6.3 Reproducing Kernel Hilbert Spaces

In this section, we formalize the idea outlined at the end of Section 6.1 of extending finite dimensional feature maps to those that are *functions* by introducing a special type of Hilbert space of functions known as a *reproducing kernel Hilbert space* (RKHS). Although the theory extends naturally to Hilbert spaces of complex-valued functions, we restrict attention to Hilbert spaces of real-valued functions here.

To evaluate the loss of a learner g in some class of functions $\mathcal{G}$, we do not need to explicitly construct g — rather, it is only required that we can evaluate g at all the feature vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ of the training set. A defining property of an RKHS is that function evaluation at a point $\boldsymbol{x}$ can be performed by simply taking the inner product of g with some feature function $\kappa_{\boldsymbol{x}}$ associated with $\boldsymbol{x}$. We will see that this property becomes particularly useful in light of the representer theorem (see Section 6.5), which states that the learner g itself can be represented as a linear combination of the set of feature functions $\{\kappa_{\boldsymbol{x}_i}, i = 1, \ldots, n\}$. Consequently, we can evaluate a learner g at the feature vectors $\{\boldsymbol{x}_i\}$ by taking linear combinations of terms of the form $\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = \langle \kappa_{\boldsymbol{x}_i}, \kappa_{\boldsymbol{x}_j} \rangle_{\mathcal{G}}$. Collecting these inner products into a matrix $\mathbf{K} = [\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j), i, j = 1, \ldots, n]$ (the Gram matrix of the $\{\kappa_{\boldsymbol{x}_i}\}$), we will see that the feature vectors $\{\boldsymbol{x}_i\}$ only enter the loss minimization problem through $\mathbf{K}$. ☞ 230

REPRODUCING KERNEL HILBERT SPACE

Definition 6.1: Reproducing Kernel Hilbert Space

For a non-empty set $\mathcal{X}$, a Hilbert space $\mathcal{G}$ of functions $g : \mathcal{X} \to \mathbb{R}$ with inner product $\langle \cdot, \cdot \rangle_{\mathcal{G}}$ is called a *reproducing kernel Hilbert space* (RKHS) with *reproducing kernel* $\kappa : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ if:

1. for every $\boldsymbol{x} \in \mathcal{X}$, $\kappa_{\boldsymbol{x}} := \kappa(\boldsymbol{x}, \cdot)$ is in $\mathcal{G}$,
2. $\kappa(\boldsymbol{x}, \boldsymbol{x}) < \infty$ for all $\boldsymbol{x} \in \mathcal{X}$,
3. for every $\boldsymbol{x} \in \mathcal{X}$ and $g \in \mathcal{G}$, $g(\boldsymbol{x}) = \langle g, \kappa_{\boldsymbol{x}} \rangle_{\mathcal{G}}$.

REPRODUCING PROPERTY

POSITIVE SEMIDEFINITE

The reproducing kernel of a Hilbert space of functions, if it exists, is unique; see Exercise 2. The main (third) condition in Definition 6.1 is known as the *reproducing property*. This property allows us to evaluate any function $g \in \mathcal{G}$ at a point $\boldsymbol{x} \in \mathcal{X}$ by taking the inner product of g and $\kappa_{\boldsymbol{x}}$; as such, $\kappa_{\boldsymbol{x}}$ is called the *representer of evaluation*. Further, by taking $g = \kappa_{\boldsymbol{x}'}$ and applying the reproducing property, we have $\langle \kappa_{\boldsymbol{x}'}, \kappa_{\boldsymbol{x}} \rangle_{\mathcal{G}} = \kappa(\boldsymbol{x}, \boldsymbol{x}')$, and so by symmetry of the inner product it follows that $\kappa(\boldsymbol{x}, \boldsymbol{x}') = \kappa(\boldsymbol{x}', \boldsymbol{x})$. As a consequence, reproducing kernels are necessarily *symmetric* functions. Moreover, a reproducing kernel κ is a *positive semidefinite* function, meaning that for every $n \geqslant 1$ and every choice of $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$ and $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in \mathcal{X}$, it holds that

$$\sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \, \kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) \, \alpha_j \geqslant 0. \tag{6.13}$$

In other words, *every* Gram matrix $\mathbf{K}$ associated with κ is a positive semidefinite matrix; that is $\boldsymbol{\alpha}^\top \mathbf{K} \boldsymbol{\alpha} \geqslant 0$ for all $\boldsymbol{\alpha}$. The proof is addressed in Exercise 1.

The following theorem gives an alternative characterization of an RKHS. The proof uses the Riesz representation Theorem A.17. Also note that in the theorem below we could ☞ 390 have replaced the word “bounded” with “continuous”, as the two are equivalent for linear functionals; see Theorem A.16.

Theorem 6.1: Continuous Evaluation Functionals Characterize a RKHS

An RKHS $\mathcal{G}$ on a set $\mathcal{X}$ is a Hilbert space in which every *evaluation functional* $\delta_{\boldsymbol{x}} : g \mapsto g(\boldsymbol{x})$ is bounded. Conversely, a Hilbert space $\mathcal{G}$ of functions $\mathcal{X} \to \mathbb{R}$ for which every evaluation functional is bounded is an RKHS.

EVALUATION FUNCTIONAL

Proof: Note that, since evaluation functionals $\delta_{\boldsymbol{x}}$ are linear operators, showing boundedness is equivalent to showing continuity. Given an RKHS with reproducing kernel κ, suppose that we have a sequence $g_n \in \mathcal{G}$ converging to $g \in \mathcal{G}$, that is $\|g_n - g\|_{\mathcal{G}} \to 0$. We apply the Cauchy–Schwarz inequality (Theorem A.15) and the reproducing property of κ to find that for every $\boldsymbol{x} \in \mathcal{X}$ and any n:

☞ 389

$$|\delta_{\boldsymbol{x}} g_n - \delta_{\boldsymbol{x}} g| = |g_n(\boldsymbol{x}) - g(\boldsymbol{x})| = |\langle g_n - g, \kappa_{\boldsymbol{x}} \rangle_{\mathcal{G}}| \leqslant \|g_n - g\|_{\mathcal{G}}\, \|\kappa_{\boldsymbol{x}}\|_{\mathcal{G}} = \|g_n - g\|_{\mathcal{G}} \sqrt{\langle \kappa_{\boldsymbol{x}}, \kappa_{\boldsymbol{x}} \rangle_{\mathcal{G}}}$$
$$= \|g_n - g\|_{\mathcal{G}} \sqrt{\kappa(\boldsymbol{x}, \boldsymbol{x})}.$$

Noting that $\sqrt{\kappa(\boldsymbol{x}, \boldsymbol{x})} < \infty$ by definition for every $\boldsymbol{x} \in \mathcal{X}$, and that $\|g_n - g\|_{\mathcal{G}} \to 0$ as $n \to \infty$, we have shown continuity of $\delta_{\boldsymbol{x}}$, that is $|\delta_{\boldsymbol{x}} g_n - \delta_{\boldsymbol{x}} g| \to 0$ as $n \to \infty$ for every $\boldsymbol{x} \in \mathcal{X}$.

Conversely, suppose that evaluation functionals are bounded. Then from the Riesz representation Theorem A.17, there exists some $g_{\delta_{\boldsymbol{x}}} \in \mathcal{G}$ such that $\delta_{\boldsymbol{x}} g = \langle g, g_{\delta_{\boldsymbol{x}}} \rangle_{\mathcal{G}}$ for all $g \in \mathcal{G}$ — the *representer* of evaluation. If we define $\kappa(\boldsymbol{x}, \boldsymbol{x}') = g_{\delta_{\boldsymbol{x}}}(\boldsymbol{x}')$ for all $\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}$, then $\kappa_{\boldsymbol{x}} := \kappa(\boldsymbol{x}, \cdot) = g_{\delta_{\boldsymbol{x}}}$ is an element of $\mathcal{G}$ for every $\boldsymbol{x} \in \mathcal{X}$ and $\langle g, \kappa_{\boldsymbol{x}} \rangle_{\mathcal{G}} = \delta_{\boldsymbol{x}} g = g(\boldsymbol{x})$, so that the reproducing property in Definition 6.1 is verified. □

The fact that an RKHS has continuous evaluation functionals means that if two functions $g, h \in \mathcal{G}$ are “close” with respect to $\|\cdot\|_{\mathcal{G}}$, then their evaluations $g(\boldsymbol{x}), h(\boldsymbol{x})$ are close *for every* $\boldsymbol{x} \in \mathcal{X}$. Formally, convergence in $\|\cdot\|_{\mathcal{G}}$ norm implies pointwise convergence for all $\boldsymbol{x} \in \mathcal{X}$.

The following theorem shows that any finite function $\kappa : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ can serve as a reproducing kernel as long as it is finite, symmetric, and positive semidefinite. The corresponding (unique!) RKHS $\mathcal{G}$ is the completion of the set of all functions of the form $\sum_{i=1}^{n} \alpha_i \kappa_{\boldsymbol{x}_i}$ where $\alpha_i \in \mathbb{R}$ for all $i = 1, \ldots, n$.

Theorem 6.2: Moore–Aronszajn

Given a non-empty set $\mathcal{X}$ and any finite symmetric positive semidefinite function $\kappa : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, there exists an RKHS $\mathcal{G}$ of functions $g : \mathcal{X} \to \mathbb{R}$ with reproducing kernel κ. Moreover, $\mathcal{G}$ is unique.

Proof: (Sketch) As the proof of uniqueness is treated in Exercise 2, the objective is to prove existence. The idea is to construct a pre-RKHS $\mathcal{G}_0$ from the given function κ that has the essential structure and then to extend $\mathcal{G}_0$ to an RKHS $\mathcal{G}$.

In particular, define $\mathcal{G}_0$ as the set of finite linear combinations of functions $\kappa_{\boldsymbol{x}}$, $\boldsymbol{x} \in \mathcal{X}$:

$$\mathcal{G}_0 := \left\{ g = \sum_{i=1}^{n} \alpha_i \kappa_{\boldsymbol{x}_i} \;\middle|\; \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in \mathcal{X},\ \alpha_i \in \mathbb{R},\ n \in \mathbb{N} \right\}.$$

Define on $\mathcal{G}_0$ the following inner product:

$$\langle f, g\rangle_{\mathcal{G}_0} := \left\langle \sum_{i=1}^{n} \alpha_i \, \kappa_{x_i}, \sum_{j=1}^{m} \beta_j \, \kappa_{x'_j} \right\rangle_{\mathcal{G}_0} := \sum_{i=1}^{n} \sum_{j=1}^{m} \alpha_i \beta_j \, \kappa(\boldsymbol{x}_i, \boldsymbol{x}'_j).$$

Then $\mathcal{G}_0$ is an inner product space. In fact, $\mathcal{G}_0$ has the essential structure we require, namely that (i) evaluation functionals are bounded/continuous (Exercise 4) and (ii) Cauchy sequences in $\mathcal{G}_0$ that converge pointwise also converge in norm (see Exercise 5).

We then enlarge $\mathcal{G}_0$ to the set $\mathcal{G}$ of all functions $g : \mathcal{X} \to \mathbb{R}$ for which there exists a Cauchy sequence in $\mathcal{G}_0$ converging pointwise to g and define an inner product on $\mathcal{G}$ as the limit

$$\langle f, g\rangle_{\mathcal{G}} := \lim_{n\to\infty} \langle f_n, g_n\rangle_{\mathcal{G}_0}, \tag{6.14}$$

where $f_n \to f$ and $g_n \to g$. To show that $\mathcal{G}$ is an RKHS it remains to be shown that (1) this inner product is well defined; (2) evaluation functionals remain bounded; and (3) the space $\mathcal{G}$ is complete. A detailed proof is established in Exercises 6 and 7. □

6.4 Construction of Reproducing Kernels

In this section we describe various ways to construct a reproducing kernel $\kappa : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ for some feature space $\mathcal{X}$. Recall that κ needs to be a finite, symmetric, and positive semidefinite function (that is, it satisfies (6.13)). In view of Theorem 6.2, specifying the space $\mathcal{X}$ and a reproducing kernel $\kappa : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ corresponds to *uniquely* specifying an RKHS.

