Electronic Journal of Statistics
Vol. 3 (2009) 824–850
ISSN: 1935-7524
DOI: 10.1214/08-EJS316

Penalized empirical risk minimization over Besov spaces

Sébastien Loustau

Université Aix-Marseille 1,
LATP, 39 Rue Joliot Curie, 13453 Marseille Cedex, France,
e-mail: loustau@cmi.univ-mrs.fr

Abstract: Kernel methods are closely related to the notion of reproducing kernel Hilbert space (RKHS). A kernel machine is based on the minimization of an empirical cost and a stabilizer (usually the norm in the RKHS). In this paper we propose to use Besov spaces as alternative hypothesis spaces. We study statistical performances of a penalized empirical risk minimization for classification where the stabilizer is a Besov norm. More precisely, we state fast rates of convergence to the Bayes rule. These rates are adaptive with respect to the regularity of the Bayes.

Received October 2008.

1. Introduction

1.1. *Classification framework*

We consider the binary classification setting. Let (X, Y) be a random variable with unknown probability distribution P over $\mathcal{X} \times \{-1, +1\}$. $X \in \mathcal{X}$ is called the *input* variable. It is a feature vector, whereas $Y \in \{-1, 1\}$ is the corresponding *class* or *label*. The goal of classification is to predict class Y when only X is observed. In other words, a classification algorithm builds a decision rule from $\mathcal{X}$ to $\{-1, 1\}$. A classifier is a function $f : \mathcal{X} \to \mathbb{R}$ where the sign of $f(x)$ determines the class of an input x. The performance of a classifier is measured by the *generalization error*, given by:

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

If we assume that the joint distribution P is known, the best classifier is defined by:

$$f^*(x) := 2\mathbb{1}_{\{\eta(x) \geq 1/2\}} - 1, \tag{1}$$

where $\eta(x) := \mathbb{P}(Y = 1|X = x)$. Classifier (1) is called the Bayes rule. It is easy to see that it minimizes the generalization error.

Unfortunately, in practice η is unknown and then f^* is not available. A natural way to overcome this difficulty is to provide an empirical classifier based on training data. Suppose we have at our disposal a *training set* $D_n = \{(X_i, Y_i),\ i = 1, \dots, n\}$ made of i.i.d. realizations of the random variable (X, Y) of law P. Now classification can be seen as a standard estimation problem where we have to estimate f^* from i.i.d. observations. The efficiency of an empirical classifier $\hat{f}_n$ is measured via its *excess risk*:

$$R(\hat{f}_n, f^*) := R(\hat{f}_n) - R(f^*), \tag{2}$$

where $R(\hat{f}_n) := \mathbb{P}(\text{sign}(\hat{f}_n(X)) \neq Y | D_n)$. Here we are interested in consistent classifier $\hat{f}_n$, i.e. such that (2) tends to zero as $n \to \infty$. Finally, 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,

$$\mathbb{E}R(\hat{f}_n, f^*) \leq C\psi_n, \tag{3}$$

where $\mathbb{E}$ is the expectation with respect to the training set.

Without any assumption over the joint distribution P, [11] gives arbitrary slow rates. However several authors propose different rates restricting the class of distributions P. Pioneering works of Vapnik [30, 31] investigate the statistical performances of the Empirical Risk Minimization (ERM). The idea is very simple: we are looking at the minimizer of the empirical risk:

$$R_n(f) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{\text{sign}(f(X_i))\neq Y_i\}}. \tag{4}$$

If we suppose that the class of possible Bayes rules has finite VC dimension, ERM reaches the parametric rate $n^{-\frac{1}{2}}$ in (3). Moreover, if P is noise-free (i.e. $R(f^*) = 0$), the rate becomes n^{-1}. This is a fast rate. More recently, [29] or [19] describes intermediate situations using margin assumptions. These assumptions add a control on the behaviour of the conditional probability function η at the level $\frac{1}{2}$. Under this condition, they get minimax fast rates of convergence between $n^{-\frac{1}{2}}$ and n^{-1} for ERM estimators in classification. At the present time, there exists a vast literature about the fast rates phenomenon. Fast rates have been obtained for different procedure such as Boosting ([7]), Plug-in rules ([1]), SVM ([26]), or dyadic decision trees ([15]). In this work we propose to state fast rates of convergence for a penalized empirical risk minimization using the hinge loss.

1.2. SVM regularization

Support Vector Machines was first proposed by Boser, Guyon and Vapnik ([8]) for pattern recognition. Given a training set D_n, the SVM classifier (without offset) $\hat{f}_n$ solves the following minimization:

$$\min_{f\in H_K} \left(\frac{1}{n}\sum_{i=1}^{n}(1 - Y_i f(X_i))_+ + \alpha_n\|f\|_K^2\right), \tag{5}$$

where H_K denotes the reproducing kernel Hilbert space (RKHS) associated to the kernel K. The first term in (5) is an empirical cost using the hinge loss $l(y, f(x)) := (1 - yf(x))_+$. This term allows the solution to fit the data. The second term regularizes the solution with $\|f\|_K^2$, the square norm of f in the RKHS. The parameter α_n is called the smoothing parameter. It has to be determined explicitly to make the trade-off between these two terms. To get statistical performances of the method, we need to take a closer look at the couple $(H_K, \|\cdot\|_K)$.

For any Borel measure μ over $\mathcal{X}$, consider $L_K : L^2(\mu) \to L^2(\mu)$ the integral operator defined as:

$$L_K : f \mapsto \int_{\mathcal{X}} K(x, \cdot) f(x) \mu(dx).$$

This operator is closely related to the kernel K. If $\mathcal{X}$ is compact and K is continuous (K is called a Mercer kernel), L_K is compact. From spectral theorem, there exist $(\phi_k)_{k\geq 1}$, orthonormal basis of $L^2(\mu)$ of eigenfunctions of L_K with $(\lambda_k)_{k\geq 1}$ corresponding eigenvalues. It allows us to get a representation of H_K in a sequence space as follows:

$$H_K = \left\{ f \in L^2(\mu) : f = \sum a_k \phi_k \, , \sum_{k\geq 1} \frac{a_k^2}{\lambda_k} < +\infty \right\}. \tag{6}$$

In this case, the regularization in (5) can be written:

$$\|f\|_K^2 = \sum_{k\geq 1} \frac{a_k^2}{\lambda_k}, \tag{7}$$

where $(a_k)_{k\geq 1}$ gives a representation of f in the basis $(\phi_k)_{k\geq 1}$. For instance, consider a convolution kernel $K(x, y) = \Phi(x - y)$. Then in (7), coefficients $(a_k)_{k\geq 1}$ are the Fourier coefficients of f whereas $(\lambda_k)_{k\geq 1}$ are the Fourier coefficients of Φ.

Representation (6) holds for Mercer kernels. One can generalize this fact to $\mathcal{X} = \mathbb{R}^d$ in the following case. Suppose $K(x, y) = \Phi(x-y)$ is a convolution kernel. If Φ has some mild properties, the RKHS associated to K can be written:

$$H_K = \left\{ f \in L^2(\mathbb{R}^d) : \|f\|_K^2 = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \frac{|\hat{f}(\omega)|^2}{\hat{\Phi}(\omega)} d\omega < \infty \right\}, \tag{8}$$

where $\hat{\Phi}$ is the Fourier transform of Φ. In this case the regularity is expressed by the asymptotic behaviour of the Fourier transform. For example, if we suppose that $\hat{\phi}$ decreases polynomially with ω, H_K is a Sobolev space.

[16] uses representation (8) to state learning rates for SVM minimization when $H_K = \mathcal{W}_s^2(\mathbb{R}^d)$ with $s > d/2$. It corresponds to a Sobolev space of continuous functions as RKHS. If $f^* \in \mathcal{B}_{r2\infty}(\mathbb{R}^d)$, one gets:

$$\mathbb{E}R(\hat{f}_n, f^*) \leq C n^{-\beta(q,r,s)}, \tag{9}$$

where β is a function of:

- q the margin parameter,
- s the exponent of the Sobolev space $\mathcal{W}_s^2(\mathbb{R}^d)$,
- r the smoothness of $f^* \in \mathcal{B}_{r2\infty}(\mathbb{R}^d)$.

Parameter r describes the regularity of f^* in the Besov space $\mathcal{B}_{r2\infty}(\mathbb{R}^d)$. This assumption is strongly related to the use of Sobolev space as hypothesis space. In these functional spaces, the smoothness is related to the asymptotic behaviour of the Fourier transform. This criterion depends on the variations of the function in $\mathbb{R}^d$. It is a global criterion. For instance, the derivability can be expressed in terms of Fourier transform. For a given f such that $\hat{f} \in L^1(\mathbb{R})$, the following elementary inequalities:

$$|f^{(k)}(t)| \leq \int |e^{i\omega t}(i\omega)^k \hat{f}(\omega)| d\omega \leq \int |\omega|^k |\hat{f}(\omega)| d\omega$$

show that if $\omega \mapsto \omega^p \hat{f}(\omega) \in L^1(\mathbb{R}^d)$, then $f \in \mathcal{C}^p$. Unfortunately the Bayes rule is not continuous. With previous remark, its Fourier transform decreases slowly (in $O(|\omega|^{-1})$). From this point of view, Sobolev spaces are not really adapted to the shape of f^*. As a result, $f^* \in \mathcal{B}_{r2\infty}(\mathbb{R}^d)$ holds for small values of r (namely $r < \frac{d}{2}$). That's why fast rates are not reached in [16].

An alternative is to take into account the regularity of f^*: it is piecewise constant with local discontinuities. This can be done using a multiresolution analysis and considering Besov spaces as hypothesis spaces.

