Fast learning rates for plug-in classifiers under the margin condition

Jean-Yves Audibert, Alexandre Tsybakov

Fast learning rates for plug-in classifiers under the margin condition

Jean-Yves AUDIBERT[1] and Alexandre B. TSYBAKOV[2]

[1]Ecole Nationale des Ponts et Chaussées, [2]Université Pierre et Marie Curie

May 24, 2011

Abstract

It has been recently shown that, under the margin (or low noise) assumption, there exist classifiers attaining fast rates of convergence of the excess Bayes risk, i.e., the rates faster than $n^{-1/2}$. The works on this subject suggested the following two conjectures: (i) the best achievable fast rate is of the order n^{-1}, and (ii) the plug-in classifiers generally converge slower than the classifiers based on empirical risk minimization. We show that both conjectures are not correct. In particular, we construct plug-in classifiers that can achieve not only the fast, but also the *super-fast* rates, i.e., the rates faster than n^{-1}. We establish minimax lower bounds showing that the obtained rates cannot be improved.

AMS 2000 Subject classifications. Primary 62G07, Secondary 62G08, 62H05, 68T10.

Key words and phrases. classification, statistical learning, fast rates of convergence, excess risk, plug-in classifiers, minimax lower bounds.

Short title. Fast Rates for Plug-in Classifiers

1 Introduction

Let (X, Y) be a random couple taking values in $\mathcal{Z} \triangleq \mathbb{R}^d \times \{0, 1\}$ with joint distribution P. We regard $X \in \mathbb{R}^d$ as a vector of features corresponding to an object and

$Y \in \{0,1\}$ as a label indicating that the object belongs to one of the two classes. Consider the sample $(X_1, Y_1), \ldots, (X_n, Y_n)$, where (X_i, Y_i) are independent copies of (X, Y). We denote by $P^{\otimes n}$ the product probability measure according to which the sample is distributed, and by P_X the marginal distribution of X.

The goal of a classification procedure is to predict the label Y given the value of X, i.e., to provide a decision rule $f : \mathbb{R}^d \to \{0,1\}$ which belongs to the set $\mathcal{F}$ of all Borel functions defined on $\mathbb{R}^d$ and taking values in $\{0,1\}$. The performance of a decision rule f is measured by the misclassification error

$$R(f) \triangleq P(Y \neq f(X)).$$

The Bayes decision rule is a minimizer of the risk $R(f)$ over all the decision rules $f \in \mathcal{F}$, and one of such minimizers has the form $f^*(X) = \mathbb{I}_{\{\eta(X) \geq \frac{1}{2}\}}$ where $\mathbb{I}_{\{\cdot\}}$ denotes the indicator function and $\eta(X) \triangleq P(Y = 1|X)$ is the regression function of Y on X (here $P(dY|X)$ is a regular conditional probability which we will use in the following without further mention).

An empirical decision rule (a classifier) is a random mapping $\hat{f}_n : \mathcal{Z}^n \to \mathcal{F}$ measurable w.r.t. the sample. Its accuracy can be characterized by the excess risk

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

where $\mathbb{E}$ is the sign of expectation. A key problem in classification is to construct classifiers with small excess risk for sufficiently large n [cf. Devroye, Györfi and Lugosi (1996), Vapnik (1998)]. Optimal classifiers can be defined as those having the best possible rate of convergence of $\mathcal{E}(\hat{f}_n)$ to 0, as $n \to \infty$. Of course, this rate, and thus the optimal classifier, depend on the assumptions on the joint distribution of (X, Y). A standard way to define optimal classifiers is to introduce a class of joint distributions of (X, Y) and to declare $\hat{f}_n$ optimal if it achieves the best rate of convergence in a minimax sense on this class.

Two types of assumptions on the joint distribution of (X, Y) are commonly used: complexity assumptions and margin assumptions.

Complexity assumptions are stated in two possible ways. First of them is to suppose that the regression function η is smooth enough or, more generally, belongs to a class of functions Σ having a suitably bounded ε-entropy. This is called a *complexity assumption on the regression function* (CAR). Most commonly it is of the following form.

Assumption (CAR). *The regression function* η *belongs to class* Σ *of functions on* $\mathbb{R}^d$ *such that*

$$\mathcal{H}(\varepsilon, \Sigma, L_p) \le A_* \varepsilon^{-\rho}, \quad \forall \varepsilon > 0,$$

with some constants $\rho > 0$, $A_* > 0$. *Here* $\mathcal{H}(\varepsilon, \Sigma, L_p)$ *denotes the* ε*-entropy of the set* Σ *w.r.t. an* L_p *norm with some* $1 \le p \le \infty$.

At this stage of discussion we do not identify precisely the value of p for the L_p norm in Assumption (CAR), nor the measure with respect to which this norm is defined. Examples will be given later. If Σ is a class of smooth functions with smoothness parameter β on a compact in $\mathbb{R}^d$, for example, a Hölder class, as described below, a typical value of ρ in Assumption (CAR) is $\rho = d/\beta$.

Assumption (CAR) is well adapted for the study of *plug-in rules*, i.e. of the classifiers having the form

$$\hat{f}_n^{PI}(X) = \mathbb{I}_{\{\hat{\eta}_n(X) \ge \frac{1}{2}\}} \tag{1.1}$$

where $\hat{\eta}_n$ is a nonparametric estimator of the function η. Indeed, Assumption (CAR) typically reads as a smoothness assumption on η implying that a good nonparametric estimator (kernel, local polynomial, orthogonal series or other) $\hat{\eta}_n$ converges with some rate to the regression function η, as $n \to \infty$. In turn, closeness of $\hat{\eta}_n$ to η implies closeness of $\hat{f}_n$ to f: for any plug-in classifier $\hat{f}_n^{PI}$ we have

$$\mathbb{E}R(\hat{f}_n^{PI}) - R(f^*) \le 2\mathbb{E} \int |\hat{\eta}_n(x) - \eta(x)| P_X(dx) \tag{1.2}$$

(cf. Devroye, Györfi and Lugosi (1996), Theorem 2.2). For various types of estimators $\hat{\eta}_n$ and under rather general assumptions it can be shown that, if (CAR) holds, the RHS of (1.2) is uniformly of the order $n^{-1/(2+\rho)}$, and thus

$$\sup_{P:\eta \in \Sigma} \mathcal{E}(\hat{f}_n^{PI}) = O(n^{-1/(2+\rho)}), \quad n \to \infty, \tag{1.3}$$

[cf. Yang (1999)]. In particular, if $\rho = d/\beta$ (which corresponds to a class of smooth functions with smoothness parameter β), we get

$$\sup_{P:\eta \in \Sigma} \mathcal{E}(\hat{f}_n^{PI}) = O(n^{-\beta/(2\beta+d)}), \quad n \to \infty. \tag{1.4}$$

Note that (1.4) can be easily deduced from (1.2) and standard results on the L_1 or L_2 convergence rates of usual nonparametric regression estimators on β-smoothness classes Σ. The rates in (1.3), (1.4) are quite slow, always slower than $n^{-1/2}$. In (1.4) they deteriorate dramatically as the dimension d increases. Moreover, Yang (1999) shows that, under general assumptions, the bound (1.4) cannot be improved in a minimax sense. These results raised some pessimism about the plug-in rules.

The second way to describe complexity is to introduce a structure on the class of possible decision sets $G^* = \{x : f^*(x) = 1\} = \{x : \eta(x) \geq 1/2\}$ rather than on that of regression functions η. A standard *complexity assumption on the decision set* (CAD) is the following.

Assumption (CAD). *The decision set G^* belongs to a class $\mathcal{G}$ of subsets of $\mathbb{R}^d$ such that*

$$\mathcal{H}(\varepsilon, \mathcal{G}, d_\triangle) \leq A_* \varepsilon^{-\rho}, \quad \forall \varepsilon > 0,$$

with some constants $\rho > 0$, $A_* > 0$. Here $\mathcal{H}(\varepsilon, \mathcal{G}, d_\triangle)$ denotes the ε-entropy of the class $\mathcal{G}$ w.r.t. the measure of symmetric difference pseudo-distance between sets defined by $d_\triangle(G, G') = P_X(G \triangle G')$ for two measurable subsets G and G' in $\mathbb{R}^d$.

The parameter ρ in Assumption (CAD) typically characterizes the smoothness of the boundary of G^* [cf. Tsybakov (2004a)]. Note that, in general, there is no connection between Assumptions (CAR) and (CAD). Indeed, the fact that G^* has a smooth boundary does not imply that η is smooth, and vice versa. The values of ρ closer to 0 correspond to smoother boundaries (less complex sets G^*). As a limit case when $\rho \to 0$ one can consider the Vapnik-Chervonenkis classes (VC-classes) for which the ε-entropy is logarithmic in $1/\varepsilon$.

Assumption (CAD) is suited for the study of empirical risk minimization (ERM) type classifiers introduced by Vapnik and Chervonenkis (1974), see also Devroye, Györfi and Lugosi (1996), Vapnik (1998). As shown in Tsybakov (2004a), for every $0 < \rho < 1$ there exist ERM classifiers $\hat{f}_n^{ERM}$ such that, under Assumption (CAD),

$$\sup_{P: G^* \in \mathcal{G}} \mathcal{E}(\hat{f}_n^{ERM}) = O(n^{-1/2}), \quad n \to \infty. \tag{1.5}$$

The rate of convergence in (1.5) is better than that for plug-in rules, cf. (1.3) – (1.4), and it does not depend on ρ (respectively, on the dimension d). Note that the comparison between (1.5) and (1.3) – (1.4) is not quite legitimate, because there is no inclusion between classes of joint distributions P of (X, Y) satisfying Assumptions (CAR) and (CAD). Nevertheless, such a comparison have been often interpreted as an argument in disfavor of the plug-in rules. Indeed, Yang's lower bound shows that the $n^{-1/2}$ rate cannot be attained under Assumption (CAD) suited for the plug-in rules. Recently, advantages of the ERM type classifiers, including penalized ERM methods, have been further confirmed by the fact that, under the margin (or low noise) assumption, they can attain *fast rates of convergence*, i.e. the rates that are faster than $n^{-1/2}$ [Mammen and Tsybakov (1999), Tsybakov (2004a), Massart and Nédélec (2003), Tsybakov and van de Geer (2005), Koltchinskii (2005), Audibert (2004)].

The *margin assumption* (or low noise assumption) is stated as follows.

Assumption (MA). *There exist constants* $C_0 > 0$ *and* $\alpha \geq 0$ *such that*

$$P_X\big(0 < |\eta(X) - 1/2| \leq t\big) \leq C_0 t^{\alpha}, \qquad \forall\, t > 0. \tag{1.6}$$

The case $\alpha = 0$ is trivial (no assumption) and is included for notational convenience. Assumption (MA) provides a useful characterization of the behavior of regression function η in a vicinity of the level $\eta = 1/2$ which turns out to be crucial for convergence of classifiers (for more discussion of the margin assumption see Tsybakov (2004a)). The main point is that, under (MA), fast classification rates up to n^{-1} are achievable. In particular, for every $0 < \rho < 1$ and $\alpha > 0$ there exist ERM type classifiers $\hat{f}_n^{ERM}$ such that

$$\sup_{P:(CAD),(MA)} \mathcal{E}(\hat{f}_n^{ERM}) = O(n^{-\frac{1+\alpha}{2+\alpha+\alpha\rho}}), \quad n \to \infty, \tag{1.7}$$

where $\sup_{P:(CAD),(MA)}$ denotes the supremum over all joint distributions P of (X, Y) satisfying Assumptions (CAD) and (MA). The RHS of (1.7) can be arbitrarily close to $O(n^{-1})$ for large α and small ρ. Result (1.7) for direct ERM classifiers on ε-nets is proved by Tsybakov (2004a), and for some other ERM type classifiers by Tsybakov and van de Geer (2005), Koltchinskii (2005) and Audibert (2004) (in some of these papers the rate of convergence (1.7) is obtained with an extra log-factor).

Comparison of (1.5) and (1.7) with (1.2) and (1.3) seems to confirm the conjecture that the plug-in classifiers are inferior to the ERM type ones. The main message of the present paper is to disprove this conjecture. We will show that there exist plug-in rules that converge with fast rates, and even with *super-fast rates*, i.e. faster than n^{-1} under the margin assumption (MA). The basic idea of the proof is to use exponential inequalities for the regression estimator $\hat{\eta}_n$ (see Section 3 below) or the convergence results in the L_∞ norm (see Section 5), rather than the usual

L_1 or L_2 norm convergence of $\hat{\eta}_n$, as previously described (cf. (1.2)). We do not know whether the super-fast rates are attainable for ERM rules or, more precisely, under Assumption (CAD) which serves for the study of the ERM type rules. It is important to note that our results on fast rates cover more general setting than just classification with plug-in rules. These are rather results about *classification in the regression complexity context under the margin assumption.* In particular, we establish minimax lower bounds valid for all classifiers, and we construct a "hybrid" plug-in/ ERM procedure (ERM based on a grid on a set regression functions η) that achieves optimality. Thus, the point is mainly not about the type of procedure (plug-in or ERM) but about the type of complexity assumption (on the regression function (CAR) or on the decision set (CAD)) that should be natural to impose. Assumption (CAR) on the regression function arises in a natural way in the analysis of several practical procedures of plug-in type, such as boosting and SVM [cf. Blanchard, Lugosi and Vayatis (2003), Bartlett, Jordan and McAuliffe (2003), Scovel and Steinwart (2003), Blanchard, Bousquet and Massart (2004), Tarigan and van de Geer (2004)]. These procedures are now intensively studied but, to our knowledge, only suboptimal rates of convergence have been proved in the regression complexity context under the margin assumption. The results in Section 4 point out this fact (see also Section 5), and establish the best achievable rates of classification that those procedures should expectedly attain.

2 Notation and definitions

In this section we introduce some notation, definitions and basic facts that will be used in the paper.

We denote by $C, C_1, C_2, \dots$ positive constants whose values may differ from line to line. The symbols $\mathbb{P}$ and $\mathbb{E}$ stand for generic probability and expectation signs, and E_X is the expectation w.r.t. the marginal distribution P_X. We denote by $\mathcal{B}(x, r)$ the closed Euclidean ball in $\mathbb{R}^d$ centered at $x \in \mathbb{R}^d$ and of radius $r > 0$.

