Lower bounds for the rate of convergence in nonparametric pattern recognition *

András Antos †

July 5, 1999

Abstract

We show that there exist individual lower bounds corresponding to the upper bounds for the rate of convergence of nonparametric pattern recognition which are arbitrarily close to Yang's minimax lower bounds, for certain "cubic" classes of regression functions used by Stone and others. The rates are equal to the ones of the corresponding regression function estimation problem. Thus for these classes classification is not easier than regression function estimation.

Keywords: Individual rates of convergence, nonparametric pattern recognition.

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

†A. Antos is with the Informatics Laboratory, Computer and Automation Research Institute of the Hungarian Academy of Sciences, H-1111 Lágymányosi u. 11, Budapest, Hungary. (email: antos@szit.bme.hu).

1 Introduction

Let (X, Y), (X_1, Y_1), (X_2, Y_2),... 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), that is, 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. Introduce the Bayes-decision

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

and let

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

be the Bayes-error. Denote the distribution of X by μ. Introduce $\|f\|_q \stackrel{\text{def}}{=} (\int |f|^q \, d\mu)^{1/q}$. It is well-known (see Devroye, Györfi and Lugosi [8]), 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, where I_A denotes the indicator function of the event A. Therefore the function g^* achieves the minimum in (1) and the minimum is L^*.

In the classification problem we consider here, the distribution of (X, Y) (and therefore also η and g^*) is unknown. Given only the independent 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 [8]). 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, that is, in a sense, classification is not more complex than regression function estimation.

It is well-known, that there exist regression function estimates, and so classification rules, which are universally consistent, that is, which satisfy

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

for all distributions of (X, Y). This was first shown in Stone [14] for nearest neighbor estimates (see also [8] for a list of references).

Classification is actually easier than regression function 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 [8] Chapter 6), that is, the relative expected error of g_n decreases faster than the expected $L_2(\mu)$ 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]), that is, the relation also holds for strong consistency. However the value of the ratio above cannot 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 $\{\eta_n\}$.

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 Devroye et al. [8] imply the following slow-rate-of-convergence result: Let $\{a_n\}$ be a positive sequence converging to zero with $1/16 \geq a_1 \geq a_2 \geq \ldots$. For every sequence $\{g_n\}$ of decision rules, 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, in order to obtain nontrivial rate-of-convergence results, one has to restrict the class of distributions. 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 positive sequence* $\{a_n\}$ *is called* **lower rate of convergence for a class** $\mathcal{D}$ *of distributions of* (X, Y) *if*

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

Remark. In many cases the limit superior in this definition could be replaced by limit inferior or infimum, because the lower bound for the minimax loss holds for all (sufficiently large) n.

Since μ and η determine the distribution of (X, Y), the class $\mathcal{D}$ of distributions is often given as a product of a class $\mathcal{H}$ of allowed distributions of X and a class $\mathcal{F}$ of allowed regression functions. For example, $\mathcal{H}$ may consist of all absolute continuous distributions, or all distributions (distribution-free approach), or one particular distribution (distribution-sensitive approach).

In this paper we give lower bounds for some classes $\mathcal{D}_\beta$ in the last (and strongest) format, when $\mathcal{H}$ contains one (uniform) distribution, and $\mathcal{F} = \mathcal{F}_\beta$ is some "cubic" class of functions (with a parameter β), defined later. For minimax lower-rate results on other types of distribution classes (e.g., Vapnik-Chervonenkis classes), see Devroye et al. [8] and the references therein. For related results on the general minimax theory of statistical estimates see Ibragimov and Khasmiskii [9], [10], [11], Korostelev and Tsybakov [12].

Yang [16] points out that while (3) holds for every fixed *distribution* for which $\{\eta_n\}$ is consistent, the optimal rate of convergence for many usual *classes* is the same in classification and regression function estimation. He shows many examples and some counterexamples to this phenomenon with rates of convergence in terms of metric entropy. Classification seems to have the same complexity as regression function estimation for classes which are rich near 1/2. (See also Mammen and Tsybakov [13].)

For example, it was shown in Yang [16] that for the distribution classes $\mathcal{D}_\beta$ above, the optimal rate of convergence is $\{n^{-\frac{\beta}{2\beta+d}}\}$, the same as for regression function estimation (see also Stone [15]).

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 previous lower bounds for the classes $\mathcal{D}_\beta$ does not exclude the possibility that there exists a sequence $\{g_n\}$ such that for *every* distribution in $\mathcal{D}_\beta$, the expected probability of error $\mathbf{E}L_n - L^*$ decreases at an exponential rate in n.

In this paper, we are interested also 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 2 *A positive sequence* $\{a_n\}$ *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 [7] concerning density estimation. For individual lower-rate results concerning pattern recognition see Antos and Lugosi [5].

We will show that for every sequence $\{b_n\}$ tending to zero, $\{b_n n^{-\frac{\beta}{2\beta+d}}\}$ is an individual lower rate of convergence for the classes $\mathcal{D}_\beta$. 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 $L_2(\mu)$ error of regression function estimation for these classes (see also Antos, Györfi and Kohler [4] and Antos [3]). Both for regression function estimation and pattern recognition, the individual lower rates are optimal, hence we extend 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 $\{\eta_n\}$).

Next we give the definitions of the function classes, for which we derive lower rates of convergence. Let $\mathcal{N} = \{1, 2, \ldots\}$ and $\mathcal{N}_0 = \mathcal{N} \cup \{0\}$, and denote the L_q norm regarding to the Lebesgue-measure λ on $\mathcal{R}^d$ by $\|\cdot\|_{L_q(\lambda)}$.

