Classification with Minimax Fast Rates for Classes of Bayes Rules with Sparse Representation

Guillaume Lecué
Université Paris VI

Abstract

We construct a classifier which attains the rate of convergence $\log n/n$ under sparsity and margin assumptions. An approach close to the one met in approximation theory for the estimation of function is used to obtain this result. The idea is to develop the Bayes rule in a fundamental system of $L^2([0,1]^d)$ made of indicator of dyadic sets and to assume that coefficients, equal to $-1, 0$ or 1, belong to a kind of L^1-ball. This assumption can be seen as a sparsity assumption, in the sense that the proportion of coefficients non equal to zero decreases as "frequency" grows. Finally, rates of convergence are obtained by using an usual trade-off between a bias term and a variance term.

1 Introduction

Consider a measurable space $(\mathcal{X}, \mathcal{A})$ and π a probability measure on this space. Denote by $D_n = (X_i, Y_i)_{1\leq i\leq n}$ n observations of (X, Y) a random variable with values in $\mathcal{X} \times \{-1, 1\}$ distributed according to π. We want to construct measurable functions which associate a label $y \in \{-1, 1\}$ to each point x of $\mathcal{X}$, such functions are called *prediction rules*. The quality of a prediction rule f is given by the value

$$R(f) = \mathbb{P}(f(X) \neq Y)$$

called *misclassification error of* f. It is well known (e.g. Devroye et al. [1996]) that there exists an optimal prediction rule which attains the minimum of R over all measurable functions with values in $\{-1,1\}$. It is called *Bayes rule* and defined by

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

where η is the *conditional probability function of* $Y = 1$ *knowing* X defined by

$$\eta(x) = \mathbb{P}(Y = 1|X = x).$$

The value

$$R^* = R(f^*) = \min_f R(f)$$

is known as the *Bayes risk*. The aim of classification is to construct a prediction rule, using the observations D_n, which has a risk as close to R^* as possible. Such a construction is called a *classifier*. Performance of a classifier $\hat{f}_n$ is measured by the value

$$\mathcal{E}_\pi(\hat{f}_n) = \mathbb{E}_\pi[R(\hat{f}_n) - R^*]$$

called *excess risk of* $\hat{f}_n$. In this case $R(\hat{f}_n) = \mathbb{P}(\hat{f}_n(X) \neq Y|D_n)$ and $\mathbb{E}_\pi$ denotes the expectation w.r.t. D_n when the probability distribution of (X_i, Y_i) is π for any $i = 1, \ldots, n$. We say that a classifier $\hat{f}_n$ learns with the convergence rate $\phi(n)$, where $(\phi(n))_{n\in\mathbb{N}}$ is a decreasing sequence, if an absolute constant $C > 0$ exists such that for any integer n, $\mathbb{E}_\pi[R(\hat{f}_n) - R^*] \leq C\phi(n)$.

We introduce a loss function on the set of all prediction rules:

$$d_\pi(f, g) = |R(f) - R(g)|.$$

This loss is a *semi-distance* (it is symmetric, satisfies the triangle inequality and $d_\pi(f, f) = 0$). For all classifiers $\hat{f}_n$, it is linked to the excess risk by

$$\mathcal{E}_\pi(\hat{f}_n) = \mathbb{E}_\pi[d_\pi(\hat{f}_n, f^*)],$$

where the RHS is the risk of $\hat{f}_n$ associated to the loss d_π. In classification we can consider three estimation problems. The first one is estimation of the Bayes rule f^*, the second one is estimation of the conditional probability function η and the last one is estimation of the probability π. Usually, estimation of η involves smoothness assumption on the conditional function η. However, global smoothness assumptions on η are somehow too restrictive for the estimation of f^* since the behavior of η away from the decision boundary $\{x \in \mathcal{X} : \eta(x) = 1/2\}$ may have no effect on the estimation of f^*.

In this paper we deal directly with estimation of f^*. But, in this case, the main difficulty of the classification problem is the dependence on π of the loss d_π (usually, we use a loss free from π, which upper bounds d_π to obtain rates of convergence). Moreover, using the loss d_π, we don't have the usual bias/variance trade-off, unlike many other estimation problems. This is due to the fact that we do not have an approximation theory in classification for the loss d_π. This gap is due to the difficulty that d_π depends on π, thus, this theory has to be uniform on π. We need approximation results of the form:

$$\forall \pi = (P^X, \eta) \in \mathcal{P}, \forall \epsilon > 0, \exists f_\epsilon \in \mathcal{F}_\epsilon : d_\pi(f_\epsilon, f^*) \leq \epsilon, \tag{1}$$

where P^X is the marginal distribution of π on $\mathcal{X}$, $f^* = \text{sign}(2\eta - 1)$, $\mathcal{P}$ is a set of probability measures on $\mathcal{X} \times \{-1, 1\}$ and the family of classes of prediction rules $(\mathcal{F}_\epsilon)_{\epsilon>0}$ is decreasing ($\mathcal{F}_\epsilon \subset \mathcal{F}_{\epsilon'}$ if $\epsilon' < \epsilon$) and $\mathcal{F}_\epsilon$ is less complex than $\{f^* : \pi \in \mathcal{P}\}$, in fact we expect $\mathcal{F}_\epsilon$ to be parametric. Similar results appear in density estimation literature, where, for instance, $\mathcal{P}$ is replaced by the set of all probability measures with a density with respect to the Lebegue measure lying in an L^1-ball and $\mathcal{F}_\epsilon$ is replaced by the set of all functions with a finite number (depending on ϵ) of coefficients non equal to zero in the decomposition in the chosen orthogonal basis. But approximation theory in density estimation does not depend on the underlying probability measure since the loss functions used there are generally independent of the underlying statistical problem. In this paper, we deal directly with the estimation of the Bayes rule and obtain convergence result w.r.t. the loss d_π by using an approximation approach of the Bayes rules w.r.t. d_π. Theorems in Section 7 of Devroye et al. [1996] show that no classifier can learn with a given convergence rate for arbitrary underlying probability distribution π. Thus, assumption on f^* has to be done to obtain convergence rates. In this paper, assumption on f^* is close to the one met in density estimation when we assume that the underlying density belongs to an L^1-ball.

Usually, a model (set of measurable functions with values in $\{-1, 1\}$) is considered and we assume that the Bayes rule belongs to this model. In this case the bias is equal to zero and no bound on the approximation term is considered. In Blanchard et al. [2003], question on the control of the approximation error for a class of models in the boosting framework is asked. In this paper, it is assumed that the Bayes rule belongs to the model and nature of distribution satisfying such condition is explored. Another related work is Lugosi and Vayatis [2004], where, under general conditions, it can be guaranteed that the approximation error converges to zero for some specific models. In the present paper, bias term is not taken equal to zero and convergence rates for the approximation error are obtained depending on the complexity of the considered model (cf. Theorem 2).

We consider the classification problem on $\mathcal{X} = [0,1]^d$. All the results can be generalized to a given compact of $\mathbb{R}^d$. Like in many other works on the classification problem an upper bound for the loss d_π is used. But, in our case we still work directly with the estimation of f^*. For a prediction rule f we have

$$d_\pi(f_1, f^*) = \mathbb{E}[|2\eta(X) - 1| \mathbb{1}_{f_1(X) \neq f^*(X)}] \leq ||f_1 - f^*||_{L^1(P^X)}. \qquad (2)$$

In order to get a distribution-free loss function, we assume that the following assumption holds

(A1) *The marginal P^X is absolutely continuous w.r.t. the Lebesgue measure λ_d and* $0 < a \leq dP^X(x)/d\lambda_d \leq A < +\infty, \quad \forall x \in [0,1]^d$.

This is a technical assumption used for the control of the P^X measure of some subset of $[0,1]^d$. In recent years some assumptions have been introduced to measure a statistical quality of classification problems. The behavior of the regression function η near the level $1/2$ is a key point of the classification's quality (cf. e.g. Tsybakov [2004]). In fact, the closest is η to $1/2$, the more difficult is the classification problem, nevertheless when we have $\eta \equiv 1/2$ the classification is trivial since all prediction rules are Bayes rules. Here, we measure the quality of the classification problem thanks to the following assumption introduced by Massart and Nédélec [2003]:

Strong Margin Assumption (SMA): There exists an absolute constant $0 < h \leq 1$ such that:

$$\mathbb{P}\left(|2\eta(X) - 1| > h\right) = 1.$$

Under assumptions (A1) and (SMA) we have

$$ah||f_1 - f^*||_{L_1(\lambda_d)} \leq d_\pi(f_1, f^*) \leq A||f_1 - f^*||_{L_1(\lambda_d)}.$$

Thus, estimation of f^* w.r.t. the loss d_π is the same as estimation w.r.t. $L_1(\lambda_d)-$norm, where λ_d is the Lebesgue measure on $[0,1]^d$.

The paper is organized as follows. In the next section we propose a representation for functions with values in $\{-1,1\}$ in a fundamental system of $L^2([0,1]^d)$. The third section is devoted to approximation and estimation of Bayes rules having a sparse representation in this system. In the fourth section we discuss about this approach. Proofs are given in the last section.

2 Classes of Bayes Rules with Sparse Representation

Theorem 2 of Subsection 3.1 is about the approximation of the Bayes rules when we assume that f^* belongs to a kind of "L^1-ball" for functions with values in $\{-1,1\}$. The idea is to develop f^* in a fundamental system of $L^2([0,1]^d, P^X)$ (that is a countable family of functions such that the set of all finite linear combinations is dense in $L^2([0,1]^d, P^X)$) inherited from the Haar basis and to control the number of coefficients non equal to zero. In this paper we only consider the case where P^X satisfies (A1). We can extend the study to a more general case by taking another partition of $[0,1]^d$ adapted to P^X.

First we construct such a fundamental system. We consider a sequence of partitions of $\mathcal{X}=[0,1]^d$ by setting for any integer j,

$$\mathcal{I}_{\mathbf{k}}^{(j)} = I_{k_1}^{(j)} \times \ldots \times I_{k_d}^{(j)},$$

where $\mathbf{k}$ is the multi-index

$$\mathbf{k}=(k_1,\ldots,k_d) \in I_d(j) = \{0,1,\ldots,2^j-1\}^d,$$

and for any integer j and any $k \in \{1,\ldots,2^j-1\}$,

$$I_k^{(j)} = \begin{cases} \left[\frac{k}{2^j}, \frac{k+1}{2^j}\right) & \text{if } k=0,\ldots,2^j-2 \\ \left[\frac{2^j-1}{2^j}, 1\right] & \text{if } k=2^j-1 \end{cases}.$$

We consider the family $\mathcal{S} = \left(\phi_{\mathbf{k}}^{(j)} : j \in \mathbb{N}, \mathbf{k} \in I_d(j)\right)$ where

$$\phi_{\mathbf{k}}^{(j)} = \mathbb{1}_{\mathcal{I}_{\mathbf{k}}^{(j)}}, \quad \forall j \in \mathbb{N}, \mathbf{k} \in I_d(j),$$

where $\mathbb{1}_A$ denotes the indicator of a set A. Set $\mathcal{S}$ is a fundamental system of $L^2([0,1]^d, P^X)$. This is the class of indicators of the dyadic sets of $[0,1]^d$.

