BANDWIDTH ENABLES GENERALIZATION IN QUANTUM KERNEL MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Quantum computers are known to provide speedups over classical state-of-the-art machine learning methods in some specialized settings. For example, quantum kernel methods have been shown to provide an exponential speedup on a learning version of the discrete logarithm problem. Understanding the generalization of quantum models is essential to realizing similar speedups on problems of practical interest. Recent results demonstrate that generalization is hindered by the exponential size of the quantum feature space. Although these results suggest that quantum models cannot generalize when the number of qubits is large, in this paper we show that these results rely on overly restrictive assumptions. We consider a wider class of models by varying a hyperparameter that we call quantum kernel bandwidth. We analyze the large-qubit limit and provide explicit formulas for the generalization of a quantum model that can be solved in closed form. Specifically, we show that changing the value of the bandwidth can take a model from provably not being able to generalize to any target function to good generalization for well-aligned targets. Our analysis shows how the bandwidth controls the spectrum of the kernel integral operator and thereby the inductive bias of the model. We demonstrate empirically that our theory correctly predicts how varying the bandwidth affects generalization of quantum models on challenging datasets, including those far outside our theoretical assumptions. We discuss the implications of our results for quantum advantage in machine learning.

1 INTRODUCTION

Quantum computers have the potential to provide computational advantage over their classical counterparts (Nielsen & Chuang, 2011), with machine learning commonly considered one of the most promising application domains. Many approaches to leveraging quantum computers for machine learning problems have been proposed. In this work, we focus on quantum machine learning methods that only assume classical access to the data. Lack of strong assumptions on the data input makes such methods a promising candidate for realizing quantum computational advantage. Specifically, we consider an approach that has gained prominence in recent years wherein a classical data point is embedded into some subspace of the quantum Hilbert space and learning is performed using this embedding. This class of methods includes so-called quantum neural networks (Mitarai et al., 2018; Farhi & Neven, 2018) and quantum kernel methods (Havlíček et al., 2019; Schuld & Killoran, 2019). Quantum neural networks are parameterized quantum circuits that are trained by optimizing the parameters to minimize some loss function. In quantum kernel methods, only the inner products of the embeddings of the data points are evaluated on the quantum computer. The values of these inner products (kernel values) are then used in a model optimized on a classical computer (e.g., support vector machine or kernel ridge regression). The two approaches are deeply connected and can be shown to be equivalent reformulations of each other in many cases (Schuld, 2021). Since the kernel perspective is more amenable to theoretical analysis, in this work we focus only on the subset of models that can be reformulated as kernel methods. A support vector machine (SVM) with a quantum kernel based on Shor's algorithm has been shown to provide exponential (in the problem size) speedup over any classical algorithm for a version of the discrete logarithm problem (Liu et al., 2021), suggesting that a judicious embedding of classical data into the quantum Hilbert space can enable a quantum kernel method to learn functions that would be hard to learn otherwise.

While the quantum kernels provide a much larger class of learnable functions compared to their classical counterpart, the ability of quantum kernels to generalize when the number of qubits is large has been called into question. Informally, Kübler et al. (2021) show that the generalization is impossible if the largest eigenvalue of the kernel integral operator is small, and Huang et al. (2021) show that generalization is unlikely if the rank of the kernel matrix is large. The two conditions are connected since for a positive-definite kernel with fixed trace, a small value of the largest eigenvalue implies that the spectrum of the kernel is "flat" with many nonzero eigenvalues. Under the assumptions used by Kübler et al. (2021); Huang et al. (2021), as the number of qubits grows, the largest eigenvalue of the integral operator gets smaller and the spectrum becomes "flat". Therefore, Kübler et al. (2021); Huang et al. (2021) conclude that learning is impossible for models with a large number of qubits unless the amount of training data provided grows exponentially with qubit count. This causes the curse of "exponential" dimensionality (Schölkopf et al., 2002) inherent in quantum kernels. However, Shaydulin & Wild (2021) show that if the class of quantum embeddings is extended by allowing a hyperparameter (denoted "kernel bandwidth") to vary, learning is possible even for high qubit counts. While extensive numerical evidence for the importance of bandwidth is provided, no analytical results are known that explain the mechanism by which bandwidth enables generalization.

In this work, we show analytically that quantum kernel models can generalize even in the limit of large numbers of qubits (and exponentially large feature space). The generalization is enabled by the bandwidth hyperparameter (Schölkopf et al., 2002; Silverman, 2018) which controls the inductive bias of the quantum model. We study the impact of the bandwidth on the spectrum of the kernel using the framework of task-model alignment developed in Canatar et al. (2021), which is based on the replica method of statistical physics (Seung et al., 1992; Dietrich et al., 1999; Mezard & Montanari, 2009; Advani et al., 2013). While nonrigorous, this framework was shown to capture various generalization phenomena accurately compared with the vacuous bounds from statistical learning theory. Together with the spectral biases of the model, task-model alignment quantifies the required amount of samples to learn a task correctly. A "flat" kernel with poor spectral bias implies large sample complexities to learn each mode in a task, while poor task-model alignment implies a large number of modes to learn. On an analytically tractable quantum kernel, we use this framework to show generalization of bandwidth-equipped models in the limit of an infinite number of qubits. Generalization in this infinite-dimensional limit contrasts sharply with previous results suggesting that high dimensionality of quantum Hilbert spaces precludes learning.

Our main contribution is an analysis showing explicitly the impact of quantum kernel bandwidth on the spectrum of the corresponding integral operator and on the generalization of the overall model. On a toy quantum model, we first demonstrate this analytically by deriving closed-form formulas for the spectrum of the integral operator, and show that larger bandwidth leads to larger values of the top eigenvalue and to a less "flat" spectrum. We show that for an aligned target function the kernel can generalize if bandwidth is optimized, whereas if bandwidth is chosen poorly, generalization requires an exponential number of samples on any target. Furthermore, we provide numerical evidence that the same mechanism allows for successful learning for a much broader class of quantum kernels, where analytical derivation of the integral operator spectrum is impossible. While our results do not necessarily imply quantum advantage, the evidence we provide suggests that, even with a compatible, well-aligned task, the quantum machine learning methods require a form of spectral bias to escape the curse of dimensionality and enable generalization.

2 Background

We begin by reviewing relevant classical and quantum machine learning concepts and establishing the notation used throughout the paper. We study the problem of regression, where the goal is to learn a target function from data. Specifically, the input is the training set $\mathcal{D} = \{\mathbf{x}^\mu, y^\mu\}_{\mu=1}^P$ containing P observations, with $\mathbf{x}$ drawn from some marginal probability density function $p : \mathcal{X} \to \mathbb{R}$ defined on $\mathcal{X} \subset \mathbb{R}^n$ and y produced by a target function $\bar{f} : \mathcal{X} \to \mathbb{R}$ as $y = \bar{f}(\mathbf{x})$.

