Centre de Recherche en Économie et Statistique
CREST

Série des Documents de Travail

n° 2013-39

Optimal Exponential Bounds on the Accuracy of Classification

G. KERKYACHARIAN[1] – A. B. TSYBAKOV[2]
V. TEMLYAKOV[3] – D. PICARD[4]
V. KOLTCHINSKII[5]

November 2013

1 Université Paris-Diderot, CNRS-LPMA.
2 CREST- Laboratoire de Statistique.
3 University of South Carolina, Department of Mathematics.
4 Université Paris-Diderot, CNRS-LPMA.
5 School of Mathematics, Georgia Institute of Technology, Atlanta, USA.

Optimal exponential bounds on the accuracy of classification

G. Kerkyacharian

Université Paris-Diderot, CNRS-LPMA, Bâtiment Sophie Germain,

5 rue Thomas Mann, 75205 Paris CEDEX 13, France

A.B. Tsybakov

CREST-ENSAE, 3 av. Pierre Larousse, 92240 Malakoff,France

V. Temlyakov,

Department of Mathematics, University of South Carolina, Columbia, SC29208, USA

D. Picard

Université Paris-Diderot, CNRS-LPMA, Bâtiment Sophie Germain

5 rue Thomas Mann, 75205 Paris CEDEX 13, France

V. Koltchinskii

School of Mathematics, Georgia Institute of Technology, Atlanta, GA 30332 USA

November 19, 2013

Abstract

Consider a standard binary classification problem, in which (X, Y) is a random couple in $\mathcal{X} \times \{0, 1\}$ and the training data consists of n i.i.d. copies of (X, Y). Given a binary classifier $f : \mathcal{X} \mapsto \{0, 1\}$, the generalization error of f is defined by $R(f) = \mathbb{P}\{Y \neq f(X)\}$. Its minimum R^* over all binary classifiers f is called the Bayes risk and is attained at a Bayes classifier. The performance of any binary classifier $\hat{f}_n$ based on the training data is characterized by the excess risk $R(\hat{f}_n) - R^*$. We study Bahadur's type exponential bounds on the following minimax accuracy confidence function based on the excess risk:

$$AC_n(\mathcal{M}, \lambda) = \inf_{\hat{f}_n} \sup_{P \in \mathcal{M}} \mathbb{P}\left(R(\hat{f}_n) - R^* \geq \lambda\right), \lambda \in [0, 1],$$

where the supremum is taken over all distributions P of (X, Y) from a given class of distributions $\mathcal{M}$ and the infimum is over all binary classifiers $\hat{f}_n$ based on the training data. We study how this quantity depends on the complexity of the class of distributions $\mathcal{M}$ characterized by exponents of entropies of the class of regression functions or of the class of Bayes classifiers corresponding to the distributions from $\mathcal{M}$. We also study its dependence on margin parameters of the classification problem. In particular, we show that, in the case when $\mathcal{X} = [0,1]^d$ and $\mathcal{M}$ is the class all distributions satisfying the margin condition with exponent $\alpha > 0$ and such that the regression function η belongs to a given Hölder class of smoothness $\beta > 0$,

$$-\frac{\log AC_n(\mathcal{M}, \lambda)}{n} \asymp \lambda^{\frac{2+\alpha}{1+\alpha}}, \ \lambda \in [Dn^{-\frac{1+\alpha}{2+\alpha+d/\beta}}, \lambda_0]$$

for some constants $D, \lambda_0 > 0$.

AMS classification: 62G08, 62G07, 62H05, 68T10

Key words and phrases: statistical learning, classification, fast rates, optimal rate of convergence, excess risk, margin condition, Bahadur efficiency

1 Introduction

Let $(\mathcal{X}, \mathcal{A})$ be a measurable space. We consider a random variable (X, Y) in $\mathcal{X} \times \{0,1\}$ with probability distribution denoted by P. Denote by μ_X the marginal distribution of X in $\mathcal{X}$ and by

$$\eta(x) \triangleq \eta_P(x) \triangleq P(Y = 1 | X = x) = E(Y | X = x)$$

the conditional probability of $Y = 1$ given $X = x$, which is also the regression function of Y on X. Assume that we have n i.i.d. observations of the pair (X, Y) denoted by $\mathcal{D}_n = ((X_i, Y_i))_{i=1,\dots,n}$. The aim is to predict the output label Y for any input X in $\mathcal{X}$ from the observations $\mathcal{D}_n$.

We recall some standard facts of classification theory. A *prediction rule* is a measurable function $f : \mathcal{X} \longmapsto \{0,1\}$. To any prediction rule we associate the *classification error* (probability of misclassification):

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

It is well known (see, e.g., Devroye *et al.* [4]) that

$$\min_{f\,:\,\mathcal{X} \longmapsto \{0,1\}} R(f) = R(f^*) \triangleq R^*,$$

where the prediction rule f^*, called the *Bayes rule*, is defined by

$$f^*(x) \triangleq f_P^*(x) \triangleq I_{\{\eta(x)\geq 1/2\}}, \qquad \forall x \in \mathcal{X},$$

where I_A denotes the indicator function of A. The minimal risk R^* is called the *Bayes risk*. A *classifier* is a function, $\hat{f}_n = \hat{f}_n(X, \mathcal{D}_n)$, measurable with respect to $\mathcal{D}_n$ and X with values in $\{0, 1\}$, that assigns to the sample $\mathcal{D}_n$ a prediction rule $\hat{f}_n(\cdot, \mathcal{D}_n) : \mathcal{X} \longmapsto \{0, 1\}$. A key characteristic of $\hat{f}_n$ is its risk $\mathbb{E}[R(\hat{f}_n)]$, where

$$R(\hat{f}_n) \triangleq P\big(Y \neq \hat{f}_n(X)|\mathcal{D}_n\big).$$

The aim of statistical learning is to construct a classifier $\hat{f}_n$ such that $R(\hat{f}_n)$ is as close to R^* as possible. The accuracy of a classifier $\hat{f}_n$ is usually measured by the quantity $\mathbb{E}[R(\hat{f}_n) - R^*]$ called the (expected) *excess risk* of $\hat{f}_n$, where the expectation $\mathbb{E}$ is taken with respect to the distribution of $\mathcal{D}_n$. We say that the classifier $\hat{f}_n$ learns with the convergence rate $\psi(n)$, if there exists an absolute constant $C > 0$ such that for any integer n, $\mathbb{E}[R(\hat{f}_n) - R^*] \leq C\psi(n)$.

Given a convergence rate, Theorem 7.2 of Devroye *et al.* [4] shows that no classifier can learn with this rate for *all* underlying probability distributions P. To achieve some rates of convergence, we need to restrict the class of possible distributions P. For instance, Yang [19] provides examples of classifiers learning with a given convergence rate under complexity assumptions expressed via the smoothness properties of the regression function η. Under complexity assumptions alone, no matter how strong they are, the rates cannot be faster than $n^{-1/2}$ (cf. Devroye *et al.* [4]). Nevertheless, they can approach n^{-1} if we add a control on the behavior of the regression function η at the level $1/2$ (the distance $|\eta(\cdot) - 1/2|$ is sometimes called the margin). This behavior is usually characterized by the following condition, cf. [15].

Margin condition. *The probability distribution* P *on the space* $\mathcal{X} \times \{0, 1\}$ *satisfies the Margin condition with exponent* $0 < \alpha < \infty$ *if there exists* $C_M > 0$ *such that*

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

Equivalently, one can assume that (1) holds only for $t \in [0, t_0]$ with some $t_0 \in [0, 1)$. This would imply (1) for all $t \in [0, 1)$ (with a larger value of C_M). Under the margin condition, *fast rates*, that is, rates faster than $n^{-1/2}$ can be obtained for different classifiers, cf. Tsybakov [15], Blanchard *et al.* [2], Bartlett *et al.* [3], Tsybakov and van de Geer [17], Koltchinskii [9],

Massart and Nédélec [12], Audibert and Tsybakov [1], Scovel and Steinwart [13] among others.

In this paper, we will study the closeness of $R(\widehat{f}_n)$ to R^* in a more refined way. Our measure of performance is inspired by the Bahadur efficiency of estimation procedures but in contrast to the classical Bahadur approach (cf., e.g., [7]) our results are non-asymptotic.

For a classifier $\widehat{f}_n$ and for a tolerance $\lambda > 0$, define *the accuracy confidence function* (or, shortly, the AC-function):

$$AC_n(\widehat{f}_n, \lambda) = \mathbb{P}\left(R(\widehat{f}_n) - R^* \geq \lambda\right).$$

Here $\mathbb{P}$ denotes the probability distribution of the observed sample $\mathcal{D}_n$. Note that $AC_n(\widehat{f}_n, \lambda) = 0$ for $\lambda > 1$ since $0 \leq R(f) \leq 1$ for all classifiers f. Moreover, $R(\widehat{f}_n) - R^* \leq 1/2$ for all interesting classifiers $\widehat{f}_n$. Indeed, it makes no sense to deal with the probabilities of error $R(\widehat{f}_n)$ greater than $1/2$ (note that $R(\widehat{f}_n) = 1/2$ is achieved when $\widehat{f}_n$ is the simple random guess classifier). Therefore, without loss of generality we can consider only $\lambda \leq 1/2$. In fact,we will sometimes use a slightly stronger restriction $\lambda \leq \lambda_0$ for some $\lambda_0 < 1/2$ independent of n.

It is intuitively clear that if the tolerance is low (λ under some critical value λ_n), the probability $AC_n(\widehat{f}_n, \lambda)$ is kept larger than some fixed level. On the opposite, for $\lambda \geq \lambda_n$, the quality of the procedure $\widehat{f}_n$ can be characterized by the rate of convergence of $AC_n(\widehat{f}_n, \lambda)$ towards zero as $n \to \infty$. Observe that evaluating the critical value λ_n yields, as a consequence, bounds and the associated rates for the excess risk $\mathbb{E}R(\widehat{f}_n) - R^*$, which is a commonly used measure of performance.

