An L_2-boosting algorithm for estimation of a regression function *

Adil Bagirov[1], Conny Clausen[2] and Michael Kohler[3]

[1] School of Information Technology and Mathematical Sciences, University of Ballarat, PO Box 663, Ballarat Victoria 3353 Australia, email: a.bagirov@ballarat.edu.au

[2] Department of Mathematics, Universität des Saarlandes, Postfach 151150, D-66041 Saarbrücken, Germany, email: clausen@math.uni-sb.de

[3] Department of Mathematics, Technische Universität Darmstadt, Schloßgartenstr. 7, D-64289 Darmstadt, Germany, email: kohler@mathematik.tu-darmstadt.de

September 29, 2008

Abstract

An L_2-boosting algorithm for estimation of a regression function from random design is presented, which consists of fitting repeatedly a function from a fixed nonlinear function space to the residuals of the data by least squares and by defining the estimate as a linear combination of the resulting least squares estimates. Splitting of the sample is used to decide after how many iterations of smoothing of the residuals the algorithm terminates. The rate of convergence of the algorithm is analyzed in case of an unbounded response variable. The method is used to fit a sum of maxima of minima of linear functions to a given data set, and is compared with other nonparametric regression estimates using simulated data.

Key words and phrases: regression, statistical learning, L_2-boosting, Greedy algorithm, rate of convergence.

*Running title: *L_2-boosting regression*

Please send correspondence and proofs to: Michael Kohler, Department of Mathematics, Technische Universität Darmstadt, Schlossgartenstr. 7, D-64289 Darmstadt, Germany, email: kohler@mathematik.tu-darmstadt.de, phone: +49-6151-16-6846, fax: +49-6151-16-6822.

1 Introduction

In regression analysis an $\mathbb{R}^d \times \mathbb{R}$-valued random vector (X,Y) with $\mathbf{E}Y^2 < \infty$ is considered and the dependency of Y on the value of X is of interest. More precisely, the goal is to find a function $f : \mathbb{R}^d \to \mathbb{R}$ such that $f(X)$ is a "good approximation" of Y. In the sequel we assume that the main aim of the analysis is minimization of the mean squared prediction error or L_2 risk

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

In this case the optimal function is the so-called regression function $m : \mathbb{R}^d \to \mathbb{R}$, $m(x) = \mathbf{E}\{Y|X = x\}$, i.e.,

$$\mathbf{E}\{|m(X) - Y|^2\} = \min_{f:\mathbb{R}^d \to \mathbb{R}} \mathbf{E}\{|f(X) - Y|^2\}, \tag{2}$$

because for an arbitrary (measurable) function $f : \mathbb{R}^d \to \mathbb{R}$ we have

$$\mathbf{E}\{|f(X) - Y|^2\} \quad = \quad \mathbf{E}\{|m(X) - Y|^2\} + \int |f(x) - m(x)|^2 \mathbf{P}_X(dx) \tag{3}$$

(cf., e.g., Section 1.1 in Györfi et al. (2002)). In addition, equation (3) implies that any function f is a good predictor in the sense that its L_2 risk is close to the optimal value, if and only if the so-called L_2 error

$$\int |f(x) - m(x)|^2 \mathbf{P}_X(dx) \tag{4}$$

is small. This motivates to measure the error caused by using a function f instead of the regression function by the L_2 error (4).

In applications, usually the distribution of (X,Y) (and hence also the regression function) is unknown. But often it is possible to observe a sample of the underlying distribution. This leads to the regression estimation problem. Here (X,Y), (X_1,Y_1), (X_2,Y_2), $\dots$ are independent and identically distributed random vectors. The set of data

$$\mathcal{D}_n = \{(X_1,Y_1), \dots, (X_n,Y_n)\}$$

is given, and the goal is to construct an estimate $m_n(\cdot) = m_n(\cdot, \mathcal{D}_n) : \mathbb{R}^d \to \mathbb{R}$ of the regression function such that the L_2 error

$$\int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx)$$

is small. For a detailed introduction to nonparametric regression we refer the reader to the monograph Györfi et al. (2002).

In this paper we are mainly interested in results which hold under very weak assumptions on the underlying distribution. In particular we do not assume that a density of the distribution of X exists or that the conditional distribution of Y given X is a normal distribution. Related results in this respect can be found, e.g., in Devroye (1981), Györfi and Walk (1997), Kohler (2008), Krzyżak, Linder and Lugosi (1996) or Walk (2001).

A closely related problem to nonparametric regression is pattern recognition, where Y takes on values only in a finite set (cf., e.g., Devroye, Györfi and Lugosi (1996)). One of the main achievements in pattern recognition in the last fifteen years was boosting (cf. Freund (1995) and Freund and Schapire (1997)), where the outputs of many "weak" classifiers are combined to produce a new powerful classification rule. Boosting can be considered as a way of fitting an additive expansion in a set of "elementary" basis functions (cf. Friedman, Hastie and Tibshirani (2000)). This view enables to extend the whole idea to regression by repeatedly fitting of functions of some fixed function space to residuals and by using the sum of the fitted functions as final estimate (cf. Friedman (2001)). Bühlmann (2006) showed that this so-called L_2-boosting is able to estimate very high-dimensional linear models well. Barron et al. (2008) analyzed the rate of convergence of corresponding Greedy algorithms, where iteratively functions of a fixed function space are fitted to the residuals of the previous estimate, and the estimates are defined by an linear combination of these functions. In Barron et al. (2008) this algorithm was used to fit a linear combination of perceptrons to the data, and under the assumption of a bounded first moment of the Fourier transform of the regression function and of boundedness of the response variable it was shown that these estimates are able to achieve (up to some logarithmic factors) the same dimension-free parametric rate of convergence as Barron (1994) showed for least squares neural networks.

In this paper we modify the general algorithm from Barron et al. (2008) by combining it with splitting of the sample in order to determine how often the residuals are smoothed. We analyze the modified general algorithm in the context of an unbounded response variable satisfying a Sub-Gaussian condition. We use it to fit a sum of maxima of minima of linear functions to the data. Since this function class contains in particular perceptrons, we get as a corollary the rate of convergence mentioned already above, but this time for unbounded response variables, too. We use an algorithm from Bagirov, Clausen and Kohler (2008) to compute our estimate, apply our new method to simulated data and compare it to other nonparametric regression estimates.

The outline of the paper is as follows: Section 2 contains the definition and our theoretical result on the general L_2-boosting algorithm. In Section 3 we apply it to estimate the regression function by a sum of maxima of minima of linear functions. This algorithm is applied to simulated data and compared to other nonparametric regression estimates in Section 4 . Finally, Section 5 contains the proofs.

2 A general L_2-boosting algorithm

Let $n_l, n_t \in \mathbb{N}$ be such that $n = n_l + n_t$, and let $\mathcal{F}_n$ be a (nonlinear) class of functions $f : \mathbb{R}^d \to \mathbb{R}$. Depending on a parameter $k_0 \in \mathbb{N}$ we define estimates

$$\tilde{m}_{n,k} \quad (k \in \{k_0, k_0+1, \dots, n\})$$

as follows: Set

$$\tilde{m}_{n,k_0} = \arg\min_{f \in \mathcal{F}_n} \frac{1}{n_l} \sum_{i=1}^{n_l} |Y_i - f(X_i)|^2 \tag{5}$$

and

$$\tilde{m}_{n,k+1} = \left(1 - \frac{2}{k+1}\right) \cdot \tilde{m}_{n,k} + f_{n_l,k} \tag{6}$$

where

$$f_{n_l,k} = \arg\min_{f \in \mathcal{F}_n} \frac{1}{n_l} \sum_{i=1}^{n_l} \left| Y_i - \left(1 - \frac{2}{k+1}\right) \cdot \tilde{m}_{n,k}(X_i) - f(X_i) \right|^2 . \tag{7}$$

Here we assume for simplicity that the above minima exist, however we do not require that they are unique. Next we truncate the estimate at heights $\pm\beta_n$, where $\beta_n \in \mathbb{R}_+$ is given and will later be chosen such that $\beta_n \to \infty$ $(n \to \infty)$. More precisely, we choose $k_0 \in \{1, \dots, n\}$ and set

$$m_{n,k}(x) = T_{\beta_n} \tilde{m}_{n,k}(x) \quad (x \in \mathbb{R}^d), \tag{8}$$

where $T_\beta(z) = \max\{-\beta, \min\{\beta, z\}\}$ for $z \in \mathbb{R}$. Finally we use splitting of the sample to select the parameter k of the estimate. To do this, we set

$$m_n(x) = m_{n,k^*}(x) \quad (x \in \mathbb{R}^d) \tag{9}$$

where

$$k^* = \arg \min_{k \in \{k_0, k_0+1, \ldots, n\}} \frac{1}{n_t} \sum_{i=n_l+1}^{n} |Y_i - m_{n,k}(X_i)|^2. \tag{10}$$

In order to be able to formulate our main theoretical result we need the notion of covering numbers.

Definition 1 *Let $x_1, ..., x_n \in \mathbb{R}^d$ and set $x_1^n = (x_1, ..., x_n)$. Let $\mathcal{F}$ be a set of functions $f : \mathbb{R}^d \to \mathbb{R}$. An L_p-ϵ-cover of $\mathcal{F}$ on x_1^n is a finite set of functions $f_1, ..., f_k : \mathbb{R}^d \to \mathbb{R}$ with the property*

$$\min_{1 \leq j \leq k} \left(\frac{1}{n} \sum_{i=1}^{n} |f(x_i) - f_j(x_i)|^p \right)^{1/p} < \epsilon \qquad \text{for all } f \in \mathcal{F}. \tag{11}$$

The L_p-ϵ-covering number $\mathcal{N}_p(\epsilon, \mathcal{F}, x_1^n)$ of $\mathcal{F}$ on x_1^n is the minimal size of an L_p-ϵ-cover of $\mathcal{F}$ on x_1^n. In case that there exist no finite L_p-ϵ-cover of $\mathcal{F}$ the L_p-ϵ-covering number of $\mathcal{F}$ on x_1^n is defined by $\mathcal{N}_p(\epsilon, \mathcal{F}, x_1^n) = \infty$.

