ESANN 2025 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 23-25 April 2025, i6doc.com publ., ISBN 9782875870933. Available from http://www.i6doc.com/en/.

Expressivity vs. Generalization in Quantum Kernel Methods

Markus Gross, Markus Lange, Bogusz Bujnowski, Hans-Martin Rieser

DLR Institute for AI Safety and Security
Sankt Augustin and Ulm, Germany

Abstract. We analytically and numerically investigate the expressivity and generalization ability of quantum kernel models. We consider prototypical parallel encoding strategies and show that they give rise to simple universal forms of quantum kernels. By using qubit-dependent data rescaling schemes, we can exponentially vary the spectral content of the kernel and thereby control its simplicity bias. We obtain analytical results on the kernel eigenspectrum and connect it to theories of kernel generalization, which allow us to study the influence of expressivity on generalization error.

1 Quantum Kernel Theory

Quantum kernels have been recognized as an attractive approach to NISQ-era quantum machine learning that avoids the optimization challenges often encountered in parameterized quantum models [1]. A kernel is a measure for similarity between two input data points x, y, evaluated using an inner product after a projection into a suitable high dimensional feature space [2]. In the quantum case, the mapping into the feature space is achieved by encoding an input datum $\mathbf{x} \in \mathbb{R}^d$ in a quantum state $|\psi\rangle \in \mathbb{C}^{2^N}$ of an N-qubit Hilbert space $\mathcal{H}$ via a feature map

$$|\psi(\mathbf{x})\rangle = U_R S(\mathbf{x})|\Phi_0\rangle. \tag{1}$$

Here $|\Phi_0\rangle$ denotes the initial state of the quantum system, $S(\mathbf{x})$ is a data encoding unitary and U_R is an arbitrary unitary operator which represents gate operations or time evolution. The corresponding quantum kernel [1] is given as

$$K(\mathbf{x}, \mathbf{y}) = |\langle\psi(\mathbf{x})|\psi(\mathbf{y})\rangle|^2 = |\langle\Phi_0|S(\mathbf{x})^\dagger S(\mathbf{y})|\Phi_0\rangle|^2. \tag{2}$$

Note that any data-independent unitary operator, such as U_R, acting after the data-dependent embedding, cancels out in the kernel.

We focus in the following on parallel Pauli encoding schemes of the form

$$S(\mathbf{x}) = \bigotimes_{j=1}^{N} \exp(-i2\pi x_j c_j \sigma_j^p), \tag{3}$$

where each qubit encodes one datum x_j (thus $d = N$) and the Pauli matrix σ_j^p ($p \in \{x, y, z\}$, identical for all qubits) acts on the jth qubit and trivially on all others. The scaling factors $c_j \in \mathbb{R}^+$ represent characteristic data length scales

ESANN 2025 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 23-25 April 2025, i6doc.com publ., ISBN 9782875870933. Available from http://www.i6doc.com/en/.

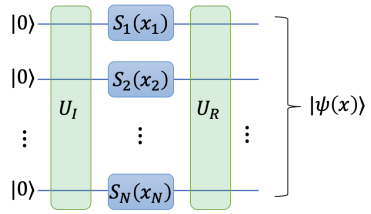


Figure 1: Illustration of the feature map considered here. A non-trivial unitary U_I can be chosen to transform the N-qubit ground state $|0\rangle$ into a different initial state (e.g., $U_I = H^{\otimes N}$, resulting in an equal superposition state). Data $\mathbf{x} \in \mathbb{R}^N$ is encoded in parallel via the unitaries $S_j(x_j)$ [see Eq. (3)]. Any additional data-independent unitaries such as U_R are immaterial for the kernel.

(possibly different for each qubit), which, in the context of kernel methods, are also known as bandwidths [2, 3]. The basic feature map is illustrated in Figure 1.

