Generalization despite overfitting in quantum machine learning models

Evan Peters[1,2,3] and Maria Schuld[4]

[1]Department of Physics, University of Waterloo, Waterloo, ON, N2L 3G1, Canada

[2]Institute for Quantum Computing, Waterloo, ON, N2L 3G1, Canada

[3]Perimeter Institute for Theoretical Physics, Waterloo, Ontario, N2L 2Y5, Canada

[4]Xanadu, Toronto, ON, M5G 2C8, Canada

The widespread success of deep neural networks has revealed a surprise in classical machine learning: very complex models often generalize well while simultaneously overfitting training data. This phenomenon of benign overfitting has been studied for a variety of classical models with the goal of better understanding the mechanisms behind deep learning. Characterizing the phenomenon in the context of quantum machine learning might similarly improve our understanding of the relationship between overfitting, overparameterization, and generalization. In this work, we provide a characterization of benign overfitting in quantum models. To do this, we derive the behavior of a classical interpolating Fourier features models for regression on noisy signals, and show how a class of quantum models exhibits analogous features, thereby linking the structure of quantum circuits (such as data-encoding and state preparation operations) to overparameterization and overfitting in quantum models. We intuitively explain these features according to the ability of the quantum model to interpolate noisy data with locally "spiky" behavior and provide a concrete demonstration example of benign overfitting.

1 Introduction

A long-standing paradigm in machine learning is the trade-off between the complexity of a model family and the model's ability to generalize: more expressive model classes contain better candidates to fit complex trends in data, but are also prone to overfitting noise [1, 2]. *Interpolation*, defined for our purposes as choosing a model with zero training error, was hence long considered bad practice [3]. The success of deep learning – machine learning in a specific regime of extremely complex model families with vast amounts of tunable parameters – seems to contradict this notion; here, consistent evidence shows that among some interpolating models, more complexity tends *not to harm* the generalisation performance[1], a phenomenon described as "benign overfitting" [4].

In recent years, a surge of theoretical studies have reproduced benign overfitting in simplified settings with the hope of isolating the essential ingredients of the phenomenon

---

Evan Peters: e6peters@uwaterloo.ca

[1]The fact that interpolation does not always harm generalization performance is in fact well-known for models like boosting and nearest neighbours, which are both interpolating models.

[4, 5]. For example, Ref. [6] showed how interpolating linear models in a high complexity regime (more dimensions than datapoints) could generalize just as well as their lower-complexity counterparts on new data, and analyzed the properties of the data that lead to the "absorption" of noise by the interpolating model without harming the model's predictions. Ref. [7] showed that there are model classes of simple functions that change quickly in the vicinity of the noisy training data, but recover a smooth trend elsewhere in data space (see Figure 1). Such functions have also been used to train nearest neighbor models that perfectly overfit training data while generalizing well, thereby directly linking "spiking models" to benign overfitting [8]. Recent works try to recover the basic mechanism of such spiking models using the language of Fourier analysis [9, 10, 11].

In parallel to these exciting developments in the theory of deep learning, quantum computing researchers have proposed families of parametrised quantum algorithms as model classes for machine learning (e.g. Ref. [12]). These quantum models can be optimised similarly to neural networks [13, 14] and have interesting parallels to kernel methods [15, 16] and generative models [17, 18]. Although researchers have taken some first steps to study the expressivity [19, 20, 21, 22], trainability [23, 24] and generalisation [25, 26, 27, 28] of quantum models, we still know relatively little about their behaviour. In particular, the interplay of overparametrisation, interpolation, and generalisation that seems so important for deep learning is yet largely unexplored.

In this paper we develop a simplified framework in which questions of overfitting in quantum machine learning can be investigated. Essentially, we exploit the observation that quantum models can often be described in terms of Fourier series where well-defined components of the quantum circuit influence the selection of Fourier modes and their respective Fourier coefficients [29, 30, 31]. We link this description to the analysis of spiking models and benign overfitting by building on prior works analyzing these phenomena using Fourier methods. In this approach, the complexity of a model is related to the number of Fourier modes that its Fourier series representation consists of, and overparametrised model classes have more modes than needed to interpolate the training data (i.e., to have zero training error). After deriving the generalization error for such model classes these "superfluous" modes lead to spiking models, which have large oscillations around the training data while keeping a smooth trend everywhere else. However, large numbers of modes can also harm the recovery of an underlying signal, and we therefore balance this trade-off to produce an explicit example of benign overfitting in a quantum machine learning model.

The mathematical link described above allows us to probe the impact of important design choices for a simplified class of quantum models on this trade-off. For example, we find why a measure of redundancy in the spectrum of the Hamiltonian that defines standard data encoding strategies strongly influences this balance; in fact to an extent that is difficult to counterbalance by other design choices of the circuit.

The remainder of the paper proceeds as follows. We will first review the classical Fourier framework for the study of interpolating models and develop explicit formulae for the error in these models to produce a basic example of benign overfitting (Sec. 2). We will then construct a quantum model with analogous components to the classical model, and demonstrate how each of these components is related to the structure of the corresponding quantum circuit and measurement (Sec. 3). We then analyze specific cases that give rise to "spikiness" and benign overfitting in these quantum models (Sec. 3.2).
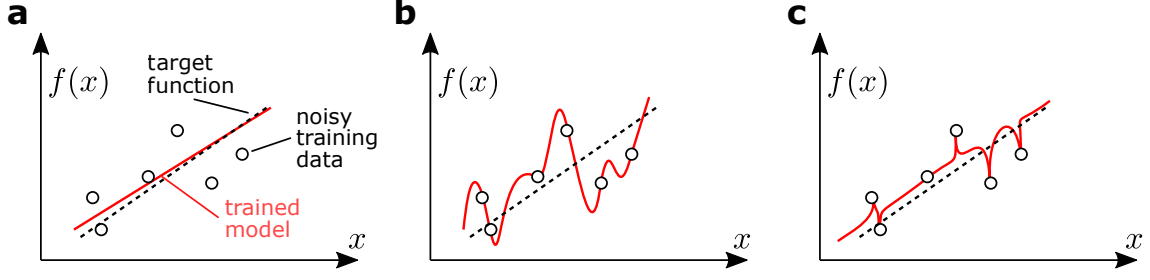
Figure 1: Intuition behind the phenomenon of benign overfitting with spiking models. **a** Training a model typically involves a trade-off between fitting noisy training data well and recovering an underlying target function **b** Traditional learning theory associates interpolating models (reaching zero training error by fitting every data point) with low generalization capability (or high test error) by failing to recover a simple target function. **c** "spiking models" that change quickly in the vicinity of training data but otherwise exhibit simple behavior can explain how both kinds of errors may be kept low.

2 Interpolating models in the Fourier framework

In this section we will provide the essential tools to probe the phenomenon of over-parametrized models that exhibit the spiking behaviour from Figure 1 using the language of Fourier series. We will review and formalize the problem setting and several examples from Refs. [9, 10, 11], before extending their framework by incorporating standard results from linear regression to derive closed-form error behavior and examples of benign overfitting.

2.1 Setting up the learning problem

We are interested in functions g defined on a finite interval that may be written in terms of a linear combination of Fourier basis functions or *modes* $e^{i2\pi kx}$ (each describing a complex sinusoid with integer-valued frequency k) weighted by their corresponding Fourier coefficients $\hat{g}_k$:

$$g(x) = \sum_{k=-\infty}^{\infty} \hat{g}_k e^{i2\pi kx}. \tag{1}$$

We restrict our attention to well-behaved functions that are sufficiently smooth and continuous to be expressed in this form. We will now set up a simple learning problem whose basic components – the model and target function to be learned – can be expressed as Fourier series with only few non-zero coefficients, and define concepts such as over-parametrization and interpolation.

Consider a machine learning problem in which data is generated by a target function of the form

$$g(x) = \sum_{k\in\Omega_{n_0}} \hat{g}_k e^{i2\pi kx} \tag{2}$$

which only contains frequencies in the discrete, integer-valued spectrum

$$\Omega_{n_0} = \left\{ -\frac{n_0-1}{2}, -\frac{n_0-1}{2}+1, \ldots, 0, \ldots, \frac{n_0-1}{2} \right\} \tag{3}$$

for some odd integer n_0. We call functions of this form *band-limited* with bandwidth $n_0/2$ (that is, $|k| < n_0/2$ for all frequencies $k \in \Omega_{n_0}$). The bandwidth limits the complexity of a function, and will therefore be important for exploring scenarios where a complex model is used to overfit a less complex target function. We are provided with n noisy samples $y_j = g(x_j) + \epsilon$ of the target function g evaluated at points $x_j = j/n$ spaced uniformly

on interval $[0, 1]$ (we will assume input data has been rescaled to $[0, 1]$ without loss of generality), where $n > n_0$ and we require n to be odd for simplicity.

The model class we consider likewise contains band-limited Fourier series, and since we are interested in interpolating models, we always assume that they have enough modes to interpolate the noisy training data, namely:

$$f(x) = \sum_{k \in \Omega_d} \alpha_k \sqrt{\nu_k} e^{i2\pi kx}, \tag{4}$$

$$\text{such that} \qquad f(x_j) = y_j \text{ for all } j \in [n].$$

Similarly to Eq. (3) we define the spectrum

$$\Omega_d = \left\{ -\frac{d-1}{2}, -\frac{d-1}{2} + 1, \ldots, 0, \ldots, \frac{d-1}{2} \right\}. \tag{5}$$

Following Ref. [9], this model class has two components: The set of weighted Fourier basis functions $\sqrt{\nu_k} e^{i2\pi kx}$ describe a feature map applied to x for some set of feature weights $\nu_k \in \mathbb{R}^+$, while the trainable weights $\alpha_k \in \mathbb{C}$ are optimized to ensure that f interpolates the data. The theory of trigonometric polynomial interpolation [32] ensures that f can always interpolate the training data for some choice of trainable weights α_k under these conditions. In the following, we will therefore usually consider the α_k as being determined by the data and interpolation condition, while the ν_k serve as our "turning knobs" to create different settings in which to study spiking properties and benign overfitting. We call the model class described in Eq. (4) *overparameterized* when the degree of the Fourier series is much larger than the degree of the target function, $d \gg n_0$, in which case the model has many more frequencies available to perform fitting than there are present in the underlying signal to fit.

Note that one can rewrite the Fourier series of Eq. (4) in a linear form

$$f(x) = \langle \boldsymbol{\alpha}, \phi(x) \rangle \tag{6}$$

where

$$\boldsymbol{\alpha} = (\alpha_{-(d-1)/2}, \ldots, \alpha_k, \ldots, \alpha_{(d-1)/2}) \in \mathbb{C}^d, \tag{7}$$

$$\phi(x) = \left(\sqrt{\nu_{-(d-1)/2}} e^{i\pi(d-1)x} \ldots, \sqrt{\nu_k} e^{i2\pi kx}, \ldots, \sqrt{\nu_{(d-1)/2}} e^{i\pi(d-1)x} \right) \in \mathbb{C}^d. \tag{8}$$

From this perspective, optimizing f amounts to learning trainable weights $\boldsymbol{\alpha}$ by performing regression on observations $\phi(x)$ sampled from a *random Fourier features* model [33] for which ν_k and $\phi(x)_k$ are precisely the eigenvalues and eigenfunctions of a shift-invariant kernel [34].

To complete the problem setup, we have to impose one more constraint. Consider that $\exp(i2\pi kx_j)$ with frequency $k < n/2$ and uniformly spaced points x_j is equal to $\exp(i2\pi k'x_j)$ for any choice of *alias frequency* $k' = k \mod n$. The presence of these aliases means that the model class described in Eq. (6) contains many interpolating solutions in the overparameterized regime. Motivated by prior work exploring benign overfitting for linear features [6], Fourier features [9, 35], and other nonlinear features [36, 37], we will study the minimum-ℓ_2 norm interpolating solution,

$$\boldsymbol{\alpha}^{opt} = \underset{\boldsymbol{\alpha}: f(x_j) = y_j \, \forall j}{\arg \min} \|\boldsymbol{\alpha}\|_2. \tag{9}$$

Minimizing the ℓ_2 norm is a typical choice for imposing a penalty on complex functions (regularization) in the underparameterized regime, though we will see that this intuition

does not carry over to the overparameterized regime. The remainder of this section will explore how this learning problem results in a trade-off in interpolating Fourier models: Overparameterization introduces alias frequencies that increase the error in fitting simple target functions but can also reduce error by absorbing noise into high-frequency modes with spiking behavior.

2.2 Two extreme cases to understand generalization

To better understand the trade-off that overparametrization – or in our case, a much larger number of Fourier modes than needed to interpolate the data – introduces between fitting noise and generalization error, we revisit two extreme cases explored in Ref. [9], involving a pure-noise signal and a noiseless signal.

Case 1: Noise only

The first case demonstrates how alias modes can help to fit noise without disturbing the (here trivial) signal. We set $g = 0$ and consider n observations $y_j = \epsilon_j$ of zero-mean noise with known variance $\mathbb{E}[\epsilon^2] = \sigma^2$. After making n uniformly spaced observations, we compute the discrete Fourier transform of the observations as the sequence of values $\hat{\epsilon}_j$ satisfying

$$\hat{\epsilon}_j = \frac{1}{n} \sum_{k \in \Omega_n} \epsilon_k e^{-i2\pi jk/n}, \tag{10}$$

which characterizes the frequency content of the noisy signal that can be captured and learned from using only n evenly spaced samples. Suppose that the degree of the model (controlling the model complexity) is given by $d = n(m+1)$ for some even integer m and that $\nu_k = 1$ for every mode, so that there are exactly m equally-weighted aliases for each frequency in the spectrum of the Fourier series for g. Then the optimal (i.e., the minimum ℓ_2-norm, interpolating) trainable weight vector $\boldsymbol{\alpha}^{opt}$ has entries

$$\alpha_{k+n\ell}^{opt} = \frac{\hat{\epsilon}_k}{n(m+1)} \tag{11}$$

for $\ell = -m/2, \ldots, m/2$, with all other entries being zero (see Appendix A.2). From Eq. (11), the minimum-$\|\boldsymbol{\alpha}\|_2$ solution distributes noise Fourier coefficients $\hat{\epsilon}_k$ evenly into many alias frequencies $k + n\ell$, while enforcing that the sum of trainable weights $\alpha_{k+n\ell}$ for all of these aliases is $\hat{\epsilon}_k$ to guarantee interpolation. As shown in Figure 2, the higher-frequency aliases suppress the optimal model $f^{opt}(x) = \langle \boldsymbol{\alpha}^{opt}, \phi(x) \rangle$ to near-zero at every point away from the interpolated points, resulting in a test error of $\mathcal{O}(\sigma^2/m)$ that decreases monotonically with the complexity of the model. As m increases, the optimal model f^{opt} remains close to the true signal $y = 0$ while becoming "spiky" near the noisy samples. By conforming to the true signal everywhere except in the vicinity of noise, this behavior embodies the mechanism of how overparameterized models can absorb noise into high frequency modes. In this case the generalization error, measuring how close the model is to the target function on average, *decreases* with increasing complexity of the model class.

Case 2: Signal only

While the above case shows how overparametrization can help to absorb noise to reduce error without harming the signal, the second case will illustrate how alias frequencies in the overparameterized model can harm the model's ability to learn the target function.
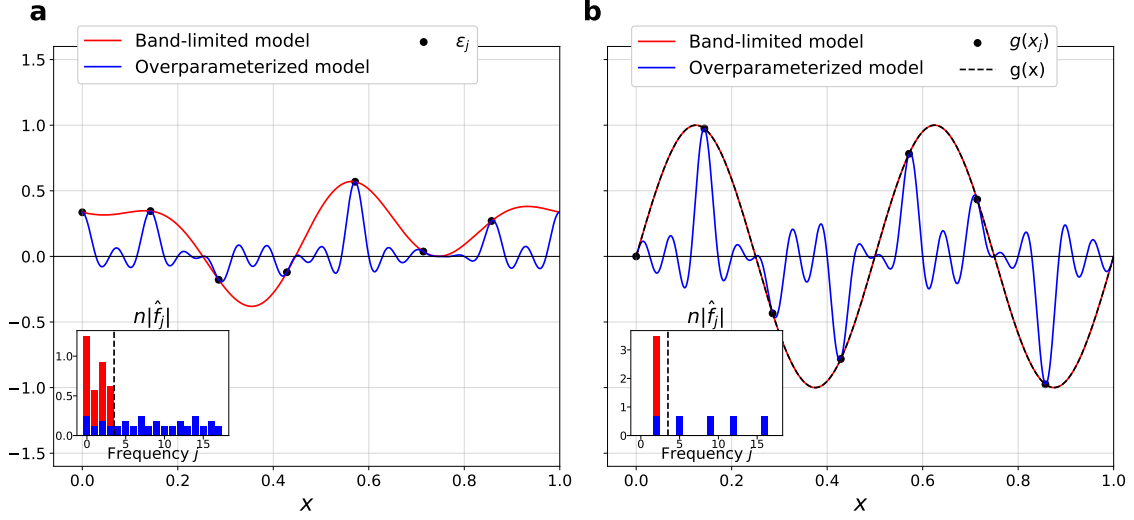
Figure 2: Comparison of overparametrized and simple models that interpolate data from different target functions. **a** *Noise only:* The overparameterized model ($n = 7$, $d = 35$, blue) is more effective at recovering the target function $g(x) = 0$ when provided with noisy data, while a band-limited model (red) cannot do so effectively. Inset: The distribution of trainable weights for the minimum $\|\boldsymbol{\alpha}\|_2$ model are distributed among many aliases suppressing the error of the optimized model (proportional to $\sum_j |\hat{f}_j|^2$) in this case. **b** *Signal only:* The opposite occurs in the case of noiseless input, for which a band-limited model ($n_0 = 5$) perfectly recovers $g(x) = \sin(2\pi(2x))$ while the overparameterized model fails to capture the behavior of the input signal. Inset: Minimizing the ℓ_2 norm of $\boldsymbol{\alpha}$ distributes Fourier coefficients $\hat{g}_2$ among aliases $\hat{f}_{2+\ell n}$ (and $\hat{g}_{-2}$ among aliases $\hat{f}_{-2+\ell n}$); this bleeding of the signal g into higher frequencies results in higher error in the overparameterized model [9, 11].

To demonstrate this, we now consider a noiseless, single-mode signal $g(x) := \hat{g}_p e^{i2\pi px}$ of frequency $p \le n_0/2$. The data is hence of the form

$$y_j = g(x_j) := \hat{g}_p e^{i2\pi px_j}. \tag{12}$$

Once again we choose $d = n(m+1)$ and for simplicity we assume an unweighted model, $\nu_k = 1$ for $k \in \Omega_d$. By orthonormality of Fourier basis functions, the interpolation condition requires that only the modes of the model f with integer multiples of the frequency p are retained. The interpolation constraint can then be rewritten as

$$\sum_{\ell=-m/2}^{m/2} \alpha_{p+n\ell} = \hat{g}_p. \tag{13}$$

The choice of trainable weights α_k that satisfy Eq. (13) while minimizing ℓ_2-norm is

$$\alpha_{p+n\ell}^{opt} = \frac{\hat{g}_p}{m+1} \tag{14}$$

for $k = p + n\ell$ and $\alpha_k = 0$ otherwise (see Appendix A.2). Eq. (13) distributes the Fourier coefficient $\hat{g}_p$ among the trainable weights $\alpha_{p+n\ell}$ corresponding to frequencies $p+n\ell$. Therefore, minimizing $\|\boldsymbol{\alpha}\|_2$ in this case "bleeds" the target function into higher frequency aliases and results in the opposite effect compared to fitting a noisy signal (see Fig. 2b): The generalization error of the overparameterized model now *increases* with the number of aliases m and the complexity of the model harms its ability to fit a noiseless target function.

In order to recover a trade-off in generalization error for more general cases, we will need to consider more interesting distributions of feature weights ν_k (instead of $\nu_k = 1$) that provide finer control over fitting the target function with low-frequency modes while spiking in the vicinity of noise with high-frequency aliases.

