Functional learning through kernel

Stéphane Canu, Xavier Mary and Alain Rakotomamonjy
PSI
INSA de Rouen,
St Etienne du Rouvray, France
stephane.canu,xavier.mary,alain.rakotomamonjy@insa-rouen.fr

2nd December 2002

Abstract

This paper reviews the functional aspects of statistical learning theory. The main point under consideration is the nature of the hypothesis set when no prior information is available but data. Within this framework we first discuss about the hypothesis set: it is a vectorial space, it is a set of pointwise defined functions, and the evaluation functional on this set is a continuous mapping. Based on these principles an original theory is developed generalizing the notion of reproduction kernel Hilbert space to non hilbertian sets. Then it is shown that the hypothesis set of any learning machine has to be a generalized reproducing set. Therefore, thanks to a general "representer theorem", the solution of the learning problem is still a linear combination of a kernel. Furthermore, a way to design these kernels is given. To illustrate this framework some examples of such reproducing sets and kernels are given.

1 Some questions regarding machine learning

Kernels and in particular Mercer or reproducing kernels play a crucial role in statistical learning theory and functional estimation. But very little is known about the associated hypothesis set, the underlying functional space where learning machines look for the solution. How to choose it? How to build it? What is its relationship with regularization? The machine learning community has been interested in tackling the problem the other way round. For a given learning task, therefore for a given hypothesis set, is there a learning machine capable of learning it? The answer to such a question allows to distinguish between learnable and non-learnable problem. The remaining question is: is there a learning machine capable of learning any learnable set.

We know since [13] that learning is closely related to the approximation theory, to the generalized spline theory, to regularization and, beyond, to the notion of reproducing kernel Hilbert space ($r.k.h.s$). This framework is based on the minimization of the empirical cost plus a stabilizer (*i.e.* a norm is some Hilbert space). Then, under these conditions, the solution to the learning task is a linear combination of some positive kernel whose shape depends on the nature of the stabilizer. This solution is characterized by strong and nice properties such as universal consistency.

But within this framework there remains a gap between theory and practical solutions implemented by practitioners. For instance, in $r.k.h.s$, kernels are positive. Some practitioners use hyperbolic tangent kernel $\tanh(\mathbf{w}^\top \mathbf{x} + w_0)$ while it is not a positive kernel: but it works. Another example is given by practitioners using non-hilbertian framework. The sparsity upholder uses absolute values such as $\int |f| d\mu$ or $\sum_j |\alpha_j|$: these are L^1 norms. They are not hilbertian. Others escape the hilbertian approximation orthodoxy by introducing prior knowledge (*i.e.* a stabilizer) through information type criteria that are not norms.

This paper aims at revealing some underlying hypothesis of the learning task extending the reproducing kernel Hilbert space framework. To do so we begin with reviewing some learning principle. We will stress that the hilbertian nature of the hypothesis set is not necessary while the reproducing property is. This leads us to define a non hilbertian framework for reproducing kernel allowing non positive kernel, non-hilbertian norms and other kinds of stabilizers.

The paper is organized as follows. The first point is to establish the three basic principles of learning. Based on these principles and before entering the non-hilbertian framework, it appears necessary to recall some basic elements of the theory of reproducing kernel Hilbert space and how to build them from non reproducing Hilbert space. Then the construction of non-hilbertian reproducing space is presented by replacing the dot (or inner) product by a more general duality map. This implies distinguishing between two different sets put in duality, one for hypothesis and the other one for measuring. In the hilbertian framework these two sets are merged in a single Hilbert space.

But before going into technical details we think it advisable to review the use of $r.k.h.s$ in the learning machine community.

2 $r.k.h.s$ perspective

2.1 Positive kernels

The interest of $r.k.h.s$ arises from its associated kernel. As it were, a $r.k.h.s$ is a set of functions entirely defined by a kernel function. A Kernel may be characterized as a function from $\mathcal{X} \times \mathcal{X}$ to $\mathbb{R}$ (usually $\mathcal{X} \subseteq R^d$). Mercer [11] first establishes some remarkable properties of a particular class of kernels: positive kernels defining an integral operator. These kernels have to belong to some functional space (typically $L^2(\mathcal{X} \times \mathcal{X})$, the set of square integrable functions on $\mathcal{X} \times \mathcal{X}$) so that the associated integral operator is compact. The positivity of kernel K is defined as follows:

$$K(x, y) \text{ positive } \Leftrightarrow \forall f \in L^2, \ \langle\langle K, f\rangle_{L^2}, f\rangle_{L^2} \geq 0$$

where $\langle ., .\rangle_{L^2}$ denotes the dot product in L^2. Then, because it is compact, the kernel operator admits a countable spectrum and thus the kernel can be decomposed. Based on that, the work by Aronszajn [2] can be presented as follows. Instead of defining the kernel operator from L^2 to L^2 Aronszajn focuses on the $r.k.h.s$ H embeded with its dot product $\langle ., .\rangle_H$. In this framework the kernel has to be a pointwise defined function. The positivity of kernel K is then defined as follows:

$$K(x, y) \text{ positive } \Leftrightarrow \forall g \in H, \ \langle\langle K, g\rangle_H, g\rangle_H \geq 0 \tag{1}$$

Aronszajn first establishes a bijection between kernel and $r.k.h.s$. Then L. Schwartz [16] shows that this was a particular case of a more general situation. The kernel doesn't have to be a genuine function. He generalizes the notion of positive kernels to weakly continuous linear application from the dual set E^* of a vector space E to itself. To share interesting properties the kernel has to be positive in the following sense:

$$K \text{ positive } \Leftrightarrow \forall h \in E^* \ ((K(h), h)_{E,E^*} \geq 0$$

where $(., .)_{E,E^*}$ denotes the duality product between E and its dual set E^*. The positivity is no longer defined in terms of scalar product. But there is still a bijection between positive Schwartz kernels and Hilbert spaces.

Of course this is only a short part of the story. For a detailed review on $r.k.h.s$ and a complete literature survey see [3, 14]. Moreover some authors consider non-positive kernels. A generalization to Banach sets has been introduced [4] within the framework of the approximation theory. Non-positive kernels have been also introduced in Kreĭn spaces as the difference between two positive ones ([1] and [16] section 12).

2.2 $r.k.h.s$ and learning in the literature

The first contribution of $r.k.h.s$ to the statistical learning theory is the regression spline algorithm. For an overview of this method see Wahba's book [20]. In this book two important hypothesis regarding the application of the $r.k.h.s$ theory to statistics are stressed. These are the nature of pointwise defined functions and the continuity of the evaluation functional[1]. An important and general result in this framework is the

[1] These definition are formaly given section 3.5, definition 3.1 and equation (3)

so-called representer theorem [9]. This theorem states that the solution of some class of approximation problems is a linear combination of a kernel evaluated at the training points. But only applications in one or two dimensions are given. This is due to the fact that, in that work, the way to build $r.k.h.s$ was based on some derivative properties. For practical reason only low dimension regressors were considered by this means.
Poggio and Girosi extended the framework to large input dimension by introducing radial functions through regularization operator [13]. They show how to build such a kernel as the green functions of a differential operator defined by its Fourier transform.
Support vector machines (SVM) perform another important link between kernel, sparsity and bounds on the generalization error [19]. This algorithm is based on Mercer's theorem and on the relationship between kernel and dot product. It is based on the ability for positive kernel to be separated and decomposed according to some generating functions. But to use Mercer's theorem the kernel has to define a compact operator. This is the case for instance when it belongs to L^2 functions defined on a compact domain.
Links between green functions, SVM and reproducing kernel Hilbert space were introduced in [8] and [17]. The link between $r.k.h.s$ and bounds on a compact learning domain has been presented in a mathematical way by Cucker and Smale [5].

Another important application of $r.k.h.s$ to learning machines comes from the bayesian learning community. This is due to the fact that, in a probabilistic framework, a positive kernel is seen as a covariance function associated to a gaussian process.

