Lower Bounds on the Rate of Convergence of Nonparametric Pattern Recognition *

András Antos[1]

Informatics Laboratory, Computer and Automation Research Institute of the Hungarian Academy of Sciences, H-1111 Lágymányosi u. 11, Budapest, Hungary,
email: antos@inf.bme.hu,
WWW home page: http://www.szit.bme.hu/~antos/

Abstract. We show that there exist individual lower bounds corresponding to the upper bounds on the rate of convergence of nonparametric pattern recognition which are arbitrarily close to Yang's minimax lower bounds, if the a posteriori probability function is in the classes used by Stone and others. The rates equal to the ones on the corresponding regression estimation problem. Thus for these classes classification is not easier than regression estimation either in individual sense.

1 Introduction

Let (X,Y), (X_1,Y_1), $(X_2,Y_2),\ldots$ be independent identically distributed $\mathcal{R}^d \times \{0,1\}$ –valued random variables. In pattern recognition (or classification) one wishes to decide whether the value of Y (the label) is 0 or 1 given the (d–dimensional) value of X (the observation), i.e., one wants to find a decision function g defined on the range of X taking values 0 or 1 so that $g(X)$ equals to Y with high probability. Assume that the main aim of the analysis is to minimize the probability of error :

$$\min_g L(g) \stackrel{\text{def}}{=} \min_g \mathbf{P}\{g(X) \neq Y\} \ . \tag{1}$$

Let $\eta(x) \stackrel{\text{def}}{=} \mathbf{P}\{Y=1|X=x\} = \mathbf{E}\{Y|X=x\}$ be the a posteriori probability (or regression) function, let

$$g^*(x) \stackrel{\text{def}}{=} \begin{cases} 1 \text{ if } \eta(x) > 1/2, \\ 0 \text{ else} \end{cases}$$

be the Bayes-decision, let

$$L^* \stackrel{\text{def}}{=} L(g^*) = \mathbf{P}\{g^*(X) \neq Y\}$$

* The author's work was supported by a grant from the Hungarian Academy of Sciences (MTA SZTAKI).

be the Bayes-error and denote the distribution of X by μ. Introduce $||f||_q \stackrel{\text{def}}{=} (\int |f|^q \, d\mu)^{1/q}$. It is well–known, that for each measurable function $g : \mathcal{R}^d \mapsto \{0,1\}$ the relation

$$L(g) - L^* = 2 \int \left| \eta - \frac{1}{2} \right| I_{\{g \neq g^*\}} \, d\mu \tag{2}$$

holds. Therefore the function g^* achieves the minimum in (1) and the minimum is L^*.

For the classification problem, the distribution of (X,Y) (and so η and g^*) is unknown. Given only a sample $D_n = \{(X_1,Y_1),\ldots,(X_n,Y_n)\}$ of the distribution of (X,Y), one wants to construct a decision rule $g_n(x) = g_n(x,D_n) : \mathcal{R}^d \times (\mathcal{R}^d \times \{0,1\})^n \mapsto \{0,1\}$ such that

$$L_n \stackrel{\text{def}}{=} L(g_n) = \mathbf{P}\{g_n(X) \neq Y | D_n\}$$

is close to L^*. In this paper we study asymptotic properties of $\mathbf{E}L_n - L^*$.

If we have an estimate η_n of the regression function η and we derive a plug-in rule g_n from η_n quite naturally by

$$g_n(x) = \begin{cases} 1 \text{ if } \eta_n(x) > 1/2, \\ 0 \text{ else,} \end{cases}$$

then from (2) we get easily

$$L_n - L^* \leq 2||\eta_n - \eta||_1 \leq 2||\eta_n - \eta||_2$$

(see Devroye, Györfi and Lugosi [6]). This shows that if $||\eta_n - \eta||_1 \to 0$ then $L_n \to L^*$ in the same sense, and the latter has at least the same rate, i.e., classification is not more complex than regression estimation.

It is well–known, that there exist regression estimates, and so classification rules, which are universally consistent, i.e., which satisfy

$$\mathbf{E}L_n \to L^* \quad (n \to \infty)$$

for all distributions of (X,Y). This was first shown in Stone [8] for nearest neighbor estimates (see also [6] for a list of references).

Classification is actually easier than regression estimation in the sense that if $\mathbf{E}||\eta_n - \eta||_1 \to 0$, then for the plug-in rule

$$\frac{\mathbf{E}L_n - L^*}{\sqrt{\mathbf{E}\{||\eta_n - \eta||_2^2\}}} \to 0 \ . \tag{3}$$

(see [6] Chapter 6), i.e. the relative expected error of g_n decreases faster than the expected $\mathcal{L}_2(\mu)$ error of η_n. (Moreover, if $||\eta_n - \eta||_1 \to 0$ a.s., then for the plug-in rule

$$\frac{L_n - L^*}{||\eta_n - \eta||_2} \to 0 \quad \text{a.s.}$$

(see Antos [1]), i.e. the relation also holds for strong consistency.) However the value of the ratio above can not be universally bounded, the convergence can be arbitrary slow. It depends on the behavior of η near $1/2$ and the rate of convergence of η_n.