6.4.1 Reproducing Kernels via Feature Mapping

Perhaps the most fundamental way to construct a reproducing kernel κ is via a feature map $\boldsymbol{\phi} : \mathcal{X} \to \mathbb{R}^p$. We define $\kappa(\boldsymbol{x}, \boldsymbol{x}') := \langle \boldsymbol{\phi}(\boldsymbol{x}), \boldsymbol{\phi}(\boldsymbol{x}')\rangle$, where $\langle \, , \, \rangle$ denotes the Euclidean inner product. The function is clearly finite and symmetric. To verify that κ is positive semidefinite, let $\boldsymbol{\Phi}$ be the matrix with rows $\boldsymbol{\phi}(\boldsymbol{x}_1)^\top, \ldots, \boldsymbol{\phi}(\boldsymbol{x}_n)^\top$ and let $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_n]^\top \in \mathbb{R}^n$. Then,

$$\sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \, \kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) \, \alpha_j = \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \, \boldsymbol{\phi}^\top(\boldsymbol{x}_i) \, \boldsymbol{\phi}(\boldsymbol{x}_j) \, \alpha_j = \boldsymbol{\alpha}^\top \boldsymbol{\Phi}\boldsymbol{\Phi}^\top \boldsymbol{\alpha} = \|\boldsymbol{\Phi}^\top \boldsymbol{\alpha}\|^2 \geqslant 0.$$

■ **Example 6.4 (Linear Kernel)** Taking the identity feature map $\boldsymbol{\phi}(\boldsymbol{x}) = \boldsymbol{x}$ on $\mathcal{X} = \mathbb{R}^p$, gives the *linear kernel*

LINEAR KERNEL

$$\kappa(\boldsymbol{x}, \boldsymbol{x}') = \langle \boldsymbol{x}, \boldsymbol{x}'\rangle = \boldsymbol{x}^\top \boldsymbol{x}'.$$

As can be seen from the proof of Theorem 6.2, the RKHS of functions corresponding to the linear kernel is the space of *linear* functions on $\mathbb{R}^p$. This space is isomorphic to $\mathbb{R}^p$ itself, as discussed in the introduction (see also Exercise 12). ■

It is natural to wonder whether a given kernel function corresponds uniquely to a feature map. The answer is no, as we shall see by way of example.

■ **Example 6.5 (Feature Maps and Kernel Functions)** Let $\mathscr{X} = \mathbb{R}$ and consider feature maps $\phi_1 : \mathscr{X} \to \mathbb{R}$ and $\boldsymbol{\phi}_2 : \mathscr{X} \to \mathbb{R}^2$, with $\phi_1(x) := x$ and $\boldsymbol{\phi}_2(x) := [x, x]^\top/\sqrt{2}$. Then

$$\kappa_{\phi_1}(x, x') = \langle \phi_1(x), \phi_1(x') \rangle = xx',$$

but also

$$\kappa_{\boldsymbol{\phi}_2}(x, x') = \langle \boldsymbol{\phi}_2(x), \boldsymbol{\phi}_2(x') \rangle = xx'.$$

Thus, we arrive at the same kernel function defined for the same underlying set $\mathscr{X}$ via two different feature maps. ■

6.4.2 Kernels from Characteristic Functions

Another way to construct reproducing kernels on $\mathscr{X} = \mathbb{R}^p$ makes use of the properties of *characteristic functions*. In particular, we have the following result. We leave its proof as Exercise 10.

☞ 441

Theorem 6.3: Reproducing Kernel from a Characteristic Function

Let $\boldsymbol{X} \sim \mu$ be an $\mathbb{R}^p$-valued random vector that is symmetric about the origin (that is, $\boldsymbol{X}$ and $-\boldsymbol{X}$ are identically distributed), and let ψ be its characteristic function: $\psi(\boldsymbol{t}) = \mathbb{E}\, \mathrm{e}^{\mathrm{i}\boldsymbol{t}^\top \boldsymbol{X}} = \int \mathrm{e}^{\mathrm{i}\boldsymbol{t}^\top \boldsymbol{x}}\, \mu(\mathrm{d}\boldsymbol{x})$ for $\boldsymbol{t} \in \mathbb{R}^p$. Then $\kappa(\boldsymbol{x}, \boldsymbol{x}') := \psi(\boldsymbol{x} - \boldsymbol{x}')$ is a valid reproducing kernel on $\mathbb{R}^p$.

■ **Example 6.6 (Gaussian Kernel)** The multivariate normal distribution with mean vector $\mathbf{0}$ and covariance matrix $b^2\, \mathbf{I}_p$ is clearly symmetric around the origin. Its characteristic function is

$$\psi(\boldsymbol{t}) = \exp\left(-\frac{1}{2} b^2\, \|\boldsymbol{t}\|^2\right), \quad \boldsymbol{t} \in \mathbb{R}^p.$$

Taking $b^2 = 1/\sigma^2$, this gives the popular *Gaussian kernel* on $\mathbb{R}^p$:

GAUSSIAN KERNEL

$$\kappa(\boldsymbol{x}, \boldsymbol{x}') = \exp\left(-\frac{1}{2} \frac{\|\boldsymbol{x} - \boldsymbol{x}'\|^2}{\sigma^2}\right). \tag{6.15}$$

The parameter σ is sometimes called the *bandwidth*. Note that in the machine learning literature, the Gaussian kernel is sometimes referred to as "the" *radial basis function (rbf) kernel*.[1]

BANDWIDTH

RADIAL BASIS FUNCTION (RBF) KERNEL

From the proof of Theorem 6.2, we see that the RKHS $\mathscr{G}$ determined by the Gaussian kernel κ is the space of pointwise limits of functions of the form

$$g(\boldsymbol{x}) = \sum_{i=1}^{n} \alpha_i \exp\left(-\frac{1}{2} \frac{\|\boldsymbol{x} - \boldsymbol{x}_i\|^2}{\sigma^2}\right).$$

We can think of each point $\boldsymbol{x}_i$ having a feature $\kappa_{\boldsymbol{x}_i}$ that is a scaled multivariate Gaussian pdf centered at $\boldsymbol{x}_i$. ■

[1] The term radial basis function is sometimes used more generally to mean kernels of the form $\kappa(\boldsymbol{x}, \boldsymbol{x}') = f(\|\boldsymbol{x} - \boldsymbol{x}'\|)$ for some function $f : \mathbb{R} \to \mathbb{R}$.

■ **Example 6.7 (Sinc Kernel)** The characteristic function of a Uniform$[-1, 1]$ random variable (which is symmetric around 0) is $\psi(t) = \text{sinc}(t) := \sin(t)/t$, so $\kappa(x, x') = \text{sinc}(x-x')$ is a valid kernel. ■

☞ 131 Inspired by kernel density estimation (Section 4.4), we may be tempted to use the pdf of a random variable that is symmetric about the origin to construct a reproducing kernel. However, doing so will not work in general, as the next example illustrates.

■ **Example 6.8 (Uniform pdf Does not Construct a Valid Reproducing Kernel)** Take the function $\psi(t) = \frac{1}{2}\mathbb{1}\{|t| \leqslant 1\}$, which is the pdf of $X \sim$ Uniform$[-1, 1]$. Unfortunately, the function $\kappa(x, x') = \psi(x - x')$ is not positive semidefinite, as can be seen for example by constructing the matrix $\mathbf{A} = [\kappa(t_i, t_j), i, j = 1, 2, 3]$ for the points $t_1 = 0$, $t_2 = 0.75$, and $t_3 = 1.5$ as follows:

$$\mathbf{A} = \begin{pmatrix} \psi(0) & \psi(-0.75) & \psi(-1.5) \\ \psi(0.75) & \psi(0) & \psi(-0.75) \\ \psi(1.5) & \psi(0.75) & \psi(0) \end{pmatrix} = \begin{pmatrix} 0.5 & 0.5 & 0 \\ 0.5 & 0.5 & 0.5 \\ 0 & 0.5 & 0.5 \end{pmatrix}.$$

The eigenvalues of $\mathbf{A}$ are $\{1/2 - \sqrt{1/2}, 1/2, 1/2 + \sqrt{1/2}\} \approx \{-0.2071, 0.5, 1.2071\}$ and so ☞ 367 by Theorem A.9, $\mathbf{A}$ is not a positive semidefinite matrix, since it has a negative eigenvalue. Consequently, κ is not a valid reproducing kernel. ■

UNIVERSAL APPROXIMATION PROPERTY

One of the reasons why the Gaussian kernel (6.15) is popular is that it enjoys the *universal approximation property* [88]: the space of functions spanned by the Gaussian kernel is dense in the space of continuous functions with support $\mathscr{Z} \subset \mathbb{R}^p$. Naturally, this is a desirable property especially if there is little prior knowledge about the properties of g^*. However, note that *every* function g in the RKHS $\mathscr{G}$ associated with a Gaussian kernel κ is infinitely differentiable. Moreover, a Gaussian RKHS does not contain non-zero constant functions. Indeed, if $A \subset \mathscr{Z}$ is non-empty and open, then the only function of the form $g(\boldsymbol{x}) = c\,\mathbb{1}\{\boldsymbol{x} \in A\}$ contained in $\mathscr{G}$ is the zero function ($c = 0$).

MATÉRN KERNEL

Consequently, if it is known that g is differentiable only to a certain order, one may prefer the *Matérn kernel* with parameters $\nu, \sigma > 0$:

$$\kappa_\nu(\boldsymbol{x}, \boldsymbol{x}') = \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\sqrt{2\nu}\, \|\boldsymbol{x} - \boldsymbol{x}'\|/\sigma\right)^\nu K_\nu\left(\sqrt{2\nu}\, \|\boldsymbol{x} - \boldsymbol{x}'\|/\sigma\right), \tag{6.16}$$

which gives functions that are (weakly) differentiable to order $\lfloor \nu \rfloor$ (but not necessarily to order $\lceil \nu \rceil$). Here, K_ν denotes the modified Bessel function of the second kind; see (4.49). ☞ 163 The particular form of the Matérn kernel appearing in (6.16) ensures that $\lim_{\nu\to\infty} \kappa_\nu(\boldsymbol{x}, \boldsymbol{x}') = \kappa(\boldsymbol{x}, \boldsymbol{x}')$, where κ is the Gaussian kernel appearing in (6.15).

We remark that Sobolev spaces are closely related to the Matérn kernel. Up to constants (which scale the unit ball in the space), in dimension p and for a parameter $s > p/2$, these spaces can be identified with $\psi(\boldsymbol{t}) = \frac{2^{1-s}}{\Gamma(s)}\|\boldsymbol{t}\|^{s-p/2} K_{p/2-s}(\|\boldsymbol{t}\|)$, which in turn can be viewed as the characteristic function corresponding to the (radially symmetric) multivariate Student's ☞ 162 t distribution with s degrees of freedom: that is, with pdf $f(\boldsymbol{x}) \propto (1 + \|\boldsymbol{x}\|^2)^{-s}$.

6.4.3 Reproducing Kernels Using Orthonormal Features

We have seen in Sections 6.4.1 and 6.4.2 how to construct reproducing kernels from feature maps and characteristic functions. Another way to construct kernels on a space $\mathcal{X}$ is to work directly from the function class $L^2(\mathcal{X}; \mu)$; that is, the set of square-integrable[2] functions on $\mathcal{X}$ with respect to μ; see also Definition A.4. ☞ 385 For simplicity, in what follows, we will consider μ to be the Lebesgue measure, and will simply write $L^2(\mathcal{X})$ rather than $L^2(\mathcal{X}; \mu)$. We will also assume that $\mathcal{X} \subseteq \mathbb{R}^p$.

