4

Kernels and Reproducing Kernel Hilbert Spaces

> **Overview.** *We saw in Section 1.3 that kernels and their feature spaces are the devices by which the linear SVM approach produces non-linear decision functions. However, so far we only have a vague notion of kernels and hence we investigate them in more detail in this chapter.*
>
> **Prerequisites.** *This chapter requires basic knowledge in functional analysis, which is provided in Section A.5. Section 4.4 on Gaussian kernels also needs some complex analysis from Section A.7.*
>
> **Usage.** *Sections 4.1, 4.2, 4.3, and 4.6, which deal with the core material on kernels, are essential for the rest of this book. Moreover, Section 4.4 is needed for binary classification discussed in Chapter 8.*

As we have described in the introduction, one of the major steps in constructing a support vector machine is mapping the input space X into a feature space H that is equipped with an inner product. The benefit of this step is that for non-linear feature maps $\Phi : X \to H$, support vector machines can produce non-linear decision functions, although SVMs are only based on a linear discriminant approach. Furthermore, we have seen that SVMs only require computing the inner products $k(x, x') := \langle \Phi(x), \Phi(x') \rangle_H$ rather than Φ itself. Thus, instead of first constructing Φ and then computing the inner products, one can use SVMs with *efficiently* computable functions $k : X \times X \to \mathbb{R}$ that realize inner products of (possibly unknown) feature maps. We called such functions k kernels, and the approach described was called the kernel trick. Of course, this trick immediately raises some questions:

- When is a function $k : X \times X \to \mathbb{R}$ a kernel?
- How can we construct kernels?
- Given a kernel k, can we find a feature map and a feature space of k in a constructive way?
- How much does the kernel trick increase the expressive power of support vector machines?

The aim of this chapter is to answer these questions. To this end, we formalize the definition of kernels in Section 4.1. Moreover, in this section we also present some simple but useful examples of kernels. Then, in Section 4.2 we describe a canonical form of feature spaces, the so-called reproducing kernel Hilbert spaces. Basic properties of the functions contained in these spaces are presented in Section 4.3. Moreover, for an important type of kernel we determine these spaces explicitly in Section 4.4. In Section 4.5, we derive a specific series representation for continuous kernels on compact spaces. Finally, in Section 4.6 we describe a class of kernels for which SVMs have a large expressive power.

4.1 Basic Properties and Examples of Kernels

In this section, we introduce the notions *kernel*, *feature space*, and *feature map.* Then we show how to construct new kernels from given kernels and present some important examples of kernels that will be used frequently in this book. Finally, we establish a criterion that characterizes kernels with the help of positive definite matrices related to these kernels.

Although in the context of machine learning one is originally only interested in real-valued kernels, we will develop the basic theory on kernels also for complex-valued kernels. This more general approach does not require any additional technical effort, but it will enable us in Section 4.4 to discover some features of the Gaussian RBF kernels that are widely used in practice.

Before we begin with the basic definitions, let us recall that every complex number $z \in \mathbb{C}$ can be represented in the form $z = x + iy$, where $x, y \in \mathbb{R}$ and $i := \sqrt{-1}$. Both x and y are uniquely determined, and in the following we thus write $\operatorname{Re} z := x$ and $\operatorname{Im} z := y$. Moreover, the conjugate of z is defined by $\bar{z} := x - iy$ and its absolute value is $|z| := \sqrt{z\bar{z}} = \sqrt{x^2 + y^2}$. In particular, we have $\bar{x} = x$ and $|x| = \sqrt{x^2}$ for all $x \in \mathbb{R}$. Furthermore, we use the symbol $\mathbb{K}$ whenever we want to treat the real and the complex cases simultaneously. For example, a $\mathbb{K}$-Hilbert space H is a real Hilbert space when considering $\mathbb{K} = \mathbb{R}$ and a complex one when $\mathbb{K} = \mathbb{C}$. Recall from Definition A.5.8 that in the latter case the inner product $\langle \cdot\,, \cdot \rangle_H : H \times H \to \mathbb{C}$ is sesqui-linear, i.e., $\langle x, \alpha x' \rangle_H = \bar{\alpha} \langle x, x' \rangle_H$, and anti-symmetric, i.e., $\langle x, x' \rangle_H = \overline{\langle x', x \rangle_H}$.

With the help of these preliminary considerations, we can now formalize the notion of kernels.

Definition 4.1. *Let X be a non-empty set. Then a function $k : X \times X \to \mathbb{K}$ is called a* ***kernel*** *on X if there exists a $\mathbb{K}$-Hilbert space H and a map $\Phi : X \to H$ such that for all $x, x' \in X$ we have*

$$k(x, x') \;=\; \langle \Phi(x'), \Phi(x) \rangle \,. \tag{4.1}$$

We call Φ a ***feature map*** *and H a* ***feature space*** *of k.*

Note that in the real case condition (4.1) can be replaced by the more natural equation $k(x, x') = \langle \Phi(x), \Phi(x') \rangle$. In the complex case, however, $\langle \cdot\,, \cdot \rangle$ is anti-symmetric and hence (4.1) is equivalent to $k(x, x') = \overline{\langle \Phi(x), \Phi(x') \rangle}$.

Given a kernel, neither the feature map nor the feature space are uniquely determined. Let us illustrate this with a simple example. To this end, let

$X := \mathbb{R}$ and $k(x,x') := xx'$ for all $x,x' \in \mathbb{R}$. Then k is a kernel since obviously the identity map $\mathrm{id}_{\mathbb{R}}$ on $\mathbb{R}$ is a feature map with feature space $H := \mathbb{R}$. However, the map $\Phi : X \to \mathbb{R}^2$ defined by $\Phi(x) := (x/\sqrt{2}, x/\sqrt{2})$ for all $x \in X$ is also a feature map of k since we have

$$\langle \Phi(x'), \Phi(x) \rangle = \frac{x'}{\sqrt{2}} \cdot \frac{x}{\sqrt{2}} + \frac{x'}{\sqrt{2}} \cdot \frac{x}{\sqrt{2}} = xx' = k(x,x')$$

for all $x,x' \in X$. Moreover, note that a similar construction can be made for arbitrary kernels, and consequently every kernel has many different feature spaces. Finally, a less trivial example for different feature maps and spaces is discussed in Exercise 4.9.

Let us now present some commonly used kernels. To this end, we need some methods to construct kernels from scratch. We begin with a simple but instructive and fundamental observation.

Lemma 4.2. *Let X be a non-empty set and $f_n : X \to \mathbb{K}$, $n \in \mathbb{N}$, be functions such that $(f_n(x)) \in \ell_2$ for all $x \in X$. Then*

$$k(x,x') := \sum_{n=1}^{\infty} f_n(x)\overline{f_n(x')}, \qquad x,x' \in X, \tag{4.2}$$

defines a kernel on X.

Proof. Using Hölder's inequality for the sequence spaces ℓ_1 and ℓ_2, we obtain

$$\sum_{n=1}^{\infty} |f_n(x) f_n(x')| \le \|(f_n(x))\|_{\ell_2} \, \|(f_n(x'))\|_{\ell_2},$$

and hence the series in (4.2) converges absolutely for all $x,x' \in X$. Now, we write $H := \ell_2$ and define $\Phi : X \to H$ by $\Phi(x) := (\overline{f_n(x)})$, $x \in X$. Then (4.2) immediately gives the assertion. □

We will see in the following that almost all kernels we are interested in have a series representation of the form (4.2). However, before we present some examples of such kernels, we first need to establish some results that allow us to construct new kernels from given ones. We begin with the following simple lemma, whose proof is left as an exercise.

Lemma 4.3 (Restriction of kernels). *Let k be a kernel on X, $\tilde{X}$ be a set, and $A : \tilde{X} \to X$ be a map. Then $\tilde{k}$ defined by $\tilde{k}(x,x') := k(A(x), A(x'))$, $x,x' \in X$, is a kernel on $\tilde{X}$. In particular, if $\tilde{X} \subset X$, then $k_{|\tilde{X}\times\tilde{X}}$ is a kernel.*

For a kernel $k : \mathbb{C}^d \times \mathbb{C}^d \to \mathbb{C}$, Lemma 4.3 shows that the restriction $k_{|\mathbb{R}^d\times\mathbb{R}^d}$ is a kernel in the *complex* sense. The following result shows that it is also a kernel in the *real* sense if it satisfies $k(x,x') \in \mathbb{R}$ for all $x,x' \in \mathbb{R}^d$.

Lemma 4.4 (Real vs. complex kernels). *Let $k : X \times X \to \mathbb{C}$ be a kernel, H be a $\mathbb{C}$-Hilbert space, and $\Phi : X \to H$ be a feature map of k. Assume that we have $k(x,x') \in \mathbb{R}$ for all $x, x' \in X$. Then $H_0 := H$ equipped with the inner product*

$$\langle w, w' \rangle_{H_0} := \operatorname{Re} \langle w, w' \rangle_H , \qquad w,\, w' \in H_0,$$

is an $\mathbb{R}$-Hilbert space, and $\Phi : X \to H_0$ is a feature map of k.

Proof. It is elementary to check that $\langle \cdot\,, \cdot \rangle_{H_0}$ is a real inner product. Furthermore, we obviously have

$$k(x,x') = \bigl\langle \Phi(x'), \Phi(x) \bigr\rangle_H = \operatorname{Re} \bigl\langle \Phi(x'), \Phi(x) \bigr\rangle_H + i \operatorname{Im} \bigl\langle \Phi(x'), \Phi(x) \bigr\rangle_H$$

for all $x, x' \in X$. Consequently, $k(x,x') \in \mathbb{R}$ shows $\operatorname{Im} \langle \Phi(x'), \Phi(x) \rangle_H = 0$ for all $x, x' \in X$, and hence we obtain the assertion. □

Let us now establish some algebraic properties of the set of kernels on X. We begin with a simple lemma, whose proof is again left as an exercise.

Lemma 4.5 (Sums of kernels). *Let X be a set, $\alpha \geq 0$, and k, k_1, and k_2 be kernels on X. Then αk and $k_1 + k_2$ are also kernels on X.*

The preceding lemma states that the set of kernels on X is a cone. It is, however, not a vector space since in general differences of kernels are *not* kernels. To see this, let k_1 and k_2 be two kernels on X such that $k_1(x,x) - k_2(x,x) < 0$ for some $x \in X$. Then $k_1 - k_2$ is not a kernel since otherwise we would have a feature map $\Phi : X \to H$ of $k_1 - k_2$ with $0 \leq \langle \Phi(x), \Phi(x) \rangle = k_1(x,x) - k_2(x,x) < 0$. Let us now consider products of kernels.

Lemma 4.6 (Products of kernels). *Let k_1 be a kernel on X_1 and k_2 be a kernel on X_2. Then $k_1 \cdot k_2$ is a kernel on $X_1 \times X_2$. In particular, if $X_1 = X_2$, then $k(x,x') := k_1(x,x')k_2(x,x')$, $x, x' \in X$, defines a kernel on X.*

Proof. Let H_i be a feature space and $\Phi_i : X_i \to H_i$ be a feature map of k_i, $i = 1, 2$. Using the definition of the inner product in the tensor product space $H_1 \otimes H_2$ and its completion $H_1 \hat{\otimes} H_2$, see Appendix A.5.2, we obtain

$$\begin{aligned} k_1(x_1, x_1') \cdot k_2(x_2, x_2') &= \bigl\langle \Phi_1(x_1'), \Phi_1(x_1) \bigr\rangle_{H_1} \cdot \bigl\langle \Phi_2(x_2'), \Phi_2(x_2) \bigr\rangle_{H_2} \\ &= \bigl\langle \Phi_1(x_1') \otimes \Phi_2(x_2'),\, \Phi(x_1) \otimes \Phi_2(x_2) \bigr\rangle_{H_1 \hat{\otimes} H_2} , \end{aligned}$$

which shows that $\Phi_1 \otimes \Phi_2 : X_1 \times X_2 \to H_1 \hat{\otimes} H_2$ is a feature map of $k_1 \cdot k_2$. For the second assertion, we observe that k is a restriction of $k_1 \cdot k_2$. □

With the lemmas above, it is easy to construct non-trivial kernels. To illustrate this, let us assume for simplicity that $X := \mathbb{R}$. Then, for every integer $n \geq 0$, the map k_n defined by $k_n(x,x') := (xx')^n$, $x, x' \in X$, is a kernel by Lemma 4.2. Consequently, if $p : X \to \mathbb{R}$ is a polynomial of the form $p(t) = a_m t^m + \cdots + a_1 t + a_0$ with non-negative coefficients a_i, then $k(x,x') := p(xx')$,

$x, x' \in X$, defines a kernel on X by Lemma 4.5. In general, computing this kernel needs its feature map $\Phi(x) := (\sqrt{a_m}x^m, \dots, \sqrt{a_1}x, \sqrt{a_0})$, $x \in X$, and consequently the computational requirements are determined by the degree m. However, for some polynomials, these requirements can be substantially reduced. Indeed, if for example we have $p(t) = (t + c)^m$ for some $c > 0$ and all $t \in \mathbb{R}$, then the time needed to compute k is independent of m. The following lemma, whose proof is left as an exercise, generalizes this idea.

Lemma 4.7 (Polynomial kernels). *Let $m \geq 0$ and $d \geq 1$ be integers and $c \geq 0$ be a real number. Then k defined by $k(z, z') := (\langle z, z'\rangle + c)^m$, $z, z' \in \mathbb{C}^d$, is a kernel on $\mathbb{C}^d$. Moreover, its restriction to $\mathbb{R}^d$ is an $\mathbb{R}$-valued kernel.*

Note that the polynomial kernels defined by $m = 1$ and $c = 0$ are called **linear kernels**. Instead of using polynomials for constructing kernels, one can use functions that can be represented by Taylor series. This is done in the following lemma.

Lemma 4.8. *Let $\mathring{B}_{\mathbb{C}}$ and $\mathring{B}_{\mathbb{C}^d}$ be the open unit balls of $\mathbb{C}$ and $\mathbb{C}^d$, respectively. Moreover, let $r \in (0, \infty]$ and $f : r\mathring{B}_{\mathbb{C}} \to \mathbb{C}$ be holomorphic with Taylor series*

$$f(z) = \sum_{n=0}^{\infty} a_n z^n, \qquad z \in r\mathring{B}_{\mathbb{C}}\,. \tag{4.3}$$

If $a_n \geq 0$ for all $n \geq 0$, then

$$k(z, z') := f(\langle z, z'\rangle)_{\mathbb{C}^d} = \sum_{n=0}^{\infty} a_n \langle z, z'\rangle^n_{\mathbb{C}^d}\,, \qquad z, z' \in \sqrt{r}\mathring{B}_{\mathbb{C}^d},$$

*defines a kernel on $\sqrt{r}\mathring{B}_{\mathbb{C}^d}$ whose restriction to $X := \{x \in \mathbb{R}^d : \|x\|_2 < \sqrt{r}\}$ is a real-valued kernel. We say that k is a kernel of **Taylor type**.*

Proof. For $z, z' \in \sqrt{r}\mathring{B}_{\mathbb{C}^d}$, we have $|\langle z, z'\rangle| \leq \|z\|_2\|z'\|_2 < r$ and thus k is well-defined. Let z_i denote the i-th component of $z \in \mathbb{C}^d$. Since (4.3) is absolutely convergent, the multinomial formula (see Lemma A.1.2) then yields

$$\begin{aligned} k(z, z') = \sum_{n=0}^{\infty} a_n \Bigl(\sum_{j=1}^{d} z_j \bar{z}'_j\Bigr)^n &= \sum_{n=0}^{\infty} a_n \sum_{\substack{j_1,\dots,j_d \geq 0 \\ j_1+\dots+j_d = n}} c_{j_1,\dots,j_d} \prod_{i=1}^{d} (z_i \bar{z}'_i)^{j_i} \\ &= \sum_{j_1,\dots,j_d \geq 0} a_{j_1+\dots+j_d} c_{j_1,\dots,j_d} \prod_{i=1}^{d} (\bar{z}'_i)^{j_i} \prod_{i=1}^{d} z_i^{j_i}\,, \end{aligned}$$

where $c_{j_1,\dots,j_d} := \frac{n!}{\prod_{i=1}^{d} j_i!}$. Let us define $\Phi : X \to \ell_2(\mathbb{N}_0^d)$ by

$$\Phi(z) := \Bigl(\sqrt{a_{j_1+\dots+j_d} c_{j_1,\dots,j_d}} \prod_{i=1}^{d} \bar{z}_i^{j_i}\Bigr)_{j_1,\dots,j_d \geq 0}\,, \qquad z \in \sqrt{r}\mathring{B}_{\mathbb{C}^d}.$$

Then we have $k(z, z') = \langle \Phi(z'), \Phi(z)\rangle_{\ell_2(\mathbb{N}_0^d)}$ for all $z, z' \in \sqrt{r}\mathring{B}_{\mathbb{C}^d}$, and hence k is a kernel. The assertion for the restriction is obvious. □

With the help of Lemma 4.8 we can now present some more examples of commonly used kernels.

Example 4.9. For $d \in \mathbb{N}$ and $x, x' \in \mathbb{K}^d$, we define $k(x,x') := \exp(\langle x, x'\rangle)$. Then k is a $\mathbb{K}$-valued kernel on $\mathbb{K}^d$ called the **exponential kernel**. $\triangleleft$

Example 4.9 can be used to introduce the following kernel that is often used in practice and will be considered in several parts of the book.

Proposition 4.10. *For $d \in \mathbb{N}$, $\gamma > 0$, $z = (z_1, \dots, z_d) \in \mathbb{C}^d$, and $z' = (z'_1, \dots, z'_d) \in \mathbb{C}^d$, we define*

$$k_{\gamma,\mathbb{C}^d}(z,z') := \exp\Bigl(-\gamma^{-2}\sum_{j=1}^{d}(z_j - \bar{z}'_j)^2\Bigr).$$

*Then $k_{\gamma,\mathbb{C}^d}$ is a kernel on $\mathbb{C}^d$, and its restriction $k_\gamma := (k_{\gamma,\mathbb{C}^d})_{|\mathbb{R}^d\times\mathbb{R}^d}$ is an $\mathbb{R}$-valued kernel, which is called the **Gaussian RBF kernel** with width γ. Moreover, k_γ can be computed by*

$$k_\gamma(x,x') = \exp\Bigl(-\frac{\|x-x'\|_2^2}{\gamma^2}\Bigr), \qquad x, x' \in \mathbb{R}^d.$$

Proof. Let us fix $z, z' \in \mathbb{C}^d$. Decomposing $k_{\gamma,\mathbb{C}^d}$ into

$$k_{\gamma,\mathbb{C}^d}(z,z') = \frac{\exp(2\gamma^{-2}\langle z, z'\rangle)}{\exp\bigl(\gamma^{-2}\sum_{j=1}^{d} z_j^2\bigr)\exp\bigl(\gamma^{-2}\sum_{j=1}^{d}(\bar{z}'_j)^2\bigr)}$$

and applying Lemmas 4.3 and 4.6, and Example 4.9 then yields the first assertion. The second assertion is trivial. □

Besides the Gaussian RBF kernel, many other $\mathbb{R}$-valued kernels can be constructed using Lemma 4.8. Here we only give one more example and refer to Exercise 4.1 for some more examples.

Example 4.11. Let $X := \{x \in \mathbb{R}^d : \|x\|_2 < 1\}$ and $\alpha > 0$. Then $k(x,x') := (1 - \langle x, x'\rangle)^{-\alpha}$ defines a kernel on X called a **binomial kernel**. Indeed, the binomial series $(1-t)^{-\alpha} = \sum_{n=0}^{\infty}\binom{-\alpha}{n}(-1)^n t^n$ holds for all $|t| < 1$, where $\binom{\beta}{n} := \prod_{i=1}^{n}(\beta - i + 1)/i$. Now the assertion follows from $\binom{-\alpha}{n}(-1)^n > 0$. $\triangleleft$

The results above are based on Taylor series expansions. Instead of these expansions, one can also employ Fourier series expansions for constructing kernels. In the case $\mathbb{K} = \mathbb{R}$, the corresponding result reads as follows.

Lemma 4.12. *Let $f : [-2\pi, 2\pi] \to \mathbb{R}$ be a continuous function that can be expanded in a pointwise convergent Fourier series of the form*

$$f(t) = \sum_{n=0}^{\infty} a_n \cos(nt). \tag{4.4}$$

*If $a_n \geq 0$ holds for all $n \geq 0$, then $k(x,x') := \prod_{i=1}^{d} f(x_i - x'_i)$ defines a kernel on $[0, 2\pi)^d$. We say that k is a kernel of **Fourier type**.*

Proof. By induction and Lemma 4.6, we may restrict ourselves to $d = 1$. Then, letting $t = 0$ in (4.4), we get $(a_n)_{n\geq 0} \in \ell_1$, and thus the definition of k yields

$$k(x, x') = a_0 + \sum_{n=1}^{\infty} a_n \sin(nx)\sin(nx') + \sum_{n=1}^{\infty} a_n \cos(nx)\cos(nx')$$

for all $x, x' \in [0, 2\pi)$. Now the assertion follows from Lemma 4.2. □

The following two examples can be treated with the help of Lemma 4.12.

Example 4.13. For fixed $0 < q < 1$ and all $t \in [-2\pi, 2\pi]$, we define

$$f(t) := \frac{1 - q^2}{2 - 4q\cos t + 2q^2}\,.$$

Then $k(x, x') := \prod_{i=1}^{d} f(x_i - x_i')$, $x, x' \in [0, 2\pi]^d$, is a kernel since we have $f(t) = 1/2 + \sum_{n=1}^{\infty} q^n \cos(nt)$ for all $t \in [0, 2\pi]$. ◁

Example 4.14. For fixed $1 < q < \infty$ and all $t \in [-2\pi, 2\pi]$, we define

$$f(t) := \frac{\pi q \cosh(\pi q - qt)}{2\sinh(\pi q)}\,.$$

Then $k(x, x') := \prod_{i=1}^{d} f(x_i - x_i')$, $x, x' \in [0, 2\pi]^d$, is a kernel since we have $f(t) = 1/2 + \sum_{n=1}^{\infty}(1 + q^{-2}n^2)^{-1} \cos(nt)$ for all $t \in [0, 2\pi]$. ◁

Although we have already seen several techniques to construct kernels, in general we still have to find a feature space in order to decide whether a given function k is a kernel. Since this can sometimes be a difficult task, we will now present a criterion that characterizes $\mathbb{R}$-valued kernels in terms of *inequalities.* To this end, we need the following definition.

Definition 4.15. *A function $k : X \times X \to \mathbb{R}$ is called* ***positive definite*** *if, for all $n \in \mathbb{N}$, $\alpha_1, \dots, \alpha_n \in \mathbb{R}$ and all $x_1, \dots, x_n \in X$, we have*

$$\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j k(x_j, x_i) \geq 0\,. \tag{4.5}$$

Furthermore, k is said to be ***strictly positive definite*** *if, for mutually distinct $x_1, \dots, x_n \in X$, equality in (4.5) only holds for $\alpha_1 = \dots = \alpha_n = 0$. Finally, k is called* ***symmetric*** *if $k(x, x') = k(x', x)$ for all $x, x' \in X$.*

Unfortunately, there is no common use of the preceding definitions in the literature. Indeed, some authors call positive definite functions *positive semi-definite*, and strictly positive definite functions are sometimes called positive definite. Moreover, for fixed $x_1, \dots, x_n \in X$, the $n \times n$ matrix $K := (k(x_j, x_i))_{i,j}$ is often called the **Gram matrix**. Note that (4.5) is equivalent to saying that the Gram matrix is positive definite.

Obviously, if k is an $\mathbb{R}$-valued kernel with feature map $\Phi : X \to H$, then k is symmetric since the inner product in H is symmetric. Moreover, k is also positive definite since for $n \in \mathbb{N}$, $\alpha_1, \dots, \alpha_n \in \mathbb{R}$, and $x_1, \dots, x_n \in X$ we have

$$\sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j k(x_j, x_i) = \Bigl\langle \sum_{i=1}^{n} \alpha_i \Phi(x_i), \sum_{j=1}^{n} \alpha_j \Phi(x_j) \Bigr\rangle_H \geq 0 . \tag{4.6}$$

Now the following theorem shows that symmetry and positive definiteness are not only necessary for k to be a kernel but also sufficient.