2 Lower Bounds

Unfortunately, there do not exist rules for which $\mathbf{E}L_n - L^*$ tends to zero with a guaranteed rate of convergence for all distributions of (X, Y). Theorem 7.2 and Problem 7.2 in [6] imply the following slow rate of convergence result: Let $\{a_n\}$ be a sequence of positive numbers converging to zero with $1/16 \geq a_1 \geq a_2 \geq \ldots$. For every sequence of decision rules $\{g_n\}$, there exists a distribution of (X, Y), such that X is uniformly distributed on $[0, 1]$, $\eta \in \{0, 1\}$ $(L^* = 0)$ and

$$\mathbf{E}L_n \geq a_n$$

for all n.

Therefore one has to restrict the class of distributions, which one considers, in order to obtain nontrivial rate of convergence results. Then it is natural to ask what the fastest achievable rate is for a given class of distributions. This is usually done by considering minimax rate of convergence results, where one derives lower bounds according to the following definition.

Definition 1. *A sequence of positive numbers a_n is called* lower rate of convergence for a class $\mathcal{D}$ *of distributions of (X, Y), if*

$$\limsup_{n\to\infty} \inf_{g_n} \sup_{(X,Y)\in\mathcal{D}} \frac{\mathbf{E}L_n - L^*}{a_n} > 0 \ .$$

Yang [10] pointed out that while (3) holds for every fixed *distribution* for which η_n is consistent, the optimal rates of convergence for many usual *classes* are the same in classification and regression estimation. He shows many examples and some counterexamples to this phenomenon with rates of convergence in term of metric entropy. Classification seems to have the same complexity as regression estimation for classes which contain functions with many values near $1/2$. (See also Mammen and Tsybakov [7].)

E.g. it was shown in [10] that for Lipschitz classes $\mathcal{D}^{(p,M)}$ below, the optimal lower rate of convergence is $n^{-\frac{p}{2p+d}}$, the same as for regression estimation (see also Stone [9]). For lower bound results on other types of distribution classes (e.g. Vapnik-Chevonenkis classes) see [6] and the references there.

Definition 2. *Let $\mathcal{F}^{(p,M)}$ be the set of functions $f : \mathcal{R}^d \to \mathcal{R}$ such that for $p = k + \beta$, $k \in \mathcal{N}_0$, $0 < \beta \leq 1$*

$$|D^\alpha f(x) - D^\alpha f(z)| \leq M\|x - z\|^\beta,$$

where D^α denotes the partial derivatives for $\alpha = (\alpha_1, \ldots, \alpha_d)$, $\alpha_i \in \mathcal{N}_0$, $\sum_{j=1}^d \alpha_j = k$.

Definition 3. *Let $\mathcal{D}^{(p,M)}$ be the class of distributions of (X, Y) such that*
(i) X is uniformly distributed on $[0, 1]^d$,
(ii) $\eta \in \mathcal{F}^{(p,M)}$.

It is well–known, that there exist regression estimates η_n, which satisfy

$$\limsup_{n\to\infty} \sup_{(X,Y)\in\mathcal{D}^{(p,M)}} \frac{\sqrt{\mathbf{E}\{||\eta_n-\eta||_2^2\}}}{n^{-\frac{p}{2p+d}}} < \infty \tag{4}$$

(see, e.g., Barron, Birge and Massart [4]), thus for the plug-in rules g_n

$$\limsup_{n\to\infty} \sup_{(X,Y)\in\mathcal{D}^{(p,M)}} \frac{\mathbf{E}L_n - L^*}{n^{-\frac{p}{2p+d}}} < \infty, \tag{5}$$

i.e., $\sup_{(X,Y)\in\mathcal{D}^{(p,M)}}(\mathbf{E}L_n - L^*) = O\left(n^{-\frac{p}{2p+d}}\right)$. (This remains true leaving condition (i) from Definition 3. Note that the rate does not depend on M.)

Theorem 1. (YANG [10]) *The sequence*

$$a_n = n^{-\frac{p}{2p+d}}$$

is a lower rate of convergence for the class $\mathcal{D}^{(p,M)}$.

In some sense, such lower bounds are not satisfactory. They do not tell us anything about the way the probability of error decreases as the sample size is increased for a given classification problem. These bounds, for each n, give information about the maximal probability of error within the class, but not about the behavior of the probability of error for a single fixed distribution as the sample size n increases. In other words, the "bad" distribution, causing the largest probability of error for a decision rule, may be different for each n. For example, the lower bound for the class $\mathcal{D}^{(p,M)}$ does not exclude the possibility that there exists a sequence of rules $\{g_n\}$ such that for *every* distribution in $\mathcal{D}^{(p,M)}$, the expected probability of error $\mathbf{E}L_n - L^*$ decreases at an exponential rate in n.

In this paper, we are interested in "individual" minimax lower bounds that describe the behavior of the probability of error for a fixed distribution (X,Y) as the sample size n grows.

