m)}, \boldsymbol{\theta})$, if it is greater than 0.5, we classify $\boldsymbol{x}^{(m)}$ as belonging to class 1; otherwise, if it is less than or equal to 0.5, we classify it as class 0. Therefore, we define the output of the learned hypothesis h_S as follows: for a feature $\boldsymbol{x}^{(m)}$, $h_S(\boldsymbol{x}^{(m)}) = h_1(\boldsymbol{x}^{(m)}, \boldsymbol{\theta}^*)$, where $\boldsymbol{\theta}^*$ denotes the optimized parameters.

To evaluate the performance of a quantum classifier on dataset S, we define the training error of the hypothesis h_S as

$$\widehat{R}_S(h_S) = \frac{1}{M}\sum_{m=1}^{M}\left|y^{(m)} - h_S(\boldsymbol{x}^{(m)})\right|.$$

This error indicates that for data belonging to class 1, we expect $h_1(\boldsymbol{x}^{(m)}, \boldsymbol{\theta}^*)$ to be close to 1, while for data belonging to class 0, we expect $h_1(\boldsymbol{x}^{(m)}, \boldsymbol{\theta}^*)$ to be close to 0 (i.e., $h_0(\boldsymbol{x}^{(m)}, \boldsymbol{\theta}^*)$ to be close to 1). Generally, a smaller value of the loss function corresponds to a lower training error.

Recall that the ultimate goal of the hypothesis h_S is to classify new samples drawn from the distribution $\mathcal{D}$. We define its prediction error $R(h_S)$ as

$$R(h_S) = \mathop{\mathbb{E}}_{(\boldsymbol{x},y)\sim\mathcal{D}}\left[|y - h_S(\boldsymbol{x})|\right].$$

Since the true labels of new samples and the distribution $\mathcal{D}$ are unknown, the prediction error $R(h_S)$ cannot be obtained directly. Therefore, we usually evaluate the test error on a test set that is drawn from the same distribution as the training data but consists of different samples, and use this as a proxy for the prediction error.

3 Main Results

Subsec. 3.1 presents the criterion for linear classification limitations in the encoded feature space. Subsec. 3.2 provides three examples of nearby or identical class-conditional mean states, involving superposition states, the maximally mixed state, or other states. Subsec. 3.3 uses Bloch sphere visualization to highlight the fundamental difference from angle encoding. Subsec. 3.4 clarifies that the limitation is not due to ℓ_2 normalization. Finally, Subsec. 3.5 examines class-conditional feature-mean separation on real-world datasets.

3.1 Criteria for Linear Classification Limitations

We present a criterion for limitations of linear classification in the encoded feature space in the following theorem:

Theorem 3.1. *Suppose the data distributions of class 0 and class 1 are denoted by $\mathcal{D}_{\mathcal{X}_0}$ and $\mathcal{D}_{\mathcal{X}_1}$, respectively, and the overall data distribution $\mathcal{D}_{\mathcal{X}}$ is a uniform mixture of these two classes. If the expectations of the encoded quantum states under $\mathcal{D}_{\mathcal{X}_0}$ and $\mathcal{D}_{\mathcal{X}_1}$ are close to each other, i.e.,*

$$T\left(\mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{D}_{\mathcal{X}_0}}[\rho(\boldsymbol{x})], \mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{D}_{\mathcal{X}_1}}[\rho(\boldsymbol{x})]\right) \leqslant \epsilon, \tag{4}$$

where $T(\rho_1,\rho_2) = 1/2\|\rho_1 - \rho_2\|_1$ is trace distance and $\|\cdot\|_1$ is Schatten-1 norm, then for any training set S sampled from the overall distribution $\mathcal{D}_{\mathcal{X}}$, the hypothesis h_S learned on S will yield predictions that tend toward random guessing, that is, $R(h_S) \geqslant \frac{1}{2} - \epsilon$.

The detailed proof of Theorem 3.1 is provided in App. A. Notably, the simplicity of our proofs reflects the strength of our criterion, which cuts through the complexity that characterized prior analyses [41] and yet yields strictly more general condition and stronger conclusion

Compared to prior work on angle encoding concentration [41], our theorem enjoys (1) broader conditions: prior work only considers concentration around the maximally mixed state, a state carrying no information, which may suggest that the limitation stems from this absence of information. We show that the limitation persists whenever the class-conditional expected encoded states are close, even if their common reference state is informative; and (2) a stronger conclusion: Ref. [41] showed, via Hoeffding's inequality, that the gradient of the cross-entropy loss becomes very small at initialization when the average encoded state approaches the maximally mixed state. However, this argument only applies at the first step, since parameters become data-dependent once training begins. In contrast, our theorem shows that regardless of the loss function, optimizer, or PQC architecture, as long as the class-conditional expected encoded states are sufficiently close, test performance inevitably approaches random guessing, even if the model fits the training set well.

3.2 Illustrative Examples of Normalized Quadratic-Feature Limitations

Below, we use three synthetic datasets to intuitively illustrate the normalized quadratic-feature limitations of amplitude encoding.

Proximity to the Superposition State.

Proposition 3.2. *Assume that all elements in the feature $\boldsymbol{x} \in \mathbb{R}^{2^n}$ have the same sign, and the elements satisfy $|x_i| \in [m, M]$. If $\left|\frac{m}{M} - 1\right| < \epsilon$, then after amplitude encoding, we have*

$$T\left(\rho(\boldsymbol{x}), \frac{1}{2^n}\mathbf{1}\mathbf{1}^\top\right) \leqslant \sqrt{2\epsilon},$$

where the state $\frac{1}{2^n}\mathbf{1}\mathbf{1}^\top = |+\rangle_n \langle +|_n$, with $|+\rangle_n = H^{\otimes n}|0\rangle_n$ the superposition of all computational basis states.

It should be emphasized that this proposition refers to the situation where samples within a single class are clustered, while samples from different classes can still be well separated. It does not mean that samples from the two classes are clustered together. In fact, the goal of classical machine learning algorithms such as linear discriminant analysis (LDA) is to maximize the between-class variance while minimizing the within-class variance [60]. Moreover, recent studies have shown that deep learning exhibits the "law of data separation": in well-trained deep neural networks, each layer progressively increases the between-class variance and decreases the within-class variance [61], ultimately resulting in "Neural Collapse," where the within-class covariance approaches zero [62].

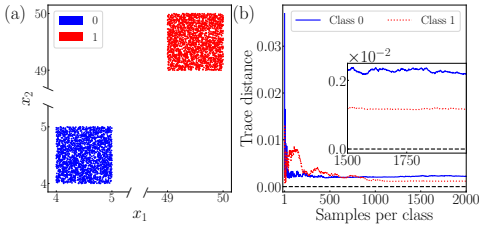

Figure 2: (a) For class 0, features x_1 and x_2 follow $\mathcal{U}[4, 5]$; for class 1, x_1 and x_2 follow $\mathcal{U}[49, 50]$. Each class has 2000 samples. (b) As the sample size increases, the trace distance between the average encoded state of each class and the superposition state $\frac{1}{2}\mathbf{1}^\top\mathbf{1}$ decreases.

As illustrated in Fig. 2(a), consider two classes that are clearly distinguishable. It is straightforward to verify that the features of each class satisfy the conditions in Proposition 3.2. As shown in Fig. 2(b), when the per-class sample size increases (with balanced classes maintained), the trace distance between the average encoded state of each class and the superposition state gradually decreases toward zero. By the triangle inequality, this implies that the trace distance between the two class-conditional average encoded states also approaches zero.

Identical Mean States at the Maximally Mixed State.

