Benign Overfitting with Quantum Kernels

Joachim Tomasi[1,2] **Sandrine Anthoine**[2] **Hachem Kadri**[1]

[1]Aix-Marseille Univ, CNRS, LIS, Marseille, France.
[2]Aix Marseille Univ, CNRS, I2M, Marseille, France.

Abstract

Kernel methods compare inputs through feature maps. Quantum kernels follow the same principle: input data are encoded into quantum states, which define *quantum* feature representations in Hilbert spaces. Kernel values are then obtained by estimating inner products between these states using suitable quantum circuit measurements. As a result, quantum kernels may be intractable to compute classically[1]while remaining efficiently computable on quantum hardware, potentially leading to a quantum advantage. However, designing effective quantum kernels remains a major challenge. Many quantum kernels, such as the fidelity kernel, suffer from exponential concentration. This results in near-identity kernel matrices that fail to capture meaningful data correlations and lead to overfitting and poor generalization. In this paper, we propose a novel strategy for constructing quantum kernels that achieve good generalization performance, drawing inspiration from benign overfitting in classical machine learning. We introduce the concept of *Local-Global* quantum kernels, which combine two components: a local quantum kernel based on measurements of small subsystems, and a global quantum kernel derived from full-system measurements. To support the effectiveness of the proposed construction, we show theoretically and empirically that Local-Global quantum kernels exhibit benign overfitting.

1 INTRODUCTION

Quantum machine learning (QML) has recently attracted significant attention [Biamonte et al., 2017, Ciliberto et al., 2018, Huang et al., 2021, Cerezo et al., 2022], particularly as Noisy Intermediate-Scale Quantum (NISQ) devices have become increasingly accessible [Preskill, 2018]. Variational Quantum Algorithms (VQAs), which operate under a hybrid quantum-classical paradigm, have received considerable interest, mainly due to the possibility of achieving quantum advantage under NISQ constraints [Peruzzo et al., 2014, Farhi et al., 2014, Du et al., 2020]. These algorithms utilize Parameterized Quantum Circuits (PQCs) to compute expectation values of quantum circuit measurements, which are then processed classically to update circuit parameters and optimize cost functions, enabling PQCs to serve as machine learning models [Benedetti et al., 2019, Mitarai et al., 2018].

Quantum kernels, such as quantum fidelity, offer alternative approaches that align with the constraints of NISQ-compatible hardware [Havlíček et al., 2019, Schuld and Killoran, 2019, Schuld, 2021]. Data are encoded into quantum systems through parameterized unitary gates that maps classical data into quantum states. This mapping, known as a quantum feature map, is the basis of the concept of *quantum kernels*: by encoding classical data into quantum feature states, the quantum fidelity kernel function is defined as the squared magnitude of the inner product between the feature states. This can be regarded as a quantum analog of the kernel trick. All quantum operations in the quantum kernel space remain linear, akin to classical kernel methods that exploit linear separability in kernel feature spaces [Havlíček et al., 2019, Schuld and Killoran, 2019]. The kernel trick enables nonlinear decision boundaries in the original data space while avoiding explicit computation of high-dimensional features.

While VQAs optimize explicit parameterized models, quantum kernels act as their implicit counterparts, leveraging the inner products of embedded quantum states. This duality establishes a theoretical bridge between certain VQA architectures and kernel methods [Schuld, 2021, Huang et al., 2021, Jerbi et al., 2023]. Quantum kernel methods are founded on the potential to achieve a computational advantage in feature representations [Mengoni and Di Pierro, 2019]. In

[1]Here, 'classical' refers to methods or computations that do not rely on quantum computing resources.

Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence (UAI 2026), PMLR 337:6740–6766.

this paradigm, certain quantum feature maps are proposed to encode data into high-dimensional Hilbert spaces that are classically intractable but remain efficiently computable on quantum hardware [Liu et al., 2021, Kübler et al., 2021, Havlíček et al., 2019, Schuld and Killoran, 2019]. For example, as shown in Liu et al. [2021], specific quantum circuits can generate feature maps with provable classical intractability under complexity-theoretic assumptions, potentially enabling learning tasks that are beyond the reach of conventional classical methods. Furthermore, recent work has shown that quantum kernels can approximate broad classes of functions [Gil-Fuster et al., 2024b].

However, despite their computational promise, it remains unclear whether quantum advantages in quantum kernels directly translate into practical improvements in learning performance. One concern is that quantum-enhanced features do not necessarily capture the intrinsic structure of the target learning problems. While these quantum models exhibit high expressivity, they often give rise to challenges such as the exponential concentration of kernel values near zero, resulting in kernel matrices that closely approximate the identity matrix [Thanasilp et al., 2024]. Such behavior undermines the model's ability to capture meaningful data correlations and thus limits its generalization capability.

To achieve generalization in the context of quantum kernels, some research has focused on mitigating overfitting through regularization techniques that avoid mere training set interpolation. For example, methods for tuning kernel bandwidths in quantum kernels and reducing the dimensionality of the quantum feature space have been proposed [Shaydulin and Wild, 2022, Canatar et al., 2023, Huang et al., 2021, Kübler et al., 2021]. Another promising direction for mitigating overfitting is to leverage PQCs to learn task-adapted quantum feature maps with the aim of improving generalization performance in quantum kernel methods [Hubregtsen et al., 2022, Rodriguez-Grasa et al., 2025].

In this work, we study how to enable generalization in interpolating quantum kernel machines, drawing inspiration from recent advances in the modern classical literature where overparameterized models could generalize well despite achieving near-zero training error. Very few studies have examined the generalization abilities of overparameterized quantum models [Gil-Fuster et al., 2024a]. Larocca et al. [2023] have investigated the generalization of overparameterized quantum neural networks (QNNs), while Kempkes et al. [2026] and Kamisoyama et al. [2026] have studied the double descent phenomenon in least-squares linear regression within quantum features. Peters and Schuld [2023] have examined benign overfitting in linear quantum models for uniformly spaced training data. In our work, we advance this line of research by proposing a framework for quantum kernels that inherently promotes *benign overfitting*. Drawing on the concept of spiky-smooth kernels [Haas et al., 2023], we propose a *Local-Global* quantum kernel constructed as a weighted sum of two quantum kernels. One component is derived from a local measurement (i.e., low-dimensional relative to the full quantum embedding space) that mimics the smooth kernel behavior, while the other comes from a global measurement of the entire quantum feature space, mimicking the spiky component. This Local-Global quantum kernel allows the learned model to have good generalization even when it interpolates training data, giving rise to benign overfitting.

Contributions. This work introduces benign overfitting as a design principle for quantum kernels. In classical machine learning, benign overfitting is typically studied as a phenomenon explaining how highly overparameterized models can generalize despite interpolating training data [Bartlett et al., 2020]. In quantum kernel methods, however, interpolation can arise naturally from kernel concentration: as the number of qubits increases, highly expressive quantum kernels may approach the identity matrix [Thanasilp et al., 2024]. Rather than treating this concentration solely as a drawback, we show that it can be made useful. Our Local-Global construction combines a spiky global component, which induces localized peaks and enables interpolation, with a smooth local component, which preserves meaningful correlations and supports generalization. This provides, to the best of our knowledge, the first theoretical and empirical characterization of benign overfitting for quantum kernels. From a classical perspective, our analysis extends spiky-smooth benign-overfitting mechanisms beyond shrinking-bandwidth constructions [Haas et al., 2023] by showing that localization can instead arise from power concentration of the global kernel component. This perspective broadens the scope of benign-overfitting analyses to other families of kernels, including periodic translation-invariant kernels, for which localization is not naturally captured by bandwidth shrinkage.

The main contributions are as follows:

- We introduce *Local-Global quantum kernels*, which combine local subsystem and global full-system measurements to obtain a quantum spiky-smooth structure: the global component produces localized peaks, while the local component preserves smooth correlations (Definition 3.1).
- We prove that, for a class of periodic translation-invariant quantum kernels, ridgeless regression with the Local-Global kernel can interpolate noisy training data while remaining statistically consistent (Theorems 3.2 and D.1).
- We provide quantum circuit implementations of the proposed Local-Global kernel construction, including both fully quantum and hybrid quantum-classical implementations (Section 3.3).
- We empirically demonstrate on synthetic and real-

world datasets that the proposed quantum kernel construction can interpolate the training data while achieving good generalization performance (Section 4 and Appendix F). We release the code at `github.com/jotomasi/benign-overfitting-with-quantum-kernels`.

2 BACKGROUND

We recall here prior results on benign overfitting in kernel regression and quantum kernels that underpin the framework considered in this paper.

2.1 BENIGN OVERFITTING IN KERNEL REGRESSION

Let $\mathcal{X} \subset \mathbb{R}^d$ be a bounded domain and let μ be a probability measure on $\mathcal{X}$ admitting a density with full support[2]. We observe an i.i.d. dataset $\mathcal{D}_n = \{(x_i, y_i)\}_{i=1}^n$, generated according to

$$x_i \sim \mu, \qquad y_i = f^\star(x_i) + \varepsilon_i,$$

where $f^\star$ is an unknown target function and ε_i are independent, zero-mean random variables with finite variance.

Kernel Regression (KR) extends linear ridge regression by mapping input data x in $\mathcal{X}$ into a high-dimensional Hilbert space $\mathcal{H}_k$ via a feature map $\phi : \mathcal{X} \to \mathcal{H}_k$. $\mathcal{H}_k$ is a Reproducing Kernel Hilbert Space (RKHS) defined through a positive definite kernel k satisfying $k(x, z) = \langle \phi(x), \phi(z) \rangle_{\mathcal{H}}$. Kernel regression seeks a function $f \in \mathcal{H}_k$ minimizing the regularized empirical risk

$$\mathcal{L}_\lambda(f, \mathcal{D}_n) := \frac{1}{n} \sum_{i=1}^n \big(y_i - f(x_i)\big)^2 + \lambda \|f\|_{\mathcal{H}_k}^2, \quad (1)$$

where $\lambda > 0$ is the ridge regularization parameter. By the representer theorem, the solution admits the expansion

$$\hat{f}(x) = \sum_{i=1}^n \alpha_i k(x_i, x), \quad \text{with } \alpha = (K + \lambda I_n)^{-1} Y,$$

where $K \in \mathbb{R}^{n \times n}$ is the kernel matrix ($K_{ij} = k(x_i, x_j)$) and $Y = (y_1, \ldots, y_n)^\top$. The case $\lambda = 0$ is referred to as *ridgeless regression*. In this case, the kernel regression problem reduces to solving the linear system $K\alpha = Y$. When K is not invertible, the ridgeless solution is defined as the minimum-norm interpolant

$$\alpha = K^+ Y,$$

where K^+ denotes the Moore-Penrose pseudoinverse of K. If the kernel matrix has sufficiently high rank so that Y lies in its column space, this estimator exactly interpolates the training data. Throughout the paper, we assume that $f^\star$ is in $\mathcal{H}_k$.

[2] i.e., $\forall x \in \mathcal{X},\ \forall t > 0, \quad \mu(B(x,t)) > 0$, where $B(x,t)$ is the Euclidean ball of radius t centered at x.

From generalization of overparameterized models to benign overfitting in kernel regression Overparameterized neural networks can demonstrate strong generalization capabilities, even when they nearly interpolate training data, challenging the classical bias-variance tradeoff. This observation has prompted a reevaluation of generalization in modern machine learning [Zhang et al., 2021], leading to various interpretations of the phenomenon. One such interpretation is benign overfitting, or harmless interpolation, where models can overfit the training data without compromising their ability to generalize [Bartlett et al., 2020, Muthukumar et al., 2020]. In this context, *overfitting* refers to a model trained to (nearly) interpolate the training data, while the terms *benign* or *harmless* describe scenarios in which this overfitting does not negatively impact the model's generalization ability.

Motivated by the desire to understand this behavior in deep networks, early works explored benign overfitting in simpler models. Notably, Bartlett et al. [2020] studied benign overfitting in overparameterized linear regression, demonstrating that this phenomenon is not limited to deep networks. Furthermore, Mallinar et al. [2022] refined the classification of overfitting by distinguishing between benign, tempered, and catastrophic regimes, which are based on the behavior of the expected risk of interpolating predictors. This phenomenon has also been observed in kernel regression [Belkin et al., 2019, Liang and Rakhlin, 2020], which can be seen as overparameterized linear regression in kernel feature spaces. When the input (or feature) dimension increases with the number of data points, the effective dimension of the kernel matrix grows, and its eigenvalues decay slowly enough that noise gets spread across many less important directions [Bartlett et al., 2020, Tsigler and Bartlett, 2023, Bartlett et al., 2021, Hastie et al., 2022]. In this context, the minimum-norm interpolant, i.e., the ridgeless regression solution that minimizes its norm, can memorize noise in redundant directions while still capturing the underlying signal. This behavior has been rigorously characterized in linear regression and extended to kernel methods [Tsigler and Bartlett, 2023, Zhang et al., 2025, Mallinar et al., 2022]. In fixed-dimensional settings, some studies showed that benign overfitting can occur under specific conditions, such as constraints on the kernel's eigenvalue spectrum [Barzilai and Shamir, 2024, Cheng et al., 2024, Mallinar et al., 2022] or on the kernel function form [Haas et al., 2023].

The intuition behind benign overfitting in both linear and kernel regression is provided by the *simple-spiky* decomposition of the minimum-norm interpolant [Bartlett et al., 2021]. In this view, the learned predictor decomposes into a *simple* component that captures the underlying signal and a *spiky* component helping to interpolate training data without significantly affecting predictions away from the training data [Li et al., 2023, Tsigler and Bartlett, 2023, Mallinar et al., 2022, Medvedev et al., 2024, Liang and Rakhlin,

2020].

Building on this intuition, Haas et al. [2023] introduced *spiky-smooth kernels* of the form $k_{\rho,\gamma}(x,z) = \tilde{k}(x,z) + \rho\,\hat{k}_\gamma(x,z)$, where $\tilde{k}$ is a smooth universal kernel and $\hat{k}_\gamma$ is a spiky kernel, enabling benign overfitting in fixed-dimensional setting. This simple-spiky perspective will guide our construction of quantum kernels.

2.2 QUANTUM KERNELS

Notation According to the bra-ket notation adopted in quantum computing, column vectors are denoted as a "ket", $|\cdot\rangle$, while row vectors are represented as a "bra", $\langle\cdot|$. These two are dual to each other, with the bra defined as $\langle\cdot| := |\cdot\rangle^\dagger$, where $\dagger$ denotes the adjoint (or conjugate transpose). The inner product of two states ψ and ϕ is written as $\langle\psi|\phi\rangle$, and their tensor product is denoted either as $|\psi\phi\rangle$, $|\psi\rangle\,|\phi\rangle$, or equivalently as $|\psi\rangle\otimes|\phi\rangle$. Also, the t-fold tensor product of a state $|\psi\rangle$, denoted by $\bigotimes_{i=1}^{t}|\psi\rangle$, can be compactly expressed as $|\psi\rangle^{\otimes t}$ or $|\psi^t\rangle$.