3 Three principles on the nature of the hypothesis set

3.1 The learning problem

A supervised learning problem is defined by a learning domain $\mathcal{X} \subseteq \mathbb{R}^d$ where d denotes the number of explicative variables, the learning codomain $\mathcal{Y} \subseteq \mathbb{R}$ and a n dimensional sample $\{(\mathbf{x}_i, y_i), i = 1, n\}$: the training set.

Main stream formulation of the learning problem considers the loading of a learning machine based on empirical data as the minimization of a given criterion with respect to some hypothesis lying in a hypothesis set $\mathcal{H}$. In this framework hypotheses are functions f from $\mathcal{X}$ to $\mathcal{Y}$ and the hypothesis space $\mathcal{H}$ is a functional space.

Hypothesis H_1 : $\mathcal{H}$ is a functional vector space

Technically a convergence criterion is needed in $\mathcal{H}$, *i.e.* $\mathcal{H}$ has to be embedded with a topology. In the remaining, we will always assumed $\mathcal{H}$ to be a convex topological vector space.

Learning is also the minimization of some criterion. Very often the criterion to be minimized contains two terms. The first one, C, represents the fidelity of the hypothesis with respect to data while Ω, the second one, represents the compression required to make a difference between memorizing and learning. Thus the learning machine solves the following minimization problem:

$$\min_{f \in \mathcal{H}} \; C(f(x_1), ..., f(x_n), \mathbf{y}) + \Omega(f) \tag{2}$$

The fact is, while writing this cost function, we implicitly assume that the value of function f at any point x_i is known. We will now discuss the important consequences this assumption has on the nature of the hypothesis space $\mathcal{H}$.

3.2 The evaluation functional

By writing $f(x_i)$ we are assuming that function f can be evaluated at this point. Furthermore if we want to be able to use our learning machine to make a prediction for a given input x, $f(x)$ has to exist for all $x \in \mathcal{X}$: we want pointwise defined functions. This property is far from being shared by all functions. For instance function $sin(1/t)$ is not defined in 0. Hilbert space L^2 of square integrable functions is a quotient space of functions defined only almost everywhere (*i.e.* not on the singletons $\{x\}, x \in \mathcal{X}$). L^2 functions are not pointwise defined because the L^2 elements are equivalence classes.

To formalize our point of view we need to define $\mathbb{R}^{\mathcal{X}}$ as the set of all pointwise defined functions from $\mathcal{X}$ to $\mathbb{R}$. For instance when $\mathcal{X} = \mathbb{R}$ all finite polynomials (including constant function) belong to $\mathbb{R}^{\mathcal{X}}$. We can lay down our second principle:

Hypothesis H_2 : $\mathcal{H}$ is a set of pointwise defined function (*i.e.* a subset of $\mathbb{R}^{\mathcal{X}}$)

Of course this is not enough to define a hypothesis set properly and at least another fundamental property is required.

3.3 Continuity of the evaluation functional

The pointwise evaluation of the hypothesis function is not enough. We want also the pointwise convergence of the hypothesis. If two functions are closed in some sense we don't want them to disagree on any point. Assume t is our unknown target function to be learned. For a given sample of size n a learning algorithm provides a hypothesis f_n. Assume this hypothesis converges in some sense to the target hypothesis. Actually the reason for hypothesis f_n is that it will be used to predict the value of t at a given x. For any x we want $f_n(x)$ to converge to $t(x)$ as follows:

$$f_n \stackrel{\mathcal{H}}{\longrightarrow} t \Longrightarrow \forall x \in \mathcal{X},\ f_n(x) \stackrel{\mathbb{R}}{\longrightarrow} t(x)$$

We are not interested in global convergence properties but in local convergence properties. Note that it may be rather dangerous to define a learning machine without this property. Usually the topology on $\mathcal{H}$ is defined by a norm. Then the pointwise convergence can be restated as follow:

$$\forall x \in \mathcal{X},\ \exists M_x \in \mathbb{R}^+ \text{ such that } |f(x) - t(x)| \ \leq\ M_x\ ||f - t||_{\mathcal{H}} \tag{3}$$

At any point x, the error can be controlled.

It is interesting to restate this hypothesis with the evaluation functional

Definition 3.1 *the evaluation functional*

$$\begin{array}{rccl} \delta_x : & \mathcal{H} & \longrightarrow & \mathbb{R} \\ & f & \longmapsto & \delta_x f = f(x) \end{array}$$

Applied to the evaluation functional our prerequisite of pointwise convergence is equivalent to its continuity.

Hypothesis H_3 : the evaluation functional is continuous on $\mathcal{H}$

Since the evaluation functional is linear and continuous, it belongs to the topological dual of $\mathcal{H}$. We will see that this is the key point to get the reproducing property.

Note that the continuity of the evaluation functional does not necessarily imply uniform convergence. But in many practical cases it does. To do so one additional hypothesis is needed, the constants M_x have to be bounded: $\sup_{x\in\mathcal{X}} M_x < \infty$. For instance this is the case when the learning domain $\mathcal{X}$ is bounded. Differences between uniform convergence and evaluation functional continuity is a deep and important topic for learning machine but out of the scope of this paper.

3.4 Important consequence

To build a learning machine we do need to choose our hypothesis set as a reproducing space to get the pointwise evaluation property and the continuity of this evaluation functional. But the Hilbertian structure is not necessary. Embedding a set of functions with the property of continuity of the evaluation functional has many interesting consequences. The most useful one in the field of learning machine is the existence of a kernel K, a two-variable function with generation property[2]:

$$\forall f \in \mathcal{H},\ \exists \ell \in \mathbb{N},\ (\alpha_i)_{i=1,\ell} \text{ such that } f(x) \approx \sum_{i=1}^{\ell} \alpha_i K(x, x_i)$$

I being a finite set of indices. Note that for practical reasons f may have a different representation.

If the evaluation set is also a Hilbert space (a vector space embedded with a dot product) it is a reproducing kernel Hilbert space ($r.k.h.s$). Although not necessary, $r.k.h.s$ are widly used for learning because they have a lot of nice practical properties. Before moving on more general reproducing sets, let's review the most important properties of $r.k.h.s$ for learning.

3.5 $\mathbb{R}^{\mathcal{X}}$ the set of the pointwise defined functions on $\mathcal{X}$

In the following, the function space of the pointwise defined functions $\mathbb{R}^{\mathcal{X}} = \{f : \mathcal{X} \to \mathbb{R}\}$ will be seen as a topological vector space embedded with the topology of simple convergence.
$\mathbb{R}^{\mathcal{X}}$ will be put in duality with $\mathbb{R}^{[\mathcal{X}]}$ the set of all functions on $\mathcal{X}$ equal to zero everywhere except on a finite subset $\{x_i, i \in I\}$ of $\mathcal{X}$. Thus all functions belonging to $\mathbb{R}^{[\mathcal{X}]}$ can be written in the following way:

$$g \in \mathbb{R}^{[\mathcal{X}]} \iff \exists \{\alpha_i\}, i = 1, n \text{ such that } g(x) = \sum_i \alpha_i \mathbb{I}_{x_i}(x)$$

were the indicator function $\mathbb{I}_{x_i}(x)$ is null everywhere except on x_i where it is equal to one.

$$\forall x \in \mathcal{X} \quad \mathbb{I}_{x_i}(x) = 0 \text{ if } x \neq x_i \text{ and } \mathbb{I}_{x_i}(x) = 1 \text{ if } x = x_i$$

Note that the indicator function is closely related to the evaluation functional since they are in bijection through:

$$\forall f \in \mathbb{R}^{\mathcal{X}}, \forall x \in \mathcal{X}, \quad \delta_x(f) = \sum_{y \in \mathcal{X}} \mathbb{I}_x(y) f(y) = f(x)$$

But formally, $\left(\mathbb{R}^{\mathcal{X}}\right)' = \text{span}\{\delta_x\}$ is a set of linear forms while $\mathbb{R}^{[\mathcal{X}]}$ is a set of pointwise defined functions.

