# Spectral Bias in Variational Quantum Machine Learning

**Callum Duffy** [*1] **Marcin Jastrzebski** [*2] **Sarah Malik** [1]

## Abstract

In this work, we investigate the phenomenon of spectral bias in quantum machine learning (QML), in classical machine learning this manifests as models fitting low-frequency components of a target function before high-frequency ones. We perform the first comprehensive study characterising how this phenomena manifests in parameterised quantum circuits (PQCs) both theoretically and empirically. Leveraging the established formulation of PQCs as Fourier series, we prove that spectral bias in this setting can arise from the "redundancy" of the Fourier coefficients, which denotes the number of terms in the analytical form of the model contributing to the same frequency component. The choice of data encoding scheme dictates the degree of redundancy for a Fourier coefficient. We then further demonstrate this empirically with several different encoding schemes. We show this has direct consequences for quantum-assisted simulation of physical systems such as PDEs. Additionally, we demonstrate that PQCs with greater redundancy exhibit increased robustness to random perturbations in their parameters at the corresponding frequencies.

## 1. Introduction

Parameterised quantum circuits (PQCs) are a class of machine learning models often utilised in quantum machine learning (QML), one of the most promising applications of quantum computing Wiebe et al. (2015). These models resemble classical neural networks in that they consist of trainable quantum gates whose parameters are optimized via hybrid quantum-classical algorithms. Classical data can be encoded into the quantum circuit similarly, through parametrised gates. As with deep neural networks (DNNs), developing a theoretical understanding of PQCs is essential for advancing their practical capabilities and understanding their limitations. In classical settings, it is well known that DNNs exhibit spectral bias, a tendency to learn low frequency functions more readily than high frequency ones (Cao et al., 2020). This bias may elucidate why large neural networks have demonstrated low generalization error Cao et al. (2020); Xu (2018); Xu et al. (2024). However, in certain problems, models capable of capturing high-frequency components in the data have been found to be beneficial. Such instances can be found in domains of image recognition Fang et al. (2024); Lu et al. (2010) and solutions to PDEs Krishnapriyan et al. (2021); Wang et al. (2020). Strides have been taken to mitigate the effects of spectral bias in deep neural networks Ziqi Liu et al. (2020); Jagtap et al. (2020); Cai et al. (2019); Tancik et al. (2020); Fang & Xu (2024). Recent work has shown that PQC outputs are expressible as Fourier series, with the data encoding strategy governing the frequency spectrum accessible to the circuit Schuld et al. (2021). In this work, we leverage this framework to study the training dynamics of PQCs in learning target functions with specific frequency content.

## 2. Related Work

Since the relation between PQCs and Fourier series was established in (Schuld et al., 2021), it has been a powerful framework to understand their expressivity and properties. It has been demonstrated that PQCs following a reuploader scheme are universal function approximators (Schuld et al., 2021; Pérez-Salinas et al., 2020). Crucially, prior Fourier-based analyses characterise expressivity, investigating which functions a PQC can represent (Mhiri et al., 2025), whereas the present work characterises training dynamics; the order in which those representable components are actually learned. The distinction being a circuit may be sufficiently expressive to represent a target function yet systematically fail to learn it due to encoding-induced spectral bias, as we demonstrate concretely in Section 5.4. Certain studies have aimed to explore how the choice of data encoding affects the spectrum (Shin et al., 2023) and how the encoding subsequently impacts generalisation bounds (Caro et al., 2021). Further work has looked into the learning capabilities of PQCs from a Fourier perspective (Heimann et al., 2024).

[1]Department of Physics and Astronomy, University College London, London, United Kingdom [2]School of Mathematical and Physical Sciences, University of Sheffield, Sheffield, United Kingdom. Correspondence to: Callum Duffy <callum.duffy.22@ucl.ac.uk>.

*Proceedings of the AI4Physics Workshop at the $43^{rd}$ International Conference on Machine Learning (AI4Physics@ICML 2026)*, Seoul, South Korea. 2026. Copyright 2026 by the author(s).

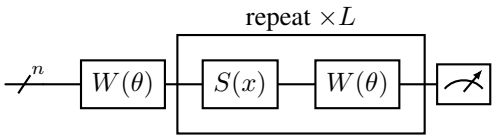



*Figure 1.* General reuploader circuit design with trainable gates $W(\theta)$ and data encoding gates $S(x)$.

This framework is widely used in the search for quantum advantage in PQCs, shedding light on the conditions under which PQCs can be dequantized (Sweke et al., 2025; Landman et al., 2022).

In parallel to this, the classical machine learning community has extensively studied spectral bias. This phenomenon has been demonstrated in fully connected networks (Xu, 2018; Cao et al., 2020), convolutional (Xu et al., 2024) and physics-informed neural networks (Krishnapriyan et al., 2021). Theoretical results have supported these findings, connecting spectral bias to neural tangent kernels, Fourier feature mappings and gradient flow (Geifman et al., 2022; Cao et al., 2020; Tancik et al., 2020; Bordelon et al., 2021; Basri et al., 2019). Motivated by the limitations this imposes on high-frequency generalisation, certain strategies have been proposed such as dynamically increasing network capacity (Fang & Xu, 2024), phase-shifted activation functions (Cai et al., 2020) and adaptive Fourier bases (Jagtap et al., 2020).

**Contributions** of this paper: Our work extends these studies by uniting the Fourier perspective of PQCs with spectral bias understood as a property of learning dynamics, explicitly relating loss gradients to the rate at which different frequency components of target functions are learned. We first establish a theoretical link between spectral bias and the redundancy structure of Fourier coefficients present in PQCs, showing that frequency components with high redundancy can exhibit larger gradients than those with lesser redundancy and do so on average when parameters are small. We then verify this numerically by comparing the learning dynamics across various encoding schemes and assess the robustness of each encoding strategy to perturbations of the trainable parameters. We then empirically demonstrate how this can affect using PQCs to learn partial differential equations (PDEs).

## 3. Background

We will consider circuits of the general form seen in Figure 1, known as reuploader circuits, and we denote the function this circuit outputs as $f$. Circuits are defined on $n$ qubits with data encoding unitaries $S(x)$, trainable unitaries $W(\theta)$ with trainable parameters $\theta \in \Theta$ and a Hermitian observable $O$. This framework considers classical data $\mathcal{X} = (\vec{x}^1, \ldots, \vec{x}^m)$, along with encoding unitaries which encode each element $x_k$ of $\vec{x}$ onto one qubit via a gate $G(x_k) = e^{-i\beta x_k H_k}$, where $\beta$ is some scale factor one may wish to apply. The full encoding unitary is given by $S(x) = \prod_{k=1}^{n} e^{-i\beta_k x_k H_k}$. combining all unitaries into a single unitary $U(x, \theta)$ of dimension $2^n$, the output of the circuit then takes the form

$$f_\theta = \langle 0|U^\dagger(x,\theta) O U(x,\theta)|0\rangle, \tag{1}$$

whereby $U(x, \theta)$ is defined as

$$U(x,\theta) = \prod_{l=1}^{L} \left(W^l(\theta_l) S^l(x)\right) W^0(\theta_0), \tag{2}$$

where $L$ is the total number of reuploading layers in the circuit. By considering the construction above we know the circuit can be functionally represented as a Fourier series (Schuld et al., 2021):

$$f(x,\theta) = \sum_{\vec{\omega}\in\Omega} c_{\vec{\omega}}(\theta) e^{i\vec{\omega}\cdot\vec{x}}. \tag{3}$$

For simplicity, we focus on a one-dimensional input case:

$$f(x) = \sum_{\omega\in\Omega} c_\omega e^{i\omega x}. \tag{4}$$

The set of frequencies $\Omega$ to which the model can have access is determined by the eigenvalues of the encoding unitaries which can be assumed to be diagonal $S^l(x) = diag(\lambda_1^l, ..., \lambda_d^l)$. More specifically:

$$\Omega = \{\omega = \Lambda_{\mathbf{k}} - \Lambda_{\mathbf{j}}, \mathbf{k}, \mathbf{j} \in [d]^L\}. \tag{5}$$

Here, the multi-index notation $\mathbf{j} = \{j_1, ..., j_L\} \in [d]^L$, has been introduced, where $[d]^L$ denotes the set of $L$ integers between $1, ..., d$ and $d = 2^n$, with $n$ denoting the number of qubits. The sum of eigenvalues coming from $L$ embedding gates can then be stated as $\Lambda_{\mathbf{j}} = \lambda_{j_1} + ... + \lambda_{j_L}$ for a given $\mathbf{j}$. The coefficients $c_{\vec{\omega}}$ are generally nontrivial and depend on the non-embedding layers of the circuit and the observable $O$, which can also be assumed diagonal (Casas & Cervera-Lierta, 2023):

$$c_\omega = \sum_{\substack{\mathbf{k},\mathbf{j}\in[d]^L \\ \Lambda_k - \Lambda_j = \omega}} a_{\mathbf{k},\mathbf{j}}, \tag{6}$$

with

$$a_{\mathbf{k},\mathbf{j}} = \sum_i \left(O_i W^{*(L)}_{k_L,i} W^{(L)}_{i,j_L}\right) W^{*(0)}_{1,k} W^{(0)}_{j_1,1} \prod_{p=2}^{L} W^{*(p-1)}_{k_{p-1},k_p} W^{(p-1)}_{j_p,j_{p-1}}. \tag{7}$$

It is important to note multiple combinations of $\mathbf{k}, \mathbf{j}$ lead to identical values of $\Lambda_{\mathbf{k}} - \Lambda_{\mathbf{j}}$, which we shall refer to as the redundancy $R(\omega)$ of a frequency $\omega$.

## 4. Spectral bias

In this section we present theoretical results about gradients of PQCs in regression tasks, based on their frequency spectra. Let us begin with stating the upper bounds of these gradients in two scenarios. First, Theorem 4.1 describes a simple upper bound for models whose spectra consist of integer-valued frequencies. This is true for the most basic and proliferated version of the reuploader model (Schuld et al., 2021). In this scenario, the terms in Equation 4 form an orthogonal basis for the space of square-integrable functions on $[0, 2\pi]$:

$$\int_{-\pi}^{\pi} e^{-ikx} e^{ijx} = 0 \Rightarrow j \neq k \quad j, k \in \mathbb{Z}. \tag{8}$$

**Theorem 4.1.** ***Upper bound on the gradient of the loss at frequency $\omega$ for integer-frequency models***

*Let $f(x, \theta)$ denote the output of a PQC with an integer-valued spectrum, trained to minimise the mean squared loss with respect to a target function $h(x)$, both of which can be expressed as Fourier series with the same, integer-valued, spectrum. Under gradient descent, frequencies $\omega$ with larger redundancies $R(\omega)$ can induce larger gradients in the loss. Specifically, for any parameter $\theta$, the magnitude of the gradient of the loss at frequency $\omega$ satisfies:*

$$|\partial_\theta L(\omega)| \leq 4R(\omega)||O||_{tr} \left|c_{D_\omega}\right|, \tag{9}$$

*where $||O||_{tr}$ is the trace norm of $O$, which, for Hermitian matrices, is the sum of absolute values of their eigenvalues. $c_{D_\omega}$ indicates the difference between the coefficients $c_\omega$ of the model and the target.*

Theorem 4.2 is an upper bound for a more general case, where the spectrum $\Omega$ contains arbitrary frequencies. This scenario is closer to what occurs, in models with trainable embeddings (Jaderberg et al., 2024). We note, however, that the theorem does not cover that case fully, as the target spectrum is assumed to match exactly that of the model. We consider the understanding of training dynamics of trainable-spectra models to be an interesting and important research direction.

**Theorem 4.2.** ***Upper bound on the gradient of the loss at frequency $\omega$ for arbitrary-frequency models***

*Let $f_{\not\perp}(x, \theta)$ denote the output of a PQC with a non-integer spectrum, trained to minimise the mean squared loss with respect to a target function $h(x)_{\not\perp}$, whose spectrum matches that of $f_{\not\perp}(x, \theta)$. Under gradient descent, frequencies $\omega$ induce gradients boosted by other frequencies $\omega'$, based on their proximity. Frequencies with larger redundancies $R(\omega)$ can contribute larger values to the gradients of the loss. Specifically, for any parameter $\theta$, the magnitude of the gradient of the loss at frequency $\omega$ satisfies:*

$$|\partial_\theta L(\omega)| \leq 2||O||_{tr} \sum_{\omega'} |sinc(\pi(\omega - \omega'))| \times \left( \left| c^*_{D_{\omega'}} \right| R(\omega) + |c_{D_\omega}| R(\omega') \right). \tag{10}$$

To analyse this expression, we begin by noting the behaviour of the sinc(x) function, which acts as a weighing term in the sum. Immediately, we see that the cross-term contributions will be suppressed for frequencies far from $\omega$. In the local neighbourhood of $\omega$, once again, frequencies with the largest redundancies can achieve largest gradients. For well-spread-out frequencies ($|\omega - \omega'| \gg 1$), the cross-terms disappear and we recover the result for orthogonal frequencies (Equation 9).

This shows that the gradient at a given frequency is allowed to be greater when $R(\omega)$ is greater. With the upper bounds for frequency-component gradients established, we now turn to finding their values in expectation.

The following theorems provide expected values of the gradients at each frequency, under the assumption of trainable single-qubit non-encoding unitaries with small parameters. Such an assumption holds, for example, for small-angle initialisation schemes (Wang et al., 2024; Zhang et al., 2025). For this approach we use the formalism of PQCs provided in (Wiedmann et al., 2024), whereby the circuits in consideration must consist of non-trainable Clifford gates and single-qubit Pauli rotations. The model can then be described as:

$$f(\boldsymbol{\theta}, x) = \sum_{\substack{s,c \in \mathbb{N}_0^d \\ s',c' \in \mathbb{N}_0^w}} k_{s,c,s',c'} \, 2^{-\sum_{j=1}^d (s_j + c_j)} (-i)^{\sum_{j=1}^d s_j} \prod_{k=1}^{w} \sin^{s'_k}(\theta_k) \cos^{c'_k}(\theta_k) \times \prod_{j=1}^{d} \sum_{a_j=0}^{s_j} \sum_{b_j=0}^{c_j} \binom{s_j}{a_j} \binom{c_j}{b_j} (-1)^{s_j - a_j} \, e^{i(2a_j + 2b_j - s_j - c_j)x_j}, \tag{11}$$

The sum over nodes $(s, c, s', c')$, each of which are non-negative integer vectors $s, c \in \mathbb{N}_0^d, s', c' \in \mathbb{N}_0^w$. The variables $s, c$ denote the number of sine and cosine terms for each input $x_j$ and $s', c'$ likewise but for each variational parameter $\theta_k$. The parameters $k_{s,c,s',c'}$ are either 0 or 1 in absolute value, originating from expectation values. The trainable variational parameters are given by $\theta_k$ and the data $x_j = \beta_j x$, where $\beta_j$ is some constant. This is a slight modification from the original formulation in (Wiedmann et al., 2024) which considered $d$-dimensional data, instead we consider one-dimensional data but multiplied by some

coefficient $\beta_j$. The frequencies of the Fourier decomposition are then given by

$$\omega(\mathbf{a},\mathbf{b}) = \sum_{j=1}^{d} m_j(a_j,b_j)\beta_j = \sum_{j=1}^{d}\omega_j. \tag{12}$$

$\mathbf{a} = (a_1,\ldots,a_d)$ and $\mathbf{b} = (b_1,\ldots,b_d)$ are multi-indices describing which element of the double sum at each index $j$ was used to create a given frequency. The total frequency $\omega$ associated with the specific choice $(\mathbf{a},\mathbf{b})$ is the sum of the local frequency contributions determined by $a_j$ and $b_j$ and $m_j(a_j,b_j) := 2a_j + 2b_j - s_j - c_j \in \mathbb{Z}$. The Fourier coefficients are then given by

$$c_\omega(\boldsymbol{\theta}) = \sum_{s,c,s',c'} k_{s,c,s',c'}\, 2^{-\sum_{j=1}^{d}(s_j+c_j)}(-i)^{\sum_{j=1}^{d} s_j} \times p(s,c,\omega)\prod_{k=1}^{w}\sin^{s'_k}(\theta_k)\cos^{c'_k}(\theta_k). \tag{13}$$

where

$$p(s,c,\omega) = \prod_{j=1}^{d}\sum_{a_j=0}^{s_j}\sum_{b_j=0}^{c_j}\binom{s_j}{a_j}\binom{c_j}{b_j}(-1)^{s_j-a_j}\delta^{\omega_j}_{m_j(a_j,b_j)\beta_j}. \tag{14}$$

allows us to group the contributions to a given frequency $\omega$. The overall model spectrum is the set of frequencies that survive possible cancellations across nodes (different nodes with the same variational polynomial can cancel for all $\theta$ only if their total prefactor vanishes). The final spectrum is then

$$\Omega = \left\{\omega \mid \exists\, s', c' \text{ such that } \sum_{s,c} k_{s,c,s',c'}\, p(s,c,\omega) \neq 0\right\}.$$

The redundancy $R(\omega)$ is defined as the number of distinct variational terms contributing to the frequency $\omega$. It corresponds to the number of non-vanishing summands in the expression for $c_\omega(\boldsymbol{\theta})$. Formally, it is the cardinality of the set of active multi-indices:

$$R(\omega) := \left|\left\{(s,c,s',c') \;\middle|\; k_{s,c,s',c'}\, p(s,c,\omega) \neq 0\right\}\right|.$$

This value $R(\omega)$ determines the number of components summing to form the coefficient.