Beside the bandwidths, another source of expressiveness of a quantum model is provided by the choice of initial state $|\Phi_0\rangle$. We consider here the N-qubit ground state $|\Phi_0\rangle = |0\rangle^{\otimes N}$ as well as the uniform superposition state $|\Phi_0\rangle = \frac{1}{\sqrt{N}} \sum_{k=0}^{N-1} |k\rangle = H^{\otimes N}|0\rangle^{\otimes N}$ (with H denoting the Hadamard operator). If $p = z$ and $|\Phi_0\rangle$ is the ground state as well as if $p = x$ and $|\Phi_0\rangle$ is the uniform superposition state, one finds $K(\mathbf{x}, \mathbf{x}) = 1$. In all other cases, one obtains a non-trivial translation invariant quantum kernel:

$$K(\mathbf{x}, \mathbf{y}) = K(\mathbf{x} - \mathbf{y}) = \prod_{j=1}^{N} \cos^2\left((x_j - y_j)\, c_j \pi\right). \tag{4}$$

This form of the kernel, in fact, applies to all initial states and encoding schemes considered here. In principle, a kernel is a purely classical object and can in simple cases be studied further with established techniques [2]. A quantum advantage may arise for kernels derived from complex circuits which are classically intractable but may be efficiently estimated using quantum hardware.

1.1 Expressivity

In order to study the expressiveness of the quantum model, we encode the same datum into each qubit, i.e., we take $x_{j=1,\ldots,N} = x \in \mathbb{R}$, but allow for different bandwidths, characterized by the three encoding schemes listed in Table 1 [4]. The kernel then takes the form of a one-dimensional Fourier series:

$$K(x - y) = \sum_{\omega \in \Omega} b_\omega e^{\mathrm{i}\pi\omega(x-y)} = \sum_{\omega \in \Omega, \omega \geq 0} B_\omega \cos(\pi\omega(x - y)) \tag{5}$$

where the frequency values are listed in Table 1. For the ternary encoding the ω-dependence of b_ω is given by

$$A(\omega, N) := \begin{cases} 2^N\,, & \text{for } \omega \in B(0) \\ 2^{N-1}\,, & \text{for } \omega \in B(1) \\ 2^{N-j}\,, & \text{for } \omega \in B(j) \end{cases} \tag{6}$$

ESANN 2025 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 23-25 April 2025, i6doc.com publ., ISBN 9782875870933. Available from http://www.i6doc.com/en/.

Encoding	Coefficients ($k = 1, 2, \ldots$)	Spectrum $\omega \in \Omega$	Coefficients b_ω
Hamming	$c_k = \frac{1}{2}$	$\{-N, -N+1, \ldots, N\}$	$4^{-N}\binom{2N}{N+\omega}$
Binary	$c_k = \frac{1}{2} \cdot 2^{k-1}$	$\{-2^N + 1, \ldots, 2^N - 1\}$	$4^{-N}(2^N - \lvert\omega\rvert)$
Ternary	$c_k = \frac{1}{2} \cdot 3^{k-1}$	$\{-\lfloor \frac{3^N}{2} \rfloor, \ldots, \lfloor \frac{3^N}{2} \rfloor\}$	$4^{-N} A(\omega, N)$

Table 1: Encoding schemes and frequency content of the quantum kernel. The degeneracy of each frequency ω is given by the value of the coefficient b_ω in the expansion in Eq. (5). $\lfloor x \rfloor$ represents the largest integer less or equal to x and $A(\omega, N)$ is defined in Eq. (6).

where $B(0) \coloneqq 0$, $B(1) \coloneqq \{3^\alpha | \alpha \in \{0, 1, \ldots, N-1\}\}$ and for $j \geq 2$

$$B(j) \coloneqq \{\pm 3^{\alpha_1} \pm \cdots \pm 3^{\alpha_j} \,|\, 0 \leq \alpha_n \leq N-1 \text{ and } \alpha_n < \alpha_m \text{ for } 1 \leq m < n \leq j\}. \quad (7)$$

For all encodings, the (real-valued) coefficients resulting from the trigonometric expansion of Eq. (4) have the property $b_\omega = b_{-\omega}$, hence $B_\omega = 2b_\omega$ for $\omega \neq 0$, and $B_0 = b_0$. Interestingly, the kernel has the same Fourier spectrum as the corresponding quantum model [4]. A generalization to handle multidimensional data x_j is provided by sequential data encoding circuits [5], which, however, break translational invariance of the kernel, i.e., $K(\mathbf{x}, \mathbf{y}) \neq K(\mathbf{x} - \mathbf{y}, 0)$.