For any multi-index $s = (s_1, \dots, s_d) \in \mathbb{N}^d$ and any $x = (x_1, \dots, x_d) \in \mathbb{R}^d$, we define $|s| = \sum_{i=1}^d s_i$, $s! = s_1! \dots s_d!$, $x^s = x_1^{s_1} \dots x_d^{s_d}$ and $\|x\| \triangleq (x_1^2 + \dots + x_d^2)^{1/2}$. Let D^s denote the differential operator $D^s \triangleq \frac{\partial^{s_1+\dots+s_d}}{\partial x_1^{s_1} \cdots \partial x_d^{s_d}}$.

Let $\beta > 0$. Denote by $\lfloor \beta \rfloor$ the maximal integer that is strictly less than β. For any $x \in \mathbb{R}^d$ and any $\lfloor\beta\rfloor$ times continuously differentiable real valued function g on $\mathbb{R}^d$, we denote by g_x its Taylor polynomial of degree $\lfloor\beta\rfloor$ at point x:

$$g_x(x') \triangleq \sum_{|s|\le\lfloor\beta\rfloor} \frac{(x'-x)^s}{s!} D^s g(x).$$

Let $L > 0$. The $\big(\beta, L, \mathbb{R}^d\big)$-*Hölder class* of functions, denoted $\Sigma(\beta, L, \mathbb{R}^d)$, is defined as the set of functions $g : \mathbb{R}^d \to \mathbb{R}$ that are $\lfloor\beta\rfloor$ times continuously differentiable and satisfy, for any $x, x' \in \mathbb{R}^d$, the inequality

$$|g(x') - g_x(x')| \le L\|x - x'\|^\beta.$$

Fix some constants $c_0, r_0 > 0$. We will say that a Lebesgue measurable set $A \subset \mathbb{R}^d$ is (c_0, r_0)-*regular* if

$$\lambda\big[A \cap \mathcal{B}(x, r)\big] \ge c_0 \lambda\big[\mathcal{B}(x, r)\big], \quad \forall\, 0 < r \le r_0, \quad \forall\, x \in A, \tag{2.1}$$

where $\lambda[S]$ stands for the Lebesgue measure of $S \subset \mathbb{R}^d$. To illustrate this definition, consider the following example. Let $d \ge 2$. Then the set $A = \big\{x = (x_1, \dots, x_d) \in \mathbb{R}^d : \sum_{j=1}^d |x_j|^q \le 1\big\}$ is (c_0, r_0)-regular with some $c_0, r_0 > 0$ for $q \ge 1$, and there are no $c_0, r_0 > 0$ such that A is (c_0, r_0)-regular for $0 < q < 1$.

Introduce now two assumptions on the marginal distribution P_X that will be used in the sequel.

Definition 2.1 *Fix* $0 < c_0, r_0, \mu_{\max} < \infty$ *and a compact* $\mathcal{C} \subset \mathbb{R}^d$*. We say that the* **mild density assumption** *is satisfied if the marginal distribution* P_X *is supported on a compact* (c_0, r_0)*-regular set* $A \subseteq \mathcal{C}$ *and has a uniformly bounded density* μ *w.r.t. the Lebesgue measure:* $\mu(x) \le \mu_{\max}, \forall\, x \in A$.

Definition 2.2 *Fix some constants* $c_0, r_0 > 0$ *and* $0 < \mu_{\min} < \mu_{\max} < \infty$ *and a compact* $\mathcal{C} \subset \mathbb{R}^d$*. We say that the* **strong density assumption** *is satisfied if the marginal distribution* P_X *is supported on a compact* (c_0, r_0)*-regular set* $A \subseteq \mathcal{C}$ *and has a density* μ *w.r.t. the Lebesgue measure bounded away from zero and infinity on* A*:*

$$\mu_{\min} \le \mu(x) \le \mu_{\max} \quad \textit{for} \;\; x \in A, \quad \textit{and} \;\; \mu(x) = 0 \quad \textit{otherwise.}$$

We finally recall some notions related to locally polynomial estimators.

Definition 2.3 *For $h > 0$, $x \in \mathbb{R}^d$, for an integer $l \geq 0$ and a function $K : \mathbb{R}^d \to \mathbb{R}_+$, denote by $\hat{\theta}_x$ a polynomial on $\mathbb{R}^d$ of degree l which minimizes*

$$\sum_{i=1}^{n} \left[Y_i - \hat{\theta}_x(X_i - x)\right]^2 K\left(\frac{X_i - x}{h}\right). \tag{2.2}$$

The **locally polynomial estimator $\hat{\eta}_n^{LP}(x)$ of order** *l, or LP(l) estimator, of the value $\eta(x)$ of the regression function at point x is defined by: $\hat{\eta}_n^{LP}(x) \triangleq \hat{\theta}_x(0)$ if $\hat{\theta}_x$ is the unique minimizer of (2.2) and $\hat{\eta}_n^{LP}(x) \triangleq 0$ otherwise. The value h is called the bandwidth and the function K is called the kernel of the LP(l) estimator.*

Let T_s denote the coefficients of $\hat{\theta}_x$ indexed by multi-index $s \in \mathbb{N}^d$: $\hat{\theta}_x(u) = \sum_{|s| \leq l} T_s u^s$. Introduce the vectors $T \triangleq \left(T_s\right)_{|s| \leq l}$, $V \triangleq \left(V_s\right)_{|s| \leq l}$ where

$$V_s \triangleq \textstyle\sum_{i=1}^n Y_i (X_i - x)^s K\left(\frac{X_i - x}{h}\right), \tag{2.3}$$

$U(u) \triangleq \left(u^s\right)_{|s| \leq l}$ and the matrix $Q \triangleq \left(Q_{s_1, s_2}\right)_{|s_1|, |s_2| \leq l}$ where

$$Q_{s_1, s_2} \triangleq \textstyle\sum_{i=1}^n (X_i - x)^{s_1 + s_2} K\left(\frac{X_i - x}{h}\right). \tag{2.4}$$

The following result is straightforward (cf. Section 1.7 in Tsybakov (2004b) where the case $d = 1$ is considered).

Proposition 2.1 *If the matrix Q is positive definite, there exists a unique polynomial on $\mathbb{R}^d$ of degree l minimizing (2.2). Its vector of coefficients is given by $T = Q^{-1}V$ and the corresponding LP(l) regression function estimator has the form*

$$\hat{\eta}_n^{LP}(x) = U^T(0) Q^{-1} V = \sum_{i=1}^{n} Y_i K\left(\frac{X_i - x}{h}\right) U^T(0) Q^{-1} U(X_i - x).$$

3 Fast rates for plug-in rules: the strong density assumption

We first state a general result showing how the rates of convergence of plug-in classifiers can be deduced from exponential inequalities for the corresponding regression estimators.

In the sequel, for an estimator $\hat{\eta}_n$ of η, we write

$$\mathbb{P}\left(\left|\hat{\eta}_n(X) - \eta(X)\right| \geq \delta\right) \triangleq \int P^{\otimes n}\left(\left|\hat{\eta}_n(x) - \eta(x)\right| \geq \delta\right) P_X(dx), \quad \forall\, \delta > 0,$$

i.e., we consider the probability taken with respect to the distribution of the sample $(X_1, Y_1, \dots X_n, Y_n)$ and the distribution of the input X.

Theorem 3.1 *Let $\hat{\eta}_n$ be an estimator of the regression function η and $\mathcal{P}$ a set of probability distributions on $\mathcal{Z}$ such that for some constants $C_1 > 0$, $C_2 > 0$, for some positive sequence a_n, for $n \geq 1$ and any $\delta > 0$, and for almost all x w.r.t. P_X, we have*

$$\sup_{P \in \mathcal{P}} P^{\otimes n}\Big(|\hat{\eta}_n(x) - \eta(x)| \geq \delta\Big) \leq C_1 \exp\big(-C_2 a_n \delta^2\big). \tag{3.1}$$

Consider the plug-in classifier $\hat{f}_n = \mathrm{1\!I}_{\{\hat{\eta}_n \geq \frac{1}{2}\}}$. If all the distributions $P \in \mathcal{P}$ satisfy the margin assumption (MA), we have

$$\sup_{P \in \mathcal{P}} \Big\{\mathbb{E}R(\hat{f}_n) - R(f^*)\Big\} \leq C a_n^{-\frac{1+\alpha}{2}}$$

for $n \geq 1$ with some constant $C > 0$ depending only on α, C_0, C_1 and C_2.

Proof. Consider the sets $A_j \subset \mathbb{R}^d, j = 1, 2, \dots,$ defined as

$$\begin{aligned} A_0 &\triangleq \big\{x \in \mathbb{R}^d : 0 < |\eta(x) - \tfrac{1}{2}| \leq \delta\big\}, \\ A_j &\triangleq \big\{x \in \mathbb{R}^d : 2^{j-1}\delta < |\eta(x) - \tfrac{1}{2}| \leq 2^j \delta\big\}, \qquad \text{for } j \geq 1. \end{aligned}$$

For any $\delta > 0$, we may write

$$\begin{aligned} \mathbb{E}R(\hat{f}_n) - R(f^*) &= \mathbb{E}\big(|2\eta(X) - 1| \mathrm{1\!I}_{\{\hat{f}_n(X) \neq f^*(X)\}}\big) \\ &= \textstyle\sum_{j=0}^{\infty} \mathbb{E}\big(|2\eta(X) - 1| \mathrm{1\!I}_{\{\hat{f}_n(X) \neq f^*(X)\}} \mathrm{1\!I}_{\{X \in A_j\}}\big) \\ &\leq 2\delta P_X\big(0 < |\eta(X) - \tfrac{1}{2}| \leq \delta\big) \\ &\quad + \textstyle\sum_{j \geq 1} \mathbb{E}\big(|2\eta(X) - 1| \mathrm{1\!I}_{\{\hat{f}_n(X) \neq f^*(X)\}} \mathrm{1\!I}_{\{X \in A_j\}}\big). \end{aligned} \tag{3.2}$$

On the event $\{\hat{f}_n \neq f^*\}$ we have $|\eta - \frac{1}{2}| \leq |\hat{\eta}_n - \eta|$. So, for any $j \geq 1$, we get

$$\begin{aligned} &\mathbb{E}\big(|2\eta(X) - 1| \mathrm{1\!I}_{\{\hat{f}_n(X) \neq f^*(X)\}} \mathrm{1\!I}_{\{X \in A_j\}}\big) \\ &\quad \leq 2^{j+1}\delta\, \mathbb{E}\big[\mathrm{1\!I}_{\{|\hat{\eta}_n(X) - \eta(X)| \geq 2^{j-1}\delta\}} \mathrm{1\!I}_{\{0 < |\eta(X) - \frac{1}{2}| \leq 2^j \delta\}}\big] \\ &\quad \leq 2^{j+1}\delta\, E_X\Big[P^{\otimes n}\big(|\hat{\eta}_n(X) - \eta(X)| \geq 2^{j-1}\delta\big) \mathrm{1\!I}_{\{0 < |\eta(X) - \frac{1}{2}| \leq 2^j \delta\}}\Big] \\ &\quad \leq C_1 2^{j+1}\delta \exp\Big(-C_2 a_n (2^{j-1}\delta)^2\Big) P_X\Big(0 < |\eta(X) - \tfrac{1}{2}| \leq 2^j \delta\Big) \\ &\quad \leq 2C_1 C_0 2^{j(1+\alpha)} \delta^{1+\alpha} \exp\big(-C_2 a_n (2^{j-1}\delta)^2\big) \end{aligned}$$

where in the last inequality we used Assumption (MA). Now, from inequality (3.2), taking $\delta = a_n^{-1/2}$ and using Assumption (MA) to bound the first term of the right hand side of (3.2), we get

$$\begin{aligned} \mathbb{E}R(\hat{f}_n) - R(f^*) &\leq 2C_0 a_n^{-\frac{1+\alpha}{2}} + C a_n^{-\frac{1+\alpha}{2}} \textstyle\sum_{j\geq 2} 2^{j(1+\alpha)} \exp\big(-C_2 2^{2j-2}\big) \\ &\leq C a_n^{-\frac{1+\alpha}{2}}. \end{aligned}$$

■

Inequality (3.1) is crucial to obtain the above result. This inequality holds true for various types of estimators and various sets of probability distributions $\mathcal{P}$. Here we focus on a standard case where η belongs to the Hölder class $\Sigma(\beta, L, \mathbb{R}^d)$ and the marginal law of X satisfies the strong density assumption. We are going to show that in this case there exist estimators satisfying inequality (3.1) with $a_n = n^{\frac{2\beta}{2\beta+d}}$. These can be, for example, locally polynomial estimators. Specifically, assume from now on that K is a kernel satisfying

$$\exists c > 0 : \quad K(x) \geq c \mathbb{I}_{\{\|x\| \leq c\}}, \quad \forall x \in \mathbb{R}^d, \tag{3.3}$$

$$\int_{\mathbb{R}^d} K(u) du = 1, \tag{3.4}$$

$$\int_{\mathbb{R}^d} \big(1 + \|u\|^{4\beta}\big) K^2(u) du < \infty, \tag{3.5}$$

$$\sup_{u \in \mathbb{R}^d} \big(1 + \|u\|^{2\beta}\big) K(u) < \infty. \tag{3.6}$$

Let $h > 0$, and consider the matrix $\bar{B} \triangleq \big(\bar{B}_{s_1,s_2}\big)_{|s_1|,|s_2| \leq \lfloor \beta \rfloor}$ where $\bar{B}_{s_1,s_2} = \frac{1}{nh^d} \sum_{i=1}^n \left(\frac{X_i - x}{h}\right)^{s_1+s_2} K\left(\frac{X_i - x}{h}\right)$. Define the regression function estimator $\hat{\eta}_n^*$ as follows. If the smallest eigenvalue of the matrix $\bar{B}$ is greater than $(\log n)^{-1}$ we set $\hat{\eta}_n^*(x)$ equal to the projection of $\hat{\eta}_n^{LP}(x)$ on the interval $[0,1]$, where $\hat{\eta}_n^{LP}(x)$ is the LP($\lfloor \beta \rfloor$) estimator with a bandwidth $h > 0$ and a kernel K satisfying (3.3) – (3.6). If the smallest eigenvalue of $\bar{B}$ is less than $(\log n)^{-1}$ we set $\hat{\eta}_n^*(x) = 0$.