Learning with kernels Given data in $\mathcal{X}$ distributed according to a probability density function $p : \mathcal{X} \to \mathbb{R}$, we consider a finite-dimensional complex reproducing kernel Hilbert space (RKHS) $\mathcal{H}$ and a corresponding feature map $\psi : \mathcal{X} \to \mathcal{H}$. This feature map gives rise to a kernel function $k(\mathbf{x}, \mathbf{x}') = \langle \psi(\mathbf{x}), \psi(\mathbf{x}') \rangle_{\mathcal{H}}$. The RKHS $\mathcal{H}$ associated with k is endowed with an inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$

satisfying the reproducing property and comprises functions $f : \mathcal{X} \to \mathbb{R}$ such that $\langle f, f \rangle_{\mathcal{H}} < \infty$ (Schölkopf et al., 2002). Given a set of P data points $\mathbf{x}^\mu \in \mathcal{X}$, the positive semidefinite Gram matrix is defined elementwise by $\mathbf{K}_{\mu\nu} = k(\mathbf{x}^\mu, \mathbf{x}^\nu)$. The continuous analogue to the Gram matrix $\mathbf{K}$ is the integral kernel operator $T_k : L_2(\mathcal{X}) \to L_2(\mathcal{X})$ defined according to its action:

$$(T_k f)(\mathbf{x}) = \int_{\mathcal{X}} k(\mathbf{x}, \mathbf{x}') f(\mathbf{x}') p(\mathbf{x}) d\mathbf{x}. \tag{1}$$

By Mercer's theorem (Schölkopf et al., 2002), the eigenfunctions of T_k satisfying $T_k \phi_k = \eta_k \phi_k$ are orthonormal (i.e., $\langle \phi_k, \phi_l \rangle = \delta_{kl}$), span $L_2(\mathcal{X})$, and give rise to an eigendecomposition of k given by $k(\mathbf{x}, \mathbf{x}') = \sum_k \eta_k \phi_k(\mathbf{x}) \phi_k^*(\mathbf{x}')$, where $\{\eta_k\}$ are real-valued, nonnegative eigenvalues of the integral operator due to its Hermiticity. The inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ in the RKHS of k may be computed with respect to the integral kernel operator of Eq. 1 as $\langle f, g \rangle_{\mathcal{H}} = \langle f, T_k^{-1} g \rangle_{L_2(\mathcal{X})}$, with the null space of T_k ignored in computing T_k^{-1}. From the kernel eigendecomposition, any target $\bar{f} \in L_2(\mathcal{X})$ that lies in the RKHS may therefore be decomposed as $\bar{f}(\mathbf{x}) = \sum_k \bar{a}_k \phi_k(\mathbf{x})$, where $\bar{a}_k = \langle \bar{f}, \phi_k \rangle_{\mathcal{H}}$ are the target weights. We comment on the case where the target lies outside of the RKHS in Appendix C.

Kernel ridge regression (KRR) is a convex optimization problem over functions that belong to a Hilbert space $\mathcal{H}$ and is stated as follows:

$$f^* = \arg\min_{f \in \mathcal{H}} \frac{1}{2} \sum_{\mu=1}^{P} \left(f(\mathbf{x}^\mu) - y^\mu \right)^2 + \frac{\lambda}{2} \|f\|_{\mathcal{H}}^2, \tag{2}$$

where $\|\cdot\|_{\mathcal{H}}$ denotes the norm with respect to the inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ defined on the Hilbert space and $\lambda \geq 0$ is the ridge parameter introduced for regularizing the solution.

Using these definitions, one can show that the solution to the regression problem takes the form $f^*(\mathbf{x}) = \mathbf{k}(\mathbf{x})^\top (\mathbf{K} + \lambda \mathbf{I})^{-1} \bar{\mathbf{y}}$, where $\mathbf{k}(\mathbf{x})$ is a vector with elements $\mathbf{k}(\mathbf{x})_\mu = k(\mathbf{x}, \mathbf{x}^\mu)$ and $\bar{\mathbf{y}}_\mu = \bar{y}^\mu$. The kernel trick (Schölkopf et al., 2002) allows one to perform regression without explicitly computing the features if one has access to the analytical kernel function. While providing a rich class of kernels, quantum kernels are typically not expressible analytically and hence require explicit representations of the feature maps.

Machine learning with quantum computers The central motivation for applying quantum computers to problems in machine learning is to leverage the ability of quantum systems to efficiently perform computation in a high-dimensional quantum Hilbert space. We consider quantum systems defined on n qubits whose dynamics may be described using complex-valued linear operators. A general quantum state on n qubits may be described by a positive definite $2^n \times 2^n$ *density matrix* with unit trace and contained in the quantum Hilbert space $\mathcal{H} = \{\rho \mid \rho \succ 0, \mathrm{Tr}(\rho) = 1, \rho \in L(\mathbb{C}^{2^n})\}$, where $L(\mathbb{C}^d)$ denotes bounded linear operators of the form $\mathbb{C}^d \to \mathbb{C}^d$ or, equivalently, $d \times d$ complex matrices. $\mathcal{H}$ is endowed with the inner product $\langle \rho, \rho' \rangle_{\mathcal{H}}$ given by the Hilbert–Schmidt inner product $\langle A, B \rangle_{\mathrm{HS}} = \mathrm{Tr}\{A^\dagger B\}$ for $A, B \in L(\mathbb{C}^d)$.

The dynamics of quantum states are described by applying linear operators to ρ, and we here will specifically consider unitary operations (represented by a $2^n \times 2^n$ unitary matrix U). Then, the quantum states that we are interested in may be prepared by applying a unitary operator to the vacuum state $|0\rangle$, where the "ket" notation $|j\rangle$ represents the jth standard basis vector $\hat{e}_j$ in $\mathbb{R}^{2^n}$. When $n = 1$, quantum states may be represented by using the *Bloch sphere*: the constraints $\mathrm{Tr}\{\rho\} = 1$ and $\rho = \rho^\dagger$ mean that the components of a density matrix are parameterized by three real parameters $\mathbf{n} = (n_x, n_y, n_z)$. Since $\|\mathbf{n}\| \leq 1$ is required for $\rho \geq 0$ subject to $\mathrm{Tr}\{\rho\} = 1$, we can identify any single-qubit state ρ with a vector $\mathbf{n}$ in the unit sphere $S^2 \subset \mathbb{R}^3$. Similarly, any unitary operation acting on a single qubit may be represented as a sequence of rotations on the Bloch sphere (Nielsen & Chuang, 2011), thus making this representation convenient for visualizing feature maps associated with quantum kernels.