Definition 3 *For given* $1 \le q \le \infty$, $R \in \mathcal{N}_0$, $(1/q - 1/2)^+ < \alpha \le 1$, $\beta = R + \alpha$ *and* $M_0, \ldots, M_R > 0$, *let* $\mathrm{Lip}_{\beta,d} = \mathrm{Lip}_{\beta,d}(q, M_0, \ldots, M_R)$ *be the class of functions* $f : \mathcal{R}^d \to \mathcal{R}$ *such that for* $R' = 0,\ldots,R$, *for every* $\gamma = (\gamma_1, \ldots, \gamma_d)$ *with* $\gamma_i \in \mathcal{N}_0$, $\sum_{i=1}^d \gamma_i = R'$,

$$\|D^\gamma f(x)\|_{L_q(\lambda)} \le M_{R'} ,$$

moreover, for every such γ *with* $\sum_{i=1}^d \gamma_i = R$ *for every* $\delta \in \mathcal{R}^d$,

$$\|D^\gamma f(x+\delta) - D^\gamma f(x)\|_{L_q(\lambda)} \le M_R \|\delta\|^\alpha \ ,$$

where D^γ *denotes the partial derivative with respect to* γ*.*

For $q < \infty$, we assume $R \ge 1$. For $q = \infty$, all the rate-of-convergence results below hold if some of $M_0, \ldots, M_{R-1}$ are infinite, that is, if we omit some of the conditions of the first kind. These classes are generalizations of the Lipschitz classes of Example 2 in Yang [16], Antos, Györfi and Kohler [4], Birgé [7] and Stone [15], and also generalizations of a special case of Example 4 in Yang [16] for polinomial modulus of continuity.

Definition 4 *For given* $\vec{R} = (R_1, \ldots, R_d) \in \mathcal{N}^d$, $0 < \beta_i \le R_i$ $(1 \le i \le d)$, $\vec{\beta} = (\beta_1, \ldots, \beta_d)$, $\beta = \frac{d}{\sum_{i=1}^d 1/\beta_i}$ *and* $M > 0$, *let* $V_{\beta,d} = V_{\beta,d}(\vec{R}, \vec{\beta}, M)$ *be the class of functions* $f : [0,1]^d \to \mathcal{R}$ *such that* $\|f\|_\infty \le M$ *and for* $\delta > 0$

$$\|\Delta_{i,\delta}^{R_i} f\|_{L_2(\lambda)} \le M \delta^{\beta_i} \ , \quad 1 \le i \le d \ ,$$

where

$$\Delta_{i,\delta}^{R} f(x) = \sum_{r=0}^{R} \binom{R}{r} (-1)^{R-r} f(x + r\delta e_i) \ ,$$

and e_i *is the* i^{th} *unit vector in* $\mathcal{R}^d$*.*

These classes are just Example 5 in Yang [16]. We assume that $M, M_0 > 1/2$.

Definition 5 *Denote one of the classes* $\mathrm{Lip}_{\beta,d}$ *and* $V_{\beta,d}$ *by* $\mathcal{F}_\beta$*. Let* $\mathcal{D}_\beta$ *be the class of distributions of* (X,Y) *such that*
(i) X *is uniformly distributed on* $[0,1]^d$,
(ii) $\eta \in \mathcal{F}_\beta$*.*

It is well-known, that there exist regression function estimates $\{\eta_n\}$, which satisfy

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

(see, e.g., [16] and Barron, Birgé and Massart [6]). (This remains true replacing condition (i) in Definition 5, e.g., with the assumption that μ is in a class of distributions with uniformly bounded density funtions. Note that the rate depends only on β and d.) Thus for the plug-in rules $\{g_n\}$

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

that is, $\sup_{(X,Y)\in\mathcal{D}_\beta}(\mathbf{E}L_n - L^*) = O\left(n^{-\frac{\beta}{2\beta+d}}\right)$. (The rate $n^{-\frac{\beta}{2\beta+d}}$ might be formulated as $n^{-\frac{1}{2+\delta}}$, where $\delta = d/\beta$, the *exponent of dimension* is just the exponent of $1/\epsilon$ in the ϵ-metric entropy of the classes, see [7].)

To handle the two types of classes together, let $\vec{\beta} = (\beta_1, \ldots, \beta_d) = (\beta, \ldots, \beta)$ for the classes $\mathrm{Lip}_{\beta,d}$. Thus

$$\beta = \frac{d}{\sum_{i=1}^d 1/\beta_i} \tag{6}$$

holds in all cases. (We may call β as *the exponent of global smoothness* following [7].)

Now we give conditions for a general function class $\mathcal{F}_\beta$, which assure lower rates of convergence for the corresponding distribution class $\mathcal{D}_\beta$. A subclass of regression functions will be indexed by the vectors

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

of $+1$ or -1 components, where $S_1, S_2, \ldots$ are introduced below. Denote the set of all such vectors by $\mathcal{C}$. Let μ be the uniform distribution on $[0,1]^d$.

Assumption 1 *With some $\vec{\beta}$ and β satisfying (6), for any probability distribution $\{p_j\}$ for every $j \in \mathcal{N}$, there is a function $m_j : \mathcal{R}^d \mapsto [0, 1/4]$ with support on the brick $I_j = \otimes_{i=1}^d [0, p_j^{\beta/\beta_i}]$, and S_j disjoint translates $A_{j,k} = I_j + a_{j,k}$ ($k = 1, \ldots, S_j$) in $[0,1]^d$, such that these are disjoint even for different j and j', and for every $c \in \mathcal{C}$,*

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

where $m_{j,k}(x) = m_j(x - a_{j,k})$. Also $\|m_j\|_2^2 \le K_1 p_j^{2\beta}\mu(I_j)$, $\|m_j\|_1 \ge K p_j^{\beta}\mu(I_j)$ and $K_2 p_j \le S_j \mu(I_j) \le p_j$ for some constants K,K_1,$K_2 > 0$ (which are independent of $\{p_j\}$ and j).

Note that $\mu(I_j) = p_j^d$ by (6).

Theorem 1 *If Assumption 1 holds for a function class $\mathcal{F}_\beta$, then the sequence*

$$\left\{ a_n = n^{-\frac{\beta}{2\beta+d}} \right\}$$

is a lower rate of convergence for the corresponding distribution class $\mathcal{D}_\beta$.

