

# Chapter 7

## Kernel methods

### Chapter summary

- Kernels and representer theorem: learning with infinite-dimensional linear models can be done in a time that depends on the number of observations using a kernel function.
- Kernels on  $\mathbb{R}^d$ : such models include polynomials and classical Sobolev spaces (functions with square-integrable partial derivatives).
- Algorithms: convex optimization algorithms can be applied with theoretical guarantees and many dedicated developments to avoid the quadratic complexity of computing the kernel matrix.
- Analysis of well-specified models: When the target function is in the associated function space, learning can be done with rates that are independent of dimension.
- Analysis of misspecified models: if the target function is not in the function space, the curse of dimensionality cannot be avoided in the worst-case situations of few existing derivatives of the target function, but the methods are adaptive to any amount of intermediate smoothness.
- Sharp analysis of ridge regression: for the square loss, a more involved analysis leads to optimal rates in various situations in  $\mathbb{R}^d$ .

In this chapter, we consider positive-definite kernel methods, with only a brief account of the main results. For more details, see [Schölkopf and Smola \(2001\)](#); [Shawe-Taylor and Cristianini \(2004\)](#); [Christmann and Steinwart \(2008\)](#), and teaching slides from Jean-Philippe Vert (available from <https://jpvert.github.io/>).

## 7.1 Introduction

In this chapter, we study empirical risk minimization for linear models, that is, prediction functions  $f_\theta : \mathcal{X} \rightarrow \mathbb{R}$  which are *linear in their parameters*  $\theta$ , that is, functions of the form  $f_\theta(x) = \langle \theta, \varphi(x) \rangle_{\mathcal{H}}$ , where  $\varphi : \mathcal{X} \rightarrow \mathcal{H}$  and  $\mathcal{H}$  is a Hilbert space (essentially a Euclidean space with potentially infinite dimension)<sup>1</sup>, and  $\theta \in \mathcal{H}$ . We will often use the notation  $\langle \theta, \varphi(x) \rangle$  in this chapter instead of  $\langle \theta, \varphi(x) \rangle_{\mathcal{H}}$  when this is not ambiguous.

The key difference with Chapter 3 on least-squares estimation is that (1) we are not restricted to the square loss (although many of the same concepts will play a role, in particular, in the analysis of ridge regression), and (2), we will explicitly allow infinite-dimensional models, thus extending the dimension-free bounds from Chapter 3. The notion of *kernel function* (or simply kernel)  $k(x, y) = \langle \varphi(x), \varphi(y) \rangle_{\mathcal{H}}$  will be particularly fruitful.

**Why is this relevant?** The study of infinite-dimensional linear methods is important for several reasons:

- Understanding linear models in finite but very large input dimensions requires tools from infinite-dimensional analysis.
- Kernel methods lead to simple and stable algorithms, with theoretical guarantees and adaptivity to the smoothness of the target function (as opposed to local averaging techniques). They can be applied in high dimensions, with good practical performance (note that for supervised learning problems with many observations in domains such as computer vision and natural language processing, they do not achieve the state of the art anymore, which is achieved by neural networks presented in Chapter 9).
- They can be easily applied when input observations are not vectors (see Section 7.3.4).
- They are helpful to understand other models such as neural networks (see Chapter 9).



The type of kernel we consider here is different from the ones in Chapter 6. The ones here are “positive definite,” the ones from Chapter 6 are “non-negative”. See more details in <https://francisbach.com/cursed-kernels/>.

## 7.2 Representer theorem

Dealing with infinite-dimensional models initially seems impossible because algorithms cannot be run in infinite dimensions. In this section, we show how the kernel function plays a crucial role in achieving lower-dimensional algorithms.

---

<sup>1</sup>More precisely, this is a vector space endowed with a inner product and which is complete for the associated normed space topology. See [https://en.wikipedia.org/wiki/Hilbert\\_space](https://en.wikipedia.org/wiki/Hilbert_space) for more details.

As a motivation, we consider the optimization problem coming from machine learning with linear models, with data  $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$ ,  $i = 1, \dots, n$ :

$$\min_{\theta \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n \ell(y_i, \langle \varphi(x_i), \theta \rangle) + \frac{\lambda}{2} \|\theta\|^2, \quad (7.1)$$

assuming the loss function  $\ell$  is already from  $\mathcal{Y} \times \mathbb{R} \rightarrow \mathbb{R}$  and not from  $\mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}$  (e.g., hinge loss, logistic loss or least-squares, see Chapter 4).

The key property of the objective function in Eq. (7.1) is that it accesses the input observations  $x_1, \dots, x_n \in \mathcal{X}$ , only through dot-products  $\langle \theta, \varphi(x_i) \rangle$ ,  $i = 1, \dots, n$ , and that we penalize using the Hilbertian norm  $\|\theta\|$ . The following proposition is crucial and has an elementary proof, due to [Kimeldorf and Wahba \(1971\)](#) for Corollary 7.1, and to [Schölkopf et al. \(2001\)](#) for the general form presented below.

**Proposition 7.1 (Representer theorem)** *Consider a feature map  $\varphi : \mathcal{X} \rightarrow \mathcal{H}$ . Let  $(x_1, \dots, x_n) \in \mathcal{X}^n$ , and assume that the functional  $\Psi : \mathbb{R}^{n+1} \rightarrow \mathbb{R}$  is strictly increasing with respect to the last variable, then the infimum of*

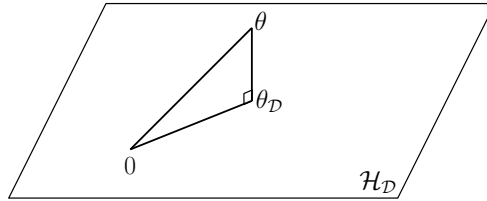
$$\Psi(\langle \theta, \varphi(x_1) \rangle, \dots, \langle \theta, \varphi(x_n) \rangle, \|\theta\|^2)$$

*can be obtained by restricting to a vector  $\theta$  of the form*

$$\theta = \sum_{i=1}^n \alpha_i \varphi(x_i),$$

*with  $\alpha \in \mathbb{R}^n$ .*

**Proof** Let  $\theta \in \mathcal{H}$ , and  $\mathcal{H}_{\mathcal{D}} = \{ \sum_{i=1}^n \alpha_i \varphi(x_i), \alpha \in \mathbb{R}^n \} \subset \mathcal{H}$ , the linear span of the observed feature vectors. Let  $\theta_{\mathcal{D}} \in \mathcal{H}_{\mathcal{D}}$  and  $\theta_{\perp} \in \mathcal{H}_{\mathcal{D}}^{\perp}$  be such that  $\theta = \theta_{\mathcal{D}} + \theta_{\perp}$ , a decomposition which is using the Hilbertian structure of  $\mathcal{H}$ . Then  $\forall i \in \{1, \dots, n\}$ ,  $\langle \theta, \varphi(x_i) \rangle = \langle \theta_{\mathcal{D}}, \varphi(x_i) \rangle + \langle \theta_{\perp}, \varphi(x_i) \rangle$  with  $\langle \theta_{\perp}, \varphi(x_i) \rangle = 0$ , by definition of the orthogonal.



From the Pythagorean theorem, we get:  $\|\theta\|^2 = \|\theta_{\mathcal{D}}\|^2 + \|\theta_{\perp}\|^2$ . Therefore, we have:

$$\begin{aligned} \Psi(\langle \theta, \varphi(x_1) \rangle, \dots, \langle \theta, \varphi(x_n) \rangle, \|\theta\|^2) &= \Psi(\langle \theta_{\mathcal{D}}, \varphi(x_1) \rangle, \dots, \langle \theta_{\mathcal{D}}, \varphi(x_n) \rangle, \|\theta_{\mathcal{D}}\|^2 + \|\theta_{\perp}\|^2) \\ &\geq \Psi(\langle \theta_{\mathcal{D}}, \varphi(x_1) \rangle, \dots, \langle \theta_{\mathcal{D}}, \varphi(x_n) \rangle, \|\theta_{\mathcal{D}}\|^2), \end{aligned}$$

with equality if and only if  $\theta_{\perp} = 0$  (since  $\Psi$  is strictly increasing with respect to the last variable). Thus

$$\inf_{\theta \in \mathcal{H}} \Psi(\langle \theta, \varphi(x_1) \rangle, \dots, \langle \theta, \varphi(x_n) \rangle, \|\theta\|^2) = \inf_{\theta \in \mathcal{H}_{\mathcal{D}}} \Psi(\langle \theta, \varphi(x_1) \rangle, \dots, \langle \theta, \varphi(x_n) \rangle, \|\theta\|^2),$$

which is exactly the desired result. ■

This implies that the minimizer of Eq. (7.1) can be found among the vectors of the form  $\theta = \sum_{i=1}^n \alpha_i \varphi(x_i)$ :

**Corollary 7.1 (Representer theorem for supervised learning)** *For  $\lambda > 0$ , the infimum of  $\frac{1}{n} \sum_{i=1}^n \ell(y_i, \langle \theta, \varphi(x_i) \rangle) + \frac{\lambda}{2} \|\theta\|^2$  can be obtained by restricting to a vector  $\theta$  of the form  $\theta = \sum_{i=1}^n \alpha_i \varphi(x_i)$ , with  $\alpha \in \mathbb{R}^n$ .*

It is important to note that there is no assumption on the loss function  $\ell$ . In particular, no convexity is assumed. This is to be contrasted to the use of duality in Section 7.4.4, where convexity will play a major role and similar  $\alpha$ 's will be defined (but with some notable differences).

Given Corollary 7.1, we can reformulate the learning problem. We will need the *kernel function*  $k$ , which is the dot product between feature vectors:

$$k(x, x') = \langle \varphi(x), \varphi(x') \rangle.$$

We then have:

$$\forall j \in \{1, \dots, n\}, \langle \theta, \varphi(x_j) \rangle = \sum_{i=1}^n \alpha_i k(x_i, x_j) = (K\alpha)_j,$$

where  $K \in \mathbb{R}^{n \times n}$  is the *kernel matrix*, such that  $K_{ij} = \langle \varphi(x_i), \varphi(x_j) \rangle = k(x_i, x_j)$ , and

$$\|\theta\|^2 = \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j \langle \varphi(x_i), \varphi(x_j) \rangle = \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j K_{ij} = \alpha^\top K \alpha.$$

We can then write:

$$\inf_{\theta \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n \ell(y_i, \langle \theta, \varphi(x_i) \rangle) + \frac{\lambda}{2} \|\theta\|^2 = \inf_{\alpha \in \mathbb{R}^n} \frac{1}{n} \sum_{i=1}^n \ell(y_i, (K\alpha)_i) + \frac{\lambda}{2} \alpha^\top K \alpha. \quad (7.2)$$

For a test point  $x \in \mathcal{X}$ , we have  $f(x) = \sum_{i=1}^n \alpha_i k(x, x_i)$ .

Thus, the input observations are summarized in the kernel matrix and the kernel function, regardless of the dimension of  $\mathcal{H}$ . Moreover, explicitly computing the feature vector  $\varphi(x)$  is never needed! This is the *kernel trick*, which allows to:

- replace the search space  $\mathcal{H}$  by  $\mathbb{R}^n$ ; this is interesting computationally when the dimension of  $\mathcal{H}$  is very large (see more details in Section 7.4),
- separate the representation problem (design of kernels on a set  $\mathcal{X}$ ) and the design of algorithms and their analysis (which only use the kernel matrix  $K$ ); this is interesting because a wide range of kernels can be defined for many data types (see more details in Section 7.3).

**Minimum norm interpolation.** The representer theorem can be extended to an interpolating estimator with essentially the same proof (see proposition below).

**Proposition 7.2** *Given  $x_1, \dots, x_n \in \mathcal{X}$ , and  $y \in \mathbb{R}^n$  such that there exists at least one  $\theta \in \mathcal{H}$  such that  $y_i = \langle \theta, \varphi(x_i) \rangle$  for all  $i \in \{1, \dots, n\}$ , then among all these  $\theta \in \mathcal{H}$  that interpolate the data, the one of minimum norm can be expressed as  $\theta = \sum_{i=1}^n \alpha_i \varphi(x_i)$ , with  $\alpha \in \mathbb{R}^n$  is such that  $y = K\alpha$  (this system must then have a solution).*

## 7.3 Kernels

In the section above, we have introduced the kernel function  $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$  as obtained from a dot product  $k(x, x') = \langle \varphi(x), \varphi(x') \rangle$ . The associated kernel matrix is then a matrix of dot-products (often called a “Gram matrix”) and is thus symmetric positive semi-definite (see proof below), that is, all of its eigenvalues are non-negative, or, equivalently,  $\forall \alpha \in \mathbb{R}^n, \alpha^\top K \alpha \geq 0$ . It turns out that this simple property is enough to impose the existence of a feature function.

$\triangle$  If  $\mathcal{H} = \mathbb{R}^d$ , and  $\Phi \in \mathbb{R}^{n \times d}$  is the matrix of features (design matrix in the context of regression) with  $i$ -th row composed of  $\varphi(x_i)$ , then  $K = \Phi \Phi^\top \in \mathbb{R}^{n \times n}$  is the kernel matrix, while  $\frac{1}{n} \Phi^\top \Phi \in \mathbb{R}^{d \times d}$  is the empirical covariance matrix.

**Definition 7.1 (Positive definite kernels)** *A function  $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$  is a positive definite kernel if and only if all kernel matrices are symmetric positive semi-definite.*

The following important proposition dates back to [Aronszajn \(1950\)](#) and comes with an elegant constructive proof. Note the total absence of assumptions on the set  $\mathcal{X}$ .

**Proposition 7.3 ([Aronszajn, 1950](#))**  *$k$  is a positive definite kernel if and only if there exists a Hilbert space  $\mathcal{H}$ , and a function  $\varphi : \mathcal{X} \rightarrow \mathcal{H}$  such that for all  $x, x' \in \mathcal{X}$ ,  $k(x, x') = \langle \varphi(x), \varphi(x') \rangle_{\mathcal{H}}$ .*

**Partial proof** We first assume that  $k(x, x') = \langle \varphi(x), \varphi(x') \rangle_{\mathcal{H}}$ . Then for any  $\alpha \in \mathbb{R}^n$ , and points  $x_1, \dots, x_n \in \mathcal{X}$ , we have:

$$\alpha^\top K \alpha = \sum_{i,j=1}^n \alpha_i \alpha_j \langle \varphi(x_i), \varphi(x_j) \rangle_{\mathcal{H}} = \left\| \sum_{i=1}^n \alpha_i \varphi(x_i) \right\|_{\mathcal{H}}^2 \geq 0.$$

Thus,  $k$  is a positive definite kernel.

For the other direction, we consider a positive-definite kernel, and we will construct a space of functions explicitly from  $\mathcal{X}$  to  $\mathbb{R}$  with a dot-product. We define the set  $\mathcal{H}' \subset \mathbb{R}^{\mathcal{X}}$  as the set of linear combinations of kernel functions  $\sum_{i=1}^n \alpha_i k(\cdot, x_i)$  for any integer  $n$ , any set of  $n$  points, and any  $\alpha \in \mathbb{R}^n$ . This is a vector space on which we can define a dot-product through

$$\left\langle \sum_{i=1}^n \alpha_i k(\cdot, x_i), \sum_{j=1}^m \beta_j k(\cdot, x'_j) \right\rangle = \sum_{i=1}^n \sum_{j=1}^m \alpha_i \beta_j k(x_i, x'_j). \quad (7.3)$$

We first check that this is a well-defined function on  $\mathcal{H}' \times \mathcal{H}'$ , that is, the value does not depend on the chosen representation as a linear combination of kernel functions. Indeed,

if we denote  $f = \sum_{i=1}^n \alpha_i k(\cdot, x_i)$ , then the dot-product is equal to  $\sum_{j=1}^m \beta_j f(x'_j)$  and depends only on the values of  $f$ , and not on its representation (and similarly for the function on the right of the dot-product).

This dot-product is bi-linear and always non-negative when applied to the same function (that is, in Eq. (7.3) above, when  $\alpha = \beta$  and the points  $(x_i)$  and  $(x'_j)$  are the same, we get a positive number because of the positivity of the kernel  $k$ ). To obtain a dot-product, we only need to show that  $\langle f, f \rangle = 0$  implies  $f = 0$ . This can be shown using Cauchy-Schwarz inequality,<sup>2</sup> leading to for any  $x \in \mathcal{X}$ , to the sequence of bounds  $f(x)^2 = \langle f, k(\cdot, x) \rangle^2 \leq \langle f, f \rangle \langle k(\cdot, x), k(\cdot, x) \rangle = \langle f, f \rangle k(x, x)$ , leading to  $f = 0$  as soon as  $\langle f, f \rangle = 0$ .

Moreover, this dot product satisfies the two properties for any  $f \in \mathcal{H}'$ ,  $x, x' \in \mathcal{X}$ :

$$\langle k(\cdot, x), f \rangle = f(x) \text{ and } \langle k(\cdot, x), k(\cdot, x') \rangle = k(x, x').$$

These are called “reproducing properties” and correspond to an explicit construction of the feature map  $\varphi(x) = k(\cdot, x)$ .

