Reproducing Kernel Hilbert Spaces

Myriam Frikha, Guillaume Sallé

January 12, 2024

Contents

1 Introduction

The journey of Reproducing Kernel Hilbert spaces (RKHS) begins with the foundational work in functional analysis and Hilbert spaces, gradually evolving into the modern form we know today. The idea of a 'reproducing kernel' was introduced as a function that allows the evaluation of inner products in a more general setting. The formal definition of a reproducing kernel was given by Nachman Aronszajn in his seminal article [1] of 1950, but its roots can be traced back to earlier works in the 20th century with for example J. Mercer's work on integral operators in 1909 and S. Bergman's work on complex analysis. Since Aronszajn's work, the theory of RKHS has found applications in a wide range of fields: In statistics, it provided a new approach to the theory of estimation and prediction. In machine learning, kernels derived from RKHS theory are central to the functioning of Support Vector Machines.

In the first part of this article, we present the theoretical foundations of RKHS and their reproducing kernels. In the second part, we focus on their applications by discussing an optimized margin-based performance bound as well as an introduction to Support Vector Machines.

2 Notations

We will consider Hilbert spaces over either the field real numbers $\mathbb{R}$ or the field of complex numbers $\mathbb{C}$. We will use $\mathbb{F}$ to denote either $\mathbb{R}$ or $\mathbb{C}$, in order to state definition and results that hold in both cases. For the real case, the conjugate operator will be the identity function, and anti-linearity will just be linearity.
In this whole section, we consider a set $\mathcal{X}$. From a statistical point of view, it represent an input space where lies some data.
The vector space of functions from $\mathcal{X}$ to $\mathbb{F}$ with pointwise operations is noted $\mathcal{F}(\mathcal{X}, \mathbb{F})$, while the set of such functions is noted $\mathbb{F}^{\mathcal{X}}$.
If a space $\mathcal{H}$ has an inner product, it will be noted $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ and its associated norm is noted $\| \cdot \|_{\mathcal{H}}$.
The open ball of center c and radius r in a vector space E is noted $B_E(0, 1)$.
The set of integer comprised between some integers n and m is noted $[\![n, m]\!]$.
Let $d \geq 1$ be an integer which will be used for the dimension of finite dimension vector spaces.

3 RKHS

3.1 Definition

Definition 1 (RKHS)**.** *Let $\mathcal{H} \subset \mathbb{F}^{\mathcal{X}}$ be a $\mathbb{F}$-Hilbert space of functions over $\mathcal{X}$. We say that $\mathcal{H}$ is a* Reproducing Kernel Hilbert space *(RKHS) over $\mathcal{X}$ if there exists a function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{F}$ verifying:*

i) $\forall x \in \mathcal{X}, \quad K(\cdot, x) \in \mathcal{H}$

ii) $\forall f \in \mathcal{H}, \quad \forall x \in \mathcal{X}, \quad \langle f, K(\cdot, x) \rangle_{\mathcal{H}} = f(x)$ *(Reproducing property)*

The function K is unique and is called a reproducing kernel *of $\mathcal{H}$.*

The reproducing property implies that the evaluation functions are linear on $\mathcal{H}$. In other words, addition and scalar multiplication in $\mathcal{H}$ are pointwise operations on functions, so $\mathcal{H}$ is a subspace of $\mathcal{F}(\mathcal{X}, \mathbb{F})$.

Example 1. *We give an example of a RKHS and a counter-example (in some sense):*

- *For any set $\mathcal{X}$, we can define the vector space*

$$\ell^2(\mathcal{X}) := \{f : \mathcal{X} \to \mathbb{F} \, : \sum_{x \in \mathcal{X}} |f(x)|^2 < +\infty\}$$

endowed with the following inner product

$$\forall f, g \in \ell^2(\mathcal{X}), \qquad \langle f, g \rangle_{\ell^2(\mathcal{X})} := \sum_{x \in \mathcal{X}} f(x) \overline{g(x)}.$$

With these definitions $\ell^2(\mathcal{X})$ becomes a Hilbert space of functions on $\mathcal{X}$. Note that $\mathcal{X}$ is not assumed to be countable: the sum over any family of positive terms $(a_i)_{i \in I}$ is defined as

$$\sum_{i \in I} a_i = \sup_{J \subset I, |J| < \infty} \sum_{j \in J} a_j$$

For any x, y in $\mathcal{X}$, let $K(x, y) := \delta_{i,j}$ (it is the Kronecker symbol which is equal to 1 if $i = j$ and 0 otherwise). It is easy to see that for all x in $\mathcal{X}$, $K(\cdot, x)$ belongs to $\ell^2(\mathcal{X})$ and that K has the reproducing property over $\ell^2(\mathcal{X})$. Therefore, $\ell^2(\mathcal{X})$ is a RKHS with K as reproducing kernel.
Two important particular cases are:

  - *the finite dimensional case $X = [\![1, n]\!]$, where $\ell^2([\![1, n]\!])$ is $\mathbb{F}^n$,*
  - *the countable case $X = \mathbb{N}$, where $\ell^2(\mathbb{N})$ is the Hilbert space of square integrable sequences of $\mathbb{F}$.*

- *The Hilbert space $L^2(0,1)$, which can be seen as the completion of the space of continuous functions $C([0,1])$, is not a space of functions: its elements are equivalence classes of functions that differ at most on a set of measure 0, and therefore we can't talk of the evaluation in one point of a "function" f in $L^2(0,1)$.*
However, one may wonder if the RKHS reproducing property can be verified almost surely, i.e. if there exists a function measurable $K : [0,1] \times [0,1] \to \mathbb{F}$ with the property that for each $x \in [0,1]$ the function $K(\cdot\,, x)$ is square-integrable and such that

$$\forall f \in L^2(0,1), \qquad f(x) = \langle f, K(\cdot\,, x)\rangle_{L^2(0,1)} = \int_{[0,1]} f\overline{K(\cdot\,, x)}\, d\lambda \qquad \text{a.s.}$$

With this inner product, K defines an operator T_K that maps linearly the functions $f \in L^2(0,1)$ to functions over $[0,1]$:

$$\forall f \in L^2([0,1]), \qquad T_K(f) := \int_{[0,1]} f\overline{K(\cdot\,, x)}\, d\lambda.$$

Such an operator is called an integral operator *and the function K is called the* integral kernel *of the integral operator. To ensure that $T_K(f) \in L^2([0,1])$, we can add the condition that $K \in L^2([0,1]^2)$ and by the Cauchy-Schwarz inequality we have*

$$\|T_K(f)\|^2_{L^2([0,1])} = \int_{[0,1]} \left| \int_{[0,1]} f(y)K(y,x)\, dy \right|^2 dx \leq \int_{[0,1]} \|f\|_2^2 \|K(\cdot\,, x)\|_2^2\, dx = \|f\|_2^2 \|K\|_2^2 < \infty.$$

Such a kernel $K \in L^2([0,1]^2)$ is called a Hilbert-Schmidt kernel. *With this vocabulary, the question of the existence of an "almost surely" reproducing kernel over $L^2([0,1])$ becomes whether the identity operator over $L^2(0,1)$ is an integral operator. J. von Neumann [9] was the first to show that the answer is no. A standard argument is to use the fact that operators defined with Hilbert-Schmidt kernels are a particular case of* Hilbert-Schmidt operators*, and then to use the fact that Hilbert-Schmidt operators are* compact operators*. We give these definitions because they are important notions in the theory of RKHS, but we will not prove this compact property.*

Definition 2. *(Hilbert-Schmidt operator) Let $\mathcal{H}$ be a Hilbert space, and $(e_i)_{i\in I}$ an orthonormal basis. A bounded linear operator $T : \mathcal{H} \to \mathcal{H}$ is called a* Hilbert-Schmidt operator *if it has a finite* Hilbert-Schmidt norm*, which is defined by:*

$$\|T\|^2_{HS} := \sum_{i\in I} \|T(e_i)\|^2_{\mathcal{H}}.$$

The existence of an orthonormal basis is always assured with the axiom of choice, but the index set I does not need to be countable. This definition is independent of the choice of the orthonormal basis.

