Gradients and frequency profiles of quantum re-loading models

Barthe, A.M.; Pérez Salinas, A.

Gradients and frequency profiles of quantum re-uploading models

Alice Barthe[1,2,3] and Adrián Pérez-Salinas[1,3]

[1]⟨aQa^L⟩ Applied Quantum Algorithms, Universiteit Leiden
[2]Quantum Technology Initiative, CERN, Geneva, Switzerland
[3]Instituut-Lorentz, Universiteit Leiden, the Netherlands

Quantum re-uploading models have been extensively investigated as a form of machine learning within the context of variational quantum algorithms. Their trainability and expressivity are not yet fully understood and are critical to their performance. In this work, we address trainability through the lens of the magnitude of the gradients of the cost function. We prove bounds for the differences between gradients of the better-studied dataless parameterized quantum circuits and re-uploading models. We coin the concept of *absorption witness* to quantify such difference. For the expressivity, we prove that quantum re-uploading models output functions with vanishing high-frequency components and upper-bounded derivatives with respect to data. As a consequence, such functions present limited sensitivity to fine details and offer protection against overfitting. We performed numerical experiments extending the theoretical results to more relaxed and realistic conditions. Overall, future designs of quantum re-uploading models will benefit from the strengthened knowledge delivered by the uncovering of absorption witnesses and vanishing high frequencies.

1 Introduction

Variational Quantum Algorithms (VQAs) have emerged as a prominent paradigm in the realm of quantum computing as a hybrid computational model suited for NISQ (Noisy Intermediate-Scale Quantum) [1] devices in conjunction with classical optimization techniques [2, 3]. These algorithms rely on the minimization of cost functions [4, 5], which encode specific computational problems. VQAs have been used to solve a variety of problems, including approximating ground states [6, 7, 8], combinatorial challenges [9], chemistry problems [10] and simulation of quantum systems [11, 12, 13, 14, 15]. Furthermore, VQAs have served as quantum computing engines for tackling various machine learning (ML) tasks, such as function regression [16, 17], classification [18, 19, 20] or generative models [21, 22, 23]. We specifically make a distinction between linear models on the one side, introducing data either as input states or through encoding maps [24], and quantum re-uploading (QRU) schemes on the other side, which introduce data iteratively throughout the execution of the quantum circuit [25, 26, 27, 28].

The performance of VQAs hinges on two critical properties: expressivity and trainability. Expressivity embodies the model's ability to represent precise solutions to the underlying problem, while trainability is a measure of the difficulty in finding the parameter set that yields the optimal attainable solution within the model. In the case of data-independent VQAs, expressivity can be intuitively understood as the proportion of attainable output states within the Hilbert space, quantified through closeness to t-designs [29]. In the context of data-dependent ML, expressivity pertains to the suitability of the output function in fitting the data [30, 31]. The universality of QRU models has been proven even with a single qubit [26, 32]. Trainability in VQAs, on the other hand, is closely linked to characteristics of the cost function, such as non-convexity [33] or vanishing gradients [34]. The relationship between the trainability of VQAs and QML schemes has been previously explored, in the absence of re-uploading [35]. Importantly, trainability and expressivity are usually mutually exclusive, and for VQAs in particular there exists a well-studied trade-off between these two properties [36, 37].

In this work, we specifically explore QRU models with a focus on exploring trainability and expressivity. Our investigation into trainability focuses on the on-average behavior of gradients, which can be related to the flatness of the cost function. We compare the cost functions of QRU models and base PQCs, which are circuits with the same architecture and observable as the QRU, but where the data gates are removed. The difference between the flatness of both cost functions is upper bounded by a quantity we refer to as *absorption witness*, which quantifies the influence of data gates on the quantum circuit when averaged over the parameter space. Such derivation opens a path to transfer existing knowledge about the flatness of PQCs [34, 38, 39] to guide the design of QRU models.

The second segment of our findings is related to

the expressivity of data-dependent output functions generated by QRU models. It is known that any hypothesis function output by QRU models can be expressed as a generalized trigonometric polynomial, with the range of available frequencies contingent on the data encoding scheme [40, 26]. We show that, under reasonable assumptions, the average magnitude of individual frequency components in the hypothesis function rapidly tends to a Gaussian profile, with a variance scaling as $\sim \sqrt{L}$, with L being the number of re-uploading steps, while the support in frequencies scales as $\sim L$. This property inherently biases the attainable hypothesis functions as being heavily dominated by lower-frequency components. This has a direct consequence on the Lipschitz constants of these output functions.

The paper is organized as follows. Section 2 introduces relevant concepts and notation for the paper. Section 3 delves into the expected norm of gradients in QRU models. Section 4 delves into the expressive capabilities of output functions within QRU models in terms of spectrum.Both sections are supported by numerical experiments showcasing agreement with our theoretical findings. Section 5 engages in a discussion of the implications and potential avenues opened up by our research. Conclusions are summarized in Section 6.

2 Background

In this work, we refer to a PQC as a sequence of parameterized gates and fixed gates applied to an initial state, namely

$$U(\boldsymbol{\theta}) = \prod_{j=1}^{M} W_j e^{iV_j\theta_i}, \tag{1}$$

where $\{V_j\}$ are, without loss of generality, traceless Hermitian matrices known as generators, $\{W_j\}_{j=1}^M$ are fixed unitary operations, and $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subset \mathbb{R}^M$. We use these PQCs as a baseline for QRU models [27, 32, 26]. This model consists of a PQC where data-encoding gates have been added in the form

$$U(\boldsymbol{\theta}, x) = \prod_{j=1}^{M} e^{ig_j x} W_j e^{iV_j\theta_i}, \tag{2}$$

where $\{g_j\}_{j=1}^M$ are, without loss of generality, traceless Hermitian generators. We do not impose constraints in $\{V_j\} \cap \{g_j\}$. The input x is a real number. Extensions to multidimensional values of x are available by adding extra terms to the model, although this case will not be considered in this work. Notice that there exists a mapping between PQCs with data as initial state and QRU models, thus making both computations formally equivalent [41], up to overheads.

QRU models yield $\boldsymbol{\theta}$-dependent hypothesis functions

$$h_{\boldsymbol{\theta}}(x) = \langle 0| U^\dagger(\boldsymbol{\theta}, x) H U(\boldsymbol{\theta}, x) |0\rangle, \tag{3}$$

when applied to an initial quantum state and measured with an observable H. Notice that $h_{\boldsymbol{\theta}}(x = x_0)$ for a fixed value x_0 is the standard definition of the cost function of a PQC. In the case $x_0 \neq 0$, we can recover our formulation of PQC by adapting the fixed gates W_j. The values of $\boldsymbol{\theta}$ are trainable to match data coming in pairs $X = \{(x, y(x)\}$, such that $h_{\boldsymbol{\theta}}(x) \approx y(x)$. These hypothesis functions can be expressed as a generalized trigonometric polynomial [26, 40], namely

$$h_{\boldsymbol{\theta}}(x) = \sum_{\omega \in \Omega} a_\omega(\boldsymbol{\theta}) e^{i\omega x}, \tag{4}$$

where $a_\omega(\boldsymbol{\theta}) = a^*_{-\omega}(\boldsymbol{\theta})$ to ensure real valued hypothesis functions, and Ω is the set of available frequencies.

The training of VQAs involves an optimization procedure where a parameter set minimizing a cost function is searched. The difficulty of this optimization task is enclosed under the broad concept of trainability. A paradigmatic example of optimizing a PQC is minimizing $h_{\boldsymbol{\theta}}(x = 0)$ with respect to $\boldsymbol{\theta}$ to find an approximation to the ground state of the corresponding Hamiltonian H. Trainability has been extensively studied in the context of PQCs [33, 34]. Training a QRU model involves finding the optimal set of parameters $\boldsymbol{\theta}$ for which $h_{\boldsymbol{\theta}}(x)$ approximately matches some target function given by data. Trainability may depend on several features of the cost function landscape [42], such as small gradients [43] or non-convexity, e.g. the existence of (many) local minima [44]. In this work, we focus on average behaviors of gradients of the cost function, inspired by the well-studied phenomenon of vanishing gradients barren plateaus (BP) [34, 3, 36].

Expressivity is another crucial aspect of parameterized models, capturing their ability to represent various solutions. For PQCs, expressivity entails the existence of parameter sets $\boldsymbol{\theta}^*$ making $U(\boldsymbol{\theta}^*, 0)$ close to some unitary operations $V \in \mathcal{SU}(2^n)$, within a specified tolerance and respect to some distance. Expressivity is often measured relative to unitary t-designs [29]. In contrast, expressivity in the context of ML (e.g. QRU models) is related to the output function and its capability. A model is expressive if its output is able to match a variety of target functions to fit some data [45].

3 Gradients in QRU models

3.1 Losses for PQC and QRU

In this section, we focus on characterizing gradients of QRU models, as compared to those of PQCs. In the case of PQC, the gradients of interest are typically defined in relationship to their cost function $h_{\boldsymbol{\theta}}(0)$. However, the optimization of QRU models involves a cost function that depends on both the quantum circuit, expressed through $h_{\boldsymbol{\theta}}(x)$, and the available data,

provided in pairs as $(x, y(x))$. Such cost function is usually given by averaging a distance between functions $\Delta(\cdot,\cdot)$ as

$$\mathcal{L}_X(\boldsymbol{\theta}) = \mathbb{E}_X\left(\Delta(h_{\boldsymbol{\theta}}(x), y(x))\right), \tag{5}$$

where $\mathbb{E}_X(\cdot)$ denotes expectation value over the training dataset $X = \{(x, y(x))\}$, usually composed by a discrete set of points. Notice that $\mathcal{L}_X$ is empirical as the training dataset is drawn from an unknown data distribution $X \sim \mathcal{D}$, and approximates the true unaccessible risk averaged over $\mathcal{D}$. In regression tasks, a common choice for the distance metric $\Delta(\cdot,\cdot)$ is the mean squared error.

Our interest lies in examining the gradients of the loss function of QRU models, expressed as

$$\partial_j \mathcal{L}(\boldsymbol{\theta}) = \mathbb{E}_X\left(\frac{\partial \Delta(h_{\boldsymbol{\theta}}(x), y(x))}{\partial h_{\boldsymbol{\theta}}(x)} \partial_j h_{\boldsymbol{\theta}}(x)\right). \tag{6}$$

The influence imposed by the choice of distance function $\Delta(\cdot,\cdot)$ can be readily bounded e. g. using its Lipschitz constant L_Δ. In particular,

$$\mathrm{Var}_{\boldsymbol{\Theta}}\left(\partial_j \mathcal{L}(\boldsymbol{\theta})\right) \leq L_\Delta^2 \,\mathrm{Var}_{\boldsymbol{\Theta}}\left(\mathbb{E}_X\left(\partial_j h_{\boldsymbol{\theta}}(x)\right)\right). \tag{7}$$

Therefore, we can bound vanishing gradients by studying only $\mathrm{Var}_{\boldsymbol{\Theta}}\left(\mathbb{E}_X\left(\partial_j h_{\boldsymbol{\theta}}(x)\right)\right)$. It is important to highlight that all results presented in this section are applicable for any distribution over parameters $\boldsymbol{\Theta}$.

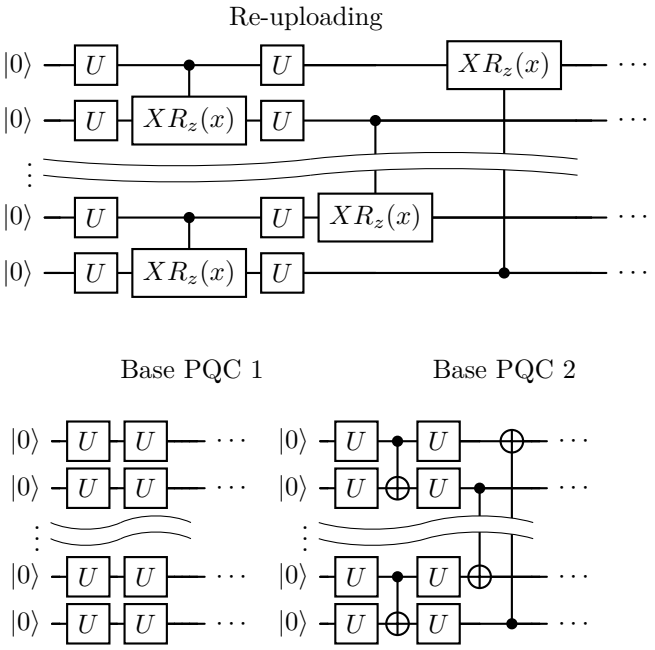


Figure 1: Quantum circuits for the first experiments. The re-uploading model is depicted on top, and compared to the PQCs described in the bottom line. The U gates here described correspond to arbitrary parameterized single-qubit operations. The circuit here described corresponds to one layer, and the depth is determined by the number of repetitions.

3.2 Gradient of the loss function

We connect now the gradients of loss functions for QRU models and PQCs. First, the average of derivatives of hypothesis functions are zero, namely [36]

$$\mathbb{E}_{\boldsymbol{\Theta}}\left(\mathbb{E}_X\left(\partial_j h_{\boldsymbol{\theta}}(x)\right)\right) = \mathbb{E}_X\left(\mathbb{E}_{\boldsymbol{\Theta}}\left(\partial_j h_{\boldsymbol{\theta}}(x)\right)\right) = 0, \tag{8}$$

if the parameters $\boldsymbol{\theta}$ are sampled uniformly from $\boldsymbol{\Theta}$. As a consequence, due to the convexity of the square function, we have $\mathbb{E}(x)^2 \leq \mathbb{E}\left(x^2\right)$. By combining these two observations and the definition $\mathrm{Var}(x) = \mathbb{E}\left(x^2\right) - \mathbb{E}(x)^2$, we derive

$$\mathrm{Var}_{\boldsymbol{\Theta}}\left(\mathbb{E}_X\left(\partial_j h_{\boldsymbol{\theta}}(x)\right)\right) \leq \mathbb{E}_X\left(\mathrm{Var}_{\boldsymbol{\Theta}}\left(\partial_j h_{\boldsymbol{\theta}}(x)\right)\right), \tag{9}$$

which can be readily connected to $\mathrm{Var}_{\boldsymbol{\Theta}}\left(\partial_j \mathcal{L}(\boldsymbol{\theta})\right)$ via Equation (7). Therefore, we can bound the variance of cost functions in QRU by the average of variances cost functions of several PQC, defined by different fixed x_0. Bounds on $\mathrm{Var}_{\boldsymbol{\Theta}}\left(\partial_j h_{\boldsymbol{\theta}}(0)\right)$ have been studied in the context of PQC. In particular, BPs are defined for exponentially vanishing bounds to $\mathrm{Var}_{\boldsymbol{\Theta}}\left(\partial_j h_{\boldsymbol{\theta}}(0)\right)$ [34].

Equation (9) suggests that QRU models present vanishing gradients if the base PQC presents BPs, which means that it is recommendable to use architectures that avoid vanishing gradients such as in [39, 46] when designing QRUs. This statement is made from a purely trainability point of view, as it has been shown that such architectures are indeed classically simulable. We are only highlighting that given a QRU model, if removing the reuploading gates yields a PQC architecture known to suffer from vanishing gradients, then vanishing gradients will also affect the QRU architecture. In addition, Equation (9) does not guarantee non-vanishing gradients for QRU models derived from BP-free PQCs. As an example, consider a re-uploading model with parameterized single-qubit gates and data-encoding entangling gates arranged in an alternate-layered structure, measured by a sum of 1-local observables, as illustrated in Figure 1, base PQC 1. A compatible base PQC is composed only of single-qubit parameterized gates. In this case, $h_{\boldsymbol{\theta}}(x_0)$ from the base PQC has large gradients [47]. The inclusion of data increases the accessible Hilbert space due to the presence of entangling operations, entanglement, which causes BPs [36].

Consider again the previous example, this time with a different base PQC which includes entangling gates, see Figure 1, base PQC 2. The gates U are considered distributed according to the Haar measure for single-qubit operations. In this new scenario, the PQC cost function $h_{\boldsymbol{\theta}}(0)$ suffers from BPs for sufficient depth [36, 47]. Notice that it is possible to decompose a data-dependent entangling gate as a fixed entangling gate and tunable single-qubit [48], allowing for introducing data through single-qubit operations. Intuitively, the data can be re-absorbed by the param-

eters to generate a new circuit with the same ansatz as the base PQC. As a direct consequence, the gradients of the PQC and those of the QRU are of the same magnitude, $\mathbb{E}_X\left(\mathrm{Var}_{\boldsymbol{\Theta}}\left(\partial_j h_{\boldsymbol{\theta}}(x)\right)\right) \approx \mathrm{Var}_{\boldsymbol{\Theta}}\left(\partial_j h_{\boldsymbol{\theta}}(x_0)\right)$, for any x_0. This intuition motivates the newly coined concept of *absorption witnesses* in Definition 3.1 as the capability of the circuit to absorb the data into the parameters.