The space  $\mathcal{H}'$  is called “pre-Hilbertian” because it is not complete.<sup>3</sup> It can be “completed” into a Hilbert space  $\mathcal{H}$  with the same reproducing property. See Aronszajn (1950); Berlinet and Thomas-Agnan (2004) for more details. ■

We can make the following observations:


- $\mathcal{H}$  is called the “feature space,” and  $\varphi$  the “feature map,” that goes from the “input space”  $\mathcal{X}$  to the feature space  $\mathcal{H}$ .
- No assumption is needed about the input space  $\mathcal{X}$ , and no regularity assumption is needed for  $k$ . Up to isomorphisms, the feature map and space happen to be unique. The particular space of functions we built is called the *reproducing kernel Hilbert space (RKHS)* associated with  $k$ , for which  $\varphi(x) = k(\cdot, x)$ .
- A classical intuitive interpretation of the identity  $\langle k(\cdot, x), f \rangle = f(x)$  is that the function evaluation is the dot-product with a function (this is, in fact, another characterization, see the exercise below). This implies that not all Hilbert spaces of real-valued functions on  $\mathcal{X}$  are RKHSs. Indeed, for example, if  $L_2(\mathbb{R}^d)$  was an RKHS, this would mean that there exists a function  $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$  such that  $\int_{\mathbb{R}^d} k(x, x') f(x') dx' = f(x)$ . In other words,  $k(x, x') dx'$  would be a Dirac measure at  $x$ , which is impossible (as Dirac measures have no density with respect to the Lebesgue measure). Thus  $L_2(\mathbb{R}^d)$  is a Hilbert space that is too large to be an RKHS. We will see below that smaller spaces of functions, with square-integrable derivatives of sufficiently high order, will be RKHSs.
- Given a positive-definite kernel  $k$ , we can thus associate it to some feature map  $\varphi$  such that  $k(x, y) = \langle \varphi(x), \varphi(y) \rangle_{\mathcal{H}}$ , but also to a *space of functions on  $\mathcal{X}$  with a given*

---

<sup>2</sup>Cauchy-Schwarz inequality applies to bilinear forms that are symmetric positive semi-definite, but may not be positive definite, that is, such that  $\langle f, f \rangle = 0$  may not imply  $f = 0$ .

<sup>3</sup>See [https://en.wikipedia.org/wiki/Complete\\_metric\\_space](https://en.wikipedia.org/wiki/Complete_metric_space) for definitions.

norm, either directly through the RKHS above or by looking at all functions  $f_\theta$  of the form  $f_\theta(x) = \langle \theta, \varphi(x) \rangle_{\mathcal{H}}$ , with a regularization term  $\|\theta\|_{\mathcal{H}}^2$ . These two views are equivalent.

 From now on, we will denote elements of the Hilbert space  $\mathcal{H}$  through the notation  $f \in \mathcal{H}$  to highlight the fact that we are considering a space of functions from  $\mathcal{X}$  to  $\mathbb{R}$ , except for optimization algorithms in Section 7.4, where we will use the notation  $\langle \theta, \varphi(x) \rangle_{\mathcal{H}}$  instead of  $f(x)$ .

**Exercise 7.1** ( $\blacklozenge$ ) *Let  $\mathcal{H}$  be a Hilbert space of real-valued functions on  $\mathcal{X}$ , such that for any  $x \in \mathcal{X}$ , the linear form  $x \mapsto f(x)$  is bounded (that is,  $\sup_{f \in \mathcal{H}, \|f\|_{\mathcal{H}} \leq 1} |f(x)|$  is finite). Using the Riesz representation theorem, show that this is a reproducing kernel Hilbert space.*

**Kernel calculus.** The set of positive definite kernels on a set  $\mathcal{X}$  is a cone, that is, it is closed under addition and multiplication by a positive constant. In other words, if  $k_1$  and  $k_2$  are two positive definite kernels and  $\lambda_1, \lambda_2 > 0$ , then so is  $\lambda_1 k_1 + \lambda_2 k_2$ . A simple proof follows from considering two feature maps  $\varphi_1 : \mathcal{X} \rightarrow \mathcal{H}_1$  and  $\varphi_2 : \mathcal{X} \rightarrow \mathcal{H}_2$ , and noticing that  $x \mapsto \begin{pmatrix} \lambda_1^{1/2} \varphi_1(x) \\ \lambda_2^{1/2} \varphi_2(x) \end{pmatrix}$  is a feature map for  $\lambda_1 k_1 + \lambda_2 k_2$ .

Moreover, positive definite kernels are closed under pointwise multiplication, that is, if  $k_1$  and  $k_2$  are positive definite kernels on the set  $\mathcal{X}$ , so is,  $(x, x') \mapsto k_1(x, x')k_2(x, x')$ . For finite-dimensional kernels, where we can consider feature spaces  $\mathcal{H}_1 = \mathbb{R}^{d_1}$  and  $\mathcal{H}_2 = \mathbb{R}^{d_2}$ , the product kernel is associated with a feature space of dimension  $d_1 d_2$  and the feature map  $x \mapsto [\varphi_1(x)_{i_1} \varphi_2(x)_{i_2}]_{i_1 \in \{1, \dots, d_1\}, i_2 \in \{1, \dots, d_2\}}$ . The general proof is left as an exercise.

**Exercise 7.2** *Show that if  $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$  is a positive definite kernel, so is the function  $(x, x') \mapsto e^{k(x, x')}$ .*

**Kernels = features and functions.** A positive-definite kernel thus defines a feature map and a space of functions. Sometimes, the feature map is easy to find, and sometimes it is not. In the next section, we will look at the main examples and describe the associated spaces of functions (and the corresponding norms).

We now look at different ways of building the kernels, by starting first from the feature vector (e.g., linear kernels), from the kernel and explicit feature map (polynomial kernel), from the norm (translation-invariant kernel on  $[0, 1]$ ), or from the kernel without explicit features (translation-invariant kernel on  $\mathbb{R}^d$ ).

### 7.3.1 Linear and polynomial kernels

We start with the most obvious kernels on  $\mathcal{X} = \mathbb{R}^d$ , for which feature maps are easily found.

**Linear kernel.** We define  $k(x, x') = x^\top x'$ . It corresponds to a function space composed of linear functions  $f_\theta(x) = \theta^\top x$ , with an  $\ell_2$ -penalty  $\|\theta\|_2^2$ . The kernel trick can be useful when the input data have huge dimension  $d$  but are quite sparse (many zeros), such as in text processing, so that the dot-product  $x^\top x'$  can be computed in time  $o(d)$ .

**Polynomial kernel.** For  $s$  a positive integer, the kernel  $k(x, x') = (x^\top x')^s$  is positive-definite as a integer power of a kernel and can be explicitly expanded as (with the binomial theorem<sup>4</sup>):

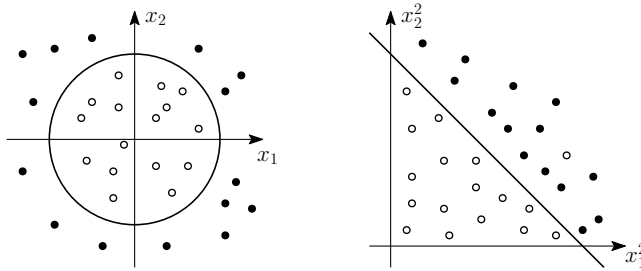
$$k(x, x') = \left( \sum_{i=1}^d x_i x'_i \right)^s = \sum_{\alpha_1 + \dots + \alpha_d = s} \binom{s}{\alpha_1, \dots, \alpha_d} \underbrace{(x_1 x'_1)^{\alpha_1} \dots (x_d x'_d)^{\alpha_d}}_{(x_1^{\alpha_1} \dots x_d^{\alpha_d})(x'_1)^{\alpha_1} \dots (x'_d)^{\alpha_d}},$$

where the sum is over all non-negative integer vectors  $(\alpha_1, \dots, \alpha_d)$  that sum to  $s$ . We have an explicit feature map:  $\varphi(x) = \left( \binom{s}{\alpha_1, \dots, \alpha_d} \right)^{\frac{1}{2}} x_1^{\alpha_1} \dots x_d^{\alpha_d}$ , and the set of functions is the set of degree- $s$  homogeneous<sup>5</sup> polynomials on  $\mathbb{R}^d$ , which has dimension  $\binom{d+s-1}{s}$ .

When  $d$  and  $s$  grow, the feature space dimension grows as  $d^s$ , an explicit representation is not desirable, and the kernel trick can be advantageous. Note, however, that the associated norm (which penalizes coefficients of the polynomials) is hard to interpret (as a small change in a single high-order coefficient can lead to significant changes).

**Exercise 7.3** Show that the kernel  $k(x, x') = (1 + x^\top x')^s$  corresponds to the set of all monomials  $x_1^{\alpha_1} \dots x_d^{\alpha_d}$  such that  $\alpha_1 + \dots + \alpha_d \leq s$ . Show that the dimension of the feature space is  $\binom{d+s}{s}$ .

As an illustration, when using a polynomial kernel of degree 2, the set of functions that are linear in the feature map is, therefore, the set of quadratic functions. Thus, in a binary classification problem where data can be separated by an ellipsoid, this can be obtained by linear separation in the feature space. See the illustration below.



<sup>4</sup>See [https://en.wikipedia.org/wiki/Binomial\\_theorem](https://en.wikipedia.org/wiki/Binomial_theorem).

<sup>5</sup>A function  $f : \mathbb{R}^d \rightarrow \mathbb{R}$  is said homogeneous if there exists  $s \in \mathbb{R}_+$  such that for all  $x \in \mathbb{R}^d$ , and  $\lambda \in \mathbb{R}_+$ ,  $f(\lambda x) = \lambda^s f(x)$ .



### 7.3.2 Translation-invariant kernels on $[0, 1]$

We consider  $\mathcal{X} = [0, 1]$ , and kernels of the form  $k(x, x') = q(x - x')$  with a function  $q : [0, 1] \rightarrow \mathbb{R}$ , which is 1-periodic. We will show how they emerge from penalties on the Fourier coefficients of functions, which we now quickly review.<sup>6</sup>

**Fourier series.** We will consider complex-valued functions and use complex exponentials, but all developments could be carried out with cosines and sines. Fourier series corresponds simply to an orthonormal decomposition of square-integrable functions on  $[0, 1]$ , which are naturally extended to 1-periodic functions on  $\mathbb{R}$ . More precisely, the set of functions  $x \mapsto e^{2im\pi x}$ , for  $m \in \mathbb{Z}$  is an orthonormal basis of  $L_2([0, 1])$ . Therefore, any squared integrable functions which are 1-periodic can be expanded in this orthonormal basis, that is,  $q(x) = \sum_{m \in \mathbb{Z}} e^{2im\pi x} \hat{q}_m$ , with  $\hat{q}_m = \int_0^1 q(x) e^{-2im\pi x} dx \in \mathbb{C}$ , for  $m \in \mathbb{Z}$ , obtained by projection  $q$  to the element of the basis. The function  $q$  is real-valued if and only if for all  $m \in \mathbb{Z}$ ,  $\hat{q}_{-m} = \hat{q}_m^*$  (the complex conjugate of  $\hat{q}_m$ ). We will also need Parseval's identity, which is exactly the Pythagorean theorem in the orthonormal basis, that is,  $\int_0^1 |q(x)|^2 dx = \sum_{m \in \mathbb{Z}} |\hat{q}_m|^2$ .

**Translation-invariant kernels.** When presenting translation-invariant kernels, we can choose to start from the kernel or the associated squared norm. In this section, we start from the squared norm, while in the next one, we start from the kernel.

Given a 1-periodic function  $f$  decomposed into its Fourier series as

$$f(x) = \sum_{m \in \mathbb{Z}} e^{2im\pi x} \hat{f}_m,$$

we consider the penalty

$$\sum_{m \in \mathbb{Z}} c_m |\hat{f}_m|^2,$$

with  $c \in \mathbb{R}_+^{\mathbb{Z}}$ ; this penalty can be interpreted through a feature map and the  $\ell_2$ -norm on  $\mathbb{C}^{\mathbb{Z}}$  (the space of functions from  $\mathbb{Z}$  to  $\mathbb{C}$ ). Indeed, it corresponds to the feature vector  $\varphi(x)_m = e^{2im\pi x} / \sqrt{c_m}$ , and  $\theta \in \mathbb{C}^{\mathbb{Z}}$ , such that  $\theta_m = \hat{f}_m \sqrt{c_m}$  (we can easily consider complex-valued features instead of real-valued features if Hermitian dot-products are considered), so that  $f(x) = \langle \theta, \varphi(x) \rangle$  and  $\sum_{m \in \mathbb{Z}} |\theta_m|^2$  is equal to the norm  $\sum_{m \in \mathbb{Z}} c_m |\hat{f}_m|^2$ .

Thus the associated kernel is

$$k(x, x') = \sum_{m \in \mathbb{Z}} \varphi(x)_m \varphi(x')_m^* = \sum_{m \in \mathbb{Z}} \frac{e^{2im\pi x}}{\sqrt{c_m}} \frac{e^{-2im\pi x'}}{\sqrt{c_m}} = \sum_{m \in \mathbb{Z}} \frac{1}{c_m} e^{2im\pi(x-x')},$$

which is thus of the form  $q(x - x')$  for a 1-periodic function  $q$ .

What we showed above is that any penalty of the form  $\sum_{m \in \mathbb{Z}} c_m |\hat{f}_m|^2$  defines a squared RKHS norm as soon as  $c_m$  is strictly positive for all  $m \in \mathbb{Z}$ , and  $\sum_{m \in \mathbb{Z}} \frac{1}{c_m}$  is

---

<sup>6</sup>See [https://en.wikipedia.org/wiki/Fourier\\_series](https://en.wikipedia.org/wiki/Fourier_series) for more details.

finite. The kernel function is then of the form  $k(x, y) = q(x - y)$  with  $q$  being 1-periodic, and such that the Fourier series has non-negative real values  $\hat{q}_m = c_m^{-1}$ . In the other direction, all such kernels are positive-definite (see an extension to  $\mathbb{R}^d$  in Section 7.3.3).

**Penalization of derivatives.** For certain penalties based on  $(c_m)_m$ , there is a natural link with penalties on derivatives, as if  $f$  is  $s$ -times differentiable with squared integrable derivative, we have, by differentiating the Fourier series representation above,  $f^{(s)}(x) = \sum_{m \in \mathbb{Z}} (2im\pi)^s e^{2im\pi x} \hat{f}_m$ , and thus, from Parseval's theorem (which states that the squared  $L_2$ -norm of a function is equal to the sum of the square modulus of its Fourier coefficients):

$$\int_0^1 |f^{(s)}(x)|^2 dx = (2\pi)^{2s} \sum_{m \in \mathbb{Z}} m^{2s} |\hat{f}_m|^2.$$

In this chapter, we will consider penalizing such derivatives, leading to Sobolev spaces on  $[0, 1]$  (see extensions in sections below). The following examples are often considered:

- **Bernoulli polynomials:** we can consider  $c_0 = 1$  and  $c_m = |m|^{2s}$  for  $m \neq 0$ , for which the associated norm is  $\|f\|_{\mathcal{H}}^2 = \frac{1}{(2\pi)^{2s}} \int_0^1 |f^{(s)}(x)|^2 dx + \left( \int_0^1 f(x) dx \right)^2$ . The corresponding kernel  $k(x, x')$  can then be written as

$$k(x, x') = \sum_{m \in \mathbb{Z}} c_m^{-1} e^{2im\pi(x-x')} = 1 + \sum_{m \geq 1} \frac{2 \cos[2\pi m(x-x')]}{m^{2s}}.$$

In order to have an expression for  $q$  in closed form we notice that if we define  $\{x\} = x - \lfloor x \rfloor \in [0, 1)$  the fractional part of  $x$ , the function  $x \mapsto \{x\}$  has (by integration by part) an  $m$ -th Fourier coefficient equal to  $\int_0^1 e^{-2im\pi x} x dx = \frac{i}{2m\pi}$ . Similarly, the  $s$ -th power of  $\{x\}$  has an  $m$ -th Fourier coefficient which is an order  $s$  polynomial in  $m^{-1}$ . This implies that  $k(x, x')$  has to be an order  $s$  polynomial in  $\{x - x'\}$ , which happens to be related to classical polynomials, as we now show.

For  $s = 1$ , we have  $k(x, x) = 1 + 2 \sum_{m \geq 1} m^{-2} = 1 + \pi^2/3$ ; moreover by using the Fourier series expansion  $\{t\} = \frac{1}{2} - \frac{1}{2\pi} \sum_{m \geq 1} \frac{2 \sin[2\pi mt]}{m}$ , and integrating, we get

$$k(x, x') = 2\pi^2 \{x - x'\}^2 - 2\pi^2 \{x - x'\} + \pi^2/3 + 1 = q(x - x'),$$

with  $q$  plotted in Figure 7.1. For  $s \geq 1$ , we have the closed-form expression  $k(x, x') = 1 + (-1)^{s-1} \frac{(2\pi)^{2s}}{(2s)!} B_{2s}(\{x - x'\})$ , where  $B_{2s}$  the  $(2s)$ -th Bernoulli polynomial,<sup>7</sup> from which we can “check” the computation above since  $B_2(t) = t^2 - t + 1/6$ .

