4 Regularization

Minimizing the empirical risk can lead to numerical instabilities and bad generalization performance. A possible way to avoid this problem is to restrict the class of admissible solutions, for instance to a compact set. This technique was introduced by Tikhonov and Arsenin [538] for solving inverse problems and has since been applied to learning problems with great success. In statistics, the corresponding estimators are often referred to as *shrinkage estimators* [262].

Kernel methods are best suited for two special types of regularization: a coefficient space constraint on the *expansion coefficients* of the weight vector in feature space [343, 591, 37, 517, 189], or, alternatively, a function space regularization *directly* penalizing the weight vector in feature space [573, 62, 561]. In this chapter we will discuss the connections between regularization, Reproducing Kernel Hilbert Spaces (RKHS), feature spaces, and regularization operators. The connection to Gaussian Processes will be explained in more detail in Section 16.3. These different viewpoints will help us to gain insight into the success of kernel methods.

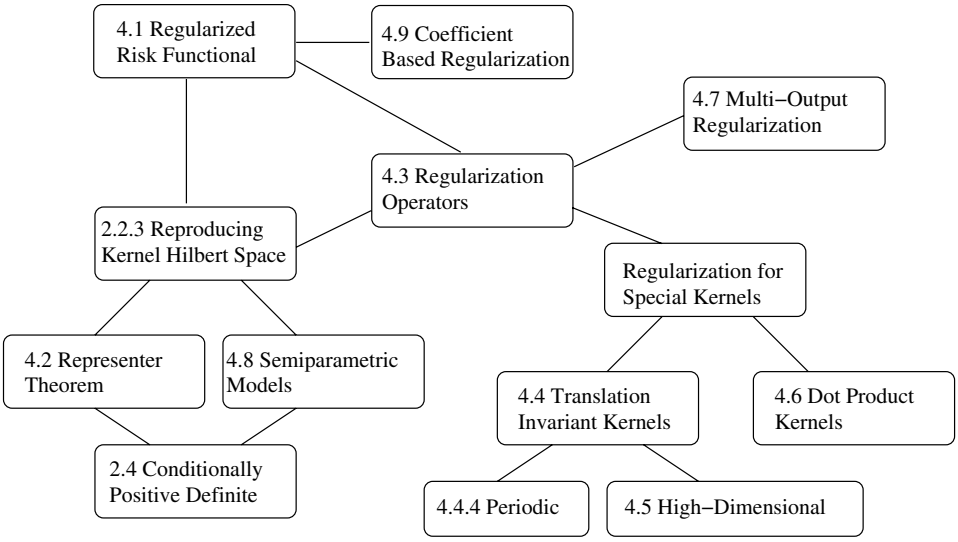
Overview

We start by introducing regularized risk functionals (Section 4.1), followed by a discussion of the Representer Theorem describing the functional form of the minimizers of a certain class of such risk functionals (Section 4.2). Section 4.3 introduces regularization operators and details their connection to SV kernels. Sections 4.4 through 4.6 look at this connection for specific classes of kernels. Following that, we have several sections dealing with various regularization issues of interest for machine learning: vector-valued functions (Section 4.7), semiparametric regularization (Section 4.8), and finally, coefficient-based regularization (Section 4.9).

Prerequisites

This chapter may not be be easy to digest for some of our readers. We recommend that most readers should nevertheless consider going through Sections 4.1 and 4.2. Those two sections are accessible with the background given in Chapters 1 and Chapter 2. The following Section 4.3 is somewhat more technical, since it is using the concept of Green's functions and operators, but should nevertheless still be looked at. A background in functional analysis will be helpful.

Sections 4.4, 4.5, and 4.6 are more difficult, and require a solid knowledge of Fourier integrals and elements of the theory of special functions. To understand Section 4.7, some basic notions of group theory are beneficial. Finally, Sections 4.8 and Section 4.9 do not require additional knowledge beyond the basic concepts put forward in the introductory chapters. Yet, some readers may find it beneficial to read these two last sections after they gained a deeper insight into classification, regression and mathematical programming, as provided by Chapters 6, 7, and 9.



4.1 The Regularized Risk Functional

The key idea in regularization is to restrict the class of possible minimizers $\mathcal{F}$ (with $f \in \mathcal{F}$) of the empirical risk functional $R_{\text{emp}}[f]$ such that $\mathcal{F}$ becomes a compact set. While there exist various characterizations for compact sets and we may define a large variety of such sets which will suit different assumptions on the type of estimates we get, the common key idea is compactness. In addition, we will assume that $R_{\text{emp}}[f]$ is continuous in f.

Continuity Assumption

Note that this is a stronger assumption than it may appear at first glance. It is easily satisfied for many regression problems, such as those using squared loss or the ε-insensitive loss. Yet binary valued loss functions, as are often used in classification (such as $c(x, y, f(x)) = \frac{1}{2}(1 - \operatorname{sgn} yf(x))$), do not meet the requirements. Since both the exact minimization of $R_{\text{emp}}[f]$ for classification problems [367], even with very restricted classes of functions, and also the approximate solution to this problem [20] have been proven to be NP-hard, we will not bother with this case any further, but rather attempt to minimize a continuous approximation of the $0-1$ loss, such as the one using a soft margin loss function (3.3).

We may now apply the operator inversion lemma to show that for compact $\mathcal{F}$, the inverse map from the minimum of the empirical risk functional $R_{\text{emp}}[f] : \mathcal{F} \to \mathbb{R}$ to its minimizer $\hat{f}$ is continuous and the optimization problem well-posed.

Theorem 4.1 (Operator Inversion Lemma (e.g., [431])) *Let X be a compact set and let the map $f : X \to Y$ be continuous. Then there exists an inverse map $f^{-1} : f(X) \to X$ that is also continuous.*

We do not directly specify a compact set $\mathcal{F}$, since this leads to a constrained

Regularization Term

optimization problem, which can be cumbersome in practice. Instead, we add a stabilization (regularization) term $\Omega[f]$ to the original objective function; the latter could be $R_{\text{emp}}[f]$, for instance. This, too, leads to better conditioning of the problem. We consider the following class of regularized risk functionals (see also Problem 4.1)

$$R_{\text{reg}}[f] := R_{\text{emp}}[f] + \lambda\Omega[f]. \tag{4.1}$$

Here $\lambda > 0$ is the so-called regularization parameter which specifies the trade-off between minimization of $R_{\text{emp}}[f]$ and the smoothness or simplicity which is enforced by small $\Omega[f]$. Usually one chooses $\Omega[f]$ to be convex, since this ensures that there exists only one global minimum, provided $R_{\text{emp}}[f]$ is also convex (see Lemma 6.3 and Theorem 6.5).

Maximization of the margin of classification in feature space by using the regularizing term $\frac{1}{2}||\mathbf{w}||^2$, and thus minimizing

$$\Omega[f] := \frac{1}{2}||\mathbf{w}||^2, \text{ and therefore } R_{\text{reg}}[f] = R_{\text{emp}}[f] + \frac{\lambda}{2}||\mathbf{w}||^2, \tag{4.2}$$

is the common choice in SV classification [573, 62]. In regression, the geometrical interpretation of minimizing $\frac{1}{2}||\mathbf{w}||^2$ is to find the *flattest* function with sufficient approximation qualities. Unless stated otherwise, we will limit ourselves to this type of regularizer in the present chapter. Other methods, e.g., minimizing the ℓ_p norm (where $||\mathbf{x}||_p^p = \sum_i x_i^p$) of the expansion coefficients for $\mathbf{w}$, will be discussed in Section 4.9.

Quadratic Regularizer

As described in Section 2.2.3, we can equivalently think of the feature space as a reproducing kernel Hilbert space. It is often useful, and indeed it will be one of the central themes of this chapter, to rewrite the risk functional (4.2) in terms of the RKHS representation of the feature space. In this case, we equivalently minimize

Regularized Risk in RKHS

$$R_{\text{reg}}[f] := R_{\text{emp}}[f] + \frac{\lambda}{2}||f||_{\mathcal{H}}^2 \tag{4.3}$$

over the whole space $\mathcal{H}$. The next section will study the properties of minimizers of (4.3), and similar regularizers that depend on $||f||_{\mathcal{H}}$.

4.2 The Representer Theorem

History of the Representer Theorem

The explicit form of a minimizer of $R_{\text{reg}}[f]$ is given by the celebrated representer theorem of Kimeldorf and Wahba [296] which plays a central role in solving practical problems of statistical estimation. It was first proven in the context of squared loss functions, and later extended to general pointwise loss functions [115]. For a machine learning point of view of the representer theorem, and variational proofs, see [205, 512]. The linear case has also been dealt with in [300]. We present a new and slightly more general version of the theorem with a simple proof [473]. As above, $\mathcal{H}$ is the RKHS associated to the kernel k.

Theorem 4.2 (Representer Theorem) *Denote by* $\Omega : [0, \infty) \to \mathbb{R}$ *a strictly monotonic increasing function, by* $\mathcal{X}$ *a set, and by* $c : (\mathcal{X} \times \mathbb{R}^2)^m \to \mathbb{R} \cup \{\infty\}$ *an arbitrary loss function. Then each minimizer* $f \in \mathcal{H}$ *of the regularized risk*

$$c\left((x_1, y_1, f(x_1)), \ldots, (x_m, y_m, f(x_m))\right) + \Omega\left(\|f\|_{\mathcal{H}}\right) \tag{4.4}$$

admits a representation of the form

$$f(x) = \sum_{i=1}^{m} \alpha_i k(x_i, x). \tag{4.5}$$

Note that this setting is slightly more general than Definition 3.1 since it allows *coupling* between the samples (x_i, y_i).

Before we proceed with the actual proof, let us make a few remarks. The original form, with pointwise mean squared loss

$$c((x_1, y_1, f(x_1)), \ldots, (x_m, y_m, f(x_m))) = \frac{1}{m} \sum_{i=1}^{m} (y_i - f(x_i))^2, \tag{4.6}$$

or hard constraints (i.e., hard limits on the maximally allowed error, incorporated formally by using a cost function that takes the value ∞), and $\Omega(\|f\|) = \frac{\lambda}{2}\|f\|^2_{\mathcal{H}}$ ($\lambda > 0$), is due to Kimeldorf and Wahba [296].

Requirements on $\Omega[f]$

Monotonicity of Ω is necessary to ensure that the theorem holds. It does not prevent the regularized risk functional (4.4) from having multiple local minima. To ensure a single minimum, we would need to require convexity. If we discard the strictness of the monotonicity, then it no longer follows that each minimizer of the regularized risk admits an expansion (4.5); it still follows, however, that there is always another solution that is as good, and that *does* admit the expansion.

Note that the freedom to use regularizers other than $\Omega(\|f\|) = \frac{\lambda}{2}\|f\|^2_{\mathcal{H}}$ allow us in principle to design algorithms that are more closely aligned with recommendations given by bounds derived from statistical learning theory, as described below (cf. Problem 5.7).

Significance

The significance of the Representer Theorem is that although we might be trying to solve an optimization problem in an infinite-dimensional space $\mathcal{H}$, containing linear combinations of kernels centered on *arbitrary* points of $\mathcal{X}$, it states that the solution lies in the span of m particular kernels — those centered on the training points. In the Support Vector community, (4.5) is called the *Support Vector expansion*. For suitable choices of loss functions, it has empirically been found that

Sparsity and Loss Function

many of the α_i often equal 0 (see Problem 4.6 for more detail on the connection between sparsity and loss functions).

Proof For convenience we will assume that we are dealing with $\bar{\Omega}(\|f\|^2) := \Omega(\|f\|)$ rather than $\Omega(\|f\|)$. This is no restriction at all, since the quadratic function is strictly monotonic on $[0, \infty)$, and therefore $\bar{\Omega}$ is strictly monotonic on $[0, \infty)$ if and only if Ω also satisfies this requirement.

We may decompose any $f \in \mathcal{H}$ into a part contained in the span of the kernel

functions $k(x_1, \cdot), \cdots, k(x_m, \cdot)$, and one in the orthogonal complement;

$$f(x) = f_{\|}(x) + f_{\perp}(x) = \sum_{i=1}^{m} \alpha_i k(x_i, x) + f_{\perp}(x). \tag{4.7}$$

Here $\alpha_i \in \mathbb{R}$ and $f_{\perp} \in \mathcal{H}$ with $\langle f_{\perp}, k(x_i, \cdot)\rangle_{\mathcal{H}} = 0$ for all $i \in [m] := \{1, \ldots, m\}$. By (2.34) we may write $f(x_j)$ (for all $j \in [m]$) as

$$f(x_j) = \big\langle f(\cdot), k(x_j, .)\big\rangle = \sum_{i=1}^{m} \alpha_i k(x_i, x_j) + \big\langle f_{\perp}(\cdot), k(x_j, .)\big\rangle_{\mathcal{H}} = \sum_{i=1}^{m} \alpha_i k(x_i, x_j). \tag{4.8}$$

Second, for all $f_{\perp}$,

$$\Omega(\|f\|_{\mathcal{H}}) = \bar{\Omega}\left(\left\|\sum_{i}^{m} \alpha_i k(x_i, \cdot)\right\|_{\mathcal{H}}^2 + \|f_{\perp}\|_{\mathcal{H}}^2\right) \geq \bar{\Omega}\left(\left\|\sum_{i}^{m} \alpha_i k(x_i, \cdot)\right\|_{\mathcal{H}}^2\right). \tag{4.9}$$

Thus for any fixed $\alpha_i \in \mathbb{R}$ the risk functional (4.4) is minimized for $f_{\perp} = 0$. Since this also has to hold for the solution, the theorem holds. ∎

Let us state two immediate extensions of Theorem 4.2. The proof of the following theorem is left as an exercise (see Problem 4.3).

Prior Knowledge by Parametric Expansions

Theorem 4.3 (Semiparametric Representer Theorem) *Suppose that in addition to the assumptions of the previous theorem we are given a set of M real-valued functions $\{\psi_p\}_{p=1}^{M} : \mathcal{X} \to \mathbb{R}$, with the property that the $m \times M$ matrix $(\psi_p(x_i))_{ip}$ has rank M. Then any $\tilde{f} := f + h$, with $f \in \mathcal{H}$ and $h \in \operatorname{span}\{\psi_p\}$, minimizing the regularized risk*

$$c\left((x_1, y_1, \tilde{f}(x_1)), \ldots, (x_m, y_m, \tilde{f}(x_m))\right) + \Omega\left(\|f\|_{\mathcal{H}}\right) \tag{4.10}$$

admits a representation of the form

$$\tilde{f}(x) = \sum_{i=1}^{m} \alpha_i k(x_i, x) + \sum_{p=1}^{M} \beta_p \psi_p(x), \tag{4.11}$$

with $\beta_p \in \mathbb{R}$ for all $p \in [M]$.

We will discuss applications of the semiparametric extension in Section 4.8.

