Journal of Machine Learning Research 6 (2005) 1485–1515 Submitted 2/03; Revised 11/04; Published 9/05

Frames, Reproducing Kernels, Regularization and Learning

Alain Rakotomamonjy ALAIN.RAKOTOMAMONJY@INSA-ROUEN.FR
Stéphane Canu STEPHANE.CANU@INSA-ROUEN.FR
Perception, Systèmes et Information CNRS FRE2645
INSA de Rouen
76801 Saint Etienne du Rouvray, France

Editor: Alex Smola

Abstract

This work deals with a method for building a reproducing kernel Hilbert space (RKHS) from a Hilbert space with frame elements having special properties. Conditions on existence and a method of construction are given. Then, these RKHS are used within the framework of regularization theory for function approximation. Implications on semiparametric estimation are discussed and a multiscale scheme of regularization is also proposed. Results on toy and real-world approximation problems illustrate the effectiveness of such methods.

Keywords: regularization, kernel, frames, wavelets

1. Introduction

A reproducing kernel Hilbert space (RKHS) is a Hilbert space of functions with special properties (Aronszajn, 1950). It plays an important role in approximation and regularization theory as it allows writing in a simple way the solution of a learning from empirical data problem (Wahba, 1990, 2000). Since the development of support vector machines (SVMs) (Vapnik, 1995; Vapnik et al., 1997; Burges, 1998; Vapnik, 1998) as a machine learning for data classification and functional estimation, there is a growing interest around reproducing kernel Hilbert spaces. In fact, for nonlinear classification or approximation, SVMs map the input space into a high dimensional feature space by means of a nonlinear transformation Φ (Boser et al., 1992). Usually in SVMs, the mapping function is related to an integral operator kernel $K(x,y)$ which corresponds to the dot product of the mapped data:

$$K(x,y) = \langle \Phi(x), \Phi(y) \rangle$$

where x and y belong to the input space.

In regularization theory (Tikhonov and Arsénin, 1977; Groetsch, 1993; Morosov, 1984), the ill-conditioned estimation from data problem is transformed into a well-conditioned problem by means of a stabilizer, which is a functional with specific properties.

For both SVMs and regularization theory, one can consider special cases of kernel and stabilizer: the kernel and the norm associated with an RKHS (Girosi, 1998; Smola et al., 1998; Evgeniou et al., 2000). This justifies the appeal of RKHS as it allows the development of a general framework that includes several approximation schemes.

One of the most important issues in a learning problem is the choice of the data representation. For instance, in SVMs this corresponds to the selection of the nonlinear mapping Φ. It is a key problem since the mapping has a direct influence on the kernel and thus, it has an influence on the solution of the approximation or classification problem. In practical cases, the choice of an appropriate data representation is as important as the choice of the learning machine. In fact, prior information on a specific problem can be used for choosing an efficient input representation, or for choosing a good hypothesis space that leads to enhanced performance of the learning machine (Scholkopf et al., 1998; Jaakkola and Haussler, 1999; Niyogi et al., 1998).

The purpose of this paper is to present a method for constructing an RKHS and its associated kernel by means of frame theory (Duffin and Schaeffer, 1952; Daubechies, 1992). A frame of a Hilbert space spans any vector of the space by linear combination of the frame elements. But unlike a basis, a frame is not necessarily linear independent although it achieves stable representation. Since a frame is a more general way to represent elements of Hilbert space, it allows flexibility in the representation of any vector of the space. By giving conditions for constructing arbitrary RKHS from frame elements, our goal is to widen the choice of kernel so that in future applications, one can adapt its RKHS to prior information available concerning a problem at hand.

The paper is organized as follows: in Section 2, we recall the problem of estimating function from data and the way of solving it owing to regularization theory. Section 3 deals with frame. After a short introduction about frame theory, we give conditions for a Hilbert space described by a frame to be an RKHS and then derive the corresponding kernel. In Section 4, a practical way for building RKHS is given. Section 5 discusses implication of these results on regularization technique and proposes an algorithm for multiscale approximation. Section 6 presents estimation results on numerical experiments on toy and real-world problems while Section 7 concludes the paper and contains remarks and other issues about this work.

2. Regularized Approximation

As argued by Girosi et al. (1995), learning from data can be viewed as a multivariate function approximation from sparse data. Supposing that one has a set of data $\{(x_i,y_i), x_i \in \mathbb{R}^d, y_i \in \mathbb{R}, i = 1 \dots \ell\}$ provided by the random sampling of a noisy function f, the goal is to recover the unknown function f, from the knowledge of the data set. It is well-known that such a problem is ill-posed as there exists an infinity of functions that pass perfectly through the data. One way to transform this problem into a well-posed one is to assume that the function f presents some smoothness properties and hence, the problem becomes a variational problem of finding the function f^* that minimizes the functional (Tikhonov and Arsénin, 1977):

$$H[f] = \frac{1}{\ell}\sum_{i=1}^{\ell} C(y_i, f(x_i)) + \lambda\Omega[f] \tag{1}$$

where λ is a positive number, C a cost function which determines how differences between $f(x_i)$ and y_i should be penalized and $\Omega[f]$ a functional which denotes the prior information on the function f. λ balances the trade-off between fitness of f to the data and smoothness of f. This regularization principle leads to different approximation schemes depending on the cost function $C(\cdot,\cdot)$. Classical L_2 cost function $(C(y_i), f(x_i)) = (y_i - f(x_i))^2$ leads to the so-called Regularization Networks (Girosi et al., 1995; Evgeniou et al., 2000) whereas cost function like Vapnik's $\varepsilon-$insensitive function leads to SVMs.

When the functional $\Omega[f]$ is defined as $\|f\|^2_{\mathcal{H}}$, the square norm of f in a reproducing kernel Hilbert space $\mathcal{H}$ associated to a positive definite function K (the square norm in a Hilbert space being related to the inner product by $\|f\|^2_{\mathcal{H}} = \langle f, f\rangle_{\mathcal{H}})$, the solution of Equation (1) is under general conditions

$$f^*(x) = \sum_{i=1}^{\ell} c_i K(x, x_i). \tag{2}$$

The case of $\|f\|_{\mathcal{H}}$ being a seminorm leads to a minimizer with the following form:

$$f^*(x) = \sum_{i=1}^{\ell} c_i K(x, x_i) + \sum_{j=1}^{m} d_j g_j(x) \tag{3}$$

where $\{g_j\}_{j=1\dots m}$ span the null space of the functional $\|f\|^2_{\mathcal{H}}$.

In a nutshell, looking for a function f of the form (3) is equivalent to minimizing the functional $H[f]$, and thus the solution which depends on λ is the “best” balance between smoothness in $\mathcal{H}$ and fitness to the data. Choosing a kernel K is equivalent to specifying a prior information on the RKHS, therefore having a large choice of RKHS should be fruitful for the approximation accuracy, if overfitting is properly controlled, since one can adapt its hypothesis space to each specific data set.

3. Frames and Reproducing Kernel Hilbert Spaces

In this section, we give an introduction to frame theory that will be useful for the remainder of the paper.

3.1 A Brief Review of Frame Theory

Frame theory was introduced by Duffin and Schaeffer (1952) (Daubechies, 1992) in order to establish general conditions under which one can reconstruct perfectly a function f in a Hilbert space $\mathcal{H}$ from its inner product $(\langle \cdot, \cdot \rangle_{\mathcal{H}})$ with a family of vectors $\{\phi_n\}_{n\in\Gamma}$ with Γ being a finite or infinite countable index set.

Definition 1 *A set of vectors $\{\phi_n\}_{n\in\Gamma}$ is a frame of a Hilbert space $\mathcal{H}$ if there exists two constants $A > 0$ and $\infty > B \geq A > 0$ so that*

$$\forall f \in \mathcal{H}, \qquad A||f||^2_{\mathcal{H}} \leq \sum_{n\in\Gamma} |\langle f, \phi_n\rangle_{\mathcal{H}}|^2 \leq B||f||^2_{\mathcal{H}}. \tag{4}$$

The frame is said to be tight if A and B are equal.

If the set $\{\phi_n\}_{n\in\Gamma}$ satisfies the frame condition then the frame operator U can be defined as

$$U : \begin{array}{ccc} \mathcal{H} & \longrightarrow & \ell^2 \\ f & \longrightarrow & \{\langle f, \phi_n\rangle_{\mathcal{H}}\}_{n\in\Gamma}. \end{array} \tag{5}$$

The reconstruction of f from its frame coefficients needs the definition of a dual frame. For this purpose, one introduces the adjoint operator U^* of U which exists and is unique because it lies on a Hilbert space:

$$U^* : \begin{array}{ccc} \ell^2 & \longrightarrow & \mathcal{H} \\ \{c_n\}_{n\in\Gamma} & \longrightarrow & \sum_{n\in\Gamma} c_n \phi_n. \end{array} \tag{6}$$

Theorem 1 *(Daubechies, 1992) Let $\{\phi_n\}_{n\in\Gamma}$ be a frame of $\mathcal{H}$ with frame bounds A and B. Let us define the dual frame $\{\overline{\phi_n}\}_{n\in\Gamma}$ as $\overline{\phi_n} = (U^\star U)^{-1}\phi_n$. For all $f \in \mathcal{H}$, we have*

$$\frac{1}{B}\|f\|^2_{\mathcal{H}} \leq \sum_{n\in\Gamma} |\langle f, \overline{\phi_n}\rangle_{\mathcal{H}}|^2 \leq \frac{1}{A}\|f\|^2_{\mathcal{H}} \tag{7}$$

and

$$f = \sum_{n\in\Gamma} \langle f, \overline{\phi_n}\rangle_{\mathcal{H}}\phi_n = \sum_{n\in\Gamma} \langle f, \phi_n\rangle_{\mathcal{H}}\overline{\phi_n}. \tag{8}$$

If the frame is tight then $\overline{\phi_n} = \frac{1}{A}\phi_n$.

■

This theorem also shows that the dual frame $\{\overline{\phi_n}\}_{n\in\Gamma}$ is a family of vectors which allows to recover any $f \in \mathcal{H}$, and consequently one can write each vector of the frame and the dual frame as

$$\forall m \in \Gamma, \quad \overline{\phi_m} = \sum_{n\in\Gamma} \langle \overline{\phi_m}, \phi_n\rangle_{\mathcal{H}}\overline{\phi_n} \tag{9}$$

and

$$\forall m \in \Gamma, \quad \phi_m = \sum_{n\in\Gamma} \langle \phi_m, \phi_n\rangle_{\mathcal{H}}\overline{\phi_n}. \tag{10}$$

According to this theorem and the above equations, one can note that an orthonormal basis of $\mathcal{H}$ is a special case of frame where $A = B = 1$, $\overline{\phi_n} = \phi_n$ and $\|\phi_n\| = 1$. However, as stated by Daubechies (1992), frame redundancy can be statistically useful. Also note that in the general case, we do not have an analytical expression of the dual frame, and thus it has to be computed numerically. Grochenig has proposed such an algorithm (Grochenig, 1993) which is based on a iterative conjugate gradient method. We have briefly described this algorithm in the appendix but for further details, one should refer to the original paper.

For the sake of simplicity, in the following we will call frameable Hilbert space, a Hilbert space $\mathcal{H}$ for which there exists a set of vector of $\mathcal{H}$ that forms a frame of $\mathcal{H}$. Note that all separable Hilbert spaces are frameable since by definition they have a countable orthonormal basis.

3.2 A Reproducing Kernel Hilbert Space and Its Frame

After this short introduction on frame theory, let us look at the conditions under which a frameable Hilbert space is also a reproducing kernel Hilbert space.

First of all, we introduce some notations that will be used throughout the rest of the paper: let $\mathbb{R}^\Omega$ be the set of all functions defined on a domain $\Omega \subset \mathbb{R}^d$ with values in $\mathbb{R}$.

For the purpose of being self-contained, we propose here some useful definitions and properties concerning RKHS. However, the reader who is interested in deeper details can refer to books describing mathematical aspects (Atteia, 1992; Berlinet and Agnan, 2004).