By associating the quantum Hilbert space with a feature space $\mathcal{H}$, we can define a feature map that gives rise to a quantum kernel (Havlíček et al., 2019; Schuld & Killoran, 2019). We consider a data-dependent unitary operator $U(\mathbf{x})$ and prepare a density matrix $\rho(\mathbf{x}) = U(\mathbf{x})|0\rangle\langle 0|U^\dagger(\mathbf{x})$, where $|0\rangle\langle 0|$ represents the $2^n \times 2^n$ matrix with "1" in the top left corner and zeros elsewhere; later we discuss examples of how to construct data-dependent unitary operators. A natural choice for a feature map is then $\psi(\mathbf{x}) \to \rho(\mathbf{x})$ with the corresponding inner product $\langle \rho, \rho' \rangle_{\mathcal{H}} = \mathrm{Tr}\{\rho\rho'\}$. Under

this feature map, the quantum kernel is defined as $k(\mathbf{x}, \mathbf{x}') = \langle \psi(\mathbf{x}), \psi(\mathbf{x}') \rangle_{\mathcal{H}} = \mathrm{Tr}\{\rho(\mathbf{x})\rho(\mathbf{x}')\}$, which inherits symmetry in its arguments from $\langle \cdot, \cdot \rangle_{\mathcal{H}}$. With this association between the feature map $\psi(\mathbf{x})$ and the quantum state $\rho(\mathbf{x})$, we can freely apply existing theory for kernel methods in terms of complex vector spaces to study the spectra and generalization behavior of quantum kernels. In Appendix A we provide further details on the theory of quantum states and kernels, such as construction of the RKHS from Hermitian linear operators (or *observables*) and the geometry and probabilistic nature of quantum operations. In practice, a quantum kernel method computes the kernel matrix entries on a quantum computer by evaluating the value of the observable $|0\rangle\langle 0|$ on the state $U(\mathbf{x})U(\mathbf{x}')|0\rangle\langle 0| U^\dagger(\mathbf{x}')U^\dagger(\mathbf{x})$.

3 Motivating example: no generalization without bandwidth

Despite existing techniques for empirically evaluating the potential performance of quantum kernels on classical datasets (Huang et al., 2021) and examples of successful implementation on currently available quantum computers (Glick et al., 2021; Peters et al., 2021; Wu et al., 2021; Hubregtsen et al., 2021), it is often still unclear how to construct a quantum kernel that will be suitable for learning on a real-world classical dataset. This uncertainty in the potential performance of quantum machine learning methods is compounded by regimes in which generalization is apparently impossible with a subexponential amount of training data. These regimes arise from the same feature of quantum computing that originally motivated quantum kernel methods: the availability of exponentially large feature spaces. In this section we discuss a simple example where the high dimensionality of the feature space precludes learning with fixed quantum embeddings, and we show that the introduction of bandwidth enables generalization.

One example of how high dimensionality of the feature space precludes learning is provided by random feature maps. Given a feature map ψ consisting entirely of independent Gaussian features, the operator $\Sigma = \mathbb{E}_{\mathcal{X}}[\psi\psi^\dagger]$ is proportional to the identity operator. Since Σ shares eigenvalues with T_k of Eq. 1 (Appendix A), the eigenvalues of $\mathbf{K}$ concentrate around unity at a rate $\mathcal{O}(P^{-1/2})$ (Rosasco et al., 2010). Analogously, we consider a quantum feature map where states $\rho(\mathbf{x})$ are prepared by 2^n-dimensional random unitaries $U(\mathbf{x})$, with the uniform distribution over unitaries being described by the *Haar measure* (e.g., Collins & Nechita (2016)). Then, identifying a correspondence $\Sigma \to \mathbb{E}_{U\sim U(2^n)}[\rho(\mathbf{x}) \otimes \rho^T(\mathbf{x})]$ in the quantum case and applying standard results from measure theory, one can directly compute the spectrum of Σ (Appendix A). This computation yields $2^{n-1}(2^n+1)$ nonzero eigenvalues with magnitude $2^{1-n}(2^n+1)^{-1}$, and thus the nonzero eigenvalues of $\mathbf{K}$ again become uniform as $n \to \infty$. Since the largest eigenvalue is exponentially small in n, generalization requires the number of data points to be exponentially large in n. The connection between the magnitude of the largest eigenvalue and the required number of training samples can be seen directly from Kübler et al. (2021, Theorem 3), although it is a straightforward consequence of many older results, for example, (Dietrich et al., 1999; Bengio et al., 2005; Liang et al., 2019; Bordelon et al., 2020; Canatar et al., 2021). In other words, efficient generalization becomes impossible when the quantum feature map uses the full extent of the quantum state space uniformly. Our results demonstrate that the converse is true: restricting embeddings to a smaller region of quantum state space recovers the possibility of efficient learning.

3.1 Limitations of fixed quantum embeddings

To make the example concrete, we consider the following learning problem. This learning problem and the failure of fixed quantum embeddings on it were considered in Huang et al. (2021, Supplementary Information 9). For an input $\mathbf{x} \in \{0, \pi\}^n$, the goal is to learn a target function $\bar{f}(\mathbf{x}) = \cos\left(x^{(n)}\right)$, where $x^{(n)}$ is the last element of the vector $\mathbf{x}$. Since learning this function is equivalent to learning the value of the last element of $\mathbf{x}$, it is trivial classically, and a simple linear regression succeeds. We now show how KRR with a badly designed quantum kernel fails on this trivial task, and we show how the introduction of bandwidth allows the KRR with a quantum kernel to learn the target.

Consider a quantum kernel equipped with feature map

$$U(\mathbf{x}) = \bigotimes_{j=1}^{n} U(x^{(j)}), \qquad U(x^{(j)}) = R_x\left(x^{(j)}\right) = \begin{pmatrix} \cos\left(x^{(j)}/2\right) & i\sin\left(x^{(j)}/2\right) \\ i\sin\left(x^{(j)}/2\right) & \cos\left(x^{(j)}/2\right) \end{pmatrix}, \tag{3}$$

