12

Reproducing kernel Hilbert spaces

Many problems in statistics—among them interpolation, regression and density estimation, as well as nonparametric forms of dimension reduction and testing—involve optimizing over function spaces. Hilbert spaces include a reasonably broad class of functions, and enjoy a geometric structure similar to ordinary Euclidean space. A particular class of function-based Hilbert spaces are those defined by reproducing kernels, and these spaces—known as reproducing kernel Hilbert spaces (RKHSs)—have attractive properties from both the computational and statistical points of view. In this chapter, we develop the basic framework of RKHSs, which are then applied to different problems in later chapters, including nonparametric least-squares (Chapter 13) and density estimation (Chapter 14).

12.1 Basics of Hilbert spaces

Hilbert spaces are particular types of vector spaces, meaning that they are endowed with the operations of addition and scalar multiplication. In addition, they have an inner product defined in the usual way:

Definition 12.1 An inner product on a vector space $\mathbb{V}$ is a mapping $\langle \cdot, \cdot \rangle_{\mathbb{V}} : \mathbb{V} \times \mathbb{V} \to \mathbb{R}$ such that

$$\langle f, g \rangle_{\mathbb{V}} = \langle g, f \rangle_{\mathbb{V}} \qquad \text{for all } f, g \in \mathbb{V}, \tag{12.1a}$$

$$\langle f, f \rangle_{\mathbb{V}} \geq 0 \qquad \text{for all } f \in \mathbb{V}, \text{ with equality iff } f = 0, \tag{12.1b}$$

$$\langle f + \alpha g, h \rangle_{\mathbb{V}} = \langle f, h \rangle_{\mathbb{V}} + \alpha \langle g, h \rangle_{\mathbb{V}} \qquad \text{for all } f, g, h \in \mathbb{V} \text{ and } \alpha \in \mathbb{R}. \tag{12.1c}$$

A vector space equipped with an inner product is known as an *inner product space*. Note that any inner product induces a norm via $\|f\|_{\mathbb{V}} := \sqrt{\langle f, f \rangle_{\mathbb{V}}}$. Given this norm, we can then define the usual notion of *Cauchy sequence*—that is, a sequence $(f_n)_{n=1}^{\infty}$ with elements in $\mathbb{V}$ is Cauchy if, for all $\epsilon > 0$, there exists some integer $N(\epsilon)$ such that

$$\|f_n - f_m\|_{\mathbb{V}} < \epsilon \qquad \text{for all } n, m \geq N(\epsilon).$$

Definition 12.2 A *Hilbert space* $\mathbb{H}$ is an inner product space $(\langle \cdot, \cdot \rangle_{\mathbb{H}}, \mathbb{H})$ in which every Cauchy sequence $(f_n)_{n=1}^{\infty}$ in $\mathbb{H}$ converges to some element $f^* \in \mathbb{H}$.

A metric space in which every Cauchy sequence $(f_n)_{n=1}^{\infty}$ converges to an element f^* of the space is known as *complete*. Thus, we can summarize by saying that a Hilbert space is a complete inner product space.

Example 12.3 (Sequence space $\ell^2(\mathbb{N})$) Consider the space of square-summable real-valued sequences, namely

$$\ell^2(\mathbb{N}) := \Big\{ (\theta_j)_{j=1}^{\infty} \mid \sum_{j=1}^{\infty} \theta_j^2 < \infty \Big\}.$$

This set, when endowed with the usual inner product $\langle \theta, \gamma \rangle_{\ell^2(\mathbb{N})} = \sum_{j=1}^{\infty} \theta_j \gamma_j$, defines a classical Hilbert space. It plays an especially important role in our discussion of eigenfunctions for reproducing kernel Hilbert spaces. Note that the Hilbert space $\mathbb{R}^m$, equipped with the usual Euclidean inner product, can be obtained as a finite-dimensional subspace of $\ell^2(\mathbb{N})$: in particular, the space $\mathbb{R}^m$ is isomorphic to the "slice"

$$\{\theta \in \ell^2(\mathbb{N}) \mid \theta_j = 0 \qquad \text{for all } j \geq m+1\}. \qquad \clubsuit$$

Example 12.4 (The space $L^2[0,1]$) Any element of the space $L^2[0,1]$ is a function $f\colon [0,1] \to \mathbb{R}$ that is Lebesgue-integrable, and whose square satisfies the bound $\|f\|_{L^2[0,1]}^2 = \int_0^1 f^2(x)\,dx < \infty$. Since this norm does not distinguish between functions that differ only on a set of zero Lebesgue measure, we are implicitly identifying all such functions. The space $L^2[0,1]$ is a Hilbert space when equipped with the inner product $\langle f, g \rangle_{L^2[0,1]} = \int_0^1 f(x)g(x)\,dx$. When the space $L^2[0,1]$ is clear from the context, we omit the subscript in the inner product notation. In a certain sense, the space $L^2[0,1]$ is equivalent to the sequence space $\ell^2(\mathbb{N})$. In particular, let $(\phi_j)_{j=1}^{\infty}$ be any complete orthonormal basis of $L^2[0,1]$. By definition, the basis functions satisfy $\|\phi_j\|_{L^2[0,1]} = 1$ for all $j \in \mathbb{N}$, and $\langle \phi_i, \phi_j \rangle = 0$ for all $i \neq j$, and, moreover, any function $f \in L^2[0,1]$ has the representation $f = \sum_{j=1}^{\infty} a_j \phi_j$, where $a_j := \langle f, \phi_j \rangle$ is the jth basis coefficient. By Parseval's theorem, we have

$$\|f\|_{L^2[0,1]}^2 = \sum_{j=1}^{\infty} a_j^2,$$

so that $f \in L^2[0,1]$ if and only if the sequence $a = (a_j)_{j=1}^{\infty} \in \ell^2(\mathbb{N})$. The correspondence $f \leftrightarrow (a_j)_{j=1}^{\infty}$ thus defines an isomorphism between $L^2[0,1]$ and $\ell^2(\mathbb{N})$. $\clubsuit$

All of the preceding examples are instances of *separable Hilbert spaces*, for which there is a countable dense subset. For such Hilbert spaces, we can always find a collection of functions $(\phi_j)_{j=1}^{\infty}$, orthonormal in the Hilbert space—meaning that $\langle \phi_i, \phi_j \rangle_{\mathbb{H}} = \delta_{ij}$ for all positive integers i, j—such that any $f \in \mathbb{H}$ can be written in the form $f = \sum_{j=1}^{\infty} a_j \phi_j$ for some sequence of coefficients $(a_j)_{j=1}^{\infty} \in \ell^2(\mathbb{N})$. Although there do exist non-separable Hilbert spaces, here we focus primarily on the separable case.

The notion of a linear functional plays an important role in characterizing reproducing kernel Hilbert spaces. A *linear functional* on a Hilbert space $\mathbb{H}$ is a mapping $L: \mathbb{H} \to \mathbb{R}$ that is linear, meaning that $L(f + \alpha g) = L(f) + \alpha L(g)$ for all $f, g \in \mathbb{H}$ and $\alpha \in \mathbb{R}$. A linear functional is said to be *bounded* if there exists some $M < \infty$ such that $|L(f)| \leq M\|f\|_{\mathbb{H}}$ for all $f \in \mathbb{H}$. Given any $g \in \mathbb{H}$, the mapping $f \mapsto \langle f, g\rangle_{\mathbb{H}}$ defines a linear functional. It is bounded, since by the Cauchy–Schwarz inequality we have $|\langle f, g\rangle_{\mathbb{H}}| \leq M\|f\|_{\mathbb{H}}$ for all $f \in \mathbb{H}$, where $M := \|g\|_{\mathbb{H}}$. The Riesz representation theorem guarantees that every bounded linear functional arises in exactly this way.

Theorem 12.5 (Riesz representation theorem) *Let L be a bounded linear functional on a Hilbert space. Then there exists a unique $g \in \mathbb{H}$ such that $L(f) = \langle f, g\rangle_{\mathbb{H}}$ for all $f \in \mathbb{H}$. (We refer to g as the representer of the functional L.)*

Proof Consider the nullspace $\mathbb{N}(L) = \{h \in \mathbb{H} \mid L(h) = 0\}$. Since L is a bounded linear operator, the nullspace is closed (see Exercise 12.1). Moreover, as we show in Exercise 12.3, for any such closed subspace, we have the direct sum decomposition $\mathbb{H} = \mathbb{N}(L) + [\mathbb{N}(L)]^{\perp}$, where $[\mathbb{N}(L)]^{\perp}$ consists of all $g \in \mathbb{H}$ such that $\langle h, g\rangle_{\mathbb{H}} = 0$ for all $h \in \mathbb{N}(L)$. If $\mathbb{N}(L) = \mathbb{H}$, then we take $g = 0$. Otherwise, there must exist a non-zero element $g_0 \in [\mathbb{N}(L)]^{\perp}$, and by rescaling appropriately, we may find some $g \in [\mathbb{N}(L)]^{\perp}$ such that $\|g\|_{\mathbb{H}} = L(g) > 0$. We then define $h := L(f)g - L(g)f$, and note that $L(h) = 0$ so that $h \in \mathbb{N}(L)$. Consequently, we must have $\langle g, h\rangle_{\mathbb{H}} = 0$, which implies that $L(f) = \langle g, f\rangle_{\mathbb{H}}$ as desired. As for uniqueness, suppose that there exist $g, g' \in \mathbb{H}$ such that $\langle g, f\rangle_{\mathbb{H}} = L(f) = \langle g', f\rangle_{\mathbb{H}}$ for all $f \in \mathbb{H}$. Rearranging yields $\langle g - g', f\rangle_{\mathbb{H}} = 0$ for all $f \in \mathbb{H}$, and setting $f = g - g'$ shows that $\|g - g'\|_{\mathbb{H}}^2 = 0$, and hence $g = g'$ as claimed. □

12.2 Reproducing kernel Hilbert spaces

We now turn to the notion of a reproducing kernel Hilbert space, or RKHS for short. These Hilbert spaces are particular types of function spaces—more specifically, functions f with domain $\mathcal{X}$ mapping to the real line $\mathbb{R}$. There are many different but equivalent ways in which to define an RKHS. One way is to begin with the notion of a positive semidefinite kernel function, and use it to construct a Hilbert space in an explicit way. A by-product of this construction is the reproducing property of the kernel. An alternative, and somewhat more abstract, way is by restricting attention to Hilbert spaces in which the evaluation functionals—that is, the mappings from the Hilbert space to the real line obtained by evaluating each function at a given point—are bounded. These functionals are particularly relevant in statistical settings, since many applications involve sampling a function at a subset of points on its domain. As our development will clarify, these two approaches are equivalent in that the kernel acts as the representer for the evaluation functional, in the sense of the Riesz representation theorem (Theorem 12.5).

12.2.1 Positive semidefinite kernel functions

Let us begin with the notion of a positive semidefinite kernel function. It is a natural generalization of the idea of a positive semidefinite matrix to the setting of general functions.

Definition 12.6 (Positive semidefinite kernel function) A symmetric bivariate function $\mathcal{K}: \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is positive semidefinite (PSD) if for all integers $n \geq 1$ and elements $\{x_i\}_{i=1}^n \subset \mathcal{X}$, the $n \times n$ matrix with elements $\mathbf{K}_{ij} := \mathcal{K}(x_i, x_j)$ is positive semidefinite.

This notion is best understood via some examples.

Example 12.7 (Linear kernels) When $\mathcal{X} = \mathbb{R}^d$, we can define the linear kernel function $\mathcal{K}(x, x') := \langle x, x' \rangle$. It is clearly a symmetric function of its arguments. In order to verify the positive semidefiniteness, let $\{x_i\}_{i=1}^n$ be an arbitrary collection of points in $\mathbb{R}^d$, and consider the matrix $\mathbf{K} \in \mathbb{R}^{n \times n}$ with entries $K_{ij} = \langle x_i, x_j \rangle$. For any vector $\alpha \in \mathbb{R}^n$, we have

$$\alpha^T \mathbf{K} \alpha = \sum_{i,j=1}^n \alpha_i \alpha_j \langle x_i, x_j \rangle = \Big\| \sum_{i=1}^n a_i x_i \Big\|_2^2 \geq 0.$$

Since $n \in \mathbb{N}$, $\{x_i\}_{i=1}^n$ and $\alpha \in \mathbb{R}^n$ were all arbitrary, we conclude that $\mathcal{K}$ is positive semidefinite. ♣

Example 12.8 (Polynomial kernels) A natural generalization of the linear kernel on $\mathbb{R}^d$ is the *homogeneous polynomial kernel* $\mathcal{K}(x, z) = (\langle x, z \rangle)^m$ of degree $m \geq 2$, also defined on $\mathbb{R}^d$. Let us demonstrate the positive semidefiniteness of this function in the special case $m = 2$. Note that we have

$$\mathcal{K}(x, z) = \Big(\sum_{j=1}^d x_j z_j \Big)^2 = \sum_{j=1}^d x_j^2 z_j^2 + 2 \sum_{i<j} x_i x_j (z_i z_j).$$

Setting $D = d + \binom{d}{2}$, let us define a mapping $\Phi: \mathbb{R}^d \to \mathbb{R}^D$ with entries

$$\Phi(x) = \begin{bmatrix} x_j^2, & \text{for } j = 1, 2, \ldots, d \\ \sqrt{2} x_i x_j, & \text{for } i < j \end{bmatrix}, \tag{12.2}$$

corresponding to all polynomials of degree two in $(x_1, \ldots, x_d)$. With this definition, we see that $\mathcal{K}$ can be expressed as a Gram matrix—namely, in the form $\mathcal{K}(x, z) = \langle \Phi(x), \Phi(z) \rangle_{\mathbb{R}^D}$. Following the same argument as Example 12.7, it is straightforward to verify that this Gram representation ensures that $\mathcal{K}$ must be positive semidefinite.

An extension of the homogeneous polynomial kernel is the *inhomogeneous polynomial kernel* $\mathcal{K}(x, z) = (1 + \langle x, z \rangle)^m$, which is based on all polynomials of degree m or less. We leave it as an exercise for the reader to check that it is also a positive semidefinite kernel function. ♣

Example 12.9 (Gaussian kernels) As a more exotic example, given some compact subset $\mathcal{X} \subseteq \mathbb{R}^d$, consider the Gaussian kernel $\mathcal{K}(x, z) = \exp\left(-\frac{1}{2\sigma^2} \|x - z\|_2^2\right)$. Here, unlike the linear

kernel and polynomial kernels, it is not immediately obvious that $\mathcal{K}$ is positive semidefinite, but it can be verified by building upon the PSD nature of the linear and polynomial kernels (see Exercise 12.19). The Gaussian kernel is a very popular choice in practice, and we return to study it further in the sequel. ♣

12.2.2 Feature maps in $\ell^2(\mathbb{N})$

The mapping $x \mapsto \Phi(x)$ defined for the polynomial kernel in equation (12.2) is often referred to as a *feature map*, since it captures the sense in which the polynomial kernel function embeds the original data into a higher-dimensional space. The notion of a feature mapping can be used to define a PSD kernel in far more generality. Indeed, any function $\Phi : \mathcal{X} \to \ell^2(\mathbb{N})$ can be viewed as mapping the original space $\mathcal{X}$ to some subset of the space $\ell^2(\mathbb{N})$ of all square-summable sequences. Our previously discussed mapping (12.2) for the polynomial kernel is a special case, since $\mathbb{R}^D$ is a finite-dimensional subspace of $\ell^2(\mathbb{N})$.