Definition 2 *A Hilbert space $\mathcal{H}$ with inner product $\langle\cdot,\cdot\rangle_{\mathcal{H}}$ is a reproducing kernel Hilbert space of $\mathbb{R}^\Omega$ if:*

- *$\mathcal{H}$ is a subspace of $\mathbb{R}^\Omega$*

- $\forall t \in \Omega \quad, \exists M_t > 0$ *so that*
$$\forall f \in \mathcal{H}, \qquad |f(t)| \leq M_t ||f||. \tag{11}$$
This latter property means that for any $t \in \Omega$, the linear functional $\mathcal{F}_t$ (also called the evaluation functional) defined as
$$\mathcal{F}_t(f): \begin{array}{ccc} \mathcal{H} & \longrightarrow & \mathbb{R} \\ f & \longrightarrow & \mathcal{F}_t(f) = f(t) \end{array}$$
is a bounded linear functional.

Note that for any Hilbert space of functions, the evaluation functional is linear, thus the important point for having the reproducing kernel property is this evaluational functional being bounded.

Definition 3 *We call* $\mathrm{Hilb}(\mathbb{R}^\Omega)$ *the set of all RKHS of* $\mathbb{R}^\Omega$.

Owing to the Riesz theorem, one can state that:

Theorem 4 *Let $\mathcal{H} \in \mathrm{Hilb}(\mathbb{R}^\Omega)$, there exists an unique symmetric function $K(\cdot,t)$ of $\mathcal{H}$ called the reproducing kernel of $\mathcal{H}$ so that*

$$\forall t \in \Omega, \quad \forall f \in \mathcal{H}, \quad f(t) = \langle f | K(\cdot,t) \rangle_{\mathcal{H}}. \tag{12}$$

■

Theorem 5 *Let $\mathcal{H}$ be a Hilbert space and $\{\phi_n\}_{n\in\Gamma}$ be a frame of this space. If $\{\phi_n\}_{n\in\Gamma}$ is a (finite or infinite) set of functions of $\mathbb{R}^\Omega$, so that:*

$$\forall t \in \Omega, \qquad \left\| \sum_{n\in\Gamma} \bar{\phi}_n(\cdot)\phi_n(t) \right\|_{\mathcal{H}} < \infty. \tag{13}$$

Then $\mathcal{H}$ is a reproducing kernel Hilbert space.

■

Proof

Step 1 Any ϕ_n is both an element of $\mathbb{R}^\Omega$ and $\mathcal{H}$. Hence the equation

$$\forall f \in \mathcal{H}, \quad f = \sum_{n\in\Gamma} \langle f, \bar{\phi}_n \rangle_{\mathcal{H}} \phi_n$$

holds in $\mathcal{H}$ according to the frame property given in Equation (8) (Mallat, 1998; Daubechies, 1992). Now since, $\mathbb{R}^\Omega$ has a structure of vector space, $f = \sum_{n\in\Gamma} \langle f, \bar{\phi}_n \rangle \phi_n$ is also valid in $\mathbb{R}^\Omega$ and thus f also belongs to $\mathbb{R}^\Omega$. Now, if for each $t \in \Omega$, we define the seminorm on the vector space $\mathbb{R}^\Omega$ as

$$\forall f \in \mathbb{R}^\Omega, \; \|f\|_t = |f(t)|.$$

According to this seminorm, we get the following pointwise convergence:

$$f = \sum_{n\in\Gamma} \langle f, \bar{\phi}_n \rangle_{\mathcal{H}} \phi_n \Leftrightarrow f(t) = \sum_{n\in\Gamma} \langle f, \bar{\phi}_n \rangle_{\mathcal{H}} \phi_n(t). \tag{14}$$

Step 2 Now let's show that $\forall t \in \Omega, \quad \exists M_t > 0$ so that

$$\forall f \in \mathcal{H}, \qquad |f(t)| \leq M_t \|f\|_{\mathcal{H}}. \tag{15}$$

All elements of $\mathcal{H}$ can be expanded with regards to the frame elements, so according to Equation (14), we have for all f in $\mathcal{H}$ and $\mathbb{R}^\Omega$:

$$|f(t)| = \left| \sum_{n \in \Gamma} \langle f(\cdot), \bar{\phi}_n(\cdot) \rangle_{\mathcal{H}} \phi_n(t) \right| \tag{16}$$

and consequently,

$$\begin{aligned} |f(t)| &= \left| \left\langle f(\cdot), \sum_{n \in \Gamma} \bar{\phi}_n(\cdot) \phi_n(t) \right\rangle_{\mathcal{H}} \right| \\ &\leq \|f\|_{\mathcal{H}} \left\| \sum_{n \in \Gamma} \bar{\phi}_n(\cdot) \phi_n(t) \right\|_{\mathcal{H}} \end{aligned} \tag{17}$$

by defining $M_t \triangleq \| \sum_{n \in \Gamma} \bar{\phi}_n(\cdot) \phi_n(t) \|_{\mathcal{H}}$ one can conclude that $\mathcal{H}$ is a reproducing kernel Hilbert space since M_t is finite by hypothesis and therefore, $\mathcal{H}$ admits an unique reproducing kernel.

■

Remark 6 *In this proof, we have chosen to expand a function f of $\mathcal{H}$ according to $f = \sum_{n \in \Gamma} \langle f, \bar{\phi}_n \rangle \phi_n$. However choosing the relationship $f = \sum_{n \in \Gamma} \langle f, \phi_n \rangle \bar{\phi}_n$ would have led to the following equivalent condition to Equation (13):*

$$\forall t \in \Omega, \qquad \left\| \sum_{n \in \Gamma} \bar{\phi}_n(t) \phi_n(\cdot) \right\|_{\mathcal{H}} < \infty. \tag{18}$$

Now let's try to express the reproducing kernel of such a Hilbert space.

Theorem 7 *Let $\mathcal{H}$ be a reproducing kernel Hilbert space and $\mathcal{H} \in Hilb(\mathbb{R}^\Omega)$, and the family $\{\phi_n\}_{n \in \Gamma}$ be a frame of this space, the reproducing kernel is $K(s,t)$ defined by:*

$$K : \left| \begin{array}{l} \Omega \times \Omega \to \mathbb{R} \\ s \times t \to K(s,t) = \sum_{n \in \Gamma} \bar{\phi}_n(s) \phi_n(t) \end{array} \right. \tag{19}$$

Proof

At first, note that according to the frame inequality:

$$\sum_{n \in \Gamma} \phi_n^2(t) = \sum_{n \in \Gamma} |\langle K(t,\cdot), \phi_n(\cdot) \rangle_{\mathcal{H}}|^2 \leq B \|K(t,\cdot)\|^2_{\mathcal{H}} < \infty.$$

Furthermore, according to Theorem (1) we know that $\{\bar{\phi}_n\}_{n \in \Gamma}$ is another frame of $\mathcal{H}$. Thus, similarly to Equation (6), we can define the adjoint operator $U^*_{\bar{\phi}}$ associated to this dual frame. And, applying U_ϕ^* to the ℓ_2 sequence $\{\phi_n(t)\}$ shows that the function $\sum_{n\in\Gamma}\bar{\phi}_n(\cdot)\phi_n(t)$ is a well-defined function of $\mathcal{H}$.

Furthermore, any $f \in \mathcal{H}$ can be expanded by means of the frame of $\mathcal{H}$, thus according to Equation (14):

$$\begin{aligned} f(t) &= \sum_{n\in\Gamma}\langle f,\bar{\phi}_n\rangle_{\mathcal{H}}\phi_n(t) \\ &= \left\langle f(\cdot), \sum_{n\in\Gamma}\bar{\phi}_n(\cdot)\phi_n(t)\right\rangle_{\mathcal{H}} \end{aligned} \tag{20}$$

and since $\mathcal{H}$ is an RKHS, we have

$$\forall f \in \mathcal{H}, \quad \forall t \in \Omega, \qquad f(t) = \langle f(\cdot), K(\cdot,t)\rangle_{\mathcal{H}}. \tag{21}$$

Hence, by identifying Equation (20) and (21) due to the unicity of the reproducing kernel, we have

$$K(\cdot,t) = \sum_{n\in\Gamma}\bar{\phi}_n(\cdot)\phi_n(t)$$

and thus, we can conclude that

$$K(s,t) = \sum_{n\in\Gamma}\bar{\phi}_n(s)\phi_n(t).$$

■

These propositions show that a Hilbert space which can be described by its frame is under general conditions, a reproducing kernel Hilbert space and its reproducing kernel is given by a linear combination of its frame and dual frame product.

A simple corollary to Theorem (7) is that for any RKHS $\mathcal{H}$ with family $\{\phi_n\}_{n\in\Gamma}$ as a frame, the inequality (13) holds. This naturally stems from the fact that $K(\cdot,t) = \sum_{n\in\Gamma}\bar{\phi}_n(\cdot)\phi_n(t)$ is a well-defined function of $\mathcal{H}$ (as stated in the proof of Theorem 7) and thus it has a finite norm in $\mathcal{H}$.

The symmetry and the positivity of the kernel $K(s,t)$ are direct consequences of $K(\cdot,\cdot)$ being a kernel of an RKHS. However, these properties can also be easily shown owing to the frame representation. In fact, according to Equation (8) and (14), we get:

$$\begin{aligned} x(t) &= \sum_{n\in\Gamma}\langle x,\bar{\phi}_n\rangle_{\mathcal{H}}\phi_n(t) = \sum_{n\in\Gamma}\langle x,\phi_n\rangle_{\mathcal{H}}\bar{\phi}_n(t) \\ &= \left\langle x(\cdot), \sum_{n\in\Gamma}\bar{\phi}_n(\cdot)\phi_n(t)\right\rangle_{\mathcal{H}} = \left\langle x(\cdot), \sum_{n\in\Gamma}\phi_n(\cdot)\bar{\phi}_n(t)\right\rangle_{\mathcal{H}} \end{aligned} \tag{22}$$

thus, owing to the uniqueness of the functional evaluation in a RKHS, one can deduce from Equation (22) that

$$K(s,t) = \sum_{n\in\Gamma}\bar{\phi}_n(s)\phi_n(t) = \sum_{n\in\Gamma}\bar{\phi}_n(t)\phi_n(s) = K(t,s).$$

The positivity can also be proved from the following reasoning. Let $x_1,\cdots,x_\ell$ be some vectors of Ω and $a_1,\cdots,a_\ell$ some scalar values in $\mathbb{R}$, we want to show that for any set $\{x_i\}$ and $\{a_i\}$:

$$\sum_{i,j}^{\ell} a_i a_j K(x_i,x_j) \geq 0.$$

According to Equation (10), we can write

$$K(x_i,x_j) = \sum_{n\in\Gamma} \overline{\phi}_n(x_i) \sum_{m\in\Gamma} \overline{\phi}_m(x_j) \langle \phi_n,\phi_m\rangle_{\mathcal{H}}.$$

Thus, we have

$$\begin{aligned}
\sum_{i,j}^{\ell} a_i a_j K(x_i,x_j) &= \sum_{i,j}^{\ell} a_i a_j \sum_{n,m\in\Gamma} \overline{\phi}_n(x_i)\overline{\phi}_m(x_j)\langle \phi_n,\phi_m\rangle_{\mathcal{H}} \\
&= \left\langle \sum_{i}^{\ell}\sum_{n\in\Gamma} a_i\overline{\phi}_n(x_i)\phi_n(\cdot), \sum_{j}^{\ell}\sum_{m\in\Gamma} a_j\overline{\phi}_m(x_j)\phi_m(\cdot) \right\rangle_{\mathcal{H}} \\
&= \left\| \sum_{i}^{\ell}\sum_{n\in\Gamma} a_i\overline{\phi}_n(x_i)\phi_n(\cdot) \right\|_{\mathcal{H}}^2 \\
&\geq 0.
\end{aligned}$$

4. Learning Schemes Using Frames

In the previous section, conditions for a frameable Hilbert space being an RKHS were given. Here, we are interested in constructing a reproducing kernel Hilbert space together with its frame and discuss about the implications of such result in a functional estimation framework.

4.1 Learning on Frameable Hilbert Spaces

An interesting point of frameable Hilbert space is that under weak conditions, it becomes easy to build RKHS. The following theorem proves such point.

Theorem 8 *Let $N \in \mathbb{N}$ and $\{\phi_n\}_{n=1\ldots N}$ be a finite set of non-zero functions of a Hilbert space $(\mathcal{B},\langle\cdot,\cdot\rangle)$ with $\mathcal{B} \subset \mathbb{R}^{\Omega}$ so that*

$$\exists M, \forall t \in \Omega,\ \forall n\ \ 1 \leq n \leq N, \qquad |\phi_n(t)| \leq M.$$

Let $\mathcal{H}$ be the set of functions so that

$$\mathcal{H} = \{ f = \sum_{n=1}^{N} a_n\phi_n : a_n \in \mathbb{R}, \quad n = 1,\ldots,N\}$$

$(\mathcal{H},\langle\cdot,\cdot\rangle_{\mathcal{B}})$ is an RKHS and its reproducing kernel is

$$K(s,t) = \sum_{n=1}^{N} \overline{\phi}_n(s)\phi_n(t),$$

where $\{\overline{\phi}_n\}_{n=1,\ldots,N}$ is the dual frame of $\{\phi_n\}_{n=1,\ldots,N}$ in $\mathcal{H}$.