3.3 Absorption Witnesses

Before further expanding on absorption witnesses, it is convenient to introduce some auxiliary quantities in the context of QRU models. We take derivatives with respect to the j-th parameter. All operations preceding the j-th parameter (not included) are considered the right part of the circuit, operationally attached to the input state ρ_0. Operations that include and follow the j-th parameter are on the left side of the circuit, attached to the observable. This description is given by

$$\rho_j(\boldsymbol{\theta}_{R,j}, x) = U_{R,j}(\boldsymbol{\theta}_{R,j}, x)\ \rho_0\ U^\dagger_{R,j}(\boldsymbol{\theta}_{R,j}, x) \quad (10)$$

$$H_j(\boldsymbol{\theta}_{L,j}, x) = U^\dagger_{L,j}(\boldsymbol{\theta}_{L,j}, x)\ H\ U_{L,j}(\boldsymbol{\theta}_{L,j}, x). \quad (11)$$

The left/right parameters $\boldsymbol{\Theta}_{R/L,j}$ are assumed to be independent. For each of the right and left parts of the circuit, we can define the difference with respect to the reference data value $x = 0$ (corresponding to the PQC) as

$$B^{(t)}_{R,j}(\boldsymbol{\theta}_{R,j}, x; \rho_0) = \rho_j^{\otimes t}(\boldsymbol{\theta}_{R,j}, x) - \rho_j^{\otimes t}(\boldsymbol{\theta}_{R,j}, 0) \quad (12)$$

$$B^{(t)}_{L,j}(\boldsymbol{\theta}_{L,j}, x; H) = H_j^{\otimes t}(\boldsymbol{\theta}_{L,j}, x) - H_j^{\otimes t}(\boldsymbol{\theta}_{L,j}, 0) \quad (13)$$

We define the absorption witness as follows.

Definition 3.1 (Absorption witness)**.** *Let $U(\boldsymbol{\theta}, x)$ be a re-uploading model as defined in Equation* (2)*. Let $U_{R/L,j}(\boldsymbol{\theta}, x)$ be the right and left parts of the circuit with respect to the j-th gate. The right/left absorption witnesses are*

$$\mathcal{B}^{(2)}_{R,j}(\rho_0) = \mathbb{E}_X\left(\left\|\mathbb{E}_{\boldsymbol{\Theta}_{R,j}}\left(B^{(2)}_{R,j}(\boldsymbol{\theta}_{R,j}, x; \rho_0)\right)\right\|_1\right), \quad (14)$$

$$\mathcal{B}^{(2)}_{L,j}(H) = \mathbb{E}_X\left(\left\|\mathbb{E}_{\boldsymbol{\Theta}_{L,j}}\left(B^{(2)}_{L,j}(\boldsymbol{\theta}_{L,j}, x; H)\right)\right\|_1\right). \quad (15)$$

The absorption witness defined above captures the effect of including data when averaging over the parameter space $\boldsymbol{\Theta}$. If $\mathcal{B}^{(2)}_{R,j}(\rho_0) = 0$, the input x yields an effect on $\rho_j(\boldsymbol{\theta}_{R,j}, x)$ equivalent to some change $\boldsymbol{\theta}_{R,j} \to \boldsymbol{\theta}^*_{R,j}$. This effect is compensated when averaging over $\boldsymbol{\Theta}$. The logic is analogous to the left part of the circuit. As an illustrative example, assume a single-layer re-uploading model composed by applying any data-encoding layer after a PQC forming a t-design. By definition, t-designs approximate up to the t-th statistical moment of Haar measure and are thus insensitive (on average) to adding extra operations, in particular any operation given by data-encoding. However, this closeness to t-designs is no longer possible for ansatzes with (several) data-encoding gates interspersed between parameterized layers.

The absorption witnesses from Definition 3.1 bound the differences between variances for PQCs and QRU models as follows.

Theorem 3.1. *Let $U(\boldsymbol{\theta}, x)$ be a re-uploading model as defined in Equation* (2)*. Then*

$$\begin{aligned}&\left|\mathbb{E}_X\left(\mathrm{Var}_{\boldsymbol{\Theta}}\left(\partial_j h_{\boldsymbol{\theta}}(x)\right)\right) - \mathrm{Var}_{\boldsymbol{\Theta}}\left(\partial_j h_{\boldsymbol{\theta}}(0)\right)\right| \\ &\leq 4\|V_j\|^2_\infty \left(\|H\|^2_\infty\ \mathcal{B}^{(2)}_{R,j}(\rho_0) + \ \|\rho_0\|^2_\infty\ \mathcal{B}^{(2)}_{L,j}(H)\right)\end{aligned} \quad (16)$$

where ρ_0 is the initial state, H is the observable to measure, and $\mathcal{B}^{(2)}_{L,j}(H), \mathcal{B}^{(2)}_{R,j}(\rho_0)$ are the absorption witnesses from Definition 3.1.

The proof can be found in Appendix A.1.

Computing the absorption witness is not easy, nor computationally efficient. Alternatively, it can be estimated by comparing variances of the magnitudes of gradients with and without data, as will be done in the numerical calculations of Section 3.5. Nevertheless, it provides a useful interpretation of the relationship between vanishing gradients for data-dependent QRU models, as compared to their base PQCs, where the BP phenomenon has been already studied [34, 38, 39]. The absorption witnesses quantify the expressivity difference between the PQC where the data uploading gates of the QRU are removed and that of the PQC where the data uploading gates are replaced by parameterized gates. As an illustrative example, consider an arbitrary Hamiltonian H, and a quantum re-uploading model composed of a gate $e^{iH\theta_0}$ immediately followed by a data uploading gate e^{iHx_0}. These gates can be combined as $e^{iH\theta_1}, \theta_1 = \theta_0 + x$. Since the relevant quantities are variances of, any shift of this kind does not affect the average behavior, yielding an absorption witness of exactly 0. On the other hand, a QRU model composed by two arbitrary Hamiltonians $e^{iH_1\theta}$, $e^{iH_2 x}$ does not admit a shift in θ to absorb x, yielding an absorption witnwess depending on $H_{1,2}$. This result is formulated in the same fashion as the ones in [36], extending their applicability to quantum machine learning.

Finally, note that it is in principle possible to construct pathological datasets such that the base PQC suffers from BP while the QRU model does not. In these cases, the data uploading gates need a careful design to cancel out the structures responsible for vanishing gradients. It is thus reasonable to assume that real-world datasets would not result in such behavior.

3.4 Gradients in layered QRU models

We consider in this section layered QRU models, in contrast to the results we presented earlier that apply to all QRU structures. In many practical scenarios there are several parameterized gates between each pair of encoding gates [38, 39]. An encoding gate and all preceding parameterized gates is referred to as a layer as

$$U(\boldsymbol{\theta}, x) = \prod_{l=1}^{L} V_l(x) u_l(\boldsymbol{\theta}_l). \tag{17}$$

In this representation, the parameterized gates $u(\boldsymbol{\theta}_l)$ are no longer defined by a single generator, and $\boldsymbol{\theta}_l$ is no longer one-dimensional. We can in this case study the absorption capability of each individual layer by defining the corresponding absorption witnesses as follows.

Definition 3.2 (Layerwise absorption witness)**.** *Let $u(\boldsymbol{\theta}_l)$ be the l-th layer of a re-uploading model from Equation* (2)*, and let $V(x)$ be the data-encoding operation applied immediately after $u(\boldsymbol{\theta}_l)$. The absorption witness for the l-th layer is*

$$\mathcal{A}_l^{(t)} = \mathbb{E}_X \left(\left\| \mathbb{E}_{\boldsymbol{\Theta}_l} \left(V_l(x)^{\otimes 2} u(\boldsymbol{\theta}_l)^{\otimes 2} - u_l(\boldsymbol{\theta}_l)^{\otimes 2} \right) \right\|_1 \right). \tag{18}$$

We provide some examples where $\mathcal{A}_l^{(t)} = 0$. First, assume a data-encoding layer sharing the generator with the corresponding parameterized gates. In this case, we can read data-encoding as a simple shift of parameters $\boldsymbol{\theta}^* \to \boldsymbol{\theta} - x$ (recall that $\boldsymbol{\theta}$ is now multi-dimensional), and averages do not change as long as $\boldsymbol{\theta}$ is sampled uniformly. Another example is the case where the ansatz is composed by k-local 2-designs located in consecutively alternated qubits, as in [47], where any k-local data-encoding gates can be re-absorbed by definition.

The use of layered ansatzes and layerwise absorption witnesses allows for further simplifications of Theorem 3.1 by bounding the complete absorption witnesses.

Lemma 3.1. *Consider a layered re-uploading model as in Equation* (17)*. Then*

$$\mathcal{B}_{R,l+1}^{(2)}(\rho_0) \leq \mathcal{B}_{R,l}^{(2)}(\rho_0) + \|\rho_0\|_\infty^2 \mathcal{A}_{l+1}^{(2)} \tag{19}$$

$$\mathcal{B}_{L,l}^{(2)}(H) \leq \mathcal{B}_{L,l+1}^{(2)}(\rho_0) + \|H\|_\infty^2 \mathcal{A}_l^{(2)}. \tag{20}$$

The proof can be found in Appendix A.2.

Consider now a layered circuit where $u_l(\cdot) = u_k(\cdot)$ for any pair (l, k). The previous result can be further simplified since $\mathcal{A}_l^{(2)} = \mathcal{A}^{(2)}$ for all values of l. Therefore

Corollary 3.1. *For a layered re-uploading model, the absorption witnesses of large parts of the circuits can be bounded by absorption witnesses of small pieces, by*

$$\mathcal{B}_{R,l}^{(2)}(\rho_0) \leq L \|\rho_0\|_\infty^2 \mathcal{A}^{(2)}, \tag{21}$$

$$\mathcal{B}_{L,l}^{(2)}(\rho_0) \leq L \|H\|_\infty^2 \mathcal{A}^{(2)}. \tag{22}$$

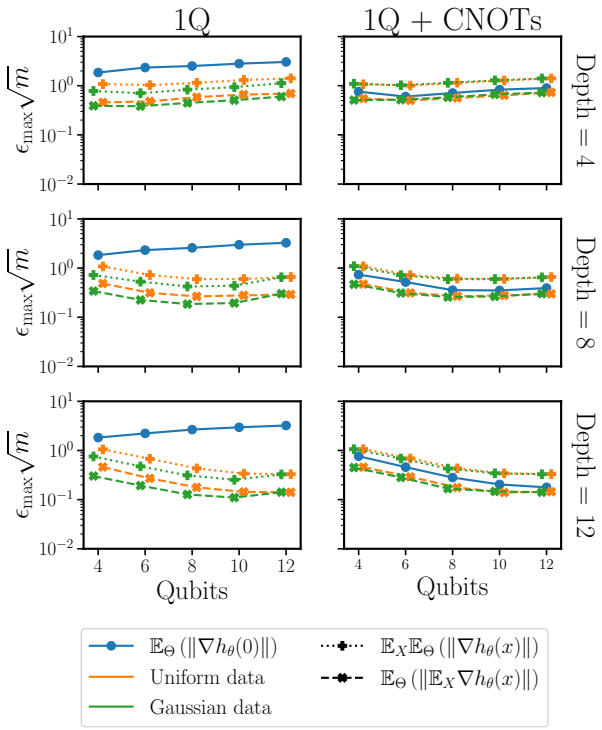


Figure 2: Results for $\epsilon_{\text{MAX}}\sqrt{m} \approx \mathbb{E}_X\left(\mathbb{E}_{\boldsymbol{\Theta}}\left(\|\nabla h_{\boldsymbol{\theta}}(x)\|\right)\right)$ (see Equation (24)) for QRU models with alternating layered ansatzes. Data is introduced through controlled-rotation gates. Parameterized gates are single-qubit arbitrary operations. Each row has an increasing depth in the circuit. The right column includes CNOT gates for the base PQC. In the left column, gradients follow different trends for the cases with and without data, implying data can not be re-absorbed into a reparametrization, in the sense of Theorem 3.1. In the right column, similar trends indicate large absorption capabilities.

The proof follows by repeated application of Lemma 3.1, together with the observation $\mathcal{B}_{R,l}^{(2)}(\rho_0) = 0$ if no data-encoding layer is considered in the absorption witness. We can therefore give a simplified bound for the results from Equation (16) in the case of layered ansatz as

$$\left| \mathbb{E}_X \left(\mathrm{Var}_{\boldsymbol{\Theta}} \left(\partial_j h_{\boldsymbol{\theta}}(x) \right) \right) - \mathrm{Var}_{\boldsymbol{\Theta}} \left(\partial_j h_{\boldsymbol{\theta}}(0) \right) \right| \leq 8L \|V_j\|_\infty^2 \|H\|_\infty^2 \|\rho_0\|_\infty^2 \mathcal{A}^{(2)}. \tag{23}$$

The result from Equation (23) is more loose than Theorem 3.1, but easier to compute, since it depends only on the layerwise absorption witness $\mathcal{A}^{(2)}$ corresponding to shallow circuits.

3.5 Numerical results

In this section, we present our numerical results, focusing on the average gradient magnitudes of hypothesis functions generated by QRU models in comparison to base PQCs. This analysis serves to validate

the findings presented in Theorem 3.1 regarding gradient variances and can be considered as a proxy for evaluating the absorption witnesses defined in Definition 3.1. We explore various ansatzes and use different data distributions for the experiments. Our code for these experiments is available in [49], and the data can be provided upon request.

For the numerical results we need to compute the magnitudes of the gradients on average. In order to reduce the computational complexity of this task, we will make use of the information content (IC) $I(\epsilon)$ [50]. The IC is a statistical measure of the variability of the optimization landscape. In a nutshell, if $I(\epsilon)$ is close to 1, then random displacements in $\boldsymbol{\theta}$ in the landscape change the value in $h_{\boldsymbol{\theta}}(x)$ in approximately ϵ, conveniently re-normalized by the norm of the displacement itself. The value ϵ_{MAX} at which $I(\epsilon)$ is maximized serves as a numerical proxy for the average norm of the gradient, that is

$$\mathbb{E}_{\boldsymbol{\Theta}}\left(\|\nabla h_{\boldsymbol{\theta}}(x)\|\right) \sim \epsilon_{\text{MAX}}\sqrt{m}, \tag{24}$$

where $\sqrt{m}$ is the number of parameters. While this approximation is not capable of computing the exact value of $\mathbb{E}_{\boldsymbol{\Theta}}\left(\|\nabla h_{\boldsymbol{\theta}}(x)\|\right)$, it is robust against statistical fluctuations and provides reliable scalings. We refer the interested reader to Ref. [50] for an in-detail explaination of the validity and utility of IC to estimate gradients.

As a first example, we compare a re-uploading ansatz, consisting of single-qubit rotations and data-encoding entangling gates, with two different PQCs (see Figure 1). In both cases, we construct the hypothesis function measuring sums of single-qubit X Pauli measurements. The addition of entangling gates in PQC2 with respect to PQC1 is essential to exploring entangled states, and it plays a role in addressing the issue of vanishing gradients [47]. The data-encoding layer is a controlled operation $C-(XR_z(x))$, which can be absorbed into single-qubit and controlled rotations [48].

The results are shown in Figure 2. The columns correspond to the respective models (1-2), and the rows correspond to different depths of the ansatz. In the left column, corresponding to a PQC with only single-qubit gates, we observe a qualitative difference in the average gradients of the PQC and re-uploading circuits. The PQC is relatively insensitive to the number of qubits, as a consequence of the redundancy of multiple consecutive single-qubit gates. Adding data modifies this behavior. In the case of the entangling PQC (right column), the results without data align with previous works [47, 50]. The addition of data drawn from different distributions (Gaussian and uniform) introduces negligible differences in the results.

