Journal of Machine Learning Research 9 (2008) 1559-1582 Submitted 10/07; Revised 2/08; Published 7/08

Aggregation of SVM Classifiers Using Sobolev Spaces

Sébastien Loustau LOUSTAU@CMI.UNIV-MRS.FR
Laboratoire d'Analyse, Topologie et Probabilités (UMR CNRS 6632)
Université Aix-Marseille 1
CMI-39 rue Joliot-Curie 13453 Marseille, France

Editor: John Shawe-Taylor

Abstract

This paper investigates statistical performances of Support Vector Machines (SVM) and considers the problem of adaptation to the margin parameter and to complexity. In particular we provide a classifier with no tuning parameter. It is a combination of SVM classifiers.

Our contribution is two-fold: (1) we propose learning rates for SVM using Sobolev spaces and build a numerically realizable aggregate that converges with same rate; (2) we present practical experiments of this method of aggregation for SVM using both Sobolev spaces and Gaussian kernels.

Keywords: classification, support vector machines, learning rates, approximation, aggregation of classifiers

1. Introduction

We consider the binary classification setting. Let $\mathcal{X} \times \{-1,1\}$ be a measurable space endowed with P an unknown probability distribution on $\mathcal{X} \times \{-1,1\}$. Let $D_n = \{(X_i,Y_i), i = 1,\dots n\}$ be n realizations of a random variable (X,Y) with law P (in the sequel we also write P_X for the marginal distribution of X). Given this training set D_n, the goal of Learning is to predict class Y of new observation X. In other words, a classification algorithm builds a decision rule from $\mathcal{X}$ to $\{-1,1\}$ or more generally a function f from $\mathcal{X}$ to $\mathbb{R}$ where the sign of $f(x)$ determines the class of an input x.

The efficiency of a classifier is measured by the *generalization error*

$$R(f) := \mathbb{P}(\text{sign}(f(X)) \neq Y),$$

where sign(y) denotes the sign of $y \in \mathbb{R}$ with the convention sign$(0) = 1$. A well-known minimizer over all measurable functions of the generalization error is called the *Bayes rule*, defined by

$$f^*(x) := \text{sign}(2\eta(x) - 1)$$

where $\eta(x) := \mathbb{P}(Y = 1|X = x)$ for all $x \in \mathcal{X}$. Unfortunately, the dependence of f^* on the unknown conditional probability function η makes it uncomputable in practice.

A natural way to overcome this difficulty is to provide an empirical decision rule or classifier based on the data D_n. It has to mimic the Bayes. The way one measures the efficiency of a classifier $\hat{f}_n := \hat{f}_n(D_n)$ is via its *excess risk*:

$$R(\hat{f}_n, f^*) := R(\hat{f}_n) - R(f^*), \tag{1}$$

where here $R(\hat{f}_n) := \mathbb{P}(\text{sign}(\hat{f}_n(X)) \neq Y|D_n)$. Given P, we hence say that a classifier $\hat{f}_n$ is consistent if the expectation of (1) with respect to $P^{\otimes n}$ (the distribution of the training set) goes to zero as n goes to infinity. Finally, we can look for a way of quantifying this convergence. A classifier $\hat{f}_n$ learns with rate $(\psi_n)_{n\in\mathbb{N}}$ if there exists an absolute constant $C > 0$ such that for all integer n,

$$\mathbb{E}R(\hat{f}_n, f^*) \leq C\psi_n, \tag{2}$$

where in the sequel $\mathbb{E}$ is the expectation with respect to $P^{\otimes n}$. Of course (2) ensures consistency of $\hat{f}_n$ whenever (ψ_n) goes to zero with n.

It has been shown in Devroye (1982) that no classifier can learn with a given rate for all distributions P. However several authors propose different rates reached by restricting the class of joint distributions. Pionneering works of Vapnik (Vapnik and Chervonenkis, 1971, 1974) investigate the statistical procedure called Empirical Risk Minimization (ERM). The ERM estimator consists in searching for a classifier that minimizes the empirical risk

$$R_n(f) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{\text{sign}(f(X_i))\neq Y_i\}}, \tag{3}$$

over a class of prediction rules $\mathcal{F}$, where $\mathbb{1}_A$ denotes the indicator function of the set A. If we suppose that the class of decision rules $\mathcal{F}$ has finite VC dimension, ERM reaches the parametric rate $n^{-\frac{1}{2}}$ in (2) when f^* belongs to the class $\mathcal{F}$. Moreover, if P is noise-free (i.e., $R(f^*) = 0$), the rate becomes n^{-1}. This is a fast rate.

More recently, Tsybakov (2004) describes intermediate situations using a margin assumption. This assumption adds a control on the behavior of the conditional probability function η at the level $\frac{1}{2}$ (see (10) below). Under this condition, Tsybakov (2004) gets minimax fast rates of convergence for classification with ERM estimators over a class $\mathcal{F}$ with controlled complexity (in terms of entropy). These rates depend on two parameters : the margin parameter and the complexity of the class of candidates f^* (see also Massart and Nédélec, 2006). Another study of the behavior of ERM is presented in Bartlett and Mendelson (2006).

It is well known, however, that minimizing (3) is computationally intractable for many non trivial classes of functions (Arora et al., 1997). It comes from the non convexity of the functional (3). It suggests that we must use a convex surrogate Φ for the loss. The main idea is to minimize an empirical Φ-risk

$$A_n^\Phi(f) = \frac{1}{n}\sum_{i=1}^{n} \Phi(Y_i f(X_i)),$$

over a class $\mathcal{F}$ of real-valued functions. Then $\hat{f}_n = \text{sign}(\hat{F}_n)$ where $\hat{F}_n \in \text{Arg}\min_{f\in\mathcal{F}} A_n^\Phi(f)$ has a small excess risk. Recently a number of methods have been proposed, such as boosting (Freund, 1995) or Support Vector Machines. The statistical consequences of choosing a convex surrogate is well treated by Zhang (2004) and Bartlett et al. (2006). In this paper it is proposed to use the hinge loss $\Phi(v) = (1-v)_+$ (where $(\cdot)_+$ denotes the positive part) as surrogate, that is, to focus on the SVM algorithm.

SVM was first proposed by Boser et al. (1992) for pattern recognition. It consists in minimizing a regularized empirical Φ-risk over a Reproducing Kernel Hilbert Space (RKHS for short in the sequel). Given a training set D_n, the SVM optimization problem without offset can be written:

$$\min_{f\in\mathcal{H}_K} \left(\frac{1}{n}\sum_{i=1}^{n} l(Y_i, f(X_i)) + \alpha_n \|f\|_K^2\right), \tag{4}$$

where in the sequel:

1. The functional l is called the hinge loss and is now written $l(y, f(x)) = (1 - yf(x))_+$. The first term of the minimization (4) is then the empirical Φ-risk A_n^Φ for $\Phi(v) = (1 - v)_+$.

2. The space $\mathcal{H}_K$ is a RKHS with reproducing kernel K. Under some mild conditions over K, it consists of continuous functions from $\mathcal{X}$ to $\mathbb{R}$ or $\mathbb{C}$ with the reproducing property:

$$\forall f \in \mathcal{H}_K, \forall x \in \mathcal{X}, f(x) =< K(x, \cdot), f >_{\mathcal{H}_K} .$$

   Recall that every positive definite kernel has an essentially unique RKHS (Aronszajn, 1950).

3. The sequence α_n is a decreasing sequence that depends on n. This smoothing parameter has to be determined explicitly. Such a problem will be studied in this work.

4. The norm $\|.\|_K$ is the norm associated to the inner product in the Hilbert space $\mathcal{H}_K$.

For a survey on this kernel method we refer to Cristianini and Shawe-Taylor (2000).

This algorithm is at the heart of many theoretical considerations. However, its good practical performances are not yet completely understood. The study of statistical consistency of the algorithm and approximation properties of kernels can be found in Steinwart (2001) or more recently in Steinwart (2005). Blanchard et al. (2006) propose a model selection point of view for SVM. Finally, several authors provide learning rates to the Bayes for SVM (Wu and Zhou, 2006; Wu et al., 2007; Steinwart and Scovel, 2007, 2005). In these papers, both approximation power of kernels and estimation results are presented. Wu and Zhou (2006) state slow rates (logarithmic with the sample size) for SVM using a Gaussian kernel with fixed width. It holds under no margin assumption for Bayes rule with a given regularity. Steinwart and Scovel (2007) give, under a margin assumption, fast rates for SVM using a decreasing width (which depends on the sample size). An additional geometric hypothesis over the joint distribution is necessary to get a control of the approximation using Gaussian kernels.

These results focus on SVM using Gaussian kernels. The goal of this work is to clarify both practical and theoretical performances of the algorithm using two different classes of kernels. In a first theoretical part, we consider a family of kernels generating Sobolev spaces as RKHS. It gives an alternative to the extensively studied Gaussian kernels. We quantify the approximation power of these kernels. It depends on the regularity of the Bayes prediction rule in terms of Besov space. Then under the margin assumption, we give learning rates of convergence for SVM using Sobolev spaces. It holds by choosing optimally the tuning parameter α_n in (4). This choice strongly depends on the regularity assumption over the Bayes and the margin assumption. As a result, it is non-adaptive. Then we turn out into more practical considerations. Following Lecué (2007a), we give a procedure to construct directly from the data a classifier with similar statistical performances. It uses a method called aggregation with exponential weights. Finally, we show practical performances of this aggregate and compare it with a similar classifier using Gaussian kernels and results of Steinwart and Scovel (2007).

The paper is organized as follows. In Section 2, we give statistical performances of SVM using Sobolev spaces. Section 3 presents the adaptive procedure of aggregation and show the performances of the data-dependent aggregate. This procedure does not damage the learning rates stated in Section 2. We show practical experiments in Section 4 and conclude in Section 5 with a discussion. Section 6 is devoted to the proofs.