Definition 3. *(Compact operator) Let E be a Banach space. An operator $T \in \mathcal{L}(E)$ is called a* compact operator *if the image by T of the unit ball $T(\overline{B_E})$ has compact closure in E.*

Since balls in infinite-dimensional Banach spaces are not compact, this is a much stronger condition than just the boundness of T. In particular, the identity application in $L^2([0,1])$ is not a compact operator. This ensures that the identity operator does not have an integral kernel.

We now give a first characterization of the notion of RKHS. Moreoever, the uniqueness of the reproducing kernel K is shown in the proof of this theorem. From now on, we will talk about *the* reproducing kernel of a RKHS.

Theorem 1. *A Hilbert space $\mathcal{H} \subset \mathbb{F}^{\mathcal{X}}$ of functions on $\mathcal{X}$ is a RKHS if and only if all the evaluation functions $(e_x)_{x\in\mathcal{X}}$ are linear and continuous on $\mathcal{H}$.*

Proof. If $\mathcal{H}$ is a RKHS with reproducing kernel K, then for any $x \in \mathcal{X}$, we have by the Cauchy-Schwarz inequality:

$$\forall f \in \mathcal{H}, \quad |e_x(f)| = |f(x)| = |\langle f, K(\cdot\,, x)\rangle_{\mathcal{H}}| \leq \|f\| \cdot \|K(\cdot\,, x)\|$$

Conversely, for any $x \in \mathcal{X}$, if the evaluation function e_x is a continuous linear form, by Riesz representation theorem there exists a unique function N_x in $\mathcal{H}$ such that

$$\forall f \in \mathcal{H}, \quad \langle f, N_x\rangle_{\mathcal{H}} = f(x).$$

Then the function $K(y, x) := N_x(y)$ verifies the properties i) and ii) of the definition of RKHS, and is the only one to do so by unicity in the Riesz theorem. □

An important consequence of this theorem is that in a RKHS, a sequence converging in the norm sense converges also pointwise to the same limit.
Let's give a simple example of an inner space of functions where it is not the case. Take $X = [0, 1]$, and $\mathcal{H}_0$ the space of continuous functions over $\mathcal{X}$ with the inner product:

$$\forall (f, g) \in \mathcal{H}^2, \qquad \langle f, g\rangle_{\mathcal{H}} := \int_0^1 f(x)\overline{g(x)}\, dx.$$

Then the polynomial sequence $(x^n)_{n\in\mathbb{N}}$ converges to the null function in the norm sense but not pointwise.

3.2 Characterisation of kernels

If a function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{F}$ is the reproducing kernel of some space $\mathcal{H}$, it will inherit some properties of its inner product. Recall that an inner product over a vector space is a sesquilinear form that is conjugate symmetric (which is just symmetry in the real case $\mathbb{F} = \mathbb{R}$) and positive definite. We can deduce from the relation

$$\forall (x, y) \in \mathcal{X} \times \mathcal{X}, \quad K(x, y) = \langle K(\cdot\,, y), K(\cdot\,, x)\rangle_{\mathcal{H}}$$

that K is conjugate symmetric and also that K inherits in some way of the property of positive definiteness of the inner product of $\mathcal{H}$ (this will be precised in the next definition). However, there is no need for $x \mapsto K(\cdot\,, x)$ to be linear (and neither for $x \mapsto K(x, \cdot)$ to be antilinear by conjugate symmetry). Actually, although $\mathcal{X}$ will often naturally have one, we did not even require a vector space structure on $\mathcal{X}$! This potential non-linearity is a crucial element of reproducing kernels and allows K to explore non-linear relations between points in $\mathcal{X}$, by embedding them into a Hilbert space and looking at linear relations through the inner product. Of course, a reproducing kernel can be linear in its second variable (and anti-linear in its first variable), but then it just boils down to an inner product on $\mathcal{X}$.

Definition 4. *A function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{F}$ is called* positive definite *if*

$$\forall n \in \mathbb{N}^*, \quad \forall (a_1, ..., a_n) \in \mathbb{F}^n, \quad \forall (x_1, ..., x_n) \in \mathcal{X}^n, \quad \sum_{i=1}^{n}\sum_{j=1}^{n} a_i\overline{a_j}K(x_i, x_j) \in \mathbb{R}^+. \tag{1}$$

In the complex case $\mathbb{F} = \mathbb{C}$, the property of conjugate symmetry can be deduced from (1). Indeed, assume that K verifies (1), and let $(x, y) \in \mathcal{X} \times \mathcal{X}$. Then for any $(a_1, a_2) \in \mathbb{C}^2$, the following number is real:

$$C(a_1, a_2) := |a_1|^2K(x, x) + a_1\overline{a_2}K(x, y) + a_2\overline{a_1}K(y, x) + |a_2|^2K(y, y) \in \mathbb{R}$$

Thus, putting first $a_1 = a_2 = 1$ and then $a_1 = i$ and $a_2 = 1$, we get

$$A := K(x, y) + K(y, x) = C(1, 1) - C(1, 0) - C(0, 1) \in \mathbb{R}$$

$$B := iK(x, y) - iK(y, x) = C(i, 1) - C(1, 0) - C(0, 1) \in \mathbb{R}$$

It follows that

$$A + iB = 2K(y, x)$$

$$A - iB = 2K(x, y)$$

hence $K(x, y)$ is the conjugate of $K(y, x)$.
However in the real case, one cannot deduce the symmetry from (1).

The condition (1) is equivalent to the positive semi-definiteness of the matrix $(K(x_i, x_j))_{1 \leq i,j \leq n}$ for any choice of $n \in \mathbb{N}^*$ and $(x_1, ..., x_n) \in \mathcal{X}^n$. If K is a reproducing kernel, this matrix is the Gram matrix of the vectors $(K(\cdot\,, x_i))_{1 \leq i \leq n}$ which is always positive semi-definite :

Lemma 1. *Let K be the reproducing kernel of a Hilbert space $\mathcal{H}$. Then K is positive definite.*

Proof. We have

$$\sum_{i=1}^{n} \sum_{j=1}^{n} a_i \overline{a_j} K(x_i, x_j) = \left\| \sum_{i=1}^{n} a_i K(\cdot\,, x_i) \right\|_{\mathcal{H}}^2 \in \mathbb{R}^+.$$

□

The following theorem shows that these two properties of positive definiteness and conjugate symmetry verified by reproducing kernels are enough to characterize all reproducing kernels.

Theorem 2. *(Moore-Aronszajn) Let K be a conjugate symmetric and positive definite function on $\mathcal{X} \times \mathcal{X}$. There exists a unique Hilbert space $\mathcal{H}$ of functions on $\mathcal{X}$ with K as reproducing kernel.*

We give a proof inspired from [4], but using instead a general result on abstract completion of a metric space (see appendix A for a sketch of this construction).

