arXiv:2309.11647v4 [quant-ph] 18 Mar 2025

Potential and limitations of random Fourier features for dequantizing quantum machine learning

Ryan Sweke[1], Erik Recio-Armengol[2,3], Sofiene Jerbi[4], Elies Gil-Fuster[4,5], Bryce Fuller[7], Jens Eisert[4,5,6], and Johannes Jakob Meyer[4]

[1]IBM Quantum, Almaden Research Center, San Jose, CA, USA
[2]ICFO-Institut de Ciencies Fotoniques, The Barcelona Institute of Science and Technology, 08860 Castelldefels, Spain
[3]Eurecat, Centre Tecnologic de Catalunya, Multimedia Technologies, Barcelona, Spain
[4]Dahlem Center for Complex Quantum Systems, Freie Universität Berlin, Berlin, Germany
[5]Fraunhofer Heinrich Hertz Institute, 10587 Berlin, Germany
[6]Helmholtz-Zentrum Berlin für Materialien und Energie, 14109 Berlin, Germany
[7]IBM Quantum, IBM T.J. Watson Research Center, Yorktown Heights, NY 10598

Quantum machine learning is arguably one of the most explored applications of near-term quantum devices. Much focus has been put on notions of variational quantum machine learning where *parameterized quantum circuits* (PQCs) are used as learning models. These PQC models have a rich structure which suggests that they might be amenable to efficient dequantization via *random Fourier features* (RFF). In this work, we establish necessary and sufficient conditions under which RFF does indeed provide an efficient dequantization of variational quantum machine learning for regression. We build on these insights to make concrete suggestions for PQC architecture design, and to identify structures which are necessary for a regression problem to admit a potential quantum advantage via PQC based optimization.

1 Introduction

In recent years, the technique of using *parameterized quantum circuits* (PQCs) to define a model class, which is then optimized over via a classical optimizer, has emerged as one of the primary methods of using near-term quantum devices for machine learning tasks [CAB+21; BLS+19]. We will refer to this approach as *variational quantum machine learning* (variational QML), although it is often also referred to as hybrid quantum/classical optimization. While a large amount of effort has been invested in both understanding the theoretical properties of variational QML, and experimenting on trial datasets, it remains unclear whether variational QML on near-term quantum devices can offer any meaningful advantages over state-of-the-art classical methods.

One approach to answering this question is via *dequantization*. In this context, the idea is to use insights into the structure of PQCs, and the model classes that they define, to design quantum-inspired classical methods which can be proven to match the performance of variational QML. Ultimately, the goal is to understand when and why variational QML can be dequantized, in order to better identify the PQC architectures, optimization algorithms and problem types for which one might obtain a meaningful quantum advantage via variational QML.

In order to discuss notions of dequantization of variational QML, we note that for typical applications variational QML consists of two distinct phases. Namely, a *training* stage and an

inference stage. In the training stage, one uses the available training data to identify an optimal PQC model, and in the inference stage one uses the identified model to make predictions on previously unseen data, or in the case of generative modelling, to generate new samples from the unknown data distribution.

A variety of works have recently proposed dequantization methods for *inference* with PQC models. The first such work was Ref. [SEM23], which used insights into the functional analytic structure of PQC model classes to show that, given a trained quantum model, one can sometimes efficiently extract a purely classical model – referred to as a *classical surrogate* – which performs inference just as well as the PQC model. More recently, Ref. [JGM+24] used insights from shadow tomography to show that it is sometimes possible to extract a classical shadow of a trained PQC model – referred to as a *shadow model* – which is again guaranteed to perform as well as the PQC model for inference. Interestingly, however, Ref. [JGM+24] also proved that, under reasonable complexity-theoretic assumptions, there exist PQC models whose inference *cannot* be dequantized by any efficient method – i.e., *all* efficient methods for the dequantization of PQC inference possess some fundamental limitations.

In this work, we are concerned with dequantization of the *training* stage of variational QML – i.e., the construction of efficient classical learning algorithms which can be proven to match the performance of variational QML in learning from data. To this end, we start by noting that Ref. [JGM+24] constructed a learning problem which admits an efficient variational QML algorithm but, again under complexity-theoretic assumptions, cannot be dequantized by any efficient classical learning algorithm. As such, we know that all methods for the dequantization of variational QML must posses some fundamental limitations, and for any given method we would like to understand its domain of applicability.

With this in mind, one natural idea is to ask whether the effect of *noise* allows direct efficient classical simulation of the the PQC model, and, therefore, of the entire PQC training process and subsequent inference. Indeed, a series of recent works has begun to address this question, and to delineate the conditions under which noise renders PQC models classically simulatable [FRD+22; FRD+23; SWC+23]. Another recent result has shown that the presence of *symmetries*, often introduced to improve PQC model performance for symmetric problems [MMG+23; LSS+22], can also constrain PQC models in a way which allows for efficient classical simulation [ABK+23].

In this work however, inspired by the dequantization of inference with PQC models, we start from the idea of "training the surrogate model". More specifically, Ref. [SEM23] used the insight that the class of functions realizable by PQC models is a subset of trigonometric polynomials with a specific set of frequencies [SSM21], to "match" the trained PQC model to the closest trigonometric polynomial with the correct frequency set (which is then the classical surrogate for inference). This, however, immediately suggests the following approach to dequantizing the *training* stage of variational QML – simply directly optimize from data over the "PQC-inspired" model class of trigonometric polynomials with the appropriate frequency set. Indeed, this is in some sense what happens during variational QML!

The above idea was explored numerically in the original work on classical surrogates for inference [SEM23]. Unfortunately, for typical PQC architectures the number of frequencies in the corresponding frequency set grows exponentially with the problem size (defined as the dimensionality of the input data), which prohibits efficient classical optimization over the relevant class of trigonometric polynomials. However, for the PQC architectures for which numerical experiments were possible, direct classical optimization over the PQC-inspired classical model class yielded trained models which outperformed those obtained from variational QML.

Inspired by Ref. [SEM23], the subsequent work of Ref. [LTD+22] introduced a method for addressing the efficiency bottleneck associated with exponentially growing frequency sets. The authors of Ref. [LTD+22] noticed that all PQC models are linear models with respect to a trigonometric polynomial feature map. As such, one can optimize over all PQC-inspired models via kernel ridge regression, which will be efficient if one can efficiently evaluate the kernel defined from the feature map. While naively evaluating the appropriate kernel classically will be inefficient – again due to the exponential growth in the size of the frequency set – the clever insight of Ref. [LTD+22] was to see that one can gain efficiency improvements by using the technique of *random Fourier features* [RR07] to *approximate* the PQC-inspired kernel. Using this technique, Ref. [LTD+22] obtained a variety of theoretical results concerning the sample complexity required for RFF-based regression with the PQC-inspired kernel to yield a model whose performance matches that of variational QML. However, the analysis of Ref. [LTD+22] applied only to PQC architectures with *universal* parameterized circuit blocks – i.e., PQC models which can realize *any* trigonometric polynomial with the appropriate frequencies. This contrasts with the PQC models arising from practically relevant PQC architectures, which due to depth constraints, can only realize a subset of trigonometric polynomials.

In light of the above, the idea of this work is to further explore the potential and limitations of RFF-based linear regression as a method for dequantizing the training stage of variational QML, with the goal of providing an analysis which is applicable to practically relevant PQC architectures. In particular, we identify a collection of necessary and sufficient requirements – of the PQC architecture, the regression problem, and the RFF procedure – for RFF-based linear regression to provide an efficient classical dequantization method for PQC based regression. This allows us to show clearly that:

1. RFF-based linear regression *cannot* be a generic dequantization technique. At least, there exist regression problems and PQC architectures for which RFF-based linear regression *cannot* efficiently dequantize PQC based regression. As mentioned before, we already knew from the results of Ref. [JGM+24] that *all* dequantization techniques must posses some limitations, and our results shed light on the specific nature of these limitations for RFF-based dequantization.

2. There exist practically relevant PQCs, and regression problems, for which RFF-based linear regression can be *guaranteed* to efficiently produce output models which perform as well as the best possible output of PQC based optimization. In other words, there exist problems and PQC architectures for which PQC dequantization via RFF is indeed possible.

Additionally, using the necessary and sufficient criteria that we identify, we are able to provide concrete recommendations for PQC architecture design, in order to mitigate the possibility of dequantization via RFF. Moreover, we are able to identify a necessary condition on the structure of a regression problem which ensures that dequantization via RFF is *not* possible. This, therefore, provides a guideline for the identification of problems which admit a potential quantum advantage (or at least, cannot be dequantized via RFF-based linear regression).

This paper is structured as follows: We begin in Section 2 by providing all the necessary preliminaries and background material. Following this, we proceed in Section 3 to motivate and present RFF-based linear regression with PQC-inspired kernels as a method for the dequantization of variational QML. Given this, we then go on in Section 4 to provide a detailed theoretical analysis of RFF-based linear regression with PQC-inspired kernels. Finally, we conclude in Section 5 with a discussion of the consequences of the previous analysis, and with an overview of natural directions for future research.

2 Setting and preliminaries

Here we provide the setting and required background material.

2.1 Statistical learning framework

Let $\mathcal{X}$ denote a set of all possible data points, and $\mathcal{Y}$ a set of all possible labels. In this work, we will set $\mathcal{X} := [0, 2\pi)^d \subset \mathbb{R}^d$ for some integer d and $\mathcal{Y} = \mathbb{R}$. We assume the existence of some unknown probability distribution P over $\mathcal{X} \times \mathcal{Y}$, which we refer to as a regression problem. Note that we consider d, the dimensionality of the input data, as the size of the problem, and as such the relevant scaling parameter for the analysis of algorithms which aim to solve the problem. Additionally, we assume a parameterized class of functions $\mathcal{F} = \{f_\theta \colon \mathcal{X} \to \mathcal{Y} \,|\, \theta \in \Theta\}$, which we call hypotheses. Given access to some finite dataset $S = \{(x_i, y_i) \sim P\}|_{i=1}^n$ the goal of the regression problem specified by P is to identify the optimal hypothesis f_{θ^*}, i.e., the hypothesis which minimizes the *true risk*, defined via

$$R(f) := \mathop{\mathbb{E}}_{(x,y)\sim P} [\mathcal{L}(y, f(x))], \tag{1}$$

where $\mathcal{L} \colon \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ is some loss function. In this work, we will consider only the quadratic loss defined via

$$\mathcal{L}(y, y') := (y - y')^2. \tag{2}$$

We also define the *empirical risk* with respect to the dataset S as

$$\hat{R}(f) := \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}(y_i, f(x_i)). \tag{3}$$

2.2 Linear and kernel ridge regression

Linear ridge regression and kernel ridge regression are two popular classical learning algorithms. In linear ridge regression we consider *linear* functions $f_{\mathrm{w}}(x) = \langle \mathrm{w}, x \rangle$, and given a dataset $S = \{(x_i, y_i)\}|_{i=1}^n$, we proceed by minimizing the empirical risk, regularized via the 2-norm, i.e.,

$$\hat{R}_\lambda(f_{\mathrm{w}}) := \frac{1}{n} \sum_{i=1}^{n} (y_i - \langle \mathrm{w}, x_i \rangle)^2 + \lambda \|\mathrm{w}\|_2^2. \tag{4}$$

With this regularization, which is added to prevent over-fitting, minimizing Eq. (4) becomes a convex quadratic problem, which admits the closed form solution

$$\mathrm{w} = \left(\hat{X}^T \hat{X} + \lambda n \mathbb{1}\right)^{-1} \hat{X}^T \hat{Y}, \tag{5}$$

where $\hat{X}$ is the $n \times d$ "data matrix" with x_i as rows, and $\hat{Y}$ is the n dimensional "target vector" with y_i as the i'th component [MRT18]. Linear ridge regression requires $\mathcal{O}(nd)$ space and $\mathcal{O}(nd^2 + d^3)$ time. As linear functions are often not sufficiently expressive, a natural approach is to consider linear functions in some higher dimensional feature space. More specifically, one assumes a feature map $\phi \colon \mathbb{R}^d \to \mathbb{R}^D$, and then considers linear functions of the form $f_v(x) = \langle v, \phi(x) \rangle$, where v is an element of the feature space $\mathbb{R}^D$. Naively, one could do linear regression at a space and time cost of $\mathcal{O}(nD)$ and $\mathcal{O}(nD^2 + D^3)$, respectively. However, often we would like to consider D extremely large (or infinite) and this is therefore infeasible. The solution is to use "the kernel trick" and consider instead a kernel function $K \colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ which satisfies

$$K(x, x') = \langle \phi(x), \phi(x') \rangle, \tag{6}$$

but which can ideally be evaluated more efficiently than by explicitly constructing $\phi(x)$ and $\phi(x')$ and taking the inner product. Given such a function, we know that the minimizer of the regularized empirical risk is given by

$$\begin{aligned} f_\alpha(x) &= \sum_{i=1}^{n} \alpha_i K(x_i, x) \\ &= \left\langle \sum_{i=1}^{n} \alpha_i \phi(x_i), \phi(x) \right\rangle \\ &= \langle v, \phi(x) \rangle, \end{aligned} \tag{7}$$

where

$$\alpha = \left(\hat{K} + n\lambda \mathbb{1}\right)^{-1} \hat{Y}, \tag{8}$$

with $\hat{K}$ the kernel matrix (or Gram matrix) with entries $\hat{K}_{i,j} = K(x_i, x_j)$. Solving Eq. (8) is known as *kernel ridge regression*. If one assumes that evaluating $K(x, x')$ requires constant time, then kernel ridge regression has space and time cost $\mathcal{O}(n^2)$ and $\mathcal{O}(n^3)$, respectively. We note that in practice one hardly ever specifies the feature map ϕ, and instead works directly with a suitable kernel function K.

2.3 Random Fourier features

For many applications, in which the number of samples n can be extremely large, a space and time cost of $\mathcal{O}(n^2)$ and $\mathcal{O}(n^3)$, respectively, prohibits the implementation of kernel ridge regression. This has motivated the development of methods which can bypass these complexity bottlenecks. The method of *random Fourier features* (RFF) is one such method [RR07]. To illustrate this method we follow the presentation of Ref. [RR17], and start by assuming that the kernel $K\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, has *an integral representation*. More specifically, we assume there exists some probability space (Φ, π) and some function $\psi\colon \mathcal{X} \times \Phi \to \mathbb{R}$ such that for all $x, x' \in \mathcal{X}$ one has that

$$K(x, x') = \int_\Phi \psi(x, \nu)\psi(x', \nu)\, \mathrm{d}\pi(\nu). \tag{9}$$

The method of random Fourier features is then based around the idea of approximating $K(x, x')$ by using Monte-Carlo type integration to perform the integral in Eq. (9). More specifically, one uses

$$K(x, x') \approx \langle \tilde{\phi}_M(x), \tilde{\phi}_M(x') \rangle, \tag{10}$$

where $\tilde{\phi}_M\colon \mathcal{X} \to \mathbb{R}^M$ is a randomized feature map of the form

$$\tilde{\phi}_M(x) = \frac{1}{\sqrt{M}}(\psi(x, \nu_1), \ldots, \psi(x, \nu_M)), \tag{11}$$

where $\nu_1, \ldots, \nu_m$ are M features sampled randomly from π. Using the approximation in Eq. (10) allows one to replace kernel ridge regression via K with linear regression with respect to the random feature map $\tilde{\phi}_M$. This yields a learning algorithm with time and space cost $\mathcal{O}(nM)$ and $\mathcal{O}(nM^2 + M^3)$, respectively, which is more efficient than kernel ridge regression whenever $M < n$. Naturally, the quality (i.e., true risk) of the output solution will depend heavily on how large M is chosen. However, we postpone until later a detailed discussion of this issue, which is central to the results and observations of this work.

So far we have *assumed* the existence of an integral representation for the kernel, as in Eq. (9). However, such a representation is not typically provided, and as such the first step towards the

implementation of linear regression with RFF is the derivation of an integral representation for the kernel of interest. Luckily however, for *shift-invariant* kernels, which are kernels of the form $K(x, x') = \overline{K}(x - x')$ for some function $\overline{K}\colon \mathcal{X} \to \mathbb{R}$, the integral representation can be easily derived from the Fourier transform of $\overline{K}$ [RR07; SS15b]. More specifically, for any shift-invariant kernel we can write

$$\overline{K}(x - x') = \int_{\omega \in \mathcal{X}} e^{i\langle \omega, x - x' \rangle} q(\omega) \mathrm{d}\omega, \tag{12}$$

where $q : \mathcal{X} \to \mathbb{R}$ is the Fourier transform of $\overline{K}$. Bochner's theorem ensures that q is a non-negative measure, and when the kernel is scaled such that $\overline{K}(0) = 1$, then it additionally ensures that q is indeed a proper probability distribution. Additionally, as the kernel is real-valued, we can replace the integrand $e^{i\langle \omega, x - x' \rangle}$ with $\cos(\langle \omega, x - x' \rangle)$. Doing this, we then have

$$\begin{aligned} K(x - x') &= \int_{\omega \in \mathcal{X}} \cos(\langle \omega, x - x' \rangle) q(\omega) \mathrm{d}\omega \\ &= \frac{1}{2\pi} \int_{\omega \in \mathcal{X}} \int_{\gamma \in [0, 2\pi)} \sqrt{2} \cos(\langle \omega, x \rangle + \gamma) \sqrt{2} \cos(\langle \omega, x' \rangle + \gamma) q(\omega) \, \mathrm{d}\omega \, \mathrm{d}\gamma \\ &\coloneqq \int_{\Phi = \mathcal{X} \times [0, 2\pi)} \psi(x, \nu) \psi(x', \nu) \, \mathrm{d}\pi(\nu), \end{aligned} \tag{13}$$

with $\psi(x, \nu) \coloneqq \sqrt{2} \cos(\langle \omega, x \rangle + \gamma)$, and $\pi = q \times \mu$ where μ is the uniform measure over $[0, 2\pi)$. As such, we have indeed arrived at an integral representation for K. Given this derivation, we note that the name *random Fourier features* comes from the fact that the measure π is proportional to the Fourier transform of $\overline{K}$.

2.4 PQC models for variational QML

As discussed in Section 1, variational QML is based on the classical optimization of models defined via *parameterized quantum circuits* (PQCs) [BLS+19]. In the context of regression, one begins by fixing a parameterized quantum circuit C, whose gates can depend on both data points $x \in \mathcal{X}$ and components of a vector $\theta \in \Theta$ of variational parameters, where typically $\Theta = [0, 2\pi)^c$ for some c. For each data point x and each vector of variational parameters θ, this circuit realizes the unitary $U(x, \theta)$. Given this, we then choose an observable O, and define the associated PQC model class $\mathcal{F}_{(C,O)}$ as the set of all functions $f_\theta\colon \mathcal{X} \to \mathbb{R}$ defined via

$$f_\theta(x) = \langle 0 | U^\dagger(x, \theta) O U(x, \theta) | 0 \rangle \tag{14}$$

for all $x \in \mathcal{X}$, i.e.,

$$\mathcal{F}_{(C,O)} = \{ f_\theta(\cdot) = \langle 0 | U^\dagger(\cdot, \theta) O U(\cdot, \theta) | 0 \rangle \,|\, \theta \in \Theta \}. \tag{15}$$

One then proceeds by using a classical optimization algorithm to optimize over the variational parameters θ. In this work, we consider an important sub-class of PQC models in which the classical data x enters only via Hamiltonian time evolutions, whose duration is controlled by a single component of x. To be more precise, for $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$, we assume that each gate in the circuit C which depends on x is of the form

$$V_{(j,k)}(x_j) = e^{-i H_k^{(j)} x_j}, \tag{16}$$

for some Hamiltonian $H_k^{(j)}$. We stress that we exclude here more general encoding schemes, such as those which allow for time evolutions parameterized by functions of x, or time evolutions of parameterized linear combinations of Hamiltonians. We denote by $\mathcal{D}^{(j)} = \{ H_k^{(j)} \,|\, k \in [L_j] \}$ the

set of all L_j Hamiltonians which are used to encode the component x_j at some point in the circuit, and we call the tuple

$$\mathcal{D} \coloneqq \left(\mathcal{D}^{(1)}, \ldots, \mathcal{D}^{(d)}\right) \tag{17}$$

the *data-encoding strategy*. It is by now well known that these models admit a succinct "classical" description [SSM21; GT20; CGM+21] given by

$$f_\theta(x) = \sum_{\omega \in \tilde{\Omega}_\mathcal{D}} c_\omega(\theta) e^{i\langle \omega, x\rangle}, \tag{18}$$

where

1. the set of frequency vectors $\tilde{\Omega}_\mathcal{D} \subseteq \mathbb{R}^d$ is completely determined by the data-encoding strategy. We describe the construction of $\tilde{\Omega}_\mathcal{D}$ from $\mathcal{D}$ in Appendix A.
2. the frequency coefficients $c_\omega(\theta)$ depend on the trainable parameters θ, but in a way which usually does not admit a concise expression.