Proposition 3.3. *Denote by $\mathcal{D}_\mathcal{X}$ the distribution of feature $\boldsymbol{x} \in \mathbb{R}^{2^n}$. Assume that the elements of $\boldsymbol{x}$ are i.i.d. with an expected value of 0. In addition, the distribution is symmetric, i.e., the probability density function satisfies $p(x_j) = p(-x_j)$ for all $j \in [0 : 2^n - 1]$. Then after amplitude encoding, the expectation of encoded states is the maximally mixed state, i.e.,*

$$\mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{D}_\mathcal{X}} [\rho(\boldsymbol{x})] = \frac{\boldsymbol{I}}{2^n}.$$

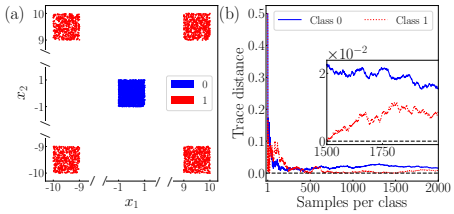

Figure 3: (a) For class 0, x_1 and x_2 follow $\mathcal{U}[-1, 1]$; for class 1, x_1 and x_2 follow the uniform distribution on $[-10, -9] \cup [9, 10]$. Each class has 2000 samples. (b) As the sample size increases, the trace distance between the average encoded state of each class and the maximally mixed state $\frac{1}{2}\boldsymbol{I}$ decreases.

Note that when the elements of the feature vector $\boldsymbol{x}$ are i.i.d. and symmetric about their mean, we can employ the standard z-score normalization [63] to ensure that the distribution satisfies the assumptions in Proposition 3.3. Specifically, this involves transforming each element as $z = \frac{x-\mu}{\sigma}$, where μ is the mean of x and σ is the standard deviation. z-score normalization is a commonly used normalization technique.

We illustrate Proposition 3.3 by using a binary classification dataset in Fig. 3(a). It is clear that the features of class 0 and class 1 can be distinguished by their distances to the origin. However, after

amplitude encoding, as the per-class sample size increases, the trace distance between the average encoded state of each class and the maximally mixed state gradually decreases toward zero. By the triangle inequality, their mutual trace distance also approaches zero.

Identical Mean States Beyond the Maximally Mixed State.

Proposition 3.4. *For a binary classification dataset, denote by $\mathcal{D}_{\mathcal{X}_0}$ and $\mathcal{D}_{\mathcal{X}_1}$ the distributions of feature $\boldsymbol{x} \in \mathbb{R}^{2^n}$ in class 0 and class 1, respectively. Assume the data are i.i.d. If for any $j \in [0 : 2^n - 1]$, the probability density functions $p_{0,j}(x)$ and $p_{1,j}(x)$ of the j-th element in class 0 and class 1 satisfy $p_{0,j}(x) = p_{1,j}(-x)$, then after amplitude encoding, the expected states of the two classes are identical, i.e.,*

$$\mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{D}_{\mathcal{X}_0}} [\rho(\boldsymbol{x})] = \mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{D}_{\mathcal{X}_1}} [\rho(\boldsymbol{x})] .$$

Proposition 3.4 highlights that the relevant condition is equality or proximity of the class-conditional expected encoded states, without requiring a particular common reference state.

We exemplify Proposition 3.4 in Fig. 4 by showing an example where the features of class 0 and class 1 are linearly separable, yet the trace distance between their average encoded states converges to zero as the sample size increases.

These examples extend to any number of qubits for which the stated assumptions hold. The detailed proofs of the propositions are provided in App. A.

These examples show that the relevant class-conditional mean states can be close to the superposition state, equal to the maximally mixed state, or coincide at another state.

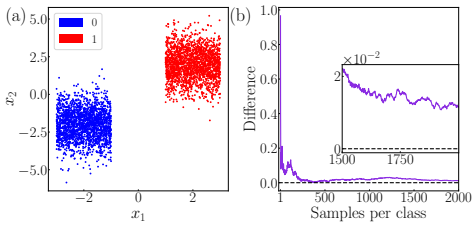

Figure 4: (a) For class 0, x_1 follows $\mathcal{U}[-3,-1]$ and x_2 follows $\mathcal{N}(-2,1)$; for class 1, x_1 follows $\mathcal{U}[1,3]$ and x_2 follows $\mathcal{N}(2,1)$. Each class has 2000 samples. (b) As the sample size increases, the trace distance between the average encoded states of the two classes decreases.

3.3 Distinction between Amplitude Encoding and Angle Encoding

Fig. 5 visualizes the distributions of encoded states on the Bloch sphere for the amplitude-encoding examples and an angle-encoding concentration example. Following Ref. [41], we use a single qubit angle encoding circuit composed of 6 layers of R_z–R_y–R_z gates to reproduce the concentration phenomenon.

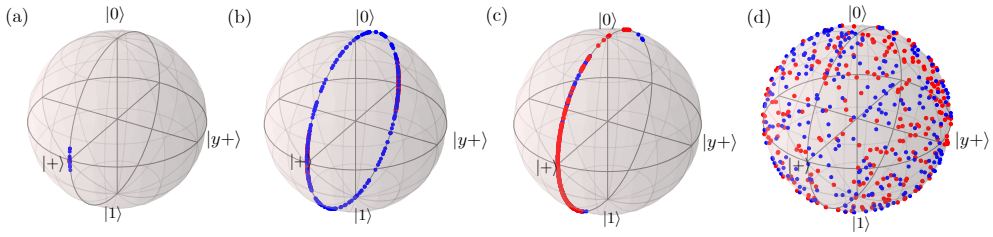

Figure 5: (a)–(c) Quantum states via amplitude encoding for datasets in Fig. 2(a), 3(a), and 4(a). (d) Quantum states via angle encoding with each element in class 0 sampled from $\mathcal{N}(0,1)$ (blue) and each element in class 1 sampled from $\mathcal{N}(\pi,1)$ (red).

We observe that only in the case of angle encoding do the generated quantum states fully occupy the Bloch sphere. In contrast, the quantum states produced by amplitude encoding do not span the entire Bloch sphere. For example, as shown in Proposition 3.2, the quantum states generated by amplitude encoding are localized near the $|+\rangle$ state and this effect is not caused by high expressivity. This figure illustrates that nearby class-conditional mean states can arise under different geometries of encoded states.

3.4 Distinction between Amplitude Encoding and the ℓ_2 Norm

From the definition of amplitude encoding in Eq. (1), the mapping $\boldsymbol{x} \mapsto |\phi(\boldsymbol{x})\rangle$ appears merely to apply ℓ_2 normalization, i.e., to erase the magnitude of $\boldsymbol{x}$, since $|\phi(\boldsymbol{x})\rangle$ must lie on the unit sphere. A natural misconception is that the normalized quadratic-feature limitation stems solely from discarding this magnitude information. Fig. 6 shows the three illustrative examples from Subsec. 3.1 after ℓ_2 normalization. Although some points overlap in Fig. 6(a) and (b), the two classes are by no means completely inseparable. Such scenarios, where a subset of points have non-unique labels, known as stochastic scenarios, are in fact more representative of real-world problems [64]. A small fraction of overlapping points does not imply that the two classes are entirely indistinguishable. Moreover, Fig. 6(c) remains linearly separable.

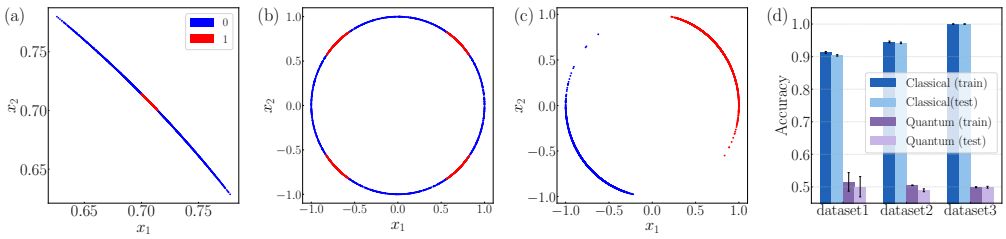

Figure 6: (a)–(c) The datasets from Fig. 2(a), 3(a), and 4(a) after ℓ_2 normalization. (d) "Classical": a two-layer MLP (hidden size 32) on ℓ_2-normalized data; "Quantum": amplitude encoding with the circuit in Fig. 8.

We further compare two approaches on the three datasets in Fig. 6: (1) an MLP with two hidden layers of size 32 trained on ℓ_2-normalized data, and (2) amplitude encoding with the quantum model in Fig. 8. Both use Adam with learning rate 0.05, and each dataset is split into 1000 training and 1000 test points with balanced classes. As shown in Fig. 6(d), the classical MLP achieves over 90% accuracy, while amplitude encoding yields only around 0.5 (random guessing). This confirms that the limitation of amplitude encoding is not due to ℓ_2 normalization.