Let $\{\xi_1, \xi_2, \ldots\}$ be an orthonormal basis of $L^2(\mathcal{X})$ and let $c_1, c_2, \ldots$ be a sequence of positive numbers. As discussed in Section 6.4.1, the kernel corresponding to a feature map $\boldsymbol{\phi} : \mathcal{X} \to \mathbb{R}^p$ is $\kappa(\boldsymbol{x}, \boldsymbol{x}') = \boldsymbol{\phi}(\boldsymbol{x})^\top \boldsymbol{\phi}(\boldsymbol{x}') = \sum_{i=1}^p \phi_i(\boldsymbol{x})\, \phi_i(\boldsymbol{x}')$. Now consider a (possibly infinite) sequence of feature functions $\phi_i = c_i\, \xi_i, i = 1, 2, \ldots$ and define

$$\kappa(\boldsymbol{x}, \boldsymbol{x}') := \sum_{i \geqslant 1} \phi_i(\boldsymbol{x})\, \phi_i(\boldsymbol{x}') = \sum_{i \geqslant 1} \lambda_i\, \xi_i(\boldsymbol{x})\, \xi_i(\boldsymbol{x}'), \tag{6.17}$$

where $\lambda_i = c_i^2, i = 1, 2, \ldots$. This is well-defined as long as $\sum_{i \geqslant 1} \lambda_i < \infty$, which we assume from now on. Let $\mathcal{H}$ be the linear space of functions of the form $f = \sum_{i \geqslant 1} \alpha_i \xi_i$, where $\sum_{i \geqslant 1} \alpha_i^2 / \lambda_i < \infty$. As every function $f \in L^2(\mathcal{X})$ can be represented as $f = \sum_{i \geqslant 1} \langle f, \xi_i \rangle \xi_i$, we see that $\mathcal{H}$ is a linear subspace of $L^2(\mathcal{X})$. On $\mathcal{H}$ define the inner product

$$\langle f, g \rangle_{\mathcal{H}} := \sum_{i \geqslant 1} \frac{\langle f, \xi_i \rangle \langle g, \xi_i \rangle}{\lambda_i}.$$

With this inner product, the squared norm of $f = \sum_{i \geqslant 1} \alpha_i\, \xi_i$ is $\|f\|_{\mathcal{H}}^2 = \sum_{i \geqslant 1} \alpha_i^2 / \lambda_i < \infty$. We show that $\mathcal{H}$ is actually an RKHS with kernel κ by verifying the conditions of Definition 6.1. First,

$$\kappa_{\boldsymbol{x}} = \sum_{i \geqslant 1} \lambda_i\, \xi_i(\boldsymbol{x})\, \xi_i \in \mathcal{H},$$

as $\sum_i \lambda_i < \infty$ by assumption, and so κ is finite. Second, the reproducing property holds. Namely, let $f = \sum_{i \geqslant 1} \alpha_i\, \xi_i$. Then,

$$\langle \kappa_{\boldsymbol{x}}, f \rangle_{\mathcal{H}} = \sum_{i \geqslant 1} \frac{\langle \kappa_{\boldsymbol{x}}, \xi_i \rangle \langle f, \xi_i \rangle}{\lambda_i} = \sum_{i \geqslant 1} \frac{\lambda_i\, \xi_i(\boldsymbol{x})\, \alpha_i}{\lambda_i} = \sum_{i \geqslant 1} \alpha_i \xi_i(\boldsymbol{x}) = f(\boldsymbol{x}).$$

The discussion above demonstrates that kernels can be constructed via (6.17). In fact, (under mild conditions) any given reproducing kernel κ can be written in the form (6.17), where this series representation enjoys desirable convergence properties. This result is known as Mercer's theorem, and is given below. We leave the full proof including the precise conditions to, e.g., [40], but the main idea is that a reproducing kernel κ can be thought of as a generalization of a positive semidefinite matrix $\mathbf{K}$, and can also be written in spectral form (see also Section A.6.5). ☞ 367 In particular, by Theorem A.9, we can write $\mathbf{K} = \mathbf{V}\mathbf{D}\mathbf{V}^\top$, where $\mathbf{V}$ is a matrix of orthonormal eigenvectors $[\boldsymbol{v}_\ell]$ and $\mathbf{D}$ the diagonal matrix of the (positive) eigenvalues $[\lambda_\ell]$; that is,

$$\mathbf{K}(i, j) = \sum_{\ell \geqslant 1} \lambda_\ell\, v_\ell(i)\, v_\ell(j).$$

[2] A function $f : \mathcal{X} \to \mathbb{R}$ is said to be square-integrable if $\int f^2(\boldsymbol{x})\, \mu(\mathrm{d}\boldsymbol{x}) < \infty$, where μ is a measure on $\mathcal{X}$.

In (6.18) below, $\boldsymbol{x}, \boldsymbol{x}'$ play the role of i, j, and ξ_ℓ plays the role of $\boldsymbol{v}_\ell$.

Theorem 6.4: Mercer

Let $\kappa : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a reproducing kernel for a compact set $\mathcal{X} \subset \mathbb{R}^p$. Then (under mild conditions) there exists a countable sequence of non-negative numbers $\{\lambda_\ell\}$ decreasing to zero and functions $\{\xi_\ell\}$ orthonormal in $L^2(\mathcal{X})$ such that

$$\kappa(\boldsymbol{x}, \boldsymbol{x}') = \sum_{\ell \geqslant 1} \lambda_\ell \, \xi_\ell(\boldsymbol{x}) \, \xi_\ell(\boldsymbol{x}') , \qquad \text{for all } \ \boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}, \tag{6.18}$$

where (6.18) converges absolutely and uniformly on $\mathcal{X} \times \mathcal{X}$.

Further, if $\lambda_\ell > 0$, then (λ_ℓ, ξ_ℓ) is an (eigenvalue, eigenfunction) pair for the integral operator $K : L^2(\mathcal{X}) \to L^2(\mathcal{X})$ defined by $[Kf](\boldsymbol{x}) := \int_{\mathcal{X}} \kappa(\boldsymbol{x}, \boldsymbol{y}) f(\boldsymbol{y}) \, \mathrm{d}\boldsymbol{y}$ for $\boldsymbol{x} \in \mathcal{X}$.

Theorem 6.4 holds if (i) the kernel κ is continuous on $\mathcal{X} \times \mathcal{X}$, (ii) the function $\widetilde{\kappa}(\boldsymbol{x}) := \kappa(\boldsymbol{x}, \boldsymbol{x})$ defined for $\boldsymbol{x} \in \mathcal{X}$ is integrable. Extensions of Theorem 6.4 to more general spaces $\mathcal{X}$ and measures μ hold; see, e.g., [115] or [40].

The key importance of Theorem 6.4 lies in the fact that the series representation (6.18) converges absolutely and uniformly on $\mathcal{X} \times \mathcal{X}$. The uniform convergence is a much stronger condition than pointwise convergence, and means for instance that properties of the sequence of partial sums, such as continuity and integrability, are transferred to the limit.

■ **Example 6.9 (Mercer)** Suppose $\mathcal{X} = [-1, 1]$ and the kernel is $\kappa(x, x') = 1 + xx'$ which corresponds to the RKHS $\mathcal{G}$ of affine functions from $\mathcal{X} \to \mathbb{R}$. To find the (eigenvalue, eigenfunction) pairs for the integral operator appearing in Theorem 6.4, we need to find numbers $\{\lambda_\ell\}$ and orthonormal functions $\{\xi_\ell(x)\}$ that solve

$$\int_{-1}^{1} (1 + xx') \, \xi_\ell(x') \, \mathrm{d}x' = \lambda_\ell \, \xi_\ell(x) , \quad \text{for all } x \in [-1, 1].$$

Consider first a constant function $\xi_1(x) = c$. Then, for all $x \in [-1, 1]$, we have that $2c = \lambda_1 c$, and the normalization condition requires that $\int_{-1}^{1} c^2 \, \mathrm{d}x = 1$. Together, these give $\lambda_1 = 2$ and $c = \pm 1/\sqrt{2}$. Next, consider an affine function $\xi_2(x) = a + bx$. Orthogonality requires that

$$\int_{-1}^{1} c(a + bx) \, \mathrm{d}x = 0,$$

which implies $a = 0$ (since $c \neq 0$). Moreover, the normalization condition then requires

$$\int_{-1}^{1} b^2 x^2 \, \mathrm{d}x = 1,$$

or, equivalently, $2b^2/3 = 1$, implying $b = \pm\sqrt{3/2}$. Finally, the integral equation reads

$$\int_{-1}^{1} (1 + xx') \, bx' \, \mathrm{d}x' = \lambda_2 \, bx \iff \frac{2bx}{3} = \lambda_2 bx,$$

implying that $\lambda_2 = 2/3$. We take the positive solutions (i.e., $c > 0$ and $b > 0$), and note that

$$\lambda_1 \xi_1(x)\xi_1(x') + \lambda_2 \xi_2(x)\xi_2(x') = 2\frac{1}{\sqrt{2}}\frac{1}{\sqrt{2}} + \frac{2}{3}\frac{\sqrt{3}}{\sqrt{2}}x\frac{\sqrt{3}}{\sqrt{2}}x' = 1 + xx' = \kappa(x, x'),$$

and so we have found the decomposition appearing in (6.18). As an aside, observe that ξ_1 and ξ_2 are orthonormal versions of the first two Legendre polynomials. The corresponding feature map can be explicitly identified as $\phi_1(x) = \sqrt{\lambda_1}\,\xi_1(x) = 1$ and $\phi_2(x) = \sqrt{\lambda_2}\,\xi_2(x) = x$. ■

☞ 387

6.4.4 Kernels from Kernels

The following theorem lists some useful properties for constructing reproducing kernels from existing reproducing kernels.

Theorem 6.5: Rules for Constructing Kernels from Other Kernels

1. If $\kappa : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$ is a reproducing kernel and $\boldsymbol{\phi} : \mathcal{X} \to \mathbb{R}^p$ is a function, then $\kappa(\boldsymbol{\phi}(\boldsymbol{x}), \boldsymbol{\phi}(\boldsymbol{x}'))$ is a reproducing kernel from $\mathcal{X} \times \mathcal{X} \to \mathbb{R}$.
2. If $\kappa : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a reproducing kernel and $f : \mathcal{X} \to \mathbb{R}_+$ is a function, then $f(\boldsymbol{x})\kappa(\boldsymbol{x}, \boldsymbol{x}')f(\boldsymbol{x}')$ is also a reproducing kernel from $\mathcal{X} \times \mathcal{X} \to \mathbb{R}$.
3. If κ_1 and κ_2 are reproducing kernels from $\mathcal{X} \times \mathcal{X} \to \mathbb{R}$, then so is their sum $\kappa_1 + \kappa_2$.
4. If κ_1 and κ_2 are reproducing kernels from $\mathcal{X} \times \mathcal{X} \to \mathbb{R}$, then so is their product $\kappa_1\kappa_2$.
5. If κ_1 and κ_2 are reproducing kernels from $\mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and $\mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ respectively, then $\kappa_+((\boldsymbol{x}, \boldsymbol{y}), (\boldsymbol{x}', \boldsymbol{y}')) := \kappa_1(\boldsymbol{x}, \boldsymbol{x}') + \kappa_2(\boldsymbol{y}, \boldsymbol{y}')$ and $\kappa_\times((\boldsymbol{x}, \boldsymbol{y}), (\boldsymbol{x}', \boldsymbol{y}')) := \kappa_1(\boldsymbol{x}, \boldsymbol{x}')\kappa_2(\boldsymbol{y}, \boldsymbol{y}')$ are reproducing kernels from $(\mathcal{X} \times \mathcal{Y}) \times (\mathcal{X} \times \mathcal{Y}) \to \mathbb{R}$.

Proof: For Rules 1, 2, and 3 it is easy to verify that the resulting function is finite, symmetric, and positive semidefinite, and so is a valid reproducing kernel by Theorem 6.2. For example, for Rule 1 we have $\sum_{i=1}^n \sum_{j=1}^n \alpha_i\, \kappa(\boldsymbol{y}_i, \boldsymbol{y}_j)\alpha_j \geqslant 0$ for every choice of $\{\alpha_i\}_{i=1}^n$ and $\{\boldsymbol{y}_i\}_{i=1}^n \in \mathbb{R}^p$, since κ is a reproducing kernel. In particular, it holds true for $\boldsymbol{y}_i = \boldsymbol{\phi}(\boldsymbol{x}_i)$, $i = 1, \ldots, n$. Rule 4 is easy to show for kernels κ_1, κ_2 that admit a representation of the form (6.17), since

$$\begin{aligned}
\kappa_1(\boldsymbol{x}, \boldsymbol{x}')\,\kappa_2(\boldsymbol{x}, \boldsymbol{x}') &= \left(\sum_{i\geqslant 1} \phi_i^{(1)}(\boldsymbol{x})\,\phi_i^{(1)}(\boldsymbol{x}')\right)\left(\sum_{j\geqslant 1} \phi_j^{(2)}(\boldsymbol{x})\,\phi_j^{(2)}(\boldsymbol{x}')\right) \\
&= \sum_{i,j\geqslant 1} \phi_i^{(1)}(\boldsymbol{x})\,\phi_j^{(2)}(\boldsymbol{x})\,\phi_i^{(1)}(\boldsymbol{x}')\,\phi_j^{(2)}(\boldsymbol{x}') \\
&= \sum_{k\geqslant 1} \phi_k(\boldsymbol{x})\,\phi_k(\boldsymbol{x}') =: \kappa(\boldsymbol{x}, \boldsymbol{x}'),
\end{aligned}$$

showing that $\kappa = \kappa_1\kappa_2$ also admits a representation of the form (6.17), where the new (possibly infinite) sequence of features (ϕ_k) is identified in a one-to-one way with the sequence $(\phi_i^{(1)}\phi_j^{(2)})$. We leave the proof of rule 5 as an exercise (Exercise 8). □

■ **Example 6.10 (Polynomial Kernel)** Consider $\boldsymbol{x}, \boldsymbol{x}' \in \mathbb{R}^2$ with

$$\kappa(\boldsymbol{x}, \boldsymbol{x}') = (1 + \langle \boldsymbol{x}, \boldsymbol{x}' \rangle)^2,$$

POLYNOMIAL KERNEL

where $\langle \boldsymbol{x}, \boldsymbol{x}' \rangle = \boldsymbol{x}^\top \boldsymbol{x}'$. This is an example of a *polynomial kernel*. Combining the fact that sums and products of kernels are again kernels (rules 3 and 4 of Theorem 6.5), we find that, since $\langle \boldsymbol{x}, \boldsymbol{x}' \rangle$ and the constant function 1 are kernels, so are $1 + \langle \boldsymbol{x}, \boldsymbol{x}' \rangle$ and $(1 + \langle \boldsymbol{x}, \boldsymbol{x}' \rangle)^2$. By writing

$$\begin{aligned}\kappa(\boldsymbol{x}, \boldsymbol{x}') &= (1 + x_1 x_1' + x_2 x_2')^2 \\ &= 1 + 2x_1 x_1' + 2x_2 x_2' + 2x_1 x_2 x_1' x_2' + (x_1 x_1')^2 + (x_2 x_2')^2,\end{aligned}$$

we see that $\kappa(\boldsymbol{x}, \boldsymbol{x}')$ can be written as the inner product in $\mathbb{R}^6$ of the two feature vectors $\boldsymbol{\phi}(\boldsymbol{x})$ and $\boldsymbol{\phi}(\boldsymbol{x}')$, where the feature map $\boldsymbol{\phi} : \mathbb{R}^2 \to \mathbb{R}^6$ can be explicitly identified as

$$\boldsymbol{\phi}(\boldsymbol{x}) = [1,\ \sqrt{2}x_1,\ \sqrt{2}x_2,\ \sqrt{2}x_1 x_2, x_1^2, x_2^2]^\top.$$

Thus, the RKHS determined by κ can be explicitly identified with the space of functions $\boldsymbol{x} \mapsto \boldsymbol{\phi}(\boldsymbol{x})^\top \boldsymbol{\beta}$ for some $\boldsymbol{\beta} \in \mathbb{R}^6$. ■

In the above example we could explicitly identify the feature map. However, in general a feature map need not be explicitly available. Using a particular reproducing kernel corresponds to using an *implicit* (possibly infinite dimensional!) feature map that never needs to be explicitly computed.