Definition 4. *A sequence of positive numbers* a_n *is called* individual lower rate of convergence for a class $\mathcal{D}$ *of distributions of* (X,Y), *if*

$$\inf_{\{g_n\}} \sup_{(X,Y)\in\mathcal{D}} \limsup_{n\to\infty} \frac{\mathbf{E}L_n - L^*}{a_n} > 0,$$

where the infimum is taken over all sequences $\{g_n\}$ *of decision rules.*

The concept of individual lower rate has been introduced in Birgé [5] concerning density estimation. For individual lower rate results concerning pattern recognition see Antos and Lugosi [3].

We will show that for every sequence $\{b_n\}$ tending to zero, $b_n n^{-\frac{p}{2p+d}}$ is an individual lower rate of convergence of the class $\mathcal{D}^{(p,M)}$. Hence there exist individual lower rates of these classes, which are arbitrarily close to the optimal lower rates.

These rates are the same as the individual lower rates for the expected $\mathcal{L}_2(\mu)$ error of regression estimation for these classes in Antos, Györfi and Kohler [2]. (Actually the results here imply some results there.) Both individual lower rates are optimal by (4) and (5), hence we extended Yang's observation for the individual rates for these classes.

Our results also imply that the ratio $\frac{\mathbf{E}L_n - L^*}{\sqrt{\mathbf{E}\{||\eta_n - \eta||_2^2\}}}$ can tend to zero arbitrary slowly (even for a fixed sequence $\{\eta_n\}$).

Theorem 2. *Let $\{b_n\}$ be an arbitrary positive sequence tending to zero. Then the sequence*

$$b_n a_n = b_n n^{-\frac{2p}{2p+d}}$$

is an individual lower rate of convergence for the class $\mathcal{D}^{(p,M)}$.

REMARK 1. Applying for the sequence $\{\sqrt{b_n}\}$, Theorem 2 implies that for all $\{g_n\}$ there is $(X,Y) \in \mathcal{D}^{(p,M)}$ such that

$$\limsup_{n\to\infty} \frac{\mathbf{E}L_n - L^*}{b_n a_n} = \infty \ .$$

REMARK 2. Certainly Theorems 1 and 2 hold if we increase the class by leaving condition (i) from Definition 3.

Call a sequence c_n an *upper rate of convergence for a class $\mathcal{D}$*, if there exist rules g_n which satisfy

$$\limsup_{n\to\infty} \sup_{(X,Y)\in\mathcal{D}} \frac{\mathbf{E}L_n - L^*}{c_n} < \infty,$$

i.e., $\sup_{(X,Y)\in\mathcal{D}}(\mathbf{E}L_n - L^*) = O(c_n)$, and call it an *individual upper rate of convergence for a class $\mathcal{D}$*, if there exist rules g_n which satisfy

$$\sup_{(X,Y)\in\mathcal{D}} \limsup_{n\to\infty} \frac{\mathbf{E}L_n - L^*}{c_n} < \infty \ .$$

This implies only that for every distribution in $\mathcal{D}$, $\mathbf{E}L_n - L^* = O(c_n)$, possibly with different constants. Then (5) implies that $n^{-\frac{p}{2p+d}}$ is an upper rate of convergence, and thus also an individual upper rate of convergence for $\mathcal{D}^{(p,M)}$. While Theorem 1 shows only that there is no upper rate of convergence for $\mathcal{D}^{(p,M)}$ better than $n^{-\frac{p}{2p+d}}$, it follows from Theorem 2 that $n^{-\frac{p}{2p+d}}$ is even the optimal individual upper rate for $\mathcal{D}^{(p,M)}$ in the sense, that there doesn't exist an individual upper rate c_n of convergence for $\mathcal{D}^{(p,M)}$, which satisfies

$$\lim_{n\to\infty} \frac{c_n}{n^{-\frac{p}{2p+d}}} = 0 \ .$$

Moreover (3) and (4) imply

$$\inf_{\{g_n\}} \sup_{(X,Y)\in\mathcal{D}^{(p,M)}} \limsup_{n\to\infty} \frac{\mathbf{E}L_n - L^*}{n^{-\frac{p}{2p+d}}} = 0, \tag{6}$$

which shows that Theorem 2 can not be improved by dropping b_n. This shows the strange nature of individual lower bounds, that while every sequence tending to zero faster than $n^{-\frac{p}{2p+d}}$ is an individual lower rate for $\mathcal{D}^{(p,M)}$, $n^{-\frac{p}{2p+d}}$ itself is not that.