As described in Ref. [SSM21], we know that $\omega_0 \coloneqq (0, \ldots, 0) \in \tilde{\Omega}_\mathcal{D}$ and that the non-zero frequencies in $\tilde{\Omega}_\mathcal{D}$ come in mirror pairs – i.e., $\omega \in \tilde{\Omega}_\mathcal{D}$ implies $-\omega \in \tilde{\Omega}_\mathcal{D}$. Additionally, one has $c_\omega(\theta) = c^*_{-\omega}(\theta)$ for all $\omega \in \tilde{\Omega}_\mathcal{D}$ and all θ, which ensures the function f_θ evaluates to a real number. As a result, we can perform an arbitrary splitting of pairs to redefine $\tilde{\Omega}_\mathcal{D} \coloneqq \Omega_\mathcal{D} \cup (-\Omega_\mathcal{D})$, where $\Omega_\mathcal{D} \cap (-\Omega_\mathcal{D}) = \{\omega_0\}$. It will also be convenient to define $\Omega^+_\mathcal{D} \coloneqq \Omega_\mathcal{D} \setminus \{\omega_0\}$. Given this, by defining

$$a_\omega(\theta) \coloneqq c_\omega(\theta) + c_{-\omega}(\theta), \tag{19}$$
$$b_\omega(\theta) \coloneqq i(c_\omega(\theta) - c_{-\omega}(\theta)) \tag{20}$$

for all $\omega \in \Omega^+_\mathcal{D}$, and writing $\Omega_\mathcal{D} = \{\omega_0, \omega_1, \ldots, \omega_{|\Omega^+_\mathcal{D}|}\}$, we can rewrite Eq. (18) as

$$\begin{aligned} f_\theta(x) &= c_{\omega_0}(\theta) + \sum_{i=1}^{|\Omega^+_\mathcal{D}|} \left(a_{\omega_i}(\theta)\cos(\langle \omega_i, x\rangle) + b_{\omega_i}(\theta)\sin(\langle \omega_i, x\rangle)\right) \\ &= \langle c(\theta), \phi_\mathcal{D}(x)\rangle \end{aligned} \tag{21}$$

where

$$c(\theta) \coloneqq \sqrt{|\Omega_\mathcal{D}|}\left(c_{\omega_0}(\theta), a_{\omega_1}(\theta), b_{\omega_1}(\theta), \ldots, a_{\omega_{|\Omega^+_\mathcal{D}|}}(\theta), b_{\omega_{|\Omega^+_\mathcal{D}|}}(\theta)\right), \tag{22}$$

$$\phi_\mathcal{D}(x) \coloneqq \frac{1}{\sqrt{|\Omega_\mathcal{D}|}}\left(1, \cos(\langle \omega_1, x\rangle), \sin(\langle \omega_1, x\rangle), \ldots, \cos\left(\langle \omega_{|\Omega^+_\mathcal{D}|}, x\rangle\right), \sin\left(\langle \omega_{|\Omega^+_\mathcal{D}|}, x\rangle\right)\right), \tag{23}$$

and the normalization constant has been chosen to ensure that $\langle \phi_\mathcal{D}(x), \phi_\mathcal{D}(x)\rangle = 1$, which will be required shortly. The formulation in Eq. (21) makes it clear that f_θ is a *linear* model in $\mathbb{R}^{|\tilde{\Omega}_\mathcal{D}|}$, taken with respect to the feature map $\phi_\mathcal{D} \colon \mathcal{X} \to \mathbb{R}^{|\tilde{\Omega}_\mathcal{D}|}$. We can now define the model class of all linear models realizable by the parameterized quantum circuit with data-encoding strategy $\mathcal{D}$, variational parameter set Θ and observable O via

$$\mathcal{F}_{(\Theta,\mathcal{D},O)} = \{f_\theta(\cdot) = \langle 0|U^\dagger(\theta,\cdot)OU(\theta,\cdot)|0\rangle \,|\, \theta \in \Theta\} \tag{24}$$
$$= \{f_\theta(\cdot) = \langle c(\theta), \phi_\mathcal{D}(\cdot)\rangle \,|\, \theta \in \Theta\}. \tag{25}$$

In what follows, we will use a tuple $(\Theta, \mathcal{D}, O)$ to represent a PQC architecture, as we have done above. We note that for all $f \in \mathcal{F}_{(\Theta,\mathcal{D},O)}$ we have that $\|f\|_\infty \leq \|O\|_\infty$.

It is critical to note that due to the constraints imposed by the circuit architecture $(\Theta, \mathcal{D}, O)$, the class $\mathcal{F}_{(\Theta,\mathcal{D},O)}$ may not contain *all* possible linear functions with respect to the feature map $\phi_{\mathcal{D}}$. Said another way, the circuit architecture gives rise to an *inductive bias* in the set of functions which can be realized. However, for the analysis that follows, it will be very useful for us to define the set of *all* linear functions $\mathcal{F}_{\mathcal{D}}$ with respect to the feature map $\phi_{\mathcal{D}}$, i.e.,

$$\mathcal{F}_{\mathcal{D}} = \left\{ f_v(\cdot) = \langle v, \phi_{\mathcal{D}}(\cdot) \rangle \,|\, v \in \mathbb{R}^{|\tilde{\Omega}_{\mathcal{D}}|} \right\}. \tag{26}$$

As shown in Figure. 1, we stress that for any architecture $(\Theta, \mathcal{D}, O)$, we have that

$$\mathcal{F}_{(\Theta,\mathcal{D},O)} \subset \mathcal{F}_{\mathcal{D}}. \tag{27}$$

The inclusion is strict due to the fact that for all $f \in \mathcal{F}_{(\Theta,\mathcal{D},O)}$ we know that $\|f\|_\infty \leq \|O\|_\infty$, whereas $\mathcal{F}_{\mathcal{D}}$ contains functions of arbitrary infinity norm. However, we note that if one defines the set

$$\mathcal{F}_{(\mathcal{D},O)} = \{ f \in \mathcal{F}_{\mathcal{D}} \,|\, \|f\|_\infty \leq \|O\|_\infty \}, \tag{28}$$

then in principle there could exist architectures for which $\mathcal{F}_{(\Theta,\mathcal{D},O)} = \mathcal{F}_{(\mathcal{D},O)}$, and therefore one has $\mathcal{F}_{(\Theta,\mathcal{D},O)} \subseteq \mathcal{F}_{(\mathcal{D},O)}$ for all architectures $(\Theta, \mathcal{D}, O)$. As such, one can in some sense think of $\mathcal{F}_{(\mathcal{D},O)}$ as the "closure" of $\mathcal{F}_{(\Theta,\mathcal{D},O)}$.

2.5 PQC feature map and PQC-kernel

Given the observation from Section 2.4 that all PQC models are linear in some high-dimensional feature space fully defined by the data-encoding strategy, we can very naturally associate to each data-encoding strategy both a feature map and an associated kernel function, which we call the *PQC-kernel*:

Definition 1: (PQC feature map and PQC-kernel) Given a data-encoding strategy $\mathcal{D}$, we define the PQC feature map $\phi_{\mathcal{D}} \colon \mathcal{X} \to \mathbb{R}^{|\tilde{\Omega}_{\mathcal{D}}|}$ via Eq. (23), i.e.,

$$\phi_{\mathcal{D}}(x) \coloneqq \frac{1}{\sqrt{|\Omega_{\mathcal{D}}|}} \left(1, \cos(\langle \omega_1, x \rangle), \sin(\langle \omega_1, x \rangle), \ldots, \cos\left(\omega_{|\Omega_{\mathcal{D}}^+|}, x\right\rangle\right), \sin\left(\omega_{|\Omega_{\mathcal{D}}^+|}, x\right\rangle)\right) \tag{29}$$

for $\omega_i \in \Omega_{\mathcal{D}}^+$. We then define the PQC-kernel $K_{\mathcal{D}}$ via

$$K_{\mathcal{D}}(x, x') \coloneqq \langle \phi_{\mathcal{D}}(x), \phi_{\mathcal{D}}(x') \rangle. \tag{30}$$

It is crucial to stress that the classical PQC-kernel defined in Definition 1 is fundamentally *different* from the so called "quantum kernels" often considered in QML – see, for example, Refs. [Sch21; JFN+23] – which are defined from a data-parameterized unitary $U(x)$ via $K(x, x') = \mathrm{Tr}[\rho(x)\rho(x')]$ with $\rho(x) = U(x)|0\rangle\langle 0|U^\dagger(x)$. Additionally, we note that the feature map $\phi_{\mathcal{D}}$ defined in Definition 1 is *not* the unique feature map with the property that all functions in $\mathcal{F}_{(\Theta,\mathcal{D},O)}$ are linear with respect to the feature map. Indeed, we will see in Section 4.3 that any "re-weighting" of $\phi_{\mathcal{D}}$ will preserve this property.

3 Potential of RFF-based linear regression for dequantizing variational QML

Let us now imagine that we have a regression problem to solve. More precisely, imagine that we have a dataset S, with n elements drawn from some distribution P, as per Section 2.1. One option is for us to use hybrid quantum classical optimization. More specifically, we choose

a PQC circuit architecture $(\Theta, \mathcal{D}, O)$ – consisting of data-encoding gates, trainable gates and measurement operator – and then variationally optimize over the trainable parameters. We can summarize this as follows.

Algorithm 1: (Variational QML) Choose a PQC architecture $(\Theta, \mathcal{D}, O)$ and optimize over the parameters $\theta \in \Theta$. The output is some linear function $f_\theta \in \mathcal{F}_{(\Theta,\mathcal{D},O)}$.

We note that Algorithm 1 essentially performs a variational search (typically via gradient based optimization) through $\mathcal{F}_{(\Theta,\mathcal{D},O)}$, which as per Lemma 4, is some parameterized subset of $\mathcal{F}_{\mathcal{D}}$, the set of all linear functions with respect to the feature map $\phi_{\mathcal{D}}$. But this begs the question: Why run Algorithm 1, when we could just do classical linear regression with respect to the feature map $\phi_{\mathcal{D}}$? More specifically, instead of running Algorithm 1, why not just run the following purely *classical* algorithm:

Algorithm 2: (Classical linear regression over $\mathcal{F}_{\mathcal{D}}$)) Given a PQC architecture $(\Theta, \mathcal{D}, O)$, construct $\mathcal{D}$ and the feature map $\phi_{\mathcal{D}}$, and then perform linear regression with respect to the feature map $\phi_{\mathcal{D}}$. The output is some $f_v \in \mathcal{F}_{\mathcal{D}}$.

Unfortunately, Algorithm 2 has the following shortcomings:

1. **Exponential complexity**: Recall that $\phi_{\mathcal{D}} \colon \mathcal{X} \to \mathbb{R}^{|\tilde{\Omega}_{\mathcal{D}}|}$. As such, the space and time complexity of Algorithm 2 is $\mathcal{O}(n|\tilde{\Omega}_{\mathcal{D}}|)$ and $\mathcal{O}(n|\tilde{\Omega}_{\mathcal{D}}|^2 + |\tilde{\Omega}_{\mathcal{D}}|^3)$, respectively. Unfortunately, as detailed in Table 1 of Ref. [CGM+21] and discussed in Section 4.4, the Cartesian product structure of $\tilde{\Omega}_{\mathcal{D}}$ results in a curse of dimensionality which leads to $|\tilde{\Omega}_{\mathcal{D}}|$ scaling *exponentially* with respect to the dimension of the input data d (which gives the size of the problem). For example, if one uses a data encoding strategy consisting only of Pauli Hamiltonians, and if each component is encoded via L encoding gates, then one obtains $|\tilde{\Omega}_{\mathcal{D}}| = \mathcal{O}\left(L^d\right)$.

2. **Potentially poor generalization**: As we have noted in Eq. (27), and illustrated in Figure. 1, due to the constrained depth/expressivity of the trainable parts of any PQC architecture which uses the data-encoding strategy $\mathcal{D}$, we have that

$$\mathcal{F}_{(\Theta,\mathcal{D},O)} \subset \mathcal{F}_{\mathcal{D}}, \tag{31}$$

   i.e., that $\mathcal{F}_{(\Theta,\mathcal{D},O)}$ is a *subset* of $\mathcal{F}_{\mathcal{D}}$. Now, let the output of Algorithm 1 be some $f_\theta \in \mathcal{F}_{(\Theta,\mathcal{D},O)}$ and the output of Algorithm 2 be some $f_v \in \mathcal{F}_{\mathcal{D}}$. Due to the fact that linear regression is a perfect empirical risk minimizer, and the inclusion in Eq. (31), we are guaranteed that

$$\hat{R}(f_v) \leq \hat{R}(f_\theta), \tag{32}$$

   i.e., the *empirical risk* achieved by Algorithm 2 will always be better than the empirical risk achieved by Algorithm 1. However, it could be that Algorithm 2 "overfits" – more precisely, it could be the case that the *true risk* achieved by f_θ is better than that achieved by f_v, i.e., that

$$R(f_v) \geq R(f_\theta). \tag{33}$$

   Said another way, the PQC architecture results in an *inductive bias*, which constrains the set of linear functions which are accessible to Algorithm 1. As illustrated in Figure 1, it could be the case that this inductive bias leads the output of Algorithm 1 to generalize better than that of Algorithm 2.

In light of the above, the natural question is then as follows.

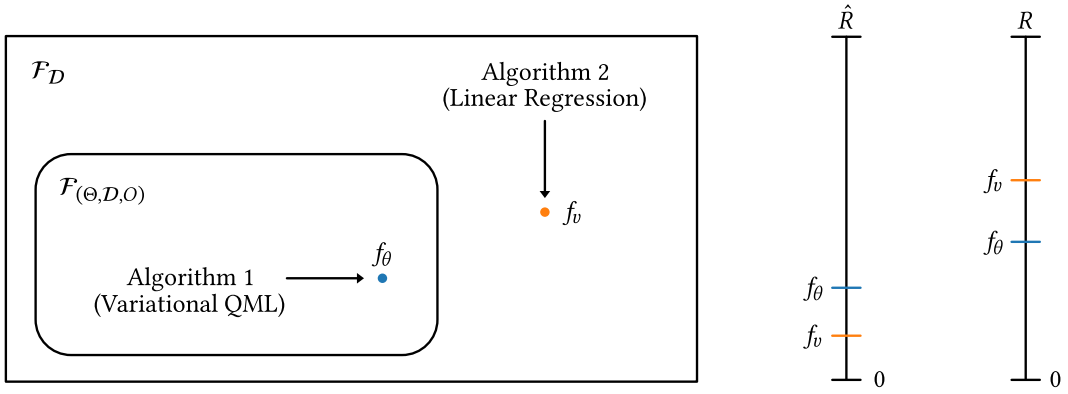


Figure 1: Illustration of the relationship between $\mathcal{F}_{\mathcal{D}}$ and $\mathcal{F}_{(\Theta,\mathcal{D},O)}$, and the output of Algorithms 1 and 2. In particular, we always have that $\mathcal{F}_{(\Theta,\mathcal{D},O)} \subset \mathcal{F}_{\mathcal{D}}$. As a consequence of this, and the fact that Algorithm 2 is a perfect empirical risk minimizer, we always have that $\hat{R}(f_v) \leq \hat{R}(f_\theta)$. However, it might be the case that $R(f_v) \geq R(f_\theta)$.

Question 1: (Existence of efficient linear regression) *Can we modify Algorithm 2 (classical linear regression) such that it is both efficient with respect to d, and with high probability outputs a function which is just as good as the output of Algorithm 1 (variational QML), with respect to the true risk?*

As we have already hinted at in the preliminaries, one possible approach – at least for addressing the issue of poor complexity – is to use random Fourier features to approximate the classical PQC-kernel $K_{\mathcal{D}}(x, x') = \langle \phi_{\mathcal{D}}(x), \phi_{\mathcal{D}}(x') \rangle$ which is used implicitly in Algorithm 2. Indeed, this has already been suggested and explored in Ref. [LTD+22]. More specifically Ref. [LTD+22], at a very high level, suggests the following algorithm:

Algorithm 3: (Classical linear regression over $\mathcal{F}_{\mathcal{D}}$ with random Fourier features) Given a data-encoding strategy $\mathcal{D}$, implement RFF-based λ-regularized linear regression using the classical PQC-kernel $K_{\mathcal{D}}$, and obtain some function $h_\nu \in \mathcal{F}_{\mathcal{D}}$. More specifically:

1. Sample M "features" $\nu_i = (\omega_i, \gamma_i) \in \mathbb{R}^d \times [0, 2\pi)$ from the distribution $\pi = p_{\mathcal{D}} \times \mu$, which as per Eq. (13) is the product distribution appearing in the integral representation of the shift-invariant kernel $K_{\mathcal{D}}$.
2. Construct the randomized feature map $\tilde{\phi}_M(x) = \frac{1}{\sqrt{M}} (\psi(x, \nu_1), \ldots, \psi(x, \nu_M))$, where $\psi(x, \nu) = \sqrt{2} \cos(\langle \omega, x \rangle + \gamma)$.
3. Implement λ-regularized linear regression with respect to the feature map $\tilde{\phi}_M$.

In the above description of the algorithm we have omitted details of how to sample frequencies from the distribution $p_{\mathcal{D}}$, which will depend on the kernel $K_{\mathcal{D}}$. We note that in order for Algorithm 3 to be efficient with respect to d, it is necessary for this sampling procedure to be efficient with respect to d. For simplicity, we assume for now that this is the case, and postpone a detailed discussion of this issue to Section 4.4. As discussed in Section 2.3, the space and time complexity of Algorithm 3 (post-sampling) is $\mathcal{O}(nM)$ and $\mathcal{O}(nM^2 + M^3)$, respectively, where M is the number of frequencies which have been sampled. Given this setup, the natural question is then as follows:

Question 2: (Efficiency of RFF for PQC dequantization) *Given a regression problem P and a circuit architecture $(\Theta, \mathcal{D}, O)$, how many data samples n and frequency samples M are necessary to ensure that, with high probability, the true risk of the function h_v output by Algorithm 3 (RFF) is no more than ϵ worse than the true risk of the function f_θ output by Algorithm 1 (variational QML) – i.e., to ensure that $R(h_\nu) - R(f_\theta) \leq \epsilon$?*

Said another way, Question 2 is asking when classical RFF-based regression (Algorithm 3) can be used to efficiently *dequantize* variational QML (Algorithm 1). In Ref. [LTD+22] the authors addressed a similar question, but with two important differences:

1. It was implicitly assumed that $(\Theta, \mathcal{D}, O)$ is universal – i.e., that using the PQC one can realize all functions in $\mathcal{F}_{(\mathcal{D},O)}$. Recall however from the discussion above that we are precisely interested in the case in which $(\Theta, \mathcal{D}, O)$ is not universal – i.e., the case in which $\mathcal{F}_{(\Theta,\mathcal{D},O)} \subset \mathcal{F}_{(\mathcal{D},O)}$ due to constraints on the circuit architecture. This is for two reasons: Firstly, because this is the case for practically realizable near-term circuit architectures. Secondly, because it is in this regime in which the circuit architecture induces an *inductive bias* which may lead to better generalization than the output of linear regression over $\mathcal{F}_{\mathcal{D}}$.

2. Instead of considering true risk, Ref. [LTD+22] considered the complexity necessary to achieve $|h_\nu(x) - f_\theta(x)| \leq \epsilon$ for all x. We note that this is *stronger* than $|R(h_\nu) - R(f_\theta)| \leq \epsilon$, due to the latter comparing the functions with respect to the data distribution P. Here we are concerned with the latter, which is the typical goal in statistical learning theory.

In light of these considerations, we proceed in Section 4 to provide answers to Question 2.

Note on recent related work: As per the discussion above, the two motivations for introducing Algorithm 3 were the poor efficiency and potentially poor generalization associated to Algorithm 2. However, we note that in Ref. [STJ24], which appeared recently, the authors show via a tensor network (TN) analysis that to every PQC architecture one can associate a feature map – different from the PQC feature map $\phi_{\mathcal{D}}$ – for which (a) all functions in the PQC model class are linear with respect to the feature map, and (b) the associated kernel *can* be evaluated efficiently classically. As such, using this feature map, for any number of data-samples n, one *can* run Algorithm 2 classically efficiently with respect to d – i.e., there is no need to approximate the kernel via RFF! Indeed, this approach to dequantization is suggested by the authors of Ref. [STJ24]. However as above, and discussed in Ref. [STJ24], this does not immediately yield an efficient dequantization of variational QML, due to the potentially poor generalization of linear regression over the entire function space. In light of this, the generalization of linear regression with respect to the tensor network kernel introduced in Ref. [STJ24] certainly deserves attention, and we hope that the methods and tools introduced in this work, and others such as Ref. [KR21], can be useful in that regard.

4 Generalization and efficiency guarantees for RFF-based linear regression

In this section, we attempt to provide rigorous answers to Question 2 – i.e., for which PQC architectures and for which regression problems does RFF-based linear regression yield an efficient dequantization of variational QML?

In particular, this section is structured as follows: We begin in Section 4.1 with a brief digression, providing definitions for some important kernel notions. With these in hand, we continue in Section 4.2 to state Theorem 1 which provides a concrete answer to Question 2. In particular, Theorem 1 provides an upper bound on both the number of data samples n, and the number of

frequency samples M, which are sufficient to ensure that, with high probability, the output of RFF-based linear regression with respect to the kernel $K_{\mathcal{D}}$ is a good approximation to the best possible PQC function, with respect to true risk. As we will see, these upper bounds depend crucially on two quantities which are defined in Section 4.1, namely the operator norm of the kernel integral operator associated to $K_{\mathcal{D}}$, and the reproducing kernel Hilbert space norm of the optimal PQC function.