Theorem 3.2 *Let $\mathcal{P}$ be a class of probability distributions P on $\mathcal{Z}$ such that the regression function η belongs to the Hölder class $\Sigma(\beta, L, \mathbb{R}^d)$ and the marginal law of X satisfies the strong density assumption. Then there exist constants $C_1, C_2, C_3 > 0$ such that for any $0 < h \leq r_0/c$, any $C_3 h^\beta < \delta$ and any $n \geq 1$ the estimator $\hat{\eta}_n^*$ satisfies*

$$\sup_{P \in \mathcal{P}} P^{\otimes n}\Big(\big|\hat{\eta}_n^*(x) - \eta(x)\big| \geq \delta\Big) \leq C_1 \exp\big(-C_2 n h^d \delta^2\big) \tag{3.7}$$

for almost all x w.r.t. P_X. As a consequence, there exist $C_1, C_2 > 0$ such that for $h = n^{-\frac{1}{2\beta+d}}$ and any $\delta > 0$, $n \geq 1$ we have

$$\sup_{P \in \mathcal{P}} P^{\otimes n}\Big(\big|\hat{\eta}_n^*(x) - \eta(x)\big| \geq \delta\Big) \leq C_1 \exp\big(-C_2 n^{\frac{2\beta}{2\beta+d}} \delta^2\big) \tag{3.8}$$

for almost all x w.r.t. P_X. The constants C_1, C_2, C_3 depend only on β, d, L, c_0, r_0, $\mu_{\min}, \mu_{\max}$, and on the kernel K.

Proof. See Section 6.1. ■

Remark 3.1 *We have chosen here the LP estimators of η because for them the exponential inequality* (3.1) *holds without additional smoothness conditions on the marginal density of X. For other popular regression estimators, such as kernel or orthogonal series ones, similar inequality can be also proved if we assume that the marginal density of X is as smooth as the regression function.*

Definition 3.1 *For a fixed parameter $\alpha \geq 0$, fixed positive parameters $c_0, r_0, C_0, \beta, L, \mu_{\max} > \mu_{\min} > 0$ and a fixed compact $\mathcal{C} \subset \mathbb{R}^d$, let $\mathcal{P}_\Sigma$ denote the class of all probability distributions P on $\mathcal{Z}$ such that*

(i) the margin assumption (MA) is satisfied,

(ii) the regression function η belongs to the Hölder class $\Sigma(\beta, L, \mathbb{R}^d)$,

(iii) the strong density assumption on P_X is satisfied.

Theorem 3.1 and (3.8) immediately imply the next result.

Theorem 3.3 *For any $n \geq 1$ the excess risk of the plug-in classifier $\hat{f}_n^* = \mathbb{I}_{\{\hat{\eta}_n^* \geq \frac{1}{2}\}}$ with bandwidth $h = n^{-\frac{1}{2\beta+d}}$ satisfies*

$$\sup_{P \in \mathcal{P}_\Sigma} \left\{ \mathbb{E}R(\hat{f}_n^*) - R(f^*) \right\} \leq C n^{-\frac{\beta(1+\alpha)}{2\beta+d}}$$

where the constant $C > 0$ depends only on α, C_0, C_1 and C_2.

For $\alpha\beta > d/2$ the convergence rate $n^{-\frac{\beta(1+\alpha)}{2\beta+d}}$ obtained in Theorem 3.3 is a *fast rate*, i.e., it is faster than $n^{-1/2}$. Furthermore, it is a *super-fast rate* (i.e., is faster than n^{-1}) for $(\alpha - 1)\beta > d$. We must note that if this condition is satisfied, the class $\mathcal{P}_\Sigma$ is rather poor, and thus super-fast rates can occur only for very particular joint distributions of (X, Y). Intuitively, this is clear. Indeed, to have a very smooth regression function η (i.e., very large β) implies that when η hits the level 1/2, it cannot "take off" from this level too abruptly. As a consequence, when the density of the distribution P_X is bounded away from 0 at a vicinity of the hitting point, the margin assumption cannot be satisfied for large α since this assumption puts an upper bound on the "time spent" by the regression function near 1/2. So, α and β cannot be simultaneously very large. It can be shown that the cases of "too large" and "not too large" (α, β) are essentially described by the condition $(\alpha - 1)\beta > d$.

To be more precise, observe first that $\mathcal{P}_\Sigma$ is not empty for $(\alpha - 1)\beta > d$, so that the super-fast rates can effectively occur. Examples of laws $P \in \mathcal{P}_\Sigma$ under this condition can be easily given, such as the one with P_X equal to the uniform distribution on a ball centered at 0 in $\mathbb{R}^d$, and the regression function defined by $\eta(x) = 1/2 - C\|x\|^2$ with an appropriate $C > 0$. Clearly, η belongs to Hölder classes with arbitrarily large β and Assumption (MA) is satisfied with $\alpha = d/2$. Thus, for $d \geq 3$ and β large enough super-fast rates can occur. Note that in this example the decision set $\{x : \eta(x) \geq 1/2\}$ has the Lebesgue measure 0 in $\mathbb{R}^d$. It turns out that this condition is necessary to achieve classification with super-fast rates when the Hölder classes of regression functions are considered.

To explain this and to have further insight into the problem of super-fast rates, consider the following two questions:

- for which parameters α, β and d is there a distribution $P \in \mathcal{P}_\Sigma$ such that the regression function associated with P hits[1] 1/2 in the support of P_X?
- for which parameters α, β and d is there a distribution $P \in \mathcal{P}_\Sigma$ such that the regression function associated with P crosses[2] 1/2 in the interior of the support of P_X?

The following result gives a precise description of the constraints on (α, β) leading to possibility or impossibility of the super-fast rates.

Proposition 3.4
- *If $\alpha(1 \wedge \beta) > d$, there is no distribution $P \in \mathcal{P}_\Sigma$ such that the regression function η associated with P hits 1/2 in the interior of the support of P_X.*
- *For any $\alpha, \beta > 0$ and integer $d \geq \alpha(1 \wedge \beta)$, any positive parameter L and any compact $\mathcal{C} \subset \mathbb{R}^d$ with non-empty interior, for appropriate positive parameters*

[1] A function $f : \mathbb{R}^d \to \mathbb{R}$ is said to hit the level $a \in \mathbb{R}$ at $x_0 \in \mathbb{R}^d$ if and only if $f(x_0) = a$ and for any $r > 0$ there exists $x \in \mathcal{B}(x_0, r)$ such that $f(x) \neq a$.

[2] A function $f : \mathbb{R}^d \to \mathbb{R}$ is said to cross the level $a \in \mathbb{R}$ at $x_0 \in \mathbb{R}^d$ if and only if for any $r > 0$, there exists x_- and x_+ in $\mathcal{B}(x_0, r)$ such that $f(x_-) < a$ and $f(x_+) > a$.

$C_0, c_0, r_0, \mu_{\max} > \mu_{\min} > 0$, there are distributions $P \in \mathcal{P}_\Sigma$ such that the regression function η associated with P hits $1/2$ in the boundary of the support of P_X.

- *For any $\alpha, \beta > 0$, any integer $d \geq 2\alpha$, any positive parameter L and any compact $\mathcal{C} \subset \mathbb{R}^d$ with non-empty interior, for appropriate positive parameters $C_0, c_0, r_0, \mu_{\max} > \mu_{\min} > 0$, there are distributions $P \in \mathcal{P}_\Sigma$ such that the regression function η associated with P hits $1/2$ in the interior of the support of P_X.*

- *If $\alpha(1 \wedge \beta) > 1$ there is no distribution $P \in \mathcal{P}_\Sigma$ such that the regression function η associated with P crosses $1/2$ in the interior of the support of P_X. Conversely, for any $\alpha, \beta > 0$ such that $\alpha(1 \wedge \beta) \leq 1$, any integer d, any positive parameter L and any compact $\mathcal{C} \subset \mathbb{R}^d$ with non-empty interior, for appropriate positive parameters $C_0, c_0, r_0, \mu_{\max} > \mu_{\min} > 0$, there are distributions $P \in \mathcal{P}_\Sigma$ such that the regression function η associated with P crosses $1/2$ in the interior of the support of P_X.*

Note that the condition $\alpha(1 \wedge \beta) > 1$ appearing in the last assertion is equivalent to $\frac{\beta(1+\alpha)}{2\beta+d} > \frac{(2\beta) \vee (\beta+1)}{2\beta+d}$, which is necessary to have super-fast rates. As a consequence, in this context, super-fast rates cannot occur when the regression function crosses $1/2$ in the interior of the support. The third assertion of the proposition shows that super-fast rates can occur with regression functions hitting $1/2$ in the interior of the support of P_X provided that the regression function is highly smooth and defined on a highly dimensional space and that a strong margin assumption holds (i.e. α large).

Proof. See Section 6.3. ■

The following lower bound shows optimality of the rate of convergence for the Hölder classes obtained in Theorem 3.3.

Theorem 3.5 *Let $d \geq 1$ be an integer, and let L, β, α be positive constants, such that $\alpha\beta \leq d$. Then there exists a constant $C > 0$ such that for any $n \geq 1$ and any classifier $\hat{f}_n : \mathcal{Z}^n \to \mathcal{F}$, we have*

$$\sup_{P \in \mathcal{P}_\Sigma} \left\{ \mathbb{E}R(\hat{f}_n) - R(f^*) \right\} \geq C n^{-\frac{\beta(1+\alpha)}{2\beta+d}}.$$

Proof. See Section 6.2. ■

Note that the lower bound of Theorem 3.5 does not cover the case of super-fast rates ($(\alpha - 1)\beta > d$).

Finally, we discuss the case where "$\alpha = \infty$", which means that there exists $t_0 > 0$ such that

$$P_X\big(0 < |\eta(X) - 1/2| \le t_0\big) = 0. \tag{3.9}$$

This is a very favorable situation for classification. The rates of convergence of the ERM type classifiers under (3.9) are, of course, faster than under Assumption (MA) with $\alpha < \infty$ [cf. Massart and Nédélec (2003)], but they are not faster than n^{-1}. Indeed, Massart and Nédélec (2003) provide a lower bound showing that, even if Assumption (CAD) is replaced by a very strong assumption that the true decision set belongs to a VC-class (note that both assumptions are naturally linked to the study the ERM type classifiers), the best achievable rate is of the order $(\log n)/n$. We show now that for the plug-in classifiers much faster rates can be attained. Specifically, if the regression function η has some (arbitrarily low) Hölder smoothness β the rate of convergence can be exponential in n. To show this, we first state a simple lemma which is valid for any plug-in classifier $\hat{f}_n$.

Lemma 3.6 *Let assumption (3.9) be satisfied, and let $\hat{\eta}_n$ be an estimator of the regression function η. Then for the plug-in classifier $\hat{f}_n = \mathbb{1}_{\{\hat{\eta}_n \ge \frac{1}{2}\}}$ we have*

$$\mathbb{E}R(\hat{f}_n) - R(f^*) \le \mathbb{P}\big(|\hat{\eta}_n(X) - \eta(X)| > t_0\big).$$

Proof. Following the argument similar to the proof of Theorem 3.1 and using condition (3.9) we get

$$\begin{aligned}
\mathbb{E}R(\hat{f}_n) - R(f^*) &\le 2t_0 P_X\big(0 < |\eta(X) - 1/2| \le t_0\big) \\
&\qquad + \mathbb{E}\big(|2\eta(X) - 1|\mathbb{1}_{\{\hat{f}_n(X) \ne f^*(X)\}}\mathbb{1}_{\{|\eta(X)-1/2|>t_0\}}\big) \\
&= \mathbb{E}\big(|2\eta(X) - 1|\mathbb{1}_{\{\hat{f}_n(X) \ne f^*(X)\}}\mathbb{1}_{\{|\eta(X)-1/2|>t_0\}}\big) \\
&\le \mathbb{P}\big(|\hat{\eta}_n(X) - \eta(X)| > t_0\big).
\end{aligned}$$

■

Lemma 3.6 and Theorem 3.2 immediately imply that, under assumption (3.9), the rate of convergence of the plug-in classifier $\hat{f}_n^* = \mathbb{1}_{\{\hat{\eta}_n^* \ge \frac{1}{2}\}}$ with a small enough fixed (independent of n) bandwidth h is exponential. To state the result, we denote by $\mathcal{P}_{\Sigma,\infty}$ the class of probability distributions P defined in the same way as $\mathcal{P}_\Sigma$, with the only difference that in Definition 3.1 the margin assumption (MA) is replaced by condition (3.9).

Proposition 3.7 *There exists a fixed (independent of n) $h > 0$ such that for any $n \geq 1$ the excess risk of the plug-in classifier $\hat{f}_n^* = \mathbb{I}_{\{\hat{\eta}_n^* \geq \frac{1}{2}\}}$ with bandwidth h satisfies*

$$\sup_{P \in \mathcal{P}_{\Sigma,\infty}} \left\{ \mathbb{E}R(\hat{f}_n^*) - R(f^*) \right\} \leq C_4 \exp(-C_5 n)$$

where the constants $C_4, C_5 > 0$ depend only on t_0, β, d, L, c_0, r_0, $\mu_{\min}, \mu_{\max}$, and on the kernel K.

Proof. Use Lemma 3.6, choose $h > 0$ such that $h < \min(r_0/c, (t_0/C_3)^{1/\beta})$, and apply (3.7) with $\delta = t_0$. ■

Koltchinskii and Beznosova (2005) prove a result on exponential rates for the plug-in classifier with some penalized regression estimator in place of the locally polynomial one that we use here. Their result is stated under a less general condition, in the sense that they consider only the Lipschitz class of regression functions η, while in Proposition 3.7 the Hölder smoothness β can be arbitrarily close to 0. Note also that we do not impose any complexity assumption on the decision set. However, the class of distributions $\mathcal{P}_{\Sigma,\infty}$ is quite restricted in a different sense. Indeed, for such distributions condition (3.9) should be compatible with the assumption that η belongs to a Hölder class. A sufficient condition for that is the existence of a band or a "corridor" of zero P_X-measure separating the sets $\{x : \eta(x) > 1/2\}$ and $\{x : \eta(x) < 1/2\}$. We believe that this condition is close to the necessary one.