Quantum states can also be represented using density matrices. For a pure state $|\psi\rangle$, the density matrix is defined as $\rho = |\psi\rangle\langle\psi|$. For single-qubit systems the states $\{|0\rangle, |1\rangle\}$ define the computational basis of $\mathbb{C}^2$, i.e. $|0\rangle = \begin{pmatrix}1 & 0\end{pmatrix}^T$ and $|1\rangle = \begin{pmatrix}0 & 1\end{pmatrix}^T$. For n-qubit systems, the computational basis is generated by the states $|i\rangle$ for $i \in \{0,1,\ldots,2^n-1\}$, which corresponds to the tensor product of individual qubit states derived from the bit decomposition of i, namely if $i = \sum_{k=0}^{n-1} i_k 2^k$ with $i_k \in \{0,1\}$, then $|i\rangle := |i_0\rangle\otimes\cdots\otimes|i_{n-1}\rangle$. Quantum observables are represented by Hermitian operators, and the expectation value of an observable O in the state ρ is given by $\langle O\rangle_\rho = \mathrm{Tr}(\rho O)$. For clarity, the subscript ρ is often omitted when the context is clear.

Quantum circuits provide a diagrammatic representation of quantum computations. They consist of a sequence of quantum gates – unitary operations applied to qubits – followed by measurements of quantum observables. These circuits offer a structured framework for manipulating quantum states and designing quantum algorithms. For a comprehensive introduction to quantum computing and quantum circuits, readers may refer to Nielsen and Chuang [2010].

Quantum encoding A quantum machine learning algorithm needs data in the form of quantum states. So classical data should be first encoded into quantum states, i.e., the transformation of a classical data x to a quantum state $|\phi_x\rangle$. Most of the interest in quantum kernels comes from the observation that encoding classical data into a quantum computer defines an explicit feature representation of the data. Consider a quantum feature map that encodes a data point x into a quantum state represented by the density matrix ρ_x. A quantum kernel $k(x,z)$ is defined as the Hilbert-Schmidt

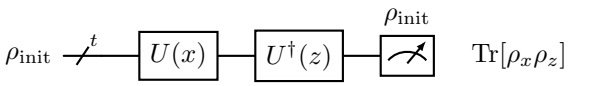


Figure 1: Quantum circuit for evaluating the quantum kernel. The t-qubit register is initialized in the reference state ρ_{init}. Given inputs x and z, the circuit prepares $\sigma_{x,z} = U^\dagger(z)U(x)\rho_{\text{init}}U^\dagger(x)U(z)$. Measuring ρ_{init} yields the expectation value $\mathrm{Tr}[\sigma_{x,z}\rho_{\text{init}}] = \mathrm{Tr}[\rho_x\rho_z]$, by cyclicity of the trace, where $\rho_x = U(x)\rho_{\text{init}}U^\dagger(x)$. This equals the kernel value $k(x,z)$. For pure-state encoding, this reduces to $k(x,z) = |\langle\phi_x|\phi_z\rangle|^2$.

inner product between the states ρ_x and ρ_z, given by

$$k(x,z) = \mathrm{Tr}[\rho_x\rho_z]. \tag{2}$$

Figure 1 depicts a circuit implementing the quantum kernel where each data point x is encoded as a density matrix of the form $\rho_x = U(x)\,\rho_{\text{init}}\,U^\dagger(x)$.

In the case where ρ_{init} is a pure state, for example $\rho_{\text{init}} = |0^t\rangle\langle 0^t|$ the quantum kernel in (2) simplifies to

$$k(x,z) = |\,\langle\phi_x|\phi_z\rangle\,|^2 = |\,\langle 0^t|U^\dagger(z)U(x)|0^t\rangle\,|^2. \tag{3}$$

This corresponds to measuring the fidelity between the quantum states $|\phi_x\rangle$ and $|\phi_z\rangle$, making this kernel known as the quantum fidelity kernel [Schuld and Killoran, 2019, Mengoni and Di Pierro, 2019]. Table 1 summarizes the classical-quantum correspondence for kernel methods.

Recent studies analyzed generalization error bounds for learning with quantum fidelity kernels and the results appear to be negative [Huang et al., 2021, Kübler et al., 2021]. The expressive power of quantum models can hinder generalization. Finding suitable quantum kernels is therefore challenging, since highly expressive quantum embeddings may induce a geometry that is poorly aligned with the learning task [Huang et al., 2021]; in particular, kernel values may concentrate as the number of qubits increases [Thanasilp et al., 2024]. In other words, when using a large number of qubits, the kernel matrix gets close to the identity matrix, resulting in overfitting and poor generalization performance [Suzuki et al., 2024]. To tackle this issue, Canatar et al. [2023] proposed the use of a bandwidth parameter for quantum fidelity kernels, while Huang et al. [2021], Kübler et al. [2021] considered constructions based on reduced density matrices.

3 BENIGN OVERFITTING WITH LOCAL-GLOBAL QUANTUM KERNELS

In this section, we introduce our main contribution: Local-Global quantum kernels, which are quantum analogs of

Table 1: Classical-quantum correspondence for kernel methods. $\mathcal{X}$ is the input space, $\mathcal{F}$ is a classical feature Hilbert space, and $\mathcal{H}_t = (\mathbb{C}^2)^{\otimes t}$ is the Hilbert space of t qubits. We denote by $\mathcal{D}(\mathcal{H}_t) = \{\rho : \mathcal{H}_t \to \mathcal{H}_t \mid \rho = \rho^\dagger,\ \rho \succeq 0,\ \mathrm{Tr}(\rho) = 1\}$ the set of t-qubit states represented as density matrices. In quantum kernel methods, a circuit $U(x)$ encodes a classical input x into a quantum state ρ_x. Similarities are then computed using the Hilbert-Schmidt inner product between density matrices.

Concept	**Classical**	**Quantum**	**Interpretation**
Input data	$(x, z) \in \mathcal{X} \times \mathcal{X}$	$(x, z) \in \mathcal{X} \times \mathcal{X}$	Inputs are classical data points in both settings.
Feature space	$\mathcal{F}$	$\mathcal{D}(\mathcal{H}_t)$	Classical features live in a Hilbert space; quantum features are density matrices acting on the t-qubit Hilbert space.
Feature map	$x \mapsto \phi(x) \in \mathcal{F}$	$x \mapsto \rho_x \in \mathcal{D}(\mathcal{H}_t)$	A feature map sends each input to a feature representation.
Feature representation	$\phi(x)$	ρ_x	The quantum state ρ_x plays the role of the feature representation.
Implementation	$\phi : \mathcal{X} \to \mathcal{F}$	$\rho_x = U(x)\rho_{\text{init}}U(x)^\dagger$	We consider quantum feature maps implemented by data-dependent unitary circuits $U(x)$.
Kernel function	$k(x, z) = \langle \phi(x), \phi(z) \rangle_{\mathcal{F}}$	$k(x, z) = \mathrm{Tr}[\rho_x \rho_z]$	Both kernels are inner products between feature representations; in the quantum case this is the Hilbert-Schmidt inner product.

spiky-smooth kernels that naturally promote benign overfitting. These kernels explicitly decompose into smooth and spiky components while remaining compatible with the quantum circuit-based measurement formalism for kernel evaluation.

3.1 LOCAL-GLOBAL QUANTUM KERNEL

A central challenge in quantum kernel methods is the trade-off between expressivity and generalization as the number of qubits increases. When kernels are constructed from similarities between full quantum feature states, inner products tend to concentrate around zero in high-dimensional Hilbert spaces and consequently yield kernel matrices that approach the identity [Thanasilp et al., 2024]. Such *spiky* kernels enable interpolation of training data but may harm generalization. In contrast, kernels constructed from reduced quantum states of a subsystem retain a smoother structure and thereby preserve nontrivial correlations between distinct inputs and can better capture the underlying signal [Huang et al., 2021].

Motivated by this phenomenon and by the spiky-smooth decomposition underlying benign overfitting in classical kernel regression [Haas et al., 2023], we introduce a *Local-Global quantum kernel* that combines: (i) a *global* kernel, defined from similarities between full quantum feature states and exhibiting concentration effects, and (ii) a *local* kernel, defined from similarities between reduced quantum feature states on a subsystem and inducing smoother behavior. This construction combines a spiky and a smooth kernel component and should support both expressivity and generalization.

Definition 3.1 (Local–Global Quantum Kernel)**.** *Let $U(x)$ be a unitary operator acting on t qubits. Define the local and global quantum feature maps*

$$\rho_x^L := U(x)\Big(L_s \otimes \frac{1}{2^{t-s}} I_{t-s}\Big)U^\dagger(x), \rho_x^G := U(x)G_t U^\dagger(x),$$

where L_s is a fixed pure quantum state (rank-one projector) on $s < t$ qubits and G_t is a fixed pure quantum state on t qubits. The Local-Global quantum kernel *is defined as*

$$k_{LG}(x, z) = \lambda_L\, k_L(x, z) + \lambda_G\, k_G(x, z), \tag{4}$$

with

$$k_L(x, z) := \mathrm{Tr}\big[\rho_x^L \rho_z^L\big], \qquad k_G(x, z) := \mathrm{Tr}\big[\rho_x^G \rho_z^G\big],$$

where $\lambda_L, \lambda_G \geq 0$ are scalar weights.

The terminology *local* and *global* refers to the structure of the quantum measurement underlying the kernel evaluation. The global kernel compares full t-qubit quantum feature states, whereas the local kernel compares reduced states obtained by tracing out $t - s$ qubits and retaining only a subsystem of size s.

More formally, the *global kernel* is constructed from G_t which is a rank-one projector, i.e., $G_t = |\psi\rangle\langle\psi|$. For instance, choosing

$$G_t = \rho_{\text{init}}^G = |0^t\rangle\langle 0^t|$$

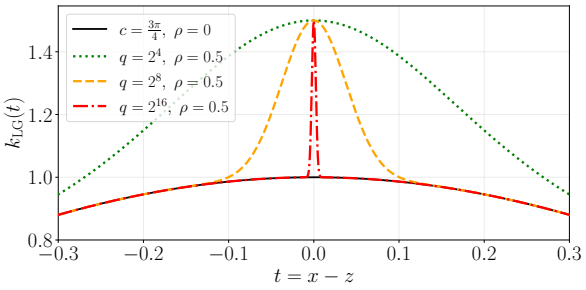


Figure 2: Local-Global kernels with $k_{LG}(x,z) = \cos^2(\frac{c(x-z)}{2}) + \rho\cos^{2q}(\frac{c(x-z)}{2})$, where $c = \frac{3\pi}{4}$, $\rho = 0.5$ and $q = 2^4, 2^8, 2^{16}$. The kernel $k(x,z) = \cos^2(\frac{c(x-z)}{2})$ is the fidelity quantum kernel obtained by angle quantum encoding [Canatar et al., 2023].

yields the standard quantum fidelity kernel as defined in (3). In contrast, the *local kernel* is obtained by initializing only a subset of $s < t$ qubits in a pure state, while the remaining qubits are maximally mixed, i.e.,

$$\rho^L_{\text{init}} = L_s \otimes \frac{1}{2^{t-s}} I_{t-s}. \tag{5}$$

Letting $\rho^L_x = U(x)\rho^L_{\text{init}}U^\dagger(x)$ and using the cyclic property of the trace together with properties of the partial trace [Filipiak et al., 2018], we obtain

$$\begin{aligned} k_L(x,z) &= \mathrm{Tr}\big[\rho^L_x \rho^L_z\big] \\ &= \mathrm{Tr}\big[U^\dagger(z)U(x)\rho^L_{\text{init}}U^\dagger(x)U(z)\rho^L_{\text{init}}\big] \\ &= \mathrm{Tr}\big[\sigma_{x,z}\,\rho^L_{\text{init}}\big] = \frac{1}{2^{t-s}}\mathrm{Tr}[\widetilde{\sigma}_{x,z}L_s]\,, \end{aligned} \tag{6}$$

where $\sigma_{x,z} := U^\dagger(z)U(x)\rho^L_{\text{init}}U^\dagger(x)U(z)$ and $\widetilde{\sigma}_{x,z}$ denotes its partial trace over the last $t-s$ qubits. This reformulation makes explicit that k_L depends only on local measurements performed on the first s qubits, while the remaining qubits are traced out and do not affect the kernel. As a result, k_L typically yields smoother kernel matrices with non-negligible off-diagonal entries.

The key intuition behind our construction is that as the number of qubits t increases, the global kernel k_G becomes increasingly spiky, whereas the local kernel k_L remains smooth. The Local-Global kernel combines these two regimes: the local component supports generalization when the model class becomes highly expressive with increasing qubits, whereas the global component introduces narrow spikes that promote the interpolation of training data. This realizes a quantum analogue of the spiky–smooth decomposition of Haas et al. [2023] and provides a mechanism for benign overfitting with quantum kernels.

3.2 SEPARABLE GLOBAL ENCODING

To analyze the Local-Global quantum kernel in a simple and analytically tractable setting, we introduce a structured quantum encoding scheme, which makes the spiking mechanism explicit and enables a precise study of benign overfitting. In particular, this construction clarifies how the choice of the subsystem size s, the total number of qubits t, and the encoding unitary $U(\cdot)$ controls the smoothness of the local kernel and the spikiness of the global kernel.

Let $s \geq 1$ and consider that $L_s = |0^s\rangle\langle 0^s|$. Suppose that the total number of qubits satisfies $t = q\,s$ for some integer $q \geq 1$. We assume that both the encoding unitary and the initial state of the global kernel factorize across q blocks of s qubits, i.e.,

$$U(x) = V_s(x)^{\otimes q}, \qquad \rho^G_{\text{init}} := |0^t\rangle\langle 0^t| = (|0^s\rangle\langle 0^s|)^{\otimes q},$$

where $V_s(x)$ acts on s qubits. We refer to this scheme as *Separable Global Encoding*.

Under the separable global encoding scheme, the reduced density matrix of the local quantum feature map becomes

$$\tilde{\rho}^L_x := \mathrm{Tr}_{t-s:t}[\rho^L_x] = V_s(x)\,|0^s\rangle\langle 0^s|\,V_s^\dagger(x),$$

which defines a *base kernel* on s qubits

$$k_B(x,z) := \mathrm{Tr}[\tilde{\rho}^L_x\tilde{\rho}^L_z] = |\,\langle 0^s|V_s(x)V_s^\dagger(z)|0^s\rangle\,|^2. \tag{7}$$

This base kernel is itself a quantum fidelity kernel, acting on a smaller subsystem than the global kernel and captures low-dimensional structure of the data.

Using the tensor structure and linearity of the trace, the local kernel reduces to

$$k_L(x,z) := \mathrm{Tr}[\rho^L_x\rho^L_z] = \frac{1}{2^{t-s}}\,k_B(x,z).$$

Thus, the local kernel is proportional to the base kernel and depends only on the reduced s-qubit subsystem.

Similarly, the global kernel factorizes over the q blocks, i.e.,

$$k_G(x,z) := \mathrm{Tr}[\rho^G_x\rho^G_z] = \mathrm{Tr}[(\tilde{\rho}^L_x\tilde{\rho}^L_z)^{\otimes q}] = k_B(x,z)^q.$$

The global kernel amplifies the spikiness through the exponent q, which grows with the total number of qubits relative to the subsystem size. The Local-Global kernel in Definition 3.1 then reduces to

$$k_{LG}(x,z) = \tilde{\lambda}_L k_B(x,z) + \lambda_G k_B(x,z)^q, \tag{8}$$

where $\tilde{\lambda}_L = \frac{\lambda_L}{2^{t-s}}$.