Given the results of Section 4.2, in order to make any concrete statements it is necessary to gain a deeper quantitative understanding of both the operator norm of the kernel integral operator and the RKHS norm of functions in the PQC model class. However, before doing this, we show in Section 4.3 that the PQC feature map $\phi_{\mathcal{D}}$ is in fact a special instance of an entire family of feature maps – which we call "re-weighted PQC feature maps" – and that Theorem 1 holds not only for $K_{\mathcal{D}}$, but for the kernel induced by any such re-weighted feature map. With this in hand, we then proceed in Section 4.4 to show how the re-weighting determines the distribution over frequencies π from which one needs to sample in the RFF procedure, and we discuss in detail for which feature maps/kernels one can and cannot *efficiently* sample from this distribution. This immediately allows us to rule out efficient RFF dequantization for a large class of re-weighted PQC feature maps, and therefore allows us to focus our attention on only those feature maps (re-weightings) which admit efficient sampling procedures.

With this knowledge, we then proceed in Sections 4.5 and 4.6 to discuss in detail the quantitative behaviour of the kernel integral operator and the RKHS norm of PQC functions, for different re-weighted PQC kernels. This allows us to place further restrictions on the circuit architectures and re-weightings which yield efficient PQC dequantization via Theorem 1. Finally, in Section 4.7 we show that the properties we have identified in Sections 4.5 and 4.6 as *sufficient* for the application of Theorem 1 are in some sense *necessary*. In particular, we prove *lower bounds* on the number of frequency samples necessary to achieve a certain *average* error via RFF, and use this to delineate rigorously when efficient dequantization via RFF is *not* possible.

4.1 Preliminary kernel theory

In order to present Theorem 1 we require a few definitions. To start, we need the notion of a *reproducing kernel Hilbert space* (RKHS), and the associated RKHS norm.

Definition 2: (RKHS and RKHS norm) Given a kernel $K\colon \mathcal{X}\times\mathcal{X}\to\mathbb{R}$ we define the associated reproducing kernel Hilbert space (RKHS) as the tuple $(\mathcal{H}_K, \langle\cdot,\cdot\rangle_K)$ where $\mathcal{H}_K$ is the set of functions defined as the completion (including limits of Cauchy series) of

$$\operatorname{span}\{K_x(\cdot) \coloneqq K(x,\cdot)\,|\,x\in\mathcal{X}\}, \tag{34}$$

and $\langle\cdot,\cdot\rangle_K$ is the inner product on $\mathcal{H}_K$ defined via

$$\langle g,h\rangle_{\mathcal{H}_K} \coloneqq \sum_{i,j}\alpha_i\beta_j K(x_i,x_j) \tag{35}$$

for any two functions $g=\sum_i \alpha_i K_{x_i}$ and $h=\sum_j \beta_j K_{x_j}$ in $\mathcal{H}_K$. This inner product then induces the RKHS norm $\|\cdot\|_K$ defined via

$$\|g\|_K \coloneqq \sqrt{\langle g,g\rangle_{\mathcal{H}_K}}. \tag{36}$$

It is crucial to note that for two kernels K_1 and K_2 it may be the case that $\mathcal{H}_{K_1}=\mathcal{H}_{K_2}$ but $\|\cdot\|_{K_1}\neq\|\cdot\|_{K_2}$. We will make heavy use of this fact shortly. In particular, the re-weighted

PQC kernels introduced in Section 4.3 have precisely this property. In addition to the definition above, we also need a definition of the kernel integral operator associated to a kernel, which we define below.

Definition 3: (Kernel integral operator) Given a kernel $K\colon \mathcal{X}\times\mathcal{X}\to\mathbb{R}$ and a probability distribution $P_{\mathcal{X}}$ over $\mathcal{X}$, we start by defining the space of square-integrable functions with respect to $P_{\mathcal{X}}$ via

$$L^2(\mathcal{X},P_{\mathcal{X}}) = \left\{ f\in\mathbb{R}^{\mathcal{X}} \text{ such that } \int_{\mathcal{X}} |f(x)|^2 \,\mathrm{d}P_{\mathcal{X}}(x) < \infty \right\}. \tag{37}$$

The kernel integral operator $T_K\colon L^2(\mathcal{X},P_{\mathcal{X}})\to L^2(\mathcal{X},P_{\mathcal{X}})$ is then defined via

$$(T_K g)(x) = \int_{\mathcal{X}} K(x,x')g(x')\,\mathrm{d}P_{\mathcal{X}}(x') \tag{38}$$

for all $g\in L^2(\mathcal{X},P_{\mathcal{X}})$.

In addition, we note that we will mostly be concerned with the *operator norm* of the kernel integral operator, which we denote with $\|T_K\|$ – i.e., when no subscript is used to specify the norm, we assume the operator norm.

4.2 Efficiency of RFF for matching variational QML

Given the definitions of Section 4.1, we proceed in this section to state Theorem 1, which provides insight into the number of data samples n and the number of frequency samples M which are sufficient to ensure that, with probability at least $1-\delta$, the true risk of the output hypothesis of RFF-based linear regression is no more than ϵ worse than the true risk of the best possible function realizable by the PQC architecture. To this end, we require first a preliminary definition of the "best possible PQC function":

Definition 4: (Optimal PQC function) Given a regression problem $P\sim\mathcal{X}\times\mathbb{R}$, and a PQC architecture $(\Theta,\mathcal{D},O)$, we define $f^*_{(\Theta,\mathcal{D},O)}$, the optimal PQC model for P (with respect to *true risk*), via

$$f^*_{(\Theta,\mathcal{D},O)} = \operatorname*{arg\,min}_{f\in\mathcal{F}_{(\Theta,\mathcal{D},O)}} \left[R(f)\right]. \tag{39}$$

With this in hand we can state Theorem 1. We note however that this follows via a straightforward application of the RFF generalization bounds provided in Ref. [RR17] to the PQC-kernel, combined with the insight that the set of PQC functions $\mathcal{F}_{(\Theta,\mathcal{D},O)}$ is contained within $\mathcal{H}_{K_{\mathcal{D}}}$, the RKHS associated with the PQC kernel. A detailed proof is provided in Appendix B. Additionally we note that the generalization bound given in Theorem 1 below holds for 2-norm *regularized* RFF-based linear regression – as per Eq. (4) – with a a very specific regularization $\lambda = 1/\sqrt{n}$. This regularization is inherited from the proof of the generalization bounds in Ref. [RR17], and we refer to there for details.

Theorem 1: (RFF vs. variational QML) *Let R be the risk associated with a regression problem $P\sim\mathcal{X}\times\mathbb{R}$. Assume the following:*

1. $\|f^*_{(\Theta,\mathcal{D},O)}\|_{K_{\mathcal{D}}} \leq C$
2. $|y|\leq b$ *almost surely when* $(x,y)\sim P$, *for some* $b>0$.

Additionally, define

$$n_0 \coloneqq \max\left\{4\|T_{K_\mathcal{D}}\|^2, \left(528\log\frac{1112\sqrt{2}}{\delta}\right)^2\right\}, \tag{40}$$

$$c_0 \coloneqq 36\left(3+\frac{2}{\|T_{K_\mathcal{D}}\|}\right), \tag{41}$$

$$c_1 \coloneqq 8\sqrt{2}(4b+\frac{5}{\sqrt{2}}C+2\sqrt{2C}). \tag{42}$$

Then, let $\delta \in (0,1]$, $\epsilon > 0$, $n \geq n_0$, *set* $\lambda_n = 1/\sqrt{n}$, *and let* $\hat{f}_{M_n,\lambda_n}$ *be the output of* λ_n*-regularized linear regression with respect to the feature map*

$$\phi_{M_n}(x) = \frac{1}{\sqrt{M_n}}(\psi(x,\nu_1),\ldots,\psi(x,\nu_{M_n})) \tag{43}$$

constructed from the integral representation of $K_\mathcal{D}$ *by sampling* M_n *elements from* π. *Then, with probability at least* $1-\delta$, *one achieves*

$$R(\hat{f}_{M_n,\lambda_n}) - R(f^*_{(\Theta,\mathcal{D},O)}) \leq \epsilon, \tag{44}$$

by ensuring that

$$n \geq \max\left\{\frac{c_1^2\log^4\frac{1}{\delta}}{\epsilon^2}, n_0\right\} \tag{45}$$

and

$$M \geq c_0\sqrt{n}\log\frac{108\sqrt{n}}{\delta}. \tag{46}$$

Let us now try to unpack what insights can be gained from Theorem 1. Firstly, recall that here we consider $\mathcal{X} \subseteq \mathbb{R}^d$, in which case d provides the relevant asymptotic scaling parameter. With this in mind, in order to gain intuition, assume for now that n_0, c_0 and c_1 are constants with respect to d. Assume additionally that one can sample efficiently from π. In this case we see via Theorem 1 that both the number of data points n, and the number of samples M, which are sufficient to ensure that with probability $1-\delta$ there is a gap of at most ϵ between the true risk of PQC optimization and the true risk of RFF, is independent of d, and polylogarithmic in both $1/\epsilon$ and $1/\delta$. Given that the time and space complexity of RFF-based linear regression is $\mathcal{O}(nM)$ and $\mathcal{O}(nM^2+M^3)$, respectively, in this case Theorem 1 guarantees that Algorithm 3 (RFF-based linear regression) provides an efficient dequantization of variational QML.

However, in general n_0, c_0 and c_1 will *not* be constants, and one will *not* be able to sample efficiently from π. In particular, n_0, c_0 and c_1 depend on both $\|T_{K_\mathcal{D}}\|$, the operator norm of the kernel integral operator, and C, the upper bound on the RKHS norm of the optimal PQC function. Given the form of n_0, c_0 and c_1, in order for Theorem 1 to yield a polynomial upper bound on n and M, it is sufficient that $\|T_{K_\mathcal{D}}\| = \Omega(1/\text{poly}(d))$ and that $C = \mathcal{O}(\text{poly}(d))$.

Summary: In order to use Theorem 1 to make any concrete statement concerning the efficiency of RFF for dequantizing variational QML, we need to obtain the following.

1. *Lower bounds* on $\|T_{K_\mathcal{D}}\|$, the operator norm of the kernel integral operator.
2. *Upper bounds* on $\|f^*_{(\Theta,\mathcal{D},O)}\|_{K_\mathcal{D}}$, the RKHS norm of the optimal PQC function.
3. An understanding of the complexity of sampling from π, the distribution appearing in the integral representation of $K_\mathcal{D}$.

We address the above requirements in the following sections. However, before doing that, we note in Section 4.3 below that Theorem 1 applies not only to $K_{\mathcal{D}}$, but to an entire family of "re-weighted" kernels, of which $K_{\mathcal{D}}$ is a specific instance. We will see that this is important as different re-weightings will lead to substantially different sampling complexities, as well as lower and upper bounds on $\|T_{K_{\mathcal{D}}}\|$ and C, respectively.

4.3 Re-weighted PQC kernels

As mentioned in Section 4.2, the proof of Theorem 1 is essentially a straightforward application of the RFF generalization bound from Ref. [RR17] to the classical PQC kernel $K_{\mathcal{D}}$. In particular, as shown in Appendix. B, the key insight that allows one to leverage the generalization bound from Ref. [RR17] into a bound on the difference in true risk between the output of variational QML and the output of RFF-based linear regression with the PQC kernel $K_{\mathcal{D}}$, is the fact that $\mathcal{F}_{(\Theta,\mathcal{D},O)} \subseteq \mathcal{H}_{K_{\mathcal{D}}}$ – i.e., the fact that the set of all PQC functions $\mathcal{F}_{(\Theta,\mathcal{D},O)}$ is contained within the RKHS associated with the PQC kernel $\mathcal{H}_{K_{\mathcal{D}}}$. With this in mind we can ask whether there exists any *other* kernel K for which $\mathcal{F}_{(\Theta,\mathcal{D},O)} \subseteq \mathcal{H}_K$, as Theorem 1 would then immediately also hold for RFF-based linear regression via K. The motivation for asking this question is the following.

1. As discussed in Section. 4.2 above, the number of sufficient data points n and the number of sufficient samples M specified by Theorem 1 depends on the RKHS norm and the kernel integral operator respectively, both of which depend on the kernel. As such, by using a different kernel, one may require less data-points n and less samples M to get the desired guarantee on the output of RFF-based linear regression.

2. The implementation of RFF based linear regression requires one to sample from the distribution π, which for shift invariant kernels is proportional to the Fourier transform of the kernel. In practice, sampling from this distribution for the kernel $K_{\mathcal{D}}$ may be hard, and using a different kernel might allow us to sample from a different distribution, for which efficient sampling is possible.

In this section we show that there is indeed a whole family of kernels – which we call *re-weighted* PQC kernels – for which the set of all PQC functions $\mathcal{F}_{(\Theta,\mathcal{D},O)}$ is contained within the RKHS of the kernel, and for which Theorem 1 is valid. This will turn out to be important and useful for a variety of reasons. Firstly, in Section 4.4 we will show that indeed, sampling from the distribution π associated with the original PQC kernel $K_{\mathcal{D}}$ can *not* be done efficiently, whereas sampling from the distribution associated with certain re-weighted kernels can be done efficiently. Additionally, we then show in Sections 4.5 and 4.6 that for certain regression problems, using the original PQC kernel $K_{\mathcal{D}}$ for RFF-based linear will *not* give an efficient dequantization method, whereas certain re-weighted kernels will.

We stress that in typical applications of RFF the kernel is fixed, however for the purposes of dequantization via RFF we do not necessarily need to restrict ourselves to any particular kernel and can therefore exploit the fact that Theorem 1 is valid for *any* kernel k for which $\mathcal{F}_{(\Theta,\mathcal{D},O)} \subseteq \mathcal{H}_K$ to *choose* the optimal kernel for dequantization. Here we discuss the family of "re-weighted" PQC kernels for which $\mathcal{F}_{(\Theta,\mathcal{D},O)} \subseteq \mathcal{H}_K$, and in subsequent sections we provide insights into how to make a well-informed choice as to which kernel should be used for dequantization of a specific PQC architecture and regression problem.

Given this motivation, we can now make the notion of a re-weighted PQC kernel precise. For any "re-weighting vector" $\mathrm{w} \in \mathbb{R}^{|\Omega_{\mathcal{D}}|}$, we define the re-weighted PQC feature map via

$$\phi_{(\mathcal{D},\mathrm{w})}(x) \coloneqq \frac{1}{\|\mathrm{w}\|_2}\big(\mathrm{w}_0, \mathrm{w}_1\cos(\langle\omega_1, x\rangle), \mathrm{w}_1\sin(\langle\omega_1, x\rangle), \ldots, \\ \ldots, \mathrm{w}_{|\Omega^+_{\mathcal{D}}|}\cos\Big(\langle\omega_{|\Omega^+_{\mathcal{D}}|}, x\rangle\Big), \mathrm{w}_{|\Omega^+_{\mathcal{D}}|}\sin\Big(\langle\omega_{|\Omega^+_{\mathcal{D}}|}, x\rangle\Big)\big), \tag{47}$$

along with the associated set of linear functions with respect to $\phi_{(\mathcal{D},\mathrm{w})}$, defined via

$$\mathcal{F}_{(\mathcal{D},\mathrm{w})} = \{f_\theta(\cdot) = \langle v, \phi_{(\mathcal{D},\mathrm{w})}(\cdot)\rangle \,|\, v \in \mathbb{R}^{|\tilde{\Omega}_{\mathcal{D}}|}\}, \tag{48}$$

and the associated re-weighted PQC kernel

$$K_{(\mathcal{D},\mathrm{w})}(x, x') \coloneqq \langle\phi_{(\mathcal{D},\mathrm{w})}(x), \phi_{(\mathcal{D},\mathrm{w})}(x')\rangle. \tag{49}$$

Note that we recover the previous definitions when w is the vector of all 1's (in which case $\|\mathrm{w}\|_2 = \sqrt{|\Omega_{\mathcal{D}}|}$). With this in hand, as per the motivating discussion above, we would now like to show that there exists a set of re-weighting vectors w such that $\mathcal{F}_{(\Theta,\mathcal{D},O)} \subseteq \mathcal{H}_{K_{(\mathcal{D},\mathrm{w})}}$ for all w in the set. To this end, we start with the following observation:

Observation 1: (Invariance of $\mathcal{F}_{\mathcal{D}}$ under non-zero feature map re-weighting) For a re-weighting vector $\mathrm{w} \in \mathbb{R}^{|\Omega_{\mathcal{D}}|}$ satisfying $\mathrm{w}_i \neq 0$ for all $i \in [|\Omega_{\mathcal{D}}|]$, we have that

$$\mathcal{F}_{(\mathcal{D},\mathrm{w})} = \mathcal{F}_{\mathcal{D}}. \tag{50}$$

Proof. Define the matrix M_{w} as the matrix with w as its diagonal, and the matrix

$$M \coloneqq \left(\sqrt{|\Omega|_{\mathcal{D}}}/\|\mathrm{w}\|_2\right) M_{\mathrm{w}}. \tag{51}$$

By the assumptions on w, the matrix M is invertible. Now, let $f \in \mathcal{F}_{(\mathcal{D},\mathrm{w})}$. We have that

$$f(\cdot) = \langle v, \phi_{(\mathcal{D},\mathrm{w})}(\cdot)\rangle = \langle v, M\phi_{\mathcal{D}}(\cdot)\rangle = \langle vM, \phi_{\mathcal{D}}(\cdot)\rangle = \langle\tilde{v}, \phi_{\mathcal{D}}(\cdot)\rangle, \tag{52}$$

i.e., $f \in \mathcal{F}_{\mathcal{D}}$, and, therefore, $\mathcal{F}_{(\mathcal{D},\mathrm{w})} \subset \mathcal{F}_{\mathcal{D}}$. Similarly, for all $g \in \mathcal{F}_{\mathcal{D}}$ we have that

$$g(\cdot) = \langle v, \phi_{\mathcal{D}}(\cdot)\rangle = \langle v, M^{-1}M\phi_{\mathcal{D}}(\cdot)\rangle = \langle vM^{-1}, M\phi_{\mathcal{D}}(\cdot)\rangle = \langle\tilde{v}, \phi_{(\mathcal{D},\mathrm{w})}(\cdot)\rangle, \tag{53}$$

i.e., $g \in \mathcal{F}_{(\mathcal{D},\mathrm{w})}$, and hence $\mathcal{F}_{\mathcal{D}} \subset \mathcal{F}_{(\mathcal{D},\mathrm{w})}$. □

Now, the fact that for any feature map the set of all linear functions with respect to the feature map is a subset of the RKHS of the associated kernel [SC08], tells us that $\mathcal{F}_{(\mathcal{D},\mathrm{w})} \subseteq \mathcal{H}_{K_{(\mathcal{D},\mathrm{w})}}$ for all re-weighting vectors w. Combing this with Observation 1 then gives

$$\mathcal{F}_{(\Theta,\mathcal{D},O)} \subset \mathcal{F}_{\mathcal{D}} = \mathcal{F}_{(\mathcal{D},\mathrm{w})} \subseteq \mathcal{H}_{K_{(\mathcal{D},\mathrm{w})}}. \tag{54}$$

for all re-weighting vectors with no zero elements. As such we indeed have $\mathcal{F}_{(\Theta,\mathcal{D},O)} \subseteq \mathcal{H}_{K_{(\mathcal{D},\mathrm{w})}}$ for all w with no zero elements. As discussed above, we therefore also have that Theorem 1 actually holds for *any* PQC kernel $K_{(\mathcal{D},\mathrm{w})}$, re-weighted via a re-weighting vector w with no zero elements, and as such we can *choose* which such kernel to use for our dequantization procedure. We note that allowing re-weighting vectors with zero elements has the effect of "shrinking" the set $F_{(\mathcal{D},\mathrm{w})}$, which might result in the existence of functions f satisfying both $f \in \mathcal{F}_{(\Theta,\mathcal{D},O)}$ and $f \notin F_{(\mathcal{D},\mathrm{w})}$. Intuitively, this is problematic because for regression problems in which f is the optimal solution, we know that the PQC architecture $(\Theta, \mathcal{D}, O)$ can realize f, but we cannot

hope for the RFF procedure via $K_{(\mathcal{D},\mathrm{w})}$ to do the same, as it is limited to hypotheses within $F_{(\mathcal{D},\mathrm{w})}$.

In light of the above observations, and as discussed above, from this point on we can broaden our discussion of the application of Theorem 1 to include all appropriately re-weighted PQC kernels. This insight is important because of the following.

1. We will see that while all appropriately re-weighted PQC kernels give rise to the same function set $\mathcal{F}_{\mathcal{D}}$, they give rise to *different* RKHS norms. As a result, the RKHS norm of the optimal PQC function – which as we have seen is critical to the complexity of the RFF method – will depend on which re-weighting we choose.

2. Similarly, we will see that the operator norm of the kernel integral operator – and therefore again the complexity of RFF linear regression – depends heavily on the re-weighting chosen.

3. We will see that the re-weighting of the PQC kernel completely determines the probability distribution π, from which it is necessary to sample in order to implement RFF-based linear regression. Therefore, once again, the efficiency of RFF will depend on the re-weighting chosen. This perspective will also allow us to see why zero elements are not allowed in the re-weighting vector. Namely, because doing this will cause the probability of sampling the associated frequency to be zero, which is problematic if that frequency is required to represent the regression function.

