Approximation and Generalization Capacities of Parametrized Quantum Circuits for Functions in Sobolev Spaces

Alberto Manzano[1], David Dechant[2,3], Jordi Tura[2,3], and Vedran Dunjko[2,4]

[1]Department of Mathematics and CITIC, Universidade da Coruña, Campus de Elviña s/n, A Coruña, Spain
[2]$\langle aQa^L \rangle$ Applied Quantum Algorithms Leiden, The Netherlands
[3]Instituut-Lorentz, Universiteit Leiden, P.O. Box 9506, 2300 RA Leiden, The Netherlands
[4]LIACS, Universiteit Leiden, P.O. Box 9512, 2300 RA Leiden, Netherlands

Parametrized quantum circuits (PQC) are quantum circuits which consist of both fixed and parametrized gates. In recent approaches to quantum machine learning (QML), PQCs are essentially ubiquitous and play the role analogous to classical neural networks. They are used to learn various types of data, with an underlying expectation that if the PQC is made sufficiently deep, and the data plentiful, the generalization error will vanish, and the model will capture the essential features of the distribution. While there exist results proving the approximability of square-integrable functions by PQCs under the L^2 distance, the approximation for other function spaces and under other distances has been less explored. In this work we show that PQCs can approximate the space of continuous functions, p-integrable functions and the H^k Sobolev spaces under specific distances. Moreover, we develop generalization bounds that connect different function spaces and distances. These results provide a theoretical basis for different applications of PQCs, for example for solving differential equations. Furthermore, they provide us with new insight on the role of the data normalization in PQCs and of loss functions which better suit the specific needs of the users.

Alberto Manzano: alberto.manzano.herrero@udc.es

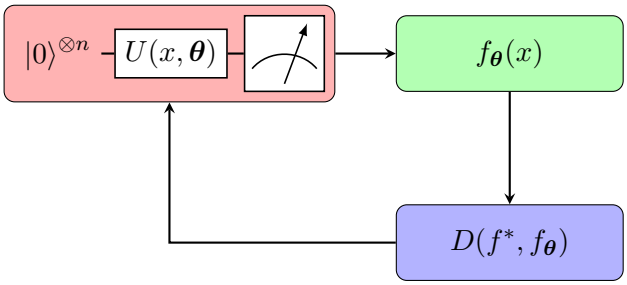

Figure 1: Sketch of a hybrid variational algorithm. $U(x, \theta)$ represents a quantum circuit that takes x as input and with variational parameters θ, $f_\theta(x)$ is the expected value of some observable and $D(f^*, f_\theta)$ is the expected loss that we want to minimize.

1 Introduction

Machine learning has gained significant attention in recent years for its practical applications and transformative impact in various fields. As a consequence, there has been a rising interest in exploring the use of quantum circuits as machine learning models, capitalizing on the advancements in both fields to unlock new possibilities and potential breakthroughs. Among the various possibilities for leveraging quantum circuits in machine learning, our particular focus lies on parametrized quantum circuits (PQC). These quantum circuits consist of both fixed and adjustable (hence 'parametrized') gates. When used for a learning task such as learning a function [14], a classical optimizer updates the parameters of the PQC in order to minimize a cost function depending on measurement results from this quantum circuit (see Figure 1).

In this context, a growing line of research studies the expressivity of PQCs. More precisely, the capacity of PQCs to approximate any function belonging to a particular *function space* defined in a prescribed *domain* up to arbitrary precision with respect to a specific *distance*. In [21], they showed that PQCs can be written as a generalized trigonometric series in the following way:

$$\begin{aligned} f_{\boldsymbol{\theta}}(\boldsymbol{x}) &= \langle 0| U^\dagger(\boldsymbol{x};\boldsymbol{\theta}) M U(\boldsymbol{x};\boldsymbol{\theta}) |0\rangle \\ &= \sum_{\boldsymbol{\omega}\in\Omega} c_{\boldsymbol{\omega}}(\theta) e^{i\boldsymbol{\omega}\boldsymbol{x}} \ . \end{aligned} \tag{1}$$

We would like to emphasize that although similar, the form of the PQC in above equation is more general than a Fourier series. This will become relevant for the results of this work. Using this formulation, it was further shown in [21] that, if the PQC is chosen carefully, the increase of its depth and number of parameter can arbitrarily reduce the L^2 distance between the expected value $f_{\boldsymbol{\theta}}(\boldsymbol{x})$ of the PQC and any square-integrable function with the domain $[0, 2\pi]^N$. Throughout the paper we will refer to the PQC as the one approximating the functions to make the text more fluent, although technically it is the expectation value of the PQC that approximates the function.
This result had a significant impact on the motivation to study PQC-based QML, analogous to the impact that the famous Universality theorem for neural networks of Cybenko [6] had on the domain of classical machine learning. Previously, different results on universality for PQC have been established. In [7], the power of PQCs in expressing matrix product states and instantaneous quantum polynomial circuits was shown. Later, the universal approximation of PQCs was studied in regression problems, for single-qubit circuits with multiple layers [17, 24] and for both single- and multiple-qubit circuits [5, 11]. However, as it turns out, there are numerous different notions of universality, and not all are useful for all applications. For instance, as will be discussed later, in the context of Physics-Inspired Neural Networks (PINN) the "vanilla" universality does not suffice. This raises the question of whether PQCs can approximate functions belonging to other function spaces or in terms of other distances.

In this paper, we present two novel results. The first result of this paper is that PQCs can arbitrarily approximate the space of continuous functions, the space of p-integrable functions and the H^k space, which is the set of functions whose derivatives up to order k are square integrable. Furthermore, we explain how these properties can be easily achieved in practice by a simple min-max feature rescaling (see (20)) of the input data. In practice, this leads to an improved expressivity of PQCs, if the input data is normalized accordingly.

The second result of the paper are generalization bounds that connect distances with loss functions which are not built via the discretization of the integrals present in the definition of the distance. To make it more clear, we recall that in a machine learning problem one needs to choose an architecture, which defines the class of functions that can be approximated, and a target distance, which is intimately connected with the generalization error[1]. However, in general it is not possible to compute the target distance, as we would need to have available infinitely many data points. Instead, one chooses a different distance function which can be computed from the available data: a loss function. This loss function is a different function than the target distance but it should be chosen in such a way that we call *consistent* with the target distance, i.e., that the minimization of the expectation value of the loss function, the expected loss, yields the minimization of the target distance up to an error which asymptotically tends to zero *when the number of samples and the expressivity* (here meant architecturally, as e.g. depth) of the PQC increases. For example, the mean square error (as a loss function) is consistent with the L^2 error (as a target distance) but is inconsistent with the supremum distance. The usual generalization bounds connect target distances which are continuous with expected losses which are their discrete version.
The generalization bounds that we derive give a mapping *across* different distances and loss functions, i.e., they relate distances with loss functions which are not built via the discretization of the integrals present in the definition

[1] In practice we may not explicitly think about the target distance, i.e. with respect to which distance we wish to approximate the "true" labeling function. But this decision is implicitly made, once the loss is chosen.

of the distance. A particular loss function we shall define, denoted ℓ_{h^1}, which consists of the sum of the mean square errors of the values of the functions and its derivative, is consistent with the supremum distance in one-dimensional problems. In the described case, this allows us to reduce the supremum distance while choosing a loss function which is differentiable.

Our results apply in many settings. For example, our first result has a direct consequence in that it allows one to approximate not most, but all function values with satisfying quality. For instance, the minimization of the ubiquitous L^2 distance may allow functions to dramatically differ from the target function in some regions where we have plenty of data points available, whereas the minimization of the supremum norm in Theorem 3 will force the PQC to converge for any given point in the domain of the target function. This is of high relevance in cases where we are interested in having a good approximation at any given point. For instance, when learning the shape of a probability distribution from samples, a good fit in the bulk of the distribution but not in its tails can lead to significant underestimation or overestimation of the probability of extreme events. In real-world applications, this could have severe consequences in risk assessment applications, where accurate estimation of tail probabilities is essential for developing appropriate contingency measures against rare but significant events, such as the COVID-19 pandemic or the 2008 economic crisis. Our second result has direct applications, e.g., in settings where we have access to data of the function and its derivatives. One case where this is standard is in settings involving solving differential equations. For example in physics-informed neural networks (PINN) problems [18] and differential machine learning (DML) [13], both function values and derivatives are accessible and in fact critical.

This paper is organized as follows: in Section 2 we explain the new results on the expressivity of PQCs. In Section 3 we discuss the proposed generalization bounds. Then, in Section 4 we illustrate the theoretical result of Sections 2 and 3 by means of some numerical experiments. Lastly, in Section 5 we wrap up with the conclusions.

During the final stages of our work, we became aware of the paper [10] which overlaps in some parts with our own results in Section 2. However, the results presented here were developed independently and follow a different line of reasoning.

2 PQCs and universal approximation

In this section, we will review the established result on universality in [21] and then present our new universality results in Theorems 2, 3 and 4.

Schuld et al. showed in [21], how a quantum machine learning model of the form $f_\theta(x) = \langle 0| U^\dagger(x;\boldsymbol{\theta}) M U(x;\boldsymbol{\theta}) |0\rangle$ can be written as a univariate generalized trigonometric series:

$$\langle 0| U^\dagger(x;\boldsymbol{\theta}) M U(x;\boldsymbol{\theta}) |0\rangle = f_m(x;\boldsymbol{\theta}) \tag{2}$$

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

where M is an observable, $U(x;\boldsymbol{\theta})$ is a quantum circuit modeled as a unitary that depends on input x and the variational parameters $\boldsymbol{\theta} = (\theta_0, \theta_1, ..., \theta_T)$. In the above, $\boldsymbol{\omega} \in \Omega$ denotes the set of available frequencies which always contain 0. The quantum circuit consists of L layers each consisting of a trainable circuit block $W_i(\boldsymbol{\theta}), i \in \{1, ..., L+1\}$ and a data encoding block $S(x)$ as shown in Figure 2. The data encoding blocks determine which frequencies $\boldsymbol{\omega}$ are accessible in the sum and are implemented as Pauli rotations. The blocks $W(\boldsymbol{\theta})$ can be built from single-qubit rotation gates and CNOT gates and they determine the coefficients $c_{\boldsymbol{\omega}}(\theta)$ of the sum. It is possible to both implement this model with $L > 1$ layers, such as data re-uploading PQC [9, 16], where the encoding is repeated on the same subsystems in sequence, or with parallel encodings [19] and $L = 1$, where the encoding is repeated on several different subsystems.

For the needs of our discussion, we will briefly describe a more specific set-up under which the authors of [21] proved a universality theorem of these quantum models for the multivariate case with inputs $\boldsymbol{x} = (x_0, x_1, ..., x_N)$.