4 Reproducing Kernel Hilbert Space ($r.k.h.s$)

Definition 4.1 (Hilbert space) *A vector space H embedded with the positive definite dot product $\langle ., . \rangle_H$ is a Hilbert space if it is complete for the induced norm $\|f\|_H^2 = \langle f, f \rangle_H$ (*i.e.* all Cauchy sequences converge in H).*

For instance $\mathbb{R}^n$, $\mathcal{P}_k$ the set of polynomials of order lower or equals to k, L^2, ℓ^2 the set of square sumable sequences seen as functions on $\mathbb{N}$ are Hilbert spaces. L^1 and the set of bounded functions L^∞ are not.

Definition 4.2 (reproduction kernel Hilbert space ($r.k.h.s$)) *A Hilbert space $(\mathcal{H}, \langle ., . \rangle_{\mathcal{H}})$ is a $r.k.h.s$ if it is defined on $\mathbb{R}^{\mathcal{X}}$ (pointwise defined functions) and if the evaluation functional is continuous on H (see the definition of continuity equation 3).*

For instance $\mathbb{R}^n$, $\mathcal{P}_k$ as any finite dimensional set of genuine functions are $r.k.h.s$. ℓ^2 is also a $r.k.h.s$. The Cameron-Martin space defined example 8.1.2 is a $r.k.h.s$ while L^2 is not because it is not a set of pointwise functions.

Definition 4.3 (positive kernel) *A function from $\mathcal{X} \times \mathcal{X}$ to $\mathbb{R}$ is a positive kernel if it is symmetric and if for any finite subset $\{x_i\}, i = 1, n$ of $\mathcal{X}$ and any sequence of scalar $\{\alpha_i\}, i = 1, n$*

$$\sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j K(x_i, y_j) \ \geq\ 0$$

[2]this property means that the set of all finite linear combinations of the kernel is dense in $\mathcal{H}$. See proposition 4.1 for a more precise statement.

This defi nition is equivalent to Aronszajn defi nition of positive kernel given equation (1).

Proposition 4.1 (bijection between $r.k.h.s$ and Kernel) *Corollary of proposition 23 in [16] and theorem 1.1.1 in [20]. There is a bijection between the set of all possible $r.k.h.s$ and the set of all positive kernels.*

Proof.

$\Rightarrow$ from $r.k.h.s$ to Kernel. Let $(\mathcal{H}, \langle ., . \rangle_{\mathcal{H}})$ be a $r.k.h.s.$ By hypothesis the evaluation functional δ_x is a continuous linear form so that it belongs to the topological dual of $\mathcal{H}$. Thanks to the Riesz theorem we know that for each $x \in \mathcal{X}$ there exists a function $K_x(.)$ belonging to $\mathcal{H}$ such that for any function $f(.) \in \mathcal{H}$:

$$\delta_x(f(.)) = \langle K_x(.), f(.) \rangle_{\mathcal{H}}$$

$K_x(.)$ is a function from $\mathcal{X} \times \mathcal{X}$ to $\mathbb{R}$ and thus can be written as a two variable function $K(x, y)$. This function is symmetric and positive since, for any real finite sequence $\{\alpha_i\}, i = 1, \ell$, $\sum_{i=1}^{\ell} \alpha_i K(x, x_i) \in \mathcal{H}$, we have:

$$\begin{aligned} \| \textstyle\sum_{i=1}^{\ell} \alpha_i K(., x_i) \|_{\mathcal{H}}^2 &= \langle \textstyle\sum_{i=1}^{\ell} \alpha_i K(., x_i), \sum_{j=1}^{\ell} \alpha_j K(., x_j) \rangle_{\mathcal{H}} \\ &= \sum_{i=1}^{\ell} \sum_{j=1}^{\ell} \alpha_i \alpha_j K(x_i, x_j) \end{aligned}$$

$\Leftarrow$ from kernel to $r.k.h.s.$ For any couple $(f(.), g(.))$ of $\mathbb{R}^{[\mathcal{X}]}$ (there exist two finite sequences $\{\alpha_i\} i = 1, \ell$ and $\{\beta_j\}, j = 1, m$ and two sequence of $\mathcal{X}$ points $\{x_i\} i = 1, \ell$, $\{y_j\}, j = 1, m$ such that $f(x) = \sum_{i=1}^{\ell} \alpha_{i=1}^{\ell} \mathbb{1}_{x_i}(x)$ and $g(x) = \sum_{j=1}^{m} \beta_j \mathbb{1}_{y_j}(x)$) we define the following bilinear form:

$$\langle f(.), g(.) \rangle_{[\mathcal{X}]} = \sum_{i=1}^{\ell} \sum_{j=1}^{m} \alpha_i \beta_j K(x_i, y_j)$$

Let $\mathcal{H}_0 = \{f \in \mathbb{R}^{[\mathcal{X}]}; | \langle f(.), f(.) \rangle_{[\mathcal{X}]} = 0\}$. $\langle ., . \rangle_{[\mathcal{X}]}$ defines a dot product on the quotient set $\mathbb{R}^{[\mathcal{X}]} / \mathcal{H}_0$. Now let's define $\mathcal{H}$ as the $\mathbb{R}^{[\mathcal{X}]}$ completion for the corresponding norm. $\mathcal{H}$ is a $r.k.h.s$ with kernel K by construction.

Proposition 4.2 (from basis to Kernel) *Let $\mathcal{H}$ be a $r.k.h.s.$ Its kernel K can be written:*

$$K(x, y) = \sum_{i \in I} e_i(x)\ e_i(y)$$

for all orthonormal basis $\{e_i\}_{i \in I}$ of $\mathcal{H}$, I being a set of indices possibly infinite and non-countable.

Proof. $K \in \mathcal{H}$ implies there exits a real sequence $\{\alpha_i\}_{i \in I}$ such that $K(x, .) = \sum_{i \in I} \alpha_i e_i(x)$. Then for all $e_i(x)$ element of the orthonormal basis:

$$\begin{aligned} \langle K(., y), e_i(.) \rangle_{\mathcal{H}} &= e_i(y) && \text{because of } K \text{ reproducing property} \\ \text{and} \quad \langle K(., y), e_i(.) \rangle_{\mathcal{H}} &= \langle \textstyle\sum_{j \in I} \alpha_j e_j(.), e_i(.) \rangle_{\mathcal{H}} \\ &= \textstyle\sum_{j \in I} \alpha_j \langle e_j(.), e_i(.) \rangle_{\mathcal{H}} \\ &= \alpha_i && \text{because } \{e_i\}_{i \in I} \text{ is an orthonormal basis} \end{aligned}$$

by identification we have $\alpha_i = e_i(y)$.

Remark 4.1 *Thanks to this results it is also possible to associate to any positive kernel a basis, possibly uncountable. Consequenty to proposition 4.1 we now how to associate a $r.k.h.s$ to any positive kernel and we get the result because every Hilbert space admit an orthonormal basis.*

The fact that the basis is countable or uncountable (that the corresponding $r.k.h.s$ is separable or not) has no consequences on the nature of the hypothesis set (see example 8.1.7). Thus Mercer kernels are a particlar case of a more general situation since every Mercer kernel is positive in the Aronszajn sense (defi nition 4.3) while the converse is false. Consequenty, when possible functionnal formulation is preferible to kernel formulation of learning algorithm.

5 Kernel and kernel operator

5.1 How to build $r.k.h.s$?

It is possible to build $r.k.h.s$ from a $L^2(G,\mu)$ Hilbert space where G is a set (usualy $G = \mathcal{X}$) and μ a measure. To do so, an operator S is defined to map L^2 functions onto the set of the pointwise valued functions $\mathbb{R}^{\mathcal{X}}$. A general way to define such an operator consists in remarking that the scalar product performs such a linear mapping. Based on that remark this operator is built from a family Γ_x of $L^2(G,\mu)$ functions when $x \in \mathcal{X}$ in the following way:

Definition 5.1 (Carleman operator) *Let $\Gamma = \{\Gamma_x, x \in \mathcal{X}\}$ be a family of $L^2(G,\mu)$ functions. The associated Carleman operator S is*

$$\begin{array}{rccl} S: & L^2 & \longrightarrow & \mathbb{R}^{\mathcal{X}} \\ & f & \longmapsto & g(.) = (Sf)(.) = \langle \Gamma_{(.)}, f\rangle_{L^2} = \displaystyle\int_G \Gamma_{(.)}\, f\, d\mu \end{array}$$

That is to say $\forall x \in \mathcal{X},\ g(x) = \langle \Gamma_x, f\rangle_{L^2}$. To make apparent the bijective restriction of S it is convenient to factorize it as follows:

$$S : L^2 \longrightarrow L^2/\text{Ker}(S) \xrightarrow{T} \text{Im}(S) \xrightarrow{i} \mathbb{R}^{\mathcal{X}} \tag{4}$$

where $L^2/\text{Ker}(S)$ is the quotient set, T the bijective restriction of S and i the cannonical injection.

