On the Inconsistency of Kernel Ridgeless Regression in Fixed Dimensions

Daniel Beaglehole[*] Mikhail Belkin[†,*] Parthe Pandit[†]

Abstract. "*Benign overfitting*", the ability of certain algorithms to interpolate noisy training data and yet perform well out-of-sample, has been a topic of considerable recent interest. We show, using a fixed design setup, that an important class of predictors, kernel machines with translation-invariant kernels, does not exhibit benign overfitting in fixed dimensions. In particular, the estimated predictor does not converge to the ground truth with increasing sample size, for any non-zero regression function and any (even adaptive) bandwidth selection. To prove these results, we give exact expressions for the generalization error, and its decomposition in terms of an approximation error and an estimation error that elicits a trade-off based on the selection of the kernel bandwidth. Our results apply to commonly used translation-invariant kernels such as Gaussian, Laplace, and Cauchy.

1. Introduction.

Recent empirical evidence has shown that certain algorithms, contrary to classical learning theory, can interpolate noisy data, i.e., achieve zero training error, while still generalizing well out-of-sample, that is, exhibiting low test error [2, 20, 25]. This phenomenon of "benign overfitting" (using the terminology of [1]) has been rigorously analyzed for certain parametric methods such as linear regression, and random feature regression [1, 3, 11, 19], as well as non-parametric methods such as kernel regression with singular kernels [4, 6, 8].

Many theoretical results in this direction assume a high-dimensional regime where the data dimension d grows with the sample size n. However, it remains unclear whether this phenomenon is common when the data dimension is fixed. In particular, it has been an open question whether popular practical algorithms, such as kernel machines [10, 24], exhibit benign overfitting.

Indeed, the work of [13] showed that interpolating kernel machines, also known as kernel ridgeless regression, can be consistent in high dimension, i.e., can converge to an optimal predictor given enough data. On the other hand, the work of Rakhlin and Zhai [22] showed that for the specific case of Laplace kernel, kernel ridgeless regression is inconsistent in fixed dimensions even with a data-adaptive bandwidth. This is significant as the kernel bandwidth hyperparameter can have a large effect on the estimated predictor, and indeed can be set adaptively in high dimensions to achieve consistency.

In this work, we show that this lack of benign overfitting in fixed dimension is in fact a general property of a broad class of kernel machines. Specifically, we prove that consistency does not hold for the widely used class of translation-invariant kernels, i.e., kernels that depend only on the difference of the two inputs, under mild spectral conditions. Important examples of such kernels include the Gaussian, Laplace, and Cauchy kernels.

Our counterexample uses a simple data model of the grid on the unit circle for $d = 1$, and, in higher dimensions, a multidimensional torus, i.e., the product of unit circles, when $d > 1$. For clarity, we outline the $d = 1$ case in the main body of the paper, and generalize to $d > 1$ in

[*] Computer Science and Engineering, University of California, San Diego

[†] Halıcıoğlu Data Science Institute, University of California, San Diego

Appendix D.

To prove these results, we derive exact expressions for the generalization mean-squared-error in terms of the Fourier series of the chosen kernel. These exact expressions elucidate the trade-off between approximation and estimation errors when choosing the bandwidth parameter. Our key insight is that while a small bandwidth reduces the estimation error, it worsens the approximation error. Our exact expressions enable us to provide a constant lower bound on the generalization error as the number of samples grows to infinity.

Related work. Several recent works have demonstrated the existence of benign interpolation in high dimensions (e.g., when dimension is scales linearly with the number of samples). In this setting, the generalization bounds for linear and random feature interpolation depend on the rate of decay of eigenvalues [1, 11]. For example, [17] derives asymptotic risk curves in high dimensions for linear ridge regression and featurized linear ridge regression. Similarly [18] describes the asymptotic behavior of random feature regression, deriving double descent curves. As these works showcase, interpolation is benign in these high dimensional settings, typically proportional asymptotics. Another work considers the consistency of rotation-invariant kernels in high-dimensions [9].

In contrast, we consider the case of fixed dimensions. We show that in fixed dimensions, interpolation with kernel machines is inconsistent. We also strengthen our result by showing that this conclusion holds regardless of an adaptive bandwidth selection, which is often necessary to achieve consistency for high dimensional settings, e.g. [4, 6, 8, 13].

2. Problem setup.

Notation. We denote functions by lowercase letters a, sequences by uppercase letters A, vectors by lowercase bold letters $\boldsymbol{a}$, matrices by uppercase bold letters $\boldsymbol{A}$. Sequences are indexed using square-brackets, $A[k]$ where $k \in \mathbb{Z}$. For vectors, functions, sequences, $\langle \boldsymbol{a}, \boldsymbol{b} \rangle, \langle a, b \rangle, \langle A, B \rangle$ denote their Euclidean, L^2, and $\ell^2(\mathbb{Z})$ inner products respectively, while $\|\boldsymbol{a}\|, \|a\|, \|A\|$ denote corresponding induced norms, and $\|\boldsymbol{a}\|_1, \|a\|_1, \|A\|_1$ denote their respective 1-norms. Like the L^1 norm, other norms or inner products will be pointed out explicitly. For a nonnegative integer N, we denote the set $\{0, 1, \ldots, N-1\}$ by $[N]$. We use j to denote $\sqrt{-1}$, and an overline, $\overline{\boldsymbol{a}}$, to denote elementwise complex conjugation. The asymptotic *big-Oh* notation $O_n(\cdot), \Omega_n(\cdot), o_n(\cdot), \omega_n(\cdot)$, have their usual meaning where the limit is with respect to n.

We use $N \in \mathbb{N}$ as a *resolution* hyperparameter (explicitly defined in Equation (3.1)). For a sequence $G \in \ell^1(\mathbb{Z})$, and a fixed N, we define an N-*hop subsequence* $G_\ell \in \ell^1(\mathbb{Z})$ as

$$G_\ell = \{G[mN + \ell]\}_{m \in \mathbb{Z}} \qquad \text{defined for } \ell \in \{0, 1, \ldots, N-1\}. \tag{2.1}$$

Nonparametric regression. We consider a supervised learning problem in the fixed design setting where we have n labeled samples $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y} \subseteq \mathbb{R}^d \times \mathbb{R}$, with labels generated as,

$$y_i = f^*(x_i) + \xi_i, \qquad \xi_i \overset{\text{i.i.d.}}{\sim} \mathbb{P}_\xi, \qquad \forall\, i \in [n],$$

for some unknown target function f^*. The noise distribution $\mathbb{P}_\xi$ is centered with a finite variance $\sigma^2 > 0$. We assume this distribution is independent of the chosen data $\{x_i\}$ and target f^*.

For a sequence of datapoints X_n, the estimation task is to propose an estimator $\widehat{f}_n = \widehat{f}_n(X_n, \boldsymbol{y}) : \mathcal{X} \to \mathbb{R}$, where $\boldsymbol{y} = (y_i) \in \mathbb{R}^n$ is the vector of all labels on these data. An estimator's performance (or generalization error) is measured in terms of its mean squared error,

$$\mathsf{MSE}\left(\widehat{f}_n, f^*\right) := \left\|\widehat{f}_n - f^*\right\|^2 = \int_{\mathcal{X}} \left(\widehat{f}_n(x) - f^*(x)\right)^2 \mathrm{d}x.$$

Weak consistency [10]. For a target function f^*, a sequence of estimators $\left\{\widehat{f}_n\right\}$ is said to be weakly consistent if,

$$\lim_{n\to\infty} \mathbb{E}_{\boldsymbol{\xi}} \, \mathsf{MSE}\left(\widehat{f}_n, f^*\right) = 0.$$

In this paper we show that a certain sequence – kernel ridgeless regression estimators – is weakly inconsistent, i.e., $\lim_{n\to\infty} \mathbb{E}_{\boldsymbol{\xi}} \, \mathsf{MSE}\left(\widehat{f}_n, f^*\right) > 0$. Note that weak inconsistency implies inconsistency in the strong sense as well.

Kernel interpolation. (also known as kernel ridgeless regression) For an RKHS $\mathcal{H}$, the kernel interpolation estimator is given by,

$$\widehat{f}_n = \underset{f\in\mathcal{H}}{\operatorname{argmin}} \; \|f\|_{\mathcal{H}} \qquad \text{subject to} \quad f(x_i) = y_i \quad \text{for } i = \{1, 2, \ldots, n\}. \tag{2.2}$$

The name ridgeless is due to the fact that the solution is equivalent to the following *kernel ridge regression* problem in the limit

$$\widehat{f}_n = \lim_{\lambda\to 0^+} \underbrace{\left(\underset{f\in\mathcal{H}}{\operatorname{argmin}} \sum_{i=1}^{n} (f(x_i) - y_i)^2 + \lambda \|f\|_{\mathcal{H}}^2 \right)}_{:=\widehat{f}_{n,\lambda}}. \tag{2.3}$$

Every RKHS is in one-to-one correspondence with a positive definite kernel function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. Define the kernel matrix $\mathbf{K} = (k(x_i, x_j))$ of pairwise evaluations of the kernel on the training data. Due to the representer theorem [23], the solution to (2.2) lies in the span of n basis functions $K(x_i, x)$ and can be written as

$$\text{(Kernel interpolation)} \qquad \widehat{f}_n(x) = \sum_{i=1}^{n} \widehat{\alpha}_i K(x_i, x), \qquad \widehat{\boldsymbol{\alpha}} = (\widehat{\alpha}_i) \in \mathbb{R}^n \qquad \widehat{\boldsymbol{\alpha}} := \mathbf{K}^{-1}\boldsymbol{y},$$

where $\boldsymbol{y} \in \mathbb{R}^n$ is the vector of all labels. The above follows as a direct consequence of $\widehat{f}_{n,\lambda} = K(\cdot, X)(\mathbf{K} + \lambda I_n)^{-1}\boldsymbol{y}$, and that $\widehat{f}_n = \lim_{\lambda\to 0^+} f_{n,\lambda}$. The matrix $\mathbf{K}$ is invertible because the kernel is positive definite, otherwise interpolation in an RKHS is not always possible. The (Riesz) representer of a given kernel K at a datum $x_\star$ is an element of $\mathcal{H}$, denoted by $K(x_\star, \cdot) : \mathcal{X} \to \mathbb{R}$. It is the evaluation functional of $x_\star \in \mathcal{X}$, i.e., $\langle f, K(x_\star, \cdot)\rangle_{\mathcal{H}} = f(x_\star)$ for all $f \in \mathcal{H}$. The basis functions above are thus the representers of the training data $\{x_1, x_2, \ldots x_n\}$.

We define the *restriction operator* R_n, and its adjoint, the *extension operator* R_n^*, as follows:

$$R_n : \mathcal{H} \to \mathbb{R}^n \qquad R_n f := (f(x_i)) \in \mathbb{R}^n, \qquad \forall\, f \in \mathcal{H} \tag{2.4}$$

$$R_n^* : \mathbb{R}^n \to \mathcal{H} \qquad R_n^* \boldsymbol{\alpha} := \sum_{i=1}^{n} \alpha_i K(x_i, \cdot) \in \mathcal{H}, \qquad \forall\, \boldsymbol{\alpha} = (\alpha_i) \in \mathbb{R}^n \tag{2.5}$$

that evaluates the function on the data. Here, since $L_n^2 \cong \mathbb{R}^n$ are isometric, we are abusing notation in favour of simpler expressions. This gives us the following equations

$$\boldsymbol{y} = R_n f^* + \boldsymbol{\xi}, \qquad \text{and} \qquad \widehat{f}_n = R_n^* \mathbf{K}^{-1} \boldsymbol{y}.$$

For an RKHS we have two data dependent operators, the *integral operator* and the *empirical operator*, respectively given by,

$$\mathcal{T}_K f(x) = \int_{\mathcal{X}} K(x,z) f(z)\,\mathrm{d}z, \tag{2.6}$$

$$\mathcal{T}_K^n f(x) = \sum_{z \in X_n} K(x,z) f(z). \tag{2.7}$$

Eigenfunctions of $\mathcal{T}_K$ that form a countable orthonormal basis of $L^2(\mathcal{X})$ can be used to provide an alternate representation for the $\mathcal{H}$-norm via the identity,

$$\langle f, g \rangle_{\mathcal{H}} = \sum_{k \in \mathbb{Z}} \frac{\langle f, \varphi_k \rangle \langle g, \varphi_k \rangle}{\sigma_k} \qquad \|f\|_{\mathcal{H}}^2 = \sum_{k \in \mathbb{Z}} \frac{\langle f, \varphi_k \rangle^2}{\sigma_k}$$

where (σ_k, φ_k) is an eigen-pair, i.e., $\mathcal{T}_K \varphi_k = \sigma_k \cdot \varphi_k$, with $\sigma_k \in \mathbb{R}_+$ and $\varphi_k \in L^2$.

Fourier analysis:. We recall some useful quantities from Fourier analysis to be used later.

Definition 2.1 (Fourier basis). *Let $\phi_k(x) = e^{jkx}$ for $k \in \mathbb{Z}$, which satisfy*

$$\langle \phi_k, \phi_\ell \rangle := \int_{-\pi}^{\pi} \phi_k(t) \overline{\phi_\ell(t)} \frac{\mathrm{d}t}{2\pi} = \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{j(k-\ell)t}\,\mathrm{d}t = \begin{cases} 1 & k = \ell \\ 0 & k \neq \ell \end{cases}.$$

The normalization factor $\frac{1}{2\pi}$ comes from the uniform density on $[-\pi, \pi)$. An important tool in our analysis is the Fourier series representation of functions $\mathcal{X} \mapsto \mathbb{R}$. In general, any integrable function $\mathbb{R} \to \mathbb{R}$ periodic with period 2π, admits such a representation.

Definition 2.2 (Fourier Series). *For $f \in L^1_{[-\pi,\pi)}$, let F be the Fourier series indexed by $k \in \mathbb{Z}$,*

$$f(t) = \sum_{k \in \mathbb{Z}} F[k] \phi_k(t) = \sum_{k \in \mathbb{Z}} F[k] e^{jkt}, \qquad \forall\, t \in [-\pi, \pi)$$

$$F[k] = \langle f, \phi_k \rangle = \int f(t) \overline{\phi_k(t)} \frac{\mathrm{d}t}{2\pi} = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(t) e^{-jkt}\,\mathrm{d}t, \qquad \forall\, k \in \mathbb{Z}\ .$$

Definition 2.3 (DFT Matrix). *The normalized discrete Fourier transform (DFT) matrix is*

$$\boldsymbol{U} = \begin{bmatrix} \boldsymbol{u}_0 & \cdots & \boldsymbol{u}_{N-1} \end{bmatrix}, \qquad \boldsymbol{u}_\ell = \frac{1}{\sqrt{N}} \begin{bmatrix} 1 & e^{-j\frac{2\pi}{N}\ell} & \ldots & e^{-j\frac{2\pi}{N}(N-1)\ell} \end{bmatrix}^\top, \quad \ell \in [N].$$

Notice that $\boldsymbol{U}\boldsymbol{U}^{\mathsf{H}} = \boldsymbol{U}^{\mathsf{H}}\boldsymbol{U} = \boldsymbol{I}$, where we use $^{\mathsf{H}}$ to denote the conjugate transpose (hermitian) of a matrix.

Proposition 2.4 (Parseval's theorem). *For a continuous function $f : [-\pi, \pi) \to \mathbb{R}$ with Fourier series F,*

$$\frac{1}{2\pi} \int_{-\pi}^{\pi} |f(t)|^2\,\mathrm{d}t = \sum_{k \in \mathbb{Z}} |F[k]|^2.$$

3. Model.

We now describe our setting and state our main result: kernel interpolation is weakly inconsistent.

Data design (grid on the unit circle). We describe the case of $d = 1$ and focus on $\mathcal{X} = [-\pi, \pi)$, viewed as the unit circle. An extension to $d > 1$ is deferred to Appendix D where we consider $[-\pi, \pi)^d$. We consider discrete, evenly-spaced grids indexed by a *resolution* hyperparameter $N \in \mathbb{N}$, given by

$$X_N = \{x_0, \dots, x_{N-1}\} \qquad x_i := \frac{2\pi}{N} i - \pi \qquad \forall\, i = 0, \dots, N-1. \tag{3.1}$$

We call N the resolution parameter of the grid on $[-\pi, \pi)$, and assume N is even for simplicity. Observe that Riemannian sums over the grid X_N for integrable functions converge to integrals on the continuum $[-\pi, \pi)$. Alternatively, the empirical distribution on the grid weakly converges to the uniform measure on the continuum. Note for $d = 1$, the total number of samples n equals the resolution N.