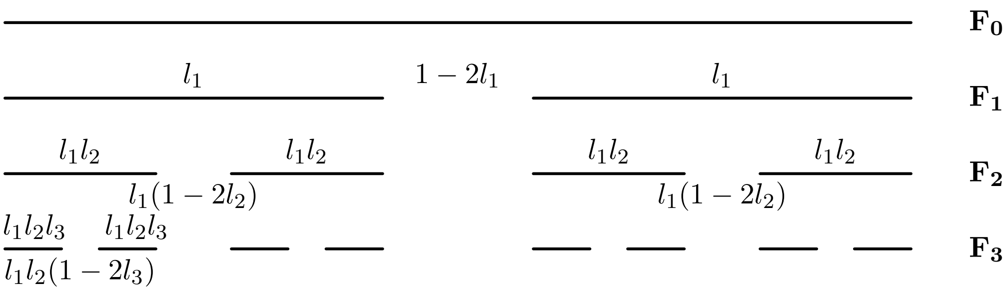
We consider the class of functions f defined $P^X - a.s.$ from $[0,1]^d$ to $\{-1,1\}$ which can be written in this system by

$$f = \sum_{j=0}^{+\infty} \sum_{\mathbf{k} \in I_d(j)} a_{\mathbf{k}}^{(j)} \phi_{\mathbf{k}}^{(j)}, P^X - a.s., \text{ where } a_{\mathbf{k}}^{(j)} \in \{-1,0,1\},$$

where, for any point $x \in [0,1]^d$, the right hand side applied in x is a finite sum. Denote this class by $\mathcal{F}^{(d)}$. In what follows, we use the vocabulary appearing in the wavelet literature. The index "j" of $a_{\mathbf{k}}^{(j)}$ and $\phi_{\mathbf{k}}^{(j)}$ is called "level of frequency". Since $\mathcal{S}$ is not an orthogonal basis of $L^2([0,1]^d, P^X)$, the expansion of f w.r.t. this system is not unique. Therefore, to avoid any ambiguity, we define an unique writing for any mapping f in $\mathcal{F}^{(d)}$ by taking $a_{\mathbf{k}}^{(j)} \in \{-1,1\}$ with preferences for low frequencies when it is possible. Roughly speaking, for $f \in \mathcal{F}^{(d)}$, denoted by $f = \sum_{j=0}^{+\infty} \sum_{\mathbf{k} \in I_d(j)} a_{\mathbf{k}}^{(j)} \phi_{\mathbf{k}}^{(j)}, P^X - a.s.$ where $a_{\mathbf{k}}^{(j)} \in \{-1,0,1\}$, it means that, we construct $A_{\mathbf{k}}^{(j)} \in \{-1,0,1\}, j \in \mathbb{N}, \mathbf{k} \in I_d(j)$, such that, if there exists $J \in \mathbb{N}$ and $\mathbf{k} \in I_d(J)$ such that for all $\mathbf{k}' \in I_d(J+1)$ satisfying $\phi_{\mathbf{k}}^{(J)} \phi_{\mathbf{k}'}^{(J+1)} \neq 0$ we have $a_{\mathbf{k}'}^{(J+1)} = 1$, then we take $A_{\mathbf{k}'}^{(J)} = 1$ and the 2^d other coefficients of higher frequency $A_{\mathbf{k}'}^{(J+1)} = 0$ instead of having these 2^d coefficients equal to 1, and the same convention holds for -1. Moreover if we have $A_{\mathbf{k}}^{(J_0)} \neq 0$ then $A_{\mathbf{k}'}^{(J)} = 0$ for all $J > J_0$ and $\mathbf{k}' \in I_d(J)$ satisfying $\phi_{\mathbf{k}}^{(J_0)} \phi_{\mathbf{k}'}^{(J)} \neq 0$. We can describe a mapping $f \in \mathcal{F}^{(d)}$ satisfying this convention by using a tree. Each knot corresponds to a coefficient $A_{\mathbf{k}}^{(J)}$. The root is $A_{0,\ldots,0}^{(0)}$. If a knot, describing the coefficient $A_{\mathbf{k}}^{(J)}$, equals to 1 or -1 then it has no branches, otherwise it has 2^d branches, corresponding to the 2^d coefficients at the following frequency, describing the coefficients $A_{\mathbf{k}'}^{(J+1)}$ for $\mathbf{k}'$ satisfying $\phi_{\mathbf{k}}^{(J)} \phi_{\mathbf{k}'}^{(J+1)} \neq 0$. At the end all the leaves of the tree equals to 1 or -1, and the depth of a leaf is the frequency of the coefficient associated. The writing convention says that a knot can not have all his leaves equal to 1 together (or -1). In this case we write this mapping by putting a 1 at the knot (or -1). In what follows we say that a function $f \in \mathcal{F}^{(d)}$ satisfies the writing convention (W) when f is written in $\mathcal{S}$ using the writing convention describes in this paragraph. Remark that this writing convention is not an assumption on the function since we can write all $f \in \mathcal{F}$ using this convention. Representation of the Bayes rules using Dyadic decision trees has been explored by

Nowak and Scott [2004].

Is it possible to write every measurable functions from $[0,1]^d$ to $\{-1,1\}$ in the fundamental system $\mathcal{S}$ using coefficients with values in $\{-1,0,1\}$? Since the family of set $(\mathring{\mathcal{I}}^{(j)}_{\mathbf{k}} : j \in \mathbb{N}, \mathbf{k} \in I_d(j))$, where $\mathring{A}$ denotes the interior of A, is a basis of open subsets of $[0,1]^d$, this question is equivalent to this one: "Take A a Borel of $[0,1]^d$, is it possible to find an open subset $\mathcal{O}$ of $[0,1]^d$ such that the symmetrical difference between A and $\mathcal{O}$ has a Lebesgue measure 0?" Unfortunately, the answer to this last question is negative. There exists $F \subset [0,1]^d$ a Borel, closed, with an empty interior and a positive Lebesgue measure $\lambda_d(F) > 0$. For example, in the one dimension case, the following algorithm yields such a set. Take $(l_k)_{k\geq 1}$ a sequence of numbers defined by $l_k = 1/2 - 1/(k+1)^2$ for any integer k. Denote by F_0 the interval $[0,1]$ and construct a sequence of closed sets $(F_k)_{k\geq 0}$ like in the following picture.



It is easy to check that $F = \cap_{k\geq 0} F_k$ is closed, with an empty interior and a positive Lebesgue measure. For the d-dimensional case, the set $F \times [0,1]^{d-1}$ satisfies the required assumptions. Thus, take F such a set and $\mathcal{O}$ an open subset of $[0,1]^d$. If $\mathcal{O} \subseteq F$ then $\mathcal{O} = \emptyset$ because $\mathring{F} = \emptyset$ and $\lambda_d(F\Delta\mathcal{O}) = \lambda_d(F) > 0$. If $\mathcal{O} \nsubseteq F$ then $\mathcal{O} \cap F^c$ is an open subset of $[0,1]$ none empty, so $\lambda_d(\mathcal{O}\Delta F) \geq \lambda_d(\mathcal{O}\cap F^c) > 0$. Thus, every measurable functions from $[0,1]^d$ to $\{-1,1\}$ can not be written in $\mathcal{S}$ using only coefficients with values in $\{-1,0,1\}$. Nevertheless, the Lebesgue measure satisfies the property of regularity, which says that for any Borel $B \in [0,1]^d$ and any $\epsilon > 0$, there exists a compact subset K and an open subset $\mathcal{O}$ such that $K \subseteq A \subseteq \mathcal{O}$ and

$\lambda_d(\mathcal{O} - K) \le \epsilon$. Hence, one can easily check that for any measurable function f from $[0,1]^d$ to $\{-1,1\}$ and any $\epsilon > 0$, there exists a function $g \in \mathcal{F}^{(d)}$ such that $\lambda_d(\{x \in [0,1]^d : f(x) \neq g(x)\}) \le \epsilon$. Thus, $\mathcal{F}^{(d)}$ is dense in $L^2(\lambda_d)$ intersected with the set of all measurable functions from $[0,1]^d$ to $\{-1,1\}$. Now, we exhibit some usual prediction rules which belong to $\mathcal{F}^{(d)}$.

Definition 1. *Let A be a Borel subset of $[0,1]^d$. We say that A is* **almost everywhere open** *if there exists an open subset $\mathcal{O}$ of $[0,1]^d$ such that $\lambda_d(A\Delta\mathcal{O}) = 0$, where λ_d is the Lebesgue measure on $[0,1]^d$ and $A\Delta\mathcal{O}$ is the symmetrical difference.*

Theorem 1. *Let η be a function from $[0,1]^d$ to $[0,1]$. We consider*

$$f_\eta(x) = \begin{cases} 1 & \text{if } \eta(x) \ge 1/2 \\ -1 & \text{otherwise.} \end{cases}$$

We assume that $\{\eta \ge 1/2\}$ and $\{\eta < 1/2\}$ are almost everywhere open. Thus, there exists $g \in \mathcal{F}$ such that for λ_d-almost every $x \in [0,1]^d$, $\quad g = f_\eta, \lambda_d - a.s.$. For instance, if $\lambda_d(\partial\{\eta = 1/2\}) = 0$ and, either η is λ_d-almost everywhere continuous (it means that there exists an open subset of $[0,1]^d$ with a Lebesgue measure equals to 1 such that η is continuous on this open subset) or if η is λ_d-almost everywhere equal to a continuous function, then $f_\eta \in \mathcal{F}^{(d)}$.

Now, we define a model for the Bayes rule by taking a subset of $\mathcal{F}^{(d)}$. For all functions w defined on $\mathbb{N}$ and with values in $\mathbb{R}^+$, we consider $\mathcal{F}_w^{(d)}$, the model for Bayes rules, made of all prediction rules f which can be written, using the previous writing convention (W), by

$$f = \sum_{j=0}^{+\infty} \sum_{\mathbf{k} \in I_d(j)} a_{\mathbf{k}}^{(j)} \phi_{\mathbf{k}}^{(j)},$$

where $a_{\mathbf{k}}^{(j)} \in \{-1,0,1\}$ and

$$\mathrm{card}\left\{\mathbf{k} \in \mathrm{I_d(j)} : \mathrm{a}_{\mathbf{k}}^{(\mathrm{j})} \neq 0\right\} \le \mathrm{w(j)}, \quad \forall \mathrm{j} \in \mathbb{N}.$$

The class $\mathcal{F}_w^{(d)}$ depends on the choice of the function w. If w is too small then the class $\mathcal{F}_w^{(d)}$ is not very rich, that is the subject of the following Proposition 1. If w is too large then $\mathcal{F}_w^{(d)}$ would be too complex for a good estimation of $f^* \in \mathcal{F}_w^{(d)}$, that is why we introduce Definition 2 in what follows.

Proposition 1. *Let w be a mapping from $\mathbb{N}$ to $\mathbb{R}^+$ such that $w(0) \geq 1$. The two following assertions are equivalent:*

(i) $\mathcal{F}_w^{(d)} \neq \{\mathbb{1}_{[0,1]^d}\}$.

(ii) $\sum_{j=1}^{+\infty} 2^{-dj} \lfloor w(j) \rfloor \geq 1$.

And if w is too large then the approximation by a parametric model will be impossible, that is why we give a particular look on the class of function introduced in the following Definition 2.