This class of integral operators is known as Carleman operators [18]. Note that this operator unlike Hilbert-Schmidt operators need not be compact neither bounded. But when G is a compact set or when $\Gamma_x \in L^2(G \times G)$ (it is a square integrable function with respect to both of its variables) S is a Hilbert-Schmidt operator. As an illustration of this property, see the gaussian example on $G = \mathcal{X} = \mathbb{R}$ in table 1. In that case $\Gamma_x(\tau) \notin L^2(\mathcal{X} \times \mathcal{X})$[3].

Proposition 5.1 (bijection between Carleman operators and the set of $r.k.h.s$) - *Proposition 21 in [16] or theorems 1 and 4 in [14]. Let S be a Carleman operator. Its image set $\mathcal{H} = Im(S)$ is a r.k.h.s. If $\mathcal{H}$ is a $r.k.h.s$ there exists a measure μ on some set G and a Carleman operator S on $L^2(G,\mu)$ such that $\mathcal{H} = Im(S)$.*

Proof.

$\Rightarrow$ Consider T the bijective restriction of S defined in equation (4). $\mathcal{H} = \text{Im}(S)$ can be embedded with the induced dot product defined as follows:

$$\begin{array}{rcll} \forall g_1(.), g_2(.) \in \mathcal{H}^2, \quad \langle g_1(.), g_2(.)\rangle_{\mathcal{H}} & = & \langle T^{-1}g_1, T^{-1}g_2\rangle_{L^2} & \\ & = & \langle f_1, f_2\rangle_{L^2} & \text{where } g_1(.) = Tf_1 \text{ and } g_2(.) = Tf_2 \end{array}$$

With respect to the induced norm, T is an isometry. To prove $\mathcal{H}$ is a $r.k.h.s$, we have to check the continuity of the evaluation functional. This works as follows:

$$\begin{array}{rclcl} g(x) & = & (Tf)(x) & & \\ & = & \langle \Gamma_x, f\rangle_{L^2} & \leq & \|\Gamma_x\|_{L^2}\ \|f\|_{L^2} \\ & & & \leq & M_x\ \ \|g(.)\|_{\mathcal{H}} \end{array}$$

with $M_x = \|\Gamma_x\|_{L^2}$. In this framework $\mathcal{H}$ reproducing kernel K verifies $S\Gamma_x = K(x,.)$. It can be built based on Γ:

$$\begin{array}{rcl} K(x,y) & = & \langle K(x,.), K(y,.)\rangle_{\mathcal{H}} \\ & = & \langle \Gamma_x, \Gamma_y\rangle_{L^2} \end{array}$$

$\Leftarrow$ Let $\{e_i\}, i \in I$ be a $L^2(G,\mu)$ orthonormal basis and $\{h_j(.)\}, j \in J$ an orthonormal basis of $\mathcal{H}$. We admit there exists a couple (G,μ) such that card(I) $\geq$ card(J) (take for instance the counting measure on the suitable

[3] To clarify the not so obvious notion of pointwise defined function, whenever possible, we use the notation f when the function is not a pointwise defined function and $f(.)$ denotes $\mathbb{R}^{\mathcal{X}}$ functions. Here $\Gamma_x(\tau)$ is a pointwise defined function with respect to variable x but not with respect to variable τ. Thus, whenever possible, the confusing notation (τ) is omitted.

Name	$\Gamma_x(u)$	$K(x,y)$
Cameron Martin	$\mathbb{I}_{\{x \leq u\}}$	$\min(x,y)$
Polynomial	$e_0(u) + \sum_{i=1}^{d} x_i e_i(u)$	$\mathbf{x}^\top \mathbf{y} + 1$
Gaussian	$1/Z exp^{-\frac{(x-u)^2}{2}}$	$1/Z' exp^{-\frac{(x-y)^2}{4}}$

Table 1: Examples of Carleman operator and their associated reproducing kernel. Note that functions $\{e_i\}_{i=1,d}$ are a finite subfamily of a L^2 orthonormal basis. Z and Z' are two constants.

set). Define $\Gamma_x = \sum_{j \in J} h_j(x) e_j$ as a L^2 family. Let T be the associated Carleman operator. The image of this Carleman operator is the $r.k.h.s$ span by $h_j(.)$ since:

$$
\begin{aligned}
\forall f \in L^2, \quad (Tf)(x) &= \langle \Gamma_x, f \rangle_{L^2} \\
&= \langle \sum_{j \in J} h_j(x) e_j, \sum_{i \in I} \alpha_i e_i \rangle_{L^2} \qquad \text{because } f = \sum_{i \in I} \alpha_i e_i \\
&= \sum_{j \in J} h_j(x) \sum_{i \in I} \alpha_i \langle e_j, e_i \rangle_{L^2} \\
&= \sum_{j \in J} \alpha_j h_j(x)
\end{aligned}
$$

and family $\{h_i(.)\}$ is orthonormal since $h_i(.) = Te_i$.

To put this framework at work the relevant function Γ_x has to be found. Some examples with popular kernels illustrating this definition are shown table 1.

5.2 Carleman operator and the regularization operator

The same kind of operator has been introduced by Poggio and Girosi in the regularization framework [13]. They proposed to define the regularization term $\Omega(f)$ (defined equation 2) by introducing a regularization operator P from hypothesis set $\mathcal{H}$ to L^2 such that $\Omega(f) = \|Pf\|^2_{L^2}$. This framework is very attractive since operator P models the prior knowledge about the solution defining its regularity in terms of derivative or Fourier decomposition properties. Furthermore the authors show that, in their framework, the solution of the learning problem is a linear combination of a kernel (a representer theorem). They also give a methodology to build this kernel as the green function of a differential operator. Following [2] in its introduction the link between green function and $r.k.h.s$ is straightforward when green function is a positive kernel. But a problem arises when operator P is chosen as a derivative operator and the resulting kernel is not derivable (for instance when P is the simple derivation, the associated kernel is the non-derivable function $\min(x,y)$). A way to overcome this technical difficulty is to consider things the other way round by defining the regularization term as the norm of the function in the $r.k.h.s$ built based on Carleman operator T. In this case we have $\Omega(f) = \|f\|_H = \|T^{-1}g\|^2_{L^2}$. Thus since T is bijective we can define operator P as: $P = T^{-1}$. This is no longer a derivative operator but a generalized derivative operator where the derivation is defined as the inverse of the integration (P is defined as T^{-1}).

5.3 Generalization

It is important to notice that the above framework can be generalized to non L^2 Hilbert spaces. A way to see this is to use Kolmogorov's dilation theorem [7]. Furthermore, the notion of reproducing kernel itself can be generalized to non-pointwise defined function by emphasizing the role played by continuity through positive generalized kernels called Schwartz or hilbertian kernels [16]. But this is out of the scope of our work.

6 Reproducing kernel spaces (RKS)

By focusing on the relevant hypothesis for learning we are going to generalize the above framework to non-hilbertian spaces.