For a class $\mathcal{M}$ of probability measures P, we define the minimax AC-function

$$AC_n(\mathcal{M}, \lambda) \triangleq \inf_{\widehat{f}_n \in \mathbb{S}_n} \sup_{P \in \mathcal{M}} \mathbb{P}\left(R(\widehat{f}_n) - R^* \geq \lambda\right),$$

where $\mathbb{S}_n$ is the set of all classifiers. We will consider classes $\mathcal{M} = \mathcal{M}(r, \alpha)$ defined by the following conditions:

(a) A margin assumption with exponent α.

(b) A complexity assumption expressed in terms of the rate of decay $r > 0$ of an ε-entropy.

The main results of this paper can be summarized as follows. Fix $r, \alpha > 0$ and set $\lambda_n = Dn^{-\frac{1+\alpha}{2+\alpha+r}}$ where $D > 0$. Then, we have an upper bound:

There exist positive constants $C,\ c$ such that, for all classes $\mathcal{M} = \mathcal{M}(r, \alpha)$ satisfying the above two conditions,

$$AC_n(\mathcal{M}, \lambda) \le C \exp\{-cn\lambda^{\frac{2+\alpha}{1+\alpha}}\}, \quad \forall\, \lambda \ge \lambda_n. \tag{2}$$

Furthermore, we prove the corresponding lower bound: there exists a class $\mathcal{M}$ satisfying the same conditions (a) and (b) such that

$$AC_n(\mathcal{M}, \lambda) \ge p_0, \qquad 0 < \lambda \le \lambda_n^- \asymp \lambda_n, \tag{3}$$

$$AC_n(\mathcal{M}, \lambda) \ge C' \exp\{-c'n\lambda^{\frac{2+\alpha}{1+\alpha}}\}, \quad \lambda_n \asymp \lambda_n^+ \le \lambda \le \lambda_0 \tag{4}$$

for some positive constants $p_0,\ C',\ c'$ and $0 < \lambda_0 < 1/2$ depending only on C_M and α. Thus, we quantify the critical level phenomenon discussed above and we derive the exact exponential rate $\exp\{-cn\lambda^{\frac{2+\alpha}{1+\alpha}}\}$ for minimax AC-function over the critical level. In particular, this implies the following bounds on the minimax AC-function in the case when $\mathcal{X} = [0,1]^d$ and $\mathcal{M}$ is the class all distributions satisfying the margin condition with exponent $\alpha > 0$ and such that the regression function η belongs to a Hölder class of smoothness $\beta > 0$ (see Section 5):

$$\begin{aligned}
&AC_n(\mathcal{M}, \lambda) \ge p_0,\ 0 < \lambda \le D_1 n^{-\frac{1+\alpha}{2+\alpha+d/\beta}},\\
&C' \exp\{-c'n\lambda^{\frac{2+\alpha}{1+\alpha}}\} \le AC_n(\mathcal{M}, \lambda) \le C \exp\{-cn\lambda^{\frac{2+\alpha}{1+\alpha}}\},\\
&D_2 n^{-\frac{1+\alpha}{2+\alpha+d/\beta}} \le \lambda \le \lambda_0.
\end{aligned}$$

Here, $r = d/\beta$. As an immediate consequence of (2) – (4) we get the minimax rate for the excess risk:

$$\inf_{\hat{f}_n \in \mathbb{S}_n} \sup_{P \in \mathcal{M}} \left[\mathbb{E}R(\hat{f}_n) - R^*\right] \asymp n^{-\frac{1+\alpha}{2+\alpha+r}} \tag{5}$$

for appropriate classes $\mathcal{M}$, which implies the results previously obtained in Tsybakov [15] and Audibert and Tsybakov [1].

It is interesting to compare (2) – (4) to the results for the regression problem in a similar setting (see DeVore *et al.* [5] and Temlyakov [14]) since there are similarities and differences. Let us quote these former results: suppose, in a supervised learning setting, that we observe n i.i.d. observations of the pair (X, Y), but here Y is valued in $[-M, M]$ instead of $\{0, 1\}$ and we want to estimate

$$\xi(x) = \mathbb{E}(Y|X = x).$$

Let $\hat{\xi}_n(x)$ denote an estimator of $\xi(x)$ and consider the loss

$$\|\hat{\xi}_n - \xi\|_{\mathbb{L}_2(\mu_X)}.$$

Here and in what follows, $\|\cdot\|_{\mathbb{L}_p(\mu_X)}$, $p \geq 1$, denotes the $\mathbb{L}_p(\mu_X)$-norm with respect to the measure μ_X on $\mathcal{X}$. In this context, $AC_n(\mathcal{M}, \lambda)$ denotes the quantity

$$\inf_{\hat{\xi}_n} \sup_{P \in \mathcal{M}} \mathbb{P}\left(\|\hat{\xi}_n - \xi\|_{\mathbb{L}_2(\mu_X)} \geq \lambda\right).$$

It is proved in [5] and [14] that if $\mathcal{M} = \mathcal{M}(\Theta, \mu_X)$ is the set of probability measures having μ_X as marginal distribution and such that ξ belongs to the set Θ, and the entropy numbers of Θ with respect to $\mathbb{L}_2(\mu_X)$ are of order n^{-r} (see [5] and [14] for details), then there exist λ_n^-, λ_n^+, with $\lambda_n^- \asymp \lambda_n^+ \asymp n^{-r/(1+2r)}$, and constants δ_0, C_1, c_1, C_2, c_2 such that

$$AC_n(\mathcal{M}(\Theta, \mu_X), \lambda) \geq \delta_0, \qquad \forall\, \lambda \leq \lambda_n^-, \tag{6}$$

$$C_1 e^{-c_1 n \lambda^2} \leq AC_n(\mathcal{M}(\Theta, \mu_X), \lambda) \leq C_2 e^{-c_2 n \lambda^2}, \qquad \forall\, \lambda \geq \lambda_n^+. \tag{7}$$

These inequalities describe accurately the behavior of the minimax AC-function for classes $\mathcal{M}(\Theta, \mu_X)$ with any marginal distribution μ_X. The same inequalities hold for the following quantity

$$\sup_{\mu_X} AC_n(\mathcal{M}(\Theta, \mu_X), \lambda).$$

Our results for the classification problem are somewhat weaker than the above results for the regression problem. In Sections 3 and 4, we prove the upper bounds for the corresponding classes in the case of any marginal distribution μ_X such that the Margin assumption holds. This is analogous to what was obtained for the regression problem. However, in Section 5, we only prove the matching lower bounds for a special marginal distribution μ_X. Thus we obtain an accurate description of the behavior of the supremum over marginal distributions $\sup_{\mu_X} AC_n(\mathcal{M}, \lambda)$ and not of the individual AC-functions for each marginal distribution μ_X.

The similarity of the results in the two different settings is that there is a regime of exponential concentration, which holds for any λ greater than a critical level. This critical level, which is also the minimax rate, depends on the complexity of the class characterized by r. We can also observe that the exponents in the bounds ($\frac{2+\alpha}{1+\alpha}$ in classification, 2 in regression) do not depend on the complexity parameter r.

The differences lie in two facts since the margin condition is entering the game at two levels. The first one is the critical value itself, $n^{-\frac{1+\alpha}{2+\alpha+\tau}}$. Note that here α is appearing in a favorable way (the larger it is, the better the rate). This is intuitively clear since larger α correspond to sharper decision boundaries.

The second place where a difference occurs is the rate in the exponent $\lambda^{\frac{2+\alpha}{1+\alpha}}$ compared to λ^2 in a regression setting. The margin condition influences the rate $\frac{2+\alpha}{1+\alpha}$, and this time again in a favorable way with respect to α (the rate improves as α grows). For $\alpha \to 0$, that is, when there is no margin condition we approach the same rate as in regression.

2 Properties related to the Margin condition

In this section, we discuss some facts related to the Margin condition. We first recall that it can be equivalently defined in the following way, cf. [15]. Let $G_0 = \{x : \eta(x) = 1/2\}$ denote the decision boundary.

