Frame, Reproducing Kernel, Regularization and Learning

Alain Rakotomamonjy and Stéphane Canu

P.S.I INSA de Rouen
Avenue de l'université
76801 Saint Etienne du Rouvray France
(alain.rakotomamonjy,stephane.canu)@insa-rouen.fr

Abstract

This works deals with a method for building Reproducing Kernel Hilbert Space (RKHS) from a Hilbert Space with frame elements having special properties. Conditions on existence and 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 approximation problems illustrate the effectiveness of such methods.

1 Introduction

A Reproducing Kernel Hilbert Space is a Hilbert Space of functions with special properties (Aronszajn 1950). It plays an important role in approximation and regularization theory as it allows to write in a simple way the solution of a learning from empirical data problem. (Wahba 1990, Wahba 2000). Since the development of the Support Vector Machine (SVM), technique proposed by Vapnik et al. (Vapnik 1998, Vapnik, Golowich & Smola 1997) as a machine learning for data classification and approximation, there is a new growing interest around Reproducing Kernel Hilbert Space (RKHS). In fact, for nonlinear classification or approximation, SVM maps the input space into a high dimensional feature space by means of a nonlinear transformation Φ (Boser, Guyon & Vapnik 1992, Vapnik 1995, Burges 1998). Usually, in SVM, 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 & Arsénin 1977, Canu n.d.), the ill-posed approximation from data problem is transformed into a well-posed problem by means of a stabilizer, which is a functional with specific properties. For both SVM and regularization problem, one can respectively consider special cases of kernel and stabilizer : the kernel and the norm associated to a RKHS (Girosi 1998, Smola, Scholkopf & Muller 1998, Evgeniou, Pontil & Poggio 2000). This justifies the attractivity of the RKHS as it allows to develop a general framework that includes several approximation schemes.

One of the most important issue in a learning problem is the choice of the data representation. For instance, in SVM this corresponds to the selection of the nonlinear mapping Φ. This is a key problem as 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 a practical case, 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 allows to enhance performance of the learning machine (Scholkopf, Simard, Smola & Vapnik 1998, Jaakkola & Haussler 1999, Niyogi, Girosi & Poggio 1998).

The purpose of this paper is to present a method for constructing a RKHS and its associated kernel by means of the Frame theory (Duffin & Schaeffer 1952, Daubechies 1992). A frame of a Hilbert Space allows to represent any vector of the space by linear combination of the frame elements. Unlike a basis, a frame is not necessarily linear independant, but, nevertheless it achieves stable representation. As frame is a more general way to represent elements of Hilbert Space, it allows flexibility in the representation, and hence, broadens the choice of the RKHS. Then, by giving conditions for constructing arbitrary RKHS, 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 given problem.

The paper is organized as follows : in section 2, we recall the problem of approximating function from data and the way to solving such problem owing to regularization theory.

Section 3 deals with frame theory. After a short introduction about frame theory, we give conditions for a Hilbert Space described by a frame for being a RKHS and then derive the corresponding kernel. In section 4, a pratical 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 approximation results on numerical experiments on a toy problem while Section 7 concludes the paper and contains remarks and other issues on this work.

2 Regularized Approximation

As argued by Girosi et al.(Girosi, Jones & Poggio 1995), learning from data can be viewed as a multivariate function approximation from sparse data. The problem is : supposing that one has a set of data $\{(x_i, y_i), x_i \in \mathbb{R}^d, y_i \in \mathbb{R}, i = 1 \dots N\}$ 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 in 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 & Arsénin 1977) :

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

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 smoothness of f. λ balances the trade-off between fitness of f to the data and smoothness of f. Using this regularization principles leads to different approximation schemes that depends on the choice of the cost function C. 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). whereas cost function like Vapnik's $\epsilon-$insensitive norm leads to SVM. Besides, according to the expression of the functional $\Omega[f]$, one can obtain differents kinds of networks such as Radial Basis Function or Gaussian SVM.

When the functional $\Omega[f]$ is defined as $\|f\|^2_{\mathcal{H}}$, the square norm in a Reproducing Kernel Hilbert Space $\mathcal{H}$, defined by a positive definite function K, the form of the solution of equation (1) is, under general condition :

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

The case of semi-definite positive function K leads to a minimizer with the following form :

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

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

The form of solution given in (3) is the so-called dual form of f^* because the solution can also be written with regards to the basis function of the RKHS. In a nutshell, looking for a function f of the form (3) is equivalent to minimizing the functional $H(f)$, and thus, depending on λ the solution is the "best" balance between smoothness in $\mathcal{H}$ and fitness to the data. The choice of the kernel K is equivalent to choosing a specific RKHS, therefore having a large choice of RKHS should be fruitful for the accuracy of approximation, as one can adapt its RKHS to each specific data set.

3 Frame and Reproducing Kernel Hilbert Space

3.1 A Brief review on Frame Theory

The frame theory was introduced by Duffin et al (Duffin & Schaeffer 1952, Daubechies 1992) in order to establish general conditions under which one can reconstruct perfectly a vector f in a Hilbert space $\mathcal{H}$ from its inner product with a family of vectors $\{\phi_n\}_{n\in\Gamma}$. Γ is an index set of either finite or infinite dimension.

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*

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

a so-called tight frame for which frame bounds 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

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

The frame decomposition allows to represent vector f in another way. The reconstruction of f from its frame coefficients need the definition of a dual frame. This needs the introduction of the adjoint operator U^* of U, which exists and is unique because it lies on a Hilbert Space :

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

Theorem 3.1.1 *(Daubechies, 1992)*
Let $\{\phi_n\}_{n\in\Gamma}$ be a frame of $\mathcal{H}$ with frame bounds A and B. Let define the dual frame $\{\bar{\phi}_n\}_{n\in\Gamma}$ as $\bar{\phi}_n = (U^*U)^{-1}\phi_n$. For all $f \in \mathcal{H}$

$$\frac{1}{B}\|f\|^2 \leq \sum_{n\in\Gamma} |\langle f, \bar{\phi}_n\rangle|^2 \leq \frac{1}{A}\|f\|^2 \tag{7}$$

and

$$f = \sum_{n\in\Gamma} \langle f, \bar{\phi}_n \rangle \phi_n = \sum_{n\in\Gamma} \langle f, \phi_n \rangle \bar{\phi}_n \tag{8}$$

if the frame is said to be tight, then $A = B$ *and* $\bar{\phi}_n = \frac{1}{A}\phi_n$

This theorem proves that the dual frame $\{\bar{\phi}_n\}_{n\in\Gamma}$ is a family of vectors that allows to recover any $f \in \mathcal{H}$, and consequently, one can write each vector of the frame and the dual frame as

$$\bar{\phi}_p = \sum_n \langle \bar{\phi}_p, \phi_n \rangle \bar{\phi}_n \tag{9}$$

and

$$\phi_p = \sum_n \langle \phi_p, \phi_n \rangle \bar{\phi}_n \tag{10}$$

An orthonormal basis of $\mathcal{H}$ is a special case of frame where $A = B = 1$. However, redundancy brought by frame are statistically useful (Daubechies 1992, Soltani 1999)

In a sake of simplicity, we will call Frameable Hilbert Space, a Hilbert Space $\mathcal{H}$ which admits a frame i.e there exists a set of vector of $\mathcal{H}$ that forms a frame of $\mathcal{H}$.

3.2 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, and then we give an expression of the kernel.

First of all, we introduce some notations, that will be used through the rest of the paper : let Ω be a compact domain included in $\mathbb{R}^d$ and $\mathbb{R}^\Omega$ be the set of function $f : \Omega \longrightarrow \mathbb{R}$.

For the purpose of being self-containing, we propose here some useful definition and properties concerning RKHS. However, the reader who is interested in deeper details can refer to books containing rigorous mathematical aspects (Atteia & Gaches 1999).