2.3 Generalization trade-offs and benign overfitting

The opposing effects of higher-frequency modes in overparameterized models in the cases discussed above hint at a trade-off in model performance that depends on the underlying signal and the feature weights of the Fourier feature map. Returning to the more general case of input samples $y_j = g(x_j) + \epsilon_j$, in Appendix A we show that the task of fitting uniformly spaced samples using weighted Fourier features may be transformed into a linear regression problem, thereby generalizing the results of [9] to derive the following general solution to the minimum-ℓ_2 interpolating problem of Eq. (9):

$$\alpha_{k+\ell n}^{opt} = \hat{y}_k \frac{\sqrt{\nu_{k+\ell n}}}{\sum_{j \in S(k)} \nu_j},\tag{15}$$

where $\hat{y}_k$ is the discrete Fourier transform of y_k and $k \in \Omega_{n_0}$, where

$$S(k) = \{\ell : k + n\ell \in \Omega_d\}\tag{16}$$

denotes the set of alias frequencies of k appearing in the overparameterized model with spectrum Ω_d. The optimal model is then expressed as

$$f^{opt}(x) = \sum_{k \in \Omega_{n_0}} \tilde{y}_k \sum_{\ell \in S(k)} e^{i2\pi \ell x} \frac{\nu_\ell}{\sum_{j \in S(k)} \nu_j}.\tag{17}$$

Recalling that our model f is trained on n noisy samples $(y_0, \ldots, y_{n-1})$ of the target function g, we are interested in the squared error of the model f averaged over (noisy) samples over the input domain,

$$L(f) := \mathbb{E}_{x,\mathbf{y}}(f(x) - g(x))^2.\tag{18}$$

and we call L the *generalization error* of f, as it captures the behavior of f with respect to g over the entire input domain $x \in [0,1]$ instead of just the uniformly spaced training points x_j. In Appendix A we derive

$$L(f^{opt}) = \underbrace{\frac{\sigma^2}{n} \sum_{k \in \Omega_n} \frac{\sum_{\ell \in S(k)} \nu_\ell^2}{\left(\sum_{j \in S(k)} \nu_j\right)^2}}_{\text{VAR}} + \underbrace{\sum_{k \in \Omega_{n_0}} |\hat{g}_k|^2 \left[\left(\frac{\sum_{\ell \in S(k) \backslash k} \nu_\ell}{\sum_{j \in S(k)} \nu_j}\right)^2 + \frac{\sum_{\ell \in S(k) \backslash k} \nu_\ell^2}{\left(\sum_{j \in S(k)} \nu_j\right)^2} \right]}_{\text{BIAS}^2}\tag{19}$$

We use this generalization error now to explore two interesting behaviors of the interpolating model in our setting: The tradeoff between noise absorption and signal recovery exemplified by the cases in Sec. 2.2, and the ability of an overparameterized Fourier features model to *benignly* overfit the training data.

The first behavior involves a trade-off in the *generalization error* $L(f^{opt})$ between absorbing noise (reducing VAR) and capturing the target function signal (reducing BIAS^2) that recovers and generalizes the behavior of the two cases in Sec. 2.2. This trade-off is controlled by three components: The noise variance σ^2, the input signal Fourier coefficients $\hat{g}_k$, and the distribution of feature weights $\sqrt{\nu_k}$. As described in the two cases above, when

$\sigma^2 \to 0$ (signal only) the variance term VAR vanishes and the model is biased for any choice of nonzero ν_k where $k > n$. Conversely, when $\hat{g} \to 0$ (noise only) the bias term BIAS^2 vanishes, and the variance term is minimized by choosing uniform ν_k for all $k \in \Omega_d$.

The second interesting behavior occurs when the generalization error of the overparameterized model decreases at a nearly optimal rate as the number of samples n increases, known as *benign overfitting*. Prior work on benign overfitting in linear regression studied scenarios where the distribution of input data varied with the dimensionality of data and size of the training set in such a way that the excess generalization error of the overparameterized model (compared to a simple model) vanished [6]. However, since the dimensionality of the input data for our model is fixed, we instead consider sequences of feature weights that vary with respect to the problem parameters (n_0, n, d) in a way that results in BIAS^2 and VAR vanishing as $n \to \infty$. In this case, by fitting an increasing number of samples n using such a sequence of feature weights, the overparameterized model both perfectly fits the training data and generalizes well for unseen test points, and therefore exhibits benign overfitting.

These behaviors are exemplified by a straightforward choice of feature weights that incorporate some prior knowledge of the spectrum Ω_{n_0} available to the target function g. For all $k \in \Omega_{n_0}$, let $\nu_k = c/n_0$ for some positive c and normalize the feature weights so that $\sum_{k \in \Omega_k} \nu_k = 1$. We show in Appendix A.3.1 that the error terms of $L(f^{opt})$ scale as

$$\text{VAR} = \mathcal{O}\left(\frac{1}{n} + \frac{n}{d}\right), \tag{20}$$

$$\text{BIAS}^2 = \mathcal{O}\left(\frac{1}{n^2}\right). \tag{21}$$

Thus, as long as the dimension of the overparameterized Fourier features model grows strictly faster than n (i.e., $d = \omega(n)$), the model exhibits benign overfitting. In Appendix A.3.2 we demonstrate how this simple example actually captures the mechanisms of benign overfitting for much more general choices of feature weights. Fig. 3 summarizes this behavior and provides an example of the bias-variance tradeoff that occurs for overparameterized models. In particular, Fig. 3a exemplifies the setting in which benign overfitting occurs, wherein the feature weights of the Fourier features model are strongly concentrated over frequencies in Ω_{n_0} but extend over a large range of alias frequencies for each $k \in \Omega_{n_0}$.

The generalization behavior described here is fundamentally different from many generalization guarantees typically found in statistical learning theory. While prior work has derived guarantees for the generalization of quantum models by constructing bounds on the complexity of the model class [25], Eqs. 20-21 demonstrate that generalization may occur as the complexity (i.e. dimension) of a model grows arbitrarily large.

So far, we have reviewed the Fourier perspective on fitting periodic functions in a classical setting and extended the analysis to characterize benign overfitting. However, if we can link the basic components of quantum models to the terms appearing in the error of Eq. (19), then we will be able to study a similar trade-off in the error of overparameterized quantum models and the conditions necessary for benign overfitting. The remainder of this work is devoted to showing that analogous mechanisms exist in certain quantum machine learning models, and to studying the choices of feature weights for which quantum models can exhibit tradeoffs in generalization error and benign overfitting.
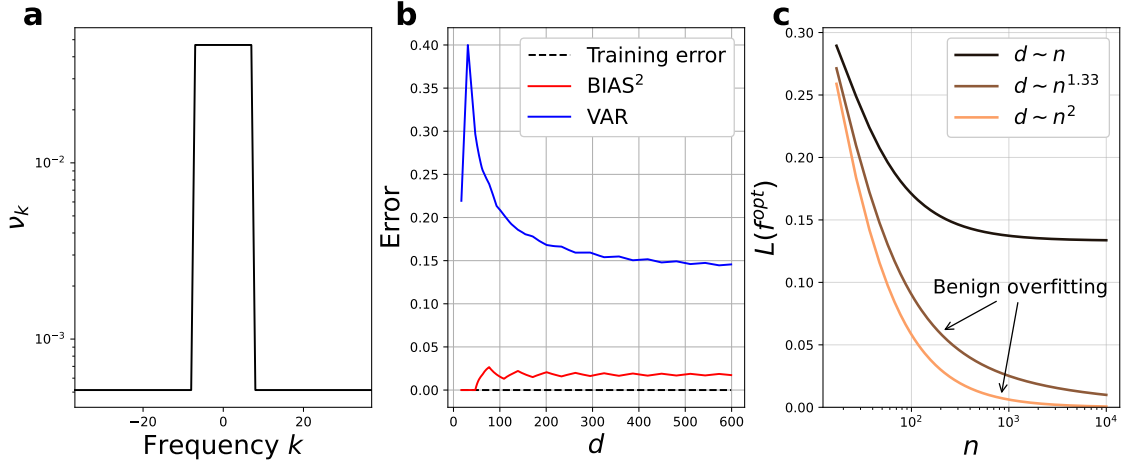
Figure 3: The feature weights ν_k may be engineered to exhibit a bias-variance tradeoff and benign overfitting. **a** A demonstration of feature weights ν_k which account for prior knowledge of the target function g, with large ν_k for $k \in \Omega_{n_0}$ and small ν_k for $k \in \Omega_d \backslash \Omega_{n_0}$. **b** After sampling a bandlimited target function ($n_0 = 15$) at a fixed number of uniformly spaced points ($n = 31$), the generalization error of the minimum ℓ_2-norm model f^{opt} consists of a tradeoff between decreasing VAR and increasing BIAS^2 for larger d. Note the peaking behavior for VAR at $d/n = 1$. **c** This choice of feature weights ν_k benignly overfits the input signal such that $\lim_{n,d\to\infty} L(f^{opt}) = 0$ for any scaling of $d = n^\alpha$ with $\alpha > 1$, and fails to generalize on the interval $[0,1]$ when $\alpha \le 1$.

3 Benign overfitting in single-layer quantum models

In the previous section we have seen that the feature weights $\sqrt{\nu_k}$ balance the trade-off between absorbing the noise and hurting the signal of overparametrized models. To understand how different design choices of quantum models impact this balance, we need to link their mathematical structure to the model class defined in Eq. (4), and in particular to the feature weights, which is what we do now.

The type of quantum models we consider here are simplified versions of parametrized quantum classifiers (also known as *quantum neural networks*) that have been heavily investigated in recent years [13, 38, 39]. They are represented by quantum circuits that consist of two steps: first, we encode a datapoint $x \in [0,1]$ into the state of the quantum system by applying a d-dimensional unitary operator $V(x)$, and then we measure the expectation value of some d-dimensional (Hermitian) observable M. This gives rise to a general class of quantum models of the form

$$f(x) = \langle 0|V^\dagger(x) M V(x)|0\rangle. \tag{22}$$

To simplify the analysis, we will consider a quantum circuit $V(x)$ that consists of a data-independent unitary U and a diagonal data-encoding unitary generated by a d-dimensional Hermitian operator H,

$$S(x) = \exp(i2\pi \cdot \mathrm{diag}(\lambda_0, \ldots, \lambda_{d-1})x) = \exp(i2\pi H x), \tag{23}$$

which includes a large class of quantum models commonly studied but excludes schemes involving data re-uploading [40, 41]. Defining $U|0\rangle = |\Gamma\rangle = \sum_{j=1}^d \gamma_j|j\rangle$, the output of this quantum model becomes

$$f(x) = \langle\Gamma|S^\dagger(x) M S(x)|\Gamma\rangle, \tag{24}$$

where $|\Gamma\rangle$ can be treated as an arbitrary input quantum state. We call the corresponding quantum circuit for this model single-layer in the sense that it contains a single diagonal data-encoding layer in which all data-dependent circuit operations could theoretically be executed simultaneously (though in general the operation U and measurement M may require significant depth to implement).

Applying insights from Refs. [29, 30], quantum models of this form can be expressed in terms of a Fourier series

$$f(x) = \sum_{k \in \Omega} e^{i2\pi kx} \sum_{\ell,m \in R(k)} \gamma_\ell \gamma_m^* M_{m\ell}, \tag{25}$$

where the *spectrum* Ω as well as the *partitions* $R(k)$ depend on the eigenspectrum $\lambda(H)$ of the data-encoding Hamiltonian H:

$$\Omega = \{(\lambda_\ell - \lambda_m) : \lambda_\ell, \lambda_m \in \lambda(H)\} \tag{26}$$
$$R(k) = \{(\ell,m) : \lambda_\ell - \lambda_m = k;\ \lambda_\ell, \lambda_m \in \lambda(H)\}. \tag{27}$$

Comparing Eq. (25) to Eq. (4) we see that that the quantum model may be expressed as a linear combination of weighted Fourier modes, but it is not yet clear how the input state γ_j and the trainable observable M of the quantum model correspond to feature weights ν_k for each Fourier mode. To reveal this correspondence, we will need to first find the minimum-norm interpolating observable that solves the optimization problem

$$M^{opt} = \underset{M=M^\dagger : f(x_j)=y_j\ \forall j}{\arg\min} \|M\|_F, \tag{28}$$

where $\|M\|_F = \sqrt{\mathrm{Tr}\left(M^\dagger M\right)}$ denotes the Frobenius norm of M. Solving Eq. (28) is analogous to the minimization the ℓ_2 norm of $\boldsymbol{\alpha}$ in the classical optimization problem of Eq. (9), and serves a role similar to regularization commonly applied to quantum models by introducing a penalty term proportional to $\|M\|_F^2$ [26, 42, 43]. In Appendix B we prove that subject to the condition that $\gamma_i > 0$, the minimum-$\|\cdot\|_F$ interpolating observable that solves Eq. (28) is given as

$$(M^{opt})_{m\ell} = \hat{y}_k \frac{\gamma_\ell^* \gamma_m}{\sum_{i,j \in R(S_k)} |\gamma_i|^2 |\gamma_j|^2}, \tag{29}$$

for all $\ell, m \in R(S_k)$, and the corresponding optimal quantum model is

$$f^{opt}(x) = \sum_{k \in \Omega_{n_0}} \hat{y}_k \sum_{\ell \in S(k)} e^{i2\pi \ell x} \frac{\nu_\ell^{opt}}{\sum_{j \in S(k)} \nu_j^{opt}}, \tag{30}$$

where $S(k)$ denotes the set of aliases of k appearing in Ω from Eq. (16). By comparison to the optimal classical model of Eq. (17) we have identified the feature weights of the optimized quantum model as

$$\nu_k^{opt} := \sum_{i,j \in R(k)} |\gamma_i|^2 |\gamma_j|^2. \tag{31}$$

Interestingly, while there was initially no clear way to separate the building blocks of the quantum model in Eq. (25) into trainable weights α_k and feature weights ν_k, this separation has now appeared after solving for the optimal observable M^{opt}. Furthermore, the optimal quantum model depends on $|\gamma_i|$ and is independent of phases associated with

amplitudes γ_i (an effect that stems from using only a single data-encoding layer $S(x)$ in the quantum model).

From Eq. (31) it is clear that the partitions $R(k)$ of Eq. (27) arising from the choice of data-encoding unitary $S(x)$ have a strong relationship with the feature weights ν_k of the quantum model. We will now consider a simplified quantum model to highlight this relationship, thereby identifying a tradeoff between noise absorption and target signal recovery and the possibility of observing benign overfitting in quantum models.

3.1 Simplified quantum model

To explicitly highlight the role of $R(k)$ in controlling the feature weights ν_k^{opt} of the optimized quantum model, we will now simplify the model by using an equal superposition input state $|\Gamma\rangle = \frac{1}{\sqrt{d}} \sum_{j=0}^{d-1} |j\rangle$ and by restricting the set of observables considered during optimization. If we fix every entry of the observable with respect to elements in a partition $R(k)$ to be proportional to some complex constant $M(k)$:

$$M_{m\ell} = \frac{M(k)}{\sqrt{|R(k)|}} \qquad \text{for all } \ell, m \in R(k), \tag{32}$$

then we can simplify the quantum model of Eq. (25) to

$$f(x) = \sum_{k \in \Omega} M(k) \frac{\sqrt{|R(k)|}}{d} e^{i2\pi kx}. \tag{33}$$

Comparing Eq. (33) to Eq. (4) we identify a direct correspondence between the trainable weights α_k in the classical model with $M(k)$, as well as a correspondence between the feature weights $\sqrt{\nu_k}$ and the the *degeneracy* $|R(k)|$ of the quantum model. Making the substitutions $\alpha_k \to M(k)$, $\sqrt{\nu_k} \to \sqrt{|R(k)|}/d$, one can verify that $\sum_{k \in \Omega} |M(k)|^2 = \|M\|_F^2$ for this restricted choice of M and so the solution to the optimization of Eq. (28) is essentially the same as that of the classical problem in Eq. (15).

The crucial property of the simplified model is that the degeneracy $|R(k)|$ – and hence the combinatorial structure introduced by the data encoding Hamiltonian's eigenvalues – completely controls the trade-off in the generalization error (Eq. (19)). We can hence study different types of partitions $R(k)$ to show a direct effect of the data-encoding unitary $S(x)$ on the fitting and generalization error behaviors for this simplified, overparameterized quantum model.

To study these behaviors we will now consider specific families of H which we call *encoding strategies* since the choice of H completely determines how the data is encoded into the quantum model. While $R(k)$ and Ω may be computed for an arbitrary $S(x)$ using brute-force combinatorics, some encoding strategies lead to particularly simple solutions. We have derived a few such examples of simple degeneracies and spectra for different encoding strategies in Appendix C and present the results in Table 1. These choices highlight the extreme variation in Ω resulting from minor changes to $S(x)$, for example $|\Omega| \propto n_q$ for the "Hamming" encoding strategy compared to $|\Omega| \propto 3^{n_q}$ for the "Ternary" encoding strategy. These examples also highlight the limitations in constructing Hamiltonians with specific properties such as uniform $|R(k)|$ or evenly-spaced frequencies in Ω.

Since the feature weights ν_k of the Fourier modes are fixed by the choice of the data-encoding unitary, we can understand a choice of $S(x)$ as providing a structural bias of a quantum model towards different overfitting behaviors, and conversely the choices of feature weights available to quantum models are limited and are particular to the structure of the associated quantum circuit. Figure 4 shows distributions for feature weights arising

Encoding strategy	Hamiltonian example	Degeneracy $|R(k)|$	Spectrum Ω	$|\Omega|$
Hamming	$\frac{1}{2}(Z_0 + Z_1 + Z_2 + \dots)$	$\binom{2n_q}{n_q - |k|}$	$\{-n_q, 1 - n_q, \dots, n_q\}$	$2n_q + 1$
Binary	$\frac{1}{2}(Z_0 + 2Z_1 + 4Z_2 + \dots)$ $\sim \text{diag}\,(0, 1, 2, \dots, 2^{n_q} - 1)$	$2^{n_q} - |k|$	$\{-2^{n_q} + 1, \dots 2^{n_q} - 1\}$	$2^{n_q+1} - 1$
Ternary	$\frac{1}{2}(Z_0 + 3Z_1 + 9Z_2 + \dots)$	$2^{n_q - \|T(k)\|_1}$	$\left\{ -\left\lfloor \frac{3^{n_q}}{2} \right\rfloor, \dots, \left\lfloor \frac{3^{n_q}}{2} \right\rfloor \right\}$	3^{n_q}
Golomb	$\text{diag}\,(0, 1, 4, 6)$	$\begin{cases} d & k = 0 \\ 1 & k \neq 0 \end{cases}$	varies	$2\binom{d}{2} + 1$

Table 1: Spectra Ω and degeneracies $R(k)$ computed for various data-encoding Hamiltonians H defined for either d dimensions or n_q qubits. The Hamming, Binary, and Ternary data encoding strategies are realized on n_q qubits using a separable Hamiltonian consisting of Pauli-Z operators with different prefactor schemes. The Ternary encoding strategy results in the largest $|\Omega|$ possible for a separable data encoding Hamiltonian (see also Ref. [44]), while the Golomb encoding encoding strategy (named in reference to Golomb rulers, e.g. [45]) results in the largest $|\Omega|$ possible for any choice of d-dimensional data-encoding Hamiltonian. Note that the spectrum is preserved under permutations and additive shifts of the diagonal of the Hamiltonian and so we use "$\sim$" to denote equivalence up to these operations. The function T converts an integer to a signed ternary string, as defined in Eq. (234) (see Appendix C).

from the example encoding strategies presented in Table 2, and demonstrates a broad connection between the degeneracies $|R(k)|$ of the model (giving rise to feature weights ν_k^{opt}) and the generalization error $L(f^{opt})$.
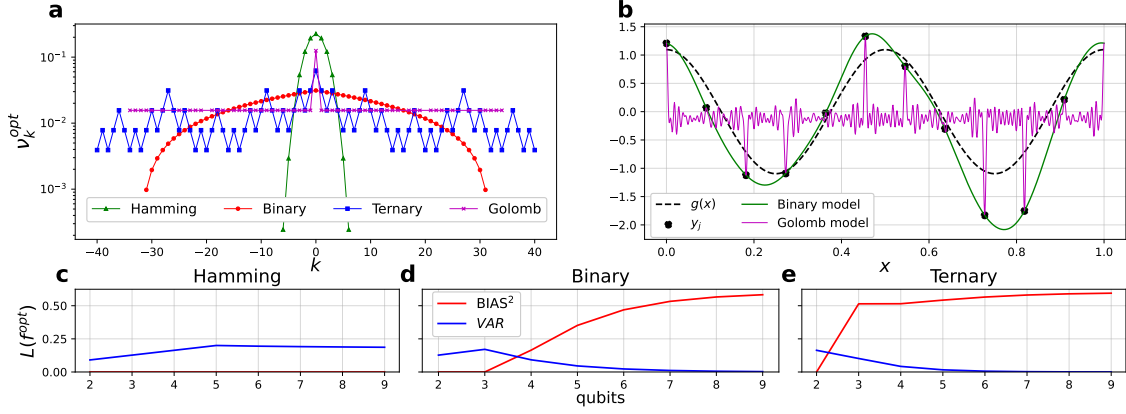
Figure 4: The degeneracies $|R(k)|$ for select quantum models directly influence the tendency of the models to overfit noise or bleed the underlying signal. **a** The degeneracies $|R(k)|$ for each frequency k for the four models introduced in Table 1 induce significantly different feature weights in the simplified quantum model. **b** Two models exemplify the extreme cases of fitting behaviors from Sec. 2.2: The Golomb encoding strategy (magenta) has significant weight on feature weights ν_k for large frequencies k, resulting in a "spiky" overfitting behavior that interpolates each noisy sample y_j but sacrifices recovery of the underlying signal $g(x)$; the Hamming encoding strategy (green) fits noisy samples smoothly and the error of the model is sensitive to the noise in the sampled data. **c-e** The BIAS^2 and VAR of the optimized Hamming encoding strategy (**c**), Binary encoding strategy (**d**), and Ternary encoding strategy (**e**) demonstrate the general effect of $|R(k)|$ and ν_k on the generalization error. The concentration of feature weights ν_k on $k \in \Omega_{n_0}$ in the Hamming encoding strategy reduces the BIAS (**c**), while ν_k having broad support over $k \in \Omega$ (e.g. Ternary, **e**) reduces VAR as the resulting f is close to its average (with respect to noise) except for becoming "spiky" in the vicinity of samples y_j, but increases the BIAS by introducing alias frequencies into the optimized model f^{opt}.

