A VC-Dimension View of Capacity and Generalization in Quantum Machine Learning with Classical Data

Anonymous authors
Paper under double-blind review

Abstract

Quantum machine learning (QML) models are often motivated by the large Hilbert spaces accessible to parametrized quantum circuits, but their capacity is ultimately determined by their trainable parameters. We study model capacity through the Vapnik-Chervonenkis (VC) dimension for several QML families. For purely univariate QML classifiers, we derive linear VC dimensions, analogous to univariate classical models. Multivariate uniform re-uploading analyzes the VC dimension from the encoding side, providing an upper bound of $O\big(\big(\frac{d}{p}+1\big)^p\big)$ for d trainable parameters and input dimension p, which is consistent with empirical evaluations. Finally, the bounded-reuse trainable Pauli classifier analyzes the VC dimension from the trainable parameter side, obtaining a $O(d \log d)$ upper bound, similar to a feedforward neural net. We conclude that asymptotic scaling depends on input dimension and architecture rather than on the model's classical or quantum nature. Empirically, we find that QML models trained on classical data exhibit generalization trends consistent with what VC-bounds predict. Conversely, previous studies have observed strong generalization for very few quantum-data, suggesting that QML generalization behavior is not yet fully understood.

1 Introduction

Quantum machine learning (QML) is often motivated by access to high-dimensional Hilbert spaces, but a learner does not train a Hilbert space – it optimizes parameters. Although a parameterized quantum circuit (PQC) may act on an exponentially large state space, its classification capacity is determined by which input labelings can be realized by varying the parameters in its gates. This raises a basic learning-theoretic question: how much classification capacity does a QML model actually gain per trainable parameter?

In the current era of big data, modern machine learning models are trained on huge, high-dimensional datasets, and quantum computers promise speedups for certain data-driven tasks compared to classical counterparts (Zhao et al., 2026; Liu et al., 2021). This has sparked substantial interest in QML, although a practical advantage remains to be demonstrated. One way to study how efficiently models use their trainable parameters is through the concept of Vapnik-Chervonenkis (VC) dimension, a distribution-free measure of the capacity of a model class, which is typically used in the context of generalization bounds (Vapnik, 1999).

However, seminal work showed that deep neural networks generalize far better than VC-based theory predicts (Zhang et al., 2017), since VC-based bounds typically require unrealistically large datasets and ignore data distribution. Similarly, recent studies indicate that standard complexity measures may fail to predict the generalization behavior of QML models on quantum data: Gil-Fuster et al. (2024) demonstrate that Quantum Convolutional Neural Networks can fit random quantum data, implying any uniform VC- or Rademacher-type bound is essentially vacuous for dataset sizes up to some N' and hence cannot explain the observed good generalization of QML models for small quantum data (Caro et al., 2022; Kottmann et al., 2021). Peters & Schuld (2023), on the other hand, analyzed the behavior of "benign" overfitting in QML models that can be represented as Fourier series. They considered target functions made up of Fourier series, which the QML models can interpolate extremely well. Yet investigating the generalization behavior of QML models on classical, practical datasets remains an open research question.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Another line of QML research has focused on model capacity and generalization, but has not examined these aspects in terms of parameter efficiency (Caro et al., 2021; Gyurik et al., 2023; Du et al., 2022; Bu et al., 2023; Hur & Park, 2025). The same applies to the key work by Abbas et al. (2021) introducing the *effective dimension*, a Fisher information-based measure that estimates the size of a model in model space – the space of all possible functions for a given model class. The study finds that quantum neural networks achieve a larger effective dimension than classical networks with the same number of trainable parameters.

Contributions. This work investigates whether certain QML architectures use trainable parameters more efficiently than classical baselines – measured by VC dimension – and whether their generalization behavior on classical data follows uniform generalization bounds. Our main contributions are:

- We derive VC dimension bounds for several QML hypothesis classes, showing that model capacity is data-dimension- and architecture-dependent rather than a property of the model's quantum or classical nature: univariate QML models scale linearly like simple classical baselines, while multivariate inputs or a trainable parameters-based analysis admits stronger asymptotic bounds, the latter similar to those of feedforward neural networks.
- We observe that the empirical generalization behavior of QML models on classical data follows trends predicted by VC generalization bounds: larger training sets reduce the generalization gap, while higher-capacity re-uploading models tend to generalize worse.
- However, other studies have reported good generalization of QML from very few quantum-data (Caro et al., 2022; Kottmann et al., 2021). Hence, we conclude that the generalization behavior of QML models is more nuanced and we do not yet fully understand it.

2 Background

2.1 Binary Classification

We consider supervised binary classification with classical data, meaning samples are vectors that must be encoded into quantum states by a quantum feature map.[1] The resulting state after encoding sample $\mathbf{x_i}$ is of the form $|\phi(\mathbf{x_i})\rangle$.[2] Quantum data, on the other hand, is naturally a quantum state: sample i is given by $|\psi_i\rangle$.

Let $(\mathbf{x}, y) \sim \mathcal{D}$ be input-output pairs sampled from an unknown joint distribution $\mathcal{D}$, where $\mathbf{x} \in \mathcal{X} \subset \mathbb{R}^p$ and $y \in \{-1, +1\}$. A parameterized model $f_{\boldsymbol{\theta}}$ induces a function class $\mathcal{F} = \left\{ f_{\boldsymbol{\theta}} : \mathcal{X} \to \mathbb{R} \;\middle|\; \boldsymbol{\theta} \in \Theta \subset \mathbb{R}^d \right\}$ which gives rise to the binary hypothesis class

$$\mathcal{H} \;=\; \Big\{ h_{\boldsymbol{\theta}} : \mathcal{X} \to \{-1, +1\} \;\Big|\; h_{\boldsymbol{\theta}}(\mathbf{x}) = \operatorname{sgn}\big(f_{\boldsymbol{\theta}}(\mathbf{x})\big),\; \boldsymbol{\theta} \in \Theta \Big\}, \tag{1}$$

where we use the convention $\operatorname{sgn}(t) = \begin{cases} +1, & t \geq 0, \\ -1, & t < 0. \end{cases}$ The goal of a learning algorithm in statistical learning theory is to select a hypothesis h^* that minimizes the true error or (expected) *risk* $R(h) := \Pr_{(x,y)\sim\mathcal{D}}[h(x) \neq y]$. In practice, the true distribution $\mathcal{D}$ is unknown, which is why we use a training set $S = \{(x_i, y_i)\}_{i=1}^N$ over which we minimize the empirical risk $\hat{R}_S(h) := \frac{1}{N}\sum_{i=1}^N \mathbf{1}\{h(x_i) \neq y_i\}$. According to Hoeffding's inequality (Hoeffding, 1963), the true error can be approximated using another empirical risk $\hat{R}_T$ with a large test set T.

The difference $|R(h) - \hat{R}_S(h)|$ is referred to as the absolute *generalization gap* and can be bound based on the VC inequality (Mohri et al., 2012): If $\mathcal{H}$ has finite VC dimension $\mathrm{VCdim}(\mathcal{H})$, then for any $\delta \in (0, 1)$, with probability at least $1 - \delta$,

$$\sup_h |R(h) - \hat{R}_S(h)| \;\leq\; B(N, \mathcal{H}, \delta), \tag{2}$$

[1] A quantum feature map is a function $\phi : \mathcal{X} \to \mathcal{K}$ where $\mathcal{K}$ is a Hilbert space and the feature vectors are quantum states. This map transforms $\mathbf{x} \to |\phi(\mathbf{x})\rangle$ through a unitary transformation $U(\mathbf{x})$. For more details on (quantum) feature maps, see `https://pennylane.ai/qml/glossary/quantum_feature_map`.

[2] A quantum state is denoted by $|\cdot\rangle$. A very brief explanation of this notation is given in Section 2.3

where

$$B(N, \mathcal{H}, \delta) = \sqrt{\frac{8}{N}\left[\ln\left(\frac{4}{\delta}\right) + \mathrm{VCdim}(\mathcal{H}) \ln\left(\frac{2eN}{\mathrm{VCdim}(\mathcal{H})}\right)\right]}. \tag{3}$$

This bound decreases with sample size N and increases with the VC dimension for $N >> \mathrm{VCdim}$, capturing the trade-off between data and model complexity.

2.2 VC Dimension

The VC dimension is a standard measure of the capacity of a binary hypothesis class. A finite set $\{x_1, \ldots, x_m\} \subset \mathcal{X}$ is said to be *shattered* by $\mathcal{H}$ if, for every labeling $(y_1, \ldots, y_m) \in \{-1, +1\}^m$, there exists $h \in \mathcal{H}$ such that $h(x_i) = y_i$ for all $i = 1, \ldots, m$. Equivalently, $\mathcal{H}$ can realize all 2^m binary labelings on that set. The VC dimension of $\mathcal{H}$ is the largest size of a set that can be shattered:

$$\mathrm{VCdim}(\mathcal{H}) := \sup\Big\{m \in \mathbb{N} : \exists x_1, \ldots, x_m \in \mathcal{X} \text{ shattered by } \mathcal{H}\Big\}.$$

It is a distribution-free measure of classification capacity which we interpret as an effective parameter count and compare its scaling with the number of trainable parameters.

2.3 Quantum models.

PQCs define real-valued functions by preparing an n-qubit quantum state, applying a data- and parameter-dependent unitary transformation, and measuring a fixed observable. Such a model is defined as

$$f_{\boldsymbol{\theta}}^{A}(\mathbf{x}) = \langle 0|^{\otimes n} U_A^{\dagger}(\mathbf{x}; \boldsymbol{\theta})\, \mathcal{M}_A\, U_A(\mathbf{x}; \boldsymbol{\theta})\, |0\rangle^{\otimes n}. \tag{4}$$

Here, $|0\rangle^{\otimes n} = (1 \quad 0 \quad \ldots \quad 0)^T \in \mathbb{C}^{2^n \times 1}$ denotes a state vector of an n-qubit system, $\langle 0|^{\otimes n} = (|0\rangle^{\otimes n})^{\dagger} \in \mathbb{C}^{1 \times 2^n}$ describes its conjugate transpose, $U_A(\mathbf{x}; \boldsymbol{\theta}) \in \mathbb{C}^{2^n \times 2^n}$ is a parameterized unitary operator also called trainable feature map, and $\mathcal{M}_A \in \mathbb{C}^{2^n \times 2^n}$ indicates a Hermitian observable represented by a self-adjoint matrix. A is the fixed circuit architecture, specifying n, d, p, the unitary and the measurement. These models can often be interpreted as Fourier series in the input (Schuld et al., 2021).

2.3.1 Trainable Pauli Classifier.

The trainable Pauli classifier family $\mathcal{F}_{\mathrm{pauli}}^{A}$ is a family of generic PQCs that allow for analysis of its VC dimension from the trainable parameter side. The data encoding in A can be arbitrary, and includes IQP-style encodings and Pauli-decomposable QCNN or butterfly ansätze. The only constraints are that every trainable gate is generated by an n-qubit Pauli-word, $R_P(\theta_j) = e^{-i\theta_j P/2}$ with $P^2 = I$, $P = P_1 \otimes \cdots \otimes P_n$, $P_k \in \{I, X, Y, Z\}$, and that all remaining gates are independent of $\boldsymbol{\theta}$. The total number of parameters is d and an individual parameter θ_j occurs q_j times. We define $Q := \sum_{j=1}^{d} q_j$ as the total number of occurrences of trainable parameters in the circuit. The hypothesis class $\mathcal{H}_{\mathrm{pauli}}^{A}$ is defined as in 1.

2.3.2 Bounded-Reuse Trainable Pauli Classifier.

This subclass $\mathcal{F}_{\mathrm{bound-pauli}}^{A} \subset \mathcal{F}_{\mathrm{pauli}}^{A}$ assumes that each trainable parameter occurs at most $q_{\max} = O(1)$ times. Consequently, $Q = \sum_{j=1}^{d} q_j \leq q_{\max} d = O(d)$.

2.3.3 General Re-Uploading Classifier.

Another approach to analyzing the VC dimension of quantum classifiers is from the encoding side. For this, we define the function class of general re-uploading encoding classifiers (Pérez-Salinas et al., 2020), where data is encoded repeatedly in parameter-independent gates. With fixed input dimension p, number of qubits n, and total parameter count d all specified by A, The classifier family is

$$\mathcal{F}_{\mathrm{reup}}^{A} := \left\{ f_{\boldsymbol{\theta}} : [0, 2\pi]^p \to [-1, 1] \;\middle|\; \boldsymbol{\theta} \in [0, 2\pi]^d \right\}, \tag{5}$$

where $f_{\boldsymbol{\theta}}^A$ is given by Equation (4). In the architecture A, input component x_i is encoded r_i times through single-qubit Pauli rotations $R_a(x_i)$, $a \in \{x, y, z\}$. A may additionally contain trainable and entangling gates. For any A, every function $f_{\boldsymbol{\theta}}^A \in \mathcal{F}_{\text{reup}}^A$ can be represented as a partial Fourier series of the form

$$f_{\boldsymbol{\theta}}^A(\mathbf{x}) = \sum_{\boldsymbol{\omega} \in \Omega} c_{\boldsymbol{\omega}}(A, \boldsymbol{\theta}) e^{i\boldsymbol{\omega} \cdot \mathbf{x}} \quad \text{with} \quad \Omega \subseteq \prod_{i=1}^{p} \{-r_i, -r_i + 1, \ldots, r_i\}. \tag{6}$$

where the frequency spectrum Ω is the Cartesian product of the individual frequency sets. It is determined solely by the data-encoding operations, while the coefficients $c_{\boldsymbol{\omega}}(A, \boldsymbol{\theta})$ depend on the full circuit and satisfy $c_{\boldsymbol{\omega}}(A, \boldsymbol{\theta}) = c_{-\boldsymbol{\omega}}^*(A, \boldsymbol{\theta})$ (Schuld et al., 2021).

2.3.4 Uniform Re-Uploading Classifier.

The subclass $\mathcal{F}_{\text{unif}-\text{reup}}^A \subset \mathcal{F}_{\text{reup}}^A$ assumes that all input features are re-uploaded the same number of times, and the trainable unitary belonging to each encoding has the same number of trainable parameters, i.e. $r_1 = r_2 = \cdots = r_p = r = \frac{d'}{cp} \in \mathbb{N}$ for some constant $c \in \mathbb{N}$ specifying the number of trainable parameters in each trainable unitary. The model may include d_{init} parameters before the first encoding block, which prepare a different initial state. Hence, $d' = d - d_{\text{init}}$. An exemplary architecture of the uniform re-uploading classifier is shown in Figure 3 in Appendix C.3.

2.3.5 Univariate n-Qubit Re-Uploading Classifier.

For $n, r \geq 1$, the subclass $\mathcal{F}_{1\text{p}-\text{reup}}^A \subset \mathcal{F}_{\text{unif}-\text{reup}}^A$ uses the fixed univariate architecture

$$U_A(x; \boldsymbol{\theta}) = E_n \bigotimes_{j=0}^{n-1} \left[W_j^{(r)} R_{a_{j,r-1}}(x) \cdots W_j^{(1)} R_{a_{j,0}}(x) W_j^{(0)} \right], \tag{7}$$

where A specifies only $a_{j,\ell} \in \{x, y, z\}$, n and r. Moreover, $W_j^{(s)} = R_z(\theta_{j,s,2}) R_y(\theta_{j,s,1}) R_z(\theta_{j,s,0})$, with $s = 0, \ldots, r$. The model measures $M^{(n)} = Z_{n-1}$ and has $d = 3n(r+1)$ trainable parameters. The final entangling layer is

$$E_n = \begin{cases} I, & n = 1, \\ \text{CNOT}_{n-1 \to 0} \, \text{CNOT}_{n-2 \to n-1} \cdots \text{CNOT}_{0 \to 1}, & n \geq 2, \end{cases} \tag{8}$$

where the rightmost gate acts first.

2.3.6 Univariate n-Qubit Trainable-Encoding Classifier.

A trainable-encoding model combining data and parameters within the same gates. It is defined by the fixed unitary

$$U_A(x; \boldsymbol{\theta}) = E_n \bigotimes_{j=0}^{n-1} R_{a_j}(\theta_j x), \qquad a_j \in \{x, y\}, \tag{9}$$

with the same E_n as in Equation (8), measurement $M^{(n)} = Z_{n-1}$, and $d = n$ and $\boldsymbol{\theta} \in [0, 2\pi]^d$. A defines n, a_j. By Lemma A.30, the model function is $f_{\boldsymbol{\theta}}^A(x) = \prod_{j=0}^{n-1} \cos(\theta_j x)$.

2.3.7 Univariate Shallow Fourier Model.

As shown in Equation (6), a PQC output can be expressed as a real Fourier series which is why Fourier models are commonly used as an intermediate class between QML and classical ML methods (Wendlinger et al., 2024). Motivated by this, we consider the univariate shallow Fourier model

$$f_{\boldsymbol{\theta}}(x) = \sum_{\omega \in \Omega_+} \left(c_\omega e^{i\omega x} + c_{-\omega} e^{-i\omega x} \right) = \sum_{j=1}^{k} \left(a_j \cos(\omega_j x) + b_j \sin(\omega_j x) \right), \tag{10}$$

where $\boldsymbol{\theta} = (\boldsymbol{\omega}, \boldsymbol{a}, \boldsymbol{b})$ and $c_\omega = c_{-\omega}^*$. The function class is $\mathcal{F}_{\text{fourier}} = \left\{ f_{\boldsymbol{\theta}} : [0, 2\pi] \to \mathbb{R} \;\middle|\; \boldsymbol{\theta} \in [0, \Omega_{\max}]^k \times \mathbb{R}^k \times \mathbb{R}^k \right\}$, where the total number of trainable parameters is $d = 3k$ and the maximum possible frequency is $\Omega_{\max}$.

2.4 Classical Baselines

We use three standard ML models as baselines: a perceptron, a univariate shallow ReLU network and the generic feedforward neural network with piecewise linear activation function. For these models, the input domain is $\mathbf{x} \in [0,1]^p$ and the trainable parameters are in $\mathbb{R}^d$. The model definitions can be found in Appendix C.3.

3 Capacity Analysis

This section analyzes the model capacity in terms of the VC dimension for selected QML and ML model families. First, Section 3.1 provides the main result of this work, a summary of the VC dimension bounds of the considered model classes. Afterwards, Section 3.3 derives VC dimension bounds of the trainable Pauli classifier. Next, Section 3.4 obtains an exact expression for the VC dimension of the univariate n-qubit re-uploading classifier, and Section 3.5 establishes VC dimension bounds for the multivariate uniform re-uploading classifier. Derivations for the other QML model families can be found in Appendix A and the VC dimensions of the classical models are listed in Appendix B.

3.1 Summary of Model Capacities

Table 1 compares the VC dimension as a function of the number of trainable parameters d. In short, VC-based parameter efficiency is architecture-dependent rather than a property of the quantum or classical nature of the model: the univariate QML models match the linear scaling in d of the shallow ReLu network up to constant factors, while the bounded-reuse trainable Pauli classifier admits a loose $O(d \log d)$ upper bound, similar to the nearly-tight $O(dL \log d)$ bound of the feedforward neural network. The loose uniform re-uploading classifier's upper bound $O\big(\big(\frac{d}{p}+1\big)^p\big)$ depends on the input dimension p, as confirmed empirically in Section 4.
While the simple univariate models studied here all have linear VC scaling in d, their constants differ. The univariate n-qubit re-uploading classifier has a coefficient of $2/3$, exceeding the shallow ReLU's coefficient of $1/3$ but falling below the perceptron's coefficient of 1. Finally, the trainable-encoding coefficient lies in $[1, 10]$ and the Fourier network's in $[2/3, C\Omega_{\max}]$. Both families can be highly expressive because their trigonometric terms carry trainable frequencies. If these are unbounded, even $x \mapsto \operatorname{sgn}(\cos(\theta x))$ has infinite VC dimension (Vapnik, 1999). However, this capacity is not automatically useful in practice, since realizing highly oscillatory decision boundaries may require large parameter values and can make optimization difficult (Novello et al., 2025; Parascandolo et al., 2017).