Definition 3.2.1 *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* $\mathbb{R}^\Omega$
- $\forall t \in \Omega \quad , \exists M_t > 0$ *so that*

$$\forall x \in \mathcal{H}, \qquad |x(t)| < M_t ||x|| \tag{11}$$

This latter property means that for any $t \in \Omega$, the functionals $\mathcal{F}_t$ defined as :

$$\mathcal{F}_t = f(t) \qquad \forall f \in \mathcal{H}$$

are linear and bounded functionals.

Definition 3.2.2 *we call* $Hilb(\mathbb{R}^\Omega)$ *the set of RKHS of* $\mathbb{R}^\Omega$.

then owing to the Riesz theorem (Atteia & Gaches 1999), one can state that :

Definition 3.2.3 *Let $\mathcal{H} \in Hilb(\mathbb{R}^\Omega)$, there exists an unique application $K : \Omega \times \Omega \to \mathbb{R}$, called the Reproducing Kernel of $\mathcal{H}$ so that :*

$$\forall t \in \Omega, \quad \forall x \in \mathcal{H}, \quad x(t) =< x|K(\cdot, t) > \tag{12}$$

Theorem 3.2.1 *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 function of $\mathbb{R}^\Omega$, so that*

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

then $\mathcal{H}$ is a Reproducing Kernel Hilbert Space.

Proof Owing to the frame property, one can show that $\mathcal{H} \subset \mathbb{R}^\Omega$. In fact,

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

hence as

$$\forall n \in \Gamma, \quad \phi_n \in \mathbb{R}^\Omega$$

and $\mathbb{R}^\Omega$ has a structure of vector space, any f in $\mathcal{H}$ belongs to $\mathbb{R}^\Omega$.

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

$$\forall x \in \mathcal{H}, \qquad |x(t)| \leq M_t ||x|| \tag{14}$$

all the elements of $\mathcal{H}$ can be written with regards to the frame elements, so

$$|x(t)| = \left| \sum_{n\in\Gamma} \langle x(\cdot), \bar{\phi}_n(\cdot)\rangle_{\mathcal{H}} \phi_n(t) \right| \tag{15}$$

$$\tag{16}$$

and consequently,

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

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

■

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

Theorem 3.2.2 *Let $\mathcal{H}$ be a separable 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 $K(s,t)$ given by*

$$K : \left| \begin{array}{l} \Omega \times \Omega \to \mathbb{R} \\ s \times t \to K(s,t) = \sum_{n\in\Gamma} \bar{\phi}_n(s)\phi_n(t) \end{array} \right. \tag{18}$$

proof Any $x \in \mathcal{H}$ can be expanded by means of the frame of $\mathcal{H}$, thus

$$\begin{aligned} x(t) &= \sum_{n\in\Gamma} \langle x, \bar{\phi}_n \rangle \phi_n(t) \\ &= \left\langle x(\cdot), \sum_{n\in\Gamma} \bar{\phi}_n(\cdot)\phi_n(t) \right\rangle \end{aligned} \tag{19}$$

and besides, the property given in (12) holds, so

$$\forall x \in \mathcal{H}, \quad \forall t \in \Omega, \qquad x(t) = \langle x(\cdot), K(\cdot, t) \rangle \tag{20}$$

As the space $\mathcal{H}$ is separable, the Reproducing Kernel can be expressed as :

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

Hence, by identifying equation (19) and (20)

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

and thus, we can conclude that

$$\alpha_n(s) = \bar{\phi}_n(s)$$

and

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

■

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

3.3 Frame expansion and Reproducing Kernel expansion

From the previous paragraph, we can deduce that a frameable Hilbert Space with frame and dual frame elements satisfying equation (13) can be used as a framework for a regularized approximation, and thus, the solution of equation (1) with the norm in $\mathcal{H}$ as a stabilizer, has the form of linear combination of the Reproducing Kernel. However, one can show that using a frame expansion of the solution is equivalent to the Reproducing Kernel expansion. In fact, one have :

$$\begin{aligned} f^*(x) &= \sum_{i=1}^{N} c_i K(x_i, x) \\ &= \sum_{i=1}^{N} c_i \sum_{n\in\Gamma} \bar{\phi}_n(x_i)\phi_n(x) \\ &= \sum_{n\in\Gamma} \left(\sum_{i=1}^{N} c_i \bar{\phi}_n(x_i) \right) \phi_n(x) \\ &= \sum_{n\in\Gamma} d_n \phi_n(x) \end{aligned} \tag{21}$$

where $d_n = \sum_{i=1}^{N} c_i \bar{\phi}_n(x_i)$

This shows the equivalence between frame vectors expansion and kernel expansion. And thus in a frameable Hilbert Space, solutions of the regularized approximation have two forms. This may be useful for a better understanding of the solution of the regularization problem.

4 Approximation schemes using frame

In the previous section, condition for a frameable Hilbert Space being a RKHS were given. Here, we are interested in constructing a Hilbert Space together with its frame and discuss about the implications of such result in function approximation framework.

4.1 Approximation on frameable Hilbert Space

One of the most interesting point of frameable Hilbert Space is that under weak conditions, it becomes easy to build RKHS. The following theorem proves such point.

Theorem 4.1.1 *Let $\{\phi_n\}_{n=1\dots N}$ be a finite set of non-zero functions of a Hilbert Space $\mathcal{B}$ of $\mathbb{R}^\Omega$ so that :*

$$\forall n \;\; 1 \leq n \leq N, \qquad \|\phi_n\| < \infty$$

and

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

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

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

$(\mathcal{H}, \langle \cdot, \cdot \rangle_{\mathcal{B}})$ is a RKHS and its Reproducing Kernel is

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

Proof

Step 1 $\mathcal{H}$ is a Hilbert Space

This is straightforward as $\mathcal{H}$ is a closed subspace of a Hilbert Space $\mathcal{B}$, and is provided of $\mathcal{B}$ 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 it exists A and B satisfying equation (4).

Let us consider the non trivial case that $span\{\phi_n\}_{n=1..N} \neq 0$

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

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

thus,

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

thus by taking $B = \sum_{n=1}^{N} \|\phi_n\|^2$ $(B < \infty)$ satisfies the right-hand of the inequality (4).

Let $\mathcal{H}^* \triangleq \{f \in \mathcal{H} : \|f\|_{\mathcal{H}} > 0\}$ and $S(f)$ be the functional

$$S : \left| \begin{array}{lll} \mathcal{H}^* & \longrightarrow & \mathbb{R} \\ f & \longrightarrow & \Omega(f) = \frac{1}{\|f\|^2} \sum_{n \in \Gamma} |\langle f, \phi_n \rangle|^2 \end{array} \right. \tag{22}$$

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

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

let A be $\sum_{n \in \Gamma} |\langle g, \phi_n \rangle|^2$. Hence $A > 0$, and as $\|g\| = 1$, one has :

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

Step 3 Now let's prove that $\mathcal{H}$ is a RKHS

for that it suffices to prove that the frame $\{\phi_n\}$ satisfies condition given in theorem 3.2.1

It is straightformard as $\{\phi_n\}_{n=1\cdots N}$ is a frame of $\mathcal{H}$ and owing to theorem 3.1.1, the dual frame is also a frame of $\mathcal{H}$. Hence, norm of each $\bar{\phi}_n$ is finite. Besides, $\phi_n(t)$ is supposed to be bounded. Thus

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

Thus, $\mathcal{H}$ is a RKHS with a kernel equals to :

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

■

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

Example 1 *Any finite set of bounded functions of $L^2(\Omega)$ spans a RKHS. For instance, the set of functions which expression are given below spans a RKHS.*

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

Example 2 *Any finite set of bounded function belonging to Sobolev space spans a RKHS. The set of functions, given in previous example, spans also a RKHS in a Sobolev inner product sense.*

Example 3 *Consider a finite set of wavelet*

$$\left\{\psi_{j,k}(t) = \frac{1}{\sqrt{a^j}}\psi\left(\frac{t - nu_0a^j}{a^j}\right), j \in [j_{min}, j_{max}], k \in [k_{min}, k_{max}]\right\}$$

where $(a, u_0) \in \mathbb{R}^2$, and $(j_{min}, j_{max}, k_{min}, k_{max}) \in \mathbb{Z}^4$, then the span of these functions endowed with L^2 inner product is a RKHS. Figure (1) and (2) plot an example of wavelet frame elements and their dual frame elements for a dilation $j = -7$.