For $d > 1$, we consider $\mathcal{X} = [-\pi, \pi)^d$, the product of d unit circles, and the respective grids, along each dimension. Thus N is the number of samples per dimension, whereby the total number of samples $n = N^d$.

Translation-invariant kernels. We consider (periodic) kernels parameterized by a positive bandwidth parameter[1] M,

$$K(x, x') = g\left(M(x - x' \bmod [-\pi, \pi))\right), \qquad x, x' \in \mathcal{X}$$

for some even function $g : \mathbb{R} \to \mathbb{R}$, where we denote,

$$\theta \bmod [-\pi, \pi) = ((\theta + \pi) \bmod 2\pi) - \pi \in [-\pi, \pi)\ . \tag{3.2}$$

We denote the RKHS corresponding to K by $\mathcal{H}$. For ease of notation, when $M = 1$, we refer to this as the base kernel and the base RKHS $\mathcal{H}_0$. Define $G_0, G : \mathbb{Z} \to \mathbb{C}$ as the Fourier series of g, i.e.,

$$g(M(\theta \bmod [-\pi, \pi))) = \sum_{k \in \mathbb{Z}} G[k] \exp(jk\theta)\ , \qquad G[k] = \frac{1}{2\pi} \int_{-\pi}^{\pi} g(M\theta) \exp(-jk\theta)\, \mathrm{d}\theta\ , \tag{3.3a}$$

$$g(\theta \bmod [-\pi, \pi)) = \sum_{k \in \mathbb{Z}} G_0[k] \exp(jk\theta)\ , \quad G_0[k] = \frac{1}{2\pi} \int_{-\pi}^{\pi} g(\theta) \exp(-jk\theta)\, \mathrm{d}\theta\ . \tag{3.3b}$$

While usually the bandwidth scales the input, we note our analysis also holds for different mechanisms that satisfy the kernel assumptions given later. For symmetric positive definite kernels, g is an even function whereby we have that G is real. Furthermore g is real whereby,

$$G[-k] = G[k] > 0 \quad \forall\, k \in \mathbb{Z}\ .$$

Proposition 3.1. *$\boldsymbol{u}_\ell$ and $\overline{\boldsymbol{u}_\ell}$ are eigenvectors of $\mathbf{K} = (K(x_i, x_j)) \in \mathbb{R}^{N \times N}$, with eigenvalue $\lambda_\ell = N \|G_\ell\|_1$,* i.e., $\mathbf{K}\boldsymbol{u}_\ell = \lambda_\ell \boldsymbol{u}_\ell$ and $\mathbf{K}\overline{\boldsymbol{u}_\ell} = \lambda_\ell \overline{\boldsymbol{u}_\ell}$. Furthermore,

$$\mathbf{K} = \sum_{\ell=0}^{N-1} \lambda_\ell \boldsymbol{u}_\ell \boldsymbol{u}_\ell^{\mathsf{H}}, \qquad \mathbf{K}^{-1} = \sum_{\ell=0}^{N-1} \frac{1}{\lambda_\ell} \boldsymbol{u}_\ell \boldsymbol{u}_\ell^{\mathsf{H}}, \qquad \mathbf{K}^{-2} = \sum_{\ell=0}^{N-1} \frac{1}{\lambda_\ell^2} \boldsymbol{u}_\ell \boldsymbol{u}_\ell^{\mathsf{H}}.$$

[1] In machine learning literature, the bandwidth may often be denoted as $1/M$ instead.

Proposition 3.2. *For any $M > 0$, the Fourier basis are eigenfunctions of the kernel integral operator $\mathcal{T}_K$ with eigenvalues G,* i.e., we have,

$$\mathcal{T}_K \phi_k = G[k] \cdot \phi_k \ .$$

The proofs to these propositions are provided in Appendix E.1.

For X_N, we define the *restriction operator* R_N, and its adjoint, the *extension operator*,

$$R_N : \mathcal{H} \to \mathbb{R}^N \ , \qquad R_N f = \left(f\left(\frac{2\pi}{N}(i-1) - \pi\right)\right)_i \in \mathbb{R}^N \ , \tag{3.4}$$

$$R_N^* : \mathbb{R}^N \to \mathcal{H} \ , \qquad R_N^* \boldsymbol{\alpha} := \sum_{i=0}^{N-1} \alpha_i K(x_i, \cdot) \in \mathcal{H} \ . \tag{3.5}$$

We also use the notation

$$\langle \boldsymbol{\alpha}, K(X_N, \cdot)\rangle_N := \sum_{i=0}^{N-1} \alpha_i K(x_i, \cdot)$$

to keep expressions simple. With this notation, the labels and the kernel interpolator can be written as

$$\widehat{f}_N = R_N^* \mathbf{K}^{-1} \boldsymbol{y} = \left\langle \mathbf{K}^{-1} \boldsymbol{y}, K(X_N, \cdot)\right\rangle_N \ . \tag{3.6}$$

Definition 3.3 (Span of Riesz Representers). *Functions in the range of R_N^*, and of $\mathcal{T}_K^N$, are in the span of the representers $\{K(x_i, \cdot)\}_{i=1}^N$.*

Target function. We assume the target function lies in the base RKHS $\mathcal{H}_0$, i.e., $\mathcal{H}$ with $M = 1$, and has a norm $\|f^*\|_{\mathcal{H}_0} = O_{M,N}(1)$. As the target function is defined on the unit circle, it admits a Fourier series,

$$f^* = \sum_{k \in \mathbb{Z}} V[k] \phi_k \ . \tag{3.7}$$

To keep derivations simple, we will assume, without loss of generality, that $V[k] \in \mathbb{R}$ for all k (i.e. the target function is even). It is straightforward to extend this argument to all f^*. We can decompose f^* into an even and odd component (by $f^*(x) = \frac{f^*(x)+f^*(-x)}{2} + \frac{f^*(x)-f^*(-x)}{2}$). The even component will only have a cosine series (and hence real $V[k]$), and the odd component will only have a sine series (imaginary $V[k]$). The argument for the case of targets with imaginary $V[k]$ is identical to that for targets with real $V[k]$. Even and odd functions are in orthogonal subspaces of L^2 and of $\mathcal{H}$, whereby for general complex $V[k]$, the errors we derive are the sum of the errors of the even and odd components, and the arguments go through.

Recall the definition of the restriction and extension operators in (3.4). Let P_X be the L^2-projection operator onto the span of the representers, i.e.,

$$P_X f := \operatorname*{argmin}_{h\in\mathcal{H}} \left\{ \|f-h\| \;\middle|\; h = \sum_{i=1}^{N} \alpha_i K(x_i,\cdot) \text{ for some } (\alpha_i) \in \mathbb{R}^N \right\} , \tag{3.8a}$$

$$\boldsymbol{\alpha}^* := (\alpha_i^*) \qquad \text{such that} \qquad P_X f^* = \sum_{i=1}^{N} \alpha_i^* K(x_i,\cdot) , \tag{3.8b}$$

$$f_\perp^* := f^* - P_X f^*, \tag{3.8c}$$

where $f_\perp^*$ is orthogonal to all functions in $\operatorname{Span}\{K(x_i,\cdot)\}$. An immediate identity using the evaluation operator R_N is,

$$R_N P_X f^* = \mathbf{K}\boldsymbol{\alpha}^* \qquad \text{and} \qquad R_N^* \boldsymbol{\alpha}^* = P_X f^* .$$

We can decompose the target function as

$$f^* = P_X f^* + f_\perp^* = \sum_{i=0}^{N-1} \alpha_i^* K(x_i,\cdot) + f_\perp^* = \langle \boldsymbol{\alpha}^*, K(X_N,\cdot)\rangle_N + f_\perp^* .$$

Using this, the vector of labels, and the kernel interpolation estimator can be written as,

$$\boldsymbol{y} = R_N f^* + \boldsymbol{\xi} = R_N P_X f^* + R_N f_\perp^* + \boldsymbol{\xi} = \mathbf{K}\boldsymbol{\alpha}^* + R_N f_\perp^* + \boldsymbol{\xi} , \tag{3.9}$$

$$\widehat{f}_N = R_N^* \mathbf{K}^{-1} \boldsymbol{y} = P_X f^* + \left\langle \mathbf{K}^{-1} R_N f_\perp^*, K(X_N,\cdot)\right\rangle_N + \left\langle \mathbf{K}^{-1}\boldsymbol{\xi}, K(X_N,\cdot)\right\rangle_N , \tag{3.10}$$

where we have used the expression from Equation (3.6).

4. Main result: Inconsistency of kernel interpolation.

Our main result holds under certain assumptions on the translation-invariant kernels. Below, we assume M', i, i^* are all non-negative integers. Recall that G is the Fourier series of the kernel function, see Equation (3.3a). Note G depends on M but G_0, the Fourier series of the kernel corresponding to $\mathcal{H}_0$ - the base RKHS, does not.

Assumption 1 (Integrability). *We assume the kernel is integrable. In particular, the integral* $\int_{-\pi}^{\pi} g(Mx)\,\mathrm{d}x$ *exists and is finite for all* $0 < M < \infty$.

Assumption 2 (Spectral Tail). *For all* $k \in \mathbb{Z}_{\geq 0}$*, there exists a constant* $C_1 > 0$ *such that,*

$$|G[M'k+i]| \leq \frac{C_1 |G[i]|}{1+k^2} \tag{4.1}$$

holds for all $M' \geq M > 0$ *and for all* $i \leq M'$*, except* $o_{M'}(M')$ *many.*

Assumption 3 (Spectral Head). *There exist constants* $C_2, C_3 \in \mathbb{R}_+$ *and* $i^* \in \mathbb{Z}_{\geq 0}$ *such that for* $M \geq C_2$*, we have that for all* $0 \leq M' < M$, $|G[i^*]| \leq C_3 |G[i^* + M']|$ *and* $|G_0[i^*]| > 0$.

To simplify analysis for many kernel functions, we give a sufficient condition that implies Assumptions 1-3, and is easy to verify for many functions.

Condition 1 (Monotonic Boundedness). *There exist constants (independent of the bandwidth M) $c, C, C', C'' > 0$, a constant $c'(M) > 0$ (that may depend on M), and a bounded, monotonically decreasing function $f : \mathbb{R}_{\geq 0} \to \mathbb{R}$ with (i) $0 < \frac{f(x+k)}{f(x)} \leq \frac{C''}{1+k^2}$ for all $x \in \mathbb{R}_{\geq 0}, k \in \mathbb{Z}$, and (ii) $\frac{f(1+x)}{f(x)} \geq C'$ for $0 \leq x \leq 1$, such that*

$$cf(\tfrac{k}{M}) \leq \frac{G[k]}{c'(M)} \leq Cf(\tfrac{k}{M})$$

for all $k \in \mathbb{Z}_{\geq 0}$, $M \in \mathbb{R}^+$.

The proof of the following propositions are provided in Appendix E.

Proposition 4.1. *If the Fourier series coefficients $G[i]$ satisfy Condition 1 (Monotonic Boundedness), then the kernel satisfies Assumptions 1-3.*

Proposition 4.2. *The Gaussian $G[k] = \exp(-\frac{k^2}{M^2})$, Laplacian $G[k] = \frac{1}{1+\frac{k^2}{M^2}}$, and Cauchy $G[k] = \exp(-\frac{|k|}{M})$ kernels (wrapped on the circle) satisfy Condition 1 (Monotonic Boundedness).*

We comment on each of the Assumptions 1-3 below.

Remark 1 (Note on Assumption 1). *A sufficient condition for our result is $|G[k]| < \infty$ for all $k \in \mathbb{Z}$. Assumption 1 implies this inequality by the definition of the Fourier coefficients.*

Remark 2 (Square-integrable derivative $\implies$ Assumption 2). *Combined with Assumption 1, the exchange formula [14] for Sobolev spaces implies Assumption 2 is equivalent to the kernel and its first derivative being L^2-integrable (for fixed bandwidth). Therefore, this assumption can be viewed as a condition on the smoothness of the kernel.*

Remark 3 (Interpretation of Assumption 3). *Intuitively, Assumption 3 enforces flatness in the frequency domain, or equivalently, sharpness in the $\mathcal{X}$-domain. A larger bandwidth M leads to a longer sequence of similar coefficients $G[k]$ for $k \in \{i^*, \dots, i^* + M\}$, giving a sharper kernel in the $\mathcal{X}$-domain.*

We present the main results in the following theorems. Recall that $\mathcal{H}_0$ is the base RKHS.

Theorem 4.3 (Inconsistency for all functions when G is monotonically bounded). *Consider a fixed non-zero regression function f^* that (i) has square-integrable zeroth and first derivatives, and (ii) can be expressed as a convergent Fourier series. Then, interpolation with a real-valued translation-invariant kernel satisfying Condition 1 is inconsistent for f^*, for any bandwidth, even if chosen adaptively.*

Recall the definition of the base RKHS $\mathcal{H}_0$, above equation (3.3), corresponding to the kernel with bandwidth $M = 1$.

Theorem 4.4 (Inconsistency for all Bandwidths). *For any translation-invariant kernel satisfying Assumptions 1-3, there exists a function with constant $\mathcal{H}_0$-norm for which kernel interpolation is inconsistent for any bandwidth, even adaptive to the data set.*

Theorem 4.5 (Inconsistency for all Functions ($M < N$)). *For any translation-invariant kernel satisfying Assumptions 1-2, with any (even data-adaptive) bandwidth $M \leq N$, kernel* *interpolation is inconsistent for all targets that can be expressed as convergent Fourier series. In particular, kernel interpolation with a fixed bandwidth is inconsistent for all such targets.*

To prove these results we apply Fourier analysis to compute an exact expression for the MSE of kernel interpolation. We decompose the MSE for a target function into three components - (i) an approximation error, measuring how close the target function is to the span of the representers, (ii) a noiseless estimation error, measuring the error in the absence of noise, and (iii) a noisy estimation error, measuring the average error if the target function is 0.

We then apply Parseval's Theorem, which relates these errors terms to the Fourier series of the target function, and of the kernel. Proving that the MSE is bounded away from 0 will rely on our assumptions on the tail and the head of the kernel spectrum.

5. Decomposition of the mean squared error. We now derive an exact expression for the MSE as a sum of three error terms: the approximation error, the noise-free estimation error, and the noisy estimation error. This useful expression will allow us to prove the main theorems of the previous section. Recall the definition of $f_\perp^*, \boldsymbol{\alpha}^*, P_X f^*$ from Equation (3.8).

Lemma 5.1 (MSE Decomposition). *For any square integrable target function f^*, the kernel interpolation $\widehat{f}_N$ estimator satisfies,*

$$\mathbb{E}_{\boldsymbol{\xi}} \operatorname{MSE}\left(\widehat{f}_N, f^*\right) = \underbrace{\|f^* - P_X f^*\|^2}_{\text{Approximation Error}} + \underbrace{\left\|\left\langle \mathbf{K}^{-1} R_N \left\{f^* - P_X f^*\right\}, K(X_N, \cdot)\right\rangle_N\right\|^2}_{\text{Noise-free Estimation Error}}$$
$$+ \underbrace{\mathbb{E}_{\boldsymbol{\xi}} \left\|\left\langle \mathbf{K}^{-1}\boldsymbol{\xi}, K(X_N, \cdot)\right\rangle_N\right\|^2}_{\text{Averaged Noisy Estimation Error}} .$$

Proof. Since $P_X f^* - \widehat{f}_N \in \operatorname{Span}\{K(x_i, \cdot)\}$, the Pythagorean theorem for the triangle $\left\{f^*, P_X f^*, \widehat{f}_N\right\}$, yields,

$$\operatorname{MSE}\left(\widehat{f}_N, f^*\right) = \left\|f^* - \widehat{f}_N\right\|^2 = \underbrace{\|f^* - P_X f^*\|^2}_{\text{Approximation Error}} + \underbrace{\left\|P_X f^* - \widehat{f}_N\right\|^2}_{\text{Estimation Error}} .$$

Notice that the estimation error above is random, due to the randomness in $\boldsymbol{\xi}$, which affects $\widehat{f}_N$. Using Equation (3.10), we can further decompose the average estimation error into two error terms,

$$\underbrace{\mathbb{E}_{\boldsymbol{\xi}} \left\|P_X f^* - \widehat{f}_N\right\|^2}_{\text{Estimation Error}} = \mathbb{E}_{\boldsymbol{\xi}} \left\|\left\langle \mathbf{K}^{-1} R_N f_\perp^*, K(X_N, \cdot)\right\rangle_N + \left\langle \mathbf{K}^{-1}\boldsymbol{\xi}, K(X_N, \cdot)\right\rangle_N\right\|^2 ,$$
$$= \underbrace{\left\|\left\langle \mathbf{K}^{-1} R_N f_\perp^*, K(X_N, \cdot)\right\rangle_N\right\|^2}_{\text{Noise-free Estimation Error}} + \underbrace{\mathbb{E}_{\boldsymbol{\xi}} \left\|\left\langle \mathbf{K}^{-1}\boldsymbol{\xi}, K(X_N, \cdot)\right\rangle_N\right\|^2}_{\text{Noisy Estimation Error}} .$$

where the cross term cancels out since the noise is centered. This concludes the proof. ■

Computing each of these terms individually, we derive the following expression for the unit circle. Recall the definition of the *N-hop subsequences* from Equation (2.1).