Theorem 4.16 (Symmetric, positive definite functions are kernels). *A function $k : X \times X \to \mathbb{R}$ is a kernel if and only if it is symmetric and positive definite.*

Proof. In view of the discussion above, it suffices to show that a symmetric and positive definite function k is a kernel. To this end, we write

$$H_{\mathrm{pre}} := \Bigl\{ \sum_{i=1}^{n} \alpha_i k(\,\cdot\,, x_i) : n \in \mathbb{N},\, \alpha_1, \dots, \alpha_n \in \mathbb{R},\, x_1, \dots, x_n \in X \Bigr\} ,$$

and for $f := \sum_{i=1}^{n} \alpha_i k(\,\cdot\,, x_i) \in H_{\mathrm{pre}}$ and $g := \sum_{j=1}^{m} \beta_j k(\,\cdot\,, x'_j) \in H_{\mathrm{pre}}$, we define

$$\langle f, g \rangle := \sum_{i=1}^{n} \sum_{j=1}^{m} \alpha_i \beta_j k(x'_j, x_i) .$$

Note that this definition is independent of the representation of f since we have $\langle f, g \rangle = \sum_{j=1}^{m} \beta_j f(x'_j)$. Furthermore, by the symmetry of k, we find $\langle f, g \rangle = \sum_{i=1}^{n} \alpha_i g(x_i)$, i.e., the definition is also independent of the representation of g. Moreover, the definition immediately shows that $\langle \cdot\,, \cdot \rangle$ is bilinear and symmetric, and since k is positive definite, $\langle \cdot\,, \cdot \rangle$ is also positive, i.e., $\langle f, f \rangle \geq 0$ for all $f \in H_{\mathrm{pre}}$. Conequently (see Exercise 4.3), $\langle \cdot\,, \cdot \rangle$ satisfies the Cauchy-Schwarz inequality, i.e.,

$$|\langle f, g \rangle|^2 \leq \langle f, f \rangle \cdot \langle g, g \rangle , \qquad f, g \in H_{\mathrm{pre}} .$$

Now let $f := \sum_{i=1}^{n} \alpha_i k(\,\cdot\,, x_i) \in H_{\mathrm{pre}}$ with $\langle f, f \rangle = 0$. Then, for all $x \in X$, we have

$$|f(x)|^2 = \Bigl| \sum_{i=1}^{n} \alpha_i k(x, x_i) \Bigr|^2 = \bigl| \langle f, k(\,\cdot\,, x) \rangle \bigr|^2 \leq \bigl\langle k(\,\cdot\,, x), k(\,\cdot\,, x) \bigr\rangle \cdot \langle f, f \rangle = 0 ,$$

and hence we find $f = 0$. Therefore we have shown that $\langle \cdot\,, \cdot \rangle$ is an inner product on H_{pre}. Let H be a completion of H_{pre} and $I : H_{\mathrm{pre}} \to H$ be the corresponding isometric embedding. Then H is a Hilbert space and we have

$$\langle Ik(\,\cdot\,, x'), Ik(\,\cdot\,, x) \rangle_H = \langle k(\,\cdot\,, x'), k(\,\cdot\,, x) \rangle_{H_{\mathrm{pre}}} = k(x, x')$$

for all $x, x' \in X$, i.e., $x \mapsto Ik(\,\cdot\,, x)$, $x \in X$, defines a feature map of k. □

The characterization above is often useful for checking whether a given function is a kernel. Let us illustrate this with the following example.

Corollary 4.17 (Limits of kernels are kernels). *Let (k_n) be a sequence of kernels on the set X that converges pointwise to a function $k : X \times X \to \mathbb{R}$, i.e., $\lim_{n\to\infty} k_n(x, x') = k(x, x')$ for all $x, x' \in X$. Then k is a kernel on X.*

Proof. Every k_n is symmetric and satisfies (4.5). Therefore, the same is true for the pointwise limit k. □

4.2 The Reproducing Kernel Hilbert Space of a Kernel

In this section, we will introduce reproducing kernel Hilbert spaces (RKHSs) and describe their relation to kernels. In particular, it will turn out that the RKHS of a kernel is in a certain sense the smallest feature space of this kernel and consequently can serve as a canonical feature space.

Let us begin with the following fundamental definitions.

Definition 4.18. *Let $X \neq \emptyset$ and H be a $\mathbb{K}$-Hilbert function space over X, i.e., a $\mathbb{K}$-Hilbert space that consists of functions mapping from X into $\mathbb{K}$.*

*i) A function $k : X \times X \to \mathbb{K}$ is called a **reproducing kernel** of H if we have $k(\,\cdot\,, x) \in H$ for all $x \in X$ and the **reproducing property***

$$f(x) = \langle f, k(\,\cdot\,, x)\rangle$$

holds for all $f \in H$ and all $x \in X$.

*ii) The space H is called a **reproducing kernel Hilbert space (RKHS)** over X if for all $x \in X$ the Dirac functional $\delta_x : H \to \mathbb{K}$ defined by*

$$\delta_x(f) := f(x)\,, \qquad\qquad f \in H,$$

is continuous.

Note that $L_2(\mathbb{R}^d)$ does *not* consist of functions and consequently it is not an RKHS. For a generalization of this statement, we refer to Exercise 4.2.

Reproducing kernel Hilbert spaces have the remarkable and, as we will see later, important property that norm convergence implies pointwise convergence. More precisely, let H be an RKHS, $f \in H$, and $(f_n) \subset H$ be a sequence with $\|f_n - f\|_H \to 0$ for $n \to \infty$. Then, for all $x \in X$, we have

$$\lim_{n\to\infty} f_n(x) = \lim_{n\to\infty} \delta_x(f_n) = \delta_x(f) = f(x) \tag{4.7}$$

by the assumed continuity of the Dirac functionals. Furthermore, reproducing kernels are actually kernels in the sense of Definition 4.1, as the following lemma shows.

Lemma 4.19 (Reproducing kernels are kernels). *Let H be a Hilbert function space over X that has a reproducing kernel k. Then H is an RKHS and H is also a feature space of k, where the feature map $\Phi : X \to H$ is given by*

$$\Phi(x) = k(\,\cdot\,, x)\,, \qquad\qquad x \in X.$$

We call Φ the ***canonical feature map****.*

Proof. The reproducing property says that each Dirac functional can be represented by the reproducing kernel, and consequently we obtain

$$|\delta_x(f)| = |f(x)| = |\langle f, k(\,\cdot\,, x)\rangle| \le \|k(\,\cdot\,, x)\|_H\, \|f\|_H \tag{4.8}$$

for all $x \in X$, $f \in H$. This shows the continuity of the functionals δ_x, $x \in X$.

In order to show the second assertion, we fix an $x' \in X$ and write $f := k(\,\cdot\,, x')$. Then, for $x \in X$, the reproducing property yields

$$\langle \Phi(x'), \Phi(x)\rangle = \langle k(\,\cdot\,, x'), k(\,\cdot\,, x)\rangle = \langle f, k(\,\cdot\,, x)\rangle = f(x) = k(x, x')\,. \qquad \square$$

We have just seen that every Hilbert function space with a reproducing kernel is an RKHS. The following theorem now shows that, conversely, every RKHS has a (unique) reproducing kernel, and that this kernel can be determined by the Dirac functionals.

Theorem 4.20 (Every RKHS has a unique reproducing kernel). *Let H be an RKHS over X. Then $k : X \times X \to \mathbb{K}$ defined by*

$$k(x, x') := \langle \delta_x, \delta_{x'}\rangle_H\,, \qquad x, x' \in X,$$

is the only reproducing kernel of H. Furthermore, if $(e_i)_{i\in I}$ is an orthonormal basis of H, then for all $x, x' \in X$ we have

$$k(x, x') = \sum_{i\in I} e_i(x)\overline{e_i(x')}\,. \tag{4.9}$$

Proof. We first show that k is a reproducing kernel. To this end, let $I : H' \to H$ be the isometric anti-linear isomorphism derived from Theorem A.5.12 that assigns to every functional in H' the representing element in H, i.e., $g'(f) = \langle f, Ig'\rangle$ for all $f \in H$, $g' \in H'$. Then, for all $x, x' \in X$, we have

$$k(x, x') = \langle \delta_x, \delta_{x'}\rangle_{H'} = \langle I\delta_{x'}, I\delta_x\rangle_H = \delta_x(I\delta_{x'}) = I\delta_{x'}(x)\,,$$

which shows $k(\,\cdot\,, x') = I\delta_{x'}$ for all $x' \in X$. From this we immediately obtain

$$f(x') = \delta_{x'}(f) = \langle f, I\delta_{x'}\rangle = \langle f, k(\,\cdot\,, x')\rangle\,,$$

i.e., k has the reproducing property. Now let $\tilde{k}$ be an *arbitrary* reproducing kernel of H. For $x' \in X$, the basis representation of $\tilde{k}(\,\cdot\,, x')$ then yields

$$\tilde{k}(\,\cdot\,,x') = \sum_{i\in I}\langle \tilde{k}(\,\cdot\,,x'), e_i\rangle e_i = \sum_{i\in I}\overline{e_i(x')}e_i\,,$$

where the convergence is with respect to $\|\cdot\|_H$. Using (4.7), we thus obtain (4.9) for $\tilde{k}$. Finally, since $\tilde{k}$ and $(e_i)_{i\in I}$ were arbitrarily chosen, we find $\tilde{k} = k$, i.e., k is the only reproducing kernel of H. □

Theorem 4.20 shows that an RKHS uniquely determines its reproducing kernel, which is actually a kernel by Lemma 4.19. The following theorem now shows that, conversely, every kernel has a unique RKHS. Consequently, we have a one-to-one relation between kernels and RKHSs. In addition, the following theorem shows that the RKHS of a kernel is in some sense the smallest feature space, and hence it can be considered as the "natural" feature space.

Theorem 4.21 (Every kernel has a unique RKHS). *Let $X \neq \emptyset$ and k be a kernel over X with feature space H_0 and feature map $\Phi_0 : X \to H_0$. Then*

$$H := \bigl\{f : X \to \mathbb{K}\;\big|\;\exists\, w \in H_0 \text{ with } f(x) = \langle w, \Phi_0(x)\rangle_{H_0} \text{ for all } x \in X\bigr\} \qquad (4.10)$$

equipped with the norm

$$\|f\|_H := \inf\bigl\{\|w\|_{H_0} : w \in H_0 \text{ with } f = \langle w, \Phi_0(\,\cdot\,)\rangle_{H_0}\bigr\} \qquad (4.11)$$

is the only RKHS for which k is a reproducing kernel. Consequently, both definitions are independent of the choice of H_0 and Φ_0. Moreover, the operator $V : H_0 \to H$ defined by

$$Vw := \langle w, \Phi_0(\,\cdot\,)\rangle_{H_0}\,, \qquad\qquad w \in H_0,$$

is a metric surjection, i.e. $V\mathring{B}_{H_0} = \mathring{B}_H$, where $\mathring{B}_{H_0}$ and $\mathring{B}_H$ are the open unit balls of H_0 and H, respectively. Finally, the set

$$H_{\mathrm{pre}} := \left\{\sum_{i=1}^n \alpha_i k(\,\cdot\,,x_i) : n \in \mathbb{N},\ \alpha_1,\dots,\alpha_n \in \mathbb{K},\ x_1,\dots,x_n \in X\right\} \qquad (4.12)$$

is dense in H, and for $f := \sum_{i=1}^n \alpha_i k(\,\cdot\,,x_i) \in H_{\mathrm{pre}}$ we have

$$\|f\|_H^2 = \sum_{i=1}^n\sum_{j=1}^n \alpha_i\overline{\alpha_j}k(x_j,x_i)\,. \qquad (4.13)$$

Proof. Let us first show that H is a Hilbert function space over X. To this end, observe that H is obviously a vector space of functions from X to $\mathbb{K}$, and V is a surjective linear operator. Furthermore, for all $f \in H$, we have

$$\|f\|_H = \inf_{w\in V^{-1}(\{f\})}\|w\|_{H_0}\,,$$

where $V^{-1}(\{f\})$ denotes the pre-image of f under V. Let us show that $\|\cdot\|_H$ is a Hilbert space norm on H. To this end, let $(w_n) \subset \ker V$ be a convergent sequence in the null space $\ker V := \{w \in H_0 : Vw = 0\}$ of V and $w \in H_0$ its limit. Then we have $\langle w, \Phi(x)\rangle = \lim_{n\to\infty}\langle w_n, \Phi(x)\rangle = 0$ for all $x \in X$. Since this shows $w \in \ker V$, the null space $\ker V$ is a closed subspace of H_0. Let $\hat{H}$ denote its orthogonal complement so that we have the orthogonal decomposition $H_0 = \ker V \oplus \hat{H}$. Then the restriction $V_{|\hat{H}} : \hat{H} \to H$ of V to $\hat{H}$ is injective by construction. Let us show that it is also surjective. To this end, let $f \in H$ and $w \in H_0$ with $f = Vw$. Since this w can be decomposed into $w = w_0 + \hat{w}$ for suitable $w_0 \in \ker V$ and $\hat{w} \in \hat{H}$, we get $f = V(w_0 + \hat{w}) = V_{|\hat{H}}\hat{w}$, which shows the surjectivity of $V_{|\hat{H}}$. Furthermore, a similar reasoning gives

$$\begin{aligned}\|f\|_H^2 = \inf_{\substack{w_0\in\ker V,\, \hat{w}\in\hat{H}\\ w_0+\hat{w}\in V^{-1}(\{f\})}} \|w_0+\hat{w}\|_{H_0}^2 &= \inf_{\substack{w_0\in\ker V,\, \hat{w}\in\hat{H}\\ w_0+\hat{w}\in V^{-1}(\{f\})}} \|w_0\|_{H_0}^2 + \|\hat{w}\|_{H_0}^2 \\ &= \big\| (V_{|\hat{H}})^{-1} f \big\|_{\hat{H}}^2 ,\end{aligned}$$

where $(V_{|\hat{H}})^{-1}$ denotes the inverse operator of $V_{|\hat{H}}$. From the equation above and the fact that $\hat{H}$ is a Hilbert space, we can immediately deduce that $\|\cdot\|_H$ is a Hilbert space norm on H and that $V_{|\hat{H}} : \hat{H} \to H$ is an isometric isomorphism. Furthermore, from the definition of V and $\|\cdot\|_H$, we can easily conclude that V is a *metric* surjection.

Let us now show that k is a reproducing kernel of H. To this end, observe that for $x \in X$ we have $k(\,\cdot\,, x) = \langle \Phi_0(x), \Phi_0(\,\cdot\,)\rangle = V\Phi_0(x) \in H$. Furthermore, we have $\langle w, \Phi_0(x)\rangle = 0$ for all $w \in \ker V$, which shows $\Phi_0(x) \in (\ker V)^\perp = \hat{H}$. Since $V_{|\hat{H}} : \hat{H} \to H$ is isometric, we therefore obtain

$$f(x) = \big\langle (V_{|\hat{H}})^{-1} f, \Phi_0(x)\big\rangle_{H_0} = \big\langle f, V_{|\hat{H}}\Phi_0(x)\big\rangle_H = \langle f, k(\,\cdot\,, x)\rangle_H$$

for all $f \in H$, $x \in X$, i.e., k has the reproducing property. By Lemma 4.19 we conclude that H is an RKHS.

Let us now show that the assertions on H_{pre} are true for an *arbitrary* RKHS $\tilde{H}$ for which k is a reproducing kernel. To this end, we first observe that $k(\,\cdot\,, x) \in \tilde{H}$ for all $x \in X$ implies $H_{\mathrm{pre}} \subset \tilde{H}$. Now let us suppose that H_{pre} was not dense in $\tilde{H}$. This assumption yields $(H_{\mathrm{pre}})^\perp \neq \{0\}$, and hence there would exist an $f \in (H_{\mathrm{pre}})^\perp$ and an $x \in X$ with $f(x) \neq 0$. Since this would imply

$$0 = \langle f, k(\,\cdot\,, x)\rangle = f(x) \neq 0\,,$$

we see that H_{pre} is dense in $\tilde{H}$. Finally, for $f := \sum_{i=1}^n \alpha_i k(\,\cdot\,, x_i) \in H_{\mathrm{pre}}$, the reproducing property implies

$$\|f\|_{\tilde{H}}^2 = \sum_{i=1}^n \sum_{j=1}^n \alpha_i \overline{\alpha_j} \big\langle k(\,\cdot\,, x_i), k(\,\cdot\,, x_j)\big\rangle_{\tilde{H}} = \sum_{i=1}^n \sum_{j=1}^n \alpha_i \overline{\alpha_j} k(x_j, x_i)\,.$$

Let us now prove that k has only one RKHS. To this end, let H_1 and H_2 be two RKHSs of k. We have seen in the previous step that H_{pre} is dense in both

H_1 and H_2 and that the norms of H_1 and H_2 coincide on H_{pre}. Let us choose an $f \in H_1$. Then there exists a sequence $(f_n) \subset H_{\mathrm{pre}}$ with $\|f_n - f\|_{H_1} \to 0$. Since $H_{\mathrm{pre}} \subset H_2$, the sequence (f_n) is also contained in H_2, and since the norms of H_1 and H_2 coincide on H_{pre}, the sequence (f_n) is a Cauchy sequence in H_2. Therefore, there exists a $g \in H_2$ with $\|f_n - g\|_{H_2} \to 0$. Since convergence with respect to an RKHS norm implies pointwise convergence, see (4.7), we then find $f(x) = g(x)$ for all $x \in X$, i.e., we have shown $f \in H_2$. Furthermore, $\|f_n - f\|_{H_1} \to 0$ and $\|f_n - f\|_{H_2} \to 0$ imply

$$\|f\|_{H_1} = \lim_{n\to\infty} \|f_n\|_{H_1} = \lim_{n\to\infty} \|f_n\|_{H_{\mathrm{pre}}} = \lim_{n\to\infty} \|f_n\|_{H_2} = \|f\|_{H_2} \,,$$

i.e., H_1 is isometrically included in H_2. Since the converse inclusion $H_2 \subset H_1$ can be shown analogously, we obtain $H_1 = H_2$ with equal norms. □

Theorem 4.21 describes the RKHS H of a given kernel k as the "smallest" feature space of k in the sense that there is a canonical metric surjection V from any other feature space H_0 of k onto H. However, for kernelized algorithms, it is more the specific *form* (4.10) that makes the RKHS important. To illustrate this, recall from the introduction that the soft margin SVM produces decision functions of the form $x \mapsto \langle w, \Phi_0(x)\rangle$, where $\Phi_0 : X \to H_0$ is a feature map of k and $w \in H_0$ is a suitable weight vector. Now, (4.10) states that the RKHS associated to k consists exactly of all possible functions of this form. Moreover, (4.10) shows that this set of functions does not change if we consider different feature spaces or feature maps of k.

Theorem 4.21 can often be used to determine the RKHS of a given kernel and its modifications such as restrictions and normalization (see Exercise 4.4 for more details). To illustrate this, let us recall that every $\mathbb{C}$-valued kernel on X that is actually $\mathbb{R}$-valued has an $\mathbb{R}$-feature space by Lemma 4.4. The following corollary of Theorem 4.21 describes the corresponding $\mathbb{R}$-RKHS.

Corollary 4.22. *Let $k : X \times X \to \mathbb{C}$ be a kernel and H its corresponding $\mathbb{C}$-RKHS. If we actually have $k(x, x') \in \mathbb{R}$ for all $x, x' \in X$, then*

$$H_{\mathbb{R}} := \bigl\{ f : X \to \mathbb{R} \,\big|\, \exists\, g \in H \ \textit{with} \ \mathrm{Re}\, g = f \bigr\}$$

equipped with the norm

$$\|f\|_{H_{\mathbb{R}}} := \inf\bigl\{ \|g\|_H : g \in H \ \textit{with} \ \mathrm{Re}\, g = f \bigr\} \,, \qquad\qquad f \in H_{\mathbb{R}},$$

is the $\mathbb{R}$-RKHS of the $\mathbb{R}$-valued kernel k.

Proof. We have already seen in Lemma 4.4 that $H_0 := H$ equipped with the inner product

$$\langle f, f'\rangle_{H_0} := \mathrm{Re}\, \langle f, f'\rangle_H \,, \qquad\qquad f,\, f' \in H_0,$$

is an $\mathbb{R}$-feature space of the $\mathbb{R}$-valued kernel k. Moreover, for $f \in H_0$ and $x \in X$, we have

$$f(x) = \langle f, \Phi(x)\rangle_H = \mathrm{Re}\,\langle f, \Phi(x)\rangle_H + i\,\mathrm{Im}\,\langle f, \Phi(x)\rangle_H = \langle f, \Phi(x)\rangle_{H_0} + i\,\mathrm{Im}\, f(x),$$

i.e., we have found $\langle f, \Phi(x)\rangle_{H_0} = \mathrm{Re}\, f(x)$. Now, the assertion follows from Theorem 4.21. □

Let us finally assume that we have an RKHS H with kernel $k : \mathbb{C}^d \times \mathbb{C}^d \to \mathbb{C}$ whose restriction to $\mathbb{R}^d$ is $\mathbb{R}$-valued, i.e., $k_{|\mathbb{R}^d \times \mathbb{R}^d} : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$. Then combining the preceding corollary with Exercise 4.4 shows that

$$H_{\mathbb{R}} := \bigl\{ f : \mathbb{R}^d \to \mathbb{R} \,\big|\, \exists\, g : \mathbb{C}^d \to \mathbb{C} \text{ with } g \in H \text{ and } \mathrm{Re}\, g_{|\mathbb{R}^d} = f \bigr\}$$

equipped with the norm

$$\|f\|_{H_{\mathbb{R}}} := \inf\bigl\{ \|g\|_H : g \in H \text{ with } \mathrm{Re}\, g_{|\mathbb{R}^d} = f \bigr\}, \qquad f \in H_{\mathbb{R}},$$

is the $\mathbb{R}$-RKHS of the restriction $k_{|\mathbb{R}^d \times \mathbb{R}^d}$.

4.3 Properties of RKHSs

Usually, a kernel has additional properties such as measurability, continuity, or differentiability. In this section, we investigate whether the functions of its associated RKHS share these properties.

Let us begin by observing that for a kernel k on X with RKHS H the Cauchy-Schwarz inequality and the reproducing property imply

$$\begin{aligned} |k(x,x')|^2 = \bigl|\bigl\langle k(\,\cdot\,,x'), k(\,\cdot\,,x)\bigr\rangle_H\bigr|^2 &\le \|k(\,\cdot\,,x')\|_H^2 \cdot \|k(\,\cdot\,,x)\|_H^2 \\ &= k(x',x') \cdot k(x,x) \end{aligned} \tag{4.14}$$

for all $x, x' \in X$. This yields $\sup_{x,x' \in X} |k(x,x')| = \sup_{x \in X} k(x,x)$, and hence k is **bounded** if and only if

$$\|k\|_\infty := \sup_{x \in X} \sqrt{k(x,x)} < \infty\,. \tag{4.15}$$

Now, let $\Phi : X \to H_0$ be a feature map of k. Then we find $\|\Phi(x)\|_{H_0} = \sqrt{k(x,x)}$ for all $x \in X$, and hence Φ is bounded if and only if k is. The following lemma provides another important characterization of bounded kernels.

Lemma 4.23 (RKHSs of bounded kernels). *Let X be a set and k be a kernel on X with RKHS H. Then k is bounded if and only if every $f \in H$ is bounded. Moreover, in this case the inclusion* $\mathrm{id} : H \to \ell_\infty(X)$ *is continuous and we have* $\|\,\mathrm{id} : H \to \ell_\infty(X)\| = \|k\|_\infty$.