For a given class $\mathcal{F}$ of functions $f : \mathbb{R}^d \to \mathbb{R}$, and fixed $N \in \mathbb{N}$, we define $\mathcal{H}_N = \mathcal{H}_N^{\mathcal{F}}$ as the class of functions $h : \mathbb{R}^d \to \mathbb{R}$ with $h(x) = \alpha_1^h g_1(x) + ... + \alpha_N^h g_N(x)$, where $\alpha_i^h \geq 0$ and $g_i \in \mathcal{F}$ $(i \in \{1, ..., N\})$ are such, that the two conditions

$$\left(\frac{2}{l} \sum_{i=1}^{N} \alpha_i^h \right) \cdot g_j \in \mathcal{F} \qquad \text{for all} \quad j \in \{1, \ldots, N\}, l \in \{1, \ldots, k\}, \tag{12}$$

$$||g_j||_\infty = \sup_{x \in \mathbb{R}^d} |g_j(x)| \leq 1, \qquad \text{for all} \quad j \in \{1, \ldots, N\}, \text{ and } x \in \mathbb{R}^d \tag{13}$$

are satisfied.

Our main theoretical result is the following theorem.

Theorem 1 *Let $\mathcal{F}_n$ be a class of functions $f : \mathbb{R}^d \to \mathbb{R}$ with the property $\alpha \cdot f \in \mathcal{F}_n$ for all $f \in \mathcal{F}_n$ and all $0 \leq \alpha \leq 1$. Let $\mathcal{N}_1(\epsilon, \mathcal{F}_n)$ be an upper bound on the L_1-ϵ-covering number of $\mathcal{F}_n$ on any finite set of points, i.e., assume*

$$\mathcal{N}_1(\epsilon, \mathcal{F}_n, x_1^n) \leq \mathcal{N}_1(\epsilon, \mathcal{F}_n) \quad \text{for all } x_1^n \in \mathbb{R}^{d \cdot n}.$$

Define the estimate m_n by (5) - (10) with $\beta_n = c_1 \cdot \log(n)$. Furthermore assume that the distribution of (X, Y) satisfies

$$\mathbf{E} \left(\exp \left(c_2 \cdot |Y|^2 \right) \right) < \infty \tag{14}$$

for some constant $c_2 > 0$ and that the regression function m is bounded in absolute value by some constant. Then

$$\begin{aligned}
&\mathbf{E} \int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx) \\
&\leq \min_{k \in \{k_0, k_0+1, \dots, n\}} \Bigg(c_3 \left(\frac{k \cdot \log(n)^2 \cdot \log \mathcal{N}_1\left(\frac{1}{80\beta_n \cdot k \cdot n}, \mathcal{F}_n\right)}{n_l} \right) \\
&\quad + \inf_{N \in \mathbb{N}} \inf_{h \in \mathcal{H}_N^{\mathcal{F}_n}} \left(16 \cdot k_0 \cdot \frac{(\alpha_1^h + \ldots + \alpha_N^h)^2}{k} + 4 \int |h(x) - m(x)|^2 \mathbf{P}_X(dx) \right) \Bigg) \\
&\quad + c_4 \frac{\log(n)^3}{n_t}
\end{aligned}$$

holds for sufficiently large constants $c_3, c_4 > 0$, which do not depend on n, β_n, k or k_0.

3 Fitting of a sum of maxima of minima of linear functions to the data

In this section we apply our general algorithm to classes of functions consisting of maxima of minima of linear functions as introduced in Bagirov, Clausen and Kohler (2006), i.e. we apply it to a truncated version of

$$\begin{aligned}
\mathcal{F}_{r_1, r_2} = \Bigg\{ f : \mathbb{R}^d \to \mathbb{R} \, : \, f(x) = \max_{k=1,\dots,r_1} \min_{l=1,\dots,r_2} (a_{k,l} \cdot x + b_{k,l}) \quad (x \in \mathbb{R}^d) \\
\text{for some } a_{k,l} \in \mathbb{R}^d, b_{k,l} \in \mathbb{R} \Bigg\}
\end{aligned} \tag{15}$$

where

$$a_{k,l} \cdot x = a_{k,l}^{(1)} \cdot x^{(1)} + \ldots + a_{k,l}^{(d)} \cdot x^{(d)}$$

denotes the scalar product between $a_{k,l} = (a_{k,l}^{(1)}, \dots, a_{k,l}^{(d)})^T$ and $x = (x^{(1)}, \dots, x^{(d)})^T$.

This class of functions consists of continuous piecewise linear functions. For $r_1, r_2 \geq 2$ it contains in particular perceptrons of the form

$$f(x) = \sigma\left(a \cdot x + b\right) \quad (x \in \mathbb{R}^d)$$

for a suitable chosen squashing function σ (i.e., for a suitable chosen monotone increasing function $\sigma : \mathbb{R} \to \mathbb{R}$ satisfying $\sigma(x) \to 0$ $(x \to -\infty)$ and $\sigma(x) \to 1$ $(x \to \infty)$). This is obvious, if we choose for σ the so-called ramp squasher

$$\sigma(z) = \max\{0, \min\{z, 1\}\} \quad (z \in \mathbb{R}).$$

In the sequel we will choose as function class for the general algorithm of Section 2

$$\mathcal{F}_n = T_{\beta_n}\mathcal{F}_{l,l} = \{T_{\beta_n} f \quad : \quad f \in \mathcal{F}_{l,l}\}$$

for some $l \geq 2$.

It is well-known that in order to derive non-trivial rate of convergence results we have to make some smoothness assumptions on the regression function (cf., e.g., Theorem 7.2 and Problem 7.2 in Devroye, Györfi and Lugosi (1996) and Section 3 in Devroye and Wagner (1980)). In the sequel we will impose such smoothness conditions implicitely on the regression function by imposing conditions on its Fourier transform. More precisely, we will consider functions $f \in L_1(\mathbb{R}^d)$, which satisfy

$$f(x) = f(0) + \frac{1}{(2\pi)^{d/2}} \int \left(e^{i(\omega \cdot x)} - 1\right) \hat{F}(\omega) d\omega, \tag{16}$$

where $\hat{F}$ is the Fourier transform of f, that is

$$\hat{F}(\omega) = \frac{1}{(2\pi)^{d/2}} \int e^{-i(\omega \cdot x)} f(x) dx \quad (\omega \in \mathbb{R}^d),$$

and we assume

$$\int ||\omega|| \cdot |\hat{F}(\omega)| d\omega \leq C \tag{17}$$

for some $C \in \mathbb{R}_+$. We denote the class of functions $f : \mathbb{R}^d \to \mathbb{R}$, which satisfy (16) and (17) by $\mathcal{F}_C$.

Corollary 1 *Let $\beta_n = c_1 \cdot \log(n)$ and assume that the distribution of (X, Y) satisfies (14) for some constant $c_2 > 0$, $X \in [-a, a]^d$ a.s. for some $a \in \mathbb{R}_+$ and that the regression function is bounded in absolute value by some constant less than or equal to β_n and that it satisfies $m \in \mathcal{F}_C$ for some $0 < C < \infty$. Let the estimate m_n be defined by (5) - (10), with $\mathcal{F} = T_{\beta_n}\mathcal{F}_{l,l}$ for some $l \geq 2$, and with $n_l = \lceil \frac{n}{2} \rceil$. Then we have for $\beta_n \geq 6 \cdot \sqrt{d} \cdot a \cdot C$*

$$\mathbf{E} \int |m_n(x) - m(x)| \mathbf{P}_X(dx) \leq c_5 \cdot C^2 \left(\frac{\log(n)^3}{n}\right)^{1/2}$$

for a sufficiently large constant $c_5 > 0$, that does not depend on n or C.

4 Application to simulated data

In this section we want to compare our new L_2–boosting estimate with other nonparametric regression estimates. To do this, we use results from a simulation study conducted in Bagirov, Clausen and Kohler (2008). There data was generated according to

$$Y = m(X) + \sigma \cdot \epsilon,$$

where ϵ is standard normally distributed and independent of X and $\sigma \in \{0, 0.5, 1\}$, and where X is uniformly distributed on $[-2,2]^d$ with $d \in \{1, 2, 10\}$, and where $\sigma \in \{0, 0.2, 1\}$. As regression functions the following 11 function have been considered:

- $m_1(x) = 2 * \max(1, \min(3 + 2 * x, 3 - 8 * x))$,
- $m_2(x) = \begin{cases} 1 & , x \leq 0, \\ 3 & , \text{ else}, \end{cases}$
- $m_3(x) = \begin{cases} 10 * \sqrt{-x} * \sin(8 * \pi * x) & , -0.25 \leq x < 0, \\ 0 & , \text{ else}, \end{cases}$
- $m_4(x) = 3 * \sin(\pi * x/2)$,
- $m_5(x_1, x_2) = x_1 * \sin(x_1^2) - x_2 * \sin(x_2^2)$,
- $m_6(x_1, x_2) = \frac{4}{1+4*x_1^2+4*x_2^2}$,
- $m_7(x_1, x_2) = 6 - 2 * \min(3, 4 * x_1^2 + 4 * |x_2|)$,
- $m_8(x_1, ..., x_{10}) = \sum_{j=1}^{10} (-1)^{j-1} * x_j * \sin(x_j^2)$,
- $m_9(x_1, ..., x_{10}) = m_7(x_1, x_2)$,
- $m_{10}(x_1, ..., x_{10}) = m_6(x_1 + ... + x_5, x_6 + ... + x_{10})$,
- $m_{11}(x_1, ..., x_{10}) = m_2(x_1 + ... + x_{10})$.