6.1 Evaluation spaces

Definition 6.1 (ES)
Let $\mathcal{H}$ be a real topological vector space (t.v.s.) on an arbitrary set $\mathcal{X}$, $\mathcal{H} \subset \mathbb{R}^{\mathcal{X}}$. $\mathcal{H}$ is an evaluation space if and only if:

$$\forall x \in \mathcal{X}, \quad \begin{array}{rcl} \delta_x : \mathcal{H} & \longrightarrow & \mathbb{R} \\ f & \longmapsto & \delta_x(f) = f(x) \end{array} \quad \textit{is continuous}$$

ES are then topological vector spaces in which δ_t (the evaluation functional at t) is continuous, *i.e.* belongs to the topological dual $\mathcal{H}^*$ of $\mathcal{H}$.

Remark 6.1 *Topological vector space $\mathbb{R}^{\mathcal{X}}$ with the topology of simple convergence is by construction an ETS (evaluation topological space).*

In the case of normed vector space, another characterization can be given:

Proposition 6.1 (normed ES or BES)
Let $(\mathcal{H}, \|.\|_{\mathcal{H}})$ be a real normed vector space on an arbitrary set $\mathcal{X}$, $\mathcal{H} \subset \mathbb{R}^{\mathcal{X}}$. $\mathcal{H}$ is an evaluation kernel space if and only if the evaluation functional:

$$\forall x \in \mathcal{X}, \ \exists M_x \in \mathbb{R}, \ \forall f \in \mathcal{H}, \ |f(x)| \leq M_x \|f\|_{\mathcal{H}}$$

if it is complete for the corresponding norme it is a Banach evaluation space (BES).

Remark 6.2 *In the case of a Hilbert space, we can identify $\mathcal{H}^*$ and $\mathcal{H}$ and, thanks to the Riesz theorem, the evaluation functional can be seen as a function belonging to $\mathcal{H}$: it is called the reproducing kernel.*

This is an important point: thanks to the Hilbertian structure the evaluation functional can be seen as a hypothesis function and therefore the solution of the learning problem can be built as a linear combination of this reproducing kernel taken different points. Representer theorem [9] demonstrates this property when the learning machine minimizes a regularized quadratic error criterion. We shall now generalize these properties to the case when no hilbertian structure is available.

6.2 Reproducing kernels

The key point when using Hilbert space is the dot product. When no such bilinear positive functional is available its role can be played by a duality map. Without dot product, the hypothesis set $\mathcal{H}$ is no longer in self duality. We need another set $\mathcal{M}$ to put in duality with $\mathcal{H}$. This second set $\mathcal{M}$ is a set of functions measuring how the information I have at point x_1 helps me to measure the quality of the hypothesis at point x_2. These two sets have to be in relation through a specifi c bilinear form. This relation is called a duality.

Definition 6.2 (Duality between two sets) *Two sets $(\mathcal{H}, \mathcal{M})$ are in duality if there exists a bilinear form $\mathcal{L}$ on $\mathcal{H} \times \mathcal{M}$ that separates $\mathcal{H}$ and $\mathcal{M}$ (see [10] for details on the topological aspect of this definition).*

Let $\mathcal{L}$ be such a bilinear form on $\mathcal{H} \times \mathcal{M}$ that separate them. Then we can defi ne a linear application $\gamma_{\mathcal{H}}$ and its reciprocal $\theta_{\mathcal{H}}$ as follows:

$$\begin{array}{rrcl} \gamma_{\mathcal{H}} : & \mathcal{M} & \longrightarrow & \mathcal{H}^* \\ & f & \longmapsto & \gamma_{\mathcal{H}} f = \mathcal{L}(.,f) \end{array} \qquad \begin{array}{rrcl} \theta_{\mathcal{H}} : & \mathrm{Im}\,(\gamma_{\mathcal{H}}) & \longrightarrow & \mathcal{M} \\ & g = \mathcal{L}(.,f) & \longmapsto & \theta_{\mathcal{H}} g = f \end{array}$$

where $\mathcal{H}^*$ (resp. $\mathcal{M}^*$) denotes the dual set of $\mathcal{H}$ (resp. $\mathcal{M}$).

Let's take an important example of such a duality.

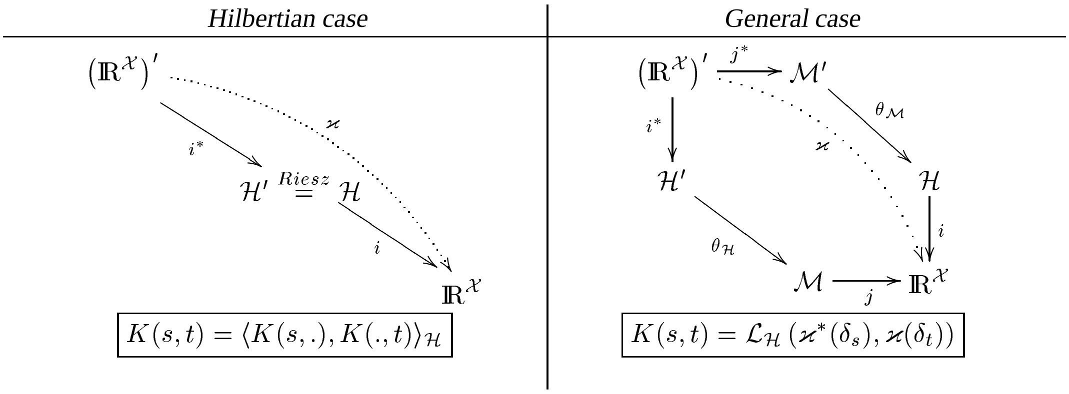

Figure 1: illustration of the subduality map.

Proposition 6.2 (duality of pointwise defined functions) *Let $\mathcal{X}$ be any set (not necessarily compact). $\mathbb{R}^{\mathcal{X}}$ and $\mathbb{R}^{[\mathcal{X}]}$ are in duality*

Proof. Let's define the bilinear application $\mathcal{L}$ as follows:

$$\begin{array}{rccl} \mathcal{L}: & \mathbb{R}^{\mathcal{X}} \times \mathbb{R}^{[\mathcal{X}]} & \longrightarrow & \mathbb{R} \\ & \left(f(.), g(.) = \sum_{i\in I} \alpha_i \mathbb{I}_{x_i}(.)\right) & \longmapsto & \sum_{i\in I} \alpha_i f(x_i) = \sum_{x\in\mathcal{X}} f(x)g(x) \end{array}$$

Another example is shown in the two following functional spaces:

$$L^1 = \left\{ f \;\Big|\; \int_{\mathcal{X}} |f|\, d\mu < \infty \right\} \qquad \text{and} \qquad L^\infty = \left\{ f \;\Big|\; \operatorname*{ess\,sup}_{x\in\mathcal{X}} |f| < \infty \right\}$$

where for instance μ denotes the Lebesgue measure. Theses two spaces are put in duality through the following duality map:

$$\begin{array}{rccl} \mathcal{L}: & L^1 \times L^\infty & \longrightarrow & \mathbb{R} \\ & f, g & \longmapsto & \mathcal{L}(f,g) = \int_{\mathcal{X}} f\ g\ d\mu \end{array}$$

Definition 6.3 (Evaluation subduality) *Two sets $\mathcal{H}$ and $\mathcal{M}$ form an evaluation subduality iff:*

- *they are in duality through their duality map $\gamma_{\mathcal{H}}$,*
- *they both are subsets of $\mathbb{R}^{\mathcal{X}}$*
- *the continuity of the evaluation functional is preserved through:*

$$Span(\delta_x) = \gamma_{\mathbb{R}^{\mathcal{X}}}\left(\left(\mathbb{R}^{\mathcal{X}}\right)'\right) \subseteq \gamma_{\mathcal{H}}(\mathcal{M}) \quad \textit{and} \quad \gamma_{\mathbb{R}^{\mathcal{X}}}\left(\left(\mathbb{R}^{\mathcal{X}}\right)'\right) \subseteq \theta_{\mathcal{H}}(\mathcal{H})$$

The key point is the way of preserving the continuity. Here the strategy to do so is fi rst to consider two sets in duality and then to build the (weak) topology such that the dual elements are (weakly) continuous.