Lemma 5.2. *For a target function $f^* = \sum_{k\in\mathbb{Z}} V[k]\phi_k$, we have*

(a) Approximation error:

$$\mathcal{E}^{\mathsf{apx}} := \|f^* - P_X f^*\|^2 = \left(\sum_{i=0}^{N-1} \|V_i\|^2 - \frac{\langle G_i, V_i\rangle^2}{\|G_i\|^2}\right) = \sum_{i=0}^{N-1} \mathcal{E}_i^{\mathsf{apx}} \,.$$

(b) Noise-free estimation error:

$$\mathcal{E}^{\mathsf{free}} := \left\|\left\langle \mathbf{K}^{-1} R_N f_\perp^*, K(X_N, \cdot)\right\rangle_N\right\|^2 = \sum_{i=0}^{N-1} \frac{1}{N}\left(\frac{\langle V_i, \mathbf{1}\rangle}{\langle G_i, \mathbf{1}\rangle} - \frac{\langle G_i, V_i\rangle}{\|G_i\|^2}\right)^2 \|G_i\|^2 = \sum_{i=0}^{N-1} \mathcal{E}_i^{\mathsf{free}} \,.$$

(c) Averaged noisy estimation error:

$$\mathcal{E}^{\mathsf{noisy}} := \mathbb{E}_{\boldsymbol{\xi}} \left\|\left\langle \mathbf{K}^{-1}\boldsymbol{\xi}, K(X_N, \cdot)\right\rangle_N\right\|^2 = \sum_{i=0}^{N-1} \frac{\sigma^2}{N}\left(\frac{\|G_i\|}{\|G_i\|_1}\right)^2 = \sum_{i=0}^{N-1} \mathcal{E}_i^{\mathsf{noisy}} \,.$$

Together, this yields that the MSE for the function f^* is,

$$\mathbb{E}_{\boldsymbol{\xi}} \,\mathsf{MSE}\left(\widehat{f}_N, f^*\right) = \sum_{i=0}^{N-1} \mathcal{E}_i = \sum_{i=0}^{N-1} \mathcal{E}_i^{\mathsf{apx}} + \mathcal{E}_i^{\mathsf{free}} + \mathcal{E}_i^{\mathsf{noisy}} \,, \tag{5.2}$$

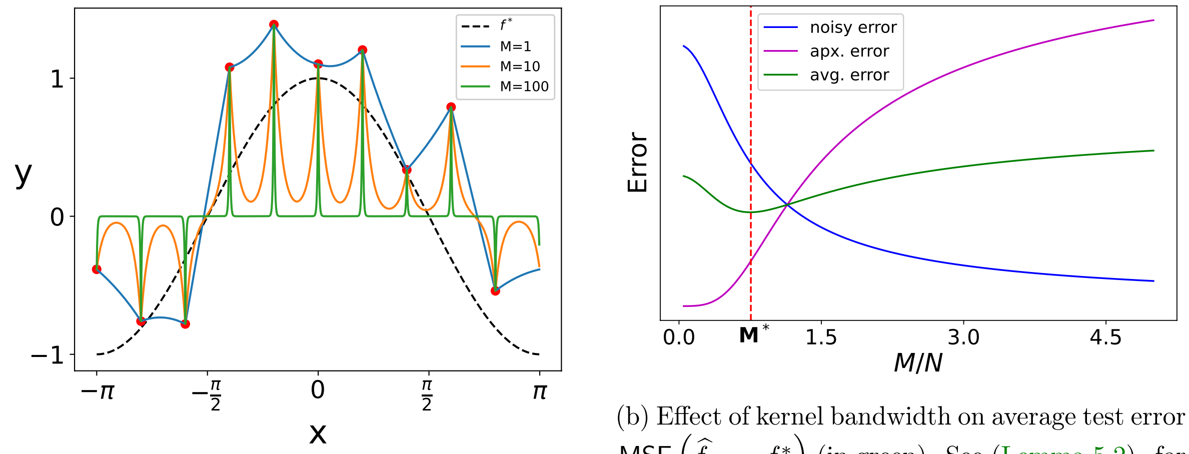
$$\mathcal{E}_i := \|V_i\|^2 - \frac{\langle G_i, V_i\rangle^2}{\|G_i\|^2} + \frac{1}{N}\left(\frac{\langle V_i, \mathbf{1}\rangle}{\langle G_i, \mathbf{1}\rangle} - \frac{\langle G_i, V_i\rangle}{\|G_i\|^2}\right)^2 \|G_i\|^2 + \frac{\sigma^2}{N}\left(\frac{\|G_i\|}{\|G_i\|_1}\right)^2 \,. \tag{5.3}$$

Appendices B.1, B.2, and B.3 provide proofs for Lemma 5.2 (a), (b), and (c) respectively.

6. Numerical experiments. We present experimental results that corroborate our theory. We visualize the effect of kernel bandwidth and regularization on the predictor and test error.

Effect of bandwidth on predictor. We visualize the effect of bandwidth on kernel interpolator with the Laplace kernel in one dimension (Figure 6.1a). On the y-axis we show the predicted values of our estimator (in blue) and the target function $f^*(x) = \cos(x)$ (in orange) with noise level $\sigma^2 = 1$. We notice that for small bandwidth (see $M = 2$ plot) kernel interpolation resembles piecewise linear interpolation. Meanwhile, interpolation with high bandwidth converges pointwise (except on a set of measure 0) to the 0 function (see $M = 200$ plot). Choosing an intermediate bandwidth does not recover the target function either ($M = 20$).

Effect of bandwidth on error. We also plot the effect of bandwidth on the exact expected error predicted by our theory (Figure 6.1b). In this experiment, we study the predicted error of our theory using Laplace kernel interpolation with a noise level $\sigma^2 = 1$ on a target function $f^*(x) = \cos(x)$. We plot the approximation and noisy estimation errors. We omit the noise-free estimation error as this is typically correlated with approximation error. Our theory predicts that the optimal bandwidth is roughly $M/N = 1$, exactly the point we use to split the cases in the proofs of the main theorems. Interestingly, our theory predicts a trade-off between the approximation error and the error due to noise (noisy estimation error). Larger bandwidths M allow you to fit noise benignly, at the cost of increased approximation error. Smaller bandwidths allow you to approximate well, but suffer in estimation error.



(a) Effect of kernel bandwidth M on predictor $\widehat{f}_{N,M}$. Here resolution is $N = 10$, noise variance is $\sigma^2 = 0.25$, and target function is $f^*(x) = \cos(x)$.

(b) Effect of kernel bandwidth on average test error $\mathsf{MSE}\left(\widehat{f}_{N,M}, f^*\right)$ (in green). See (Lemma 5.2). for a detailed definition of each error term in the error decomposition. Here resolution $N = 20$, noise variance $\sigma^2 = 1$, and target function is $f^* = \cos(x)$.

Figure 6.1: Effect of kernel bandwidth on (6.1a) predictor and (6.1b) test error.

Effect of regularization. Standard kernel ridge regression (KRR) will prevent interpolation and enable consistent estimation of the target function. However, one can perform interpolation with a modified kernel that mimics regularization to improve generalization while continuing to interpolate. For example, we modify the laplace kernel K on the unit circle to create a new kernel $\widetilde{K}(x, x') = K(x, x') + \lambda K(M(x - x'))$ for $M = 50$ and regularization parameter $\lambda = 1$. We compare this modified kernel to Laplace KRR with regularization parameters $\lambda \in \{0, 1\}$ in Figure 6.2.

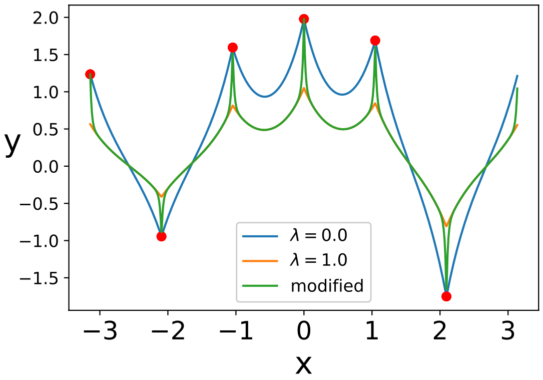


Figure 6.2: Modified kernel to mimic regularization (for $f^*(x) = \cos(x)$).

Benefits of Regularization. Our results show that positive regularization allows one to decrease the noisy estimation error at the expense of additional approximation error. Moreover, the faster the decay of the target function's Fourier coefficients, the less regularization worsens the approximation error. To understand this, we note that adding positive regularization, in effect, adds a Dirac δ-function to the kernel at the origin. In the Fourier domain, this is equivalent to adding an infinitesimal to all of the Fourier coefficients. To understand how this may help generalization, consider adding a small quantity $\Delta > 0$ to each of the first $1/\Delta$ Fourier coefficients. We analyze our MSE expression in Lemma

5.2. For a fixed function, adding this Δ will have a vanishing effect on the approximation error as $\Delta \to 0$. However, adding this Δ will decrease the noisy estimation error by extending the tail of G, making the ratio $\|G_i\|/\|G_i\|_1$ smaller. We show how a similar modification to the Laplace kernel will cause the interpolated solution to resemble the regularized solution in Figure 6.2.

7. Discussion and Outlook.

Following the connection of wide neural networks to kernel methods [12], the theory of kernel methods has seen a renewed interest as a tool to better understand deep neural networks [5]. Kernel methods, being analytically more tractable than neural networks, can yield significant insights about the behavior of deep networks. However several questions remain unanswered about the behavior of kernel methods themselves.

In this paper, we investigated the consistency of kernel methods in fixed dimensions. We showed that kernel interpolation, or kernel ridgeless regression, is inconsistent in fixed dimension even with adaptive bandwidth. This provides a generalization of the main result in [22], which considered the special case of the Laplace kernel, to a broad class of translation-invariant kernels including the Gaussian, Lapalce, and Cauchy kernels.

Our work suggests that infinitely-wide neural networks are inconsistent in fixed dimensions, as these networks are equivalent to kernel machines [12]. It is an interesting direction for future work if feature learning in finite-width networks [21] can enable consistency.

Further, while our result may be perceived as a negative result about kernel methods, it still leaves open the possibility of bounded inconsistency under interpolation, also called *tempered overfitting* in [15]. It remains unclear when interpolation may be an acceptable solution concept. In any case, consistency can be enabled using appropriate regularization.

The Role of Data Dimension. When the dimension of the inputs scales with the number of samples, kernel ridgeless regression can generalize [13]. Our results provide additional evidence that high dimensions can dissipate the error due to noise. In particular, under our assumptions on the kernel, for the expression of the noisy estimation error (Lemma D.3(c)), the constants decay exponentially with dimension. This dependence was also observed in [22] for the Laplace kernel. As an additional effect, for target functions with norm that is invariant to dimension (before scaling by $(2\pi)^{-d}$), the 0-estimator has approximation error that vanishes exponentially with dimension. Further, to counteract the error due to noise, the bandwidth should be much larger than the data resolution in each dimension, i.e., $M > N$. However, when the dimensions grow with the number of samples, say $d = \omega(\log n)$, the resolution $N = n^{1/d} = O(1)$ in each dimension is approximately constant, and therefore the bandwidth does not need to increase with d, n to satisfy $M > N$. As increasing the bandwidth in general will worsen the approximation error, the constancy of the bandwidth is a form of the blessing of dimensionality.

Acknowledgements. We are grateful for support of the NSF, the Simons Institute for the Theory of Computing, and the Simons Foundation for the Collaboration on the Theoretical Foundations of Deep Learning[2] through awards DMS-2031883 and #814639. We also acknowledge NSF support through IIS-1815697 and the TILOS institute (NSF CCF-2112665). We also thank Sam Buchanan, Jamie Simon, Jonathan Shi, and the anonymous reviewers for useful conversations, questions, and feedback on this work.

[2] https://deepfoundations.ai/

REFERENCES

[1] 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.

[2] Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019.

[3] Mikhail Belkin, Daniel Hsu, and Ji Xu. Two models of double descent for weak features. *SIAM Journal on Mathematics of Data Science*, 2(4):1167–1180, 2020.

[4] Mikhail Belkin, Daniel J Hsu, and Partha Mitra. Overfitting or perfect fitting? risk bounds for classification and regression rules that interpolate. *Advances in neural information processing systems*, 31, 2018.

[5] Mikhail Belkin, Siyuan Ma, and Soumik Mandal. To understand deep learning we need to understand kernel learning. In *International Conference on Machine Learning*, pages 541–549. PMLR, 2018.

[6] Mikhail Belkin, Alexander Rakhlin, and Alexandre B Tsybakov. Does data interpolation contradict statistical optimality? In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1611–1619. PMLR, 2019.

[7] Guorui Bian and James M Dickey. Properties of multivariate cauchy and poly-cauchy distributions with bayesian g-prior applications. Technical report, University of Minnesota, 1991.

[8] Luc Devroye, Laszlo Györfi, and Adam Krzyżak. The hilbert kernel regression estimate. *Journal of Multivariate Analysis*, 65(2):209–227, 1998.

[9] Konstantin Donhauser, Mingqi Wu, and Fanny Yang. How rotational invariance of common kernels prevents generalization in high dimensions. In *International Conference on Machine Learning*, pages 2804–2814. PMLR, 2021.

[10] László Györfi, Michael Kohler, Adam Krzyzak, Harro Walk, et al. *A distribution-free theory of nonparametric regression*, volume 1. Springer, 2002.

[11] Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *The Annals of Statistics*, 50(2):949–986, 2022.

[12] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31, 2018.

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

[14] Elliott H Lieb and Michael Loss. *Analysis*, volume 14. American Mathematical Soc., 2001.

[15] Neil Mallinar, James B Simon, Amirhesam Abedsoltan, Parthe Pandit, Mikhail Belkin, and Preetum Nakkiran. Benign, tempered, or catastrophic: A taxonomy of overfitting. *arXiv preprint arXiv:2207.06569*, 2022.

[16] Kanti V Mardia, Peter E Jupp, and KV Mardia. *Directional statistics*, volume 2. Wiley Online Library, 2000.

[17] Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Generalization error of random feature and kernel methods: hypercontractivity and kernel matrix concentration. *Applied and Computational Harmonic Analysis*, 59:3–84, 2022.

[18] Song Mei and Andrea Montanari. The generalization error of random features regression: Precise asymptotics and the double descent curve. *Communications on Pure and Applied Mathematics*, 75(4):667–766, 2022.

[19] 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.

[20] Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. In search of the real inductive bias: On the role of implicit regularization in deep learning. *arXiv preprint arXiv:1412.6614*, 2014.

[21] Adityanarayanan Radhakrishnan, Daniel Beaglehole, Parthe Pandit, and Mikhail Belkin. Feature learning in neural networks and kernel machines that recursively learn features. *arXiv preprint arXiv:2212.13881*, 2022.

[22] Alexander Rakhlin and Xiyu Zhai. Consistency of interpolation with laplace kernels is a high-dimensional phenomenon. In *Conference on Learning Theory*, pages 2595–2623. PMLR, 2019.

[23] Bernhard Schölkopf, Ralf Herbrich, and Alex J Smola. A generalized representer theorem. In *International conference on computational learning theory*, pages 416–426. Springer, 2001.

[24] Bernhard Scholkopf and Alexander J Smola. *Learning with kernels: support vector machines, regularization, optimization, and beyond.* MIT press, 2018.

[25] 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.