For these 11 different regression functions and each value $\sigma \in \{0, 0.5, 1\}$ data sets of size $n \in \{500, 5000\}$ have been generated, so altogether $3 \cdot 11 = 33$ different distributions have been considered, and for each of these distributions the estimates have been compared for 2 different sample sizes. The *maxmin*–estimate proposed in Bagirov, Clausen and

Kohler (2008) has been compared for $d = 1$ with with kernel estimates (with Gaussian kernel) (see, e.g., Chapter 5 in Györfi et al. (2002)), local linear kernel estimates (see, e.g., Section 5.4 in Györfi et al. (2002)), smoothing splines (see, e.g., Chapter 20 in Györfi et al. (2002)), neural networks and regression trees (as implemented in the freely available statistics software R). Since for $d > 1$ not all of these estimates are easily applicable in R, for $d > 1$ the *maxmin*–estimate has been compared only with neural networks and regression trees.

In order to compute the L_2 errors of the estimates, Monte Carlo integration was used, i.e.,

$$\int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx) = \mathbf{E}\{|m_n(U) - m(U)|^2 | \mathcal{D}_n\}$$

was approximated by

$$\frac{1}{N} \sum_{j=1}^{N} |m_n(\tilde{U}_j) - m(\tilde{U}_j)|^2,$$

where the random variables $\tilde{U}_1, \tilde{U}_2, \ldots$ are i.i.d. with distribution $\mathbf{P}_U = \mathbf{P}_X$ and independent of $\mathcal{D}_n$, and where $N = 3000$. Since this error is a random variable itself, the experiment was 25 times repeated with independent realizations of the sample, and the mean and the standard deviation of the Monte Carlo estimates of the L_2 error was reported.

In the sequel we make the same simulations with our newly propsed L_2–boosting estimate. Here we set $l = 4$ for $d \in \{1, 2\}$ and $l = 5$ for $d = 10$, $k_0 = 1000$, repeat 7 boosting steps and use splitting of the sample with $n_l = n_t = n/2$ to choose one of these seven estimates as final estimate. In the sequel we present the mean and the standard deviation of the Monte Carlo estimates of the L_2 error of our estimates. In order to save space, we do not repeat the error values already published in Bagirov, Clausen and Kohler (2008), instead we just summarize them by reporting whether the error of the L_2–boosting estimate is better, worse or the same as the error of the *maxmin*–estimate (coded by $+$, $-$ and $=$, resp.), and by reporting which position the error of the L_2–boosting estimate achieves, if we order the mean error values of all estimates (except the *maxmin*–estimate) increasingly (which gives us a number between 1 and 6 in case of $d = 1$, and a number between 1 and 3 in case of $d > 1$).

Table 1 summarizes the results for the four univariate regression functions $m_1, \ldots, m_4$, Table 2 summarizes the results for the three bivariate regression functions m_5, m_6 and m_7

	l	500			5000		
	σ	0	0.5	1	0	0.5	1
m_1	Error Std. deviation Comparison	0.0000 (0.0000) = / 1	0.0078 (0.0043) + / 1	0.0290 (0.0166) + / 1	0.0000 (0.0000) = / 1	0.0007 (0.0004) = / 1	0.0025 (0.0014) + / 1
m_2	Error Std. Comparison	0.0041 (0.0039) + / 1	0.0157 (0.0128) − / 1	0.0389 (0.0157) + / 2	0.0007 (0.0010) = / 1	0.0012 (0.0008) + / 1	0.0040 (0.0020) − / 1
m_3	Error Std. Comparison	0.0155 (0.0343) + / 3	0.0272 (0.0129) − / 1	00.1070 (0.0459) + / 2	0.0015 (0.0012) − / 3	0.0041 (0.0022) − / 2	0.0072 (0.0031) − / 1
m_4	Error Std. Comparison	0.0006 (0.0003) + / 5	0.0184 (0.0054) + / 4	0.0508 (0.0151) + / 3	0.0006 (0.0004) + / 6	0.0030 (0.0007) + / 5	0.0088 (0.0027) + / 6

Table 1: Simulation results and comparison with six other nonparametric regression estimates for four univariate regression functions.

and Table 3 summarizes the results for the four regression functions $m_8, \ldots, m_{11}$ where $d = 10$. Considering the results in Table 1,2 and 3 we can firstly see, that the error of our L_2–boosting estimate was 47-times less than but only 15–times bigger than the error of the original *maxmin*–estimate. Taking into account that the newly proposed estimates requires on average three to four times less time for computation of the estimate, we can say that L_2–boosting clearly leads to an improvement of the *maxmin*–estimate.

Secondly, by looking at Table 3 we can see that the L_2–boosting estimate is especially suited for high-dimensional data sets and large sample size in comparison with other nonparametric regression estimates.

	l	500			5000		
	σ	0	0.5	1	0	0.5	1
m_5	Error Std. Comparison	0.0322 (0.0075) + / 2	0.1036 (0.0226) + / 2	0.2076 (0.0426) + / 1	0.0089 (0.0024) + / 2	0.0212 (0.0037) + / 2	0.0445 (0.0118) + / 2
m_6	Error Std. Comparison	0.0143 (0.0045) − / 2	0.0645 (0.0143) − / 2	0.1486 (0.0330) + / 1	0.0064 (0.0013) + / 2	0.0123 (0.0015) + / 2	0.0311 (0.0039) + / 1
m_7	Error Std. Comparison	0.0317 (0.0150) + / 2	0.1192 (0.0310) − / 1	0.1952 (0.0469) − / 1	0.0049 (0.0018) + / 1	0.0234 (0.0123) − / 1	0.0392 (0.0133) − / 1

Table 2: Simulation results and comparison with three other nonparametric regression estimates for three bivariate regression functions.

5 Proofs

5.1 A deterministic lemma

Let $\mathcal{F}$ be a class of functions $f : \mathbb{R}^d \to \mathbb{R}$ and let $(x_1, y_1), \ldots, (x_n, y_n) \in \mathbb{R}^d \times \mathbb{R}$, $k_0 \in \mathbb{N}$ and define $m_{n,k}$ $(k \geq k_0)$ recursively by

$$m_{n,k_0} = \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} |y_i - f(x_i)|^2 \tag{18}$$

and

$$m_{n,k+1} = \left(1 - \frac{2}{k+1}\right) \cdot m_{n,k} + f_{n,k} \tag{19}$$

where

$$f_{n,k} = \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \left| y_i - \left(1 - \frac{2}{k+1}\right) \cdot m_{n_1,k}(x_i) - f(x_i) \right|^2. \tag{20}$$

Lemma 1 *Let $m_{n,k}$ be defined by (18) - (20). Then for any $k \geq k_0$, $N \in \mathbb{N}$, $g_1, \ldots, g_N \in \mathcal{F}$ and $\alpha_1, \ldots, \alpha_N > 0$, such that*

$$\left(\frac{2}{l} \sum_{i=1}^{N} \alpha_i\right) \cdot g_j \in \mathcal{F}, \qquad \textit{for all} \quad j \in \{1, \ldots, N\}, l \in \{1, \ldots, k\}, \tag{21}$$

$$\textit{and} \quad ||g_j||_\infty = \sup_{x \in \mathbb{R}^d} |g_j(x)| \leq 1, \qquad \textit{for all} \quad j \in \{1, \ldots, N\}, \tag{22}$$

	l	500			5000		
	σ	0	0.5	1	0	0.5	1
m_8	Error Std. Comparison	4.4171 (0.1606) + / 1	4.4991 (0.1774) − / 1	4.5242 (0.1687) + / 1	1.1063 (0.1292) + / 1	1.1496 (0.1190) + / 1	1.1246 (0.1622) + / 1
m_9	Error Std. Comparison	0.5656 (0.1602) + / 3	0.6867 (0.1287) + / 3	0.8648 (0.0629) + / 3	0.0348 (0.0142) − / 2	0.0461 (0.0073) + / 2	0.1247 (0.0327) + / 3
m_{10}	Error Std. Comparison	0.1529 (0.0309) + / 1	0.2078 (0.0271) + / 2	0.2441 (0.0614) + / 3	0.0185 (0.0029) + / 1	0.0410 (0.0053) + / 1	0.0962 (0.0173) + / 1
m_{11}	Error Std. Comparison	0.0698 (0.0252) + / 1	0.1983 (0.0662) + / 1	0.4128 (0.0425) + / 1	0.0169 (0.0034) − / 1	0.0251 (0.0052) + / 1	0.0514 (0.0174) + / 1

Table 3: Simulation results and comparison with three other nonparametric regression estimates for regression functions where $d = 10$.

we have:

$$\frac{1}{n}\sum_{i=1}^{n}|y_i - m_{n,k}(x_i)|^2 \leq \frac{1}{n}\sum_{i=1}^{n}|y_i - (\alpha_1 g_1 + \ldots + \alpha_N g_N)(x_i)|^2 + 4 \cdot k_0 \cdot \frac{\left(\sum_{i=1}^{N}\alpha_i\right)^2}{k}.$$

The proof of the above lemma is a modification of the proof of Theorem 2.4 in Barron et al. (2008). For the sake of completeness we repeat it below.

