An Introduction to Kernel Methods[1]

Yuri Kalnishkan

Technical Report

CLRC — TR — 09 — 01

May 2009

Royal Holloway
University of London

Department of Computer Science
Egham, Surrey TW20 0EX, England

[1] This paper has been written for wiki project onlineprediction.net; see http://onlineprediction.net/?n=Main.KernelMethods for the up-to-date version. The author has been supported by the EPSRC grant EP/F002998 'Practical competitive prediction'.

Abstract

Kernel methods are a powerful tool of modern learning. This article provides an introduction to kernel methods through a motivating example of kernel ridge regression, defines reproducing kernel Hilbert spaces (RKHS), and then sketches a proof of the fundamental existence theorem.

Some results that appear to be important in the context of learning are also discussed.

Contents

1 A Motivating Example: Kernel Ridge Regression

In this section we will introduce kernels in the context of ridge regression. The reader may skip this section and proceed straight to the next session if he is only interested in the formal theory of RKHSs.

A more detailed discussion of Ridge Regression and kernels can be found in Section 3 of Steve Busuttil's dissertation.

1.1 The Problem

Suppose we are given a set of T examples $(x_1, y_1), (x_2, y_2), \ldots, (x_T, y_T)$, where $x_i \in \mathbb{R}^n$ are *signals* and $y_i \in \mathbb{R}$ are *outcomes* or *labels*. We want to find a dependency between signals and outcomes and to be able to predict y given a new x. This problem is often referred to as the *regression problem*[1].

Let us start by restricting ourselves to linear dependencies of the form $y = \langle w, x \rangle = w'x$, where $w \in \mathbb{R}^n$, $\langle \cdot, \cdot \rangle$ is the standard scalar product in $\mathbb{R}^n$, and the prime stands for transposition (by default all vectors are assumed to be column vectors). The class of linear functions is not too reach, and we will need to progress to more sophisticated classes later.

1.2 Least Squares and Ridge Regression

The *least squares* is a natural, popular, and time-honoured (apparently going back to Legendre and Gauss) approach to finding w. Let us take an w minimising the sum of squared discrepancies

$$L_{\mathrm{SQ}}(w) = \sum_{i=1}^{T} (w'x_i - y_i)^2 .$$

It is easy to find such a w; it can be interpreted as a projection of the vector $Y = (y_1, y_2, \ldots, y_T)$ on the subspace of $\mathbb{R}^T$ generated by n vectors $v_1, v_2, \ldots, v_T$, where v_i consists of i-th coordinates of xs.

We will derive the exact formula, but for a more general problem. Let us take $a \geq 0$ and consider the expression

$$L_{\mathrm{RR}}(w) = a\|w\|^2 + \sum_{i=1}^{T} (w'x_i - y_i)^2 .$$

A vector w minimising this is called a solution of the *ridge regression* problem (for reasons that will become apparent later). The least squares approach is a special case of the ridge regression approach, namely, that of $a = 0$.

Why would anyone want to use $a > 0$? There are two main reasons. First, the term $a\|w\|^2$ performs the regularisation function. It penalises the growth of coefficients of w and urges us to look for 'simpler' solutions. Secondly, it makes the problem easier computationally, as will be shown below.

1.3 Solution: Primary Form

Let us find the solution of the ridge regression problem with $a > 0$. It is convenient to introduce a $T \times n$ matrix $X = (x_1, x_2, \ldots, x_t)'$; the rows of X are vectors x_i (and the columns are vectors v_i mentioned above). We get

$$L_{\mathrm{RR}}(w) = a\|w\|^2 + \|Xw - Y\|^2.$$

[1] As different, for example, to the classification problem, where ys belong to a finite set.

By differentiating $L_{\mathrm{RR}}(w)$ w.r.t. w and equating the result to 0 we get

$$2aw - 2X'Y + X'Xw = 0$$

and

$$w = (aI + X'X)^{-1}X'Y,$$

where I is the identity matrix. This must be a solution; indeed, as coefficients of w approach infinity, $a\|w\|^2$ and therefore $L_{\mathrm{RR}}(w)$ must go to infinity.

Let us analyse this expression. The matrices $X'X$, I, and $aI + X'X$ have the size $n \times n$. The matrix $X'X$ is positive semi-definite, i.e., $\xi'(X'X)\xi \geq 0$ for all $\xi \in \mathbb{R}^n$ (this follows from $\xi'(X'X)\xi = (X\xi)'(X\xi) = \|X\xi\|^2$). By adding aI we make the matrix positive definite, i.e., we have $\xi'(aI + X'X)\xi > 0$ for all $\xi \neq 0$ (indeed, $\xi'(aI + X'X)\xi = \|X\xi\|^2 + a\|\xi\|^2$). Because every positive definite matrix is non-singular [2], $aI + X'X$ must have the inverse. If $a > 0$, a solution to the ridge regression problem always exists and it is unique.

If $a = 0$, the matrix may become singular. In fact, this will always happen in the case $T < n$. The singularity simply means that the solution to least squares is not unique. The ridge regression is thus theoretically simple and we will concentrate on it below.

As a approach 0, the matrix $aI + X'X$ may become close to singular. The numerical routines for finding w will then become less and less stable: they will have to deal with very big or very small values and make large round-up errors. Taking a larger $a > 0$ thus stabilises the computation as mentioned earlier.

Let us attempt a rough hypothetical analysis of the predictive performance of ridge regression for different values of a. If a is very big, the term aI completely overshadows $X'X$ and the predictive performance deteriorates. If a is very small, we may encounter numerical problems. An optimal value should thus be neither too big no too small. In some sense it must be comparable in size to elements of $X'X$. The exact choice of a depends on the particular dataset.

Finally, let us go back to the term 'ridge regression'. One of the versions of its etymology is that the diagonal of aI forms a 'ridge' added on top of the least squares matrix $X'X$.

1.4 Solution: Dual Form

Using the matrix identity $A(aI + A'A)^{-1} = (aI + AA')^{-1}A$ we can rewrite the ridge regression solution as follows. For an arbitrary $x \in \mathbb{R}^n$ the outcome suggested by ridge regression is $w'x$ and this can be rewritten as

$$\begin{aligned} w'x &= ((aI + X'X)^{-1}X'Y)'x, \\ &= Y'X(aI + X'X)^{-1}x, \\ &= Y'(aI + XX')^{-1}Xx. \end{aligned}$$

This formula is called the dual form of the ridge regression solution.

Similar arguments concerning non-singularity apply to $aI + XX'$. The matrix has the size $T \times T$. This might seem a disadvantage compared to the primary form: it is natural to expect that in practice n would be fixed and not too big, while the size of the sample T may be quite large. However this formula allows us to develop important generalisations.

[2] Indeed, let A be positive definite. If A is singular, $Av = 0$ for some $v \neq 0$, but this implies $v'Av = 0$.

We can say that $w = Y'(aI + XX')^{-1}X$ in the dual form. However there is a more interesting way to interpret the dual form formula. We have

$$w'x = Y'(aI + K)^{-1}k,$$

where K is the matrix of mutual scalar products

$$K = \begin{pmatrix} \langle x_1, x_1 \rangle & \langle x_1, x_2 \rangle & \dots & \langle x_1, x_T \rangle \\ \langle x_2, x_1 \rangle & \langle x_2, x_2 \rangle & \dots & \langle x_2, x_T \rangle \\ \vdots & \vdots & \ddots & \vdots \\ \langle x_T, x_1 \rangle & \langle x_T, x_2 \rangle & \dots & \langle x_T, x_T \rangle \end{pmatrix}$$

and $k = k(x)$ is the vector of scalar products of x_i by x:

$$k = \begin{pmatrix} \langle x_1, x \rangle \\ \langle x_2, x \rangle \\ \dots \\ \langle x_T, x \rangle \end{pmatrix}.$$

Note that all x_is and x appear in this formula only in mutual scalar product. This observation has important consequences.

1.5 Non-linear Regression

Now let us try to extend the class of functions we use and consider a wider class. Suppose that $n = 1$, i.e., all xs are numbers and we are interested in approximations by polynomials of degree 3, i.e., functions of the form $w_0 + w_1 x + w_2 x^2 + w_3 x^3$. Of course we can write down $L(w)$ for this case, perform the differentiation and find the solution as we did before. However there is a simpler argument based on the dual form.