The main interest in this theorem is the flexibility that it allows in the choice of the RKHS and in the functions which generate the hypothesis space function for learning. Conversely to some definite positive kernel for which we do not have any knowledge about the basis functions, here, the kernel is built from these basis functions. This way of creating RKHS provides advantages and drawbacks as discussed in the next section.

5 Discussions

Propositions presented here describe a way for easily building RKHS and its associate Reproducing Kernel. Hence, the space or the kernel can be used within the framework or regularization networks or SVM for function approximation.
For SVM, 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 the Mercer condition :

$$\int_\Omega \int_\Omega k(u,v)x(u)x(v)dudv \geq 0 \tag{23}$$

for all $x \in L^2(\Omega)$.

Now, one may ask what are the advantages and inconvenients in using kernel built by means of theorem 4.1.1.

- Both Mercer condition and frameable RKHS allows to obtain a definite positive function. However, it is obvious that conditions for having frameable RKHS is easier to verify than Mercer 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 4 *Consider the set of functions on* $\mathbb{R}$ $\{\phi_n(t) = \frac{sin(t-n)}{t-n}\}_{n=1...N}$. *The space spanned by these frame elements associated to* $L^2(\mathbb{R})$ *inner product form an RKHS. Thus, as a direct corollary of theorem 4.1.1, the kernel*

  $$k(x,y) = \sum_{i=1}^{N} \bar{\phi}_i(x)\phi_i(y)$$

  is an admissible kernel for SVM.

  This conclusion is not so straightforward using Mercer's condition.

- As the condition for obtaining a frameable RKHS holds mainly for finite dimensional space (although, it may exists infinite dimensional Hilbert space which frame elements satisfy condition 3.2.1), it is fairest to compare the frameable kernel to a finite dimensional kernel. According to Mercer condition, or other more detailed papers on the subject (Aronszajn 1950, Wahba 2000), This latter can be expanded as follows :

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

  where s and t belongs to Ω, λ_l is a positive real number and $\{\psi_l\}_{i=l..N}$ is a set of orthogonal functions. The condition for constructing frameable kernel is 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 same expansion as noted above, since dual frame elements is equal to frame elements up to a multiplicative constant.

- Using a frame-based kernel, for instance in SVM, allows easier capacity control capability. Indeed, using a large (or low) number of non redundant frame elements increases (or decreases) the capacity control of the set of approximating functions. Hence, removing high capacity frame elements (i.e highly oscillating elements) from the kernel expansion is likely to have beneficial effects since the data will be approximated by the lower capacity function and thus, the solution will be flatter in the feature space.

- Besides, since the kernel has an expansion with regards to the frame elements, the solution of equation (1) can be more understandable. In fact, the solution depends on the kernel expression but can be rewritten as a linear combination of the frame elements. Thus, compared to other kernels for which basis functions remain unknown (e.g the gaussian kernel), using frame-based kernel increases interpretability of the data model.

- Drawbacks of using frame-based kernel lie mainly in the temporal complexity burden that is added for constructing the data model. For both SVM 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 the appendix). This, by its own, may be time-consuming. But besides, the construction of the matrix K, need the processing of the sum. Hence, if the number M of frame elements describing the kernel and the number N of data are large, building K becomes rapidly very time-consuming (of an order of $M^2 \cdot N^2$)

These points are arguments that may suggest that frame based kernels can be useful by themselves. However, within the context of semiparametric estimation, this flexibility for building kernel offers some interesting perspectives. Semiparametric estimation can be introduced by the following theorem :

Theorem 5.0.2 *(Kimerdolf & Wahba 1971)*
Let $\mathcal{H}_K$ be a RKHS of real valued functions on Ω with reproducing kernel K. Denote by $\{(x_i, y_i), i = 1 \dots n\}$ the training set and let $\{g_j, j = 1 \dots m\}$ 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

$$\min_{f \in span(g)+h, h \in \mathcal{H}_K} \frac{1}{n} \sum_{i=1}^{n} C(y_i, f(x_i)) + \lambda \|f\|^2_{\mathcal{H}_K} \tag{24}$$

has a representation of the form

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

The solution of this problem can be interpreted as a semiparametric estimation as 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 allows to exploit some prior knowledge on the estimation problem (e.g major properties of the data are described by linear combination of a small set of functions), and making a "good" guess (on this set) can have a large effect on performance.

Again in this context, the flexibility of frame based kernel can be exploited. In fact, let $G = \{g_i\}_{i=1\dots n}$ be a set of n linearly independent functions that satisfy theorem 4.1.1, hence, any subset of G, $\{g_i\}_{i\in\Gamma}$, Γ being an index set of size $n_o < n$ can be used for building a RKHS $\mathcal{H}_K$ while the remaining vectors can be used in the parametric part of the Kimerdolf-Wahba Theorem. Hence in this case, the solution of (24) is written

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

The flexibility comes from the fact that, in the approximation problem, any elements of G can be regularized (if involved in the span of $\mathcal{H}_K$) or be kept as it is (if used in the parametric part). Intuitively, one should move any vector that comes from "good" prior knowledge, in the parametric part of the approximation while leaves in the kernel expansion the others frame elements. Notice also that only the subset of G which is used in the parametric part has to be linearly independent.

Another perspective which follows directly from this findings is a new technique of regularization that we call multiscale regularization which is inspired from the Multi Resolution Analysis of Mallat (Mallat 1998). Here, we just sketch the idea behind this concept and in no way, the following paragraph should be considered as a complete study of this new technique as the analysis of all its properties goes beyond the scope of this paper. Consider the same problem as the one described in theorem 5.0.2. Now, suppose that $\{g_i\}$ is a set of N linearly independent functions verifying theorem (4.1.1). Let $\{\Gamma_i\}_{i=0\dots m}$ be a set of index set such that $\{\cup_{i=0}^{m}\Gamma_i\} = \{1\dots n\}$. By subdividing the set $\{g_i\}$ by means of the index set $\{\Gamma_i\}_{i=0\dots m}$, one can construit m RKHS $\{\mathcal{F}_i\}_{i=0\dots m-1}$ in such a way that

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

and reproducing kernel of $\mathcal{F}_i$ is noted K_i. Now, denotes 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 $\mathcal{H}_0 = span\ \{g_k\}_{k\in\Gamma_0}$. By construction, the space $\mathcal{H}_i$ are nested spaces :

$$\mathcal{H}_0 \subset \mathcal{H}_1 \subset \dots \subset \mathcal{H}_{m-1}$$

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, as $\mathcal{H}_i = \mathcal{H}_{i-1} + \mathcal{F}_{i-1}$, one can think of $\mathcal{F}_i$ as

the details needed to be added to $\mathcal{H}_i$ to obtain $\mathcal{H}_{i+1}$, thus we will call $\mathcal{F}_i$ as the "details" space whereas $\mathcal{H}_i$ are the "trend" space. Any of this space $\mathcal{F}_i$ and $\mathcal{H}_i$ are a RKHS as any subset of $\{g_i\}$ satisfies theorem (4.1.1).

