Lower bounds on the rate of convergence of nonparametric regression estimates [1]

András Antos[a], László Györfi[a], Michael Kohler[b,*]

[a] *Dept. of Computer Science and Information Theory, Technical University of Budapest, 1521 Stoczek u. 2, Budapest, Hungary*

[b] *Mathematisches Institut A, Universität Stuttgart, Pfaffenwaldring 57, D-70569 Stuttgart, Germany*

June 1, 1998

Abstract

We show that there exist individual lower bounds on the rate of convergence of nonparametric regression estimates, which are arbitrarily close to Stone's minimax lower bounds.

AMS classification: Primary 62G20; secondary 62G05.

Keywords: Individual rates of convergence, nonparametric regression.

1 Introduction

Let (X,Y), (X_1,Y_1), (X_2,Y_2), ... be independent identically distributed $\mathcal{R}^d \times \mathcal{R}$ –valued random variables with $\mathbf{E}Y^2 < \infty$. In regression analysis one wishes to estimate the (real) value of Y given the (d–dimensional) value of X, i.e., one wants to find a function f defined on the range of X so that $f(X)$ is "close" to Y. Assume that the main aim of the analysis is to minimize the mean squared error :

$$\min_f \mathbf{E}\{(f(X)-Y)^2\}. \tag{1}$$

[1] The first and the second authors' work was supported by a grant from the Hungarian Academy of Sciences (MTA SZTAKI). The third author's research was supported by DFG.

* Correspondence address: Mathematisches Institut A, Universität Stuttgart, Pfaffenwaldring 57, D-70569 Stuttgart, Germany.

Abbreviated title: Lower bounds for nonparametric regression.

Let $m(x) = \mathbf{E}(Y \mid X = x)$ be the regression function, and denote the distribution of X by μ. It is well–known, that for each measurable function $f : \mathcal{R}^d \to \mathcal{R}$ the relation

$$\mathbf{E}\{(f(X) - Y)^2\} = \int \mid f(x) - m(x) \mid^2 \mu(dx) + \mathbf{E}\{(m(X) - Y)^2\}$$

holds. Therefore the regression function m achieves the minimum in (1), and the mean squared error of an arbitrary function f is close to its minimum (1) if and only if

$$\|f - m\|^2 = \int |f(x) - m(x)|^2 \mu(dx)$$

is close to zero.

For the regression estimation problem, the distribution of (X, Y) (and therefore also m) 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 an estimate $m_n(x) = m_n(x, D_n)$ of $m(x)$ such that $\|m_n - m\|^2$ is small. In this article we study asymptotic properties of $\mathbf{E}\|m_n - m\|^2$.

It is well–known, that there exists estimates which are universally consistent, i.e., which satisfy

$$\mathbf{E}\|m_n - m\|^2 \to 0 \quad (n \to \infty)$$

for all distributions of (X, Y) with $\mathbf{E}Y^2 < \infty$. This was first shown in Stone (1977) for nearest neighbor estimates, and was later extended by a lot of additional results (see, e.g., Devroye et al (1994) for a list of references).

Unfortunately, there don't exist estimates for which $\mathbf{E}\|m_n - m\|^2$ tends to zero with a guaranteed rate of convergence for all distributions of (X, Y). Theorem 7.2 and Problem 7.2 in Devroye, Györfi, Lugosi (1996) and section 3 in Devroye and Wagner (1980) imply the following slow rate of convergence result: Let $\{a_n\}$ be a sequence of positive numbers converging to zero with $1/64 \geq a_1 \geq a_2 \geq \ldots$. For every sequence of regression estimates, there exists a distribution of (X, Y), such that X is uniformly distributed, $Y = m(X)$ and

$$\mathbf{E}\{\|m_n - m\|^2\} \geq a_n$$

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*

$$\limsup_{n\to\infty} \inf_{m_n} \sup_{(X,Y)\in\mathcal{D}} \frac{\mathbf{E}\{\|m_n - m\|^2\}}{a_n} > 0.$$

It was shown in Stone (1982), that for classes $\mathcal{D}^{(p,M)}$ of distributions, where (roughly speaking, see below for exact definition) $p \in \mathcal{N}$ and the regression function m is p–times continuously differentiable, the optimal lower rate of convergence is $n^{-\frac{2p}{2p+d}}$. For related results on the general minimax theory of statistical estimates see Ibragimov and Khasmiskii (1980, 1981, 1982), Korostelev and Tsybakov (1993).