**Theorem 4.3.** ***Expected upper bound on the gradient of the loss at frequency $\omega$ for integer frequencies.***

*Under the assumption models are initialized with small-angles from a Gaussian distribution $\mu = 0, \sigma \ll 1$. The magnitude of the gradient of the loss at frequency $\omega$ to first-order satisfies:*

$$\mathbb{E}\left[|\partial_{\theta_k} L(\omega)|\right] \lesssim 2|c_{\omega_h}| \sum_{r=1}^{R(\omega)} \frac{|p(\mathbf{s}^{(r)},\mathbf{c}^{(r)},\omega)|}{2^{\sum_{j=1}^{d}(s_j^{(r)}+c_j^{(r)})}} \left[\frac{s_k'^{(r)} 2^{\frac{s_k'^{(r)}-1}{2}}\Gamma(\frac{s_k'^{(r)}}{2})}{\sqrt{\pi}}\prod_{j\neq k}\frac{2^{\frac{s_j'^{(r)}}{2}}\Gamma(\frac{s_j'^{(r)}+1}{2})}{\sqrt{\pi}}\right]\sigma^{S^{(r)}-1}. \tag{15}$$

*$\Gamma$ functions represent the absolute moments of Gaussian random variables, $S(r)$ is the total sine-degree from summing over all paths. In the case of no weight-sharing thus $s_j'^{(r)}$=1 at most.*

$$\mathbb{E}\left[|\partial_{\theta_k} L(\omega)|\right] \lesssim 2|c_{\omega_h}| \sum_{r=1}^{R(\omega)} \frac{|p(\mathbf{s}^{(r)},\mathbf{c}^{(r)},\omega)|}{2^{\sum(s_j^{(r)}+c_j^{(r)})}} \left(\frac{2}{\pi}\right)^{\frac{w-1}{2}} \sigma^{S^{(r)}-1}. \tag{16}$$

*Demonstrating Fourier coefficients with high redundancy can potentially elicit stronger gradient signals in expectation. Refer to Appendix D.3 for the full proof.*

We can similarly, demonstrate this for the case of non-integer frequencies.

**Theorem 4.4.** ***Expected upper bound on the gradient of the loss at frequency $\omega$ for non-integer frequencies.***

$$\mathbb{E}[|\partial_{\theta_k} L(\omega)|] \lesssim \sum_{\omega'} |sinc(\pi(\omega-\omega'))| \left(\|h_{\omega'}| \sum_{r=1}^{R(\omega)} \mathcal{G}^{(r)}(\omega)\sigma^{S^{(r)}-1} + |h_\omega| \sum_{q=1}^{R(\omega')} \mathcal{G}^{(q)}(\omega')\sigma^{S^{(q)}-1}\right), \tag{17}$$

*where $\mathcal{G}^{(r)}(\omega)$ represents the explicit geometric and combinatorial prefactor for the $r$-th path of frequency $\omega$:*

$$\mathcal{G}^{(r)}(\omega) = \frac{|p(\mathbf{s}^{(r)},\mathbf{c}^{(r)},\omega)|}{2^{\sum(s_j^{(r)}+c_j^{(r)})}} \times \left[\frac{s_k'^{(r)} 2^{\frac{s_k'^{(r)}-1}{2}}\Gamma(\frac{s'^{(r)}k}{2})}{\sqrt{\pi}}\prod_{j\neq k}\frac{2^{\frac{s_j'^{(r)}}{2}}\Gamma(\frac{s_j'^{(r)}+1}{2})}{\sqrt{\pi}}\right]. \tag{18}$$

*Refer to Appendix D.4 for a full proof.*

While the preceding results establish upper bounds, the expected upper bounds derived here further emphasise the role of redundancy in spectral bias, providing sharper guarantees on average. Furthermore, redundancy analysis allows us to highlight that terms with high redundancy can be expected

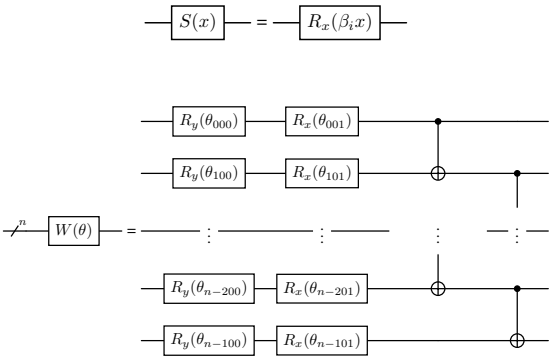



*Figure 2.* The specific components of the reuploader model used for the experiments, containing trainable parameters $\theta$ and data $x$. The choice of coefficients $\beta$ determines the nature of the encoding and $L$ is the number of circuit layers.

to exhibit higher gradients. Thus, one can use the redundancy distribution of an encoding scheme to boost gradients of frequencies of interest. On the other hand, the Heisenberg evolution picture used in the derivation shows a suppression of terms which pick up many trigonometric terms. Highest frequencies are always formed from such paths, indicating the inevitable tail-off of any encoding scheme.

## 5. Experimental results

Models used in this section follow the general structure of Fig. 2 with either 5 or 10 qubits, $L = 20$ and $O = Z_1$. The following results presented follow a similar analysis to that presented in the work (Cao et al., 2020). We perform a regression task with each of the models, introducing a target function of the form

$$h(x) = \sum_{\omega \in \Omega_h} A_\omega \sin(\omega x + \phi_\omega), \tag{19}$$

with frequencies $\omega$ forming the set $\Omega_h = \{5, 10, 15, 20, 25, 30, 35, 40, 45, 50\}$, all $A_\omega$ set to 1 before normalization and the phases $\phi_\omega$ drawn from the uniform distribution over $U(0, 2\pi)$ for each training instance. We sample 2048 equally spaced points in $x$ in the interval $x \in [0, 2\pi]$. We train the models using `PyTorch` (Ansel et al., 2024) and `Pennylane` (Bergholm et al., 2022) using the Adam optimizer (Kingma & Ba, 2017), evaluating every 5 epochs. The batch size is equal to the train size and the learning rate is fixed at 0.0005. All results reported are statistical means computed from ten separate runs. Compute resources can be found in Appendix A.

### 5.1. Spectral bias

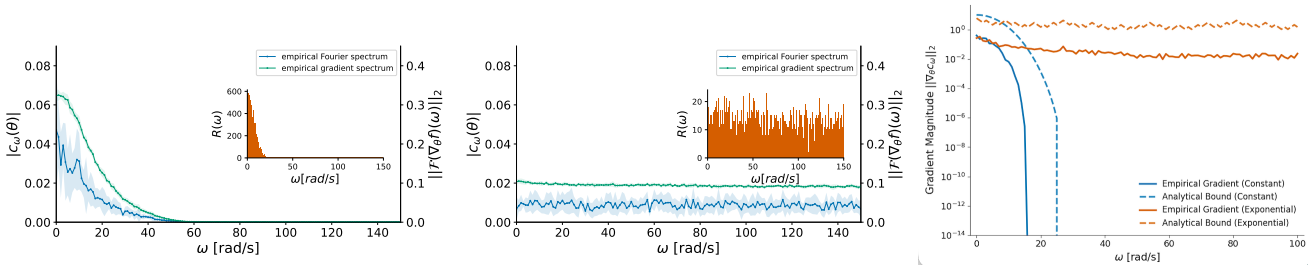
To demonstrate that PQC spectral bias is tied to frequency redundancy, we introduce two models (Figure 2) differing only by their embedding coefficients $\beta_i$. The "constant Pauli" embedding ($\beta_i = 1$) yields a concentrated spectrum, while the "ternary Pauli" embedding ($\beta_i = 3^i$) produces a much more uniform dense spectrum, the widest possible per reupload layer for single-qubit gates (Shin et al., 2023). In our $L$-layer setup, $\beta_i$ values repeat per layer to avoid overflow from large coefficients. Initial analysis in Figure 3 confirms that both model coefficients and gradient spectra follow the redundancy trends with Figures 3a and 3b showing constant and ternary encodings respectively. Furthermore, Figure 3c confirms that our analytical upper bounds closely track empirical gradients; here, we limit the analysis to $L = 5$ layers to maintain the computational tractability of the full Pauli propagation tree, though the redundancy-based logic holds for arbitrary depth. Target function fitting in Figure 4 clarifies the resulting trend: higher-redundancy frequencies are learned faster. Specifically, constant encoding exhibits a sharp redundancy drop-off and delayed high-frequency learning, whereas ternary encoding's near-constant redundancy ensures uniform learning rates. Appendices E and F extend these results to additional encodings, while Appendix G demonstrates that replacing Adam with vanilla gradient descent (GD) yields identical learning orders (Figure 10). This confirms that Theorems 4.1 and 4.3 characterize the loss landscape geometry rather than optimizer-specific properties, as Adam modulates step sizes without altering the underlying gradient magnitudes governed by our bounds.

We also study how trainable encodings alter spectral bias, specifically regarding how Theorems 4.2 and 4.4 manifest in practice. Maintaining the previous setup but with trainable encoding coefficients, we explore constant, ternary, and uniform initialisations from $\mathcal{U}(0, 0.1)$. Figure 5 demonstrates that while trainable frequencies dynamically shift the accessible spectrum, they partially mitigate rather than eliminate spectral bias, as the instantaneous redundancy distribution continues to govern relative gradient magnitudes. The sinc cross-term allows neighbouring frequencies to boost mutual gradients, yet this cannot override the ordering induced by the initial redundancy structure; consequently, constant initialisation retains a persistent low-frequency bias, whereas ternary initialisation's near-uniform redundancy remains robust to frequency perturbations. Furthermore, the fact that random small-angle initialisation severely hampers training dynamics proves that principled initialisation remains critical even when embeddings are fully trainable.

### 5.2. Shot Noise

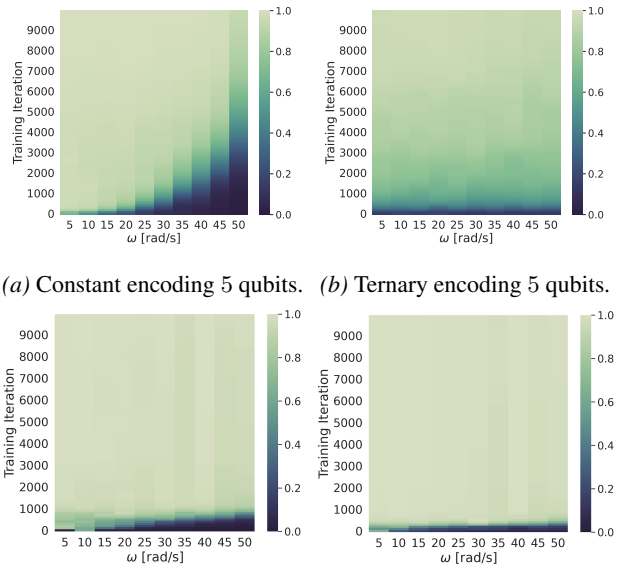
When deploying on hardware, one must rely on finite-shot estimates of the gradient. In this practical regime, our redundancy-based framework makes concrete predictions regarding the resolvability of frequencies.

**Proposition 5.1** (Finite-Shot Estimation Exacerbates Spectral Bias)**.** *Under any unbiased finite-shot gradient esti-*




*(a)* Constant Pauli encoding *(b)* Ternary Pauli encoding *(c)* Theoretical Bounds

*Figure 3.* The Fourier spectra of the (a) constant and (b) ternary encoding schemes with empirical data taken as the sampled mean over ten models. Depicted are the sampled theoretically accessible frequencies (light red), the mean Fourier coefficient (blue), and the total gradient of trainable parameters at each Fourier coefficient (green). The empirical gradient spectrum (green) is computed by projecting the full parameter gradient $\nabla_\theta f$ onto the frequency domain: for each frequency $\omega$, we compute the Fourier coefficient of $f$ with respect to $x$ at $\omega$, then take the $\ell_2$ norm across all parameters $\theta$, yielding the total gradient energy attributable to each frequency component. Plot (c) depicts the empirical gradient spectrum against and theoretical bounds imposed in Theorem 4.3, for 5 qubit circuits with 5 layers.




*(a)* Constant encoding 5 qubits. *(b)* Ternary encoding 5 qubits.

*(c)* Constant encoding 10 qubits. *(d)* Ternary encoding 10 qubits.

*Figure 4.* The rate at which frequencies (x-axis) are learnt during the course of training (y-axis), the colorbar measures the PQC spectrum normalised by the target amplitude at a given frequency ($|\tilde{f}_\omega|/A_i$). Each subplot depicts the training dynamics for a different encoding schemes and qubit number.

