Electronic Journal of Statistics
Vol. 2 (2008) 741–773
ISSN: 1935-7524
DOI: 10.1214/07-EJS015

Classification with minimax fast rates for classes of Bayes rules with sparse representation

Guillaume Lecué

Laboratoire de Probabilités et Modèles Aléatoires (UMR CNRS 7599),
Université Paris VI
4, pl.Jussieu, Boîte courrier 188,
75252 Paris, France
e-mail: lecue@ccr.jussieu.fr

Abstract: We consider the classification problem on the cube $[0,1]^d$ when the Bayes rule is known to belong to some new functions classes. These classes are made of prediction rules satisfying some conditions regarding their coefficients when developed over the (overcomplete) basis of indicator functions of dyadic cubes of $[0,1]^d$. The main concern of the paper is on the thorough analysis of the approximation term, which is in general bypassed in the classification literature. An adaptive classifier is designed to achieve the minimax rate of convergence (up to a logarithmic factor) over these functions classes. Lower bounds on the convergence rate over these classes are established when the underlying marginal of the design is comparable to the Lebesgue measure. Connections with some existing models for classification (RKHS and "boundary fragements") are established.

AMS 2000 subject classifications: Primary 62G05; secondary 62C20.
Keywords and phrases: Classification; Sparsity; Decision dyadic trees; Minimax rates; Aggregation.

1. Introduction

We denote by $D_n = (X_i, Y_i)_{1\leq i\leq n}$ a sample of n i.i.d. observations of a couple (X, Y) of random variables with values in $[0,1]^d \times \{-1, 1\}$. We denote by π the probability distribution of the couple (X, Y). We want to construct some measurable functions which associate a label $y \in \{-1, 1\}$ to each point x of $[0,1]^d$. 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 $R(\cdot)$ over all measurable functions with values in $\{-1, 1\}$. This function is called the *Bayes rule* and is defined by

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

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 only the observations D_n, with 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 the *excess risk of* $\hat{f}_n$. In this case, $R(\hat{f}_n) = \mathbb{P}(\hat{f}_n(X) \neq Y|D_n)$ and the symbol $\mathbb{E}_\pi$ denotes the expectation w.r.t. D_n, when the probability distribution of (X_i, Y_i) is π for any $i = 1, \ldots, n$. Consider $(\phi(n))_{n\in\mathbb{N}}$ a decreasing sequence of positive numbers. We say that a classifier $\hat{f}_n$ *learns at the convergence rate* $\phi(n)$, if there exists an absolute constant $C > 0$ 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 function is a *semi-distance* (it is symmetric, satisfies the triangle inequality and $d_\pi(f, f) = 0$). The excess risk of any classifier $\hat{f}_n$ can be written as

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

where the RHS is the risk of $\hat{f}_n$ associated with the loss d_π.

Theorem 7.2 of Devroye et al. [1996] shows that no classifier can learn with a given convergence rate for arbitrary underlying probability distribution π. To achieve a rate of convergence, we need to assume that the Bayes rule belongs to some functions classes with a small complexity. For instance, Yang [1999a,b] provide examples of classifiers learning, with a given convergence rate, under complexity assumptions on the set of conditional probability functions. Other rates of convergence have been obtained under the assumption that the Bayes rule belongs to a class of prediction rules with a finite dimension of Vapnik and Chervonenkis (cf. Devroye et al. [1996]). In both cases, the problem of a direct approximation of f^* is not treated. In the first case, the problem of approximation of f^* is shifted to the problem of approximation of the regression function η. In fact, if $\bar{f}$ denotes the plug-in rule $\mathbb{1}_{\bar{\eta}\geq 1/2}$, where $\bar{\eta}$ is a function with values in $[0, 1]$ then the loss of $\bar{f}$ satisfies

$$d_\pi(\bar{f}, f^*) \leq 2\mathbb{E}[|\bar{\eta}(X) - \eta(X)|] \tag{1}$$

Thus, under smoothness assumption on the conditional function η, we can control the approximation term. 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 $\{x \in [0,1]^d : \eta(x) = 1/2\}$ has no effect on the estimation of f^*. In the second case, the approximation term equals to zero, since the Bayes rule is assumed to belong to a class with a finite VC dimension and so we don't need to approach the Bayes rule by a simpler object.

The main difficulty of a direct approximation of f^* relies on the fact that the loss function d_π depends on the unknown probability measure π. Given a model $\mathcal{P}$ (a set of probability measures on $[0,1]^d \times \{-1,1\}$) with a known complexity, we want to be able to construct a decreasing family $(\mathcal{F}_\epsilon)_{\epsilon>0}$ of classes of prediction rules, such that the following approximation result holds:

$$\forall \pi = (P^X, \eta) \in \mathcal{P}, \forall \epsilon > 0, \exists f_\epsilon \in \mathcal{F}_\epsilon : d_\pi(f_\epsilon, f^*) \leq \epsilon, \tag{2}$$

where P^X is the marginal distribution of π on $[0,1]^d$ and $f^* = \mathrm{Sign}(2\eta - 1)$ is the Bayes rule associated with the regression function η of π. In fact, we want the classes $\mathcal{F}_\epsilon$ to be parametric in such a way that, for the estimation problem, we just have to estimate a parametric object in a class $\mathcal{F}_{\epsilon_n}$, for a well chosen ϵ_n (generally obtained by a trade-off between the bias/approximation term and the variance term, coming from the estimation of the best parametric object in $\mathcal{F}_{\epsilon_n}$ approaching f^*).

We upper bound the loss d_π, but we still work with a direct approximation of f^*. For a prediction rule f we have

$$d_\pi(f, f^*) = \mathbb{E}[|2\eta(X) - 1| \mathbb{1}_{f(X) \neq f^*(X)}] \leq (1/2)||f - f^*||_{L^1(P^X)}. \tag{3}$$

In order to get a distribution-free loss function, we assume that the following assumption holds. This assumption is close to assuming that the marginal distribution of X is the Lebesgue measure on $[0,1]^d$.

Assumption 1 (A1). *The marginal probability distribution P^X of X is absolutely continuous w.r.t. the Lebesgue measure λ_d and there exist two constants $0 < a < A < +\infty$ such that one version of the density function $dP^X(\cdot)/d\lambda_d$ satisfies $a \leq dP^X(x)/d\lambda_d \leq A, \quad \forall x \in [0,1]^d$.*

The behavior of the regression function η near the level $1/2$ is a key characteristic of the classification's quality (cf. e.g. Tsybakov [2004]). In fact, the closer is η to $1/2$, the more difficult is the classification problem. Here, we work under the following assumption introduced by Massart and Nédélec [2006].

Assumption 2 (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, for any prediction rule f,

$$\frac{ah}{2}||f - f^*||_{L_1(\lambda_d)} \leq d_\pi(f, f^*) \leq \frac{A}{2}||f - f^*||_{L_1(\lambda_d)}.$$

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

The paper is organized as follows. In the next section, we present in a semi-informal way, the gist of the approach and an overview of the results. Then, in Section 3, we introduce a class of functions, with values in $\{-1, 1\}$, developed in a fundamental system of $L^2([0,1]^d)$. Section 4 is devoted to the approximation and the estimation of Bayes rules having a sparse representation in this system. In Section 5, we discuss this approach. Proofs are postponed to Section 6.

2. Related works and overview of the results

Many authors pointed out the need for developing a suitable approximation theory for classification. Given a model $\mathcal{C}$ of prediction rules, it is written in p.34 of Boucheron et al. [2005]: "estimating the model bias $\min_{f\in\mathcal{C}}(R(f) - R^*)$ seems to be beyond the reach of our understanding. In fact, estimating R^* is known to be a difficult statistical problem, see Devroye et al. [1996] and Antos et al. [1999]." 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 form 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 Tsybakov [2004] and Tsybakov and van de Geer [2005], the author examine classes that are indexed by a complexity exponent that reflects the smoothness of the Bayes decision boundary. An argument of entropy is then used to upper bound the bias term. A generalization of these classes is given in Scott and Nowak [2006]. In Steinwart et al. [2006], the authors present necessary and sufficient conditions to know wether a function class approximates the Bayes risk for some general loss functions.

In the present work, we develop the Bayes rule in the overcomplete basis $\mathcal{S}$ made of all the indicator functions of dyadic cubes of $[0,1]^d$ assuming that the coefficients of this development take their values in $\{-1, 0, 1\}$. Then, we estimate these coefficients by doing a simple majority vote in each dyadic cell of a partition of $[0,1]^d$. The resulting estimator can be seen as a decision tree algorithm, since we can make a link between the analytical development of f^* in $\mathcal{S}$ and the dyadic decision trees representation of the Bayes rule. We obtain minimax rates of convergence (up to a logarithm factor) for these classes of Bayes rules.

The best known decision tree algorithms are CART (cf. Breiman et al. [1984]) and C4.5 (cf. Quinlan [1993]). These methods use a growing and pruning algorithm. First, a large tree is grown by splitting recursively nodes along coordinates axes according to an "impurity" criterion. Next, this tree is pruned using a penalty function. Penalties are usually based on standard complexity regularization like the square root of the size of the tree. Spatially adaptive penalties depend not only on the complexity of the tree, but also on the spatial distribution of training samples. More recent constructions of decision trees have been proposed in Scott and Nowak [2006] and Blanchard et al. [2007]. In Scott

and Nowak [2006], the authors consider, in the multi-class framework, dyadic decision trees and exhibit near-minimax rates of convergence by considering spatial adaptive penalties. They obtained rates of convergence over classes of prediction functions having a complexity defined in the same spirit as Mammen and Tsybakov [1999] and Tsybakov [2004]. In Blanchard et al. [2007], a general framework is worked out including classification for different loss functions. The authors select among a set of dyadic trees having a finite depth, the best tree realizing an optimal trade-off between the empirical risk and a penalty term. Here, the penalty term is proportional to the number of leaves in the tree. They obtained oracle inequalities and derived rates of convergence in the regression setup under a regularity assumption on the underlying regression function to estimate. Rates of convergence, for the classification problem, are not derived from these oracle inequalities, since they do not treat the bias term.