Let us map x into $\mathbb{R}^4$ as follows: $x \to (1, x, x^2, x^3)$. Once we have done this, we can do linear regression on new 'long' signals. If we use the dual form, we do not even have to perform the transformations explicitly. Because we only need scalar products, we can compute all the necessary products $1 + x_1 x_2 + x_1^2 x_2^2 + x_1^3 x_2^3$ and substitute them into the dual form formula.

Let us write down a formal generalisation. The signals x do not have to come from $\mathbb{R}^n$ any longer. Let them be drawn from some arbitrary set[3] X. Suppose that we have a mapping $\Phi : X \to \mathcal{S}$, where $\mathcal{S}$ is some vector space equipped with a scalar product $\langle \cdot, \cdot \rangle$ (dot-product space); the space $\mathcal{S}$ is sometimes referred to as the *feature space*. We can use ridge regression in the feature space. The prediction of ridge regression on a signal x can be written as

$$\gamma_{\mathrm{RR}} = Y'(aI + K)^{-1}k,$$

where

$$K = \begin{pmatrix} \mathcal{K}(x_1, x_1) & \mathcal{K}(x_1, x_2) & \dots & \mathcal{K}(x_1, x_T) \\ \mathcal{K}(x_2, x_1) & \mathcal{K}(x_2, x_2) & \dots & \mathcal{K}(x_2, x_T) \\ \vdots & \vdots & \ddots & \vdots \\ \mathcal{K}(x_T, x_1) & \mathcal{K}(x_T, x_2) & \dots & \mathcal{K}(x_T, x_T) \end{pmatrix}$$

and

$$k = \begin{pmatrix} \mathcal{K}(x_1, x) \\ \mathcal{K}(x_2, x) \\ \dots \\ \mathcal{K}(x_T, x) \end{pmatrix};$$

[3] It is important that no particular structure is postulated on X; throughout the most of this article it is just a set.

the function $\mathcal{K} : X^2 \to \mathbb{R}$ is given by $\mathcal{K}(x_1, x_2) = \langle \Phi(x_1), \Phi(x_2) \rangle$.

The space $\mathcal{S}$ does not have to be finite-dimensional. However since every vector space with a scalar product can be embedded into a Hilbert space (see below for a definition) we can assume that it is Hilbert.

The transformation Φ is of no particular importance to us. Once we know the $\mathcal{K}$, we can perform regression with it.

1.6 Mappings and Kernels

It would be nice to have a characterisation of all $\mathcal{K}$ without a reference to Φ. A characterisation of this kind can be given.

It is easy to see that $\mathcal{K}(x_1, x_2) = \langle \Phi(x_1), \Phi(x_2) \rangle$ has the following properties:

- it is *symmetric*: $\mathcal{K}(x_1, x_2) = \mathcal{K}(x_2, x_1)$ for all $x_1, x_2 \in X$; this follows from the symmetry of the scalar product $\langle \cdot, \cdot \rangle$;
- it is *positive semi-definite*: for every positive integer T and every $x_1, x_2, \ldots, x_T \in X$ the matrix

$$K = \begin{pmatrix} \mathcal{K}(x_1, x_1) & \mathcal{K}(x_1, x_2) & \ldots & \mathcal{K}(x_1, x_T) \\ \mathcal{K}(x_2, x_1) & \mathcal{K}(x_2, x_2) & \ldots & \mathcal{K}(x_2, x_T) \\ \vdots & \vdots & \ddots & \vdots \\ \mathcal{K}(x_T, x_1) & \mathcal{K}(x_T, x_2) & \ldots & \mathcal{K}(x_T, x_T) \end{pmatrix}$$

  is positive semi-definite[4]; indeed, K is the Gram matrix of the images $\Phi(x_1), \Phi(x_2), \ldots, \Phi(x_T)$[5].

Surprisingly these two simple properties are sufficient. Let us call a function $\mathcal{K} : X^2 \to \mathbb{R}$ satisfying these two properties a *kernel*. Then the following theorem can be formulated.

Theorem 1. *For any set X a function $\mathcal{K} : X^2 \to \mathbb{R}$ is a kernel, i.e., it is symmetric and positive semi-definite, if and only if there is a mapping Φ from X into a Hilbert space $\mathcal{H}$ with a scalar product $\langle \cdot, \cdot \rangle$ such that $\mathcal{K}(x_1, x_2) = \langle \Phi(x_1), \Phi(x_2) \rangle$ for all $x_1, x_2 \in X$.*

We proved the 'if' part when we defined kernels. The 'only if' part follows from the results of the next sections, where we will show that the class of kernels coincides with the class of so called reproducing kernels.

2 Reproducing Kernel Hilbert Spaces

In this section we introduce reproducing kernel Hilbert spaces (RKHS) and show some of their basic properties. The presentation is based mainly on [Aro43] and [Aro50] and the reader may consult these papers for more details; note that the former paper is in French.

2.1 Hilbert Space

A set of some elements H is a *Hilbert space* if

1. H is a vector space over $\mathbb{R}$ (Hilbert spaces over the complex plain $\mathbb{C}$ can also be considered, but we shall restrict ourselves to $\mathbb{R}$ in this article);

[4] A definition and a discussion were given above.

[5] We have $\xi' K \xi = \| \sum_{i=1}^{T} \xi_i \Phi(x_i) \|^2 \geq 0$.

2. H is equipped with a scalar product $\langle \cdot, \cdot \rangle$ (i.e., with a symmetric positive definite bilinear form);

3. H is complete w.r.t. the metric generated by the scalar product, i.e., every fundamental sequence of elements of H converges.

Some authors require a Hilbert space to be separable, t.e., to have a countable dense subset. For example, [Aro43] reserves the name ‘Hilbert’ for separable spaces and calls general Hilbert spaces ‘generalised Euclidean’. We shall not impose this requirement by default.

As a matter of fact all separable Hilbert spaces are isomorphic (the situation is similar to that with finite-dimensional spaces; the separable Hilbert space is ‘countable-dimensional’).

Typical (though not particularly relevant to this article) examples of Hilbert spaces are provided by $L_2(X, \mu)$, which is the space of all real-valued functions f on X such that f^2 is Lebesgue-integrable w.r.t. the measure μ on X with the scalar product $\langle f, g \rangle = \int_X fgd\mu$, and l_2, which is the set of infinite sequences $(x_1, x_2, \ldots)$, $x_i \in \mathbb{R}$, such that the sum $\sum_{i=1}^{+\infty} x_i^2$ converges. Both l_2 and L_2 on $[0, 1]$ with the standard Lebesgue measure are separable; therefore they are isomorphic.

2.2 Reproducing Kernel Hilbert Spaces: a Definition

Let $\mathcal{F}$ be a Hilbert space consisting of functions on a set X. A function $\mathcal{K}(x_1, x_2)$ is a *reproducing kernel* (r.k.) for $\mathcal{F}$ if

◇ for every $x \in X$ the function $\mathcal{K}(x, \cdot)$ (i.e., $\mathcal{K}(x, x_2)$ as the function of the second argument with x fixed) belongs to $\mathcal{F}$

◇ the *reproducing property* holds: for every $f \in \mathcal{F}$ and every $x \in X$ we have $f(x) = \langle f, \mathcal{K}(x, \cdot) \rangle$.

A space $\mathcal{F}$ admitting a reproducing kernel is called a *reproducing kernel Hilbert space (RKHS).*

2.3 Reproducing Kernel Hilbert Spaces: Some Properties

Let us formulate and prove some basic properties of reproducing kernels.

Theorem 2. *1. If a r.k. for $\mathcal{F}$ exists, it is unique.*

2. If $\mathcal{K}$ is a reproducing kernel for $\mathcal{F}$, then for all $x \in X$ and $f \in \mathcal{F}$ we have $|f(x)| \leq \sqrt{\mathcal{K}(x, x)} \|f\|_{\mathcal{F}}$.

3. If $\mathcal{F}$ is a RKHS, then convergence in $\mathcal{F}$ implies pointwise convergence of corresponding functions.