4 Optimal learning rates without the strong density assumption

In this section we show that if P_X does not admit a density bounded away from zero on its support the rates of classification are slower than those obtained in Section 3. In particular, super-fast rates, i.e., the rates faster than n^{-1}, cannot be achieved. Introduce the following class of probability distributions.

Definition 4.1 *For a fixed parameter $\alpha \geq 0$, fixed positive parameters $c_0, r_0, C_0, \beta, L, \mu_{\max} > 0$ and a fixed compact $\mathcal{C} \subset \mathbb{R}^d$, let $\mathcal{P}'_\Sigma$ denote the class of all probability distributions P on $\mathcal{Z}$ such that* *(i) the margin assumption (MA) is satisfied,*

(ii) the regression function η belongs to the Hölder class $\Sigma(\beta, L, \mathbb{R}^d)$,

(iii) the mild density assumption on P_X is satisfied.

In this section we mainly assume that the distribution P of (X, Y) belongs to $\mathcal{P}'_\Sigma$, but we also consider larger classes of distributions satisfying the margin assumption (MA) and the complexity assumption (CAR).

Clearly, $\mathcal{P}_\Sigma \subset \mathcal{P}'_\Sigma$. The only difference between $\mathcal{P}'_\Sigma$ and $\mathcal{P}_\Sigma$ is that for $\mathcal{P}'_\Sigma$ the marginal density of X is not bounded away from zero. The optimal rates for $\mathcal{P}'_\Sigma$ are slower than for $\mathcal{P}_\Sigma$. Indeed, we have the following lower bound for the excess risk.

Theorem 4.1 *Let $d \geq 1$ be an integer, and let L, β, α be positive constants. Then there exists a constant $C > 0$ such that for any $n \geq 1$ and any classifier $\hat{f}_n : \mathcal{Z}^n \to \mathcal{F}$ we have*

$$\sup_{P \in \mathcal{P}'_\Sigma} \left\{ \mathbb{E} R(\hat{f}_n) - R(f^*) \right\} \geq C n^{-\frac{(1+\alpha)\beta}{(2+\alpha)\beta + d}}.$$

Proof. See Section 6.2. ■

In particular, when $\alpha = d/\beta$, we get slow convergence rate $1/\sqrt{n}$, instead of the fast rate $n^{-\frac{\beta+d}{2\beta+d}}$ obtained in Theorem 3.3 under the strong density assumption. Nevertheless, the lower bound can still approach n^{-1}, as the margin parameter α tends to ∞.

We now show that the rate of convergence given in Theorem 4.1 is optimal in the sense that there exist estimators that achieve this rate. This will be obtained as a consequence of a general upper bound for the excess risk of classifiers over a larger set $\mathcal{P}$ of distributions than $\mathcal{P}'_\Sigma$.

Fix a Lebesgue measurable set $\mathcal{C} \subset \mathbb{R}^d$ and a value $1 \leq p \leq \infty$. Let Σ be a class of regression functions η on $\mathbb{R}^d$ such that Assumption (CAR) is satisfied where the ε-entropy is taken w.r.t. the $L_p(\mathcal{C}, \lambda)$ norm (λ is the Lebesgue measure on $\mathbb{R}^d$). Then for every $\varepsilon > 0$ there exists an ε-net $\mathcal{N}_\varepsilon$ on Σ w.r.t. this norm such that

$$\log \left(\operatorname{card} \mathcal{N}_\varepsilon \right) \leq A' \varepsilon^{-\rho},$$

where A' is a constant. Consider the empirical risk

$$R_n(f) = \frac{1}{n} \sum_{i=1}^{n} \mathrm{I\!I}_{\{f(X_i) \neq Y_i\}}, \qquad f \in \mathcal{F},$$

and set

$$\varepsilon_n = \varepsilon_n(\alpha, \rho, p) \triangleq \begin{cases} n^{-\frac{1}{2+\alpha+\rho}} & \text{if} \quad p = \infty, \\ n^{-\frac{p+\alpha}{(2+\alpha)p+\rho(p+\alpha)}} & \text{if} \quad 1 \le p < \infty. \end{cases}$$

Define a sieve estimator $\hat{\eta}_n^S$ of the regression function η by the relation

$$\hat{\eta}_n^S \in \text{Argmin}_{\bar{\eta} \in \mathcal{N}_{\varepsilon_n}} R_n(f_{\bar{\eta}}) \tag{4.1}$$

where $f_{\bar{\eta}}(x) = \mathbb{I}_{\{\bar{\eta}(x) \ge 1/2\}}$, and consider the classifier $\hat{f}_n^S = \mathbb{I}_{\{\hat{\eta}_n^S \ge 1/2\}}$. Note that $\hat{f}_n^S$ can be viewed as a "hybrid" plug-in/ ERM procedure: the ERM is performed on a set of plug-in rules corresponding to a grid on the class of regression functions η.

Theorem 4.2 *Let $\mathcal{P}$ be a set of probability distributions P on $\mathcal{Z}$ such that*

(i) the margin assumption (MA) is satisfied,

(ii) the regression function η belongs to a class Σ which satisfies the complexity assumption (CAR) with the ε-entropy taken w.r.t. the $L_p(\mathcal{C}, \lambda)$ norm for some $1 \le p \le \infty$,

(iii) for all $P \in \mathcal{P}$ the supports of marginal distributions P_X are included in $\mathcal{C}$.

Consider the classifier $\hat{f}_n^S = \mathbb{I}_{\{\hat{\eta}_n^S \ge 1/2\}}$. If $p = \infty$ for any $n \ge 1$ we have

$$\sup_{P \in \mathcal{P}} \Big\{ \mathbb{E} R(\hat{f}_n^S) - R(f^*) \Big\} \le C n^{-\frac{1+\alpha}{2+\alpha+\rho}}. \tag{4.2}$$

If $1 \le p < \infty$ and, in addition, for all $P \in \mathcal{P}$ the marginal distributions P_X are absolutely continuous w.r.t. the Lebesgue measure and their densities are uniformly bounded from above by some constant $\mu_{\max} < \infty$, then for any $n \ge 1$ we have

$$\sup_{P \in \mathcal{P}} \Big\{ \mathbb{E} R(\hat{f}_n^S) - R(f^*) \Big\} \le C n^{-\frac{(1+\alpha)p}{(2+\alpha)p+\rho(p+\alpha)}}. \tag{4.3}$$

Proof. See Section 6.4. ■

Theorem 4.2 allows one to get fast classification rates without any density assumption on P_X. Namely, define the following class of distributions P of (X, Y).

Definition 4.2 *For fixed parameters $\alpha \ge 0$, $C_0 > 0, \beta > 0, L > 0$, and for a fixed compact $\mathcal{C} \subset \mathbb{R}^d$, let $\mathcal{P}_\Sigma^0$ denote the class of all probability distributions P on $\mathcal{Z}$ such that* *(i) the margin assumption (MA) is satisfied,*

(ii) the regression function η belongs to the Hölder class $\Sigma(\beta, L, \mathbb{R}^d)$,

(iii) for all $P \in \mathcal{P}$ the supports of marginal distributions P_X are included in $\mathcal{C}$.

If $\mathcal{C}$ is a compact the estimates of ε-entropies of Hölder classes $\Sigma(\beta, L, \mathbb{R}^d)$ in the $L_\infty(\mathcal{C}, \lambda)$ norm can be obtained from Kolmogorov and Tikhomorov (1961), and they yield Assumption (CAR) with $\rho = d/\beta$. Therefore, from (4.2) we easily get the following upper bound.

Theorem 4.3 *Let $d \geq 1$ be an integer, and let L, β, α be positive constants. For any $n \geq 1$ the classifier $\hat{f}_n^S = \mathbb{I}_{\{\hat{\eta}_n^S \geq 1/2\}}$ defined by (4.1) with $p = \infty$ satisfies*

$$\sup_{P \in \mathcal{P}_\Sigma^0} \left\{ \mathbb{E} R(\hat{f}_n^S) - R(f^*) \right\} \leq C n^{-\frac{(1+\alpha)\beta}{(2+\alpha)\beta + d}}$$

with some constant $C > 0$ depending only on α, β, d, L and C_0.

Since $\mathcal{P}'_\Sigma \subset \mathcal{P}_\Sigma^0$, Theorems 3.5 and 4.3 show that $n^{-\frac{(1+\alpha)\beta}{(2+\alpha)\beta + d}}$ is optimal rate of convergence of the excess risk on the class of distributions $\mathcal{P}_\Sigma^0$.

5 Comparison lemmas

In this section we give some useful inequalities between the risks of plug-in classifiers and the L_p risks of the corresponding regression estimators under the margin assumption (MA). These inequalities will be helpful in the proofs. They also illustrate a connection between the two complexity assumptions (CAR) and (CAD) defined in the Introduction and allow one to compare our study of plug-in estimators with that given by Yang (1999) who considered the case $\alpha = 0$ (no margin assumption), as well as with the developments in Bartlett, Jordan and McAuliffe (2003) and Blanchard, Lugosi and Vayatis (2003).

Throughout this section $\bar{\eta}$ is a Borel function on $\mathbb{R}^d$ and

$$\bar{f}(x) = \mathbb{I}_{\{\bar{\eta}(x) \geq 1/2\}}.$$

For $1 \leq p \leq \infty$ we denote by $\|\cdot\|_p$ the $L_p(\mathbb{R}^d, P_X)$ norm. We first state some comparison inequalities for the L_∞ norm.

Lemma 5.1 *For any distribution P of (X, Y) satisfying Assumption (MA) we have*

$$R(\bar{f}) - R(f^*) \leq 2C_0 \|\bar{\eta} - \eta\|_\infty^{1+\alpha}, \tag{5.1}$$

and

$$P_X\big(\bar{f}(X) \neq f^*(X),\, \eta(X) \neq 1/2\big) \leq C_0 \|\bar{\eta} - \eta\|_\infty^{\alpha}. \tag{5.2}$$

Proof. To show (5.1) note that

$$\begin{aligned} R(\bar{f}) - R(f^*) &= \mathbb{E}\big(|2\eta(X) - 1| \mathbb{1}_{\{\bar{f}(X) \neq f^*(X)\}}\big) \\ &\leq 2\mathbb{E}\big(|\eta(X) - \tfrac{1}{2}| \mathbb{1}_{0 < \{|\eta(X) - \frac{1}{2}| \leq |\eta(X) - \bar{\eta}(X)|\}}\big) \\ &\leq 2\|\eta - \bar{\eta}\|_\infty P_X\big(0 < |\eta(X) - \tfrac{1}{2}| \leq \|\eta - \bar{\eta}\|_\infty\big) \\ &\leq 2C_0 \|\eta - \bar{\eta}\|_\infty^{1+\alpha}. \end{aligned}$$

Similarly,

$$\begin{aligned} P_X\big(\bar{f}(X) \neq f^*(X),\, \eta(X) \neq 1/2\big) &\leq P_X\big(0 < |\eta(X) - \tfrac{1}{2}| \leq |\eta(X) - \bar{\eta}(X)|\big) \\ &\leq P_X\big(0 < |\eta(X) - \tfrac{1}{2}| \leq \|\eta - \bar{\eta}\|_\infty\big) \\ &\leq C_0 \|\eta - \bar{\eta}\|_\infty^{\alpha}. \end{aligned}$$

■

Remark 5.1 *Lemma 5.1 offers an easy way to obtain the result of Theorem 3.3 in a slightly less precise form, with an extra logarithmic factor in the rate. In fact, under the strong density assumption, there exist nonparametric estimators $\hat{\eta}_n$ (for instance, suitably chosen locally polynomial estimators) such that*

$$\mathbb{E}\Big(\|\hat{\eta}_n - \eta\|_\infty^q\Big) \leq C \left(\frac{\log n}{n}\right)^{\frac{q\beta}{2\beta+d}}, \quad \forall\, q > 0,$$

uniformly over $\eta \in \Sigma(\beta, L, \mathbb{R}^d)$ [see, e.g., Stone (1982)]. Taking here $q = 1 + \alpha$ and applying the comparison inequality (5.1) we immediately get that the plug-in classifier $\hat{f}_n = \mathbb{1}_{\{\hat{\eta}_n \geq 1/2\}}$ has the excess risk $\mathcal{E}(\hat{f}_n)$ of the order $(n/\log n)^{-\beta(1+\alpha)/(2\beta+d)}$.

Another immediate application of Lemma 5.1 is to get lower bounds on the risks of regression estimators in the L_∞ norm from the corresponding lower bounds on the excess risks of classifiers (cf. Theorems 3.5 and 4.1). But here again we loose a logarithmic factor required for the best bounds.

We now consider the comparison inequalities for L_p norms with $1 \leq p < \infty$.