2. Statistical Performances

As a regularization procedure, minimization (4) generates two types of errors: the estimation error and the approximation error. The use of a finite sample size produces the estimation error. The approximation error can be seen as the distance between the hypothesis space and the Bayes decision rule. It comes from the use of a RKHS of continuous functions in the minimization whereas the Bayes is not continuous. The first one is random and depends on the fluctuation of the training set. The second one is deterministic and depends on the size of the RKHS. We can see coarsely that these errors are antagonist. Theorem 7 gives a choice of the regularization parameter α_n that makes the trade-off between these two errors.

For the estimation error, we will state an oracle-type inequality of the form :

$$\mathbb{E}R_l(\hat{f}_n, f^*) \leq C \inf_{f\in\mathcal{H}_K} \left(R_l(f, f^*) + \alpha_n \|f\|_K^2\right) + \varepsilon_n, \tag{5}$$

where $R_l(f, f^*) := \mathbb{E}_P l(Y, f(X)) - \mathbb{E}_P l(Y, f^*(X))$ is the excess l-risk of f. The term ε_n must be a residual term and satisfies:

$$\varepsilon_n \leq C' \inf_{f\in\mathcal{H}_K} \left(R_l(f, f^*) + \alpha_n \|f\|_K^2\right),$$

where $C' > 0$. Inequality (5) deals with the estimation error. It depends on the complexity of the class of functions $\mathcal{H}_K$ and the difficulty of the problem.

Hence it remains to control the infimum in the right hand side (RHS for short) of (5). Steinwart and Scovel (2007) define the approximation error function as:

$$a(\alpha_n) := \inf_{f\in\mathcal{H}_K} \left(R_l(f, f^*) + \alpha_n \|f\|_K^2\right). \tag{6}$$

This function represents the theoretical version of the empirical minimization (4). It depends on the chosen $\mathcal{H}_K$ and the behaviour of α_n as a function of n.

Using this approach, Steinwart and Scovel (2007) study the statistical performances of SVM minimization (4) with the parametric family of Gaussian kernels. For $\sigma \in \mathbb{R}$, we define the Gaussian kernel $K_\sigma(x,y) = \exp\left(-\sigma^2\|x-y\|^2\right)$ on the closed unit ball of $\mathbb{R}^d$ (denoted $\mathcal{X}$). The parameter σ^{-1} is called the width of the Gaussian kernel. In this paper, under a margin assumption and a geometric assumption over the distribution, they state fast learning rates for SVM. These rates hold under some specific choices of tuning parameters recalled in Sect. 4. Following Lecué (2007a), we will use this result and more precisely these choices of tuning parameters to implement the aggregate using Gaussian kernels.

2.1 Sobolev Smooth Kernels

We propose to deal with other class of kernels than the Gaussian kernels. First we need to introduce some notations. Let us consider the set of complex-valued and integrable (resp. square-integrable) functions on $\mathbb{R}^d$ denoted as $L^1(\mathbb{R}^d)$ (resp. $L^2(\mathbb{R}^d)$). On this set, we define the Fourier transform of f to be:

$$\hat{f}(\omega) = \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} f(t) e^{-i\omega.t} dt, \forall \omega \in \mathbb{R}^d,$$

where $x.y$ denotes the usual scalar product of $\mathbb{R}^d$ between two points $x,y \in \mathbb{R}^d$.

After the usual extension from $L^1(\mathbb{R}^d)$ to $L^2(\mathbb{R}^d)$ with Plancherel, this operator is an isometry on $L^2(\mathbb{R}^d)$. It allows us to define, for any $s \in \mathbb{R}^+$, the Sobolev space $\mathcal{W}_s^2$ (often called fractional Sobolev space) as the following subspace of $L^2(\mathbb{R}^d)$ (Malliavin, 1974):

$$\mathcal{W}_s^2 := \{f \in L^2(\mathbb{R}^d) : \|f\|_s^2 = \int_{\mathbb{R}^d} |\hat{f}(\omega)|^2 (1+\|\omega\|^2)^s d\omega < \infty\}. \tag{7}$$

We refer to Triebel (1992) or Adams (1975) for a large study of this well-known functional space. With such a norm, $\mathcal{W}_s^2$ is a Hilbert space endowed with the inner product defined as:

$$< f,g >_s = \int_{\mathbb{R}^d} \hat{f}(\omega)\overline{\hat{g}(\omega)}(1+\|\omega\|^2)^s d\omega,$$

where $\bar{z}$ is the complex conjugate of z in $\mathbb{C}$. Moreover it is a Hilbert space of continuous functions for any $s > \frac{d}{2}$ (due to the embedding between $\mathcal{W}_s^2$ and $C(\mathbb{R}^d)$ for any $s > \frac{d}{2}$). It can be seen as a RKHS.

In this framework, a kernel is a symmetric and positive definite function $K : \mathbb{R}^d \times \mathbb{R}^d \mapsto \mathbb{C}$. For $r \in \mathbb{R}^+$, a kernel K_r will be called *Sobolev smooth kernel* with exponent $r > d$ if the associated RKHS $\mathcal{H}_{K_r}$ is such that

$$\mathcal{H}_{K_r} = \mathcal{W}_{\frac{r}{2}}^2,$$

where $\mathcal{W}_{\frac{r}{2}}^2$ is defined in (7). The restriction $r > d$ ensures that the RKHS consists of continuous functions from $\mathbb{R}^d$ to $\mathbb{C}$. Corollary 2 provides a way of constructing such a kernel.

We say that a kernel K is a *translation invariant kernel* (or RBF kernel), if for all $x,y \in \mathbb{R}^d$,

$$K(x,y) = \Phi(x-y) \tag{8}$$

for a given $\Phi : \mathbb{R}^d \mapsto \mathbb{C}$. Function Φ is often called RB function for Radial Basis function. The most popular example of translation invariant kernel is the Gaussian kernel $K_\sigma(x,y) = \exp(-\sigma^2\|x-y\|^2)$. This kernel is not a Sobolev smooth kernel (see below).

Under suitable assumptions on Φ, the following theorem gives a Fourier representation of a RKHS associated to a translation invariant kernel. The proof is given in Section 6.

Theorem 1 *Let $K : \mathbb{R}^d \times \mathbb{R}^d \mapsto \mathbb{C}$ be a translation invariant kernel where in* (8) *Φ belongs to $L^1(\mathbb{R}^d) \cap L^2(\mathbb{R}^d)$ and such that $\widehat{\Phi}$ is integrable. Then the RKHS associated to K can be written*

$$\mathcal{H}_K = \{f \in L^2(\mathbb{R}^d) : \|f\|_K^2 = \frac{1}{(2\pi)^{d/2}} \int_S \frac{|\hat{f}(\omega)|^2}{\widehat{\Phi}(\omega)} d\omega < \infty \text{ and } \hat{f} = 0 \text{ on } \mathbb{R}^d \backslash S\}$$

with the inner product

$$< f,g >_K = \frac{1}{(2\pi)^{d/2}} \int_S \frac{\hat{f}(\omega)\overline{\hat{g}(\omega)}}{\widehat{\Phi}(\omega)} d\omega,$$

where $S := \{\omega \in \mathbb{R}^d : \widehat{\Phi}(\omega) \neq 0\}$ is the support of $\widehat{\Phi}$.

Sufficient conditions to have a Sobolev smooth kernel are:

Corollary 2 *Let K satisfying assumptions of Theorem 1. Suppose moreover that there exist constants $C, c > 0$ and a real number $s > \frac{d}{2}$ such that*

$$\widehat{\Phi}(\omega) = \frac{C}{(c + \|\omega\|^2)^s}, \forall \omega \in \mathbb{R}^d. \tag{9}$$

Then K is a Sobolev smooth kernel with exponent $r = 2s > d$.

In Section 5 we propose an example of Sobolev smooth kernel and use it into the SVM procedure.

Remark 3 (Gaussian kernels are not Sobolev smooth) *Theorem 1 can be used to define Gaussian kernels in terms of Fourier transform. Indeed, the Gaussian kernel defined above is a translation invariant kernel with RB function $\Phi(x) = \exp(-\sigma^2 \|x\|^2)$. Its Fourier transform is given by*

$$\widehat{\Phi}(\omega) = \frac{1}{(\sqrt{2}\sigma)^d} \exp\left(-\frac{\|\omega\|^2}{4\sigma^2}\right).$$

Then Φ satisfies assumptions of Theorem 1. The Fourier representation of $\mathcal{H}_\sigma$ is given by:

$$\mathcal{H}_\sigma = \{ f \in L^2(\mathbb{R}^d) : \int_{\mathbb{R}^d} |\hat{f}(\omega)|^2 \sigma^d \exp\left(\frac{\|\omega\|^2}{4\sigma^2}\right) d\omega < \infty \}.$$

From definition (7), it is clear that $\mathcal{H}_\sigma$ is not a Sobolev space. This integral representation of a Gaussian RKHS illustrates the smoothness of functions $f \in \mathcal{H}_\sigma$. Indeed we can see trivially that $\mathcal{H}_\sigma \subset \mathcal{H}_{K_r}$ for any fixed $\sigma, r > 0$ (because the Fourier transform of Φ is rapidly decreasing in this case). Moreover the parameter σ can be seen as a regularization parameter : the fewer is σ, the smoother are the functions in $\mathcal{H}_\sigma$. More precisely, $\sigma < \sigma'$ entails $\mathcal{H}_\sigma \subset \mathcal{H}_{\sigma'}$.

2.2 Approximation Efficiency of Sobolev Smooth Kernels

Here we are interested in approximation properties of $\mathcal{H}_{K_r}$. We aim at bounding the approximation function $a(\alpha_n)$ defined in (6) for the procedure (4). The best case appears when $f^* \in \mathcal{H}_K$. Then we get coarsely $a(\alpha_n) \leq C\alpha_n$ where C is an absolute constant. This case is not realizable considering a continuous RKHS since the Bayes classifier is not. In this paper, we get a control of the approximation function when f^* does not belong to the RKHS. Theorem 4 provides such a result using a Sobolev smooth kernel.