Proof. In order to prove (1) suppose that there are two r.k. $\mathcal{K}_1$ and $\mathcal{K}_2$ for the same space $\mathcal{F}$. For every $x \in X$ the function $\mathcal{K}_1(x, \cdot) - \mathcal{K}_2(x, \cdot)$ belongs to $\mathcal{F}$ and, applying linearity and the reproducing property, we get

$$\begin{aligned}
\|\mathcal{K}_1(x, \cdot) - \mathcal{K}_2(x, \cdot)\|_{\mathcal{F}}^2 &= \langle \mathcal{K}_1(x, \cdot) - \mathcal{K}_2(x, \cdot), \mathcal{K}_1(x, \cdot) - \mathcal{K}_2(x, \cdot) \rangle \\
&= \langle \mathcal{K}_1(x, \cdot) - \mathcal{K}_2(x, \cdot), \mathcal{K}_1(x, \cdot) \rangle - \\
&\qquad \langle \mathcal{K}_1(x, \cdot) - \mathcal{K}_2(x, \cdot), \mathcal{K}_2(x, \cdot) \rangle \\
&= (\mathcal{K}_1(x, x) - \mathcal{K}_2(x, x)) - (\mathcal{K}_1(x, x) - \mathcal{K}_2(x, x)) \\
&= 0.
\end{aligned}$$

The definition of a Hilbert space implies that $\mathcal{K}_1(x, \cdot)$ coincides with $\mathcal{K}_2(x, \cdot)$ and therefore they are equal everywhere as functions.

Property (2) follows immediately from the reproducing property and the Cauchy (-Schwarz-Bunyakovsky) inequality.

Property (3) follows from (2). Indeed, for all $f_1, f_2 \in \mathcal{F}$ and $x \in X$ we have

$$|f_1(x) - f_2(x)| \leq \sqrt{\mathcal{K}(x,x)}\|f_1 - f_2\|_{\mathcal{F}}.$$

□

We shall now give an important 'internal' characterisation of reproducing kernel Hilbert spaces.

Let $\mathcal{F}$ consisting of real-valued functions on X be a Hilbert space. Take $x \in X$ and consider the functional $\mathcal{F} \to \mathbb{R}$ mapping $f \in \mathcal{F}$ into $f(x)$. It is linear (in f) and is called the *evaluation functional*.

Note that the evaluation functional is not defined on L_2: the elements of L_2 are in fact equivalence classes of functions that coincide everywhere up to a set of measure 0, and thus they are not really defined at every point.

Theorem 3. *A Hilbert space $\mathcal{F}$ consisting of real-valued functions on X is a RKHS if and only if for every $x \in X$ the corresponding evaluation functional is continuous.*

Proof. The 'only if' part follows from (2) from the previous theorem.

In order to prove the 'if' part we need the Riess-Fischer Representation Theorem, which states that every continuous linear functional on a Hilbert space can be represented as the scalar product by some element of the space.

Take $x \in X$. Because the evaluation functional is continuous, there is a unique $k_x \in \mathcal{F}$ such that $f(x) = \langle f, k_x \rangle$. We can define a mapping $F : X \to \mathcal{F}$ by $F(x) = k_x$. Let $\mathcal{K}(x_1, x_2) = \langle F(x_1), F(x_2) \rangle$.

We have

$$\begin{aligned}\mathcal{K}(x_1, x_2) &= \langle k_{x_1}, k_{x_2} \rangle \\ &= k_{x_1}(x_2)\end{aligned}$$

and thus $\mathcal{K}(x_1, \cdot) = k_{x_1}(\cdot) \in \mathcal{F}$. On the other hand for every $f \in \mathcal{F}$ and $x \in X$ we have

$$\begin{aligned}f(x) &= \langle f, k_x \rangle \\ &= \langle f, \mathcal{K}(x, \cdot) \rangle.\end{aligned}$$

Therefore $\mathcal{K}$ is a r.k. for $\mathcal{F}$. □

This criterion is quite important. The continuity of the evaluation functional means that it is consistent with the norm: functions f_1 and f_2 that are close with respect to the norm evaluate to values $f_1(x)$ and $f_2(x)$ that are close. If we consider functions from some space as hypotheses in machine learning and the norm on the space as a measure of complexity, it is natural to require the continuity of the evaluation functional. The theorem shows that all 'natural' Hilbert spaces of functions are in fact reproducing kernel Hilbert spaces.

2.4 Existence Theorem

We have shown that a r.k. $\mathcal{K}(x_1, x_2)$ can be represented as $\langle F(x_1), F(x_2) \rangle$. This implies that $\mathcal{K}$ is

- ⋄ symmetric due to the symmetry of the scalar product;

- ◇ positive semi-definite, i.e., for all $x_1, x_2, \ldots, x_T \in X$ the matrix

$$K = \begin{pmatrix} \mathcal{K}(x_1, x_1) & \mathcal{K}(x_1, x_2) & \ldots & \mathcal{K}(x_1, x_T) \\ \mathcal{K}(x_2, x_1) & \mathcal{K}(x_2, x_2) & \ldots & \mathcal{K}(x_2, x_T) \\ \vdots & \vdots & \ddots & \vdots \\ \mathcal{K}(x_T, x_1) & \mathcal{K}(x_T, x_2) & \ldots & \mathcal{K}(x_T, x_T) \end{pmatrix}$$

  is positive semi-definite; this holds since K is the Gram matrix.

Thus $\mathcal{K}$ is a kernel according to the definition from the previous section. The following theorem shows that the classes of kernels and reproducing kernels coincide.

Theorem 4. *Let $\mathcal{K} : X^2 \to \mathbb{R}$ is a real-valued function of two arguments on X. Then $\mathcal{K}$ is a reproducing kernel for some Hilbert space of functions $\mathcal{F}$ on X if and only if*

- ◇ *$\mathcal{K}$ is symmetric*
- ◇ *$\mathcal{K}$ is positive semi-definite.*

If there is a space admitting $\mathcal{K}$ as its reproducing kernel, it is unique.

3 Proof of the Existence Theorem

In this section we will prove the existence theorem. Let $\mathcal{K} : X^2 \to \mathbb{R}$ be a kernel.

3.1 Linear Combinations: A Dot Product Space

We start the proof by constructing a linear space of functions $\mathcal{F}_1$ consisting of linear combinations $\sum_{i=1}^n a_i \mathcal{K}(x_i, \cdot)$, where n is a positive integer, $a_i \in \mathbb{R}$ and $x_i \in X$. The linearity follows by construction.

The scalar product is defined after the following fashion. Let

$$\left\langle \sum_{i=1}^n a_i \mathcal{K}(x_i, \cdot), \sum_{i=1}^n b_i \mathcal{K}(x_i, \cdot) \right\rangle = \sum_{i,j=1}^n a_i b_j \mathcal{K}(x_i, x_j)$$

(by adding terms with zero coefficients we can ensure that the linear combinations have equal numbers of terms and that all x_i in the combinations are the same). We need to prove that the scalar product is well-defined, i.e., to show that it is independent of particular representations of factors (recall that we are constructing a space of functions rather than formal linear combinations).

Let $f(x) = \sum_{i=1}^n a_i \mathcal{K}(x_i, x)$ and $g(x) = \sum_{i=1}^n b_i \mathcal{K}(x_i, x)$. We have

$$\begin{aligned} \langle f, g \rangle &= \sum_{i,j=1}^n a_i b_j \mathcal{K}(x_i, x_j) \\ &= \sum_{i=1}^n a_i \left(\sum_{j=1}^n b_j \mathcal{K}(x_i, x_j) \right) \\ &= \sum_{i=1}^n a_i g(x_j). \end{aligned}$$

We see that the scalar product can be expressed in terms of values of g and thus is independent of a particular representation of g as a linear combination. A similar argument can be applied to f. The independence follows.

The function $\langle\cdot,\cdot\rangle$ is symmetric because $\mathcal{K}$ is symmetric. For f from above we have

$$\langle f, f\rangle = \sum_{i,j=1}^{n} a_i a_j \mathcal{K}(x_i, x_j) \geq 0$$

because $\mathcal{K}$ is positive semi-definite. Therefore $\langle\cdot,\cdot\rangle$ is positive semi-definite. We have shown that it is a positive semi-definite symmetric bilinear form. One final step is necessary to prove that it is positive definite and therefore a scalar product.