*mator with $M$ shots, the signal-to-noise ratio (SNR) for frequency $\omega$ is proportional to $R(\omega)$. Consequently, finite-shot estimation acts as a statistical filter that enforces the encoding-induced spectral bias: frequencies with lower redundancy fall below the noise floor first.*

*Proof.* The parameter-shift rule yields the exact analytical gradient in expectation, so Theorems 4.1– 4.4 apply natively to hardware-estimated gradients. Let the variance of a single-shot measurement be bounded by a constant $V$, and let $\hat{g}_\omega^{(M)}$ denote an unbiased $M$-shot estimator of the gradient at frequency $\omega$. Then $\mathrm{Var}(\hat{g}_\omega^{(M)}) \leq V/M$, giving noise floor $\sigma_{\text{noise}} \propto M^{-1/2}$. The SNR is:

$$\mathrm{SNR}(\omega) = \frac{|\partial_\theta \mathcal{L}(\omega)|}{\sigma_{\text{noise}}} \leq \frac{4R(\omega)\|O\|_{\text{tr}}\,|c_{D_\omega}|}{\sigma_{\text{noise}}} \propto R(\omega), \tag{20}$$

where the upper bound follows from Theorem 4.1. For any fixed shot budget $M$, frequencies for which $4R(\omega)\|O\|_{\text{tr}}|c_{D_\omega}| < \sigma_{\text{noise}}$ are physically erased from the learning process. Encodings with concentrated redundancy (e.g. constant Pauli) achieve high SNR at low frequencies but are noise-limited at high frequencies; encodings with uniform redundancy (e.g. ternary Pauli) distribute robustness evenly across the spectrum. □

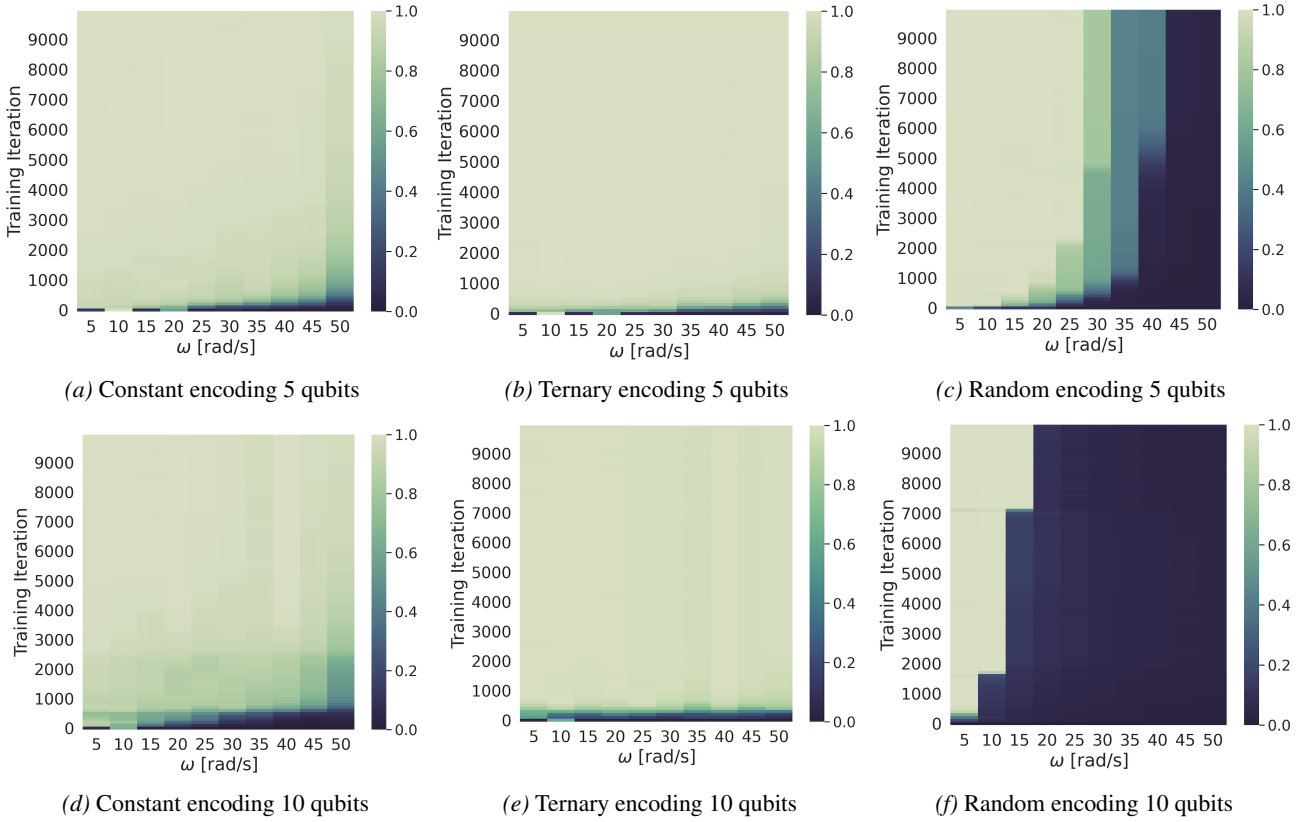
### 5.3. Robustness

We continue our exploration of the properties of PQCs by evaluating the robustness of the models' parameters to perturbations. Let $c_\omega(\theta^*)$ be the learned Fourier coefficient at frequency $\omega$, and suppose it decomposes as a sum of $R(\omega)$ contributing terms $c_\omega(\theta^*) = \sum_{i=1}^{R(\omega)} a_i(\theta^*)$. We then can study the robustness of the models under small isotropic parameter perturbations.

**Theorem 5.2** (Upper bound on the robustness of PQCs to isotropic parameter perturbations)**.** *The normalized root mean square (RMS) deviation is upper-bounded by*

$$\mathcal{R} = \frac{\sqrt{\mathbb{E}[|\Delta C|^2]}}{|c(\theta^*)|} \leq \frac{\sigma_a}{\kappa\,\bar{a}}\sqrt{\frac{1+(R(\omega)-1)\rho}{R(\omega)}}. \tag{21}$$

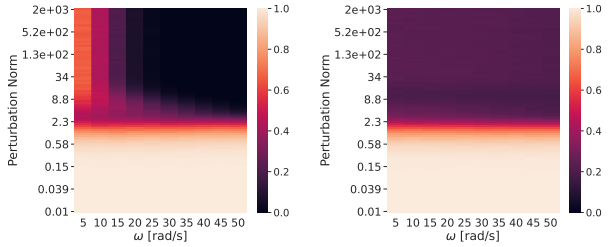
*where $\sigma_a$ is the typical perturbation scale of each summand and $\bar{a}$ is a typical per-summand magnitude, $\kappa$ the coherence between terms $a_i$ and $\rho$ the correlation between terms $a_i$. The full proof can be found in Appendix H*




*(a)* Constant encoding 5 qubits *(b)* Ternary encoding 5 qubits *(c)* Random encoding 5 qubits

*(d)* Constant encoding 10 qubits *(e)* Ternary encoding 10 qubits *(f)* Random encoding 10 qubits

*Figure 5.* Spectral training dynamics for trainable encoding schemes. Each heatmap shows the rate at which frequencies (x-axis) are learnt during training (y-axis), measured as $|\tilde{f}_\omega|/A_i$. Rows correspond to (top) 5 qubit simulations and (bottom) 10. Columns correspond to (a,d) fixed constant initialisation, (b,e) fixed ternary initialisation, and (c,f) random small-angle initialisation $\beta_i \sim \mathcal{U}(0, 0.1)$. Results are sampled means over ten models.



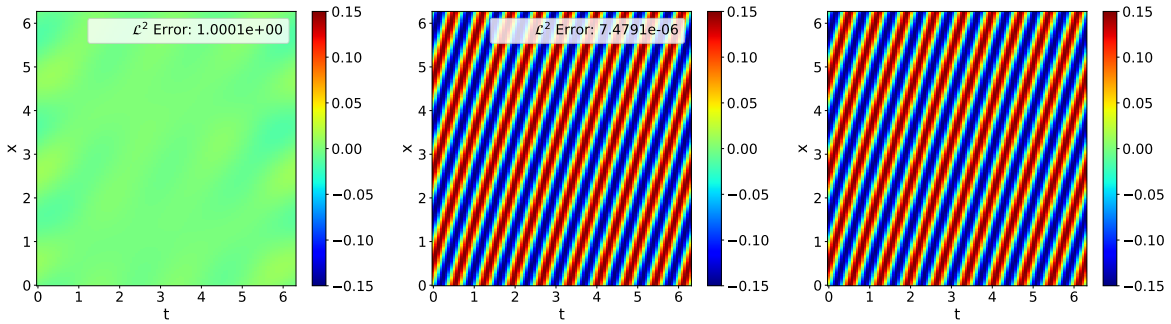

*(a)* Constant Pauli encoding *(b)* Ternary Pauli encoding

*Figure 6.* Normalised Fourier spectrum of the model output (x-axis: frequency, colourbar: magnitude) as a function of parameter perturbation (y-axis). Each subplot depicts the effects of parameter perturbations for a different encoding scheme.

In particular, if $\rho \approx 0$ (perturbation effects approximately uncorrelated across $i$)) and $\kappa$ is bounded away from zero, then

$$\mathcal{R} \lesssim \frac{1}{\kappa} \frac{\sigma_a}{\bar{a}} \frac{1}{\sqrt{R(\omega)}}.$$

Thus, in the case where per-term perturbations are approximately uncorrelated and the summands add with non-negligible coherence, redundancy reduces relative sensitivity roughly as $R(\omega)^{-1/2}$. If instead the perturbation responses are strongly correlated or the summands cancel in phase, the redundancy benefit vanishes. Continuing with the setup from section 5.1, we evaluate each model at the end of training and examine the effect of random isotropic perturbations $\vec{\theta} = \vec{\theta}^* + \delta\hat{\theta}$ to the learnt function. Here, $\delta$ is a chosen magnitude and $\hat{\theta}$ is a random unit vector in parameter space. Once $\vec{\theta}^*$ is perturbed we compute $f$ at the new parameters $\vec{\theta}$ and find the coefficients of the model at the frequencies of interest from the target function. Averages are taken over 100 samples of $\hat{\theta}$. A final average is taken over the phases $\phi$ of the target function. Figure 6 depicts the results for two encoding schemes with their trainable final parameters perturbed. It is evident lower frequencies are more robust to perturbations for the constant Pauli model whereas the exponential Pauli model shows a uniformity to the robustness at each frequency. We can see here the circuits operate in the regime where redundancy plays a key




*(a)* Constant encoding *(b)* Ternary encoding *(c)* Exact solution

*Figure 7.* Model predictions for the $1D$ convection equation across different encoding strategies. Subfigures (a) and (b) show the predicted solutions for constant and ternary encoding, respectively, compared against (c) the ground truth exact solution. The relative $\mathcal{L}^2$ error for each model is displayed in the upper right corner.

role in the robustness of the model at each frequency.

### 5.4. Convection Equation

The convection equation is a prototypical model for transport-dominated phenomena and provides a stringent test for learning oscillatory dynamics, as it preserves frequency content without dissipation. Here we consider the one-dimensional hyperbolic partial differential equation

$$\frac{\partial u}{\partial t} + \beta \frac{\partial u}{\partial x} = 0, \tag{22}$$

with domain $x \in [0, 2\pi]$, $t \in [0, 2\pi]$, initial condition $u(x, 0) = 0.15 \sin(3x)$, and convection coefficient $\beta = 3$. We dynamically sample $500$ collocation points per epoch, while boundary and initial condition losses are evaluated on fixed grids of $256$ and $100$ points in space and time, respectively. This equation has previously been used to expose failure modes in PINN models, where oscillatory solutions are difficult to learn despite sufficient model expressivity (Krishnapriyan et al., 2021). This problem therefore serves as a concrete downstream test of the spectral bias mechanisms identified in the preceding sections. Our models consist of six qubits with three reuploads of both $x$ and $t$, and are trained for $30{,}000$ epochs using the Adam optimizer with a learning rate of $10^{-4}$. Figure 7 shows the final learned solutions and corresponding $\mathcal{L}_2$ relative errors. When using constant encoding (Figure 7a), the model fails to recover the solution, yielding an error of $\mathcal{L}_2 = 0.936$. In contrast, ternary encoding (Figure 7b) produces an accurate solution with $\mathcal{L}_2 = 3.53 \times 10^{-4}$. Importantly, this discrepancy is not due to a lack of expressivity. Both models are capable of representing the target solution. Rather, the difference arises from the encoding-induced spectral learning dynamics, with ternary encoding enabling more effective learning of the required frequency components. This mirrors known PINN failure modes in classical settings and highlights the role of spectral bias in quantum models for physics-informed learning.

## 6. Conclusion

In this work, we presented the first rigorous analysis of spectral bias in parameterised quantum circuits, grounded in their Fourier structure. We established a theoretical link between PQC training dynamics and frequency redundancy (the number of terms contributing to a given Fourier component) showing that the gradient magnitude at a frequency is upper bounded by its redundancy. This provides a concrete mechanism by which spectral bias arises in PQCs. Numerical experiments confirmed these predictions across multiple encoding schemes. Encodings that distribute redundancy uniformly, such as exponential Pauli encoding, learn all frequencies at comparable rates, while encodings with rapidly decaying redundancy, such as constant Pauli encoding, exhibit slower convergence at higher frequencies. Increased redundancy was also shown to improve robustness to global parameter perturbations. Despite these insights, our study has limitations. We focus on single-qubit encodings. Extending this analysis to broader classes of PQCs including trainable-frequency encodings, parameter sharing, mid-circuit measurements, and more complex circuit architectures remains an important direction for future work. Ultimately, a deeper understanding of spectral bias will help clarify where PQCs are most effective in practice, beyond the convection example considered here.

## Impact Statement

This paper presents work whose goal is to advance the field of Quantum Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

## References


Ansel, J., Yang, E., He, H., Gimelshein, N., Jain, A., Voznesensky, M., Bao, B., Bell, P., Berard, D., Burovski, E., Chauhan, G., Chourdia, A., Constable, W., Desmaison, A., DeVito, Z., Ellison, E., Feng, W., Gong, J., Gschwind, M., Hirsh, B., Huang, S., Kalambarkar, K., Kirsch, L., Lazos, M., Lezcano, M., Liang, Y., Liang, J., Lu, Y., Luk, C., Maher, B., Pan, Y., Puhrsch, C., Reso, M., Saroufim, M., Siraichi, M. Y., Suk, H., Suo, M., Tillet, P., Wang, E., Wang, X., Wen, W., Zhang, S., Zhao, X., Zhou, K., Zou, R., Mathews, A., Chanan, G., Wu, P., and Chintala, S. PyTorch 2: Faster Machine Learning Through Dynamic Python Bytecode Transformation and Graph Compilation. In *29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2 (ASPLOS '24)*. ACM, April 2024. doi: 10.1145/3620665.3640366. URL https://docs.pytorch.org/assets/pytorch2-2.pdf.

Basri, R., Jacobs, D., Kasten, Y., and Kritchman, S. The convergence rate of neural networks for learned functions of different frequencies, 2019. URL https://arxiv.org/abs/1906.00425.

Bergholm, V., Izaac, J., Schuld, M., Gogolin, C., Ahmed, S., Ajith, V., Alam, M. S., Alonso-Linaje, G., Akash-Narayanan, B., Asadi, A., Arrazola, J. M., Azad, U., Banning, S., Blank, C., Bromley, T. R., Cordier, B. A., Ceroni, J., Delgado, A., Matteo, O. D., Dusko, A., Garg, T., Guala, D., Hayes, A., Hill, R., Ijaz, A., Isacsson, T., Ittah, D., Jahangiri, S., Jain, P., Jiang, E., Khandelwal, A., Kottmann, K., Lang, R. A., Lee, C., Loke, T., Lowe, A., McKiernan, K., Meyer, J. J., Montañez-Barrera, J. A., Moyard, R., Niu, Z., O'Riordan, L. J., Oud, S., Panigrahi, A., Park, C.-Y., Polatajko, D., Quesada, N., Roberts, C., Sá, N., Schoch, I., Shi, B., Shu, S., Sim, S., Singh, A., Strandberg, I., Soni, J., Száva, A., Thabet, S., Vargas-Hernández, R. A., Vincent, T., Vitucci, N., Weber, M., Wierichs, D., Wiersema, R., Willmann, M., Wong, V., Zhang, S., and Killoran, N. Pennylane: Automatic differentiation of hybrid quantum-classical computations, 2022. URL https://arxiv.org/abs/1811.04968.