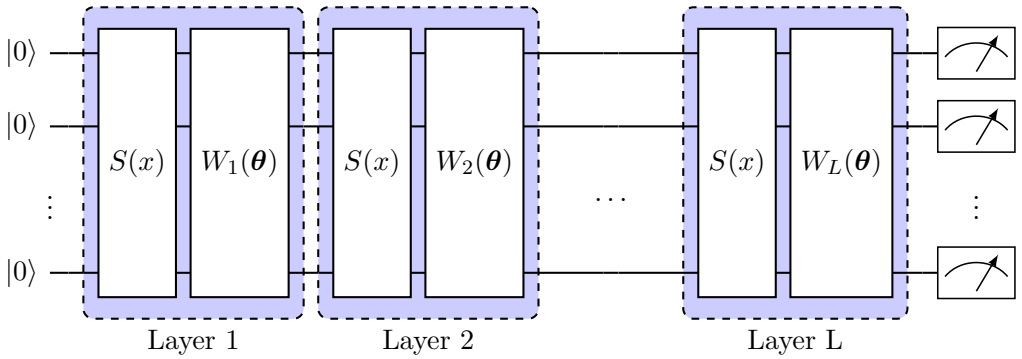

Figure 2: parametrized quantum circuit that can be written as a generalized trigonometric series as in (1). It consists of L layers, each layer is composed by a trainable circuit block $W_i(\boldsymbol{\theta}), i \in \{1, ..., L+1\}$ and a data encoding block $S(x)$. The data encoding blocks $S(x)$ are identical for all layers, they determine which frequencies $\boldsymbol{\omega}$ are accessible and are implemented as Pauli rotations. The blocks $W_i(\boldsymbol{\theta})$ can be built from local rotation gates and $CNOT$ gates. They determine the coefficients $c_{\boldsymbol{\omega}}(\theta)$.

Let us construct a model of the form in (1), with the measurement M and a quantum circuit of one layer, $L = 1$:

$$f_{\boldsymbol{\theta}} = \langle 0| U^\dagger(\boldsymbol{\theta}, \boldsymbol{x}) M U(\boldsymbol{\theta}, \boldsymbol{x}) |0\rangle \ , \text{ with} \quad (4)$$

$$U(\boldsymbol{\theta}, \boldsymbol{x}) = W^{(2)}(\boldsymbol{\theta}^{(2)}) S(\boldsymbol{x}) W^{(1)}(\boldsymbol{\theta}^{(1)}) \ , \quad (5)$$

where $\boldsymbol{\theta}^{(1)}$ and $\boldsymbol{\theta}^{(2)}$ are those parameters in $\boldsymbol{\theta}$ that affect $W^{(1)}$ and $W^{(2)}$, respectively. Let us further make the following two assumptions: Firstly, we assume that the data-encoding blocks $S(\boldsymbol{x})$ are written in the following way:

$$S(\boldsymbol{x}) = e^{-x_1 H} \otimes \cdots \otimes e^{-x_N H} \quad (6)$$

$$=: S_H(\boldsymbol{x}) \ , \quad (7)$$

where H is a Hamiltonian that we specify later. Secondly, we assume that the trainable circuit blocks $W^{(1)}(\boldsymbol{\theta}^{(1)})$ and $W^{(2)}(\boldsymbol{\theta}^{(2)})$ are able to represent arbitrary global unitaries. In practice, this may require exponential circuit depth. With this assumption, we drop the dependence on $\boldsymbol{\theta}$ and reformulate the assumption as being able to prepare an arbitrary initial state $|\Gamma\rangle := W^{(1)}(\boldsymbol{\theta}^{(1)}) |0\rangle$ and by absorbing $W^{(2)}(\boldsymbol{\theta}^{(2)})$ into the measurement M. We can then write the above quantum model as:

$$f(\boldsymbol{x}) = \langle \Gamma| S_H^\dagger(\boldsymbol{x}) M S_H(\boldsymbol{x}) |\Gamma\rangle \ . \quad (8)$$

Let us further present the notion of a universal Hamiltonian family, as defined in [21]:

Definition 1. *Let $\{H_m | m \in \mathbb{N}\}$ be a Hamiltonian family where H_m acts on m subsystems of dimension d.*
Such a Hamiltonian family gives rise to a family of models $\{f_m\}$ in the following way:

$$f_m(\boldsymbol{x}) = \langle \Gamma| S_{H_m}^\dagger(\boldsymbol{x}) M S_{H_m}(\boldsymbol{x}) |\Gamma\rangle \ . \quad (9)$$

Further, we call the set

$$\Omega_{H_m} := \{\lambda_j - \lambda_k | j, k \in \{1, ..., d^m\}\} \quad (10)$$

where $\{\lambda_1, ..., \lambda_{d^m}\}$ are the eigenvalues of H_m, the frequency spectrum of H_m.

Remark. *We call a Hamiltonian family $\{H_m\}$ a universal Hamiltonian family, if for all $K \in \mathbb{N}$, there exists an $m \in \mathbb{N}$, such that:*

$$\mathbb{Z}_K = \{-K, ..., 0, ..., K\} \subseteq \Omega_{H_m} \ , \quad (11)$$

hence if the frequency spectrum of $\{H_m\}$ asymptotically contains any integer frequency.

As shown in [21], a simple example of a universal Hamiltonian family is one which consists of tensor products of single-qubit Pauli gates:

$$H_m = \sum_{i=1}^{m} \sigma_q^{(i)} \ , \quad (12)$$

with $\sigma_q^{(i)}$, $q \in \{X, Y, Z\}$ and $d = 2$. The scaling of the frequency spectrum for this example goes as $K = m$.
With these definitions, we can give the following theorem:

Theorem 1 (Convergence in L^2)**.** *[21] Let $\{H_m\}$ be a universal Hamiltonian family, and $\{f_m\}$ the associated quantum model family, defined via (9). For all functions $f^* \in L^2\left([0,2\pi]^N\right)$, and for all $\epsilon > 0$, there exists some $m' \in \mathbb{N}$, some state $|\Gamma\rangle \in \mathbb{C}^{m'}$ and some observable M such that*

$$\|f_{m'} - f^*\|_{L^2} < \epsilon. \tag{13}$$

Here, we clearly see that there are two conditions on the target function f^* that must be fulfilled in order for the theorem to work properly. The first condition is that f^* belongs to L^2. This is not surprising, we need to assume certain regularity on the target function to make the theorem work. The second condition is that the target function f^* needs to be restricted to the domain $[0,2\pi]^N$. However, as suggested in the original paper [21], if the function f^* does not belong to this domain, we can easily map $[a,b]^N$ to the required domain $[0,2\pi]^N$ (or $[-\pi,\pi]^N$ equivalently).

We would like to highlight the fact that the distance we use to bring the approximator closer to the target function is the L^2 distance. Note that convergence in the L^2 sense does not imply other modes of convergence. For example, this does not give us information about the general case of L^p-distances, with $1 \leq p < \infty$. We explicitly address this more general case in the following theorem:

Theorem 2 (Convergence in L^p)**.** *Let $\{H_m\}$ be a universal Hamiltonian family, and $\{f_m\}$ the associated quantum model family, defined via (1). For all functions $f^* \in L^p\left([0,2\pi]^N\right)$ where $1 \leq p < \infty$, and for all $\epsilon > 0$, there exists some $m' \in \mathbb{N}$, some state $|\Gamma\rangle \in \mathbb{C}^{m'}$, and some observable M such that:*

$$\|f_{m'} - f^*\|_{L^p} < \epsilon. \tag{14}$$

The proof of Theorem 2 is given in Appendix A.

Let us emphasize the difference between Theorems 1 and 2: The target function can belong to any L^p space with $1 \leq p < \infty$ in contrast to the previous requirement of being square-integrable (L^2). This is essentially achieved by the fact that PQCs are not only able to represent Fourier series as it is discussed in [21] but they are also able to represent more general trigonometric series. This allow us to identify the expectation value of the quantum circuit with the Cèsaro summation of the partial Fourier series of f^* and leverage the power of Fejér-like theorems [8]. See Appendix A for more details.

Nevertheless, the ability to approximate functions in L^p does not prevent us from having arbitrarily big errors in certain points. Intimately related to this problem is the so-called Gibbs phenomenon [12]. Namely, the approximation of a continuous, but non-periodic function by a Fourier series is increasingly better in the interior of the domain but increasingly poorer on its boundaries. That leads to the fundamental question if we can approximate f^* in a stronger sense, so that we ensure that the target function f^* is well approximated in any given point. We answer this question in the next theorem.

Theorem 3 (Convergence in C^0)**.** *Let $\{H_m\}$ be a universal Hamiltonian family, and $\{f_m\}$ the associated quantum model family, defined via (1). For all functions $f^* \in C^0(U)$ where U is compactly contained in the closed cube $[0,2\pi]^N$, and for all $\epsilon > 0$, there exists some $m' \in \mathbb{N}$, some state $|\Gamma\rangle \in \mathbb{C}^{m'}$, and some observable M such that $f_{m'}$ converges uniformly to f^*:*

$$\|f_{m'} - f^*\|_{C^0} < \epsilon, \tag{15}$$

with [2]

$$\|f_{m'} - f^*\|_{C^0} := \sup_{\mathbf{x}\in[0,2\pi]^N} \|f_{m'}(\mathbf{x}) - f^*(\mathbf{x})\| \ . \tag{16}$$

The proof of Theorem 3 can be found in Appendix A.

A set $U \subset \mathbb{R}^N$ is compactly contained in another set $V \subset \mathbb{R}^N$, if the closure of U is compact and contained in the interior of V.

Simply stated, this theorem means that $f_{m'}$ converges uniformly to f^*. In other words, if we select a given target error ϵ we are always able to find a finite PQC such that the error on any point is smaller than the prescribed ϵ. Let us emphasize again the differences between Theorems 1 and 3. The first difference is that the function f^* has to be defined in a domain U which is compactly contained in $[0,2\pi]^N$. A simple example of U is the interval $\left([-\frac{\pi}{2},\frac{\pi}{2}]^N\right)$

[2] Since f^* is defined on a compact domain U, the supremum is equivalent to the maximum in this case.

(or $\left([0,\pi]^N\right)$, equivalently). By restricting ourselves to half of the original space we can always find a C^0 extension of the function f^* in $\mathbb{T}^N$. The second difference is that the target function now belongs to the class of continuous functions in contrast to the previous requirement of being square-integrable (L^2).

A last result that we will show in this regard is about the approximation of the function and its derivatives by the parametrized quantum circuit. This might seem as a purely synthetic question but it has many implications in practice. When we approximate a target function, in many occasions we not only want to recover its value but also its dynamics. This is particularly relevant for problems in physics, where we typically have a differential equation which describes the behavior of the system. As we will see in the following theorem, the universality results translate to functions defined in the Sobolev space H^k as well:

Definition 2. *The Sobolev space $H^k(\Omega)$ is defined as the space of square integrable functions on a domain $\Omega \subseteq \mathbb{R}^N$ which derivatives up to order k are square integrable as well:*

$$f^\alpha = D^\alpha f \ , \ \textit{and} \ \left\| f^{(\alpha)} \right\|_2 < \infty \ , \tag{17}$$

for all $0 \leq |\alpha| \leq k$ and $D^\alpha := \frac{\partial^{|\alpha|}}{\partial x_1^{\alpha_1} ... \partial x_N^{\alpha_N}}$. The Sobolev norm $\|\cdot\|_{H^k}$ is defined as

$$\|f\|_{H^k} := \left(\sum_{|\alpha| \leq k} \int_\Omega |D^\alpha f|^2 \right)^{1/2} . \tag{18}$$

Theorem 4 (Convergence in H^k)**.** *Let $\{H_m\}$ be a universal Hamiltonian family, and $\{f_m\}$ the associated quantum model family, defined via (1). For all functions $f^* \in H^k(U)$ where U is compactly contained in the closed cube $[0, 2\pi]^N$, and for all $\epsilon > 0$, there exists some $m' \in \mathbb{N}$, some state $|\Gamma\rangle \in \mathbb{C}^{m'}$, and some observable M such that $f_{m'}$ converges to f^* with respect to the H^k-distance:*

$$\|f_{m'} - f^*\|_{H^k} < \epsilon. \tag{19}$$

The proof of Theorem 4 is given in Appendix A.