Bias

Remark 4.4 (Biased Regularization) *Another extension of the representer theorems can be obtained by including a term $-\langle f_0, f\rangle$ in (4.4) or (4.10), where $f_0 \in \mathcal{H}$. In this case, if a solution to the minimization problem exists, it admits an expansion which differs from those described above in that it additionally contains a multiple of f_0. To see this, decompose $f_{\perp}(\cdot)$ used in the proof of Theorem 4.2 into a part orthogonal to f_0 and the remainder.*

Biased regularization means that we do not assume that the function $f = 0$ is the most simple of all estimates. This is a convenient way of incorporating prior knowledge about the type of solution we expect from our estimation procedure.

After this rather abstract and formal treatment of regularization, let us consider some practical cases where the representer theorem can be applied. First consider

the problem of regression, where the solution is chosen to be an element of a Reproducing Kernel Hilbert Space.

Application of Semiparametric Expansion

Example 4.5 (Support Vector Regression) *For Support Vector regression with the ε-insensitive loss (Section 1.6) we have*

$$c\left(\left(x_i, y_i, f(x_i)\right)_{i\in[m]}\right) = \frac{1}{m}\sum_{i=1}^{m}|y_i - f(x_i)|_\varepsilon \tag{4.12}$$

and $\Omega\left(\|f\|\right) = \frac{\lambda}{2}\|f\|^2$, where $\lambda > 0$ and $\varepsilon \geq 0$ are fixed parameters which determine the trade-off between regularization and fit to the training set. In addition, a single ($M = 1$) constant function $\psi_1(x) = 1$ is used as an offset, and is not regularized by the algorithm.

Section 4.8 and [507] contain details how the case of $M > 1$, for which more than one parametric function is used, can be dealt with algorithmically. Theorem 4.3 also applies in this case.

Example 4.6 (Support Vector Classification) *Here, the targets consist of $y_i \in \{\pm 1\}$, and we use the soft margin loss function (3.3) to obtain*

$$c\left(\left(x_i, y_i, f(x_i)\right)_i\right) = \frac{1}{m}\sum_{i}^{m}\max\left(0, 1 - y_i f(x_i)\right). \tag{4.13}$$

The regularizer is $\Omega\left(\|f\|\right) = \frac{\lambda}{2}\|f\|^2$, and $\psi_1(x) = 1$. For $\lambda \to 0$, we recover the hard margin SVM, for which the minimizer must correctly classify each training point $\left(x_i, y_i\right)$. Note that after training, the actual classifier will be $\operatorname{sgn}\left(f(.)\right)$.

Kernel Principal Component Analysis

Example 4.7 (Kernel PCA) *Principal Component Analysis (see Chapter 14 for details) in a kernel feature space can be shown to correspond to the case of*

$$c((x_i, y_i, f(x_i))_i) = \begin{cases} 0 & \text{if } \frac{1}{m}\sum_i \left(f(x_i) - \frac{1}{m}\sum_j f(x_j)\right)^2 = 1 \\ \infty & \text{otherwise} \end{cases} \tag{4.14}$$

with $\Omega(.)$ an arbitrary function that is strictly monotonically increasing [480]. The constraint ensures that we only consider linear feature extraction functionals that produce outputs of unit empirical variance. In other words, the task is to find the simplest function with unit variance. Note that in this case of unsupervised learning, there are no labels y_i to consider.

4.3 Regularization Operators

Curse of Dimensionality

The RKHS framework proved useful in obtaining the explicit functional form of minimizers of the regularized risk functional. It still does not explain the good performance of kernel algorithms, however. In particular, it seems counter-intuitive that estimators using very high dimensional feature spaces (easily with some 10^{10} features as in optical character recognition with polynomial kernels, or even infinite dimensional spaces in the case of Gaussian RBF-kernels) should exhibit good

performance. It seems as if kernel methods are defying the curse of dimensionality [29], which requires the number of samples to increase with the dimensionality of the space in which estimation is performed. However, the distribution of capacity in these spaces is not isotropic (cf. Section 2.2.5).

Regularization Operator Viewpoint

The basic idea of the viewpoint described in the present section is simple: rather than dealing with an abstract quantity such as an RKHS, which is defined by means of its corresponding kernel k, we take the converse approach of obtaining a kernel via the corresponding Hilbert space. Unless stated otherwise, we will use $L_2(\mathcal{X})$ as the Hilbert space (cf. Section B.3) on which the regularization operators will be defined. Note that $L_2(\mathcal{X})$ is *not* the feature space $\mathcal{H}$.

Recall that in Section 2.2.2, we showed that one way to think of the kernel mapping is as a map that takes a point $x \in \mathcal{X}$ to a function $k(x, .)$ living in an RKHS. To do this, we constructed a dot product $\langle ., . \rangle_{\mathcal{H}}$ satisfying

$$k(x, x') = \langle k(x, .), k(x, .)' \rangle_{\mathcal{H}}. \tag{4.15}$$

Physically, however, it is still unclear what the dot product $\langle f, g \rangle_{\mathcal{H}}$ actually does. Does it compute some kind of "overlap" of the functions, similar to the usual dot product between functions in $L_2(\mathcal{X})$? Recall that, assuming we can define an integral on $\mathcal{X}$, the latter is (cf. (B.60))

$$\langle f, g \rangle_{L_2(\mathcal{X})} = \int_{\mathcal{X}} f \cdot g. \tag{4.16}$$

Main Idea

In the present section, we will show that whilst our dot product in the RKHS is not quite a simple as (4.16), we can at least write it as

$$\langle f, g \rangle_{\mathcal{H}} = \langle \Upsilon f, \Upsilon g \rangle_{L_2} = \int_{\mathcal{X}} \Upsilon f(x) \Upsilon g(x) dx \tag{4.17}$$

in a suitable L_2 space of functions. This space contains transformed versions or the original functions, where the transformation Υ "extracts" those parts that should be affected by the regularization. This gives a much clearer physical understanding of the dot product in the RKHS (and thus of the similarity measure used by SVMs). It becomes particularly illuminating once one sees that for common kernels, the associated transformation Υ extracts properties like *derivatives* of functions. In other words, these kernels induce a form of regularization that penalizes non-smooth functions.

Definition 4.8 (Regularization Operator) *A regularization operator Υ is defined as a linear map from the space of functions $\mathcal{F} := \{f | f : \mathcal{X} \to \mathbb{R}\}$ into a space equipped with a dot product. The regularization term $\Omega[f]$ takes the form*

$$\Omega[f] := \frac{1}{2} \langle \Upsilon f, \Upsilon f \rangle. \tag{4.18}$$

Positive Definite Operator

Without loss of generality, we may assume that Υ is positive definite. This can be seen as follows: all that matters for the definition of $\Omega[f]$ is the positive definite operator $\Upsilon^* \Upsilon$ (since $\langle \Upsilon f, \Upsilon f \rangle = \langle f, \Upsilon^* \Upsilon f \rangle$). Hence we may always define a positive definite operator $\Upsilon_h := (\Upsilon^* \Upsilon)^{\frac{1}{2}}$ (cf. Section B.2.2) which has the same regulariza-

tion properties as Υ. Next, we formally state the equivalence between RKHS and regularization operator view.

Theorem 4.9 (RKHS and Regularization Operators) *For every RKHS $\mathcal{H}$ with reproducing kernel k there exists a corresponding regularization operator $\Upsilon : \mathcal{H} \to \mathcal{D}$ such that for all $f \in \mathcal{H}$*

$$\langle \Upsilon k(x, \cdot), \Upsilon f(\cdot) \rangle_{\mathcal{D}} = f(x), \tag{4.19}$$

and in particular,

$$\langle \Upsilon k(x, \cdot), \Upsilon k(x', \cdot) \rangle_{\mathcal{D}} = k(x, x'). \tag{4.20}$$

Matching RKHS

Likewise, for every regularization operator $\Upsilon : \mathcal{F} \to \mathcal{D}$, where $\mathcal{F}$ is some function space equipped with a dot product, there exists a corresponding RKHS $\mathcal{H}$ with reproducing kernel k such that (4.19) and (4.20) are satisfied.

Equation (4.20) will become the central tool to analyze smoothness properties of kernels, in particular if we pick $\mathcal{D}$ to be $L_2(\mathcal{X})$. In this case we will obtain an explicit form of the dot product induced by the RKHS which will thereby clarify why kernel methods work.

From Section 2.2.4 we can see that minimization of $||\mathbf{w}||^2$ is equivalent to minimization of $\Omega[f]$ (4.18), due to the feature map $\Phi(x) := k(x, \cdot)$.

Proof We prove the first part by explicitly constructing an operator that takes care of the mapping. One can see immediately that $\Upsilon = \mathbf{1}$ and $\mathcal{D} = \mathcal{H}$ will satisfy all requirements.[1]

For the converse statement, we have to obtain k from $\Upsilon^*\Upsilon$ and show that this is, in fact, the kernel of an RKHS (note that this does not imply that $\mathcal{D} = \mathcal{H}$ since it may be equipped with a different dot product than $\mathcal{H}$).

A function $G_x(\cdot)$ satisfying the first equality in

$$f(x) = \langle \Upsilon^*\Upsilon G_x(\cdot), f \rangle_{\mathcal{F}} = \langle \Upsilon G_x, \Upsilon f \rangle_{\mathcal{F}} \tag{4.21}$$

for all $f \in \Upsilon^*\Upsilon\mathcal{F}$ is called *Green's function* of the operator $\Upsilon^*\Upsilon$ on $\mathcal{D}$. It is known that such functions exist [448]. Note that this amounts to our desired reproducing property (4.19), on the set $\Upsilon^*\Upsilon\mathcal{F}$. The second equality in (4.21) follows from the definition of the adjoint operator Υ^*.

By applying (4.21) to G_x it follows immediately that G is symmetric,

$$G_x(x') = \langle \Upsilon^*\Upsilon G_{x'}, G_x \rangle_{\mathcal{D}} = \langle \Upsilon G_{x'}, \Upsilon G_x \rangle_{\mathcal{D}} = \langle \Upsilon G_x, \Upsilon G_{x'} \rangle_{\mathcal{D}} = G_{x'}(x). \tag{4.22}$$

We will write it as $G(x, x')$. Observe that (4.22) actually tells us that $x \mapsto \Upsilon G_x$ is actually a valid feature map for G. Therefore, we may identify $G(x, x')$ with $k(x, x')$.

1. $\Upsilon = \mathbf{1}$ is not the most useful operator. Typically we will seek an operator Υ corresponding to a *specific* dot product space $\mathcal{D}$. Note that this need not always be possible if $\mathcal{D}$ is not suitably chosen, e.g., for $\mathcal{D} = \mathbb{R}$.

The corresponding RKHS is the closure of the set $\{f \in \Upsilon^*\Upsilon\mathcal{F} | \|\Upsilon f\|^2 < \infty\}$. ■

This means that $\mathcal{D}$ is an RKHS with inner product $\langle \Upsilon\cdot, \Upsilon\cdot\rangle_{\mathcal{D}}$. Furthermore, Theorem 4.9 means that fixing the regularization operator Υ determines the possible set of functions that we might obtain, independently[2] of the class of functions in which we expand the estimate f. Thus Support Vector Machines are simply a very convenient way of specifying the regularization and a matching class of basis functions via one kernel function. This is done mainly for algorithmic advantages when formulating the corresponding optimization problem (cf. Chapter 7). The case where the two do not match is discussed in detail in [512].

Kernel Function ≙ Regularization Operator

Given the eigenvector decomposition of a regularization operator we can define a class of kernels that satisfy the self consistency condition (4.20).

Proposition 4.10 (A Discrete Counterpart) *Given a regularization operator Υ with an expansion of $\Upsilon^*\Upsilon$ into a discrete eigenvector decomposition (λ_n, ψ_n), and a kernel k with*

$$k(x, x') := \sum_{n, \lambda_n \neq 0} \frac{d_n}{\lambda_n} \psi_n(x)\psi_n(x'), \tag{4.23}$$

where $d_n \in \{0, 1\}$ for all m, and $\sum_n \frac{d_n}{\lambda_n}$ convergent, then k satisfies (4.20). Moreover, the corresponding RKHS is given by span$\{\psi_i | d_i = 1$ and $i \in \mathbb{N}\}$.

Proof We evaluate (4.21) and use the orthonormality of the system $(\frac{d_n}{\lambda_n}, \psi_n)$.

$$\begin{aligned} &\left\langle k(x_i, .), (\Upsilon^*\Upsilon k)(x_j, .)\right\rangle \qquad (4.24)\\ &= \left\langle \sum_n \frac{d_n}{\lambda_n}\psi_n(x_i)\psi_n(.), \Upsilon^*\Upsilon\left(\sum_{n'} \frac{d_{n'}}{\lambda_{n'}}\psi_{n'}(x_j)\psi_{n'}(.)\right)\right\rangle \\ &= \sum_{n,n'} \frac{d_n}{\lambda_n}\frac{d_{n'}}{\lambda_{n'}}\psi_n(x_i)\psi_{n'}(x_j)\left\langle \psi_n(.), \Upsilon^*\Upsilon\psi_{n'}(.)\right\rangle \\ &= \sum_n \frac{d_n}{\lambda_n}\psi_n(x_i)\psi_n(x_j) = k(x_i, x_j). \end{aligned}$$

The statement about the span follows immediately from the construction of k.

■

The summation coefficients are permitted to be rearranged, since the eigenfunctions are orthonormal and the series $\sum_n \frac{d_n}{\lambda_n}$ converges absolutely. Consequently a large class of kernels can be associated with a given regularization operator (and vice versa), thereby restricting us to a subspace of the eigenvector decomposition of $\Upsilon^*\Upsilon$.

Null Space of $\Upsilon^*\Upsilon$

In other words, there exists a one to one correspondence between kernels and regularization operators only on the image of $\mathcal{H}$ under the integral operator

2. Provided that no $f \in \mathcal{D}$ contains directions of the null space of the regularization operator $\Upsilon^*\Upsilon$, and that the kernel functions k span the whole space $\mathcal{D}$. If this is not the case, simply define the space to be the span of $k(x, \cdot)$.

$(T_k f)(x) := \int k(x, x') f(x') dx$, namely that T_k and $\Upsilon^* \Upsilon$ are inverse to another. On the null space of T_k, however, the regularization operator $\Upsilon^* \Upsilon$ may take on an arbitrary form. In this case k still will fulfill the self consistency condition.

Excluding eigenfunctions of $\Upsilon^* \Upsilon$ from the kernel expansion effectively decreases the expressive power of the set of approximating functions, and limits the capacity of the system of functions. Removing low capacity (i.e. very flat) eigenfunctions from the expansion will have an adverse effect, though, as the data will then be approximated by the higher capacity functions.

We have now covered the main insights of the present chapter. The following sections are more technical and can be skipped if desired. Recall that at the beginning of the present section, we explained that regularization operators can be thought of as extracting those parts of the functions that should be affected by the regularization. In the next section, we show that for a specific class of kernels, this extraction coincides with the Fourier transform.

4.4 Translation Invariant Kernels

An important class of kernels $k(x, x')$, such as Gaussian RBF kernels or Laplacian kernels only depends on the difference between x and x'. For the sake of simplicity and with slight abuse of notation we will use the shorthand

$$k(x, x') = k(x - x') \tag{4.25}$$

or simply $k(x)$. Since such k are independent of the *absolute* position of x but depend only on $x - x'$ instead, we will refer to them as *translation invariant* kernels.

