1. Kernel function basics

Kernels methods are a useful family of techniques for regression, interpolation, and classification. Given a "data" space X (which we will always take, in these notes, to be $\mathbf{R}^d$ or a subset thereof), we will consider (real[1])"kernels" to be functions $k\colon X \times X \to \mathbf{R}$ with the follow properties:

- *Symmetry*: $k(x, y) = k(y, x)$.
- *Positive definiteness*: for all $x_1, \dots x_N \in X$, the *Gram matrix* K given by $K_{ij} = k(x_i, x_j)$ is positive semidefinite (which we write $K \succeq 0$), and if the x_i's are distinct, K is (strictly) positive definite (which we write $K \succ 0$).

A famous result of kernel theory is that a kernel k is positive definite if and only if it can be expressed in the following form:

$$k(x, y) = \langle \Phi(x), \Phi(y) \rangle,$$

where $\Phi\colon X \to \mathcal{H}$ is a map from X into some Hilbert space $\mathcal{H}$. However, if we are only given the function k, finding such a Hilbert space map Φ is often difficult.

Many of the kernels we use in practice (on $\mathbf{R}^d$) are *stationary*: i.e., $k(x, y) = \phi(x - y)$, where $\phi\colon \mathbf{R}^d \to \mathbf{R}$ is a function. A single-variable function such that a kernel constructed from it is positive definite is called a *positive definite function*. The following famous theorem characterizes all positive definite funcitons on $\mathbf{R}^d$:

Theorem 1.1 (Bochner's theorem [1, Theorem 6.6])**.** *A function $\phi\colon \mathbf{R}^d \to \mathbf{R}$ is positive definite if and only if it is the Fourier transform of a nonnegative and nonzero Radon measure on $\mathbf{R}^d$.*

A comprehensive treatment of what kinds of functions are "positive definite" on more general spaces is the book by Berg, Christensen, and Ressel [2].

Most of the stationary kernels we use in practice have an even more specific form: *radial basis functions* (RBFs) are kernels of the form $k(x, y) = \phi(\|x -$

[1]The complex-valued case is also very useful, but for simplicity we skip it; the biggest difficulty is keeping track of complex conjugates and the ordering of the arguments of k.

$y\|$), where $\|\cdot\|$ is the Euclidean (ℓ_2) norm on $\mathbf{R}^d$. Bochner's theorem also has versions for radial functions (again, see [1, Chapter 6]). A function $\phi\colon [0,\infty) \to \mathbf{R}$ which gives rise to a positive definite kernel on $\mathbf{R}^d$ is also called positive definite. In general, this will depend on the dimension d. However, certain functions form positive definite radial basis functions for all dimensions: these are called *completely monotone* functions. See [1, Chapter 7] for a characterization.

Examples of kernels from completely monotone functions (which are probably the most common used in practice) are the squared-exponential RBF

$$k(x,y) = e^{-\|x-y\|^2/2\ell^2}$$

and the Matérn kernels

$$k(x,y) = \frac{2^{1-\nu}}{\Gamma(\nu)}\left(\frac{\sqrt{2\nu}\|x-y\|}{\ell}\right)^{\nu} K_\nu\left(\frac{\sqrt{2\nu}\|x-y\|}{\ell}\right),$$

where $\nu, \ell > 0$, and K_ν denotes the modified Bessel function of the second kind of order ν (notation is from [3]). ν represents the smoothness of the kernel; it is a fact that as $\nu \to \infty$, the Matérn function converges to a squared-exponential function.

2. The RKHS

Given a kernel k, we can define its reproducing kernel Hilbert space (RKHS). We define a set of functions on X

$$\mathcal{H}_0 = \left\{ \sum_{i=1}^{N} a_i k(\cdot, x_i) : a_1, \dots, a_N \in \mathbf{R}, x_1, \dots, x_N \in X \right\},$$

with an inner product

$$\left\langle \sum_{i=1}^{M} a_i k(\cdot, x_i), \sum_{j=1}^{N} b_j k(\cdot, y_j) \right\rangle = \sum_{i=1}^{M}\sum_{i=1}^{N} a_i b_j k(x_i, x_j).$$

The fact that k is positive definite implies that this is a true inner product, so $\mathcal{H}_0$ is an inner product space. We then take our RKHS $\mathcal{H}$ to be the

topological completion of $\mathcal{H}_0$ with respect to the norm induced by the inner product.[2]

Note that if $f = \sum a_i k(\cdot, x_i)$, then

$$f(x) = \sum a_i k(x, x_i) = \sum a_i \langle k(\cdot, x), k(\cdot, x_i) = \langle f, k(\cdot x) \rangle.$$

In other words, taking an inner product with a kernel function centered at x produces the value of a function at x. This phenomenon is what gives us the term "reproducing kernel."