We turn our attention to translation-invariant ansatzes. These circuits are not capable of freely exploring the Hilbert space, but only its invariant subspace. This restriction reduces the freedom in

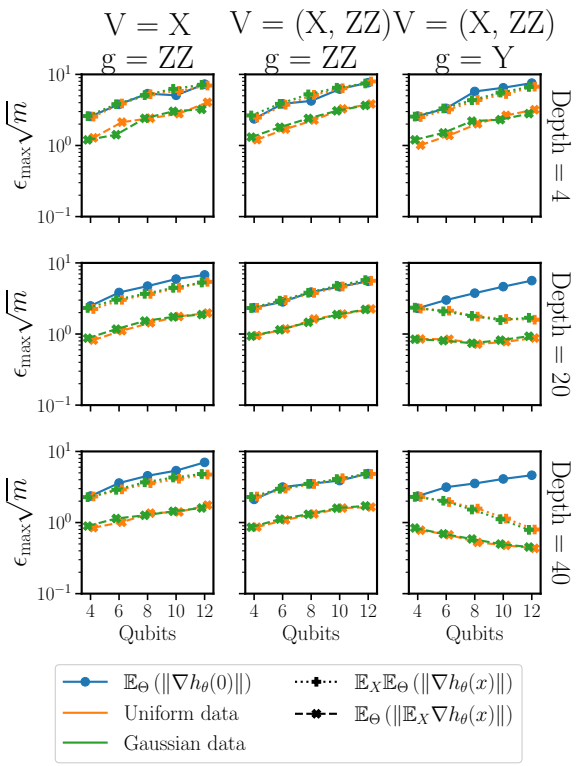


Figure 3: Results for $\epsilon_{\text{MAX}}\sqrt{m} \approx \mathbb{E}_X\left(\mathbb{E}_{\boldsymbol{\Theta}}\left(\|\nabla h_{\boldsymbol{\theta}}(x)\|\right)\right)$ (see Equation (24)) for QRU models with translation-invariant layered ansatzes. Data is introduced through the generators g, and data through the generators V, indicated at the top for every column. Each row has an increasing depth in the circuit. In the left column, gradients follow approximately the same trends with and without data, implying high absorption capabilities in the sense of Theorem 3.1. For the middle column, the absorption is total since it can be done by a simple shift of parameters. The right column reveals different trends for the cases with and without data, implying low absorption.

these circuits, leading to an increase in the average gradients of the cost function for PQCs [39]. We choose three layered models, based on the generators $X = \sum_q X_q$, $Y = \sum_q Y_q$, $ZZ = \sum_q Z_q Z_{q+1}$, where q cyclically iterates over all qubits. In the first model, the generator associated with parameters is $V_i = X$, and data-encoding is conducted through $g = ZZ$. The second model is given by $\{V\} = \{X, ZZ\}$, $g = ZZ$. The third model is defined by $\{V\} = \{X, ZZ\}$, $g = Y$. In all cases, the observable considered is X. Among these models, only the second one can automatically absorb data into parameters through shifts. Gaussian-distributed data is used in all cases.

Results are detailed in Figure 3. The columns correspond to the respective models, and the rows correspond to different circuit depths. For each model, the average norm of the gradient scales differently with the number of qubits, with and without data. Models 1 and 2 present similar behavior when including data. In particular, for model 2 results show no differ-

ence between the re-uploading model and PQC since the data can be perfectly re-absorbed through a simple shift. A significant difference is noticeable in the third model. In this case, the absence of BPs in all instances makes the QRU models trainable by construction.

4 Expressivity in QRU models

The hypothesis class of QRU can always be expressed as a generalized trigonometric polynomial [40], see Equation (4). In QRU models, the set of frequencies Ω is generated through the sequential Minkowski sum of the spectrum of the data encoding generators $\{\lambda_j\}_j$. In the general case, $\{\lambda_j\}_j$ consists of incommensurable real numbers, i. e. with non-rational ratios, and each new encoding step makes Ω combinatorially denser. In this section we first consider harmonic generators, i. e. with integer eigenvalues, and extend the results later to generic generators. As a main observation of this work, the behavior is similar in both cases.

4.1 Harmonic representation of quantum states

In this section, we introduce a representation of QRU models based on the Fourier decomposition of the hypothesis function. Such representation is useful for subsequent analytical results. Starting from Equation (2) and assuming the generators g_i possess an integer spectrum, we can express the state before measurement as

$$U(\boldsymbol{\theta}, x)\,|0\rangle = \sum_{k=-K}^{K} \sum_{j=1}^{2^n} c_{j,k}(\boldsymbol{\theta}) e^{i\mu k x}\,|j\rangle\,. \tag{25}$$

The coefficients $c_{j,k}$ form a matrix $\boldsymbol{C} \in \mathbb{C}^{2^n \times (2K+1)}$ that defines uniquely (up to a global phase) the output state of the re-uploading circuit before measurement. The matrix $\boldsymbol{C}$ depends only on the parameters $\boldsymbol{\theta}$ and the generators of the ansatz, but not on the data x. The value K corresponds to the largest attainable frequency, namely the sum of the largest eigenvalue for each generator used in the circuit. The recipe to construct $\boldsymbol{C}$ from the description of the circuit is detailed in Appendix A.3. Notice this approach is equivalent to adding an extra dimension (frequency)to the standard brute-force state vector simulation, which is not efficient from a computational point of view. This harmonic representation of QRU models simulator is available on [49].

4.2 Vanishing high frequencies in QRU models

We use the above representations and the intuition that adding a data encoding layer corresponds to a convolution operation with the data encoding generator spectrum as defined below in 4.1. For proof purposes, we assume that Haar random matrices are interleaved in between reuploading layers, as is common in most papers exploring barren plateaus. We examine the statistical properties of the amplitude of the coefficients as a function of the frequency. We begin by defining the spectrum kernel of a harmonic Hermitian matrix.

Definition 4.1 (Harmonic spectrum kernel)**.** *Let H be a $N \times N$ Hermitian matrix with integer eigenvalues $\{\lambda\}$ with multiplicities $m(\lambda)$. The spectrum kernel of H is the vector (indexed by k)*

$$\mathcal{K}_H(k) = \begin{cases} m(k\mu)/N & \text{if } k\mu \in \{\lambda\} \\ 0 & \text{Otherwise} \end{cases}, \tag{26}$$

where μ is the largest value in $\mathbb{R}$ compatible with this description.

This function simply maps the eigenvalues of a Hermitian matrix into the normalized dimensionality of the corresponding eigenspace. For readability, we will refer to the *spectrum multiplicity function* simply as the *spectrum* for the remainder of the paper.

In the case of layered QRU models, the spectra of their data-encoding generators and the number of layers L directly determine the set of attainable frequencies. The maximum attainable frequency is bounded by $L\|g\|_2$. We provide now some insight into how the coefficients are expected to behave.

Lemma 4.1 (Harmonic convolution)**.** *Let $|\psi_{\boldsymbol{\theta}}(x)\rangle = U(\boldsymbol{\theta}, x)\,|\psi_0\rangle$ be the output state of a re-uploading model, with data-encoded through the generator set $\{g_j\}$, each with spectrum $\mathcal{K}_{g_j}$, encoded as in Equation (25). Assuming that each parameterized step is drawn from the Haar measure of unitaries, then $\sum_j |c_{j,k}|^2$ is a random variable satisfying*

$$\sum_j |c_{j,k}|^2 \sim \operatorname{Dir}\left(\left(\mathcal{K}_{g_1} * \ldots * \mathcal{K}_{g_j} * \ldots * \mathcal{K}_{g_L}\right)(k)\right), \tag{27}$$

where $\operatorname{Dir}(\alpha_1, \alpha_2, \ldots)$ is the Dirichlet distribution [51] and $*$ denotes the convolution.

The proof can be found in Appendix A.4. The Dirichlet distribution is a family of probability distributions for multidimensional variables $\boldsymbol{x} \in [0,1]^N$, subject to $\|\boldsymbol{x}\|_1 = 1$. The Dirichlet distribution over N variables is fully described by N parameters $\alpha_i \in \mathbb{R}_{>0}$. Dirichlet is the multidimensional extension of the beta distribution. A detailed definition of the Dirichlet distribution and auxiliary results are given in Appendix A.5. For completeness, we define convolution as

$$(f * g)(k) = \sum_{l=-\infty}^{\infty} f[k]g[l-k]. \tag{28}$$

In other words, this lemma gives statistical properties of the frequency content, expressed as the norm of the 2^n quantum vector corresponding to each frequency (seee Equation (25)) for QRU models composed of a sequence of data uploading gates interleaved with gates drawn from the Haar distribution. It states that the vector of frequency content follows a multidimensional distribution whose mean is the result of the successive convolution of the multiplicity kernels of the data encoding gates. It follows a Dirichlet distribution because all values are positive and sum up to one as per the normalization of a quantum state.

It is worth discussing the role of the Haar distribution in this result. First, choosing random unitaries allows us to scramble the inner quantum state in the QRU model at each step, thus transforming the QRU circuit into a random walk in the space of frequencies, where the parameters in Dirichlet only account for the number of paths leading to the same outputs. Second, random choices of unitaries is in alignment with other works exploring trainability and expressivity in VQAs [34, 47, 36], which rely on sampling unitaries from a t-design. The difference between the Haar distribution and a t-design is rather technical, since t-designs are sets of unitaries with the same statistical moments as the Haar measure, up to degree t [29]. With respect to Lemma 4.1, lowering the requirements in the parameterized steps from Haar distribution to t-design would imply to substitute the Dirichlet distribution with another probability distribution with the same t-statistical moments. Technical descriptions of this transformations are left as open questions for future research. Note that our results lose their validity if the parameterized steps are drawn with respect to other distributions of unitaries.

The previous result immediately implies the following.

Theorem 4.1 (Single-generator convolution)**.** *Let $|\psi_{\boldsymbol{\theta}}(x)\rangle = U(\boldsymbol{\theta}, x)\,|\psi_0\rangle$ be the output state of a re-uploading model, with data-encoded through the generator g, with spectrum $\mathcal{K}_g$, encoded as in Equation* (25)*. Assuming that each parameterized step is drawn from the Haar measure of unitaries, then*

$$\sum_j |c_{j,k}|^2 = \mathrm{Dir}\left(\left(\mathcal{K}_g^{*L}\right)(k)\right), \tag{29}$$

where $(\cdot)^{*L}$ denotes the L-fold convolution.

The proof is immediate from extending Lemma 4.1.

We provide two explicit examples to distinguish the cases captured by Lemma 4.1 and Theorem 4.1. Consider the single-qubit generator $g = (Z_0 + I)/2$, with spectrum $\mathcal{K}_g = (0, 1)$. To illustrate Lemma 4.1, we choose the list of generators as $\{2^l g\}_{l=0}^{L}$, yielding a convolution

$$\left(\mathcal{K}_{g_1} * \ldots * \mathcal{K}_{g_j} * \ldots * \mathcal{K}_{g_L}\right)(k) = 1, \quad \forall k \in \{0, \ldots, 2^L - 1\}. \tag{30}$$

On the other hand, illustrating Theorem 4.1 we consider a repeated application of g, yielding

$$\left(\mathcal{K}_g^{*L}\right)(k) = \binom{k}{L}, \forall k \in \{0, \ldots, L\}. \tag{31}$$

The behavior in the two cases of the random variable $\sum_j |c_{j,k}|^2$ is significantly different. In the first case, the output is a flat distribution of exponential size. On the contrary, the second case is a distribution of linear size with high concentration in its mean values.

Notice that, provided that the data generator is known, it is possible to classically store $\mathcal{K}_g^{*L}$ within memory of size $\mathcal{O}\left(L\|g\|_2/\mu\right)$, with computational cost $\mathcal{O}\left((L\|g\|_2/\mu)^3\right)$. This allows us to classically characterize the frequency profile prior to executing the QRU model in quantum hardware for harmonic generators with only polynomially many eigenvalues.

The previous theorem can be readily interpreted in the limit of large L by virtue of the central limit theorem [52]. The repeated convolution of any random variable with a variance of σ and a probability distribution in the spaces L^1 and L^2, tends to a normal distribution in a weak sense. We can thus obtain the following result.

Corollary 4.1 (Vanishing high frequencies)**.** *In the conditions of Theorem 4.1 and for large number of re-uploading L,*

$$\lim_{L\to\infty} \sum_j |c_{j,k}|^2 \sim \mathrm{Dir}\left(\mathcal{N}\left(0, \sigma_g^2 L\right)(k)\right), \tag{32}$$

where σ_g is the standard deviation of the spectrum $\mathcal{K}_g$, and $\mathcal{N}(\mu, \sigma^2)$ is the normal distribution.

This observation implies that tails of the distribution vanish exponentially for large frequencies and there is a concentration in the low-frequency terms as the magnitudes of high-frequency terms vanish. In asymptotic scaling the available spectrum reduces from $\|g\|_\lambda^2 L$ to $\sigma_g\sqrt{L}$. For interpretability, recall the example $g = (Z_0 + I)/2$, yielding a binomial distribution in the convolution of the subsequent spectra. The binomial distribution rapidly tends to a gaussian distribution.

The results from Theorem 4.1 and Corollary 4.1 can be extended to non-harmonic generators. For readability, we postpone this result until Section 4.4.

The previous discussion considers the effects of the spectrum on the internal state of the re-uploading model in its harmonic representation. We are however not interested in the state itself, but rather in $h_{\boldsymbol{\theta}}(x)$ measured as an expectation value of this internal state. The Fourier components $h_{\boldsymbol{\theta}}(x)$ satisfy the following corollary.

Corollary 4.2. *Let $|\psi_{\boldsymbol{\theta}}(x)\rangle$ be the output state of a re-uploading model, with a single data-encoding generator g with spectrum $\mathcal{K}_g$. Let $h_{\boldsymbol{\theta}}(x)$ be the hypothesis function induced by the observable H in the re-uploading model, as in Equation* (3)*, and let $a_k(\boldsymbol{\theta})$ be*

their corresponding Fourier coefficients as in Equation (4)*. In the conditions of Theorem 4.1, and for symmetric spectra* $\mathcal{K}_g(k) = \mathcal{K}_g(-k)$*,*

$$\|H\|_\lambda^{-2} |a_k(\boldsymbol{\theta})|^2 \leq p_k \tag{33}$$

$$p_k \sim \mathrm{Dir}(\mathcal{K}_g^{*2L}(k)), \tag{34}$$

where p_k *is a multidimensional probability distribution sampled from the Dirichlet distribution defined by the L-fold convoluted spectrum* $\mathrm{Dir}(\mathcal{K}_g^{*2L}(k))$ *[51].*

Additionally, this result extends to the Gaussian distribution in the limit of large L as for Corollary 4.1.

The results from this section show that the frequency terms of $h_\theta(x)$ tend to follow a Gaussian profile of width $\sim L$, in the assumption that the generator of data-encoding gates is repeated in the QRU model. However, the frequency support of these functions scales linearly in L. As an immediate consequence, only frequencies $\omega \in \mathcal{O}(\sqrt{L})$ have practical support on average, while larger frequencies have exponentially vanishing weight in the hypothesis function. Note, that the Gaussian profile described by Corollary 4.1 does not imply a dense frequency space, which is still restricted to integer frequencies. This result holds even in the case where the generator provides exponentially-in-qubits many frequencies. It is then possible to have exponentially large frequency sets even with a small number of re-uploading steps, and the Gaussian approximation still holds with $\sigma \sim \sqrt{L}e^n$.

4.3 Lipschtitz expressivity

In this section, we delve into a more practical understanding of the expressivity of hypothesis functions in terms of the magnitude of their derivatives. The ability to capture fine-grained data patterns depends on the function's ability to access high rates of change, i.e., the magnitude of its derivative. This concept can be quantified through the maximum value of the derivative, known as the optimal Lipschitz constant. For a function f, the optimal Lipschitz constant is defined as

$$\mathcal{L}(f) = \max_x |\partial_x f(x)| . \tag{35}$$

The Lipschitz constant is closely related to Fourier analysis, as high derivatives can only be achieved if the Fourier spectrum includes high frequencies with significant coefficients. Specifically,

$$\mathcal{L}(f) \leq \sum_{k=-K}^{K} |k|\mu |a_k e^{ik\mu x}|, \tag{36}$$

where a_k represents the Fourier coefficients of the hypothesis function.