Let us evaluate $\langle f(\cdot), \mathcal{K}(x,\cdot)\rangle$, where $f \in \mathcal{F}_1$ and x is some element from X. We get

$$\langle f(\cdot), \mathcal{K}(x,\cdot)\rangle = f(x).$$

The form $\langle\cdot,\cdot\rangle$ and $\mathcal{K}$ thus satisfy the reproducing property.

Because the form $\langle\cdot,\cdot\rangle$ is positive semi-definite, the Cauchy inequality holds for it and

$$\langle f, g\rangle \leq \|f\| \cdot \|g\|,$$

where $\|f\|$ is defined as $\sqrt{\langle f, f\rangle}$. Combining this with the reproducing property yields

$$\begin{aligned}\langle f(\cdot), \mathcal{K}(x,\cdot)\rangle &\leq \|f\| \cdot \|\mathcal{K}(x,\cdot)\| \\ &= \|f\|\sqrt{\mathcal{K}(x,x)}.\end{aligned}$$

Therefore $\|f\| = 0$ implies that $f(x) = 0$ for an arbitrary $x \in X$. We have thus shown that $\langle\cdot,\cdot\rangle$ is actually positive definite and therefore a scalar product.

The construction is not finished yet because $\mathcal{F}_1$ is not necessarily complete. It remains to construct a completion of $\mathcal{F}_1$. It is well known that every linear space with a scalar product has a completion, which is a Hilbert space. However this argument cannot be applied here[6]: we need a completion of a specific form, namely, consisting of functions $X \to \mathbb{R}$. Note that, however, we have already proved Theorem 1 from Section 1: we can map X into *some* Hilbert space H so that the value of the kernel is given by the scalar product of images. The mapping $\Phi : X \to \mathcal{F}_1$ is given by the obvious $\Phi(x) = \mathcal{K}(x,\cdot)$.

3.2 Completion

In this subsection we will construct a completion of $\mathcal{F}_1$.

Let $f_1, f_2, \ldots \in \mathcal{F}_1$ be a fundamental sequence. For every $x \in X$ the inequalities

$$\begin{aligned}|f_n(x) - f_m(x)| &= |\langle f_n - f_m, \mathcal{K}(x,\cdot)\rangle| \\ &\leq \|f_n - f_m\|\sqrt{\mathcal{K}(x,x)},\end{aligned}$$

which follow from the previous subsection, imply that the sequence $f_1(x), f_2(x), \ldots$ is fundamental and therefore has a limit. We can define a function $f : X \to \mathbb{R}$ by $f(x) = \lim_{n\to\infty} f_n(x)$.

Let $\mathcal{F}$ consist of all functions thus obtained. Clearly, $\mathcal{F}_1 \subseteq \mathcal{F}$ since each f from $\mathcal{F}_1$ is the pointwise limit of the sequence $f, f, \ldots$.

The scalar product on $\mathcal{F}$ can be introduced as follows. If f is the pointwise limit of $f_1, f_2, \ldots \in \mathcal{F}_1$ and g is the pointwise limit of $g_1, g_2, \ldots \in \mathcal{F}_1$, then $\langle f, g\rangle_{\mathcal{F}} = \lim_{n\to\infty}\langle f_n, g_n\rangle_{\mathcal{F}_1}$.

[6] The book [SS02] uses this argument in Section 2.2.3, p. 35-36, in a rather misleading way.

Let us show that this limit exists. For all positive integers n_1, n_2, m_1 and m_2 we have

$$\begin{aligned}&|\langle f_{n_1}, g_{m_1}\rangle - \langle f_{n_2}, g_{m_2}\rangle| \leq \\ &\qquad |\langle f_{n_1}, g_{m_1}\rangle - \langle f_{n_1}, g_{m_2}\rangle| + |\langle f_{n_1}, g_{m_2}\rangle - \langle f_{n_2}, g_{m_2}\rangle| \leq \\ &\qquad\qquad \|f_{n_1}\| \cdot \|g_{m_1} - g_{m_2}\| + \|g_{m_2}\| \cdot \|f_{n_1} - f_{n_2}\|.\end{aligned}$$

Because the norms of elements of a fundamental sequence are uniformly bounded, the difference can be made as close to 0 as necessary for sufficiently large n_1, n_2, m_1 and m_2. Thus there even a double limit $\lim_{n,m\to\infty}\langle f_n, g_m\rangle = s$ in the sense that for all sufficiently big n and m the difference $|\langle f_n, g_m\rangle - s|$ becomes arbitrarily small.

Let us show that the scalar product is independent of a choice of fundamental sequences converging to f and g. Consider two pairs of fundamental sequences, $f_1, f_2, \ldots$ and $f'_1, f'_2, \ldots$ converging to f and $g_1, g_2, \ldots$ and $g'_1, g'_2, \ldots$ converging to g.

Consider the expression $\langle f_m - f'_m, g_n\rangle$. The sequence consisting of functions $f_n - f'_n$ is clearly fundamental, therefore, as shown above, there must exist a limit $\lim_{n,m\to\infty}\langle f_m - f'_m, g_n\rangle$. Let us evaluate this limit. There are coefficients $b_1^m, b_2^m, \ldots, b_p^m$ and elements $z_1^m, z_2^m, \ldots, z_p^m$ such that $f(\cdot) = \sum_{i=1}^p b_i^m \mathcal{K}(z_i^m, \cdot)$ ($p = p(m)$ may change as m varies). We have

$$\begin{aligned}\langle f_m - f'_m, g_n\rangle &= \langle f_m - f'_m, \sum_{i=1}^p b_i^m \mathcal{K}(z_i^m, \cdot)\rangle \\ &= \sum_{i=1}^m b_i^m (f_m(z_i) - f'_m(z_i)).\end{aligned}$$

Since f_m and f'_m converge pointwise to 0, this expression converges to zero as $m \to \infty$. Thus

$$\lim_{n,m\to\infty}\langle f_m - f'_m, g_n\rangle = 0.$$

Similarly

$$\lim_{n,m\to\infty}\langle f'_m, g_n - g'_n\rangle = 0.$$

Therefore the difference

$$\langle f_m, g_n\rangle - \langle f'_m, g'_n\rangle = \langle f_m - f'_m, g_n\rangle + \langle f'_m, g_n - g'_n\rangle$$

converges to 0 as $n, m \to \infty$. Our definition is thus independent of a particular choice of fundamental sequences.

The bilinearity of $\langle\cdot,\cdot\rangle$ on $\mathcal{F}$ is easy to check. The number $\|f\| = \langle f, f\rangle$ is non-negative as a limit of non-negative numbers. More precisely, let $f_1, f_2, \ldots \in \mathcal{F}_1$ be a fundamental sequence converging to f pointwise. Because

$$\begin{aligned}|f(x)| &= \lim_{n\to\infty}|f_n(x)| \\ &\leq \lim_{n\to\infty}\sqrt{\mathcal{K}(x,x)}\|f_n\| \\ &= \sqrt{\mathcal{K}(x,x)}\|f\|\end{aligned}$$

the equality $\|f\|$ implies that $f(x) = 0$ for all $x \in X$.

We have shown that $\mathcal{F}$ is indeed a linear space with a scalar product. Clearly, $\mathcal{F}_1 \subseteq \mathcal{F}$ and the scalar product on $\mathcal{F}$ extends that on $\mathcal{F}_1$.

Let us show that $\mathcal{F}$ is complete. First, let $f_1, f_2 \ldots$ be a fundamental sequence of elements of $\mathcal{F}_1$ converging pointwise to f. We have

$$\begin{aligned}\|f - f_n\| &= \sqrt{\langle f - f_n, f - f_n\rangle} \\ &= \sqrt{\lim_{m\to\infty} \langle f_m - f_n, f_m - f_n\rangle} \\ &= \sqrt{\lim_{m\to\infty} \|f_m - f_n\|}.\end{aligned}$$

This converges to 0 as $n \to 0$ and thus f is the limit of f_n in $\mathcal{F}$. Secondly, consider a fundamental sequence $f_1, f_2 \ldots$ of elements of $\mathcal{F}$. For each n there is $g_n \in \mathcal{F}_1$ such that $\|f_n - g_n\| \leq 1/2^n$. The sequence $g_1, g_2, \ldots$ is fundamental in $\mathcal{F}_1$ and therefore has a limit in $\mathcal{F}$. It must be the limit of $f_1, f_2 \ldots$ too.

