Universität Stuttgart

On the rate of convergence of a neural network regression estimate learned by gradient descent

Aline Braun
Michael Kohler
Harro Walk

Stuttgarter Mathematische Berichte

2019-003

mathematik

Fachbereich Mathematik
Fakultät Mathematik und Physik
Universität Stuttgart
Pfaffenwaldring 57
D-70 569 Stuttgart

E-Mail: preprints@mathematik.uni-stuttgart.de
WWW: http://www.mathematik.uni-stuttgart.de/preprints

ISSN **1613-8309**

On the rate of convergence of a neural network regression estimate learned by gradient descent *

Alina Braun[1], Michael Kohler[1,†] and Harro Walk[2]

[1] *Fachbereich Mathematik, Technische Universität Darmstadt, Schlossgartenstr. 7, 64289 Darmstadt, Germany, email: braun@mathematik.tu-darmstadt.de, kohler@mathematik.tu-darmstadt.de*

[2] *Fachbereich Mathematik, Universität Stuttgart, Pfaffenwaldring 57, 70569 Stuttgart, Germany, email: walk@mathematik.uni-stuttgart.de*

April 25, 2019

Abstract

Nonparametric regression with random design is considered. Estimates are defined by minimzing a penalized empirical L_2 risk over a suitably chosen class of neural networks with one hidden layer via gradient descent. Here, the gradient descent procedure is repeated several times with randomly chosen starting values for the weights, and from the list of constructed estimates the one with the minimal empirical L_2 risk is chosen. Under the assumption that the number of randomly chosen starting values and the number of steps for gradient descent are sufficiently large it is shown that the resulting estimate achieves (up to a logarithmic factor) the optimal rate of convergence in a projection pursuit model. The final sample size performance of the estimates is illustrated by using simulated data.

AMS classification: Primary 62G08; secondary 62G20.

Key words and phrases: gradient descent, neural networks, nonparametric regression, rate of convergence, projection pursuit.

1. Introduction

1.1. Scope of this article

Motivated by the huge success of multilayer neural networks in applications (see, e.g., Schmidhuber (2015) and the literature cited therein) there has been an increasing interest in the theoretical analysis of such estimates. Often this is done in the area of nonparametric regression, and recently there has been a tremendous progress in the theoretical understanding of least squares regression estimates based on deep neural networks, i.e., neural networks with many hidden layers. The corresponding theoretical results are based

*Running title: *Neural network regression estimates*

†Corresponding author. Tel: +49-6151-16-23382, Fax:+49-6151-16-23381

on the derivation of new approximation results for piecewise polynomials by neural networks, and they make extensive use of the network structure, which allows to exploit compository assumptions on the structure of the regression function in order to circumvent the curse of dimensionality (cf., Kohler and Krzyżak (2017), Bauer and Kohler (2017), Schmidt-Hieber (2017), Imaizumi and Fukumizu (2018), Kohler and Langer (2018), Eckle and Schmidt-Hieber (2018) and Kohler, Krzyżak and Langer (2019)).

In all the articles above the neural network regression estimate is defined as a nonlinear least squares estimate, i.e., as a function which minimizes the empirical L_2 risk over a nonlinear class of neural networks. In practice, it is usually not possible to find the global minimum of the empirical L_2 risk over a nonlinear class of neural networks and one usually tries to find a local minimum using, for instance, the steepest descent algorithm. So although the above theoretical results are quite impressive, there is a big gap between the estimates studied theoretically and the estimates used in practice.

The purpose of this paper is to narrow this gap. To do this, we consider the following question: If we define a neural network regression estimate theoretically exactly as it is implemented in practice, can we show a rate of convergence result for this estimate? The ultimative goal is to analyze theoretically neural network regression estimates which are actually used in practice. As a first step in this direction we define a simple neural network regression estimate where we use gradient descent in order to learn the weights of a neural network with one hidden layer in a projection pursuit model. We show that if we repeatedly apply this procedure to starting values, which are chosen randomly from a special structure, then, for sufficiently many starting values and steps in each procedure, we will find an estimate which achieves the optimal rate of convergence up to a logarithmic factor in this projection pursuit model.

1.2. Nonparametric regression

We study neural network estimates in the context of nonparametric regression with random design. Here, (X, Y) is an $\mathbb{R}^d \times \mathbb{R}$–valued random vector satisfying $\mathbf{E}\{Y^2\} < \infty$, and given a sample of (X, Y) of size n, i.e., given a data set

$$\mathcal{D}_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}, \tag{1}$$

where (X, Y), (X_1, Y_1), $\ldots$, (X_n, Y_n) are i.i.d. random variables, the aim is to construct an estimate

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

of the regression function $m : \mathbb{R}^d \to \mathbb{R}$, $m(x) = \mathbf{E}\{Y|X = x\}$ such that the L_2 error

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

is "small" (see, e.g., Györfi et al. (2002) for a systematic introduction to nonparametric regression and a motivation for the L_2 error).

It is well–known that one needs smoothness assumptions on the regression function in order to derive non–trivial results on the rate of convergence of nonparametric regression

estimates (cf., e.g., Theorem 7.2 and Problem 7.2 in Devroye, Györfi and Lugosi (1996) and Section 3 in Devroye and Wagner (1980)). To do this we will use the following definition.

Definition 1 *Let $p = q + s$ for some $q \in \mathbb{N}_0$ and $0 < s \leq 1$, where $\mathbb{N}_0$ is the set of nonnegative integers. A function $f : \mathbb{R}^d \to \mathbb{R}$ is called (p, C)-**smooth**, if for every $\alpha = (\alpha_1, \ldots, \alpha_d) \in \mathbb{N}_0^d$ with $\sum_{j=1}^d \alpha_j = q$ the partial derivative $\frac{\partial^q f}{\partial x_1^{\alpha_1} \ldots \partial x_d^{\alpha_d}}$ exists and satisfies*

$$\left| \frac{\partial^q f}{\partial x_1^{\alpha_1} \ldots \partial x_d^{\alpha_d}}(x) - \frac{\partial^q f}{\partial x_1^{\alpha_1} \ldots \partial x_d^{\alpha_d}}(z) \right| \leq C \cdot \|x - z\|^s$$

for all $x, z \in \mathbb{R}^d$, where $\| \cdot \|$ denotes the Euclidean norm.