Definition 2. *Let w be a mapping from $\mathbb{N}$ to $\mathbb{R}^+$. If w satisfies*

$$\sum_{j=0}^{+\infty} \frac{\lfloor w(j) \rfloor}{2^{dj}} < +\infty, \tag{3}$$

then we say that $\mathcal{F}_w^{(d)}$ is a $\mathbf{L^1}$**−ball of prediction rules***.*

Remark 1. *We say that $\mathcal{F}_w^{(d)}$ is a "L^1-ball" for a function w satisfying (3), because , the sequence $(\lfloor w(j) \rfloor)_{j\in\mathbb{N}}$ belongs to a L^1-ball of $\mathbb{N}^{\mathbb{N}}$, with radius $(2^{dj})_{j\in\mathbb{N}}$. Moreover, definition 2 can be link to the definition of a L^1-ball for real valued functions, since we have a kind of base, given by $\mathcal{S}$, and we have a control on coefficients which increases with the frequency. Control on coefficients, given by (3), is close to the one for coefficients of a real valued function in L^1-ball since it deals with the quality of approximation of the class $\mathcal{F}_w^{(d)}$ by a parametric model.*

Remark 2. *A L^1-ball of prediction rules is made of "sparse" prediction rules. In fact, for $f \in \mathcal{F}_w^{(d)}$, the repartition of coefficients non equal to zero in the decomposition of f at a given frequency becomes sparse as the frequency grows. That is the reason why $\mathcal{F}_w^{(d)}$ can be called a* **sparse class of prediction rules***. For exemple, if $(\lfloor w(j) \rfloor / 2^{dj})_{j\geq 1}$ decreases and (3) holds then number of coefficients non equal to* 0 *at the frequency j is smaller than j^{-1} per cent of the maximal number of coefficients (that is 2^{dj}).*

Remark 3. *If we assume that P^X is known then we can work with any measurable space $\mathcal{X}$ endowed with a Lebesgue measure λ, while assuming that $P^X << \lambda$. In this case, we take $\left(\mathcal{I}_{\mathbf{k}}^{(j)} : j \in \mathbb{N}, \mathbf{k} \in I_d(j)\right)$, such that for any $j \in \mathbb{N}$, $\left(I_{\mathbf{k}}^{(j)} : \mathbf{k} \in I_d(j)\right)$ is* *a partition of $\mathcal{X}$ adapted to the previous one $\left(I_{\mathbf{k}}^{(j-1)} : \mathbf{k} \in I_d(j-1)\right)$ and satisfying $P^X(I_{\mathbf{k}}^{(j)}) = 2^{-jd}$. All the results below can be obtained in this framework.*

Now, examples of functions satisfying (3) are given. Classes $\mathcal{F}_w^{(d)}$ associated to these functions are used in what follows to define statistical models. As an introduction we define the minimal infinite class of prediction rules, by $\mathcal{F}_0^{(d)}$ which is the class $\mathcal{F}_w^{(d)}$ for $w = w_0^{(d)}$ where $w_0^{(d)}(0) = 1$ and $w_0^{(d)}(j) = 2^d - 1$, for all $j \geq 1$. To understand why this class is important we introduce a notion of local oscillation of a prediction rule. This concept defines a kind of "regularity" for functions with values in $\{-1, 1\}$.

Definition 3. *Let f be a prediction rule from $[0,1]^d$ to $\{-1,1\}$ in $\mathcal{F}^{(d)}$. We consider the writing of f in the fundamental system introduce in Section 3.1 with writing convention (W):*

$$f = \sum_{j=0}^{+\infty} \sum_{\mathbf{k} \in I_d(j)} a_{\mathbf{k}}^{(j)} \phi_{\mathbf{k}}^{(j)}, P^X - a.s..$$

Let $J \in \mathbb{N}$ and $\mathbf{k} \in I_d(j)$. We say that $I_{\mathbf{k}}^{(J)}$ is a **low oscillating block** *of f when f has exactly $2^d - 1$ coefficients, in this block, non equal to zero at each level of frequencies greater than $J+1$. In this case we say that f* **has a low oscillating block of frequency** *J.*

Remark that, if f has an oscillating block of frequency J, then f has an oscillating block of frequency J', for all $J' \geq J$. The function class $\mathcal{F}_0^{(d)}$ is made of all prediction rules with one oscillate block at level 1 and of the indicator function $\mathbb{1}_{[0,1]^d}$. If we have $w(j_0) < w_0^{(d)}(j_0)$ for one $j_0 \geq 1$ and $w(j) = w_0^{(d)}(j)$ for $j \neq j_0$ then the associated class $\mathcal{F}_w^{(d)}$ contains only the indicator function $\mathbb{1}_{[0,1]^d}$, that is the reason why we say that $\mathcal{F}_0^{(d)}$ is "minimal".

Nevertheless, the following proposition shows that $\mathcal{F}_0^{(d)}$ is a rich class of prediction rules from a combinatorial point of view. We recall some quantities which measure a combinatorial richness of a class of prediction rules. For any class $\mathcal{F}$ of prediction rules from $\mathcal{X}$ to $\{-1,1\}$, we consider

$$N(\mathcal{F}, (x_1, \ldots, x_m)) = card\left(\{(f(x_1), \ldots, f(x_m)) : f \in \mathcal{F}\}\right)$$

where $x_1, \ldots, x_m \in \mathcal{X}$ and $m \in \mathbb{N}$,

$$S(\mathcal{F}, m) = \max\left(N(\mathcal{F}, (x_1, \ldots, x_m)) : x_1, \ldots, x_m \in \mathcal{X}\right)$$

and the VC-dimension of $\mathcal{F}$ is

$$VC(\mathcal{F}) = \min\left(m \in \mathbb{N} : S(\mathcal{F}, m) \neq 2^m\right).$$

Consider $x_j = \left(\frac{2^j+1}{2^{j+1}}, \frac{1}{2^{j+1}}, \ldots, \frac{1}{2^{j+1}}\right)$, for any $j \in \mathbb{N}$. Thus, for any integer m, we have $N(\mathcal{F}_0^{(d)}, (x_1, \ldots, x_m)) = 2^m$. Hence, the following proposition holds.

Proposition 2. *The class of prediction rules $\mathcal{F}_0^{(d)}$ has an infinite VC-dimension.*

Thus every class $\mathcal{F}_w^{(d)}$ such that $w \geq w_0^{(d)}$ has an infinite VC-dimension (since $w \leq w' \Rightarrow \mathcal{F}_w^{(d)} \subseteq \mathcal{F}_{w'}^{(d)}$), which is the case for the following classes.

Now, we introduce some examples of L^1-ball of Bayes rules. We denote by $\mathcal{F}_K^{(d)}$, for a $K \in \mathbb{N}^*$, the class $\mathcal{F}_w^{(d)}$ of prediction rules where w is equal to the function

$$w_K^{(d)}(j) = \begin{cases} 2^{dj} & \text{if } j \leq K, \\ 2^{dK} & \text{otherwise.} \end{cases}$$

This class is called the **truncated class of level K**.

We consider **exponential classes**. These sets of prediction rules are denoted by $\mathcal{F}_\alpha^{(d)}$, where $0 < \alpha < 1$, and are equal to $\mathcal{F}_w^{(d)}$ when $w = w_\alpha^{(d)}$ and

$$w_\alpha^{(d)}(j) = \begin{cases} 2^{dj} & \text{if } j \leq N^{(d)}(\alpha) \\ 2^{d\alpha j} & \text{otherwise} \end{cases},$$

where $N^{(d)}(\alpha) = \inf\left(N \in \mathbb{N} : 2^{d\alpha N} \geq 2^d - 1\right)$, that is for $N^{(d)}(\alpha) = \lceil \log(2^d - 1)/(d\alpha \log 2) \rceil$.

Remark 4. *For the one-dimensional case, an other point of view is to consider $f^* \in L^2([0,1])$ and to develop f^* in an orthogonal basis of $L^2([0,1])$. Namely,*

$$f^* = \sum_{j \in \mathbb{N}} \sum_{k=0}^{2^j-1} a_k^{(j)} \psi_k^{(j)},$$

where $a_k^{(j)} = \int_0^1 f^*(x) \psi_k^{(j)}(x) dx$ for any $j \in \mathbb{N}$ and $k = 0, \ldots, 2^j - 1$. For the control of the bias term we assume that the family of coefficients $(a_k^{(j)}, j \in \mathbb{N}, k = 0, \ldots, 2^j - 1)$

belongs to a L^1-ball. But this point of view leads to analysis and estimation issues. First problem: Which functions with values in $\{-1,1\}$ have wavelet coefficients in a L^1-ball and which wavelet basis is more adapted to our problem (maybe the Haar basis)? Second problem: Which kind of estimators could be used for the estimation of these coefficients? As we can see, the main problem is that there is no approximation theory for functions with values in $\{-1,1\}$. We do not know how to approach, in $L^2([0,1])$, measurable functions with values in $\{-1,1\}$ by "parametric" functions with values in $\{-1,1\}$. Methods developed in this paper may be seen as a first step in this field. We can generalize this approach to functions with values in $\mathbb{Z}$. Remark that when functions take values in $\mathbb{R}$, that is for the regression problem, usual approximation theory is used to obtain a control on the bias term.

Remark 5. *Other sets of prediction rules are described by the classes $\mathcal{F}_w^{(d)}$ where w is from $\mathbb{N}$ to $\mathbb{R}^+$ and satisfies*

$$\sum_{j\geq 1} a_j \frac{\lfloor w(j) \rfloor}{2^{dj}} \leq L,$$

where $(a_j)_{j\geq 1}$ is an increasing sequence of positive numbers.

3 Rates of Convergence over $\mathcal{F}_w^{(d)}$ under (SMA)

3.1 Approximation Result

Let w be a function from $\mathbb{N}$ to $\mathbb{R}^+$ and $A > 1$, we denote by $\mathcal{P}_{w,A}$ the set of all probability measures π on $[0,1]^d \times \{-1,1\}$ such that the Bayes rules f^*, associated to π, belongs to $\mathcal{F}_w^{(d)}$ and the marginal of π on $[0,1]^d$ is absolutely continuous and one version of its Lebesgue density is upper bounded by A. The following Theorem can be seen as an approximation Theorem for the Bayes rules w.r.t. the loss d_π uniformly in $\pi \in \mathcal{P}_{w,A}$.

Theorem 2 (Approximation Theorem). *Let $\mathcal{F}_w^{(d)}$ be a L^1-ball of prediction rules. We have:*

$$\forall \epsilon > 0, \exists J_\epsilon \in \mathbb{N} : \forall \pi \in \mathcal{P}_{w,A}, \exists f_\epsilon = \sum_{\mathbf{k} \in I_d(J_\epsilon)} B_{\mathbf{k}}^{(J_\epsilon)} \phi_{\mathbf{k}}^{(J_\epsilon)}$$

where $B_{\mathbf{k}}^{(J_\epsilon)} \in \{-1, 1\}$ *and*

$$d_\pi(f^*, f_\epsilon) \leq \epsilon,$$

where f^* *is the Bayes rule associated to* π. *For example,* J_ϵ *can be the smallest integer* J *satisfying* $\sum_{j=J+1}^{+\infty} 2^{-dj} \lfloor w(j) \rfloor < \epsilon/A$.