As in Theorems 3 and 4, we require that the target function is defined on a compactly contained subset of $[0, 2\pi]^N$, we propose to perform a min-max feature scaling of the input data:

$$\boldsymbol{x} = (x_1, ..., x_n) \longrightarrow \ \tilde{\boldsymbol{x}} = (\tilde{x}_1, ..., \tilde{x}_n) \ , \tag{20}$$

where $\boldsymbol{x} \in [a, b]^N$, $\tilde{\boldsymbol{x}} \in \left[-\frac{\pi}{2}, \frac{\pi}{2}\right]^N$, and

$$\tilde{\boldsymbol{x}} = \left(-\frac{\pi}{2} + \pi \frac{x_1 - a}{b - a}, ..., -\frac{\pi}{2} + \pi \frac{x_n - a}{b - a} \right) . \tag{21}$$

This simple recipe allows the PQC to approximate a much wider set of function spaces as shown throughout this section. This normalization strategy works very well in practice as can be seen in Section 4. However, we would like to emphasize that this particular normalization is not the only choice. The classical strategy in machine learning of normalizing the input data to lie in the $[-1, 1]^N$ domain is also completely valid.
Throughout this section, we have discussed the expressive power of PQCs, but when we do machine learning, we have more ingredients that we need to take into account. In the next section we will discuss the role that the loss function plays in accordance with the type of approximation that our PQC can get.

3 Connections between different generalization bounds

As we have seen in the previous section, the notion of approximation depends on a prescribed distance. This distance is not given by the problem itself, but instead chosen by the user, this is why we refer to it as target distance. In general, it is however not possible to compute the target distance, which for example is the case for the L^p and H^k distances. This is why one needs to choose a distance function which can be computed from data, a loss function. It has to be chosen in such a way that it is consistent with the target function. To discuss the topic in more depth, let us formally introduce the continuous regression problem, which is the problem that we are most interested in.

In general, we can describe the continuous regression problem in the following way: assume that there is some target function $f^* \in \mathcal{F} \subseteq H^k$ mapping inputs $x \in \mathcal{X}$ to target labels $y \in \mathcal{Y}$.

Moreover, assume that the points in $\mathcal{X}$ are sampled according to a bounded[3] density function p. Our goal is to find the best approximation $f \in \mathcal{M} \subseteq H^k$ of the target function f^*.
The notion of what is understood as a "good" approximation as clarified, allows for some freedom. For this reason, one has to make a choice by specifying a functional $D : H^k \times H^k \longrightarrow \mathbb{R}^+ \cup \{0\}$ which defines a distance between the elements of $\mathcal{F}$ and $\mathcal{M}$. The problem can then be stated as:

$$f = \arg\min_{\hat{f} \in \mathcal{M}} D(f^*, \hat{f}). \tag{22}$$

The most common distance in the literature for continuous regression problems is the one induced by the $L^2(\mathcal{X}, P)$ norm:

$$\begin{aligned} D_{L^2}(f^*, f) &= ||f^* - f||_{L^2} \\ &= \left(\int_{\mathcal{X}} (f^*(\boldsymbol{x}) - f(\boldsymbol{x}))^2 dP \right)^{\frac{1}{2}}. \end{aligned} \tag{23}$$

However, in regression we do not typically have access to the full information (i.e., we cannot compute the integral). It is for this reason that instead we work with the empirical risk minimization problem, which uses the discrete version l^2 of the L^2 distance as a loss function. The difference with the previous setup is that, for the empirical risk minimization problem, we are given a finite training set S of I inputs sampled from the same probability density p, together with their target labels $\{(x_1, y_1), ..., (x_I, y_I)\}$ with $(x, y) \in \mathcal{X} \times \mathcal{Y}$, according to the target function $f^* : \mathcal{X} \to \mathcal{Y}$, $f \in \mathcal{F}$. Now, instead of minimizing a continuous functional, we will minimize a discrete one. We call

$$D_\ell(f^*, f) = \frac{1}{I} \sum_{i=0}^{I-1} \ell\left(f^*(\boldsymbol{x}^i), f(\boldsymbol{x}^i)\right) \tag{24}$$

the expected loss according to a loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$. Similarly to the continuous case, we are concerned with the expected loss of the l^2 distance, which is defined as:

$$D_{l^2}(f^*, f) := \left(\frac{1}{I} \sum_{i=0}^{I-1} \left(f^*(\boldsymbol{x}^i) - f(\boldsymbol{x}^i)\right)^2 \right)^{\frac{1}{2}}, \tag{25}$$

[3]It is possible to have more general density functions. However, we restrict ourselves with this one since it simplifies the analysis.

with $\boldsymbol{x}^i$ denoting the i-th input.

Although we are solving the minimization problem associated with the expected loss defined in (25), in general we are interested in the generalization performance, i.e., the distance in terms of (23). Using generalization bounds [15] we can relate the performance in terms of the distance given by (25) with the distance given by (23). However, these classical results in machine learning do in general not relate the l^2 distance with other distances, like the C^0 distance. In other words, even a solution which, as the model and the number of points grow larger asymptotically makes the D_{L^2} go to zero, does not necessarily make the D_{C^0} distance vanish, which is defined as:

$$D_{C^0}(f^*, f) := \sup_{\mathbf{x} \in \mathcal{X}} |f^*(\mathbf{x}) - f(\mathbf{x})|. \tag{26}$$

In such cases, we could find points where there is an arbitrarily large discrepancy between the solution and the target function.
One possible solution would be to use a different distance than D_{l^2}. For example one could try the discrete form of the D_{C^0} distance:

$$D_{l^\infty} = \max_{i \in \{0,...,I-1\}} \left| f^*(\mathbf{x}^i) - f(\mathbf{x}^i) \right|, \tag{27}$$

but this distance is not differentiable, making the optimization process much harder.

Thus, we identify two desirable features for a distance in order to be able to approximate with the C^0 distance. The first requirement is that the solution of the minimization problem that it defines, tends uniformly to the target function f^* as we increase the number of given points I and we increase the size of our PQC. The second one is that it has to be differentiable in order to make minimization easier.
The solution that we propose here is to use a distance motivated by discretizing the Sobolev distance H^k on a fixed finite training set $\left\{ \left(x_1, f^*(\boldsymbol{x}^1), \{D^\alpha f^*(\boldsymbol{x}^1)\}_{|\alpha| \leq k}\right), ..., \left(x_I, f^*(\boldsymbol{x}^I), \{D^\alpha f^*(\boldsymbol{x}^I)\}_{|\alpha| \leq k}\right) \right\}$, $(x, y) \in \mathcal{X} \times \mathcal{Y}$, according to the target function $f^* : \mathcal{X} \to \mathcal{Y}$, $f \in \mathcal{F}$. The sets $\{D^\alpha f(\boldsymbol{x})\}_{|\alpha| \leq k}$ and $\{D^\alpha f^*(\boldsymbol{x})\}_{|\alpha| \leq k}$ consists of the $M(N, k) := \sum_{\alpha=1}^{k} \binom{\alpha+N-1}{N-1}$ different partial derivatives up to order k evaluated at point $\boldsymbol{x}$. We write N for the number of input

dimensions. Note that for being able to apply this distance, one needs to have access to training data containing the required partial derivatives additionally to the function values.
We show the expected loss of the discretized version of H^1 and H^k, respectively, in the following two equations:

$$D_{h^1}(f^*, f) := \left[\frac{1}{I}\sum_{i=0}^{I-1}\left(f^*(\boldsymbol{x}^i) - f(\boldsymbol{x}^i)\right)^2 + \sum_{j=0}^{N-1}\sum_{i=0}^{I-1}\frac{1}{I}\left(\frac{\partial f^*}{\partial x_j}(\boldsymbol{x}^i) - \frac{\partial f}{\partial x_j}(\boldsymbol{x}^i)\right)^2\right]^{\frac{1}{2}}, \tag{28}$$

$$D_{h^k}(f^*, f) := \left[\frac{1}{I}\sum_{i=0}^{I-1}\left(f^*(\boldsymbol{x}^i) - f(\boldsymbol{x}^i)\right)^2 + \sum_{|\alpha|\leq k}\sum_{i=0}^{I-1}\frac{1}{I}\left(D^\alpha f^*(\boldsymbol{x}^i) - D^\alpha f(\boldsymbol{x}^i)\right)^2\right]^{\frac{1}{2}}. \tag{29}$$

The expected loss as given in (28) was first introduced in [13] and gives rise to a new subfield of machine learning known in the literature as differential machine learning (DML). Its generalization, the discretization of the distance H^k, is given in (29), and can be applied when the required higher-dimensional derivatives are available as well. The derivatives $\frac{\partial^{(p)} f^*}{\partial x_j^{(p)}}$ and $\frac{\partial^{(p)} f}{\partial x_j^{(p)}}$ are the p-th order derivative functions in direction x_j of f^* and f, respectively. The corresponding loss function is thus defined as

$$\ell_{h^k} : \mathbb{R}^{M(N,k)+1} \times \mathbb{R}^{M(N,k)+1} \to \mathbb{R}, \tag{30}$$

$$\left(f(\boldsymbol{x}), \{D^\alpha f(\boldsymbol{x})\}_{|\alpha|\leq k}, f^*(\boldsymbol{x}), \{D^\alpha f^*(\boldsymbol{x}) :\}_{|\alpha|\leq k}\right) \mapsto$$

$$\ell_{h^k}(f(\boldsymbol{x}), f^*(\boldsymbol{x})) = (f^*(\boldsymbol{x}) - f(\boldsymbol{x})))^2 \tag{31}$$

$$+ \sum_{|\alpha|\leq k} (D^\alpha f^*(\boldsymbol{x}) - D^\alpha f(\boldsymbol{x})))^2 \ .$$

With classical neural networks, DML has proven to yield better generalization results in terms of the D_{l^2} distance than the solution of the D_{l^2} itself. This means that, if we take the solutions f_{h^1} and f_{l^2} of the minimization problems defined by Equations (28) with the same number of labels and (25) respectively and evaluate their performance in terms of the D_{L^2}, in practice f_{h^1} performs better than f_{l^2}:

$$D_{L^2}\left(f^*, f_{h^1}\right) \leq D_{L^2}\left(f^*, f_{l^2}\right). \tag{32}$$

However, to the best of our knowledge there is no theoretical explanation in the literature on why this happens or under which condition we might expect this behavior. In the following theorems we present generalization bounds that shed some lights onto it.
Before stating them, we will define two function families to which the generalization bounds apply:

Definition 3. *[4] By $\mathcal{F}_\Omega^B$, we denote the function family defined as*

$$\mathcal{F}_\Omega^B = \Big\{[0, 2\pi]^N \ni \boldsymbol{x} \mapsto f(\boldsymbol{x}) = \sum_{\boldsymbol{\omega}\in\Omega} c_{\boldsymbol{\omega}} \exp(-i\boldsymbol{\omega}\boldsymbol{x}) : \{c_{\boldsymbol{\omega}}\}_{\boldsymbol{\omega}\in\Omega} \text{ s.t. } \|f\|_\infty \leq B \text{ and } |\Omega| < \infty\Big\}.$$

By $\mathcal{H}_\Omega^{\tilde{B}}$, we denote the function family defined as

$$\mathcal{H}_\Omega^{\tilde{B}} = \Big\{[0, 2\pi]^N \ni \boldsymbol{x} \mapsto \frac{a_0}{2} + \sum_{\boldsymbol{\omega}\in\Omega_+} (a_{\boldsymbol{\omega}} \cos(\boldsymbol{\omega}\boldsymbol{x}) + b_{\boldsymbol{\omega}} \sin(\boldsymbol{\omega}\boldsymbol{x})) : \sqrt{a_0^2 + \sum_{\boldsymbol{\omega}\in\Omega_+} a_{\boldsymbol{\omega}}^2 + b_{\boldsymbol{\omega}}^2} \leq \tilde{B} \text{ and } |\Omega_+| < \infty\Big\},$$

where the frequency set Ω is divided into the disjoint parts $\Omega = \Omega_+ \cup \Omega_- \cup \{0\}$, where $\Omega_+ \cap \Omega_- = \emptyset$ and such that for every $\boldsymbol{\omega} \in \Omega_+$, it holds that $-\boldsymbol{\omega} \in \Omega_-$.