3.5 Normalized Quadratic-Feature Mean Separation on Real-World Data

To examine class-conditional normalized quadratic-feature mean separation in real-world datasets, we conduct experiments on four image classification datasets: EuroSAT [65] (forest vs. sea/lakes), PathMNIST [66] (adipose tissue vs. background), CIFAR-10 [19] (airplane vs. bird), and MNIST [67] (digit 3 vs. 6). All images are first converted to grayscale and then resized to 32×32 pixels, resulting in 1024-dimensional feature vectors that are encoded using 10 qubits. We compute the trace distance between the two classes, with results shown in Fig. 7.

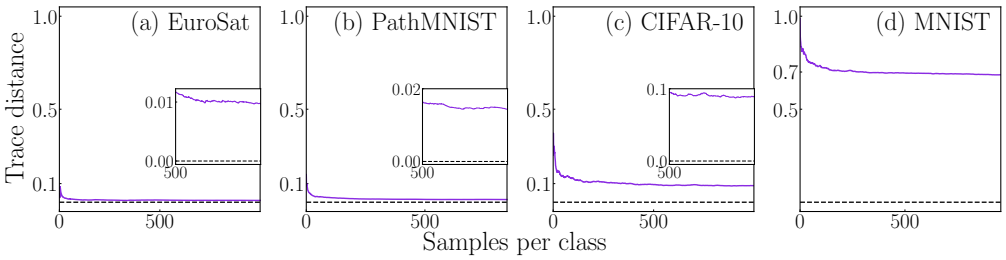

Figure 7: Trace distance between the average encoded states of the two classes on different datasets: (a) EuroSAT, (b) PathMNIST, (c) CIFAR-10, (d) MNIST.

As shown in Fig. 7, as the sample size increases, the trace distance between the average encoded states of the two classes drops below 0.01 on the test sets of EuroSAT and PathMNIST, and below 0.1 on CIFAR10. This indicates that after amplitude encoding, any QML model would struggle to achieve good prediction performance on these datasets. In contrast, on the MNIST dataset, the trace distance between the two classes consistently remains above 0.7, ensuring that favorable prediction results can be attained.

4 Experiments

4.1 Synthetic Datasets

We first use the datasets in Fig. 2 to study the test error under the normalized quadratic-feature limitation of amplitude encoding. The test error remains unaffected by variations in either the samples per class in the training set or the parameter count of the PQC.

In our experiments, we employ the single-qubit circuit shown in Fig. 8, which consists of L layers, each composed of three rotation gates: R_z, R_x, and R_z. All parameters are initialized from a standard Gaussian distribution. Training is performed by minimizing the cross-entropy loss (Eq. (3)) using the Adam optimizer with a learning rate of 0.001 for a total of 100 epochs. The test set consists of 1000 samples per class and remains consistent throughout all experimental configurations. We conduct two sets of complementary experiments: (i) fixing the number of PQC layers to 20 while varying the number of samples per class in $\{1, 20, 200, 2000\}$, and (ii) fixing the number of samples per class to 20 while varying the number of PQC layers in $\{2, 20, 200, 2000\}$. Each configuration is repeated five times. All experiments are run on a CPU platform (Intel Core i9-13900KF, 64GB RAM).

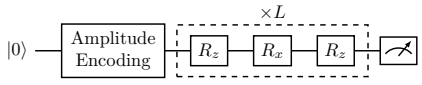

Figure 8: Single-qubit PQC used in synthetic experiments, consisting of L layers of R_z–R_x–R_z gates.

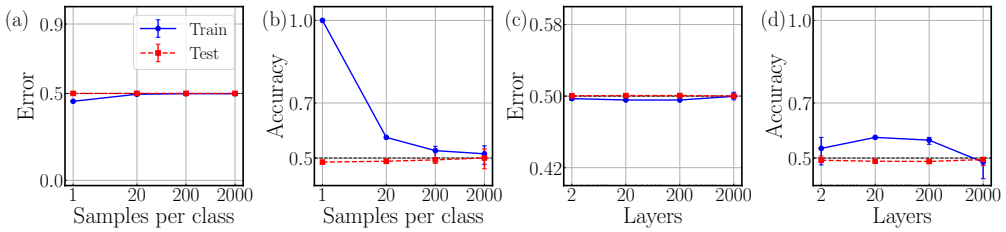

Figure 9: (a) Train/test error vs. training samples (20 layers); (b) corresponding accuracy for (a); (c) train/test error vs. circuit layers (20 samples per class); (d) corresponding accuracy for (c). Solid lines indicate the mean over five runs; error bars span the minimum and maximum values.

The results, shown in Fig. 9, reveal that the test error stays close to 0.5 no matter how many training samples or circuit layers are used. The corresponding accuracy curves confirm this behavior: a test error near 0.5 yields an accuracy that fluctuates around 0.5, indistinguishable from random guessing. As the dataset grows while the prediction error remains stuck at 0.5, the generalization gap shrinks and the training error is also pulled toward 0.5. The underlying normalized quadratic feature map remains unchanged as the number of training samples increases.

4.2 Real-World Datasets

On real-world datasets we selected EuroSAT and MNIST, introduced in Subsec. 3.5 as representatives of, respectively, small and large separation between class-conditional normalized quadratic-feature means, and examined their resulting test errors.

To avoid barren plateaus, we employed a 10-qubit QCNN circuit (App. D), which is immune to this issue [33]. Each training set contains 2000 samples per class; the EuroSAT test set has 1000 samples per class, while MNIST uses its original test set (1010 samples for digit 3 and 958 samples for digit 6). We minimized the cross-entropy loss (Eq. (3)) with Adam (learning rate 0.001), initialized parameters from a standard Gaussian, trained for 100 epochs, and repeated each experiment 5 times. All experiments are run on a CPU platform (Intel Core i9-13900KF, 64GB RAM).

The results are displayed in Fig. 10. On EuroSAT, where the class-conditional mean states are close, both training and test errors hover around 0.5 while the corresponding accuracy oscillates sharply between 0.5 and 0.6. In contrast, on MNIST, where their separation is larger, the training and test errors rapidly drop below 0.5 as epochs progress, and the accuracies quickly climb above 0.9. These MNIST results also corroborate the observation that amplitude encoding can learn tasks that are hard to train under angle encoding [41].

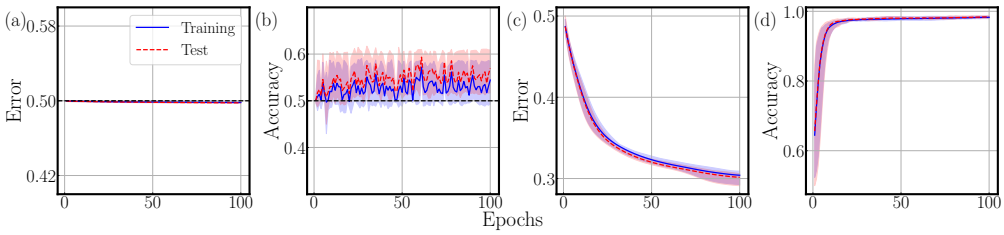

Figure 10: (a)–(b) Training and test errors together with the corresponding accuracies on the EuroSAT dataset; (c)–(d) training and test errors together with the corresponding accuracies on the MNIST dataset. Solid lines indicate the mean over five runs; shaded regions span the minimum and maximum values.

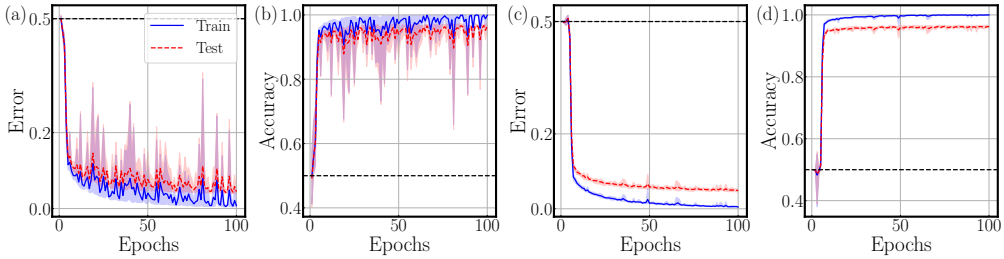

Figure 11: Training and test errors together with the corresponding accuracies on EuroSAT using ResNet. (a)–(b) with preprocessing (c)–(d) with preprocessing and ℓ_2 normalization. Solid lines indicate the mean over five runs; shaded regions span the minimum and maximum values.