3.2 Trainable unitaries reweight the general quantum model

We now return to the general quantum model of Eq. (25) to understand how different choices of the state preparation unitary U (giving rise to input state $|\Gamma\rangle$) affect the feature weights ν_k^{opt} of the general quantum model, thereby influencing benign overfitting of the target function. While Eq. (31) differs from the correspondence $\nu_k = |R(k)|/d^2$ that we observed for the simplified quantum model, we see that the feature weights of the optimal quantum model still depend heavily on the degeneracies $|R(k)|$. For instance, the *average* ν_k^{opt} with respect to Haar-random d-dimensional input states $|\Gamma\rangle$ is proportional to $|R(k)|$ whenever $k \neq 0$:

$$\underset{U \sim \mathrm{U}(d)}{\mathbb{E}} \nu_k^{opt} = \frac{|R(k)| + d\delta_0^k}{d(d+1)}. \tag{34}$$

where δ_i^j denotes the Kronecker delta. Furthermore, in Appendix B we observe that the variance of ν_k^{opt} around its average tends to be small for encoding strategies considered in this work, for instance scaling like $\mathcal{O}\left(d^{-3}\right)$ for the Binary encoding strategy and $\mathcal{O}\left(d^{-4}\right)$ for the Golomb encoding strategy. This demonstrates that the feature weights of generic quantum models (i.e., ones for which U is randomly sampled) will be dominated by the degeneracy $|R(k)|$ introduced by the data-encoding operation $S(x)$.

Despite the behavior of ν_k being dominated by $|R(k)|$ in an average sense, there are specific choices of U for which the feature weights deviate significantly from this average. We will now use one such choice of U to provide a concrete example of an interpolating quantum model that exhibits benign overfitting. Suppose we choose U such that the

elements of $|\Gamma\rangle$ are given by

$$\gamma_j = \begin{cases} \sqrt{a}, & j \in [c_1, c_2), \\ \sqrt{b}, & \text{otherwise} \end{cases} \tag{35}$$

for $j \in [d]$, and some integers $0 < c_1 < c_2 < d$ and amplitudes a, b subject to normalization. We show that given a band-limited target function g with access to spectrum Ω_{n_0}, there is a specific choice of c_1, c_2 dependent on d and n_0 for which the interpolating quantum model f^{opt} of Eq. (30) also benefits from vanishing generalization error in the limit of many samples, namely we show in Appendix D that

$$L(f^{opt}) = \mathcal{O}\left(\frac{1}{n} + \frac{n}{d}\right). \tag{36}$$

Thus, by perfectly fitting the training data and exhibiting good generalization in the $n \to \infty$ limit, the quantum model exhibits benign overfitting. This behavior is outlined in Figure 5, which highlights the role that $|\Gamma\rangle$ plays in concentrating the feature weights ν_k within the spectrum Ω_{n_0} of g while preserving a long tail that provides the model with low-variance "spiky" behavior in the vicinity of noisy samples. In contrast, the feature weights for the Binary encoding strategy with a uniform input state has little support on Ω_{n_0} and resulting in a large BIAS error.

The above discussion shows how the input state amplitudes γ_i provide additional degrees of freedom with which the feature weights ν_k^{opt} can be tuned in order to modify the generalization behavior of the interpolating quantum model, and to exhibit benign overfitting in particular. It is therefore worthwhile to consider what other kinds of feature weights ν_k might be prepared by some choice of input state $|\Gamma\rangle$. We may use simple counting arguments to demonstrate the restrictions in designing particular distributions of feature weights. Suppose we define $\Omega_+ = \{k : k \in \Omega, k > 0\}$ containing the positive frequencies of a quantum model. Then the introduction of an arbitrary input state $|\Gamma\rangle$ provides us with $2^{n_q} - 1$ free parameters with which to tune $|\Omega_+|$-many terms in the distribution of ν_k^{opt} (subject to $\nu_k = \nu_{-k}$ and $\sum_k \nu_k = 1$). Clearly, there are distributions of feature weights ν_k^{opt} that can not be achieved for models where $|\Omega_+| \geq 2^{n_q}$.

Conversely, the condition $|\Omega_+| < 2^{n_q}$ does not necessarily mean that we can thoroughly explore the space of possible feature weights by modifying the input state $|\Gamma\rangle$. For example, consider the Hamming encoding strategy for which the number of free parameters controlling the distribution of feature weights ν_k^{opt} is $|\Omega_+| = n_q$, which is exponentially smaller the number of parameters in $|\Gamma\rangle$. While this might suggest significant freedom in adjusting ν_k^{opt}, the opposite is true: For any choice of input state $|\Gamma\rangle$, there is another state of the form

$$|\Gamma'\rangle = \sum_{i=0}^{n_q} \phi_i |\Phi_i\rangle, \tag{37}$$

that achieves exactly the same distribution of feature weights ν_k. In Eq. (37), $|\Phi_i\rangle$ describes a uniform superposition over all computational basis state bitstrings with weight i, and so the distribution of ν_k^{opt} actually only depends on $n_q + 1$ real parameters ϕ_i, $i = 0, 1, \ldots, n_q$, and the feature weights are invariant under any operations in U that preserve $|\Phi_i\rangle$ (see Appendix B). An example of such operations are the *particle-conserving unitaries* well-known in quantum chemistry, which act to preserve the number of set bits (i.e., the Hamming weight) when each bit represents a mode occupied by a particle in second quantization [46, 47]. This example demonstrates how symmetry in the data-encoding Hamiltonian (e.g. Refs. [48, 49]) can have a profound influence on the ability to prepare
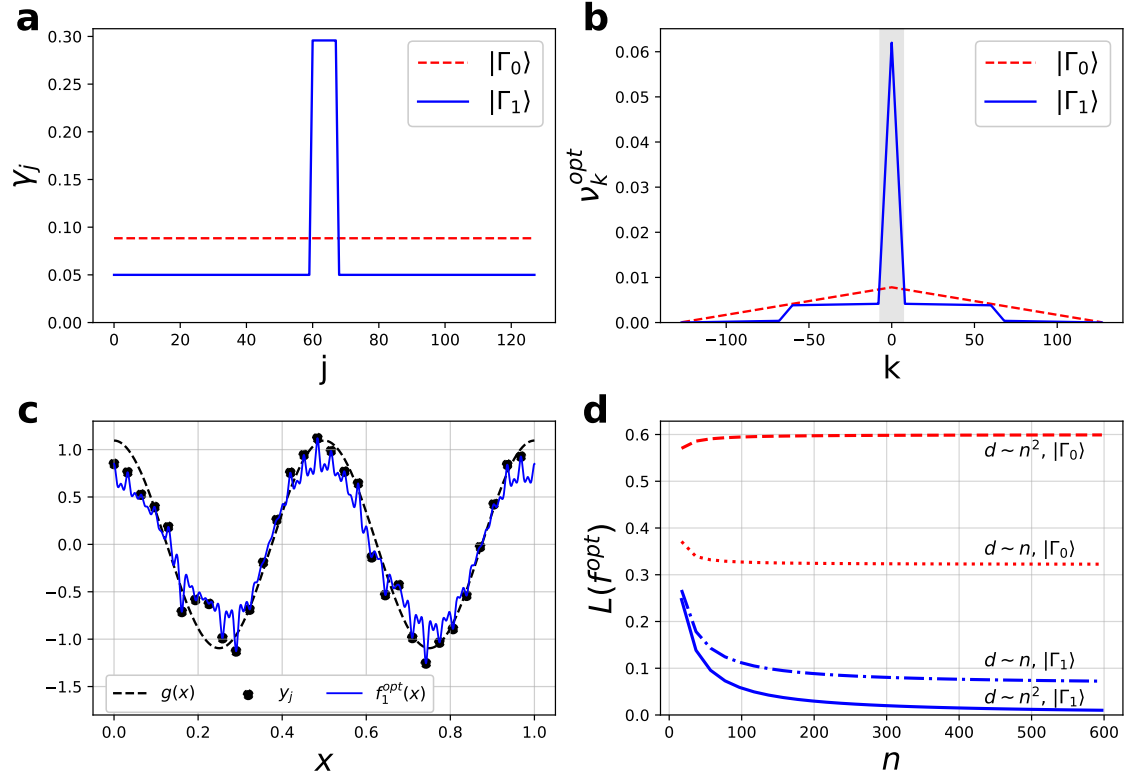
Figure 5: Benign overfitting for the general quantum model. **a** For the Hamiltonian $H \sim \operatorname{diag}((0, 1, \ldots, d-1))$ we construct a simple input state $|\Gamma_1\rangle$ (Eq. (35)) that is dependent on n_0 and d (see Appendix D), compared to the uniform state $|\Gamma_0\rangle = \frac{1}{\sqrt{d}} \sum_k |k\rangle$. **b** $|\Gamma_1\rangle$ induces feature weights ν_k^{opt} in the optimized model that are heavily concentrated in the target function spectrum Ω_{n_0} (gray band) and decay like $1/d$ and $1/d^2$ elsewhere for $k \in \Omega_d \backslash \Omega_{n_0}$, compared to the feature weights induced by $|\Gamma_0\rangle$ which have similar magnitude over all of Ω_d. **c** Plotting the $\|M\|_F$-minimizing quantum model f_1^{opt} with input state $|\Gamma_1\rangle$ for a sample signal ($d = 128$, $n = 31$) highlights how $|\Gamma_1\rangle$ and H work together to interpolate y_j via high-frequency "spiky" behavior near the sampled points, but otherwise closely matches the underlying signal $g(x)$. **d** The resulting quantum model benignly overfits the band-limited target function: As long as $d = \omega(n)$, the generalization error $L(f_1^{opt})$ of the minimum-$\|M\|_F$ interpolating quantum model vanishes for large enough sample size n.

specific distributions of feature weights ν_k^{opt}, and consequently affect the generalization and overparameterization behavior of the associated quantum models.

4 Conclusion

In this work we have taken a first step towards characterizing the phenomenon of benign overfitting in a quantum machine learning setting. We derived the error for an overparameterized Fourier features model that interpolates the (noisy) input signal with minimum ℓ_2-norm trainable weights and connected the feature weights associated with each Fourier mode to a trade-off in the generalization error of the model. We then demonstrated an analogous simplified quantum model for which the feature weights are induced by the choice of data-encoding unitary $S(x)$. Finally, we discussed how introducing an arbitrary state-preparation unitary U gives rise to effective feature weights in the optimized general quantum model, presenting the possibility of connecting U and $S(x)$ to benign overfitting in more general quantum models.

Our discussion of interpolating quantum models presents an interpretation of overparameterization (i.e., the size of the model spectrum Ω) that departs from other measures of quantum circuit complexity discussed in the literature [19, 50, 51], as even the simplified quantum models studied here are able to interpolate training data using a fixed circuit U and optimized measurements. We also reemphasize that – unlike much of the quantum machine learning literature – we do not consider a setting where the model is optimized with respect to a trainable circuit, as the model of Eq. (30) is constructed to exhibit zero training error (and can therefore not be improved via optimization). Finding the input state $|\Gamma\rangle$ that will result in a specific distribution of feature weights ν_k^{opt} generally requires solving a $|\Omega_+|$-dimensional system of equations that are second order in 2^{n_q} many real parameters $|\gamma_i|^2$ (i.e., inverting the map of the form $\mathbb{R}^{2^{n_q}} \to \mathbb{R}^{|\Omega_+|}$ in Eq. (31)) or otherwise performing a variational optimization that will likely fail due to the familiar phenomenon of *barren plateaus* [23, 24, 52, 53].

While we have shown an example of benign overfitting by a quantum model in a relatively restricted context, future work may lead to more general characterizations of this phenomenon. Similar behavior likely exists for quantum kernel methods and may complement existing studies on these methods' generalization power [54]. An exciting possibility would be to demonstrate benign overfitting in quantum models trained on distributions of quantum states which are hard to learn classically [55, 56], thereby extending the growing body of statistical learning theory for quantum learning algorithms [27, 28, 57].

5 Code availability

Code to reproduce the figures and analysis is available at the following repository: `https://github.com/peterse/benign-overfitting-quantum`.

6 Acknowledgements

by the Province of Ontario through the Ministry of Colleges and Universities. Circuit simulations were performed in PennyLane [58].

References

[1] Michael A Nielsen. "Neural networks and deep learning". Determination Press. (2015). url: http://neuralnetworksanddeeplearning.com/.

[2] Stuart Geman, Elie Bienenstock, and René Doursat. "Neural networks and the bias/variance dilemma". Neural Comput. **4**, 1–58 (1992).

[3] Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. "The elements of statistical learning: data mining, inference, and prediction". Volume 2. Springer. (2009).

[4] Peter L. Bartlett, Andrea Montanari, and Alexander Rakhlin. "Deep learning: a statistical viewpoint". Acta Numerica **30**, 87–201 (2021).

[5] Mikhail Belkin. "Fit without fear: remarkable mathematical phenomena of deep learning through the prism of interpolation". Acta Numerica **30**, 203–248 (2021).

[6] Peter L. Bartlett, Philip M. Long, Gábor Lugosi, and Alexander Tsigler. "Benign overfitting in linear regression". Proc. Natl. Acad. Sci. **117**, 30063–30070 (2020).

[7] Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. "Reconciling modern machine-learning practice and the classical bias-variance trade-off". Proc. Natl. Acad. Sci. **116**, 15849–15854 (2019).

[8] Mikhail Belkin, Alexander Rakhlin, and Alexandre B. Tsybakov. "Does data interpolation contradict statistical optimality?". In Proceedings of Machine Learning Research. Volume 89, pages 1611–1619. PMLR (2019). url: https://proceedings.mlr.press/v89/belkin19a.html.

[9] Vidya Muthukumar, Kailas Vodrahalli, Vignesh Subramanian, and Anant Sahai. "Harmless interpolation of noisy data in regression". IEEE Journal on Selected Areas in Information Theory **1**, 67–83 (2020).

[10] Vidya Muthukumar, Adhyyan Narang, Vignesh Subramanian, Mikhail Belkin, Daniel Hsu, and Anant Sahai. "Classification vs regression in overparameterized regimes: Does the loss function matter?". J. Mach. Learn. Res. **22**, 1–69 (2021). url: http://jmlr.org/papers/v22/20-603.html.

[11] Yehuda Dar, Vidya Muthukumar, and Richard G. Baraniuk. "A farewell to the bias-variance tradeoff? an overview of the theory of overparameterized machine learning" (2021). arXiv:2109.02355.

[12] Marcello Benedetti, Erika Lloyd, Stefan Sack, and Mattia Fiorentini. "Parameterized quantum circuits as machine learning models". Quantum Sci. Technol. **4**, 043001 (2019).

[13] K. Mitarai, M. Negoro, M. Kitagawa, and K. Fujii. "Quantum circuit learning". Phys. Rev. A **98**, 032309 (2018).

[14] Maria Schuld, Ville Bergholm, Christian Gogolin, Josh Izaac, and Nathan Killoran. "Evaluating analytic gradients on quantum hardware". Phys. Rev. A **99**, 032331 (2019).

[15] Maria Schuld and Nathan Killoran. "Quantum machine learning in feature hilbert spaces". Phys. Rev. Lett. **122**, 040504 (2019).

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

[17] Seth Lloyd and Christian Weedbrook. "Quantum generative adversarial learning". Phys. Rev. Lett. **121**, 040502 (2018).

[18] Pierre-Luc Dallaire-Demers and Nathan Killoran. "Quantum generative adversarial networks". Phys. Rev. A **98**, 012324 (2018).

[19] Amira Abbas, David Sutter, Christa Zoufal, Aurelien Lucchi, Alessio Figalli, and Stefan Woerner. "The power of quantum neural networks". Nat. Comput. Sci. **1**, 403–409 (2021).

[20] Logan G. Wright and Peter L. McMahon. "The capacity of quantum neural networks". In 2020 Conference on Lasers and Electro-Optics (CLEO). Pages 1–2. (2020). url: https://ieeexplore.ieee.org/document/9193529.

[21] Sukin Sim, Peter D. Johnson, and Alán Aspuru-Guzik. "Expressibility and entangling capability of parameterized quantum circuits for hybrid quantum-classical algorithms". Adv. Quantum Technol. **2**, 1900070 (2019).

[22] Thomas Hubregtsen, Josef Pichlmeier, Patrick Stecher, and Koen Bertels. "Evaluation of parameterized quantum circuits: on the relation between classification accuracy, expressibility and entangling capability". Quantum Mach. Intell. **3**, 1 (2021).

[23] Jarrod R McClean, Sergio Boixo, Vadim N Smelyanskiy, Ryan Babbush, and Hartmut Neven. "Barren plateaus in quantum neural network training landscapes". Nat. Commun. **9**, 4812 (2018).

[24] Marco Cerezo, Akira Sone, Tyler Volkoff, Lukasz Cincio, and Patrick J Coles. "Cost function dependent barren plateaus in shallow parametrized quantum circuits". Nat. Commun. **12**, 1791 (2021).

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

[26] Hsin-Yuan Huang, Michael Broughton, Masoud Mohseni, Ryan Babbush, Sergio Boixo, Hartmut Neven, and Jarrod R McClean. "Power of data in quantum machine learning". Nat. Commun. **12**, 2631 (2021).

[27] 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". Nat. Commun. **13**, 4919 (2022).

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

[29] Francisco Javier Gil Vidal and Dirk Oliver Theis. "Input redundancy for parameterized quantum circuits". Front. Phys. **8**, 297 (2020).

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

[31] David Wierichs, Josh Izaac, Cody Wang, and Cedric Yen-Yu Lin. "General parameter-shift rules for quantum gradients". Quantum **6**, 677 (2022).

[32] Kendall E Atkinson. "An introduction to numerical analysis". John Wiley & Sons. (2008).

[33] Ali Rahimi and Benjamin Recht. "Random features for large-scale kernel machines". In Advances in Neural Information Processing Systems. Volume 20. (2007). url: https://papers.nips.cc/paper_files/paper/2007/hash/013a006f03dbc5392effeb8f18fda755-Abstract.html.

[34] Walter Rudin. "The basic theorems of fourier analysis". John Wiley & Sons, Ltd. (1990).

[35] Song Mei and Andrea Montanari. "The generalization error of random features regression: Precise asymptotics and the double descent curve". Commun. Pure Appl. Math. 75, 667–766 (2022).