3.2 Proof strategies.

Fixed encoding frequencies result in re-uploading outputs with bounded Fourier support and, consequently, a finite-dimensional trigonometric polynomial space. For the univariate n-qubit re-uploading classifier every bounded polynomial is realizable (Fejér-Riesz factorization enables interpolation), giving an exact bound. For uniform multivariate re-uploading, the frequency coefficients are not necessarily independent, so only a loose upper bound follows. Trainable-encoding and shallow Fourier models have trainable frequencies and their upper bounds instead use sign-pattern and short-interval zero-count arguments, respectively. The classical baselines (perceptron, shallow ReLU, feedforward NN) use their standard bounds.

3.3 VC Dimension of Trainable Pauli Classifier

This section derives an upper bound of the VC dimension of the trainable Pauli classifier based on the trainable parameters. The full proof can be found in Appendix A.1.

Theorem 3.1 (VC upper bound of trainable Pauli classifier)**.** *Let $\mathcal{H}^A_{\text{pauli}}$ be the hypothesis class of the trainable Pauli classifier with fixed architecture A defined in Section 2.3.1. Then the VC dimension is*

[3] If the frequencies of the Fourier network are bounded in $[0, 2\pi]$, as it is the case for the trainable-encoding model class, the effective coefficient is $C\Omega_{\max} \approx 17$.

Table 1: Comparison of VC-dimension-based capacity results. Models with VC dimension derived in this work are highlighted in **bold** and purely univariate models in *italic*. d denotes the number of trainable parameters, d_{init} the number of trainable parameters before the first encoding layer, p the input dimension, e is Euler's number, Q the total parameter count including parameter re-use, and $\Omega_{\max}$ is the maximum allowed frequency. The $\cdot$ is a placeholder for the VC dimension.

Model class	d	VCdim	Ref.
Trainable Pauli	d	$< 2d\log_2(64eQ)$	Theorem 3.1
Bounded-Reuse Pauli	d	$\leq O(d\log d)$	Theorem 3.1
Univ. n-qubit Re-up.	$3n(r+1)$	$\frac{2}{3}d - 2n + 1$	Theorem 3.2
Uniform Re-up.	d	$\frac{2}{3}(d-d_{\text{init}})+1 \leq \cdot \leq O\big(\big(\frac{d}{p}+1\big)^p\big)$	Theorem 3.3
Trainable-Encoding	n	$d \leq \cdot \leq 10d$	Theorem A.35
Fourier network[3]	$3k$	$\frac{2}{3}d \leq \cdot \leq C\Omega_{\max}d$	Theorem A.39
Perceptron	$p+1$	d	Appendix B.1, Vapnik (1999)
Shallow ReLU network	$3h+1$	$\frac{d}{3}+\frac{2}{3}$	Appendix B.2, DeVore et al. (2021)
Feedforward NN	d	$O(dL\log d)$	Appendix B.3, Bartlett et al. (2019)

given by

$$\boxed{\text{VCdim}(\mathcal{H}^A_{\text{pauli}}) < 2d\log_2(64eQ).}$$

For the bounded-reuse trainable Pauli classifier with $Q = O(d)$, we obtain

$$\boxed{\text{VCdim}(\mathcal{H}^A_{\text{bound-pauli}}) \leq O(d\log d)}$$

Proof sketch. By Lemma A.2 and Proposition A.3, for every fixed input $\mathbf{x}$, the circuit output is a real polynomial in $\cos\theta_1, \sin\theta_1, \ldots, \cos\theta_d, \sin\theta_d$, with degree at most q_j in each parameter pair and total degree at most $Q = \sum_{j=1}^d q_j$.
The parameter domain is the torus $(S^1)^d$, which is covered by 2^d rational charts by Lemma A.4. On each chart, Proposition A.5 converts the outputs into real polynomials in d variables of degree at most $2Q$, without changing their signs. Hence, for $m \geq d$, Proposition A.6 gives, $\Pi_{\mathcal{H}_{\text{pauli}}}(m) \leq \left(\frac{32eQm}{d}\right)^d$. If m points are shattered, then $2^m \leq \left(\frac{32eQm}{d}\right)^d$. Setting $t = m/d$ and applying Lemma A.7 yields $t < 2\log_2(64eQ)$, and therefore $m < 2d\log_2(64eQ)$. The case $m < d$ follows immediately since $Q \geq 1$ implies $m < d < 2d\log_2(64eQ)$. □

This upper bound holds for every A satisfying the assumptions in Section 2.3.1 and Section 2.3.2, respectively. As stated in Proposition A.1, no universal positive lower bound can exist for all A. Nevertheless, for $n = 1$ we have $\mathcal{F}^A_{\text{1p-reup}} \subset \mathcal{F}^A_{\text{bound-pauli}} \subset \mathcal{F}^A_{\text{pauli}}$ and hence, $\text{VCdim}(\mathcal{H}^{A'}_{\text{pauli}}) \geq \text{VCdim}(\mathcal{H}^{A'}_{\text{bound-pauli}}) \geq \frac{2}{3}d - 1$ for some A', meaning linear VC dimension is achievable within this circuit family.

3.4 VC Dimension of Univariate n-Qubit Re-Uploading Classifier

This section derives an exact expression for the VC dimension of the univariate n-qubit re-uploading classifier based on the data encoding. The detailed proof can be found in Appendix A.2

Theorem 3.2 (VC dimension of the univariate n-qubit re-uploading classifier)**.** *Let $\mathcal{H}^{(n)}_{\text{1p-reup}}$ be the hypothesis class of the univariate n-qubit re-uploading classifier with $n \geq 1$ qubits and $r \geq 1$ re-uploadings per qubit. Then*

$$\boxed{\text{VCdim}\Big(\mathcal{H}^{(n)}_{\text{1p-reup}}\Big) = 2nr + 1 = \frac{2}{3}d - 2n + 1,}$$

since the total number of trainable parameters is $d = 3n(r+1)$.

Proof sketch. By Lemmas A.21 and A.22, the model output factorizes into n bounded trigonometric polynomials of degree at most r. Hence it has degree at most nr, and Proposition A.23 gives

$$\mathrm{VCdim}\Big(\mathcal{H}^{(n)}_{\mathrm{1p-reup}}\Big) \leq 2nr + 1.$$

For the matching lower bound, Lemma A.24 constructs n factors in $T^{(1)}_r$ whose product realizes every labeling of $2nr + 1$ points. By Lemma A.20, each factor is realizable by one re-uploading branch, yielding Proposition A.25. Thus

$$\mathrm{VCdim}\Big(\mathcal{H}^{(n)}_{\mathrm{1p-reup}}\Big) = 2nr + 1.$$

□

3.5 VC Dimension Bounds of Multivariate Uniform Re-Uploading Classifier

We establish an upper bound of the VC dimension of the uniform re-uploading classifier by using the bound derived by Chen et al. (2022). A lower bound follows by considering the subclass that realizes a univariate 1-qubit re-uploading classifier in Proposition A.27.

Theorem 3.3 (VC bounds of uniform re-uploading model)**.** *Let $\mathcal{F}^{(p)}_{\mathrm{unif-reup}}$ be the function family of the uniform re-uploading model with input dimension p defined in Section 2.3.4 and $\mathcal{H}^{(p)}_{\mathrm{unif-reup}}$ be the corresponding hypothesis class, as defined in Equation* (1)*. Then*

$$\boxed{2pr + 1 \leq \mathrm{VCdim}(\mathcal{H}^{(p)}_{\mathrm{unif-reup}}) \leq \prod_{1}^{p}(2r + 1) = O\big((\frac{d}{p} + 1)^p\big).}$$

Proof sketch. **Lower bound.** Proposition A.27 restricts the uniform model to the diagonal $\Delta(t) = (t, \ldots, t)$, where all pr encodings embed the same scalar input and can realize the 1-qubit univariate re-uploading subclass with pr re-uploadings. By Theorem 3.2, this subclass has VC dimension $2pr + 1$.

Upper bound. By Proposition A.28, data re-uploading restricts the model outputs to a finite-dimensional space of multivariate trigonometric polynomials with bounded frequencies, so the VC dimension is controlled by the size of this space, leading to $\mathrm{VCdim}(\mathcal{H}^{(p,n)}_{\mathrm{unif-reup}}) \leq \prod_{i=1}^{p}(2r_i+1)$. Per model definition in Section 2.3.4, we have $\mathcal{H}^{(p,n)}_{\mathrm{unif-reup}} \subset \mathcal{H}^{(p,n)}_{\mathrm{unif-reup}}$, hence $\mathrm{VCdim}(\mathcal{H}^{(p,n)}_{\mathrm{unif-reup}}) \leq \mathrm{VCdim}(\mathcal{H}^{(p,n)}_{\mathrm{unif-reup}}) \leq \prod_{i=1}^{p}(2r_i + 1)$. Additionally, per definition $r_1 = r_2 = \cdots = r_p = r$, leading to

$$\mathrm{VCdim}(\mathcal{H}^{(p,n)}_{\mathrm{unif-reup}}) \leq \prod_{i=1}^{p}\big(2r_i + 1\big) = \big(2r + 1\big)^p = \big(\frac{2(d - d_{\mathrm{init}})}{cp} + 1\big)^p = O\big((\frac{d}{p} + 1)^p\big).$$

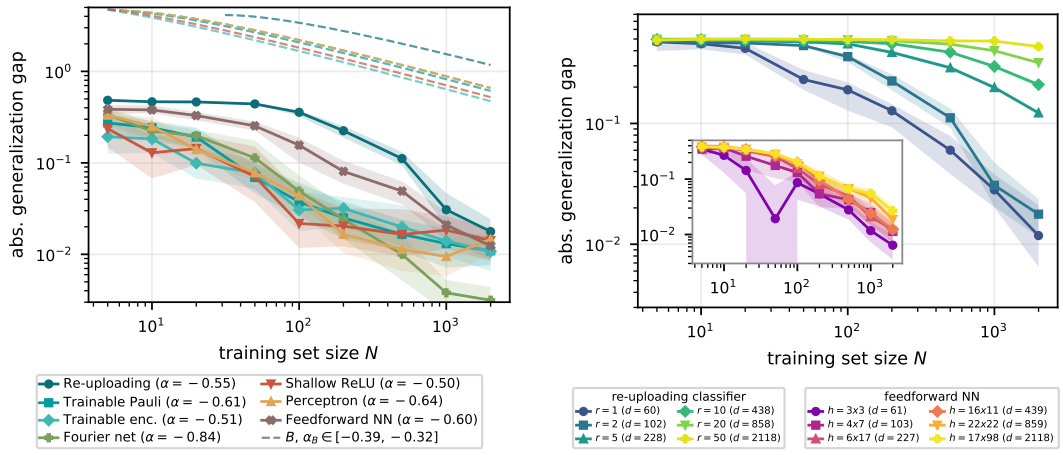
□

4 Empirical Generalization Capability

This section empirically evaluates the generalization ability of QML classifiers and places it within the context of statistical learning theory. Experimental details can be found in Appendix C. The reported accuracy gap $|\mathrm{acc}_{\mathrm{train}} - \mathrm{acc}_{\mathrm{test}}|$ equals the empirical 0-1 risk gap $|\hat{R}_S(h) - \hat{R}_T(h)|$. The hinge loss used during training serves only as a differentiable surrogate for selecting h. The uniform bound over $\mathcal{H}$, applies regardless of the optimization objective. The analysis of the generalization gap as function of training set size for quantum and classical models in Figure 1 reveals three trends predicted by the VC dimension bound in Equations (2) and (3). First, the generalization gap decreases with increasing training set size across model types and capacities. Second, Figure 1a shows that the empirical generalization gap is indeed bounded by the theoretical bound B, which is confirmed for a total of eight datasets and three quantum model classes in Figure 4 in Appendix D.1. However, the VC dimension bound B is rather loose: for low N, the generalization gap is bounded by 0.5 and B is

about one order of magnitude larger. Additionally, the mean slope of the empirical generalization gap is larger in absolute value than that of B which means the empirical generalization gap decreases faster with increasing N. Third, Figure 1b reveals that re-uploading classifiers with larger VC dimension (i.e. more re-uploadings of the data) typically exhibit larger generalization gaps, a trend similar to that found in classical models. This pattern is observed across different re-uploading architectures in Figure 5 in Appendix D.2.

Gil-Fuster et al. (2024) showed that QML models can fit random quantum data up to some N', yielding large generalization gaps and making uniform generalization bounds vacuous for training sizes $N \leq N'$. In contrast, empirical studies have reported that QML models can generalize from very few quantum-data samples (Caro et al., 2022; Kottmann et al., 2021), a phenomenon not captured by such bounds. Our plots in Figures 1a and 4 in Appendix D.1 are consistent with the results by Gil-Fuster et al. (2024) and show that QML models do not always generalize as well as observed by (Caro et al., 2022; Kottmann et al., 2021): Across all eight evaluated datasets, quantum classifiers have large generalization gaps for small N. The train and test accuracies in Appendix D.3 reveal that the re-uploading classifier is more sensitive to overfitting than the feedforward NN. Especially large capacity re-uploading models fit train data perfectly while they behave like a random classifier on test data, the behavior expected when fitting random data. Moreover, the generalization gap bound B or its proxies are above the generalization gap for random guessing, i.e. 0.5, for up to $N = 100$, and for most models and datasets even for up to $N = 2000$, rendering exact bounds uninformative for typical training set sizes currently used in QML.

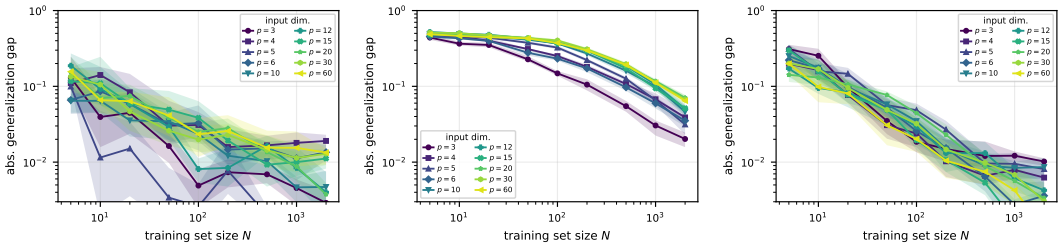


(a) Comparison of quantum and classical models. The bounds B from Equations (2) and (3) are plotted as dashed line. For the perceptron and shallow ReLU network they are exact, for the other models the lower bounds from Table 1 are used as conservative proxies. No bound is shown for the feedforward NN as only the order of its bound is known. The average slope of the curves is given by α. The parameter counts and datasets are listed in Table 4 in Appendix C.3.

(b) Influence of parameter count d. The figure shows the uniform re-uploading classifier and the inset the feedforward neural net with ReLU activation both on census_income. Both models use approximately the same parameter counts d. The high variance of the feedforward NN at $d = 61$ and $N = 50$ is due to nine out of ten trials achieving a train acc. of 50% and only one trial learning from the data.

Figure 1: Absolute generalization gap over training set size. The shaded regions are 95% BCa bootstrap confidence intervals.

The influence of the input dimension p on the absolute generalization gap is evaluated in Figure 2. The re-uploading classifier (Figure 2b) generally exhibits larger empirical train-test gaps for larger input dimensions, whereas the trainable Pauli and feedforward neural net baselines show no such trend. These results suggest that the VC dimension of the uniform re-uploading classifier depends on the input dimension p, although the $(\frac{d}{p} + 1)^p$ upper bound from Theorem 3.3 might be too extreme.



(a) Trainable Pauli, d=20, 7 qubits. (b) Re-uploading classifier, d=198, 6 qubits. (c) Feedforward NN, d=198.

Figure 2: Influence of input dimension p on the absolute generalization gap versus training set size N on the hidden-manifold dataset (m=3).

5 Conclusion

In this work, we derived VC dimension bounds w.r.t. the number of trainable parameters d for several QML model families. For univariate input, these bounds scale linearly with d, matching the behavior of small classical models up to constants. The multivariate bounded-reuse trainable Pauli classifier, on the other hand, admits a $O(d \log d)$ upper bound, similar to the $O(dL \log d)$ scaling of deep feedforward NNs with piecewise-linear activation functions. Moreover, the multivariate uniform re-uploading classifier's upper bound $O((\frac{d}{p}+1)^p)$ has a dependency on input dimension p, which has not been encountered in other model classes. While these upper bounds are not tight, empirical studies of the generalization gap also indicate a p-dependence of the uniform re-uploading classifier's VC dimension. We therefore conclude that parameter efficiency – measured by VC dimension – depends on the data and the model architecture, rather than the classical or quantum nature of the model.

Our experiments show that the empirical generalization behavior of QML models on classical data follows trends predicted by uniform generalization bounds: While VC dimension-based generalization bounds are loose, they consistently bound the empirical generalization gap. Increasing the capacity of the re-uploading classifier reduces bias but leads to larger generalization gaps, while increasing the input dimension produces a trend consistent with the derived capacity bounds.

The strong overfitting of the re-uploading classifier on classical data is consistent with the results of Gil-Fuster et al. (2024) and shows that QML models do not always generalize as well as observed by Caro et al. (2022) and Kottmann et al. (2021). We conclude that the generalization behavior of QML is more nuanced than what uniform generalization bounds predict and that the data has a large influence on it. However, the examples of poor and good generalization on quantum data presented by Gil-Fuster et al. (2024) and by Caro et al. (2022); Kottmann et al. (2021), respectively, show that good generalization of QML models cannot be attributed solely to the use of quantum data. Disentangling which properties of the data-generating process govern generalization and what other factors influence the generalization behavior of QML are important directions for future work.

6 Limitations and Future Directions

The VC dimension bounds for the trainable Pauli classifiers, uniform re-uploading classifier, trainable encoding classifier and Fourier network derived in this work are loose and should be tightened in future work. Second, we use the VC dimension as a proxy for the model's effective parameter usage. However, the VC dimension has its limitations, as it is distribution-free, characterizes entire hypothesis classes rather than individual trained predictors, and is insensitive to the training and the learning process. Nevertheless, VC dimension-based measures have been useful in classical learning theory for explaining qualitative generalization trends, which motivates their use as a capacity proxy in our setting. Third, QML models can generalize well or poorly. Analyzing the factors influencing the generalization behavior of QML models is an important step in understanding their potential. While the data appears to have a significant impact, it might not be the sole variable.

AI USE STATEMENT

In this work, we used generative AI tools to

- create or edit software code.
- describe the experimental details in Appendix C. All other sections of the paper are human-authored.
- assist in the writing of proofs.
- find and summarize related literature.
- format the paper and perform spelling/grammar checks.

We have reviewed all AI-assisted work. AI generated software code has been reviewed; additionally all test cases have been reviewed and a plausibility check of the results has been performed. Proofs have been verified for correctness. All references suggested by AI have been verified by reading the abstract and the essential parts relevant to our work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

Details required to reproduce the experiments are given in Appendix C. Additionally, the source code can be found in the supplementary materials.

REFERENCES

Amira Abbas, David Sutter, Christa Zoufal, Aurelien Lucchi, Alessio Figalli, and Stefan Woerner. The power of quantum neural networks. *Nature Computational Science*, 1(6):403–409, June 2021. ISSN 2662-8457. doi: 10.1038/s43588-021-00084-1. URL http://dx.doi.org/10.1038/s43588-021-00084-1.

Noga Alon and Edward R. Scheinerman. Degrees of freedom versus dimension for containment orders. *Order*, 5(1):11–16, Mar 1988. ISSN 1572-9273. doi: 10.1007/BF00143892. URL https://doi.org/10.1007/BF00143892.

Martin Anthony and Peter L. Bartlett. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, 1999.

Peter L. Bartlett, Nick Harvey, Christopher Liaw, and Abbas Mehrabian. Nearly-tight vc-dimension and pseudodimension bounds for piecewise linear neural networks. *Journal of Machine Learning Research*, 20(63):1–17, 2019. URL http://jmlr.org/papers/v20/17-612.html.

R. Bock. MAGIC Gamma Telescope. UCI Machine Learning Repository, 2004. DOI: https://doi.org/10.24432/C52C8B.

Joseph Bowles, Shahnawaz Ahmed, and Maria Schuld. Better than classical? the subtle art of benchmarking quantum machine learning models, 2024. URL https://arxiv.org/abs/2403.07059.

Kaifeng Bu, Dax Enshan Koh, Lu Li, Qingxian Luo, and Yaobo Zhang. Effects of quantum resources and noise on the statistical complexity of quantum circuits. *Quantum Science and Technology*, 8(2):025013, February 2023. ISSN 2058-9565. doi: 10.1088/2058-9565/acb56a. URL http://dx.doi.org/10.1088/2058-9565/acb56a.