Stone (1982) showed that the optimal minimax rate of convergence in nonparametric regression for (p, C)-smooth functions is $n^{-2p/(2p+d)}$. In case that d is large compared to p this rate of convergence is rather slow (so called curse of dimensionality). In the sequel we want to circumvent this curse of dimensionality by imposing the additional constraint on the regression function that it satisfies a projection pursuit model, i.e., by assuming that it satisfies

$$m(x) = \sum_{s=1}^{r} g_s(\mathbf{c}_s^T x) \quad (x \in \mathbb{R}^d) \tag{2}$$

for some $r \in \mathbb{N}$, $\mathbf{c}_s \in \mathbb{R}^d$ and (p, C)-smooth functions $g_s : \mathbb{R} \to \mathbb{R}$ $(s = 1, \ldots, r)$. Under this assumption our aim is to show that suitably defined neural network estimates, which can be actually implemented in an application, can achieve the one-dimensional rate of convergence.

1.3. Main result of this article

In this paper we study neural network regression estimates using neural networks with one hidden layer in the above projection pursuit model, i.e., we assume that the regression function satisfies (2). We learn the weights of our neural network regression estimate by choosing in a first step randomly vectors for the directions $\mathbf{c}_s$ of our projection pursuit model, by defining in a second step an appropriate starting value for the weights of our neural network regression estimate based on the randomly chosen directions, and by applying in a third step successively many gradient descent steps in order to optimze the weights of our neural network. Then we repeat this whole procedure several times and choose from the list of estimates which we get the one with the minimal error on our training data.

Our main result is that for a sufficiently large number of repititions of this procedure and a sufficiently large number of gradient descent steps the expected L_2 error of a truncated version of our estimate converges towards zero in the projection pursuit model (2) in case of (p, C)–smoth functions g_s (where $p \leq 1$) with the rate of convergence

$$\left(\frac{(\log n)^3}{n} \right)^{\frac{2p}{2p+1}},$$

i.e., with the optimal rate of convergence up to a logarithmic factor. Here, the rate of convergence is independent of the dimension d of X. Hence, our neural network regression estimate is able to circumvent the curse of dimensionality in the projection pursuit model (2).

We achieve this result by choosing our initial weights such that the initial network basically computes a piecewise constant function and by showing that in this case the gradient descent is able to choose the outer weights in the neural network in an optimal way (provided the number of gradient descent steps is sufficiently large).

1.4. Discussion of related results

It is well-known that it is possible to circumvent the curse of dimensionality by imposing additional constraints on the structure of the regression function. Stone (1985) assumed that the regression function is additive, i.e., that $m : \mathbb{R}^d \to \mathbb{R}$ satisfies

$$m(x^{(1)}, \dots, x^{(d)}) = m_1(x^{(1)}) + \cdots + m_d(x^{(d)}) \quad (x^{(1)}, \dots, x^{(d)} \in \mathbb{R})$$

for some (p, C)–smooth univariate functions $m_1, \dots, m_d : \mathbb{R} \to \mathbb{R}$, and showed that in this case suitably defined spline estimates achieve the corresponding univariate rate of convergence. Stone (1994) extended this results to interaction models, where the regression function is assumed to be a sum of functions applied to at most $d^* < d$ components of x and showed in this case that suitably defined spline estimates achieve the d^*–dimensional rate of convergence. Other classes of functions which enable us to achieve a better rate of convergence results include single index models, where

$$m(x) = g(\mathbf{c}^T x) \quad (x \in \mathbb{R}^d)$$

for some $\mathbf{c} \in \mathbb{R}^d$ and $g : \mathbb{R} \to \mathbb{R}$ (cf., e.g., Härdle and Stoker (1989), Härdle, Hall and Ichimura (1993), Yu and Ruppert (2002), Kong and Xia (2007) and Lepski and Serdyukova (2014)) and projection pursuit, where it is assumed that (2) holds for some $r \in \mathbb{N}$, $\mathbf{c}_s \in \mathbb{R}^d$ and $g_s : \mathbb{R} \to \mathbb{R}$ $(s = 1, \dots, r)$ (cf., e.g., Friedman and Stuetzle (1981) and Huber (1985)). Horowitz and Mammen (2007) studied the case of a regression function, which satisfies

$$m(x) = g\left(\sum_{l_1=1}^{L_1} g_{l_1}\left(\sum_{l_2=1}^{L_2} g_{l_1,l_2}\left(\cdots \sum_{l_r=1}^{L_r} g_{l_1,\dots,l_r}(x^{l_1,\dots,l_r})\right)\right)\right),$$

where $g, g_{l_1}, \dots, g_{l_1,\dots,l_r}$ are (p, C)-smooth univariate functions and $x^{l_1,\dots,l_r}$ are single components of $x \in \mathbb{R}^d$ (not necessarily different for two different indices $(l_1, \dots, l_r)$). With the use of a penalized least squares estimate for smoothing splines, they proved the rate $n^{-2p/(2p+1)}$.

For the L_2 error of a single hidden layer neural network, Barron (1993, 1994) proves the dimensionless rate of convergence $n^{-1/2}$ (up to some logarithmic factor), provided the Fourier transform has a finite first moment (which basically requires that the function becomes smoother with increasing dimension d of X). McCaffrey and Gallant (1994)

showed a rate of $n^{-\frac{2p}{2p+d+5}+\varepsilon}$ for the L_2 error of suitably defined single hidden layer neural network estimate for (p, C)-smooth functions, but their study was restricted to the use of a certain cosine squasher as the activation function.

Recently it was shown in several papers that neural networks can achieve a dimensionality reduction in case that the regression function is a composition of (sums of) functions, where each of the function is a function of at most $d^* < d$ variables. The first paper in this respect was Kohler and Krzyżak (2017), where it was shown that under a corresponding assumption suitably defined multilayer neural networks achieve the rate of convergence $n^{-2p/(2p+d^*)}$ (up to some logarithmic factor) in case $p \leq 1$. Bauer and Kohler (2017) showed that this result even holds for $p > 1$ provided the squashing function is suitably chosen. Schmidt-Hieber (2017) showed similar results for neural networks with ReLU activation function, and Kohler and Langer (2018) showed that the results of Bauer and Kohler (2017) also hold for very simply constructed fully connected feedforward neural networks. Eckle and Schmidt-Hieber (2018) showed that neural networks with ReLU activation function can approximate well piecewise polynomials with rather general partitions based on the intersection of hyperplanes and used this result to relate the error of neural network estimates to the error of piecewise polynomial partitioning estimates. Kohler, Krzyżak and Langer (2019) derived a similar result for neural networks with squashing functions as activation function and used this result to prove that neural networks are able to circumvent the curse of dimensionality in case that the regression function has a low local dimensionality. Results concerning the approximation of piecewise polynomials with partitions with rather general smooth boundaries by neural networks have been derived in Imaizumi and Fukamizu (2018).

