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

that also holds true in more general cases: what kind of nonlinear dissimilarity measure do we have to substitute for $\|. - .\|^2$ on the right hand side of (2.79), to ensure that the left hand side becomes positive definite? To state the answer, we first need to define the appropriate class of kernels.

The following definition differs from Definition 2.4 only in the additional constraint on the sum of the c_i. Below, $\mathbb{K}$ is a shorthand for $\mathbb{C}$ or $\mathbb{R}$; the definitions are the same in both cases.

Definition 2.20 (Conditionally Positive Definite Matrix) *A symmetric $m \times m$ matrix K ($m \geq 2$) taking values in $\mathbb{K}$ and satisfying*

$$\sum_{i,j=1}^{m} c_i \bar{c}_j K_{ij} \geq 0 \text{ for all } c_i \in \mathbb{K}, \text{ with } \sum_{i=1}^{m} c_i = 0, \tag{2.80}$$

is called conditionally positive definite (cpd).

Definition 2.21 (Conditionally Positive Definite Kernel) *Let $\mathcal{X}$ be a nonempty set. A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{K}$ which for all $m \geq 2, x_1, \ldots, x_m \in \mathcal{X}$ gives rise to a conditionally positive definite Gram matrix is called a* conditionally positive definite (cpd) kernel.

Note that symmetry is also required in the complex case. Due to the additional constraint on the coefficients c_i, it does not follow automatically anymore, as it did in the case of complex positive definite matrices and kernels. In Chapter 4, we will revisit cpd kernels. There, we will actually introduce cpd kernels of different orders. The definition given in the current chapter covers the case of kernels which are cpd *of order* 1.

Connection PD — CPD

Proposition 2.22 (Constructing PD Kernels from CPD Kernels [42]) *Let $x_0 \in \mathcal{X}$, and let k be a symmetric kernel on $\mathcal{X} \times \mathcal{X}$. Then*

$$\tilde{k}(x, x') := \frac{1}{2}(k(x, x') - k(x, x_0) - k(x_0, x') + k(x_0, x_0))$$

is positive definite if and only if k is conditionally positive definite.

The proof follows directly from the definitions and can be found in [42]. This result does generalize (2.79): the negative squared distance kernel is indeed cpd, since $\sum_i c_i = 0$ implies $-\sum_{i,j} c_i c_j \|x_i - x_j\|^2 = -\sum_i c_i \sum_j c_j \|x_j\|^2 - \sum_j c_j \sum_i c_i \|x_i\|^2 + 2\sum_{i,j} c_i c_j \langle x_i, x_j \rangle = 2\sum_{i,j} c_i c_j \langle x_i, x_j \rangle = 2\|\sum_i c_i x_i\|^2 \geq 0$. In fact, this implies that all kernels of the form

$$k(x, x') = -\|x - x'\|^\beta, 0 \leq \beta \leq 2 \tag{2.81}$$

are cpd (they are not pd),[10] by application of the following result (note that the case $\beta = 0$ is trivial):

10. Moreover, they are not cpd if $\beta > 2$ [42].

Proposition 2.23 (Fractional Powers and Logs of CPD Kernels [42]) *If $k : \mathcal{X} \times \mathcal{X} \to (-\infty, 0]$ is cpd, then so are $-(-k)^\alpha$ $(0 < \alpha < 1)$ and $-\ln(1-k)$.*

To state another class of cpd kernels that are not pd, note first that as a trivial consequence of Definition 2.20, we know that (i) sums of cpd kernels are cpd, and (ii) any constant $b \in \mathbb{R}$ is a cpd kernel. Therefore, any kernel of the form $k + b$, where k is cpd and $b \in \mathbb{R}$, is also cpd. In particular, since pd kernels are cpd, we can take any pd kernel and offset it by b, and it will still be at least cpd. For further examples of cpd kernels, cf. [42, 578, 205, 515].