**Exercise 7.4** Show that for  $s = 2$ , we have for all  $x, x' \in [0, 1]$ ,  $k(x, x') = q(x - x')$  with  $q(t) = 1 - \frac{(2\pi)^4}{24} (\{t\}^4 - 2\{t\}^3 + \{t\}^2 - \frac{1}{30})$ .

<sup>7</sup>See [https://en.wikipedia.org/wiki/Bernoulli\\_polynomials](https://en.wikipedia.org/wiki/Bernoulli_polynomials).

- **Periodic exponential kernel:** we can consider  $c_m = 1 + \alpha^2|m|^2$ , for which we also have a closed-form formula, with the penalty  $\|f\|_{\mathcal{H}}^2 = \frac{\alpha^2}{(2\pi)^2} \int_0^1 |f'(x)|^2 dx + \int_0^1 |f(x)|^2 dx$ .

**Exercise 7.5** (◆◆◆) Show that we have  $k(x, x') = \sum_{m \in \mathbb{Z}} \frac{e^{2im\pi(x-x')}}{1 + \alpha^2|m|^2} = q(x - x')$  for  $q(t) = \frac{\pi}{\alpha} \frac{\cosh \frac{\pi}{\alpha}(1-2|\{t+1/2\}-1/2|)}{\sinh \frac{\pi}{\alpha}}$ . Hint: use the Cauchy residue formula.<sup>8</sup>

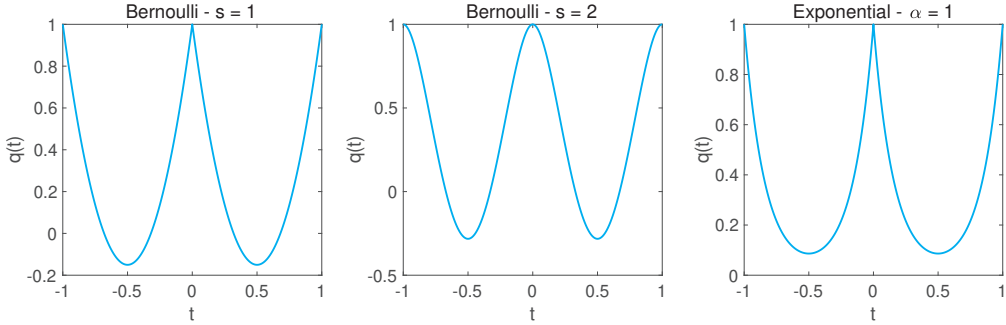


Figure 7.1: Translation-invariant kernels on  $[0, 1]$ , of the form  $k(x, x') = q(x - x')$ , with  $q$  1-periodic, for the kernels based on Bernoulli polynomials, and the periodic exponential kernel. Kernels are normalized so that  $k(x, x) = 1$ .

These kernels are mainly used for their simplicity and explicit feature map, which are simpler than the kernels described below which are most used in practice (with similar links with Sobolev spaces). Note also that for the uniform distribution on  $[0, 1]$ , the Fourier basis will be an orthogonal eigenbasis of the covariance operator with eigenvalues  $c_m^{-1}$  (see Section 7.6.6).

We saw that for the kernel  $q(x - x')$  with Fourier series  $\hat{q}_m$  for  $q$ , the associated norm is  $\sum_{m \in \mathbb{Z}} \frac{|\hat{f}_m|^2}{\hat{q}_m}$ . We now extend this to Fourier transforms (instead of Fourier series).

### 7.3.3 Translation-invariant kernels on $\mathbb{R}^d$

We consider  $\mathcal{X} = \mathbb{R}^d$ , and a kernel of the form  $k(x, x') = q(x - x')$  with a function  $q : \mathbb{R}^d \rightarrow \mathbb{R}$ , which we refer to as translation-invariant as it is invariant by the addition of the same constant to both arguments. We start with a short review of Fourier transforms.<sup>9</sup>

**Fourier transforms.** The Fourier transform  $\hat{f} : \mathbb{R}^d \rightarrow \mathbb{C}$  of an integrable function  $f : \mathbb{R}^d \rightarrow \mathbb{C}$  can be defined through

$$\hat{f}(\omega) = \int_{\mathbb{R}^d} f(x) e^{-i\omega^\top x} dx,$$

<sup>8</sup>See <https://francisbach.com/cauchy-residue-formula/>.

<sup>9</sup>See [https://en.wikipedia.org/wiki/Fourier\\_transform](https://en.wikipedia.org/wiki/Fourier_transform) for more details.

which is then a continuous function of  $\omega$ . It can naturally be extended to an operator on all square-integrable functions, and under appropriate conditions on  $f$ , we can recover  $f$  from its Fourier transform, that is,

$$f(x) = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \hat{f}(\omega) e^{i\omega^\top x} d\omega.$$

Moreover, Parseval's identity leads to  $\int_{\mathbb{R}^d} |f(x)|^2 dx = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} |\hat{f}(\omega)|^2 d\omega$ .

**Translation-invariant kernels.** The following proposition gives conditions under which we obtain a positive definite kernel.

**Proposition 7.4 (Bochner's theorem)** *The kernel  $k$  is positive definite if and only if  $q$  is the Fourier transform of a non-negative Borel measure. Consequently, if  $q$  is integrable (for the Lebesgue measure) and its Fourier transform only has non-negative real values, then  $k$  is positive definite.*

**Partial proof** We only give the proof of the consequence, which is the only one we need. Since  $q$  is integrable,  $\hat{q}(\omega) = \int_{\mathbb{R}^d} e^{-i\omega^\top x} q(x) dx$  is defined on  $\mathbb{R}^d$  and continuous, and we have through the inverse Fourier transform formula:

$$q(x - x') = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \hat{q}(\omega) e^{i(x-x')^\top \omega} d\omega.$$

Let  $x_1, \dots, x_n \in \mathbb{R}^d$ , let  $\alpha_1, \dots, \alpha_n \in \mathbb{R}$ . We have:

$$\begin{aligned} \sum_{s,j=1}^n \alpha_s \alpha_j k(x_s, x_j) &= \sum_{s,j=1}^n \alpha_s \alpha_j q(x_s - x_j) = \frac{1}{(2\pi)^d} \sum_{s,j=1}^n \alpha_s \alpha_j \int_{\mathbb{R}^d} e^{i\omega^\top (x_s - x_j)} \hat{q}(\omega) d\omega \\ &= \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \left( \sum_{s,j=1}^n \alpha_s \alpha_j e^{i\omega^\top x_s} (e^{i\omega^\top x_j})^* \right) \hat{q}(\omega) d\omega \\ &= \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \left| \sum_{s=1}^n \alpha_s e^{i\omega^\top x_s} \right|^2 \hat{q}(\omega) d\omega \geq 0, \end{aligned}$$

which shows the positive-definiteness. See [Reed and Simon \(1978\)](#) and [Varadhan \(2001, Theorem 2.7\)](#) for a proof of the other direction. ■

**Construction of the associated norm.** We give an intuitive (non-rigorous) reasoning: if  $q$  is integrable, then  $\hat{q}(\omega)$  exists and, we have an explicit representation as

$$k(x, x') = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \sqrt{\hat{q}(\omega)} e^{i\omega^\top x} (\sqrt{\hat{q}(\omega)} e^{i\omega^\top x'})^* d\omega = \int_{\mathbb{R}^d} \varphi(x)_\omega \varphi(x')_\omega^* d\omega,$$

which is of the form  $\langle \varphi(x), \varphi(y) \rangle$ , with  $\varphi(x)_\omega = \frac{1}{(2\pi)^{d/2}} \sqrt{\hat{q}(\omega)} e^{i\omega^\top x}$  (it is non rigorous because the index  $\omega$  belongs to  $\mathbb{R}^d$ , which is not countable). If we consider a function

$f$  defined as  $f(x) = \int_{\mathbb{R}^d} \varphi(x)_\omega \theta_\omega d\omega = \langle \varphi(x), \theta \rangle$ , then  $\theta_\omega = \frac{1}{(2\pi)^{d/2}} \hat{f}(\omega) / \sqrt{\hat{q}(\omega)}$ , and the squared norm of  $\theta$  is equal to  $\frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \frac{|\hat{f}(w)|^2}{\hat{q}(w)} d\omega$ , where  $\hat{f}$  denotes the Fourier transform of  $f$ . Therefore, the norm of a function  $f \in \mathcal{H}$  should be:

$$\|f\|_{\mathcal{H}}^2 = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \frac{|\hat{f}(w)|^2}{\hat{q}(w)} d\omega.$$

Given the candidate for the norm and the associated dot-product, we can simply check that this is the correct one by showing the reproducing property  $\langle f, k(\cdot, x) \rangle$  for this dot-product. Note the similarity with the penalty for the kernel on  $[0, 1]$  (see more similarity below).

**Link with derivatives.** When  $f$  has partial derivatives, then the Fourier transform of  $\frac{\partial f}{\partial x_j}$  is equal to  $i\omega_j$  times the Fourier transform of  $f$ . This leads to, using Parseval's theorem,  $\frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} |\omega_j|^2 |\hat{f}(w)|^2 d\omega = \int_{\mathbb{R}^d} \left| \frac{\partial f}{\partial x_j}(x) \right|^2 dx$ , which extends to higher order derivatives:

$$\frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} |\omega_1^{j_1} \cdots \omega_d^{j_d}|^2 |\hat{f}(w)|^2 d\omega = \int_{\mathbb{R}^d} \left| \frac{\partial^j f}{\partial x_1^{j_1} \cdots \partial x_d^{j_d}}(x) \right|^2 dx, \quad (7.4)$$

for a vector  $j \in \mathbb{N}^d$ . This will allow us to find corresponding norms by expanding  $\hat{q}(\omega)^{-1}$  as sums of monomials. We now consider the main classical examples.

**Exponential kernel.** This is the kernel  $q(x - x') = \exp(-\|x - x'\|_2/r)$ , where  $r$  is often referred to as the kernel bandwidth (homogeneous to  $x$ ), for which the Fourier transform can be computed as  $\hat{q}(\omega) = 2^d \pi^{(d-1)/2} \Gamma((d+1)/2) \frac{r^d}{(1+r^2\|\omega\|_2^2)^{(d+1)/2}}$ . See [Rasmussen and Williams \(2006, page 84\)](#). Thus, for  $d$  odd,  $\hat{q}(\omega)^{-1}$  is a sum of monomials, and looking at their orders, we see that the corresponding RKHS norm (that is, the norm on the space of functions on  $\mathbb{R}^d$  that our kernel defines) is penalizing all derivatives up to total order  $(d+1)/2$ , that is, in Eq. (7.4), for all  $j \in \mathbb{N}^d$  such that  $j_1 + \cdots + j_d \leq (d+1)/2$ , which is a Sobolev space (fractional for  $d$  even).<sup>10</sup>

In particular, for  $d = 1$ , we have  $\hat{q}(\omega) = \frac{2r}{1+r^2\omega^2}$ , and thus

$$\begin{aligned} \|f\|_{\mathcal{H}}^2 &= \frac{1}{2\pi} \int_{\mathbb{R}} \frac{|\hat{f}(w)|^2}{\hat{q}(w)} d\omega = \frac{1}{2r} \frac{1}{2\pi} \int_{\mathbb{R}} |\hat{f}(w)|^2 d\omega + \frac{r}{2} \frac{1}{2\pi} \int_{\mathbb{R}} |\omega \hat{f}(w)|^2 d\omega \\ &= \frac{1}{2r} \int_{\mathbb{R}} |f(x)|^2 dx + \frac{r}{2} \int_{\mathbb{R}} |f'(x)|^2 dx, \end{aligned}$$

---

<sup>10</sup>See [https://en.wikipedia.org/wiki/Sobolev\\_space](https://en.wikipedia.org/wiki/Sobolev_space).

and we recover the Sobolev space of functions with squared-integrable derivatives.




The constant  $r$  is homogeneous to the input  $x$ , while the constant  $R$  will be homogeneous to features  $\varphi(x)$  (that is, square roots of kernel values). A common rule of thumb is to choose  $r$  to be a quantile (such as the median) of all pairwise distances  $\|x_i - x_j\|_2$  of the training data.

**Gaussian kernel.** This is the kernel  $q(x - x') = \exp(-\|x - x'\|_2^2/r^2)$  (still with a kernel bandwidth  $r$ ), for which the Fourier transform can be explicitly computed as  $\hat{q}(\omega) = (\pi r^2)^{d/2} \exp(-r^2 \|\omega\|_2^2/4)$ . By expanding  $\hat{q}(\omega)^{-1}$  through its power series as  $\hat{q}(\omega)^{-1} = (\pi r^2)^{d/2} \sum_{s=0}^{\infty} \frac{(r \|\omega\|_2^2)^{2s}}{4^s s!}$ , this corresponds to an RKHS norm which is penalizing all derivatives. Note that all members of this RKHS (the associated function space) are infinitely differentiable and thus much smoother than functions coming from the exponential kernel (the RKHS is smaller).

**Matern kernels and Sobolev spaces.** More generally, one can define a series of kernels so that  $\hat{q}(\omega)$  is proportional to  $(1 + r^2 \|\omega\|_2^2)^{-s}$  for  $s > d/2$ , to ensure integrability of the Fourier transform. These so-called “Matern kernels” all correspond to Sobolev spaces of order  $s$  and can be computed in closed form; see [Rasmussen and Williams \(2006, page 84\)](#). A key fact is that to be an RKHS, a Sobolev space has to have many derivatives when  $d$  grows; in particular, having only first-order derivatives ( $s = 1$ ) only leads to an RKHS for  $d = 1$ , and having  $s = 0$  never does.

For  $s = \frac{d+3}{2}$ , we have  $k(x, x') \propto (1 + \sqrt{3} \|x - x'\|_2/r) \exp(-\sqrt{3} \|x - x'\|_2/r)$ , and for  $s = \frac{d+5}{2}$ , we have  $k(x, x') \propto (1 + \sqrt{5} \|x - x'\|_2/r + \frac{5}{3} \|x - x'\|_2^2/r^2) \exp(-\sqrt{5} \|x - x'\|_2/r)$ . General values  $s$  also lead to closed-form formulas (through Bessel functions).

**Density in  $L_2(\mathbb{R}^d)$ .** For all the kernels below, the set  $\mathcal{H}$  is dense in  $L_2(\mathbb{R}^d)$  (the set of square-integrable functions with respect to the Lebesgue measure), meaning that all functions in  $L_2(\mathbb{R}^d)$  can be approached (with respect to their corresponding norm) by a function in  $\mathcal{H}$ . This is made quantitative in Section 7.5.2.

 In this chapter, we will consider two spaces of integrable functions, with respect to the Lebesgue measure (which is not a probability measure), which we denote  $L_2(\mathbb{R}^d)$ , and with respect to the probability measure of the input data, which we denote  $L_2(p)$ . If  $p$  has a density with respect to the Lebesgue measure and this density  $\frac{dp}{dx}(x)$  is uniformly bounded, then  $L_2(\mathbb{R}^d) \subset L_2(p)$ ; more precisely,  $\|f\|_{L_2(p)} \leq \left\| \frac{dp}{dx} \right\|_{\infty}^{1/2} \|f\|_{L_2(\mathbb{R}^d)}$ . However, the converse is not true, simply because being an element of  $L_2(\mathbb{R}^d)$  imposes a zero limit at infinity, which being an element of  $L_2(p)$  does not impose (moreover, non zero constants are in  $L_2(p)$  but not in  $L_2(\mathbb{R}^d)$ ). Note moreover that  $\left\| \frac{dp}{dx} \right\|_{\infty}$  is typically exponential in  $d$ , and is homogeneous to  $r^{-d}$  (in terms of units), where,  $r$  is homogeneous to  $x$ .

**Examples of members of RKHS.** Below, we sampled  $n = 10$  random points in  $[-1, 1]$  with 10 random responses, and we look for the function  $f \in \mathcal{H}$  such that  $f(x_i) = y_i$  for

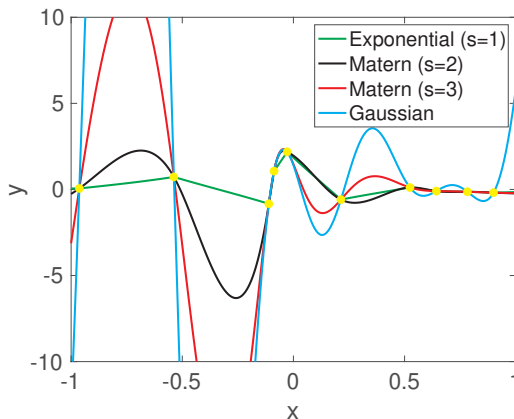


Figure 7.2: Examples of functions in the reproducing kernel Hilbert spaces (RKHS) for several kernels. All functions are the minimum norm interpolators of the yellow points.

all  $i \in \{1 \dots, n\}$  and with minimum norm. Given the representer theorem, we can write  $f(x) = \sum_{i=1}^n \alpha_i k(x, x_i)$ , and the interpolation condition implies that  $K\alpha = y$ , and thus  $\alpha = K^{-1}y$  (see Prop. 7.2).