Proof. We are going to build a Hilbert space $\mathcal{H}$ that satisfies the properties of the RKHS definition with K as kernel. Let $\mathcal{H}_0$ be the linear span of $\{K(\cdot\,, x),\, x \in \mathcal{X}\}$ in $\mathcal{F}(\mathcal{X}, \mathbb{F})$. The first RKHS property implies that $\mathcal{H}_0 \subset \mathcal{H}$. Now we want to define an inner product over $\mathcal{H}_0$. For any f, g in $\mathcal{H}_0$ and any decomposition

$$f = \sum_{i=1}^{n} \alpha_i K(\cdot\,, x_i), \qquad g = \sum_{j=1}^{m} \beta_j K(\cdot\,, y_j),$$

the reproducing property of a RKHS requires the inner product of $\mathcal{H}_0$ to verify by linearity

$$\langle f, g \rangle_{\mathcal{H}_0} = \sum_{i=1}^{n} \sum_{j=1}^{m} \alpha_i \overline{\beta_j} \langle K(\cdot\,, x_i), K(\cdot\,, y_j) \rangle_{\mathcal{H}_0} = \sum_{i=1}^{n} \sum_{j=1}^{m} \alpha_i \overline{\beta_j} K(y_j, x_i).$$

Lets take this last equality as definition of the inner product of f and g in $\mathcal{H}_0$. Although the representations of f and g are not necessarily unique, this definition only depends on f and g through their values since we have

$$\langle f, g \rangle_{\mathcal{H}_0} = \sum_{i=1}^{n} \alpha_i \overline{g(x_i)} = \sum_{j=1}^{m} \overline{\beta_j} f(y_j).$$

The first sum shows that this quantity does not depend on the decomposition of g, and the second sum shows the same for f.
In particular, for any x in $\mathcal{X}$, by taking $g = K(\cdot\,, x)$ we get the reproducing property on $\mathcal{H}_0$

$$\langle f, K(\cdot\,, x) \rangle_{\mathcal{H}_0} = \sum_{i=1}^{n} \alpha_i \overline{K(x_i, x)} = \sum_{i=1}^{n} \alpha_i K(x, x_i) = f(x). \tag{2}$$

Let's prove that it defines an inner product on $\mathcal{H}_0$. From the definition of $\langle \cdot\,, \cdot \rangle_{\mathcal{H}_0}$, it is clear that it is linear in its first variable and anti-linear in its second variable. As K is positive definite and conjugate symmetric, $\langle \cdot\,, \cdot \rangle_{\mathcal{H}_0}$ is positive semi-definite and conjugate symmetric.
If $\|f\|_{\mathcal{H}_0} = 0$, from the Cauchy-Schwarz inequality we have

$$\forall x \in \mathcal{X}, \quad |f(x)| = |\langle f, K(\cdot\,, x) \rangle_{\mathcal{H}_0}| \leq \sqrt{\|f\|_{\mathcal{H}_0}} \sqrt{\|K(x, x)\|_{\mathcal{H}_0}} = 0. \tag{3}$$

Thus $f = 0$ and $\langle \cdot, \cdot \rangle_{\mathcal{H}_0}$ is a inner product on $\mathcal{H}_0$.

Let $\overline{\mathcal{H}_0}$ be an abstract completion of $\mathcal{H}_0$. $\overline{\mathcal{H}_0}$ is a Hilbert space containing $\mathcal{H}_0$ as a dense subspace, but the added elements of $\overline{\mathcal{H}_0} \setminus \mathcal{H}_0$ are not necessarily functions! However, equation (2) shows that the evaluation functions on H_0 extend continuously to $\overline{\mathcal{H}_0}$. We can associate to each element of $\overline{\mathcal{H}_0}$ a function over $\mathcal{X}$ this way

$$\forall f \in \overline{\mathcal{H}_0}, \qquad Af := \quad x \in \mathcal{X} \mapsto \langle f, K(\cdot, x) \rangle_{\overline{\mathcal{H}_0}}$$

such that A is the identity on $\mathcal{H}_0$ and the evaluations $(e_x \circ A)_{x \in \mathcal{X}}$ are continuous on $\overline{\mathcal{H}_0}$.
In order to transfer the Hilbert structure of $\overline{\mathcal{H}_0}$ to $\mathcal{H} := A(\overline{\mathcal{H}_0})$, it remains to prove that this association is injective. By linearity of A, we only need to prove that $A(f) = 0$ implies $f = 0$. By density, f is the limit in $\overline{\mathcal{H}_0}$ of a sequence (f_n) of $\mathcal{H}_0$, and in particular (f_n) is a Cauchy sequence in $\mathcal{H}_0$. Moreover, (f_n) converges pointwise to 0 since

$$\forall x \in \mathcal{X}, \qquad f_n(x) = Af_n(x) \underset{n \to \infty}{\longrightarrow} Af(x) = 0$$

Lemma 2. *Let (f_n) be a Cauchy sequence in $\mathcal{H}_0$ converging pointwise to 0. Then (f_n) converges to 0 in the norm sense.*

Proof. Let $\epsilon \geq 0$ and $N(\epsilon)$ such that

$$n \geq N(\epsilon) \Rightarrow \|f_{N(\epsilon)} - f_n\|_{\mathcal{H}_0} \leq \epsilon.$$

There exists $k, \alpha_1, ..., \alpha_k$ and $x_1, ..., x_k$ such that

$$f_{N(\epsilon)} = \sum_{i=1}^{k} \alpha_i K(\cdot, x_i).$$

We have, for $n \geq N(\epsilon)$,

$$\|f_n\|^2_{\mathcal{H}_0} = \langle f_n - f_{N(\epsilon)}, f_n \rangle_{\mathcal{H}_0} + \langle f_{N(\epsilon)}, f_n \rangle_{\mathcal{H}_0} \leq A\epsilon + \sum_{i=1}^{k} \alpha_i f_n(x_i),$$

with A an upper bound for the Cauchy sequence $(\|f_n\|_{\mathcal{H}_0})$. The terms $f_n(x_i)$ tend to 0 by hypothesis, thus (f_n) converges to 0 in the norm sense. □

Now we can define a Hilbert structure on $\mathcal{H}$ by defining the norm:

$$\forall f \in \mathcal{H}, \qquad \|f\|_{\mathcal{H}} := \|A^{-1} f\|_{\overline{\mathcal{H}_0}}$$

Finally, $\mathcal{H}$ has the reproducing property with K

$$f(x) = A(A^{-1} f)(x) = \langle A^{-1} f, K(\cdot, x) \rangle_{\overline{\mathcal{H}_0}} = \langle f, K(\cdot, x) \rangle_{\mathcal{H}}.$$

thus $\mathcal{H}$ is a RKHS with kernel K.

To prove uniqueness, let $\mathcal{G}$ be another Hilbert space of functions for which K is a reproducing kernel. $\mathcal{G}$ contains the functions of $\mathcal{H}_0$ and by construction, $\mathcal{H}_0$ has the only inner product that verifies the reproducing property, so $\mathcal{H}_0 \subset \mathcal{G}$. Moreover, $\mathcal{G}$ is complete and the evaluation functions are continuous over $\mathcal{G}$, so $\mathcal{H} \subset \mathcal{G}$. To prove the converse, let f be an element of $\mathcal{G}$. $\mathcal{H}$ is complete, thus $\mathcal{H}$ is closed in $\mathcal{G}$ and we can decompose $f = f_{\mathcal{H}} + f_{\mathcal{H}^\perp}$, with $f_{\mathcal{H}} \in \mathcal{H}$ and $f_{\mathcal{H}^\perp} \in \mathcal{H}^\perp$. For any x in $\mathcal{X}$, since K is a reproducing kernel of $\mathcal{G}$ and $\mathcal{H}$, we have

$$f(x) = \langle f, K(\cdot, x) \rangle_{\mathcal{G}} = \langle f_{\mathcal{H}}, K(\cdot, x) \rangle_{\mathcal{G}} + \langle f_{\mathcal{H}^\perp}, K(\cdot, x) \rangle_{\mathcal{G}} = \langle f_{\mathcal{H}}, K(\cdot, x) \rangle_{\mathcal{H}} = f_{\mathcal{H}}(x)$$

Thus $f \in \mathcal{H}$ and $\mathcal{H} = \mathcal{G}$. □

The following characterization of a reproducing kernel is very simple and close to lemma 1. It gives an interpretation of the kernel K popular in machine learning as an inner product in a feature space, as well as a practical way of constructing reproducing kernels by defining a mapping from $\mathcal{X}$ to a Hilbert space.

Proposition 1. *A function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{F}$ is a reproducing kernel if and only if there exists a mapping φ from $\mathcal{X}$ to some $\mathbb{F}$-Hilbert space $\mathcal{G}$ such that*

$$\forall (x, y) \in \mathcal{X} \times \mathcal{X}, \qquad K(x, y) = \langle \varphi(y), \varphi(x) \rangle_{\mathcal{G}}.$$

The function φ is called a feature map *and the space $\mathcal{G}$ is called a* feature space.