4.4 RFF implementation

Recall from Section 2.3, and from our presentation of Algorithm 3 in Section 3, that when given a shift-invariant kernel K, in order to implement RFF-based linear regression, one has to sample from the probability measure $\pi = p \times \mu$, which one obtains from the Fourier transform of K. As such, given a re-weighted PQC kernel $K_{(\mathcal{D},\mathrm{w})}$, we need to

1. understand the structure of the probability distribution p, which should depend on both $\mathcal{D}$ and the re-weighting vector w, and

2. understand when and how – i.e., for which data-encoding strategies and which re-weightings – one can efficiently sample from p.

Let us begin with point 1. To this end we start by noting that the re-weighted PQC kernels $K_{(\mathcal{D},\mathrm{w})}$ have a particularly simple integral representation, from which one can read off the required distribution. In particular, note that

$$
\begin{aligned}
K_{(\mathcal{D},\mathrm{w})}(x,x') &= \langle \phi_{(\mathcal{D},\mathrm{w})}(x), \phi_{(\mathcal{D},\mathrm{w})}(x') \rangle \\
&= \frac{1}{\|\mathrm{w}\|_2^2} \left(\mathrm{w}_0^2 + \sum_{i=1}^{|\Omega_{\mathcal{D}}^+|} \mathrm{w}_i^2 \left[\cos(\langle \omega_i, x \rangle) \cos(\langle \omega_i, x' \rangle) + \sin(\langle \omega_i, x \rangle) \sin(\langle \omega_i, x' \rangle) \right] \right) \qquad (55) \\
&= \frac{1}{\|\mathrm{w}\|_2^2} \sum_{i=0}^{|\Omega_{\mathcal{D}}^+|} \mathrm{w}_i^2 \left[\cos(\langle \omega_i, x \rangle) \cos(\langle \omega_i, x' \rangle) + \sin(\langle \omega_i, x \rangle) \sin(\langle \omega_i, x' \rangle) \right] \\
&= \frac{1}{\|\mathrm{w}\|_2^2} \sum_{i=0}^{|\Omega_{\mathcal{D}}^+|} \mathrm{w}_i^2 \left[\cos(\langle \omega_i, (x - x') \rangle) \right], \\
&= \frac{1}{2\pi} \int_{\mathcal{X}} \int_0^{2\pi} \sqrt{2} \cos(\langle \omega, x \rangle + \gamma) \sqrt{2} \cos(\langle \omega, x' \rangle + \gamma) q_{(\mathcal{D},\mathrm{w})}(\omega) \, \mathrm{d}\gamma \mathrm{d}\nu, \qquad (56)
\end{aligned}
$$

where

$$q_{(\mathcal{D},\mathrm{w})}(\omega) = \sum_{i=0}^{|\Omega_{\mathcal{D}}^{+}|} \frac{\mathrm{w}_i^2}{\|\mathrm{w}\|_2^2} \delta(\omega - \omega_i), \tag{57}$$

and δ is the Dirac delta function. By comparison with Eq. (13) we, therefore, see that $\pi = q_{(\mathcal{D},\mathrm{w})} \times \mu$, where as before, μ is the uniform distribution over $[0, 2\pi)$. For convenience, we refer to $q_{(\mathcal{D},\mathrm{w})}$ as the probability distribution associated to $K_{(\mathcal{D},\mathrm{w})}$.

Let us now move on to point 2 – in particular, for which data encoding strategies and re-weighting vectors can we *efficiently* sample from $q_{(\mathcal{D},\mathrm{w})}$? Firstly, note that sampling from the *continuous* distribution $q_{(\mathcal{D},\mathrm{w})}$ can be done by sampling from the *discrete* distribution $p_{(\mathcal{D},\mathrm{w})}$ over $\Omega_{\mathcal{D}}$ defined via

$$p_{(\mathcal{D},\mathrm{w})}(\omega_i) = \frac{\mathrm{w}_i^2}{\|\mathrm{w}\|_2^2} \tag{58}$$

for all $\omega_i \in \Omega_{\mathcal{D}}$. As a result, from this point on we focus on the distribution $p_{(\mathcal{D},\mathrm{w})}$, and when clear from the context we drop the subscript and just use p to refer to $p_{(\mathcal{D},\mathrm{w})}$. Also, we note that we can in principle just work directly with the choice of probability distribution, as opposed to the underlying weight vector, as we know for any probability distribution over $\Omega_{\mathcal{D}}$ there exists an appropriate weight vector.

As p is a distribution over $\Omega_{\mathcal{D}}$, in order to discuss the efficiency of sampling from p, it is necessary to briefly recall some facts about the sets $\Omega_{\mathcal{D}}$ and $\tilde{\Omega}_{\mathcal{D}}$. In particular, as discussed in Appendix A, given a data-encoding strategy $\mathcal{D} = \left(\mathcal{D}^{(1)}, \ldots, \mathcal{D}^{(d)}\right)$, where $\mathcal{D}^{(j)}$ contains the Hamiltonians used to encode the j'th data component x_j, we know that

$$\tilde{\Omega}_{\mathcal{D}} = \tilde{\Omega}_{\mathcal{D}}^{(1)} \times \ldots \times \tilde{\Omega}_{\mathcal{D}}^{(d)}, \tag{59}$$

where $\tilde{\Omega}_{\mathcal{D}}^{(j)} \subset \mathbb{R}$ depends only on $\mathcal{D}^{(j)}$ – i.e., $\tilde{\Omega}_{\mathcal{D}}$ has a Cartesian product structure. Additionally, as discussed in Section 2.4 we know that $\tilde{\Omega}_{\mathcal{D}} := \Omega_{\mathcal{D}} \cup (-\Omega_{\mathcal{D}})$, where $\Omega_{\mathcal{D}} \cap (-\Omega_{\mathcal{D}}) = \{\omega_0\}$. Taken together, we see that

$$|\tilde{\Omega}_{\mathcal{D}}| = \prod_{j=1}^{d} |\tilde{\Omega}_{\mathcal{D}}^{(j)}|, \tag{60}$$

$$|\Omega_{\mathcal{D}}| = \frac{|\tilde{\Omega}_{\mathcal{D}}| - 1}{2} + 1. \tag{61}$$

Now, let us define $N_j := |\tilde{\Omega}_{\mathcal{D}}^{(j)}|$, and make the assumption that N_j is independent of d[1]. Furthermore, let us define $N_{\min} = \min\{N_j\}$. We then have that

$$|\tilde{\Omega}_{\mathcal{D}}| \geq N_{\min}^{d} \tag{62}$$

i.e., that the number of frequencies in $\tilde{\Omega}_{\mathcal{D}}$, and therefore $\Omega_{\mathcal{D}}$, scales *exponentially* with respect to d. From this we can immediately make our first observation:

[1]One can see from the discussion in Appendix A that N_j depends directly only on L_j, the number of encoding gates in $\mathcal{D}^{(j)}$, and the spectra of the encoding Hamiltonians in $\mathcal{D}^{(j)}$. This assumption is therefore justified for all data-encoding strategies in which both L_j and the Hamiltonian spectra are independent of d, which is standard practice. One can see Table 1 in Ref. [CGM+21] for a detailed list of asymptotic upper bounds on N_j for different encoding strategies.

Observation 2: Given that the number of elements in $\Omega_{\mathcal{D}}$ scales exponentially with d, one *cannot* efficiently store and sample from arbitrary distributions supported on $\Omega_{\mathcal{D}}$.

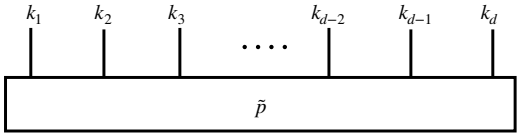
As such, we have to restrict ourselves to *structured distributions*, whose structure facilitates efficient sampling. One such subset of distributions are those which are supported only on a polynomial (in d) size subset of $\Omega_{\mathcal{D}}$. Another suitable set of distributions is what we call *product-induced* distributions. Specifically, let $\tilde{p}^{(j)}$ be an arbitrary distribution over $\tilde{\Omega}_{\mathcal{D}}^{(j)}$, and define the product distribution $\tilde{p}$ over $\tilde{\Omega}_{\mathcal{D}}$ via

$$\tilde{p}(\omega = (\omega_1, \ldots, \omega_d)) = \tilde{p}^{(1)}(\omega_1) \times \ldots \times \tilde{p}^{(d)}(\omega_d). \tag{63}$$

Note that, due to the d-independence of $|\tilde{\Omega}_{\mathcal{D}}^{(j)}|$ we can store and sample from $\tilde{p}^{(j)}$ efficiently, which then allows us to sample from $\tilde{p}$ by simply drawing $\omega_j \sim \tilde{p}^{(j)}$ for all $j \in [d]$ and then outputting $\omega = (\omega_1, \ldots, \omega_d)$. However, it may be the case that $\omega \notin \Omega_{\mathcal{D}}$. As such, the natural thing to do is simply output ω if $\omega \in \Omega_{\mathcal{D}}$, and if not, output $-\omega$. If one does this, then one samples from the distribution p over $\Omega_{\mathcal{D}}$ defined via

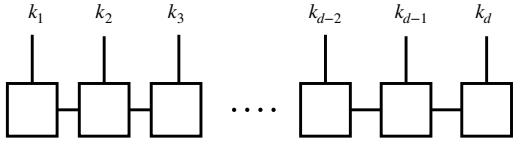
$$p(\omega) := \begin{cases} \tilde{p}(\omega) & \text{if } \omega = \omega_0, \\ \tilde{p}(\omega) + \tilde{p}(-\omega) & \text{else,} \end{cases} \tag{64}$$

which we refer to as a *product-induced* distribution. We can in fact however go further, and use the Cartesian product structure of $\tilde{\Omega}_{\mathcal{D}}$ to generalize product-induced distributions to *matrix-product-state-induced* (MPS-induced) distributions. To do this, let us label the elements of $\tilde{\Omega}_{\mathcal{D}}^{(j)}$ via $\tilde{\Omega}_{\mathcal{D}}^{(j)} = \{\tilde{\Omega}_{k_j}^{(j)}\}$ for $k_j \in [N_j]$. We can then write any $\omega \in \tilde{\Omega}_{\mathcal{D}}$ via $\omega = (\omega_{k_1}^{(1)}, \ldots, \omega_{k_d}^{(d)})$, for some indexing $(k_1, \ldots, k_d)$. From this, we see that *any* distribution $\tilde{p}$ over $\tilde{\Omega}_{\mathcal{D}}$ can be naturally represented as a d-tensor - i.e., as as a tensor with d legs, where the j'th leg is N_j dimensional. Graphically, we have that

$\tilde{p}[(\omega_{k_1}^{(1)}, \ldots, \omega_{k_d}^{(d)})] =$ 

(65)

Now we can consider the subset of distributions which can be represented by a *matrix product state* [Sch11], i.e., those distributions for which

$\tilde{p}[(\omega_{k_1}^{(1)}, \ldots, \omega_{k_d}^{(d)})] =$ 

(66)

We refer to distributions which admit such a representation as MPS distributions [FV12; SW10; GSP+19]. One can efficiently store such distributions whenever the bond dimension χ is polynomial in d, and as described in Refs. [FV12; SW10], one can sample from such distributions with complexity $dN_{\max}\chi^3$. Note that the product distribution in Eq. (63) is a special case of an MPS distribution,with $\chi = 1$. Now, given an MPS distribution $\tilde{p}$ over $\tilde{\Omega}_{\mathcal{D}}$, we define the induced distribution over $\Omega_{\mathcal{D}}$ via Eq. (64).

Summary: In order to *efficiently* implement the RFF procedure for a kernel $K_{(\mathcal{D},\mathrm{w})}$, it is necessary that one can sample efficiently from $p_{(\mathcal{D},\mathrm{w})}$, the discrete probability distribution associated

to the kernel. However, the number of frequency vectors $|\Omega_{\mathcal{D}}|$ typically scales exponentially in d, and as such one *cannot* efficiently store and sample from arbitrary probability distributions over $\Omega_{\mathcal{D}}$. As such, *efficiently* implementing RFF is only possible for the subset of kernels whose associated distributions have a structure which facilitates efficient sampling. Due to the Cartesian product structure of $\tilde{\Omega}_{\mathcal{D}}$ one such set of distributions (amongst others) are those induced by MPS with polynomial bond dimension.

4.5 Kernel integral operator for PQC kernels

Recall from Theorem 1 that

$$M \geq c_o \sqrt{n} \log \frac{108\sqrt{n}}{\delta} \tag{67}$$

frequency samples are sufficient to guarantee, with probability greater than $1-\delta$, an error of at most ϵ between the output of the RFF procedure and the optimal PQC model. As such, in order to fully understand the complexity of RFF-based regression, it is necessary for us to gain a better understanding of c_0, which is given by

$$c_0 = 9\left(\frac{29}{4} + \frac{4}{\|T_{K_{(\mathcal{D},\mathrm{w})}}\|}\right). \tag{68}$$

In particular, in order to find the smallest number of sufficient frequency samples, it is necessary for us to obtain an upper bound on c_0, which in turn requires a *lower bound* on $\|T_{K_{(\mathcal{D},\mathrm{w})}}\|$, the operator norm of the kernel integral operator associated with $K_{(\mathcal{D},\mathrm{w})}$. We achieve this with the following lemma, whose proof can be found in Appendix C.

Lemma 1: (Operator norm of $T_{(K_{\mathcal{D}},\mathrm{w})}$) *Let $K_{(\mathcal{D},\mathrm{w})}$ be the re-weighted PQC kernel defined via Eq. (49), and let $T_{K_{(\mathcal{D},\mathrm{w})}}$ be the associated kernel integral operator (as per Definition 3). Assume that (a) the marginal distribution $P_{\mathcal{X}}$ appearing in the definition of the kernel integral operator is fixed to the uniform distribution, and (b) The frequency set $\Omega_{\mathcal{D}}$ consists only of integer vectors – i.e., $\Omega_{\mathcal{D}} \subset \mathbb{Z}^d$. Then, we have that*

$$\begin{aligned}\|T_{K_{(\mathcal{D},\mathrm{w})}}\| &= \max_{i\in|\Omega_{\mathcal{D}}|}\left\{\frac{1}{2}\frac{\mathrm{w}_i^2}{\|\mathrm{w}\|_2^2}\right\} \\ &= \max_{\omega\in\Omega_{\mathcal{D}}}\left\{\frac{1}{2}p_{(\mathcal{D},\mathrm{w})}(\omega)\right\}.\end{aligned} \tag{69}$$

In light of this, let us again drop the subscript for convenience, and define

$$p_{\max} \coloneqq \max_{\omega\in\Omega_{\mathcal{D}}}\left\{p_{(\mathcal{D},\mathrm{w})}(\omega)\right\}. \tag{70}$$

With this in hand we can immediately make the following observation:

Observation 3: In order to achieve $c_0 = \mathcal{O}(\mathrm{poly}(d))$, which is necessary for Theorem 1 to imply that $M = \mathcal{O}(\mathrm{poly}(d))$ frequency samples are sufficient, we require that

$$p_{\max} = \Omega\left(\frac{1}{\mathrm{poly}(d)}\right), \tag{71}$$

i.e., the maximum probability of the probability distribution associated to $K_{(\mathcal{D},\mathrm{w})}$ must decay at most inversely polynomially in d. This is illustrated in Figure 2.

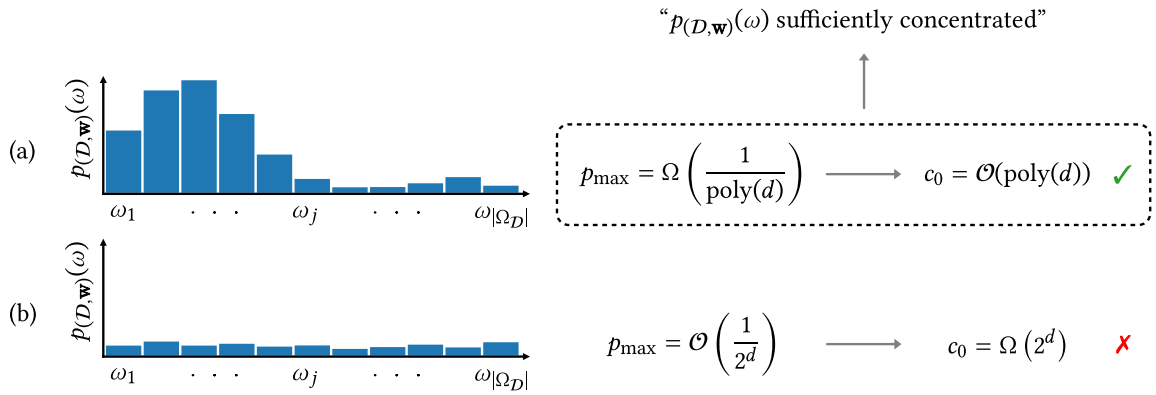


Figure 2: A graphical illustration of Observation 3. In particular, we see that if the re-weighting probability distribution $p_{(\mathcal{D},\mathrm{w})}$ is sufficiently concentrated, then c_0 will scale polynomially in d, which implies via Theorem 1 that polynomially many frequency samples M are sufficient to achieve the desired guarantee from RFF-based linear regression.

It is important to stress that we have *not* yet established the *necessity* that $p_{\max}$ decays inversely polynomially for the RFF procedure to be efficient. In particular, we have only established that this is required for the guarantee of Theorem 1 to be meaningful. However, we will show shortly, in Section 4.7, that Eq. (71) is indeed also necessary, at least in order to obtain a small *average* error.

In light of Observation 3, we can immediately rule out the meaningful applicability of Theorem 1 for kernels with the following associated distributions:

The uniform distribution: As discussed in Section 4.4, we have that $|\tilde{\Omega}_\mathcal{D}| \geq N_{\min}^d$, and therefore, for the uniform distribution over $|\Omega_\mathcal{D}|$ one has that

$$p_{\max} \leq \frac{2}{N_{\min}^d}, \tag{72}$$

i.e., $p_{\max}$ scales inverse exponentially with d.

Product-induced distributions: Consider a probability distribution p over $\Omega_\mathcal{D}$ induced by the product distribution $\tilde{p}$ over $\tilde{\Omega}_\mathcal{D}$, defined as per Eq. (63). We have that

$$\begin{aligned} p_{\max} &\leq 2\tilde{p}_{\max} && \text{[via Eq. (64)]} \\ &= 2\prod_{j\in[d]} \tilde{p}_{\max}^j && \text{[via Eq. (63)]} \\ &\leq 2\left(\max_{j\in[d]}\left\{\tilde{p}_{\max}^j\right\}\right)^d. \end{aligned} \tag{73}$$

Therefore, whenever $\max_{j\in[d]}\left\{\tilde{p}_{\max}^j\right\} < 1$, there exists some constant $c > 1$ such that

$$p_{\max} \leq \frac{2}{c^d}. \tag{74}$$

Summary: In order for Theorem 1 to be meaningfully applicable – i.e., to guarantee the efficiency of RFF for approximating variational QML – one requires that the operator norm of the kernel integral operator decays at most inverse polynomially with respect to d, which via Lemma 1 requires that the maximum probability of the distribution associated with the kernel decays

at most inversely polynomially with d. Unfortunately, this rules out any efficiency guarantee, via Theorem 1, for kernels $K_{(\mathcal{D},\mathrm{w})}$ whose associated distribution $p_{(\mathcal{D},\mathrm{w})}$ is either the uniform distribution or a product-induced distribution (with all component probability distributions non-trivial).

4.6 RKHS norm for PQC kernels

Recall from Theorem 1 that one requires

$$n \geq \max\left\{\frac{c_1^2 \log^4 \frac{1}{\delta}}{\epsilon^2}, n_0\right\} \tag{75}$$

data samples, in order for Theorem 1 to guarantee, with probability greater than $1-\delta$, an error of at most ϵ between the output of the RFF procedure and the optimal PQC model. As such, in order to fully understand the complexity of RFF-based regression, it is necessary for us to gain a better understanding of c_1, which is given by

$$c_1 \leq 8(4b + 3C + 2\sqrt{C}), \tag{76}$$

where b is set by the regression problem (and we can assume to be constant), and C is an upper bound on the RKHS norm of the optimal PQC model, with respect to the kernel used for RFF – i.e.,

$$\|f^*_{(\Theta,\mathcal{D},O)}\|_{K_{(\mathcal{D},\mathrm{w})}} \leq C. \tag{77}$$

Therefore, we see that in order to determine the smallest number of sufficient data samples, it is necessary for us to obtain a concrete upper bound on the RKHS norm of the optimal PQC function with respect to the kernel $K_{(\mathcal{D},\mathrm{w})}$. More specifically, we would like to understand, for which PQC architectures, and for which kernels, one obtains

$$C = \mathcal{O}\left(\mathrm{poly}(d)\right) \tag{78}$$

as for any such kernel and architecture, we can guarantee, via Theorem 1, the sample efficiency of the RFF procedure for approximating the optimal PQC model.

Ideally we would like to obtain results and insights which are *problem-independent* and therefore we focus here not on the optimal PQC function (which requires knowing the solution to the problem) but on the maximum RKHS norm over the entire PQC architecture. More specifically, given an architecture $(\Theta, \mathcal{D}, O)$ we would like to place upper bounds on

$$C_{(\Theta,\mathcal{D},O)} := \max\left\{\|f\|_{K_{(\mathcal{D},\mathrm{w})}} \,|\, f \in \mathcal{F}_{(\Theta,\mathcal{D},O)}\right\}, \tag{79}$$

as this clearly provides an upper bound on $\|f^*_{(\Theta,\mathcal{D},O)}\|_{K_{(\mathcal{D},\mathrm{w})}}$ for *any* regression problem.