3 Proofs

The proofs of the theorems apply the following lemma:

Lemma 1. *Let $u = (u_1, \ldots, u_l)$ be a l-dimensional real vector taking values in $[-1/4, 1/4]^l$, let C be a zero mean random variable taking values in $\{-1, +1\}$, and let $Y_1, \ldots, Y_l$ be indepedent binary variables given C with*

$$\mathbf{P}\{Y_i = 1|C\} = \frac{1}{2} + Cu_i \quad i = 1, \ldots, l \ .$$

Then for the error probability of the Bayes decision for C based on $Y = (Y_1, \ldots, Y_l)$

$$L^* \geq \frac{1}{4} e^{-10\sqrt{\sum_i u_i^2 + 4\sum_{i \neq i'} u_i^2 u_{i'}^2}} \ .$$

PROOF. The Bayes decision is 1 if $\mathbf{P}\{C = 1|Y\} \geq \frac{1}{2}$ and -1 otherwise, therefore

$$L^* = \mathbf{E}\{\min(\mathbf{P}\{C = 1|Y\}, \mathbf{P}\{C = -1|Y\})\} = \mathbf{E}\{\pi\} \ .$$

One can verify that

$$\mathbf{P}\{C = 1|Y\} = \frac{T}{T+1},$$

where

$$\begin{aligned} T &\stackrel{\text{def}}{=} \prod_{i \leq l} \left(\frac{\frac{1}{2} + u_i}{\frac{1}{2} - u_i} I_{\{Y_i = 1\}} + \frac{\frac{1}{2} - u_i}{\frac{1}{2} + u_i} I_{\{Y_i = 0\}} \right) = \prod_{i \leq l} \left(\frac{\frac{1}{2} + u_i}{\frac{1}{2} - u_i} \right)^{2Y_i - 1} \\ &= e^{\sum_{i \leq l} (2Y_i - 1) \log \frac{1 + 2u_i}{1 - 2u_i}} = e^{\sum_{i \leq l} Z_i}, \end{aligned}$$

where $Z_i \stackrel{\text{def}}{=} (2Y_i - 1) \log((1 + 2u_i)/(1 - 2u_i))$. For arbitrary $0 < q < 1/2$, $\pi \geq q$ if and only if $|\log T| \leq \log \frac{1-q}{q}$, therefore

$$\mathbf{E}\{\pi\} \geq q\mathbf{P}\{\pi \geq q\} = q\mathbf{P}\left\{ |\log T| \leq \log \frac{1-q}{q} \right\} \ .$$

By Markov's inequality

$$\mathbf{P}\left\{ |\log T| \leq \log \frac{1-q}{q} \right\} \geq 1 - \frac{\mathbf{E}|\log T|}{\log \frac{1-q}{q}} \ .$$

Moreover because of $|\log T| = \left|\sum_{i\le l} Z_i\right|$ we get

$$\mathbf{E}|\log T| = \mathbf{E}\left|\sum_{i\le l} Z_i\right| \le \sqrt{\mathbf{E}\left(\sum_{i\le l} Z_i\right)^2} = \sqrt{\mathbf{E}\sum_i Z_i^2 + \sum_{i\ne i'} Z_i Z_{i'}}$$
$$= \sqrt{\sum_i \mathbf{E}\{Z_i^2\} + \sum_{i\ne i'} \mathbf{E}\{Z_i Z_{i'}\}} \ .$$

Using the inequality for $-1/4 \le x \le 1/4$

$$\left|\log\frac{1+2x}{1-2x}\right| = |\log(1+2x) - \log(1-2x)| = \log(1+2|x|) - \log(1-2|x|)$$
$$\le 2|x| + \log 4 \cdot 2|x| \le 5|x|,$$

on the one hand

$$\mathbf{E}\{Z_i^2\} = \log^2\frac{1+2u_i}{1-2u_i} \le 25u_i^2$$

and on the other hand

$$\mathbf{E}\{(2Y_i-1)(2Y_{i'}-1)\} = 4\mathbf{E}\{Y_iY_{i'}\} - 2\mathbf{E}\{Y_i\} - 2\mathbf{E}\{Y_{i'}\} + 1$$
$$= 4(\mathbf{E}\{Y_iY_{i'}|C=1\}\mathbf{P}\{C=1\} + \mathbf{E}\{Y_iY_{i'}|C=-1\}\mathbf{P}\{C=-1\}) - 1$$
$$= 4\left(\left(\frac{1}{2}+u_i\right)\left(\frac{1}{2}+u_{i'}\right)\frac{1}{2} + \left(\frac{1}{2}-u_i\right)\left(\frac{1}{2}-u_{i'}\right)\frac{1}{2}\right) - 1 = 4u_iu_{i'},$$

so

$$\mathbf{E}\{Z_iZ_{i'}\} = \mathbf{E}\{(2Y_i-1)(2Y_{i'}-1)\}\log\frac{1+2u_i}{1-2u_i}\log\frac{1+2u_{i'}}{1-2u_{i'}}$$
$$= 4u_iu_{i'}\log\frac{1+2u_i}{1-2u_i}\log\frac{1+2u_{i'}}{1-2u_{i'}}$$
$$\le 4|u_i|\left|\log\frac{1+2u_i}{1-2u_i}\right||u_{i'}|\left|\log\frac{1+2u_{i'}}{1-2u_{i'}}\right| \le 100u_i^2u_{i'}^2 \ .$$

Hence

$$\mathbf{E}|\log T| \le \sqrt{25\sum_i u_i^2 + 100\sum_{i\ne i'} u_i^2u_{i'}^2} \ .$$