According to [4], both of these function families can be modeled by the quantum model given in Equation (9). As can be seen by this equation, the bounds B and $\tilde{B}$ depend on the chosen circuit and observable, and they determine the scaling in the following generalization bounds. Note as well that the truncated Fourier series as defined in $\mathcal{F}_\Omega^B$ and $\mathcal{H}_\Omega^{\tilde{B}}$ are differentiable, and their derivatives form truncated Fourier series as well. If one chooses the frequency set Ω' and the bounds B' and $\tilde{B}'$ large enough, for a given function family

$\mathcal{F}$, both the functions and their derivatives are contained in $\mathcal{F}_{\Omega'}^{B'}$ and $\mathcal{H}_{\Omega'}^{\tilde{B}'}$.

Theorem 5 (Generalization bound for H^k)**.** *Let $f^* \in \mathcal{F} \subseteq H^k([0,2\pi]^N)$ be a target function, and let there be a $B > 0$ and a $\tilde{B} > 0$, such that $\mathcal{F}_\Omega^B \subseteq \mathcal{H}_\Omega^{\tilde{B}}$ is a suitable model family. Let us further assume that $\ell_{h^k}(f_1(\boldsymbol{x}), f_2(\boldsymbol{x})) \leq c$ for all $\mathbf{x} \in [0,2\pi]^N$, and for all $f_1, f_2 \in \mathcal{F}_\Omega^B$ or $\mathcal{F}$. For any $\delta \in (0,1)$ and the empirical risk $D_{h^k}(f^*, f)$ trained on an i.i.d. training data S with size I and containing data of ξ partial derivatives, the following holds for all functions $f \in \mathcal{F}_\Omega^B$ with probability at least $1-\delta$:*

$$D_{H^k}(f^*, f) \leq D_{h^k}(f^*, f) + r(|\Omega|, \xi, B, \tilde{B}, c, I, \delta), \tag{33}$$

where $r(|\Omega|, \xi, B, \tilde{B}, c, I, \delta) \to 0$ as $I \to \infty$.

Theorem 6 (Generalization bound for L^p)**.** *Let $f^* \in \mathcal{F} \subseteq H^k([0,2\pi]^N)$ be a target function, and let there be a $B > 0$ and a $\tilde{B} > 0$, such that $\mathcal{F}_\Omega^B \subseteq \mathcal{H}_\Omega^{\tilde{B}}$ is a suitable model family. Let us further assume that $\ell_{h^k}(f_1(\boldsymbol{x}), f_2(\boldsymbol{x})) \leq c$ for all $\mathbf{x} \in [0,2\pi]^N$, and for all $f_1, f_2 \in \mathcal{F}_\Omega^B$ or $\mathcal{F}$. Assume that $k, p \in \mathbb{N}$ satisfy one of the two following cases:*

1. *$N\left(\frac{1}{2} - \frac{1}{p}\right) < k < N/2$ and $1 \leq p < N$.*
2. *$k \geq N/2$ and $1 \leq p < \infty$.*

For any $\delta \in (0,1)$ and the empirical risk $D_{h^k}(f^*, f)$ trained on an i.i.d. training data S with size I and containing data of ξ partial derivatives, the following holds for all functions $f \in \mathcal{F}_\Omega^B$ with probability at least $1-\delta$:

$$\frac{1}{C} D_{L^p}(f^*, f) \leq D_{h^k}(f^*, f) + r(|\Omega|, \xi, B, \tilde{B}, c, I, \delta), \tag{34}$$

where C is a constant and $r(|\Omega|, \xi, B, \tilde{B}, c, I, \delta) \to 0$ as $I \to \infty$.

Theorem 7 (Generalization bound for C^0)**.** *Let $f^* \in \mathcal{F} \subseteq H^k([0,2\pi]^N)$ be a target function, and let there be a $B > 0$ and a $\tilde{B} > 0$, such that $\mathcal{F}_\Omega^B \subseteq \mathcal{H}_\Omega^{\tilde{B}}$ is a suitable model family. Let us further assume that $\ell_{h^k}(f_1(\boldsymbol{x}), f_2(\boldsymbol{x})) \leq c$ for all $\mathbf{x} \in [0,2\pi]^N$, and for all $f_1, f_2 \in \mathcal{F}_\Omega^B$ or $\mathcal{F}$ and that $\|f\|_\infty \leq B$ for all $f \in \mathcal{F}_\Omega^B$. Assume, that $k \in \mathbb{N}$ satisfies $k > N/2$. For any $\delta \in (0,1)$ and the empirical risk $D_{h^k}(f^*, f)$ trained on an i.i.d. training data S with size I and containing data of ξ partial derivatives, the following holds for all functions $f \in \mathcal{F}_\Omega^B$ with probability at least $1-\delta$:*

$$\frac{1}{C} D_{C^0}(f^*, f) \leq D_{h^k}(f^*, f) + r(|\Omega|, \xi, B, \tilde{B}, c, I, \delta), \tag{35}$$

where C is a constant and $r(|\Omega|, \xi, B, \tilde{B}, c, I, \delta) \to 0$ as $I \to \infty$.

The proofs of Theorems 5, 6 and 7 can be found in Appendix B.

A consequence of Theorem 7 is that, if the order of the derivatives that we have at our disposal are higher than half the number of input dimensions ($k > N/2$), our solution of the D_{h^k} problem is also a solution of the D_{C^0} problem, corresponding to uniform convergence. It means that training with the ℓ_{h^k} loss function (for $k > N/2$), which sums the ℓ_{l^2} losses of function and derivative values, is sufficient for an approximation in C^0. This would not be possible by a training with ℓ_{l^2} loss function and more practical than the training with the ℓ_{l^∞} loss function, as described above.

Note that we face a curse of dimensionality-like phenomenon as the dimension of the input grows. In this case, the number of terms that go into the ℓ_{h^k} loss function grows exponentially with k, as we have to take into account mixed derivatives. Hence, for high dimensional problems the demand on data of partial derivatives is higher and only if they are available, this generalization bound holds.

Further, the requirement of quantum resources for evaluating $D_{h^1}(f^*, f)$ is higher than for the evaluation of $D_{l^2}(f^*, f)$. If we use the parameter shift rule for the evaluation of the derivatives, we need to evaluate $I(1+2N)$ different PQCs. Similar to the demand on training data, this number of PQCs to evaluate $D_{h^k}(f^*, f)$ grows exponentially in k. However, even if the amount of training data is the same (and implying an increase of required PQC evaluations up to a factor of 2), the training with the ℓ_{h^k} loss function shows the promised advantages, as presented in [13].

The last property we wish to highlight is the fact that the generalization bounds connect the empirical risk with the full risk, but they do not give us information of whether they can both tend to zero. In order to tackle that

question we need to combine the results of the theorems present in this section with the ones present in Section 2. For example, if we try to fit a one-dimensional function which is not periodic on $[0, 2\pi]$, using model families $\mathcal{F}_\Omega^B$ and $\mathcal{H}_\Omega^{\tilde{B}}$ and the ℓ_{h^1} loss function, as we increase the number of sample points both sides of the inequality will tend to the same constant but they will not converge to zero. In this regard, observe that the fact that the empirical risk goes to zero is a sufficient but not a necessary condition for the target distance to also tend to zero. Following the same example, if instead of training the model using the ℓ_{h^1} loss function we trained the model using the ℓ_{l^2} loss function, then the L^2 distance will vanish. This idea is illustrated in Figure 5. The bottom line is that more information in the training data does not always equate to a better approximation, if we are not very careful with the necessary data normalization.

4 Numerical experiments

In this section we illustrate the theoretical discussion of Sections 2 and 3 with an illustrative example: the approximation of function $f^* = \frac{x}{2\pi}$, $x \in [-\pi, \pi]$ by the PQC in Figure 3: We conduct two different numerical experiments and show them in Figures 4 and 5. We chose a linear function to show that even in this simple case, the numerical tests fail utterly if the results of Sections 2 and 3 are not applied.

All simulations have been performed using 10 points (10 for the labels plus 10 for the derivative values when they are present) uniformly distributed along the domain for the training phase. Each experiment has been repeated 100 times and we depict the 25, 50 and 75 percentiles in colored solid lines in Figure 4. The legends call the result of the PQCs as $f_\bullet(\cdot)$, where the subscript denotes under which loss function we have done the training and in the parentheses we indicate which normalization we have chosen.

In Figure 4 we compare the performance of our PQC under different normalizations. We normalize the data to lie in the domains $[-\frac{\pi}{2}, \frac{\pi}{2}]$, $[-\pi, \pi]$ and $[-2\pi, 2\pi]$, respectively. When we normalized our data to lie in the range $[-\frac{\pi}{2}, \frac{\pi}{2}]$ we get the best results, as we expected due to Theorem 3. In contrast, when the data is normalized to lie in the range $[-2\pi, 2\pi]$ we obtain very poor approximation results, because in this case, it is not possible to approximate with the C^0-distance or the L^2-distance. The intermediate regime happens when we normalize the data to lie in the range $[-\pi, \pi]$, here we obtain a reasonable approximation except for the boundaries. This is a consequence of approximating with the L^2-distance instead of the C^0-distance: we cannot guarantee that the error will be reduced on any given point. This behavior remains even when we increase the size of the circuit and the number of given points.

In Figure 5, we study the impact of the different loss functions with different normalizations in the learning problem. We simulated the regression using two different loss functions, ℓ_{h^1} and ℓ_{l^2} under two different normalizations, with the domains $[-\frac{\pi}{2}, \frac{\pi}{2}]$ and $[-\pi, \pi]$. The first noticeable phenomenon that we can see is that using the h^1 norm instead of the l^2 norm when the data is normalized to lie in the interval $[-\frac{\pi}{2}, \frac{\pi}{2}]$ not only reduces the variance stemming from repeating the experiments 100 times, but also has some impact on the bias. What might be more surprising is the effect of the h^1 norm when the data is normalized to lie in the interval $[-\pi, \pi]$. Instead of getting a better approximation w.r.t. the l^2 we worsen it. We explain it with the fact that, when we normalize the data to lie in the interval $[-\pi, \pi]$, our PQC is not an approximator of H^1 but it is an approximator of L^2, i.e., it can approximate the function but it cannot simultaneously approximate the function and the derivatives. Thus, in the minimization process the PQC tries to find a balance between the error in the function and the error in the derivatives, worsening the results with respect to the quality of the function approximation.