■

Proof

Step 1 $\mathcal{H}$ is a Hilbert space.

This is straightforward since $\mathcal{H}$ is a closed subspace of a Hilbert space $\mathcal{B}$, and is endowed with $\mathcal{B}$ inner product. Hence $\mathcal{H}$ is a Hilbert space.

Step 2 $\{\phi_n\}$ is a frame of $\mathcal{H}$. A proof of this step is also given in Christensen (1993). We have to show that there exists A and B satisfying equation (4). Let us consider the non trivial case that $\text{span}\{\phi_n\}_{n=1..N} \neq 0$.

The existence of B is straightforward applying Cauchy-Schwartz inequality. In fact, for all $f \in \mathcal{H}$

$$|\langle f, \phi_n \rangle|^2 \leq \|f\|^2 \|\phi_n\|^2$$

and thus

$$\sum_{n=1}^{N} |\langle f, \phi_n \rangle|^2 \leq \|f\|^2 \sum_{n=1}^{N} \|\phi_n\|^2.$$

Thus by taking $B = \sum_{n=1}^{N} \|\phi_n\|^2$, we have $B < \infty$ and B satisfies the right-hand inequality of Equation (4).
Let $\mathcal{H}^* \triangleq \{f \in \mathcal{H} : \|f\|_{\mathcal{H}} > 0\}$ and $S(f)$ be the following mapping:

$$S : \left| \begin{array}{lll} \mathcal{H}^* & \longrightarrow & \mathbb{R} \\ f & \longrightarrow & S(f) = \sum_{n \in \Gamma} |\langle f, \phi_n \rangle|^2. \end{array} \right. \tag{23}$$

This mapping is continuous and because $\mathcal{H}^*$ is of finite dimension the restriction of S to the unit ball in $\text{span}\{\phi_n\}_{n=1..N}$ reach its infimum (Brezis, 1983): there is $g \in \text{span}\{\phi_n\}_{n=1,\ldots,N}$ with $\|g\| = 1$ such that

$$\sum_{n \in \Gamma} |\langle g, \phi_n \rangle|^2 = \inf \left\{ \sum_{n \in \Gamma} |\langle f, \phi_n \rangle|^2, \quad f \in \mathcal{H}^* \text{ so that } \|f\| = 1 \right\}.$$

Let A be $\sum_{n \in \Gamma} |\langle g, \phi_n \rangle|^2$. Hence $A > 0$, and as $\|g\| = 1$, one has for any $f \in \mathcal{H}^*$:

$$A\|f\|^2 \leq \sum_{n=1}^{N} |\langle f, \phi_n \rangle|^2.$$

Step 3 Now let's prove that $\mathcal{H}$ is an RKHS. For that it suffices to prove that the frame $\{\phi_n\}$ satisfies condition given in Theorem 5.

This is straightforward since $\{\phi_n\}_{n=1,\ldots,N}$ is a frame of $\mathcal{H}$ and owing to Theorem 1, the dual frame $\{\bar{\phi}_n\}_{n=1,\ldots,N}$ is also a frame of $\mathcal{H}$. Hence, the norm of each $\bar{\phi}_n$ is finite. Besides, $|\phi_n(t)|$ is supposed to be bounded by M. Hence,

$$\left\| \sum_{n=1}^{N} \bar{\phi}_n(\cdot)\phi_n(t) \right\| \leq M \sum_{n=1}^{N} \|\bar{\phi}_n(\cdot)\| < \infty$$

and consequently, $\mathcal{H}$ is an RKHS with a kernel equal to:

$$K(s,t) = \sum_{n=1}^{N} \bar{\phi}_n(s)\phi_n(t).$$

■

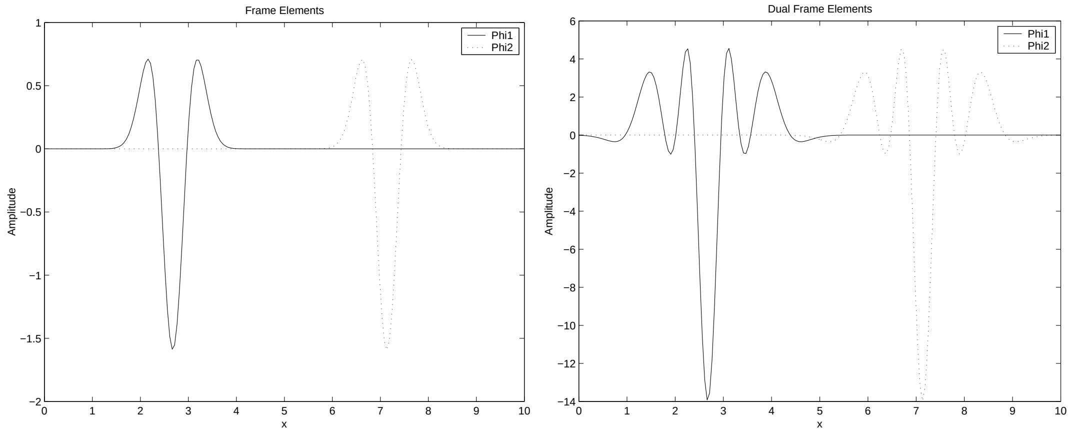


Figure 1: Examples of wavelet frame elements (left) anf their dual elements (right).

Here, we give some examples of RKHS that have been derived from the direct application of this theorem.

Example 1 *Any finite set of bounded, real-valued, pointwise-defined and square integrable functions on Ω endowed with the inner product $\langle f,g\rangle = \int_\Omega f(t)g(t)dt$ spans a RKHS. For instance, the set of functions which expressions are given below spans an RKHS.*

$$\forall t \in \Omega, \phi_n(t) = t \cdot e^{-(t-n)^2}, \qquad n \in [n_{min}, n_{max}] \text{ with } (n_{min}, n_{max}) \in \mathbb{N}^2$$

Example 2 *Any finite set of bounded and pointwise-defined functions belonging to Sobolev space (Berlinet and Agnan, 2004) spans an RKHS. The set of functions, given in the previous example spans also an RKHS in a Sobolev inner product sense.*

Example 3 *Consider a finite set of wavelet on $\mathbb{R}$*

$$\left\{\psi_{j,k}(t) = \frac{1}{\sqrt{a^j}}\psi\left(\frac{t - ku_0a^j}{a^j}\right), j \in \mathbb{Z} : j_{min} \leq j \leq j_{max}, k \in \mathbb{Z} : k_{min} \leq k \leq k_{max}\right\}$$

where $(a, u_0) \in \mathbb{R}_+^* \times \mathbb{R}$, and $(j_{min}, j_{max}, k_{min}, k_{max}) \in \mathbb{Z}^4$. Then the span of these functions endowed with the inner product $\langle f,g\rangle = \int_\mathbb{R} f(t)g(t)dt$ is an RKHS. Figure (1) plots an example of wavelet frame and dual frame elements for a dilation $j = -7$.

The main interest of Theorem (8) is the flexibility it introduces in the RKHS choice or in the choice of the functions that span the hypothesis space. However, this theorem only deals with *finite* dimension RKHS. For building infinite dimensional RKHS, Theorem (5) has to be used. The main difference between the finite and *infinite* dimensional case and thus between Theorems (5) and (8) is that a finite set of functions $\{\phi_n\}_{n=1,\ldots,N}$, if endowed with an adequate inner product, is always a frame of the space it spans (see step 2 of the proof of Theorem (8)) . This is not always true for an infinite set of functions and in this case, the frame condition given in Equation (4) and the boundedness of the evaluation functional in Equation (13) have to be verified. Next examples are examples of infinite dimension RKHS which kernels are given explicitly by their frame elements.

Example 4 *Let us consider $\mathcal{H}$ as the space of continuous and differentiable functions on $\Omega = [0,1]$ with the constraints that for any $f \in \mathcal{H}$, $f(0) = f(1) = 0$ and $\partial f \in L_2(\Omega)$ where ∂f is the usual derivative of f. Endowed with the inner product:*

$$\forall f \text{ and } g \in \mathcal{H},\ \langle f,g\rangle_{\mathcal{H}} = \int_\Omega \partial_x f(x) \partial_x g(x) dx$$

one can show that $\mathcal{H}$ is a Hilbert space of functions on Ω and that the set

$$\{\phi_n(t)\}_{n\in\mathbb{N}^*} = \left\{ \frac{\sqrt{2}}{n\pi} \sin(n\pi t) \right\}_{n\in\mathbb{N}^*}$$

is an orthonormal basis of $\mathcal{H}$ (Debnath and Mikusinki, 1998; Atteia and Gaches, 1999). Hence, $\{\phi_n(x)\}_{n\in\mathbb{N}^*}$ is a tight frame of $\mathcal{H}$ with the frame constant A equals to 1. Let us show that this frame verify the condition given in Theorem (5) in order to prove that $\mathcal{H}$ is an RKHS.

At first, let us prove that for all $t \in \Omega$, the sequence $\{\phi_n(t)\}_{n\in\mathbb{N}^*}$ belongs to ℓ_2. Because $\overline{\phi_n} = \phi_n$, we have for any $t \in \Omega$:

$$\begin{aligned}\sum_{n\in\mathbb{N}^*} \phi_n^2(t) = \sum_{n\in\mathbb{N}^*} \overline{\phi}_n^2(t) &= \sum_{n\in\mathbb{N}^*} \frac{2}{n^2\pi^2} \sin^2(n\pi t) \\ &\leq \frac{2}{\pi^2} \sum_{n\in\mathbb{N}^*} \frac{1}{n^2} \\ &< \infty.\end{aligned}$$

Hence, according to the adjoint frame operator U^* given in equation (6), for any $t \in \Omega$, the function $\sum_{n\in\mathbb{N}^*} \phi_n(\cdot)\phi_n(t)$ is a well-defined function of $\mathcal{H}$. Thus,

$$\left\| \sum_{n\in\mathbb{N}^*} \phi_n(\cdot)\phi_n(t) \right\|_{\mathcal{H}}^2 = \sum_{n\in\mathbb{N}^*} \phi_n^2(t) < \infty.$$

Hence $\mathcal{H}$ is a infinite dimensional RKHS with kernel

$$\forall s,t \in \Omega,\ K(s,t) = \sum_{n=1}^{\infty} \frac{2}{n^2\pi^2} \sin(n\pi s)\sin(n\pi t).$$

Example 5 *This other example shows a way for constructing an infinite dimensional RKHS from its frame. Let $\{\alpha_n\}_{n\in\Gamma}$ be a set of strictly positive real values and define the subspace ℓ^2_α of ℓ^2 as*

$$\ell^2_\alpha = \left\{ c = \{c_n\}_{n\in\Gamma},\ c_n \in \mathbb{R} : \sum_{n\in\Gamma} \frac{c_n^2}{\alpha_n} < \infty \right\}.$$

Endowed with the inner product $\langle c,d\rangle_{\ell^2_\alpha} \equiv \sum_{n\in\Gamma} \frac{c_n d_n}{\alpha_n}$, one can show that ℓ^2_α is a Hilbert space. Now, let $\{\phi_n\}_{n\in\Gamma}$ be a set of functions on $\mathbb{R}^\Omega$ so that:

$$\forall t \in \Omega,\ \sum_{n\in\Gamma} \alpha_n \phi_n^2(t) < \infty$$

and T the mapping:

$$T: \begin{array}{lcl} \ell^2_\alpha & \rightarrow & \mathcal{H} \subset \mathbb{R}^\Omega \\ c & \rightarrow & f = \sum_{n\in\Gamma} c_n\phi_n. \end{array}$$

It is simple to show that $\mathcal{H}$ is a space of functions on Ω since for all $t \in \Omega, \{\alpha_n\phi_n(t)\}_{n\in\Gamma}$ belongs to ℓ^2_α. Then we have,

$$\langle c, \{\alpha_n\phi_n(t)\}_{n\in\Gamma}\rangle_{\ell^2_\alpha} = \sum_{n\in\Gamma} \frac{c_n\alpha_n\phi_n(t)}{\alpha_n} = \sum_{n\in\Gamma} c_n\phi_n(t) < \infty.$$

Suppose furthermore for simplicity and clarity that $\{\phi_n\}_{n\in\Gamma}$ has been chosen so that T is an injective mapping. Then the range of the mapping T also defined as

$$\mathcal{H} = \left\{ f = \sum_{n\in\Gamma} c_n\phi_n : \{c_n\}_{n\in\Gamma} \in \ell^2_\alpha \right\}$$

and endowed with the inner product:

$$\langle f, g\rangle_{\mathcal{H}} \equiv \langle c, d\rangle_{\ell^2_\alpha} = \sum_{n\in\Gamma} \frac{c_n d_n}{\alpha_n} \quad \textit{with } f = \sum_{n\in\Gamma} c_n\phi_n \textit{and } g = \sum_{n\in\Gamma} d_n\phi_n$$

is also a Hilbert space since in this case T is an isometric isomorphism between ℓ^2_α and $\mathcal{H}$ (Debnath and Mikusinki, 1998). Note that this way of building a Hilbert space is also described by Opfer (2004a) and Amato et al. (2004). However, none of them has presented the following frame-based point of view for showing that under some weak hypothesis $\mathcal{H}$ can be an RKHS.