We introduce an upper bound to the Lipschitz constant inspired by Equation (35), adapted for QRU models and properly normalized with respect to the measured observable as

$$\Lambda(h_{\boldsymbol{\theta}}) = \sum_{k=-K}^{K} \mu|k||a_k|. \tag{37}$$

It is straightforward to see that $\Lambda(h_{\boldsymbol{\theta}}) \geq \mathcal{L}(h_{\boldsymbol{\theta}})$, and therefore we are going to use this quantity as a proxy for it. For readability, this optimal Lipschitz constant upper bound will be referred to LB in the subsequent sections of this paper and be noted $\Lambda(h_{\boldsymbol{\theta}})$ unless otherwise specified.

Using results from previous sections we study $\Lambda(h_{\boldsymbol{\theta}})$, starting with a result giving tight bounds on the asymptotic average of the LB over the parameters $\boldsymbol{\Theta}$. The results stated in the next and subsequent propositions stem from the conditions discussed in Section 4.2, namely tunable gates are drawn from the Haar distribution.

Theorem 4.2 (Average LB)**.** *Let* $h_{\boldsymbol{\theta}}(x)$ *be the hypothesis function of a re-uploading model for which Theorem 4.1 applies. Let* $\Lambda(h_{\boldsymbol{\theta}})$ *be the LB as defined in Equation* (37)*. Then,*

$$\|H\|_\lambda \sqrt{2L}\mu\sigma_g \leq \lim_{L\to\infty} \mathbb{E}_{\boldsymbol{\Theta}} \left(\Lambda(h_{\boldsymbol{\theta}})\right) \leq \|H\|_\lambda \frac{4}{\sqrt{\pi}} \sqrt{L}\mu\sigma_g \tag{38}$$

The proof can be found in Appendix A.7. Notice the tightness of the bounds above since $2\sqrt{2}/\sqrt{\pi} \approx 1.6$.

The following subsection quantifies the likelihood of the LB different from the average. Notice that values smaller than the average are not relevant due to the definition of $\Lambda(h_{\boldsymbol{\theta}})$. We can leverage the insights from previous results, particularly the role of Dirichlet distributions, to derive the following result:

Theorem 4.3 (Deviation of LB)**.** *Let* $h_{\boldsymbol{\theta}}(x)$ *be the hypothesis function of a re-uploading model for which Theorem 4.1 applies, with data-encoding generator* g*. Let* $\Lambda(h_{\boldsymbol{\theta}})$ *be its LB as defined in Equation* (37)*. Then,*

$$\lim_{L\to\infty} \mathrm{Prob}\left(\Lambda(h_{\boldsymbol{\theta}}) - \|H\|_\lambda \sqrt{2L}\sigma_g\mu \geq t\right) \in \mathcal{O}\left(\exp\left(-\frac{t^2}{\mathrm{poly}(L\mu)}\right)\right). \tag{39}$$

The proof can be found in Appendix A.8. Notice this result automatically bounds the probability of the optimal Lipschitz constant itself of being bigger than $\sqrt{2L\mu}\sigma_g$.

The previous theorem can be further refined to provide a tighter bound on the likelihood of large deviations from the LB. As mentioned in the detailed proof, when Theorem 4.1 holds the weight of each frequency and tends to follow a Gaussian-like profile, with central frequencies having exponentially larger probabilities than the extremal ones. It is expected that the

primary contributions to $\Lambda(h_{\boldsymbol{\theta}})$ come from these central frequencies, which also have the smallest prefactors. Taking this into account, we can update the results from Theorem 4.3 to provide a more precise bound,

$$\lim_{L\to\infty} \text{Prob}\left(\Lambda(h_{\boldsymbol{\theta}}) - \|H\|_\lambda \sqrt{2L}\mu\sigma_g \geq t\right) \in \mathcal{O}\left(\exp\left(-\frac{t^2}{(\sigma_g \mu\sqrt{L})^3}\right)\right). \quad (40)$$

The vanishing high frequencies from Section 4.2 have consequences on the properties of the attainable hypothesis functions. In particular, its maximal derivative with respect x, given by the Lipschitz constant, scales in average with $\sqrt{L}$, and the probability of finding larger Lipschitz constants vanishes super-exponentially fast. This imposes in practice constraints on the capability of the hypothesis functions to capture fine details in the data, effectively restricting target functions that can be approximated by QRU models.

4.4 Extension to generic data generators

In previous subsections, we have proven the phenomenon of vanishing high frequencies and its consequences on the Lipshitz constant for harmonic data generators. In this subsection, we extend the results from Section 4.2 and 4.3 to any data generator. We start by defining the spectrum kernel for generic Hermitian matrices.

Definition 4.2 (Hermitian spectrum kernel)**.** *Let H be a $N \times N$ Hermitian matrix with integer (positive or negative) eigenvalues $\{\lambda\}$ with multiplicities $m(\lambda)$. We define the vector $\vec{\mu} \in \mathbb{R}^D$, with $\mu_i/\mu_j \in \mathbb{R}\backslash\mathbb{Q}\ \forall(i,j)$, such that any eigenvalue can be written as $\lambda = \vec{\mu}\cdot\vec{k}$, with $\vec{k} \in \mathbb{Z}^D$. We refer to the number of anharmonic dimensions as $D \leq 2^n$, where n is the number of qubits. We define the spectrum kernel of H as $\mathcal{K}_H$ such that*

$$\mathcal{K}_H(\vec{k}) = \begin{cases} m(\lambda)/N & \text{if } \vec{k}\cdot\vec{\mu} \in \{\lambda\} \\ 0 & \text{Otherwise} \end{cases} \quad (41)$$

Where each μ_j is the largest value in $\mathbb{R}$ compatible with this description.

We note the covariance of this spectrum as the $D \times D$ matrix Σ_g. The average of this spectrum is 0 since we consider traceless generators. The results from Lemma 4.1 and Theorem 4.1 hold in the non-harmonic case. In this scenario, the convolution must be done in a D-dimensional space, leading to D-dimensional frequency profiles. The convoluted spectrum, for the single-generator case, can be stored in a memory structure of size $\mathcal{O}\left((L\|g\|_2/\min_j\mu_j)^D\right)$. Notice that for each eigenvalue λ there exists only one compatible $\vec{k}$, due to the irrational ratios between elements in $\vec{\mu}$.

The central limit theorem still applies in the non-harmonic case as well, leading to the following result.

Corollary 4.3 (Vanishing high frequencies)**.** *Given the conditions of Theorem 4.1 for non-harmonic generators and for large number of re-uploadings L,*

$$\lim_{L\to\infty} \sum_j \left|c_{j,\vec{k}}\right|^2 \sim \text{Dir}\left(\mathcal{N}\left(0, \Sigma_g L\right)\left(\vec{k}\right)\right). \quad (42)$$

Following the reasoning from the harmonic case, we focus now on the Lipschitz constant of the hypothesis functions. The definition from Equation (37) can be extended to

$$\Lambda(h_{\boldsymbol{\theta}}) = \sum_{\omega\in\Omega} |\omega||a_\omega|, \quad (43)$$

with $\omega = \vec{\mu}\cdot\vec{k}$. Since $\vec{k}$ has integer values and $\vec{\mu}$ has irrational ratios among its elements, there is at most one solution of $\vec{k}$ for each ω. With this definition, we can formulate results analogous to Theorem 4.2 and Theorem 4.3.

Corollary 4.4 (Lipschitz bounds for non-harmonic generators)**.** *Let $h_{\boldsymbol{\theta}}(x)$ be the hypothesis function of a re-uploading model for which Corollary 4.3 applies. Let $\Lambda(h_{\boldsymbol{\theta}})$ be the LB as defined in Equation* (43)*. Then,*

$$\lim_{L\to\infty} \mathbb{E}_{\boldsymbol{\Theta}}\left(\Lambda(h_{\boldsymbol{\theta}})\right) \leq \frac{4}{\sqrt{\pi}} \|H\|_\lambda \sqrt{\text{Tr}(\Sigma)}\ \|\vec{\mu}\|_2 \sqrt{L} \quad (44)$$

$$\lim_{L\to\infty} \mathbb{E}_{\boldsymbol{\Theta}}\left(\Lambda(h_{\boldsymbol{\theta}})\right) \geq \sqrt{2}\ \|H\|_\lambda \sqrt{\min_\lambda(\Sigma)}\, \|\vec{\mu}\|_2 \sqrt{L}. \quad (45)$$

The proof of Corollary 4.4 can be found in Appendix A.9. In addition, following the same reasoning leading to Theorem 4.3, we can infer exponential concentrations of $\Lambda(h_{\boldsymbol{\theta}})$ around its average values, by

$$\lim_{L\to\infty} \text{Prob}\left(\Lambda(h_{\boldsymbol{\theta}}) - \|H\|_\lambda \sqrt{2L}\|\vec{\mu}\|_2 \sqrt{\min_\lambda(\Sigma)} \geq t\right) \in \mathcal{O}\left(\exp\left(-\frac{t^2}{\left(\sqrt{\max_\lambda(\Sigma_g)}\max_j(\mu_j)\sqrt{L}\right)^3}\right)\right). \quad (46)$$

In light of the previous theorem, we can observe that the vanishing high frequencies phenomenon extends to non-harmonic generators, with minor changes with respect to the harmonic case, rooting from norm bounds in the multi-dimensional space. The tightness of these bounds depends on the regularity of the anharmonic spaces, which is reflected into the values of $\vec{\mu}$ and the eigenvalues of Σ.

An immediate consequence of this section is that QRU models can have a dense frequency spectrum without significantly modifying the envelope of the

frequency profile. The only elements of QRU models allowing to increase the set of available frequencies in practice are L and the spectrum profile Σ, while $\vec{\mu}$, which can be related to $\|g\|_\lambda$ have a more modest effect. It is possible to reach exponentially many different frequencies by using generators with exponentially large $\mathcal{K}_g$.

The number of different frequencies directly affects the surrogability of the studied QML models In the case the frequency space is polynomial in the number of qubits, it is possible to construct a classical model fitting the corresponding generalized trigonometric polynomial [53]. On the other hand, exponentially large frequency spaces do not admit arbitrary efficient classical representations. The findings detailed in this work provide methods to circumvent surrogability. This can come from generators with exponentially large spectra, or designed in such a way that the frequency space scales exponentially with L, for instance with convolutions of highly non-harmonic spectra.

4.5 Numerical results

In this section, we show the results of a series of numerical experiments in which such conditions are relaxed and show that the theoretical results still apply. We use three models to test different situations. The first two models are constructed with permutation-invariant generators, which correspond to PQC that have been proven to be trainable [46]. Those models express only the symmetric subspace in the available Hilbert space. In the first model (A), $g = X, V = ZZ$, and the second model (B) $g = X, V = \{Y, X, ZZ\}$. For the third model, $g = X$, and the parameterized pieces are sampled from the Haar measure, that is the set V is free. We choose these models to have full control of the spectrum of the generator $\mathcal{K}_{g=X}$, which allows us to informatively compare to the theorems. All experiments were conducted with systems of 4 qubits unless explicitly stated, without affecting the scaling of the obtained results.

The first experiment tests Theorem 4.1 and Corollary 4.1, and the results can be seen in Figure 4. In model (A), the spectrum spreads towards large frequencies with the number of data re-uploadings. The parameterized gates are not general enough to support the theoretical results. For models (B) and (C), the Gaussian limit is matched even for a moderately small number of layers. This means that even though the theorem is proven for Haar random unitaries, the vanishing high-frequencies behavior still holds for model (B), even though the Haar condition is not guaranteed. Notice the difference in spreads for models (B) and (C). This is a consequence of the space explored by the ansatz. Model (B) is composed of a permutation invariant ansatz, and it is as general as possible only in the symmetric subspace, of dimension $n+1$. In this scenario, the spectrum of the corresponding g is flat (see the results for 1 layer in figure 4), and the spread depends on the number of qubits n as $\sigma_g = \mathcal{O}(n)$. For the model (C), the spectrum of g with no restriction follows a binomial distribution, centered in $k = 0$, with $\sigma_g \in \mathcal{O}(\sqrt{n})$. A comparison between the theoretical and observed variances can be found in Figure 5, showing high agreement with the theoretical results.

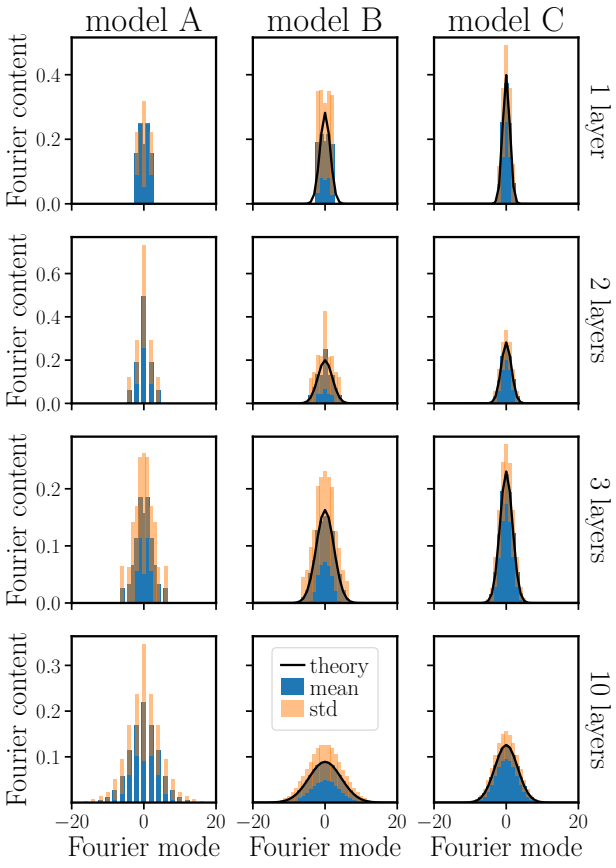


Figure 4: Evolution of frequency spectrum with the number of layers, for models (A, B, C) detailed in the first paragraph of the section. The Fourier content refers to (average and standard deviation of) $|c_k|^2$. Model (A) is not general enough to follow Theorem 4.1, but still, a spread in frequencies is observed. Model (B) is permutation-invariant and almost fully general in the symmetric space, and model (C) is general with no restrictions in the Hilbert space. As L increases, the Fourier spectrum approximates a Gaussian profile with increasing variance according Equation (32). The values σ_g for models (B) and (C) change due to the constraint in the available space.

We numerically check Theorem 4.3 in Figure 6. We depict in this figure the observed cumulative distribution functions (CDF) of both the numerically found LB, and $\Lambda(h_{\boldsymbol{\theta}})$ defined in Equation (35). These CDFs are compared to the upper bound from Theorem 4.3.

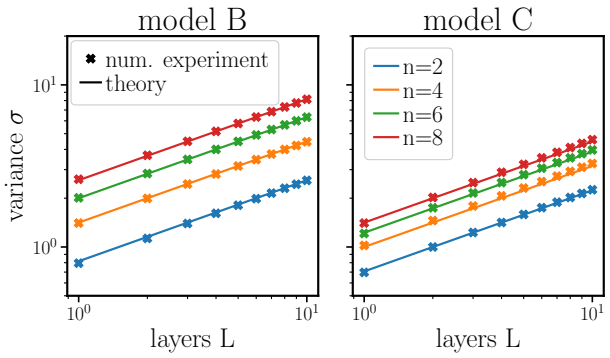


Figure 5: Variances for the tending-to-Gaussian profiles from Figure 4, for different numbers of qubits, in the y-axis, while the x-axis corresponds to the number of layers. The left figure corresponds to model (B), with $\sigma_g^2 = n(n+2)/12$. The right figure corresponds to model (C), with $\sigma_g^2 = n/4$. The scaling in $\sqrt{L}$ is in agreement with Equation (32).

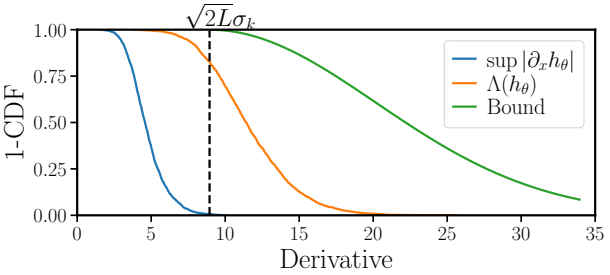


Figure 6: Inverse CDFs for the optimal Lipschitz constants and $\Lambda(h_{\boldsymbol{\theta}})$, as compared to the bounds from Theorem 4.3. The x-axis indicate the value for each CDF, respectively $\sup_x |\partial_x h_{\boldsymbol{\theta}}(x)|$, $\Lambda(h_{\boldsymbol{\theta}})$ and t for each line.