Proof. Let us first assume that k is bounded. Then the Cauchy-Schwarz inequality and the reproducing property imply

$$|f(x)| = |\langle f, k(\,\cdot\,,x)\rangle| \le \|f\|_H \sqrt{k(x,x)} \le \|f\|_H \|k\|_\infty$$

for all $f \in H, x \in X$. Hence we obtain $\|f\|_\infty \leq \|k\|_\infty \|f\|_H$ for all $f \in H$, which shows that $\mathrm{id} : H \to \ell_\infty(X)$ is well-defined and $\|\mathrm{id} : H \to \ell_\infty(X)\| \leq \|k\|_\infty$.

Conversely, let us now assume that every $f \in H$ is bounded. Then the inclusion $\mathrm{id} : H \to \ell_\infty(X)$ is well-defined. Our first goal is to show that the inclusion is continuous. To this end, we fix a sequence $(f_n) \subset H$ for which there exist an $f \in H$ and a $g \in \ell_\infty(X)$ such that $\lim_{n\to\infty} \|f_n - f\|_H = 0$ and $\lim_{n\to\infty} \|\mathrm{id}\, f_n - g\|_\infty = 0$. Then the first convergence implies $f_n(x) \to f(x)$ for all $x \in X$, while the second convergence implies $f_n(x) \to g(x)$ for all $x \in X$. We conclude $f = g$ and hence $\mathrm{id} : H \to \ell_\infty(X)$ is continuous by the closed graph theorem, see Theorem A.5.4. For $x \in X$, we then have

$$|k(x,x)| \leq \|k(\cdot\,,x)\|_\infty \leq \|\mathrm{id} : H \to \ell_\infty(X)\| \cdot \|k(\cdot\,,x)\|_H = \|\mathrm{id}\|\sqrt{k(x,x)}\,,$$

i.e., $\sqrt{k(x,x)} \leq \|\mathrm{id}\|$. This shows $\|k\|_\infty \leq \|\mathrm{id} : H \to \ell_\infty(X)\|$. □

Our next goal is to investigate measurable kernels and their integrability. We begin with the following lemma.

Lemma 4.24 (RKHSs of measurable kernels). *Let X be a measurable space and k be a kernel on X with RKHS H. Then all $f \in H$ are measurable if and only if $k(\cdot\,,x) : X \to \mathbb{R}$ is measurable for all $x \in X$.*

Proof. If all $f \in H$ are measurable, then $k(\cdot\,,x) \in H$ is measurable for all $x \in X$. Conversely, if $k(\cdot\,,x)$ is measurable for all $x \in X$, then all functions $f \in H_{\mathrm{pre}}$ are measurable, where H_{pre} is defined by (4.12). Let us now fix an $f \in H$. By Theorem 4.21, there then exists a sequence $(f_n) \subset H_{\mathrm{pre}}$ with $\lim_{n\to\infty} \|f_n - f\|_H = 0$, and since all Dirac functionals are continuous, we obtain $\lim_{n\to\infty} f_n(x) = f(x)$, $x \in X$. This gives the measurability of f. □

The next lemma investigates the measurability of canonical feature maps.

Lemma 4.25 (Measurability of the canonical feature map). *Let X be a measurable space and k be a kernel on X such that $k(\cdot\,,x) : X \to \mathbb{R}$ is measurable for all $x \in X$. If the RKHS H of k is separable, then both the canonical feature map $\Phi : X \to H$ and $k : X \times X \to \mathbb{R}$ are measurable.*

Proof. Let $w \in H'$ be a bounded linear functional. By the Fréchet-Riesz representation theorem (see Theorem A.5.12) there then exists an $f \in H$ with

$$\langle w, \Phi(x)\rangle_{H',H} = \langle f, \Phi(x)\rangle_H = f(x)\,, \qquad\qquad x \in X,$$

and hence $\langle w, \Phi(\cdot\,)\rangle_{H',H} : X \to \mathbb{R}$ is measurable by Lemma 4.24. By Petti's measurability theorem (see Theorem A.5.19), we then obtain the measurability of Φ. The second assertion now follows from $k(x,x') = \langle \Phi(x'), \Phi(x)\rangle$ and the fact that the inner product is continuous. □

Our next goal is to investigate under which assumptions on the kernel k the functions of its RKHS are integrable. To this end, recall that $x \mapsto k(x,x)$ is a non-negative function, and hence its integral is always defined, though in general it may not be finite.

Theorem 4.26 (Integral operators of kernels I). *Let X be a measurable space, μ be a σ-finite measure on X, and H be a separable RKHS over X with measurable kernel $k : X \times X \to \mathbb{R}$. Assume that there exists a $p \in [1, \infty)$ such that*

$$\|k\|_{L_p(\mu)} := \left(\int_X k^{p/2}(x,x) d\mu(x) \right)^{1/p} < \infty\,. \tag{4.16}$$

Then H consists of p-integrable functions and the inclusion $\mathrm{id} : H \to L_p(\mu)$ *is continuous with* $\|\,\mathrm{id} : H \to L_p(\mu)\| \leq \|k\|_{L_p(\mu)}$. *Moreover, the adjoint of this inclusion is the operator $S_k : L_{p'}(\mu) \to H$ defined by*

$$S_k g(x) := \int_X k(x,x')g(x')d\mu(x')\,, \qquad g \in L_{p'}(\mu),\ x \in X, \tag{4.17}$$

where p' is defined by $\frac{1}{p} + \frac{1}{p'} = 1$. Finally, the following statements are true:

i) H is dense in $L_p(\mu)$ if and only if $S_k : L_{p'}(\mu) \to H$ is injective.
ii) $S_k : L_{p'}(\mu) \to H$ has a dense image if and only if $\mathrm{id} : H \to L_p(\mu)$ *is injective.*

Proof. Let us fix an $f \in H$. Using $\|k(\,\cdot\,,x)\|_H = \sqrt{k(x,x)}$, we then find

$$\int_X |f(x)|^p d\mu(x) = \int_X \big|\langle f, k(\,\cdot\,,x)\rangle\big|^p d\mu(x) \leq \|f\|_H^p \int_X k^{p/2}(x,x) d\mu(x)\,,$$

which shows the first two assertions. Furthermore, for $g \in L_{p'}(\mu)$, inequality (4.14) together with Hölder's inequality yields

$$\begin{aligned} \int_X \big|k(x,x')g(x')\big|\, d\mu(x') &\leq \sqrt{k(x,x)} \int_X \sqrt{k(x',x')}\, |g(x')|\, d\mu(x') \\ &\leq \sqrt{k(x,x)}\, \|k\|_{L_p(\mu)}\, \|g\|_{L_{p'}(\mu)}\,, \end{aligned} \tag{4.18}$$

and hence $x' \mapsto k(x,x')g(x')$ is integrable. In other words, the integral defining $S_k g(x)$ exists for all $x \in X$. Moreover, since $\sqrt{k(x',x')} = \|\Phi(x')\|_H$, the second inequality in (4.18) shows $(x' \mapsto \|g(x')\Phi(x')\|_H) \in L_1(\mu)$, i.e., this function is Bochner integrable and

$$\bar{g} := \int_X g(x')\Phi(x')\, d\mu(x') \in H\,.$$

Moreover, (A.32) applied to the bounded linear operator $\langle \,\cdot\,, \Phi(x)\rangle : H \to \mathbb{R}$ yields

$$S_k g(x) = \int_X \langle \Phi(x'), \Phi(x)\rangle_H\, g(x')\, d\mu(x') = \left\langle \int_X g(x')\Phi(x')\, d\mu(x'), \Phi(x) \right\rangle_H$$

for all $x \in X$, and hence we conclude that $S_k g = \bar{g} \in H$. For $f \in H$, another application of (A.32) yields

$$\begin{aligned}\langle g, \mathrm{id}\, f\rangle_{L_{p'}(\mu),L_p(\mu)} = \int_X g(x)\langle f, k(\,\cdot\,,x)\rangle_H \, d\mu(x) &= \left\langle f, \int_X g(x)k(\,\cdot\,,x)\, d\mu(x)\right\rangle_H \\ &= \langle f, S_k g\rangle_H \\ &= \langle \iota S_k g, f\rangle_{H',H}\,,\end{aligned}$$

where $\iota : H \to H'$ is the isometric isomorphism described in Theorem A.5.12. By identifying H' with H via ι, we then find $\mathrm{id}' = S_k$. Finally, the last two assertions follow from the fact that a bounded linear operator has a dense image if and only if its adjoint is injective, as mentioned in Section A.5.1 around (A.19). □

One may be tempted to think that the "inclusion" $\mathrm{id} : H \to L_p(\mu)$ is always injective. However, since this map assigns every f to its *equivalence class* $[f]_\sim$ in $L_p(\mu)$, see (A.33), the opposite is true. To see this, consider for example an infinite-dimensional RKHS (see Section 4.6 for examples of such spaces) and an empirical measure μ. Then $L_p(\mu)$ is finite-dimensional and hence the map $\mathrm{id} : H \to L_p(\mu)$ cannot be injective. For a simple condition ensuring that $\mathrm{id} : H \to L_p(\mu)$ *is* injective, we refer to Exercise 4.6.

Let us now have a closer look at the case $p = 2$ in the preceding theorem. The following theorem shows that in this case the Hilbert space structure of $L_2(\mu)$ provides some additional features of the operator S_k which will be of particular interest in Chapter 7.

Theorem 4.27 (Integral operators of kernels II). *Let X be a measurable space with σ-finite measure μ and H be a separable RKHS over X with measurable kernel $k : X \times X \to \mathbb{R}$ satisfying $\|k\|_{L_2(\mu)} < \infty$. Then $S_k : L_2(\mu) \to H$ defined by (4.17) is a Hilbert-Schmidt operator with*

$$\|S_k\|_{\mathsf{HS}} = \|k\|_{L_2(\mu)}\,. \tag{4.19}$$

Moreover, the integral operator $T_k = S_k^* S_k : L_2(\mu) \to L_2(\mu)$ is compact, positive, self-adjoint, and nuclear with $\|T_k\|_{\mathsf{nuc}} = \|S_k\|_{\mathsf{HS}} = \|k\|_{L_2(\mu)}$.

Proof. Let us first show that $S_k^* : H \to L_2(\mu)$ is a Hilbert-Schmidt operator. To this end, let $(e_i)_{i\geq 1}$ be an ONB of H. By Theorem 4.20, we then find

$$\sum_{i=1}^{\infty} \|S_k^* e_i\|_{L_2(\mu)}^2 = \int_X \sum_{i=1}^{\infty} |S_k^* e_i(x)|^2 \, d\mu(x) = \int_X \sum_{i=1}^{\infty} e_i^2(x)\, d\mu(x) = \|k\|_{L_2(\mu)}^2 < \infty,$$

i.e., S_k^* is indeed Hilbert-Schmidt with the desired norm. Consequently, S_k is Hilbert-Schmidt, too. Now the remaining assertions follow from the spectral theory recalled around Theorem A.5.13. □

Since $S_k^* = \mathrm{id} : H \to L_2(\mu)$, one may be tempted to think that the operators T_k and S_k are the same modulo their image space. However, recall that in general $L_2(\mu)$ does not consist of functions, and hence $S_k f(x)$ is defined, while $T_k f(x)$ is *not*.

Our next goal is to investigate continuity properties of kernels. To this end, we say that a kernel k on a topological space X is **separately continuous** if $k(\,\cdot\,,x) : X \to \mathbb{R}$ is continuous for all $x \in X$. Now, our first lemma characterizes RKHSs consisting of continuous functions.

Lemma 4.28 (RKHSs consisting of continuous functions). *Let X be topological space and k be a kernel on X with RKHS H. Then k is bounded and separately continuous if and only if every $f \in H$ is a bounded and continuous function. In this case, the inclusion* $\mathrm{id} : H \to C_b(X)$ *is continuous and we have* $\|\,\mathrm{id} : H \to C_b(X)\| = \|k\|_\infty$.

Proof. Let us first assume that k is bounded and separately continuous. Then H_{pre} only contains continuous functions since k is separately continuous. Let us fix an arbitrary $f \in H$. By Theorem 4.21, there then exists a sequence $(f_n) \subset H_{\mathrm{pre}}$ with $\lim_{n\to\infty} \|f_n - f\|_H = 0$. Since k is bounded, this implies $\lim_{n\to\infty} \|f_n - f\|_\infty = 0$ by Lemma 4.23 and hence f, as a uniform limit of continuous functions, is continuous. Finally, both the continuity of $\mathrm{id} : H \to C_b(X)$ and $\|\,\mathrm{id} : H \to C_b(X)\| = \|k\|_\infty$ follow from Lemma 4.23, too.

Conversely, let us now assume that H only contains continuous functions. Then $k(\,\cdot\,,x) : X \to \mathbb{K}$ is continuous for all $x \in X$, i.e., k is separately continuous. Furthermore, the boundedness of k follows from Lemma 4.23. □

Lemma 4.28 in particular applies to continuous kernels. Let us now discuss these kernels in more detail. To this end, let k be a kernel on X with feature map $\Phi : X \to H$. Then the **kernel metric** d_k on X is defined by

$$d_k(x,x') := \|\Phi(x) - \Phi(x')\|_H\,, \qquad\qquad x,x' \in X. \tag{4.20}$$

Obviously, d_k is a pseudo-metric on X, and if Φ is injective it is even a metric. Moreover, since

$$d_k(x,x') = \sqrt{k(x,x) - 2k(x,x') + k(x',x')}\,, \tag{4.21}$$

the definition of d_k is actually *independent* of the choice of Φ. Furthermore, the kernel metric can be used to characterize the continuity of k.

Lemma 4.29 (Characterization of continuous kernels). *Let (X,τ) be a topological space and k be a kernel on X with feature space H and feature map $\Phi : X \to H$. Then the following statements are equivalent:*

i) k is continuous.
ii) k is separately continuous and $x \mapsto k(x,x)$ is continuous.
iii) Φ is continuous.
iv) $\mathrm{id} : (X,\tau) \to (X,d_k)$ *is continuous.*

Proof. i) ⇒ ii). Trivial.

ii) ⇒ iv). By (4.21) and the assumptions, we see that $d_k(\,\cdot\,,x) : (X,\tau) \to \mathbb{R}$ is continuous for every $x \in X$. Consequently, $\{x' \in X : d_k(x',x) < \varepsilon\}$ is open with respect to τ and therefore $\mathrm{id} : (X,\tau) \to (X,d_k)$ is continuous.

iv) ⇒ iii). This implication follows from the fact that $\Phi : (X, d_k) \to H$ is continuous.

iii) ⇒ i). Let us fix $x_1, x_1' \in X$ and $x_2, x_2' \in X$. Then we have

$$\begin{aligned}|k(x_1, x_1') - k(x_2, x_2')| &\leq |\langle \Phi(x_1'), \Phi(x_1) - \Phi(x_2)\rangle| + |\langle \Phi(x_1') - \Phi(x_2'), \Phi(x_2)\rangle| \\ &\leq \|\Phi(x_1')\| \cdot \|\Phi(x_1) - \Phi(x_2)\| + \|\Phi(x_2)\| \cdot \|\Phi(x_1') - \Phi(x_2')\|,\end{aligned}$$

and from this we can easily deduce the assertion. □

As discovered by Lehto (1952), separately continuous, bounded kernels are not necessarily continuous, even if one only considers $X = [-1, 1]$. However, since his example is out of the scope of this book, we do not present it here.

We have seen in Lemma 4.23 that an RKHS over X is continuously included in $\ell_\infty(X)$ if its kernel is bounded. The next proposition gives a condition that ensures that this inclusion is even compact. This compactness will play an important role when we investigate the statistical properties of support vector machines in Chapter 6.

Proposition 4.30 (RKHSs compactly included in $\ell_\infty(\mathbf{X})$). *Let X be a set and k be a kernel on X with RKHS H and canonical feature map $\Phi : X \to H$. If $\Phi(X)$ is compact in H, then the inclusion* $\mathrm{id} : H \to \ell_\infty(X)$ *is compact.*

Proof. Since $\Phi(X)$ is compact, k is bounded and the space (X, d_k) is compact, where d_k is the semi-metric defined in (4.20). We write $C(X, d_k)$ for the space of functions from X to $\mathbb{R}$ that are continuous with respect to d_k. Obviously, $C(X, d_k)$ is a subspace of $\ell_\infty(X)$. Moreover, for $x, x' \in X$ and $f \in H$, we obtain

$$|f(x) - f(x')| = |\langle f, \Phi(x) - \Phi(x')\rangle| \leq \|f\|_H \cdot d_k(x, x'),$$

i.e., f is Lipschitz continuous on (X, d_k) with Lipschitz constant not larger than $\|f\|_H$. In particular, the unit ball B_H of H is equicontinuous, and since B_H is also $\|\cdot\|_\infty$-bounded by the boundedness of k, the theorem of Arzelà-Ascoli shows that $\overline{B_H}$ is compact in $C(X, d_k)$ and thus in $\ell_\infty(X)$. □

Obviously, the proposition above remains true if one only assumes the compactness of $\Phi(X)$ for an *arbitrary* feature map Φ. Furthermore, continuous images of compact sets are compact, and hence Proposition 4.30 has the following direct consequence.

Corollary 4.31. *Let X be a compact topological space and k be a continuous kernel on X with RKHS H. Then the inclusion* $\mathrm{id} : H \to C(X)$ *is compact.*

We emphasize that in general one cannot expect compactness of the inclusion $\mathrm{id} : H \to C_b(X)$ if k is bounded and continuous but X is *not* compact. The following example illustrates this.

Example 4.32. Let k_γ be the **Gaussian RBF kernel** on $\mathbb{R}$ with width $\gamma > 0$ and RKHS $H_\gamma(\mathbb{R})$. Obviously, k_γ is bounded and continuous, and hence the inclusion $\mathrm{id} : H_\gamma(\mathbb{R}) \to C_b(\mathbb{R})$ is well-defined and continuous. Moreover, since $\|k_\gamma\|_\infty = 1$, we also have $k_\gamma(\,\cdot\,, x) \in B_{H_\gamma(\mathbb{R})}$ for all $x \in \mathbb{R}$. However, for all $n, m \in \mathbb{N}$ with $n \neq m$, we obtain

$$\|k_\gamma(\,\cdot\,, n) - k_\gamma(\,\cdot\,, m)\|_\infty \;\geq\; |k_\gamma(n,n) - k_\gamma(n,m)| \;\geq\; |1 - \exp(-\gamma^{-2})|\,,$$

and thus $B_{H_\gamma(\mathbb{R})}$ cannot be relatively compact in $C_b(\mathbb{R})$.

Let us end the discussion on continuous kernels with the following lemma that gives a sufficient condition for the separability of RKHSs.

Lemma 4.33 (Separable RKHSs). *Let X be a separable topological space and k be a continuous kernel on X. Then the RKHS of k is separable.*

Proof. By Lemma 4.29, the canonical feature map $\Phi : X \to H$ into the RKHS H of k is continuous and hence $\Phi(X)$ is separable. Consequently, H_{pre}, considered in Theorem 4.21, is also separable, and hence so is H by Theorem 4.21. □

Our last goal in this section is to investigate how the differentiability of a kernel is inherited by the functions of its RKHS. In order to formulate the next lemma, which to some extent is of its own interest, we need to recall Banach space valued differentiation from Section A.5.3. Moreover, note that we can interpret a kernel $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ as a function $\tilde{k} : \mathbb{R}^{2d} \to \mathbb{R}$. Consequently, considering the mixed partial derivative of the kernel $k(x, x')$ with respect to the i-th coordinates in x *and* x' is the same as considering the mixed partial derivative $\partial_i \partial_{i+d} \tilde{k}$ at (x, x'). In the following, we make this identification implicitly by writing $\partial_i \partial_{i+d} k := \partial_i \partial_{i+d} \tilde{k}$. Moreover, we extend this notation to kernels defined on open subsets of $\mathbb{R}^d$ in the obvious way.

Lemma 4.34 (Differentiability of feature maps). *Let $X \subset \mathbb{R}^d$ be an open subset, k be a kernel on X, H be a feature space of k, and $\Phi : X \to H$ be a feature map of k. Let $i \in \{1, \ldots, d\}$ be an index such that the mixed partial derivative $\partial_i \partial_{i+d} k$ of k with respect to the coordinates i and $i + d$ exists and is continuous. Then the partial derivative $\partial_i \Phi$ of $\Phi : X \to H$ with respect to the i-th coordinate exists, is continuous, and for all $x, x' \in X$ we have*

$$\big\langle \partial_i \Phi(x), \partial_i \Phi(x') \big\rangle_H = \partial_i \partial_{i+d} k(x, x') = \partial_{i+d} \partial_i k(x, x')\,. \qquad (4.22)$$

Proof. Without loss of generality, we may assume $X = \mathbb{R}^d$. For $h \in \mathbb{R}$ and $e_i \in \mathbb{R}^d$ being the i-th vector of the canonical basis of $\mathbb{R}^d$, we then define $\Delta_h \Phi(x) := \Phi(x + he_i) - \Phi(x)$, $x \in X$. In order to show that $\partial_i \Phi(x)$ exists for an arbitrary $x \in X$, it obviously suffices to show that $h_n^{-1} \Delta_{h_n} \Phi(x)$ converges for all sequences $(h_n) \subset \mathbb{R}^d \setminus \{0\}$ with $h_n \to 0$. Since a feature space is complete, it thus suffices to show that $(h_n^{-1} \Delta_{h_n} \Phi(x))$ is a Cauchy sequence. To this end, we first observe that

$$\begin{aligned}\big\| h_n^{-1}\Delta_{h_n}\Phi(x) - h_m^{-1}\Delta_{h_m}\Phi(x)\big\|_H^2 &= \big\langle h_n^{-1}\Delta_{h_n}\Phi(x), h_n^{-1}\Delta_{h_n}\Phi(x)\big\rangle_H \\ &\quad+\big\langle h_m^{-1}\Delta_{h_m}\Phi(x), h_m^{-1}\Delta_{h_m}\Phi(x)\big\rangle_H \\ &\quad-2\big\langle h_n^{-1}\Delta_{h_n}\Phi(x), h_m^{-1}\Delta_{h_m}\Phi(x)\big\rangle_H\end{aligned}$$

for all $x \in X$ and $n, m \in \mathbb{N}$. For the function $K(x') := k(x+h_n e_i, x')-k(x,x')$, $x' \in X$, we further have $\langle \Delta_{h_n}\Phi(x), \Delta_{h_m}\Phi(x')\rangle_H = K(x'+h_m e_i) - K(x')$, and hence the mean value theorem yields a $\xi_{m,n} \in [-|h_m|, |h_m|]$ such that

$$\begin{aligned}&\big\langle \Delta_{h_n}\Phi(x), h_m^{-1}\Delta_{h_m}\Phi(x')\big\rangle_H \\ &= \partial_i K(x' + \xi_{m,n}e_i) \\ &= \partial_{i+d}k(x+h_n e_i, x'+\xi_{m,n}e_i) - \partial_{i+d}k(x, x'+\xi_{m,n}e_i).\end{aligned}$$

Another application of the mean value theorem yields an $\eta_{m,n} \in [-|h_n|, |h_n|]$ such that

$$\big\langle h_n^{-1}\Delta_{h_n}\Phi(x), h_m^{-1}\Delta_{h_m}\Phi(x')\big\rangle_H = \partial_i\partial_{i+d}k(x+\eta_{m,n}e_i, x'+\xi_{m,n}e_i)\,.$$

By the continuity of $\partial_i\partial_{i+d}k$, we conclude that for a given $\varepsilon > 0$ there exists an $n_0 \in \mathbb{N}$ such that for all $n, m \geq n_0$ we have

$$\Big|\big\langle h_n^{-1}\Delta_{h_n}\Phi(x), h_m^{-1}\Delta_{h_m}\Phi(x')\big\rangle_H - \partial_i\partial_{i+d}k(x,x')\Big| \leq \varepsilon\,. \tag{4.23}$$

Applying (4.23) for $x = x'$ to the three terms on the right-hand side of our first equation, we see that $(h_n^{-1}\Delta_{h_n}\Phi(x))$ is a Cauchy sequence. By definition, its limit is $\partial_i\Phi$, and the first equality in (4.22) is then a direct consequence of (4.23). The second equality follows from the symmetry of k. □

A direct consequence of the result above is that $\partial_i\partial_{i+d}k$ is a kernel on $X \times X$ with feature map $\partial_i\Phi$. Now assume that even $\partial_j\partial_{j+d}\partial_i\partial_{i+d}k$ exists and is continuous. Then an iterated application of the preceding lemma shows that $\partial_j\partial_i\Phi$ exists, is continuous, and is a feature map of the kernel $\partial_j\partial_{j+d}\partial_i\partial_{i+d}k$. Obviously, we can further iterate this procedure if even higher-order derivatives exist. In order to describe such situations, we write $\partial^{\alpha,\alpha} := \partial_1^{\alpha_1}\dots\partial_d^{\alpha_d}\partial_{1+d}^{\alpha_1}\dots\partial_{2d}^{\alpha_d}$, where $\alpha = (\alpha_1, \dots, \alpha_d) \in \mathbb{N}_0^d$ is a multi-index and arbitrary reorderings of the partial derivatives are allowed.