Theorem 4 *Consider the approximation function $a(\alpha_n)$ defined in (6), with Sobolev smooth kernel K_r such that $r > 2s > 0$. Suppose P_X satisfies $\frac{dP_X}{dx} \leq C_0$.*

Then if $f^* \in \mathcal{B}^2_{s,\infty}(\mathbb{R}^d)$, we have:

$$a(\alpha_n) \leq C_0^{\frac{r-2s}{r-s}} \|f^*\|_{s2\infty}^{\frac{r}{r-s}} \alpha_n^{\frac{s}{r-s}},$$

where $\|.\|_{s2\infty}$ defines the norm in the Besov space $\mathcal{B}^2_{s,\infty}(\mathbb{R}^d)$.

The proof is detailed in Section 6 where we define explicitly Besov spaces $\mathcal{B}^2_{s,\infty}(\mathbb{R}^d)$.

Remark 5 *(*BAYES REGULARITY*) Here we get a control of the approximation function under an assumption on the smoothness of the Bayes classifier. Of course large values of s are not possible because $f^*(x) = \operatorname{sign}(2\eta(x) - 1)$ is not even continuous (except for the trivial case $\eta(x) < \frac{1}{2}$ a.s. or $\eta(x) > \frac{1}{2}$). More precisely, the Besov space $\mathcal{B}^p_{s,q}(\mathbb{R}^d)$ is included in the space of continuous functions for $s > \frac{d}{p}$ and $q > 1$. Here $p = 2$ then parameter s must satisfy $s < \frac{d}{2}$ to have $f^* \in \mathcal{B}^2_{s,\infty}(\mathbb{R}^d)$. In Remark 10 we give an example of Bayes rule verifying this smoothness assumption.*

Remark 6 *(*COMPARISON WITH STEINWART AND SCOVEL, 2007*) Steinwart and Scovel (2007) propose a same type of result using Gaussian kernels. Under a geometric assumption over the distribution, they get*

$$a(\alpha_n) \leq C\alpha_n^{\frac{\gamma}{\gamma+1}},$$

where γ is the geometric noise exponent. Here we propose a same type of result under a regularity assumption over the possible f^*. Theorem 17 in Section 6 shows that this result can be generalized to any other kernel, using interpolation spaces.

2.3 Learning Rates

In this work, we restrict the class of considered distributions P. We add a control on the local slope of the conditional probability function η at the level $\frac{1}{2}$. This margin assumption (we often call $|\eta - \frac{1}{2}|$ the margin) is originally due to Mammen and Tsybakov (1999) for discriminant analysis. We will use throughout this paper the following formulation: we say that P has *margin parameter* $q > 0$ if there exists a constant $c_0 > 0$ such that

$$\mathbb{P}(|2\eta(X) - 1| \leq t) \leq c_0 t^q, \tag{10}$$

for all sufficiently small t.

According to Boucheron et al. (2005), this hypothesis is equivalent to the low noise or margin assumption in Tsybakov (2004). Best situation for learning appears when the conditional probability makes a jump at the level $\frac{1}{2}$. Hence (10) holds true for any positive q. It corresponds to a margin parameter $q = +\infty$, that is, $\kappa = 1$ in the sense of Tsybakov (2004).

Finally, last step of modelling consists in clipping the solution of minimization (4). For any classifier $\hat{f}$, we hence define the *clipped version* $\hat{f}^C$ with values in $[-1, 1]$ by

$$\hat{f}^C(x) = \begin{cases} -1 \text{ for } x : \hat{f}(x) < -1, \\ \\ f(x) \text{ for } x : \hat{f}(x) \in [-1, 1], \\ \\ 1 \text{ for } x : \hat{f}(x) > 1. \end{cases}$$

This operation does not modify the classification property of $\hat{f}$ since $\operatorname{sign}(\hat{f}) = \operatorname{sign}(\hat{f}^C)$. It produces classifiers with bounded norm $\|.\|_\infty$. It appears in several works (Bartlett, 1998; Steinwart et al., 2007). We stress that the clip does not modify the algorithm. It is done after the training as a part of the theoretical study of the algorithm. We are now on time to state the main result of this section.

Theorem 7 *Let P be a distribution over $\mathbb{R}^d \times \{-1,1\}$ such that P_X satisfies $\frac{dP_X}{dx} \leq C_0$ and* (10) *holds for $q \in [0,+\infty]$. Let $s > 0$ and suppose $f^* \in \mathcal{B}^2_{s,\infty}(\mathbb{R}^d)$.*

Consider the SVM minimization (4) *with Sobolev smooth kernel K_r, with $r > 2s \vee d$, built on the i.i.d. sequence $(X_i,Y_i), i = 1 \dots n$ according to P.*

If we choose α_n such that

$$\alpha_n \sim n^{-\frac{r(r-s)(q+1)}{s(r(q+2)-d)+d(r-s)(q+1)}}, \tag{11}$$

then there exists a constant C which depends on r, s, d, c_0, q and C_0 such that

$$\mathbb{E}R(\hat{f}_n^C, f^*) \leq Cn^{-\gamma(q,s)},$$

where

$$\gamma(q,s) = \frac{rs(q+1)}{s(r(q+2)-d)+d(r-s)(q+1)}. \tag{12}$$

The proof of this theorem is given in Section 6.

Remark 8 *(*FAST RATES*) Rate* (12) *is a fast rate (i.e., faster than $n^{-\frac{1}{2}}$) if $\frac{rs(q+1)}{s(r(q+2)-d)+d(r-s)(q+1)} > \frac{1}{2}$. In particular, for $q = +\infty$, it corresponds to $s > \frac{rd}{r+d}$. The presence of fast rates depends on the regularity of the Bayes classifier. Unfortunately the behaviour of f^* (see Remark 4) entails $s < \frac{d}{2}$. As a result, $\frac{sr}{sr+d(r-s)} < \frac{1}{2}$ and fast rates can not be reached.*

Remark 9 *(*COMPARISON WITH STEINWART AND SCOVEL, 2007*) This theorem gives performances of SVM using a fixed kernel. On the contrary, according to Steinwart and Scovel (2007), the bandwidth of the kernel has to be chosen as a function of n. Nevertheless, rates of convergences are fast for sufficiently large geometric noise parameter. Here we cannot get fast rates for reasonable assumption over f^*.*

Remark 10 *(*OPTIMAL SMOOTHING PARAMETER*) Theorem 7 provides a particular choice of α_n to reach rates* (12)*. Other definitions for the sequence α_n give other rates of convergence. We only mention the best possible rates. It holds for a regularization parameter optimizing the statistical performances. Indeed, α_n in* (11) *makes the balance between the estimation error and the approximation error.*

Remark 11 (EXAMPLE) *Consider the one-dimensional case where $\mathcal{X} = \mathbb{R}$. Suppose f^* is such that:*

$$card\{x \in \mathbb{R} : f^* \text{ jumps at } x\} = N < \infty. \tag{13}$$

It means that the Bayes rule changes only a finite number of times over the real line. Using standard analysis, we get

$$\|f^*\|_{TV} = \int_{\mathbb{R}} |Df^*(x)|dx = 2N$$

where Df^* is the generalized derivative of f^*. Moreover, for any f, $|\hat{f}(\omega)| \leq \|f\|_{VT}/|\omega|$. Then f^* belongs to $\mathcal{W}_{s,2}$ only for $s < 1/2$. Finally, with basic properties of Besov spaces (Triebel, 1992), we have $\mathcal{W}_{s,2} = \mathcal{B}^2_{s,2} \subset \mathcal{B}^2_{s,\infty}$.

Consequently, f^* verifying (13) *belongs to $\mathcal{B}^2_{s,\infty}$ for any $s < \frac{1}{2}$. If we consider a margin parameter $q = +\infty$, we hence cannot reach the rate of convergence*

$$n^{-\frac{r}{3r-1}}$$

which corresponds to a regularity $s = \frac{1}{2}$ in the Besov space. Then the SVM using Sobolev smooth kernel H_{K_r} with $r > 1$ cannot learn with fast rate in this simple case.

3. Aggregation

Theorem 7 provides the optimal value of α_n to reach rates of convergence (12) in the context of Sobolev spaces. It holds under two ad-hoc assumptions: a margin assumption over the distribution and a regularity assumption over the Bayes rule. Hence the choice of the smoothing parameter depends on two unknown parameters: the margin parameter q and the exponent s in the Besov space. Consequently the classifier $\hat{f}_n$ of Theorem 7 cannot be constructed from the data. It is called non-adaptive.

The goal of this section is to overcome this difficulty. We propose a classifier that adapts automatically both to the margin and to regularity. In other words, we will build a decision rule from D_n which does not depend on the unknown parameters s and q. Moreover, Theorem 12 shows that this procedure of adaptation will not damage the learning rates of Theorem 7.

We use a technique called aggregation (Nemirovski, 1998; Yang, 2000). We apply the method presented in Lecué (2007a) to our framework of Sobolev smooth kernel. It consists of splitting the data into two parts : the first part in used to construct a family of classifiers. The second part is used to make a convex combination of these classifiers. We obtain an adaptive decision rule which mimics the best one over the family. Let us first describe the method.