A useful property of the RKHS is the the *evaluation functional* $f \mapsto f(x)$ is bounded:

$$|f(x)| = |\langle f, k(\cdot, x) \rangle| \leq \|f\| \|k(\cdot, x)\| = \|f\| \sqrt{k(x, x)}.$$

Of course, this suggests that elements of $\mathcal{H}$ must have some kind of smoothness or regularity to them, since arbitrary sets of functions (such as L_2 functions on a subset of $\mathbf{R}^d$) do not have this property. Via the Riesz representation theorem, it can be shown that this is also a *sufficient* condition for a Hilbert space of functions on X to be an RKHS.

A classic paper on the abstract theory of RKHSs is [4].

3. Mercer's theorem—RKHS on sets with finite measure

The RKHS $\mathcal{H}$ as defined above is still a very abstract object. If we make some additional assumptions about the set on which the functions are defined, we can get a much more precise characterization.

Suppose X is compact and has a finite measure μ (e.g., X is a closed and bounded subset of $\mathbf{R}^d$, and μ is standard Lebesgue measure), and k is a continuous positive definite kernel. Then the integral operator $T\colon L_2(X) \to L_2(X)$ defined by

$$(Tf)(x) = \int_X k(x, y) f(y)\, dy$$

is a compact, self-adjoint operator. Therefore, it has an orthogonal decomposition

$$Tf = \sum_i \lambda_i f_i \otimes f_i,$$

[2]Essentially, this lets us consider infinite linear combinations of shifted kernel functions.

where $\{\lambda_i\}_{i=1}^{\infty}$ is the set of eigenvalues of T (which are necessarily nonnegative, and which we put in decreasing order), and the eigenfunctions $\{f_i\}$ form an orthonormal basis for $L_2(X)$. Furthermore, we can write

$$k(x,y) = \sum_{i=1}^{\infty} \lambda_i f_i(x) f_i(y),$$

which converges uniformly in x and y.

It is easily shown that

$$\operatorname{trace}(T) = \sum_{i=1}^{\infty} \lambda_i = \int_X k(x,x)\ dx,$$

and the squared Hilbert-Schmidt norm

$$\|T\|_{\mathrm{HS}}^2 = \sum_{i=1}^{\infty} \lambda_i^2 = \int_{X\times X} k^2(x,y)\ dx\ dy.$$

Note that because X has finite measure, and k is continuous (and therefore bounded), both quantities are finite.

The above-stated facts, which are generally called Mercer's theorem, are purely a result from functional analysis and the theory of integral operators. We can, however, prove some interesting consequences for the RKHS.

Proposition 3.1. *The RKHS $\mathcal{H}$ is given by*

$$\mathcal{H} = \left\{ \sum a_i f_i : \sum \frac{a_i^2}{\lambda_i} < \infty \right\},$$

and the RKHS inner product is given by

$$\left\langle \sum a_i f_i, \sum b_i f_i \right\rangle_{\mathcal{H}} = \sum \frac{a_i b_i}{\lambda_i}.$$

Proof. We define $\mathcal{H}'$ to be the Hilbert space of functions defined above with the above inner product. We first show that for each $x \in X$, $k(\cdot, x) \in \mathcal{H}'$. Indeed $k(\cdot,x) = \sum \lambda_i f_i(x) f_i$, and

$$\begin{aligned} \|k(\cdot,x)\|_{\mathcal{H}'}^2 &= \sum \frac{(\lambda_i f_i(x))^2}{\lambda_i} \\ &= \sum \lambda_i f_i^2(x) \\ &= k(x,x) \\ &< \infty. \end{aligned}$$