Assumption 1 is similar to A'2(k) in Birgé [7]. In this form, it will be required only for the individual lower rates. It seems from the proof (choice of $\{p_j\}$) that for the minimax rates above, we only use the weaker form:

Assumption 2 *With some $\vec{\beta}$ and β satisfying (6), for any $\epsilon \in (0,1]$, there is a function $m_\epsilon : \mathcal{R}^d \mapsto [0,1/4]$ with support on the brick $I = I_\epsilon = \otimes_{i=1}^d [0, \epsilon^{1/\beta_i}]$, and $r = r_\epsilon$ disjoint translates $I + a_j$ ($j = 1, \ldots, r$) in $[0,1]^d$, such that for every $c = (c_1, c_2, \ldots, c_r) \in \{-1, +1\}^r$,*

$$\eta^{(c)}(x) = \frac{1}{2} + \sum_{j=1}^{r} c_j m_\epsilon(x - a_j) \in \mathcal{F}_\beta \ .$$

Also $\|m_\epsilon\|_2^2 \leq K_1 \epsilon^2 \mu(I)$, $\|m_\epsilon\|_1 \geq K\epsilon\mu(I)$ and $K_2 \leq r\mu(I) \leq 1$ for some constants K,K_1,$K_2 > 0$.

This is almost the same as Assumption 3 in Yang [16] and A2(k) in Birgé [7] in case of polinomial metric entropy, and Theorem 1 is analogous to Yang's Theorem 2. Both Yang's and our theorems give as a special case:

Corollary 1 *If $\mathcal{F}_\beta$ is $\mathrm{Lip}_{\beta,d}$ or $V_{\beta,d}$, then the sequence*

$$\left\{ a_n = n^{\frac{\beta}{2\beta+d}} \right\}$$

is a lower rate of convergence for the class $\mathcal{D}_\beta$.

Our main result is the following extention of Theorem 1 to individual lower rates (see Antos [2] for $\mathrm{Lip}_{\beta,d}$ if $q = \infty$):

Theorem 2 *Let $\{b_n\}$ be an arbitrary positive sequence tending to zero. If Assumption 1 holds for a function class $\mathcal{F}_\beta$, then the sequence*

$$\left\{ b_n a_n = b_n n^{-\frac{2\beta}{2\beta+d}} \right\}$$

is an individual lower rate of convergence for the corresponding distribution class $\mathcal{D}_\beta$.

REMARK 1. Applying for the sequence $\{\sqrt{b_n}\}$, Theorem 2 implies that for all $\{g_n\}$ there is $(X, Y) \in \mathcal{D}_\beta$ 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 5.

Focusing again on the classes $\mathrm{Lip}_{\beta,d}$ and $V_{\beta,d}$:

Corollary 2 *Let $\{b_n\}$ be an arbitrary positive sequence tending to zero. If $\mathcal{F}_\beta$ is $\mathrm{Lip}_{\beta,d}$ or $V_{\beta,d}$, then the sequence*

$$\left\{ b_n a_n = b_n n^{-\frac{2\beta}{2\beta+d}} \right\}$$

is an individual lower rate of convergence for the class $\mathcal{D}_\beta$.

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

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

that is, $\sup_{(X,Y)\in\mathcal{D}}(\mathbf{E}L_n - L^*) = O(c_n)$, 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 $n^{-\frac{\beta}{2\beta+d}}$ is an upper rate of convergence, and thus also an individual upper rate of convergence for $\mathcal{D}_\beta$. While Theorem 1 shows only that there is no upper rate of convergence for $\mathcal{D}_\beta$ better than $n^{-\frac{\beta}{2\beta+d}}$, it follows from Theorem 2 that $n^{-\frac{\beta}{2\beta+d}}$ is even the optimal individual upper rate for $\mathcal{D}_\beta$ in the sense, that there doesn't exist an individual upper rate c_n of convergence for $\mathcal{D}_\beta$, which satisfies

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

Moreover (3) and (4) imply

$$\inf_{\{g_n\}} \sup_{(X,Y)\in\mathcal{D}_\beta} \limsup_{n\to\infty} \frac{\mathbf{E}L_n - L^*}{n^{-\frac{\beta}{2\beta+d}}} = 0 , \tag{7}$$

which shows that Theorem 2 cannot be improved by dropping b_n. This shows the strange nature of individual lower bounds, that while every sequence tending to zero faster than $n^{-\frac{\beta}{2\beta+d}}$ is an individual lower rate for $\mathcal{D}_\beta$, $n^{-\frac{\beta}{2\beta+d}}$ 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 $\vec{Y} = (Y_1, \ldots, Y_l)$

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

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

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

(see [8]). One can verify that

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

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

$$Z_i \stackrel{\text{def}}{=} (2Y_i - 1)\log\frac{1+2u_i}{1-2u_i} \ .$$

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 $|\log T| = \left|\sum_{i\leq l} Z_i\right|$, we get

$$\begin{aligned} \mathbf{E}\{|\log T|\} &= \mathbf{E}\left\{ \left| \sum_{i\leq l} Z_i \right| \right\} \leq \sqrt{\mathbf{E}\left\{ \left(\sum_{i\leq l} Z_i \right)^2 \right\}} = \sqrt{\mathbf{E}\left\{ \sum_i Z_i^2 + \sum_{i\neq i'} Z_i Z_{i'} \right\}} \\ &= \sqrt{\sum_i \mathbf{E}\{Z_i^2\} + \sum_{i\neq i'} \mathbf{E}\{Z_i Z_{i'}\}} \ . \end{aligned}$$