We consider several kernels in Figure 7.2, going from close to piecewise affine interpolation to infinitely differentiable functions (for the Gaussian kernel).

### 7.3.4 Beyond vectorial input spaces (♦)

While our theoretical analysis of kernel methods focuses a lot on kernels on  $\mathbb{R}^d$  and their link with differentiability properties of the target function, kernels can be applied to a wide variety of problems with various input types. We give below classic examples (see more details by Shawe-Taylor and Cristianini, 2004)

- Set of subsets of a given set  $V$ : for example, the function  $k$  defined as  $k(A, B) = \frac{|A \cap B|}{|A \cup B|}$  is a positive definite kernel (classically referred to as the Jaccard index<sup>11</sup>).
- Point clouds: a point cloud in  $\mathbb{R}^d$  is a finite subset of  $\mathbb{R}^d$ , thus with no particular ordering. They occur, for example, in computer vision or graphics. To build a kernel for such objects, a simple first idea is to compute the empirical average of a certain feature vector  $\varphi : \mathbb{R}^d \rightarrow \mathcal{H}$ , and then use a kernel on  $\mathcal{H}$ . Other kernels may be obtained as functions of the concatenation of the two point clouds (see more details by Cuturi et al., 2005). These constructions extend to probability distributions.
- Text documents/web pages: with the usual “bag of words” assumption, we represent a text document or a web page by considering a vocabulary of “words” (this could be groups of letters, single original words, or groups of words or letters), and counting the number of occurrences of this word in the corresponding document. This gives a typically high-dimensional feature vector  $\varphi(x)$  (with dimension the

<sup>11</sup>See [https://en.wikipedia.org/wiki/Jaccard\\_index](https://en.wikipedia.org/wiki/Jaccard_index).

vocabulary size). Using linear functions on this feature provides cheap and stable predictors on such data types (better models that take into account the word order can be obtained, such as neural networks, at the expense of significantly more computational resources). See, e.g., [Joulin et al. \(2017\)](#) for examples.

- Sequences: given some finite alphabet  $\mathcal{A}$ , we consider the set  $\mathcal{X}$  of finite sequences in  $\mathcal{A}$  with arbitrary length. A classical infinite-dimensional feature space is indexed by  $\mathcal{X}$  itself, and for  $y \in \mathcal{X}$ ,  $\varphi(x)_y$  is equal to 1 if  $y$  is a subsequence of  $x$  (we could also count the number of times the subsequence  $y$  appears in  $x$ , or we could add a weight that depends on  $y$ , e.g., to penalize longer subsequences). This kernel has an infinite-dimensional feature space, but for two sequences  $x$  and  $x'$ , we can enumerate all subsequences of  $x$  and  $x'$  and compare them in polynomial time (there exist much faster algorithms, see [Gusfield \(1997\)](#)). These kernels have many applications in bioinformatics.

The same techniques can be extended to more general combinatorial objects such as trees and graphs (see [Shawe-Taylor and Cristianini, 2004](#)).

- Images: before neural networks took over in the years 2010s with the use of large amounts of data, several kernels were designed for images, with often a “bag-of-words” assumption that provides for free invariance by translation. The key is what to consider as “words”, i.e., the presence of specific local patterns in the image and the regions under which this assumption is made. See [Zhang et al. \(2007\)](#) for details.

## 7.4 Algorithms

In this section, we briefly mention algorithms aimed at solving

$$\min_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n \ell(y_i, f(x_i)) + \frac{\lambda}{2} \|f\|_{\mathcal{H}}^2, \quad (7.5)$$

for  $\ell$  being convex with respect to its second variable. We assume that for all  $i \in \{1, \dots, n\}$ ,  $k(x_i, x_i) = \|\varphi(x_i)\|^2 \leq R^2$ .

### 7.4.1 Representer theorem

We can directly apply the representer theorem, as done in Eq. (7.2) and try to solve

$$\min_{\alpha \in \mathbb{R}^n} \frac{1}{n} \sum_{i=1}^n \ell(y_i, (K\alpha)_i) + \frac{\lambda}{2} \alpha^\top K \alpha,$$

which is a convex optimization problem since  $\ell$  is assumed convex with respect to the second variable, and  $K$  is positive-semidefinite.

In the particular case of the square loss (ridge regression), this leads to

$$\min_{\alpha \in \mathbb{R}^n} \frac{1}{2n} \|y - K\alpha\|_2^2 + \frac{\lambda}{2} \alpha^\top K \alpha,$$



and setting the gradient to zero, we get  $(K^2 + n\lambda K)\alpha = Ky$ , with a solution  $\alpha = (K + n\lambda I)^{-1}y$ , which is not unique as soon as  $K$  is not invertible.

However, in general (for the square loss and beyond), it is an ill-conditioned optimization problem because  $K$  often has very small eigenvalues (more on this later). When the loss is smooth, the Hessians are equal to  $\frac{1}{n}K \text{Diag}(h)K + \lambda K$ , where  $h \in \mathbb{R}^n$  is a vector of second-order derivatives of  $\ell$ , so that the Hessians are ill-conditioned.

A better alternative is to first compute a square root of  $K$  as  $K = \Phi\Phi^\top$ , where  $\Phi \in \mathbb{R}^{n \times m}$ , and  $m$  the rank of  $K$ , and solve

$$\min_{\beta \in \mathbb{R}^m} \frac{1}{n} \sum_{i=1}^n \ell(y_i, (\Phi\beta)_i) + \frac{\lambda}{2} \|\beta\|_2^2,$$

with  $\beta = \Phi^\top \alpha$ . Note that this corresponds to an explicit feature space representation (that is, the rows of  $\Phi$  correspond to features in  $\mathbb{R}^m$  for the corresponding data point). For ridge regression, the objective function's Hessian is equal to  $\frac{1}{n}\Phi^\top \Phi + \lambda I$ , which is well-conditioned because its lowest eigenvalue is greater than  $\lambda$  and is thus directly controlled by regularization.

Computing a square root can be done in several ways (through Cholesky decomposition or SVD) (Golub and Loan, 1996), in running time  $O(m^2n)$ .

### 7.4.2 Column sampling

To approximate  $K$ , approximate square roots are a very useful tool, and among various algorithms, approximating  $K \in \mathbb{R}^{n \times n}$  from a subset of its columns can be done as  $K \approx K(V, I)K(I, I)^{-1}K(I, V)$ , where  $K(A, B)$  is the sub-matrix of  $K$  obtained by taking rows from the set  $A \subset \{1, \dots, n\}$  and columns from  $B \subset \{1, \dots, n\}$ , and  $V = \{1, \dots, n\}$ . See below for an illustration when  $I = \{1, \dots, m\}$  and a partition of the kernel matrix.

$K(I, I)$	$K(I, J)$
$K(J, I)$	$K(J, J)$

This corresponds to an approximate square root  $\Phi = K(V, I)K(I, I)^{-1/2} \in \mathbb{R}^{n \times m}$ , with  $m = |I|$ , and it can be computed in time  $O(m^2n)$  (computing the entire kernel matrix is not even needed). Then, the complexity is typically  $O(m^2n)$  instead of  $O(n^3)$  (e.g., when using matrix inversion for ridge regression, for faster algorithms, see below), and is thus linear in  $n$ .

**Exercise 7.6** (♦) *Show that column sampling corresponds to approximating optimally each  $\varphi(x_j)$ ,  $j \notin I$ , by a linear combination of  $\varphi(x_i)$ ,  $i \in I$ .*

This approximation technique, often called “Nyström approximation,” can be analyzed when the columns are chosen randomly (Rudi et al., 2015).

### 7.4.3 Random features

Some kernels have a special form that leads to specific approximation schemes, that is,

$$k(x, x') = \int_{\mathcal{V}} \varphi(x, v) \varphi(x', v) d\tau(v),$$


where  $\tau$  is a probability distribution on some space  $\mathcal{V}$  and  $\varphi(x, v) \in \mathbb{R}$ . We can then approximate the expectation by an empirical average

$$\hat{k}(x, x') = \frac{1}{m} \sum_{j=1}^m \varphi(x, v_j) \varphi(x', v_j),$$

where the  $v_j$ ’s are sampled i.i.d. from  $\tau$ . We can thus use an explicit feature representation  $\hat{\varphi}(x) = (\frac{1}{\sqrt{m}} \varphi(x, v_j))_{j \in \{1, \dots, m\}}$ , and solve

$$\min_{\beta \in \mathbb{R}^m} \frac{1}{n} \sum_{i=1}^n \ell(y_i, \hat{\varphi}(x_i)^\top \beta) + \frac{\lambda}{2} \|\beta\|_2^2.$$

For this scheme to make sense, the number  $m$  of random features has to be significantly smaller than  $n$ , which is often sufficient in practice (see an analysis by Rudi and Rosasco, 2017).

 Note that dimension reduction is here performed independently of the input data (that is, the random feature functions  $\varphi(\cdot, v_j)$  are selected before the data are observed, as opposed to column sampling, which is a data-dependent dimension reduction scheme.

The two classic examples are:

- **Translation-invariant kernels:**  $k(x, y) = q(x - y) = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \hat{q}(\omega) e^{i\omega^\top (x-y)} d\omega$ , for which we can take  $\varphi(x, \omega) = \sqrt{q(0)} e^{i\omega^\top x} \in \mathbb{C}$ , where  $\omega$  is sampled from the distribution with density  $\frac{1}{(2\pi)^d} \frac{\hat{q}(\omega)}{q(0)}$ , which is a Gaussian distribution for the Gaussian kernel. Alternatively, one can use a real-valued feature (instead of a complex-valued one) by using  $\sqrt{2} \cos(\omega^\top x + b)$  with  $b$  sampled uniformly in  $[0, 2\pi]$  (Rahimi and Recht, 2008).
- **Neural networks with random weights:** we can start from an expectation, for which the sampled features are classical, e.g.,  $\varphi(x, v) = \sigma(v^\top x)$  for some function  $\sigma : \mathbb{R} \rightarrow \mathbb{R}$ . For the “rectified linear unit”, that is,  $\sigma(\alpha) = \max\{0, \alpha\}$ , and for  $v$  sampled uniformly on the sphere, we have (proof left as an exercise)  $k(x, x') = \frac{\|x\|_2 \|x'\|_2}{2(d+1)\pi} [(\pi - \eta) \cos \eta + \sin \eta]$ , where  $\cos \eta = \frac{x^\top x'}{\|x\|_2 \|x'\|_2}$  (Le Roux and Bengio, 2007). Therefore, we can view a neural network with a large number of hidden neurons,

with random input weights and not optimized as a kernel method. See a thorough discussion in Chapter 9.

#### 7.4.4 Dual algorithms (♦)

For the following two algorithms, we go back to the notation  $f(x) = \langle \varphi(x), \theta \rangle$  with  $\theta \in \mathcal{H}$  because it is more adapted (and is a direct infinite-dimensional extension of the algorithms from Chapter 5). To solve  $\min_{\theta \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n \ell(y_i, \langle \varphi(x_i), \theta \rangle) + \frac{\lambda}{2} \|\theta\|^2$ , for a loss which is convex with respect to the second variable, we can derive a Lagrange dual in the following way (for an introduction to Lagrange duality, see [Boyd and Vandenberghe, 2004](#)). We start by reformulating the problem as a constrained problem:

$$\begin{aligned} & \min_{\theta \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n \ell(y_i, \langle \varphi(x_i), \theta \rangle) + \frac{\lambda}{2} \|\theta\|^2 \\ &= \min_{\theta \in \mathcal{H}, u \in \mathbb{R}^n} \frac{1}{n} \sum_{i=1}^n \ell(y_i, u_i) + \frac{\lambda}{2} \|\theta\|^2 \text{ such that } \forall i \in \{1, \dots, n\}, \langle \varphi(x_i), \theta \rangle = u_i. \end{aligned}$$

By Lagrange duality, this is equal to (with  $\lambda$  added on top of the regular multiplier  $\alpha$  for convenience):

$$\begin{aligned} & \max_{\alpha \in \mathbb{R}^n} \min_{\theta \in \mathcal{H}, u \in \mathbb{R}^n} \frac{1}{n} \sum_{i=1}^n \ell(y_i, u_i) + \frac{\lambda}{2} \|\theta\|^2 + \lambda \sum_{i=1}^n \alpha_i (u_i - \langle \varphi(x_i), \theta \rangle) \\ &= \max_{\alpha \in \mathbb{R}^n} \left\{ \frac{1}{n} \sum_{i=1}^n \min_{u_i \in \mathbb{R}} \{ \ell(y_i, u_i) + n\lambda\alpha_i u_i \} + \min_{\theta \in \mathcal{H}} \left\{ \frac{\lambda}{2} \|\theta\|^2 - \lambda \sum_{i=1}^n \alpha_i \langle \varphi(x_i), \theta \rangle \right\} \right\} \\ & \quad \text{by reordering,} \\ &= \max_{\alpha \in \mathbb{R}^n} \frac{1}{n} \sum_{i=1}^n \min_{u_i \in \mathbb{R}} \{ \ell(y_i, u_i) + n\lambda\alpha_i u_i \} - \frac{\lambda}{2} \left\| \sum_{i=1}^n \alpha_i \varphi(x_i) \right\|^2 \text{ with } \theta = \sum_{i=1}^n \alpha_i \varphi(x_i), \\ &= \max_{\alpha \in \mathbb{R}^n} \frac{1}{n} \sum_{i=1}^n \min_{u_i \in \mathbb{R}} \{ \ell(y_i, u_i) + n\lambda\alpha_i u_i \} - \frac{\lambda}{2} \alpha^\top K \alpha, \end{aligned}$$

with  $\theta = \sum_{i=1}^n \alpha_i \varphi(x_i)$  at optimum. Since the functions  $\alpha_i \mapsto \min_{u_i \in \mathbb{R}} \{ \ell(y_i, u_i) + n\lambda\alpha_i u_i \}$  are concave (as minima of affine functions), this is a concave maximization problem.

Note the similarity with the representer theorem (existence of  $\alpha \in \mathbb{R}^n$  such that  $\theta = \sum_{i=1}^n \alpha_i \varphi(x_i)$ ) and the dissimilarity (one is a minimization problem, one is maximization problem). Moreover, when the loss is smooth, one can show that the function  $\alpha_i \mapsto \min_{u_i \in \mathbb{R}} \{ \ell(y_i, u_i) + n\lambda\alpha_i u_i \}$  is a strongly concave function, and thus relatively easy to optimize (in other words, the associated condition numbers of dual problems are smaller than when using the representer theorem).

**Exercise 7.7** (a) For ridge regression, compute the dual problem and compare the condition number of the primal problem and the condition number of the dual problem; (b) compare the two formulations to the use of normal equations as in Chapter 3, and relate the two using the matrix inversion lemma  $(\Phi\Phi^\top + n\lambda I)^{-1}\Phi = \Phi(\Phi^\top\Phi + n\lambda I)^{-1}$ .

### 7.4.5 Stochastic gradient descent (◆)

When minimizing an expectation

$$\min_{\theta \in \mathcal{H}} \mathbb{E}[\ell(y, \langle \varphi(x), \theta \rangle)] + \frac{\lambda}{2} \|\theta\|^2$$

as in Chapter 5, the stochastic gradient algorithm leads to the recursion

$$\theta_t = \theta_{t-1} - \gamma_t [\ell'(y_t, \langle \varphi(x_t), \theta_{t-1} \rangle) \varphi(x_t) + \lambda \theta_{t-1}],$$

where  $(x_t, y_t)$  is an i.i.d. sample from the distribution defining the expectation, and  $\ell'$  is the derivative with respect to the second variable.

When initializing at  $\theta_0 = 0$ ,  $\theta_t$  is a linear combination of all  $\varphi(x_i)$ ,  $i = 1, \dots, t$ , and thus we can write

$$\theta_t = \sum_{i=1}^t \alpha_i^{(t)} \varphi(x_i),$$

with  $\alpha^{(0)} = 0$ , and the recursion in  $\alpha$  as

$$\alpha_i^{(t)} = (1 - \gamma_t \lambda) \alpha_i^{(t-1)} \text{ for } i \in \{1, \dots, t-1\}, \text{ and } \alpha_t^{(t)} = -\gamma_t \ell' \left( y_t, \sum_{i=1}^{t-1} \alpha_i^{(t-1)} k(x_t, x_i) \right).$$

The complexity after  $t$  iterations is  $O(t^2)$  kernel evaluations. The convergence rates from Chapter 5 apply. More precisely, if the loss is  $G$ -Lipschitz continuous, then, for  $F(\theta) = \mathbb{E}[\ell(y, \langle \varphi(x), \theta \rangle)] + \frac{\lambda}{2} \|\theta\|^2$ , we have, for the averaged iterate  $\bar{\theta}_t$  (from Prop. 5.8):

$$\mathbb{E}[F(\bar{\theta}_t)] - \inf_{\theta \in \mathcal{H}} F(\theta) \leq \frac{2G^2 R^2 (1 + \log t)}{\lambda t}.$$



When doing a single pass with  $t = n$ , then  $F(\theta)$  is the regularized expected risk, and we obtain a generalization bound, leading to  $\mathbb{E}[\mathcal{R}(f_{\bar{\theta}_t})] \leq \frac{G^2 R^2}{\lambda n} + \inf_{f \in \mathcal{H}} \left\{ \mathcal{R}(f) + \frac{\lambda}{2} \|f\|_{\mathcal{H}}^2 \right\}$ . These bounds are similar to the ones in Section 7.5 below (which assume a regularized empirical risk minimizer is available).