The above mentioned results show that least squares neural network regression estimates are able to circumvent the curse of dimensionality under much more general assumptions than the projection pursuit model assumed in this paper. However, these estimates cannot be computed in practice, whereas our result shows that in the projection pursuit model we can achieve this with neural networks even in the case where we restrict ourselves to estimates which can be computed much easier.

Gradient descent has been studied in many different papers, see, e.g., Karimi, Nutini and Schmidt (2018) and the literature cited therein. A standard reference is the monograph Luenberger and Ye (2016). We also mention Poljak (1981) as an early paper, where the case of noise corrupted function values is considered, too. Stochastic approximation deals with the latter field, see, e.g., the monograph Kushner and Yin (2003), and here in a classic situation the constant factor at the gradient is replaced by a decreasing factor at a vector of divided differences (multidimensional Kiefer-Wolfowitz method). The paper of White (1989, 1992) brings together the two fields of stochastic approximation and neural network models (see also Fabian (1994)). In Dippon and Fabian (1994) and Dippon (1998) it is explained how gradient descent in stochastic approximation can be combined with a slowly convergent global optimizer in order to find not only a local but even a global minimum of a general function. The main difficulty of using such results to derive rate of convergence results for neural network regression estimates lies in the fact that for neural network regression estimates the neural network is using more and more

neurons with increasing sample size. This means that it is not sufficient to analyze gradient descent applied to a fixed function where the number of steps is tending to infinity. Instead the function is changing for increasing number of steps. Basically, this requires the ability to analyze the behaviour of gradient descent for a finite number of steps. As far as we know such results do not exist in the literature in case of a general function like the empirical L_2 risk of a neural network (which is neither convex nor has a global minimum or an easily analysable Hessian matrix considered as a real-valued function of the weight vector).

1.5. Notation

Throughout the paper, the following notation is used: The sets of natural numbers, natural numbers including 0 and real numbers are denoted by $\mathbb{N}$, $\mathbb{N}_0$ and $\mathbb{R}$, respectively. For $z \in \mathbb{R}$, we denote the smallest integer greater than or equal to z by $\lceil z \rceil$. The Euclidean norm of $x \in \mathbb{R}^d$ is denoted by $\|x\|$, and $\|x\|_\infty$ denotes its supremum norm. For $f : \mathbb{R}^d \to \mathbb{R}$ let

$$\|f\|_\infty = \sup_{x \in \mathbb{R}^d} |f(x)|$$

denote its supremum norm. A finite collection $f_1, \ldots, f_N : \mathbb{R}^d \to \mathbb{R}$ is called an ε–L_1–cover of $\mathcal{F}$ on $x_1^n = (x_1, \ldots, x_n) \in (\mathbb{R}^d)^n$ if for any $f \in \mathcal{F}$ there exists $i \in \{1, \ldots, N\}$ such that

$$\frac{1}{n} \sum_{j=1}^{n} |f(x_j) - f_i(x_j)| < \varepsilon.$$

The ε–L_1- covering number of $\mathcal{F}$ on x_1^n is the size N of the smallest ε–L_1– cover of $\mathcal{F}$ on x_1^n and is denoted by $\mathcal{N}_1(\varepsilon, \mathcal{F}, x_1^n)$.

1.6. Outline

The outline of this paper is as follows: In Section 2 we define our neural network regression estimates and in Section 3 we present our main theoretical result. The finite sample size performance of our newly proposed estimate is illustrated in Section 4 by applying it to simulated data. The proofs are given in Section 5.

2. Definition of the estimate

In the construction of our estimate we assume that the regression function m satisfies (2) and that the support of X is contained in the cube $[-A, A]^d$ for some given $A \geq 1$. We approximate each $g_s : \mathbb{R} \to \mathbb{R}$ by a neural network with logistic squasher

$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

chosen such that it is close to a piecewise constant function of the form

$$u \mapsto \sum_{l=1}^{K} a_{s,l} \cdot 1_{[b_l, \infty)}.$$

As we will show in Lemma 5 below, such a neural network can be chosen of the form

$$u \mapsto \sum_{l=1}^{K} a_{s,l} \cdot \sigma(\rho_n \cdot (u - b_l)),$$

where $\rho_n > 0$ is a large constant, and the error of this approximation will be small at all those points, where $\rho_n \cdot |u - b_l|$ is large. By replacing u with $\mathbf{c}_s^T x$ we see that we can approximate m by networks with one hidden layer and $K \cdot r$ neurons in this hidden layer defined by

$$f_{net,(\mathbf{a},\mathbf{b})}(x) = \sum_{k=1}^{K \cdot r} a_k \cdot \sigma \left(\sum_{j=1}^{d} b_{k,j} \cdot x^{(j)} + b_{k,0} \right) + a_0. \tag{3}$$

Here, $K \cdot r \in \mathbb{N}$ is the number of neurons, $\sigma : \mathbb{R} \to \mathbb{R}$ is the activation function and

$$a_k \in \mathbb{R} \quad (k = 0, \ldots, K \cdot r) \quad \text{and} \quad b_{k,j} \in \mathbb{R} \quad (k = 1, \ldots, K \cdot r, j = 0, \ldots, d)$$

are the weights. The above condition that $\rho_n \cdot |u - b_l|$ is large in order to achieve a small error at point u of the above neural network approximation of the piecewise constant function is replaced by the assumption that

$$\min_{i=1,\ldots,n} |\sum_{j=1}^{d} b_{k,j} \cdot X_i^{(j)} + b_{k,0}|$$

is large, which will enable us to show that our approximation is good at all x-values of the data points. And this condition in turn will be ensured by a proper choice of the initial weights described below.

We will learn the weights by gradient descent. More precisely, we minimize the penalized empirical L_2 risk

$$F(\mathbf{a}, \mathbf{b}) = \frac{1}{n} \sum_{i=1}^{n} |f_{net,(\mathbf{a},\mathbf{b})}(X_i) - Y_i|^2 + \frac{c_1}{n} \cdot \sum_{k=0}^{K \cdot r} a_k^2, \tag{4}$$

where $c_1 > 0$ is a constant, by choosing an appropriate starting value $(\mathbf{a}^{(0)}, \mathbf{b}^{(0)})$ and by setting

$$\begin{pmatrix} \mathbf{a}^{(t+1)} \\ \mathbf{b}^{(t+1)} \end{pmatrix} = \begin{pmatrix} \mathbf{a}^{(t)} \\ \mathbf{b}^{(t)} \end{pmatrix} - \lambda_n \cdot (\nabla_{(\mathbf{a},\mathbf{b})} F)(\mathbf{a}^{(t)}, \mathbf{b}^{(t)}) \tag{5}$$

for some $\lambda_n > 0$ chosen below and $t = 0, 1, \ldots, t_n - 1$.