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, November 2021. ISSN 2521-327X. doi: 10.22331/q-2021-11-17-582. URL http://dx.doi.org/10.22331/q-2021-11-17-582.

Matthias C. Caro, Hsin-Yuan Huang, M. Cerezo, Kunal Sharma, Andrew Sornborger, Lukasz Cincio, and Patrick J. Coles. Generalization in quantum machine learning from few training data. *Nature Communications*, 13(1), August 2022. ISSN 2041-1723. doi: 10.1038/s41467-022-32550-3. URL http://dx.doi.org/10.1038/s41467-022-32550-3.

Chih-Chieh Chen, Masaru Sogabe, Kodai Shiba, Katsuyoshi Sakamoto, and Tomah Sogabe. General vapnik–chervonenkis dimension bounds for quantum circuit learning. *Journal of Physics: Complexity*, 3(4):045007, nov 2022. doi: 10.1088/2632-072X/ac9f9b. URL https://doi.org/10.1088/2632-072X/ac9f9b.

Ronald DeVore, Boris Hanin, and Guergana Petrova. Neural network approximation. *Acta Numerica*, 30:327–444, 2021. doi: 10.1017/S0962492921000052.

Michael A. Dritschel and James Rovnyak. The operator fejér-riesz theorem. In Sheldon Axler, Peter Rosenthal, and Donald Sarason (eds.), *A Glimpse at Hilbert Space Operators: Paul R. Halmos in Memoriam*, pp. 223–254. Springer Basel, Basel, 2010. ISBN 978-3-0346-0347-8. doi: 10.1007/978-3-0346-0347-8_14. URL https://doi.org/10.1007/978-3-0346-0347-8_14.

Yuxuan Du, Zhuozhuo Tu, Xiao Yuan, and Dacheng Tao. Efficient measure for the expressivity of variational quantum algorithms. *Physical Review Letters*, 128(8), February 2022. ISSN 1079-7114. doi: 10.1103/physrevlett.128.080506. URL http://dx.doi.org/10.1103/PhysRevLett.128.080506.

R. M. Dudley. Central Limit Theorems for Empirical Measures. *The Annals of Probability*, 6(6): 899 – 929, 1978. doi: 10.1214/aop/1176995384. URL https://doi.org/10.1214/aop/1176995384.

Elies Gil-Fuster, Jens Eisert, and Carlos Bravo-Prieto. Understanding quantum machine learning also requires rethinking generalization. *Nature Communications*, 15(1), March 2024. ISSN 2041-1723. doi: 10.1038/s41467-024-45882-z. URL http://dx.doi.org/10.1038/s41467-024-45882-z.

Casper Gyurik, Dyon Vreumingen, van, and Vedran Dunjko. Structural risk minimization for quantum linear classifiers. *Quantum*, 7:893, January 2023. ISSN 2521-327X. doi: 10.22331/q-2023-01-13-893. URL http://dx.doi.org/10.22331/q-2023-01-13-893.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963. doi: 10.1080/01621459.1963.10500830.

Tak Hur and Daniel K. Park. Understanding generalization in quantum machine learning with margins. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=iQQ2zuWhFM.

Ron Kohavi. Census Income. UCI Machine Learning Repository, 1996. DOI: https://doi.org/10.24432/C5GP7S.

Korbinian Kottmann, Friederike Metz, Joana Fraxanet, and Niccolò Baldelli. Variational quantum anomaly detection: Unsupervised mapping of phase diagrams on a physical quantum computer. *Phys. Rev. Res.*, 3:043184, Dec 2021. doi: 10.1103/PhysRevResearch.3.043184. URL https://link.aps.org/doi/10.1103/PhysRevResearch.3.043184.

Yunchao Liu, Srinivasan Arunachalam, and Kristan Temme. A rigorous and robust quantum speed-up in supervised machine learning. *Nature Physics*, 17(9):1013–1017, July 2021. ISSN 1745-2481. doi: 10.1038/s41567-021-01287-z. URL http://dx.doi.org/10.1038/s41567-021-01287-z.

M. Mohri, A. Rostamizadeh, and A. Talwalkar. *Foundations of Machine Learning*. Adaptive Computation and Machine Learning series. MIT Press, 2012. ISBN 9780262018258.

Tiago Novello, Diana Aldana, Andre Araujo, and Luiz Velho. Tuning the frequencies: Robust training for sinusoidal neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3071–3080, June 2025.

Giambattista Parascandolo, Heikki Huttunen, and Tuomas Virtanen. Taming the waves: sine as activation function in deep neural networks, 2017. URL https://openreview.net/forum?id=Sks3zF9eg.

Adrián Pérez-Salinas, Alba Cervera-Lierta, Elies Gil-Fuster, and José I. Latorre. Data re-uploading for a universal quantum classifier. *Quantum*, 4:226, February 2020. ISSN 2521-327X. doi: 10.22331/q-2020-02-06-226. URL https://doi.org/10.22331/q-2020-02-06-226.

Evan Peters and Maria Schuld. Generalization despite overfitting in quantum machine learning models. *Quantum*, 7:1210, December 2023. ISSN 2521-327X. doi: 10.22331/q-2023-12-20-1210. URL http://dx.doi.org/10.22331/q-2023-12-20-1210.

M. J. D. Powell. *Approximation Theory and Methods*. Cambridge University Press, 1981.

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), March 2021. ISSN 2469-9934. doi: 10.1103/physreva.103.032430. URL http://dx.doi.org/10.1103/PhysRevA.103.032430.

Oved Shisha. Tchebycheff systems and best partial bases. *Pacific Journal of Mathematics*, 86(2): 563–580, 1980.

Vladimir N. Vapnik. *The Nature of Statistical Learning Theory*, chapter 3, pp. 80–83. Springer New York, 2 edition, 1999.

Hugh E. Warren. Lower bounds for approximation by nonlinear manifolds. *Transactions of the American Mathematical Society*, 133(1):167–178, 1968. ISSN 00029947. URL http://www.jstor.org/stable/1994937.

Maximilian Wendlinger, Kilian Tscharke, and Pascal Debus. A comparative analysis of adversarial robustness for quantum and classical machine learning models. In *2024 IEEE International Conference on Quantum Computing and Engineering (QCE)*, volume 01, pp. 1447–1457, 2024. doi: 10.1109/QCE60285.2024.00171.

Zhan Yu, Hongshun Yao, Mujin Li, and Xin Wang. Power and limitations of single-qubit native quantum neural networks. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 27810–27823. Curran Associates, Inc., 2022.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=Sy8gdB9xx.

Haimeng Zhao, Alexander Zlokapa, Hartmut Neven, Ryan Babbush, John Preskill, Jarrod R. McClean, and Hsin-Yuan Huang. Exponential quantum advantage in processing massive classical data, 2026. URL https://arxiv.org/abs/2604.07639.

A Proofs for Quantum Models

This appendix contains the proofs for the VC dimension bounds of the quantum models in Section 3. First, Appendix A.1 derives an upper bound for the VC dimension of the trainable Pauli classifier. Second, Appendix A.2 gives the exact VC dimension of the univariate n-qubit re-uploading classifier. Third, Appendix A.3 gives VC bounds of the multivariate uniform re-uploading classifier based on the number of re-uploadings, which is ultimately converted to a d-dependent bound in Theorem 3.3 and Appendix A.2. Third, linear lower and upper VC dimension bounds of the univariate n-qubit trainable-encoding classifier are given in Appendix A.4. Finally, linear VC dimension bounds are provided for the univariate shallow Fourier network in Appendix A.5.

A.1 VC Dimension of Trainable Pauli Classifier

The trainable Pauli classifier is a generic PQC of form $f_{\boldsymbol{\theta}}^{A}(\mathbf{x}) = \langle 0|^{\otimes n} U_A^{\dagger}(\mathbf{x};\boldsymbol{\theta})\, \mathcal{M}_A\, U_A(\mathbf{x};\boldsymbol{\theta})\, |0\rangle^{\otimes n}$ where the architecture A is restricted such that every trainable gate in the unitary $U_A(\mathbf{x};\boldsymbol{\theta})$ is generated by an n-qubit Pauli-word,

$$R_P(\theta_j) = e^{-i\theta_j P/2}$$

with

$$P^2 = I,\ P = P_1 \otimes \cdots \otimes P_n,\ P_k \in \{I, X, Y, Z\}.$$

$\mathcal{M}_A$ is a fixed Hermitian observable. Additionally, all remaining gates are independent of $\boldsymbol{\theta}$. The total number of parameters is d and an individual parameter θ_j occurs q_j times. We define $Q := \sum_{j=1}^{d} q_j$ as the total parameter count including parameter re-use. The classifier makes no assumptions about the data encoding.

We first show in Proposition A.1 that no universal positive lower VC bound can exist. Next, Lemma A.2 and Proposition A.3 reveals that the circuit output is a polynomial in the sines and cosines of the trainable parameters, with degree controlled by their number of occurrences. Then, Lemma A.4 and Proposition A.5 converts this representation on each parameter-space chart into an ordinary polynomial in d real variables. Using the polynomial sign-pattern bound, Proposition A.6 bounds the number of realizable labelings. Combining this result with Lemma A.7 yields the VC-dimension bound in Theorem 3.1.

Proposition A.1 (No universal positive VC lower bound)**.** *For every $d \geq 1$ and every choice of integers $1 \leq q_1, \ldots, q_d \leq q_{\max} = O(1)$, there exists some architecture A' for the bounded-reuse trainable Pauli classifier, such that*

$$\mathrm{VCdim}(\mathcal{H}^{A'}_{\mathrm{bound-pauli}}) = 0.$$

From $\mathcal{H}^{A}_{\mathrm{bound-pauli}} \subset \mathcal{H}^{A}_{\mathrm{pauli}}$ follows

$$\mathrm{VCdim}(\mathcal{H}^{A'}_{\mathrm{pauli}}) = 0.$$

Proof. Consider one qubit, an input $x \in [0, 2\pi]$, the observable $\mathcal{M}_A = Z$, and the circuit

$$U_A(x; \boldsymbol{\theta}) = \left(\prod_{j=1}^{d} R_z(\theta_j)^{q_j} \right) R_z(x).$$

Each θ_j occurs in exactly q_j trainable Pauli rotations, while $R_z(x)$ is independent of $\boldsymbol{\theta}$. Since $R_z(\varphi)|0\rangle = e^{-i\varphi/2}|0\rangle$, we have

$$U_A(x; \boldsymbol{\theta})|0\rangle = e^{-\frac{i}{2}(x + \sum_{j=1}^{d} q_j \theta_j)}|0\rangle.$$

This differs from $|0\rangle$ only by a global phase. Hence $f^A_{\boldsymbol{\theta}}(x) = \langle Z \rangle = 1$ for all x and $\boldsymbol{\theta}$. The hypothesis class contains only the constant $+1$ function and therefore has VC dimension zero. □

Lemma A.2 (Conjugation by a Pauli rotation)**.** *Let $R_P(\theta)$ be a trainable gate as defined in Appendix A.1. Then, for every operator B,*

$$R_P(\theta)^\dagger B R_P(\theta) = \frac{B + PBP}{2} + \frac{\cos\theta}{2}(B - PBP) + \frac{i \sin\theta}{2}(PB - BP).$$

In particular, conjugation by $R_P(\theta)$ results in a linear combination of 1, $\cos\theta$, and $\sin\theta$.

Proof. Since $P^2 = I$, we have

$$R_P(\theta) = \cos\left(\frac{\theta}{2}\right) I - i \sin\left(\frac{\theta}{2}\right) P.$$

Writing

$$c = \cos\left(\frac{\theta}{2}\right), \qquad s = \sin\left(\frac{\theta}{2}\right),$$

gives

$$\begin{aligned} R_P(\theta)^\dagger B R_P(\theta) &= (cI + isP) B (cI - isP) \\ &= c^2 B + s^2 PBP + ics(PB - BP). \end{aligned}$$

Using

$$c^2 = \frac{1 + \cos\theta}{2}, \qquad s^2 = \frac{1 - \cos\theta}{2}, \qquad 2cs = \sin\theta,$$

yields the stated identity. □

Proposition A.3 (Polynomial dependence on trainable parameters)**.** *Consider the trainable Pauli classifier from above. Then, for every fixed input* $\mathbf{x}$*, there exists a real polynomial* $P_{\mathbf{x}}$ *such that*

$$f_{\boldsymbol{\theta}}^{A}(\mathbf{x}) = P_{\mathbf{x}}\left(\cos\theta_1, \sin\theta_1, \ldots, \cos\theta_d, \sin\theta_d\right).$$

Moreover, the total (i.e. combined) degree in the two variables $\cos\theta_j$ *and* $\sin\theta_j$*,*

$$\deg_{(\cos\theta_j, \sin\theta_j)} P_{\mathbf{x}} \leq q_j$$

for every j*, and*

$$\deg P_{\mathbf{x}} \leq Q.$$

Proof. Fix $\mathbf{x}$, so all parameter-independent gates become fixed unitaries. We propagate the observable $\mathcal{M}_A$ backwards through the circuit $\langle 0|^{\otimes n} U_A^{\dagger}(\mathbf{x};\boldsymbol{\theta})\, \mathcal{M}_A\, U_A(\mathbf{x};\boldsymbol{\theta})\, |0\rangle^{\otimes n}$.

Conjugation by a gate that is independent of $\boldsymbol{\theta}$ acts linearly on the observable and does not change its dependence on the trainable parameters. By Lemma A.2, conjugation by one occurrence of $R_P(\theta_j)$ replaces every operator coefficient by a linear combination of terms multiplied by

$$1, \qquad \cos\theta_j, \qquad \sin\theta_j.$$

Thus, each occurrence of θ_j increases the degree in the variable pair $(\cos\theta_j, \sin\theta_j)$ by at most one.

Induction over the trainable-gate occurrences therefore shows that, after all q_j occurrences of θ_j, the degree in the j-th variable pair is at most q_j. Summing over all parameter occurrences gives total degree at most

$$\sum_{j=1}^{d} q_j = Q.$$

The propagated observable is Hermitian at every step. Its expectation value in the fixed initial state is therefore real, so the resulting polynomial $P_{\mathbf{x}}$ may be chosen to have real coefficients. □

Lemma A.4 (Rational charts of the circle)**.** *For* $\varepsilon \in \{-1, +1\}$*, define* $\phi_\varepsilon : \mathbb{R} \to S^1$ *by*

$$\phi_\varepsilon(t) = \left(\frac{\varepsilon(1-t^2)}{1+t^2}, \frac{2t}{1+t^2}\right).$$

Then

$$S^1 = \phi_{+1}(\mathbb{R}) \cup \phi_{-1}(\mathbb{R}).$$

Consequently, the torus $(S^1)^d$ *is covered by* 2^d *rational charts indexed by* $\boldsymbol{\varepsilon} \in \{-1, +1\}^d$*.*

Proof. For every $t \in \mathbb{R}$ and $\varepsilon \in \{-1, +1\}$,

$$\left(\frac{\varepsilon(1-t^2)}{1+t^2}\right)^2 + \left(\frac{2t}{1+t^2}\right)^2 = \frac{(1-t^2)^2 + 4t^2}{(1+t^2)^2} = \frac{(1+t^2)^2}{(1+t^2)^2} = 1.$$

Hence, $\phi_\varepsilon(t) \in S^1$.

For every $(u, v) \in S^1 \setminus \{(-1, 0)\}$, the coordinate

$$t = \frac{v}{1+u}$$

satisfies

$$\phi_{+1}(t) = (u, v).$$

Therefore,

$$\phi_{+1}(\mathbb{R}) = S^1 \setminus \{(-1, 0)\}.$$

Similarly, for every $(u, v) \in S^1 \setminus \{(1, 0)\}$, the coordinate

$$t = \frac{v}{1-u}$$

satisfies

$$\phi_{-1}(t) = (u, v).$$

Thus,

$$\phi_{-1}(\mathbb{R}) = S^1 \setminus \{(1, 0)\}.$$

The two images therefore cover the circle:

$$S^1 = \phi_{+1}(\mathbb{R}) \cup \phi_{-1}(\mathbb{R}).$$

Taking Cartesian products, one independently chooses one of the two circle charts for each of the d factors. Hence, the torus $(S^1)^d$ is covered by 2^d rational product charts indexed by

$$\boldsymbol{\varepsilon} \in \{-1, +1\}^d.$$

□

Proposition A.5 (Chartwise polynomialization)**.** *Let $P_{\mathbf{x}}$ be the polynomial from Proposition A.3. For a chart $\boldsymbol{\varepsilon} \in \{-1, +1\}^d$, set*

$$u_j(t_j) = \frac{\varepsilon_j(1 - t_j^2)}{1 + t_j^2}, \qquad v_j(t_j) = \frac{2t_j}{1 + t_j^2}.$$

Define

$$\begin{aligned}\widetilde{P}_{\mathbf{x},\boldsymbol{\varepsilon}}(\mathbf{t}) :=& P_{\mathbf{x}}\big(u_1(t_1), v_1(t_1), \ldots, u_d(t_d), v_d(t_d)\big) \\ &\times \prod_{j=1}^{d} (1 + t_j^2)^{q_j}.\end{aligned}$$

Then $\widetilde{P}_{\mathbf{x},\boldsymbol{\varepsilon}}$ is a real polynomial in d real variables satisfying

$$\deg_{t_j} \widetilde{P}_{\mathbf{x},\boldsymbol{\varepsilon}} \le 2q_j$$

and

$$\deg \widetilde{P}_{\mathbf{x},\boldsymbol{\varepsilon}} \le 2Q.$$

Furthermore,

$$\operatorname{sgn} \widetilde{P}_{\mathbf{x},\boldsymbol{\varepsilon}}(\mathbf{t}) = \operatorname{sgn} P_{\mathbf{x}}\big(u_1(t_1), v_1(t_1), \ldots, u_d(t_d), v_d(t_d)\big).$$

Proof. By Proposition A.3, the degree of $P_{\mathbf{x}}$ in the variable pair (u_j, v_j) is at most q_j. Substituting the rational expressions for u_j and v_j therefore produces a denominator dividing

$$\prod_{j=1}^{d} (1 + t_j^2)^{q_j}.$$

Multiplication by this product clears all denominators, so $\widetilde{P}_{\mathbf{x},\boldsymbol{\varepsilon}}$ is a real polynomial. To see the degree bound, consider a monomial containing $u_j^{a_j} v_j^{b_j}$, where $a_j + b_j \le q_j$. After substitution and multiplication by $(1 + t_j^2)^{q_j}$, its t_j-dependent part is

$$(1 - t_j^2)^{a_j} (2t_j)^{b_j} (1 + t_j^2)^{q_j - a_j - b_j}.$$

Its degree is at most

$$2a_j + b_j + 2(q_j - a_j - b_j) = 2q_j - b_j \le 2q_j.$$

Hence,

$$\deg_{t_j} \widetilde{P}_{\mathbf{x},\boldsymbol{\varepsilon}} \le 2q_j.$$

Summing these bounds over j yields

$$\deg \widetilde{P}_{\mathbf{x},\boldsymbol{\varepsilon}} \le 2 \sum_{j=1}^{d} q_j = 2Q.$$

Finally,

$$\prod_{j=1}^{d}(1+t_j^2)^{q_j} > 0$$

for every $\mathbf{t} \in \mathbb{R}^d$. Clearing the denominators therefore does not change the sign. □

Proposition A.6 (Growth-function bound)**.** *Let $\mathcal{H}^A_{\text{pauli}}$ be the thresholded hypothesis class of the trainable Pauli classifier defined in Equation* (1)*. Under the assumptions of Proposition A.3, the growth function Π of $\mathcal{H}^A_{\text{pauli}}$ satisfies, for $m \geq d$,*

$$\Pi_{\mathcal{H}^A_{\text{pauli}}}(m) \leq \left(\frac{32eQm}{d}\right)^d.$$

Proof. Fix arbitrary inputs

$$\mathbf{x}_1, \ldots, \mathbf{x}_m.$$

On each rational chart of the parameter torus, Proposition A.5 represents their classifications through the signs of the m real polynomials

$$\widetilde{P}_{\mathbf{x}_1,\boldsymbol{\varepsilon}}, \ldots, \widetilde{P}_{\mathbf{x}_m,\boldsymbol{\varepsilon}}$$

in d real variables, each of degree at most $2Q$.