Proposition 1. *Fix $0 < \alpha < \infty$. A probability measure P satisfies the Margin condition (1) if there exists a positive constant c_M such that, for any Borel set $G \subset \mathcal{X}$,*

$$\int_G |2\eta(x) - 1| \mu_X(dx) \geq c_M \mu_X(G)^{\varkappa}, \tag{8}$$

where $\varkappa = (1+\alpha)/\alpha$. Conversely, if the Margin condition (1) holds, then there exists a positive constant c_M such that, for any Borel set $G \subset \mathcal{X}$,

$$\int_G |2\eta(x) - 1| \mu_X(dx) \geq c_M \mu_X(G \setminus G_0)^{\varkappa}. \tag{9}$$

Proof: We prove first (9). Let G be given. Clearly, it suffices to assume that $\mu_X(G \setminus G_0) > 0$. Choose t from the equation $\mu_X(G \setminus G_0) = 2C_M t^{\alpha}$, and set $A = \{x : |\eta(x) - 1/2| \leq t\}$. Then by the Margin condition (1),

$$\mu_X(G \setminus A) \geq \mu_X(G \setminus G_0) - \mu_X(A \setminus G_0) \geq \mu_X(G \setminus G_0) - C_M t^{\alpha} \geq C_M t^{\alpha}.$$

Therefore,

$$\begin{aligned} \int_G |2\eta(x) - 1| \mu_X(dx) &\geq 2 \int_{G \setminus A} t \mu_X(dx) \\ &\geq 2C_M t^{\alpha+1} = (2C_M)^{-1/\alpha} \mu_X(G \setminus G_0)^{1+1/\alpha}. \end{aligned}$$

Conversely, assume that for some $\varkappa > 1$ inequality (8) holds for any Borel set G. Take $G = \{x : 0 < |\eta(x) - 1/2| \le t\}$. Then (8) yields

$$\begin{aligned}\mu_X(0 < |\eta(X) - 1/2| \le t) &\le \left(c_M^{-1} \int_{0<|\eta(x)-1/2|\le t} |2\eta(x) - 1| \mu_X(dx)\right)^{1/\varkappa} \\ &\le \left(2c_M^{-1} t\, \mu_X(0 < |\eta(X) - 1/2| \le t)\right)^{1/\varkappa}.\end{aligned}$$

Solving this inequality with respect to $\mu_X(0 < |\eta(X) - 1/2| \le t)$ we obtain the Margin condition (1). □

Remark 1. *In what follows we will distinguish between the Margin condition (1) and the Margin condition (8) with $\varkappa \ge 1$. If $\varkappa = 1$, then (8) still makes sense but it cannot be directly linked to Margin condition (1) in terms of α; formally, one would set $\alpha = +\infty$ but this lacks rigor. As suggested in [12], it is more appropriate to define the analog of (1) for $\varkappa = 1$ in the form $\mu_X\big(0 < |\eta(X) - 1/2| \le t_0\big) = 0$, which means that the regression function η has a jump at the decision boundary. The case $\varkappa = 1$ will be treated separately in Section 4.*

We now state an easy consequence of Proposition 1. For any prediction rule f, set $f^*_{P,f} = f^*_P I_{\{\eta \ne 1/2\}} + f I_{\{\eta = 1/2\}}$.

Lemma 1. *If the probability measure P satisfies the Margin condition (1), then for any prediction rule f,*

$$R(f) - R^* \ge (2C_M)^{-1/\alpha} \|f - f^*_{P,f}\|^{\frac{1+\alpha}{\alpha}}_{L_1(\mu_X)}. \tag{10}$$

Analogously, if the probability measure P satisfies the Margin condition (8) with some $\varkappa \ge 1$, then for any prediction rule f,

$$R(f) - R^* \ge c_M \|f - f^*_P\|^{\varkappa}_{L_1(\mu_X)}. \tag{11}$$

Proof: Note that, for any prediction rule f,

$$R(f) - R^* = \int_{D_P(f)} |2\eta(x) - 1| \mu_X(dx) = \int_{D'_P(f)} |2\eta(x) - 1| \mu_X(dx),$$

where $D_P(f) \triangleq \{x : f^*_P(x) \ne f(x)\}$ and $D'_P(f) \triangleq \{x : f^*_{P,f}(x) \ne f(x)\}$. Thus, (10) follows from (9) and the relations $D'_P(f) \setminus G_0 = D'_P(f)$, and

$$\mu_X(D'_P(f)) = \|f - f^*_{P,f}\|_{L_1(\mu_X)}.$$

Similar argument for $D_P(f)$ together with (8) imply (11).

Finally, we have the following property.

Proposition 2. *For any Borel function* $\bar{\eta} : \mathcal{X} \to [0,1]$ *and any distribution* P *of* (X,Y) *satisfying the Margin condition (8) with some* $\varkappa > 1$*, we have*

$$\|f_{\bar{\eta}} - f_P^*\|_{L_1(\mu_X)} \le 2C_M \|\bar{\eta} - \eta_P\|^{\alpha}_{L_\infty(\mu_X)}$$

where $f_{\bar{\eta}}(x) = I_{\{\bar{\eta}(x) \ge 1/2\}}$ *and* $\alpha = (1-\varkappa)^{-1}$.

Proof: By Lemma 5.1 in [1],

$$R(f_{\bar{\eta}}) - R^* \le 2C_M \|\bar{\eta} - \eta_P\|^{1+\alpha}_{L_\infty(\mu_X)}. \tag{12}$$

This and Lemma 1 yield the result.

Corollary 1. *Let* $\mathcal{P}$ *be a class of joint distributions of* (X,Y) *satisfying the Margin condition (8) with some* $\varkappa > 1$ *and all having the same marginal* μ_X*. Then, for any pair* $P, \bar{P} \in \mathcal{P}$ *with the corresponding regression functions* $\eta, \bar{\eta}$ *and decision rules* $f_\eta(x) = I_{\{\eta(x) \ge 1/2\}}$*,* $f_{\bar{\eta}}(x) = I_{\{\bar{\eta}(x) \ge 1/2\}}$*, we have*

$$\|f_{\bar{\eta}} - f_\eta\|_{L_1(\mu_X)} \le 2C_M \|\bar{\eta} - \eta\|^{\alpha}_{L_\infty(\mu_X)}.$$

3 Upper bound under complexity assumption on the regression function

In this section, we prove an upper bound of the form (2) for a class of probability distributions P, for which the complexity assumption (b) (cf. the Introduction) is expressed in terms of the entropy of the class of underlying regression functions η_P.

For $g : \mathcal{X} \to \mathbb{R}$, define the sup-norm $\|g\|_\infty = \sup_{x \in \mathcal{X}} |g(x)|$.

Fix some positive constants r, α, C_M, B. Let $\mathcal{M}(r,\alpha) = \mathcal{M}(r,\alpha,C_M,B)$ be any set of joint distributions P of (X,Y) satisfying the following two conditions.

(i) The Margin condition (1) with exponent α *and constant* C_M.

(ii) The regression function $\eta = \eta_P$ *belongs to a known class of functions* $\mathcal{U}$*, which admits the* ε*-entropy bound*

$$\mathcal{H}(\varepsilon, \mathcal{U}, \|\cdot\|_\infty) \le B\varepsilon^{-r}, \ \forall \epsilon > 0. \tag{13}$$

Here, the ε*-entropy* $\mathcal{H}(\varepsilon, \mathcal{U}, \|\cdot\|_\infty)$ *is defined as the natural logarithm of the minimal number of* ε*-balls in the* $\|\cdot\|_\infty$ *norm needed to cover* $\mathcal{U}$*.*

For any prediction rule f, we define the empirical risk

$$R_n(f) = \frac{1}{n}\sum_{i=1}^{n} I_{\{f(X_i)\neq Y_i\}}.$$

We consider the classifier $\hat{f}_{n,1}(x) = I_{\{\hat{\eta}_n(x)\ge 1/2\}}$, where

$$\hat{\eta}_n = \operatorname{argmin}_{\eta' \in \mathcal{N}_\varepsilon} R_n(f_{\eta'}).$$

Here $f_{\eta'}(x) = I_{\{\eta'(x)\ge 1/2\}}$ and $\mathcal{N}_\varepsilon$ denotes a minimal ε-net on $\mathcal{U}$ in the $\|\cdot\|_\infty$ norm, i.e., $\mathcal{N}_\varepsilon$ is the minimal subset of $\mathcal{U}$ such that the union of ε-balls in the $\|\cdot\|_\infty$ norm centered at the elements of $\mathcal{N}_\varepsilon$ covers $\mathcal{U}$.

Theorem 1. *Let* r, α, C_M, B *be finite positive constants. Set* $\varepsilon = \varepsilon_n = n^{-\frac{1}{2+\alpha+r}}$*. Then there exist positive constants* c *and* c' *depending only on* r, α, C_M, B *such that*

$$\sup_{P\in\mathcal{M}(r,\alpha)} \mathbb{P}\{R(\hat{f}_{n,1}) - R(f_P^*) \ge \lambda\} \le 2\exp\{-cn\lambda^{\frac{2+\alpha}{1+\alpha}}\}$$

for $\lambda \ge c' n^{-\frac{1+\alpha}{2+\alpha+r}}$*.*

This theorem has an immediate consequence in terms of AC-functions.

Corollary 2. *There exist* $d > 0$*,* $c > 0$ *such that for* $\lambda_n = dn^{-\frac{1+\alpha}{2+\alpha+r}}$ *we have*

$$AC_n(\mathcal{M}(r,\alpha),\lambda) \le 2e^{-cn\lambda^{\frac{2+\alpha}{1+\alpha}}}, \qquad \forall\, \lambda \ge \lambda_n.$$

Proof of Theorem 1. We follow the argument of Theorem 4.2 in [1] with suitable modifications. Set $d(\eta') \triangleq R(f_{\eta'}) - R(f_P^*) \equiv R(f_{\eta'}) - R(f_{P,\eta'}^*)$ where, for brevity, $f_{P,\eta'}^* \triangleq f_{P,f_{\eta'}}^*$. Let $\bar{\eta} \in \mathcal{N}_\varepsilon$ be such that $\|\bar{\eta} - \eta_P\|_\infty \le \varepsilon$. Using Lemma 5.1 in [1], cf. (12) above, we get

$$d(\bar{\eta}) = R(f_{\bar{\eta}}) - R^* \le 2C_M\|\bar{\eta} - \eta_P\|_\infty^{1+\alpha} \le 2C_M\varepsilon^{1+\alpha} \le \lambda/2 \tag{14}$$

for any $\lambda \geq 4C_M n^{-\frac{1+\alpha}{2+\alpha+r}}$. Define a set of functions $\mathcal{G}_\varepsilon = \{\eta' \in \mathcal{N}_\varepsilon : d(\eta') \geq \lambda\}$, and introduce the centered empirical increments

$$\mathcal{Z}_n(\eta') = (R_n(f_{\eta'}) - R_n(f^*_{P,\eta'})) - (R(f_{\eta'}) - R(f^*_{P,\eta'})).$$