Definition 4.35. *Let k be a kernel on an open $X \subset \mathbb{R}^d$. For $m \geq 0$, we say that k is m-**times continuously differentiable** if $\partial^{\alpha,\alpha}k : X \times X \to \mathbb{R}$ exists and is continuous for all multi-indexes $\alpha \in \mathbb{N}_0^d$ with $|\alpha| \leq m$.*

Iteratively applying Lemma 4.34 to an m-times continuously differentiable kernel yields the following result.

Corollary 4.36 (RKHSs of differentiable kernels). *Let $X \subset \mathbb{R}^d$ be an open subset, $m \geq 0$, and k be an m-times continuously differentiable kernel on X with RKHS H. Then every $f \in H$ is m-times continuously differentiable, and for $\alpha \in \mathbb{N}_0^d$ with $|\alpha| \leq m$ and $x \in X$ we have*

$$\big|\partial^\alpha f(x)\big| \;\leq\; \|f\|_H \cdot \big(\partial^{\alpha,\alpha}k(x,x)\big)^{1/2}\,. \tag{4.24}$$

Proof. Let us write $\Phi : X \to H$ for the canonical feature map of k. By iteratively applying Lemma 4.34, we see that $\partial^\alpha \Phi : X \to H$ is a feature map of the kernel $\partial^{\alpha,\alpha} k : X \times X \to \mathbb{R}$. By the continuity of $\langle f, \cdot \rangle_H : H \to \mathbb{R}$, we then conclude that $\langle f, \partial^\alpha \Phi(x)\rangle_H = \partial^\alpha \langle f, \Phi(x)\rangle_H = \partial^\alpha f(x)$, i.e, the latter partial derivative exists and is continuous. Finally, (4.24) follows from

$$\big|\partial^\alpha f(x)\big| = \big|\langle f, \partial^\alpha \Phi(x)\rangle_H\big| \leq \|f\|_H \cdot \sqrt{\langle \partial^\alpha \Phi(x), \partial^\alpha \Phi(x)\rangle_H} \tag{4.25}$$

and an iterated application of (4.22) to the right-hand side of (4.25). □

4.4 Gaussian Kernels and Their RKHSs

The goal of this section is to use the developed theory on RKHSs to investigate the Gaussian RBF kernels and their RKHSs in more detail. In particular, we will present two representations of these RKHSs and discuss some consequences. We begin, however, with a simple result that describes the effect of the kernel parameter γ on the input domain.

Proposition 4.37. *Let $X \subset \mathbb{R}^d$ be a non-empty subset and $s, \gamma > 0$ be real numbers. Given a function $f : sX \to \mathbb{R}$, we define $\tau_s f(x) := f(sx)$ for $x \in X$. Then, for all $f \in H_{s\gamma}(sX)$, we have $\tau_s f \in H_\gamma(X)$, and the corresponding linear operator $\tau_s : H_{s\gamma}(sX) \to H_\gamma(X)$ is an isometric isomorphism.*

Proof. We define $\Phi : X \to H_{s\gamma}(sX)$ by $\Phi(x) := \Phi_{s\gamma}(sx)$, where $x \in X$ and $\Phi_{s\gamma} : sX \to H_{s\gamma}(sX)$ is the canonical feature map of $k_{s\gamma}$, i.e., $\Phi_{s\gamma}(y) = k_{s\gamma}(\cdot, y)$ for all $y \in sX$. For $x, x' \in X$, we then have

$$\begin{aligned}\langle \Phi(x'), \Phi(x)\rangle_{H_{s\gamma}(sX)} = \langle \Phi_{s\gamma}(sx'), \Phi_{s\gamma}(sx)\rangle_{H_{s\gamma}(sX)} &= k_{s\gamma}(sx', sx) \\ &= \exp\big(-(s\gamma)^{-2}\|sx - sx'\|_2^2\big) \\ &= k_\gamma(x, x'),\end{aligned}$$

and hence $\Phi : X \to H_{s\gamma}(sX)$ is a feature map of $k_\gamma : X \times X \to \mathbb{R}$. Let us now fix an $f \in H_{s\gamma}(sX)$. By Theorem 4.21, we then know that $\langle f, \Phi(\cdot)\rangle_{H_{s\gamma}(sX)} \in H_\gamma(X)$ and

$$\big\| \langle f, \Phi(\cdot)\rangle_{H_{s\gamma}(sX)} \big\|_{H_\gamma(sX)} \leq \|f\|_{H_{s\gamma}(X)} .$$

Moreover, for $x \in X$, the reproducing property in $H_{s\gamma}(sX)$ yields

$$\langle f, \Phi(x)\rangle_{H_{s\gamma}(sX)} = \langle f, \Phi_{s\gamma}(sx)\rangle_{H_{s\gamma}(sX)} = f(sx) = \tau_s f(x) ,$$

and hence we have found $\tau_s f \in H_\gamma(X)$ with $\|\tau_s f\|_{H_\gamma(X)} \leq \|f\|_{H_{s\gamma}(sX)}$. Finally, we obtain the converse inequality by applying the results above to the dilation operator $\tau_{1/s}$. □

Portions of Section 4.4 are based on material originally published in "I. Steinwart, D. Hush, and C. Scovel (2006), 'An explicit description of the reproducing kernel Hilbert spaces of Gaussian RBF kernels.' *IEEE Trans. Inf. Theory*, **52**, 4635–4643"

Roughly speaking, the preceding proposition states that scaling the kernel parameter has the same effect on the RKHSs as scaling the input space. Considering the definition of the Gaussian RBF kernels, this is not really surprising.

Our next goal is to determine an explicit formula for the RKHSs of Gaussian RBF kernels. To this end, let us fix $\gamma > 0$ and $d \in \mathbb{N}$. For a given holomorphic function $f : \mathbb{C}^d \to \mathbb{C}$, we define

$$\|f\|_{\gamma,\mathbb{C}^d} := \Bigl(\frac{2^d}{\pi^d \gamma^{2d}} \int_{\mathbb{C}^d} |f(z)|^2 e^{\gamma^{-2} \sum_{j=1}^d (z_j - \bar{z}_j)^2} dz\Bigr)^{1/2}, \tag{4.26}$$

where z_j is the j-th component of $z \in \mathbb{C}^d$, $\bar{z}_j$ its conjugate, and dz stands for the complex Lebesgue measure on $\mathbb{C}^d$. Furthermore, we write

$$H_{\gamma,\mathbb{C}^d} := \bigl\{ f : \mathbb{C}^d \to \mathbb{C} \,|\, f \text{ holomorphic and } \|f\|_{\gamma,\mathbb{C}^d} < \infty \bigr\}. \tag{4.27}$$

Obviously, $H_{\gamma,\mathbb{C}^d}$ is a $\mathbb{C}$-vector space with pre-Hilbert norm $\|\cdot\|_{\gamma,\mathbb{C}^d}$. Now, our first result shows that $H_{\gamma,\mathbb{C}^d}$ is the RKHS of the *complex* Gaussian RBF kernel $k_{\gamma,\mathbb{C}^d}$ defined in Proposition 4.10.

Theorem 4.38 (RKHS of the complex Gaussian RBF). *Let $\gamma > 0$ and $d \in \mathbb{N}$. Then $(H_{\gamma,\mathbb{C}^d}, \|\cdot\|_{H_{\gamma,\mathbb{C}^d}})$ is an RKHS and $k_{\gamma,\mathbb{C}^d}$ is its reproducing kernel. Furthermore, for $n \in \mathbb{N}_0$, let $e_n : \mathbb{C} \to \mathbb{C}$ be defined by*

$$e_n(z) := \sqrt{\frac{2^n}{\gamma^{2n} n!}}\, z^n e^{-\gamma^{-2} z^2}, \qquad z \in \mathbb{C}. \tag{4.28}$$

Then the system $(e_{n_1} \otimes \cdots \otimes e_{n_d})_{n_1,\dots,n_d \geq 0}$ of functions $e_{n_1} \otimes \cdots \otimes e_{n_d} : \mathbb{C}^d \to \mathbb{C}$ defined by

$$e_{n_1} \otimes \cdots \otimes e_{n_d}(z_1, \dots, z_d) := \prod_{j=1}^d e_{n_j}(z_j), \qquad (z_1, \dots, z_d) \in \mathbb{C}^d,$$

is an orthonormal basis of $H_{\gamma,\mathbb{C}^d}$.

For the proof of Theorem 4.38, we need the following technical lemma.

Lemma 4.39. *For all $d \in \mathbb{N}$, all holomorphic functions $f : \mathbb{C}^d \to \mathbb{C}$, all $r_1, \dots, r_d \in [0, 1)$, and all $z \in \mathbb{C}^d$, we have*

$$|f(z)|^2 \leq \frac{1}{(2\pi)^d} \int_0^{2\pi} \cdots \int_0^{2\pi} \bigl|f(z_1 + r_1 e^{i\theta_1}, \dots, z_d + r_d e^{i\theta_d})\bigr|^2 d\theta_1 \cdots d\theta_d, \tag{4.29}$$

where $i := \sqrt{-1}$ denotes the imaginary unit.

Proof. We proceed by induction over d. For $d = 1$, Hardy's convexity theorem (see Theorem A.7.3) states that the function

$$r \mapsto \frac{1}{2\pi} \int_0^{2\pi} \left|f(z + re^{i\theta})\right|^2 d\theta$$

is non-decreasing on $[0, 1)$, and hence we obtain the assertion in this case.

Now let us suppose that we have already shown the assertion for $d \in \mathbb{N}$. Let $f : \mathbb{C}^{d+1} \to \mathbb{C}$ be a holomorphic function, and choose $r_1, \dots, r_{d+1} \in [0, 1)$. Since for fixed $(z_1, \dots, z_d) \in \mathbb{C}^d$ the function $z_{d+1} \mapsto f(z_1, \dots, z_d, z_{d+1})$ is holomorphic, we already know that

$$|f(z_1, \dots, z_{d+1})|^2 \leq \frac{1}{2\pi} \int_0^{2\pi} \left|f(z_1, \dots, z_d, z_{d+1} + r_{d+1}e^{i\theta_{d+1}})\right|^2 d\theta_{d+1} \, .$$

Now applying the induction hypothesis to the holomorphic functions

$$(z_1, \dots, z_d) \mapsto f(z_1, \dots, z_d, z_{d+1} + r_{d+1}e^{i\theta_{d+1}})$$

on $\mathbb{C}^d$ gives the assertion for $d + 1$. □

Proof (of Theorem 4.38). We first prove that $H_{\gamma,\mathbb{C}}$ is an RKHS. To this end, we begin by showing that for all compact subsets $K \subset \mathbb{C}^d$ there exists a constant $c_K > 0$ with

$$|f(z)| \leq c_K \, \|f\|_{\gamma,\mathbb{C}^d} \, , \qquad z \in K, \; f \in H_{\gamma,\mathbb{C}^d}. \tag{4.30}$$

In order to establish (4.30), we define

$$c := \max\{e^{-\gamma^{-2}\sum_{j=1}^d (z_j - \bar{z}_j)^2} : (z_1, \dots, z_d) \in K + (B_\mathbb{C})^d\} \, ,$$

where $B_\mathbb{C}$ denotes the closed unit ball of $\mathbb{C}$. Now, by Lemma 4.39, we have

$$2^d r_1 \cdots r_d |f(z)|^2 \leq \frac{r_1 \cdots r_d}{\pi^d} \int_0^{2\pi} \!\! \cdots \! \int_0^{2\pi} \left|f(z_1 + r_1e^{i\theta_1}, \dots, z_d + r_de^{i\theta_d})\right|^2 d\theta_1 \cdots d\theta_d,$$

and integrating this inequality with respect to $r = (r_1, \dots, r_d)$ over $[0, 1)^d$ then yields

$$\begin{aligned} |f(z)|^2 \leq \frac{1}{\pi^d} \int_{z + (B_\mathbb{C})^d} |f(z')|^2 dz' &\leq \frac{c}{\pi^d} \int_{z + (B_\mathbb{C})^d} |f(z')|^2 e^{\gamma^{-2}\sum_{j=1}^d (z'_j - \bar{z}'_j)^2} dz' \\ &\leq \frac{c\gamma^{2d}}{2^d} \, \|f\|^2_{\gamma,\mathbb{C}^d} \, , \qquad z \in K, \end{aligned}$$

by the continuity of f. This means that we have established (4.30). Now, (4.30) obviously shows that the Dirac functionals are bounded on $H_{\gamma,\mathbb{C}^d}$. Furthermore, (4.30) also shows that convergence in $\| \cdot \|_{\gamma,\mathbb{C}}$ implies *compact convergence*, i.e., uniform convergence on every compact subset. Using the fact that a compactly convergent sequence of holomorphic functions has a holomorphic limit (see, e.g., Theorem A.7.2), we then immediately find that $H_{\gamma,\mathbb{C}^d}$ is complete. Therefore $H_{\gamma,\mathbb{C}^d}$ is an RKHS.

To show that the system $(e_{n_1} \otimes \cdots \otimes e_{n_d})_{n_1,\dots,n_d \geq 0}$ is an ONB of $H_{\gamma,\mathbb{C}^d}$, we first consider the case $d = 1$. To this end, we observe that for $n \in \mathbb{N}_0$ we have

$$\begin{aligned}\int_{\mathbb{C}} z^n (\bar{z})^n e^{-2\gamma^{-2} z\bar{z}} dz &= \int_0^\infty r \int_0^{2\pi} r^{2n} e^{-2\gamma^{-2} r^2} d\theta dr \\ &= 2\pi \int_0^\infty r^{2n+1} e^{-2\gamma^{-2} r^2} dr \\ &= \frac{\pi \gamma^{2(n+1)}}{2^{n+1}} \int_0^\infty t^n e^{-t} dt \\ &= \frac{\pi \gamma^{2(n+1)} n!}{2^{n+1}}, \end{aligned} \tag{4.31}$$

where in the last step we used the gamma function, see Section A.1. Furthermore, for $n, m \in \mathbb{N}_0$ with $n \neq m$, a simple calculation gives

$$\int_{\mathbb{C}} z^n (\bar{z})^m e^{-2\gamma^{-2} z\bar{z}} dz = \int_0^\infty r \int_0^{2\pi} r^{n+m} e^{i(n-m)\theta} e^{-2\gamma^{-2} r^2} d\theta dr = 0\,. \tag{4.32}$$

In addition, for $z, \bar{z} \in \mathbb{C}$ and $n, m \geq 0$, we have

$$\begin{aligned} e_n(z)\overline{e_m(z)} e^{\gamma^{-2}(z-\bar{z})^2} &= \sqrt{\frac{2^{n+m}}{n!\,m!\,\gamma^{2(n+m)}}}\, z^n (\bar{z})^m e^{-\gamma^{-2} z^2 - \gamma^{-2} \bar{z}^2} e^{\gamma^{-2}(z-\bar{z})^2} \\ &= \sqrt{\frac{2^{n+m}}{n!\,m!\,\gamma^{2(n+m)}}}\, z^n (\bar{z})^m e^{-2\gamma^{-2} z\bar{z}}\,, \end{aligned}$$

and consequently we obtain

$$\langle e_n, e_m \rangle = \frac{2}{\pi\gamma^2} \int_{\mathbb{C}} e_n(z)\overline{e_m(z)} e^{\gamma^{-2}(z-\bar{z})^2} dz = \begin{cases} 1 & \text{if } n = m \\ 0 & \text{otherwise} \end{cases}$$

by (4.31) and (4.32), i.e., $(e_n)_{n\geq 0}$ is an ONS. Now, let us show that this system is actually an ONB. To this end, let $f \in H_{\gamma,\mathbb{C}}$. Then $z \mapsto e^{\gamma^{-2} z^2} f(z)$ is an entire function, and therefore there exists a sequence $(a_n) \subset \mathbb{C}$ such that

$$f(z) = \sum_{n=0}^\infty a_n z^n e^{-\gamma^{-2} z^2} = \sum_{n=0}^\infty a_n \sqrt{\frac{\gamma^{2n} n!}{2^n}}\, e_n(z) \tag{4.33}$$

for all $z \in \mathbb{C}$. Obviously, it suffices to show that the convergence above also holds with respect to $\|\cdot\|_{\gamma,\mathbb{C}}$. To prove this, we first recall from complex analysis that the series in (4.33) converges absolutely and compactly. Therefore, for $n \geq 0$ equations (4.31), (4.32), and (4.33) yield

$$\begin{aligned}\langle f, e_n\rangle &= \frac{2}{\pi\gamma^2}\int_{\mathbb{C}} f(z)\overline{e_n(z)}e^{\gamma^{-2}(z-\bar{z})^2}dz \\ &= \frac{2}{\pi\gamma^2}\sum_{m=0}^{\infty} a_m \int_{\mathbb{C}} z^m e^{-\gamma^{-2}z^2}\overline{e_n(z)}e^{\gamma^{-2}(z-\bar{z})^2}dz \\ &= \frac{2}{\pi\gamma^2}\sqrt{\frac{2^n}{\gamma^{2n}n!}}\sum_{m=0}^{\infty} a_m \int_{\mathbb{C}} z^m(\bar{z})^n e^{-2\gamma^{-2}z\bar{z}}dz \\ &= a_n\sqrt{\frac{\gamma^{2n}n!}{2^n}}\,. \end{aligned} \tag{4.34}$$

Furthermore, since (e_n) is an ONS, there exists a function $g \in H_{\gamma,\mathbb{C}}$ with $g = \sum_{n=0}^{\infty}\langle f, e_n\rangle e_n$, where the convergence takes place in $H_{\sigma,\mathbb{C}}$. Now, using (4.33), (4.34), and the fact that norm convergence in RKHSs implies pointwise convergence, we find $g = f$, i.e., the series in (4.33) converges with respect to $\|\cdot\|_{\sigma,\mathbb{C}}$.

Now, let us briefly treat the general, d-dimensional case. In this case, a simple calculation shows

$$\langle e_{n_1}\otimes\cdots\otimes e_{n_d}, e_{m_1}\otimes\cdots\otimes e_{m_d}\rangle_{H_{\gamma,\mathbb{C}^d}} = \prod_{j=1}^{d}\langle e_{n_j}, e_{m_j}\rangle_{H_{\gamma,\mathbb{C}}}\,,$$

and hence we find the orthonormality of $(e_{n_1}\otimes\cdots\otimes e_{n_d})_{n_1,\dots,n_d\geq 0}$. In order to check that this orthonormal system is an ONB, let us fix an $f \in H_{\sigma,\mathbb{C}^d}$. Then $z \mapsto f(z)\exp(\sigma^2\sum_{i=1}^{d} z_i^2)$ is an entire function, and hence there exist $a_{n_1,\dots,n_d} \in \mathbb{C}$, $(n_1,\dots,n_d) \in \mathbb{N}_0^d$, such that

$$\begin{aligned} f(z) &= \sum_{(n_1,\dots,n_d)\in\mathbb{N}_0^d} a_{n_1,\dots,n_d}\prod_{i=1}^{d} z_i^{n_i}e^{-\sigma^2 z_i^2} \\ &= \sum_{(n_1,\dots,n_d)\in\mathbb{N}_0^d} a_{n_1,\dots,n_d}\prod_{i=1}^{d}\sqrt{\frac{n_i!}{(2\sigma^2)^{n_i}}}e_{n_i}(z) \end{aligned}$$

for all $z = (z_1,\dots,z_d) \in \mathbb{C}^d$. From this we easily derive

$$\langle f, e_{n_1}\otimes\cdots\otimes e_{n_d}\rangle = a_{n_1,\dots,n_d}\prod_{i=1}^{d}\sqrt{\frac{n_i!}{(2\sigma^2)^{n_i}}}\,.$$

Now we see that $(e_{n_1}\otimes\cdots\otimes e_{n_d})_{n_1,\dots,n_d\geq 0}$ is an ONB as in the one-dimensional case.

Finally, let us show that $k_{\gamma,\mathbb{C}^d}$ is the reproducing kernel of $H_{\gamma,\mathbb{C}^d}$. To this end, we write k for the reproducing kernel of $H_{\gamma,\mathbb{C}^d}$. Then (4.9) and the Taylor series expansion of the exponential function yield

$$\begin{aligned}
k(z,z') &= \sum_{n_1,\dots,n_d=0}^{\infty} e_{n_1}\otimes\cdots\otimes e_{n_d}(z)\overline{e_{n_1}\otimes\cdots\otimes e_{n_d}(z')} \\
&= \sum_{n_1,\dots,n_d=0}^{\infty} \prod_{j=1}^{d} \frac{2^{n_j}}{\gamma^{2n_j} n_j!}(z_j\bar z_j')^{n_j} e^{-\gamma^{-2}z_j^2-\gamma^{-2}(\bar z_j')^2} \\
&= \prod_{j=1}^{d}\sum_{n_j=0}^{\infty} \frac{2^{n_j}}{\gamma^{2n_j} n_j!}(z_j\bar z_j')^{n_j} e^{-\gamma^{-2}z_j^2-\gamma^{-2}(\bar z_j')^2} \\
&= \prod_{j=1}^{d} e^{-\gamma^{-2}z_j^2-\gamma^{-2}(\bar z_j')^2+2\gamma^{-2}z_j\bar z_j'} \\
&= e^{-\gamma^{-2}\sum_{j=1}^{d}(z_j-\bar z_j')^2}\,. \qquad \square
\end{aligned}$$

With the help of Theorem 4.38, we can obtain some interesting information on the RKHSs of the *real-valued* Gaussian RBF kernels k_γ. Let us begin with the following corollary that describes their RKHSs.

Corollary 4.40 (RKHS of Gaussian RBF). *For $X \subset \mathbb{R}^d$ and $\gamma > 0$, the RKHS $H_\gamma(X)$ of the real-valued Gaussian RBF kernel k_γ on X is*

$$H_\gamma(X) = \bigl\{ f : X \to \mathbb{R} \,|\, \exists\, g \in H_{\gamma,\mathbb{C}^d} \text{ with } \operatorname{Re} g_{|X} = f \bigr\}\,,$$

and for $f \in H_\gamma(X)$ the norm $\|\cdot\|_{H_\gamma(X)}$ in $H_\gamma(X)$ can be computed by

$$\|f\|_{H_\gamma(X)} = \inf\bigl\{ \|g\|_{\gamma,\mathbb{C}^d} : g \in H_{\gamma,\mathbb{C}^d} \text{ with } \operatorname{Re} g_{|X} = f \bigr\}\,.$$

Proof. The assertion directly follows from Theorem 4.38, Proposition 4.10, and the discussion following Corollary 4.22. □

The preceding corollary shows that every $f \in H_\gamma(X)$ of the Gaussian RBF kernel k_γ originates from the complex RKHS $H_{\gamma,\mathbb{C}^d}$, which consists of entire functions. Consequently, every $f \in H_\gamma(X)$ can be represented by a power series that converges on $\mathbb{R}^d$. This observation suggests that there may be an intimate relationship between $H_\gamma(X)$ and $H_\gamma(\mathbb{R}^d)$ if X contains an open set. In order to investigate this conjecture, we need some additional notation. For a multi-index $\nu := (n_1,\dots,n_d) \in \mathbb{N}_0^d$, we write $|v| := n_1+\cdots+n_d$. Furthermore, for $X \subset \mathbb{R}$ and $n \in \mathbb{N}_0$, we define $e_n^X : X \to \mathbb{R}$ by

$$e_n^X(x) := \sqrt{\frac{2^n}{\gamma^{2n} n!}}\, x^n e^{-\gamma^{-2}x^2}\,, \qquad x \in X\,, \tag{4.35}$$

i.e., we have $e_n^X = (e_n)_{|X} = (\operatorname{Re} e_n)_{|X}$, where $e_n : \mathbb{C} \to \mathbb{C}$ is an element of the ONB of $H_{\gamma,\mathbb{C}}$ defined by (4.28). Furthermore, for a multi-index $\nu := (n_1,\dots,n_d) \in \mathbb{N}_0^d$, we write

$$e_\nu^X := e_{n_1}^X \otimes \cdots \otimes e_{n_d}^X$$

and $e_\nu := e_{n_1} \otimes \cdots \otimes e_{n_d}$. Given an $x := (x_1, \ldots, x_d) \in \mathbb{R}^d$, we also adopt the notation $x^\nu := x_1^{n_1} \cdot \ldots \cdot x_d^{n_d}$. Finally, recall that $\ell_2(\mathbb{N}_0^d)$ denotes the set of all *real*-valued square-summable families, i.e.,

$$\ell_2(\mathbb{N}_0^d) := \left\{ (a_\nu)_{\nu \in \mathbb{N}_0^d} : a_\nu \in \mathbb{R} \text{ for all } \nu \in \mathbb{N}_0^d \text{ and } \|(a_\nu)\|_2^2 := \sum_{\nu \in \mathbb{N}_0^d} a_\nu^2 < \infty \right\}.$$

With the help of these notations, we can now show an intermediate result.