Our estimation procedure does not provide any tree algorithm in the same spirit as these previous works. The main reason is that, we obtain results under the assumption on the marginal distribution given by (A1). This assumption allows us to work at a given "frequency" and we do not need a multi-scale construction of the dyadic tree as in the previous related work. Once the optimal frequency is obtained (by trade off), the estimation procedure is a regular histogram rule as considered in Chapter 9 of Devroye et al. [1996].

The present work focuses on the control of the approximation term and the introduction of classes of prediction rules having different complexities and approximation qualities (the complexity of a class is given by the way the number of coefficients, in the development of the Bayes rule in the basis $\mathcal{S}$, non equal to zero, is controlled in function of the "depth"(or "frequency") of these coefficients). As we shall see, one crucial difference of our estimator is that it is able to deal with infinite trees. Such infinite trees can be considered since we control the bias term. Nevertheless, when the complexity is unknown, we use a multi-scale approach to construct an adaptive procedure.

3. Classes of Bayes rules with sparse representation

In this section, we introduce a class of prediction rules. For that, we consider two different representations of a prediction rule.

The first way is to represent a prediction rule as an infinite dyadic tree. An *infinite dyadic decision tree* is defined as a partitioning of the hypercube $[0,1]^d$ obtained by cutting in half perpendicular to one of the axis coordinates, then cutting recursively the two pieces obtained in half again, and so on. Most of the time, finite dyadic trees are considered (cf. Blanchard et al. [2007] and Scott and Nowak [2006]). It means that the previous constructions stop at an arbitrary point along every branch. For a survey on decision trees we refer to Murthy [1998]. Here, we consider also infinite dyadic trees.

The other way is more "analytic". First, we consider the representation of prediction rules in a fundamental system of $L^2([0,1]^d, \lambda_d)$ (that is a countable family of functions such that the set made of all their finite linear combinations

is dense in $L^2([0,1]^d, \lambda_d)$) inherited from the Haar basis. Then, we control the number of non-zero coefficients (which are restricted here to take only the values $-1, 0$ or 1).

3.1. Analytic representation of decision trees

First, we construct a fundamental system of $L^2([0,1]^d, \lambda_d)$. We consider a sequence of partitions of $[0,1]^d$ by setting for any integer j,

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

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 $k \in \{1, \ldots, 2^j - 1\}$,

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

Some examples of sets $\mathcal{I}_{\mathbf{k}}^{(j)}$ in the 2-dimensional case are given in Figure 1.

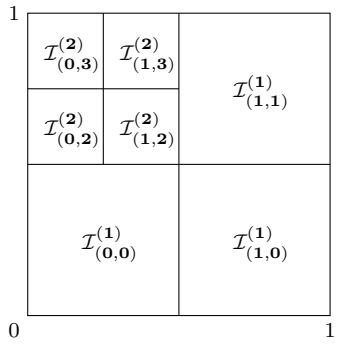

FIG 1. *Example of a finite dyadic partition of* $[0,1]^2$.

We use these dyadic partitions to define a fundamental system of $L^2([0,1]^d, \lambda_d)$.

Definition 1. *The* **class of indicators functions of the dyadic sets** *of* $[0,1]^d$ *is*

$$\mathcal{S} = \left(\phi_{\mathbf{k}}^{(j)} : j \in \mathbb{N}, \mathbf{k} \in I_d(j)\right) \text{ where } \phi_{\mathbf{k}}^{(j)} = \mathbb{1}_{\mathcal{I}_{\mathbf{k}}^{(j)}}, \quad \forall j \in \mathbb{N}, \mathbf{k} \in I_d(j)$$

and $\mathbb{1}_A$ *denotes the indicator of a set* A.

We use this fundamental system of $L^2([0,1]^d, \lambda_d)$ to construct some statistical models of prediction rules.

Definition 2. *We define the class of functions* $\boldsymbol{\mathcal{F}^{(d)}}$ *to be the closure set in* $L^2([0,1]^d, \lambda_d)$ *of all the finite linear combinations of elements in* $\mathcal{S}$ *having coefficients only in* $\{-1,0,1\}$ *intersected with the set of all the prediction rules.*

Namely, the functions class $\mathcal{F}^{(d)}$ is the set of all the prediction rules $f : [0,1]^d \longmapsto \{-1,1\}$ such that there exists a sequence $(a_{\mathbf{k}}^{(j)})_{j,\mathbf{k}}$ of numbers in $\{-1,0,1\}$ satisfying

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

Since $\mathcal{S}$ is not an orthogonal basis of $L^2([0,1]^d, \lambda_d)$, the expansion of a function f of $\mathcal{F}^{(d)}$ in $\mathcal{S}$ is not unique. Therefore, to avoid any ambiguity, we define a convention on the choice of the coefficients of f when developed in the system $\mathcal{S}$.

Definition 3. *Let* f *be a prediction rule in* $\mathcal{F}^{(d)}$ *and* $(a_{\mathbf{k}}^{(j)})_{j,\mathbf{k}}$ *be a sequence of numbers in* $\{-1,0,1\}$ *such that* $f = \sum_{j=0}^{+\infty} \sum_{\mathbf{k} \in I_d(j)} a_{\mathbf{k}}^{(j)} \phi_{\mathbf{k}}^{(j)}$. *We say that the function* f *or the sequence* $(a_{\mathbf{k}}^{(j)})_{j,\mathbf{k}}$ *satisfies the* **writing convention (W)** *when*

- *for any* $j \in \mathbb{N}$ *and* $\mathbf{k} \in I_d(j)$, *if* $a_{\mathbf{k}}^{(j)} \neq 0$ *then for any* $j' > j$ *and* $\mathbf{k}' \in I_d(j')$ *such that* $\phi_{\mathbf{k}}^{(j)} \phi_{\mathbf{k}'}^{(j')} \neq 0$ *we have* $a_{\mathbf{k}'}^{(j')} = 0$
- *for any* $j \in \mathbb{N}$ *and* $\mathbf{k} \in I_d(j)$, *the set of coefficients* $\{a_{\mathbf{k}'}^{(j+1)}$ *s.t.* $\phi_{\mathbf{k}}^{(j)} \phi_{\mathbf{k}'}^{(j+1)} \neq 0\}$ *is not equal to* $\{1\}$ *or* $\{-1\}$.

In what follows, we use the vocabulary appearing in the wavelet literature. The index "j" of the coefficient $a_{\mathbf{k}}^{(j)}$ and the function $\phi_{\mathbf{k}}^{(j)}$ is called "level of frequency". We can describe a mapping $f \in \mathcal{F}^{(d)}$ satisfying the writing convention (W) by using an infinite dyadic decision tree with some restriction on the nodes.

Each node is associated with a coefficient $a_{\mathbf{k}}^{(j)}$. The root is $a_{(0,\ldots,0)}^{(0)}$. If a node, describing the coefficient $a_{\mathbf{k}}^{(j)}$, is equal to 1 or -1 then it has no branches (this is the first point of Definition 3), otherwise it has 2^d branches, corresponding to the 2^d coefficients $\{a_{\mathbf{k}'}^{(j+1)}$ s.t. $\phi_{\mathbf{k}}^{(j)} \phi_{\mathbf{k}'}^{(j+1)} \neq 0\}$ of the next frequency. The second point of Definition 3 imposes that a node cannot have all his leaves equal to 1 together (or -1). At the end, all the leaves of the tree equal to 1 or -1 (because f takes its values only in $\{-1,1\}$) and the depth of a leaf is the frequency of the associated coefficient.

Remark that, this writing convention is not an assumption. The following proposition proves that all functions in $\mathcal{F}^{(d)}$ can be written using this convention.

Proposition 1. *Let* f *be a function in* $\mathcal{F}^{(d)}$. *There exists a sequence of coefficients* $(a_{\mathbf{k}}^{(j)})_{j,\mathbf{k}}$ *with values in* $\{-1,0,1\}$ *satisfying the writing convention (W)*

such that

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

An example of the development of a prediction rule in the basis $\mathcal{S}$ and its associated decision tree representation is given in Figure 2. This figure illustrate the overcomplete basis $\mathcal{S}$ and the writing convention on this simple example.

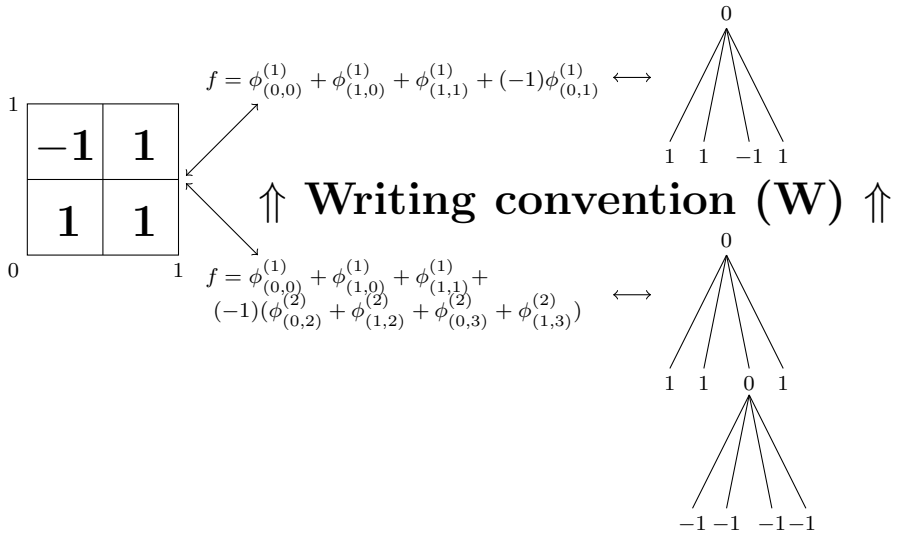

FIG 2. *Example of the writing convention (W) on the analytical and dyadic tree representations in the* 2*-dimensional case.*

We can avoid the problem of the non-uniqueness of the expansion of a function in the overcomplete system $\mathcal{S}$. For instance, by replacing $\mathcal{S}$ by the wavelet tensor product of the Haar basis (cf. Meyer [1990]), we obtain an orthonormal wavelet basis of $L^2([0,1]^d)$. In that case, the link with dyadic decision trees is much more complicated and the obtained results are not easily interpretable.