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

$$\begin{aligned} \left| \log\frac{1+2x}{1-2x} \right| &= |\log(1+2x) - \log(1-2x)| = \log(1+2|x|) - \log(1-2|x|) \\ &\leq 2|x| + \log 4 \cdot 2|x| \leq 5|x| \ , \end{aligned}$$

on the one hand

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

and on the other hand

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

so

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

Hence

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

Thus

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

By choosing

$$
q = \frac{1}{1+e^{1+5\sqrt{\sum_i u_i^2 + 4\sum_{i\neq i'} u_i^2u_{i'}^2}}} \ ,
$$

$$
\begin{aligned}
\mathbf{E}\{\pi\} &\ge \frac{1}{1+\exp(1+5\sqrt{\sum_i u_i^2 + 4\sum_{i\neq i'} u_i^2u_{i'}^2})}\frac{1}{1+5\sqrt{\sum_i u_i^2 + 4\sum_{i\neq i'} u_i^2u_{i'}^2}} \\
&\ge \frac{1}{(e+1)e^{10\sqrt{\sum_i u_i^2 + 4\sum_{i\neq i'} u_i^2u_{i'}^2}}} \ge \frac{1}{4}e^{-10\sqrt{\sum_i u_i^2 + 4\sum_{i\neq i'} u_i^2u_{i'}^2}} \ . \quad \square
\end{aligned}
$$

PROOF OF THEOREM 1. The method of this proof differs from that of Yang, it can be easily modified to individual lower bound in Theorem 2. Assumption 1 and $0 \le \eta^{(c)} \le 1$ imply that each distribution (X,Y) with $X \sim \mathrm{Unif}[0,1]^d$, $Y \in \{0,1\}$ and $\mathbf{E}\{Y|X=x\} = \eta^{(c)}(x)$

for all $x \in [0,1]^d$ for some $c \in \mathcal{C}$ is contained in $\mathcal{D}_\beta$, which implies

$$\limsup_{n\to\infty} \inf_{g_n} \sup_{(X,Y)\in\mathcal{D}_\beta} \frac{\mathbf{E}L_n - L^*}{a_n} \geq$$
$$\limsup_{n\to\infty} \inf_{g_n} \sup_{(X,Y):X\sim\mathrm{Unif}[0,1]^d, \mathbf{E}\{Y|X=x\}=\eta^{(c)}(x), c\in\mathcal{C}} \frac{\mathbf{E}L_n - L^*}{a_n} . \tag{8}$$

Let g_n be an arbitrary rule. By definition, $\{I_{A_{j,k}}/2 : j,k\}$ is an orthogonal system in $L_2(\nu)$ for the measure $\nu(A) = \int_A \sum_{j,k} m_{j,k}\, d\mu$, 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

$$\begin{aligned}
\hat{c}_{n,j,k} &= \frac{\int (g_n - 1/2) I_{A_{j,k}}/2\, d\nu}{\int \left(I_{A_{j,k}}/2\right)^2 d\nu} = 2\frac{\int_{A_{j,k}} (g_n - 1/2)\, d\nu}{\int_{A_{j,k}} 1\, d\nu} \\
&= 2\frac{\int_{A_{j,k}} (g_n - 1/2) m_{j,k}\, d\mu}{\int_{A_{j,k}} m_{j,k}\, d\mu} = 2\frac{\int_{A_{j,k}} g_n m_{j,k}\, d\mu}{\int_{A_{j,k}} m_{j,k}\, d\mu} - 1 .
\end{aligned}$$

(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)

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

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

$$L_n - L^* \geq \frac{K}{2} \sum_{j,k} I_{\{\tilde{c}_{n,j,k} \neq c_{j,k}\}} p_j^{\beta+d} .$$

This proves

$$\mathbf{E}L_n - L^* \geq \frac{K}{2} R_n(c) \ , \tag{9}$$

where

$$R_n(c) = \sum_{j: np_j^{2\beta+d} \leq 1} \sum_{k=1}^{S_j} p_j^{\beta+d} \cdot \mathbf{P}\{\tilde{c}_{n,j,k} \neq c_{j,k}\} \ . \tag{10}$$

Equations (8) and (9) imply

$$\limsup_{n\to\infty} \inf_{g_n} \sup_{(X,Y)\in\mathcal{D}^{(\beta,M)}} \frac{EL_n - L^*}{a_n} \geq \frac{K}{2} \limsup_{n\to\infty} \inf_{g_n} \sup_{c\in\mathcal{C}} \frac{R_n(c)}{a_n} \ . \tag{11}$$

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, $\ldots$, which satisfy $\mathbf{P}\{C_{1,1} = 1\} = \mathbf{P}\{C_{1,1} = -1\} = 1/2$. Next we derive a lower bound for

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

The sign $\tilde{c}_{n,j,k}$ 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

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

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

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

By Jensen-inequality

$$
\begin{aligned}
\mathbf{P}\{\bar C_{n,j,k} \neq C_{j,k}\} &= \mathbf{E}\{\mathbf{P}\{\bar C_{n,j,k} \neq C_{j,k} | X_1,\dots,X_n\}\} \\
&\geq \frac{1}{4}\mathbf{E}\left\{ e^{-10\sqrt{\sum_i m_{j,k}^2(X_i)+4\sum_{i\neq i'} m_{j,k}^2(X_i)m_{j,k}^2(X_{i'})}} \right\} \\
&\geq \frac{1}{4} e^{-10\mathbf{E}\left\{ \sqrt{\sum_i m_{j,k}^2(X_i)+4\sum_{i\neq i'} m_{j,k}^2(X_i)m_{j,k}^2(X_{i'})} \right\}} \\
&\geq \frac{1}{4} e^{-10\sqrt{\sum_i \mathbf{E}\{m_{j,k}^2(X_i)\}+4\sum_{i\neq i'} \mathbf{E}\{m_{j,k}^2(X_i)m_{j,k}^2(X_{i'})\}}} \\
&\geq \frac{1}{4} e^{-10\sqrt{K_1 n p_j^{2\beta+d}+4K_1^2 n(n-1)p_j^{4\beta+2d}}}
\end{aligned}
$$