Bordelon, B., Canatar, A., and Pehlevan, C. Spectrum dependent learning curves in kernel regression and wide neural networks, 2021. URL https://arxiv.org/abs/2002.02561.

Cai, W., Li, X., and Liu, L. A phase shift deep neural network for high frequency approximation and wave problems, 2019. URL https://arxiv.org/abs/1909.11759.

Cai, W., Li, X., and Liu, L. A phase shift deep neural network for high frequency approximation and wave problems. *SIAM Journal on Scientific Computing*, 42(5): A3285–A3312, 2020. doi: 10.1137/19M1310050. URL https://doi.org/10.1137/19M1310050.

Cao, Y., Fang, Z., Wu, Y., Zhou, D.-X., and Gu, Q. Towards understanding the spectral bias of deep learning, 2020. URL https://arxiv.org/abs/1912.01198.

Caro, M. C., Gil-Fuster, E., Meyer, J. J., Eisert, J., and Sweke, R. 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.

Casas, B. and Cervera-Lierta, A. Multidimensional fourier series with quantum circuits. *Physical Review A*, 107(6), June 2023. ISSN 2469-9934. doi: 10.1103/physreva.107.062612. URL http://dx.doi.org/10.1103/PhysRevA.107.062612.

Fang, R. and Xu, Y. Addressing spectral bias of deep neural networks by multi-grade deep learning, 2024. URL https://arxiv.org/abs/2410.16105.

Fang, R., Xu, Y., and Yan, M. Inexact fppa for the $\ell_0$ sparse regularization problem, 2024. URL https://arxiv.org/abs/2404.17689.

Geifman, A., Galun, M., Jacobs, D., and Basri, R. On the spectral bias of convolutional neural tangent and gaussian process kernels, 2022. URL https://arxiv.org/abs/2203.09255.

Heimann, D., Schönhoff, G., Mounzer, E., Hohenfeld, H., and Kirchner, F. Learning capability of parametrized quantum circuits, 2024. URL https://arxiv.org/abs/2209.10345.

Jaderberg, B., Gentile, A. A., Berrada, Y. A., Shishenina, E., and Elfving, V. E. Let quantum neural networks choose their own frequencies. *Physical Review A*, 109(4), April 2024. ISSN 2469-9934. doi: 10.1103/physreva.109.042421. URL http://dx.doi.org/10.1103/PhysRevA.109.042421.

Jagtap, A. D., Kawaguchi, K., and Karniadakis, G. E. Adaptive activation functions accelerate convergence in deep and physics-informed neural networks. *Journal of Computational Physics*, 404:109136, 2020. ISSN 0021-9991. doi: https://doi.org/10.1016/j.jcp.2019.109136. URL https://www.sciencedirect.com/science/article/pii/S0021999119308411.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization, 2017. URL https://arxiv.org/abs/1412.6980.

Krishnapriyan, A. S., Gholami, A., Zhe, S., Kirby, R. M., and Mahoney, M. W. Characterizing possible failure modes in physics-informed neural networks, 2021. URL https://arxiv.org/abs/2109.01050.

Landman, J., Thabet, S., Dalyac, C., Mhiri, H., and Kashefi, E. Classically approximating variational quantum machine learning with random fourier features, 2022. URL https://arxiv.org/abs/2210.13200.

Lu, Y., Shen, L., and Xu, Y. Integral equation models for image restoration: high accuracy methods and fast algorithms. *Inverse Problems*, 26(4): 045006, mar 2010. doi: 10.1088/0266-5611/26/4/045006. URL https://dx.doi.org/10.1088/0266-5611/26/4/045006.

Mhiri, H., Monbroussou, L., Herrero-Gonzalez, M., Thabet, S., Kashefi, E., and Landman, J. Constrained and vanishing expressivity of quantum fourier models. *Quantum*, 9:1847, 2025. ISSN 2521-327X. doi: 10.22331/q-2025-09-03-1847. URL http://dx.doi.org/10.22331/q-2025-09-03-1847.

Papoulis, A. and Pillai, S. *Probability, Random Variables, and Stochastic Processes*. McGraw-Hill series in electrical engineering: Communications and signal processing. Tata McGraw-Hill, 2002. ISBN 9780070486584. URL https://books.google.com.sg/books?id=g6eUoWOlcQMC.

Pérez-Salinas, A., Cervera-Lierta, A., Gil-Fuster, E., and Latorre, J. I. 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 http://dx.doi.org/10.22331/q-2020-02-06-226.

Recio-Armengol, E., Ahmed, S., and Bowles, J. Train on classical, deploy on quantum: scaling generative quantum machine learning to a thousand qubits, 2025. URL https://arxiv.org/abs/2503.02934.

Schuld, M., Sweke, R., and Meyer, J. J. Effect of data encoding on the expressive power of variational quantum-machine-learning models. *Phys. Rev. A*, 103: 032430, Mar 2021. doi: 10.1103/PhysRevA.103.032430. URL https://link.aps.org/doi/10.1103/PhysRevA.103.032430.

Shin, S., Teo, Y. S., and Jeong, H. Exponential data encoding for quantum supervised learning. *Phys. Rev. A*, 107:012422, Jan 2023. doi: 10.1103/PhysRevA.107.012422. URL https://link.aps.org/doi/10.1103/PhysRevA.107.012422.

Stein, E. M. and Shakarchi, R. Fourier analysis: An introduction. 2003. URL https://api.semanticscholar.org/CorpusID:60789172.

Sweke, R., Recio-Armengol, E., Jerbi, S., Gil-Fuster, E., Fuller, B., Eisert, J., and Meyer, J. J. Potential and limitations of random fourier features for dequantizing quantum machine learning. *Quantum*, 9:1640, February 2025. ISSN 2521-327X. doi: 10.22331/q-2025-02-20-1640. URL http://dx.doi.org/10.22331/q-2025-02-20-1640.

Tancik, M., Srinivasan, P. P., Mildenhall, B., Fridovich-Keil, S., Raghavan, N., Singhal, U., Ramamoorthi, R., Barron, J. T., and Ng, R. Fourier features let networks learn high frequency functions in low dimensional domains, 2020. URL https://arxiv.org/abs/2006.10739.

Wang, S., Teng, Y., and Perdikaris, P. Understanding and mitigating gradient pathologies in physics-informed neural networks, 2020. URL https://arxiv.org/abs/2001.04536.

Wang, Y., Qi, B., Ferrie, C., and Dong, D. Trainability enhancement of parameterized quantum circuits via reduced-domain parameter initialization. *Phys. Rev. Appl.*, 22:054005, Nov 2024. doi: 10.1103/PhysRevApplied.22.054005. URL https://link.aps.org/doi/10.1103/PhysRevApplied.22.054005.

Wiebe, N., Kapoor, A., and Svore, K. M. Quantum deep learning, 2015. URL https://arxiv.org/abs/1412.3489.

Wiedmann, M., Periyasamy, M., and Scherer, D. D. Fourier analysis of variational quantum circuits for supervised learning, 2024. URL https://arxiv.org/abs/2411.03450.

Xu, Z. J. Understanding training and generalization in deep learning by fourier analysis, 2018. URL https://arxiv.org/abs/1808.04295.

Xu, Z.-Q. J., Zhang, Y., and Luo, T. Overview frequency principle/spectral bias in deep learning, 2024. URL https://arxiv.org/abs/2201.07395.

Zhang, K., Liu, L., Hsieh, M.-H., and Tao, D. Escaping from the barren plateau via gaussian initializations in deep variational quantum circuits, 2025. URL https://arxiv.org/abs/2203.09376.

Ziqi Liu, Z. L., Wei Cai, W. C., and Zhi-Qin John Xu, Z.-Q. J. X. Multi-scale deep neural network (mscalednn) for solving poisson-boltzmann equation in complex domains. *Communications in Computational Physics*, 28 (5):1970–2001, January 2020. ISSN 1815-2406. doi: 10.4208/cicp.oa-2020-0179. URL http://dx.doi.org/10.4208/cicp.OA-2020-0179.

## A. Compute Resources

All experiments were conducted using two eight-core Intel Xeon E5-2620 3.0GHz CPU along with 64GB of RAM and 16GB of storage for the results.

## B. Proof of Theorem 4.1

*Proof.* Assuming the target function $h(x)$ lies in the family of models expressible by $f(x, \theta)$ (and thus described by Equation 4), define the difference between the model output and the target

$$D(x) = |h(x) - f(x)| = \sum_{\omega=-N}^{N} c_{D_\omega} e^{i\omega x}, \tag{23}$$

where $c_{D_\omega}$ is the difference between the Fourier coefficients $c_\omega$ of $f(x, \theta)$ and $h(x)$ respectively. The mean-squared loss over the domain of $x \in [0, 2\pi]$ is thus

$$L = \frac{1}{2\pi} \int_0^{2\pi} D(x)^2 dx. \tag{24}$$

By Parseval's theorem (Stein & Shakarchi, 2003), this is simply equal to:

$$L = \sum_{\omega=-N}^{N} |c_{D_\omega}|^2 = \sum_{\omega=-N}^{N} L(\omega). \tag{25}$$

As a result, $L(\omega) = |c_{D_\omega}|^2$. The gradient of the loss with respect to a circuit parameter $\theta$ is then

$$\partial_\theta L = \sum_{\omega=-N}^{N} \partial_\theta L(\omega) \tag{26}$$

To ascertain whether a spectral bias exists within PQCs we focus on the terms in the summand of equation 26 to find the contribution from an $\omega$ at $\theta$, to identify whether gradients at certain values of $\omega$ are larger than others. Each term satisfies

$$\partial_\theta L(\omega) = c_{D_\omega}(\theta) \partial_\theta c^*_{D_\omega}(\theta) + c^*_{D_\omega}(\theta) \partial_\theta c_{D_\omega(\theta)}. \tag{27}$$

Since the coefficients of $h(x)$ have no dependence on $\theta$ their derivatives are zero. What remains are derivatives for the Fourier coefficients of $f(x, \theta)$ which we shall denote $c_{\omega_f}$:

$$\partial_\theta L(\omega) = c_{D_\omega}(\theta) \partial_\theta c^*_{\omega_f}(\theta) + c^*_{D_\omega}(\theta) \partial_\theta c_{\omega_f}(\theta), \tag{28}$$

The magnitude of the gradient is bounded as follows,

$$|\partial_\theta L(\omega)| = \left| 2\text{Re}\left( c^*_{D_\omega}(\theta) \partial_\theta c_{\omega_f}(\theta) \right) \right| \leq 2 \left| c_{D_\omega}(\theta) \right| \left| \partial_\theta c_{\omega_f}(\theta) \right|. \tag{29}$$

Which depends, intuitively, on the difference between the target and model coefficients, and the derivative of the model's coefficient itself. The derivative of $c_{\omega_f}$ can be further upper-bounded from its decomposition:

$$|\partial_\theta c_\omega| \leq \sum_{\substack{\mathbf{k},\mathbf{j} \in [d]^L \\ \Lambda_k - \Lambda_j = \omega}} |\partial_\theta a_{\mathbf{k},\mathbf{j}}(\theta)|. \tag{30}$$

$a_{\mathbf{k},\mathbf{j}}$ can be seen as a weighted sum over the eigenvalues of the Hermitian observable $O$, with weights formed by products of unitary matrix entries from the parameterised gates. Since any given $\theta$ is present only in a single layer, it can be shown that

$$|\partial_\theta a_{\mathbf{k},\mathbf{j}}(\theta)| \leq 2||O||_{\text{tr}}. \tag{31}$$

Altogether we find

$$|\partial_\theta L(\omega)| \leq 4R(\omega)||O||_{\text{tr}} \left| c_{D_\omega}(\theta) \right|. \tag{32}$$

□

## C. Proof of Theorem 4.2

*Proof.* Following the analysis for integer-frequency models (Proof B), we study the squared error loss in the frequency domain:

$$
\begin{aligned}
L &= \frac{1}{2\pi}\int_0^{2\pi} (f_{\not\perp}(x) - h(x))^2 dx \\
&= \frac{1}{2\pi}\int_0^{2\pi} (\sum_{\omega\in\Omega_{\not\perp}} c_{D_\omega} e^{i\omega x})^2 dx \\
&= \frac{1}{2\pi}\int_0^{2\pi} \sum_{\omega}\sum_{\omega'} c_{D_\omega} c^*_{D_{\omega'}} e^{ix(\omega-\omega')} dx \\
&= \sum_{\omega}\sum_{\omega'} c_{D_\omega} c^*_{D_{\omega'}} e^{i\pi(\omega-\omega')} \mathrm{sinc}(\pi(\omega-\omega')),
\end{aligned}
\tag{33}
$$

with cross-terms ($\omega \neq \omega'$). There is now an ambiguity in assigning part of the total sum loss to a certain frequency $L(\omega)$, which was not present in the integer-only spectrum models. Without loss of generality, we choose to assign $L(\omega) = \mathrm{Re}\left(c_\omega \sum_{\omega'} c^*_{\omega'} e^{i\pi(\omega-\omega')}\mathrm{sinc}(\pi(\omega-\omega'))\right)$. To determine an upper bound for $|\partial_\theta L(\omega)|$, let us first denote $\pi(\omega-\omega') = \beta_{\omega'}$ and thus:

$$
|\partial_\theta L(\omega)| = \left|\partial_\theta \mathrm{Re}\left(\sum_{\omega'} e^{i\beta_{\omega'}}\mathrm{sinc}(\beta_{\omega'}) c_{D_\omega} c^*_{D_{\omega'}}\right)\right| = \left|\mathrm{Re}\left(\sum_{\omega'} e^{i\beta_{\omega'}}\mathrm{sinc}(\beta_{\omega'})\partial_\theta\left(c_{D_\omega} c^*_{D_{\omega'}}\right)\right)\right| \tag{34}
$$

and using properties of the absolute value:

$$
|\partial_\theta L(\omega)| \leq \left|\sum_{\omega'} e^{i\beta_{\omega'}}\mathrm{sinc}(\beta_{\omega'})\partial_\theta\left(c_{D_\omega} c^*_{D_{\omega'}}\right)\right| \leq \sum_{\omega'} |\mathrm{sinc}(\beta_{\omega'})| \times \left|\partial_\theta\left(c_{D_\omega} c^*_{D_{\omega'}}\right)\right|. \tag{35}
$$

The derivative in a single element in this sum satisfies:

$$
\left|\partial_\theta(c_{D_\omega} c^*_{D_{\omega'}})\right| = \left|c^*_{D_{\omega'}}\partial_\theta c_{f_\omega} + c_{D_\omega}\partial_\theta c^*_{f\omega'}\right| \leq |\partial_\theta c_{f_\omega}|\left|c^*_{D_{\omega'}}\right| + |c_{D_\omega}|\left|\partial_\theta c^*_{f_{\omega'}}\right|, \tag{36}
$$

and using $|\partial_\theta c_\omega| \leq 2R(\omega)||O||_{\mathrm{tr}}$, established in Proof B,

$$
\left|\partial_\theta(c_{D_\omega} c^*_{D_{\omega'}})\right| \leq 2||O||_{\mathrm{tr}}\left(\left|c^*_{D_{\omega'}}\right| R(\omega) + |c_{D_\omega}| R(\omega')\right). \tag{37}
$$