Furthermore, for $f = \sum a_i f_i \in \mathcal{H}'$,

$$\begin{aligned}\langle f, k(\cdot, x)\rangle_{\mathcal{H}'} &= \Big\langle \sum a_i f_i \sum \lambda_i f_i(x) f_i \Big\rangle_{\mathcal{H}'} \\ &= \sum \frac{a_i \lambda_i f_i(x)}{\lambda_i} \\ &= \sum a_i f_i(x) \\ &= f(x).\end{aligned}$$

Clearly, then, $\mathcal{H} \subset \mathcal{H}'$, and k satisfies the reproducing property on $\mathcal{H}'$ (which implies that the inner product of $\mathcal{H}'$ coincides with that of $\mathcal{H}$ on $\mathcal{H}$).

To show that we cannot have strict inclusion, suppose that f is in the orthogonal complement of $\mathcal{H}$ in $\mathcal{H}'$. All of the functions $k(\cdot, x)$ are in $\mathcal{H}$, so $\langle f, k(\cdot, x)\rangle = 0$. But k satisfies the reproducing property on all of $\mathcal{H}'$, so this means that $f(x) = 0$ for all x, i.e., $f = 0$. Because $\mathcal{H}$ is closed, this implies $\mathcal{H} = \mathcal{H}'$. □

A simple consequence of this fact is the following: $\mathcal{H} = T^{1/2}(L_2(X))$, and $T^{1/2}$ is, in fact, an isometry (preserving distances and inner products) between the two spaces. In fact, the set of vectors $\{\sqrt{\lambda_i} f_i\}_{i=1}^{\infty}$ is an orthonormal basis for $\mathcal{H}$.

4. Regression and Interpolation

We now consider how to estimate a function f^* given measurements of the form $y_i = f^*(x_i), i \in \{1, \ldots, N\}$. We will assume that $f^* \in \mathcal{H}$ (not always a reasonable assumption, considering the comment at the end of the last section!).

Since values of f^* can be expressed as inner products, we consider a more general framework: we make (potentially noisy) observation of the form $y = \mathcal{A} f^* + \epsilon \in \mathbf{R}^N$, where ϵ is a noise vector, and $\mathcal{A} \colon \mathcal{H} \to R^N$ is defined by

$$\mathcal{A} f = \begin{bmatrix} \langle g_1, f\rangle \\ \vdots \\ \langle g_N, f\rangle \end{bmatrix}$$

for some elements $g_1, \dots, g_N \in \mathcal{H}$. We try to estimate f^* by the following optimization problem:

$$\min_{f \in \mathcal{H}} \sum_{i=1}^{N} (y_i - \langle g_i, f \rangle)^2 + \alpha \|f\|^2 = \min_{f \in \mathcal{H}} \|y - \mathcal{A}\, f\|_{\ell_2^N}^2 + \alpha \|f\|^2,$$

where $\alpha \geq 0$ is a regularization parameter. The objective function F is strictly convex in f, so we can solve it by setting the gradient equal to 0:

$$\nabla F = 2\alpha f - 2\,\mathcal{A}^*(y - \mathcal{A}\, f) = 0. \tag{1}$$

There are two standard ways to solve (1). One way is to gather all the terms involving f on one side and solving, which results in the common ridge regression formula

$$\hat{f} = (\alpha\, \mathcal{I} + \mathcal{A}^*\, \mathcal{A})^{-1}\, \mathcal{A}^*\, y. \tag{2}$$

This formula can be useful for theoretical analysis, but, since it involves the inversion of an infinite-dimensional operator, it is not usually very tractable to compute directly. Furthermore, it is, in general, not even well-defined for $\alpha = 0$, since $\mathcal{A}^*\, \mathcal{A}$ cannot have full rank unless $\mathcal{H}$ is finite-dimensional.