In some sense, such lower bounds are not satisfactory. They do not tell us anything about the way the error decreases as the sample size is increased for a given regression problem. These bounds, for each n, give information about the maximal error within the class, but not about the behavior of the error for a single fixed distribution as the sample size n increases. In other words, the "bad" distribution, causing the largest error for an estimator, 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 estimators $\{m_n\}$ such that for *every* distribution in $\mathcal{D}^{(p,M)}$, the expected error $\mathbf{E}\{\|m_n - m\|^2\}$ decreases at an exponential rate in n.

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

Definition 2 *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*

$$\inf_{\{m_n\}} \sup_{(X,Y)\in\mathcal{D}} \limsup_{n\to\infty} \frac{\mathbf{E}\{\|m_n - m\|^2\}}{a_n} > 0 \ ,$$

where the infimum is taken over all sequences $\{m_n\}$ of estimates.

The concept of individual lower rate has been introduced in Birgé (1986) concerning density estimation.

We will show that for every sequence $\{b_n\}$ tending to zero $b_n n^{-\frac{2p}{2p+d}}$ 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.

The main result is formulated in section 2 and proved in section 3.

2 Main result

Next we will give a precise definition of the class $\mathcal{D}^{(p,M)}$, for which we derive (individual) lower rates of convergence.

Definition 3 *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 4 *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) $Y = m(X) + N$, where X and N are independent and N is standard normal,
(iii) $m \in \mathcal{F}^{(p,M)}$.

It was shown in Stone (1982), that the optimal lower rate of convergence for $\mathcal{D}^{(p,M)}$ is $n^{-\frac{2p}{2p+d}}$. Our main result is the following theorem, which states that there even exist individual lower rates of convergence for $\mathcal{D}^{(p,M)}$ arbitrarily close to $n^{-\frac{2p}{2p+d}}$.

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

$$\limsup_{n\to\infty} \frac{\mathbf{E}\{||m_n - m||^2\}}{\sqrt{b_n} a_n} > 0 \ ,$$

thus

$$\limsup_{n\to\infty} \frac{\mathbf{E}\{||m_n - m||^2\}}{b_n a_n} = \infty \ .$$

Remark 2. It is well–known, that there exist estimates m_n, which satisfy

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

(see, e.g., Barron, Birge and Massart (1995)). Call a sequence c_n an *individual upper rate of convergence for a class $\mathcal{D}$* of distributions of (X, Y), if there exist estimates m_n which satisfy

$$\sup_{(X,Y)\in\mathcal{D}} \limsup_{n\to\infty} \frac{\mathbf{E}\{\|m_n - m\|^2\}}{c_n} < \infty.$$

Then (2) implies that $n^{-\frac{2p}{2p+d}}$ is an individual upper rate of convergence for $\mathcal{D}^{(p,M)}$. It follows from Theorem 1 that $n^{-\frac{2p}{2p+d}}$ 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 satisfy

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

REMARK 3. The proof of Theorem 1 also implies that $n^{-\frac{2p}{2p+d}}$ is a lower rate of convergence for $\mathcal{D}^{(p,M)}$. To see this, set

$$p_j = p_{j,n} = \left(\frac{1}{n}\right)^{\frac{1}{2p+d}} \text{ for } j \leq n^{\frac{1}{2p+d}}.$$

Then (4) and (8) imply the assertion.

REMARK 4. One can show the same (individual) lower rate of convergence result for classes of distributions of (X, Y) with $|Y|$ bounded a.s. For example, Theorem 1 remains valid if one replaces $\mathcal{D}^{(p,M)}$ by the class of all distributions of (X, Y) which satisfy

(I') X is uniformly distributed on $[0,1]^d$,

(II') $Y \in \{0,1\}$ a.s. and $\mathbf{P}\{Y = 1|X = x\} = 1 - \mathbf{P}\{Y = 0|X = x\} = m(x)$ for all $x \in [0,1]$,

(III') $m \in \mathcal{F}^{(p,M)}$ and $m(x) \in [0,1]$ for all $x \in [0,1]^d$.

This can be shown by using in the proof of Theorem 1 results of section 2 in Stone (1982) instead of (7) .