Next, we explain how we choose the initial values $(\mathbf{a}^{(0)}, \mathbf{b}^{(0)})$ for our weights. As explained above, our choice is motivated by the structure of m in the projection pursuit model (2). Here the number r of terms in this model is a parameter of our estimate (which we will choose data-dependent in any application, cf., Remark 2 below). In a first step we randomly choose values

$$\bar{\mathbf{c}}_1, \ldots, \bar{\mathbf{c}}_r \in [-1, 1]^d \tag{6}$$

as an independent sample from a uniform distribution on $[-1,1]^d$. Using these values as approximation of the directions $\mathbf{c}_1, \dots, \mathbf{c}_r$ of our projection pursuit model, we define our initial inner weights as follows: For $s \in \{1, \dots, r\}$ we define

$$b_{(s-1)\cdot K+1,0}, \dots, b_{(s-1)\cdot K+1,d}, \dots, b_{s\cdot K,0}, \dots, b_{s\cdot K,d}$$

according to $\mathbf{c}_s$: First we choose $b_1, \dots, b_K \in \mathbb{R}$ such that

$$b_1 \leq -A \cdot \sqrt{d},$$

$$b_K \geq A \cdot \sqrt{d} - \frac{4 \cdot \sqrt{d} \cdot A}{K-1},$$

$$|b_{k+1} - b_k| \leq \frac{4 \cdot \sqrt{d} \cdot A}{K-1} \quad (k = 1, \dots, K-1)$$

and

$$\min_{i=1,\dots,n, k=1,\dots,K} \left| \bar{\mathbf{c}}_s^T X_i - b_k \right| \geq \frac{\sqrt{d} \cdot A}{(n+1) \cdot (K-1)}.$$

Such a choice is always possible, e.g., we can set $b_1 = -\sqrt{d} \cdot A - 2 \cdot \sqrt{d} \cdot A / ((n+1) \cdot (K-1))$ and define b_k $(k = 2, \dots, K)$ by subdividing the interval

$$\left[-\sqrt{d} \cdot A + (k-2) \cdot \frac{2 \cdot \sqrt{d} \cdot A}{K-1}, -d \cdot A + (k-1) \cdot \frac{2 \cdot \sqrt{d} \cdot A}{K-1} \right]$$

into $(n+1)$ equidistant subintervals of length $2 \cdot \sqrt{d} \cdot A / ((K-1) \cdot (n+1))$ and by choosing b_k as the midpoint of one of those intervals which does not contain any of the n values $\mathbf{c}_s^T X_i$ (such an interval must exists since not every one of the $n+1$ disjoint intervals can contain one of the above n points). As soon as we have chosen $b_1, \dots, b_K$ we define $b_{(s-1)\cdot K+k,j}$ $(s = 1, \dots, r,\ k = 1, \dots, K,\ j = 0, \dots, d)$ such that we have for some $\rho_n > 0$ chosen below (cf., Theorem 1 below)

$$\sum_{j=1}^{d} b_{(s-1)\cdot K+k,j} \cdot x^{(j)} + b_{(s-1)\cdot K+k,0} = \rho_n \cdot (\bar{\mathbf{c}}_s^T x - b_k) \quad \text{for all } x \in \mathbb{R}^d,$$

namely, we set

$$b_{(s-1)\cdot K+k,j} = \rho_n \cdot \bar{c}_s^{(j)} \quad \text{and} \quad b_{(s-1)\cdot K+k,0} = -\rho_n \cdot b_k$$

$(s = 1, \dots, r,\ k = 1, \dots, K,\ j = 0, \dots, d)$. Then we choose $a_l = 0$ for all $l \in \{0, \dots, K \cdot r\}$.

After this choice of $(\mathbf{a}^{(0)}, \mathbf{b}^{(0)})$ we define $(\mathbf{a}^{(t+1)}, \mathbf{b}^{(t+1)})$ recursively by (5) for $\lambda_n > 0$ and $t = 0, 1, \dots, t_n - 1$.

We repeat this whole procedure L_n times, and let

$$\tilde{m}_n$$

be the neural network which achieves the smallest penalized empirical L_2 error (4) among all the t_n networks. Finally we truncate our estimate by selecting some $\beta_n > 0$ and by setting

$$m_n(x) = T_{\beta_n} \tilde{m}_n(x),$$

where $T_{\beta_n} z = \max\{\min\{z, \beta_n\}, -\beta_n\}$ for $z \in \mathbb{R}$.

3. Main result

Our main result is the following theorem.

Theorem 1 *Let $n \geq 2$, let $A \geq 1$ and let (X,Y), (X_1,Y_1), ..., (X_n,Y_n) be independent and identically distributed random variables with values in $[-A,A]^d \times \mathbb{R}$. Set $m(x) = \mathbf{E}\{Y|X=x\}$ and assume that (X,Y) satisfies*

$$\mathbf{E}\left(e^{c_2 \cdot |Y|^2}\right) < \infty \tag{7}$$

for some constant $c_2 > 0$, and that m satisfies

$$m(x) = \sum_{s=1}^{r} g_s(\mathbf{c}_s^T x) \quad (x \in \mathbb{R}^d)$$

for some $r \in \mathbb{N}$, $\mathbf{c}_s \in [-1,1]^d$ and $g_s : \mathbb{R} \to \mathbb{R}$ $(s = 1,\dots,r)$. Assume that g_s is (p,C)-smooth for $s \in \{1,\dots,r\}$, where $p \in (0,1]$ and $C > 0$ are fixed. Define the regression estimate m_n as in Section 2 with

$$\sigma(x) = \frac{1}{1+e^{-x}},$$

with parameter r as in the above projection pursuit model, and with the other parameters chosen by

$$\beta_n = c_3 \cdot \log n, \quad K = K_n = \lceil (n/(\log n)^3)^{1/(2p+1)}] \rceil, \quad \lambda_n = \frac{1}{3 \cdot K}, \quad \rho_n = n^2 \cdot K,$$

and

$$t_n = K_n \cdot n \cdot (\log n)^2 \quad \text{and} \quad L_n = \lceil (\log n) \cdot n^{r \cdot d/(2p+1)} \rceil.$$

Then m_n satisfies

$$\mathbf{E} \int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx) \leq c_4 \cdot \left(\frac{(\log n)^3}{n} \right)^{\frac{2p}{2p+1}}$$

for some constant $c_4 > 0$ which does not depend on n.