Proof of Lemma 1. *In the first step of the proof* we show

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}|y_i - m_{n,k}(x_i)|^2 - \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \sum_{l=1}^{N}\alpha_l g_l(x_i)\right)^2 \\
&\leq \left(1 - \frac{2}{k}\right)\cdot\left(\frac{1}{n}\sum_{i=1}^{n}(y_i - m_{n,k-1}(x_i))^2 - \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \sum_{l=1}^{N}\alpha_l g_l(x_i)\right)^2\right) \\
&\quad + \frac{4}{k^2}\cdot\left((\sum_{j=1}^{N}\alpha_j)^2 - \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{l=1}^{N}\alpha_l g_l(x_i)\right)^2\right).
\end{aligned}$$

To do this, let $j \in \{1,\ldots,N\}$ and set $\beta_k = \frac{2}{k}\cdot\sum_{i=1}^{N}\alpha_i$. Because of $\beta_k \cdot g_j \in \mathcal{F}$ we have by definition of the estimate

$$\frac{1}{n}\sum_{i=1}^{n}|y_i - m_{n,k}(x_i)|^2$$

$$
\begin{aligned}
&\leq \frac{1}{n}\sum_{i=1}^{n}\left|\left(1-\frac{2}{k}\right)\cdot(y_i-m_{n,k-1}(x_i))+\frac{2}{k}\cdot y_i-\beta_k\cdot g_j(x_i)\right|^2\\
&= \left(1-\frac{2}{k}\right)^2\cdot\frac{1}{n}\sum_{i=1}^{n}|y_i-m_{n,k-1}(x_i)|^2\\
&\quad +2\left(1-\frac{2}{k}\right)\cdot\frac{1}{n}\sum_{i=1}^{n}(y_i-m_{n,k-1}(x_i))\cdot\left(\frac{2}{k}\cdot y_i-\beta_k\cdot g_j(x_i)\right)\\
&\quad +\frac{1}{n}\sum_{i=1}^{n}\left(\frac{2}{k}\left(y_i-\sum_{l=1}^{N}\alpha_l g_l(x_i)\right)+\frac{2}{k}\cdot\sum_{l=1}^{N}\alpha_l g_l(x_i)-\beta_k\cdot g_j(x_i)\right)^2\\
&\leq \left(1-\frac{2}{k}\right)^2\cdot\frac{1}{n}\sum_{i=1}^{n}|y_i-m_{n,k-1}(x_i)|^2\\
&\quad +2\left(1-\frac{2}{k}\right)\cdot\frac{1}{n}\sum_{i=1}^{n}(y_i-m_{n,k-1}(x_i))\cdot\left(\frac{2}{k}\cdot y_i-\beta_k\cdot g_j(x_i)\right)\\
&\quad +\left(\frac{2}{k}\right)^2\frac{1}{n}\sum_{i=1}^{n}\left(y_i-\sum_{l=1}^{N}\alpha_l g_l(x_i)\right)^2\\
&\quad +\frac{4}{k}\frac{1}{n}\sum_{i=1}^{n}\left(y_i-\sum_{l=1}^{N}\alpha_l g_l(x_i)\right)\cdot\left(\frac{2}{k}\cdot\sum_{l=1}^{N}\alpha_l g_l(x_i)-\beta_k\cdot g_j(x_i)\right)\\
&\quad +\left(\frac{2}{k}\right)^2\frac{1}{n}\sum_{i=1}^{n}\left(\sum_{l=1}^{N}\alpha_l g_l(x_i)\right)^2-2\beta_k\cdot\frac{2}{k}\cdot\frac{1}{n}\sum_{i=1}^{n}\left(\sum_{l=1}^{N}\alpha_l g_l(x_i)\right)\cdot g_j(x_i)+\beta_k^2\\
&=: L_j,
\end{aligned}
$$

where we have used

$$
\frac{1}{n}\sum_{i=1}^{n}\beta_k^2 g_j^2(x_i)\leq\beta_k^2\|g_j\|_\infty^2\leq\beta_k^2.
$$

Since $\alpha_j\geq 0$ and $\sum_{j=1}^{N}(2/k)\cdot\alpha_j=\beta_k$ we can conclude

$$
\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}|y_i-m_{n,k}(x_i)|^2\\
&\leq\sum_{j=1}^{N}\frac{2\cdot\alpha_j}{k\cdot\beta_k}\cdot L_j\\
&=\left(1-\frac{2}{k}\right)^2\cdot\frac{1}{n}\sum_{i=1}^{n}|y_i-m_{n,k-1}(x_i)|^2\\
&\quad +2\left(1-\frac{2}{k}\right)\cdot\frac{1}{n}\sum_{i=1}^{n}(y_i-m_{n,k-1}(x_i))\cdot\left(\frac{2}{k}\cdot y_i-\frac{2}{k}\sum_{j=1}^{N}\alpha_j g_j(x_i)\right)\\
&\quad +\left(\frac{2}{k}\right)^2\frac{1}{n}\sum_{i=1}^{n}\left(y_i-\sum_{l=1}^{N}\alpha_l g_l(x_i)\right)^2-\left(\frac{2}{k}\right)^2\frac{1}{n}\sum_{i=1}^{n}\left(\sum_{l=1}^{N}\alpha_l g_l(x_i)\right)^2+\beta_k^2
\end{aligned}
$$

$$
\begin{aligned}
&= \left(1-\frac{2}{k}\right)^2 \frac{1}{n}\sum_{i=1}^n (y_i - m_{n,k-1}(x_i))^2 + \left(\frac{2}{k}\right)^2 \frac{1}{n}\sum_{i=1}^n \left(y_i - \sum_{l=1}^N \alpha_l g_l(x_i)\right)^2 \\
&\quad + \left(1-\frac{2}{k}\right)\cdot \frac{2}{k}\cdot \frac{1}{n}\sum_{i=1}^n 2\cdot (y_i - m_{n,k-1}(x_i)) \left(y_i - \sum_{l=1}^N \alpha_l g_l(x_i)\right) \\
&\quad - \left(\frac{2}{k}\right)^2 \frac{1}{n}\sum_{i=1}^n \left(\sum_{l=1}^N \alpha_l g_l(x_i)\right)^2 + \beta_k^2.
\end{aligned}
$$

Using $2 \cdot a \cdot b \leq a^2 + b^2$ we get

$$
\begin{aligned}
&\frac{1}{n}\sum_{i=1}^n |y_i - m_{n,k}(x_i)|^2 \\
&\leq \left(1-\frac{2}{k}\right)\cdot \frac{1}{n}\sum_{i=1}^n (y_i - m_{n,k-1}(x_i))^2 + \frac{2}{k}\cdot \frac{1}{n}\sum_{i=1}^n \left(y_i - \sum_{l=1}^N \alpha_l g_l(x_i)\right)^2 \\
&\quad + \beta_k^2 - \left(\frac{2}{k}\right)^2 \frac{1}{n}\sum_{i=1}^n \left(\sum_{l=1}^N \alpha_l g_l(x_i)\right)^2,
\end{aligned}
$$

which implies the assertion of the first step.

In the second step of the proof we show

$$
\frac{1}{n}\sum_{i=1}^n (y_i - m_{n,k_0}(x_i))^2 - \frac{1}{n}\sum_{i=1}^n \left(y_i - \sum_{l=1}^N \alpha_l g_l(x_i)\right)^2 \leq 4\left(\sum_{l=1}^N \alpha_l\right)^2.
$$

To do this, let $j \in \{1, \ldots, N\}$ and set $\gamma_{k_0} = \sum_{j=1}^N \alpha_j$. Then

$$
\frac{1}{n}\sum_{i=1}^n (y_i - m_{n,k_0}(x_i))^2 \leq \frac{1}{n}\sum_{i=1}^n \left(y_i - \sum_{l=1}^N \alpha_l g_l(x_i) + \sum_{l=1}^N \alpha_l g_l(x_i) - \gamma_{k_0}\cdot g_j(x_i)\right)^2,
$$

and arguing as above we get

$$
\begin{aligned}
&\frac{1}{n}\sum_{i=1}^n (y_i - m_{n,k_0}(x_i))^2 \leq \sum_{j=1}^N \frac{\alpha_j}{\gamma_{k_0}}\cdot \frac{1}{n}\sum_{i=1}^n \left(y_i - \sum_{l=1}^N \alpha_l g_l(x_i) + \sum_{l=1}^N \alpha_l g_l(x_i) - \gamma_{k_0}\cdot g_j(x_i)\right)^2 \\
&= \frac{1}{n}\sum_{i=1}^n (y_i - \sum_{l=1}^N \alpha_l g_l(x_i))^2 + \sum_{j=1}^N \frac{\alpha_j}{\gamma_{k_0}}\cdot \frac{1}{n}\sum_{i=1}^n \left(\sum_{l=1}^N \alpha_l g_l(x_i) - \gamma_{k_0}\cdot g_j(x_i)\right)^2 \\
&\leq \frac{1}{n}\sum_{i=1}^n (y_i - \sum_{l=1}^N \alpha_l g_l(x_i))^2 - \frac{1}{n}\sum_{i=1}^n \left(\sum_{l=1}^N \alpha_l g_l(x_i)\right)^2 + \gamma_{k_0}^2,
\end{aligned}
$$

from which we conclude the assertion of the second step.

In the third step of the proof we finish the proof. To do this, we observe that by the results of the previous steps we know already that

$$a_k \;:=\; \frac{1}{n}\sum_{i=1}^{n}|y_i - m_{n,k}(x_i)|^2 - \frac{1}{n}\sum_{i=1}^{n}(y_i - \sum_{l=1}^{N}\alpha_l g_l(x_i))^2$$

satisfies

$$a_{k_0} \le \frac{4M}{k_0} \quad \text{and} \quad a_k \le \left(1 - \frac{2}{k}\right) a_{k-1} + \frac{4}{k^2}M$$

where M is defined as $M := k_0 \cdot \left(\sum_{j=1}^{N} \alpha_j\right)^2$.