Remark 1. *An interesting fact is that, we can consider the set $\mathcal{S}$, as a dictionary of basic functions. Considering prediction rules as linear combinations of the functions in this dictionary with coefficients in $\{-1,0,1\}$ (using the convention of writing (W)), we obtain that, the LASSO estimator (cf. Tibshirani [1996]) is given, in this framework, by*

$$\operatorname{Arg} \max_{f \in \mathcal{F}^{(d)}} \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{f(X_i) \neq Y_i} + \gamma \sum_{j,\mathbf{k}} |a_{\mathbf{k}}^{(j)}|,$$

where $f = \sum_{j=0}^{+\infty} \sum_{\mathbf{k} \in I_d(j)} a_{\mathbf{k}}^{(j)} \phi_{\mathbf{k}}^{(j)}$, $\lambda_d - a.s.$ Since the coefficients $a_{\mathbf{k}}^{(j)}$ take their values in $\{-1,0,1\}$, the l_1-type penalty $\sum_{j,\mathbf{k}} |a_{\mathbf{k}}^{(j)}|$ is exactly the number of leaves

of the dyadic tree associated with the prediction rule f. Thus, LASSO estimator, in this framework and for the dictionary $\mathcal{S}$, is the same as the estimator considered in Blanchard et al. [2007].

It is easy to see that all measurable functions from $[0,1]^d$ to $\{-1,1\}$ cannot be represented in the overcomplete system $\mathcal{S}$. For example, in the one dimensional case, the set of open subsets $\{\mathring{\mathcal{I}}_{\mathbf{k}}^{(j)} : j \in \mathbb{N}, \mathbf{k} \in I_1(j)\}$ is a basis of open subsets of $[0,1]$. Thus, saying that any measurable function from $[0,1]$ to $\{-1,1\}$ can be written in the system $\mathcal{S}$ with $\{-1,0,1\}$ valued coefficients is equivalent to say that any measurable subset of $[0,1]$ is almost everywhere equal to an open subset of $[0,1]$. This fact is not true. A simple example is given by the following construction. Consider $(q_k)_{k\geq 1}$ an enumeration of the rational numbers of $(0,1)$. Denote by A the union, over $k \in \mathbb{N}$, of the open balls $\mathcal{B}(q_k, 2^{-(k+1)})$. This is a dense open set of Lebesgue measure bounded by $1/2$. Now, let $\mathcal{O}$ be an open subset of $[0,1]$. If $\mathcal{O} \subseteq A^c$ then $\mathcal{O} = \emptyset$ because A is dense (so the interior set of A^c is empty) so $\lambda_1(\mathcal{O}\Delta A^c) = \lambda(A^c) \geq 1/2$. If $\mathcal{O} \nsubseteq A^c$ then $\mathcal{O} \cap A$ is a non-empty open subset so $\lambda_1(\mathcal{O}\Delta A^c) \geq \lambda(\mathcal{O} \cap A) > 0$. In both cases, we have $\lambda_1(\mathcal{O}\Delta A^c) > 0$. Thus, A^c cannot be equal almost everywhere to any open subset of $[0,1]$. In particular, the prediction rule $f = 2\mathbb{1}_A - 1$ cannot be written in the fundamental system $\mathcal{S}$ using coefficients with values only in $\{-1,0,1\}$ ($f \notin \mathcal{F}^{(1)}$). "Fat Cantor" are other examples of measurable subsets which cannot be equal almost everywhere to any open subset. Nevertheless, under a mild assumption (cf. the following definition) a prediction rule belongs to $\mathcal{F}^{(d)}$.

Definition 4. *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 Δ stands for the symbol of symmetric difference.*

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

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

We assume that $\{\eta \geq 1/2\}$ and $\{\eta < 1/2\}$ are almost everywhere open. Then, there exists $g \in \mathcal{F}^{(d)}$ such that $g = f_\eta, \lambda_d - a.s.$.

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 $f_\eta \in \mathcal{F}^{(d)}$. Moreover, 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 $\lambda_d(\mathcal{O} - K) \leq \epsilon$. 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 $\lambda_d(\{x \in [0,1]^d : f(x) \neq g(x)\}) \leq \epsilon$. Thus, $\mathcal{F}^{(d)}$ is dense in $L^2(\lambda_d)$ intersected with the set of all measurable functions from $[0,1]^d$ to $\{-1,1\}$.

3.2. Class of Bayes rules

In this subsection, we construct some models of prediction rules by taking some subsets of $\mathcal{F}^{(d)}$.

Definition 5. *For any function $w : \mathbb{N} \longmapsto \mathbb{N}$, we denote by $\mathcal{F}_w^{(d)}$ the functions class made of all prediction rules $f \in \mathcal{F}^{(d)}$ satisfying the writing convention (W) with*

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

where $a_{\mathbf{k}}^{(j)} \in \{-1, 0, 1\}$ and such that the number of non zero coefficients at each frequency level satisfies

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

The richness of the functions 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 poor. That is the subject of the following proposition.

Proposition 2. *Let w be a mapping from $\mathbb{N}$ to $\mathbb{N}$ such that $w(0) \geq 1$. The two following assertions are equivalent:*

(i) $\mathcal{F}_w^{(d)} \neq \{\mathbb{1}_{[0,1]^d}\}$.
(ii) $\sum_{j=1}^{+\infty} 2^{-dj} w(j) \geq 1$.

This proposition is strongly connected to the Kraft's inequality (see e.g. Cover and Thomas [1991]). Indeed, Kraft's inequality for a binary tree $\mathcal{T}$ says that $\sum_{l \in \mathcal{L}} 2^{-\mathrm{depth}(l)} \leq 1$, where $\mathcal{L}$ is the set of all the leaves of $\mathcal{T}$ (note that a leaf of a tree is a terminal node). This quantity can be written as

$$\sum_{l \in \mathcal{L}} 2^{-\mathrm{depth}(l)} = \sum_{j=1}^{\infty} \sum_{l \in \mathcal{L}_j} 2^{-j} = \sum_{j=1}^{\infty} 2^{-j} w(j), \tag{4}$$

where $\mathcal{L}_j$ is the set of all the leaves of $\mathcal{T}$ of level (or depth) j and $w(j)$ its cardinality. Kraft's inequality is a restriction on the sequence $\mathcal{N}_l = (w(j))_j$ of number of leaves in a tree. This restriction comes from the fact that, with the notations here, if we want to fill the interval $[0, 1]$ with disjoint dyadic intervals there is necessarily a restriction on the sequence $\mathcal{N}_l$ (the cumulative mass of dyadic intervals cannot be greater than 1). In our case, we have to fill all the interval $[0, 1]$ to construct $\{-1, 1\}$-valued functions. So we need the sequence $\mathcal{N}_s$ to be rich enough to fill all this space. That is why the quantity in equation (4) is a break point in both approaches.

If w is too large then, the approximation of the model $\mathcal{F}_w^{(d)}$, by a parametric model will be impossible. That is why we give a particular look on the functions classes introduced in the following definition.

Definition 6. *Let w be a mapping from $\mathbb{N}$ to $\mathbb{N}$. If w satisfies*

$$\sum_{j=0}^{+\infty} \frac{w(j)}{2^{dj}} < +\infty \tag{5}$$

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

We say that $\mathcal{F}_w^{(d)}$ is a "L^1-ellipsoid" for a function w satisfying (5), because, the sequence $(w(j))_{j\in\mathbb{N}}$ belongs to a L^1-ellipsoid of $\mathbb{N}^{\mathbb{N}}$, with sequence of radius $(2^{dj})_{j\in\mathbb{N}}$. Moreover, Definition 6 can be linked to the definition of a L^1-ellipsoid for real valued functions (cf. Meyer [1990]), since we have a kind of basis, given by $\mathcal{S}$, and we have a control on coefficients which increases with the frequency. Control on coefficients, given in equation (5), is close to the one for coefficients of a real valued function in a L^1-ellipsoid of Sobolev (cf. Korostelev and Tsybakov [1993]), since it deals with the quality of approximation of the class $\mathcal{F}_w^{(d)}$ by a parametric model.

Remark 2. *A L^1-ellipsoid of prediction rules is made of "sparse" prediction rules. In fact, for $f \in \mathcal{F}_w^{(d)}$ with w satisfying (5), the proportion of non-zero coefficients in the decomposition of f (using the writing convention (W)), at a given frequency, becomes small (w.r.t. the 2^{dj} coefficients of level j) as the frequency grows. That is the reason why $\mathcal{F}_w^{(d)}$ can be called a* **sparse class of prediction rules**.

Next, we provide examples of functions satisfying (5). Functions Classes $\mathcal{F}_w^{(d)}$ associated with these functions are used in what follows as statistical models. We first define the minimal infinite class of prediction rules $\mathcal{F}_0^{(d)}$ which is the class $\mathcal{F}_w^{(d)}$ when $w = w_0^{(d)}$ 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 concept of local oscillation of a prediction rule. This concept defines a kind of "regularity" for functions with values in $\{-1, 1\}$. Let f be a function from $[0,1]^d$ to $\{-1,1\}$ in $\mathcal{F}^{(d)}$, we consider the writing of f in the fundamental system introduced in Section 4.1 with writing convention (W):

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

Let $J \in \mathbb{N}$ and $\mathbf{k} \in I_d(J)$. We say that $\mathcal{I}_{\mathbf{k}}^{(J)}$ is a **low oscillating block** of f when f has exactly $2^d - 1$ non-zero coefficients, in this block, at each level of frequencies greater or equal to $J + 1$. 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 oscillating 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". Figure 3 provides an example of function in $\mathcal{F}_0^{(2)}$. In this example the top left cells of the cube $[0,1]^2$ is a low oscillating block.