At first, note that due to the one-to-one mapping between ℓ^2_α and $\mathcal{H}$, the following equality holds:

$$\forall k, n \in \Gamma, \quad \langle \phi_k, \phi_n\rangle_{\mathcal{H}} = \frac{\delta_{k,n}}{\alpha_k}$$

where $\delta_{k,n}$ is the Kronecker symbol.

Let us show that $\{\phi_n\}_{n\in\Gamma}$ is a frame of $\mathcal{H}$. Owing to the above property, we have $\sum_{n\in\Gamma} |\langle f, \phi_n\rangle|^2 = \sum_{n\in\Gamma} \frac{c_n^2}{\alpha_n^2}$ and $\|f\|^2_{\mathcal{H}} = \sum_{n\in\Gamma} \frac{c_n^2}{\alpha_n}$, then it is clear that the following inequality holds:

$$\frac{1}{\alpha_{max}} \|f\|^2_{\mathcal{H}} \leq \sum_{n\in\Gamma} |\langle f, \phi_n\rangle|^2 \leq \frac{1}{\alpha_{min}} \|f\|^2_{\mathcal{H}}$$

where $\alpha_{max} = \max_{n\in\Gamma}\alpha_n$ and $\alpha_{min} = \min_{n\in\Gamma}\alpha_n$. Since according to the frame property given in Equation (8), each frame element can be expanded as $\phi_k = \sum_{n\in\Gamma}\langle\phi_k, \phi_n\rangle\bar{\phi}_n$, we have $\phi_k = \frac{1}{\alpha_k}\bar{\phi}_k$. Hence since the frame and dual frame elements are so that for any $t \in \Omega$, we have

$$\sum_{n\in\Gamma} \alpha_n\phi_n^2(t) = \sum_{n\in\Gamma} \frac{(\bar{\phi}_n(t))^2}{\alpha_n} < \infty. \tag{24}$$

Then $\{\bar{\phi}_n(t)\}_{n\in\Gamma} \in \ell^2_\alpha$ and consequently, the function $K(\cdot, t) = \sum_{n\in\Gamma} \bar{\phi}_n(t)\phi_n(\cdot)$ is well-defined, belongs by construction to $\mathcal{H}$ and is so that

$$\left\| \sum_{n\in\Gamma} \bar{\phi}_n(t)\phi_n(\cdot) \right\|_{\mathcal{H}} < \infty,$$

and $\mathcal{H}$ is an RKHS whose kernel is

$$K(s,t) = \sum_{n \in \Gamma} \overline{\phi}_n(t)\phi_n(s) = \sum_{n \in \Gamma} \alpha_n \phi_n(s)\phi_n(t).$$

A practical example of such an infinite dimensional RKHS can be obtained as follows. Let us consider that $\Omega = \mathbb{R}$ and each $\phi_n(t) = \frac{1}{\sqrt{2^J}}\varphi\left(\frac{t-n}{2^J}\right)$ with $n \in \mathbb{Z}$, $J \in \mathbb{Z}$ and $\varphi(t)$ a pointwise-defined on Ω and compactly supported function so that $\{\phi_n\}_{n\in\Gamma}$ are linearly independent. Examples of such functions $\varphi(t)$ are functions that are classically used in wavelet-based multiresolution analysis (Mallat, 1998). Since each ϕ_n is a compactly supported shift of a function φ, for any t, the sum in Equation (24) becomes a finite sum of non-zero terms which convergence is consequently guaranteed for any $\{\alpha_n\}_{n\in\Gamma}$. At this point, we can state that the space

$$\mathcal{H} = \left\{ f = \sum_{n\in\Gamma} \frac{c_n}{\sqrt{2^J}} \varphi\left(\frac{t-n}{2^J}\right) : \sum_{n\in\Gamma} \frac{c_n^2}{\alpha_n} < \infty \right\}$$

is a reproducing kernel Hilbert space.

If we want $\mathcal{H}$ to be the span of different dilations and shifts of φ, we can also show that $\mathcal{H}$ is an RKHS by choosing the $\{\alpha_n\}_{n\in\Gamma}$ to be related to the dilation parameter J so that the inequality in (24) holds.

4.2 Other Classes of Frame-Based Kernels

Recently, Gao et al. (2001) have proposed another class of frame-based kernels. Their approach is based on the connection between regularization operator and support vector kernel as described in Smola et al. (1998). Supposing that U is the frame operator of a either finite or infinite dimensional RKHS, their kernel is based on the statement that the operator $Q = U^*U$ is a symmetric positive definite operator and the Green function associated to this operator is a Mercer kernel. Thus, the kernel they proposed, named the frame operator kernel, can be expanded with respect to the dual frame elements as

$$K(s,t) = \sum_{n\in\Gamma} \overline{\phi}_n(s)\overline{\phi}_n(t).$$

A detailed proof of this equation is given in Gao et al. (2001).

From the point of view of the regularization theory (Smola et al., 1998), this frame-operator kernel of Gao et al. is different from the one we propose as the regularization operator associated to each of them are different. In fact, in our case the regularization operator can be considered as the projector of any function space on $\mathcal{H}$ whereas in the Gao et al. case, it can be seen as the frame operator U.

More recently, Opfer (2004b) has shown that the kernel associated to an RKHS $\mathcal{H}$ can be expanded as

$$K(s,t) = \sum_{n\in\Gamma} \phi_n(s)\phi_n(t)$$

if and only if the set of functions $\{\phi_n\}_{n\in\Gamma}$ is a super tight frame (which is a tight frame with frame bounds equal to 1) of $\mathcal{H}$. This results is a particular case of Theorem (7) since for a super tight frame each dual frame element is $\overline{\phi}_n = \phi_n$. Furthermore, compared to Opfer's work, our Theorem (5) gives a frame-based condition for a Hilbert space to be an RKHS.

The works of Amato et al. (2004) and Opfer (2004a) where they both proposed the concept of multiscale kernels can also be related to our work. Interestingly, they have both shown that a Hilbert space spanned by wavelet can be under some weak hypotheses an RKHS. The way they build their RKHS $\mathcal{H}$ is very similar to the one we described in example (5) and the related reproducing kernel is naturally

$$K(s,t) = \sum_{n\in\Gamma} \alpha_n \phi_n(s)\phi_n(t),$$

where each α_n is a strictly positive real value. On one hand, Amato et al. ended up with this kernel by considering that $\{\phi_n\}_{n\in\Gamma}$ are a orthonormal wavelet basis of $L_2([0,1])$ and showing that for their space $\mathcal{H}$, the evaluation functional is continuous. On the other hand, for achieving this result, Opfer has shown that the function $K(\cdot,t)$ belongs to $\mathcal{H}$ and satisfies the reproducing property without explicit explanations on how this kernel has been obtained. Hence, although very similar to the work of Opfer, the example (5) gives the functional setting on how the kernel in (Opfer, 2004a) can be derived.

5. Discussions

Propositions presented in previous sections describe a way for easily building RKHS and its associate reproducing kernel. Hence, this kernel can be used within the framework of regularization networks or SVMs for functional estimation.

For SVMs, one usually chooses as a kernel a continuous symmetric function K in $L_2(\Omega)$ (Ω being a compact subset of $\mathbb{R}^d$) that has to satisfy the following condition, known as Mercer's condition:

$$\int_\Omega \int_\Omega K(x,y) f(x) f(y) dx dy \geq 0 \tag{25}$$

for all $f \in L_2(\Omega)$.

Now, one may ask what are the advantages and drawbacks of using kernels built by means of Theorem (5) or (8).

- Both Mercer's condition and frameable RKHS allow to obtain a positive definite function. However, it is obvious that conditions for having frameable RKHS are easier to verify than Mercer's condition. Thus, this can be interpreted as a flexibility for adapting kernel to a particular problem. Examples of this flexibility will be given below within the context of semiparametric estimation. Notice that methods for choosing the appropriate frame elements of the RKHS are not given here.

  Example 6 *Consider the set of functions on* $\mathbb{R}$ $\left\{\phi_n(s) = \frac{sin(\pi(s-n))}{\pi(s-n)}\right\}_{n=1\dots N}$. *The space spanned by these frame elements associated to* $L_2(\mathbb{R})$ *inner product is an RKHS. Thus, as a direct corollary of Theorem 8, the kernel*

  $$K(s,t) = \sum_{i=1}^{N} \overline{\phi}_i(s)\phi_i(t)$$

  is an admissible kernel for SVMs.

  A representation of such a kernel with $N = 9$ *is given in Figure (2).*

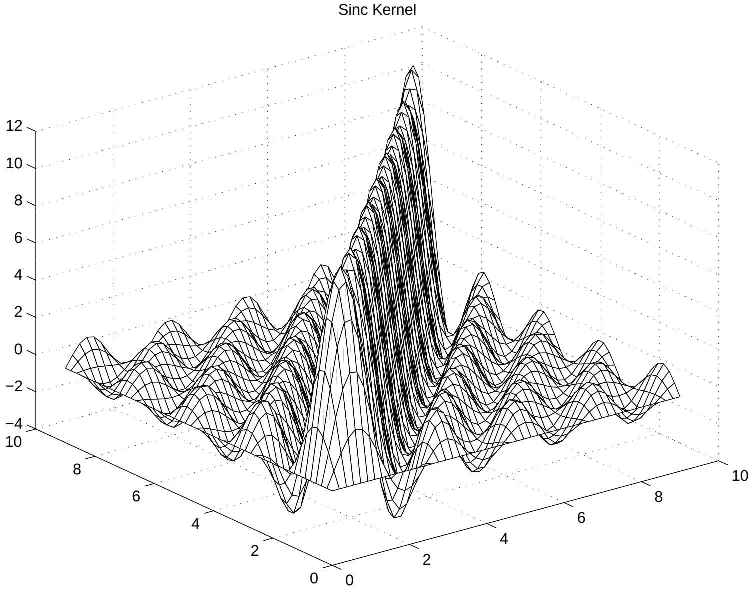


Figure 2: The sinc kernel.

- Since conditions for obtaining a frameable RKHS hold mainly for finite dimensional space (although, it may exists infinite dimensional Hilbert space which frame elements satisfy hypotheses of Theorem (5)), it is fairest to compare the frameable kernel to a finite dimensional kernel. According to Mercer's condition, or other more detailed papers on the subject (Aronszajn, 1950; Wahba, 2000), Mercer's kernel can be expanded as follows:

  $$K(s,t) = \sum_{n=1}^{N} \frac{1}{\lambda_n} \psi_n(s)\psi_n(t)$$

  where s and t belong to Ω, λ_l is a positive real number and $\{\psi_l\}_{i=1..N}$ is a set of orthogonal functions. Conditions for constructing frameable kernel are less restricting since the orthogonality of the frame elements are not needed. One can note that for tight frame or orthonormal basis, frameable kernel leads to the following expansion:

  $$K(s,t) = \sum_{n=1}^{N} \frac{1}{A} \psi_n(s)\psi_n(t)$$

  since dual frame elements is equal to frame elements up to a multiplicative constant depending on the frame bound A . Tightness of a frame is a very interesting property since in this case processing the dual frame is no more needed. However, unless we explicitly build the RKHS $\mathcal{H}$ so that it is spanned by a tight frame (as in example (5) or in Opfer (2004b)), tightness of a frame needs more constraints on the frame elements than other frames. Thus a tight frame of a space is harder to build than other frame of the same space.