5 Conclusions

In this work, we have developed a broader theory of approximation capacities of PQCs. We have shown how an appropriate choice of the data normalization greatly improves the expressivity of the PQCs. More specifically, we showed that a min-max feature scaling that normalizes the input data along each dimension

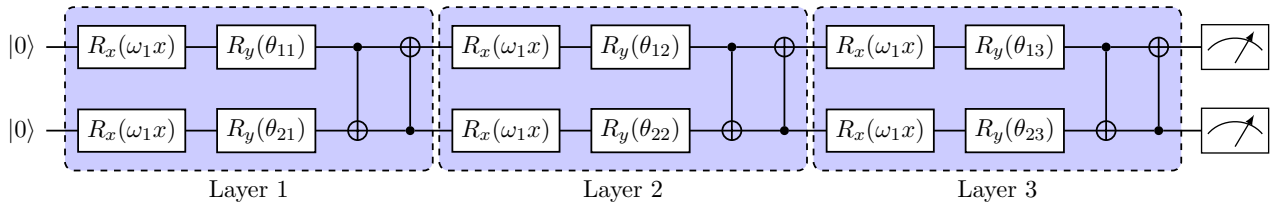

Figure 3: Architecture $U(x, \boldsymbol{\theta})$ used in the experiments. The parameters θ_{ij} are variational parameters. Each qubit is measured in the Pauli-Z basis.

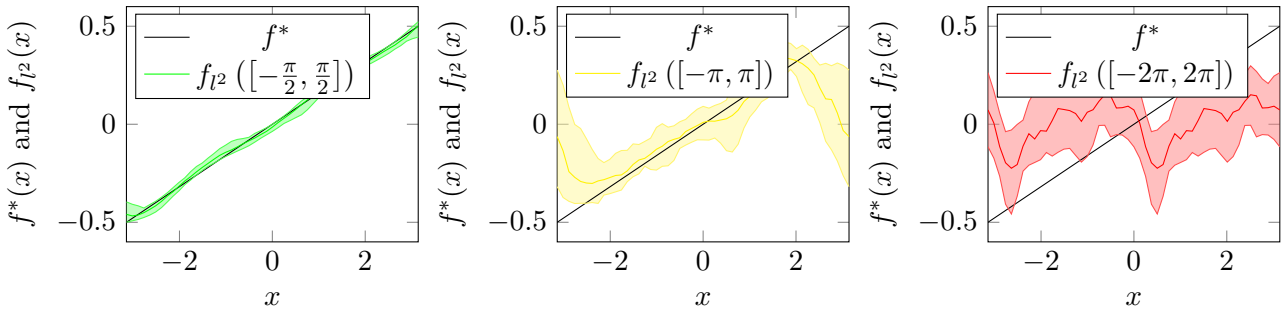

Figure 4: In this picture we have trained the PQC of Figure 3 to approximate the function $f^* = \frac{x}{2\pi}$. We have used 10 training points, the ℓ_{l^2} loss function and 100 epochs with the Adam optimizer. The experiments have been repeated 100 times. In the left panel we have normalized the data to lie in the interval $\left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$. In the central panel we have normalized the data to lie in the interval $[-\pi, \pi]$. In the right panel we have normalized the data to lie in the interval $[-2\pi, 2\pi]$.

to lie in the range $[-\frac{\pi}{2}, \frac{\pi}{2}]$ makes PQCs universal approximators in the L^p space with $1 \leq p < \infty$, the continuous function space and the H^k space. Moreover, since with this normalization we are able to approximate functions in the sense of the L^p, the C^0 and the H^k distance, we discussed that a loss function which is consistent with those distances in training the models might be more appropriate than other choices. In particular, the natural choice for the C^0 would be the l^∞ distance. However, since the l^∞ distance is not differentiable, which makes the optimization of PQCs harder, we leveraged Sobolev inequalities to show that the h^1 distance is consistent with the C^0 distance in $\mathbb{R}$ while being differentiable. We showed further, that the the h^k distances are consistent with the L^p and the H^k distances.
Lastly, we performed some numerical experiments to illustrate how this simple choice of normalization and loss function can vastly improve the results in practice.

The data normalization technique can be seen as a complementary result to the work of [21]. Nevertheless, there is still much work to do in this direction. For example, if instead of only taking a min-max feature scaling, we can combine it with a mapping of the form $\tilde{x} = \arcsin(x)$ to end up with a series that closely resembles Chebyshev polynomials, which are better suited for certain problems. In analogy with neural networks, the data encoding strategy is playing a similar role to that of the activation functions.

The relation between the ℓ_{h^k} loss functions and the L^p generalization bounds can be seen as a complementary result to differential machine learning [13] and to generalization bounds for PQCs as derived in [4]. This is the first work that gives some insight on why differential machine learning leads to better generalization results. From the relations that we derived, one would expect this technique to fail as we increase the input dimension. However, in practice it has demonstrated very good results, as shown in [13], where a 7-dimensional Basket option was trained using the ℓ_{h^1} loss function. An interesting line of research would be to study the threshold at which differential machine learning starts to fail.
Since a natural application are physical systems governed by differential equations where data on the derivatives of a target function are available, another open question remains regarding how our

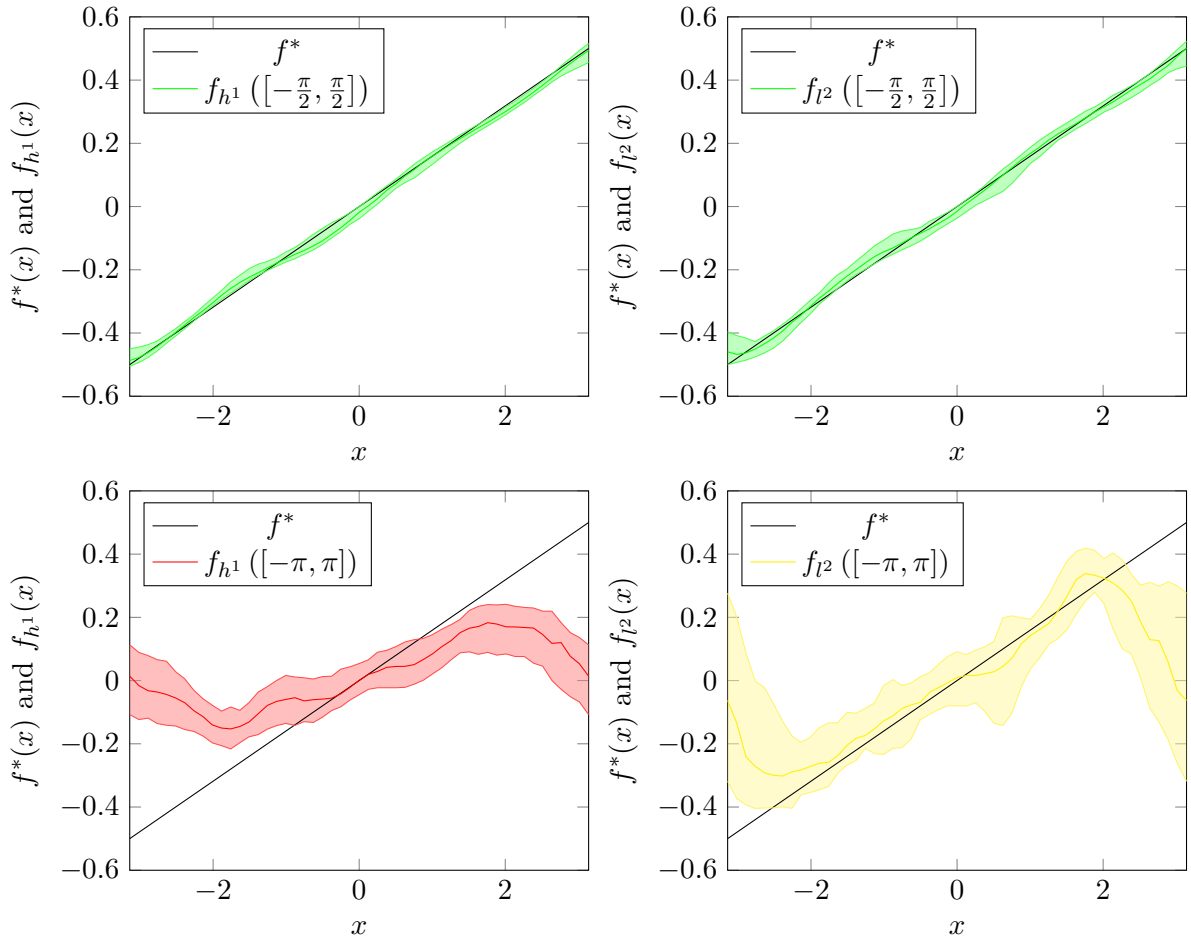

Figure 5: In this picture we have trained the PQC of Figure 3 to approximate the function $f^* = \frac{x}{2\pi}$, using the two different loss functions ℓ_{h^1} and ℓ_{l^2}. We have used 10 training points (10 for the labels plus 10 for the derivative values when they are present) and 100 epochs with the Adam optimizer. The experiments have been repeated 100 times. In the upper panel, we have normalized the data to lie in the interval $\left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$. In the lower panel we have normalized the data to lie in the interval $[-\pi, \pi]$.

approach compares to standard differential equation solvers in these scenarios.

6 Acknowledgments

The authors would like to thank Adrián Pérez Salinas for useful feedback on an earlier version of this paper and Hao Wang and Carlos Vázquez for helpful discussions.
JT, VD and DD acknowledge the support received by the Dutch National Growth Fund (NGF), as part of the Quantum Delta NL programme.
JT acknowledges the support received from the European Union's Horizon Europe research and innovation programme through the ERC StG FINE-TEA-SQUAD (Grant No. 101040729).'
VD and AM acknowledge the support by the project NEASQC funded from the European Union's Horizon 2020 research and innovation programme (grant agree- ment No 951821).
VD acknowledges by the Dutch Research Council (NWO/OCW), as part of the Quantum Software Consortium programme (project number 024.003.037).
AM acknowledges the support received from the Centro de Investigación de Galicia "CITIC", funded by Xunta de Galicia and the European Union (European Regional Development Fund-Galicia 2014-2020 Program), by grant ED431G 2019/01.
The views and opinions expressed here are solely those of the authors and do not necessarily reflect those of the funding institutions. Neither of the funding institution can be held responsible for them.

References

[1] Robert A Adams and John JF Fournier. *Sobolev spaces*. Elsevier, 2003.

[2] Victor Burenkov. Extension theorems for sobolev spaces. In *The Maz'ya Anniversary Collection: Volume 1: On Maz'ya's work in functional analysis, partial differential equations and applications*, pages 187–200. Springer, 1999. DOI: 10.1007/978-3-0348-8675-8_13.

[3] Claudio Canuto and Alfio Quarteroni. Approximation results for orthogonal polynomials in sobolev spaces. *Mathematics of Computation*, 38(157):67–86, 1982. DOI: 10.1090/S0025-5718-1982-0637287-3.

[4] Matthias C. Caro, Elies Gil-Fuster, Johannes Jakob Meyer, Jens Eisert, and Ryan Sweke. Encoding-dependent generalization bounds for parametrized quantum circuits. *Quantum*, 5:582, November 2021. ISSN 2521-327X. DOI: 10.22331/q-2021-11-17-582.

[5] Berta Casas and Alba Cervera-Lierta. Multidimensional fourier series with quantum circuits. *Physical Review A*, 107(6):062612, 2023. DOI: 10.1103/PhysRevA.107.062612.

[6] George Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, 2(4):303–314, 1989. DOI: 10.1007/BF02551274.

[7] Yuxuan Du, Min-Hsiu Hsieh, Tongliang Liu, and Dacheng Tao. Expressive power of parametrized quantum circuits. *Physical Review Research*, 2(3):033125, 2020. DOI: 10.1103/PhysRevResearch.2.033125.

[8] Leopold Fejér. Untersuchungen über fouriersche reihen. *Mathematische Annalen*, 58 (1-2):51–69, 1903.

[9] Francisco Javier Gil Vidal and Dirk Oliver Theis. Input redundancy for parameterized quantum circuits. *Frontiers in Physics*, 8: 297, 2020. DOI: 10.3389/fphy.2020.00297.