To verify that the poor QML performance on EuroSAT is not due to our preprocessing, we trained a ResNet-18 on the same data, adapting its first layer for single-channel input and its output layer for binary classification. Using SGD with cross-entropy loss (learning rate 0.001) for 100 epochs over five runs, both training and test errors converged to zero while accuracies approached 1.0 (Fig. 11.(a)(b)). This confirms that the data retains sufficient information; the near-random-guessing behavior stems from amplitude encoding itself.

Moreover, training ResNet-18 on ℓ_2-normalized EuroSAT with identical settings yields not worse but more stable results (see Fig. 11(c)(d)), showing that ℓ_2 normalization brings a benefit rather than a drawback. This further confirms that the limitation of amplitude encoding does not arise from ℓ_2 normalization. All classical experiments are run on a single GPU RTX 4090.

5 Discussion

Since the typical use case of amplitude encoding is to harness qubit entanglement and superposition for high-dimensional data intractable to classical computers, and complex preprocessing such as PCA or cross-sample interactions is often infeasible in such regimes, we do not pursue a dedicated discussion of preprocessing in this work. Moreover, this work aims to advance the end-to-end quantum machine learning paradigm and thus does not consider hybrid classical-quantum architectures. Nonetheless, even in hybrid settings, the criteria established in this paper can serve as a diagnostic tool to assess whether the quantum component becomes a performance bottleneck.

Acknowledgments and Disclosure of Funding

The authors thank the reviewers for their constructive feedback. The work of Xin Wang and Rebing Wu was supported by the Innovation Program for Quantum Science and Technology (No. 2021ZD0300201), the Tsinghua-Toyota Joint Research Fund, and the Fundamental and Interdisciplinary Disciplines Breakthrough Plan of the Ministry of Education of China. Yabo Wang was supported in part by the Fundamental Research Funds for the Central Universities under Grant No. 2682026CX133.

References

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

[2] Tailong Xiao, Xinliang Zhai, Xiaoyan Wu, Jianping Fan, and Guihua Zeng. Practical advantage of quantum machine learning in ghost imaging. *Communications Physics*, 6:171, 2023.

[3] Pei-Lin Zheng, Jia-Bao Wang, and Yi Zhang. Efficient and quantum-adaptive machine learning with fermion neural networks. *Physical Review Applied*, 20:044002, 2023.

[4] Pierre Azam and Robin Kaiser. Optically accelerated extreme learning machine using hot atomic vapors. *Physical Review Applied*, 22:034041, 2024.

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

[6] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet classification with deep convolutional neural networks. In *Proceedings of the 26th Conference on Neural Information Processing Systems (NeurIPS 2012)*, Red Hook, NY, USA, 2012. Curran Associates, Inc.

[7] Vinod Raman and Ambuj Tewari. Online classification with predictions. In *Proceedings of the 38th Conference on Neural Information Processing Systems (NeurIPS 2024)*, Red Hook, NY, USA, 2024. Curran Associates, Inc.

[8] Ming-Ming Wang and Xiao-Ying Zhang. Quantum Bayes classifiers and their application in image classification. *Physical Review A*, 110:012433, 2024.

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

[10] 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*, 12:2631, 2021.

[11] Thomas Hubregtsen, David Wierichs, Elies Gil-Fuster, Peter-Jan H. S. Derks, Paul K. Faehrmann, and Johannes Jakob Meyer. Training quantum embedding kernels on near-term quantum computers. *Physical Review A*, 106:042431, 2022.

[12] Sofiene Jerbi, Lukas J. Fiderer, Hendrik Poulsen Nautrup, Jonas M. Kübler, Hans J. Briegel, and Vedran Dunjko. Quantum machine learning beyond kernel methods. *Nature Communications*, 14:517, 2023.

[13] Yabo Wang, Xin Wang, Bo Qi, and Daoyi Dong. Supervised-learning guarantee for quantum AdaBoost. *Physical Review Applied*, 22:054001, 2024.

[14] 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.

[15] Maxwell Henderson, Samriddhi Shakya, Shashindra Pradhan, and Tristan Cook. Quanvolutional neural networks: powering image recognition with quantum circuits. *Quantum Machine Intelligence*, 2:2, 2020.

[16] Li Xu, Xiao-yu Zhang, Ming Li, and Shu-qian Shen. Quantum classifiers with a trainable kernel. *Physical Review Applied*, 21:054056, 2024.

[17] Naoki Mitsuda, Tatsuhiro Ichimura, Kouhei Nakaji, Yohichi Suzuki, Tomoki Tanaka, Rudy Raymond, Hiroyuki Tezuka, Tamiya Onodera, and Naoki Yamamoto. Approximate complex amplitude encoding algorithm and its application to data classification problems. *Physical Review A*, 109:052423, 2024.

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

[19] Alex Krizhevsky. Learning multiple layers of features from tiny images, 2009.

[20] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of the 9th International Conference on Learning Representations (ICLR 2021)*, 2021.

[21] Maxwell T West, Martin Sevior, and Muhammad Usman. Reflection equivariant quantum neural networks for enhanced image classification. *Machine Learning: Science and Technology*, 4(3):035027, 2023.

[22] Arsenii Senokosov, Alexandr Sedykh, Asel Sagingalieva, Basil Kyriacou, and Alexey Melnikov. Quantum machine learning for image classification. *Machine Learning: Science and Technology*, 5(1):015040, 2024.

[23] Sreetama Das, Stefano Martina, and Filippo Caruso. The role of data embedding in equivariant quantum convolutional neural networks. *Quantum Machine Intelligence*, 6(2):82, 2024.

[24] Koustubh Phalak, Junde Li, and Swaroop Ghosh. Dataset distillation for quantum neural networks. *arXiv preprint arXiv:2503.17935*, 2025.

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

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

[27] Xuchen You and Xiaodi Wu. Exponentially many local minima in quantum neural networks. In *Proceedings of the 38th International Conference on Machine Learning (ICML 2021)*, New York, NY, USA, 2021. PMLR.

[28] Carlos Ortiz Marrero, Mária Kieferová, and Nathan Wiebe. Entanglement-induced barren plateaus. *PRX Quantum*, 2:040316, 2021.

[29] Supanut Thanasilp, Samson Wang, Nhat Anh Nghiem, Patrick Coles, and Marco Cerezo. Subtleties in the trainability of quantum machine learning models. *Quantum Machine Intelligence*, 5:21, 2023.

[30] Michael Ragone, Bojko N Bakalov, Frédéric Sauvage, Alexander F Kemper, Carlos Ortiz Marrero, Martín Larocca, and M Cerezo. A Lie algebraic theory of barren plateaus for deep parameterized quantum circuits. *Nature Communications*, 15:7172, 2024.

[31] Eric R. Anschuetz and Bobak T. Kiani. Quantum variational algorithms are swamped with traps. *Nature Communications*, 13:7760, 2022.

[32] Martin Larocca, Nathan Ju, Diego García-Martín, Patrick J. Coles, and Marco Cerezo. Theory of overparametrization in quantum neural networks. *Nature Computational Science*, 3:542–551, 2023.

[33] Arthur Pesah, Marco Cerezo, Samson Wang, Tyler Volkoff, Andrew T Sornborger, and Patrick J Coles. Absence of barren plateaus in quantum convolutional neural networks. *Physical Review X*, 11:041011, 2021.

[34] M. Cerezo, Akira Sone, Tyler Volkoff, Lukasz Cincio, and Patrick J. Coles. Cost function dependent barren plateaus in shallow parametrized quantum circuits. *Nature Communications*, 12:1791, 2021.

[35] Samson Wang, Enrico Fontana, Marco Cerezo, Kunal Sharma, Akira Sone, Lukasz Cincio, and Patrick J Coles. Noise-induced barren plateaus in variational quantum algorithms. *Nature Communications*, 12(1):6961, 2021.

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

[37] Nathan Wiebe. Key questions for the quantum machine learner to ask themselves. *New Journal of Physics*, 22:091001, 2020.

