determined by the m conditions (2.63).

For the converse, assume an arbitrary $\boldsymbol{\alpha} \in \mathbb{R}^m$, and compute

$$\sum_{i,j=1}^{m} \alpha_i \alpha_j K_{ij} = \left\langle \sum_{i=1}^{m} \alpha_i \Phi(x_i), \sum_{j=1}^{m} \alpha_j \Phi(x_j) \right\rangle = \left\| \sum_{i=1}^{m} \alpha_i \Phi(x_i) \right\|^2 \geq 0. \tag{2.64}$$

■

In particular, this result implies that given data $x_1, \ldots, x_m$, and a kernel k which gives rise to a positive definite matrix K, it is always possible to construct a feature space $\mathcal{H}$ of dimension at most m that we are implicitly working in when using kernels (cf. Problem 2.32 and Section 2.2.6).

If we perform an algorithm which requires k to correspond to a dot product in some other space (as for instance the SV algorithms described in this book), it is possible that even though k is not positive definite in general, it still gives rise to a positive definite Gram matrix K with respect to the training data at hand. In this case, Proposition 2.16 tells us that nothing will go wrong during training when we work with these data. Moreover, if k leads to a matrix with some small negative eigenvalues, we can add a small multiple of some strictly positive definite kernel k' (such as the identity $k'(x_i, x_j) = \delta_{ij}$) to obtain a positive definite matrix. To see this, suppose that $\lambda_{\min} < 0$ is the minimal eigenvalue of k's Gram matrix. Note that being strictly positive definite, the Gram matrix K' of k' satisfies

$$\min_{\|\boldsymbol{\alpha}\|=1} \langle \boldsymbol{\alpha}, K'\boldsymbol{\alpha} \rangle \geq \lambda'_{\min} > 0, \tag{2.65}$$

where $\lambda'_{\min}$ denotes its minimal eigenvalue, and the first inequality follows from Rayleigh's principle (B.57). Therefore, provided that $\lambda_{\min} + \lambda\lambda'_{\min} \geq 0$, we have

$$\langle \boldsymbol{\alpha}, (K + \lambda K')\boldsymbol{\alpha} \rangle = \langle \boldsymbol{\alpha}, K\boldsymbol{\alpha} \rangle + \lambda \langle \boldsymbol{\alpha}, K'\boldsymbol{\alpha} \rangle \geq \|\boldsymbol{\alpha}\|^2 \left(\lambda_{\min} + \lambda\lambda'_{\min}\right) \geq 0 \tag{2.66}$$

for all $\boldsymbol{\alpha} \in \mathbb{R}^m$, rendering $(K + \lambda K')$ positive definite.

2.3 Examples and Properties of Kernels

Polynomial

For the following examples, let us assume that $\mathcal{X} \subset \mathbb{R}^N$. Besides homogeneous polynomial kernels (cf. Proposition 2.1),

$$k(x, x') = \langle x, x' \rangle^d, \tag{2.67}$$

Gaussian

Boser, Guyon, and Vapnik [62, 223, 561] suggest the usage of Gaussian radial basis function kernels [26, 4],

$$k(x, x') = \exp\left(-\frac{\|x - x'\|^2}{2\,\sigma^2}\right), \tag{2.68}$$

Sigmoid

where $\sigma > 0$, and sigmoid kernels,

$$k(x, x') = \tanh(\kappa \langle x, x' \rangle + \vartheta), \tag{2.69}$$

where $\kappa > 0$ and $\vartheta < 0$. By applying Theorem 13.4 below, one can check that the latter kernel is not actually positive definite (see Section 4.6 and [85, 511] and the discussion in Example 4.25). Curiously, it has nevertheless successfully been used in practice. The reasons for this are discussed in [467].