Given any such feature map, we can then define a symmetric kernel via the inner product $\mathcal{K}(x, z) = \langle \Phi(x), \Phi(z) \rangle_{\ell_2(\mathbb{N})}$. It is often the case, for suitably chosen feature maps, that this kernel has a closed-form expression in terms of the pair (x, z). Consequently, we can compute inner products between the embedded data pairs $(\Phi(x), \Phi(z))$ without actually having to work in $\ell^2(\mathbb{N})$, or some other high-dimensional space. This fact underlies the power of RKHS methods, and goes under the colloquial name of the "kernel trick". For example, in the context of the mth-degree polynomial kernel on $\mathbb{R}^d$ from Example 12.8, evaluating the kernel requires on the order of d basic operations, whereas the embedded data lies in a space of roughly d^m (see Exercise 12.11). Of course, there are other kernels that implicitly embed the data in some infinite-dimensional space, with the Gaussian kernel from Example 12.9 being one such case.

Let us consider a particular form of feature map that plays an important role in subsequent analysis:

Example 12.10 (PSD kernels from basis expansions) Consider the sinusoidal Fourier basis functions $\phi_j(x) := \sin\left(\frac{(2j-1)\pi x}{2}\right)$ for all $j \in \mathbb{N} = \{1, 2, \ldots\}$. By construction, we have

$$\langle \phi_j, \phi_k \rangle_{L^2[0,1]} = \int_0^1 \phi_j(x)\phi_k(x)\, dx = \begin{cases} 1 & \text{if } j = k, \\ 0 & \text{otherwise,} \end{cases}$$

so that these functions are orthonormal in $L^2[0, 1]$. Now given some sequence $(\mu_j)_{j=1}^{\infty}$ of non-negative weights for which $\sum_{j=1}^{\infty} \mu_j < \infty$, let us define the feature map

$$\Phi(x) := \left(\sqrt{\mu_1}\phi_1(x), \ \sqrt{\mu_2}\phi_2(x), \ \sqrt{\mu_3}\phi_3(x), \ \ldots \right).$$

By construction, the element $\Phi(x)$ belongs to $\ell^2(\mathbb{N})$, since

$$\|\Phi(x)\|^2_{\ell^2(\mathbb{N})} = \sum_{j=1}^{\infty} \mu_j \phi_j^2(x) \leq \sum_{j=1}^{\infty} \mu_j \ < \infty.$$

Consequently, this particular choice of feature map defines a PSD kernel of the form

$$\mathcal{K}(x, z) := \langle \Phi(x), \Phi(z) \rangle_{\ell^2(\mathbb{N})} = \sum_{j=1}^{\infty} \mu_j \phi_j(x) \phi_j(z).$$

As our development in the sequel will clarify, a very broad class of PSD kernel functions can be generated in this way. ♣

12.2.3 Constructing an RKHS from a kernel

In this section, we show how any positive semidefinite kernel function $\mathcal{K}$ defined on the Cartesian product space $\mathcal{X} \times \mathcal{X}$ can be used to construct a particular Hilbert space of functions on $\mathcal{X}$. This Hilbert space is unique, and has the following special property: for any $x \in \mathcal{X}$, the function $\mathcal{K}(\cdot, x)$ belongs to $\mathbb{H}$, and satisfies the relation

$$\langle f, \mathcal{K}(\cdot, x) \rangle_{\mathbb{H}} = f(x) \qquad \text{for all } f \in \mathbb{H}. \tag{12.3}$$

This property is known as the *kernel reproducing property* for the Hilbert space, and it underlies the power of RKHS methods in practice. More precisely, it allows us to think of the kernel itself as defining a feature map[1] $x \mapsto \mathcal{K}(\cdot, x) \in \mathbb{H}$. Inner products in the embedded space reduce to kernel evaluations, since the reproducing property ensures that $\langle \mathcal{K}(\cdot, x), \mathcal{K}(\cdot, z) \rangle_{\mathbb{H}} = \mathcal{K}(x, z)$ for all $x, z \in \mathcal{X}$. As mentioned earlier, this computational benefit of the RKHS embedding is often referred to as the kernel trick.

How does one use a kernel to define a Hilbert space with the reproducing property (12.3)? Recalling the definition of a Hilbert space, we first need to form a vector space of functions, and then we need to endow it with an appropriate inner product. Accordingly, let us begin by considering the set $\widetilde{\mathbb{H}}$ of functions of the form $f(\cdot) = \sum_{j=1}^{n} \alpha_j \mathcal{K}(\cdot, x_j)$ for some integer $n \geq 1$, set of points $\{x_j\}_{j=1}^{n} \subset \mathcal{X}$ and weight vector $\alpha \in \mathbb{R}^n$. It is easy to see that the set $\widetilde{\mathbb{H}}$ forms a vector space under the usual definitions of function addition and scalar multiplication.

Given any pair of functions $f, \bar{f}$ in our vector space—let us suppose that they take the form $f(\cdot) = \sum_{j=1}^{n} \alpha_j \mathcal{K}(\cdot, x_j)$ and $\bar{f}(\cdot) = \sum_{k=1}^{\bar{n}} \bar{\alpha}_k \mathcal{K}(\cdot, \bar{x}_k)$—we propose to define their inner product as

$$\left\langle f, \bar{f} \right\rangle_{\widetilde{\mathbb{H}}} := \sum_{j=1}^{n} \sum_{k=1}^{\bar{n}} \alpha_j \bar{\alpha}_k \mathcal{K}(x_j, \bar{x}_k). \tag{12.4}$$

It can be verified that this definition is independent of the particular representation of the functions f and $\bar{f}$. Moreover, this proposed inner product does satisfy the kernel reproducing property (12.3), since by construction, we have

$$\langle f, \mathcal{K}(\cdot, x) \rangle_{\widetilde{\mathbb{H}}} = \sum_{j=1}^{n} \alpha_j \mathcal{K}(x_j, x) = f(x).$$

Of course, we still need to verify that the definition (12.4) defines a valid inner product. Clearly, it satisfies the symmetry (12.1a) and linearity requirements (12.1c) of an inner

[1] This view—with the kernel itself defining an embedding from $\mathcal{X}$ to $\mathbb{H}$—is related to but slightly different than our earlier perspective, in which the feature map Φ was a mapping from $\mathcal{X}$ to $\ell^2(\mathbb{N})$. Mercer's theorem allows us to connect these two points of view; see equation (12.14) and the surrounding discussion.

product. However, we need to verify the condition (12.1b)—namely, that $\langle f, f\rangle_{\mathbb{H}} \geq 0$ with equality if and only if $f = 0$. After this step, we will have a valid inner product space, and the final step is to take closures of it (in a suitable sense) in order to obtain a Hilbert space. With this intuition in place, we now provide a formal statement, and then prove it:

Theorem 12.11 *Given any positive semidefinite kernel function $\mathcal{K}$, there is a unique Hilbert space $\mathbb{H}$ in which the kernel satisfies the reproducing property* (12.3)*. It is known as the reproducing kernel Hilbert space associated with $\mathcal{K}$.*

Proof As outlined above, there are three remaining steps in the proof, and we divide our argument accordingly.

Verifying condition (12.1b)*:* The positive semidefiniteness of the kernel function $\mathcal{K}$ implies that $\|f\|_{\widetilde{\mathbb{H}}^2} = \langle f, f\rangle_{\widetilde{\mathbb{H}}} \geq 0$ for all f, so we need only show that $\|f\|^2_{\widetilde{\mathbb{H}}} = 0$ if and only if $f = 0$. Consider a function of the form $f(\cdot) = \sum_{i=1}^n \alpha_i \mathcal{K}(\cdot, x_i)$, and suppose that

$$\langle f, f\rangle_{\widetilde{\mathbb{H}}} = \sum_{i,j=1}^n \alpha_i \alpha_j \mathcal{K}(x_j, x_i) = 0.$$

We must then show that $f = 0$, or equivalently that $f(x) = \sum_{i=1}^n \alpha_i \mathcal{K}(x, x_i) = 0$ for all $x \in \mathcal{X}$. Let $(a, x) \in \mathbb{R} \times \mathcal{X}$ be arbitrary, and note that by the positive semidefiniteness of $\mathcal{K}$, we have

$$0 \leq \|a\mathcal{K}(\cdot, x) + \sum_{i=1}^n \alpha_i \mathcal{K}(\cdot, x_i)\|^2_{\mathbb{H}} = a^2 \mathcal{K}(x, x) + 2a \sum_{i=1}^n \alpha_i \mathcal{K}(x, x_i).$$

Since $\mathcal{K}(x, x) \geq 0$ and the scalar $a \in \mathbb{R}$ is arbitrary, this inequality can hold only if $\sum_{i=1}^n \alpha_i \mathcal{K}(x, x_i) = 0$. Thus, we have shown that the pair $(\widetilde{\mathbb{H}}, \langle \cdot, \cdot \rangle_{\widetilde{\mathbb{H}}})$ is an inner product space.

Completing the space: It remains to extend $\widetilde{\mathbb{H}}$ to a complete inner product space—that is, a Hilbert space—with the given reproducing kernel. If $(f_n)_{n=1}^\infty$ is a Cauchy sequence in $\widetilde{\mathbb{H}}$, then for each $x \in \mathcal{X}$, the sequence $(f_n(x))_{n=1}^\infty$ is Cauchy in $\mathbb{R}$, and so must converge to some real number. We can thus define the pointwise limit function $f(x) := \lim_{n\to\infty} f_n(x)$, and we let $\mathbb{H}$ be the completion of $\widetilde{\mathbb{H}}$ by these objects. We define the norm of the limit function f as $\|f\|_{\mathbb{H}} := \lim_{n\to\infty} \|f_n\|_{\widetilde{\mathbb{H}}}$.

In order to verify that this definition is sensible, we need to show that for any Cauchy sequence $(g_n)_{n=1}^\infty$ in $\widetilde{\mathbb{H}}$ such that $\lim_{n\to\infty} g_n(x) = 0$ for all $x \in \mathcal{X}$, we also have $\lim_{n\to\infty} \|g_n\|_{\widetilde{\mathbb{H}}} = 0$. Taking subsequences as necessary, suppose that $\lim_{n\to\infty} \|g_n\|^2_{\widetilde{\mathbb{H}}} = 2\epsilon > 0$, so that for n, m sufficiently large, we have $\|g_n\|^2_{\widetilde{\mathbb{H}}} \geq \epsilon$ and $\|g_m\|^2_{\widetilde{\mathbb{H}}} > \epsilon$. Since the sequence $(g_n)_{n=1}^\infty$ is Cauchy, we also have $\|g_n - g_m\|_{\widetilde{\mathbb{H}}} < \epsilon/2$ for n, m sufficiently large. Now since $g_m \in \widetilde{\mathbb{H}}$, we can write $g_m(\cdot) = \sum_{i=1}^{N_m} \alpha_i \mathcal{K}(\cdot, x_i)$, for some finite positive integer N_m and vector $\alpha \in \mathbb{R}^{N_m}$. By the

reproducing property, we have

$$\langle g_m, g_n \rangle_{\widetilde{\mathbb{H}}} = \sum_{i=1}^{N_m} \alpha_i g_n(x_i) \to 0 \qquad \text{as } n \to +\infty,$$

since $g_n(x) \to 0$ for each fixed x. Hence, for n sufficiently large, we can ensure that $|\langle g_m, g_n \rangle_{\widetilde{\mathbb{H}}}| \leq \epsilon/2$. Putting together the pieces, we have

$$\|g_n - g_m\|_{\widetilde{\mathbb{H}}}^2 = \|g_n\|_{\widetilde{\mathbb{H}}}^2 + \|g_m\|_{\widetilde{\mathbb{H}}}^2 - 2\langle g_n, g_m \rangle_{\widetilde{\mathbb{H}}} \geq \epsilon + \epsilon - \epsilon = \epsilon.$$

But this lower bound contradicts the fact that $\|g_n - g_m\|_{\widetilde{\mathbb{H}}} \leq \epsilon/2$.

Thus, the norm that we have defined is sensible, and it can be used to define an inner product on $\mathbb{H}$ via the polarization identity

$$\langle f, g \rangle_{\mathbb{H}} := \tfrac{1}{2} \left\{ \|f + g\|_{\mathbb{H}}^2 - \|f\|_{\mathbb{H}}^2 + \|g\|_{\mathbb{H}}^2 \right\}.$$

With this definition, it can be shown that $\langle \mathcal{K}(\cdot, x), f \rangle_{\mathbb{H}} = f(x)$ for all $f \in \mathbb{H}$, so that $\mathcal{K}(\cdot, x)$ is again reproducing over $\mathbb{H}$.

Uniqueness: Finally, let us establish uniqueness. Suppose that $\mathbb{G}$ is some other Hilbert space with $\mathcal{K}$ as its reproducing kernel, so that $\mathcal{K}(\cdot, x) \in \mathbb{G}$ for all $x \in \mathcal{X}$. Since $\mathbb{G}$ is complete and closed under linear operations, we must have $\mathbb{H} \subseteq \mathbb{G}$. Consequently, $\mathbb{H}$ is a closed linear subspace of $\mathbb{G}$, so that we can write $\mathbb{G} = \mathbb{H} \oplus \mathbb{H}^\perp$. Let $g \in \mathbb{H}^\perp$ be arbitrary, and note that $\mathcal{K}(\cdot, x) \in \mathbb{H}$. By orthogonality, we must have $0 = \langle \mathcal{K}(\cdot, x), g \rangle_{\mathbb{G}} = g(x)$, from which we conclude that $\mathbb{H}^\perp = \{0\}$, and hence that $\mathbb{H} = \mathbb{G}$ as claimed. □

12.2.4 A more abstract viewpoint and further examples

Thus far, we have seen how any positive semidefinite kernel function can be used to build a Hilbert space in which the kernel satisfies the reproducing property (12.3). In the context of the Riesz representation theorem (Theorem 12.5), the reproducing property is equivalent to asserting that the function $\mathcal{K}(\cdot, x)$ acts as the representer for the *evaluation functional* at x—namely, the linear functional $L_x : \mathbb{H} \to \mathbb{R}$ that performs the operation $f \mapsto f(x)$. Thus, it shows that in any reproducing kernel Hilbert space, the evaluation functionals are all bounded. This perspective leads to the natural question: How large is the class of Hilbert spaces for which the evaluation functional is bounded? It turns out that this class is exactly equivalent to the class of reproducing kernel Hilbert spaces defined in the proof of Theorem 12.11. Indeed, an alternative way in which to define an RKHS is as follows:

Definition 12.12 A *reproducing kernel Hilbert space* $\mathbb{H}$ is a Hilbert space of real-valued functions on $\mathcal{X}$ such that for each $x \in \mathcal{X}$, the evaluation functional $L_x : \mathbb{H} \to \mathbb{R}$ is bounded (i.e., there exists some $M < \infty$ such that $|L_x(f)| \leq M \|f\|_{\mathbb{H}}$ for all $f \in \mathbb{H}$).

Theorem 12.11 shows that any PSD kernel can be used to define a reproducing kernel Hilbert space in the sense of Definition 12.12. In order to complete the equivalence, we need

to show that all Hilbert spaces specified by Definition 12.12 can be equipped with a reproducing kernel function. Let us state this claim formally, and then prove it:

Theorem 12.13 *Given any Hilbert space $\mathbb{H}$ in which the evaluation functionals are all bounded, there is a unique PSD kernel $\mathcal{K}$ that satisfies the reproducing property* (12.3).