Results show agreement with Theorem 4.3, and even indicate the possibility of finding tighter bounds, at least in terms of prefactors.

4.5.1 Training

All the LB results describe an average behavior for $\boldsymbol{\Theta}$. In this subsection, we briefly explore the effect of previous results in the training. We task model (B) to learn functions whose Fourier coefficients follow a step function of increasing width. This approximately corresponds to a cardinal sinus of decreasing width.

We display results of trained QRU models in Figure 7. In the top figure, we show the Fourier components of different functions to be fitted (in dashed lines), and the hypothesis functions after training (solid lines). The target function is learned by the model for $K \leq 20$ but the hypothesis function fails to capture high frequencies from $k > 25$. Notice that the obtained hypothesis functions for $K = \{30, 40\}$ seem to saturate the expressivity capabilities of the model. The bottom figure represents the functions in the data domain for $K = \{4, 40\}$.

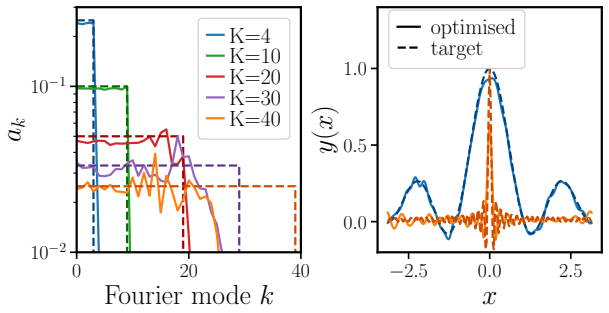


Figure 7: Time and frequency domains representations of the functions of circuits (B) trained to match increasingly sharp cardinal sinus that yields increasing high-frequency content. From Fourier mode $k = 20$, the model is not able to match the amplitude, exhibiting a consequence of vanishing high frequencies.

5 Discussion

We turn our attention first to gradients. From our results, we can infer that vanishing gradients are avoided in QRU models if the base PQC is BP-free, and data can be absorbed into the parameterized gates. We can refer to existing literature on avoiding BPs for PQCs by restricting the dimensionality of the search space, by means of the dynamical Lie algebra [39, 46]. In a nutshell, the Lie algebra depends on the generators of the quantum model. Absorption witnesses can only be maintained close to 0 if the base PQC and the derived QRU model share a common Lie algebra. This observation allows one to choose data-encoding generators avoiding the emergence of BPs.

The average of $\Lambda(h_{\boldsymbol{\theta}})$ is a consequence of the vanishing high frequencies behavior that grows as $\sim \sigma_g \sqrt{L}$, as imposed by the central limit theorem. Deviating from this average is exponentially unlikely, as proven in theorem 4.3. As discussed later, it is in principle possible to amplify high-frequency components, at the expense of losing all degrees of freedom in the process. Therefore, for practical purposes, we need to adjust the number of re-uploading layers according to the scaling $\sim \sqrt{L}$, and not $\sim L$, as suggested by other theoretical works on expressivity via generalization bounds [40].

In this work we derived the the scaling of the Fourier spectrum of hypothesis functions with the number of layers, but not with the number of qubits n. Our numerical simulations focus on frequency spaces increasing polynomially with n. However, it is possible to construct data-generators with exponentially many equally probable different accessible frequencies [54]. In this scenario, Theorem 4.1 still holds, leading to a Gaussian profile of frequencies with variance $\sigma \in \Theta(2^n \sqrt{L})$. Note that exponentially many frequencies require exponentially many tunable parameters to match the number of degrees of freedom. Therefore, data-encoding generators with

$\sim e^n$ different frequencies can only aim to efficiently learn functions with sparse Fourier representations, i.e. with only $\mathcal{O}(\text{poly}(n))$ non-zero Fourier coefficients in a $\mathcal{O}(e^n)$ frequency space.

The expressivity results from Section 4 imply direct limitations in the attainable hypothesis functions, but also give an intuition on how to amplify high-frequency Fourier components, or in other words how to maximize the Lipschitz constant. The only recipe to obtain a high-frequency Fourier profile is by repeatedly amplifying the eigenvectors corresponding to extreme eigenvalues of the data-encoding generator. This yields an extremal case as far as possible from the average case where unitaries are sampled from the Haar distribution. Without loss of generality, we may choose the ground state. The first step of the circuit would have to transform the initial state into the ground state of the data-encoding generator. For k-local Hamiltonians with $k \geq 2$, this problem is QMA-complete [55], and finding the hypothesis function with maximum high-frequency content implies repeatedly solving a QMA-complete problem. An example is choosing the data generator to be the Hamiltonian of a transverse-field Ising model, constructed on an arbitrary graph. Such PQC does not suffer from BPs if the parameters respect the permutation invariance of the graph [39]. Therefore, reaching the hypothesis function with maximized high-frequency content is in general hard. A notable exception appears in layered circuits with one g, and $W_i = I$, where setting all parameters $\theta = 0$ suffices to maintain the quantum state aligned with the ground state of g. Notably, maximizing the Lipschitz constant in the experiments from Section 4.5 is feasible.

We have seen that hypothesis functions produced by layered QRU models have naturally vanishing high-frequency components, therefore limiting their Lipschitz constant. Regularization of the Lipschitz constant yields increased generalization and robustness of classical Neural Networks [56, 57]. As a consequence, this could hint toward a better generalization capacity of QRU models, in agreement with existing literature [58].

The scope of this work is the average behavior of QRU models. It shows concentration properties, similar to other existing results [59, 60, 34], and provides useful insights on the internal working principles of the model. This interpretability will be useful to develop QML models with specific properties. It will be possible to investigate protection against dequantization through peaked generalized trigonometric polynomials, in alignment with peaked circuits [61]. Restrictions of generators and parameterized steps in the circuits can be applied to constraint the behavior of output models, allowing for systematic exploration of ingredients in QRU.

6 Conclusions

We have explored the features of QRU models to understand the implications of injecting data into the better-studied PQCs models. Two main features are studied, first the magnitude of the gradients of the loss function, and second the frequency profile of the hypothesis function output by QRU models. Results were proven analytically and extended to more practical scenarios numerically.

We give analytical bounds for the connection between the variance of gradients in the hypothesis functions of QRU models and the cost functions on corresponding PQCs. Vanishing gradients of hypothesis functions imply vanishing gradients for any cost function to train ML models with, thus preventing trainability. The difference between QRU models and PQCs can be quantified by measuring the effect of adding data to the circuit, averaging over the parameter space. This is quantified by the coined concept of *absorption witness*. If data can be re-absorbed in a shift of parameters, then the gradients for PQCs and QRU models take similar value ranges. Results can be further simplified for the case of layered ansatzes. These results provide insights into the construction of QRU models protected against the phenomenon of BP, by using existing knowledge on PQC that do not exhibit BPs.

We prove also that QRU models suffer from vanishing high frequencies. Each additional data encoding operation corresponds to an additional convolution of the current spectrum with the data generator spectrum. As the number of layers, denoted as L, becomes large the span of attainable frequencies grows as $\mathcal{O}(L)$. However, the central limit theorem dictates that the frequency profile follows a Gaussian distribution, spreading out proportionally to $\sqrt{L}$. Therefore, in practice, only frequencies (approximately) bounded by $\sqrt{L}$ are available, with the contribution of higher frequencies exponentially vanishing. The vanishing high frequencies have direct consequences on the class of functions attainable by the QRU models. The average of the optimal Lipschitz constant scales with $\sqrt{L}$, exhibiting an exponentially decaying probability of exceeding this value. These findings offer insights into the inherent limitations of expressivity in QRU models and provide tools for estimating the computational resources required to represent specific datasets effectively.

The results derived in this work broaden our understanding of the properties of QRU models and provide guidelines for the design of re-uploading schemes. As an example, the concept of absorption witness can be employed to select generators ensuring an ansatz with the necessary characteristics to be trainable. For expressivity, adjusting the depth of the model can strike a balance between capturing intricate details in the data and avoiding overfitting.

Consequently, we anticipate that these tools and insights will contribute to enhancing the applicability and performance of QRU models.

Acknowledgments

The authors thank Elies Gil-Fuster, Vedran Dunjko, Patrick Emonts, and Xavier Bonet-Monroig for their useful comments on this manuscript. This work was supported by CERN through the CERN Quantum Technology Initiative. This work was carried out as part of the quantum computing for earth observation (QC4EO) initiative of ESA ϕ-lab, partially funded under contract 4000135723/21/I-DT-lr, in the FutureEO program. This work was supported by the Dutch National Growth Fund (NGF), as part of the Quantum Delta NL programme.

References

[1] John Preskill. "Quantum Computing in the NISQ era and beyond". Quantum **2**, 79 (2018).

[2] Kishor Bharti, Alba Cervera-Lierta, Thi Ha Kyaw, Tobias Haug, Sumner Alperin-Lea, Abhinav Anand, Matthias Degroote, Hermanni Heimonen, Jakob S. Kottmann, Tim Menke, Wai-Keong Mok, Sukin Sim, Leong-Chuan Kwek, and Alán Aspuru-Guzik. "Noisy intermediate-scale quantum algorithms". Reviews of Modern Physics **94**, 015004 (2022).

[3] M. 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". Nature Reviews Physics **3**, 625–644 (2021).

[4] Jarrod R. McClean, Matthew P. Harrigan, Masoud Mohseni, Nicholas C. Rubin, Zhang Jiang, Sergio Boixo, Vadim N. Smelyanskiy, Ryan Babbush, and Hartmut Neven. "Low depth mechanisms for quantum optimization". PRX Quantum **2**, 030312 (2021). arXiv:2008.08615.

[5] Lennart Bittel and Martin Kliesch. "Training variational quantum algorithms is NP-hard". Physical Review Letters **127**, 120502 (2021). arXiv:2101.07267.

[6] Alberto Peruzzo, Jarrod McClean, Peter Shadbolt, Man-Hong Yung, Xiao-Qi Zhou, Peter J. Love, Alán Aspuru-Guzik, and Jeremy L. O'Brien. "A variational eigenvalue solver on a photonic quantum processor". Nature Communications **5**, 4213 (2014).

[7] Jarrod R. McClean, Jonathan Romero, Ryan Babbush, and Alán Aspuru-Guzik. "The theory of variational hybrid quantum-classical algorithms". New Journal of Physics **18**, 023023 (2016).

[8] Ilya G. Ryabinkin, Scott N. Genin, and Artur F. Izmaylov. "Constrained Variational Quantum Eigensolver: Quantum Computer Search Engine in the Fock Space". Journal of Chemical Theory and Computation **15**, 249–255 (2019).

[9] Edward Farhi, Jeffrey Goldstone, and Sam Gutmann. "A Quantum Approximate Optimization Algorithm" (2014). arXiv:1411.4028.

[10] Yudong Cao, Jonathan Romero, Jonathan P. Olson, Matthias Degroote, Peter D. Johnson, Mária Kieferová, Ian D. Kivlichan, Tim Menke, Borja Peropadre, Nicolas P. D. Sawaya, Sukin Sim, Libor Veis, and Alán Aspuru-Guzik. "Quantum Chemistry in the Age of Quantum Computing". Chemical Reviews **119**, 10856–10915 (2019).

[11] Ying Li and Simon C. Benjamin. "Efficient Variational Quantum Simulator Incorporating Active Error Minimization". Physical Review X **7**, 021050 (2017).

[12] Cristina Cirstoiu, Zoe Holmes, Joseph Iosue, Lukasz Cincio, Patrick J. Coles, and Andrew Sornborger. "Variational Fast Forwarding for Quantum Simulation Beyond the Coherence Time". npj Quantum Information **6**, 82 (2020). arXiv:1910.04292.

[13] Kishor Bharti and Tobias Haug. "Quantum-assisted simulator". Physical Review A **104**, 042418 (2021).

[14] Sam McArdle, Tyson Jones, Suguru Endo, Ying Li, Simon C. Benjamin, and Xiao Yuan. "Variational ansatz-based quantum simulation of imaginary time evolution". npj Quantum Information **5**, 1–6 (2019).

[15] Xiao Yuan, Suguru Endo, Qi Zhao, Ying Li, and Simon C. Benjamin. "Theory of variational quantum simulation". Quantum **3**, 191 (2019).

[16] Kosuke Mitarai, Makoto Negoro, Masahiro Kitagawa, and Keisuke Fujii. "Quantum Circuit Learning". Physical Review A **98**, 032309 (2018). arXiv:1803.00745.

[17] Maria Schuld, Alex Bocharov, Krysta Svore, and Nathan Wiebe. "Circuit-centric quantum classifiers". Physical Review A **101**, 032308 (2020). arXiv:1804.00633.

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

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

[20] J. S. Otterbach, R. Manenti, N. Alidoust, A. Bestwick, M. Block, B. Bloom, S. Caldwell, N. Didier, E. Schuyler Fried, S. Hong, P. Karalekas, C. B. Osborn, A. Papageorge, E. C. Peterson, G. Prawiroatmodjo, N. Rubin, Colm A. Ryan, D. Scarabelli, M. Scheer, E. A. Sete, P. Sivarajah, Robert S. Smith, A. Staley, N. Tezak, W. J. Zeng, A. Hudson, Blake R. Johnson, M. Reagor, M. P. da Silva, and C. Rigetti. "Unsupervised Machine Learning on a Hybrid Quantum Computer" (2017). arXiv:1712.05771.

[21] Christa Zoufal, Aurélien Lucchi, and Stefan Woerner. "Quantum Generative Adversarial Networks for learning and loading random distributions". npj Quantum Information **5**, 1–9 (2019).

[22] Christa Zoufal, Aurélien Lucchi, and Stefan Woerner. "Variational quantum Boltzmann machines". Quantum Machine Intelligence **3**, 7 (2021).

[23] Pierre-Luc Dallaire-Demers and Nathan Killoran. "Quantum generative adversarial networks". Physical Review A **98**, 012324 (2018). arXiv:1804.08641.

[24] Maria Schuld and Nathan Killoran. "Quantum machine learning in feature Hilbert spaces". Physical Review Letters **122**, 040504 (2019). arXiv:1803.07128.

[25] Javier Gil Vidal and Dirk Oliver Theis. "Input Redundancy for Parameterized Quantum Circuits" (2020). arXiv:1901.11434.

[26] Maria Schuld, Ryan Sweke, and Johannes Jakob Meyer. "The effect of data encoding on the expressive power of variational quantum machine learning models". Physical Review A **103**, 032430 (2021). arXiv:2008.08605.

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

[28] Adrián Pérez-Salinas, Juan Cruz-Martinez, Abdulla A. Alhajri, and Stefano Carrazza. "Determining the proton content with a quantum computer". Physical Review D **103**, 034027 (2021).

[29] Sukin Sim, Peter D. Johnson, and Alán Aspuru-Guzik. "Expressibility and Entangling Capability of Parameterized Quantum Circuits for Hybrid Quantum-Classical Algorithms". Advanced Quantum Technologies **2**, 1900070 (2019).

[30] G. Cybenko. "Approximation by superpositions of a sigmoidal function". Mathematics of Control, Signals and Systems **2**, 303–314 (1989).

[31] Kurt Hornik. "Approximation capabilities of multilayer feedforward networks". Neural Networks **4**, 251–257 (1991).

[32] Adrián Pérez-Salinas, David López-Núñez, Artur García-Sáez, P. Forn-Díaz, and José I. Latorre. "One qubit as a universal approximant". Physical Review A **104**, 012405 (2021).

[33] Eric R. Anschuetz and Bobak T. Kiani. "Beyond Barren Plateaus: Quantum Variational Algorithms Are Swamped With Traps" (2022). arXiv:2205.05786.

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

[35] Supanut Thanasilp, Samson Wang, Nhat A. Nghiem, Patrick J. Coles, and M. Cerezo. "Subtleties in the trainability of quantum machine learning models". Quantum Machine Intelligence **5**, 21 (2023). arXiv:2110.14753.

[36] Zoë Holmes, Kunal Sharma, M. Cerezo, and Patrick J. Coles. "Connecting Ansatz Expressibility to Gradient Magnitudes and Barren Plateaus". PRX Quantum **3**, 010313 (2022).

[37] Thomas Hubregtsen, Josef Pichlmeier, Patrick Stecher, and Koen Bertels. "Evaluation of parameterized quantum circuits: On the relation between classification accuracy, expressibility, and entangling capability". Quantum Machine Intelligence **3**, 9 (2021).