Instead, we note that the solution $\hat{f}$ to (1) must have the form

$$\hat{f} = \mathcal{A}^*\, a = \sum_{i=1}^{N} a_i g_i$$

for some $a \in \mathbf{R}^N$. For any solution $\hat{a}$ to the equation

$$\alpha a - (y - \mathcal{A}\, \mathcal{A}^*\, a) = 0,$$

$\hat{f} = \mathcal{A}^*\, \hat{a}$ solves (1). We can solve this in terms of a as

$$\hat{a} = (\alpha I_N + \mathcal{A}\, \mathcal{A}^*)^{-1} y, \tag{3}$$

where I_N denotes the $N \times N$ identity matrix. Solving for $\hat{a}$ simply (or not, if N is large) involves inverting an $N \times N$ matrix. We can easily check that $\mathcal{A}\, \mathcal{A}^*$ is just the familiar Gram matrix of the set $\{g_i\}$; its (i, j)-th entry is the inner product $\langle g_i, g_j \rangle$.

For $\alpha > 0$, the optimization problem is strictly convex, so its solution is unique, and both formulas above give the same solution. If $\alpha = 0$ we can instead consider the limiting (as $\alpha \downarrow 0$) problem

$$\min_{f \in \mathcal{H}} \|f\| \text{ s.t. } \mathcal{A}\, f = y,$$

and it is easily seen by linear algebra arguments that, if $\mathcal{A}\,\mathcal{A}^*$ has full rank, the interpolant $\widehat{f} = \mathcal{A}^*(\mathcal{A}\,\mathcal{A}^*)^{-1}y$ is indeed the unique solution.

In our case, the linear measurements are simply point evaluations, so we can write $g_i = k(\cdot, x_i)$. Then $(\mathcal{A}\,\mathcal{A}^*)_{ij} = \langle k(\cdot, x_i), k(\cdot, x_j)\rangle = k(x_i, x_j)$, so $K = \mathcal{A}\,\mathcal{A}^*$ is simply the Gram matrix. Because k is positive definite, this matrix is full-rank whenever all of the $x_i's$ are distinct.

Another useful version of the formulas above is to write

$$\widehat{f}(x) = \sum_{i=1}^{N} y_i u_i(x),$$

where the "Lagrange functions" $\{u_i\}$ are defined by $u_i = \mathcal{A}^*(\alpha I_N + \mathcal{A}\,\mathcal{A}^*)^{-1}e_i$, where e_i is the ith standard basis vector in $\mathbf{R}^N$. This is equivalent to

$$\begin{bmatrix} u_1(x) \\ \vdots \\ u_N(x) \end{bmatrix} = (\alpha I_N + K)^{-1} \begin{bmatrix} k(x, X_1) \\ \vdots \\ k(x, X_N) \end{bmatrix}.$$

One can easily check that if $\alpha = 0$, and the x_i's are distinct, then $u_i(x_j) = \mathbf{1}_{\{i=j\}}$; thus $\widehat{f}$ indeed interpolates the observed values of f^*.

4.1. Some error analysis

Given a (deterministic) set of points $x_1, \ldots, x_N$, there is a quick way to get a pointwise error bound of our estimate in terms of $\|f^*\|_{\mathcal{H}}$. For $x_0 \in X$, we have

$$\begin{aligned}
|f^*(x_0) - \widehat{f}(x_0)| &= \left| f^*(x_0) - \sum_{i=1}^{N} y_i u_i(x_0) \right| \\
&= \left| f^*(x_0) - \sum_{i=1}^{N} (f^*(x_i) + \xi_i) u_i(x_0) \right| \\
&= \left| \left\langle f^*, k(x_0, \cdot) - \sum_{i=1}^{N} u_i(x_0) k(x_i, \cdot) \right\rangle_{\mathcal{H}} + \sum_{i=1}^{N} \xi_i u_i(x_0) \right| \\
&\leq \left\| k(x_0, \cdot) - \sum_{i=1}^{N} u_i(x_0) k(x_i, \cdot) \right\|_{\mathcal{H}} \cdot \|f^*\|_{\mathcal{H}} + \left| \sum_{i=1}^{N} \xi_i u_i(x_0) \right|.
\end{aligned}$$