1.3. Besov regularization

It seems interesting to consider minimization (5) with more general hypothesis spaces. We propose to use Besov spaces as hypothesis spaces and study the minimization procedure:

$$\min_{f \in \mathcal{B}_{spq}(\mathbb{R}^d)} \left(\frac{1}{n} \sum_{i=1}^{n} (1 - Y_i f(X_i))_+ + \alpha_n \|f\|_{spq}^2 \right), \tag{10}$$

where $\|\cdot\|_{spq}$ denotes the norm in $\mathcal{B}_{spq}(\mathbb{R}^d)$. We replace H_K by Besov spaces. The advantage of Besov spaces as compared to Sobolev spaces is that they give a more general description of the smoothness properties of functions. An explicit description of $\mathcal{B}_{spq}(\mathbb{R}^d)$ and $\|\cdot\|_{spq}$ is given in Section 2. There exist several motivations to introduce Besov spaces in (5):

- We have $\mathcal{B}_{s22}(\mathbb{R}^d) = \mathcal{W}_s^2(\mathbb{R}^d)$ is a Sobolev space of order s. For $p = q = 2$, (10) corresponds to the standard SVM using Sobolev spaces as RKHS (see [16]). Then (10) generalizes the Sobolev case.
- The use of the hinge loss $l(y, f(x)) = (1 - yf(x))_+$ in (10) is related to the SVM algorithm. The statistical consequences of minimizing such a loss is well-treated in [4]. The principal advantage when we are expecting rates

of convergence is the control of the excess risk (2) by the excess risk using the hinge loss. It gives in this paper fast rates of convergence. Another caracteristic of the hinge loss concerns the regularity of the minimizer of the risk. We have:

$$\arg\min R_l(f) = \arg\min R(f) = f^*,$$

which is a non continuous function with values +1 or −1. For this reason, fast rates cannot be reached in [16] for SVM with Sobolev spaces. $\mathcal{B}_{spq}(\mathbb{R}^d)$ with $p < 2$ gives more flexibility. It contains for instance piecewise regular functions. In this case it will be easier to approximate the Bayes classifier, which leads to better rates of convergence.

- There is a large theory around Besov spaces, such as a characterization using wavelet coefficients. It gives a representation of the norm in (10) in a sequence space as follows:

$$\|f\|_{spq} = \left(\sum_{k\in\mathbb{Z}^d} |\alpha_k|^p\right)^{\frac{1}{p}} + \left(\sum_{j\in\mathbb{N}} \left(2^{j\left(s+d\left(\frac{1}{2}-\frac{1}{p}\right)\right)} \sum_{l=1}^{2^d-1} \left(\sum_{k\in\mathbb{Z}^d} |\beta_{jkl}|^p\right)^{\frac{1}{p}}\right)^{q}\right)^{\frac{1}{q}},$$

  where (α_k) and (β_{jkq}) are the wavelet coefficients.

This representation can be compared to the sequence space representation (7) of a RKHS norm. In the standard SVM case, the regularization can be expressed with respect to the spectrum of L_K. It allows to control the complexity of the RKHS. [20] or [6] control the local Rademacher average of balls in RKHS in this sequence space. It depends on the asymptotic behaviour of the sequence $(\lambda_k)_{k\geq 1}$ and affects the statistical performances of the method. In this paper we point out that similar facts can be derived for Besov spaces using a wavelet analysis.

Minimization (10) is strongly related to the SVM minimization. However as a kernel method, SVM uses a RKHS norm as regularization. It allows to define SVM as a large margin hyperplane in some Hilbert feature space. Here the hypothesis space is a Besov space. Besov spaces are not Hilbertian, and then cannot be represented as RKHS. This penalized empirical risk minimization is not an SVM minimization. However an interesting open problem is to express (10) as a kernel method. This problem is connected to recent developments on the theory of reproducing kernels. In this direction, a short discussion is proposed at the end of this work.

The remainder of this paper is organized as follows: In Section 2, we introduce the wavelet theory. We characterize Besov spaces in terms of wavelet coefficients. It reduces the control of the Rademacher average to a problem in a sequence space, leading to very natural proofs. An oracle inequality is deduced in Section 3. It is a direct application of a general model selection theorem due to [6]. We finally control the approximation power of Besov balls to state fast learning rates for the procedure (10). The solution is adaptive with respect to the regularity of the Bayes. We conclude in Section 4 with a discussion. Section 5 is dedicated to the proofs of the main results.

2. Wavelet framework

For the mathematical aspects of wavelets, we refer for example to [21], while [17] proposes comprehensive expositions for signal processing. Wavelet applications in statistical settings are given for instance in [13]. For a complete study of minimax rates of convergence for density estimation by wavelet thresholding, we refer to [12].

Here recall some definitions and notations for wavelets and Besov spaces. Going back to statistical learning theory, one proposes a control of the local Rademacher average of Besov balls.

2.1. Besov spaces and wavelets

2.1.1. Wavelet bases of $L^2(\mathbb{R}^d)$

For the one-dimensional case, we refer for instance to [12]. To introduce the d-dimensional case, we begin with an example in dimension 2 using the tensor product. Write $(V_j^1)_{j\in\mathbb{Z}}$ a multiresolution analysis (MRA for short in the sequel) of $L^2(\mathbb{R})$ generated by ϕ. Write $V_j = V_j^1 \times V_j^1$ for all $j \in \mathbb{Z}$. Then the system $(\phi(x-k)\phi(y-l))_{k,l\in\mathbb{Z}}$ is an orthonormal basis of V_0 in $L^2(\mathbb{R}^2)$. Let consider W_j, $j \in \mathbb{Z}$ such that $V_{j+1} = V_j \oplus W_j$. Then we have for $j = 0$:

$$W_0 = \overline{V_0^1 \otimes W_0^1} \oplus \overline{W_0^1 \otimes V_0^1} \oplus \overline{W_0^1 \otimes W_0^1}.$$

A basis of W_0 is obtained with the three collections $\phi(x-k)\psi(y-l)$, $\psi(x-k)\phi(y-l)$ and $\psi(x-k)\psi(y-l)$ for $(k,l) \in \mathbb{Z}^2$. More generally for all $j \in \mathbb{Z}$:

$$W_j = \overline{V_j^1 \otimes W_j^1} \oplus \overline{W_j^1 \otimes V_j^1} \oplus \overline{W_j^1 \otimes W_j^1}.$$

Then the two-dimensional mother wavelets are $2^j\phi(2^jx-k)\psi(2^jy-l)$, $2^j\psi(2^jx-k)\phi(2^jy-l)$ and $2^j\psi(2^jx-k)\psi(2^jy-l)$ for $(k,l) \in \mathbb{Z}^2$. This means that there are three wavelets in the two-dimensional case. This fact is illustrated in [21] with a geometrical point of view. We can generalize this result in higher dimensions with the following lemma.

Lemma 1. *Let V_j, $j \in \mathbb{Z}$ a MRA r-regular of $L^2(\mathbb{R}^d)$. Then there exist $L = 2^d - 1$ functions $\psi_1, \ldots \psi_L \in V_1$ such that:*

1. *for all $l \in \{1 \ldots L\}$, for all $\alpha \in \mathbb{N}^d : |\alpha| \leq r$, for all $x \in \mathbb{R}^d$ and $N \geq 1$,*

$$|\partial^\alpha \psi_l(x)| \leq C_N(1+|x|)^{-N};$$

2. *the system $\{\psi_l(x-k),\, 1 \leq l \leq L,\, k \in \mathbb{Z}^d\}$ is an ONB of W_0.*

As a result, the system given by:

$$2^{\frac{dj}{2}}\psi_l(2^jx-k),\ 1 \leq l \leq L,\ k \in \mathbb{Z}^d,\ j \in \mathbb{Z} \tag{11}$$

is an orthonormal basis of $L^2(\mathbb{R}^d)$.

This lemma generalizes the one-dimensional case. From a scaling function r-regular and rapidly decreasing generating a MRA, we can construct $2^d - 1$ mother wavelets with the same regularity. The existence of such a wavelet basis is proved in [21].

As a consequence, any $f \in L^2(\mathbb{R}^d)$ can be decomposed as:

$$f = \sum_{k \in \mathbb{Z}^d} \alpha_{0k}\phi_{0k} + \sum_{j \geq 0} \sum_{k \in \mathbb{Z}^d} \sum_{l=1}^{2^d-1} \beta_{jkl}\psi_{jkl}, \tag{12}$$

where

$$\alpha_{0k} = \int_{\mathbb{R}^d} f(x)\phi_{0k}(x)dx \text{ and } \beta_{jkl} = \int_{\mathbb{R}^d} f(x)\psi_{jkl}(x)dx.$$

In the case of tensor product, we have for all $k \in \mathbb{Z}^d$ and $x \in \mathbb{R}^d$:

$$\phi_{0k}(x) = \phi(x_1 - k_1)\dots\phi(x_d - k_d).$$

Moreover for all $j \geq 0$, $k \in \mathbb{Z}^d$, $l \in \{1, \dots, 2^d - 1\}$ and $x \in \mathbb{R}^d$:

$$\psi_{jkl}(x) = 2^{\frac{dj}{2}}\psi^{e_1}(2^j x_1 - k_1)\dots\psi^{e_d}(2^j x_d - k_d),$$

for $e \in \{0,1\}^d \backslash 0_{\mathbb{R}^d} = E$ and where we write for simplicity $\psi^0 = \phi$ and $\psi^1 = \psi$.

Here we are interested in compactly supported wavelet bases. [10] has shown that in dimension $d = 1$, there exists an orthonormal basis of compactly supported wavelets satisfying conditions of Lemma 1, for any integer $r \geq 1$ (for $r = 0$, it corresponds to the Haar basis). Using the tensor product, this result gives a compactly supported d-dimensional wavelet basis of $L^2(\mathbb{R}^d)$ (see [21] for details).

2.1.2. Besov spaces

Besov spaces were introduced by O.V. Besov in the 60s. Here we propose to characterize Besov spaces $\mathcal{B}_{spq}(\mathbb{R}^d)$ in terms of wavelet coefficients.

Recall $P_j : L^2(\mathbb{R}^d) \to V_j$ is the projection operator into V_j and $D_j = P_{j+1} - P_j$. We know that for $f \in L^p(\mathbb{R}^d)$, $f \in \mathcal{B}_{spq}(\mathbb{R}^d)$ if and only if $P_0(f) \in L^p(\mathbb{R}^d)$ and if there exists a positive sequence $(\epsilon_j)_{j \in \mathbb{N}}$ such that:

$$\|D_j(f)\|_p \leq 2^{-js}\epsilon_j. \tag{13}$$

To express the L^p-norm of $D_j(f)$ in terms of the AMR of $L^2(\mathbb{R}^d)$, we need the following lemma.

Lemma 2. *Let $g_1, \dots g_L$ compactly supported on $\mathbb{R}^d$ satisfying assumptions 1. and 2. of Lemma 1 for $L = 2^d - 1$. Let $f(x) = \sum_{l=1}^L \sum_{k \in \mathbb{Z}^d} \lambda_{kl} 2^{\frac{dj}{2}} g_l(2^j x - k)$. Then there exist $0 < c_1 < c_2$ such that for all $1 \leq p$,*

$$c_1 2^{dj\left(\frac{1}{2}-\frac{1}{p}\right)} \sum_{l=1}^{2^d-1} \left(\sum_{k \in \mathbb{Z}^d} |\lambda_{kl}|^p \right)^{\frac{1}{p}} \leq \|f\|_p \leq c_2 2^{dj\left(\frac{1}{2}-\frac{1}{p}\right)} \sum_{l=1}^{2^d-1} \left(\sum_{k \in \mathbb{Z}^d} |\lambda_{kl}|^p \right)^{\frac{1}{p}}.$$

Lemma 2 is a direct consequence of [21, Lemma 8], using the d-dimensional change of variables formula.