The Local-Global quantum kernel in this case includes a parameter q, which determines the kernel's degree and serves as a tuning parameter for controlling the behavior of the global component. Figure 2 illustrates this idea. The Local-Global kernel closely follows the local component but deviates in specific narrow regions. As the parameter q increases, these deviations become more localized, reinforcing the analogy with spiky-smooth kernels.

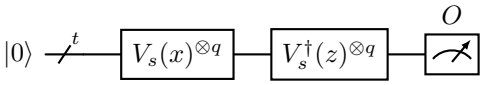


Figure 3: A quantum circuit for computing the Local-Global kernel $k_{LG} = \langle O \rangle$ with $O = \tilde{\lambda}_L O_L + \lambda_G O_G$, $O_L = |0^s\rangle\langle 0^s| \otimes I_s^{\otimes q-1}$ and $O_G = (|0^s\rangle\langle 0^s|)^{\otimes q}$.

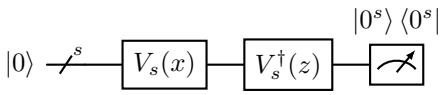


Figure 4: A quantum circuit for computing the base (local) kernel $k_B = \langle |0^s\rangle\langle 0^s| \rangle$. The Local-Global quantum kernel $k_{LG}(x, z)$ is then obtained from k_B using (8).

3.3 QUANTUM CIRCUIT IMPLEMENTATION

We now describe how the Local-Global quantum kernel defined in (8) can be implemented on quantum hardware.

Fully quantum implementation. The reduced operator $\tilde{\sigma}_{x,z}$ associated with the local kernel, see (6), satisfies

$$\tilde{\sigma}_{x,z} := \mathrm{Tr}_{s+1:t}[\sigma_{x,z}] = V_s^\dagger(z)V_s(x)\ |0^s\rangle\langle 0^s|\ V_s^\dagger(x)V_s(z). \tag{9}$$

Using the factorization properties of the trace over tensor products and the fact that $\mathrm{Tr}[\tilde{\sigma}_{x,z}] = 1$, the Local-Global kernel can be expressed as

$$\begin{aligned} k_{LG}(x, z) &= \tilde{\lambda}_L \mathrm{Tr}[\tilde{\sigma}_{x,z} |0^s\rangle\langle 0^s|] + \lambda_G \mathrm{Tr}[\tilde{\sigma}_{x,z} |0^s\rangle\langle 0^s|]^q \\ &= \mathrm{Tr}[\tilde{\sigma}_{x,z}^{\otimes q} O_{LG}], \end{aligned} \tag{10}$$

where the measurement operator O is defined as $O = \tilde{\lambda}_L O_L + \lambda_G O_G$ with $O_L = |0^s\rangle\langle 0^s| \otimes I_s^{\otimes q-1}$ and $O_G = (|0^s\rangle\langle 0^s|)^{\otimes q}$.

This decomposition shows that k_{LG} can be computed entirely through quantum measurements on t qubits using a weighted sum of local and global observables. Figure 3 illustrates the corresponding quantum circuit.

Hybrid classical-quantum implementation. A more resource-efficient approach exploits the separable structure of the kernel. Since $k_{LG}(x, z)$ is fully determined by the base kernel $k_B(x, z)$, we can compute $k_B(x, z)$ using a quantum circuit on only s qubits, and then classically compute the kernel via (8). This hybrid scheme avoids preparing q tensor copies and reduces the required quantum resources from $t = qs$ qubits to just s qubits. Figure 4 shows this implementation.

3.4 THEORETICAL ANALYSIS

We consider *Local-Global quantum kernels* arising from the global separable encoding introduced in Section 3.2, for which the kernel admits the decomposition given in (8). Throughout this section, we fix $\tilde{\lambda}_L = 1$ and $\lambda_G = \rho$.

The objective here is to demonstrate that *quantum kernels can be constructed to yield interpolating estimators that nevertheless remain statistically consistent*, thereby exhibiting *benign overfitting* in the sense of Mallinar et al. [2022], Haas et al. [2023]. Specifically, we show that for a broad class of quantum fidelity kernels, the associated Local-Global kernel induces a ridgeless estimator that interpolates the training data while remaining consistent. For clarity, we present the result for the cosine quantum kernel as base kernel

$$k_B(x, z) = \prod_{i=1}^{d} \cos^2\Big(\tfrac{c(x^{(i)} - z^{(i)})}{2}\Big), \tag{11}$$

which arises from angle encoding [Canatar et al., 2023] (for more details, see Appendix A.1). This concrete instance shows the essential mechanism underlying benign overfitting in the Local-Global quantum kernel construction.

Theorem 3.2 (Benign overfitting for the cosine quantum kernel)**.** *Let $k(x, z)$ be the quantum fidelity kernel defined in* (11)*. Assume that the data are generated according to the model of Section 2.1, and that the input distribution μ admits a density $g \in L^1(\mathbb{R}^d) \cap L^p(\mathbb{R}^d)$ for some $p > 1$. Let $\alpha > 0$ and $0 < \alpha_0 < 1$.*

Let $\hat{f}_{\rho,q,n}$ denote the ridgeless regression estimator trained on n samples with the Local-Global kernel $k + \rho k^q$. Suppose that the real positive sequences $(q_n)_n$ and $(\rho_n)_n$ satisfy

$$n^{\alpha_0}\rho_n \xrightarrow[n\to\infty]{} 0, \qquad n\rho_n \xrightarrow[n\to\infty]{} \infty,$$

$$q_n \geq Cn^\xi \ln n, \qquad C > 6 + \alpha, \qquad \xi > \tfrac{2(2+\alpha)p}{d(p-1)},$$

and with high probability the estimator interpolates the data: $\hat{f}_{\rho_n,q_n,n}(x_i) = y_i, \quad \forall i.$

Then the sequence $\{\hat{f}_{\rho_n,q_n,n}\}_n$ is consistent in probability:

$$\mathbb{P}\Big(\mathbb{E}_x\Big[\big(\hat{f}_{\rho_n,q_n,n}(x) - f^*(x)\big)^2\Big] > \varepsilon\Big) \xrightarrow[n\to\infty]{} 0, \quad \forall \varepsilon > 0.$$

Theorem 3.2 shows that the cosine quantum fidelity kernel admits a Local-Global construction that interpolates the training data while remaining statistically consistent. The sequence (ρ_n) ensures the consistency of the underlying kernel ridge estimator based on k, while the sequence (q_n) concentrates the global component k^{q_n} increasingly tightly around the training inputs. This enables exact interpolation of the sample without degrading generalization performance. Consequently, consistency is achieved in the ridgeless regime.

This phenomenon is not specific to the cosine quantum kernel. In Appendix A.2, we show how to build periodic, translation-invariant quantum fidelity kernels. We then show in Appendices B to D that this whole class of quantum kernels exhibits the same properties as the cosine kernel.

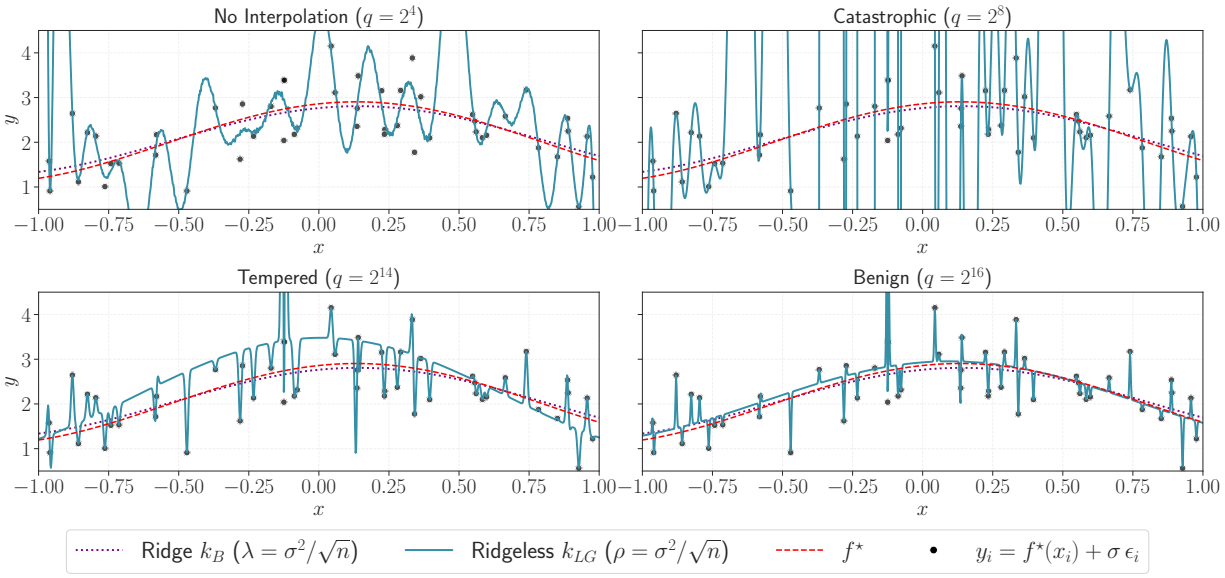


Figure 5: Ridgeless regression using the Local-Global kernel $k_{LG}(x, z) = k_B(x, z) + \rho k_B(x, z)^q$ with different values of q.

Sketch of proof. Benign overfitting requires two properties: interpolation of the training data and consistency of the estimator.

Under the data model of Subsection 2.1, the regression function satisfies $f^* \in \mathcal{H}_k$. Since $\mathcal{H}_k \subseteq \mathcal{H}_{k+\rho_n k^{q_n}}$, the learning problem is well specified for both kernels. Consistency therefore reduces to showing

$$\mathbb{E}_x\left[\left(\hat{f}_{\rho_n,q_n,n}(x) - f^*(x)\right)^2\right] \underset{n\to\infty}{\longrightarrow} 0.$$

Let $\hat{f}_{\rho_n,n}$ denote the kernel ridge estimator with kernel k and regularization parameter ρ_n. We have

$$\mathbb{E}_x[(\hat{f}_{\rho_n,q_n,n}(x) - f^*(x))^2] \le 2\mathbb{E}_x[(\hat{f}_{\rho_n,n}(x) - f^*(x))^2] + 2\mathbb{E}_x[(\hat{f}_{\rho_n,q_n,n}(x) - \hat{f}_{\rho_n,n}(x))^2]. \quad (12)$$

The first term on the right-hand side of (12) corresponds to classical kernel ridge regression with kernel k and vanishing regularization. Standard results guarantee its consistency under the stated conditions on ρ_n [Bach, 2024, chapter 7].

The second term on the right-hand side of (12) quantifies the deviation between the ridgeless estimator based on $k + \rho_n k^{q_n}$ and the ridge estimator based on k. Following the proof technique of Haas et al. [2023], the key observation is that k^q acts as a *highly localized* kernel: for any fixed (x, z) with $0 \le k(x, z) < 1$, we have $k(x, z)^q \to 0$ as $q \to \infty$. As a result, for large q_n, the Hadamard power (element-wise product) $K^{\odot q_n}$ becomes diagonally dominant, and consequently the matrix $K + \rho_n K^{\odot q_n}$ is close to $K + \rho_n I_n$.

To formalize this, we introduce the δ-neighborhood $S(\delta)$ of the spike set $Z = \{(x, z) \in \mathcal{X}^2 : k(x, z) = 1\}$ (see Appendix C). Choosing $\delta_n \to 0$ sufficiently slowly ensures that with high probability no pair of distinct training points lies in $S(\delta_n)$ (Lemma C.5). Outside $S(\delta_n)$, the kernel k is uniformly bounded away from 1 by $m(\delta_n) := \sup_{(x,z)\in S(\delta_n)^c} k(x, z) < 1$, so that for $i \neq j$,

$$k(x_i, x_j)^{q_n} \le m(\delta_n)^{q_n}.$$

Choosing q_n sufficiently large relative to $m(\delta_n)$ ensures that the second term on the right-hand side of (12) vanishes. Since both terms vanish, consistency in probability follows. A complete proof of the theorem for periodic translation-invariant quantum kernels is provided in Appendix D, while Appendix E details its specialization to the cosine quantum kernel.

Comparison with Haas et al. [2023]. In the spiky–smooth framework of Haas et al. [2023], the smooth component is chosen to be universal, while the spiky component admits a bandwidth parameter that controls localization around the diagonal. This shrinking-bandwidth mechanism simplifies the analysis of interpolation and consistency. Such a construction is not available for quantum fidelity kernels. For instance, kernels of the form $k_\gamma(x, z) = \cos^2\left(\frac{x-z}{\gamma}\right)$ are periodic and do not converge as $\gamma \to 0$. The Local-Global construction introduced here circumvents this limitation: the global component k^q induces localization through *power concentration* rather than bandwidth shrinkage, enabling benign overfitting without altering the quantum encoding.

4 NUMERICAL EXPERIMENTS

We evaluate the proposed Local-Global kernel construction (8) with $\tilde{\lambda}_L = 1$ and $\lambda_G = \rho$ with the cosine quantum kernel k_B defined in (11). We compare ridgeless regression

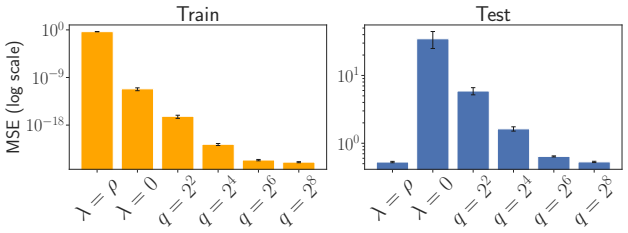


Figure 6: Regression with synthetic data ($n_{\text{train}} = 500$, $d = 7$). Train and test MSE for k_B (ridgeless $\lambda = 0$ & ridge $\lambda = \rho$) and k_{LG} (ridgeless) with varying q.

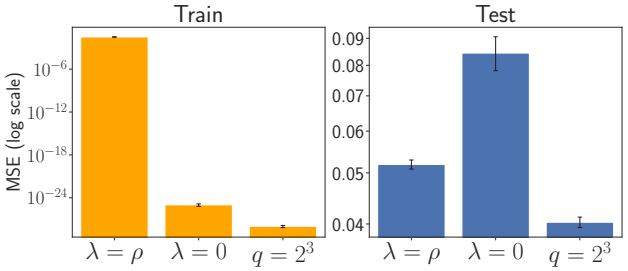


Figure 7: Regression with Airfoil dataset ($n_{\text{train}} = 500$, $d = 20$). Train and test MSE for k_B (ridgeless $\lambda = 0$ & ridge $\lambda = \rho$) and k_{LG} (ridgeless).

with the Local-Global kernel k_{LG} to regression with the kernel k_B, both in the ridgeless regime ($\lambda = 0$) and with ridge regularization ($\lambda = \rho$). Experiments are conducted on synthetic and real-world datasets, with performance measured by mean squared error (MSE) on independent test sets and averaged over multiple random splits. Further details on the datasets and experimental setup are provided in Appendix F.

Illustration of overfitting regimes in 1D We first illustrate the behavior of the Local-Global kernel ridgeless regression in one dimension as the exponent q varies. The target function is constructed as a linear combination of the base kernel k_B evaluated on uniformly sampled points in $[-1, 1]$. Figure 5 reveals four distinct regimes as q increases: (top left) for small q, the estimator fails to interpolate and exhibits large oscillations (poor bias-variance tradeoff); (top right) for intermediate q, interpolation is achieved but strong oscillations between samples lead to severe overfitting; (bottom left) for larger q, interpolation becomes spatially localized while predictions remain smooth away from the data; (bottom right) for very large q, the interpolating predictor remains close to $f^\star$, illustrating benign overfitting. This 1D experiment exposes the full overfitting phase diagram induced by quantum kernel powering in the Local-Global construction.