Proposition 6.3 (Subduality kernel) *A unique weakly continuous linear application $\varkappa$ is associated to each subduality. This linear application, called the subduality kernel, is defined as follows:*

$$\begin{array}{rccl} \varkappa: & \left(\mathbb{R}^{\mathcal{X}}\right)' & \longrightarrow & \mathbb{R}^{\mathcal{X}} \\ & \sum_{i\in I} \delta_{x_i} & \longmapsto & i \circ \theta_{\mathcal{M}} \circ j^*(\sum_{i\in I} \delta_{x_i}) \end{array}$$

where i and j^* are the canonical injections from $\mathcal{H}$ to $\mathbb{R}^{\mathcal{X}}$ and respectively from $\left(\mathbb{R}^{\mathcal{X}}\right)'$ to $\mathcal{M}'$ (figure 1).

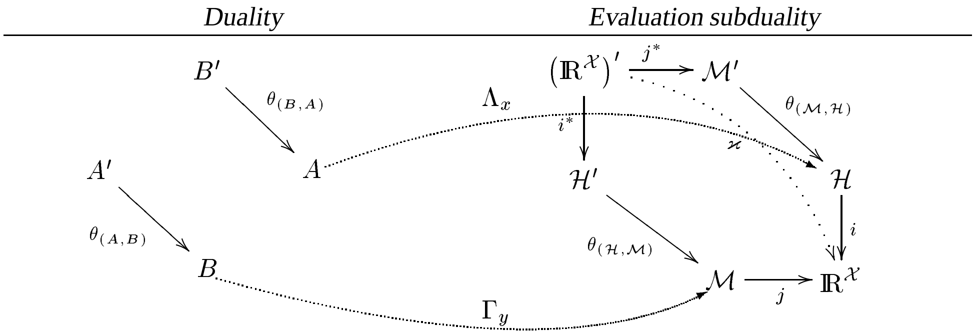

Figure 2: illustration of the building operators for reproducing kernel subduality from a duality (A, B).

Proof. for details see [10].
We can illustrate this mapping detailing all performed applications as in figure 1:

$$\begin{array}{ccccccccc} \left(\mathbb{R}^{\mathcal{X}}\right)' & \overset{\text{see 3.5}}{\longrightarrow} & \mathbb{R}^{[\mathcal{X}]} & \overset{j^*}{\longrightarrow} & \mathcal{M}' & \overset{\theta_{\mathcal{M}}}{\longrightarrow} & \mathcal{H} & \overset{i}{\longrightarrow} & \mathbb{R}^{\mathcal{X}} \\ \delta_x & \longmapsto & \mathbb{I}_{\{x\}} & \longmapsto & \mathcal{L}(K_x, .) & \longmapsto & K_x(.) & \longmapsto & K(x, .) \end{array}$$

Definition 6.4 (Reproducing kernel of an evaluation subduality) *Let $(\mathcal{H}, \mathcal{M})$ be an evaluation subduality with respect to map $\mathcal{L}_{\mathcal{H}}$ associated with subduality kernel $\varkappa$. The reproducing kernel associated with this evaluation subduality is the function of two variables defined as follows:*

$$\begin{array}{rccl} K: & \mathcal{X} \times \mathcal{X} & \longrightarrow & \mathbb{R} \\ & (x, y) & \longmapsto & K(x, y) = \mathcal{L}_{\mathcal{H}}\left(\varkappa^*(\delta_y), \varkappa(\delta_x)\right) \end{array}$$

This structure is illustrated in figure 1. Note that this kernel no longer needs to be definite positive. If the kernel is definite positive it is associated with a unique $r.k.h.s$. However, as shown in example 8.2.1 it can also be associated with evaluation subdualities. A way of looking at things is to define $\varkappa$ as the generalization of the Schwartz kernel while K is the generalization of the Aronszajn kernel to non hilbertian structures. Based on these definitions the important expression property is preserved.

Proposition 6.4 (generation property) $\forall f \in \mathcal{H},\ \exists (\alpha_i)_{i \in I}$ *such that* $f(x) \approx \sum_{i \in I} \alpha_i K(x, x_i)$ *and* $\forall g \in \mathcal{M},\ \exists (\alpha_i)_{i \in I}$ *such that* $g(x) \approx \sum_{i \in I} \alpha_i K(x_i, x)$

Proof. This property is due to the density of Span$\{K(., x), x \in \mathcal{X}\}$ in $\mathcal{H}$. For more details see [10] Lemma 4.3.

Just like $r.k.h.s$, another important point is the possibility to build an evaluation subduality, and of course its kernel, starting from any duality.

Proposition 6.5 (building evaluation subdualities) *Let (A, B) be a duality with respect to map $\mathcal{L}_A$. Let $\{\Gamma_x, x \in \mathcal{X}\}$ be a total family in A and $\{\Lambda_x, x \in \mathcal{X}\}$ be a total family in B. Let S (reps. T) be the linear mapping from A (reps. B) to $\mathbb{R}^{\mathcal{X}}$ associated with Γ_x (reps. Λ_x) as follows:*

$$\begin{array}{rccl} S: & A & \longrightarrow & \mathbb{R}^{\mathcal{X}} \\ & g & \longmapsto & Sg(x) = \mathcal{L}_A\left(g, \Lambda_x\right) \end{array} \qquad \begin{array}{rccl} T: & B & \longrightarrow & \mathbb{R}^{\mathcal{X}} \\ & f & \longmapsto & Tf(x) = \mathcal{L}_A\left(\Gamma_x, f\right) \end{array}$$

Then S and T are injective and $(S(A), T(B))$ is an evaluation subduality with the reproducing kernel K defined by:

$$K(x, y) = \mathcal{L}_A(\Gamma_x, \Lambda_y)$$

Proof. see [10] Lemma 4.5 and proposition 4.6

An example of such subduality is obtained by mapping the (L^1, L^∞) duality to $\mathbb{R}^{\mathcal{X}}$ using *injective* operators defined by the families $\Gamma_x(\tau) = \mathbb{I}_{\{x<\tau\}}$ and $\Lambda_y(\tau) = \mathbb{I}_{\{y<\tau\}}$:

$$\begin{array}{rccl} T: & L^1 & \longrightarrow & \mathbb{R}^{\mathcal{X}} \\ & f & \longmapsto & Tf(x) = (\Gamma_x, f)_{L^\infty, L^1} = \int \mathbb{I}_{\{x<\tau\}} f(\tau)\, d\tau \end{array}$$

and

$$\begin{array}{llll} S: & L^\infty & \longrightarrow & \mathbb{R}^{\mathcal{X}} \\ & g & \longmapsto & Sg(y) = (g, \Lambda_y)_{L^\infty, L^1} = \int g(\tau) \mathbb{I}_{\{y<\tau\}} \, d\tau \end{array}$$

In this case $\mathcal{H} = Im(T)$, $\mathcal{M} = Im(S)$ and $K(y,x) = \displaystyle\int \Lambda(y,\tau)\Gamma(x,\tau)\, d\tau = \min(x,y)$. We define the duality map between $\mathcal{H}$ and $\mathcal{M}$ through:

$$\mathcal{L}_{\mathcal{X}}(g_1, g_2) = \mathcal{L}_{\mathcal{X}}(Sf_1, Tf_2) = \mathcal{L}(f_1, f_2)$$

See example 8.2.1 for details.

All useful properties of $r.k.h.s$ – pointwise evaluation, continuity of the evaluation functional, representation and building technique – are preserved. A missing dot product has no consequence on this functional aspect of the learning problem.

7 Representer theorem

Another issue is of paramount practical importance: determining the shape of the solution. To this end representer theorem states that, when $\mathcal{H}$ is a $r.k.h.s$, the solution of the minimization of the regularized cost defined equation (2) is a linear combination of the reproducing kernel evaluated at the training examples [9, 15]. When hypothesis set $\mathcal{H}$ is a reproducing space associated with a subduality we have the same kind of result. The solution lies in a finite n-dimensional subspace of $\mathcal{H}$. But we don't know yet how to systematically build a convenient generating family in this subspace.