Gathering with (13), we arrive at the following characterization of Besov spaces.

Lemma 3. *Let $p \geq 1$ and $f \in L^p(\mathbb{R}^d)$. Then $f \in \mathcal{B}_{spq}(\mathbb{R}^d)$ if and only if:*

$$\left(\sum_{k\in\mathbb{Z}^d}|\alpha_{0k}|^p\right)^{\frac{1}{p}} + \left(\sum_{j\in\mathbb{N}}\left(2^{j\left(s+d\left(\frac{1}{2}-\frac{1}{p}\right)\right)}\sum_{l=1}^{2^d-1}\left(\sum_{k\in\mathbb{Z}^d}|\beta_{jkl}|^p\right)^{\frac{1}{p}}\right)^{q}\right)^{\frac{1}{q}} < +\infty, \quad (14)$$

where:

$$\alpha_{0k} = \int_{\mathbb{R}^d} f(x)\phi_{0k}(x)dx \text{ and } \beta_{jkl} = \int_{\mathbb{R}^d} f(x)\psi_{jkl}(x)dx.$$

First term in (14) corresponds to the L^p-norm of $P_0(f)$ whereas the second term corresponds to the l^q-norm of $2^{js}\|D_j(f)\|_p$.

This characterization of Besov spaces will be useful to control the complexity of $\mathcal{B}_{spq}(\mathbb{R}^d)$ in this sequence space. For other characterizations, we refer to [22] or [28], including the most usual definition in terms of modulus of continuity.

2.2. Local complexity of Besov balls

First error bounds for empirical risk minimization go back to Vapnik (see [31]). Consider an ERM estimator $\hat{f}_{ERM}$ over a collection of classifiers $\mathcal{F}$, [31] states that:

$$R(\hat{f}_{ERM}) - \inf_{f\in\mathcal{F}} R(f) \leq 2\sup_{f\in\mathcal{F}}|R_n(f) - R(f)|\,. \quad (15)$$

This leads to the study of the supremum of an empirical process. With concentration inequalities, this random process can be controled by its expectation, up to some residual terms. The behaviour of the maximum of the empirical process gives rise to a specific notion of size fot the class $\mathcal{F}$, called global size. This measure is related to the worst deviation of the empirical error to the true error, and the obtained bounds might be loose.

Recently, sharp bounds have been established using different localized versions of (15). It is now common to use localized averages. Considering the penalized empirical minimization (10) using the hinge loss $l(y, f(x)) = (1 - yf(x))_+$, we are interesting in:

$$\mathbb{E}\sup_{f\in B(R):\mathbb{E}f(X)^2\leq r}\left|\frac{1}{n}\sum_{i=1}^{n} l(Y_i, f(X_i)) - \mathbb{E}l(Y, f(X))\right|, \quad (16)$$

where in the sequel $B(R) = \{f \in \mathcal{B}_{spq}(\mathbb{R}^d) : \|f\|_{spq} \leq R\}$. Parameter r allows us to identify locally the scale of richness of the function class. It really measures the magnitude of the error deviation of functions with small variance, which are the one that are likely to be picked by the learning algorithm. From the lipschitz property of the hinge loss, gathering with the well-known symmetrization device (originally in [30]), it turns out that there is a tight connection between such a quantity and the following expectation:

$$\mathbb{E} \sup_{f \in B(R): \mathbb{E}f(X)^2 \leq r} \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i f(X_i) \right|, \tag{17}$$

where ϵ_i, $i = 1, \dots, n$ are i.i.d. with $P(\epsilon_1 = 1) = P(\epsilon_1 = -1) = \frac{1}{2}$. The ϵ_i are called Rademacher variables. (17) is called the local Rademacher average of $B(R)$. The use of Rademacher averages in Classification goes back to [14] (see also [5, 3, 2]).

[20] has proved that the local Rademacher average of a kernel class is determined by the spectrum of its integral operator (see also [6]). Under assumptions on the law of X, we propose a same type of result for Besov classes. The following theorem is the meaty part to deduce statistical performances of minimization (10). It allows us to control the local average (16) and to obtain an oracle inequality (Proposition 1).

Theorem 1. *Suppose P_X admits a density ρ such that:*

- *$a \leq \rho(x) \leq A$ for any $x \in \mathcal{X}$;*
- *ρ has compact support $\mathcal{P} = \{x \in \mathcal{X} : \rho(x) \neq 0\}$.*

Then if $s > \frac{d}{p}$ and $1 \leq p \leq 2$, there exists a constant c depending on a and A such that:

$$\forall r > 0, \, \mathbb{E} \sup_{f \in B(R): \mathbb{E}f(X)^2 \leq r} \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i f(X_i) \right| \leq \frac{c}{\sqrt{n}} R^{\frac{d}{2u}} r^{\frac{s-\frac{d}{p}}{2u}},$$

where $u = s + d\left(\frac{1}{2} - \frac{1}{p}\right)$.

A detailed proof is presented in Section 5. As mentioned in the introduction, we use wavelet theory presented above. More precisely, Lemma 3 allows us to control (17) in a sequence space.

Remark 1. Consider the Sobolev case in dimension 1. For $p = q = 2$ and $d = 1$, the upper bound becomes $R^{\frac{1}{2s}} r^{\frac{2s-1}{4s}}$. It corresponds to the upper bound of [20, Theorem 2.1] for eigenvalues of the integral operator such that $\lambda_k \leq k^{-2s}$. It illustrates that this result generalizes the Sobolev case $p = q = 2$.

Remark 2. This result holds for parameter range of Besov spaces such that $s > \frac{d}{p}$. In this case, there exists a continuous embedding from $\mathcal{B}_{spq}(\mathbb{R}^d)$ into $C(\mathbb{R}^d)$. It ensures that the evaluation functional $\delta_x : f \mapsto f(x)$ is continuous on $\mathcal{B}_{spq}(\mathbb{R}^d)$, exactly as in the RKHS case. In the sequel we consider Besov spaces with such a restriction.

3. Statistical performances

To state learning rates to the Bayes, we act in two steps. First step is to state an oracle inequality of the form:

$$\mathbb{E}R_l(\hat{f}_n, f^*) \leq C \inf_{f \in \mathcal{B}_{spq}(\mathbb{R}^d)} \left(R_l(f, f^*) + \alpha_n \|f\|_{spq}^2 \right) + \delta_n.$$

The statistical sense of this inequality is rather transparent. It ensures classifier $\hat{f}_n$ to have comparable performances with the best classifier called oracle (which minimizes the true risk), up to a residual term δ_n such that $\delta_n \to 0$ as $n \to \infty$. Constant C has to be close to 1.

It remains to control the right hand side of the oracle inequality. The main term of this bound is called the approximation function, defined in this case as:

$$a(\alpha_n) = \inf_{f \in \mathcal{B}_{spq}(\mathbb{R}^d)} \left(R_l(f, f^*) + \alpha_n \|f\|_{spq}^2 \right).$$

Following [16], we use the theory of interpolation spaces to control this function under an assumption over the Bayes f^*. Finally, to get rates of convergence such as (3), it remains to note that:

$$\mathbb{E}R(\hat{f}_n, f^*) \leq \mathbb{E}R_l(\hat{f}_n, f^*).$$

3.1. Oracle inequality

To obtain good statistical properties, we need to restrict the class of considered distributions P. A standard way is to impose a margin hypothesis over the conditional probability function η. In this work we will assume that there exist $\eta_0, \eta_1 > 0$ such that:

$$\forall x \in \mathcal{X}, \left| \eta(x) - \frac{1}{2} \right| \geq \eta_0 \text{ and } \min(\eta(x), 1 - \eta(x)) \geq \eta_1. \tag{18}$$

This assumption is closely related to the margin assumption originally due to [29]. The first part ensures a jump of the probability η at the level $\frac{1}{2}$. The second part is not natural. It avoids the no noise case where $\eta(x) \in \{0, 1\}$. It appears for some technical reasons discussed in Section 5 (see also [6]).

Proposition 1 (Oracle inequality). *Let P the joint distribution such that the marginal of X satisfies assumptions of Theorem 1. Suppose (18) holds for some $\eta_0, \eta_1 > 0$. Consider a non-decreasing function ϕ on $\mathbb{R}^+$ such that $\phi(0) = 0$ and $\phi(x) \geq x$ for $x \geq \frac{1}{2}$.*

Given (X_i, Y_i), $i = 1, \ldots, n$ i.i.d. from P, we define:

$$\hat{g}_n = \arg \min_{f \in \mathcal{B}_{spq}(\mathbb{R}^d)} \left(\frac{1}{n} \sum_{i=1}^{n} l(Y_i, f(X_i)) + \alpha_n \phi(\|f\|_{spq}) \right), \tag{19}$$

where $s > \frac{d}{p}$ and $1 \leq p \leq 2$. If we choose α_n such that:

$$\alpha_n \geq c_1 n^{-\frac{2u}{2u+d}} + \eta_1^{-1} \left(c_2 \frac{\log n}{n} + c_3 \frac{\log \log n}{n} + \frac{c_4}{n} \right), \tag{20}$$

then the estimator $\hat{g}_n$ is such that:

$$\begin{aligned} \mathbb{E} R_l(\hat{g}_n, f^*) &\leq 2 \inf_{f \in \mathcal{B}_{spq}(\mathbb{R}^d)} \left(R_l(f, f^*) + \alpha_n \phi(\|f\|_{spq}) \right) \\ &+ 4\alpha_n \left(2\phi(2) + c\frac{\eta_1}{\eta_0} \right) + \frac{2}{n}, \end{aligned}$$

where c, c_1, c_2, c_3 and c_4 are absolute constants and $u = s + d\left(\frac{1}{2} - \frac{1}{p}\right)$.

Remark 3. It holds whatever $\phi : \phi(0) = 0$ and $\phi(x) \geq x$ for $x \geq \frac{1}{2}$. From the model selection approach, the minimum required regularization is of order $\|f\|_{spq}$. In the standard SVM, a regularization of order $\|f\|_{spq}^2$ is used. Thus we only consider in Corollary 1 the two cases $\phi(x) = x$ and $\phi(x) = 2x^2$. These two orders of regularization will lead to different statistical performances.

Remark 4. This inequality is independent of the approximation term. The choice of α_n in (20) only depends on the hypothesis set we consider. A control of the approximation power of Besov spaces will give adaptive learning rates.