It remains to show that the reproducing property holds of $\mathcal{F}$. It follows by continuity. Let $f_1, f_2 \ldots$ be a fundamental sequence of elements of $\mathcal{F}_1$ converging pointwise to f. We have

$$\begin{aligned}f(x) &= \lim_{n\to\infty} f_n(x) \\ &= \lim_{n\to\infty} \langle f_n(\cdot), \mathcal{K}(x, \cdot) \\ &= \langle f(\cdot), \mathcal{K}(x, \cdot)\rangle\end{aligned}$$

We have constructed a RKHS for $\mathcal{K}$. Note that $\mathcal{F}_1$ constructed in the previous subsection is dense in it.

3.3 Uniqueness

Let us show that the RKHS for a particular kernel $\mathcal{K}$ is unique. Let $\mathcal{F}$ be the RKHS constructed above and $\mathcal{H}$ be some other RKHS for the same kernel $\mathcal{K}$.

The definition of an RKHS implies that all functions $\mathcal{K}(x, \cdot)$ must belong to $\mathcal{H}$. The same must be true of their linear combinations $\sum_{i=1}^{n} a_i \mathcal{K}(x_i, \cdot)$. Thus $\mathcal{F}_1 \subseteq \mathcal{H}$ as a set.

Since the reproducing property holds on $\mathcal{H}$, on elements of $\mathcal{F}_1$ the scalar product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ must coincide with scalar product we constructed above. Thus $\mathcal{F}_1$ is a subspace of $\mathcal{H}$.

Because $\mathcal{H}$ is complete, all fundamental sequences from $\mathcal{F}_1$ should have limits in $\mathcal{H}$. In RKHSs convergence implies pointwise convergence and thus all pointwise limits of fundamental sequences from $\mathcal{F}_1$ belong to $\mathcal{H}$. Thus $\mathcal{F} \subseteq \mathcal{H}$ as a set. Because the scalar product is continuous w.r.t. itself, we have

$$\lim_{n\to\infty} \langle f_n, g_n\rangle = \langle f, g\rangle$$

for all sequences $f_1, f_2, \ldots$ and $g_1, g_2, \ldots$ such that $f_n \to f$ and $g_n \to g$ in $\mathcal{H}$ as $n \to \infty$. Thus the scalar product on $\mathcal{F}$ coincides with that on $\mathcal{H}$, or, in other terms, $\mathcal{F}$ is a closed subspace of $\mathcal{H}$.

Let $h \in \mathcal{H}$. We can represent it as $h = h_{\mathcal{F}} + h^{\perp}$, where $h_{\mathcal{F}} \in \mathcal{F}$ and $h^{\perp}$ is orthogonal to $\mathcal{F}$ and therefore to all functions $\mathcal{K}(x, \cdot)$, which belong to $\mathcal{F}_1 \subseteq \mathcal{F}$. Because the reproducing property holds on $\mathcal{H}$, we get

$$\begin{aligned}h(x) &= \langle h, \mathcal{K}(x, \cdot)\rangle \\ &= \langle h_{\mathcal{F}}, \mathcal{K}(x, \cdot)\rangle + \langle h^{\perp}, \mathcal{K}(x, \cdot)\rangle \\ &= \langle h_{\mathcal{F}}, \mathcal{K}(x, \cdot)\rangle \\ &= h_{\mathcal{F}}(x).\end{aligned}$$

Thus h coincides with $h_{\mathcal{F}}$ everywhere on X and $\mathcal{H} = \mathcal{F}$.

4 RKHSs and Prediction in Feature Spaces

We have shown that the three definitions of a kernel $\mathcal{K} : X^2 \to \mathbb{R}$ are equivalent:

- ◇ a positive semi-definite symmetric function;
- ◇ a reproducing kernel;
- ◇ the scalar product in a feature space, i.e., $\langle \Phi(x_1), \Phi(x_2) \rangle$, where Φ maps X into a Hilbert space H.

The RKHS for a particular kernel is unique. Note that uniqueness holds in a very strong sense: it is a unique set of functions with a uniquely defined scalar product; there are no isomorphisms or equivalences involved.

The mapping Φ in the third definition is by no means unique. Indeed, let $H = l_2$. Consider a right shift $R : l_2 \to l_2$ defined by $R(x_1 x_2 \ldots) = 0 x_1 x_2 \ldots$. The composition $R(\Phi)$ will produce the same kernel as Φ.

However there is some degree of uniqueness. Let $S \subseteq H$ be the closure of the linear span of all images $\Phi(x)$, $x \in X$. It is isomorphic to the RKHS.

4.1 RKHS Inside a Feature Space

Theorem 5. *For every mapping $\Phi : X \to H$, where H is a Hilbert space, the closure of the linear span of the image of X, i.e., $\overline{\mathrm{span}}(\Phi(X)) \subseteq H$, is isomorphic to the RKHS of the kernel $\mathcal{K}(x_1, x_2) = \langle \Phi(x_1), \Phi(x_2) \rangle$. There is a canonical isomorphism mapping $\Phi(x) \in H$ onto $\mathcal{K}(x, \cdot)$ from the RKHS.*

Proof. Let us denote the closure of the span by S and the RKHS by $\mathcal{F}$. Let $\mathcal{F}_1 \subseteq \mathcal{F}$ be the set of finite sums of the form $\sum_i a_i \mathcal{K}(x_i, \cdot)$, where $x_i \in X$ and a_i are some coefficients, as in the construction above.

We start by constructing the isomorphism L of $\mathrm{span}(\Phi(X))$ and $\mathcal{F}_1$. Put $L(\Phi(x)) = \mathcal{K}(x, \cdot)$ and, by linearity, $L(\sum_{i=1}^n a_i \Phi(x_i)) = \sum_{i=1}^n a_i \mathcal{K}(x_i, \cdot)$. We need to show that L is well-defined. Let $\sum_{i=1}^n a_i \Phi(x_i) = \sum_{j=1}^m b_j \Phi(z_i)$ for some coefficients a_i and b_i and elements $x_i, z_i \in X$. Then for every $x \in X$ we have

$$\left\langle \sum_{i=1}^n a_i \Phi(x_i), x \right\rangle = \left\langle \sum_{j=1}^m b_j \Phi(z_i), x \right\rangle,$$

i.e.,

$$\sum_{i=1}^n a_i \mathcal{K}(x_i, x) = \sum_{j=1}^m b_j \mathcal{K}(z_j, x)$$

by the definition of $\mathcal{K}$. The functions $\sum_{i=1}^n a_i \mathcal{K}(x_i, \cdot)$ and $\sum_{j=1}^m b_j \mathcal{K}(z_j, \cdot)$ coincide everywhere and thus L is well-defined.

The mapping L preserves the scalar product:

$$\begin{aligned}\left\langle \sum_{i=1}^n a_i \mathcal{K}(x_i, \cdot), \sum_{j=1}^m b_j \mathcal{K}(z_i, \cdot) \right\rangle &= \sum_{i,j} a_i b_j \mathcal{K}(x_i, z_j) \\ &= \left\langle \sum_{i=1}^n a_i \Phi(x_i), \sum_{j=1}^m b_j \Phi(z_j) \right\rangle.\end{aligned}$$

The mapping $L : \mathrm{span}(\Phi(X)) \to \mathcal{F}_1$ is surjective. Indeed, each $\sum_{i=1}^n a_i \mathcal{K}(x_i, \cdot)$ has an inverse image. The mapping is also injective. Assume the converse. Then there is a point $z = \sum_{i=1}^n a_i \Phi(x_i) \in \mathrm{span}(\Phi(X))$ such that $z \neq 0$ but $L(z) = 0$

in the RKHS. This is a contradiction because L preserves the scalar product and therefore the norm. Thus L is a bijection.