Multiscale regularization is an iterative technique that consists at step $k = 1 \dots m$ to look for

$$f_{m-k}(\cdot) = \arg \min_{f \in \mathcal{H}_{m-k+1}} \frac{1}{n} \sum_{i=1}^{n} C(y_{i,m-k}, f(x_i)) + \lambda_{m-k} \|f\|^2_{\mathcal{F}_{m-k}} \tag{25}$$

which expression is :

$$f_{m-k}(\cdot) = \sum_{i=1}^{n} c_{i,m-k} K_{m-k}(x_i, \cdot) + \sum_{j \in \cup_{l=0}^{m-k} \Gamma_l} d_{j,m-k} g_j(\cdot) \tag{26}$$

where $y_{i,m-1} = y_i$, $y_{i,m-(k+1)} = y_{i,m-k} - \sum_{j=1}^{n} c_{j,m-k} K_{m-k}(x_j, x_i)$ and, the solution of the so-called multiscale regularization is :

$$\hat{f}(\cdot) = \sum_{k=1}^{m} \sum_{i=1}^{n} c_{i,m-k} K_{m-k}(x_i, \cdot) + \sum_{j \in \Gamma_0} d_{j,0} g_j(\cdot) \tag{27}$$

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 substracted to the data and one tries to approximate these residuals 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) shows an example of how the algorithm works for a 3 level approximation. Illustrations of this technique is given in the next section.

6 Numerical Experiments

This section describes some experiments that compare frame-based kernels to classical one (*e.g* gaussian kernel) in some simulated approximation 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 behaviour of different kernels using regularization networks and support vector regression. The function to be approximated is

$$f(x) = \sin x + sinc(\pi(x-5)) + sinc(5\pi(x-2)) \tag{28}$$

where $\text{sinc}(x) = \frac{\sin x}{x}$. Data used for the approximation is corrupted by an additive noise, thus $y_i = f(x_i) + \epsilon_i$ where ϵ_i is a gaussian noise of standard deviation 0.2 . Points x_i are

Table 1: True Generalization Error for Gaussian, Wavelet, Sin/Sinc Kernels with Regularization Networks and Support Vector Regression for the best hyperparameters.

	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

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 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 span 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 $\phi_i(x) = \{1, \sin(x), \cos(x), sinc(j\pi(x - k)) : j \in \{1, 3, 6\}, k \in [0 \dots 10])\}$.

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

For both regularization network and Support Vector Regression, some hyperparameters have to be tuned. Different approachs are available for solving this model selection problem. In this study, we have run the experiment a 100 time, and for each dataset, the true generalization error has been evaluated for a range of finely sampled values of hyperparameters. Then, averaging is done over all the datasets.

Table 1 depicts the generalization error evaluated on 200 datapoints for the two learning machines and the different kernels using the best hyperparameter setting found by the cross validation procedure. Figure 5 shows the contour or plot of the mean-square error with regards to the regularization parameter's value. Analysis of this table and figure leads to the following observations :

- The best performance has been achieved by regularization networks associated with a gaussian kernel. Parameters of the model are the following : $\lambda = 0.3$, $\sigma = 0.5$.

- Others kernels give results of similar order. Using prior knowledge on the problem in this context does not give better performance (Sin/Sin kernel or wavelet kernel compared to gaussian kernel). A justification of this observation can be that, such kernel uses strong prior knowledge (the sin frame element) that is included in the kernel expansion and thus, get regularized as much as other frame elements that are of higher frequency. This suggest that semiparametric regularization should be more appropriate to get advantage of such a kernel.

- As expected, regularization networks gives better results than SVM regardless to the kernels as a proper loss function is used with regards to the noise distribution.

6.2 Experience 2

In this experience, 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 $\left\{\psi_{j,k}(x) = \frac{1}{\sqrt{a^j}}\psi\left(\frac{x - k u_0 a^j}{a^j}\right), j \in \{0, 5\}\right\}$
- Wavelet kernel and wavelet basis functions : the latter are the same as the previous kernel, whereas 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 (experience 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 :

$$k(x, y) = \sum_{i \in \Gamma} \bar{\phi}_i(x)\phi_i(y)$$

with $\phi_i(x) = \{sinc(j\pi(x - k)) : j \in \{3, 6\},\ k \in [0 \dots 10]\}$

and the basis functions are $\{1, \sin x, \cos x, sinc(\pi(x - k) : k \in [1 \dots 10]\}$.

For each kernel, model selection has been solved by crossvalidation using 50 datasets. The results of this step are depicted in figure 6. Then, after having spotted the best hyperparameters, the experiment has been 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

Table 2: Generalization performance for semiparametric regression networks for differents setting of kernel and basis functions. The number in parentheses reflects the number of trials for which the model has been the best model.

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

the gaussian kernel and sin basis functions. Comparing the different results lead to the following remarks :

- First of all, we can note that exploiting prior knowledge on the function to be approximated leads immediatly 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 using the kernel of the RKHS spanned by the "bad" prior knowledge for the approximation problem seems to be fruitful in this case. (The frame elements $sinc(3\pi(x-k))$ and $sinc(6\pi(x-k))$ 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 correspond 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 results in poorer performance. Thus, prior knowledge on structure, which may be easiest to get than prior knowledge on basis function, may be easily exploited by means of wavelet span and wavelet kernel.

- Analysis of Mean Square Error with regards to hyperparameters on figure (6) shows that the gaussian kernel and sin basis function approximation is very sensitive to hyperparameter tuning.

- Other kernels and basis functions result in a approximation performance varying up to 30% within the explored range of hyperparameters value compared to the 300% variation for the gaussian kernel.

6.3 Experience 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. Both the first two methods, a gaussian kernel is used whereas for the two latter, wavelet kernel and basis functions are taken. The true functions used for benchmarking are the following :

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

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 $\{x_i, f(x_i) + \epsilon_i\}$. Here again, a range of finely sampled values of hyperparameters has been tested for the model selection. In each case, an averaging of the true error generalization over 100 datasets 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

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

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

$$\begin{aligned} \mathcal{H}_0 &= \text{span}\left\{\frac{1}{\sqrt{a^j}} \psi\left(\frac{t - ku_0a^j}{a^j}\right), j = 5\right\} \\ \mathcal{F}_0 &= \text{span}\left\{\frac{1}{\sqrt{a^j}} \psi\left(\frac{t - ku_0a^j}{a^j}\right), j = 0\right\} \\ \mathcal{F}_1 &= \text{span}\left\{\frac{1}{\sqrt{a^j}} \psi\left(\frac{t - ku_0a^j}{a^j}\right), j = -10\right\} \end{aligned}$$

These dilation parameters have been set in a *ad hoc* way, but their choices can be justified by the following reasoning : Three disctinct levels has been used for separating the approximation in three structures which should be smooth($j = 5$), irregular ($j = 0$) and highly irregular ($j = -10$), and dilation levels are strictly related to frequency contents (Mallat 1998). The same values of j has been 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 $j_{SPL} = 5$.

Figure (8) and (9) depict the mean square error during the cross validation processing while Table 3 presents the average of the mean-square error of the different learning machines for the two functions and for the best hyperparameters value.

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

	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

Comments and analysis of this experiment validating the concept of multiscale approximation are :

- From table 3, notice that semiparametric 2 and Multiscale approximation gives 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 to be regularized. In this experiment, it seems that leaving the space spanned by wavelet of dilation $j = 0$ leads to overfitting and thus to degradation of performance.

- Analysis of figure (12) and (13) 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.

- Multiscale approximation is able to catch all the structures of the signal (see figure (10) and (11)). 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 $sinc(3x)$ function whereas the sin is located on the highest dilation $j = 5$.