What we will show in the following is that for kernels defined via (4.25) there exists a simple recipe how to find a regularization operator $\Upsilon^* \Upsilon$ corresponding to k and vice versa. In particular, we will show that the Fourier transform of $k(x)$ will provide us with the representation of the regularization operator in the frequency domain.

Fourier Transformation For this purpose we need a few definitions. For the sake of simplicity we assume $\mathcal{X} \subset \mathbb{R}^N$. In this case the Fourier transformation of f is given by

$$F[f](\boldsymbol{\omega}) := (2\pi)^{-\frac{N}{2}} \int_{\mathcal{X}} f(\mathbf{x}) \exp(-i \langle \mathbf{x}, \boldsymbol{\omega} \rangle) d\mathbf{x}. \tag{4.26}$$

Note that here i is the imaginary unit and that, in general, $F[f](\boldsymbol{\omega}) \in \mathbb{C}$ is a complex number. The inverse Fourier transformation is then given by

$$f(x) = F^{-1}[f](\boldsymbol{\omega}) = (2\pi)^{-\frac{N}{2}} \int_{\mathcal{X}} F[f](\boldsymbol{\omega}) \exp(i \langle \mathbf{x}, \boldsymbol{\omega} \rangle) d\boldsymbol{\omega}. \tag{4.27}$$

Regularization Operator in Fourier Domain We now specifically consider regularization operators Υ that may be written as multiplications in Fourier space (i.e. $\Upsilon^* \Upsilon$ is diagonalized in the Fourier basis).

Denote by $\upsilon(\boldsymbol{\omega})$ a nonnegative, symmetric function defined on $\mathcal{X}$, i.e. $\upsilon(-\boldsymbol{\omega}) = \upsilon(\boldsymbol{\omega}) \geq 0$ which converges to 0 for $\|\boldsymbol{\omega}\| \to \infty$. Moreover denote by Ω the support of $\upsilon(\boldsymbol{\omega})$ and by $\bar{x}$ the complex conjugate of x. Now we introduce a regularization operator by

$$\langle \Upsilon f, \Upsilon g \rangle_{\mathcal{D}} = (2\pi)^{\frac{N}{2}} \int_{\Omega} \frac{\overline{F[f](\boldsymbol{\omega})} F[g](\boldsymbol{\omega})}{\upsilon(\boldsymbol{\omega})} d\boldsymbol{\omega}. \tag{4.28}$$

The goal of regularized risk minimization in RKHS is to find a function which minimizes $R_{\mathrm{reg}}[f]$ while keeping $\langle \Upsilon f, \Upsilon f \rangle_{\mathcal{D}}$ reasonably small. In the context of (4.28) this means the following:

Small nonzero values of $\upsilon(\boldsymbol{\omega})$ correspond to a *strong* attenuation of the corresponding frequencies. Hence small values of $\upsilon(\boldsymbol{\omega})$ for large $\boldsymbol{\omega}$ are desirable, since high frequency components of $F[f]$ correspond to rapid changes in f. It follows that $\upsilon(\boldsymbol{\omega})$ describes the filter properties of $\Upsilon^* \Upsilon$ — note that no attenuation takes place for $\upsilon(\boldsymbol{\omega}) = 0$, since these frequencies have been excluded from the integration domain Ω.

Our next step is to construct kernels k corresponding to Υ as defined in (4.28).

Green's Functions and Fourier Transformations We show that

$$G(x, x') = (2\pi)^{-\frac{N}{2}} \int_{\Omega} e^{i\boldsymbol{\omega}(x - x')} \upsilon(\boldsymbol{\omega}) d\boldsymbol{\omega}, \tag{4.29}$$

is a Green's function for $\Upsilon, \mathcal{D}$ and that it can be used as a kernel. For a function f, whose support of its Fourier transform is contained in Ω, we have

$$\begin{aligned} \langle G(x, \cdot), f \rangle_{\mathcal{D}} &= (2\pi)^{-\frac{N}{2}} \int_{\Omega} \frac{\overline{F[G(x, \cdot)](\boldsymbol{\omega})} F[f](\boldsymbol{\omega})}{\upsilon(\boldsymbol{\omega})} d\boldsymbol{\omega} && (4.30) \\ &= (2\pi)^{-\frac{N}{2}} \int_{\Omega} \frac{\overline{\upsilon(\boldsymbol{\omega})} \exp(i \langle x, \boldsymbol{\omega} \rangle) F[f](\boldsymbol{\omega})}{\upsilon(\boldsymbol{\omega})} d\boldsymbol{\omega} && (4.31) \\ &= (2\pi)^{-\frac{N}{2}} \int_{\Omega} \exp(i \langle x, \boldsymbol{\omega} \rangle) F[f](\boldsymbol{\omega}) d\boldsymbol{\omega} = f(x). && (4.32) \end{aligned}$$

From Theorem 4.9 it now follows that G is a Green's function and that it can be used as an RKHS kernel.

Eq. (4.29) provides us with an efficient tool for analyzing SV kernels and the types of capacity control they exhibit: we may also read (4.29) backwards and, in doing so, find the regularization operator for a given kernel, simply by applying the Fourier transform to $k(x)$. As expected, kernels with high frequency components will lead to less smooth estimates.

Note that (4.29) is a special case of Bochner's theorem [60], which states that the Fourier transform of a positive measure constitutes a positive definite kernel.

In the remainder of this section we will now apply our new insight to a wide range of popular kernels such as B_n-splines, Gaussian kernels, Laplacian kernels, and periodic kernels. A discussion of the multidimensional case which requires additional mathematical techniques is left to Section 4.5.

4.4.1 B_n-Splines

As was briefly mentioned in Section 2.3, splines are an important tool in interpolation and function estimation. They excel at problems of low dimensional interpolation. Computational problems become increasingly acute, however, as the dimensionality of the patterns (i.e. of x) increases; yet there exists a way to circumvent these difficulties. In [501, 572], a method is proposed for using B_n-splines (see Figure 4.1) as building blocks for kernels, i.e.,

Splines in $\mathbb{R}$

$$k(x) = B_n(x). \tag{4.33}$$

We start with $\mathcal{X} = \mathbb{R}$ (higher dimensional cases can also be obtained, for instance by taking products over the individual dimensions). Recall that B_n splines are defined as $n+1$ convolutions[3] of the centered unit interval (cf. (2.71) and [552]);

$$B_n = \bigotimes_{i=1}^{n+1} I_{[-0.5,0.5]}. \tag{4.34}$$

Given this kernel, we now use (4.29) in order to obtain the corresponding Fourier representation. In particular, we must compute the Fourier transform of $B_n(x)$. The following theorem allows us to do this conveniently for functions represented by convolutions.

Convolutions and Products

Theorem 4.11 (Fourier-Plancherel, e.g. [306, 112]) *Denote by f, g two functions in $L_2(\mathcal{X})$, by $F[f], F[g]$ their corresponding Fourier transforms, and by $\otimes$ the convolution operation. Then the following identities hold.*

$$F[f \otimes g] = F[f] \cdot F[g], \text{ and } F[f] \otimes F[g] = F[f \cdot g] \tag{4.35}$$

In other words, convolutions in the original space become products in the Fourier domain and vice versa. Hence we may jump from one representation to the other depending on which space is most convenient for our calculations.

Repeated application of Theorem 4.11 shows that in the case of B_n splines, the Fourier representation is conveniently given by the $n+1$st power of the Fourier transform of B_0. Since the Fourier transform of B_n equals $\upsilon(\omega)$, we obtain (up to a multiplicative constant)

$$\upsilon(\boldsymbol{\omega}) = F[k](\boldsymbol{\omega}) = \prod_{i=1}^{N} \operatorname{sinc}^{(n+1)}\left(\frac{\omega_i}{2}\right), \text{ where } \operatorname{sinc} x := \frac{\sin x}{x}. \tag{4.36}$$

3. A convolution $f \otimes g$ of two functions $f, g : \mathcal{X} \to \mathbb{R}$ is defined as

$$f \otimes g = (2\pi)^{-\frac{N}{2}} \int_{\mathcal{X}} f(x')g(x - x')dx'.$$

The normalization factor of $(2\pi)^{-\frac{N}{2}}$ serves to make the convolution compatible with the Fourier transform. We will need this property in Theorem 4.11. Note that $f \otimes g = g \otimes f$, as can be seen by exchange of variables.

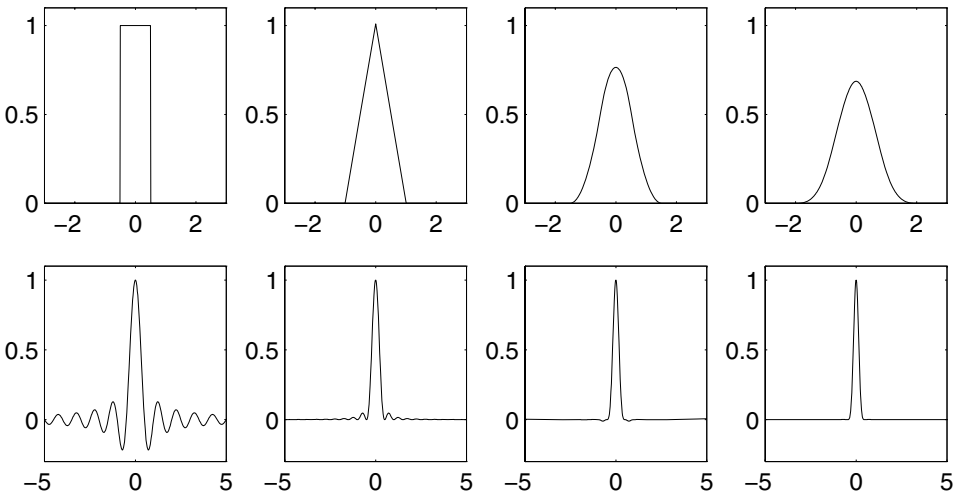


Figure 4.1 From left to right: B_n splines of order 0 to 3 (top row) and their Fourier transforms (bottom row). The length of the support of B_n is $n+1$, and the degree of continuous differentiability increases with $n-1$. Note that the higher the degree of B_n, the more peaked the Fourier transform (4.36) becomes. This is due to the increasing support of B_n. The frequency axis labels of the Fourier transform are multiples of 2π.

Only B_{2n+1} Splines Admissible

This illustrates why only B_n splines of odd order are positive definite kernels (cf. (2.71)):[4] The even ones have negative components in the Fourier spectrum (which would result in an amplification of the corresponding frequencies). The zeros in $F[k]$ stem from the fact that B_n has compact support; $\left[-\frac{n+1}{2}, \frac{n+1}{2}\right]$. See Figure 4.2 for details.

By using this kernel, we trade reduced computational complexity in calculating f (we need only take points into account whose distance $\|x_i - x_j\|$ is smaller than the support of B_n), for a potentially decreased performance of the regularization operator, since it completely removes (i.e., disregards) frequencies $\boldsymbol{\omega}_p$ with $F[k](\boldsymbol{\omega}_p) = 0$. Moreover, as we shall see below, in comparison to other kernels, such as the Gaussian kernel, $F[k](\boldsymbol{\omega})$ decays rather slowly.

4.4.2 Gaussian Kernels

Another class of kernels are Gaussian radial basis function kernels (Figure 4.3). These are widely popular in Neural Networks and approximation theory [80, 203, 201, 420]. We have already encountered $k(x, x') = \exp\left(-\frac{\|x-x'\|^2}{2\sigma^2}\right)$ in (2.68); we now investigate the regularization and smoothness properties of these kernels.

For a Fourier representation we need only compute the Fourier transform of

4. Although both even and odd order B_n splines converge to a Gaussian as $n \to \infty$ due to the law of large numbers.

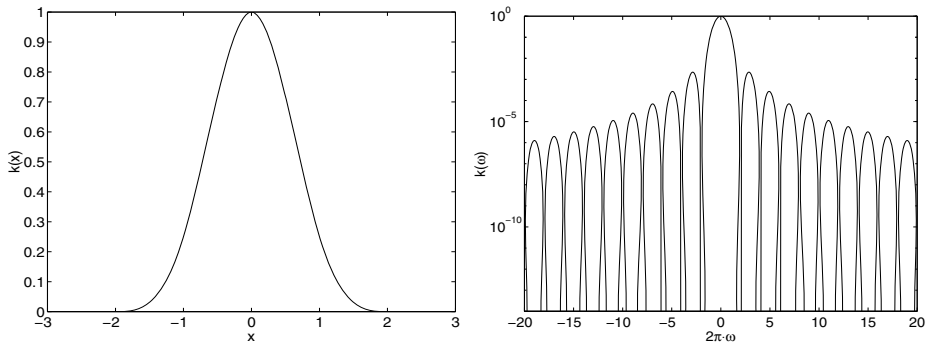


Figure 4.2 Left: B_3-spline kernel. Right: Fourier transform of k (in log-scale). Note the zeros and the rapid decay in the spectrum of B_3.

(2.68), which is given by

$$F[k](\omega) = \upsilon(\omega) = |\sigma| \exp\left(-\frac{\sigma^2\omega^2}{2}\right). \tag{4.37}$$

Uncertainty Relation

In other words, the smoother k is in pattern space, the more peaked its Fourier transform becomes. In particular, the product between the width of k and its Fourier transform is constant.[5] This phenomenon is also known as the uncertainty relation in physics and engineering.

Equation (4.37) also means that the contribution of high frequency components in estimates is relatively small, since $\upsilon(\omega)$ decays extremely rapidly. It also helps explain why Gaussian kernels produce full rank kernel matrices (Theorem 2.18).

We next determine an explicit representation of $\|\Upsilon f\|^2$ in terms of differential operators, rather than a pure Fourier space formalism. While this is not possible by using only "conventional" differential operators, we may achieve our goal by using *pseudo-differential* operators.

Pseudo-Differential Operators

Roughly speaking, a pseudo-differential operator differs from a differential operator in that it may contain an infinite sum of differential operators. The latter correspond to a Taylor expansion of the operator in the Fourier domain. There is an additional requirement that the arguments lie inside the radius of convergence, however.

Following the exposition of Yuille and Grzywacz [612] one can see that

$$\|\Upsilon f\|^2 = \int_{\mathcal{X}} \sum_n \frac{\sigma^{2n}}{n!\, 2^n} (O^n f(x))^2 dx, \tag{4.38}$$

with $O^{2n} = \Delta^n$ and $O^{2n+1} = \nabla\Delta^n$, Δ being the Laplacian and ∇ the Gradient operator, is equivalent to a regularization with $\upsilon(\omega)$ as in (4.37). The key observation in this context is that derivatives in $\mathcal{X}$ translate to multiplications in the frequency

5. The multidimensional case is completely analogous, since it can be decomposed into a product of one-dimensional Gaussians. See also Section 4.5 for more details.