For $m \geq d$, the polynomial sign-pattern bound implies that m real polynomials of degree at most D in d variables realize at most

$$\left(\frac{8eDm}{d}\right)^d$$

sign conditions in $\{-1, 0, +1\}^m$ Warren (1968); Alon & Scheinerman (1988). With $D = 2Q$, the number of sign conditions on one chart is at most

$$\left(\frac{16eQm}{d}\right)^d.$$

Under the adopted binary sign convention, every binary labeling is determined by one of these ternary sign conditions: both 0 and $+1$ correspond to the positive class. Hence the number of binary labelings is no larger than the number of ternary sign conditions.

By Lemma A.4, the parameter torus is covered by 2^d charts. Summing the chartwise bounds gives

$$\begin{aligned}\Pi_{\mathcal{H}^A_{\text{pauli}}}(m) &\leq 2^d \left(\frac{16eQm}{d}\right)^d \\ &= \left(\frac{32eQm}{d}\right)^d.\end{aligned}$$

Since the inputs were arbitrary, the result holds for the growth function. □

Lemma A.7 (Exponential inequality)**.** *Let $B \geq 2$ and $t > 0$. If*

$$2^t \leq Bt,$$

then

$$t < 2\log_2(2B).$$

Proof. Set

$$t_0 = 2\log_2(2B).$$

Since $B \geq 2$, we have $t_0 \geq 4$. Moreover,

$$2^{t_0} = (2B)^2 = 4B^2,$$

whereas

$$Bt_0 = 2B\log_2(2B).$$

Since

$$\log_2(2B) < 2B$$

for $B \geq 2$, it follows that

$$2^{t_0} > Bt_0.$$

The function $t \mapsto 2^t/t$ is strictly increasing for $t > 1/\ln 2$. Therefore,

$$2^t > Bt$$

for every $t \geq t_0$. Taking the contrapositive proves that $2^t \leq Bt$ implies $t < t_0$. □

Theorem A.8 (Restatement of Theorem 3.1: VC upper bound of trainable Pauli classifier)**.** *Let $\mathcal{H}^A_{\text{pauli}}$ be the hypothesis class of the trainable Pauli classifier defined in Section 2.3.1 with fixed architecture A. Then an upper bound of the VC dimension is given by*

$$\boxed{\text{VCdim}(\mathcal{H}^A_{\text{pauli}}) < 2d\log_2(64eQ).}$$

For the bounded-reuse trainable Pauli classifier with $Q = O(d)$, we obtain

$$\boxed{\text{VCdim}(\mathcal{H}^A_{\text{bound-pauli}}) \leq O(d\log d)}$$

Proof. By Lemma A.2 and Proposition A.3, for every fixed input $\mathbf{x}$, the circuit output can be written as a real polynomial

$$f_{\boldsymbol{\theta}}(\mathbf{x}) = P_{\mathbf{x}}\left(\cos\theta_1, \sin\theta_1, \ldots, \cos\theta_d, \sin\theta_d\right),$$

whose degree in each variable pair $(\cos\theta_j, \sin\theta_j)$ is at most q_j, and whose total degree is at most

$$Q = \sum_{j=1}^{d} q_j.$$

The parameter vector

$$(\cos\theta_1, \sin\theta_1, \ldots, \cos\theta_d, \sin\theta_d)$$

lies on the torus $(S^1)^d$. By Lemma A.4, this torus is covered by 2^d rational charts. On each chart, Proposition A.5 converts $P_{\mathbf{x}}$ into a real polynomial in d real variables of degree at most $2Q$, while preserving its sign. Applying the polynomial sign-pattern bound over all 2^d charts therefore gives, by Proposition A.6, for $m \geq d$,

$$\Pi_{\mathcal{H}_{\text{pauli}}}(m) \leq \left(\frac{32eQm}{d}\right)^d.$$

Now let $m \geq d$ points be shattered by $\mathcal{H}_{\text{pauli}}$. Shattering requires all 2^m binary labelings to be realizable, and hence

$$2^m \leq \Pi_{\mathcal{H}_{\text{pauli}}}(m) \leq \left(\frac{32eQm}{d}\right)^d.$$

Taking the d-th root and setting

$$t := \frac{m}{d}$$

gives

$$2^t \leq 32eQ\, t.$$

Since $Q \geq 1$, the choice $A = 32eQ$ satisfies the assumptions of Lemma A.7. Thus,

$$t < 2\log_2(64eQ),$$

and therefore

$$m < 2d\log_2(64eQ).$$

If $m < d$, the same conclusion follows immediately since $Q \geq 1$ implies

$$m < d < 2d\log_2(64eQ).$$

Hence no set of size $m \geq 2d\log_2(64eQ)$ can be shattered, and consequently

$$\text{VCdim}(\mathcal{H}_{\text{pauli}}) < 2d\log_2(64eQ).$$

□

This upper bound holds for every A satisfying the assumptions in Section 2.3.1 and Section 2.3.2, respectively. As stated in Proposition A.1, no universal positive lower bound can exist for all A. Nevertheless, for $n = 1$ we have $\mathcal{F}^A_{\mathrm{1p-reup}} \subset \mathcal{F}^A_{\mathrm{bound-pauli}} \subset \mathcal{F}^A_{\mathrm{pauli}}$ and hence, $\mathrm{VCdim}(\mathcal{H}^{A'}_{\mathrm{pauli}}) = \mathrm{VCdim}(\mathcal{H}^{A'}_{\mathrm{bound-pauli}}) = \frac{2}{3}d - 1$ for some A', meaning linear VC dimension is achievable within this circuit family.

A.2 VC Dimension of Univariate n-Qubit Re-Uploading Classifier

We first establish the VC dimension of real trigonometric polynomials in Appendix A.2.1. We then show that every single-qubit branch of the n-qubit re-uploading classifier realizes exactly the bounded trigonometric polynomials of degree at most r. Using the structure of the entangling CNOT ring, we express the n-qubit model output as a product of these single-qubit functions. This yields a trigonometric degree of at most nr and hence an upper bound on the VC dimension. Finally, we construct a factorized trigonometric polynomial realizing arbitrary labelings of $2nr + 1$ points, which gives the matching lower bound and the exact result in Theorem 3.2.

A.2.1 VC Dimension of Bounded Trigonometric Polynomials

This section establishes the VC dimension of thresholded real trigonometric polynomials by first characterizing the structure of T_r via Lemmas A.9 and A.10, then showing exact interpolation through Proposition A.11, which yields the lower bound in Proposition A.12, while Proposition A.13 provides the matching upper bound; combining both results gives Proposition A.14.

Let

$$T_r := \left\{ a_0 + \sum_{n=1}^{r} a_n \cos(nx) + b_n \sin(nx) : \ a_0, a_n, b_n \in \mathbb{R} \right\}$$

be the space of real trigonometric polynomials of degree at most r.

Let further $C \in (0, \infty]$, and define

$$T_r^{(C)} := \{ g \in T_r : \|g\|_\infty \leq C \}$$

as the class of bounded real trigonometric polynomials of degree at most r.

Lemma A.9 (Dimension of T_r)**.** *The space T_r is a real vector space of dimension $2r + 1$.*

Proof. A basis is given by

$$1, \ \cos(x), \ \sin(x), \ \ldots, \ \cos(rx), \ \sin(rx).$$

□

Lemma A.10 (Zero bound)**.** *If $t \in T_r$ is nonzero, then it has at most $2r$ distinct zeros in $[0, 2\pi)$.*

Proof. This is a standard property of trigonometric polynomials, see e.g. (Powell, 1981). □

Proposition A.11 (Trigonometric interpolation)**.** *Let $m := 2r + 1$, and let $x_1, \ldots, x_m \in [0, 2\pi)$ be distinct. Then the evaluation map*

$$E : T_r \to \mathbb{R}^m, \qquad E(t) = (t(x_1), \ldots, t(x_m))$$

is a linear isomorphism. Equivalently, for every $(y_1, \ldots, y_m) \in \mathbb{R}^m$, there exists a unique $t \in T_r$ such that

$$t(x_i) = y_i, \qquad i = 1, \ldots, m.$$

Proof. By Lemma A.9, both T_r and $\mathbb{R}^m$ have dimension $2r + 1$. It is therefore enough to show that E is injective, which is true if the kernel is trivial. If $E(t) = 0$, then t vanishes at the $2r + 1$ distinct points $x_1, \ldots, x_m$. By Lemma A.10, this forces $t \equiv 0$. So E is injective, and hence bijective. □

Proposition A.12 (VC lower bound for bounded trigonometric polynomials)**.** *For every $C \in (0, \infty]$,*

$$\mathrm{VCdim}(\mathcal{H}_r^{(C)}) \geq 2r + 1.$$

Proof. Let $m := 2r + 1$, and choose distinct points $x_1, \dots, x_m \in [0, 2\pi)$. We show that $\{x_1, \dots, x_m\}$ is shattered.

Fix arbitrary labels $y_1, \dots, y_m \in \{-1, +1\}$. By Proposition A.11, there exists $u \in T_r$ such that

$$u(x_i) = y_i, \qquad i = 1, \dots, m.$$

If $C = \infty$, then $u \in T_r^{(C)}$, so the labeling is realized.

If $C < \infty$, define

$$A := \max\{C, \|u\|_\infty\}, \qquad g := \frac{C}{A} u.$$

Then $g \in T_r$ and

$$\|g\|_\infty = \frac{C}{A} \|u\|_\infty \le C,$$

so $g \in T_r^{(C)}$. Since $C/A > 0$, we also have

$$\operatorname{sgn}(g(x_i)) = \operatorname{sgn}(u(x_i)) = y_i, \qquad i = 1, \dots, m.$$

Thus every labeling of $\{x_1, \dots, x_m\}$ is realized by $\mathcal{H}_r^{(C)}$. Hence

$$\operatorname{VCdim}(\mathcal{H}_r^{(C)}) \ge 2r + 1.$$

□

Proposition A.13 (VC upper bound for bounded trigonometric polynomials)**.** *For every $C \in (0, \infty]$, the hypothesis class $\mathcal{H}_r^{(C)}$ satisfies*

$$\operatorname{VCdim}(\mathcal{H}_r^{(C)}) \le \operatorname{VCdim}(\mathcal{H}_r) \le 2r + 1.$$

Proof. The first inequality follows from $\mathcal{H}_r^{(C)} \subseteq \mathcal{H}_r$. Every hypothesis in $\mathcal{H}_r$ is the sign of a function from the $(2r + 1)$-dimensional vector space T_r. Dudley's theorem for thresholded finite-dimensional real vector spaces (Anthony & Bartlett, 1999; Dudley, 1978) implies that such a class has VC dimension at most the dimension of the underlying space. Using Lemma A.9, we obtain

$$\operatorname{VCdim}(\mathcal{H}_r) \le \dim T_r = 2r + 1.$$

□

Proposition A.14 (VC dimension of bounded trigonometric polynomials)**.** *For every $C \in (0, \infty]$,*

$$\operatorname{VCdim}(\mathcal{H}_r^{(C)}) = 2r + 1.$$

Proof. Combine Propositions A.12 and A.13. □

A.2.2 Fourier Characterization of a Single Re-Uploading Branch

This section shows that the function class realized by one single-qubit branch of the univariate n-qubit re-uploading classifier is independent of the choices of Pauli encoding axes and coincides with the class of bounded real trigonometric polynomials $T_r^{(1)}$ defined in Appendix A.2.1

For a fixed qubit j, denote its re-uploading unitary by

$$V_j(x; \boldsymbol{\theta}_j) = W_j^{(r)} R_{a_{j,r-1}}(x) W_j^{(r-1)} \cdots W_j^{(1)} R_{a_{j,0}}(x) W_j^{(0)},$$

and its corresponding single-qubit function class by

$$\mathcal{F}_{\text{branch}}^{(r)} = \left\{ x \mapsto \langle 0 | V_j^\dagger(x; \boldsymbol{\theta}_j) Z V_j(x; \boldsymbol{\theta}_j) | 0 \rangle \;\middle|\; \boldsymbol{\theta}_j \in [0, 2\pi]^{3(r+1)} \right\}.$$

Lemma A.15 (Invariance under fixed Pauli encoding axes)**.** *Let $a_0, \dots, a_{r-1} \in \{x, y, z\}$ be arbitrary fixed encoding axes. The set of unitaries realized by*

$$W^{(r)} R_{a_{r-1}}(x) W^{(r-1)} \cdots W^{(1)} R_{a_0}(x) W^{(0)},$$

where all $W^{(s)}$ are general single-qubit rotations, is identical to the set realized when every encoding gate is $R_z(x)$.

Proof. For every $a \in \{x, y, z\}$, choose a fixed single-qubit unitary S_a satisfying

$$S_a Z S_a^\dagger = \sigma_a.$$

Then

$$R_a(x) = S_a R_z(x) S_a^\dagger.$$

Writing $S_\ell := S_{a_\ell}$, Lemma A.15 becomes

$$\begin{aligned} &W^{(r)} S_{r-1} R_z(x) S_{r-1}^\dagger W^{(r-1)} S_{r-2} R_z(x) \cdots S_0^\dagger W^{(0)} \\ &\quad = \widetilde{W}^{(r)} R_z(x) \widetilde{W}^{(r-1)} \cdots R_z(x) \widetilde{W}^{(0)}, \end{aligned} \tag{11}$$

where

$$\widetilde{W}^{(r)} := W^{(r)} S_{r-1}, \tag{12}$$

$$\widetilde{W}^{(s)} := S_s^\dagger W^{(s)} S_{s-1}, \qquad s = 1, \ldots, r-1, \tag{13}$$

$$\widetilde{W}^{(0)} := S_0^\dagger W^{(0)}. \tag{14}$$

Since $W^{(s)}$ can realize any single-qubit unitary, so can $\widetilde{W}^{(s)}$. Hence the fixed basis rotations can be absorbed into the trainable layers without changing the realizable circuit family. □

Proposition A.16 (Re-uploading branch outputs are bounded trigonometric polynomials)**.** *The single-branch function class satisfies*

$$\mathcal{F}_{\text{branch}}^{(r)} \subseteq T_r^{(1)}.$$

Proof. Each branch contains exactly r Pauli encoding gates $R_{a_{j,\ell}}(x)$. By Equation (6), every resulting model output has Fourier support contained in $\{-r, \ldots, r\}$, independently of the fixed choices of Pauli axes. Since the model output is real-valued, it therefore belongs to T_r.

Moreover, the output is the expectation value of the Pauli observable Z, whose spectrum is $\{\pm 1\}$. Hence

$$|f(x)| \le 1$$

for every x, and therefore $\|f\|_\infty \le 1$. □

Lemma A.17 (Fejér–Riesz factorization for the shifted polynomial)**.** *Let $g \in T_r$ satisfy $\|g\|_\infty \le 1$, and define*

$$q(x) := \frac{1 + g(x)}{2}.$$

Then q is a nonnegative trigonometric polynomial of degree at most r. Moreover, there exists a polynomial

$$p(z) = \sum_{k=0}^{r} \alpha_k z^k$$

such that

$$q(x) = |p(e^{ix})|^2.$$

Proof. Since $g \in T_r$, the function $q = (1 + g)/2$ is again a trigonometric polynomial of degree at most r. The bound $\|g\|_\infty \le 1$ implies

$$0 \le q(x) = \frac{1 + g(x)}{2} \le 1 \qquad \text{for all } x.$$

Hence q is nonnegative on the unit circle. The Fejér–Riesz theorem (Dritschel & Rovnyak, 2010) therefore yields a polynomial $p(z) = \sum_{k=0}^{r} \alpha_k z^k$ such that

$$q(x) = |p(e^{ix})|^2.$$

□

Proposition A.18 (Laurent lift and parity)**.** *Under the assumptions of Lemma A.17, define*

$$P(x) := e^{-irx/2}p(e^{ix}).$$

Then $P \in \mathbb{C}[e^{ix/2}, e^{-ix/2}]$, $\deg(P) \le r$, P *has parity* $r \bmod 2$*, and*

$$|P(x)|^2 = \frac{1+g(x)}{2} \le 1 \qquad \textit{for all } x.$$

Proof. Using $p(z) = \sum_{k=0}^{r} \alpha_k z^k$, we obtain

$$P(x) = e^{-irx/2}p(e^{ix}) = \sum_{k=0}^{r} \alpha_k e^{i(2k-r)x/2}.$$

Hence P is a Laurent polynomial in $e^{ix/2}$ and $e^{-ix/2}$ with exponents in $\{-r, -r+2, \dots, r-2, r\}$. Therefore $\deg(P) \le r$, and all exponents have parity $r \bmod 2$, which proves the parity statement. Finally,

$$|P(x)|^2 = |e^{-irx/2}|^2\,|p(e^{ix})|^2 = |p(e^{ix})|^2 = \frac{1+g(x)}{2},$$

since $|e^{-irx/2}| = 1$. □

Proposition A.19 (Realization by a single re-uploading branch)**.** *Let* $P \in \mathbb{C}[e^{ix/2}, e^{-ix/2}]$ *satisfy* $\deg(P) \le r$*, parity* $r \bmod 2$*, and* $|P(x)| \le 1$ *for all* x*. Then there exists a parameter choice for a single re-uploading branch such that its unitary satisfies*

$$V(x) = \begin{pmatrix} P(x) & -Q(x) \\ Q^*(x) & P^*(x) \end{pmatrix}$$

for some $Q \in \mathbb{C}[e^{ix/2}, e^{-ix/2}]$ *with* $\deg(Q) \le r$*, the same parity, and*

$$|P(x)|^2 + |Q(x)|^2 = 1.$$

Consequently,

$$\langle 0|\, V^\dagger(x) Z V(x)\, |0\rangle = 2|P(x)|^2 - 1.$$

Proof. By Lemma 3 of Yu et al. (2022), a single-qubit re-uploading circuit with r encoding layers and $R_z(x)$ encoding realizes exactly the unitaries

$$\begin{pmatrix} P(x) & -Q(x) \\ Q^*(x) & P^*(x) \end{pmatrix},$$

where

$$\deg(P), \deg(Q) \le r, \qquad \mathrm{parity}(P) = \mathrm{parity}(Q) = r \bmod 2,$$

and

$$|P(x)|^2 + |Q(x)|^2 = 1.$$

The condition $|P(x)| \le 1$ guarantees the existence of such a Q. By Lemma A.15, the same set of unitaries is realizable for any fixed sequence of Pauli encoding axes.

Finally,

$$V(x)\,|0\rangle = \begin{pmatrix} P(x) \\ Q^*(x) \end{pmatrix},$$

and hence

$$\langle 0|\, V^\dagger Z V\, |0\rangle = |P(x)|^2 - |Q(x)|^2 = 2|P(x)|^2 - 1.$$

□

Lemma A.20 (Characterization of a re-uploading branch)**.** *The function class of a single re-uploading branch satisfies*

$$\mathcal{F}^{(r)}_{\mathrm{branch}} = T^{(1)}_r.$$

Proof. We prove both inclusions.

(i) Inclusion $\mathcal{F}^{(r)}_{\text{branch}} \subseteq T^{(1)}_r$. This follows from Proposition A.16.