- Figure 8 and 9 suggest that semiparametric and multiscale algorithm seem to be less dependent to hyperparameters setting. Infact, they lead to acceptable performance on a wide range of hyperparameter values.

7 Conclusions

In this paper, we showed that a RKHS can be defined by its frame elements and conversely, one can construct a RKHS from a frame. One of the key result is that the space spanned by linear combination of appropriated L^2 functions is a RKHS with a kernel that can be at least numerically described. Hence, this is another method 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 semiparametric approach with frame-based improve 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 SVM kernels, is not as fruitful as one may have expected. However, depending on the prior knowledge on the problem, one can build appropriate kernel that can enhance further 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 :

- 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, Soltani 1999). In fact, redundancy allows to "attenuate" 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 hence, a higher temporal complexity of the algorithm. Thus, optimal algorithms still have to be derived.

8 Appendix

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 define the operator S

$$S : \left| \begin{array}{ccc} \mathcal{H} & \longrightarrow & \mathcal{H} \\ f & \longrightarrow & \sum_{n\in\Gamma}\langle f, \phi_n\rangle\phi_n \end{array} \right. \tag{29}$$

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 has proposed an algorithm to compute the problem $f = S^{-1}g$ (Grochenig 1993). 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 8.0.1 *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 define by induction,

$$\lambda_n = \frac{\langle r_n, p_n\rangle}{\langle p_n, Sp_n\rangle} \tag{30}$$

$$f_{n+1} = f_n + \lambda_n p_n \tag{31}$$

$$r_{n+1} = r_n - \lambda_n Sp_n \tag{32}$$

$$p_{n+1} = Sp_n - \frac{\langle Sp_n, Sp_n\rangle}{\langle p_n, Sp_n\rangle}p_n - \frac{\langle Sp_n, Sp_{n-1}\rangle}{\langle p_{n-1}, Sp_{n-1}\rangle}p_{n-1} \tag{33}$$

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

$$\|f - f_n\|_S \leq \frac{2\sigma^n}{1+2\sigma^n}\|f\|_S \tag{34}$$

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

Proof Only some steps of the proof are highlighted here. For the complete proofs, one should refer to Grochenig.

Step 1 Let U_n be the subspace generated by $\{S^j f\}_{1\leq j\leq n}$. By induction on n, one derives from (33) that $p_j \in U_n$, for $j < n$.

Step 2 By defining the inner product in U_n as $\langle f, g\rangle_S = \langle f, Sg\rangle$. From this, by induction, one proves that $\{p_j\}_{0<j<n}$ is an orthonormal basis of U_n. Then, assuming that $\langle p_n, Sp_j\rangle = 0,$ for $j \leq n-1$, it can be shown that $\langle p_{n+1}, Sp_j\rangle = 0$ for $j \leq n$.

step 3 Then, one shows that f_n is the orthogonal projection of f onto U_n, with respect to the inner product $\langle \cdot, \cdot\rangle_S$, by proving that $\langle f - f_n, p_j\rangle_S = 0,$ for $j < n$, as $f_n \in U_n$.
As a consequence, one has

$$\forall h \in U_n, \qquad \|f - h\|_S \leq \|f - f_n\|_S$$

step 4 One computes the orthogonal projection of f in embedded spaces U_n of dimension n, and then obtain the following $\lim_{n\to+\infty} f_n = f$.

■

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

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

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

Boser, B. E., Guyon, I. M. & Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers, *in* D. Haussler (ed.), *5th Annual ACM Workshop on COLT*, ACM Press, Pittsburgh, PA, pp. 144–152.

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

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

Canu, S. (n.d.). Apprentissage et approximation : les techniques de regularisation.

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

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

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

Evgeniou, T., Pontil, M. & Poggio, T. (2000). Regularization networks and support vector machines, *Advances in Computational Mathematics* **13**(1): 1–50.

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

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

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

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

Kimerdolf, G. & Wahba, G. (1971). Some results on Tchebycheffian spline functions., *J. Math. Anal. Applic.* **33**: 82–95.

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

Niyogi, P., Girosi, F. & Poggio, T. (1998). Incorporating prior information in machine learning by creating virtual examples, *Proceedings of the IEEE*, Vol. 86, pp. 2196–2209.

Scholkopf, B., Simard, P. Y., Smola, A. J. & Vapnik, V. (1998). Prior knowledge in support vector kernels, *in* M. I. Jordan, M. J. Kearns & S. A. Solla (eds), *Advances in Neural information processings systems*, Vol. 10, MIT Press, Cambridge, MA, pp. 640–646.

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

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

Soltani, S. (1999). *Application de la transformée en ondelettes en reconnaissances de formes*, PhD thesis, Univ. Tech. Compiegne.

Tikhonov, A. & Arsénin, V. (1977). *Solutions of Ill-posed problems*, W.H. Winston, Washington D.C.

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

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

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

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

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

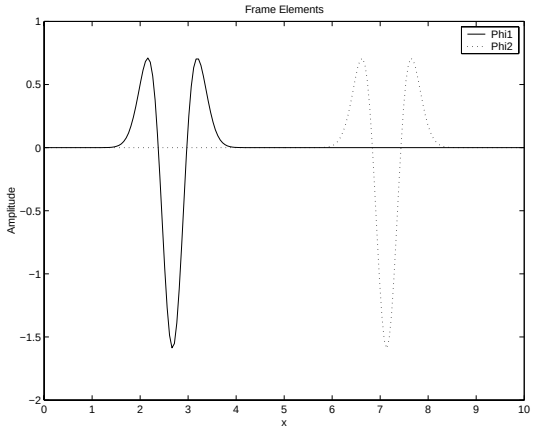


Figure 1: Examples of wavelet frame elements.

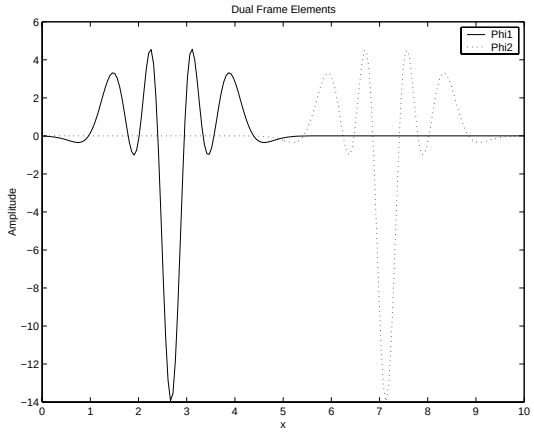


Figure 2: Examples of wavelet dual frame elements.

