

An L_2 -Boosting Algorithm for Estimation of a Regression Function

Adil M. Bagirov, Conny Clausen, and Michael Kohler

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.

Index Terms— L_2 -boosting, greedy algorithm, rate of convergence, regression, statistical learning.

I. 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 $f: \mathbb{R}^d \rightarrow \mathbb{R}$ 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

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

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

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

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

$$\begin{aligned} \mathbf{E}\{|f(X) - Y|^2\} \\ = \mathbf{E}\{|m(X) - Y|^2\} + \int |f(x) - m(x)|^2 \mathbf{P}_X(dx) \end{aligned}$$

Manuscript received September 29, 2008; revised October 14, 2009. Current version published March 12, 2010.

A. Bagirov is with the School of Information Technology and Mathematical Sciences, University of Ballarat, Ballarat Victoria 3353, Australia (e-mail: a.bagirov@ballarat.edu.au).

C. Clausen is with the Department of Mathematics, Universität des Saarlandes, D-66041 Saarbrücken, Germany (e-mail: clausen@math.uni-sb.de).

M. Kohler is with the Department of Mathematics, Technische Universität Darmstadt, D-64289 Darmstadt, Germany (e-mail: kohler@mathematik.tu-darmstadt.de).

Communicated by A. Krzyzak, Associate Editor for Pattern Recognition, Statistical Learning and Inference.

Digital Object Identifier 10.1109/TIT.2009.2039161

(cf., e.g., [14, Sec. 1.1]). In addition, (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

$$\int |f(x) - m(x)|^2 \mathbf{P}_X(dx) \quad (3)$$

is small. This motivates to measure the error caused by using a function f instead of the regression function by the L_2 error (3).

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) , ... are independent and identically distributed i.i.d. 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

$$m_n(\cdot) = m_n(\cdot, \mathcal{D}_n): \mathbb{R}^d \rightarrow \mathbb{R}$$

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 [14].

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., [7], [15], [16], [17], or [18].

A closely related problem to nonparametric regression is pattern recognition, where Y takes on values only in a finite set (cf., e.g., [8]). One of the main achievements in pattern recognition in the last fifteen years was boosting (cf. [10] and [11]), 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. [13]). 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. [12]). [6] showed that this so-called L_2 -boosting is able to estimate very high-dimensional linear models well. Reference [5] 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 a linear combination of these functions. In [5], 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 [4] showed for least squares neural networks.

In this paper, we modify the general algorithm from [5] 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 [16] 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 II contains the definition and our theoretical result on the general L_2 -boosting algorithm. In Section III, 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 IV. Finally, Section V contains the proofs.

II. A GENERAL L_2 -BOOSTING ALGORITHM

Let $n_l, n_t \in \mathbb{N}$ be such that $n = n_l + n_t$, let $\beta_n \in \mathbb{R}_+$ (which will later be chosen such that $\beta_n \rightarrow \infty$ ($n \rightarrow \infty$)), and let \mathcal{F} be a (nonlinear) class of functions $f : \mathbb{R}^d \rightarrow \mathbb{R}$. Set

$$T_{\beta_n} \mathcal{F} = \{T_{\beta_n} f : f \in \mathcal{F}\}$$

where $(T_{\beta_n} f)(x) = T_{\beta_n}(f(x))$ and

$$T_{\beta_n} z = \max\{-\beta, \min\{\beta, z\}\} \quad (z \in \mathbb{R})$$

and define

$$\mathcal{F}_n = T_{\beta_n} \mathcal{F}. \quad (4)$$

Depending on a parameter $k_0 \in \{1, \dots, n\}$, we define estimates

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

as follows. Set

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

and

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

where

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

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 $\pm\beta_n$. More precisely, we set for $k \in \{k_0, k_0 + 1, \dots, n\}$

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

Finally we use splitting of the sample to select the parameter k of the estimate. To do this, we set

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

where

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

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

Definition 1: Let $x_1, \dots, x_n \in \mathbb{R}^d$ and set $x_1^n = (x_1, \dots, x_n)$. Let \mathcal{G} be a set of functions $g : \mathbb{R}^d \rightarrow \mathbb{R}$. An L_p - ϵ -cover of \mathcal{G} on x_1^n is a finite set of functions $g_1, \dots, g_k : \mathbb{R}^d \rightarrow \mathbb{R}$ with the property

$$\min_{1 \leq j \leq k} \left(\frac{1}{n} \sum_{i=1}^n |g(x_i) - g_j(x_i)|^p \right)^{1/p} < \epsilon \quad \text{for all } g \in \mathcal{G}. \quad (11)$$

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

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

$$\left(\frac{2}{l} \sum_{i=1}^N \alpha_i^h \right) \cdot g_j \in \mathcal{G} \quad (12)$$

for all $j \in \{1, \dots, N\}$, $l \in \{1, \dots, n\}$, and

$$\|g_j\|_{\infty} = \sup_{x \in \mathbb{R}^d} |g_j(x)| \leq 1 \quad (13)$$

for all $j \in \{1, \dots, N\}$ and $x \in \mathbb{R}^d$ are satisfied. Our main theoretical result is the following theorem.