3.2. Rates of convergence

Last step is to control the approximation term in the oracle inequality of Proposition 1. The theory of interpolation spaces allows us to measure how well the models approximate the target function f^*. We finally get the following rates of convergence.

Corollary 1 (Rates of convergence). *Let P satisfying assumptions of Proposition 1. Then for any $1 \leq p \leq 2$ and $s > \frac{d}{p}$, define the estimators*

$$\hat{f}_n := \arg \min_{f \in \mathcal{B}_{spq}(\mathbb{R}^d)} \left(\frac{1}{n} \sum_{i=1}^n l(Y_i, f(X_i)) + \alpha_n \|f\|_{spq} \right)$$

and

$$\hat{g}_n := \arg \min_{f \in \mathcal{B}_{spq}(\mathbb{R}^d)} \left(\frac{1}{n} \sum_{i=1}^n l(Y_i, f(X_i)) + 2\alpha_n \|f\|_{spq}^2 \right).$$

Suppose that $f^* \in \mathcal{B}_{rp\infty}(\mathbb{R}^d)$ for some $r > 0$. Then there exist absolute constants C, $C' > 0$ such that:

$$\mathbb{E} R(\hat{f}_n, f^*) \leq C n^{-\frac{r}{s}\frac{2u}{2u+d}}, \tag{21}$$

and

$$\mathbb{E} R(\hat{g}_n, f^*) \leq C' n^{-\frac{r}{2s-r}\frac{2u}{2u+d}}, \tag{22}$$

where we choose α_n such that an equality holds in (20).

Remark 5. We consider two special cases for the function ϕ of Proposition 1. Estimator $\hat{f}_n$ is the penalized empirical minimizer using the weakest regularization (linear with respect to the norm) whereas $\hat{g}_n$ uses the standard SVM penalization (of order $\|f\|^2$). We can see coarsely that the rate of $\hat{f}_n$ outperforms the one of $\hat{g}_n$ since $\frac{r}{s} > \frac{r}{2s-r}$. With this approach, a lighter regularization results in a better bound.

Remark 6. The construction of these estimators does not depend on the regularity of the Bayes. The smoothing parameter α_n is chosen independently of the parameter r appearing in the assumption $f^* \in \mathcal{B}_{rp\infty}(\mathbb{R}^d)$. As a result, estimators $\hat{f}_n$ and $\hat{g}_n$ are called adaptive. They adapt to the regularity of the Bayes.

Remark 7. [16] gives learning rates for SVM using Sobolev spaces. In particular, under a strong margin assumption, we obtain $n^{-\frac{2rs}{2rs+d(2s-r)}}$. We can compare this bound with (22) for $p = q = 2$. In this case we have $n^{-\frac{r}{2s-r}\frac{2s}{2s+d}}$. This rate is clearly slower than $n^{-\frac{2rs}{2rs+d(2s-r)}}$ since $s > r$. However it gives similar results when $s \to r$.

3.3. Fast rates and optimality

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

$$\text{card}\{x \in \mathbb{R} : f^* \text{ jumps at } x\} = N < \infty. \tag{23}$$

It means that the Bayes rule changes only a finite number of times over the real line. Under this assumption, SVM algorithm using Sobolev spaces cannot reach fast rates (see [16]). In this paper Besov spaces allow us to consider values of $p < 2$. With (23), if $1 \leq p = q < 2$, we have using [17]:

$$f^* \in \mathcal{B}_{rpq}(\mathbb{R}) \text{ for } r = \frac{1}{2} + \frac{1}{p}.$$

Consequently, f^* such that (23) holds belongs to $\mathcal{B}_{r11}(\mathbb{R}) \subset \mathcal{B}_{r1\infty}(\mathbb{R})$ for $r = 3/2$. Substituing into (22), the rate becomes $n^{-\frac{6s-3}{2s(4s-3)}}$ which is a fast rate for s small enough. This example illustrates the importance to consider Besov spaces with $p < 2$ as hypothesis space. For $p < 2$, these spaces contain piecewise regular functions with local discontinuities. It gives fast rates of convergence.

An interesting question thought it is out of the scope of this paper is the optimality of Corollary 1. To answer to this question, it should be possible to link the assumption of regularity of f^* in the d-dimensional case with more standard complexity assumption for classification. For example, a more classical model for possible Bayes rule is the Hölder Boundary Fragment assumption over the decision boundary ([29]). Using the characterization of Besov norms with wavelet coefficients, it should be interesting to link Corollary 1 to this framework. It may give a direction to deduce minimax facts in our framework, using for instance lower bounds of [24].

4. Conclusion

We have studied a new procedure of penalized empirical risk minimization using Besov spaces. This method generalizes SVM algorithm using Sobolev spaces as RKHS. The introduction of Besov spaces gives more flexibility to study the approximation power of the procedure. For the estimation part of the analysis, we adopt the model selection approach of [6]. We propose a control of the local Rademacher average of Besov balls. We hence obtain fast learning rates to the Bayes. Moreover, the construction of these estimators does not depend on the regularity of the Bayes. They are adaptive with respect to the regularity of f^*.

From technical point of view, this paper generalizes the control of Rademacher to a non Hilbertian functional space. It is well-known that local Rademacher of RKHS balls can be controled using RKHS formalism. Here we propose to use a wavelet analysis to get a similar result for Besov spaces. A compactly supported wavelet basis allows us to work in a sequence space.

This contribution could be compared with another introduction of wavelet theory in classification. [15] studies the statistical performances of the LASSO estimator, solving the minimization:

$$\min_{f\in\mathcal{F}^d}\left(\sum_{i=1}^n \mathbb{1}(f(X_i)\neq Y_i)+\alpha\|f\|_{L^1}\right).$$

The hypothesis space $\mathcal{F}^d$ is made of piecewise constant classifiers on a dyadic regular grid of $[0,1]^d$. It allows to decompose each classifier into a fundamental system of indicators on dyadic sets of $[0,1]^d$. This system is closely related to the wavelet tensor product of the Haar basis. As a consequence, in all the proofs, similitudes with the technics used in the wavelet literature are granted. From this point of view, the present work can be compared to [15].

Unfortunately, from practical point of view, the presence of Besov norms in our procedure leads to some computational problems. Besov spaces are not Hilbert spaces. As a result, our method cannot be embedded into a kernel method and computed as SVM algorithm. The feature space is not a RKHS in this case.

Recently, several authors investigate learning algorithm with non Hilbertian hypothesis space. [9] underlines the main principles of an hypothesis space in a learning problem. The hypothesis set must be composed of pointwise defined functions. Moreover the evaluation functional $\delta_x : f \mapsto f(x)$ must be continuous. Due to the embedding theorem, Besov spaces $\mathcal{B}_{spq}(\mathbb{R}^d)$ with $s>\frac{d}{p}$ have this property. In the RKHS case, it corresponds to the reproducing property. It gives a reproducing kernel lying in the RKHS. However the Hilbertian structure is not necessary. To generalize the notion of RKHS to RKS (Reproducing Kernel Space), we need a bilinear form corresponding to the scalar product for RKHS. It could be done with the duality map, considering a duality couple $(\mathcal{H},\mathcal{H}^*)$. [18] establishes an equivalence between particular dualities called evaluation subdualities and a set of weakly continuous applications called reproducing kernels. [9] also provides an explicit construction of both subdualities and the associated reproducing kernel. It is a generalization to the construction of RKHS using Carleman operator. The construction is based on the duality map.

Finally we know from [21] that $(\mathcal{B}_{spq}(\mathbb{R}^d), \mathcal{B}_{-sp'q'}(\mathbb{R}^d))$ are in duality through the duality map:

$$<f,g>_{\mathcal{B}_{spq}(\mathbb{R}^d),\mathcal{B}_{-sp'q'}(\mathbb{R}^d)}=<P_0(f),P_0(g)>_{L^2(\mathbb{R}^d)}+\sum_{j\geq 0}<D_j(f),D_j(g)>_{L^2(\mathbb{R}^d)},$$

where $\mathcal{B}_{-sp'q'}(\mathbb{R}^d)$ is the space of distributions such that (13) holds for $-s<0$. As a result, it will be interesting in this direction to find a kernel generating Besov spaces as RKS. Last step would be to implement our procedure with such a kernel.

5. Proofs

This section is dedicated to the proofs of the main results of this paper. Throughout this section, c denotes a constant that may vary from line to line. For p, $q>0$, we write p', q' such that $1/p+1/p'=1/q+1/q'=1$. Finally, with some abuse of notations, we write $\mathbb{E}f$ for $\mathbb{E}_{P_X}f(X)$ and $\mathbb{E}l(f)$ for $\mathbb{E}_P l(Y,f(X))$.

5.1. *Proof of Theorem 1*

Since the marginal of X admits a bounded density ρ with compact support $\mathcal{P}$, we have:

$$\left\{f\in B(R):\mathbb{E}f^2\leq r\right\}\subseteq\left\{f\in B(R):\int_{\mathcal{P}}f^2(x)dx\leq\frac{r}{a}\right\}:=\mathcal{F}(R,r).$$

Moreover $X_1,\dots X_n$ are i.i.d. from ρ. Then:

$$\mathbb{E}\sup_{f\in\mathcal{F}(R,r)}\left|\frac{1}{n}\sum_{i=1}^n\epsilon_i f(X_i)\right|=\mathbb{E}\sup_{f\in B(R):\|f\|^2_{L^2(\mathbb{R}^d)}\leq\frac{r}{a}}\left|\frac{1}{n}\sum_{i=1}^n\epsilon_i f(X_i)\right|. \tag{24}$$

We then have to bound the RHS of (24).

Let begin with the one-dimensional case, i.e. when the input domain $\mathcal{X}\subset\mathbb{R}$. From wavelet decomposition, we can write $f\in L^2(\mathbb{R})$ as:

$$f=\sum_{k\in\mathbb{Z}}\alpha_{0k}\phi_{0k}+\sum_{j\geq 0}\sum_{k\in\mathbb{Z}}\beta_{jk}\psi_{jk}=f_{\alpha,\beta}. \tag{25}$$

The description of Besov spaces using wavelets leads to the following equivalent norm:

$$\|f\|_{spq}=\left(\sum_{k\in\mathbb{Z}}|\alpha_{0k}|^p\right)^{\frac{1}{p}}+\left(\sum_{j\geq 0}2^{jq(s+\frac{1}{2}-\frac{1}{p})}\left(\sum_{k\in\mathbb{Z}}|\beta_{jk}|^p\right)^{\frac{q}{p}}\right)^{\frac{1}{q}}.$$

Moreover from Lemma 2,

$$\|f\|_2 \approx \left(\sum_{k\in\mathbb{Z}} |\alpha_{0k}|^2\right)^{\frac{1}{2}} + \sum_{j\geq 0}\left(\sum_{k\in\mathbb{Z}} |\beta_{jk}|^2\right)^{\frac{1}{2}},$$

where $x \approx y$ means there exist c, $C > 0$ such that $cy \leq x \leq Cy$.