Then

$$\begin{aligned}
\mathbb{P}(R(\widehat{f}_{n,1}) - R(f^*_P) \geq \lambda) &\leq \mathbb{P}(\exists \eta' \in \mathcal{G}_\varepsilon : R_n(f_{\eta'}) - R_n(f_{\bar{\eta}}) \leq 0) \\
&\leq \sum_{\eta' \in \mathcal{G}_\varepsilon} \mathbb{P}(d(\eta') + \mathcal{Z}_n(\eta') - d(\bar{\eta}) - \mathcal{Z}_n(\bar{\eta}) \leq 0).
\end{aligned}$$

Note that for any $\eta' \in \mathcal{G}_\varepsilon$ we have

$$d(\eta') - d(\bar{\eta}) \geq d(\eta')/2 \geq \lambda/2.$$

Using this remark and (13) we find

$$\begin{aligned}
\mathbb{P}(R(\widehat{f}_{n,1}) - R(f^*_P) \geq \lambda) &\leq \sum_{\eta' \in \mathcal{G}_\varepsilon} \mathbb{P}(\mathcal{Z}_n(\eta') \leq -d(\eta')/4) \\
&\quad + \mathbb{P}(\mathcal{Z}_n(\bar{\eta}) \geq \lambda/4) \\
&\leq \exp(B\varepsilon^{-r}) \max_{\eta' \in \mathcal{G}_\varepsilon} \mathbb{P}(\mathcal{Z}_n(\eta') \leq -d(\eta')/4) \\
&\quad + \mathbb{P}(\mathcal{Z}_n(\bar{\eta}) \geq \lambda/4).
\end{aligned} \tag{15}$$

Now, $\mathcal{Z}_n(\eta') = \frac{1}{n}\sum_{i=1}^n \xi_i(\eta')$, where

$$\xi_i(\eta') = I_{\{f_{\eta'}(X_i) \neq Y_i\}} - I_{\{f^*_{P,\eta'}(X_i) \neq Y_i\}} - \mathbb{E}\Big(I_{\{f_{\eta'}(X_i) \neq Y_i\}} - I_{\{f^*_{P,\eta'}(X_i) \neq Y_i\}}\Big).$$

Clearly, $|\xi_i(\eta')| \leq 2$ and, using (10) of Lemma 1,

$$\begin{aligned}
\mathbb{E}(\xi_i(\eta')^2) &\leq \mathbb{E}\Big(\big[I_{\{f_{\eta'}(X_i) \neq Y_i\}} - I_{\{f^*_{P,\eta'}(X_i) \neq Y_i\}}\big]^2\Big) \\
&= \|f_{\eta'} - f^*_{P,\eta'}\|_{L_1(\mu_X)} \\
&\leq \big[(2C_M)^{1/\alpha}(R(f_{\eta'}) - R(f^*_P))\big]^{\frac{\alpha}{1+\alpha}} \\
&= (2C_M)^{\frac{1}{1+\alpha}}(d(\eta'))^{\frac{\alpha}{1+\alpha}}.
\end{aligned}$$

Therefore, we can apply Bernstein's inequality to get

$$\begin{aligned}
\mathbb{P}(\mathcal{Z}_n(\eta') \leq -d(\eta')/4) &\leq \exp\left(-\frac{nd^2(\eta')/16}{2((2C_M)^{\frac{1}{1+\alpha}}(d(\eta'))^{\frac{\alpha}{1+\alpha}} + d(\eta')/3)}\right) \\
&\leq \exp\left(-\frac{nd^2(\eta')}{c_1'(d(\eta'))^{\frac{\alpha}{1+\alpha}}}\right)
\end{aligned}$$

where $c_1' = 2((2C_M)^{\frac{1}{1+\alpha}} + 1/3)$ and we used that $d(\eta') \leq d^{\frac{\alpha}{1+\alpha}}(\eta')$ since $d(\eta') \leq 1$. Thus, for any $\eta' \in \mathcal{G}_\varepsilon$ we obtain

$$\mathbb{P}(\mathcal{Z}_n(\eta') \leq -d(\eta')/4) \leq \exp\left(-n\lambda^{\frac{2+\alpha}{1+\alpha}}/c_1'\right).$$

As a consequence,

$$\begin{aligned} \exp(B\varepsilon^{-r}) \max_{\eta' \in \mathcal{G}_\varepsilon} \mathbb{P}(\mathcal{Z}_n(\eta') \leq -d(\eta')/4) &\leq \exp(Bn^{\frac{r}{2+\alpha+r}} - n\lambda^{\frac{2+\alpha}{1+\alpha}}/c_1') \\ &\leq \exp(-n\lambda^{\frac{2+\alpha}{1+\alpha}}/2c_1') \qquad (16) \end{aligned}$$

where we used that $\lambda \geq c' n^{-\frac{1+\alpha}{2+\alpha+r}}$ for some large enough $c' > 0$. Another application of Bernstein's inequality and (14) yields

$$\begin{aligned} \mathbb{P}(\mathcal{Z}_n(\bar{\eta}) \geq \lambda/4) &\leq \exp\left(-\frac{n\lambda^2/16}{2((2C_M)^{\frac{1}{1+\alpha}}(d(\bar{\eta}))^{\frac{\alpha}{1+\alpha}} + \lambda/3)}\right) \\ &\leq \exp\left(-\frac{n\lambda^2}{c_1'(\lambda^{\frac{\alpha}{1+\alpha}} + \lambda)}\right). \end{aligned}$$

For $\lambda \leq 1$ the last inequality implies

$$\mathbb{P}(\mathcal{Z}_n(\bar{\eta}) \geq \lambda/4) \leq \exp\left(-\frac{n\lambda^{\frac{2+\alpha}{1+\alpha}}}{2c_1'}\right).$$

This, together with (15) and (16), yields result of the theorem for $\lambda \leq 1$. If $\lambda > 1$ it holds trivially since $d(\eta') \leq 1$ for all η'.

4 Upper bound under complexity assumption on the Bayes classifier

This section provides a result analogous to that of Section 3 when the complexity assumption (b) (cf. the Introduction) is expressed in terms of the entropy of the class of underlying Bayes classifiers f_P^* rather than of that of regression functions η_P.

First, introduce some definitions. Let $\mathcal{F}$ be a class of measurable functions from a measurable space $(S, \mathcal{A}_S, \mu)$ into $[0, 1]$. Here μ is a σ-finite measure.

For $1 \leq q \leq \infty$, and $\varepsilon > 0$, let $N_{[\,]}(\varepsilon, \mathcal{F}, \|\cdot\|_{L_q(\mu)})$ denote the $L_q(\mu)$-bracketing numbers of $\mathcal{F}$. That is, $N_{[\,]}(\varepsilon, \mathcal{F}, \|\cdot\|_{L_q(\mu)})$ is the minimal number N of functional brackets

$$[f_j^-, f_j^+] \triangleq \{g : f_j^- \leq g \leq f_j^+\},\ j = 1, \ldots, N,$$

such that

$$\mathcal{F} \subset \bigcup_{j=1}^{N} [f_j^-, f_j^+] \quad \text{and} \quad \|f_j^+ - f_j^-\|_{L_q(\mu)} \leq \varepsilon,\ j = 1, \ldots, N.$$

The bracketing ε-entropy of $\mathcal{F}$ in the $\|\cdot\|_{L_q(\mu)}$-norm is defined by

$$\mathcal{H}_{[\,]}(\varepsilon, \mathcal{F}, \|\cdot\|_{L_q(\mu)}) \triangleq \log N_{[\,]}(\varepsilon, \mathcal{F}, \|\cdot\|_{L_q(\mu)}).$$

We will consider a class of probability distributions P of (X, Y) characterized by the complexity of the corresponding Bayes classifiers. Specifically, fix some $\rho \in (0, 1), \alpha > 0, C_M > 0, c_\mu > 0, B' > 0$, and let $\mathcal{M}^*(\rho, \alpha) = \mathcal{M}^*(\rho, \alpha, C_M, c_\mu, B')$ be any set of joint distributions P of (X, Y) satisfying the following conditions.

(i) The marginal distribution μ_X of X is absolutely continuous with respect to a σ-finite measure μ on $(\mathcal{X}, \mathcal{A})$, and $(d\mu_X/d\mu)(x) \leq c_\mu$ for μ-almost all $x \in \mathcal{X}$.

(ii) The Margin condition (8) with exponent $\varkappa = (1 + \alpha)/\alpha$, $0 < \alpha < \infty$, and constant c_M is satisfied.

(iii) The Bayes classifier f_P^* belongs to a known class of prediction rules $\mathcal{F}$ satisfying the bracketing entropy bound

$$\mathcal{H}_{[\,]}(\varepsilon, \mathcal{F}, \|\cdot\|_{L_1(\mu)}) \leq B' \varepsilon^{-\rho},\ \forall \varepsilon > 0. \tag{17}$$

We consider a classifier $\hat{f}_{n,2}$ that minimizes the empirical risk over the class $\mathcal{F}$:

$$\hat{f}_{n,2} \triangleq \operatorname{argmin}_{f \in \mathcal{F}} R_n(f).$$

The main result of this section is that for $\hat{f}_{n,2}$ we have the following exponential upper bound.

Theorem 2. *Let* $\rho \in (0,1)$*, and let* α, c_M, c_μ, B' *be positive constants. Then there exist positive constants* c *and* c' *depending only on* $\rho, \alpha, c_M, c_\mu, B'$ *such that*

$$\sup_{P \in \mathcal{M}^*(\rho,\alpha)} \mathbb{P}\{R(\widehat{f}_{n,2}) - R(f_P^*) \geq \lambda\} \leq e \exp\{-cn\lambda^{\frac{2+\alpha}{1+\alpha}}\}$$

for $\lambda \geq c' n^{-\frac{1+\alpha}{2+\alpha(1+\rho)}}$.