Remark 1. According to Stone (1982) the rate of convergence in the above theorem is optimal up to a logarithmic factor in case of a (p,C)-smooth projection pursuit model. Because of the fact that this rate of convergence is independent of the dimension d of X, the above theorem shows that our newly proposed computable neural network regression estimate is able to circumvent the curse of dimensionality in case that the regression function satisfies the assumption of projection pursuit. We should however mention that the number of repitions L_n of the initial random choices of the directions $\mathbf{c}_s$ and correspondingly the number of repititions of the t_n gradient descent steps is rather huge.
Remark 2. The parameters r and K_n of the above algorithm depend on the projection pursuit model and hence are unknown in any application. However, it is easy to choose them data-dependent by using, e.g., the splitting of the sample technique es explained in the next section. In this way it is possible to define an estimate which does not depend on the value of r of the projection pursuit model and which is nevertheless able to achieve the rate of convergence in Theorem 1.

4. Application to simulated data

In this section we illustrate the finite sample size performance of our newly proposed estimate by applying it to simulated data.

The simulated data which we use is defined as follows: We choose $d = 4$, X uniformly distributed on $[-1, 1]^d$, ϵ standard normal and independent of X, and we define Y by

$$Y = m_j(X) + \sigma \cdot \lambda_j \cdot \epsilon, \tag{8}$$

where $m_j : [-1, 1]^d \to \mathbb{R}$ is described below, $\lambda_j > 0$ is a scaling value defined below and σ is chosen from $\{0.05, 0.2\}$ ($j \in \{1, 2\}$). As regression function we use

$$m_1(x_1, x_2, x_3, x_4) = 2 \cdot \sin\left(\frac{2 \cdot \pi}{\sqrt{4}} \cdot (-x_1 + x_2 - x_3 + x_4)\right),$$

so m_1 satisfies a single index model, and

$$\begin{aligned} &m_2(x_1, x_2, x_3, x_4) \\ &= 4 \cdot \sin\left(\frac{2 \cdot \pi}{\sqrt{4}} \cdot (-x_1 + x_2 - x_3 + x_4)\right) + \frac{7}{2 + \frac{1}{\sqrt{30}} \cdot (x_1 - 2 \cdot x_2 + 3 \cdot x_3 - 4 \cdot x_4)}, \end{aligned}$$

hence m_2 satisfies a single index model with $r = 2$ terms. λ_j is chosen approximately as IQR of a sample of size $100,000$ of $m(X)$, and we use the concrete values $\lambda_1 = 2.8289$ and $\lambda_2 = 5.2841$. From this distribution we generate a sample of size $n = 100$ and apply our newly proposed neural network regression estimate and two alternative regression estimates to this sample. Then we compute the L_2 errors of these three estimate approximately by using the empirical L_2 error $\varepsilon_{L_2,\bar{N}}(\cdot)$ on an independent sample of X of size $\bar{N} = 10,000$. Since this error strongly depends on the behavior of the correct function m_j, we consider it in relation to the error of the simplest estimate for m_j we can think of, a completely constant function (whose value is the average of the observed data according to the least squares approach). Thus, the scaled error measure we use for evaluation of the estimates is $\varepsilon_{L_2,\bar{N}}(m_{n,i})/\bar{\varepsilon}_{L_2,\bar{N}}(avg)$, where $\bar{\varepsilon}_{L_2,\bar{N}}(avg)$ is the median of 50 independent realizations of the value one obtains if one plugs the average of n observations into $\varepsilon_{L_2,\bar{N}}(\cdot)$. To a certain extent, this quotient can be interpreted as the relative part of the error of the constant estimate that is still contained in the more sophisticated approaches. The resulting scaled errors of course depend on the random sample of (X, Y), and to be able to compare these values nevertheless we repeat the whole computation 20 times and report the median and the interquartile range of the 20 scaled errors for each of our three estimates.

Our first estimate *Tps* is a smoothing spline estimate with parameter chosen by generalized cross validation as implemented in the routine *Tps()* of the library *fields* in *R*.

Our second estimate *neighbor* is a nearest neighbor estimate where the number of nearest neighbors is chosen from the set $\{1, 2, 4, 8, 16, 32\}$ by splitting of the sample. Here we split our sample in a learning sample of size $n_l = 0.8 \cdot n$ and a testing sample of size $n_t = 0.2 \cdot n$. We compute the estimate for all parameter values from the above set

using the learning sample, compute the corresponding empirical L_2 risk on the testing sample and choose the parameter value which leads to the minimal empirical L_2 risk on the testing sample.

Our third estimate *neural* is our newly proposed neural network estimate presented in this paper, which we have implemented in *R*. Here the parameters r and K of the estimate are chosen via splitting of the sample (as described above) from the set $\{1, 2\}$ and $\{5, 10, 20\}$, respectively. In order to accelerate the computation of this estimate we use only $L_n = 50$ random choices for the vectors of directions in the computation of the estimate for each parameter value.

The results are summarized in Table 1. As we can see from the reported scaled errors, our newly proposed neural network estimate outperforms in all four settings both the smoothing spline estimate and the nearest neighbor estimate.

	m_1		m_2	
noise	5%	20%	5%	20%
$\bar{\varepsilon}_{L_2,\bar{N}}(avg)$	2.0109	2.0157	10.3565	11.1144
approach	median (IQR)	median (IQR)	median (IQR)	median (IQR)
Tps	1.18 (0.52)	1.19 (0.13)	0.89 (0.08)	0.91 (0.14)
neighbor	1.06 (0.14)	1.13 (0.26)	0.91 (0.06)	0.86 (0.07)
neural	0.52 (0.33)	0.46 (0.24)	0.42 (0.14)	0.53 (0.13)

Table 1: Median and IQR of the scaled empirical L_2 error of estimates for m_1 and m_2 for sample size $n = 100$.

5. Proofs

5.1. Learning of linear penalized least squares estimates by gradient descent

Let $(x_1, y_1), \ldots, (x_n, y_n) \in \mathbb{R}^d \times \mathbb{R}$, let $K \in \mathbb{N}$, let $B_1, \ldots, B_K : \mathbb{R}^d \to \mathbb{R}$ and let $c_1 > 0$. In this subsection we consider the problem to minimize

$$F(\mathbf{a}) = \frac{1}{n} \sum_{i=1}^{n} |\sum_{k=1}^{K} a_k \cdot B_k(x_i) - y_i|^2 + \frac{c_1}{n} \cdot \|a\|^2, \tag{9}$$

where

$$\mathbf{a} = (a_1, \ldots, a_K)^T \quad \text{and} \quad \|a\|^2 = \sum_{j=1}^{K} a_j^2,$$

by gradient descent. To do this, we choose $\mathbf{a}^{(0)} \in \mathbb{R}^K$ and set

$$\mathbf{a}^{(t+1)} = \mathbf{a}^{(t)} - \lambda_n \cdot (\nabla_{\mathbf{a}} F)(\mathbf{a}^{(t)}) \tag{10}$$

for some properly chosen $\lambda_n > 0$.