We start off with an alternative definition of the RKHS norm, which turns out to be much more convenient to work with than the one we have previously encountered.

Lemma 2: (Alternative definition of RKHS norm – Adapted from Theorem 4.21 in Ref. [SC08]) *Given some kernel $K\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ defined via*

$$K(x, x') = \langle \phi(x), \phi(x') \rangle, \tag{80}$$

for some feature map $\phi\colon \mathcal{X} \to \mathcal{X}'$, one has that

$$\|f\|_K = \inf\left\{\|v\|_2 \,|\, v \in \mathcal{X}' \text{ such that } f(\cdot) = \langle v, \phi(\cdot)\rangle\right\} \tag{81}$$

for all $f \in \mathcal{H}_K$.

In words, Lemma 2 says that the RKHS norm is defined as the *infimum* over the 2-norms of all hyperplanes in feature space which realize f - i.e., the infimum over $\|v\|_2$ for all v such that $f(x) = \langle v, \phi(x)\rangle$. We stress that in general functions in the reproducing kernel Hilbert space *do not* have a unique hyper-plane representation with respect to the feature map. However, as detailed in Observation 4 below, for PQC feature maps and data-encoding strategies giving rise to integer frequency vectors (such as encoding strategies using only Pauli Hamiltonians [SSM21]), the hyperplane representation is indeed *unique*.

Observation 4: (Hyperplane uniqueness for integer frequencies) Let $\mathcal{D}$ be an encoding strategy for which $\Omega_{\mathcal{D}}^{+} \subset \mathbb{Z}^d$. In this case, one has that

$$\left\{1, \cos(\langle \omega_1, x\rangle), \sin(\langle \omega_1 x\rangle), \ldots, \cos\left(\langle \omega_{|\Omega_{\mathcal{D}}^{+}|}, x\rangle\right), \sin\left(\langle \omega_{|\Omega_{\mathcal{D}}^{+}|}, x\rangle\right)\right\} \tag{82}$$

is a mutually orthogonal set of functions. Therefore for any strictly positive re-weighting w, and any $u, v \in \mathbb{R}^{|\tilde{\Omega}_{\mathcal{D}}|}$, if $f(\cdot) = \langle v, \phi_{(\mathcal{D},\mathrm{w})}(\cdot)\rangle$ and $f(\cdot) = \langle u, \phi_{(\mathcal{D},\mathrm{w})}(\cdot)\rangle$ then $u = v$. Specifically, there exists only one hyperplane in feature space which realizes f. As a consequence one has, via Lemma 2, that if $f(\cdot) = \langle v, \phi_{(\mathcal{D},\mathrm{w})}(\cdot)\rangle$ then

$$\|f\|_{K_{(\mathcal{D},\mathrm{w})}} = \|v\|_2. \tag{83}$$

Additionally, we note that for any such encoding strategy, the Fourier transform of any PQC function $f \in \mathcal{F}_{(\Theta,\mathcal{D},O)}$ will immediately yield the hyper-plane v such that $f(\cdot) = \langle v, \phi_{(\mathcal{D},\mathrm{w})}(\cdot)\rangle$ for the *uniform* weight vector. The hyper-plane representation with respect to any other weight vector (from which the RKHS norm can be calculated) can then be extracted by re-scaling the hyper-plane components appropriately. With the above insights in hand, we can do some examples to gain intuition into the behaviour of the RKHS norm.

Example 1: Given a data-encoding strategy $\mathcal{D}$, and the uniform weight vector

$$\mathrm{w} = \frac{1}{\sqrt{|\Omega_{\mathcal{D}}|}}(1, \ldots, 1), \tag{84}$$

consider the function $f(x) = \cos(\langle \omega_1, x\rangle)$. In this case one has that $f(\cdot) = \langle v, \phi_{(\mathcal{D},\mathrm{w})}(\cdot)\rangle$ with

$$v = \left(0, \sqrt{|\Omega_{\mathcal{D}}|}, 0, \ldots, 0\right), \tag{85}$$

and therefore $\|f\|_{K_{(\mathcal{D},\mathrm{w})}} \leq \sqrt{|\Omega_{\mathcal{D}}|}$, with an equality in the case of encoding strategies with integer frequency vectors. Note that we obtain the same result for $f(x) = \cos(\langle \omega, x\rangle)$ and $f(x) = \sin(\langle \omega, x\rangle)$ for any $\omega \in \Omega_{\mathcal{D}}$.

Example 2: Let us consider the same function as in Example 1 – i.e., $f(x) = \cos(\langle \omega_1, x\rangle)$ – but this time let us consider the weight vector $\mathrm{w} = (0, 1, 0, \ldots, 0)$. In this case one has $f(\cdot) = \langle v, \phi_{(\mathcal{D},\mathrm{w})}(\cdot)\rangle$ with

$$v = (0, 1, 0 \ldots, 0), \tag{86}$$

and therefore $\|f\|_{K_{(\mathcal{D},\mathrm{w})}} \leq 1$, with an equality in the case of encoding strategies with integer frequency vectors. We again get the same result for $f(x) = \cos(\langle \omega, x\rangle)$ and $f(x) = \sin(\langle \omega, x\rangle)$ for any $\omega \in \Omega_{\mathcal{D}}$, if one uses the weight vector with $\mathrm{w}_\omega = 1$.

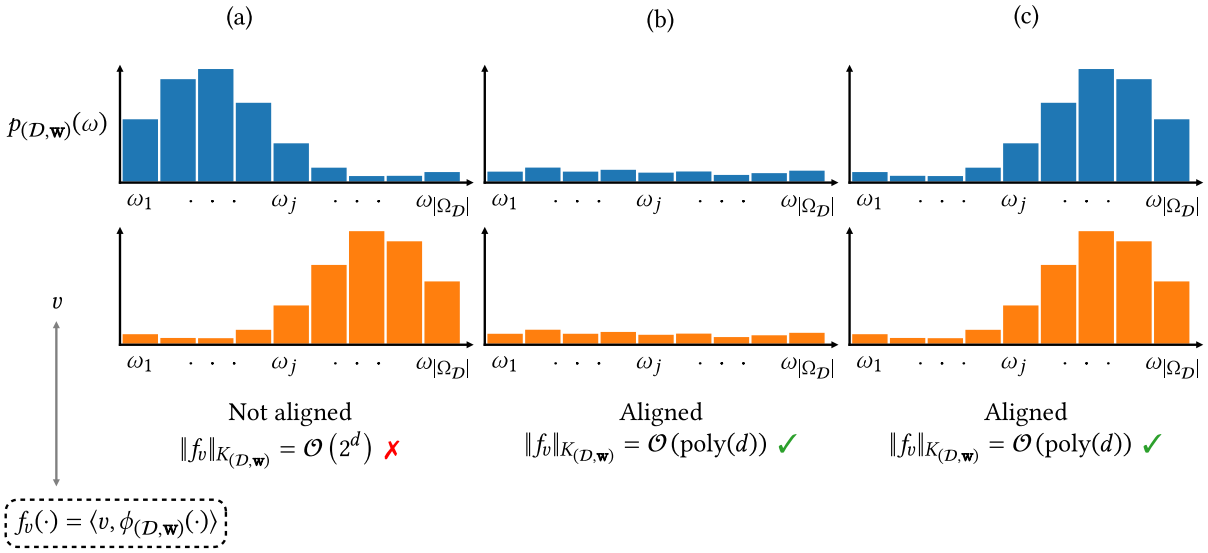


Figure 3: A graphical illustration of the conditions necessary for the RKHS norm of a function to scale polynomially in d. At a high level, we see that for a function $f_v(\cdot) = \langle v, \phi_{(\mathcal{D},\mathrm{w})}(\cdot)\rangle$, the hyperplane vector v (equivalently frequency spectrum of f), needs to be sufficiently well-aligned with the re-weighting vector w, which determines the kernel $K_{(\mathcal{D},\mathrm{w})}$ with respect to which the RKHS norm is taken.

Example 3: Given a data-encoding strategy $\mathcal{D}$, and the uniform weight vector

$$\mathrm{w} = \frac{1}{\sqrt{|\Omega_\mathcal{D}|}}(1,\ldots,1), \tag{87}$$

consider the function

$$f(x) = \frac{1}{|\Omega_\mathcal{D}|}\sum_{\omega\in\Omega_\mathcal{D}} \cos(\langle\omega, x\rangle). \tag{88}$$

In this case one has $f(\cdot) = \langle v, \phi_{(\mathcal{D},\mathrm{w})}(\cdot)\rangle$ with

$$v = \frac{1}{\sqrt{|\Omega_\mathcal{D}|}}\left(1,1,0,1,0\ldots,1,0\right), \tag{89}$$

and, therefore, $\|f\|_{K_{(\mathcal{D},\mathrm{w})}} \leq 1$, with an equality in the case of encoding strategies with integer frequency vectors.

Given these examples, we can extract the following important observations:

1. As per Example 1, there exist functions, and re-weighted PQC kernels, for which the RKHS norm of the function scales with the number of frequencies in $\Omega_\mathcal{D}$, and therefore exponentially in d. As such, we *cannot* hope to place a universally applicable (architecture independent) polynomial (in d) upper bound on $C_{(\Theta,\mathcal{D},O)}$. On the contrary, as per Examples 2 and 3 there do exist functions and reweightings for which the RKHS norm is *constant*. Therefore, while we cannot hope to place an architecture-independent upper bound on the RKHS norm, it may be the case that there exist specific circuit architectures and kernel re-weightings for which $C_{(\Theta,\mathcal{D},O)}$ is upper bounded by a polynomial in d. Note that this can be interpreted as an *expressivity constraint* on $(\Theta,\mathcal{D},O)$, as the more expressive an architecture is, the more likely it contains a function with large RKHS norm (with this likelihood becoming a certainty for the case of universal architectures).

2. By comparing Examples 1 and 2 we see that, as expected, the RKHS norm of a function depends strongly on the reweighting which defines the kernel. In particular, the same function can have a very different RKHS norm with respect to different feature map re-weightings.

3. By looking at all examples together, we see that informally, what seems to determine the RKHS norm of a given function f_v is the "alignment" between (a) the frequency distribution of the function f_v, i.e., the components of the vector v, and (b) the re-weighting vector w (or alternatively, the probability distribution $p_{(\mathcal{D},\mathrm{w})}$). In particular, in Example 1, the frequency representation of the function f is peaked on a single frequency, whereas the probability distribution $p_{(\mathcal{D},\mathrm{w})}$ is uniform over all frequencies. In this example, v and $p_{(\mathcal{D},\mathrm{w})}$ are *non-aligned*, and we find that the RKHS norm of f with respect to $K_{(\mathcal{D},\mathrm{w})}$ scales exponentially in D. On the contrary, in both Examples 2 and 3 we have that the frequency representation of f is well aligned with the probability distribution p, and we find that we can place a *constant* upper bound on the RKHS norm of the function. This is illustrated in Figure 3.

4. At an intuitive level, one should expect the "alignment" between the frequency representation of a target function and the probability distribution associated with the kernel to play a role in the complexity of RFF. Informally, in order to learn an approximation of the function f_v via RFF, when constructing the approximate kernel via frequency sampling we need to sample frequencies present in v. Therefore, if the distribution is supported mainly on frequencies *not* present in v, we cannot hope to achieve a good approximation via RFF. On the contrary, if the distribution is supported on frequencies present in v, with the correct weighting, then we can hope to approximate f_v using our approximate kernel. In this sense, our informal observation that the RKHS norm depends on the alignment of target function with kernel probability distribution squares well with our intuitive understanding of RFF-based linear regression. We will make this intuition much more precise in Section 4.7.

Summary: In order for the statement of Theorem 1 to imply that a polynomial number of data samples is sufficient, we require that $\|f^*_{(\Theta,\mathcal{D},O)}\|_{K_{(\mathcal{D},\mathrm{w})}}$, the RKHS norm of the optimal PQC function, scales polynomially with respect to d. Unfortunately, in the worst case, $\|f^*_{(\Theta,\mathcal{D},O)}\|_{K_{(\mathcal{D},\mathrm{w})}}$ can scale exponentially with respect to d, and therefore we cannot hope for efficient dequantization of variational QML via RFF for *all* possible circuit architectures. However, given a specific circuit architecture, and a re-weighting which leads to a distribution $p_{(\mathcal{D},\mathrm{w})}$ with an efficient sampling algorithm, it may be the case that $C_{(\Theta,\mathcal{D},O)}$ scales polynomially in d, in which case Theorem 1 yields a meaningful sample complexity for the RFF dequantization of optimization of $(\Theta,\mathcal{D},O)$ for *any* regression problem P. Unfortunately however, it seems unlikely that an expressive circuit architecture will not contain *any* functions with large RKHS norm. Ultimately though, all that is required by Theorem 1 is that the RKHS norm of the *optimal* PQC function scales polynomially in d, and this may be the case even when $C_{(\Theta,\mathcal{D},O)}$ scales superpolynomially. Unfortunately, given a regression problem P it is not clear how to assess the RKHS norm of the optimal PQC function without knowing this function in advance, which seems to require running the PQC optimization.

4.7 Lower bounds for RFF efficiency

By this point we have seen that the following is *sufficient* for Theorem 1 to imply the efficient dequantization of variational QML via RFF-based linear regression:

1. We need to be able to efficiently sample from $p_{(\mathcal{D},\mathrm{w})}$.

2. The distribution $p_{(\mathcal{D},\mathrm{w})}$ needs to be sufficiently concentrated. In particular, we need $p_{\max} = \Omega(1/\mathrm{poly}(d))$ in order to place a sufficiently strong lower bound on the operator norm of the kernel integral operator.

3. The frequency representation of the optimal PQC function needs to be "well aligned" with the probability distribution $p_{(\mathcal{D},\mathrm{w})}$. This is required to ensure a sufficiently strong upper bound on the RKHS norm of the optimal PQC function.

It is clear that point 1 above is a *necessary* criterion for efficient dequantization via RFF. However, it is less clear to which extent points 2 and 3 are necessary. In particular, it could be the case that the bounds provided by Theorem 1 are not tight, and that efficient PQC dequantization via RFF is possible even when not guaranteed by Theorem 1. In this section we address this issue to some extent, by proving *lower bounds* on the complexity of RFF which show that both points 2 and 3 are also necessary conditions in some sense, when using certain encoding strategies. To be more specific, Theorem 1 provides sufficient conditions for the output of RFF-based linear regression to be (a) with high probability, (b) no more than ϵ worse than the optimal PQC function with respect to true risk. In this section we show necessary conditions – when using encoding strategies with integer frequencies – to ensure that (a) the expected, (b) L^2-norm difference, is small between the output of RFF-based linear regression and the optimal PQC function. As such, the necessary conditions we provide here are different from the sufficient conditions of Theorem 1 in the following ways:

1. They ensure that the L^2-norm difference is small between the output of RFF-regression and the optimal PQC function, as opposed to the true risk difference.

2. They ensure the above difference is small *in expectation*, as opposed to with high probability.

3. They apply only to circuit architectures using encoding strategies with integer-valued frequency vectors (such as those using Pauli word Hamiltonians).

To make this more precise, we need to introduce some additional notation. We consider a data encoding strategy, giving rise to an integer-valued frequency set $\tilde{\Omega}_{\mathcal{D}}$, as well as a weight vector w, giving rise to the sampling distribution $p_{(\mathcal{D},\mathrm{w})}$, which we abbreviate as p. Now, given a regression problem, we define the following notions:

1. Let $f^*_{(\Theta,\mathcal{D},O)}$ represent the optimal PQC model. In this section, for convenience we abbreviate this as f^*.

2. As per Section 2.3 and the description of Algorithm 3 in Section 3, we consider running RFF regression by sampling M frequencies from the distribution $\pi = p \times \mu$. Let $\vec{\omega} = (\omega_1, \ldots, \omega_M) \in \Omega_{\mathcal{D}}^M$ be the random variable of M frequencies sampled from p, and let $g_{\vec{\omega}}$ be the output of linear regression using these frequencies to approximate the kernel $K_{(\mathcal{D},\mathrm{w})}$.

As discussed above, in this section we are concerned with lower bounding the *expected* L^2-norm of the difference between the optimal PQC function and the output of the RFF procedure, with respect to multiple runs of RFF-based linear regression. In particular, we want to place lower bounds on the quantity

$$\begin{aligned} \hat{\epsilon} &:= \mathop{\mathbb{E}}_{\vec{\omega}\sim p^M} \|f^* - g_{\vec{\omega}}\|_2^2 \\ &= \sum_{\vec{\omega}\in\Omega_{\mathcal{D}}^M} \|f^* - g_{\vec{\omega}}\|_2^2 \xi(\vec{\omega}), \end{aligned} \tag{90}$$

where $\xi = p^M$, i.e., $\xi(\vec{\omega}) = p(\omega_1) \times \ldots \times p(\omega_M)$. In order to lower bound $\hat{\epsilon}$, recall from Section 2.4 that f^* can be written as

$$f^*(x) = \sum_{\omega \in \tilde{\Omega}_{\mathcal{D}}} \hat{f}^*(\omega) e^{i\langle \omega, x \rangle}. \tag{91}$$

We abuse notation slightly and use the notation $\hat{f}^*$ to denote the vector with entries $\hat{f}^*(\omega)$. Finally, we denote by $p_{\max}$ the maximum probability in p. With this in hand, we have the following lemma (whose proof can be found in Appendix D):

Lemma 3: (Lower bound on average relative error) *Given any encoding strategy $\mathcal{D}$ for which all $\omega \in \Omega_{\mathcal{D}}$ have only integer-valued components, the expected L^2-norm of the difference between the optimal PQC function and the output of RFF-based linear regression can be lower bounded as*

$$\begin{aligned} \hat{\epsilon} &\geq (2\pi)^d \|\hat{f}^*\|_2^2 - (2\pi)^d 2M \sum_{\omega \in \Omega_{\mathcal{D}}} |\hat{f}^*(\omega)|^2 p(\omega) \qquad (92) \\ &\geq (2\pi)^d \|\hat{f}^*\|_2^2 - (2\pi)^d 2M \sum_{\omega \in \Omega_{\mathcal{D}}} |\hat{f}^*(\omega)|^2 p_{\max} \\ &= \|f^*\|_2^2 \left(1 - 2M p_{\max}\right). \qquad (93) \end{aligned}$$

Using this Lemma, we can now see that, at least for some class of regression problems, both concentration of the probability distribution p, and "alignment" of the frequency representation of the optimal function with p, are *necessary* conditions to achieve a small *expected* relative error $\hat{\epsilon}$, when using encoding strategies with integer-valued frequencies.

Concentration of p: To this end, note that we can rewrite Eq. (93) as

$$M \geq \frac{1}{2p_{\max}} \left(1 - \frac{\hat{\epsilon}}{\|f^*\|_2^2}\right). \tag{94}$$

Given this, we see that when $(1 - \hat{\epsilon}/\|f^*\|_2^2) = \Omega(1)$ then one requires $M = \Omega(1/p_{\max})$ frequency samples to achieve expected relative error $\hat{\epsilon}$. We stress however that the condition $(1 - \hat{\epsilon}/\|f^*\|_2^2) = \Omega(1)$ will *not* always be satisfied. More specifically, one requires that $\|f^*\|_2^2 \geq c + \hat{\epsilon}$ for all d, for some *constant* c. In particular, we note that the bound of Eq. (94) is vacuous whenever $\|f^*\|_2^2 \leq \hat{\epsilon}$. However when $(1 - \hat{\epsilon}/\|f^*\|_2^2) = \Omega(1)$ is satisfied, the RFF procedure *cannot* be efficient whenever $p_{\max}$ is a negligible function – i.e., decays faster than any inverse polynomial. More specifically, when p is not sufficiently concentrated, one will require super-polynomially many samples M. Recall from Section 4.5 that for all product-induced distributions, including the uniform distribution, $p_{\max}$ is a negligible function of d - and therefore we *cannot* achieve efficient RFF dequantization of variational QML via any re-weighting giving rise to such a distribution, when $(1 - \hat{\epsilon}/\|f^*\|_2^2) = \Omega(1)$ is satisfied. Complimenting the lower bound, we recall that as discussed in Section 4.5, whenever $p_{\max} = \Omega(1/\text{poly}(d))$ – i.e., when p is sufficiently concentrated – then one *can* apply Theorem 1 to place a polynomial upper bound on M.

Alignment of $\hat{f}^*$ and p: Note that we can rewrite Eq. (92) as

$$M \geq \frac{1}{2\sum_{\omega \in \Omega_{\mathcal{D}}} |\hat{f}^*(\omega)|^2 p(\omega)} \left(\|\hat{f}^*\|_2^2 - \frac{\hat{\epsilon}}{(2\pi)^d}\right). \tag{95}$$

Therefore, whenever $\|\hat{f}^*\|_2^2 - \hat{\epsilon}/(2\pi)^d = \Omega(1)$ one has that

$$M = \Omega\left(\frac{1}{\sum_{\omega \in \Omega_{\mathcal{D}}} |\hat{f}^*(\omega)|^2 p(\omega)}\right) \tag{96}$$

where

$$\sum_{\omega \in \Omega_{\mathcal{D}}} |\hat{f}^*(\omega)|^2 p(\omega) \tag{97}$$

is the inner product between the frequency vector $\hat{f}^*$ and the probability distribution p, which we interpret as the "alignment" between the frequency representation and the sampling distribution. We therefore see that the smaller this overlap, the larger number of frequencies are required to achieve a given relative error. Again, we therefore see that "large alignment" between $\hat{f}^*$ and the probability distribution p is a necessary condition to achieve a smaller expected relative error.