Furthermore, the same classifier $\widehat{f}_{n,2}$ satisfies a similar exponential bound under the Margin condition (8) with $\varkappa = 1$. To consider this case, we define the class of distributions $\mathcal{M}^*(\rho,\infty) = \mathcal{M}^*(\rho, c_M, c_\mu, B')$ as being a set of joint distributions P of (X,Y) satisfying the same assumptions as $\mathcal{M}^*(\rho,\alpha), 0 < \alpha < \infty$, with the only difference that assumption (ii) is replaced by

(ii') *The Margin condition (8) with exponent* $\varkappa = 1$ *and constant* $c_M > 0$ *is satisfied.*

The upper bound for this class is as follows.

Theorem 3. *Let* $\rho \in (0,1)$*, and let* c_M, c_μ, B' *be positive constants. Then there exist positive constants* c *and* c' *depending only on* ρ, c_M, c_μ, B' *such that*

$$\sup_{P \in \mathcal{M}^*(\rho,\infty)} \mathbb{P}\{R(\widehat{f}_{n,2}) - R(f_P^*) \geq \lambda\} \leq e \exp\{-cn\lambda\}$$

for $\lambda \geq c' n^{-\frac{1}{1+\rho}}$.

The proof of Theorems 2 and 3 is given in the Appendix. Note that this proof can be deduced in several different ways from well known general excess risk bounds in learning theory (see, e.g., Massart [11] or Koltchinskii [10] and references therein). The version of the proof given below follows [10].

Inspection of the proof shows that Theorems 2 and 3 remain valid if we drop condition (i) and replace (iii) by the following more general condition:

(iii') *The Bayes classifier* f_P^* *belongs to a known class of prediction rules* $\mathcal{F}$ *satisfying the bracketing entropy bound*

$$\mathcal{H}_{[\,]}(\varepsilon, \mathcal{F}, \|\cdot\|_{L_1(\mu_X)}) \leq B' \varepsilon^{-\rho}, \ \forall \varepsilon > 0.$$

Condition (iii') is, in fact, an assumption on both $\mathcal{F}$ and the class of possible marginal densities μ_X. The reason why we have introduced conditions (i)

and (iii) instead of (iii') is that they are easily interpretable. Indeed, in this way we decouple assumptions on $\mathcal{F}$ and assumptions on μ_X. The case that is even easier corresponds to considering a subclass of $\mathcal{M}^*(\rho, \alpha)$ composed of measures $P \in \mathcal{M}^*(\rho, \alpha)$ with the same marginal μ_X. Then again we only need to assume (ii) and (iii') but now (iii') should hold for one fixed measure μ_X and not simultaneously for a set of possible marginal measures.

We finish this section by a comparison of Theorems 1 and 2. They differ in imposing entropy assumptions on different objects, regression function η_P and Bayes classifier f_P^* respectively. Also, in Theorem 1 the complexity is measured by the usual entropy for the sup-norm, whereas in Theorem 2 it is done in terms of the bracketing entropy for the L_1-norm. Note that for many classes the bracketing and the usual ε-entropies behave similarly, so that the relationship between the corresponding rates of decay r in (13) and ρ in (17) is only determined by the relationship between the sup-norm of the regression function η and the L_1-norm on the induced Bayes classifier. In this respect, Corollary 1 is insightful suggesting the correspondence

$$\rho = \frac{r}{\alpha}\,.$$

Finally, note that the ranges of the complexity parameters as well as the assumptions on the measure μ_X in Theorems 1 and 2 are somewhat different. Namely, Theorem 1 holds under no additional assumption on μ_X except for the Margin condition and covers classes with high complexity (all $r > 0$ are allowed). Theorem 2 needs a relatively mild additional assumption (i) on μ_X and restricts the complexity by the condition $\rho < 1$. The classifier $\widehat{f}_{n,2}$ of Theorem 2 does not require the knowledge of the margin parameter α. Thus, $\widehat{f}_{n,2}$ is adaptive to the margin parameter. On the other hand, the classifier $\widehat{f}_{n,1}$ of Theorem 1 does require the knowledge of α which is involved in the definition of parameter ε of the net $\mathcal{N}_\varepsilon$. Note that for classes $\mathcal{F}$ of high complexity (with $\rho > 1$) the empirical risk minimization over the whole class $\mathcal{F}$ usually does not provide optimal convergence rates. In such cases, some form of regularization is needed. It could be based on penalized empirical risk minimization (see, e.g., [10]) over proper sieves of subclasses of $\mathcal{F}$ (for instance, sieves of ε-nets for $\mathcal{F}$).

5 Minimax lower bounds

In this section, we will assume that the regression function η belongs to a Hölder class defined as follows.

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

For $\beta > 0$, let $\lfloor \beta \rfloor$ be the maximal integer that is strictly less than β. For any $x \in [0,1]^d$ and any $\lfloor \beta \rfloor$ times continuously differentiable real valued function g on $[0,1]^d$, we denote by g_x its Taylor polynomial of degree $\lfloor \beta \rfloor$ at point $x \in [0,1]^d$:

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

Let $\beta > 0$, $L > 0$. The *Hölder class* of functions $\Sigma(\beta, L, [0,1]^d)$ is defined as the set of all functions $g : [0,1]^d \to \mathbb{R}$ that are $\lfloor \beta \rfloor$ times continuously differentiable and satisfy, for any $x, y \in [0,1]^d$, the inequality

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

Fix $\alpha > 0, \beta > 0, L > 0$, and a probability distribution μ_X on $[0,1]^d$. Denote by $M'(\mu_X, \alpha, \beta)$ the class of all joint distributions P of (X, Y) such that:

(i) The marginal distribution of X *is* μ_X*;*

(ii) The Margin condition (1) is satisfied with some constant $C_M > 0$*;*

(iii) The regression function $\eta = \eta_P$ *belongs to the Hölder class* $\Sigma(\beta, L, [0,1]^d)$*.*

Theorem 4. *There exist a marginal distribution* μ_X^* *and positive constants* C_1', C_2', c', d_1' , d_2', λ_0' *depending only on* α, β, L, d*, and* C_M *such that for any classifier* $\hat{f}_n$*,*

$$\sup_{P \in \mathcal{M}'(\mu_X^*, \alpha, \beta)} \mathbb{P}\{R(\hat{f}_n) - R^* \ge \lambda\} \ge C_1'$$

for any $0 < \lambda \le d_1' n^{-\frac{1+\alpha}{2+\alpha+d/\beta}}$*, and*

$$\sup_{P \in \mathcal{M}'(\mu_X^*, \alpha, \beta)} \mathbb{P}\{R(\hat{f}_n) - R^* \ge \lambda\} \ge C_2' \exp\{-c' n \lambda^{\frac{2+\alpha}{1+\alpha}}\}$$

for any $d_2' n^{-\frac{1+\alpha}{2+\alpha+d/\beta}} \le \lambda \le \lambda_0'$*.*

The proof of Theorem 4 including the explicit form of the marginal distribution μ_X^* is given in Section 6.

Note that there exists a constant $B > 0$ such that the set of regression functions $\mathcal{U} = \{\eta_P,\ P \in \mathcal{M}'(\mu_X^*, \alpha, \beta)\}$ satisfies the entropy bound

$$\mathcal{H}(\varepsilon, \mathcal{U}, \|\cdot\|_\infty) \le B\varepsilon^{-r},\ \forall \varepsilon > 0, \tag{18}$$

where $r = d/\beta$. Indeed, (18) holds since $\mathcal{U} = \{\eta \in \Sigma(\beta, L, [0,1]^d) : 0 \le \eta(x) \le 1\}$, and

$$\mathcal{H}(\varepsilon, \Sigma(\beta, L, [0,1]^d), \|\cdot\|_\infty) \le B\varepsilon^{-d/\beta},$$

cf. Kolmogorov and Tikhomirov [8]. Thus, for the choice of μ_X^* described in Section 6, the class of probability distributions $\mathcal{M}'(\mu_X^*, \alpha, \beta)$ is a particular case of $\mathcal{M}(r, \alpha)$ (with $r = d/\beta$) defined in Section 3. Theorem 4 shows that, for this particular case, it is impossible to obtain faster rates for AC-functions than those established in Theorem 1. In this sense, Theorem 4 provides a lower bound that matches the upper bound of Theorem 1.

6 Proof of Theorem 4

The proof of Theorem 4 will be divided in steps. First, we construct a finite family $P_1, \dots, P_N$ of probability distributions of the pair (X, Y). Second, we apply the general tools for minimax lower bounds (cf. the Appendix) to prove a minimax lower bound on this finite family. Finally, we choose the parameters of the family $P_1, \dots, P_N$ in order to embed it into the class $\mathcal{M}'(\mu_X^*, \alpha, \beta)$, which leads to the result of Theorem 4.

Construction of a finite family of probability measures. We proceed here similarly to [1]. Let $\sigma = (\sigma_1, \dots, \sigma_b)$ be a binary vector of length b with elements $\sigma_j \in \{-1, 1\}$. Let φ be an infinitely differentiable function with compact support in $\mathbb{R}^d$ such that $0 \le \varphi(x) \le c$ for some constant $c \in (0, 1/2)$. Consider functions $\varphi_1, \dots, \varphi_b$ on $\mathbb{R}^d$ satisfying:

a) φ_j is a shift of φ, $j = 1, \dots, b$,

b) the supports Δ_j of functions φ_j are disjoint.