Note that

$$\begin{aligned}\left\| k(x_0,\cdot) - \sum_{i=1}^N u_i(x_0) k(x_i,\cdot) \right\|_{\mathcal{H}}^2 &= \left\| k(x_0,\cdot) - (\alpha + \mathcal{A}^*\mathcal{A})^{-1}\mathcal{A}^*\mathcal{A}\, k(x_0,\cdot)\right\|_{\mathcal{H}}^2 \\ &= k(x_0,x_0) \\ &\quad - 2\langle (\alpha+\mathcal{A}^*\mathcal{A})^{-1}\mathcal{A}^*\mathcal{A}\, k(x_0,\cdot), k(x_0,\cdot)\rangle_{\mathcal{H}} \\ &\quad + \langle ((\alpha+\mathcal{A}^*\mathcal{A})^{-1}\mathcal{A}^*\mathcal{A})^2 k(x_0,\cdot), k(x_0,\cdot)\rangle_{\mathcal{H}} \\ &\le k(x_0,x_0) - \langle (\alpha+\mathcal{A}^*\mathcal{A})^{-1}\mathcal{A}^*\mathcal{A}\, k(x_0,\cdot), k(x_0,\cdot)\rangle_{\mathcal{H}} \\ &= k(x_0,x_0) - \sum_{i=1}^N k(x_0,x_i) u_i(x_0).\end{aligned}$$

Thus we can bound the "bias" error at x_0 (that depends on $\|f^*\|_{\mathcal{H}}$) in terms of how well the kernel regression/interpolation procedure recovers the function $k(x_0,\cdot)$.

5. Gaussian processes

Reproducing kernel Hilbert spaces have many similarities to Gaussian processes. A fairly digestible introduction to Gaussian processes, including their connection to RKHSs, can be found in the book [3]. An extremely technical overview of the close relationship between these concepts can be found in [5].

5.1. Review of (multivariate) normal distributions

The standard normal distribution $\mathcal{N}(0,1)$ has density $p(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$. A more general normal distribution with mean $\mu \in \mathbf{R}$ and variance $\sigma^2 > 0$ has the distribution of $\mu + \sigma X$, where $X \sim \mathcal{N}(0,1)$.

There are several ways to define a multivariate normal random variable; we use the following, which is fairly simple to work with:

Definition 5.1. A random variable X has a multivariate normal distribution if it can be written $X = AW + \mu$, where $\mu \in \mathbf{R}^d$, and, for some m, $A \in \mathbf{R}^{d\times m}$, and W is a vector of m i.i.d. standard normal random variables.

Besides the mean μ, the other characteristic quantity of the multivariate normal distribution is its covariance $\Sigma = \mathbf{E}\, XX^T$. One can easily verify

that, from the above definition, we have $\Sigma = AA^T$. If Σ is nonsingular (which is equivalent to A having linearly independent rows), the distribution of X has the familiar density

$$p(x) = \frac{1}{(2\pi)^{d/2}\sqrt{\det \Sigma}} \exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right).$$

An essential property of the multivariate normal random variable is that the vector $\langle X, z\rangle = \sum_{i=1}^d z_i X_i$ has a (univariate) normal distribution for any $z \in \mathbf{R}^d$.[3] This can easily be verified by the fact that a linear combination of i.i.d. standard normal variables is normal.[4]

5.2. Gaussian process definition

The next, more general step is to consider *functions* whose values are Gaussian:

Definition 5.2. A (centered) Gaussian process on a space X is a random function $Z\colon X \to \mathbf{R}$ such that, for every integer $N \geq 1$ and every $x_1, \dots, x_N \in X$, the vector $(Z(x_1), \dots, Z(x_N))$ has a zero-mean multivariate normal distribution.

The distribution of a Gaussian process is completely determined by its *covariance function* $k(x,y) := \mathbf{E}\, Z(x)Z(y)$. Given fixed $x_1, \dots, x_N \in X$, the normal random vector $(Z(x_1), \dots, Z(x_N))$ has covariance matrix K, where $K_{ij} = k(x_i, x_j)$. One can easily check that k is a positive semidefinite kernel on X.

5.3. Bayesian inference

Given $x_1, \dots, x_N \in X$, suppose we observe $y_i = Z(x_i) + \xi_i$ for each $i \in \{1, \dots, N\}$, where the ξ_i's are i.i.d. $\mathcal{N}(0, \sigma^2)$ random variables independent of Z. Because the posterior distribution of Z is a (non-zero-mean) Gaussian process, we can completely characterize it by computing its mean and covariance. Let $\bar{x}_1, \dots, \bar{x}_m \in X$. We want to find the distribution of $\bar{y} =$

[3] $\langle X, z\rangle \sim \mathcal{N}(\langle \mu, z\rangle, z^T\Sigma z)$, to be precise.