[36] Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J. Tibshirani. "Surprises in high-dimensional ridgeless least squares interpolation". Ann. Stat. 50, 949 – 986 (2022).

[37] Tengyuan Liang, Alexander Rakhlin, and Xiyu Zhai. "On the multiple descent of minimum-norm interpolants and restricted lower isometry of kernels". In Proceedings of Machine Learning Research. Volume 125, pages 1–29. PMLR (2020). url: http://proceedings.mlr.press/v125/liang20a.html.

[38] Edward Farhi and Hartmut Neven. "Classification with quantum neural networks on near term processors" (2018). arXiv:1802.06002.

[39] Maria Schuld, Alex Bocharov, Krysta M. Svore, and Nathan Wiebe. "Circuit-centric quantum classifiers". Phys. Rev. A 101, 032308 (2020).

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

[41] Sofiene Jerbi, Lukas J Fiderer, Hendrik Poulsen Nautrup, Jonas M Kübler, Hans J Briegel, and Vedran Dunjko. "Quantum machine learning beyond kernel methods". Nat. Commun. 14, 517 (2023).

[42] Casper Gyurik, Dyon Vreumingen, van, and Vedran Dunjko. "Structural risk minimization for quantum linear classifiers". Quantum 7, 893 (2023).

[43] Maria Schuld. "Supervised quantum machine learning models are kernel methods" (2021). arXiv:2101.11020.

[44] S. Shin, Y. S. Teo, and H. Jeong. "Exponential data encoding for quantum supervised learning". Phys. Rev. A 107, 012422 (2023).

[45] Sophie Piccard. "Sur les ensembles de distances des ensembles de points d'un espace euclidien.". Memoires de l'Universite de Neuchatel. Secretariat de l'Universite. (1939).

[46] Dave Wecker, Matthew B. Hastings, Nathan Wiebe, Bryan K. Clark, Chetan Nayak, and Matthias Troyer. "Solving strongly correlated electron models on a quantum computer". Phys. Rev. A 92, 062318 (2015).

[47] Ian D. Kivlichan, Jarrod McClean, Nathan Wiebe, Craig Gidney, Alán Aspuru-Guzik, Garnet Kin-Lic Chan, and Ryan Babbush. "Quantum simulation of electronic structure with linear depth and connectivity". Phys. Rev. Lett. 120, 110501 (2018).

[48] Martín Larocca, Frédéric Sauvage, Faris M. Sbahi, Guillaume Verdon, Patrick J. Coles, and M. Cerezo. "Group-invariant quantum machine learning". PRX Quantum 3, 030341 (2022).

[49] Johannes Jakob Meyer, Marian Mularski, Elies Gil-Fuster, Antonio Anna Mele, Francesco Arzani, Alissa Wilms, and Jens Eisert. "Exploiting symmetry in variational quantum machine learning". PRX Quantum **4**, 010328 (2023).

[50] Martin Larocca, Nathan Ju, Diego García-Martín, Patrick J Coles, and Marco Cerezo. "Theory of overparametrization in quantum neural networks". Nat. Comput. Sci. **3**, 542–551 (2023).

[51] Yuxuan Du, Min-Hsiu Hsieh, Tongliang Liu, and Dacheng Tao. "Expressive power of parametrized quantum circuits". Phys. Rev. Res. **2**, 033125 (2020).

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

[53] Samson Wang, Enrico Fontana, Marco Cerezo, Kunal Sharma, Akira Sone, Lukasz Cincio, and Patrick J Coles. "Noise-induced barren plateaus in variational quantum algorithms". Nat. Commun. **12**, 6961 (2021).

[54] Abdulkadir Canatar, Evan Peters, Cengiz Pehlevan, Stefan M. Wild, and Ruslan Shaydulin. "Bandwidth enables generalization in quantum kernel models". Transactions on Machine Learning Research (2023). url: https://openreview.net/forum?id=A1N2qp4yAq.

[55] Hsin-Yuan Huang, Michael Broughton, Jordan Cotler, Sitan Chen, Jerry Li, Masoud Mohseni, Hartmut Neven, Ryan Babbush, Richard Kueng, John Preskill, and Jarrod R. McClean. "Quantum advantage in learning from experiments". Science **376**, 1182–1186 (2022).

[56] Sitan Chen, Jordan Cotler, Hsin-Yuan Huang, and Jerry Li. "Exponential separations between learning with and without quantum memory". In 2021 IEEE 62nd Annual Symposium on Foundations of Computer Science (FOCS). Pages 574–585. (2022).

[57] Hsin-Yuan Huang, Richard Kueng, and John Preskill. "Information-theoretic bounds on quantum advantage in machine learning". Phys. Rev. Lett. **126**, 190505 (2021).

[58] Ville Bergholm, Josh Izaac, Maria Schuld, Christian Gogolin, M. Sohaib Alam, Shahnawaz Ahmed, Juan Miguel Arrazola, Carsten Blank, Alain Delgado, Soran Jahangiri, Keri McKiernan, Johannes Jakob Meyer, Zeyue Niu, Antal Száva, and Nathan Killoran. "Pennylane: Automatic differentiation of hybrid quantum-classical computations" (2018). arXiv:1811.04968.

[59] Peter L. Bartlett, Philip M. Long, Gábor Lugosi, and Alexander Tsigler. "Benign overfitting in linear regression". Proc. Natl. Acad. Sci. **117**, 30063–30070 (2020).

[60] Vladimir Koltchinskii and Karim Lounici. "Concentration inequalities and moment bounds for sample covariance operators". Bernoulli **23**, 110 – 133 (2017).

[61] Zbigniew Puchała and Jarosław Adam Miszczak. "Symbolic integration with respect to the haar measure on the unitary group". Bull. Pol. Acad. Sci. **65**, 21–27 (2017).

[62] Daniel A. Roberts and Beni Yoshida. "Chaos and complexity by design". J. High Energy Phys. **2017**, 121 (2017).

[63] Wallace C. Babcock. "Intermodulation interference in radio systems frequency of occurrence and control by channel selection". Bell Syst. tech. j. **32**, 63–73 (1953).

[64] M. Atkinson, N. Santoro, and J. Urrutia. "Integer sets with distinct sums and differences and carrier frequency assignments for nonlinear repeaters". IEEE Trans. Commun. **34**, 614–617 (1986).

[65] J. Robinson and A. Bernstein. "A class of binary recurrent codes with limited error propagation". IEEE Trans. Inf. **13**, 106–113 (1967).

[66] R. J. F. Fang and W. A. Sandrin. "Carrier frequency assignment for nonlinear repeaters". COMSAT Technical Review **7**, 227–245 (1977).

A Solution for the classical overparameterized model

In this section we derive the optimal solution and generalization error for the classical overparameterized weighted Fourier functions model. We then discuss the conditions under which benign overfitting may be observed and construct examples of the phenomenon.

A.1 Linearization of overparameterized Fourier model

We first show that the classical overparameterized Fourier features model may be cast as a *linear* model under an appropriate orthogonal transformation. We are interested in learning a target function of the form

$$g(x) = \sum_{k \in \Omega_{n_0}} \hat{g}_k e^{i2\pi kx}, \tag{38}$$

with the additional constraint that the Fourier coefficients satisfy $\hat{g}_k = \hat{g}^*_{-k}$ such that g is real. The spectrum of Eq. (38) for odd n_0 only contains integer frequencies,

$$\Omega_{n_0} = \left\{ -\frac{n_0 - 1}{2}, \dots, 0, \dots, \frac{n_0 - 1}{2} \right\}, \tag{39}$$

and we accordingly call g a *bandlimited* function with bandlimit $n_0/2 - 1$. To learn g, we first sample n equally spaced datapoints on the interval $[0, 1]$,

$$x_j = \frac{j}{n}, \qquad j \in [n], \tag{40}$$

where $[n] = \{0, 1, 2, \dots, n-1\}$ and we assume n is odd, and we then evaluate $g(x_j)$ with additive error. This noisy sampling process yields n observations of the form $y_j = g(x_j) + \epsilon$ with $\mathbb{E}[\epsilon^2] = \sigma^2$. We will fit the observations y_j using overparameterized Fourier features models of the form

$$f(x) = \langle \phi(x), \boldsymbol{\alpha} \rangle = \sum_{k \in \Omega_d} \alpha_k \sqrt{\nu_k} e^{i2\pi kx}, \tag{41}$$

with $\boldsymbol{\alpha} \in \mathbb{C}^d$, and we have introduced weighted Fourier features $\phi : \mathbb{R} \to \mathbb{C}^d$ defined elementwise as

$$\phi(x)_k = \sqrt{\nu_k} e^{i2\pi kx}. \tag{42}$$

In Eq. (41), Ω_d describes the set of frequencies available to the model for any choice of $d \geq n \geq n_0$. We are interested in the case where f interpolates the observations y_j, i.e., $f(x_j) = y_j$ for all $j = 0, \dots, n-1$. To this end, we define a $n \times d$ feature matrix Φ whose rows are given by $\phi(x_j)^\dagger$:

$$\Phi_{jk} = [\phi(x_j)]^*_k = \sqrt{\nu_k} e^{-i2\pi jk/n}. \tag{43}$$

The interpolation condition may then be stated in matrix form as

$$\Phi \boldsymbol{\alpha} = \mathbf{y}, \tag{44}$$

where $(\mathbf{y})_j = y_j$ is the vector of noisy observations. Ω_d contains alias frequencies of Ω_{n_0}, and so there the choice of $\boldsymbol{\alpha}$ that satisfies Eq. (44) is not unique. Here we will focus on the minimum-ℓ_2 norm interpolating solution,

$$\boldsymbol{\alpha}^{opt} = \arg\min_{\Phi \boldsymbol{\alpha} = \mathbf{y}} \|\boldsymbol{\alpha}\|_2. \tag{45}$$

A.1.1 Fourier transform into the linear model

We will now show that Eq. (45) with uniformly-spaced datapoints x_j can be solved using methods from ordinary linear regression under a suitable choice of transformation. Defining the n-th root of identity as $\omega = e^{i2\pi/n}$, then the j-th row of the LHS of Eq. (44) is equivalent to

$$\sum_{k \in \Omega_d} \alpha_k \sqrt{\nu_k} \omega^{-kj} = \sum_{k \in \Omega_n} \sum_{\ell \in S(k)} \alpha_{k+n\ell} \sqrt{\nu_{k+n\ell}} \omega^{-(k+n\ell)j} \tag{46}$$

$$= \sum_{k \in \Omega_n} \omega^{-kj} \sum_{\ell \in S(k)} \alpha_{k+n\ell} \sqrt{\nu_{k+n\ell}} \tag{47}$$

$$:= \sum_{k \in \Omega_n} \omega^{-kj} \langle P_{S(k)} \boldsymbol{\alpha}, P_{S(k)} \mathbf{w} \rangle. \tag{48}$$

where $S(k) = \{j : j \in \Omega_d, j \bmod n = k\}$ is the set of alias frequencies of k appearing in Ω_d, i.e. the set of frequencies $k + \ell n$ with $\ell \in \mathbb{Z}$ that obey

$$e^{i2\pi k x_j} = e^{i2\pi kj/n} = e^{i2\pi(k+\ell n)j/n} = e^{i2\pi(k+\ell n)x_j}, \tag{49}$$

which follows from $e^{i2\pi \ell k} = 1$ since $k \in \mathbb{Z}$. The operator $P_{S(k)} : \mathbb{C}^d \to \mathbb{C}^d$ is the projector onto the set of standard basis vectors in $\mathbb{C}^d$ with indices in $S(k)$, and $(\mathbf{w})_k := \sqrt{\nu_k}$. The roots of unity are orthonormal since

$$\sum_{k=0}^{n-1} \omega^{k(p-q)} = n\delta_p^q \tag{50}$$

for $p, q \in [n]$. This implies

$$\sum_{k \in \Omega_n} \omega^{k(p-q)} = \omega^{\min(\Omega_n)(p-q)} n\delta_p^q = n\delta_p^q, \tag{51}$$

for $p, q \in \Omega_n$. Defining the discrete Fourier Transform $\hat{y}_k$ of $\mathbf{y}$ according to

$$\hat{y}_j = \frac{1}{n} \sum_{k \in \Omega_n} y_k \omega^{jk}, \tag{52}$$

then using the identity of Eq. (51), we evaluate the j-th row of Eq. (44) as

$$\sum_{k \in \Omega_n} \omega^{-kj} \langle P_{S(k)} \boldsymbol{\alpha}, P_{S(k)} \mathbf{w} \rangle = y_j, \tag{53}$$

$$\sum_{j \in \Omega_n} \omega^{pj} \sum_{k \in \Omega_n} \omega^{-kj} \langle P_{S(k)} \boldsymbol{\alpha}, P_{S(k)} \mathbf{w} \rangle = \sum_{j \in \Omega_n} \omega^{pj} \left(\sum_{k \in \Omega_n} \hat{y}_k \omega^{-kj} \right), \tag{54}$$

$$\sum_{k \in \Omega_n} n\delta_k^p \langle P_{S(k)} \boldsymbol{\alpha}, P_{S(k)} \mathbf{w} \rangle = \sum_{k \in \Omega_n} n\delta_k^p \hat{y}_k, \tag{55}$$

$$\langle P_{S(p)} \boldsymbol{\alpha}, P_{S(p)} \mathbf{w} \rangle = \hat{y}_p. \tag{56}$$

Inspecting this final line yields a new matrix equation: Let X be an $n \times d$ matrix X with elements

$$X_{jk} = (P_{S(j)} \mathbf{w})_k = \sqrt{\nu_k} \mathbb{1}\{k \in S(j)\}, \tag{57}$$

where the conditional operator $\mathbb{1}\{Z\}$ evaluates to 1 if the predicate Z is true, and 0 otherwise. Then we may express Eq. (56) for all $p \in \Omega_n$ as a matrix equation

$$X\boldsymbol{\alpha} = \hat{\mathbf{y}}, \tag{58}$$

where $(\hat{\mathbf{y}})_j = \hat{y}_j$. We have shown that Eq. (44) is exactly equivalent to Eq. (58) for uniformly spaced inputs, and as $\boldsymbol{\alpha}$ is unchanged between these two representations this implies that the solution to Eq. (45) is also given by

$$\boldsymbol{\alpha}^{opt} = \underset{X\boldsymbol{\alpha}=\hat{\mathbf{y}}}{\arg\min} \|\boldsymbol{\alpha}\|_2. \tag{59}$$

Therefore, the minimum ℓ_2-norm solution to interpolate the input signal using weighted Fourier functions provided as samples y_j is exactly the same as the minimum ℓ_2-norm solution for an equivalent *linear regression* problem on the matrix X with targets $\hat{\mathbf{y}}$. Furthermore, this linear regression problem is related to the original problem via Fourier transform. Let F be the (nonunitary) discrete Fourier transform defined on $\mathbb{C}^n$ elementwise as

$$F_{jk} = \omega^{jk}, \tag{60}$$

$$F^\dagger F = n\mathbb{I}_n. \tag{61}$$

Then for $k \in \Omega_d$, $j \in \Omega_n$,

$$\frac{1}{n}(F\Phi)_{jk} = \frac{1}{n}\sum_{\ell \in \Omega_n} \omega^{(k-j)\ell}\sqrt{\nu_k} = \frac{1}{n}\left(n\delta_j^{k \bmod n}\right)\sqrt{\nu_k} = \sqrt{\nu_k}\mathbb{1}\{k \in S(j)\} = X_{jk}, \tag{62}$$

implying

$$\frac{1}{n}F\Phi = X. \tag{63}$$

We may similarly recover $\frac{1}{n}F\mathbf{y} = \hat{\mathbf{y}}$ to show that the coefficients $\hat{y}_k$ are given by a discrete Fourier transform of y_j.

A.1.2 Error analysis of the linear model

Having shown that the discrete Fourier transform F relates the original system of trigonometric equations $\Phi\boldsymbol{\alpha} = \mathbf{y}$ to the system of linear equations (or ordinary least squares problem $X\boldsymbol{\alpha} = \hat{\mathbf{y}}$, where we treat the rows of X as observations in $\mathbb{R}^d$), we now derive the error of the problem in the Fourier representation. Standard treatment for ordinary least squares (OLS) gives the minimum ℓ_2-norm solution to Eq. (59) as

$$\boldsymbol{\alpha}^{opt} = X^T(XX^T)^{-1}\hat{\mathbf{y}} \tag{64}$$

in which case the optimal interpolating overparameterized Fourier model is

$$f^{opt}(x) = \langle \phi(x), \boldsymbol{\alpha}^{opt} \rangle \tag{65}$$

$$= \sum_{k \in \Omega_n} \frac{\hat{y}_k}{\sum_{j \in S(k)} \nu_j} \sum_{\ell \in S(k)} \nu_\ell e^{i2\pi\ell x}. \tag{66}$$

Once we have trained a model on noisy samples $\mathbf{y}$ of g using uniformly spaced values of x, we would like to evaluate how well the model performs for arbitrary x in $[0, 1]$. Given some function $f = \langle \phi(x), \boldsymbol{\alpha} \rangle$ the mean squared error $\mathbb{E}_x(f(x) - g(x))^2$ may be evaluated

with respect to the interval $[0, 1]$. We define the *generalization error* of the model as the expected mean squared error evaluated with respect to $\mathbf{y}$:

$$L(f) := \mathbb{E}_{x,\mathbf{y}}(f(x) - g(x))^2. \tag{67}$$

We decompose the generalization error of the model as

$$\mathbb{E}_{x,\mathbf{y}}(f(x) - g(x))^2 = \mathbb{E}_{x,\mathbf{y}}(f(x) - \mathbb{E}_{\mathbf{y}}f(x) + \mathbb{E}_{\mathbf{y}}f(x) - g(x))^2 \tag{68}$$

$$= \mathbb{E}_{x,\mathbf{y}}(f(x) - \mathbb{E}_{\mathbf{y}}f(x))^2 + \mathbb{E}_{x,\mathbf{y}}(\mathbb{E}_{\mathbf{y}}f(x) - g(x))^2 \tag{69}$$

$$+ 2\mathbb{E}_{x,\mathbf{y}}\left[(f(x) - \mathbb{E}_{\mathbf{y}}f(x))(\mathbb{E}_{\mathbf{y}}f(x) - g(x))\right].$$

Because of orthonormality of Fourier basis functions, the cross-terms cancel resulting in a decomposition of the generalization error of the optimal standard bias and variance terms:

$$L(f^{opt}) = \underbrace{\mathbb{E}_{x,\mathbf{y}}(f^{opt}(x) - \mathbb{E}_{\mathbf{y}}f^{opt}(x))^2}_{\text{VAR}} + \underbrace{\mathbb{E}_{x,\mathbf{y}}(\mathbb{E}_{\mathbf{y}}f^{opt}(x) - g(x))^2}_{\text{BIAS}^2}. \tag{70}$$

We now evaluate VAR and BIAS^2 using the linear representation developed in the previous section. Beginning with the variance, conditional on constructing Φ from the set of uniformly spaced points x_j we apply the discrete Fourier transformation to yield X and compute $\boldsymbol{\alpha}^{opt}$ using Eq. (64):

$$\text{VAR} = \mathbb{E}_{x,\mathbf{y}}(f^{opt}(x) - \mathbb{E}_{\mathbf{y}}f^{opt}(x))^2 \tag{71}$$

$$= \mathbb{E}_{x,\mathbf{y}}|\langle\phi(x), \boldsymbol{\alpha}^{opt} - \mathbb{E}_{\mathbf{y}}\boldsymbol{\alpha}^{opt}\rangle|^2 \tag{72}$$

$$= \mathbb{E}_{x,\mathbf{y}}|\langle\phi(x), X^T(XX^T)^{-1}(\hat{\mathbf{y}} - \mathbb{E}_{\mathbf{y}}\hat{\mathbf{y}})\rangle|^2 \tag{73}$$

$$= \text{Tr}\left(\mathbb{E}_{\mathbf{y}}\left[(\hat{\mathbf{y}} - \mathbb{E}_{\mathbf{y}}\hat{\mathbf{y}})(\hat{\mathbf{y}} - \mathbb{E}_{\mathbf{y}}\hat{\mathbf{y}})^\dagger\right](XX^T)^{-1}X\mathbb{E}_x\left[\phi(x)\phi(x)^\dagger\right]X^T(XX^T)^{-1}\right) \tag{74}$$

$$= \frac{\sigma^2}{n}\text{Tr}\left((XX^T)^{-2}X\Sigma_\phi X^T\right). \tag{75}$$

Letting $\boldsymbol{\epsilon} := (\mathbf{y} - \mathbb{E}_{\mathbf{y}}\mathbf{y})$ we have simplified the above using the following[2]:

$$\mathbb{E}_{\mathbf{y}}\left[(\hat{\mathbf{y}} - \mathbb{E}_{\mathbf{y}}\hat{\mathbf{y}})(\hat{\mathbf{y}} - \mathbb{E}_{\mathbf{y}}\hat{\mathbf{y}})^\dagger\right]_{jk} = \frac{1}{n^2}\mathbb{E}_{\mathbf{y}}(F\boldsymbol{\epsilon})_j(F\boldsymbol{\epsilon})_k^* \tag{76}$$

$$= \frac{1}{n^2}\mathbb{E}_{\mathbf{y}}\left(\sum_{p\in\Omega_n}\omega^{-jp}\epsilon_p\right)\left(\sum_{q\in\Omega_n}\omega^{qk}\epsilon_q\right) \tag{77}$$

$$= \frac{1}{n^2}\sum_{p,q\in\Omega_n}\mathbb{E}_{\mathbf{y}}[\epsilon_p\epsilon_q]\omega^{-jp+kq} \tag{78}$$

$$= \frac{\sigma^2}{n}\delta_j^k, \tag{79}$$

where we have used the fact that the errors $\boldsymbol{\epsilon}$ are independent and zero mean, $\mathbb{E}_{\mathbf{y}}[\epsilon_p\epsilon_q] = \mathbb{E}_{\mathbf{y}}[\epsilon_p^2]\delta_p^q$. We have defined the feature covariance matrix as $\Sigma_\phi = \mathbb{E}_x[\phi(x)\phi(x)^\dagger]$, which may be computed elementwise using the orthonormality of Fourier features on $[0, 1]$:

$$(\Sigma_\phi)_{jk} = \mathbb{E}_x[\phi(x)_j\phi(x)_k^*] = \sqrt{\nu_j\nu_k}\langle e^{i2\pi jx}, e^{i2\pi kx}\rangle_{L_2^{[0,1]}} = \delta_j^k\nu_j. \tag{80}$$

---

[2]In Eq. (75), $\mathbb{E}[X^TX] = \Sigma_\phi$, so that Σ_ϕ differs from the covariance matrix for the rows of X by a factor of $1/n$.