Proof When L_x is a bounded linear functional, the Riesz representation (Theorem 12.5) implies that there must exist some element R_x of the Hilbert space $\mathbb{H}$ such that

$$f(x) = L_x(f) = \langle f, R_x \rangle_{\mathbb{H}} \qquad \text{for all } f \in \mathbb{H}. \tag{12.5}$$

Using these representers of evaluation, let us define a real-valued function $\mathcal{K}$ on the Cartesian product space $\mathcal{X} \times \mathcal{X}$ via $\mathcal{K}(x, z) := \langle R_x, R_z \rangle_{\mathbb{H}}$. Symmetry of the inner product ensures that $\mathcal{K}$ is a symmetric function, so that it remains to show that $\mathcal{K}$ is positive semidefinite. For any $n \geq 1$, let $\{x_i\}_{i=1}^n \subseteq \mathcal{X}$ be an arbitrary collection of points, and consider the $n \times n$ matrix $\mathbf{K}$ with elements $K_{ij} = \mathcal{K}(x_i, x_j)$. For an arbitrary vector $\alpha \in \mathbb{R}^n$, we have

$$\alpha^{\mathrm{T}} \mathbf{K} \alpha = \sum_{j,k=1}^{n} \alpha_j \alpha_k \mathcal{K}(x_j, x_k) = \left\langle \sum_{j=1}^{n} \alpha_j R_{x_j}, \sum_{j=1}^{n} \alpha_j R_{x_j} \right\rangle_{\mathbb{H}} = \| \sum_{j=1}^{n} \alpha_j R_{x_j} \|_{\mathbb{H}}^2 \geq 0,$$

which proves the positive semidefiniteness.

It remains to verify the reproducing property (12.3). It actually follows easily, since for any $x \in \mathcal{X}$, the function $\mathcal{K}(\cdot, x)$ is equivalent to $R_x(\cdot)$. In order to see this equivalence, note that for any $y \in \mathcal{X}$, we have

$$\mathcal{K}(y, x) \overset{\text{(i)}}{=} \langle R_y, R_x \rangle_{\mathbb{H}} \overset{\text{(ii)}}{=} R_x(y),$$

where step (i) follows from our original definition of the kernel function, and step (ii) follows since R_y is the representer of evaluation at y. It thus follows that our kernel satisfies the required reproducing property (12.3). Finally, in Exercise 12.4, we argue that the reproducing kernel of an RKHS must be unique. □

Let us consider some more examples to illustrate our different viewpoints on RKHSs.

Example 12.14 (Linear functions on $\mathbb{R}^d$) In Example 12.7, we showed that the linear kernel $\mathcal{K}(x, z) = \langle x, z \rangle$ is positive semidefinite on $\mathbb{R}^d$. The constructive proof of Theorem 12.11 dictates that the associated RKHS is generated by functions of the form

$$z \mapsto \sum_{i=1}^{n} \alpha_i \langle z, x_i \rangle = \left\langle z, \sum_{i=1}^{n} \alpha_i x_i \right\rangle.$$

Each such function is linear, and therefore the associated RKHS is the class of all linear functions—that is, functions of the form $f_\beta(\cdot) = \langle \cdot, \beta \rangle$ for some vector $\beta \in \mathbb{R}^m$. The induced inner product is given by $\left\langle f_\beta, f_{\widetilde{\beta}} \right\rangle_{\mathbb{H}} := \left\langle \beta, \widetilde{\beta} \right\rangle$. Note that for each $z \in \mathbb{R}^d$, the function

$\mathcal{K}(\cdot, z) = \langle \cdot, z \rangle \equiv f_z$ is linear. Moreover, for any linear function f_β, we have

$$\left\langle f_\beta, \mathcal{K}(\cdot, z) \right\rangle_{\mathbb{H}} = \langle \beta, z \rangle = f_\beta(z),$$

which provides an explicit verification of the reproducing property (12.3). ♣

Definition 12.12 and the associated Theorem 12.13 provide us with one avenue of verifying that a given Hilbert space is *not* an RKHS, and so cannot be equipped with a PSD kernel. In particular, the boundedness of the evaluation functionals R_x in an RKHS has a very important consequence: in particular, it ensures that convergence of a sequence of functions in an RKHS implies pointwise convergence. Indeed, if $f_n \to f^*$ in the Hilbert space norm, then for any $x \in \mathcal{X}$, we have

$$\left| f_n(x) - f^*(x) \right| = \left| \langle R_x, f_n - f^* \rangle_{\mathbb{H}} \right| \leq \|R_x\|_{\mathbb{H}} \|f_n - f^*\|_{\mathbb{H}} \to 0, \tag{12.6}$$

where we have applied the Cauchy–Schwarz inequality. This property is not shared by an arbitrary Hilbert space, with the Hilbert space $L^2[0, 1]$ from Example 12.4 being one case where this property fails.

Example 12.15 (The space $L^2[0, 1]$ is not an RKHS) From the argument above, it suffices to provide a sequence of functions $(f_n)_{n=1}^{\infty}$ that converge to the all-zero function in $L^2[0, 1]$, but do not converge to zero in a pointwise sense. Consider the sequence of functions $f_n(x) = x^n$ for $n = 1, 2, \ldots$. Since $\int_0^1 f_n^2(x)\, dx = \frac{1}{2n+1}$, this sequence is contained in $L^2[0, 1]$, and moreover $\|f_n\|_{L^2[0,1]} \to 0$. However, $f_n(1) = 1$ for all $n = 1, 2, \ldots$, so that this norm convergence does not imply pointwise convergence. Thus, if $L^2[0, 1]$ were an RKHS, then this would contradict inequality (12.6).

An alternative way to see that $L^2[0, 1]$ is not an RKHS is to ask whether it is possible to find a family of functions $\{R_x \in L^2[0, 1],\ x \in [0, 1]\}$ such that

$$\int_0^1 f(y) R_x(y)\, dy = f(x) \qquad \text{for all } f \in L^2[0, 1].$$

This identity will hold if we define R_x to be a "delta-function"—that is, infinite at x and zero elsewhere. However, such objects certainly do not belong to $L^2[0, 1]$, and exist only in the sense of generalized functions. ♣

Although $L^2[0, 1]$ itself is too large to be a reproducing kernel Hilbert space, we can obtain an RKHS by imposing further restrictions on our functions. One way to do so is by imposing constraints on functions and their derivatives. The *Sobolev spaces* form an important class that arise in this way: the following example describes a first-order Sobolev space that is an RKHS.

Example 12.16 (A simple Sobolev space) A function f over $[0, 1]$ is said to be *absolutely continuous* (or abs. cts. for short) if its derivative f' exists almost everywhere and is Lebesgue-integrable, and we have $f(x) = f(0) + \int_0^x f'(z)\, dz$ for all $x \in [0, 1]$. Now consider the set of functions

$$\mathbb{H}^1[0, 1] := \{f \colon [0, 1] \to \mathbb{R} \mid f(0) = 0, \text{ and } f \text{ is abs. cts. with } f' \in L^2[0, 1]\}. \tag{12.7}$$

Let us define an inner product on this space via $\langle f, g \rangle_{\mathbb{H}^1} := \int_0^1 f'(z) g'(z)\, dz$; we claim that the resulting Hilbert space is an RKHS.

One way to verify this claim is by exhibiting a representer of evaluation: for any $x \in [0, 1]$, consider the function $R_x(z) = \min\{x, z\}$. It is differentiable at every point $z \in [0, 1] \setminus \{x\}$, and we have $R'_x(z) = \mathbb{I}_{[0,x]}(z)$, corresponding to the binary-valued indicator function for membership in the interval $[0, x]$. Moreover, for any $z \in [0, 1]$, it is easy to verify that

$$\min\{x, z\} = \int_0^z \mathbb{I}_{[0,x]}(u)\, du, \tag{12.8}$$

so that R_x is absolutely continuous by definition. Since $R_x(0) = 0$, we conclude that R_x is an element of $\mathbb{H}^1[0, 1]$. Finally, to verify that R_x is the representer of evaluation, we calculate

$$\langle f, R_x \rangle_{\mathbb{H}^1} = \int_0^1 f'(z) R'_x(z)\, dz = \int_0^x f'(z)\, dz = f(x),$$

where the final equality uses the fundamental theorem of calculus.

As shown in the proof of Theorem 12.13, the function $\mathcal{K}(\cdot, x)$ is equivalent to the representer $R_x(\cdot)$. Thus, the kernel associated with the first-order Sobolev space on $[0, 1]$ is given by $\mathcal{K}(x, z) = R_x(z) = \min\{x, z\}$. To confirm that is positive semidefinite, note that equation (12.8) implies that

$$\mathcal{K}(x, z) = \int_0^1 \mathbb{I}_{[0,x]}(u) \mathbb{I}_{[0,z]}(u)\, du = \langle \mathbb{I}_{[0,x]}, \mathbb{I}_{[0,z]} \rangle_{L^2[0,1]},$$

thereby providing a Gram representation of the kernel that certifies its PSD nature. We conclude that $\mathcal{K}(x, z) = \min\{x, z\}$ is the unique positive semidefinite kernel function associated with this first-order Sobolev space. ♣

Let us now turn to some higher-order generalizations of the first-order Sobolev space from Example 12.16.

Example 12.17 (Higher-order Sobolev spaces and smoothing splines) For some fixed integer $\alpha \geq 1$, consider the class $\mathbb{H}^\alpha[0, 1]$ of real-valued functions on $[0, 1]$ that are α-times differentiable (almost everywhere), with the α-derivative $f^{(\alpha)}$ being Lebesgue-integrable, and such that $f(0) = f^{(1)}(0) = \cdots = f^{(\alpha-1)}(0) = 0$. (Here $f^{(k)}$ denotes the kth-order derivative of f.) We may define an inner product on this space via

$$\langle f, g \rangle_{\mathbb{H}} := \int_0^1 f^{(\alpha)}(z) g^{(\alpha)}(z)\, dz. \tag{12.9}$$

Note that this set-up generalizes Example 12.16, which corresponds to the case $\alpha = 1$.

We now claim that this inner product defines an RKHS, and more specifically, that the kernel is given by

$$\mathcal{K}(x, y) = \int_0^1 \frac{(x - z)_+^{\alpha-1}}{(\alpha - 1)!} \frac{(y - z)_+^{\alpha-1}}{(\alpha - 1)!}\, dz,$$

where $(t)_+ := \max\{0, t\}$. Note that the function $R_x(\cdot) := \mathcal{K}(\cdot, x)$ is α-times differentiable almost everywhere on $[0, 1]$ with $R_x^{(\alpha)}(y) = (x - y)_+^{\alpha-1}/(\alpha - 1)!$. To verify that R_x acts as the representer of evaluation, recall that any function $f : [0, 1] \to \mathbb{R}$ that is α-times differentiable

almost everywhere has the Taylor-series expansion

$$f(x) = \sum_{\ell=0}^{\alpha-1} f^{(\ell)}(0)\frac{x^\ell}{\ell!} + \int_0^1 f^{(\alpha)}(z)\frac{(x-z)_+^{\alpha-1}}{(\alpha-1)!}\,dz. \tag{12.10}$$

Using the previously mentioned properties of R_x and the definition (12.9) of the inner product, we obtain

$$\langle R_x,\, f\rangle_{\mathbb{H}} = \int_0^1 f^{(\alpha)}(z)\frac{(x-z)_+^{\alpha-1}}{(\alpha-1)!}\,dz = f(x),$$

where the final equality uses the Taylor-series expansion (12.10), and the fact that the first $(\alpha-1)$ derivatives of f vanish at 0.

In Example 12.29 to follow, we show how to augment the Hilbert space so as to remove the constraint on the first $(\alpha-1)$ derivatives of the functions f. ♣

12.3 Mercer's theorem and its consequences

We now turn to a useful representation of a broad class of positive semidefinite kernel functions, namely in terms of their eigenfunctions. Recall from classical linear algebra that any positive semidefinite matrix has an orthonormal basis of eigenvectors, and the associated eigenvalues are non-negative. The abstract version of Mercer's theorem generalizes this decomposition to positive semidefinite kernel functions.

Let $\mathbb{P}$ be a non-negative measure over a compact metric space $\mathcal{X}$, and consider the function class $L^2(\mathcal{X};\mathbb{P})$ with the usual squared norm

$$\|f\|^2_{L^2(\mathcal{X};\mathbb{P})} = \int_{\mathcal{X}} f^2(x)\,d\mathbb{P}(x).$$

Since the measure $\mathbb{P}$ remains fixed throughout, we frequently adopt the shorthand notation $L^2(\mathcal{X})$ or even just L^2 for this norm. Given a symmetric PSD kernel function $\mathcal{K}\colon \mathcal{X}\times\mathcal{X}\to\mathbb{R}$ that is continuous, we can define a linear operator $T_{\mathcal{K}}$ on $L^2(\mathcal{X})$ via

$$T_{\mathcal{K}}(f)(x) := \int_{\mathcal{X}} \mathcal{K}(x,\,z)f(z)\,d\mathbb{P}(z). \tag{12.11a}$$

We assume that the kernel function satisfies the inequality

$$\int_{\mathcal{X}\times\mathcal{X}} \mathcal{K}^2(x,\,z)\,d\mathbb{P}(x)\,d\mathbb{P}(z) < \infty, \tag{12.11b}$$

which ensures that $T_{\mathcal{K}}$ is a bounded linear operator on $L^2(\mathcal{X})$. Indeed, we have

$$\begin{aligned}\|T_{\mathcal{K}}(f)\|^2_{L^2(\mathcal{X})} &= \int_{\mathcal{X}}\Big(\int_{\mathcal{X}} \mathcal{K}(x,\,y)f(x)\,d\mathbb{P}(x)\Big)^2 d\mathbb{P}(y)\\ &\le \|f\|^2_{L^2(\mathcal{X})}\int_{\mathcal{X}\times\mathcal{X}} \mathcal{K}^2(x,y)\,d\mathbb{P}(x)\,d\mathbb{P}(y),\end{aligned}$$

where we have applied the Cauchy–Schwarz inequality. Operators of this type are known as *Hilbert–Schmidt operators*.

Let us illustrate these definitions with some examples.

Example 12.18 (PSD matrices) Let $\mathcal{X} = [d] := \{1, 2, \ldots, d\}$ be equipped with the Hamming metric, and let $\mathbb{P}(\{j\}) = 1$ for all $j \in \{1, 2, \ldots, d\}$ be the counting measure on this discrete space. In this case, any function $f : \mathcal{X} \to \mathbb{R}$ can be identified with the d-dimensional vector $(f(1), \ldots, f(d))$, and a symmetric kernel function $\mathcal{K} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ can be identified with the symmetric $d \times d$ matrix $\mathbf{K}$ with entries $K_{ij} = \mathcal{K}(i, j)$. Consequently, the integral operator (12.11a) reduces to ordinary matrix–vector multiplication

$$T_{\mathcal{K}}(f)(x) = \int_{\mathcal{X}} \mathcal{K}(x, z) f(z) \, d\mathbb{P}(z) = \sum_{z=1}^{d} \mathcal{K}(x, z) f(z).$$

By standard linear algebra, we know that the matrix $\mathbf{K}$ has an orthonormal collection of eigenvectors in $\mathbb{R}^d$, say $\{v_1, \ldots, v_d\}$, along with a set of non-negative eigenvalues $\mu_1 \geq \mu_2 \geq \cdots \geq \mu_d$, such that

$$\mathbf{K} = \sum_{j=1}^{d} \mu_j v_j v_j^{\mathrm{T}}. \tag{12.12}$$

Mercer's theorem, to be stated shortly, provides a substantial generalization of this decomposition to a general positive semidefinite kernel function. ♣

Example 12.19 (First-order Sobolev kernel) Now suppose that $\mathcal{X} = [0, 1]$, and that $\mathbb{P}$ is the Lebesgue measure. Recalling the kernel function $\mathcal{K}(x, z) = \min\{x, z\}$, we have

$$T_{\mathcal{K}}(f)(x) = \int_0^1 \min\{x, z\} f(z) \, dz = \int_0^x z f(z) \, dz + \int_x^1 x f(z) \, dz.$$

We return to analyze this particular integral operator in Example 12.23. ♣

Having gained some intuition for the general notion of a kernel integral operator, we are now ready for the statement of the abstract Mercer's theorem.