Evaluation on synthetic and real-world data We next evaluate the Local-Global kernel on a synthetic regression task (see Figure 6) and on the Airfoil real-world dataset (see Figure 7); additional results on other datasets are provided in Appendix F. Ridgeless regression with the quantum kernel k_B interpolates the training data yet fails to generalize. The Local-Global kernel preserves interpolation while maintaining strong generalization, and in some cases even outperforms ridge regression. This improvement does not imply that interpolation is always preferable to regularization. The two estimators search different hypothesis spaces: ridge regression with the local kernel in $\mathcal{H}_L$, and ridgeless regression with the Local-Global kernel in the larger space $\mathcal{H}_{LG}$, with $\mathcal{H}_L \subseteq \mathcal{H}_{LG}$. When the target is better approximated in $\mathcal{H}_{LG}$, the Local-Global estimator can improve test performance while still interpolating the training data.

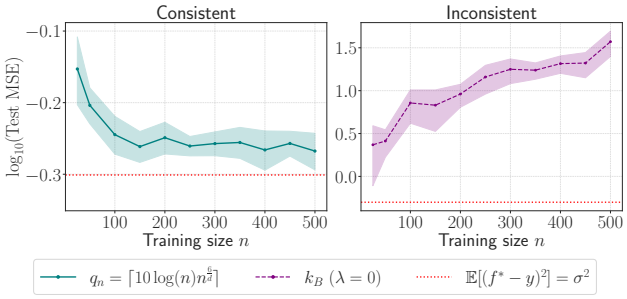


Figure 8: Consistency on synthetic data for ridgeless k_{LG} with $\rho_n = \sigma^2/\sqrt{n}$ and $q_n = \lceil 10 \log(n) n^{6/d} \rceil$. The test error converges to the noise level σ^2. Ridgeless regression with k_B ($\lambda = 0$) is inconsistent.

Consistency Finally, we illustrate the consistency of the Local-Global estimator on the synthetic data from the previous experiment as the training size n increases. Figure 8 shows that the test MSE of the Local-Global kernel decreases and converges to the noise level σ^2, which is in agreement with Theorem 3.2. Ridgeless regression with the cosine quantum kernel clearly exhibits inconsistency.

5 CONCLUSION

We introduced a novel approach to quantum kernel construction, called *Local-Global* kernels, which combines local and global measurements to enable benign overfitting in quantum machine learning. By leveraging separable global encoding, the kernel can be naturally implemented on a quantum device while achieving strong generalization. Our theoretical results establish conditions under which ridgeless regression with this kernel achieves benign overfitting and consistency. This shows that carefully designed quantum features can interpolate data without compromising predictive performance. Empirical evaluations on both synthetic and real-world datasets confirm these findings. While this approach is designed to promote benign overfitting, it also presents a promising strategy for constructing effective quantum kernels.

Acknowledgements

H.K. was supported in part by the French National Research Agency under grant ANR-19-CE23-0011. J.T. acknowledges financial support from the QuanTEdu-France program under grant ANR-22-CMAS-0001. J.T. also thanks Kais Hariz and Ugo Nzongani for helpful discussions.

References

Francis Bach. *Learning theory from first principles*. MIT press, 2024.

Peter L. Bartlett, Philip M. Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48): 30063–30070, 2020.

Peter L. Bartlett, Andrea Montanari, and Alexander Rakhlin. Deep learning: a statistical viewpoint. *Acta numerica*, 30: 87–201, 2021.

Daniel Barzilai and Ohad Shamir. Generalization in kernel regression under realistic assumptions. In *International Conference on Machine Learning (ICML)*, volume 235, pages 3096–3132, 2024.

Mikhail Belkin, Alexander Rakhlin, and Alexandre B. Tsybakov. Does data interpolation contradict statistical optimality? In *International conference on artificial intelligence and statistics (AISTAT)*, pages 1611–1619, 2019.

Marcello Benedetti, Erika Lloyd, Stefan Sack, and Mattia Fiorentini. Parameterized quantum circuits as machine learning models. *Quantum science and technology*, 4(4): 043001, 2019.

Jacob Biamonte, Peter Wittek, Nicola Pancotti, Patrick Rebentrost, Nathan Wiebe, and Seth Lloyd. Quantum machine learning. *Nature*, 549(7671):195–202, 2017.

Abdulkadir Canatar, Evan Peters, Cengiz Pehlevan, Stefan M. Wild, and Ruslan Shaydulin. Bandwidth enables generalization in quantum kernel models. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856.

Marco Cerezo, Guillaume Verdon, Hsin-Yuan Huang, Lukasz Cincio, and Patrick J. Coles. Challenges and opportunities in quantum machine learning. *Nature computational science*, 2(9):567–576, 2022.

Tin Sum Cheng, Aurelien Lucchi, Anastasis Kratsios, and David Belius. Characterizing overfitting in kernel ridgeless regression through the eigenspectrum. In *International Conference on Machine Learning (ICML)*, volume 235, pages 8141–8162, 2024.

Carlo Ciliberto, Mark Herbster, Alessandro Davide Ialongo, Massimiliano Pontil, Andrea Rocchetto, Simone Severini, and Leonard Wossnig. Quantum machine learning: a classical perspective. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 474 (2209):20170551, 2018.

Yuxuan Du, Min-Hsiu Hsieh, Tongliang Liu, and Dacheng Tao. Expressive power of parametrized quantum circuits. *Physical Review Research*, 2(3):033125, 2020.

Edward Farhi, Jeffrey Goldstone, and Sam Gutmann. A quantum approximate optimization algorithm. *arXiv preprint arXiv:1411.4028*, 2014.

Katarzyna Filipiak, Daniel Klein, and Erika Vojtková. The properties of partial trace and block trace operators of partitioned matrices. *The Electronic Journal of Linear Algebra*, 33:3–15, 2018.

Elies Gil-Fuster, Jens Eisert, and Carlos Bravo-Prieto. Understanding quantum machine learning also requires rethinking generalization. *Nature Communications*, 15(1): 2277, 2024a.

Elies Gil-Fuster, Jens Eisert, and Vedran Dunjko. On the expressivity of embedding quantum kernels. *Machine Learning: Science and Technology*, 5(2):025003, 2024b.

Moritz Haas, David Holzmüller, Ulrike Luxburg, and Ingo Steinwart. Mind the spikes: Benign overfitting of kernels and neural networks in fixed dimension. *Advances in Neural Information Processing Systems (NeurIPS)*, 36: 20763–20826, 2023.

Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J. Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *Annals of statistics*, 50 (2):949, 2022.

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, 2019.

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):2631, 2021.

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. *Physical Review A*, 106(4):042431, 2022.

Sofiene Jerbi, Lukas J. Fiderer, Hendrik Poulsen Nautrup, Jonas M. Kübler, Hans J. Briegel, and Vedran Dunjko. Quantum machine learning beyond kernel methods. *Nature Communications*, 14(1):517, 2023.

Kensuke Kamisoyama, Lento Nagano, and Koji Terashi. Double descent in quantum kernel ridge regression. *arXiv preprint arXiv:2604.17202*, 2026.

Marie Kempkes, Aroosa Ijaz, Elies Gil-Fuster, Carlos Bravo-Prieto, Jakob Spiegelberg, Evert van Nieuwenburg, and Vedran Dunjko. Double descent in quantum kernel methods. *PRX Quantum*, 7(1):010312, 2026.

Jonas Kübler, Simon Buchholz, and Bernhard Schölkopf. The inductive bias of quantum kernels. *Advances in Neural Information Processing Systems (NeurIPS)*, 34: 12661–12673, 2021.

Martin Larocca, Nathan Ju, Diego García-Martín, Patrick J. Coles, and Marco Cerezo. Theory of overparametrization in quantum neural networks. *Nature Computational Science*, 3(6):542–551, 2023.

Zhu Li, Weijie J. Su, and Dino Sejdinovic. Benign overfitting and noisy features. *Journal of the American Statistical Association*, 118(544):2876–2888, 2023.

Tengyuan Liang and Alexander Rakhlin. Just interpolate: Kernel "Ridgeless" regression can generalize. *The Annals of Statistics*, 48(3):1329 – 1347, 2020.

Yunchao Liu, Srinivasan Arunachalam, and Kristan Temme. A rigorous and robust quantum speed-up in supervised machine learning. *Nature physics*, 17(9):1013–1017, 2021.

Neil Mallinar, James Simon, Amirhesam Abedsoltan, Parthe Pandit, Misha Belkin, and Preetum Nakkiran. Benign, tempered, or catastrophic: Toward a refined taxonomy of overfitting. *Advances in Neural Information Processing Systems (NeurIPS)*, 35:1182–1195, 2022.

Marko Medvedev, Gal Vardi, and Nathan Srebro. Overfitting behaviour of gaussian kernel ridgeless regression: Varying bandwidth or dimensionality. *Advances in Neural Information Processing Systems (NeurIPS)*, 37:52624–52669, 2024.

Riccardo Mengoni and Alessandra Di Pierro. Kernel methods in quantum machine learning. *Quantum Machine Intelligence*, 1(3):65–71, 2019.

Kosuke Mitarai, Makoto Negoro, Masahiro Kitagawa, and Keisuke Fujii. Quantum circuit learning. *Physical Review A*, 98(3):032309, 2018.

Vidya Muthukumar, Kailas Vodrahalli, Vignesh Subramanian, and Anant Sahai. Harmless interpolation of noisy data in regression. *IEEE Journal on Selected Areas in Information Theory*, 1(1):67–83, 2020.

Michael A. Nielsen and Isaac L. Chuang. *Quantum computation and quantum information*. Cambridge university press, 2010.

Alberto Peruzzo, Jarrod McClean, Peter Shadbolt, Man-Hong Yung, Xiao-Qi Zhou, Peter J. Love, Alán Aspuru-Guzik, and Jeremy L. O'brien. A variational eigenvalue solver on a photonic quantum processor. *Nature communications*, 5(1):4213, 2014.

Evan Peters and Maria Schuld. Generalization despite overfitting in quantum machine learning models. *Quantum*, 7: 1210, 2023.

John Preskill. Quantum Computing in the NISQ era and beyond. *Quantum*, 2:79, August 2018. ISSN 2521-327X.

Pablo Rodriguez-Grasa, Yue Ban, and Mikel Sanz. Neural quantum kernels: training quantum kernels with quantum neural networks. *Physical Review Research*, 7(2):023269, 2025.

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

Maria Schuld and Nathan Killoran. Quantum machine learning in feature hilbert spaces. *Physical review letters*, 122 (4):040504, 2019.

Ruslan Shaydulin and Stefan M. Wild. Importance of kernel bandwidth in quantum machine learning. *Physical Review A*, 106(4):042407, 2022.

Yudai Suzuki, Hideaki Kawaguchi, and Naoki Yamamoto. Quantum Fisher kernel for mitigating the vanishing similarity issue. *Quantum Science and Technology*, 9(3): 035050, 2024.

Supanut Thanasilp, Samson Wang, Marco Cerezo, and Zoë Holmes. Exponential concentration in quantum kernel methods. *Nature communications*, 15(1):5200, 2024.

Alexander Tsigler and Peter L. Bartlett. Benign overfitting in ridge regression. *Journal of Machine Learning Research (JMLR)*, 24(123):1–76, 2023.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021.

Haobo Zhang, Weihao Lu, and Qian Lin. The phase diagram of kernel interpolation in large dimensions. *Biometrika*, 112(1):asae057, 02 2025. ISSN 1464-3510.

Julien Zylberman, Ugo Nzongani, Andrea Simonetto, and Fabrice Debbasch. Efficient quantum circuits for non-unitary and unitary diagonal operators with space-time-accuracy trade-offs. *ACM Transactions on Quantum Computing*, 6(2):1–43, 2025.

Benign Overfitting with Quantum Kernels (Supplementary Material)

Joachim Tomasi[1,2] **Sandrine Anthoine**[2] **Hachem Kadri**[1]

[1]Aix-Marseille Univ, CNRS, LIS, Marseille, France.
[2]Aix Marseille Univ, CNRS, I2M, Marseille, France.

A QUANTUM ENCODING FOR PERIODIC TRANSLATION-INVARIANT QUANTUM KERNELS

This appendix formalizes a class of quantum feature maps inducing periodic, translation-invariant kernels. These quantum kernels yield interpolating estimators that are statistically consistent (and thus benign overfitting) as described in Theorem D.1 and shown in Appendices B to D.

A.1 ILLUSTRATIVE EXAMPLE: ANGLE ENCODING

Consider d qubits initialized in $\rho_0 = (|0\rangle\langle 0|)^{\otimes d}$. Given $x = (x_1, \dots, x_d) \in \mathbb{R}^d$, define single-qubit rotations over the σ_X-Pauli matrix on the jth-qubits

$$U_j(x_j) = \exp\Big(-i\frac{x_j}{2}\sigma_X^{(j)}\Big) = R_X^{(j)}(x_j),$$

and the encoding unitary

$$U_{\cos^2}(x) = \prod_{j=1}^{d} U_j(x_j).$$

The corresponding quantum feature map, expressed as a density matrix, is

$$\rho_{\cos^2}(x) = U_{\cos^2}(x)\rho_0 U_{\cos^2}(x)^\dagger.$$

The induced quantum kernel

$$k_{\cos^2}(x, z) = \mathrm{Tr}\big[\rho_{\cos^2}(x)\rho_{\cos^2}(z)\big]$$

factorizes and satisfies

$$k_{\cos^2}(x, z) = \prod_{j=1}^{d} \cos^2\Big(\frac{x_j - z_j}{2}\Big).$$

For the details of the computations, see Canatar et al. [2023]. This quantum kernel is translation invariant and 2π-periodic in each coordinate.

Let $\tau = (\tau_1, \dots, \tau_d) \in (\mathbb{R}_*^+)^d$. Replacing x_j by $\frac{2\pi x_j}{\tau_j}$ yields the kernel

$$k(x, z) = \prod_{j=1}^{d} \cos^2\Big(\frac{\pi(x_j - z_j)}{\tau_j}\Big),$$

which is periodic with period τ_j in the j-th coordinate. The spike set Z is

$$Z := \{(x, z) \in \mathcal{X}^2 : k(x, z) = 1\} = \{(x, z) \in \mathcal{X}^2 : x - z \in \tau \odot \mathbb{Z}^d\},$$

where $\odot$ is the element-wise product. The spike set is formed of the input data points such that the pairwise difference lies in the lattice parametrized by τ.

A.2 HAMILTONIAN ENCODING

The above construction admits a Hamiltonian representation

$$U_{\cos^2}(x) = \exp\Big(-i\sum_{j=1}^{d} H_j x_j\Big), \qquad \text{with } H_j = \tfrac{1}{2}\,\sigma_X^{(j)}.$$

Since the H_i act nontrivially on different qubits, the H_i commute, i.e. $[H_j, H_k] = 0$. One has $U_{\cos^2}(x)U_{\cos^2}^{\dagger}(z) = U_{\cos^2}(x-z)$ which leads to translation invariance of the kernel.