(ii) Inclusion $T^{(1)}_r \subseteq \mathcal{F}^{(r)}_{\text{branch}}$. Let $g \in T_r$ with $\|g\|_\infty \le 1$ and define

$$q(x) := \frac{1+g(x)}{2}.$$

By Lemma A.17, there exists $p(z) = \sum_{k=0}^{r} \alpha_k z^k$ such that

$$q(x) = |p(e^{ix})|^2.$$

By Proposition A.18, this yields a Laurent polynomial $P(x)$ of degree at most r and parity $r \bmod 2$ satisfying

$$|P(x)|^2 = \frac{1+g(x)}{2}.$$

Finally, Proposition A.19 shows that this P is realizable by the branch and gives the output

$$f(x) = 2|P(x)|^2 - 1 = g(x).$$

Thus every $g \in T^{(1)}_r$ is realizable. □

A.2.3 Exact VC Dimension of the n-Qubit Re-Uploading Classifier

For fixed $n \ge 1$, re-uploading count $r \ge 1$, and input dimension $p = 1$, the univariate n-qubit re-uploading classifier is defined as

$$\mathcal{F}^{(n)}_{\text{1p}-\text{reup}} := \Big\{ f_{\boldsymbol{\theta}} : [0,2\pi] \to [-1,1] \;\Big|\; f_{\boldsymbol{\theta}}(x) = \langle 0|^{\otimes n} U^\dagger(x;\boldsymbol{\theta}) M U(x;\boldsymbol{\theta}) |0\rangle^{\otimes n}, \\ \boldsymbol{\theta} \in [0,2\pi]^d \Big\}.$$

The circuit unitary is

$$\begin{aligned} U(x;\boldsymbol{\theta}) = E_n V(x;\boldsymbol{\theta}) &= E_n \bigotimes_{j=0}^{n-1} V_j(x;\boldsymbol{\theta}) \\ &= E_n \bigotimes_{j=0}^{n-1} \left[W_j^{(r)} R_{a_{j,r-1}}(x) W_j^{(r-1)} \cdots W_j^{(1)} R_{a_{j,0}}(x) W_j^{(0)} \right], \end{aligned} \tag{15}$$

where $a_{j,\ell} \in \{x, y, z\}$, $j = 0, \dots, n-1$, and $\ell = 0, \dots, r-1$ are arbitrary fixed encoding axes, with $R_{a_{j,\ell}}(x) = \exp[-ix\sigma_{a_{j,\ell}}/2]$. Each trainable layer is a general single-qubit rotation

$$W_j^{(s)} = R_z(\theta_{j,s,2})\, R_y(\theta_{j,s,1})\, R_z(\theta_{j,s,0}), \qquad s = 0, \dots, r,$$

leading to $d = 3n(r+1)$. The fixed entangling layer is

$$E_n = \begin{cases} I, & n = 1, \\ \mathrm{CNOT}_{n-1\to 0}\, \mathrm{CNOT}_{n-2\to n-1} \cdots \mathrm{CNOT}_{1\to 2}\, \mathrm{CNOT}_{0\to 1}, & n \ge 2, \end{cases}$$

where the rightmost gate acts first. The fixed measurement observable is

$$M = Z_{n-1} = I^{\otimes(n-1)} \otimes Z.$$

Lemma A.21 (CNOT-ring conjugation)**.** *Let*

$$E_n = \begin{cases} I, & n = 1, \\ \mathrm{CNOT}_{n-1\to 0}\, \mathrm{CNOT}_{n-2\to n-1} \cdots \mathrm{CNOT}_{1\to 2}\, \mathrm{CNOT}_{0\to 1}, & n \ge 2, \end{cases}$$

be a ring of CNOT gates, where the rightmost gate acts first. Then

$$E_n^\dagger Z_{n-1} E_n = \bigotimes_{j=0}^{n-1} Z_j, \qquad n \ge 1.$$

Proof. For $n = 1$, $E_1 = I$, and the claim follows immediately.

For $n \geq 2$, conjugation by a CNOT with control c and target t satisfies

$$\mathrm{CNOT}_{c\to t}^{\dagger}\, Z_c\, \mathrm{CNOT}_{c\to t} = Z_c, \qquad \mathrm{CNOT}_{c\to t}^{\dagger}\, Z_t\, \mathrm{CNOT}_{c\to t} = Z_c Z_t.$$

Propagating Z_{n-1} backwards through E_n, the first gate $\mathrm{CNOT}_{n-1\to 0}$ leaves Z_{n-1} unchanged. The subsequent gates successively introduce $Z_{n-2}, Z_{n-3}, \ldots, Z_0$. Hence

$$E_n^{\dagger} Z_{n-1} E_n = Z_0 Z_1 \cdots Z_{n-1} = \bigotimes_{j=0}^{n-1} Z_j.$$

□

Lemma A.22 (Factorization of the n-qubit model output)**.** *Every $f_{\boldsymbol{\theta}} \in \mathcal{F}_{\mathrm{1p-reup}}^{(n)}$ can be written as*

$$f_{\boldsymbol{\theta}}(x) = \prod_{j=0}^{n-1} g_j(x), \qquad g_j \in T_r^{(1)}.$$

Conversely, every product of n independently chosen functions $g_j \in T_r^{(1)}$ is realizable by the model.

Proof. Using Lemma A.21 to conjugate the CNOT-ring through the measurement,

$$\begin{aligned} f_{\boldsymbol{\theta}}(x) &= \langle 0|^{\otimes n}\, V^{\dagger} E_n^{\dagger} Z_{n-1} E_n V\, |0\rangle^{\otimes n} \\ &= \langle 0|^{\otimes n}\, V^{\dagger} \left(\bigotimes_{j=0}^{n-1} Z_j \right) V\, |0\rangle^{\otimes n}. \end{aligned} \tag{16}$$

By Equation (15), $V(x;\boldsymbol{\theta})$ is a tensor product, and so is $|0\rangle^{\otimes n}$. Consequently, the expectation value factorizes:

$$f_{\boldsymbol{\theta}}(x) = \prod_{j=0}^{n-1} \langle 0|\, V_j^{\dagger}(x;\boldsymbol{\theta}_j) Z V_j(x;\boldsymbol{\theta}_j)\, |0\rangle.$$

By Lemma A.20, each factor belongs to $T_r^{(1)}$.

Conversely, the parameter vectors of the n branches are independent. Thus, again by Lemma A.20, any collection $g_0, \ldots, g_{n-1} \in T_r^{(1)}$ can be realized simultaneously. □

Proposition A.23 (VC upper bound)**.** *For $n, r \geq 1$,*

$$\mathrm{VCdim}\Big(\mathcal{H}_{\mathrm{1p-reup}}^{(n)}\Big) \leq 2nr + 1.$$

Proof. By Lemma A.22, every model output is a product of n real trigonometric polynomials of degree at most r. Hence its Fourier degree is at most nr, and therefore

$$\mathcal{F}_{\mathrm{1p-reup}}^{(n)} \subseteq T_{nr}^{(1)}.$$

Consequently, the corresponding signed hypothesis class is a subclass of $\mathcal{H}_{nr}^{(1)}$. By Proposition A.14,

$$\mathrm{VCdim}\Big(\mathcal{H}_{\mathrm{1p-reup}}^{(n)}\Big) \leq \mathrm{VCdim}\Big(\mathcal{H}_{nr}^{(1)}\Big) = 2nr + 1.$$

□

Lemma A.24 (Factorized trigonometric sign interpolation)**.** *Let $n, r \geq 1$ and $m := 2nr + 1$. For any m distinct points $x_1, \ldots, x_m \in [0, 2\pi)$ and any labels $y_1, \ldots, y_m \in \{-1, +1\}$, there exist*

$$g_0, \ldots, g_{n-1} \in T_r^{(1)}$$

such that

$$\mathrm{sgn}\left(\prod_{j=0}^{n-1} g_j(x_i) \right) = y_i, \qquad i = 1, \ldots, m.$$

Proof. Relabel the points such that

$$0 \le x_1 < x_2 < \cdots < x_m < 2\pi$$

and set $y_{m+1} := y_1$.

Let

$$C := \{i \in \{1, \dots, m\} : y_i \neq y_{i+1}\}$$

denote the set of label changes around the circle. Since

$$\prod_{i=1}^{m} y_i y_{i+1} = 1,$$

the cardinality of C is even; write $|C| = 2s$. Since $m = 2nr + 1$ is odd,

$$2s \le m - 1 = 2nr,$$

and hence $s \le nr$.

For each $i \in C$, choose a point z_i strictly inside the circular interval from x_i to x_{i+1}, with x_{m+1} identified with $x_1 + 2\pi$. Define

$$q(x) = \prod_{i \in C} \sin\left(\frac{x - z_i}{2}\right),$$

with the empty product equal to 1.

The zeros z_i are simple, so the sign of q changes exactly in the intervals in which the prescribed labels change. Moreover, since $|C| = 2s$ is even, q is 2π-periodic. Thus there exists $\eta \in \{-1, +1\}$ such that

$$\operatorname{sgn}(\eta q(x_i)) = y_i$$

for all i.

Since $s \le nr$, choose integers $s_0, \dots, s_{n-1} \in \{0, \dots, r\}$ satisfying

$$\sum_{j=0}^{n-1} s_j = s.$$

Partition the $2s$ zeros into disjoint sets $C_0, \dots, C_{n-1}$ with $|C_j| = 2s_j$, and define

$$q_j(x) = \prod_{z \in C_j} \sin\left(\frac{x - z}{2}\right).$$

Each q_j is 2π-periodic and, by expanding the $2s_j$ half-angle sine factors in complex exponentials, has integer Fourier frequencies between $-s_j$ and s_j. Hence

$$q_j \in T_{s_j} \subseteq T_r.$$

Furthermore, $|q_j(x)| \le 1$, so $q_j \in T_r^{(1)}$.

Set

$$g_0 := \eta q_0, \qquad g_j := q_j, \quad j = 1, \dots, n-1.$$

Then $g_j \in T_r^{(1)}$ and

$$\prod_{j=0}^{n-1} g_j(x) = \eta q(x),$$

which gives the required labels. □

Proposition A.25 (VC lower bound)**.** *For $n, r \ge 1$,*

$$\operatorname{VCdim}\left(\mathcal{H}_{\mathrm{1p-reup}}^{(n)}\right) \ge 2nr + 1.$$

Proof. Choose any $m = 2nr + 1$ distinct points $x_1, \dots, x_m \in [0, 2\pi)$ and fix an arbitrary labeling $y_1, \dots, y_m \in \{-1, +1\}$.

By Lemma A.24, there exist $g_0, \dots, g_{n-1} \in T_r^{(1)}$ whose product realizes the prescribed signs. By the converse part of Lemma A.22, this product is realizable by the n-qubit re-uploading classifier. Hence every labeling of the m points is realizable, so they are shattered. Therefore

$$\mathrm{VCdim}\Big(\mathcal{H}^{(n)}_{\mathrm{1p-reup}}\Big) \geq 2nr + 1.$$

□

Theorem A.26 (Restatement of Theorem 3.2: VC dimension of the univariate n-qubit re-uploading classifier)**.** *Let $\mathcal{H}^{(n)}_{\mathrm{1p-reup}}$ be the hypothesis class of the univariate n-qubit re-uploading classifier with $n \geq 1$ qubits and $r \geq 1$ re-uploadings per qubit. Then*

$$\boxed{\mathrm{VCdim}\Big(\mathcal{H}^{(n)}_{\mathrm{1p-reup}}\Big) = 2nr + 1.}$$

Since the total number of trainable parameters is $d = 3n(r + 1)$, equivalently

$$\boxed{\mathrm{VCdim}\Big(\mathcal{H}^{(n)}_{\mathrm{1p-reup}}\Big) = \frac{2}{3}d - 2n + 1.}$$

For fixed n, the VC dimension therefore scales as

$$\mathrm{VCdim}\Big(\mathcal{H}^{(n)}_{\mathrm{1p-reup}}\Big) = \frac{2}{3}d + O(1).$$

Proof. Combine Propositions A.23 and A.25. Substituting $r = d/(3n) - 1$ into $2nr + 1$ gives

$$2nr + 1 = 2n\left(\frac{d}{3n} - 1\right) + 1 = \frac{2}{3}d - 2n + 1.$$

□

A.3 VC Dimension Bounds for the Multivariate Uniform Re-Uploading Classifier

In this section, we derive lower and upper bounds for the VC dimension of the uniform re-uploading classifier. The general re-uploading classifier is defined by

$$\mathcal{F}^{(p,n)}_{\mathrm{reup}} = \Big\{ f_{\boldsymbol{\theta}} : [0, 2\pi]^p \to [-1, 1] \;\Big|\; f_{\boldsymbol{\theta}}(\mathbf{x}) = \langle 0|^{\otimes n}\, U^\dagger(\mathbf{x}; \boldsymbol{\theta})\, M\, U(\mathbf{x}; \boldsymbol{\theta})\, |0\rangle^{\otimes n}, \\ \mathbf{x} \in [0, 2\pi]^p,\ \boldsymbol{\theta} \in [0, 2\pi]^d \Big\}.$$

$U(\mathbf{x}; \boldsymbol{\theta})$ is a parameterized quantum circuit consisting of data-encoding layers and trainable unitaries. The i-th input component x_i is encoded r_i times via single-qubit Pauli rotations $R_a(x_i)$ for $a \in \{x, y, z\}$.

For the uniform subclass $\mathcal{F}^{(p,n)}_{\mathrm{uni-reup}}$, all input features are re-uploaded the same number of times, and the trainable unitary belonging to each encoding has the same number of trainable parameters, i.e. $r_1 = r_2 = \cdots = r_p = r = \frac{d'}{cp} \in \mathbb{N}$ for some constant $c \in \mathbb{N}$ specifying the number of trainable parameters in each trainable unitary. The model may include d_{init} parameters before the first encoding block, which prepare a different initial state. The re-uploading parameter count is $d' = d - d_{\mathrm{init}}$.

The lower bound considers the subclass of the uniform re-uploading classifier that realizes the univariate n-qubit re-uploading classifier, and uses the latter's exact VC dimension as lower bound for the uniform re-uploading classifier. The upper bound extends the theorem of Chen et al. (2022) by including R_x in the encoding gate set.

Proposition A.27 (Lower bound for the uniform re-uploading classifier)**.** *Let $\mathcal{H}^{(p,n)}_{\mathrm{uni-reup}}$ be the hypothesis class of the uniform p-dimensional re-uploading classifier on $n \geq 1$ qubits, where each input feature is encoded r times. Then*

$$\mathrm{VCdim}\Big(\mathcal{H}^{(p,n)}_{\mathrm{uni-reup}}\Big) \geq 2pr + 1.$$

Proof. Consider the diagonal embedding

$$\Delta : [0, 2\pi] \to [0, 2\pi]^p, \qquad \Delta(t) = (t, \ldots, t).$$

On the diagonal, all input features satisfy $x_i = t$. Hence the pr encoding gates of the uniform model become Pauli rotations encoding the same scalar input t.

Consider the subclass in which all pr encoding gates act sequentially on a single qubit, while the remaining qubits and all remaining degrees of freedom are fixed. This realizes the $n = 1$ special case of the univariate n-qubit re-uploading classifier with pr re-uploadings.

By Theorem 3.2, this subclass has VC dimension

$$2pr + 1.$$

Therefore, if $t_1, \ldots, t_{2pr+1}$ are shattered by this univariate subclass, their diagonal embeddings

$$\Delta(t_1), \ldots, \Delta(t_{2pr+1})$$

are shattered by $\mathcal{H}^{(p,n)}_{\mathrm{uni-reup}}$. Hence

$$\mathrm{VCdim}\Big(\mathcal{H}^{(p,n)}_{\mathrm{uni-reup}}\Big) \geq 2pr + 1.$$

□

Proposition A.28 (VC dimension upper bound for general re-uploading classifier)**.** *Let $\mathcal{F}_{\mathrm{reup}}$ be the class of general re-uploading classifiers defined above and $\mathcal{H}_{\mathrm{reup}}$ being the corresponding signed-hypothesis class as defined in Equation* (1)*. Then*

$$\mathrm{VCdim}(\mathcal{H}_{\mathrm{reup}}) \leq \prod_{i=1}^{p}(2r_i + 1).$$

Proof. Our proof is a slightly more general version of (Chen et al., 2022), since we include R_x in the encoding gate set. Every Hermitian observable $M = \sum_\alpha m_\alpha P_\alpha$ can be decomposed into a sum of Pauli operators P_α and constant coefficients m_α. For any Pauli rotation $R_a(x) = e^{-ix\sigma_a/2}$, conjugation maps Pauli operators to linear combinations of Pauli operators with coefficients in $\{1, \cos x, \sin x\}$. For example,

$$R_z(x)^\dagger X R_z(x) = \cos(x)X - \sin(x)Y,$$
$$R_z(x)^\dagger Y R_z(x) = \sin(x)X + \cos(x)Y,$$

and

$$R_z(x)^\dagger Z R_z(x) = Z.$$

Analogous identities hold for rotations around the x- and y-axes. Trainable unitaries and entangling gates do not introduce any dependence on $\mathbf{x}$; they only change the Pauli coefficients.

Therefore, each time the input coordinate x_i is encoded, the trigonometric degree in x_i increases by at most one. Since x_i is encoded r_i times, the final expectation value belongs to the real vector space spanned by products of the form

$$\prod_{i=1}^{p}\psi_i(x_i), \qquad \psi_i \in \{1, \cos(x_i), \sin(x_i), \ldots, \cos(r_i x_i), \sin(r_i x_i)\}.$$

Thus every hypothesis in $\mathcal{H}_{\mathrm{reup}}$ is the sign of a function from a real vector space of dimension

$$\prod_{i=1}^{p}(2r_i + 1).$$

By Dudley's theorem for thresholded finite-dimensional real vector spaces (Anthony & Bartlett, 1999; Dudley, 1978),

$$\mathrm{VCdim}(\mathcal{H}_{\mathrm{reup}}) \leq \prod_{i=1}^{p}(2r_i + 1).$$

□

Theorem A.29 (Restatement of Theorem 3.3: VC bounds of uniform re-uploading model)**.** *Let $\mathcal{F}^{(p)}_{\mathrm{uni-reup}}$ be the function family of the uniform re-uploading model with input dimension p defined in Section 2.3.4 and $\mathcal{H}^{(p)}_{\mathrm{uni-reup}}$ be the corresponding hypothesis class, as defined in Equation* (1)*. Then*

$$\boxed{2pr+1 \le \mathrm{VCdim}(\mathcal{H}^{(p)}_{\mathrm{uni-reup}}) \le \prod_{1}^{p}(2r+1) = O\big((\frac{d}{p}+1)^p\big).}$$

Proof. The lower bound is given by Proposition A.27. The upper bound of the general re-uploading classifier is derived in Proposition A.28. Since $\mathcal{H}^{(p,n)}_{\mathrm{uni-reup}} \subset \mathcal{H}^{(p,n)}_{\mathrm{reup}}$ we have $\mathrm{VCdim}(\mathcal{H}^{(p,n)}_{\mathrm{uni-reup}}) \le \mathrm{VCdim}(\mathcal{H}^{(p,n)}_{\mathrm{reup}}) \le \prod_{i=1}^{p}(2r_i+1)$. The uniform re-uploading classifier uses $r_1 = r_2 = \cdots = r_p = r$, hence

$$\mathrm{VCdim}(\mathcal{H}^{(p,n)}_{\mathrm{uni-reup}}) \le \prod_{i=1}^{p}\big(2r_i+1\big) = \big(2r+1\big)^p = \big(\frac{2(d-d_{\mathrm{init}})}{cp}+1\big)^p = O\big((\frac{d}{p}+1)^p\big).$$

□

A.4 VC Dimension of Univariate n-Qubit Trainable-Encoding Classifier

This section estimates the VC dimension of the univariate n-qubit trainable-encoding classifier defined by

$$\mathcal{F}^{(n)}_{\mathrm{tr-enc}} = \Big\{ f_{\boldsymbol{\theta}} : [0,2\pi] \to [-1,1] \;\Big|\; f_{\boldsymbol{\theta}}(x) = \langle 0|^{\otimes n} U_n^\dagger(x;\boldsymbol{\theta}) Z_{n-1} U_n(x;\boldsymbol{\theta}) |0\rangle^{\otimes n}, \\ x \in [0,2\pi],\ \boldsymbol{\theta} \in [0,2\pi]^n \Big\},$$

where

$$U_n(x;\boldsymbol{\theta}) = E_n \left(\bigotimes_{j=0}^{n-1} R_{a_j}(\theta_j x) \right), \qquad a_j \in \{x, y\},$$

and

$$E_n = \begin{cases} I, & n = 1, \\ \mathrm{CNOT}_{n-1\to 0}\,\mathrm{CNOT}_{n-2\to n-1}\cdots\mathrm{CNOT}_{1\to 2}\,\mathrm{CNOT}_{0\to 1}, & n \ge 2, \end{cases}$$

We first derive an analytical expression for the model output in Lemma A.30. Then, Proposition A.31 establishes a linear lower bound. For the upper bound, Lemma A.32 bounds the number of sign patterns generated by one cosine factor, and Proposition A.33 extends this result to products of cosine factors. Combining these results yields the VC-dimension bounds in Theorem A.35.