The following may be computed directly:

$$(XX^T)_{jk} = \delta_j^k \sum_{\ell \in S(j)} \nu_\ell \tag{81}$$

$$(X\Sigma_{\mathbf{x}}X^T)_{jk} = \sum_{\ell \in \Omega_d} X_{j\ell} \left(\frac{\nu_\ell}{n} \right) X_{k\ell} \tag{82}$$

$$= \frac{1}{n} \sum_{\ell \in \Omega_d} \nu_p^2 \mathbb{1}\{\ell \in S(j)\}\mathbb{1}\{\ell \in S(k)\} \tag{83}$$

$$= \frac{\delta_j^k}{n} \sum_{\ell \in S(k)} \nu_\ell^2. \tag{84}$$

In line (84) we have used the identity

$$\mathbb{1}\{\ell \in S(j)\}\mathbb{1}\{\ell \in S(k)\} = \mathbb{1}\{\ell \in S(j)\}\delta_k^j, \tag{85}$$

since $\ell \in S(k) \Rightarrow k = \ell \bmod n$ and therefore $\ell \in S(j) \Rightarrow j = \ell \bmod n = k$. We now compute the variance as

$$\text{VAR} = \frac{\sigma^2}{n} \text{Tr}\left((XX^T)^{-2} X\Sigma_\phi X^T) \right) \tag{86}$$

$$= \frac{\sigma^2}{n} \sum_{j,k \in \Omega_n} \left(\delta_k^j \sum_{\ell \in S(k)} \nu_\ell \right)^{-2} \left(\delta_j^k \sum_{\ell \in S(k)} \nu_\ell^2 \right) \tag{87}$$

$$= \frac{\sigma^2}{n} \sum_{k \in \Omega_n} \frac{\sum_{\ell \in S(k)} \nu_\ell^2}{\left(\sum_{\ell \in S(k)} \nu_\ell \right)^2}. \tag{88}$$

To evaluate the BIAS^2, we will first rewrite $g(x)$ in terms of its Fourier representation,

$$g(x) = \langle \phi(x), \boldsymbol{\alpha}_0 \rangle, \tag{89}$$

$$(\boldsymbol{\alpha}_0)_k = \frac{\hat{g}_k}{\sqrt{\nu_k}} \mathbb{1}\{k \in \Omega_{n_0}\}. \tag{90}$$

Noting that $(X\boldsymbol{\alpha}_0)_k = \hat{g}_k$ implies $X\boldsymbol{\alpha}_0 = \mathbb{E}_{\mathbf{y}}\hat{\mathbf{y}}$ we evaluate

$$\text{BIAS}^2 = \mathbb{E}_{x,\mathbf{y}}(\mathbb{E}_{\mathbf{y}} f^{opt}(x) - g(x))^2 \tag{91}$$

$$= \mathbb{E}_{x,\mathbf{y}} |\langle \phi(x), \mathbb{E}_{\mathbf{y}}\boldsymbol{\alpha}^{opt} - \boldsymbol{\alpha}_0 \rangle|^2 \tag{92}$$

$$= \text{Tr}\left(\mathbb{E}_x \left[\phi(x)\phi(x)^\dagger \right] (\mathbb{E}_{\mathbf{y}}\boldsymbol{\alpha}^{opt} - \boldsymbol{\alpha}_0)(\mathbb{E}_{\mathbf{y}}\boldsymbol{\alpha}^{opt} - \boldsymbol{\alpha}_0)^\dagger \right) \tag{93}$$

$$= \text{Tr}\left(\Sigma_\phi (X^T(XX^T)^{-1}X - \mathbb{I}_d)\boldsymbol{\alpha}_0\boldsymbol{\alpha}_0^\dagger (X^T(XX^T)^{-1}X - \mathbb{I}_d)^\dagger \right) \tag{94}$$

$$= \|\Sigma_\phi^{1/2}(\mathbb{I}_d - X^T(XX^T)^{-1}X)\boldsymbol{\alpha}_0\|_2^2. \tag{95}$$

While Eq. (95) already completely characterizes the bias error in terms of the choice of feature weights and input data, it may be greatly simplified by taking advantage of the sparseness X. We have

$$(X^T(XX^T)^{-1}X)_{jk} = \sum_{\ell \in \Omega_n} \left(\sum_{m \in S(\ell)} \nu_m \right)^{-1} X_{\ell j} X_{\ell k} \tag{96}$$

$$= \left(\sum_{m \in S(j \bmod n)} \nu_m \right)^{-1} \sqrt{\nu_k}\sqrt{\nu_j} \mathbb{1}\{k \in S(j \bmod n)\}, \tag{97}$$

where in line (97) we have used

$$\sum_{\ell \in \Omega_n} \mathbb{1}\{k \in S(\ell)\}\mathbb{1}\{j \in S(\ell)\} = \mathbb{1}\{k \in S(j \bmod n)\} \tag{98}$$

for $k, j \in \Omega_d$ and $\ell \in \Omega_n$, which follows from similar reasoning as Eq. (84). And so, writing

$$(\Sigma_\phi^{1/2}(\mathbb{I}_d - X^T(XX^T)^{-1}X)\boldsymbol{\alpha}_0)_\ell \tag{99}$$

$$= \sum_{j \in \Omega_d} \Bigg[\delta_\ell^j \hat{g}_j \mathbb{1}\{j \in \Omega_{n_0}\}$$

$$- \sum_{k \in \Omega_d} \sqrt{\nu_\ell} \delta_j^\ell \left(\sum_{m \in S(j \bmod n)} \nu_m \right)^{-1} \hat{g}_k \sqrt{\nu_j} \mathbb{1}\{k \in S(j \bmod n)\}\mathbb{1}\{k \in \Omega_{n_0}\} \Bigg] \tag{100}$$

$$= \left(\sum_{j \in \Omega_{n_0}} \delta_\ell^j \hat{g}_j - \sum_{k \in \Omega_{n_0}} \nu_\ell \left(\sum_{m \in S(\ell \bmod n)} \nu_m \right)^{-1} \hat{g}_\ell \mathbb{1}\{k \in S(\ell \bmod n)\} \right) \tag{101}$$

$$= \left(\mathbb{1}\{\ell \in \Omega_{n_0}\}\hat{g}_j - \sum_{k \in \Omega_{n_0}} \nu_\ell \left(\sum_{m \in S(k)} \nu_m \right)^{-1} \hat{g}_k \mathbb{1}\{\ell \in S(k)\} \right). \tag{102}$$

Letting $Q_k := \sum_{m \in S(k)} \nu_m$ the bias term of the error evaluates to

$$\text{BIAS}^2 = \sum_{\ell \in \Omega_d} |(\Sigma_{\mathbf{x}}^{1/2}(\mathbb{I}_d - X^T(XX^T)^{-1}X)\boldsymbol{\alpha}_0)_\ell|^2 \tag{103}$$

$$= \sum_{\ell \in \Omega_d} \left| \mathbb{1}\{\ell \in \Omega_{n_0}\}\hat{g}_\ell - \sum_{k \in \Omega_{n_0}} \nu_\ell Q_k^{-1}\hat{g}_k \mathbb{1}\{\ell \in S(k)\} \right|^2 \tag{104}$$

$$= \sum_{\ell \in \Omega_d} \Bigg(\mathbb{1}\{\ell \in \Omega_{n_0}\}|\hat{g}_\ell|^2 - \sum_{k \in \Omega_{n_0}} \mathbb{1}\{\ell \in \Omega_{n_0}\}\hat{g}_\ell \hat{g}_k^* \nu_\ell Q_k^{-1}\mathbb{1}\{\ell \in S(k)\}$$

$$- \sum_{k \in \Omega_{n_0}} \mathbb{1}\{\ell \in \Omega_{n_0}\}\hat{g}_\ell^* \hat{g}_k \nu_\ell Q_k^{-1}\mathbb{1}\{\ell \in S(k)\}$$

$$+ \sum_{j,k \in \Omega_{n_0}} \nu_\ell^2 Q_j^{-1} Q_k^{-1}\hat{g}_j^* \hat{g}_k \mathbb{1}\{\ell \in S(j)\}\mathbb{1}\{\ell \in S(k)\} \Bigg) \tag{105}$$

$$= \left(\sum_{k \in \Omega_n} |\hat{g}_k|^2 - 2\sum_{k \in \Omega_{n_0}} \nu_k |\hat{g}_k|^2 Q_k^{-1} + \sum_{k \in \Omega_{n_0}} \sum_{\ell \in S(k)} \nu_\ell^2 Q_k^{-2}|\hat{g}_k|^2 \right) \tag{106}$$

$$= \sum_{k \in \Omega_{n_0}} \frac{|\hat{g}_k|^2}{Q_k^2} \left((Q_k - \nu_k)^2 + \sum_{\ell \in S(k)\backslash k} \nu_\ell^2 \right) \tag{107}$$

$$= \sum_{k \in \Omega_{n_0}} |\hat{g}_k|^2 \frac{\left(\sum_{\ell \in S(k)\backslash k} \nu_\ell\right)^2 + \sum_{\ell \in S(k)\backslash k} \nu_\ell^2}{\left(\sum_{\ell \in S(k)} \nu_\ell\right)^2}. \tag{108}$$

With Eqs. (88) and (108) we have recovered a closed form expression for generalization error in terms of feature weights ν_k, the target function Fourier coefficients $\hat{g}_k$, and the

noise variance σ^2. This was possible in part because of the orthonormality of the Fourier features $\phi(x)_k$ and the choice to sample x_j uniformly on $[0,1]$, resulting in diagonal Σ_ϕ and XX^T respectively. This simplicity is more advantageous for studying scenarios where benign overfitting may exist compared to prior works [6, 59]. We will now analyze choices of feature weights ν_k that may give rise to benign overfitting for overparameterized weighted Fourier Features models.

We remark that the results of this section may also be derived using the methods of Ref. [9], though we have opted here to use a language more reminiscent of linear regression to highlight connections between the analysis of weighted Fourier features models and ordinary least squares errors.

A.2 Solutions for special cases

A.2.1 Noise-only minimum norm estimator

We now derive the cases considered in Sec. 2.2 of the main text. We first consider when the target function is given by $g = 0$ and we attempt to predict on a pure noise signal

$$y_j = \epsilon_j, \tag{109}$$

with $\mathbb{E}[\epsilon^2] = \sigma^2$. We can recover an unaliased, "simple" fit to this pure noise signal by reconstructing $\boldsymbol{\epsilon}$ from the DFT using Eq. (52):

$$\epsilon_j = \sum_{k \in \Omega_n} \hat{\epsilon}_k e^{i2\pi jk/n}. \tag{110}$$

By setting all weights $\nu_k = 1$, an immediate choice for an interpolating f with access to frequencies in Ω_d is found by setting

$$\alpha_k = \frac{\hat{\epsilon}_{k \bmod n}}{n(m+1)}, \tag{111}$$

where we have assumed $d = n(m+1)$. Eq. (111) is the minimum ℓ_2-norm estimator and evenly distributes the weight of each Fourier coefficient $\hat{\epsilon}_k$ over m aliases frequencies in in $S(k)$. The effect of higher-frequency aliases is to reduce the f to near-zero everywhere except at the interpolated training data. We can directly compute the generalization error of the interpolating estimator as

$$\mathbb{E}_{x,\mathbf{y}}|f(x) - g(x)|^2 = \mathbb{E}_x |f(x)|^2 \tag{112}$$

$$= \mathbb{E}_{\mathbf{y}} \left| \sum_{k=0}^{d-1} \alpha_k e^{-i2\pi kx} \right|^2 \tag{113}$$

$$= \mathbb{E}_{\mathbf{y}} \sum_{k=0}^{d-1} \left| \frac{\hat{\epsilon}_{k \bmod n}}{n(m+1)} \right|^2 \tag{114}$$

$$= \mathbb{E}_{\mathbf{y}} \frac{n(m+1)\|\hat{\boldsymbol{\epsilon}}\|_2^2}{n^2(m+1)^2} \tag{115}$$

$$= \frac{\sigma^2}{n(m+1)}. \tag{116}$$

Using $d = n(m+1)$ as the dimensionality of the feature space, we have recovered the lower bound scaling for overparameterized models derived in Ref. [9]. In line (113) we have used the independence of ϵ from x_k and y_k and in line (114) we have used orthonormality

of Fourier basis functions on $[0, 1]$. line (116) uses Parseval's relation for the Fourier coefficients, namely:

$$\sum_{k \in \Omega_n} |\epsilon_k|^2 = n \sum_{k \in \Omega_n} |\hat{\epsilon}_k|^2, \tag{117}$$

which implies $\mathbb{E}\|\hat{\boldsymbol{\epsilon}}\|_2^2 = n^{-1}\mathbb{E}\|\boldsymbol{\epsilon}\|_2^2 = \sigma^2$. Figure 2a of the main text shows the effect of the number of cohorts m on the behavior of f, which interpolates a pure noise signal with very little bias. As the number of cohorts increases, the function deviates very little away from the true "signal" $y = 0$, and becomes very "spiky" in the vicinity of noise

A.2.2 Signal-only minimum norm estimator

Now we study the opposite situation in which the pure tone is noiseless, and aliases in the spectrum of f interfere in prediction of f. In this case, we set $\sigma = 0$ and interpolate target labels

$$y_j = \hat{g}_p e^{i2\pi p x_j}, \tag{118}$$

with $-n/2 < p < n/2$. When we set $d = n(m+1)$ and predict on y, there are exactly $|S(p)| - 1 = m$ aliases for the target function with frequency p. We again assume all feature weights are equal, $\nu_k = 1$, and by orthonormality of Fourier basis functions, only the components of f with frequency in $S(p)$ are retained:

$$f(x) = \alpha_p e^{i2\pi p x} + \sum_{k \in S(p)} \alpha_k e^{i2\pi k x}, \tag{119}$$

which will interpolate the training points for any choice of $\boldsymbol{\alpha}$ satisfying

$$\sum_{k \in S(p)} \alpha_k = \hat{g}_p. \tag{120}$$

The choice of trainable weights α_k that satisfy Eq. (120) while minimizing ℓ_2-norm is

$$\alpha_k = \begin{cases} \frac{\hat{g}_p}{m+1}, & k \in S(p), \\ 0, & \text{otherwise.} \end{cases} \tag{121}$$

The problem with minimizing the ℓ_2 norm in this case is that it spreads the true signal into higher frequency aliases: The generalization error of this model is

$$\mathbb{E}_{x,\mathbf{y}}|f(x) - g(x)|^2 = \left|\hat{g}_p \left(\frac{m}{m+1}\right)\right|^2 + m\frac{|\hat{g}_p|^2}{(m+1)^2} = \mathcal{O}(1). \tag{122}$$

We see that this model generally fails to generalize. This poor generalization was described as "signal bleeding" by Ref. [9]: Using n samples, there is no way to distinguish the signal due to an alias of p from the signal due to the true frequency p, so the coefficients α_k become evenly distributed over aliases with very little weight allocated to the true Fourier coefficient $\hat{g}_p$ in the model f. Fig. 2b in the main text shows the effect of "signal bleed" for learning a pure tone in the absence of noise.

A.3 Conditions for Benign overfitting

The behavior of the error of the overparameterized Fourier model (Eq. (19) of the main text) depends on an interplay between noise variance σ^2, signal Fourier coefficients $\hat{g}_k$, feature weights ν_k, and the size of the model d. A desirable property of the feature weights

ν_k is that they should result in a model that both interpolates the sampled data while also achieving good generalization error in the limit $n \to \infty$. For our purposes we will consider cases where $\lim_{n \to \infty} L(f^{opt}) = 0$, though this condition could be relaxed to allow for more interesting or natural choices of ν_k. We now analyze the error arising from a simple weighting scheme to demonstrate benign overfitting using the overparameterized Fourier models discussed in this work.

A.3.1 A simple demonstration of benign overfitting

Here we demonstrate a simple example of benign overfitting when the feature weights ν_k are chosen with direct knowledge of the spectrum Ω_{n_0} of g. For some $n_0 < n < d$, fix $c \in (0, 1)$ and use the feature weights given as

$$\nu_k = \begin{cases} \frac{c}{n_0}, & k \in \Omega_{n_0}, \\ \frac{1-c}{d-n_0}, & \text{otherwise}, \end{cases} \tag{123}$$

for all $k \in \Omega_d$. For simplicity, suppose $d = n(m+1)$ such that $|S(k)| = m+1$ for all $k \in \Omega_n$. Defining the *signal power* as

$$P := \sum_{k \in \Omega_{n_0}} |\hat{g}_k|^2, \tag{124}$$

we can directly evaluate VAR of Eq. (88) and BIAS^2 of Eq. (108):

$$\text{VAR} \to \frac{\sigma^2}{n} \left(n_0 \frac{\left(\frac{c}{n_0}\right)^2 + m \left(\frac{1-c}{d-n_0}\right)^2}{\left(\frac{c}{n_0} + m\frac{1-c}{d-n_0}\right)^2} + \frac{n-n_0}{m+1} \right). \tag{125}$$

$$\text{BIAS}^2 \to \frac{P n_0^2 (m+1) m}{\left(\frac{c}{1-c}(d-n_0) + mn_0\right)^2}. \tag{126}$$

Fixing n_0, we can bound generalization error in the asymptotic limit $n \to \infty$ as:

$$\text{VAR} = \mathcal{O}\left(\frac{1}{n} + \frac{n}{d}\right) \tag{127}$$

$$\text{BIAS}^2 = \mathcal{O}\left(\frac{1}{n^2}\right). \tag{128}$$

Therefore, by setting $d = \omega(n)$ the model perfectly interpolates the training data and also achieves vanishing generalization error in the limit of large number of data. A similar example was considered in Ref. [9], though a rigorous error analysis (and relationship to benign overfitting) was not considered there. This benign overfitting behavior is entirely due to the feature weights of Eq. (123): As $d, n \to \infty$ with $d = \omega(n)$, the feature weights are concentrated on Ω_{n_0} (suppressing BIAS) while becoming increasingly small and evenly distributed over all aliases of Ω_{n_0} (suppressing VAR).