To obtain a larger class of quantum kernels that have the same translation invariance and periodicity properties, we construct the unitary encoding from its Hamiltonian representation and the quantum feature map in the following way.

Let

$$U(x) = \exp\Big(-i\sum_{j=1}^{d} H_j x_j\Big),\ \rho(x) = U(x)\,|0^m\rangle\langle 0^m|\,U(x)^{\dagger},$$

where $\{H_j\}_{j=1}^{d}$ are Hermitian operators acting on m qubits satisfying

(F1) $[H_j, H_k] = 0$ for all j, k;

(F2) $|0^m\rangle$ is not an eigenstate of H_j for all j;

(F3) For each j, there exists a minimal $2\pi \geq \tau_j > 0$ such that $e^{-iH_j\tau_j}\,|0^m\rangle = e^{-i\theta_j}\,|0^m\rangle$ for some $\theta_j \in \mathbb{R}$.

Under **(F1)**, the induced kernel satisfies

$$k(x,z) = \mathrm{Tr}\big[\rho(x)\rho(z)\big] = \big|\,\langle 0^m|U(x-z)|0^m\rangle\,\big|^2 = \kappa(x-z),$$

and is therefore translation invariant.

Assumption **(F2)** ensures that the kernel function does not depend trivially on each coordinate. For example if $H_j\,|0^m\rangle = \theta_j\,|0^m\rangle$ then $\exp(-iH_j)\,|0^m\rangle = \exp(-i\theta_j)\,|0^m\rangle$. We have $|\,\langle 0^m|\exp\big(-iH_j(x_j - z_j)\big)|0^m\rangle\,|^2 = |\exp(-i\theta_j(x_j - z_j))|^2 = 1$ which is independent of the j-th coordinates of $x - z$.

If we add assumption **(F3)**, we have:

$$\kappa(x-z) = 1 \quad \Longleftrightarrow \quad x_j - z_j \in \tau_j\mathbb{Z} \quad \forall j \in \{1,\dots,d\}.$$

Hence the kernel is periodic with period vector $\tau = (\tau_1, \dots, \tau_d)$ and its spike set is parametrized by a lattice in $\mathbb{R}^d$ determined by the hamiltonians $\{H_j\}$.

Remark A.1. *A simple implementation is obtained by choosing the Hamiltonians H_j as local operators acting on distinct qubits, so that **(F1)** holds automatically. More generally, since commuting Hamiltonians share a common eigenbasis, one can construct efficiently implementable examples by combining an efficiently implementable unitary change of basis with a data-parametrized diagonal operator, which can be implemented efficiently using techniques from Zylberman et al. [2025].*

B SETTING AND NOTATIONS

B.1 FORM OF THE LOCAL KERNEL

Let

$$|\psi\rangle : \mathcal{X} \subset \mathbb{R}^d \to \mathbb{C}^{2^m}, \qquad |\psi(x)\rangle = U(x)\,|0^m\rangle\,,$$

be a quantum feature map with $\|\,|\psi(x)\rangle\,\|_2 = 1$ for all $x \in \mathcal{X}$. Define the associated density matrix

$$\rho(x) = |\psi(x)\rangle\,\langle\psi(x)|\,, \quad \|\rho(x)\|_{\mathrm{HS}} := \sqrt{\mathrm{Tr}[\rho(x)\rho^{\dagger}(x)]} = 1.$$

The induced quantum kernel is

$$k(x,z) = \mathrm{Tr}\big[\rho(x)\rho(z)\big] = \big|\langle\psi(x) \mid \psi(z)\rangle\big|^2. \tag{13}$$

Note that $0 \le k(x,z) \le 1, \forall (x,z)$.

Define the spike set as

$$Z := \{(x,z) \in \mathcal{X}^2 : k(x,z) = 1\}, \tag{14}$$

and for $\delta > 0$ its δ-neighborhood

$$S(\delta) := \Big\{(x,z) \in \mathcal{X}^2 : \inf_{(x',z')\in Z} \|(x,z) - (x',z')\|_2 \le \delta\Big\}. \tag{15}$$

We restrict attention to periodic translation-invariant kernels satisfying:

$$\exists\, \tau \in (\mathbb{R}_*^+)^d \text{ s.t. } \quad k(x,z) = 1 \iff x - z \in \tau \odot \mathbb{Z}^d. \tag{H1}$$

Note that the kernels described in Appendix A all have this property. Under *(H1)*, the spike set admits the explicit characterization

$$Z = \{(x,z) \in \mathcal{X}^2 : x - z \in \tau \odot \mathbb{Z}^d\}$$

and we have: $0 \le k(x,z) < 1, \forall (x,z) \notin Z$.

B.2 DATA SETTING

Let $\mathcal{X} \subset \mathbb{R}^d$ be bounded. Let μ be a probability measure on $\mathcal{X}$ admitting a density $g \in L^1(\mathbb{R}^d) \cap L^p(\mathbb{R}^d)$ for some $p > 1$ (possibly $p = +\infty$), with full support:

$$\forall x \in \mathcal{X},\ \forall t > 0, \quad \mu(B(x,t)) > 0.$$

Let $\{(x_i, y_i)\}_{i=1}^n$ be i.i.d. samples generated as

$$x_i \sim \mu, \qquad y_i = f^*(x_i) + \varepsilon_i, \tag{16}$$

where $\mathbb{E}[\varepsilon_i \mid x_i] = 0$, $\mathbb{E}[\varepsilon_i^2 \mid x_i] = \sigma^2$, and $f^* \in \mathcal{H}_k$ where $\mathcal{H}_k$ is the RKHS of the local kernel k.

Denote by

$$X_n = \{x_1, \ldots, x_n\}, \qquad Y_n = (y_1, \ldots, y_n)^\top.$$

Define the minimal pairwise spike set distance as

$$\Delta(X_n) = \min_{1\le i<j\le n} \inf_{(x',z')\in Z} \|(x_i, x_j) - (x', z')\|_2. \tag{17}$$

B.3 KERNEL REGRESSORS

Let

$$K = (k(x_i, x_j))_{i,j}, \qquad K^{\odot q} = (k(x_i, x_j)^q)_{i,j}.$$

For $x \in \mathcal{X}$, define

$$k(x, X_n) = (k(x, x_1), \ldots, k(x, x_n))^\top,$$
$$k^{\odot q}(x, X_n) = (k(x, x_1)^q, \ldots, k(x, x_n)^q)^\top.$$

The ρ-ridge regressor obtained with the kernel k is

$$\hat{f}_{\rho,n}(x) = k(x, X_n)^\top (K + \rho I_n)^{-1} Y_n.$$

For $q \in \mathbb{N}^*$ and $\rho > 0$, using the separable global encoding in Fig. 3, obtain a Local-Global kernel of the form

$$k_{\rho,q}(x,z) = k(x,z) + \rho\, k(x,z)^q,$$

with the Gram matrix $K_{\rho,q} = K + \rho K^{\odot q}$.

The ridgeless regressor obtained with the kernel $k_{\rho,q}$ is

$$\hat{f}_{\rho,q,n}(x) = k_{\rho,q}(x, X_n)^\top K_{\rho,q}^+ Y_n,$$

where $(\cdot)^+$ denotes the Moore-Penrose pseudoinverse.

C AUXILIARY RESULTS ON THE SPIKE SET

Lemma C.1. *Assume* (H1) *and $\mathcal{X} \subset \mathbb{R}^d$ bounded. Then there exists a finite set $\{a_i\}_{i=1}^{N_\mathcal{X}} \subset \mathbb{Z}^d$, with $a_1 = 0$, such that*

$$Z = \bigcup_{i=1}^{N_\mathcal{X}} \big\{(x,z) \in \mathcal{X}^2 :\ x - z = \tau \odot a_i\big\}.$$

Proof. By (H1),

$$Z = \{(x,z) \in \mathcal{X}^2 :\ x - z = \tau \odot \mathbb{Z}^d\} = \bigcup_{m \in \mathbb{Z}^d} \{(x,z) \in \mathcal{X}^2 :\ x - z = \tau \odot m\}. \tag{18}$$

Consider $m \in \mathbb{Z}^d$ such that $\{(x,z) \in \mathcal{X}^2 :\ x - z = \tau \odot m\} \neq \emptyset$. There exists $(x,z) \in \mathcal{X}^2$ such that $x - z = \tau \odot m$. Thus

$$\|\tau \odot m\|_2 \leq 2 \sup_{x \in \mathcal{X}} \|x\|_2$$

Also $\|\tau \odot m\|_2 \geq \tau_{\min} \|m\|_2$ with $\tau_{\min} = \min_j \tau_j > 0$ by (H1).

Since $\mathcal{X}$ is bounded, we conclude that $\|m\|_2 \leq \frac{2 \sup_{x \in \mathcal{X}} \|x\|_2}{\tau_{\min}} < +\infty$. So there is only a finite number of m in $\mathbb{Z}^d$ yielding non-empty sets in (18).

The claim follows, noting that $m = 0_d$ is always included since $(x,x) \in \mathcal{X}^2$ for all $x \in \mathcal{X}$. □

Lemma C.2. *Assume* (H1) *and $\mathcal{X} \subset \mathbb{R}^d$ bounded. Then there exists a finite collection $\{a_i\}_{i=1}^{N_\mathcal{X}} \subset \mathbb{Z}^d$, with $a_1 = 0$, such that for every $\delta > 0$,*

$$S(\delta) \subset \bigcup_{i=1}^{N_\mathcal{X}} \Big\{(x,z) \in \mathcal{X}^2 : \|x - z - \tau \odot a_i\|_2 \leq \sqrt{2}\,\delta\Big\}.$$

Proof. By Lemma C.1, the spike set admits a finite decomposition

$$Z = \bigcup_{i=1}^{N_\mathcal{X}} Z_{a_i}, \qquad Z_{a_i} := \{(x,z) \in \mathcal{X}^2 :\ x - z = \tau \odot a_i\}.$$

Hence,

$$S(\delta) = \bigcup_{i=1}^{N_\mathcal{X}} \Big\{(x,z) \in \mathcal{X}^2 : \inf_{(x',z') \in Z_{a_i}} \|(x,z) - (x',z')\|_2 \leq \delta\Big\}.$$

For each i, define the set

$$W_{a_i} := \{(x',z') \in \mathbb{R}^{2d} :\ x' - z' = \tau \odot a_i\},$$

so that $Z_{a_i} \subset W_{a_i}$ and

$$\inf_{(x',z') \in Z_{a_i}} \|(x,z) - (x',z')\|_2 \geq \inf_{(x',z') \in W_{a_i}} \|(x,z) - (x',z')\|_2.$$

Note that

$$\inf_{(x',z')\in W_{a_i}} \|(x,z)-(x',z')\|_2^2 = \inf_{w\in\mathbb{R}^d} \|(x,z)-(w,w-\tau\odot a_i)\|_2^2.$$

The latter is a strictly convex quadratic form in w, a direct computation yields

$$\inf_{(x',z')\in W_{a_i}} \|(x,z)-(x',z')\|_2^2 = \tfrac{1}{2}\,\|x-z-\tau\odot a_i\|_2^2,$$

where the minimizer is $(x',z')=(w^*,w^*-\tau\odot a_i)$ and $w^*=\frac{1}{2}(x+z+\tau\odot a_i)$. Therefore,

$$\inf_{(x',z')\in Z_{a_i}} \|(x,z)-(x',z')\|_2 \le \delta \Rightarrow\ \|x-z-\tau\odot a_i\|_2 \le \sqrt{2}\,\delta,$$

which concludes the proof.

□

Lemma C.3. *Assume $\mathcal{X}\subset\mathbb{R}^d$ is bounded. Let $x_1,x_2\sim\mu$ i.i.d., where μ admits a density g fully supported on $\mathcal{X}$ with $g\in L^1(\mathbb{R}^d)\cap L^p(\mathbb{R}^d)$, for $p>1$. Let p' satisfy $\frac{1}{p}+\frac{1}{p'}=1$. Then*

$$\forall z\in\mathcal{X},\qquad \mu\big(\{x:(x,z)\in S(\delta)\}\big)=O\big(\delta^{\frac{d}{p'}}\big).$$

Proof. Fix $z\in\mathcal{X}$. By Lemma C.2,

$$\{x:(x,z)\in S(\delta)\}\subset\bigcup_{i=1}^{N_{\mathcal{X}}}\big\{x\in\mathcal{X}:\ \|x-z-\tau\odot a_i\|_2\le\sqrt{2}\,\delta\big\}=\bigcup_{i=1}^{N_{\mathcal{X}}}\Big(B(z+\tau\odot a_i,\sqrt{2}\delta)\cap\mathcal{X}\Big).$$

Therefore,

$$\mu\big(\{x:(x,z)\in S(\delta)\}\big)=\int_{\mathcal{X}}\mathbf{1}_{\{(x,z)\in S(\delta)\}}(x)\,g(x)\,dx\le\sum_{i=1}^{N_{\mathcal{X}}}\int_{\mathcal{X}}\mathbf{1}_{B(z+\tau\odot a_i,\sqrt{2}\delta)}(x)\,g(x)\,dx.$$

Applying Hölder's inequality with exponents (p,p') gives

$$\int_{\mathcal{X}}\mathbf{1}_{B(z+\tau\odot a_i,\sqrt{2}\delta)}(x)\,g(x)\,dx\le\|g\|_{L^p}\,\|\mathbf{1}_{B(z+\tau\odot a_i,\sqrt{2}\delta)}\|_{L^{p'}}=\|g\|_{L^p}\,\big(v_d(\sqrt{2}\delta)^d\big)^{1/p'},$$

where v_d denotes the volume of the unit ball in $\mathbb{R}^d$.