Lemma A.30 (Analytical form of the trainable-encoding classifier)**.** *Consider the univariate trainable-encoding classifier defined above. Its output satisfies*

$$f_{\boldsymbol{\theta}}(x) = \prod_{j=0}^{n-1} \cos(\theta_j x).$$

Proof. By Lemma A.21, conjugating the measurement observable through the CNOT-ring yields

$$E_n^\dagger Z_{n-1} E_n = \bigotimes_{j=0}^{n-1} Z_j.$$

Let

$$R(x;\boldsymbol{\theta}) = \bigotimes_{j=0}^{n-1} R_{a_j}(\theta_j x),$$

so that $U_n(x;\boldsymbol{\theta}) = E_n R(x;\boldsymbol{\theta})$. Since both $R(x;\boldsymbol{\theta})$ and the initial state are tensor products, the expectation value factorizes:

$$f_{\boldsymbol{\theta}}(x) = \prod_{j=0}^{n-1} \langle 0| R_{a_j}(\theta_j x)^\dagger Z R_{a_j}(\theta_j x) |0\rangle .$$

For $a_j \in \{x, y\}$,

$$\langle 0| R_{a_j}(\phi)^\dagger Z R_{a_j}(\phi) |0\rangle = \cos\phi.$$

Hence,

$$f_{\boldsymbol{\theta}}(x) = \prod_{j=0}^{n-1} \cos(\theta_j x).$$

□

Proposition A.31 (Lower bound for the n-qubit trainable-encoding classifier)**.** *Let $\mathcal{H}^{(n)}_{\mathrm{tr-enc}}$ be the signed hypothesis class of the n-qubit trainable-encoding classifier. Then*

$$\mathrm{VCdim}(\mathcal{H}^{(n)}_{\mathrm{tr-enc}}) \geq n.$$

Proof. Set $m := n$ and choose

$$1 < x_1 < \cdots < x_m < 2.$$

We show that these m points are shattered.

Fix any labeling

$$\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_m) \in \{-1, +1\}^m$$

and define

$$J(\boldsymbol{\sigma}) = \{i \in \{1, \ldots, m-1\} : \sigma_i \neq \sigma_{i+1}\}.$$

For every $i \in J(\boldsymbol{\sigma})$, choose

$$\tau_i \in (x_i, x_{i+1}), \qquad \omega_i = \frac{\pi}{2\tau_i}.$$

Since all sample points lie in $(1, 2)$,

$$0 < \omega_i x_\ell < \pi,$$

and the first zero of $\cos(\omega_i x)$ occurs at $x = \tau_i$. Thus,

$$\mathrm{sgn}(\cos(\omega_i x_\ell)) = \begin{cases} +1, & x_\ell < \tau_i, \\ -1, & x_\ell > \tau_i. \end{cases}$$

Each such factor therefore creates exactly the required sign change between x_i and x_{i+1}.

If $\sigma_1 = -1$, choose additionally

$$\tau_\star \in (x_m/3, x_1), \qquad \omega_\star = \frac{\pi}{2\tau_\star}.$$

Then

$$\frac{\pi}{2} < \omega_\star x_\ell < \frac{3\pi}{2}$$

for all ℓ, so this factor is negative on every sample point and globally reverses the labeling.

At most

$$|J(\boldsymbol{\sigma})| + \mathbf{1}\{\sigma_1 = -1\} \leq (m-1) + 1 = m$$

nonconstant cosine factors are required. By Lemma A.30, the model output is

$$f_{\boldsymbol{\theta}}(x) = \prod_{j=0}^{n-1} \cos(\theta_j x),$$

and hence provides exactly $m = n$ independently trainable cosine factors. Assign the constructed frequencies to distinct parameters θ_j and set all unused parameters to zero, so that the corresponding factors equal 1.

Thus every labeling of the $m = n$ points can be realized, and therefore

$$\mathrm{VCdim}(\mathcal{H}^{(n)}_{\mathrm{tr-enc}}) \geq n.$$

□

Lemma A.32 (Sign patterns of a single cosine factor)**.** *Fix N sample points*

$$x_1, \dots, x_N \in [0, 2\pi].$$

Define the ternary sign function

$$\operatorname{sgn}_0(t) = \begin{cases} +1, & t > 0, \\ 0, & t = 0, \\ -1, & t < 0, \end{cases}$$

and, for $\theta \in [0, 2\pi]$, let

$$v_\theta = (\operatorname{sgn}_0(\cos(\theta x_1)), \dots, \operatorname{sgn}_0(\cos(\theta x_N))) \in \{-1, 0, +1\}^N.$$

Let

$$L_N := |\{v_\theta : \theta \in [0, 2\pi]\}|$$

denote the number of distinct ternary sign patterns generated by a single cosine factor on the N sample points. Then

$$L_N \le 26N + 1.$$

Proof. For $x_i > 0$, the zeros of $\theta \mapsto \cos(\theta x_i)$ are

$$\theta = \frac{\pi/2 + k\pi}{x_i}, \qquad k \in \mathbb{N}_0.$$

Such a zero lies in $[0, 2\pi]$ only if

$$k \le 2x_i - \frac{1}{2} \le 4\pi - \frac{1}{2} < 13.$$

Thus, each sample point contributes at most 13 breakpoints. If $x_i = 0$, there are no breakpoints.

Let B denote the number of distinct breakpoints generated by all N sample points. Then

$$B \le 13N.$$

The vector v_θ is constant on every connected component of $[0, 2\pi]$ after removing the breakpoints, giving at most $B + 1$ distinct vectors. At each breakpoint, at most one additional ternary vector is obtained. Therefore,

$$L_N \le (B + 1) + B \le 26N + 1.$$

□

Proposition A.33 (Growth-function bound for the trainable-encoding classifier)**.** *Let $n \ge 1$ and set*

$$m := n.$$

Then, for every $N \ge 1$, the growth function of the trainable-encoding hypothesis class satisfies

$$\Pi_{\mathcal{H}^{(n)}_{\mathrm{tr-enc}}}(N) \le \binom{26N + m}{m} \le \left(e\left(\frac{26N}{m} + 1\right)\right)^m.$$

Proof. Fix arbitrary sample points

$$x_1, \dots, x_N \in [0, 2\pi].$$

By Lemma A.30, every hypothesis in $\mathcal{H}^{(n)}_{\mathrm{tr-enc}}$ is of the form

$$x \mapsto \operatorname{sgn}\left(\prod_{j=0}^{m-1} \cos(\theta_j x)\right), \qquad m = n,$$

with

$$\theta_0, \dots, \theta_{m-1} \in [0, 2\pi].$$

For a single cosine factor, define its ternary sign vector on the sample by

$$v_\theta = (\operatorname{sgn}_0(\cos(\theta x_1)), \dots, \operatorname{sgn}_0(\cos(\theta x_N))) \in \{-1, 0, +1\}^N.$$

By Lemma A.32, the number L_N of distinct vectors v_θ obtained as $\theta \in [0, 2\pi]$ varies satisfies

$$L_N \leq 26N + 1.$$

For parameters $\boldsymbol{\theta} = (\theta_0, \ldots, \theta_{m-1})$, the ternary sign vector of

$$\prod_{j=0}^{m-1} \cos(\theta_j x)$$

on the sample is the componentwise product

$$v_{\theta_0} \odot \cdots \odot v_{\theta_{m-1}}.$$

Since componentwise multiplication is commutative, this product depends only on the multiset of the m individual vectors and not on their order. Since there are at most L_N possible vectors for each factor, the number of such multisets is

$$\binom{L_N + m - 1}{m}.$$

Consequently, the number of distinct ternary product vectors is at most

$$\binom{L_N + m - 1}{m} \leq \binom{26N + m}{m}.$$

The binary classifier uses the convention

$$\operatorname{sgn}(t) = \begin{cases} +1, & t \geq 0, \\ -1, & t < 0. \end{cases}$$

Hence its binary label vector is obtained from the corresponding ternary product vector by replacing every zero entry by $+1$. This mapping cannot increase the number of distinct label vectors. Therefore,

$$\Pi_{\mathcal{H}^{(n)}_{\mathrm{tr-enc}}}(N) \leq \binom{26N + m}{m}.$$

Finally, using

$$\binom{a}{b} \leq \left(\frac{ea}{b}\right)^b,$$

we obtain

$$\Pi_{\mathcal{H}^{(n)}_{\mathrm{tr-enc}}}(N) \leq \left(\frac{e(26N + m)}{m}\right)^m = \left(e\left(\frac{26N}{m} + 1\right)\right)^m.$$

□

Proposition A.34 (Upper bound for the trainable-encoding classifier)**.** *Let $\mathcal{H}^{(n)}_{\mathrm{tr-enc}}$ be the signed hypothesis class of the n-qubit trainable-encoding classifier. Then*

$$\mathrm{VCdim}(\mathcal{H}^{(n)}_{\mathrm{tr-enc}}) < 10n.$$

Proof. By Lemma A.30, the trainable-encoding classifier has n effective cosine factors. Suppose, for contradiction, that a set of $N \geq 10n$ points is shattered by $\mathcal{H}^{(n)}_{\mathrm{tr-enc}}$. Then

$$2^N \leq \Pi_{\mathcal{H}^{(n)}_{\mathrm{tr-enc}}}(N).$$

By Proposition A.33,

$$\Pi_{\mathcal{H}^{(n)}_{\mathrm{tr-enc}}}(N) \leq \left(e\left(\frac{26N}{n} + 1\right)\right)^n.$$

Hence,

$$2^N \leq \left(e\left(\frac{26N}{n} + 1\right)\right)^n.$$

Setting

$$\alpha := \frac{N}{n} \geq 10$$

and taking the n-th root yields

$$2^\alpha \leq e(26\alpha + 1).$$

Define the logarithmic difference of this expression as

$$g(\alpha) = \alpha \ln 2 - 1 - \ln(26\alpha + 1).$$

At $\alpha = 10$,

$$2^{10} = 1024 > 261e,$$

and therefore $g(10) > 0$. Moreover,

$$g'(\alpha) = \ln 2 - \frac{26}{26\alpha + 1} > 0 \qquad \text{for all } \alpha \geq 10.$$

Thus $g(\alpha) > 0$ for every $\alpha \geq 10$, or equivalently,

$$2^\alpha > e(26\alpha + 1),$$

contradicting the necessary shattering inequality.

Therefore, no set of $N \geq 10n$ points can be shattered. Hence,

$$\mathrm{VCdim}(\mathcal{H}^{(n)}_{\mathrm{tr-enc}}) < 10n.$$

□

Theorem A.35 (VC-dimension bounds for the n-qubit trainable-encoding classifier)**.** *Let $d = n \geq 1$, and let $\mathcal{H}^{(n)}_{\mathrm{tr-enc}}$ be the hypothesis class induced by the univariate n-qubit trainable-encoding classifier defined above. Then*

$$\boxed{d \leq \mathrm{VCdim}(\mathcal{H}^{(n)}_{\mathrm{tr-enc}}) < 10d.}$$

Consequently,

$$\boxed{\mathrm{VCdim}(\mathcal{H}^{(n)}_{\mathrm{tr-enc}}) = \Theta(d).}$$

Proof. The lower bound follows from Proposition A.31, and the upper bound from Proposition A.34. □

A.5 Shallow Fourier Network

The Fourier network defined as

$$f_{\boldsymbol{\theta}}(x) = \sum_{j=1}^{k} \left(a_j \cos(\omega_j x) + b_j \sin(\omega_j x)\right)$$

can be seen as a middle ground between PQCs and classical neural networks. We derive the VC dimension of shallow Fourier networks with trainable frequencies in A.39. The key idea is to combine a lower bound via distinct admissible frequencies (Lemma A.36) with an upper bound based on the Chebyshev structure of such models on short intervals (Proposition A.38).

Lemma A.36 (Lower bound for every positive frequency range)**.** *Let $k \geq 1$, $\Omega_{\max} > 0$, and let*

$$\mathcal{H}_{k,\Omega_{\max}} := \left\{ x \mapsto \mathrm{sgn}\left(\sum_{j=1}^{k} \big(a_j \cos(\omega_j x) + b_j \sin(\omega_j x)\big) \right) : a_j, b_j \in \mathbb{R}, \quad \omega_j \in [0, \Omega_{\max}] \right\}, \qquad x \in [0, 2\pi].$$

Then $\mathrm{VCdim}(\mathcal{H}_{k,\Omega_{\max}}) \geq 2k$.

Proof. Fix the admissible distinct positive frequencies $\omega_j = j\Omega_{\max}/(k+1)$ for $j = 1, \dots, k$. The $2k$ real functions $\cos(\omega_j x), \sin(\omega_j x)$, $j = 1, \dots, k$, are linearly independent on $[0, 2\pi]$. Indeed, a vanishing real linear combination would give a vanishing linear combination of the exponentials $e^{\pm i\omega_j x}$ with $2k$ distinct exponents. Differentiating at $x = 0$ through order $2k-1$ gives a Vandermonde system, whose coefficients must all vanish.

Linear independence implies that there are $2k$ points $x_1, \dots, x_{2k} \in [0, 2\pi]$ for which the evaluation matrix of these functions is invertible: otherwise all evaluation vectors would span a proper subspace of $\mathbb{R}^{2k}$, yielding a nonzero linear combination that vanishes throughout the interval. For any labels $y_i \in \{-1, +1\}$, choose the real coefficients a_j, b_j by solving the evaluation system so that $f(x_i) = y_i$ for every i. Thus all labelings of these $2k$ points occur, proving the claim. □

Lemma A.37 (Zero-multiplicity bound on VC dimension)**.** *Let $\mathcal{F}$ be a class of real-valued analytic functions on an interval $I \subseteq \mathbb{R}$, and let $\mathcal{H}$ be the corresponding binary hypothesis class*

$$\mathcal{H} \;=\; \Big\{ h_{\boldsymbol{\theta}} : I \to \{-1, +1\} \;\Big|\; h_{\boldsymbol{\theta}}(\mathbf{x}) = \operatorname{sgn}\big(f_{\boldsymbol{\theta}}(\mathbf{x})\big),\ \boldsymbol{\theta} \in \Theta \Big\}.$$

For all nonzero $f \in \mathcal{F}$, let $Z_{\mathrm{mult}}(f; I) \le N$ be the zero-count of f on interval I with multiplicity. Then

$$\mathrm{VCdim}(\mathcal{H}) \le N + 1.$$

Proof. Let $x_1 < \cdots < x_{N+2}$ be arbitrary distinct points in I. We show that no hypothesis $h_f \in \mathcal{H}$ can realize the alternating labeling

$$y_i = (-1)^{i+1}, \qquad i = 1, \dots, N+2,$$

which has $N+1$ sign changes along the ordered set. If $f \equiv 0$, then $h_f \equiv +1$, so h_f has no sign changes. Now assume $f \not\equiv 0$. Since f is analytic, its zeros are isolated, and each zero has a well-defined finite multiplicity.

Consider the sign sequence

$$h_f(x_1), \dots, h_f(x_{N+2}).$$

Every sign change between two consecutive sample points must be caused by a zero of f in the corresponding closed interval. If the zero lies strictly between two consecutive sample points, it contributes at least one to $Z_{\mathrm{mult}}(f; I)$. If the zero coincides with a sample point, then it can create at most two adjacent sign changes. This can only require two sign changes when the zero has even multiplicity, in which case its multiplicity is at least 2. Hence the total number of sign changes along any ordered sample is bounded by the number of zeros counted with multiplicity:

$$\#\{i : h_f(x_i) \neq h_f(x_{i+1})\} \le Z_{\mathrm{mult}}(f; I) \le N.$$

However, the alternating labeling on $N+2$ ordered points has $N+1$ sign changes. Therefore it cannot be realized by any $h_f \in \mathcal{H}$. Hence no set of $N+2$ points can be shattered, and so

$$\mathrm{VCdim}(\mathcal{H}) \le N + 1.$$

□

Proposition A.38 (VC dimension upper bound for trainable Fourier networks)**.** *For $\Omega_{\max} \ge 0$, let*

$$\mathcal{H}_{k,\Omega_{\max}} := \left\{ x \mapsto \operatorname{sgn}\left(\sum_{j=1}^{k} a_j \cos(\omega_j x) + b_j \sin(\omega_j x) \right) :\ a_j, b_j \in \mathbb{R},\ \omega_j \in [0, \Omega_{\max}] \right\}, \quad x \in [0, 2\pi].$$

Then

$$\mathrm{VCdim}(\mathcal{H}_{k,\Omega_{\max}}) \le 2k\big(\lfloor 4\Omega_{\max} \rfloor + 1\big) + 1.$$

In particular, if $d = 3k$, then

$$\mathrm{VCdim}(\mathcal{H}_{k,\Omega_{\max}}) = O(d(1 + \Omega_{\max})).$$

Proof. Fix any nonzero

$$f(x) = \sum_{j=1}^{k} \big(a_j \cos(\omega_j x) + b_j \sin(\omega_j x)\big), \qquad \omega_j \in [0, \Omega_{\max}].$$

If $\Omega_{\max} = 0$, all outputs are constant and the claimed bound follows directly. Hence assume $\Omega_{\max} > 0$. After merging repeated frequencies, f can be written in the form

$$f(x) = c + \sum_{k'=1}^{r} \big(A_{k'} \cos(\nu_{k'} x) + B_{k'} \sin(\nu_{k'} x)\big),$$

where

$$0 < \nu_1 < \cdots < \nu_r \le \Omega_{\max}, \qquad r \le k,$$

and $c \in \mathbb{R}$ is a possible constant term coming from zero frequency.

We first bound the number of zeros of f on $[0, 2\pi]$.

STEP 1: PARTITION INTO SHORT INTERVALS. Let

$$N := \lfloor 4\Omega_{\max} \rfloor + 1.$$

Partition $[0, 2\pi]$ into N consecutive closed intervals

$$I_\ell = [x_{\ell-1}, x_\ell], \qquad \ell = 1, \dots, N,$$

so that each interval has length at most

$$|I_\ell| \le \frac{2\pi}{N} < \frac{\pi}{2\Omega_{\max}}.$$

STEP 2: CHEBYSHEV STRUCTURE ON SHORT INTERVALS. Fix one interval $I = [u, v]$ and set $t = x - u \in [0, v - u]$. Then

$$f(u + t) = c + \sum_{k'=1}^{r} \Big(\widetilde{A}_{k'} \cos(\nu_{k'} t) + \widetilde{B}_{k'} \sin(\nu_{k'} t)\Big),$$

for suitable coefficients $\widetilde{A}_{k'}, \widetilde{B}_{k'} \in \mathbb{R}$. Rescaling to $s \in [0, \pi]$ via $t = \frac{v-u}{\pi} s$ yields

$$f(u + t) = c + \sum_{k'=1}^{r} \Big(\widehat{A}_{k'} \cos(\alpha_{k'} s) + \widehat{B}_{k'} \sin(\alpha_{k'} s)\Big),$$

where

$$\alpha_{k'} = \nu_{k'} \frac{v - u}{\pi} \le \Omega_{\max} \frac{v - u}{\pi} < \tfrac{1}{2}.$$

Let

$$V := \operatorname{span}\Big\{1, \cos(\alpha_1 s), \sin(\alpha_1 s), \dots, \cos(\alpha_r s), \sin(\alpha_r s)\Big\}.$$

By Theorem 8 of (Shisha, 1980), V is an extended Chebyshev system on $[0, \pi]$. Hence every nontrivial $g \in V$ has at most $\dim(V) - 1$ zeros, counted with multiplicity. Since $r \le k$, we have $\dim(V) \le 2k + 1$, and every nonzero Fourier network output satisfies

$$Z_{\text{mult}}(f; I) \le 2k.$$

Therefore, over all $N = \lfloor 4\Omega_{\max} \rfloor + 1$ intervals,

$$Z_{\text{mult}}(f; [0, 2\pi]) \le 2k \left(\lfloor 4\Omega_{\max} \rfloor + 1\right).$$

STEP 3: CONVERT THE ZERO BOUND INTO A VC DIMENSION BOUND. By Lemma A.37 we obtain

$$\mathrm{VCdim}(\mathcal{H}_{k,\Omega_{\max}}) \le 2k\big(\lfloor 4\Omega_{\max}\rfloor + 1\big) + 1.$$

Since $d = 3k$ and $\lfloor 4\Omega_{\max}\rfloor + 1 \le 4\Omega_{\max} + 1$, this implies uniformly for $\Omega_{\max} \ge 0$ that

$$\mathrm{VCdim}(\mathcal{H}_{k,\Omega_{\max}}) = O\big(d(1+\Omega_{\max})\big).$$

□

Theorem A.39 (VC dimension of univariate shallow Fourier network)**.** *Let*

$$\mathcal{H}_{\mathrm{fourier}} := \left\{ x \mapsto \mathrm{sgn}\left(\sum_{j=1}^{k} a_j \cos(\omega_j x) + b_j \sin(\omega_j x) \right) : \; a_j, b_j \in \mathbb{R},\; \omega_j \in [0, \Omega_{\max}] \right\}, \quad x \in [0, 2\pi],$$

and let $d = 3k$ be the number of trainable parameters. If $\Omega_{\max} > 0$, then

$$\boxed{\frac{2}{3}d \;\le\; \mathrm{VCdim}(\mathcal{H}_{\mathrm{fourier}}) \;\le\; 2k\big(\lfloor 4\Omega_{\max}\rfloor + 1\big) + 1.}$$

Consequently, for each fixed $\Omega_{\max} > 0$, $\mathrm{VCdim}(\mathcal{H}_{\mathrm{fourier}}) = \Theta(d)$. If $\Omega_{\max} = 0$, then $\mathrm{VCdim}(\mathcal{H}_{\mathrm{fourier}}) = 1$.