Theorem 1: Let \mathcal{F} be a class of functions $f : \mathbb{R}^d \rightarrow \mathbb{R}$. Let $\mathcal{N}_1(\epsilon, T_{\beta_n} \mathcal{F})$ be an upper bound on the L_1 - ϵ -covering number of $T_{\beta_n} \mathcal{F}$ on any finite set of points, i.e., assume

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

Define the estimate m_n by (4)–(10) with $\beta_n = c_1 \cdot \log(n)$ for some $c_1 > 0$. Furthermore assume that the distribution of (X, Y) satisfies

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

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

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

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

The upper bound on the expected L_2 error in Theorem 1 can be interpreted as follows: If we ignore the minimum over k then the first term in the sum is the usual bound for the estimation error of a least squares estimate in case that a sum of k functions from $T_{\beta_n} \mathcal{F}$ is fitted to the data. The second term in the sum measures the approximation error, where besides the usual bound

$$\inf_{h \in \mathcal{H}_N^{T_{\beta_n} \mathcal{F}}} \int |h(x) - m(x)|^2 \mathbf{P}_X(dx)$$

an additional term occurs which comes from the fact that we use a greedy algorithm to minimize the empirical L_2 risk of the estimate. Finally the minimum in front of the sum of these two terms shows that by splitting of the sample our estimate behaves in view of the above error bound (up to some constant factor) as good as if we have chosen the value of k optimally according to the underlying distribution. In this sense our estimate is able to adapt to the underlying distribution.

Remark 1: In principle it is also possible to choose the parameter k_0 of the estimate by splitting of the sample. But in case of the simulated data in Section IV it turned out that the estimate improves always for large values of k_0 . Therefore we choose in our simulations a fixed very large value for k_0 ($k_0 = 1000$).

III. 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 [1], i.e., we apply it to

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

where

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

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

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

for a suitable chosen squashing function σ (i.e., for a suitable chosen monotone increasing function $\sigma : \mathbb{R} \rightarrow \mathbb{R}$ satisfying $\sigma(x) \rightarrow 0$ ($x \rightarrow -\infty$) and $\sigma(x) \rightarrow 1$ ($x \rightarrow \infty$)). 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 $\mathcal{F}_{l,l}$ as function class for the general algorithm of Section II for some $l \geq 2$. Here l is independent of the sample size. In the application in Section IV, we will choose l depending on the dimension d of X and we will use larger values of l in case of larger d .

It is well known that in order to derive nontrivial rate of convergence results, we have to make some smoothness assumptions on the regression function (cf., e.g., [8, Th. 7.2 and Prob. 7.2] and [9, Sec. 3]). In the sequel we will impose such smoothness conditions implicitly 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

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

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 $C \in \mathbb{R}_+$ (cf. [3]). We denote the class of functions $f : \mathbb{R}^d \rightarrow \mathbb{R}$, which satisfy (16) and (17) by \mathcal{F}^C .

Condition (17) is often used for the analysis of the rate of convergence of neural network regression estimates. It is an extremely strong assumption, in particular it implies that the

smoothness of the function f increases more and more as dimension d of X grows. By imposing it on the regression function we are able to derive the following rate of convergence result for our estimate.

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 m is bounded in absolute value by some constant less than or equal to β_n and that $m \in \mathcal{F}^C$ for some $0 < C < \infty$. Let the estimate m_n be defined by (4)–(10), with $\mathcal{F} = \mathcal{F}_{l,l}$ for some $l \geq 2$, and with $n_l = \lceil \frac{n}{2} \rceil$. Then we have for $\beta_n \geq 6 \cdot \sqrt{d} \cdot a \cdot C$

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

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

Remark 2: By using standard approximation result for neural networks (e.g., [14]) it is easy to see that the proof of Corollary 1 implies

$$\mathbf{E} \int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx) \rightarrow 0 \quad (n \rightarrow \infty)$$

for all distributions of (X, Y) satisfying (14) and m bounded in absolute value. By a careful analysis of the proof of Lemma 2 it should be possible to show the same result even for all distributions of (X, Y) satisfying $\mathbf{E}Y^2 < \infty$.

IV. 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 [2]. There data were 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 functions have been considered:

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

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 two different sample sizes. The *maxmin*-estimate proposed in [2], which uses splitting of the sample and the principle of least squares to fit a maximum of minima of linear functions to a data set, has been compared for $d = 1$ with kernel estimates (with Gaussian kernel) (see, e.g., [14, Ch. 5]), local linear kernel estimates (see, e.g., [14, Sec. 5.4]), smoothing splines (see, e.g., [14, Ch. 20]), 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 $\tilde{U}_1, \tilde{U}_2, \dots$ 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 were reported.