[38] Martin Larocca, Piotr Czarnik, Kunal Sharma, Gopikrishnan Muraleedharan, Patrick J. Coles, and M. Cerezo. "Diagnosing Barren Plateaus with Tools from Quantum Optimal Control". Quantum **6**, 824 (2022). arXiv:2105.14377.

[39] Martin Larocca, Nathan Ju, Diego García-Martín, Patrick J. Coles, and M. Cerezo. "Theory of overparametrization in quantum neural networks" (2021). arXiv:2109.11676.

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

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

[42] Mario A. Muñoz, Michael Kirley, and Saman K. Halgamuge. "Exploratory Landscape Analysis of Continuous Space Optimization Problems Using Information Content". IEEE Transactions on Evolutionary Computation **19**, 74–87 (2015).

[43] Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. "On the difficulty of training Recurrent Neural Networks" (2013). arXiv:1211.5063.

[44] Tobias Friedrich, Timo Kötzing, Martin S. Krejca, and Amirhossein Rajabi. "Escaping Local Optima with Local Search: A Theory-Driven Discussion". In Günter Rudolph, Anna V. Kononova, Hernán Aguirre, Pascal Kerschke, Gabriela Ochoa, and Tea Tušar, editors, Parallel Problem Solving from Nature – PPSN XVII. Pages 442–455. Lecture Notes in Computer ScienceCham (2022). Springer International Publishing.

[45] Maithra Raghu, Ben Poole, Jon Kleinberg, Surya Ganguli, and Jascha Sohl-Dickstein. "On the Expressive Power of Deep Neural Networks" (2017). arXiv:1606.05336.

[46] Louis Schatzki, Martin Larocca, Quynh T. Nguyen, Frederic Sauvage, and M. Cerezo. "Theoretical Guarantees for Permutation-Equivariant Quantum Neural Networks" (2022). arXiv:2210.09974.

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

[48] Adriano Barenco, Charles H. Bennett, Richard Cleve, David P. DiVincenzo, Norman Margolus, Peter Shor, Tycho Sleator, John A. Smolin, and Harald Weinfurter. "Elementary gates for quantum computation". Physical Review A **52**, 3457–3467 (1995).

[49] Alice Barthe and Adrián Pérez-Salinas. "Github repository: `QRU_average`" (2023).

[50] Adrián Pérez-Salinas, Hao Wang, and Xavier Bonet-Monroig. "Analyzing variational quantum landscapes with information content" (2023). arXiv:2303.16893.

[51] Ingram Olkin and Herman Rubin. "Multivariate Beta Distributions and Independence Properties of the Wishart Distribution". The Annals of Mathematical Statistics **35**, 261–269 (1964).

[52] Patrick Billingsley. "Probability and Measure". Wiley. (1995).

[53] Franz J. Schreiber, Jens Eisert, and Johannes Jakob Meyer. "Classical surrogates for quantum learning models". Physical Review Letters **131**, 100803 (2023). arXiv:2206.11740.

[54] S. Shin, Y. S. Teo, and H. Jeong. "Exponential data encoding for quantum supervised learning". Physical Review A **107**, 012422 (2023).

[55] Julia Kempe, Alexei Kitaev, and Oded Regev. "The Complexity of the Local Hamiltonian Problem" (2005). arXiv:quant-ph/0406180.

[56] Henry Gouk, Eibe Frank, Bernhard Pfahringer, and Michael J. Cree. "Regularisation of neural networks by enforcing Lipschitz continuity". Machine Learning **110**, 393–416 (2021).

[57] Peter Bartlett, Dylan J. Foster, and Matus Telgarsky. "Spectrally-normalized margin bounds for neural networks" (2017). arXiv:1706.08498.

[58] Evan Peters and Maria Schuld. "Generalization despite overfitting in quantum machine learning models" (2022). arXiv:2209.05523.

[59] Enrico Fontana, Manuel S. Rudolph, Ross Duncan, Ivan Rungger, and Cristina Cîrstoiu. "Classical simulations of noisy variational quantum circuits" (2023). arXiv:2306.05400.

[60] Manuel S. Rudolph, Enrico Fontana, Zoë Holmes, and Lukasz Cincio. "Classical surrogate simulation of quantum systems with LOWESA" (2023). arXiv:2308.09109.

[61] Sergey Bravyi, David Gosset, and Yinchen Liu. "Classical simulation of peaked shallow quantum circuits" (2023). arXiv:2309.08405.

[62] Sergio Boixo, Sergei V. Isakov, Vadim N. Smelyanskiy, Ryan Babbush, Nan Ding, Zhang Jiang, Michael J. Bremner, John M. Martinis, and Hartmut Neven. "Characterizing Quantum Supremacy in Near-Term Devices". Nature Physics **14**, 595–600 (2018). arXiv:1608.00263.

[63] Ralph W. Bailey. "Distributional Identities of Beta and Chi-Squared Variates: A Geometrical Interpretation". The American Statistician **46**, 117–120 (1992). arXiv:2684178.

[64] Wassily Hoeffding. "Probability Inequalities for Sums of Bounded Random Variables". Journal of the American Statistical Association **58**, 13–30 (1963).

[65] A. Zee. "Quantum field theory in a nutshell". In a Nutshell. Princeton University Press. Princeton, N.J (2010). 2nd ed edition.

[66] user26872. "Answer to "reference for multidimensional gaussian integral"" (2012).

A Proofs

A.1 Proof of Theorem 3.1

We begin by explicitly writing the derivatives of the hypothesis function

$$\partial_j h_{\boldsymbol{\theta}}(x) = \mathrm{Tr}\Big\{U_{R,j}(\boldsymbol{\theta}_{R,j}, x)\ \rho_0 U_{R,j}^{\dagger}(\boldsymbol{\theta}_{R,j}, x)\left[V_j, U_{L,j}^{\dagger}(\boldsymbol{\theta}_{L,j}, x) H U_{L,j}(\boldsymbol{\theta}_{L,j}, x)\right]\Big\}. \tag{47}$$

In this equation, the indices R, L indicate all operations before and after the $j-$th operation. We redefine the quantities for readability

$$\rho_j(\boldsymbol{\theta}_{R,j}, x) = U_{R,j}(\boldsymbol{\theta}_{R,j}, x)\ \rho_0\ U_{R,j}^{\dagger}(\boldsymbol{\theta}_{R,j}, x) \tag{48}$$

$$H_j(\boldsymbol{\theta}_{L,j}, x) = U_{L,j}^{\dagger}(\boldsymbol{\theta}_{L,j}, x)\ H\ U_{L,j}(\boldsymbol{\theta}_{L,j}, x) \tag{49}$$

The variance of these derivatives over $\boldsymbol{\Theta}$ is given by

$$\mathbb{E}_X\left(\mathrm{Var}_{\boldsymbol{\Theta}}\left(\partial_j h_{\boldsymbol{\theta}}(x)\right)\right) = \mathbb{E}_{\boldsymbol{\Theta}}\left(\mathrm{Var}_X\left(\partial_j h_{\boldsymbol{\theta}}(x)\right)\right) = \mathbb{E}_{\boldsymbol{\Theta}_{R,j}}\left(\mathbb{E}_{\boldsymbol{\Theta}_{L,j}}\left(\mathbb{E}_X\left(\left(\partial_j h_{\boldsymbol{\theta}}(x)\right)^2\right)\right)\right), \tag{50}$$

where we assume no correlation between the parameters in the left and right parts of the circuit. By calling the property $\mathrm{Tr}\{A \otimes B\} = \mathrm{Tr}\, A\, \mathrm{Tr}\, B$, we can plug Equation (47) into Equation (50) to obtain

$$\mathrm{Var}_{\boldsymbol{\Theta},X}\left(\partial_j h_{\boldsymbol{\theta}}(x)\right) = \mathbb{E}_{\boldsymbol{\Theta}_{R,j}}\left(\mathbb{E}_{\boldsymbol{\Theta}_{L,j}}\left(\mathbb{E}_X\left(\mathrm{Tr}\Big\{\rho_j(\boldsymbol{\theta}_{R,j}, x)^{\otimes 2}\left[V_j, H_j(\boldsymbol{\theta}_{L,j}, x)\right]^{\otimes 2}\Big\}\right)\right)\right) \tag{51}$$

We aim to describe this quantity in terms of the difference between the QML models, partially described by data x, and their corresponding PQC models, where $x = 0$. We define the corresponding difference operators

$$B_{R,j}^{(t)}(\boldsymbol{\theta}_{R,j}, x; \rho_0) = \rho_j^{\otimes t}(\boldsymbol{\theta}_{R,j}, x) - \rho_j^{\otimes t}(\boldsymbol{\theta}_{R,j}, 0) \tag{52}$$

$$B_{L,j}^{(t)}(\boldsymbol{\theta}_{L,j}, x; H) = H_j^{\otimes t}(\boldsymbol{\theta}_{L,j}, x) - H_j^{\otimes t}(\boldsymbol{\theta}_{L,j}, 0). \tag{53}$$

We rearrange the terms in the integrand of Equation (51) as

$$\mathrm{Tr}\Big\{\rho_j(\boldsymbol{\theta}_{R,j}, x)^{\otimes 2}\left[V_j, H_j(\boldsymbol{\theta}_{L,j}, x)\right]^{\otimes 2}\Big\} = \tag{54}$$

$$\mathrm{Tr}\big\{\rho_j(\boldsymbol{\theta}_{R,j}, 0)^{\otimes 2}[V_j, H_j(\boldsymbol{\theta}_{R,j}, x)]^{\otimes 2}\big\} + \mathrm{Tr}\Big\{B_{R,j}^{(2)}(\boldsymbol{\theta}_{R,j}, x; \rho_0)[V_j, H_j(\boldsymbol{\theta}_{L,j}, x)]^{\otimes 2}\Big\} = \tag{55}$$

$$\mathrm{Tr}\big\{H_j(\boldsymbol{\theta}_{R,j}, x)^{\otimes 2}[\rho_j(\boldsymbol{\theta}_{R,j}, 0), V_j]^{\otimes 2}\big\} + \mathrm{Tr}\Big\{B_{R,j}^{(2)}(\boldsymbol{\theta}_{R,j}, x; \rho_0)[V_j, H_j(\boldsymbol{\theta}_{L,j}, x)]^{\otimes 2}\Big\} = \tag{56}$$

$$\mathrm{Tr}\big\{\rho_j(\boldsymbol{\theta}_{R,j}, 0)^{\otimes 2}[V_j, H_j(\boldsymbol{\theta}_{R,j}, 0)]^{\otimes 2}\big\} + \tag{57}$$

$$\mathrm{Tr}\Big\{B_{R,j}^{(2)}(\boldsymbol{\theta}_{R,j}, x; \rho_0)[V_j, H_j(\boldsymbol{\theta}_{L,j}, x)]^{\otimes 2}\Big\} + \tag{58}$$

$$\mathrm{Tr}\Big\{B_{L,j}^{(2)}(\boldsymbol{\theta}_{L,j}, x; H)[\rho_j(\boldsymbol{\theta}_{R,j}, 0), V_j]^{\otimes 2}\Big\} \tag{59}$$

by recalling the identities $\mathrm{Tr}\big\{A^{\otimes 2}\big\} = \mathrm{Tr}\{A\}^2$ and $\mathrm{Tr}\{A[B, C]\} = \mathrm{Tr}\{B[C, A]\} = \mathrm{Tr}\{C[A, B]\}$. The term in Equation (57) corresponds to the standard variance in PQC. We denote it simply as $\mathrm{Var}_{\boldsymbol{\Theta}}\left(\partial_j h_{\boldsymbol{\theta}}(0)\right)$.

We move our attention now Equation (58). This term measures the difference between QRU models and PQC in the right part of the quantum circuit. Assuming that the right and left parameters are uncorrelated, we can rewrite

$$\mathbb{E}_{\boldsymbol{\Theta}_{R,j}}\left(\mathbb{E}_{\boldsymbol{\Theta}_{L,j}}\left(\mathbb{E}_X\left(\mathrm{Tr}\Big\{B_{R,j}^{(2)}(\boldsymbol{\theta}_{R,j}, x; \rho_0)\left[V_j, H(\boldsymbol{\theta}_{L,j}, x)\right]^{\otimes 2}\Big\}\right)\right)\right) = \\ \mathrm{Tr}\Big\{\left(\mathbb{E}_X\left(\mathbb{E}_{\boldsymbol{\Theta}_{R,j}}\left(B_{R,j}^{(2)}(\boldsymbol{\theta}_{R,j}, x; \rho_0)\right)\right)\right)\left[V_j, \mathbb{E}_{\boldsymbol{\Theta}_{L,j}}\left(H(\boldsymbol{\theta}_{L,j}, x)\right)\right]^{\otimes 2}\Big\} \tag{60}$$

Using von Neumann's trace and Hölder inequalities, with Schatten norms

$$\left|\mathrm{Tr}\{AB\}\right| \leq \|A\|_1 \|B\|_{\infty}, \tag{61}$$

in Equation (60) together with the triangular and Cauchy-Schwarz inequality we obtain

$$\left|\mathbb{E}_X\left(\mathrm{Tr}\Big\{\mathbb{E}_{\boldsymbol{\Theta}_{R,j}}\left(B_{R,j}^{(2)}(\boldsymbol{\theta}_{R,j}, x; \rho_0)\right)\Big\}\left[V_j, \mathbb{E}_{\boldsymbol{\Theta}_{L,j}}\left(H(\boldsymbol{\theta}_{L,j}, x)\right)\right]^{\otimes 2}\right)\right| \leq \\ \mathbb{E}_X\left(\left|\mathrm{Tr}\Big\{\mathbb{E}_{\boldsymbol{\Theta}_{R,j}}\left(B_{R,j}^{(2)}(\boldsymbol{\theta}_{R,j}, x; \rho_0)\right)\Big\}\left[V_j, \mathbb{E}_{\boldsymbol{\Theta}_{L,j}}\left(H(\boldsymbol{\theta}_{L,j}, x)\right)\right]^{\otimes 2}\right|\right) \leq \\ \mathbb{E}_X\left(\left\|\mathbb{E}_{\boldsymbol{\Theta}_{L,j}}\left(\left[V_j, H(\boldsymbol{\theta}_{L,j}, x)\right]^{\otimes 2}\right)\right\|_{\infty}\left\|\mathbb{E}_{\boldsymbol{\Theta}_{R,j}}\left(B_{R,j}(\boldsymbol{\theta}_{R,j}, x; \rho_0)\right)\right\|_1\right). \tag{62}$$

Substituting in the previous equation the property [36]

$$\left\|[V_j, H(\boldsymbol{\theta}_{L,j}, x)]^{\otimes 2}\right\|_\infty \leq 4\|V_j\|_\infty^2 \|H\|_\infty^2, \tag{63}$$

and defining

$$\mathcal{B}_{R,j}^{(2)}(\rho_0) = \mathbb{E}_X\left(\left\|\mathbb{E}_{\boldsymbol{\Theta}_{R,j}}\left(B\right)_{R,j}^{(2)}(\boldsymbol{\theta}_{R,j}, x; \rho_0)\right\|_1\right), \tag{64}$$

we obtain

$$\mathbb{E}_{\boldsymbol{\Theta}_{R,j}}\left(\mathbb{E}_{\boldsymbol{\Theta}_{L,j}}\left(\mathbb{E}_X\left(\mathrm{Tr}\left\{B_{R,j}^{(2)}(\boldsymbol{\theta}_{R,j}, x; \rho_0)\left[V_j, H(\boldsymbol{\theta}_{L,j}, x)\right]^{\otimes 2}\right\}\right)\right)\right) \leq 4\|V_j\|_\infty^2\|H\|_\infty^2\ \mathcal{B}_{R,j}^{(2)}(\rho_0). \tag{65}$$

Following the same steps for Equation (59), we can bound this quantity as

$$\mathbb{E}_{\boldsymbol{\Theta}_{R,j}}\left(\mathbb{E}_{\boldsymbol{\Theta}_{L,j}}\left(\mathbb{E}_X\left(\mathrm{Tr}\left\{B_{L,j}^{(2)}(\boldsymbol{\theta}_{L,j}, x; H)[V_j, \rho_R(\boldsymbol{\theta}_{R,j}, 0)]^{\otimes 2}\right\}\right)\right)\right) \leq 4\|V_j\|_\infty^2\|\rho_0\|_\infty^2\ \mathcal{B}_{L,j}^{(2)}(H), \tag{66}$$

where we defined analogously

$$\mathcal{B}_{L,j}^{(2)}(H) = \mathbb{E}_X\left(\left\|\mathbb{E}_{\boldsymbol{\Theta}_{L,j}}\left(B_{L,j}^{(2)}(\boldsymbol{\theta}_{L,j}, x; H)\right)\right\|_1\right). \tag{67}$$