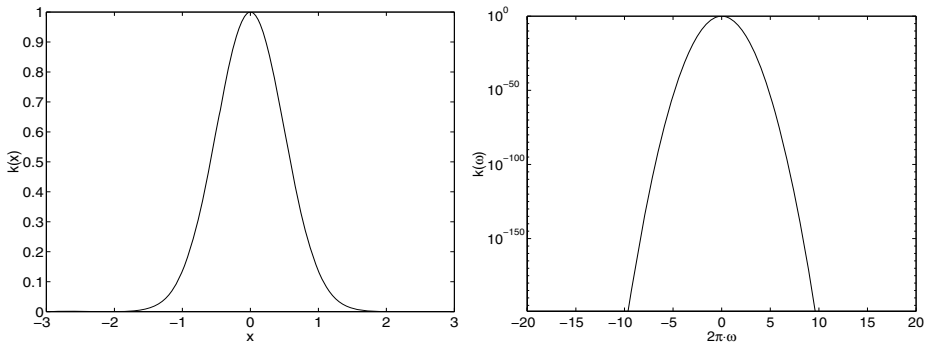


Figure 4.3 Left: Gaussian kernel with standard deviation 0.5. Right: Fourier transform of the kernel.

Taylor Expansion in Differential Operators

domain and vice versa.[6] Therefore a Taylor expansion of $\upsilon(\omega)$ in ω, can be rewritten as a Taylor expansion in $\mathcal{X}$ in terms of differential operators. See [612] and the references therein for more detail.

On the practical side, training an SVM with Gaussian RBF kernels [482] corresponds to minimizing the specific loss function with a regularization operator of type (4.38). Recall that (4.38) causes all derivatives of f to be penalized, to obtain a very smooth estimate. This also explains the good performance of SVMs in this case, since it is by no means obvious that choosing a flat function in *some* high dimensional space will correspond to a simple function in a low dimensional space (see Section 4.4.3 for a counterexample).

4.4.3 Dirichlet Kernels

Proposition 4.10 can also be used to generate practical kernels. In particular, [572] introduced a class of kernel based on Fourier expansions by

$$k(x) := 2\sum_{j=0}^{n} \cos jx = \frac{\sin(2n+1)\frac{x}{2}}{\sin\frac{x}{2}} \tag{4.39}$$

As in Section 4.4.1, we consider $x \in \mathbb{R}$ to avoid tedious notation. By construction, this kernel corresponds to $\upsilon(\omega) = \frac{1}{2}\sum_{i=-n}^{n} \delta_i(\omega)$, with δ_i being Dirac's delta function.

A regularization operator with these properties may not be desirable, however, as it only damps a finite number of frequencies (see Figure 4.4), and leaves all other frequencies unchanged, which can lead to overfitting (Figure 4.5).

6. Integrability considerations aside, one can see this by

$$\frac{d}{dx}f = \frac{d}{dx}\int_{\Omega} F[f](\omega)\exp(i\omega x)d\omega = \int_{\Omega} i\omega F[f](\omega)\exp(i\omega x)d\omega.$$

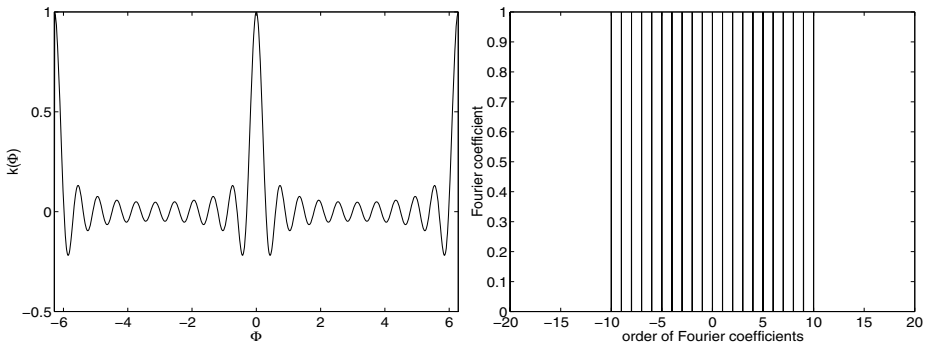


Figure 4.4 Left: Dirichlet kernel of order 10. Note that this kernel is periodic. Right: Fourier transform of the kernel.

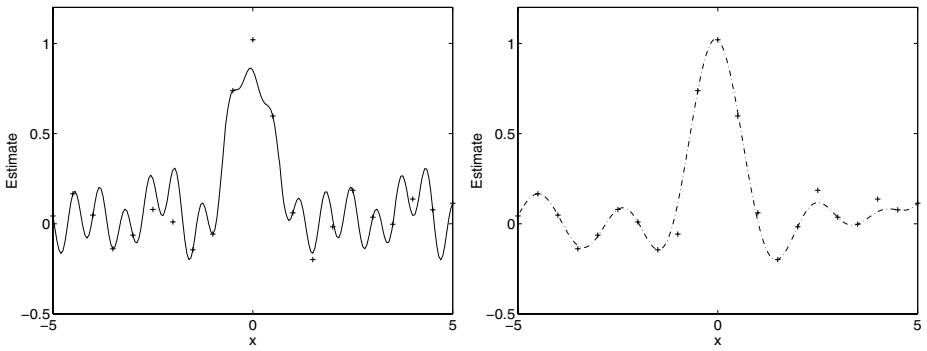


Figure 4.5 Left: Regression with a Dirichlet Kernel of order $N = 10$. One can clearly observe the overfitting (solid line: interpolation, '+': original data). Right: Regression based on the same data with a Gaussian Kernel of width $\sigma^2 = 1$ (dash dotted line: interpolation, '+': original data).

In other words, this kernel only describes band-limited periodic functions where no distinction is made between the different components of the frequency spectrum. Section 4.4.4 will present an example of a periodic kernel with a more suitable distribution of capacity over the frequency spectrum.

Types of Invariances

In some cases, it might be useful to approximate periodic functions, for instance functions defined on a circle. This leads to the second possible type of translation invariant kernel function, namely functions defined on factor spaces[7]. It is not reasonable to define translation invariant kernels on a bounded interval, since the data will lie beyond the boundaries of the specified interval when translated by a large amount. Therefore unbounded intervals and factor spaces are the only possible domains.

7. Factor spaces are vector spaces $\mathcal{X}$, with the additional property that for at least one nonzero element $\hat{x} \in \mathcal{X}$, we have $x + \hat{x} = x$ for all $x \in \mathcal{X}$. For instance, the modulo operation on $\mathbb{Z}$ forms such a space. We denote this space by $\mathbb{Z}/\hat{x}$.

We assume a period of 2π without loss of generality, and thus consider translation invariance on $\mathbb{R}/2\pi$. The next section shows how this setting affects the operator defined in section 4.4.2.

4.4.4 Periodic Kernels

Regularization Operator on $[0, 2\pi]$

One way of dealing with periodic invariances is to begin with a translation invariant regularization operator, defined similarly to (4.38), albeit on $L_2([0, 2\pi])$ (where the points 0 and π are identified) rather than on $L_2(\mathbb{R})$, and to find a matching kernel function. We start with the regularization operator;

$$||\Upsilon f||^2 := \pi^{-N} \int_{[0,2\pi]^N} \sum_n \frac{\sigma^{2n}}{n! 2^n} (O^n f(x))^2 dx, \tag{4.40}$$

with O defined as in Section 4.4.2. For the sake of simplicity, assume dim $\mathcal{X} = 1$. A generalization to multidimensional kernels is straightforward.

To obtain the eigensystem of Υ we start with the Fourier basis, which is dense on $L_2([0, 2\pi])$ [69], the space of functions we are interested in. One can check that the Fourier basis $\{\frac{1}{2\pi}, \sin(nx), \cos(nx), n \in \mathbb{N}\}$ is an eigenvector decomposition of the operator defined in (4.40), with eigenvalues $\exp(\frac{n^2\sigma^2}{2})$, by substitution into (4.40). Due to the Fourier basis being dense in $L_2([0, 2\pi])$, we have thus identified all eigenfunctions of Υ. Next we apply Proposition 4.10, taking into account all eigenfunctions except the constant function with $n = 0$. This yields the following kernel,

Periodic Kernels via Fourier Coefficients

$$\begin{aligned} k(x, x') &= \sum_{n=1}^{\infty} e^{-\frac{n^2\sigma^2}{2}} (\sin(nx)\sin(nx') + \cos(nx)\cos(nx')) \\ &= \sum_{n=1}^{\infty} e^{-\frac{n^2\sigma^2}{2}} \cos(n(x - x')). \end{aligned} \tag{4.41}$$

For practical purposes, one may truncate the expansion after a finite number of terms. Since the expansion coefficients decay rapidly, this approximation is very good. If necessary, k can be rescaled to have a range of exactly $[0, 1]$.

While this is a convenient way of building kernels if the Fourier expansion is known, we would also like to be able to render arbitrary translation invariant kernels on $\mathbb{R}$ periodic. The method is rather straightforward, and works as follows.

Periodic Kernels via Translation

Given any translation invariant kernel k we obtain k_p by

$$k_p(x, x') := \sum_{n \in \mathbb{Z}} k(x - x' + 2\pi n). \tag{4.42}$$

Again, we can approximate (4.42) by truncating the sum after a finite number of terms. The question is whether the definition of k_p leads to a positive definite kernel at all, and if so, which regularization properties it exhibits.

Proposition 4.12 (Spectrum of Periodized Kernels) *Denote by k a translation invariant kernel in $L_2(\mathcal{X})$, and by k_p its periodization according to (4.42). Moreover denote*

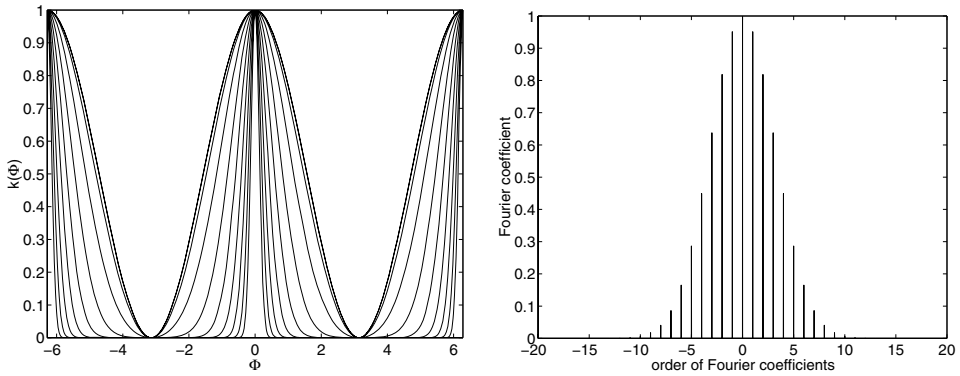


Figure 4.6 Left: Periodic Gaussian kernel for several values of σ (normalized to 1 as its maximum and 0 as its minimum value). Peaked functions correspond to small σ. Right: Fourier coefficients of the kernel for $\sigma^2 = 0.1$.

by $F[f]$ the Fourier transform of f. Then k_p can be expanded into the series

$$k_p(x, x') = (2\pi)^{-\frac{1}{2}} \left(F[f](0) + 2 \sum_{j=1}^{\infty} F[f](j) \cos(j(x - x')) \right). \tag{4.43}$$

Proof The proof makes use of the fact that for Lebesgue integrable functions k the integral over $\mathcal{X}$ can be split up into a sum over segments of size 2π. Specifically, we obtain

$$(2\pi)^{-\frac{1}{2}} \int_{\mathbb{R}} k(x) e^{-i\omega x} dx = (2\pi)^{-\frac{1}{2}} \sum_{j \in \mathbb{Z}} \int_{[0,2\pi]} k(x + 2\pi j) e^{-i\omega(x+2\pi j)} dx \tag{4.44}$$

$$= (2\pi)^{-\frac{1}{2}} \int_{[0,2\pi]} e^{-i\omega x} \sum_{j \in \mathbb{Z}} k(x + 2\pi j) dx \tag{4.45}$$

$$= (2\pi)^{-\frac{1}{2}} \int_{[0,2\pi]} e^{-i\omega x} k_p(x) dx. \tag{4.46}$$

The latter, however, is the Fourier transform of k_p over the interval $[0, 2\pi]$. Hence we have $F[k](j) = F[k_p](j)$ for $j \in \mathbb{Z}$, where $F[k_p](j)$ denotes the Fourier transform over the compact set $[0, 2\pi]$.

Now we may use the inverse Fourier transformation on $[0, 2\pi]$, to obtain a decomposition of k_p into a trigonometric series. Due to the symmetry of k, the imaginary part of $F[f]$ vanishes, and thus all contributions of $\sin jx$ cancel out. Moreover, we obtain (4.43) since $\cos x$ is a symmetric function. ■

In some cases, the full summation of k_p can be computed in closed form. See Problem 4.10 for an application of this reasoning to Laplacian kernels.

In the context of periodic functions, the difference between this kernel and the Dirichlet kernel of Section 4.4.3 is that the latter does not distinguish between the different frequency components in $\omega \in \{-n\pi, \ldots, n\pi\}$.

4.4.5 Practical Implications

We are now able to draw some useful conclusions regarding the practical application of translation invariant kernels. Let us begin with two extreme situations.

- Suppose that the shape of the power spectrum $\mathrm{Pow}[f](\boldsymbol{\omega})$ of the function we would like to estimate is known beforehand. In this case, we should choose k such that $F[k]$ matches the expected value of the power spectrum of f. The latter is given by the squared absolute value of the Fourier transformation of f, i.e.,

$$\mathrm{Pow}[f](\omega) := \left|F[f](\omega)\right|^2. \tag{4.47}$$

One may check, using the Fourier-Plancherel equality (Theorem 4.11) that $\mathrm{Pow}[f]$ equals the Fourier transformation of the autocorrelation function of f, given by $f(x) \otimes f(-x)$. In signal processing this is commonly known as the problem of "matched filters" [581]. It has been shown that the optimal filter for the reconstruction of signals corrupted with white noise, has to match the frequency distribution of the signal which is to be reconstructed. (White noise has a uniform distribution over the frequency band occupied by the useful signal.)

Matched Filters

- If we know very little about the given data, however, it is reasonable to make a general smoothness assumption. Thus a Gaussian kernel as in Section 4.4.2 or 4.4.4 is recommended. If computing time is important, we might instead consider kernels with compact support, such as the B_n-spline kernels of Section 4.4.1. This choice will cause many matrix elements $k_{ij} = k(x_i - x_j)$ to vanish.

The usual scenario will be in between these two extremes, and we will have some limited prior knowledge available, which should be used in the choice of kernel. The goal of the present reasoning is to give a guide to selection of kernels through a deeper understanding of the regularization properties. For more information on using prior knowledge for choosing kernels, e.g. by explicit construction of kernels exhibiting only a limited amount of interaction, see Chapter 13.

Prior Knowledge

Finally, note that the choice of the kernel width may be more important than the actual functional form of the kernel. For instance, there may be little difference in the relevant filter properties close to $\omega = 0$ between a B-spline and a Gaussian kernel (cf. Figure 4.7). This heuristic holds if we are interested only in uniform convergence results of a certain degree of precision, in which case only a small part of the power spectrum of k is relevant (see [604, 606] and also Section 12.4.1).