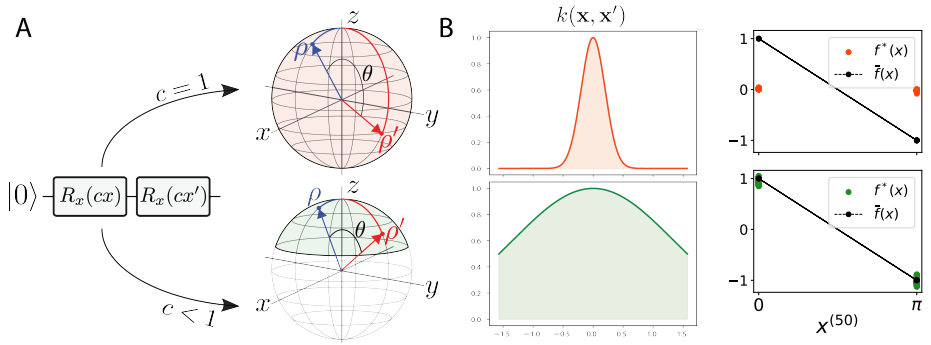

Figure 1: **A** A quantum kernel method involves embedding data using a quantum circuit, often involving rotation by some *angle* about an axis. When angles are not rescaled properly, data can be embedded far apart in a space with dimensionality $O(2^n)$ (Sec. 3). Similarly, $\lambda_{\max}$ is suppressed as the mean embedding $\mathbb{E}_{\mathcal{X}}[\rho(\mathbf{x})]$ approaches the center of the Bloch sphere (Kübler et al., 2021). **B** High-dimensional feature space results in $k(\mathbf{x}, \mathbf{x}')$ with narrow width (Eq. 6 with $n = 50$ and (top) $c = 1$ (bottom) $c = 0.25$) Tuning bandwidth escapes the "curse of dimensionality" associated with high-dimensional feature space. For $n = 50$ and $\bar{f}(\mathbf{x}) = \cos\left(x^{(50)}\right)$, quantum features that are nearly orthogonal result in a *narrow* kernel ($c = 1$, top) and failure to generalize, while tuning bandwidth ($c < 1$, bottom) recovers KRR generalization performance.

where $\bigotimes_{j=1}^{n}$ is the tensor product of single-qubit operations and $R_x(\theta)$ represents a rotation of θ about the x-axis of the single-qubit Bloch sphere (see Fig. 1A). For this feature map, the embedding factorizes over qubits as $\rho(\mathbf{x}) = U(\mathbf{x})|0\rangle\langle 0|U^\dagger(\mathbf{x}) = \bigotimes_{j=1}^{n} \rho(x^{(j)})$, and

$$\rho(x^{(j)}) = \begin{pmatrix} \cos^2(x^{(j)}/2) & -i\cos\left(x^{(j)}/2\right)\sin\left(x^{(j)}/2\right) \\ i\cos\left(x^{(j)}/2\right)\sin\left(x^{(j)}/2\right) & \sin^2(x^{(j)}/2) \end{pmatrix}, \tag{4}$$

with the kernel given by

$$k(\mathbf{x}, \mathbf{x}') = \mathrm{Tr}\big(\rho(\mathbf{x})\rho(\mathbf{x}')\big) = \prod_{j=1}^{n} \cos^2\left(\left(x^{(j)} - x'^{(j)}\right)/2\right). \tag{5}$$

While this kernel is obtained via quantum operations, the fact that it has a closed form expression makes it classically easy to simulate. Nevertheless, it is a useful toy model (Kübler et al., 2021; Huang et al., 2021) for the analytical analysis of exponentially large quantum feature spaces. Since the input data is $\mathbf{x} \in \{0, \pi\}^n$, the kernel becomes a delta function: $k(\mathbf{x}, \mathbf{x}') = \delta_{\mathbf{x},\mathbf{x}'}$. Therefore, any two points in the feature space $\rho(\mathbf{x})$ and $\rho(\mathbf{x}')$ for $\mathbf{x} \neq \mathbf{x}'$ are orthogonal and the kernel cannot capture the correlations in data. KRR with this kernel simply memorizes the training set and cannot generalize to any target function with a subexponential (in n) training set size.

3.2 Bandwidth enables learning

The preceding example highlights how quantum kernel methods utilizing high-dimensional spaces can fail to generalize. Our central technique for mitigating this limitation will be to introduce bandwidth to quantum kernels (Shaydulin & Wild, 2021). We now reconsider the kernel with the feature map of Eq. 3 and introduce a scaling parameter $c \in [0, 1]$ that controls the bandwidth of the kernel. The feature map and the kernel become

$$U(x^{(j)}) = R_x\left(cx^{(j)}\right), \qquad k(\mathbf{x}, \mathbf{x}') = \prod_{j=1}^{n} \cos^2\left(c\left(x^{(j)} - x'^{(j)}\right)/2\right). \tag{6}$$

Geometrically, the factor c restricts features $\rho(\mathbf{x})$ to a smaller region of the Bloch sphere (Fig. 1A). Consequently, the kernel matrix is no longer diagonal, and we can straightforwardly check that simply tuning c allows KRR with kernel Eq. 6 to generalize (see Fig. 1B).

4 EFFECT OF BANDWIDTH ON SCALING AND SPECTRA

In the preceding section we provided a qualitative mechanism for how bandwidth improves the generalization of quantum kernels. We now analyze the expected generalization error of quantum kernels equipped with bandwidth. We derive explicitly the spectrum of the bandwidth-equipped quantum kernel with the feature map Eq. 6 and show how the bandwidth makes the spectrum less "flat," thereby enabling learning. Our main tool is the theory developed in Bordelon et al. (2020); Canatar et al. (2021) where the generalization error, as a function of the training set size, is analytically obtained from the eigenvalues of the kernel and the projection of the target function on the RKHS defined by the kernel.

4.1 EXPLICIT FORMULAS FOR GENERALIZATION OF BANDWIDTH-EQUIPPED KERNELS

We consider the quantum kernel with the feature map given by Eq. 6 and the input distribution $\mathbf{x} \sim \text{Unif}([-\pi, \pi]^n)$, as previously studied by Kübler et al. (2021). For a single qubit, the kernel becomes $k(x, x') = \cos^2(c(x - x')/2)$ with bandwidth parameter c, and the eigenvalues can be computed as follows (see Appendix B):

$$\lambda_1 = \frac{3}{8} + \frac{1}{8}\text{sinc}(2\pi c) + \frac{1}{8}\sqrt{(1 - \text{sinc}(2\pi c))^2 + 16\text{sinc}(\pi c)^2}, \qquad \lambda_2 = \frac{1}{4} - \frac{1}{4}\text{sinc}(2\pi c),$$
$$\lambda_3 = \frac{3}{8} + \frac{1}{8}\text{sinc}(2\pi c) - \frac{1}{8}\sqrt{(1 - \text{sinc}(2\pi c))^2 + 16\text{sinc}(\pi c)^2}, \qquad \lambda_4 = 0. \tag{7}$$

For an n-qubit system, the largest eigenvalue $\eta_{\max}$ of the kernel in Eq. 6 falls exponentially with n for $c \sim \mathcal{O}_n(1)$ and makes generalization impossible with $P \sim \text{poly}(n)$ amount data (Kübler et al., 2021) (see Appendix. B). To prevent $\eta_{\max}$ decreasing with n, we choose a bandwidth that scales with n (i.e., $c = an^{-\alpha}$ with $a \sim \mathcal{O}(1)$ and $\alpha > 0$). In Appendix B we show that $\alpha \geq \frac{1}{2}$ is required for generalization. We consider $\alpha = 1/2$ henceforth. Then, all nonzero eigenvalues of the n-qubit kernel Eq. 6 can be expressed as $\lambda_1^{k_1}\lambda_2^{k_2}\lambda_3^{k_3}$ with $k_1 + k_2 + k_3 = n$, where the single-qubit eigenvalues asymptotically (with n) look like

$$\lambda_1 \approx 1, \quad \lambda_2 \approx \frac{a^2\pi^2}{6n}, \quad \lambda_3 \approx \frac{a^4\pi^4}{180n^2}. \tag{8}$$