Lemma 5.2 *For any $1 \le p < \infty$ and any distribution P of (X,Y) satisfying Assumption (MA) with $\alpha > 0$ we have*

$$R(\bar f) - R(f^*) \le C_1(\alpha,p)\|\eta - \bar\eta\|_p^{\frac{p(1+\alpha)}{p+\alpha}}, \tag{5.3}$$

and

$$P_X\big(\bar f(X) \neq f^*(X),\, \eta(X) \neq 1/2\big) \le C_2(\alpha,p)\|\eta - \bar\eta\|_p^{\frac{p}{p+\alpha}}, \tag{5.4}$$

where $C_1(\alpha,p) = 2(\alpha+p)p^{-1}\big(\frac{p}{\alpha}\big)^{\frac{\alpha}{\alpha+p}} C_0^{\frac{p-1}{\alpha+p}}$, $C_2(\alpha,p) = (\alpha+p)p^{-1}\big(\frac{p}{\alpha}\big)^{\frac{\alpha}{\alpha+p}} C_0^{\frac{p}{\alpha+p}}$. In particular,

$$R(\bar f) - R(f^*) \le C_1(\alpha,2)\left(\int [\bar\eta(x) - \eta(x)]^2 P_X(dx)\right)^{\frac{1+\alpha}{2+\alpha}}. \tag{5.5}$$

Proof. For any $t > 0$ we have

$$\begin{aligned}
&R(\bar f) - R(f^*)\\
&\quad = \mathbb{E}\big[|2\eta(X) - 1|\mathbb{1}_{\{\bar f(X)\neq f^*(X)\}}\big]\\
&\quad = 2\mathbb{E}\big[|\eta(X) - 1/2|\mathbb{1}_{\{\bar f(X)\neq f^*(X)\}}\mathbb{1}_{\{0<|\eta(X)-1/2|\le t\}}\big]\\
&\quad\ + 2\mathbb{E}\big[|\eta(X) - 1/2|\mathbb{1}_{\{\bar f(X)\neq f^*(X)\}}\mathbb{1}_{\{|\eta(X)-1/2|>t\}}\big]\\
&\quad \le 2\mathbb{E}\big[|\eta(X) - \bar\eta(X)|\mathbb{1}_{\{0<|\eta(X)-1/2|\le t\}}\big] + 2\mathbb{E}\big[|\eta(X) - \bar\eta(X)|\mathbb{1}_{\{|\eta(X)-\bar\eta(X)|>t\}}\big]\\
&\quad \le 2\|\eta - \bar\eta\|_p\big[P_X(0 < |\eta(X) - 1/2| \le t)\big]^{\frac{p-1}{p}} + \frac{2\|\eta - \bar\eta\|_p^p}{t^{p-1}}
\end{aligned} \tag{5.6}$$

by Hölder and Markov inequalities. So, for any $t > 0$, introducing $E \triangleq \|\eta - \bar\eta\|_p$ and using Assumption (MA) to bound the probability in (5.6) we obtain

$$R(\bar f) - R(f^*) \le 2\left(C_0^{\frac{p-1}{p}} t^{\frac{\alpha(p-1)}{p}} E + \frac{E^p}{t^{p-1}}\right).$$

Minimizing in t the RHS of this inequality we get (5.3). Similarly,

$$\begin{aligned}
P_X\big(\bar f(X) \neq f^*(X),\, \eta(X) \neq 1/2\big) &\le P_X\big(0 < |\eta(X) - 1/2| \le t\big) + P_X\big(|\eta(X) - \bar\eta(X)| > t\big)\\
&\le C_0 t^\alpha + \frac{\|\eta - \bar\eta\|_p^p}{t^p},
\end{aligned}$$

and minimizing this bound in t we obtain (5.4). ■

If the regression function η belongs to the Hölder class $\Sigma\big(\beta, L, \mathbb{R}^d\big)$ there exist estimators $\hat\eta_n$ such that, uniformly over the class,

$$\mathbb{E}\Big\{\big[\hat\eta_n(X) - \eta(X)\big]^2\Big\} \le C n^{-\frac{2\beta}{2\beta+d}} \tag{5.7}$$

for some constant $C > 0$. This has been shown by Stone (1982) under the additional strong density assumption and by Yang (1999) with no assumption on P_X. Using (5.7) and (5.5) we get that the excess risk of the corresponding plug-in classifier $\hat{f}_n = \mathbb{I}_{\{\hat{\eta}_n \geq 1/2\}}$ admits a bound of the order $n^{-\frac{2\beta}{2\beta+d}\frac{1+\alpha}{2+\alpha}}$ which is suboptimal when $\alpha \neq 0$ (cf. Theorems 4.2, 4.3). In other words, under the margin assumption, Lemma 5.2 is not the right tool to analyze the convergence rate of plug-in classifiers. On the contrary, when no margin assumption is imposed (i.e., $\alpha = 0$ in our notation) inequality (1.2), which is a version of (5.5) for $\alpha = 0$, is precise enough to give the optimal rate of classification [Yang (1999)].

Another way to obtain (5.5) is to use Bartlett, Jordan and McAuliffe (2003): it is enough to apply their Theorem 10 with (in their notation) $\phi(t) = (1-t)^2, \psi(t) = t^2$ and to note that for this choice of ϕ we have $R_\phi(\bar{\eta}) - R^*_\phi = \|\eta - \bar{\eta}\|_2^2$. Blanchard, Lugosi and Vayatis (2003) used the result of Bartlett, Jordan and McAuliffe (2003) to prove fast rates of the order $n^{-\frac{2(1+\alpha)}{3(2+\alpha)}}$ for a boosting procedure over the class of regression functions η of bounded variation in dimension $d = 1$. Note that the same rates can be obtained for other plug-in classifiers using (5.5). Indeed, if η is of bounded variation, there exist estimators of η converging with the mean squared L_2 rate $n^{-2/3}$[cf. van de Geer (2000), Györfi et al. (2002)], and thus application of (5.5) immediately yields the rate $n^{-\frac{2(1+\alpha)}{3(2+\alpha)}}$ for the corresponding plug-in rule. However, Theorem 4.2 shows that this is not an optimal rate (here again we observe that inequality (5.5) fails to establish the optimal properties of plug-in classifiers). In fact, let $d = 1$ and let the assumptions of Theorem 4.2 be satisfied, where instead of assumption (ii) we use its particular instance: η belongs to a class of functions on $[0, 1]$ whose total variation is bounded by a constant $L < \infty$. It follows from Birman and Solomjak (1967) that Assumption (CAR) for this class is satisfied with $\rho = 1$ for any $1 \leq p < \infty$. Hence, we can apply (4.3) of Theorem 4.2 to find that

$$\sup_{P \in \mathcal{P}} \left\{ \mathbb{E}R(\hat{f}_n^S) - R(f^*) \right\} \leq C n^{-\frac{(1+\alpha)p}{(2+\alpha)p+(p+\alpha)}} \tag{5.8}$$

for the corresponding class $\mathcal{P}$. If $p > 2$ (recall that the value $p \in [1, \infty)$ is chosen by the statistician), the rate in (5.8) is faster than $n^{-\frac{2(1+\alpha)}{3(2+\alpha)}}$ obtained under the same conditions by Blanchard, Lugosi and Vayatis (2003).

6 Proofs

6.1 Proof of Theorem 3.2

Consider a distribution P in $\mathcal{P}_\Sigma$. Let A be the support of P_X. Fix $x \in A$ and $\delta > 0$. Consider the matrix $B \triangleq \big(B_{s_1,s_2}\big)_{|s_1|,|s_2|\le\lfloor\beta\rfloor}$ with elements $B_{s_1,s_2} \triangleq \int_{\mathbb{R}^d} u^{s_1+s_2}K(u)\mu(x+hu)du$. The smallest eigenvalue $\lambda_{\bar{B}}$ of $\bar{B}$ satisfies

$$\begin{aligned}\lambda_{\bar{B}} &= \min_{\|W\|=1} W^T\bar{B}W \\ &\ge \min_{\|W\|=1} W^TBW + \min_{\|W\|=1} W^T(\bar{B}-B)W \\ &\ge \min_{\|W\|=1} W^TBW - \textstyle\sum_{|s_1|,|s_2|\le\lfloor\beta\rfloor} \big|\bar{B}_{s_1,s_2} - B_{s_1,s_2}\big|.\end{aligned} \tag{6.1}$$

Let $A_n \triangleq \big\{u \in \mathbb{R}^d : \|u\| \le c;\ x+hu \in A\big\}$ where c is the constant appearing in (3.3). Using (3.3), for any vector W satisfying $\|W\|=1$, we obtain

$$\begin{aligned}W^TBW &= \textstyle\int_{\mathbb{R}^d} \big(\sum_{|s|\le\lfloor\beta\rfloor} W_s u^s\big)^2 K(u)\mu(x+hu)du \\ &\ge c\mu_{\min}\textstyle\int_{A_n} \big(\sum_{|s|\le\lfloor\beta\rfloor} W_s u^s\big)^2 du.\end{aligned}$$

By assumption of the theorem, $ch \le r_0$. Since the support of the marginal distribution is (c_0, r_0)-regular we get

$$\lambda[A_n] \ge h^{-d}\lambda\big[\mathcal{B}(x,ch)\cap A\big] \ge c_0 h^{-d}\lambda\big[\mathcal{B}(x,ch)\big] \ge c_0 v_d c^d,$$

where $v_d \triangleq \lambda\big[\mathcal{B}(0,1)\big]$ is the volume of the unit ball and $c_0 > 0$ is the constant introduced in the definition (2.1) of the (c_0, r_0)-regular set.

Let $\mathcal{A}$ denote the class of all compact subsets of $\mathcal{B}(0,c)$ having the Lebesgue measure $c_0 v_d c^d$. Using the previous displays we obtain

$$\min_{\|W\|=1} W^TBW \ge c\mu_{\min} \min_{\|W\|=1; S\in\mathcal{A}} \int_S \Big(\sum_{|s|\le\lfloor\beta\rfloor} W_s u^s\Big)^2 du \triangleq 2\mu_0. \tag{6.2}$$

By the compactness argument, the minimum in (6.2) exists and is strictly positive.

For $i = 1, \dots, n$ and any multi-indices s_1, s_2 such that $|s_1|, |s_2| \le \lfloor\beta\rfloor$, define

$$T_i \triangleq \tfrac{1}{h^d}\left(\tfrac{X_i-x}{h}\right)^{s_1+s_2} K\left(\tfrac{X_i-x}{h}\right) - \textstyle\int_{\mathbb{R}^d} u^{s_1+s_2}K(u)\mu(x+hu)du.$$

We have $\mathbb{E}T_i = 0$, $|T_i| \le h^{-d}\sup_{u\in\mathbb{R}^d}\big(1+\|u\|^{2\beta}\big)K(u) \triangleq \kappa_1 h^{-d}$ and the following bound for the variance of T_i:

$$\begin{aligned}\mathbb{V}\mathrm{ar}\, T_i &\le \tfrac{1}{h^{2d}}\mathbb{E}\left(\tfrac{X_i-x}{h}\right)^{2s_1+2s_2} K^2\left(\tfrac{X_i-x}{h}\right) \\ &= \tfrac{1}{h^d}\textstyle\int_{\mathbb{R}^d} u^{2s_1+2s_2}K^2(u)\mu(x+hu)du \\ &\le \tfrac{\mu_{\max}}{h^d}\textstyle\int_{\mathbb{R}^d}\big(1+\|u\|^{4\beta}\big)K^2(u)du \triangleq \tfrac{\kappa_2}{h^d}.\end{aligned}$$

From Bernstein's inequality, we get

$$P^{\otimes n}\left(|\bar{B}_{s_1,s_2} - B_{s_1,s_2}| > \epsilon\right) = P^{\otimes n}\left(\left|\frac{1}{n}\sum_{i=1}^n T_i\right| > \epsilon\right) \le 2\exp\left\{-\frac{nh^d\epsilon^2}{2\kappa_2 + 2\kappa_1\epsilon/3}\right\}.$$

This and (6.1) – (6.2) imply that

$$P^{\otimes n}(\lambda_{\bar{B}} \le \mu_0) \le 2M^2 \exp\left(-Cnh^d\right) \tag{6.3}$$

where M^2 is the number of elements of the matrix $\bar{B}$. Assume in what follows that n is large enough, so that $\mu_0 > (\log n)^{-1}$. Then for $\lambda_{\bar{B}} > \mu_0$ we have $|\hat{\eta}_n^*(x) - \eta(x)| \le |\hat{\eta}_n^{LP}(x) - \eta(x)|$. Therefore,

$$P^{\otimes n}\left(\left|\hat{\eta}_n^*(x) - \eta(x)\right| \ge \delta\right) \le P^{\otimes n}\left(\lambda_{\bar{B}} \le \mu_0\right) + P^{\otimes n}\left(\left|\hat{\eta}_n^{LP}(x) - \eta(x)\right| \ge \delta,\ \lambda_{\bar{B}} > \mu_0\right). \tag{6.4}$$

We now evaluate the second probability on the right hand side of (6.4). For $\lambda_{\bar{B}} > \mu_0$ we have $\hat{\eta}_n^{LP}(x) = U^T(0)Q^{-1}V$ (where V is given by (2.3)). Introduce the matrix $Z \triangleq \left(Z_{i,s}\right)_{1\le i\le n, |s|\le\lfloor\beta\rfloor}$ with elements

$$Z_{i,s} \triangleq (X_i - x)^s \sqrt{K\left(\frac{X_i - x}{h}\right)}.$$

The s-th column of Z is denoted by Z_s, and we introduce $Z^{(\eta)} \triangleq \sum_{|s|\le\lfloor\beta\rfloor} \frac{\eta^{(s)}(x)}{s!} Z_s$. Since $Q = Z^T Z$, we get

$$\forall |s| \le \lfloor\beta\rfloor: \quad U^T(0)Q^{-1}Z^T Z_s = \mathbb{I}_{\{s=(0,\dots,0)\}},$$

hence $U^T(0)Q^{-1}Z^T Z^{(\eta)} = \eta(x)$. So we can write

$$\hat{\eta}_n^{LP}(x) - \eta(x) = U^T(0)Q^{-1}(V - Z^T Z^{(\eta)}) = U^T(0)\bar{B}^{-1}\mathbf{a}$$

where $\mathbf{a} \triangleq \frac{1}{nh^d} H(V - Z^T Z^{(\eta)}) \in \mathbb{R}^M$ and H is a diagonal matrix $H \triangleq \left(H_{s_1,s_2}\right)_{|s_1|,|s_2|\le\lfloor\beta\rfloor}$ with $H_{s_1,s_2} \triangleq h^{-s_1}\mathbb{I}_{\{s_1=s_2\}}$. For $\lambda_{\bar{B}} > \mu_0$ we get

$$\left|\hat{\eta}_n^{LP}(x) - \eta(x)\right| \le \|\bar{B}^{-1}\mathbf{a}\| \le \lambda_{\bar{B}}^{-1}\|\mathbf{a}\| \le \mu_0^{-1}\|\mathbf{a}\| \le \mu_0^{-1} M \max_s |a_s|, \tag{6.5}$$

where a_s are the components of the vector $\mathbf{a}$ given by

$$a_s = \frac{1}{nh^d}\sum_{i=1}^n \left[Y_i - \eta_x(X_i)\right]\left(\frac{X_i - x}{h}\right)^s K\left(\frac{X_i - x}{h}\right).$$