3 Proofs

The proof of Theorem 1 applies the following lemma:

Lemma 1 *Let u be a l-dimensional real vector, let C be a zero mean random variable taking values in $\{-1, +1\}$, and let N be a l-dimensional standard normal random variable, independent of C, and*

$$Z = Cu + N.$$

Then the error probability of the Bayes decision for C based on Z is

$$L^* = \Phi(-\|u\|),$$

where Φ is the standard normal distribution function.

PROOF. The case $u = 0$ is obvious. For $u \neq 0$ the Bayes decision for C based on Z is given by

$$g^*(z) = \begin{cases} 1 & \text{if } \mathbf{P}\{C = 1|Z = z\} > \mathbf{P}\{C = -1|Z = z\} \ , \\ -1 & \text{otherwise.} \end{cases}$$

Let φ be the density of a l–dimensional standard normal random variable. The conditional density of Z given $C = 1$ is $\varphi(z - u)$, and given $C = -1$ is $\varphi(z + u)$, therefore the equivalent version of Bayes decision is

$$g^*(z) = \begin{cases} 1 & \text{if } \varphi(z - u) > \varphi(z + u) \ , \\ -1 & \text{otherwise.} \end{cases}$$

This together with

$$\varphi(z-u) > \varphi(z+u) \Leftrightarrow z \text{ is closer to } u \text{ than to } -u \Leftrightarrow (u,z) > 0$$

implies that $g^*(z)$ is the sign of the inner product (u,z).

Set $\tilde{N} = (u,N)/\|u\|$. Then $\tilde{N}$ is a one–dimensional standard normal variable independent of C, and $Z = Cu + N$ implies

$$\frac{(u,Z)}{\|u\|} = C \cdot \|u\| + \frac{(u,N)}{\|u\|} = C \cdot \|u\| + \tilde{N}.$$

Therefore the error of the Bayes decision for C based on Z is given by

$$\begin{aligned} L^* &= \mathbf{P}\{C=-1, (u,Z)>0\} + \mathbf{P}\{C=+1,(u,Z)<0\} \\ &= \mathbf{P}\{C=-1\}\cdot\mathbf{P}\{(u,Z)>0|C=-1\} + \mathbf{P}\{C=+1\}\cdot\mathbf{P}\{(u,Z)<0|C=+1\} \\ &= \frac{1}{2}\mathbf{P}\{C\cdot\|u\|+\tilde{N}>0|C=-1\} + \frac{1}{2}\mathbf{P}\{C\cdot\|u\|+\tilde{N}<0|C=+1\} \\ &= \frac{1}{2}\mathbf{P}\{\tilde{N}>\|u\|\} + \frac{1}{2}\mathbf{P}\{\tilde{N}<-\|u\|\} = \Phi(-\|u\|)\ . \end{aligned}$$

□

Proof of Theorem 1. 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

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

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 center of $A_{j,k}$. Choose a function $g : \mathcal{R}^d \to \mathcal{R}$ such that
(I) the support of g is a subset of $[-\frac{1}{2},\frac{1}{2}]^d$,
(II) $\int g^2(x)\,dx > 0$,
(III) $g \in \mathcal{F}^{(p,M2^{\beta-1})}$.
The class of regression 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

$$m^{(c)}(x) = \sum_{j=1}^{\infty} \sum_{k=1}^{S_j} c_{j,k} g_{j,k}(x),$$

where

$$g_{j,k}(x) = p_j^p g(p_j^{-1}(x - a_{j,k})).$$

Then it is easy to check (cf. Stone (1982), p. 1045) that because of (III)

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

Hence, each distribution (X, Y) with $Y = m^{(c)}(X) + N$ for some $c \in \mathcal{C}$ is contained in $\mathcal{D}^{(p,M)}$, which implies

$$\inf_{\{m_n\}} \sup_{(X,Y)\in\mathcal{D}^{(p,M)}} \limsup_{n\to\infty} \frac{\mathbf{E}\{\|m_n - m\|^2\}}{b_n a_n} \geq$$

$$\inf_{\{m_n\}} \sup_{(X,Y):Y=m^{(c)}(X)+N, c\in\mathcal{C}} \limsup_{n\to\infty} \frac{\mathbf{E}\{\|m_n - m^{(c)}\|^2\}}{b_n a_n} \ . \tag{3}$$