Remark 6. *No assumption on the quality of the classification problem, like an assumption on the margin, is needed to state Theorem 2. Only assumption on the "number of oscillations" of* f^* *is used. Theorem 2 deals with approximation of functions in the* L^1-*ball* $\mathcal{F}_w^{(d)}$ *by functions with values in* $\{-1, 1\}$ *and no estimation issues are met.*

Remark 7. *Theorem 2 is the first step to prove an estimation theorem using a trade-off between a bias term and a variance term. We write*

$$\mathcal{E}_\pi(\hat{f}_n) = \mathbb{E}_\pi \left[d_\pi(\hat{f}_n, f^*) \right] \leq \mathbb{E}_\pi \left[d_\pi(\hat{f}_n, f_\epsilon) \right] + d_\pi(f_\epsilon, f^*).$$

Since f_ϵ *belongs to a parametric model we expect to have a control of the variance term,* $\mathbb{E}_\pi \left[d_\pi(\hat{f}_n, f_\epsilon) \right]$, *depending on the dimension of the parametric model which is linked to the quality of the approximation in the bias term.*

Remark 8. *Since* $d_\pi(f^*, f_\epsilon) = \mathbb{E}\left[|2\eta(X) - 1| \mathbb{1}_{f^*(X) \neq f_\epsilon(X)} \right]$, *the closest to* $1/2$ η *is, the smallest the bias is. Especially, we have a bias equal to zero when* $\eta = 1/2$ *(in this case any prediction rule is a Bayes rules). Thus, more difficult the problem of estimation is (that is for underlying probability measure* $\pi = (P^X, \eta)$ *with* η *close to* $1/2$*), the smallest the bias is. This behavior does not appear clearly in density estimation.*

3.2 Estimation Result

We consider the following class of estimators indexed by the frequency rank $J \in \mathbb{N}$:

$$\hat{f}_n^{(J)} = \sum_{\mathbf{k} \in I_d(J)} \hat{A}_{\mathbf{k}}^{(J)} \phi_{\mathbf{k}}^{(J)}, \tag{4}$$

where coefficients are defined by

$$\hat{A}_{\mathbf{k}}^{(J)} = \begin{cases} 1 & \text{if } \exists X_i \in I_{\mathbf{k}}^{(J)} \text{ and } \mathrm{card} \left\{ \mathrm{i} : \begin{array}{c} X_i \in I_{\mathbf{k}}^{(J)}, \\ Y_i = 1 \end{array} \right\} > \mathrm{card} \left\{ \mathrm{i} : \begin{array}{c} X_i \in I_{\mathbf{k}}^{(J)}, \\ Y_i = -1 \end{array} \right\}, \\ -1 & \text{otherwise} \end{cases}$$

To obtain a good control of the variance term, we need to assure a good quality of the estimation problem. Therefore, estimation results are obtained in Theorem 3 under (SMA) assumption. In recent years we have understood that (SMA) assumption can lead to fast rates but is not enough to assure any rate of convergence (cf. corolary 1 at the end of section 3.3), thus we have to define a model for η or f^*, here we use a L^1-ball of prediction rules as a model for f^*.

Theorem 3 (estimation Theorem)**.** *Let $\mathcal{F}_w^{(d)}$ be a L^1-ball of prediction rules. Let π be a probability measure on $[0,1]^d \times \{-1,1\}$ satisfying assumptions (A1) and (SMA), and such that the Bayes rule, associated to ^pi, belongs to $\mathcal{F}_w^{(d)}$. The excess risk of the classifier $\hat{f}_n^{(J_\epsilon)}$ satisfies for any positive number ϵ,*

$$\mathcal{E}_\pi(\hat{f}_n^{(J_\epsilon)}) = \mathbb{E}_\pi\left[d_\pi(\hat{f}_n^{(J_\epsilon)}, f^*)\right] \leq (1+A)\epsilon + \exp\left(-na(1-\exp(-h^2/2))2^{-dJ_\epsilon}\right),$$

where J_ϵ is the smallest integer satisfying $\sum_{j=J_\epsilon+1}^{+\infty} 2^{-dj}\lfloor w(j)\rfloor < \epsilon/A$. Parameters a, A appear in Assumption (A1) and h is used in (SMA).

Remark 9. *The upper bound can be split in the bias term: ϵ and the variance term: $A\epsilon + \exp\left(-na(1-\exp(-h^2/2))2^{-dJ_\epsilon}\right)$. Remark that a bias term appears in the variance term.*

3.3 Optimality

This section is devoted to the optimality, in a minimax sense, of estimation in classification models such that $f^* \in \mathcal{F}_w^{(d)}$. Let $0 < h < 1$, $0 < a \leq 1 \leq A < +\infty$ and w a mapping from $\mathbb{N}$ to $\mathbb{R}^+$. we denote by $\mathcal{P}_{w,h,a,A}$ the set of all probability measures $\pi = (P^X, \eta)$ on $[0,1]^d \times \{-1,1\}$ such that

1. The marginal P^X satisfies (A1).
2. The Assumption (SMA) is satisfied.
3. The Bayes rule f^*, associated to π, belongs to $\mathcal{F}_w^{(d)}$.

We use the version of Lemma of Assouad in the appendix of Lecué [2006c] to lower bound the minimax risk on $\mathcal{P}_{w,h,a,A}$. From Theorem 3 and Theorem 4, we can deduce the optimality (up to a logarithm term) of the estimator $\hat{f}_n^{(J_n)}$ where the rank J_n is obtained by an optimal trade-off between the bias term and the variance term.

Theorem 4. *Let w be a function from $\mathbb{N}$ to $\mathbb{R}^+$ such that*

(i) $\lfloor w(0) \rfloor \geq 1$ *and* $\forall j \geq 1, \quad \lfloor w(j) \rfloor \geq 2^d - 1$

(ii) $\forall j \geq 1, \quad \lfloor w(j-1) \rfloor \geq 2^{-d} \lfloor w(j) \rfloor$.

We have for all $n \in \mathbb{N}$,

$$\inf_{\hat{f}_n} \sup_{\pi \in \mathcal{P}_{w,h,a,A}} \mathcal{E}_\pi(\hat{f}_n) \geq C_0 n^{-1} \left(\lfloor w \left(\lfloor \log n / (d \log 2) \rfloor + 1 \right) \rfloor - (2^d - 1) \right),$$

and if $\lfloor w(j) \rfloor \geq 2^d, \quad \forall j \geq 1$ *then* $\inf_{\hat{f}_n} \sup_{\pi \in \mathcal{P}_{w,h,a,A}} \mathcal{E}_\pi(\hat{f}_n) \geq C_0 n^{-1}$ *where* $C_0 = (h/8) \exp\left(-(1 - \sqrt{1 - h^2})\right)$.

Remark 10. *For a function w satisfying assumptions of Theorem 4 and under (SMA), we can not expect a convergence rate faster than $1/n$, which is the usual lower bound for the classification problem under (SMA).*

From the previous Theorem we obtain immediately Theorem 7.1 of Devroye et al. [1996]. We denote by $\mathcal{P}_1$ the class of all probability measures on $[0,1]^d \times \{-1, 1\}$ such that the marginal distribution P^X is λ_d (the Lebesgue probability distribution on $[0,1]^d$) and (SMA) is satisfied with the margin $h = 1$. The case "$h = 1$" is equivalent to $R^* = 0$. That is for a perfect classification problem, where Y is an exact function of X given by $Y = f^*(X) = \eta(X)$.

Corollary 1. *For any integer n we have*

$$\inf_{\hat{f}_n} \sup_{\pi \in \mathcal{P}_1} \mathcal{E}(\hat{f}_n) \geq \frac{1}{8e}.$$

It means that no classifier can achieve a rate of convergence in the classification models $\mathcal{P}_1$, even if these classification problems are all very good (Y is given by $f^*(X)$ without any noise and there are no spot of low probability).

3.4 Rates of Convergence for Different Classes of Prediction Rules

In this section we apply results stated in Theorem 3 and Theorem 4 to different L^1-ball classes $\mathcal{F}_w^{(d)}$ introduced at the end of Section 2. We give rates of convergence and lower bounds for these models. Using notations introduced in Section 2 and subsection 3.3, we consider the following models. For $w = w_K^{(d)}$ denote by $\mathcal{P}_K^{(d)}$ the set $\mathcal{P}_{w_K^{(d)},h,a,A}$ of probability measures on $[0,1]^d \times \{-1,1\}$ and $\mathcal{P}_\alpha^{(d)}$ for $w = w_\alpha^{(d)}$.

Theorem 5. *For the truncated class $\mathcal{F}_K^{(d)}$, we have*

$$\sup_{\pi \in \mathcal{P}_K^{(d)}} \mathcal{E}_\pi(\hat{f}_n^{(J_n)}) \leq C_{K,h,a,A} \frac{\log n}{n},$$

where $C_{K,h,a,A} > 0$ is depending only on K, h, a, A and for the lower bound, there exists $C_{0,K,h,a,A} > 0$ depending only on K, h, a, A such that, for all $n \in \mathbb{N}$,

$$\inf_{\hat{f}_n} \sup_{\pi \in \mathcal{P}_K^{(d)}} \mathcal{E}_\pi(\hat{f}_n) \geq C_{0,K,h,a,A} n^{-1}.$$

For the exponential class $\mathcal{F}_\alpha^{(d)}$ where $0 < \alpha < 1$, we have for any integer n

$$\sup_{\pi \in \mathcal{P}_\alpha^{(d)}} \mathcal{E}_\pi(\hat{f}_n^{(J_n)}) \leq C'_{\alpha,h,a,A} \left(\frac{\log n}{n} \right)^{1-\alpha},$$

where $C'_{\alpha,h,a,A} > 0$ and for the lower bound, there exists $C'_{0,\alpha,h,a,A} > 0$ depending only on α, h, a, A such that, for all $n \in \mathbb{N}$,

$$\inf_{\hat{f}_n} \sup_{\pi \in \mathcal{P}_\alpha^{(d)}} \mathcal{E}_\pi(\hat{f}_n) \geq C'_{0,\alpha,h,a,A} n^{-1+\alpha}.$$

In both classes, order of J_n is $\lceil \log\left(an/(2^d \log n)\right)/(d \log 2) \rceil$, up to a multiplying constant.

A remarkable point is that the class $\mathcal{F}_K^{(d)}$ has an infinite VC-dimension (cf. Section 2). Nevertheless, the rate $\log n/n$ is achieved on this model.

4 Discussion

In this section we discuss about representation and estimation of "simple" prediction rules in our framework. In considering the classification problem over the square $[0,1]^2$, a classifier has to be able to approach, for instance, the "simple" Bayes rule $f_{\mathcal{C}}^*$ which is equal to 1 inside $\mathcal{C}$, where $\mathcal{C}$ is a disc of $[0,1]^2$, and -1 outside $\mathcal{C}$. In our framework, two questions need to be considered:

- How is the representation of the simple function $f_{\mathcal{C}}^*$ in our fundamental system, using only coefficients with values in $\{-1, 0, 1\}$ and with the writing convention (W)?

- Is the estimate $\hat{f}_n^{(J_n)}$, where $J_n = \lceil \log\left(an/(2^d \log n)\right) / (d \log 2) \rceil$ is the frequency rank appearing in Theorem 5, a good classifier when the underlying probability measure has $f_{\mathcal{C}}^*$ for Bayes rule?

At a first glance, our point of view is not the right way to estimate $f_{\mathcal{C}}^*$. In this regular case (the border is an infinite differentiable curve), the direct estimation of the border is a better approach. The main reason is that a 2-dimensional estimation problem becomes a 1-dimensional problem. Such reduction of dimension makes estimation easier (in passing, our approach is specifically good in the 1-dimensional case, since the notion of border does not exist in this case). Nevertheless, our approach is applicable for the estimation of such functions (cf. Theorem 6). Actually, direct estimation of the border reduces the dimension but there is a big waste of observations since observations far from the border are not used for this estimation point of view. It may explain why our approach is applicable. Denote by

$$\mathcal{N}(A, \epsilon, ||.||_\infty) = \min\left(N : \exists x_1, \ldots, x_N \in \mathbb{R}^2 : A \subseteq \cup_{j=1}^N B_\infty(x_j, \epsilon)\right)$$

the $\epsilon-$covering number of a subset A of $[0,1]^2$, w.r.t. the infinity norm of $\mathbb{R}^2$. For example, the circle $\mathcal{C} = \{(x,y) \in \mathbb{R}^2 : (x-1/2)^2 + (y-1/2)^2 = (1/4)^2\}$ satisfies $\mathcal{N}(\mathcal{C}, \epsilon, ||.||_\infty) \leq (\pi/4)\epsilon^{-1}$. For any set A of $[0,1]^2$, denote by ∂A the border of A.