6.5 Representer Theorem

Recall the setting discussed at the beginning of this chapter: we are given training data $\tau = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ and a loss function that measures the fit to the data, and we wish to find a function g that minimizes the training loss, with the addition of a regularization term, as described in Section 6.2. To do this, we assume first that the class $\mathcal{G}$ of prediction functions can be decomposed as the direct sum of an RKHS $\mathcal{H}$, defined by a kernel function $\kappa : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, and another linear space of real-valued functions $\mathcal{H}_0$ on $\mathcal{X}$; that is,

$$\mathcal{G} = \mathcal{H} \oplus \mathcal{H}_0,$$

meaning that any element $g \in \mathcal{G}$ can be written as $g = h + h_0$, with $h \in \mathcal{H}$ and $h_0 \in \mathcal{H}_0$. In minimizing the training loss we wish to penalize the h term of g but not the h_0 term. Specifically, the aim is to solve the functional optimization problem

$$\min_{g \in \mathcal{H} \oplus \mathcal{H}_0} \frac{1}{n} \sum_{i=1}^n \mathrm{Loss}(y_i, g(\boldsymbol{x}_i)) + \gamma \, \|g\|_{\mathcal{H}}^2. \tag{6.19}$$

Here, we use a slight abuse of notation: $\|g\|_{\mathcal{H}}$ means $\|h\|_{\mathcal{H}}$ if $g = h + h_0$, as above. In this way, we can view $\mathcal{H}_0$ as the null space of the functional $g \mapsto \|g\|_{\mathcal{H}}$. This null space may be empty, but typically has a small dimension m; for example it could be the one-dimensional space of constant functions, as in Example 6.2. ☞ 217

■ **Example 6.11 (Null Space)** Consider again the setting of Example 6.2, for which we have feature vectors $\widetilde{\boldsymbol{x}} = [1, \boldsymbol{x}^\top]^\top$ and $\mathcal{G}$ consists of functions of the form $g : \widetilde{\boldsymbol{x}} \mapsto \beta_0 + \boldsymbol{x}^\top \boldsymbol{\beta}$. Each function g can be decomposed as $g = h + h_0$, where $h : \widetilde{\boldsymbol{x}} \mapsto \boldsymbol{x}^\top \boldsymbol{\beta}$, and $h_0 : \widetilde{\boldsymbol{x}} \mapsto \beta_0$.

Given $g \in \mathcal{G}$, we have $\|g\|_{\mathcal{H}} = \|\boldsymbol{\beta}\|$, and so the null space $\mathcal{H}_0$ of the functional $g \mapsto \|g\|_{\mathcal{H}}$ (that is, the set of all functions $g \in \mathcal{G}$ for which $\|g\|_{\mathcal{H}} = 0$) is the set of constant functions here, which has dimension $m = 1$. ■

Regularization favors elements in $\mathcal{H}_0$ and penalizes large elements in $\mathcal{H}$. As the regularization parameter γ varies between zero and infinity, solutions to (6.19) vary from "complex" ($g \in \mathcal{H} \oplus \mathcal{H}_0$) to "simple" ($g \in \mathcal{H}_0$).

A key reason why RKHSs are so useful is the following. By choosing $\mathcal{H}$ to be an RKHS in (6.19) this *functional* optimization problem effectively becomes a *parametric* optimization problem. The reason is that any solution to (6.19) can be represented as a finite-dimensional linear combination of kernel functions, evaluated at the training sample. This is known as the *kernel trick*.

KERNEL TRICK

Theorem 6.6: Representer Theorem

The solution to the penalized optimization problem (6.19) is of the form

$$g(\boldsymbol{x}) = \sum_{i=1}^{n} \alpha_i \, \kappa(\boldsymbol{x}_i, \boldsymbol{x}) + \sum_{j=1}^{m} \eta_j \, q_j(\boldsymbol{x}), \tag{6.20}$$

where $\{q_1, \ldots, q_m\}$ is a basis of $\mathcal{H}_0$.

Proof: Let $\mathcal{F} = \operatorname{Span}\{\kappa_{\boldsymbol{x}_i}, i = 1, \ldots, n\}$. Clearly, $\mathcal{F} \subseteq \mathcal{H}$. Then, the Hilbert space $\mathcal{H}$ can be represented as $\mathcal{H} = \mathcal{F} \oplus \mathcal{F}^\perp$, where $\mathcal{F}^\perp$ is the orthogonal complement of $\mathcal{F}$. In other words, $\mathcal{F}^\perp$ is the class of functions

$$\{f^\perp \in \mathcal{H} : \langle f^\perp, f \rangle_{\mathcal{H}} = 0, \ f \in \mathcal{F}\} \equiv \{f^\perp : \langle f^\perp, \kappa_{\boldsymbol{x}_i} \rangle_{\mathcal{H}} = 0, \ \forall i\}.$$

It follows, by the reproducing kernel property, that for all $f^\perp \in \mathcal{F}^\perp$:

$$f^\perp(\boldsymbol{x}_i) = \langle f^\perp, \kappa_{\boldsymbol{x}_i} \rangle_{\mathcal{H}} = 0, \quad i = 1, \ldots, n.$$

Now, take any $g \in \mathcal{H} \oplus \mathcal{H}_0$, and write it as $g = f + f^\perp + h_0$, with $f \in \mathcal{F}$, $f^\perp \in \mathcal{F}^\perp$, and $h_0 \in \mathcal{H}_0$. By the definition of the null space $\mathcal{H}_0$, we have $\|g\|_{\mathcal{H}}^2 = \|f + f^\perp\|_{\mathcal{H}}^2$. Moreover, by Pythagoras' theorem, the latter is equal to $\|f\|_{\mathcal{H}}^2 + \|f^\perp\|_{\mathcal{H}}^2$. It follows that

$$\begin{aligned}
\frac{1}{n} \sum_{i=1}^{n} \operatorname{Loss}(y_i, g(\boldsymbol{x}_i)) + \gamma \|g\|_{\mathcal{H}}^2 &= \frac{1}{n} \sum_{i=1}^{n} \operatorname{Loss}(y_i, f(\boldsymbol{x}_i) + h_0(\boldsymbol{x}_i)) + \gamma \left(\|f\|_{\mathcal{H}}^2 + \|f^\perp\|_{\mathcal{H}}^2 \right) \\
&\geqslant \frac{1}{n} \sum_{i=1}^{n} \operatorname{Loss}(y_i, f(\boldsymbol{x}_i) + h_0(\boldsymbol{x}_i)) + \gamma \, \|f\|_{\mathcal{H}}^2.
\end{aligned}$$

Since we can obtain equality by taking $f^\perp = 0$, this implies that the minimizer of the penalized optimization problem (6.19) lies in the subspace $\mathcal{F} \oplus \mathcal{H}_0$ of $\mathcal{G} = \mathcal{H} \oplus \mathcal{H}_0$, and hence is of the form (6.20). □

Substituting the representation (6.20) of g into (6.19) gives the finite-dimensional optimization problem:

$$\min_{\boldsymbol{\alpha}\in\mathbb{R}^n,\,\boldsymbol{\eta}\in\mathbb{R}^m} \frac{1}{n}\sum_{i=1}^{n} \mathrm{Loss}(y_i, (\mathbf{K}\boldsymbol{\alpha} + \mathbf{Q}\boldsymbol{\eta})_i) + \gamma\, \boldsymbol{\alpha}^\top \mathbf{K}\boldsymbol{\alpha}, \tag{6.21}$$

where

- $\mathbf{K}$ is the $n \times n$ (Gram) matrix with entries $[\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j), i = 1, \ldots, n,\ j = 1, \ldots, n]$.
- $\mathbf{Q}$ is the $n \times m$ matrix with entries $[q_j(\boldsymbol{x}_i), i = 1, \ldots, n,\ j = 1, \ldots, m]$.

In particular, for the squared-error loss we have

$$\min_{\boldsymbol{\alpha}\in\mathbb{R}^n,\,\boldsymbol{\eta}\in\mathbb{R}^m} \frac{1}{n}\left\|\boldsymbol{y} - (\mathbf{K}\boldsymbol{\alpha} + \mathbf{Q}\boldsymbol{\eta})\right\|^2 + \gamma\, \boldsymbol{\alpha}^\top \mathbf{K}\boldsymbol{\alpha}. \tag{6.22}$$

This is a convex optimization problem, and its solution is found by differentiating (6.22) with respect to $\boldsymbol{\alpha}$ and $\boldsymbol{\eta}$ and equating to zero, leading to the following system of $(n + m)$ linear equations:

$$\begin{bmatrix} \mathbf{K}\mathbf{K}^\top + n\gamma\mathbf{K} & \mathbf{K}\mathbf{Q} \\ \mathbf{Q}^\top\mathbf{K}^\top & \mathbf{Q}^\top\mathbf{Q} \end{bmatrix} \begin{bmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\eta} \end{bmatrix} = \begin{bmatrix} \mathbf{K}^\top \\ \mathbf{Q}^\top \end{bmatrix} \boldsymbol{y}. \tag{6.23}$$

As long as $\mathbf{Q}$ is of full column rank, the minimizing function is unique.

■ **Example 6.12 (Ridge Regression (cont.))** We return to Example 6.2 and identify that $\mathcal{H}$ is the RKHS with linear kernel function $\kappa(\boldsymbol{x}, \boldsymbol{x}') = \boldsymbol{x}^\top\boldsymbol{x}'$ and $\mathcal{C} = \mathcal{H}_0$ is the linear space of constant functions. In this case, $\mathcal{H}_0$ is spanned by the function $q_1 \equiv 1$. Moreover, $\mathbf{K} = \mathbf{X}\mathbf{X}^\top$ and $\mathbf{Q} = \mathbf{1}$.