Denote by $\Sigma(b)$ the set of all binary vectors σ of length b. For every $\sigma \in \Sigma(b)$ define

$$\phi_\sigma(x) \triangleq \sum_{j=1}^{b} \sigma_j \varphi_j(x), \quad \eta_\sigma(x) \triangleq (1 + \phi_\sigma(x))/2.$$

Consider the following class Θ of regression functions

$$\Theta \triangleq \{\eta_\sigma, \sigma \in \Sigma(b)\}.$$

In what follows we assume without loss of generality that $b \geq 16$. By the Varshamov-Gilbert lemma (cf. [16], p. 104), there is a subset S of $\Sigma(b)$ such that cardinality $|S| \geq 2^{b/8}$, and for any two different elements σ and σ' from S we have

$$\|\sigma - \sigma'\|_{\ell_1} \geq b/4.$$

Let $\mathcal{X} = [0,1]^d$, $q \in \mathbb{N}$, and $b \triangleq q^d$. Let ψ be a nonnegative infinitely differentiable function with support $(0,1)^d$ such that $\psi \leq c < 1/2$ and $\int_{(0,1)^d} \psi(x)dx > 0$. For given parameters $\delta \in (0,1)$ (small parameter) and $\alpha \in [0,\infty)$, define

$$\varphi(x) \triangleq \delta^{1/(1+\alpha)}\psi(qx).$$

For a vector $k = (k_1, \ldots, k_d)$, $k_j \in \{0, \ldots, q-1\}$, $j = 1, \ldots, d$, define a grid point

$$x^k \triangleq (x_1^k, \ldots, x_d^k), \quad x_j^k = k_j/q, \ j = 1, \ldots, d.$$

We now consider b functions $\varphi_k(x) = \varphi(x - x^k)$ and the corresponding class Θ of regression functions defined above. We set $N \triangleq |S|$ and consider a subset $\Theta' \subset \Theta$:

$$\Theta' \triangleq \{\eta_\sigma, \sigma \in S\} = \{\eta_i\}_{i=1}^N.$$

Now, recalling that the regression function $\eta(X)$ is the conditional probability of $Y = 1$ given X, we define the joint probability measures P_σ, $\sigma \in S$, of (X, Y) (these measures will be also denoted by $P_i, i = 1, \ldots, N$) :

$$P_\sigma(Y = 1, X \in A) = \int_A \eta_\sigma(x)\mu_X(dx)$$

for any Borel set A, where the marginal distribution $\mu_X = \mu_X^*$ is specified as follows. First, for all x such that

$$1/(4q) \leq x_j - x_j^k \leq 3/(4q), \quad j = 1, \ldots, d,$$

the distribution μ_X^* has a density w.r.t. the Lebesgue measure

$$\frac{d\mu_X^*}{dx}(x) \triangleq \frac{w}{\mathrm{Leb}(B(0, 1/(4q)))} = 2^d b w$$

where $B(x,r)$ is the ℓ_∞-ball of radius r centered at x, $\mathrm{Leb}(\cdot)$ denotes the Lebesgue measure, and $w = C\delta^{\alpha/(1+\alpha)}/b$ for some $C \in (0,1]$. Second, we set $d\mu_X^*(x)/dx = 0$ for all other x such that at least one of $\eta_i(x)$ is not $1/2$. Finally, on the complementary set $A_0 \subset [0,1]^d$ where all $\eta_i(x)$ are equal to $1/2$, we set $d\mu_X^*(x)/dx \triangleq (1-bw)/\mathrm{Leb}(A_0)$ to ensure that $\int_{\mathbb{R}^d} d\mu_X^*(x) = 1$ (we assume that the support of the function ψ belongs to the set $[\gamma, 1-\gamma]$ for a small $\gamma > 0$; then, it is easy to see that $\mathrm{Leb}(A_0) > 0$).

We now impose an extra restriction on φ and prove that under this restriction the measures P_i satisfy the Margin condition with parameter α. Assume that $\psi(x) = c_2 > 0$ for x satisfying the inequalities $1/4 \le x_j \le 3/4$, $j = 1, \dots, d$, and $\psi(x) < c_2$ for other x. Here $c_2 \in (0, 1/2)$. Then

$$\begin{aligned}\mu_X^*(0 < |\eta_\sigma(X) - 1/2| \le t) &= \mu_X^*(0 < |\sum_{j=1}^{b} \sigma_j \varphi_j(X)| \le 2t) \\ &= b\mu_X^*(0 < \varphi(X) \le 2t),\end{aligned}$$

because the supports Δ_j of functions φ_j are disjoint. Then, using the definition $\varphi(x) \triangleq \delta^{1/(1+\alpha)}\psi(qx)$ we obtain that

$$\mu_X^*(0 < \varphi(X) \le 2t) = w \quad \text{if} \quad c_2\delta^{1/(1+\alpha)} \le 2t$$

and $\mu_X^*(0 < \varphi(X) \le 2t) = 0$ for all other $t > 0$. Therefore,

$$b\mu_X^*(0 < \varphi(X) \le 2t) \le C\delta^{\alpha/(1+\alpha)} I_{\{c_2\delta^{1/(1+\alpha)} \le 2t\}} \le C(2t/c_2)^\alpha, \quad t > 0.$$

Thus, all P_i satisfy the Margin condition with parameter α and constant $C_M = C(2/c_2)^\alpha$.

Minimax lower bound for the finite set of measures $P_1, \dots, P_N$. Let us check the assumptions of Theorem 5 in the Appendix for the set of probability measures $P_1, \dots, P_N$ defined above. Since $0 < c < 1/2$ we have $1/4 \le \eta_i(x) \le 3/4$ for all $\delta \in (0,1)$ and all $x \in (0,1)^d$. Next, for any $\sigma, \sigma' \in S$ we have

$$\|\eta_{P_\sigma} - \eta_{P_{\sigma'}}\|^2_{L_2(\mu_X^*)} \le b\|\varphi\|^2_{L_\infty(\mu_X^*)} w \le C\delta^{(2+\alpha)/(1+\alpha)},$$

and for $\sigma \ne \sigma'$, in view of (6) and (6),

$$\begin{aligned}\|f^*_{P_\sigma} - f^*_{P_{\sigma'}}\|_{L_1(\mu_X^*)} &= 2\sum_{j=1}^{b} I_{\{\sigma_j \ne \sigma'_j\}} \int_{B(0,1/(4q))} 2^d b w \, dx \\ &= \|\sigma - \sigma'\|_{\ell_1} w \ge c_1 \delta^{\alpha/(1+\alpha)},\end{aligned}$$

where $c_1 = C/4$. Thus, the assumptions of Theorem 5 in the Appendix are satisfied with $N = |S| \geq 2^{b/8} \geq 2^{b/16} + 1$, and

$$\gamma^2 = C\delta^{(2+\alpha)/(1+\alpha)}, \qquad s = c_1\delta^{\alpha/(1+\alpha)}. \tag{19}$$

Therefore, we get the following result.

Proposition 3. *Fix $\alpha > 0$, $\delta \in (0,1)$ and $q \in \mathbb{N}$ such that $b = q^d \geq 16$. Let $P_1, \ldots, P_N$ be the family of probability measures defined above. Then for any classifier $\hat{f}_n$ we have*

$$\max_{1\leq k\leq N} \mathbb{P}_k\Big\{\|\hat{f}_n - f^*_{P_k}\|_{L_1(\mu^*_X)} \geq \frac{C\delta^{\frac{\alpha}{1+\alpha}}}{8}\Big\} \geq \frac{1}{12}\min(1, 2^{\frac{b}{16}}\exp\{-c_3 n\delta^{\frac{2+\alpha}{1+\alpha}}\}) \tag{20}$$

where $C \in (0,1)$ is the constant used in the construction of $P_1, \ldots, P_N$, and $c_3 > 0$ is a constant depending only on C. Furthermore, for $0 < \lambda < \lambda_0$,

$$\max_{1\leq k\leq N} \mathbb{P}_k\{R(\hat{f}_n) - R(f^*_{P_k}) \geq \lambda\} \geq \frac{1}{12}\min(1, 2^{\frac{b}{16}}\exp\{-c_4 n\lambda^{\frac{2+\alpha}{1+\alpha}}\}) \tag{21}$$

where $\lambda_0 = 16^{-(1+\alpha)/\alpha}Cc_2$, and $c_4 > 0$ is a constant depending only on C, c_2 and α.

Proof: Bound (20) follows from Theorem 5 and (19). To prove (21), we combine (20) with Lemma 1, set $\lambda = \lambda_0\delta$, and use that $C_M = C(2/c_2)^\alpha$ by the construction of $P_1, \ldots, P_N$.