Theorem 7.1 (representer) *Assume $(\mathcal{H}, \mathcal{M})$ is a subduality of $\mathbb{R}^{\mathcal{X}}$ with kernel $K(x,y)$. Assume the stabilizer Ω is convex and differentiable (∂_Ω denotes its subdifferential set).*
If $\partial_\Omega(\sum \alpha_i K(x_i, x)) \subseteq \{\sum \beta_i \delta_{x_i}\} \in \mathcal{H}^*$ then the solution of cost minimization lies in a n-dimensional subspace of $\mathcal{H}$.

Proof. Define a $\mathcal{M}$ subset $M_1 = \left\{\sum_{i=1}^n \alpha_i K(x_i, .)\right\}$. Let $H_2 \subset \mathcal{H}$ be the M_1 orthogonal in the sense of the duality map (*i.e.* $\forall f \in H_2, \forall g \in M_1\ \mathcal{L}(f,g) = 0$). Then for all $f \in H_2, f(x_i) = 0, i = 1, n$. Now let H_1 be the complement vector space defined such that

$$\mathcal{H} = H_1 \oplus H_2 \quad \Leftrightarrow \forall f \in \mathcal{H}\ \exists f_1 \in H_1 \text{ and } f_2 \in H_2 \quad \text{such that } f = f_1 + f_2$$

The solution of the minimizing problem lies in H_1 since:

- $\forall f_2 \in H_2, C(f_2) = \text{constant}$
- $\Omega(f_1 + f_2) \geq \Omega(f_1) + (\partial_\Omega(f_1), f_2)_{\mathcal{M},\mathcal{H}}$ (thanks to the convexity of Ω)
- and $\forall f_2 \in H_2, ; (\partial_\Omega(f_1), f_2)_{\mathcal{M},\mathcal{H}} = 0$ by hypothesis

By construction H_1 a n-dimensional subspace of $\mathcal{H}$.

The nature of vector space H_1 depends on kernel K and on regularizer Ω. In some cases it is possible to be more precise and retrieve the nature of H_1. Let's assume regularizer $\Omega(f)$ is given. $\mathcal{H}$ may be chosen as the set of function such that $\Omega(f) < \infty$. Then, if it is possible to build a subduality $(\mathcal{H}, \mathcal{M})$ with kernel K such that

$$E = \underbrace{\text{Vect}\{K(x_i, .)\}}_{H_1} \oplus \underbrace{(\ \text{Vect}\{K(., x_i)\})^\top}_{M_1^\top}$$

and if the vector space spaned by the kernel belongs to the regularizer subdifferential $\partial\Omega(f)$:

$$\forall f \in H_1, \quad \exists g \in M_1 \text{ such that } g \in \partial\Omega(f)$$

then solution f^* of the minimization of the regularized empirical cost is a linear combination of the kernel:

$$f^*(x) = \sum_{i=1}^n \alpha_i K(x_i, x)$$

An example of such result is given with the following regularizer based on the p-norm on $G = [0,1]$:

$$\Omega(f) = \int_0^1 (f')^p \, d\mu$$

The hypothesis set is Sobolev space H^p (the set of functions defined on $[0,1]$ whose generalized derivative is p-integrable) put in duality with H^q (with $1/p + 1/q = 1$) through the following duality map:

$$\mathcal{L}(f,g) = \int_0^1 f'g' \, d\mu$$

The associated kernel is just like in Cameron Martin case $K(x,y) = \min(x,y)$. Some tedious derivations lead to:

$$\forall h \in \mathcal{H} \quad \mathcal{L}(h, \partial\Omega(f)) = \int_0^1 h' \, p(f')^{p-1} \, d\mu$$

Thus the kernel verifies $p(K(.,y)')^{p-1} \propto K(x,.)$

This question of the representer theorem is far from being closed. We are still looking for a way to derive a generating family from the kernel and the regularizer. To go more deeply into general and constructive results, a possible way to investigate is to go through Ω Fenchel dual.

8 Examples

8.1 Examples in Hilbert space

The examples in this section all deal with r.k.h.s included in a L^2 space.

1. Schmidt ellipsoid:
   Let $(\mathcal{X},\mu)$ be a measure space, $\{e_i, i \in I\}$ a basis of $L^2(\mathcal{X},\mu)$ I being a countable set of indices. Any sequence $\{\alpha_i, i \in I, \quad \sum_{i\in I} \alpha_i^2 < +\infty\}$ defines a Hilbert-Schmidt operator on $L^2(\mathcal{X},\mu)$ with kernel function $\Gamma(x,y) = \sum_{i\in I} \alpha_i e_i(x) e_i(y)$, thus a reproducing kernel Hilbert space with kernel function:

   $$\forall (x,y) \in \mathcal{X}^2, \quad K(x,y) = \sum_{i\in I} \alpha_i^2 e_i(x) e_i(y)$$

   The closed unit ball $\mathfrak{B}_H$ of the $r.k.h.s$ verifies

   $$\mathfrak{B}_H = T(\mathfrak{B}_{L^2}) = \left\{ f \in L^2, f = \sum_{i\in I} f_i e_i, \quad \sum_{i\in I} \left(\frac{f_i}{\alpha_i}\right)^2 \leq 1 \right\}$$

   and is then a Schmidt ellipsoid in L^2. An interesting discussion about Schmidt ellipsoids and their applications to sample continuity of Gaussian measures may be found in [6].

2. Cameron-Martin space:
   Let T be the Carleman integral operator on $L^2([0,1]\mu)$ (μ is the Lebesgue measure) with kernel function

   $$\Gamma(x,y) = Y(x-y) = \mathbb{1}_{\{y \leq x\}}$$

   it defines a r.k.h.s with reproducing kernel $K(x,y) = \min(x,y)$. The space $(H; \langle .,. \rangle_H)$ is the Sobolev space of degree 1, also called the Cameron-Martin space.

   $$\begin{cases} H = \{f \text{ absolutely continuous}, \exists f' \in L^2([0,1]),\ f(x) = \int_0^x f' d\mu\} \\ \langle f,g \rangle_H = \langle f', g' \rangle_{L^2} \end{cases}$$

3. A Carleman but non Hilbert-Schmidt operator:
Let T be the integral operator on $L^2(\mathbb{R},\mu)$ (μ is the Lebesgue measure) with kernel function
$$\Gamma(x,y) = \exp^{-\frac{1}{2}(x-y)^2}$$
It is a Carleman integral operator, thus we can define a r.k.h.s $(H;\langle .,.\rangle_H) = Im(T)$, but T is not a Hilbert-Schmidt operator. H reproducing kernel is:
$$K(x,y) = \frac{1}{Z}\exp^{-\frac{1}{4}(x-y)^2}$$
where Z is a suitable constant.

4. Continuous kernel:
This example is based on theorem 3.11 in [12]. Let $\mathcal{X}$ be a compact subspace of $\mathbb{R}$, $K(.,.)$ a continuous symmetric positive definite kernel. It defines a $r.k.h.s$ $(H;\langle .,.\rangle_H)$ and any Radon measure μ of full support is kernel-injective. Then, for any such μ, there exists a Carleman operator T on $L^2(\mathcal{X},\mu)$ such that $(H;\langle .,.\rangle_H) = Im(T)$.

5. Hilbert space of constants:
Let $(H;\langle .,.\rangle_H)$ be the Hilbert space of constant functions on $\mathbb{R}$ with scalar product $\langle f,g\rangle_H = f(0)g(0)$. It is obviously a $r.k.h.s$ with reproducing kernel $K(.,.)\equiv 1$. For any probability measure μ on $\mathbb{R}$ let:
$$\forall f \in L^2(\mathbb{R},\mu), \quad Tf = \int_{\mathbb{R}} f(s)\mu(ds)$$
Then $H = T(L^2(\mathbb{R},\mu))$ and $\forall f,g \in H,\ \langle f,g\rangle_H = \langle f,g\rangle_{L^2}$.