If we appeal to the representer theorem directly, then the problem in (6.6) becomes, as a result of (6.21):

$$\min_{\boldsymbol{\alpha},\eta_0} \frac{1}{n}\left\|\boldsymbol{y} - \eta_0\,\mathbf{1} - \mathbf{X}\mathbf{X}^\top\boldsymbol{\alpha}\right\|^2 + \gamma\,\|\mathbf{X}^\top\boldsymbol{\alpha}\|^2.$$

This is a convex optimization problem, and so the solution follows by taking derivatives and setting them to zero. This gives the equations

$$\mathbf{X}\mathbf{X}^\top((\mathbf{X}\mathbf{X}^\top + n\gamma\,\mathbf{I}_n)\,\boldsymbol{\alpha} + \eta_0\,\mathbf{1} - \boldsymbol{y}) = 0,$$

and

$$n\,\eta_0 = \mathbf{1}^\top(\boldsymbol{y} - \mathbf{X}\mathbf{X}^\top\boldsymbol{\alpha}).$$

Note that these are equivalent to (6.8) and (6.9) (once again assuming that $n \geqslant p$ and $\mathbf{X}$ has full rank p). Equivalently, the solution is found by solving (6.23):

$$\begin{bmatrix} \mathbf{X}\mathbf{X}^\top\mathbf{X}\mathbf{X}^\top + n\gamma\,\mathbf{X}\mathbf{X}^\top & \mathbf{X}\mathbf{X}^\top\mathbf{1} \\ \mathbf{1}^\top\mathbf{X}\mathbf{X}^\top & n \end{bmatrix} \begin{bmatrix} \boldsymbol{\alpha} \\ \eta_0 \end{bmatrix} = \begin{bmatrix} \mathbf{X}\mathbf{X}^\top \\ \mathbf{1}^\top \end{bmatrix} \boldsymbol{y}.$$

This is a system of $(n + 1)$ linear equations, and is typically of much larger dimension than the $(p + 1)$ linear equations given by (6.8) and (6.9). As such, one may question the practicality of reformulating the problem in this way. However, the benefit of this formulation is that the problem can be expressed entirely through the Gram matrix $\mathbf{K}$, without having to explicitly compute the feature vectors — in turn permitting the (implicit) use of infinite dimensional feature spaces. ■

■ **Example 6.13 (Estimating the Peaks Function)** Figure 6.4 shows the surface plot of the *peaks* function:

$$f(x_1, x_2) = 3(1 - x_1)^2 \mathrm{e}^{-x_1^2-(x_2+1)^2} - 10\left(\frac{x_1}{5} - x_1^3 - x_2^5\right)\mathrm{e}^{-x_1^2-x_2^2} - \frac{1}{3}\mathrm{e}^{-(x_1+1)^2-x_2^2}. \qquad (6.24)$$

The goal is to learn the function $y = f(\boldsymbol{x})$ based on a small set of training data (pairs of $(\boldsymbol{x}, y)$ values). The red dots in the figure represent data $\tau = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{20}$, where $y_i = f(\boldsymbol{x}_i)$ and the $\{\boldsymbol{x}_i\}$ have been chosen in a *quasi-random* way, using *Hammersley points* (with bases 2 and 3) on the square $[-3, 3]^2$. Quasi-random point sets have better space-filling properties than either a regular grid of points or a set of pseudo-random points. We refer to [71] for details. Note that there is no observation noise in this particular problem.

QUASI-RANDOM

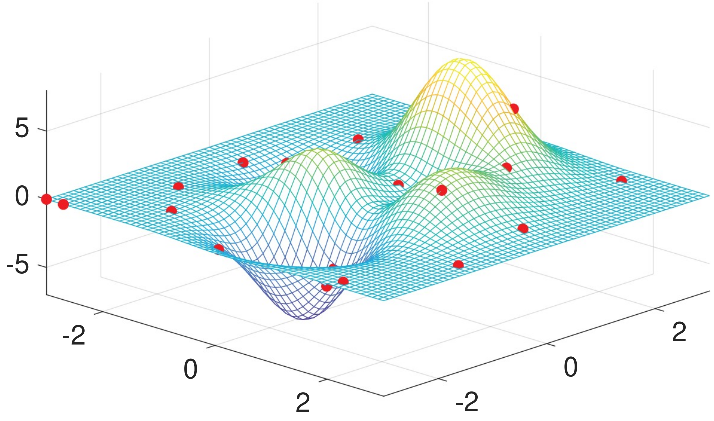

Figure 6.4: Peaks function sampled at 20 Hammersley points.

The purpose of this example is to illustrate how, using the small data set of size $n = 20$, the entire *peaks* function can be approximated well using kernel methods. In particular, we use the Gaussian kernel (6.15) on $\mathbb{R}^2$, and denote by $\mathscr{H}$ the unique RKHS corresponding to this kernel. We omit the regularization term in (6.19), and thus our objective is to find the solution to

$$\min_{g \in \mathscr{H}} \frac{1}{n} \sum_{i=1}^{n} (y_i - g(\boldsymbol{x}_i))^2.$$

By the representer theorem, the optimal function is of the form

$$g(\boldsymbol{x}) = \sum_{i=1}^{n} \alpha_i \exp\left(-\frac{1}{2}\frac{\|\boldsymbol{x} - \boldsymbol{x}_i\|^2}{\sigma^2}\right),$$

where $\boldsymbol{\alpha} := [\alpha_1, \ldots, \alpha_n]^\top$ is, by (6.23), the solution to the set of linear equations $\mathbf{K}\mathbf{K}^\top\boldsymbol{\alpha} = \mathbf{K}\boldsymbol{y}$.

Note that we are performing regression over the class of functions $\mathscr{H}$ with an implicit feature space. Due to the representer theorem, the solution to this problem coincides with the solution to the linear regression problem for which the i-th feature (for $i = 1, \ldots, n$) is chosen to be the vector $[\kappa(\boldsymbol{x}_1, \boldsymbol{x}_i), \ldots, \kappa(\boldsymbol{x}_n, \boldsymbol{x}_i)]^\top$.

The following code performs these calculations and gives the contour plots of g and the `peaks` functions, shown in Figure 6.5. We see that the two are quite close. Code for the generation of Hammersley points is available from the book's GitHub site as `genham.py`.

peakskernel.py

```
from genham import hammersley
import numpy as np
import matplotlib.pyplot as plt
from mpl_toolkits.mplot3d import Axes3D
from matplotlib import cm
from numpy.linalg import norm

import numpy as np
def peaks(x,y):
    z =  (3*(1-x)**2 * np.exp(-(x**2) - (y+1)**2)
          - 10*(x/5 - x**3 - y**5) * np.exp(-x**2 - y**2)
          - 1/3 * np.exp(-(x+1)**2 - y**2))
    return(z)

n = 20
x = -3 + 6*hammersley([2,3],n)
z = peaks(x[:,0],x[:,1])
xx, yy = np.mgrid[-3:3:150j,-3:3:150j]
zz = peaks(xx,yy)
plt.contour(xx,yy,zz,levels=50)

fig=plt.figure()
ax = fig.add_subplot(111,projection='3d')
ax.plot_surface(xx,yy,zz,rstride=1,cstride=1,color='c',alpha=0.3,
   linewidth=0)
ax.scatter(x[:,0],x[:,1],z,color='k',s=20)
plt.show()

sig2 = 0.3 # kernel parameter
def k(x,u):
    return(np.exp(-0.5*norm(x- u)**2/sig2))
K = np.zeros((n,n))
for i in range(n):
    for j in range(n):
        K[i,j] = k(x[i,:],x[j])
alpha = np.linalg.solve(K@K.T, K@z)



N, = xx.flatten().shape
Kx = np.zeros((n,N))
for i in range(n):
    for j in range(N):
        Kx[i,j] = k(x[i,:],np.array([xx.flatten()[j],yy.flatten()[j
            ]]))

g = Kx.T @ alpha
dim = np.sqrt(N).astype(int)
yhat = g.reshape(dim,dim)
plt.contour(xx,yy,yhat,levels=50)
```

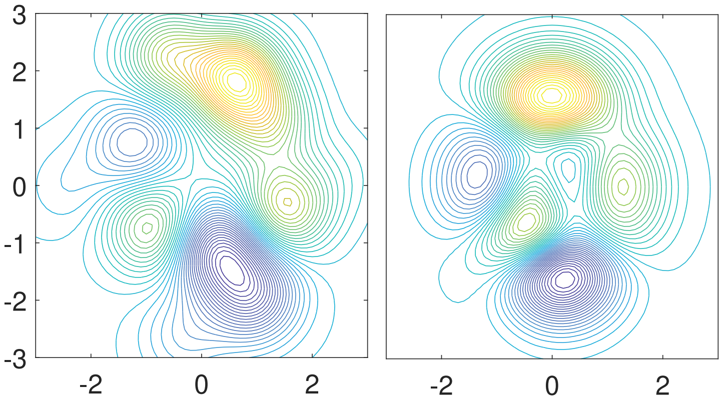

Figure 6.5: Contour plots for the prediction function g (left) and the *peaks* function given in (6.24) (right).

6.6 Smoothing Cubic Splines

A striking application of kernel methods is to fitting "well-behaved" functions to data. Key examples of "well-behaved" functions are those that do not have large second-order derivatives. Consider functions $g : [0, 1] \to \mathbb{R}$ that are twice differentiable and define $\|g''\|^2 := \int_0^1 (g'')^2 \, \mathrm{d}x$ as a measure of the size of the second derivative.

■ **Example 6.14 (Behavior of $\|g''\|^2$)** Intuitively, the larger $\|g''\|^2$ is, the more "wiggly" the function g will be. As an explicit example, consider $g(x) = \sin(\omega x)$ for $x \in [0, 1]$, where ω is a free parameter. We can explicitly compute $g''(x) = -\omega^2 \sin(\omega x)$, and consequently

$$\|g''\|^2 = \int_0^1 \omega^4 \sin^2(\omega x) \, \mathrm{d}x = \frac{\omega^4}{2} (1 - \mathrm{sinc}(2\omega)) .$$

As $|\omega| \to \infty$, the frequency of g increases and we have $\|g''\|^2 \to \infty$. ■

Now, in the context of data fitting, consider the following penalized least-squares optimization problem on $[0, 1]$:

$$\min_{g \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^{n} (y_i - g(x_i))^2 + \gamma \, \|g''\|^2, \tag{6.25}$$

where we will specify $\mathcal{G}$ in what follows. In order to apply the kernel machinery, we want to write this in the form (6.19), for some RKHS $\mathcal{H}$ and null space $\mathcal{H}_0$. Clearly, the norm on $\mathcal{H}$ should be of the form $\|g\|_{\mathcal{H}} = \|g''\|$ and should be well-defined (i.e., finite and ensuring g and g' are absolutely continuous). This suggests that we take

$$\mathcal{H} = \{g \in L^2[0, 1] : \|g''\| < \infty, g, g' \text{ absolutely continuous}, \; g(0) = g'(0) = 0\},$$

with inner product

$$\langle f, g \rangle_{\mathcal{H}} := \int_0^1 f''(x) \, g''(x) \, \mathrm{d}x.$$

One rationale for imposing the boundary conditions $g(0) = g'(0) = 0$ is as follows: when expanding g about the point $x = 0$, Taylor's theorem (with integral remainder term) states that

$$g(x) = g(0) + g'(0)\,x + \int_0^x g''(s)\,(x - s)\,\mathrm{d}s.$$

Imposing the condition that $g(0) = g'(0) = 0$ for functions in $\mathcal{H}$ will ensure that $\mathcal{G} = \mathcal{H} \oplus \mathcal{H}_0$ where the null space $\mathcal{H}_0$ contains only linear functions, as we will see.

To see that this $\mathcal{H}$ is in fact an RKHS, we derive its reproducing kernel. Using integration by parts (or directly from the Taylor expansion above), write

$$g(x) = \int_0^x g'(s)\,\mathrm{d}s = \int_0^x g''(s)\,(x - s)\,\mathrm{d}s = \int_0^1 g''(s)\,(x - s)_+\,\mathrm{d}s.$$

If κ is a kernel, then by the reproducing property it must hold that

$$g(x) = \langle g, \kappa_x \rangle_{\mathcal{H}} = \int_0^1 g''(s)\,\kappa_x''(s)\,\mathrm{d}s,$$

so that κ must satisfy $\frac{\partial^2}{\partial s^2}\kappa(x, s) = (x - s)_+$, where $y_+ := \max\{y, 0\}$. Therefore, noting that $\kappa(x, u) = \langle \kappa_x, \kappa_u \rangle_{\mathcal{H}}$, we have (see Exercise 15)

$$\kappa(x, u) = \int_0^1 \frac{\partial^2 \kappa(x, s)}{\partial s^2}\frac{\partial^2 \kappa(u, s)}{\partial s^2}\,\mathrm{d}s = \frac{\max\{x, u\}\min\{x, u\}^2}{2} - \frac{\min\{x, u\}^3}{6}.$$

The last expression is a cubic function with quadratic and cubic terms that misses the constant and linear monomials. This is not surprising considering the Taylor's theorem interpretation of a function $g \in \mathcal{H}$. If we now take $\mathcal{H}_0$ as the space of functions of the following form (having zero second derivative):

$$h_0 = \eta_1 + \eta_2\,x, \quad x \in [0, 1],$$

then (6.25) is exactly of the form (6.19).

As a consequence of the representer Theorem 6.6, the optimal solution to (6.25) is a linear combination of piecewise cubic functions:

$$g(x) = \eta_1 + \eta_2\,x + \sum_{i=1}^{n} \alpha_i\,\kappa(x_i, x). \tag{6.26}$$

CUBIC SPLINE

Such a function is called a *cubic spline* with *n knots* (with one knot at each data point x_i) — so called, because the piecewise cubic function between knots is required to be "tied together" at the knots. The parameters $\boldsymbol{\alpha}, \boldsymbol{\eta}$ are determined from (6.21) for instance by solving (6.23) with matrices $\mathbf{K} = [\kappa(x_i, x_j)]_{i,j=1}^n$ and $\mathbf{Q}$ with i-th row of the form $[1, x_i]$ for $i = 1, \dots, n$.