Lemma 1 *Let $F : \mathbb{R}^K \to \mathbb{R}$ be differentiable and define $\mathbf{a}^{(t+1)}$ by (10), where*

$$\lambda_n = \frac{1}{L_n} \tag{11}$$

for some $L_n > 0$. Let $\mathbf{a}_{opt} \in \mathbb{R}^K$ be arbitrary.
a) *If*

$$\|(\nabla_{\mathbf{a}} F)(\mathbf{a}_1) - (\nabla_{\mathbf{a}} F)(\mathbf{a}_2)\| \le L_n \cdot \|\mathbf{a}_1 - \mathbf{a}_2\| \quad (\mathbf{a}_1, \mathbf{a}_2 \in \mathbb{R}^K) \tag{12}$$

holds, then we have

$$F(\mathbf{a}^{(t+1)}) - F(\mathbf{a}^{(t)}) \le -\frac{1}{2 \cdot L_n} \cdot \|(\nabla_{\mathbf{a}} F)(\mathbf{a}^{(t)})\|^2.$$

b) *If inequality (12) and, in addition,*

$$\|(\nabla_{\mathbf{a}} F)(\mathbf{a})\|^2 \ge \rho_n \cdot (F(\mathbf{a}) - F(\mathbf{a}_{opt})) \quad (\mathbf{a} \in \mathbb{R}^K) \tag{13}$$

hold, then we have

$$F(\mathbf{a}^{(t+1)}) - F(\mathbf{a}_{opt}) \le \left(1 - \frac{\rho_n}{2 \cdot L_n}\right) \cdot (F(\mathbf{a}^{(t)}) - F(\mathbf{a}_{opt})).$$

Proof. Lemma 1 follows from well-known bounds in the literature, see, e.g., Karimi, Nutini and Schmidt (2018). For the sake of completeness a complete proof is given in the supplementary material. □

Lemma 2 *Let F be defined by (9). Then we have for any $\mathbf{a}_1, \mathbf{a}_2 \in \mathbb{R}^K$*

$$\|(\nabla_{\mathbf{a}} F)(\mathbf{a}_1) - (\nabla_{\mathbf{a}} F)(\mathbf{a}_2)\| \le \left(2 \cdot \sum_{k=1}^{K} \frac{1}{n} \sum_{i=1}^{n} B_k(x_i)^2 + \frac{2 \cdot c_1}{n}\right) \cdot \|\mathbf{a}_1 - \mathbf{a}_2\|.$$

Proof. We have

$$F(\mathbf{a}) = \frac{1}{n} \cdot (\mathbf{B} \cdot \mathbf{a} - \mathbf{y})^T \cdot (\mathbf{B} \cdot \mathbf{a} - \mathbf{y}) + \frac{c_1}{n} \cdot \mathbf{a}^T \cdot \mathbf{a}$$

where

$$\mathbf{B} = (\mathcal{B}_j(x_i))_{1 \le i \le n, 1 \le j \le K} \quad \text{and} \quad \mathbf{y} = (y_1, \ldots, y_n)^T.$$

Consequently,

$$(\nabla_{\mathbf{a}} F)(\mathbf{a}) = \frac{2}{n} \cdot \left(\mathbf{B}^T \mathbf{B} \mathbf{a} - \mathbf{B}^T \mathbf{y}\right) + \frac{2 \cdot c_1}{n} \cdot \mathbf{a}$$

and

$$\|(\nabla_{\mathbf{a}} F)(\mathbf{a}_1) - (\nabla_{\mathbf{a}} F)(\mathbf{a}_2)\| \le \|\frac{2}{n} \cdot \mathbf{B}^T \mathbf{B} \cdot (\mathbf{a}_1 - \mathbf{a}_2)\| + \frac{2 \cdot c_1}{n} \cdot \|\mathbf{a}_1 - \mathbf{a}_2\|.$$

By applying twice the inequality of Cauchy–Schwarz we get

$$\left\|\frac{2}{n} \cdot \mathbf{B}^T \mathbf{B} \cdot \mathbf{a}\right\|^2 = \sum_{j=1}^{K} \left(\sum_{k=1}^{K} (\frac{2}{n} \sum_{i=1}^{n} B_j(x_i) \cdot B_k(x_i)) \cdot a_k\right)^2$$

$$
\begin{aligned}
&\leq \sum_{j=1}^{K}\sum_{k=1}^{K}(\frac{2}{n}\sum_{i=1}^{n}B_j(x_i)\cdot B_k(x_i))^2\cdot\|\mathbf{a}\|^2\\
&\leq \sum_{j=1}^{K}\sum_{k=1}^{K}4\cdot\frac{1}{n}\sum_{i=1}^{n}B_j(x_i)^2\cdot\frac{1}{n}\sum_{i=1}^{n}B_k(x_i))^2\cdot\|\mathbf{a}\|^2\\
&= \left(2\cdot\sum_{k=1}^{K}\frac{1}{n}\sum_{i=1}^{n}B_k(x_i)^2\right)^2\cdot\|\mathbf{a}\|^2,
\end{aligned}
$$

which implies the assertion. □

Lemma 3 *Let F be defined by (9) and choose $\mathbf{a}_{opt}$ such that*

$$F(\mathbf{a}_{opt}) = \min_{\mathbf{a}\in\mathbb{R}^K} F(\mathbf{a}).$$

Then for any $\mathbf{a}\in\mathbb{R}^K$ we have

$$\|(\nabla_{\mathbf{a}}F)(\mathbf{a})\|^2 \geq \frac{c_1}{n}\cdot(F(\mathbf{a})-F(\mathbf{a}_{opt})).$$