Appendices

Appendix A. Proof of main result ($d = 1$).

We prove the main results. The proof strategy is to obtain an $\Omega(1)$ lower bound on $\mathbb{E}_{\boldsymbol{\xi}} \mathsf{MSE}\left(\widehat{f}_N, f^*\right)$. Equation (5.2) expressed this quantity as a sum of N non-negative quantities. We show that at least $\Omega(N)$ of these quantities, $\mathcal{E}_i$, are $\Omega(1/N)$.

Proof of Theorem 4.4. When $M > N$, we show that the approximation error is large in the base RKHS $\mathcal{H}_0$. On the other hand, when $M \le N$ we show that the averaged noisy estimation error has a constant lower bound.

Case 1, $M \le N$. In this case we show that the noisy estimation error is bounded away from 0. Define, $\Delta_i := \|G_i\|_1 - |G[i]| = \sum_{m\neq 0} |G[mN+i]| \ge 0$. Assumption 2 says that for all but $o(N)$ terms corresponding to $i \in [N]$, we have,

$$\Delta_i = \sum_{m\neq 0} |G[mN+i]| \le C_1 |G[i]| \sum_{m\neq 0} \frac{1}{1+m^2} \le 4C_1 |G[i]| \ . \tag{A.1}$$

For such an i, we can lower bound the noisy estimation error term $\mathcal{E}_i^{\mathsf{noisy}}$ as,

$$\mathcal{E}_i^{\mathsf{noisy}} = \frac{\sigma^2}{N} \frac{\|G_i\|^2}{\|G_i\|_1^2} = \frac{\sigma^2}{N} \frac{\|G_i\|^2}{(|G[i]| + \Delta_i)^2} \ge \frac{\sigma^2}{N} \frac{|G[i]|^2}{2|G[i]|^2 + 2|\Delta_i|^2} \ge \frac{\sigma^2}{2N(1+4C_1)}$$

$$\mathcal{E}^{\mathsf{noisy}} = \sum_{i=1}^{N} \mathcal{E}_i^{\mathsf{noisy}} = \Omega(\sigma^2),$$

since there are $\Omega(N)$ such indices $i \in [N]$ for which Equation (A.1) holds.

Case 2, $M > N$. In this case we show the approximation error will be bounded from 0. Since $M > N$, by Assumption 3, there exists a fixed integer i^*, such that $|G[i^*]| \le C_3 |G[N+i^*]|$. Now let $f^*(x) = \sqrt{2}\cos(i^* x)$ be the (real-valued) function with Fourier coefficients $V[i^*] = V[-i^*] = \frac{1}{\sqrt{2}}$, and $V[k] = 0$ for $|k| \neq i^*$. Using this, we can lower bound the approximation error as,

$$\mathcal{E}^{\mathsf{apx}} \ge \mathcal{E}_{i^*}^{\mathsf{apx}} \ge V[i^*]^2 \left(1 - \frac{\frac{1}{2}|G[i^*]|^2}{\sum_{m\in\mathbb{Z}} |G[mN+i^*]|^2}\right) \ge \frac{1}{2}\left(1 - \frac{\frac{1}{2}|G[i^*]|^2}{|G[N+i^*]|^2 + |G[i^*]|^2}\right)$$

$$\ge \frac{1+2C_3^2}{2+2C_3^2} = \Omega(1) \ .$$

The fact that $G_0[i^*] > 0$ from the Assumption 3, also allows us to conclude that,

$$\|f^*\|_{\mathcal{H}_0}^2 = \frac{|V[i^*]|^2}{G_0[i^*]} + \frac{|V[-i^*]|^2}{G_0[-i^*]} < \infty.$$

Proof of Theorem 4.5. This follows from the proof of Theorem 4.4. We showed that for $M \leq N$ (Case 1 above), the noisy estimation error $\mathcal{E}^{\mathsf{noisy}}$ satisfies $\mathcal{E}^{\mathsf{noisy}} = \Omega(\sigma^2)$. Since $\mathcal{E}^{\mathsf{noisy}}$ does not involve the target function f^*, the statement of Theorem 4.5 follows. ■

Proof of Theorem 4.3. For $M \leq N$, we know from Case 1 in the proof of Theorem 4.4 that $\mathcal{E}^{\mathsf{noisy}} = \Omega(1)$. For $M > N$, we will show that $\mathcal{E}^{\mathsf{apx}} = \Omega(1)$ which sufficies to prove the claim.

Note we have $G[k] = G[-k]$. We start by the monotonic boundedness of G, we have

$$\frac{|G[i]|^2}{\|G_i\|^2} = \frac{|G[i]|^2}{|G[i]|^2 + \sum_{k\in\mathbb{Z}*} |G[i+kN]|^2} \leq \frac{1}{1+\upsilon}$$

for $|i| < N/2$, where $\upsilon = C^2(C'')^2/c^2 > 0$.

Now consider $f^* = \sum_{k\in\mathbb{Z}} V[k]\phi_k$, and suppose $\|f^*\|^2 = \sum_{k\in\mathbb{Z}} |V[k]|^2 = 1$. As f^* has square-integrable zeroth and first derivatives, its Fourier series coefficients have decay $|V[i]| \leq B/(1+i^2)$ for all i, for a sufficiently large constant B. This implies for all $N > 0$,

$$\sum_{|i|>N/2} |V[i]| < \frac{4B}{N}, \qquad \text{and} \tag{A.2}$$

$$\sum_{|i|<N/2} |V[i]|^2 \geq 1 - \frac{48B^2}{N^3}. \tag{A.3}$$

One can show by contradiction (to equation (A.3)) – for some $\varepsilon \in [0, \frac{1}{2}]$ there exists a subset $S_\varepsilon \subset \{i : |i| < N/2\}$ such that $|S_\varepsilon| = \Omega(N^{2\varepsilon})$ and $|V[i]| \geq \Omega(N^{-\varepsilon})$ for all $i \in S_\varepsilon$. For any such ε, consider the following set of inequalities for $i \in S_\varepsilon$,

$$\begin{aligned}
\mathcal{E}_i^{\mathsf{apx}} &= \|V_i\|^2 - \frac{\langle G_i, V_i\rangle^2}{\|G_i\|^2} \\
&\geq \|V_i\|^2 - \frac{|G[i]|^2}{\|G_i\|^2}\left(\sum_{k\in\mathbb{Z}} |V[i+kN]|\right)^2 \geq \|V_i\|^2 - \frac{1}{1+\upsilon}\left(\sum_{k\in\mathbb{Z}} |V[i+kN]|\right)^2 \\
&\geq |V[i]|^2 - \frac{1}{1+\upsilon}\left(\sum_{k\in\mathbb{Z}} |V[i+kN]|\right)^2 = |V[i]|^2 - \frac{1}{1+\upsilon}\left(|V[i]| + \sum_{k\in\mathbb{Z}*} |V[i+kN]|\right)^2 \\
&\geq |V[i]|^2 - \frac{1}{1+\upsilon}\left(|V[i]| + \frac{4B}{N}\right)^2 = |V[i]|^2\left(1 - \frac{1}{1+\upsilon}\left(1 + \frac{4B}{N|V[i]|}\right)^2\right) \\
&\geq \Omega(N^{-2\varepsilon})\left(1 - \frac{\left(1 + 4B\cdot O(N^{\varepsilon-1})\right)^2}{1+\upsilon}\right)
\end{aligned}$$

Since $\varepsilon - 1 < 0$ due to the range of ε, the term in the inner parenthesis always approaches 1 for large enough N. Hence we have,

$$\mathcal{E}^{\mathsf{apx}} = \sum_{i=1}^{N} \mathcal{E}_i^{\mathsf{apx}} \geq \sum_{\substack{i\in S_\varepsilon \\ i>0}}^{N} \mathcal{E}_i^{\mathsf{apx}} + \sum_{\substack{i\in S_\varepsilon \\ i<0}}^{N} \mathcal{E}_{N-i}^{\mathsf{apx}} \geq \Omega(N^{-2\varepsilon}) \cdot |S_\varepsilon| = \Omega(1). \tag{A.4}$$

This proves the claim. ■

Appendix B. Decomposition of MSE: Proof of Lemma 5.2.

Appendices B.1, B.2, and B.3 provide proofs for Lemma 5.2 (a), (b), and (c) respectively. Recall that P_X is the L^2 projection operator onto $\operatorname{span}(\{K(x_i,\cdot)\})$

B.1. Approximation error: Proof of Lemma 5.2(a).

The proof proceeds by applying the Pythagorean theorem to the triangle $\{0, f^*, P_X f^*\}$ in L^2_μ. The following lemma gives exact expressions for projection of the target function and its norm.

Lemma B.1 (Projection). *For* $f^* = \sum_{k\in\mathbb{Z}} V[k]\phi_k$

$$P_X f^* = \sum_{\ell=0}^{N-1} \frac{\langle V_\ell, G_\ell\rangle}{\|G_\ell\|^2} \sum_{m\in\mathbb{Z}} G[mN+\ell]\phi_{mN+\ell}, \qquad \text{and} \qquad \|P_X f^*\|^2 = \sum_{\ell=0}^{N-1} \frac{|\langle V_\ell, G_\ell\rangle|^2}{\|G_\ell\|^2}$$

We get,

$$\|f^* - P_X f^*\|^2 = \|f^*\|^2 - \|P_X f^*\|^2 = \|V\|^2 - \sum_{\ell=0}^{N-1} \frac{\langle V_\ell, G_\ell\rangle^2}{\|G_\ell\|^2}$$

Proof of Lemma B.1. Note that Lemma E.5 shows that $\left\{\frac{\psi_\ell}{\|\psi_\ell\|}\right\}_{\ell=0}^{N-1}$ is an orthonormal basis for $\operatorname{Span}\{K(x_\ell,\cdot)\}$. Consequently, we have

$$P_X f^* = \sum_{\ell=0}^{N-1} \left\langle f^*, \frac{\psi_\ell}{\|\psi_\ell\|}\right\rangle \frac{\psi_\ell}{\|\psi_\ell\|}, \qquad \text{and} \qquad \|P_X f^*\|^2 = \sum_{\ell=0}^{N-1} \left\langle f^*, \frac{\psi_\ell}{\|\psi_\ell\|}\right\rangle^2$$

We compute these projections below. For $\ell \in [N]$,

$$\begin{aligned}
\sqrt{\|G_\ell\|_1}\,\langle f^*, \psi_\ell\rangle &= \left\langle \sum_{k\in\mathbb{Z}} V[k]\phi_k, \sum_{m\in\mathbb{Z}} G[mN+\ell]\phi_{mN+\ell}\right\rangle \\
&= \sum_{m,k\in\mathbb{Z}} G[mN+\ell]V[k]\mathbb{1}_{\{k=mN+\ell\}} \\
&= \sum_{m\in\mathbb{Z}} G[mN+\ell]V[mN+\ell] = \langle V_\ell, G_\ell\rangle
\end{aligned}$$

Thus, we get that,

$$\left\langle f^*, \frac{\psi_\ell}{\|\psi_\ell\|}\right\rangle \frac{\psi_\ell}{\|\psi_\ell\|} = \frac{\langle V_\ell, G_\ell\rangle}{\|G_\ell\|^2} \sum_{m\in\mathbb{Z}} G[mN+\ell]\phi_{mN+\ell}$$

The claims follow immediately. ■

B.2. Noise-free estimation error: Proof of Lemma 5.2(b).

Let E be the fourier series of $\left\langle \mathbf{K}^{-1} R_N\{f^* - P_X f^*\}, K(X_N,\cdot)\right\rangle$. From Lemma E.3 we have,

$$E[k] = \sqrt{N} R_N \{f^* - P_X f^*\}^\top \mathbf{K}^{-1} \boldsymbol{u}_{k \bmod N} \cdot G[k]$$

By Parseval's theorem (Proposition 2.4), we conclude,

$$\frac{1}{2\pi}\int_{-\pi}^{\pi}\left(\left\langle \boldsymbol{K}^{-1}R_N\left\{f^*-P_Xf^*\right\},K(X_N,t)\right\rangle_N\right)^2\,\mathrm{d}t=\sum_{k\in\mathbb{Z}}|E[k]|^2$$

First, we show that

$$R_N\left\{f^*-P_Xf^*\right\}^\top\boldsymbol{K}^{-1}\boldsymbol{u}_\ell=\left(\frac{\langle V_i,\mathbf{1}\rangle}{\langle G_i,\mathbf{1}\rangle}-\frac{\langle G_\ell,V_\ell\rangle}{\|G_\ell\|^2}\right)$$

From Proposition 3.1 we have $\mathbf{K}^{-1}\boldsymbol{u}_\ell=\boldsymbol{u}_\ell\cdot\frac{1}{N\|G_\ell\|_1}$. Thus by Lemma B.1, we can write P_Xf^* on the data as

$$P_Xf^*(x_i)=\sum_{\ell=0}^{N-1}\frac{\langle G_\ell,V_\ell\rangle}{\|G_\ell\|^2}\sum_{m\in\mathbb{Z}}G[mN+\ell]\phi_{mN+\ell}(x_i)=\sum_{\ell=0}^{N-1}\frac{\langle G_\ell,V_\ell\rangle}{\|G_\ell\|^2}\|G_\ell\|_1\,\overline{u_{\ell i}}$$

$$f^*(x_i)=\sum_{k\in\mathbb{Z}}V[k]\phi_k(x_i)=\sum_{\ell=0}^{N-1}\langle V_i,\mathbf{1}\rangle\,\overline{u_{\ell i}}$$

We thus have

$$\begin{aligned}R_N\left\{f^*-P_Xf^*\right\}^\top\mathbf{K}^{-1}\boldsymbol{u}_\ell&=\sum_{i,\ell'=0}^{N-1}\left(\langle V_{\ell'},\mathbf{1}\rangle-\frac{\langle G_{\ell'},V_{\ell'}\rangle}{\|G_{\ell'}\|^2}\|G_{\ell'}\|_1\right)\frac{\overline{u_{\ell' i}}u_{\ell i}}{N\,\|G_\ell\|_1}\\&=\frac{1}{N}\left(\frac{\langle V_\ell,\mathbf{1}\rangle}{\langle G_\ell,\mathbf{1}\rangle}-\frac{\langle V_\ell,G_\ell\rangle}{\|G_\ell\|^2}\right)\end{aligned}$$

This gives,

$$\sum_{k\in\mathbb{Z}}|E[k]|^2=\frac{1}{N}\sum_{\ell=0}^{N-1}\left(\frac{\langle V_\ell,\mathbf{1}\rangle}{\langle G_\ell,\mathbf{1}\rangle}-\frac{\langle V_\ell,G_\ell\rangle}{\|G_\ell\|^2}\right)^2\|G_\ell\|^2\,.$$

B.3. Noisy estimation error: Proof of Lemma 5.2(c). We derive this by an application of Parseval's theorem. Define the Fourier series,

$$\left\langle\mathbf{K}^{-1}\boldsymbol{\xi},K(X_N,t)\right\rangle_N=\sum_{k\in\mathbb{Z}}\widetilde{E}[k]e^{jkt}$$

By Proposition 2.4 (Parseval's theorem), we have,

$$\begin{aligned}&\mathbb{E}_{\boldsymbol{\xi}}\left[\frac{1}{2\pi}\int_{-\pi}^{\pi}|\left\langle\mathbf{K}^{-1}\boldsymbol{\xi},K(X_N,t)\right\rangle_N|^2\,\mathrm{d}t\right]=\sum_{k\in\mathbb{Z}}\mathbb{E}_{\boldsymbol{\xi}}\,|\widetilde{E}[k]|^2=\sum_{i=0}^{N-1}\sum_{m\in\mathbb{Z}}\mathbb{E}_{\boldsymbol{\xi}}\left|\widetilde{E}[mN+i]\right|^2\\&\overset{\text{(a)}}{=}\sum_{i=0}^{N-1}\sum_{m\in\mathbb{Z}}|G[mN+i]|^2\,\mathbb{E}_{\boldsymbol{\xi}}\left|\boldsymbol{\xi}^\top\mathbf{K}^{-1}\boldsymbol{u}_i\right|^2\cdot N\overset{\text{(b)}}{=}\sigma^2\sum_{i=0}^{N-1}\|G_i\|^2\left(\boldsymbol{u}_i^{\mathsf{H}}\mathbf{K}^{-2}\boldsymbol{u}_i\right)\cdot N\\&\overset{\text{(c)}}{=}\sigma^2\sum_{i=0}^{N-1}\|G_i\|^2\,\frac{1}{N^2\,\|G_i\|_1^2}N=\sigma^2\sum_{i=0}^{N-1}\frac{\|G_i\|^2}{N\,\|G_i\|_1^2}\end{aligned}$$

where we have used Lemma E.3 in (a), and Lemma E.2 in (b), and Proposition 3.1 in (c).