5 Discussion and conclusions

In this work, we have provided a detailed analysis of classical linear regression with random Fourier features, using re-weighted PQC kernels, as a method for the dequantization of PQC based regression. Intuitively, as discussed in Section 3, this method is motivated by the fact that it optimizes over a natural extension of the same function space used by PQC models – i.e., the method has to some extent an *inductive bias* which is comparable to that of PQC regression. At a very high level, given a PQC architecture $(\Theta, \mathcal{D}, O)$ and a regression problem $P \sim \mathcal{X} \times \mathbb{R}$, the method consists of:

1. Choosing a re-weighting w of the PQC feature map, or equivalently, choosing a distribution p over frequencies appearing in $\Omega_{\mathcal{D}}$.
2. Sampling M frequencies from p, and using them to construct an approximation of the PQC feature map.
3. Being given, or sampling, n training data points from P and running regularized linear regression with the approximate feature map.

The main intuitive take-aways from this work are that in order for this dequantization procedure to output a good approximation to the optimal PQC function f^* one needs to sample frequencies present in the multivariate Fourier decomposition of f^*, and in order to be able to do this *efficiently* only a polynomial number of these frequencies should be sufficient for a good approximation.

To be more precise, we know that Step 3 above has time and space complexity $\mathcal{O}(nM)$ and $\mathcal{O}(nM^2 + M^3)$, respectively. Given this, the question we have addressed is whether it is possible to use some number of frequencies $M = \mathcal{O}(\text{poly(d)})$ and some number of data samples $n = \mathcal{O}(\text{poly}(d))$, such that, with high probability, the output of classical RFF-based linear regression is guaranteed to achieve a true risk which is no more than ϵ worse than the output of PQC based optimization. In other words, whether we can *efficiently* dequantize PQC regression via architecture $(\Theta, \mathcal{D}, O)$ for regression problem P. In order to answer this question we have been interested in obtaining necessary and sufficient conditions on n and M – in terms of properties of the circuit architecture $(\Theta, \mathcal{D}, O)$, re-weighting w and regression problem P – in order to achieve the desired guarantee. To this end we have seen, via Theorem 1 and the subsequent analysis, that RFF-based linear regression with re-weighting w provides an efficient dequantization of PQC regression over the circuit architecture $(\Theta, \mathcal{D}, O)$, for regression problem P, if:

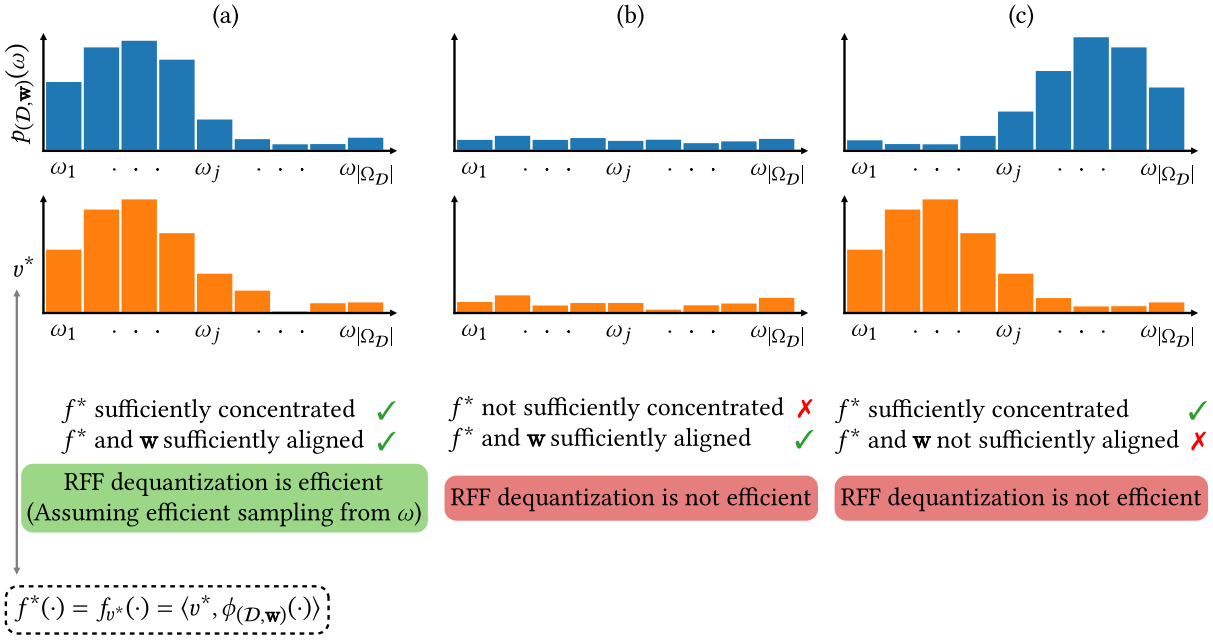


Figure 4: A graphical illustration of the conditions sufficient for RFF-based linear regression with re-weighting w to provide an efficient dequantization of PQC regression via circuit architecture $(\Theta, \mathcal{D}, O)$, for regression problem P. One requires that the optimal PQC function f^* for P is both sufficiently concentrated and sufficiently well-aligned with the re-weighting distribution $p_{(\mathcal{D},\mathrm{w})}$.

1. The re-weighting distribution $p_{(\mathcal{D},\mathrm{w})}$ is sufficiently concentrated. In particular, $p_{\max}$ should decay at most inversely polynomially in d.
2. The re-weighting distribution $p_{(\mathcal{D},\mathrm{w})}$ is sufficiently "well-aligned" with the optimal PQC function f^* for the regression problem P. Technically, the RKHS norm of the optimal PQC function $\|f^*_{(\Theta,\mathcal{D},O)}\|_{K_{(\mathcal{D},\mathrm{w})}}$ should scale polynomially in d.
3. There exists an efficient algorithm to sample from $p_{(\mathcal{D},\mathrm{w})}$, given as input only the data encoding strategy $\mathcal{D}$.

We stress again the intuitive perspective on the above conditions: In order for the dequantization procedure to output a good approximation to the optimal PQC function f^* we need to sample frequencies present in f^* (condition 2), and in order to be able to do this efficiently only a polynomial number of these frequencies should be sufficient for a good approximation (condition 1), and the sampling procedure itself should be efficient (condition 3). Additionally, we note that in order to satisfy conditions 1 and 2 above, it is necessary that the frequency representation of the optimal PQC function f^* is sufficiently concentrated. If this is not the case, then it is impossible for the re-weighting distribution $p_{(\mathcal{D},\mathrm{w})}$ to be both sufficiently concentrated *and* well-aligned with f^*. Given this, we have summarized the sufficient conditions given above graphically in Figure. 4. On the other hand, we have seen, via Lemma 3, that these conditions are also *necessary* in a certain sense – i.e., that if they are not satisfied then – at least for integer-valued encoding strategies and certain regression problems – RFF-based linear regression will *not* provide an efficient dequantization of PQC regression (on average).

Given these insights, perhaps the most interesting questions are the following:

1. Given a PQC architecture $(\Theta, \mathcal{D}, O)$, as well as (samples from) a regression problem P, how do we evaluate whether PQC regression over $(\Theta, \mathcal{D}, O)$ can be efficiently dequantized via RFF-based linear regression?

2. If PQC regression can be efficiently dequantized, how do we identify an appropriate re-weighting w?

Figure 5 provides a methodology for answering this question, using the results of this work as a guide. We elaborate on this below.

The first step is to ask whether variational QML with the PQC architecture $(\Theta, \mathcal{D}, O)$ can be dequantized via RFF-based linear regression, *irrespective* of the regression problem of interest. More specifically, for any regression problem P, there will exist some optimal PQC function f^*. Previously, we have focused our analysis on this specific function f^*, and noted that it needs to be both sufficiently concentrated and well aligned with a re-weighting whose associated distribution is efficient to sample from. However, if there exists a sufficiently concentrated re-weighting w, which can be efficiently sampled from, and which is well aligned with *all functions* $f \in \mathcal{F}_{(\Theta,\mathcal{D},O)}$, then using w for the re-weighting will ensure that the sufficient conditions for dequantization will be satisfied *for any* regression problem P. More specifically, no matter which $f \in \mathcal{F}_{(\Theta,\mathcal{D},O)}$ is the optimal PQC function f^*, it will be well aligned with w, which is sufficiently concentrated. From a more technical perspective, this step of the methodology is equivalent to asking whether there exists a sufficiently concentrated re-weighting w, which can be efficiently sampled from, such that for all $f \in \mathcal{F}_{(\Theta,\mathcal{D},O)}$ the RKHS norm $\|f_{(\Theta,\mathcal{D},O)}\|_{K_{(\mathcal{D},\mathrm{w})}}$ scales polynomially in d. As illustrated in Figure. 5, if the answer to this question is "Yes", then we know that dequantization of $(\Theta, \mathcal{D}, O)$ is possible via re-weighting w for *any* regression problem P. If the answer is "No" (or the question is too hard to answer in practice) then we proceed to examine the specific regression problem of interest.

Before continuing it is worth commenting on a few aspects of the above discussion. Firstly, in order for all $f \in \mathcal{F}_{(\Theta,\mathcal{D},O)}$ to be well-aligned with a single sufficiently concentrated re-weighting w, they should all be sufficiently concentrated and mutually well-aligned. In principle, we can use this insight to guide PQC architecture design: in order to avoid being dequantizable via RFF for any regression problem, a PQC architecture should contain functions whose frequency spectra are either not sufficiently concentrated, or not mutually well-aligned. While useful in principle, we stress that in practice this may be difficult to assess. In particular, determining this property of $\mathcal{F}_{(\Theta,\mathcal{D},O)}$ may in the worst-case require analyzing the frequency spectrum of *all* functions $f \in \mathcal{F}_{(\Theta,\mathcal{D},O)}$. We elaborate more on the challenge and potential of using these insights for PQC architecture design in Section 6.

While it may be the case that for certain restricted circuit architectures problem-independent dequantization is possible, we expect that for any sufficiently expressive circuit architecture this will not be the case, and therefore it will be necessary to perform an evaluation which depends on the regression problem of interest. To this end, as discussed before, we need to assess whether the *optimal* PQC function f^* for the regression problem is well-aligned with a sufficiently concentrated re-weighting distribution, from which one can efficiently sample. Unfortunately, it is not a-priori clear how to do this without first solving the regression problem and identifying f^*!

However, as illustrated in Figure 5, we note that a natural way to approach this problem is via the following sequence of questions: Firstly, given samples (i.e., data) from the regression problem, is the frequency spectrum of the optimal PQC function is sufficiently concentrated? If yes, can we identify the frequencies on which the support is concentrated? In principle, it is possible that answering these two questions is easier than learning the function itself, which would also require

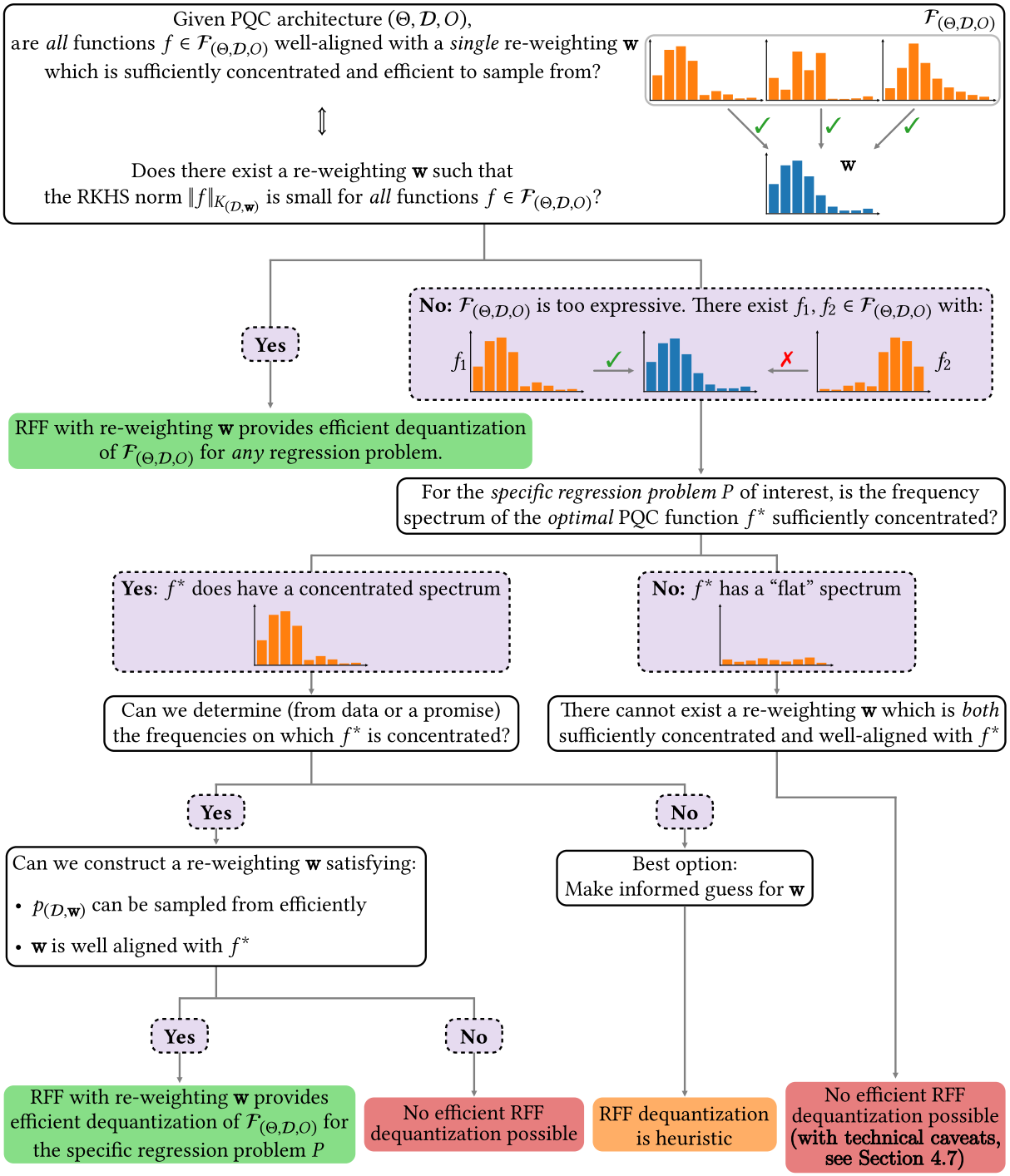


Figure 5: A methodology for determining whether, and via which re-weighting, linear regression via RFF can provide an efficient means for dequantizing PQC regression over circuit architecture $(\Theta, \mathcal{D}, O)$, for regression problem P?

learning the actual frequency co-efficients. This is in line with the observation underlying the field of property testing, that often *testing* properties of a function can be done more efficiently than *learning* the function [Gol17]. As discussed in Section 6 below, the development of techniques for answering these two property-testing questions is therefore a natural and well-defined direction of future research. Moreover, it may be the case that for certain regression problems one has a *structural promise* which allows one to answer these questions. Indeed, even a partial answer may allow one to make a well-informed *guess* for the re-weighting distribution, which leads to a well-informed *heuristic* method for dequantization. Additionally, as also illustrated in Figure. 5, if we are able to identify that the optimal PQC function f^* is *not* sufficiently concentrated, then we can immediately conclude that efficient dequantization via RFF is *not* possible (provided that the technical caveats discussed in detail in Section 4.7, on the encoding strategy and 2-norm of f^*, are satisfied). As discussed in Section 6 below, this therefore provides a potential tool for identifying candidate problems for "quantum advantage" – i.e., problems for which efficient dequantization via RFF-based linear regression is *not* possible.

Finally, as illustrated in Figure. 5, we note that even if we are able to identify the frequencies on which the optimal PQC function f^* is concentrated, efficient RFF dequantization still requires the construction of a re-weighting distribution supported on these frequencies, from which one can efficiently sample. Once again, it is not immediately a-priori clear how one should construct such distributions. As such, the development of methods for the identification and design of such distributions is once again a natural avenue for future research.

6 Future directions

Given the discussion above, the following are natural avenues for future research:

Identification of problems admitting potential quantum advantage: Can we identify a class of scientifically, industrially or socially relevant regression problems, whose regression functions are well aligned with anti-concentrated distributions over exponentially large frequency sets, and are therefore good candidates for quantum advantage via variational QML?

Testing frequency concentration and support: A key step of the methodology outlined in Figure 5 involves determining whether the optimal PQC function, for a specific regression problem, has a sufficiently concentrated frequency spectrum. As such, the development of efficient algorithms for *testing* this property of the optimal PQC function from a specific architecture, from samples of the regression problem at hand, would facilitate the practical assessment of whether PQC regression via a specific architecture can be dequantized, for a specific regression problem.

Design of suitable sampling distributions: We have seen that a *necessary* condition for efficient dequantization via RFF is that the distribution p is both efficiently sampleable and sufficiently concentrated. This immediately *rules out* a large class of natural distributions - namely the uniform distribution and product-induced distributions with non-trivial components. Given this, in order for RFF-based dequantization to be useful, it is important to identify and motivate suitable sampling distributions. We note that ideally one would choose the distribution based on knowledge of the regression function of the problem P, however in practice it is more likely that one would first choose a distribution p, which will then determine the class of problems for which RFF will be an efficient dequantization method – namely those problems whose regression function is well aligned with p.

PQC architecture design guided by RKHS norm: As we have discussed, for any circuit architecture for which *all* expressible functions are well aligned with a suitable distribution p, one

cannot obtain a quantum advantage, as RFF-based linear regression will provide an efficient dequantization method. Given this, it is of interest to investigate circuit architectures from this perspective, to understand which architectures might facilitate a quantum advantage, and which architectures are prone to dequantization. Unfortunately, we note that gaining *analytic* insight into which hyperplanes (i.e., frequency representations) are expressible by a given PQC architecture is a hard problem, which has so far resisted progress. However, as observed in Section 4.6, for data-encoding strategies with integer frequencies – i.e., all encoding strategies using Pauli word Hamiltonians – one can in principle numerically evaluate the RKHS norm of a given PQC function via its Fourier transform. Of course, in practice this will be limited by the fact that the frequency set typically scales exponentially with the dimensionality of the data. However, the hope is that one may be able to extract meaningful and useful insights by studying both smaller PQC model classes, for which the RKHS norm calculations are tractable numerically, and more structured PQC model classes for which analytic calculations may be possible.

Effect of noise on RKHS norm of PQC architectures: As we have pointed out, for any sufficiently expressive circuit architecture we expect the worst case RKHS norm to scale superpolynomially – i.e., that there exist PQC functions whose frequency representation is aligned with an anti-concentrated distribution over frequencies. It would, however, be of interest to understand the effect of noise on architectures realizing such functions. In particular, it could be the case that realistic circuit noise causes a concentration of the frequencies which are expressible by a PQC architecture, and therefore facilitates dequantization via RFF-based linear regression.

Extension to classification problems: The analysis we have performed here has focused on *regression* problems. Extending this analysis to classification problems would be both natural and interesting.

Improved RFF methods for sparse data: In this work, we have provided an analysis of "standard" RFF-based linear regression, for regression problems with no promised structure. However, when one is promised that the distribution P has some particular structure, then one can devise variants of RFF with improved efficiency guarantees. One such example we have already seen – namely, if one can guarantee that the regression function has a frequency representation supported on a subset of possible frequencies, then on can design the sampling distribution appropriately, which leads to improved RFF efficiencies [RR17]. In a similar vein, it is known that when one has a promise on the sparsity of the vectors in the support of P, then one can devise variants of RFF with improved efficiency [CSK16]. As this is a natural promise for application-relevant distributions, understanding the potential and limitations of "Sparse-RFF" as a dequantization method is an interesting research direction.

PQC dequantization without RFF: Here we have discussed only *one* potential method for the dequantization of variational QML. As we have noted in Section 3, recently Ref. [STJ24] has noted that to each PQC one can associate a feature map for which the associated kernel can be evaluated efficiently classically *without* requiring any approximations. As such, understanding the extent to which one can place relative error guarantees on linear regression using such a kernel is a natural avenue for investigation. Additionally, as mentioned in the introduction, a variety of recent works have shown that PQCs can be efficiently simulated, *in an average case sense*, in the presence of certain types of circuit noise [FRD+23; SWC+23]. Again, understanding the extent to which this allows one to classically emulate noisy variational QML is another natural approach to dequantization of *realistic* variational QML. Finally, quite recently Ref. [JGM+24] has shown that one can sometimes efficiently extract from a trained PQC a "shadow model" which can be used for efficient classical inference. Given this, it would be interesting to understand the extent to which one can train classical shadow models directly from data.