But from this we get the assertion, since $a_k \le 4M/k$ implies

$$a_{k+1} \le \left(1 - \frac{2}{k+1}\right) \cdot \frac{4M}{k} + \frac{4}{(k+1)^2}M \le \frac{4M}{k+1},$$

where the last inequality follows from

$$\left(1 - \frac{2}{k+1}\right) \cdot \frac{1}{k} + \frac{1}{(k+1)^2} = \frac{k^2+k-1}{k^2+k} \cdot \frac{1}{k+1} \le \frac{1}{k+1}.$$

□

5.2 Splitting of the Sample for Unbounded Y

The following lemma is an extension of Theorem 7.1 in Györfi et al. (2002) to unbounded data. It is about bounding the L_2 error of estimates, which are defined by splitting of the sample. Let $n = n_l + n_t$, let $\mathcal{Q}_n$ be a finite set of parameters and assume that for each parameter $h \in \mathcal{Q}_n$ an estimate

$$m_{n_l}^{(h)}(\cdot) = m_{n_l}^{(h)}(\cdot, \mathcal{D}_{n_l})$$

is given, which depends only on the training data $\mathcal{D}_{n_l} = \{(X_1, Y_1), \ldots, (X_{n_l}, Y_{n_l})\}$. Then we define

$$m_n(x) = m_{n_l}^{(H)}(x) \qquad \text{for all } x \in \mathbb{R}^d, \tag{23}$$

where $H \in \mathcal{Q}_n$ is chosen such that

$$\frac{1}{n_t}\sum_{i=n_l+1}^{n}|m_{n_l}^{(H)}(X_i) - Y_i|^2 \;=\; \min_{h \in \mathcal{Q}_n}\frac{1}{n_t}\sum_{i=n_l+1}^{n}|m_{n_l}^{(h)}(X_i) - Y_i|^2. \tag{24}$$

Lemma 2 *Let $\beta_n = c_1 \cdot \log(n)$ for some constant $c_1 > 0$ and assume that the estimates $m_{n_l}^{(h)}$ are bounded in absolute value by β_n for $h \in \mathcal{Q}_n$. Assume furthermore that the distribution of (X,Y) satisfies the Sub-Gaussian condition (14) for some constant $c_2 > 0$, and that the regression function fulfils $||m||_\infty < L$ for some $L \in \mathbb{R}_+$, with $L \leq \beta_n$. Then, for every estimate m_n defined by (23) and (24) and any $\delta > 0$,*

$$
\begin{aligned}
&\mathbf{E} \int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx) \\
&\leq (1+\delta) \min_{h \in \mathcal{Q}} \mathbf{E} \int |m_{n_l}^{(h)}(x) - m(x)|^2 \mathbf{P}_X(dx) + c_6 \cdot \beta_n^2 \cdot \frac{1 + \log|\mathcal{Q}_n|}{n_t} + c_7 \frac{\log(n)}{n}
\end{aligned}
$$

holds, with $c_6 = 16/\delta + 35 + 19\delta$ and a sufficiently large constant $c_7 > 0$.

Proof. We use the following error decomposition

$$
\begin{aligned}
&\mathbf{E}\left(\int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx) \bigg| \mathcal{D}_{n_l} \right) \\
&\quad = \mathbf{E}\left(\int |m_{n_l}^{(H)}(x) - m(x)|^2 \mathbf{P}_X(dx) \bigg| \mathcal{D}_{n_l} \right) \\
&\quad = \Bigg[\mathbf{E}\left(|m_{n_l}^{(H)}(X) - Y|^2 \Big| \mathcal{D}_{n_l} \right) - \mathbf{E}\left(|m(X) - Y|^2 \right) \\
&\qquad\quad - \mathbf{E}\left(|m_{n_l}^{(H)}(X) - T_{\beta_n} Y|^2 \Big| \mathcal{D}_{n_l} \right) - \mathbf{E}\left(|m_{\beta_n}(X) - T_{\beta_{n_t}} Y|^2 \right) \Bigg] \\
&\qquad + \Bigg[\mathbf{E}\left(|m_{n_l}^{(H)}(X) - T_{\beta_n} Y|^2 \Big| \mathcal{D}_{n_l} \right) - \mathbf{E}\left(|m_{\beta_n}(X) - T_{\beta_{n_t}} Y|^2 \right) \\
&\qquad\quad - (1+\delta) \cdot \frac{1}{n_t} \sum_{i=n_l+1}^{n} \left(|m_{n_l}^{(H)}(X_i) - T_{\beta_n} Y_i|^2 - |m_{\beta_n}(X_i) - T_{\beta_n} Y_i|^2 \right) \Bigg] \\
&\qquad + \Bigg[(1+\delta) \cdot \frac{1}{n_t} \sum_{i=n_l+1}^{n} \left(|m_{n_l}^{(H)}(X_i) - T_{\beta_n} Y_i|^2 - |m_{\beta_n}(X_i) - T_{\beta_n} Y_i|^2 \right) \\
&\qquad\quad - \left((1+\delta) \cdot \frac{1}{n_t} \sum_{i=n_l+1}^{n} \left(|m_{n_l}^{(H)}(X_i) - Y_i|^2 - |m(X_i) - Y_i|^2 \right) \right) \Bigg] \\
&\qquad + \left[(1+\delta) \cdot \frac{1}{n_t} \sum_{i=n_l+1}^{n} \left(|m_{n_l}^{(H)}(X_i) - Y_i|^2 - |m(X_i) - Y_i|^2 \right) \right] = \sum_{i=1}^{4} T_{i,n},
\end{aligned}
$$

where $T_{\beta_n} Y$ denotes the truncated version of Y and $m_{\beta_n}(x) = \mathbf{E}\{T_{\beta_n} Y | X = x\}$. Due to equality (24) we can bound the last term $T_{4,n}$ by

$$
(1+\delta) \left(\frac{1}{n_t} \sum_{i=n_l+1}^{n} \left(|m_{n_l}^{(h)}(X_i) - Y_i|^2 - |m(X_i) - Y_i|^2 \right) \right),
$$

for every $h \in \mathcal{Q}_n$, and this entails for its conditional expectation

$$
\begin{aligned}
\mathbf{E}(T_{4,n}|\mathcal{D}_{n_l}) &\leq (1+\delta)\min_{h\in\mathcal{Q}_n}\left(\mathbf{E}\left(|m_{n_l}^{(h)}(X)-Y|^2\middle|\mathcal{D}_{n_l}\right)-\mathbf{E}\left(|m(X)-Y|^2\right)\right)\\
&= (1+\delta)\min_{h\in\mathcal{Q}_n}\int|m_{n_l}^{(h)}(x)-m(x)|^2\mathbf{P}_X(dx).
\end{aligned}
$$

By using $a^2-b^2=(a-b)(a+b)$ we get for $T_{1,n}$

$$
\begin{aligned}
T_{1,n} &= \mathbf{E}\Big(|m_{n_l}^{(H)}(X)-Y|^2-|m_{n_l}^{(H)}(X)-T_{\beta_n}Y|^2\Big|\mathcal{D}_{n_l}\Big)\\
&\quad -\mathbf{E}\Big(|m(X)-Y|^2-|m_{\beta_n}(X)-T_{\beta_n}Y|^2\Big)\\
&= \mathbf{E}\Big((T_{\beta_n}Y-Y)(2m_{n_l}^{(H)}(X)-Y-T_{\beta_n}Y)\Big|\mathcal{D}_{n_l}\Big)\\
&\quad -\mathbf{E}\Big(\Big((m(X)-m_{\beta_n}(X))+(T_{\beta_n}Y-Y)\Big)\\
&\qquad \cdot\Big(m(X)+m_{\beta_n}(X)-Y-T_{\beta_n}Y\Big)\Big)\\
&= T_{5,n}+T_{6,n}.
\end{aligned}
$$

With the Cauchy-Schwarz inequality and

$$
I_{\{|Y|>\beta_n\}}\leq\frac{\exp(c_2/2\cdot|Y|^2)}{\exp(c_2/2\cdot\beta_n^2)}, \tag{25}
$$

it follows

$$
\begin{aligned}
|T_{5,n}| &\leq \sqrt{\mathbf{E}(|T_{\beta_n}Y-Y|^2)}\cdot\sqrt{\mathbf{E}(|2m_{n_l}^{(H)}(X)-Y-T_{\beta_n}Y|^2|\mathcal{D}_{n_l})}\\
&\leq \sqrt{\mathbf{E}(|Y|^2\cdot I_{\{|Y|>\beta_n\}})}\cdot\sqrt{\mathbf{E}(2\cdot|2m_{n_l}^{(H)}(X)-T_{\beta_n}Y|^2+2\cdot|Y|^2|\mathcal{D}_{n_l})}\\
&\leq \sqrt{\mathbf{E}\left(|Y|^2\cdot\frac{\exp(c_2/2\cdot|Y|^2)}{\exp(c_2/2\cdot\beta_n^2)}\right)}\cdot\sqrt{2(3\beta_n)^2+2\mathbf{E}(|Y|^2)}\\
&\leq \sqrt{\mathbf{E}\Big(|Y|^2\exp(c_2/2\cdot|Y|^2)\Big)}\exp\left(-\frac{c_2\cdot\beta_n^2}{4}\right)\sqrt{2(3\beta_n)^2+2\mathbf{E}(|Y|^2)},
\end{aligned}
$$

owing to the boudedness of $m_{n_l}^{(H)}$. With $x\leq\exp(x)$ for $x\in\mathbb{R}$ we get

$$
|Y|^2\leq\frac{2}{c_2}\cdot\exp\left(\frac{c_2}{2}|Y|^2\right)
$$

and hence $\mathbf{E}\Big(|Y|^2\cdot\exp(c_2/2\cdot|Y|^2)\Big)$ is bounded by

$$
\mathbf{E}\left(\frac{2}{c_2}\cdot\exp\left(c_2/2\cdot|Y|^2\right)\cdot\exp(c_2/2\cdot|Y|^2)\right)\leq\mathbf{E}\left(\frac{2}{c_2}\cdot\exp\left(c_2\cdot|Y|^2\right)\right)\leq c_6,
$$

which is less than infinity by the assumptions of the theorem. Furthermore the third term is bounded by $\sqrt{18\beta_n^2 + c_8}$, because

$$\mathbf{E}(|Y|^2) \leq \mathbf{E}(1/c_2 \cdot \exp(c_2 \cdot |Y|^2) \leq c_8 < \infty,$$

which follows again as above. With the setting $\beta_n = c_1 \cdot \log(n)$ it follows for some constants $c_9, c_{10} > 0$

$$|T_{5,n}| \leq \sqrt{c_5} \cdot \exp\left(-c_9 \cdot \log(n)^2\right) \cdot \sqrt{(18 \cdot c_1^2 \cdot \log(n)^2 + 2c_7)} \leq c_{10} \cdot \frac{\log(n)}{n}.$$