2.4.2 Hilbert Space Representation of CPD Kernels

We now return to the main flow of the argument. Proposition 2.22 allows us to construct the feature map for k from that of the pd kernel $\tilde{k}$. To this end, fix $x_0 \in \mathcal{X}$ and define $\tilde{k}$ according to Proposition 2.22. Due to Proposition 2.22, $\tilde{k}$ is positive definite. Therefore, we may employ the Hilbert space representation $\Phi : \mathcal{X} \to \mathcal{H}$ of $\tilde{k}$ (cf. (2.32)), satisfying $\langle \Phi(x), \Phi(x') \rangle = \tilde{k}(x, x')$; hence,

$$\|\Phi(x) - \Phi(x')\|^2 = \tilde{k}(x, x) + \tilde{k}(x', x') - 2\tilde{k}(x, x'). \tag{2.82}$$

Substituting Proposition 2.22 yields

$$\|\Phi(x) - \Phi(x')\|^2 = -k(x, x') + \frac{1}{2}\left(k(x, x) + k(x', x')\right). \tag{2.83}$$

This implies the following result [465, 42].

Feature Map for CPD Kernels

Proposition 2.24 (Hilbert Space Representation of CPD Kernels) *Let k be a real-valued CPD kernel on $\mathcal{X}$, satisfying $k(x, x) = 0$ for all $x \in \mathcal{X}$. Then there exists a Hilbert space $\mathcal{H}$ of real-valued functions on $\mathcal{X}$, and a mapping $\Phi : \mathcal{X} \to \mathcal{H}$, such that*

$$\|\Phi(x) - \Phi(x')\|^2 = -k(x, x'). \tag{2.84}$$

If we drop the assumption $k(x, x) = 0$, the Hilbert space representation reads

$$\|\Phi(x) - \Phi(x')\|^2 = -k(x, x') + \frac{1}{2}\left(k(x, x) + k(x', x')\right). \tag{2.85}$$

It can be shown that if $k(x, x) = 0$ for all $x \in \mathcal{X}$, then

$$d(x, x') := \sqrt{-k(x, x')} = \|\Phi(x) - \Phi(x')\| \tag{2.86}$$

is a semi-metric: clearly, it is nonnegative and symmetric; additionally, it satisfies the triangle inequality, as can be seen by computing $d(x, x') + d(x', x'') = \|\Phi(x) - \Phi(x')\| + \|\Phi(x') - \Phi(x'')\| \geq \|\Phi(x) - \Phi(x'')\| = d(x, x'')$ [42].

It is a metric if $k(x, x') \neq 0$ for $x \neq x'$. We thus see that we can rightly think of k as the negative of a distance measure.

We next show how to represent *general* symmetric kernels (thus in particular cpd kernels) as symmetric bilinear forms Q in feature spaces. This generalization of the previously known feature space representation for pd kernels comes at a

cost: Q will no longer be a dot product. For our purposes, we can get away with this. The result will give us an intuitive understanding of Proposition 2.22: we can then write $\tilde{k}$ as $\tilde{k}(x, x') := Q(\Phi(x) - \Phi(x_0), \Phi(x') - \Phi(x_0))$. Proposition 2.22 thus essentially adds an origin in feature space which corresponds to the image $\Phi(x_0)$ of one point x_0 under the feature map.

Feature Map for General Symmetric Kernels

Proposition 2.25 (Vector Space Representation of Symmetric Kernels) *Let k be a real-valued symmetric kernel on $\mathcal{X}$. Then there exists a linear space $\mathcal{H}$ of real-valued functions on $\mathcal{X}$, endowed with a symmetric bilinear form $Q(.,.)$, and a mapping $\Phi : \mathcal{X} \to \mathcal{H}$, such that $k(x, x') = Q(\Phi(x), \Phi(x'))$.*