Proposition 4.41. *Let $\gamma > 0$, $X \subset \mathbb{R}^d$ be a subset with non-empty interior, i.e., $\mathring{X} \neq \emptyset$, and $f \in H_\gamma(X)$. Then there exists a unique element $(b_\nu) \in \ell_2(\mathbb{N}_0^d)$ such that*

$$f(x) \;=\; \sum_{\nu \in \mathbb{N}_0^d} b_\nu e_\nu^X(x)\,, \qquad\qquad x \in X, \tag{4.36}$$

where the convergence is absolute. Furthermore, for all functions $g : \mathbb{C}^d \to \mathbb{C}$, the following two statements are equivalent:

i) We have $g \in H_{\gamma,\mathbb{C}^d}$ and $\operatorname{Re} g_{|X} = f$.
ii) There exists an element $(c_\nu) \in \ell_2(\mathbb{N}_0^d)$ with

$$g \;=\; \sum_{\nu \in \mathbb{N}_0^d} (b_\nu + i c_\nu) e_\nu\,. \tag{4.37}$$

Finally, we have the identity $\|f\|_{H_\gamma(X)}^2 = \sum_{\nu \in \mathbb{N}_0^d} b_\nu^2$.

Proof. *i)* $\Rightarrow$ *ii)*. Let us fix a $g \in H_{\gamma,\mathbb{C}^d}$ with $\operatorname{Re} g_{|X} = f$. Since (e_ν) is an ONB of $H_{\gamma,\mathbb{C}^d}$, we then have

$$g = \sum_{\nu \in \mathbb{N}_0^d} \langle g, e_\nu \rangle \, e_\nu\,,$$

where the convergence is with respect to $H_{\gamma,\mathbb{C}^d}$. In addition, recall that the family of Fourier coefficients is square-summable and satisfies Parseval's identity, see Lemma A.5.11,

$$\|g\|_{H_{\gamma,\mathbb{C}^d}}^2 \;=\; \sum_{\nu \in \mathbb{N}_0^d} \left| \langle g, e_\nu \rangle \right|^2.$$

Since convergence in $H_{\gamma,\mathbb{C}^d}$ implies pointwise convergence, we then obtain

$$f(x) = \operatorname{Re} g_{|X}(x) = \operatorname{Re}\left(\sum_{\nu \in \mathbb{N}_0^d} \langle g, e_\nu \rangle \, e_\nu(x) \right) = \sum_{\nu \in \mathbb{N}_0^d} \operatorname{Re}\big(\langle g, e_\nu \rangle\big) e_\nu^X(x)$$

for all $x \in X$, where in the last step we used $e_\nu(x) \in \mathbb{R}$ for $x \in X$. In order to prove *ii)*, it consequently remains to show that $b_\nu := \operatorname{Re} \langle g, e_\nu \rangle$ only depends on f but not on g. To this end, let $\tilde{g} \in H_{\gamma,\mathbb{C}^d}$ be another function with $\operatorname{Re} \tilde{g}_{|X} = f$. By repeating the argument above for $\tilde{g}$, we then find

$$f(x) = \sum_{\nu\in\mathbb{N}_0^d} \operatorname{Re}\big(\langle\tilde g, e_\nu\rangle\big) e_\nu^X(x)\,, \qquad\qquad x\in X.$$

Using the definition (4.35) of e_n^X, we then obtain

$$\sum_{\nu\in\mathbb{N}_0^d} \operatorname{Re}\big(\langle\tilde g, e_\nu\rangle\big)\, a_\nu\, x^\nu = \sum_{\nu\in\mathbb{N}_0^d} \operatorname{Re}\big(\langle g, e_\nu\rangle\big)\, a_\nu\, x^\nu\,, \qquad x\in X,$$

where $a_\nu := a_{n_1}\cdots a_{n_d}$ and $a_n := \big(\frac{2^n}{\gamma^{2n}n!}\big)^{1/2}$. Since X has a non-empty interior, the identity theorem for power series and $a_\nu \neq 0$ then give $\operatorname{Re}\langle\tilde g, e_\nu\rangle = \operatorname{Re}\langle g, e_\nu\rangle$ for all $\nu\in\mathbb{N}_0^d$. This shows both (4.36) and (4.37). Finally, Corollary 4.40 and Parseval's identity give

$$\begin{aligned}\|f\|^2_{H_\gamma(X)} &= \inf\big\{\|g\|_{\gamma,\mathbb{C}^d} : g\in H_{\gamma,\mathbb{C}^d} \text{ with } \operatorname{Re} g_{|X} = f\big\}\\ &= \inf\Bigg\{\sum_{\nu\in\mathbb{N}_0^d} b_\nu^2 + c_\nu^2 : (c_\nu)\in\ell_2(\mathbb{N}_0^d)\Bigg\}\\ &= \sum_{\nu\in\mathbb{N}_0^d} b_\nu^2\,.\end{aligned}$$

ii) ⇒ *i)*. Since $(b_\nu)\in\ell_2(\mathbb{N}_0^d)$ and $(c_\nu)\in\ell_2(\mathbb{N}_0^d)$ imply $\big(|b_\nu+ic_\nu|\big)\in\ell_2(\mathbb{N}_0^d)$, we have $g\in H_{\gamma,\mathbb{C}^d}$. Furthermore, $\operatorname{Re} g_{|X} = f$ follows from

$$\operatorname{Re} g(x) = \operatorname{Re}\sum_{\nu\in\mathbb{N}_0^d}(b_\nu+ic_\nu)e_\nu(x) = \sum_{\nu\in\mathbb{N}_0^d} b_\nu e_\nu^X(x) = f(x)\,, \qquad\qquad x\in X.$$

□

With the help of the preceding proposition, we can now establish our main result on $H_\gamma(X)$ for input spaces X having a non-empty interior. Roughly speaking, this result states that $H_\gamma(X)$ is isometrically embedded into $H_{\gamma,\mathbb{C}^d}$ via a canonical extension procedure based on a specific ONB of $H_\gamma(X)$.

Theorem 4.42 (ONB of real Gaussian RKHS). *Let $\gamma>0$ and $X\subset\mathbb{R}^d$ be a subset with a non-empty interior. Furthermore, for an $f\in H_\gamma(X)$ represented by (4.36), we define*

$$\hat f := \sum_{\nu\in\mathbb{N}_0^d} b_\nu e_\nu\,.$$

Then the extension operator $\hat{\ }: H_\gamma(X)\to H_{\gamma,\mathbb{C}^d}$ defined by $f\mapsto\hat f$ satisfies

$$\begin{aligned}\operatorname{Re}\hat f_{|X} &= f\,,\\ \|\hat f\|_{H_{\gamma,\mathbb{C}^d}} &= \|f\|_{H_\gamma(X)}\end{aligned}$$

for all $f\in H_\gamma(X)$. Moreover, (e_ν^X) is an ONB of $H_\gamma(X)$, and for $f\in H_\gamma(X)$ having the representation (4.36), we have $b_\nu = \langle f, e_\nu^X\rangle$ for all $\nu\in\mathbb{N}_0^d$.

Proof. By (4.36), the extension operator is well-defined. The identities then follow from Proposition 4.41 and Parseval's identity. Moreover, the extension operator is obviously $\mathbb{R}$-linear and satisfies $\hat{e}_\nu^X = e_\nu$ for all $\nu \in \mathbb{N}_0^d$. Consequently, we obtain

$$\|e_{\nu_1}^X \pm e_{\nu_2}^X\|_{H_\gamma(X)} = \|\hat{e}_{\nu_1}^X \pm \hat{e}_{\nu_2}^X\|_{H_{\gamma,\mathbb{C}^d}} = \|e_{\nu_1} \pm e_{\nu_2}\|_{H_{\gamma,\mathbb{C}^d}}$$

for $\nu_1, \nu_2 \in \mathbb{N}_0^d$. Using the first polarization identity of Lemma A.5.9, we then see that (e_ν^X) is an ONS in $H_\gamma(X)$. To see that it actually is an ONB we fix an $f \in H_\gamma(X)$. Furthermore, let $(b_\nu) \in \ell_2(\mathbb{N}_0^d)$ be the family that satisfies (4.36). Then

$$\tilde{f} := \sum_{\nu \in \mathbb{N}_0^d} b_\nu e_\nu^X$$

converges in $H_\gamma(X)$. Since convergence in $H_\gamma(X)$ implies pointwise convergence, (4.36) then yields $\tilde{f}(x) = f(x)$ for all $x \in X$. Consequently, (e_ν^X) is an ONB of $H_\gamma(X)$. Finally, the identity $b_\nu = \langle f, e_\nu^X\rangle$, $\nu \in \mathbb{N}_0^d$, follows from the fact that the representation of f by (e_ν^X) is unique. □

In the following, we present some interesting consequences of the preceding theorem.

Corollary 4.43. *Let $X \subset \mathbb{R}^d$ be a subset with non-empty interior, $\gamma > 0$, and $\hat{}\, : H_\gamma(X) \to H_{\gamma,\mathbb{C}^d}$ be the extension operator defined in Theorem 4.42. Then the extension operator $I : H_\gamma(X) \to H_\gamma(\mathbb{R}^d)$ defined by $If := \operatorname{Re} \hat{f}_{|\mathbb{R}^d}$, $f \in H_\gamma(X)$, is an isometric isomorphism.*

Proof. For $f \in H_\gamma(X)$, we have $(\langle f, e_\nu^X\rangle) \in \ell_2(\mathbb{N}_0^d)$, and hence

$$\tilde{f} := \sum_{\nu \in \mathbb{N}_0^d} \langle f, e_\nu^X\rangle e_\nu^{\mathbb{R}^d}$$

is an element of $H_\gamma(\mathbb{R}^d)$. Moreover, for $\nu \in \mathbb{N}_0^d$, we have $(\operatorname{Re} e_\nu)_{|\mathbb{R}^d} = e_\nu^{\mathbb{R}^d}$ and $\langle f, e_\nu^X\rangle \in \mathbb{R}$, and hence we find $If = \tilde{f}$. Furthermore, $\|f\|_{H_\gamma(X)} = \|If\|_{H_\gamma(\mathbb{R}^d)}$ immediately follows from Parseval's identity. Consequently, I is isometric, linear, and injective. The surjectivity finally follows from the fact that, given an $\tilde{f} \in H_\gamma(\mathbb{R}^d)$, the function

$$f := \sum_{\nu \in \mathbb{N}_0^d} \langle f, e_\nu^{\mathbb{R}^d}\rangle e_\nu^X$$

obviously satisfies $f \in H_\gamma(X)$ and $If = \tilde{f}$. □

Roughly speaking, the preceding corollary means that $H_\gamma(\mathbb{R}^d)$ does not contain "more" functions than $H_\gamma(X)$ if X has a non-empty interior. Moreover, Corollary 4.43 in particular shows that $H_\gamma(X_1)$ and $H_\gamma(X_2)$ are isometrically isomorphic via a simple extension-restriction mapping going through

$H_\gamma(\mathbb{R}^d)$ whenever both input spaces $X_1, X_2 \subset \mathbb{R}^d$ have a non-empty interior. Consequently, we sometimes use the notation $H_\gamma := H_\gamma(X)$ and $\|\cdot\|_\gamma := \|\cdot\|_{H_\gamma(X)}$ if X has a non-empty interior and no confusion can arise.

Besides the isometry above, Theorem 4.42 also yields the following interesting observation.

Corollary 4.44 (Gaussian RKHSs do not contain constants). *Let $\gamma > 0$, $X \subset \mathbb{R}^d$ be a subset with a non-empty interior, and $f \in H_\gamma(X)$. If f is constant on a non-empty open subset A of X, then $f = 0$.*

Proof. Let $c \in \mathbb{R}$ be a constant with $f(x) = c$ for all $x \in A$. Let us define $a_n := (\frac{2^n}{\gamma^{2n} n!})^{1/2}$ for all $n \in \mathbb{N}_0$. Furthermore, for a multi-index $\nu := (n_1, \ldots, n_d) \in \mathbb{N}_0^d$, we write $b_\nu := \langle f, e_\nu^X \rangle$ and $a_\nu := a_{n_1} \cdot \ldots \cdot a_{n_d}$. For $x := (x_1, \ldots, x_d) \in A$, the definition (4.35) and the representation (4.36) then yield

$$c \exp\Biggl(\gamma^{-2} \sum_{j=1}^d x_j^2\Biggr) = f(x) \exp\Biggl(\gamma^{-2} \sum_{j=1}^d x_j^2\Biggr) = \sum_{\nu \in \mathbb{N}_0^d} b_\nu a_\nu x^\nu . \qquad (4.38)$$

Moreover, for $x \in \mathbb{R}^d$, a simple calculation shows

$$\begin{aligned} \exp\Biggl(\gamma^{-2} \sum_{j=1}^d x_j^2\Biggr) = \prod_{j=1}^d e^{\gamma^{-2} x_j^2} &= \prod_{j=1}^d \Biggl(\sum_{n_j=0}^\infty \frac{x_j^{2n_j}}{n_j!\, \gamma^{2n_j}} \Biggr) \\ &= \sum_{n_1, \ldots, n_d = 0}^\infty \prod_{j=1}^d \frac{x_j^{2n_j}}{n_j!\, \gamma^{2n_j}} . \end{aligned}$$

Using (4.38) and the identity theorem for power series, we hence obtain

$$b_\nu a_\nu = \begin{cases} c\,\gamma^{-|\nu|} \prod_{j=1}^d \frac{1}{n_j!} & \text{if } \nu = (2n_1, \ldots, 2n_d) \text{ for some } (n_1, \ldots, n_d) \in \mathbb{N}_0^d \\ 0 & \text{otherwise}, \end{cases}$$

or in other words

$$b_\nu = \begin{cases} c \prod_{j=1}^d \frac{\sqrt{(2n_j)!}}{n_j!}\, 2^{-n_j} & \text{if } \nu = (2n_1, \ldots, 2n_d) \text{ for some } (n_1, \ldots, n_d) \in \mathbb{N}_0^d \\ 0 & \text{otherwise}. \end{cases}$$

Consequently, Parseval's identity yields

$$\begin{aligned} \|f\|_{H_\gamma(X)}^2 &= \sum_{\nu \in \mathbb{N}_0^d} b_\nu^2 = \sum_{n_1, \ldots, n_d = 0}^\infty c^2 \prod_{j=1}^d \frac{(2n_j)!}{(n_j!)^2}\, 2^{-2n_j} \\ &= \prod_{j=1}^d \Biggl(\sum_{n_j=0}^\infty c^{2/d} \frac{(2n_j)!}{(n_j!)^2}\, 2^{-2n_j} \Biggr) \\ &= \Biggl(\sum_{n=0}^\infty c^{2/d} \frac{(2n)!}{(n!)^2}\, 2^{-2n} \Biggr)^d . \end{aligned}$$

Let us write $\alpha_n := \frac{(2n)!}{(n!)^2} 2^{-2n}$ for $n \in \mathbb{N}_0$. By an easy calculation, we then obtain

$$\frac{\alpha_{n+1}}{\alpha_n} = \frac{(2(n+1))!\;(n!)^2\;2^n}{(2n)!\;((n+1)!\,)^2\;2^{2(n+1)}} = \frac{(2n+1)(2n+2)}{4(n+1)^2} = \frac{2n+1}{2n+2} \geq \frac{n}{n+1}$$

for all $n \geq 1$. In other words, $(n\alpha_n)$ is an increasing, positive sequence. Consequently there exists an $\alpha > 0$ with $\alpha_n \geq \frac{\alpha}{n}$ for all $n \geq 1$, and hence we find $\sum_{n=0}^{\infty} \alpha_n = \infty$. Therefore, $\|f\|^2_{H_\gamma(X)} < \infty$ implies $c = 0$, and thus we have $f = 0$. □

The preceding corollary shows in particular that $\mathbf{1}_A \notin H_\gamma(X)$ for all open subsets $A \subset X$. Some interesting consequences of this observation with respect to the hinge loss used in classification are discussed in Exercise 4.8.

Let us now compare the norms $\|\cdot\|_\gamma$ for different values of γ. To this end, we first observe that the weight function in the definition of $\|\cdot\|_{\gamma,\mathbb{C}^d}$ satisfies

$$e^{\gamma^{-2}\sum_{j=1}^d (z_j - \bar{z}_j)^2} = e^{-4\gamma^{-2}\sum_{j=1}^d y_j^2},$$

where $y_j := \operatorname{Im} z_j$, $j = 1, \ldots, d$. For $\gamma_1 \leq \gamma_2$, we hence find $H_{\gamma_2,\mathbb{C}^d} \subset H_{\gamma_1,\mathbb{C}^d}$ and

$$\|f\|_{H_{\gamma_1,\mathbb{C}^d}} \leq \left(\frac{\gamma_2}{\gamma_1}\right)^d \|f\|_{H_{\gamma_2,\mathbb{C}^d}}, \qquad f \in H_{\gamma_2,\mathbb{C}^d}.$$

This suggests that a similar relation holds for the RKHSs of the real Gaussian kernels. In order to investigate this conjecture, let us now present another feature space and feature map for k_γ. To this end, recall that $L_2(\mathbb{R}^d)$ denotes the space of Lebesgue square-integrable functions $\mathbb{R}^d \to \mathbb{R}$ equipped with the usual norm $\|\cdot\|_2$. Our first result shows that $L_2(\mathbb{R}^d)$ is a feature space of k_γ.

Lemma 4.45. *For $0 < \gamma < \infty$ and $X \subset \mathbb{R}^d$, we define $\Phi_\gamma : X \to L_2(\mathbb{R}^d)$ by*

$$\Phi_\gamma(x) := \frac{2^{\frac{d}{2}}}{\pi^{\frac{d}{4}}\gamma^{\frac{d}{2}}} e^{-2\gamma^{-2}\|x - \cdot\,\|_2^2}, \qquad x \in X.$$

Then $\Phi_\gamma : X \to L_2(\mathbb{R}^d)$ is a feature map of k_γ.

Proof. Let us first recall that, using the density of the normal distribution, we have

$$\int_{\mathbb{R}^d} e^{-t^{-1}\|z-x\|_2^2} dz = (\pi t)^{\frac{d}{2}} \tag{4.39}$$

for all $t > 0$ and $x \in \mathbb{R}^d$. Moreover, for $\alpha \geq 0$, an elementary calculation shows that

$$\|y - x\|_2^2 + \alpha\|y - x'\|_2^2 = \frac{\alpha}{1+\alpha}\|x - x'\|_2^2 + (1+\alpha)\left\| y - \frac{x + \alpha x'}{1+\alpha} \right\|_2^2 \tag{4.40}$$

for all $y, x, x' \in \mathbb{R}^d$. By using (4.39) and setting $\alpha := 1$ in (4.40), we now obtain

$$\begin{aligned}\langle \Phi_\gamma(x), \Phi_\gamma(x')\rangle_{L_2(\mathbb{R}^d)} &= \frac{2^d}{\pi^{\frac{d}{2}}\gamma^d}\int_{\mathbb{R}^d} e^{-2\gamma^{-2}\|x-z\|_2^2} e^{-2\gamma^{-2}\|x'-z\|_2^2} dz \\ &= \frac{2^d}{\pi^{\frac{d}{2}}\gamma^d} e^{-\gamma^{-2}\|x-x'\|_2^2} \int_{\mathbb{R}^d} e^{-4\gamma^{-2}\|z-\frac{x+x'}{2}\|_2^2} dz \\ &= \frac{2^d}{\pi^{\frac{d}{2}}\gamma^d} \cdot e^{-\gamma^{-2}\|x-x'\|_2^2} \Bigl(\frac{\pi\gamma^2}{4}\Bigr)^{\frac{d}{2}} \\ &= k_\gamma(x,x')\,,\end{aligned}$$

and hence Φ_γ is a feature map and $L_2(\mathbb{R}^d)$ is a feature space of k_γ. □

Having the feature map $\Phi_\gamma : X \to L_2(\mathbb{R}^d)$ of k_γ, we can now give another description of the RKHS of k_γ. To this end, we need the integral operators $W_t : L_2(\mathbb{R}^d) \to L_2(\mathbb{R}^d)$, $t > 0$, defined by

$$W_t g(x) := (\pi t)^{-\frac{d}{2}} \int_{\mathbb{R}^d} e^{-t^{-1}\|y-x\|_2^2} g(y)dy\,, \qquad g \in L_2(\mathbb{R}^d),\ x \in \mathbb{R}^d. \quad (4.41)$$

Note that W_t is actually a convolution operator, i.e., for $g \in L_2(\mathbb{R}^d)$ we have $W_t g = k * g$, where $k := (\pi t)^{-\frac{d}{2}} e^{-t^{-1}\|\cdot\|_2^2}$. Moreover, we have $\|k\|_1 = 1$ by (4.39), and hence Young's inequality (see Theorem A.5.23) that shows

$$\|W_t g\|_2 \le \|g\|_2\,, \qquad\qquad g \in L_2(\mathbb{R}^d),\ t > 0. \quad (4.42)$$

In other words, we have $\|W_t : L_2(\mathbb{R}^d) \to L_2(\mathbb{R}^d)\| \le 1$ for all $t > 0$.

With the help of the operator family $(W_t)_{t>0}$, we can now give another description of the spaces $H_\gamma(X)$.