In the sequel, we make the same simulations with our newly proposed L_2 —boosting estimate. Here we set $l = 4$ for $d \in \{1, 2\}$ and $l = 5$ for $d = 10$, $k_0 = 1000$, repeat seven 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 [2], 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 $d > 1$).

Tables I and II summarize the results for the four univariate regression functions m_1, \dots, m_4 , Tables III and IV summarize the results for the three bivariate regression functions m_5, m_6 and m_7 and Tables V and VI summarize the results for the four regression functions m_8, \dots, m_{11} where $d = 10$.

Considering the results in Tables I–VI we can first see, that the error of our L_2 —boosting estimate is in 47 cases less than but only in 15 cases 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.

Second, by looking at Table VI we can see that the L_2 —boosting estimate is especially suited for high-dimen-

TABLE I
SIMULATION RESULTS AND COMPARISON WITH SIX OTHER NONPARAMETRIC REGRESSION ESTIMATES FOR FOUR UNIVARIATE REGRESSION FUNCTIONS AND SAMPLE SIZE $n = 500$

	σ	0	0.5	1
m_1	Error	0.0000	0.0078	0.0290
	Std. deviation	(0.0000)	(0.0043)	(0.0166)
	Comparison	= / 1	+ / 1	+ / 1
m_2	Error	0.0041	0.0157	0.0389
	Std. deviation	(0.0039)	(0.0128)	(0.0157)
	Comparison	+ / 1	- / 1	+ / 2
m_3	Error	0.0155	0.0272	0.01070
	Std. deviation	(0.0343)	(0.0129)	(0.0459)
	Comparison	+ / 3	- / 1	+ / 2
m_4	Error	0.0006	0.0184	0.0508
	Std. deviation	(0.0003)	(0.0054)	(0.0151)
	Comparison	+ / 5	+ / 4	+ / 3

TABLE II
SIMULATION RESULTS AND COMPARISON WITH SIX OTHER NONPARAMETRIC REGRESSION ESTIMATES FOR FOUR UNIVARIATE REGRESSION FUNCTIONS AND SAMPLE SIZE $n = 5000$

	σ	0	0.5	1
m_1	Error	0.0000	0.0007	0.0025
	Std. deviation	(0.0000)	(0.0004)	(0.0014)
	Comparison	= / 1	= / 1	+ / 1
m_2	Error	0.0007	0.0012	0.0040
	Std. deviation	(0.0010)	(0.0008)	(0.0020)
	Comparison	= / 1	+ / 1	- / 1
m_3	Error	0.0015	0.0041	0.0072
	Std. deviation	(0.0012)	(0.0022)	(0.0031)
	Comparison	- / 3	- / 2	- / 1
m_4	Error	0.0006	0.0030	0.0088
	Std. deviation	(0.0004)	(0.0007)	(0.0027)
	Comparison	+ / 6	+ / 5	+ / 6

TABLE III
SIMULATION RESULTS AND COMPARISON WITH THREE OTHER NONPARAMETRIC REGRESSION ESTIMATES FOR THREE BIVARIATE REGRESSION FUNCTIONS AND SAMPLE SIZE $n = 500$

	σ	0	0.5	1
m_5	Error	0.0322	0.1036	0.2076
	Std. deviation	(0.0075)	(0.0226)	(0.0426)
	Comparison	+ / 2	+ / 2	+ / 1
m_6	Error	0.0143	0.0645	0.1486
	Std. deviation	(0.0045)	(0.0143)	(0.0330)
	Comparison	- / 2	- / 2	+ / 1
m_7	Error	0.0317	0.1192	0.1952
	Std. deviation	(0.0150)	(0.0310)	(0.0469)
	Comparison	+ / 2	- / 1	- / 1

TABLE IV
SIMULATION RESULTS AND COMPARISON WITH THREE OTHER NONPARAMETRIC REGRESSION ESTIMATES FOR THREE BIVARIATE REGRESSION FUNCTIONS AND SAMPLE SIZE $n = 5000$

	σ	0	0.5	1
m_5	Error	0.0089	0.0212	0.0445
	Std. deviation	(0.0024)	(0.0037)	(0.0118)
	Comparison	+ / 2	+ / 2	+ / 2
m_6	Error	0.0064	0.0123	0.0311
	Std. deviation	(0.0013)	(0.0015)	(0.0039)
	Comparison	+ / 2	+ / 2	+ / 1
m_7	Error	0.0049	0.0234	0.0392
	Std. deviation	(0.0018)	(0.0123)	(0.0133)
	Comparison	+ / 1	- / 1	- / 1

sional data sets and large sample size in comparison with other nonparametric regression estimates.