Denote $D^1_{n_1}$ (resp. $D^2_{n_2}$) the first subsample of size n_1 (resp. second subsample of size n_2) with $n_1 + n_2 = n$. The choice of n_1 and n_2 will be discussed later. We construct a set of classifiers $(\hat{f}^{\alpha}_{n_1})_{\alpha \in \mathcal{G}(n_2)}$ defined by $\hat{f}^{\alpha}_{n_1} = sign\left(\hat{F}^{\alpha}_{n_1}\right)$ where

$$\hat{F}^{\alpha}_{n_1} := \arg\min_{f \in \mathcal{H}_{K_r}} \left(\frac{1}{n_1} \sum_{i=1}^{n_1} l(Y_i, f(X_i)) + \alpha \|f\|^2_K \right).$$

The grid $\mathcal{G}(n_2)$ is defined by

$$\mathcal{G}(n_2) := \left\{ \alpha_k = n_2^{-\phi_k} : \phi_k = \frac{1}{2} + k\Delta^{-1}, k = 0, \ldots, \lfloor \frac{(2r-d)\Delta}{2d} \rfloor \right\},$$

with $\Delta = n_2^b$ for some $b > 0$. We hence have $\left\lfloor \frac{(2r-d)\Delta}{2d} \right\rfloor + 1$ classifiers to aggregate.

The procedure of aggregation uses the second subsample $D^2_{n_2}$ to construct a convex combination with exponential weights. Namely, the aggregate $\tilde{f}_n$ is defined by

$$\tilde{f}_n = \sum_{\alpha \in \mathcal{G}(n_2)} \omega^{(n)}_{\alpha} \hat{f}^{\alpha}_{n_1}, \tag{14}$$

where

$$\omega^{(n)}_{\alpha} = \frac{\exp\left(\sum_{i=n_1+1}^{n} Y_i \hat{f}^{\alpha}_{n_1}(X_i)\right)}{\sum_{\alpha' \in \mathcal{G}(n_2)} \exp\left(\sum_{i=n_1+1}^{n} Y_i \hat{f}^{\alpha'}_{n_1}(X_i)\right)}.$$

We hence have the following result.

Theorem 12 *Consider the classifier $\tilde{f}_n$ defined in* (14) *where $n_2 = \lceil a\frac{n}{\log n}\rceil$ for $a > 0$. Let K a compact of $(0,\infty)^2$. Then there exists a constant C which depends on r,d,c_0,K,a,b,L and C_0 such that for all $(q,s) \in K$*

$$\sup_{P\in Q_{q,s}} \mathbb{E}R(\tilde{f}_n, f^*) \leq Cn^{-\gamma(q,s)},$$

where

$$\gamma(q,s) = \frac{rs(q+1)}{s(r(q+2)-d)+d(r-s)(q+1)}$$

and $Q_{q,s}$ is the set of distributions P satisfying $\frac{dP_X}{dx} < C_0$, (10) *with parameter q and such that $f^* \in \mathcal{B}^2_{s,\infty}(\mathbb{R}^d,L) = \{f \in \mathcal{B}^2_{s,\infty}(\mathbb{R}^d) : \|f\| \leq L\}$.*

Remark 13 *Same rates as in Theorem 7 are attained. Here we deal with an implementable classifier. In Section 5 we sum up practical performances of this aggregate.*

Remark 14 *Instead of aggregating a power of n classifiers, only* $\log n$ *classifiers are enough to obtain this result. Lecué (2007b) states an oracle inequality such as* (22) *without any restriction on the number of estimators to aggregate.*

Remark 15 *(*AVERAGE OF AGGREGATES*) This method supposes, for a given n_1 and n_2, an arbitrary choice for the subsample $D^1_{n_1}$ and $D^2_{n_2}$. However we can use different splits of the training set. We get an average of aggregates, namely*

$$\overline{f}_n = \frac{1}{M}\sum_{k=1}^{M} \tilde{f}^k_n.$$

It does not depend on a particular split. Each $\tilde{f}^k_n$ is defined in (14) *for the split number k. With (Lecué, 2007a, Theorem 2.4), this average satisfies the oracle inequality* (22)*. Then Theorem 12 holds for $\overline{f}_n$ for any family of M splits, for $M \leq C^{n_1}_n$.*

4. Practical Experiments

We now propose experiments illustrating performances of the aggregate of Section 3. We study SVM classifiers using both Sobolev spaces and Gaussian kernels. The aggregates were implemented in **R** using the free library *kernlab*. It contains implementations of support vector machines. For a description of this package for kernel-based learning methods in **R**, we refer to Karatzoglou et al. (2007). We use real world data sets from benchmark repository[1] used by Rätsch et al. (1998). We consider 9 data sets called ”Banana”, ”Titanic”, ”Thyroid”, ”Diabetes”, ”Breast-Cancer”, ”Flare-solar”, ”Heart”, ”Image” and ”Waveform”. These data sets are explained in Table 1. For each data set, we have several realizations of training and test set. The dimension of the input space is denoted by d whereas the number of observations for the training set is n. It follows the notations used in the previous sections. On each realization, we train and test our classifiers. The results presented in Table 2,3,4 show the average test errors over these realizations and the standard deviations.

[1] Data sets are available online at this address http://ida.first.fraunhofer.de/projects/bench/benchmarks.htm.

Data Set	d	n	test sample	realizations
Banana	2	400	4900	100
Titanic	3	150	2051	100
Thyroid	5	140	75	100
Diabetis	8	468	300	100
Breast-cancer	9	200	77	100
Flare-solar	9	666	400	100
Heart	13	170	100	100
Image	18	1300	1010	20
Waveform	21	400	4600	100

Table 1: Description of the data sets

4.1 SVM Using Sobolev Smooth Kernel

The first step is to pick up a Sobolev smooth kernel. Consider the following class of RBF kernels, with Radial Basis function Φ:

$$K(x,y)=\Phi(x-y)=\exp\left(-\sigma\|x-y\|\right),\forall\sigma\in\mathbb{R}. \tag{15}$$

For a given σ, this kernel is called a Laplacian kernel. It is clear that $\Phi\in L^1(\mathbb{R}^d)\cap L^2(\mathbb{R}^d)$. Recall the Fourier transform of $\Phi:\mathbb{R}^d\mapsto\mathbb{R}$ (see Williamson et al., 2001):

$$\widehat{\Phi}(\omega)=2^{\frac{d}{2}}\pi^{-\frac{1}{2}}\Gamma(\frac{d}{2}+1)\frac{\sigma}{(\sigma^2+\|\omega\|^2)^{\frac{d+1}{2}}},\forall\omega\in\mathbb{R}^d,$$

where $\Gamma(x)=\int_{\mathbb{R}^+}e^{-t}t^{x-1}dt$ is the Gamma function.

With Corollary 2, for any fixed σ, the Laplacian kernel defined in (15) is a Sobolev smooth kernel with exponent $r=d+1$. It satisfies assumptions of Theorem 7 and can be used in the implementation of the algorithm.

It is worth noticing that the parameter σ is constant. If we take a significantly small value for σ, as $\sigma=n^{-u}$, $u>0$, (9) holds for C and c depending on n. Thus Corollary 2 does not hold. To avoid this problem, we choose in our aggregation step using this class of kernels a constant $\sigma=5$. In the sequel the Laplacian kernel used is precisely $K(x,y)=\exp(-5\|x-y\|)$.

Table 2 shows the first experiments. For each realization of training set, we use previous section to build

- the set of classifiers $(\hat{f}_{n_1}^{\alpha})$ for α belonging to $\mathcal{G}(n_2)$;
- exponential weights $\omega_\alpha^{(n)}$ to deduce aggregate $\tilde{f}_n$.

Recall the definition of $\mathcal{G}(n_2)$ in this case:

$$\mathcal{G}(n_2):=\left\{\alpha_k=n_2^{-\phi_k}:\phi_k=\frac{1}{2}+k\Delta^{-1},k=0,\ldots,\lfloor\frac{(2r-d)\Delta}{2d}\rfloor\right\},$$

where $\Delta=n_2^b$. We take $b=1$ in the construction. Instead of a step $\Delta=n_2^b$, it is possible to take only $\Delta=\log n_2$ (see Remark 14) . The value of b governs the size of the grid. The cardinal is given in Table 2 for each data set. Note that growing b does not improve significantly the performances whereas it adds computing time. Indeed, whatever b, $\mathcal{G}(n_2)$ is contained in this case into $[n_2^{-\frac{d+1}{d}}, n_2^{-\frac{1}{2}}]$. This location is motivated by Theorem 7, namely equation (11).The value of b only deals with the distance between each point of $\mathcal{G}(n_2)$. It does not change the location of the grid.

Table 2 relates the average test errors and the standard deviations. We first collect the performances of the family of weak estimators $(\hat{f}^{\alpha}_{n_1}), \alpha \in \mathcal{G}(n_2)$. We mention in order the performances of the worst estimator, the mean over the family and the best over the family. It gives an idea of the estimators to aggregate. Then the performances of the aggregate using exponential weights are given in the last column.

Data Set	card$\mathcal{G}(n_2)$	max	mean	min	Laplace Aggregate
Banana	102	11.41±0.58	11.33± 0.57	11.12±0.59	11.31± 0.57
Titanic	38	22.80±1.16	22.80±1.14	22.77±1.13	22.77±1.13
Thyroid	31	5.97±2.61	5.45±2.56	4.77±2.63	5.45±2.68
Diabetis	72	29.56±2.03	28.40±2.00	27.33±1.96	28.34±2.27
Breast-cancer	35	35.10±5.34	33.26±5.06	31.49±5.05	32.74±5.16
Flare-solar	95	35.97±1.94	35.68±1.90	35.52±1.90	35.69±1.93
Heart	29	22.38±3.97	22.11±3.98	21.76±3.99	22.12±3.98
Image	152	4.35±0.87	4.06±0.74	3.79±0.74	3.95±0.74
Waveform	56	14.51±0.70	14.16±0.67	13.78±0.65	14.12±0.72

Table 2: Performances using Laplacian kernel

Note that the amplitude in the family is not very important. It may be explain by its construction. Indeed, $\mathcal{G}(n_2)$ is motivated by Theorem 7, which gives the location of the grid (see above). This family has a mathematical justification. The test errors of the aggregate are located between the average over the family and the oracle of the family.