Proposition 4.46. *For $0 < \gamma_1 < \gamma_2 < \infty$, we define $t := \frac{1}{2}(\gamma_2^2 - \gamma_1^2)$. Then, for all non-empty $X \subset \mathbb{R}^d$, we obtain a commutative diagram*

$$\begin{tikzcd}
H_{\gamma_2}(X) \arrow[r, "\mathrm{id}"] & H_{\gamma_1}(X) \\
L_2(\mathbb{R}^d) \arrow[u, "V_{\gamma_2}"] \arrow[r, "(\frac{\gamma_2}{\gamma_1})^{\frac{d}{2}} W_t"'] & L_2(\mathbb{R}^d) \arrow[u, "V_{\gamma_1}"']
\end{tikzcd}$$

where the vertical maps V_{γ_1} and V_{γ_2} are the metric surjections of Theorem 4.21. Moreover, these metric surjections are of the form

$$V_\gamma g(x) = \frac{2^{\frac{d}{2}}}{\gamma^{\frac{d}{2}}\pi^{\frac{d}{4}}} \int_{\mathbb{R}^d} e^{-2\gamma^{-2}\|x-y\|_2^2} g(y)dy\,, \qquad g \in L_2(\mathbb{R}^d),\ x \in X, \quad (4.43)$$

where $\gamma \in \{\gamma_1, \gamma_2\}$. Finally, we have

$$\|\,\mathrm{id} : H_{\gamma_2}(X) \to H_{\gamma_1}(X)\| \le \Bigl(\frac{\gamma_2}{\gamma_1}\Bigr)^{\frac{d}{2}}\,. \quad (4.44)$$

Proof. For $\gamma > 0$, let $V_\gamma : L_2(\mathbb{R}^d) \to H_\gamma(X)$ be the metric surjection of Theorem 4.21. Furthermore, let Φ_γ be the feature map defined in Lemma 4.45. For $g \in L_2(\mathbb{R}^d)$ and $x \in X$, we then have $V_\gamma g(x) = \langle g, \Phi_\gamma(x)\rangle_{L_2(\mathbb{R}^d)}$, and hence we obtain (4.43). In order to establish the diagram, let us first consider the case $X = \mathbb{R}^d$. Then (4.41) together with (4.43) gives the relation

$$V_\gamma g = (\pi\gamma^2)^{\frac{d}{4}} W_{\frac{\gamma^2}{2}} g\,, \qquad\qquad g \in L_2(\mathbb{R}^d). \tag{4.45}$$

Let us now show that the operator family $(W_t)_{t>0}$ is a semi-group, i.e., it satisfies

$$W_{t_1+t_2} = W_{t_1} W_{t_2}\,, \qquad\qquad t_1, t_2 > 0. \tag{4.46}$$

To this end, let us fix a $g \in L_2(\mathbb{R}^d)$ and an $x_0 \in \mathbb{R}^d$. Then, for $\alpha := \frac{t_1}{t_2}$, equations (4.40) and (4.39) yield

$$\begin{aligned} W_{t_1} W_{t_2} g(x_0) &= (\pi t_1)^{-\frac{d}{2}} \int_{\mathbb{R}^d} e^{-t_1^{-1}\|x_0-y\|_2^2} W_{t_2} g(y)dy \\ &= (\pi^2 t_1 t_2)^{-\frac{d}{2}} \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} e^{-t_1^{-1}\|x_0-y\|_2^2} e^{-t_2^{-1}\|x-y\|_2^2} g(x)dx\,dy \\ &= (\pi^2 t_1 t_2)^{-\frac{d}{2}} \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} e^{-\frac{\|x_0-x\|_2^2}{t_1+t_2} - \frac{t_1+t_2}{t_1t_2}\|y-\frac{x_0+\alpha x}{1+\alpha}\|_2^2} g(x)dy\,dx \\ &= W_{t_1+t_2} g(x_0)\,, \end{aligned}$$

i.e., (4.46) is verified. Combining (4.45) and (4.46) then gives the diagram in the case of $X = \mathbb{R}^d$. The general case $X \subset \mathbb{R}^d$ follows from the fact that the *computation* of V_γ in (4.43) is independent of X. Finally, since V_{γ_2} is a metric surjection, we obtain

$$\|\operatorname{id}\circ V_{\gamma_2} : L_2(\mathbb{R}^d) \to H_{\gamma_1}(X)\| = \|\operatorname{id} : H_{\gamma_2}(X) \to H_{\gamma_1}(X)\|\,,$$

and hence the commutativity of the diagram implies

$$\|\operatorname{id} : H_{\gamma_2}(X) \to H_{\gamma_1}(X)\| = \left(\frac{\gamma_2}{\gamma_1}\right)^{\frac{d}{2}} \|V_{\gamma_1}\circ W_t\| \le \left(\frac{\gamma_2}{\gamma_1}\right)^{\frac{d}{2}} \|W_t\|\,.$$

Moreover, we have $\|W_t\| \le 1$ by (4.42), and thus we find the assertion. □

If the set X in the preceding proposition has a non-empty interior, then the metric surjections V_{γ_1} and V_{γ_2} are actually isometric isomorphisms. This is a direct consequence of the following theorem, (4.43), and the fact that the restriction operator mapping $H_\gamma(\mathbb{R}^d)$ to $H_\gamma(X)$ is an isometric isomorphism.

Theorem 4.47 (Injectivity of Gaussian integral operators). *Let μ be either a finite measure on $\mathbb{R}^d$ or the Lebesgue measure on $\mathbb{R}^d$, and $p \in (1,\infty)$. Moreover, let k_γ be the Gaussian RBF kernel with width $\gamma > 0$. Then the operator $S_{k_\gamma} : L_p(\mu) \to H_\gamma(\mathbb{R}^d)$ defined by (4.17) is injective.*

Proof. Let us write $S_\gamma := S_{k_\gamma}$. We fix an $f \in L_p(\mu)$ with $S_\gamma f = 0$. Obviously, our goal is to show that $f = 0$. To this end, our first intermediate goal is to prove that the map $g : \mathbb{R}^d \times (0, \infty) \to \mathbb{R}$ defined by

$$g(x,t) := \int_{\mathbb{R}^d} e^{-t\|x-x'\|_2^2} f(x')\, d\mu(x')\,, \qquad\qquad x \in \mathbb{R}^d,\ t \in (0,\infty),$$

is real-analytic in t for all fixed $x \in \mathbb{R}^d$. Here we note that $e^{-t\|x-\cdot\|_2^2} \in L_{p'}(\mu)$ together with Hölder's inequality ensures that the integral above is defined and finite. To show the analyticity, we now fix a $t_0 \in (0, \infty)$ and define

$$a_i(x,x',t) := \frac{(-\|x-x'\|_2^2)^i e^{-t_0\|x-x'\|_2^2}}{i!}(t-t_0)^i\, f(x')$$

for all $x, x' \in \mathbb{R}^d$, $t \in (0, t)$, and $i \geq 0$. Obviously, we have

$$g(x,t) = \int_{\mathbb{R}^d} \sum_{i=0}^{\infty} a_i(x,x',t)\, d\mu(x') \tag{4.47}$$

for all $x \in \mathbb{R}^d$ and $t \in (0,\infty)$. Moreover, for $t \in (0, t_0]$, we find

$$\sum_{i=0}^{\infty} \bigl|a_i(x,x',t)\bigr| = \sum_{i=0}^{\infty} \frac{\|x-x'\|_2^{2i} e^{-t_0\|x-x'\|_2^2}}{i!}(t_0-t)^i f(x') = e^{-t\|x-x'\|_2^2} f(x')\,,$$

and hence Hölder's inequality yields

$$\int_{\mathbb{R}^d} \sum_{i=0}^{\infty} \bigl|a_i(x,x',t)\bigr|\, d\mu(x') < \infty\,. \tag{4.48}$$

On the other hand, for $t \in [t_0, \infty)$, we have

$$\begin{aligned}\sum_{i=0}^{\infty} \bigl|a_i(x,x',t)\bigr| &= \sum_{i=0}^{\infty} \frac{\|x-x'\|_2^{2i} e^{-t_0\|x-x'\|_2^2}}{i!}(t-t_0)^i f(x') \\ &= e^{-(2t_0-t)\|x-x'\|_2^2} f(x')\,,\end{aligned}$$

and from this it is easy to conclude by Hölder's inequality that (4.48) also holds for $t \in [t_0, 2t_0)$. By Fubini's theorem, we can then change the order of integration and summation in (4.47) to obtain

$$g(x,t) = \sum_{i=0}^{\infty} \biggl(\int_{\mathbb{R}^d} \frac{(-\|x-x'\|_2^2)^i e^{-t_0\|x-x'\|_2^2}}{i!}\, f(x')\, d\mu(x') \biggr) (t-t_0)^i$$

for all $t \in (0, 2t_0)$. In other words, $g(x, \cdot\,)$ can be locally expressed by a power series, i.e., it is real-analytic. Let us now define

$$u(x,t) := t^{-\frac{d}{2}} g\Big(x, \frac{1}{4t}\Big) = \int_{\mathbb{R}^d} t^{-\frac{d}{2}} e^{-\frac{\|x-x'\|_2^2}{4t}}\, f(x')\, d\mu(x')\,, \qquad x \in \mathbb{R}^d,\, t > 0\,.$$

Obviously, $u(x, \cdot\,)$ is again real-analytic for all $x \in \mathbb{R}^d$. Moreover, for fixed $x' := (x'_1, \dots, x'_d) \in \mathbb{R}^d$, the map

$$u_0(x,t) := t^{-\frac{d}{2}} e^{-\frac{\|x-x'\|_2^2}{4t}}\,, \qquad x \in \mathbb{R}^d,\, t > 0,$$

which appears in the integral above, satisfies

$$\begin{aligned} \frac{\partial u_0}{\partial t}(x,t) &= t^{-\frac{d}{2}-2} e^{-\frac{\|x-x'\|_2^2}{4t}} \left(\frac{\|x-x'\|_2^2}{4t} - \frac{d\,t}{2}\right), \\ \frac{\partial^2 u_0}{\partial^2 x_i}(x,t) &= t^{-\frac{d}{2}-2} e^{-\frac{\|x-x'\|_2^2}{4t}} \left(\frac{(x_i - x'_i)^2}{4t} - \frac{t}{2}\right), \end{aligned}$$

for all $t > 0$ and all $x = (x_1, \dots, x_d) \in \mathbb{R}^d$. Consequently, u_0 satisfies the partial differential equation

$$\frac{\partial u_0}{\partial t} = \varDelta u_0 := \sum_{i=1}^d \frac{\partial^2 u_0}{\partial^2 x_i}\,.$$

Moreover, as a function of x', all derivatives of u_0 are contained in $L_{p'}(\mu)$, and these derivatives are continuous with respect to the variables x and t. Another application of Hölder's inequality, together with Corollary A.3.7, shows that the function u satisfies the same partial differential equation. This leads to

$$\frac{\partial^2 u}{\partial^2 t} = \frac{\partial}{\partial t} \sum_{i=1}^d \frac{\partial^2 u}{\partial^2 x_i} = \sum_{i=1}^d \frac{\partial^3 u}{\partial^2 x_i \partial t} = \sum_{i=1}^d \sum_{j=1}^d \frac{\partial^4 u}{\partial^2 x_i \partial^2 x_j} = \varDelta^2 u\,,$$

and by iterating this procedure we obtain $\frac{\partial^n u}{\partial^n t} = \varDelta^n u$ for all $n \geq 1$. Let us now recall that our $f \in L_p(\mu)$ satisfies $S_\gamma f = 0$. For $t_0 := \gamma^2/4$, we then have $u(x, t_0) = (2/\gamma)^d\, S_\gamma f(x) = 0$ for all $x \in \mathbb{R}^d$, and hence we obtain

$$\frac{\partial^n u}{\partial^n t}(x, t_0) = \varDelta^n u(x, t_0) = 0\,, \qquad\qquad x \in \mathbb{R}^d.$$

By the analyticity of $u(x, \cdot\,)$, we thus conclude that $u(x,t) = 0$ for all $x \in \mathbb{R}^d$ and all $t > 0$. Now let $h : \mathbb{R}^d \to \mathbb{R}$ be a continuous function with compact support. Then we obviously have $\|h\|_\infty < \infty$, $h \in L_p(\mu)$, and

$$0 = \int_{\mathbb{R}^d} h(x) u(x,t) dx = t^{-\frac{d}{2}} \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} h(x) e^{-\frac{\|x-x'\|_2^2}{4t}}\, f(x')\, d\mu(x') dx \tag{4.49}$$

for all $t > 0$. Now note that if μ is finite, we easily find that

$$(x, x') \mapsto h(x) e^{-\frac{\|x-x'\|_2^2}{4t}}\, f(x') \tag{4.50}$$

is integrable with respect to the product of μ and the Lebesgue measure on $\mathbb{R}^d$. Moreover, if μ is the Lebesgue measure, its translation invariance yields

$$\begin{aligned}
&\int_{\mathbb{R}^d}\int_{\mathbb{R}^d}\bigl|h(x)e^{-\frac{\|x-x'\|_2^2}{4t}} f(x')\bigr|\, d\mu(x')dx \\
&\leq \int_{\mathbb{R}^d}|h(x)|\cdot\|f\|_{L_p(\mu)}\biggl(\int_{\mathbb{R}^d} e^{-\frac{p'\|x-x'\|_2^2}{4t}}\, d\mu(x')\biggr)^{1/p'} dx \\
&< \infty\,,
\end{aligned}$$

i.e., the function in (4.50) is integrable in this case, too. For

$$h_t(x') \;:=\; t^{-\frac{d}{2}}\int_{\mathbb{R}^d} h(x)e^{-\frac{\|x-x'\|_2^2}{4t}}\, dx, \qquad\qquad x'\in\mathbb{R}^d,\, t>0,$$

Fubini's theorem and (4.49) then yield

$$0 = t^{-\frac{d}{2}}\int\limits_{\mathbb{R}^d}\int\limits_{\mathbb{R}^d} h(x)e^{-\frac{\|x-x'\|_2^2}{4t}}\, f(x')\, dx\, d\mu(x') = \int\limits_{\mathbb{R}^d} f(x')h_t(x')d\mu(x')\,. \tag{4.51}$$

Now fix an $x\in\mathbb{R}^d$ and an $\varepsilon>0$. Then there exists a $\delta>0$ such that, for all $x'\in\mathbb{R}^d$ with $\|x'-x\|_2\leq\delta$, we have $|h(x')-h(x)|\leq(4\pi)^{-d/2}\epsilon$. Since

$$(4\pi t)^{-\frac{d}{2}}\int_{\mathbb{R}^d} e^{-\frac{\|x-x'\|_2^2}{4t}}dx' = 1\,, \qquad\qquad t>0,$$

we hence obtain

$$\begin{aligned}
h_t(x)-(4\pi)^{\frac{d}{2}}h(x) &= t^{-\frac{d}{2}}\int_{\mathbb{R}^d}\bigl(h(x')-h(x)\bigr)e^{-\frac{\|x-x'\|_2^2}{4t}}dx' \\
&\leq \varepsilon + t^{-\frac{d}{2}}\int\limits_{\|x'-x\|_2>\delta}\bigl(h(x')-h(x)\bigr)e^{-\frac{\|x-x'\|_2^2}{4t}}dx' \\
&\leq \varepsilon + 2\|h\|_\infty\, t^{-\frac{d}{2}}\int\limits_{\|x'\|_2>\delta} e^{-\frac{\|x'\|_2^2}{4t}}dx' \\
&\leq \varepsilon + 8\pi^{d/2}\frac{\max\{1,d/2\}}{\Gamma(d/2)}\|h\|_\infty\,\delta^{d-2}t^{1-d/2}e^{-\frac{\delta^2}{4t}}
\end{aligned}$$

for all $0<t\leq\delta^2/(2d)$, where in the last step we used (A.3) and (A.5). Since the last term of this estimate tends to 0 for $t\to 0$, we conclude that $\lim_{t\to 0}h_t(x) = (4\pi)^{\frac{d}{2}}h(x)$ for all $x\in\mathbb{R}^d$. Therefore the dominated convergence theorem and (4.51) yield

$$0 = \lim_{t\to 0}\int_{\mathbb{R}^d} f(x')h_t(x')d\mu(x') = \int_{\mathbb{R}^d} f(x')h(x')d\mu(x') = \langle f,h\rangle_{L_{p'}(\mu),L_p(\mu)}.$$

Since for finite measures it follows from Theorem A.3.15 and Theorem A.5.25 that the continuous functions with compact support are dense in $L_p(\mu)$, we find $f=0$. Finally, the Lebesgue measure is also regular, and hence we find the assertion in this case analogously. □

Our last goal is to compute Sobolev norms for functions in $H_\gamma(X)$. This is done in the following theorem.

Theorem 4.48 (Sobolev norms for Gaussian RKHSs). *Let $X \subset \mathbb{R}^d$ be a bounded non-empty open set, $\gamma > 0$, and $m \geq 1$. Then there exists a constant $c_{m,d} > 0$ only depending on m and d such that for all $f \in H_\gamma(X)$ we have*

$$\|f\|_{W^m(X)} \leq c_{m,d}\sqrt{\mathrm{vol}(X)} \Biggl(\sum_{\substack{\alpha\in\mathbb{N}_0^d \\ |\alpha|\leq m}} \gamma^{-2|\alpha|} \Biggr)^{1/2} \|f\|_{H_\gamma(X)} \,.$$

Proof. Let us fix a multi-index $\alpha = (\alpha_1, \dots, \alpha_d) \in \mathbb{N}_0^d$ with $|\alpha| = m$. Moreover, let $V_\gamma : L_2(X) \to H_\gamma(X)$ be the metric surjection defined by (4.43). For a fixed $f \in H_\gamma(X)$ and $\varepsilon > 0$, there then exists a $g \in L_2(\mathbb{R}^d)$ such that $V_\gamma g = f$ and $\|g\|_{L_2(\mathbb{R}^d)} \leq (1 + \varepsilon)\|f\|_{H_\gamma(X)}$. By Hölder's inequality, we then have

$$\begin{aligned}\| \,\partial^\alpha f \,\|^2_{\mathcal{L}_2(X)} &= \frac{2^d}{\gamma^d \pi^{\frac{d}{2}}} \int_X \left(\partial_x^\alpha \int_{\mathbb{R}^d} e^{-2\gamma^{-2}\|x-y\|_2^2} g(y)dy \right)^2 dx \\ &\leq \frac{2^d}{\gamma^d \pi^{\frac{d}{2}}} \int_X \left(\int_{\mathbb{R}^d} \partial_x^\alpha e^{-2\gamma^{-2}\|x-y\|_2^2} \,|g(y)|\, dy \right)^2 dx \\ &\leq \frac{2^d}{\gamma^d \pi^{\frac{d}{2}}} \|g\|^2_{L_2(\mathbb{R}^d)} \int_X \int_{\mathbb{R}^d} \bigl| \partial_x^\alpha e^{-2\gamma^{-2}\|x-y\|_2^2} \bigr|^2 dy\, dx \,. \qquad (4.52)\end{aligned}$$

Now recall that the Hermite polynomials h_n, $n \geq 0$, defined in (A.1) satisfy

$$\frac{\partial^n}{\partial t^n} e^{-t^2} = (-1)^n e^{-t^2} h_n(t) \,, \qquad\qquad t \in \mathbb{R} \,,$$

and hence we have

$$\frac{\partial^n}{\partial t^n} e^{-2\gamma^{-2}(t-s)^2} = \bigl(-\sqrt{2}\,\gamma^{-1}\bigr)^n e^{-2\gamma^{-2}(t-s)^2} h_n\bigl(\sqrt{2}\,\gamma^{-1}(t-s)\bigr)$$

for all $s, t \in \mathbb{R}$. Using the translation invariance of the Lebesgue measure, $h_n(-s) = (-1)^n h_n(s)$, a change of variables, and (A.2), we conclude that

$$\begin{aligned}\int_{\mathbb{R}} \left| \frac{d^n}{dt^n} e^{-2\gamma^{-2}(t-s)^2} \right|^2 ds &= \bigl(2\gamma^{-2}\bigr)^n \int_{\mathbb{R}} e^{-4\gamma^{-2}(t-s)^2} h_n^2\bigl(\sqrt{2}\,\gamma^{-1}(t-s)\bigr) ds \\ &= \bigl(2\gamma^{-2}\bigr)^n \int_{\mathbb{R}} e^{-4\gamma^{-2}s^2} h_n^2\bigl(\sqrt{2}\,\gamma^{-1}s\bigr) ds \\ &= \bigl(\sqrt{2}\,\gamma^{-1}\bigr)^{2n-1} \int_{\mathbb{R}} e^{-2s^2} h_n^2(s) ds \\ &\leq \sqrt{\pi}\; 2^{2n-1/2} n!\, \gamma^{1-2n} \,.\end{aligned}$$

Since $e^{-2\gamma^{-2}\|x-y\|_2^2} = \prod_{i=1}^d e^{-2\gamma^{-2}(x_i-y_i)^2}$, we hence find

$$\int_{\mathbb{R}^d} \bigl| \partial_x^\alpha e^{-2\gamma^{-2}\|x-y\|_2^2} \bigr|^2 dy \leq \pi^{m/2} 2^{2m-d/2} \alpha!\ \gamma^{d-2m} ,$$

where $\alpha! := \alpha_1! \cdots \alpha_d!$. Combining this estimate with (4.52), we obtain

$$\bigl\| \, \partial^\alpha f \, \bigr\|_{\mathcal{L}_2(X)}^2 \leq (1+\varepsilon)\, 2^{2m+d/2} \pi^{(m-d)/2} \alpha!\ \operatorname{vol}(X) \gamma^{-2m} \|f\|_{H_\gamma(X)}^2 .$$

Finally, since f is a restriction of an analytic function defined on $\mathbb{R}^d$, see Corollary 4.40, we have $\partial^{(\alpha)} f = \partial^\alpha f$, where $\partial^{(\alpha)} f$ denotes the weak α-derivative defined in Section A.5.5. From this we easily obtain the assertion. □

4.5 Mercer's Theorem (*)

In this section, we present Mercer's theorem, which provides a series representation for continuous kernels on compact domains. This series representation is then used to describe the corresponding RKHSs.

Let us begin with some preliminary considerations. To this end, let X be a measurable space, μ be a σ-finite measure on X, and k be a measurable kernel on X with $\|k\|_{L_2(\mu)} < \infty$. Moreover, recall the following factorization of the operators defined in Theorem 4.26 and Theorem 4.27:

$$\begin{tikzcd}
L_2(\mu) \arrow[rr, "T_k"] \arrow[dr, "S_k"'] & & L_2(\mu) \\
& H \arrow[ur, "S_k^*"'] &
\end{tikzcd}$$

Theorem 4.27 showed that $T_k = S_k^* S_k$ is compact, positive, and self-adjoint, and hence the Spectral Theorem A.5.13 shows that there exist an at most countable ONS $(e_i)_{i\in I}$ and a family $(\lambda_i)_{i\in I} \subset \mathbb{R}$ converging to 0 such that $|\lambda_1| \geq |\lambda_2| \geq \cdots > 0$ and

$$T_k f = \sum_{i\in I} \lambda_i \langle f, e_i \rangle e_i , \qquad\qquad f \in L_2(\mu) .$$

Moreover, $\{\lambda_i : i \in I\}$ is the set of non-zero eigenvalues of T_k. Let us write $\tilde{e}_i := \lambda_i^{-1} S_k e_i \in H$ for $i \in I$. Then the diagram shows $\tilde{e}_i = \lambda_i^{-1} T_k e_i$ almost surely, and hence we have $e_i = \lambda_i^{-1} T_k e_i = \tilde{e}_i$ almost surely. Consequently, we may assume without loss of generality that $e_i \in H$ and $\lambda_i e_i = S_k e_i$ for all $i \in I$. From this we conclude that

$$\begin{aligned}
\lambda_i \lambda_j \langle e_i, e_j \rangle_H = \langle S_k e_i, S_k e_j \rangle_H = \langle e_i, S_k^* S_k e_j \rangle_{L_2(\mu)} &= \langle e_i, T_k e_j \rangle_{L_2(\mu)} \\
&= \lambda_j \langle e_i, e_j \rangle_{L_2(\mu)} .
\end{aligned}$$

In other words, $(\sqrt{\lambda_i} e_i)_{i \in I}$ is an ONS in H. The goal of this section is to show that under certain circumstances this family is even an ONB. To this end, we need the following theorem, whose proof can be found, for example, in Riesz and Nagy (1990).

Theorem 4.49 (Mercer's theorem). *Let X be a compact metric space and $k : X \times X \to \mathbb{R}$ be a continuous kernel. Furthermore, let μ be a finite Borel measure with $\operatorname{supp} \mu = X$. Then, for $(e_i)_{i \in I}$ and $(\lambda_i)_{i \in I}$ as above, we have*

$$k(x, x') = \sum_{i \in I} \lambda_i e_i(x) e_i(x') \,, \qquad x, x' \in X, \tag{4.53}$$

where the convergence is absolute and uniform.

Note that (4.53) together with the proof of Lemma 4.2 shows that $\Phi : X \to \ell_2$ defined by $\Phi(x) := (\sqrt{\lambda_i} e_i(x))_{i \in I}$, $x \in X$, is a feature map of k. In order to show that $(\sqrt{\lambda_i} e_i)_{i \in I}$ is an ONB of H, we need the following corollary.