Thus

$$\mathbf{E}\{\pi\} \ge q\left(1 - \frac{\mathbf{E}|\log T|}{\log\frac{1-q}{q}}\right) \ge q\left(1 - \frac{5\sqrt{\sum_i u_i^2 + 4\sum_{i\ne i'} u_i^2u_{i'}^2}}{\log\frac{1-q}{q}}\right) \ .$$

By choosing $q = \left(1 + \exp\left(1 + 5\sqrt{\sum_i u_i^2 + 4\sum_{i\neq i'} u_i^2 u_{i'}^2}\right)\right)^{-1}$,

$$\mathbf{E}\{\pi\} \geq \frac{1}{1+\exp(1+5\sqrt{\sum_i u_i^2 + 4\sum_{i\neq i'} u_i^2 u_{i'}^2})} \frac{1}{1+5\sqrt{\sum_i u_i^2 + 4\sum_{i\neq i'} u_i^2 u_{i'}^2}}$$

$$\geq \frac{1}{(e+1)e^{10\sqrt{\sum_i u_i^2 + 4\sum_{i\neq i'} u_i^2 u_{i'}^2}}} \geq \frac{1}{4} e^{-10\sqrt{\sum_i u_i^2 + 4\sum_{i\neq i'} u_i^2 u_{i'}^2}} \quad .$$

□

PROOF OF THEOREM 1. This proof differs from that of Yang, it can be easily modified to individual lower bound in Theorem 2. First we define a subclass of distributions (X,Y) contained in $\mathcal{D}^{(p,M)}$. We pack infinitely many disjoint cubes into $[0,1]^d$ in the following way: For a given probability distribution $\{p_j\}$, let $\{B_j\}$ be a partition of $[0,1]$ such that B_j is an interval of length p_j. We pack disjoint cubes of volume p_j^d into the rectangle

$$B_j \times [0,1]^{d-1} \quad .$$

Denote these cubes by $A_{j,1}, \ldots, A_{j,S_j}$, where $S_j = \left\lfloor \frac{1}{p_j} \right\rfloor^{d-1}$. Let $a_{j,k}$ be the center of $A_{j,k}$. Choose a function $m : \mathcal{R}^d \to [0, 1/4]$ such that
(I) the support of m is a subset of $[-\frac{1}{2}, \frac{1}{2}]^d$,
(II) $\int m(x)\,dx > 0$,
(III) $m \in \mathcal{F}^{(p,M2^{\beta-1})}$.
The class of a posteriori probability functions is indexed by a vector

$$c = (c_{1,1}, c_{1,2}, \ldots, c_{1,S_1}, c_{2,1}, c_{2,2}, \ldots, c_{2,S_2}, \ldots)$$

of $+1$ or -1 components. Denote the set of all such vectors by $\mathcal{C}$. For $c \in \mathcal{C}$ define the function

$$\eta^{(c)}(x) = \frac{1}{2} + \sum_{j=1}^{\infty} \sum_{k=1}^{S_j} c_{j,k} m_{j,k}(x),$$

where $m_{j,k}(x) = p_j^p m(p_j^{-1}(x - a_{j,k}))$. Then it is easy to check (cf. [9], p. 1045) that $\eta^{(c)}(x) \in [0,1]$ for all $x \in [0,1]^d$ and because of (III)

$$\eta^{(c)} \in \mathcal{F}^{(p,M)} \quad .$$

Hence, each distribution (X,Y) with $Y \in \{0,1\}$ and $\mathbf{E}\{Y|X=x\} = \mathbf{P}\{Y = 1|X=x\} = \eta^{(c)}(x)$ for all $x \in [0,1]^d$ for some $c \in \mathcal{C}$ is contained in $\mathcal{D}^{(p,M)}$, which implies

$$\limsup_{n\to\infty} \inf_{g_n} \sup_{(X,Y)\in\mathcal{D}^{(p,M)}} \frac{\mathbf{E}L_n - L^*}{a_n}$$

$$\geq \limsup_{n\to\infty} \inf_{g_n} \sup_{(X,Y):\mathbf{E}\{Y|X=x\}=\eta^{(c)}(x), c\in\mathcal{C}} \frac{\mathbf{E}L_n - L^*}{a_n} \quad . \tag{7}$$

Let g_n be an arbitrary rule. By definition, $\{I_{A_{j,k}}/2 : j,k\}$ is an orthogonal system in $\mathcal{L}_2(\nu)$ for the measure $\nu(A) = \int\limits_A \sum_{j,k} m_{j,k}\, d\mu$, therefore the projection $\hat{g}_n - \frac{1}{2}$ of $g_n - \frac{1}{2}$ is given by

$$\hat{g}_n(x) - \frac{1}{2} = \sum_{j,k} \hat{c}_{n,j,k} \frac{I_{A_{j,k}}(x)}{2},$$

where

$$\hat{c}_{n,j,k} = \frac{\int (g_n - 1/2) I_{A_{j,k}}/2 \, d\nu}{\int \left(I_{A_{j,k}}/2\right)^2 \, d\nu} = 2 \frac{\int\limits_{A_{j,k}} (g_n - 1/2)\, d\nu}{\int\limits_{A_{j,k}} 1 \, d\nu}$$

$$= 2 \frac{\int\limits_{A_{j,k}} (g_n(x) - 1/2) m_{j,k}(x)\, dx}{\int\limits_{A_{j,k}} m_{j,k}(x)\, dx} = 2 \frac{\int\limits_{A_{j,k}} g_n(x) m_{j,k}(x)\, dx}{\int\limits_{A_{j,k}} m_{j,k}(x)\, dx} - 1 \ .$$