Starting from $k_1 = n$ and $k_2 = k_3 = 0$, the hierarchy of eigenvalues obtained in this way are given in Table 1. We denote each of these eigenvalues as $\eta_{k,z}$ and their corresponding eigenfunctions as

Table 1: Hierarchy of eigenvalues based on their scaling. $N(n, k)$ denotes the degeneracy of eigenvalues with scaling n^{-k}, and $|n, k\rangle$ denotes the form of the corresponding eigenstates.

n^{-k}	Degeneracy $N(n,k)$	Eigenstate $\lvert n,k\rangle$
n^0	1	$\lvert\psi_1\rangle^{\otimes n}$
n^{-1}	$\binom{n}{1}$	$\lvert\psi_1\rangle^{\otimes(n-1)}\lvert\psi_2\rangle$
n^{-2}	$\binom{n}{2} + \binom{n}{1}$	$\lvert\psi_1\rangle^{\otimes(n-2)}\lvert\psi_2\rangle^{\otimes 2}, \lvert\psi_1\rangle^{\otimes(n-1)}\lvert\psi_3\rangle$
n^{-3}	$\binom{n}{3} + \binom{n-1}{1}\binom{n}{1}$	$\lvert\psi_1\rangle^{\otimes(n-3)}\lvert\psi_2\rangle^{\otimes 3}, \lvert\psi_1\rangle^{\otimes(n-2)}\lvert\psi_2\rangle\lvert\psi_3\rangle$

$\phi_{k,z}(x)$, where k indexes the overall scaling of the eigenvalue and z indexes each of the individual eigenvalues with scaling n^{-k}. The degeneracy $N(n, k) \sim \mathcal{O}(n^k)$ denotes the number of eigenvalues $\eta_{k,z} \sim \mathcal{O}(n^{-k})$ for each k. Notice that the quantity $\bar{\eta}_{k,z} = N(n, k)\eta_{k,z} \sim \mathcal{O}_n(1)$ with respect to the input dimension and its value depends on a and k. For each k, we will further make the approximation $\bar{\eta}_{k,z} \approx \bar{\eta}_{k,z'}$ for all pairs z, z' since the spectrum is almost flat at each scaling k (see Figure 2A).

We obtain the projections of target function $\bar{f}(x)$ on the kernel eigenfunctions as $\bar{a}_{k,z} = \int \bar{f}(x)\phi_{k,z}(x)p(x)dx$. We also define the total target power at each scaling $\bar{a}_k^2 \equiv \sum_{z=1}^{N(n,k)} \bar{a}_{k,z}^2$. With the eigenvalues and the target weights $\bar{a}_k^2$, the generalization error is given by (Canatar et al., 2021) (Appendix C):

$$E_g = \frac{\kappa^2}{1-\gamma}\sum_k \frac{\bar{a}_k^2}{(\kappa + \alpha_k\bar{\eta}_k)^2}, \quad \kappa = \lambda + \kappa\sum_k \frac{\bar{\eta}_k}{\kappa + \alpha_k\bar{\eta}_k}, \quad \gamma = \sum_k \frac{\alpha_k\bar{\eta}_k^2}{(\kappa + \alpha_k\bar{\eta}_k)^2}, \tag{9}$$

where λ is the KRR regularization parameter, κ should be solved self-consistently, and we have defined $\alpha_k = \frac{P}{N(n,k)}$. Taking the large qubit and large data limit ($n \to \infty$ and $P \to \infty$) while keeping $\alpha_l \sim \mathcal{O}_n(1)$ for some mode l (meaning $P \sim \mathcal{O}(n^l)$), the generalization error decouples across different scaling limits and becomes

$$E_g(\alpha_l) = \left(\frac{\kappa^2}{1-\gamma} \frac{\bar{a}_l^2}{\left(\kappa + \alpha_l \bar{\eta}_l\right)^2} + \frac{\gamma}{1-\gamma} \sum_{k>l} \bar{a}_k^2 \right) + \sum_{k>l} \bar{a}_k^2. \tag{10}$$

The target modes with $k > l$ remain unlearned since $\alpha_{k>l} = 0$ and the modes $k < l$ have already been learned using the provided data since $\alpha_{k<l} = \infty$. This scaling defines *learning stages* where at each stage a single mode l is being learned, and the remaining modes contribute as constant error. The term in the parentheses goes to zero as $\alpha_l \to \infty$ (see Appendix C). Hence, when the mode l is completely learned, the ratio of the generalization error to its initial value becomes $\frac{E_g(\alpha_l=\infty)}{E_g(0)} = (\sum_{k>l} \bar{a}_k^2)/(\sum_k \bar{a}_k^2)$. Therefore, given a data budget $P \sim \mathcal{O}(n^l)$, the quantum kernel is guaranteed to generalize to target functions whose weights beyond mode l are vanishing. The quantity $C(l) = 1 - \frac{E_g(\alpha_l=\infty)}{E_g(0)}$, called the *cumulative power* (Canatar et al., 2021), describes the amount of power placed in the first l modes and quantifies the task-model alignment.

It was shown in Canatar et al. (2021) that kernels generalize better for target functions with sharply rising $C(l)$. In our case, for generalizability at $P \sim \mathcal{O}(n^l)$ samples, it is a necessary but not sufficient condition for target functions to have good task-model alignment for which $C(l) \approx 1$. Note that the trace of this kernel is $\int k(\mathbf{x}, \mathbf{x}) p(\mathbf{x}) d\mathbf{x} = 1$, and therefore the eigenvalues satisfy $\sum_{k,z} \eta_{k,z} = 1$. Flatness of the spectrum, then, implies $\eta_{k,z} \sim \mathcal{O}(3^{-n})$ since there are 3^n nonzero modes of the kernel in Eq. 6 (see Appendix B). Equation 9 suggests that even if the target is aligned well with the kernel, it requires $P \sim \mathcal{O}(3^n)$ samples to learn each mode, and so generalization becomes impossible with polynomial sample complexity $P \sim \mathcal{O}(n^l)$.

On the other hand, bandwidth enables sufficient decay in the spectrum of the kernel, and Eq. 9 shows that $P \sim \mathcal{O}(n^l)$ samples yield an excess generalization error $E_g \approx 1 - C(l)$. In Figure 2, we show the results of simulating the kernel of Eq. 6 with $n = 40$ qubits for a target function given by $\bar{f}(\mathbf{x}) = e^{-\|\mathbf{x}\|^2/n^2}$. For $c = 1$, the kernel has a flat spectrum (Figure 2A) with poor alignment with the task (Figure 2B). On the other hand, bandwidth introduces sufficient decay in the eigenspectrum so that polynomial time learning becomes possible. Surprisingly, bandwidth also improves the task-model alignment which, together with spectral bias, implies generalizability with better sample efficiency. In Figure 2C, we perform kernel regression with our toy kernel and confirm that generalization improves with bandwidth up to an optimal value after which it degrades again. This is due to the fact that larger bandwidths cause underfitting of the data (Silverman, 2018) since only a very few eigenmodes become learnable while the target cannot be fully explained by those modes. In Figure 2, we find that the optimal bandwidth parameter is $c^* \approx \frac{2}{n}$ (see Appendix E).