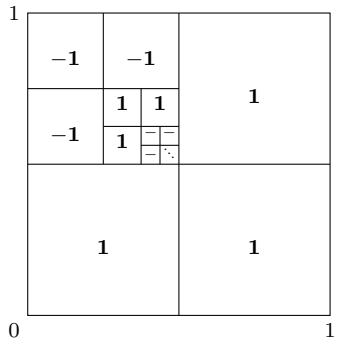

FIG 3. *Example of a prediction rule in* $\mathcal{F}_0^{(2)}$.

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 (cf. Devroye et al. [1996]). For any class $\mathcal{F}$ of prediction rules from $[0,1]^d$ 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 $x_1,\ldots,x_m\in[0,1]^d$ and $m\in\mathbb{N}$,

$$S(\mathcal{F},m) = \max\left(N(\mathcal{F},(x_1,\ldots,x_m)) : x_1,\ldots,x_m\in[0,1]^d\right)$$

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

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

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 $j\in\mathbb{N}$. For any integer m, we have $N(\mathcal{F}_0^{(d)},(x_1,\ldots,x_m)) = 2^m$. Hence, the following proposition holds.

Proposition 3. *The class of prediction rules* $\mathcal{F}_0^{(d)}$ *has an infinite* VC*-dimension.*

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

Let $K\in\mathbb{N}^*$. We denote by $\mathcal{F}_K^{(d)}$ the class $\mathcal{F}_w^{(d)}$ of prediction rules where w is 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 $w = w_\alpha^{(d)}$ for

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

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

The classes $\mathcal{F}_0^{(d)}$, $\mathcal{F}_K^{(d)}$ and $\mathcal{F}_\alpha^{(d)}$ are examples of L^1-ellipsoid of prediction rules.

4. Rates of convergence over $\mathcal{F}_w^{(d)}$ under (SMA)

4.1. Approximation result

Let w be a function from $\mathbb{N}$ to $\mathbb{N}$ 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 with π, belongs to $\mathcal{F}_w^{(d)}$ and the marginal of π on $[0,1]^d$ is absolutely continuous and a 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-ellipsoid 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 with π. For instance, J_ϵ can be the smallest integer J satisfying $\sum_{j=J+1}^{+\infty} 2^{-dj} w(j) < \epsilon/A$.

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

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

Since f_ϵ belongs to a parametric model we expect a good control of the variance term $\mathbb{E}_\pi[d_\pi(\hat{f}_n, f_\epsilon)]$, depending on the dimension of the parametric model which is linked to the quality of the approximation in the bias term. Remark that, no assumption on the quality of the classification problem (like an assumption on the margin) is required to obtain Theorem 2. Only assumption on the "number of oscillations" of f^* is used. Theorem 2 deals with approximation of functions in the L^1-ellipsoid $\mathcal{F}_w^{(d)}$ by functions with values in $\{-1,1\}$ and no estimation issues are considered.

4.2. Estimation result

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

$$\hat{f}_n^{(J)} = \sum_{\mathbf{k} \in I_d(J)} \hat{A}_{\mathbf{k}}^{(J)} \phi_{\mathbf{k}}^{(J)}, \tag{6}$$

where the coefficients are defined by

$$\hat{A}_{\mathbf{k}}^{(J)} = \begin{cases} 1 & \text{if } \exists X_i \in \mathcal{I}_{\mathbf{k}}^{(J)} \text{ and } N_{\mathbf{k}}^{(J)+} > N_{\mathbf{k}}^{(J)-} \\ -1 & \text{otherwise,} \end{cases}$$

where, for any $\mathbf{k} \in I_d^{(J)}$, $N_{\mathbf{k}}^{(J)+} = \text{Card}\{i : X_i \in \mathcal{I}_{\mathbf{k}}^{(J)} \text{ and } Y_i = 1\}$ and $N_{\mathbf{k}}^{(J)-} = \text{Card}\{i : X_i \in \mathcal{I}_{\mathbf{k}}^{(J)} \text{ and } Y_i = -1\}$. The estimator $\hat{f}_n^{(J)}$ realizes a simple majority vote in each cell $\mathcal{I}_{\mathbf{k}}^{(J)}$ for any $\mathbf{k} \in I_d^{(J)}$ (for an example we refer to Figure 6.1 p.96 in Devroye et al. [1996]).

To obtain a good control of the variance term, we need to insure a good quality of the estimation problem. Therefore, estimation results are obtained in Theorem 3 under the assumption (SMA). Nevertheless, the assumption (SMA) is not enough to obtain any rate of convergence (cf. Chapter 7 of Devroye et al. [1996] or corollary 1 at the end of section 4.3). We have to define a model for η or f^* with a finite complexity. Here, we assume that the underlying Bayes rule f^*, associated with π, belongs to a L^1-ellipsoid of prediction rules. For that, we introduce the following models. Let $0 < h < 1$, $0 < a \leq 1 \leq A < +\infty$ and w a mapping from $\mathbb{N}$ to $\mathbb{N}$. 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 with π, belongs to $\mathcal{F}_w^{(d)}$.

Theorem below provides an estimation theorem in the model $\mathcal{P}_{w,h,a,A}$.

Theorem 3 (Estimation theorem). *Let $\mathcal{F}_w^{(d)}$ be a L^1-ellipsoid of prediction rules. Let π be a probability measure on $[0,1]^d \times \{-1,1\}$ in $\mathcal{P}_{w,h,a,A}$. The excess risk of the classifier $\hat{f}_n^{(J_\epsilon)}$ satisfies, for any $\epsilon > 0$,*

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

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

4.3. Optimality

This section is devoted to the optimality, in a minimax sense, of estimation in the classification models $\mathcal{F}_w^{(d)}$. We apply a version of the Assouad Lemma to lower bound the risk over $\mathcal{P}_{w,h,a,A}$.

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

(i) $w(0) \geq 1$ *and* $\forall j \geq 1, \quad w(j) \geq 2^d - 1$
(ii) $\forall j \geq 1, \quad w(j-1) \geq 2^{-d} w(j)$.

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

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

where $C_0 = (h/8) \exp\left(-(1 - \sqrt{1-h^2})\right)$. *Moreover, if* $w(j) \geq 2^d, \forall j \geq 1$ *then*

$$\inf_{\hat{f}_n} \sup_{\pi \in \mathcal{P}_{w,h,a,A}} \mathcal{E}_\pi(\hat{f}_n) \geq C_0 n^{-1}.$$

Remark 3. *For a function w satisfying assumptions of Theorem 4 and under (SMA), we cannot expect a convergence rate faster than $1/n$, which is the usual lower bound for the classification problem under (SMA).*

Theorem 7.1 of Devroye et al. [1996] can be deduced from Theorem 4. 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 $R^* = 0$.

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 any rate of convergence in the classification model $\mathcal{P}_1$.

4.4. 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-ellipsoid classes $\mathcal{F}_w^{(d)}$ introduced at the end of Section 3. Rates of convergence and lower bounds for these models are stated. Using notation introduced in Section 3 and Subsection 4.3, we consider the following models. For $w = w_K^{(d)}$, we denote by $\mathcal{P}_K^{(d)}$ the set of probability measures $\mathcal{P}_{w_K^{(d)},h,a,A}$ and by $\mathcal{P}_\alpha^{(d)}$ for the exponential model with $w = w_\alpha^{(d)}$.

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

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

where $C_{K,h,a,A} > 0$ is depending only on K, h, a, A. 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

$$\sup_{\pi \in \mathcal{P}_\alpha^{(d)}} \mathcal{E}_\pi(\hat{f}_n^{(J_n(\alpha))}) \leq C'_{\alpha,h,a,A} \left(\frac{\log n}{n}\right)^{1-\alpha}, \tag{7}$$

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

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

The levels of frequency $J_n(\alpha)$ *and* $J_n(K)$ *are, up to a multiplying constant, of the order of* $\lceil \log\left(an/(2^d \log n)\right)/(d \log 2) \rceil$*.*

A remarkable point is that the class $\mathcal{F}_K^{(d)}$ has an infinite VC-dimension (cf. Proposition 3). Nevertheless, the rate $\log n/n$ is achieved in this model. Existence of classes of prediction rules with infinite VC dimension that are consistent when the marginal distribution of the design X is without atoms has been remarked in Devroye et al. [1996].

4.5. Adaptation to the complexity

In this section, we provide an adaptive estimator for the exponential classes. The estimator $\hat{f}_n^{(J_n(\alpha))}$, appearing in equation (7), depends on the complexity parameter α, since

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

and $\epsilon_n = (\log n/(nC))^{1-\alpha}$, where $C = a(1-e^{-h^2/2})2^{-d}(A^{-1}(2^{d(1-\alpha)}-1))^{1/(1-\alpha)}$. In practice, we do not have access to this parameter. Thus, it is important to construct an estimator free from this parameter and which can learn at the near-optimal rate $((\log n)/n)^{1-\alpha}$ if the underlying probability distribution belongs to $\mathcal{P}_\alpha^{(d)}$ for any α. This is the problem of adaptation to the complexity parameter α.

To construct an adaptive estimator, we use an aggregation procedure. We split the sample in two parts. Denote by $D_m^{(1)}$ the subsample containing the first m observations and $D_l^{(2)}$ the one containing the $l(= n - m)$ last ones. Subsample $D_m^{(1)}$ is used to construct classifiers $\hat{f}_m^{(J)}$ for different frequency levels $J \in [0, J^{(n)}]$, for an integer $J^{(n)}$ chosen later. Subsample $D_l^{(2)}$ is used to construct the exponential weights of our aggregation procedure (cf. Lecué [2007]). We aggregate the basis classifiers $\hat{f}_m^{(J)}$, $J \in [1, J^{(n)}]$, by the procedure

$$\tilde{f}_n = \sum_{J=1}^{J^{(n)}} w_J^{(l)} \hat{f}_m^{(J)}, \tag{8}$$

where the weights

$$w_J^{(l)} = \frac{\exp\left(\sum_{i=m+1}^n Y_i \hat{f}_m^{(J)}(X_i)\right)}{\sum_{J'=1}^{J^{(n)}} \exp\left(\sum_{i=m+1}^n Y_i \hat{f}_m^{(J')}(X_i)\right)}, \quad \forall J = 1, \ldots, J^{(n)} \tag{9}$$

are called the exponential weights.