For $f \in \mathcal{B}_{spq}(\mathbb{R})$ with $s > \frac{1}{p}$, the wavelet expansion (25) is pointwise since the evaluation functionals are continuous in these spaces (see Remark 2). Thus we obtain:

$$\mathbb{E} \sup_{f\in B(R):\|f\|^2_{L^2}\leq \frac{r}{a}} \left|\frac{1}{n}\sum_{i=1}^n \epsilon_i f(X_i)\right| \leq \mathbb{E} \sup_{(\alpha,\beta)\in\Gamma(R,r)} \left|\frac{1}{n}\sum_{i=1}^n \epsilon_i f_{\alpha,\beta}(X_i)\right|,$$

where $f_{\alpha,\beta}$ is defined in (25) and

$$\Gamma(R,r) = \left\{(\alpha,\beta) : \gamma_{pq}(\alpha,\beta) \leq R \text{ and } \gamma_2(\alpha,\beta) \leq \frac{\sqrt{r}}{\sqrt{ac}}\right\},$$

for

$$\gamma_{pq}(\alpha,\beta) = \left(\sum_{k\in\mathbb{Z}} |\alpha_{0k}|^p\right)^{\frac{1}{p}} + \left(\sum_{j\geq 0}\left(2^{j\left(s+d\left(\frac{1}{2}-\frac{1}{p}\right)\right)}\left(\sum |\beta_{jk}|^p\right)^{\frac{1}{p}}\right)^q\right)^{\frac{1}{q}},$$

and

$$\gamma_2(\alpha,\beta) = \left(\sum_{k\in\mathbb{Z}} |\alpha_{0k}|^2\right)^{\frac{1}{2}} + \sum_{j\geq 0}\left(\sum_{k\in\mathbb{Z}} |\beta_{jk}|^2\right)^{\frac{1}{2}}.$$

Hence we get for any integer d':

$$\begin{aligned}
\left|\sum_{i=1}^n \epsilon_i f_{\alpha,\beta}(X_i)\right| &= \left|\sum_{k\in\mathbb{Z}} \alpha_{0k}\sum_{i=1}^n \epsilon_i\phi_{0k}(X_i) + \sum_{j\geq 0}\sum_{k\in\mathbb{Z}}\beta_{jk}\sum_{i=1}^n \epsilon_i\psi_{jk}(X_i)\right| \\
&\leq \left|\sum_{k\in\mathbb{Z}} \alpha_{0k}\sum_{i=1}^n \epsilon_i\phi_{0k}(X_i) + \sum_{j=0}^{d'}\sum_{k\in\mathbb{Z}}\beta_{jk}\sum_{i=1}^n \epsilon_i\psi_{jk}(X_i)\right| \\
&+ \left|\sum_{j>d'}\sum_{k\in\mathbb{Z}}\beta_{jk}\sum_{i=1}^n \epsilon_i\psi_{jk}(X_i)\right| := T_1 + T_2.
\end{aligned}$$

To prove the inequality, we will bound this two terms separately.

We begin applying the Hölder (twice) and Jensen inequalities to T_2:

$$\begin{aligned}\mathbb{E}[T_2] &\leq \mathbb{E}\sum_{j>d'}\left(\sum_{k\in\mathbb{Z}}|\beta_{jk}|^p\right)^{\frac{1}{p}}\left(\sum_{k\in\mathbb{Z}}|\sum_{i=1}^n\epsilon_i\psi_{jk}(X_i)|^{p'}\right)^{\frac{1}{p'}}\\ &\leq \sum_{j>d'}\left(\sum_{k\in\mathbb{Z}}|\beta_{jk}|^p\right)^{\frac{1}{p}}\left(\sum_{k\in\mathbb{Z}}\mathbb{E}|\sum_{i=1}^n\epsilon_i\psi_{jk}(X_i)|^{p'}\right)^{\frac{1}{p'}}\\ &\leq \left(\sum_{j>d'}\left(2^{j(s+\frac{1}{2}-\frac{1}{p})}\left(\sum_{k\in\mathbb{Z}}|\beta_{jk}|^p\right)^{\frac{1}{p}}\right)^q\right)^{\frac{1}{q}}\times\\ &\quad\left(\sum_{j>d'}\left(2^{-j(s+\frac{1}{2}-\frac{1}{p})}\left(\sum_{k\in\mathbb{Z}}\mathbb{E}\left|\sum_{i=1}^n\epsilon_i\psi_{jk}(X_i)\right|^{p'}\right)^{\frac{1}{p'}}\right)^{q'}\right)^{\frac{1}{q'}}.\end{aligned}$$

The definition of $\Gamma(R,r)$ leads to:

$$\mathbb{E}[T_2]\leq R\left(\sum_{j>d'}\left(2^{-j(s+\frac{1}{2}-\frac{1}{p})}\left(\sum_{k\in\mathbb{Z}}\mathbb{E}\left|\sum_{i=1}^n\epsilon_i\psi_{jk}(X_i)\right|^{p'}\right)^{\frac{1}{p'}}\right)^{q'}\right)^{\frac{1}{q'}},\tag{26}$$

where $\frac{1}{p}+\frac{1}{p'}=\frac{1}{q}+\frac{1}{q'}=1$.

Next step is to control, for all $j>d'$, the serie:

$$\sum_{k\in\mathbb{Z}}\mathbb{E}\left|\sum_{i=1}^n\epsilon_i\psi_{jk}(X_i)\right|^{p'}.$$

Lemma 4. *Let $Y_1,\dots Y_n$ i.i.d. with zero mean and σ^2 variance. Then for all $u\geq 2$, there exists $c_u>0$ such that:*

$$\mathbb{E}\left|\frac{1}{n}\sum_{i=1}^nY_i\right|^u\leq c_u\left[\frac{\sigma^u}{n^{\frac{u}{2}}}+\frac{\mathbb{E}|Y_1|^u}{n^{u-1}}\right].$$

This concentration inequality is due to Rosenthal ([23]).

Putting $Y_i=\epsilon_i\psi_{jk}(X_i)$, we have with Lemma 2, gathering with conditions on the density ρ:

$$\mathbb{E}|Y_i|^p\leq A\|\psi_{jk}\|_p^p\leq c2^{j\left(\frac{p}{2}-1\right)},$$

for an absolute constant c depending on A. As a result, applying Lemma 4 for $u=p'\geq 2$, we obtain:

$$\mathbb{E}\left|\frac{1}{n}\sum_{i=1}^n\epsilon_i\psi_{jk}(X_i)\right|^{p'}\leq c_{p'}n^{-\frac{p'}{2}}\left[c^{\frac{p'}{2}}+c\left(\frac{2^j}{n}\right)^{\frac{p'}{2}-1}\right].$$

Now it is worth noticing that since p and the wavelets functions ψ_{jk} are compactly supported, the quantity:

$$\mathbb{E}\psi_{jk}^{p'} = \int_{\mathbb{R}} |\psi_{jk}(x)|^{p'} p(x)dx$$

is zero whatever $k \in \mathcal{S}^C(j) := \{k \in \mathbb{Z} : \text{supp}\psi_{jk} \cap \mathcal{P} = \varnothing\}$. We know from [21] that there exists a constant $c > 0$ such that $\sharp\mathcal{S}(j) \leq c2^j$. Then:

$$\begin{aligned}
\sum_{k\in\mathbb{Z}} \mathbb{E}\left|\frac{1}{n}\sum_{i=1}^n \epsilon_i \psi_{jk}(X_i)\right|^{p'} &= \sum_{k\in\mathcal{S}(j)} \mathbb{E}\left|\frac{1}{n}\sum_{i=1}^n \epsilon_i \psi_{jk}(X_i)\right|^{p'} \\
&\leq c2^j n^{-\frac{p'}{2}} \left[c^{\frac{p'}{2}} + c\left(\frac{2^j}{n}\right)^{\frac{p'}{2}-1}\right].
\end{aligned}$$

Gathering with (26), we obtain:

$$\begin{aligned}
\mathbb{E}[T_2] &\leq c_{p'}\frac{R}{\sqrt{n}}\left(\sum_{j>d'} 2^{-j\left(s+\frac{1}{2}-\frac{1}{p}\right)q'}\left(\sum_{k\in\mathcal{S}(j)} c^{\frac{p'}{2}} + c\left(\frac{2^j}{n}\right)^{\frac{p'}{2}-1}\right)^{\frac{q'}{p'}}\right)^{\frac{1}{q'}} \\
&\leq c_{p'}\frac{R}{\sqrt{n}}\left(\sum_{j>d'} 2^{-jq'\left(s-\frac{1}{2}\right)}\left(c^{\frac{p'}{2}} + c\left(\frac{2^j}{n}\right)^{\frac{p'}{2}-1}\right)^{\frac{q'}{p'}}\right)^{\frac{1}{q'}} \\
&\leq c\frac{R}{\sqrt{n}}\left(\sum_{j>d'} 2^{-jq'\left(s-\frac{1}{2}\right)} + 2^{-jq'\left(s-\frac{1}{p}\right)} n^{\frac{q'}{p'}\left(1-\frac{p'}{2}\right)}\right)^{\frac{1}{q'}} \\
&\leq c\frac{R}{\sqrt{n}}\left(2^{-d'q'\left(s-\frac{1}{2}\right)} + 2^{-d'q'\left(s-\frac{1}{p}\right)} n^{\frac{q'}{p'}\left(1-\frac{p'}{2}\right)}\right)^{\frac{1}{q'}},
\end{aligned}$$

where the convergence of the geometric serie comes from the condition $s > \frac{1}{p}$. Moreover we have $p \leq 2$. Then $s - \frac{1}{2} \geq s - \frac{1}{p}$ and $1 - \frac{p'}{2} \leq 0$. We obtain:

$$\begin{aligned}
\mathbb{E}[T_2] &\leq c\frac{R}{\sqrt{n}}\left(2^{-d'q'\left(s-\frac{1}{p}\right)}\left(1 + n^{\frac{q'}{p'}\left(1-\frac{p'}{2}\right)}\right)\right)^{\frac{1}{q'}} \\
&\leq c\frac{R}{\sqrt{n}} 2^{-d'\left(s-\frac{1}{p}\right)}.
\end{aligned}$$