The typical shape of the kernel is illustrated in Figure 2(a,b). Asymptotically for large N, the kernel becomes insensitive to the data as it approaches a constant $K(x, y) \to 1/2^N$ for $x \neq y$, while still $K(x, x) = 1$ due to normalization of the state. This concentration effect can be mitigated by scaling the data accordingly [3]. With increasing order of encoding, the spectrum not only covers a broader range of frequencies, i.e., larger expressiveness, but also becomes flatter, giving similar weight to high and low frequency components [see Figure 2(c)].

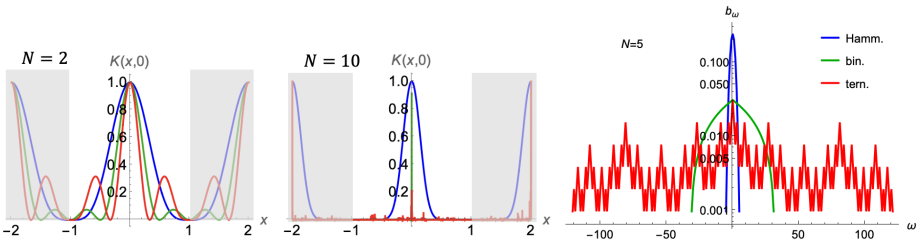


Figure 2: (a,b) Shape of the kernel (4) for $N = 2$ and 10 qubits and various encodings. (c) Fourier amplitudes b_ω of the kernel for various encodings.

ESANN 2025 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 23-25 April 2025, i6doc.com publ., ISBN 9782875870933. Available from http://www.i6doc.com/en/.

Encoding	Eigenvalues λ_s	$\lambda_s =$	corresp. $\omega(s)$	Degeneracy	$\mathcal{N}_\lambda$ (# nonzero)
Hamming	$\{0\} \cup \left\{4^{-N}\binom{2N}{s}\right\},\ s = 0, 1, \ldots, N$	b_{s-N}	$s - N$	2-fold†	$2N + 1$
Binary	$\{0\} \cup \left\{4^{-N}s\right\},\ s = 1, 2, \ldots, 2^N$	b_{2^N-s}	$2^N - s$	2-fold†	$2^{N+1} - 1$
Ternary	$\{0\} \cup \{4^{-N}2^s\},\ s = 0, 1, \ldots, N$	$b_{B(N-s)}$	$B(N-s)$	$2^{N-s}\binom{N}{s}$	3^N

Table 2: Eigenspectrum of the kernel. The possible eigenvalues (indexed by $s \in \mathbb{N}_0$) directly correspond to the Fourier expansion coefficients b_ω in Eq. (5). The mapping between s and ω is not simple in the ternary case due to the non-trivial degeneracy (multiplicity), see Eq. (7). †Exceptions from the 2-fold degeneracy of each eigenvalue occur for $\omega = 0$, which is 1-fold degenerate (corresponding to the largest eigenvalue), and $\lambda = 0$, which has a degeneracy of $4^N - \mathcal{N}_\lambda$. In the ternary case, the 1-fold degeneracy of the largest eigenvalue is accounted for by the stated expression.

1.2 Eigenspectrum

The generalization behavior of a kernel model is controlled by its eigenspectrum [6]. Mercer's theorem allows to diagonalize the kernel with respect to the eigenfunctions of the corresponding kernel-dependent integral operator

$$T[f](y) = \int p(x)K(x,y)f(x)\mathrm{d}x, \tag{8}$$

with data distribution $p(x)$. Hence, if $T[\phi_s](y) = \lambda_s\phi_s(y)$ we get

$$K(x,y) = \sum_s \lambda_s \phi_s(x)\phi_s^*(y) \tag{9}$$

with L_p^2-orthonormal eigenfunctions ϕ_s.