From the Cauchy-Schwarz inequality we get

$$\begin{aligned} T_{6,n} &\leq \sqrt{2\mathbf{E}\Big(|(m(X) - m_{\beta_n}(X))|^2\Big) + 2\mathbf{E}\Big(|(T_{\beta_n}Y - Y)|^2\Big)} \\ &\quad \cdot \sqrt{\mathbf{E}\left(\Big|m(X) + m_{\beta_n}(X) - Y - T_{\beta_n}Y\Big|^2\right)}, \end{aligned}$$

where we can bound the second factor on the right hand-side in the above inequality in the same way we have bounded the second factor from $T_{5,n}$, because by assumption $||m||_\infty$ is bounded, and m_{β_n} is clearly also bounded, namely by β_n. Thus, we get for some constant $c_{11} > 0$,

$$\sqrt{\mathbf{E}\left(\Big|m(X) + m_{\beta_n}(X) - Y - T_{\beta_n}Y\Big|^2\right)} \leq c_{11} \cdot \log(n).$$

Next we consider the first term. With the inequality from Jensen it follows

$$\mathbf{E}\Big(|m(X) - m_{\beta_n}(X)|^2\Big) \leq \mathbf{E}\left(\mathbf{E}\Big(|Y - T_{\beta_n}Y|^2\Big|X\Big)\right) = \mathbf{E}\Big(|Y - T_{\beta_n}Y|^2\Big).$$

Hence we get,

$$T_{6,n} \leq \sqrt{4\mathbf{E}\left(|Y - T_{\beta_n}Y|^2\right)} \cdot c_{11} \cdot \log(n),$$

and therefore the calculations from $T_{5,n}$ imply $T_{6,n} \leq c_{12} \cdot \log(n)/n$, for some constant $c_{12} > 0$. Altogether we get $T_{1,n} \leq c_{13} \cdot \log(n)/n$ for some constant $c_{13} > 0$.

With the same arguments we get also

$$\mathbf{E}\{T_{3,n}|\mathcal{D}_{n_l}\} \leq c_{13}\frac{\log(n)}{n},$$

for sufficiently large $c_{13} > 0$. Hence it suffices to show

$$\mathbf{E}(T_{2,n}|\mathcal{D}_{n_l}) \leq c_6\beta_n^2 \cdot \frac{1+\log(|\mathcal{Q}_n|)}{n_t},$$

to complete this proof. But a bound on $\mathbf{E}(T_{2,n}|\mathcal{D}_{n_l})$ can be derived analogously to the bounding of the corresponding term in the proof of Theorem 7.1 in Györfi et al. (2007) by an application of Bernstein inequality, because $T_{2,n}$ contains only the bounded versions of Y and the belonging bounded regression function. Hence this yields to the desired assertion and closes this proof. □

5.3 Proof of Theorem 1

By Lemma 2 applied with $\delta = 1$ and with $\mathcal{Q}_n = \{k_0, k_0+1, ..., n\}$ we get

$$\begin{aligned}
&\mathbf{E}\int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx) \\
&\leq 2 \min_{k \in \{k_0, k_0+1, ..., n\}} \mathbf{E}\int |m_{n_l,k}(x) - m(x)|^2 \mathbf{P}_X(dx) + 70\beta_n^2 \frac{1+\log(n)}{n_t} + c_7 \frac{\log(n)}{n}.
\end{aligned}$$

For $k \in \{k_0, k_0+1, ..., n\}$, we now use the following error decomposition:

$$\begin{aligned}
&\int |m_{n_l,k}(x) - m(x)|^2 \mathbf{P}_X(dx) \\
&= \Big[\mathbf{E}\Big(|m_{n_l,k}(X) - Y|^2 |\mathcal{D}_{n_l}\Big) - \mathbf{E}\Big(|m(X) - Y|^2\Big) \\
&\qquad -\mathbf{E}\Big(|m_{n_l,k}(X) - T_{\beta_n}Y|^2 |\mathcal{D}_{n_l}\Big) - \mathbf{E}\Big(|m_{\beta_n}(X) - T_{\beta_n}Y|^2\Big)\Big] \\
&\quad + \Big[\mathbf{E}\Big(|m_{n_l,k}(X) - T_{\beta_n}Y|^2 |\mathcal{D}_{n_l}\Big) - \mathbf{E}\Big(|m_{\beta_n}(X) - T_{\beta_n}Y|^2\Big) \\
&\qquad -2 \cdot \frac{1}{n_l}\sum_{i=1}^{n_l}\Big(|m_{n_l,k}(X_i) - T_{\beta_n}Y_i|^2 - |m_{\beta_n}(X_i) - T_{\beta_n}Y_i|^2\Big)\Big] \\
&\quad + \Big[2 \cdot \frac{1}{n_l}\sum_{i=1}^{n_l}|m_{n_l,k}(X_i) - T_{\beta_n}Y_i|^2 - 2 \cdot \frac{1}{n_l}\sum_{i=1}^{n_l}|m_{\beta_n}(X_i) - T_{\beta_n}Y_i|^2 \\
&\qquad - \Big(2 \cdot \frac{1}{n_l}\sum_{i=1}^{n_l}|m_{n_l,k}(X_i) - Y_i|^2 - 2 \cdot \frac{1}{n_l}\sum_{i=1}^{n_l}|m(X_i) - Y_i|^2\Big)\Big] \\
&\quad + \Big[2\Big(\frac{1}{n_l}\sum_{i=1}^{n_l}|m_{n_l,k}(X_i) - Y_i|^2 - \frac{1}{n_l}\sum_{i=1}^{n_l}|m(X_i) - Y_i|^2\Big)\Big] \\
&= \sum_{i=1}^{4} T_{i,n},
\end{aligned}$$

where again $T_{\beta_n}Y$ is the truncated version of Y and m_{β_n} is the regression function of $T_{\beta_n}Y$.

Both terms $T_{1,n}$ and $T_{3,n}$ can be bounded like their corresponding terms in the proof of Lemma 2, and hence we have

$$T_{1,n} \leq c_{14}\frac{\log n}{n} \quad \text{and} \quad \mathbf{E}\{T_{3,n}|\mathcal{D}_{n_l}\} \leq c_{14}\frac{\log n}{n},$$

for a constant $c_{14} > 0$. Next we consider $T_{4,n}$. Let A_{n_l} be the event, that there exists $i \in \{1, ..., n_l\}$ such that $|Y_i| > \beta_n$ and let $I_{A_{n_l}}$ be the indicator function of A_{n_l}. Then we get

$$\begin{aligned}
\mathbf{E}(T_{4,n}) &\leq 2 \cdot \mathbf{E}\left(\frac{1}{n_l}\sum_{i=1}^{n_l} |m_{n_l,k}(X_i) - Y_i|^2 \cdot I_{A_{n_l}}\right) \\
&\quad +2 \cdot \mathbf{E}\left(\frac{1}{n_l}\sum_{i=1}^{n_l} |m_{n_l,k}(X_i) - Y_i|^2 \cdot I_{A^c_{n_l}} - \frac{1}{n_l}\sum_{i=1}^{n_l} |m(X_i) - Y_i|^2\right) \\
&= 2 \cdot \mathbf{E}\left(|m_{n_l,k}(X_1) - Y_1|^2 \cdot I_{A_{n_l}}\right) \\
&\quad +2 \cdot \mathbf{E}\left(\frac{1}{n_l}\sum_{i=1}^{n_l} |m_{n_l,k}(X_i) - Y_i|^2 \cdot I_{A^c_{n_l}} - \frac{1}{n_l}\sum_{i=1}^{n_l} |m(X_i) - Y_i|^2\right) \\
&= T_{7,n} + T_{8,n}.
\end{aligned}$$