Inhomogeneous Polynomial

Other useful kernels include the inhomogeneous polynomial,

$$k(x, x') = \left(\langle x, x' \rangle + c\right)^d, \tag{2.70}$$

($d \in \mathbb{N}, c \geq 0$) and the B_n-spline kernel [501, 572] (I_X denoting the indicator (or characteristic) function on the set X, and $\otimes$ the convolution operation, $(f \otimes g)(x) := \int f(x')g(x' - x)dx'$),

B_n-Spline of Odd Order

$$k(x, x') = B_{2p+1}(\|x - x'\|) \text{ with } B_n := \bigotimes_{i=1}^{n} I_{[-\frac{1}{2}, \frac{1}{2}]}. \tag{2.71}$$

The kernel computes B-splines of order $2p + 1$ ($p \in \mathbb{N}$), defined by the $(2p + 1)$-fold convolution of the unit interval $[-1/2, 1/2]$. See Section 4.4.1 for further details and a regularization theoretic analysis of this kernel.

Invariance of Kernels

Note that all these kernels have the convenient property of unitary invariance, $k(x, x') = k(Ux, Ux')$ if $U^\top = U^{-1}$, for instance if U is a rotation. If we consider complex numbers, then we have to use the adjoint $U^* := \overline{U}^\top$ instead of the transpose.

RBF Kernels

Radial basis function (RBF) kernels are kernels that can be written in the form

$$k(x, x') = f(d(x, x')), \tag{2.72}$$

where d is a metric on $\mathcal{X}$, and f is a function on $\mathbb{R}_0^+$. Examples thereof are the Gaussians and B-splines mentioned above. Usually, the metric arises from the dot product; $d(x, x') = \|x - x'\| = \sqrt{\langle x - x', x - x' \rangle}$. In this case, RBF kernels are unitary invariant, too. In addition, they are translation invariant; in other words, $k(x, x') = k(x + x_0, x' + x_0)$ for all $x_0 \in \mathcal{X}$.

In some cases, invariance properties alone can distinguish particular kernels: in Section 2.1, we explained how using polynomial kernels $\langle x, x' \rangle^d$ corresponds to mapping into a feature space whose dimensions are spanned by all possible dth order monomials in input coordinates. The different dimensions are scaled with the square root of the number of ordered products of the respective d entries (e.g., $\sqrt{2}$ in (2.13)). These scaling factors precisely ensure invariance under the group of all orthogonal transformations (rotations and mirroring operations). In many cases, this is a desirable property: it ensures that the results of a learning procedure do not depend on which orthonormal coordinate system (with fixed origin) we use for representing our input data.

Proposition 2.17 (Invariance of Polynomial Kernels [480]) *Up to a scaling factor, the kernel $k(x, x') = \langle x, x' \rangle^d$ is the only kernel inducing a map into a space of* all *monomials of degree d which is invariant under orthogonal transformations of $\mathbb{R}^N$.*

Properties of RBF Kernels

Some interesting additional structure exists in the case of a Gaussian RBF kernel k (2.68). As $k(x, x) = 1$ for all $x \in \mathcal{X}$, each mapped example has unit length, $\|\Phi(x)\| =$

1 (Problem 2.18 shows how to achieve this for general kernels). Moreover, as $k(x, x') > 0$ for all $x, x' \in \mathcal{X}$, all points lie inside the same orthant in feature space. To see this, recall that for unit length vectors, the dot product (1.3) equals the cosine of the enclosed angle. We obtain

$$\cos(\angle(\Phi(x), \Phi(x'))) = \langle \Phi(x), \Phi(x') \rangle = k(x, x') > 0, \tag{2.73}$$

which amounts to saying that the enclosed angle between any two mapped examples is smaller than $\pi/2$.

The above seems to indicate that in the Gaussian case, the mapped data lie in a fairly restricted area of feature space. However, in another sense, they occupy a space which is as large as possible:

Theorem 2.18 (Full Rank of Gaussian RBF Gram Matrices [360]) *Suppose that $x_1, \ldots, x_m \subset \mathcal{X}$ are distinct points, and $\sigma \neq 0$. The matrix K given by*

$$K_{ij} := \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right) \tag{2.74}$$

has full rank.

In other words, the points $\Phi(x_1), \ldots, \Phi(x_m)$ are linearly independent (provided no two x_i are the same). They span an m-dimensional subspace of $\mathcal{H}$. Therefore a Gaussian kernel defined on a domain of infinite cardinality, with no a priori restriction on the number of training examples, produces a feature space of *infinite* dimension. Nevertheless, an analysis of the shape of the mapped data in feature space shows that capacity is distributed in a way that ensures smooth and simple estimates whenever possible (see Section 12.4).