Proof. If K is a reproducing kernel, there is a canonical feature map from $\mathcal{X}$ to its RKHS $\mathcal{H}$ defined as follow:

$$\forall x \in \mathcal{X}, \qquad \varphi(x) := K(\cdot\,, x) \in \mathcal{H}.$$

Conversely, if K can be expressed as an inner product in a feature space through a feature map, then K is conjugate symmetric and for any $(x_i)_{1 \leq n} \subset \mathcal{X}$ the matrix $(K(x_i, x_j))_{1 \leq i,j \leq n}$ is a Gram matrix. Thus K is definite positive and a reproducing kernel by theorem 2. □

An important consequence of this proposition is that *any* application from $\mathcal{X}$ to *any* Hilbert space defines a reproducing kernel and a RKHS. Since any Hilbert space $\mathcal{H}$ is isometric to some space $\ell^2(\Omega)$ (the space of square-summable families indexed by a set Ω we defined in the first example of RKHS), with Ω a set of cardinal the dimension of $\mathcal{H}$, the notion of reproductive kernel over $\mathcal{X}$ boils down to the notion of mapping from $\mathcal{X}$ to some ℓ^2 space.

Example 2. *With $\mathcal{X} = \mathbb{R}^+$, the application*

$$\begin{aligned} T : \quad \mathbb{R}^+ &\longrightarrow L^2(\mathbb{R}^+, \lambda) \\ x &\mapsto \mathbb{1}_{[0,x]}(\cdot) \end{aligned}$$

defines a reproducing kernel K, and we have

$$K(x, y) = \langle T(y), T(x) \rangle_{L^2(\mathbb{R}^+} = \int_{\mathbb{R}^+} \mathbb{1}_{[0,y]}(t) \mathbb{1}_{[0,x]}(t)\, d\lambda(t) = \min(x, y).$$

This reproducing kernel is linked with the ReLU function widely used as activation function in neural networks:

$$\forall x \in \mathbb{R}, \qquad ReLU(x) = \max(0, x).$$

The following theorem shows that the RKHS $\mathcal{H}$ of a kernel is in some sense the smallest feature space since it is isomorphic to a subspace of any feature space $\mathcal{G}$. It also gives an expression of the norm of $\mathcal{H}$ with a feature map and a feature space; this will be useful in applications of RKHS to optimisation problems.

Theorem 3. *Let $\mathcal{G}$ be a Hilbert space $\varphi : X \to \mathcal{G}$ and K the kernel associated to the feature map φ. Then the RKHS associated to K is:*

$$\mathcal{H} = \{\langle w, \varphi(\cdot) \rangle_{\mathcal{G}} : w \in \mathcal{G}\}$$

equipped with the norm:

$$\|f\|_{\mathcal{H}} = \min\{\|w\|_{\mathcal{G}} : w \in \mathcal{G}, f = \langle w, \varphi(\cdot) \rangle_{\mathcal{G}}\}.$$

There is a proof of this theorem in [7] which proves at the same time the Moore-Aronszajn theorem 2. Since we followed the presentation of [4] and already proved theorem 2, we give an adapted proof.

Proof. Define the linear application

$$\begin{aligned} A : \mathcal{G} &\longrightarrow \mathcal{F}(\mathcal{X}, \mathbb{F}) \\ w &\mapsto \langle w, \varphi(\cdot) \rangle_{\mathcal{G}} \end{aligned}$$

Since for all x, we have

$$\langle \varphi(x), \varphi(x) \rangle_{\mathcal{G}} = K(x, x) = \langle K(\cdot\,, x), K(\cdot\,, x) \rangle_{\mathcal{H}},$$

the restriction of A defines an isomorphism between the following subspaces:

$$\mathcal{G}_0 := \text{span}\{\varphi(x), x \in \mathcal{X}\} \quad \cong \quad \mathcal{H}_0 := (\text{span}_{\mathcal{H}}\{K(\cdot, x), x \in \mathcal{X}\}, \|\cdot\|_{\mathcal{H}}).$$

Being uniformly continuous into a complete space, this restriction has a unique continuous extension $\overline{A}$ to the closure of $\mathcal{G}_0$, which is given for any limit g of a sequence (g_n) of $\mathcal{G}_0$ by

$$\forall x \in \mathcal{X}, \qquad \overline{A}(g)(x) = \lim_{n\to\infty} A(g_n)(x) = \lim_{n\to\infty} \langle g_n, \varphi(x)\rangle_{\mathcal{G}} = \langle g, \varphi(x)\rangle_{\mathcal{G}} = A(g)(x).$$

Thus $\overline{A}$ coincides with A, and A defines an isomorphism of Hilbert spaces between $\overline{\mathcal{G}_0}$ and $\mathcal{H} = \overline{\mathcal{H}_0}$. Clearly, A is null on $(\mathcal{G}_0)^\perp$, so $\text{im}(A) = \mathcal{H}$ and $\ker(A) = (\mathcal{G}_0)^\perp$. For any f in $\mathcal{H}$, there exists a unique w_0 in $\overline{\mathcal{G}_0}$ such that

$$A^{-1}(f) = \{w_0 + w_\perp, w_\perp \in (\mathcal{G}_0)^\perp\}, \qquad \text{and then} \qquad \|f\|_{\mathcal{H}} = \|w_0\|_{\mathcal{G}} = \min_{w\in A^{-1}(f)} \|w\|_{\mathcal{G}}.$$

□

3.3 Basic properties and examples of kernels

In this section, we give some properties of kernels that allow us to construct new kernels from given ones. However, we do not try to describe the new RKHS from the given ones, as it is a bit technical: for this analysis, see [4]. We can then use the characterization of reproducing kernel as either positive definite functions or as scalar product in a feature space through a feature map.

Example 3. *If $\mathcal{X}$ is a subset of $\mathbb{F}^d$, the simplest reproducing kernel is the* linear kernel *defined by*

$$\forall (x, y) \in \mathcal{X}^2, \qquad K(x, y) = \langle x, y\rangle_{\mathbb{F}^d}.$$

The identity function is a feature map for this kernel. This kernel is linear in its first variable and (anti-)linear in its second variable. It will be used to build more complex kernels.

Proposition 2. *(Restriction of kernels) Let K be a reproducing kernel on $\mathcal{X}$, and $\mathcal{X}'$ be a subset of $\mathcal{X}$. Then the restriction $K|_{\mathcal{X}'\times\mathcal{X}'}$ is a RKHS on $\mathcal{X}' \times \mathcal{X}'$.*

Proof. It is an application of theorem 2, since $K|_{\mathcal{X}'\times\mathcal{X}'}$ is positive definite. □

Let us now establish some algebraic properties of the set of reproducing kernels on $\mathcal{X}$.

Proposition 3. *(Sums of kernels) Let K_1 and K_2 be reproducing kernels on $\mathcal{X}$ and $\lambda \in \mathbb{R}^+$. Then λK_1 and $(K_1 + K_2)$ are also reproducing kernels on $\mathcal{X}$.*

Proof. It is an application of theorem 2, since λK_1 and $(K_1 + K_2)$ are positive definite. □

Proposition 4. *(Tensor product of kernels) Let K_1 be a reproducing kernel on $\mathcal{X}_1$ and K_2 be a reproducing kernel on $\mathcal{X}_2$. Then the tensor product $K = K_1 \otimes K_2$ is a reproducing kernel on $\mathcal{X}_1 \times \mathcal{X}_2$.*

Proof. We use feature maps in this proof, but the same can be done using properties of tensor product of matrices to show that any Gram matrix of $K_1 \otimes K_2$ is positive definite.
Let $\varphi_1 : \mathcal{X}_1 \to \mathcal{G}_1$ and $\varphi_2 : \mathcal{X}_2 \to \mathcal{G}_2$ be feature maps of respectively K_1 and K_2. Using the definition of the inner product in the tensor product of Hilbert space $\mathcal{G}_1 \otimes \mathcal{G}_2$ (see appendix A), we get

$$\begin{aligned} K_1(x_1, x_1') \cdot K_2(x_2, x_2') &= \langle \varphi_1(x_1'), \varphi_1(x_1)\rangle_{\mathcal{G}_1} \cdot \langle \varphi_2(x_2'), \varphi_2(x_2)\rangle_{\mathcal{G}_2} \\ &= \langle \varphi_1(x_1') \otimes \varphi_2(x_2'), \varphi_1(x_1) \otimes \varphi_2(x_2)\rangle_{\mathcal{G}_1\otimes\mathcal{G}_2}. \end{aligned}$$