Which results in

$$
|\partial_\theta L(\omega)| \leq 2||O||_{\mathrm{tr}} \sum_{\omega'} |\mathrm{sinc}(\pi(\omega-\omega'))| \times \left(\left|c^*_{D_{\omega'}}\right| R(\omega) + |c_{D_\omega}| R(\omega')\right) \tag{38}
$$

□

## D. Small angle approximation

### D.1. Expected value

We analyse the expected gradient magnitude of the Fourier coefficients under a small-angle initialization of the variational parameters. Assuming each parameter is drawn independently from a Gaussian distribution $\theta \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$ with $\sigma \ll 1$, our goal is to compute (maybe add second too?)

$$
\mathbb{E}[\,|\partial_{\theta_k} c_\omega(\boldsymbol{\theta})|\,], \tag{39}
$$

which is the expected absolute gradient of a Fourier coefficient with respect to a single parameter $\theta_k$. We factorise the monomial representation of $c_\omega$ and focus initially only on the terms dependent on trainable parameters $\boldsymbol{\theta}$ and let

$$g(\boldsymbol{\theta}) = \prod_{j=1}^{w} g_j(\theta_j), \qquad g_j(\theta_j) = \sin(\theta_j)^{s'_j} \cos(\theta_j)^{c'_j}, \tag{40}$$

where $s'_j, c'_j \in \mathbb{Z}_{\geq 0}$. Fixing an index $k$, taking the partial derivative with respect to $\theta_k$ isolates the $k$-th term

$$\frac{\partial g}{\partial \theta_k} = \partial_{\theta_k} g_k(\theta_k) \prod_{j \neq k} g_j(\theta_j), \tag{41}$$

with

$$\partial_{\theta_k} g_k(\theta_k) = s'_k \, \sin^{s'_k - 1}(\theta_k) \cos^{c'_k + 1}(\theta_k) - c'_k \, \sin^{s'_k + 1}(\theta_k) \cos^{c'_k - 1}(\theta_k). \tag{42}$$

By independence of the $\theta_j$ and multiplicativity of the absolute value,

$$\mathbb{E}\left[ \left| \frac{\partial g}{\partial \theta_k} \right| \right] = \mathbb{E}[\,|g'_k(\theta_k)|\,] \prod_{j \neq k} \mathbb{E}[\,|g_j(\theta_j)|\,]. \tag{43}$$

We now begin with applying the small-angle approximation, firstly by writing $\partial_{\theta_k} g_k(\theta_k) = A(\theta_k) - B(\theta_k)$ with

$$A(\theta_k) = s'_k \, \sin^{s'_k - 1}(\theta_k) \cos^{c'_k + 1}(\theta_k), \tag{44}$$

$$B(\theta_k) = c'_k \, \sin^{s'_k + 1}(\theta_k) \cos^{c'_k - 1}(\theta_k). \tag{45}$$

Inserting the expansions of $\sin$ and $\cos$ around zero:

$$\sin^r(\theta) = \theta^r \left(1 - \frac{r}{6}\theta^2 + O(\theta^4)\right), \tag{46}$$

$$\cos^t(\theta) = 1 - \frac{t}{2}\theta^2 + O(\theta^4), \tag{47}$$

Applying these approximations to the undifferentiated monomials $g_j(\theta_j) = \sin^{s'_j}(\theta_j) \cos^{c'_j}(\theta_j)$ yields the local expansion:

$$g_j(\theta_j) = \theta_j^{s'_j} \left(1 - \left(\frac{s'_j}{6} + \frac{c'_j}{2}\right) \theta_j^2 + O(\theta_j^4)\right). \tag{48}$$

For the differentiated term $\partial_{\theta_k} g_k$, we proceed by substituting the trigonometric expansions into the components $A(\theta_k)$ and $B(\theta_k)$ defined above:

$$A(\theta_k) = s'_k \, \theta^{s'_k - 1} \Big(1 - \tfrac{s'_k - 1}{6}\theta_k^2 + O(\theta_k^4)\Big)\Big(1 - \tfrac{c'_k + 1}{2}\theta_k^2 + O(\theta_k^4)\Big) \tag{49}$$

$$= s'_k \, \theta^{s'_k - 1} \left(1 - \left(\tfrac{s'_k - 1}{6} + \tfrac{c'_k + 1}{2}\right)\theta_k^2 + O(\theta_k^4)\right), \tag{50}$$

$$B(\theta_k) = c'_k \, \theta^{s'_k + 1} \Big(1 - \left(\tfrac{s'_k + 1}{6} + \tfrac{c'_k - 1}{2}\right)\theta_k^2 + O(\theta_k^4)\Big). \tag{51}$$

Subtracting gives

$$\partial_{\theta_k} g_k(\theta) = s'_k \, \theta^{s'_k - 1} - \frac{(s'_k + 3c'_k)(s'_k + 2)}{6} \, \theta^{s'_k + 1} + O(\theta^{s'_k + 3}). \tag{52}$$

Factorising out the leading term and taking the absolute value,

$$|\partial_{\theta_k} g_k(\theta)| = s'_k |\theta|^{s'_k - 1} \left(1 - \frac{(s'_k + 3c'_k)(s'_k + 2)}{6 s'_k} \, \theta_k^2 + O(\theta_k^4)\right). \tag{53}$$

This assumes the term in the bracket is positive which requires

$$|\theta_k| \lesssim \sqrt{\frac{6s'_k}{(s'_k+3c'_k)(s'_k+2)}}, \quad s'_k \geq 1. \tag{54}$$

Under regimes with no weight sharing ($s'_k, c'_k \leq 1$), the tightest constraint occurs at $s'_k = 1, c'_k = 0$, requiring $|\theta_k| \lesssim 1/\sqrt{2} \approx 0.71$. This condition is satisfied with high probability for Gaussian initialization with $\sigma \ll 1$. For the case $s'_k = 0$, the sine power vanishes and the expansion simplifies to

$$\partial_{\theta_k} g_k(\theta_k) = -c'_k \, \sin(\theta_k) \cos^{c'_k-1}(\theta_k) = -c'_k \theta_k + O(\theta_k^3), \tag{55}$$

yielding the magnitude

$$|\partial_{\theta_k} g_k(\theta_k)| \approx c'_k \, |\theta_k| + O(\theta_k^3). \tag{56}$$

The absolute moments of Gaussian variables for $\theta \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$ are (Papoulis & Pillai, 2002)

$$M_r := \mathbb{E}[|\theta|^r] = \sigma^r 2^{r/2} \frac{\Gamma\left(\frac{r+1}{2}\right)}{\sqrt{\pi}}. \tag{57}$$

Along with the recurrence relation

$$\frac{M_{r+2}}{M_r} = \sigma^2 (r+1). \tag{58}$$

We can then use these moments to find the expectation value of $\mathbb{E}[\,|\partial_{\theta_k} g(\theta_k)|\,]$ for $s'_k \geq 1$:

$$\mathbb{E}[\,|\partial_{\theta_k} g_k(\theta_k)|\,] = s'_k M_{s'_k-1} \left(1 - \frac{(s'_k+3c'_k)(s'_k+2)}{6}\sigma^2 + O(\sigma^4)\right) \tag{59}$$

For $s'_k = 0$,

$$\mathbb{E}[\,|\partial_{\theta_k} g_k(\theta_k)|\,] = c'_k \, \mathbb{E}[|\theta_k|] + O(\sigma^3) = c'_k \, \sigma \sqrt{\tfrac{2}{\pi}} + O(\sigma^3). \tag{60}$$

Now we can combine this, with the full expression for $g(\theta)$, whereby the expectations of $g_j(\theta_j)$ follows very similarly from above (maybe also state expansion above). For $s'_k \geq 1$, combining Equation 59 with the expectation value of Equation 48 gives:

$$\begin{aligned}\mathbb{E}[\,|\partial_{\theta_k} g(\boldsymbol{\theta})|\,] = s'_k M_{s'_{k-1}} \left(1 - \tfrac{(s'_k+3c'_k)(s'_k+2)}{6}\sigma^2 + O(\sigma^4)\right) \\ \times \prod_{j\neq k} \left(M_{s_j}\left(1 - \left(\tfrac{s_j}{6} + \tfrac{c_j}{2}\right)(s_j+1)\sigma^2 + O(\sigma^4)\right)\right).\end{aligned} \tag{61}$$

and for $s'_k = 0$

$$\mathbb{E}[\,|\partial_{\theta_k} g(\boldsymbol{\theta})|\,] = (c'_k \, M_1) \prod_{j\neq k} \left(M_{s_j}\left(1 - \left(\tfrac{s_j}{6} + \tfrac{c_j}{2}\right)(s_j+1)\sigma^2 + O(\sigma^4)\right)\right). \tag{62}$$

The Fourier coefficient $c_\omega$ (Equation 13)is a linear combination of such monomials, with prefactors $k_{s,c,s',c'}$, phases, and combinatorial factors $p(s,c,\omega)$, thus taking the expectation value its gradient:

$$\mathbb{E}[\,|\partial_{\theta_k} c_\omega(\boldsymbol{\theta})|\,] \leq \sum_{s,c,s',c'} |k_{s,c,s',c'}| \, 2^{-\sum_j (s_j+c_j)} |(-i)^{\sum_j s_j}| \, p(s,c,\omega) \, \mathbb{E}[\,|\partial_{\theta_k} g(\boldsymbol{\theta})|\,]. \tag{63}$$

Likewise, $\mathbb{E}[c_\omega(\theta)]$:

$$\mathbb{E}[\,|c_\omega(\boldsymbol{\theta})|\,] \leq \sum_{s,c,s',c'} |k_{s,c,s',c'}| \, 2^{-\sum_j (s_j+c_j)} (-i)^{\sum_j s_j} \, p(s,c,\omega) \, \mathbb{E}[\,|g(\boldsymbol{\theta})|\,]. \tag{64}$$

Noting that the prefactors $|k_{s,c,s',c'}|$ are either 0 or 1, we rewrite the expectation in Eq. 63 by explicitly enumerating the $R(\omega)$ non-zero contributing variational paths. Let the index $r$ denote the $r$-th active path contributing to frequency $\omega$, with associated exponents $\mathbf{s}^{(r)}, \mathbf{c}^{(r)}, \mathbf{s}'^{(r)}, \mathbf{c}'^{(r)}$. For the dominant case ($s'_k \geq 1$), the expected gradient is bounded by the sum over these $R(\omega)$ components to leading order, explicitly subsituting in the moments from Equation 57:

$$\mathbb{E}[\,|\partial_{\theta_k} c_\omega(\boldsymbol{\theta})|\,] \lesssim \sum_{r=1}^{R(\omega)} \frac{|p(\mathbf{s}^{(r)}, \mathbf{c}^{(r)}, \omega)|}{2^{\sum(s_j^{(r)}+c_j^{(r)})}} \left[ \frac{s_k'^{(r)} 2^{\frac{s_k'^{(r)}-1}{2}} \Gamma(\frac{s_k'^{(r)}}{2})}{\sqrt{\pi}} \prod_{j\neq k}^{w^{(r)}} \frac{2^{\frac{s_j'^{(r)}}{2}} \Gamma(\frac{s_j'^{(r)}+1}{2})}{\sqrt{\pi}} \right] \sigma^{S^{(r)}-1}, \tag{65}$$

and similarly for the suppressed case ($s'_k = 0$):

$$\mathbb{E}[\,|\partial_{\theta_k} c_\omega(\boldsymbol{\theta})|\,] \lesssim \sum_{r=1}^{R(\omega)} \frac{|p(\mathbf{s}^{(r)}, \mathbf{c}^{(r)}, \omega)|}{2^{\sum(s_j^{(r)}+c_j^{(r)})}} \left[ c_k'^{(r)} \sqrt{\frac{2}{\pi}} \prod_{j\neq k}^{w^{(r)}} \frac{2^{\frac{s_j'^{(r)}}{2}} \Gamma(\frac{s_j'^{(r)}+1}{2})}{\sqrt{\pi}} \right] \sigma^{S^{(r)}+1}. \tag{66}$$

Here $S^{(r)} = \sum_j s_j'^{(r)}$ is the total sine-degree of the $r$-th monomial. These expressions explicitly link the expected value of the gradient magnitude of frequency components to the redundancy $R(\omega)$, showing it is a coherent sum of $R(\omega)$ terms, each scaled by the initialization variance $\sigma$. For the case of no weight sharing and thus $s'_k, c'_k$ at most are one. To first order

$$\mathbb{E}[\,|\partial_{\theta_k} c_\omega(\boldsymbol{\theta})|\,] \lesssim \sum_{r=1}^{R(\omega)} \frac{|p(\mathbf{s}^{(r)}, \mathbf{c}^{(r)}, \omega)|}{2^{\sum(s_j^{(r)}+c_j^{(r)})}} \left[ \left(\frac{2}{\pi}\right)^{\frac{w^{(r)}-1}{2}} \right] \sigma^{S^{(r)}-1}. \tag{67}$$

### D.2. Second Moment

To compute the second moments, we require the moments of the squared Gaussian variable $\theta^2$ (where $\theta \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$), defined as:

$$\mu_{2n} := \mathbb{E}[\theta^{2n}] = (2n-1)!!\,\sigma^{2n} \tag{68}$$

The recurrence relation for these moments is $\mu_{2n+2} = (2n+1)\sigma^2 \mu_{2n}$. We evaluate the expected squared gradient for the monomial $g(\boldsymbol{\theta})$:

$$\mathbb{E}\left[\left(\frac{\partial g}{\partial \theta_k}\right)^2\right] = \mathbb{E}\left[(\partial_{\theta_k} g_k(\theta_k))^2\right] \prod_{j\neq k} \mathbb{E}\left[g_j(\theta_j)^2\right], \tag{69}$$

as well as $\mathbb{E}[g(\theta)]$.

**Local Parameter Moments:** First we find the expectation values for the individual terms comprising $g(\theta)$.

*Undifferentiated Terms ($g_j$):* Using the local expansion for $g_j(\theta_j)$ derived previously in (48), $g_j(\theta_j) \approx \theta_j^{s'_j}(1 - (s'_j/6 + c'_j/2)\theta_j^2)$, we square the expression and retain terms up to $\mathcal{O}(\theta_j^2)$ relative to the leading order:

$$g_j(\theta_j)^2 \approx \theta_j^{2s'_j} \left(1 - \left(\frac{s'_j}{3} + c'_j\right)\theta_j^2\right). \tag{70}$$

Taking the expectation yields:

$$\mathbb{E}[g_j(\theta_j)^2] \approx \mu_{2s'_j} - \left(\frac{s'_j}{3} + c'_j\right)\mu_{2s'_j+2} \tag{71}$$

$$= (2s'_j - 1)!!\,\sigma^{2s'_j}\left(1 - \left(\frac{s'_j}{3} + c'_j\right)(2s'_j+1)\sigma^2 + \mathcal{O}(\sigma^4)\right). \tag{72}$$

*Differentiated Term ($\partial_{\theta_k} g_k$):* We consider the two cases for the sine-degree $s'_k$, $s'_k \geq 1$ and $s'_k = 0$.