Combining the above bounds concludes the proof:

$$\mu\big(\{x:(x,z)\in S(\delta)\}\big)\le N_{\mathcal{X}}\|g\|_{L^p}(v_d\sqrt{2})^{\frac{d}{p'}}\,\delta^{\frac{d}{p'}}.$$

□

Lemma C.4. *Under the assumptions of Lemma C.3,we have*

$$p(\delta):=(\mu\times\mu)\big(S(\delta)\big)=O\big(\delta^{\frac{d}{p'}}\big).$$

Proof.

$$\begin{aligned}
p(\delta)&:=(\mu\times\mu)(\{(x,z)\in S(\delta)\})\\
&=\int_{\mathcal{X}^2}\mathbf{1}_{S(\delta)}(x,z)g(x)g(z)dxdz\\
&=\int_{\mathcal{X}}\Big(\mu\big(\{x:(x,z)\in S(\delta)\}\big)\Big)g(z)dz\\
&\le\int_{\mathcal{X}}N_{\mathcal{X}}\|g\|_{L^p}(v_d\sqrt{2})^{\frac{d}{p'}}\,\delta^{\frac{d}{p'}}g(z)dz\qquad\text{(using Lemma C.3)}\\
&\le N_{\mathcal{X}}\|g\|_{L^p}(v_d\sqrt{2})^{\frac{d}{p'}}\,\delta^{\frac{d}{p'}},
\end{aligned}$$

which concludes the proof. □

Lemma C.5. *Assume the same conditions as in Lemma C.3. Fix $\alpha > 0$. If the sequence $\{\delta_n\}_{n\in\mathbb{N}}$ verifies $\delta_n = O\left(n^{-(2+\alpha)\frac{p'}{d}}\right)$ and if for all n, $X_n = \{x_i\}_{i=1}^n$ are i.i.d. samples from μ, then*

$$\mathbb{P}\big(\Delta(X_n) > \delta_n\big) \xrightarrow[n\to\infty]{} 1.$$

where

$$\Delta(X_n) = \min_{1\le i<j\le n} \inf_{(x',z')\in Z} \|(x_i,x_j)-(x',z')\|_2.$$

Proof. Noticing that

$$\begin{aligned}\Delta(X_n) \le \delta_n &\Leftrightarrow \exists\,(i,j)\in\{1,..,n\}^2, i<j, \text{ s.t. } \inf_{(x',z')\in Z}\|(x_i,x_j)-(x',z')\|_2 \le \delta_n\\ &\Leftrightarrow \exists\,(i,j)\in\{1,..,n\}^2, i<j, \text{ s.t. } (x_i,x_j)\in S(\delta_n).\end{aligned}$$

We have : $\forall (x_1,..,x_n),\ \mathbf{1}_{\{\Delta(X_n)\le\delta_n\}}(x_1,..,x_n) \le \sum_{i=1}^n\sum_{j=i+1}^n \mathbf{1}_{S(\delta_n)}(x_i,x_j)$, therefore

$$\begin{aligned}\mathbb{P}(\Delta(X_n)\le\delta_n) &\le \int_{\mathcal{X}^n}\left(\sum_{i=1}^n\sum_{j=i+1}^n \mathbf{1}_{S(\delta_n)}(x_i,x_j)\right) g(x_1)..g(x_n)dx_1..dx_n\\ &\le \sum_{i=1}^n\sum_{j=i+1}^n \int_{\mathcal{X}^n}\mathbf{1}_{S(\delta_n)}(x_i,x_j)g(x_1)..g(x_n)dx_1..dx_n\\ &\le \sum_{i=1}^n\sum_{j=i+1}^n p(\delta_n)\\ &\le \tfrac{n(n-1)}{2}p(\delta_n)\\ &\le C\tfrac{n(n-1)}{2}(\delta_n)^{\frac{d}{p'}},\end{aligned}$$

where the last line follows from Lemma C.4. By assumption $\delta_n \le C'\left(n^{-(2+\alpha)\frac{p'}{d}}\right)$ so that $C\frac{n(n-1)}{2}(\delta_n)^{\frac{d}{p'}} \le \frac{CC'^{\frac{d}{p'}}}{2}n^{-\alpha}$. We conclude that $\lim_{n\to+\infty}\mathbb{P}(\Delta(X_n)\le\delta_n)=0$ and $\lim_{n\to+\infty}\mathbb{P}(\Delta(X_n)>\delta_n)=1$. □

D PROOF OF THE MAIN THEOREM

Theorem D.1 (Benign overfitting for periodic quantum kernels)**.** *Let k be a quantum fidelity kernel satisfying* (H1)*, let the data satisfy the hypotheses in Appendix B.2. Write $\beta = \frac{d}{p'}$ with $\frac{1}{p'}+\frac{1}{p}=1$. Choose $\alpha>0$ and $0<\alpha_0<1$.*

Then choosing sequences $\{\delta_n\}_{n\ge1}$, $\{q_n\}_{n\ge1}$ and $\{\rho_n\}_{n\ge1}$ satisfying:

$$\delta_n = o\left(n^{-\frac{2+\alpha}{\beta}}\right), \qquad \rho_n n^{\alpha_0} \xrightarrow[n\to\infty]{} 0, \quad \rho_n n \xrightarrow[n\to\infty]{} +\infty,$$

and

$$q_n \ge \frac{c_\alpha \log n}{-\log m(\delta_n)}, \qquad \textit{with} \qquad c_\alpha > \tfrac{6+\alpha}{2},$$

where

$$m(\delta) \overset{\textit{def}}{=} \sup_{(x,z)\in S(\delta)^c} k(x,z), \quad \forall\delta>0.$$

$$\textit{Write} \qquad \Delta(X_n) = \min_{1\le i<j\le n}\inf_{(x',z')\in Z}\|(x_i,x_j)-(x',z')\|_2 \qquad \textit{and} \qquad P_n = \mathbb{P}(\Delta(X_n)<\delta_n).$$

Then $P_n \xrightarrow[n\to\infty]{} 0$ and with probability at least $(1-O(n^{-\alpha_0}))(1-O(n^{-\alpha}))(1-P_n)$, the ridgeless estimator $\hat{f}_{\rho_n,q_n,n}$, built on the kernel k_{ρ_n,q_n}, satisfies

$$\mathbb{E}_x\left[\left|\hat{f}_{\rho_n,q_n,n}(x)-f^*(x)\right|^2\right] \le B_{0,n}+B_{1,n}+B_{2,n}\ . \tag{19}$$

with

$$B_{0,n} = 2\sigma^2 4^m n^{\alpha_0 - 1}\Big(1 + \frac{1}{\rho_n n}\Big) \quad + \quad 2\|f^*\|^2_{\mathcal{H}_k}\rho_n n^{\alpha_0}\Big(1 + \frac{1}{\rho_n n}\Big)^2, \tag{20}$$

$$B_{1,n} = 4C_y\,\frac{n^{4+\alpha} m(\delta_n)^{2q_n}}{\rho_n^2\big(1 - nm(\delta_n)^{q_n}\big)^2}, \qquad \textit{where} \quad C_y = \mathbb{E}_x[f^*(x)^2] + \sigma^2, \tag{21}$$

$$B_{2,n} = 4C_y\,\frac{n^{2+\alpha}\left(C_\mu \delta_n^\beta + m(\delta_n)^{2q_n}\right)}{\big(1 - nm(\delta_n)^{q_n}\big)^2}, \qquad \textit{where} \quad C_\mu = N_{\mathcal{X}}\|g\|_{L^p}(v_d\sqrt{2})^{\frac{d}{p'}}. \tag{22}$$

Furthermore $B_{i,n} \xrightarrow[n\to\infty]{} 0$ *for* $i = 0, 1, 2$.

Hence, the sequence of estimators $\{\hat{f}_{\rho_n, q_n, n}\}_n$ *is consistent in probability.*

Proof of Theorem D.1. We follow the same error decomposition as in Haas et al. [2023].

$$\mathbb{E}_x\Big[\big|\hat{f}_{\rho,q,n}(x) - f^*(x)\big|^2\Big] \le 2\mathbb{E}_x\Big[\big|\hat{f}_{\rho,n}(x) - f^*(x)\big|^2\Big] + 2\mathbb{E}_x\Big[\big|\hat{f}_{\rho,q,n}(x) - \hat{f}_{\rho,n}(x)\big|^2\Big]. \tag{23}$$

and

$$\mathbb{E}_x\Big[\big|\hat{f}_{\rho,q,n}(x) - \hat{f}_{\rho,n}(x)\big|^2\Big] \le 2\mathbb{E}_x\Big[\big|k(x, X_n)^\top\big[(K + \rho I_n)^{-1} - K^+_{\rho,q}\big]Y_n\big|^2\Big] + 2\rho^2\mathbb{E}_x\Big[\big|k^{\odot q}(x, X_n)^\top K^+_{\rho,q} Y_n\big|^2\Big].$$

Applying Cauchy-Schwarz' inequality and operator norm inequality gives

$$\mathbb{E}_x\Big[\big|\hat{f}_{\rho,q,n} - \hat{f}_{\rho,n}\big|^2\Big] \le 2\,\mathbb{E}_x\big[\|k(x, X_n)\|_2^2\big]\,\big\|K^+_{\rho,q} - (K + \rho I_n)^{-1}\big\|^2_{\mathrm{op}}\,\|Y_n\|_2^2 + 2\rho^2\,\mathbb{E}_x\big[\|k^{\odot q}(x, X_n)\|_2^2\big]\,\|K^+_{\rho,q}\|^2_{\mathrm{op}}\,\|Y_n\|_2^2 \tag{24}$$

Combining (23) and (24), we obtain

$$\mathbb{E}_x\Big[\big|\hat{f}_{\rho,q,n}(x) - f^*(x)\big|^2\Big] \le A_{0,n} + A_{1,n} + A_{2,n}, \tag{25}$$

where

$$A_{0,n} \overset{\text{def}}{=} 2\mathbb{E}_x\Big[\big|\hat{f}_{\rho,n}(x) - f^*(x)\big|^2\Big], \tag{26}$$

$$A_{1,n} \overset{\text{def}}{=} 4\,\mathbb{E}_x\big[\|k(x, X_n)\|_2^2\big]\,\big\|K^+_{\rho,q} - (K + \rho I_n)^{-1}\big\|^2_{\mathrm{op}}\,\|Y_n\|_2^2, \tag{27}$$

$$A_{2,n} \overset{\text{def}}{=} 4\rho^2\,\mathbb{E}_x\big[\|k^{\odot q}(x, X_n)\|_2^2\big]\,\|K^+_{\rho,q}\|^2_{\mathrm{op}}\,\|Y_n\|_2^2. \tag{28}$$

To show the consistency in probability a sufficient condition is to obtain $A_{i,n} \xrightarrow[n\to\infty]{} 0$ in probability for $i = 0, 1, 2$.

Bounding $A_{0,n}$

By construction, the quantum feature map of k in Eq (13) takes values in the Hilbert space of linear operators on $(\mathbb{C}^2)^{\otimes m}$ endowed with the Hilbert-Schmidt inner product. Since $\rho(x)$ is a rank-one density matrix, we have

$$\forall x \in \mathcal{X}, \quad \|\rho(x)\|^2_{\mathrm{HS}} = \mathrm{Tr}\big[\rho(x)\rho^\dagger(x)\big] = \mathrm{Tr}\big[\rho^2(x)\big] = 1$$

so the feature map is uniformly bounded by 1. Moreover, this Hilbert space has finite dimension $p \le 4^m$.

So, under the quantum feature map in Eq (13) and data hypothesis in Appendix B.2, we can apply the Proposition 7.6 of Bach [2024] with regularization $\lambda = \rho_n, R = 1$ in the finite dimensional case

$$\mathbb{E}_{(X_n, Y_n)}\Big[\mathbb{E}_x\Big[|\hat{f}_{\rho_n,n}(x) - f^*(x)|^2\Big]\Big] \le \frac{\sigma^2 p}{n}\Big(1 + \frac{1}{\rho_n n}\Big) + \rho_n\Big(1 + \frac{1}{\rho_n n}\Big)^2\|f^*\|^2_{\mathcal{H}_k}. \tag{29}$$

When choosing a regularization sequence satisfying

$$\rho_n \to 0 \qquad \text{and} \qquad \rho_n n \to \infty,$$

both terms on the right-hand side vanish:

$$\mathbb{E}_{(X_n,Y_n)}\left[\mathbb{E}_x\left[|\hat{f}_{\rho_n,n}(x)-f^*(x)|^2\right]\right] \xrightarrow[n\to\infty]{} 0.$$

Let

$$C_n(\epsilon)=\left\{(X_n,Y_n)\in(\mathcal{X}\times\mathbb{R})^n : \mathbb{E}_x\left[|\hat{f}_{\rho_n,n}(x)-f^*(x)|^2\right]\geq\epsilon\right\}$$

By Markov's inequality, for any fixed $\alpha_0>0$,

$$\mathbb{P}\Big(C_n\big(\mathbb{E}_x\left[|\hat{f}_{\rho_n,n}(x)-f^*(x)|^2\right]\, n^{\alpha_0}\big)\Big)\leq n^{-\alpha_0}.$$

Hence, with probability at least $(1-n^{-\alpha_0})$,

$$A_{0,n}\leq\frac{2\sigma^2 pn^{\alpha_0}}{n}\Big(1+\frac{1}{\rho_n n}\Big)+2n^{\alpha_0}\rho_n\Big(1+\frac{1}{\rho_n n}\Big)^2\|f^*\|^2_{\mathcal{H}_k}=B_{0,n}. \tag{30}$$

Then choosing $0<\alpha_0<1$ and $\{\rho_n\}_n$ such that

$$n^{\alpha_0}\rho_n \xrightarrow[n\to\infty]{} 0, \quad n\rho_n \xrightarrow[n\to\infty]{} +\infty$$

guarantees that $B_{0,n}\xrightarrow[n\to\infty]{}0$ and $A_{0,n}\xrightarrow[n\to\infty]{}0$ with probability at least $(1-n^{-\alpha_0})$.

Bounding $A_{2,n}$

We have $A_{2,n}\overset{\text{def}}{=}4\rho^2\,\mathbb{E}_x\big[\|k^{\odot q}(x,X_n)\|_2^2\big]\,\|K^+_{\rho,q}\|^2_{\text{op}}\,\|Y_n\|_2^2$. We bound each term.

Bounding $\mathbb{E}_x\big[\|k^{\odot q_n}(x,X_n)\|_2^2\big]$

For each i:

$$\begin{aligned}k^{2q_n}(x,x_i)&=k^{2q_n}(x,x_i)\left(\mathbf{1}_{\{(x,x_i)\in S(\delta_n)\}}+\mathbf{1}_{\{(x,x_i)\in S^c(\delta_n)\}}\right)\\&\leq\mathbf{1}_{\{(x,x_i)\in S(\delta_n)\}}+m(\delta_n)^{2q_n}\mathbf{1}_{\{(x,x_i)\in S^c(\delta_n)\}}\\&\leq\mathbf{1}_{\{(x,x_i)\in S(\delta_n)\}}+m(\delta_n)^{2q_n}\end{aligned}$$

Then summing over i and taking the expectation gives

$$\mathbb{E}_x\big[\|k^{\odot q_n}(x,X_n)\|_2^2\big]=\sum_{i=1}^n\mathbb{E}_x[k^{2q_n}(x,x_i)]\leq\sum_{i=1}^n\mu\big(\{x:(x,x_i)\in S(\delta_n)\}\big)+nm(\delta_n)^{2q_n}.$$

Then using Lemma C.3, we obtain

$$\mathbb{E}_x\big[\|k^{\odot q_n}(x,X_n)\|_2^2\big]\leq n(C_\mu\delta_n^\beta+m(\delta_n)^{2q_n})\quad\text{where}\quad C_\mu=N_{\mathcal{X}}\|g\|_{L^p}(v_d\sqrt{2})^{\frac{d}{p'}}.$$

Bounding $\|K^+_{\rho_n,q_n}\|_{op}$

Since $\delta_n=O\left(n^{-\frac{2+\alpha}{\beta}}\right)$, we can use Lemma C.5: with probability $(1-P_n)$, $\Delta(X_n)\geq\delta_n$ i.e. $(x_i,x_j)\notin S(\delta_n)$ for all $i\neq j$.