Then $\varphi_1 \otimes \varphi_2 : \mathcal{X}_1 \times \mathcal{X}_2 \to \mathcal{G}_1 \otimes \mathcal{G}_2$ is a feature map of $K_1 \cdot K_2$. □

Proposition 5. *(Product of kernels) Let K_1 and K_2 be two reproducing kernels on $\mathcal{X}$. Then $K := K_1 \cdot K_2$ is a reproducing kernel on $\mathcal{X}$.*

Proof. K is the restriction of the reproducing kernel $K_1 \otimes K_2$ on $\mathcal{X} \times \mathcal{X}$ to the diagonal $\{(x, x), x \in \mathcal{X}\}$. □

Example 4. *We assume that $X \subset \mathbb{F}^d$. With the sum, the multiplication by a non-negative real number, and the product of kernels, we can construct the polynomials kernels. For any polynomial P with non-negative real coefficients, we can define the following kernel*

$$\forall (x,y) \in \mathcal{X}^2, \qquad K(x,x') := P(\langle x,y\rangle_{\mathbb{F}^d})$$

as it is a linear combination of powers of the linear kernel with non-negative real coefficients.

Proposition 6. *(Limit of kernels) Let $(K_n)_n$ be a sequence of kernels that converge pointwise to $K : \mathcal{X}\times\mathcal{X} \to \mathbb{F}$. Then K is a kernel.*

As a consequence, we have the following important property:

Proposition 7. *(Taylor type kernel) Let $r \in \mathbb{R}_+^*$ and $g : B_{\mathbb{C}}(0,r) \to \mathbb{C}$ be a holomorphic function with only non-negative coefficients in its Taylor series:*

$$\forall z \in rB_{\mathbb{C}}, \qquad g(z) = \sum_{n=0}^{\infty} a_n z^n \qquad \text{and} \qquad \forall n \in \mathbb{N}, a_n \geq 0.$$

Then for any x, y in $\mathbb{C}^d$ with norm strictly smaller than $\sqrt{r}$, the function

$$K(x,y) := f(\langle x,y\rangle_{\mathbb{C}^d} = \sum_{n=0}^{\infty} a_n \langle x,y\rangle_{\mathbb{C}^d}^n$$

defines a kernel on $B_{\mathbb{C}^d}(0,\sqrt{r})$ whose restriction to $X := B_{\mathbb{R}^d}(0,\sqrt{r})$ is a real-valued kernel. We say that K is a kernel of Taylor type*.*

Proof. The Cauchy-Schwarz inequality ensures that $\langle x,y\rangle_{\mathbb{F}^d}$ belongs to $B_{\mathbb{F}}(0,1)$, and K is the pointwise limit of the partial sums of its Taylor series, which are polynomial kernels since all the coefficients a_n are non-negative. □

With this property, we can build some more commonly used kernels.

Example 5.

- *If $\mathcal{X} \subset \mathbb{F}^d$, the function*

  $$\forall (x,y)\mathcal{X}^2, \qquad K(x,y) := \exp(\langle x,y\rangle_{\mathbb{F}^d})$$

  is a $\mathbb{F}$-valued kernel on $\mathcal{X}$ called the exponential kernel*.*

- *Let $X := B_{\mathbb{C}}(0,1)$ and $\alpha > 0$. Then the function*

  $$\forall (x,y) \in \mathcal{X}^2, \qquad K(x,y) := (1 - \langle x,y\rangle_{\mathbb{R}^d})^{-\alpha}$$

  defines a kernel on $\mathcal{X}$ called a binomial kernel*.*
  Two important particular cases are $\alpha = 1$ which is the Szegö kernel*, and $\alpha = 2$ which is the* Bergman kernel*.*

3.4 RBF Kernels

Definition 5. *We call* radial basis function *(RBF) kernels the kernels which we can write as $K(x,y) = \varphi(\|x-y\|)$.*

Let E be a subspace of $\mathbb{R}^d$. A function of one variable $f : E \to \mathbb{F}$ is said to be of positive type if the function of two variables defined by

$$\forall (x,y) \in E^2, \qquad (x,y) \mapsto f(x-y)$$

is a function of positive type. This will allows us to define RKHS from a function of a single variable.
In order to define RBF kernels, the following property gives a sufficient condition for a function of one variable to be of positive type:

Proposition 8. *The Fourier transform of a positive function $g : \mathbb{R} \to \mathbb{R}^+$ is a continuous function of positive type.*

A more general theorem due to Bochner states that the converse is also true if we generalize from a positive function g to bounded positive measure μ.

We can now define

Example 6. *For any $\gamma > 0$, the function*

$$\forall (x, y) \in \mathbb{R}^d, \qquad K(x, y) := \exp\left(-\frac{\|x - y\|^2}{\gamma^2}\right)$$

is a $\mathbb{R}^d$-valued reproducing kernel which is called the Gaussian RBF kernel *with width γ. This kernel is used by default in many machine learning libraries such as scikit-learn.*

3.5 Universal kernels

Definition 6. *(Universal kernel) Let $\mathcal{X}$ a compact metric space. A continuous reproducing kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{F}$ is called a* universal kernel *if its RKHS $\mathcal{H}$ is dense in the $\mathbb{F}$-vector space of continuous fonctions $C(\mathcal{X})$ with the uniform norm.*

If K is a universal kernel, it means that for every continuous function g over $\mathcal{X}$ and any $\epsilon > 0$, there exists a (continuous) function f in $\mathcal{H}$ such that $\|f - g\|_\infty < \epsilon$. Therefore, in order to be universal, a reproducing kernal must have a very large RKHS $\mathcal{H}$.

Example 7. *The binomial, exponential and gaussian kernels are universal kernels.*

4 Classification with kernel methods

The introduction of new methods handling high dimensional problems, such as support vector machines have impacted classification algorithms.

4.1 Margin-based performance bounds

As suggested in [6], the most basic way to approach classification is to consider a binary set $\{-1, +1\}$, and a space $\mathcal{X}$ such that $(X, Y) \in \mathcal{X} \times \{-1, +1\}$ is the modeling observation and its corresponding class. We can then consider a class $\mathcal{C}$ of classifiers $g : \mathcal{X} \to \{-1, +1\}$, the probability of error $L(g) = \mathbb{P}(g(X) \neq Y)$ and the empirical error $Ln(g) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{(g(X_i) \neq Y_i)}$. We aim at minimizing the empirical risk over a class of $\mathcal{C}$ using the Vapnik-Chervonenkis inequality (see [6]- section 3 for more details).

This technique has two flaws. The first one lies in the fact that the minimization bound depends on the choice of the dimension of VC. By requiring limitation on the dimension of VC, one imposes limitations on the approximation properties of the class. Second, the minimization problem is NP hard and is difficult to compute. In order to solve these problems, we introduce a *cost function*. We consider classifiers in the form :

$$g_f(x) = \begin{cases} 1 & \text{if } f(x) \geq 0, \\ -1 & \text{otherwise.} \end{cases}$$

where $f : \mathcal{C} \to \mathbb{R}$ is a real-valued function. The probability of error of g is written as

$$L(g_f) = \mathbb{P}(sgn(f(x)) \neq Y) \leq \mathbb{E}\mathbb{1}_{f(X)Y<0}$$

Let $\phi : \mathcal{R} \to \mathcal{R}_+$ be a non negative cost function such that $\phi(x) \geq \mathbb{1}_{x>0}$. We introduce the cost functional and its empirical version by

$$A(f) = \mathbb{E}\phi(-f(X)Y) \qquad \text{and} \qquad A_n(f) = \frac{1}{n}\sum_{1}^{n} \phi(-f(X_i)Y_i).$$

Theorem 4. *Assume that the function f_n is chosen from a class $\mathcal{F}$ based on the data $(Z_1, ..., Z_n) := (X_1, Y_1), ..., (X_n, Y_n)$. Let B denote a uniform upper bound on $\phi(-f(x)y)$ and let L_ϕ be the Lipschitz constant of ϕ. Then the probability of error of the corresponding classifier may be bounded, with probability at least $1 - \delta$, by*

$$L(g_{f_n}) \leq A_n(f_n) + 2L_\phi \mathbb{E}R_n(\mathcal{F}(X_1^n)) + B\sqrt{\frac{2log\frac{1}{\delta}}{n}}$$

Where $R_n(A)$ is the Rademacher average associated with A. Thus, the Rademacher average of the class of real-valued functions f bounds the performance of the classifier.