[10] Lukas Gonon and Antoine Jacquier. Universal approximation theorem and error bounds for quantum neural networks and quantum reservoirs. *arXiv preprint arXiv:2307.12904*, 2023. DOI: 10.48550/arXiv.2307.12904.

[11] Takahiro Goto, Quoc Hoan Tran, and Kohei Nakajima. Universal approximation property of quantum machine learning models in quantum-enhanced feature spaces. *Physical Review Letters*, 127(9):090506, 2021. DOI: 10.1103/PhysRevLett.127.090506.

[12] David Gottlieb and Chi-Wang Shu. On the gibbs phenomenon and its resolution. *SIAM Review*, 39(4):644–668, 1997. DOI: 10.1137/S0036144596301390.

[13] Brian Huge and Antoine Savine. Differential machine learning. *arXiv preprint arXiv:2005.02347*, 2020. DOI: 10.48550/arXiv.2005.02347.

[14] Kosuke Mitarai, Makoto Negoro, Masahiro Kitagawa, and Keisuke Fujii. Quantum circuit learning. *Physical Review A*, 98

(3):032309, 2018. DOI: 10.1103/PhysRevA.98.032309.

[15] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2018. DOI: 10.1007/s00362-019-01124-9.

[16] Adrián Pérez-Salinas, Alba Cervera-Lierta, Elies Gil-Fuster, and José I Latorre. Data re-uploading for a universal quantum classifier. *Quantum*, 4:226, 2020. DOI: 10.22331/q-2020-02-06-226.

[17] Adrián Pérez-Salinas, David López-Núñez, Artur García-Sáez, and José I Latorre. One qubit as a universal approximant. *Physical Review A*, 104(1):012405, 2021. DOI: 10.1103/PhysRevA.104.012405.

[18] Maziar Raissi, Paris Perdikaris, and George Em Karniadakis. Physics informed deep learning (part i): Data-driven solutions of nonlinear partial differential equations. *arXiv preprint arXiv:1711.10561*, 2017. DOI: https://doi.org/10.48550/arXiv.1711.10561.

[19] Patrick Rebentrost, Masoud Mohseni, and Seth Lloyd. Quantum support vector machine for big data classification. *Physical review letters*, 113(13):130503, 2014. DOI: 10.1103/PhysRevLett.113.130503.

[20] Halsey Lawrence Royden and Patrick Fitzpatrick. *Real analysis*, volume 2. Macmillan New York, 1968.

[21] Maria Schuld, Ryan Sweke, and Johannes Jakob Meyer. Effect of data encoding on the expressive power of variational quantum-machine-learning models. *Physical Review A*, 103(3):032430, 2021. DOI: 10.1103/PhysRevA.103.032430.

[22] Ferenc Weisz. ℓ_1-summability of higher-dimensional fourier series. *Journal of Approximation Theory*, 163(2):99–116, 2011. ISSN 0021-9045. DOI: https://doi.org/10.1016/j.jat.2010.07.011.

[23] Michael M. Wolf. Mathematical foundations of supervised learning. Lecture Notes, 2023.

[24] Zhan Yu, Hongshun Yao, Mujin Li, and Xin Wang. Power and limitations of single-qubit native quantum neural networks. *Advances in Neural Information Processing Systems*, 35:27810–27823, 2022.

A Proof of Theorems 2, 3 and 4

For proving Theorems 2, 3 and 4, we need two preliminary results. Firstly, we need to show that a quantum circuit can realize the ℓ^1-Fejér's mean of $C^0\left(\mathbb{T}^N\right)$ and $L^p\left(\mathbb{T}^N\right), \forall\, 1 \leq p < \infty$ functions. Secondly, we need to prove that we can define periodic extensions of functions belonging to $C^0\left(U\right)$ and $H^k\left(U\right), \forall\, 1 \leq k < \infty$, where U is compactly contained in $\mathbb{T}^N$ to functions belonging to $C^0\left(\mathbb{T}^N\right)$ and $H^k\left(\mathbb{T}^N\right), \forall\, 1 \leq k < \infty$ respectively . The combination of both results plus Fejér's theorem in multiple dimensions naturally yields Theorems 2 and 3. Theorem 4 can be proven by a standard approximation Theorem of the Fourier series.

A.1 Féjer's mean

We call the function

$$\sigma_{NK}(f_{m'}) = \sum_{\mathbf{j}\in\mathbb{Z}_K^N} \left(1 - \frac{\|\mathbf{j}\|_1}{NK}\right) \hat{f}_{\mathbf{j}} e^{i\mathbf{x}\cdot\mathbf{j}} \,, \tag{36}$$

where $\hat{f}_{\mathbf{j}}$ is the $\mathbf{j}$-th Fourier coefficient of $f_{m'}$, the ℓ^1-Fejér's mean of $f_{m'}$.

We will show that our PQC can realize the Fejér's mean of any well-defined function. In Appendix C of [21], the authors showed that the quantum model family $f_{m'}$ can be written as a generalized trigonometric series of the form

$$f_{m'}(\mathbf{x}) = \sum_{\mathbf{j}\in\mathbb{Z}_K^N} c_{\mathbf{j}} e^{i\mathbf{x}\cdot\mathbf{j}} \,, \tag{37}$$

where $\mathbb{Z}_K^N = \{-K, -K+1, ..., 0, ..., K-1, K\}^N$ is contained in the Cartesian product of the frequency spectrum associated with H_m, as defined in Definition 1 and that the coefficients $c_{\mathbf{j}}$ are completely determined by the observable freely up to the complex-conjugation symmetry that guarantees that the model output is a real-valued function. Note that we can choose the coefficients $c_{\mathbf{j}}$ as:

$$c_{\mathbf{j}} = \left(1 - \frac{\|\mathbf{j}\|_1}{NK}\right) \hat{f}_{\mathbf{j}}, \tag{38}$$

which are the coefficients of the ℓ^1-Fejér's mean in Equation (36).

A.2 Periodic extension for C^0 functions

By the Tietze extension theorem [20], there exists a function $g_1 \in C^0(\mathbb{R}^N)$ with $g_1|_U = f^*$. Then, we define a function $g_2 \in C^0(\mathbb{R}^N)$ with $g_2|_{\overline{U}} = 1$ and $g_2|_{\mathbb{R}^N \setminus V} = 0$, where V is defined as $\overline{U} \subset V \subset (0, 2\pi)^N$. This set V exists since U is compactly contained in $[0, 2\pi]^N$.
We can explicitly construct the function g_2 in the following way: Let $\delta > 0$, such that the closure $\overline{\omega_{2\delta}}$ of the 2δ-neighborhood of ω, is contained in $[0, 2\pi]^N$, which is possible due to U being compactly contained in $[0, 2\pi]^N$. We define $V := \omega_{2\delta}$ and a function $\psi_\delta \in C^0(\mathbb{R}^N)$, supported on the δ Ball in $\mathbb{R}^N$ centered around 0 and normalized as $\int_{\mathbb{R}^N} \psi_\delta(x)dx = 1$. Then, we define g_2 as the convolution of $\mathbb{1}_{U_\delta}$ and ψ_δ:

$$g_2(x) = \int_{\mathbb{R}^N} \mathbb{1}_{U_\delta}(\tau)\psi_\delta(\tau - x)d\tau \ . \tag{39}$$

With this construction, g_2 satisfies the asked properties. We define the extension f_{ext} as the product $g_1 g_2$, which yields a function f_{ext}^* with

$$f_{ext}^*|_U = f^* \ , \tag{40}$$
$$f_{ext}^*|_{\mathbb{R}^N \setminus V} = 0 \ , \text{ hence} \tag{41}$$
$$f_{ext}^*(x) = f_{ext}^*(y) \quad \forall x, y \in \partial\mathbb{T}^N \ . \tag{42}$$

The such defined extension f_{ext}^* is thus an element of $C^0([0, 2\pi]^N)$ with periodic boundary conditions, so we can map it onto the N-dimensional torus $\mathbb{T}^N$.

A.3 Periodic extension for H^k functions

By the extension theorems for Sobolev functions [2, Theorem 2.2, Part 2], there exists a function $g_1 \in H^k(\mathbb{R}^N)$ with $g_1|_U = f^*$. Then, we define a function $g_2 \in H^k(\mathbb{R}^N)$ with $g_2|_{\overline{U}} = 1$ and $g_2|_{\mathbb{R}^N \setminus V} = 0$, where V is defined as $\overline{U} \subset V \subset (0, 2\pi)^N$. This set V exists since U is compactly contained in $[0, 2\pi]^N$.
We can explicitly construct the function g_2 in the following way: Let $\delta > 0$, such that the closure $\overline{\omega_{2\delta}}$ of the 2δ-neighborhood of ω, is contained in $[0, 2\pi]^N$, which is possible due to U being compactly contained in $[0, 2\pi]^N$. We define $V := \omega_{2\delta}$ and a function $\psi_\delta \in H^k(\mathbb{R}^N)$, supported on the δ Ball in $\mathbb{R}^N$ centered around 0 and normalized as $\int_{\mathbb{R}^N} \psi_\delta(x)dx = 1$. Then, we define g_2 as the convolution of $\mathbb{1}_{U_\delta}$ and ψ_δ:

$$g_2(x) = \int_{\mathbb{R}^N} \mathbb{1}_{U_\delta}(\tau)\psi_\delta(\tau - x)d\tau \ . \tag{43}$$

With this construction, g_2 satisfies the asked properties. We define the extension f_{ext} as the product $g_1 g_2$, which yields a function f_{ext}^* with

$$f_{ext}^*|_U = f^* \ , \tag{44}$$
$$f_{ext}^*|_{\mathbb{R}^N \setminus V} = 0 \ , \text{ hence} \tag{45}$$
$$f_{ext}^*(x) = f_{ext}^*(y) \quad \forall x, y \in \partial\mathbb{T}^N \ . \tag{46}$$

The such defined extension f_{ext}^* is thus an element of $H^k([0, 2\pi]^N)$ with periodic boundary conditions, so we can map it onto the N-dimensional torus $\mathbb{T}^N$.

A.4 Proof of Theorems 2, 3 and 4

The final step leverages the power of Fejèr's theorem in multiple dimensions:

Theorem 8. *[22][Theorem 2] For all functions $f^* \in L^p\left(\mathbb{T}^N\right)$ with $1 \leq p < \infty$, and for all $\epsilon > 0$, there exists some $t \in \mathbb{N}$, such that*

$$\|\sigma_t(f) - f^*\|_{L^p} < \epsilon. \tag{47}$$

Combining Theorem 8 with the fact that quantum circuits can recover any ℓ^1-Fejér's mean as shown in Appendix A.1 directly implies Theorem 2.

Similarly, for continuous functions we have another version of Fejér's theorem for continuous functions:

Theorem 9. *[22][Theorem 2] For all functions $f^* \in C^0\left(\mathbb{T}^N\right)$, and for all $\epsilon > 0$, there exists some $t \in \mathbb{N}$, such that*

$$\|\sigma_t(f) - f^*\|_{\infty} < \epsilon. \tag{48}$$

Combining Theorem 9 with the fact that quantum circuits can recover any ℓ^1-Fejér's mean as shown in Appendix A.1 and the fact that we can extend any function in $C^0(U)^N, \forall\, 1 \leq p < \infty$ where U is compactly contained in $\mathbb{T}^N$ to a function in $C^0\left(\mathbb{T}^N\right), \forall\, 1 \leq p < \infty$ as shown in Appendix A.2 directly implies Theorem 3.