■ **Example 6.15 (Smoothing Spline)** Figure 6.6 shows various cubic smoothing splines for the data $(0.05, 0.4), (0.2, 0.2), (0.5, 0.6), (0.75, 0.7), (1, 1)$. In the figure, we use the reparameterization $r = 1/(1 + n\,\gamma)$ for the smoothing parameter. Thus $r \in [0, 1]$, where $r = 0$ means an infinite penalty for curvature (leading to the ordinary linear regression solution) and $r = 1$ does not penalize curvature at all and leads to a perfect fit via the so-called *natural spline*. Of course the latter will generally lead to overfitting. For r from 0 up to 0.8 the solutions will be close to the simple linear regression line, while only for r very close to 1, the shape of the curve changes significantly.

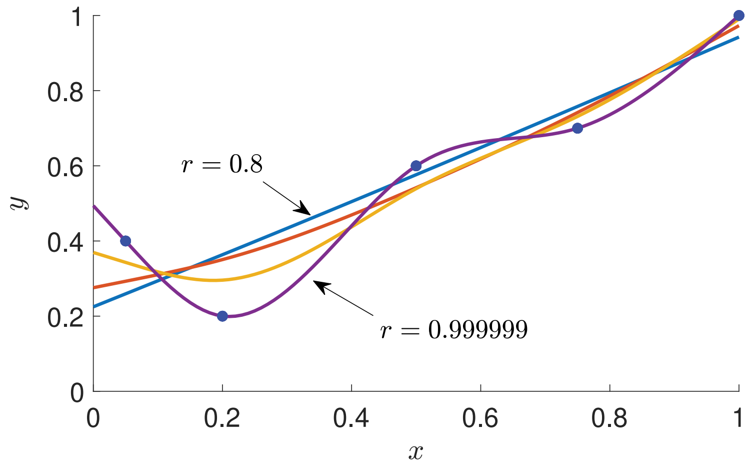

Figure 6.6: Various cubic smoothing splines for smoothing parameter $r = 1/(1 + n\gamma) \in \{0.8, 0.99, 0.999, 0.999999\}$. For $r = 1$, the natural spline through the data points is obtained; for $r = 0$, the simple linear regression line is found.

The following code first computes the matrices $\mathbf{K}$ and $\mathbf{Q}$, and then solves the linear system (6.23). Finally, the smoothing curve is determined via (6.26), for selected points, and then plotted. Note that the code plots only a single curve corresponding to the specified value of p.

smoothspline.py

```python
import matplotlib.pyplot as plt
import numpy as np

x = np.array([[0.05, 0.2, 0.5, 0.75, 1.]]).T
y = np.array([[0.4, 0.2, 0.6, 0.7, 1.]]).T

n = x.shape[0]
r = 0.999
ngamma = (1-r)/r

k = lambda x1, x2 : (1/2)* np.max((x1,x2)) * np.min((x1,x2)) ** 2 \
                           - ((1/6)* np.min((x1,x2))**3)
K = np.zeros((n,n))
for i in range(n):
    for j in range(n):
        K[i,j] = k(x[i], x[j])

Q = np.hstack((np.ones((n,1)), x))

m1 = np.hstack((K @ K.T + (ngamma * K), K @ Q))
m2 = np.hstack((Q.T @ K.T, Q.T @ Q))
```

```
M = np.vstack((m1,m2))

c = np.vstack((K, Q.T)) @ y

ad = np.linalg.solve(M,c)

# plot the curve
xx = np.arange(0,1+0.01,0.01).reshape(-1,1)

g = np.zeros_like(xx)
Qx = np.hstack((np.ones_like(xx), xx))
g = np.zeros_like(xx)
N = np.shape(xx)[0]

Kx = np.zeros((n,N))
for i in range(n):
    for j in range(N):
        Kx[i,j] = k(x[i], xx[j])

g = g + np.hstack((Kx.T, Qx)) @ ad

plt.ylim((0,1.15))
plt.plot(xx, g, label = 'r = {}'.format(r), linewidth = 2)
plt.plot(x,y, 'b.', markersize=15)
plt.xlabel('$x$')
plt.ylabel('$y$')
plt.legend()
```

6.7 Gaussian Process Regression

GAUSSIAN PROCESS

Another application of the kernel machinery is to Gaussian process regression. A *Gaussian process* (GP) on a space $\mathscr{X}$ is a stochastic process $\{Z_{\boldsymbol{x}}, \boldsymbol{x} \in \mathscr{X}\}$ where, for any choice of indices $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$, the vector $[Z_{\boldsymbol{x}_1}, \ldots Z_{\boldsymbol{x}_n}]^\top$ has a multivariate Gaussian distribution. As such, the distribution of a GP is completely specified by its mean and covariance functions $\mu : \mathscr{X} \to \mathbb{R}$ and $\kappa : \mathscr{X} \times \mathscr{X} \to \mathbb{R}$, respectively. The covariance function is a finite positive semidefinite function, and hence, in view of Theorem 6.2, can be viewed as a reproducing kernel on $\mathscr{X}$.

☞ 168

As for ordinary regression, the objective of GP regression is to learn a regression function g that predicts a response $y = g(\boldsymbol{x})$ for each feature vector $\boldsymbol{x}$. This is done in a Bayesian fashion, by establishing (1) a prior pdf for g and (2) the likelihood of the data, for a given g. From these two we then derive, via Bayes' formula, the posterior distribution of g given the data. We refer to Section 2.9 for the general Bayesian framework.

☞ 47

A simple Bayesian model for GP regression is as follows. First, the prior distribution of g is taken to be the distribution of a GP with some known mean function μ and covariance function (that is, kernel) κ. Most often μ is taken to be a constant, and for simplicity of exposition, we take it to be 0. The Gaussian kernel (6.15) is often used for the covariance function. For radial basis function kernels (including the Gaussian kernel), points that are closer will be more highly correlated or "similar" [97], independent of translations in space.

Second, similar to standard regression, we view the observed feature vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ as fixed and the responses $y_1, \ldots, y_n$ as outcomes of random variables $Y_1, \ldots, Y_n$. Specifically, given g, we model the $\{Y_i\}$ as

$$Y_i = g(\boldsymbol{x}_i) + \varepsilon_i\,, \quad i = 1, \ldots, n, \tag{6.27}$$

where $\{\varepsilon_i\} \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$. To simplify the analysis, let us assume that σ^2 is known, so no prior needs to be specified for σ^2. Let $\boldsymbol{g} = [g(\boldsymbol{x}_1), \ldots, g(\boldsymbol{x}_n)]^\top$ be the (unknown) vector of regression values. Placing a GP prior on the function g is equivalent to placing a multivariate Gaussian prior on the vector $\boldsymbol{g}$:

$$\boldsymbol{g} \sim \mathcal{N}(\boldsymbol{0}, \mathbf{K}), \tag{6.28}$$

where the covariance matrix $\mathbf{K}$ of $\boldsymbol{g}$ is a Gram matrix (implicitly associated with a feature map through the kernel κ), given by:

$$\mathbf{K} = \begin{bmatrix} \kappa(\boldsymbol{x}_1, \boldsymbol{x}_1) & \kappa(\boldsymbol{x}_1, \boldsymbol{x}_2) & \ldots & \kappa(\boldsymbol{x}_1, \boldsymbol{x}_n) \\ \kappa(\boldsymbol{x}_2, \boldsymbol{x}_1) & \kappa(\boldsymbol{x}_2, \boldsymbol{x}_2) & \ldots & \kappa(\boldsymbol{x}_2, \boldsymbol{x}_n) \\ \vdots & \vdots & \ddots & \vdots \\ \kappa(\boldsymbol{x}_n, \boldsymbol{x}_1) & \kappa(\boldsymbol{x}_n, \boldsymbol{x}_2) & \ldots & \kappa(\boldsymbol{x}_n, \boldsymbol{x}_n) \end{bmatrix}. \tag{6.29}$$

The likelihood of our data given $\boldsymbol{g}$, denoted $p(\boldsymbol{y} \,|\, \boldsymbol{g})$, is obtained directly from the model (6.27):

$$(\boldsymbol{Y} \,|\, \boldsymbol{g}) \sim \mathcal{N}(\boldsymbol{g}, \sigma^2 \mathbf{I}_n). \tag{6.30}$$

Solving this Bayesian problem involves deriving the posterior distribution of $(\boldsymbol{g} \,|\, \boldsymbol{Y})$. To do so, we first note that since $\boldsymbol{Y}$ has covariance matrix $\mathbf{K} + \sigma^2 \mathbf{I}_n$ (which can be seen from (6.27)), the joint distribution of $\boldsymbol{Y}$ and $\boldsymbol{g}$ is again normal, with mean $\boldsymbol{0}$ and covariance matrix:

$$\mathbf{K}_{y,g} = \begin{bmatrix} \mathbf{K} + \sigma^2 \mathbf{I}_n & \mathbf{K} \\ \mathbf{K} & \mathbf{K} \end{bmatrix}. \tag{6.31}$$

The posterior can then be found by conditioning on $\boldsymbol{Y} = \boldsymbol{y}$, via Theorem C.8, giving ☞ 436

$$(\boldsymbol{g} \,|\, \boldsymbol{y}) \sim \mathcal{N}\left(\mathbf{K}^\top (\mathbf{K} + \sigma^2 \mathbf{I}_n)^{-1} \boldsymbol{y},\ \mathbf{K} - \mathbf{K}^\top (\mathbf{K} + \sigma^2 \mathbf{I}_n)^{-1} \mathbf{K}\right).$$

This only gives information about g at the observed points $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$. It is more interesting to consider the posterior predictive distribution of $\widetilde{g} := g(\widetilde{\boldsymbol{x}})$ for a new input $\widetilde{\boldsymbol{x}}$. We can find the corresponding posterior predictive pdf $p(\widetilde{g} \,|\, \boldsymbol{y})$ by integrating out the joint posterior pdf $p(\widetilde{g}, \boldsymbol{g} \,|\, \boldsymbol{y})$, which is equivalent to taking the expectation of $p(\widetilde{g} \,|\, \boldsymbol{g})$ when $\boldsymbol{g}$ is distributed according to the posterior pdf $p(\boldsymbol{g} \,|\, \boldsymbol{y})$; that is,

$$p(\widetilde{g} \,|\, \boldsymbol{y}) = \int p(\widetilde{g} \,|\, \boldsymbol{g})\, p(\boldsymbol{g} \,|\, \boldsymbol{y})\, \mathrm{d}\boldsymbol{g}.$$

To do so more easily than direct evaluation via the above integral representation of $p(\widetilde{g} \,|\, \boldsymbol{y})$, we can begin with the joint distribution of $[\boldsymbol{y}^\top, \widetilde{g}]^\top$, which is multivariate normal with mean $\boldsymbol{0}$ and covariance matrix

$$\widetilde{\mathbf{K}} = \begin{bmatrix} \mathbf{K} + \sigma^2 \mathbf{I}_n & \boldsymbol{\kappa} \\ \boldsymbol{\kappa}^\top & \kappa(\widetilde{\boldsymbol{x}}, \widetilde{\boldsymbol{x}}) \end{bmatrix}, \tag{6.32}$$

where $\boldsymbol{\kappa} = [\kappa(\widetilde{\boldsymbol{x}}, \boldsymbol{x}_1), \ldots, \kappa(\widetilde{\boldsymbol{x}}, \boldsymbol{x}_n)]^\top$. It now follows, again by using Theorem C.8, that $(\widetilde{g} \mid \boldsymbol{y})$ has a normal distribution with mean and variance given respectively by

$$\mu(\widetilde{\boldsymbol{x}}) = \boldsymbol{\kappa}^\top (\mathbf{K} + \sigma^2 \mathbf{I}_n)^{-1} \boldsymbol{y} \tag{6.33}$$

and

$$\sigma^2(\widetilde{\boldsymbol{x}}) = \kappa(\widetilde{\boldsymbol{x}}, \widetilde{\boldsymbol{x}}) - \boldsymbol{\kappa}^\top (\mathbf{K} + \sigma^2 \mathbf{I}_n)^{-1} \boldsymbol{\kappa}. \tag{6.34}$$

PREDICTIVE

These are sometimes called the *predictive* mean and variance. It is important to note that we are predicting the *expected* response $\mathbb{E}\widetilde{Y} = g(\widetilde{\boldsymbol{x}})$ here, and not the actual response $\widetilde{Y}$.