4.2 Kernel methods

Algorithms working with kernels aim at performing minimization of a cost functional on a ball of the associated reproducing kernel Hilbert space of the form

$$\mathcal{F}_\lambda = \left\{ f(x) = \sum_{j=1}^N c_j k(x_j, x) : N \in \mathbb{N}, \sum_{i=1}^N \sum_{j=1}^N c_i c_j k(x_i, x_j) \leq \lambda^2, x_1, ..., x_N \in \mathcal{X} \right\}$$

One of the main ideas we are going to discuss further is that any linear algorithm based on computing inner products can be extended to a non linear version if we replace the inner product by a kernel function. Thus, the algorithm can encompass a broad set of functions while having relatively low complexity.

As a first illustration, consider the example when γ is a fixed positive parameter and

$$\phi(x) = \begin{cases} 0 & \text{if } x \leq -\gamma \\ 1 & \text{if } x \geq 0 \\ 1 + \frac{x}{\gamma} & \text{otherwise} \end{cases}$$

In this case, $B = 1$ and $L_\phi = \frac{1}{\gamma}$. Notice also that $\mathbb{1}_{x>0} \leq \phi(x) \leq \mathbb{1}_{x>-\gamma}$ and therefore $A_n(f) \leq L_n^\gamma(f)$ where $L_n^\gamma(f)$ is the so-called *margin error* defined by

$$L_n^\gamma(f) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{f(X_i)Y_i<\gamma}.$$

Finally, the *reproducing property* allows to have a precise Rademacher average of $\mathcal{F}_\lambda$ and provides bound that is computationally feasible :

Corollary 1. *Let f_n be any function chosen from the ball $\mathcal{F}_\lambda$. Then, with probability at least $1 - \delta$,*

$$L(g_{f_n}) \leq L_n^\gamma(f_n) + 2\frac{\lambda}{\gamma n}\sqrt{\sum_{i=1}^n k(X_i, X_i)} + \sqrt{\frac{2log\frac{2}{\delta}}{n}}.$$

Proof. Let $\mathbb{E}_\sigma$ be an expectation with respect to the Rademacher variables $(\sigma_1, ..., \sigma_n)$ we have

$$\begin{aligned} R_n(\mathcal{F}_\lambda(X_1^n)) &= \frac{1}{n}\mathbb{E}_\sigma \sup_{\|f\| \leq \lambda} \sum_{i=1}^n \sigma_i f(X_i) \\ &= \frac{1}{n}\mathbb{E}_\sigma \sup_{\|f\| \leq \lambda} \sum_{i=1}^n \sigma_i \langle f, k(X_i, .)\rangle \quad \text{by the reproducing property} \\ &= \frac{\lambda}{n}\mathbb{E}_\sigma \left\| \sum_{i=1}^n \sigma_i k(X_i, .) \right\| \end{aligned}$$

by the Cauchy-Schwarz inequality, where $\|.\|$ denotes the norm in the reproducing kernel Hilbert space. The Kahane-Khinchine inequality states that for any vectors $a_1, ..., a_n$ in a Hilbert space,

$$\frac{1}{\sqrt{2}}\mathbb{E}\left\| \sum_{i=1}^n \sigma_i a_i \right\|^2 \leq \left(\mathbb{E}\left\| \sum_{i=1}^n \sigma_i a_i \right\| \right)^2 \leq \mathbb{E}\left\| \sum_{i=1}^n \sigma_i a_i \right\|^2.$$

It is also easy to see that

$$\mathbb{E}\left\|\sum_{i=1}^{n}\sigma_i a_i\right\|^2 = \mathbb{E}\sum_{i,j=1}^{n}\sigma_i\sigma_j\left\langle a_i, a_j\right\rangle = \sum_{i=1}^{n}\left\|a_i\right\|^2,$$

so we obtain

$$\frac{\lambda}{n\sqrt{2}}\sqrt{\sum_{i=1}^{n}k(X_i,X_i)} \leq R_n(\mathcal{F}_\lambda(X_1^n)) \leq \frac{\lambda}{n}\sqrt{\sum_{i=1}^{n}k(X_i,X_i)}\sqrt{\sum_{i=1}^{n}k(X_i,X_i)}.$$

□

We have considered the case of minimization of a loss function on a ball of the reproducing kernel Hilbert space. However, it is computationally more convenient to formulate the problem as the minimization of a regularized functional of the form

$$\min_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^{n}\phi(-Y_i f(X_i)) + \lambda\left\|f\right\|^2$$

The standard *Support Vector Machine* (SVM) algorithm corresponds to the choice of $\phi_{hinge}(x) = (1+x)_+$. Let's note that this cost functional is the surrogate of the $0/1$ classification loss. Its convexity is particularly relevant since it leads to even better rates of convergence ([6], section 4.2). One motivation of this regularization term is to reduce problems related to overfitting : Complex functions $f \in \mathcal{F}$ which compute a model too closely to the output values of the training data set tend to have large $\mathcal{F}$-norms. The regularization term penalizes such functions more than "simple" functions.

4.3 Linear SVM

Let's now consider a more specific minimization problem that introduces the *generalized portrait algorithm* invented by Vapnik and Lerner [8]. To this end, let's assume that our input space X is a subset of the Euclidian Space $\mathbb{R}^d$. Moreover, we assume that we have a training set $D = ((x_1,y_1),...,(x_n,y_n))$ for which there exists an element $w \in \mathbb{R}^d$ with $\left\|w\right\|_2 = 1$ and a real number $b \in \mathbb{R}$ such that

$$\langle w, x_i\rangle + b > 0, \text{ for all i with } y_i = 1$$
$$\langle w, x_i\rangle + b < 0, \text{ for all i with } y_i = -1$$

In other words, the affine linear hyperplane described by (w,b) separates the training set D into two groups $\{(X_i,Y_i)\in D : y_i = +1\}$ and $\{(X_i,Y_i)\in D : y_i = -1\}$. The *generalized portrait algorithm* constructs a separating hyperplane, described by (w^*,b^*) that are solutions to :

$$\min_{w\in\mathbb{R}^d, b\in\mathbb{R}}\frac{1}{n}\sum_{i=1}^{n}\phi_{hinge}(-Y_i\left\langle w, X_i\right\rangle + b) + \lambda\left\|w\right\|^2$$

This gives rise to a novel classifier $g := sgn(\langle w^*, .\rangle) + b^*)$ called a *Linear Support Vector Machine*. Thus, the classifier g assigns negative labels to one affine half-space defined by the hyperplane (w^*,b^*) and positive labels to the other. This algorithm is geometrically compelling and offers a clear visualisation of classification. However, it has two main issues :

- A linear form of the decision function is not always a good choice for classification, especially if the training set cannot be linearly separated at all.
- Overfitting might still occur.