(Note that $\hat{c}_{n,j,k} \in [-1,1]$.) Let $c \in \mathcal{C}$ be arbitrary. Note that $g^* = \frac{1}{2} + \sum_{j,k} c_{j,k} \frac{I_{A_{j,k}}}{2}$. Then by (2)

$$\begin{aligned}
L_n - L^* &= 2 \int \left| \eta^{(c)} - \frac{1}{2} \right| I_{\{g_n \neq g^*\}}\, d\mu = 2 \int \sum_{j,k} m_{j,k} (g_n - g^*)^2 \, d\mu \\
&= 2 \int \left(\left(g_n - \frac{1}{2} \right) - \left(g^* - \frac{1}{2} \right) \right)^2 \, d\nu \\
&\geq 2 \int \left(\left(\hat{g}_n - \frac{1}{2} \right) - \left(g^* - \frac{1}{2} \right) \right)^2 \, d\nu \\
&= 2 \int \sum_{j,k} m_{j,k} (\hat{g}_n - g^*)^2 \, d\mu = 2 \sum_{j,k} \frac{(\hat{c}_{n,j,k} - c_{j,k})^2}{4} \int\limits_{A_{j,k}} m_{j,k}\, d\mu \\
&= \frac{1}{2} ||m||_1 \sum_{j,k} (\hat{c}_{n,j,k} - c_{j,k})^2 p_j^{p+d} \ .
\end{aligned}$$

Let $\tilde{c}_{n,j,k}$ be 1 if $\hat{c}_{n,j,k} \geq 0$ and -1 otherwise. Because of $|\hat{c}_{n,j,k} - c_{j,k}| \geq \frac{|\tilde{c}_{n,j,k} - c_{j,k}|}{2} = I_{\{\tilde{c}_{n,j,k} \neq c_{j,k}\}}$, we get

$$L_n - L^* \geq \frac{1}{2} ||m||_1 \sum_{j,k} I_{\{\tilde{c}_{n,j,k} \neq c_{j,k}\}} p_j^{p+d} \ .$$

This proves

$$\mathbf{E} L_n - L^* \geq \frac{1}{2} ||m||_1 R_n(c), \tag{8}$$

where

$$R_n(c) = \sum_{j : n p_j^{2p+d} \leq 1} \sum_{k=1}^{S_j} p_j^{p+d} \cdot \mathbf{P}\{\tilde{c}_{n,j,k} \neq c_{j,k}\} \ . \tag{9}$$

Equations (7) and (8) imply

$$\limsup_{n\to\infty} \inf_{g_n} \sup_{(X,Y)\in\mathcal{D}^{(p,M)}} \frac{EL_n - L^*}{a_n} \geq \frac{1}{2}||m||_1 \limsup_{n\to\infty} \inf_{g_n} \sup_{c\in\mathcal{C}} \frac{R_n(c)}{a_n} \quad . \tag{10}$$

To bound the last term, we fix the rules g_n and choose $c \in \mathcal{C}$ randomly. Let

$$C = (C_{1,1}, \ldots, C_{1,S_1}, C_{2,1}, \ldots, C_{2,S_2}, \ldots)$$

be a sequence of independent identically distributed random variables independent of X_1, X_2, ..., which satisfy $\mathbf{P}\{C_{1,1} = 1\} = \mathbf{P}\{C_{1,1} = -1\} = 1/2$. Next we derive a lower bound for

$$\mathbf{E}R_n(C) = \sum_{j:np_j^{2p+d}\leq 1} \sum_{k=1}^{S_j} p_j^{p+d} \cdot \mathbf{P}\{\tilde{c}_{n,j,k} \neq C_{j,k}\} \quad .$$

$\tilde{c}_{n,j,k}$ can be interpreted as a decision on $C_{j,k}$ using D_n. Its error probability is minimal for the Bayes decision $\bar{C}_{n,j,k}$, which is 1 if $\mathbf{P}\{C_{j,k} = 1|D_n\} \geq \frac{1}{2}$ and -1 otherwise, therefore

$$\mathbf{P}\{\tilde{c}_{n,j,k} \neq C_{j,k}\} \geq \mathbf{P}\{\bar{C}_{n,j,k} \neq C_{j,k}\} \quad .$$