Theorem 6. *Let A be a subset of $[0,1]^2$ such that $\mathcal{N}(\partial A, \epsilon, ||.||_\infty) \leq \delta(\epsilon)$, for any $\epsilon > 0$, where δ is a decreasing function from $\mathbb{R}_+^*$ with values in $\mathbb{R}^+$ satisfying $\epsilon^2\delta(\epsilon) \longrightarrow 0$ when ϵ tends to zero. Consider the prediction rule $f_A = 2\mathbb{1}_A - 1$. For any $\epsilon > 0$, denote by ϵ_0 the greatest positive number satisfying $\delta(\epsilon_0)\epsilon_0^2 \leq \epsilon$. There exists a prediction rule constructed in the fundamental system $\mathcal{S}$ at the frequency rank J_{ϵ_0} with coefficients in $\{-1, 1\}$ denoted by*

$$f_{\epsilon_0} = \sum_{\mathbf{k} \in I_2(J_{\epsilon_0})} a_{\mathbf{k}}^{(J_{\epsilon_0})} \phi_{\mathbf{k}}^{(J_{\epsilon_0})},$$

with $J_{\epsilon_0} = \lfloor \log(1/\epsilon_0)/\log 2 \rfloor$ such that

$$||f_{\epsilon_0} - f_A||_{L^1(\lambda_2)} \leq 36\epsilon.$$

For instance, there exists a function f_n, written in the fundamental system $\mathcal{S}$ at the frequency level $J_n = \lfloor \log(4n/(\pi \log n))/\log 2 \rfloor$, which approaches the prediction rule $f_{\mathcal{C}}$ with a $L^1(\lambda_2)$ error upper bounded by $36(\log n)/n$. This frequency level is, up to a multiplying constant, the same one appearing in Theorem 5. In a more general way, any prediction rule with a border having a finite perimeter (for instance polygons) is approached by a function written in the fundamental system at the same frequency rank J_n and the same order of $L^1(\lambda_2)$ error $(\log n)/n$. Remark that for this frequency level J_n, we have to estimate $n/\log n$ coefficients. Estimations of one of these coefficients $a_{\mathbf{k}}^{(J_n)}$, where $\mathbf{k} \in I_2(J_n)$, depends on the number of observation in the square $\mathcal{I}_{\mathbf{k}}^{(J_n)}$ associated this coefficient. The probability that no observation "falls" in $\mathcal{I}_{\mathbf{k}}^{(J_n)}$ is smaller than n^{-1}. Thus, number of coefficients estimated with no observations is small compare to the order of approach $(\log n)/n$ and is taken into account in the variance term. Now, the problem is about finding a L^1-ball of prediction rules such that for any integer n the approximation function f_n belongs to such a ball. This problem depends on the geometry of the border set ∂A. It arises naturally since we chose a particular geometry for our partition: dyadic partitions of the space $[0,1]^d$, and we have to pay a price for this choice which has been made independently of the type of functions to estimate. But this choice of geometry in our case is the same as the one met in density approximation using approximation theory while choosing a particular wavelet basis. Depending on the type of Bayes rules we have to estimate, a special partition can be considered. For example our "dyadic approach" is very well adapted for the estimation of Bayes rules associated to chessboard (with the value 1 for black square and -1 for white square). This kind of Bayes rules are very bad estimated by classification procedure estimating the border since most of these procedure have regularity assumptions which are not fulfilled in the case of chessboard.

We can extend our approach in several different ways. Consider the dyadic partition of $[0,1]^d$ with frequency J_n. Instead of choosing 1 or -1 for each square of this partition (like in our approach), we can do a least square regression in each cell of the partition. Inside a square $Sq = \mathcal{I}_{\mathbf{k}}^{(J_n)}$, where $\mathbf{k} \in I_2(J_n)$, we can compute the line minimizing

$$\sum_{i=1}^{n} \mathbb{1}_{(2f(X_i)-1 \neq Y_i, X_i \in Sq)},$$

where f is taken in the set of all indicators of half spaces of $[0,1]^d$ intersecting Sq. Of course, depending on the number of observations inside the cell Sq we can consider bigger classes of functions than the one made of the indicators of half spaces. Our classifier is close to the histogram estimator in density or regression framework, which has been extend to smoother procedure. The other way to extend our approach deals with the problem of the underlying choice of geometry by taking $\mathcal{S}$ for fundamental system. One possible solution is to consider classifiers "adaptive to the geometry". Using an adaptive procedure, for instance aggregation procedure (cf. Lecué [2005]), we can construct classifiers adaptive to the "rotation" and "translation". Consider the dyadic partition of $[0,1]^2$ at the frequency level J_n. We can construct classifiers using the same procedure as (4) but for partitions obtained by translation of the dyadic partition by $(n_1/(2^{J_n}\log n), n_2/(2^{J_n}\log n))$, where $n_1, n_2 = 0, \ldots, \lceil \log n \rceil$. We can do the same thing by aggregating classifiers obtained by the procedure (4) for partitions obtained by rotation of center $(1/2, 1/2)$ with angle $n_3\pi/(2\log n)$, where $n_3 = 0, \ldots, \lceil \log n \rceil$, of the initial dyadic partition. In this heuristic we don't discuss about the way to solve problems near the border of $[0,1]^2$.

5 Proofs

Proof of Theorem 1: Since $\{\eta \geq 1/2\}$ is almost everywhere open there exists an open subset $\mathcal{O}$ of $[0,1]^d$ such that $\lambda_d(\{\eta \geq 1/2\}\Delta\mathcal{O}) = 0$. If $\mathcal{O}$ is the empty set then take $g = -1$, otherwise, for all $x \in \mathcal{O}$ denote by $\mathcal{I}_x$ the biggest subset $\mathcal{I}_{\mathbf{k}}^{(j)}$ for $j \in \mathbb{N}$ and $\mathbf{k} \in I_d(j)$ such that $x \in \mathcal{I}_{\mathbf{k}}^{(j)}$ and $\mathcal{I}_{\mathbf{k}}^{(j)} \subseteq \mathcal{O}$. Remark that $\mathcal{I}_x$ exists because $\mathcal{O}$ is open. We can see that for any $y \in \mathcal{I}_x$ we have $\mathcal{I}_y = \mathcal{I}_x$, thus, $(\mathcal{I}_x : x \in \mathcal{O})$ is a partition of $\mathcal{O}$. We denote by $I_{\mathcal{O}}$ a subset of index $(j,\mathbf{k})$, where $j \in \mathbb{N}, \mathbf{k} \in I_d(j)$ such that $\{\mathcal{O}_x : x \in \mathcal{O}\} = \{\mathcal{I}_{\mathbf{k}}^{(j)} : (j,\mathbf{k}) \in I_{\mathcal{O}}\}$. For any $(j,\mathbf{k}) \in I_{\mathcal{O}}$ we take $a_{\mathbf{k}}^{(j)} = 1$.

Take $\mathcal{O}_1$ an open subset λ_d-almost everywhere equal to $\{\eta < 1/2\}$. If $\mathcal{O}_1$ is the empty set then take $g = 1$. Otherwise, consider the set of index $I_{\mathcal{O}_1}$ built in the same way as previously, and for any $(j,\mathbf{k}) \in I_{\mathcal{O}_1}$ we take $a_{\mathbf{k}}^{(j)} = -1$.

For all $(j,\mathbf{k}) \notin I_{\mathcal{O}} \cup I_{\mathcal{O}_1}$, we take $a_{\mathbf{k}}^{(j)} = 0$. Consider

$$g = \sum_{j=0}^{+\infty} \sum_{\mathbf{k}\in I_d(j)} a_{\mathbf{k}}^{(j)} \phi_{\mathbf{k}}^{(j)}.$$

It is easy to check that the function g belongs to $\mathcal{F}^{(d)}$ and satisfies the writing convention (W) and that, for λ_d-almost $x \in [0,1]^d$, $g(x) = f_\eta(x)$.

Proof of Proposition 1: Assume that $\mathcal{F}_w^{(d)} \neq \{\mathbb{1}_{[0,1]^d}\}$. Take $f \in \mathcal{F}_w^{(d)} - \{\mathbb{1}_{[0,1]^d}\}$. Consider the writing of f in the system $\mathcal{S}$ using the convention (W),

$$f = \sum_{j\in\mathbb{N}} \sum_{\mathbf{k}\in I_d(j)} a_{\mathbf{k}}^{(j)} \phi_{\mathbf{k}}^{(j)},$$

where $a_{\mathbf{k}}^{(j)} \in \{-1,0,1\}$ for any $j \in \mathbb{N}, \mathbf{k} \in I_d(j)$. Consider $b_{\mathbf{k}}^{(j)} = |a_{\mathbf{k}}^{(j)}|$ for any $j \in \mathbb{N}, \mathbf{k} \in I_d(j)$. Take $f_2 = \sum_{j\in\mathbb{N}} \sum_{\mathbf{k}\in I_d(j)} b_{\mathbf{k}}^{(j)} \phi_{\mathbf{k}}^{(j)}$. Remark that the function $f_2 \in \mathcal{F}^{(d)}$ does not satisfy the writing convention (W). We have $f_2 = \mathbb{1}_{[0,1]^d}$. For any $j \in \mathbb{N}$ we have

$$\text{card}\left\{\mathbf{k} \in I_d(j) : b_{\mathbf{k}}^{(j)} \neq 0\right\} = \text{card}\left\{\mathbf{k} \in I_d(j) : a_{\mathbf{k}}^{(j)} \neq 0\right\}. \tag{5}$$

Moreover, one coefficient $b_{\mathbf{k}}^{(j)} \neq 0$ contributes to fill a cell of Lebesgue measure 2^{-dj} among the hypercube $[0,1]^d$. Since the mass total of $[0,1]^d$ is 1, we have

$$1 = \sum_{j\in\mathbb{N}} \sum_{\mathbf{k}\in I_d(j)} 2^{-dj} \text{card}\left\{\mathbf{k} \in I_d(j) : b_{\mathbf{k}}^{(j)} \neq 0\right\}. \tag{6}$$

Moreover, $f \in \mathcal{F}^{(d)}$ thus, for any $j \in \mathbb{N}$,

$$\lfloor w(j) \rfloor \geq \text{card}\left\{\mathbf{k} \in I_d(j) : a_{\mathbf{k}}^{(j)} \neq 0\right\}.$$

We obtain the second assertion of Proposition 1 by using the last inequality and the both assertions (5) and (6).

Assume that $\sum_{j=1}^{+\infty} 2^{-dj} \lfloor w(j) \rfloor \geq 1$. For any integer $j \neq 0$, denote by $\mathcal{I}(j)$ the set of indexes $\{(j,\mathbf{k}) : \mathbf{k} \in I_d(j)\}$.