We finally prove Theorem 4, which uses the setup in [21] as described in section 2: We note firstly that the quantum model family $f_{m'}$ generates a truncated Fourier series $\tilde{f}$ in the domain $[0, 2\pi]^N$ of the form

$$\tilde{f}(\mathbf{x}) = \sum_{\mathbf{j} \in \mathbb{Z}_K^N} c_{\mathbf{j}} e^{i\mathbf{x}\cdot\mathbf{j}}\ , \tag{49}$$

where $\mathbb{Z}_K^N = \{-K, -K+1, ..., 0, ..., K-1, K\}^N$ is contained in the Cartesian product of the frequency spectrum associated with H_m, as defined in Definition 1. The proof of that is written in Appendix C of [21].
Secondly, we can extend the function f^* defined on U to a periodic function f^*_{ext} on $[0, 2\pi]^N$ via the construction shown in Appendix A.3. As written in Theorem 1.1 in [3], the Fourier series of f^*_{ext}, which we can write in the form of equation 49, converges in the H^k-distance to f^*_{ext}. As $f^*_{ext}(x) = f^*(x)$ for all $x \in U$, the Fourier series of f^*_{ext} converges in the H^k-distance to f^* on U. This implies Theorem 4.

B Proof of Theorems 5, 6 and 7

In this appendix, we prove Theorems 5, 6 and 7, for which we need several preliminary definitions and results:

Definition 4 (L-Lipschitz loss function)**.** *Let $(\mathcal{Y}, d_{\mathcal{Y}})$ be a metric space with metric $d_{\mathcal{Y}}$ and let $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ be a loss function. We call it $L-$ Lipschitz with regard to a fixed $y \in \mathcal{Y}$, if there exists a constant $L \geq 0$, such that for all $z_1, z_2 \in \mathbb{R}$,*

$$d_{\mathcal{Y}}(\ell(y, z_1), \ell(y, z_2)) \leq L\,|z_1 - z_2|\ . \tag{50}$$

Theorem 10 (Generalization bound for general trigonometric series)**.** *[4][Theorem 11] Let $N, I \in \mathbb{N}$. Let $B > 0$ and $\tilde{B} > 0$ be such that $\mathcal{F}_\Omega^B \subseteq \mathcal{H}_\Omega^{\tilde{B}}$, for the function families $\mathcal{F}_\Omega^B$ and $\mathcal{H}_\Omega^{\tilde{B}}$ as defined in Definition 3. Let $\ell : \mathbb{R} \times \mathbb{R} \to [0, c]$ be a bounded loss function such that $\mathbb{R} \ni z \mapsto \ell(y, z)$ is L- Lipschitz for all $y \in \mathbb{R}$. For any $\delta \in (0, 1)$ and for any probability measure P on $[0, 2\pi]^N \times \mathbb{R}$, with probability*

at least $1-\delta$ *over the choice of i.i.d. training data* $S \in ([0,2\pi]^N \times \mathbb{R})^I$ *of size* I*, for every* $f \in \mathcal{F}_\Omega^B$*, the generalization error can be upper-bounded as*

$$\int_{[0,2\pi]^N \times \mathbb{R}} \ell(f^*(\boldsymbol{x}), f(\boldsymbol{x})) dP((x), f^*(\boldsymbol{x})) - \frac{1}{|I|} \sum_{\boldsymbol{x}_i, f(\boldsymbol{x}_i) \in S} \ell(f^*(\boldsymbol{x}_i), f(\boldsymbol{x}_i)) \tag{51}$$

$$\leq \mathcal{O}\left(BL\sqrt{\frac{|\Omega|(\log(|\Omega|) + \log(\tilde{B}))}{I}} + c\sqrt{\frac{\log(1/\delta)}{I}} \right) , \tag{52}$$

for a target function $f^* : [0,2\pi]^N \to \mathbb{R}$ *.*

This theorem is written for loss functions that take two real values as an input, which is the case for most loss functions. We show that the theorem holds as well for the loss function ℓ_{h^k}:

Lemma 1. *Theorem 10 holds as well for the loss function* $\ell_{h^k} : \mathbb{R}^{\binom{N}{k}+1} \times \mathbb{R}^{\binom{N}{k}+1} \to [0,c]$ *with* $N, k \in \mathbb{N}$ *by choosing the frequency set* Ω *and the bounds* B *and* $\tilde{B}$ *large enough, such that both the functions* f *of a considered function family* $\mathcal{F}$ *and their derivatives* $D^\alpha f$ *for* $|\alpha| \leq k$ *are contained in the families* $\mathcal{F}_\Omega^B \subseteq \mathcal{H}_\Omega^{\tilde{B}}$*.*

Proof. The proof goes analogous to the proof of Theorem 11 in [4]. There are two points which require special care:
Firstly, we need to adapt the application of Talagrand's lemma which is used to upper bound the Rademacher complexity. Let us use the ℓ_{l^2} loss function

$$\ell_{l^2}(f^*(\boldsymbol{x}), f(\boldsymbol{x})) = (f^*(\boldsymbol{x}) - f(\boldsymbol{x}))^2 \ , \tag{53}$$

which is related to the loss function ℓ_{h^k} by

$$\ell_{h^k}(f^*(\boldsymbol{x}), f(\boldsymbol{x})) = \sum_{|\alpha| \leq k} \ell_{l^2}\left(D^\alpha f^*(\boldsymbol{x}), D^\alpha f(\boldsymbol{x})\right) \ . \tag{54}$$

By using the reverse triangle inequality, we can prove the lipschitzness of the loss function ℓ_{l^2}, for a fixed $f^*(\boldsymbol{x}) \in L^2([0,2\pi]^N)$:

$$\left|\ell_{l^2}(f_1(\boldsymbol{x}), f^*(\boldsymbol{x})) - \ell_{l^2}(f_2(\boldsymbol{x}), f^*(\boldsymbol{x}))\right| = \left| |f^*(\boldsymbol{x}) - f_1(\boldsymbol{x}))|^2 - |f^*(\boldsymbol{x}) - f_2(\boldsymbol{x}))|^2 \right| \tag{55}$$

$$\leq \left| |f^*(\boldsymbol{x}) - f_1(\boldsymbol{x}) - (f^*(\boldsymbol{x}) - f_2(\boldsymbol{x}))|^2 \right| \tag{56}$$

$$= \left| |(f^*(\boldsymbol{x}) - f^*(\boldsymbol{x})) - (f_1(\boldsymbol{x})) - f_2(\boldsymbol{x}))|^2 \right| \tag{57}$$

$$= |(f_1(\boldsymbol{x})) - f_2(\boldsymbol{x}))|^2 \ . \tag{58}$$

Thus, the loss function ℓ_{l^2} is L-Lipschitz with the Lipschitz constant $L = 1$. Note that this is the Lipschitz constant of the loss function ℓ_{l^2}, which is not related to the Lipschitz constant of functions of the function space $L^2([0,2\pi]^N)$.
Parallel to the proof of Theorem 11 in [4], we now define the set

$$\mathcal{G} = \left\{ [0,2\pi]^N \times [0,2\pi]^N \ni (\boldsymbol{x}, \boldsymbol{x}) \mapsto \ell_{h^k}(f^*(\boldsymbol{x}), f(\boldsymbol{x})) \middle| f^* \in H^k([0,2\pi]^N) \text{ and } f \in \mathcal{F}_\Omega^B \right\} \tag{59}$$

We can now upper bound the Rademacher complexity $\hat{\mathcal{R}}_S(\mathcal{G})$ for a training set S with I data points and a target function f^* as

$$\hat{\mathcal{R}}_S(\mathcal{G}) = \frac{1}{I}\mathbb{E}_\sigma \left[\sup_{f\in\mathcal{F}_\Omega^B} \sum_{i=1}^{I} \sigma_i \ell_{h^k}(f(\boldsymbol{x}_i), f^*(\boldsymbol{x}_i)) \right] \tag{60}$$

$$= \frac{1}{I}\mathbb{E}_\sigma \left[\sup_{f\in\mathcal{F}_\Omega^B} \sum_{i=1}^{I} \sigma_i \sum_{|\alpha|\leq k} \ell_{l^2}(D^\alpha f(\boldsymbol{x}_i), D^\alpha f^*(\boldsymbol{x}_i)) \right] \tag{61}$$

$$\leq \frac{1}{I}\mathbb{E}_\sigma \left[\sup_{D^\alpha f(\boldsymbol{x})\in\mathcal{F}_\Omega^B, |\alpha|\leq k} \sum_{i=1}^{I} \sigma_i \sum_{|\alpha|\leq k} \ell_{l^2}(D^\alpha f(\boldsymbol{x}_i), D^\alpha f^*(\boldsymbol{x}_i)) \right] \tag{62}$$

$$= \sum_{|\alpha|\leq k} \frac{1}{I}\mathbb{E}_\sigma \left[\sup_{D^\alpha f(\boldsymbol{x})\in\mathcal{F}_\Omega^B} \sum_{i=1}^{I} \sigma_i \ell_{l^2}(D^\alpha f(\boldsymbol{x}_i), D^\alpha f^*(\boldsymbol{x}_i)) \right] \tag{63}$$

$$\leq \xi \sup_{|\alpha|\leq k} \frac{1}{I}\mathbb{E}_\sigma \left[\sup_{D^\alpha f(\boldsymbol{x})\in\mathcal{F}_\Omega^B} \sum_{i=1}^{I} \sigma_i \ell_{l^2}(D^\alpha f(\boldsymbol{x}_i), D^\alpha f^*(\boldsymbol{x}_i)) \right] . \tag{64}$$

The i.i.d. random variables $\sigma_i \in \{-1, 1\}$ are the Rademacher random variables and ξ is the number of derivatives D^α with $|\alpha| \leq k$. Here, we first used the relation between the loss functions ℓ_{l^2} and ℓ_{h^k}. Then, we used the fact that the supremum over functions and derivatives $D^\alpha f(\boldsymbol{x}) \in \mathcal{F}_\Omega^B, |\alpha| \leq k$ which are independent from each other is larger than the supremum which is only taken over the functions $f \in \mathcal{F}_\Omega^B$, in which case the derivatives that are taken account in the loss functions have to be the derivatives of these functions. In the last inequality, we used that each of the ξ terms in the sum $\sum_{|\alpha|\leq k}$ can be upper bounded by its supremum.
We can now apply Talagrand's lemma on the quantity $\frac{1}{I}\mathbb{E}_\sigma \left[\sup_{D^\alpha f(\boldsymbol{x})\in\mathcal{F}_\Omega^B} \sum_{i=1}^{I} \sigma_i \ell_{l^2}(D^\alpha f(\boldsymbol{x}_i), D^\alpha f^*(\boldsymbol{x}_i))\right]$, for a fixed $|\alpha| \leq k$ in which way we obtain the upper bound

$$\hat{\mathcal{R}}_S(\mathcal{G}) \leq \xi \hat{\mathcal{R}}_{S|_x}(\mathcal{F}_\Omega^B) , \tag{65}$$

where we used that the loss function ℓ_{l^2} has the Lipschitz constant $L = 1$ and where $S|_x := \{\boldsymbol{x}_i\}_{i=0}^I$ is the set of the unlabeled training data points. The supremum $\sup_{|\alpha|\leq k}$ can be omitted on the right hand side of the bound, since the subset $S|_x$ of the training set does not include the labels $D^\alpha f^*(\boldsymbol{x}_i)$ and since we assumed the function family $\mathcal{F}_\Omega^B$ to contain the relevant derivatives as well. This upper bound corresponds to equation (97) in the proof of Theorem 11 in [4], apart from the additional factor ξ.
Secondly, in the last step of the proof in [4], the authors use standard generalization bounds as stated in Theorem 1.15 in [23]. The formulation of this standard generalization bound theorem allows for the loss function ℓ_{h^k} as well. □

Definition 5 (Compact embedding)**.** *[1][Definition 1.25] Let X and Y be normed spaces with the norms $\|\cdot\|_X$ and $\|\cdot\|_Y$, respectively, and X a subspace of Y. Let $I : X \to Y$, $Ix = x$ for all $x \in X$ be the embedding operator from X to Y. We say that X is continuously embedded in Y, and write $X \to Y$, if there exists a constant C, such that*

$$\|Ix\|_Y \leq C \|x\|_X , \forall x \in X . \tag{66}$$

We call the embedding compact, if X is continuously embedded in V and the embedding operator I is compact.