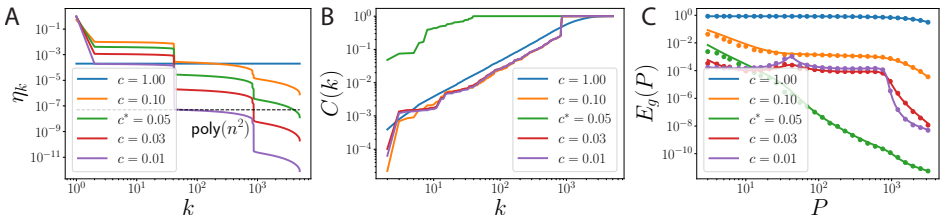

Figure 2: **A** Eigenvalues of the kernel Eq. 6 with respect to data $\mathbf{x} \sim \text{Unif}([-\pi, \pi]^{40})$. Learnability is explained by the preservation of large-eigenvalue eigenspaces; without bandwidth, the model provably cannot learn in poly(n) complexity due to the flat spectrum. **B** The projections of the target $\bar{f}(\mathbf{x}) = e^{-\|\mathbf{x}\|^2/n^2}$ on the eigenvectors of the kernel for each bandwidth. Apart from the flatness of the $c = 1$ kernel, its eigenfunctions align poorly with the target. **C** Generalization error as a function of the number of training samples computed by using theory (solid lines) (Eq. 9) and performing kernel ridge regression empirically (dots). Bandwidth $c = 1$ yields a constant learning curve. While all $c < 1$ kernels provide improvement, an optimal bandwidth parameter $c^* \approx 2/n$ gives the best task-model alignment.

4.2 EVIDENCE OF PERFORMANCE GAINS IN REAL DATASETS

To evaluate our theory in a practical setting, we consider two previously proposed quantum kernels that have been conjectured to be hard to simulate classically: a kernel with a feature map inspired by instantaneous quantum polynomial-time (IQP) circuit (Shepherd & Bremner, 2009; Havlíček et al., 2019; Huang et al., 2021) and a kernel with Hamiltonian evolution (EVO) feature map (Huang et al., 2021; Shaydulin & Wild, 2021) (see Appendix E for details). Unlike the kernel considered in the preceding section, the spectrum cannot be derived analytically for these kernels.

We evaluate these kernels on binary classification versions of FMNIST (Xiao et al., 2017), KMNIST (Clanuwat et al., 2018), and PLAsTiCC (The PLAsTiCC team et al., 2018) datasets with the input data downsized to $n = 22$ dimensions, which were previously used to evaluate quantum kernel performance in Shaydulin & Wild (2021); Huang et al. (2021); Peters et al. (2021). We use the kernel values reported in Shaydulin & Wild (2021), which were evaluated with high precision using an idealized (noiseless) simulator. In practice, an additive error is introduced when evaluating the kernel values on a fault-tolerant quantum computer. Bandwidth-equipped kernels are robust against this error; see the discussion in Shaydulin & Wild (2021). We perform SVM for binary classification using these kernels with varying bandwidths. In Table 2, we report the test accuracies with bandwidth parameter $c = 1$ for the IQP and EVO kernels. We also report the test performance with bandwidth parameters $c^* < c$ optimized by hyperparameter tuning for each kernel using cross validation (see Appendix E). For both kernels, bandwidth significantly improves the test performance.

	IQP		EVO		Random Guess
	c^*	$c = 1$	c^*	$c = 1$	
FMNIST	0.926	0.542	0.916	0.643	0.5
KMNIST	0.915	0.629	0.914	0.600	0.5
PLAsTiCC	0.789	0.5	0.788	0.613	0.5

Table 2: Bandwidth tuning recovers significant performance gains over out-of-the-box quantum models, opening up the possibility of better workflows for general quantum machine learning on many qubits via hyperparameter tuning.

To test our intuition, we present in Fig. 3A the shape of the IQP kernel where the kernel clearly sharpens for larger values of bandwidth parameter implying flat spectrum. In Fig. 3B, we further confirm that the spectrum decay improves with bandwidth when the IQP kernel is evaluated on the FMNIST dataset. We also find that the task-model alignment improves greatly with the bandwidth (Fig. 3C), thus implying, together with the previous point, possible generalization.

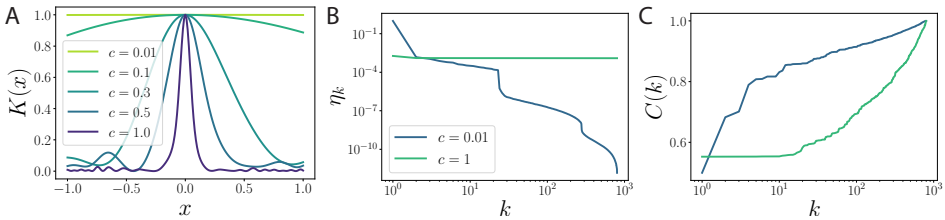

Figure 3: **A** IQP kernel function example: Intuitive behavior of bandwidth persists even when quantum kernel is not strictly shift-invariant. **B** IQP with the addition of bandwidth is capable of recovering significant target alignment with FMNIST dataset. **C** While the spectrum for $c = 1$ IQP is flat, it also has poor alignment with the target (see Appendix E).

5 RELATED WORK

Huang et al. (2021) introduce the idea that the exponential dimensionality of quantum feature spaces precludes generalization of quantum kernel methods, and they provide an upper bound on generalization error that includes the dimension of the space spanned by the training set. They connect their results to the spectrum of the kernel by introducing a measure of "approximate dimension" of

the span of the training set given by $\sum_{k=1}^{n}\left(\frac{1}{n-k}\sum_{l=k}^{m} t_l\right)$, where $\{t_l\}_{l=1}^{n}$ are the eigenvalues of the $n \times n$ kernel matrix. This number is effectively a measure of the flatness of the spectrum of the kernel and is used in numerical experiments in Huang et al. (2021). An alternative analysis is given by Banchi et al. (2021), who derive bounds on generalization error using quantum information techniques and show that near-identity kernels resulting from large dimensionality of the feature space preclude generalization. We use the construction from Huang et al. (2021, Appendix I) as our motivating example in Sec. 3.