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

$$
\mathbf{P}\{\bar C_{n,j,k} \neq C_{j,k}\} \geq \frac{e^{-10\sqrt{K_1+4K_1^2}}}{4} > \frac{e^{-20K_1-3}}{4} \ ,
$$

and

$$
\begin{aligned}
\mathbf{E}R_n(C) &\geq \sum_{j: np_j^{2\beta+d}\leq 1} \sum_{k=1}^{S_j} p_j^{\beta+d} \cdot \mathbf{P}\{\bar C_{n,j,k} \neq C_{j,k}\} \\
&\geq \frac{e^{-20K_1-3}}{4} \sum_{j: np_j^{2\beta+d}\leq 1} S_j p_j^{\beta+d} \geq K_3 \sum_{j: np_j^{2\beta+d}\leq 1} p_j^{\beta+1} \ , \qquad (12)
\end{aligned}
$$

where $K_3 = \frac{e^{-20K_1-3}}{4}K_2$. Setting $p_j = p_{j,n} = n^{-\frac{1}{2\beta+d}}$ for $j \leq n^{\frac{1}{2\beta+d}}$,

$$
\mathbf{E}R_n(C) \geq K_3 \lfloor n^{\frac{1}{2\beta+d}} \rfloor n^{-\frac{\beta+1}{2\beta+d}} = K_3 a_n(1-o(1)) \ ,
$$

so

$$
\limsup_{n\to\infty} \inf_{g_n} \sup_{c\in\mathcal{C}} \frac{R_n(c)}{a_n} \geq \limsup_{n\to\infty} \inf_{g_n} \frac{\mathbf{E}R_n(C)}{a_n} \geq K_3 > 0 \ . \qquad (13)
$$

This together with (11) implies the assertion. □

Proof of Theorem 2. We use the notations and results of the proof of Theorem 1. Now we have by (9)

$$
\inf_{\{g_n\}} \sup_{(X,Y)\in\mathcal{D}_\beta} \limsup_{n\to\infty} \frac{\mathbf{E}L_n - L^*}{b_n a_n} \geq \frac{K}{2} \inf_{\{g_n\}} \sup_{c\in\mathcal{C}} \limsup_{n\to\infty} \frac{R_n(c)}{b_n a_n} \ . \qquad (14)
$$

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 $b_{n_t} \leq 2^{-t}$ and $a_{n_t}^{1/\beta} \leq 2^{-t}$. Define q_t such that

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

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

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

by $a_{n_t}^{1/\beta} \le 2^{-t}$, and specially for $n = n_s$ (12) implies

$$\mathbf{E} R_{n_s}(C) \ge K_3 \sum_{j: n_s p_j^{2\beta+d} \le 1} p_j^{\beta+1} \ge \frac{K_3}{2^{\beta}} \sum_{t \ge s} b_{n_t} a_{n_t} \ge \frac{K_3}{2^{\beta}} b_{n_s} a_{n_s} \ . \tag{15}$$

We finish the proof in the spirit of Lemma 1 in Antos and Lugosi [5]. Using (15) one gets

$$\begin{aligned}
\inf_{\{g_n\}} \sup_{c \in \mathcal{C}} \limsup_{n \to \infty} \frac{R_n(c)}{b_n a_n} &\ge \inf_{\{g_n\}} \sup_{c \in \mathcal{C}} \limsup_{s \to \infty} \frac{R_{n_s}(c)}{b_{n_s} a_{n_s}} \\
&\ge \frac{K_3}{2^{\beta}} \inf_{\{g_n\}} \sup_{c \in \mathcal{C}} \limsup_{s \to \infty} \frac{R_{n_s}(c)}{\mathbf{E} R_{n_s}(C)} \\
&\ge \frac{K_3}{2^{\beta}} \inf_{\{g_n\}} \mathbf{E} \left(\limsup_{s \to \infty} \frac{R_{n_s}(C)}{\mathbf{E} R_{n_s}(C)} \right) \ .
\end{aligned}$$

Because of (12) and the fact that for all $c \in \mathcal{C}$

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

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

$$\inf_{\{g_n\}} \sup_{c \in \mathcal{C}} \limsup_{n \to \infty} \frac{R_n(c)}{b_n a_n} \ge \frac{K_3}{2^{\beta}} \inf_{\{g_n\}} \limsup_{s \to \infty} \mathbf{E} \left(\frac{R_{n_s}(C)}{\mathbf{E} R_{n_s}(C)} \right) = \frac{K_3}{2^{\beta}} > 0 \ .$$

This together with (14) implies the assertion. □

PROOF OF COROLLARY 1 AND 2. We must prove that the classes $\mathrm{Lip}_{\beta,d}$ and $V_{\beta,d}$ satisfy Assumption 1. The parameters $\vec{\beta}$ and β are given as in the text satisfying (6). For any $\{p_j\}$, we give the required functions m_j and sets $A_{j,k}$.

First we pack the disjoint sets $A_{j,k}$ into $[0,1]^d$ in the following way: Assume for simplicity, that $\beta_1 = \min_i \beta_i$. (If not, then the index of the minimal β_i takes the role of the first dimension in the construction below.) For a given $\{p_j\}$, let $\{B_j\}$ be a partition of $[0,1]$ such that B_j is an interval of length p_j. We pack disjoint translates of I_j into the brick

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

This gives

$$S_j = \left\lfloor \frac{p_j}{p_j^{\beta/\beta_1}} \right\rfloor \prod_{i=2}^{d} \left\lfloor \frac{1}{p_j^{\beta/\beta_i}} \right\rfloor ,$$

and since, for $x \geq 1$, $\lfloor x \rfloor \geq x/2$ and $p_j/p_j^{\beta/\beta_1} \geq 1$ by $\beta_1 \leq \beta$,

$$K_2 p_j = \frac{p_j^{d+1}}{2^d p_j^{\sum_{i=1}^d \beta/\beta_i}} = \frac{p_j p_j^d}{2 p_j^{\beta/\beta_1}} \prod_{i=2}^{d} \frac{1}{2 p_j^{\beta/\beta_i}} \leq S_j \mu(I_j) \leq \frac{p_j p_j^d}{p_j^{\beta/\beta_1}} \prod_{i=2}^{d} \frac{1}{p_j^{\beta/\beta_i}} = p_j ,$$

where $K_2 = 2^{-d}$.