6. A non-separable $r.k.h.s$ - the L^2 space of almost surely null functions:
Define the positive definite kernel function on $\mathcal{X}\subset\mathbb{R}$ by $\forall s,t\in\mathcal{X},\ K(s,t) = \mathbb{I}_{\{s=t\}}$. It defines a $r.k.h.s$ $(H;\langle .,.\rangle_H)$ and its functions are null except on a countable set. Define a measure μ on $(\mathcal{X},\mathcal{B})$ where $\mathcal{B}$ is the Borel σ-algebra on $\mathcal{X}$ by $\mu(t) = 1\ \forall t\in\mathcal{X}$. μ verifies: $\mu(\{t_1,\cdots,t_n\}) = n$ and $\mu(A) = +\infty$ for any non-finite $A\in\mathcal{B}$. The kernel function is then square integrable and H is injectively included in $L^2(\mathcal{X},\mathcal{B},\mu)$. Moreover, $K(s,t) = \int_{\mathcal{X}} K(t,u)K(u,s)d\mu(u)$ with K Carleman integrable and $T = Id_{L^2}$ (note that the identity is a non-compact Carleman integral operator). Finally, $(H;\langle .,.\rangle_H) = L^2(\mathcal{X},\mathcal{B},\mu)$.

7. Separable $r.k.h.s$:
Let H be a separable $r.k.h.s$. It is well known that any separable Hilbert space is isomorphic to ℓ^2. Then there exists T kernel operator $Im(T) = H$. It is easy to construct effectively such a T: let $\{h_n(.),\ n\in\mathbb{N}\}$ be an orthonormal basis of H and define T kernel operator on ℓ^2 with kernel $\Gamma_x \to \{h_n(x),\ n\in\mathbb{N}\}(\in l^2)$. Then $Im(T) = H$.

8.2 Other examples

Applications to non-hilbertian spaces are also feasible:

1. (L^1,L^∞) - “Cameron-Martin" evaluation subduality:
Let T be the kernel operator on $L^1([0,1]\mu)$ (μ is the Lebesgue measure) with kernel function
$$\Gamma(t,s) = Y(t-s) = \mathbb{I}_{\{s\le t\}}, \quad \Gamma(t,.)\in L^\infty$$
it defines an evaluation duality $(H_1;H_\infty)$ with reproducing kernel
$$\forall (s,t)\in\mathcal{X}^2, \quad K(s,t) = \min(s,t)$$
$$\begin{cases} H_1 = \{f \text{ absolutely continuous}, \exists f'\in L^1([0,1]),\ f(t) = \int_0^t f'(s)ds\} \\ \|f\|_{H_1} = \|f'\|_{L^1} \end{cases}$$
and
$$\begin{cases} H_\infty = \{f \text{ absolutely continuous}, \exists f'\in L^\infty([0,1]),\ f(t) = \int_0^t f'(s)ds\} \\ \|f\|_{H_\infty} = \|f'\|_{L^\infty} \end{cases}$$

2. $\left(\mathbb{R}^{\mathcal{X}}, \mathbb{R}^{[\mathcal{X}]}\right)$:
   We have seen that $\mathbb{R}^{\mathcal{X}}$ endowed with the topology of simple convergence is an ETS. However, $\mathbb{R}^{\mathcal{X}}$ endowed with the topology of almost sure convergence is never an ETS unless every singleton of $\mathcal{X}$ has strictly positive measure.

9 Conclusion

It is always possible to learn without kernel. But even if it is not visible, one is hidden somewhere! We have shown, from some basic principles (we want to be able to compute the value of a hypothesis at any point and we want the evaluation functional to be continuous), how to derive a framework generalizing $r.k.h.s$ to non-hilbertian spaces. In our reproducing kernel dualities, all $r.k.h.s$ nice properties are preserved except the dot product replaced by a duality map. Based on the generalization of the hilbertian case, it is possible to build associated kernels thanks to simple operators. The construction of evaluation subdualities without Hilbert structure is easy within this framework (and rather new). The derivation of evaluation subdualities from any kernel operator has many practical outcome. First, such operators on separable Hilbert spaces can be represented by matrices, and we can build any separable $r.k.h.s$ from well-known ℓ^2 structures (like wavelets in a L^2 space for instance). Furthermore, the set of kernel operators is a vector space whereas the set of evaluation subdualities is not (the set of $r.k.h.s$ is for instance a convex cone), hence practical combination of such operators are feasible. On the other hand, from the bayesian point of view, this result may have many theoretical and practical implications in the theory of Gaussian or Laplacian measures and abstract Wiener spaces.

Unfortunately, even if some work has been done, a general representer theorem is not available yet. We are looking for an automatic mechanism designing the *shape* of the solution of the learning problem in the following way:

$$f(\mathbf{x}) = \sum_{i=1}^{m} \alpha_i K(\mathbf{x}_i, \mathbf{x}) + \sum_{j=1}^{k} \beta_j \varphi_j(\mathbf{x})$$

where Kernel K, number of component m and functions $\varphi_k(\mathbf{x}), j = 1, k$ are derivated from regularizer Ω. The remaining questions being: how to learn the coeffi cients and how to determine cost function?

Acknowledgements

Part of this work has been realized while the authors were visiting B. Schölkopf in Tuebingen. The section dedicated to the representer theorem benefi ts from O. Bousquet ideas. This work also benefi ts from comments and discussion with NATO ASI on Learning Theory and Practice students in Leuven.

References

[1] Daniel Aron Alpay. Some krein spaces of analytic functions and an inverse scattering problem. *Michigan Journal of Mathematics*, 34:349–359, 1987.

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

[3] Marc Attéia. *Hilbertian kernels and spline functions*. North-Holland, 1992.

[4] Marc Attéia and Jean Audounet. Inf-compact potentials and banachic kernels. In *Banach space theory and its applications*, volume 991 of *Lecture notes in mathematics*, pages 7–27. Springer-Verlag, 1981.

[5] Felipe Cucker and Steve Smale. On the mathematical foundations of learning. *Bulletin of the American Mathematical society*, pages 1–49, 2001.

[6] R. M. Dudley. *Uniform central limit theorems*. Cambridge university press, 1999.

[7] D. Evans and J.T. Lewis. Dilations of irreversible evolutions in algebraic quantum theory. *Communications of the Dublin Institute for advanced Studies. Series A.*, 24, 1977.

[8] Federico Girosi. An equivalence between sparse approximation and support vector machines. *Neural Computation*, 10(6):1455–1480, 1998.

[9] G. Kimeldorf and G. Wahba. Some results on Tchebycheffi an spline functions. *J. Math. Anal. Applic.*, 33:82–95, 1971.

[10] Xavier Mary, Denis De Brucq, and Stéphane Canu. Sous-dualités et noyaux (reproduisants) associés. Technical report PSI 02-006.

[11] J. Mercer. Functions of positive and negative type and their connection with the theory of integral equations. *Transactions of the London Philosophical Society A*, 209:415–446, 1909.

[12] Jacques Neveu. *Processus aléatoires gaussiens*. Séminaires de mathématiques supérieures. Les presses de l'université de Montréal, 1968.

[13] Tomaso Poggio and Federico Girosi. A theory of networks for approximation and learning. Technical Report AIM-1140, 1989.

[14] Saburou Saitoh. *Theory of reproducing kernels and its applications*, volume 189. Longman scientifi c and technical, 1988.

[15] B. Scholkopf. A generalized representer theorem. Technical Report 2000-81, NeuroColt2 Technical Report Series, 2000.

[16] Laurent Schwartz. Sous espaces hilbertiens d'espaces vectoriels topologiques et noyaux associés. *Journal d'Analyse Mathématique*, pages 115–256, 1964.

[17] Alex J. Smola and Bernhard Schölkopf. From regularization operators to support vector kernels. In Michael I. Jordan, Michael J. Kearns, and Sara A. Solla, editors, *Advances in Neural Information Processing Systems*, volume 10. The MIT Press, 1998.

[18] Gyorgy I. Targonski. On Carleman integral operators. *Proceedings of the American Mathematical Society*, 18(3):450–456, 1967.

[19] Vladimir Vapnik. *The Nature of Statistical Learning Theory*. Springer, N.Y, 1995.

[20] Grace Wahba. *Spline Models for Observational Data*. Series in Applied Mathematics, Vol. 59, SIAM, Philadelphia, 1990.