TABLE V
SIMULATION RESULTS AND COMPARISON WITH THREE OTHER NONPARAMETRIC REGRESSION ESTIMATES FOR REGRESSION FUNCTIONS WHERE $d = 10$ FOR SAMPLE SIZE $n = 500$

	σ	0	0.5	1
m_8	Error	4.4171	4.4991	4.5242
	Std. deviation	(0.1606)	(0.1774)	(0.1687)
	Comparison	+ / 1	- / 1	+ / 1
m_9	Error	0.5656	0.6867	0.8648
	Std. deviation	(0.1602)	(0.1287)	(0.0629)
	Comparison	+ / 3	+ / 3	+ / 3
m_{10}	Error	0.1529	0.2078	0.2441
	Std. deviation	(0.0309)	(0.0271)	(0.0614)
	Comparison	+ / 1	+ / 2	+ / 3
m_{11}	Error	0.0698	0.1983	0.4128
	Std. deviation	(0.0252)	(0.0662)	(0.0425)
	Comparison	+ / 1	+ / 1	+ / 1

TABLE VI
SIMULATION RESULTS AND COMPARISON WITH THREE OTHER NONPARAMETRIC REGRESSION ESTIMATES FOR REGRESSION FUNCTIONS WHERE $d = 10$ FOR SAMPLE SIZE $n = 5000$

	σ	0	0.5	1
m_8	Error	1.1063	1.1496	1.1246
	Std. deviation	(0.1292)	(0.1190)	(0.1622)
	Comparison	+ / 1	+ / 1	+ / 1
m_9	Error	0.0348	0.0461	0.1247
	Std. deviation	(0.0142)	(0.0073)	(0.0327)
	Comparison	- / 2	+ / 2	+ / 3
m_{10}	Error	0.0185	0.0410	0.0962
	Std. deviation	(0.0029)	(0.0053)	(0.0173)
	Comparison	+ / 1	+ / 1	+ / 1
m_{11}	Error	0.0169	0.0251	0.0514
	Std. deviation	(0.0034)	(0.0052)	(0.0174)
	Comparison	- / 1	+ / 1	+ / 1

V. PROOFS

A. A Deterministic Lemma

Let \mathcal{F} be a class of functions $f : \mathbb{R}^d \rightarrow \mathbb{R}$ and let $(x_1, y_1), \dots, (x_n, y_n) \in \mathbb{R}^d \times \mathbb{R}$, $k_0 \in \mathbb{N}$ 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

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

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

$$\left(\frac{2}{l} \sum_{i=1}^N \alpha_i\right) \cdot g_j \in \mathcal{F}$$

for all $j \in \{1, \dots, N\}$, $l \in \{1, \dots, k\}$, and

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

we have

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

The proof of the above lemma is a modification of [5, Proof of Th. 2.4]. For the sake of completeness, we repeat it as follows.

Proof of Lemma 1: In the first step of the proof we show

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

To do this, let $j \in \{1, \dots, N\}$ 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

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

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

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

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

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

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

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

which implies the assertion of the first step.

In the second step of the proof we show

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

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

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

and arguing as above we get

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

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

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

satisfies

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

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 $a_k \leq 4M/k$ implies

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

where the last inequality follows from

$$\begin{aligned}
& \left(1 - \frac{2}{k+1}\right) \cdot \frac{1}{k} + \frac{1}{(k+1)^2} = \frac{k^2 + k - 1}{k^2 + k} \cdot \frac{1}{k+1} \\
& \leq \frac{1}{k+1}.
\end{aligned}$$

□

B. Splitting of the Sample for Unbounded Y

The following lemma is an extension of [14, Th. 7.1] 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 $\mathcal{D}_{n_l} = \{(X_1, Y_1), \dots, (X_{n_l}, Y_{n_l})\}$. Then, we define

$$m_n(x) = m_{n_l}^{(H)}(x) \text{ for all } x \in \mathbb{R}^d, \quad (21)$$

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

$$\begin{aligned}
& \frac{1}{n_t} \sum_{i=n_l+1}^n \left| m_{n_l}^{(H)}(X_i) - Y_i \right|^2 \\
& = \min_{h \in \mathcal{Q}_n} \frac{1}{n_t} \sum_{i=n_l+1}^n \left| m_{n_l}^{(h)}(X_i) - Y_i \right|^2. \quad (22)
\end{aligned}$$

Lemma 2: Let $\beta_n = c_1 \cdot \log(n)$ for some constant $c_1 > 0$ and assume that the estimates $m_{n_i}^{(h)}$ 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 (21) and (22) and any $\delta > 0$,