Let $X_{i_1}, \ldots, X_{i_l}$ be those $X_i \in A_{j,k}$. Then given $X_1, \ldots, X_n$, $(Y_{i_1}, \ldots, Y_{i_l})$ is distributed as $(Y_1, \ldots, Y_l)$ in the conditions of Lemma 1 with $u_r = m_{j,k}(X_{i_r})$, while

$$(Y_1, \ldots, Y_n) \setminus (Y_{i_1}, \ldots, Y_{i_l})$$

depends only on $C \setminus \{C_{j,k}\}$ and on X_r's with $r \notin \{i_1, \ldots, i_l\}$, therefore is independent of $C_{j,k}$ given $X_1, \ldots, X_n$. Now conditioning on $X_1, \ldots, X_n$, the error of the conditional Bayes decision for $C_{j,k}$ based on $(Y_1, \ldots, Y_n)$ depends only on $(Y_{i_1}, \ldots, Y_{i_l})$, hence Lemma 1 implies

$$\mathbf{P}\{\bar{C}_{n,j,k} \neq C_{j,k}|X_1, \ldots, X_n\} \geq \frac{1}{4} e^{-10\sqrt{\sum_i m_{j,k}^2(X_i) + 4\sum_{i\neq i'} m_{j,k}^2(X_i) m_{j,k}^2(X_{i'})}} \quad .$$

By Jensen-inequality

$$\begin{aligned}
\mathbf{P}\{\bar{C}_{n,j,k} \neq C_{j,k}\} &= \mathbf{E}\{\mathbf{P}\{\bar{C}_{n,j,k} \neq C_{j,k}|X_1, \ldots, X_n\}\} \\
&\geq \frac{1}{4}\mathbf{E}\left\{ e^{-10\sqrt{\sum_i m_{j,k}^2(X_i) + 4\sum_{i\neq i'} m_{j,k}^2(X_i) m_{j,k}^2(X_{i'})}} \right\} \\
&\geq \frac{1}{4} e^{-10\sqrt{\sum_i \mathbf{E}\{m_{j,k}^2(X_i)\} + 4\sum_{i\neq i'} \mathbf{E}\{m_{j,k}^2(X_i) m_{j,k}^2(X_{i'})\}}} \\
&= \frac{1}{4} e^{-10\sqrt{\int m^2\, d\mu \cdot np_j^{2p+d} + 4\left(\int m^2\, d\mu\right)^2 \cdot n(n-1)p_j^{4p+2d}}}
\end{aligned}$$

independently of k. Thus if $np_j^{2p+d} \le 1$

$$\mathbf{P}\{\bar{C}_{n,j,k} \neq C_{j,k}\} \ge \frac{1}{4} e^{-10||m||_2\sqrt{1+4||m||_2^2}} > 0,$$

and

$$\begin{aligned}\mathbf{E}R_n(C) &\ge \sum_{j:np_j^{2p+d}\le 1} S_j p_j^{p+d} \cdot \mathbf{P}\{\bar{C}_{n,j,1} \neq C_{j,1}\} \\ &\ge \frac{1}{4} e^{-10||m||_2\sqrt{1+4||m||_2^2}} \sum_{j:np_j^{2p+d}\le 1} S_j p_j^{p+d} \ge K_1 \sum_{j:np_j^{2p+d}\le 1} p_j^{p+1}, \quad (11)\end{aligned}$$

where $K_1 = \frac{1}{4} e^{-10||m||_2\sqrt{1+4||m||_2^2}} (1/2)^{d-1}$. Setting $p_j = p_{j,n} = (1/n)^{\frac{1}{2p+d}}$ for $j \le n^{\frac{1}{2p+d}}$,

$$\mathbf{E}R_n(C) \ge K_1 \lfloor n^{\frac{1}{2p+d}} \rfloor n^{-\frac{p+1}{2p+d}} = K_1 a_n (1 - o(1)),$$

so

$$\limsup_{n\to\infty} \inf_{g_n} \sup_{c\in\mathcal{C}} \frac{R_n(c)}{a_n} \ge \limsup_{n\to\infty} \inf_{g_n} \frac{\mathbf{E}R_n(C)}{a_n} \ge K_1 > 0 \ . \quad (12)$$

This together with (10) implies the assertion. □

Proof of Theorem 2. We use the notations and results of the proof of Theorem 1. Now we have by (8)

$$\inf_{\{g_n\}} \sup_{(X,Y)\in\mathcal{D}^{(p,M)}} \limsup_{n\to\infty} \frac{EL_n - L^*}{b_n a_n} \ge \frac{1}{2}||m||_1 \inf_{\{g_n\}} \sup_{c\in\mathcal{C}} \limsup_{n\to\infty} \frac{R_n(c)}{b_n a_n} \ . \quad (13)$$