Theorem 12.20 (Mercer's theorem) *Suppose that $\mathcal{X}$ is compact, the kernel function $\mathcal{K}$ is continuous and positive semidefinite, and satisfies the Hilbert–Schmidt condition* (12.11b)*. Then there exist a sequence of eigenfunctions $(\phi_j)_{j=1}^{\infty}$ that form an orthonormal basis of $L^2(\mathcal{X}; \mathbb{P})$, and non-negative eigenvalues $(\mu_j)_{j=1}^{\infty}$ such that*

$$T_{\mathcal{K}}(\phi_j) = \mu_j \phi_j \qquad \text{for } j = 1, 2, \ldots. \tag{12.13a}$$

Moreover, the kernel function has the expansion

$$\mathcal{K}(x, z) = \sum_{j=1}^{\infty} \mu_j \phi_j(x) \phi_j(z), \tag{12.13b}$$

where the convergence of the infinite series holds absolutely and uniformly.

Remarks: The original theorem proved by Mercer applied only to operators defined on $L^2([a, b])$ for some finite $a < b$. The more abstract version stated here follows as a consequence of more general results on the eigenvalues of compact operators on Hilbert spaces; we refer the reader to the bibliography section for references.

Among other consequences, Mercer's theorem provides intuition on how reproducing kernel Hilbert spaces can be viewed as providing a particular embedding of the function domain $\mathcal{X}$ into a subset of the sequence space $\ell^2(\mathbb{N})$. In particular, given the eigenfunctions and eigenvalues guaranteed by Mercer's theorem, we may define a mapping $\Phi : \mathcal{X} \to \ell^2(\mathbb{N})$ via

$$x \mapsto \Phi(x) := \Big(\sqrt{\mu_1}\, \phi_1(x), \quad \sqrt{\mu_2}\, \phi_2(x), \quad \sqrt{\mu_3}\, \phi_3(x), \quad \ldots \Big). \tag{12.14}$$

By construction, we have

$$\|\Phi(x)\|^2_{\ell^2(\mathbb{N})} = \sum_{j=1}^{\infty} \mu_j \phi_j^2(x) = \mathcal{K}(x, x) < \infty,$$

showing that the map $x \mapsto \Phi(x)$ is a type of (weighted) feature map that embeds the original vector into a subset of $\ell^2(\mathbb{N})$. Moreover, this feature map also provides an explicit inner product representation of the kernel over $\ell^2(\mathbb{N})$—namely

$$\langle \Phi(x), \Phi(z) \rangle_{\ell^2(\mathbb{N})} = \sum_{j=1}^{\infty} \mu_j \, \phi_j(x) \, \phi_j(z) \;=\; \mathcal{K}(x, z).$$

Let us illustrate Mercer's theorem by considering some examples:

Example 12.21 (Eigenfunctions for a symmetric PSD matrix) As discussed in Example 12.18, a symmetric PSD d-dimensional matrix can be viewed as a kernel function on the space $[d] \times [d]$, where we adopt the shorthand $[d] := \{1, 2, \ldots, d\}$. In this case, the eigenfunction $\phi_j : [d] \to \mathbb{R}$ can be identified with the vector $v_j := (\phi_j(1), \ldots, \phi_j(d)) \in \mathbb{R}^d$. Thus, in this special case, the eigenvalue equation $T_{\mathcal{K}}(\phi_j) = \mu_j \phi_j$ is equivalent to asserting that $v_j \in \mathbb{R}^d$ is an eigenvector of the kernel matrix. Consequently, the decomposition (12.13b) then reduces to the familiar statement that any symmetric PSD matrix has an orthonormal basis of eigenfunctions, with associated non-negative eigenvalues, as previously stated in equation (12.12). ♣

Example 12.22 (Eigenfunctions of a polynomial kernel) Let us compute the eigenfunctions of the second-order polynomial kernel $\mathcal{K}(x, z) = (1 + xz)^2$ defined over the Cartesian product $[-1, 1] \times [-1, 1]$, where the unit interval is equipped with the Lebesgue measure. Given a function $f : [-1, 1] \to \mathbb{R}$, we have

$$\begin{aligned} \int_{-1}^{1} \mathcal{K}(x, z) f(z)\, dz &= \int_{-1}^{1} \Big(1 + 2xz + x^2 z^2\Big) f(z)\, dz \\ &= \left\{ \int_{-1}^{1} f(z)\, dz \right\} + \left\{ 2 \int_{-1}^{1} z f(z)\, dz \right\} x + \left\{ \int_{-1}^{1} z^2 f(z)\, dz \right\} x^2, \end{aligned}$$

showing that any eigenfunction of the kernel integral operator must be a polynomial of degree at most two. Consequently, the eigenfunction problem can be reduced to an ordinary

eigenvalue problem in terms of the coefficients in the expansion $f(x) = a_0 + a_1 x + a_2 x^2$. Following some simple algebra, we find that, if f is an eigenfunction with eigenvalue μ, then these coefficients must satisfy the linear system

$$\begin{bmatrix} 2 & 0 & 2/3 \\ 0 & 4/3 & 0 \\ 2/3 & 0 & 2/5 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \end{bmatrix} = \mu \begin{bmatrix} a_0 \\ a_1 \\ a_2 \end{bmatrix}.$$

Solving this ordinary eigensystem, we find the following eigenfunction–eigenvalue pairs

$$\begin{aligned} \phi_1(x) &= -0.9403 - 0.3404x^2, && \text{with } \mu_1 = 2.2414, \\ \phi_2(x) &= x, && \text{with } \mu_2 = 1.3333, \\ \phi_3(x) &= -0.3404 + 0.9403x^2, && \text{with } \mu_3 = 0.1586. \end{aligned}$$ ♣

Example 12.23 (Eigenfunctions for a first-order Sobolev space) In Example 12.16, we introduced the first-order Sobolev space $\mathbb{H}^1[0, 1]$. In Example 12.19, we found that its kernel function takes the form $\mathcal{K}(x, z) = \min\{x, z\}$, and determined the form of the associated integral operator. Using this previous development, if $\phi\colon [0, 1] \to \mathbb{R}$ is an eigenfunction of $T_{\mathcal{K}}$ with eigenvalue $\mu \neq 0$, then it must satisfy the relation $T_{\mathcal{K}}(\phi) = \mu\phi$, or equivalently

$$\int_0^x z\phi(z)\, dz + \int_x^1 x\phi(z)\, dz = \mu\phi(x) \qquad \text{for all } x \in [0, 1].$$

Since this relation must hold for all $x \in [0, 1]$, we may take derivatives with respect to x. Doing so twice yields the second-order differential equation $\mu\phi''(x) + \phi(x) = 0$. Combined with the boundary condition $\phi(0) = 0$, we obtain $\phi(x) = \sin(x/\sqrt{\mu})$ as potential eigenfunctions. Now using the boundary condition $\int_0^1 z\phi(z)\, dz = \mu\phi(1)$, we deduce that the eigenfunction–eigenvalue pairs are given by

$$\phi_j(t) = \sin\frac{(2j-1)\pi t}{2} \quad \text{and} \quad \mu_j = \left(\frac{2}{(2j-1)\,\pi}\right)^2 \qquad \text{for } j = 1, 2, \ldots.$$ ♣

Example 12.24 (Translation-invariant kernels) An important class of kernels have a translation-invariant form. In particular, given a function $\psi\colon [-1, 1] \to \mathbb{R}$ that is even (meaning that $\psi(u) = \psi(-u)$ for all $u \in [-1, 1]$), let us extend its domain to the real line by the periodic extension $\psi(u + 2k) = \psi(u)$ for all $u \in [-1, 1]$ and integers $k \in \mathbb{Z}$.

Using this function, we may define a *translation-invariant* kernel on the Cartesian product space $[-1, 1] \times [-1, 1]$ via $\mathcal{K}(x, z) = \psi(x - z)$. Note that the evenness of ψ ensures that this kernel is symmetric. Moreover, the kernel integral operator takes the form

$$T_{\mathcal{K}}(f)(x) = \underbrace{\int_{-1}^1 \psi(x - z) f(z)\, dz}_{(\psi * f)(x)},$$

and thus is a convolution operator.

A classical result from analysis is that the eigenfunctions of convolution operators are given by the Fourier basis; let us prove this fact here. We first show that the cosine functions

$\phi_j(x) = \cos(\pi j x)$ for $j = 0, 1, 2, \ldots$ are eigenfunctions of the operator $T_{\mathcal{K}}$. Indeed, we have

$$T_{\mathcal{K}}(\phi_j)(x) = \int_{-1}^{1} \psi(x - z)\cos(\pi j z)\, dz = \int_{-1-x}^{1-x} \psi(-u)\cos(2\pi j(x + u))\, du,$$

where we have made the change of variable $u = z - x$. Note that the interval of integration $[-1 - x,\ 1 - x]$ is of length 2, and since both $\psi(-u)$ and $\cos(2\pi(x + u))$ have period 2, we can shift the interval of integration to $[-1, 1]$. Combined with the evenness of ψ, we conclude that $T_{\mathcal{K}}(\phi_j)(x) = \int_{-1}^{1} \psi(u)\cos(2\pi j(x + u))\, du$. Using the elementary trigonometric identity

$$\cos(\pi j(x + u)) = \cos(\pi j x)\cos(\pi j u) - \sin(\pi j x)\sin(\pi j u),$$

we find that

$$\begin{aligned} T_{\mathcal{K}}(\phi_j)(x) &= \left\{\int_{-1}^{1} \psi(u)\cos(\pi j u)\, du\right\}\cos(\pi j x) - \left\{\int_{-1}^{1} \psi(u)\sin(\pi j u)\, du\right\}\sin(\pi j x) \\ &= c_j \cos(\pi j x), \end{aligned}$$

where $c_j = \int_{-1}^{1} \psi(u)\cos(\pi j u)\, du$ is the jth cosine coefficient of ψ. In this calculation, we have used the evenness of ψ to argue that the integral with the sine function vanishes.

A similar argument shows that each of the sinusoids

$$\widetilde{\phi}_j(x) = \sin(j\pi x) \qquad \text{for } j = 1, 2, \ldots$$

are also eigenfunctions with eigenvalue c_j. Since the functions $\{\phi_j,\ j = 0, 1, 2, \ldots\} \cup \{\widetilde{\phi}_j,\ j = 1, 2, \ldots\}$ form a complete orthogonal basis of $L^2[-1, 1]$, there are no other eigenfunctions that are not linear combinations of these functions. Consequently, by Mercer's theorem, the kernel function has the eigenexpansion

$$\mathcal{K}(x, z) = \sum_{j=0}^{\infty} c_j\{\cos(\pi j x)\cos(\pi j z) + \sin(\pi j x)\sin(\pi j z)\} = \sum_{j=0}^{\infty} c_j \cos(\pi j(x - z)),$$

where c_j are the (cosine) Fourier coefficients of ψ. Thus, we see that $\mathcal{K}$ is positive semidefinite if and only if $c_j \geq 0$ for $j = 0, 1, 2, \ldots$. ♣

Example 12.25 (Gaussian kernel) As previously introduced in Example 12.9, a popular choice of kernel on some subset $\mathcal{X} \subseteq \mathbb{R}^d$ is the Gaussian kernel given by $\mathcal{K}(x, z) = \exp(-\frac{\|x-z\|_2^2}{2\sigma^2})$, where $\sigma > 0$ is a bandwidth parameter. To keep our calculations relatively simple, let us focus here on the univariate case $d = 1$, and let $\mathcal{X}$ be some compact interval of the real line. By a rescaling argument, we can restrict ourselves to the case $\mathcal{X} = [-1, 1]$, so that we are considering solutions to the integral equation

$$\int_{-1}^{1} e^{-\frac{(x-z)^2}{2\sigma^2}} \phi_j(z)\, dz = \mu_j \phi_j(x). \tag{12.15}$$

Note that this problem cannot be tackled by the methods of the previous example, since we are *not* performing the periodic extension of our function.[2] Nonetheless, the eigenvalues of the Gaussian integral operator are very closely related to the Fourier transform.

[2] If we were to consider the periodically extended version, then the eigenvalues would be given by the cosine coefficients $c_j = \int_{-1}^{1} \exp(-\frac{u^2}{2\sigma^2})\cos(\pi j u)\, du$, with the cosine functions as eigenfunctions.

In the remainder of our development, let us consider a slightly more general integral equation. Given a bounded, continuous and even function $\Psi\colon \mathbb{R} \to [0, \infty)$, we may define its (real-valued) Fourier transform $\psi(u) = \int_{-\infty}^{\infty} \Psi(\omega) e^{-i\omega u}\, d\omega$, and use it to define a translation-invariant kernel via $\mathcal{K}(x, z) := \psi(x - z)$. We are then led to the integral equation

$$\int_{-1}^{1} \psi(x - z)\phi_j(z)\, dz = \mu_j \phi_j(x). \tag{12.16}$$

Classical theory on integral operators can be used to characterize the spectrum of this integral operator. More precisely, for any operator such that $\log \Psi(\omega) \asymp -\omega^{\alpha}$ for some $\alpha > 1$, there is a constant c such that the eigenvalues $(\mu_j)_{j=1}^{\infty}$ associated with the integral equation (12.16) scale as $\mu_j \asymp e^{-cj \log j}$ as $j \to +\infty$. See the bibliographic section for further discussion of results of this type.

The Gaussian kernel is a special case of this set-up with the pair $\Psi(\omega) = \exp(-\frac{\sigma^2\omega^2}{2})$ and $\psi(u) = \exp(-\frac{u^2}{2\sigma^2})$. Applying the previous reasoning guarantees that the eigenvalues of the Gaussian kernel over a compact interval scale as $\mu_j \asymp \exp(-cj \log j)$ as $j \to +\infty$. We thus see that the Gaussian kernel class is relatively small, since its eigenvalues decay at exponential rate. (The reader should contrast this fast decay with the significantly slower $\mu_j \asymp j^{-2}$ decay rate of the first-order Sobolev class from Example 12.23.) ♣

An interesting consequence of Mercer's theorem is in giving a relatively explicit characterization of the RKHS associated with a given kernel.

Corollary 12.26 *Consider a kernel satisfying the conditions of Mercer's theorem with associated eigenfunctions $(\phi_j)_{j=1}^{\infty}$ and non-negative eigenvalues $(\mu_j)_{j=1}^{\infty}$. It induces the reproducing kernel Hilbert space*

$$\mathbb{H} := \left\{ f = \sum_{j=1}^{\infty} \beta_j \phi_j \mid \text{for some } (\beta_j)_{j=1}^{\infty} \in \ell^2(\mathbb{N}) \text{ with } \sum_{j=1}^{\infty} \frac{\beta_j^2}{\mu_j} < \infty \right\}, \tag{12.17a}$$

along with inner product

$$\langle f, g \rangle_{\mathbb{H}} := \sum_{j=1}^{\infty} \frac{\langle f, \phi_j \rangle \langle g, \phi_j \rangle}{\mu_j}, \tag{12.17b}$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product in $L^2(\mathcal{X}; \mathbb{P})$.