### 7.4.6 “Kernelization” of linear algorithms

Beyond supervised learning, many unsupervised learning algorithms can be “kernelized,” such as principal component analysis (as presented in Section 3.9),  $K$ -means, or canonical correlation analysis.<sup>12</sup> Indeed, these algorithms can be cast only through the matrices of dot-products between observations and can thus be applied after the feature transformation  $\varphi : \mathcal{X} \rightarrow \mathcal{H}$ , and run implicitly only using the kernel function  $k(x, x') = \langle \varphi(x), \varphi(x') \rangle$ . See Schölkopf and Smola (2001); Shawe-Taylor and Cristianini (2004) for details and exercises below.

<sup>12</sup>See [https://en.wikipedia.org/wiki/Canonical\\_correlation](https://en.wikipedia.org/wiki/Canonical_correlation).

**Exercise 7.8 (Kernel principal component analysis)** We consider  $n$  observations  $x_1, \dots, x_n$  in a set  $\mathcal{X}$  equipped with a positive definite kernel and feature map  $\varphi$  from  $\mathcal{X}$  to  $\mathcal{H}$ . Show that the largest eigenvector of the empirical non-centered covariance operator  $\frac{1}{n} \sum_{i=1}^n \varphi(x_i) \otimes \varphi(x_i)$  is proportional to  $\sum_{i=1}^n \alpha_i \varphi(x_i)$  where  $\alpha \in \mathbb{R}^n$  is an eigenvector of the  $n \times n$  kernel matrix associated with the largest eigenvalue. Given the RKHS  $\mathcal{H}$  associated with the kernel  $k$ , relate this eigenvalue problem to the maximizer of  $\frac{1}{n} \sum_{i=1}^n f(x_i)^2$  subject to  $\|f\|_{\mathcal{H}} = 1$ .

**Exercise 7.9 (Kernel  $K$ -means)** Show that the  $K$ -means clustering algorithm<sup>13</sup> can be expressed only using dot-products.

**Exercise 7.10 (Kernel quadrature)** We consider a probability distribution  $p$  on a set  $\mathcal{X}$  equipped with a positive definite kernel  $k$  with feature map  $\varphi : \mathcal{X} \rightarrow \mathcal{H}$ . For a function  $f$  which is linear in  $\varphi$ , we want to approximate  $\int_{\mathcal{X}} f(x) dp(x)$  from a linear combination  $\sum_{i=1}^n \alpha_i f(x_i)$  with  $\alpha \in \mathbb{R}^n$ .  
(a) Show that

$$\left| \int_{\mathcal{X}} f(x) dp(x) - \sum_{i=1}^n \alpha_i f(x_i) \right| \leq \|f\| \cdot \left\| \int_{\mathcal{X}} \varphi(x) dp(x) - \sum_{i=1}^n \alpha_i \varphi(x_i) \right\|.$$

(b) Express the square of the right-hand side with the kernel function and show how to minimize with respect to  $\alpha \in \mathbb{R}^n$ .

(c) Show that if the points  $x_1, \dots, x_n$  are sampled i.i.d. from  $p$  and  $\alpha_i = 1/n$  for all  $i$ , then  $\mathbb{E} \left\| \int_{\mathcal{X}} \varphi(x) dp(x) - \sum_{i=1}^n \alpha_i \varphi(x_i) \right\|^2 \leq \frac{1}{n} \mathbb{E}[k(x, x)]$ .

**Exercise 7.11** Consider a binary classification problems with data  $(x_1, y_1), \dots, (x_n, y_n)$  in  $\mathcal{X} \times \{-1, 1\}$ , with a positive kernel  $k$  defined on  $\mathcal{X}$  with feature map  $\varphi : \mathcal{X} \rightarrow \mathcal{H}$ . Let  $\mu_+$  (resp.  $\mu_-$ ) be the mean of all feature vectors for positive (resp. negative) labels. We consider the classification rule that predicts 1 if  $\|\varphi(x) - \mu_+\|_{\mathcal{H}}^2 < \|\varphi(x) - \mu_-\|_{\mathcal{H}}^2$  and  $-1$  otherwise. Compute the classification rule only using kernel functions and compare it to local averaging methods from Chapter 6.

## 7.5 Generalization guarantees - Lipschitz-continuous losses

In this section, we consider a  $G$ -Lipschitz-continuous loss function, and consider a minimizer  $\hat{f}_D^{(c)}$  of the constrained problem

$$\min_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n \ell(y_i, f(x_i)) \text{ such that } \|f\|_{\mathcal{H}} \leq D, \quad (7.6)$$

---

<sup>13</sup>See [https://en.wikipedia.org/wiki/K-means\\_clustering](https://en.wikipedia.org/wiki/K-means_clustering).

and the unique minimizer  $\hat{f}_\lambda^{(r)}$  of the regularized problem

$$\min_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n \ell(y_i, f(x_i)) + \frac{\lambda}{2} \|f\|_{\mathcal{H}}^2. \quad (7.7)$$

We denote by  $\mathcal{R}(f) = \mathbb{E}[\ell(y, f(x))]$  the expected risk, and by  $f^*$  one of its minimizers (which we assume to be square integrable). We assume  $k(x, x) \leq R^2$  almost surely.

We can first relate the excess risk to the  $L_2$ -norm of  $f - f^*$ , as (using Jensen's inequality)

$$\begin{aligned} \mathcal{R}(f) - \mathcal{R}(f^*) &\leq \mathbb{E}[|\ell(y, f(x)) - \ell(y, f^*(x))|] \leq G \mathbb{E}[|f(x) - f^*(x)|] \\ &\leq G \sqrt{\mathbb{E}[|f(x) - f^*(x)|^2]} = G \|f - f^*\|_{L_2(p)}, \end{aligned}$$

that is, the excess risk is dominated by the  $L_2(p)$ -norm of  $f - f^*$ . For  $\mathcal{X} = \mathbb{R}^d$ , and probability measures with bounded density with respect to the Lebesgue measure, we have shown that  $\|f\|_{L_2(p)} \leq \left\| \frac{dp}{dx} \right\|_\infty^{1/2} \|f\|_{L_2(\mathbb{R}^d)}$ , so we can replace in upper-bounds the quantity  $G \|f - f^*\|_{L_2(p)}$  by  $G \left\| \frac{dp}{dx} \right\|_\infty^{1/2} \|f - f^*\|_{L_2(\mathbb{R}^d)}$ .

### 7.5.1 Risk decomposition

We now assume that  $\sup_{x \in \mathcal{X}} k(x, x) \leq R^2$ , compatible with the convention in earlier chapters on linear models (e.g., Section 4.5.3) that  $\|\varphi(x)\|_{\mathcal{H}}^2 \leq R^2$  for all  $x \in \mathcal{X}$ .

**Constrained problem.** Dimension-free results from Chapter 4 (Prop. 4.5), based on Rademacher complexities, immediately apply, and we obtain that the estimation error is bounded from above by  $\frac{4GDR}{\sqrt{n}}$ , leading to:

$$\mathbb{E}[\mathcal{R}(\hat{f}_D^{(c)})] - \mathcal{R}(f^*) \leq \frac{4GDR}{\sqrt{n}} + G \inf_{\|f\|_{\mathcal{H}} \leq D} \|f - f^*\|_{L_2(p)},$$

(the first term is the **estimation error** of using the empirical risk minimizer constrained to the ball of RKHS norm less than  $D$ , the second term is the **approximation error**).

To find the optimal  $D$  (to balance estimation and approximation error), we can minimize the bound with respect to  $D$ , leading to (using  $|a| + |b| \leq \sqrt{2(a^2 + b^2)}$ ):

$$\begin{aligned} \inf_{D \geq 0} \frac{4GRD}{\sqrt{n}} + G \inf_{\|f\|_{\mathcal{H}} \leq D} \|f - f^*\|_{L_2(p)} &= \inf_{f \in \mathcal{H}} \frac{4GR\|f\|_{\mathcal{H}}}{\sqrt{n}} + G \|f - f^*\|_{L_2(p)} \\ &\leq G \sqrt{2 \inf_{f \in \mathcal{H}} \left\{ \|f - f^*\|_{L_2(p)}^2 + \frac{16R^2}{n} \|f\|_{\mathcal{H}}^2 \right\}}. \quad (7.8) \end{aligned}$$

Note that if we consider  $D$  equal to  $\frac{\sqrt{n}}{4R} \sqrt{\inf_{f \in \mathcal{H}} \left\{ \|f - f^*\|_{L_2(p)}^2 + \frac{16R^2}{n} \|f\|_{\mathcal{H}}^2 \right\}}$ , we can obtain a bound proportional to what we obtained above.

Overall, we need to understand how the deterministic quantity

$$A(\mu, f^*) = \inf_{f \in \mathcal{H}} \{ \|f - f^*\|_{L_2(p)}^2 + \mu \|f\|_{\mathcal{H}}^2 \}$$

goes to zero when  $\mu$  goes to zero (note that we define  $A(\mu, f^*)$  above through a regularized estimation problem to study trade-offs between estimation and approximation errors, and that this is not a justification to use  $16R^2/n$  as a regularization parameter in practice). A few situations are possible:

- If the target function  $f^*$  happens to be in  $\mathcal{H}$  (well-specified problem), then we have  $A(\mu, f^*) \leq \mu \|f^*\|_{\mathcal{H}}^2$ , and thus it tends to zero as  $O(\mu)$ . This is the best-case scenario and requires that the target function is sufficiently regular (e.g., with at least  $d/2$  derivatives for  $\mathcal{X} = \mathbb{R}^d$ ). Then, using it with  $\mu = 16R^2/n$  above, the overall excess risk goes to zero as  $O(1/\sqrt{n})$ . Moreover, the suggested value of  $D$  not surprisingly is exactly  $\|f^*\|_{\mathcal{H}}$ .
- The target function  $f^*$  is not in  $\mathcal{H}$  (mis-specified problem), but can be approached arbitrarily closely in  $L_2(p)$ -norm by a function in  $\mathcal{H}$ ; in other words,  $f^*$  is in the closure of  $\mathcal{H}$  in  $L_2(p)$ . In this situation,  $A(\mu, f^*)$  goes to zero as  $\mu$  goes to zero, but without an explicit rate if no further assumptions are made.

For  $\mathcal{X} = \mathbb{R}^d$ , and the distribution  $p$  of inputs with a bounded density with respect to the Lebesgue measure, and for the translation-invariant kernels from Section 7.3.3, this closure includes all of  $L_2(\mathbb{R}^d)$ , so this case includes most potential functions. See Section 7.5.2 for explicit rates.

- Otherwise, denoting  $\Pi_{\bar{\mathcal{H}}}(f^*)$  the orthogonal projection in  $L_2(p)$  of  $f^*$  on the closure of  $\mathcal{H}$ , by the Pythagorean theorem,  $A(\mu, f^*) = A(\mu, \Pi_{\bar{\mathcal{H}}}(f^*)) + \|f^* - \Pi_{\bar{\mathcal{H}}}(f^*)\|_{L_2(p)}^2$ , that is, there is an incompressible error due to a choice of function space which is not large enough.

Note that we will use the same reasoning for neural networks in Section 9.4.

**Regularized problem (♦).** For the regularized problem, we can use the bound from Chapter 4 (Prop. 4.6):

$$\mathbb{E}[\mathcal{R}(\hat{f}_\lambda^{(r)})] - \mathcal{R}(f^*) \leq \frac{32G^2R^2}{\lambda n} + \inf_{f \in \mathcal{H}} \left\{ G\|f - f^*\|_{L_2(p)} + \frac{\lambda}{2}\|f\|_{\mathcal{H}}^2 \right\}.$$

We can now minimize the bound with respect to  $\lambda$ , with  $f$  fixed, as  $\lambda = \frac{8RG}{\|f\|_{\mathcal{H}}\sqrt{n}}$ , to obtain the bound:

$$G \inf_{f \in \mathcal{H}} \left\{ \|f - f^*\|_{L_2(p)} + \frac{8R}{\sqrt{n}}\|f\|_{\mathcal{H}} \right\} \leq G \sqrt{2 \inf_{f \in \mathcal{H}} \left\{ \|f - f^*\|_{L_2(p)}^2 + \frac{64R^2}{n}\|f\|_{\mathcal{H}}^2 \right\}},$$

which is the same bound as for the constrained problem but on a more commonly used optimization problem in practice. Note that for well-specified problems, the suggested regularization parameter is  $\lambda = \frac{8RG}{\|f^*\|_{\mathcal{H}}\sqrt{n}}$ .

### 7.5.2 Approximation error for translation-invariant kernels on $\mathbb{R}^d$

We first start by analyzing kernel methods' approximation error for translation invariant kernels. Given a distribution  $p$  of inputs, the goal is to compute

$$A(\mu, f^*) = \inf_{f \in \mathcal{H}} \|f - f^*\|_{L_2(p)}^2 + \mu \|f\|_{\mathcal{H}}^2,$$

where  $f^*$  is the target function (e.g., the minimizer of the test risk), which we assume is squared-integrable. If  $A(\mu, f^*)$  tends to zero when  $\mu$  tends to zero for any fixed  $f^*$ , kernel-based supervised learning leads to universally consistent algorithms.

We assume that  $\|f - f^*\|_{L_2(p)}^2 \leq \frac{C}{r^d} \|f - f^*\|_{L_2(\mathbb{R}^d)}^2$  (e.g., with  $C = r^d \|dp/dx\|_\infty$  where  $dp/dx$  is the density of  $p$ ), where we have introduced a constant  $r$  to preserve homogeneity. Moreover, for simplicity, we assume that  $\|f^*\|_{L_2(\mathbb{R}^d)}$  is finite (which implies that  $f^*$  has to go to zero at infinity). We now give bounds on

$$\tilde{A}(\mu, f^*) = \inf_{f \in \mathcal{H}} \frac{1}{r^d} \|f - f^*\|_{L_2(\mathbb{R}^d)}^2 + \mu \|f\|_{\mathcal{H}}^2,$$

keeping in mind that  $A(\mu, f^*) \leq C \tilde{A}(\mu/C, f^*)$ . Remember from Section 7.5.1 that if  $f^* \in \mathcal{H}$  (best case scenario), then both  $A(\mu, f^*)$  and  $\tilde{A}(\mu, f^*)$  are less than  $\mu \|f^*\|_{\mathcal{H}}^2$ .

**Explicit approximation.** We have, for translation-invariant kernels defined in Section 7.3.3, an explicit formulation of the norm  $\|\cdot\|_{\mathcal{H}}$  as  $\|f\|_{\mathcal{H}}^2 = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \frac{|\hat{f}(\omega)|^2}{\hat{q}(\omega)} d\omega$ , and thus

$$\tilde{A}(\mu, f^*) = \inf_{\hat{f} \in L_2(\mathbb{R}^d)} \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \left[ \frac{1}{r^d} |\hat{f}(\omega) - \hat{f}^*(\omega)|^2 + \mu \frac{|\hat{f}(\omega)|^2}{\hat{q}(\omega)} \right] d\omega.$$

The optimization can be performed independently for each  $\omega$ , which is a quadratic problem. Setting the derivative with respect to  $\hat{f}(\omega)$  to zero leads to  $0 = 2\frac{1}{r^d}(\hat{f}(\omega) - \hat{f}^*(\omega)) + 2\mu \frac{\hat{f}(\omega)}{\hat{q}(\omega)}$ , and thus  $\hat{f}_\mu(\omega) = \frac{\hat{f}^*(\omega)}{1 + \mu r^d \hat{q}(\omega)^{-1}}$ . In terms of the objective function, we get:

$$\tilde{A}(\mu, f^*) = \frac{1}{(2\pi r)^d} \int_{\mathbb{R}^d} |\hat{f}^*(\omega)|^2 \left( 1 - \frac{1}{1 + \mu r^d \hat{q}(\omega)^{-1}} \right) d\omega = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} |\hat{f}^*(\omega)|^2 \frac{\mu}{\hat{q}(\omega) + \mu r^d} d\omega.$$

When  $\mu$  goes to zero, we see that for each  $\omega$ ,  $\hat{f}_\mu(\omega)$  tends to  $\hat{f}^*(\omega)$ . By the dominated convergence theorem,  $\tilde{A}(\mu, f^*)$  goes to zero when  $\mu$  goes to zero.

Without further assumptions, it is impossible to obtain a convergence rate (otherwise, the no-free lunch theorem from Chapter 2 would be invalidated). However, this is possible when assuming regularity properties for  $f^*$ .

**⚠** Note that the universal approximation properties of translation-invariant kernels do not require the kernel bandwidth  $r$  to go to zero (as opposed to smoothing kernels from Chapter 6).