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

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

Proof: We use the following error decomposition

$$\begin{aligned} & \mathbf{E} \left(\int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx) \middle| \mathcal{D}_{n_i} \right) \\ & = \mathbf{E} \left(\int \left| m_{n_i}^{(H)}(x) - m(x) \right|^2 \mathbf{P}_X(dx) \middle| \mathcal{D}_{n_i} \right) \\ & = \sum_{i=1}^4 T_{i,n}, \end{aligned}$$

where

$$\begin{aligned} T_{1,n} & = \mathbf{E} \left(\left| m_{n_i}^{(H)}(X) - Y \right|^2 \middle| \mathcal{D}_{n_i} \right) - \mathbf{E} \left(|m(X) - Y|^2 \right) \\ & \quad - \mathbf{E} \left(\left| m_{n_i}^{(H)}(X) - T_{\beta_n} Y \right|^2 \middle| \mathcal{D}_{n_i} \right) \\ & \quad - \mathbf{E} \left(|m_{\beta_n}(X) - T_{\beta_n} Y|^2 \right) \end{aligned}$$

$$\begin{aligned} T_{2,n} & = \mathbf{E} \left(\left| m_{n_i}^{(H)}(X) - T_{\beta_n} Y \right|^2 \middle| \mathcal{D}_{n_i} \right) \\ & \quad - \mathbf{E} \left(|m_{\beta_n}(X) - T_{\beta_n} Y|^2 \right) - (1 + \delta) \\ & \quad \cdot \frac{1}{n_t} \sum_{i=n_l+1}^n \left(\left| m_{n_i}^{(H)}(X_i) - T_{\beta_n} Y_i \right|^2 \right. \\ & \quad \left. - |m_{\beta_n}(X_i) - T_{\beta_n} Y_i|^2 \right) \end{aligned}$$

$$\begin{aligned} T_{3,n} & = (1 + \delta) \cdot \frac{1}{n_t} \sum_{i=n_l+1}^n \left(\left| m_{n_i}^{(H)}(X_i) - T_{\beta_n} Y_i \right|^2 \right. \\ & \quad \left. - |m_{\beta_n}(X_i) - T_{\beta_n} Y_i|^2 \right) \\ & \quad - (1 + \delta) \cdot \frac{1}{n_t} \sum_{i=n_l+1}^n \left(\left| m_{n_i}^{(H)}(X_i) - Y_i \right|^2 \right. \\ & \quad \left. - |m(X_i) - Y_i|^2 \right) \end{aligned}$$

$$\begin{aligned} T_{4,n} & = (1 + \delta) \cdot \frac{1}{n_t} \sum_{i=n_l+1}^n \left(\left| m_{n_i}^{(H)}(X_i) - Y_i \right|^2 \right. \\ & \quad \left. - |m(X_i) - Y_i|^2 \right) \end{aligned}$$

and 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 (22) we can bound the last term $T_{4,n}$ by

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

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

$$\mathbf{E}(T_{4,n} | \mathcal{D}_{n_i}) \leq (1 + \delta) \min_{h \in \mathcal{Q}_n} \int \left| m_{n_i}^{(h)}(x) - m(x) \right|^2 \mathbf{P}_X(dx).$$

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

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

With the Cauchy–Schwarz inequality and

$$I_{\{|Y| > \beta_n\}} \leq \frac{\exp(c_2/2 \cdot |Y|^2)}{\exp(c_2/2 \cdot \beta_n^2)} \quad (23)$$

it follows that

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

owing to the boundedness of $m_{n_i}^{(H)}$. 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 $\mathbf{E}(|Y|^2 \cdot \exp(c_2/2 \cdot |Y|^2))$ is bounded by

$$\begin{aligned} \mathbf{E}\left(\frac{2}{c_2} \cdot \exp(c_2/2 \cdot |Y|^2) \cdot \exp(c_2/2 \cdot |Y|^2)\right) \\ \leq \mathbf{E}\left(\frac{2}{c_2} \cdot \exp(c_2 \cdot |Y|^2)\right) \leq c_8 \end{aligned}$$

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 $c_9, c_{10} > 0$

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

From the Cauchy–Schwarz inequality, we get

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

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 $c_{11} > 0$

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

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

$$\begin{aligned} \mathbf{E}\left(|m(X) - m_{\beta_n}(X)|^2\right) &\leq \mathbf{E}\left(\mathbf{E}\left(|Y - T_{\beta_n}Y|^2 \middle| X\right)\right) \\ &= \mathbf{E}\left(|Y - T_{\beta_n}Y|^2\right). \end{aligned}$$

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

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

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

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

to complete this proof. But a bound on $\mathbf{E}(T_{2,n}|\mathcal{D}_{n_i})$ can be derived analogously to the bounding of the corresponding term in the [14, Proof of Th. 7.1] 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. \square

C. Proof of Theorem 1

By Lemma 2 applied with $\delta = 1$ and with

$$\mathcal{Q}_n = \{k_0, k_0 + 1, \dots, n\}$$

we get

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

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

$$\begin{aligned} \int |m_{n_i,k}(x) - m(x)|^2 \mathbf{P}_X(dx) \\ = \sum_{i=1}^4 T_{i,n} \end{aligned}$$

where

$$\begin{aligned} T_{1,n} &= \mathbf{E}\left(|m_{n_i,k}(X) - Y|^2|\mathcal{D}_{n_i}\right) - \mathbf{E}\left(|m(X) - Y|^2\right) \\ &\quad - \left(\mathbf{E}\left(|m_{n_i,k}(X) - T_{\beta_n}Y|^2|\mathcal{D}_{n_i}\right)\right) \\ &\quad - \mathbf{E}\left(|m_{\beta_n}(X) - T_{\beta_n}Y|^2\right) \end{aligned}$$