[38] Minati Rath and Hema Date. Quantum data encoding: a comparative analysis of classical-to-quantum mapping techniques and their impact on machine learning accuracy. *EPJ Quantum Technology*, 11:72, 2024.

[39] Maria Schuld and Francesco Petruccione. *Supervised Learning with Quantum Computers*. Springer, Cham, Switzerland, 2018.

[40] 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: 032430, 2021.

[41] Guangxi Li, Ruilin Ye, Xuanqiang Zhao, and Xin Wang. Concentration of data encoding in parameterized quantum circuits. In *Proceedings of the 35th Conference on Neural Information Processing Systems (NeurIPS 2022)*, Red Hook, NY, USA, 2022. Curran Associates, Inc.

[42] Supanut Thanasilp, Samson Wang, M Cerezo, and Zoë Holmes. Exponential concentration in quantum kernel methods. *Nature Communications*, 15:5200, 2024.

[43] Xin Wang, Hanxiao Tao, and Rebing Wu. Predictive performance of deep quantum data re-uploading models. In *Proceedings of the 42nd International Conference on Machine Learning (ICML 2024)*, New York, NY, USA, 2024. PMLR.

[44] Vittorio Giovannetti, Seth Lloyd, and Lorenzo Maccone. Quantum random access memory. *Physical Review Letters*, 100:160501, 2008.

[45] Daniel K Park, Francesco Petruccione, and June-Koo Kevin Rhee. Circuit-based quantum random access memory for classical data. *Scientific Reports*, 9:3949, 2019.

[46] N Jiang, Y-F Pu, W Chang, C Li, S Zhang, and L-M Duan. Experimental realization of 105-qubit random access quantum memory. *npj Quantum Information*, 5:28, 2019.

[47] Martin Plesch and Časlav Brukner. Quantum-state preparation with universal gate decompositions. *Physical Review A*, 83:032302, 2011.

[48] Gui-Lu Long and Yang Sun. Efficient scheme for initializing a quantum register with an arbitrary superposed state. *Physical Review A*, 64:014303, 2001.

[49] Israel F Araujo, Daniel K Park, Francesco Petruccione, and Adenilton J da Silva. A divide-and-conquer algorithm for quantum state preparation. *Scientific Reports*, 11:1–12, 2021.

[50] Javier Gonzalez-Conde, Thomas W Watts, Pablo Rodriguez-Grasa, and Mikel Sanz. Efficient quantum amplitude encoding of polynomial functions. *Quantum*, 8:1297, 2024.

[51] Kouhei Nakaji, Shumpei Uno, Yohichi Suzuki, Rudy Raymond, Tamiya Onodera, et al. Approximate amplitude encoding in shallow parameterized quantum circuits and its application to financial market indicators. *Physical Review Research*, 4:023136, 2022.

[52] Edward Grant, Marcello Benedetti, Shuxiang Cao, Andrew Hallam, Joshua Lockhart, Vid Stojevic, Andrew G Green, and Simone Severini. Hierarchical quantum classifiers. *npj Quantum Information*, 4:65, 2018.

[53] Tak Hur, Leeseok Kim, and Daniel K Park. Quantum convolutional neural network for classical data classification. *Quantum Machine Intelligence*, 4:3, 2022.

[54] Qingyu Li, Yuhan Huang, Xiaokai Hou, Ying Li, Xiaoting Wang, and Abolfazl Bayat. Ensemble-learning error mitigation for variational quantum shallow-circuit classifiers. *Physical Review Research*, 6:013027, 2024.

[55] Paolo Alessandro Xavier Tognini, Leonardo Banchi, and Giacomo De Palma. Solving MNIST with a globally trained mixture of quantum experts. *arXiv preprint arXiv:2505.14789*, 2025.

[56] Walid El Maouaki, Nouhaila Innan, Alberto Marchisio, Taoufik Said, Mohamed Bennai, and Muhammad Shafique. QFAL: Quantum federated adversarial learning. *arXiv preprint arXiv:2502.21171*, 2025.

[57] A De Lorenzis, MP Casado, MP Estarellas, N Lo Gullo, T Lux, F Plastina, A Riera, and J Settino. Harnessing quantum extreme learning machines for image classification. *Physical Review Applied*, 23(4):044024, 2025.

[58] Kevin Shen, Bernhard Jobst, Elvira Shishenina, and Frank Pollmann. Classification of the Fashion-MNIST dataset on a quantum computer. *arXiv preprint arXiv:2403.02405*, 2024.

[59] Florian J Kiwit, Bernhard Jobst, Andre Luckow, Frank Pollmann, and Carlos A Riofrío. Typical machine learning datasets as low-depth quantum circuits. *arXiv preprint arXiv:2505.03399*, 2025.

[60] Ronald A Fisher. The use of multiple measurements in taxonomic problems. *Annals of eugenics*, 7(2):179–188, 1936.

[61] Hangfeng He and Weijie J Su. A law of data separation in deep learning. *Proceedings of the National Academy of Sciences*, 120(36):e2221704120, 2023.

[62] 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.

[63] Nanyi Fei, Yizhao Gao, Zhiwu Lu, and Tao Xiang. Z-score normalization, hubness, and few-shot learning. In *Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV 2021)*, Los Alamitos, CA, USA, 2021. IEEE Computer Society.

[64] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of Machine Learning*. The MIT Press, Cambridge, MA, USA, 2nd edition, 2018.

[65] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. EuroSAT: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12:2217–2226, 2019. doi: 10.1109/JSTARS.2019.2918242.

[66] Jiancheng Yang, Rui Shi, Donglai Wei, Zequan Liu, Lin Zhao, Bilian Ke, Hanspeter Pfister, and Bingbing Ni. MedMNIST v2-A large-scale lightweight benchmark for 2D and 3D biomedical image classification. *Scientific Data*, 10:41, 2023. doi: https://doi.org/10.1038/s41597-022-01721-8.

[67] Yann LeCun. The MNIST database of handwritten digits. `http://yann.lecun.com/exdb/mnist/`.

A Proof of Theorem 3.1

Theorem A.1 (Theorem 3.1 in the main text)**.** *Suppose the data distributions of class 0 and class 1 are denoted by $\mathcal{D}_{\mathcal{X}_0}$ and $\mathcal{D}_{\mathcal{X}_1}$, respectively, and the overall data distribution $\mathcal{D}_{\mathcal{X}}$ is a uniform mixture of these two classes. If the expectations of the encoded quantum states under $\mathcal{D}_{\mathcal{X}_0}$ and $\mathcal{D}_{\mathcal{X}_1}$ are close to each other, i.e.,*

$$T\left(\mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{D}_{\mathcal{X}_0}}[\rho(\boldsymbol{x})], \mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{D}_{\mathcal{X}_1}}[\rho(\boldsymbol{x})]\right) \leqslant \epsilon, \tag{5}$$

where $T(\rho_1,\rho_2) = 1/2\|\rho_1-\rho_2\|_1$ is trace distance and $\|\cdot\|_1$ is Schatten-1 norm, then for any training set S sampled from the overall distribution $\mathcal{D}_{\mathcal{X}}$, the hypothesis h_S learned on S will yield predictions that tend toward random guessing, that is,$R(h_S) \geqslant \frac{1}{2}-\epsilon$.

Proof. By definition, we have

$$h_S(\boldsymbol{x}) = \mathrm{Tr}\left[|1\rangle\langle 1| U(\boldsymbol{\theta}^*)\rho(\boldsymbol{x})U^\dagger(\boldsymbol{\theta}^*)\right]$$

where $\boldsymbol{\theta}^*$ denotes the parameters obtained from training.