With the Cauchy-Schwarz inequality we get, for $T_{7,n}$,

$$\begin{aligned}
\frac{1}{2} \cdot T_{7,n} &\leq \sqrt{\mathbf{E}\left((|m_{n_l,k}(X_1) - Y_1|^2)^2\right)} \cdot \sqrt{\mathbf{P}(A_{n_l})} \\
&\leq \sqrt{\mathbf{E}\left((2|m_{n_l,k}(X_1)|^2 + 2|Y_1|^2)^2\right)} \cdot \sqrt{n_l \cdot \mathbf{P}\{|Y_1| > \beta_n\}} \\
&\leq \sqrt{\mathbf{E}\left(4|m_{n_l,k}(X_1)|^4 + 4|Y_1|^4\right)} \cdot \sqrt{n_l \cdot \frac{\mathbf{E}\left(\exp(c_2 \cdot |Y_1|^2)\right)}{\exp(c_2 \cdot \beta_n^2)}},
\end{aligned}$$

where the last inequality follows from inequality (25). Because $x \leq \exp(x)$ holds for all $x \in \mathbb{R}$, we get

$$\begin{aligned}
\mathbf{E}\left(|Y|^4\right) &= \mathbf{E}\left(|Y|^2 \cdot |Y|^2\right) \leq \mathbf{E}\left(\frac{2}{c_2} \cdot \exp\left(\frac{c_2}{2} \cdot |Y|^2\right) \cdot \frac{2}{c_2} \cdot \exp\left(\frac{c_2}{2} \cdot |Y|^2\right)\right) \\
&= \frac{4}{c_2^2} \cdot \mathbf{E}\left(\exp\left(c_2 \cdot |Y|^2\right)\right),
\end{aligned}$$

which is less than infinity by the assumption (14). Furthermore $||m_{n_l,k}||_\infty$ is bounded by β_n and therefore the first factor is bounded by

$$c_{15} \cdot \beta_n^2 = c_{16} \cdot \log(n)^2,$$

for some constant $c_{15}, c_{16} > 0$. The second factor is bounded by $1/n$, because (14) yields to

$$\sqrt{n_l \cdot \frac{\mathbf{E}\left(\exp(c_2 \cdot |Y_1|^2)\right)}{\exp(c_2 \cdot \beta_n^2)}} \quad \leq \quad \sqrt{n_l} \cdot \frac{\sqrt{c_{17}}}{\sqrt{\exp(c_2 \cdot \beta_n^2)}} \leq \sqrt{n_l}\sqrt{c_{17}} \cdot \exp\left(-\frac{c_{18} \cdot \log(n)^2}{2}\right).$$

Since $\exp(-c_{18} \cdot \log(n)^2) = O(n^{-2})$, further on this leads to

$$T_{7,n} \quad \leq \quad c_{19} \cdot \frac{\log(n)^2 \sqrt{n_l}}{n^2} \leq c_{20} \cdot \frac{\log(n)}{n}. \tag{26}$$

With the definition of $A_{n_l}^c$ and $m_{n_l,k}$ defined as in (8), it follows for $T_{8,n}$

$$\begin{aligned} T_{8,n} \quad &\leq \quad 2 \cdot \mathbf{E}\left(\frac{1}{n_l}\sum_{i=1}^{n_l} |\tilde{m}_{n_l,k}(X_i) - Y_i|^2 \cdot I_{A_{n_l}^c} - \frac{1}{n_l}\sum_{i=1}^{n_l} |m(X_i) - Y_i|^2\right) \\ &\leq \quad 2 \cdot \mathbf{E}\left(\frac{1}{n_l}\sum_{i=1}^{n_l} |\tilde{m}_{n_l,k}(X_i) - Y_i|^2 - \frac{1}{n_l}\sum_{i=1}^{n_l} |m(X_i) - Y_i|^2\right). \end{aligned}$$

Lemma 1 yields for arbitrary $N \in \mathbb{N}$ and $h \in \mathcal{H}_N^{\mathcal{F}}$

$$\begin{aligned} T_{8,n} \quad &\leq \quad 2 \cdot \mathbf{E}\left(4 \cdot k_0 \cdot \frac{(\alpha_1^h + \ldots + \alpha_N^h)^2}{k} + \frac{1}{n_l}\sum_{i=1}^{n_l} |h(X_i) - Y_i|^2 - \frac{1}{n_l}\sum_{i=1}^{n_l} |m(X_i) - Y_i|^2\right) \\ &= \quad 8k_0 \frac{(\alpha_1^h + \ldots + \alpha_N^h)^2}{k} + 2\int |h(x) - m(x)|^2 \mathbf{P}_X(dx), \end{aligned}$$

which together with (26) implies

$$\begin{aligned} &\mathbf{E}(T_{4,n}) \leq \\ &c_{21} \cdot \frac{\log(n)}{n} + \inf_{N \in \mathbb{N}} \inf_{h \in \mathcal{H}_N^{\mathcal{F}_n}} \left(8 \cdot k_0 \cdot \frac{(\alpha_1^h + \ldots + \alpha_N^h)^2}{k} + 2\int |h(x) - m(x)|^2 \mathbf{P}_X(dx)\right). \end{aligned}$$

The last part of the proof considers $T_{2,n}$. To get bounds on the expectation of $T_{2,n}$ we need conclusions for the covering numbers of $\mathcal{F}_n$. With the notation

$$\bigoplus_{k=1}^{K} \mathcal{F}_n \quad = \quad \left\{g : \mathbb{R}^d \to \mathbb{R}, g(x) = \sum_{k=1}^{K} g_k(x),\ (x \in \mathbb{R}^d), \text{for some } g_k \in \mathcal{F}_n, 1 \leq k \leq K\right\} \tag{27}$$

it is clear that $m_{n_l,k} \in T_\beta(\bigoplus_{i=1}^{k} \mathcal{F}_n)$. Furthermore for an arbitrary class $\mathcal{G}$ of real functions on $\mathbb{R}^d$

$$\mathcal{N}_p\left(\epsilon, T_\beta \mathcal{G}, z_1^n\right) \leq \mathcal{N}_p\left(\epsilon, \mathcal{G}, z_1^n\right) \tag{28}$$

holds, because whenever $g_1, ..., g_N$ is an L_p-ϵ-cover of $\mathcal{G}$ on z_1^n then $T_\beta g_1, ..., T_\beta g_N$ is an L_p-ϵ-cover of $T_\beta \mathcal{G}$ on z_1^n, too. Together with Lemma 16.4 in Györfi et al. (2002) this yields

$$\mathcal{N}_1\left(\epsilon, T_\beta \bigoplus_{i=1}^k \mathcal{F}_n, z_1^n\right) \leq \mathcal{N}_1\left(\frac{\epsilon}{k}, \mathcal{F}_n, z_1^n\right)^k \leq \mathcal{N}_1\left(\frac{\epsilon}{k}, \mathcal{F}_n\right)^k.$$

This bound will be used to get a bound on the following probability. We have, for arbitrary $t > 1/n$,

$$\begin{aligned}\mathbf{P}\{T_{2,n} > t\} \leq \mathbf{P}\Bigg\{ & \exists f \in T_{\beta_n} \bigoplus_{i=1}^k \mathcal{F}_n : \mathbf{E}\left(\left|\frac{f(X)}{\beta_n} - \frac{T_{\beta_n} Y}{\beta_n}\right|^2\right) - \mathbf{E}\left(\left|\frac{m_{\beta_n}(X)}{\beta_n} - \frac{T_{\beta_n} Y}{\beta_n}\right|^2\right) \\ & - \frac{1}{n_l} \sum_{i=1}^{n_l} \left(\left|\frac{f(X_i)}{\beta_n} - \frac{T_{\beta_n} Y_i}{\beta_n}\right|^2 - \left|\frac{m_{\beta_n}(X_i)}{\beta_n} - \frac{T_{\beta_n} Y_i}{\beta_n}\right|^2\right) \\ & > \frac{1}{2}\left(\frac{t}{\beta_n^2} + \mathbf{E}\left(\left|\frac{f(X)}{\beta_n} - \frac{T_{\beta_n} Y}{\beta_n}\right|^2\right) - \mathbf{E}\left(\left|\frac{m_{\beta_n}(X)}{\beta_n} - \frac{T_{\beta_n} Y}{\beta_n}\right|^2\right)\right)\Bigg\}.\end{aligned}$$

Thus with Theorem 11.4 in Györfi et al. (2002), the above derived bound, and

$$\mathcal{N}_1\left(\delta, \left\{\frac{1}{\beta_n} f : f \in \mathcal{F}_n\right\}, z_1^n\right) \leq \mathcal{N}_1\left(\delta \cdot \beta_n, \mathcal{F}_n, z_1^n\right),$$

we get for $z_1^n = (z_1, ..., z_n) \in \mathbb{R}^d \times ... \times \mathbb{R}^d$

$$\begin{aligned}\mathbf{P}\{T_{2,n} > t\} & \leq 14 \sup_{z_1^n} \mathcal{N}_1\left(\frac{t}{80\beta_n}, T_{\beta_n} \bigoplus_{i=1}^k \mathcal{F}_n, z_1^n\right) \cdot \exp\left(-\frac{n_l}{5136 \cdot \beta_n^2} t\right) \\ & \leq 14 \cdot \mathcal{N}_1\left(\frac{t}{80\beta_n \cdot k}, \mathcal{F}_n\right)^k \cdot \exp\left(-\frac{n_l}{5136 \cdot \beta_n^2} t\right).\end{aligned}$$

Using this we get for arbitrary $\epsilon \geq 1/n$

$$\begin{aligned}\mathbf{E}(T_{2,n}) & \leq \epsilon + \int_\epsilon^\infty \mathbf{P}\{T_{2,n} > t\} dt \\ & = \epsilon + 14 \cdot \mathcal{N}_1\left(\frac{1}{80\beta_n \cdot n \cdot k}, \mathcal{F}_n\right)^k \cdot \frac{5136\beta_n^2}{n_l} \cdot \exp\left(-\frac{n_l}{5136\beta_n^2}\epsilon\right).\end{aligned}$$

With

$$\epsilon = \frac{5136 \cdot \beta_n^2}{n_l} \cdot \log\left(14 \cdot \mathcal{N}_1\left(\frac{1}{80\beta_n \cdot n \cdot k}, \mathcal{F}_n\right)^k\right)$$

we get

$$\mathbf{E}(T_{2,n}) \leq \frac{c_{22} \cdot \beta_n^2 \cdot k \cdot \log\left(\mathcal{N}_1\left(\frac{1}{80\beta_n \cdot n \cdot k}, \mathcal{F}_n\right)\right)}{n_l}$$

for some sufficient large constant $c_{22} > 0$, which does not depend on n, β_n or k. Gathering the above results, the proof is complete. $\square$

5.4 Proof of Corollary 1

In the proof we will use the following bound on the covering number of $T_\beta \mathcal{F}_{l,l}$ shown in Lemma 2 in Bagirov, Clausen and Kohler (2006).