- The conditions for a frameable Hilbert space being an RKHS is given in Equation (13) and they hold also for infinite dimensional case for which the kernel is written

  $$K(s,t) = \sum_n \overline{\phi}_n(s)\phi_n(t).$$

  Again in this case, the frame kernel expansion is similar to the Mercer's kernel one. The main difference between the finite and infinite dimensional case relies on the fact that a finite set of functions $\{\phi_n\}$ is always a frame of the space it spans (provided that this latter is endowed with an adequate inner product). This is not always true for an infinite set of functions. However, we have shown in example (5) that under some mild conditions, it is possible to build an infinite dimensional RKHS.

- In the SVMs algorithm, the kernel realizes the dot product of the data points mapped in some feature space:

  $$K(s,t) = \langle \Phi(s), \Phi(t) \rangle$$

  with Φ being the mapping. Usually, this mapping is not explicitly given since one only needs for computing the optimal hyperplane the dot product in the feature space. With frame-based kernels, we have the relation

  $$\begin{aligned} K(s,t) &= \sum_{n=1}^{N} \overline{\phi}_n(s)\phi_n(t) \\ &= \sum_{n=1}^{N} \overline{\phi}_n(s) \sum_{j=1}^{N} \overline{\phi}_j(t) \langle \phi_j(\cdot), \phi_n(\cdot) \rangle_{\mathcal{H}} \qquad \text{according to Equation (10)} \\ &= \left\langle \sum_{n=1}^{N} \overline{\phi}_n(s)\phi_n(\cdot), \sum_{j=1}^{N} \overline{\phi}_j(t)\phi_j(\cdot) \right\rangle_{\mathcal{H}}. \end{aligned}$$

  Thus the data embedding can be defined as

  $$\Phi : \begin{array}{ccc} \Omega & \longrightarrow & \mathcal{H} \\ t & \longrightarrow & \sum_{n=1}^{N} \overline{\phi}_n(t)\phi_n(\cdot). \end{array}$$

  The data points are mapped to a function belonging to $\mathcal{H}$. The mapping is consequently strictly related to the frame elements $\{\phi_n\}$ and is implicitly defined by them.

- Besides, since the kernel has an expansion with regards to the frame elements, the solution of Equation (1) is of easier interpretation. Indeed, although the solution depends on the kernel expression, it can be rewritten as a linear combination of the frame elements. Thus, compared to other kernels for which basis functions are unknown, using frame-based kernel increases model interpretability.

- Drawbacks of using frame-based kernel rely mainly on the time complexity burden that is added for constructing the data model. For both SVMs and regularization networks, one has to process the kernel matrix K with elements $K_{i,j} = K(x_i,x_j)$. Thus, with frame-based kernel, one has to compute the dual frame elements, (for instance, by means of an iterative algorithm, as the one described in (Grochenig, 1993)). This by its own may be time-consuming. Furthermore, the construction of the matrix K needs the processing of the sum. Hence, if the number N of frame elements describing the kernel and the number ℓ of data are large, building K becomes rapidly very time-consuming (of an order of $N^2 \cdot \ell^2$).

Some of these points suggest that frame-based kernels can be useful by themselves. However, within the context of semiparametric estimation, this flexibility for building kernel offers some other interesting perspectives. Semiparametric estimation can be introduced by the following theorem.

Theorem 9 *(Kimeldorf and Wahba, 1971)*
Let $\mathcal{H}_K$ be an RKHS of real valued functions on Ω with reproducing kernel K. Denote by $\{(x_i,y_i)_{i=1\ldots\ell}\}$ the training set and let $\{g_j, j=1\ldots N\}$ be a set of functions on Ω such that the matrix $G_{i,j} = g_j(x_i)$ has maximal rank. Then, the solution to the problem

$$\min_{f\in span(g)+h,h\in\mathcal{H}_K} \frac{1}{\ell}\sum_{i=1}^{\ell} C(y_i, f(x_i)) + \lambda\|f\|^2_{\mathcal{H}_K} \tag{26}$$

has a representation of the form

$$f(\cdot) = \sum_{i=1}^{\ell} c_i K(x_i,\cdot) + \sum_{j=1}^{N} d_j g_j(\cdot).$$

■

The solution of this problem can be interpreted as a semiparametric estimation since one part of the solution (the first sum) comes from a non-parametric estimation (the regularization problem) while the other term is due to the parametric expansion (the span of $\{g_j\}$). As stated by Smola in his thesis (Smola, 1998), semiparametric estimation can be advantageous with regards to a fully non parametric estimation as it exploits some prior knowledge on the estimation problem (for instance major properties of the data are described by linear combination of a small set of functions), and making a “good” guess (on the set of functions $\{g_j\}$) can have a large effect on performance.

Again in this context, the flexibility of frame-based kernel can be exploited. In fact, let $G = \{g_i\}_{i=1\ldots N}$ be a set of N linearly independent functions that satisfies Theorem 8, hence, any subset of G, $\{g_i\}_{i\in\Gamma}$,(Γ being an index set of size $n_0 < N$) can be used for building an RKHS $\mathcal{H}_K$ while the remaining vectors can be used in the parametric part of the Kimeldorf-Wahba theorem. Hence in this case, the solution of (26) is written

$$f(\cdot) = \sum_{i=1}^{\ell} c_i \sum_{k\in\Gamma} \bar{g}_k(x_i) g_k(\cdot) + \sum_{j\in C_\Gamma} d_j g_j(\cdot).$$

The flexibility comes from the fact that in a learning problem, any elements of G can be regularized (if involved in the span of $\mathcal{H}_K$) or can be kept as it is (if used in the parametric part). Intuitively, one should use any vector that comes from “good” prior knowledge, in the parametric part of the approximation while leaving in the kernel expansion the other frame elements. Notice also that only

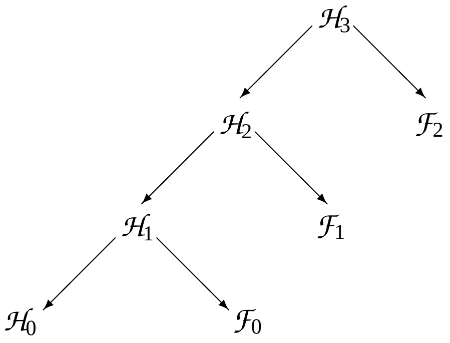


Figure 3: Example of multiscale approximation on 3 levels. Each space $\mathcal{H}_j$ can be decomposed in a trend space $\mathcal{H}_{j-1}$ and a detail space $\mathcal{F}_{j-1}$. In this case, $\mathcal{H}_3$ can be considered as the sum of $\mathcal{H}_0$, $\mathcal{F}_0$, $\mathcal{F}_1$ and $\mathcal{F}_2$.

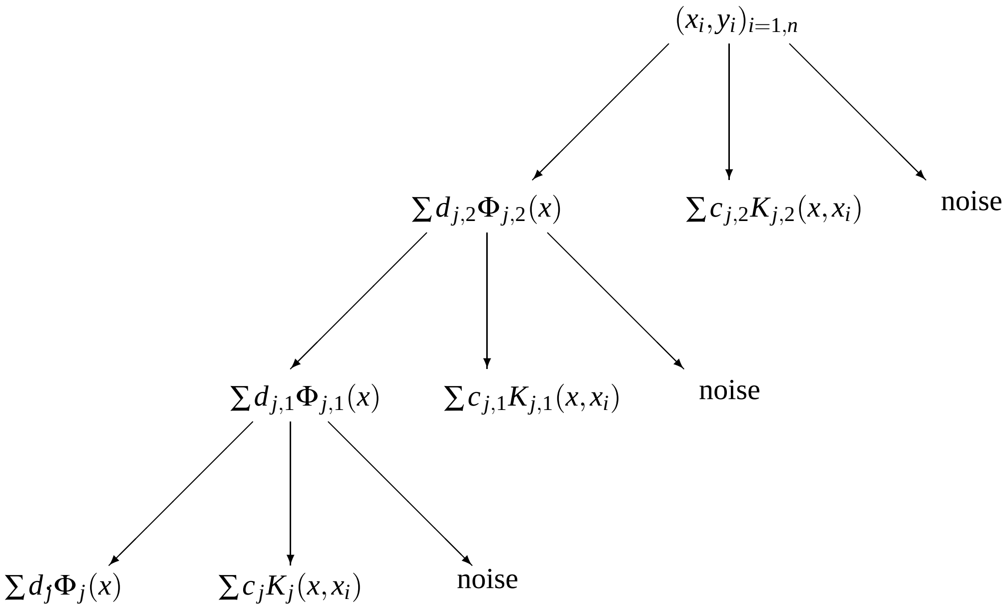

$(x_i, y_i)_{i=1,n}$
$\sum d_{j,2}\Phi_{j,2}(x)$
$\sum c_{j,2}K_{j,2}(x,x_i)$
noise
$\sum d_{j,1}\Phi_{j,1}(x)$
$\sum c_{j,1}K_{j,1}(x,x_i)$
noise
$\sum d_j\Phi_j(x)$
$\sum c_jK_j(x,x_i)$
noise

Figure 4: Example of multiscale approximation on 3 levels: the kernel point of view. For instance, here we want to learn a function $f(x)$ that has generated the samples $(x_i, y_i)_{i=1,n}$ under some noisy condition. The first step consists of decomposing the hypothesis space into a parametric part spanned by $\{\Phi_{j,2}(x)\}$ and a non parametric part spanned by $K_{j,2}(x,x_i)$. Then the resulting parametric approximation is decomposed again in two parts and so on. The multiscale approximation of $f(x)$ is then $\hat{f}(x) = \sum d_j\Phi_j(x) + \sum c_jK_j(x,x_i) + \sum c_{j,1}K_{j,1}(x,x_i) + \sum c_{j,2}K_{j,2}(x,x_i)$.

the subset of G which is used in the parametric part has to be linearly independent.

Another perspective which follows directly from this finding is a technique of regularization that we call multiscale regularization which is inspired from the multiresolution analysis of Mallat (1998). Here, we just sketch the idea behind this concept and in no way, the following paragraph should be considered a complete study of this new technique since the analysis of its properties goes beyond the scope of this paper. Consider the same problem as the one described in Theorem 9. Now, suppose that $\{g_i\}$ is a set of N linearly independent functions verifying Theorem (8). Let $\{\Gamma_i\}_{i=0\dots m}$ be a set of index set such that $\cup_{i=0}^{m}\Gamma_i = \{1,\dots,N\}$ and $\Gamma_i \cap \Gamma_j = \emptyset$ for $i \neq j$ and $\mathcal{H}$ being the RKHS spanned by $\{g_i\}$. By subdividing the set $\{g_i\}$ with the index set $\{\Gamma_i\}_{i=0\dots m}$, one can construct m RKHS $\{\mathcal{F}_i\}_{i=0\dots m-1}$ in such a way that

$$\forall i = 1\dots m, \quad \mathcal{F}_{i-1} = \text{span}\,\{g_k\}_{k\in\Gamma_i}$$

and reproducing kernel of $\mathcal{F}_i$ is noted K_i. Now, denote as $\mathcal{H}_i$ the RKHS such that

$$\forall i = 1\dots m, \quad \mathcal{H}_i = \mathcal{H}_{i-1} + \mathcal{F}_{i-1}$$

with $\mathcal{H}_0 = \text{span}\,\{g_k\}_{k\in\Gamma_0}$. By construction, the space $\mathcal{H}_i$ are nested spaces:

$$\mathcal{H}_0 \subset \mathcal{H}_1 \subset \dots \subset \mathcal{H}_m = \mathcal{H}.$$

In this case, one can interpret $\mathcal{H}_0$ as the space of lower approximation capacity whereas $\mathcal{H}_m$ is the space with higher capacity. Besides, since $\mathcal{H}_i = \mathcal{H}_{i-1} + \mathcal{F}_{i-1}$, one can think of $\mathcal{F}_{i-1}$ as the details needed to be added to $\mathcal{H}_{i-1}$ to obtain $\mathcal{H}_i$, thus we will call spaces $\mathcal{F}_i$ the "details" spaces whereas spaces $\mathcal{H}_i$ are the "trend" spaces. Every of these spaces $\mathcal{F}_i$ and $\mathcal{H}_i$ are an RKHS since any subset of $\{g_i\}$ satisfies Theorem (8).