Next, we have:

$$\begin{aligned}
&\left|\mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{D}_{\mathcal{X}_0}}[h_S(\boldsymbol{x})] - \mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{D}_{\mathcal{X}_1}}[h_S(\boldsymbol{x})]\right| \\
=&\left|\mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{D}_{\mathcal{X}_0}}\left[\mathrm{Tr}\left[|1\rangle\langle 1| U(\boldsymbol{\theta}^*)\rho(\boldsymbol{x})U^\dagger(\boldsymbol{\theta}^*)\right]\right] - \mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{D}_{\mathcal{X}_1}}\left[\mathrm{Tr}\left[|1\rangle\langle 1| U(\boldsymbol{\theta}^*)\rho(\boldsymbol{x})U^\dagger(\boldsymbol{\theta}^*)\right]\right]\right| \\
=&\left|\mathrm{Tr}\left[|1\rangle\langle 1| U(\boldsymbol{\theta}^*)\mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{D}_{\mathcal{X}_0}}[\rho(\boldsymbol{x})]U^\dagger(\boldsymbol{\theta}^*)\right] - \mathrm{Tr}\left[|1\rangle\langle 1| U(\boldsymbol{\theta}^*)\mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{D}_{\mathcal{X}_1}}[\rho(\boldsymbol{x})]U^\dagger(\boldsymbol{\theta}^*)\right]\right| \\
\leqslant&\left\|\mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{D}_{\mathcal{X}_0}}[\rho(\boldsymbol{x})] - \mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{D}_{\mathcal{X}_1}}[\rho(\boldsymbol{x})]\right\|_1 \\
=&2T\left(\mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{D}_{\mathcal{X}_0}}[\rho(\boldsymbol{x})], \mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{D}_{\mathcal{X}_1}}[\rho(\boldsymbol{x})]\right) \leqslant 2\epsilon
\end{aligned}$$

The second equality follows from the linearity of the trace and expectation operators, and the inequality utilizes Hölder's inequality.

Since the distribution $\mathcal{D}_{\mathcal{X}}$ is a uniform mixture of $\mathcal{D}_{\mathcal{X}_0}$ and $\mathcal{D}_{\mathcal{X}_1}$, we have

$$\begin{aligned}
R(h_S) &= \mathop{\mathbb{E}}_{(\boldsymbol{x},y)\sim\mathcal{D}}[|y-h(\boldsymbol{x})|] \\
&= \frac{1}{2}\left(\mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{D}_{\mathcal{X}_1}}[1-h_S(\boldsymbol{x})] + \mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{D}_{\mathcal{X}_0}}[h_S(\boldsymbol{x})]\right) \\
&= \frac{1}{2} + \frac{1}{2}\left(\mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{D}_{\mathcal{X}_0}}[h_S(\boldsymbol{x})] - \mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{D}_{\mathcal{X}_1}}[h_S(\boldsymbol{x})]\right) \geqslant \frac{1}{2}-\epsilon,
\end{aligned}$$

where $\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}_{\mathcal{X}_k}}[h_S(\boldsymbol{x})] \in [0,1]$ for $k\in\{0,1\}$. □

B Proof of Illustrative Example

Proposition B.1 (Proposition 3.2 in the main text)**.** *Assume that all elements in the feature $\boldsymbol{x}\in\mathbb{R}^{2^n}$ have the same sign, and the elements satisfy $|x_i|\in[m,M]$. If $\left|\frac{m}{M}-1\right| < \epsilon$, then after amplitude encoding, we have*

$$T\left(\rho(\boldsymbol{x}), \frac{1}{2^n}\mathbf{1}\mathbf{1}^\top\right) \leqslant \sqrt{2\epsilon},$$

where the state $\frac{1}{2^n}\mathbf{1}\mathbf{1}^\top = |+\rangle_n\langle+|_n$, with $|+\rangle_n = H^{\otimes n}|0\rangle_n$ being the superposition of all computational basis states.

Proof. For the feature vector $\boldsymbol{x} = [x_0, x_1, \cdots, x_{2^n-1}]^\top$, denote the encoded state vector as $|\psi(\boldsymbol{x})\rangle$, with the density matrix as $\rho(\boldsymbol{x})$. Then, the probability amplitude of the encoded state corresponding to the basis vector $|i\rangle$ reads

$$\alpha_i = \sqrt{\frac{x_i^2}{x_0^2 + x_1^2 + \cdots + x_{2^n-1}^2}}.$$

The corresponding probability of projection into the basis vector $|i\rangle$ is

$$p_i = |\alpha_i|^2 = \frac{x_i^2}{x_0^2 + x_1^2 + \cdots + x_{2^n-1}^2} = \frac{1}{1 + (x_0/x_i)^2 + \cdots + (x_{2^n-1}/x_i)^2}.$$

If $|x_i| \in [m, M]$, then

$$p_i = \frac{1}{1 + (x_0/x_i)^2 + \cdots + (x_{2^n-1}/x_i)^2} \geqslant \frac{1}{1 + (2^n - 1)(M/m)^2}.$$

Let $x = \frac{m}{M}$, where $x \in (0, 1]$, and define $f_{\min}(x) = \frac{x^2}{x^2 + (2^n - 1)}$. Thus, we have

$$|2^n f_{\min}(x) - 1| = \left| \frac{2^n x^2 - x^2 - (2^n - 1)}{x^2 + (2^n - 1)} \right| = \left| \frac{(2^n - 1)(x^2 - 1)}{x^2 + (2^n - 1)} \right| \leqslant |x^2 - 1| \leqslant 2\epsilon.$$

Thus, $f_{\min}(\frac{m}{M}) = \frac{1}{1 + (2^n - 1)(M/m)^2} \geqslant \frac{1}{2^n} \cdot (1 - 2\epsilon)$.

The absolute value of the inner product between the encoded state vector $|\psi(\boldsymbol{x})\rangle$ and the n-qubit superposition state $|+\rangle_n$ satisfies

$$\left| \langle \psi(\boldsymbol{x}) | + \rangle_n \right| = \frac{1}{\sqrt{2^n}} \sum_{i=0}^{2^n-1} \sqrt{p_i} \geqslant \sum_{i=0}^{2^n-1} \frac{1}{2^n} \cdot \sqrt{1 - 2\epsilon} \geqslant \sqrt{1 - 2\epsilon}.$$

Since both the encoded state $|\psi(\boldsymbol{x})\rangle$ and the superposition state $|+\rangle_n$ are pure, we have

$$T\left(\rho(\boldsymbol{x}), \frac{1}{2^n} \mathbf{1}\mathbf{1}^\top\right) = \frac{1}{2} \left\| \rho(\boldsymbol{x}) - \frac{1}{2^n} \mathbf{1}\mathbf{1}^\top \right\|_1 = \sqrt{1 - |\langle \psi(\boldsymbol{x}) | + \rangle_n|^2} \leqslant \sqrt{2\epsilon}.$$

□

Proposition B.2 (Proposition 3.3 in the main text)**.** *Denote by $\mathcal{D}_{\mathcal{X}}$ the distribution of the feature $\boldsymbol{x} \in \mathbb{R}^{2^n}$. Assume that the elements of $\boldsymbol{x}$ are i.i.d. with an expected value of 0. In addition, the distribution is symmetric, i.e., the probability density function satisfies $p(x_j) = p(-x_j)$, for all $j \in [0 : 2^n - 1]$. Then after amplitude encoding, the expectation of encoded state is the maximally mixed state, i.e.,*

$$\mathop{\mathbb{E}}_{\boldsymbol{x} \sim \mathcal{D}_{\mathcal{X}}} [\rho(\boldsymbol{x})] = \frac{\boldsymbol{I}}{2^n}.$$

Proof. For the feature vector $\boldsymbol{x} = [x_0, x_1, \cdots, x_{2^n-1}]^\top$, according to the definition of amplitude encoding in Eq. (2), the expectation of encoded states reads

$$\begin{aligned}
\mathop{\mathbb{E}}_{\boldsymbol{x} \sim \mathcal{D}_{\mathcal{X}}} [\rho(\boldsymbol{x})] &= \mathop{\mathbb{E}}_{\boldsymbol{x} \sim \mathcal{D}_{\mathcal{X}}} \left(\sum_{i=0}^{2^n-1} \sum_{j=0}^{2^n-1} \frac{x_i x_j}{\|\boldsymbol{x}\|_2^2} |i\rangle \langle j| \right) \\
&= \sum_{i=0}^{2^n-1} \sum_{j=0}^{2^n-1} \mathop{\mathbb{E}}_{\boldsymbol{x} \sim \mathcal{D}_{\mathcal{X}}} \left(\frac{x_i x_j}{x_0^2 + \cdots + x_{2^n-1}^2} \right) |i\rangle \langle j| .
\end{aligned}$$