The classifier that we propose is

$$\hat{f}_n = \mathrm{Sign}(\tilde{f}_n). \tag{10}$$

Theorem 6. *Assume that $J^{(n)}$ is greater than $(\log n)^2$ and choose $l = \lceil n/\log n \rceil$ for the learning sample size. For any $\alpha \in (0,1)$, we have, for n large enough,*

$$\sup_{\pi \in \mathcal{P}_\alpha^{(d)}} \mathcal{E}_\pi(\hat{f}_n) \leq 6C'_{\alpha,h,a,A} \left(\frac{\log n}{n}\right)^{1-\alpha}, \tag{11}$$

where $C'_{\alpha,h,a,A} > 0$ has been introduced in Theorem 5.

The classifier $\hat{f}_n$ does not require the knowledge of the parameter α neither of a, A, h. Thus, it is also adaptive to the parameters a, A and h.

Remark 4. *We may compare our method with the ERM type aggregate defined by*

$$\bar{f}_n \in \mathrm{Arg} \min_{f \in \{\hat{f}_m^{(0)}, \ldots, \hat{f}_m^{(J^{(n)})}\}} \sum_{i=m+1}^n \mathbb{I}_{(f(X_i) \neq Y_i)}.$$

This aggregate also satisfies (11), if we replace $\hat{f}_n$ by $\bar{f}_n$ (cf. Lecué [2007]). The difference is that the aggregate (8) uses a multi-scale approach (it associates a weight to each frequency), whereas the adaptive classifier $\bar{f}_n$ selects the best "empirical frequency".

The other way to extend our approach deals with the problem of choice of the geometry by taking $\mathcal{S}$ as fundamental system. One possible solution is to consider classifiers "adaptive to the geometry". Using an adaptive procedure, for instance the same as in (8), we can construct classifiers adaptive to the "rotation" and "translation". Consider, for example, the dyadic partition of $[0,1]^2$ at the frequency level J_n. We can construct classifiers using the same procedure as (6) but for partitions obtained by translation of the dyadic partitions by the vector $(n_1/(2^{J_n}\log n), n_2/(2^{J_n}\log n))$, where $n_1, n_2 = 0, \ldots, \lceil \log n \rceil$. We can do the same thing by aggregating classifiers obtained by the procedure (6) 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 boundary of $[0,1]^2$.

5. Related structures and perspectives

5.1. RKHS

Reproducing kernel Hilbert spaces (RKHS) (cf. Aronszajn [1950]) are functions spaces which are usually used as statistical models for the classification problem (and other problems). They are usually associated with the support vectors machines (SVM) estimators. Some of these functions space have a structure close to the L^1-ellipsoids of this paper.

We consider the one-dimensional Haar basis $(\psi_{j,k})_{j,k}$ defined by

$$\psi_{-1,0}=\phi_0^{(0)} \text{ and } \psi_{j,k}=2^{j/2}(\phi_{2k}^{(j+1)}-\phi_{2k+1}^{(j+1)}), \forall j\in\mathbb{N}, k=0,\dots,2^j-1$$

Any function $f\in\mathcal{L}^2([0,1],\lambda_1)$ can be expanded in the Haar basis. We denote by $f_{j,k}$ the (j,k)-th coefficient $<f,\psi_{j,k}>$ of f. We consider a mapping $\Gamma:\mathbb{N}\cup\{-1\}\longmapsto\mathbb{N}$ such that $\sum_j 2^j\Gamma(j)^{-1}<\infty$ and the set of functions

$$\mathcal{H}_\Gamma=\{f\in\mathcal{L}^2([0,1]):\forall x\in[0,1], f(x)=\sum_{j,k}f_{j,k}\psi_{j,k}(x) \text{ and } \sum_{j,k}\Gamma(j)|f_{j,k}|^2<\infty\}$$

endowed with the inner product $<f,h>_\Gamma=\sum_{j,k}\Gamma(j)f_{j,k}h_{j,k}$ and the associated norm denoted by $\|\cdot\|_\Gamma$. It is easy to see that $\mathcal{H}_\Gamma$ is a Hilbert space. Note that, for any point $x\in[0,1]$, the sum $\sum_{j=-1}^{\infty}\sum_{k=0}^{2^j-1}f_{j,k}\psi_{j,k}(x)$ converges when $\sum_{j,k}\Gamma(j)|f_{j,k}|^2<\infty$.

We now prove that $\mathcal{H}_\Gamma$ is a RKHS. For that, we just have to prove that any point evaluation is a bounded linear functional. Let $x\in[0,1]$ and $f\in\mathcal{H}_\Gamma$. We have $|f(x)|=\left|\sum_{j,k}f_{j,k}\psi_{j,k}(x)\right|\leq\left(\sum_{j=0}^{\infty}2^j\Gamma(j)^{-1}\right)^{1/2}\|f\|_\Gamma$. Thus, for any point x, the linear functional $f\in\mathcal{H}_\Gamma\longmapsto f(x)$ is continuous and so $\mathcal{H}_\Gamma$ is a RKHS. The reproducing kernel associated with is given by

$$K_w(x,y)=\sum_{j,k}\Gamma(j)\psi_{j,k}(x)\psi_{j,k}(y), \forall x,y\in[0,1].$$

Let $f:[0,1]\longmapsto\{-1,1\}$ be a prediction rule such that $\forall x\in[0,1], f(x)=\sum_{j,k}f_{j,k}\psi_{j,k}(x)$ with the coefficients $f_{j,k}$ only in $\{-2^{-j/2},0,2^{-j/2}\}$ for any j,k. Let $w:\mathbb{N}\longmapsto\mathbb{N}$ be such that $\sum_j 2^{-j}\Gamma(j)w(j+1)\leq\infty$ and assume that $\text{card}\{k\in\{0,\dots,2^j-1\}:f_{j,k}\neq 0\}\leq w(j)$. Then, it is easy to see that $f\in\mathcal{F}^{(1)}\cap\mathcal{H}_\Gamma$. Comparing the set $\mathcal{H}_\Gamma$ intersected with the set of all prediction rules and some L^1-ellipsoids $\mathcal{F}_w^{(1)}$ is not an easy task. We do not investigate this comparison further long in this paper.

5.2. Boundary fragments

Considering the classification problem on 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 included in $[0,1]^2$, and -1 outside $\mathcal{C}$. In our framework, two questions need to be considered:

- What is the representation of the simple function $f_{\mathcal{C}}^*$ in the fundamental system $\mathcal{S}$ using only coefficients with values in $\{-1,0,1\}$?
- 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 boundary is an infinite differentiable curve), the direct estimation of the boundary is a better approach. The main reason is that a 2-dimensional estimation problem becomes a 1-dimensional problem. Such a reduction of the dimension makes the estimation easier (note that, our approach is specifically good in the 1-dimensional case, since the notion of boundary does not exist). Nevertheless, our approach is applicable for the estimation of such a function (cf. Theorem 7). Actually, a direct estimation of the boundary reduces the dimension but there is a loss of observations since observations far from the boundary are not used by this estimation point of view (they are only used to detect the boundary). This may explain why our approach is applicable. Denote by

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

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 boundary of A.

Theorem 7. *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 on $\mathbb{R}_+^*$ with values in $\mathbb{R}^+$ satisfying $\epsilon^2\delta(\epsilon)\longrightarrow 0$ when $\epsilon\longrightarrow 0$. 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 constant factor, the same as the one appearing in Theorem 5. In a more general way, any prediction rule with a boundary having a finite perimetry (for instance polygons) is close (w.r.t. the $L^1(\lambda_2)$-norm) to a function developed in the fundamental system $\mathcal{S}$ at the frequency rank J_n, with an error of the order $(\log n)/n$.

Now, the problem is about finding a L^1-ellipsoid of prediction rules such that for any integer n the approximation function f_n belongs to this ellipsoid. This problem depends on the geometry of the boundary set ∂A. It comes naturally since we made the choice of a particular geometry for the basis partitions: 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 is, in our case, the same as the choice "a prior" of a wavelet basis, for instance, in the density estimation problem. Depending on the type of Bayes rules we have to estimate, a special partition can be considered. Here, "dyadic approach" is very well adapted for the estimation of Bayes rules associated with chessboard (with the value 1 for black square and -1 for white square). This kind of Bayes rules are very badly estimated by classification procedures estimating the boundary since most of these procedures require regularity assumptions which are not fulfilled in the case of chessboards.

In the general case, the ideal choice of the geometry is adapted to the particular geometry induced by the measure μ on $[0,1]^d$, defined by

$$\mu(A) = \int_A |2\eta(x) - 1| P^X(dx),$$

for any measurable set $A \subseteq [0,1]^d$. Namely, we do not need a good resolution of the partition for the regions of $[0,1]^d$ with a low $\mu-$probability. However, we need a sharper resolution for regions with a high $\mu-$probability. In our case (under assumptions (A1) and (SMA)), the measure μ is equivalent to the Lebesgue measure. Thus, we do not need different scale of resolution for different areas of the square $[0,1]^d$.

Nevertheless, in some cases, it is possible to make some connections between the "estimation of the boundary" point of view and the geometrical point of view presented here. For that, we consider the models introduced in Mammen and Tsybakov [1999]. They obtain minimax rates of convergence for classes of Bayes rules $f = 2\mathbb{1}_G - 1$ where $G \subset [0,1]^d$ is a *boundary fragment* of smoothness γ. That is $G = \mathrm{epi}(g)$ is the epigraph $\{(s,t) \in [0,1]^d : g(s) \leq t\}$ of a function $g : [0,1]^{d-1} \longmapsto [0,1]$ with Hölder regularity γ and Lipschitz constant c (that is $\forall s,t \in \mathbb{R}^{d-1}, |g(s) - g(t)| \leq c|s-t|^\gamma$). We denote by $\mathcal{F}_{BF}^{(d)}(\gamma, c)$ this class of prediction rules. For $\gamma \geq 1$, the boundary ∂G of a boundary fragment G with smoothness γ has a finite perimetry and thus is close to an element of $\mathcal{F}^{(d)}$ at frequency J_n (cf. Theorem 7). In this particular case, it is possible to obtain more precise results written in the following theorem.