Corollary 4.50. *With the assumptions and notations of Theorem 4.49, the series $\sum_{i \in I} a_i \sqrt{\lambda_i} e_i(x)$ converges absolutely and uniformly for all $(a_i) \in \ell_2(I)$.*

Proof. For $x \in X$ and $J \subset I$, Hölder's inequality and Mercer's theorem imply

$$\sum_{i \in J} |a_i \sqrt{\lambda_i} e_i(x)| \leq \biggl(\sum_{i \in J} a_i^2\biggr)^{1/2} \biggl(\sum_{i \in J} \lambda_i e_i^2(x)\biggr)^{1/2} = \|(a_i)\|_{\ell_2(I)} \cdot \sqrt{k(x,x)} \,.$$

From this the assertion easily follows. □

With the help of the Corollary 4.50, we can now give an explicit representation of the RKHSs of continuous kernels on a compact metric space X.

Theorem 4.51 (Mercer representation of RKHSs). *With the assumptions and notations of Theorem 4.49, we define*

$$H := \biggl\{ \sum_{i \in I} a_i \sqrt{\lambda_i} e_i \ : \ (a_i) \in \ell_2(I) \biggr\} .$$

Moreover, for $f := \sum_{i \in I} a_i \sqrt{\lambda_i} e_i \in H$ and $g := \sum_{i \in I} b_i \sqrt{\lambda_i} e_i \in H$, we write

$$\langle f, g \rangle_H := \sum_{i \in I} a_i b_i \ .$$

Then H equipped with inner product $\langle \,\cdot\, , \,\cdot\, \rangle_H$ is the RKHS of the kernel k. Furthermore, the operator $T_k^{1/2} : L_2(\mu) \to H$ is an isometric isomorphism.

Proof. Routine work shows that $\langle\,\cdot\,,\,\cdot\,\rangle$ is a well-defined inner product and hence H is a Hilbert function space. Now, for fixed $x \in X$, Mercer's theorem implies

$$k(\,\cdot\,,x) = \sum_{i\in I} \sqrt{\lambda_i} e_i(x) \sqrt{\lambda_i} e_i(\,\cdot\,)\,,$$

and since Mercer's theorem also yields

$$\|(\sqrt{\lambda_i} e_i(x))\|^2_{\ell_2(I)} = \sum_{i\in I} \lambda_i e_i^2(x) = k(x,x) < \infty\,,$$

we find $k(\,\cdot\,,x) \in H$. Moreover, for $f := \sum_{i\in I} a_i \sqrt{\lambda_i} e_i \in H$, we have

$$\langle f, k(\,\cdot\,,x)\rangle_H = \sum_{i\in I} a_i \sqrt{\lambda_i} e_i(x) = f(x)\,, \qquad x \in X,$$

i.e., k is the reproducing kernel of H.

Let us now consider the operator $T_k^{1/2}$. To this end, let us fix an $f \in L_2(\mu)$. Since (e_i) is an orthonormal basis in $L_2(\mu)$, we then find $f = \sum_{i\in I} \langle f, e_i\rangle_{L_2(\mu)} e_i$, where the convergence takes place in $L_2(\mu)$. Consequently, we have

$$T_k^{1/2} f = \sum_{i\in I} \langle f, e_i\rangle_{L_2(\mu)} \sqrt{\lambda_i} e_i\,, \tag{4.54}$$

where the convergence is again with respect to the $L_2(\mu)$-norm. Now, Parseval's identity gives $(\langle f, e_i\rangle_{L_2(\mu)}) \in \ell_2(I)$, and hence we find $T_k^{1/2} f \in H$ for all $f \in L_2(\mu)$. Moreover, this also shows by Corollary 4.50 that the convergence in (4.54) is absolute and uniform and that

$$\|T_k^{1/2} f\|_H^2 = \sum_{i\in I} |\langle f, e_i\rangle_{L_2(\mu)}|^2 = \|f\|^2_{L_2(\mu)}\,.$$

In other words, $T_k^{1/2} : L_2(\mu) \to H$ is isometric. Finally, to check that the operator is surjective, we fix an $f \in H$. Then there exists an $(a_i) \in \ell_2$ such that $f(x) = \sum_{i\in I} a_i \sqrt{\lambda_i} e_i(x)$ for all $x \in X$. Now we obviously have $g := \sum_{i\in I} a_i e_i \in L_2(\mu)$ with convergence in $L_2(\mu)$, and thus $\langle g, e_i\rangle_{L_2(\mu)} = a_i$. Furthermore, we have already seen that the convergence in (4.54) is pointwise, and hence for all $x \in X$ we finally obtain

$$T_k^{1/2} g(x) = \sum_{i\in I} \langle g, e_i\rangle_{L_2(\mu)} \sqrt{\lambda_i} e_i(x) = \sum_{i\in I} a_i \sqrt{\lambda_i} e_i(x) = f(x)\,. \qquad \square$$

4.6 Large Reproducing Kernel Hilbert Spaces

We saw in Section 1.2 that SVMs are based on minimization problems over RKHSs. Moreover, we will see in the following chapters that the size of the chosen RKHS has a twofold impact on the generalization ability of the SVM: on the one hand, a "small size" inhibits the learning machine to produce highly complex decision functions and hence can prevent the SVM from overfitting in the presence of noise. On the other hand, for complex distributions, a "small" RKHS may not be sufficient to provide an accurate decision function, so the SVM underfits. In this section, we thus investigate RKHSs that are rich enough to provide *arbitrarily accurate* decision functions for *all* distributions. The reason for introducing these RKHSs is that their flexibility is necessary to guarantee learning in the absence of assumptions on the data-generating distribution. However, as we have indicated above, this flexibility also carries the danger of overfitting. We will thus investigate in Chapters 6 and 7 how regularized learning machines such as SVMs use the regularizer to avoid this overfitting.

Let us now begin by introducing a class of particularly large RKHSs.

Definition 4.52. *A continuous kernel k on a compact metric space X is called **universal** if the RKHS H of k is dense in $C(X)$, i.e., for every function $g \in C(X)$ and all $\varepsilon > 0$ there exists an $f \in H$ such that*

$$\|f - g\|_\infty \leq \varepsilon\,.$$

Instead of using the RKHS in the preceding definition, one can actually consider an arbitrary feature space H_0 of k. Indeed, if $\Phi_0 : X \to H_0$ is a corresponding feature map, then the RKHS of k is given by (4.10) and hence k is universal if and only if for all $g \in C(X)$ and $\varepsilon > 0$ there exists a $w \in H_0$ such that $\|\langle w, \Phi_0(\,\cdot\,)\rangle - g\|_\infty \leq \varepsilon$. Although this is a rather trivial observation, we will see below that it is very useful for finding universal kernels.

One may wonder whether the preceding definition also makes sense for compact *topological* spaces. At first glance, this is indeed the case, but some further analysis shows that there exists no universal kernel if the topology is not generated by a metric (see Exercise 4.13).

Let us now discuss some of the surprising geometric properties of universal kernels. To this end, we need the following definition.

Definition 4.53. *Let k be a kernel on a metric space X with RKHS H. We say that k **separates the disjoint sets** $A, B \subset X$ if there exists an $f \in H$ with $f(x) > 0$ for all $x \in A$, and $f(x) < 0$ for all $x \in B$. Furthermore, we say that k **separates all finite (or compact) sets** if k separates all finite (or compact) disjoint sets $A, B \subset X$.*

It can be shown (see Exercise 4.11) that strictly positive definite kernels separate all finite sets. Furthermore, every kernel that separates all compact sets obviously also separates all finite sets, but in general the converse is not true (see Exercise 4.14). Moreover, every universal kernel separates all compact sets, as the following proposition shows.

Proposition 4.54. *Let X be a compact metric space and k be a universal kernel on X. Then k separates all compact sets.*

Proof. Let $A, B \subset X$ be disjoint compact subsets and d be the metric of X. Then, for all $x \in X$, we define

$$g(x) := \frac{\operatorname{dist}(x, B)}{\operatorname{dist}(x, A) + \operatorname{dist}(x, B)} - \frac{\operatorname{dist}(x, A)}{\operatorname{dist}(x, A) + \operatorname{dist}(x, B)} ,$$

where we used the distance function $\operatorname{dist}(x, C) := \inf_{x' \in C} \operatorname{dist}(x, x')$ for $x \in X$ and $C \subset X$. Since this distance function is continuous, we see that g is a continuous function. Furthermore, we have $g(x) = 1$ for all $x \in A$ and $g(x) = -1$ for all $x \in B$. Now, let H be the RKHS of k. Then there exists an $f \in H$ with $\|f - g\|_\infty \leq 1/2$, and by our previous considerations this f then satisfies $f(x) \geq 1/2$ for all $x \in A$ and $f(x) \leq 1/2$ for all $x \in B$. □

Although Proposition 4.54 easily follows from the notion of universality, it has surprising consequences for the geometric interpretation of the shape of the feature maps of universal kernels. Indeed, let k be a universal kernel on X with feature space H_0 and feature map $\Phi_0 : X \to H_0$. Furthermore, let us suppose that we have a finite subset $\{x_1, \dots, x_n\}$ of X. Then Proposition 4.54 ensures that for *every* choice of signs $y_1, \dots, y_n \in \{-1, 1\}$ we find a function f in the RKHS H of k with $y_i f(x_i) > 0$ for all $i = 1, \dots, n$. By (4.10), this f can be represented by $f = \langle w, \Phi_0(\,\cdot\,)\rangle$ for a suitable $w \in H_0$. Consequently, the mapped training set $((\Phi_0(x_1), y_1), \dots, (\Phi_0(x_n), y_n))$ can be correctly separated in H_0 by the hyperplane defined by w. Moreover, a closer look at the proof of Proposition 4.54 shows that this can even be done by a hyperplane that has almost the same distance to every point of $\Phi(x_i)$, $i = 1, \dots, n$. Obviously, all these phenomena are impossible for general training sets in $\mathbb{R}^2$ or $\mathbb{R}^3$, and hence every two- or three-dimensional illustration of the feature space of universal kernels such as Figure 1.1 can be misleading. In particular, it seems to be very difficult to *geometrically* understand the learning mechanisms of both hard- and soft margin SVMs when these SVMs use universal kernels.

The geometric interpretation above raises the question of whether universal kernels can exist. As we will see below, the answer to this question is "yes" and in addition, many standard kernels, including the Gaussian RBF kernels, are universal. To establish these results, we need the following simple lemma.

Lemma 4.55 (Properties of universal kernels). *Let X be a compact metric space and k be a universal kernel on X. Then the following statements are true:*

i) Every feature map of k is injective.
ii) We have $k(x, x) > 0$ for all $x \in X$.
iii) Every restriction of k onto some compact $X' \subset X$ is universal.
iv) The ***normalized kernel*** *$k^* : X \times X \to \mathbb{R}$ defined by*

$$k^*(x, x') := \frac{k(x, x')}{\sqrt{k(x, x) k(x', x')}} , \qquad x, x' \in X,$$

is universal.

Proof. The first three assertions are direct consequences of Proposition 4.54 and the definition. To prove the fourth assertion, let $\Phi : X \to H$ be the canonical feature map of k into its RKHS H. Defining $\alpha(x) := k(x,x)^{-1/2}$ for all $x \in X$, we see that $\alpha\Phi : X \to H$ is a feature map of k^* and thus k^* is a kernel. To show that k^* is universal, we fix a function $g \in C(X)$ and an $\varepsilon > 0$. For $c := \|\alpha\|_\infty < \infty$, we then get an $f \in H$ with $\|f - \frac{g}{\alpha}\|_\infty \leq \frac{\varepsilon}{c}$. This yields

$$\big\| \langle f, \alpha(\,\cdot\,)\Phi(\,\cdot\,)\rangle - g \,\big\|_\infty \leq \|\alpha\|_\infty \,\Big\| f - \frac{g}{\alpha} \Big\|_\infty \leq \varepsilon\,,$$

and thus k^* is universal by the observation following Definition 4.52. □

Let us now investigate the existence of universal kernels. We begin by presenting a simple sufficient condition for the universality of kernels.

Theorem 4.56 (A test for universality). *Let X be a compact metric space and k be a continuous kernel on X with $k(x,x) > 0$ for all $x \in X$. Suppose that we have an injective feature map $\Phi : X \to \ell_2$ of k. We write $\Phi_n : X \to \mathbb{R}$ for its n-th component, i.e., $\Phi(x) = (\Phi_n(x))_{n\in\mathbb{N}}$, $x \in X$. If $\mathcal{A} := \operatorname{span}\{\Phi_n : n \in \mathbb{N}\}$ is an algebra, then k is universal.*

Proof. We will apply Stone-Weierstraß' theorem (see Theorem A.5.7). To this end, we first observe that the algebra $\mathcal{A}$ does not vanish since $\|(\Phi_n(x))\|_{\ell_2}^2 = k(x,x) > 0$ for all $x \in X$. Moreover, k is continuous and thus every $\Phi_n : X \to \mathbb{R}$ is continuous by Lemma 4.29. This shows that $\mathcal{A} \subset C(X)$. Moreover, the injectivity of Φ implies that $\mathcal{A}$ separates points, and thus Stone-Weierstraß' theorem shows that $\mathcal{A}$ is dense in $C(X)$. Now we fix a $g \in C(X)$ and an $\varepsilon > 0$. Then there exists a function $f \in \mathcal{A}$ of the form

$$f = \sum_{j=1}^{m} \alpha_j \Phi_{n_j}$$

with $\|f - g\|_\infty \leq \varepsilon$. For $n \in \mathbb{N}$, we define $w_n := \alpha_j$ if there is an index j with $n_j = n$ and $w_n := 0$ otherwise. This yields $w := (w_n) \in \ell_2$ and $f = \langle w, \Phi(\,\cdot\,)\rangle_{\ell_2}$, and thus k is universal by the observation following Definition 4.52. □

With the help of the preceding theorem, we are now in a position to give examples of universal kernels. Let us begin with kernels of Taylor type.

Corollary 4.57 (Universal Taylor kernels). *Fix an $r \in (0,\infty]$ and a C^∞-function $f : (-r,r) \to \mathbb{R}$ that can be expanded into its Taylor series at 0, i.e.,*

$$f(t) = \sum_{n=0}^{\infty} a_n t^n\,, \qquad t \in (-r,r).$$

Let $X := \{x \in \mathbb{R}^d : \|x\|_2 < \sqrt{r}\}$. If we have $a_n > 0$ for all $n \geq 0$, then k given by

$$k(x,x') := f(\langle x, x'\rangle)\,, \qquad x,x' \in X,$$

is a universal kernel on every compact subset of X.

Proof. We have already seen in Lemma 4.8 and its proof that k is a kernel with feature space $\ell_2(\mathbb{N}_0^d)$ and feature map $\Phi : X \to \ell_2(\mathbb{N}_0^d)$ defined by

$$\Phi(x) := \Bigl(\sqrt{a_{j_1+\cdots+j_d} c_{j_1,\ldots,j_d}} \prod_{i=1}^d x_i^{j_i}\Bigr)_{j_1,\ldots,j_d \geq 0}, \qquad x \in X.$$

Obviously, k is also continuous and $a_0 > 0$ implies $k(x,x) > 0$ for all $x \in X$. Furthermore, it is easy to see that Φ is injective. Finally, since polynomials form an algebra, $\operatorname{span}\{\Phi_{j_1,\ldots,j_d} : j_1,\ldots,j_d \geq 0\}$ is an algebra, and thus we obtain by Theorem 4.56 that k is universal. □

Recall that we presented some examples of Taylor kernels in Section 4.1. The following corollary shows that all these kernels are universal.

Corollary 4.58 (Examples of universal kernels). *Let X be a compact subset of $\mathbb{R}^d$, $\gamma > 0$, and $\alpha > 0$. Then the following kernels on X are universal:*

$$\begin{aligned} \textit{exponential kernel:} &\quad k(x,x') := \exp(\langle x, x'\rangle)\,, \\ \textit{Gaussian RBF kernel:} &\quad k_\gamma(x,x') := \exp(-\gamma^{-2}\|x-x'\|_2^2)\,, \\ \textit{binomial kernel:} &\quad k(x,x') := (1-\langle x, x'\rangle)^{-\alpha}\,, \end{aligned}$$

where for the last kernel we additionally assume $X \subset \{x \in \mathbb{R}^d : \|x\|_2 < 1\}$.

Proof. The assertion follows from Examples 4.9 and 4.11, Proposition 4.10, Corollary 4.57, and part *iv)* of Lemma 4.55. □

Note that a result similar to Corollary 4.57 can be established for Fourier type kernels (see Exercise 4.12 for details). Furthermore, it is obvious that polynomial kernels cannot be universal whenever $|X| = \infty$. By Proposition 5.41, it will thus be easy to show that there do exist learning problems that are extremely underfitted by these types of kernels.

We will see in Corollary 5.29 that the universality of a kernel with RKHS H guarantees

$$\inf_{f\in H} \mathcal{R}_{L,\mathrm{P}}(f) = \mathcal{R}_{L,\mathrm{P}}^* \tag{4.55}$$

for all continuous P-integrable Nemitski losses. However, this result requires the input space X to be a compact metric space, and hence many interesting spaces, such as $\mathbb{R}^d$ and *infinite* discrete sets, are excluded. On the other hand, Theorem 5.31 will show that, for almost all interesting loss functions, it suffices to know that H is dense in $L_p(\mathrm{P}_X)$ for some $p \geq 1$ in order to establish (4.55). In the rest of this section, we will therefore investigate RKHSs that are dense in $L_p(\mathrm{P}_X)$. To this end, our main tool will be Theorem 4.26, which characterized this type of denseness by the injectivity of the associated integral operator $S_k : L_{p'}(\mathrm{P}_X) \to H$ defined by (4.17).

We begin by considering distributions P_X that are absolutely continuous with respect to a suitable reference measure μ.

Lemma 4.59. *Let X be a measurable space, μ be a measure on X, and k be a measurable kernel on X with RKHS H and $\|k\|_{L_p(\mu)} < \infty$ for some $p \in [1,\infty)$. Assume that the integral operator $S_k : L_{p'}(\mu) \to H$ is injective. Then H is dense in $L_q(h\mu)$ for all $q \in [1,p]$ and all measurable $h : X \to [0,\infty)$ with $h \in L_s(\mu)$, where $s := \frac{p}{p-q}$.*

Proof. Let us fix an $f \in L_{q'}(h\mu)$. Then we have $f|h|^{\frac{1}{q'}} \in L_{q'}(\mu)$ and, for r defined by $\frac{1}{q'} + \frac{1}{r} = \frac{1}{p'}$, Hölder's inequality and $\frac{r}{q} = s$ thus yield

$$\|fh\|_{L_{p'}(\mu)} \;=\; \Big\| f|h|^{\frac{1}{q'}}\,|h|^{\frac{1}{q}}\Big\|_{L_{p'}(\mu)} \;\le\; \Big\| f|h|^{\frac{1}{q'}}\Big\|_{L_{q'}(\mu)}\Big\| |h|^{\frac{1}{q}}\Big\|_{L_r(\mu)} < \infty\,.$$

Moreover, if $f \neq 0$ in $L_{q'}(h\mu)$, we have $fh \neq 0$ in $L_{p'}(\mu)$, and hence we obtain

$$0 \neq S_k(fh) = \int_X f(x)h(x)k(\,\cdot\,,x)\,d\mu(x) = \int_X f(x)k(\,\cdot\,,x)\,d(h\mu)(x)\,.$$

Since the latter integral describes the integral operator $L_{q'}(h\mu) \to H$, we then obtain the assertion by Theorem 4.26. □

Let us now investigate denseness properties of RKHSs over discrete spaces X. To this end, let us write $\ell_p(X) := L_p(\nu)$, where $p \in [1,\infty]$ and ν is the counting measure on X, which is defined by $\nu(\{x\}) = 1$, $x \in X$. Note that these spaces obviously satisfy the inclusion $\ell_p(X) \subset \ell_q(X)$ for $p \le q$, which is used in the proof of the following result.

Proposition 4.60 (Large RKHSs on discrete spaces I). *Let X be a countable set and k be a kernel on X with $\|k\|_{\ell_p(X)} < \infty$ for some $p \in [1,\infty)$. If k satisfies*

$$\sum_{x,x' \in X} k(x,x')f(x)f(x') > 0 \tag{4.56}$$

for all $f \in \ell_{p'}(X)$ with $f \neq 0$, then the RKHS of k is dense in $L_q(\mu)$ for all $q \in [1,\infty)$ and all distributions μ on X.

Proof. Recall that the counting measure ν is σ-finite since X is countable. Let us fix an $f \in \ell_{p'}(X)$ with $f \neq 0$. For the operator $S_k : \ell_{p'}(X) \to H$ defined by (4.17), we then have $S_k f \in H \subset \ell_p(X)$ and hence we obtain

$$\langle S_k f, f\rangle_{\ell_p(X),\ell_{p'}(X)} = \sum_{x,x' \in X} k(x,x')f(x)f(x') > 0\,.$$

This shows that $S_k : \ell_{p'}(X) \to H$ is injective. Now let μ be a distribution on X. Then there exists a function $h \in \ell_1(X)$ with $\mu = h\nu$. Since for $q \in [1,p]$ we have $s := \frac{p}{p-q} \ge 1$, we then find $h \in \ell_s(X)$ and hence we obtain the assertion by applying Lemma 4.59. In addition, for $q > p$, we have $\|k\|_{\ell_q(X)} \le \|k\|_{\ell_p(X)} < \infty$ and $\ell_{q'}(X) \subset \ell_{p'}(X)$, and consequently this case follows from the case $q = p$ already shown . □

Note that the case $p = \infty$ is excluded in Proposition 4.60. The reason for this is that the dual of $\ell_\infty(X)$ is *not* $\ell_1(X)$. However, if instead we consider the *pre*-dual of $\ell_1(X)$, namely the Banach **space of functions vanishing at infinity**,

$$c_0(X) := \bigl\{ f : X \to \mathbb{R} \,\big|\, \forall \varepsilon > 0 \, \exists \text{ finite } A \subset X \ \forall x \in X \backslash A : |f(x)| \leq \varepsilon \bigr\} ,$$

which is equipped with the usual $\|\cdot\|_\infty$-norm, we obtain the following result.

Theorem 4.61 (Large RKHSs on discrete spaces II). *Let X be a countable set and k be a bounded kernel on X that satisfies both $k(\,\cdot\,, x) \in c_0(X)$ for all $x \in X$ and (4.56) for all $f \in \ell_1(X)$ with $f \neq 0$. Then the RKHS of k is dense in $c_0(X)$.*

Proof. Since $k(\,\cdot\,, x) \in c_0(X)$ for all $x \in X$, we see $H_{\text{pre}} \subset c_0(X)$, where H_{pre} is the space defined in (4.12). Let us write H for the RKHS of k. Since k is bounded, the inclusion $I : H \to \ell_\infty(X)$ is well-defined and continuous by Lemma 4.23. Now let us fix an $f \in H$. By Theorem 4.21, there then exists a sequence $(f_n) \subset H_{\text{pre}}$ with $\lim_{n\to\infty} \|f - f_n\|_H = 0$, and the continuity of $I : H \to \ell_\infty(X)$ then yields $\lim_{n\to\infty} \|f - f_n\|_\infty = 0$. Now the completeness of $c_0(X)$ shows that $c_0(X)$ is a closed subspace of $\ell_\infty(X)$, and since we already know $f_n \in c_0(X)$ for all $n \geq 1$, we can conclude that $f \in c_0(X)$. In other words, the inclusion $I : H \to c_0(X)$ is well-defined and continuous. Moreover, a simple calculation analogous to the one in the proof of Theorem 4.26 shows that its adjoint operator is the integral operator $S_k : \ell_1(X) \to H$. Since this operator is injective by (4.56), we see that H is dense in $c_0(X)$ by Theorem 4.26. □

One may be tempted to assume that condition (4.56) is already satisfied if it holds for all functions $f : X \to \mathbb{R}$ with $0 < |\{x \in X : f(x) \neq 0\}| < \infty$, i.e., for strictly positive definite kernels. The following result shows that this is not the case in a strong sense.