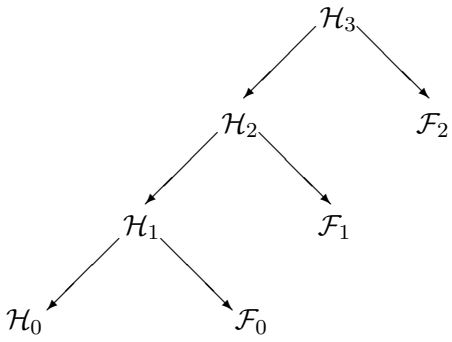


Figure 3: Example of multiscale approximation on 3 levels

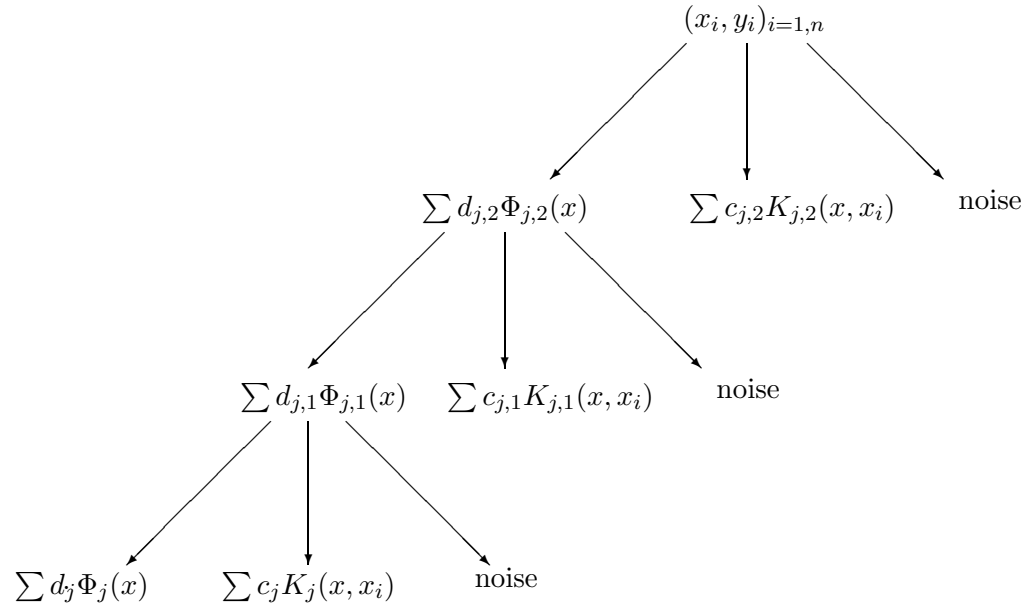


Figure 4: Example of multiscale approximation on 3 levels : the kernel point of view

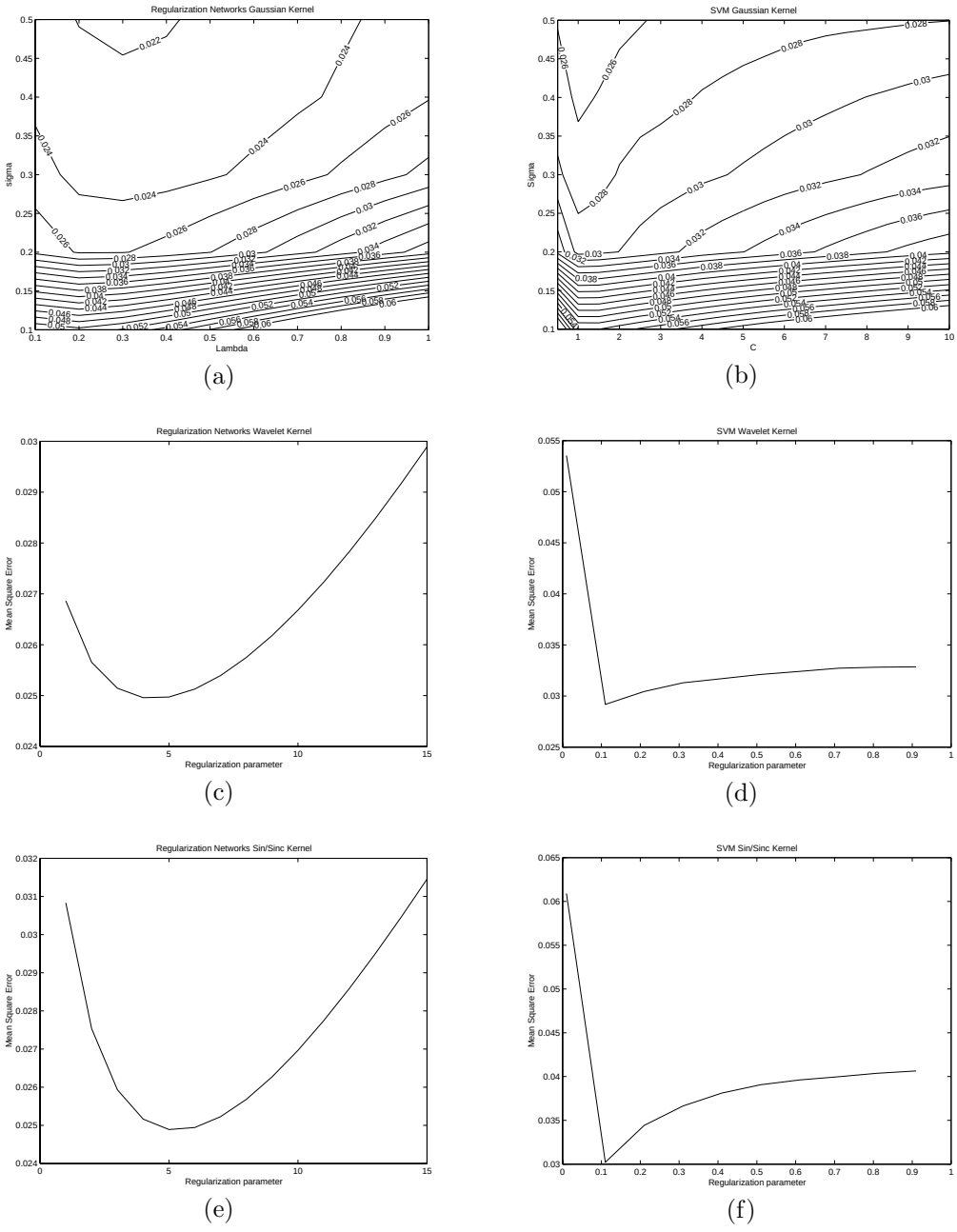


Figure 5: Generalization Mean Square Error with regards to model parameters for different kernels and learning machines. (a) Regularization Networks with gaussian kernel. (b) Support Vector Regression with gaussian kernel.(c) Regularization Networks with wavelet kernel. (d) Support Vector Regression with wavelet kernel.(e) Regularization Networks with sin/sinc kernel. (f) Support Vector Regression with sin/sinc kernel.