Notice that we could interchange the x-dependency in Equation (59) and Equation (58) with no effect in the final bounds. The reason is that Equation (63) eliminates the x-dependency in the term where it is applied. We can compact the results from Equations (65) and (66) using the triangular inequality in

$$\left|\mathrm{Var}_{\boldsymbol{\Theta}}\left(\partial_j h_{\boldsymbol{\theta}}(0)\right) - \mathrm{Var}_{\boldsymbol{\Theta}}\left(\mathbb{E}_X\left(\partial_j h_{\boldsymbol{\theta}}(x)\right)\right)\right| \leq 4\|V_j\|_\infty^2\left(\|H\|_\infty^2\ \mathcal{B}_{R,j}^{(2)}(\rho_0) + \ \|\rho_0\|_\infty^2\ \mathcal{B}_{L,j}^{(2)}(h)\right) \tag{68}$$

□

A.2 Proof of Lemma 3.1

We start by bounding the right absorption witness of the $(l+1)$-th layer with the triangular and Hölder's inequality as

$$\begin{aligned}\mathcal{B}_{R,l+1}^{(2)}(\rho_0) = \mathbb{E}_X\left(\left\|\mathbb{E}_{\boldsymbol{\Theta}_{l+1}}\left(u(\boldsymbol{\theta}_{l+1})^{\otimes 2}V(x)^{\otimes 2}\ \mathbb{E}_{\boldsymbol{\Theta}_{R,l}}\left(\rho_l(\boldsymbol{\theta}_{R,l}, x)\right)V^\dagger(x)^{\otimes 2}u^\dagger(\boldsymbol{\theta}_{l+1})^{\otimes 2}\right)\right\|_2\right) \leq \\ \mathbb{E}_X\left(\left\|\mathbb{E}_{\boldsymbol{\Theta}_{l+1}}\left(u(\boldsymbol{\theta}_{l+1})^{\otimes 2}V(x)^{\otimes 2}\ \mathbb{E}_{\boldsymbol{\Theta}_{R,l}}\left(B_{R,l}^{(2)}(\boldsymbol{\theta}_{R,l}, x; \rho_0)\right)V^\dagger(x)^{\otimes 2}u^\dagger(\boldsymbol{\theta}_{l+1})^{\otimes 2}\right)\right\|_1\right) + \\ \mathbb{E}_X\left(\left\|\mathbb{E}_{\boldsymbol{\Theta}_{l+1}}\left(u(\boldsymbol{\theta}_{l+1})^{\otimes 2}V(x)^{\otimes 2} - u(\boldsymbol{\theta}_{l+1})^{\otimes 2}\right)\right\|_1\left\|\mathbb{E}_{\boldsymbol{\Theta}_{R,l}}\left(\rho_l(\boldsymbol{\theta}_{R,l}, 0)\right)\right\|_\infty\right)\end{aligned} \tag{69}$$

The second term can be identified as the layerwise absorption witness from Definition 3.2. The first term of the equation above can be bounded as

$$\begin{aligned}\mathbb{E}_X\left(\left\|\mathbb{E}_{\boldsymbol{\Theta}_{l+1}}\left(u(\boldsymbol{\theta}_{l+1})^{\otimes 2}V(x)^{\otimes 2}\ \mathbb{E}_{\boldsymbol{\Theta}_{R,l}}\left(B_{R,l}^{(2)}(\boldsymbol{\theta}_{R,l}, x; \rho_0)\right)V^\dagger(x)^{\otimes 2}u^\dagger(\boldsymbol{\theta}_{l+1})^{\otimes 2}\right)\right\|_1\right) \leq \\ \mathbb{E}_X\left(\mathbb{E}_{\boldsymbol{\Theta}_{l+1}}\left(\left\|u(\boldsymbol{\theta}_{l+1})^{\otimes 2}V(x)^{\otimes 2}\ \mathbb{E}_{\boldsymbol{\Theta}_{R,l}}\left(B_{R,l}^{(2)}(\boldsymbol{\theta}_{R,l}, x; \rho_0)\right)V^\dagger(x)^{\otimes 2}u^\dagger(\boldsymbol{\theta}_{l+1})^{\otimes 2}\right\|_1\right)\right) = \mathcal{B}_{R,l}^{(2)}(\rho_0).\end{aligned} \tag{70}$$

Arranging both results together we can find

$$\mathcal{B}_{R,l+1}^{(2)}(\rho_0) \leq \mathcal{B}_{R,l}^{(2)}(\rho_0) + \|\rho_0\|_\infty^2\mathcal{A}_{l+1}^{(2)}. \tag{71}$$

Equivalently for the left part of the circuit, and counting layers backward we find

$$\mathcal{B}_{L,l}^{(2)}(H) \leq \mathcal{B}_{L,l+1}^{(2)}(\rho_0) + \|H\|_\infty^2\mathcal{A}_l^{(2)}. \tag{72}$$

□

A.3 Details on harmonic representation of QRU models

As stated in Equation (25), the wavefunction after a re-uploading circuit and before measurement can be expressed as

$$|\psi(x)\rangle = \sum_{j=1}^{2^n} \sum_{k=-K}^{K} c_{k,j} e^{ikx} |j\rangle \,. \tag{73}$$

The coefficients $c_{k,j}$ form the matrix $\boldsymbol{C} \in \mathbb{C}^{2^n \times (2K+1)}$. Each column (indexed with k) represents the corresponding term e^{ikx}. The states $|j\rangle$ are elements of any basis of choice. Each row (indexed with j) corresponds to the x-dependent amplitude attached $|j\rangle$. The quantity $p_j(x) = \langle j|\psi(x)\rangle$ is a trigonometric polynomial, $p_j(x) = \sum_{k=-K}^{K} c_{k,j} e^{ikx}$. Such polynomial can be represented as a vector $p_j = \{c_{k,j}\}_{-K \leq k \leq +K}$. In this vector representation, the multiplication of polynomials corresponds to convolution as

$$p(x)q(x) = \sum_{k=-K_p}^{K_p} p_k e^{ikx} \sum_{k=-K_q}^{K_q} q_k e^{ikx} = (p * q)(x), \tag{74}$$

We consider the three operations that can be applied to the harmonic representation.

Parametrized gates Applying a parameterized gate on the quantum state maps into applying the unitary representation of that gate to each column individually, the same way it would be done in state vector simulation for each (fixed) k.

Data-encoding gates Adding a data-encoding gate involves convolution, when $\boldsymbol{C}$ is expressed in the eigenbasis of the data generator. Each row j corresponds to the j-th (integer) eigenvalue λ_j from the spectrum of the data-encoding generator, $\mathcal{K}_g$. The vector representation of polynomials p_j is convoluted with the vector $e_{\lambda_j} = [\delta_{k=\lambda_j}]_{-K \leq k \leq +K}$.

Measurement For the measurement, we express $\boldsymbol{C}$ the basis of the observable. Secondly the representation the transpose conjugate of the wavefunction is computed from $\boldsymbol{C}$, by taking its conjugate and reverting the rows indexing $c_{j,k} = c_{j,-k}$. Finally the hypothesis function $h_{\boldsymbol{\theta}}(x)$ can be obtained as a linear combination of the rows of the result of the convolution weighted by the corresponding eigenvalue of the measurement operator.

A.4 Proof of Theorem 4.1

We assume now that the parameterized gates between two consecutive encoding steps are random unitaries sampled from the Haar measure of the group $\mathcal{SU}(N)$. Notice that data-independent operations leave the norm of coefficients associated with the same frequency invariant. Random unitaries output random states for any input. Random states give rise to a probability distribution sampled from a uniform Dirichlet [51], also known as Porter-Thomas [62] distribution. Under this assumption, no basis has any preference over any other, and the application of the data encoding layer *transports* as many coefficients as dictated by the spectrum $\mathcal{K}_{g_j}$ to the corresponding new frequencies. Notice that it is irrelevant which coefficients are transported, and also they are randomly chosen. The weights for the elements of each frequency are then described by a Dirichlet distribution with parameters given by the convoluted spectrum. Formally,

$$\sum_j |c_{k,j}|^2 \sim \mathrm{Dir}\left(\left(\mathcal{K}_{g_1} * \ldots * \mathcal{K}_{g_j} * \ldots * \mathcal{K}_{g_L}\right)(k)\right), \tag{75}$$

where $*$ denotes the convolution operator. Notice that the index k runs over all non-zero entries of the convoluted spectra. □

A.5 Dirichlet distribution

Definition A.1 (Dirichlet distribution [51])**.** *The Dirichlet distribution $\boldsymbol{x} \sim \mathrm{Dir}(\boldsymbol{\alpha})$ parameterized by $\alpha \in \mathbb{R}^N_{>0}$ is supported on the $(N-1)$-standard simplex, i.e., $\boldsymbol{x} = (x_1, x_2, \ldots, x_N)$, $\|\boldsymbol{x}\|_1 = 1$. It has the following probability density function with respect to the Lebesgue measure on $\mathbb{R}^{N-1}$:*

$$f_{\mathrm{Dir}}(\boldsymbol{x}, \boldsymbol{\alpha}) = \frac{\Gamma(\|\boldsymbol{\alpha}\|_1)}{\prod_{i=1}^N \Gamma(\alpha_i)} \prod_{i=1}^N x_i^{\alpha_i - 1}. \tag{76}$$

In this definition, $\Gamma(\cdot)$ is defined as

$$\Gamma(z) = \int_0^\infty t^{z-1} e^{-t} dt, \tag{77}$$

being the complex extension of the factorial for positive integers

$$\Gamma(n) = (n-1)!. \tag{78}$$

The Dirichlet distribution admits straightforward analytical calculations for the statistical moments of arbitrary order $\boldsymbol{k} = (k_1, k_2, \ldots, k_N)$,

$$\mathbb{E}_{\boldsymbol{x}\sim\mathrm{Dir}(\boldsymbol{\alpha})}\left(\prod_{i=1}^{N} x_i^{k_i}\right) = \frac{\Gamma(\|\boldsymbol{\alpha}\|_1)}{\Gamma(\|\boldsymbol{\alpha}\|_1 + \|\boldsymbol{k}\|_1)} \prod_{i=1}^{N} \frac{\Gamma(\alpha_i + k_i)}{\Gamma(\alpha_i)}. \tag{79}$$

In particular

$$\mathbb{E}_{\boldsymbol{x}\sim\mathrm{Dir}(\boldsymbol{\alpha})}(x_i) = \frac{\alpha_i}{\|\boldsymbol{\alpha}\|_1} \tag{80}$$

$$\mathrm{Var}_{\boldsymbol{x}\sim\mathrm{Dir}(\boldsymbol{\alpha})}(x_i) = \frac{\alpha_i\left(1 - \frac{\alpha_i}{\|\boldsymbol{\alpha}\|_1}\right)}{\|\boldsymbol{\alpha}\|_1(\|\boldsymbol{\alpha}\|_1 + 1)} \tag{81}$$

$$\mathrm{Cov}_{\boldsymbol{x}\sim\mathrm{Dir}(\boldsymbol{\alpha})}(x_i, x_j) = \frac{-\alpha_i\alpha_j}{\|\boldsymbol{\alpha}\|_1^2(\|\boldsymbol{\alpha}\|_1 + 1)} \tag{82}$$

A.6 Proof of Corollary 4.2

Let us express the wavefunction as the $\boldsymbol{C}$ matrix, such that the wavefunction is reconstructed from its elements as

$$|\psi(x)\rangle = \sum_{j=1}^{2^n} \sum_{k=-K}^{K} c_{k,j} e^{ik\mu x} |j\rangle, \tag{83}$$

where $|j\rangle$ is expressed in this case the eigenbasis of the observable of interest H. We are interested in the function

$$h(x) = \langle\psi(x)| H |\psi(x)\rangle, \tag{84}$$

which in the eigenbasis of H is

$$h(x) = \sum_j \sum_k \sum_l \lambda_j c_{k,j} c_{l,j}^* e^{i\mu(k-l)\ x}. \tag{85}$$

We give a bound now on the terms sharing the same frequencies

$$\left|\sum_j \sum_{k-l=\omega} \lambda_j c_{k,j} c_{l,j}^*\right|^2 \leq \|H\|_\lambda^2 \left|\sum_j \sum_{k-l=\omega} c_{k,j} c_{l,j}^*\right|^2 \leq \|H\|_\lambda^2 \sum_{k-l=\omega} \left(\sum_j |c_{k,j}|^2\right)\left(\sum_j |c_{l,j}|^2\right), \tag{86}$$

where we used the triangular inequality and the Cauchy-Schwarz inequality. The last term is related to Theorem 4.1. Each of these elements is drawn from the Dirichlet distribution imposed by the spectrum $\mathcal{K}_g^{*L}$. The aggregation property of Dirichlet distributions allows us to directly work with the spectrums. The spectrum of interest is a modified convolution of $\mathcal{K}_g^{*L}$ with itself under an inversion of the variable, namely

$$\sum_{k-l=\omega} \mathcal{K}_g^{*L}(k)\mathcal{K}_g^{*L}(l) = \sum_k \mathcal{K}_g^{*L}(k)\mathcal{K}_g^{*L}(k-\omega) = \sum_k \mathcal{K}_g^{*L}(k)\mathcal{K}_g^{*L}(-(k-\omega)) = \left(\mathcal{K}_g^{*L} * \mathcal{K}_g'^{*L}\right)(\omega), \tag{87}$$

with $\mathcal{K}_g^{*L}(x) = \mathcal{K}_g'^{*L}(-x)$. In the case of symmetric spectra, both functions are equivalent. Recalling the properties of Dirichlet distributions, we can bound

$$\|H\|_\lambda^{-2} |a_\omega(\boldsymbol{\theta})|^2 \leq p_\omega \sim \mathrm{Dir}(\mathcal{K}_g^{*2L}(\omega)), \tag{88}$$

with

$$a_\omega(\boldsymbol{\theta}) = \sum_j \sum_{k-l=\omega} \lambda_j c_{k,j} c_{l,j}^* \tag{89}$$

□

A.7 Proof of Theorem 4.2

We use tools from statistics to compute upper and lower bounds to the Lipschitz constant of a hypothesis function $\Lambda(h_{\boldsymbol{\theta}})$. We first recall the definition

$$\Lambda(h_{\boldsymbol{\theta}}) := \sum_{k=-K}^{K} \mu|k||a_k|, \tag{90}$$

and we recall the result from Corollary 4.2 in the limit of many re-uploadings. We know that

$$\mathbb{E}_{\boldsymbol{\Theta}}\left(\Lambda(h_{\boldsymbol{\theta}})\right) = \sum_{k=-K}^{K} \mu|k|\mathbb{E}_{\boldsymbol{\Theta}}\left(|a_k|\right). \tag{91}$$

We can use the known bound for the probability distribution underlying $|a_k|$, $|a_k|^2 \leq p_k \sim \mathrm{Dir}(\mathcal{K}_g^{*L})$. In particular, the marginals of the Dirichlet distribution are beta distributions [63]. The beta probability distribution with parameters α and β is defined as

$$\mathrm{Beta}_{\alpha,\beta}(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{(\alpha-1)}(1-x)^{(\beta-1)}. \tag{92}$$

In our case, see Equation (34), α is given by the Gaussian spectrum and $\beta = 1$. The expectation value of each element is given by

$$\mathbb{E}_{\boldsymbol{\Theta}}\left(|a_k|\right) \leq \int_0^1 dx \frac{\Gamma(\alpha_k+1)}{\Gamma(\alpha_k)} x^{(\alpha_k-1/2)} = \frac{\alpha_k}{\alpha_k+1/2} \leq 2\alpha_k, \tag{93}$$

using the property of the gamma function $\Gamma(1+x) = x\Gamma(x)$. The last inequality allows us to compute an upper bound in the limit of many re-uploadings by just computing

$$\sum_{k=-K}^{K} \mu k\mathbb{E}_{\boldsymbol{\Theta}}\left(|a_k|\right) \leq 2\|H\|_\lambda \sum_{k=-K}^{K} \mu k\mathcal{K}_g^{*2L}(k) \approx 4\|H\|_\lambda \int_0^\infty \frac{\mu k}{\sqrt{4\pi L}\sigma_g} \exp\left(-\frac{k^2}{4\sigma_g^2 L}\right) = \|H\|_\lambda \frac{4}{\sqrt{\pi}}\sigma_g\mu\sqrt{L}, \tag{94}$$

leading to the first result of the theorem.