Write $K_{\rho_n,q_n}=K+\rho_n K^{\odot q_n}=K+\rho_n I_n+\rho_n E(q_n)$. By Weyl's inequality,

$$\lambda_{\min}(K_{\rho_n,q_n})\geq\lambda_{\min}(K+\rho_n I_n)+\rho_n\,\lambda_{\min}(E(q_n)).$$

For $i\neq j$, $|E(q_n)_{ij}|=k^{q_n}(x_i,x_j)\leq m(\delta_n)^{q_n}$ and $E(q_n)_{ii}=0$. The Gershgorin circle theorem allows to bound all the eigenvalues of $E(q_n)$

$$|\lambda_\ell(E(q_n))|\leq(n-1)m(\delta_n)^{q_n}.$$

Hence $\|E(q_n)\|_{\mathrm{op}} \leq nm(\delta_n)^{q_n}$ and $\lambda_{\min}(E(q_n)) \geq -\|E(q_n)\|_{\mathrm{op}} \geq -nm(\delta_n)^{q_n}$ and

$$\lambda_{\min}(K_{\rho_n,q_n}) \geq \rho_n - \rho_n nm(\delta_n)^{q_n} = \rho_n\big(1 - nm(\delta_n)^{q_n}\big).$$

Notice that $0 \leq m(\delta_n) < 1$. So since $q_n \geq -\frac{\log n}{\log m(\delta_n)}$ then $0 \leq nm(\delta_n)^{q_n} < 1$. This implies that $\lambda_{\min}(K_{\rho_n,q_n}) > 0$ and $K^+_{\rho_n,q_n} = K^{-1}_{\rho_n,q_n}$. Consequently

$$\|K^+_{\rho_n,q_n}\|^2_{\mathrm{op}} \leq \frac{1}{\rho_n^2\big(1 - nm(\delta_n)^{q_n}\big)^2}, \quad \text{with probability } (1 - P_n).$$

Bounding $\|Y_n\|_2^2$

As in the proof of Theorem G.8 of Haas et al. [2023]:

$$\mathbb{E}\big[\|Y_n\|_2^2\big] = \sum_{i=1}^{n} \mathbb{E}[y_i^2] = n\big(\mathbb{E}_x[f^*(x)^2] + \sigma^2\big) = nC_y.$$

By Markov's inequality, for any fixed $\alpha > 0$,

$$\mathbb{P}\Big(\|Y_n\|_2^2 \geq \mathbb{E}[\|Y_n\|_2^2]\, n^{\alpha}\Big) \leq \frac{\mathbb{E}[\|Y_n\|_2^2]}{\mathbb{E}[\|Y_n\|_2^2]\, n^{\alpha}} = n^{-\alpha}.$$

Hence, with probability at least $(1 - n^{-\alpha})$, $\|Y_n\|_2^2 \leq C_y\, n^{1+\alpha}$.

Final bound for $A_{2,n}$

Collecting the three previous bounds yields with probability at least $(1 - n^{-\alpha})(1 - P_n)$

$$A_{2,n} \leq 4\rho_n^2 n\big(C_\mu \delta_n^\beta + m(\delta_n)^{2q_n}\big)\, \frac{C_y n^{1+\alpha}}{\rho_n^2\big(1 - nm(\delta_n)^{q_n}\big)^2} = 4C_y\, \frac{n^{2+\alpha}\big(C_\mu \delta_n^\beta + m(\delta_n)^{2q_n}\big)}{\big(1 - nm(\delta_n)^{q_n}\big)^2} = B_{2,n}. \tag{31}$$

Bounding $A_{1,n}$

$$A_{1,n} \stackrel{\text{def}}{=} 4\,\mathbb{E}_x\big[\|k(x, X_n)\|_2^2\big]\, \big\|K^+_{\rho,q} - (K + \rho I_n)^{-1}\big\|^2_{\mathrm{op}}\, \|Y_n\|_2^2.$$

As before, we bound each term composing $A_{1,n}$.

Since $k(x, x_i)^2 \leq 1$, $\mathbb{E}_x\big[\|k(x, X_n)\|^2\big] = \mathbb{E}_x\big[\sum_{i=1}^n k(x, x_i)^2\big] \leq n$.

Using Lemma G.8 of Haas et al. [2023] with $A = K + \rho_n I_n$, $B = K_{\rho_n,q_n}$, we obtain

$$\big\|K^+_{\rho_n,q_n} - (K + \rho_n I_n)^{-1}\big\|_{\mathrm{op}} \leq \big\|(K + \rho_n I_n)^{-1}\big\|_{\mathrm{op}}\, \frac{\|E(q_n)\|_{\mathrm{op}}}{1 - \|E(q_n)\|_{\mathrm{op}}} \leq \frac{1}{\rho_n}\, \frac{nm(\delta_n)^{q_n}}{(1 - nm(\delta_n)^{q_n})}.$$

Combining the previous bounds shows that with probability at least $(1 - n^{-\alpha})(1 - P_n)$

$$\begin{aligned} A_{1,n} &\leq 4n\, \frac{n^2 m(\delta_n)^{2q_n}}{\rho_n^2\big(1 - nm(\delta_n)^{q_n}\big)^2}\, C_y n^{1+\alpha} \\ &= 4C_y\, \frac{n^{4+\alpha} m(\delta_n)^{2q_n}}{n^2\rho_n^2\big(1 - nm(\delta_n)^{q_n}\big)^2} = B_{1,n}. \end{aligned} \tag{32}$$

The bounds hold also with probability at least $(1 - n^{-\alpha})(1 - P_n)$.

Limits of $B_{1,n}$ and $B_{2,n}$

To conclude the proof, it remains to show that the hypotheses on $\{\delta_n\}_n$, $\{\rho_n\}_n$ and $\{q_n\}_n$ ensure that $B_{1,n} \xrightarrow[n\to\infty]{} 0$ and $B_{2,n} \xrightarrow[n\to\infty]{} 0$.

Since $q_n \geq \frac{c_\alpha \log n}{-\log m(\delta_n)}$, we have $m(\delta_n)^{q_n} \leq n^{-C_\alpha}$ with $C_\alpha > \frac{6+\alpha}{2}$, thus

$$n^{2+\alpha} m(\delta_n)^{2q_n} \leq n^{2+\alpha-2C_\alpha} \leq n^{-4} \xrightarrow[n\to\infty]{} 0,$$

$$n^{4+\alpha} m(\delta_n)^{2q_n} \leq n^{4+\alpha-2C_\alpha} \leq n^{-2} \xrightarrow[n\to\infty]{} 0,$$

and

$$n m(\delta_n)^{q_n} \leq n^{1-C_\alpha} \xrightarrow[n\to\infty]{} 0.$$

By assumption $n\rho_n \xrightarrow[n\to\infty]{} +\infty$ and $n^{2+\alpha}\delta_n^\beta \xrightarrow[n\to\infty]{} 0$. This gives all the ingredients to conclude that $B_{1,n} \xrightarrow[n\to\infty]{} 0$ and $B_{2,n} \xrightarrow[n\to\infty]{} 0$.

□

E INSTANTIATION FOR THE COSINE QUANTUM KERNEL

The purpose of this appendix is to connect the general periodic result of Appendix D with Theorem 3.2, which is stated specifically for the cosine-squared kernel $k(x,z) = \prod_{j=1}^d \cos^2\Big(\frac{x_j - z_j}{2}\Big)$.

Appendix D establishes benign overfitting for periodic translation-invariant kernels when (q_n) grows at a rate which depends on $m(\delta)$. In contrast, Theorem 3.2 is stated with an explicit lower bound on (q_n), which no longer involves $m(\delta)$. To bridge both results, we derive an explicit upper bound on $m(\delta)$ for the cosine kernel. Substituting this bound into the general condition of Appendix D.1 yields the explicit requirement on (q_n) appearing in Theorem 3.2.

We now establish such a bound.

Lemma E.1. *Let $\delta > 0$ and*

$$k(x,z) = \prod_{j=1}^d \cos^2\Big(\frac{x_j - z_j}{2}\Big).$$

Then

$$m(\delta) \leq \exp\Big(-\frac{2\delta^2}{\pi^2}\Big).$$

Proof. Since the kernel is 2π-periodic in each coordinate, Lemma C.2 implies that if $(x,z) \notin S(\delta)$, then for some $a \in \mathbb{Z}^d$,

$$h := x - z - 2\pi a \quad \text{satisfies} \quad \|h\|_2 > \sqrt{2}\,\delta \quad \text{and} \quad h \in [-\pi,\pi]^d.$$

By convexity of $u \mapsto e^{-u}$ on $\mathbb{R}_+$, the function lies above its tangent at $u = 0$. Hence,

$$\forall u \geq 0, \qquad 1 - u \leq e^{-u}.$$

Moreover, since $v \mapsto \sin v$ is concave on $[0,\pi]$, it lies above the chord joining $(0,0)$ and $(\pi/2, 1)$. Therefore,

$$\forall v \in \Big[0, \frac{\pi}{2}\Big], \qquad \frac{2v}{\pi} \leq \sin v.$$

Let $t \in [-\pi,\pi]$. Setting $v = |t|/2 \in [0,\pi/2]$, we obtain

$$\sin\Big(\frac{|t|}{2}\Big) \geq \frac{|t|}{\pi}.$$

Squaring both sides yields

$$\sin^2\left(\frac{t}{2}\right) \geq \frac{t^2}{\pi^2}.$$

Using $\cos^2(t/2) = 1 - \sin^2(t/2)$, we deduce

$$\cos^2(t/2) \leq 1 - \frac{t^2}{\pi^2}.$$

Applying the inequality $1 - u \leq e^{-u}$ with $u = t^2/\pi^2$, we conclude

$$\cos^2(t/2) \leq \exp\left(-\frac{t^2}{\pi^2}\right).$$

Hence

$$k(x,z) = \prod_{j=1}^{d} \cos^2(h_j/2) \leq \exp\Big(-\frac{\|h\|_2^2}{\pi^2}\Big) \leq \exp\Big(-\frac{2\delta^2}{\pi^2}\Big),$$

which proves the claim. □

Appendix D proves benign overfitting for periodic translation-invariant kernels under the condition

$$q_n \;\geq\; \frac{c_\alpha \log n}{-\log m(\delta_n)}, \qquad c_\alpha > \tfrac{6+\alpha}{2}.$$

For the cosine kernel, we have shown E.1 that

$$m(\delta) \leq \widetilde{m}(\delta) := \exp(-\frac{2\delta^2}{\pi^2}).$$

Because $m(\delta_n) \leq \widetilde{m}(\delta_n) < 1$, we obtain

$$-\log m(\delta_n) \;\geq\; -\log \widetilde{m}(\delta_n) = \frac{2\delta_n^2}{\pi^2}.$$

Therefore

$$\frac{c_\alpha \log n}{-\log m(\delta_n)} \;\leq\; \frac{c_\alpha \pi^2 \log n}{2\delta_n^2},$$

and it is sufficient to choose

$$q_n \;\geq\; \frac{c_\alpha \pi^2 \log n}{2\delta_n^2},$$

Since $n^{-\xi} = o\Big(n^{-\frac{2(2+\alpha)p}{d(p-1)}}\Big)$ for $\xi > \frac{2(2+\alpha)p}{d(p-1)}$, we can set $\delta_n = \frac{\pi}{2} n^{\frac{-\xi}{2}} = o\Big(n^{-\frac{(2+\alpha)p}{d(p-1)}}\Big)$, giving $q_n \geq 2c_\alpha n^\xi \ln n$, in accordance with Theorem D.1. This yields the condition on (q_n) stated in Theorem 3.2.

F ADDITIONAL EXPERIMENTAL DETAILS

F.1 SYNTHETIC EXPERIMENTS

The synthetic experiments (a-c) illustrate the behavior of the Local-Global construction in controlled settings and complement the theoretical analysis.

Table 2: Summary of experimental settings. Dimension d; training set size n; test set size n_{test}; noise variance σ^2; kernel parameters ρ, q, c; target function.

Experiment	d	n	n_{test}	σ^2	ρ	q	c	Target
a	1	50	–	0.5	$\sigma^2/\sqrt{n}$	$\{2^p : p \in \{4, 8, 14, 16\}\}$	$\frac{3\pi}{4}$	f_1 (33)
b1	7	500	2000	0.5	$\sigma^2/\sqrt{n}$	$\{2^p : p \in \{2, 4, 6, 8\}\}$	$\frac{1}{\sqrt{d}}$	f_2 (34) (35)
b2	7	500	2000	$7 \cdot 10^{-3}$	$\sigma^2/\sqrt{n}$	$\{2^p : p \in \{2, 4, 6, 8\}\}$	$\frac{1}{\sqrt{d}}$	f_3 (34) (36)
c	7	25-500	2000	0.5	$\sigma^2/\sqrt{n}$	$\lceil 10 \log(n) n^{6/d} \rceil$	$\frac{1}{\sqrt{d}}$	f_2 (34) (35)

General setting. Inputs are sampled i.i.d. from $\mathcal{U}([-1, 1]^d)$ and outputs are generated as

$$y_i = f^\star(x_i) + \sigma \varepsilon_i, \qquad \varepsilon_i \sim \mathcal{N}(0, 1).$$

The base kernel is

$$k_B(x, z) = \prod_{j=1}^{d} \cos^2\left(\frac{c(x_j - z_j)}{2}\right),$$

and the Local-Global kernel is

$$k_{LG}(x, z) = k_B(x, z) + \rho\, k_B(x, z)^q,$$

with $\rho = \sigma^2/\sqrt{n}$ as in the table 2. For experiments (b) and (c), results are averaged over 10 independent splits and evaluated using MSE on an independent test set.

Target functions. Experiment (a) uses the one-dimensional target

$$f_1(x) = \frac{2}{n_z} \sum_{j=1}^{n_z} k_B(x, z_j) + 1, \qquad z_j \overset{\text{i.i.d.}}{\sim} \mathcal{U}([-1, 1]), \quad n_z = 5. \tag{33}$$

For experiments (b1), (b2), and (c), we generate targets as:

$$\tilde{f}(x; k^\star) = \sum_{j=1}^{n_z} \alpha_j\, k^\star(x, z_j) - \mathbb{E}_{x'}\left[\sum_{j=1}^{n_z} \alpha_j\, k^\star(x', z_j)\right], \qquad n_z = 50, \qquad z_j \overset{\text{i.i.d.}}{\sim} \mathcal{U}([-1, 1]^d) \tag{34}$$

where α_j are i.i.d. Gaussian coefficients. We then rescale $\{\alpha_j\}$ so that, for $x \sim \mathcal{U}([-1, 1]^d)$ we have $\mathbb{E}[\tilde{f}(x; k^\star)] = 0$, $\mathrm{Var}(\tilde{f}(x; k^\star)) = 1$.

In (b1) and (c), we take the base kernel:

$$k^\star(x, z) = k_B(x, z), \tag{35}$$

and denote the resulting target by $f_2(x) = \tilde{f}(x; k_B) \in \mathcal{H}_{k_B}$.

In (b2), we take the Local-Global kernel

$$k^\star(x, z) = k_B(x, z) + \rho^\star k_B(x, z)^{q^\star}, \qquad (\rho^\star, q^\star) = (5, 8), \tag{36}$$

and denote the resulting target by $f_3(x) = \tilde{f}(x; k^\star) \notin \mathcal{H}_{k_B}$.

Experiment (a). This one-dimensional setting provides a minimal illustration of the behavior of the Local-Global construction as q varies, as depicted in Figure 5. We set the parameters as described in Table 2. We compare ridge regression with k_B ($\lambda = \rho$) and ridgeless regression with k_{LG} for different values of q. For numerical stability, a small regularization $\lambda_\varepsilon = 10^{-14}$ is used when $q = 2^4$; all other models are trained in the ridgeless regime. Predictions are evaluated on a dense grid over $[-1, 1]$ for visualization.