Let us make a few comments on this claim. First, in order to assuage any concerns regarding division by zero, we can restrict all sums to only indices j for which $\mu_j > 0$. Second, note that Corollary 12.26 shows that the RKHS associated with a Mercer kernel is isomorphic to an infinite-dimensional ellipsoid contained with $\ell^2(\mathbb{N})$—namely, the set

$$\mathcal{E} := \left\{ (\beta_j)_{j=1}^{\infty} \in \ell^2(\mathbb{N}) \mid \sum_{j=1}^{\infty} \frac{\beta_j^2}{\mu_j} \leq 1 \right\}. \tag{12.18}$$

We study the properties of such ellipsoids at more length in Chapters 13 and 14.

Proof For the proof, we take $\mu_j > 0$ for all $j \in \mathbb{N}$. This assumption entails no loss of generality, since otherwise the same argument can be applied with relevant summations truncated to the positive eigenvalues of the kernel function. Recall that $\langle \cdot, \cdot \rangle$ denotes the inner product on $L^2(\mathcal{X}; \mathbb{P})$.

It is straightforward to verify that $\mathbb{H}$ along with the specified inner product $\langle \cdot, \cdot \rangle_{\mathbb{H}}$ is a Hilbert space. Our next step is to show that $\mathbb{H}$ is in fact a reproducing kernel Hilbert space, and satisfies the reproducing property with respect to the given kernel. We begin by showing that for each fixed $x \in \mathcal{X}$, the function $\mathcal{K}(\cdot, x)$ belongs to $\mathbb{H}$. By the Mercer expansion, we have $\mathcal{K}(\cdot, x) = \sum_{j=1}^{\infty} \mu_j \phi_j(x) \phi_j(\cdot)$, so that by definition (12.17a) of our Hilbert space, it suffices to show that $\sum_{j=1}^{\infty} \mu_j \phi_j^2(x) < \infty$. By the Mercer expansion, we have

$$\sum_{j=1}^{\infty} \mu_j \phi_j^2(x) = \mathcal{K}(x, x) < \infty,$$

so that $\mathcal{K}(\cdot, x) \in \mathbb{H}$.

Let us now verify the reproducing property. By the orthonormality of $(\phi_j)_{j=1}^{\infty}$ in $L^2(\mathcal{X}; \mathbb{P})$ and Mercer's theorem, we have $\langle \mathcal{K}(\cdot, x), \phi_j \rangle = \mu_j \phi_j(x)$ for each $j \in \mathbb{N}$. Thus, by definition (12.17b) of our Hilbert inner product, for any $f \in \mathbb{H}$, we have

$$\langle f, \mathcal{K}(\cdot, x) \rangle_{\mathbb{H}} = \sum_{j=1}^{\infty} \frac{\langle f, \phi_j \rangle \langle \mathcal{K}(\cdot, x), \phi_j \rangle}{\mu_j} = \sum_{j=1}^{\infty} \langle f, \phi_j \rangle \phi_j(x) = f(x),$$

where the final step again uses the orthonormality of $(\phi_j)_{j=1}^{\infty}$. Thus, we have shown that $\mathbb{H}$ is the RKHS with kernel $\mathcal{K}$. (As discussed in Theorem 12.11, the RKHS associated with any given kernel is unique.) □

12.4 Operations on reproducing kernel Hilbert spaces

In this section, we describe a number of operations on reproducing kernel Hilbert spaces that allow us to build new spaces.

12.4.1 Sums of reproducing kernels

Given two Hilbert spaces $\mathbb{H}_1$ and $\mathbb{H}_2$ of functions defined on domains $\mathcal{X}_1$ and $\mathcal{X}_2$, respectively, consider the space

$$\mathbb{H}_1 + \mathbb{H}_2 := \Big\{ f_1 + f_2 \mid f_j \in \mathbb{H}_j, \ j = 1, 2 \Big\},$$

corresponding to the set of all functions obtained as sums of pairs of functions from the two spaces.

Proposition 12.27 *Suppose that $\mathbb{H}_1$ and $\mathbb{H}_2$ are both RKHSs with kernels $\mathcal{K}_1$ and $\mathcal{K}_2$, respectively. Then the space $\mathbb{H} = \mathbb{H}_1 + \mathbb{H}_2$ with norm*

$$\|f\|_{\mathbb{H}}^2 := \min_{\substack{f=f_1+f_2 \\ f_1\in\mathbb{H}_1,\, f_2\in\mathbb{H}_2}} \{\|f_1\|_{\mathbb{H}_1}^2 + \|f_2\|_{\mathbb{H}_2}^2\} \tag{12.19}$$

is an RKHS with kernel $\mathcal{K} = \mathcal{K}_1 + \mathcal{K}_2$.

Remark: This construction is particularly simple when $\mathbb{H}_1$ and $\mathbb{H}_2$ share only the constant zero function, since any function $f \in \mathbb{H}$ can then be written as $f = f_1 + f_2$ for a unique pair (f_1, f_2), and hence $\|f\|_{\mathbb{H}}^2 = \|f_1\|_{\mathbb{H}_1}^2 + \|f_2\|_{\mathbb{H}_2}^2$. Let us illustrate the use of summation with some examples:

Example 12.28 (First-order Sobolev space and constant functions) Consider the kernel functions on $[0,1] \times [0,1]$ given by $\mathcal{K}_1(x,z) = 1$ and $\mathcal{K}_2(x,z) = \min\{x,z\}$. They generate the reproducing kernel Hilbert spaces

$$\mathbb{H}_1 = \text{span}\{1\} \quad \text{and} \quad \mathbb{H}_2 = \mathbb{H}^1[0,1],$$

where $\text{span}\{1\}$ is the set of all constant functions, and $\mathbb{H}^1[0,1]$ is the first-order Sobolev space from Example 12.16. Note that $\mathbb{H}_1 \cap \mathbb{H}_2 = \{0\}$, since $f(0) = 0$ for any element of $\mathbb{H}_2$. Consequently, the RKHS with kernel $\mathcal{K}(x,z) = 1 + \min\{x,z\}$ consists of all functions

$$\bar{\mathbb{H}}^1[0,1] := \{f\colon [0,1] \to \mathbb{R} \mid f \text{ is absolutely continuous with } f' \in L^2[0,1]\},$$

equipped with the squared norm $\|f\|_{\bar{\mathbb{H}}^1[0,1]}^2 = f^2(0) + \int_0^1 (f'(z))^2\, dz$. ♣

As a continuation of the previous example, let us describe an extension of the higher-order Sobolev spaces from Example 12.17:

Example 12.29 (Extending higher-order Sobolev spaces) For an integer $\alpha \geq 1$, consider the kernel functions on $[0,1] \times [0,1]$ given by

$$\mathcal{K}_1(x,z) = \sum_{\ell=0}^{\alpha-1} \frac{x^\ell}{\ell!}\frac{z^\ell}{\ell!} \quad \text{and} \quad \mathcal{K}_2(x,z) = \int_0^1 \frac{(x-y)_+^{\alpha-1}}{(\alpha-1)!}\frac{(z-y)_+^{\alpha-1}}{(\alpha-1)!}\, dy.$$

The first kernel generates an RKHS $\mathbb{H}_1$ of polynomials of degree $\alpha - 1$, whereas the second kernel generates the α-order Sobolev space $\mathbb{H}_2 = \mathbb{H}^\alpha[0,1]$ previously defined in Example 12.17.

Letting $f^{(\ell)}$ denote the ℓth-order derivative, recall that any function $f \in \mathbb{H}^\alpha[0,1]$ satisfies the boundary conditions $f^{(\ell)}(0) = 0$ for $\ell = 0, 1, \ldots, \alpha - 1$. Consequently, we have $\mathbb{H}_1 \cap \mathbb{H}_2 = \{0\}$ so that Proposition 12.27 guarantees that the kernel

$$\mathcal{K}(x,z) = \sum_{\ell=0}^{\alpha-1} \frac{x^\ell}{\ell!}\frac{z^\ell}{\ell!} + \int_0^1 \frac{(x-y)_+^{\alpha-1}}{(\alpha-1)!}\frac{(z-y)_+^{\alpha-1}}{(\alpha-1)!}\, dy \tag{12.20}$$

generates the Hilbert space $\bar{\mathbb{H}}^\alpha[0,1]$ of all functions that are α-times differentiable almost

everywhere, with $f^{(\alpha)}$ Lebesgue-integrable. As we verify in Exercise 12.15, the associated RKHS norm takes the form

$$\|f\|_{\mathbb{H}}^2 = \sum_{\ell=0}^{\alpha-1} \left(f^{(\ell)}(0)\right)^2 + \int_0^1 \left(f^{(\alpha)}(z)\right)^2 dz. \tag{12.21}$$

♣

Example 12.30 (Additive models) It is often convenient to build up a multivariate function from simpler pieces, and additive models provide one way in which to do so. For $j = 1, 2, \ldots, M$, let $\mathbb{H}_j$ be a reproducing kernel Hilbert space, and let us consider functions that have an additive decomposition of the form $f = \sum_{j=1}^M f_j$, where $f_j \in \mathbb{H}_j$. By Proposition 12.27, the space $\mathbb{H}$ of all such functions is itself an RKHS equipped with the kernel function $\mathcal{K} = \sum_{j=1}^M \mathcal{K}_j$. A commonly used instance of such an additive model is when the individual Hilbert space $\mathbb{H}_j$ corresponds to functions of the jth coordinate of a d-dimensional vector, so that the space $\mathbb{H}$ consists of functions $f: \mathbb{R}^d \to \mathbb{R}$ that have the additive decomposition

$$f(x_1, \ldots, x_d) = \sum_{j=1}^d f_j(x_j),$$

where $f_j: \mathbb{R} \to \mathbb{R}$ is a univariate function for the jth coordinate. Since $\mathbb{H}_j \cap \mathbb{H}_k = \{0\}$ for all $j \neq k$, the associated Hilbert norm takes the form $\|f\|_{\mathbb{H}}^2 = \sum_{j=1}^d \|f_j\|_{\mathbb{H}_j}^2$. We provide some additional discussion of these additive decompositions in Exercise 13.9 and Example 14.11 to follow in later chapters.

More generally, it is natural to consider expansions of the form

$$f(x_1, \ldots, x_d) = \sum_{j=1}^d f_j(x_j) + \sum_{j \neq k} f_{jk}(x_j, x_k) + \cdots.$$

When the expansion functions are chosen to be mutually orthogonal, such expansions are known as *functional ANOVA* decompositions. ♣

We now turn to the proof of Proposition 12.27.

Proof Consider the direct sum $\mathbb{F} := \mathbb{H}_1 \oplus \mathbb{H}_2$ of the two Hilbert spaces; by definition, it is the Hilbert space $\{(f_1, f_2) \mid f_j \in \mathbb{H}_j, j = 1, 2\}$ of all ordered pairs, along with the norm

$$\|(f_1, f_2)\|_{\mathbb{F}}^2 := \|f_1\|_{\mathbb{H}_1}^2 + \|f_2\|_{\mathbb{H}_2}^2. \tag{12.22}$$

Now consider the linear operator $L: \mathbb{F} \to \mathbb{H}$ defined by $(f_1, f_2) \mapsto f_1 + f_2$, and note that it maps $\mathbb{F}$ onto $\mathbb{H}$. The nullspace $\mathbb{N}(L)$ of this operator is a subspace of $\mathbb{F}$, and we claim that it is closed. Consider some sequence $((f_n, -f_n))_{n=1}^\infty$ contained within the nullspace $\mathbb{N}(L)$ that converges to a point $(f, g) \in \mathbb{F}$. By the definition of the norm (12.22), this convergence implies that $f_n \to f$ in $\mathbb{H}_1$ (and hence pointwise) and $-f_n \to g$ in $\mathbb{H}_2$ (and hence pointwise). Overall, we conclude that $f = -g$, meaning $(f, g) \in \mathbb{N}(L)$.

Let $\mathbb{N}^\perp$ be the orthogonal complement of $\mathbb{N}(L)$ in $\mathbb{F}$, and let $L_\perp$ be the restriction of L to

$\mathbb{N}^\perp$. Since this map is a bijection between $\mathbb{N}^\perp$ and $\mathbb{H}$, we may define an inner product on $\mathbb{H}$ via

$$\langle f, g\rangle_{\mathbb{H}} := \langle L_\perp^{-1}(f), L_\perp^{-1}(g)\rangle_{\mathbb{F}}.$$

It can be verified that the space $\mathbb{H}$ with this inner product is a Hilbert space.

It remains to check that $\mathbb{H}$ is an RKHS with kernel $\mathcal{K} = \mathcal{K}_1 + \mathcal{K}_2$, and that the norm $\|\cdot\|_{\mathbb{H}}^2$ takes the given form (12.19). Since the functions $\mathcal{K}_1(\cdot, x)$ and $\mathcal{K}_2(\cdot, x)$ belong to $\mathbb{H}_1$ and $\mathbb{H}_2$, respectively, the function $\mathcal{K}(\cdot, x) = \mathcal{K}_1(\cdot, x) + \mathcal{K}_2(\cdot, x)$ belongs to $\mathbb{H}$. For a fixed $f \in \mathbb{F}$, let $(f_1, f_2) = L_\perp^{-1}(f) \in \mathbb{F}$, and for a fixed $x \in \mathcal{X}$, let $(g_1, g_2) = L_\perp^{-1}(\mathcal{K}(\cdot, x)) \in \mathbb{F}$. Since $(g_1 - \mathcal{K}_1(\cdot, x),\ g_2 - \mathcal{K}_2(\cdot, x))$ must belong to $\mathbb{N}(L)$, it must be orthogonal (in $\mathbb{F}$) to the element $(f_1, f_2) \in \mathbb{N}^\perp$. Consequently, we have $\langle (g_1 - \mathcal{K}_1(\cdot, x),\ g_2 - \mathcal{K}_2(\cdot, x)),\ (f_1, f_2)\rangle_{\mathbb{F}} = 0$, and hence

$$\begin{aligned}\langle f_1, \mathcal{K}_1(\cdot, x)\rangle_{\mathbb{H}_1} + \langle f_2, \mathcal{K}_2(\cdot, x)\rangle_{\mathbb{H}_2} &= \langle f_1, g_1\rangle_{\mathbb{H}_1} + \langle f_2, g_2\rangle_{\mathbb{H}_2} \\ &= \langle f, \mathcal{K}(\cdot, x)\rangle_{\mathbb{H}}.\end{aligned}$$

Since $\langle f_1, \mathcal{K}_1(\cdot, x)\rangle_{\mathbb{H}_1}\, \langle f_2, \mathcal{K}_2(\cdot, x)\rangle_{\mathbb{H}_2} = f_1(x) + f_2(x) = f(x)$, we have established that $\mathcal{K}$ has the reproducing property.