Proof. **Lower bound.** For $\Omega_{\max} > 0$, the lower bound follows from Lemma A.36, since $d = 3k$. For $\Omega_{\max} = 0$, the output is constant and both labels are attainable, so the VC dimension is 1. **Upper bound.** By Proposition A.38 we have $\mathrm{VCdim}(\mathcal{H}_{k,\Omega_{\max}}) = O\big(d(1+\Omega_{\max})\big)$ □

B VC DIMENSION OF CLASSICAL MODELS

This section gives the VC dimensions of the classical baselines listed in Table 1. The VC dimensions of these models were derived by other authors and are included here solely for the sake of completeness.

B.1 PERCEPTRON

A perceptron given by $f_{\boldsymbol{\theta}}(\tilde{\mathbf{x}}) = \boldsymbol{\theta}^T \tilde{\mathbf{x}}$ with $\tilde{\mathbf{x}} = (1, x_1, \ldots, x_p)^T \in [0,1]^{p+1}$, $\boldsymbol{\theta} \in \mathbb{R}^d$ and $d = p + 1$ has the VC dimension (Vapnik, 1999):

$$\boxed{\mathrm{VCdim}(\mathcal{H}_{\mathrm{perc}}) = p + 1 = d}$$

B.2 UNIVARIATE SHALLOW RELU NEURAL NETWORK

We consider a fully connected one-hidden-layer ReLU network with h hidden units and one-dimensional input.

Proposition B.1. *The VC dimension of the univariate shallow ReLU model*

$$\mathcal{F}_{\mathrm{ReLU}} := \left\{ f : \mathbb{R} \to \mathbb{R} \;\middle|\; f(x) = a_0 + \sum_{j=1}^{h} a_j\, \sigma(w_j x + b_j),\; a_0, a_j, w_j, b_j \in \mathbb{R} \right\},$$

with $\sigma(t) = \max\{0, t\}$ is given by

$$\boxed{\mathrm{VCdim}\big(\mathcal{H}_{\mathrm{ReLU}}\big) = h + 1 = \frac{d}{3} + \frac{2}{3}.}$$

Proof sketch. The full proof is given in Lemma 3.11(i) in (DeVore et al., 2021). For completeness, we briefly recall the key argument. Each function $f \in \mathcal{F}_{\mathrm{ReLU}}$ has the same approximation power as a continuous piecewise linear function with at most h knots.

Lower bound. By induction it can be shown that the function $S(t) := c + \sum_{j=1}^{h} a_j (t - \xi_j)_+$ can interpolate arbitrary labels on $h + 1$ ordered points, hence $\mathrm{VCdim} \ge h + 1$.

Upper bound. By induction it can be shown that there is no function in the model family that can realize the alternating labeling on a set of size $h + 2$. □

B.3 Feedforward Neural Network

An arbitrary feedforward NN with piecewise-linear activation function (e.g. ReLU) with L layers and d trainable parameters has the nearly-tight VC dimension (Bartlett et al., 2019):

$$\boxed{\mathrm{VCdim}(\mathcal{H}_{\mathrm{ffnn}}) = O(dL \log d)}$$

C Experimental Details

This section describes the datasets, models, training protocol, and evaluation procedure used for the experiments in Section 4. The source code and the exact JSON configuration files are provided in the supplementary materials.

C.1 Training and Evaluation Protocol

Sample sizes and trials. For every model–dataset combination we evaluate the training-set sizes $N \in \{5, 10, 20, 50, 100, 200, 500, 1000, 2000\}$. For each N we run 10 independent trials. In each trial we draw a balanced training subset of $N/2$ positive and $N/2$ negative samples (up to rounding for odd N) from the fixed training pool; the balanced test set of size 2000 is held fixed across all trials and sample sizes.

Input rescaling. Features are first min–max scaled to $[0, 1]$. For quantum models whose paper domain is $[0, 2\pi]$, the scaled features are then multiplied by 2π (rescaling `zero_to_two_pi`); for the classical baselines, whose paper domain is $[0, 1]^p$, no further rescaling is applied (rescaling `identity`). The rescaling is chosen per model family (`model_default`). Labels are mapped to $\{-1, +1\}$.

Optimization. All models are trained with the Adam optimizer and the hinge loss $\mathcal{L}(\boldsymbol{\theta}) = \mathbb{E}\big[\max(0,\, 1 - y\, f_{\boldsymbol{\theta}}(\mathbf{x}))\big]$, with learning rate 10^{-3}, no weight decay, and minibatch size 64 (or the full training set if it is smaller). Training runs for at most 2000 epochs and stops early when the training loss does not improve by at least 10^{-4} for 20 consecutive epochs. The model state with the lowest training loss is retained. We use a single training restart per trial. All experiments run on CPU (PyTorch-based batched state-vector simulator for the quantum models).

Parameter initialization. Trainable angles are initialized by independent zero-mean Gaussians with standard deviation `init_std`. All configs use a fixed `init_std` $= 0.05$, except the trainable-encoding classifier initializes its frequencies to $\theta_j = 1.0$.

Metrics. The reported quantity is the absolute generalization gap

$$\mathrm{gap}(f) = \big|\text{train accuracy}(f) - \text{test accuracy}(f)\big|.$$

Accuracies use the convention $\mathrm{sgn}(t) = +1$ for $t \geq 0$ and -1 otherwise. Shaded regions in all figures are 95% BCa bootstrap confidence intervals over the 10 trials (`scipy.stats.bootstrap`, 1000 resamples, fixed seed). For the margin analyses (Appendices D.4 and D.5), the per-sample margin is $y \cdot \langle Z \rangle$ for quantum models (expectation values already bounded in $[-1, 1]$) and $y \cdot \tanh(s/2)$ for classical models, where s is the real-valued model score; both are bounded to $[-1, 1]$.

VC bound overlays. The bound $B(N, \mathcal{H}, \delta)$ from Equation (3) is plotted with $\delta = 0.05$ using the VC dimension lower bounds of Table 1, listed for convenience in Table 2. No bound is shown for the feedforward NN, for which only the order $O(dL \log d)$ is known. The bounds are only plotted for $N > \mathrm{VCdim}$.

C.2 Datasets

We use eight datasets, listed in Table 3. Three are univariate ($p = 1$) synthetic tasks, three are multivariate synthetic and two are multivariate real-world datasets. Each dataset config fixes a seed and draws a deterministic pool of 3000 training and 2000 test samples; the multivariate datasets are

Table 2: VC dimension lower bounds used for the $B(N, \mathcal{H}, \delta)$ overlays in the figures.

Model	VCdim used
Univariate n-qubit re-uploading	$\frac{2}{3}d - 2n + 1$
Fourier network	$\frac{2}{3}d$
Perceptron	d
Uniform re-uploading (multivariate)	$\frac{2}{3}(d - d_{\text{init}}) + 1$
Bounded-reuse Pauli (proxy)	$\frac{2}{3}d$

Table 3: Datasets used in the experiments. Every dataset uses a balanced train/test pool of 3000/2000 samples, from which balanced subsets of size N are drawn. Datasets written in *italic* are synthetic.

Name	Type / notes	p	Train / Test
sine	$f = \sin(2\pi \cdot 2x)$; $x \in [0, 1]$	1	3000 / 2000
centered interval	$y = \text{sgn}(r^2 - (x - 0.5)^2)$, $r = 0.25$	1	
spline	piecewise-linear; 6 knots on $[0, 1]$, alternating labels	1	
two curves	two Fourier manifolds (Bowles et al., 2024)	$3 - 60$	
hidden manifold	Gaussian data on a 3-dim. manifold (Bowles et al., 2024)	$3 - 60$	
make_classification	`sklearn`[4]; 1 cluster/class, `class_sep`=1, `flip_y`=0.02	$3 - 60$	
census_income	UCI Census Income (Kohavi, 1996)	14	
gamma_telescope	UCI MAGIC Gamma Telescope (Bock, 2004)	10	

generated for $p \in \{3, 4, 5, 6, 10, 12, 15, 20, 30, 60\}$ where applicable. All datasets are balanced to 50% positive / 50% negative.

Univariate synthetic datasets. Samples $x \sim U[x_{\min}, x_{\max}]$ (with $[0, 1]$ here) and labels $y = \text{sgn}(g(x))$ for the target score function g. The spline uses knots $\{0, 0.2, 0.4, 0.6, 0.8, 1.0\}$ with values $\{-1, +1, -1, +1, -1, +1\}$, leading to decision boundaries $x \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$. The centered-interval task uses $g(x) = 0.25^2 - (x - 0.5)^2$, i.e. positive inside the interval $x \in [0.25, 0.75]$.

Two-curves and hidden-manifold. These two datasets are based on the dataset generators from the QML benchmarking suite (Bowles et al., 2024). Two-curves embeds samples on two offset Fourier manifolds: we use degree 5, offset 0.1, and additive Gaussian noise $\sigma = 0.01$. The hidden-manifold model ($m = 3$) maps Gaussian inputs of dimension m through a random weight matrix, a $\tanh$ nonlinearity, and a random readout, then median-splits the output into balanced classes. Both standardize features after generation.

Make-classification. This dataset is generated with `sklearn.datasets.make_classification` as a standard linearly-separable baseline with controlled difficulty. We use `n_informative` $= p$ (all features informative), `n_redundant` $=$ `n_repeated` $= 0$, one cluster per class, `class_sep` $= 1.0$, and label noise `flip_y` $= 0.02$. A stratified train/test split of 3000/2000 samples is drawn, then features are min–max scaled.

Census Income. The two official UCI Census Income files are read with the 14 standard columns. Numeric columns are median-imputed and standardized; categorical columns are most-frequent-imputed and ordinal-encoded. A balanced pool of 3000 train / 2000 test samples is drawn from the respective splits. All features are then min–max scaled to $[0, 1]$ (and to $[0, 2\pi]$ for quantum models).

MAGIC Gamma Telescope. The 10 numeric features and the class labels ($g \to 1,\ h \to -1$) are read from `magic04.data`. A disjoint balanced pool of 3000 train / 2000 test samples is drawn, then min–max scaled.

[4] `https://scikit-learn.org/stable/modules/generated/sklearn.datasets.make_classification.html`

C.3 Model Architectures and Hyperparameters

All models share the training protocol of Appendix C.1; per-model architectural choices are summarized below and in Table 4. The quantum models are simulated with a custom PyTorch-based batched state-vector simulator (`StateVecSimTorch`) running on CPU. The single-qubit trainable gate is the general rotation $\mathrm{Rot}(\phi, \theta, \omega) = R_z(\omega)\, R_y(\theta)\, R_z(\phi)$. Entangling layers use the CNOT-ring E_n from Equation (8) (rightmost gate applied first), and the measurement observable is $M = Z_{n-1}$ unless stated otherwise. Data is encoded with Pauli-Y rotations throughout.

Table 4: Model configurations used in the main experiments (Figures 1 and 2). "Main dataset" is the dataset on which the model appears in Figure 1a; the input dimension p and parameter count d are those of that configuration.

Model	Main dataset	p	d	n	Capacity sweep
Bounded-reuse Pauli	census_income	14	20	7	$d \in \{2, 4, 8, 12, 16, 20\}$; reuse $\in \{1, 2, 4, 8\}$
Uniform re-up. (multi.)	census_income	14	102	6	$r \in \{1, 2, 5, 10, 20, 50\}$
Trainable-encoding	spline	1	7	4	$n \in \{3, \ldots, 10\}$
Fourier network	spline	1	24	–	$k \in \{1, 2, 4, 8, 16, 32\}$
Shallow ReLU	spline	1	25	–	$h \in \{1, 2, 4, 8, 16, 32\}$
Perceptron	make_classification	15	16	–	none ($d = p + 1$)
Feedforward NN	census_income	14	227	–	matched to re-up. $r \in \{1, 2, 5, 10, 20, 50\}$

Bounded-reuse trainable Pauli classifier. A concrete subclass of $\mathcal{F}_{\mathrm{bound-pauli}}$ using single-qubit Pauli-word rotations $R_{a_j}(\theta_j)$. Each of the d parameters is placed on a qubit q_j (round-robin over the n qubits) with axis $a_j \in \{x, y, z\}$ (default y). The circuit is $W^{(0)}\,[R_a(x)\, W^{(\ell)}]^{\mathrm{reuse}}\, E_n$, i.e. $(\mathrm{reuse}+1)$ parameter layers interleaved with parameter-free $R_y(x)$ data encodings, so each parameter occurs $q_j = (\mathrm{reuse} + 1)$ times and $Q = (\mathrm{reuse} + 1)\, d$. The data encoding places one $R_{a_j}(x_{j \bmod p})$ per parameter slot on its qubit, so every feature is reachable regardless of n. The main configuration uses $n = 7$ qubits, $d = 20$, reuse $= 1$; the Q-probe sweep varies reuse $\in \{1, 2, 4, 8\}$ at fixed $d = 20$ to probe the $O(d \log d)$ upper bound (via Q), and the univariate capacity sweep varies $d \in \{2, 4, 8, 12, 16, 20\}$. Parameter initialization: `init_std` $= 0.05$.

Uniform re-uploading classifier (multivariate). This is the multivariate model of Section 2.3.4, whose architecture is shown in Figure 3. The circuit consists of an initial trainable Rot block and entangling layer on all n qubits, followed by r re-upload layers each consisting of a data-encoding $R_y(x)$ layer, a trainable Rot layer, and a CNOT entangling layer; a final entangling layer precedes measurement. Encoding occurrences are scheduled round-robin over the p features and the n qubits; unused slots in the final layer are fixed identities. The main configuration (Figures 1a and 2b) uses $n = 6$ qubits with a *ring* CNOT block, $d_{\mathrm{init}} = 3n = 18$ initial parameters, and 2 re-uploads, giving $d = 102$ at $p = 14$. The capacity sweep on census_income and gamma_telescope varies $r \in \{1, 2, 5, 10, 20, 50\}$ at fixed $n \in \{4, 6, 10\}$ and both *ring* and *full* CNOT blocks to disentangle qubit count and entangling topology from capacity. For census_income this leads to $d \in \{60, 102, 228, 438, 858, 2118\}$. The *full* block applies all $\binom{n}{2}$ CNOTs; the *ring* block applies nearest-neighbor CNOTs plus a wrap-around $(n{-}1 \to 0)$. Parameter initialization: `init_std` $=$ 0.05. The p-sweep uses $n = 6$ (full) and $n = 10$ (ring) with 60 encoding occurrences and cyclic parameter reuse to keep d constant across p.

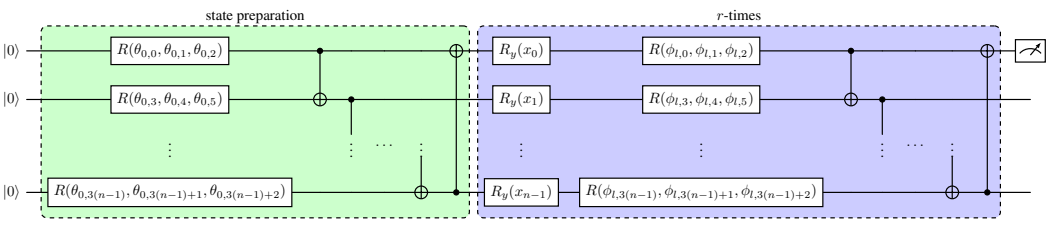


Figure 3: Exemplary architecture of the uniform re-uploading classifier with n-qubits and $l \in \{0, \ldots, r-1\}$ re-uploads.

Univariate n-qubit re-uploading classifier. The fixed architecture of Equation (7) with $d = 3n(r+1)$ and exact VC dimension $\frac{2}{3}d - 2n + 1$. The trainable block $W_j^{(s)} = R_z(\theta_{j,s,2})R_y(\theta_{j,s,1})R_z(\theta_{j,s,0})$ is applied per qubit; data is encoded as $R_y(x)$ on every qubit in lockstep. The `n_reuses` option repeats the r re-upload pairs (same parameters reused) to lengthen the circuit without changing d; all reported univariate re-uploading results use `n_reuses` $= 1$ unless noted. The capacity sweep varies $r \in \{1, 2, 5, 10, 20, 50\}$ at $n \in \{1, 2, 4\}$.

Trainable-encoding classifier. The univariate model of Equation (9) with output $f_{\boldsymbol{\theta}}(x) = \prod_{j=0}^{n-1} \cos(\theta_j x)$ and $d = n$. Parameters are constrained to $(0, 2\pi)$ via the sigmoid reparameterization $\theta_j = 2\pi\,\sigma(r_j)$ with unconstrained raw variables $r_j \in \mathbb{R}$, and initialized to $\theta_j = 1.0$. Encoding axes are y for all qubits. The capacity sweep varies $n \in \{3, 4, 5, 6, 7, 8, 9, 10\}$.

Univariate shallow Fourier network. The model of Equation (10) with $d = 3k$ and frequencies constrained to $[0, \Omega_{\max}]$ via the reparameterization $\omega_j = \Omega_{\max}\,\sigma(\omega_j^{\text{raw}})$ with unconstrained raw variables $\omega_j^{\text{raw}} \in \mathbb{R}$, initialized so the realized frequencies spread linearly across $[0, \Omega_{\max}]$. The main configuration uses $\Omega_{\max} = 10$; the capacity sweep varies $k \in \{1, 2, 4, 8, 16, 32\}$. Amplitudes a_j, b_j are unconstrained and initialize with `init_std` $= 0.05$.

Perceptron. The linear model $f(\mathbf{x}) = \boldsymbol{\theta}^\top \tilde{\mathbf{x}}$ with $\tilde{\mathbf{x}} = (1, x_1, \ldots, x_p)$ and $d = p + 1$. It has no internal capacity parameter; capacity changes only through p. Initialization: `init_std` $= 0.05$.

Univariate Shallow ReLU Network. We consider the family $\mathcal{F}_{\text{relu}}$ of fully connected one-hidden-layer ReLU networks with h hidden units and one-dimensional input, given by

$$f_{\boldsymbol{\theta}}(x) = a_0 + \sum_{j=1}^{h} a_j\,\sigma(w_j x + b_j), \qquad \sigma(t) = \max\{0, t\}, \qquad x \in [0, 1], \tag{17}$$

where $\boldsymbol{\theta} = (a_0, \{a_j, w_j, b_j\}_{j=1}^h) \in \mathbb{R}^d$ with $d = 3h + 1$. The univariate capacity sweep varies $h \in \{1, 2, 4, 8, 16, 32\}$. Initialization: `init_std` $= 0.05$.