$$\begin{aligned} T_{2,n} &= \mathbf{E}\left(|m_{n_i,k}(X) - T_{\beta_n}Y|^2|\mathcal{D}_{n_i}\right) \\ &\quad - \mathbf{E}\left(|m_{\beta_n}(X) - T_{\beta_n}Y|^2\right) \\ &\quad - 2 \cdot \frac{1}{n_i} \sum_{i=1}^{n_i} \left(|m_{n_i,k}(X_i) - T_{\beta_n}Y_i|^2\right. \\ &\quad \left. - |m_{\beta_n}(X_i) - T_{\beta_n}Y_i|^2\right) \end{aligned}$$

$$\begin{aligned} T_{3,n} \\ = 2 \cdot \frac{1}{n_i} \sum_{i=1}^{n_i} |m_{n_i,k}(X_i) - T_{\beta_n}Y_i|^2 \end{aligned}$$

$$\begin{aligned}
& -2 \cdot \frac{1}{n_l} \sum_{i=1}^{n_l} |m_{\beta_n}(X_i) - T_{\beta_n} Y_i|^2 \\
& - \left(2 \cdot \frac{1}{n_l} \sum_{i=1}^{n_l} |m_{n_l,k}(X_i) - Y_i|^2 \right. \\
& \quad \left. - 2 \cdot \frac{1}{n_l} \sum_{i=1}^{n_l} |m(X_i) - Y_i|^2 \right)
\end{aligned}$$

and

$$\begin{aligned}
T_{4,n} & = 2 \left(\frac{1}{n_l} \sum_{i=1}^{n_l} |m_{n_l,k}(X_i) - Y_i|^2 \right. \\
& \quad \left. - \frac{1}{n_l} \sum_{i=1}^{n_l} |m(X_i) - Y_i|^2 \right).
\end{aligned}$$

Here, 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

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

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

$$\begin{aligned}
\mathbf{E}(T_{4,n}) & \leq 2 \cdot \mathbf{E} \left(\frac{1}{n_l} \sum_{i=1}^{n_l} |m_{n_l,k}(X_i) - Y_i|^2 \cdot I_{A_{n_l}} \right) \\
& \quad + 2 \cdot \mathbf{E} \left(\frac{1}{n_l} \sum_{i=1}^{n_l} |m_{n_l,k}(X_i) - Y_i|^2 \cdot I_{A_{n_l}^c} \right. \\
& \quad \quad \left. - \frac{1}{n_l} \sum_{i=1}^{n_l} |m(X_i) - Y_i|^2 \right) \\
& = 2 \cdot \mathbf{E} \left(|m_{n_l,k}(X_1) - Y_1|^2 \cdot I_{A_{n_l}} \right) \\
& \quad + 2 \cdot \mathbf{E} \left(\frac{1}{n_l} \sum_{i=1}^{n_l} |m_{n_l,k}(X_i) - Y_i|^2 \cdot I_{A_{n_l}^c} \right. \\
& \quad \quad \left. - \frac{1}{n_l} \sum_{i=1}^{n_l} |m(X_i) - Y_i|^2 \right) \\
& = T_{5,n} + T_{6,n}.
\end{aligned}$$

By the Cauchy-Schwarz inequality we get, for $T_{5,n}$

$$\begin{aligned}
\frac{1}{2} \cdot T_{5,n} & \leq \sqrt{\mathbf{E} \left((|m_{n_l,k}(X_1) - Y_1|^2)^2 \right)} \cdot \sqrt{\mathbf{P}(A_{n_l})} \\
& \leq \sqrt{\mathbf{E} \left((2|m_{n_l,k}(X_1)|^2 + 2|Y_1|^2)^2 \right)} \\
& \quad \cdot \sqrt{n_l \cdot \mathbf{P}\{|Y_1| > \beta_n\}} \\
& \leq \sqrt{\mathbf{E} \left(4|m_{n_l,k}(X_1)|^4 + 4|Y_1|^4 \right)} \\
& \quad \cdot \sqrt{n_l \cdot \frac{\mathbf{E}(\exp(c_2 \cdot |Y_1|^2))}{\exp(c_2 \cdot \beta_n^2)}}
\end{aligned}$$

where the last inequality follows analogously to inequality (23). Because $x \leq \exp(x)$ holds for all $x \in \mathbb{R}$, we get

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

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) leads to

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

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

$$T_{5,n} \leq c_{19} \cdot \frac{\log(n)^2 \sqrt{n_l}}{n^2} \leq c_{20} \cdot \frac{\log(n)}{n}. \quad (24)$$