Finally, let us verify that the norm $\|f\|_{\mathbb{H}} := \|L_\perp^{-1}(f)\|_{\mathbb{F}}$ that we have defined is equivalent to the definition (12.19). For a given $f \in \mathbb{H}$, consider some pair $(f_1, f_2) \in \mathbb{F}$ such that $f = f_1 + f_2$, and define $(v_1, v_2) = (f_1, f_2) - L_\perp^{-1}(f)$. We have

$$\|f_1\|_{\mathbb{H}_1}^2 + \|f_2\|_{\mathbb{H}_2}^2 \overset{\text{(i)}}{=} \|(f_1, f_2)\|_{\mathbb{F}}^2 \overset{\text{(ii)}}{=} \|(v_1, v_2)\|_{\mathbb{F}}^2 + \|L_\perp^{-1}(f)\|_{\mathbb{F}}^2 \overset{\text{(iii)}}{=} \|(v_1, v_2)\|_{\mathbb{F}}^2 + \|f\|_{\mathbb{H}}^2,$$

where step (i) uses the definition (12.22) of the norm in $\mathbb{F}$, step (ii) follows from the Pythagorean property, as applied to the pair $(v_1, v_2) \in \mathbb{N}(L)$ and $L_\perp^{-1}(f) \in \mathbb{N}^\perp$, and step (iii) uses our definition of the norm $\|f\|_{\mathbb{H}}$. Consequently, we have shown that for any pair f_1, f_2 such that $f = f_1 + f_2$, we have

$$\|f\|_{\mathbb{H}}^2 \leq \|f_1\|_{\mathbb{H}_1}^2 + \|f_2\|_{\mathbb{H}_2}^2,$$

with equality holding if and only if $(v_1, v_2) = (0, 0)$, or equivalently $(f_1, f_2) = L_\perp^{-1}(f)$. This establishes the equivalence of the definitions. □

12.4.2 Tensor products

Consider two separable Hilbert spaces $\mathbb{H}_1$ and $\mathbb{H}_2$ of functions, say with domains $\mathcal{X}_1$ and $\mathcal{X}_2$, respectively. They can be used to define a new Hilbert space, denoted by $\mathbb{H}_1 \otimes \mathbb{H}_2$, known as the tensor product of $\mathbb{H}_1$ and $\mathbb{H}_2$. Consider the set of functions $h\colon \mathcal{X}_1 \times \mathcal{X}_2 \to \mathbb{R}$ that have the form

$$\{h = \sum_{j=1}^{n} f_j g_j \mid \text{for some } n \in \mathbb{N} \text{ and such that } f_j \in \mathbb{H}_1,\ g_j \in \mathbb{H}_2 \text{ for all } j \in [n]\}.$$

If $h = \sum_{j=1}^n f_j g_j$ and $\widetilde{h} = \sum_{k=1}^m \widetilde{f}_k \widetilde{g}_k$ are two members of this set, we define their inner product

$$\langle h, \widetilde{h}\rangle_{\mathbb{H}} := \sum_{j=1}^{n}\sum_{k=1}^{m} \langle f_j, \widetilde{f}_k\rangle_{\mathbb{H}_1}\, \langle g_j, \widetilde{g}_k\rangle_{\mathbb{H}_2}. \tag{12.23}$$

Note that the value of the inner product depends neither on the chosen representation of h nor on that of $\widetilde{h}$; indeed, using linearity of the inner product, we have

$$\langle h, \widetilde{h}\rangle_{\mathbb{H}} = \sum_{k=1}^{m} \langle (h \odot \widetilde{f_k}), \widetilde{g}_k\rangle_{\mathbb{H}_2},$$

where $(h \odot \widetilde{f_k}) \in \mathbb{H}_2$ is the function given by $x_2 \mapsto \langle h(\cdot, x_2), \widetilde{f_k}\rangle_{\mathbb{H}_1}$. A similar argument shows that the inner product does not depend on the representation of $\widetilde{h}$, so that the inner product (12.23) is well defined.

It is straightforward to check that the inner product (12.23) is bilinear and symmetric, and that $\langle h, h\rangle_{\mathbb{H}}^2 = \|h\|_{\mathbb{H}}^2 \geq 0$ for all $h \in \mathbb{H}$. It remains to check that $\|h\|_{\mathbb{H}} = 0$ if and only if $h = 0$. Consider some $h \in \mathbb{H}$ with the representation $h = \sum_{j=1}^{n} f_j g_j$. Let $(\phi_j)_{j=1}^{\infty}$ and $(\psi_k)_{k=1}^{\infty}$ be complete orthonormal bases of $\mathbb{H}_1$ and $\mathbb{H}_2$, respectively, ordered such that

$$\text{span}\{f_1, \ldots, f_n\} \subseteq \text{span}\{\phi_1, \ldots, \phi_n\} \quad \text{and} \quad \text{span}\{g_1, \ldots, g_n\} \subseteq \text{span}\{\psi_1, \ldots, \psi_n\}.$$

Consequently, we can write f equivalently as the double summation $f = \sum_{j,j=1}^{n} \alpha_{j,k}\phi_j\psi_k$ for some set of real numbers $\{\alpha_{j,k}\}_{j,k=1}^{n}$. Using this representation, we are guaranteed the equality $\|f\|_{\mathbb{H}}^2 = \sum_{j=1}^{n}\sum_{k=1}^{n} \alpha_{j,k}^2$, which shows that $\|f\|_{\mathbb{H}} = 0$ if and only if $\alpha_{j,k} = 0$ for all (j, k), or equivalently $f = 0$.

In this way, we have defined the tensor product $\mathbb{H} = \mathbb{H}_1 \otimes \mathbb{H}_2$ of two Hilbert spaces. The next result asserts that when the two component spaces have reproducing kernels, then the tensor product space is also a reproducing kernel Hilbert space:

Proposition 12.31 *Suppose that $\mathbb{H}_1$ and $\mathbb{H}_2$ are reproducing kernel Hilbert spaces of real-valued functions with domains $\mathcal{X}_1$ and $\mathcal{X}_2$, and equipped with kernels $\mathcal{K}_1$ and $\mathcal{K}_2$, respectively. Then the tensor product space $\mathbb{H} = \mathbb{H}_1 \otimes \mathbb{H}_2$ is an RKHS of real-valued functions with domain $\mathcal{X}_1 \times \mathcal{X}_2$, and with kernel function*

$$\mathcal{K}((x_1, x_2), (x_1', x_2')) = \mathcal{K}_1(x_1, x_1')\,\mathcal{K}_2(x_2, x_2'). \tag{12.24}$$

Proof In Exercise 12.16, it is shown that $\mathcal{K}$ defined in equation (12.24) is a positive semidefinite function. By definition of the tensor product space $\mathbb{H} = \mathbb{H}_1 \otimes \mathbb{H}_2$, for each pair $(x_1, x_2) \in \mathcal{X}_1 \times \mathcal{X}_2$, the function $\mathcal{K}((\cdot, \cdot), (x_1, x_2)) = \mathcal{K}_1(\cdot, x_1)\mathcal{K}_2(\cdot, x_2)$ is an element of the tensor product space $\mathbb{H}$. Let $f = \sum_{j,k=1}^{n} \alpha_{j,k}\phi_j\psi_k$ be an arbitrary element of $\mathbb{H}$. By definition of the inner product (12.23), we have

$$\begin{aligned}\langle f, \mathcal{K}((\cdot, \cdot), (x_1, x_2))\rangle_{\mathbb{H}} &= \sum_{j,k=1}^{n} \alpha_{j,k}\langle \phi_j, \mathcal{K}_1(\cdot, x_1)\rangle_{\mathbb{H}_1}\, \langle \psi_k, \mathcal{K}_2(\cdot, x_2)\rangle_{\mathbb{H}_2} \\ &= \sum_{j,k=1}^{n} \alpha_{j,k}\phi_j(x_1)\psi_k(x_2) = f(x_1, x_2),\end{aligned}$$

thereby verifying the reproducing property. □

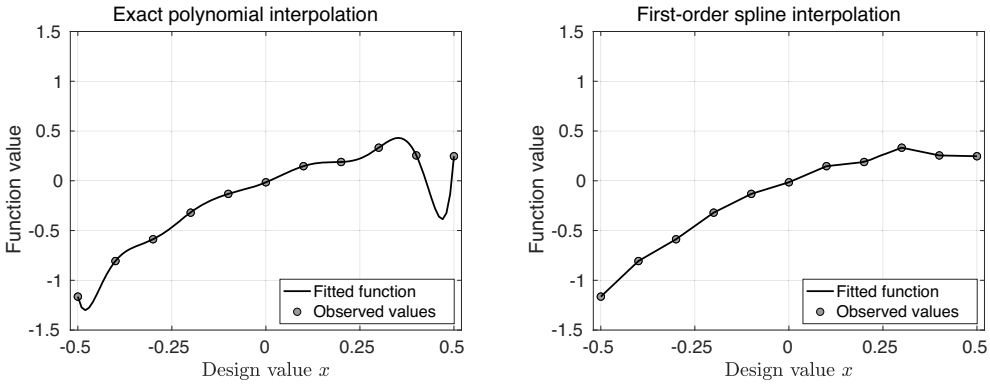
12.5 Interpolation and fitting

Reproducing kernel Hilbert spaces are useful for the classical problems of interpolating and fitting functions. An especially attractive property is the ease of computation: in particular, the representer theorem allows many optimization problems over the RKHS to be reduced to relatively simple calculations involving the kernel matrix.

12.5.1 Function interpolation

Let us begin with the problem of function interpolation. Suppose that we observe n samples of an unknown function $f^* : \mathcal{X} \to \mathbb{R}$, say of the form $y_i = f^*(x_i)$ for $i = 1, 2, \ldots, n$, where the design sequence $\{x_i\}_{i=1}^n$ is known to us. Note that we are assuming for the moment that the function values are observed without any noise or corruption. In this context, some questions of interest include:

- For a given function class $\mathscr{F}$, does there exist a function $f \in \mathscr{F}$ that exactly fits the data, meaning that $f(x_i) = y_i$ for all $i = 1, 2, \ldots, n$?
- Of all functions in $\mathscr{F}$ that exactly fit the data, which does the "best" job of interpolating the data?



(a) (b)

Figure 12.1 Exact interpolation of $n = 11$ equally sampled function values using RKHS methods. (a) Polynomial kernel $\mathcal{K}(x, z) = (1 + xz)^{12}$. (b) First-order Sobolev kernel $\mathcal{K}(x, z) = 1 + \min\{x, z\}$.

The first question can often be answered in a definitive way—in particular, by producing a function that exactly fits the data. The second question is vaguely posed and can be answered in multiple ways, depending on our notion of "best". In the context of a reproducing kernel Hilbert space, the underlying norm provides a way of ordering functions, and so we are led to the following formalization: of all the functions that exactly fit the data, choose the one with minimal RKHS norm. This approach can be formulated as an optimization problem in Hilbert space—namely,

$$\text{choose} \quad \widehat{f} \in \arg\min_{f \in \mathbb{H}} \|f\|_{\mathbb{H}} \quad \text{such that} \quad f(x_i) = y_i \text{ for } i = 1, 2, \ldots, n. \tag{12.25}$$

This method is known as *minimal norm interpolation*, and it is feasible whenever there exists at least one function $f \in \mathbb{H}$ that fits the data exactly. We provide necessary and sufficient conditions for such feasibility in the result to follow. Figure 12.1 illustrates this minimal Hilbert norm interpolation method, using the polynomial kernel from Example 12.8 in Figure 12.1(a), and the first-order Sobolev kernel from Example 12.23 in Figure 12.1(b).

For a general Hilbert space, the optimization problem (12.25) may not be well defined, or may be computationally challenging to solve. Hilbert spaces with reproducing kernels are attractive in this regard, as the computation can be reduced to simple linear algebra involving the kernel matrix $\mathbf{K} \in \mathbb{R}^{n \times n}$ with entries $K_{ij} = \mathcal{K}(x_i, x_j)/n$. The following result provides one instance of this general phenomenon:

Proposition 12.32 *Let $\mathbf{K} \in \mathbb{R}^{n \times n}$ be the kernel matrix defined by the design points $\{x_i\}_{i=1}^n$. The convex program* (12.25) *is feasible if and only if $y \in \text{range}(\mathbf{K})$, in which case any optimal solution can be written as*

$$\widehat{f}(\cdot) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \widehat{\alpha}_i \mathcal{K}(\cdot, x_i), \qquad \text{where } \mathbf{K}\widehat{\alpha} = y/\sqrt{n}.$$

Remark: Our choice of normalization by $1/\sqrt{n}$ is for later theoretical convenience.

Proof For a given vector $\alpha \in \mathbb{R}^n$, define the function $f_\alpha(\cdot) := \frac{1}{\sqrt{n}} \sum_{i=1}^n \alpha_i \mathcal{K}(\cdot, x_i)$, and consider the set $\mathbb{L} := \{f_\alpha \mid \alpha \in \mathbb{R}^n\}$. Note that for any $f_\alpha \in \mathbb{L}$, we have

$$f_\alpha(x_j) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \alpha_i \mathcal{K}(x_j, x_i) = \sqrt{n}(\mathbf{K}\alpha)_j,$$

where $(\mathbf{K}\alpha)_j$ is the jth component of the vector $\mathbf{K}\alpha \in \mathbb{R}^n$. Thus, the function $f_\alpha \in \mathbb{L}$ satisfies the interpolation condition if and only if $\mathbf{K}\alpha = y/\sqrt{n}$. Consequently, the condition $y \in \text{range}(\mathbf{K})$ is sufficient. It remains to show that this range condition is necessary, and that the optimal interpolating function must lie in $\mathbb{L}$.

Note that $\mathbb{L}$ is a finite-dimensional (hence closed) linear subspace of $\mathbb{H}$. Consequently, any function $f \in \mathbb{H}$ can be decomposed uniquely as $f = f_\alpha + f_\perp$, where $f_\alpha \in \mathbb{L}$ and $f_\perp$ is orthogonal to $\mathbb{L}$. (See Exercise 12.3 for details of this direct sum decomposition.) Using this decomposition and the reproducing property, we have

$$f(x_j) = \langle f, \mathcal{K}(\cdot, x_j) \rangle_{\mathbb{H}} = \langle f_\alpha + f_\perp, \mathcal{K}(\cdot, x_j) \rangle_{\mathbb{H}} = f_\alpha(x_j),$$

where the final equality follows because $\mathcal{K}(\cdot, x_j)$ belongs to $\mathbb{L}$, and we have $\langle f_\perp, \mathcal{K}(\cdot, x_j) \rangle_{\mathbb{H}} = 0$ due to the orthogonality of $f_\perp$ and $\mathbb{L}$. Thus, the component $f_\perp$ has no effect on the interpolation property, showing that the condition $y \in \text{range}(\mathbf{K})$ is also a necessary condition. Moreover, since f_α and $f_\perp$ are orthogonal, we are guaranteed to have $\|f_\alpha + f_\perp\|_{\mathbb{H}}^2 = \|f_\alpha\|_{\mathbb{H}}^2 + \|f_\perp\|_{\mathbb{H}}^2$. Consequently, for any Hilbert norm interpolant, we must have $f_\perp = 0$. □

12.5.2 Fitting via kernel ridge regression

In a statistical setting, it is usually unrealistic to assume that we observe noiseless observations of function values. Rather, it is more natural to consider a noisy observation model, say of the form

$$y_i = f^*(x_i) + w_i, \qquad \text{for } i = 1, 2, \ldots, n,$$

where the coefficients $\{w_i\}_{i=1}^n$ model noisiness or disturbance in the measurement model. In the presence of noise, the exact constraints in our earlier interpolation method (12.25) are no longer appropriate; instead, it is more sensible to minimize some trade-off between the fit to the data and the Hilbert norm. For instance, we might only require that the mean-squared differences between the observed data and fitted values be small, which then leads to the optimization problem

$$\min_{f \in \mathbb{H}} \|f\|_{\mathbb{H}} \qquad \text{such that } \frac{1}{2n} \sum_{i=1}^{n} \big(y_i - f(x_i)\big)^2 \leq \delta^2, \tag{12.26}$$

where $\delta > 0$ is some type of tolerance parameter. Alternatively, we might minimize the mean-squared error subject to a bound on the Hilbert radius of the solution, say

$$\min_{f \in \mathbb{H}} \frac{1}{2n} \sum_{i=1}^{n} \big(y_i - f(x_i)\big)^2 \qquad \text{such that } \|f\|_{\mathbb{H}} \leq R \tag{12.27}$$

for an appropriately chosen radius $R > 0$. Both of these problems are convex, and so by Lagrangian duality, they can be reformulated in the penalized form

$$\widehat{f} = \arg\min_{f \in \mathbb{H}} \Big\{ \frac{1}{2n} \sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda_n \|f\|_{\mathbb{H}}^2 \Big\}. \tag{12.28}$$

Here, for a fixed set of observations $\{(x_i, y_i)\}_{i=1}^n$, the regularization parameter $\lambda_n \geq 0$ is a function of the tolerance δ or radius R. This form of function estimate is most convenient to implement, and in the case of a reproducing kernel Hilbert space considered here, it is known as the *kernel ridge regression* estimate, or KRR estimate for short. The following result shows how the KRR estimate is easily computed in terms of the kernel matrix $\mathbf{K} \in \mathbb{R}^{n \times n}$ with entries $K_{ij} = \mathcal{K}(x_i, x_j)/n$.