**Sobolev spaces (♦).** If we assume that

$$\frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} (1 + r^2 \|\omega\|_2^2)^t |\hat{f}^*(\omega)|^2 d\omega < +\infty \quad (7.9)$$

for some  $t > 0$ , that is,  $f^*$  with squared integrable partial derivatives up to order  $t$ , then we can further bound:

$$\tilde{A}(\mu, f^*) \leq \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} (1 + r^2 \|\omega\|_2^2)^t |\hat{f}^*(\omega)|^2 d\omega \times \sup_{\omega \in \mathbb{R}^d} \left\{ \frac{\mu}{\hat{q}(\omega) + \mu r^d} \frac{1}{(1 + r^2 \|\omega\|_2^2)^t} \right\}.$$

If we now assume  $\hat{q}(\omega) \propto r^d (1 + r^2 \|\omega\|_2^2)^{-s}$  (Matern kernels), with  $s > d/2$  to get an RKHS, then with  $t \geq s$ ,  $f^* \in \mathcal{H}$ , and have  $\tilde{A}(\mu, f^*) = \mu \|f^*\|_{\mathcal{H}}^2$ . With  $t < s$ , that is the function is not inside the RKHS  $\mathcal{H}$ , then we get a bound proportional to (using  $a + b \geq \frac{t}{s}a + (1 - \frac{t}{s})b \geq a^{t/s} b^{1-t/s}$ ):

$$\sup_{\omega \in \mathbb{R}^d} \left\{ \frac{\mu}{\hat{q}(\omega) + \mu r^d} \frac{1}{(1 + r^2 \|\omega\|_2^2)^t} \right\} \leq \sup_{\omega \in \mathbb{R}^d} \left\{ \frac{\mu}{\hat{q}(\omega)^{t/s} (\mu r^d)^{1-t/s}} \frac{1}{(1 + r^2 \|\omega\|_2^2)^t} \right\} = O(\mu^{t/s}).$$

**Exercise 7.12 (♦)** Find an upper-bound of  $\tilde{A}(\mu, f^*)$  for the same assumption on  $f^*$  but with the Gaussian kernel.



There are two regularities, with two different constraints:  $t \geq 0$  for the target function, and  $s > d/2$  for the kernel.

**Putting things together.** Thus, for Lipschitz-continuous losses and target functions that satisfy Eq. (7.9), we get an expected excess risk of the order  $(\tilde{A}(R^2/n, f^*))^{1/2} = O(n^{-t/(2s)})$ , when  $t \leq s$ . For example, when  $t = 1$ , that is, only first-order derivatives are assumed to be square integrable, then for  $s = d/2 + 1/2$  (exponential kernel), we obtain a rate of  $O(n^{-1/(d+1)})$ , which is similar to the rate obtained with local averaging techniques in Chapter 6 (note here that we are in Lipschitz-loss set-up, which leads to worse rates, see the square loss in Section 7.6). Thus, kernel methods do not escape the curse of dimensionality (which is unavoidable anyway). *However*, with the proper choice of the regularization parameter, they can benefit from extra smoothness of the target function: in the very favorable case, where  $f^* \in \mathcal{H}$ , that is  $t \geq s$ , then we obtain a dimension-independent rate of  $1/\sqrt{n}$ . In intermediate scenarios, the rates are in between. This is why kernel methods are said to be *adaptive to the smoothness* of the target function.

**Approximation bounds (♦).** In some analysis set-ups (such as those explored in Chapter 9), it is required to approximate some  $f^*$  up to  $\varepsilon$  with the minimum possible RKHS norm. This can be done as follows.

A bound on the quantity  $A(\mu, f^*) = \inf_{f \in \mathcal{H}} \{ \|f - f^*\|_{L_2(p)}^2 + \mu \|f\|_{\mathcal{H}}^2 \}$  of the form  $c\mu^\alpha$

for  $\alpha \in (0, 1)$  leads to the following bound:

$$\begin{aligned}
& \inf_{f \in \mathcal{H}} \|f\|_{\mathcal{H}}^2 \text{ such that } \|f - f^*\|_{L_2(p)} \leq \varepsilon \\
&= \inf_{f \in \mathcal{H}} \sup_{\mu \geq 0} \|f\|_{\mathcal{H}}^2 + \mu(\|f - f^*\|_{L_2(p)}^2 - \varepsilon^2) \text{ using Lagrangian duality,} \\
&= \sup_{\mu \geq 0} \mu A(\mu^{-1}, f^*) - \mu \varepsilon^2 \leq \sup_{\mu \geq 0} \mu c \mu^{-\alpha} - \mu \varepsilon^2.
\end{aligned}$$

The optimal  $\mu$  is such that  $(1 - \alpha)c\mu^{-\alpha} = \varepsilon^2$ , leading to an approximation bound proportional to  $\varepsilon^{2(1-1/\alpha)} = \varepsilon^{-2(1-\alpha)/\alpha}$ .

Applied to  $\alpha = t/s$  like before, this leads to an RKHS norm proportional to  $\varepsilon^{-(1-\alpha)/\alpha}$  to get an error less than  $\|f - f^*\|_{L_2(\mathbb{R}^d)}$ . So when  $t = 1$  (single derivative for the target function), and  $s > d/2$  (for the Sobolev kernel), we get a norm of the order  $\varepsilon^{-(1/\alpha-1)} = \varepsilon^{-(s-1)} \geq \varepsilon^{-d/2+1}$ , which explodes exponentially in dimension, which is another way of formulating the curse of dimensionality.

**Relationship between Lipschitz-continuous functions and Sobolev spaces on  $\mathbb{R}^d$  (♦♦).** In the previous chapter on local averaging methods, as well as for neural networks (Chapter 9), we will consider Lipschitz-continuous functions on a subset of  $\mathbb{R}^d$ , which we take here to be the ball of center 0 and radius  $r$ . To apply results from the current chapter, we need to extend them to a function  $g$  on  $\mathbb{R}^d$  with controlled squared Sobolev norm with order  $t = 1$ , that is,  $\int_{\mathbb{R}^d} (|g(x)|^2 + r^2 \|g'(x)\|_2^2) dx$ . Then, the estimation rates for Sobolev space of order  $t$ , that is,  $O(n^{-1/(1+d)})$ , applies to Lipschitz-continuous functions on an Euclidean ball.

For this, we also need to impose a bound on the value of  $f$  at 0, that is, we assume  $|f(0)| \leq rD$ , and  $f$  is  $D$ -Lipschitz-continuous on the ball of center 0 and radius  $r$ . We now show that we can extend it to a function  $g$  with squared Sobolev norm less than a constant  $c_d$  (that depends on  $d$ ) times  $R^{d+2}D^2$ .

We define the function  $g$  which is equal to  $f$  on the ball of radius  $r$ , equal to 0 outside of the ball of radius  $2r$ , and equal to  $g(x) = f(rx/\|x\|_2)(2 - \|x\|_2/r)$  for  $\|x\|_2 \in [r, 2r]$ , that is, on each ray  $\{ty, t \in [r, 2r]\}$ , for  $y \in \mathbb{R}^d$  of unit norm, the function  $g$  goes linearly from  $f(y)$  to 0. The function  $g$  is continuous and has almost everywhere bounded derivatives. On the ball of radius  $2r$ ,  $|g(x)| \leq 2rD$ , while when  $\|x\|_2 \in [r, 2r]$ ,  $g'(x) = -\frac{1}{r}f(rx/\|x\|_2)x/\|x\|_2 + \frac{r}{\|x\|_2}(I - xx^\top/\|x\|_2^2)f'(rx/\|x\|_2)(2 - \|x\|_2/r)$ , leading to, by the Pythagorean theorem,  $\|g'(x)\|_2^2 = \frac{1}{r^2}|f(rx/\|x\|_2)|^2 + \frac{r^2}{\|x\|_2^2}(2 - \|x\|_2/r)^2\|(I - xx^\top/\|x\|_2^2)f'(rx/\|x\|_2)\|_2^2 \leq \frac{1}{r^2}|2rD|^2 + D^2 = 5D^2$ . Thus,  $\int_{\mathbb{R}^d} (|g(x)|^2 + r^2 \|g'(x)\|_2^2) dx \leq 9r^2D^2(2r)^d \frac{\pi^{d/2}}{\Gamma(1+d/2)}$ , since the volume of the Euclidean unit ball is equal to  $\frac{\pi^{d/2}}{\Gamma(1+d/2)}$ . Thus the constant  $c_d$  is less than  $\frac{9 \cdot 2^d \pi^{d/2}}{\Gamma(1+d/2)}$ .

## 7.6 Theoretical analysis of ridge regression (◆)

In this section, we provide finer results for ridge regression (that is, square loss and penalization by squared norm) used within kernel methods. Compared to the analysis performed in Section 3.6, there are three difficulties:

- (1) we go from fixed design to random design: this will require finer probabilistic arguments to relate population and empirical covariance operators,
- (2) we need to go infinite-dimensional: in terms of notations, this will mean not using transposes of matrices, but dot-products, which is a minor modification,
- (3) the infimum of the expected risk over linear functions parameterized by  $\theta \in \mathcal{H}$  may not be attained by an element of  $\mathcal{H}$ , but by an element of its closure in  $L_2(p)$ . This is important, as this allows access to a potentially large set of functions and requires more care.

### 7.6.1 Kernel ridge regression as a “linear” estimator

We consider  $n$  i.i.d. observations  $(x_i, y_i) \in \mathcal{X} \times \mathbb{R}$ ,  $i = 1, \dots, n$ , and we aim at minimizing, for  $\lambda > 0$ ,

$$\frac{1}{n} \sum_{i=1}^n (y_i - f(x_i))^2 + \lambda \|f\|_{\mathcal{H}}^2.$$

Like local averaging methods in Chapter 6, the ridge regression estimator happens to be a “linear” estimator that depends linearly on the response vector (but of course non-linearly in  $x$  in general). Indeed, using the representer theorem from Eq. (7.2), the estimator is  $f(x) = \sum_{i=1}^n \alpha_i k(x, x_i)$ , with  $\alpha \in \mathbb{R}^n$  defined as  $\alpha = (K + n\lambda I)^{-1}y$ , where  $K \in \mathbb{R}^{n \times n}$  is the kernel matrix. We can then write

$$f(x) = \sum_{i=1}^n \hat{w}_i(x) y_i,$$

with  $\hat{w}(x) = (K + n\lambda I)^{-1}q(x) \in \mathbb{R}^n$ , where  $q(x) \in \mathbb{R}^n$  is defined as  $q_i(x) = k(x, x_i)$ . The smoothing matrix  $H$  is then equal to  $H = K(K + n\lambda I)^{-1}$ .

The key differences with local averaging are that (a) the weights do not sum to one, that is,  $\sum_{i=1}^n \hat{w}_i(x)$  may be different from one, and (b) the weights are not constrained to be non-negative. While the first difference can be removed using centering (see exercise below), the second one is more fundamental: allowing the weights to be negative will enable the adaptivity to smoothness, which local averaging methods missed (see Section 6.5).

**Exercise 7.13** We consider the optimization problem  $\frac{1}{2n} \|y - \Phi\theta - \eta \mathbf{1}_n\|_2^2 + \frac{\lambda}{2} \|\theta\|_2^2$ , where  $\Phi \in \mathbb{R}^{n \times d}$  is the design matrix obtained from feature map  $\varphi$  and data points  $x_1, \dots, x_n$ , and  $y \in \mathbb{R}^n$ , and  $\mathbf{1}_n \in \mathbb{R}^n$  is the vector of all ones. Show that the optimal value of  $\theta$  and  $\eta$  are:  $\theta = \Phi^\top \alpha$ , and  $\eta = \frac{1}{n} \mathbf{1}_n^\top (y - \Phi\theta)$ , with  $\alpha = \Pi_n (\Pi_n K \Pi_n + n\lambda I)^{-1} \Pi_n y$ , and  $\Pi_n = I - \frac{1}{n} \mathbf{1}_n \mathbf{1}_n^\top$ . Show that the prediction function  $f(x) = \varphi(x)^\top \theta + \eta$  is of the form  $\sum_{i=1}^n \hat{w}_i(x) y_i$  with weights that sum to one.

**Exercise 7.14 (♦)** For  $x_1, \dots, x_n$  equally spaced in  $[0, 1]$  and for a translation-invariant kernel from Section 7.3.2, compute the eigenvalues of the kernel matrix and the smoothing matrix.

## 7.6.2 Bias and variance decomposition (♦)

**Beyond fixed-design finite-dimensional analysis.** In Chapter 3, we considered ridge regression in the fixed design setting (where the input data were assumed deterministic) and a finite-dimensional feature space  $\mathcal{H}$ , and obtained in Prop. 3.7 the following *exact* expression of the excess risk of the ridge regression estimator  $\hat{\theta}_\lambda$ , assuming  $y_i = \langle \theta_*, \varphi(x_i) \rangle + \varepsilon_i$ , with  $\varepsilon_i$  independent from  $x_i$ , and where  $\mathbb{E}[\varepsilon_i] = 0$ ,  $\mathbb{E}[\varepsilon_i^2] = \sigma^2$ :

$$\mathbb{E}[(\hat{\theta}_\lambda - \theta_*)^\top \hat{\Sigma}(\hat{\theta}_\lambda - \theta_*)] = \lambda^2 \theta_*^\top (\hat{\Sigma} + \lambda I)^{-2} \hat{\Sigma} \theta_* + \frac{\sigma^2}{n} \text{tr} [\hat{\Sigma}^2 (\hat{\Sigma} + \lambda I)^{-2}]. \quad (7.10)$$

For the random design assumption (the usual machine learning setting), we first need to obtain a value for the expected risk. Moreover, we need to replace the matrix notation to apply to infinite dimensional  $\mathcal{H}$  where the minimizer has a potentially infinite norm.

**Modeling assumptions.** We assume that

$$y_i = f^*(x_i) + \varepsilon_i,$$

with for simplicity  $\mathbb{E}[\varepsilon_i | x_i] = 0$ , and  $\mathbb{E}[\varepsilon_i^2 | x_i] \leq \sigma^2$  almost surely, for some target function  $f^* \in L_2(p)$ , so that  $f^*(x) = \mathbb{E}[y|x]$  is exactly the conditional expectation of  $y|x$ .



The target function  $f^*$  may not be in  $\mathcal{H}$ . All dot-products will always be in  $\mathcal{H}$ , while we will specify the corresponding space for norms.

We thus consider the optimization problem:

$$\min_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n (y_i - f(x_i))^2 + \lambda \|f\|_{\mathcal{H}}^2, \quad (7.11)$$

with the solution found with algorithms in Section 7.4.



The theoretical analysis of kernel methods typically does not involve the parameters  $\alpha \in \mathbb{R}^n$  obtained from the representer theorem.

We have, with  $\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^n \varphi(x_i) \otimes \varphi(x_i)$  a self-adjoint operator from  $\mathcal{H}$  to  $\mathcal{H}$  (the empirical covariance operator), a cost function equal to

$$\frac{1}{n} \sum_{i=1}^n y_i^2 + \langle f, \hat{\Sigma} f \rangle - 2 \left\langle \frac{1}{n} \sum_{i=1}^n y_i \varphi(x_i), f \right\rangle + \lambda \langle f, f \rangle,$$

leading to the minimizer  $\hat{f}_\lambda$  of Eq. (7.11) equal to:

$$\hat{f}_\lambda = (\hat{\Sigma} + \lambda I)^{-1} \frac{1}{n} \sum_{i=1}^n y_i \varphi(x_i) = (\hat{\Sigma} + \lambda I)^{-1} \frac{1}{n} \sum_{i=1}^n f^*(x_i) \varphi(x_i) + (\hat{\Sigma} + \lambda I)^{-1} \frac{1}{n} \sum_{i=1}^n \varepsilon_i \varphi(x_i).$$

We can now compute the (expected) excess risk equal to  $\mathbb{E}[\|\hat{f}_\lambda - f^*\|_{L_2(p)}^2]$  as (using that  $\mathbb{E}(\varepsilon_i | x_i) = 0$ ):

$$\begin{aligned} & \mathbb{E}[\|\hat{f}_\lambda - f^*\|_{L_2(p)}^2] \\ &= \mathbb{E}\left[\left\|(\hat{\Sigma} + \lambda I)^{-1} \frac{1}{n} \sum_{i=1}^n \varepsilon_i \varphi(x_i)\right\|_{L_2(p)}^2\right] + \mathbb{E}\left[\left\|(\hat{\Sigma} + \lambda I)^{-1} \frac{1}{n} \sum_{i=1}^n f^*(x_i) \varphi(x_i) - f^*\right\|_{L_2(p)}^2\right]. \end{aligned}$$

The first term is the usual **variance** term (that depends on the noise on top of the optimal predictions). In contrast, the second is the **(squared) bias** term (which depends on the regularity of the target function). Before developing the probabilistic argument, we give simplified upper bounds of the two terms.