Define

$$\begin{aligned} T_i^{(s,1)} &\triangleq \frac{1}{h^d}\left[Y_i - \eta(X_i)\right]\left(\frac{X_i - x}{h}\right)^s K\left(\frac{X_i - x}{h}\right), \\ T_i^{(s,2)} &\triangleq \frac{1}{h^d}\left[\eta(X_i) - \eta_x(X_i)\right]\left(\frac{X_i - x}{h}\right)^s K\left(\frac{X_i - x}{h}\right). \end{aligned}$$

We have

$$|a_s| \le \left|\tfrac{1}{n}\sum_{i=1}^n T_i^{(s,1)}\right| + \left|\tfrac{1}{n}\sum_{i=1}^n \left[T_i^{(s,2)} - \mathbb{E}T_i^{(s,2)}\right]\right| + \left|\mathbb{E}T_i^{(s,2)}\right|. \qquad (6.6)$$

Note that $\mathbb{E}T_i^{(s,1)} = 0$, $\left|T_i^{(s,1)}\right| \le \kappa_1 h^{-d}$, and

$$\begin{array}{lcl} \mathbb{V}\mathrm{ar}\, T_i^{(s,1)} & \le & 4^{-1}h^{-d}\int u^{2s}K^2(u)\mu(x+hu)du \le (\kappa_2/4)h^{-d}, \\ \left|T_i^{(s,2)} - \mathbb{E}T_i^{(s,2)}\right| & \le & L\kappa_1 h^{\beta-d} + L\kappa_2 h^{\beta} \le Ch^{\beta-d}, \\ \mathbb{V}\mathrm{ar}\, T_i^{(s,2)} & \le & h^{-d}L^2\int h^{2\beta}\|u\|^{2s+2\beta}K^2(u)\mu(x+hu)du \le L^2\kappa_2 h^{2\beta-d}. \end{array}$$

From Bernstein's inequality, for any $\epsilon_1, \epsilon_2 > 0$, we obtain

$$P^{\otimes n}\left(\left|\tfrac{1}{n}\sum_{i=1}^n T_i^{(s,1)}\right| \ge \epsilon_1\right) \le 2\exp\left\{-\tfrac{nh^d\epsilon_1^2}{\kappa_2/2 + 2\kappa_1\epsilon_1/3}\right\}$$

and

$$P^{\otimes n}\left(\left|\tfrac{1}{n}\sum_{i=1}^n \left[T_i^{(s,2)} - \mathbb{E}T_i^{(s,2)}\right]\right| \ge \epsilon_2\right) \le 2\exp\left\{-\tfrac{nh^d\epsilon_2^2}{2L^2\kappa_2 h^{2\beta} + 2Ch^{\beta}\epsilon_2/3}\right\}.$$

Since also

$$\left|\mathbb{E}T_i^{(s,2)}\right| \le Lh^{\beta}\int \|u\|^{s+\beta}K^2(u)\mu(x+hu)du \le L\kappa_2 h^{\beta}$$

we get, using (6.6), that if $3\mu_0^{-1}ML\kappa_2 h^{\beta} \le \delta \le 1$ the following inequality holds

$$\begin{array}{lcl} P^{\otimes n}\left(|a_s| \ge \frac{\mu_0\delta}{M}\right) & \le & P^{\otimes n}\left(\left|\frac{1}{n}\sum_{i=1}^n T_i^{(s,1)}\right| > \frac{\mu_0\delta}{3M}\right) + P^{\otimes n}\left(\left|\frac{1}{n}\sum_{i=1}^n \left[T_i^{(s,2)} - \mathbb{E}T_i^{(s,2)}\right]\right| > \frac{\mu_0\delta}{3M}\right) \\ & \le & 4\exp\left(-Cnh^d\delta^2\right). \end{array}$$

Combining this inequality with (6.3), (6.4) and (6.5), we obtain

$$P^{\otimes n}\left(\left|\hat{\eta}_n^*(x) - \eta(x)\right| \ge \delta\right) \le C_1\exp\Big(-C_2 nh^d\delta^2\Big) \qquad (6.7)$$

for $3m^{-1}ML\kappa_2 h^{\beta} \le \delta$ (for $\delta > 1$ inequality (6.7) is obvious since $\hat{\eta}_n^*, \eta$ take values in $[0,1]$). The constants C_1, C_2 in (6.7) do not depend on the distribution P_X, on its support A and on the point $x \in A$, so that we get (3.7). Now, (3.7) implies (3.8) for $Cn^{-\frac{\beta}{2\beta+d}} \le \delta$, and thus for all $\delta > 0$ (with possibly modified constants C_1 and C_2).

6.2 Proof of Theorems 3.5 and 4.1

The proof of both theorems is based on Assouad's lemma [see, e.g., Korostelev and Tsybakov (1993), Chapter 2 or Tsybakov (2004b), Chapter 2]. We apply it in a form adapted for the classification problem (Lemma 5.1 in Audibert (2004)).

For an integer $q \geq 1$ we consider the regular grid on $\mathbb{R}^d$ defined as

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

Let $n_q(x) \in G_q$ be the closest point to $x \in \mathbb{R}^d$ among points in G_q (we assume uniqueness of $n_q(x)$: if there exist several points in G_q closest to x we define $n_q(x)$ as the one which is closest to 0). Consider the partition $\mathcal{X}'_1, \dots, \mathcal{X}'_{q^d}$ of $[0,1]^d$ canonically defined using the grid G_q (x and y belong to the same subset if and only if $n_q(x) = n_q(y)$). Fix an integer $m \leq q^d$. For any $i \in \{1, \dots, m\}$, we define $\mathcal{X}_i \triangleq \mathcal{X}'_i$ and $\mathcal{X}_0 \triangleq \mathbb{R}^d \setminus \cup_{i=1}^m \mathcal{X}_i$, so that $\mathcal{X}_0, \dots, \mathcal{X}_m$ form a partition of $\mathbb{R}^d$.

Let $u : \mathbb{R}_+ \to \mathbb{R}_+$ be a nonincreasing infinitely differentiable function such that $u = 1$ on $[0, 1/4]$ and $u = 0$ on $[1/2, \infty)$. One can take, for example, $u(x) = \left(\int_{1/4}^{1/2} u_1(t)dt \right)^{-1} \int_x^\infty u_1(t)dt$ where the infinitely differentiable function u_1 is defined as

$$u_1(x) = \begin{cases} \exp\left\{ -\frac{1}{(1/2-x)(x-1/4)} \right\} & \text{for } x \in (1/4, 1/2), \\ 0 & \text{otherwise.} \end{cases}$$

Let $\phi : \mathbb{R}^d \to \mathbb{R}_+$ be the function defined as

$$\phi(x) \triangleq C_\phi u(\|x\|),$$

where the positive constant C_ϕ is taken small enough so ensure that $|\phi(x') - \phi_x(x')| \leq L\|x' - x\|^\beta$ for any $x, x' \in \mathbb{R}^d$. Thus, $\phi \in \Sigma(\beta, L, \mathbb{R}^d)$.

Define the hypercube $\mathcal{H} = \left\{ \mathbb{P}_{\vec{\sigma}} : \vec{\sigma} = (\sigma_1, \dots, \sigma_m) \in \{-1, 1\}^m \right\}$ of probability distributions $\mathbb{P}_{\vec{\sigma}}$ of (X, Y) on $\mathcal{Z} = \mathbb{R}^d \times \{0, 1\}$ as follows.

For any $\mathbb{P}_{\vec{\sigma}} \in \mathcal{H}$ the marginal distribution of X does not depend on $\vec{\sigma}$, and has a density μ w.r.t. the Lebesgue measure on $\mathbb{R}^d$ defined in the following way. Fix $0 < w \leq m^{-1}$ and a set A_0 of positive Lebesgue measure included in $\mathcal{X}_0$ (the particular choices of A_0 will be indicated later), and take: (i) $\mu(x) = w/\lambda[\mathcal{B}(0, (4q)^{-1})]$ if x belongs to a ball $\mathcal{B}(z, (4q)^{-1})$ for some $z \in G_d$, (ii) $\mu(x) = (1 - mw)/\lambda[A_0]$ for $x \in A_0$, (iii) $\mu(x) = 0$ for all other x.

Next, the distribution of Y given X for $\mathbb{P}_{\vec{\sigma}} \in \mathcal{H}$ is determined by the regression function $\eta_{\vec{\sigma}}(x) = P(Y = 1|X = x)$ that we define as $\eta_{\vec{\sigma}}(x) = \frac{1+\sigma_j\varphi(x)}{2}$ for any $x \in \mathcal{X}_j$, $j = 1, \dots, m$, and $\eta_{\vec{\sigma}} \equiv 1/2$ on $\mathcal{X}_0$, where $\varphi(x) \triangleq q^{-\beta}\phi\big(q[x - n_q(x)]\big)$. We will assume that $C_\phi \leq 1$ to ensure that φ and $\eta_{\vec{\sigma}}$ take values in $[0, 1]$.

For any $s \in \mathbb{N}^d$ such that $|s| \le \lfloor \beta \rfloor$, the partial derivative $D^s\varphi$ exists, and $D^s\varphi(x) = q^{|s|-\beta} D^s\phi\big(q[x - n_q(x)]\big)$. Therefore, for any $i \in \{1, \dots, m\}$ and any $x, x' \in \mathcal{X}_i$, we have

$$|\varphi(x') - \varphi_x(x')| \le L\|x - x'\|^\beta.$$

This implies that for any $\vec{\sigma} \in \{-1,1\}^m$ the function $\eta_{\vec{\sigma}}$ belongs to the Hölder class $\Sigma\big(\beta, L, \mathbb{R}^d\big)$.

We now check the margin assumption. Set $x_0 = \big(\frac{1}{2q}, \dots, \frac{1}{2q}\big)$. For any $\vec{\sigma} \in \{-1,1\}^m$ we have

$$\begin{aligned}
\mathbb{P}_{\vec{\sigma}}\big(0 < \big|\eta_{\vec{\sigma}}(X) - 1/2\big| \le t\big) &= m\mathbb{P}_{\vec{\sigma}}\big(0 < \phi[q(X - x_0)] \le 2tq^\beta\big) \\
&= m \int_{\mathcal{B}(x_0,(4q)^{-1})} 1\!\!1_{\{0<\phi[q(x-x_0)]\le 2tq^\beta\}} \frac{w}{\lambda[\mathcal{B}(0,(4q)^{-1})]} dx \\
&= \frac{mw}{\lambda[\mathcal{B}(0,1/4)]} \int_{\mathcal{B}(0,1/4)} 1\!\!1_{\{\phi(x)\le 2tq^\beta\}} dx \\
&= mw 1\!\!1_{\{t \ge C_\phi/(2q^\beta)\}}.
\end{aligned}$$

Therefore, the margin assumption (MA) is satisfied if $mw = O(q^{-\alpha\beta})$.

According to Lemma 5.1 in Audibert (2004), for any classifier $\hat{f}_n$ we have

$$\sup_{P\in\mathcal{H}} \big\{\mathbb{E}R(\hat{f}_n) - R(f^*)\big\} \ge mwb'(1 - b\sqrt{nw})/2 \tag{6.8}$$

where

$$\begin{aligned}
b &\triangleq \Big[1 - \big(\textstyle\int_{\mathcal{X}_1} \sqrt{1 - \varphi^2(x)}\, \mu_1(x) dx\big)^2\Big]^{1/2} = C_\phi q^{-\beta}, \\
b' &\triangleq \textstyle\int_{\mathcal{X}_1} \varphi(x)\mu_1(x)dx = C_\phi q^{-\beta}
\end{aligned}$$

with $\mu_1(x) = \mu(x)/\int_{\mathcal{X}_1} \mu(z)dz$.

We now prove Theorem 3.5. Take $q = \big\lfloor \bar{C} n^{\frac{1}{2\beta+d}} \big\rfloor$, $w = C'q^{-d}$ and $m = \big\lfloor C''q^{d-\alpha\beta} \big\rfloor$ with some positive constants $\bar{C}, C', C''$ to be chosen, and set $A_0 = [0,1]^d \setminus \cup_{i=1}^m \mathcal{X}_i$. The condition $\alpha\beta \le d$ ensures that the above choice of m is not degenerate: we have $m \ge 1$ for C'' large enough. We now prove that $\mathcal{H} \subset \mathcal{P}_\Sigma$ under the appropriate choice of $\bar{C}, C', C''$. In fact, select these constants so that the triplet (q, w, m) meets the conditions $m \le q^d$, $0 < w \le m^{-1}$, $mw = O(q^{-\alpha\beta})$. Then, in view of the argument preceding (6.8), for any $\vec{\sigma} \in \{-1,1\}^m$ the regression function $\eta_{\vec{\sigma}}$ belongs to $\Sigma\big(\beta, L, \mathbb{R}^d\big)$ and Assumption (MA) is satisfied. We now check that P_X obeys the strong density assumption. First, the density $\mu(x)$ equals to a positive constant for x belonging to the union of balls $\cup_{i=1}^m \mathcal{B}(z_i, (4q)^{-1})$ where z_i is the center of $\mathcal{X}_i$, and $\mu(x) = (1 - mw)/(1 - mq^{-d}) = 1 + o(1)$, as $n \to \infty$, for $x \in A_0$. Thus, $\mu_{\min} \le \mu(x) \le \mu_{\max}$ for some positive $\mu_{\min}$ and $\mu_{\max}$. (Note that this construction does not allow to choose any prescribed values of $\mu_{\min}$ and $\mu_{\max}$, because $\mu(x) = 1 + o(1)$. The problem can be fixed via a straightforward but cumbersome modification of the definition of A_0 that we skip here.) Second, the (c_0, r_0)-regularity of the support A of P_X with some $c_0 > 0$ and $r_0 > 0$ follows from the fact that, by construction, $\lambda(A \cap \mathcal{B}(x,r)) = (1 + o(1))\lambda([0,1]^d \cap \mathcal{B}(x,r))$ for all $x \in A$ and $r > 0$ (here again we skip the obvious generalization allowing to get any prescribed $c_0 > 0$). Thus, the strong density assumption is satisfied, and we conclude that $\mathcal{H} \subset \mathcal{P}_\Sigma$. Theorem 3.5 now follows from (6.8) if we choose C' small enough.