Proposition 12.33 *For all $\lambda_n > 0$, the kernel ridge regression estimate* (12.28) *can be written as*

$$\widehat{f}(\cdot) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \widehat{\alpha}_i \mathcal{K}(\cdot, x_i), \tag{12.29}$$

where the optimal weight vector $\widehat{\alpha} \in \mathbb{R}^n$ is given by

$$\widehat{\alpha} = (\mathbf{K} + \lambda_n \mathbf{I}_n)^{-1} \frac{y}{\sqrt{n}}. \tag{12.30}$$

Remarks: Note that Proposition 12.33 is a natural generalization of Proposition 12.32, to which it reduces when $\lambda_n = 0$ (and the kernel matrix is invertible). Given the kernel matrix $\mathbf{K}$, computing $\widehat{\alpha}$ via equation (12.30) requires at most $O(n^3)$ operations, using standard routines in numerical linear algebra (see the bibliography for more details). Assuming that the kernel function can be evaluated in constant time, computing the $n \times n$ matrix requires an additional $O(n^2)$ operations. See Figure 12.2 for some illustrative examples.

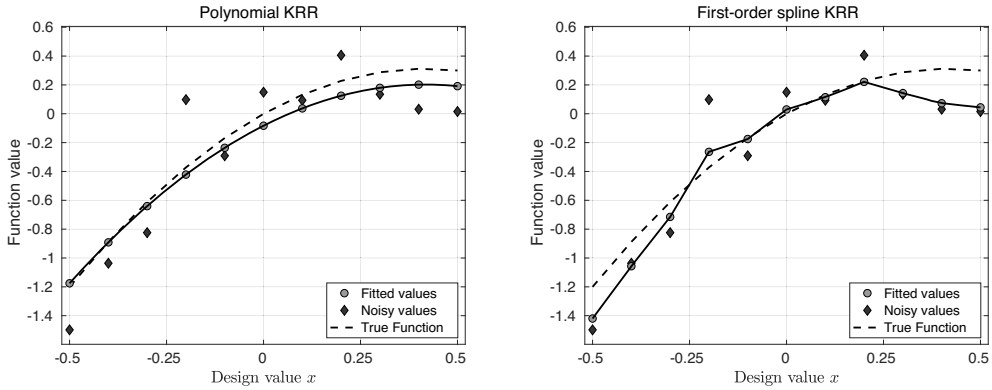


Figure 12.2 Illustration of kernel ridge regression estimates of function $f^*(x) = \frac{3x}{2} - \frac{9}{5}x^2$ based on $n = 11$ samples, located at design points $x_i = -0.5 + 0.10\,(i-1)$ over the interval $[-0.5, 0.5]$. (a) Kernel ridge regression estimate using the second-order polynomial kernel $\mathcal{K}(x, z) = (1 + xz)^2$ and regularization parameter $\lambda_n = 0.10$. (b) Kernel ridge regression estimate using the first-order Sobolev kernel $\mathcal{K}(x, z) = 1 + \min\{x, z\}$ and regularization parameter $\lambda_n = 0.10$.

We now turn to the proof of Proposition 12.33.

Proof Recall the argument of Proposition 12.32, and the decomposition $f = f_\alpha + f_\perp$. Since $f_\perp(x_i) = 0$ for all $i = 1, 2, \ldots, n$, it can have no effect on the least-squares data component of the objective function (12.28). Consequently, following a similar line of reasoning to the proof of Proposition 12.32, we again see that any optimal solution must be of the specified form (12.29).

It remains to prove the specific form (12.30) of the optimal $\widehat{\alpha}$. Given a function f of the form (12.29), for each $j = 1, 2, \ldots, n$, we have

$$f(x_j) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \alpha_i \mathcal{K}(x_j, x_i) = \sqrt{n}\, e_j^{\mathrm{T}} \mathbf{K} \alpha,$$

where $e_j \in \mathbb{R}^n$ is the canonical basis vector with 1 in position j, and we have recalled that $K_{ji} = \mathcal{K}(x_j, x_i)/n$. Similarly, we have the representation

$$\|f\|_{\mathbb{H}}^2 = \frac{1}{n} \left\langle \sum_{i=1}^{n} \alpha_i \mathcal{K}(\cdot, x_i), \sum_{j=1}^{n} \alpha_j \mathcal{K}(\cdot, x_j) \right\rangle_{\mathbb{H}} = \alpha^{\mathrm{T}} \mathbf{K} \alpha.$$

Substituting these relations into the cost function, we find that it is a quadratic in the vector α, given by

$$\frac{1}{n}\|y - \sqrt{n}\mathbf{K}\alpha\|_2^2 + \lambda\,\alpha^{\mathrm{T}}\mathbf{K}\alpha = \frac{1}{n}\|y\|_2^2 + \alpha^{\mathrm{T}}(\mathbf{K}^2 + \lambda\mathbf{K})\alpha - \frac{2}{\sqrt{n}}y^{\mathrm{T}}\mathbf{K}\alpha.$$

In order to find the minimum of this quadratic function, we compute the gradient and set it equal to zero, thereby obtaining the stationary condition

$$\mathbf{K}\,(\mathbf{K} + \lambda\mathbf{I}_n)\alpha = \mathbf{K}\frac{y}{\sqrt{n}}.$$

Thus, we see that the vector $\widehat{\alpha}$ previously defined in equation (12.30) is optimal. Note that any vector $\beta \in \mathbb{R}^n$ such that $\mathbf{K}\beta = 0$ has no effect on the optimal solution. □

We return in Chapter 13 to study the statistical properties of the kernel ridge regression estimate.

12.6 Distances between probability measures

There are various settings in which it is important to construct distances between probability measures, and one way in which to do so is via measuring mean discrepancies over a given function class. More precisely, let $\mathbb{P}$ and $\mathbb{Q}$ be a pair of probability measures on a space $\mathcal{X}$, and let $\mathscr{F}$ be a class of functions $f : \mathcal{X} \to \mathbb{R}$ that are integrable with respect to $\mathbb{P}$ and $\mathbb{Q}$. We can then define the quantity

$$\rho_{\mathscr{F}}(\mathbb{P}, \mathbb{Q}) := \sup_{f \in \mathscr{F}} \left| \int f(d\mathbb{P} - d\mathbb{Q}) \right| = \sup_{f \in \mathscr{F}} \left| \mathbb{E}_{\mathbb{P}}[f(X)] - \mathbb{E}_{\mathbb{Q}}[f(Z)] \right|. \tag{12.31}$$

It can be verified that, for any choice of function class $\mathscr{F}$, this always defines a pseudometric, meaning that $\rho_{\mathscr{F}}$ satisfies all the metric properties, except that there may exist pairs $\mathbb{P} \neq \mathbb{Q}$ such that $\rho_{\mathscr{F}}(\mathbb{P}, \mathbb{Q}) = 0$. When $\mathscr{F}$ is sufficiently rich, then $\rho_{\mathscr{F}}$ becomes a metric, known as an *integral probability metric*. Let us provide some classical examples to illustrate:

Example 12.34 (Kolmogorov metric) Suppose that $\mathbb{P}$ and $\mathbb{Q}$ are measures on the real line. For each $t \in \mathbb{R}$, let $\mathbb{I}_{(-\infty,t]}$ denote the $\{0, 1\}$-valued indicator function for the event $\{x \leq t\}$, and consider the function class $\mathscr{F} = \{\mathbb{I}_{(-\infty,t]} \mid t \in \mathbb{R}\}$. We then have

$$\rho_{\mathscr{F}}(\mathbb{P}, \mathbb{Q}) = \sup_{t \in \mathbb{R}} \left| \mathbb{P}(X \leq t) - \mathbb{Q}(X \leq t) \right| = \|F_{\mathbb{P}} - F_{\mathbb{Q}}\|_\infty,$$

where $F_{\mathbb{P}}$ and $F_{\mathbb{Q}}$ are the cumulative distribution functions of $\mathbb{P}$ and $\mathbb{Q}$, respectively. Thus, this choice leads to the *Kolmogorov distance* between $\mathbb{P}$ and $\mathbb{Q}$. ♣

Example 12.35 (Total variation distance) Consider the class $\mathscr{F} = \{f : \mathcal{X} \to \mathbb{R} \mid \|f\|_\infty \leq 1\}$ of real-valued functions bounded by one in the supremum norm. With this choice, we have

$$\rho_{\mathscr{F}}(\mathbb{P}, \mathbb{Q}) = \sup_{\|f\|_\infty \leq 1} \Big| \int f(d\mathbb{P} - d\mathbb{Q}) \Big|.$$

As we show in Exercise 12.17, this metric corresponds to (two times) the total variation distance

$$\|\mathbb{P} - \mathbb{Q}\|_1 = \sup_{A \subset \mathcal{X}} |\mathbb{P}(A) - \mathbb{Q}(A)|,$$

where the supremum ranges over all measurable subsets of $\mathcal{X}$. ♣

When we choose $\mathscr{F}$ to be the unit ball of an RKHS, we obtain a mean discrepancy pseudometric that is easy to compute. In particular, given an RKHS with kernel function $\mathcal{K}$, consider the associated pseudometric

$$\rho_{\mathbb{H}}(\mathbb{P}, \mathbb{Q}) := \sup_{\|f\|_{\mathbb{H}} \leq 1} \left| \mathbb{E}_{\mathbb{P}}[f(X)] - \mathbb{E}_{\mathbb{Q}}[f(Z)] \right|.$$

As verified in Exercise 12.18, the reproducing property allows us to obtain a simple closed-form expression for this pseudometric—namely,

$$\rho^2_{\mathbb{H}}(\mathbb{P}, \mathbb{Q}) = \mathbb{E}[\mathcal{K}(X, X') + \mathcal{K}(Z, Z') - 2\mathcal{K}(X, Z)], \tag{12.32}$$

where $X, X' \sim \mathbb{P}$ and $Z, Z' \sim \mathbb{Q}$ are all mutually independent random vectors. We refer to this pseudometric as a *kernel means discrepancy*, or KMD for short.

Example 12.36 (KMD for linear and polynomial kernels) Let us compute the KMD for the linear kernel $\mathcal{K}(x, z) = \langle x, z \rangle$ on $\mathbb{R}^d$. Letting $\mathbb{P}$ and $\mathbb{Q}$ be two distributions on $\mathbb{R}^d$ with mean vectors $\mu_p = \mathbb{E}_{\mathbb{P}}[X]$ and $\mu_q = \mathbb{E}_{\mathbb{Q}}[Z]$, respectively, we have

$$\begin{aligned} \rho^2_{\mathbb{H}}(\mathbb{P}, \mathbb{Q}) &= \mathbb{E}\Big[\langle X, X' \rangle + \langle Z, Z' \rangle - 2\langle X, Z \rangle \Big] \\ &= \|\mu_p\|_2^2 + \|\mu_q\|_2^2 - 2 \left\langle \mu_p, \mu_q \right\rangle \\ &= \|\mu_p - \mu_q\|_2^2. \end{aligned}$$

Thus, we see that the KMD pseudometric for the linear kernel simply computes the Euclidean distance of the associated mean vectors. This fact demonstrates that KMD in this very special case is not actually a metric (but rather just a pseudometric), since $\rho_{\mathbb{H}}(\mathbb{P}, \mathbb{Q}) = 0$ for any pair of distributions with the same means (i.e., $\mu_p = \mu_q$).

Moving onto polynomial kernels, let us consider the homogeneous polynomial kernel of degree two, namely $\mathcal{K}(x, z) = \langle x, z \rangle^2$. For this choice of kernel, we have

$$\mathbb{E}[\mathcal{K}(X, X')] = \mathbb{E}\left[\Big(\sum_{j=1}^d X_j X'_j \Big)^2 \right] = \sum_{i,j=1}^d \mathbb{E}[X_i X_j] \mathbb{E}[X'_i X'_j] = |||\mathbf{\Gamma}_p|||_F^2,$$

where $\mathbf{\Gamma}_p \in \mathbb{R}^{d \times d}$ is the second-order moment matrix with entries $[\mathbf{\Gamma}_p]_{ij} = \mathbb{E}[X_i X_j]$, and the squared Frobenius norm corresponds to the sum of the squared matrix entries. Similarly, we have $\mathbb{E}[\mathcal{K}(Z, Z')] = |||\mathbf{\Gamma}_q|||_F^2$, where $\mathbf{\Gamma}_q$ is the second-order moment matrix for $\mathbb{Q}$. Finally, similar calculations yield that

$$\mathbb{E}[\mathcal{K}(X, Z)] = \sum_{i,j=1}^d [\mathbf{\Gamma}_p]_{ij} [\mathbf{\Gamma}_q]_{ij} = \langle\!\langle \mathbf{\Gamma}_p, \mathbf{\Gamma}_q \rangle\!\rangle,$$

where $\langle\!\langle \cdot, \cdot \rangle\!\rangle$ denotes the trace inner product between symmetric matrices. Putting together the pieces, we conclude that, for the homogeneous second-order polynomial kernel, we have

$$\rho^2_{\mathbb{H}}(\mathbb{P}, \mathbb{Q}) = |||\mathbf{\Gamma}_p - \mathbf{\Gamma}_q|||_F^2. \qquad \clubsuit$$

Example 12.37 (KMD for a first-order Sobolev kernel) Let us now consider the KMD induced by the kernel function $\mathcal{K}(x, z) = \min\{x, z\}$, defined on the Cartesian product $[0, 1] \times [0, 1]$. As seen previously in Example 12.16, this kernel function generates the first-order Sobolev space

$$\mathbb{H}^1[0, 1] = \left\{ f : \mathbb{R}[0, 1] \to \mathbb{R} \mid f(0) = 0 \text{ and } \int_0^1 (f'(x))^2 \, dx < \infty \right\},$$

with Hilbert norm $\|f\|^2_{\mathbb{H}^1[0,1]} = \int_0^1 (f'(x))^2 \, dx$. With this choice, we have

$$\rho^2_{\mathbb{H}}(\mathbb{P}, \mathbb{Q}) = \mathbb{E}\Big[\min\{X, X'\} + \min\{Z, Z'\} - 2 \min\{X, Z\} \Big]. \qquad \clubsuit$$

12.7 Bibliographic details and background

The notion of a reproducing kernel Hilbert space emerged from the study of positive semidefinite kernels and their links to Hilbert space structure. The seminal paper by Aronszajn (1950) develops a number of the basic properties from first principles, including Propositions 12.27 and 12.31 as well as Theorem 12.11 from this chapter. The use of the kernel trick for computing inner products via kernel evaluations dates back to Aizerman et al. (1964), and underlies the success of the support vector machine developed by Boser et al. (1992), and discussed in Exercise 12.20. The book by Wahba (1990) contains a wealth of information on RKHSs, as well as the connections between splines and penalized methods for regression. See also the books by Berlinet and Thomas-Agnan (2004) as well as Gu (2002). The book by Schölkopf and Smola (2002) provides a number of applications of kernels in the setting of machine learning, including the support vector machine (Exercise 12.20) and related methods for classification, as well as kernel principal components analysis. The book by Steinwart and Christmann (2008) also contains a variety of theoretical results on kernels and reproducing kernel Hilbert spaces.