Multiscale regularization is an iterative technique that at step $k = 1,\dots,m$ consists of looking for the solution $f_{m-k}(\cdot)$ of the following minimization problem:

$$\min_{f\in\mathcal{H}_{m-k+1}} \frac{1}{n}\sum_{i=1}^{n} C(y_{i,m-k}, f(x_i)) + \lambda_{m-k}\|f\|^2_{\mathcal{F}_{m-k}} \tag{27}$$

where $y_{i,m-1} = y_i$, $y_{i,m-(k+1)} = y_{i,m-k} - \sum_{j=1}^{n} c_{j,m-k}K_{m-k}(x_j,x_i)$. According to the representer Theorem (9), $f_{m-k}(\cdot)$ can be written:

$$f_{m-k}(\cdot) = \sum_{i=1}^{n} c_{i,m-k}K_{m-k}(x_i,\cdot) + \sum_{j\in\cup_{l=0}^{m-k}\Gamma_l} d_{j,m-k}g_j(\cdot) \tag{28}$$

and thus the overall solution of the so-called multiscale regularization is

$$\hat{f}(\cdot) = \sum_{k=1}^{m}\sum_{i=1}^{n} c_{i,m-k}K_{m-k}(x_i,\cdot) + \sum_{j\in\Gamma_0} d_{j,0}g_j(\cdot). \tag{29}$$

The solution $\hat{f}$ of the multiscale regularization is the sum of different approximations on nested spaces. At first, one seeks to approximate the data on the highest approximation capacity space by regularizing only the details. Then, these details are subtracted to the data and one tries to approximate this residual on the next space by keeping regularizing the details on this space, and so on. Thus at each step, one can control the "amount" of regularization brought to each details space, increasing in this way the capacity control capability of the model. Figure (3) and (4) show an example of how the algorithm works for a 3-level approximation scheme.

The framework of additive models of Hastie et al. (Hastie and Tibshirani (1990)) can give other insights to multiscale regularization. In fact, if we suppose that the family $\{g_i\}_{i=1,\dots,N}$ forms an orthonormal basis of $\mathcal{H}$ and build the spaces $\mathcal{H}_0$ and $\mathcal{F}_m$ in the same way as described above, then by construction, we have

$$\mathcal{H} = \mathcal{H}_0 \oplus \mathcal{F}_0 \oplus \cdots \oplus \mathcal{F}_{m-1}.$$

Hence any function $f \in \mathcal{H}$ can be written as $f(x) = \sum_{i=0}^{m} f_i(x)$ with $f_0 \in \mathcal{H}_0$ and $f_i \in \mathcal{F}_{i-1}$ for $i = 1,\dots,m$. Thus, the multiscale regularization algorithm can be interpreted as an algorithm which looks for the function f that minimizes the following empirical risk:

$$R_{reg}[f] = \frac{1}{\ell}\sum_{i=1}^{\ell} C(y_i, \sum_{j=0}^{m} f_j(x_i)) + \sum_{j=1}^{m} \lambda_j \|f_j\|^2_{\mathcal{F}_{j-1}} \tag{30}$$

where each λ_j is a hyperparameter that controls the amount of regularization for $\mathcal{F}_{j-1}$. This minimization problem is typically the problem of fitting an additive model as proposed by Hastie and Tibshirani (1990).

Illustrations of the multiscale regularization algorithm on both toy and real-world problems are given in the next section.

6. Numerical Experiments

This section describes some experiments that compare frame-based kernels to classical one (for instance gaussian kernel) on some regression problems. Besides, illustrations of some points raised in the discussion such as the multiscale approximation algorithm are given.

6.1 Experiment 1

This first experiment aims at comparing the behavior of different kernels using regularization networks and support vector regression. The function to be approximated is

$$f(x) = \sin x + \operatorname{sinc}(\pi(x-5)) + \operatorname{sinc}(5\pi(x-2)) \tag{31}$$

where $\operatorname{sinc}(x) = \frac{\sin x}{x}$. Data used for the approximation is corrupted by an additive noise, thus $y_i = f(x_i) + \varepsilon_i$ where ε_i is a zero-mean gaussian noise of standard deviation 0.2 . Points x_i are drawn from uniform random sampling of interval $[0,10]$. Three kernels have been used for the approximation:

- Gaussian kernel:
$$K(x,y) = e^{-\frac{\|x-y\|^2}{2\sigma^2}}$$
- Wavelet kernel:
$$K(x,y) = \sum_{i\in\Gamma} \bar{\psi}_i(x)\psi_i(y)$$
where i denote a multi index and $\psi_i(x) = \psi_{j,k}(x) = \frac{1}{\sqrt{a^j}}\psi\left(\frac{x-ku_0a^j}{a^j}\right)$. $\psi(x)$ is the mother wavelet which in this experiment is a mexican hat wavelet. Dilation parameter j takes value

	Regularization Networks	Support vector regression
Gaussian kernel	0.0218 ± 0.0049	0.0248 ± 0.0058
Wavelet kernel	0.0249 ± 0.0078	0.0291 ± 0.0086
Sin/Sinc kernel	0.0249 ± 0.0122	0.0302 ± 0.0176

Table 1: True generalization error for Gaussian, Wavelet, Sin/Sinc kernels with Regularization Networks and support vector regression for the best hyperparameters.

in the set $\{-5,0,5\}$ whereas k is chosen so that a given wavelet $\psi_{j,k}(x)$ has its support in the interval $[0,10]$. For now on, we set $u_0 = 1$ and $a = 2^{0.25}$. These values are those proposed by Daubechies (Daubechies, 1992) so that a wavelet set is a frame of $L_2(\mathbb{R})$. Notice that in our case, we only use a subset of this frame.

- Sin/Sinc kernel:
$$K(x,y) = \sum_{i\in\Gamma} \bar{\phi}_i(x)\phi_i(y)$$
where $\phi_i(x) = \{1, \sin(x), \cos(x), \mathrm{sinc}(j\pi(x-k)) : j \in \{1,3,6\}, k \in [1\dots 9]\}$.

For frame-based kernel, if necessary the dual frame is processed using Grochenig's algorithm.

For both regularization network and support vector regression, some hyperparameters have to be tuned. Different approaches are possible for solving this model selection problem. In this study, the true generalization error has been evaluated for a range of finely sampled values of hyperparameters. This is repeated for a hundred different data sets, and the mean and standard deviation of the generalization error are thus obtained. Table 1 depicts the true generalization error evaluated on 200 datapoints for the two learning machines and the different kernels using the best hyperparameters setting. Analysis of this table leads to the following observation: The different kernels and learning machines give comparable results (all averages are within one standard deviation from each other). Using prior knowledge on the problem in this context does not improve performance (Sin/Sinc kernel or wavelet kernel compared to gaussian kernel). A justification can be that such kernels use strong prior knowledge (the *sin* frame element) that is included in the kernel expansion and thus this prior knowledge gets regularized as much as other frame elements. This suggests that semiparametric regularization should be more appropriate to get advantage of such a kernel.

6.2 Experiment 2

In this experiment, we suppose that some additional knowledge on the approximation problem is available, and thus its exploitation using semiparametric approximation should lead to better performance. We have kept the same experimental setup as the one used in the first example but we have restricted our study to regularization networks.

Basis functions and kernel used are the following:

- Gaussian kernel and sinusoidal basis functions $\{1, \sin(x), \cos(x)\}$.
- Gaussian kernel and wavelet basis functions $\left\{\psi_{j,k}(x) = \frac{1}{\sqrt{a^j}}\psi\left(\frac{x-ku_0a^j}{a^j}\right), j \in \{0,5\}\right\}$

- Wavelet kernel and wavelet basis functions: these functions are the same as in the previous case but the kernel is built only with low dilation wavelet ($j = -10$). In a nutshell, we can consider that the RKHS associated to the kernel used in the non- parametric context (experiment 1) has been splitted in two RKHS. One that leads to a hypothesis space that have to be regularized and another one that does not have to be controlled.

- Sinc kernel and Sin/Sinc basis functions: in this setting, the kernel is given by the following equation:
$$K(x,y) = \sum_{i \in \Gamma} \overline{\phi}_i(x)\phi_i(y)$$
with $\phi_i(x) = \{\text{sinc}(j\pi(x-k)) : j \in \{3,6\}, k \in [1\dots9]\}$

  and the basis functions are $\{1, \sin x, \cos x, \text{sinc}(\pi(x-k) : k \in [1\dots9]\}$.

For each kernel, model selection has been solved by cross-validation using 50 data sets. Then, after having spotted the best hyperparameters, the experiment was run a hundred times and the true generalization error in a mean-square sense, was evaluated. Table 2 summarizes all these trials and describes the performance improvement achieved by different kernels compared to the gaussian kernel and sin basis functions. From this table, one can note that:

- exploiting prior knowledge on the function to be approximated leads immediately to a lower generalization error (compare Table 1 and Table 2).

- as one may have expected, using strong prior knowledge on the hypothesis space and the related kernel gives considerably higher performances than gaussian kernel. In fact, the sinc-based kernel achieves by far the lower mean square error. The idea of including the “good” knowledge in a non-regularized hypothesis space while including the “bad” prior knowledge in the RKHS span seems to be fruitful in this case (the frame elements $\text{sinc}(3\pi(x-k))$ and $\text{sinc}(6\pi(x-k))$ can be termed as “bad” knowledge as, they are not used in the target function).

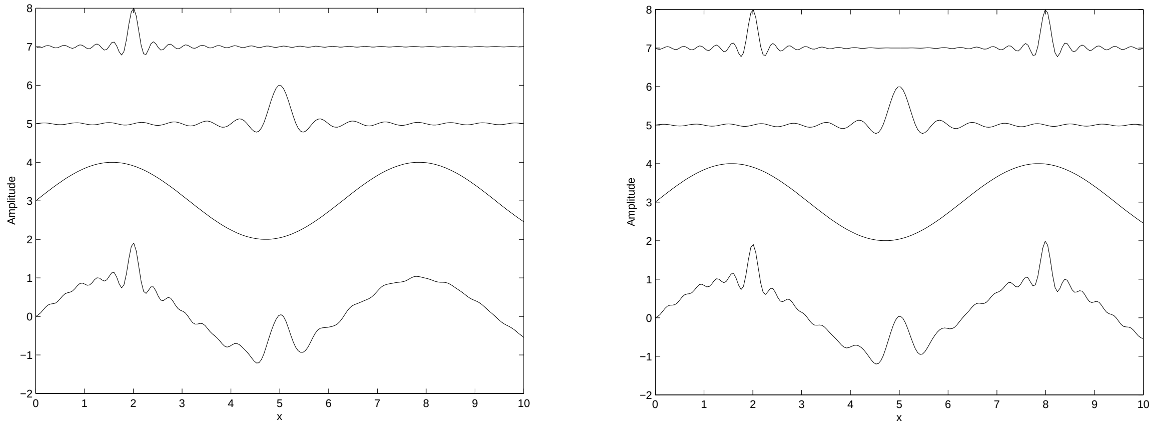
- wavelet kernel achieves minor improvement of performance compared to gaussian kernel. However, this is still of interest as using wavelet kernel and basis functions does corresponds to prior knowledge that can be reformulated as: “the function to be approximated contains smooth structure (the *sin* part), irregular structures (the *sinc* part) and noise”. It is obvious that knowing the true basis function leads to better performance, however that information is not always available and using bad knowledge may result in poorer performance. Thus, prior knowledge on structures which may be easiest to get than prior knowledge on basis function can be easily exploited by means of wavelet span and wavelet kernel.

6.3 Experiment 3

This last simulated example targets at illustrating the concept of multiscale regularization. We have compared several learning algorithms in function approximation problems. The learning machines are: regularization networks, SVM, semiparametric regularization and multiscale regularization. For the two first methods, a gaussian kernel is used whereas for the two latter, wavelet kernel

Kernel / Basis Functions	M.S.E	Improvement (%)
Gaussian / Sin	0.0216 ± 0.0083 (6)	0
Gaussian / Wavelet	0.0202 ± 0.0072 (4)	4.6
Wavelet / Wavelet	0.0195 ± 0.0077 (2)	9.7
Sinc / Sin	0.0156 ± 0.0076 (88)	27.8

Table 2: True generalization performance for semiparametric regression networks and different settings of kernel and basis functions. The number in parentheses reflects the number of trials for which the model has been the best model.