Theorem 8. *We consider the function* $w : \mathbb{N} \longmapsto \mathbb{N}$ *defined by* $w(j) = (2c\sqrt{d-1}+3)2^{j(d-1)}, \forall j \in \mathbb{N}$. *We have*

$$\mathcal{F}_{BF}^{(d)}(1,c) \subset \mathcal{F}_w^{(d)}.$$

By applying Theorem 3 and Theorem 4, we can easily obtain that the minimax rate of convergence over $\mathcal{F}_w^{(d)}$ for $w(j) \sim 2^{j(d-1)}$ is between $((\log n)/n)^{1/d}$ and $n^{-1/d}$. Moreover, we know by Tsybakov [2004] that the minimax rate of

convergence over $\mathcal{F}_{BF}(1,c)$ (under (SMA)) is $n^{-1/d}$. First, we can conclude that the L^1-ellipsoid $\mathcal{F}_w^{(d)}$ is a larger class of prediction rules than $\mathcal{F}_{BF}^{(d)}(1,c)$ (it is easy to construct examples of prediction rules which are in $\mathcal{F}_w^{(d)}$ but not in $\mathcal{F}_{BF}^{(d)}(1,c)$) and there is at most a logarithm loss in the minimax rate. Second, under the (SMA) and (A1), histogram classifiers provide another way to achieve the minimax rate of convergence for the class $\mathcal{F}_{BF}^{(d)}(1,c)$. Finally, "geometric" and "boundary estimation" points of view can provide the same results in some particular cases.

For previous work on the connections between boundaries fragments and dyadic trees, we refer the reader to Scott and Nowak [2006].

5.3. Perspectives

We can extend our approach in several ways. First, it seems possible to avoid assumption (A1) by using the same tools as those used for Histogram rules (cf. Section 6 in Devroye et al. [1996]). Similarly, assumption (SMA) may be relaxed in favor of a genral Tsybakov's noise condition.

Next, 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 cell $\mathcal{I}_{\mathbf{k}}^{(J_n)}$, where $\mathbf{k} \in I_d(J_n)$, we can compute the line minimizing

$$\sum_{i=1}^{n} (f(X_i) - Y_i)^2 \mathbb{1}_{(X_i \in \mathcal{I}_{\mathbf{k}}^{(J_n)})},$$

where f is taken in the set of all indicators of half spaces of $[0,1]^d$ intersecting $\mathcal{I}_{\mathbf{k}}^{(J_n)}$. Of course, depending on the number of observations inside the cell $\mathcal{I}_{\mathbf{k}}^{(J_n)}$, we can consider larger classes of indicators 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 extended to smoother procedures.

In this paper, we start by considering a model of prediction rules. Then, we provide an approximation theorem for these models. The form of object approaching the Bayes rule in these models leads to a particular form of estimators (here the histogram estimators). Finally, the way the estimator depends on the complexity of the underlying model (here the level of frequency) impose a way to construct adaptive estimators. As we can see everything depends on the starting model we consider.

For the one-dimensional case, another point of view is to consider $f^* \in L^2([0,1])$ and to develop f^* in an orthonormal wavelet basis of $L^2([0,1])$. Namely, $f^* = \sum_{j,k} f_{j,k}\psi_{j,k}$,where $f_{j,k} =< f^*, \psi_{j,k} >$ for any $j \in \mathbb{N} \cup \{-1\}$ and $k = 0, \ldots, 2^j - 1$. For the control of the bias term, a classical assumption is to take the family of coefficients $(f_{j,k})_{j,k}$ in a L^1-ellipsoid of $\mathbb{R}^{\mathbb{N}}$. This point of view leads to functional analysis and estimation issues. First problem: which functions with values in $\{-1,1\}$ have wavelet coefficients in a L^1-ellipsoid and which

wavelet basis is more adapted to this problem (maybe the Haar basis)? Second problem: what 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 direction. We can generalize this approach to functions with values in $\mathbb{Z}$. When functions take values in $\mathbb{R}$, for instance in the regression problem, usual approximation theory is used to obtain a control on the bias term. Finally, remark that functions with values in $\{-1,1\}$ can be approximated by real-valued (possibly smooth) functions; this is for example what is used for SVM or boosting. In those cases, control of the approximation term is still an open question (cf. Steinwart and Scovel [April 2007] and Lugosi and Vayatis [2004]).

6. Proofs

In all the proofs, we use the analytical representation of the predictions rules to underly the similitude with the technics used in the wavelet literature. Nevertheless, these proofs can be obtained by using the dyadic decision tree representation.

Proof of Proposition 1. Let $\epsilon > 0$. We construct recursively an integer N and a family $A = (A_{\mathbf{k}}^{(j)})_{j=0,\dots,N;\mathbf{k}\in I_d(j)}$ of elements in $\{-1,0,1\}$ satisfying the writing convention (W) and such that $\|f - \sum_{j=0}^{N}\sum_{\mathbf{k}\in I_d(j)} A_{\mathbf{k}}^{(j)}\phi_{\mathbf{k}}^{(j)}\|_{L^2(\lambda_d)} \leq \epsilon$.

First, we initialize the construction with a family $A_0 = (A_{\mathbf{k}}^{(j,0)})_{j=0,\dots,N;\mathbf{k}\in I_d(j)}$ of elements in $\{-1,0,1\}$ such that $\|f - \sum_{j=0}^{N}\sum_{\mathbf{k}\in I_d(j)} A_{\mathbf{k}}^{(j,0)}\phi_{\mathbf{k}}^{(j)}\|_{L^2(\lambda_d)} \leq \epsilon$. The integer N and the family A_0 exist because f belongs to $\mathcal{F}^{(d)}$. Without loss of generality we can take A_0 such that $\forall j < N, \mathbf{k} \in I_d(j), A_{\mathbf{k}}^{(j,0)} = 0$ and $\forall \mathbf{k} \in I_d(N), A_{\mathbf{k}}^{(j,0)} \neq 0$ (we can split the cell of frequency $j < N$ up to the frequency N).

Then, let p be an integer in $\{0,\dots,N-1\}$. Given is a family A_p of elements $(A_{\mathbf{k}}^{(j,p)})_{j=0,\dots,J;\mathbf{k}\in I_d(j)}$ in $\{-1,0,1\}$ such that $\sum_{j=0}^{N}\sum_{\mathbf{k}\in I_d(j)} A_{\mathbf{k}}^{(j,0)}\phi_{\mathbf{k}}^{(j)}$ takes its values in $\{-1,1\}$ a.s., we construct a family $A_{p+1} = (A_{\mathbf{k}}^{(j,p+1)})_{j=0,\dots,J;\mathbf{k}\in I_d(j)}$ of elements in $\{-1,0,1\}$ such that: at the frequency $J = N-p-1$, for any multi-index $\mathbf{k} \in I_d(J)$ such that $A_{\mathbf{k}'}^{(J+1,p)} = 1$ for all $\mathbf{k}' \in I_d(J+1)$ satisfying $\phi_{\mathbf{k}}^{(J)}\phi_{\mathbf{k}'}^{(J+1)} \neq 0$, we define $A_{\mathbf{k}}^{(J,p+1)} = 1$ and the other 2^d coefficients $A_{\mathbf{k}'}^{(J+1,p+1)} = 0$ such that $\phi_{\mathbf{k}}^{(J)}\phi_{\mathbf{k}'}^{(J+1)} \neq 0$. The same construction holds when we replace 1 by -1. Otherwise we take $A_{\mathbf{k}}^{(J,p+1)} = A_{\mathbf{k}}^{(J,p)}$ and $A_{\mathbf{k}'}^{(J+1,p+1)} = A_{\mathbf{k}'}^{(J+1,p)}$ for all $\mathbf{k}' \in I_d(J+1)$ satisfying $\phi_{\mathbf{k}}^{(J)}\phi_{\mathbf{k}'}^{(J+1)} \neq 0$.

After N iteration, it is easy to see that $A = A_N$ satisfies the writing conven-

tion (W) and by construction $\sum_{j=0}^{N}\sum_{\mathbf{k}\in I_d(j)} A_{\mathbf{k}}^{(j,0)}\phi_{\mathbf{k}}^{(j)} = \sum_{j=0}^{N}\sum_{\mathbf{k}\in I_d(j)} A_{\mathbf{k}}^{(j,N)}\phi_{\mathbf{k}}^{(j)}$. □

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 $\{\mathcal{O}_x : x \in \mathcal{O}\} = \{\mathcal{I}_k^{(j)} : (j,\mathbf{k}) \in I_{\mathcal{O}}\}$. 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. For any $(j,\mathbf{k}) \in I_{\mathcal{O}_1}$ we take $a_{\mathbf{k}}^{(j)} = -1$.

For any $(j,\mathbf{k}) \notin I_{\mathcal{O}} \cup I_{\mathcal{O}_1}$, 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 $\mathcal{F}^{(d)}$, satisfies the writing convention (W) and, for λ_d-almost $x \in [0,1]^d$, $g(x) = f_\eta(x)$. □

Proof of Proposition 2. 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 $j \in \mathbb{N}, \mathbf{k} \in I_d(j)$. Consider $f_2 = \sum_{j\in\mathbb{N}}\sum_{\mathbf{k}\in I_d(j)} b_{\mathbf{k}}^{(j)}\phi_{\mathbf{k}}^{(j)}$. Remark that the function $f_2 \in \mathcal{F}^{(d)}$ but does not satisfy the writing convention (W). We have $f_2 = \mathbb{1}_{[0,1]^d}$ a.s.. For any $j \in \mathbb{N}$ we have

$$\operatorname{card}\left\{\mathbf{k} \in I_d(j) : b_{\mathbf{k}}^{(j)} \neq 0\right\} = \operatorname{card}\left\{\mathbf{k} \in I_d(j) : a_{\mathbf{k}}^{(j)} \neq 0\right\}. \tag{12}$$

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

$$1 = \sum_{j\in\mathbb{N}} 2^{-dj}\operatorname{card}\left\{\mathbf{k} \in I_d(j) : b_{\mathbf{k}}^{(j)} \neq 0\right\}. \tag{13}$$

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

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

We obtain the second assertion of Proposition 2 by using the last inequality and both of the assertions (12) and (13).