- *Case ($s'_k \geq 1$):* The derivative scales with the sine power reduced by one. Squaring the expansion derived in Eq. (52) retaining terms up to $\mathcal{O}(\theta_k^2)$ relative to leading order:

$$\mathbb{E}[(\partial_{\theta_k} g_k)^2] \approx (s'_k)^2 \mu_{2(s'_k-1)} - \frac{s'_k(s'_k+3c'_k)(s'_k+2)}{3}\mu_{2s'_k} \tag{73}$$

$$= (s'_k)^2 (2s'_k-3)!!\, \sigma^{2(s'_k-1)} \left(1 - \frac{(s'_k+3c'_k)(s'_k+2)(2s'_k-1)}{3s'_k}\sigma^2\right). \tag{74}$$

- *Case ($s'_k = 0$):* Squaring Eq.(56) gives $(\partial_{\theta_k} g_k)^2 \approx (c'_k)^2 \theta_k^2$, thus:

$$\mathbb{E}[(\partial_{\theta_k} g_k)^2] \approx (c'_k)^2 \sigma^2. \tag{75}$$

**Global Monomial Moments:** We now combine these local results to find the expected moments for the full monomial $g(\boldsymbol{\theta}) = \prod g_j(\theta_j)$.

*Undifferentiated Monomial Magnitude ($g(\theta)$):* The expected squared magnitude of the monomial is the product of the undifferentiated expectations:

$$\mathbb{E}[g(\boldsymbol{\theta})^2] = \prod_{j=1}^{w} \mathbb{E}[g_j(\theta_j)^2] \approx \left[\prod_{j=1}^{w} (2s'_j-1)!!\right] \sigma^{2S}. \tag{76}$$

where $S = \sum s'_j$ is the total sine-degree.

*Monomial Gradient Magnitude:* The expected squared gradient combines the differentiated term $k$ with the product of undifferentiated terms $j \neq k$.

- *Case $s'_k \geq 1$:*

$$\mathbb{E}\left[(\partial_{\theta_k} g)^2\right] = \left((s'_k)^2 (2s'_k-3)!!\, \sigma^{2(s'_k-1)} \left[1 - \mathcal{O}(\sigma^2)\right]\right) \times \prod_{j\neq k} \left((2s'_j-1)!!\, \sigma^{2s'_j} \left[1 - \mathcal{O}(\sigma^2)\right]\right). \tag{77}$$

- *Case $s'_k = 0$:*

$$\mathbb{E}\left[(\partial_{\theta_k} g)^2\right] = \left((c'_k)^2 \sigma^2\right) \prod_{j\neq k} \left((2s'_j-1)!!\, \sigma^{2s'_j} \left[1 - \mathcal{O}(\sigma^2)\right]\right). \tag{78}$$

**Fourier Coefficient:** The Fourier coefficient $c_\omega$ is a linear combination of these monomials. We will want to find the Root Mean Square (RMS) magnitudes. To bound the RMS amplitudes, we apply the triangle inequality for the $L_2$ norm ($\sqrt{\mathbb{E}[(\sum X)^2]} \leq \sum \sqrt{\mathbb{E}[X^2]}$).

*RMS Coefficient Magnitude:* Summing the RMS of the monomials $\sqrt{\mathbb{E}[g^2]} \sim \sigma^S$:

$$\sqrt{\mathbb{E}[|c_\omega(\boldsymbol{\theta})|^2]} \lesssim \sum_{r=1}^{R(\omega)} \frac{|p(\mathbf{s}^{(r)}, \mathbf{c}^{(r)}, \omega)|}{2^{\sum(s_j^{(r)}+c_j^{(r)})}} \left[\prod_j \sqrt{(2s_j'^{(r)}-1)!!}\right] \sigma^{S^{(r)}}. \tag{79}$$

*RMS Gradient Magnitude:* Summing the RMS of the gradients $\sqrt{\mathbb{E}[(\partial g)^2]}$: For the dominant case ($s'_k \geq 1$), we obtain:

$$\sqrt{\mathbb{E}[|\partial_{\theta_k} c_\omega|^2]} \lesssim \sum_{r=1}^{R(\omega)} \frac{|p(\mathbf{s}^{(r)}, \mathbf{c}^{(r)}, \omega)|}{2^{\sum(s_j^{(r)}+c_j^{(r)})}} \left[s_k'^{(r)} \sqrt{(2s_k'^{(r)}-3)!!} \prod_{j\neq k} \sqrt{(2s_j'^{(r)}-1)!!}\right] \sigma^{S^{(r)}-1}. \tag{80}$$

For the suppressed case ($s'_k = 0$), the derivative scales linearly with $\theta_k$, increasing the total order in $\sigma$:

$$\sqrt{\mathbb{E}[|\partial_{\theta_k} c_\omega|^2]} \lesssim \sum_{r=1}^{R(\omega)} \frac{|p(\mathbf{s}^{(r)}, \mathbf{c}^{(r)}, \omega)|}{2^{\sum(s_j^{(r)}+c_j^{(r)})}} \left[ c'^{(r)} k \prod_{j\neq k} \sqrt{(2s_j'^{(r)}-1)!!} \right] \sigma^{S^{(r)}-s'_k+1}. \tag{81}$$

For the standard case of no weight sharing ($s'_j = 1, c'_j \in \{0,1\}$), the double factorials simplify ($(-1)!! = 1$ and $1!! = 1$), and the expression simplifies purely to the redundancy and scaling factors:

$$\sqrt{\mathbb{E}[|\partial_{\theta_k} c_\omega|^2]} \lesssim \sum_{r=1}^{R(\omega)} \frac{|p(\mathbf{s}^{(r)}, \mathbf{c}^{(r)}, \omega)|}{2^{\sum(s_j^{(r)}+c_j^{(r)})}} \sigma^{w-1}. \tag{82}$$

This confirms that the second moment (RMS) follows the exact same $\sigma^{w-1}$ scaling law as the first moment (Mean Absolute), validating the tightness of the bounds used in the main theorem.

### D.3. Expected Upper Bound

*Proof.* We seek a rigorous upper bound on the expected magnitude of the loss gradient. We begin with the expression for the gradient of the loss at a specific frequency $\omega$, derived in Eq. (26):

$$|\partial_{\theta_k} L(\omega)| = \left| 2\mathrm{Re}\left( c^*_{D_\omega} \partial_{\theta_k} c_{\omega_f} \right) \right|. \tag{83}$$

First, we apply the inequality $|\mathrm{Re}(z)| \leq |z|$ to bound the magnitude:

$$|\partial_{\theta_k} L(\omega)| \leq 2\, |c_{D_\omega}(\boldsymbol{\theta})| \left| \partial_{\theta_k} c_{\omega_f}(\boldsymbol{\theta}) \right|. \tag{84}$$

Next, we substitute the definition of the model error $c_{D_\omega} = c_{\omega_h} - c_{\omega_f}(\boldsymbol{\theta})$ and apply the triangle inequality $|A - B| \leq |A| + |B|$:

$$|\partial_{\theta_k} L(\omega)| \leq 2 \left( |c_{\omega_h}| + |c_{\omega_f}(\boldsymbol{\theta})| \right) \left| \partial_{\theta_k} c_{\omega_f}(\boldsymbol{\theta}) \right|. \tag{85}$$

We now take the expectation over the parameter distribution $\boldsymbol{\theta} \sim \mathcal{N}(0, \sigma^2 \mathbf{I})$. Using the linearity of the expectation operator, we split the bound into two distinct terms:

$$\mathbb{E}\left[ |\partial_{\theta_k} L(\omega)| \right] \leq 2|c_{\omega_h}| \mathbb{E}\left[ \left| \partial_{\theta_k} c_{\omega_f} \right| \right] + 2\mathbb{E}\left[ \left| c_{\omega_f} \right| \left| \partial_{\theta_k} c_{\omega_f} \right| \right]. \tag{86}$$

Applying Cauchy-Schwarz:

$$\mathbb{E}\left[ |\partial_{\theta_k} L(\omega)| \right] \leq 2|c_{\omega_h}| \mathbb{E}\left[ \left| \partial_{\theta_k} c_{\omega_f} \right| \right] + 2\sqrt{\mathbb{E}[|c_{\omega_f}|^2] \mathbb{E}[|\partial_{\theta_k} c_{\omega_f}|^2]}. \tag{87}$$

We evaluate this bound by substituting the explicit moment sums derived over the $R(\omega)$ active variational paths. We consider the dominant regime where the differentiated parameter $\theta_k$ has sine-dependence ($s'_k \geq 1$), which provides the leading-order contribution.

*First term:* Substituting the explicit expression for the expected gradient magnitude eq..:

$$2|c_{\omega_h}| \mathbb{E}[|\partial_{\theta_k} c_{\omega_f}|] \lesssim 2|c_{\omega_h}| \sum_{r=1}^{R(\omega)} \mathcal{C}^{(r)} \left[ \frac{s_k'^{(r)} 2^{\frac{s_k'^{(r)}-1}{2}} \Gamma(\frac{s_k'^{(r)}}{2})}{\sqrt{\pi}} \prod_{j\neq k} \frac{2^{\frac{s_j'^{(r)}}{2}} \Gamma(\frac{s_j'^{(r)}+1}{2})}{\sqrt{\pi}} \right] \sigma^{S^{(r)}-1}, \tag{88}$$

where $\mathcal{C}^{(r)} = |p(\mathbf{s}^{(r)}, \mathbf{c}^{(r)}, \omega)|\, 2^{-\sum(s_j^{(r)}+c_j^{(r)})}$ encapsulates the combinatorial and trigonometric scaling factors for the $r$-th path.

*Second term:* This term involves the product of the RMS coefficient magnitude and the RMS gradient. The RMS coefficient magnitude scales as $\sigma^{S^{(r)}}$:

$$\sqrt{\mathbb{E}[|c_{\omega_f}|^2]} \lesssim \sum_{r=1}^{R(\omega)} \mathcal{C}^{(r)} \left[ \prod_j \sqrt{(2s_j'^{(r)}-1)!!} \right] \sigma^{S^{(r)}}. \tag{89}$$

The RMS gradient is given by the explicit second-moment sum:

$$\sqrt{\mathbb{E}[|\partial_{\theta_k} c_{\omega_f}|^2]} \lesssim \sum_{r=1}^{R(\omega)} \mathcal{C}^{(r)} \left[ s_k'^{(r)} \sqrt{(2s_k'^{(r)}-3)!!} \prod_{j\neq k} \sqrt{(2s_j'^{(r)}-1)!!} \right] \sigma^{S^{(r)}-1}. \tag{90}$$

Multiplying these two factors, the self-interaction term scales as $\sigma^{S^{(r)}} \times \sigma^{S^{(r)}-1} = \sigma^{2S^{(r)}-1}$.

Comparing the two components of the upper bound, we observe that the target-interaction term scales as $\sigma^{S-1}$ while the self-interaction term scales as $\sigma^{2S-1}$. In the small-angle initialization regime ($\sigma \ll 1$), the first term dominates. Thus, the expected gradient is bounded by the explicit sum:

$$\mathbb{E}\left[|\partial_{\theta_k} L(\omega)|\right] \lesssim 2|c_{\omega_h}| \sum_{r=1}^{R(\omega)} \frac{|p(\mathbf{s}^{(r)}, \mathbf{c}^{(r)}, \omega)|}{2^{\sum(s_j^{(r)}+c_j^{(r)})}} \left[ \frac{s_k'^{(r)} 2^{\frac{s_k'^{(r)}-1}{2}} \Gamma(\frac{s_k'^{(r)}}{2})}{\sqrt{\pi}} \prod_{j\neq k} \frac{2^{\frac{s_j'^{(r)}}{2}} \Gamma(\frac{s_j'^{(r)}+1}{2})}{\sqrt{\pi}} \right] \sigma^{S^{(r)}-1}. \tag{91}$$

This result confirms that the gradient signal is a sum over $R(\omega)$ coherent paths. While the geometric prefactors (Gamma functions) grow with the sine-powers $s_j'$, they are dominated by the exponential suppression factor $\sigma^{S^{(r)}-1}$.

*No weight sharing:* In the case of with no weight sharing of variational parameters, we have $s_j'^{(r)} = 1$ at most for all active parameters. The sine-degree $S^{(r)}$ becomes the variational depth $d_\omega$. The geometric prefactors simplify significantly:

$$\frac{\Gamma(1/2)}{\sqrt{\pi}} = 1 \quad \text{and} \quad \frac{2^{1/2}\Gamma(1)}{\sqrt{\pi}} = \sqrt{\frac{2}{\pi}}. \tag{92}$$

The bound reduces to the concise scaling law:

$$\mathbb{E}\left[|\partial_{\theta_k} L(\omega)|\right] \lesssim 2|c_{\omega_h}| \sum_{r=1}^{R(\omega)} \frac{|p(\mathbf{s}^{(r)}, \mathbf{c}^{(r)}, \omega)|}{2^{\sum(s_j^{(r)}+c_j^{(r)})}} \left(\frac{2}{\pi}\right)^{\frac{d_\omega-1}{2}} \sigma^{d_\omega-1}. \tag{93}$$

In the standard case of Pauli encoding with fixed circuit depth $d$, the exponential suppression factor $\sigma^{d-1}$ applies uniformly across the spectrum. However, the result demonstrates that the expected gradient magnitude for a frequency $\omega$ is directly proportional to its redundancy $R(\omega)$ and the target coefficient magnitude $|c_{\omega_h}|$. This scaling establishes the mechanism of spectral bias through combinatorial redundancy. □

### D.4. Expected Non-Integer Upper Bound

*Proof.* We extend the expected gradient analysis to the case of non-integer frequencies. Starting from the deterministic bound derived in Proof C:

$$|\partial_{\theta_k} L(\omega)| \leq \sum_{\omega'} |\text{sinc}(\pi(\omega-\omega'))| \left( \left|{c^*}_{D_{\omega'}}\right| \left|\partial_{\theta_k} c_{\omega_f}\right| + |c_{D_\omega}| \left|\partial_{\theta_k} c_{\omega' f}\right| \right). \tag{94}$$

We take the expectation over the parameter distribution $\boldsymbol{\theta} \sim \mathcal{N}(0, \sigma^2\mathbf{I})$. Using the linearity of expectation, we bring the operator inside the summation:

$$\mathbb{E}[|\partial\theta_k L(\omega)|] \leq \sum_{\omega'} |\text{sinc}(\pi(\omega-\omega'))| \left( \mathbb{E}\left[|c_{D_{\omega'}}||\partial_{\theta_k} c_{\omega_f}|\right] + \mathbb{E}\left[|c_{D_\omega}||\partial_{\theta_k} c_{\omega' f}|\right] \right). \tag{95}$$

We apply the same bounding strategy used for the integer case. For each interaction term $\mathbb{E}[|c_{D_\omega}||\partial c_{\omega_f}|]$, we use Cauchy-Schwarz and the triangle inequality to separate the target contribution from the model self-interaction. For small initialization variance $\sigma \ll 1$, the terms are dominated by the target coefficients $|h\omega|$ and $|h_{\omega'}|$. Neglecting the higher-order self-interaction terms ($\mathcal{O}(\sigma^{2d-1})$), the expectation simplifies to:

$$\mathbb{E}[|\partial_{\theta_k} L(\omega)|] \lesssim \sum_{\omega'} |\text{sinc}(\pi(\omega-\omega'))| \left( |h_{\omega'}| \mathbb{E}[|\partial_{\theta_k} c_{\omega_f}|] + |h_\omega| \mathbb{E}[|\partial_{\theta_k} c_{\omega'_f}|] \right). \tag{96}$$

We now substitute the explicit moment sums derived in Sections D.1 and D.2. For the dominant regime where the differentiated parameter has sine-dependence ($s'_k \geq 1$):