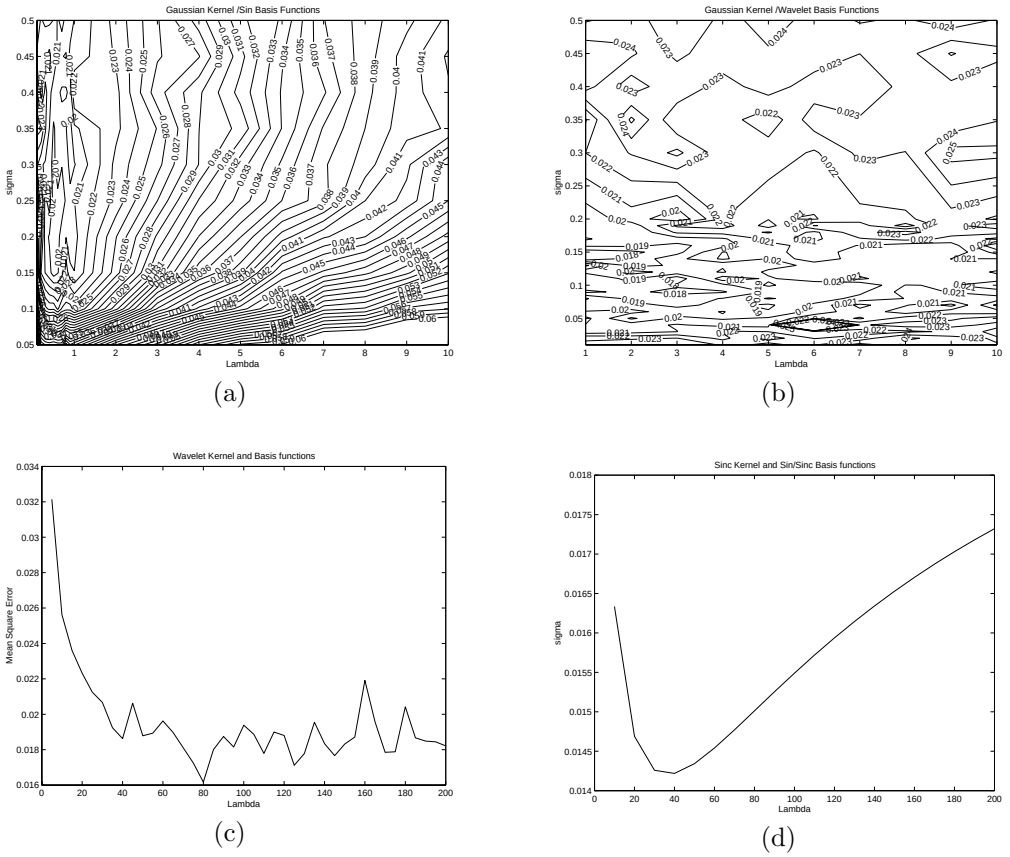
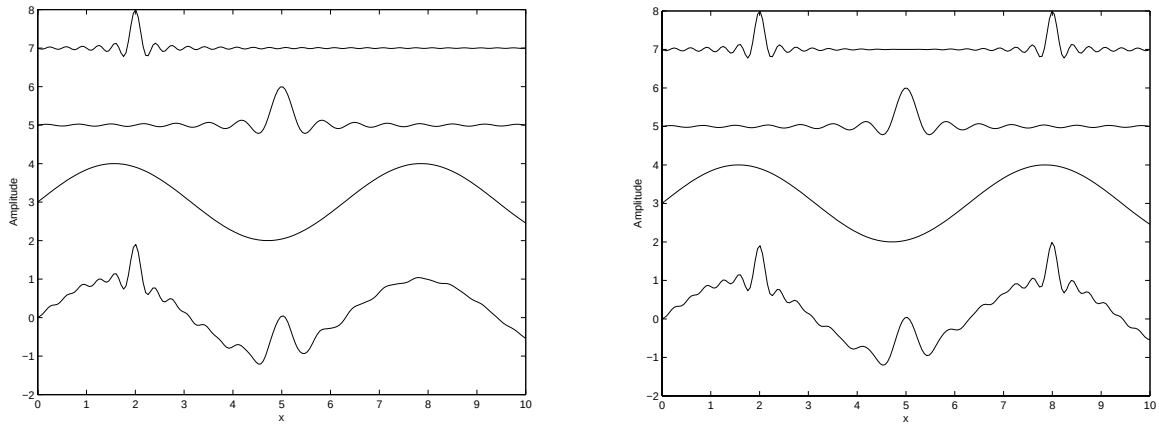


Figure 6: True generalization Mean Square Error for semiparametric regularization networks. (a) Gaussian kernel and Sin basis functions (b) Gaussian kernel and Wavelet Basis functions. (c) wavelet kernel and basis functions. (d) Sinc Kernel and Sin/Sinc basis functions



(a) (b)

Figure 7: Original functions used for benchmarking in experience 3. (a) f_1 (b) f_2. Top : multiscale structure on 3 levels. Bottom : Complete function.

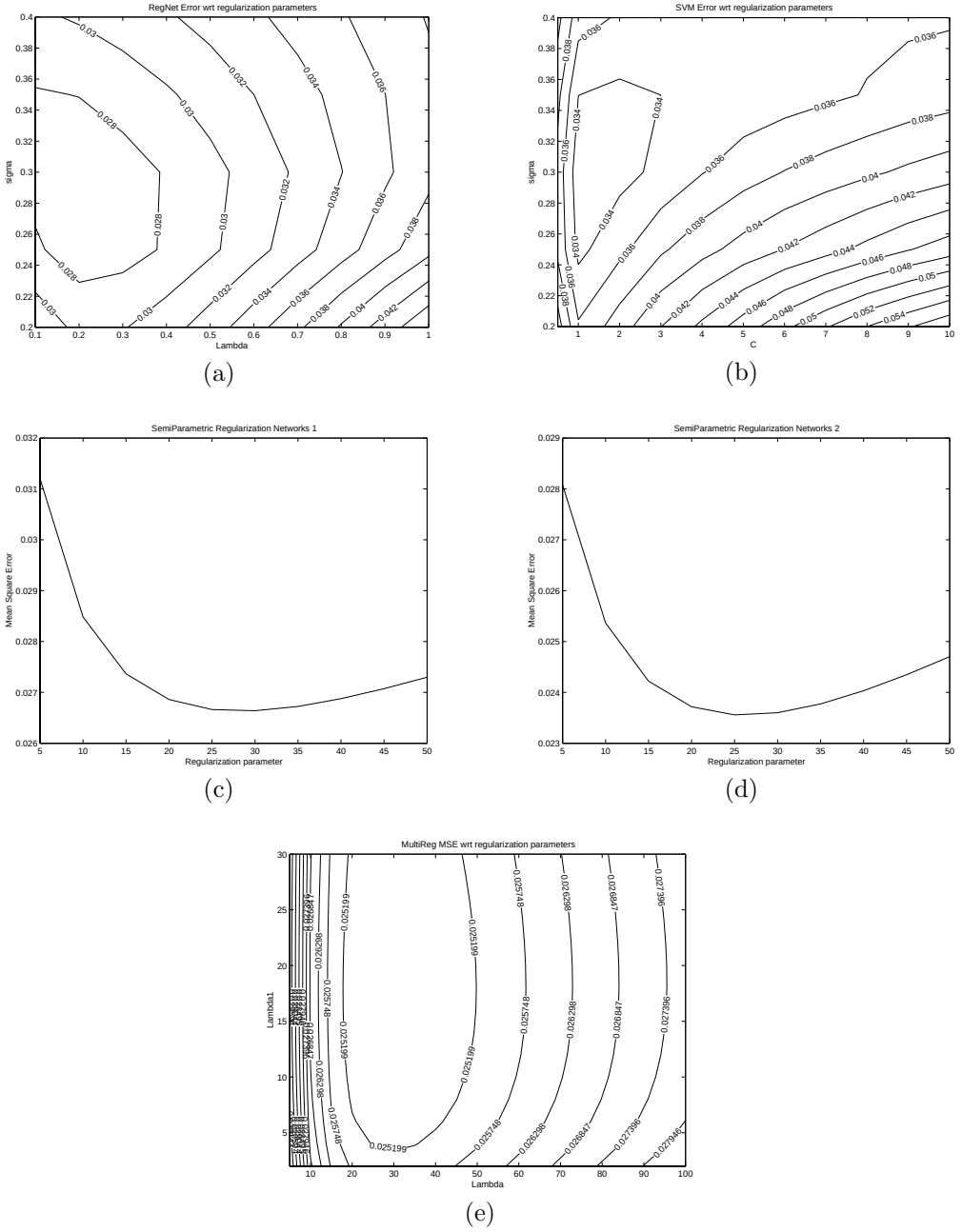


Figure 8: True generalization Mean Square Error for approximation of f_1 wrt hyperparameters. (a) Regularization Networks with Gaussian kernel (b) Support Vector Machines with Gaussian kernel. (c) Semiparametric Regularization networks with $j_{SPH} = -10$ and $j_{SPL} = 0, 5$ (d) Semiparametric Regularization networks with $j_{SPH} = -10, 0$ and $j_{SPL} = 5$. (e) Multiscale Regularization with $j_0 = 5$, $j_1 = 0$ and $j_2 = -10$