Unless otherwise mentioned, we assume the data to be uniformly distributed on $[-1, 1]$. For $m, n \in \mathbb{Z}$ we get $\frac{1}{2}\int_{-1}^{1}\mathrm{d}x e^{\mathrm{i}\pi m(x-y)}e^{-\mathrm{i}\pi nx} = \delta_{m,n}e^{-\mathrm{i}\pi ny}$, which implies $\phi_s(x) = e^{\mathrm{i}\pi\omega(s)x}$ as kernel eigenfunctions with nonzero eigenvalues $\lambda_s = b_{\omega(s)}$ (see Table 2). The nullspace of T, corresponding to $\lambda = 0$, is infinite-dimensional and consists of functions orthogonal to the ϕ_s including, e.g., all higher-order Fourier modes $e^{\mathrm{i}\pi\rho x}$ with $\rho \notin \Omega$. This directly leads via the first equation in (5) to the Mercer decomposition (9). An alternative set of eigenfunctions is given by $\sin(\pi\omega x)$ and $\cos(\pi\omega x)$, for $\omega > 0$. This follows from the fact that, for $g \in \{\sin, \cos\}$, $\frac{1}{2}\int_{-1}^{1}\mathrm{d}x \cos(\pi m(x-y))g(\pi nx) = \frac{1}{2}\delta_{m,n}g(\pi my)$.

The eigenfunctions of a kernel define the type of target functions that can be learned, which in this case are functions representable by their Fourier expansion up to degree $|\Omega|$. The distribution of kernel eigenvalues, see Table 2(c), indicates the simplicity bias of the model: kernels with Hamming encoding are dominated by a few large eigenvalues and thus able to learn only rather simple functions, whereas kernels with ternary (or higher order) encodings have many small eigenvalues and can thus represent complex functions, at the risk of overfitting noise

ESANN 2025 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 23-25 April 2025, i6doc.com publ., ISBN 9782875870933. Available from http://www.i6doc.com/en/.

[2]. Note that the maximum possible number of nonzero eigenvalues $\mathcal{N}_\lambda = 3^N$ is reached only for ternary and higher-order encodings $c_k \propto m^{k-1}$ with $m > 3$.

2 Kernel regression

We now study the interplay between the expressivity of the kernel, as characterized by its frequency content, and its generalization behavior using the theory of [6, 3] and the results of Section 1.2. To this end, we consider (noiseless) kernel regression on a random Fourier series of degree D,

$$F(x) = \sum_{n=-D}^{D} r_n e^{\mathrm{i}\pi n x} = r_0 + 2\sum_{n=1}^{D} r_n \cos(\pi n x), \qquad r_{-n} = r_n \in \mathbb{R} \tag{10}$$

i.e., we fit the model

$$f(x) = \sum_{j=1} \alpha_j K(x_j, x) \tag{11}$$

via the weights $\boldsymbol{\alpha}$ to a dataset $\{x_j, y_j = F(x_j)\}_{j=1,\ldots,P}$, where the x_j's are uniformly sampled within $[-1, 1]$ and the r_n are fixed Gaussian i.i.d. random numbers with a variance of $\mathcal{O}(1)$. The optimal solution of this problem is given by [2]

$$\boldsymbol{\alpha}^* = \mathbf{y}^T(\mathbf{K} + \lambda \mathbb{1})^{-1}, \quad \text{with} \quad K_{ij} \equiv K(x_i, x_j), \tag{12}$$

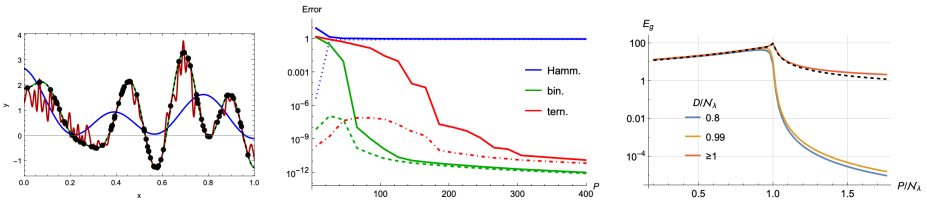
where λ is a regularization parameter. Notably, due to the structure of the kernel [Eqs. (5) and (9)], this is mathematically equivalent to performing Ridge regression with the model $f(x) = \mathbf{w}^T \boldsymbol{\phi}(x)$, where $\mathbf{w} \in \mathbb{R}^{|\Omega|}$ are adjustable weights and the feature vector $\boldsymbol{\phi}(x) = [e^{\mathrm{i}\pi\omega_n x}]_{n=1,\ldots,|\Omega|}$ is constructed from all $|\Omega|$ Fourier modes of the kernel (see Table 1).