4.5 Translation Invariant Kernels in Higher Dimensions

Things get more complicated in higher dimensions. There are basically two ways to construct kernels in $\mathbb{R}^N \times \mathbb{R}^N \to \mathbb{R}$ with $N > 1$, if no particular assumptions on the data are made. First, we could construct kernels $k : \mathbb{R}^N \times \mathbb{R}^N \to \mathbb{R}$, by

Product Kernels

$$k(\mathbf{x} - \mathbf{x}') = k(x_1 - x'_1) \cdot \ldots \cdot k(x_N - x'_N). \tag{4.48}$$

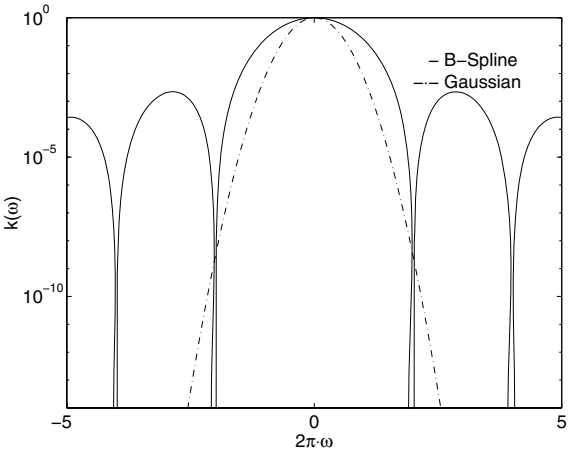


Figure 4.7 Comparison of regularization properties in the low frequency domain of the B_3-spline kernel and Gaussian kernel ($\sigma^2 = 20$). Down to an attenuation factor of $5 \cdot 10^{-3}$, i.e. in the interval $[-4\pi, 4\pi]$, both types of kernels exhibit somewhat similar filter characteristics.

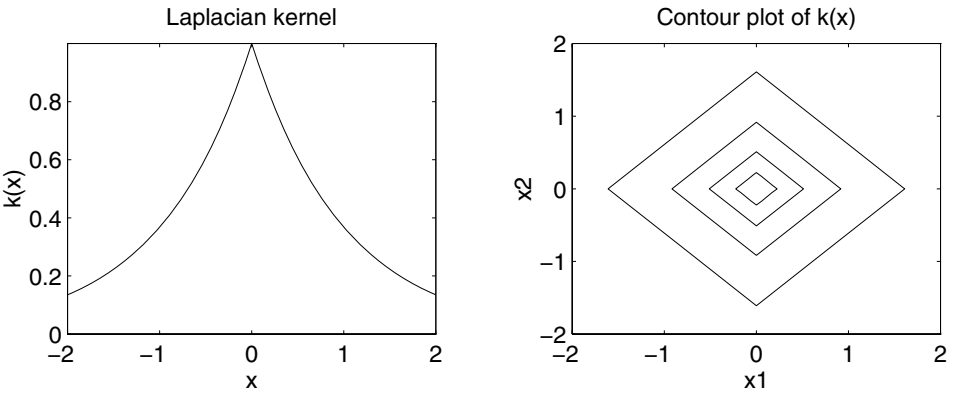


Figure 4.8 Laplacian product kernel in $\mathbb{R}$ and $\mathbb{R}^2$. Note the preferred directions in the two dimensional case.

Note that we have deviated from our usual notation in that in the present section, we use bold face letters to denote elements of the input space. This will help to simplify the notation, using $\mathbf{x} = (x_1, \ldots, x_N)$ and $\mathbf{w} = (w_1, \ldots, w_d)$ below.

The choice (4.48) usually leads to preferred directions in input space (see Figure 4.8), since the kernels are not generally rotation invariant, the exception being Gaussian kernels. This can also be seen from the corresponding regularization operator. Since k factorizes, we can apply the Fourier transform to k on a per-dimension basis, to obtain

$$F[k](\boldsymbol{\omega}) = F[k](\omega_1) \cdot \ldots \cdot F[k](\omega_N). \tag{4.49}$$

Kernels on Distance Matrices

The second approach is to assume $k(\mathbf{x} - \mathbf{x}') = k(||\mathbf{x} - \mathbf{x}'||_{\ell_2})$. This leads to kernels which are both translation invariant and rotation invariant. It is quite straightforward to generalize the exposition to the rotation asymmetric case, and norms other than the ℓ_2 norm. We now recall some basic results which will be useful later.

4.5.1 Basic Tools

Fourier Transform

The N-dimensional Fourier transform is defined as

$$F : L_2(\mathbb{R}^N) \to L_2(\mathbb{R}^N) \text{ with } F[f](\boldsymbol{\omega}) := \frac{1}{(2\pi)^{N/2}} \int_{\mathbb{R}^N} e^{-i\langle \boldsymbol{\omega}, \mathbf{x} \rangle} f(\mathbf{x}) d\mathbf{x}. \tag{4.50}$$

Its inverse transform is given by

$$F^{-1} : L_2(\mathbb{R}^N) \to L_2(\mathbb{R}^N) \text{ with } F^{-1}[f](x) = \frac{1}{(2\pi)^{N/2}} \int_{\mathbb{R}^N} e^{i\langle \boldsymbol{\omega}, \mathbf{x} \rangle} f(\boldsymbol{\omega}) d\boldsymbol{\omega}. \tag{4.51}$$

For radially symmetric functions, i.e. $f(\mathbf{x}) = f(\|\mathbf{x}\|)$, we can explicitly carry out the integration on the sphere to obtain a Fourier transform which is also radially symmetric (cf. [520, 373]):

$$F[f](\|\boldsymbol{\omega}\|) = \omega^{-\nu} H_\nu[r^\nu f(r)](\|\boldsymbol{\omega}\|), \tag{4.52}$$

where $\nu := \frac{1}{2}d - 1$, and H_ν is the Hankel transform over the positive real line (we use the shorthand $\omega = \|\boldsymbol{\omega}\|$). The latter is defined as

Hankel Transform

$$H_\nu[f](\omega) := \int_0^\infty r f(r) J_\nu(\omega r) dr. \tag{4.53}$$

Bessel Function

Here J_ν is the Bessel function of the first kind, which is given by

$$J_\nu(r) := r^\nu 2^{-\nu} \sum_{j=0}^{\infty} \frac{(-1)^j r^{2j}}{2^{2j} j! \Gamma(j + \nu + 1)} \tag{4.54}$$

and $\Gamma(x)$ is the Gamma function, satisfying $\Gamma(n+1) = n!$ for $n \in \mathbb{N}$.

Note that $H_\nu = H_\nu^{-1}$, i.e. $f = H_\nu[H_\nu[f]]$ (in L_2) due to the Hankel inversion theorem [520] (see also Problem 4.11), which is just another way of writing the inverse Fourier transform in the rotation symmetric case. Based on the results above, we can now use (4.29) to compute the Green's functions in $\mathbb{R}^N$ directly from the regularization operators given in Fourier space.

4.5.2 Regularization Properties of Kernels in $\mathbb{R}^N$

We now give some examples of kernels typically used in SVMs, this time in $\mathbb{R}^N$. We must first compute the Fourier/Hankel transform of the kernels.

Gaussian → Gaussian

Example 4.13 (Gaussian RBFs) *For Gaussian RBFs in N dimensions, $k(r) = \sigma^{-N} e^{-\frac{r^2}{2\sigma^2}}$, and correspondingly (as before we use the shorthand $\omega := \|\boldsymbol{\omega}\|$),*

$$F[k](\omega) = \omega^{-\nu} \sigma^{-N} H_\nu \left[r^\nu e^{-\frac{r^2}{2\sigma^2}} \right](\omega) = \omega^{-\nu} \sigma^{2(\nu+1)-N} \omega^\nu e^{-\frac{\omega^2\sigma^2}{2}} = e^{-\frac{\omega^2\sigma^2}{2}}.$$

In other words, the Fourier transform of a Gaussian is also a Gaussian, in higher dimensions.

Example 4.14 (Exponential RBFs) *In the case of* $k(r) = e^{-ar}$,

$$\begin{aligned} F[k](\omega) &= \omega^{-\nu} H_\nu \left[r^\nu e^{-ar} \right](\omega) \\ &= \omega^{-\nu} 2^{\nu+1} \omega^\nu a \pi^{-\frac{1}{2}} \Gamma\left(\nu + \tfrac{3}{2}\right) \left(a^2 + \omega^2\right)^{-\nu-\frac{3}{2}} \\ &= 2^{\frac{N}{2}} a \pi^{-\frac{1}{2}} \Gamma\left(\tfrac{N}{2} + 1\right) \left(a^2 + \omega^2\right)^{-\frac{N+1}{2}} \end{aligned} \tag{4.55}$$

Exponential → Inverse Polynomial

For $N = 1$ *we recover the damped harmonic oscillator in the frequency domain. In general, a decay in the Fourier spectrum approximately proportional to* $\omega^{-(N+1)}$ *can be observed. Moreover the Fourier transform of k, viewed itself as a kernel,* $k(r) = \left(1 + r^2\right)^{-\frac{N+1}{2}}$, *yields the initial kernel as its corresponding Fourier transform.*

Example 4.15 (Damped Harmonic Oscillator) *Another way to generalize the harmonic oscillator, this time so that k does not depend on the dimensionality N, is to set* $k(r) = \frac{1}{a^2+r^2}$. *Following [586, Section 13.6],*

Inverse Polynomial → Exponential

$$F[k](\omega) = \omega^{-\nu} H_\nu \left[\frac{r^\nu}{a^2 + r^2} \right](\omega) = \omega^{-\nu} a^\nu K_\nu(\omega a), \tag{4.56}$$

where K_ν *is the Bessel function of the second kind, defined by (see [520])*

$$K_\nu(x) = \int_0^\infty e^{-x \cosh t} \cosh(\nu t)\, dt. \tag{4.57}$$

It is possible to upper bound $F[k]$ *using*

$$K_\nu(x) = \sqrt{\frac{\pi}{2x}} e^{-x} \left[\sum_{j=0}^{p-1} (2x)^{-j} \frac{\Gamma\left(\nu + j + \frac{1}{2}\right)}{j! \Gamma\left(\nu - j + \frac{1}{2}\right)} + \theta \cdot (2x)^{-p} \frac{\Gamma\left(\nu + p + \frac{1}{2}\right)}{j! \Gamma\left(\nu - p + \frac{1}{2}\right)} \right], \tag{4.58}$$

with $p > \nu - \frac{1}{2}$ *and* $\theta \in [0, 1]$ *[209, eq. (8.451.6)]).The term in brackets* $[\cdot]$ *converges to* 1 *as* $x \to \infty$, *and thus results in an exponential decay of the Fourier spectrum.*

Example 4.16 (Modified Bessel Kernels) *In the previous example, we defined a kernel via* $k(r) = \frac{1}{a^2+r^2}$. *Since* $k(r)$ *is a nonnegative function with acceptable decay properties. Therefore we could also use this function to define a kernel in Fourier space via* $\upsilon(\boldsymbol{\omega}) = \frac{1}{a^2+||\boldsymbol{\omega}||^2}$. *The consequence thereof is that (4.56) will now be a kernel, i.e.,*

$$k(r) := r^{-\nu} a^\nu K_\nu(ra). \tag{4.59}$$

This is a popular kernel in Gaussian Process estimation [599] (see Section 16.3), since for $\nu > n$ *the corresponding Gaussian process is a mean-square differentiable stochastic processes. See [3] for more detail on this subject. For our purposes, it is sufficient to know that for* $\nu > n$, $k(||\mathbf{x} - \mathbf{x}'||)$ *is differentiable in* $\mathbb{R}^N$.

Example 4.17 (Generalized B_n Splines) *Finally, we generalize* B_n*-splines to N dimensions. One way is to define*

$$B_n^N := \bigotimes_{j=0}^{n} I_{U_N}, \tag{4.60}$$

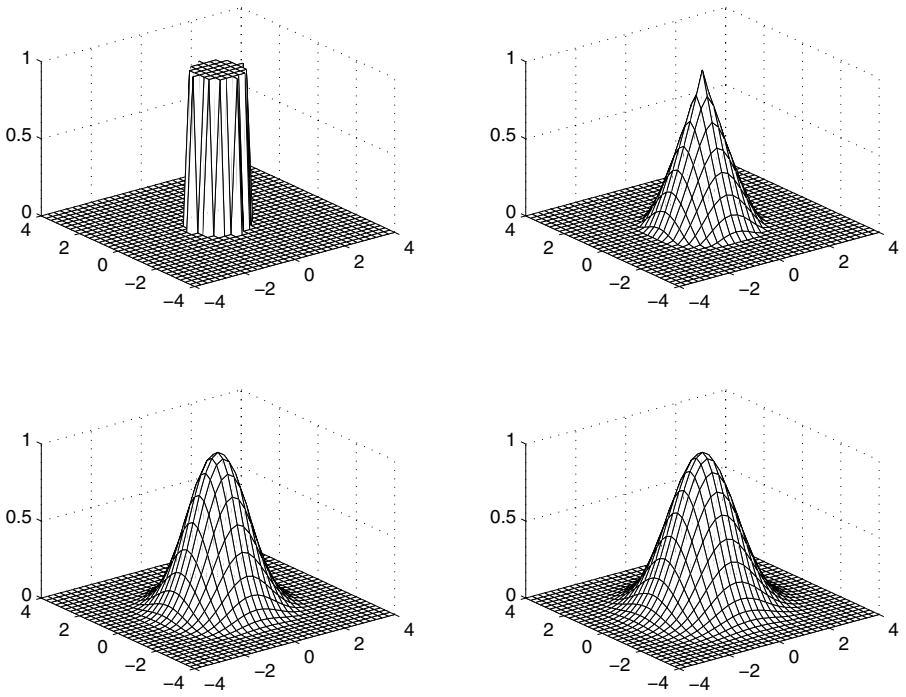


Figure 4.9 B_n splines in 2 dimensions. From left to right and top to bottom: Splines of order 0 to 3. Again, note the increasing degree of smoothness and differentiability with increasing order of the splines.

B_n Splines $\rightarrow$ Bessel Functions

so that B_n^N is the $n+1$-times convolution of the indicator function of the unit ball U_N in N dimensions. See Figure 4.9 for examples of such functions. Employing the Fourier-Plancherel Theorem (Theorem 4.11), we find that its Fourier transform is the $(n+1)$st power of the Fourier transform of the unit ball,

$$F[B_0^N](\omega) = \omega^{-(\nu+1)} J_{\nu+1}(\omega), \tag{4.61}$$

and therefore,

$$F[B_n^N](\omega) = \omega^{-(n+1)(\nu+1)} J_{\nu+1}^{n+1}(\omega). \tag{4.62}$$

Only odd n generate positive definite kernels, since it is only then that the kernel has a nonnegative Fourier transform.