Proof The proof is a direct modification of the pd case. We use the map (2.21) and linearly complete the image as in (2.22). Define $Q(f, g) := \sum_{i=1}^{m} \sum_{j=1}^{m'} \alpha_i \beta_j k(x_i, x'_j)$. To see that it is well-defined, although it explicitly contains the expansion coefficients (which need not be unique), note that $Q(f, g) = \sum_{j=1}^{m'} \beta_j f(x'_j)$, independent of the α_i. Similarly, for g, note that $Q(f, g) = \sum_i \alpha_i g(x_i)$, hence it is independent of β_j. The last two equations also show that Q is bilinear; clearly, it is symmetric. ■

Note, moreover, that by definition of Q, k is a reproducing kernel for the feature space (which is not a Hilbert space): for all functions f (2.22), we have $Q(k(., x), f) = f(x)$; in particular, $Q(k(., x), k(., x')) = k(x, x')$.

Rewriting $\tilde{k}$ as $\tilde{k}(x, x') := Q(\Phi(x) - \Phi(x_0), \Phi(x') - \Phi(x_0))$ suggests an immediate generalization of Proposition 2.22: in practice, we might want to choose other points as origins in feature space — points that do not have a pre-image x_0 in the input domain, such as the mean of a set of points (cf. [543]). This will be useful when considering kernel PCA. It is only crucial that the behavior of our reference point under translation is identical to that of individual points. This is taken care of by the constraint on the sum of the c_i in the following proposition.

Matrix Centering

Proposition 2.26 (Exercise 2.23 in [42]) *Let K be a symmetric matrix, $\mathbf{e} \in \mathbb{R}^m$ be the vector of all ones, $\mathbf{1}$ the $m \times m$ identity matrix, and let $\mathbf{c} \in \mathbb{C}^m$ satisfy $\mathbf{e}^*\mathbf{c} = 1$. Then*

$$\tilde{K} := (\mathbf{1} - \mathbf{e}\mathbf{c}^*)K(\mathbf{1} - \mathbf{c}\mathbf{e}^*) \tag{2.87}$$

is positive definite if and only if K is conditionally positive definite.[11]

Proof "$\Longrightarrow$": suppose $\tilde{K}$ is positive definite. Thus for any $\mathbf{a} \in \mathbb{C}^m$ which satisfies $\mathbf{a}^*\mathbf{e} = \mathbf{e}^*\mathbf{a} = 0$, we have $0 \leq \mathbf{a}^*\tilde{K}\mathbf{a} = \mathbf{a}^*K\mathbf{a} + \mathbf{a}^*\mathbf{e}\mathbf{c}^*K\mathbf{c}\mathbf{e}^*\mathbf{a} - \mathbf{a}^*K\mathbf{c}\mathbf{e}^*\mathbf{a} - \mathbf{a}^*\mathbf{e}\mathbf{c}^*K\mathbf{a} = \mathbf{a}^*K\mathbf{a}$. This means that $0 \leq \mathbf{a}^*K\mathbf{a}$, proving that K is conditionally positive definite.

"$\Longleftarrow$": suppose K is conditionally positive definite. This means that we have to show that $\mathbf{a}^*\tilde{K}\mathbf{a} \geq 0$ for all $\mathbf{a} \in \mathbb{C}^m$. We have

$$\mathbf{a}^*\tilde{K}\mathbf{a} = \mathbf{a}^*(\mathbf{1} - \mathbf{e}\mathbf{c}^*)K(\mathbf{1} - \mathbf{c}\mathbf{e}^*)\mathbf{a} = \mathbf{s}^*K\mathbf{s} \text{ for } \mathbf{s} = (\mathbf{1} - \mathbf{c}\mathbf{e}^*)\mathbf{a}. \tag{2.88}$$

11. $\mathbf{c}^*$ is the vector obtained by transposing and taking the complex conjugate of $\mathbf{c}$.