Appendix C. Main Results for $d > 1$.

We can perform a similar analysis for $d > 1$. To generalize the main results, we also generalize Assumptions 1-3 for the kernel to Assumptions 4-6 for dimensions greater than one, as well as the monotonicity condition (Condition 2). Under these assumptions, we show the main results hold.

Theorem C.1 (Inconsistency for all Functions (when G is monotonically bounded)). *Consider a fixed non-zero regression function f^* (i) in the Sobolev space of order $\frac{d}{2} + 1$ (i.e. whose derivatives of orders $\alpha \in \mathbb{Z}^d_{\geq 0}$ for $\|\alpha\|_1 \leq \frac{d}{2} + 1$ are square-integrable) and (ii) that can be expressed as a convergent Fourier series. Then, interpolation with a real-valued translation-invariant kernel satisfying Condition 2 is inconsistent for f^*, for any bandwidth, even if chosen adaptively.*

See [14] for a definition of an order α derivative.

Theorem C.2 (Inconsistency for all Bandwidths). *For any translation-invariant kernel satisfying Assumptions 4-6, there exists a function with constant $\mathcal{H}_0$-norm for which kernel interpolation will be inconsistent for any bandwidth, even adaptive to the data set.*

Theorem C.3 (Inconsistency for all Functions). *For any translation-invariant kernel satisfying Assumptions 4-5, with a bandwidth $M \leq N$, kernel interpolation will be inconsistent for all target functions that can be expressed as convergent Fourier series. In particular, kernel interpolation with any fixed bandwidth will be inconsistent for all such functions.*

Further, these results hold for the Gaussian, Laplace, and Cauchy kernels.

Proposition C.4. *The Gaussian $G[k] = \exp(-\frac{\|\boldsymbol{k}\|^2}{M^2})$, Laplace $G[\boldsymbol{k}] = \left(1 + \frac{\|\boldsymbol{k}\|^2}{M^2}\right)^{-\frac{d+1}{2}}$, and Cauchy $G[\boldsymbol{k}] = \exp(-\frac{\|\boldsymbol{k}\|}{M})$ kernels (wrapped on the unit circle) satisfy Condition 2.*

Supplementary materials:

On the Inconsistency of Kernel Ridgeless Regression in Fixed Dimensions

Appendix D. Extending proofs to higher dimensions. Proofs missing from this section are provided in the supplementary materials.

Notation. In this section $d \geq 1$ and $n = N^d$. By $[N]^d$ we denote the d-fold Cartesian product of $[N] := \{0, 1, \ldots, N-1\}$. For vectors $\boldsymbol{p}, \boldsymbol{q}$, we write $\boldsymbol{p} \leq \boldsymbol{q}$ to indicate a coordinate-wise inequality, i.e., for all coordinates i, we have $p_i \leq q_i$. Similarly, $\boldsymbol{p} \nleq \boldsymbol{q}$ indicates $\boldsymbol{p} \leq \boldsymbol{q}$ is violated, i.e., there exists a coordinate i for which $p_i > q_i$. We similarly define $\boldsymbol{p} \geq \boldsymbol{q}$ and $\boldsymbol{p} \ngeq \boldsymbol{q}$. We also denote $\mathbf{0}$ and $\mathbf{1}$ to be the vectors of all 0's and all 1's respectively in a dimension compatible with the expression. For a scalar C, the expression $\boldsymbol{p} \leq C$ means $\boldsymbol{p} \leq C \cdot \mathbf{1}$, and similarly $\boldsymbol{p} \geq C, \boldsymbol{p} \nleq C, \boldsymbol{p} \ngeq C$.

We consider sequences indexed by $\mathbb{Z}^d$, and for such sequences we extend the definition of *$N-$ hop subsequences* from Equation (2.1) in the following manner. For a fixed $N \in \mathbb{N}$, and a sequence $G \in \ell^1(\mathbb{Z}^d)$, and for $\boldsymbol{\ell} \in [N]^d$, let

$$G_{\boldsymbol{\ell}} \in \ell^1(\mathbb{Z}^d) \qquad G_{\boldsymbol{\ell}}[\boldsymbol{m}] = G[\boldsymbol{m}N + \boldsymbol{\ell}], \qquad \forall\, \boldsymbol{m} \in \mathbb{Z}^d$$

be the *N-hop subsequence* with entries given as above. For $\boldsymbol{k} \in \mathbb{Z}^d$, and $\boldsymbol{x} \in \mathbb{R}^d$

(D.1)

$$\boldsymbol{k} \bmod N := (k_1 \pmod N),\ k_2 \pmod N),\ \ldots,\ k_d \pmod N)) \in [N]^d$$

(D.2)

$$\boldsymbol{x} \bmod [-\pi,\pi) := (x_1 \ (\mathrm{mod}\ [-\pi,\pi)),\ x_2 \ (\mathrm{mod}\ [-\pi,\pi)),\ \ldots,\ x_d \ (\mathrm{mod}\ [-\pi,\pi))) \in [-\pi,\pi)^d$$

where we remind the reader of notation from Equation (3.2). We denote by $[-\pi,\pi)^d$ the Cartesian product of d unit circles $[-\pi,\pi)$ along each dimension. We refer to this as the unit torus.

Definition D.1 (Fourier basis). *For $k \in \mathbb{Z}^d$ and $\boldsymbol{x} \in [-\pi,\pi)^d$, define $\phi_{\boldsymbol{k}}(\boldsymbol{x}) := \exp\left(j \langle \boldsymbol{k}, \boldsymbol{x} \rangle\right) = \prod_{i=1}^d \exp(jk_i x_i)$. This basis satisfies $\langle \phi_{\boldsymbol{k}}, \phi_{\boldsymbol{\ell}} \rangle = \frac{1}{(2\pi)^d} \int_{[-\pi,\pi)^d} \exp(j \langle \boldsymbol{k} - \boldsymbol{\ell}, \boldsymbol{x} \rangle) \, \mathrm{d}\boldsymbol{x} = \mathbb{1}_{\{\boldsymbol{k}=\boldsymbol{\ell}\}}$.*

A target function defined on the unit torus admits a Fourier series,

$$f^* = \sum_{\boldsymbol{k} \in \mathbb{Z}^d} V[\boldsymbol{k}] \phi_{\boldsymbol{k}} \qquad V[\boldsymbol{k}] = \langle f^*, \phi_{\boldsymbol{k}} \rangle\,. \tag{D.3}$$

Definition D.2 (DFT Matrix $d > 1$). *The normalized DFT matrix in $d > 1$ is*

$$\boldsymbol{U}_d = \begin{bmatrix} \boldsymbol{u}_{\mathbf{0}} & \cdots & \boldsymbol{u}_{(N-1)\mathbf{1}} \end{bmatrix} \in \mathbb{C}^{N^d \times N^d}, \qquad \boldsymbol{u}_{\boldsymbol{\ell},\boldsymbol{p}} := N^{-d/2} \exp\left(-j \frac{2\pi}{N} \langle \boldsymbol{\ell}, \boldsymbol{p} \rangle\right), \quad \boldsymbol{\ell}, \boldsymbol{p} \in [N]^d$$

Data distribution. For $d > 1$, the continuous distribution is $\mu = \mathrm{Uniform}([-\pi,\pi)^d)$ and the discrete distribution over $\boldsymbol{x} \in [-\pi,\pi)^d$, with $n = N^d$ samples, to be

$$\mu_n(\boldsymbol{x}) := \frac{1}{N^d} \sum_{\ell \in [N]^d} \delta(\boldsymbol{x} - \boldsymbol{x}_{\boldsymbol{\ell}}), \qquad (\boldsymbol{x}_{\boldsymbol{\ell}})_i = \frac{2\pi}{N} \ell_i - \pi, \qquad \forall\, \boldsymbol{\ell} \in [N]^d, \text{ and } \forall\, i \in [N].$$

The $n = N^d$ samples $\{\boldsymbol{x_\ell}\}$ are indexed by elements of $[N]^d$, where $\boldsymbol{x_\ell} \in [-\pi, \pi)^d$ and has coordinates given by the expression above. Note again that μ_n weakly converges to μ. In the rest of this section we use

$$\langle \boldsymbol{\alpha}, K(X_n, \cdot)\rangle_n := \sum_{\boldsymbol{p}\in[N]^d} \alpha_{\boldsymbol{p}} K(\boldsymbol{x_p}, \cdot)$$

to keep the notation simple.

Translation-invariant kernels. As in $d = 1$, we can define translation-invariant kernels with the following property,

$$K(\boldsymbol{x}, \boldsymbol{x}') = g\left(M(\boldsymbol{x} - \boldsymbol{x}' \bmod [-\pi, \pi))\right), \qquad \boldsymbol{x}, \boldsymbol{x}' \in [-\pi, \pi)^d$$

for some even function $g : \mathbb{R}^d \to \mathbb{R}$ (see Definition (D.1)).

Define $G_0, G : \mathbb{Z}^d \to \mathbb{C}$ as the Fourier series, i.e.

$$g(M(\boldsymbol{\theta} \bmod [-\pi, \pi))) = \sum_{\boldsymbol{k}\in\mathbb{Z}^d} G[\boldsymbol{k}] \exp(j \langle \boldsymbol{k}, \boldsymbol{\theta}\rangle) \tag{D.4a}$$

$$G[\boldsymbol{k}] = \frac{1}{(2\pi)^d} \int_{[-\pi,\pi)^d} g(M\boldsymbol{\theta}) \exp(-j \langle \boldsymbol{k}, \boldsymbol{\theta}\rangle) \,\mathrm{d}\boldsymbol{\theta} \tag{D.4b}$$

Note that while usually the bandwidth scales the input (as detailed here), our analysis also holds for different mechanisms.

As in $d = 1$, for positive definite kernels, g is an even function whereby we have,

$$G[\boldsymbol{k}] = G[-\boldsymbol{k}] \geq 0 \quad \forall \boldsymbol{k} \in \mathbb{Z}^d$$

D.1. MSE Decomposition.

We start with a result analogous to Lemma 5.1, when $d > 1$.

Lemma D.3 (Decomposition of MSE for $d > 1$). *For a target function $f^* = \sum_{\boldsymbol{k}\in\mathbb{Z}^d} V[\boldsymbol{k}]\phi_{\boldsymbol{k}}$,*

(a) Approximation error:

$$\mathcal{E}^{\mathsf{apx}} := \|f^* - P_X f^*\|^2 = \sum_{\boldsymbol{p}\in[N]^d} \|V_{\boldsymbol{p}}\|^2 - \frac{\langle G_{\boldsymbol{p}}, V_{\boldsymbol{p}}\rangle^2}{\|G_{\boldsymbol{p}}\|^2} = \sum_{\boldsymbol{p}\in[N]^d} \mathcal{E}^{\mathsf{apx}}_{\boldsymbol{p}}$$

(b) Noise-free estimation error:

$$\mathcal{E}^{\mathsf{free}} := \left\|\left\langle \mathbf{K}^{-1} R_n \left\{f^* - P_X f^*\right\}, K(X_n, \cdot)\right\rangle_n\right\|^2 = \sum_{\boldsymbol{p}\in[N]^d} \frac{\|G_{\boldsymbol{p}}\|^2}{N^d} \left(\frac{\langle V_{\boldsymbol{p}}, \mathbf{1}\rangle}{\langle G_{\boldsymbol{p}}, \mathbf{1}\rangle} - \frac{\langle G_{\boldsymbol{p}}, V_{\boldsymbol{p}}\rangle}{\|G_{\boldsymbol{p}}\|^2}\right)^2$$

$$= \sum_{\boldsymbol{p}\in[N]^d} \mathcal{E}^{\mathsf{free}}_{\boldsymbol{p}}$$

(c) Averaged noisy estimation error:

$$\mathcal{E}^{\mathsf{noisy}} := \mathbb{E}_{\boldsymbol{\xi}} \left\|\left\langle \mathbf{K}^{-1}\boldsymbol{\xi}, K(X_n, \cdot)\right\rangle_n\right\|^2 = \sum_{\boldsymbol{p}\in[N]^d} \frac{\sigma^2}{N^d} \left(\frac{\|G_{\boldsymbol{p}}\|}{\|G_{\boldsymbol{p}}\|_1}\right)^2 = \sum_{\boldsymbol{p}\in[N]^d} \mathcal{E}^{\mathsf{noisy}}_{\boldsymbol{p}}$$

Together, this yields that the MSE for this function is,

$$\mathbb{E}_{\boldsymbol{\xi}} \operatorname{MSE}\left(\widehat{f}_N, f^*\right) = \sum_{\boldsymbol{p}\in[N]^d} \mathcal{E}_{\boldsymbol{p}} = \sum_{\boldsymbol{p}\in[N]^d} \mathcal{E}_{\boldsymbol{p}}^{\mathsf{apx}} + \mathcal{E}_{\boldsymbol{p}}^{\mathsf{free}} + \mathcal{E}_{\boldsymbol{p}}^{\mathsf{noisy}} \tag{D.5}$$

$$\mathcal{E}_{\boldsymbol{p}} := \|V_{\boldsymbol{p}}\|^2 - \frac{\langle G_{\boldsymbol{p}}, V_{\boldsymbol{p}}\rangle^2}{\|G_{\boldsymbol{p}}\|^2} + \frac{\|G_{\boldsymbol{p}}\|^2}{N^d}\left(\frac{\langle V_{\boldsymbol{p}}, \mathbf{1}\rangle}{\langle G_{\boldsymbol{p}}, \mathbf{1}\rangle} - \frac{\langle G_{\boldsymbol{p}}, V_{\boldsymbol{p}}\rangle}{\|G_{\boldsymbol{p}}\|^2}\right)^2 + \frac{\sigma^2}{N^d}\left(\frac{\|G_{\boldsymbol{p}}\|}{\|G_{\boldsymbol{p}}\|_1}\right)^2 \tag{D.6}$$

The derivation for $d > 1$ is similar to the $d = 1$ case. Appendix F.1, Appendix F.2, and Appendix F.3 in the supplementary materials provide proofs for Lemma D.3 Item (a), Item (b), and Item (c) respectively.

D.2. Spectral Assumptions. We assume spectral conditions for kernels in $d > 1$ that are analogous to those in $d = 1$.

Assumption 4 (Integrability). *We assume the kernel is integrable. In particular, the integral $\int_{[-\pi,\pi)^d} g(M\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$ exists and is finite for all $0 < M < \infty$.*

Assumption 5 (Spectral Tail). *For all $\boldsymbol{k} \in \mathbb{Z}^d_{\geq 0}$, there exists a dimension-dependent constant $C_{1,d} > 0$ such that,*

$$|G[\boldsymbol{k}M' + \boldsymbol{p}]| \leq C_{1,d}|G[\boldsymbol{p}]|\left(1 + \|\boldsymbol{k}\|^2\right)^{-\frac{d+1}{2}} \tag{D.7}$$

holds for all $M' \geq M$ and for all $0 \leq \boldsymbol{p} \leq M'$, except $o_{M'}((M')^d)$ many.

Assumption 6 (Spectral Head). *There exist dimension-dependent constants $C_{2,d}, C_{3,d} \in \mathbb{R}_+$, $\boldsymbol{p}^* \in \mathbb{Z}_{\geq 0}$, and $\mathbf{0} \leq \boldsymbol{m}^* \leq \mathbf{1}$ with $\boldsymbol{m}^* \neq 0$, such that for $M \geq C_{2,d}$, we have that for all $M' \leq M$, $|G[\boldsymbol{p}^*]| \leq C_{3,d}|G[\boldsymbol{p}^* + M'\boldsymbol{m}^*]|$ and $|G_0[\boldsymbol{p}^*]| > 0$.*

We give a condition that implies all three of these assumptions:

Condition 2 (Monotonic Boundedness ($d > 1$)). *There exist constants (that are independent of the bandwidth M) $c_d, C_d, C'_d, C''_d > 0$, a constant $c'_d(M)$ (that may depend on M) and a monotonically decreasing function $f(\|\boldsymbol{x}\|)$ with (i) $0 < \frac{f(\|\boldsymbol{x}+\boldsymbol{k}\|)}{f(\|\boldsymbol{x}\|)} \leq C''_d\left(1 + \|\boldsymbol{k}\|^2\right)^{-\frac{d+1}{2}}$ for all $x \in \mathbb{R}^d_{\geq 0}$,$\boldsymbol{k} \in \mathbb{Z}^d_{\geq 0}$, and (ii) $\frac{f(\|\boldsymbol{e}_i+\boldsymbol{x}\|)}{f(\|\boldsymbol{x}\|)} \geq C'_d$ for $\mathbf{0} \leq \boldsymbol{x} \leq \mathbf{1}$ and standard basis vectors $\boldsymbol{e}_i$ for $i \in [d]$, such that $c_d f(\|\boldsymbol{k}\|/M) \leq G[\boldsymbol{k}]/c'_d(M) \leq C_d f(\|\boldsymbol{k}\|/M)$ for all $\boldsymbol{k} \in \mathbb{Z}^d_{\geq 0}$.*

With the assumptions defined, we can prove the main results for $d > 1$.