Proof. Set

$$\mathbf{B} = (\mathcal{B}_j(x_i))_{1\leq i\leq n, 1\leq j\leq K} \quad \text{and} \quad \mathbf{A} = \frac{1}{n}\cdot\mathbf{B}^T\cdot\mathbf{B}+\frac{c_1}{n}\cdot\mathbf{1},$$

where $\mathbf{1}$ is the unit matrix. Then $\mathbf{A}$ is positive definite and hence regular, from which we can conlcude

$$
\begin{aligned}
F(\mathbf{a}) &= \frac{1}{n}\cdot(\mathbf{B}\cdot\mathbf{a}-\mathbf{y})^T\cdot(\mathbf{B}\cdot\mathbf{a}-\mathbf{y})+\frac{c_1}{n}\cdot\mathbf{a}^T\cdot\mathbf{a}\\
&= \mathbf{a}^T\mathbf{A}\mathbf{a}-2\mathbf{y}^T\frac{1}{n}\mathbf{B}\mathbf{a}+\frac{1}{n}\mathbf{y}^T\mathbf{y}\\
&= (\mathbf{a}-\mathbf{A}^{-1}\frac{1}{n}\mathbf{B}^T\mathbf{y})^T\mathbf{A}(\mathbf{a}-\mathbf{A}^{-1}\frac{1}{n}\mathbf{B}^T\mathbf{y})+F(\mathbf{a}_{opt}),
\end{aligned}
$$

where

$$F(\mathbf{a}_{opt}) = \frac{1}{n}\mathbf{y}^T\mathbf{y}-\mathbf{y}^T\cdot\frac{1}{n}\cdot\mathbf{B}\mathbf{A}^{-1}\cdot\frac{1}{n}\cdot\mathbf{B}^T\mathbf{y}.$$

Using

$$\mathbf{b}^T\mathbf{A}\mathbf{b}\geq\frac{c_1}{n}\cdot\mathbf{b}^T\mathbf{b}$$

and $\mathbf{A}^T=\mathbf{A}$ we conclude

$$
\begin{aligned}
&F(\mathbf{a})-F(\mathbf{a}_{opt})\\
&= ((\mathbf{A}^{1/2})^T(\mathbf{a}-\mathbf{A}^{-1}\frac{1}{n}\mathbf{B}^T\mathbf{y}))^T\mathbf{A}^{1/2}(\mathbf{a}-\mathbf{A}^{-1}\frac{1}{n}\mathbf{B}^T\mathbf{y})\\
&\leq \frac{n}{c_1}\cdot((\mathbf{A}^{1/2})^T(\mathbf{a}-\mathbf{A}^{-1}\frac{1}{n}\mathbf{B}^T\mathbf{y}))^T\mathbf{A}\mathbf{A}^{1/2}(\mathbf{a}-\mathbf{A}^{-1}\frac{1}{n}\mathbf{B}^T\mathbf{y})
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{n}{c_1} \cdot ((\mathbf{A})^T(\mathbf{a} - \mathbf{A}^{-1}\frac{1}{n}\mathbf{B}^T\mathbf{y}))^T\mathbf{A}(\mathbf{a} - \mathbf{A}^{-1}\frac{1}{n}\mathbf{B}^T\mathbf{y}) \\
&= \frac{n}{c_1} \cdot ((\mathbf{A}\mathbf{a} - \frac{1}{n}\mathbf{B}^T\mathbf{y}))^T(\mathbf{A}\mathbf{a} - \frac{1}{n}\mathbf{B}^T\mathbf{y}) \\
&= \frac{n}{4 \cdot c_1} \cdot ((2\mathbf{A}\mathbf{a} - \frac{2}{n}\mathbf{B}^T\mathbf{y}))^T(2\mathbf{A}\mathbf{a} - \frac{2}{n}\mathbf{B}^T\mathbf{y}) \\
&= \frac{n}{4 \cdot c_1} \cdot \|(\nabla_{\mathbf{a}} F)(\mathbf{a})\|^2,
\end{aligned}
$$

where the last equality follows from

$$
(\nabla_{\mathbf{a}} F)(\mathbf{a}) = \nabla_{\mathbf{a}} \left(\mathbf{a}^T\mathbf{A}\mathbf{a} - 2\mathbf{y}^T\frac{1}{n}\mathbf{B}\mathbf{a} + \frac{1}{n}\mathbf{y}^T\mathbf{y} \right) = 2\mathbf{A}\mathbf{a} - \frac{2}{n}\mathbf{B}^T\mathbf{y}.
$$

□

5.2. Result for neural networks with one hidden layer

In this subsection we study neural networks with one hidden layer, which are defined by

$$
f_{net,(\mathbf{a},\mathbf{b})}(x) = \sum_{k=1}^{K} a_k \cdot \sigma \left(\sum_{j=1}^{d} b_{k,j} \cdot x^{(j)} + b_{k,0} \right) + a_0 \tag{14}
$$

(compare (3)), where $K \in \mathbb{N}$ is the number of neurons, $\sigma : \mathbb{R} \to \mathbb{R}$ is the activation function and where the weights

$$
a_k \quad (k = 0, \ldots, K) \quad \text{and} \quad b_{k,j} \in \mathbb{R} \quad (k = 1, \ldots, K, j = 0, \ldots, d)
$$

are learned by gradient descent. More precisely, we minimize

$$
F(\mathbf{a}, \mathbf{b}) = \frac{1}{n} \sum_{i=1}^{n} |f_{net,(\mathbf{a},\mathbf{b})}(x_i) - y_i|^2 + \frac{c_1}{n} \cdot \sum_{k=0}^{K} a_k^2 \tag{15}
$$

(compare (4)) by choosing an appropriate starting value $(\mathbf{a}^{(0)}, \mathbf{b}^{(0)})$ and by setting

$$
\begin{pmatrix} \mathbf{a}^{(t+1)} \\ \mathbf{b}^{(t+1)} \end{pmatrix} = \begin{pmatrix} \mathbf{a}^{(t)} \\ \mathbf{b}^{(t)} \end{pmatrix} - \lambda_n \cdot (\nabla_{(\mathbf{a},\mathbf{b})} F)(\mathbf{a}^{(t)}, \mathbf{b}^{(t)}) \tag{16}
$$

for some $\lambda_n > 0$ chosen below.

Our main idea is, that in the case of the logistic squasher

$$
\sigma(x) = \frac{1}{1 + e^{-x}} \quad (x \in \mathbb{R}),
$$

the neural network (14) is for appropriate weigths $b_{k,j}$ close to a linear combination of indicator functions, and in this case the gradient descent will change the inner weights $b_{k,j}$ only slightly. From this we will conclude from our results for linear least squares

estimates that for such networks the gradient descent leads to estimates where the outer weights a_k are chosen optimally.

In Lemma 5 below we study the approximation of Hölder continuous functions by neural networks of the above form in the case of univariate functions and networks. To do this, we will need the following auxiliary result.