The lower bound is easy to obtain by recalling the property $\|a\|_1 \geq \|a\|_2$. In our context, and in the limit of Gaussian processes

$$\mathbb{E}_{\boldsymbol{\Theta}}\left(\Lambda(h_{\boldsymbol{\theta}})\right)^2 \geq \|H\|_\lambda^2 \sum_{k=-K}^{K} \mu^2k^2\mathbb{E}_{\boldsymbol{\Theta}}\left(|a_k|^2\right) \approx \|H\|_\lambda^2 2L\mu^2\sigma_g^2, \tag{95}$$

leading to the second result of the theorem: $\mathbb{E}_{\boldsymbol{\Theta}}\left(\Lambda(h_{\boldsymbol{\theta}})\right) \geq \|H\|_\lambda\sigma_g\mu\sqrt{2L}$. □

A.8 Proof of Theorem 4.3

We are interested in knowing the probability of $\Lambda(h_{\boldsymbol{\theta}})$ to be larger than a certain reference value by some distance. We take this reference value to be the average $\mathbb{E}_{\boldsymbol{\Theta}}\left(\Lambda(h_{\boldsymbol{\theta}})\right)$, as in many statistics results. Consider now the lower-bound on the expectation value from Theorem 4.2. Since

$$\Lambda(h_{\boldsymbol{\theta}}) - \mathbb{E}_{\boldsymbol{\Theta}}\left(\Lambda(h_{\boldsymbol{\theta}})\right) \geq t \implies \Lambda(h_{\boldsymbol{\theta}}) - \|H\|_\lambda\mu\sqrt{2L}\sigma_g \geq t, \tag{96}$$

but not in the opposite direction, then

$$\mathrm{Prob}_{\boldsymbol{\Theta}}\left(\Lambda(h_{\boldsymbol{\theta}}) - \mathbb{E}_{\boldsymbol{\Theta}}\left(\Lambda(h_{\boldsymbol{\theta}})\right) \geq t\right) \leq \mathrm{Prob}_{\boldsymbol{\Theta}}\left(\Lambda(h_{\boldsymbol{\theta}}) - \|H\|_\lambda\mu\sqrt{2L}\sigma_g \geq t\right). \tag{97}$$

We can bound the right hand by considering Hoeffding's inequality [64]. Let X_i be a set of independent random variables, and let $X_i \in [a_i, b_i]$ *almost surely*, then

$$\mathrm{Prob}\left(\sum_i X_i - \mathbb{E}\left(\sum_i X_i\right) \geq t\right) \leq \frac{1}{2}\exp\left(\frac{-t^2}{\sum_i (a_i-b_i)^2}\right). \tag{98}$$

Hoeffding's inequality cannot be directly applied to a Dirichlet distribution since the variables are not independent. However, this problem can be overcome for this particular case by recalling the following property. If $X_i \sim \mathrm{Dir}(\alpha_i)$, then

$$X_i \sim \frac{Y_i}{V}, \tag{99}$$

with

$$Y_i \sim \text{Gamma}(\alpha_i, \theta) \tag{100}$$

$$V = \sum_i Y_i \sim \text{Gamma}\left(\sum_i \alpha_i, \theta\right) \tag{101}$$

By changing the description of the Dirichlet distribution to the quotient of gamma distributions we can now apply Hoeffding's inequality. Without loss of generality, we can assume that all probabilities are bounded between 0 and 1, thus we can find a first bound by recalling

$$\sum_n n^2 \in \mathcal{O}((\|g\|_\lambda L)^3), \tag{102}$$

and thus

$$\text{Prob}_{\boldsymbol{\Theta}}\left(\Lambda(h_{\boldsymbol{\theta}}) - \mathbb{E}_{\boldsymbol{\Theta}}\left(\Lambda(h_{\boldsymbol{\theta}})\right) \geq t\right) \leq \text{Prob}_{\boldsymbol{\Theta}}\left(\Lambda(h_{\boldsymbol{\theta}}) - \|H\|_\lambda \sqrt{2L}\mu\sigma_g \geq t\right) \in \mathcal{O}\left(\exp\left(\frac{-t^2}{(\|g\|_\lambda L)^3}\right)\right) \tag{103}$$

□

A.8.1 A tighter numerical bound

This bound can be however easily improved by recalling subgaussianity properties of the Gamma distribution. A random variable X is subgaussian if its cumulative distribution function decays faster than exponentially

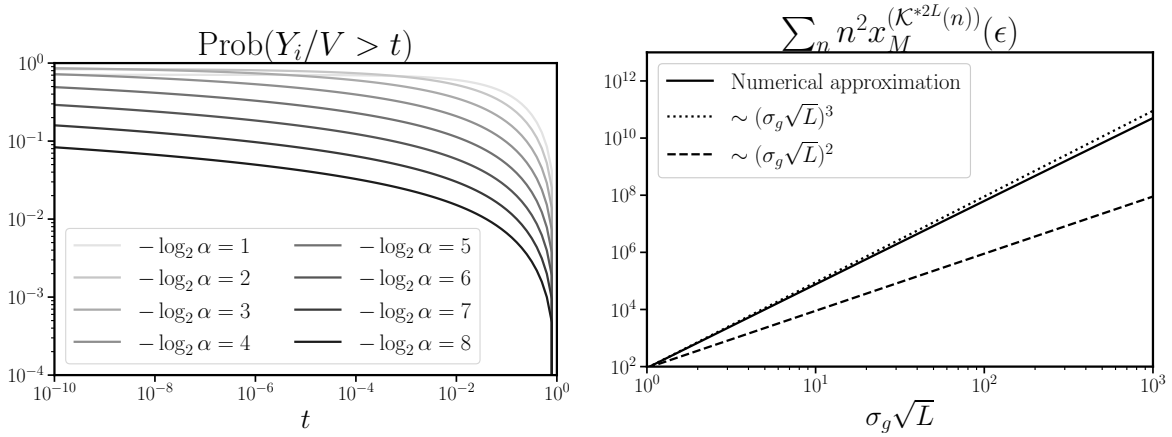
$$\text{Prob}\left(|X| \geq t\right) \in \mathcal{O}\left(\exp\left(-t^2\right)\right), \tag{104}$$

for some positive constant K. We can compute this cumulative probability for the quotient of Gamma distributions as

$$\text{Prob}\left(\frac{Y_i}{V} \geq t\right) = \int_0^\infty dx \int_x^{x/t} dy \frac{x^{\alpha-1}e^{-x}e^{-y}}{\Gamma(\alpha)} = \frac{1}{2^\alpha} - \frac{1}{(1+t^{-1})^\alpha} \in \mathcal{O}\left(\exp\left(-t^2\right)\right). \tag{105}$$

These functions take the value 1 for $t = 0$ and decay until vanishing for $t = 1$. The decay is faster as $\alpha \to 0$, as it can be seen in Figure 8(a). We can thus recover Hoeffding's inequality with the observation that each X_i is bounded by the function in Equation (105). In particular, the variable X_i is, with probability $1 - \epsilon$, smaller than

$$x_M^{(\alpha_i)}(\epsilon) = \left(\left(\frac{1}{2^{\alpha_i}} - \epsilon\right)^{-1/\alpha_i} - 1\right)^{-1}. \tag{106}$$



(a) Numerical calculations for Equation (105) for decreasing values of α. The values of the random variable concentrate in small values for t as α decays.

(b) Numerical approximation to Equation (107) for increasing $\sigma_g\sqrt{L}$. The value ϵ is set in this calculations to 10^{-10}

Figure 8: Numerical auxiliary calculations for Equation (105) and Equation (107). These results substitute the non-accessible analytical treatment of the probability distributions of interest to obtain the bound given in Equation (40)

For a sufficiently small ϵ, the denominator of the exponent of Hoeffding's inequality becomes

$$\sum_n (a_n - b_n)^2 = 2 \sum_{n=1}^{\|g\|_2 L} n^2 x_M^{(\mathcal{K}_g^{*2L}(n))}(\epsilon), \tag{107}$$

with $\mathcal{K}_g^{*2L}$ a Gaussian spectrum in the limit of large R. The Gaussian limit forces the intuition that only a small number of elements will contribute effectively, while for large values of n the corresponding Dirichlet variable is always so small that it has negligible influence in the Lipschitz constant. The description of the variable bounds in Equation (106) and the sum in Equation (107) prevent a straightforward analysis in terms of the relevant quantity $\sigma_g\sqrt{R}$. We can however make a numerical analysis, depicted in Figure 8(b). This calculation shows that the sum in Equation (107) follows a polynomial trend in $\sigma_g\sqrt{R}$, which is the variance of the resulting Gaussian spectrum. Therefore, we can update our previous version of Hoeffding's inequality to

$$\text{Prob}_{\boldsymbol{\Theta}}\left(\Lambda(h_{\boldsymbol{\theta}}) - \|H\|_\lambda \sqrt{2L}\sigma_g \geq t\right) \in \mathcal{O}\left(\exp\left(\frac{-t^2}{(\sigma_g\sqrt{L})^3}\right)\right). \tag{108}$$

A.9 Extension to non-harmonic spectrum

The non-harmonic extension leads to an average of the elements in the trigonometric polynomial given by

$$\mathbb{E}_{\boldsymbol{\Theta}}\left(|a_{\vec{k}}|^2\right) = \frac{1}{\sqrt{(4\pi L)^D|\Sigma|}} \exp\left(-\frac{\vec{k}^T\Sigma^{-1}\vec{k}}{4L}\right), \tag{109}$$

where $|\Sigma|$ is the determinant of Σ. We compute now $\Lambda(h_{\boldsymbol{\theta}})$ following the steps from Appendix A.7, given by

$$\mathbb{E}_{\boldsymbol{\Theta}}\left(\Lambda(h_{\boldsymbol{\theta}})\right) = \sum_{\vec{k}} |\vec{\mu}\cdot\vec{k}|\mathbb{E}_{\boldsymbol{\Theta}}\left(|a_{\vec{k}}|\right) \leq \frac{2}{\sqrt{(4\pi L)^D|\Sigma|}} \int_{\mathbb{R}^D} d\vec{k}|\vec{\mu}\cdot\vec{k}| \exp\left(-\frac{\vec{k}^T\Sigma^{-1}\vec{k}}{4L}\right). \tag{110}$$

Notice that $d\vec{k}$ integrates over D-dimensional space. We perform now a change the variables to diagonalize $\Sigma = U^\dagger SU$, and consequently choose $\vec{l} = U\vec{k}$. The diagonal elements of S are denoted $\{s_j^2\}_j$ The quantity of interest is now $\vec{\mu}\cdot(U^\dagger\vec{l}) = (U\vec{\mu})\cdot\vec{l}$. Since U is unitary $d\vec{l} = d\vec{k}$

$$\mathbb{E}_{\boldsymbol{\Theta}}\left(\Lambda(h_{\boldsymbol{\theta}})\right) \leq \frac{2}{\sqrt{(4\pi L)^D|\Sigma|}} \int_{\mathbb{R}^D} d\vec{l}|(U\vec{\mu})\cdot\vec{l}| \exp\left(-\frac{\vec{l}^T S^{-1}\vec{l}}{4L}\right) \leq \tag{111}$$

$$\frac{2}{\sqrt{(4\pi L)^D|\Sigma|}} \sum_{j=1}^{D} |(U\vec{\mu})_j| \int_{\mathbb{R}^D} d\vec{l}|l_j| \exp\left(-\frac{\vec{l}^T S^{-1}\vec{l}}{4L}\right) \tag{112}$$

We focus now on the integral.

$$\int_{\mathbb{R}^D} d\vec{l}|l_j| \exp\left(-\frac{\vec{l}^T S^{-1}\vec{l}}{4L}\right) = \int_{\mathbb{R}} dl_j|l_j| \exp\left(-\frac{\vec{l}_j^2}{4Ls_j^2}\right) \prod_{i\neq j} \int_{\mathbb{R}} dl_i \exp\left(-\frac{\vec{l}_i^2}{4Ls_i^2}\right) \tag{113}$$

$$= 4Ls_j^2 \prod_{i\neq j} \sqrt{4\pi L s_i^2} = \frac{1}{\pi}\sqrt{(4\pi L)^{D+1}}\sqrt{|\Sigma|}s_j \tag{114}$$

Plugging this result into Equation (112) we obtain

$$\mathbb{E}_{\boldsymbol{\Theta}}\left(\Lambda(h_{\boldsymbol{\theta}})\right) \leq \frac{(2\sqrt{L\pi})^{D+1}}{\pi} \frac{\sqrt{|\Sigma|}|\vec{\mu}U^\dagger|\cdot\sqrt{\vec{S}}}{\sqrt{(4\pi L)^D|\Sigma|}} = \frac{4\sqrt{L}}{\sqrt{\pi}} \sum_{j=1}^{D} |U\vec{\mu}|_j s_j \tag{115}$$

By means of Cauchy-Schwarz inequality, we can give a looser yet more comprehensive bound as

$$\mathbb{E}_{\boldsymbol{\Theta}}\left(\Lambda(h_{\boldsymbol{\theta}})\right) \leq \frac{4}{\sqrt{\pi}}\|\vec{\mu}\|_2\sqrt{\text{Tr}(\Sigma)}\sqrt{L}. \tag{116}$$

For the lower bound we follow Appendix A.7 to obtain

$$\mathbb{E}_{\boldsymbol{\Theta}}\left(\Lambda(h_{\boldsymbol{\theta}})\right)^2 \geq \|H\|_\lambda^2 \sum_{k=-K}^{K} (\vec{\mu}\cdot\vec{k})^2 \mathbb{E}_{\boldsymbol{\Theta}}\left(|a_k|^2\right). \tag{117}$$

We recall the property [65, 66]

$$\int d^D\vec{k} f(\vec{k}) \exp\left(-\frac{\vec{k}^T\Sigma^{-1}k}{2}\right) = \sqrt{(2\pi)^D|\Sigma|}\exp\left(\frac{\vec{\nabla}^T\Sigma^{-1}\vec{\nabla}}{2}\right) f(\vec{k})\Big|_{\vec{k}=0}, \tag{118}$$

where $\vec{\nabla}_j = \partial/\partial\vec{k}_j$. Since $f(\vec{k}) = (\vec{\mu}\cdot\vec{k})^2$, we can reduce

$$\exp\left(\frac{\vec{\nabla}^T\Sigma^{-1}\vec{\nabla}}{2}\right) f(\vec{k})\Big|_{\vec{k}=0} = \vec{\mu}^T\Sigma\vec{\mu} \geq \|\vec{\mu}\|_2^2 \min_\Lambda(\Sigma), \tag{119}$$

yielding a result

$$\Lambda^2(h_{\boldsymbol{\theta}}) \geq \|H\|_\lambda^2 2L\|\mu\|_2^2 \min_\lambda(\Sigma). \tag{120}$$

□

A.9.1 A simple example

We illustrate the spectral convolution with an example. Consider a data generator whose spectrum and multiplicities are

$$\lambda = \{-\sqrt{2}-1, -\sqrt{2}, -1, 0, +1, +\sqrt{2}, +\sqrt{2}+1\} \tag{121}$$

$$m(\lambda) = \{1, 1, 1, 2, 1, 1, 1\} \tag{122}$$

Any frequency resulting from the L-fold application of such data generator can be written as $\lambda_{k,l} = k\sqrt{2} + l$ where $-L \leq k, l \leq L$ are integers. The corresponding frequency content can therefore be represented as a two-dimensional tensor A. The elements of A follow a 2-dimensional Dirichlet distribution, in the sense of Theorem 4.1, given by the convoluted kernel

$$\mathcal{K}_g^{*L} = \frac{1}{8}\begin{bmatrix} 1 & 1 & 0 \\ 1 & 2 & 1 \\ 0 & 1 & 1 \end{bmatrix}^{*L} \tag{123}$$

In the limit of large L, the central limit theorem applies exactly in the same way as in the harmonic case, and the L-fold convolution tends towards a multivariate Gaussian kernel with $[0, 0]$ mean and covariance matrix

$$\Sigma = \frac{L}{2}\begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}. \tag{124}$$