A temperature parameter usually appears in aggregation methods. It governs the variations of values $\omega_{\alpha}^{(n)}$, for $\alpha \in \mathcal{G}(n_2)$. In Table 2 the weak classifiers have almost the same performances. This could explain why no temperature parameter is needed here.

4.2 SVM Using Gaussian Kernels

Here we focus on the parametric class of Gaussian kernels $K_{\sigma}(x,y) = \exp\left(-\sigma^2\|x-y\|^2\right)$, for $\sigma \in \mathbb{R}$. We build an aggregate made of a convex combination of Gaussian SVM classifiers. In this case, the construction is not exactly the same. It comes from Steinwart and Scovel (2007). In this paper, they introduce a geometric noise assumption. This hypothesis deals with the concentration of the measure $|2\eta - 1|P_X$ near the decision boundary. It allows to control the approximation function (6). According to Steinwart and Scovel (2007), suppose that the probability distribution P has a geometric noise $\gamma > 0$ and assumption (10) holds with margin parameter $q > 0$. Then if we choose

$$\alpha_n = \begin{cases} n^{-\frac{\gamma+1}{2\gamma+1}} \text{ if } \gamma \leq \frac{q+2}{2q}, \\ n^{-\frac{2(\gamma+1)(q+1)}{2\gamma(q+2)+3q+4}}, \text{ otherwise} \end{cases}$$

the solution of (4) using a Gaussian kernel K_σ with $\sigma = \alpha_n^{-\frac{1}{(\gamma+1)d}}$ learns with rates

$$\begin{cases} n^{-\frac{\gamma}{2\gamma+1}+\varepsilon} \text{ if } \gamma \leq \frac{q+2}{2q}, \\ n^{-\frac{2\gamma(q+1)}{2\gamma(q+2)+3q+4}+\varepsilon} \text{ otherwise,} \end{cases}$$

for all $\varepsilon > 0$.

We can see that the variance of the Gaussian kernels is not fixed. It has to be chosen as a function of the geometric noise exponent. As a result, parameter σ must be considered in the aggregation procedure, as the smoothing parameter α. It gives a two-dimensional grid of Gaussian SVM of the following form (Lecué, 2007a):

$$\mathcal{N}(n_2) = \left\{(\sigma_{n_2,\phi}, \alpha_{n_2,\psi}) = (n_2^{\phi/d}, n_2^{-\psi}) : (\phi, \psi) \in \mathcal{M}(n_2)\right\}$$

where

$$\mathcal{M}(n_2) = \left\{(\phi_{n_2,p_1}, \psi_{n_2,p_2}) = \left(\frac{p_1}{2\Delta}, \frac{p_2}{\Delta} + \frac{1}{2}\right) : p_1 = 1, \ldots, 2\lfloor\Delta\rfloor; p_2 = 1, \ldots, \lfloor\Delta/2\rfloor\right\},$$

for $\Delta = n_2^b$. Thus we have more classifiers to aggregate and needs more time to run. As a consequence, we choose constant $b = 0.5$ in our experiments. Such as the Sobolev case, the number of classifiers to aggregate is mentioned in Table 3 for each data set.

Table 3 relates the generalization performances of the classifiers over the test samples. We first give the performances of the family of Gaussian SVM (namely the worst, the mean and the oracle over the family). The performances of the aggregate using exponential weights are given in the last column.

Data Set	card$\mathcal{N}(n_2)$	max	mean	min	gaussian aggregate
Banana	100	17.29± 3.08	12.27±0.89	10.85±0.63	11.43±0.84
Titanic	36	23.15±1.30	22.81±1.00	22.49±0.78	22.57±0.79
Thyroid	36	8.19±2.63	6.76±2.72	5.59±2.94	6.31±2.97
Diabetis	100	29.82±1.98	28.19±1.84	26.39±1.85	27.80±2.06
Breast-cancer	42	34.83±5.12	32.76±4.82	30.48±4.61	32.13±4.77
Flare-solar	144	39.06±1.92	36.01±1.54	34.09±1.69	34.87±1.82
Heart	42	23.1±3.80	22.60±3.71	21.99±3.59	22.62±3.77
Image	256	7.79±1.00	6.33±0.83	5.30±0.73	5.66±0.74
Waveform	100	15.41±0.80	15.08±0.78	14.72±0.77	15.04±0.79

Table 3: Performances using Gaussian kernels

In this case the generalization errors in the family are more disparate. It comes from a two-dimensional grid of parameters. The performances of the Gaussian aggregate, as above, are located between the average of weak estimators and the best among the family.

4.3 Comparison With Rätsch et al. (1998)

Table 4 combines the performances of the aggregates using Laplacian kernel and Gaussian kernels. The errors are comparable. Gaussian kernels and Laplacian kernel lead to similar performances. Then we mention the generalization errors of Rätsch et al. (1998).

Rätsch et al. (1998) proposes generalizations of the original Adaboost algorithm. However, extensive simulations are presented like experimental results for SVM using Gaussian kernels. The choice of the parameters (α_n, σ) are done by 5-fold-cross validation thanks to several training data sets. This approach has not any mathematical justification. Moreover their mathematical programming problems are distributed over 30 computers. We only use last column to have an idea of reasonable average test errors for these data sets.

Data Set	Laplace Aggregate	Gaussian Aggregate	Rätsch et al. (1998)
Banana	11.31± 0.57	11.43±0.84	11.53±0.66
Titanic	22.77±1.13	22.57±0.79	22.42±1.02
Thyroid	5.45±2.68	6.31±2.97	4.80±2.19
Diabetis	28.34±2.27	27.80±2.06	23.53±1.76
Breast-cancer	32.74±5.16	32.13±4.77	26.04±4.74
Flare-solar	35.69±1.93	34.87±1.82	32.43±1.82
Heart	22.12±3.98	22.62±3.77	15.95±3.26
Image	3.95±0.74	5.66±0.74	2.96±0.6
Waveform	14.12±0.72	15.04±0.79	9.88±0.83

Table 4: Comparison with Rätsch et al. (1998).

Table 4 illustrates good resistance of our aggregates when the dimension is not too large. Nevertheless, in the last columns, our estimators fail. This may have a theoretical explanation. In Theorem 7 and 12, a constant C appears in the upper bounds. This constant in front of the rates of convergence depends on the dimension of the input space. Increasing d grows this constant C and may affect the performances. Moreover, the choice of the parameters in Rätsch et al. (1998) are done with several training sets. In our approach, for each realization of a training set, we construct an adaptive classifier using n observations. The amount of information used is not the same. It may also explain this difference.

5. Conclusion

This paper gives some insights into SVM algorithm, from both theoretical and practical point of view. We have tackled several important questions such as its statistical performances, the role of the kernel and the choice of the tuning parameters.

The first part of the paper focuses on the statistical performances of the method. In this study, we consider Sobolev smooth kernels as an alternative to the Gaussian kernels. It allows us to bring out a functional class of Bayes rule (namely Besov spaces $\mathcal{B}^2_{s,\infty}$) ensuring good approximation properties for our hypothesis space. Explicit rates of convergence have been given depending on the margin and the regularity (Theorem 7). Nevertheless, this result was non-adaptive.

Then it has been necessary to consider the problem of adaptation. The aggregation method appeared suitable in this context to construct directly from the data a competitive decision rule: it has the same statistical performances as the non-adaptive classifier (Theorem 12). In this procedure, we use explicitly the theoretical part to choose the scale of tuning parameters. For completeness, we have finally implemented the method and gave practical performances over real benchmark data sets. These practical experiments are to be considered as preliminary. However it shows similar performances for SVM using Gaussian or non-Gaussian kernel. Moreover it illustrates rather well the importance of constructing a classifier with some mathematical background.

6. Proofs

This section contains proofs of the results presented in this paper.

6.1 Proof of Theorem 1 and Corollary 2

We consider a translation invariant kernel $K : \mathbb{R}^d \times \mathbb{R}^d \mapsto \mathbb{C}$ with RB function Φ satisfying assumptions of Theorem 1. The following lemma will be useful.

Lemma 16 *For any $y \in \mathbb{R}^d$, consider the function $k_y : x \mapsto K(x,y)$ defined in $\mathbb{R}^d$. Then we have the following statements:*

1. *$k_y(x) = \overline{\hat{g}_y(x)}$ where $g_y(\omega) = e^{i\omega.y}\widehat{\Phi}(\omega)$.*
2. *$\hat{k}_y(\omega) = e^{-i\omega.y}\widehat{\Phi}(\omega)$.*

Proof

1. $\Phi \in L^2(\mathbb{R}^d)$ hence the inverse Fourier formula allows us to write :

$$\begin{aligned} k_y(x) = \Phi(x-y) &= \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} e^{i\omega.(x-y)}\widehat{\Phi}(\omega)d\omega \\ &= \frac{1}{(2\pi)^{d/2}} \overline{\int_{\mathbb{R}^d} e^{-i\omega.x}e^{i\omega.y}\widehat{\Phi}(\omega)d\omega} \\ &= \frac{1}{(2\pi)^{d/2}} \overline{\int_{\mathbb{R}^d} e^{-i\omega.x}g_y(\omega)d\omega}. \end{aligned}$$

2. Now using 1. one gets

$$\hat{k}_y(\omega) = \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} e^{-i\omega.x}k_y(x)dx = \frac{1}{(2\pi)^{d/2}} \overline{\int_{\mathbb{R}^d} e^{i\omega.x}\hat{g}_y(x)dx}.$$

Gathering with the inverse Fourier transform of $g_y \in L^2(\mathbb{R}^d)$, we have

$$\hat{k}_y(\omega) = \overline{g_y(\omega)} = e^{-i\omega.y}\widehat{\Phi}(\omega).$$

■

Proof (of Theorem 1)
We write

$$\mathcal{H}_0 = \{f \in L^2(\mathbb{R}^d) : \int_S \frac{|\hat{f}(\omega)|^2}{\widehat{\Phi}(\omega)} d\omega < \infty \text{ and } \hat{f} = 0 \text{ on } S\},$$

with the corresponding norm

$$\|f\|_{\mathcal{H}_0} := \sqrt{\frac{1}{(2\pi)^{d/2}} \int_S \frac{|\hat{f}(\omega)|^2}{\widehat{\Phi}(\omega)} d\omega}.$$

We will show that $\mathcal{H}_0$ coincides with $\mathcal{H}_K$.