4.4 Kernel trick and non linear SVM

To resolve the first issue, *Boser et al* [5] introduced an SVM that maps the input data $(X_1, ..., X_n)$ into a *feature space* $\mathcal{G}$ by the *feature map* $\varphi : \mathbb{R}^d \to \mathcal{G}$. Thus the *generalized portrait algorithm* is applied to $((\varphi(X_1), Y_1), ..., (\varphi(X_n), Y_n))$ instead, which gives rise to a new minimization problem:

$$\min_{w \in \mathcal{G}, b \in \mathbb{R}} \frac{1}{n} \sum_{i=1}^{n} \phi_{hinge}(-Y_i \langle w, \varphi(X_i) \rangle_{\mathcal{G}} + b) + \lambda \|w\|_{\mathcal{G}}^2 \tag{P1}$$

As a consequence, the separating hyperplane now lies in a high dimensional space and the second issue of generating overfitted simulations becomes even more important. To address this issue, we can add *slack variables* $\epsilon_i \geq 0$ which allow to replace the hinge loss by a linear constraint.

Lemma 3. *One has*

$$\forall x \in \mathbb{R} : \max(0, 1 - x) = \inf_{\epsilon \in \mathbb{R}_+ : x \geq 1 - \epsilon} \epsilon$$

Let $C = \frac{1}{\lambda n}$. Then (P1) can be rewritten equivalently :

$$\begin{aligned} &\min_{w \in \mathcal{G}, b \in \mathbb{R}, \epsilon \in \mathbb{R}^n} C \sum_{i=1}^{n} \epsilon_i + \|w\|_{\mathcal{G}}^2 \\ &\text{s.t.} \begin{cases} \forall i \in [n],\ \epsilon_i \geq 0 \\ \forall i \in [n],\ Y_i(\langle w, \varphi(X_i) \rangle_{\mathcal{G}} + b) \geq 1 - \epsilon_i \end{cases} \end{aligned} \tag{P2}$$

Let $\mathcal{H}$ be the RKHS associated to the kernel k. We would like to use kernels directly instead of computing the feature map itself. According to *Theorem 3*, we know that for all $w \in \mathcal{G}$, $h = \langle w, \varphi(.) \rangle_{\mathcal{G}} \in \mathcal{H}$ and $\|h\|_{\mathcal{H}} = inf \left\{ \|w'\|_{\mathcal{G}} : w' \in \mathcal{G}, f = \langle w', \varphi(.) \rangle_{\mathcal{G}} \right\}$. Therefore, by joint convexity, (P2) can be written :

$$\begin{aligned} &\min_{h \in \mathcal{H}, \epsilon \in \mathbb{R}^n, b \in \mathbb{R}} C \sum_{i=1}^{n} \epsilon_i + \|h\|_{\mathcal{H}}^2 \\ &\text{s.t.} \begin{cases} \forall i \in [n],\ \epsilon_i \geq 0 \\ \forall i \in [n],\ Y_i(h(X_i) + b) \geq 1 - \epsilon_i \end{cases} \end{aligned} \tag{P3}$$

(P3) reveals that by transforming the data with the feature map φ, a linear SVM can be used to estimate a decision function $g := h + b,\ h \in \mathcal{H},\ b \in \mathbb{R}$ that can be non linear as soon as the kernel k is not the linear kernel. This observation is what we call the "kernel trick". When k is the linear kernel, φ boils down to the identity and $\mathcal{H}$ is the set of linear functions.

Computing SVM in practice is not easy. On one hand, solving (P2) involves computing a feature map φ which is unknown for some kernels or infinite dimensional for some kernels such as the Gaussian RBF kernel. On the other hand, (P3) involves a non parametric optimization variable $h \in \mathcal{H}$.

Luckily, the next theorem provides a way to solve (P3) which consists of restricting h to the form $\sum_{i=1}^{n} \alpha_i k(., X_i),\ \text{for } \alpha \in \mathbb{R}^n$.

Theorem 5. *(Representer theorem) Let $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be a reproducing kernel and $\mathcal{H}$ the associated RKHS. Let also $\psi : \mathbb{R}^+ \to \mathbb{R}$ be a non-decreasing function and $l : \mathbb{R}^{2n} \to \mathbb{R}$ be any loss function. Given a training sample $\{(X_i, Y_i)\}_{1 \geq i \geq n}$ from $(\mathcal{X} \times \mathbb{R})^n$, if the optimization problem*

$$\min_{h \in \mathcal{H}, b \in \mathbb{R}} \psi(\|h\|_{\mathcal{H}}) + l(Y_1, ..., Y_n, h(X_1) + b, ..., h(X_n) + b) \tag{P4}$$

has a solution, then there exists a solution (h^*, b^*) *such that* h^* *has the form*

$$h^* = \sum_{i=1}^{n} \alpha_i k(., X_i).$$

where $\alpha \in \mathbb{R}^n$.

In addition, if ψ *is an increasing function, then all solutions of (P5) can be written in the form described above.*

Proof. Let $(h, b) \in \mathcal{H} \times \mathbb{R}$ be a solution to (P4). Let us consider $V = \operatorname{span}\{k(., X_i), i \in [n]\} \neq \varnothing$ as well as its orthogonal complement $V_\perp$. Let $(h_\parallel, h_\perp) \in V \times V_\perp$ such that $h = h_\parallel + h_\perp$.
Let us remark that:

- by the Pythagorean theorem, $\|h\|_{\mathcal{H}} = \sqrt{\left\|h_\parallel\right\|_{\mathcal{H}}^2 + \|h_\perp\|_{\mathcal{H}}^2} \geq \left\|h_\parallel\right\|_{\mathcal{H}}$
- for all $i \in [\![1, n]\!]$, $h_\perp(X_i) = \langle h_\perp, k(., X_i)\rangle_{\mathcal{H}} = 0$

Consequently,

$$\begin{aligned}
&\psi(\|h\|_{\mathcal{H}}) + l(Y_1, ..., Y_n, h(X_1) + b, ..., h(X_n) + b) \\
&\qquad \geq \psi(\left\|h_\parallel\right\|_{\mathcal{H}}) + l(Y_1, ..., Y_n, h(X_1) + b, ..., h(X_n) + b) \qquad (\psi \text{ non-decreasing and } \|h\|_{\mathcal{H}} \geq \left\|h_\parallel\right\|_{\mathcal{H}}) \\
&\qquad = \psi(\left\|h_\parallel\right\|_{\mathcal{H}}) + l(Y_1, ..., Y_n, h_\parallel(X_1) + b, ..., h_\parallel(X_n) + b) \qquad (h(X_i) = h_\parallel(X_i) + h_\perp(X_i) = h_\parallel(X_i))
\end{aligned}$$

which ensures that $(h_\parallel, b)$ is solution to (P4), with $h_\parallel \in V$, that is $h_\parallel = \sum_{i=1}^n \alpha_i k(., X_i)$ for some $\alpha \in \mathbb{R}^n$. In addition, since (h, b) and $(h_\parallel, b)$ are solutions to (P4), we also have

$$\begin{aligned}
&\psi(\|h\|_{\mathcal{H}}) + l(Y_1, ..., Y_n, h(X_1) + b, ..., h(X_n) + b) \\
&\qquad = \psi(\left\|h_\parallel\right\|_{\mathcal{H}}) + l(Y_1, ..., Y_n, h_\parallel(X_1) + b, ..., h_\parallel(X_n) + b) \\
&\qquad = \psi(\left\|h_\parallel\right\|_{\mathcal{H}}) + l(Y_1, ..., Y_n, h(X_1) + b, ..., h(X_n) + b),
\end{aligned}$$

Which leads to $\psi(\left\|h_\parallel\right\|_{\mathcal{H}}) = \psi(\|h\|_{\mathcal{H}})$.Therefore, as soon as ψ is increasing, one has $\left\|h_\parallel\right\|_{\mathcal{H}}^2 = \|h\|_{\mathcal{H}}^2 = \left\|h_\parallel\right\|_{\mathcal{H}}^2 + \|h_\perp\|_{\mathcal{H}}^2$. So $\|h_\perp\|_{\mathcal{H}} = 0$, i.e. $h_\perp = 0$ and $h = h_\parallel$. □

There's a specific case in which the solution actually exists and is unique: When the loss function is convex and the function ψ has the form $\lambda\|h\|_{\mathcal{H}}$ for some $\lambda > 0$.