Acknowledgments

RS is grateful for helpful conversations with Alex Nietner, Christa Zoufal and Yanting Teng. This work is supported by the Government of Spain (Severo Ochoa CEX2019-000910-S, FUNQIP and European Union NextGenerationEU PRTR-C17.I1), Fundació Cellex, Fundació Mir-Puig, Generalitat de Catalunya (CERCA program) and European Union (PASQuanS2.1, 101113690). ERA is a fellow of Eurecat's "Vicente López" PhD grant program. ERA would also like to give special thanks to Jens Eisert for inviting him to participate in his group and also thank Adan Garriga for helping him all the way with this stay. SJ thanks the BMBK (EniQma) and the Einstein Foundation (Einstein Research Unit on Quantum Devices) for their support. EGF is funded by the Einstein Foundation (Einstein Research Unit on Quantum Devices). JJM is funded by QuantERA (HQCC). JE is funded by the QuantERA (HQCC), the Einstein Foundation (Einstein Research Unit on Quantum Devices), the Munich Quantum Valley (K-8), Berlin Quantum, the QuantERA (HQCC), the MATH+ cluster of excellence, the BMWK (EniQma), and the BMBF (Hybrid++). We acknowledge support by the Open Access Publication Fund of Freie Universität Berlin.

A Construction of the frequency set $\tilde{\Omega}_{\mathcal{D}}$

We describe here the way in which the frequency set $\tilde{\Omega}_{\mathcal{D}}$ of a PQC model is constructed from the data-encoding strategy $\mathcal{D}$. We follow closely the presentation in Ref. [LTD+22], and start by noting that

$$\tilde{\Omega}_{\mathcal{D}} = \tilde{\Omega}_{\mathcal{D}}^{(1)} \times \ldots \times \tilde{\Omega}_{\mathcal{D}}^{(d)} \tag{98}$$

where $\tilde{\Omega}_{\mathcal{D}}^{(j)} \subseteq \mathbb{R}$ depends only on $\mathcal{D}^{(j)}$. We can therefore focus on the construction of $\tilde{\Omega}_{\mathcal{D}}^{(j)}$ for a single co-ordinate. In light of this, let us drop some coordinate-indicating superscripts for ease of presentation. In particular, let us write $\mathcal{D}^{(j)} = \{H_k \,|\, k \in [L_j]\}$, where we have dropped the coordinate-indicating superscripts from the Hamiltonians. We then use λ_k^i to denote the i'th eigenvalue of H_k, and N_k to denote the number of eigenvalues of H_k. We also introduce the multi-index $\vec{i} = (i_1, \ldots, i_{L_j})$, with $i_k \in [N_k]$, which allows us to define the sum of the eigenvalues indexed by $\vec{i}$, one from each Hamiltonian, as

$$\Lambda_{\vec{i}} = \lambda_1^{i_1} + \ldots + \lambda_{L_j}^{i_{L_j}}. \tag{99}$$

With this setup, we then have that the frequency set $\tilde{\Omega}_{\mathcal{D}}^{(j)}$ is given by the set of all differences of all possible sums of eigenvalues, i.e.,

$$\tilde{\Omega}_{\mathcal{D}}^{(j)} = \left\{\Lambda_{\vec{i}} - \Lambda_{\vec{j}} \,|\, \vec{i}, \vec{j}\right\}, \tag{100}$$

and as mentioned before, the total frequency set is given by Eq. (98). There is a convenient graphical way to understand this construction, which is illustrated in Figure. 6. Essentially, one notes that, in order to construct $\tilde{\Omega}_{\mathcal{D}}^{(j)}$ one can consider a tree, with depth equal to the number of data-encoding gates, whose leaves contain the eigenvalue sums $\Lambda_{\vec{i}}$. The frequency set is then given by all possible pairwise differences between leaves.

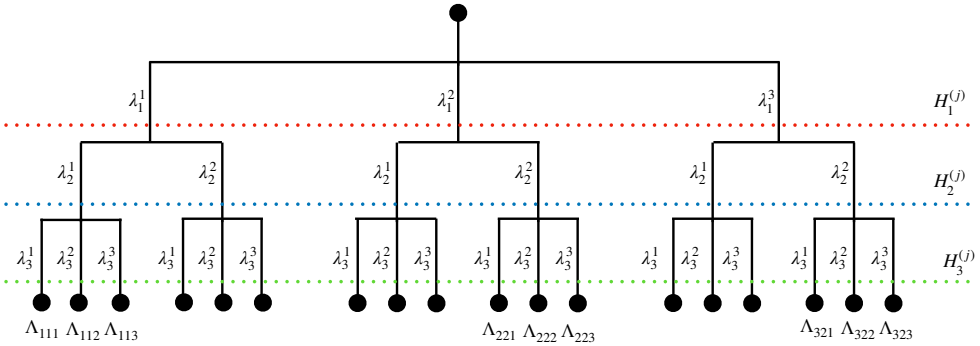


Figure 6: Construction of the frequency set $\tilde{\Omega}_{\mathcal{D}}^{(j)}$ from the data-encoding strategy $\mathcal{D}^{(j)}$.

B Proof of Theorem 1

We start by noting that Theorem 1 in the main text follows as an immediate corollary of the following Theorem:

Theorem 2: (RFF vs. variational QML – alternative form) *Let R be the risk associated with a regression problem $P \sim \mathcal{X} \times \mathbb{R}$. Assume the following:*

1. $\|f^*_{(\Theta,\mathcal{D},O)}\|_{K_{\mathcal{D}}} \leq C$,
2. $|y| \leq b$ *almost surely when* $(x, y) \sim P$, *for some* $b > 0$.

Additionally, define

$$n_0 := \max\left\{4\|T_{K_{\mathcal{D}}}\|^2, \left(528 \log \frac{1112\sqrt{2}}{\delta}\right)^2\right\}, \tag{101}$$

$$c_0 := 36\left(3 + \frac{2}{\|T_{K_{\mathcal{D}}}\|}\right), \tag{102}$$

$$c_1 := 8\sqrt{2}(4b + \frac{5}{\sqrt{2}}C + 2\sqrt{2C}). \tag{103}$$

Then, let $\delta \in (0, 1]$, let $n \geq n_0$, set $\lambda_n = 1/\sqrt{n}$, and let $\hat{f}_{M_n,\lambda_n}$ be the output of λ_n-regularized linear regression with respect to the feature map

$$\phi_{M_n}(x) = \frac{1}{\sqrt{M_n}}(\psi(x, \nu_1), \ldots, \psi(x, \nu_{M_n})) \tag{104}$$

constructed from the integral representation of $K_{\mathcal{D}}$ by sampling M_n elements from π. Then,

$$M_n \geq c_0\sqrt{n} \log \frac{108\sqrt{n}}{\delta} \tag{105}$$

is enough to guarantee, with probability at least $1 - \delta$, that if $R(\hat{f}_{M_n,\lambda_n}) \geq R(f^*_{(\Theta,\mathcal{D},O)})$, then

$$R(\hat{f}_{M_n,\lambda_n}) - R(f^*_{(\Theta,\mathcal{D},O)}) \leq \frac{c_1 \log^2 \frac{1}{\delta}}{\sqrt{n}}. \tag{106}$$

Next we note that Theorem 2 above – from which we derive Theorem 1 in the main text as an immediate corollary – is essentially a straightforward application of the generalization bound

given as Theorem 1 in Ref. [RR17]. As such, we start our proof of Theorem 2 with a presentation of this result. To this end, we require first a few definitions. Firstly, given a kernel K, with associated RKHS $(\mathcal{H}_K, \langle\cdot,\cdot\rangle_K)$, we define $\mathcal{H}_K^C = \{f \in \mathcal{H}_K \,|\, \|f\|_K \leq C\}$ as the subset of functions in $\mathcal{H}_K$ with RKHS norm bounded by C. We define $\mathcal{F}_{\mathcal{D}}^C$ and $\mathcal{F}_{(\Theta,\mathcal{D},O)}^C$ analogously. Additionally, given a regression problem P with associated risk R, we then define

$$f_{\mathcal{H}_K^C}^* = \underset{f\in\mathcal{H}_K^C}{\arg\min}\,[R(f)]\,, \tag{107}$$

as the optimal function for P in $\mathcal{H}_K^C$. Finally, recall that we denote by T_K the kernel integral operator associated with the kernel K (see Definition 3). With this in hand, we can state a slightly reformulated version of the RFF generalization bound proven in Ref. [RR17] (which in turn build on the earlier results of Ref. [SS15a]).

Theorem 3: (Theorem 1 from [RR17]) *Assume a regression problem $P \sim \mathcal{X}\times\mathbb{R}$. Let $K : \mathcal{X}\times\mathcal{X}\to\mathbb{R}$ be a kernel, and let $\mathcal{H}_K^C$ be the subset of the RKHS $\mathcal{H}_K$ consisting of functions with RKHS-norm upper bounded by some constant C. Assume the following:*

1. *K has an integral representation*

$$K(x,x') = \int_\Phi \psi(x,\nu)\psi(x',\nu)\,\mathrm{d}\pi(\nu). \tag{108}$$

2. *The function ψ is continuous in both variables and satisfies $|\psi(x,\nu)| \leq \kappa$ almost surely, for some $\kappa \in [1,\infty)$.*

3. *$|y| \leq b$ almost surely when $(x,y)\sim P$, for some $b > 0$.*

Additionally, define

$$\overline{B} \coloneqq 2b + 2\kappa\max\{1, \|f_{\mathcal{H}_K^C}^*\|_K\}, \tag{109}$$

$$\overline{\sigma} \coloneqq 2b + 2\kappa\sqrt{\max\{1, \|f_{\mathcal{H}_K^C}^*\|_K\}}, \tag{110}$$

and

$$n_0 \coloneqq \max\left\{4\|T_K\|^2, \left(264\kappa^2\log\frac{556\kappa^3}{\delta}\right)^2\right\}, \tag{111}$$

$$c_0 \coloneqq 9\left(3 + 4\kappa^2 + \frac{4\kappa^2}{\|T_K\|} + \frac{\kappa^4}{4}\right), \tag{112}$$

$$c_1 \coloneqq 8\big(\overline{B}\kappa + \overline{\sigma}\kappa + \max\{1, \|f_{\mathcal{H}_K^C}^*\|_K\}\big). \tag{113}$$

Then, let $\delta\in(0,1]$, $n \geq n_0$, assume $\lambda_n = 1/\sqrt{n}$, and let $\hat{f}_{M_n,\lambda_n}$ be the output of λ_n-regularized linear regression with respect to the feature map

$$\phi_{M_n}(x) = \frac{1}{\sqrt{M_n}}\big(\psi(x,\nu_1),\ldots\psi(x,\nu_M)\big] \tag{114}$$

constructed from the integral representation of K by sampling M_n elements from π. Then,

$$M_n \geq c_0\sqrt{n}\log\frac{108\kappa^2\sqrt{n}}{\delta} \tag{115}$$

is enough to guarantee, with probability at least $1-\delta$, that

$$R(\hat{f}_{M_n,\lambda_n}) - R(f_{\mathcal{H}_K^C}^*) \leq \frac{c_1\log^2\frac{1}{\delta}}{\sqrt{n}}. \tag{116}$$

This statement demonstrates, that under reasonable assumptions, the estimator that is obtained with a number of random features proportional to $\mathcal{O}(\sqrt{n}\log n)$ achieves a $\mathcal{O}(1/\sqrt{n})$ learning error. We would now like to prove Theorem 2 by applying Theorem 3 to the classical PQC-kernel $K_{\mathcal{D}}$. To do this, we require the following Lemma:

Lemma 4: (Function set inclusions) *For any constant C one has that*

$$\begin{array}{c} \mathcal{F}^C_{(\Theta,\mathcal{D},O)} \subseteq \mathcal{F}^C_{\mathcal{D}} \subseteq \mathcal{F}_{\mathcal{D}} \subseteq \mathcal{H}_{K_{\mathcal{D}}} \\ \rotatebox{90}{$\supseteq$} \\ \mathcal{H}^C_{K_{\mathcal{D}}}. \end{array} \tag{117}$$

Proof of Lemma 4. The inclusions $\mathcal{F}^C_{(\Theta,\mathcal{D},O)} \subseteq \mathcal{F}^C_{\mathcal{D}} \subseteq \mathcal{F}_{\mathcal{D}}$ are immediate by definition. To prove $\mathcal{F}_{\mathcal{D}} \subseteq \mathcal{H}_{K_{\mathcal{D}}}$ we consider any $f_v \in \mathcal{F}_{\mathcal{D}}$ and let $\tilde{\mathcal{F}}$ be the image of $\mathcal{X}$ under $\phi_{\mathcal{D}}$. We can then write $v = v_1 + v_2$ where $v_1 \in \tilde{\mathcal{F}}$ and v_2 lies in the remainder. We then have that

$$\begin{aligned} f_v(x) &= \langle v, \phi_{\mathcal{D}}(x)\rangle \\ &= \langle v_1, \phi_{\mathcal{D}}(x)\rangle + \langle v_2, \phi_{\mathcal{D}}(x)\rangle \\ &= \langle v_1, \phi_{\mathcal{D}}(x)\rangle. \end{aligned} \tag{118}$$

By definition – i.e., the fact that $v_1 \in \tilde{\mathcal{F}}$ – we know that we can write

$$v_1 = \sum_j \gamma_j \phi_{\mathcal{D}}(x_j) \tag{119}$$

and, therefore, we see that

$$\begin{aligned} f_v(x) &= \sum_j \gamma_j \langle \phi_{\mathcal{D}}(x_j), \phi_{\mathcal{D}}(x)\rangle_{\mathcal{F}} \\ &= \sum_j \gamma_j K_{\mathcal{D}}(x_j, x), \end{aligned} \tag{120}$$

i.e., f_v is indeed in the RKHS $\mathcal{H}_{K_{\mathcal{D}}}$. Finally, the inclusion $\mathcal{F}^C_{\mathcal{D}} \subseteq \mathcal{H}^C_{K_{\mathcal{D}}}$ now follows easily, as $\mathcal{F}^C_{\mathcal{D}} \subseteq \mathcal{F}_{\mathcal{D}} \subseteq \mathcal{H}_{K_{\mathcal{D}}}$ yields

$$f \in \mathcal{F}^C_{\mathcal{D}} \implies f \in \mathcal{H}_{K_{\mathcal{D}}}, \tag{121}$$

and by definition

$$f \in \mathcal{F}^C_{\mathcal{D}} \implies \|f\|_{K_{\mathcal{D}}} \leq C \tag{122}$$

which together means that

$$f \in \mathcal{F}^C_{\mathcal{D}} \implies f \in \mathcal{H}^C_{K_{\mathcal{D}}}. \tag{123}$$

□

With this in hand, we can now prove Theorem 2.

Proof of Theorem 2. We start by recalling that, as shown in Section 4.4, for any reweighting vector w the reweighted PQC-kernel $K_{(\mathcal{D},\mathrm{w})}$ has the integral representation

$$K_{(\mathcal{D},\mathrm{w})}(x,x') = \frac{1}{2\pi}\int_{\mathcal{X}}\int_0^{2\pi} \sqrt{2}\cos(\langle \omega, x\rangle + \gamma)\sqrt{2}\cos(\langle \omega, x'\rangle + \gamma) q_{(\mathcal{D},\mathrm{w})}(\omega)\ \mathrm{d}\gamma \mathrm{d}\nu, \tag{124}$$

where

$$q_{(\mathcal{D},\mathrm{w})}(\omega) = \sum_{i=0}^{|\Omega_{\mathcal{D}}^{+}|} \frac{\mathrm{w}_i^2}{\|\mathrm{w}\|_2^2} \delta(\omega - \omega_i). \tag{125}$$

As a result, for any reweighting, including $\mathrm{w} = (1, \ldots, 1)$, the kernel $K_{(\mathcal{D},\mathrm{w})}$ satisfies assumption (1) of Theorem 3 with

$$\psi(x, \nu) = \sqrt{2}\cos(\langle \omega, x \rangle + \gamma). \tag{126}$$

Given this, we note that ψ is continuous in both variables and that $|\psi(x, \nu)| \leq \sqrt{2}$ for all x, ν – i.e., for any kernel $K_{(\mathcal{D},\mathrm{w})}$, assumption (2) of Theorem 3 is satisified with $\kappa = \sqrt{2}$.

Next, set the C appearing in Theorem 3 to the constant C appearing in assumption (1) of Theorem 2. More specifically, we apply Theorem 3 to the subset $\mathcal{H}^C_{K_{\mathcal{D}}}$, where C is an upper bound on the RKHS norm of the optimal function for P in $\mathcal{F}_{(\Theta,\mathcal{D},O)}$ – i.e., $\|f^*_{(\Theta,\mathcal{D},O)}\|_{K_{\mathcal{D}}} \leq C$. Doing this we obtain, via Theorem 3 and the fact that $\kappa = \sqrt{2}$, that provided all the conditions of Theorem 2 are satisfied, then

$$R(\hat{f}_{M_n,\lambda_n}) - R(f^*_{\mathcal{H}^C_{K_{\mathcal{D}}}}) \leq \frac{c_1 \log^2 \frac{1}{\delta}}{\sqrt{n}}. \tag{127}$$

To achieve the statement of Theorem 1 we then use the assumption that $\|f^*_{(\Theta,\mathcal{D},O)}\|_{K_{\mathcal{D}}} \leq C$. More specifically, via Lemma 4 this assumption implies that $f^*_{(\Theta,\mathcal{D},O)} \in \mathcal{H}^C_{K_{\mathcal{D}}}$, which together with the definition of $f^*_{\mathcal{H}^C_{K_{\mathcal{D}}}}$ as the *optimal* function in $\mathcal{H}^C_{K_{\mathcal{D}}}$, allows us to conclude that

$$R(f^*_{(\Theta,\mathcal{D},O)}) \geq R(f^*_{\mathcal{H}^C_{K_{\mathcal{D}}}}). \tag{128}$$

This then implies

$$\begin{aligned} R(\hat{f}_{M_n,\lambda_n}) - R(f^*_{(\Theta,\mathcal{D},O)}) &\leq R(\hat{f}_{M_n,\lambda_n}) - R(f^*_{\mathcal{H}^C_{K_{\mathcal{D}}}}) \\ &\leq \frac{c_1 \log^2 \frac{1}{\delta}}{\sqrt{n}} \qquad \text{[via Eq. (127)]} \end{aligned} \tag{129}$$

as per the statement of Theorem 2. □

As already mentioned, Theorem 1 in the main text then follows as an immediate corollary of Theorem 2.