A.3.2 More general conditions for benign overfitting

We have derived closed-form solutions for the BIAS^2 and VAR terms that determine the total generalization error of an interpolating model f and in the previous section provided a concrete example of a model that achieves benign overfitting in this setting. We now discuss conditions under which a more general choice of model can exhibit benign overfitting.

We begin by showing that the variance of Eq. (88) splits naturally into an error due to a (simple) prediction component and a (spiky, complex) interpolating component. Following Refs. [4, 6], we will split the variance into components corresponding to eigenspaces of Σ_ϕ with large and small eigenvalues respectively. Let $S^{\leq p}$ denote the set indices for the largest p eigenvalues of Σ_ϕ (i.e., the largest p values of ν_k), and $S^{>p} = [d]\backslash S^{\leq p}$ be its complement. Define $P^{\leq p} : \mathbb{R}^d \to \mathbb{R}^d$ as the projector onto the the subspace of $\mathbb{R}^d$ spanned by basis vectors labelled by indices in $S^{\leq p}$ (and $P^{>p} = \mathbb{I}_d - P^{\leq p}$ is defined analogously). Then letting $(\boldsymbol{\nu})_k = \nu_k$ be the vector of feature weights and assuming $p \leq n$, we may rewrite the variance of Eq. (88) as

$$\text{VAR} = \frac{\sigma^2}{n} \sum_{k \in \Omega_n} \frac{\sum_{\ell \in S(k)} \nu_\ell^2}{\left(\sum_{\ell \in S(k)} \nu_\ell\right)^2} \tag{129}$$

$$= \frac{\sigma^2}{n} \sum_{k \in \Omega_n} \frac{\|P_{S(k)}\boldsymbol{\nu}\|_2^2}{\|P_{S(k)}\boldsymbol{\nu}\|_1^2} \tag{130}$$

$$= \frac{\sigma^2}{n} \sum_{k \in \Omega_n} \frac{\|P^{\leq p}P_{S(k)}\boldsymbol{\nu} + P^{>p}P_{S(k)}\boldsymbol{\nu}\|_2^2}{\|P_{S(k)}\boldsymbol{\nu}\|_1^2} \tag{131}$$

$$\leq \frac{\sigma^2}{n} \left(\sum_{k \in \Omega_n} \frac{\|P^{\leq p}P_{S(k)}\boldsymbol{\nu}\|_2^2}{\|P^{\leq p}P_{S(k)}\boldsymbol{\nu}\|_1^2} + \sum_{k \in \Omega_n} \frac{\|P^{>p}P_{S(k)}\boldsymbol{\nu}\|_2^2}{\|P^{>p}P_{S(k)}\boldsymbol{\nu}\|_1^2} \right) \tag{132}$$

$$\leq \sigma^2 \left[\frac{p}{n} + \frac{1}{n} \sum_{k \in \Omega_n} \frac{1}{R_p^{(k)}(\Sigma_\phi)} \right], \tag{133}$$

where we have used $P^{\leq p}P_{S(k)}\hat{e}_j = \delta_j^k \mathbb{1}\{j \leq p\}$ for $p \leq n$ and we have introduced an effective rank for the alias cohort of k,

$$R_p^{(k)}(\Sigma_\phi) = \frac{\left(\sum_{\ell \in S(k) \cap S^{>p}} \nu_\ell\right)^2}{\sum_{\ell \in S(k) \cap S^{>p}} \nu_\ell^2}. \tag{134}$$

Since $p = n_0$ is a relevant choice for our problem setup, we define $R^{(k)} := R_{n_0}^{(k)}(\Sigma_\phi)$ and focus on the bound

$$\text{VAR} \leq \sigma^2 \left[\frac{n_0}{n} + \frac{1}{n} \sum_{k \in \Omega_n} \frac{1}{R^{(k)}} \right], \tag{135}$$

The first term of the decomposition of Eq. (135) corresponds to the variance of a (non-interpolating) model with access to n_0 Fourier modes, while the second term corresponds to excess error introduced by high frequency components of f. Given a sequence of experiments with increasing n (while g and σ^2 remain fixed), we would like to understand the choices of feature weights ν_k for which VAR vanishes as $n \to \infty$. Given that $|\{k \in \Omega_n\}| = n$, a sufficient condition for a sequence of feature weight distributions to be benign is that $R^{(k)} = \omega(1)$ for all k while a necessary condition is that there is no k for which $R^{(k)} = \mathcal{O}(1/n)$. These conditions are not difficult to satisfy: Intuitively they require only that Σ_ϕ changes with increasing n in such a way that the values of ν_k in $\Omega_n\backslash\Omega_{n_0}$ continue "flatten out" as n increases. This is precisely the behavior engineered in the example of Sec. A.3.1.

We now proceed to bound the bias term of Eq. (95). Observe that $P^\perp := (\mathbb{I}_d - X^T(XX^T)^{-1}X)$ is a projector onto the subspace of $\mathbb{R}^d$ orthogonal to the rows of X,

therefore satisfying $\|P^\perp\| \le 1$ and $P^\perp X^T X = 0$. Then the BIAS^2 is bounded as

$$\mathrm{BIAS}^2 = \|\Sigma_\phi^{1/2} P^\perp \boldsymbol{\alpha}_0\|_2^2 \tag{136}$$

$$= \|(\Sigma_\phi - X^T X)^{1/2} P^\perp \boldsymbol{\alpha}_0\|_2^2 \tag{137}$$

$$\le \|\Sigma_\phi - X^T X\|_\infty \|\boldsymbol{\alpha}_0\|_2^2. \tag{138}$$

The term $X^T X$ can be interpreted as a finite-sample estimator for Σ_ϕ in the sense that $X^T X = n^{-1} \Phi^T \Phi = \Sigma_\phi$ for $n = d$. However, we cannot apply standard results on the convergence of sample covariance matrices (e.g. Ref. [60]) since the uniform spacing requirement for training data x violates the standard assumption of i.i.d. input data. To proceed, we will make a number of simplifying assumptions about the feature weights. First, to control for the possibility that large feature weights ν_k concentrate within a specific $S(k)$ we will assume that feature weights corresponding to any set of alias frequencies of Ω_n are close to their average. Letting $d = (m+1)n$, we define

$$\eta_\ell = \frac{1}{n} \sum_{k \in \Omega_n} \nu_{k+n\ell}, \tag{139}$$

for $\ell = 0, 1, \ldots, m$. We will impose that $|\nu_{k+n\ell} - \eta_\ell| \le t$ for all $k \in \Omega_n$, $\ell = 1, \ldots, m$, for some positive number t. We further assume normalization, $\sum_{k \in \Omega_d} \nu_k = 1 = n \sum_{\ell=0}^m \eta_\ell$. Under these assumptions we can bound the first term of (138) as

$$\|\Sigma_\phi - X^T X\|_\infty \le \max_j \nu_j^{1/2} \left(\sum_{k \in S(j \bmod n) \setminus j} \nu_k^{1/2} \right) \tag{140}$$

$$\le (m-1)^{1/2} \max_j \nu_j^{1/2} \left(\sum_{k \in S(j \bmod n) \setminus j} \nu_k \right)^{1/2} \tag{141}$$

$$\le m^{1/2} \max_j \nu_j^{1/2} \left(\sum_{\ell=1}^m (\eta_\ell + t) \right)^{1/2} \tag{142}$$

$$\le \left(\frac{m}{n} + m^2 t \right)^{1/2}, \tag{143}$$

where in line (140) we have used Gershgorin circles. Meanwhile, defining $\zeta := \left(\sum_{k \in \Omega_{n_0}} \nu_k \right)$, by Cauchy-Schwarz we have that

$$\|\boldsymbol{\alpha}_0\|^2 \le P \zeta^{-2}, \tag{144}$$

where P is the signal power of Eq. (124). A necessary condition for producing benign overfitting in overparameterized Fourier models is that ζ remains relatively large as $n, d \to \infty$. If this is accomplished, then a small enough t guarantees that all feature weights associated with frequencies $k \in \Omega_d \setminus \Omega_{n_0}$ will be uniformly suppressed. For instance, if ζ is lower bounded as a constant while $t = 0$ then combining Eqs. (143) and (144) yields a bound of

$$\mathrm{BIAS}^2 = \mathcal{O}\left(\frac{d^{1/2}}{n} \right). \tag{145}$$

Although the analysis of Sec. A.3.1 yields a significantly tighter bound, this demonstrates that the mechanisms behind that simple demonstration of benign overfitting are somewhat generic. In particular, normalization and lower bounded support of the feature weights on Ω_{n_0} is almost sufficient to control the bias term of the generalization error.

B Solution for the quantum overparameterized model

We now derive the solution to the minimum-norm interpolating quantum model,

$$M^{opt} = \underset{\substack{M=M^\dagger \\ f(x_j)=y_j}}{\arg\min} \|M\|_F, \tag{146}$$

$$f(x) = \sum_{k\in\Omega} e^{i2\pi kx} \sum_{\ell,m\in R(k)} \gamma_\ell M_{m\ell}\gamma_m^*. \tag{147}$$

We will use the following notation and definitions:

$$a_k = -\left\lfloor \frac{k + \frac{|\Omega|-n}{2}}{n} \right\rfloor, \tag{148}$$

$$b_k = \left\lfloor \frac{-k + \frac{|\Omega|+n}{2} - 1}{n} \right\rfloor, \tag{149}$$

$$S_k = \{k - a(k)n, \ldots, k, k+n, \ldots, k+b(k)n\} \tag{150}$$

$$= \{k + pn, \quad p = -a(k), \ldots, b(k)\}, \tag{151}$$

$$R(S_k) := \bigcup_{j\in S_k} R(j). \tag{152}$$

Here, a_k and b_k characterize the number positive and negative frequency aliases of k appearing in Ω (i.e., $k + n\ell \in \Omega$ for all $a_k \le \ell \le b_k$) assuming that $|\Omega|$ is odd, a requirement for any quantum model. Let $L(d)$ be the space of linear operators acting on $d \times d$ matrices. Define the linear operator $P_k : L(d) \to L(d)$ for $k \in [n]$ as

$$P_k(X) = \sum_{\ell,m\in R(k)} \langle\ell|X|m\rangle|\ell\rangle\langle m| \tag{153}$$

for any $X \in L(d)$. Importantly, P_k is not necessarily Hermitian preserving. Denoting $\Gamma := |\Gamma\rangle\langle\Gamma|$ for brevity, we may rewrite the Fourier coefficients of $f(x)$ as

$$\langle P_k(M), P_k(\Gamma)\rangle = \mathrm{Tr}\left(P_k(M)^\dagger P_k(\Gamma)\right) \tag{154}$$

$$= \mathrm{Tr}\left(\sum_{\ell,m\in R(k)} M_{\ell m}^* |m\rangle\langle\ell| \sum_{i,j\in R(k)} \gamma_i\gamma_j^*|i\rangle\langle j|\right) \tag{155}$$

$$= \sum_{\ell,m\in R(k)} \gamma_\ell^* M_{m\ell}\gamma_m, \tag{156}$$

where in the last line we have used hermiticity of M. Applying $f(x_j) = y_j \;\forall\, j \in [n]$ and substituting into Eq. (147) we find the interpolation condition

$$\hat{y}_p = \sum_{q=a_k}^{b_k} \langle P_{p+nq}(M), P_{p+nq}(\Gamma)\rangle \tag{157}$$

$$= \langle P_{S_p}(M), P_{S_p}(\Gamma)\rangle, \qquad \forall\, p \in [n], \tag{158}$$

where $P_{S_p} := \sum_{q=a_k}^{b_k} P_{p+nq}$. The equality follows from

$$\langle P_j(X), P_k(Y)\rangle = \langle P_j(X), P_j(Y)\rangle\delta_j^k \tag{159}$$

due to the fact that $R(j)$ and $R(k)$ are disjoint sets for any $j \neq k$. Following the technique of Ref. [9] we apply the Cauchy-Schwarz inequality to find

$$\|P_{S_p}(M)\|_F \geq \frac{|\langle P_{S_p}(M), P_{S_p}(\Gamma)\rangle|}{\|P_{S_p}(\Gamma)\|_F}, \tag{160}$$

with equality if and only if $P_{S_p}(M)$ is proportional to $P_{S_p}(\Gamma)$. Saturating this lower bound by setting $M_{\ell m} = c\gamma_\ell \gamma_m^*$ and solving for the proportionality c constant using Eq. (160), we find

$$\hat{y}_p = \langle cP_{S_p}(\Gamma), P_{S_p}(\Gamma)\rangle = \sum_{\ell,m \in R(S_p)} c^* |\gamma_\ell|^2 |\gamma_m|^2. \tag{161}$$

This indicates an additional requirement for interpolation that $\gamma_\ell, \gamma_m > 0$ for some pair $(\ell, m) \in R(k)$ whenever $\tilde{y}_k \neq 0$ and so for simplicity we will require that $\gamma_\ell > 0$ for all $\ell = 1, \ldots, d$. Within each set of indices $R(S_k)$, the elements of the optimal observable are defined piecewise with respect to that partition R:

$$(\ell, m) \in R(S_k) \Rightarrow M_{\ell m} = \hat{y}_k^* \frac{\gamma_\ell \gamma_m^*}{\sum_{i,j \in R(S_k)} |\gamma_i|^2 |\gamma_j|^2}. \tag{162}$$

Minimization of $\|M\|_F$ subject to the interpolation constraint is equivalent to minimization of $\|P_{p+nq}(M)\|_F$ for all $q \in [a_k, b_k]$, $k \in \Omega$, and so solving the constrained optimization over all distinct subspaces in $\bigcup_{k=0}^{n-1} R(S_k) = \{0,1\}^n \times \{0,1\}^n$ we recover the optimal observable

$$M_{opt} = \arg\min_{f(x_j)=y_j \, \forall j} \|M\|_F = \sum_{k \in \Omega_n} \hat{y}_k^* \sum_{\ell,m \in R(S_k)} \frac{\gamma_\ell \gamma_m^*}{\sum_{i,j \in R(S_k)} |\gamma_i|^2 |\gamma_j|^2} |\ell\rangle\langle m|. \tag{163}$$

We now verify that this matrix is Hermitian and therefore a valid observable. We will use the following:

$$\hat{y}_k = \hat{y}_{-k}^*, \tag{164}$$

$$R(-k) = \{(m, \ell) \text{ for all } (\ell, m) \in R(k)\} := R(k)^T. \tag{165}$$

Eq. (164) follows from our assumption that $y_j \in \mathbb{R} \, \forall j \in [n]$. And so

$$M_{\ell m} = \hat{y}_k^* \frac{\gamma_\ell \gamma_m^*}{\sum_{i,j \in R(S_k)} |\gamma_i|^2 |\gamma_j|^2} \quad \forall (\ell, m) \in R(S_k) \tag{166}$$

$$= \hat{y}_{-k} \frac{\gamma_\ell \gamma_m^*}{\sum_{j,i \in R(S_{-k})^T} |\gamma_i|^2 |\gamma_j|^2} \quad \forall (\ell, m) \in R(S_{-k})^T \tag{167}$$

$$= \left(\hat{y}_{-k}^* \frac{\gamma_m \gamma_\ell^*}{\sum_{i,j \in R(S_{-k})} |\gamma_i|^2 |\gamma_j|^2}\right)^* \quad \forall (m, \ell) \in R(S_{-k}) \tag{168}$$

$$= M_{m\ell}^*. \tag{169}$$

In line (167) we have used Eq. (164), while in line (169) we have observed that Eq. (162) holds only with respect to any fixed partition; $M_{\ell m}$ and $M_{m\ell}$ must be computed with respect to distinct partitions. The optimal model may now be rewritten in terms of base frequencies as

$$f^{opt}(x) = \sum_{k \in \Omega_n} \hat{y}_k \sum_{\ell=a_k}^{b_k} e^{i2\pi(k+n\ell)x} \frac{\sum_{i,j \in R(k+n\ell)} |\gamma_i|^2 |\gamma_j|^2}{\sum_{r=a_k}^{b_k} \sum_{c,d \in R(k+nr)} |\gamma_c|^2 |\gamma_d|^2}. \tag{170}$$

Recall that the optimal classical model derived in Eq. (65) is given by

$$f(x) = \sum_{k \in \Omega_n} \hat{y}_k \sum_{\ell=a_k}^{b_k} e^{i2\pi(k+\ell n)x} \frac{\nu_{k+\ell n}}{\sum_{r=a_k}^{b_k} \nu_{k+nr}}. \tag{171}$$

Then despite Eq. (147) not having a clear decomposition into scalar feature weights and trainable weights, we can identify the feature weights of the optimized quantum model as

$$\nu_k^{opt} := \sum_{i,j \in R(k)} |\gamma_i|^2 |\gamma_j|^2, \tag{172}$$

which recovers the same form of the optimal classical model of Eq. (65). This means that the behavior of the feature weights of the (optimal) general quantum model are strongly controlled by the degeneracy sets $R(k)$. The generalization error of the quantum model is also described by Eq. (19) of the main text under the identification of Eq. (172), and therefore exhibits a tradeoff that is predominantly controlled by the degeneracies $R(k)$ of the data-encoding Hamiltonian.

We can now substitute $\gamma_\ell = 1/\sqrt{d}$ to recover the optimal observable for the simplified model derived in Sec 3.1 of the main text using other means, namely

$$M_{m\ell}^{opt} = \hat{y}_k \frac{d}{\sum_{r=a_k}^{b_k} |R(k+nr)|}, \tag{173}$$

$$f(x) = \sum_{k \in \Omega_n} \hat{y}_k \sum_{q=a_k}^{b_k} e^{i2\pi(k+nq)x} \frac{|R(k+nq)|}{\sum_{r=a_k}^{b_k} |R(k+nr)|}, \tag{174}$$

$$\nu_k = \frac{|R(k)|}{d^2}. \tag{175}$$

B.1 Computing feature weights of typical quantum models

In Sec 3.1 we introduced a simple model with a uniform amplitude state $|\Gamma\rangle$ as input and demonstrated that the feature weights of this simple quantum model are completely determined by the sets $R(k)$ induced by the encoding strategy. We now wish to extend the intuition that the behavior of the optimized general quantum model is strongly influenced by the distribution of the degeneracies $|R(k)|$. We do so by evaluating the optimal quantum models with respect to an "average" state preparation unitary U. We can compute the average value of $|\gamma_i|^2 |\gamma_j|^2$ for U sampled uniformly from the Haar distribution using standard results from the literature [61][3]:

$$\mathbb{E}_{U \sim \mathrm{U}(d)} |\gamma_i|^2 |\gamma_j|^2 = \mathbb{E}_{U \sim \mathrm{U}(d)} U_{i0} U_{i0}^* U_{j0} U_{j0}^* = \frac{\delta_i^j + 1}{d(d+1)}. \tag{176}$$

Since $(i,i) \in R(0)$ we can then compute the feature weights of the optimal model according to[4]

$$\mathbb{E}_{U \sim \mathrm{U}(d)} \nu_k^{opt} = \mathbb{E}_{U \sim \mathrm{U}(d)} \sum_{i,j \in R(k)} |\gamma_i|^2 |\gamma_j|^2 = \frac{|R(k)|}{d(d+1)} + \delta_0^k \frac{1}{d+1}. \tag{177}$$

---

[3]Note that since ν_k^{opt} is invariant with respect to $\gamma_i \to e^{i\phi}\gamma_i$ a spherical measure would suffice here.

[4]It is implied that we compute the optimal ν with respect to each distinct U sampled independently and uniformly with respect to the Haar measure – without optimizing M with respect to each U we would find the trivial result

$$\mathbb{E}_{U \sim \mathrm{U}(d)} f(x) = \mathbb{E}_{U \sim \mathrm{U}(d)} \mathrm{Tr}\,(M\rho(x)) = d^{-1} \mathrm{Tr}\,(M)$$

.