4.5.3 A Note on Other Invariances

So far we have only been exploiting invariances with respect to the translation group in $\mathbb{R}^N$. The methods could also be applied to other symmetry transformations with corresponding canonical coordinate systems, however. This means that we use a coordinate system where invariance transformations can be represented as additions.

Lie Groups and Lie Algebras

Not all symmetries have this property. Informally speaking, those that do are called *Lie groups* (see also Section 11.3), and the parameter space where the additions take place is referred to as a *Lie algebra*. For instance, the rotation and scaling group (i.e. the product between the special orthogonal group $SO(N)$ and radial scaling), as proposed in [487, 167], corresponds to a log-polar parametrization of $\mathbb{R}^N$. The matching transform into frequency space is commonly referred to as the Fourier-Mellin transform [520].

4.6 Dot Product Kernels

A second, important family of kernels can be efficiently described in term of dot products, i.e.,

$$k(x, x') = k\left(\langle x, x' \rangle\right). \tag{4.63}$$

Here, with slight abuse of notation we use k to define dot product kernels via $k(\langle x, x' \rangle)$. Such dot product kernels include homogeneous and inhomogeneous polynomial kernel $\left(\langle x, x' \rangle + c\right)^p$ with $c \geq 0$. Proposition 2.1 shows that they satisfy Mercer's condition.

What we will do in the following is state an easily verifiable criterion, under which conditions a general kernel, as defined by (4.63), will satisfy Mercer's condition. A side-effect of this analysis will be a deeper insight into the regularization properties of the operator $\Upsilon^*\Upsilon$, when considered on the space $L_2(S_{N-1})$, where S_{N-1} is the unit sphere in $\mathbb{R}^N$. The choice of the domain S_{N-1} is made in order to exploit the symmetries inherent in k: $k(x, x')$ is rotation invariant in its arguments x, x'.

Regularization Properties via Mercer's Theorem

In a nutshell, we use Mercer's Theorem (Theorem 2.10) explicitly to obtain an expansion of k in terms of the eigenfunctions of the integral operator T_k (2.38) corresponding to k. For convenience, we briefly review the connection between T_k, the eigenvalues λ_i, and kernels k.

For a given kernel k, the integral operator $(T_k f)(x) := \int_{\mathcal{X}} k(x, x') f(x')\, d\mu(x')$ can be expanded into its eigenvector decomposition $(\lambda_i, \psi_i(x))$, such that

$$k(x, x') = \sum_j \lambda_j \psi_j(x) \psi_j(x') \tag{4.64}$$

holds. Furthermore, the eigensystem of the regularization operator $\Upsilon^*\Upsilon$ is given by $(\lambda_i^{-1}, \psi_i(x))$. The latter tells us the preference of a kernel expansion for specific types of functions (namely the eigenfunctions ψ_j), and the smoothness assumptions made via the size of the eigenvalues λ_i: for instance, large values of λ_i correspond to functions that are weakly penalized.

4.6.1 Conditions for Positivity and Eigenvector Decompositions

In the following we assume that $\mathcal{X}$ is the unit sphere $S_{N-1} \subset \mathbb{R}^N$ and that μ is the uniform measure on S_{N-1}. This takes advantage of the inherent invariances in dot product kernels and it simplifies the notation. We begin with a few definitions.

Legendre Polynomials Denote by $\mathcal{P}_n(\xi)$ the Legendre Polynomials of degree n and by $\mathcal{P}_n^N(\xi)$ the *associated* Legendre Polynomials (see [373] for more details and examples), where $\mathcal{P}_n = \mathcal{P}_n^3$. Without stating the explicit functional form we list some properties we will need:

1. The (associated) Legendre Polynomials form an orthogonal basis with

Orthonormal Basis

$$\int_{-1}^{1} \mathcal{P}_n^N(\xi)\mathcal{P}_m^N(\xi)(1-\xi^2)^{\frac{N-3}{2}}d\xi = \frac{|S_{N-1}|}{|S_{N-2}|}\frac{1}{M(N,n)}\delta_{m,n}. \tag{4.65}$$

Here $|S_{N-1}| = \frac{2\pi^{N/2}}{\Gamma(N/2)}$ denotes the surface of S_{N-1}, and $M(N,n)$ denotes the multiplicity of spherical harmonics of order n on S_{N-1}, which is given by $M(N,n) = \frac{2n+N-2}{n}\binom{n+N-3}{n-1}$.

2. We can find an expansion for any analytic function $k(\xi)$ on $[-1,1]$ into orthogonal basis functions $\mathcal{P}_n^N$, by[8]

Series Expansion

$$k(\xi) = \sum_{n=0}^{\infty} M(N,n)\frac{|S_{N-2}|}{|S_{N-1}|}\int_{-1}^{1} k(\xi)P_n^N(\xi)(1-\xi^2)^{\frac{N-3}{2}}d\xi. \tag{4.66}$$

3. The Legendre Polynomials may be expanded into an orthonormal basis of spherical harmonics $Y_{n,j}^N$ by the Funk-Hecke equation (see [373]), to obtain

$$\mathcal{P}_n^N(\langle x, x'\rangle) = \frac{|S_{N-1}|}{M(N,n)}\sum_{j=1}^{M(N,n)} Y_{n,j}^N(x)Y_{n,j}^N(x'). \tag{4.67}$$

The explicit functional form of $Y_{n,j}^N$ is not important for the further analysis.

Necessary and Sufficient Conditions Below we list conditions, as proven by Schoenberg [466], under which a function $k(\langle x, x'\rangle)$, defined on S_{N-1}, is positive definite. In particular, he proved the following two theorems:

Theorem 4.18 (Dot Product Kernels in Finite Dimensions) *A kernel $k(\langle x, x'\rangle)$ defined on $S_{N-1} \times S_{N-1}$ is positive definite if and only if its expansion into Legendre polynomials $\mathcal{P}_n^N$ has only nonnegative coefficients, i.e.*

Legendre Expansion

$$k(\xi) = \sum_{n=0}^{\infty} b_n \mathcal{P}_n^N(\xi) \text{ with } b_n \geq 0. \tag{4.68}$$

Theorem 4.19 (Dot Product Kernels in Infinite Dimensions) *A kernel $k(\langle x, x'\rangle)$ defined on the unit sphere in a Hilbert space is positive definite if and only if its Tay-*

Taylor Series Expansion

8. Typically, computer algebra programs can be used to find such expansions for given kernels k. This greatly reduces the problems in the analysis of such kernels.

lor series expansion has only nonnegative coefficients;

$$k(\xi) = \sum_{n=0}^{\infty} a_n \xi^n \text{ with } a_n \geq 0. \tag{4.69}$$

Therefore, all we have to do in order to check whether a particular kernel may satisfy Mercer's condition, is to look at its polynomial series expansion, and check the coefficients.

We note that (4.69) is a more stringent condition than (4.68). In other words, in order to prove positive definiteness for arbitrary dimensions it suffices to show that the Taylor expansion contains only positive coefficients. On the other hand, in order to prove that a candidate for a kernel function will never be positive definite, it is sufficient to show this for (4.68) where $\mathcal{P}_n^N = \mathcal{P}_n$, i.e. for the Legendre Polynomials.

Eigenvector Decomposition We conclude this section with an explicit representation of the eigensystem of $k(\langle x, x' \rangle)$. For a proof see [511].

Lemma 4.20 (Eigenvector Decomposition of Dot Product Kernels) *Denote by $k(\langle x, x' \rangle)$ a kernel on $S_{N-1} \times S_{N-1}$ satisfying condition (4.68) of Theorem 4.18. Then the eigenvectors of k are given by*

$$\psi_{n,j} = Y_{n,j}^N, \text{ with eigenvalues } \lambda_{n,j} = a_n \frac{|S_{N-1}|}{M(N,n)} \text{ of multiplicity } M(N,n). \tag{4.70}$$

In other words, $\frac{a_n}{M(N,n)}$ determines the regularization properties of $k(\langle x, x' \rangle)$.

4.6.2 Examples and Applications

In the following we will analyze a few kernels, and state under which conditions they may be used as SV kernels.

Example 4.21 (Homogeneous Polynomial Kernels $k(x, x') = \langle x, x' \rangle^p$) *As we showed Chapter 2, this kernel is positive definite for $p \in \mathbb{N}$. We will now show that for $p \notin \mathbb{N}$ this is never the case.*

We thus have to show that (4.68) cannot hold for an expansion in terms of Legendre Polynomials ($N = 3$). From [209, 7.126.1], we obtain for $k(\xi) = |\xi|^p$ (we need $|\xi|$ to make k well-defined),

$$\int_{-1}^{1} \mathcal{P}_n(\xi) |\xi|^p d\xi = \frac{\sqrt{\pi}\Gamma(p+1)}{2^p \Gamma\left(1 + \frac{p}{2} - \frac{n}{2}\right) \Gamma\left(\frac{3}{2} + \frac{p}{2} + \frac{n}{2}\right)} \text{ if } n \text{ even.} \tag{4.71}$$

For odd n, the integral vanishes, since $\mathcal{P}_n(-\xi) = (-1)^n \mathcal{P}_n(\xi)$. In order to satisfy (4.68), the integral has to be nonnegative for all n. One can see that $\Gamma\left(1 + \frac{p}{2} - \frac{n}{2}\right)$ is the only term in (4.71) that may change its sign. Since the sign of the Γ function alternates with period 1 for $x < 0$ (and has poles for negative integer arguments), we cannot find any p for which $n = 2\lfloor \frac{p}{2} + 1 \rfloor$ and $n = 2\lceil \frac{p}{2} + 1 \rceil$ correspond to positive values of the integral.

Example 4.22 (Inhomogeneous Polynomial Kernels $k(x, x') = \langle x, x' \rangle + 1)^p$) *Likewise, let us analyze $k(\xi) = (1 + \xi)^p$ for $p > 0$. Again, we expand k in a series of Legendre Polynomials, to obtain [209, 7.127]*

$$\int_{-1}^{1} \mathcal{P}_n(\xi)(\xi + 1)^p d\xi = \frac{2^{p+1}\Gamma^2(p+1)}{\Gamma(p+2+n)\Gamma(p+1-n)}. \tag{4.72}$$

For $p \in \mathbb{N}$, all terms with $n > p$ vanish, and the remainder is positive. For non-integer p, however, (4.72) may change its sign. This is due to $\Gamma(p + 1 - n)$. In particular, for any $p \notin \mathbb{N}$ (with $p > 0$), we have $\Gamma(p + 1 - n) < 0$ for $n = \lceil p \rceil + 1$. This violates condition (4.68), hence such kernels cannot be used in SV machines unless $p \in \mathbb{N}$.

Example 4.23 (Vovk's Real Polynomial $k(x, y) = \frac{1-\langle x,y\rangle^p}{1-\langle x,y\rangle}$ with $p \in \mathbb{N}$ [459]) *This kernel can be written as $k(\xi) = \sum_{n=0}^{p-1} \xi^n$, hence all the coefficients $a_i = 1$, which means that the kernel can be used regardless of the dimensionality of the input space.*

Likewise we can analyze an infinite power series.

Example 4.24 (Vovk's Infinite Polynomial $k(x, x') = (1 - (\langle x, x' \rangle))^{-1}$ [459]) *This kernel can be written as $k(\xi) = \sum_{n=0}^{\infty} \xi^n$, hence all the coefficients $a_i = 1$. The flat spectrum of the kernel suggests poor generalization properties.*

Example 4.25 (Neural Network Kernels $k(x, x') = \tanh(a + \langle x, x' \rangle)$) *We next show that $k(\xi) = \tanh(a + \xi)$ is never positive definite, no matter how we choose the parameters.*

The technique is identical to that of Examples 4.21 and 4.22: we have to show that the kernel does not satisfy the conditions of Theorem 4.18. Since this is very technical (and is best done using computer algebra programs such as Maple), we refer the reader to [401] for details, and explain how the method works in the simpler case of Theorem 4.19. Expanding $\tanh(a + \xi)$ into a Taylor series yields

$$\tanh a + \xi \tfrac{1}{\cosh^2 a} - \xi^2 \tfrac{\tanh a}{\cosh^2 a} - \tfrac{\xi^3}{3}(1 - \tanh^2 a)(1 - 3\tanh^2 a) + O(\xi^4). \tag{4.73}$$

We now analyze (4.73) coefficient-wise. Since the coefficients have to be nonnegative, we obtain $a \in [0, \infty)$ from the first term, $a \in (-\infty, 0]$ from the third term, and $|a| \in [\operatorname{arctanh} \frac{1}{3}, \operatorname{arctanh} 1]$ from the fourth term . This leaves us with $a \in \emptyset$, hence there are no parameters for which this kernel is positive definite.

4.7 Multi-Output Regularization

So far in this chapter we only considered scalar functions $f : \mathcal{X} \to \mathcal{Y}$. Below we will show that under rather mild assumptions on the symmetry properties of $\mathcal{Y}$, there exist no other vector valued extensions to $\Upsilon^*\Upsilon$ than the trivial extension, i.e., the application of a scalar regularization operator to each of the dimensions of $\mathcal{Y}$ separately. The reader not familiar with group theory may want to skip the more detailed discussion given below.

The type of regularization we study are quadratic functionals $\Omega[f]$. Ridge regression, RKHS regularizers and also Gaussian Processes are examples of such regularization. Our proofs rely on a result from [509] which is stated without proof.

Proposition 4.26 (Homogeneous Invariant Regularization [509]) *Any regularization term $\Omega[f]$ that is both homogeneous quadratic, and invariant under an irreducible orthogonal representation ρ of the group*[9] *$\mathcal{G}$ on $\mathcal{Y}$; i.e., that satisfies*

$$\Omega[f] \geq 0 \text{ for all } f \in \mathcal{F}, \tag{4.74}$$

$$\Omega[af] = |a|^2 \Omega[f] \text{ for all scalars } a, \tag{4.75}$$

$$\Omega[\rho(g)f] = \Omega[f] \text{ for all } g \in \mathcal{G}, \tag{4.76}$$

is of the form

$$\Omega[f] = \langle \Upsilon f, \Upsilon f \rangle, \text{ where } \Upsilon \text{ is a scalar operator.} \tag{4.77}$$

Positivity

The motivation for the requirements (4.74) to (4.76) can be seen as follows: the necessity that a regularization term be positive (4.74) is self evident — it must at least be bounded from below. Otherwise we could obtain arbitrarily "good" estimates by exploiting the pathological behavior of the regularization operator. Hence, via a positive offset, $\Omega[f]$ can be transformed such that it satisfies the positivity condition (4.74).

Homogeneity

Homogeneity (4.75) is a useful condition for efficient capacity control — it allows easy capacity control by noting that the entropy numbers (a quantity to be introduced in Chapter 12), which are a measure of the size of the set of possible solutions, scale in a linear (hence, homogeneous) fashion when the hypothesis class is rescaled by a constant. Practically speaking, this means that we do not need new capacity bounds for every scale the function f might assume. The requirement of being quadratic is merely algorithmic, as it allows to avoid taking absolute values in the linear or cubic case to ensure positivity, or when dealing with derivatives.