Let m_n be an arbitrary estimate. By definition, $\{g_{j,k} : j, k\}$ is an orthogonal system in L_2, therefore the projection $\hat{m}_n$ of m_n to $\{m^{(c)} : c \in \mathcal{C}\}$ is given by

$$\hat{m}_n(x) = \sum_{j,k} \hat{c}_{n,j,k} g_{j,k}(x),$$

where

$$\hat{c}_{n,j,k} = \frac{\int_{A_{j,k}} m_n(x) g_{j,k}(x)\, dx}{\int_{A_{j,k}} g_{j,k}^2(x)\, dx}.$$

Let $c \in \mathcal{C}$ be arbitrary. Then

$$\|m_n - m^{(c)}\|^2 \geq \|\hat{m}_n - m^{(c)}\|^2 = \sum_{j,k} \int_{A_{j,k}} (\hat{c}_{n,j,k} g_{j,k}(x) - c_{j,k} g_{j,k}(x))^2\, dx$$

$$= \sum_{j,k} \int_{A_{j,k}} (\hat{c}_{n,j,k} - c_{j,k})^2 g_{j,k}^2(x)\, dx = \int g^2(x)\, dx \sum_{j,k} (\hat{c}_{n,j,k} - c_{j,k})^2 p_j^{2p+d}.$$

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

$$|\hat{c}_{n,j,k} - c_{j,k}| \geq |\tilde{c}_{n,j,k} - c_{j,k}|/2,$$

we get

$$\|m_n - m^{(c)}\|^2 \geq \int g^2(x)\, dx \frac{1}{4} \sum_{j,k} (\tilde{c}_{n,j,k} - c_{j,k})^2 p_j^{2p+d} \geq \int g^2(x)\, dx \sum_{j,k} I_{\{\tilde{c}_{n,j,k} \neq c_{j,k}\}} p_j^{2p+d}.$$

This proves

$$\mathbf{E}\{\|m_n - m^{(c)}\|^2\} \geq \int g^2(x)\,dx \cdot R_n(c)\ , \tag{4}$$

where

$$R_n(c) = \sum_{j:np_j^{2p+d}\leq 1} \sum_{k=1}^{S_j} p_j^{2p+d} \cdot \mathbf{P}\{\tilde{c}_{n,j,k} \neq c_{j,k}\}. \tag{5}$$

(3) and (4) imply

$$\inf_{\{m_n\}} \sup_{(X,Y)\in\mathcal{D}^{(p,M)}} \limsup_{n\to\infty} \frac{E\{\|m_n - m\|^2\}}{b_n a_n} \geq \int g^2(x)\,dx \inf_{\{m_n\}} \sup_{c\in\mathcal{C}} \limsup_{n\to\infty} \frac{R_n(c)}{b_n a_n} \tag{6}$$

To bound the last term, we fix a sequence $\{m_n\}$ of estimates and choose $c \in \mathcal{C}$ randomly. Let $C_{1,1}, \ldots, C_{1,S_1}, C_{2,1}, \ldots, C_{2,S_2}, \ldots$ be a sequence of independent identically distributed random variables independent of (X_1, N_1), (X_2, N_2), ..., which satisfy

$$\mathbf{P}\{C_{1,1} = 1\} = \mathbf{P}\{C_{1,1} = -1\} = \frac{1}{2}.$$

Set

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

Next we derive a lower bound for

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

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

$$(Y_{i_1}, \ldots, Y_{i_l}) = C_{j,k} \cdot (g_{j,k}(X_{i_1}), \ldots, g_{j,k}(X_{i_l})) + (N_{i_1}, \ldots, N_{i_l}),$$

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 and N_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

$$\mathbf{P}\{\bar{C}_{n,j,k} \neq C_{j,k}|X_1, \ldots, X_n\} = \Phi\left(-\sqrt{\sum_{r=1}^{l} g_{j,k}^2(X_{i_r})}\right) = \Phi\left(-\sqrt{\sum_{i=1}^{n} g_{j,k}^2(X_i)}\right).$$