From Eq. (177) we see that the feature weights of a quantum model optimized with respect to random U are completely determined by the degeneracies $R(k)$. This expected value is useful but does not fully characterize the behavior of an encoding strategy. To demonstrate that this average behavior is meaningful, we would further like to verify that the feature weights corresponding to a random U concentrate around the mean of Eq. (177). We characterize this by computing the variance

$$\text{Var}(\nu_k) = \mathop{\mathbb{E}}_{U\sim\text{U}(d)}\left(\nu_k^2\right) - \left(\mathop{\mathbb{E}}_{U\sim\text{U}(d)}\nu_k\right)^2,\tag{178}$$

where we have dropped the superscript on ν_k^{opt} for brevity. This computation requires significantly more counting arguments dependent on the structure of $R(k)$. When $k \neq 0$, we identify cases for which $i \neq j$ whenever $(i,j) \in R(k)$:

$$\nu_k^2 = \sum_{i,j\in R(k)}\sum_{\ell,m\in R(k)}|\gamma_i|^2|\gamma_j|^2|\gamma_\ell|^2|\gamma_m|^2\tag{179}$$

$$= \sum_{i,j\in R(k)}\left(|\gamma_i|^4|\gamma_j|^4 + \sum_{\substack{\ell,m\in R(k)\\\ell\neq j,m=i}}|\gamma_i|^4|\gamma_j|^2|\gamma_\ell|^2\right.$$

$$\left. + \sum_{\substack{m,\ell\in R(k)\\\ell=j,m\neq i}}|\gamma_i|^2|\gamma_j|^4|\gamma_m|^2 + \sum_{\substack{m,\ell\in R(k)\\\ell\neq j,m\neq i}}|\gamma_i|^2|\gamma_j|^2|\gamma_\ell|^2|\gamma_m|^2\right).\tag{180}$$

The expected values for of these terms are evaluated using the observation that the vector $(|\gamma_1|^2,\ldots,|\gamma_s|^2)$ is distributed uniformly on the d-simplex leading to simple expressions for the following expected values [61]:

$$\mathop{\mathbb{E}}_{U\sim\text{U}(d)}|\gamma_p|^4|\gamma_q|^4 = 4/D,\tag{181}$$

$$\mathop{\mathbb{E}}_{U\sim\text{U}(d)}|\gamma_p|^4|\gamma_q|^2|\gamma_r|^2 = 2/D,\tag{182}$$

$$\mathop{\mathbb{E}}_{U\sim\text{U}(d)}|\gamma_p|^2|\gamma_q|^2|\gamma_r|^2|\gamma_s|^2 = 1/D,\tag{183}$$

where $p,q,r,s = 1,\ldots,d$ and $p \neq q \neq r \neq s$, and $D := (d+3)(d+2)(d+1)d$. In computing the expected value of Eq. (180), by linearity the terms of each sum will become a constant and only the number of items in each sum will be relevant. The first sum of Eq. (180) contains $|R(k)|$ many terms and the total number of terms is $|R(k)|^2$, and so we only need compute the number of elements in the two middle sums of Eq. (180) (which contain an equal number of terms due to the symmetry $R(k) = R(-k)^T$). These computations may be carried out by brute-force combinatorics and are summarized in Table 2 for a few of the models studied in this work.

For the Binary encoding strategy we compute

$$\mathbb{E}(\nu_k^2) = \begin{cases} \frac{5d-7k+(d-k)^2}{D}, & 0 < 2k \leq d \\ \frac{3(d-k)+(d-k)^2}{D}, & 2k > d, \end{cases}\tag{184}$$

Encoding strategy	$\|\{(i,j,\ell,m) : m \neq i, \ell = j, (i,j), (\ell,m) \in R(k)\}\|$	$\mathrm{Var}(\nu_k^{opt})$
Hamming	$\sum_{p=0}^{n_q-2k} \binom{n_q}{p} \binom{n_q}{p+|k|} \binom{n_q}{p+2|k|}$	-
Binary	$\max(2^{n_q} - 2|k|, 0)$	$O\left(2^{-3n_q}\right)$
Golomb	0	$\mathcal{O}(d^{-4})$

Table 2: The size of the subset of $R(k) \times R(k)$ with a single repeated index computed for different encoding strategies and the corresponding scaling of the variance of feature weights (all entries assume $k \neq 0$). The Hamming encoding strategy computation works as follows: Each bitstring i with weight p (of which there are n-choose-p) is paired with n-choose-$(p+k)$ many bitstrings j with weight $p+k$. Then, taking $\ell = j$ there will be n-choose-$(p+2k)$ many bitstrings m with weight $(p+k)+k = p+2k$. Summing over all such p's where $p + 2k \leq n_q$ yields the desired result, however this computation does not admit a clear closed-form expression and so we have omitted the corresponding scaling of $\mathrm{Var}(\nu_k^{opt})$ for the Binary encoding strategy.

and so

$$\mathrm{Var}(\nu_k) = \left\{ \begin{array}{ll} \dfrac{(5d - 7k)d(d+1) - (d-k)^2(4d+6)}{Dd(d+1)} & \text{if } 0 < 2k \leq d \\[2ex] \dfrac{3(d-k)d(d+1) - (d-k)^2(4d+6)}{Dd(d+1)} & \text{if } 2k > d \end{array} \right\} \leq \mathcal{O}(d^{-3}). \quad (185)$$

Importantly, $\mathbb{E}[\nu_k] \geq \mathcal{O}(d^{-2})$ for the Binary encoding strategy. Taking $d = 2^{n_q}$ corresponding to n_q qubits, then while the mean decays exponentially in the number of qubits n_q, the variance decays exponentially faster. For the Golomb encoding strategy, the calculation is comparatively straightforward, yielding for $k \neq 0$

$$\mathrm{Var}(\nu_k) = \frac{3d^2 - d - 6}{Dd(d+1)} = \mathcal{O}(d^{-4}), \quad (186)$$

with the variance again decaying significantly faster in d than the mean. Figure 6 shows the average and variance of ν_k^{opt} for the general quantum model with U sampled uniformly with respect to the Haar measure. We find tight concentration of ν_k^{opt} around its average in each of these cases

Why is it that the feature weights ν_k^{opt} of the quantum model given by Eq. (172) only appear after deriving the optimal observable? One explanation is that while the classical Fourier Features model utilizes random Fourier features such that components of $\phi(x)$ are mutually orthogonal on $[0,1]$, the quantum model does not. Consider the operator

$$\Sigma := \int_{\mathcal{X}} \mathrm{Vec}\left(\rho(x)\right) \mathrm{Vec}\left(\rho(x)\right)^{\dagger} p(x)dx = \int_{\mathcal{X}} \rho(x) \otimes \rho^T(x) p(x)dx, \quad (187)$$

where $p : \mathcal{X} \to \mathbb{R}$ is the probability density function describing the distribution of data in $\mathcal{X}$ and $\mathrm{Vec} : \mathbb{C}^{d \times d} \to \mathbb{C}^{d^2}$ is the *vectorization map* that acts by stacking transposed rows of a square matrix into a column vector. Σ is analogous to a classical covariance matrix, and here determines the correlation between components of $\rho(x)$ (with the second equality holding when $\rho(x)$ is a pure state for all x). From Eq. (8) describing the classical Fourier features it is straightforward to compute $\mathbb{E}_{x \sim [0,1]}[\phi\phi^{\dagger}] = \mathbb{I}_d$, demonstrating that the classical Fourier features are indeed orthonormal. However, under the identification of the feature vector $\phi(x) \to \mathrm{Vec}\left(\rho(x)\right)$ in the quantum case, the same is not true for the
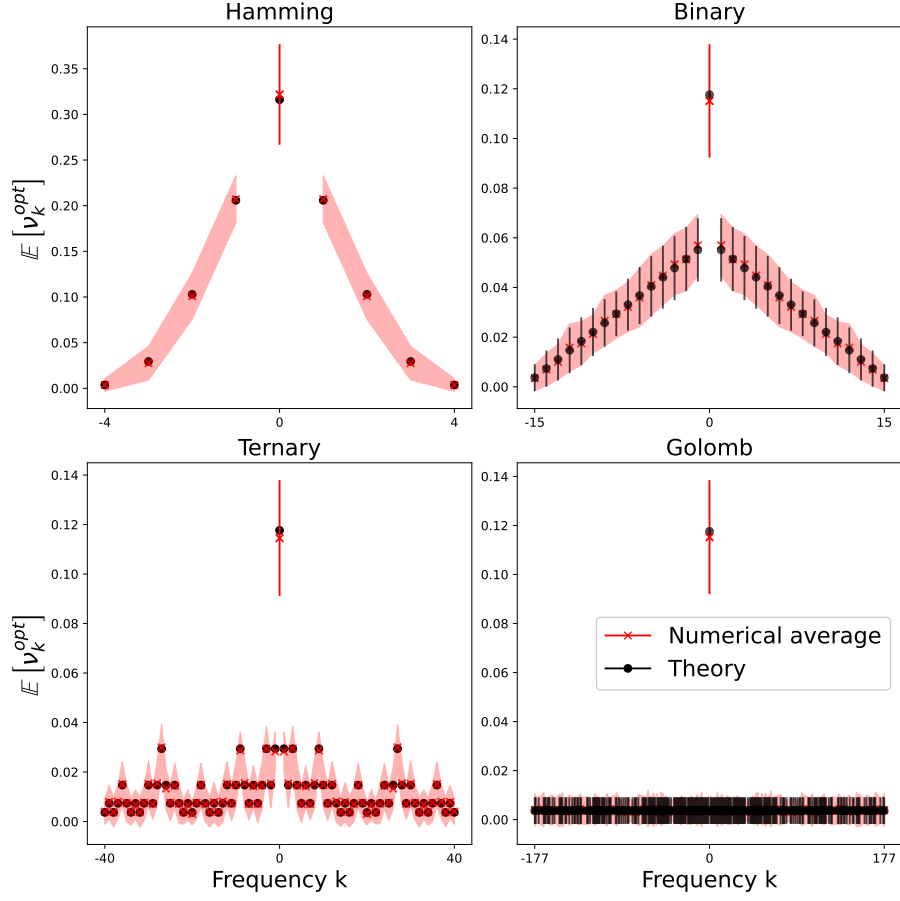
Figure 6: Plotting $\mathbb{E}\left[\nu_k^{opt}\right]$ and $\text{Var}\left[\nu_k^{opt}\right]^{1/2}$ demonstrates tight concentration of the distribution of ν_k^{opt} around its mean with respect to Haar-random U in the general quantum model. For the Binary and Golomb encoding strategies we plot Eqs. (185) and (186) respectively. For all models we also include averages and variances computed by sampling states uniformly from the Haar measure numerically.

quantum model:

$$\Sigma = \int_0^1 \left(\sum_{k \in \Omega} e^{i2\pi kx} \sum_{i,j \in R(k)} \gamma_i \gamma_j^* |ij\rangle \right) \left(\sum_{k' \in \Omega} e^{i2\pi k'x} \sum_{\ell,m \in R(k')} \gamma_\ell^* \gamma_m \langle \ell m| \right) dx \tag{188}$$

$$= \sum_k \sum_{\substack{i,j \in R(k) \\ \ell,m \in R(k)}} \gamma_i \gamma_j^* \gamma_\ell^* \gamma_m |ij\rangle \langle \ell m|. \tag{189}$$

Thus Σ is not diagonal in general, as many components of $S(x)|\Gamma\rangle$ each contribute to a single frequency k. Nor will it be possible in general to construct a quantum feature map via unitary operations that does act as an orthogonal Fourier features vector. As Σ is positive semidefinite by construction, there exists a spectral decomposition $\Sigma = WDW^\dagger$ with diagonal D and $d^2 \times d^2$ unitary W. However the linear operator $\Phi \in L(d)$ acting on d-dimensional states according to $\text{Vec}(\Phi(\rho)) = W^\dagger \text{Vec}(\rho(x))$ will not be unitary in general, and thus it may not be possible to prepare ρ in such a way that the elements of $\text{Vec}(\rho(x))$ consist of orthonormal Fourier features.

B.2 The effect of state preparation on feature weights in the general quantum model

We now discuss how the state preparation unitary U may affect the feature weights ν_k^{opt} in a quantum model, and what choices one can make to construct a U that gives rise to a specific distribution of feature weights ν_k^{opt}.

As an example, we consider the general quantum model using the Hamming encoding strategy and an input state $|\Gamma\rangle$. Computing the feature weights ν_k^{opt} for the Hamming encoding strategy depends on the amplitudes γ_i and γ_j for which the weights of the indices i, j differ by k. We have

$$\nu_k^{opt} = \sum_{i,j \in R(k)} |\gamma_i|^2 |\gamma_j|^2 = \sum_{\substack{i,j: \\ w(j)-w(i)=k}} |\gamma_i|^2 |\gamma_j|^2. \tag{190}$$

We will now show how the feature weight ν_k^{opt} may be computed with respect to a re-balanced input state which distributes each amplitude γ_i among all computational basis states with index having weight $w(i)$. We define this rebalanced state on n qubits as

$$|\Gamma'\rangle = \sum_{i \in \{0,1\}^n} \gamma_i' |i\rangle = \sum_{i=0}^n \sum_{j \in W(i)} \phi_i |j\rangle, \tag{191}$$

$$\phi_i = \frac{1}{\sqrt{|W(i)|}} \left(\sum_{j \in W(i)} |\gamma_j|^2 \right)^{1/2}. \tag{192}$$

where $W(i) = \{j : j \in \{0,1\}^n, w(j) = i\}$ is the set of weight-i indices and $|W(i)| = \binom{n}{i}$.

Observing that the amplitudes of $|\Gamma'\rangle$ satisfy $|\gamma'_i|^2 = \phi^2_{w(i)}$, we can derive for $k \geq 0$

$$\nu^{opt}_k = \sum_{\substack{i,j: \\ w(j)-w(i)=k}} |\gamma_i|^2 |\gamma_j|^2 \tag{193}$$

$$= \sum_{\ell=0}^{n-k} \left(\sum_{i:w(i)=\ell} |\gamma_i|^2 \right) \left(\sum_{j:w(j)=\ell+k} |\gamma_j|^2 \right) \tag{194}$$

$$= \sum_{\ell=0}^{n-k} \left(|W(\ell)| \frac{\sum_{i:w(i)=\ell} |\gamma_i|^2}{|W(\ell)|} \right) \left(|W(\ell+k)| \frac{\sum_{j:w(j)=\ell+k} |\gamma_j|^2}{|W(\ell+k)|} \right) \tag{195}$$

$$= \sum_{\ell=0}^{n-k} |W(\ell)| \phi^2_\ell |W(\ell+k)| \phi^2_{\ell+k} \tag{196}$$

$$= \sum_{\ell=0}^{n-k} \left(\sum_{i:w(i)=\ell} \phi^2_{w(i)} \right) \left(\sum_{j:w(j)=\ell+k} \phi^2_{w(j)} \right) \tag{197}$$

$$= \sum_{\substack{i,j: \\ w(j)-w(i)=k}} |\gamma'_i|^2 |\gamma'_j|^2. \tag{198}$$

Therefore, the feature weights computed from $|\Gamma\rangle = U|0\rangle$ and the rebalanced state $|\Gamma'\rangle$ are identical. We can emphasize the significance of this observation by rewriting $|\Gamma'\rangle$ of Eq. (191) as

$$|\Gamma'\rangle = \sum_{i=0}^n \phi_i |\Phi_i\rangle, \tag{199}$$

where $|\Phi_i\rangle = \sum_{j\in W(i)} |j\rangle$ is an (unnormalized) superposition of bitstrings with weight i. The state $|\Gamma'\rangle$ of Eq. (199) has only $n+1$ real parameters ϕ_i, and is invariant under operations that are restricted to act within the subspace spanned by components of $|\Phi\rangle$. This invariance greatly reduces the class of unitaries U that affect the distribution of feature weights ν^{opt}_k and enables some degree of tuning for these parameters.

B.2.1 Vanishing gradients in preparing feature weights

We briefly remark on the possibility of training feature weights ν^{opt}_k of the general quantum model variationally. For example, one could consider defining a desired distribution of feature weights as $\phi_k \in \mathbb{R}^{|\Omega|}$ and then attempting to tune parameters of U in order to minimize a cost function such as[5]

$$L = \sum_{k\in\Omega} |\nu^{opt}_k - \phi_k|^2. \tag{200}$$

We will now demonstrate that for a sufficiently expressive class of state preparation unitaries U, such an optimization problem will be difficult on average. For simplicity, we follow the original formulation of barren plateaus presented in Ref. [23]: Let U be defined with respect to a set of parameters $\boldsymbol{\theta} \in \mathbb{R}^L$ as

$$U = \prod_{\ell=1}^L U(\theta_\ell) W_\ell \tag{201}$$

---

[5]As ν_k represents a probability distribution, other distance measures such as relative entropy or earth-mover's distance may be considered more appropriate. However using alternative cost functions will not affect the main arguments here.

with $U(\theta_\ell) = \exp(i\theta_\ell V_\ell)$, V_ℓ being a d-dimensional Hermitian operator, and W_ℓ being a d-dimensional unitary operator. Pick a parameter θ_m and define $U_+ = \prod_{\ell=m}^{L} U(\theta_\ell) W_\ell$ and $U_- = \prod_{\ell=1}^{m} U(\theta_\ell) W_\ell$, and observe that

$$\partial_{\theta_m} |\gamma_j|^2 = i\langle 0|U_-^\dagger [V_m, U_+^\dagger |j\rangle\langle j|U_+] U_-|0\rangle. \tag{202}$$

We can compute the derivative of ν_k with respect to θ_m using the chain rule:

$$\frac{\partial \nu_k}{\partial \theta_m} = \sum_{j=1}^{d} \frac{\partial \nu_k}{\partial |\gamma_j|^2} \frac{\partial |\gamma_j|^2}{\partial \theta_m} \tag{203}$$

$$= \sum_{j=1}^{d} \left(\sum_{a,b \in R(k)} \left(|\gamma_a|^2 \delta_b^j + \delta_a^j |\gamma_b|^2 \right) \right) i\langle 0|U_-^\dagger [V_m, U_+^\dagger |j\rangle\langle j|U_+] U_-|0\rangle \tag{204}$$

$$= i \sum_{a,b \in R(k)} \left[\mathrm{Tr}\left(\rho_- H_a \right) \mathrm{Tr}\left(\rho_- [V_m, H_b] \right) + \mathrm{Tr}\left(\rho_- [V_m, H_a] \right) \mathrm{Tr}\left(\rho_- H_b \right) \right] \tag{205}$$

where $\rho_- = U_-|0\rangle\langle 0|U_-^\dagger$ and $H_x = U_+^\dagger |x\rangle\langle x|U_+$, and we used the equality

$$|\gamma_a|^2 = \langle 0|U_-^\dagger U_+^\dagger |a\rangle\langle a|U_+ U_-|0\rangle = \mathrm{Tr}\left(\rho_- H_a \right). \tag{206}$$

We now show that each term in this sum vanishes by following Ref. [23] in letting U_- be sampled from a distribution that forms a unitary 2-design:

$$\mathbb{E}_{U_- \sim U(d)} \mathrm{Tr}\left(\rho_- H_a \right) \mathrm{Tr}\left(\rho_- [V_m, H_b] \right) \tag{207}$$

$$= \mathrm{Tr}\left(\mathbb{E}_{U_- \sim U(d)} [\rho_- \otimes \rho_-] H_a \otimes [V_m, H_b] \right) \tag{208}$$

$$= \frac{1}{d(d+1)} \mathrm{Tr}\left(\left(\mathbb{I}_d + \sum_{i,j=1}^{d} |ij\rangle\langle ji| \right) H_a \otimes [V_m, H_b] \right) \tag{209}$$

$$= \frac{\mathrm{Tr}\left([V_m, H_b] H_a \right)}{d(d+1)} \tag{210}$$

$$= \frac{V_m U_+^\dagger |b\rangle\langle b|U_+ U_+^\dagger |a\rangle\langle a|U_+ - U_+^\dagger |b\rangle\langle b|U_+ V_m U_+^\dagger |a\rangle\langle a|U_+}{d(d+1)} \tag{211}$$

$$= \frac{\delta_a^b \langle a|U_+ V_m U_+^\dagger |b\rangle - \delta_a^b \langle b|U_+ V_m U_+^\dagger |a\rangle}{d(d+1)} \tag{212}$$

$$= 0, \tag{213}$$

where in line (209) we have used a common expression for $\mathbb{E}_{U \sim U(d)}[\rho \otimes \rho]$ in terms of the projector onto the symmetric subspace (e.g. [62]). Substituting this result into Eq. (205) we find

$$\mathbb{E}_{U \sim U(d)} \left(\frac{\partial \nu_k}{\partial \theta_m} \right) = 0 \tag{214}$$

By extension, the gradient of a loss function of the form of Eq. (200) will vanish for expressive enough state-preparation unitaries U, suggesting that solving for a choice of U to induce a specific distribution of feature weights ν_k^{opt} will be infeasible in practice.