Let us extend L to the isomorphism of S and $\mathcal{F}$. Let $h_1, h_2, \ldots \in \mathrm{span}(\Phi(X))$ converge to $h \in S$. The sequence $h_1, h_2, \ldots$ is fundamental in $\mathrm{span}(\Phi(X))$. Since L preserves the scalar product on $\mathrm{span}(\Phi(X))$, the images $L(h_1), L(h_2), \ldots$ form a fundamental sequence in $\mathcal{F}$. It should converge. Put $L(h) = \lim_{i\to\infty} L(h_i)$.

Suppose that there are two sequences $h_1, h_2, \ldots \in \mathrm{span}(\Phi(X))$ and $g_1, g_2, \ldots \in \mathrm{span}(\Phi(X))$ converging to h. Let us mix the sequences into u_i (e.g., by letting $u_{2i} = h_i$ and $u_{2i-1} = g_i$, $i = 1, 2, \ldots$). The sequence $u_1, u_2, \ldots$ converges to h and is therefore fundamental. The images of u_i must form a fundamental sequence in $\mathcal{F}$ and must have a limit. All its subsequences should converge to the same limit. Thus $\lim_{i\to\infty} L(h_i) = \lim_{i\to\infty} L(g_i)$ and L is well-defined.

The scalar product is preserved by continuity. The surjectivity can be shown as follows. Let $f \in \mathcal{F}$ and let $f_1, f_2, \ldots \in \mathcal{F}_1$ converge to f. The inverse images of f_i must form a fundamental sequence in $\mathrm{span}(\Phi(X))$ and must have a limit h. It follows from the definition that $L(h) = f$. The injectivity on $\overline{\mathrm{span}}(\Phi(X))$ follows from the same argument as on $\mathrm{span}(\Phi(X))$.

The theorem follows. □

The mapping L can be extended to the mapping of the whole H by letting $L_H(h) = L(\mathrm{pr}_S(h))$, where $\mathrm{pr}_S : H \to S$ is the projection operator. The mapping L_H is no longer injective (unless H coincides with S) and no longer preserves the scalar product. However we have $\|L(h)\| = \min_{g\in H: L(g)=L(h)} \|L(g)\|$, where the minimum is attained on the projection $\mathrm{pr}_S(h)$.

4.2 Another Definition of RKHS

The above construction allows us to construct an interpretation of RKHS important for machine learning.

In competitive prediction we prove consistency results of the following type. We do not assume the existence of a 'correct' hypothesis but rather show that our method competes well with a class of some other predictors, such as all linear regressors. Therefore identifying and describing such natural classes is an important task.

In Hilbert spaces we have a natural equivalent of linear regressors. Those are linear functionals, or, as the Riess-Fischer Representation Theorem shows, scalar products by an element $h \in H$. After we have mapped X into a Hilbert space H, we can consider predictors $f : X \to \mathbb{R}$ of the form $f(x) = \langle h, \Phi(x)\rangle$.

Theorem 5 immediately implies that the class of such functions coincides with the RKHS. Indeed, there is a unique decomposition $h = h_0 + h^\perp$, where $h_0 = \mathrm{pr}_S(h)$ is the projection of h on $S = \overline{\mathrm{span}}(\Phi(X))$ and $h^\perp$ is orthogonal to S. We have

$$\begin{aligned} f(x) &= \langle h, \Phi(x)\rangle \\ &= \langle h_0, \Phi(x)\rangle \\ &= \langle L(h_0), L(\Phi(x))\rangle_{\mathcal{F}} \\ &= \langle L(h_0), \mathcal{K}(x,\cdot)\rangle_{\mathcal{F}} \\ &= g(x), \end{aligned}$$

where $g = L(h_0) \in \mathcal{F}$ belongs to the RKHS.

We may want to assign the norm $\|h\|$ to the predictor $f(x) = \langle h, \Phi(x)\rangle$; clearly $\|h\| \geq \|h_0\| = \|g\|_{\mathcal{F}}$. The space of predictors thus obtained does not exceed the RKHS and the norms of predictors are equal to or greater than those of the elements of the RKHS. Thus regressors in the feature space have no more power than functions from the RKHS.

We get the following theorem as a bonus.

Theorem 6. *Let $\Phi : X \to H$ be a mapping into a Hilbert space H. The space of functions $f : X \to \mathbb{R}$ defined by $f(x) = \langle h, \Phi(x)\rangle$, where $h \in H$, equipped with the norm $\|f\| = \min_{h\in H: f(\cdot)=\langle h,\Phi(\cdot)\rangle} \|h\|$ coincides with the reproducing kernel Hilbert space for the kernel defined by $\mathcal{K}(x_1, x_2) = \langle \Phi(x_1), \Phi(x_2)\rangle$.*

4.3 Ridge Regression in Feature Spaces

In this section we revisit the ridge regression problem from Section 1 and present one argument of great importance for competitive prediction.

Suppose that we have a sequence of signals and outcomes as in Section 1. On top of that suppose that we have a mapping $\Phi : X \to H$ from the set of signals X into a Hilbert feature space H. Take $h \in H$; as we said before, it can be considered as a regressor yielding the dependency $y = \langle h, \Phi(x)\rangle$. We may ask if there is h minimising the expression

$$L_{\mathrm{RR}}(h) = a\|h\|^2 + \sum_{i=1}^{T} \left(\langle h, \Phi(x_i)\rangle - y_i\right)^2 .$$

with $a > 0$.

Consider the predictor

$$\gamma_{\mathrm{RR}} = Y'(aI + K)^{-1}k,$$

where

$$K = \begin{pmatrix} \mathcal{K}(x_1,x_1) & \mathcal{K}(x_1,x_2) & \dots & \mathcal{K}(x_1,x_T) \\ \mathcal{K}(x_2,x_1) & \mathcal{K}(x_2,x_2) & \dots & \mathcal{K}(x_2,x_T) \\ \vdots & \vdots & \ddots & \vdots \\ \mathcal{K}(x_T,x_1) & \mathcal{K}(x_T,x_2) & \dots & \mathcal{K}(x_T,x_T) \end{pmatrix}$$

and

$$k = \begin{pmatrix} \mathcal{K}(x_1,x) \\ \mathcal{K}(x_2,x) \\ \dots \\ \mathcal{K}(x_T,x) \end{pmatrix};$$

It qualifies as a regressor of the above type. Indeed, it is a linear combination of $\mathcal{K}(x_i, \cdot) = \langle \Phi(x_i), \Phi(\cdot)\rangle$. Let us show that it minimises L_{RR}.

Let $H_0 \subset H$ be a subspace of H consisting of all linear combinations of $\Phi(x_i)$ (it is closed because it is finite-dimensional). Take $h \in H$. It can be decomposed into a sum $h = h_0 + h^{\perp}$, where $h^{\perp}$ is orthogonal to H. For all $i = 1, 2, \dots, T$ we have $\langle h, \Phi(x_i)\rangle = \langle h_0, \Phi(x_i)\rangle$; we also have $\|h_0\| \le \|h\|$. Therefore $L_{\mathrm{RR}}(h_0) \le L_{\mathrm{RR}}(h_0)$. When minimising L_{RR} we can restrict ourselves to predictors from H_0, i.e., linear combinations of $\Phi(x_i)$!

Because H_0 is finite-dimensional, the arguments from Section 1 apply and the ridge regression turns out to provide the minimum.

This argument essentially repeats the representer theorem.

5 Subspaces of RKHSs and Hierarchies of Kernels

Consider an RKHS $\mathcal{F}$ corresponding to a kernel $\mathcal{K}$. Let $\mathcal{F}' \subseteq \mathcal{F}$ be a subspace of $\mathcal{F}$. Clearly, $\mathcal{F}'$ is a RKHS. This can be shown as follows. The evaluation functional is continuous on $\mathcal{F}$. Its restriction on $\mathcal{F}'$ should remain continuous and therefore

$\mathcal{F}'$ is a RKHS. This does not contradict the uniqueness theorem. If $\mathcal{F}'$ is a proper subspace of $\mathcal{F}$, it is an RKHS for a *different* kernel $\mathcal{K}'$.

Suppose that $\mathcal{F}$ itself is a subspace of some Hilbert space of functions $\mathcal{F}''$. As we discussed above, in applications such as machine learning it does not make much sense to consider spaces where the evaluation functional is not continuous, and therefore $\mathcal{F}''$ should be an RKHS with its own kernel too.