For a given $y \in \mathbb{R}^d$, from Lemma 16 it is clear that $\hat{k}_y(\omega) = 0$ for $\omega \in \mathbb{R}^d \backslash S$. Moreover using again Lemma 16:

$$\int_S \frac{|\hat{k}_y(\omega)|^2}{\widehat{\Phi}(\omega)} d\omega = \int_S \widehat{\Phi}(\omega) d\omega < \infty$$

since $\widehat{\Phi}$ is integrable. Then $k_y \in \mathcal{H}_0$ for any $y \in \mathbb{R}^d$. Now we have to establish that $\mathcal{H}_0$ is a Hilbert space. Following Matache and Matache (2002), we can show that, for any $f \in \mathcal{H}_0$:

$$\|\hat{f}\|_1 \leq \sqrt{(2\pi)^{d/2} \|\widehat{\Phi}\|_1} \|f\|_{\mathcal{H}_0} \text{ and } \|\hat{f}\|_2 \leq \sqrt{(2\pi)^{d/2} \|\Phi\|_1} \|f\|_{\mathcal{H}_0},$$

where $\|\cdot\|_1$ and $\|\cdot\|_2$ denote the norms in $L^1(\mathbb{R}^d)$ and $L^2(\mathbb{R}^d)$.

Indeed, by Cauchy-Schwarz,

$$\int_{\mathbb{R}^d} |\hat{f}(\omega)| d\omega \leq \sqrt{\frac{1}{(2\pi)^{d/2}} \int_S \frac{|\hat{f}(\omega)|^2}{\widehat{\Phi}(\omega)} d\omega} \sqrt{(2\pi)^{d/2} \int_S \widehat{\Phi}(\omega) d\omega}.$$

Moreover, since $\|\widehat{\Phi}\|_\infty \leq \|\Phi\|_1$,

$$\int_{\mathbb{R}^d} |\hat{f}(\omega)|^2 d\omega \leq \|\Phi\|_1 \int_S \frac{|\hat{f}(\omega|^2}{\widehat{\Phi}(\omega)} d\omega.$$

Then considering a Cauchy sequence $(f_n)_{n\in\mathbb{N}}$ in $\mathcal{H}_0$ endowed with $\|\cdot\|_{\mathcal{H}_0}$, $(\hat{f}_n)_{n\in\mathbb{N}}$ will be a Cauchy sequence in both $L^1(\mathbb{R}^d)$ and $L^2(\mathbb{R}^d)$. We conclude with Matache and Matache (2002) that $(f_n)_n$ is convergent in $\mathcal{H}_0$. Then $\mathcal{H}_0$ is complete and becomes a Hilbert space endowed with the following inner product:

$$< f, g >_{\mathcal{H}_0} = \frac{1}{(2\pi)^{d/2}} \int_S \frac{\hat{f}(\omega)\overline{\hat{g}(\omega)}}{\widehat{\Phi}(\omega)} d\omega.$$

Finally reproducing property holds. Indeed let $f \in \mathcal{H}_0$. Using again Lemma 16 :

$$f(x) = \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} e^{i\omega.x} \hat{f}(\omega) d\omega = \frac{1}{(2\pi)^{d/2}} \int_S \frac{\hat{f}(\omega)}{\widehat{\Phi}(\omega)} \overline{\hat{k}_x(\omega)} d\omega = < f, k_x >_{\mathcal{H}_0} .$$

We have already shown that $\forall x \in \mathbb{R}^d$, $k_x \in \mathcal{H}_0$. As a result, the unicity of the RKHS for a given kernel concludes the proof. ■

Proof (of Corollary 2)

First we have trivially that $\widehat{\Phi}$ is integrable since $s > \frac{1}{2}$. We can hence apply Theorem 1 to have

$$\mathcal{H}_K = \{ f \in L^2(\mathbb{R}^d) : \int_{\mathbb{R}^d} |\hat{f}(\omega)|^2 (c + \|\omega\|^2)^s d\omega < \infty \},$$

since the support of $\widehat{\Phi}$ is $\mathbb{R}^d$. This expression of the RKHS associated to K corresponds, up to a constant, to the Sobolev space $\mathcal{W}_s^2$ defined in (7). Then K is a Sobolev smooth kernel with exponent $r = 2s$. ■

6.2 Proof of Theorem 4

First introduce the notion of interpolation space (Bennett and Sharpley, 1988). We restrict ourselves to a description of the real interpolation method. Let $(B, \|.\|_B)$ be a Banach space and $\mathcal{H}$ a Hilbert space dense in B. The Peetre's functional for the couple $(B, \mathcal{H})$ is defined by, for $t > 0$,

$$P(f, t, B, \mathcal{H}) := \inf \left\{ \|f_0\|_B + t\|f_1\|_{\mathcal{H}}, f = f_0 + f_1 \text{ such that } f_0 \in B, f_1 \in \mathcal{H} \right\}.$$

For fixed $t > 0$, the functional P defines a norm in the Banach space B. It is therefore a simple way to define the interpolation space between B and $\mathcal{H}$ entirely in terms of this functional. Given $\theta \in]0, 1[$ and $q \in [0, \infty]$, the space $(B, \mathcal{H})_{\theta,q}$ called *interpolation space between B and $\mathcal{H}$* consists of all $f \in B$ such that

$$\|f\|_{\theta,q} := \begin{cases} \left(\int_0^{+\infty} t^{-\theta q} P(f, t, B, \mathcal{H})^q \frac{dt}{t} \right)^{\frac{1}{q}} \text{ if } q < \infty, \\ \sup_{t>0} \left\{ t^{-\theta} P(f, t, B, \mathcal{H}) \right\} \text{ if } q = \infty \end{cases}$$

is finite.

Here we are interested in the case $q = \infty$ and the following geometric explanation of interpolation space (Smale and Zhou, 2003, Theorem 3.1):

$$f \in (B, \mathcal{H})_{\theta,\infty} \Longrightarrow \inf_{g \in B_{\mathcal{H}}(R)} \|f - g\|_B \leq \|f\|_{\theta,\infty}^{\frac{1}{1-\theta}} \left(\frac{1}{R} \right)^{\frac{\theta}{1-\theta}}, \tag{16}$$

where $B_{\mathcal{H}}(R) := \left\{ f \in \mathcal{H} : \|f\|_{\mathcal{H}} \leq R \right\}$. Hence the interpolation space between $\mathcal{B}$ and $\mathcal{H}$ satisfies $\mathcal{H} \subset (\mathcal{B}, \mathcal{H})_{\theta,\infty} \subset \mathcal{B}$. To be more precise it consists of functions located at a polynomial decreasing distance in $\mathcal{B}$ from a ball in $\mathcal{H}$ of radius R as a function of R. It would be useful to control the approximation error function in our framework.

Theorem 17 *Consider $a(\alpha_n)$ defined in* (6). *Suppose the marginal of X is such that $\frac{dP_X}{dx} \leq C_0$. Then if $f^* \in (L^2(\mathbb{R}^d), \mathcal{H}_K)_{\theta,\infty}$ we have:*

$$a(\alpha_n) \leq \|f^*\|_{\theta,\infty}^{\frac{2}{2-\theta}} \alpha_n^{\frac{\theta}{2-\theta}}.$$

Proof By the lipschitz property of the hinge loss, we have clearly since $\frac{dP_X}{dx} \leq C_0$:

$$\begin{aligned} a(\alpha_n) &\leq \inf_{f \in \mathcal{H}_K} \left(\|f - f^*\|_{L^1(P_X)} + \alpha_n \|f\|_K^2 \right) \\ &\leq \inf_{R>0} \left(C_0 \inf_{f \in B_{\mathcal{H}_K}(R)} \|f - f^*\|_{L^2(\mathbb{R}^d)} + \alpha_n R^2 \right). \end{aligned}$$

Now from (16), it follows that if $f^* \in (L^2(\mathbb{R}^d), \mathcal{H}_K)_{\theta,\infty}$,

$$a(\alpha_n) \leq \inf_{R>0} \left(\|f^*\|_{\theta,\infty}^{\frac{1}{1-\theta}} \Big(\frac{1}{R}\Big)^{\frac{\theta}{1-\theta}} + \alpha_n R^2 \right).$$

Optimizing with respect to R leads to the conclusion. ■

Let introduce Besov spaces $\mathcal{B}^p_{s,q}(\mathbb{R}^d)$. A Besov space is a collection of functions with common smoothness, in terms of modulus of continuity. This is a large class of functional spaces, including in particular the Sobolev spaces defined in (7) ($\mathcal{W}^2_s = \mathcal{B}^2_{s,2}(\mathbb{R}^d)$ for any $s > 0$) and the Hölder spaces ($H^s = \mathcal{B}^s_{\infty,\infty}(\mathbb{R}^d)$ for any $s > 0$). For a large study, we refer to Triebel (1992).

Here we restrict ourselves to the spaces $\mathcal{B}^2_{s,\infty}(\mathbb{R}^d)$. For any $h \in \mathbb{R}^d$, we write I for the identity operator, T_h for the translation operator ($T_h(f,x) = f(x+h)$) and $\Delta^r_h := (T_h - I)^r$ for the difference operator. The modulus of continuity of order r of a function $f \in L^2(\mathbb{R}^d)$ is then

$$\omega_r(f,t)_2 = \sup_{|h|\leq t} \|\Delta^r_h(f)\|_{L^2(\mathbb{R}^d)}.$$

Then the Besov space $\mathcal{B}^2_{s,\infty}(\mathbb{R}^d)$ consists of all functions f such that the semi-norm

$$\|f\|_{s,\infty} = \sup_{t>0} t^{-s}\omega_r(f,t)_2$$

is finite.