[4] This property is often used in more theoretical works as the *definition* of a multivariate normal variable, but deriving the density and other properties from this definition is more complicated than is appropriate for these notes. See [6, Chapter 1] for a review of this construction.

$(Z(\bar{x}_1), \dots, Z(\bar{x}_m))$ conditioned on our data $x = (x_1, \dots, x_N)$ and $y = (y_1, \dots, y_N)$. Note that $\mathbf{E}\, y_i y_j = k(x_i, x_j) + \sigma^2 \mathbf{1}_{\{i=j\}}$, so the Gaussian random vector y has zero mean and covariance $k(x,x) + \sigma^2 I_N$. Denote $\overline{K} = [k(\bar{x}_i, \bar{x}_j)]_{i,j}$ and $\widetilde{K} = [k(x_i, \bar{x}_j)]_{i,j}$. Bayes rule gives

$$\begin{aligned} p(\bar{y} \mid y) &= \frac{p(\bar{y}, y)}{\int p(\bar{y}, y)\ d\bar{y}\ dy} \\ &= \exp\left(-\frac{1}{2} \begin{bmatrix} y^T & \bar{y}^T \end{bmatrix} \begin{bmatrix} K + \sigma^2 I_N & \widetilde{K} \\ \widetilde{K}^T & \overline{K} \end{bmatrix}^{-1} \begin{bmatrix} y \\ \bar{y} \end{bmatrix} + C(y) \right) \end{aligned}$$

A Schur complement block matrix inverse formula gives

$$\begin{aligned} \begin{bmatrix} K + \sigma^2 I_N & \widetilde{K} \\ \widetilde{K}^T & \overline{K} \end{bmatrix}^{-1} &= \begin{bmatrix} I_N & -(K + \sigma^2 I_N)^{-1} \widetilde{K} \\ 0 & I_m \end{bmatrix} \\ &\quad \times \begin{bmatrix} (K + \sigma^2 I_N)^{-1} & 0 \\ 0 & (\overline{K} - \widetilde{K}^T (K + \sigma^2 I_N)^{-1} \widetilde{K})^{-1} \end{bmatrix} \\ &\quad \times \begin{bmatrix} I_N & 0 \\ -\widetilde{K}^T (K + \sigma^2 I_N)^{-1} & I_m \end{bmatrix}. \end{aligned}$$

Then, all of the terms involving $\bar{y}$ can be collected into a quadratic form of the vector $\bar{y} - \widetilde{K}^T (K + \sigma^2 I_N)^{-1} y$ on the matrix $(\overline{K} - \widetilde{K}^T (K + \sigma^2 I_N)^{-1} \widetilde{K})^{-1}$. Thus the posterior distribution of $\bar{y}$ is

$$\bar{y} \mid y \sim \mathcal{N}(\widetilde{K}^T (K + \sigma^2 I_N)^{-1} y, \overline{K} - \widetilde{K}^T (K + \sigma^2 I_N)^{-1} \widetilde{K}).$$

Considering a single point $x_0 \in X$, the posterior distribution of $Z(x_0)$ is normal with mean

$$\mathbf{E}[Z(x_0) \mid y] = \sum_{i=1}^{N} a_i k(x_0, x_i),$$

where $a = (K + \sigma^2 I_N)^{-1} y$. This is precisely the RKHS regression estimate with kernel k and regularization parameter $\alpha = \sigma^2$! One can quickly check that the variance of $Z(x_0)$ given y is

$$\operatorname{var}(Z(x_0) \mid y) = k(x_0, x_0) - \sum_{i=1}^{N} k(x_0, x_i) u_i(x_0).$$

Note that this we have seem this exact expression before in Section 4.1!

5.4. Karhunen-Loève decomposition

The following famous theorem describes how the Gaussian process can be decomposed according to Mercer's theorem:

Theorem 5.3 (Karhunen-Loève)**.** *Let Z be the Gaussian process on X with covariance function k. Let $T = \sum_{i=1}^{N} \lambda_i f_i \otimes f_i$ be the eigenvalue decomposition of the integral operator corresponding to k. Then, we can write*

$$Z(x) = \sum_{i=1}^{\infty} \sqrt{\lambda_i} Z_i f_i(x),$$

where the Z_i's are i.i.d. standard normal random variables, and the convergence is in mean square, uniformly in x.