D.3. Proof of Theorem C.2. When $M \leq N$ we show that the averaged noisy estimation error is large. On the other hand, when $M > N$, we show that the approximation error is large for a cosine function in the base RKHS $\mathcal{H}_0$.

Case 1, $M > N$. In this case we show the approximation error is bounded away from 0. Since $M > N$, by Assumption 6, there exists a vector $\boldsymbol{p}^*$ of constant integers, and an $\mathbf{0} \leq \boldsymbol{m}^* \leq \mathbf{1}$ with $\boldsymbol{m}^* \neq 0$, such that $|G[\boldsymbol{p}^*]| \leq C_{3,d}\,|G[\boldsymbol{m}^*N + \boldsymbol{p}^*]|$. Now let f^* be the (real-valued) function with Fourier coefficients $V[\boldsymbol{p}^*] = V[-\boldsymbol{p}^*] = \sqrt{\frac{(2\pi)^{d-1}}{2}}$, and $V[k] = 0$ for

$|k| \neq i^*$. Using this, we can lower bound the approximation error as,

$$\begin{aligned}\mathcal{E}^{\mathsf{apx}}_{\boldsymbol{p}^*} \geq \frac{2V[\boldsymbol{p}^*]^2}{(2\pi)^d}\left(1 - \frac{|G[\boldsymbol{p}^*]|^2}{\sum_{\boldsymbol{m}\in\mathbb{Z}^d}|G[\boldsymbol{m}N+\boldsymbol{p}^*]|^2}\right) &\geq \frac{1}{2\pi}\left(1 - \frac{|G[\boldsymbol{p}^*]|^2}{|G[\boldsymbol{m}^*N+\boldsymbol{p}^*]|^2 + |G[\boldsymbol{p}^*]|^2}\right)\\ &\geq \frac{1}{2\pi(1+C_{3,d})}\end{aligned}$$

Case 2, $M \leq N$. In this case we show that the noisy estimation error is bounded away from 0. Define, $\Delta_{\boldsymbol{p}} := \|G_{\boldsymbol{p}}\|_1 - |G[\boldsymbol{p}]| = \sum_{\boldsymbol{m}\neq 0}|G[\boldsymbol{m}N+\boldsymbol{p}]| \geq 0$. Assumption 5 says that for all but $o(N^d)$ terms $\boldsymbol{p}\in[N]^d$, we have,

$$\Delta_{\boldsymbol{p}} = \sum_{\boldsymbol{m}\neq 0}|G[\boldsymbol{m}N+\boldsymbol{p}]| \leq C_{1,d}\,|G[\boldsymbol{p}]| \sum_{\boldsymbol{m}\neq 0}\left(1+\|\boldsymbol{m}\|^2\right)^{-\frac{d+1}{2}} \leq C_{1,d}\,|G[\boldsymbol{p}]| \tag{D.8}$$

For such an $\boldsymbol{p}$, we can lower bound the noisy estimation error term $\mathcal{E}^{\mathsf{noisy}}_{\boldsymbol{p}}$ as,

$$\begin{aligned}&\mathcal{E}^{\mathsf{noisy}}_{\boldsymbol{p}} = \frac{\sigma^2}{N^d}\frac{\|G_{\boldsymbol{p}}\|^2}{\|G_{\boldsymbol{p}}\|_1^2} = \frac{\sigma^2}{N^d}\frac{\|G_{\boldsymbol{p}}\|^2}{(|G[\boldsymbol{p}]|+\Delta_{\boldsymbol{p}})^2} \geq \frac{\sigma^2}{N^d}\frac{|G[\boldsymbol{p}]|^2}{2\,|G[\boldsymbol{p}]|^2 + 2\,|\Delta_{\boldsymbol{p}}|^2} \geq \frac{\sigma^2}{2N^d(1+C_{1,d})}\\ &\mathcal{E}^{\mathsf{noisy}} = \Omega(\sigma^2),\end{aligned}$$

since there are $\Omega(N^d)$ such $\boldsymbol{p}\in[N]^d$ for which equation (D.8) holds.

Proof of Theorem C.3. As in the $d = 1$ case, Theorem C.3 follows from the proof of Theorem C.2. We showed that for $M \leq N$ (Case 2 above), the noisy estimation error $\mathcal{E}^{\mathsf{noisy}}$ satisfies $\mathcal{E}^{\mathsf{noisy}} = \Omega(\sigma^2)$. Since $\mathcal{E}^{\mathsf{noisy}}$ is independent of the target function f^*, the statement of Theorem 4.5 follows. ■

Proof of Theorem C.1. We showed that for $M \leq N$ (Case 2 above), the noisy estimation error $\mathcal{E}^{\mathsf{noisy}}$ satisfies $\mathcal{E}^{\mathsf{noisy}} = \Omega(\sigma^2)$ for all functions. We start by the monotonic boundedness of G, we have

$$\frac{|G[\boldsymbol{p}]|^2}{\|G_{\boldsymbol{p}}\|^2} = \frac{|G[\boldsymbol{p}]|^2}{|G[\boldsymbol{p}]|^2 + \sum_{k\in(\mathbb{Z}*)^d}|G[\boldsymbol{p}+\boldsymbol{k}N]|^2} \leq \frac{1}{1+\upsilon_d}$$

for $\|\boldsymbol{p}\|_\infty < N/2$, where $\upsilon_d = C_d^2(C_d'')^2/c_d^2 > 0$.

Now consider $f^* = \sum_{\boldsymbol{k}\in\mathbb{Z}^d} V[\boldsymbol{k}]\phi_{\boldsymbol{k}}$, and suppose $\|f^*\|^2 = \sum_{\boldsymbol{k}\in\mathbb{Z}^d}|V[\boldsymbol{k}]|^2 = 1$. By the smoothness condition on the target function f^*, it has Fourier series coefficients that decay as $|V[\boldsymbol{k}]| \leq B_d(1+\|\boldsymbol{k}\|^2)^{-\frac{d+1}{2}}$ for all $\boldsymbol{k}\in\mathbb{Z}^d$, for a sufficiently large constant B_d. Using this we get the inequalities,

$$\begin{aligned}&\sum_{\|\boldsymbol{p}\|>N/2}|V[\boldsymbol{p}]| \leq \widetilde{C}_d\int_{N/2}^{\infty}\frac{r^{d-1}}{(1+r^2)^{(d+1)/2}}\,\mathrm{d}r = O(\frac{1}{N})\\ &\sum_{\|\boldsymbol{p}\|<N/2}|V[\boldsymbol{p}]|^2 \geq 1 - \widehat{C}_d\int_{N/2}^{\infty}\frac{r^{d-1}}{(1+r^2)^{d+1}}\,\mathrm{d}r = 1 - O(N^{-d-2})\end{aligned}$$ ■

where we have used Wolfram Alpha to obtain order bounds. The integral involves the special function ${}_2F_1$, also known as, the hypergeometric function. The rest of the proof proceeds in a similar manner as the $d = 1$ case. The only difference is that the range of ε is now $\varepsilon \in [0, \frac{d}{2}]$.

D.4. Special cases of kernels for $d > 1$.

Proof of Proposition C.4.

Gaussian kernel. For the wrapped Gaussian kernel, we have

$$G[\boldsymbol{k}] = \exp\left(-\|\boldsymbol{k}\|^2/M^2\right)$$

Therefore, the monotonicity is satisfied with $f(\|\boldsymbol{x}\|) = \exp(-\|\boldsymbol{x}\|^2)$.

Laplace kernel. For the wrapped Laplace kernel, we have

$$G[\boldsymbol{k}] = \left(1 + \frac{\|\boldsymbol{k}\|^2}{M^2}\right)^{-\frac{d+1}{2}}$$

(See [22] for a derivation). Therefore, the monotonicity is satisfied with $f(\|\boldsymbol{x}\|) = \left(1 + \|\boldsymbol{x}\|^2\right)^{-\frac{d+1}{2}}$.

Cauchy kernel. For the wrapped Cauchy kernel, we have (from [7]),

$$G[\boldsymbol{k}] = \exp(-\|\boldsymbol{k}\|/M)$$

Therefore, the monotonicity is satisfied with $f(\|\boldsymbol{x}\|) = \exp(-\|\boldsymbol{x}\|)$. ■

Lemma D.4. *For $\boldsymbol{\beta} = (\beta_{\boldsymbol{p}}) \in \mathbb{C}^{N^d}$, the Fourier series of the function $\langle \boldsymbol{\beta}, K(X_n, \cdot)\rangle_n$ is,*

$$B[\boldsymbol{k}] = N^{d/2} \boldsymbol{\beta}^\top \boldsymbol{u}_{\boldsymbol{k} \text{ mod } N} \cdot G[\boldsymbol{k}], \quad \boldsymbol{k} \in \mathbb{Z}^d$$

The proof for this lemma is provided in Appendix F.4 in the supplementary materials.

Appendix E. Miscellaneous results.

The proof of the following for $d > 1$ is provided in Appendix D.4.

Proof of Proposition 4.2 (d=1). We state what it means for a kernel to be wrapped on the unit circle [16]. This means that the wrapped kernel evaluation at $t \in [-\pi, \pi)$ has contributions from the Euclidean kernel at $t + 2\pi k$ for all $k \in \mathbb{Z}$. In particular,

Definition E.1 (Wrapper kernel). *A kernel $\widetilde{g} : \mathbb{R} \to \mathbb{R}^+$ can be wrapped to define a wrapped kernel $g : [-\pi, \pi) \to \mathbb{R}^+$ given by:*

$$g(t) = \sum_{k \in \mathbb{Z}} \widetilde{g}(t + 2\pi k)$$

It is a fact that the Fourier series coefficients of the wrapped on the unit circle are equal to the Fourier transform at integer values on the real line, i.e.,

$$G[k] = \int \widetilde{g}(t) \exp(-jkt) \, \mathrm{d}t \qquad \forall \ k \in \mathbb{Z}.$$ ■

Using this, we derive the proposition for the given kernels [16].

Gaussian kernel. The wrapped Gaussian kernel satisfies $G[k] = e^{-k^2/4M^2}$. Therefore, $f(x) = e^{-x^2}$ clearly satisfies the monotonicity condition.

Laplace kernel. For the wrapped Laplace kernel, $G[k] = \dfrac{1}{M^2 + k^2}$. Thus, the Laplace kernel satisfies Condition 1 with $f(x) = \frac{1}{1+x^2}$.

Cauchy kernel. For the Cauchy kernel, $G[k] = \exp(-|k|/M)$. Therefore, Condition 1 is satisfied with $f(x) = \exp(-|x|)$.

Proof of Proposition 4.1. We first show boundedness. As $f(0) < \infty$, we have $f(k) < \frac{f(0)}{1+k^2}$ for all $k \in \mathbb{R}^+$. Therefore, $\sum_{k\in\mathbb{Z}} |G[k]| \le C \int_k f(k) < \infty$. Then, $K(x, x') \le \sum_k |G[k]| < \infty$.

For the tail assumption, we have for $M' > M$,

$$\frac{G[M'k + i]}{G[i]} \le \frac{C''C}{c} \frac{f((i+k)/M)}{f(i/M)} \frac{C''C}{c} \frac{1}{1+k^2}\ .$$

For the head assumption, we have for $M' \le M$,

$$0 < \frac{G[i]}{G[i + M']} \le \frac{c}{C''C} \frac{f(i/M)}{f((i+M')/M)} \le \frac{c}{C''CC'}\ .$$

These prove the proposition. ■

E.1. Intermediate Lemmas.

Lemma E.2. *Let $\boldsymbol{\xi}$ be a random vector with $\mathbb{E}\,\boldsymbol{\xi}\boldsymbol{\xi}^{\mathsf{H}} = \sigma^2\boldsymbol{I}$, then for $\boldsymbol{u} \in \mathbb{C}^N$, $\mathbb{E}_{\boldsymbol{\xi}}\,|\boldsymbol{\xi}^{\mathsf{H}}\mathbf{K}^{-1}\boldsymbol{u}|^2 = \sigma^2 \cdot \boldsymbol{u}^{\mathsf{H}}\mathbf{K}^{-2}\boldsymbol{u}$*

Proof. $\mathbb{E}_{\boldsymbol{\xi}}\,|\boldsymbol{\xi}^{\mathsf{H}}\mathbf{K}^{-1}\boldsymbol{u}|^2 = \mathbb{E}_{\boldsymbol{\xi}}\,\boldsymbol{u}^{\mathsf{H}}\mathbf{K}^{-1}\boldsymbol{\xi}\boldsymbol{\xi}^{\mathsf{H}}\mathbf{K}^{-1}\boldsymbol{u} = \sigma^2\boldsymbol{u}^{\mathsf{H}}\mathbf{K}^{-2}\boldsymbol{u}$. ■

Lemma E.3. *For $\boldsymbol{\beta} \in \mathbb{C}^N$, let B be the Fourier series of the function $\langle \boldsymbol{\beta}, K(X_N, \cdot)\rangle_N$. Then,*

$$B[k] = \sqrt{N}\boldsymbol{\beta}^\top \boldsymbol{u}_{k\ mod\ N} \cdot G[k], \quad k \in \mathbb{Z}$$

Proof. Use the Fourier series definition to get,

$$\begin{aligned}\sum_{i=1}^N \beta_i \frac{1}{2\pi}\int_{-\pi}^{\pi} K(x_i, t)\,\mathrm{d}t &= \sum_{i=1}^N \beta_i \frac{1}{2\pi}\int_{-\pi}^{\pi} g(M(x_i - t) \bmod [-\pi, \pi))\,\mathrm{d}t \\ &= \sum_{i=1}^N \beta_i \frac{1}{2\pi}\int_{-\pi}^{\pi} g(M\tau)e^{-jk\tau}e^{-jkx_i}\,\mathrm{d}\tau = \sqrt{N}\beta^\top \boldsymbol{u}_{k \bmod N} G[k]\end{aligned}$$

since $e^{-jkx_i} = e^{-j\frac{2\pi}{N}ki} = e^{-j\frac{2\pi}{N}(k \bmod N)i} = \sqrt{N}\boldsymbol{u}_{k \bmod N}$. This concludes the proof. ■

Eigenfunctions of $\mathcal{T}_K$, $\mathcal{T}_K^N$ and eigenvectors of $\mathbf{K}$.