On top of the non-centered empirical covariance operator  $\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^n \varphi(x_i) \otimes \varphi(x_i)$ , we will need its expectation, the covariance operator (from  $\mathcal{H}$  to  $\mathcal{H}$ )

$$\Sigma = \mathbb{E}[\varphi(x) \otimes \varphi(x)],$$

for the corresponding distribution of the  $x_i$ 's. A key property relates the  $L_2(p)$ -norm and the RKHS norm, that is, that for  $g \in \mathcal{H}$ ,

$$\begin{aligned} \|g\|_{L_2(p)}^2 &= \int_{\mathcal{X}} g(x)^2 dp(x) = \int_{\mathcal{X}} \langle g, \varphi(x) \rangle^2 dp(x) = \int_{\mathcal{X}} \langle g, \varphi(x) \otimes \varphi(x) g \rangle dp(x) \\ &= \langle g, \Sigma g \rangle = \|\Sigma^{1/2} g\|_{\mathcal{H}}^2. \end{aligned} \tag{7.12}$$

**Variance term.** Starting from **variance** =  $\mathbb{E}\left[\left\|(\hat{\Sigma} + \lambda I)^{-1} \frac{1}{n} \sum_{i=1}^n \varepsilon_i \varphi(x_i)\right\|_{L_2(p)}^2\right]$ , the variance term can be upper-bounded as follows (first using independence and zero means of the variables  $\varepsilon_i$ ). Below, we use the property that for symmetric matrices such that  $A \succcurlyeq 0$  and  $B \preccurlyeq C$ , we have  $\text{tr}[AB] \leq \text{tr}[AC]$ , and Eq. (7.12):

$$\begin{aligned} & \mathbb{E}\left[\left\|(\hat{\Sigma} + \lambda I)^{-1} \frac{1}{n} \sum_{i=1}^n \varepsilon_i \varphi(x_i)\right\|_{L_2(p)}^2\right] \\ &= \frac{1}{n^2} \sum_{i=1}^n \mathbb{E}\left[\text{tr}\left((\hat{\Sigma} + \lambda I)^{-1} \Sigma (\hat{\Sigma} + \lambda I)^{-1} \varepsilon_i^2 \varphi(x_i) \otimes \varphi(x_i)\right)\right] \\ &\leq \frac{\sigma^2}{n} \mathbb{E}\left[\text{tr}\left((\hat{\Sigma} + \lambda I)^{-1} \Sigma (\hat{\Sigma} + \lambda I)^{-1} \hat{\Sigma}\right)\right] \text{ using } \mathbb{E}[\varepsilon_i^2 | x_i] \leq \sigma^2, \\ &\leq \frac{\sigma^2}{n} \mathbb{E}\left[\text{tr}\left((\hat{\Sigma} + \lambda I)^{-1} \Sigma\right)\right] \text{ using } (\hat{\Sigma} + \lambda I)^{-1} \hat{\Sigma} \preccurlyeq I. \end{aligned} \tag{7.13}$$

This will be the main expression we will bound later.

**Bias term.** We first assume that  $f^* \in \mathcal{H}$ , that is, the model is well-specified. Then, writing  $f^*(x_i) = \langle f^*, \varphi(x_i) \rangle$  (which is possible because  $f^* \in \mathcal{H}$ ), the bias term is equal to

$$\text{bias} = \mathbb{E} \left[ \left\| (\widehat{\Sigma} + \lambda I)^{-1} \frac{1}{n} \sum_{i=1}^n f^*(x_i) \varphi(x_i) - f^* \right\|_{L_2(p)}^2 \right] \quad (7.14)$$

$$= \mathbb{E} \left[ \left\| (\widehat{\Sigma} + \lambda I)^{-1} \frac{1}{n} \sum_{i=1}^n \langle f^*, \varphi(x_i) \rangle \varphi(x_i) - f^* \right\|_{L_2(p)}^2 \right] \quad (7.15)$$

$$= \mathbb{E} \left[ \left\| (\widehat{\Sigma} + \lambda I)^{-1} \widehat{\Sigma} f^* - f^* \right\|_{L_2(p)}^2 \right] \\ = \mathbb{E} \left[ \left\| \lambda \Sigma^{1/2} (\widehat{\Sigma} + \lambda I)^{-1} f^* \right\|_{\mathcal{H}}^2 \right] = \lambda^2 \mathbb{E} \left[ \langle f^*, (\widehat{\Sigma} + \lambda I)^{-1} \Sigma (\widehat{\Sigma} + \lambda I)^{-1} f^* \rangle \right], \quad (7.16)$$

where we have used Eq. (7.12) above to re-introduce the operator  $\Sigma$ . This will be the main expression we will bound later.

**Upper-bound on excess risk.** We have thus shown the following proposition:

**Proposition 7.5** *When  $f^* \in \mathcal{H}$ , the excess risk of the ridge regression estimator is upper-bounded by:*

$$\mathbb{E} [\| \hat{f}_\lambda - f^* \|_{L_2(p)}^2] \leq \frac{\sigma^2}{n} \mathbb{E} \left[ \text{tr} ((\widehat{\Sigma} + \lambda I)^{-1} \Sigma) \right] + \lambda^2 \mathbb{E} \left[ \langle f^*, (\widehat{\Sigma} + \lambda I)^{-1} \Sigma (\widehat{\Sigma} + \lambda I)^{-1} f^* \rangle \right]. \quad (7.17)$$

Given the expression of the expected variance in Eq. (7.13) and the expected bias in Eq. (7.16), we notice that both the empirical and expected covariance operators appear and that it would be important to replace the empirical one with the expected one. This is possible with extra multiplicative factors, which we now show. Then, we will bound the two terms separately and show how balancing them leads to interesting learning bounds.

### 7.6.3 Relating empirical and population covariance operators

We follow Mourtada and Rosasco (2022) and derive simple relationships between the empirical covariance operator  $\widehat{\Sigma}$  and the population operator  $\Sigma$ , by showing the following lemma dealing with expectations; for high probability bounds, see, e.g., Rudi et al. (2015); Rudi and Rosasco (2017), as well as the end of Section 7.6.4.

**Lemma 7.1 (Mourtada and Rosasco, 2022)** *Assuming i.i.d. data  $x_1, \dots, x_n \in \mathcal{X}$ , and bounded features  $\|\varphi(x)\|_{\mathcal{H}} \leq R$  for all  $x \in \mathcal{X}$ ; we have, for all  $g \in \mathcal{H}$ :*

$$\mathbb{E} \left[ \text{tr} ((\widehat{\Sigma} + \lambda I)^{-1} \Sigma) \right] \leq \left( 1 + \frac{R^2}{\lambda n} \right) \text{tr} ((\Sigma + \lambda I)^{-1} \Sigma) \quad (7.18)$$

$$\mathbb{E} \left[ \langle g, (\widehat{\Sigma} + \lambda I)^{-1} \Sigma (\widehat{\Sigma} + \lambda I)^{-1} g \rangle \right] \preceq \lambda^{-1} \left( 1 + \frac{R^2}{\lambda n} \right)^2 \langle g, (\Sigma + \lambda I)^{-1} \Sigma g \rangle. \quad (7.19)$$

**Proof (♦)** The main idea is to introduce a  $(n+1)$ -th independent observation from the same distribution, write  $\Sigma = \mathbb{E} [\varphi(x_{n+1}) \otimes \varphi(x_{n+1})]$ , and use the fact that the obser-

uations are “exchangeable”, that is, they can be permuted without changing their joint distribution.

We denote  $C = \sum_{i=1}^{n+1} \varphi(x_i) \otimes \varphi(x_i)$ , and using the matrix inversion lemma (Section 1.1.3), we have

$$\begin{aligned} (C + n\lambda I)^{-1} \varphi(x_{n+1}) &= (n\widehat{\Sigma} + n\lambda I + \varphi(x_{n+1}) \otimes \varphi(x_{n+1}))^{-1} \varphi(x_{n+1}) \\ &= \frac{(n\widehat{\Sigma} + n\lambda I)^{-1} \varphi(x_{n+1})}{1 + \langle \varphi(x_{n+1}), (n\widehat{\Sigma} + n\lambda I)^{-1} \varphi(x_{n+1}) \rangle}. \end{aligned} \quad (7.20)$$

Finally, we will use  $c = \langle \varphi(x_{n+1}), (n\widehat{\Sigma} + n\lambda I)^{-1} \varphi(x_{n+1}) \rangle \leq \frac{R^2}{\lambda n}$ . To prove Eq. (7.18), we use Eq. (7.20) above to express  $(\widehat{\Sigma} + \lambda I)^{-1} \varphi(x_{n+1})$  in terms of  $(C + n\lambda I)^{-1} \varphi(x_{n+1})$ , to get:

$$\begin{aligned} \mathbb{E} \left[ \text{tr} \left( (\widehat{\Sigma} + \lambda I)^{-1} \Sigma \right) \right] &= \mathbb{E} \left[ \text{tr} \left( (\widehat{\Sigma} + \lambda I)^{-1} \varphi(x_{n+1}) \otimes \varphi(x_{n+1}) \right) \right] \\ &= \mathbb{E} \left[ \langle \varphi(x_{n+1}), (\widehat{\Sigma} + \lambda I)^{-1} \varphi(x_{n+1}) \rangle \right] \\ &= n \mathbb{E} \left[ (1 + c) \langle \varphi(x_{n+1}), (C + n\lambda I)^{-1} \varphi(x_{n+1}) \rangle \right], \end{aligned}$$

which leads to  $\mathbb{E} \left[ \text{tr} \left( (\widehat{\Sigma} + \lambda I)^{-1} \Sigma \right) \right] \leq (1 + \frac{R^2}{\lambda n}) \mathbb{E} \left[ \langle \varphi(x_{n+1}), (C + n\lambda I)^{-1} \varphi(x_{n+1}) \rangle \right]$ . Thus, using that the variables  $(x_1, \dots, x_{n+1})$  are exchangeable:

$$\begin{aligned} &\mathbb{E} \left[ \text{tr} \left( (\widehat{\Sigma} + \lambda I)^{-1} \Sigma \right) \right] \\ &\leq \left( 1 + \frac{R^2}{\lambda n} \right) \frac{1}{n+1} \sum_{i=1}^{n+1} \mathbb{E} \left[ \langle \varphi(x_i), (C + n\lambda I)^{-1} \varphi(x_i) \rangle \right] \\ &= \left( 1 + \frac{R^2}{\lambda n} \right) \frac{1}{n+1} \mathbb{E} \left[ \text{tr} \left( C (C + n\lambda I)^{-1} \right) \right] \text{ since } C = \sum_{i=1}^{n+1} \varphi(x_i) \otimes \varphi(x_i) \\ &\leq \left( 1 + \frac{R^2}{\lambda n} \right) \frac{1}{n+1} \left[ \text{tr} \left( \mathbb{E}[C] (\mathbb{E}[C] + n\lambda I)^{-1} \right) \right] \text{ by Jensen's inequality, }^{14} \\ &= \left( 1 + \frac{R^2}{\lambda n} \right) \frac{1}{n+1} \text{tr} \left( (n+1) \Sigma ((n+1) \Sigma + n\lambda I)^{-1} \right) \\ &\leq \left( 1 + \frac{R^2}{\lambda n} \right) \text{tr} \left( \Sigma (\Sigma + \lambda I)^{-1} \right), \text{ which is exactly Eq. (7.18).} \end{aligned}$$

To prove Eq. (7.19), we use the same technique, that is,

$$\begin{aligned} &\mathbb{E} \left[ (\widehat{\Sigma} + \lambda I)^{-1} \Sigma (\widehat{\Sigma} + \lambda I)^{-1} \right] = \mathbb{E} \left[ (\widehat{\Sigma} + \lambda I)^{-1} \varphi(x_n) \otimes \varphi(x_n) (\widehat{\Sigma} + \lambda I)^{-1} \right] \\ &= n^2 (1 + c)^2 \left[ (C + n\lambda I)^{-1} \varphi(x_{n+1}) \right] \otimes \left[ (C + n\lambda I)^{-1} \varphi(x_{n+1}) \right]. \end{aligned}$$

This leads to:

$$\begin{aligned}
& \mathbb{E}[\langle g, (\widehat{\Sigma} + \lambda I)^{-1} \Sigma (\widehat{\Sigma} + \lambda I)^{-1} g \rangle] \\
&= n^2 \mathbb{E}[(1+c)^2 \langle (C + n\lambda I)^{-1} \varphi(x_{n+1}), g \rangle^2] \\
&\leq n^2 \left(1 + \frac{R^2}{\lambda n}\right)^2 \mathbb{E}[\langle (C + n\lambda I)^{-1} \varphi(x_{n+1}), g \rangle^2] \\
&= \frac{n^2}{n+1} \left(1 + \frac{R^2}{\lambda n}\right)^2 \mathbb{E}[\langle g, (C + n\lambda I)^{-1} C (C + n\lambda I)^{-1} g \rangle] \text{ by exchangeability,} \\
&\leq \frac{1}{\lambda} \frac{n}{n+1} \left(1 + \frac{R^2}{\lambda n}\right)^2 \mathbb{E}[\langle g, C (C + n\lambda I)^{-1} g \rangle] \\
&\leq \frac{1}{\lambda} \frac{n}{n+1} \left(1 + \frac{R^2}{\lambda n}\right)^2 \langle g, \mathbb{E}[C] (\mathbb{E}[C] + n\lambda I)^{-1} g \rangle \text{ by Jensen's inequality,} \\
&= \frac{1}{\lambda} n \left(1 + \frac{R^2}{\lambda n}\right)^2 \langle g, \Sigma((n+1)\Sigma + n\lambda I)^{-1} g \rangle \leq \lambda^{-1} \left(1 + \frac{R^2}{\lambda n}\right)^2 \langle g, (\Sigma + \lambda I)^{-1} \Sigma g \rangle.
\end{aligned}$$

■

#### 7.6.4 Analysis for well-specified problems (◆)

In this section, we assume that  $f^* \in \mathcal{H}$ . We have the following result for the excess risk, whose proof consists in applying Lemma 7.1 to Eq. (7.17).

**Proposition 7.6 (Well-specified model kernel ridge regression)** *Assume i.i.d. data  $(x_i, y_i) \in \mathcal{X} \times \mathbb{R}$ , for  $i = 1, \dots, n$ , and  $y_i = f^*(x_i) + \varepsilon_i$ , with  $\mathbb{E}[\varepsilon_i | x_i] = 0$  and  $\mathbb{E}[\varepsilon_i^2 | x_i] \leq \sigma^2$ , and  $f^* \in \mathcal{H}$ . Assume  $\|\varphi(x)\|_{\mathcal{H}} \leq R$ . We have:*

$$\mathbb{E}[\|\hat{f}_\lambda - f^*\|_{L_2(p)}^2] \leq \frac{\sigma^2}{n} \left(1 + \frac{R^2}{\lambda n}\right) \text{tr}((\Sigma + \lambda I)^{-1} \Sigma) + \lambda \left(1 + \frac{R^2}{\lambda n}\right)^2 \langle f^*, \Sigma(\Sigma + \lambda I)^{-1} f^* \rangle. \quad (7.21)$$

This is to be contrasted with Eq. (7.10): we obtain a similar result with  $\widehat{\Sigma}$  replaced by  $\Sigma$ , but with some extra multiplicative constants that are close to one if  $R^2/(\lambda n)$  is small. We can further bound  $\text{tr}((\Sigma + \lambda I)^{-1} \Sigma) \leq \frac{R^2}{\lambda}$  and  $\langle f^*, \Sigma(\Sigma + \lambda I)^{-1} f^* \rangle \leq \langle f^*, f^* \rangle$ , to get the bound

$$\mathbb{E}[\|\hat{f}_\lambda - f^*\|_{L_2(p)}^2] \leq \sigma^2 \frac{R^2}{\lambda n} \left(1 + \frac{R^2}{\lambda n}\right) + \lambda \left(1 + \frac{R^2}{\lambda n}\right)^2 \|f^*\|_{\mathcal{H}}^2,$$

which is a random design version of the developments in the proof of Prop. 3.8. In such a situation, the choice  $\lambda = R^2/\sqrt{n}$  (which does not impose any knowledge of  $\|f^*\|_{\mathcal{H}}$ ) leads to a bound on the excess risk proportional to  $(\sigma^2 + R^2 \|f^*\|_{\mathcal{H}}^2)/\sqrt{n}$ .

In finite feature dimensions  $d$ , we can alternatively bound  $\text{tr}((\Sigma + \lambda I)^{-1} \Sigma) \leq \frac{R^2}{\lambda}$  by  $d$ , then leading to a natural choice  $\lambda$  of order  $R^2/n$ , and an upper-bound of the excess risk proportional to  $\sigma^2 d/n + R^2 \|f^*\|_{\mathcal{H}}^2/n$ .

The last two paragraphs lead to different choices of the regularization parameter, proportional to  $R^2/\sqrt{n}$  or  $R^2/n$ , two classical rules of thumbs within kernel methods.



**Bounds in high-probability (◆◆).** Instead of obtaining bounds in expectation (with respect to the training data), we can obtain high-probability bounds, as briefly shown below for the simplest bound; see, more refined bounds by [Rudi et al. \(2015\)](#); [Rudi and Rosasco \(2017\)](#). Note that they do not rely on Rademacher averages but on direct probabilistic arguments that can only be applied to the square loss.