Proof. Let

$$Z_i = \frac{1}{\sqrt{\lambda_i}} \langle Z, f_i \rangle_{L_2} = \frac{1}{\sqrt{\lambda_i}} \int_X Z(x) f_i(x)\, dx.$$

Each Z_i is Gaussian, as an integral (i.e., a limit of finite sums) of a Gaussian process. Furthermore,

$$\begin{aligned}
\mathbf{E}\, Z_i Z_j &= \frac{1}{\sqrt{\lambda_i \lambda_j}} \mathbf{E} \int_{X \times X} Z(x) f_i(x) Z(y) f_j(y)\, dx\, dy \\
&= \frac{1}{\sqrt{\lambda_i \lambda_j}} \int_{X \times X} \mathbf{E}\, Z(x) Z(y) f_i(x) f_j(y)\, dx\, dy \\
&= \frac{1}{\sqrt{\lambda_i \lambda_j}} \int_{X \times X} k(x,y) f_i(x) f_j(y)\, dx\, dy \\
&= \frac{1}{\sqrt{\lambda_i \lambda_j}} \langle T^{1/2} f_i, T^{1/2} f_j \rangle_{L_2} \\
&= \begin{cases} 1 & \text{if } i = j \\ 0 & \text{otherwise.} \end{cases}
\end{aligned}$$

Therefore, each $Z_i \sim \mathcal{N}(0,1)$, and, being jointly Gaussian random variables that are uncorrelated, they are independent.

To show convergence, note first that, for $x \in X$,

$$\begin{aligned}
\mathbf{E}\, Z(x)Z_i &= \frac{1}{\sqrt{\lambda_i}}\, \mathbf{E}\, Z(x) \int_X Z(y) f_i(y)\, dy \\
&= \frac{1}{\sqrt{\lambda_i}} \int_X \mathbf{E}\, Z(x)Z(y) f_i(y)\, dy \\
&= \frac{1}{\sqrt{\lambda_i}} \int_X k(x,y) f_i(y)\, dy \\
&= \sqrt{\lambda_i} f_i(x).
\end{aligned}$$

Then, for $N \geq 1$ and $x \in X$,

$$\begin{aligned}
\mathbf{E} \left(Z(x) - \sum_{i=1}^{N} \sqrt{\lambda_i} Z_i f_i(x) \right)^2 &= \mathbf{E}\, Z^2(x) + \sum_{i=1}^{N} \lambda_i f_i^2(x)\, \mathbf{E}\, Z_i^2 \\
&\quad - 2 \sum_{i=1}^{N} \sqrt{\lambda_i} f_i(x)\, \mathbf{E}\, Z(x) Z_i \\
&= k(x,x) - \sum_{i=1}^{N} \lambda_i f_i^2(x),
\end{aligned}$$

which, by Mercer's theorem, converges to zero uniformly in x as $N \to \infty$. $\square$

Note that the KL theorem implies that the RKHS norm of the canonical Gaussian process associated with K is infinite! An exploration of the connections (including this interesting paradox) between Gaussian processes and the RKHS of their kernels fact can be found in [5].

References

[1] H. Wendland, *Scattered Data Approximation*. Cambridge, 2005.

[2] C. Berg, J. P. R. Christensen, and P. Ressel, *Harmonic Analysis on Semigroups*. Springer, 1984.

[3] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*. Cambridge, MA: MIT Press, 2006.

[4] N. Aronszajn, "Theory of reproducing kernels," *Trans. Amer. Math. Soc.*, vol. 68, pp. 337–404, 1950.

[5] A. van der Vaart and J. H. van Zanten, "Reproducing kernel Hilbert spaces of Gaussian priors," in *Pushing the Limits of Contemporary Statistics: Contributions in Honor of Jayanta K. Ghosh*, ser. Collections, vol. 3, Institute of Mathematical Statistics, 2008, pp. 200–222.

[6] J.-F. Le Gall, *Brownian Motion, Martingales, and Stochastic Calculus*. Springer, 2016.