One can say that RKHSs form hierarchies: larger spaces have more power than smaller spaces. However each of them has its own kernel. In competitive prediction the natural competitors of a kernel method are the functions from the corresponding RKHS. Other RKHSs require the use of a different kernel.

The rest of this section contains some results clarifying the structure of this hierarchy.

Theorem 7. *Let a space $\mathcal{F}$ of real-valued functions on X be the RKHS corresponding to a kernel $\mathcal{K} : X^2 \to \mathbb{R}$. If $\mathcal{F}' \subseteq \mathcal{F}$ is a subspace of $\mathcal{F}$, then $\mathcal{F}'$ is an RKHS and has a kernel $\mathcal{K}' : X^2 \to \mathbb{R}$. If $\mathcal{F}''$ is the orthogonal complement of $\mathcal{F}'$, then it is also an RKHS and it has the kernel $\mathcal{K}'' : X^2 \to \mathbb{R}$ such that $\mathcal{K}'(x_1, x_2) + \mathcal{K}''(x_1, x_2) = \mathcal{K}(x_1, x_2)$ for all $x_1, x_2 \in X$.*

Proof. Let pr' and pr' be the projection operators from $\mathcal{F}$ to $\mathcal{F}'$ and $\mathcal{F}''$, respectively.

Take a point $x \in X$. The evaluation functional on $\mathcal{F}$ equals the scalar product by $\mathcal{K}(x, \cdot)$. It is easy to see that $\mathrm{pr}'(\mathcal{K}(x, \cdot))$ plays the same role in $\mathcal{F}'$. Indeed, $\mathrm{pr}'(\mathcal{K}(x, \cdot)) \in \mathcal{F}'$ and for every function $f \in \mathcal{F}'$ we have

$$\begin{aligned} f(x) &= \langle \mathcal{K}(x, \cdot), f(\cdot) \rangle \\ &= \langle \mathrm{pr}'(\mathcal{K}(x, \cdot)), f(\cdot) \rangle. \end{aligned}$$

Put $\mathcal{K}'(x_1, x_2) = \langle \mathrm{pr}'(\mathcal{K}(x_1, \cdot)), \mathrm{pr}'(\mathcal{K}(x_2, \cdot)) \rangle$. Let us prove that it is the kernel for $\mathcal{F}'$

We do this by showing that $\mathcal{K}'(x_1, x_2)$ as a function of x_2 coincides with the projection $\mathrm{pr}'(\mathcal{K}(x_1, \cdot))$. Fix x_1 and denote the function $\mathrm{pr}'(\mathcal{K}(x_1, \cdot))$ by $f(\cdot)$. We have $f \in \mathcal{F}'$. The above argument implies that $\langle f(\cdot), \mathrm{pr}'(\mathcal{K}(x_2, \cdot)) \rangle = f(x_2)$ for every $x_2 \in X$ and thus $\mathcal{K}'(x_1, \cdot) = f(\cdot) \in \mathcal{F}$. The reproducing property follows.

Similarly $\mathcal{K}''(x_1, x_2) = \langle \mathrm{pr}''(\mathcal{K}(x_1, \cdot)), \mathrm{pr}''(\mathcal{K}(x_2, \cdot)) \rangle$ is the kernel for $\mathcal{F}''$.

Let $f, g \in \mathcal{F}$. We have $f = \mathrm{pr}'(f) + \mathrm{pr}''(f)$ and $g = \mathrm{pr}'(g) + \mathrm{pr}''(g)$; therefore

$$\begin{aligned} \langle f, g \rangle &= \langle \mathrm{pr}'(f) + \mathrm{pr}''(f), \mathrm{pr}'(g) + \mathrm{pr}''(g) \rangle \\ &= \langle \mathrm{pr}'(f), \mathrm{pr}'(g) \rangle + \langle \mathrm{pr}''(f), \mathrm{pr}''(g) \rangle . \end{aligned}$$

By taking $f = \mathcal{K}(x_1, \cdot)$ and $g = \mathcal{K}(x_2, \cdot)$ we get $\mathcal{K}'(x_1, x_2) + \mathcal{K}''(x_1, x_2) = \mathcal{K}(x_1, x_2)$. □

Theorem 8. *Let $\mathcal{K}, \mathcal{K}', \mathcal{K}'' : X^2 \to \mathbb{R}$ be three kernels on X such that $\mathcal{K}(x_1, x_2) = \mathcal{K}'(x_1, x_2) + \mathcal{K}''(x_1, x_2)$ for all $x_1, x_2 \in X$. Then the RKHSs $\mathcal{F}'$ and $\mathcal{F}''$ corresponding to the kernels $\mathcal{K}'$ and $\mathcal{K}''$ are subsets (but not necessarily subspaces) of the RKHS $\mathcal{F}$ corresponding to the kernel $\mathcal{K}$. For each $f \in \mathcal{F}$ there are functions $f_1 \in \mathcal{F}'$ and $f_2 \in \mathcal{F}''$ such that $f = f_1 + f_2$ and for its norm we have the equality*

$$\|f\|^2_{\mathcal{F}} = \min_{f_1 \in \mathcal{F}_1, f_2 \in \mathcal{F}_2 : f = f_1 + f_2} \left(\|f_1\|^2_{\mathcal{F}'} + \|f_2\|^2_{\mathcal{F}''} \right) .$$

If $\mathcal{F}'$ and $\mathcal{F}''$ have only the identical zero function in common (i.e., $\mathcal{F}' \cap \mathcal{F}'' = \{0\}$), then they are subspaces of $\mathcal{F}$ and each one is the orthogonal complement of the other.

It is easy to see that $\mathcal{F}'$ does not have to be a subspace of $\mathcal{F}$. Indeed, let $\mathcal{K}' = \mathcal{K}''$ and $\mathcal{K} = 2\mathcal{K}'$. Clearly if we take the set of functions constituting $\mathcal{F}$ and equip it with the scalar product $\langle \cdot, \cdot \rangle_{\mathcal{F}}/2$, we get the RKHS for $\mathcal{F}'$. It is a subset but not a subspace of $\mathcal{F}$.

Proof. Let Φ' and Φ'' mapping X into Hilbert spaces H' and H'', respectively, be feature maps giving rise to the kernels $\mathcal{K}'$ and $\mathcal{K}''$, i.e.,

$$\mathcal{K}'(x_1, x_2) = \langle \Phi'(x_1), \Phi'(x_2)\rangle$$
$$\mathcal{K}''(x_1, x_2) = \langle \Phi''(x_1), \Phi''(x_2)\rangle.$$

Let H be the Hilbert space consisting of pairs (h', h'') such that $h' \in H'$ and $h'' \in H''$ with the scalar product given by

$$\langle (h'_1, h''_1), (h'_2, h''_2)\rangle_H = \langle h'_1, h'_2\rangle_{H_1} + \langle h''_1, h''_2\rangle_{H_2}.$$

Take $\Phi : X \to H$ defined by

$$\Phi(x) = (\Phi'(x), \Phi''(x)).$$

It is easy to see that

$$\mathcal{K}(x_1, x_2) = \langle \Phi(x_1), \Phi(x_2)\rangle.$$

The results of Subsect. 4.2 imply that the RKHS for $\mathcal{K}$ coincides with the set of functions $\langle h, \Phi(\cdot)\rangle$, where $h \in H$. Similarly, the RKHSs for $\mathcal{K}'$ and $\mathcal{K}''$ consist of all functions $\langle h', \Phi'(\cdot)\rangle$ and $\langle h'', \Phi''(\cdot)\rangle$, respectively, where $h' \in H'$ and $h'' \in H''$.

For every $h' \in H'$ the element $(h', 0)$ belongs to H (0 is the zero element of H'' here). We have

$$\langle (h', 0), \Phi(\cdot)\rangle_H = \langle (h', \Phi'(\cdot))\rangle_{H_1}$$

and therefore $\mathcal{F}' \subseteq \mathcal{F}$ as a set; the same argument applies to $\mathcal{F}''$. The decomposition

$$\langle (h', h''), \Phi(\cdot)\rangle = \langle h', \Phi'(\cdot)\rangle + \langle h'', \Phi''(\cdot)\rangle$$

implies that each $f \in \mathcal{F}$ can be decomposed into the sum of $f_1 \in \mathcal{F}'$ and $f_2 \in \mathcal{F}''$.