Figure 3 illustrates the typical behavior obtained in kernel regression. In the case of $N = 4$ qubits, the Hamming encoding, having a frequency content $|\Omega| = \omega_{\max} = 4 < D$, is not expressive enough to fit a Fourier series with $D = 10$ random coefficients, in contrast to the binary and ternary encoding kernels. If $|\Omega| \geq D$ and the training error is small, it follows from kernel theory [6] that a small generalization error requires $P \gtrsim \mathcal{N}_\lambda$ samples (see Table 2). This implies that low-order encoding schemes can achieve good generalization with fewer training samples than higher order ones. As illustrated in Fig. 3(c), a target function that cannot be represented by Eq. (11) or, equivalently, has overlap with eigenfunctions pertaining to a vanishing eigenvalue ($\lambda = 0$), acts similar to noise in the test error E_g and produces a peak at $P = \mathcal{N}_\lambda$. Note the abrupt transition of $E_g(P \to \infty)$ from 0 to a nonzero value (determined by magnitude of overlap with the $\lambda = 0$ eigenmodes) at $D = \mathcal{N}_\lambda$.

3 Conclusions

We find that quantum feature maps based on parallel rotational data encoding and either ground or uniform superposition initial states generally lead to a

ESANN 2025 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 23-25 April 2025, i6doc.com publ., ISBN 9782875870933. Available from http://www.i6doc.com/en/.



(a) (b) (c)

Figure 3: (a) Illustration of under-/ overfitting behavior in kernel regression for Hamming, binary, and ternary encodings ($N = 5$ qubits and $P = 150$ samples). (b) Train (broken lines) and test error (solid lines) for regression with the kernel in Eq. (4) for $N = 4$ qubits as a function of sample number P. The target function (dashed in (a)) is a random Fourier series of degree $D = 10$. The regularization parameter is $\lambda = 10^{-5}$, but quantitatively similar results are obtained for any value $0 < \lambda \ll 1$. (c) Sample number dependence of the test error E_g (for a binary encoding kernel) for varying degrees D of the target function [Eq. (10)] (relative to the number of nonzero eigenvalues $\mathcal{N}_\lambda$). The dashed curve shows E_g for a noisy target (noise variance 1, $D = 10$).

'universal' form of the kernel [Eq. (4)]. Inspired by previous studies of quantum ML models [7], we focused here on the *encoding scheme* as our main source of expressivity and its effect on the generalization error. We have determined the frequency and eigen-spectra of the kernel for various encodings and pointed out some similarities to conventional quantum models [7, 4]. Focusing on kernel regression, we characterized the trade-off between expressivity, generalization ability, and sample number.

References

[1] Maria Schuld. Supervised quantum machine learning models are kernel methods. arXiv, 2101.11020, 2021.

[2] Bernhard Schölkopf and Alexander J. Smola. Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. The MIT Press, December 2001.

[3] Akash Canatar, Eric Peters, Cengiz Pehlevan, and Stefan Wild. Bandwidth enables generalization in quantum kernel models. Trans. Mach. Learn. Res., 2023, 2022.

[4] Eric Peters and Maria Schuld. Generalization despite overfitting in quantum machine learning models. Quantum, 6:777, 2022.

[5] Adrián Pérez-Salinas, Alba Cervera-Lierta, Elisa Gil-Fuster, et al. Data re-uploading for a universal quantum classifier. Quantum, 4:226, 2020.

[6] Abdulkadir Canatar, Blake Bordelon, and Cengiz Pehlevan. Spectral bias and task-model alignment explain generalization in kernel regression and infinitely wide neural networks. Nat. Comm., 12(1), 2021.

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