Infinite-Dimensional Feature Space

The examples given above all apply to the case of vectorial data. Let us next give an example where $\mathcal{X}$ is *not* a vector space [42].

Proposition 2.19 (Similarity of Probabilistic Events) *If $(\mathcal{X}, \mathcal{C}, \mathrm{P})$ is a probability space with σ-algebra $\mathcal{C}$ and probability measure P, then*

$$k(A, B) = \mathrm{P}(A \cap B) - \mathrm{P}(A)\mathrm{P}(B) \tag{2.75}$$

is a positive definite kernel on $\mathcal{C} \times \mathcal{C}$.

Proof To see this, we define a feature map

$$\Phi : A \mapsto (I_A - \mathrm{P}(A)), \tag{2.76}$$

where I_A is the characteristic function on A. On the feature space, which consists of functions on $\mathcal{X}$ taking values in $[-1, 1]$, we use the dot product

$$\langle f, g \rangle := \int_{\mathcal{X}} f \cdot g \, d\mathrm{P}. \tag{2.77}$$

The result follows by noticing $\langle I_A, I_B \rangle = \mathrm{P}(A \cap B)$ and $\langle I_A, \mathrm{P}(B) \rangle = \mathrm{P}(A)\mathrm{P}(B)$.

■

Further examples include kernels for string matching, as proposed by [585, 234, 23]. We shall describe these, and address the general problem of designing kernel functions, in Chapter 13.

The next section will return to the connection between kernels and feature spaces. Readers who are eager to move on to SV algorithms may want to skip this section, which is somewhat more technical.

2.4 The Representation of Dissimilarities in Linear Spaces

2.4.1 Conditionally Positive Definite Kernels

We now proceed to a larger class of kernels than that of the positive definite ones. This larger class is interesting in several regards. First, it will turn out that some kernel algorithms work with this class, rather than only with positive definite kernels. Second, its relationship to positive definite kernels is a rather interesting one, and a number of connections between the two classes provide understanding of kernels in general. Third, they are intimately related to a question which is a variation on the central aspect of positive definite kernels: the latter can be thought of as dot products in feature spaces; the former, on the other hand, can be embedded as *distance measures* arising from norms in feature spaces.

The present section thus attempts to extend the utility of the kernel trick by looking at the problem of which kernels can be used to compute distances in feature spaces. The underlying mathematical results have been known for quite a while [465]; some of them have already attracted interest in the kernel methods community in various contexts [515, 234].

Clearly, the squared distance $\|\Phi(x) - \Phi(x')\|^2$ in the feature space associated with a pd kernel k can be computed, using $k(x, x') = \langle \Phi(x), \Phi(x') \rangle$, as

$$\|\Phi(x) - \Phi(x')\|^2 = k(x, x) + k(x', x') - 2k(x, x'). \tag{2.78}$$

Positive definite kernels are, however, not the full story: there exists a *larger* class of kernels that can be used as generalized distances, and the present section will describe why and how [468].

Let us start by considering how a dot product and the corresponding distance measure are affected by a translation of the data, $x \mapsto x - x_0$. Clearly, $\|x - x'\|^2$ is translation invariant while $\langle x, x' \rangle$ is not. A short calculation shows that the effect of the translation can be expressed in terms of $\|. - .\|^2$ as

$$\langle (x - x_0), (x' - x_0) \rangle = \frac{1}{2} \left(-\|x - x'\|^2 + \|x - x_0\|^2 + \|x_0 - x'\|^2 \right). \tag{2.79}$$

Note that this, just like $\langle x, x' \rangle$, is still a pd kernel: $\sum_{i,j} c_i c_j \langle (x_i - x_0), (x_j - x_0) \rangle = \| \sum_i c_i (x_i - x_0) \|^2 \geq 0$ holds true for any c_i. For any choice of $x_0 \in \mathcal{X}$, we thus get a similarity measure (2.79) associated with the dissimilarity measure $\|x - x'\|$.

This naturally leads to the question of whether (2.79) might suggest a connection