All we need to show is $\mathbf{e}^*\mathbf{s} = 0$, since then we can use the fact that K is cpd to obtain $\mathbf{s}^* K\mathbf{s} \geq 0$. This can be seen as follows $\mathbf{e}^*\mathbf{s} = \mathbf{e}^*(\mathbf{1} - \mathbf{c}\mathbf{e}^*)\mathbf{a} = (\mathbf{e}^* - (\mathbf{e}^*\mathbf{c})\mathbf{e}^*)\mathbf{a} = (\mathbf{e}^* - \mathbf{e}^*)\mathbf{a} = 0$. ■

This result directly implies a corresponding generalization of Proposition 2.22:

Kernel Centering

Proposition 2.27 (Adding a General Origin) *Let k be a symmetric kernel, $x_1, \ldots, x_m \in \mathcal{X}$, and let $c_i \in \mathbb{C}$ satisfy $\sum_{i=1}^m c_i = 1$. Then*

$$\tilde{k}(x, x') := \frac{1}{2}\Big(k(x, x') - \sum_{i=1}^m c_i k(x, x_i) - \sum_{i=1}^m c_i k(x_i, x') + \sum_{i,j=1}^m c_i c_j k(x_i, x_j)\Big)$$

is positive definite if and only if k is conditionally positive definite.

Proof Consider a set of $m' \in \mathbb{N}$ points $x'_1, \ldots, x'_{m'} \in \mathcal{X}$, and let K be the $(m + m') \times (m + m')$ Gram matrix based on $x_1, \ldots, x_m, x'_1, \ldots, x'_{m'}$. Apply Proposition 2.26 using $c_{m+1} = \ldots = c_{m+m'} = 0$. ■

Application to SVMs

The above results show that conditionally positive definite kernels are a natural choice whenever we are dealing with a translation invariant problem, such as the SVM: maximization of the margin of separation between two classes of data is independent of the position of the origin. Seen in this light, it is not surprising that the structure of the dual optimization problem (cf. [561]) allows cpd kernels: as noted in [515, 507], the constraint $\sum_{i=1}^m \alpha_i y_i = 0$ projects out the same subspace as (2.80) in the definition of conditionally positive definite matrices.

Application to Kernel PCA

Another example of a kernel algorithm that works with conditionally positive definite kernels is Kernel PCA (Chapter 14), where the data are centered, thus removing the dependence on the origin in feature space. Formally, this follows from Proposition 2.26 for $c_i = 1/m$.

Application to Parzen Windows Classifiers

Let us consider another example. One of the simplest distance-based classification algorithms proceeds as follows. Given m_+ points labelled with $+1$, m_- points labelled with -1, and a mapped test point $\Phi(x)$, we compute the mean squared distances between the latter and the two classes, and assign it to the one for which this mean is smaller;

$$y = \operatorname{sgn}\left(\frac{1}{m_-}\sum_{y_i=-1}\|\Phi(x) - \Phi(x_i)\|^2 - \frac{1}{m_+}\sum_{y_i=1}\|\Phi(x) - \Phi(x_i)\|^2\right). \tag{2.89}$$

We use the distance kernel trick (Proposition 2.24) to express the decision function as a kernel expansion in the input domain: a short calculation shows that

$$y = \operatorname{sgn}\left(\frac{1}{m_+}\sum_{y_i=1}k(x, x_i) - \frac{1}{m_-}\sum_{y_i=-1}k(x, x_i) + b\right), \tag{2.90}$$

with the constant offset

$$b = \frac{1}{2m_-}\sum_{y_i=-1}k(x_i, x_i) - \frac{1}{2m_+}\sum_{y_i=1}k(x_i, x_i). \tag{2.91}$$

Note that for some cpd kernels, such as (2.81), $k(x_i, x_i)$ is always 0, and thus $b = 0$. For others, such as the commonly used Gaussian kernel, $k(x_i, x_i)$ is a nonzero constant, in which case b vanishes provided that $m_+ = m_-$. For normalized Gaussians, the resulting decision boundary can be interpreted as the Bayes decision based on two Parzen window density estimates of the classes; for general cpd kernels, the analogy is merely a formal one; that is, the decision functions take the same form.