We use the natural order of $\mathbb{N}^{d+1}$ to order sets of indexes. Take $\mathcal{I}_w(1)$ the family of the first $\lfloor w(1) \rfloor$ elements of $\mathcal{I}(1)$. Denote by $\mathcal{I}_w(2)$ the family made of the first $\lfloor w(1) \rfloor$ elements of $\mathcal{I}(1)$ and add, at the end of this family in the correct order, the first $\lfloor w(2) \rfloor$ elements $(2,\mathbf{k})$ of $\mathcal{I}(2)$ such that $\phi_{\mathbf{k}'}^{(1)} \phi_{\mathbf{k}}^{(2)} = 0$ for any $(1,\mathbf{k}') \in \mathcal{I}_w(1)$,..., for the step j, construct the family $\mathcal{I}_w(j)$ made of all the elements of $\mathcal{I}_w(j-1)$ in the same order and add at the end of this family the indexes $(j,\mathbf{k})$ in $\mathcal{I}(j)$ among the first $\lfloor w(j) \rfloor$ elements of $\mathcal{I}(j)$ such that $\phi_{\mathbf{k}'}^{(J)} \phi_{\mathbf{k}}^{(j)} = 0$ for any $(J,\mathbf{k}') \in \mathcal{I}_w(j-1)$. If there is no more index satisfying this condition then we stop the construction otherwise we go on. Denote by $\mathcal{I}$ the final family obtained by this construction ($\mathcal{I}$ may be finite or infinite). Then, we enumerate the indexes of $\mathcal{I}$ by $(j_1, \mathbf{k}_1) \prec (j_2, \mathbf{k}_2) \prec \cdots$. For the first $(j_1, \mathbf{k}_1) \in \mathcal{I}$ take $a_{\mathbf{k}_1}^{(j_1)} = 1$, for the second element $(j_2, \mathbf{k}_2) \in \mathcal{I}$ take $a_{\mathbf{k}_2}^{(j_2)} = -1$,etc. . Consider the function

$$f = \sum_{j \in \mathbb{N}} \sum_{\mathbf{k} \in I_d(j)} a_{\mathbf{k}}^{(j)} \phi_{\mathbf{k}}^{(j)}.$$

If the construction stops at a given iteration N then f takes its values in $\{-1, 1\}$ and the writing convention (W) is fulfilled since every cells $\mathcal{I}_{\mathbf{k}}^{(j)}$ such that $a_{\mathbf{k}}^{(j)} \neq 0$ has a neighboring cell associated to a coefficient non equals to 0 with an opposite value. Otherwise, for any integer $j \neq 0$, the number of coefficient $a_{\mathbf{k}}^{(j)}$, for $\mathbf{k} \in I_d(j)$, non equals to 0 is $\lfloor w(j) \rfloor$ and the total mass of cells $\mathcal{I}_{\mathbf{k}}^{(j)}$ such that $a_{\mathbf{k}}^{(j)} \neq 0$ is $\sum_{j \in \mathbb{N}} \sum_{\mathbf{k} \in I_d(j)} 2^{-dj} \mathrm{card} \left\{ \mathbf{k} \in I_d(j) : a_{\mathbf{k}}^{(j)} \neq 0 \right\}$ which is greater or equal to 1 by assumption. Thus, all the hypercube is filled by cells associated to coefficients non equal to 0. So f takes its values in $\{-1, 1\}$ and the writing convention (W) is fulfilled since every cells $\mathcal{I}_{\mathbf{k}}^{(j)}$ such that $a_{\mathbf{k}}^{(j)} \neq 0$ has a neighboring cell associated to a coefficient non equals to 0 with an opposite value. Moreover $f \neq \mathbb{1}_{[0,1]^d}$.

Proof of Theorem 2. Let $\pi = (P^X, \eta)$ be a probability measure on $\mathcal{X} \times \{-1, 1\}$ belonging to $\mathcal{P}_{w,A}$. Denote by f^* a Bayes classifier associated to π (for example $f^* = \mathrm{sign}(2\eta - 1)$) . We have

$$d_\pi(f, f^*) = (1/2)\mathbb{E}\left[|2\eta(X) - 1||f(X) - f^*(X)|\right] \leq (A/2)||f - f^*||_{L^1(\lambda_d)}.$$

Let $\epsilon > 0$. Define by J_ϵ the smallest integer satisfying

$$\sum_{j=J_\epsilon+1}^{+\infty} 2^{-dj} \lfloor w(j) \rfloor < \frac{\epsilon}{A}.$$

We write f^* in the fundamental system $(\phi_{\mathbf{k}}^{(j)}, j \in \mathbb{N}, \mathbf{k} \in I_d(j))$ using the convention of writing of section 3.1 but we start at the level of frequency J_ϵ:

$$f^* = \sum_{\mathbf{k} \in I_d(J_\epsilon)} A_{\mathbf{k}}^{(J_\epsilon)} \phi_{\mathbf{k}}^{(J_\epsilon)} + \sum_{j=J_\epsilon+1}^{+\infty} \sum_{\mathbf{k} \in I_d(j)} a_{\mathbf{k}}^{(j)} \phi_{\mathbf{k}}^{(j)}.$$

We consider

$$f_\epsilon = \sum_{\mathbf{k} \in I_d(J_\epsilon)} B_{\mathbf{k}}^{(J_\epsilon)} \phi_{\mathbf{k}}^{(J_\epsilon)}, \tag{7}$$

where

$$B_{\mathbf{k}}^{(J_\epsilon)} = \begin{cases} 1 & \text{if } p_{\mathbf{k}}^{(J_\epsilon)} > 1/2 \\ -1 & \text{otherwise} \end{cases} \tag{8}$$

and

$$p_{\mathbf{k}}^{(J_\epsilon)} = \mathbb{P}(Y = 1 | X \in I_{\mathbf{k}}^{(J_\epsilon)}) = \int_{I_{\mathbf{k}}^{(J_\epsilon)}} \eta(x) \frac{dP^X(x)}{P^X(I_{\mathbf{k}}^{(J_\epsilon)})}, \tag{9}$$

for all $\mathbf{k} \in I_d(J_\epsilon)$. Note that, if $A_{\mathbf{k}}^{(J_\epsilon)} \neq 0$ then $A_{\mathbf{k}}^{(J_\epsilon)} = B_{\mathbf{k}}^{(J_\epsilon)}$, moreover f^* take its values in $\{-1, 1\}$, thus ,we have

$$\begin{aligned} ||f_\epsilon - f^*||_{L^1(\lambda_d)} &= \sum_{\substack{\mathbf{k} \in I_d(J_\epsilon) \\ A_{\mathbf{k}}^{(J_\epsilon)} \neq 0}} \int_{I_{\mathbf{k}}^{(J_\epsilon)}} |f^*(x) - f_\epsilon(x)| dx + \sum_{\substack{\mathbf{k} \in I_d(J_\epsilon) \\ A_{\mathbf{k}}^{(J_\epsilon)} = 0}} \int_{I_{\mathbf{k}}^{(J_\epsilon)}} |f^*(x) - f_\epsilon(x)| dx \\ &\leq 2^{-dJ_\epsilon + 1} \text{card} \left\{ \mathbf{k} \in I_d(J_\epsilon) : A_{\mathbf{k}}^{(J_\epsilon)} = 0 \right\} \leq 2 \sum_{j=J_\epsilon+1}^{+\infty} 2^{-dj} \lfloor w(j) \rfloor < 2\epsilon/A. \end{aligned}$$

Proof of Theorem 3. Let $\pi = (P^X, \eta)$ be a probability measure on $\mathcal{X} \times \{-1, 1\}$ satisfying (A1), (SMA) and such that $f^* = \text{sign}(2\eta - 1)$, a Bayes classifier associated to π, belongs to $\mathcal{F}_w^{(d)}$ (a L^1-ball of Bayes rules).

Let $\epsilon > 0$ and J_ϵ the smallest integer satisfying $\sum_{j=J_\epsilon+1}^{+\infty} 2^{-dj} \lfloor w(j) \rfloor < \epsilon/A$. We decompose the risk in the bias term and variance term:

$$\mathcal{E}(\hat{f}_n^{(J_\epsilon)}) = \mathbb{E}\left[d_\pi(\hat{f}_n^{(J_\epsilon)}, f^*) \right] \leq \mathbb{E}\left[d_\pi(\hat{f}_n^{(J_\epsilon)}, f_\epsilon) \right] + d_\pi(f_\epsilon, f^*),$$

where $\hat{f}_n^{(J_\epsilon)}$ is introduced in (4) and f_ϵ in (7).

Using the definition of J_ϵ and according to the approximation Theorem (Theorem 1), the bias term satisfies:

$$d_\pi(f_\epsilon, f^*) \leq \epsilon.$$

For the variance term we have (using the notations introduced in (4) and (8)):

$$\begin{aligned} \mathbb{E}\left[d_\pi(\hat{f}_n^{(J_\epsilon)}, f_\epsilon) \right] &= \frac{1}{2} \left| \mathbb{E}\left[Y(f_\epsilon(X) - \hat{f}_n^{(J_\epsilon)}(X)) \right] \right| \leq \frac{1}{2} \mathbb{E}\left[\int_{[0,1]^d} |f_\epsilon(x) - \hat{f}_n^{(J_\epsilon)}(x)| dP^X(x) \right] \\ &= \frac{1}{2} \sum_{\mathbf{k} \in I_d(J_\epsilon)} \mathbb{E}\left[\int_{I_{\mathbf{k}}^{(J_\epsilon)}} |B_{\mathbf{k}}^{(J_\epsilon)} - \hat{A}_{\mathbf{k}}^{(J_\epsilon)}| dP^X \right] \\ &\leq \frac{A}{2^{dJ_\epsilon+1}} \sum_{\mathbf{k} \in I_d(J_\epsilon)} \mathbb{E}[|B_{\mathbf{k}}^{(J_\epsilon)} - \hat{A}_{\mathbf{k}}^{(J_\epsilon)}|] \leq \frac{A}{2^{dJ_\epsilon}} \sum_{\mathbf{k} \in I_d(J_\epsilon)} \mathbb{P}\left(|B_{\mathbf{k}}^{(J_\epsilon)} - \hat{A}_{\mathbf{k}}^{(J_\epsilon)}| = 2 \right). \end{aligned}$$

Let $\mathbf{k} \in I_d(J_\epsilon)$. For any $m \in \{0, \dots, n\}$, we introduce the sets

$$\Omega_{\mathbf{k}}^{(m)} = \left\{ \mathrm{Card}\{i \in \{1, \dots, n\} : X_i \in I_{\mathbf{k}}^{(J_\epsilon)}\} = m \right\}$$

and

$$\Omega_{\mathbf{k}} = \left\{ \mathrm{card} \left\{ i \in \{1, \dots, n\} : \begin{array}{l} X_i \in I_{\mathbf{k}}^{(J_\epsilon)}, \\ Y_i = 1 \end{array} \right\} \leq \mathrm{card} \left\{ i \in \{1, \dots, n\} : \begin{array}{l} X_i \in I_{\mathbf{k}}^{(J_\epsilon)}, \\ Y_i = -1 \end{array} \right\} \right\}.$$

We have

$$\mathbb{P}(\hat{A}_{\mathbf{k}}^{(J_\epsilon)} = -1) = \mathbb{P}(\Omega_{\mathbf{k}}^{(0)c} \cap \Omega_{\mathbf{k}}) + \mathbb{P}(\Omega_{\mathbf{k}}^{(0)})$$

and

$$\mathbb{P}(\Omega_{\mathbf{k}}^{(0)c} \cap \Omega_{\mathbf{k}}) = \sum_{m=1}^{n} \mathbb{P}(\Omega_{\mathbf{k}}^{(m)} \cap \Omega_{\mathbf{k}}) = \sum_{m=1}^{n} \mathbb{P}(\Omega_{\mathbf{k}} | \Omega_{\mathbf{k}}^{(m)}) \mathbb{P}(\Omega_{\mathbf{k}}^{(m)}).$$

Moreover, denote by $Z_1, \dots, Z_n$ some variables i.i.d. with a Bernoulli with parameter $p_{\mathbf{k}}^{(J_\epsilon)}$ for common probability distribution ($p_{\mathbf{k}}^{(J_\epsilon)}$ is introduced in (9) and is equal to $\mathbb{P}(Y = 1 | X \in I_{\mathbf{k}}^{(J_\epsilon)})$), we have for any $m = 1, \dots, n$,

$$\mathbb{P}(\Omega_{\mathbf{k}} | \Omega_{\mathbf{k}}^{(m)}) = \mathbb{P}\left(\frac{1}{m} \sum_{i=1}^{m} Z_i \leq \frac{1}{2} \right).$$

Concentration inequality of Hoeffding leads to

$$\mathbb{P}\left(\frac{1}{m} \sum_{i=1}^{m} Z_i \geq p_{\mathbf{k}}^{(J_\epsilon)} + t \right) \leq \exp(-2mt^2) \text{ and } \mathbb{P}\left(\frac{1}{m} \sum_{i=1}^{m} Z_i \leq p_{\mathbf{k}}^{(J_\epsilon)} - t \right) \leq \exp(-2mt^2), \tag{10}$$

for all $t > 0$ and $m = 1, \dots, n$.