**Proposition 7.7 (High-probability bound for kernel ridge regression)** *Assume i.i.d. data  $(x_i, y_i) \in \mathcal{X} \times \mathbb{R}$ , for  $i = 1, \dots, n$ , and  $y_i = f^*(x_i) + \varepsilon_i$ , with  $\mathbb{E}[\varepsilon_i|x_i] = 0$  and  $\varepsilon_i^2 \leq \sigma^2$  almost surely, and  $f^* \in \mathcal{H}$ . Assume  $\|\varphi(x)\|_{\mathcal{H}} \leq R$  and  $n \geq (\frac{4}{3} + \frac{R^2}{8\lambda}) \log \frac{14R^2}{\lambda\delta}$ . We have, with a probability greater than  $1 - \delta$ ,*

$$\|\hat{f}_\lambda - f^*\|_{L_2(p)}^2 \leq \frac{8\sigma^2 R^2}{\lambda n} + 4\lambda \|f^*\|_{\mathcal{H}}^2 + \frac{16\sigma^2 R^2}{\lambda n} \log \frac{2}{\delta}. \quad (7.22)$$

**Proof** We first apply Prop. 1.7 with  $M_i = \Sigma(\Sigma + \lambda I)^{-1} - (\Sigma + \lambda I)^{-1/2} \varphi(x_i) \otimes \varphi(x_i) (\Sigma + \lambda I)^{-1/2}$ , for which we have  $V = \frac{R^2}{\lambda} \Sigma(\Sigma + \lambda I)^{-1}$ ,  $\sigma^2 = \frac{R^2}{\lambda}$ ,  $c = 1$ , and  $t = 1$ , leading to

$$\lambda_{\max}[(\Sigma + \lambda I)^{-1/2}(\Sigma - \hat{\Sigma})(\Sigma + \lambda I)^{-1/2}] \leq \frac{1}{2}$$

with probability greater than  $1 - 7 \frac{R^2}{\lambda} \exp[-\frac{n}{4/3 + R^2/(8\lambda)}]$ , as soon as  $\frac{1}{2} \geq \frac{1}{3n} + \frac{R}{\sqrt{\lambda n}}$ . This probability is greater than  $1 - \delta/2$  as soon as  $n \geq (4/3 + R^2/(8\lambda)) \log \frac{14R^2}{\lambda\delta}$ .

This implies  $\Sigma - \hat{\Sigma} \preceq \frac{1}{2}(\Sigma + \lambda I)$ ,  $\frac{1}{2}(\Sigma + \lambda I) \preceq \hat{\Sigma} + \lambda I$ , and thus  $(\hat{\Sigma} + \lambda I)^{-1} \preceq 2(\Sigma + \lambda I)^{-1}$ . Using the Łojasiewicz inequality (Lemma 5.1) on the regularized empirical risk  $\hat{\mathcal{R}}_\lambda(f) = \frac{1}{2n} \langle f - f^*, \hat{\Sigma}(f - f^*) \rangle - \langle \frac{1}{n} \sum_{i=1}^m \varepsilon_i \varphi(x_i), f \rangle + \frac{\lambda}{2} \|f\|_{\mathcal{H}}^2$ , we get:

$$\hat{\mathcal{R}}_\lambda(f^*) - \hat{\mathcal{R}}_\lambda(\hat{f}_\lambda) \leq \frac{1}{2\lambda} \|\hat{\mathcal{R}}'_\lambda(f^*)\|_{\mathcal{H}}^2.$$

Using  $\hat{\mathcal{R}}_\lambda(f^*) - \hat{\mathcal{R}}_\lambda(\hat{f}_\lambda) = \frac{1}{2} \langle f^* - \hat{f}_\lambda, (\hat{\Sigma} + \lambda I)(f^* - \hat{f}_\lambda) \rangle \geq \frac{1}{4} \langle f^* - \hat{f}_\lambda, (\Sigma + \lambda I)(f^* - \hat{f}_\lambda) \rangle = \frac{1}{4} \|\hat{f}_\lambda - f^*\|_{L_2(p)}^2$ , we get

$$\|\hat{f}_\lambda - f^*\|_{L_2(p)}^2 \leq \frac{2}{\lambda} \left\| \frac{1}{n} \sum_{i=1}^m \varepsilon_i \varphi(x_i) + \lambda f^* \right\|_{\mathcal{H}}^2 \leq \frac{4}{\lambda} \left\| \frac{1}{n} \sum_{i=1}^n \varepsilon_i \varphi(x_i) \right\|_{\mathcal{H}}^2 + 4\lambda \|f^*\|_{\mathcal{H}}^2.$$

We thus need a high-probability bound for  $\left\| \frac{1}{n} \sum_{i=1}^n \varepsilon_i \varphi(x_i) \right\|_{\mathcal{H}}$ , which we can obtain, with probability greater than  $1 - \delta/2$ , from McDiarmid's inequality, as

$$\left\| \frac{1}{n} \sum_{i=1}^n \varepsilon_i \varphi(x_i) \right\|_{\mathcal{H}} \leq \frac{R\sigma}{\sqrt{n}} \left( 1 + \sqrt{2 \log \frac{2}{\delta}} \right).$$

■

Before analyzing the last proposition and balancing bias and variance, we show how this can be applied beyond well-specified models.

### 7.6.5 Analysis beyond well-specified problems (♦)

In the bound in Eq. (7.21), the only term that requires potentially that  $f^* \in \mathcal{H}$  is the bias term  $\lambda \langle f^*, (\Sigma + \lambda I)^{-1} \Sigma f^* \rangle$ . The following simple lemma is the key to extending to all functions  $f^*$  in the closure of  $\mathcal{H}$ .

**Lemma 7.2** *Given the covariance operator  $\Sigma$  and any function  $f^* \in \mathcal{H}$ , then*

$$\lambda \langle f^*, (\Sigma + \lambda I)^{-1} \Sigma f^* \rangle = \inf_{f \in \mathcal{H}} \left\{ \|f - f^*\|_{L_2(p)}^2 + \lambda \|f\|_{\mathcal{H}}^2 \right\}.$$

**Proof** The optimization problem above can be written as  $\inf_{f \in \mathcal{H}} \left\{ \|\Sigma^{1/2}(f - f^*)\|_{\mathcal{H}}^2 + \lambda \|f\|_{\mathcal{H}}^2 \right\}$ , using Eq. (7.12), with solution  $f = (\Sigma + \lambda I)^{-1} \Sigma f^*$  and we can simply put back the value in the objective function to get the desired result. ■

**Target function in the closure of  $\mathcal{H}$ .** By using a limiting argument, we can extend the formula of the bias term in Prop. 7.6 to the general case of  $f^* \in L_2(p)$  with Eq. (7.23), in the closure of  $\mathcal{H}$  in  $L_2(p)$  (because all functions in the closure can be approached by a function in  $\mathcal{H}$ ), leading to

$$\left(1 + \frac{R^2}{\lambda n}\right)^2 \inf_{f \in \mathcal{H}} \left\{ \|f - f^*\|_{L_2(p)}^2 + \lambda \|f\|_{\mathcal{H}}^2 \right\}, \quad (7.23)$$

For translation-invariant kernels in  $\mathbb{R}^d$  (which are dense in  $L_2(\mathbb{R}^d)$ ), this allows estimating any target function.

**Final result.** Combining the two cases above, we can now show the upper bound for kernel ridge regression in the potentially misspecified case.

**Proposition 7.8 (Mis-specified model kernel ridge regression)** *Assume i.i.d. data  $(x_i, y_i) \in \mathcal{X} \times \mathbb{R}$ , for  $i = 1, \dots, n$ , and  $y_i = f^*(x_i) + \varepsilon_i$ , with  $\mathbb{E}[\varepsilon_i | x_i] = 0$  and  $\mathbb{E}[\varepsilon_i^2 | x_i] \leq \sigma^2$ . Assume  $\|\varphi(x)\|_{\mathcal{H}} \leq R$  and  $f^*$  in the closure of  $\mathcal{H}$  in  $L_2(p)$ . We have:*

$$\mathbb{E}[\|\hat{f}_\lambda - f^*\|_{L_2(p)}^2] \leq \frac{\sigma^2}{n} \left(1 + \frac{R^2}{\lambda n}\right) \text{tr}((\Sigma + \lambda I)^{-1} \Sigma) + \left(1 + \frac{R^2}{\lambda n}\right)^2 \inf_{f \in \mathcal{H}} \left\{ \|f - f^*\|_{L_2(p)}^2 + \lambda \|f\|_{\mathcal{H}}^2 \right\}. \quad (7.24)$$



Be careful with homogeneity of formulas; e.g.,  $\frac{R^2}{\lambda n}$  is indeed a constant.

### 7.6.6 Balancing bias and variance (♦)

We can now balance the bias and variance term in the following upper-bound on the expected excess risk,

$$\frac{\sigma^2}{n} \left(1 + \frac{R^2}{\lambda n}\right) \text{tr}((\Sigma + \lambda I)^{-1} \Sigma) + \left(1 + \frac{R^2}{\lambda n}\right)^2 \inf_{f \in \mathcal{H}} \left\{ \|f - f^*\|_{L_2(p)}^2 + \lambda \|f\|_{\mathcal{H}}^2 \right\}.$$

For this section, we will assume that  $\mathcal{X} = \mathbb{R}^d$  and that the target function belongs to a Sobolev kernel of order  $t > 0$ , while the RKHS is a Sobolev space of order  $s > d/2$ .

We have seen in Section 7.5.2 that the bias term is of order  $(1 + \frac{R^2}{\lambda n})^2 \lambda^{t/s}$  when  $s \geq t$ . For the variance term, we need to study the so-called “degrees of freedom”.

**Degrees of freedom.** This is the quantity  $\text{tr} [\Sigma(\Sigma + \lambda I)^{-1}]$ , which is decreasing in  $\lambda$ , from  $+\infty$  for  $\lambda = 0$  to 0 for  $\lambda = +\infty$ . If we know that the eigenvalues  $(\lambda_m)_{m \geq 0}$  of the covariance operator satisfy

$$\lambda_m \leq C(m+1)^{-\alpha},$$

for  $\alpha > 1$ , then one has, with the change of variable  $u = \lambda C^{-1} t^\alpha$  below,

$$\begin{aligned} \text{tr} [\Sigma(\Sigma + \lambda I)^{-1}] &= \sum_{m \geq 0} \frac{\lambda_m}{\lambda_m + \lambda} \leq \sum_{m \geq 0} \frac{1}{1 + \lambda C^{-1} (m+1)^\alpha} \leq \int_0^\infty \frac{1}{1 + \lambda C^{-1} t^\alpha} \\ &\leq \int_0^\infty \lambda^{-1/\alpha} C^{1/\alpha} \frac{1}{\alpha} u^{1/\alpha-1} \frac{du}{1+u} \leq O(\lambda^{-1/\alpha}). \end{aligned}$$

It turns out that if the distribution of inputs has a bounded density with respect to the Lebesgue measure, then for our chosen Sobolev space, we have  $\alpha = 2s/d$  (see, e.g., [Harchaoui et al., 2008](#), Appendix D).

**Balancing terms (Sobolev spaces).** We thus need to balance  $\lambda^{t/s}$  with  $\frac{1}{n} \lambda^{-1/\alpha}$ , leading to an optimal  $\lambda$  proportional to  $n^{-(1/\alpha+t/s)^{-1}}$ , and a rate proportional to  $n^{-\alpha t/(\alpha t+s)}$ . This rate is only achievable through our analysis when  $\frac{R^2}{n\lambda}$  remains bounded, that is, essentially  $\lambda \geq R^2/n$ , thus,  $\frac{1}{\alpha} + \frac{t}{s} \geq 1$ .

For  $\alpha = 2s/d$ , we obtain the rate  $\frac{1}{n^{2t/(2t+d)}}$ , which is valid as long as  $\frac{d}{2} + t \geq s \geq t$ . We can make the following observations:

- Except for the constraint  $\frac{d}{2} + t \geq s \geq t$ , the upper-bound on the rate obtained after optimizing over  $\lambda$  does not depend on the kernel.
- We obtain some form of adaptivity, that is, the rate improves with the regularity of the target function, from the slow rate  $\frac{1}{n^{2t/(2t+d)}}$  when  $t = 1$  (recovering the same rate as for local averaging methods<sup>15</sup> in Chapter 6), and that can only be achieved when  $s \leq d/2 + 1$ , e.g., with the exponential kernel), to the rate  $\frac{1}{n^{2s/(2s+d)}}$  when  $t = s$ , the rate is then always better than  $1/\sqrt{n}$  because of the constraint  $s > d/2$ .
- In order to allow for regularization parameters  $\lambda$  which are less than  $1/n$ , other assumptions are needed. See, e.g., [Pillaud-Vivien et al. \(2018\)](#) and references therein.

<sup>15</sup>In Chapter 6, we assumed the target function to be Lipschitz-continuous, which can be made an element of the Sobolev space of order  $t = 1$ , with the construction at the end of Section 7.5.2.

## 7.7 Experiments

We consider one-dimensional problems to highlight the adaptivity of kernel methods to the regularity of the target function, with one smooth target and one non-smooth target, and three kernels: exponential kernel corresponding to the Sobolev space of order 1 (top of Figure 7.3), Matern kernel corresponding to the Sobolev space of order 3 (middle), and Gaussian kernel (bottom). In the right plots, dotted lines are affine fits to the log-log learning curves. The regularization parameter for ridge regression is selected to minimize expected risk, and learning curves are obtained by averaging over 20 replications. See results in Figure 7.3. The data

We observe adaptivity for the three kernels: learning is possible even with irregular functions, and the rates are better for smooth target functions. We also note that for kernels with smaller feature spaces (Matern and Gaussian), the performance on the non-smooth target function is worse than for the large feature space (exponential kernel). As highlighted by [Bach \(2013\)](#), this drop in performance is primarily due to a numerical issue (the eigenvalues of the kernel matrix decay exponentially fast, and finite precision arithmetic prevents the use of regularization parameters that are too small).

## 7.8 Conclusion

In this chapter, we have shown how models that are linear in their parameters can be made infinite-dimensional. Algorithmically, this is made possible using the kernel trick that uses only dot-products between the feature maps. Statistically, this leads to models that can adapt to complex prediction functions using the appropriate kernels.

Since algorithms presented in Section 7.4 rely on convex optimization, we obtain precise generalization guarantees that can take into account, estimation, approximation, and optimization errors. A key benefit of positive-definite kernel methods compared to local-averaging techniques is the adaptivity to the smoothness of the prediction function. What is still missing is adaptivity to problems where the optimal prediction function only depends on a subset of the original variables (when applying to inputs in  $\mathbb{R}^d$ ). This will be achieved by neural networks in Chapter 9 at the expense of non-convex optimization problems.

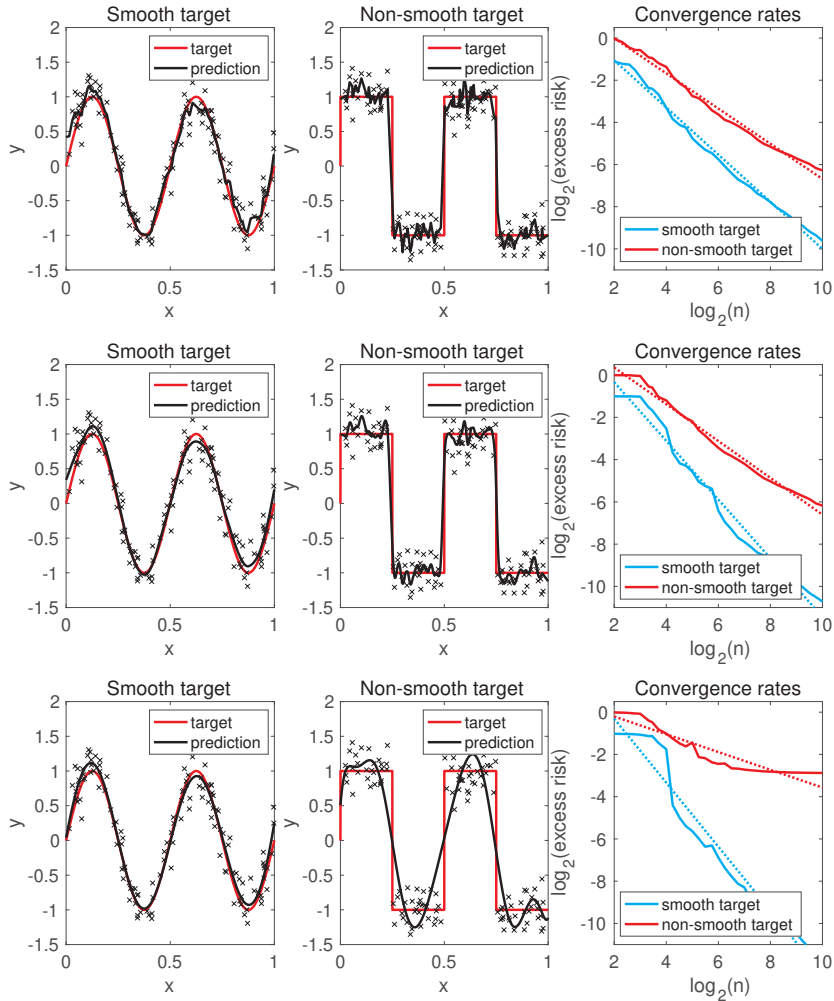


Figure 7.3: Comparison of three kernels, Sobolev space of order 1 (top), Matern kernel corresponding to the Sobolev space of order 3 (middle), and Gaussian kernel (bottom). We consider two different target functions and plot on the right plots the excess risks.