Minimax lower bound on the class $\mathcal{M}'(\mu^*_X, \alpha, \beta)$. We now prove Theorem 4 using a particular instance of the constructions introduced above in this section. Set $q = \lceil c_5\delta^{-\frac{1}{(1+\alpha)\beta}}\rceil$ where $c_5 > 0$ is a constant, and $\lceil x\rceil$ denotes the minimal integer greater than x. It is easy to see that if c_5 is small enough, then we have $\varphi \in \Sigma(\beta, L, [0,1]^d)$ implying that $\eta_\sigma \in \Sigma(\beta, L, [0,1]^d)$ for all $\sigma \in S$. Choose such a small c_5. It is also easy to see that one can always choose constants $C \in (0,1)$ and $c_2 \in (0, 1/2)$ in the construction of Section 6 in such a way that $C(2/c_2)^\alpha = C_M$ which is needed to satisfy the margin condition (ii). Then, for any fixed $\delta \in (0,1)$, the finite family of probability distributions $\{P_1, \ldots, P_N\}$ constructed above (and depending on δ) belongs to $\mathcal{M}'(\mu^*_X, \alpha, \beta)$. To indicate this dependence on δ explicitly, denote this family by $\mathcal{P}_\lambda$ where $\lambda = \lambda_0\delta$ and λ_0 is defined in Proposition 3. Since $\mathcal{P}_\lambda \subset \mathcal{M}'(\mu^*_X, \alpha, \beta)$, for any $\lambda < \lambda_0$ we can write

$$\sup_{P\in\mathcal{M}'(\mu^*_X,\alpha,\beta)} \mathbb{P}\{R(\hat{f}_n) - R^* \geq \lambda\} \geq \max_{P\in\mathcal{P}_\lambda} \mathbb{P}\{R(\hat{f}_n) - R^* \geq \lambda\}$$

and then estimate the right hand side of this inequality using (21) of Proposition 3. Note that in Proposition 3 we have the assumption $q^d \geq 16$, which is satisfied if $\delta \leq \delta_0$ where δ_0 is a small enough constant depending only on the constants in the definition of the class $\mathcal{M}'(\mu_X^*, \alpha, \beta)$. Thus we obtain

$$\begin{aligned}\sup_{P \in \mathcal{M}'(\mu_X^*, \alpha, \beta)} \mathbb{P}\{R(\widehat{f}_n) - R^* \geq \lambda\} &\geq \frac{1}{12} \min(1, 2^{b/16} \exp\{-c_4 n \lambda^{\frac{2+\alpha}{1+\alpha}}\}) \\ &\geq \frac{1}{12} \min(1, \exp\{c_6 \lambda^{-\frac{d}{(1+\alpha)\beta}} - c_4 n \lambda^{\frac{2+\alpha}{1+\alpha}}\})\end{aligned}$$

for all $0 < \lambda < \lambda_0'$ where $\lambda_0' > 0$ and $c_6 > 0$ depend only on the constants in the definition of the class $\mathcal{M}'(\mu_X^*, \alpha, \beta)$. This immediately implies the theorem. □

APPENDIX

Proof of Theorems 2 and 3. We deduce Theorems 2 and 3 from the following fact that we state here as a proposition.

Proposition 4. *Let either* $0 < \alpha < \infty$ *and* $\varkappa = \frac{1+\alpha}{\alpha}$ *or* $\alpha = \infty$ *and* $\varkappa = 1$. *Then there exists a constant* $C_* > 0$ *such that, for all* $t > 0$,

$$\sup_{P \in \mathcal{M}^*(\rho, \alpha)} \mathbb{P}\left\{ R(\widehat{f}_{n,2}) - R(f_P^*) \geq C_* \left[n^{-\frac{\varkappa}{2\varkappa - 1 + \rho}} \vee \left(\frac{t}{n}\right)^{\frac{\varkappa}{2\varkappa - 1}} \right] \right\} \leq e^{1-t}.$$

It is easy to see that Theorem 2 follows from this proposition by taking $t = cn\lambda^{\frac{2+\alpha}{1+\alpha}}$ with $\lambda \geq c' n^{-\frac{1+\alpha}{2+\alpha(1+\rho)}}$ for some constants $c, c' > 0$, and using that $\varkappa = \frac{1+\alpha}{\alpha}$. To obtain Theorem 3, we take $t = cn\lambda$ with $\lambda \geq c' n^{-\frac{1}{1+\rho}}$.

Proposition 4 will be derived from a general excess risk bound in abstract empirical risk minimization ([10], Theorem 4.3). We will state this result here for completeness. To this end, we need to introduce some notation. Let $\mathcal{G}$ be a class of measurable functions from a probability space $(S, \mathcal{A}_S, P)$ into $[0, 1]$ and let $Z_1, \ldots, Z_n$ be i.i.d. copies of an observation Z sampled from P. For any probability measure P and any $g \in \mathcal{G}$, introduce the following notation for the expectation:

$$Pg = \int_S g dP.$$

Denote by P_n the empirical measure based on $(Z_1, \ldots, Z_n)$, and consider the minimizer of the empirical risk

$$\hat{g}_n \triangleq \operatorname{argmin}_{g \in \mathcal{G}} P_n g.$$

For a function $g \in \mathcal{G}$, define the excess risk

$$\mathcal{E}_P(g) \triangleq Pg - \inf_{g' \in \mathcal{G}} Pg'.$$

The set

$$\mathcal{F}_P(\delta) \triangleq \{g \in \mathcal{G} : \mathcal{E}_P(g) \leq \delta\}$$

is called the δ-minimal set. The size of such a set will be controlled in terms of its $L_2(P)$-diameter

$$D(\delta) \triangleq \sup_{g, g' \in \mathcal{F}_P(\delta)} \|g - g'\|_{L_2(P)}$$

and also in terms of the following "localized empirical complexity":

$$\phi_n(\delta) \triangleq \mathbb{E} \sup_{g, g' \in \mathcal{F}_P(\delta)} |(P_n - P)(g - g')|.$$

We will use these complexity measures to construct an upper confidence bound on the excess risk $\mathcal{E}_P(\hat{f}_{n,2})$. For a function $\psi : \mathbb{R}_+ \mapsto \mathbb{R}_+$, define

$$\psi^\flat(\delta) \triangleq \sup_{\sigma \geq \delta} \frac{\psi(\sigma)}{\sigma}.$$

Let

$$V_n^t(\delta) \triangleq 4\left[\phi_n^\flat(\delta) + \sqrt{(D^2)^\flat(\delta)\frac{t}{n\delta}} + \frac{t}{n\delta}\right], \ \delta > 0, t > 0,$$

and define

$$\sigma_n^t \triangleq \inf\{\sigma : V_n^t(\sigma) \leq 1\}.$$

The following result is the first bound of Theorem 4.3 in [10].

Proposition 5. *For all* $t > 0$,

$$\mathbb{P}\{\mathcal{E}_P(\hat{f}_{n,2}) > \sigma_n^t\} \leq e^{1-t}.$$

In addition to this, we will use the well-known inequality for the expected sup-norm of the empirical process in terms of bracketing entropy, see Theorem 2.14.2 in [18]. More precisely, we will need the following simplified version of that result.

Lemma 2. *Let $\mathcal{T}$ be a class of functions from S into $[0,1]$ such that $\|g\|_{L_2(P)} \le a$ for all $g \in \mathcal{T}$. Assume that $H_{[\,]}(a, \mathcal{T}, \|\cdot\|_{L_2(P)}) + 1 \le a^2 n$. Then*

$$\mathbb{E} \sup_{g \in \mathcal{T}} |P_n g - Pg| \le \frac{\bar{C}}{\sqrt{n}} \int_0^a \left(H_{[\,]}(\varepsilon, \mathcal{T}, \|\cdot\|_{L_2(P)}) + 1 \right)^{1/2} d\varepsilon,$$

where $\bar{C} > 0$ is a universal constant.

Proof of Proposition 4. Note that, if $t > n$, then $(\frac{t}{n})^{\varkappa/(2\varkappa-1)} > 1$, and the result holds trivially with $C_* = 1$ since $R(\hat{f}_{n,2}) - R(f_P^*) \le 1$. Thus, it is enough to consider the case $t \le n$.

Let $S = \mathcal{X} \times \{0,1\}$ and P be the distribution of $Z = (X, Y)$. We will apply Proposition 5 to the class $\mathcal{G} \triangleq \{g_f : g_f(x,y) = I_{\{y \ne f(x)\}},\ f \in \mathcal{F}\}$. Then, clearly, $Pg_f = R(f)$ and $\mathcal{E}_P(g_f) = R(f) - R(f_P^*)$ for $g_f(x,y) = I_{\{y \ne f(x)\}}$, which implies that

$$\mathcal{F}_P(\delta) = \{g_f : f \in \mathcal{F},\ R(f) - R(f_P^*) \le \delta\}.$$

We also have $\|g_{f_1} - g_{f_2}\|_{L_2(P)}^2 = \|f_1 - f_2\|_{L_1(\mu_X)}$. Thus, it follows from Lemma 1 that, for all $g_f \in \mathcal{G}$,

$$\mathcal{E}_P(g_f) \ge c_M \|g_f - g_{f_P^*}\|_{L_2(P)}^{2\varkappa}$$

and we get a bound on the $L_2(P)$-diameter of the δ-minimal set $\mathcal{F}_P(\delta)$: with some constant $\bar{c}_1 > 0$

$$D(\delta) \le \bar{c}_1 \delta^{1/(2\varkappa)}. \tag{22}$$

To bound the function $\phi_n(\delta)$, we will apply Lemma 2 to the class $\mathcal{T} = \mathcal{F}_P(\delta)$ with $a = 1$. Note that

$$\begin{aligned} H_{[\,]}(\varepsilon, \mathcal{F}_P(\delta), \|\cdot\|_{L_2(P)}) &\le 2H_{[\,]}(\varepsilon/2, \mathcal{G}, \|\cdot\|_{L_2(P)}) \\ &\le 2H_{[\,]}(\varepsilon^2/4, \mathcal{F}, \|\cdot\|_{L_1(\mu_X)}) \\ &\le 2H_{[\,]}(\varepsilon^2/(4c_\mu), \mathcal{F}, \|\cdot\|_{L_1(\mu)}). \end{aligned}$$

Using (17) we easily get from Lemma 2 that, with some constants $\bar{c}_2, \bar{c}_3 > 0$,

$$\phi_n(\delta) \le \bar{c}_2 \delta^{\frac{1-\rho}{2\varkappa}} n^{-1/2}, \quad \delta \ge \bar{c}_3 n^{-\frac{\varkappa}{1+\rho}},$$

which implies that, with some constant $\bar{c}_4 > 0$,

$$\phi_n(\delta) \le \bar{c}_4 \max(\delta^{\frac{1-\rho}{2\varkappa}} n^{-1/2}, n^{-\frac{1}{1+\rho}}), \delta > 0.$$