**1. Self Gradient Contribution** ($\mathbb{E}[|\partial c_{\omega_f}|]$)**:**The gradient at frequency $\omega$ is a sum over its $R(\omega)$ active paths.

$$\mathbb{E}[|\partial_{\theta_k} c_{\omega_f}|] \lesssim \sum_{r=1}^{R(\omega)} \frac{|p(\mathbf{s}^{(r)}, \mathbf{c}^{(r)}, \omega)|}{2^{\sum(s_j^{(r)}+c_j^{(r)})}} \left[ \frac{s_k'^{(r)} 2^{\frac{s_k'^{(r)}-1}{2}} \Gamma(\frac{s'^{(r)}k}{2})}{\sqrt{\pi}} \prod_{j\neq k} \frac{2^{\frac{s_j'^{(r)}}{2}} \Gamma(\frac{s_j'^{(r)}+1}{2})}{\sqrt{\pi}} \right] \sigma^{S^{(r)}-1}. \tag{97}$$

**2. Leakage Gradient Contribution** ($\mathbb{E}[|\partial c_{\omega'_f}|]$)**:**The gradient at the neighbor frequency $\omega'$ is a sum over its $R(\omega')$ active paths (indexed by $q$).

$$\mathbb{E}[|\partial_{\theta_k} c_{\omega' f}|] \lesssim \sum_{q=1}^{R(\omega')} \frac{|p(\mathbf{s}^{(q)}, \mathbf{c}^{(q)}, \omega')|}{2^{\sum(s_j^{(q)}+c_j^{(q)})}} \left[ \frac{s_k'^{(q)} 2^{\frac{s_k'^{(q)}-1}{2}} \Gamma(\frac{s'^{(q)}k}{2})}{\sqrt{\pi}} \prod_{j\neq k} \frac{2^{\frac{s_j'^{(q)}}{2}} \Gamma(\frac{s_j'^{(q)}+1}{2})}{\sqrt{\pi}} \right] \sigma^{S^{(q)}-1}. \tag{98}$$

Substituting these explicit sums back into the main bound yields the final expression:

$$\mathbb{E}[|\partial_{\theta_k} L(\omega)|] \quad \lesssim \quad \sum_{\omega'} |\mathrm{sinc}(\pi(\omega \quad - \quad \omega'))| \left( \|h_{\omega'}| \sum_{r=1}^{R(\omega)} \mathcal{G}^{(r)}(\omega) \sigma^{S^{(r)}-1} \quad + \quad |h_\omega| \sum_{q=1}^{R(\omega')} \mathcal{G}^{(q)}(\omega') \sigma^{S^{(q)}-1} \right), \tag{99}$$

where $\mathcal{G}^{(r)}(\omega)$ represents the explicit geometric and combinatorial prefactor for the $r$-th path of frequency $\omega$:

$$\mathcal{G}^{(r)}(\omega) = \frac{|p(\mathbf{s}^{(r)}, \mathbf{c}^{(r)}, \omega)|}{2^{\sum(s_j^{(r)}+c_j^{(r)})}} \left[ \frac{s_k'^{(r)} 2^{\frac{s_k'^{(r)}-1}{2}} \Gamma(\frac{s'^{(r)}k}{2})}{\sqrt{\pi}} \prod_{j\neq k} \frac{2^{\frac{s_j'^{(r)}}{2}} \Gamma(\frac{s_j'^{(r)}+1}{2})}{\sqrt{\pi}} \right]. \tag{100}$$

This result generalizes spectral bias to the non-integer regime. The expected gradient for a frequency $\omega$ is a weighted sum over the spectrum. Because the sinc function decays as $1/|\omega - \omega'|$, the gradient is primarily determined by three factors: the variational complexities $S^{(r)}$ and $S^{(q)}$ (which provide exponential suppression $\sigma^{S-1}$), the target amplitudes $|h|$, and crucially, the redundancies $R(\omega)$ and $R(\omega')$. Since the inner sums scale linearly with the number of contributing paths, frequencies with high redundancy or those in close proximity to high-redundancy neighbors receive significantly larger gradient updates. Thus, even with spectral leakage, high-complexity frequencies remain exponentially suppressed unless they benefit from the redundancy of a nearby low-complexity, high-amplitude target frequency. □

## E. Extended redundancy analysis results

An extended set of results depicting the spectra of PQCs with a variety of encodings in Figure 8. The encodings are constant Pauli ($\beta_i = 1$), linear Pauli ($\beta_i = i$), binary Pauli ($\beta_i = 2^i$) and exponential Pauli ($\beta_i = 3^i$).

## F. Extended encodings Spectral bias results

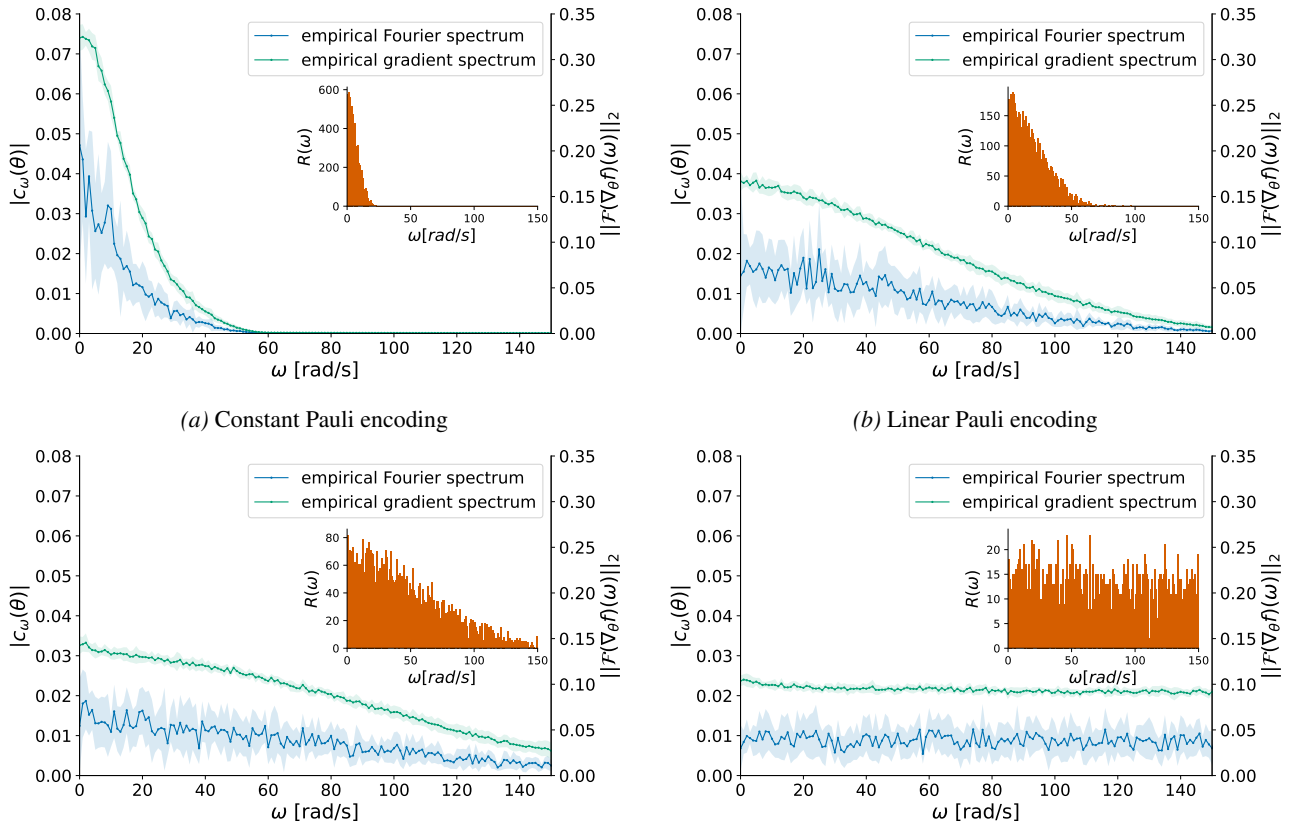
An extended set of results depicting the spectra of PQCs with a variety of encodings in Figure 9. The encodings are constant Pauli ($\beta_i = 1$), linear Pauli ($\beta_i = i$), binary Pauli ($\beta_i = 2^i$) and exponential Pauli ($\beta_i = 3^i$).

## G. Gradient descent spectral bias results

Figure 10 shows the spectral training dynamics reproduced from Section 5.1 using vanilla gradient descent in place of Adam, for constant and ternary Pauli encoding. The frequency learning order is identical to the Adam results, confirming that the spectral bias mechanism is independent of the choice of first-order optimiser.

## H. Robustness of redundant Fourier coefficients under parameter perturbation

Here we detail the proof of theorem 5.2.




*(a)* Constant Pauli encoding

*(b)* Linear Pauli encoding

*(c)* Binary Pauli encoding

*(d)* Exponential Pauli encoding

*Figure 8.* The Fourier spectra of the four Pauli encoding schemes with empirical data taken as the sampled mean over ten models. Depicted are the sampled theoretically accessible frequencies (light red), the mean Fourier coefficient (blue), and the total gradient of trainable parameters at each Fourier coefficient (green).

**Notation and setup.** For a fixed frequency $\omega$. Decompose the (learned) Fourier coefficient with trained parameters $\theta^*$ as a sum of $R(\omega)$ contributing terms.

$$c_\omega(\theta) = \sum_{i=1}^{R(\omega)} a_i(\theta^*),$$

where each $a_i(\theta^*) \in \mathbb{C}$. Let $\delta\theta$ be a random additive perturbation, and define

$$\Delta a_i := a_i(\theta^* + \delta\theta) - a_i(\theta^*), \qquad \Delta C := \sum_{i=1}^{R(\omega)} \Delta a_i.$$

We consider the normalized RMS deviation (root-mean-square fractional deviation)

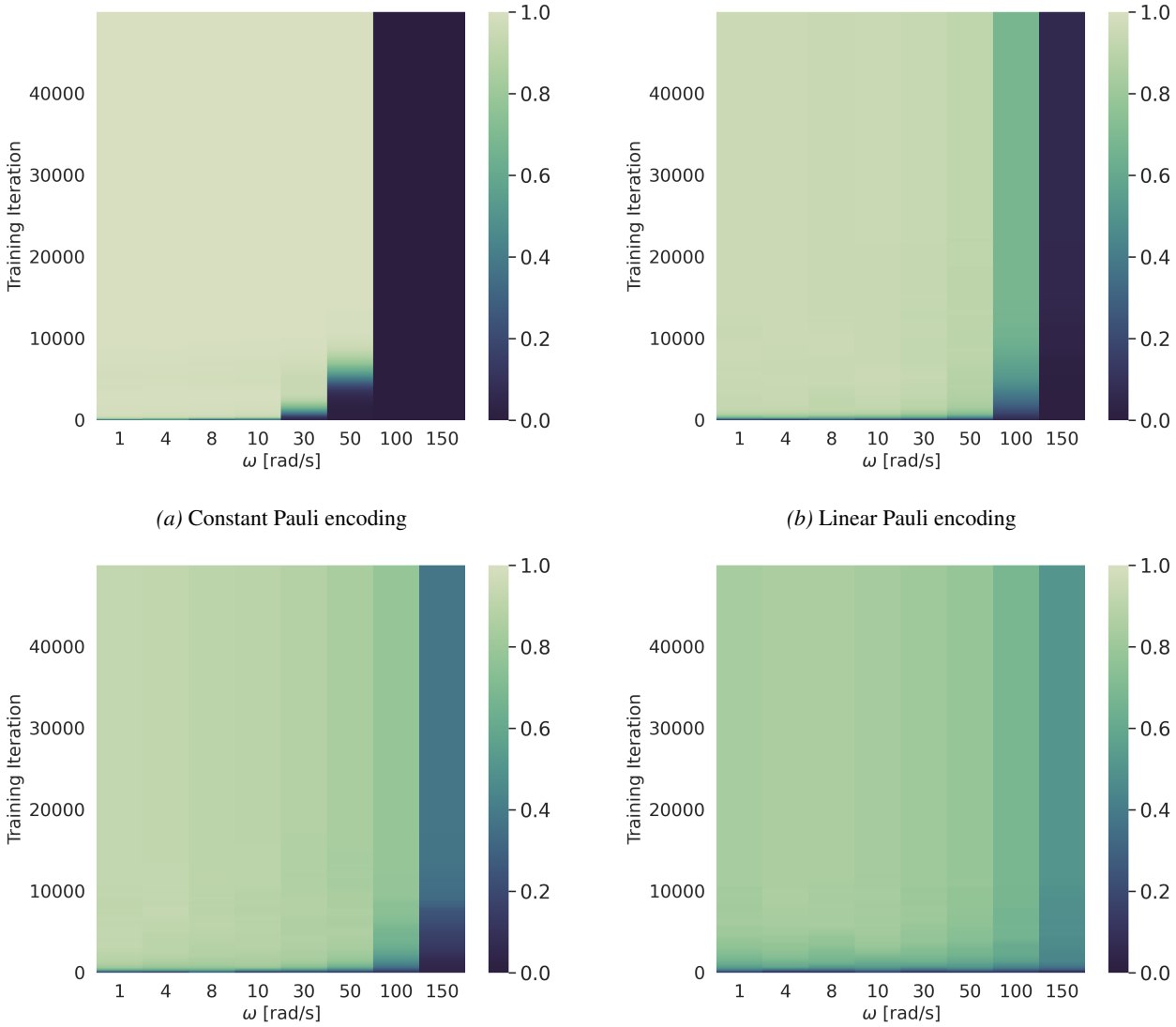
$$\mathcal{R} \;:=\; \frac{\sqrt{\mathbb{E}[\,|\Delta C|^2\,]}}{|c(\theta^*)|},$$

and we assume the coefficient of interest is non-zero $c_\omega(\theta^*)$.

**Assumptions.**

(A1) (*Partial coherence.*) There exist constants $\bar{a} > 0$ and $\kappa \in (0, 1]$ such that

$$|c_\omega(\theta^*)| \geq \kappa\, R(\omega)\, \bar{a}, \qquad \bar{a} := \frac{1}{R(\omega)} \sum_{i=1}^{R(\omega)} |a_i(\theta^*)|.$$




*(a)* Constant Pauli encoding

*(b)* Linear Pauli encoding

*(c)* Binary Pauli encoding

*(d)* Exponential Pauli encoding

*Figure 9.* The rate at which frequencies (x-axis) are learnt during the course of training (y-axis), the colorbar measures the PQC spectrum normalised by the target amplitude at a given frequency ($|\tilde{f}_\omega|/A_i$). Each subplot depicts the training dynamics for a different encoding scheme (a) constant Pauli encoding, (b) linear Pauli encoding and (c) exponential Pauli encoding.