Assume that $\sum_{j=1}^{+\infty} 2^{-dj} w(j) \geq 1$. For any integer $j \neq 0$, denote by $\mathrm{Ind}(j)$ the set of indexes $\{(j,\mathbf{k}) : \mathbf{k} \in I_d(j)\}$. We use the lexicographic order of $\mathbb{N}^{d+1}$ to order sets of indexes. Take $\mathrm{Ind}_w(1)$ the family of the first $w(1)$ elements of $\mathrm{Ind}(1)$. Denote by $\mathrm{Ind}_w(2)$ the family made of the first $w(1)$ elements of $\mathrm{Ind}(1)$ and add, at the end of this family in the correct order, the first $w(2)$ elements $(2,\mathbf{k})$ of $\mathrm{Ind}(2)$ such that $\phi_{\mathbf{k}'}^{(1)}\phi_{\mathbf{k}}^{(2)} = 0$ for any $(1,\mathbf{k}') \in \mathrm{Ind}_w(1), \ldots$, for the step j, construct the family $\mathrm{Ind}_w(j)$ made of all the elements of $\mathrm{Ind}_w(j-1)$ in the same order and add at the end of this family the indexes $(j,\mathbf{k})$ of $\mathrm{Ind}(j)$ among the first $w(j)$ elements of $\mathrm{Ind}(j)$ such that $\phi_{\mathbf{k}'}^{(J)}\phi_{\mathbf{k}}^{(j)} = 0$ for any $(J,\mathbf{k}') \in \mathrm{Ind}_w(j-1)$. If there is no more indexes satisfying this condition then, we stop the construction, otherwise, we go on. Denote by Ind the final family obtained by this construction (Ind can be finite or infinite). Then, we enumerate the indexes of Ind by $(j_1,\mathbf{k}_1) \prec (j_2,\mathbf{k}_2) \prec \cdots$. For the first $(j_1,\mathbf{k}_1) \in \mathrm{Ind}$ 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 $w(j)$ and the total mass of cells $\mathcal{I}_{\mathbf{k}}^{(j)}$ such that $a_{\mathbf{k}}^{(j)} \neq 0$ is $\sum_{j\in\mathbb{N}} 2^{-dj} \mathrm{card}\{\mathbf{k} \in I_d(j) : a_{\mathbf{k}}^{(j)} \neq 0\}$ which is greater or equal to 1 by assumption. Thus, all the hypercube is filled by cells associated with 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 with a coefficient non equals to 0 with an opposite value. Moreover f is not $\mathbb{I}_{[0,1]^d}$. □

Proof of Theorem 2. Let $\pi = (P^X, \eta)$ be a probability measure on $[0,1]^d \times \{-1,1\}$ in $\mathcal{P}_{w,A}$. Denote by f^* a Bayes rule associated with π (for example $f^* = \mathrm{sign}(2\eta - 1)$). We have

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

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

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

We write f^* in the fundamental system $(\phi_{\mathbf{k}}^{(j)}, j \geq J_\epsilon)$ using the convention of writing of section 4.1. Remark that, we start the expansion of f^* at the level of

frequency J_ϵ and then, we use the writing convention (W) on the coefficients of this expansion. Namely, we consider

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

Next, we define the best approximation of f^* at the frequency level J_ϵ by

$$f_\epsilon = \sum_{\mathbf{k}\in I_d(J_\epsilon)} B_{\mathbf{k}}^{(J_\epsilon)}\phi_{\mathbf{k}}^{(J_\epsilon)}, \text{ where } B_{\mathbf{k}}^{(J_\epsilon)} = \left\{ \begin{array}{ll} 1 & \text{if } p_{\mathbf{k}}^{(J_\epsilon)} > 1/2 \\ -1 & \text{otherwise} \end{array} \right. \tag{14}$$

and

$$p_{\mathbf{k}}^{(J_\epsilon)} = \mathbb{P}(Y=1|X\in \mathcal{I}_{\mathbf{k}}^{(J_\epsilon)}) = \int_{\mathcal{I}_{\mathbf{k}}^{(J_\epsilon)}} \eta(x)\frac{dP^X(x)}{P^X(\mathcal{I}_{\mathbf{k}}^{(J_\epsilon)})}, \tag{15}$$

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^* takes its values in $\{-1,1\}$, thus, we have

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

□

Proof of Theorem 3. Let $\pi = (P^X, \eta)$ be a probability measure on $[0,1]^d \times \{-1,1\}$ satisfying (A1), (SMA) and such that $f^* = \text{sign}(2\eta - 1)$, a Bayes classifier associated with π, belongs to $\mathcal{F}_w^{(d)}$ (an L^1-ellipsoid of Bayes rules).

Let $\epsilon > 0$ and J_ϵ the smallest integer satisfying $\sum_{j=J_\epsilon+1}^{+\infty} 2^{-dj}w(j) < \epsilon/A$. 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 (6) and f_ϵ in (14).

Using the definition of J_ϵ and according to the approximation Theorem (Theorem 2), the bias term satisfies:

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

For the variance term we have (using the notations introduced in (6) and (14)):

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

Now, we apply a concentration inequality in each cell of the dyadic partition $(\mathcal{I}_\mathbf{k}^{(J_\epsilon)} : \mathbf{k} \in I_d(J_\epsilon))$. Let $\mathbf{k} \in I_d(J_\epsilon)$. We introduce the following events:

$$\Omega_\mathbf{k}^{(m)} = \left\{\text{Card}\{i \in \{1,\dots,n\} : X_i \in \mathcal{I}_\mathbf{k}^{(J_\epsilon)}\} = m\right\}, \forall m \in \{0,\dots,n\}$$

and

$$\Omega_\mathbf{k} = \{N_\mathbf{k}^{(J_\epsilon)+} \leq N_\mathbf{k}^{(J_\epsilon)-}\},$$

where $N_\mathbf{k}^{(J_\epsilon)+}$ and $N_\mathbf{k}^{(J_\epsilon)-}$ have been defined in subsection 4.2. 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

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

Moreover, if we denote by $Z_1,\dots,Z_n$ n i.i.d. random variables with a Bernoulli with parameter $p_\mathbf{k}^{(J_\epsilon)}$ for common probability distribution (we recall that $p_\mathbf{k}^{(J_\epsilon)}$ is introduced in (15) and is equal to $\mathbb{P}(Y = 1|X \in I_\mathbf{k}^{(J_\epsilon)})$), we have for any $m = 1,\dots,n$,

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

The concentration inequality of Hoeffding leads to

$$\mathbb{P}\left(\frac{1}{m}\sum_{i=1}^m Z_i \geq p_\mathbf{k}^{(J_\epsilon)} + t\right) \leq \exp(-2mt^2) \tag{16}$$

and

$$\mathbb{P}\left(\frac{1}{m}\sum_{i=1}^m Z_i \leq p_\mathbf{k}^{(J_\epsilon)} - t\right) \leq \exp(-2mt^2), \tag{17}$$

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

Denote by $b_{\mathbf{k}}^{(J_\epsilon)}$ the probability $\mathbb{P}(X\in\mathcal{I}_{\mathbf{k}}^{(J_\epsilon)})$. If $p_{\mathbf{k}}^{(J_\epsilon)}>1/2$, applying inequality (17) leads to

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

If $p_{\mathbf{k}}^{(J_\epsilon)}<1/2$ then, similar arguments used in the previous case and inequality (16) lead to

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

If $p_{\mathbf{k}}^{(J_\epsilon)}=1/2$, we use $\mathbb{P}\left(|B_{\mathbf{k}}^{(J_\epsilon)}-\hat{A}_{\mathbf{k}}^{(J_\epsilon)}|=2\right)\leq 1$. 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:

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

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

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

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

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

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

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

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

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

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

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 $\mathcal{X'}_1^{(q)}, \ldots, \mathcal{X'}_{2^{dq}}^{(q)}$ 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 $x'^{(q)}_{k_1,\ldots,k_d} = \left(\frac{2k_1+1}{2^{q+1}}, \ldots, \frac{2k_d+1}{2^{q+1}}\right)$. We say that $x'^{(q)}_{k_1,\ldots,k_d} \prec x'^{(q)}_{k'_1,\ldots,k'_d}$ if

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

or

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

for the lexicographical order on $\mathbb{N}^d$. Thus, the partition $(\mathcal{X'}_j^{(q)} : j = 1, \ldots, 2^{dq})$ has an increasing indexation according to the order of $(x'^{(q)}_{k_1,\ldots,k_d})$ for the order defined above. This order take care of the previous partition by splitting blocks in the given right order and, inside a block of a partition, we take the lexicographic order of $\mathbb{N}^d$. We introduce an other parameter $m \in \{1, \ldots, 2^{qd}\}$ and we define for all $i = 1, \ldots, m$, $\mathcal{X}_i^{(q)} = \mathcal{X'}_i^{(q)}$ 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 $\mathcal{X}_1, \ldots, \mathcal{X}_m$ 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 $\sigma = (\sigma_1, \ldots, \sigma_m) \in \Omega = \{-1, 1\}^m$ we consider η_σ defined, for any $x \in [0,1]^d$, by

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

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 choose

$$W = 2^{-dq}, \tag{18}$$

then, $f_X = \mathbb{1}_{[0,1]^d}$ (so $P^X << \lambda$ 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

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

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

$$w(q+1) \geq \inf(x \in 2^d\mathbb{N} : x \geq m). \tag{19}$$

Take q, m and W such that (18) and (19) 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 neighboring probability measures. Let ρ be the Hamming distance on Ω. Let σ, σ' in Ω such that $\rho(\sigma, \sigma') = 1$. We have