Theorem 6. *With the same hypothesis as theorem 5, if the loss function* l *is convex and* $\lambda > 0$*, the problem*

$$\min_{h \in \mathcal{H}, b \in \mathbb{R}} \lambda \|h\|_{\mathcal{H}} + l(Y_1, ..., Y_n, h(X_1) + b, ..., h(X_n) + b) \tag{P5}$$

has a unique solution.

Proof. The evaluation functions $(e_x)_{x \in \mathbb{R}^d}$ are linear and continuous on $\mathcal{H}$, therefore l is a convex continuous function in the variable f. The objective function of this problem is then a strongly convex continuous function over $\mathcal{H}$, in particular it is a continuous coercive function. From the previous representer theorem 5, it is equivalent to look for a solution in the span of $\{k(\cdot, X_i), i \in [\![1, n]\!]\}$. Since this subspace is of finite dimension, there exists a solution to this problem: indeed, if (f_n) is a minimizing sequence of the objective function, it is a bounded sequence of a finite dimension space and any of its adherence point is a minimizer, by continuity.
Now let's show the uniqueness of the solution. Let f and g be minimizers, and let's note J the objective function of this problem. By convexity of l, and using the parallelogram law, we have

$$J(\frac{f+g}{2}) = \lambda \left\| \frac{f+g}{2} \right\|^2 + l(\frac{f+g}{2}) \leq \lambda \left(\frac{1}{2}\|f\|^2 + \frac{1}{2}\|g\|^2 - \left\| \frac{f+g}{2} \right\|^2 \right) + \frac{1}{2}l(f) + \frac{1}{2}l(g) = J(f) - \lambda \left\| \frac{f-g}{2} \right\|^2$$

Since f is a minimizer of J, it implies $\|f - g\| = 0$ and concludes this proof. □

5 Conclusion

Today, RKHS continues to be a vibrant area of research, with ongoing developments in theory and applications. Researchers are exploring new types of kernels, extending the framework of RKHS to new domains, and finding innovative applications in neural networks.

A Completion

A.1 Completion of an inner product space

It is always possible to take the completion of a metric space $(\mathcal{M}, d)$ by taking of the set of all Cauchy sequences of $\mathcal{M}$ endowed with the pseudometric

$$d((x_n),(y_n)) = \lim_{n\to\infty} d(x_n, y_n)$$

and considering the quotient of this set by the equivalence relation "having distance 0" :

$$(x_n) \sim (y_n) \quad \Leftrightarrow \quad d((x_n),(y_n)) = 0.$$

One can show that this new space $\overline{\mathcal{M}}$ is complete. The original space is embedded in $\overline{\mathcal{M}}$ via the identification of an element x of $\mathcal{M}$ with the equivalence class of sequences in $\mathcal{M}$ converging to x. This defines an isometry into a dense subspace, and we can replace this dense subspace by $\mathcal{M}$ in order to have $\mathcal{M} \subset \overline{\mathcal{M}}$.
If if this procedure is applied to an inner product space $\mathcal{H}_0$, the result is a Hilbert space. Indeed, the addition and scalar multiplication on $\mathcal{H}_0$ are compatible with the equivalence relation $\sim$, and we can define an operation on the set of Cauchy sequences of $\mathcal{H}_0$ as follow:

$$\langle (x_n),(y_n)\rangle := \lim_{n\to\infty} \langle x_n, y_n\rangle_{\mathcal{H}_0}$$

This operation is compatible with the equivalence relation $\sim$, and it defines on $\overline{\mathcal{H}_0}$ an inner product compatible with the distance d, thus making $\overline{\mathcal{H}_0}$ a Hilbert space.

A.2 Functional completion

However, if $H_0 \subset \mathcal{F}(\mathcal{X}, \mathbb{F})$ is an inner product space of functions defined on a set $\mathcal{X}$ such that the evaluation functions are continuous on $\mathcal{H}_0$, there is no guarantee that there exists a Hilbert space of functions $\mathcal{H}$ such that $\mathcal{H}_0 \subset \mathcal{H} \subset \mathcal{F}(\mathcal{X}, \mathbb{F})$ and such that the evaluations functions are continuous on $\mathcal{H}$. If that is the case, $\mathcal{H}$ is called the *functional completion* of $\mathcal{H}_0$.

We give an example of such an inner product space on which the evaluation functions are continuous but with no functional completion. As we saw in the proof of theorem 2 while searching for a candidate $\mathcal{H}$, if a Cauchy sequence (f_n) in $\mathcal{H}_0$ converges pointwise to 0, then in the completion $\mathcal{H}$ this sequence must converge and the continuity of the evaluation functions in $\mathcal{H}$ impose the limit $f = 0$. Since $f = 0$ belongs to $\mathcal{H}_0$, (f_n) must converge to 0 in the norm sense in $\mathcal{H}_0$.

We take $\mathcal{X} := B(0,1)$ the open disk of radius 1 in $\mathbb{C}$. For $\mathcal{H}_0$ we take the subspace of square integrable holomorphic functions on $\mathcal{X}$ that admit a limit in 1. We define the following product on $\mathcal{H}_0$:

$$\langle f, g\rangle := \int_{B(0,1)} f(x+iy)\overline{g(x+iy)}dxdy + \lim_{z\to 1} f(z)\overline{g(z)}.$$

With some complex analysis, we can show that the evaluations functions $(e_x)_{x\in\mathcal{X}}$ are continuous. Now if we consider the sequence $(f_n : z \mapsto z^n)_n$ then we see that for any x in $\mathcal{X}$, $e_x(f_n) = x^n \to 0$. But $\|f_n\| \geq 1$ for all $n \in \mathbb{N}$.
For more information on functional completion, see Aronszajn [2].

A.2.1 Bergman kernel

The previous example is close to an important RKHS which is associated to a kernel named the *Bergman kernel*. With the same set $\mathcal{X}$, let $L^2(\mathcal{X})$ be the Hilbert space of square integrable (classes of)-functions on $\mathcal{X}$, and let $\mathcal{H}$ be the subspace consisting of holomorphic functions in $L^2(\mathcal{X})$ (with the canonical injection of functions into classes of functions). With some complex analysis, we can show that $\mathcal{H}$ is a closed hence complete subspace of $L^2(\mathcal{X})$, and that the evaluations functions are continuous on $\mathcal{H}$. Therefore $\mathcal{H}$ is a RKHS. Its reproducing kernel is the Bergman kernel (see Bergman [3]):

$$K(x, y) = \frac{1}{\pi(1 - x\overline{y})^2}.$$

This kernel has a lot of applications in the theory of conformal mapping.

References

[1] Nachman Aronszajn. Theory of reproducing kernels. *Trans . Amer. Math. Soc.*, 68:337–404, 1950.

[2] Nachman Aronszajn and K. T. Smith. Functional spaces and functional completion. *Annales de l'Institut Fourier*, 6:125–185, 1956.

[3] S. Bergman and M. Schiffer. Kernel functions and conformal mapping. *Compositio Mathematica*, 8:205–249, 1951.

[4] Alain Berlinet and Christine Thomas-Agnan. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Kluwer Academic, 2001.

[5] B. E. Boser, I. Guyon, and V. Vapnik. A training algorithm for optimal margin classifiers. *Proceedings of the Fifth Annual ACM Workshop on Computational Learning Theory*, 1992.

[6] Stephane Boucheron, Olivier Bousquet, and Gabor Lugosi. Theory of classification : A survey of some recent advances. *ESAIM: PS*, 9:323–375, Jun 2005.

[7] Ingo Steinwart and Andreas Christmann. *Support Vector Machines*. Springer, 2008.

[8] V. N. Vapnik and A. Lerner. Pattern recognition using generalized portrait method. *Automation and Remote Control*, pages 774–780, 1963.

[9] John von Neumann. Charakterisierung des spektrums eines integraloperators. *Actualités Sc. et Industr.*, 229, 1935.