Invariance

Finally, the invariance must be chosen beforehand. If it happens to be sufficiently strong, it can rule out all operators but scalars. Permutation symmetry is such a case; in classification, for instance, this would mean that all class labels are treated equally.

No Vector Valued Regularizer

A consequence of the proposition is that there exists no vector valued regularization operator satisfying the invariance conditions. We now look at practical applications of Proposition 4.26, which will be stated in the form of corollaries.

Corollary 4.27 (Permutation and Rotation Symmetries) *Under the assumptions of Proposition 4.26, both the canonical representation of the permutation group (by permutation matrices) in a finite dimensional vector space $\mathcal{Y}$, and the group of orthogonal transformations on $\mathcal{Y}$, enforce scalar operators Υ.*

9. $\mathcal{G}$ also may be directly defined on $\mathcal{Y}$, i.e. it might be a matrix group like SU(N).

Permutation and Rotation Symmetries are Irreducible

This follows immediately from the fact that both rotations and permutations (or more precisely their representations on $\mathcal{Y}$), are unitary and irreducible on $\mathcal{Y}$ by construction. For instance if the permutation group was reducible on $\mathcal{Y}$, then there would exist subspaces on $\mathcal{Y}$ which do not change under any permutation on $\mathcal{Y}$. This is impossible, however, since we are considering the group of all possible permutations over $\mathcal{Y}$. Finally, permutations are a subgroup of the group of all possible orthogonal transformations.

Let us now address the more practical side of such operators, namely how they translate into function expansions. We need only evaluate $\langle \Upsilon \alpha f, \alpha' f' \rangle$, where f, f' are scalar function and $\alpha, \alpha' \in \mathcal{Y}$. Since Υ is also scalar, this yields $\langle \alpha, \alpha' \rangle \langle \Upsilon f, \Upsilon f' \rangle$. It then remains to evaluate $\Omega[f]$ for a kernel expansion of f. We obtain:

Corollary 4.28 (Kernel Expansions) *Under the assumptions of proposition 4.26, the regularization functional $\Omega[f]$ for a kernel expansion*

$$f(x) = \sum_i \alpha_i k(x_i, x), \quad \text{with} \quad \alpha_i \in \mathcal{Y}, \tag{4.78}$$

where $k(x_i, x)$ is a function mapping $\mathcal{X} \times \mathcal{X}$ to the space of scalars $\mathcal{S}$, compatible with the dot product space $\mathcal{Y}$ (we require that $\beta\alpha \in \mathcal{Y}$ for $\alpha \in \mathcal{Y}$ and $\beta \in \mathcal{S}$) can be stated

$$\Omega[f] = \sum_{i,j} \left\langle \alpha_i, \alpha_j \right\rangle \left\langle \Upsilon k(x_i, \cdot), \Upsilon k(x_j, \cdot) \right\rangle. \tag{4.79}$$

In particular, if k is the Green's function of $\Upsilon^* \Upsilon$, we get

$$\Omega[f] = \sum_{i,j} \left\langle \alpha_i, \alpha_j \right\rangle k(x_i, x_j). \tag{4.80}$$

For possible applications such as regularized principal manifolds, see Chapter 17.

4.8 Semiparametric Regularization

Preference for Parametric Part

Understandable Model

In some cases, we may have additional knowledge about the solution we are going to encounter. In particular, we may know that a specific *parametric* component is very likely going to be part of the solution. It would be unwise not to take advantage of this extra knowledge. For instance, it might be the case that the major properties of the data are described by a combination of a small set of linearly independent basis functions $\{\phi_1(\cdot), \ldots, \phi_n(\cdot)\}$. Or we might want to correct the data for some (e.g. linear) trends. Second, it may also be the case that the user wants to have an *understandable* model, without sacrificing accuracy. Many people in life sciences tend to have a preference for linear models. These reasons motivate the construction of *semiparametric* models, which are both easy to understand (due to the parametric part) and perform well (often thanks to the nonparametric term). For more advantages and advocacy on semiparametric models, see [47].

A common approach is to fit the data with the parametric model and train the nonparametric add-on using the errors of the parametric part; that is, we fit the

nonparametric part to the errors. We will show that this is useful only in a very restricted situation. In general, this method does not permit us to find the best model amongst a given class for different loss functions. It is better instead to solve a convex optimization problem, as in standard SVMs, but with a different set of admissible functions;

Backfitting vs. Global Solution

$$f(x) = g(x) + \sum_{i=1}^{n} \beta_i \phi_i(x). \tag{4.81}$$

Here $g \in \mathcal{H}$, where $\mathcal{H}$ is a Reproducing Kernel Hilbert Space as used in Theorem 4.3. In particular, this theorem implies that there exists a mixed expansion in terms of kernel functions $k(x_i, x)$ *and* the parametric part ϕ_i.

Keeping the standard regularizer $\Omega[f] = \frac{1}{2}\|f\|^2_{\mathcal{H}}$, we can see that there exist functions $\phi_1(\cdot), \ldots, \phi_n(\cdot)$ whose contribution is not regularized at all. This need not be a major concern if n is sufficiently smaller than m, as the VC dimension (and thus the capacity) of this additional class of linear models is n, hence the overall capacity control will still work, provided the nonparametric part is sufficiently restricted.

Capacity Control

The Algorithm

We will show, in the case of SV regression, how the semiparametric setting translates into optimization problems. The application to classification is straightforward, and is left as an exercise (see Problem 4.8).

Formulating the optimization equations for the expansion (4.81), using the ε-insensitive loss function, and introducing kernels, we arrive at the following primal optimization problem:

Primal Objective Function

$$\begin{aligned} &\text{maximize} \quad \tfrac{\lambda}{2}\|\mathbf{w}\|^2 + \sum_{i=1}^{m} \xi_i + \xi_i^*, \\ &\text{subject to} \quad \begin{cases} \langle \mathbf{w}, \psi(x_i)\rangle + \sum_{j=1}^{n} \beta_j \phi_j(x_i) - y_i & \leq \epsilon + \xi_i^*, \\ y_i - \langle \mathbf{w}, \psi(x_i)\rangle - \sum_{j=1}^{n} \beta_j \phi_j(x_i) & \leq \epsilon + \xi_i, \\ \xi_i, \xi_i^* & \geq 0. \end{cases} \end{aligned} \tag{4.82}$$

Computing the Lagrangian (we introduce $\alpha_i, \alpha_i^*, \eta_i, \eta_i^*$ for the constraints) and solving for the Wolfe dual, yields[10]

Dual Objective Function

$$\begin{aligned} &\text{maximize} \quad \begin{cases} -\frac{1}{2}\sum_{i,j=1}^{m} (\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*) k(x_i, x_j), \\ -\varepsilon \sum_{i=1}^{m} (\alpha_i + \alpha_i^*) + \sum_{i=1}^{m} y_i (\alpha_i - \alpha_i^*), \end{cases} \\ &\text{subject to} \quad \begin{cases} \sum_{i=1}^{m} (\alpha_i - \alpha_i^*)\phi_j(x_i) & = 0 \text{ for all } 1 \leq j \leq n, \\ \alpha_i, \alpha_i^* & \in [0, 1/\lambda]. \end{cases} \end{aligned} \tag{4.83}$$

10. See also (1.26) for details how to formulate the Lagrangian.

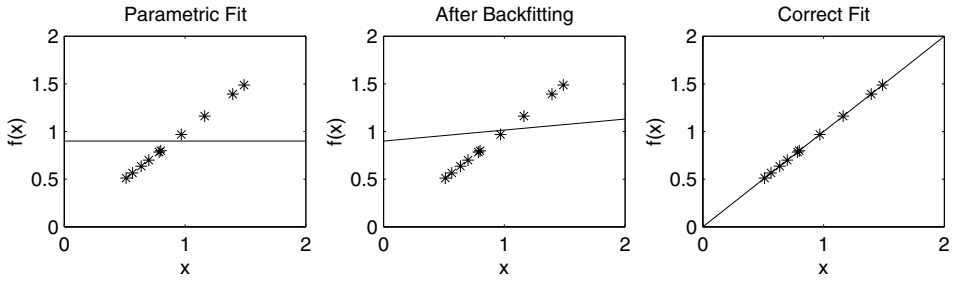


Figure 4.10 Backfitting of a model with two parameters, $f(x) = wx + \beta$. Data was generated by taking 10 samples from the uniform distribution on $\left[\frac{1}{2}, \frac{3}{2}\right]$. The target values were obtained by the dependency $y_i = x_i$. From left to right: (left) best fit with the parametric model of a constant function; (middle) after adaptation of the second parameter while keeping the first parameter fixed; (right) optimal fit with both parameters.

Note the similarity to the standard SV regression model. The objective function, and the box constraints on the Lagrange multipliers α_i, α_i^*, remain unchanged. The only modification comes from the additional un-regularized basis functions. Instead of a single (constant) function $\phi_1(x) = 1$ as in the standard SV case, we now have an expansion in the basis $\beta_i \phi_i(\cdot)$. This gives rise to n constraints instead of one. Finally, f can be found as

Semiparametric Kernel Expansion

$$f(x) = \sum_{i=1}^{m} (\alpha_i - \alpha_i^*) k(x_i, x) + \sum_{i=1}^{n} \beta_i \phi_i(x) \text{ since } \mathbf{w} = \sum_{i=1}^{m} (\alpha_i - \alpha_i^*) \psi(x_i). \tag{4.84}$$

The only difficulty remaining is how to determine β_i. This can be done by exploiting the Karush-Kuhn-Tucker optimality conditions in an analogous manner to (1.30), or more easily, by using an interior point algorithm (Section 6.4). In the latter case, the variables β_i can be obtained as the dual variables of the dual (dual dual = primal) optimization problem (4.83), as a by-product of the optimization process.

Why Backfitting Is Not Sufficient

It might seem that the approach presented above is quite unnecessary, and overly complicated for semiparametric modelling. In fact, we could try to fit the data to the parametric model first, and then fit the nonparametric part to the residuals; this approach is called backfitting. In most cases, however, this does not lead to the minimum of the regularized risk functional. We will show this using a simple example.

Backfitting for SVMs

Consider a SV regression machine as defined in Section 1.6, with linear kernel (i.e. $k(x, x') = \langle x, x' \rangle$) in one dimension, and a constant term as parametric part (i.e. $f(x) = wx + \beta$). Now suppose the data was generated by $y_i = x_i$, where x_i is uniformly drawn from $\left[\frac{1}{2}, \frac{3}{2}\right]$ without noise. Clearly, $y_i \geq \frac{1}{2}$ also holds for all i. By construction, the best overall fit of the pair (β, w) will be arbitrarily close to $(0, 1)$ if the regularization parameter λ is chosen sufficiently small. For backfitting, we first carry out the parametric fit, to find a constant β minimizing the term $\sum_{i=1}^{m} c(y_i - \beta)$. Depending on the chosen loss function $c(\cdot)$, β will be the mean (L_2-error), the median (L_1-error), a trimmed mean (related to the ε-insensitive loss), or

Coordinate Descent

some other function of the set $\{y_1 - wx_1, \ldots, y_m - wx_m\}$ (cf. Section 3.4). Since all $y_i \geq 1$, we have $\beta \geq 1$; this is not the optimal solution of the overall problem, since in the latter case β would be close to 0, as seen above.

Hence backfitting does not minimize the regularized risk functional, even in the simplest of settings; and we certainly cannot expect backfitting to work in more complex cases. There exists only one case in which backfitting suffices, namely if the function spaces spanned by the kernel expansion $\{k(x_i, \cdot)\}$ and $\{\phi_i(\cdot)\}$ are orthogonal. Consequently we must in general jointly solve for both the parametric and the nonparametric part, as done in (4.82) and (4.83).

Orthogonal Decomposition

Above, we effectively excluded a set of basis functions $\phi_1, \ldots, \phi_n$ from being regularized at all. This means that we could use regularization functionals $\Omega[f]$ that need not be positive definite on the whole Reproducing Kernel Hilbert Space $\mathcal{H}$ but only on the orthogonal complement to span $\{\phi_1, \ldots \phi_n\}$.

$\Omega[f]$ for Subspaces

This brings us back to the notion of conditional positive definite kernels, as explained in Section 2.2. These exclude the space of linear functions from the space of admissible functions f, in order to achieve a positive definite regularization term $\Omega[f]$ on the orthogonal complement.

In (4.83), this is precisely what happens with the functions ϕ_i, which are not supposed to be regularized. Consequently, if we choose ϕ_i to be the family of all linear functions, the semiparametric approach will allow us to use conditionally positive definite (cpd) kernels (see Definition 2.21 and below) without any further problems.

Connecting CPD Kernels and Semiparametric Models

4.9 Coefficient Based Regularization

Most of the discussion in the current chapter was based on regularization in Reproducing Kernel Hilbert Spaces, and explicitly avoided any specific restrictions on the type of coefficient expansions used. This is useful insofar as it provides a powerful mathematical framework to assess the quality of the estimates obtained in this process.

In some cases, however, we would rather use a regularization operator that acts *directly* on coefficient space, be it for theoretical reasons (see Section 16.5), or to satisfy the practical desire to obtain sparse expansions (Section 4.9.2); or simply by the heuristic that small coefficients generally translate into simple functions.

Function Space vs. Coefficient Space

We will now consider the situation where $\Omega[f]$ can be written as a function of the coefficients α_i, where f will again be expanded as a linear combination of kernel functions,

General Kernel Expansion

$$f(x) = \sum_{i=1}^{n} \alpha_i k(x_i', x) \text{ and } \Omega[f] = \Omega[\alpha], \tag{4.85}$$

but with the possibility that x_i' and the training patterns x_i do not coincide, and that possibly $m \neq n$.

4.9.1 Ridge Regression

A popular choice to regularize linear combinations of basis functions is by a weight decay term (see [339, 49] and the references therein), which penalizes large weights. Thus we choose

$$\Omega[f] := \frac{1}{2}\sum_{i=1}^{n}\alpha_i^2 = \frac{1}{2}\|\boldsymbol{\alpha}\|^2. \tag{4.86}$$

Weight Decay

This is also called Ridge Regression [245, 377], and is a very common method in the context of shrinkage estimators.

Similar to Section 4.3, we now investigate whether there exists a correspondence between Ridge Regression and SVMs. Although no strict equivalence holds, we will show that it is possible to obtain models generated by the same type of regularization operator. The requirement on an operator Υ for a strict equivalence would be

$$\Omega[f] = \frac{1}{2}\sum_{i,j=1}^{n}\left\langle (\Upsilon k)(x_i,.),(\Upsilon k)(x_j,.)\right\rangle \alpha_i\alpha_j = \frac{1}{2}\sum_{i=1}^{n}\alpha_i^2, \tag{4.87}$$

Equivalence Condition

and thus,

$$\left\langle (\Upsilon k)(x_i,.),(\Upsilon k)(x_j,.)\right\rangle = \delta_{ij}. \tag{4.88}$$

Unfortunately this requirement is not suitable for the case of the Kronecker δ, as (4.88) implies the functions $(\Upsilon k)(x_i,\cdot)$ to be elements of a non-separable Hilbert space. The solution is to change the finite Kronecker δ into the more appropriate δ-distribution, i.e. $\delta(x_i - x_j)$.