If we add $\|f\|_{L^2(\mathbb{R}^d)}$ to this semi-norm, we obtain the usual norm of $\mathcal{B}^2_{s,\infty}(\mathbb{R}^d)$.

Lemma 18 *Let $s > 0$ and $0 < \theta < 1$. Then,*

$$(L^2(\mathbb{R}^d), \mathcal{W}^2_s)_{\theta,\infty} = \mathcal{B}^2_{\theta s,\infty}(\mathbb{R}^d).$$

A proof is presented by Triebel (1978) in a more general framework.

Proof (of Theorem 4)

From the definition of Sobolev smooth kernels, we have $\mathcal{H}_{K_r} = \mathcal{W}^2_{\frac{r}{2}}$. Hence we obtain with Lemma 18:

$$(L^2(\mathbb{R}^d), \mathcal{H}_{K_r})_{\theta,\infty} = \mathcal{B}^2_{\frac{\theta r}{2},\infty}(\mathbb{R}^d).$$

Applying Theorem 17 with $\theta = \frac{2s}{r}$, this ends up the proof since P_X satisfies $\frac{dP_X}{dx} < C_0$. ■

6.3 Proof of Theorem 7

In order to control the generalization error, we have to state an inequality such as (5). We propose to use a stochastic oracle inequality from Steinwart et al. (2007). This result takes place under a margin assumption of the type (10) and a complexity assumption over the used RKHS.

We define the covering numbers of a subset A of a Banach space (E,d) as :

$$\mathcal{N}(A,\varepsilon,E) = \min\{n \geq 1 : \exists x_1, \ldots x_n \in E \text{ such that } A \subset \cup_{i=1}^n B_d(x_i,\varepsilon)\}.$$

Furthermore, given a realization $T = \{(x_1,y_1),\dots(x_n,y_n)\}$ of the training set, we denote by $L^2(T_X)$ the space of all equivalence classes of functions $f : \mathcal{X} \mapsto \mathbb{R}$ such that the norm

$$\|f\|_{L^2(T_X)} := \left(\frac{1}{n}\sum_{i=1}^{n} f(x_i)^2\right)^{1/2} \tag{17}$$

is finite. Then we can consider the behaviour of $\log \mathcal{N}(B_{\mathcal{H}_K}, \varepsilon, L^2(T_X))$ as a complexity measure for the used RKHS.

Proposition 19 (Steinwart and Scovel, 2007) *Let P be a distribution on $\mathcal{X} \times \{-1,1\}$ and $\mathcal{H}_K$ a RKHS of continuous functions on $\mathcal{X}$. Suppose*

1. *There exists $q \in [0,+\infty]$ and $c_0 > 0$ such that*

$$\mathbb{P}(|2\eta(X)-1| \le t) \le c_0 t^q, \, ,\forall t > 0.$$

2. *There exist $a \ge 1, 0 < p < 1$ such that*

$$\sup_{T\in(\mathcal{X}\times\mathcal{Y})^n} \log \mathcal{N}\left(B_{\mathcal{H}_K}, \varepsilon, L^2(T_X)\right) \le a\varepsilon^{-2p}, \forall \varepsilon > 0. \tag{18}$$

Then there exist constants $c \ge 1$, $\kappa, \kappa', \kappa'' > 0$ such that for all $x \ge 1$, the clipped version $\hat{f}_n^C$ of SVM classifier $\hat{f}_n$ satisfies, with probability larger than $1-e^{-x}$,

$$\begin{aligned} R_l(\hat{f}_n^C, f^*) \quad &\le \quad c \inf_{f\in\mathcal{H}_K} \left(\mathbb{E}_P\left(l(f)-l(f^*)\right)+\alpha_n\|f\|_K^2\right)+\frac{\kappa}{n\alpha_n^p} \\ &+ \quad \left(\frac{\kappa}{n\alpha_n^p}\right)^{\frac{q+1}{q+2-p}} + \frac{\kappa'}{n^{\frac{q+1}{q+2}}} + \frac{\kappa'' x}{n}. \end{aligned}$$

Proof (of Theorem 7)

The hinge loss $l(y, f(x)) = (1-yf(x))_+$ satisfies, for all classifier $\hat{f}$ (Zhang, 2004):

$$R(\hat{f}, f^*) \le R_l(\hat{f}, f^*). \tag{19}$$

Therefore, to control the excess risk of a classifier, it is sufficient to control the RHS of (19).

We apply Proposition 19 for the stochastic part and Theorem 4 for the approximation part of the analysis.

Recall a standard result for covering numbers of Sobolev spaces (Chen et al., 2004):

$$\log \mathcal{N}(B_{\mathcal{W}_r^2}, \varepsilon, C(\mathbb{R}^d)) \le a\varepsilon^{-\frac{d}{r}}, \tag{20}$$

where constant $a := a(d)$. From (17) we have $\|f\|_{L^2(T_X)} \le \|f\|_\infty$ for any $f \in C(\mathbb{R}^d)$, $T \in (\mathcal{X}\times\mathcal{Y})^n$. Then (20) holds true for $\log \mathcal{N}(B_{\mathcal{W}_r^2}, \varepsilon, L^2(T_X))$ uniformly over $T \in (\mathcal{X}\times\mathcal{Y})^n$. Gathering with $\mathcal{H}_{K_r} = \mathcal{W}_{r/2}^2$, the RKHS $\mathcal{H}_{K_r}$ satisfies (18) of Proposition 19 with $p = \frac{d}{r}$. Applying Proposition 19, there exist $c \ge 1$, $\kappa, \kappa', \kappa'' > 0$ such that, for all $x \ge 1$, with probability larger than $1-e^{-x}$,

$$\begin{aligned} R_l(\hat{f}_n^C, f^*) \quad &\le \quad c \inf_{f\in\mathcal{H}_K} \left(R_l(f, f^*)+\alpha_n\|f\|_K^2\right)+\frac{\kappa}{n\alpha_n^{\frac{d}{r}}} \\ &+ \quad \left(\frac{\kappa}{n\alpha_n^{\frac{d}{r}}}\right)^{\frac{q+1}{q+2-d/r}} + \frac{\kappa'}{n^{\frac{q+1}{q+2}}} + \frac{\kappa'' x}{n}. \end{aligned}$$

Since $f^* \in \mathcal{B}^2_{s,\infty}(\mathbb{R}^d)$, we get from Theorem 4 that with probability larger than $1-e^{-x}$,

$$R_l(\hat{f}_n^C, f^*) \leq cC_0^{\frac{r}{r-s}} \|f^*\|_{s,\infty}^{\frac{r}{r-s}} \alpha_n^{\frac{s}{r-s}} + \frac{\kappa}{n\alpha_n^{\frac{d}{r}}} + \left(\frac{\kappa}{n\alpha_n^{\frac{d}{r}}}\right)^{\frac{q+1}{q+2-d/r}} + \frac{\kappa'}{n\alpha_n^{\frac{d}{r}}} + \frac{\kappa'' x}{n}.$$

The choice of α_n in (11) optimizes the RHS. Integrating with respect to the training set, one leads to the conclusion. ■

6.4 Proof of Theorem 12

To prove Theorem 12, we use a general oracle inequality for aggregation. Let us first recall the general context of aggregation.

Suppose we have $M \geq 2$ differents classifiers $f_1, \dots, f_M$ with values in $\{-1,1\}$. The method of aggregation consists in building a new classifier $\tilde{f}_n$ from D_n called aggregate which mimics the best among $f_1, \dots f_M$. Our procedure is using exponential weights of the following form:

$$\omega_j^{(n)} = \frac{\exp\left(\sum_{i=1}^n Y_i f_j(X_i)\right)}{\sum_{k\in\{1\dots M\}} \exp\left(\sum_{i=1}^n Y_i f_k(X_i)\right)}.$$

Then we define the following aggregate:

$$\tilde{f}_n = \sum_{j=1}^M \omega_j^{(n)} f_j. \tag{21}$$

Under the margin assumption (10), we have this oracle inequality:

Theorem 20 (Lecué, 2005) *Suppose* (10) *holds for some* $q \in (0,+\infty)$. *Assume we have at least a polynomial number of classifiers to aggregate (i.e., there exist* $a \geq 1$, $b > 0$ *such that* $M \geq an^b$*). Then the aggregate defined in* (21) *satisfies, for all integer* $n \geq 1$,

$$\mathbb{E}R(\tilde{f}_n, f^*) \leq (1 + 2\log^{-1/4} M)\left(2 \min_{k\in\{1,\dots M\}} R(f_k, f^*) + Cn^{-\frac{q+1}{q+2}} \log^{7/4} M\right), \tag{22}$$

where C *depends on* a, b *and the constant* c_0 *appearing in* (10).