(a)

(b)

Figure 5: Original functions used for benchmarking in experiment 3. (a) f_1 (b) f_2. Top: multiscale structure on 3 levels. Bottom: Complete function.

and basis functions are taken. The true functions used for benchmarking are the following:

$$\begin{aligned} f_1(x) &= \sin x + \operatorname{sinc}(3\pi(x-5)) + \operatorname{sinc}(6\pi(x-2)), \\ f_2(x) &= \sin x + \operatorname{sinc}(3\pi(x-5)) + \operatorname{sinc}(6\pi(x-2)) + \operatorname{sinc}(6\pi(x-8)). \end{aligned}$$

The two functions f_1 and f_2 have been randomly sampled on the interval $[0,10]$. Gaussian noise ε_i of standard deviation 0.2 is added to the samples, thus the entries of the learning machines become $\{x_i, f(x_i) + \varepsilon_i\}$. Here again, a range of finely sampled values of hyperparameters has been tested for model selection. In each case, an averaging of the true error generalization over 100 data sets of 200 samples was evaluated using a uniform measure.

For semiparametric regularization, the kernel and basis setting was built with a wavelet set given by

$$\psi_{j,k}(x) = \frac{1}{\sqrt{a^j}}\psi\left(\frac{x - ku_0a^j}{a^j}\right).$$

	f_1	f_2
Gaussian Reg. Networks	0.0266 ± 0.0085	0.0385 ± 0.0141
Gaussian SVM	0.0328 ± 0.0093	0.0475 ± 0.0155
Semip Reg. Networks 1	0.0266 ± 0.0085	0.0397 ± 0.0113
Semip Reg. Networks 2	0.0236 ± 0.0063	0.0353 ± 0.0080
Multi. Regularization	0.0246 ± 0.0060	0.0344 ± 0.0069

Table 3: True mean-square-error generalization for regularization networks, SVM, semiparametric regularization networks, and multiscale regularization for f_1 and f_2.

The kernel is constructed from a set of wavelet frame of dilation j_{SPH} and the basis functions are given by another wavelet set described by j_{SPL}. For multiscale regularization, the setting of the nested spaces are the following:

$$\mathcal{H}_0 = \text{span}\left\{ \frac{1}{\sqrt{a^j}} \psi \left(\frac{t - ku_0 a^j}{a^j} \right), j = 5 \right\},$$
$$\mathcal{F}_0 = \text{span}\left\{ \frac{1}{\sqrt{a^j}} \psi \left(\frac{t - ku_0 a^j}{a^j} \right), j = 0 \right\},$$
$$\mathcal{F}_1 = \text{span}\left\{ \frac{1}{\sqrt{a^j}} \psi \left(\frac{t - ku_0 a^j}{a^j} \right), j = -10 \right\}.$$

These dilation parameters have been set in a *ad hoc* way, but their choices can be justified by the following reasoning: Three distinct levels have been used for separating the approximation in three structures which should be smooth ($j = 5$), irregular ($j = 0$) and highly irregular ($j = -10$). The same values of j were used in the semiparametric context. Two semiparametric settings have been tested: the first one uses $j_{SPH} = -10$ and $j_{SPL} = \{0, 5\}$ and the other one is configured as follows $j_{SPH} = \{-10, 0\}$ and $j_{SPL} = 5$.

Table 3 presents the average of the mean-square error of the different learning machines for the two functions and for the best hyperparameter value found by cross-validation. Comments and analysis of this experiment validating the concept of multiscale approximation are:

- semiparametric 2 and multiscale approximation give the best mean-square error. They achieve respectively a performance improvement with regards to gaussian regularization networks of 11.2% and 7.5% for f_1, and 8.3% and 10.6% for f_2. Also note that both learning machines give the lowest standard deviation of the mean square error.

- multiscale approximation balances loss of approximation due to error at each level (see Figure) and flexibility of regularization, thus its performance is better than semiparametric one's when the multiscale structure of the signal is more pronounced.

- comparison of the two semiparametric settings shows that the second setup outperforms the first one (especially for f_2). This highlights the importance of selecting the hypothesis space

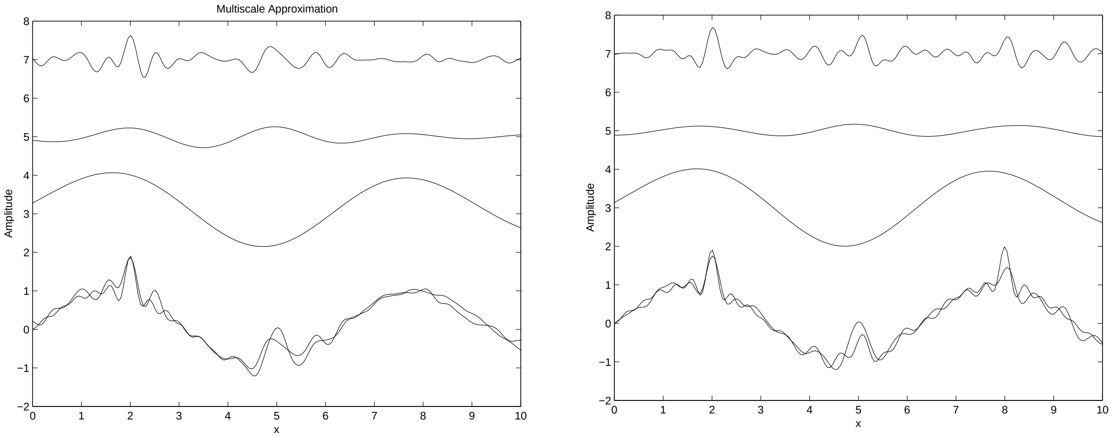


Figure 6: Top: Multiscale structure of a typical prediction of of f_1 (left) and f_2 (right) by multiscale wavelet approximation Bottom: full approximation and true function

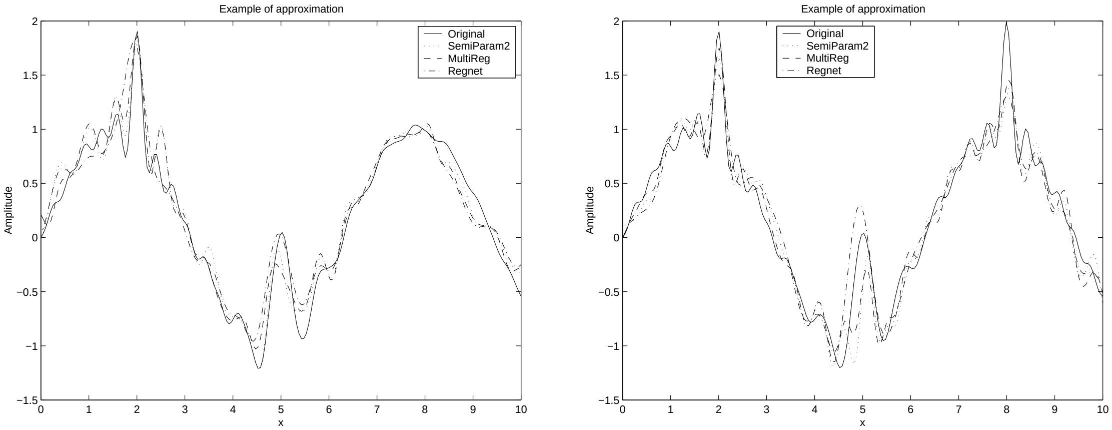


Figure 7: Examples of approximation of f_1 (left) and f_2 (right): Original function (Solid), Semiparametric 2 (dotted), Multiscale regularization (dashed), Regularization network (dash-dotted

to be regularized. In this experiment, it seems that leaving the space spanned by wavelet of dilation $j = 0$ on the parametric span (the space which is not regularized) leads to overfitting.

- multiscale approximation is able to catch all the structures of the signal (see Figure (7)). One can see that each level of approximation represents one structure of the function f_1 and f_2: the lowest dilation ($j = -10$) represents the wiggles due to the highest frequency sinc, at level $j = 0$, one has the $\mathrm{sinc}(3x)$ function whereas the *sin* is located on the highest dilation $j = 5$.

- Figure (6.3) shows that multiscale and semiparametric algorithms achieve better approximation of the "wiggles" than nonparametric methods without compromising smoothness in region of the functions where it is needed.

6.4 Experiments on Real-World Data Sets

This paragraph presents some estimation results on real-world time-series. These times-series are publicly available in a time-series data library (Hyndman and Akram (1998)) and have already been widely used in the field of statistics. The first one *engines* concerns a monthly measured ratio between the motor vehicles engines production and the consumer price index in Canada whereas the second one *basiron* deals with the monthly production of iron in Australia. The problem we want to solve is the estimation of these time-series after a zero-mean normalization.

For this purpose, two models have been compared, the first one being a regularization networks with a gaussian kernel whereas the other one is a multiscale regularization algorithm with an orthogonal wavelet kernel. The wavelet that has been used is a *Symmlet* wavelet with 4 vanishing moments (Mallat, 1998). The kernel of the corresponding hypothesis space $\mathcal{H}$ which have been split into three orthogonal spaces, is so that

$$\mathcal{H} = \mathcal{H}_0 \oplus \mathcal{F}_0 \oplus \mathcal{F}_1 \quad \text{and} \quad K_{\mathcal{H}}(x,y) = K_{\mathcal{H}_0}(x,y) + K_{\mathcal{F}_0}(x,y) + K_{\mathcal{F}_1}(x,y)$$

with

$$K_{\mathcal{H}_0}(x,y) = \sum_{j=j_{min}}^{j_1} \sum_k \psi_{j,k}(x)\psi_{j,k}(y) + \sum_k \phi_{j,k}(x)\phi_{j,k}(y), \tag{32}$$

$$K_{\mathcal{F}_0}(x,y) = \sum_{j=j_1+1}^{j_2} \sum_k \psi_{j,k}(x)\psi_{j,k}(y), \tag{33}$$

$$K_{\mathcal{F}_1}(x,y) = \sum_{j=j_2+1}^{j_{max}} \sum_k \psi_{j,k}(x)\psi_{j,k}(y), \tag{34}$$

and the dilation indexes are so that $j_{min} \le j_1 \le j_2 \le j_{max}$. For both data sets, we have set $j_{min} = -3$, $j_1 = 0$, $j_2 = 4$ and $j_{max} = 7$.

For each estimation trial, each data set has been randomly split in a learning set of 100 samples with the remaining samples being considered as the test set. The results that we present are the normalized mean-squared error averaged over 30 trials for the best hyperparameters values of each model: for the gaussian regularization networks and the multiscale regularization networks, these hyperparameters are respectively $\{\lambda, \sigma\}$ and $\{\lambda_0, \lambda_1, \lambda_2\}$ which are the regularization parameters associated to each scale. The best hyperparameters have been obtained by evaluating the test error on a large range of finely sampled values of these hyperparameters.

	basiron	*engine*
Gaussian Reg. Networks	10.55 ± 1.24 (1)	37.57 ± 5.62 (8)
Multi. Regularization	9.58 ± 1.21 (29)	36.00 ± 4.30 (22)

Table 4: Averaged normalized mean-square error of estimation of real-world time-series with a gaussian and a wavelet multiscale regularization networks. The number within parenthesis is the number of time a given model has performed better than the other.

Table (4) summarizes the performance of each model. It shows that for both time-series, the multiscale algorithm performs better than the gaussian regularization networks. Indeed, for the *basiron* data set, although the difference in normalized mean-squared error is only about 0.9%, the multiscale approach has given the best results on 29 of the 30 trials. For the *engines* time-series, although the difference in normalized mean-squared error is higher (1%), our algorithm gives better results on only 22 trials. Figure (8) depicts some examples of estimation for both time-series and algorithms. This figure shows that the best model for the gaussian regularization networks is rather a smooth model whereas the wavelet multiscale model is far less smooth. This is essentially due to the nature of the time-series which are composed of a slow-varying part denoting the trend of the series, and a fast-varying part denoting the fluctuation of the time-series around the trend. Hence, because of the particular structure of the signal to be estimated, the gaussian model is not able to estimate correctly both the trend and the fluctuation whereas the multiscale model gives a better estimate. This is particularly clear for the *basiron* data set which is composed of a slow-varying trend and fluctuations.

7. Conclusions

In this paper, we showed that an RKHS can be defined by its frame elements and conversely, one can construct an RKHS from a frame. One of the key result is that the space spanned by any finite number of functions belonging to a given Hilbert space, endowed with an adequate inner product, is an RKHS with a kernel that can be at least numerically described. We have also proposed some conditions for a infinite dimensional Hilbert space to be an RKHS. These conditions depend on the frame and the dual frame elements of the Hilbert space and under some weak hypothesis, these conditions are easy to check (see example 5) . Hence, we have essentially provided some methods for building a specific kernel adapted to a problem at hand.