Properties of CPD Kernels

Many properties of positive definite kernels carry over to the more general case of conditionally positive definite kernels, such as Proposition 13.1.

Using Proposition 2.22, one can prove an interesting connection between the two classes of kernels:

Proposition 2.28 (Connection PD — CPD [465]) *A kernel k is conditionally positive definite if and only if* $\exp(tk)$ *is positive definite for all $t > 0$.*

Positive definite kernels of the form $\exp(tk)$ ($t > 0$) have the interesting property that their nth root ($n \in \mathbb{N}$) is again a positive definite kernel. Such kernels are called *infinitely divisible*. One can show that, disregarding some technicalities, the logarithm of an infinitely divisible positive definite kernel mapping into $\mathbb{R}_0^+$ is a conditionally positive definite kernel.

2.4.3 Higher Order CPD Kernels

For the sake of completeness, we now present some material which is of interest to one section later in the book (Section 4.8), but not central for the present chapter. We follow [341, 204].

Definition 2.29 (Conditionally Positive Definite Functions of Order q) *A continuous function h, defined on* $[0, \infty)$, *is called conditionally positive definite (cpd) of order q on $\mathbb{R}^N$ if for any distinct points $x_1, \ldots, x_m \in \mathbb{R}^N$, the quadratic form,*

$$\sum_{i,j=1}^{m} \alpha_i \alpha_j h(\|x_i - x_j\|^2), \tag{2.92}$$

is nonnegative, provided that the scalars $\alpha_1, \ldots, \alpha_m$ satisfy $\sum_{i=1}^m \alpha_i p(x_i) = 0$, for all polynomials $p(\cdot)$ on $\mathbb{R}^N$ of degree lower than q.

Let Π_q^N denote the space of polynomials of degree lower than q on $\mathbb{R}^N$. By definition, every cpd function h of order q generates a positive definite kernel for SV expansions in the space of functions orthogonal to Π_q^N, by setting $k(x, x') := h(\|x - x'\|^2)$.

There exists also an analogue to the positive definiteness of the integral operator in the conditions of Mercer's theorem. In [157, 341] it is shown that for cpd functions h of order q, we have

$$\int h(\|x - x'\|^2) f(x) f(x') dx dx' \geq 0, \tag{2.93}$$

provided that the projection of f onto Π_q^N is zero.

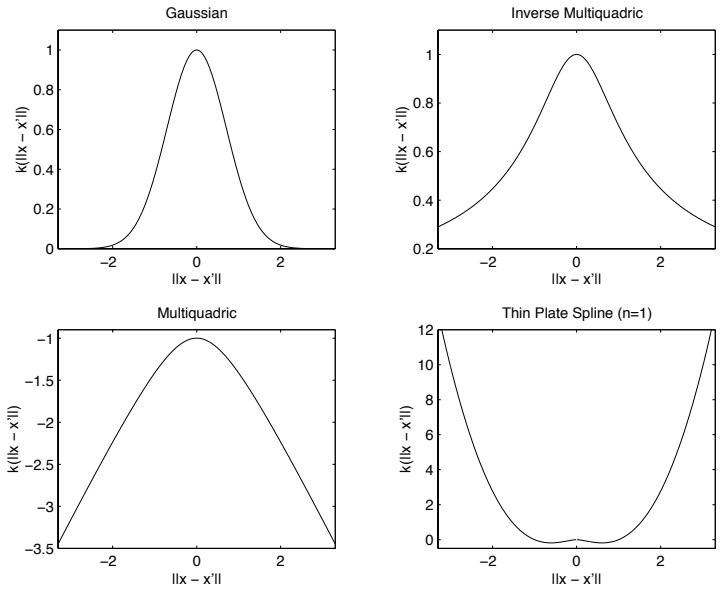

Figure 2.4 Conditionally positive definite functions, as described in Table 2.1. Where applicable, we set the free parameter c to 1; β is set to 2. Note that cpd kernels need not be positive anywhere (e.g., the Multiquadric kernel).