Proof (of Theorem 12)

Let $(q_0, s_0) \in K$ and consider $0 < q_{min} < q_{max} < +\infty$ and $0 < s_{min} < s_{max} < +\infty$ such that $K \subset [q_{min}, q_{max}] \times [s_{min}, s_{max}]$. We consider the function

$$\Phi(q,s) = \frac{r(r-s)(q+1)}{s(r(q+2)-d) + (r-s)(q+1)d}$$

defined on $[0,+\infty[\times[0,+\infty[$ with value on $[\frac{1}{2}, \frac{r}{d}]$. We denote by

$$k_0 \in \left\{0, \dots, \left\lfloor \frac{(2r-d)\Delta}{2d} \right\rfloor - 1\right\}$$

the integer such that

$$\frac{1}{2}+k_0\Delta^{-1} \le \Phi(q_0,s_0) \le \frac{1}{2}+(k_0+1)\Delta^{-1}.$$

Since $q \mapsto \Phi(q,s)$ continuously increases on $\mathbb{R}^+$, for n greater than a constant depending on b, r, d and K, there exists $\overline{q}_0 \in \left[\frac{q_{min}}{2}, q_{max}\right]$ such that $\overline{q}_0 \le q_0$ and

$$\Phi(\overline{q}_0,s_0) = \frac{1}{2}+k_0\Delta^{-1}. \tag{23}$$

Now we can apply Theorem 20 for $\overline{q}_0$. Since $\Delta = n_2^b$, putting $M = \left\lfloor \frac{(2r-d)\Delta}{2d} \right\rfloor$ we have the following oracle inequality:

$$\mathbb{E}_{P^{\otimes n_2}}\left(R(\tilde{f}_n, f^*)|D^1_{n_1}\right) \le (1+2\log^{-\frac{1}{4}} M)\left(2\min_{\alpha\in\mathcal{G}(n_2)}\left(R(\hat{f}^{\alpha}_{n_1}, f^*)\right)+C_1 n_2^{-\frac{\overline{q}_0+1}{\overline{q}_0+2}}\log^{7/4} M\right),$$

where C_1 depends on c_0, K and b. Hence we have, integrating with respect to $D^1_{n_1}$,

$$\mathbb{E}\left(R(\tilde{f}_n, f^*)\right) \le C_2\left(\mathbb{E}R(\hat{f}^{\alpha_{k_0}}_{n_1}, f^*)+n_2^{-\frac{\overline{q}_0+1}{\overline{q}_0+2}}\log^{7/4} n_2\right),$$

where $\alpha_{k_0} = m^{-\phi_{k_0}} = n_2^{-\Phi(\overline{q}_0,s_0)}$ with (23) and C_2 depends on K,b,r,d and c_0. Therefore we can apply Theorem 7 to the classifier $\hat{f}^{\alpha_k}_{n_1}$:

$$\mathbb{E}_{P^{\otimes n_1}}R(\hat{f}^{\alpha_{k_0}}_{n_1}, f^*) \le C n_1^{-\frac{s_0}{r-s_0}\Phi(\overline{q}_0,s_0)},$$

where C depends on r,d and K. Remark that C does not depend on $\overline{q}_0$ and s_0 since $(\overline{q}_0,s_0) \in [\frac{q_{min}}{2}, q_{max}] \times [s_{min}, s_{max}]$. Moreover C is uniformly bounded over (q,s) belonging to a compact in Theorem 7.

Finally suppose P satisfies (10) for q_0. Hence we obtain:

$$\mathbb{E}\left(R(\tilde{f}_n, f^*)\right) \quad \le \quad C_3\left(n_1^{-\frac{s_0}{r-s_0}\Phi(\overline{q}_0,s_0)}+n_2^{-\frac{\overline{q}_0+1}{\overline{q}_0+2}}\log^{\frac{7}{4}} n_2\right)$$

for $C_3 := C_3(K,b,c_0,r,C_0,d)$. We have $n \ge n_2 \ge \frac{an}{\log n}$ and $n_1 \ge n(\frac{2}{3}-\frac{a}{\log 3})$. Then for n greater than a constant depending on β_{min}, a, and b, there exists $C_3' := C_3'(K,b,c_0,r,C_0,d)$ such that

$$\begin{aligned}\mathbb{E}\left(R(\tilde{f}_n, f^*)\right) \quad &\le \quad C_3\left(n^{-\frac{s_0}{r-s_0}\Phi(\overline{q}_0,s_0)}+n^{-\frac{\overline{q}_0+1}{\overline{q}_0+2}}\log^{\frac{11}{4}} n\right)\\ &\le \quad C_3' n^{-\frac{s_0}{r-s_0}\Phi(\overline{q}_0,s_0)}.\end{aligned}$$

The construction of $\overline{q}_0$ and restrictions on r entail $\frac{s_0}{r-s_0}|\Phi(\overline{q}_0,s_0)-\Phi(q_0,s_0)| \le \Delta^{-1} = n_2^{-b}$. We lead to the conclusion since the sequence $(n^{n_2^{-b}})_{n\in\mathbb{N}}$ is convergent. ■

Acknowledgments

I would like to acknowledge my advisor, Laurent Cavalier, for giving me many ideas and advices for this work. I'm also grateful to Liva Ralaivola for the experimental part of this work and to the anonymous referees for interesting remarks.

References

R.A. Adams. *Sobolev Spaces*. Academic Press, 1975.

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

S. Arora, L. Babai, J. Stern, and Z. Sweedyk. The hardness of approximate optima in lattices, codes, and systems of linear equations. *Journal of Computer and System Sciences*, 54 (2):317–331, 1997.

P.L. Bartlett. The sample complexity of pattern classification with neural networks: the size of the weights is more important than the size of the network. *IEEE Transactions on Information Theory*, 44 (2):525–536, 1998.

P.L. Bartlett and S. Mendelson. Empirical minimization. *Probability Theory and Related Fields*, 135 (3):311–334, 2006.

P.L. Bartlett, M.I. Jordan, and J.D. McAuliffe. Convexity, classification, and risk bounds. *J. Amer. Statist. Assoc.*, 101 (473):138–156, 2006.

C. Bennett and R. Sharpley. *Interpolation of Operators*. Academic Press, 1988.

G. Blanchard, O. Bousquet, and P. Massart. Statistical performance of support vector machines. to appear Annals of Statistics, 2006.

B.E. Boser, I. Guyon, and V. Vapnik. A training algorithm for optimal margin classifiers. In *Computational Learning Theory*, pages 144–152, 1992.

S. Boucheron, O. Bousquet, and G. Lugosi. Theory of classification: a survey of some recent advances. *ESAIM: Probability and Statistics*, 9:323–375, 2005.

D.R. Chen, Q. Wu, Y. Ying, and D.X. Zhou. Support vector machine soft margin classifiers: error analysis. *Journal of Machine Learning Research*, 5:1143–1175, 2004.

N. Cristianini and H. Shawe-Taylor. *Introduction to Support Vector Machines, and Other Kernel-Based Learning Methods*. Cambridge University Press, 2000.

L. Devroye. Necessary and sufficient conditions for the pointwise convergence of nearest neighbor regression function estimates. *Z. Wahrsch. Vew. Gebiete*, 61 (4):467–481, 1982.

Y. Freund. Boosting a weak learning algorithm by majority. *Information and Computation*, 121 (2): 256–285, 1995.

A. Karatzoglou, A. Smola, and K. Hornik. An S4 package for kernel methods in R. Reference manual, 2007.

G. Lecué. Simultaneous adaptation to the margin and to complexity in classification. *The Annals of Statistics*, 35 (4):1698–1721, 2007a.

G. Lecué. Optimal rates of aggregation in classification under low noise assumption. *Bernoulli*, 13 (4):1000–1022, 2007b.

P. Malliavin. *Analyse de Fourier-Analyses spectrales*. Ecole Polytechnique, 1974.

E. Mammen and A.B. Tsybakov. Smooth discrimination analysis. *The Annals of Statistics*, 27 (6): 1808–1829, 1999.

P. Massart and E. Nédélec. Risk bounds for statistical learning. *The Annals of Statistics*, 34 (5): 2326–2366, 2006.

M. Matache and V. Matache. Hilbert spaces induced by Toeplitz covariance kernels. *Lecture notes in Control and Information Sciences*, 280:319–334, 2002.

A. Nemirovski. *Topics in Nonparametric Statistics*. Ecole d'été de Saint-Flour XXVIII, Springer, N.Y., 1998.

D. Rätsch, T. Onoda, and K.R. Müller. Soft margin for adaboost. Esprit Working Group in Neural and Computational Learning II, 1998.

S. Smale and D.X. Zhou. Estimating the approximation error in learning theory. *Analysis and Applications*, 1 (1):17–41, 2003.

I. Steinwart. On the influence of the kernel on the consistency of support vector machines. *Journal of Machine Learning Research*, 2:67–93, 2001.

I. Steinwart. Consistency of support vector machines and other regularized kernel classifiers. *IEEE Transactions on Information Theory*, 51 (1):128–142, 2005.

I. Steinwart and C. Scovel. Fast rates for support vector machines. In *Proc. 18th Annu. Conference on Comput. Learning Theory*, volume 3559, pages 279–294, 2005.

I. Steinwart and C. Scovel. Fast rates for support vector machines using Gaussian kernels. *The Annals of Statistics*, 35 (2):575–607, 2007.

I. Steinwart, D. Hush, and C. Scovel. An oracle inequality for clipped regularized risk minimizers. *Neural Information Processing Systems*, 19:1321–1328, 2007.

H. Triebel. *Theory of Functions Spaces II*. Birkhauser, 1992.

H. Triebel. *Interpolation Theory, Function Spaces, Differential Operators*. North-Holland Publishing Company, 1978.

A.B. Tsybakov. Optimal aggregation of classifiers in statistical learning. *The Annals of Statistics*, 32 (1):135–166, 2004.

V.N. Vapnik and A.Ya. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16 (2):264–280, 1971.

V.N. Vapnik and A.Ya. Chervonenkis. *Theory of Pattern Recognition*. Nauka, Moscow, 1974.

R.C. Williamson, A.J. Smola, and B. Schölkopf. Generalization performance of regularization networks and support vector machines via entropy numbers of compact operators. *IEEE Transactions on Information Theory*, 47 (6):2516–2532, 2001.

Q. Wu and D.X. Zhou. Analysis of support vector machine classification. *J. Comput. Anal. Appl.*, 8 (2):99–119, 2006.

Q. Wu, Y. Ying, and D.X. Zhou. Multi-kernel regularized classifiers. *Journal of Complexity*, 23 (1): 108–134, 2007.

Y. Yang. Mixing strategies for density estimation. *The Annals of Statistics*, 28 (1):75–87, 2000.

T. Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization. *The Annals of Statistics*, 32 (1):56–85, 2004.