(1) Consider the diagonal elements:

$$\begin{aligned}
\mathop{\mathbb{E}}_{\boldsymbol{x} \sim \mathcal{D}_{\mathcal{X}}} \left(\frac{x_i^2}{x_0^2 + \cdots + x_{2^n-1}^2} \right) &= \mathop{\mathbb{E}}_{\boldsymbol{x} \sim \mathcal{D}_{\mathcal{X}}} (1) - \mathop{\mathbb{E}}_{\boldsymbol{x} \sim \mathcal{D}_{\mathcal{X}}} \left(\frac{x_0^2 + \cdots + x_{i-1}^2 + x_{i+1}^2 + \cdots + x_{2^n-1}^2}{x_0^2 + \cdots + x_{2^n-1}^2} \right) \\
&= 1 - \sum_{j=0, j \neq i}^{2^n-1} \mathop{\mathbb{E}}_{\boldsymbol{x} \sim \mathcal{D}_{\mathcal{X}}} \left(\frac{x_j^2}{x_0^2 + \cdots + x_{2^n-1}^2} \right).
\end{aligned}$$

Note that the elements x_j's are independent and identically distributed. We assume that $\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}_\mathcal{X}}\left(\frac{x_j^2}{x_0^2+\cdots+x_{2^n-1}^2}\right)=a$. Then we have $a=1-(2^n-1)a$, and $a=\frac{1}{2^n}$, accordingly. Thus, for all diagonal elements, we have $\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}_\mathcal{X}}\left(\frac{x_i^2}{\|\boldsymbol{x}\|_2^2}\right)=\frac{1}{2^n}$.

(2) Consider the non-diagonal elements:

$$\mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{D}_\mathcal{X}}\left(\frac{x_ix_j}{\|\boldsymbol{x}\|_2^2}\right)=\mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{D}_\mathcal{X}}\left(\frac{x_ix_j}{x_0^2+\cdots+x_{2^n-1}^2}\right).$$

Since $\frac{(-x_i)x_j}{x_0^2+\cdots+x_{2^n-1}^2}=-\frac{x_ix_j}{x_0^2+\cdots+x_{2^n-1}^2}$, and the distribution of x_j is symmetric, we have

$$\mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{D}_\mathcal{X}}\left(\frac{x_ix_j}{\|\boldsymbol{x}\|_2^2}\right)=\mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{D}_\mathcal{X}}\left(\frac{x_ix_j}{x_0^2+\cdots+x_{2^n-1}^2}\right)=0.$$

Therefore,

$$\mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{D}_\mathcal{X}}\left(\frac{x_ix_j}{\|\boldsymbol{x}\|_2^2}\right)=\begin{cases}\frac{1}{2^n} & i=j;\\ 0 & i\neq j.\end{cases}$$

This implies that the expectation of encoded states is the maximally mixed state, i.e., $\mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{D}_\mathcal{X}}\left[\rho(\boldsymbol{x})\right]=\frac{\boldsymbol{I}}{2^n}$. □

Proposition B.3 (Proposition 3.4 in the main text)**.** *For a binary classification dataset, denote by $\mathcal{D}_{\mathcal{X}_0}$ and $\mathcal{D}_{\mathcal{X}_1}$ the distributions of feature $\boldsymbol{x}\in\mathbb{R}^{2^n}$ in class 0 and class 1, respectively. Assume the data are i.i.d. If for any $j\in[0:2^n-1]$, the probability density functions $p_{0,j}(x)$ and $p_{1,j}(x)$ of the j-th element in class 0 and class 1 satisfy $p_{0,j}(x)=p_{1,j}(-x)$, then after amplitude encoding, the expected states of the two classes are identical, i.e.,*

$$\mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{D}_{\mathcal{X}_0}}\left[\rho(\boldsymbol{x})\right]=\mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{D}_{\mathcal{X}_1}}\left[\rho(\boldsymbol{x})\right].$$

Proof. For the feature vector $\boldsymbol{x}=[x_0,x_1,\cdots,x_{2^n-1}]^\top$, let $[\rho(\boldsymbol{x})]_{ij}$ denote the (i,j)-th element of $\rho(\boldsymbol{x})$. Then, according to the definition of amplitude encoding in Eq. (2), we have

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{D}_{\mathcal{X}_0}}\left\{[\rho(\boldsymbol{x})]_{ij}\right\}=\mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{D}_{\mathcal{X}_0}}\left(\frac{x_ix_j}{\|\boldsymbol{x}\|_2^2}\right)\\
=&\underbrace{\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}}_{2^n}\frac{x_ix_j}{x_0^2+\cdots+x_{2^n-1}^2}p_{0,0}(x_0)\cdots p_{0,2^n-1}(x_{2^n-1})\,\mathrm{d}x_0\cdots\,\mathrm{d}x_{2^n-1}\\
=&\underbrace{\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}}_{2^n}\frac{x_ix_j}{x_0^2+\cdots+x_{2^n-1}^2}p_{1,0}(-x_0)\cdots p_{1,2^n-1}(-x_{2^n-1})\,\mathrm{d}x_0\cdots\,\mathrm{d}x_{2^n-1} \qquad (6)\\
=&(-1)^{2^n}\underbrace{\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}}_{2^n}\frac{(-x_i)(-x_j)}{(-x_0)^2+\cdots+(-x_{2^n-1})^2}p_{1,0}(-x_0)\cdots p_{1,2^n-1}(-x_{2^n-1})\,\mathrm{d}(-x_0)\cdots\,\mathrm{d}(-x_{2^n-1})\\
=&\underbrace{\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}}_{2^n}\frac{x_ix_j}{x_0^2+\cdots+x_{2^n-1}^2}p_{1,0}(x_0)\cdots p_{1,2^n-1}(x_{2^n-1})\,\mathrm{d}x_0\cdots\,\mathrm{d}x_{2^n-1} \qquad (7)\\
=&\mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{D}_{\mathcal{X}_1}}\left\{[\rho(\boldsymbol{x})]_{ij}\right\}.
\end{aligned}$$

where Eq. (6) follows from $p_{0,j}(x)=p_{1,j}(-x)$ and Eq. (7) follows from the change of variable $x_k\to-x_k$. Therefore, $\mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{D}_{\mathcal{X}_0}}\left[\rho(\boldsymbol{x})\right]=\mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{D}_{\mathcal{X}_1}}\left[\rho(\boldsymbol{x})\right]$.

□

C More Experiments

C.1 Synthetic data

In addition to the experiments on Fig. 2 investigating the effects of training sample size and PQC layer count on test prediction error, we also conducted the same experiments on the datasets shown in Fig. 3 and Fig. 4. Following the same experimental setup as in the main text, we sampled 3000 data points per class and selected 1000 as the test set (2000 test samples in total, 1000 per class). The same test set was used for all combinations of training sample sizes and layer counts. The loss function was cross-entropy, the optimizer was Adam with a learning rate of 0.001, training was conducted for 100 epochs, and each experiment was repeated 5 times.

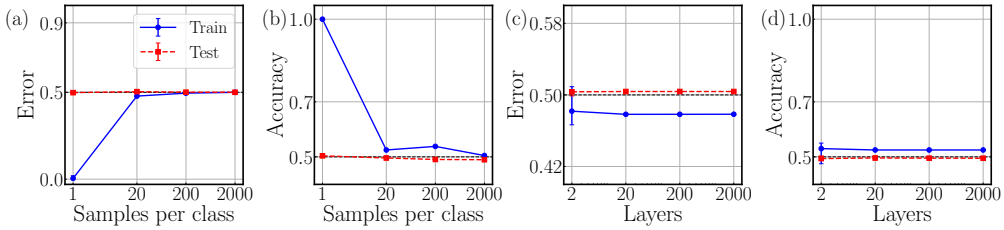

Figure C.1: (a) Train/test error vs. training samples (20 layers); (b) corresponding accuracy for (a); (c) train/test error vs. circuit layers (20 samples per class); (d) corresponding accuracy for (c). Solid lines indicate the mean over five runs; error bars span the minimum and maximum values.