Last step is to control T_1. We put $\beta_{-1,k} = \alpha_{0k}$ and $\psi_{-1k} = \phi_{0k}$. Thus we have, applying successively Cauchy-Schwarz and Jensen inequalities,

$$\begin{aligned}\mathbb{E}[T_1] &\leq \mathbb{E}\left|\sum_{j=-1}^{d'}\left(\sum_{k\in\mathbb{Z}}\beta_{jk}^2\right)^{\frac{1}{2}}\left(\sum_{k\in\mathbb{Z}}\left(\sum_{i=1}^{n}\epsilon_i\psi_{jk}(X_i)\right)^2\right)^{\frac{1}{2}}\right| \\ &\leq \sum_{j=-1}^{d'}\left(\sum_{k\in\mathbb{Z}}\beta_{jk}^2\right)^{\frac{1}{2}}\left(\sum_{k\in\mathbb{Z}}\mathbb{E}\left(\sum_{i=1}^{n}\epsilon_i\psi_{jk}(X_i)\right)^2\right)^{\frac{1}{2}}.\end{aligned}$$

Besides, $\mathbb{E}\epsilon_i\epsilon_j = 0$, $\forall i \neq j$. Then:

$$\mathbb{E}\left(\sum_{i=1}^{n}\epsilon_i\psi_{jk}(X_i)\right)^2 = \sum_{i=1}^{n}\mathbb{E}\psi_{jk}(X_i)^2.$$

We have to control, for $j \in \{-1, \dots d'\}$ the serie:

$$\sum_{k\in\mathbb{Z}}\sum_{i=1}^{n}\mathbb{E}\psi_{jk}(X_i)^2.$$

As above, since p and the wavelet mother ψ are compactly supported,

$$\sum_{k\in\mathbb{Z}}\sum_{i=1}^{n}\mathbb{E}\psi_{jk}(X_i)^2 = \sum_{k\in\mathcal{S}(j)}\sum_{i=1}^{n}\mathbb{E}\psi_{jk}(X_i)^2$$

where $\sharp\mathcal{S}(j) \leq c2^j$. Finally we obtain:

$$\begin{aligned}\mathbb{E}[T_1] &\leq c\sqrt{n}\sum_{j=-1}^{d'}\left(\sum_{k\in\mathbb{Z}}\beta_{jk}^2\right)^{\frac{1}{2}}2^{\frac{j}{2}} \\ &\leq c\sqrt{n}2^{\frac{d'}{2}}\sum_{j=-1}^{d'}\left(\sum_{k\in\mathbb{Z}}\beta_{jk}^2\right)^{\frac{1}{2}} \\ &\leq 2^{\frac{d'}{2}}c\frac{\sqrt{rn}}{\sqrt{a}},\end{aligned}$$

where last line comes from the definition of $\Gamma(R, r)$.

Then there exists a constant $c > 0$ depending on a and A such that:

$$\mathbb{E}\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i f_{\alpha,\beta}(X_i)\right| \leq \frac{c}{\sqrt{n}}\inf_{d'\in\mathbb{N}}\left(R2^{-d'\left(s-\frac{1}{p}\right)} + 2^{\frac{d'}{2}}\sqrt{\frac{r}{a}}\right).$$

Optimizing with respect to d', we obtain the following upper bound in dimension 1:

$$\frac{c}{\sqrt{n}}R^{\frac{1}{2\left(\frac{1}{2}+s-\frac{1}{p}\right)}}r^{\frac{s-\frac{1}{p}}{2\left(\frac{1}{2}+s-\frac{1}{p}\right)}}.$$

Now we turn out into the d-dimensional case. The principle of the proof follows the one dimensional case. From (12) we have, for any $f \in \mathcal{B}_{spq}(\mathbb{R}^d)$:

$$f = \sum_{k \in \mathbb{Z}^d} \alpha_{0k}\phi_{0k} + \sum_{j \geq 0} \sum_{k \in \mathbb{Z}^d} \sum_{l=1}^{2^d-1} \beta_{jkl}\psi_{jkl} = f_{\alpha,\beta}(x),$$

where the equality is pointwise since $s > \frac{d}{p}$. Then we can write:

$$\mathbb{E} \sup_{f \in B(R): \|f\|_{L^2}^2 \leq \frac{r}{a}} \left| \frac{1}{n} \sum_{i=1}^n \epsilon_i f(X_i) \right| \leq \mathbb{E} \sup_{(\alpha,\beta) \in \Gamma_d(R,r)} \left| \frac{1}{n} \sum_{i=1}^n \epsilon_i f_{\alpha,\beta}(X_i) \right|,$$

where now:

$$\Gamma_d(R,r) = \left\{ (\alpha,\beta) : \gamma_{pq}^d(\alpha,\beta) \leq R \text{ and } \gamma_2^d(\alpha,\beta) \leq \frac{\sqrt{r}}{\sqrt{ac}} \right\},$$

for

$$\gamma_{pq}^d(\alpha,\beta) = \left(\sum_{k \in \mathbb{Z}^d} |\alpha_k|^p \right)^{\frac{1}{p}} + \left(\sum_{j \geq 0} \left(2^{j\left(s+d(\frac{1}{2}-\frac{1}{p})\right)} \sum_{l=1}^{2^d-1} \left(\sum_{k \in \mathbb{Z}^d} |\beta_{jkl}|^p \right)^{\frac{1}{p}} \right)^q \right)^{\frac{1}{q}},$$

and

$$\gamma_2^d(\alpha,\beta) = \left(\sum_{k \in \mathbb{Z}^d} |\alpha_k|^2 \right)^{\frac{1}{2}} + \sum_{j \geq 0} \sum_{l=1}^{2^d-1} \left(\sum_{k \in \mathbb{Z}^d} |\beta_{jkl}|^2 \right)^{\frac{1}{2}}.$$

We proceed as in dimension 1. For any integer d':

$$\begin{aligned}\left| \sum_{i=1}^n \epsilon_i f_{\alpha,\beta}(X_i) \right| &\leq \left| \sum_{k \in \mathbb{Z}^d} \alpha_{0k} \sum_{i=1}^n \epsilon_i \phi_{0k}(X_i) + \sum_{j=0}^{d'} \sum_{l=1}^{2^d-1} \sum_{k \in \mathbb{Z}^d} \beta_{jkl} \sum_{i=1}^n \epsilon_i \psi_{jkl}(X_i) \right| \\ &+ \left| \sum_{j > d'} \sum_{l=1}^{2^d-1} \sum_{k \in \mathbb{Z}^d} \beta_{jkl} \sum_{i=1}^n \epsilon_i \psi_{jkl}(X_i) \right| := T_3 + T_4.\end{aligned}$$

We apply the Hölder (twice) and Jensen inequalities to T_4 to get:

$$\mathbb{E}[T_4] \leq R \left(\sum_{j > d'} \left(2^{-j\left(s+d\left(\frac{1}{2}-\frac{1}{p}\right)\right)} \left(\sum_{k \in \mathbb{Z}^d} \sum_{l=1}^{2^d-1} \mathbb{E} \left| \sum_{i=1}^n \epsilon_i \psi_{jkl}(X_i) \right|^{p'} \right)^{\frac{1}{p'}} \right)^{q'} \right)^{\frac{1}{q'}}.$$

Next step is to control, for all $j > d'$, the serie:

$$\sum_{k \in \mathbb{Z}^d} \sum_{l=1}^{2^d-1} \mathbb{E} \left| \sum_{i=1}^n \epsilon_i \psi_{jkl}(X_i) \right|^{p'}.$$

We have with Lemma 2:

$$\mathbb{E}|Y_i|^p \leq A\|\psi_{jkl}\|_p^p \leq c2^{dj\left(\frac{p}{2}-1\right)},$$

since ψ is compactly supported. As a result, applying the Rosenthal inequality, we obtain:

$$\mathbb{E}\left|\frac{1}{n}\sum_{i=1}^n \epsilon_i\psi_{jkl}(X_i)\right|^{p'} \leq c_{p'}n^{-\frac{p'}{2}}\left[c^{\frac{p'}{2}} + c\left(\frac{2^{dj}}{n}\right)^{\frac{p'}{2}-1}\right].$$

Now it is worth noticing that since p and the wavelets ψ_{jkl} are compactly supported, the quantity

$$\mathbb{E}|\psi_{jkl}|^{p'} = \int_{\mathbb{R}} |\psi_{jkl}(x)|^{p'}p(x)dx$$

is zero whatever $k \notin \mathcal{S}_d(j) := \{k \in \mathbb{Z}^d : supp(\psi_{jkl}) \cap \mathcal{P} \neq \varnothing\}$. There exists an absolute constant $c > 0$ which only depends on d such that $\sharp\mathcal{S}_d(j) \leq c2^{dj}$. As a result,

$$\sum_{k\in\mathbb{Z}^d}\sum_{l=1}^{2^d-1}\mathbb{E}\left|\sum_{i=1}^n \epsilon_i\psi_{jkl}(X_i)\right|^{p'} = \sum_{k\in\mathcal{S}_d(j)}\sum_{l=1}^{2^d-1}\mathbb{E}\left|\sum_{i=1}^n \epsilon_i\psi_{jkl}(X_i)\right|^{p'}.$$

With previous inequality, we hence have:

$$\begin{aligned}
\mathbb{E}[T_4] \quad \leq \quad & c_{p'}\frac{R}{\sqrt{n}}\left(\sum_{j>d'} 2^{-j\left(s+d\left(\frac{1}{2}-\frac{1}{p}\right)\right)q'}\left(\sum_{k\in\mathcal{S}_d(j)} c^{\frac{p'}{2}} + c\left(\frac{2^{dj}}{n}\right)^{\frac{p'}{2}-1}\right)^{\frac{q'}{p'}}\right)^{\frac{1}{q'}} \\
& \leq c\frac{R}{\sqrt{n}}\left(\sum_{j>d'} 2^{-jq'\left(s-\frac{d}{2}\right)} + 2^{-jq'\left(s-\frac{d}{p}\right)}n^{\frac{q'}{p'}\left(1-\frac{p'}{2}\right)}\right)^{\frac{1}{q'}} \\
& \leq c\frac{R}{\sqrt{n}}\left(2^{-d'q'\left(s-\frac{d}{2}\right)} + 2^{-d'q'\left(s-\frac{d}{p}\right)}n^{\frac{q'}{p'}\left(1-\frac{p'}{2}\right)}\right)^{\frac{1}{q'}},
\end{aligned}$$

where the condition $s > \frac{d}{p}$ ensures the convergence of the geometric serie. Moreover we have $p \leq 2$. Then, as above:

$$\mathbb{E}[T_4] \leq c\frac{R}{\sqrt{n}}2^{-d'\left(s-\frac{d}{p}\right)}.$$

It remains to control T_3. For brievity, we put $\beta_{-1k1} = \alpha_k$, $\psi_{-1k1} = \phi_k$ and for any $l > 1$, $\beta_{-1kl} = 0$ $\psi_{-1kl} = 0$. Applying successively Cauchy-Schwarz and Jensen, one has:

$$\mathbb{E}\,T_3 \leq \sum_{j=-1}^{d'}\left(\sum_{k\in\mathbb{Z}^d}\sum_{l=1}^{2^d-1}\beta_{jkl}^2\right)^{\frac{1}{2}}\left(\sum_{k\in\mathbb{Z}^d}\sum_{l=1}^{2^d-1}\mathbb{E}\left(\sum_{i=1}^n \epsilon_i\psi_{jkl}(X_i)\right)^2\right)^{\frac{1}{2}}.$$