Lemma 3 *Let $x_1^n \in \mathbb{R}^d \times ... \times \mathbb{R}^d$. Then we have for $\mathcal{F}_{r_1,r_2}$ defined by (15), that*

$$\mathcal{N}_1\left(\epsilon, T_\beta \mathcal{F}_{m_1,m_2}, x_1^n\right) \leq 3\left(\frac{6e\beta}{\epsilon} \cdot m_1 \cdot m_2\right)^{2(d+2)\cdot r_1 \cdot r_2}$$

holds for all $\epsilon > 0$.

Furthermore we need the following approximation result for neural networks, which is proven in Lemma 16.8 in Györfi et al. (2002).

Lemma 4 *Let $\sigma : \mathbb{R} \to \mathbb{R}$ be a squashing function, i.e., assume that σ is monotone increasing and satisfies $\sigma(x) \to 0$ $(x \to -\infty)$ and $\sigma(x) \to 1$ $(x \to \infty)$. Then for every probability measure μ on $\mathbb{R}^d$, every measurable $f \in \mathcal{F}_C$, every $r > 0$ and every $k \geq 1$ there exists a neural network f_k in*

$$\left\{\sum_{i=1}^{k} c_i \sigma(a_i \cdot x + b_i) + c_0;\ k \in \mathbb{N}, a_i \in \mathbb{R}^d, b_i, c_i \in \mathbb{R}\right\}$$

such that

$$\int_{S_r} (f(x) - f_k(x))^2 \mu(dx) \leq \frac{(2rC)^2}{k},$$

where S_r is the closed ball around zero with radius r. The coefficients of this neural network f_k may be chosen such that $\sum_{i=0}^{k} |c_i| \leq 3rC + f(0)$.

Proof of Corollary 1. Application Theorem 1 with the choice $n_l = \lceil \frac{n}{2} \rceil$ together with Lemma 3 yields

$$\begin{aligned}
&\mathbf{E} \int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx) \\
&\leq c_{24} \frac{\log(n)^3}{n} + \min_{k \in \{k_0, k_0+1, \ldots, n\}} \left(c_{23} \left(\frac{k \cdot \log(n)^3}{n} \right) \right. \\
&\qquad \left. + \inf_{N \in \mathbb{N}} \inf_{h \in \mathcal{H}_N^{\mathcal{F}_n}} \left(16 \cdot k_0 \cdot \frac{(\alpha_1^h + \ldots + \alpha_N^h)^2}{k} + 4 \int |h(x) - m(x)|^2 \mathbf{P}_X(dx) \right) \right),
\end{aligned}$$

for large enough constants $c_{23}, c_{24} > 0$. Choosing $k = \left(\frac{n}{\log(n)^3}\right)^{1/2}$ we can bound the minimum above by

$$c_{25}\left(\frac{\log(n)^3}{n}\right)^{1/2} + \inf_{N\in\mathbb{N}} \inf_{h\in\mathcal{H}_N^{\mathcal{F}_n}} \left(\frac{16\cdot k_0\cdot(\alpha_1^h+\ldots+\alpha_N^h)^2}{\left(\frac{n}{\log(n)^3}\right)^{1/2}} + 4\int |h(x)-m(x)|^2 \mathbf{P}_X(dx)\right)$$

for sufficiently large constant $c_{25} > 0$, that does not depend on n, β_n or k.

Hence we do only need a bound on the infimum over $h \in \mathcal{H}_N^{\mathcal{F}_n}$ to conclude this proof. For this purpose we will use Lemma 4. It is quiet easy to see that, for the so-called ramp squasher σ, defined by $\sigma(x) = \max\{0, \min\{x, 1\}\}$, functions of the form

$$\sum_{i=1}^{k} c_i\sigma(a_i\cdot x + b_i)$$

are elements of $\mathcal{H}_k^{\mathcal{F}_n}$. This results from the fact, that for arbitrary $a_i \in \mathbb{R}^d$ and $b_i \in \mathbb{R}$

$$\sigma(a_i\cdot x+b_i) \;=\; \max\Big\{0\ ,\ \min\Big\{a_i\cdot x+b_i\ ,\ 1\Big\}\Big\} := f_i^+ \in \mathcal{F}_{l,l},$$

with $||f_i^+||_\infty \le 1$ and also

$$\begin{aligned} -\sigma(a_i\cdot x+b_i) &= \begin{cases} 0, & a_i\cdot x < -b_i, \\ -(a_i\cdot x+b_i), & -b_i \le a_i\cdot x \le 1-b_i, \\ -1, & a_i\cdot x > 1-b_i, \end{cases} \\ &= \max\Big\{-1\ ,\ \min\Big\{-(a_i\cdot x+b_i)\ ,\ 0\Big\}\Big\} := f_i^- \in \mathcal{F}_{l,l}, \end{aligned}$$

with $||f_i^-||_\infty \le 1$ as well, what ensures that condition (13) holds. Therefore we can rewrite

$$\sum_{i=1}^{k} c_i\sigma(a_i\cdot x + b_i),$$

by using the algebraic sign of the c_i to choose whether f_i^+ or f_i^-, as

$$|c_1|\cdot f_1^{sign(c_1)} + |c_2|\cdot f_2^{sign(c_2)} + \ldots + |c_k|\cdot f_k^{sign(c_k)}.$$

In this notation it is now obvious, that $\sum_{i=1}^k c_i\sigma(a_i\cdot x+b_i) \in \mathcal{H}_k^{\mathcal{F}_{l,l}}$, whereas the correctness of condition (12) follows from the fact, that multiplication of a function from $\mathcal{F}_{l,l}$ with a positive factor still yields a functions from $\mathcal{F}_{l,l}$. If β_n is large enough, the same is true for $\mathcal{H}_k^{T_{\beta_n}\mathcal{F}_{l,l}}$, because in this case the boundedness of the weights in Lemma 4 together with the boundedness of the regression function imply that the truncation makes no changes at all.

We have moreover assumed $X \in [-a, a]^d$ $a.s.$ and for $r = \sqrt{d} \cdot a$ we have $X \in S_r$ $a.s.$ Thus with Lemma 4 and the assumptions $N = k + 1$ and $\beta_n \geq 2 \cdot (3rC + m(0))$ we can now bound the last term:

$$\begin{aligned}
&\inf_{h \in \mathcal{H}_N} \left(16(\alpha_1^h + \ldots + \alpha_N^h)^2 \cdot \left(\frac{\log(n)}{n} \right)^{1/2} + 4 \int |h(x) - m(x)|^2 \mathbf{P}_X(dx) \right) \Bigg) \\
&\quad \leq 16 \cdot (3rC + m(0))^2 \cdot \left(\frac{\log(n)}{n} \right)^{1/2} + 4 \cdot (2rC)^2 \cdot \left(\frac{\log(n)}{n} \right)^{1/2} \\
&\quad \leq c_{26} \cdot C^2 \cdot \left(\frac{\log(n)}{n} \right)^{1/2}
\end{aligned}$$

for a sufficiently large constant c_{26}, that does not depend an r, C, n or k. □

References

[1] Bagirov, A.M., Clausen, C., and Kohler, M. (2006). Estimation of a regression function by maxima of minima of linear functions. To appear in *IEEE Transactions on Information Theory* 2009.

[2] Bagirov, A.M., Clausen, C., and Kohler, M. (2008). An algorithm for the estimation of a regression function by continuous piecewise linear functions. To appear in *Computational Optimizations and Applications.*

[3] Barron, A. R. (1994). Approximation and estimation bounds for artificial neural networks. *Machine Learning* **14**, pp. 115-133.

[4] Barron, A. R., Cohen, A., Dahmen, W., and DeVore, R. (2008). Approximation and learning by greedy algorithm. *Annals of Statistics* **36**, pp. 64 - 94.

[5] Bühlmann, P. (2006). Boosting for high-dimensional linear models. *Annals of Statistics* **34**, pp. 559 - 583.

[6] Devroye, L. (1981). On the almost everywhere convergence of nonparametric regression function estimates. *Annals of Statistics* **9**, pp. 1310–1319.

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

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

[9] Freund, Y. (1995). Boosting a weak learning algorithm by majority. *Information and Computation* **121**, pp. 256-285.

[10] Freund, Y. and Schapire, R. (1997). A decision-theoretic generalization of online learning and an application to boosting. *Journal of Computer and System Science* **55**, pp. 119-139.

[11] Friedman, J. (2001). Greedy function approximation: the gradient boosting machine. *Annals of Statistics* **29**, pp. 1189-1232

[12] Friedman, J., Hastie, T., and Tibshirani, R. (2000). Additive logistic regression: a statistical view of boosting (with discussion). *Annals of Statistics* **28**, pp. 337-407.

[13] Györfi, L., Kohler, M., Krzyżak, A. and Walk, H. (2002). *A distribution-free theory of nonparametric regression.* Springer Series in Statistics, Springer, New York.

[14] Györfi, L. and Walk, H. (1997). On the strong universal consistency of a recursive regression estimate by Pál Révész. *Statistics and Probability Letters* **31**, 177–183.

[15] Kohler, M. (2008). Multivariate orthogonal series estimates for random design regression. *Journal of Statistical Planning and Inference* **138**, pp. 3217–3237.

[16] Krzyżak, A., Linder, T., and Lugosi, G. (1996). Nonparametric estimation and classification using radial basis function nets and empirical risk minimization. *IEEE Transactions on Neural Networks* **7**, pp. 475-487.

[17] Walk, H. (2001). Strong universal pointwise consistency of recursive regression estimates. *Annals of the Institute of Statistical Mathematics* **53**, pp. 691-707.