Kübler et al. (2021) introduce techniques for deriving the spectrum of quantum kernels and obtain the spectrum of the kernel Eq. 5. They then use these techniques to show that the purity of the mean embedding provides an upper bound on the largest eigenvalue of the quantum kernel integral operator and consequently that quantum kernel methods with feature maps that utilize the full quantum Hilbert space require an exponential number of qubits of data to learn. However, Kübler et al. (2021) did not consider bandwidth as a method for controlling the inductive bias of quantum kernels. Our contribution is using the techniques of Kübler et al. (2021) to derive the spectrum of the bandwidth-equipped kernel Eq. 6 and explicit formulas for the expected generalization error.

Shaydulin & Wild (2021) introduce the concept of quantum kernel bandwidth and provide numerical evidence that bandwidth tuning (equivalent to rescaling of the input data) improves the generalization of SVM with quantum kernels. An analogous mechanism has been observed to improve trainability of quantum neural networks (Zhang et al., 2022). However, previous results make no connection between the bandwidth and the spectrum of the kernel and provide no analytical results for generalization. We reinterpret the data from Shaydulin & Wild (2021) in Sec. 4.2 and show how the bandwidth controls the spectrum of the kernel and enables generalization.

6 Discussion

In this work, we studied the kernels induced by quantum feature embeddings of data and their generalization potential. Recent work suggests that machine learning models built in this way suffer from the curse of dimensionality, requiring exponentially large training sets to learn. Note that embedding n-dimensional data requires at least n qubits (where $n \sim 10^3$ for standard datasets) which span a 2^n dimensional feature space. While quantum models may possess potentially powerful and classically inaccessible representations for certain tasks, utilization of those necessarily requires a control over the large space they span in order to generalize.

Here we showed that the bandwidth hyperparameter enables generalization of quantum kernel methods when the numbers of qubits is large, and provided explicit formulas for the resulting expected generalization error on a toy model. Our results open up promise for quantum machine learning beyond intermediate numbers of qubits. A central lesson provided by our work is that thoughtfully chosen hyperparameters can significantly improve the performance of quantum machine learning methods. Identifying such hyperparameters that control the inductive bias of quantum models is essential to realizing the full potential of quantum machine learning methods.

Unlike prior works, we focus not just on the spectrum of the quantum kernel but on the alignment between the kernel and the real-world datasets. Our empirical results imply that scaling the bandwidth not only makes the spectrum less flat, but also improves the alignment with real-world target functions. These observations make us optimistic about the potential of quantum kernel methods to solve classically challenging problems.

An important limitation of our results is that while bandwidth scaling is guaranteed to improve the spectrum, it does not necessarily lead to good alignment. For example, in the limit of $c \to 0$ the spectrum has only one nonzero eigenvalue, although learning is still not possible (see Appendix D). This suggests that optimizing bandwidth as a hyperparameter during training can balance triviality of the feature map ($c \to 0$) with greater utilization of the quantum state space. At the same time, if the feature map is chosen poorly, varying bandwidth would not lead to good generalization.

If the kernel values are evaluated on a noisy quantum computer, the quantum hardware noise would affect the spectrum of the kernel. Hardware noise reduces the purity of embedding, leading to a trivial lower bound on generalization error from Kübler et al. (2021, Theorem 1). Heyraud et al. (2022) give a more detailed analysis. While outside the scope of this work, the impact of noise on generalization is nonetheless an important consideration for near-term prospects of quantum kernel methods.

REFERENCES

Héctor Abraham, Ismail Yunus Akhalwaya, Gadi Aleksandrowicz, Thomas Alexander, G Alexandrowics, E Arbel, A Asfaw, C Azaustre, P Barkoutsos, G Barron, et al. Qiskit: An open-source framework for quantum computing. *URL https://doi. org/10.5281/zenodo*, 2562111, 2019. doi: 10.5281/zenodo.2562111.

Madhu Advani, Subhaneil Lahiri, and Surya Ganguli. Statistical mechanics of complex neural systems and high dimensional data. *Journal of Statistical Mechanics: Theory and Experiment*, 2013(03):P03014, 2013. doi: 10.1088/1742-5468/2013/03/P03014.

Leonardo Banchi, Jason Pereira, and Stefano Pirandola. Generalization in quantum machine learning: A quantum information standpoint. *PRX Quantum*, 2(4), November 2021. doi: 10.1103/prxquantum.2.040321. URL https://doi.org/10.1103/prxquantum.2.040321.

Yoshua Bengio, Olivier Delalleau, and Nicolas Le Roux. The curse of dimensionality for local kernel machines. Technical Report TR-1258, Université de Montréal, March 2005. URL https://www.microsoft.com/en-us/research/publication/the-curse-of-dimensionality-for-local-kernel-machines/.

Blake Bordelon, Abdulkadir Canatar, and Cengiz Pehlevan. Spectrum dependent learning curves in kernel regression and wide neural networks. In *International Conference on Machine Learning*, pp. 1024–1034. PMLR, 2020. URL https://proceedings.mlr.press/v119/bordelon20a.html.

Abdulkadir Canatar, Blake Bordelon, and Cengiz Pehlevan. Spectral bias and task-model alignment explain generalization in kernel regression and infinitely wide neural networks. *Nature Communications*, 12(1):1–12, 2021. doi: 10.1038/s41467-021-23103-1.

Tarin Clanuwat, Mikel Bober-Irizar, Asanobu Kitamoto, Alex Lamb, Kazuaki Yamamoto, and David Ha. Deep learning for classical Japanese literature. *arXiv:1812.01718*, 2018. doi: 10.20676/00000341.

Benoît Collins and Ion Nechita. Random matrix techniques in quantum information theory. *Journal of Mathematical Physics*, 57(1):015215, January 2016. doi: 10.1063/1.4936880.

Rainer Dietrich, Manfred Opper, and Haim Sompolinsky. Statistical mechanics of support vector networks. *Physical Review Letters*, 82(14):2975, 1999. doi: 10.1103/PhysRevLett.82.2975.

Edward Farhi and Hartmut Neven. Classification with quantum neural networks on near term processors. *arXiv:1802.06002*, 2018. doi: 10.48550/arXiv.1802.06002.

Jennifer R. Glick, Tanvi P. Gujarati, Antonio D. Corcoles, Youngseok Kim, Abhinav Kandala, Jay M. Gambetta, and Kristan Temme. Covariant quantum kernels for data with group structure. *arXiv:2105.03406*, 2021. doi: 10.48550/arXiv.2105.03406.

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(7747):209–212, March 2019. doi: 10.1038/s41586-019-0980-2.

Valentin Heyraud, Zejian Li, Zakari Denis, Alexandre Le Boité, and Cristiano Ciuti. Noisy quantum kernel machines. *arXiv:2204.12192*, 2022. doi: 10.48550/arXiv.2204.12192.

Zoë Holmes, Kunal Sharma, M. Cerezo, and Patrick J. Coles. Connecting ansatz expressibility to gradient magnitudes and barren plateaus. *PRX Quantum*, 3(1), January 2022. doi: 10.1103/prxquantum.3.010313.