By reasoning similar to Theorem 4.9, we can see that (4.88) holds, with $k(x,x')$ the Green's function of Υ. Note that as a regularization operator, $(\Upsilon^*\Upsilon)^{\frac{1}{2}}$ is equivalent to Υ, as we can always replace the latter by the former without any difference in the regularization properties. Therefore, we assume without loss of generality that Υ is a positive definite operator. Formally, we require

Equivalent Operator

$$\left\langle (\Upsilon k)(x_i,.),(\Upsilon k)(x_j,.)\right\rangle = \left\langle \delta_{x_i}(.),\delta_{x_j}(.)\right\rangle = \delta_{x_i,x_j}. \tag{4.89}$$

Again, this allows us to connect regularization operators and kernels: the Green's function of Υ must be found in order to satisfy (4.89). For the special case of translation invariant operators represented in Fourier space, we can associate Υ with $\Upsilon_{\text{ridge}}(\boldsymbol{\omega})$ as with (4.28), leading to

$$\|\Upsilon f\|_2^2 = \int \left|\frac{F[f](\boldsymbol{\omega})}{\Upsilon_{\text{ridge}}(\boldsymbol{\omega})}\right|^2 d\boldsymbol{\omega}. \tag{4.90}$$

This expansion is possible since the Fourier transform diagonalizes the corresponding regularization operator: repeated applications of Υ become multiplications in the Fourier domain. Comparing (4.90) with (4.28) leads to the conclusion that the following relation between kernels for Support Vector Machines and

Ridge Regression holds,

$$\Upsilon_{\mathrm{SV}}(\omega) = |\Upsilon_{\mathrm{ridge}}(\omega)|^2. \tag{4.91}$$

In other words, in Ridge Regression it is the *squared* Fourier transform of the kernels that determines the regularization properties. Later on in Chapter 16, Theorem 16.9 will give a similar result, derived under the assumption that the penalties on α_i are given by a prior probability over the distribution of expansion coefficients.

This connection also explains the performance of Ridge Regression Models in a smoothing regularizer context (the squared norm of the Fourier transform of the kernel function describes its regularization properties), and allows us to "transform" Support Vector Machines to Ridge Regression models and vice versa. Note, however, that the sparsity properties of Support Vectors are lost.

4.9.2 Linear Programming Regularization (ℓ_1^m)

ℓ_1 for Sparsity

A squared penalty on the coefficients α_i has the disadvantage that even though some kernel functions $k(x_i, x)$ may not contribute much to the overall solution, they still appear in the function expansion. This is due to the fact that the gradient of α_i^2 tends to 0 for $\alpha_i \to 0$ (this can easily be checked by looking at the partial derivative of $\Omega[f]$ wrt. α_i). On the other hand, a regularizer whose derivative does not vanish in the neighborhood of 0 will not exhibit such problems. This is why we choose

$$\Omega[f] = \sum_i |\alpha_i|. \tag{4.92}$$

The regularized risk minimization problem can then be rewritten as

$$\begin{aligned} &\text{minimize} \quad R_{\mathrm{reg}}[f] = \lambda \sum_{i=1}^m |\alpha_i| + \sum_{i=1}^m (\xi_i + \xi_i^*), \\ &\text{subject to} \quad \begin{cases} y_i - \sum_{j=1}^m \alpha_j k(x_j, x_i) - \sum_{j=1}^n \phi_j(x_i) - b & \leq \varepsilon + \xi_i, \\ \sum_{j=1}^m \alpha_i k(x_j, x_i) + \sum_{j=1}^n \phi_j(x_i) + b - y_i & \leq \varepsilon + \xi_i^*, \\ \xi_i, \xi_i^* & \geq 0. \end{cases} \end{aligned} \tag{4.93}$$

Soft Margin $\to$ Linear Program

Besides replacing α_i with $\alpha_i - \alpha_i^*$, $|\alpha_i|$ with $\alpha_i + \alpha_i^*$, and requiring $\alpha_i, \alpha_i^* \geq 0$, there is hardly anything that can be done to render the problem more computationally feasible — the constraints are already linear. Moreover most optimization software can deal efficiently with problems of this kind.

4.9.3 Mixed Semiparametric Regularizers

We now investigate the use of mixed regularization functionals, with different penalties for distinct parts of the function expansion, as suggested by equations (4.92) and (4.81). Indeed, we can construct the following variant, which is a mix-

ture of linear and quadratic regularizers,

$$\Omega[f] = \frac{1}{2}||\mathbf{w}||^2 + \sum_{i=1}^{n} |\beta_i|. \tag{4.94}$$

The equation above is essentially the SV estimation model, with an additional linear regularization term added for the parametric part. In this case, the constraints on the optimization problem (4.83) become

Mixed Dual Problem

$$\begin{array}{rcccl} -1 & \leq & \sum_{i=1}^{m} (\alpha_i - \alpha_i^*)\phi_j(x_i) & \leq & 1 \qquad \text{for all } 1 \leq j \leq n, \\ & & \alpha_i, \alpha_i^* & \in & [0, 1/\lambda], \end{array} \tag{4.95}$$

and the variables β_i are obtained as the dual variables of the constraints, as discussed previously in similar cases. Finally, we could reverse the setting to obtain a regularizer,

Semiparametric and Sparse

$$\Omega[f] = \sum_{i=1}^{m} |\alpha_i - \alpha_i^*| + \frac{1}{2} \sum_{i,j=1}^{n} \beta_i \beta_j M_{ij}, \tag{4.96}$$

for some positive definite matrix M. Note that (4.96) can be reduced to the case of (4.94) by renaming variables accordingly, given a suitable choice of M.

The proposed regularizers are a simple extension of existing methods such as Basis Pursuit [104], or Linear Programming for classification (e.g. [184]). The common idea is to have two different sets of basis functions which are regularized differently, or a subset that is not regularized at all. This is an efficient way of encoding prior knowledge or user preference, since the emphasis is on the functions with little or no regularization.

Finally, one could also use a regularization functional $\Omega[f] = \|\alpha\|_0$ which simply counts the number of nonzero terms in the vector $\alpha \in \mathbb{R}^m$, or alternatively, combine this regularizer with the ℓ_1 norm to obtain $\Omega[f] = \|\alpha\|_0 + \|\alpha\|_1$. This is a *concave* function in α, which, in combination with the soft-margin loss function, leads to an optimization problem which is, as a whole, concave. Therefore one may apply Rockafellar's theorem (Theorem 6.12) to obtain an optimal solution. See [189] for further details and an explicit algorithm.

4.10 Summary

A connection between Support Vector kernels and regularization operators has been established, which can provide one key to understanding why Support Vector Machines have been found to exhibit high generalization ability. In particular, for common choices of kernels, the mapping into feature space is not arbitrary, but corresponds to useful regularization operators (see Sections 4.4.1, 4.4.2 and 4.4.4). For kernels where this is not the case, Support Vector Machines may show poor performance (Section 4.4.3). This will become more obvious in Section 12, where, building on the results of the current chapter, the eigenspectrum of integral opera-

tors is connected with generalization bounds of the corresponding Support Vector Machines.

The link to regularization theory can be seen as a tool for determining the structure, consisting of sets of functions, in which Support Vector Machines and other kernel algorithms (approximately) perform structural risk minimization [561], possibly in a data dependent manner. In other words, it allows us to choose an appropriate kernel given the data and the problem specific knowledge.

Bayesian Methods

A simple consequence of this link is a Bayesian interpretation of Support Vector Machines. In this case, the choice of a special kernel can be regarded as a prior on the hypothesis space, with $P[f] \propto \exp(-\lambda \|\Upsilon f\|^2)$. See Chapter 16 for more detail on this matter.

Vector Valued Functions

It should be clear by now that the setting of Tikhonov and Arsenin [538], whilst very powerful, is certainly not the only conceivable one. A theorem on vector valued regularization operators showed, however, that under quite generic conditions on the isotropy of the space of target values, only scalar operators are possible; an extended version of their approach is thus the only possible option.

Finally a closer consideration of the null space of regularization functionals $\Omega[f]$ led us to formulate semiparametric models. The roots of such models lie in the representer theorem (Theorem 4.2), proposed and explored in the context of smoothing splines in [296]. In fact, the SV expansion is a direct consequence of the representer theorem.

Semiparametric Models

Moreover the semiparametric setting solves a problem created by the use of *conditionally* positive definite kernels of order q (see Section 2.4.3). Here, polynomials of order lower than q are excluded. Hence, to cope with this effect, we must add polynomials back in "manually." The semiparametric approach presents a way of doing that. Another application of semiparametric models, besides the conventional approach of treating the nonparametric part as *nuisance parameters* [47], is in the domain of hypothesis testing, for instance to test whether a parametric model fits the data sufficiently well. This can be achieved in the framework of structural risk minimization [561] — given the different models (nonparametric vs. semiparametric vs. parametric), we can evaluate the bounds on the expected risk, and then choose the model with the best bound.

4.11 Problems

4.1 (Equivalent Optimization Strategies ●●●) *Denote by S a metric space and by $R, \Omega : S \to \mathbb{R}$ two strictly convex continuous maps. Let $\lambda > 0$.*

- *Show that the map $f \mapsto R[f] + \lambda\Omega[f]$ has only one minimum and a unique minimizer. Hint: assume the contrary and consider a straight line between two minima.*
- *Show that for every $\lambda > 0$, there exists an Ω_λ such that minimization of $R[f] + \lambda\Omega[f]$, is equivalent to minimizing $R[f]$ subject to $\Omega[f] \leq \Omega_\lambda$. Show that an analogous statement holds with R and Ω exchanged. Hint: consider the minimizer of $R[f] + \lambda\Omega[f]$, and keep*

the second term fixed while minimizing over the first term.

- *Consider the parametrized curve* $(\Omega(\lambda), R(\lambda))$*. What is the shape of this curve? Show that (barring discontinuities)* $-\lambda$ *is the tangent on the curve.*
- *Consider the parametrized curve* $(\ln \Omega(\lambda), \ln R(\lambda))$ *as proposed by Hansen [225]. Show that a tangent criterion similar to that imposed above is scale insensitive wrt.* Ω *and* R*. Why is this useful? What are the numerical problems with such an ansatz?*

4.2 (Orthogonality and Span ●●) *Show that the second condition of Definition 2.9 is equivalent to requiring*

$$\langle f, k(x, \cdot)\rangle_{\mathcal{H}} = 0 \text{ for all } x \in \mathcal{X} \Longleftrightarrow f = 0. \tag{4.97}$$

4.3 (Semiparametric Representer Theorem ●●) *Prove Theorem 4.3. Hint: start with a decomposition of* $\tilde{f}$ *into a parametric part, a kernel part, and an orthogonal contribution and evaluate the loss and regularization terms independently.*

4.4 (Kernel Boosting ●●●) *Show that for* $f \in \mathcal{H}$ *and* $c(x, y, f(x)) = \exp(-yf(x))$*, you can develop a boosting algorithm by performing a coefficient-wise gradient descent on the coefficients* α_i *of the expansion* $f(x) = \sum_{i=1}^{m} \alpha_i k(x_i, x)$*. In particular, show that the expansion above is optimal.*

What changes if we drop the regularization term $\Omega[f] = ||f||^2$*? See [498, 577, 221] for examples.*

4.5 (Monotonicity of the Regularizer ●●) *Give an example where, due to the fact that* $\Omega[f]$ *is not strictly monotonic the kernel expansion (4.5) is not the only minimizer of the regularized risk functional (4.4).*

4.6 (Sparse Expansions ●●) *Show that it is a sufficient requirement for the coefficients* α_i *of the kernel expansion of the minimizer of (4.4) to vanish, if for the corresponding loss functions* $c(x_i, y_i, f(x_i))$ *both the lhs and the rhs derivative with respect to* $f(x_i)$ *vanish. Hint: use the proof strategy of Theorem 4.2.*

Furthermore show that for loss functions $c(x, y, f(x))$ *this implies that we can obtain vanishing coefficients only if* $c(x_i, y_i, f(x_i)) = 0$*.*

4.7 (Biased Regularization ●●) *Show that for biased regularization (Remark 4.4) with* $g(||f||_{\mathcal{H}}) = \frac{1}{2}||f||^2_{\mathcal{H}}$*, the effective overall regularizer is given by* $\frac{1}{2}||f - f_0||^2$*.*

4.8 (Semiparametric Classification ●●) *Show that given a set of parametric basis functions* ϕ_i*, the optimization problem for SV classification has the same objective function as (1.31), however with the constraints [506]*

$$0 \leq \alpha_i \leq C \text{ for all } i \in [m] \text{ and } \sum_{i=1}^{m} \alpha_i y_i \phi_j(x_i) = 0 \text{ for all } j. \tag{4.98}$$

What happens if you combine semiparametric classification with adaptive margins (the ν*-trick)?*

4.9 (Regularization Properties of Kernels •) *Analyze the regularization properties of the Laplacian kernel $k(x, x') = e^{-|x-x'|}$. What is the rate of decay in its power spectrum? What is the kernel corresponding to the operator*

$$||\Upsilon f||^2 := ||f||^2 + ||\partial_x f||^2 + ||\partial_x^2 f||^2? \tag{4.99}$$

Hint: rewrite Υ in the Fourier domain.

4.10 (Periodizing the Laplacian Kernel •) *Show that for the Laplacian kernel $k(x, x') = e^{-|x-x'|}$, the periodization with period a results in a kernel proportional to*

$$k_p(x, x') = e^{-\left[|x-x'| \bmod a\right]} + e^{-\left[|x-x'| \bmod a\right]+a}. \tag{4.100}$$

4.11 (Hankel Transform and Inversion •••) *Show that for radially symmetric functions, the Fourier transform is given by (4.52). Moreover use (4.51) to prove the Hankel inversion theorem, stating that H_ν is its own inverse.*

4.12 (Eigenvector Decompositions of Polynomial Kernels •••) *Compute the eigenvalues of polynomial kernels on U_N. Hint: use [511] and separate the radial from the angular part in the eigenvector decomposition of k, and solve the radial part empirically via numerical analysis. Possible kernels to consider are Vovk's kernel, (in)homogeneous polynomials and the hyperbolic tangent kernel.*

4.13 (Necessary Conditions for Kernels ••) *Burges [86] shows, by using differential geometric methods, that a necessary condition for a differentiable translation invariant kernel $k(x, x') = k(||x - x'||^2)$ to be positive definite is*

$$k(0) > 0 \text{ and } k'(0) < 0. \tag{4.101}$$

Prove this using functional analytic methods.

4.14 (Mixed Semiparametric Regularizers ••) *Derive (4.96). Hint: set up the primal optimization problem as described in Section 1.4, compute the Lagrangian, and eliminate the primal variables.*

Can you find an interpretation of (4.95)? What is the effect of $\sum_{i=1}^m (\alpha_i - \alpha_i^*)\phi_j(x_i)$?