Denote by $a_{\mathbf{k}}^{(J_\epsilon)}$ the probability $\mathbb{P}\left(X \in I_{\mathbf{k}}^{(J_\epsilon)}\right)$. If $p_{\mathbf{k}}^{(J_\epsilon)} > 1/2$, applying second inequality of (10) leads to

$$\begin{aligned}
& \mathbb{P}\left(|B_{\mathbf{k}}^{(J_\epsilon)} - \hat{A}_{\mathbf{k}}^{(J_\epsilon)}| = 2 \right) = \mathbb{P}(\hat{A}_{\mathbf{k}}^{(J_\epsilon)} = -1) \\
\leq \;& \sum_{m=1}^{n} \mathbb{P}\left[\frac{1}{m} \sum_{j=1}^{m} Z_j \leq p_{\mathbf{k}}^{(J_\epsilon)} - (p_{\mathbf{k}}^{(J_\epsilon)} - 1/2) \right] \begin{pmatrix} n \\ m \end{pmatrix} (a_{\mathbf{k}}^{(J_\epsilon)})^m (1 - a_{\mathbf{k}}^{(J_\epsilon)})^{n-m} \\
+ \;& \mathbb{P}(\Omega_{\mathbf{k}}^{(0)}) \\
\leq \;& \sum_{m=0}^{n} \exp\left(-2m(p_{\mathbf{k}}^{(J_\epsilon)} - 1/2)^2 \right) \begin{pmatrix} n \\ m \end{pmatrix} (a_{\mathbf{k}}^{(J_\epsilon)})^m (1 - a_{\mathbf{k}}^{(J_\epsilon)})^{n-m} \\
= \;& \left(1 - a_{\mathbf{k}}^{(J_\epsilon)} (1 - \exp(-2(p_{\mathbf{k}}^{(J_\epsilon)} - 1/2)^2)) \right)^n \\
\leq \;& \exp\left(-na(1 - \exp(-2(p_{\mathbf{k}}^{(J_\epsilon)} - 1/2)^2)) 2^{-dJ_\epsilon} \right).
\end{aligned}$$

If $p_{\mathbf{k}}^{(J_\epsilon)} < 1/2$ then similar arguments used in the previous case and first inequality of (10) lead to

$$\begin{aligned}\mathbb{P}\left(|B_{\mathbf{k}}^{(J_\epsilon)} - \hat{A}_{\mathbf{k}}^{(J_\epsilon)}| = 2\right) &= \mathbb{P}(\hat{A}_{\mathbf{k}}^{(J_\epsilon)} = 1) \\ &\le \exp\left(-na(1-\exp(-2(p_{\mathbf{k}}^{(J_\epsilon)} - 1/2)^2))2^{-dJ_\epsilon}\right).\end{aligned}$$

If $p_{\mathbf{k}}^{(J_\epsilon)} = 1/2$, we use $\mathbb{P}\left(|B_{\mathbf{k}}^{(J_\epsilon)} - \hat{A}_{\mathbf{k}}^{(J_\epsilon)}| = 2\right) \le 1$. Like in the proof of Theorem 2, we use the writing

$$f^* = \sum_{\mathbf{k}\in I_d(J_\epsilon)} A_{\mathbf{k}}^{(J_\epsilon)} \phi_{\mathbf{k}}^{(J_\epsilon)} + \sum_{j=J_\epsilon+1}^{+\infty} \sum_{\mathbf{k}\in I_d(j)} a_{\mathbf{k}}^{(j)} \phi_{\mathbf{k}}^{(j)}.$$

Since $P^X(\eta = 1/2) = 0$, if $A_{\mathbf{k}}^{(J_\epsilon)} \neq 0$ then $p_{\mathbf{k}}^{(J_\epsilon)} \neq 1/2$. Thus, the variance term satisfies:

$$\begin{aligned}&\mathbb{E}\left[d_\pi(\hat{f}_n, f_\epsilon^*)\right] \\ &\le \frac{A}{2^{dJ_\epsilon}} \left(\sum_{\substack{\mathbf{k}\in I_d(J_\epsilon) \\ A_{\mathbf{k}}^{(J_\epsilon)} \neq 0}} \mathbb{P}\left(|B_{\mathbf{k}}^{(J_\epsilon)} - \hat{A}_{\mathbf{k}}^{(J_\epsilon)}| = 2\right) + \sum_{\substack{\mathbf{k}\in I_d(J_\epsilon) \\ A_{\mathbf{k}}^{(J_\epsilon)} = 0}} \mathbb{P}\left(|B_{\mathbf{k}}^{(J_\epsilon)} - \hat{A}_{\mathbf{k}}^{(J_\epsilon)}| = 2\right) \right) \\ &\le \frac{A}{2^{dJ_\epsilon}} \sum_{\substack{\mathbf{k}\in I_d(J_\epsilon) \\ A_{\mathbf{k}}^{(J_\epsilon)} \neq 0}} \exp\left(-na(1-\exp(-2(p_{\mathbf{k}}^{(J_\epsilon)} - 1/2)^2))2^{-dJ_\epsilon}\right) + A\epsilon.\end{aligned}$$

If $A_{\mathbf{k}}^{(J_\epsilon)} \neq 0$ then $\eta > 1/2$ or $\eta < 1/2$ over the whole set $I_{\mathbf{k}}^{(J_\epsilon)}$, so

$$\left|\frac{1}{2} - p_{\mathbf{k}}^{(J_\epsilon)}\right| = \int_{I_{\mathbf{k}}^{(J_\epsilon)}} \left|\eta(x) - \frac{1}{2}\right| \frac{dP^X(x)}{P^X(I_{\mathbf{k}}^{(J_\epsilon)})}.$$

Moreover π satisfies $\mathbb{P}\left(|2\eta(X) - 1| \ge h\right) = 1$, so

$$\left|\frac{1}{2} - p_{\mathbf{k}}^{(J_\epsilon)}\right| \ge \frac{h}{2}.$$

We have shown that for all $\epsilon > 0$,

$$\mathcal{E}(\hat{f}_n) = \mathbb{E}\left[d_\pi(\hat{f}_n, f^*)\right] \le (1+A)\epsilon + \exp\left(-na(1-\exp(-2(h/2)^2))2^{-dJ_\epsilon}\right),$$

where J_ϵ is the smallest integer satisfying $\sum_{j=J_\epsilon+1}^{+\infty} 2^{-dj} \lfloor w(j) \rfloor < \epsilon/A$.

Proof of Theorem 4. For all $q \in \mathbb{N}$ we consider G_q a net of $[0,1]^d$ defined by:

$$G_q = \left\{ \left(\frac{2k_1+1}{2^{q+1}}, \dots, \frac{2k_d+1}{2^{q+1}} \right) : (k_1, \dots, k_d) \in \{0, \dots, 2^q - 1\}^d \right\}$$

and the function η_q from $[0,1]^d$ to G_q such that $\eta_q(x)$ is the closest point of G_q from x (in the case of ex aequo, we choose the smallest point for the usual order on $\mathbb{R}^d$). Associated to this grid, the partition ${\mathcal{X}'}_1^{(q)}, \dots, {\mathcal{X}'}_{2^{dq}}^{(q)}$ of $[0,1]^d$ is defined by $x, y \in {\mathcal{X}'}_i^{(q)}$ iff $\eta_q(x) = \eta_q(y)$ and we use a special indexation for this partition: denote by ${x'}_{k_1,\dots,k_d}^{(q)} = \left(\frac{2k_1+1}{2^{q+1}}, \dots, \frac{2k_d+1}{2^{q+1}}\right)$ and we say that ${x'}_{k_1,\dots,k_d}^{(q)} \prec {x'}_{{k'}_1,\dots,{k'}_d}^{(q)}$ if

$$\eta_{q-1}({x'}_{k_1,\dots,k_d}^{(q)}) \prec \eta_{q-1}({x'}_{{k'}_1,\dots,{k'}_d}^{(q)})$$

or

$$\eta_{q-1}({x'}_{k_1,\dots,k_d}^{(q)}) = \eta_{q-1}({x'}_{{k'}_1,\dots,{k'}_d}^{(q)}) \text{ and } (k_1, \dots, k_d) < ({k'}_1, \dots, {k'}_d),$$

for the usual order on $\mathbb{N}^d$. Thus, the partition $({\mathcal{X}'}_j^{(q)} : j = 1, \dots, 2^{dq})$ has an increasing indexation according to the order of $({x'}_{k_1,\dots,k_d}^{(q)})$ for the order defined above. This order take care of the previous partition by splitting blocks in the right given order and inside a block of a partition we take the natural order of $\mathbb{N}^d$. We introduce an other parameter $m \in \{1, \dots, 2^{qd}\}$ and we define for all $i = 1, \dots, m$, $\mathcal{X}_i^{(q)} = {\mathcal{X}'}_i^{(q)}$ and $\mathcal{X}_0^{(q)} = [0,1]^d - \cup_{i=1}^m \mathcal{X}_i^{(q)}$. Parameters q and m will be chosen later. We consider $W \in [0, m^{-1}]$, chosen later, and define the function f_X from $[0,1]^d$ to $\mathbb{R}$ by $f_X = W/\lambda_d(\mathcal{X}_1)$ (where λ_d is the Lebesgue measure on $[0,1]^d$) on $\mathcal{X}_1, \dots, \mathcal{X}_m$ and $(1 - mW)/\lambda_d(\mathcal{X}_0)$ on $\mathcal{X}_0$. We denote by P^X the probability distribution on $[0,1]^d$ with the density f_X w.r.t. the Lebesgue measure. For all $\sigma = (\sigma_1, \dots, \sigma_m) \in \Omega = \{-1, 1\}^m$ we consider η_σ defined for any $x \in [0,1]^d$ by

$$\eta_\sigma(x) = \begin{cases} \frac{1+\sigma_j h}{2} & \text{if } x \in \mathcal{X}_j, j = 1, \dots, m, \\ 1 & \text{if } x \in \mathcal{X}_0. \end{cases}$$

We have a set of probability measures $\{\pi_\sigma : \sigma \in \Omega\}$ on $[0,1]^d \times \{-1, 1\}$ indexed by the hypercube Ω where P^X is the marginal on $[0,1]^d$ of π_σ and η_σ its conditional probability function of $Y = 1$ given X. We denote by f_σ^* the Bayes rule associated to π_σ, we have $f_\sigma^*(x) = \sigma_j$ if $x \in \mathcal{X}_j$ for $j = 1, \dots, m$ and 1 if $x \in \mathcal{X}_0$, for any $\sigma \in \Omega$.