Hsin-Yuan Huang, Michael Broughton, Masoud Mohseni, Ryan Babbush, Sergio Boixo, Hartmut Neven, and Jarrod R. McClean. Power of data in quantum machine learning. *Nature Communications*, 12(1), May 2021. doi: 10.1038/s41467-021-22539-9.

Thomas Hubregtsen, David Wierichs, Elies Gil-Fuster, Peter-Jan H. S. Derks, Paul K. Faehrmann, and Johannes Jakob Meyer. Training quantum embedding kernels on near-term quantum computers. *arXiv:2105.02276*, 2021. doi: 10.48550/arXiv.2105.02276.

Jonas Kübler, Simon Buchholz, and Bernhard Schölkopf. The inductive bias of quantum kernels. *Advances in Neural Information Processing Systems*, 34, 2021. URL https://proceedings.neurips.cc/paper/2021/file/69adc1e107f7f7d035d7baf04342e1ca-Paper.pdf.

Tengyuan Liang, Alexander Rakhlin, and Xiyu Zhai. On the multiple descent of minimum-norm interpolants and restricted lower isometry of kernels. *arXiv:1908.10292*, 2019. doi: 10.48550/arXiv.1908.10292.

Yunchao Liu, Srinivasan Arunachalam, and Kristan Temme. A rigorous and robust quantum speed-up in supervised machine learning. *Nature Physics*, 17(9):1013–1017, July 2021. doi: 10.1038/s41567-021-01287-z.

Jarrod R. McClean, Sergio Boixo, Vadim N. Smelyanskiy, Ryan Babbush, and Hartmut Neven. Barren plateaus in quantum neural network training landscapes. *Nature Communications*, 9(1), November 2018. doi: 10.1038/s41467-018-07090-4.

M. Mezard and A. Montanari. *Information, Physics, and Computation*. Oxford University Press, 2009. doi: 10.1093/acprof:oso/9780198570837.001.0001.

K. Mitarai, M. Negoro, M. Kitagawa, and K. Fujii. Quantum circuit learning. *Physical Review A*, 98 (3), September 2018. doi: 10.1103/physreva.98.032309.

Michael A Nielsen and Isaac L Chuang. *Quantum Computation and Quantum Information*. Cambridge University Press, 2011. doi: 10.1017/CBO9780511976667.

Evan Peters, João Caldeira, Alan Ho, Stefan Leichenauer, Masoud Mohseni, Hartmut Neven, Panagiotis Spentzouris, Doug Strain, and Gabriel N. Perdue. Machine learning of high dimensional data on a noisy quantum processor. *npj Quantum Information*, 7(1), November 2021. doi: 10.1038/s41534-021-00498-9.

Zbigniew Puchała and Jarosław Adam Miszczak. Symbolic integration with respect to the Haar measure on the unitary group. *Bulletin of the Polish Academy of Sciences: Technical Sciences*, 65 (1):21–27, 2017. doi: 10.1515/bpasts-2017-0003.

Lorenzo Rosasco, Mikhail Belkin, and Ernesto De Vito. On learning with integral operators. *Journal of Machine Learning Research*, 11(30):905–934, 2010. URL http://jmlr.org/papers/v11/rosasco10a.html.

Bernhard Schölkopf, Alexander J Smola, Francis Bach, et al. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, 2002. doi: 10.7551/mitpress/4175.001.0001.

Maria Schuld. Supervised quantum machine learning models are kernel methods. *arXiv:2101.11020*, 2021. doi: 10.48550/arXiv.2101.11020.

Maria Schuld and Nathan Killoran. Quantum machine learning in feature Hilbert spaces. *Physical Review Letters*, 122(4), February 2019. doi: 10.1103/physrevlett.122.040504.

H. S. Seung, H. Sompolinsky, and N. Tishby. Statistical mechanics of learning from examples. *Physical Review A*, 45:6056–6091, Apr 1992. doi: 10.1103/PhysRevA.45.6056.

J. Shawe-Taylor, C.K.I. Williams, N. Cristianini, and J. Kandola. On the eigenspectrum of the Gram matrix and the generalization error of kernel-pca. *IEEE Transactions on Information Theory*, 51 (7):2510–2522, 2005. doi: 10.1109/TIT.2005.850052.

Ruslan Shaydulin and Stefan M. Wild. Importance of kernel bandwidth in quantum machine learning. *arXiv:2111.05451*, 2021. doi: 10.48550/arXiv.2111.05451.

Dan J. Shepherd and Michael J. Bremner. Temporally unstructured quantum computation. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 465:1413–1439, 2009. doi: 10.1098/rspa.2008.0443.

Bernard W Silverman. *Density estimation for statistics and data analysis*. Routledge, 2018.

The PLAsTiCC team, Tarek Allam, Anita Bahmanyar, Rahul Biswas, Mi Dai, Lluís Galbany, Renée Hložek, Emille E. O. Ishida, Saurabh W. Jha, David O. Jones, Richard Kessler, Michelle Lochner, Ashish A. Mahabal, Alex I. Malz, Kaisey S. Mandel, Juan Rafael Martínez-Galarza, Jason D. McEwen, Daniel Muthukrishna, Gautham Narayan, Hiranya Peiris, Christina M. Peters, Kara Ponder, Christian N. Setzer, The LSST Dark Energy Science Collaboration, The LSST Transients, and Variable Stars Science Collaboration. The photometric LSST astronomical time-series classification challenge (PLAsTiCC): Data set. *arXiv:1810.00001*, 2018. doi: 10.48550/arXiv.1810.00001.

John Watrous. *The Theory of Quantum Information*. Cambridge University Press, 2018. doi: 10.1017/9781316848142.

Sau Lan Wu, Shaojun Sun, Wen Guan, Chen Zhou, Jay Chan, Chi Lung Cheng, Tuan Pham, Yan Qian, Alex Zeng Wang, Rui Zhang, Miron Livny, Jennifer Glick, Panagiotis Kl. Barkoutsos, Stefan Woerner, Ivano Tavernelli, Federico Carminati, Alberto Di Meglio, Andy C. Y. Li, Joseph Lykken, Panagiotis Spentzouris, Samuel Yen-Chi Chen, Shinjae Yoo, and Tzu-Chieh Wei. Application of quantum machine learning using the quantum kernel algorithm on high energy physics analysis at the LHC. *Physical Review Research*, 3(3), sep 2021. doi: 10.1103/physrevresearch.3.033221.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms. *arXiv:1708.07747*, 2017. doi: 10.48550/arXiv.1708.07747.

Kaining Zhang, Min-Hsiu Hsieh, Liu Liu, and Dacheng Tao. Gaussian initializations help deep variational quantum circuits escape from the barren plateau. *arXiv:2203.09376*, 2022. doi: 10.48550/arXiv.2203.09376.