Table 2.1 Examples of Conditionally Positive Definite Kernels. The fact that the exponential kernel is pd (i.e., cpd of order 0) follows from (2.81) and Proposition 2.28.

Kernel	Order	
$e^{-c\|x-x'\|^\beta}$, $0 \leq \beta \leq 2$	0	Exponential
$\frac{1}{\sqrt{\|x-x'\|^2+c^2}}$	0	Inverse Multiquadric
$-\sqrt{\|x-x'\|^2+c^2}$	1	Multiquadric
$\|x-x'\|^{2n} \ln \|x-x'\|$	n	Thin Plate Spline

Definition 2.30 (Completely Monotonic Functions) *A function $h(x)$ is called completely monotonic of order q if*

$$(-1)^n \frac{d^n}{dx^n} h(x) \geq 0 \text{ for all } x \in [0, \infty) \text{ and } n \geq q. \tag{2.94}$$

It can be shown [464, 465, 360] that a function $h(x^2)$ is conditionally positive definite if and only if $h(x)$ is completely monotonic of the same order. This gives a (sometimes simpler) criterion for checking whether a function is cpd or not.

If we use cpd kernels in learning algorithms, we must ensure orthogonality of the estimate with respect to Π_q^N. This is usually done via constraints $\sum_{i=1}^m \alpha_i p(x_i) = 0$ for all polynomials $p(\cdot)$ on $\mathbb{R}^N$ of degree lower than q (see Section 4.8).

2.5 Summary

The crucial ingredient of SVMs and other kernel methods is the so-called kernel trick (see (2.7) and Remark 2.8), which permits the computation of dot products in high-dimensional feature spaces, using simple functions defined on pairs of input patterns. This trick allows the formulation of nonlinear variants of any algorithm that can be cast in terms of dot products, SVMs being but the most prominent example. The mathematical result underlying the kernel trick is almost a century old [359]. Nevertheless, it was only much later that it was exploited by the machine learning community for the analysis [4] and construction of algorithms [62], and that it was described as a general method for constructing nonlinear generalizations of dot product algorithms [480].

The present chapter has reviewed the mathematical theory of kernels. We started with the class of polynomial kernels, which can be motivated as computing a combinatorially large number of monomial features rather efficiently. This led to the general question of which kernel can be used, or: which kernel can be represented as a dot product in a linear feature space. We defined this class and discussed some of its properties. We described several ways how, given such a kernel, one can construct a representation in a feature space. The most well-known representation employs Mercer's theorem, and represents the feature space as an ℓ_2 space defined in terms of the eigenfunctions of an integral operator associated with the kernel. An alternative representation uses elements of the theory of reproducing kernel Hilbert spaces, and yields additional insights, representing the linear space as a space of functions written as kernel expansions. We gave an in-depth discussion of the kernel trick in its general form, including the case where we are interested in dissimilarities rather than similarities; that is, when we want to come up with nonlinear generalizations of distance-based algorithms rather than dot-product-based algorithms.

In both cases, the underlying philosophy is the same: we are trying to express a complex nonlinear algorithm in terms of simple geometrical concepts, and we are then dealing with it in a linear space. This linear space may not always be readily available; in some cases, it may even be hard to construct explicitly. Nevertheless, for the sake of design and analysis of the algorithms, it is sufficient to know that the linear space exists, empowering us to use the full potential of geometry, linear algebra and functional analysis.

2.6 Problems

2.1 (Monomial Features in $\mathbb{R}^2$ •) *Verify the second equality in (2.9).*

2.2 (Multiplicity of Monomial Features in $\mathbb{R}^N$ [515] ••) *Consider the monomial kernel $k(x, x') = \langle x, x' \rangle^d$ (where $x, x' \in \mathbb{R}^N$), generating monomial features of order d. Prove*