Proof of Proposition 3.1. It suffices to show the eigenvector equation for the unnormalized version of $\boldsymbol{u}_\ell$. We start by noting that

$$\mathbf{K}_{im'} = g(M(x_{m'} - x_i)) = \sum_{m\in\mathbb{Z}} G[m]e^{jm(x_{m'} - x_i)} = \sum_{m\in\mathbb{Z}} G[m]e^{j\frac{2\pi}{N}m(m'-i)}.$$

Using this, we have

$$
\begin{aligned}
(\mathbf{K}\boldsymbol{u}_\ell)_i &= \sum_{m'=0}^{N-1} \mathbf{K}_{im'} e^{-j\frac{2\pi}{N}m'\ell} = \sum_{m'}\sum_{m\in\mathbb{Z}} G[m] e^{j\frac{2\pi}{N}m(m'-i)} e^{-j\frac{2\pi}{N}m'\ell} \\
&= \sum_{m\in\mathbb{Z}} G[m] e^{-j\frac{2\pi}{N}mi} \sum_{m'=0}^{N-1} e^{j\frac{2\pi}{N}(m-\ell)m'} = N \sum_{m\in\mathbb{Z}} G[mN+\ell] e^{-j\frac{2\pi}{N}(mN+\ell)i} \\
&= N e^{-j\frac{2\pi}{N}\ell i} \sum_{m\in\mathbb{Z}} G[mN+\ell] e^{-j2\pi mi} = e^{-j\frac{2\pi}{N}\ell i} N \sum_{m\in\mathbb{Z}} G[mN+\ell] = e^{-j\frac{2\pi}{N}\ell i} \cdot N \left\| G_\ell \right\|_1
\end{aligned}
$$

This proves $\mathbf{K}\boldsymbol{u}_\ell = N \left\| G_\ell \right\|_1 \boldsymbol{u}_\ell$. The rest follows from standard results on linear algebra. ■

Proof of Proposition 3.2. Observe that

$$
\begin{aligned}
\mathcal{T}_K\left\{\phi_k\right\}(x) &= \frac{1}{2\pi}\int_{-\pi}^{\pi} K(x,x')\phi_k(x')\,\mathrm{d}x' = \frac{1}{2\pi}\int_{-\pi}^{\pi} g(M(x'-x \bmod [-\pi,\pi)))e^{jkx'}\,\mathrm{d}x' \\
&= e^{jkx}\cdot\frac{1}{2\pi}\int_{-\pi}^{\pi} g(Mu)e^{-jku}\,\mathrm{d}u = G[k]\phi(x)
\end{aligned}
$$

■

This proves the claim.

The following lemma relates the eigenfunctions of the empirical covariance operator defined in equation (2.7) to the eigenvectors of the kernel matrix.

Lemma E.4 (Eigenfunctions of $\mathcal{T}_K^n$). *Let (λ,ψ) be an eigenvalue-eigenfunction pair of $\mathcal{T}_K^n$. Assume $\mathbf{K}$ is invertible. Then for $\lambda > 0$, a unit-norm eigenfunction ψ satisfies,*

$$
\psi = \sum_{i=1}^{n} \frac{e_i}{\sqrt{n\lambda}}, K(x_i,\cdot), \tag{E.2}
$$

where $\boldsymbol{e} = (e_i) \in \mathbb{C}^n$ is a unit-norm eigenvector of $\mathbf{K}$ satisfying, $\mathbf{K}\boldsymbol{e} = n\lambda\boldsymbol{e}$.

We apply the above lemma to the setting described in Section 3. The proof is provided in Appendix F.4 in the supplementary materials.

Lemma E.5 (Eigenfunctions of $\mathcal{T}_K^N$). *The eigenfunctions for $\mathcal{T}_K^N$ are,*

$$
\psi_\ell = \frac{1}{\sqrt{\left\| G_\ell \right\|_1}} \sum_{m\in\mathbb{Z}} G[mN+\ell]\phi_{mN+\ell}, \qquad \ell \in [N].
$$

They satisfy, $\mathcal{T}_K^N\psi_\ell = \left\| G_\ell \right\|_1 \psi_\ell$, and their norms satisfy $\left\|\psi_\ell\right\|_{\mathcal{H}} = 1$, and $\left\|\psi_\ell\right\| = \frac{1}{\sqrt{\left\| G_\ell \right\|_1}} \left\| G_\ell \right\|$. Furthermore, ψ_ℓ are orthogonal in L^2, i.e., $\langle\psi_\ell,\psi_k\rangle = 0$ for $k \neq \ell$.

Proof of Lemma E.5. By Lemma E.4, we have

$$
\psi_\ell = \left\langle \frac{\boldsymbol{u}_\ell}{\sqrt{N\left\| G_\ell \right\|_1}}, K(X_N,\cdot) \right\rangle_N = \left\langle \frac{\overline{\boldsymbol{u}_\ell}}{\sqrt{N\left\| G_\ell \right\|_1}}, K(X_N,\cdot) \right\rangle_N
$$

Then using Lemma E.3 we have a Fourier series expansion of the form

$$\psi_\ell = \sum_{k\in\mathbb{Z}} \sqrt{N}\frac{\boldsymbol{u}_\ell^{\mathsf{H}}}{\sqrt{N\,\|G_\ell\|_1}}\boldsymbol{u}_{k \bmod N} G[k]\phi_k = \frac{1}{\sqrt{\|G_\ell\|_1}}\sum_{k\in\mathbb{Z}} \boldsymbol{u}_\ell^{\mathsf{H}}\boldsymbol{u}_{k \bmod N} G[k]\phi_k$$
$$= \frac{1}{\sqrt{\|G_\ell\|_1}}\sum_{m\in\mathbb{Z}} G[mN+\ell]\phi_{mN+\ell}$$

To see the orthogonality, suppose $k, \ell \in \{0, 1, \dots, N-1\}$ and $k \neq \ell$. Then

$$\langle\psi_\ell, \psi_k\rangle = \frac{1}{\sqrt{\|G_\ell\|_1\,\|G_k\|_1}}\sum_{m,m'\in\mathbb{Z}} G[mN+\ell]G[m'N+k]\,\langle\phi_{mN+\ell}, \phi_{m'N+k}\rangle = 0$$

For L^2 norm, substitute $k = \ell$ above to get,

$$\langle\psi_\ell, \psi_\ell\rangle = \frac{1}{\|G_\ell\|_1}\sum_{m,m'\in\mathbb{Z}} G[mN+\ell]G[m'N+\ell]\,\langle\phi_{mN+\ell}, \phi_{m'N+\ell}\rangle = \frac{1}{\|G_\ell\|_1}\|G_\ell\|^2$$

This proves the claim. ■

Appendix F. Proofs to technical lemmas.

Proposition F.1 (Parseval's theorem in high dimensions). *For a function $f : [-\pi, \pi)^d \to \mathbb{R}$ with Fourier series coefficients $F[\boldsymbol{k}]$ for $k \in \mathbb{Z}^d$, we have*

$$\sum_{\boldsymbol{k}\in\mathbb{Z}^d} |F[\boldsymbol{k}]|^2 = \frac{1}{(2\pi)^d}\int_{[-\pi,\pi)^d} |f(\boldsymbol{t})|^2\,\mathrm{d}\boldsymbol{t}.$$

Eigenfunctions of $\mathcal{T}_K$, $\mathcal{T}_K^n$ and eigenvectors of $\mathbf{K}$ for $d > 1$. The proofs of the following two statements are provided in Appendix F.4.

Proposition F.2. *Proposition 3.1 holds with $\boldsymbol{u}_{\boldsymbol{\ell}}$ from Definition D.2 and $\mathbf{K} = (K(\boldsymbol{x}_{\boldsymbol{p}}, \boldsymbol{x}_{\boldsymbol{p}'})) \in \mathbb{R}^{N^d\times N^d}$, with eigenvalue $\lambda_{\boldsymbol{\ell}} = N^d\,\|G_{\boldsymbol{\ell}}\|_1$,* i.e., $\mathbf{K}\boldsymbol{u}_{\boldsymbol{\ell}} = \lambda_{\boldsymbol{\ell}}\boldsymbol{u}_{\boldsymbol{\ell}}$.

Lemma F.3 (Eigenfunctions of $\mathcal{T}_K^{N,d}$). *The eigenfunctions for the empirical operator $\mathcal{T}_K^{N,d}$ are,*

$$\psi_{\boldsymbol{\ell}} = \frac{1}{\sqrt{\|G_{\boldsymbol{\ell}}\|_1}}\sum_{\boldsymbol{m}\in\mathbb{Z}^d} G[\boldsymbol{m}N+\boldsymbol{\ell}]\phi_{\boldsymbol{m}N+\boldsymbol{\ell}}, \qquad \boldsymbol{\ell}\in[N]^d.$$

They satisfy, $\mathcal{T}_K^{N,d}\{\psi_{\boldsymbol{\ell}}\} = \|G_{\boldsymbol{\ell}}\|_1\,\psi_{\boldsymbol{\ell}}$, and their norms satisfy $\|\psi_{\boldsymbol{\ell}}\|_{\mathcal{H}} = 1$, as well as $\|\psi_{\boldsymbol{\ell}}\| = \frac{1}{\sqrt{\|G_{\boldsymbol{\ell}}\|_1}}\|G_{\boldsymbol{\ell}}\|$. Furthermore, $\psi_{\boldsymbol{\ell}}$ are orthogonal in L^2, i.e., $\langle\psi_{\boldsymbol{\ell}}, \psi_{\boldsymbol{k}}\rangle = 0$ for $\boldsymbol{k} \neq \boldsymbol{\ell}$.

F.1. Approximation error: Proof of Lemma D.3(a).

Once again, the proof proceeds by applying the Pythagorean theorem to the triangle $\{0, f^*, P_X f^*\}$ in L^2. The following lemma gives exact expressions for projection of the target function and its norm.

Lemma F.4 (Projection). *For $f^* = \sum_{\boldsymbol{k}\in\mathbb{Z}^d} V[\boldsymbol{k}]\phi_{\boldsymbol{k}}$*

$$P_X f^* = \sum_{\boldsymbol{\ell}\in[N]^d}\frac{\langle G_{\boldsymbol{\ell}}, V_{\boldsymbol{\ell}}\rangle}{\|G_{\boldsymbol{\ell}}\|^2}\sum_{\boldsymbol{m}\in\mathbb{Z}^d} G[\boldsymbol{m}N+\boldsymbol{\ell}]\phi_{\boldsymbol{m}N+\boldsymbol{\ell}}, \qquad \text{and} \qquad \|P_X f^*\|^2 = \sum_{\boldsymbol{\ell}\in[N]^d}\frac{\langle G_{\boldsymbol{\ell}}, V_{\boldsymbol{\ell}}\rangle^2}{\|G_{\boldsymbol{\ell}}\|^2}$$

We get,

$$\|f^* - P_X f^*\|^2 = \|f^*\|^2 - \|P_X f^*\|^2 = \|V\|^2 - \sum_{\boldsymbol{\ell}\in[N]^d} \frac{\langle G_{\boldsymbol{\ell}}, V_{\boldsymbol{\ell}}\rangle^2}{\|G_{\boldsymbol{\ell}}\|^2}$$

Proof of Lemma F.4. Note that Lemma F.3 shows that $\left\{\frac{\psi_{\boldsymbol{\ell}}}{\|\psi_{\boldsymbol{\ell}}\|}\right\}_{\boldsymbol{\ell}\in[N]^d}$ is an orthonormal basis for $\mathrm{Span}\{K(\boldsymbol{x}_{\boldsymbol{\ell}}, \cdot)\}$. Consequently, we have

$$P_X f^* = \sum_{\boldsymbol{\ell}\in[N]^d} \left\langle f^*, \frac{\psi_{\boldsymbol{\ell}}}{\|\psi_{\boldsymbol{\ell}}\|}\right\rangle \frac{\psi_{\boldsymbol{\ell}}}{\|\psi_{\boldsymbol{\ell}}\|}, \qquad \text{and} \qquad \|P_X f^*\|^2 = \sum_{\boldsymbol{\ell}\in[N]^d} \left\langle f^*, \frac{\psi_{\boldsymbol{\ell}}}{\|\psi_{\boldsymbol{\ell}}\|}\right\rangle^2$$

We compute these projections. For $\boldsymbol{\ell} \in [N]^d$,

$$\begin{aligned}
\langle f^*, \psi_{\boldsymbol{\ell}}\rangle &= \frac{1}{\sqrt{\|G_{\boldsymbol{\ell}}\|_1}} \left\langle \sum_{\boldsymbol{k}\in\mathbb{Z}^d} V[\boldsymbol{k}]\phi_{\boldsymbol{k}}, \sum_{\boldsymbol{m}\in\mathbb{Z}^d} G[\boldsymbol{m}N + \boldsymbol{\ell}]\phi_{\boldsymbol{m}N+\boldsymbol{\ell}} \right\rangle \\
&= \frac{1}{\sqrt{\|G_{\boldsymbol{\ell}}\|_1}} \sum_{\boldsymbol{m},\boldsymbol{k}\in\mathbb{Z}^d} G[\boldsymbol{m}N + \boldsymbol{\ell}]V[\boldsymbol{k}] \langle \phi_{\boldsymbol{k}}, \phi_{\boldsymbol{m}N+\boldsymbol{\ell}}\rangle \\
&= \frac{1}{\sqrt{\|G_{\boldsymbol{\ell}}\|_1}} \sum_{\boldsymbol{m}\in\mathbb{Z}^d} G[\boldsymbol{m}N + \boldsymbol{\ell}]V[\boldsymbol{m}N + \boldsymbol{\ell}] = \frac{\langle G_{\boldsymbol{\ell}}, V_{\boldsymbol{\ell}}\rangle}{\sqrt{\|G_{\boldsymbol{\ell}}\|_1}}
\end{aligned}$$

Thus, we get that,

$$\left\langle f^*, \frac{\psi_{\boldsymbol{\ell}}}{\|\psi_{\boldsymbol{\ell}}\|}\right\rangle \frac{\psi_{\boldsymbol{\ell}}}{\|\psi_{\boldsymbol{\ell}}\|} = \frac{\langle G_{\boldsymbol{\ell}}, V_{\boldsymbol{\ell}}\rangle}{\|G_{\boldsymbol{\ell}}\|^2} \sum_{\boldsymbol{m}\in\mathbb{Z}^d} G[\boldsymbol{m}N + \boldsymbol{\ell}]\phi_{\boldsymbol{m}N+\boldsymbol{\ell}}$$

The claims follow immediately. ■

F.2. Noise-free estimation error: Proof of Lemma D.3(b).

Let F be the fourier series of $\left\langle \mathbf{K}^{-1} R_n\{f^* - P_X f^*\}, K(X_n, \cdot)\right\rangle_n$. From Lemma D.4 we have,

$$F[\boldsymbol{k}] = \sqrt{N^d} R_n\{f^* - P_X f^*\}^\top \mathbf{K}^{-1}\boldsymbol{u}_{\boldsymbol{k} \bmod N^d} \cdot G[\boldsymbol{k}]$$

By Parseval's theorem (Proposition F.1), we conclude,

$$\frac{1}{(2\pi)^d}\int_{[-\pi,\pi)} \left(\left\langle \boldsymbol{K}^{-1} R_n\{f^* - P_X f^*\}, K(X_n, t)\right\rangle_n\right)^2 \mathrm{d}t = \sum_{\boldsymbol{k}\in\mathbb{Z}^d} |F[\boldsymbol{k}]|^2$$

We will show that

$$R_n\{f^* - P_X f^*\}^\top \boldsymbol{K}^{-1}\boldsymbol{u}_{\boldsymbol{\ell}} = \left(\frac{\langle V_{\boldsymbol{\ell}}, \mathbf{1}\rangle}{\langle G_{\boldsymbol{\ell}}, \mathbf{1}\rangle} - \frac{\langle G_{\boldsymbol{\ell}}, V_{\boldsymbol{\ell}}\rangle}{\|G_{\boldsymbol{\ell}}\|^2}\right)$$

We have $\mathbf{K}^{-1}\boldsymbol{u}_{\boldsymbol{\ell}} = \boldsymbol{u}_{\boldsymbol{\ell}} \cdot \frac{1}{N^d \|G_{\boldsymbol{\ell}}\|_1}$. By Lemma B.1, we can write $P_X f^*$ on the data as

$$\begin{aligned}
P_X f^*(\boldsymbol{x}_{\boldsymbol{p}}) &= \sum_{\boldsymbol{\ell}\in[N]^d} \frac{\langle G_{\boldsymbol{\ell}}, V_{\boldsymbol{\ell}}\rangle}{\|G_{\boldsymbol{\ell}}\|^2} \sum_{\boldsymbol{m}\in\mathbb{Z}^d} G[\boldsymbol{m}N+\boldsymbol{\ell}]\phi_{\boldsymbol{m}N+\boldsymbol{\ell}}(\boldsymbol{x}_{\boldsymbol{p}}) = \sum_{\boldsymbol{\ell}\in[N]^d} \frac{\langle G_{\boldsymbol{\ell}}, V_{\boldsymbol{\ell}}\rangle}{\|G_{\boldsymbol{\ell}}\|^2} \|G_{\boldsymbol{\ell}}\|_1 \overline{u_{\boldsymbol{\ell}\boldsymbol{p}}} \\
f^*(\boldsymbol{x}_{\boldsymbol{p}}) &= \sum_{\boldsymbol{k}\in\mathbb{Z}^d} V[\boldsymbol{k}]\phi_{\boldsymbol{k}}(\boldsymbol{x}_{\boldsymbol{p}}) = \sum_{\boldsymbol{\ell}\in[N]^d} \langle V_{\boldsymbol{\ell}}, \mathbf{1}\rangle \overline{u_{\boldsymbol{\ell}\boldsymbol{p}}}
\end{aligned}$$

We thus have

$$\begin{aligned}
R_n\left\{f^* - P_X f^*\right\}^\top \mathbf{K}^{-1}\boldsymbol{u}_{\boldsymbol{\ell}} &= \sum_{\boldsymbol{p},\boldsymbol{\ell}'\in[N]^d} \left(\langle V_{\boldsymbol{\ell}'}, \mathbf{1}\rangle - \frac{\langle G_{\boldsymbol{\ell}'}, V_{\boldsymbol{\ell}'}\rangle}{\|G_{\boldsymbol{\ell}'}\|^2} \|G_{\boldsymbol{\ell}'}\|_1\right) \frac{\overline{u_{\boldsymbol{\ell}' i}} u_{\boldsymbol{\ell} i}}{N^d \|G_{\boldsymbol{\ell}}\|_1} \\
&= \frac{1}{N^d}\left(\frac{\langle V_{\boldsymbol{\ell}}, \mathbf{1}\rangle}{\|G_{\boldsymbol{\ell}}\|_1} - \frac{\langle V_{\boldsymbol{\ell}}, G_{\boldsymbol{\ell}}\rangle}{\|G_{\boldsymbol{\ell}}\|^2}\right)
\end{aligned}$$

This gives,

$$\sum_{\boldsymbol{k}\in\mathbb{Z}^d} |F[\boldsymbol{k}]|^2 = \frac{1}{N^d} \sum_{\boldsymbol{\ell}\in[N]^d} \left(\frac{\langle V_{\boldsymbol{\ell}}, \mathbf{1}\rangle}{\langle G_{\boldsymbol{\ell}}, \mathbf{1}\rangle} - \frac{\langle V_{\boldsymbol{\ell}}, G_{\boldsymbol{\ell}}\rangle}{\|G_{\boldsymbol{\ell}}\|^2}\right)^2 \|G_{\boldsymbol{\ell}}\|^2 .$$

F.3. Noisy estimation error: Proof of Lemma D.3(c). Similar to $d = 1$, we derive this by an application of Parseval's theorem.