Now we give conditions on q, m and W such that for all σ in Ω, π_σ belongs to $\mathcal{P}_{w,h,a,A}$. If we take

$$W = 2^{-dq}, \tag{11}$$

then $P^X << \lambda$ and $\forall x \in [0,1]^d, a \leq dP^X/d\lambda(x) \leq A$. We have clearly $|2\eta(x)-1| \geq h$ for any $x \in [0,1]^d$. We can see that $f^*_\sigma \in \mathcal{F}^{(d)}_w$ for all $\sigma \in \{-1,1\}^m$ iff

$$\begin{array}{l}
\lfloor w(q+1) \rfloor \geq \inf(x \in 2^d\mathbb{N} : x \geq m) \\
\lfloor w(q) \rfloor \geq \left\{ \begin{array}{lll} 2^d - 1 & \text{if} & m < 2^d \\ \inf(x \in 2^d\mathbb{N} : x \geq 2^{-d}m) & \text{otherwise} & \end{array} \right. \\
\dots \\
\lfloor w(1) \rfloor \geq \left\{ \begin{array}{lll} 2^d - 1 & \text{if} & m < 2^{dq} \\ \inf(x \in 2^d\mathbb{N} : x \geq 2^{-dq}m) & \text{otherwise} & \end{array} \right. \\
\lfloor w(0) \rfloor \geq 1
\end{array}.$$

Since we have $\lfloor w(j) \rfloor \geq 2^d - 1$ for all $j \geq 1$ and $\lfloor w(0) \rfloor = 1$, and $\lfloor w(j-1) \rfloor \geq \lfloor w(j) \rfloor / 2^d$, then $f^*_\sigma \in \mathcal{F}^{(d)}_w$ for all $\sigma \in \Omega$ iff

$$\lfloor w(q+1) \rfloor \geq \inf(x \in 2^d\mathbb{N} : x \geq m). \tag{12}$$

Take q, m and W such that (11) and (12) are fulfilled then, $\{\pi_\sigma : \sigma \in \Omega\}$ is a subset of $\mathcal{P}_{w,h,a,A}$. Let $\sigma \in \Omega$ and $\hat{f}_n$ be a classifier, we have

$$\begin{aligned}
\mathbb{E}_{\pi_\sigma}\left[R(\hat{f}_n) - R^*\right] &= (1/2)\mathbb{E}_{\pi_\sigma}\left[|2\eta_\sigma(X) - 1||\hat{f}_n(X) - f^*_\sigma(X)|\right] \\
&\geq (h/2)\mathbb{E}_{\pi_\sigma}\left[|\hat{f}_n(X) - f^*_\sigma(X)|\right] \\
&\geq (h/2)\mathbb{E}_{\pi_\sigma}\left[\sum_{i=1}^m \int_{\mathcal{X}_i} |\hat{f}_n(x) - f^*_\sigma(x)| dP^X(x) + \int_{\mathcal{X}_0} |\hat{f}_n(x) - f^*_\sigma(x)| dP^X(x)\right] \\
&\geq (Wh/2)\sum_{i=1}^m \mathbb{E}_{\pi_\sigma}\left[\int_{\mathcal{X}_i} |\hat{f}_n(x) - \sigma_i| \frac{dx}{\lambda(\mathcal{X}_1)}\right] \\
&\geq (Wh/2)\mathbb{E}_{\pi_\sigma}\left[\sum_{i=1}^m \left|\sigma_i - \int_{\mathcal{X}_i} \hat{f}_n(x) \frac{dx}{\lambda(\mathcal{X}_1)}\right|\right].
\end{aligned}$$

We deduce that

$$\inf_{\hat{f}_n} \sup_{\pi \in \mathcal{P}_{w,h,a,A}} \mathcal{E}_\pi(\hat{f}_n) \geq (Wh/2) \inf_{\hat{\sigma}_n \in [-1,1]^m} \sup_{\sigma \in \{-1,1\}^m} \mathbb{E}_{\pi_\sigma}\left[\sum_{i=1}^m |\sigma_i - \hat{\sigma}_i|\right].$$

Now, we control the Hellinger distance between two neighbouring probability measures. Let ρ be the Hamming distance on Ω. Let σ, σ' in Ω such that $\rho(\sigma,\sigma')=1$. We have

$$H^2(\pi_\sigma^{\otimes n}, \pi_{\sigma'}^{\otimes n}) = 2\left(1-\left(1-\frac{H^2(\pi_\sigma,\pi_{\sigma'})}{2}\right)^n\right),$$

and a straightforward calculus leads to $H^2(\pi_\sigma,\pi_{\sigma'}) = 2W\left(1-\sqrt{1-h^2}\right)$. Take

$$W = 1/n, \tag{13}$$

thus, for any integer n, we have $H^2(\pi_\sigma^{\otimes n}, \pi_{\sigma'}^{\otimes n}) \leq \beta < 2$ where $\beta = 2\left(1-\exp(1-\sqrt{1-h^2})\right)$. The Assouad's Lemma (cf. Lecué [2006c]) yields $\inf_{\hat{\sigma}_n\in[-1,1]^m}\sup_{\sigma\in\{-1,1\}^m}\mathbb{E}_{\pi_\sigma}\left[\sum_{i=1}^m|\sigma_i-\hat{\sigma}_i|\right] \geq \frac{m}{4}\left(1-\frac{\beta}{2}\right)^2$. We conclude that

$$\inf_{\hat{f}_n}\sup_{\pi\in\mathcal{P}_{w,h,a,A}}\mathcal{E}_\pi(\hat{f}_n) \geq Wh\frac{m}{8}\left(1-\frac{\beta}{2}\right)^2.$$

According to (11), (12) and (13) we take $W=2^{-dq}=1/n, q=\lfloor \log n/(d\log 2)\rfloor, m = \lfloor w\left(\lfloor \log n/(d\log 2)\rfloor+1\right)\rfloor-(2^d-1)$. For these values we have

$$\inf_{\hat{f}_n}\sup_{\pi\in\mathcal{P}_{w,h,a,A}}\mathcal{E}_\pi(\hat{f}_n) \geq C_0 n^{-1}\left(\lfloor w\left(\lfloor \log n/(d\log 2)\rfloor+1\right)\rfloor-(2^d-1)\right).$$

where $C_0 = (h/8)\exp\left(-(1-\sqrt{1-h^2})\right)$.

Proof of Corollary 1: It suffices to apply Theorem 4 to the function w defined by $w(j)=2^{dj}$ for any integer j and $a=A=1$ for $P^X=\lambda_d$.

Proof of Theorem 5:

1. If we assume that $J_\epsilon \geq K$ then $\sum_{j=J_\epsilon+1}^{+\infty}2^{-dj}\lfloor w_K^{(d)}(j)\rfloor = (2^{dK})/(2^{dJ_\epsilon}(2^d-1))$. We take

$$J_\epsilon = \left\lceil \frac{\log\left((A2^{dK})/(\epsilon(2^d-1))\right)}{d\log 2}\right\rceil$$

and ϵ_n the unique solution of $(1+A)\epsilon_n = \exp(-nC\epsilon_n)$, where $C = a(1-e^{-h^2/2})(2^d-1)[A2^{d(K+1)}]^{-1}$. Thus, $\epsilon_n \leq (\log n)/(Cn)$. For $J_n = J_{\epsilon_n}$, we have

$$\mathcal{E}\left(\hat{f}_n^{(J_n)}\right) \leq C_{K,d,h,a,A}\frac{\log n}{n},$$

for any integer n such that $\log n \geq 2^{d(K+1)}(2^d-1)^{-1}$ and $J_n \geq K$, where $C_{K,d,h,a,A} = 2(1+A)/C$.

If we have $\lfloor \log n/(d\log 2)\rfloor \geq 2$ then $\lfloor w\left(\lfloor \log n/(d\log 2)\rfloor + 1\right)\rfloor - (2^d-1) \geq 2^d$, so we obtain the lower bound with the constant $C_{0,K} = 2^d C_0$ and if $\lfloor \log n/(d\log 2)\rfloor \geq K$ the constant can be $C_{0,K} = C_0(2^{dK} - (2^d-1))$.

2. If we have $J_\epsilon \geq N^{(d)}(\alpha)$, then $\sum_{j=J_\epsilon+1}^{+\infty} 2^{-dj}\lfloor w_\alpha^{(d)}(j)\rfloor \leq (2^{d(1-\alpha)J_\epsilon}(2^{d(1-\alpha)}-1))^{-1}$. We take

$$J_\epsilon = \left\lceil \frac{\log(A/(\epsilon(2^{d(1-\alpha)}-1)))}{d(1-\alpha)\log 2} \right\rceil .$$

Denote by ϵ_n the unique solution of $(1+A)\epsilon_n = \exp(-nC\epsilon_n^{1/(1-\alpha)})$ where $C = a(1-e^{-h^2/2})2^{-d}(A^{-1}(2^{d(1-\alpha)}-1))^{1/(1-\alpha)}$. We have $\epsilon_n \leq (\log n/(nC))^{1-\alpha}$. For $J_n = J_{\epsilon_n}$, we have

$$\mathcal{E}\left(\hat{f}_n^{(J_n)}\right) \leq \frac{2(1+A)A}{2^{d(1-\alpha)}-1}\left[\frac{2^d}{a(1-e^{-h^2/2})}\right]^{1-\alpha}\left(\frac{\log n}{n}\right)^{1-\alpha}.$$

For the lower bound we have for any integer n,

$$\inf_{\hat{f}_n}\sup_{\pi\in\mathcal{P}_\alpha^{(d)}} \mathcal{E}_\pi(\hat{f}_n) \geq C_0 \max\left(1, n^{-1}\left(2^d n^\alpha - (2^d-1)\right)\right).$$

Proof of Theorem 6: Let $\epsilon > 0$. Denote by ϵ_0 the greatest positive number satisfying $\delta(\epsilon_0)\epsilon_0^2 \leq \epsilon$. Consider $N(\epsilon_0) = \mathcal{N}(\partial A, \epsilon_0, ||.||_\infty)$ and $x_1, \ldots, x_{N(\epsilon_0)} \in \mathbb{R}^2$ such that $\partial A \subset \cup_{j=1}^{N(\epsilon_0)} B_\infty(x_j, \epsilon_0)$. Since $2^{-J_{\epsilon_0}} \geq \epsilon_0$, only nine dyadic sets of frequency J_{ϵ_0} can be used to cover a ball of radius ϵ_0 for the infinity norm of $\mathbb{R}^2$. Thus, we only need $9N(\epsilon_0)$ dyadic sets of frequency J_{ϵ_0} to cover ∂A. Consider the partition of $[0,1]^2$ by dyadic sets of frequency J_{ϵ_0}. Except on the $9N(\epsilon_0)$ dyadic sets used to cover the border ∂A, the prediction rule f_A is constant, equal to 1 or -1, on the other dyadic sets. Thus, by taking $f_{\epsilon_0} = \sum_{k_1,k_2=0}^{2^{J_{\epsilon_0}}-1} a_{k_1,k_2}^{(J_{\epsilon_0})}\phi_{k_1,k_2}^{(J_{\epsilon_0})}$, where $a_{k_1,k_2}^{(J_{\epsilon_0})}$ is equal to one value of f_A in the dyadic set $\mathcal{I}_{k_1,k_2}^{(J_{\epsilon_0})}$, we have

$$||f_{\epsilon_0} - f_A||_{L^1(\lambda_2)} \leq 9N(\epsilon_0)2^{-2J_{\epsilon_0}} \leq 36\delta(\epsilon_0)\epsilon_0^2 \leq 36\epsilon.$$

References

G. Blanchard, G. Lugosi, and N. Vayatis. On the rate of convergence of regularized boosting classifiers. *JMLR*, 4:861–894, 2003.

L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition.* Springer, New York, Berlin, Heidelberg, 1996.

G. Lecué. Optimal rates of aggregation in classification. Available at http://hal.ccsd.cnrs.fr/ccsd-00021233/en/, 2006c.

G. Lecué. Simultaneous adaptation to the marge and to complexity in classification. Available at http://hal.ccsd.cnrs.fr/ccsd-00009241/en/, 2005.

G. Lugosi and N. Vayatis. On the bayes-risk consistency of regularized boosting methods. *Ann. Statist.*, 32(1):30–55, 2004.

P. Massart and E. Nédélec. Risk bound for statistical learning. Preprint. available at http://www.math.u-psud.fr/∼massart/page5.html, 2003.

R. Nowak and C. Scott. Minimax-optimal classification with dyadic decision trees. *IEEE Transaction on Information Theory*, 2004.

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