C Proof of Lemma 1

Proof of Lemma 1. As discussed in Ref. [RBV10], the kernel integral operator is self-adjoint. In light of this, we know that $\|T_{K_{(\mathcal{D},\mathrm{w})}}\| = \rho(T_{K_{(\mathcal{D},\mathrm{w})}})$, where $\rho(T_{K_{(\mathcal{D},\mathrm{w})}})$ denotes the *spectral radius* of $\rho(T_{K_{(\mathcal{D},\mathrm{w})}})$. As such, we focus on determining the spectrum of $\rho(T_{K_{(\mathcal{D},\mathrm{w})}})$. To this end, note that under assumption (a) of the lemma statement we have that

$$\begin{aligned} (T_{K_{(\mathcal{D},\mathrm{w})}} g)(x) &= \int_{\mathcal{X}} K_{(\mathcal{D},\mathrm{w})}(x, x') g(x') dP_{\mathcal{X}}(x') \\ &= \frac{1}{(2\pi)^d} \int_{\mathcal{X}} K_{(\mathcal{D},\mathrm{w})}(x, x') g(x')\, \mathrm{d}x' \qquad \text{[via assumption (a)]} \end{aligned} \tag{130}$$

with

$$K_{(\mathcal{D},\mathrm{w})}(x,x') = \frac{1}{\|\mathrm{w}\|_2^2}\left(\mathrm{w}_0^2 + \sum_{i=1}^{|\Omega_{\mathcal{D}}^+|} \mathrm{w}_i^2 \cos\big(\langle \omega_i, (x-x')\rangle\big)\right), \tag{131}$$

where $\omega_0 = (0,\ldots,0)$. We now use Assumption (b) – i.e., that $\Omega_{\mathcal{D}^+} \subset \mathbb{Z}^d$ – to show that for any $\omega \in \mathbb{Z}^d$, the function $g(x') = \cos(\langle \omega, x'\rangle)$ is an eigenfunction of $T_{K_{(\mathcal{D},\mathrm{w})}}$. Specifically, using the following notation

$$\delta(\omega \pm \nu) \coloneqq \begin{cases} 1 & \text{if } (\omega = \nu) \vee (\omega = -\nu), \\ 0 & \text{else,} \end{cases} \tag{132}$$

and defining w_ω to be the weight associated with $\omega \in \Omega_{\mathcal{D}}$, we have that

$$\begin{aligned}
(T_{K_{(\mathcal{D},\mathrm{w})}} g)(x) &= \frac{1}{(2\pi)^d}\int_{\mathcal{X}} \left[\frac{1}{\|\mathrm{w}\|_2^2}\left(\mathrm{w}_0^2 + \sum_{i=1}^{|\Omega_{\mathcal{D}}^+|} \mathrm{w}_i^2 \cos\big(\langle \omega_i, (x-x')\rangle\big)\right)\right] \cos\big(\langle \omega, x'\rangle\big)\,\mathrm{d}x' \qquad (133)\\
&= \frac{1}{(2\pi)^d\|\mathrm{w}\|_2^2}\left[\int_{\mathcal{X}} w_0^2 \cos\big(\langle \omega, x'\rangle\big)\,\mathrm{d}x' + \int_{\mathcal{X}} \sum_{i=1}^{|\Omega_{\mathcal{D}}^+|} \mathrm{w}_i^2 \cos\big(\langle \omega_i, (x-x')\rangle\big)\cos\big(\langle \omega, x'\rangle\big)\,\mathrm{d}x'\right]\\
&= \frac{1}{(2\pi)^d\|\mathrm{w}\|_2^2}\int_{\mathcal{X}} \sum_{i=1}^{|\Omega_{\mathcal{D}}^+|} \mathrm{w}_i^2 \cos\big(\langle \omega_i, (x-x')\rangle\big)\cos\big(\langle \omega, x'\rangle\big)\,\mathrm{d}x' \qquad \text{[via assumption (b)]}\\
&= \frac{1}{(2\pi)^d\|\mathrm{w}\|_2^2}\sum_{i=1}^{|\Omega_{\mathcal{D}}^+|} \mathrm{w}_i^2 \int_{\mathcal{X}} \cos\big(\langle \omega_i, (x-x')\rangle\big)\cos\big(\langle \omega, x'\rangle\big)\,\mathrm{d}x'\\
&= \frac{1}{(2\pi)^d\|\mathrm{w}\|_2^2}\sum_{i=1}^{|\Omega_{\mathcal{D}}^+|} \mathrm{w}_i^2 \frac{(2\pi)^d}{2}\cos(\langle \omega_i, x\rangle)\delta(\omega_i \pm \omega) \qquad \text{[via assumption (b)]}\\
&= \begin{cases} \frac{\mathrm{w}_\omega^2}{2\|\mathrm{w}\|_2^2}\cos(\langle \omega, x\rangle) & \text{if } \omega \in \Omega_{\mathcal{D}}^+, \\ \frac{\mathrm{w}_{-\omega}^2}{2\|\mathrm{w}\|_2^2}\cos(\langle \omega, x\rangle) & \text{if } \omega \in -\Omega_{\mathcal{D}}^+, \\ 0 & \text{else.} \end{cases}
\end{aligned}$$

A similar calculation shows that, for all $\omega \in \mathbb{Z}^d$,

$$\Big(T_{K_{(\mathcal{D},\mathrm{w})}} \sin\big(\langle \omega, x'\rangle\big)\Big)(x) = \begin{cases} \frac{\mathrm{w}_\omega^2}{2\|\mathrm{w}\|_2^2}\sin(\langle \omega, x\rangle) & \text{if } \omega \in \Omega_{\mathcal{D}}^+, \\ \frac{\mathrm{w}_{-\omega}^2}{2\|\mathrm{w}\|_2^2}\sin(\langle \omega, x\rangle) & \text{if } \omega \in -\Omega_{\mathcal{D}}^+, \\ 0 & \text{else.} \end{cases} \tag{134}$$

As such, we have that all functions in the set $\{\sin\langle \omega, x\rangle) \,|\, \omega \in \mathbb{Z}^d\} \cup \{\cos\langle \omega, x\rangle) \,|\, \omega \in \mathbb{Z}^d\}$ are eigenfunctions of $T_{K_{\mathcal{D}}}$. However, as this set is a basis for $L^2(\mathcal{X}, P_{\mathcal{X}})$ – in the relevant case where $P_{\mathcal{X}}$ is the uniform distribution – we can conclude that

$$\begin{aligned}
\|T_{K_{(\mathcal{D},\mathrm{w})}}\| &= \rho\left(T_{K_{(\mathcal{D},\mathrm{w})}}\right) \qquad (135)\\
&= \max_{\omega\in\Omega_{\mathcal{D}}}\left\{\frac{1}{2}\frac{\mathrm{w}_\omega^2}{\|\mathrm{w}\|_2^2}\right\}\\
&= \max_{\omega\in\Omega_{\mathcal{D}}}\left\{\frac{1}{2}p_{(\mathcal{D},\mathrm{w})}(\omega)\right\}.
\end{aligned}$$

□

As an aside, it is interesting to note that the minimization of the norm $\|T_{K_{(\mathcal{D},\mathrm{w})}}\|$ subject to the constraint $\|\mathrm{w}\|_2^2 \leq c_0$ for some $c_0 > 0$ can be captured in terms of a convex semi-definite problem. This problem can be written as

$$\begin{aligned} & \text{minimize} && c && (136)\\ \text{subject to } & \frac{1}{2}\frac{\mathrm{w}_\omega^2}{\|\mathrm{w}\|_2^2} && \leq c, && (137)\\ & \|\mathrm{w}\|_2^2 && \leq c_0, \end{aligned}$$

which is easily seen to be equivalent with

$$\begin{aligned} & \text{minimize} && c && , && (138)\\ \text{subject to } & \begin{bmatrix} 2c & |\mathrm{w}_\omega| \\ |\mathrm{w}_\omega| & d \end{bmatrix} && \geq 0 \text{ for all } \omega, && (139)\\ & d && \leq c_0, \\ & \begin{bmatrix} d & \mathrm{w} \\ \mathrm{w}^T & I \end{bmatrix} && \geq 0, \end{aligned}$$

by making use of Schur complements.

D Proof of Lemma 3

Proof of Lemma 3. As $g_{\vec{\omega}}(x)$ is the output of the RFF procedure in which frequencies $\vec{\omega} = (\omega_1, \ldots, \omega_M)$ were drawn, we know that $g_{\vec{\omega}}$ can be written as

$$g_{\vec{\omega}}(x) = \sum_{\omega \in \tilde{\Omega}_{\mathcal{D}}} \hat{g}_{\vec{\omega}}(\omega) e^{i\langle \omega, x\rangle}, \tag{140}$$

where $\hat{g}_{\vec{\omega}}(\omega) = 0$ for all $\omega \notin \{\omega_i\}|_{i=1}^M$. Again we abuse notation and use $\hat{g}_{\vec{\omega}}$ to denote the vector with entries $\hat{g}_{\vec{\omega}}(\omega)$. Now, given some vector $\vec{\omega} = (\omega_1, \ldots, \omega_M) \in \Omega_{\mathcal{D}}^M$, we define the sets $\Omega_{\vec{\omega}} \coloneqq \{\omega_1, \ldots, \omega_M\} \subseteq \Omega_{\mathcal{D}}$ and $\tilde{\Omega}_{\vec{\omega}} \coloneqq \Omega_{\vec{\omega}} \cup (-\Omega_{\vec{\omega}})$. Given some $\hat{f}^*$, we then define the vectors

$$\hat{f}^*_{\vec{\omega}}(\omega) = \begin{cases} \hat{f}^*(\omega) \text{ if } \omega \in \tilde{\Omega}_{\vec{\omega}}, \\ 0 \text{ else}, \end{cases} \tag{141}$$

$$\hat{f}^*_{/\vec{\omega}}(\omega) = \begin{cases} 0 \text{ if } \omega \in \tilde{\Omega}_{\vec{\omega}}, \\ \hat{f}^*(\omega) \text{ else}. \end{cases} \tag{142}$$

Note that with these definitions, $\hat{f}^* = \hat{f}^*_{\vec{\omega}} + \hat{f}^*_{/\vec{\omega}}$. Using this, we have that

$$\begin{aligned} \|\hat{f}^* - \hat{g}_{\vec{\omega}}\|_2^2 &= \|\hat{f}^*_{\vec{\omega}} + \hat{f}^*_{/\vec{\omega}} - \hat{g}_{\vec{\omega}}\|_2^2 && (143)\\ &= \|\hat{f}^*_{/\vec{\omega}}\|_2^2 + \|\hat{f}^*_{\vec{\omega}} - \hat{g}_{\vec{\omega}}\|_2^2 \\ &\geq \|\hat{f}^*_{/\vec{\omega}}\|_2^2 && (144)\\ &= \|\hat{f}^*\|_2^2 - \|\hat{f}^*_{\vec{\omega}}\|_2^2. && (145) \end{aligned}$$

Using this expression, we can then lower-bound $\hat{\epsilon}$, the expected L^2-norm of the difference between the optimal PQC function and the output of the RFF procedure, recalling that $\xi(\vec{\omega})$ is the

probability of sampling the vector of frequencies $\vec{\omega}$,

$$\hat{\epsilon} = \sum_{\vec{\omega}\in\Omega_{\mathcal{D}}^{M}} \|f^* - g_{\vec{\omega}}\|_2^2\, \xi(\vec{\omega}) \tag{146}$$

$$= (2\pi)^d \sum_{\vec{\omega}\in\Omega_{\mathcal{D}}^{M}} \|\hat{f}^* - \hat{g}_{\vec{\omega}}\|_2^2 \xi(\vec{\omega}) \qquad \text{[via Parseval's identity]} \tag{147}$$

$$\geq (2\pi)^d \sum_{\vec{\omega}\in\Omega_{\mathcal{D}}^{M}} \left[\|\hat{f}^*\|_2^2 - \|\hat{f}^*_{\vec{\omega}}\|_2^2\right] \xi(\vec{\omega}) \qquad \text{[via Eq. (145)]}$$

$$= (2\pi)^d \|\hat{f}^*\|_2^2 - (2\pi)^d \sum_{\vec{\omega}\in\Omega_{\mathcal{D}}^{M}} \|\hat{f}^*_{\vec{\omega}}\|_2^2\, \xi(\vec{\omega}). \tag{148}$$

Note that we used the assumption of integer-valued frequency vectors – and therefore orthogonal components of the feature map – when invoking Parseval's identity to move from line (146) to (147). Using the short-hand notation 0 to denote the frequency vector $(0,\ldots,0)$, we can now analyze the final term as

$$\begin{aligned}
\sum_{\vec{\omega}\in\Omega_{\mathcal{D}}^{M}} \|\hat{f}^*_{\vec{\omega}}\|_2^2\, \xi(\vec{\omega}) &= \sum_{\substack{\vec{\omega}\in\Omega_{\mathcal{D}}^{M} \\ 0\notin\Omega_{\vec{\omega}}}} \sum_{i=1}^{M} \left(|\hat{f}^*(-\omega_i)|^2 + |\hat{f}^*(\omega_i)|^2\right) \xi(\vec{\omega}) \\
&\qquad + \sum_{\substack{\vec{\omega}\in\Omega_{\mathcal{D}}^{M} \\ 0\in\Omega_{\vec{\omega}}}} \left[\sum_{i=1}^{M-1} \left(|\hat{f}^*(-\omega_i)|^2 + |\hat{f}^*(\omega_i)|^2\right) + |\hat{f}^*(0)|^2\right] \xi(\vec{\omega}) \\
&\leq \sum_{\substack{\vec{\omega}\in\Omega_{\mathcal{D}}^{M} \\ 0\notin\Omega_{\vec{\omega}}}} \sum_{i=1}^{M} 2|\hat{f}^*(\omega_i)|^2 \xi(\vec{\omega}) \\
&\qquad + \sum_{\substack{\vec{\omega}\in\Omega_{\mathcal{D}}^{M} \\ 0\in\Omega_{\vec{\omega}}}} \left[\sum_{i=1}^{M-1} 2|\hat{f}^*(\omega_i)|^2 + 2|\hat{f}^*(0)|^2\right] \xi(\vec{\omega}) \\
&= 2\sum_{i=1}^{M} \sum_{\omega_1\in\Omega_{\mathcal{D}}} \cdots \sum_{\omega_M\in\Omega_{\mathcal{D}}} |\hat{f}^*(\omega_i)|^2 p(\omega_1)\cdots p(\omega_M) \\
&= 2\sum_{i=1}^{M} \sum_{\omega_i\in\Omega_{\mathcal{D}}} |\hat{f}^*(\omega_i)|^2 p(\omega_i) \\
&= 2M \sum_{\nu\in\Omega_{\mathcal{D}}} |\hat{f}^*(\nu)|^2 p(\nu).
\end{aligned} \tag{149}$$

Substituting Eq. (149) into Eq. (148) then gives the statement of the Lemma. □

References

[ABK+23] Eric R. Anschuetz, Andreas Bauer, Bobak T. Kiani, and Seth Lloyd. "Efficient classical algorithms for simulating symmetric quantum systems". In: *Quantum* 7 (Nov. 2023), p. 1189. ISSN: 2521-327X. URL: https://doi.org/10.22331/q-2023-11-28-1189 (page 2).

[BLS+19] Marcello Benedetti, Erika Lloyd, Stefan Sack, and Mattia Fiorentini. "Parameterized quantum circuits as machine learning models". In: *Quantum Science and Technology* 4.4 (2019), p. 043001. URL: https://doi.org/10.1088%2F2058-9565%2Fab4eb5 (pages 1, 6).

[CAB+21] Marco Cerezo, Andrew Arrasmith, Ryan Babbush, Simon C. Benjamin, Suguru Endo, Keisuke Fujii, Jarrod R. McClean, Kosuke Mitarai, Xiao Yuan, Lukasz Cincio, and Patrick J. Coles. "Variational quantum algorithms". In: *Nature Reviews Physics* 3.9 (2021), pp. 625–644. URL: https://doi.org/10.1038%2Fs42254-021-00348-9 (page 1).

[CGM+21] Matthias C. Caro, Elies Gil-Fuster, Johannes Jakob Meyer, Jens Eisert, and Ryan Sweke. "Encoding-dependent generalization bounds for parametrized quantum circuits". In: *Quantum* 5 (2021), p. 582. URL: https://doi.org/10.22331/q-2021-11-17-582 (pages 7, 9, 18).

[CSK16] Jen-Hao Rick Chang, Aswin C. Sankaranarayanan, and B. V. K. Vijaya Kumar. "Random Features for Sparse Signal Classification". In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 5404–5412. URL: https://doi.org/10.1109/CVPR.2016.583 (page 33).

[FRD+22] Enrico Fontana, Ivan Rungger, Ross Duncan, and Cristina Cîrstoiu. "Spectral analysis for noise diagnostics and filter-based digital error mitigation". In: (2022). arXiv: 2206.08811 [quant-ph]. URL: https://doi.org/10.48550/arXiv.2206.08811 (page 2).

[FRD+23] Enrico Fontana, Manuel S. Rudolph, Ross Duncan, Ivan Rungger, and Cristina Cîrstoiu. "Classical simulations of noisy variational quantum circuits". In: (2023). arXiv: 2306.05400. URL: https://doi.org/10.48550/arXiv.2306.05400 (pages 2, 33).

[FV12] Andrew J. Ferris and Guifre Vidal. "Perfect sampling with unitary tensor networks". In: *Phys. Rev. B* 85 (16 2012), p. 165146. URL: https://doi.org/10.1103/PhysRevB.85.165146 (page 19).

[Gol17] Oded Goldreich. *Introduction to property testing*. Cambridge University Press, 2017. URL: https://doi.org/10.1017/9781108135252 (page 32).

[GSP+19] Ivan Glasser, Ryan Sweke, Nicola Pancotti, Jens Eisert, and Ignacio Cirac. "Expressive power of tensor-network factorizations for probabilistic modeling". In: *Advances in Neural Information Processing Systems*. Ed. by H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett. Vol. 32. Curran Associates, Inc., 2019. URL: https://doi.org/10.48550/arXiv.1907.03741 (page 19).

[GT20] Francisco Javier Gil Vidal and Dirk Oliver Theis. "Input redundancy for parameterized quantum circuits". In: *Frontiers in Physics* 8 (2020), p. 297. URL: https://doi.org/10.3389/fphy.2020.00297 (page 7).

[JFN+23] Sofiene Jerbi, Lukas J. Fiderer, Hendrik Poulsen Nautrup, Jonas M. Kübler, Hans J. Briegel, and Vedran Dunjko. "Quantum machine learning beyond kernel methods". In: *Nature Communications* 14 (2023), p. 517. URL: https://doi.org/10.1038/s41467-023-36159-y (page 8).

[JGM+24] Sofiene Jerbi, Casper Gyurik, Simon C. Marshall, Riccardo Molteni, and Vedran Dunjko. "Shadows of quantum machine learning". In: *Nature Communications* 15.1

(July 2024). ISSN: 2041-1723. URL: https://doi.org/10.1038/s41467-024-49877-8 (pages 2, 3, 33).

[KR21] Behnoush Khavari and Guillaume Rabusseau. *Lower and Upper Bounds on the VC-Dimension of Tensor Network Models*. 2021. arXiv: 2106.11827 [cs.LG]. URL: https://doi.org/10.48550/arXiv.2106.11827 (page 11).

[LSS+22] Martín Larocca, Frédéric Sauvage, Faris M. Sbahi, Guillaume Verdon, Patrick J. Coles, and M. Cerezo. "Group-Invariant Quantum Machine Learning". In: *PRX Quantum* 3.3 (Sept. 2022). ISSN: 2691-3399. URL: https://doi.org/10.1103/PRXQuantum.3.030341 (page 2).

[LTD+22] Jonas Landman, Slimane Thabet, Constantin Dalyac, Hela Mhiri, and Elham Kashefi. "Classically approximating variational quantum machine learning with random Fourier features". In: (2022). arXiv: 2210.13200. URL: https://doi.org/10.48550/arXiv.2210.13200 (pages 3, 10, 11, 34).

[MMG+23] 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". In: *PRX Quantum* 4 (1 2023), p. 010328. URL: https://doi.org/10.1103/PRXQuantum.4.010328 (page 2).

[MRT18] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2018 (page 4).

[RBV10] Lorenzo Rosasco, Mikhail Belkin, and Ernesto De Vito. "On Learning with Integral Operators". In: *Journal of Machine Learning Research* 11.30 (2010), pp. 905–934. URL: http://jmlr.org/papers/v11/rosasco10a.html (page 38).

[RR07] Ali Rahimi and Benjamin Recht. "Random Features for Large-Scale Kernel Machines". In: *Advances in Neural Information Processing Systems*. Ed. by J. Platt, D. Koller, Y. Singer, and S. Roweis. Vol. 20. Curran Associates, Inc., 2007. URL: https://proceedings.neurips.cc/paper_files/paper/2007/file/013a006f03dbc5392effeb8f18fda755-Paper.pdf (pages 3, 5, 6).

[RR17] Alessandro Rudi and Lorenzo Rosasco. "Generalization Properties of Learning with Random Features". In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett. Vol. 30. Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/61b1fb3f59e28c67f3925f3c79be81a1-Paper.pdf (pages 5, 13, 15, 33, 36).

[SC08] Ingo Steinwart and Andreas Christmann. *Support vector machines*. Springer Science & Business Media, 2008. URL: https://doi.org/10.1007/978-0-387-77242-4 (pages 16, 22).

[Sch11] Ulrich Schollwöck. "The density-matrix renormalization group in the age of matrix product states". In: *Annals of Physics* 326.1 (Jan. 2011), pp. 96–192. ISSN: 0003-4916. URL: https://doi.org/10.1016/j.aop.2010.09.012 (page 19).

[Sch21] Maria Schuld. "Supervised quantum machine learning models are kernel methods". In: (2021). arXiv: 2101.11020. URL: https://doi.org/10.48550/arXiv.2101.11020 (page 8).

[SEM23] Franz J. Schreiber, Jens Eisert, and Johannes Jakob Meyer. "Classical surrogates for quantum learning models". In: *Physical Review Letters* 131 (2023), p. 100803. URL: https://doi.org/10.1103/PhysRevLett.131.100803 (pages 2, 3).

[SS15a] Bharath Sriperumbudur and Zoltan Szabo. "Optimal Rates for Random Fourier Features". In: *Advances in Neural Information Processing Systems*. Ed. by C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett. Vol. 28. Curran As-

sociates, Inc., 2015. URL: https://proceedings.neurips.cc/paper_files/paper/2015/file/d14220ee66aeec73c49038385428ec4c-Paper.pdf (page 36).

[SS15b] Danica J. Sutherland and Jeff Schneider. *On the Error of Random Fourier Features*. 2015. arXiv: 1506.02785 [cs.LG]. URL: https://doi.org/10.48550/arXiv.1506.02785 (page 6).

[SSM21] Maria Schuld, Ryan Sweke, and Johannes Jakob Meyer. "Effect of data encoding on the expressive power of variational quantum-machine-learning models". In: *Phys. Rev. A* 103 (3 2021), p. 032430. URL: https://doi.org/10.1103/PhysRevA.103.032430 (pages 2, 7, 23).

[STJ24] Seongwook Shin, Yong Siah Teo, and Hyunseok Jeong. "Dequantizing quantum machine learning models using tensor networks". In: *Phys. Rev. Res.* 6 (2 2024), p. 023218. URL: https://doi.org/10.1103/PhysRevResearch.6.023218 (pages 11, 33).

[SW10] E M Stoudenmire and Steven R White. "Minimally entangled typical thermal state algorithms". In: *New Journal of Physics* 12.5 (May 2010), p. 055026. ISSN: 1367-2630. URL: https://doi.org/10.1088/1367-2630/12/5/055026 (page 19).

[SWC+23] Yuguo Shao, Fuchuan Wei, Song Cheng, and Zhengwei Liu. "Simulating quantum mean values in noisy variational quantum algorithms: A polynomial-scale approach". In: (2023). arXiv: 2306.05804. URL: https://doi.org/10.48550/arXiv.2306.05804 (pages 2, 33).