Let μ be the uniform distribution on $[0,1]^d$ and $R = \max_i R_i$ for $V_{\beta,d}$. Choose a function $m : \mathcal{R}^d \to [0,1/4]$ such that

(I) the support of m is a subset of $[0,1]^d$,

(II) $\|m\|_1 > 0$,

(III) $m \in \mathrm{Lip}_{\beta,d}(q, M_0 - 1/2, \ldots, M_R/7)$, for $\mathcal{F}_\beta = \mathrm{Lip}_{\beta,d}$ and
$m \in V_{\beta,d}(\vec{R}, \vec{\beta}, (M-1/2)/(1+(4R)^{R+1}))$, for $\mathcal{F}_\beta = V_{\beta,d}$.

Set

$$m_j(x) = p_j^\beta m(< p_j^{-\beta/\beta_1}, \ldots, p_j^{-\beta/\beta_d} > \cdot x) ,$$

where $< \ldots >$ denotes a diagonal matrix. Thus m_j is a contraction of m from $[0,1]^d$ to I_j. Now we have $\|m_j\|_2^2 = p_j^{2\beta} \mu(I_j) \|m\|_2^2$ and $\|m_j\|_1 = p_j^\beta \mu(I_j) \|m\|_1$, so $K_1 = \|m\|_2^2$ and $K = \|m\|_1$ will do.

We only have to check that because of (III), for every $c \in \mathcal{C}$,

$$\eta^{(c)}(x) = \frac{1}{2} + \sum_{j=1}^{\infty} \sum_{k=1}^{S_j} c_{j,k} m_{j,k}(x) \in \mathcal{F}_\beta .$$

Case $\mathcal{F}_\beta = \mathrm{Lip}_{\beta,d}$.

Note that the functions $m_{j,k}$ have disjoint support, and this holds also for their derivatives. If $q < \infty$, then for $R' = 0$,

$$\|\sum_{j,k} c_{j,k} m_{j,k}\|_q^q = \sum_{j,k} \|m_j\|_q^q = \sum_{j,k} p_j^{q\beta+d} \|m\|_q^q \leq \|m\|_q^q \sum_{j,k} p_j^d \leq \left(M_0 - \frac{1}{2}\right)^q \ ,$$

thus

$$\|\eta^{(c)}\|_q \leq \frac{1}{2} + \|\sum_{j,k} c_{j,k} m_{j,k}\|_q \leq M_0 \ .$$

For $R' = 1,\ldots,R$, for every γ with $\sum_{i=1}^d \gamma_i = R'$,

$$\begin{aligned} \|D^\gamma \eta^{(c)}\|_q^q &= \|\sum_{j,k} c_{j,k} D^\gamma m_{j,k}\|_q^q = \sum_{j,k} \|D^\gamma m_j\|_q^q = \sum_{j,k} p_j^{q\beta - qR' + d} \|D^\gamma m\|_q^q \\ &\leq M_{R'}^q \sum_{j,k} p_j^{q\beta - qR' + d} \leq M_{R'}^q \sum_{j,k} p_j^d \leq M_{R'}^q \ . \end{aligned}$$

Moreover, for every γ with $\sum_{i=1}^d \gamma_i = R$ for every $\delta \in \mathcal{R}^d$, with $\Delta_\delta f(x) = f(x+\delta) - f(x)$,

$$\begin{aligned} &\|D^\gamma \eta^{(c)}(x+\delta) - D^\gamma \eta^{(c)}(x)\|_{L_q(\lambda)} = \|\sum_{j,k} c_{j,k} \Delta_\delta D^\gamma m_{j,k}\|_{L_q(\lambda)} \\ &\leq \|\sum_{j,k: p_j > 3\|\delta\|} c_{j,k} \Delta_\delta D^\gamma m_{j,k}\|_{L_q(\lambda)} + \|\sum_{j,k: p_j \leq 3\|\delta\|} c_{j,k} \Delta_\delta D^\gamma m_{j,k}\|_{L_q(\lambda)} \ . \end{aligned}$$

We can choose m such that its support is in $[1/3, 2/3]^d$. Now if $\|\delta\| < p_j/3$ then the support of $\Delta_\delta D^\gamma m_{j,k}$ is in $A_{j,k}$, hence they are disjoint. Thus for the first term, with $M' = M_R/7$,

$$\begin{aligned} &\|\sum_{j,k: p_j > 3\|\delta\|} c_{j,k} \Delta_\delta D^\gamma m_{j,k}\|_{L_q(\lambda)}^q \\ &= \sum_{j,k: p_j > 3\|\delta\|} \|\Delta_\delta D^\gamma m_j\|_q^q \leq \sum_{j,k} p_j^{q\beta - qR + d} \|\Delta_{p_j^{-1}\delta} D^\gamma m\|_q^q \\ &\leq \sum_{j,k} p_j^{q\beta - qR + d} M'^q \|p_j^{-1}\delta\|^{q\alpha} = M'^q \|\delta\|^{q\alpha} \sum_{j,k} p_j^d \leq M'^q \|\delta\|^{q\alpha} \ . \end{aligned}$$

For the second term,

$$\begin{aligned} &\|\sum_{j,k: p_j \leq 3\|\delta\|} c_{j,k} \Delta_\delta D^\gamma m_{j,k}\|_{L_q(\lambda)}^q \\ &= \|\sum_{j,k: p_j \leq 3\|\delta\|} c_{j,k} D^\gamma m_{j,k}(x+\delta) - \sum_{j,k: p_j \leq 3\|\delta\|} c_{j,k} D^\gamma m_{j,k}(x)\|_{L_q(\lambda)}^q \\ &\leq 2^{q-1} \|\sum_{j,k: p_j \leq 3\|\delta\|} c_{j,k} D^\gamma m_{j,k}(x+\delta)\|_{L_q(\lambda)}^q + 2^{q-1} \|\sum_{j,k: p_j \leq 3\|\delta\|} c_{j,k} D^\gamma m_{j,k}(x)\|_{L_q(\lambda)}^q \\ &= 2^q \|\sum_{j,k: p_j \leq 3\|\delta\|} c_{j,k} D^\gamma m_{j,k}\|_q^q = 2^q \sum_{j,k: p_j \leq 3\|\delta\|} \|D^\gamma m_j\|_q^q \\ &= 2^q \sum_{j,k: p_j \leq 3\|\delta\|} p_j^{q\beta - qR + d} \|D^\gamma m\|_q^q \leq 2^q M'^q \sum_{j,k: p_j \leq 3\|\delta\|} p_j^{q\alpha + d} \\ &\leq 2^q M'^q \sum_{j,k: p_j \leq 3\|\delta\|} (3\|\delta\|)^{q\alpha} p_j^d \leq 2^q 3^{q\alpha} M'^q \|\delta\|^{q\alpha} \sum_{j,k} p_j^d \leq 2^q 3^{q\alpha} M'^q \|\delta\|^{q\alpha} \ . \end{aligned}$$