Theorem 4.62. *There exists a bounded, strictly positive definite kernel k on $X := \mathbb{N}_0$ with $k(\,\cdot\,, x) \in c_0(X)$ for all $x \in X$ such that for all finite measures μ on X with $\mu(\{x\}) > 0$, $x \in X$, and all $q \in [1, \infty]$, the RKHS H of k is not dense in $L_q(\mu)$.*

Proof. Let us write $p_n := \mu(\{n\})$, $n \in \mathbb{N}_0$. Moreover, let $(b_i)_{i\geq 1} \subset (0,1)$ be a strictly positive sequence with $\|(b_i)\|_2 = 1$ and $(b_i) \in \ell_1$. Furthermore, let (e_n) be the canonical ONB of ℓ_2. We write $\Phi(0) := (b_i)$ and $\Phi(n) := e_n$, $n \geq 1$. Then we have $\Phi(n) \in \ell_2$ for all $n \in \mathbb{N}_0$, and hence

$$k(n, m) := \bigl\langle \Phi(n), \Phi(m) \bigr\rangle_{\ell_2} , \qquad\qquad n, m \geq 0,$$

defines a kernel. Moreover, an easy calculation shows $k(0,0) = 1$, $k(n,m) = \delta_{n,m}$, and $k(n,0) = b_n$ for $n, m \geq 1$. Since $b_n \to 0$, we hence find $k(\,\cdot\,, n) \in$

$c_0(X)$ for all $n \in \mathbb{N}_0$. Now let $n \in \mathbb{N}_0$ and $\alpha := (\alpha_0, \dots, \alpha_n) \in \mathbb{R}^{n+1}$ be a vector with $\alpha \neq 0$. Then the definition of k yields

$$\begin{aligned} A := \sum_{i=0}^{n}\sum_{j=0}^{n} \alpha_i\alpha_j k(i,j) &= \alpha_0^2 k(0,0) + 2\sum_{i=1}^{n} \alpha_i\alpha_0 k(i,0) + \sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i\alpha_j k(i,j) \\ &= \alpha_0^2 + 2\alpha_0 \sum_{i=1}^{n} \alpha_i b_i + \sum_{i=1}^{n} \alpha_i^2 \\ &= \alpha_0^2 + \sum_{i=1}^{n} \alpha_i (2\alpha_0 b_i + \alpha_i)\,. \end{aligned}$$

If $\alpha_0 = 0$, we hence find $A = \sum_{i=1}^n \alpha_i^2 > 0$ since we assumed $\alpha \neq 0$. Moreover, if $\alpha_0 \neq 0$, we find $t(2\alpha_0 b_i + t) \geq -\alpha_0^2 b_i^2$ for all $t \in \mathbb{R}$ by simple calculus, and hence our assumptions $\|(b_i)\|_2 = 1$ and $b_i > 0$, $i \geq 1$, imply

$$A \;\geq\; \alpha_0^2 - \sum_{i=1}^{n} \alpha_0^2 b_i^2 \;=\; \alpha_0^2 \sum_{i=n+1}^{\infty} b_i^2 \;>\; 0\,.$$

Consequently, we have $A > 0$ in any case, and from this it is easy to see that k is strictly positive definite. Let us now define $f : \mathbb{N}_0 \to \mathbb{R}$ by $f(0) := 1$ and $f(n) := -\frac{b_n}{p_n} p_0$ for $n \geq 1$. Then we have $\|f\|_{L_1(\mu)} = p_0 + p_0\|(b_n)\|_{\ell_1} < \infty$, and a simple calculation yields

$$S_k f(0) = k(0,0)f(0)p_0 + \sum_{n=1}^{\infty} k(0,n)f(n)p_n = p_0 - p_0\sum_{n=1}^{\infty} b_n^2 = 0\,.$$

Furthermore, for $m \geq 1$, our construction yields

$$S_k f(m) = k(m,0)f(0)p_0 + \sum_{n=1}^{\infty} k(m,n)f(n)p_n = b_m f(0)p_0 - f(m)p_m = 0\,,$$

and hence we have $S_k f = 0$, i.e., $S_k : L_1(\mu) \to H$ is not injective. Moreover, by (A.34), the space $L_1(\mu)$ can be interpreted as a subspace of $L'_\infty(\mu)$, and we have $S''_k f = S_k f$ for all $f \in L_1(\mu)$ as we mention in (A.20). From this we conclude that $S''_k : L'_\infty(\mu) \to H$ is not injective, and hence $S'_k : H \to L_\infty(\mu)$ does not have a dense image. Repeating the proof of Theorem 4.26, we further see that $\mathrm{id} : H \to L_\infty(\mu)$ equals S'_k, and thus H is not dense in $L_\infty(\mu)$. From this we easily find the assertion for $q \in [1,\infty)$. □

Finally, let us treat the Gaussian RBF kernels yet another time.

Theorem 4.63 (Gaussian RKHS is large). *Let $\gamma > 0$, $p \in [1,\infty)$, and μ be a finite measure on $\mathbb{R}^d$. Then the RKHS $H_\gamma(\mathbb{R}^d)$ of the Gaussian RBF kernel k_γ is dense in $L_p(\mu)$.*

Proof. Since $L_p(\mu)$ is dense in $L_1(\mu)$, it suffices to consider the case $p > 1$. Moreover, by Theorem 4.26, it suffices to show that the integral operator $S_{k_\gamma} : L_{p'}(\mu) \to H_\gamma(\mathbb{R}^d)$ of k_γ is injective. However, the latter was already established in Theorem 4.47. □

4.7 Further Reading and Advanced Topics

The idea of using kernels for pattern recognition algorithms dates back to the 1960s, when Aizerman *et al.* (1964) gave a feature space interpretation of the potential function method. However, it took almost thirty years before Boser *et al.* (1992) combined this idea with another old idea, namely the generalized portrait algorithm of Vapnik and Lerner (1963), in the hard margin SVM. Shortly thereafter, Cortes and Vapnik (1995) added slack variables to this first type of SVM, which led to soft margin SVMs. In these papers on SVMs, the feature space interpretation was based on an informal version of Mercer's theorem, which may cause some misunderstandings, as discussed in Exercise 4.10. The RKHS interpretation for SVMs was first found in 1996 and then spread rapidly; see, e.g., the books by Schölkopf (1997) and Vapnik (1998). For more information, we refer to G. Wahba's talk on multi-class SVMs given at IPAM in 2005 (see `http://www.oid.ucla.edu/Webcast/ipam/`). Since the introduction of SVMs, many kernels for specific learning tasks have been developed; for an overview, we refer to Schölkopf and Smola (2002) and Shawe-Taylor and Cristianini (2004). In addition, it was first observed by Schölkopf *et al.* (1998) that the "kernel trick", i.e., the idea of combining a linear algorithm with a kernel to obtain a non-linear algorithm, works not only for SVMs but actually for a variety of different algorithms. Many of these "kernelized" algorithms can be found in the books by Schölkopf and Smola (2002) and Shawe-Taylor and Cristianini (2004).

As indicated above, the use of kernels for machine learning methods was discovered relatively recently. However, the theory of kernels and their applications to various areas of mathematics are much older. Indeed, Mercer's theorem has been known for almost a century (see Mercer, 1909), and based on older work by Moore (1935, 1939) and others, Aronszajn (1950) developed the theory of RKHSs in the 1940s. The latter article also provides a good overview of the early history and the first applications of kernels. Since then, many new applications have been discovered. We refer to the books by Berlinet and Thomas-Agnan (2004), Ritter (2000), and Wahba (1990) for a variety of examples.

We must admit that two important types of kernels have been almost completely ignored in this chapter. The first of these are the **translation-invariant kernels**, i.e., kernels $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{K}$ for which there exists a function $\kappa : \mathbb{R}^d \to \mathbb{K}$ such that

$$k(x, x') = \kappa(x - x'), \qquad x, x' \in \mathbb{R}^d. \tag{4.57}$$

Bochner (1932, 1959) showed that, given a continuous function $\kappa : \mathbb{R}^d \to \mathbb{C}$, equation (4.57) defines a kernel k if and only if there exists a unique finite Borel measure μ on $\mathbb{R}^d$ such that

$$\kappa(x) = \int_{\mathbb{R}^d} e^{i\langle x, y\rangle} \, d\mu(y), \qquad x \in \mathbb{R}^d. \tag{4.58}$$

From this and Exercise 4.5, it is easy to conclude that for continuous functions $\kappa : \mathbb{R}^d \to \mathbb{R}$, equation (4.57) defines a kernel if there exists a unique finite Borel measure μ on $\mathbb{R}^d$ such that

$$\kappa(x) = \int_{\mathbb{R}^d} \cos\langle x, y\rangle \, d\mu(y)\,, \qquad\qquad x \in \mathbb{R}^d. \tag{4.59}$$

Note that this sufficient condition is a generalization of the Fourier kernels introduced in Lemma 4.12, and in fact one could prove this condition directly along the lines of the proof of Lemma 4.12. Finally, Cucker and Zhou (2007) showed in their Proposition 2.14 that k is a kernel if the Fourier transform of κ is non-negative. The second type of kernel we did not systematically consider are **radial kernels**, i.e., kernels $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ for which there exists a function $\kappa : \mathbb{R}^d \to \mathbb{R}$ such that

$$k(x, x') = \kappa(\|x - x'\|_2^2)\,, \qquad\qquad x, x' \in \mathbb{R}^d. \tag{4.60}$$

Schoenberg (1938), see also Section 5.2 in Berg *et al.* (1984), showed that, given a continuous function $\kappa : \mathbb{R} \to \mathbb{R}$, equation (4.60) defines a kernel k for *all* $d \geq 1$, if and only if there exists a unique finite Borel measure μ on $[0, \infty)$ such that

$$\kappa(t) = \int_{\mathbb{R}^d} e^{-ty} \, d\mu(y)\,, \qquad\qquad t \in [0, \infty). \tag{4.61}$$

Finally, it is known that if κ is completely monotonic, then (4.60) defines a kernel. For a proof, we refer to Proposition 2.18 of Cucker and Zhou (2007).

Most of the material presented in Sections 4.1, 4.2, and 4.3 is folklore and can be found in many other introductions to RKHSs (see, e.g., Hille, 1972; Meschkowski, 1962; Saitoh, 1988, 1997). Polynomial kernels were first used in the machine literature by Poggio (1975). The exponential kernel and its RKHS are closely related to the so-called *Fock space* considered in quantum mechanics (see, e.g., Bargmann, 1961; Folland, 1989). Furthermore, the binomial kernel is a generalization of the Bergmann kernel (see, e.g., Duren, 1970; Duren and Schuster, 2004; Hedenmalm *et al.*, 2000), and the examples of Fourier type kernels were considered by Vapnik (1998), who also presents some more examples of kernels of possible interest for machine learning. Finally, the notion of separately continuous kernels in Section 4.3 is taken from Hein and Bousquet (2004).

The description of $H_\gamma(X)$ follows Steinwart *et al.* (2006a), but some of the results can also be found in the book by Saitoh (1997). The operator W_t is known as the *Gauss-Weierstraß integral operator* and is used for the heat equation (see, e.g., Hille and Phillips, 1957). Since this integral operator is neither surjective nor compact, Theorem 4.47 can be used to show that the inclusion $\mathrm{id} : H_{\gamma_2}(X) \to H_{\gamma_1}(X)$ considered in Proposition 4.46 is neither surjective nor compact if X has a non-empty interior. In addition, the bound on its norm given in (4.44) turns out to be sharp for such X. We refer to Steinwart *et al.* (2006a) for more information.

The RKHS representation based on Mercer's theorem closely follows the presentation of Cucker and Smale (2002). This article also provides some other useful insights into the theory of RKHSs. For a proof of Mercer's theorem, we refer to Werner (1995) and Riesz and Nagy (1990).

The first part of Section 4.6 is taken almost completely from Steinwart (2001). It is not hard to see that Corollary 4.57 does *not* provide a necessary condition for universality. Indeed, if, for example, one only assumes $a_n > 0$ for all indexes n but one $n_0 \neq 0$, then k is still a universal kernel. This raises the question of how many non-vanishing coefficients are necessary for the universality. Surprisingly, this question was answered by Dahmen and Micchelli (1987) in a different context. Their result states that k is universal if and only if $a_0 > 0$ and

$$\sum_{a_{2n}>0} \frac{1}{2n} = \sum_{a_{2n+1}>0} \frac{1}{2n+1} = \infty\,.$$

Note that this condition implies that the sets $N_{\mathrm{even}} := \{2n \in \mathbb{N} : a_{2n} > 0\}$ and $N_{\mathrm{odd}} := \{2n+1 \in \mathbb{N} : a_{2n+1} > 0\}$ are infinite. Interestingly, Pinkus (2004) has recently shown that the latter characterize strictly positive definite kernels, i.e., he has shown that a kernel is strictly positive definite if and only if $a_0 > 0$ and $|N_{\mathrm{odd}}| = |N_{\mathrm{even}}| = \infty$. In particular, both results together show that not every strictly positive definite kernel is universal. An elementary proof of this latter observation can be found by combining Exercise 4.11 and Exercise 4.14. Moreover, it is interesting to note that this observation can also be deduced from Theorem 4.62. Recently, Micchelli *et al.* (2006) investigated under which conditions translation-invariant kernels and radial kernels are universal. Besides other results, they showed that *complex* translation-invariant kernels are universal if the support of the measure μ in (4.58) has a strictly positive Lebesgue measure. Using a feature map similar to that of the proof of Lemma 4.12, it is then easy to conclude that kernels represented by (4.59) are universal if $\mathrm{vol}(\mathrm{supp}\,\mu) > 0$. Moreover, Micchelli *et al.* (2006) showed that radial kernels are universal if the measure μ in (4.61) satisfies $\mathrm{supp}\,\mu \neq \{0\}$. Finally, the second part of Section 4.6, describing denseness results of H in $L_p(\mu)$, is taken from Steinwart *et al.* (2006b).

4.8 Summary

In this chapter, we gave an introduction to the mathematical theory of kernels. We first defined kernels via the existence of a feature map, but it then turned out that kernels can also be characterized by simple inequalities, namely the positive definiteness condition. Furthermore, we saw that certain representations of kernel functions lead directly to feature maps. This observation helped us to introduce several important kernels.

Although neither the feature map nor the feature space are uniquely determined for a given kernel, we saw in Section 4.2 that we can always construct a canonical feature space consisting of functions. We called this feature space the reproducing kernel Hilbert space. One of our major results was that there is a one-to-one relation between kernels and RKHSs. Moreover, we showed in Section 4.3 that many properties of kernels such as measurability, continuity, or differentiability are inherited by the functions in the RKHS.

We then determined the RKHSs of Gaussian RBF kernels and gained some insight into their structure. In particular, we were able to compare the RKHS norms for different widths and showed that these RKHSs do not contain constant functions. We further investigated properties of their associated integral operators, showing, e.g., that in many cases these operators are injective.

For continuous kernels on compact input spaces, Mercer's theorem provided a series representation in terms of the eigenvalues and functions of the associated integral operators. This series representation was then used in Section 4.5 to give another characterization of the functions contained in the corresponding RKHSs.

In Section 4.6, we then considered kernels whose RKHS H is large in the sense that H is dense in either $C(X)$ or a certain Lebesgue space of p-integrable functions. In particular, we showed that, among others, the Gaussian RBF kernels belong to this class. As we will see in later chapters, this denseness is one of the key reasons for the universal learning ability of SVMs.

4.9 Exercises

4.1. Some more kernels of Taylor type ($\star$)
Use Taylor expansions to show that the following functions can be used to construct kernels by Lemma 4.8: $x \mapsto \cosh x$, $x \mapsto \operatorname{arcoth} x^{-1}$, $x \mapsto \ln\bigl(\frac{1+x}{1-x}\bigr)$, and $x \mapsto \operatorname{arctanh} x$. What are the corresponding (maximal) domains of these kernels? Are these kernels universal?

4.2. Many standard Hilbert spaces are not RKHSs ($\star$)
Let μ be a measure on the non-empty set X. Show that $L_2(\mu)$ is an RKHS if and only if for all non-empty $A \subset X$ we have $\mu(A) > 0$.

4.3. Cauchy-Schwarz inequality ($\star\star$)
Let E be an $\mathbb{R}$-vector space and $\langle\,\cdot\,,\,\cdot\,\rangle : E \to \mathbb{R}$ be a positive, symmetric bilinear form, i.e., it satisfies

i) $\langle x, x\rangle \geq 0$
ii) $\langle x, y\rangle = \langle y, x\rangle$
iii) $\langle \alpha x + y, z\rangle = \alpha\langle x, z\rangle + \langle y, z\rangle$

for all $x, y, z \in E$, $\alpha \in \mathbb{R}$. Show the Cauchy-Schwarz inequality

$$|\langle x, y\rangle|^2 \leq \langle x, x\rangle \cdot \langle y, y\rangle\,, \qquad\qquad x, y \in E.$$

Hint: Start with $0 \leq \langle x + \alpha y, x + \alpha y\rangle$ and consider the cases $\alpha = 1$ and $\alpha = -1$ if $\langle x, x\rangle = \langle y, y\rangle = 0$. Otherwise, if, e.g., $\langle y, y\rangle \neq 0$, use $\alpha := -\frac{\langle x,y\rangle}{\langle y,y\rangle}$.

4.4. The RKHSs of restricted and normalized kernels $(\star\star\star)$
Let k be a kernel on X with RKHS H. Using Theorem 4.21, show that:

i) For $X' \subset X$, the RKHS of the restricted kernel $k_{|X' \times X'}$ is

$$H_{|X'} := \bigl\{f : X' \to \mathbb{R} \,|\, \exists \hat{f} \in H \text{ with } \hat{f}_{|X'} = f\bigr\}$$

with norm $\|f\|_{H_{|X'}} := \inf\{\|\hat{f}\|_H : \hat{f} \in H \text{ with } \hat{f}_{|X'} = f\}$.

ii) Suppose $k(x,x) > 0$ for all $x \in X$. Then the RKHS H^* of the normalized kernel k^* considered in Lemma 4.55 is

$$H^* = \bigl\{f : X \to \mathbb{R} \,|\, (x \mapsto k(x,x)f(x)) \in H\bigr\}$$

and has norm $\|f\|_{H^*} := \|(x \mapsto k(x,x)f(x))\|_H$.

iii) Determine the RKHS of the exponential kernel with the help of $H_{\gamma,\mathbb{C}^d}$.

4.5. Real part of complex kernels $(\star\star)$
Let $k : X \times X \to \mathbb{C}$ be a kernel. Show that $\mathrm{Re}\, k : X \times X \to \mathbb{R}$ is a kernel.

Hint: Show that $\mathrm{Re}\, k$ is symmetric and positive definite. For the latter, use $k(x,x') + k(x',x) = 2\mathrm{Re}\, k(x,x')$.

4.6. Injectivity of $\mathrm{id} : H \to L_p(\mu)$ $(\star\star)$
Let X be a Polish space and μ be a Borel measure with $\mathrm{supp}\, \mu = X$. Moreover, let k be a continuous kernel on X with $\|k\|_{L_p(\mu)} < \infty$ for some $p \in [1,\infty]$. Show that $\mathrm{id} : H \to L_p(\mu)$ is injective.

4.7. Properties of functions contained in the Gaussian RKHSs $(\star\star)$
For $\gamma > 0$, show the following statements:

i) Every $f \in H_\gamma(\mathbb{R}^d)$ is infinitely many times differentiable.

ii) Every $f \in H_\gamma(\mathbb{R}^d)$ is 2-integrable, and the inclusion $\mathrm{id} : H_\gamma(\mathbb{R}^d) \to L_2(\mathbb{R}^d)$ is continuous.

iii) Every $f \in H_\gamma(\mathbb{R}^d)$ is bounded, and the inclusion $\mathrm{id} : H_\gamma(\mathbb{R}^d) \to \ell_\infty(\mathbb{R}^d)$ is continuous.

Hint: For *ii)*, use that the integral operator $S_k : L_2(\mathbb{R}^d) \to H_\gamma(\mathbb{R}^d)$ is continuous. Then consider its adjoint.

4.8. Gaussian kernels and the hinge loss $(\star\star\star)$
Let P be a distribution on $X \times Y$, where $X \subset \mathbb{R}^d$ and $Y := \{-1,1\}$. Furthermore, let L_{hinge} be the hinge loss defined in Example 2.27 and $H_\gamma(X)$ be a Gaussian RKHS. Show that no minimizer $f^*_{L_{\mathrm{hinge}},\mathrm{P}}$ of the L_{hinge}-risk is contained in $H_\gamma(X)$ if for $\eta(x) := \mathrm{P}(y=1|x)$, $x \in X$, the set $\{x : \eta(x) \neq 0, 1/2, 1\}$ has a non-empty interior. Give some (geometric) examples for such distributions. Does a similar observation hold for P satisfying $\mathcal{R}^*_{L_{\mathrm{hinge}},\mathrm{P}} = 0$?

4.9. Different feature spaces of the Gaussian kernels $(\star\star)$
Compare the different feature spaces and maps of the Gaussian RBF kernels we presented in Corollary 4.40 and Lemma 4.45.

4.10. Discussion of Mercer's theorem ($\star\star\star$)
Using inadequate versions of Mercer's theorem can lead to mistakes. Consider the following two examples:

i) Sometimes a version of Mercer's theorem is presented that holds not only for continuous kernels but also for bounded and measurable kernels. For these kernels, the relation (4.53) is only stated μ^2-almost surely. Now, one might think that by modifying the eigenfunctions on a zero set one can actually obtain (4.53) for *all* $x, x' \in X$. Show that in general such a modification does not exist.

ii) Show that if the assumption $\operatorname{supp}\mu = X$ of Theorem 4.49 is dropped, (4.53) holds at least for all $x, x' \in \operatorname{supp}\mu$. Furthermore, give an example that demonstrates that in general (4.53) does not hold for all $x, x' \in X$.

Hint: For *for i)* Use $[0,1]$ equipped with the Lebesgue measure and consider the kernel k defined by $k(x,x) := 1$ for $x \in X$ and $k(x,x') = 0$ otherwise.

4.11. Strictly positive definite kernels separate all finite subsets ($\star\star$)
Let $k : X \times X \to \mathbb{R}$ be a kernel. Show that k separates all finite subsets if and only if it is strictly positive definite.

Hint: Recall from linear algebra that a symmetric matrix is (strictly) positive definite if and only if its eigenvalues are all real and (strictly) positive. Then express the equations $f(x_i) = y_i$, $i = 1, \dots, n$, $f \in H$, in terms of the Gram matrix $(k(x_j, x_i))_{i,j}$.

4.12. Universality of Fourier type kernels ($\star\star\star$)
Formulate and prove a condition for Fourier type kernels (see Lemma 4.12) that ensures universality. Then show that the kernels in Examples 4.13 and 4.14 are universal.

Hint: Use a condition similar to that of Corollary 4.57.

4.13. Existence of universal kernels ($\star\star\star\star$)
Let (X, τ) be a compact topological space. Show that the following statements are equivalent:

i) (X, τ) is metrizable, i.e., there exists a metric d on X such that the collection of the open subsets defined by d equals the topology τ.

ii) There exists a continuous kernel on X whose RKHS is dense in $C(X)$.

Hint: Use that X is metrizable if and only if $C(X)$ is separable (see, e.g., Theorem V.6.6 of Conway, 1990). Furthermore, for *i)* $\Rightarrow$ *ii)*, use a countable, dense subset of $C(X)$ to construct a universal kernel in the spirit of Lemma 4.2. For the other direction, use that every compact topological space is separable.

4.14. A kernel separating all finite but not all compact sets ($\star\star\star\star$)
Let $X := \{-1, 0\} \cup \{1/n : n \in \mathbb{N}\}$ and (e_n) be the canonical ONB of ℓ_2. Define the map $\Phi : X \to \ell_2 \oplus_2 \mathbb{R}$ by $\Phi(-1) := (\sum_{n=1}^{\infty} 2^{-n} e_n, 1)$, $\Phi(0) := (0, 1)$, and $\Phi(1/n) := (n^{-2} e_n, 1)$ for $n \in \mathbb{N}$. Then the kernel associated to the feature map Φ separates all finite sets but does not separate the compact sets $\{-1\}$ and $X \backslash \{-1\}$.