With the definition of $A_{n_l}^c$ and $m_{n_l,k}$ defined as in (8), it follows for $T_{6,n}$

$$\begin{aligned}
T_{6,n} & \leq 2 \cdot \mathbf{E} \left(\frac{1}{n_l} \sum_{i=1}^{n_l} |\tilde{m}_{n_l,k}(X_i) - Y_i|^2 \cdot I_{A_{n_l}^c} \right. \\
& \quad \left. - \frac{1}{n_l} \sum_{i=1}^{n_l} |m(X_i) - Y_i|^2 \right) \\
& \leq 2 \cdot \mathbf{E} \left(\frac{1}{n_l} \sum_{i=1}^{n_l} |\tilde{m}_{n_l,k}(X_i) - Y_i|^2 \right. \\
& \quad \left. - \frac{1}{n_l} \sum_{i=1}^{n_l} |m(X_i) - Y_i|^2 \right).
\end{aligned}$$

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

$$\begin{aligned}
T_{6,n} & \leq 2 \cdot \mathbf{E} \left(4 \cdot k_0 \cdot \frac{(\alpha_1^h + \cdot s + \alpha_N^h)^2}{k} \right. \\
& \quad \left. + \frac{1}{n_l} \sum_{i=1}^{n_l} |h(X_i) - Y_i|^2 \right. \\
& \quad \left. - \frac{1}{n_l} \sum_{i=1}^{n_l} |m(X_i) - Y_i|^2 \right) \\
& = 8k_0 \frac{(\alpha_1^h + \cdot s + \alpha_N^h)^2}{k} \\
& \quad + 2 \int |h(x) - m(x)|^2 \mathbf{P}_X(dx)
\end{aligned}$$

which together with (24) implies

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

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 notations $\delta_{k_0} = 1$

$$\delta_k = \left(1 - \frac{1}{k_0 + 1}\right) \cdot \left(1 - \frac{1}{k_0 + 2}\right) \cdots \left(1 - \frac{1}{k}\right) \in [0, 1]$$

and

$$\begin{aligned} \bigoplus_{k=k_0}^K \mathcal{F}_n &= \left\{ g : \mathbb{R}^d \rightarrow \mathbb{R}, g(x) = \sum_{k=k_0}^K \delta_k \cdot g_k(x) \right. \\ &\quad \left. (x \in \mathbb{R}^d), \text{ for some } g_k \in \mathcal{F}_n, 1 \leq k \leq K \right\} \quad (25) \end{aligned}$$

it is clear that $m_{n_i, k} \in T_\beta \left(\bigoplus_{i=k_0}^k \mathcal{F}_n \right)$. Furthermore, for an arbitrary class \mathcal{G} of real functions on \mathbb{R}^d

$$\mathcal{N}_p(\epsilon, T_\beta \mathcal{G}, z_1^n) \leq \mathcal{N}_p(\epsilon, \mathcal{G}, z_1^n) \quad (26)$$

holds, because whenever g_1, \dots, g_N is an L_p - ϵ -cover of \mathcal{G} on z_1^n then $T_\beta g_1, \dots, T_\beta g_N$ is an L_p - ϵ -cover of $T_\beta \mathcal{G}$ on z_1^n , too. Together with [14, Proof of Lemma 16.4] this yields

$$\begin{aligned} \mathcal{N}_1 \left(\epsilon, T_\beta \bigoplus_{i=k_0}^k \mathcal{F}_n, z_1^n \right) &\leq \mathcal{N}_1 \left(\frac{\epsilon}{k}, \mathcal{F}_n, z_1^n \right)^k \\ &\leq \mathcal{N}_1 \left(\frac{\epsilon}{k}, \mathcal{F}_n \right)^k. \end{aligned}$$

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

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

Thus by [14, Th. 11.4], the above derived bound, and

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

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

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

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

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

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

D. 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 [1, Lemma 2].

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

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

holds for all $\epsilon > 0$.

Furthermore we need the following approximation result for neural networks, which is proven [14, Lemma 16.8].

Lemma 4: Let $\sigma : \mathbb{R} \rightarrow \mathbb{R}$ be a squashing function, i.e., assume that σ is monotone increasing and satisfies $\sigma(x) \rightarrow 0$ ($x \rightarrow -\infty$) and $\sigma(x) \rightarrow 1$ ($x \rightarrow \infty$). 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 of Theorem 1 with the choice $n_l = \lceil \frac{n}{2} \rceil$ together with Lemma 3 yields

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

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

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

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

Hence we 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 quite 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 $b_i \in \mathbb{R}$

$$\begin{aligned} \sigma(a_i \cdot x + b_i) &= \max \left\{ 0, \min \{a_i \cdot x + b_i, 1\} \right\} \\ &=: f_i^+ \in \mathcal{F}_{l,l} \end{aligned}$$

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

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

with $\|f_i^-\|_\infty \leq 1$ 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

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

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 $\mathcal{H}_k^{\mathcal{F}_{\beta_n} \mathcal{F}_{l,l}}$, 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

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

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

ACKNOWLEDGMENT

The authors would like to thank two anonymous referees and the associate editor for many detailed and helpful comments.