Feedforward Neural Network with Piecewise Linear Activation Function. Finally, we have the fully connected feedforward neural network family $\mathcal{F}_{\text{ffnn}}$ with L layers, h_ℓ neurons in layer ℓ, and piecewise-linear activation function σ (e.g., ReLU). The model is given by

$$f_{\boldsymbol{\theta}}(\mathbf{x}) = W^{(L)}\sigma\big(W^{(L-1)}\sigma(\cdots\sigma(W^{(1)}\mathbf{x} + \mathbf{b}^{(1)})\cdots) + \mathbf{b}^{(L-1)}\big) + \mathbf{b}^{(L)}, \tag{18}$$

where $\boldsymbol{\theta} = \{W^{(\ell)}, \mathbf{b}^{(\ell)}\}_{\ell=1}^L \in \mathbb{R}^d$ denotes the trainable weights and biases and $\mathbf{x} \in \mathbb{R}^d$. For the capacity sweep we match the re-uploading classifier's parameter count: the target is $d = d_{\text{init}} + 3\,p\,r$ with $d_{\text{init}} = 3n = 18$ and two hidden layers (max width 512), and we search for the (h_1, h_2) closest to that target. The resulting architectures on census_income are $(h_1, h_2) \in \{(3, 3), (4, 7), (6, 17), (16, 11), (22, 22), (17, 98)\}$ for $r \in \{1, 2, 5, 10, 20, 50\}$, giving $d \in \{61, 103, 227, 439, 859, 2118\}$. The p-sweep configuration fixes a per-p architecture from a table so that $d \approx 198$ across all p. Initialization: `init_std` $= 0.05$.

C.4 Compute and Reproducibility

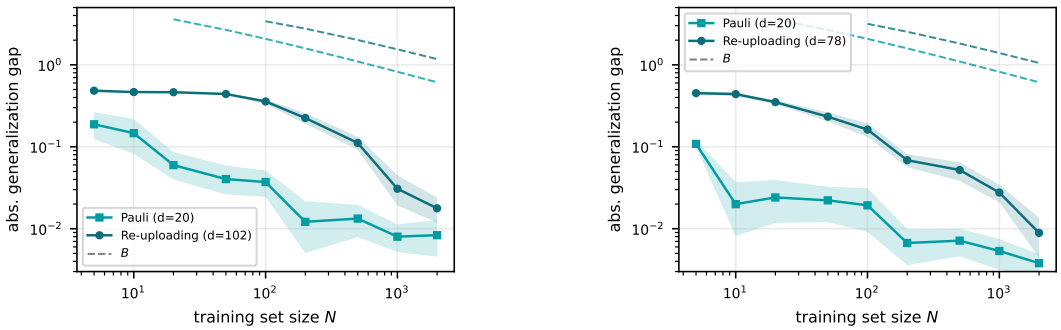
All experiments run on CPU; the quantum circuits are simulated with a batched PyTorch state-vector simulator. Each (N, trial) configuration is seeded deterministically from the base seed, the input dimension, N, and the trial index, so all results are reproducible from the provided configs. The largest configurations (high-capacity re-uploading classifiers on census_income at $r = 50$, $d \approx 2118$) took up to several days for 10 trials; smaller configurations completed in minutes to a few hours.

D Additional Results

Since over 4,000 individual configurations were evaluated, this section shows only a selection of the results. Results for all experimental runs can be found in the source code. This section contains selected further results for i) generalization gaps for different models and d for additional datasets, ii) train and test accuracies, and iii) margins.

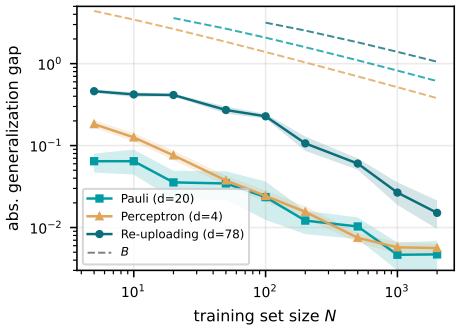
D.1 Empirical VC Dimension Bounds

This section shows that the generalization of quantum models is bounded by the generalization bounds predicted by the VC dimension Equations (2) and (3).

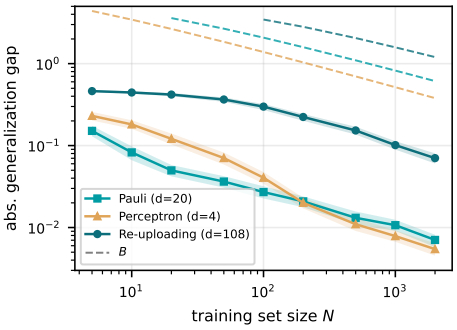


(a) Census Income

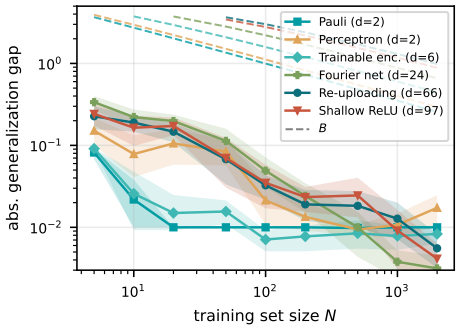
(b) Gamma Telescope



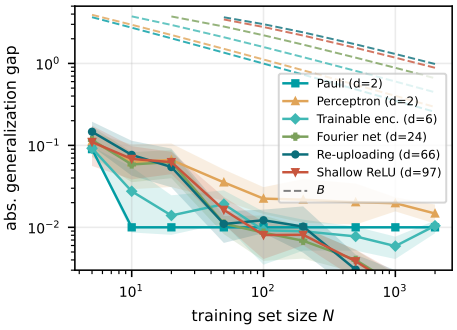
(c) Hidden Manifold



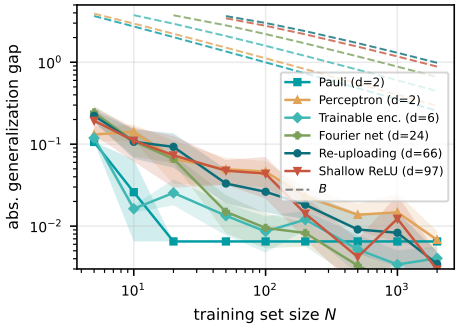
(d) Make Classification



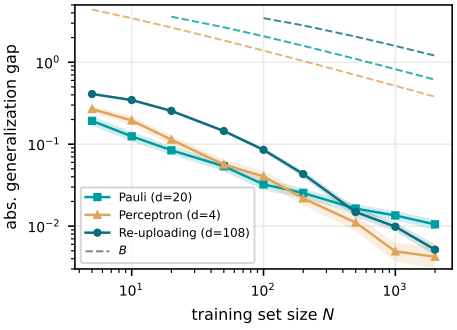
(e) Spline



(f) Centered Interval



(g) Sine

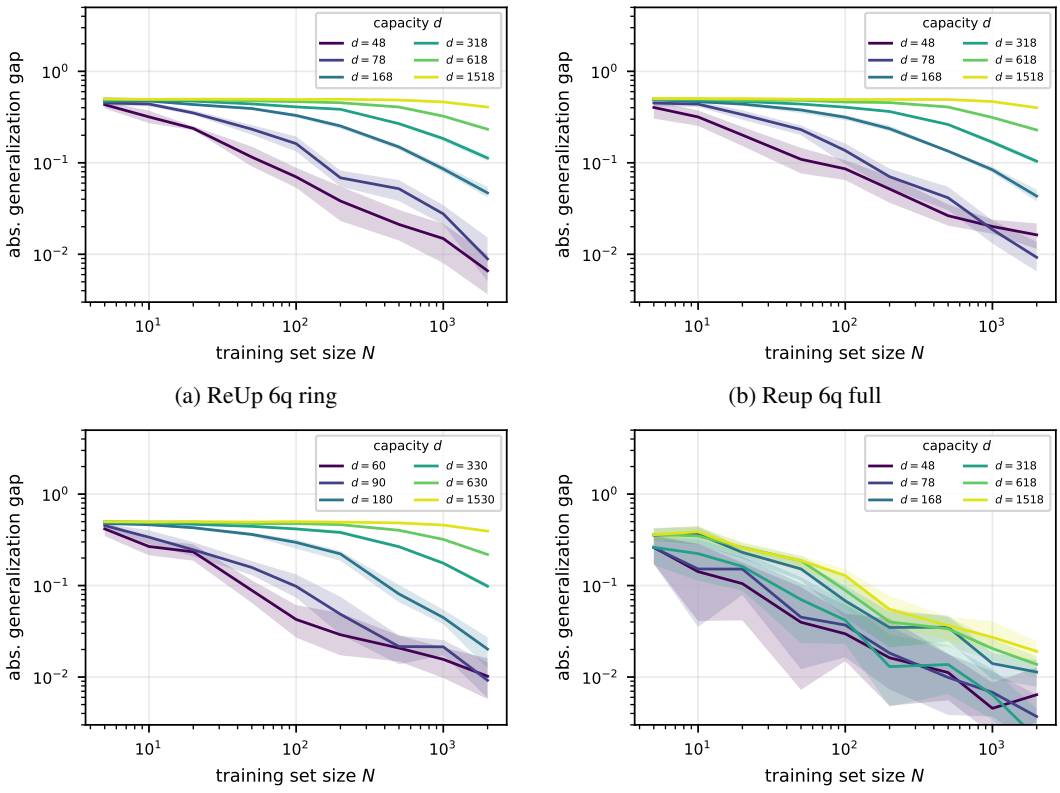


(h) Two Curves

Figure 4: Absolute generalization gaps and generalization bounds predicted by VC dimension for quantum and classical models on eight datasets. In all cases, the generalization gap is bounded by the VC-based reference curves. The VC dimensions are exact for univariate re-uploading, perceptron and shallow ReLU network. For the other models, the lower bounds are used as conservative proxies, since B increases with VC dimension for $N >> \mathrm{VCdim}$. The feedforward NN is omitted as only the order of its bound is known.

D.2 Dependence on Model Capacity

This section plots generalization gap for additional models on the gamma_telescope dataset.



(a) ReUp 6q ring

(b) Reup 6q full

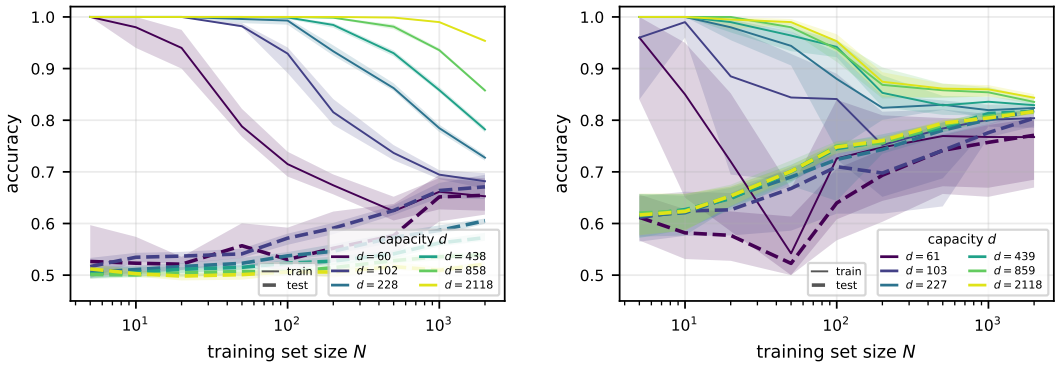
(c) Reup 10q ring

(d) Feedforward NN

Figure 5: Absolute generalization gaps for different reuploading architectures and qubit counts, as well as feedforward NN as baseline on the multivariate gamma telescope dataset. All three reuploading models exhibit near identical generalization gaps for high capacity models with $d \geq 168$ and some differences in the generalization gaps for smaller models. However, for these smaller models, the confidence intervals are large, indicating a high degree of randomness. Hence, we conclude that the entangling architecture and qubit count have no notable influence on the generalization gap. The generalization gap of the feedforward NN increases faster than for the reuploading classifiers for $d \geq 168$ and similar for smaller capacities.

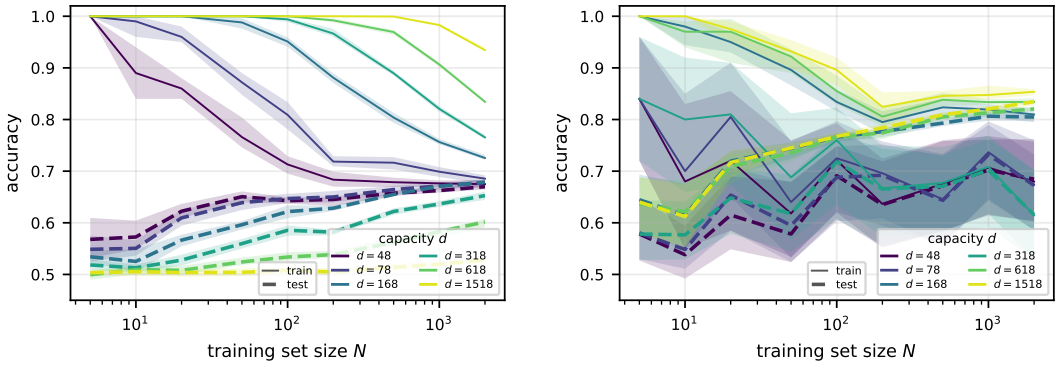
D.3 Train and Test Accuracies

This section plots the train and test accuracies for selected models on census_income, gamma_telescope, and hidden_manifold.



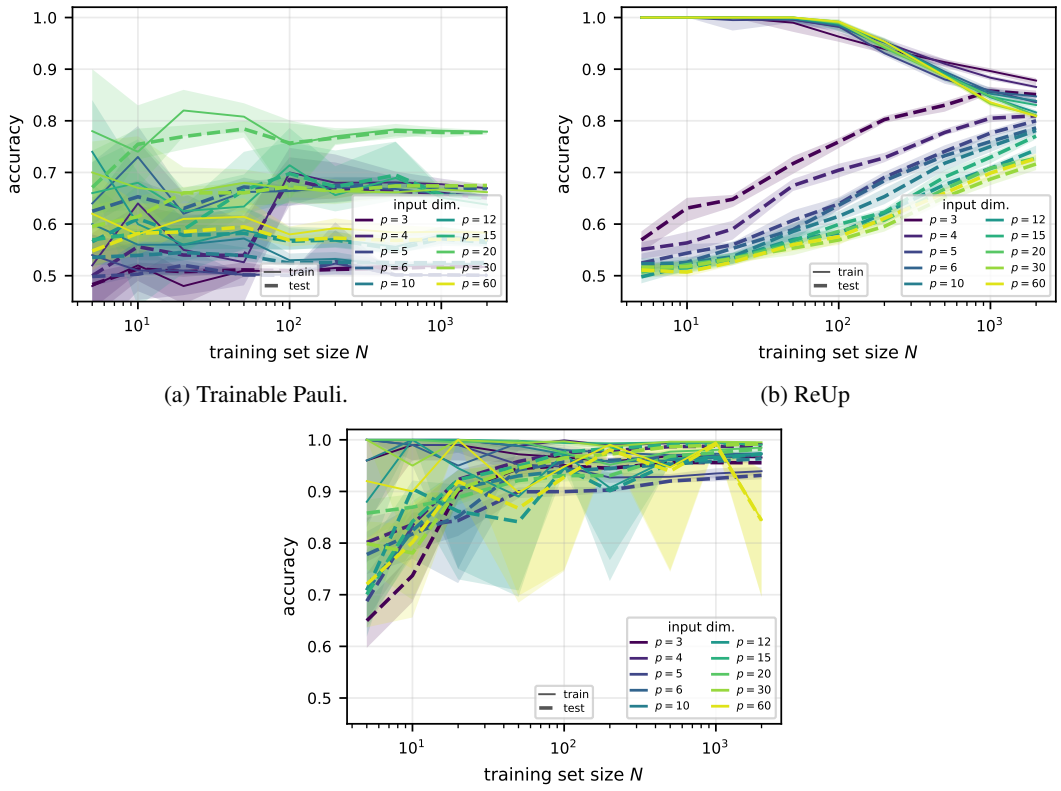
(a) ReUp 6q ring (b) Feedforward NN

Figure 6: Train and test accuracies for the reuploading classifier and the feedforward NN with different capacities on the census_income dataset. The reuploading classifier tends to overfit, as seen by the large gap between train and test accuracies. The gap is largest for small N and starts to converge with increasing N, however they only fully converge for the smallest model with $d = 60$ at $N \geq 1000$. The feedforward NN also overfits, but less extreme than the reuploading classifier. The feedforward NN's test accuracies are higher and the train accuracies decrease faster than that of the reuploading classifier.



(a) ReUp 6q ring (b) Feedforward NN

Figure 7: Train and test accuracies for selected models on the gamma_telescope dataset. The models behave similar as in Figure 6.



(a) Trainable Pauli.

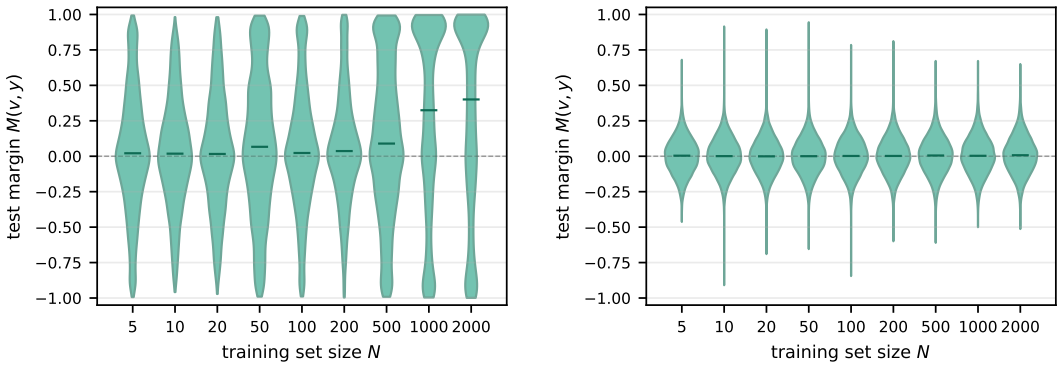
(b) ReUp

(c) Feedforward NN

Figure 8: Train and test accuracies for the models from Figure 2 on the hidden_manifold dataset. The re-uploading classifier shows a trend of increasing test accuracy with input dimension, whereas trainable Pauli classifier and feedforward NN exhibit no such trend. The reuploading classifier has small confidence intervals while the trainable Pauli classifier and feedforward NN exhibit large confidence intervals. The trainable Pauli classifier struggles to learn, achieving low train and test accuracies. The feedforward NN performs well with high train and test accuracies.

D.4 MARGINS

This section plots the margins for the Reuploading classifier on census_income.



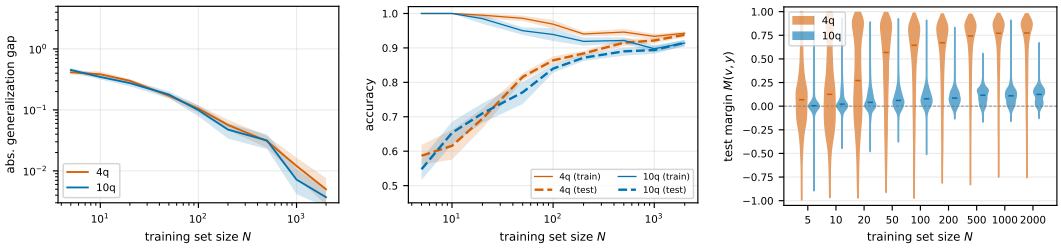
(a) d=60

(b) d=2118

Figure 9: Test margins for two different sized Reuploading classifiers on census_income. Both models are unconfident and behave randomly for $N \leq 500$, visible by their mean test margins close to zero and the approximately symmetric distributions around 0. For $N = 1000$ and $N = 2000$, the small Reuploading classifier becomes more confident, as the mean test margin becomes positive and the distribution less symmetric. The large classifier remains random and unconfident for all train set sizes. Together with the train and test accuracies in Figure 6a, the margins underscore the model's tendency to overfit and behave randomly on the test set.

D.5 Model Confidence

This section compares the generalization gap, accuracies and test margins of the Reuploading classifier with two different qubit counts on the two_curves dataset to show that good generalization alone is not a sufficient criterion for a desirable model.



(a) Absolute generalization gap. (b) Train and test accuracies. (c) Test margins.

Figure 10: Comparing the confidence of the Reuploading classifier for p=3 on two_curves. Both models attain similar generalization gaps and accuracies (a–b), yet their test-margin distributions diverge sharply (c): the 4q model becomes confidently correct (median margin 0.77 at N=2000), while the 10q model remains correct but barely confident (median 0.12). The generalization gap alone cannot distinguish these solutions, highlighting that VC-dimension bounds - which govern only the gap - characterize a necessary but not sufficient aspect of a good model.