■ **Example 6.16 (GP Regression)** Suppose the regression function is

$$g(x) = 2\sin(2\pi x), \qquad x \in [0, 1].$$

We use GP regression to estimate g, using a Gaussian kernel of the form (6.15) with bandwidth parameter 0.2. The explanatory variables $x_1, \ldots, x_{30}$ were drawn uniformly on the interval [0, 1], and the responses were obtained from (6.27), with noise level $\sigma = 0.5$. Figure 6.7 shows 10 samples from the prior distribution for g as well as the data points and the true sinusoidal regression function g.

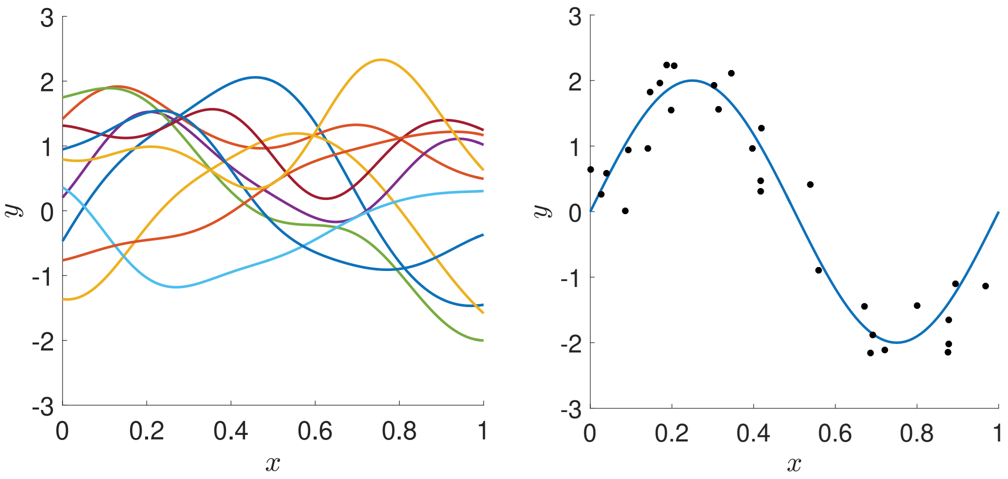

Figure 6.7: Left: samples drawn from the GP prior distribution. Right: the true regression function with the data points.

Again assuming that the variance σ^2, is known, the predictive distribution as determined by (6.33) and (6.34) is shown in Figure 6.8 for bandwidth 0.2 (left) and 0.02 (right). Clearly, decreasing the bandwidth leads to the covariance between points x and x' decreasing at a faster rate with respect to the squared distance $\|x - x'\|^2$, leading to a predictive mean that is less smooth. ■

In the above exposition, we have taken the mean function for the prior distribution of g to be identically zero. If instead we have a general mean function m and write $\boldsymbol{m} = [m(\boldsymbol{x}_1), \ldots, m(\boldsymbol{x}_n)]^\top$ then the predictive variance (6.34) remains unchanged, and the predictive mean (6.33) is modified to read

$$\mu(\widetilde{\boldsymbol{x}}) = m(\widetilde{\boldsymbol{x}}) + \boldsymbol{\kappa}^\top (\mathbf{K} + \sigma^2 \mathbf{I}_n)^{-1} (\boldsymbol{y} - \boldsymbol{m}). \tag{6.35}$$

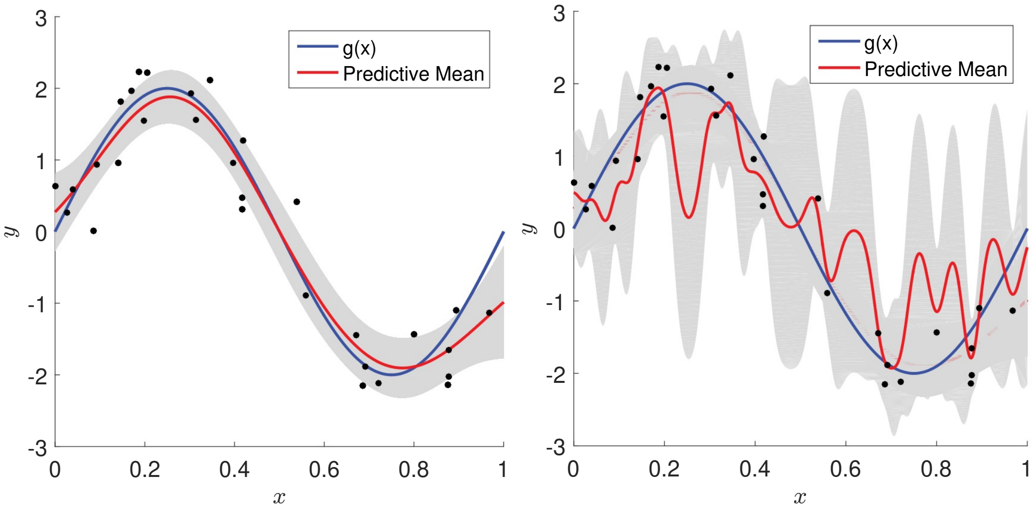

Figure 6.8: GP regression of synthetic data set with bandwidth 0.2 (left) and 0.02 (right). The black dots represent the data and the blue curve is the latent function $g(x) = 2\sin(2\pi x)$. The red curve is the mean of the GP predictive distribution given by (6.33), and the shaded region is the 95% confidence band, corresponding to the predictive variance given in (6.34).

Typically, the variance σ^2 appearing in (6.27) is not known, and the kernel κ itself depends on several parameters — for instance a Gaussian kernel (6.15) with an unknown bandwidth parameter. In the Bayesian framework, one typically specifies a hierarchical model by introducing a prior $p(\boldsymbol{\theta})$ for the vector $\boldsymbol{\theta}$ of such *hyperparameters*. Now, the GP prior $(g \,|\, \boldsymbol{\theta})$ (equivalently, specifying $p(\boldsymbol{g} \,|\, \boldsymbol{\theta})$) and the model for the likelihood of the data given $\boldsymbol{Y}|\boldsymbol{g}, \boldsymbol{\theta}$, namely $p(\boldsymbol{y} \,|\, \boldsymbol{g}, \boldsymbol{\theta})$, are both dependent on $\boldsymbol{\theta}$. The posterior distribution of $(\boldsymbol{g} \,|\, \boldsymbol{y}, \boldsymbol{\theta})$ is as before.

HYPERPARAMETERS

One approach to setting the hyperparameter $\boldsymbol{\theta}$ is to determine its posterior $p(\boldsymbol{\theta} \,|\, \boldsymbol{y})$ and obtain a point estimate, for instance via its maximum a posteriori estimate. However, this can be a computationally demanding exercise. What is frequently done in practice is to consider instead the *marginal likelihood* $p(\boldsymbol{y} \,|\, \boldsymbol{\theta})$ and maximize this with respect to $\boldsymbol{\theta}$. This procedure is called *empirical Bayes*.

EMPIRICAL BAYES

Considering again the mean function m to be identically zero, from (6.31), we have that $(\boldsymbol{Y} \,|\, \boldsymbol{\theta})$ is multivariate normal with mean $\mathbf{0}$ and covariance matrix $\mathbf{K}_y = \mathbf{K} + \sigma^2 \mathbf{I}_n$, immediately giving an expression for the marginal log-likelihood:

$$\ln p(\boldsymbol{y} \,|\, \boldsymbol{\theta}) = -\frac{n}{2}\ln(2\pi) - \frac{1}{2}\ln|\det(\mathbf{K}_y)| - \frac{1}{2}\boldsymbol{y}^\top \mathbf{K}_y^{-1}\boldsymbol{y}. \tag{6.36}$$

We notice that only the second and third terms in (6.36) depend on $\boldsymbol{\theta}$. Considering a partial derivative of (6.36) with respect to a single element θ of the hyperparameter vector $\boldsymbol{\theta}$ yields

$$\frac{\partial}{\partial \theta}\ln p(\boldsymbol{y} \,|\, \boldsymbol{\theta}) = -\frac{1}{2}\mathrm{tr}\left(\mathbf{K}_y^{-1}\left[\frac{\partial}{\partial \theta}\mathbf{K}_y\right]\right) + \frac{1}{2}\boldsymbol{y}^\top \mathbf{K}_y^{-1}\left[\frac{\partial}{\partial \theta}\mathbf{K}_y\right]\mathbf{K}_y^{-1}\boldsymbol{y}, \tag{6.37}$$

where $\left[\frac{\partial}{\partial\theta}\mathbf{K}_y\right]$ is the element-wise derivative of matrix K_y with respect to θ. If these partial derivatives can be computed for each hyperparameter θ, gradient information could be used when maximizing (6.36).

■ **Example 6.17 (GP Regression (cont.))** Continuing Example 6.16, we plot in Figure 6.9 the marginal log-likelihood as a function of the noise level σ and bandwidth parameter.

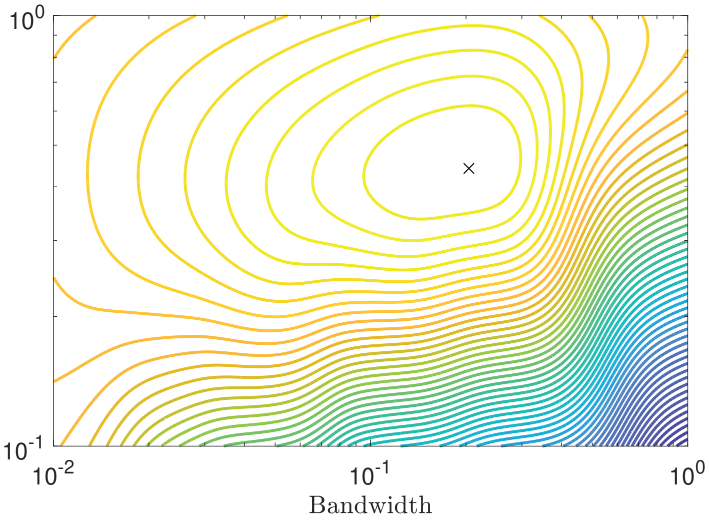

Figure 6.9: Contours of the marginal log-likelihood for the GP regression example. The maximum is denoted by a cross.

The maximum is attained for a bandwidth parameter around 0.20 and $\sigma \approx 0.44$, which is very close to the left panel of Figure 6.8 for the case where σ was assumed to be known (and equal to 0.5). We note here that the marginal log-likelihood is extremely flat, perhaps owing to the small number of points. ■

6.8 Kernel PCA

☞ 153

In its basic form, kernel PCA (principal component analysis) can be thought of as PCA in feature space. The main motivation for PCA introduced in Section 4.8 was as a dimensionality reduction technique. There, the analysis rested on an SVD of the matrix $\widehat{\boldsymbol{\Sigma}} = \frac{1}{n}\mathbf{X}^\top\mathbf{X}$, where the data in $\mathbf{X}$ was first centered via $x'_{i,j} = x_{i,j} - \overline{x}_j$ where $\overline{x}_i = \frac{1}{n}\sum_{i=1}^{n} x_{i,j}$.

What we shall do is to first re-cast the problem in terms of the Gram matrix $\mathbf{K} = \mathbf{X}\mathbf{X}^\top = [\langle \boldsymbol{x}_i, \boldsymbol{x}_j\rangle]$ (note the different order of $\mathbf{X}$ and $\mathbf{X}^\top$), and subsequently replace the inner product $\langle \boldsymbol{x}, \boldsymbol{x}'\rangle$ with $\kappa(\boldsymbol{x}, \boldsymbol{x}')$ for a general reproducing kernel κ. To make the link, let us start with an SVD of $\mathbf{X}^\top$:

$$\mathbf{X}^\top = \mathbf{U}\mathbf{D}\mathbf{V}^\top. \tag{6.38}$$

The dimensions of $\mathbf{X}^\top$, $\mathbf{U}$, $\mathbf{D}$, and $\mathbf{V}$ are $d \times n$, $d \times d$, $d \times n$, and $n \times n$, respectively. Then an SVD of $\mathbf{X}^\top\mathbf{X}$ is

$$\mathbf{X}^\top\mathbf{X} = (\mathbf{U}\mathbf{D}\mathbf{V}^\top)(\mathbf{U}\mathbf{D}\mathbf{V}^\top)^\top = \mathbf{U}(\mathbf{D}\mathbf{D}^\top)\mathbf{U}^\top$$

and an SVD of $\mathbf{K}$ is

$$\mathbf{K} = (\mathbf{U}\mathbf{D}\mathbf{V}^\top)^\top(\mathbf{U}\mathbf{D}\mathbf{V}^\top) = \mathbf{V}(\mathbf{D}^\top\mathbf{D})\mathbf{V}^\top.$$

Let $\lambda_1 \geqslant \cdots \geqslant \lambda_r > 0$ denote the non-zero eigenvalues of $\mathbf{X}^\top\mathbf{X}$ (or, equivalently, of $\mathbf{K}$) and denote the corresponding $r \times r$ diagonal matrix by $\boldsymbol{\Lambda}$. Without loss of generality we can assume that the eigenvector of $\mathbf{X}^\top\mathbf{X}$ corresponding to λ_k is the k-th column of $\mathbf{U}$ and that the k-th column of $\mathbf{V}$ is an eigenvector of $\mathbf{K}$. Similar to Section 4.8, let $\mathbf{U}_k$ and $\mathbf{V}_k$ contain ☞ 153 the first k columns of $\mathbf{U}$ and $\mathbf{V}$, respectively, and let $\mathbf{\Lambda}_k$ be the corresponding $k\times k$ submatrix of $\mathbf{\Lambda}$, $k = 1, \dots, r$.