We have

$$\begin{aligned}\|f\|^2_{\mathcal{F}} &= \min_{h\in H: f(\cdot)=\langle h,\Phi(\cdot)\rangle} \|h\|^2_H \\ &= \min_{h'\in H, h''\in H'': f(\cdot)=\langle h',\Phi'(\cdot)\rangle+\langle h'',\Phi''(\cdot)\rangle} \left(\|h'\|^2_{H_1} + \|h''\|^2_{H_2}\right)\end{aligned}$$

The minimum is taken over pairs of (h', h''); clearly, we can take the minimum over all pairs of $f_1 \in \mathcal{F}'$ and $f_2 \in \mathcal{F}''$ such that $f = f_1 + f_2$; indeed, $\|f_1\|_{\mathcal{F}'} = \min_{h'\in H_1: f_1(\cdot)=\langle h',\mathcal{K}'(\cdot)\rangle_{H_1}} \|h'\|_{H_1}$.

Now let $\mathcal{F}' \cap \mathcal{F}'' = \{0\}$. Under this assumption every $f \in \mathcal{F}$ has a unique decomposition $f = f_1 + f_2$, where $f_1 \in \mathcal{F}'$ and $f_2 \in \mathcal{F}''$. Indeed, if $f = f_1 + f_2 = g_1 + g_2$, then $f_1 - g_1 = f_2 - g_2$. The function of the left-hand side belongs to $\mathcal{F}'$ and the function on the right-hand side belongs to $\mathcal{F}''$ and therefore they are both equal to zero. Thus for every pair $f_1 \in \mathcal{F}'$ and $f_2 \in \mathcal{F}''$ we have $\|f_1 + f_2\|^2_{\mathcal{F}} = \|f_1\|^2_{\mathcal{F}'} + \|f_2\|^2_{\mathcal{F}''}$.

Take $f_2 = 0$. Then this equality implies that $\|f_1\|_{\mathcal{F}} = \|f_1\|_{\mathcal{F}'}$. Taking $f_1 = 0$ leads to $\|f_2\|_{\mathcal{F}} = \|f_2\|_{\mathcal{F}''}$. The norms on $\mathcal{F}'$ and $\mathcal{F}''$ coincide with the norm on $\mathcal{F}$; the same should apply to the scalar product and thus $\mathcal{F}'$ and $\mathcal{F}''$ are subspaces rather than just subsets of $\mathcal{F}$.

Picking arbitrary f_1 and f_2 and applying Pythagoras theorem yields $\|f_1+f_2\|^2_{\mathcal{F}} = \|f_1\|^2_{\mathcal{F}} + \|f_2\|^2_{\mathcal{F}} + 2\langle f_1, f_2\rangle_{\mathcal{F}}$. Comparing this with the above equality implies that $\langle f_1, f_2\rangle_{\mathcal{F}} = 0$, i.e., $\mathcal{F}'$ and $\mathcal{F}''$ are orthogonal subspaces.

□

Let us introduce a relation on kernels on a set X. We will write $\mathcal{K}' \ll \mathcal{K}$ if the difference $\mathcal{K}''(x_1, x_2) = \mathcal{K}(x_1, x_2) - \mathcal{K}'(x_1, x_2)$ is a kernel.

If this relation holds, then $\mathcal{K}'(x, x) \leq \mathcal{K}(x, x)$ for all $x \in X$. Indeed, since $\mathcal{K}''$ is a kernel, the 1×1 matrix $\mathcal{K}''(x, x)$ is positive semi-definite, i.e., $\mathcal{K}''(x, x) \geq$

0. This observation justifies the notation to some extent and implies that $\ll$ is antisymmetric. Indeed, if $\mathcal{K}' \ll \mathcal{K}$ and $\mathcal{K} \ll \mathcal{K}'$ then for $\mathcal{K}'' = \mathcal{K} - \mathcal{K}'$ we get $\mathcal{K}''(x,x) = 0$ for all $x \in X$. Theorem 2 implies that for every f from the RKHS of $\mathcal{K}''$ and every $x \in X$ we have $f(x) = 0$ and thus f is identically zero. This implies that $\mathcal{K}'' = 0$.

Clearly, $\ll$ is transitive: if $\mathcal{K}' \ll \mathcal{K}'' \ll \mathcal{K}''$, then $\mathcal{K}' \ll \mathcal{K}'''$.

The theorems above imply that $\ll$ for kernels is closely linked with the relation $\subseteq$ on their RKHSs. However no direct correspondence has been shown. The following results close the gap.

Theorem 9. *Let $\mathcal{K}$ and $\mathcal{K}'$ be two kernels on the same set and let $\mathcal{F}$ and $\mathcal{F}'$ be their RKHSs. If $\mathcal{K}' \ll \mathcal{K}$, then $\mathcal{F}' \subseteq \mathcal{F}$ as a set and for every $f_1 \in \mathcal{F}'$ we have*

$$\|f_1\|_{\mathcal{F}} \leq \|f_1\|_{\mathcal{F}'}.$$

This theorem follows from our previous results. Indeed, the square $\|f\|^2_{\mathcal{F}}$ is given by the minimum of $\|f_1\|^2_{\mathcal{F}'} + \|f_2\|^2_{\mathcal{F}''}$ taken over all decompositions $f = f_1 + f_2$, where $\mathcal{F}''$ is the RKHS corresponding to the difference $\mathcal{K} - \mathcal{K}'$. Every $f_1 \in \mathcal{F}'$ can be represented as $f_1 + 0$, which implies the inequality in the theorem.

The opposite result holds.

Theorem 10. *Let $\mathcal{K}$ and $\mathcal{F}$ be its RKHSs. If $\mathcal{F}' \subseteq \mathcal{F}$ as a set and $\mathcal{F}'$ forms a Hilbert space w.r.t. a norm $\|\cdot\|_{\mathcal{F}_1}$ such that*

$$\|f_1\|_{\mathcal{F}} \leq \|f_1\|_{\mathcal{F}'}$$

for every $f_1 \in \mathcal{F}'$, then $\mathcal{F}'$ is an RKHS and its reproducing kernel $\mathcal{K}'$ satisfies $\mathcal{K}' \ll \mathcal{K}$.

It is easy to see that the inequality on the norms cannot be omitted. Consider some kernel $\mathcal{K}'$ on X and let $\mathcal{K} = \mathcal{K}' + \mathcal{K}' = 2\mathcal{K}$. Let $\mathcal{F}'$ be the RKHS for $\mathcal{K}'$. For every $f \in \mathcal{F}'$ we have

$$\begin{aligned} f(x) &= \langle f(\cdot), \mathcal{K}'(x,\cdot)\rangle_{\mathcal{F}'} \\ &= \frac{1}{2}\langle f(\cdot), 2\mathcal{K}'(x,\cdot)\rangle_{\mathcal{F}'} \end{aligned}$$

and therefore $\mathcal{F}$ that coincides with $\mathcal{F}'$ as a set and has the scalar product $\langle\cdot,\cdot\rangle_{\mathcal{F}} = \frac{1}{2}\langle\cdot,\cdot\rangle_{\mathcal{F}'}$ is the RKHS for $\mathcal{K}$. We have $\|f\|_{\mathcal{F}'} = \sqrt{2}\|f\|_{\mathcal{F}} \geq \|f\|_{\mathcal{F}}$. However let us consider $\mathcal{F}$ as a subset of $\mathcal{F}'$. It satisfies the conditions of the theorem apart from the norm clause but $\mathcal{K} \not\ll \mathcal{K}'$.

The proof of the theorem is beyond the scope of this article and can be found in [Aro50], pp.355-356.

References

[Aro43] N. Aronszajn. La théorie des noyaux reproduisants et ses applications. Première partie. *Proceedings of the Cambridge Philosophical Society*, 39:133–153, 1943.

[Aro50] N. Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society.*, 68:337–404, 1950.

[SS02] B. Schölkopf and A. J. Smola. *Learning with Kernels.* MIT Press, Cambridge, MA, 2002.