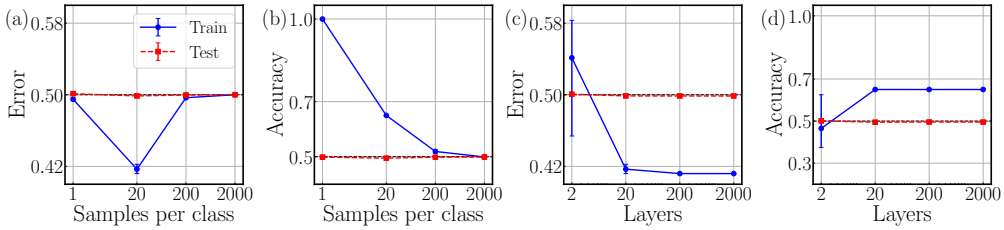

Figure C.2: (a) Train/test error vs. training samples (20 layers); (b) corresponding accuracy for (a); (c) train/test error vs. circuit layers (20 samples per class); (d) corresponding accuracy for (c). Solid lines indicate the mean over five runs; error bars span the minimum and maximum values.

The experimental results corresponding to Fig. 3 and Fig. 4 are shown in Fig. C.1 and Fig. C.2, respectively. We observe that although varying the number of training samples and PQC layers may lead to changes in training error, and the trends differ across datasets, the test error consistently remains around 0.5 (random guessing). Correspondingly, the test accuracy also approaches that of random guessing.

C.2 Real-World data

We also conduct classification experiments on the PathMNIST and CIFAR10 datasets, following the same experimental setup as described in Subsec. 4.2. The results are shown in Fig. C.3.

D QCNN Circuit Diagram

The 10-qubit QCNN circuit is shown in Fig. D.1. After amplitude encoding, the PQC consists of four convolutional layers (Conv) and pooling layers (Pool). In the pooling layers, one of the paired qubits is measured, and conditioned on the measurement result, a rotation is applied to the remaining qubit. The two-qubit gates are $XX(\theta) = e^{-i\frac{\theta}{2}(X\otimes X)}$, $YY(\theta) = e^{-i\frac{\theta}{2}(Y\otimes Y)}$, and $ZZ(\theta) = e^{-i\frac{\theta}{2}(Z\otimes Z)}$. The single-qubit gates are $R_x(\theta) = e^{-i\frac{\theta}{2}X}$, $R_z(\theta) = e^{-i\frac{\theta}{2}Z}$, where X, Y, and Z are Pauli operators. The circuit parameters are initialized from a standard Gaussian distribution.

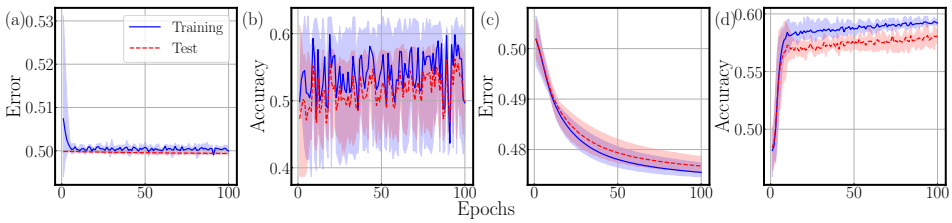

Figure C.3: (a)–(b) Training and test errors together with the corresponding accuracies on the PathMNIST dataset; (c)–(d) training and test errors together with the corresponding accuracies on the CIFAR10 dataset. Solid lines indicate the mean over five runs; shaded regions span the minimum and maximum values.

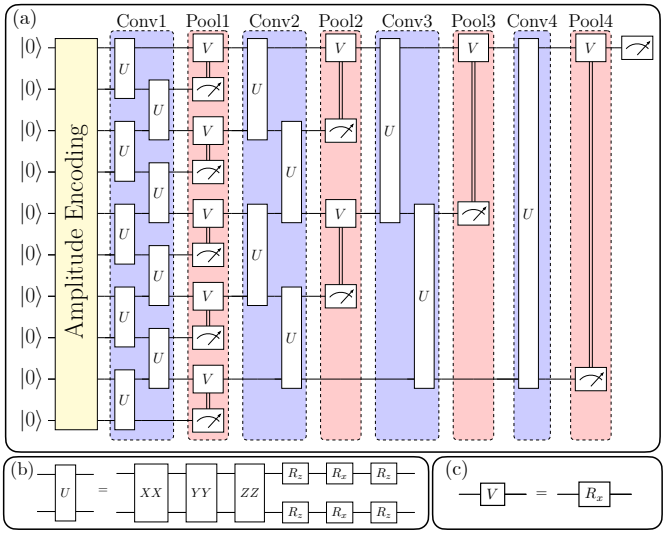

Figure D.1: (a) The architecture of QCNN. (b) The circuit for gate U. (c) The circuit for gate V.

E Limitations and Future Work

While this work has highlighted the normalized quadratic-feature limitation of amplitude encoding and motivated the design of alternative encoding schemes, the question of how to construct such schemes remains open. We plan to devote future work to the development of new encoding methods to address this challenge.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly state the main contributions: (1) analyzing linear classification limitations in the normalized quadratic feature space induced by amplitude encoding, (2) establishing theoretical results under general conditions, and (3) providing insights into why QML models perform well on simple datasets like MNIST but fail on complex real-world tasks. The paper delivers on these claims through theoretical analysis, illustrative examples, and experimental validation.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper discusses limitations in Sec. 5 and App. E. It acknowledges that while the work identifies the normalized quadratic-feature limitation, the question of how to construct new encoding schemes to mitigate these issues remains open. The paper also notes that it focuses on end-to-end quantum machine learning and does not consider hybrid classical-quantum architectures.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: All theoretical results (Theorem 3.1 and Propositions 3.2-3.4) are clearly stated with complete assumptions. Full formal proofs are provided in App. A and App. B. The main theorem includes all necessary assumptions about data distributions, and the propositions specify conditions for normalized quadratic-feature limitations under different scenarios.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper provides detailed experimental settings including: circuit architectures (Fig. 8 and App. D), hyperparameters (learning rates, optimizers, epochs), dataset specifications, train/test splits, initialization methods, and number of experimental runs. The synthetic datasets are precisely defined with distributions, and real datasets are standard benchmarks (MNIST, CIFAR-10, EuroSAT, PathMNIST) with specified preprocessing.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: We will provide the source code and data after the paper is accepted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper provides comprehensive experimental details: Adam optimizer with learning rate 0.001, 100 training epochs, specific train/test splits (e.g., 1000 samples per class for test), parameter initialization from standard Gaussian distribution, cross-entropy loss function, and detailed circuit architectures. The number of repetitions (5 runs) is also specified.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The paper reports error bars in experimental figures (Fig. 9, 10, and 11) showing the range (minimum to maximum) over 5 independent runs. The plots clearly show solid lines for means and either error bars or shaded regions spanning the min/max values, properly capturing experimental variability due to random initialization and training dynamics.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: All quantum machine learning experiments were conducted on a CPU platform (Intel Core i9-13900KF, 64GB RAM), while classical machine learning experiments were run on a GPU platform (NVIDIA RTX 4090). The runtime of each individual experiment did not exceed 5 hours.

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research is purely theoretical and experimental analysis of quantum machine learning algorithms. It involves no human subjects, personal data collection, or potentially harmful applications. The work aims to advance scientific understanding of quantum computing limitations, which is beneficial for the research community.

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: This work is foundational theoretical research identifying limitations in quantum machine learning algorithms. It does not have direct immediate societal applications or deployment concerns. The research aims to advance scientific understanding to eventually improve quantum computing effectiveness, which would have broad positive impacts for computational capabilities in various fields.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not introduce new models, datasets, or systems that pose misuse risks. It is a theoretical analysis of existing quantum machine learning algorithms using standard public datasets. No safeguards are needed as the work poses no risk for misuse.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The paper properly cites all datasets used (MNIST, CIFAR-10, EuroSAT, PathMNIST) with appropriate references to their original publications. These are all well-established public datasets with permissive usage terms for research purposes.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not release new datasets, code, or models. It presents theoretical analysis and uses existing public datasets for validation. The synthetic datasets described are simple mathematical constructions that can be easily reproduced from the provided specifications.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve any crowdsourcing or research with human subjects. All experiments are computational studies using existing datasets and synthetic data.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve any research with human subjects. All work is computational and theoretical analysis of quantum machine learning algorithms.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: The core methodology development in this research does not involve LLMs as any important, original, or non-standard components. The work is based on mathematical analysis, theoretical proofs, and quantum circuit simulations.