By the SVD (6.38), we have $\mathbf{X}^\top\mathbf{V}_k = \mathbf{U}\mathbf{D}\mathbf{V}^\top\mathbf{V}_k = \mathbf{U}_k\mathbf{\Lambda}_k^{1/2}$. Next, consider the projection of a point $\boldsymbol{x}$ onto the k-dimensional linear space spanned by the columns of $\mathbf{U}_k$ — the first k principal components. We saw in Section 4.8 that this projection simply is the linear mapping $\boldsymbol{x} \mapsto \mathbf{U}_k^\top\boldsymbol{x}$. Using the fact that $\mathbf{U}_k = \mathbf{X}^\top\mathbf{V}_k\mathbf{\Lambda}^{-1/2}$, we find that $\boldsymbol{x}$ is projected to a point $\boldsymbol{z}$ given by

$$\boldsymbol{z} = \mathbf{\Lambda}_k^{-1/2}\mathbf{V}_k^\top\mathbf{X}\boldsymbol{x} = \mathbf{\Lambda}_k^{-1/2}\mathbf{V}_k^\top\boldsymbol{\kappa}_{\boldsymbol{x}},$$

where we have (suggestively) defined $\boldsymbol{\kappa}_{\boldsymbol{x}} := [\langle\boldsymbol{x}_1, \boldsymbol{x}\rangle, \dots, \langle\boldsymbol{x}_n, \boldsymbol{x}\rangle]^\top$. The important point is that $\boldsymbol{z}$ is completely determined by the vector of inner products $\boldsymbol{\kappa}_{\boldsymbol{x}}$ and the k principal eigenvalues and (right) eigenvectors of the Gram matrix $\mathbf{K}$. Note that each component z_m of $\boldsymbol{z}$ is of the form

$$z_m = \sum_{i=1}^{n} \alpha_{m,i}\,\kappa(\boldsymbol{x}_i, \boldsymbol{x}), \quad m = 1, \dots, k. \tag{6.39}$$

The preceding discussion assumed centering of the columns of $\mathbf{X}$. Consider now an uncentered data matrix $\widetilde{\mathbf{X}}$. Then the centered data can be written as $\mathbf{X} = \widetilde{\mathbf{X}} - \frac{1}{n}\mathbf{E}_n\widetilde{\mathbf{X}}$, where $\mathbf{E}_n$ is the $n \times n$ matrix of ones. Consequently,

$$\mathbf{X}\mathbf{X}^\top = \widetilde{\mathbf{X}}\widetilde{\mathbf{X}}^\top - \frac{1}{n}\mathbf{E}_n\widetilde{\mathbf{X}}\widetilde{\mathbf{X}}^\top - \frac{1}{n}\widetilde{\mathbf{X}}\widetilde{\mathbf{X}}^\top\mathbf{E}_n + \frac{1}{n^2}\mathbf{E}_n\widetilde{\mathbf{X}}\widetilde{\mathbf{X}}^\top\mathbf{E}_n,$$

or, more compactly, $\mathbf{X}\mathbf{X}^\top = \mathbf{H}\,\widetilde{\mathbf{X}}\widetilde{\mathbf{X}}^\top\mathbf{H}$, where $\mathbf{H} = \mathbf{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top$, $\mathbf{I}_n$ is the $n\times n$ identity matrix, and $\mathbf{1}_n$ is the $n \times 1$ vector of ones.

To generalize to the kernel setting, we replace $\widetilde{\mathbf{X}}\widetilde{\mathbf{X}}^\top$ by $\mathbf{K} = [\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j), i, j = 1, \dots, n]$ and set $\boldsymbol{\kappa}_{\boldsymbol{x}} = [\kappa(\boldsymbol{x}_1, \boldsymbol{x}), \dots, \kappa(\boldsymbol{x}_n, \boldsymbol{x})]^\top$, so that $\mathbf{\Lambda}_k$ is the diagonal matrix of the k largest eigenvalues of $\mathbf{HKH}$ and $\mathbf{V}_k$ is the corresponding matrix of eigenvectors. Note that the "usual" PCA is recovered when we use the linear kernel $\kappa(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{x}^\top\boldsymbol{y}$. However, instead of having only kernels that are explicitly inner products of feature vectors, we are now permitted to implicitly use *infinite* feature maps (functions) by using kernels.

■ **Example 6.18 (Kernel PCA)** We simulated 200 points, $\boldsymbol{x}_1, \dots, \boldsymbol{x}_{200}$, from the uniform distribution on the set $B_1 \cup (B_4 \cap B_3^c)$, where $B_r := \{(x, y) \in \mathbb{R}^2 : x^2 + y^2 \leqslant r^2\}$ (disk with radius r). We apply kernel PCA with Gaussian kernel $\kappa(\boldsymbol{x}, \boldsymbol{x}') = \exp\left(-\|\boldsymbol{x} - \boldsymbol{x}'\|^2\right)$ and compute the functions $z_m(\boldsymbol{x}), m = 1, \dots, 9$ in (6.39). Their density plots are shown in Figure 6.10. The data points are superimposed in each plot. From this we see that the principal components identify the radial structure present in the data. Finally, Figure 6.11 shows the projections $[z_1(\boldsymbol{x}_i), z_2(\boldsymbol{x}_i)]^\top, i = 1, \dots, 200$ of the original data points onto the first two principal components. We see that the projected points can be separated by a straight line, whereas this is not possible for the original data; see also, Example 7.6 for a related prob- ☞ 272 lem.

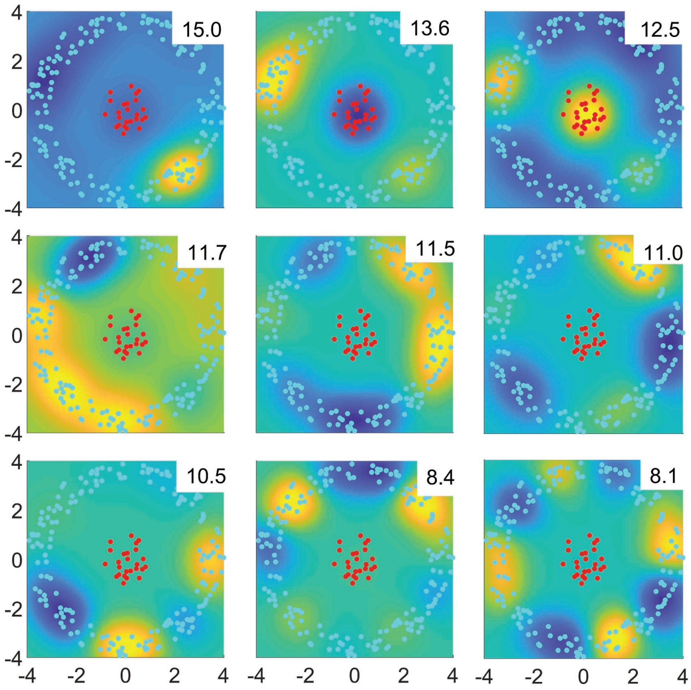

Figure 6.10: First nine eigenfunctions using a Gaussian kernel for the two-dimensional data set formed by the red and cyan points.

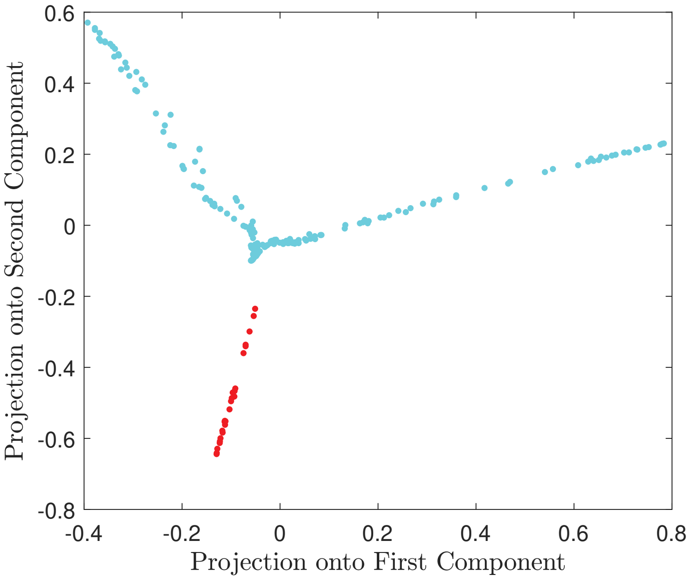

Figure 6.11: Projection of the data onto the first two principal components. Observe that already the projections of the inner and outer points are well separated.

Further Reading

For a good overview of the ridge regression and the lasso, we refer the reader to [36, 56]. For overviews of the theory of RKHS we refer to [3, 115, 126], and for in-depth background on splines and their connection to RKHSs we refer to [123]. For further details on GP regression we refer to [97] and for kernel PCA in particular we refer to [12, 92]. Finally, many facts about kernels and their corresponding RKHSs can be found in [115].

Exercises

1. Let $\mathcal{G}$ be an RKHS with reproducing kernel κ. Show that κ is a positive semidefinite function.

2. Show that a reproducing kernel, if it exists, is unique.

3. Let $\mathcal{G}$ be a Hilbert space of functions $g : \mathcal{X} \to \mathbb{R}$. Recall that the *evaluation functional* is the map $\delta_{\boldsymbol{x}} : g \mapsto g(\boldsymbol{x})$ for a given $\boldsymbol{x} \in \mathcal{X}$. Show that evaluation functionals are linear operators.

4. Let $\mathcal{G}_0$ be the pre-RKHS $\mathcal{G}_0$ constructed in the proof of Theorem 6.2. Thus, $g \in \mathcal{G}_0$ is of the form $g = \sum_{i=1}^n \alpha_i \, \kappa_{\boldsymbol{x}_i}$ and

$$\langle g, \kappa_{\boldsymbol{x}} \rangle_{\mathcal{G}_0} = \sum_{i=1}^n \alpha_i \, \langle \kappa_{\boldsymbol{x}_i}, \kappa_{\boldsymbol{x}} \rangle_{\mathcal{G}_0} = \sum_{i=1}^n \alpha_i \, \kappa(\boldsymbol{x}_i, \boldsymbol{x}) = g(\boldsymbol{x}).$$

   Therefore, we may write the evaluation functional of $g \in \mathcal{G}_0$ at $\boldsymbol{x}$ as $\delta_{\boldsymbol{x}} g := \langle g, \kappa_{\boldsymbol{x}} \rangle_{\mathcal{G}_0}$. Show that $\delta_{\boldsymbol{x}}$ is bounded on $\mathcal{G}_0$ for every $\boldsymbol{x}$; that is, $|\delta_{\boldsymbol{x}} f| < \gamma \, \|f\|_{\mathcal{G}_0}$, for some $\gamma < \infty$.

5. Continuing Exercise 4, let (f_n) be a Cauchy sequence in $\mathcal{G}_0$ such that $|f_n(\boldsymbol{x})| \to 0$ for all $\boldsymbol{x}$. Show that $\|f_n\|_{\mathcal{G}_0} \to 0$.

6. Continuing Exercises 5 and 4, to show that the inner product (6.14) is well defined, a number of facts have to be checked.

   (a) Verify that the limit converges.

   (b) Verify that the limit is independent of the Cauchy sequences used.

   (c) Verify that the properties of an inner product are satisfied. The only non-trivial property to verify is that $\langle f, f \rangle_{\mathcal{G}} = 0$ if and only if $f = 0$.

7. Exercises 4–6 show that $\mathcal{G}$ defined in the proof of Theorem 6.2 is an inner product space. It remains to prove that $\mathcal{G}$ is an RKHS. This requires us to prove that the inner product space $\mathcal{G}$ is complete (and thus Hilbert), and that its evaluation functionals are bounded and hence continuous (see Theorem A.16). This is done in a number of steps. ☞ 389

   (a) Show that $\mathcal{G}_0$ is dense in $\mathcal{G}$ in the sense that every $f \in \mathcal{G}$ is a limit point (with respect to the norm on $\mathcal{G}$) of a Cauchy sequence (f_n) in $\mathcal{G}_0$.