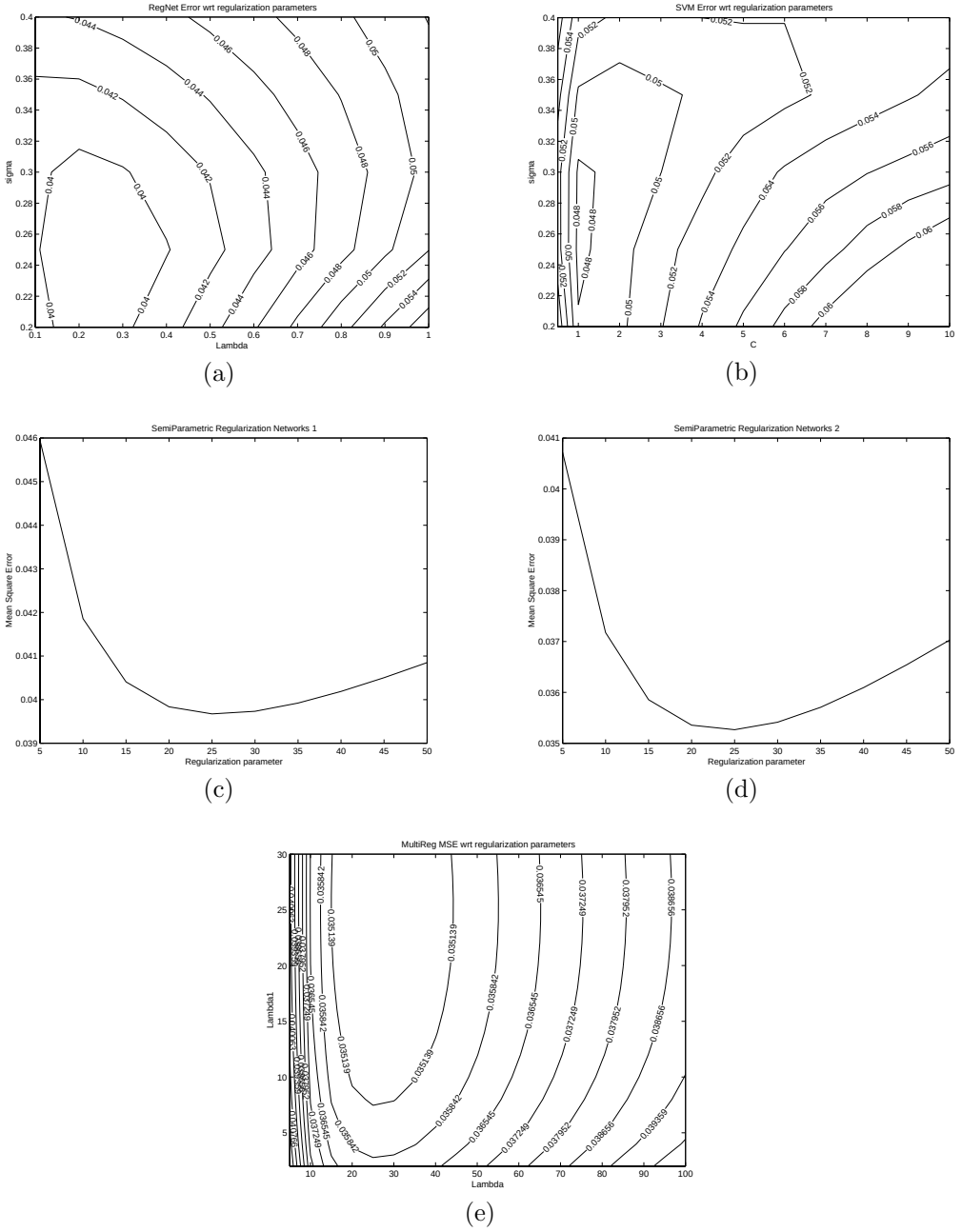


Figure 9: True generalization Mean Square Error for approximation of f_2 wrt hyperparameters. (a) Regularization Networks with Gaussian kernel (b) Support Vector Machines with Gaussian kernel. (c) Semiparametric Regularization networks with $j_{SPH} = -10$ and $j_{SPL} = 0, 5$ (d) Semiparametric Regularization networks with $j_{SPH} = -10, 0$ and $j_{SPL} = 5$. (e) Multiscale Regularization with $j_0 = 5$, $j_1 = 0$ and $j_2 = -10$

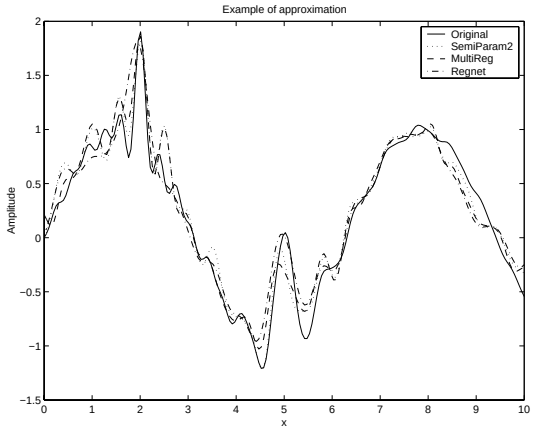


Figure 10: Examples of approximation of f_1 : Original function (Solid), Semiparametric 2 (dotted), Multiscale regularization (dashed), Regularization network (dash-dotted)

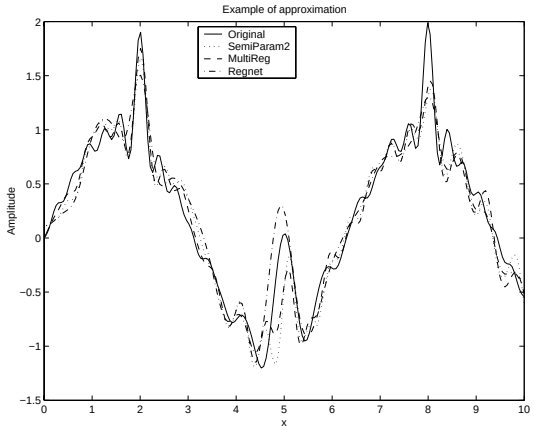


Figure 11: Examples of approximation of f_2 : Original function (Solid), Semiparametric 2 (dotted), Multiscale regularization (dashed), Regularization network (dash-dotted

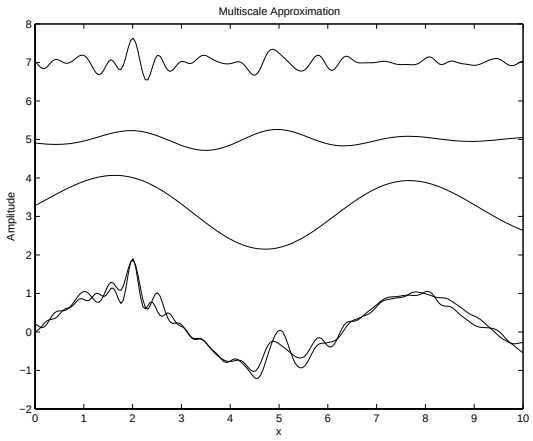


Figure 12: Top : Multiscale structure of a typical prediction of f_1 by multiscale wavelet approximation Bottom : full approximation and true function

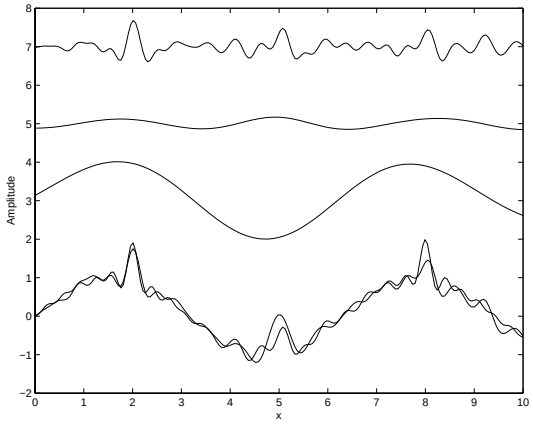


Figure 13: Top : Multiscale structure of a typical prediction of f_1 by multiscale wavelet approximation Bottom : full approximation and true function