With the same argument as in dimension one, we obtain:

$$\mathbb{E}\left(\sum_{i=1}^n \epsilon_i \psi_{jkl}(X_i)\right)^2 = \sum_{i=1}^n \mathbb{E}\psi_{jkl}(X_i)^2.$$

We have to control, for $j \in \{-1, \dots d'\}$ the serie

$$\sum_{k\in\mathbb{Z}^d} \sum_{l=1}^{2^d-1} \sum_{i=1}^n \mathbb{E}\psi_{jkl}(X_i)^2.$$

Since f and the wavelet mother ψ are compactly supported,

$$\sum_{k\in\mathbb{Z}^d} \sum_{l=1}^{2^d-1} \sum_{i=1}^n \mathbb{E}\psi_{jkl}(X_i)^2 = \sum_{k\in\mathcal{S}_d(j)} \sum_{l=1}^{2^d-1} \sum_{i=1}^n \mathbb{E}\psi_{jkl}(X_i)^2$$

where $\sharp\mathcal{S}_d(j) \leq c2^{dj}$. Finally we get:

$$\mathbb{E}[T_3] \leq c\sqrt{n} \sum_{j=-1}^{d'} \left(\sum_{k\in\mathbb{Z}^d} \sum_{l=1}^{2^d-1} |\beta_{jkl}|^2\right)^{\frac{1}{2}} 2^{\frac{dj}{2}} \leq \sqrt{nr}2^{\frac{dd'}{2}},$$

from the definition of $\Gamma_d(R,r)$.

Then the control of the two terms entails:

$$\mathbb{E}\left|\frac{1}{n}\sum_{i=1}^n \epsilon_i f_{\alpha,\beta}(X_i)\right| \leq \frac{c}{\sqrt{n}} \inf_{d'\in\mathbb{N}} \left(R2^{-d'\left(s-\frac{d}{p}\right)} + 2^{\frac{dd'}{2}}\sqrt{r}\right).$$

Optimizing with respect to d', we lead to the conclusion.

5.2. *Proof of Proposition 1*

To prove the oracle inequality, we use the following model selection approach. From [6], minimization (10) can be rewritten as $\hat{f}_n = \hat{f}_{\widehat{R}}$ where:

$$\begin{aligned} \hat{f}_R &= \arg\min_{f\in B(R)} \frac{1}{n}\sum_{i=1}^n (1-Y_i f(X_i))_+ \quad \text{and} \\ \widehat{R} &= \arg\min_{R>0} \left(\frac{1}{n}\sum_{i=1}^n (1-Y_i\hat{f}_R(X_i))_+ + \alpha_n R^2\right), \end{aligned}$$

where $B(R) = \{f \in \mathcal{B}_{spq}(\mathbb{R}^d) : \|f\|_{spq} \leq R\}$. This gives a model selection interpretation of classifier $\hat{f}_n$, where models are balls in $\mathcal{B}_{spq}(\mathbb{R}^d)$. We can then apply the following general model selection theorem (Theorem 5 in [6]). We recall it for completeness.

Theorem 2. *Let l a loss function such that $g^* \in \arg\min_{f \in L^2(P_X)} \mathbb{E}l(Y, f(X))$. Let $(\mathcal{G}_m)_{m\in\mathcal{M}}$ a countable collection of models with $\mathcal{G}_m \subset L^2(P_X)$, $\forall m \in \mathcal{M}$. We suppose there exist a pseudo-distance d on $L^2(P_X)$, a sequence of subroot[1] functions $(\phi_m)_{m\in\mathcal{M}}$, and two sequences of positive numbers $(b_m)_{m\in\mathcal{M}}$ and $(C_m)_{m\in\mathcal{M}}$ such that:*

(H1) $\forall m \in \mathcal{M}, \forall g \in \mathcal{G}_m$, $\|l(g)\|_\infty \leq b_m$.
(H2) $\forall g, g' \in L^2(P_X)$, $Var\left(l(Y, g(X)) - l(Y, g'(X))\right) \leq d^2(g, g')$.
(H3) $\forall m \in \mathcal{M}, \forall g \in \mathcal{G}_m$, $d^2(g, g^*) \leq C_m \mathbb{E}\left(l(Y, g(X)) - l(Y, g(X)^*)\right)$.
(H4) $\forall m \in \mathcal{M}, \forall g_0 \in \mathcal{G}_m, \forall r \geq r_m^*$:

$$\mathbb{E}\left[\sup_{g\in\mathcal{G}_m : d(g,g_0)^2 \leq r} (\mathbb{E} - \hat{E}_n)(l(Y, g(X)) - l(Y, g(X)_0))\right] \leq \phi_m(r),$$

where r_m^* satisfies $\phi_m(r_m^*) = r_m^*/C_m$ and $\hat{\mathbb{E}}_n X = \frac{1}{n}\sum X_i$.

Let $(x_m)_{m\in\mathcal{M}}$ a real sequence such that $\sum_{m\in\mathcal{M}} e^{-x_m} \leq 1$ and:

$$\forall m,\, m' \in \mathcal{M},\, x_m \leq x_{m'} \Rightarrow b_m \leq b_{m'} \text{ and } C_m \leq C_{m'}.$$

Let $(\rho_m)_{m\in\mathcal{M}}$ a family of positive numbers. Let $\tilde{g}$ such that there exists $\tilde{m} \in \mathcal{M}$ with $\tilde{g} \in \mathcal{G}_{\tilde{m}}$ and:

$$\frac{1}{n}\sum_{i=1}^n l(Y_i, \tilde{g}(X_i)) + pen(\tilde{m}) \leq \inf_{m\in\mathcal{M}} \inf_{g\in\mathcal{G}_m} \left(\sum_{i=1}^n l(Y_i, g(X_i)) + pen(m) + \rho_m\right).$$

Then if $m \mapsto pen(m)$ satisfies, for any $m \in \mathcal{M}$:

$$pen(m) \geq 250K\frac{r_m^*}{C_m} + \frac{B_m(x_m + \log 2)}{3n} + \frac{B_m \log B_m}{n}, \tag{27}$$

where $B_m = 75KC_m + 28b_m$, we obtain:

$$\mathbb{E}R_l(\tilde{g}, f^*) \leq \frac{K + \frac{1}{5}}{K - 1} \inf_{m\in\mathcal{M}} \left(\inf_{g\in\mathcal{G}_m} R_l(g, g^*) + 2pen(m) + \rho_m + \frac{2}{n}\right). \tag{28}$$

Theorem 2 is rather general. It can be applied to a wide variety of situations related to many statistical models. In particular it can be used to propose adaptive estimators in non-parametric regression, density estimation or classification. It gives the minimum required penalty to get an oracle inequality for a penalized empirical cost minimizer.

In our setup, we have to find constant b_R, C_R, a subroot function ϕ_R and a distance d on $L^2(P_X)$ such that:

(H1) $\forall R \in \mathbb{R}^+, \forall g \in B(R), \|l(g)\|_\infty \leq b_R$;
(H2) $\forall g, g' \in L^2(P_X), Var\left(l(g) - l(g')\right) \leq d^2(g, g')$;

[1] $\phi : \mathbb{R}^+ \to \mathbb{R}^+$ is a subroot function if it is a positive non-decreasing function such that $\phi(r)/\sqrt{r}$ is non-increasing.

(H3) $\forall R \in \mathbb{R}^+, \forall g \in B(R), d^2(g, f^*) \leq C_R \mathbb{E}\left(l(g) - l(f^*)\right)$;
(H4) $\forall R \in \mathbb{R}^+, \forall r > 0$, we have

$$\mathbb{E} \sup_{f \in B(R): d(f,0)^2 \leq r} \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i f(X_i) \right| \leq \phi_R(r).$$

Once assumptions **(H1)-(H4)** are granted, next step is to discretize the continuous family of models $(B(R))_{R \in \mathbb{R}^+}$ over a certain family of values of the radii. Following [6], we consider the set of discretized radii:

$$\mathcal{R} = \{M^{-1} 2^k, k \in \mathbb{N}, 0 \leq k \leq \lceil \log_2 n \rceil\}.$$

To apply the second part of Theorem 5 in [6], the penalty function should satisfy:

$$\operatorname{pen}(R) \geq c_1 \left(\frac{r^*}{C_R} + \frac{(C_R + b_R)(x_R + \log 2)}{3n} + \frac{(C_R + b_R)\log(C_R + b_R))}{n} \right),$$

where c_1 is a suitable constant. It can be checked that condition (20) on α_n ensures such an inequality for:

$$\operatorname{pen}(R) = \alpha_n \left(\phi\left(\frac{MR}{2}\right) + \frac{\eta_1}{\eta_0} \right).$$

Last step is to forth between the discretized framework and the continuous framework. We follow exactly [6] to write $\hat{g}_n$ defined in (19) as an approximate penalized minimum empirical risk estimator of Theorem 2 over the family $(B(R))_{R \in \mathcal{R}}$.

It only remains to prove **(H1)-(H4)**.

5.2.1. *Proof of* ***(H1)***

We only consider in Proposition 1 a parameter range of Besov spaces $\mathcal{B}_{spq}(\mathbb{R}^d)$ such that $s > \frac{d}{p}$. As a result, from the continuous embedding of $\mathcal{B}_{spq}(\mathbb{R}^d)$ into $C(\mathbb{R}^d)$ for $s > \frac{d}{p}$, one gets for any $f \in \mathcal{B}_{spq}(\mathbb{R}^d)$:

$$\|f\|_\infty \leq c \|f\|_{spq}.$$

We hence obtain (H_1) with $b_R = 1 + cR$ since $|l(y, f(x))| \leq 1 + |f(x)|$.

5.2.2. *Proof of* ***(H2)-(H3)***

To check these assumptions, we have to choose a distance d in $L^2(P_X)$. This choice has been done implicitely in Theorem 1. This theorem will prove **(H4)** with the usual distance $d(g, g') = \mathbb{E}(g - g')^2$, for any $g, g' \in L^2(P_X)$. It comes from Section 3.2.1 which allows us to write the L^2-norm of a function in $\mathcal{B}_{spq}(\mathbb{R}^d)$, using wavelet decomposition. Then we consider the same distance to check **(H2)** and **(H3)**.