(A2) (*Second-moment and pairwise covariance control.*) For the perturbation distribution,

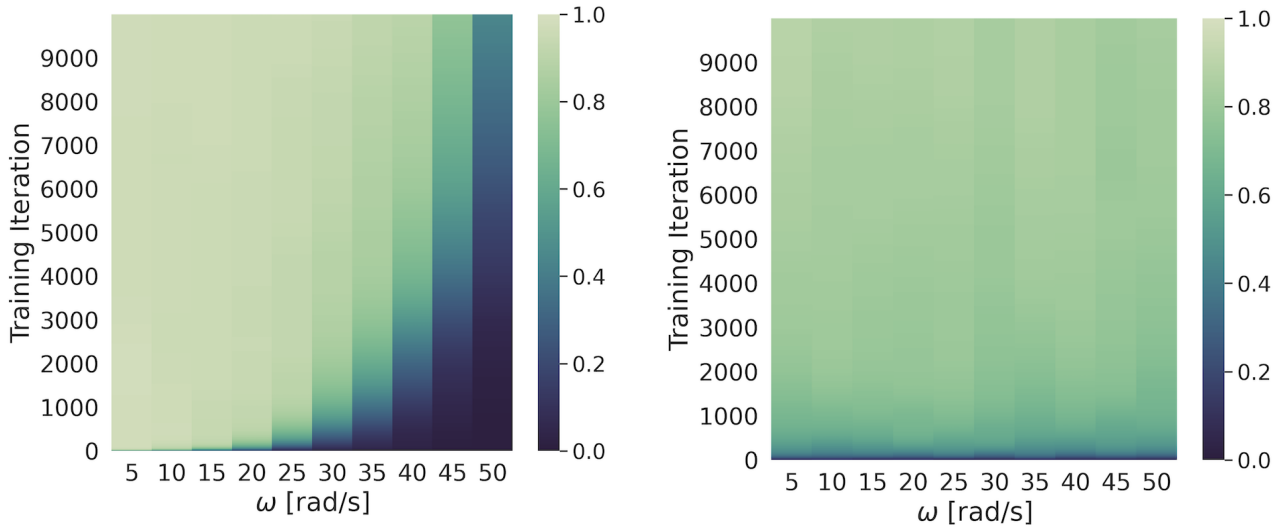
$$\mathbb{E}\big[|\Delta a_i|^2\big] = \sigma_a^2 \quad \text{for all } i,$$

and for $i \neq j$,

$$\big|\mathbb{E}[\Delta a_i \, \overline{\Delta a_j}]\big| \leq \rho \, \sigma_a^2$$

for some $\rho \in [0, 1]$.

(A3) (*Perturbation scale dependence.*) The variance $\sigma_a^2 = \sigma_a^2(\|\delta\theta\|)$ depends only on the perturbation magnitude; for small perturbations one typically has $\sigma_a \propto \|\delta\theta\|$ by linearisation.




*(a)* Constant Pauli encoding

*(b)* Ternary Pauli encoding

*Figure 10.* The rate at which frequencies (x-axis) are learnt during the course of training (y-axis), the colorbar measures the PQC spectrum normalised by the target amplitude at a given frequency ($|\tilde{f}_\omega|/A_i$). Each subplot depicts the training dynamics for a different encoding scheme using gradient descent as the optimizer, (a) constant Pauli encoding and (b) ternary Pauli encoding.

*Proof.* Compute the second moment:

$$\mathbb{E}[|\Delta C|^2] \;=\; \mathbb{E}\Big[\Big|\sum_{i=1}^{R}\Delta a_i\Big|^2\Big] \;=\; \sum_{i=1}^{R}\mathbb{E}[|\Delta a_i|^2] + \sum_{i\neq j}\mathbb{E}[\Delta a_i\overline{\Delta a_j}].$$

Applying the uniform bounds in (A2) the first sum equals $R\sigma_a^2$ and each off-diagonal term has magnitude at most $\rho\sigma_a^2$. Taking absolute values and bounding,

$$\mathbb{E}[|\Delta C|^2] \leq R(\omega)\sigma_a^2 + R(\omega)(R(\omega)-1)\rho\sigma_a^2 = R(\omega)\sigma_a^2\big(1+(R(\omega)-1)\rho\big).$$

Hence

$$\sqrt{\mathbb{E}[|\Delta C|^2]} \leq \sigma_a\sqrt{R(\omega)\big(1+(R(\omega)-1)\rho\big)}.$$

Using (A1) to lower bound the denominator,

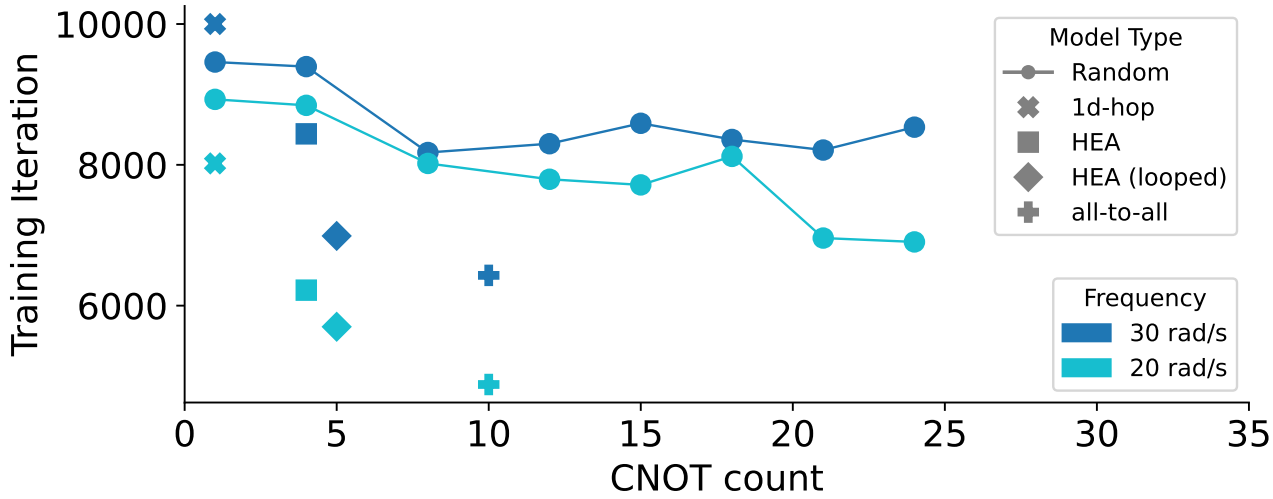
$$\mathcal{R} \leq \frac{\sigma_a\sqrt{R(\omega)(1+(R(\omega)-1)\rho)}}{\kappa R(\omega)\bar{a}} = \frac{\sigma_a}{\kappa\bar{a}}\sqrt{\frac{1+(R(\omega)-1)\rho}{R(\omega)}},$$

which is (21). □

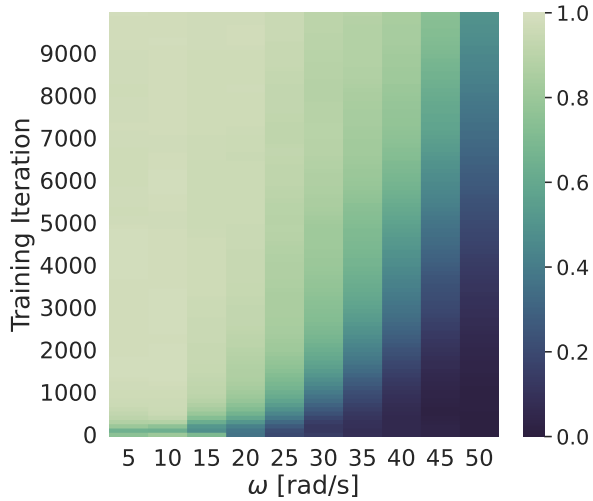
If $\rho \to 0$ (approximate independence of perturbation effects across summands) the bound reduces to the $1/\sqrt{R(\omega)}$ scaling up to multiplicative constants. If $\rho$ is close to 1 (highly correlated perturbation effects), the redundancy offers no benefit.

### H.1. Entanglement

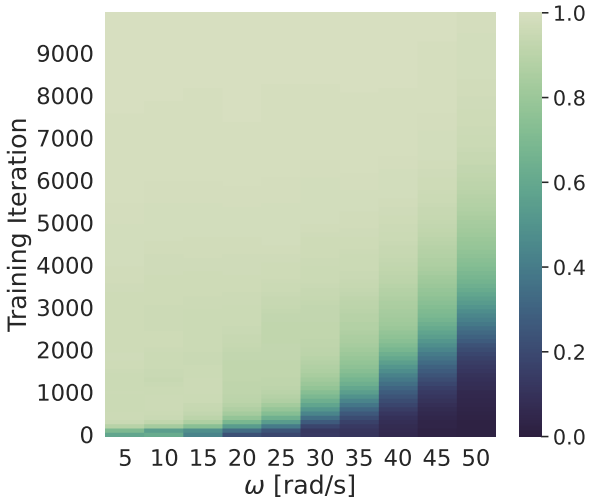
In this section we examine the effect of the entanglement structure on the spectral dynamics, the results of which are shown in Figure 11. Parallel encoding provides access to additional frequency components, but only when the added qubits are entangled, either through the circuit or via multi-qubit measurements. We therefore probe how entanglement affects the resulting spectral bias. We use the constant Pauli encoding scheme and vary the entangling scheme. For the experiment we select four well known entangling structures and compare the number of epochs each takes to converge to a given amplitude associated to a frequency component of the target. The entanglement structures include two extreme cases; the 1d-hop




*(a)*



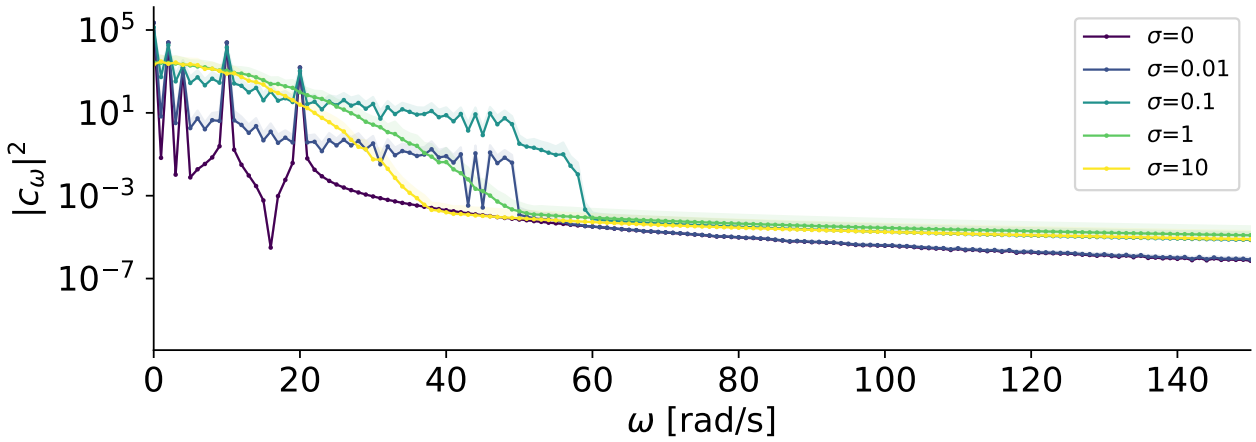

*(b)* 1d-hop entanglement



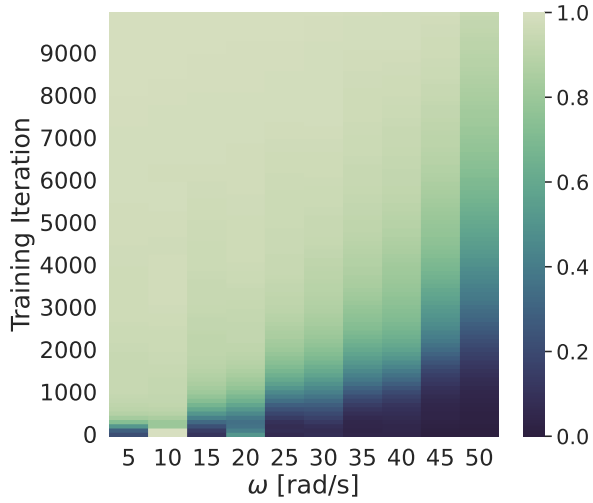

*(c)* all-to-all entanglement

*Figure 11.* Impact of entanglement on spectral training. (a) Mean epochs to convergence vs. number of CNOT gates per layer. We compare random CNOT placement (solid line) against structured entanglement topologies (points). (b, c) Spectral dynamics for the 1D-hop and all-to-all structures, respectively.

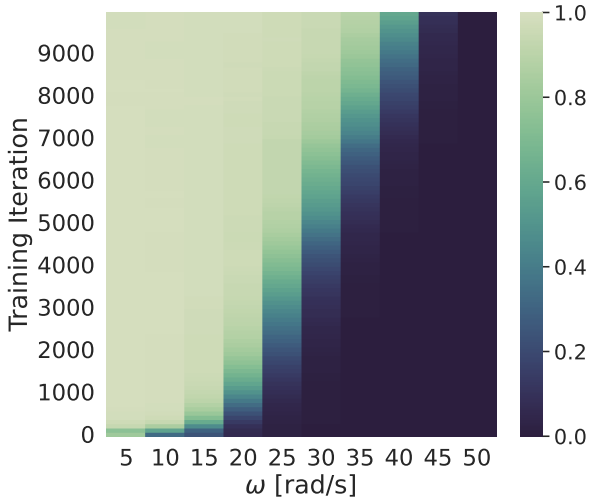
structure connects only two qubits in each reupload layer in a cascading-down fashion and the all-to-all scheme connects all pairwise qubits in each layer. The other entanglement schemes use the hardware-aware (HEA) structure, connecting only neighbouring qubits. One of these schemes does not connect the last qubit to the first qubit, which was the default in our studies so far, whilst the other one does, which we denote as HAE (looped). We also include a baseline of randomly placed CNOT gates. The structured models have their convergence values taken from a mean over ten random initialisations, while the randomised models are means over 20. Figure 11a shows that increasing the number of CNOTs systematically reduces spectral bias, even for randomly placed gates, with structured layouts yielding the strongest effect. We also show the spectral training plot for the two extreme entangling schemes, shown in Figures 11b and 11c respectively. Together with our results on redundancy, this effect suggests a route to designing circuits with task-dependent spectral bias. The reduction induced by additional CNOTs can be understood via a lightcone argument, whereby increased entangling depth enlarges the measurement lightcone and reduces parameter exclusion.




*(a)*




*(b)* $\sigma^2 = 0.01$




*(c)* $\sigma^2 = 10$

*Figure 12.* The effect of initial parameter variance on the spectrum and training. (a) Size of Fourier coefficients based on the variance (around 0) of the trainable PQC parameters. (b) and (c) depict the spectral dynamics for two initialisations with vastly different variance.

### H.2. Initialization

We investigate how the parameter initialization scale affects the rate at which different frequencies are learned in PQCs. We focus on circuits with constant Pauli encoding, initializing trainable parameters $\theta \sim \mathcal{N}(0, \sigma^2)$ and varying $\sigma^2$ across experiments. Before training, we analyze the output spectrum under different initialization scales by computing the Fourier coefficients of the circuit output. Figure 12a shows the squared coefficient magnitudes for $\sigma$ ranging from $0.01$ to $10$. As $\sigma$ increases, $|c_\omega|^2$ initially grows (peaking near $\sigma = 0.1$) before decreasing uniformly across frequencies, indicating that small initializations yield larger initial Fourier coefficients than large ones. This suppression has direct consequences for training. With $\sigma = 0.01$ (Figure 12b), the circuit starts with relatively large coefficients across many frequencies, resulting in modest spectral bias. In contrast, for $\sigma = 10$ (Figure 12c), higher-frequency coefficients are strongly suppressed, leading to smaller gradients, slower learning, and a more pronounced spectral bias. Overall, the initialization scale shapes both expressivity and spectral learning dynamics, with large initializations amplifying spectral bias. As in classical machine learning, where principled initialization schemes are essential, these findings motivate the development of informed initialization strategies for PQCs, as also argued in recent QML literature (Recio-Armengol et al., 2025).