Finally, we prove Theorem 4.1. Take $q = \lfloor C n^{\frac{1}{(2+\alpha)\beta+d}} \rfloor$, $w = C' q^{2\beta}/n$ and $m = q^d$ for some constants $C > 0$, $C' > 0$, and choose A_0 as a Euclidean ball contained in $\mathcal{X}_0$. As in the proof of Theorem 3.5, under the appropriate choice of C and C', the regression function $\eta_{\vec{\sigma}}$ belongs to $\Sigma\big(\beta, L, \mathbb{R}^d\big)$ and the margin assumption (MA) is satisfied. Moreover, it is easy to see that the marginal distribution of X obeys the mild density assumption (the (c_0, r_0)-regularity of the support of P_X follows from considerations analogous to those in the proof of Theorem 3.5). Thus, $\mathcal{H} \subset \mathcal{P}'_\Sigma$. Choosing C' small enough and using (6.8) we obtain Theorem 4.1.

6.3 Proof of Proposition 3.4

The following lemma describes how the smoothness constraint on the regression function η at some point $x \in \mathbb{R}^d$ implies that η "stays close" to $\eta(x)$ in the vicinity of x.

Lemma 6.1 *For any distribution $P \in \mathcal{P}_\Sigma$ with regression function η and for any $\kappa > 0$, there exist $L' > 0$ and $t_0 > 0$ such that for any x in the support of P_X and $0 < t \leq t_0$, we have*

$$P_X\Big[\big|\eta(X) - \eta(x)\big| \leq t; X \in \mathcal{B}\big(x, \kappa t^{\frac{1}{1\wedge\beta}}\big)\Big] \geq L' t^{\frac{d}{1\wedge\beta}}.$$

Proof of Lemma 6.1. Let A denote the support of P_X. Let us first consider the case $\beta \leq 1$. Then for any $x, x' \in \mathbb{R}^d$, we have $\big|\eta(x') - \eta(x)\big| \leq L\|x' - x\|^\beta$. Let

$\kappa' = \kappa \wedge L^{-1/\beta}$. For any $0 < t \leq Lr_0^{\beta}$, we get

$$\begin{aligned} P_X\Big[\big|\eta(X) - \eta(x)\big| \leq t; X \in \mathcal{B}\big(x, \kappa t^{\frac{1}{1\wedge\beta}}\big)\Big] \\ &= P_X\Big[\big|\eta(X) - \eta(x)\big| \leq t; X \in \mathcal{B}\big(x, \kappa t^{\frac{1}{\beta}}\big) \cap A\Big] \\ &\geq P_X\Big[X \in \mathcal{B}\big(x, \kappa t^{\frac{1}{\beta}} \wedge \big(\tfrac{t}{L}\big)^{\frac{1}{\beta}}\big) \cap A\Big] \\ &\geq \mu_{\min}\lambda\Big[\mathcal{B}\big(x, \kappa' t^{\frac{1}{\beta}}\big) \cap A\Big] \\ &\geq c_0\mu_{\min}\lambda\Big[\mathcal{B}\big(x, \kappa' t^{\frac{1}{\beta}}\big)\Big] \\ &\geq c_0\mu_{\min}v_d(\kappa')^d t^{\frac{d}{\beta}}, \end{aligned}$$

which is the desired result with $L' \leq c_0\mu_{\min}v_d(\kappa')^d$ and $t_0 \leq Lr_0^{\beta}$.

For the case $\beta > 1$, by assumption, η is continuously differentiable. Let $\mathcal{C}(A)$ be the convex hull of the support A of P_X. By compactness of $\mathcal{C}(A)$, there exists $C > 0$ such that for any $s \in \mathbb{N}^d$ with $|s| = 1$,

$$\sup_{x \in \mathcal{C}(A)} \big|D^s\eta(x)\big| \leq C.$$

So we have for any $x, x' \in A$,

$$|\eta(x) - \eta(x')| \leq C\|x - x'\|.$$

The rest of the proof is then similar to the one of the first case. ■

- We will now prove the first item of Proposition 3.4. Let $P \in \mathcal{P}_\Sigma$ such that the regression function associated with P hits $1/2$ at $x_0 \in \overset{\circ}{A}$, where $\overset{\circ}{A}$ denotes the interior of the support of P_X. Let $r > 0$ such that $\mathcal{B}(x_0, r) \subset A$. Let $x \in \mathcal{B}(x_0, r)$ such that $\eta(x) \neq \frac{1}{2}$. Let $t_1 = \big|\eta(x) - 1/2\big|$. For any $0 < t \leq t_1$, let $x_t \in [x_0; x]$ such that $\big|\eta(x_t) - 1/2\big| = t/2$. We have $x_t \in A$ so that we can apply Lemma 6.1 (with $\kappa = 1$ for instance) and obtain for any $0 < t \leq t_1 \wedge (4t_0)$

  $$P_X\Big[0 < \big|\eta(X) - 1/2\big| \leq t\Big] \geq P_X\Big[\big|\eta(X) - \eta(x_t)\big| \leq t/4\Big] \geq L'(t/4)^{\frac{d}{1\wedge\beta}}.$$

  Now from the margin assumption, we get that for any small enough $t > 0$ $C_0 t^{\alpha} \geq L'(t/4)^{\frac{d}{1\wedge\beta}}$, hence $\alpha \leq \frac{d}{1\wedge\beta}$.

- For the second item of Proposition 3.4, to skip cumbersome details, we may assume that $\mathcal{C}$ contains the unit ball in $\mathbb{R}^d$. Consider the distribution such that

– P_X is the uniform measure on $\{(x_1,\ldots,x_d) \in \mathbb{R}^d : |x_1 - 1/4| + |x_2| + \cdots + |x_d| \le 1/4\}$

– the regression function associated with P is

$$\eta(x_1,\ldots,x_d) = \frac{1 + C_\eta \mathrm{sign}(x_1)|x_1|^{\beta\wedge 1} u(x_1)}{2},$$

where

$$u(t) = \begin{cases} \exp\left(-\frac{1}{1-t^2}\right) & \text{if } |t| < 1 \\ 0 & \text{otherwise,} \end{cases}$$

and $0 < C_\eta \le 1$ is small enough so that for any $x, x' \in \mathbb{R}^d$, η satisfies

$$|\eta(x') - \eta_x(x')| \le L\|x - x'\|^\beta.$$

For appropriate positive parameters $c_0, r_0, \mu_{\max} > \mu_{\min} > 0$, the only non-trivial task in checking that P belongs to $\mathcal{P}_\Sigma$ is to check the margin assumption. For t small enough, we have

$$P_X\Big[\big|\eta(X) - 1/2\big| \le t\Big] \le P_X\Big[|X_1|^{\beta\wedge 1} \le Ct; |X_1 - 1/4| + |X_2| + \cdots + |X_d| \le 1/4\Big]$$

for some $C > 0$. Therefore, we have $P_X\Big[0 < \big|\eta(X) - 1/2\big| \le t\Big] \le Ct^{\frac{d}{\beta\wedge 1}}$. So the margin assumption is satisfied for an appropriate C_0 whenever $\alpha \le \frac{d}{\beta\wedge 1}$. Since η hits $1/2$ at $0_{\mathbb{R}^d}$ which is in boundary of the support of P_X, we have proved the second assertion.

- For the third assertion of Proposition 3.4, to avoid cumbersome details again, we may assume that $\mathcal{C}$ contains the unit ball in $\mathbb{R}^d$. Consider the distribution such that

  – P_X is the uniform measure on the unit ball,

  – the regression function associated with P is

$$\eta(x) = \frac{1 + C_\eta \|x\|^2 u(\|x\|^2/2)}{2},$$

where $0 < C_\eta \le 1$ is small enough so that for any $x, x' \in \mathbb{R}^d$, η satisfies

$$|\eta(x') - \eta_x(x')| \le L\|x - x'\|^\beta.$$

For appropriate positive parameters $C_0, c_0, r_0, \mu_{\max} > \mu_{\min} > 0$, the distribution P belongs to $\mathcal{P}_\Sigma$ provided that $\alpha \leq d/2$ (in order that the margin assumption holds). We have obtained the desired result since η hits $1/2$ at $0_{\mathbb{R}^d}$ which is in the interior of the support of P_X.

- For the last item of Proposition 3.4, let $P \in \mathcal{P}_\Sigma$ such that the regression function η associated with P crosses $1/2$ at $x_0 \in \overset{\circ}{A}$. For $d = 1$, from the first item of the theorem, we necessarily have $\alpha(\beta \wedge 1) \leq 1$. Let us now consider the case: $d > 1$.

  Figure 1 will help to keep track of the following notation.

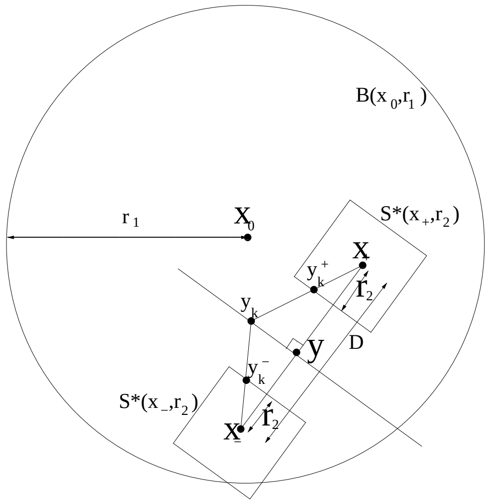

Figure 1: Notation summary

Let $r_1 > 0$ such that $\mathcal{B}(x_0, 3r_1) \subset A$. Introduce x_- and x_+ in $\mathcal{B}(x_0, r_1)$ such that $\eta(x_-) < 1/2$ and $\eta(x_+) > 1/2$. Let $t_1 = \big(1/2 - \eta(x_-)\big) \wedge \big(\eta(x_+) - 1/2\big)$. Define $y = \frac{x_- + x_+}{2}$, $e_d = \frac{x_+ - x_-}{\|x_+ - x_-\|}$ and $D = \|x_+ - x_-\|$. Let $e_1, \ldots, e_{d-1}$ be unit vectors such that $e_1, \ldots, e_d$ is an orthonormal basis of $\mathbb{R}^d$. Let $\mathcal{B}^*(x, r)$ (resp. $\mathcal{S}^*(x, r)$) denote the ball (resp. the sphere) centered at x and of radius r wrt the norm $\|x\|_* = \sup_{1 \leq i \leq d} |\langle x, e_i \rangle|$.

Since η is continuous, there exists $r_2 > 0$ such that

$$\begin{cases} \eta(x) < 1/2 - t_1/2 & \text{for any } x \in \mathcal{B}^*(x_-, r_2) \\ \eta(x) > 1/2 + t_1/2 & \text{for any } x \in \mathcal{B}^*(x_+, r_2) \end{cases}$$

Let $\zeta = \frac{1}{\beta \wedge 1}$. For any $k = (k_1, \ldots, k_{d-1}) \in \mathbb{Z}^{d-1}$, introduce

$$y_k = y + t^\zeta \sum_{i=1}^{d-1} k_i e_i.$$

For any t in $]0; t_1[$, consider the grid $G = \left\{ y_k; k \in \mathbb{Z}^{d-1}, \max_{1 \le i \le d-1} |k_i| \le \frac{D}{2\sqrt{d-1}t^\zeta} \right\}$. For any y_k in G, we have $\|y_k - y\| \le \sqrt{d-1} \max_{1 \le i \le d-1} |t^\zeta k_i| \le D/2 \le r_1$. Therefore, using that $y \in \mathcal{B}(x_0, r_1)$, the grid G is included in $\mathcal{B}(x_0, 2r_1)$. For any $y_k \in G$, let $y_k^- = [x_-; y_k] \cap \mathcal{S}^*(x_-, r_2)$ and $y_k^+ = [x_+; y_k] \cap \mathcal{S}^*(x_+, r_2)$. Since $\|y_k - y\| \le D/2$, we have $y_k^- = x_- + r_2 e_d + \frac{2r_2}{D} t^\zeta \sum_{i=1}^{d-1} k_i e_i$ and $y_k^+ = x_+ - r_2 e_d + \frac{2r_2}{D} t^\zeta \sum_{i=1}^{d-1} k_i e_i$.

For any y_k in G, consider the continuous path formed by the segments $[y_k^-; y_k]$ and $[y_k; y_k^+]$. Since η is continuous on this path, there exists $w_k \in \gamma_k \triangleq [y_k^-; y_k] \cup [y_k; y_k^+]$ such that $\eta(w_k) = 1/2 + t/2$. Now let us show that when $k \neq k'$, w_k and $w_{k'}$ are at least $\frac{\sqrt{2}r_2}{D} t^\zeta$ away from each other. The distance between w_k and $w_{k'}$ is not less than the distance between the paths γ_k and $\gamma_{k'}$. Let U denote the biggest integer smaller than or equal to $\frac{D}{2\sqrt{d-1}t^\zeta}$. When $y_k \neq y_{k'}$ in G, the distance between γ_k and $\gamma_{k'}$ is minimum for $k = K \triangleq (U, \ldots, U)$ and $k' = K' \triangleq (U-1, U, \ldots, U)$. This distance is equal to the distance between y_K^- and its orthogonal projection on $[y_{K'}^-; y_{K'}]$, which is the distance between y_K^- and the line $(x_-; y_{K'})$. Let $K'' = (0, U, \ldots, U) \in \mathbb{Z}^{d-1}$. To compute this distance V, it suffices to look at the plane $(x_-; y_{K''}; y_K)$ (see figure 2).

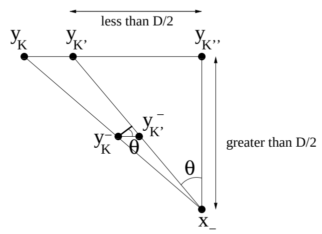

Figure 2: plane $(x_-; y_{K''}; y_K)$

We obtain that the angle θ between $y_{K'} - x_-$ and $y_{K''} - x_+$ is smaller than $\pi/4$. As a consequence, $V = \|y_K^- - y_{K'}^-\| \cos\theta \ge \sqrt{2} r_2 t^\zeta / D$.