In this case we have to choose $\{p_j\}$ independently from n. Since b_n and a_n tend to zero we can take a subsequence $\{n_t\}_{t\in\mathcal{N}}$ of $\{n\}_{n\in\mathcal{N}}$ with $b_{n_t} \le 2^{-t}$ and $a_{n_t}^{1/p} \le 2^{-t}$. Define q_t such that

$$\frac{2^{-t}}{q_t} = \left\lceil \frac{2^{-t}}{a_{n_t}^{1/p}} \right\rceil,$$

and choose $\{p_j\}$ as

$$q_1, \ldots, q_1, q_2, \ldots, q_2, \ldots, q_t, \ldots, q_t, \ldots,$$

where q_t is repeated $2^{-t}/q_t$ times. So

$$\begin{aligned}&\sum_{j:np_j^{2p+d}\le 1} p_j^{p+1} = \sum_{t:nq_t^{2p+d}\le 1} \frac{2^{-t}}{q_t} q_t^{p+1} \ge \sum_{t:nq_t^{2p+d}\le 1} b_{n_t} q_t^p \\ &= \sum_{t:\lceil 2^{-t} a_{n_t}^{-1/p}\rceil \ge 2^{-t} a_n^{-1/p}} b_{n_t} \left(\frac{2^{-t}}{\left\lceil \frac{2^{-t}}{a_{n_t}^{1/p}} \right\rceil} \right)^p \ge \sum_{t:a_{n_t}\le a_n} b_{n_t} \left(\frac{2^{-t}}{\frac{2^{-t}}{a_{n_t}^{1/p}} + 1} \right)^p \\ &= \sum_{t:n_t\ge n} b_{n_t} \left(\frac{a_{n_t}^{1/p}}{1 + 2^t a_{n_t}^{1/p}} \right)^p \ge \sum_{t:n_t\ge n} \frac{b_{n_t} a_{n_t}}{2^p}\end{aligned}$$

by $a_{n_t}^{1/p} \le 2^{-t}$, and specially for $n = n_s$ (11) implies

$$\mathbf{E}R_{n_s}(C) \ge K_1 \sum_{j:n_s p_j^{2p+d} \le 1} p_j^{p+1} \ge \frac{K_1}{2^p} \sum_{t \ge s} b_{n_t} a_{n_t} \ge \frac{K_1}{2^p} b_{n_s} a_{n_s} \quad . \tag{14}$$

Using (14) one gets

$$\inf_{\{g_n\}} \sup_{c \in \mathcal{C}} \limsup_{n \to \infty} \frac{R_n(c)}{b_n a_n} \ge \inf_{\{g_n\}} \sup_{c \in \mathcal{C}} \limsup_{s \to \infty} \frac{R_{n_s}(c)}{b_{n_s} a_{n_s}}$$
$$\ge \frac{K_1}{2^p} \inf_{\{g_n\}} \sup_{c \in \mathcal{C}} \limsup_{s \to \infty} \frac{R_{n_s}(c)}{\mathbf{E}R_{n_s}(C)} \ge \frac{K_1}{2^p} \inf_{\{g_n\}} \mathbf{E}\left(\limsup_{s \to \infty} \frac{R_{n_s}(C)}{\mathbf{E}R_{n_s}(C)} \right) \quad .$$

Because of (11) and the fact that for all $c \in \mathcal{C}$

$$R_n(c) \le \sum_{j:np_j^{2p+d} \le 1} S_j p_j^{p+d} \le \sum_{j:np_j^{2p+d} \le 1} p_j^{p+1},$$

the sequence $R_{n_s}(C)/\mathbf{E}R_{n_s}(C)$ is uniformly bounded, so we can apply Fatou's lemma to get

$$\inf_{\{g_n\}} \sup_{c \in \mathcal{C}} \limsup_{n \to \infty} \frac{R_n(c)}{b_n a_n} \ge \frac{K_1}{2^p} \inf_{\{g_n\}} \limsup_{s \to \infty} \mathbf{E}\left(\frac{R_{n_s}(C)}{\mathbf{E}R_{n_s}(C)} \right) = \frac{K_1}{2^p} > 0 \quad .$$

This together with (13) implies the assertion. □

References

1. Antos, A.: *Függvényosztályok tulajdonságai és szerepe az alakfelismerésben* (in Hungarian) (Properties of classes of functions and their roles in pattern recognition). M.Sc. Thesis, Technical University of Budapest, Budapest, Hungary (1995)
2. Antos, A., Györfi, L. and Kohler, M.: Lower bounds on the rate of convergence of nonparametric regression estimates. Preprint No. 98-11 (1998), Universität Stuttgart. Submitted
3. Antos, A. and Lugosi, G.: Strong minimax lower bounds for learning. *Machine Learning* 30 (1998) 31–56
4. Barron, A. R., Birgé, L. and Massart, P.: Risk bounds for model selection via penalization. Technical Report No. 95.54 (1995), Université Paris Sud. To appear in *Probability Theory and Related Fields*
5. Birgé, L.: On estimating a density using Hellinger distance and some other strange facts. *Probability Theory and Related Fields* 71 (1986) 271–291
6. Devroye, L. , Györfi, L. and Lugosi, G. : *A Probabilistic Theory of Pattern Recognition* Springer Verlag (1996)
7. Mammen, E. and Tsybakov, A. B.: Smooth discrimination analysis. Submitted
8. Stone, C. J.: Consistent nonparametric regression. *Annals of Statistics* 5 (1977) 595–645
9. Stone, C. J. : Optimal global rates of convergence for nonparametric regression. *Annals of Statistics* 10 (1982) 1040–1053
10. Yang, Y.: Minimax nonparametric classification — Part I: Rates of convergence. Submitted