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

and a straightforward calculus leads to $H^2(\pi_\sigma, \pi_{\sigma'}) = 2W\left(1-\sqrt{1-h^2}\right)$. If we have $W \leq 1/n$ then, $H^2(\pi_\sigma^{\otimes n}, \pi_{\sigma'}^{\otimes n}) \leq \beta < 2$ where $\beta = 2\left(1-\exp(1-\sqrt{1-h^2})\right)$. One version of the Assouad Lemma (cf. Assouad [1983] or Lecué [2007]) yields

$$\inf_{\hat{\sigma}_n \in [-1,1]^m} \sup_{\sigma \in \{-1,1\}^m} \mathbb{E}_{\pi_\sigma}\left[\sum_{i=1}^m |\sigma_i - \hat{\sigma}_i|\right] \geq (m/4)\left(1-(\beta/2)\right)^2.$$

We conclude that

$$\inf_{\hat{f}_n} \sup_{\pi \in \mathcal{P}_{w,h,a,A}} \mathcal{E}_\pi(\hat{f}_n) \geq Wh\frac{m}{8}\left(1-\frac{\beta}{2}\right)^2. \tag{20}$$

Finally, we take $m = w\left(\lfloor \log n/(d\log 2)\rfloor + 1\right) - (2^d-1)$, $q = \lfloor \log n/(d\log 2)\rfloor$ and $W = 2^{-dq} \leq 1/n$. Next, replacing these values in (20), we obtain the result. □

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. First, if we assume that $J_\epsilon \geq K$ then $\sum_{j=J_\epsilon+1}^{+\infty} 2^{-dj} w_K^{(d)}(j) = (2^{dK})/(2^{dJ_\epsilon}(2^d-1))$. 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 $J_n(K) = J_{\epsilon_n}$, we have

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

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

If we have $\lfloor \log n/(d\log 2)\rfloor \geq 2$ then $w\left(\lfloor \log n/(d\log 2)\rfloor + 1\right) - (2^d-1) \geq 2^d$, 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 $C_{0,K} = C_0(2^{dK} - (2^d-1))$.

Second, if we have $J_\epsilon \geq N^{(d)}(\alpha)$, then $\sum_{j=J_\epsilon+1}^{+\infty} 2^{-dj} w_\alpha^{(d)}(j) \leq (2^{d(1-\alpha)J_\epsilon}(2^{d(1-\alpha)}-1))^{-1}$. 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 $J_n(\alpha) = J_{\epsilon_n}$, we have

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

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 $\alpha \in (0,1)$. For n large enough, we have $J^{(n)} \geq J_m(\alpha)$. Since the (SMA) assumption is equivalent to the margin assumption introduced by Mammen and Tsybakov [1999] and Tsybakov [2004] with margin parameter equal to 1 (cf. proof of Proposition 1 of Lecué [2007]) we have, according to Corollary 1 of Lecué [2006],

$$\mathbb{E}[R(\hat{f}_n) - R^*] \leq 3 \min_{J=0,\ldots,J^{(n)}} \mathbb{E}[R(\hat{f}_m^{(J)}) - R^*] + C\frac{(\log n)\log(J^{(n)}+1)}{n}. \quad (21)$$

According to Theorem 5, we have

$$\mathbb{E}[R(\hat{f}_m^{(J)}) - R^*] \leq C'_{\alpha,h,a,A}\Big(\frac{\log m}{m}\Big)^{1-\alpha}.$$

Then, combining the last inequality, the fact that $m \leq n/2$ and (21), we complete the proof. □

Proof of Theorem 7. 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

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

□

Proof of Theorem 8. Let $g : \mathbb{R}^{d-1} \longmapsto \mathbb{R}$ be a Hölder function with regularity $\gamma = 1$ and Lipschitz constant c. We denote by G the epigraph of g and by $f = 2\mathbb{1}_G - 1$ the prediction rule associated with the set G. By continuity of g and Theorem 1, f belongs to $\mathcal{F}^{(d)}$. By Proposition 1 there exists a sequence $(a_{\mathbf{k}}^{(j)})_{j,\mathbf{k}}$ of coefficients with values in $\{-1,0,1\}$ satisfying the writing convention (W) such that $f = \sum_{j,\mathbf{k}} a_{\mathbf{k}}^{(j)} \phi_{\mathbf{k}}^{(j)}$. We denote by $N(j) = \text{card}\{\mathbf{k} \in \{0,\ldots,2^{dj}-1\} : a_{\mathbf{k}}^{(j)} \neq 0\}, \forall j \in \mathbb{N}$. We want to prove that $f \in \mathcal{F}_w^{(d)}$. For that, it is enough to prove that $N(j) \leq w(j), \forall j \in \mathbb{N}$.

Let $j \in \mathbb{N} - \{0\}$. It is easy to see that the minimal number of cells $(\mathcal{I}_{\mathbf{k}}^{(j)} : \mathbf{k} \in \{0,\ldots,2^j-1\}^d)$ intersected by ∂G is $2^{j(d-1)}$ and the maximal number of

intersected cells of frequency j is $c'2^{j(d-1)}$ where $c' = c\sqrt{d-1}+2$ (cf. lemma 2 of Scott and Nowak [2006]). We denote by G_j the union of all the cells of $(\mathcal{I}_{\mathbf{k}}^{(j)} : \mathbf{k} \in \{0,\dots,2^j-1\}^d)$ intersected by ∂G. By the writing convention all the others cells at frequency j of $[0,1]^d - G_j$ are associated with a non-negative coefficient. Thus, we have

$$\Big\| \sum_{p=0}^{j} \sum_{\mathbf{k}} a_{\mathbf{k}}^{(j)} \phi_{\mathbf{k}}^{(j)} \Big\|_2 \in [1-c'2^{-j}, 1-2^{-j}],$$

in other words, $S_j = \sum_{p=0}^{j} 2^{-dp} N(p) \in [1-c'2^{-j}, 1-2^{-j}]$. This holds for any $j \geq 1$, thus we have $2^{-d(j+1)} N(j+1) \leq 1 - S_j - 2^{-(j+1)} \leq (2c'-1)2^{-(j+1)}$. This means that $\forall j \geq 1, N(j) \leq w(j)$. □

Acknowledgements

I want to thank Albert Cohen, Lucien Birgé, Stéphane Boucheron, Gérard Kerkyacharian, my advisor: Alexandre Tsybakov, and the referees for giving me many ideas and advices for this work.

References

A. Antos, L. Devroye, and L.Györfi. Lower bounds for Bayes error estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21:643–645, 1999.

N. Aronszajn. Theory of reproducing kernels. *Trans. Am. Math. Soc.*, 68:337–404, 1950. MR0051437

P. Assouad. Deux remarques sur l'estimation. *C. R. Acad. Sci. Paris Sér. I Math.*, 296(23):1021–1024, 1983. French. MR0777600

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

G. Blanchard, C. Schäfer, Y. Rozenholc, and K-R. Müller. Optimal dyadic decision trees. *Machine Learning*, 66(2-3):209–242, 2007.

S. Boucheron, O. Bousquet, and G. Lugosi. Theory of classification: a survey of some recent advances. *ESAIM: Probability and Statistics*, 9:323–375, 2005. MR2182250

L. Breiman, J. Freidman, J. Olshen, and C. Stone. Classification and regression trees. Wadsworth, 1984.

T. M. Cover and J. A. Thomas. *Elements of Information Theory*. 1991. Second edition, 2006. MR2239987

L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer, New York, Berlin, Heidelberg, 1996. MR1383093

A.P. Korostelev and A.B. Tsybakov. *Minimax Theory of Image Reconstruction*, volume 82 of *Lecture Notes in Statistics*. NY e.a., 1993. MR1226450

G. Lecué. Optimal oracle inequality for aggregation of classifiers under low noise condition. *In Proceeding of the 19th Annual Conference on Learning Theory, COLT 2006*, 32(4):364–378, 2006. MR2280618

G. Lecué. Optimal rates of aggregation in classification under low noise assumption. *Bernoulli*, 13(4):1000–1022, 2007. MR2364224

G. Lugosi and N. Vayatis. On the bayes-risk consistency of regularized boosting methods. *Ann. Statist.*, 32(1):30–55, 2004. MR2051000

E. Mammen and A.B. Tsybakov. Smooth discrimination analysis. *Ann. Statist.*, 27:1808–1829, 1999. MR1765618

P. Massart and E. Nédélec. Risk Bound for Statistical Learning. *Ann. Statist.*, 34(5), 2006.

Y. Meyer. *Ondelettes et Opérateurs*. Hermann, Paris, 1990. MR1085487

S. Murthy. Automatic construction of decision trees from data: A multi-disciplinary survey. *Data Mining and knowledge Discovery*, 2(4):345–389, 1998.

J.R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, San Mateo, 1993.

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

I. Steinwart, D. Hush, and C. Scovel. Function classes that approximate the Bayes risk. *In Proceeding of the 19th Annual Conference on Learning Theory, COLT 2006*, 32(4):79–93, 2006. MR2277920

I. Steinwart and C. Scovel. Fast Rates for Support Vector Machines using Gaussian Kernels. *Ann. Statist.*, 35(2), April 2007. MR2336860

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series BB 58*, pages 267–288, 1996. MR1379242

A.B. Tsybakov. Optimal aggregation of classifiers in statistical learning. *Ann. Statist.*, 32(1):135–166, 2004. MR2051002

A.B. Tsybakov and S.A. van de Geer. Square root penalty: adaptation to the margin in classification and in edge estimation. *Ann. Statist.*, 33:1203–1224, 2005. MR2195633

Y. Yang. Minimax nonparametric classification—part I: Rates of convergence. *IEEE Transaction on Information Theory*, 45:2271–2284, 1999a. MR1725115

Y. Yang. Minimax nonparametric classification—partII: Model selection for adaptation. *IEEE Transaction on Information Theory*, 45:2285–2292, 1999b. MR1725116