Since $\Phi(-\sqrt{x})$ is convex, by Jensen-inequality

$$\mathbf{P}\{\bar{C}_{n,j,k} \neq C_{j,k}\} = \mathbf{E}\{\mathbf{P}\{\bar{C}_{n,j,k} \neq C_{j,k} | X_1, \ldots, X_n\}\} = \mathbf{E}\left\{\Phi\left(-\sqrt{\sum_{i=1}^{n} g_{j,k}^2(X_i)}\right)\right\}$$

$$\geq \Phi\left(-\sqrt{\mathbf{E}\{\sum_{i=1}^{n} g_{j,k}^2(X_i)\}}\right) = \Phi\left(-\sqrt{n\mathbf{E}\{g_{j,k}^2(X_1)\}}\right) = \Phi\left(-\sqrt{np_j^{2p+d}\int g^2(x)\,dx}\right) \quad (7)$$

independently of k, thus

$$\mathbf{E}R_n(C) \geq \sum_{j:np_j^{2p+d}\leq 1} \sum_{k=1}^{S_j} p_j^{2p+d}\Phi\left(-\sqrt{np_j^{2p+d}\int g^2(x)\,dx}\right)$$

$$\geq \Phi\left(-\sqrt{\int g^2(x)\,dx}\right) \sum_{j:np_j^{2p+d}\leq 1} S_j p_j^{2p+d} \geq K_1 \cdot \sum_{j:np_j^{2p+d}\leq 1} p_j^{2p+1}, \quad (8)$$

where

$$K_1 = \Phi\left(-\sqrt{\int g^2(x)\,dx}\right)\left(\frac{1}{2}\right)^{d-1} .$$

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

$$a_{n_t}^{1/2p} \leq 2^{-t}.$$

Define q_t such that

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

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

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

$$
\begin{aligned}
&\geq \sum_{t:a_{n_t}\leq a_n} b_{n_t}\left(\frac{2^{-t}}{\frac{2^{-t}}{a_{n_t}^{1/2p}}+1}\right)^{2p} \\
&= \sum_{t:n_t\geq n} b_{n_t}\left(\frac{a_{n_t}^{1/2p}}{1+2^t a_{n_t}^{1/2p}}\right)^{2p} \\
&\geq \sum_{t:n_t\geq n} \frac{b_{n_t}a_{n_t}}{2^{2p}}
\end{aligned}
$$

by $a_{n_t}^{1/2p}\leq 2^{-t}$, and specially for $n=n_s$ (8) implies

$$
\mathbf{E}R_{n_s}(C)\geq K_1\sum_{j:n_s p_j^{2p+d}\leq 1} p_j^{2p+1}\geq \frac{K_1}{2^{2p}}\sum_{t\geq s} b_{n_t}a_{n_t}\geq \frac{K_1}{2^{2p}}b_{n_s}a_{n_s}\ . \tag{9}
$$

We finish the proof in the spirit of Lemma 1 in Antos and Lugosi (1998). Using (9) one gets

$$
\begin{aligned}
\inf_{\{m_n\}}\sup_{c\in\mathcal{C}}\limsup_{n\to\infty}\frac{R_n(c)}{b_n a_n} &\geq \inf_{\{m_n\}}\sup_{c\in\mathcal{C}}\limsup_{s\to\infty}\frac{R_{n_s}(c)}{b_{n_s}a_{n_s}} \\
&\geq \frac{K_1}{2^{2p}}\inf_{\{m_n\}}\sup_{c\in\mathcal{C}}\limsup_{s\to\infty}\frac{R_{n_s}(c)}{\mathbf{E}R_{n_s}(C)} \\
&\geq \frac{K_1}{2^{2p}}\inf_{\{m_n\}}\mathbf{E}\left(\limsup_{s\to\infty}\frac{R_{n_s}(C)}{\mathbf{E}R_{n_s}(C)}\right).
\end{aligned}
$$

Because of (8) and the fact that for all $c\in\mathcal{C}$

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

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

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

This together with (6) implies the assertion. □

References

[1] Antos, A. and G. Lugosi (1998). Strong minimax lower bound for learning. *Machine Learning* 30, 31–56.

[2] Barron, A. R., L. Birgé and P. Massart (1995). Risk bounds for model selection via penalization. Technical Report No. 95.54, Université Paris Sud. To appear in *Probability Theory and Related Fields.*

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

[4] Devroye, L., L. Györfi, A. Krzyżak and G. Lugosi (1994). On the strong universal consistency of nearest neighbor regression function estimates. *Annals of Statistics* 22, 1371–1385.

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

[6] Devroye, L. and T. J. Wagner (1980). Distribution-free consistency results in nonparametric discrimination and regression function estimation. *Annals of Statistics* 8, 231–239.

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

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

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

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

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

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