(H2) is trivially satisfied because the hinge loss l is a Lipschitz function. Moreover with Lemma 11 of [6], hypothesis (18) ensures **(H3)** with constant $C_R = 2\left(\frac{MR}{\eta_1} + \frac{1}{\eta_0}\right)$. The choice of the distance above corresponds to the setting (S1) in [6]. That's why the second part of (18) is necessary in our context.

*5.2.3. Proof of **(H4)***

The proof of **(H4)** has been done in Section 3.2.2.

5.3. Proof of Corollary 1

We only treat the particular case $\phi(x) = x$. From Proposition 1, we have in this case:

$$\mathbb{E}R_l(\hat{f}_n, f^*) \leq 2 \inf_{f\in\mathcal{B}_{spq}(\mathbb{R}^d)} \left(R_l(f,f^*) + \alpha_n\|f\|_{spq}\right) + 4\alpha_n\left(4 + c\frac{\eta_1}{\eta_0}\right) + \frac{2}{n},$$

which gives, since l is the hinge loss, a control on the excess risk of $\hat{f}_n$ as follows:

$$\mathbb{E}R(\hat{f}_n, f^*) \leq 2 \inf_{f\in\mathcal{B}_{spq}(\mathbb{R}^d)} \left(R_l(f,f^*) + \alpha_n\|f\|_{spq}\right) + 4\alpha_n\left(4 + c\frac{\eta_1}{\eta_0}\right) + \frac{2}{n}. \quad (29)$$

To get Corollary 1, it remains to control the RHS of (29) called the approximation function, defined by:

$$a(\alpha_n) = \inf_{f\in\mathcal{B}_{spq}(\mathbb{R}^d)} \left(R_l(f,f^*) + \alpha_n\|f\|_{spq}\right).$$

By the Lipschitz property of the hinge loss, gathering with assumptions on the marginal of X, we have, for any $p \geq 1$:

$$\begin{aligned} a(\alpha_n) &\leq \inf_{f\in\mathcal{B}_{spq}(\mathbb{R}^d)} \left(A\|f-f^*\|_{L^p(\mathbb{R}^d)} + \alpha_n\|f\|_{spq}\right) \\ &= c \inf_{R\in\mathbb{R}^+} \left(\inf_{f\in B(R)} \|f-f^*\|_{L^p(\mathbb{R}^d)} + \alpha_n R\right), \end{aligned}$$

where c depends on A.

To control the first term above, we use the following result.

Lemma 5. *For any $r < s$,*

$$f^* \in \mathcal{B}_{rp\infty}(\mathbb{R}^d) \Rightarrow \inf_{f\in B(R)} \|f-f^*\|_{L^p(\mathbb{R}^d)} \leq \|f^*\|_{rp\infty}^{\frac{s}{s-r}} \left(\frac{1}{R}\right)^{\frac{r}{s-r}}.$$

Proof. The cornerstone idea in the proof is the use of interpolation spaces. Given two Banach spaces $\mathcal{B}$ and $\mathcal{B}'$, $\theta \in]0,1[$ and $q \in [0,\infty]$, the space $(\mathcal{B},\mathcal{B}')_{\theta,q}$ called *interpolation space between $\mathcal{B}$ and $\mathcal{B}'$* consists of all $f \in \mathcal{B}$ such that

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

is finite, where $P_t(f)$ is a norm in $\mathcal{B}$ called the Peetre's functional (see [27] for a definition).

Here we are interested in the case $q = \infty$ and the following geometric explanation of interpolation space [25, Theorem 3.1]:

$$f \in (\mathcal{B}, \mathcal{B}')_{\theta,\infty} \Longrightarrow \inf_{g \in B_{\mathcal{B}'}(R)} \|f - g\|_B \leq \|f\|_{\theta,\infty}^{\frac{1}{1-\theta}} \big(\frac{1}{R}\big)^{\frac{\theta}{1-\theta}}, \tag{30}$$

where $B_{\mathcal{B}'}(R) := \{f \in \mathcal{B}' : \|f\|_{\mathcal{B}'} \leq R\}$. It means that the distance of any function in $(\mathcal{B}, \mathcal{B}')_{\theta,\infty}$ to the ball $B_{\mathcal{B}'}(R)$ tends to zero with a given rate of convergence. This approximation problem arose from the study of approximation error in learning theory, where usually $\mathcal{B} = L^2$ and $\mathcal{B}' = \mathcal{H}_K$ a reproducing kernel Hilbert space ([25]). Here we propose a generalization to the Banach case with Besov spaces. We use in particular the following stability of Besov spaces in terms of interpolation spaces:

$$\forall 0 < \theta < 1,\ (L^p(\mathbb{R}^d), \mathcal{B}_{spq}(\mathbb{R}^d))_{\theta,\infty} = \mathcal{B}_{\gamma p \infty}(\mathbb{R}^d),$$

where $\gamma = \theta s$. From (30) with $\theta = \frac{r}{s}$, we conclude the proof of Lemma 5. ■

Using this lemma and optimizing with respect to R leads to:

$$a(\alpha_n) \leq c \inf_{R \in \mathbb{R}+} \left(\left(\frac{1}{R}\right)^{\frac{r}{s-r}} + \alpha_n R \right) \leq c\alpha_n^{\frac{r}{s}}.$$

Finally going back to (29), we arrive at:

$$\mathbb{E}R(\hat{f}_n), f^*) \leq 2c\alpha_n^{\frac{r}{s}} + 4\alpha_n \left(4 + c\frac{\eta_1}{\eta_0}\right) + \frac{2}{n}.$$

Choosing α_n such that an equality holds in (20) concludes the proof.

References

[1] Audibert, J.Y. and Tsybakov, A.B. (2007). Fast learning rates for plug-in classifiers. *The Annals of Statistics 35 (2)*, 608–633. MR2336861

[2] Bartlett, P.L., Boucheron, S., and Lugosi, G. (2002). Model selection and error estimation. *Machine Learning 48*, 85–113.

[3] Bartlett, P.L., Bousquet, O., and Mendelson, S. (2005). Local rademacher complexities. *The Annals of Statistics 33 (4)*, 1497–1537. MR2166554

[4] Bartlett, P.L., Jordan, M.I., and McAuliffe, J.D. (2006). Convexity, classification, and risk bounds. *J. Amer. Statist. Assoc. 101 (473)*, 138–156. MR2268032

[5] Bartlett, P.L. and Mendelson, S. (2002). Rademacher and Gaussian complexities: risk bounds and structural results. *Journal of Machine Learning Research 3*, 463–482. MR1984026

[6] Blanchard, G., Bousquet, O., and Massart, P. (2008). Statistical performance of support vector machines. *Annals of Statistics 36 (2)*. MR2396805

[7] Blanchard, G., Lugosi, G., and Vayatis, N. (2003). On the rate of convergence of regularized boosting classifiers. *Journal of Machine Learning Research 4*, 861–894. MR2076000

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

[9] Canu, S., Mary, X., and Rakotomamonjy, A. (2003). Functional learning through kernel. *Advances in Learning Theory: Methods, Models and Applications 190*, 89–110.

[10] Daubechies, I. (1988). Orthonormal bases of compactly supported wavelets. *Communications on Pure and Applied Mathematics 41 (7)*, 909–996. MR0951745

[11] Devroye, L., Györfi, L., and Lugosi, G. (1996). *A Probabilistic Theory of Pattern Recognition*. Springer-Verlag. MR1383093

[12] Donoho, D.L., Johnstone, I.M., Kerkyacharian, G., and Picard, D. (1996). Density estimation by wavelet thresholding. *The Annals of Statistics 24 (2)*, 508–539. MR1394974

[13] Härdle, W., Kerkyacharian, G., Picard, D., and Tsybakov, A. (1997). *Wavelets, Approximation, and Statistical Applications*. Lecture Notes in Statistics. MR1618204

[14] Koltchinskii, V. (2001). Rademacher penalties and structural risk minimization. *IEEE Transactions on Information Theory 47 (5)*, 1902–1914. MR1842526

[15] Lecué, G. (2008). Classification with minimax fast rates for classes of Bayes rules with sparse representation. *Electronic Journal of Statistics 2*, 741–773. MR2430253

[16] Loustau, S. (2008). Aggregation of SVM classifiers using Sobolev spaces. *Journal of Machine Learning Research 9*, 1559–1582. MR2426051

[17] Mallat, S. (2000). *Une exploration des signaux en ondelettes*. Ellipses.

[18] Mary, X., De Brucq, D., and Canu, S. (2003). Sous-dualités et noyaux (reproduisants) associés. *C. R. Acad. Sc. Paris 336 (1)*, 949–954. MR1994600

[19] Massart, P. and Nédélec, E. (2006). Risk bounds for statistical learning. *The Annals of Statistics 34 (5)*, 2326–2366. MR2291502

[20] Mendelson, S. (2003). On the performance of kernel classes. *Journal of Machine Learning Research 4*, 759–771. MR2075996

[21] Meyer, Y. (1990). *Ondelettes et Opérateurs 1 : Ondelettes*. Hermann. MR1085487

[22] Peetre, J. (1976). *New thoughts on Besov spaces*. Mathematics Department, Duke University, Durham, N.C. MR0461123

[23] Rosenthal, H.P. (1972). On the span in l_p of sequences of independent random variables. *Israël J. Math. 8*, 273–303. MR0271721

[24] Scott, C. and Nowak, R. (2006). Minimax-optimal classification with dyadic decision trees. *IEEE Transactions on Information Theory 52-4*, 1335–1353. MR2241192

[25] Smale, S. and Zhou, D.X. (2003). Estimating the approximation error in learning theory. *Analysis and Applications 1 (1)*, 17–41. MR1959283

[26] Steinwart, I. and Scovel, C. (2007). Fast rates for support vector machines using Gaussian kernels. *The Annals of Statistics 35 (2)*, 575–607. MR2336860

[27] Triebel, H. (1978). *Interpolation Theory, Function Spaces, Differential Operators*. North-Holland Publishing Company. MR0503903

[28] Triebel, H. (1992). *Theory of Functions Spaces II*. Birkhauser. MR1163193

[29] Tsybakov, A.B. (2004). Optimal aggregation of classifiers in statistical learning. *The Annals of Statistics 32 (1)*, 135–166. MR2051002

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

[31] Vapnik, V.N. and Chervonenkis, A.Ya. (1974). *Theory of Pattern Recognition*. Nauka, Moscow. MR0474638