This gives

$$\|D^\gamma\eta^{(c)}(x+\delta) - D^\gamma\eta^{(c)}(x)\|_{L_q(\lambda)} \leq (1+2\cdot 3^\alpha)M'\|\delta\|^\alpha \leq 7M'\|\delta\|^\alpha = M_R\|\delta\|^\alpha \ .$$

If $q=\infty$, then for $R'=0$, $\|\eta^{(c)}\|_\infty \leq 1/2 + \|m\|_\infty \leq M_0$. For $R'=1,\ldots,R$, for every γ with $\sum_{i=1}^d \gamma_i = R'$,

$$\|D^\gamma\eta^{(c)}\|_\infty = \|\sum_{j,k} c_{j,k} D^\gamma m_{j,k}\|_\infty = \sup_j \|D^\gamma m_j\|_\infty = \sup_j p_j^{\beta-R'}\|D^\gamma m\|_\infty \leq M_{R'} \ .$$

Moreover, for every $\delta \in \mathcal{R}^d$, introduce x' and x'' as a function of x the following way: If x and $x+\delta$ fall in the same $A_{j,k}$, then let $x' = x'' = x+\delta$. If x and $x+\delta$ fall in different bricks, then consider the segment $\overline{x(x+\delta)}$, and let x' and x'' be the intersections of this segment and the borders of the bricks x and $x+\delta$ fall into, respectively. Since $m_{j,k}$ is vanishing outside $A_{j,k}$, any of its partial derivatives (up to order R) is zero on the border of $A_{j,k}$, thus in both case $D^\gamma\eta^{(c)}(x') = D^\gamma\eta^{(c)}(x'')$ for every γ with $\sum_{i=1}^d \gamma_i = R$. Hence

$$\begin{aligned}
&\|D^\gamma\eta^{(c)}(x+\delta) - D^\gamma\eta^{(c)}(x)\|_\infty \\
&= \|D^\gamma\eta^{(c)}(x+\delta) - D^\gamma\eta^{(c)}(x'') + D^\gamma\eta^{(c)}(x') - D^\gamma\eta^{(c)}(x)\|_\infty \\
&\leq \|D^\gamma\eta^{(c)}(x+\delta) - D^\gamma\eta^{(c)}(x'')\|_\infty + \|D^\gamma\eta^{(c)}(x') - D^\gamma\eta^{(c)}(x)\|_\infty \ .
\end{aligned}$$

Using the fact that x and x' are in the same brick, and so the support of $D^\gamma m_{j,k}(x') - D^\gamma m_{j,k}(x)$ is in $A_{j,k}$, for the second term

$$\begin{aligned}
\|D^\gamma\eta^{(c)}(x') - D^\gamma\eta^{(c)}(x)\|_\infty &= \|\sum_{j,k} c_{j,k}(D^\gamma m_{j,k}(x') - D^\gamma m_{j,k}(x))\|_\infty \\
&= \sup_{j,k} \|(D^\gamma m_{j,k}(x') - D^\gamma m_{j,k}(x))\|_\infty \\
&= \sup_{j,k} \|p_j^{\beta-R}(D^\gamma m(p_j^{-1}(x'-a_{j,k})) - D^\gamma m(p_j^{-1}(x-a_{j,k})))\|_\infty \\
&\leq \sup_j p_j^{\beta-R} M_R \|p_j^{-1}(x'-x)\|^\alpha/7 \\
&\leq M_R\|\delta\|^\alpha/7 \ .
\end{aligned}$$

We get the same bound similarly for the first term, which gives

$$\|D^\gamma\eta^{(c)}(x+\delta) - D^\gamma\eta^{(c)}(x)\|_\infty \leq M_R\|\delta\|^\alpha \ .$$

(See also [15], p. 1045.)

Case $\mathcal{F}_\beta = V_{\beta,d}$.

Now $\|\eta^{(c)}\|_\infty \le 1/2 + \|m\|_\infty \le M$, and for $1 \le i \le d$

$$\begin{aligned}
\|\Delta_{i,\delta}^{R_i}\eta^{(c)}\|_{L_2(\lambda)} &= \|\sum_{j,k} c_{j,k}\Delta_{i,\delta}^{R_i} m_{j,k}\|_{L_2(\lambda)} \\
&\le \|\sum_{j,k:p_j^{\beta/\beta_i}>2R_i\delta} c_{j,k}\Delta_{i,\delta}^{R_i} m_{j,k}\|_{L_2(\lambda)} + \|\sum_{j,k:p_j^{\beta/\beta_i}\le 2R_i\delta} c_{j,k}\Delta_{i,\delta}^{R_i} m_{j,k}\|_{L_2(\lambda)} \ .
\end{aligned}$$

We can choose m such that its support is in $[1/2,1]^d$. Now, if $R_i\delta < p_j^{\beta/\beta_i}/2$ then the support of $\Delta_{i,\delta}^{R_i} m_{j,k}$ is in $A_{j,k}$, hence they are disjoint. For the first term, with $M' = (M-1/2)/(1+(4R)^{R+1})$,

$$\begin{aligned}
\|\sum_{j,k:p_j^{\beta/\beta_i}>2R_i\delta} c_{j,k}\Delta_{i,\delta}^{R_i} m_{j,k}\|_{L_2(\lambda)}^2 &= \sum_{j,k:p_j^{\beta/\beta_i}>2R_i\delta} \|\Delta_{i,\delta}^{R_i} m_{j,k}\|_2^2 \\
&\le \sum_{j,k} p_j^{2\beta+d}\|\Delta_{i,p_j^{-\beta/\beta_i}\delta}^{R_i} m\|_2^2 \\
&\le \sum_{j,k} p_j^{2\beta+d} M'^2 (p_j^{-\beta/\beta_i}\delta)^{2\beta_i} = M'^2\delta^{2\beta_i}\sum_{j,k} p_j^d \le M'^2\delta^{2\beta_i} \ .
\end{aligned}$$