REFERENCES

- [1] A. M. Bagirov, C. Clausen, and M. Kohler, "Estimation of a regression function by maxima of minima of linear functions," *IEEE Trans. Inf. Theory*, vol. 55, pp. 833–845, 2009.
- [2] A. M. Bagirov, C. Clausen, and M. Kohler, "An algorithm for the estimation of a regression function by continuous piecewise linear functions," *Computat. Optim. Applicat.*, 2008, to be published.
- [3] A. R. Barron, "Universal approximation bounds for superpositions of a sigmoidal function," *IEEE Trans. Inf. Theory*, vol. 39, pp. 930–944, 1993.
- [4] A. R. Barron, "Approximation and estimation bounds for artificial neural networks," *Machine Learn.*, vol. 14, pp. 115–133, 1994.

- [5] A. R. Barron, A. Cohen, W. Dahmen, and R. DeVore, "Approximation and learning by greedy algorithm," *Ann. Stat.*, vol. 36, pp. 64–94, 2008.
- [6] P. Bühlmann, "Boosting for high-dimensional linear models," *Ann. Stat.*, vol. 34, pp. 559–583, 2006.
- [7] L. Devroye, "On the almost everywhere convergence of nonparametric regression function estimates," *Ann. Stat.*, vol. 9, pp. 1310–1319, 1981.
- [8] L. Devroye, L. Györfi, and G. Lugosi, *A Probabilistic Theory of Pattern Recognition*. New York: Springer, 1996.
- [9] L. P. Devroye and T. J. Wagner, "Distribution-free consistency results in nonparametric discrimination and regression function estimation," *Ann. Stat.*, vol. 8, pp. 231–239, 1980.
- [10] Y. Freund, "Boosting a weak learning algorithm by majority," *Inf. Comput.*, vol. 121, pp. 256–285, 1995.
- [11] Y. Freund and R. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *J. Comput. Syst. Sci.*, vol. 55, pp. 119–139, 1997.
- [12] J. Friedman, "Greedy function approximation: The gradient boosting machine," *Ann. Stat.*, vol. 29, pp. 1189–1232, 2001.
- [13] J. Friedman, T. Hastie, and R. Tibshirani, "Additive logistic regression: A statistical view of boosting (with discussion)," *Ann. Stat.*, vol. 28, pp. 337–407, 2000.
- [14] L. Györfi, M. Kohler, A. Krzyżak, and H. Walk, *A Distribution-Free Theory of Nonparametric Regression*, ser. Springer Series in Statistics. New York: Springer, 2002.
- [15] L. Györfi and H. Walk, "On the strong universal consistency of a recursive regression estimate by Pál Révész," *Stat. Prob. Lett.*, vol. 31, pp. 177–183, 1997.
- [16] M. Kohler, "Multivariate orthogonal series estimates for random design regression," *J. Stat. Planning Inference*, vol. 138, pp. 3217–3237, 2008.
- [17] A. Krzyżak, T. Linder, and G. Lugosi, "Nonparametric estimation and classification using radial basis function nets and empirical risk minimization," *IEEE Trans. Neural Netw.*, vol. 7, pp. 475–487, 1996.
- [18] H. Walk, "Strong universal pointwise consistency of recursive regression estimates," *Ann. Inst. Stat. Math.*, vol. 53, pp. 691–707, 2001.

Adil M. Bagirov was born on January 7, 1960 in Bilesuvar, Azerbaijan. He received the Master degree in applied mathematics from the Baku State University, Azerbaijan, in 1983, and the Ph.D. degrees in mathematical cybernetics from the Institute of Cybernetics Azerbaijan National Academy of Sciences in 1989 and in optimization from the University of Ballarat, Australia, in 2001.

From 2001 to 2005, he was a Research Fellow at the University of Ballarat. Since 2006, he has been an Australian Research Council Research Fellow at the University of Ballarat. His main research interests are in the area of nonsmooth and global optimization and their applications in data mining, regression analysis.

Conny Clausen was born on June 8, 1980 in Flensburg, Germany. She received the degree in mathematics from Saarland University in 2005 and the Ph.D. degree in mathematics from Saarland University in 2008.

Since 2008, she has been working as an IT-Consultant at Beck *et al.* projects GmbH, Munich, Germany.

Michael Kohler was born on July 17, 1969 in Esslingen, Germany. He received degrees in computer science and mathematics from the University of Stuttgart, Germany, in 1995 and the Ph.D. degree in mathematics from the University of Stuttgart in 1997.

In 1998, he worked as a Visiting Scientist at the Stanford University, Stanford, CA. From 2005 to 2007, he was Professor of Applied Mathematics at the University of Saarbrücken, and since 2007, he has been Professor of Mathematical Statistics at the Technische Universität Darmstadt. He coauthored with L. Györfi, A. Krzyżak, and H. Walk the book *A Distribution-Free Theory of Nonparametric Regression* (New York: Springer, 2002). His main research interest are in the area of nonparametric statistics, especially curve estimation and applications in mathematical finance.