The argument underlying the proofs of Propositions 12.32 and 12.33 is known as the *representer theorem*, and is due to Kimeldorf and Wahba (1971). From the computational point of view, it is extremely important, since it allows the infinite-dimensional problem of optimizing over an RKHS to be reduced to an n-dimensional convex program. Bochner's theorem relates the positive semidefiniteness of kernel functions to the non-negativity of Fourier coefficients. In its classical formulation, it applies to the Fourier transform over $\mathbb{R}^d$, but it can be generalized to all locally compact Abelian groups (Rudin, 1990). The results used to compute the asymptotic scaling of the eigenvalues of the Gaussian kernel in Example 12.25 are due to Widom (1963; 1964).

There are a number of papers that study the approximation-theoretic properties of various types of reproducing kernel Hilbert spaces. For a given Hilbert space $\mathbb{H}$ and norm $\|\cdot\|$, such

results are often phrased in terms of the function

$$A(f^*; R) := \inf_{\|f\|_{\mathbb{H}} \leq R} \|f - f^*\|_p, \tag{12.33}$$

where $\|g\|_p := (\int_{\mathcal{X}} g^p(x)\, dx)^{1/p}$ is the usual L^p-norm on a compact space $\mathcal{X}$. This function measures how quickly the $L^p(\mathcal{X})$-error in approximating some function f^* decays as the Hilbert radius R is increased. See the papers (Smale and Zhou, 2003; Zhou, 2013) for results on this form of the approximation error. A reproducing kernel Hilbert space is said to be $L^p(\mathcal{X})$-*universal* if $\lim_{R \to \infty} A(f^*; R) = 0$ for any $f^* \in L^p(\mathcal{X})$. There are also various other forms of universality; see the book by Steinwart and Christmann (2008) for further details.

Integral probability metrics of the form (12.31) have been studied extensively (Müller, 1997; Rachev et al., 2013). The particular case of RKHS-based distances are computationally convenient, and have been studied in the context of proper scoring rules (Dawid, 2007; Gneiting and Raftery, 2007) and two-sample testing (Borgwardt et al., 2006; Gretton et al., 2012).

12.8 Exercises

Exercise 12.1 (Closedness of nullspace) Let L be a bounded linear functional on a Hilbert space. Show that the subspace $\text{null}(L) = \{f \in \mathbb{H} \mid L(f) = 0\}$ is closed.

Exercise 12.2 (Projections in a Hilbert space) Let $\mathbb{G}$ be a closed convex subset of a Hilbert space $\mathbb{H}$. In this exercise, we show that for any $f \in \mathbb{H}$, there exists a unique $\widehat{g} \in \mathbb{G}$ such that

$$\|\widehat{g} - f\|_{\mathbb{H}} = \underbrace{\inf_{g \in \mathbb{G}} \|\widehat{g} - f\|_{\mathbb{H}}}_{p^*}.$$

This element $\widehat{g}$ is known as the projection of f onto $\mathbb{G}$.

(a) By the definition of infimum, there exists a sequence $(g_n)_{n=1}^{\infty}$ contained in $\mathbb{G}$ such that $\|g_n - f\|_{\mathbb{H}} \to p^*$. Show that this sequence is a Cauchy sequence. (*Hint:* First show that $\|f - \frac{g_n + g_m}{2}\|_{\mathbb{H}}$ converges to p^*.)

(b) Use this Cauchy sequence to establish the existence of $\widehat{g}$.

(c) Show that the projection must be unique.

(d) Does the same claim hold for an arbitrary convex set $\mathbb{G}$?

Exercise 12.3 (Direct sum decomposition in Hilbert space) Let $\mathbb{H}$ be a Hilbert space, and let $\mathbb{G}$ be a closed linear subspace of $\mathbb{H}$. Show that any $f \in \mathbb{H}$ can be decomposed uniquely as $g + g^{\perp}$, where $g \in \mathbb{G}$ and $g^{\perp} \in \mathbb{G}^{\perp}$. In brief, we say that $\mathbb{H}$ has the direct sum decomposition $\mathbb{G} \oplus \mathbb{G}^{\perp}$. (*Hint:* The notion of a projection onto a closed convex set from Exercise 12.2 could be helpful to you.)

Exercise 12.4 (Uniqueness of kernel) Show that the kernel function associated with any reproducing kernel Hilbert space must be unique.

Exercise 12.5 (Kernels and Cauchy–Schwarz)

(a) For any positive semidefinite kernel $\mathcal{K}: \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, prove that

$$\mathcal{K}(x, z) \leq \sqrt{\mathcal{K}(x, x)\mathcal{K}(z, z)} \qquad \text{for all } x, z \in \mathcal{X}.$$

(b) Show how the classical Cauchy–Schwarz inequality is a special case.

Exercise 12.6 (Eigenfunctions for linear kernels) Consider the ordinary linear kernel $\mathcal{K}(x, z) = \langle x, z \rangle$ on $\mathbb{R}^d$ equipped with a probability measure $\mathbb{P}$. Assuming that a random vector $X \sim \mathbb{P}$ has all its second moments finite, show how to compute the eigenfunctions of the associated kernel operator acting on $L^2(\mathcal{X}; \mathbb{P})$ in terms of linear algebraic operations.

Exercise 12.7 (Different kernels for polynomial functions) For an integer $m \geq 1$, consider the kernel functions $\mathcal{K}_1(x, z) = (1 + xz)^m$ and $\mathcal{K}_2(x, z) = \sum_{\ell=0}^{m} \frac{x^\ell}{\ell!} \frac{z^\ell}{\ell!}$.

(a) Show that they are both PSD, and generate RKHSs of polynomial functions of degree at most m.
(b) Why does this not contradict the result of Exercise 12.4?

Exercise 12.8 True or false? If true, provide a short proof; if false, give an explicit counterexample.

(a) Given two PSD kernels $\mathcal{K}_1$ and $\mathcal{K}_2$, the bivariate function $\mathcal{K}(x, z) = \min_{j=1,2} \mathcal{K}_j(x, z)$ is also a PSD kernel.
(b) Let $f: \mathcal{X} \to \mathbb{H}$ be a function from an arbitrary space $\mathcal{X}$ to a Hilbert space $\mathbb{H}$. The bivariate function

$$\mathcal{K}(x, z) = \frac{\langle f(x), f(z) \rangle_{\mathbb{H}}}{\|f(x)\|_{\mathbb{H}} \|f(z)\|_{\mathbb{H}}}$$

defines a PSD kernel on $\mathcal{X} \times \mathcal{X}$.

Exercise 12.9 (Left–right multiplication and kernels) Let $\mathcal{K}: \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a positive semidefinite kernel, and let $f: \mathcal{X} \to \mathbb{R}$ be an arbitrary function. Show that $\widetilde{\mathcal{K}}(x, z) = f(x)\mathcal{K}(x, z)f(z)$ is also a positive semidefinite kernel.

Exercise 12.10 (Kernels and power sets) Given a finite set S, its power set $\mathcal{P}(S)$ is the set of all the subsets of S. Show that the function $\mathcal{K}: \mathcal{P}(S) \times \mathcal{P}(S) \to \mathbb{R}$ given by $\mathcal{K}(A, B) = 2^{|A \cap B|}$ is a positive semidefinite kernel function.

Exercise 12.11 (Feature map for polynomial kernel) Recall from equation (12.14) the notion of a feature map. Show that the polynomial kernel $\mathcal{K}(x, z) = (1 + \langle x, z \rangle)^m$ defined on the Cartesian product space $\mathbb{R}^d \times \mathbb{R}^d$ can be realized by a feature map $x \mapsto \Phi(x) \in \mathbb{R}^D$, where $D = \binom{d+m}{m}$.

Exercise 12.12 (Probability spaces and kernels) Consider a probability space with events $\mathcal{E}$ and probability law $\mathbb{P}$. Show that the real-valued function

$$\mathcal{K}(A, B) := \mathbb{P}[A \cap B] - \mathbb{P}[A]\mathbb{P}[B]$$

is a positive semidefinite kernel function on $\mathcal{E} \times \mathcal{E}$.

Exercise 12.13 (From sets to power sets) Suppose that $\mathcal{K}: S \times S \to \mathbb{R}$ is a symmetric PSD kernel function on a finite set S. Show that

$$\mathcal{K}'(A, B) = \sum_{x \in A,\, z \in B} \mathcal{K}(x, z)$$

is a symmetric PSD kernel on the power set $\mathcal{P}(S)$.

Exercise 12.14 (Kernel and function boundedness) Consider a PSD kernel $\mathcal{K}: \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ such that $\mathcal{K}(x, z) \leq b^2$ for all $x, z \in \mathcal{X}$. Show that $\|f\|_\infty \leq b$ for any function f in the unit ball of the associated RKHS.

Exercise 12.15 (Sobolev kernels and norms) Show that the Sobolev kernel defined in equation (12.20) generates the norm given in equation (12.21).

Exercise 12.16 (Hadamard products and kernel products) In this exercise, we explore properties of product kernels and the Hadamard product of matrices.

(a) Given two $n \times n$ matrices $\mathbf{\Gamma}$ and $\mathbf{\Sigma}$ that are symmetric and positive semidefinite, show that the Hadamard product matrix $\mathbf{\Sigma} \odot \mathbf{\Gamma} \in \mathbb{R}^{n \times n}$ is also positive semidefinite. (The Hadamard product is simply the elementwise product—that is, $(\mathbf{\Sigma} \odot \mathbf{\Gamma})_{ij} = \Sigma_{ij}\Gamma_{ij}$ for all $i, j = 1, 2, \ldots, n$.)

(b) Suppose that $\mathcal{K}_1$ and $\mathcal{K}_2$ are positive semidefinite kernel functions on $\mathcal{X} \times \mathcal{X}$. Show that the function $\mathcal{K}(x, z) := \mathcal{K}_1(x, z)\,\mathcal{K}_2(x, z)$ is a positive semidefinite kernel function. (*Hint:* The result of part (a) could be helpful.)

Exercise 12.17 (Total variation norm) Given two probability measures $\mathbb{P}$ and $\mathbb{Q}$ on $\mathcal{X}$, show that

$$\sup_{\|f\|_\infty \leq 1} \left| \int f(d\mathbb{P} - d\mathbb{Q}) \right| = 2 \sup_{A \subset \mathcal{X}} |\mathbb{P}(A) - \mathbb{Q}(A)|,$$

where the left supremum ranges over all measurable functions $f: \mathcal{X} \to \mathbb{R}$, and the right supremum ranges over all measurable subsets A of $\mathcal{X}$.

Exercise 12.18 (RKHS-induced semi-metrics) Let $\mathbb{H}$ be a reproducing kernel Hilbert space of functions with domain $\mathcal{X}$, and let $\mathbb{P}$ and $\mathbb{Q}$ be two probability distributions on $\mathcal{X}$. Show that

$$\sup_{\|f\|_{\mathbb{H}} \leq 1} \left| \mathbb{E}_{\mathbb{P}}[f(X)] - \mathbb{E}_{\mathbb{Q}}[f(Z)] \right|^2 = \mathbb{E}[\mathcal{K}(X, X') + \mathcal{K}(Z, Z') - 2\mathcal{K}(X, Z)],$$

where $X, X' \sim \mathbb{P}$ and $Z, Z' \sim \mathbb{Q}$ are jointly independent.

Exercise 12.19 (Positive semidefiniteness of Gaussian kernel) Let $\mathcal{X}$ be a compact subset of $\mathbb{R}^d$. In this exercise, we work through a proof of the fact that the Gaussian kernel $\mathcal{K}(x, z) = e^{-\frac{\|x - z\|_2^2}{2\sigma^2}}$ on $\mathcal{X} \times \mathcal{X}$ is positive semidefinite.

(a) Let $\widetilde{\mathcal{K}}$ be a PSD kernel, and let p be a polynomial with non-negative coefficients. Show that $\mathcal{K}(x, z) = p(\widetilde{\mathcal{K}}(x, z))$ is a PSD kernel.

(b) Show that the kernel $\mathcal{K}_1(x, z) = e^{\langle x, z \rangle / \sigma^2}$ is positive semidefinite. (*Hint:* Part (a) and the fact that a pointwise limit of PSD kernels is also PSD could be useful.)

(c) Show that the Gaussian kernel is PSD. (*Hint:* The result of Exercise 12.9 could be useful.)

Exercise 12.20 (Support vector machines and kernel methods) In the problem of binary classification, one observes a collection of pairs $\{(x_i, y_i)\}_{i=1}^n$, where each feature vector $x_i \in \mathbb{R}^d$ is associated with a label $y_i \in \{-1, +1\}$, and the goal is derive a classification function that can be applied to unlabelled feature vectors. In the context of reproducing kernel Hilbert spaces, one way of doing so is by minimizing a criterion of the form

$$\widehat{f} = \arg\min_{f \in \mathbb{H}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \max\{0, 1 - y_i f(x_i)\} + \frac{1}{2} \lambda_n \|f\|_{\mathbb{H}}^2 \right\}, \tag{12.34}$$

where $\mathbb{H}$ is a reproducing kernel Hilbert space, and $\lambda_n > 0$ is a user-defined regularization parameter. The classification rule is then given by $x \mapsto \operatorname{sign}(\widehat{f}(x))$.

(a) Prove that $\widehat{f}$ can be written in the form $\widehat{f}(\cdot) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \widehat{\alpha}_i \mathcal{K}(\cdot, x_i)$, for some vector $\widehat{\alpha} \in \mathbb{R}^n$.

(b) Use part (a) and duality theory to show that an optimal coefficient vector $\widehat{\alpha}$ can be obtained by solving the problem

$$\widehat{\alpha} \in \arg\max_{\alpha \in \mathbb{R}^n} \left\{ \frac{1}{n} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \alpha^T \widetilde{\mathbf{K}} \alpha \right\} \qquad \text{s.t. } \alpha_i \in [0, \tfrac{1}{\lambda_n \sqrt{n}}] \text{ for all } i = 1, \ldots, n,$$

and where $\widetilde{\mathbf{K}} \in \mathbb{R}^{n \times n}$ has entries $\widetilde{K}_{ij} := y_i y_j \mathcal{K}(x_i, x_j)/n$.