For the second term,

$$\begin{aligned}
&\|\sum_{j,k:p_j^{\beta/\beta_i}\le 2R_i\delta} c_{j,k}\Delta_{i,\delta}^{R_i} m_{j,k}\|_{L_2(\lambda)}^2 \\
&= \left\|\sum_{j,k:p_j^{\beta/\beta_i}\le 2R_i\delta} c_{j,k}\sum_{r=0}^{R_i}\binom{R_i}{r}(-1)^{R_i-r} m_{j,k}(x+r\delta e_i)\right\|_{L_2(\lambda)}^2 \\
&= \left\|\sum_{r=0}^{R_i}\binom{R_i}{r}(-1)^{R_i-r}\sum_{j,k:p_j^{\beta/\beta_i}\le 2R_i\delta} c_{j,k} m_{j,k}(x+r\delta e_i)\right\|_{L_2(\lambda)}^2 \\
&\le (R_i+1)\sum_{r=0}^{R_i}\left\|\binom{R_i}{r}(-1)^{R_i-r}\sum_{j,k:p_j^{\beta/\beta_i}\le 2R_i\delta} c_{j,k} m_{j,k}(x+r\delta e_i)\right\|_{L_2(\lambda)}^2 \\
&= (R_i+1)\sum_{r=0}^{R_i}\binom{R_i}{r}^2\|\sum_{j,k:p_j^{\beta/\beta_i}\le 2R_i\delta} c_{j,k} m_{j,k}\|_2^2 = (R_i+1)\binom{2R_i}{R_i}\sum_{j,k:p_j^{\beta/\beta_i}\le 2R_i\delta}\|m_j\|_2^2 \\
&= (R_i+1)\binom{2R_i}{R_i}\sum_{j,k:p_j^{\beta/\beta_i}\le 2R_i\delta} p_j^{2\beta+d}\|m\|_2^2 \le (R_i+1)\binom{2R_i}{R_i} M'^2 \sum_{j,k:p_j^{\beta/\beta_i}\le 2R_i\delta} p_j^{2\beta+d} \\
&\le (R_i+1)\binom{2R_i}{R_i} M'^2 \sum_{j,k:p_j^{\beta/\beta_i}\le 2R_i\delta} (2R_i\delta)^{2\beta_i} p_j^d \le (R_i+1)\binom{2R_i}{R_i}(2R_i)^{2\beta_i} M'^2\delta^{2\beta_i}\sum_{j,k} p_j^d \\
&\le (4R_i)^{2R_i+2} M'^2\delta^{2\beta_i} \ .
\end{aligned}$$

This gives

$$\|\Delta_{i,\delta}^{R_i}\eta^{(c)}\|_{L_2(\lambda)} \le (1+(4R_i)^{R_i+1})M'\delta^{\beta_i} < M\delta^{\beta_i}\ . \quad \Box$$

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.: Lower bounds on the rate of convergence of nonparametric pattern recognition. In *Computational Learning Theory: 4th European Conference, EuroCOLT'99, Proceedings*, v.1572 of *LNAI/LNCS*, pp. 241–252, Springer, Berlin (1999)

[3] Antos, A.: *Performance limits of nonparametric estimators.* Ph.D. Thesis, Technical University of Budapest, Budapest, Hungary (1999)

[4] Antos, A., Györfi, L. and Kohler, M.: Lower bounds on the rate of convergence of nonparametric regression estimates. *Journal of Statistical Planning and Inference* In print Preprint No. 98-11 (1998), Universität Stuttgart

[5] Antos, A. and Lugosi, G.: Strong minimax lower bounds for learning. *Machine Learning* 30 (1998) 31–56 Economics Working Paper 197 (1997), Universitat Pompeu Fabra

[6] Barron, A., Birgé, L., and Massart, P.: Risk bounds for model selection via penalization. *Probability Theory and Related Fields*, 113 (1999) 301–413 Technical Report No. 95.54 (1995), Université Paris Sud.

[7] Birgé, L.: On estimating a density using Hellinger distance and some other strange facts. *Probability Theory and Related Fields* 71 (1986) 271–291

[8] Devroye, L. , Györfi, L. and Lugosi, G. : *A Probabilistic Theory of Pattern Recognition* Springer Verlag (1996)

[9] Ibragimov, I. A. and Khasminskii, R. Z.: On nonparametric estimation of regression. *Doklady Acad. Nauk SSSR* 252 (1980) 780–784

[10] Ibragimov, I. A. and Khasminskii, R. Z.: *Statistical Estimation: Asymptotic Theory.* Springer Verlag, New York (1981)

[11] Ibragimov, I. A. and Khasminskii, R. Z.: On the bounds for quality of nonparametric regression function estimation. *Theory of Probability and its Applications* 27 (1982) 81–94

[12] Korostelev, A. P. and Tsybakov, A. B.: *Minimax Theory of Image Reconstruction.* Springer Verlag, Berlin (1993)

[13] Mammen, E. and Tsybakov, A. B.: Smooth discrimination analysis. *Annals of Statistics* Revised

[14] Stone, C. J.: Consistent nonparametric regression. *Annals of Statistics* 5 (1977) 595–645

[15] Stone, C. J. : Optimal global rates of convergence for nonparametric regression. *Annals of Statistics* 10 (1982) 1040–1053

[16] Yang, Y.: Minimax nonparametric classification — Part I: Rates of convergence. *IEEE Transaction on Information Theory* Accepted