Finally, focusing on the behaviour of the regression function near the w_k's, by using Lemma 6.1 with $\kappa = \frac{4^\zeta r_2}{\sqrt{2}D}$, we obtain that there exists $L' > 0$ and $t_0 > 0$

such that for any $0 < t < 4t_0 \wedge t_1$,

$$\begin{array}{rcl} C_0 t^\alpha & \geq & P_X\Big[0 < \big|\eta(X) - \frac{1}{2}\big| \leq t\Big] \\ & \geq & \sum_{k\in\mathbb{Z}^{d-1}:\ \max_{1\leq i\leq d-1}|k_i|\leq \frac{D}{2\sqrt{d-1}t^\zeta}} P_X\Big[\big|\eta(X) - \eta(w_k)\big| \leq t/4; X \in \mathcal{B}\big(w_k, \frac{r_2 t^\zeta}{\sqrt{2}D}\big)\Big] \\ & \geq & (2U+1)^{d-1} L'(t/4)^{d\zeta} \\ & \geq & \big(\frac{D}{2\sqrt{d-1}t^\zeta}\big)^{d-1} L'(t/4)^{d\zeta} \\ & \geq & Ct^\zeta, \end{array}$$

hence $\alpha \leq \zeta$ (which is the desired result).

For the converse, the proof is similar to the ones of the second and third assertions of the proposition. Without loss of generality, we may assume that $\mathcal{S} = \big\{(x_1, \ldots, x_d) \in \mathbb{R}^d : \max_{1\leq i\leq d}|x_i| \leq 1/2\big\}$ is a subset of $\mathcal{C}$. we consider the distribution P such that

- P_X is the uniform measure on $\mathcal{S}$
- the regression function associated with P is

$$\eta(x_1, \ldots, x_d) = \frac{1 + C_\eta \mathrm{sign}(x_1)|x_1|^{\beta\wedge 1}u(x_1)}{2},$$

  where $0 < C_\eta \leq 1$ is small enough so that for any $x, x' \in \mathbb{R}^d$, η satisfies

$$|\eta(x') - \eta_x(x')| \leq L\|x - x'\|^\beta.$$

For small enough $t > 0$, we have

$$P_X\Big[\big|\eta(X) - 1/2\big| \leq t\Big] \leq P_X\big[|X_1|^{\beta\wedge 1} \leq Ct\big],$$

for some constant $C > 0$, so that we have $P_X\Big[0 < \big|\eta(X) - \frac{1}{2}\big| \leq t\Big] \leq 2(Ct)^{\frac{1}{\beta\wedge 1}}$. As a consequence, for appropriate parameters $C_0, c_0, r_0, \mu_{\max} > \mu_{\min} > 0$, the distribution P belongs to $\mathcal{P}_\Sigma$ whenever $\alpha \leq \frac{1}{\beta\wedge 1}$. Since η crosses $1/2$ at $0_{\mathbb{R}^d}$ which is in the interior of the support of P_X, the converse holds.

6.4 Proof of Theorem 4.2

We prove the theorem for $p < \infty$. The proof for $p = \infty$ is analogous. For any decision rule f we set $d(f) \triangleq R(f) - R(f^*)$ and

$$f^{**}(x, f) \triangleq \begin{cases} f^*(x) & \text{if } \eta(x) \neq 1/2, \\ f(x) & \text{if } \eta(x) = 1/2, \end{cases} \qquad \forall\, x \in \mathbb{R}^d.$$

Lemma 6.2 *Under Assumption (MA) for any decision rule f we have*

$$P_X(f(X) \neq f^{**}(X,f)) \leq Cd(f)^{\alpha/(1+\alpha)}. \tag{6.9}$$

Proof. Note that $f^{**}(\cdot, f)$ is a Bayes rule, and following the same lines as in Proposition 1 of Tsybakov (2004a) we get $P_X(f(X) \neq f^{**}(X,f), \eta(X) \neq 1/2) \leq Cd(f)^{\alpha/(1+\alpha)}$. It remains to observe that $P_X(f(X) \neq f^{**}(X,f), \eta(X) \neq 1/2) = P_X(f(X) \neq f^{**}(X,f))$. ■

For a Borel function $\bar\eta$ on $\mathbb{R}^d$ define $f_{\bar\eta} \triangleq \mathbb{I}_{\{\bar\eta \geq 1/2\}}$, $f^*_{\bar\eta}(\cdot) \triangleq f^{**}(\cdot, f_{\bar\eta})$ and

$$Z_n(f_{\bar\eta}) \triangleq [R_n(f_{\bar\eta}) - R_n(f^*_{\bar\eta})] - [R(f_{\bar\eta}) - R(f^*_{\bar\eta})] = [R_n(f_{\bar\eta}) - R_n(f^*_{\bar\eta})] - d(f_{\bar\eta}).$$

Let η_n be an element of $\mathcal{N}_{\varepsilon_n}$ such that $\|\eta_n - \eta\|_{p,\lambda} \leq \varepsilon_n$, where $\|\cdot\|_{p,\lambda}$ is the $L_p(\mathcal{C}, \lambda)$ norm. In view of the assumption on $\mathcal{P}$ we have $\|\eta_n - \eta\|_p \leq \mu_{\max}^{1/p}\varepsilon_n$ where $\|\cdot\|_p$ is the $L_p(\mathbb{R}^d, P_X)$ norm. It follows from the comparison inequality (5.3) that $d(f_{\eta_n}) \leq C\varepsilon_n^{\frac{(1+\alpha)p}{p+\alpha}} \triangleq \delta_n$. Set

$$\Delta_n = Cn^{-\frac{(1+\alpha)p}{(2+\alpha)p+\rho(p+\alpha)}}$$

(i.e., Δ_n is of the order of desired rate). Fix $t > 0$ and introduce the set

$$\mathcal{N}_n^* = \{\bar\eta \in \mathcal{N}_{\varepsilon_n} : \ d(f_{\bar\eta}) \geq t\Delta_n\}.$$

For any $t > 0$ we have

$$\begin{aligned}
\mathbb{P}(d(\hat f_n^s) \geq t\Delta_n) &\leq \mathbb{P}(\min_{\bar\eta \in \mathcal{N}_n^*}[R_n(f_{\bar\eta}) - R_n(f_{\eta_n})] \leq 0) \\
&= \mathbb{P}(\min_{\bar\eta \in \mathcal{N}_n^*}[Z_n(f_{\bar\eta}) - Z_n(f_{\eta_n}) + d(f_{\bar\eta}) - d(f_{\eta_n})] \leq 0) \\
&\leq \mathbb{P}(\min_{\bar\eta \in \mathcal{N}_n^*}[Z_n(f_{\bar\eta}) - Z_n(f_{\eta_n}) + d(f_{\bar\eta})/2 + t\Delta_n/2 - d(f_{\eta_n})] \leq 0) \\
&\leq \mathbb{P}(\min_{\bar\eta \in \mathcal{N}_n^*}[Z_n(f_{\bar\eta}) + d(f_{\bar\eta})/2] \leq 0) \\
&\quad + \mathbb{P}(Z_n(f_{\eta_n}) \geq t\Delta_n/2 - d(f_{\eta_n})) \\
&\leq \mathbb{P}(\min_{\bar\eta \in \mathcal{N}_n^*}[Z_n(f_{\bar\eta}) + d(f_{\bar\eta})/2] \leq 0) \\
&\quad + \mathbb{P}(Z_n(f_{\eta_n}) \geq t\Delta_n/2 - \delta_n).
\end{aligned}$$

Since Δ_n is of the same order as δ_n, we can choose t large enough to have $t\Delta_n/2 - \delta_n \geq t\Delta_n/4$. Thus,

$$\begin{aligned}\mathbb{P}(d(\hat{f}_n^s) \geq t\Delta_n) &\leq \text{card } \mathcal{N}_n^* \max_{\bar{\eta}\in\mathcal{N}_n^*} \mathbb{P}(Z_n(f_{\bar{\eta}}) \leq -d(f_{\bar{\eta}})/2) \\ &\quad +\mathbb{P}(Z_n(f_{\eta_n}) \geq t\Delta_n/4) \\ &\leq \exp(A'\varepsilon_n^{-\rho}) \max_{\bar{\eta}\in\mathcal{N}_n^*} \mathbb{P}(Z_n(f_{\bar{\eta}}) \leq -d(f_{\bar{\eta}})/2) \\ &\quad +\mathbb{P}(Z_n(f_{\eta_n}) \geq t\Delta_n/4).\end{aligned}$$

Note that for any decision rule f the value $Z_n(f)$ is an average of n i.i.d. bounded and centered random variables whose variance does not exceed $P_X(f(X) \neq f^{**}(X, f))$. Thus, using Bernstein's inequality and (6.9) we obtain

$$\mathbb{P}(-Z_n(f) \geq a) \leq \exp\left(-\frac{Cna^2}{a + d(f)^{\alpha/(1+\alpha)}}\right), \quad \forall\, a > 0.$$

Therefore, for $\bar{\eta} \in \mathcal{N}_n^*$,

$$\begin{aligned}\mathbb{P}(Z_n(f_{\bar{\eta}}) \leq -d(f_{\bar{\eta}})/2) &\leq \exp(-Cnd(f_{\bar{\eta}})^{(2+\alpha)/(1+\alpha)}) \\ &\leq \exp(-Cn(t\Delta_n)^{(2+\alpha)/(1+\alpha)}).\end{aligned}$$

Similarly, for $t > C$,

$$\begin{aligned}\mathbb{P}(Z_n(f_{\eta_n}) \geq t\Delta_n/4) &\leq \exp\left(-\frac{Cn\Delta_n^2}{\Delta_n + d(f_{\eta_n})^{\alpha/(1+\alpha)}}\right) \\ &\leq \exp\left(-\frac{Cn\Delta_n^2}{\Delta_n + \delta_n^{\alpha/(1+\alpha)}}\right) \\ &\leq \exp\left(-Cn\Delta_n^{(2+\alpha)/(1+\alpha)}\right).\end{aligned}$$

The result of the theorem follows now from the above inequalities and the relation $n\Delta_n^{(2+\alpha)/(1+\alpha)} \asymp \varepsilon_n^{-\rho}$.

References

[1] Audibert, J.-Y. (2004). Classification using Gibbs estimators under complexity and margin assumptions. Preprint, Laboratoire de Probabilités et Modelès Aléatoires, http://www.proba.jussieu.fr/mathdoc/textes/PMA-908.pdf.

[2] Bartlett, P.L., Jordan, M.I. and McAuliffe, J.D. (2003). Convexity, classification and risk bounds. Techn. Report 638, University of California at Berkeley.

[3] Birman, M.S. and Solomjak, M.Z. (1967) Piecewise-polynomial approximation of functions of the classes W_p^α. *Mat. Sbornik* **73** 295-317.

[4] Blanchard, G., Bousquet,O. and Massart, P. (2004) Statistical performance of support vector machines. Manuscript. http://www.kyb.mpg.de/publications/pss/ps2731.ps

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

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

[7] Györfi, L., Kohler, M., Krzyżak, A. and Walk, H. (2002). *A Distribution-Free Theory of Nonparametric Regression.* Springer, New York.

[8] Kolmogorov, A.N. and Tikhomorov V.M. (1961) ϵ-entropy and ϵ-capacity of sets in function spaces. *Translations of the American Mathematical Society* **17** 277-364.

[9] Koltchinskii, V. (2005) Local Rademacher complexities and oracle inequalities in risk minimization. Manuscript.

[10] Koltchinskii, V. and Beznosova, O. (2005) Exponential convergence rates in classification. Proceedings of the 18th Conference on Learning Theory (COLT-2005).

[11] Korostelev, A. P. and Tsybakov, A. B. (1993). *Minimax Theory of Image Reconstruction.* Lecture Notes in Statistics **82**, Springer, New York, Berlin, Heidelberg.

[12] Lugosi, G. and Wegkamp, M. (2004). Complexity regularization via localized random penalties. *Ann. Statist.* **32** 1679 - 1697.

[13] Mammen, E. and Tsybakov, A. B. (1999). Smooth discrimination analysis. *Ann. Statist.* **27** 1808 - 1829.

[14] Massart, P. and Nédélec, E. (2003) Risk bounds for statistical learning. Preprint. http://www.math.u-psud.fr/ massart/margin.pdf.

[15] Scovel, J.C. and Steinwart, I. (2004). Fast Rates for Support Vector Machines. Los Alamos National Laboratory Technical Report LA-UR 03-9117. http://www.c3.lanl.gov/ ingo/publications/ann-04a.pdf.

[16] Stone, C.J. (1980) Optimal rates of convergence for nonparametric estimators, *Ann. Statist.* **8** 1348–1360.

[17] Stone, C.J. (1982) Optimal global rates of convergence for nonparametric regression. *Ann. Statist.* **10** 1040-1053.

[18] Tarigan, B. and van de Geer, S. (2004). Adaptivity of Support Vector Machines with ℓ_1 Penalty. Preprint, University of Leiden, http://www.math.leidenuniv.nl/ geer/svm4.pdf.

[19] Tsybakov, A.B. (2004a). Optimal aggregation of classifiers in statistical learning. *Ann. Statist.* **32** 135-166.

[20] Tsybakov, A. B. (2004b). *Introduction à l'estimation non–paramétrique.* Springer, Berlin.

[21] Tsybakov, A. B. and van de Geer, S. (2005) Square root penalty: adaptation to the margin in classification and in edge estimation. *Annals of Statistics* **33** 1203 - 1224.

[22] van de Geer, S. (2000). *Empirical Processes in M-Estimation*, Cambridge Univ. Press.

[23] Vapnik, V. N. (1998). *Statistical Learning Theory.* Wiley, New York.

[24] Vapnik, V. N. and Chervonenkis, A. Ja. (1974). *Theory of Pattern Recognition.* Nauka, Moscow (in Russian).

[25] Yang, Y. (1999) Minimax nonparametric classification - Part I: Rates of convergence, Part II: Model selection for adaptation. *IEEE Trans. Inf. Theory* **45** 2271-2292.

[1]Center for Education and Research in Informatics
Ecole Nationale des Ponts et Chaussées
19, rue Alfred Nobel
Cité Descartes, Champs-sur-Marne
77455 Marne-La-Valle, France
e-mail: audibert@certis.enpc.fr

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