Define the Fourier series,

$$\left\langle \mathbf{K}^{-1}\boldsymbol{\xi}, K(X_n, \boldsymbol{t})\right\rangle_n = \sum_{\boldsymbol{k}\in\mathbb{Z}^d} E[\boldsymbol{k}] \exp\left(j\left\langle \boldsymbol{k}, \boldsymbol{t}\right\rangle\right)$$

By Proposition F.1 (Parseval's theorem), we have,

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{\xi}}\left[\frac{1}{(2\pi)^d}\int_{[-\pi,\pi)} \left|\left\langle \mathbf{K}^{-1}\boldsymbol{\xi}, K(X_n, \boldsymbol{t})\right\rangle_n\right|^2 \mathrm{d}\boldsymbol{t}\right] = \sum_{\boldsymbol{k}\in\mathbb{Z}^d} \mathbb{E}_{\boldsymbol{\xi}} |E[\boldsymbol{k}]|^2 = \sum_{\boldsymbol{p}\in[N]^d}\sum_{\boldsymbol{m}\in\mathbb{Z}^d} \mathbb{E}_{\boldsymbol{\xi}} |E[\boldsymbol{m}N+\boldsymbol{p}]|^2 \\
&= \sum_{\boldsymbol{p}\in[N]^d}\sum_{\boldsymbol{m}\in\mathbb{Z}^d} |G[\boldsymbol{m}N+\boldsymbol{p}]|^2\, \mathbb{E}_{\boldsymbol{\xi}} \left|\boldsymbol{\xi}^\top \mathbf{K}^{-1}\boldsymbol{u}_i\right|^2 \cdot N^d = \sigma^2 \sum_{\boldsymbol{p}\in[N]^d} \|G_{\boldsymbol{p}}\|^2 \left(\boldsymbol{u}_{\boldsymbol{p}}^{\mathsf{H}} \mathbf{K}^{-2}\boldsymbol{u}_{\boldsymbol{p}}\right) \cdot N^d \\
&= \sigma^2 \sum_{\boldsymbol{p}\in[N]^d} \|G_{\boldsymbol{p}}\|^2 \frac{1}{N^{2d}\|G_{\boldsymbol{p}}\|_1^2} N^d = \sigma^2 \sum_{\boldsymbol{p}\in[N]} \frac{\|G_{\boldsymbol{p}}\|^2}{N^d \|G_{\boldsymbol{p}}\|_1^2}
\end{aligned}$$

where we have used Lemma D.4 in the second, and Lemma E.2 in the third, and Proposition F.2 in the last line.

F.4. Additional Proofs.

Proof of Lemma E.4. We will first show that ψ can be written as a linear combination of the n representers $\{K(x_i, \cdot)\}$.

$$\psi = \sum_{i=0}^{n-1} \beta_i K(x_i, \cdot) \tag{F.2}$$

Let $\psi \in \mathcal{H}$ be an eigenfunction of $\mathcal{T}_K^n$ with eigenvalue λ. Then by definition of $\mathcal{T}_K^n$ we have,

$$\lambda\psi = \mathcal{T}_K^n(\psi) = \frac{1}{n}\sum_{i=1}^{n}\langle K(x_i,\cdot),\psi\rangle_{\mathcal{H}} K(x_i,\cdot) = \frac{1}{n}\sum_{i=1}^{n}\psi(x_i)K(x_i,\cdot) \tag{F.3}$$

where the last equality holds due to the reproducing property of the kernel. Define $\beta_i = \frac{\psi(x_i)}{n\lambda}$ to show (F.2). Next, rewriting the equation for an eigenfunction ψ, expressed as (F.2), we get

$$\mathcal{T}_K^n\left(\sum_{i=1}^{n}\beta_i K(x_i,\cdot)\right) = \lambda\sum_{i=1}^{n}\beta_i K(x_i,\cdot). \tag{F.4}$$

By definition of $\mathcal{T}_K^n$ however we get,

$$\mathcal{T}_K^n\left(\sum_{i=0}^{n}\beta_i K(x_i,\cdot)\right) = \frac{1}{n}\sum_{i,j=1}^{n}\beta_i\langle K(x_j,\cdot),K(x_i,\cdot)\rangle_{\mathcal{H}} K(x_j,\cdot) = \frac{1}{n}\sum_{j=1}^{n}(\mathbf{K}\boldsymbol{\beta})_j K(x_j,\cdot) \tag{F.5}$$

Evaluating functions on the RHS of equations (F.2) and (F.5) at x_ℓ yields,

$$\frac{1}{n}\cdot\sum_{i=1}^{n}\sum_{j=1}^{n}\beta_i K(x_i,x_j)K(x_j,x_l) = \lambda\sum_{i=1}^{n}\beta_i K(x_i,x_l) \qquad \text{for all } \ell\in\{0,1,\ldots,n-1\}$$

Compactly these n equations can be written as:

$$\mathbf{K}^2\boldsymbol{\beta} = n\lambda\mathbf{K}\boldsymbol{\beta} \implies \mathbf{K}\boldsymbol{\beta} = n\lambda\boldsymbol{\beta}$$

since $\mathbf{K}$ is inverible. Thus $\boldsymbol{\beta}$ is a scaled eigenvector of $\mathbf{K}$. It remains to determine the scale of $\boldsymbol{\beta}$ that defines ψ.

Now, the norm of ψ can be simplified as

$$\begin{aligned}\|\psi\|_{\mathcal{H}}^2 = \left\langle\sum_{i=1}^{n}\beta_i K(x_i,\cdot),\sum_{j=1}^{n}\beta_j K(x_j,\cdot)\right\rangle_{\mathcal{H}} &= \sum_{i,j=1}^{n}\beta_i\overline{\beta_j}\,\langle K(x_i,\cdot),K(x_j,\cdot)\rangle_{\mathcal{H}} \\ &= \boldsymbol{\beta}^{\mathsf{H}}\mathbf{K}\boldsymbol{\beta} = n\lambda\,\|\boldsymbol{\beta}\|^2\,.\end{aligned}$$

Since ψ is unit norm, we have $\|\boldsymbol{\beta}\| = \frac{1}{\sqrt{n\lambda}}$. This concludes the proof. ■

Proof of Lemma D.4. Use the Fourier series definition to get,

$$\begin{aligned}&\sum_{\boldsymbol{p}\in[N]^d}\beta_{\boldsymbol{p}}\frac{1}{(2\pi)^d}\int_{[-\pi,\pi)} g(M(\boldsymbol{t}-\boldsymbol{x_p} \bmod [-\pi,\pi)))\exp\left(-j\left\langle\boldsymbol{k},\boldsymbol{t}\right\rangle\right)\,\mathrm{d}\boldsymbol{t} \\ &= \sum_{\boldsymbol{p}\in[N]^d}\frac{\beta_{\boldsymbol{p}}}{(2\pi)^d}\int_{[-\pi,\pi)} g(M\boldsymbol{\tau})\exp\left(-j\left\langle\boldsymbol{k},\boldsymbol{\tau}\right\rangle\right)\exp\left(-j\left\langle\boldsymbol{k},\boldsymbol{x_p}\right\rangle\right)\,\mathrm{d}\boldsymbol{\tau} = N^{d/2}\beta^{\top}\boldsymbol{u}_{\boldsymbol{k} \bmod N}G[\boldsymbol{k}]\end{aligned}$$

This concludes the proof. ■

Proof of Proposition F.2. It suffices to show the eigenvector equation for the unnormalized version of $\boldsymbol{u}_{\boldsymbol{\ell}}$. We start by noting that

$$\mathbf{K}_{\boldsymbol{p}\boldsymbol{m}'} = g(M(\boldsymbol{x}_{\boldsymbol{m}'} - \boldsymbol{x}_{\boldsymbol{p}})) = \sum_{\boldsymbol{m}\in\mathbb{Z}^d} G[\boldsymbol{m}]e^{j\langle \boldsymbol{m}, \boldsymbol{x}_{\boldsymbol{m}'} - \boldsymbol{x}_{\boldsymbol{p}}\rangle)} = \sum_{\boldsymbol{m}\in\mathbb{Z}^d} G[\boldsymbol{m}]e^{j\frac{2\pi}{N}\langle \boldsymbol{m}, \boldsymbol{m}' - \boldsymbol{p}\rangle}.$$

Using this, we have

$$\begin{aligned}
(\mathbf{K}\boldsymbol{u}_{\boldsymbol{\ell}})_{\boldsymbol{q}} &= \sum_{\boldsymbol{p}\in[N]^d} \mathbf{K}_{\boldsymbol{q},\boldsymbol{p}} u_{\boldsymbol{\ell}\boldsymbol{p}} = \sum_{\boldsymbol{p}} \sum_{\boldsymbol{m}'\in\mathbb{Z}^d} G[\boldsymbol{m}'] \exp\left(j\frac{2\pi}{N}\left\langle \boldsymbol{m}', \boldsymbol{p} - \boldsymbol{q}\right\rangle\right) \exp\left(\frac{-j2\pi}{N}\left\langle \boldsymbol{p}, \boldsymbol{\ell}\right\rangle\right) \\
&= \sum_{\boldsymbol{m}'\in\mathbb{Z}^d} G[\boldsymbol{m}']e^{-j\frac{2\pi}{N}\langle \boldsymbol{m}', \boldsymbol{q}\rangle} \sum_{\boldsymbol{p}\in[N]^d} e^{j\frac{2\pi}{N}\langle (\boldsymbol{m}' - \boldsymbol{\ell}), \boldsymbol{p}\rangle} = N^d \sum_{\boldsymbol{m}\in\mathbb{Z}^d} G[\boldsymbol{m}N + \boldsymbol{\ell}]e^{-j\frac{2\pi}{N}\langle \boldsymbol{m}N + \boldsymbol{\ell}, \boldsymbol{q}\rangle} \\
&= e^{-j\frac{2\pi}{N}\langle \boldsymbol{\ell}, \boldsymbol{q}\rangle} N^d \sum_{\boldsymbol{m}\in\mathbb{Z}^d} G[\boldsymbol{m}N + \boldsymbol{\ell}] = e^{-j\frac{2\pi}{N}\langle \boldsymbol{\ell}, \boldsymbol{q}\rangle} N^d \lambda_{\boldsymbol{\ell}}
\end{aligned}$$

This proves $\mathbf{K}\boldsymbol{u}_{\boldsymbol{\ell}} = N^d \left\|G_{\boldsymbol{\ell}}\right\|_1 \boldsymbol{u}_{\boldsymbol{\ell}}$. The rest follows from standard results on linear algebra. ■

Proof of Lemma F.3. By Lemma E.4, we have

$$\psi_{\boldsymbol{\ell}} = \left\langle \frac{\boldsymbol{u}_{\boldsymbol{\ell}}}{\sqrt{N^d \left\|G_{\boldsymbol{\ell}}\right\|_1}}, K(X_n, \cdot) \right\rangle_n = \left\langle \frac{\overline{\boldsymbol{u}_{\boldsymbol{\ell}}}}{\sqrt{N^d \left\|G_{\boldsymbol{\ell}}\right\|_1}}, K(X_n, \cdot) \right\rangle_n$$

Then using Lemma D.4 we have a Fourier series expansion of the form

$$\begin{aligned}
\psi_{\boldsymbol{\ell}} &= \sum_{\boldsymbol{k}\in\mathbb{Z}^d} \sqrt{N^d} \frac{\boldsymbol{u}_{\boldsymbol{\ell}}^{\mathsf{H}}}{\sqrt{N^d \left\|G_{\boldsymbol{\ell}}\right\|_1}} \boldsymbol{u}_{\boldsymbol{k} \bmod N} G[\boldsymbol{k}]\phi_{\boldsymbol{k}} = \frac{1}{\sqrt{\left\|G_{\boldsymbol{\ell}}\right\|_1}} \sum_{\boldsymbol{k}\in\mathbb{Z}^d} \boldsymbol{u}_{\boldsymbol{\ell}}^{\mathsf{H}} \boldsymbol{u}_{\boldsymbol{k} \bmod N} G[\boldsymbol{k}]\phi_{\boldsymbol{k}} \\
&= \frac{1}{\sqrt{\left\|G_{\boldsymbol{\ell}}\right\|_1}} \sum_{\boldsymbol{m}\in\mathbb{Z}^d} G[\boldsymbol{m}N + \boldsymbol{\ell}]\phi_{\boldsymbol{m}N + \boldsymbol{\ell}}
\end{aligned}$$

To see the orthogonality, suppose $\boldsymbol{k}, \boldsymbol{\ell} \in [N]^d$ and $\boldsymbol{k} \neq \boldsymbol{\ell}$. Then

$$\langle \psi_{\boldsymbol{\ell}}, \psi_{\boldsymbol{k}} \rangle = \frac{1}{\sqrt{\left\|G_{\boldsymbol{\ell}}\right\|_1 \left\|G_{\boldsymbol{k}}\right\|_1}} \sum_{\boldsymbol{m}, \boldsymbol{m}'\in\mathbb{Z}^d} G[\boldsymbol{m}N + \boldsymbol{\ell}]G[\boldsymbol{m}'N + \boldsymbol{k}] \left\langle \phi_{\boldsymbol{m}N + \boldsymbol{\ell}}, \phi_{\boldsymbol{m}'N + \boldsymbol{k}} \right\rangle = 0$$

For L^2 norm, substitute $\boldsymbol{k} = \boldsymbol{\ell}$ above to get,

$$\langle \psi_{\boldsymbol{\ell}}, \psi_{\boldsymbol{\ell}} \rangle = \frac{1}{\left\|G_{\boldsymbol{\ell}}\right\|_1} \sum_{\boldsymbol{m}, \boldsymbol{m}'\in\mathbb{Z}^d} G[\boldsymbol{m}N + \boldsymbol{\ell}]G[\boldsymbol{m}'N + \boldsymbol{\ell}] \left\langle \phi_{\boldsymbol{m}N + \boldsymbol{\ell}}, \phi_{\boldsymbol{m}'N + \boldsymbol{\ell}} \right\rangle = \frac{1}{\left\|G_{\boldsymbol{\ell}}\right\|_1} \left\|G_{\boldsymbol{\ell}}\right\|^2$$

This proves the claim. ■