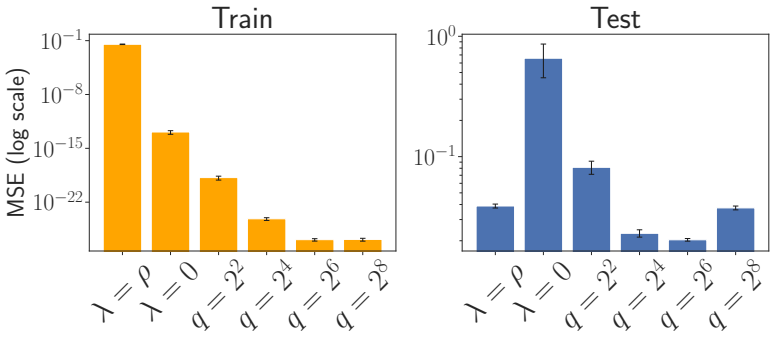


Figure 9: Synthetic regression in $d = 7$ (Experiment (b2 F.1), $n_{\text{train}} = 500$). Train and test MSE for k_B (ridgeless $\lambda = 0$ and ridge $\lambda = \rho$) and ridgeless k_{LG} as q increases. The target is generated from the Local-Global teacher in (36), so $f^\star \notin \mathcal{H}_{k_B}$.

Experiment (b). We consider the setting described in Table 2 for experiments (b1) and (b2). We compare ridge regression with k_B ($\lambda = \rho$), ridgeless regression with k_B ($\lambda = 0$), and ridgeless regression with k_{LG} for the values of q specified in the table. Results are averaged over 10 independent splits.

(b1) Well-specified case with respect to the base kernel. As reported in Figure 6, the target f_2 is generated using the base kernel k_B, hence $f_2 \in \mathcal{H}_{k_B}$. In this regime, ridge regression with k_B provides a strong reference solution. As q increases, the interpolating Local-Global estimator approaches the ridge performance while preserving exact interpolation.

(b2) Misspecified case with respect to the base kernel. The target f_3 is generated using the Local-Global kernel defined in (36), so that $f_3 \notin \mathcal{H}_{k_B}$. As q increases, the test error of the Local-Global estimator approaches that of ridge regression based on k_B, in accordance with the theoretical construction. In this misspecified regime, however, ridge regression with k_B is not optimal. Intermediate values of q yield strictly lower test error than both ridgeless regression with k_B and the large-q (ridge-like) limit, while remaining in the interpolating regime.

This observation motivates Experiment (d) on real datasets: if the best ridge estimator based on the base kernel is suboptimal due to misspecification, a Local-Global extension with a suitable value of q may achieve improved generalization, while being in the interpolating regime. Figure 7 illustrates such a scenario.

Experiment (c). This experiment examines the asymptotic behavior of the Local-Global estimator as the training size increases. We use the same target f_2 as in Experiment (b1) (see F.1), defined in (34)-(35), keeping the anchor points $\{z_j\}$ and coefficients $\{\alpha_j\}$ fixed across all values of n. Thus, the underlying function remains unchanged; only the training sample size varies.

For each n listed in Table 2, we generate a new training set of size n together with an independent test set, and compute the mean squared error averaged over 10 independent repetitions. The parameters are scaled with n as specified in Table 2.

Figure 8 shows that the Local-Global estimator using q_n exhibits decreasing test error that approaches the noise level, illustrating the benign overfitting behavior predicted by the theory. For comparison, ridgeless regression with the base kernel k_B does not display this convergence, as its test error grows with n.

F.2 REAL-WORLD DATA

We consider two complementary real-world evaluations. Experiment (d) is a controlled small-sample study designed to reproduce the behavior observed in Section F.1: after tuning the local ridge baseline to obtain the best local-kernel representation of the data, we sweep the exponent q of the Local-Global kernel and analyze the resulting evolution of the test error. Experiment (e) follows a standard validation-based protocol on larger training sets, with the goal of testing whether the ridgeless Local-Global estimator improves over the ridgeless local estimator when performance is measured on an independent test set.

Experiment (d). This experiment provides a real-data analogue of the synthetic setting in (b), where q interpolates between ridgeless regression and the ridge-like limit induced by the Local-Global construction. For each dataset, we first tune the bandwidth c and ridge parameter ρ for the local ridge estimator using repeated train/validation splits. Each tuning split

Table 3: Summary of the setting of the Experiment (d). Here d is the input dimension, n_{train} and n_{test} are the training and test sizes, and (c, ρ) are first tuned for the local ridge baseline.

Experiment	Dataset	d	n_{train}	n_{test}	ρ	q	c
d1	Airfoil	20	500	6900	tuned	$\{2^p : p \in \{1, \dots, 10\}\}$	tuned
d2	cal_housing	8	500	20140	tuned	$\{2^p : p \in \{1, \dots, 10\}\}$	tuned

Table 4: Dataset splits used in Experiment (e).

Dataset	n	d	n_{train}	n_{val}	n_{test}
cal_housing	20640	8	4128	6604	9908
energy	14980	14	2996	4793	7191
airfoil	7400	20	1480	2368	3552

uses n_{train} samples for training and $0.5 \cdot n_{\text{train}}$ additional samples for validation. The pair $(c^\star, \rho^\star)$ minimizing the average validation MSE over several such splits is selected.

We then fix $(c^\star, \rho^\star)$ and evaluate on independent train/test splits of size depicted in Table 3: (i) ridgeless regression with k_B, (ii) ridge regression with k_B using $\lambda = \rho^\star$, and (iii) ridgeless regression with k_{LG} for the values of q given in Table 3. Train and test MSE are averaged over 10 random splits.

The two datasets exhibit behaviors consistent with the two synthetic regimes of Experiment (b): for Airfoil (d1), intermediate values of q can outperform the tuned local ridge baseline (as in the misspecified pattern of (b2)); for California Housing (d2), the best performance is attained near the ridge-like limit (as in (b1)). These results motivate the practitioner-oriented comparison in the next set of experiments.

Experiment (e): validation-based real-data evaluation. We finally evaluate the proposed Local-Global construction under a more standard real-data protocol on the regression datasets listed in Table 4. All features are standardized to zero mean and unit variance.

For each dataset, we perform $n_{\text{splits}} = 5$ independent random splits. In each split, 20% of the samples are used for training. The remaining data are partitioned into a validation set (40% of the remainder) and a test set (60% of the remainder). Model selection is carried out on the validation set, and performance is reported on the independent test set.

Model selection. We compare two families of interpolating estimators.

- **Local kernel k_B.** The bandwidth c is selected by minimizing the validation MSE over the grid $c_{\text{base}} \in \{0.1, 0.2, 0.4, 0.7, 1.0\}$. The estimator is trained in the nominally ridgeless regime; however, we add a small numerical regularization $\lambda_\epsilon = 10^{-10}$.
- **Local-Global kernel k_{LG}.** The hyperparameters (c, ρ, q) are jointly selected by minimizing the validation MSE over the grid

$$c \in \{0.1, 0.2, 0.4, 0.7, 1.0\}, \quad \rho \in \{0.01, 0.1, 0.5, 1.0, 10\}, \quad q \in \{3, 4, 5, 6, 7, 8, 9, 10\}.$$

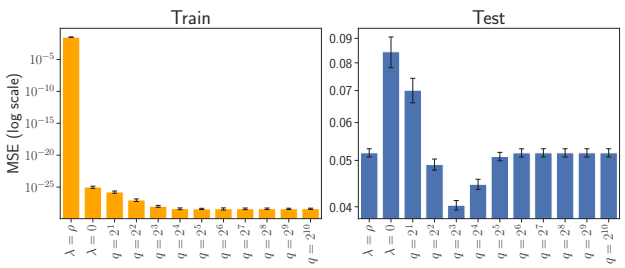
  After selection, the estimator is refit on the training set using the chosen parameters and evaluated on the test set.

For each split, we record both training and test MSE. Results are reported as mean $\pm$ standard deviation across splits.

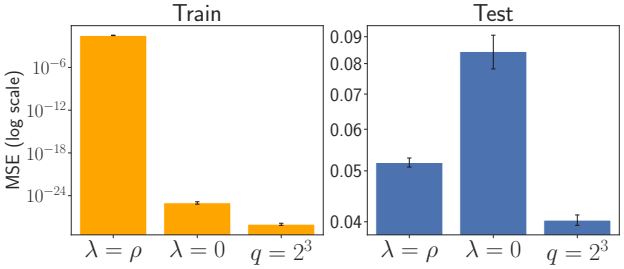
Tables 5 and 6 report the predictive performance of the two kernel families. The Local-Global estimator systematically operates in the interpolating regime, achieving near-zero training error across splits. In contrast, the local baseline does not always interpolate and exhibits noticeable training error on datasets such as California Housing and Energy.

On the test set, both methods operate in regimes that may be associated with overfitting. However, the nature of this overfitting differs. The local baseline exhibits poor and unstable generalization despite achieving low training error. By contrast, the interpolating Local-Global estimator maintains stable performance across splits (low standard deviation) together with improved test accuracy.

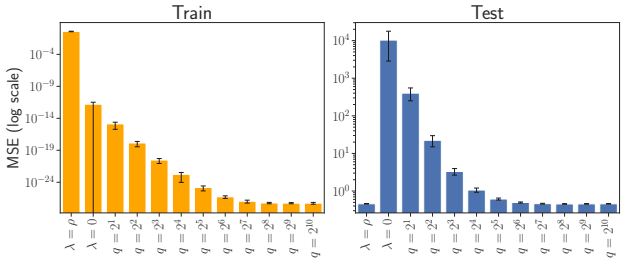
Table 7 lists the hyperparameters achieving the lowest test error among the evaluated configurations.



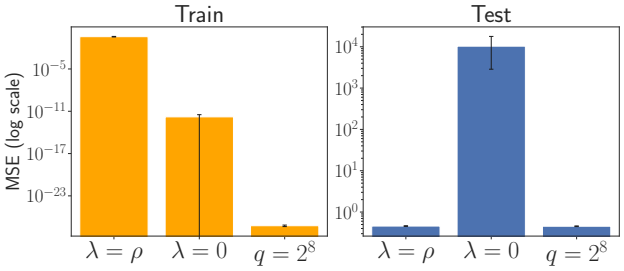
(a) Airfoil - full q sweep



(b) Airfoil - best q



(c) California Housing - full q sweep



(d) California Housing - best q

Figure 10: Real-data experiments (Experiment (d)). Top: Airfoil dataset. Bottom: California Housing dataset. Left: full evolution with q. Right: best-performing Local-Global model.

Table 5: Training MSE (mean $\pm$ standard deviation over 5 splits) for the local kernel k_B and the Local-Global kernel k_{LG} in the ridgeless regime.

Dataset	k_B MSE train	k_{LG} MSE train
airfoil	$(1.3 \cdot 10^{-19}) \pm (6.0 \cdot 10^{-20})$	$(1.4 \cdot 10^{-28}) \pm (1.0 \cdot 10^{-28})$
cal_housing	$(2.5 \cdot 10^{-1}) \pm (2.0 \cdot 10^{-2})$	$(1.1 \cdot 10^{-23}) \pm (6.0 \cdot 10^{-24})$
energy	$(8.7 \cdot 10^{-2}) \pm (3.0 \cdot 10^{-2})$	$(2.1 \cdot 10^{-22}) \pm (7.0 \cdot 10^{-23})$

G LOCAL-GLOBAL QUANTUM KERNELS AS QUANTUM EMBEDDING KERNEL

In this appendix, we show that the Local-Global quantum kernel can be interpreted as a standard quantum embedding kernel associated with a single unitary feature map acting on an enlarged Hilbert space.

Let $U(x)$ be a unitary quantum feature map acting on t qubits. To obtain the normalization we consider the case where $\sqrt{\lambda_L} + \sqrt{\lambda_G} = 1$ with $\lambda_L, \lambda_G \geq 0$.

Let ρ^L_{init} and ρ^G_{init} denote the initial states associated with the local and global kernels, see Definition 3.1, and define the corresponding embeddings

$$\rho^L_x = U(x)\rho^L_{\text{init}}U^\dagger(x), \qquad \rho^G_x = U(x)\rho^G_{\text{init}}U^\dagger(x). \tag{37}$$

We construct a single embedding acting on $t+1$ qubits. Define the enlarged initial state

$$\rho^{LG}_{\text{init}} = \sqrt{\lambda_L}\,|0\rangle\langle 0| \otimes \rho^L_{\text{init}} + \sqrt{\lambda_G}\,|1\rangle\langle 1| \otimes \rho^G_{\text{init}}, \tag{38}$$

and the unitary

$$W(x) = I_2 \otimes U(x), \tag{39}$$

where the first qubit acts as a control register.

The associated quantum embedding is

$$\rho^{LG}_x = W(x)\rho^{LG}_{\text{init}}W^\dagger(x). \tag{40}$$

Table 6: Test MSE (mean $\pm$ standard deviation over 5 splits) for the local kernel k_B and the Local-Global kernel k_{LG}.

Dataset	k_B MSE test	k_{LG} MSE test
airfoil	$(6.0 \cdot 10^{-2}) \pm (6.02 \cdot 10^{-3})$	$(3.0 \cdot 10^{-2}) \pm (1.45 \cdot 10^{-3})$
cal_housing	$(2.56 \cdot 10^{3}) \pm (3.76 \cdot 10^{3})$	$(8.0 \cdot 10^{-1}) \pm (3.0 \cdot 10^{-2})$
energy	$(5.26 \cdot 10^{2}) \pm (7.7 \cdot 10^{3})$	$(1.18 \cdot 10^{0}) \pm (8.0 \cdot 10^{-2})$

Table 7: Hyperparameters achieving the lowest test MSE across splits for each dataset. For k_B, we report the best bandwidth c^*_{base}. For k_{LG}, we report the best (c^*, ρ^*, q^*) together with the corresponding test error.

Dataset	n	d	n_{train}	n_{val}	n_{test}
cal_housing	20640	8	4128	6604	9908
energy	14980	14	2996	4793	7191
airfoil	7400	20	1480	2368	3552

By construction,

$$\rho_x^{LG} = \sqrt{\lambda_L}|0\rangle\langle 0| \otimes \rho_x^L + \sqrt{\lambda_G}|1\rangle\langle 1| \otimes \rho_x^G. \tag{41}$$

Consider the Hilbert-Schmidt inner product $\operatorname{Tr}\left[\rho_x^{LG}\rho_z^{LG}\right]$. Using the orthogonality relation $\operatorname{Tr}[|0\rangle\langle 0|\,|1\rangle\langle 1|] = 0$, the cross terms vanish and we obtain

$$\begin{aligned}
\operatorname{Tr}\left[\rho_x^{LG}\rho_z^{LG}\right] &= \lambda_L \operatorname{Tr}\left[|0\rangle\langle 0| \otimes \rho_x^L\rho_z^L\right] + \lambda_G \operatorname{Tr}\left[|1\rangle\langle 1| \otimes \rho_x^G\rho_z^G\right] && (42)\\
&= \lambda_L \operatorname{Tr}[|0\rangle\langle 0|]\operatorname{Tr}[\rho_x^L\rho_z^L] + \lambda_G \operatorname{Tr}[|1\rangle\langle 1|]\operatorname{Tr}[\rho_x^G\rho_z^G] && (43)\\
&= \lambda_L k_L(x,z) + \lambda_G k_G(x,z). && (44)
\end{aligned}$$

Therefore, the Local-Global kernel can be interpreted as a standard quantum embedding kernel obtained from a single unitary feature map acting on an enlarged Hilbert space. The additional qubit acts as a classical selector between local and global initializations, and the resulting Hilbert-Schmidt inner product exactly reproduces the weighted kernel.