This and (22) lead to the following bound on the function $V_n^t(\delta)$:

$$V_n^t(\delta) \le \bar{c}_5 \left[\delta^{\frac{1-\rho}{2\varkappa}-1} n^{-1/2} \vee \delta^{-1} n^{-\frac{1}{1+\rho}} + \delta^{\frac{1}{2\varkappa}-1} \sqrt{\frac{t}{n}} + \delta^{-1} \frac{t}{n} \right]$$

that holds with some constant $\bar{c}_5$. Thus, we end up with a bound on σ_n^t :

$$\sigma_n^t \le \bar{c}_6 \left[n^{-\frac{\varkappa}{2\varkappa-1+\rho}} \vee n^{-\frac{1}{1+\rho}} \vee \left(\frac{t}{n}\right)^{\varkappa/(2\varkappa-1)} \vee \frac{t}{n} \right]. \tag{23}$$

Note that, for $\varkappa \ge 1$, $\rho < 1$ and $t \le n$, we have

$$n^{-\varkappa/(2\varkappa-1+\rho)} \ge n^{-1/(1+\rho)} \quad \text{and} \quad \left(\frac{t}{n}\right)^{\varkappa/(2\varkappa-1)} \ge \frac{t}{n}.$$

Therefore, (23) can be simplified as follows:

$$\sigma_n^t \le \bar{c}_7 \left[n^{-\frac{\varkappa}{2\varkappa-1+\rho}} + \left(\frac{t}{n}\right)^{\varkappa/(2\varkappa-1)} \right],$$

and the result immediately follows from Proposition 5. □

Tools for the minimax lower bounds. For two probability measures μ and ν on a measurable space $(\mathcal{X}, \mathcal{A})$, we define the Kullback-Leibler divergence and the χ^2-divergence as follows:

$$\mathcal{K}(\mu, \nu) \triangleq \int_{\mathcal{X}} g \ln g \, d\nu, \quad \chi^2(\mu, \nu) \triangleq \int_{\mathcal{X}} (g-1)^2 d\nu,$$

if μ is absolutely continuous with respect to ν with Radon-Nikodym derivative $g = \frac{d\mu}{d\nu}$, and we set $\mathcal{K}(\mu, \nu) \triangleq +\infty$, $\chi^2(\mu, \nu) \triangleq +\infty$ otherwise.

We will use the following auxiliary result.

Lemma 3. *Let $(\mathcal{X}, \mathcal{A})$ be a measurable space and let $A_i \in \mathcal{A}$, $i \in \{0, 1, \dots, M\}$, $M \ge 2$, be such that $\forall i \ne j$, $A_i \cap A_j = \emptyset$. Assume that Q_i, $i \in \{0, 1 \dots, M\}$, are probability measures on $(\mathcal{X}, \mathcal{A})$ such that*

$$\frac{1}{M} \sum_{j=1}^{M} \mathcal{K}(Q_j, Q_0) \le \chi < \infty.$$

Then

$$p_* \triangleq \max_{0\le i\le M} Q_i(\mathcal{X}\setminus A_i) \ge \frac{1}{12}\min\{1,\, Me^{-3\chi}\}.$$

Proof: Proposition 2.3 in [16] yields:

$$p_* \ge \sup_{0<\tau<1} \frac{\tau M}{\tau M+1}\left(1+\frac{\chi+\sqrt{\chi/2}}{\log\tau}\right).$$

In particular, taking $\tau^* = \min(M^{-1}, e^{-3\chi})$ and using that $\sqrt{6\log M}\ge 2$ for $M\ge 2$, we obtain

$$p_* \ge \frac{\tau^* M}{\tau^* M+1}\left(1+\frac{\chi+\sqrt{\chi/2}}{\log\tau^*}\right) \ge \frac{1}{12}\min\{1,\, Me^{-3\chi}\}.$$

We now prove a classification setting analogue of the lower bound obtained by DeVore *et al.* [5] in the regression problem.

Theorem 5. *Assume that a class Θ of probability distributions P with the corresponding regression functions η_P and Bayes rules f_P^* (as defined above), contains a set $\{P_i\}_{i=1}^N \subset \Theta$, $N\ge 3$, with the following properties: the marginal distribution of X is μ_X for all P_i, independently of i, where μ_X is an arbitrary probability measure, $1/4\le \eta_{P_i}\le 3/4$, $i=1,\dots,N$, and for any $i\ne j$*

$$\|\eta_{P_i}-\eta_{P_j}\|_{L_2(\mu_X)} \le \gamma, \tag{24}$$

$$\|f^*_{P_i}-f^*_{P_j}\|_{L_1(\mu_X)} \ge s \tag{25}$$

with some $\gamma>0$, $s>0$. Then for any classifier $\hat{f}_n$ we have

$$\max_{1\le k\le N} \mathbb{P}_k\{\|\hat{f}_n - f^*_{P_k}\|_{L_1(\mu_X)} \ge s/2\} \ge \frac{1}{12}\min\left(1,\,(N-1)\exp\{-24n\gamma^2\}\right)$$

where $\mathbb{P}_k$ denotes the product probability measure associated to the i.i.d. n-sample from P_k.

Proof: We apply Lemma 3 where we set $Q_i=\mathbb{P}_i$, $M=N-1$, and define the random events A_i as follows:

$$A_i \triangleq \{\mathcal{D}_n : \|\hat{f}_n - f^*_{P_i}\|_{L_1(\mu_X)} < s/2\}, \quad i=1,\dots,N.$$

The events A_i are disjoint because of (25). Thus, the theorem follows from Lemma 3 if we prove that $\mathcal{K}(\mathbb{P}_i, \mathbb{P}_j) \le 8n\gamma^2$ for all i, j.

Let us evaluate $\mathcal{K}(\mathbb{P}_i, \mathbb{P}_j)$. For each η_{P_i}, the corresponding measure P_i is determined as follows

$$dP_i(x,y) \triangleq (\eta_{P_i}(x)d\delta_1(y) + (1-\eta_{P_i}(x))d\delta_0(y))d\mu_X(x),$$

where $d\delta_\xi$ denotes the Dirac measure with unit mass at ξ. Set for brevity $\eta_i \triangleq \eta_{P_i}$. Fix i and j. We have $dP_i(x,y) = g(x,y)dP_j(x,y)$, where

$$g(x,1) = \frac{\eta_i(x)}{\eta_j(x)}, \quad g(x,0) = \frac{1-\eta_i(x)}{1-\eta_j(x)}.$$

Therefore, using the inequalities $1/4 \le \eta_i, \eta_j \le 3/4$ and (24) we find

$$\begin{aligned}\chi^2(P_i, P_j) &= \int \left\{ \frac{(\eta_i(x)-\eta_j(x))^2}{\eta_j(x)} + \frac{(\eta_i(x)-\eta_j(x))^2}{1-\eta_j(x)} \right\} d\mu_X(x) \\ &\le 8\|\eta_i - \eta_j\|^2_{L_2(\mu_X)} \le 8\gamma^2.\end{aligned}$$

Together with inequality between the Kullback and χ^2-divergences, cf. [16], p. 90, this yields

$$\mathcal{K}(\mathbb{P}_i, \mathbb{P}_j) = n\mathcal{K}(P_i, P_j) \le n\chi^2(P_i, P_j) \le 8n\gamma^2.$$

□

Comment. The preprint version of this paper was posted on the Arxiv under the pseudonym N.I. Pentacaput [?]. Then the paper was submitted to "Constructive Approximation" and was accepted for publication under this pseudonym. However, it turns out that because of the Publisher rules no paper can be published under a pseudonym. As a result, we publish it under our real names that we have chosen to arrange in a random order.

References

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

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

[3] P.L. Bartlett, M.I. Jordan and J.D. McAuliffe (2006) Convexity, classification and risk bounds. *Journal of the American Statistical Association* **101**, 138–156.

[4] L. Devroye, L. Györfi and G. Lugosi, *A probabilistic theory of pattern recognition*, Vol. 31 of Applications of Mathematics (New York), Springer-Verlag, New York, 1996.

[5] R. DeVore, G. Kerkyacharian, D. Picard and V. Temlyakov (2006) Approximation methods for supervised learning. *Foundations of Computational Mathematics* **6**, 3–58.

[6] R. Dudley, *Uniform Central Limit Theorems*, Cambridge University Press, 1999.

[7] I.A. Ibragimov and R.Z. Hasminskii, *Statistical Estimation: Asymptotic Theory,* Springer, New York, 1981.

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

[9] V. Koltchinskii (2006) Local Rademacher Complexities and Oracle Inequalities in Risk Minimization. *Annals of Statistics*, **34**, 6, 2593–2656.

[10] V. Koltchinskii, *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems. Ecole d'été de Probabilités de Saint-Flour 2008*, Lecture Notes in Mathematics, Springer, New York, 2011.

[11] P. Massart, *Concentration Inequalities and Model Selection. Ecole d'été de Probabilités de Saint-Flour*, Lecture Notes in Mathematics, Springer, New York, 2007.

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

[13] N.I. Pentacaput (2011) Optimal exponential bounds on the accuracy of classification. `arXiv:1111.6160`

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

[15] V. N. Temlyakov (2008) Approximation in learning theory, *Constructive Approximation* 27 (1), 33–74.

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

[17] A. B. Tsybakov, *Introduction to Nonparametric Estimation.* Springer, New York, 2009.

[18] A.B. Tsybakov and S. van de Geer (2005) Square root penalty: adaptation to the margin in classification and in edge estimation. *Annals of Statistics*, **33**, 3, 1203–1224.

[19] A. van der Vaart and J. Wellner, *Weak Convergence and Empirical Processes. With Applications to Statistics.* Springer-Verlag, New York, 1996.

[20] Y. Yang (1999) Minimax nonparametric classification – part i: Rates of convergence. *IEEE Transaction on Information Theory* **45**, 2271–2284.