By exploiting this new way for constructing RKHS, a multiscale algorithm using nested RKHS has been introduced and examples given in this paper showed that using this algorithm or a semi-parametric approach with frame-based kernel improves the result of a regression problem with regards to nonparametric approximation. It has also been shown that these frame-based kernels allow better approximation only if exploited in a semiparametric context. Using them as a regularization network or SVMs kernels are not as efficient as one may have expected. However, depending on the prior knowledge on the problem, one can build appropriate kernels that can enhance the quality of the regressor within a semiparametric approach. However, for fully taking advantage of the main theorem proposed in this paper, one has to answer some open questions:

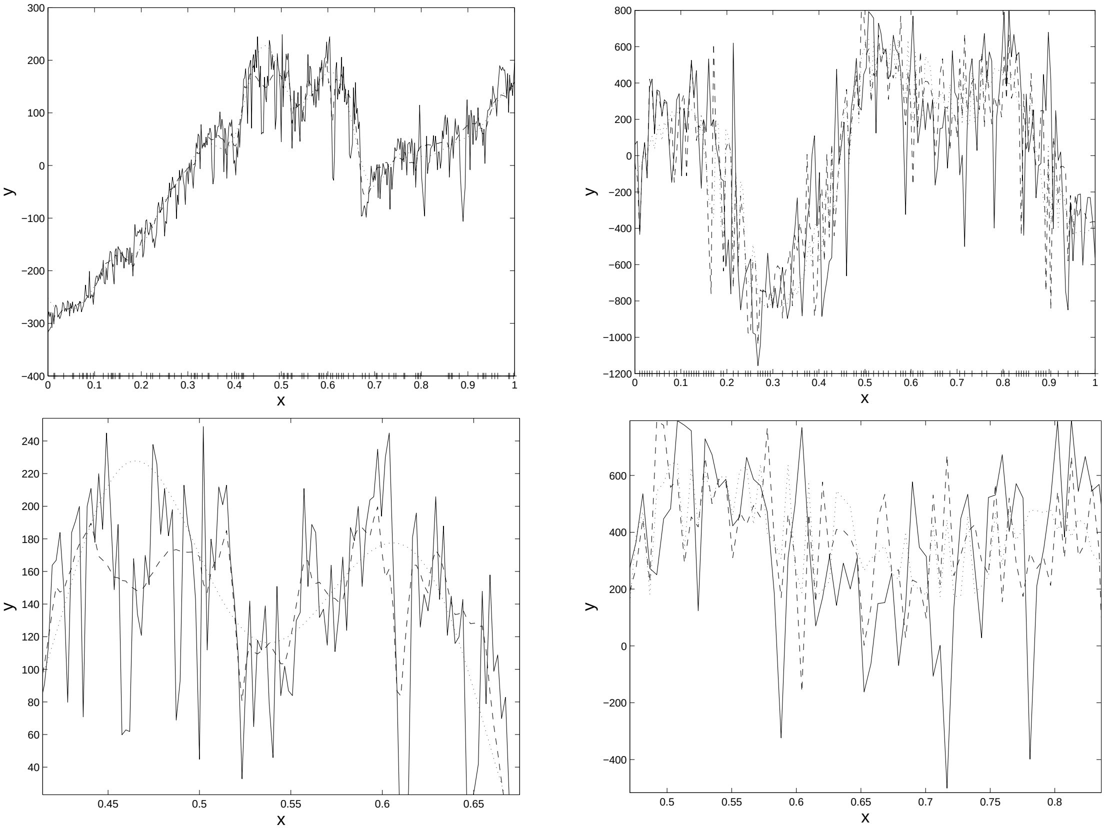


Figure 8: Examples of estimation of real-world time-series. The left and right columns respectively depict estimation of *basiron* and *engines*. The top and bottom figures respectively show the full time-series estimations and a zoomed version of these estimations. The $+$ marks at the bottom of each figure denote the position of the learning examples in the time-series. (solid) true function. (dotted) gaussian regularization networks estimation. (dashed) wavelet multiscale regularization networks estimation.

- we give conditions for building RKHS to be used for approximation. But the difficulty stands in one question: How to transform prior information on the learning problem to frame elements? This is still an open issue.

- reconstruction from frame elements has been shown to be more robust in presence of noise (Daubechies, 1992). In fact, redundancy attenuates noise effects on the frame coefficients. Thus, this is a good statistical argument for using frame with high redundancy. However, this implies the computing of the dual frame and consequently a higher time complexity of the algorithm. Hence, fast algorithms still have to be derived.

- a multiscale regularization algorithm has been sketched in this paper in order to take advantage of frame kernels. Although some experiments show that in some situations, this algorithm performs well, it is not clear whether it theoretically sounds or not. Hence, some further works have to be carried for a better theoretical understanding of this novel regularization method and for a better implementation of the algorithm and all the subsequent problems such as model selection.

Acknowlegments

We would like to thank the anonymous reviewers for their useful comments. This work was supported in part by the IST Program of the European Community, under the PASCAL Network of Excellence, IST-2002-506778. This publication only reflects the authors' views.

Appendix A.

We recall in this appendix a numerical method to process the dual frame of a frameable Hilbert space $\mathcal{H}$ with frame elements $\{\phi_n\}_{n\in\Gamma}$. Let us define the operator S

$$S:\left|\begin{array}{lcl}\mathcal{H} & \longrightarrow & \mathcal{H}\\ f & \longrightarrow & \sum_{n\in\Gamma}\langle f,\phi_n\rangle\phi_n.\end{array}\right. \tag{35}$$

One can also write the operator S as $S \triangleq U^*U$ where U is the frame operator defined in equation (5) and (6). Our goal is to process

$$\forall n, \qquad \bar{\phi}_n = S^{-1}\phi_n.$$

Grochenig (1993) has proposed an algorithm to compute the problem $f = S^{-1}g$. The idea is to calculate f with a gradient descent algorithm along orthogonal directions with respect to norm induced by the symmetric operator S:

$$\|f\|_S^2 = \|Sf\|^2.$$

This norm is useful to compute the error.

Theorem 10 *Let $g \in \mathcal{H}$. To compute $f = S^{-1}g$, one has to initialize*

$$f_0 = 0\ ,\ r_0 = p_0 = g\ ,\ p_{-1} = 0.$$

Then, for any $n \geq 0$, one defines by induction,

$$\lambda_n = \frac{\langle r_n, p_n\rangle}{\langle p_n, Sp_n\rangle} \tag{36}$$

$$f_{n+1} = f_n + \lambda_n p_n \tag{37}$$

$$r_{n+1} = r_n - \lambda_n S p_n \tag{38}$$

$$p_{n+1} = Sp_n - \frac{\langle Sp_n, Sp_n\rangle}{\langle p_n, Sp_n\rangle}p_n - \frac{\langle Sp_n, Sp_{n-1}\rangle}{\langle p_{n-1}, Sp_{n-1}\rangle}p_{n-1}. \tag{39}$$

If $\sigma = \frac{\sqrt{B}-\sqrt{A}}{\sqrt{B}+\sqrt{A}}$, then

$$\|f - f_n\|_S \leq \frac{2\sigma^n}{1+2\sigma^n}\|f\|_S \tag{40}$$

and thus, $\lim_{n\to+\infty} f_n = f$.

Then, in order to process numerically the dual frame of $\mathcal{H}$, one has to apply this algorithm on each element of the frame.

One can note that the speed of convergence is highly dependent on frame bounds A and B.

References

U. Amato, A. Antoniadis, and M. Pensky. Wavelet kernel penalized estimation for non-equispaced design regression. *Statistics and Computing*, to appear, 2004.

N. Aronszajn. Theory of reproducing kernels. *Trans. Am. Math. Soc.*, (68):337–404, 1950.

M. Atteia. *Hilbertian kernels and spline functions*. North-Holland, 1992.

M. Atteia and J. Gaches. *Approximation hilbertienne : Splines, Ondelettes, Fractales*. Presses Universitaires de Grenoble, 1999.

A. Berlinet and C. Thomas Agnan. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Kluwer Academic Publishers, 2004.

B. E. Boser, I. M. Guyon, and V. N. Vapnik. A training algorithm for optimal margin classifiers. In D. Haussler, editor, *5th Annual ACM Workshop on COLT*, pages 144–152, Pittsburgh, PA, 1992. ACM Press.

H. Brezis. *Analyse fonctionnelle, Théorie et applications*. Masson, 1983.

C. Burges. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2(2), 1998.

O. Christensen. *Frame decomposition in Hilbert Spaces*. PhD thesis, Aarhus Univ. Danemmark and Univ. of Vienna, Austria, 1993.

I. Daubechies. *Ten Lectures on Wavelet*. SIAM, CBMS-NSF regional conferences edition, 1992.

L. Debnath and P. Mikusinki. *Introduction to Hilbert Spaces with applications*. Academic Press, 1998.

R. Duffin and A. Schaeffer. A class of nonharmonic fourier series. *Trans. Amer. Math. Soc.*, 72: 341–366, 1952.

T. Evgeniou, M. Pontil, and T. Poggio. Regularization Networks and Support Vector Machines. *Advances in Computational Mathematics*, 13(1):1–50, 2000.

J. Gao, C. Harris, and S. Gunn. On a class of a support vector kernels based on frames in function hilbert spaces. *Neural Computation*, 13(9):1975–1994, 2001.

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

F. Girosi, M. Jones, and T. Poggio. Regularization theory and neural networks architectures. *Neural Computation*, 7(2):219–269, 1995.

K. Grochenig. Acceleration of the frame algorithm. *IEEE Trans. Signal Proc.*, 41(12):3331–3340, 1993.

C. Groetsch. *Inverse Problems in the mathematical sciences*. Vieweg and Sohn, 1993.

T. Hastie and R. Tibshirani. *Generalized additive models*. Chapman and Hall, 1990.

R. Hyndman and M. Akram. Time Series data Library. Technical report, University of Monash, Dept of Econometrics and Business Statistics, 1998. http://www-personal.buseco.monash.edu.au/ hyndman/TSDL/index.htm.

T. Jaakkola and D. Haussler. Probabilistic kernel regression models. In *Proceedings of the 1999 Conference on AI and Statistics*, 1999.

G. Kimeldorf and G. Wahba. Some results on Tchebycheffian spline functions. *J. Math. Anal. Applic.*, 33:82–95, 1971.

S. Mallat. *A wavelet tour of signal processing*. Academic Press, 1998.

V. Morosov. *Methods for solving incorrectly posed problems*. Springer Verlag, 1984.

P. Niyogi, F. Girosi, and T. Poggio. Incorporating prior information in machine learning by creating virtual examples. In *Proceedings of the IEEE*, volume 86, pages 2196–2209, 1998.

R. Opfer. Multiscale kernels. Technical report, Institut fur Numerische und Angewandte Mathematik, Universitt Gottingen, 2004a.

R. Opfer. Tight frame expansions of multiscale reproducing kernels in Sobolev spaces. Technical report, Institut fur Numerische und Angewandte Mathematik, Universitt Gottingen, 2004b.

B. Scholkopf, P. Y. Simard, A. J. Smola, and V. Vapnik. Prior knowledge in support vector kernels. In M. I. Jordan, M. J. Kearns, and S. A. Solla, editors, *Advances in Neural information processings systems*, volume 10, pages 640–646, Cambridge, MA, 1998. MIT Press.

A. Smola. *Learning with Kernels*. PhD thesis, Published by: GMD, Birlinghoven, 1998.

A. Smola, B. Scholkopf, and KR Muller. The connection between regularization operators and support vector kernels. *Neural Networks*, 11:637–649, 1998.

A. Tikhonov and V. Arsénin. *Solutions of Ill-posed problems*. W.H. Winston, 1977.

V. Vapnik. *The Nature of Statistical Learning Theory*. Springer, N.Y, 1995.

V. Vapnik. *Statistical Learning Theory*. Wiley, 1998.

V. Vapnik, S. Golowich, and A. Smola. *Support Vector Method for function estimation, Regression estimation and Signal processing*, volume Vol. 9. MIT Press, Cambridge, MA, neural information processing systems, edition, 1997.

G. Wahba. *Spline Models for Observational Data*. Series in Applied Mathematics, Vol. 59, SIAM, 1990.

G. Wahba. An introduction to model building with reproducing kernel hilbert spaces. Technical Report TR-1020, University of Wisconsin-Madison, 2000.