Definition 6. *We write $C_B^0(U)$ for the space of bounded, continuous functions on U.*

Definition 7 (Finite cone and Cone condition)**.** *[1][Definitions 4.4 and 4.6] Let $v, x \in \mathbb{R}^N$ be nonzero vectors, let $\angle(x, v)$ be the angle between vectors x and v. For given such v, a $\rho > 0$ and a κ such that $0 < \kappa \leq \pi$, the set*

$$C_{v,\rho,\kappa} = \{x \in \mathbb{R}^N : x = 0 \text{ or } 0 < \|x\| \leq \rho, \angle(x, v) \leq \kappa/2\} \tag{67}$$

is called a finite cone of height ρ, axis direction v and aperture angle κ with vertex at the origin. We say that $U \subseteq \mathbb{R}^N$ satisfies the cone condition, if there exists a finite cone C such that every $x \in U$ is the vertex of a finite cone C_x contained in U and congruent to C.

Theorem 11 (Rellich-Kondrachov)**.** *[1][Theorem 6.3, Part I and II] Let U be a domain in $\mathbb{R}^N$ satisfying the cone condition, let U_0 be a bounded subdomain of U, and let U_0^N be the intersection of U_0 with a N-dimensional plane in $\mathbb{R}^N$. Let $k \geq 1$ be integers. Let one of the following cases hold:*

1. *$2k < N$ and $1 \leq p < 2N/(N - 2k)$*
2. *$2k = N$ and $1 \leq p < \infty$*
3. *$2k > N$ and $1 \leq p < \infty$*

Then, the following embeddings are compact:

$$H^k(U) \to L^p(U_0^N) \ . \tag{68}$$

Additionally, in case 3, the following embedding is compact:

$$H^k(U) \to C_B^0(U_0^N) \ . \tag{69}$$

Remark. *The theorem relates to the Rellich-Kondrachov Theorem stated in [1] in the following way:*

- *Case 1 and Case 2 are the two cases stated in Part 1 of Theorem 6.3 in [1].*
- *Case 3 corresponds to the first and second case of Part 2 in Theorem 6.3 in [1].*
- *We use a different notation: The symbols Ω, j, p, q, k, n, m used in [1] are here equal to $U, 0, 2, p, N, N, k$, respectively.*
- *We formulate the theorem for the special cases $W^{k,2} = H^k$ and $W^{0,p} = L^p$ of the Sobolev spaces.*

With these preliminary results, we can prove Theorems 5, 6 and 7, which we restate here:

Theorem 5 (Generalization bound for H^k)**.** *Let $f^* \in \mathcal{F} \subseteq H^k([0, 2\pi]^N)$ be a target function, and let there be a $B > 0$ and a $\tilde{B} > 0$, such that $\mathcal{F}_\Omega^B \subseteq \mathcal{H}_\Omega^{\tilde{B}}$ is a suitable model family. Let us further assume that $\ell_{h^k}(f_1(\boldsymbol{x}), f_2(\boldsymbol{x})) \leq c$ for all $\mathbf{x} \in [0, 2\pi]^N$, and for all $f_1, f_2 \in \mathcal{F}_\Omega^B$ or $\mathcal{F}$. For any $\delta \in (0, 1)$ and the empirical risk $D_{h^k}(f^*, f)$ trained on an i.i.d. training data S with size I and containing data of ξ partial derivatives, the following holds for all functions $f \in \mathcal{F}_\Omega^B$ with probability at least $1 - \delta$:*

$$D_{H^k}(f^*, f) \leq D_{h^k}(f^*, f) + r(|\Omega|, \xi, B, \tilde{B}, c, I, \delta), \tag{70}$$

where $r(|\Omega|, \xi, B, \tilde{B}, c, I, \delta) \to 0$ as $I \to \infty$.

Proof. In the work [4], the authors developed generalization bounds for the function family defined in (9). We restated the theorem in Theorem 10. As we have shown in Corollary 1, the theorem also holds for the loss function ℓ_{h^k}.

According to the assumption, the function f^* is in $\mathcal{F}_\Omega^B$. The choice of the constant $\tilde{B}$ such that $\mathcal{F}_\Omega^B \subseteq \mathcal{H}_\Omega^{\tilde{B}}$ is satisfied depends on the encoding strategy. As written in [4], it can for example for integer valued frequencies be chosen as $\tilde{B} = 2B$. Thus, Lemma 1 can be applied and the following bound holds:

$$D_{H^k}(f^*, f_{h^k}) \leq D_{h^k}(f^*, f_{h^k}) + r(|\Omega|, B, \tilde{B}, c, I, \delta) \ , \tag{71}$$

with a function $r(|\Omega|, B, \tilde{B}, c, I, \delta)$ which tends to 0 as $I \to \infty$. □

Theorem 6 (Generalization bound for L^p)**.** *Let $f^* \in \mathcal{F} \subseteq H^k([0,2\pi]^N)$ be a target function, and let there be a $B > 0$ and a $\tilde{B} > 0$, such that $\mathcal{F}_\Omega^B \subseteq \mathcal{H}_\Omega^{\tilde{B}}$ is a suitable model family. Let us further assume that $\ell_{h^k}(f_1(\boldsymbol{x}), f_2(\boldsymbol{x})) \leq c$ for all $\mathbf{x} \in [0,2\pi]^N$, and for all $f_1, f_2 \in \mathcal{F}_\Omega^B$ or $\mathcal{F}$. Assume that $k, p \in \mathbb{N}$ satisfy one of the two following cases:*

1. *$N\left(\frac{1}{2} - \frac{1}{p}\right) < k < N/2$ and $1 \leq p < N$.*
2. *$k \geq N/2$ and $1 \leq p < \infty$.*

For any $\delta \in (0,1)$ and the empirical risk $D_{h^k}(f^*, f)$ trained on an i.i.d. training data S with size I and containing data of ξ partial derivatives, the following holds for all functions $f \in \mathcal{F}_\Omega^B$ with probability at least $1-\delta$:

$$\frac{1}{C} D_{L^p}(f^*, f) \leq D_{h^k}(f^*, f) + r(|\Omega|, \xi, B, \tilde{B}, c, I, \delta), \tag{72}$$

where C is a constant and $r(|\Omega|, \xi, B, \tilde{B}, c, I, \delta) \to 0$ as $I \to \infty$.

Proof. We will prove the theorem by proving the following two inequalities:

$$\frac{1}{C} D_{L^p}(f^*, f) \leq D_{H^k}(f^*, f) \leq D_{h^k}(f^*, f) + r(|\Omega|, B, \tilde{B}, c, I, \delta)\ . \tag{73}$$

The right hand side inequality is following directly from Theorem 5, and the left hand side inequality is a consequence of Theorem 11. Let us look at case 1 in Theorem 11: We want to rewrite the bound $p < 2N/(N-2k)$ as an upper bound for k for a given p. Let us therefore firstly check, which values p is allowed to reach. Due to k being bound from above by $k < N/2$, the upper bound on p, $p < 2N/(N-2k)$ is maximal for $k = \frac{N}{2} - 1$, in which case the upper bound on p becomes $p < N$. That means that values for p chosen in $1 \leq p < N$ are valid values. With p such chosen, the bound $p < 2N/(N-2k)$ is equivalent to bounding k in the following way:

$$N\left(\frac{1}{2} - \frac{1}{p}\right) < k\ . \tag{74}$$

For case 2 in Theorem 11, we have the inequalities $k \geq N/2$ and $1 \leq p < \infty$.
Further, because of the assumptions $\ell_{h^k}(f^*(\mathbf{x}), f(\mathbf{x})) \leq c$ for all $\mathbf{x} \in [0,2\pi]^N$, the subdomain $U = [0,2\pi]^N$ is equal to U_0, and because an N-dimensional plane in $\mathbb{R}^N$ is $\mathbb{R}^N$ itself, U is also equal to U_0^N. Let C be a cone of height at most π, angle at most $\pi/2$. Then, for each $\boldsymbol{x}$ in $U = [0,2\pi]^N$, we can choose an appropriate axis direction such that C_x lies entirely in U, so it satisfies the cone condition. To sum up, Theorem 11 states that for the cases

1. $N\left(\frac{1}{2} - \frac{1}{p}\right) < k < N/2$ and $1 \leq p < N$.
2. $k \geq N/2$ and $1 \leq p < \infty$,

the following embeddings are compact:

$$H^k([0,2\pi]^N) \to L^p([0,2\pi]^N)\ . \tag{75}$$

According to the definition of a compact embedding (Definition 5), there exists a constant C, such that

$$\|f^* - f\|_{L^p} \leq C \|f^* - f\|_{H^k}\ . \tag{76}$$

□

Theorem 7 (Generalization bound for C^0)**.** *Let $f^* \in \mathcal{F} \subseteq H^k([0,2\pi]^N)$ be a target function, and let there be a $B > 0$ and a $\tilde{B} > 0$, such that $\mathcal{F}_\Omega^B \subseteq \mathcal{H}_\Omega^{\tilde{B}}$ is a suitable model family. Let us further assume that $\ell_{h^k}(f_1(\boldsymbol{x}), f_2(\boldsymbol{x})) \leq c$ for all $\mathbf{x} \in [0,2\pi]^N$, and for all $f_1, f_2 \in \mathcal{F}_\Omega^B$ or $\mathcal{F}$ and that $\|f\|_\infty \leq B$ for all $f \in \mathcal{F}_\Omega^B$. Assume, that $k \in \mathbb{N}$ satisfies $k > N/2$. For any $\delta \in (0,1)$ and the empirical risk $D_{h^k}(f^*, f)$ trained on an i.i.d. training data S with size I and containing data of ξ partial derivatives, the following holds for all functions $f \in \mathcal{F}_\Omega^B$ with probability at least $1-\delta$:*

$$\frac{1}{C} D_{C^0}(f^*, f) \leq D_{h^k}(f^*, f) + r(|\Omega|, \xi, B, \tilde{B}, c, I, \delta), \tag{77}$$

where C is a constant and $r(|\Omega|, \xi, B, \tilde{B}, c, I, \delta) \to 0$ as $I \to \infty$.

Proof. The prove of this theorem is equivalent to the proof of Theorem 6 above. We will prove this theorem as well by proving the following two inequalities:

$$\frac{1}{C} D_{L^p}(f^*, f) \leq D_{H^k}(f^*, f) \leq D_{h^k}(f^*, f) + r(|\mathcal{M}|, I, \delta)\ . \tag{78}$$

The right hand side inequality is following directly from Theorem 5, and the left hand side inequality is a consequence of Theorem 11. As written in the proof of Theorem 6, the assumptions of Theorem 11 are satisfied, we can thus also apply it here.
The upper bound on the distance $D_{C^0}(f^*, f)$ in the supremum norm is a direct consequence of the third case in Theorem 11.

□