C Determining the degeneracy of quantum models

Here we will develop a theoretical framework for manipulating the degeneracy (and therefore feature weights) of quantum models. We will begin with choosing the data-encoding

operator $S(z) = \exp(-iHz)$ to be separable, generated from a Hamiltonian of the form

$$H = \mathbf{r} \cdot \mathbf{Z} = \sum_{j=0}^{n-1} r_j Z^{(j)}, \qquad (215)$$

and $Z^{(j)}$ denotes the operator that applies a Pauli Z operator to the jth register and acts trivially elsewhere, and $\mathbf{r} \in \mathbb{R}^n$. One can show that the diagonal elements of H are given by

$$\lambda_j = H_{jj} = 2(\mathbf{r} \cdot \mathbf{j}) - \|\mathbf{r}\|_1, \qquad (216)$$

where here and elsewhere we use interchangeably a bitstring denoted as $j = j_0 j_1 \ldots j_{n-1}$ with decimal value $j = \sum_k j_k 2^j \in [0, 2^n - 1]$, and $\mathbf{j} \in \{0,1\}^n$ as its corresponding vector representation. The Fourier spectrum and degeneracy of the encoding strategy are then computed as

$$\Omega = \{\lambda_j - \lambda_i : i, j = 0, \ldots, 2^n - 1\} \qquad (217)$$
$$= \{2(\mathbf{r} \cdot (\mathbf{j} - \mathbf{i})) : \mathbf{i}, \mathbf{j} \in \{0,1\}^n\}, \qquad (218)$$
$$R(\omega) = \{(i, j) : 2\mathbf{r} \cdot (\mathbf{j} - \mathbf{i}) = k\}. \qquad (219)$$

Note that the subtraction in the definition of Ω does not preserve $\mathbb{Z}_2$, i.e. $(\mathbf{j} - \mathbf{i}) \in \{-1, 0, 1\}^n$. From Eq. (217) we see that the largest possible size for $|\Omega|$ is 3^n, the number of unique choices for $(\mathbf{j} - \mathbf{i})$. Any encoding strategy of Eq. (215) therefore introduces a *combinatorial degeneracy*, since the set $\{(\mathbf{j} - \mathbf{i}) : \mathbf{i}, \mathbf{j} \in \{0,1\}^n\}$ is the image of a surjective map on $\{0,1\}^n \times \{0,1\}^n$, with each $(\mathbf{j} - \mathbf{i})$ occurring with multiplicity

$$\#(\mathbf{j} - \mathbf{i}) = 2^{n - \|(\mathbf{j}-\mathbf{i})\|_1}. \qquad (220)$$

To prove this, construct the set $\{(\mathbf{j} - \mathbf{i}) : \mathbf{i}, \mathbf{j} \in \{0,1\}^n\}$ by reflecting the hypercube $\{0,1\}^n$ over 2^n distinct axes, and count the number of vertices shared by among reflected images – this gives the desired degeneracy. Many choices of $\mathbf{r}$ will reduce $|\Omega|$ by reducing the size of the image of the map $(\mathbf{i}, \mathbf{j}) \rightarrow \mathbf{r} \cdot (\mathbf{j} - \mathbf{i})$. Conversely, the spectrum Ω will saturate $|\Omega| = 3^n$ only under particular conditions on $\mathbf{r}$. For example with $n = 2$, $|\Omega| = 9$ is achieved only if $\mathbf{r}$ satisfies all of the following conditions

$$r_0 \neq 0, \qquad r_0 \neq \pm\frac{1}{2}r_1, \qquad r_0 \neq \pm r_1, \qquad r_0 \neq \pm 2r_1, \qquad (221)$$

i.e., the inverse of the dot product $\mathbf{r} \cdot \mathbf{d}$ with respect to an input $\mathbf{d}$ exists if the elements of $\mathbf{r}$ satisfy $r_j \neq r_i \bmod 2^p$ for some integer p; a more general, recursive statement of this property is provided in Ref. [44]. Requirements of this form may be proven using the *frame analysis operator* $T_n \in \mathbb{Z}^{3^n \times n}$, the rows of which contain each element of $\{-1, 0, 1\}^n$. And so to demonstrate Eq. (221) above, if we let $\mathbf{k} \in \mathbb{R}^{3^n}$ be a vector containing (repeated) frequencies in Ω, we observe that $T_2 \mathbf{r} = \mathbf{k} \in \mathbb{R}^9$ will have unique entries only under specific conditions on $\mathbf{r}$. Ω being the result of a frame reconstruction of $\mathbf{r}$ also provides a direct way to tune the elements of Ω: given a length 3^n vector $\mathbf{k}$ containing the (not necessarily unique) desired elements of Ω, one can then compute H by solving $\mathbf{r} = T_n^\dagger \mathbf{k}$.

C.1 Hamming encoding strategy

We instantiate the Hamiltonian of Eq. (215) for

$$\mathbf{r} = \frac{1}{2}(1, 1, \ldots, 1) \in \mathbb{R}^n.$$

This introduces many degeneracies into the spectrum; using $T_n \mathbf{r} = \mathbf{k}$ we immediately recover the result of Ref. [30] that the unique elements of $\mathbf{k}$ are given by

$$\Omega = \{-n, -(n-1), \ldots, 0, \ldots, n-1, n\}, \tag{222}$$

with $|\Omega| = 2n + 1$. To recover the degeneracy of each frequency we compute

$$\lambda_i - \lambda_j = \frac{1}{2}\sum_{t=1}^{n}(-1)^{i_t} - \frac{1}{2}\sum_{t=1}^{n}(-1)^{j_t} = \frac{1}{2}\sum_{t=1}^{n}(1 - 2i_t) - \frac{1}{2}\sum_{t=1}^{n}(1 - 2j_t) = w(j) - w(i), \tag{223}$$

where $w : \{0,1\}^n \to [n]$ is the the weight of a bitstring $b = b_{n-1} \ldots b_1 b_0 \in \{0,1\}^n$ defined as

$$w(b) = \sum_{k=1}^{n} b_k, \tag{224}$$

and when b is written as an integer it should be interpreted according to its binary value. It will be useful to consider subsets of bitstrings having a constant weight. For each $k \in [n]$ (results for $k < 0$ follow by symmetry), we now define the k-weight set $W(k)$ as

$$W(k) = \{j \in \{0,1\}^n : w(j) = k\}. \tag{225}$$

Then the degeneracy set is exactly the set of bitstring indices differing by constant weight:

$$R(k) = \{(i,j) \in \{0,1\}^n \times \{0,1\}^n : w(j) - w(i) = k\} \tag{226}$$

$$= \bigcup_{p=k}^{n} W(p) \times W(p-k). \tag{227}$$

where $\times$ denotes the Cartesian product (this motivates the name "Hamming encoding strategy", as the frequencies and degeneracy of this model arise from index pairs having fixed Hamming weight). From counting arguments, the degeneracy of this model is then

$$|R(k)| = \sum_{j=1}^{n-k}\binom{n}{j}\binom{n}{k+j} = \binom{2n}{n-k}. \tag{228}$$

C.2 Binary encoding strategy

We instantiate the Hamiltonian of Eq. (215) for

$$\mathbf{r} = \frac{1}{2}(2^{n-1}, \ldots, 4, 2, 1) \in \mathbb{R}^n,$$

such that $\mathbf{r}$ generates the decimal value of binary vector. We know from relations like those given in Eq. (221) that $|\Omega| << 3^n$, and indeed using Eq. (216) we find that $\lambda_j = (j - (2^n - 1))/2$ resulting in a frequency spectrum of

$$\Omega = \{-2^n + 1, -2^n + 2, \ldots, 0, 1, \ldots 2^n - 1\}, \tag{229}$$

such that $|\Omega| = 2^{n+1} - 1$, and the degenerate index set of the kernel for frequency k is equivalent (up to permutations) to the indices corresponding to nonzero elements of an elementary Toeplitz matrix with 1's on the k-th diagonal:

$$R(k) = \{(i,j) : i = j - k \; i,j \in [2^n]\}. \tag{230}$$

C.3 Ternary encoding strategy

We instantiate the Hamiltonian of Eq. (215) for

$$\mathbf{r} = \frac{1}{2}(3^{n-1}, \ldots, 9, 3, 1) \in \mathbb{R}^n,$$

resulting in the spectrum

$$\Omega = \left\{ -\frac{1}{2}(3^n - 1), -\frac{1}{2}(3^n - 1) + 1, \ldots, 0, \ldots, \frac{1}{2}(3^n - 1) \right\}. \tag{231}$$

This spectrum (also studied in Ref. [44]) is interesting to study because it is the unique choice giving rise to a Fourier spectrum with frequencies spaced by 1 saturating the $|\Omega| = 3^n$ limit for spectra generated by separable Hamiltonians. Given $k = 2\mathbf{r} \cdot \mathbf{d}$, then $\mathbf{d} = \mathbf{i} - \mathbf{j}$ may be uniquely recovered from k. Suppose k has a ternary representation

$$k = k_n \ldots k_1 k_0 = \sum_j k_j 3^j, \tag{232}$$

with $n = \lceil \log_3 k \rceil$ and $k_j \in \{0, 1, 2\}$. Then defining the operation $t : \mathbb{Z} \to \{0, 1, 2\}^n$ which converts decimals to their vector-form ternary representation according to $t(k) = (k_0, k_1, \ldots, k_{n-1})$, then the degeneracy of this encoding strategy follows directly from the combinatorial degeneracy given in Eq. (220):

$$|R(k)| = 2^{n - \|T(k)\|_1}, \tag{233}$$

where the shifted ternary operation is defined for $k \in [0, 3^n - 1]$

$$T(k) = t(k + \|\mathbf{r}\|_1) - \mathbf{1}, \tag{234}$$

and $\mathbf{1} = (1, 1, \ldots, 1) \in \mathbb{Z}^n$ is the array of all ones. Intuitively, this computation converts a signed ternary expression $k = \sum_j k_j 3^j$ (with $\mathbf{k} \in \{-1, 0, 1\}$) to an unsigned ternary expression provided in Eq. (232).

C.4 Nonseparable models

Given a data-encoding Hamiltonian $H = \text{diag}(\Lambda)$ with $\Lambda \in \mathbb{R}^d$, a key feature of the associated spectrum is that it is preserved under permutations of Λ. This means that a large class of models where $S(x)$ is generated by a separable $H = H_0 \otimes H_1 \otimes H_{n-1}$ are capable of producing the same spectrum as models where $S(x)$ is generated by an nonseparable Hamiltonian. For instance, letting $\sim$ denote equivalence up to permutation and additive shift, it holds that $I \otimes Z \sim Z \otimes Z$, so that a data encoding unitaries

$$S_1(x) = \exp(i2\pi r(I \otimes Z)x), \tag{235}$$
$$S_2(x) = \exp(i2\pi r(Z \otimes Z)x), \tag{236}$$

have identical spectrum and degeneracy properties (differing only in their corresponding degeneracy index sets R). Understanding the classes of nonseparable data-encoding Hamiltonians that are equivalent to a separable encoding strategy up to permutation and translation may be an interesting problem for future work.

We therefore limit our discussion of nonseparable models to a particular choice of Hamiltonian which is provably inequivalent to any separable Hamiltonian, as it saturates the largest possible $|\Omega|$ (and smallest values of $|R(k)|$) available to any single-layer quantum

model. This can be achieved by setting the diagonal of the data-encoding Hamiltonian as the elements of a Golomb ruler [45, 63]. The resulting spectrum is nondegenerate for all nonzero frequencies, though it is straightforward to prove that one cannot achieve uniform spacing in the spectrum (i.e., a *Perfect Golomb ruler*) for $d \geq 5$. As a result, the corresponding spectrum of this model generally exhibits gaps between frequencies. Further exploration of the connections to concepts from radio engineering [64, 65] and classical coding theory [66] may enrich investigations into the spectral properties of quantum models.

D Demonstration of benign overfitting in the general quantum model

In this section we demonstrate an example of benign overfitting in the general quantum model by explicitly constructing a sequence of state preparation unitaries U for a particular choice of data-encoding unitary $S(d)$. We will consider a d-dimensional version of the Binary encoding strategy constructed from a data-encoding Hamiltonian $H = \text{diag}\,(0, 1, 2, \ldots, d - 1)$ that gives rise to

$$\Omega = \left\{ \frac{-d}{2} + 1, \ldots, 0, \ldots, \frac{d}{2} - 1 \right\}, \tag{237}$$

$$|R(k)| = d - |k|, \tag{238}$$

for $k \in \Omega$, and we will require d to be even for simplicity. Suppose the target function spectrum has size $n_0 < n < d$ for some integer n and satisfies $(n_0 + 1) \bmod 4 = 0$. Define the constants

$$c_1 = \frac{d}{2} - \frac{n_0 + 1}{4}, \tag{239}$$

$$c_2 = \frac{d}{2} + \frac{n_0 + 1}{4}. \tag{240}$$

We then choose a constant $a \in [0, 2/(n_0 + 1)]$ and prepare $|\Gamma\rangle$ with elements given by

$$\gamma_j = \begin{cases} \sqrt{a}, & j \in [c_1, c_2), \\ \sqrt{b}, & j \in [0, c_1) \cup [c_2, d), \end{cases} \tag{241}$$

where normalization requires that

$$b = \frac{1 - \frac{n_0 + 1}{2} a}{d - \frac{n_0 + 1}{2}}. \tag{242}$$

For this encoding strategy, the optimal feature weights of the quantum model corresponding to positive frequencies $k \in \Omega_+$ are given by

$$\nu_k = \sum_{j=0}^{d-k-1} |\gamma_j|^2 |\gamma_{j+k}|^2. \tag{243}$$

Counting arguments and algebraic simplification lead to

$$\nu_k = \begin{cases} (d - B_0 - 2k)b^2 + 2kab + (B_0 - k)a^2, & 0 \leq k < c_2 - c_1, \\ (d - 2B_0 - k)b^2 + 2B_0ab, & c_2 - c_1 \leq k < c_1, \\ (k - B_0)b^2 + (d - B_0 - 2k)ab, & c_1 \leq k < c_2, \\ (d - k)b^2, & c_2 \leq k < d, \end{cases} \tag{244}$$

where $B_0 = (n_0 + 1)/2$ and $\nu_{k+1} \leq \nu_k$ for all $k > 0$. Noting that $b = \mathcal{O}(d^{-1})$ and a is constant, we find that ν_k has three distinct scaling regimes: $\nu_k = \Omega(1)$ for $k \in \Omega_{n_0}$, $\nu_k = \Theta\left(d^{-2}\right)$ when $k \in [c_2, d)$, and $\nu_k = \mathcal{O}(d^{-1})$ otherwise. We can use this behavior to bound the generalization error of the quantum model in a manner analogous to Sec. A.3.1. Defining $m = |S(k)| - 1$, we bound the bias according to

$$\text{BIAS}^2 = \sum_{k \in \Omega_{n_0}} |\hat{g}_k|^2 \frac{\left(\sum_{\ell \in S(k) \backslash k} \nu_\ell\right)^2 + \sum_{\ell \in S(k) \backslash k} \nu_\ell^2}{\left(\sum_{\ell \in S(k)} \nu_\ell\right)^2} \tag{245}$$

$$\leq \sum_{k \in \Omega_{n_0}} |\hat{g}_k|^2 \frac{m(m+1)(db^2 + 2B_0 ab)^2}{(a^2 + mb^2)^2} \tag{246}$$

$$\leq \left[m(m+1)b^2\right] \left(\frac{db + 2B_0 a}{a^2 + mb^2}\right)^2 P \tag{247}$$

$$= \mathcal{O}\left(\frac{1}{n^2}\right), \tag{248}$$

where we have defined $P := \sum_{k \in \Omega_{n_0}} |\hat{g}_k|^2$. In line (246) we have used the monotonicity of ν_k to bound the components of $S(k) \backslash k$ in the numerator according to the upper bound for the interval $k \in [c_2 - c_1, c_1)$ and applied a lower bound for elements of $S(k) \backslash k$ in the denominator using the interval $k \in [c_2, d)$. The scaling of Eq. (248) follows from $m = \mathcal{O}(d/n)$, which implies that $m(m+1)b^2 = \mathcal{O}(n^{-2})$, while the remaining terms in line (247) scale as no greater than a constant. To bound the variance, we observe

$$\sum_{k \in \Omega_{n_0}} \frac{\sum_{\ell \in S(k)} \nu_\ell^2}{\left(\sum_{\ell \in S(k)} \nu_\ell\right)^2} \tag{249}$$

$$\leq \sum_{k \in \Omega_{n_0}} \frac{\left[a^2(B_0 - k) + 2kab + (d - B_0 - 2k)b^2\right]^2 + \left[(d - 2B_0 - k)b^2 + 2B_0 ab\right]^2}{(a^2 + mb^2)^2} \tag{250}$$

$$\leq n_0 \frac{(a^2 B_0 + n_0 ab + db^2)^2 + m(db^2 + 2B_0 ab)^2}{(a^2 + mb^2)^2} \tag{251}$$

$$= \mathcal{O}(1), \tag{252}$$

which follows identically to the arguments used to reach Eq. (248) and the observation that $2k \leq n_0$ for $k \in \Omega_{n_0}$. Furthermore, defining $m' = \lfloor (c_1 - n_0)/n \rfloor$ as the lower bound for the number of aliases of any $k \in \{\ell \in \Omega_n \backslash \Omega_{n_0} : c_1 \leq \ell < c_2\}$ we have

$$\sum_{k \in \Omega_n \backslash \Omega_{n_0}} \frac{\sum_{\ell \in S(k)} \nu_\ell^2}{\left(\sum_{\ell \in S(k)} \nu_\ell\right)^2} \leq \frac{(m+1)(db^2 + 2B_0 ab)^2}{(m'\left[(d - 2B_0 - k)b^2 + 2B_0 ab\right] + (m + 1 - m')b^2)^2} \tag{253}$$

$$\leq \frac{n - n_0}{(m+1)} \frac{(db + 2B_0 a)^2}{\left(\frac{m'}{m+1}\left[(d - 2B_0 - c_1)b + 2B_0 a\right] + \left(1 - \frac{m'}{m+1}\right)b\right)^2} \tag{254}$$

$$= \mathcal{O}\left(\frac{n^2}{d}\right), \tag{255}$$

where in line (254) we have split the denominator into frequencies $n_0/2 \leq k < c_1$ and $c_1 \leq k < d$, bounding ν_k for the latter as b^2. Line (255) follows from observing that $m' = \Theta(d/n) \Rightarrow m'/m = \Theta(1)$, and thus the entire second term of line (254) is bounded

as $\mathcal{O}(1)$. Combining Eqs. (252) and (255) with Eq. (88) for the variance of the Fourier features model gives

$$\text{VAR} = \mathcal{O}\left(\frac{1}{n} + \frac{n}{d}\right). \tag{256}$$

It follows that the minimum-$\|\cdot\|_F$ interpolating quantum model with state preparation unitary U satisfying Eq. (241) (up to permutations) and data encoded using the Binary encoding strategy will result in benign overfitting as long as the dimensionality of the model scales as $d = \omega(n)$ (i.e., the number of qubits satisfies $n_q = \omega(\log n)$), in which case Eqs. (248) and (256) characterizing the generalization error of the model both vanish in the large-n limit.

A convenience of this demonstration of benign overfitting in a quantum model was that the state $|\Gamma\rangle$ of Eq. (241) incorporates knowledge of the target function (namely the band-limited spectrum Ω_{n_0}). This could be considered a limitation of the approach, as it imposes an inductive bias on the resulting interpolating model f^{opt}, in contrast to other examples of benign overfitting that are more agnostic to the underlying distribution [6, 59]. Future work could reveal choices of $|\Gamma\rangle$ that are more data-independent (no explicit dependence on n_0) but give rise to feature weights ν_k^{opt} with the same desirable properties as Eq. (244). As described in Appendix A.3.2, the desirable properties include a more weight on all ν_k with $k \in \Omega_{n_0}$ and a long, thin tail of feature weights for all other $k \in \Omega_d \backslash \Omega_{n_0}$.