Lemma 4 *Let σ be the logistic squasher.*
a) *For any $x \in \mathbb{R}$ we have*

$$|\sigma(x) - 1_{[0,\infty)}(x)| \leq e^{-|x|}.$$

b) *For any $b \in \mathbb{R}$, $c > 0$ and $x \in \mathbb{R}$ we have*

$$|\sigma(c \cdot (x-b)) - 1_{[b,\infty)}(x)| \leq e^{-c \cdot |x-b|}.$$

Proof. a) For $x \geq 0$ we have

$$|\sigma(x) - 1_{[0,\infty)}(x)| = 1 - \frac{1}{1+e^{-x}} = \frac{e^{-x}}{1+e^{-x}} \leq e^{-x} = e^{-|x|}.$$

And for $x < 0$ we get

$$|\sigma(x) - 1_{[0,\infty)}(x)| = \frac{1}{1+e^{-x}} \leq e^{x} = e^{-|x|}.$$

b) From $c > 0$ and a) we get

$$|\sigma(c \cdot (x-b)) - 1_{[b,\infty)}(x)| = |\sigma(c \cdot (x-b)) - 1_{[0,\infty)}(c \cdot (x-b))| \leq e^{-|c \cdot (x-b)|} = e^{-c \cdot |x-b|}.$$

□

Lemma 5 *Let σ be the logistic squasher. Let $a, b \in \mathbb{R}$ with $a < b$ and let $m : [a,b] \to \mathbb{R}$ be (p,C)-smooth for some $p \in [0,1]$ and $C > 0$. Let $c > 0$, $K \in \mathbb{N}$ and set*

$$b_k = a + k \cdot \frac{b-a}{K} \quad (k = 0, \ldots, K-1).$$

Let

$$a_0 = m(a) \quad \textit{and} \quad a_k = m(b_k) - m(b_{k-1}) \quad (k = 1, \ldots, K-1).$$

Then we have

$$\sup_{x \in [a,b]} \left| a_0 + \sum_{k=1}^{K-1} a_k \cdot \sigma(c \cdot (x - b_k)) - m(x) \right| \leq \frac{3 \cdot (b-a)^p \cdot C}{K^p} + C \cdot (b-a)^p \cdot K^{1-p} \cdot e^{-\frac{c \cdot (b-a)}{K}}.$$

Proof. We have

$$\left| a_0 + \sum_{k=1}^{K-1} a_k \cdot \sigma(c \cdot (x - b_k)) - m(x) \right|$$

$$\leq \left| \sum_{k=1}^{K-1} a_k \cdot \sigma(c \cdot (x - b_k)) - \sum_{k=1}^{K-1} a_k \cdot 1_{[b_k,\infty)}(x) \right| + \left| a_0 + \sum_{k=1}^{K-1} a_k \cdot 1_{[b_k,\infty)}(x) - m(x) \right|.$$

For $b_k \leq x < b_{k+1}$ we can conclude from the definition of a_k and the (p, C)-smoothness of m

$$\begin{aligned} &\left| a_0 + \sum_{k=1}^{K-1} a_k \cdot 1_{[b_k,\infty)}(x) - m(x) \right| \\ &= \left| a_0 + \sum_{j=1}^{k} a_j - m(x) \right| = |m(b_k) - m(x)| \leq C \cdot |b_k - x|^p \leq \frac{C \cdot (b-a)^p}{K^p}. \end{aligned}$$

It is easy to see that this inequality is also true for $x = b$, hence we have shown

$$\sup_{x \in [a,b]} \left| a_0 + \sum_{k=1}^{K-1} a_k \cdot 1_{[b_k,\infty)}(x) - m(x) \right| \leq \frac{C \cdot (b-a)^p}{K^p}.$$

We finish the proof by showing

$$\begin{aligned} &\sup_{x \in [a,b]} \left| \sum_{k=1}^{K-1} a_k \cdot \sigma(c \cdot (x - b_k)) - \sum_{k=1}^{K-1} a_k \cdot 1_{[b_k,\infty)}(x) \right| \\ &\leq \frac{2 \cdot (b-a)^p \cdot C}{K^p} + C \cdot (b-a)^p \cdot K^{1-p} \cdot e^{-\frac{c \cdot (b-a)}{K}}. \end{aligned}$$

For $b_k \leq x \leq b_{k+1}$ we have

$$\begin{aligned} &\left| \sum_{k=1}^{K-1} a_k \cdot \sigma(c \cdot (x - b_k)) - \sum_{k=1}^{K-1} a_k \cdot 1_{[b_k,\infty)}(x) \right| \\ &\leq \sum_{j=1}^{k-1} |a_j| \cdot \left| \sigma(c \cdot (x - b_j)) - 1_{[b_j,\infty)}(x) \right| + |a_k| + |a_{k+1}| \\ &\quad + \sum_{j=k+2}^{K-1} |a_j| \cdot \left| \sigma(c \cdot (x - b_j)) - 1_{[b_j,\infty)}(x) \right| \\ &\leq \max_{j=1,\ldots,K-1} |a_j| \cdot \left(2 + (K-3) \cdot \max_{j \in \{1,2,\ldots,k-1,k+2,k+3,\ldots,K-1\}} \left| \sigma(c \cdot (x - b_j)) - 1_{[b_j,\infty)}(x) \right| \right). \end{aligned}$$

Using the (p, C)-smoothness of m, which implies

$$|a_j| \leq C \cdot \frac{(b-a)^p}{K^p},$$

together with Lemma 4 we get for $b_k \leq x \leq b_{k+1}$

$$\left| \sum_{k=1}^{K-1} a_k \cdot \sigma(c \cdot (x - b_k)) - \sum_{k=1}^{K-1} a_k \cdot 1_{[b_k,\infty)}(x) \right|$$

$$\leq C \cdot \frac{(b-a)^p}{K^p} \cdot (2 + (K-3) \cdot \max_{j \in \{1,2,\ldots,k-1,k+2,k+3,\ldots,K-1\}} e^{-c \cdot |x - b_j|})$$
$$\leq \frac{2(b-a)^p \cdot C}{K^p} + C \cdot (b-a)^p \cdot K^{1-p} \cdot e^{-\frac{c \cdot (b-a)}{K}}.$$

□

Lemma 6 *Let σ be the logistic squasher. Define F by (15) and set*

$$\bar{\mathbf{b}} = \mathbf{b} - \lambda_n \cdot (\nabla_{\mathbf{b}} F)(\mathbf{a}, \mathbf{b})$$

for some $\lambda_n > 0$. Then we have for any $k \in \{1, \ldots, K\}$ and any $j \in \{0, \ldots, d\}$:

$$|\bar{b}_{k,j} - b_{k,j}| \leq \lambda_n \cdot 2 \cdot \sqrt{F(\mathbf{a}, \mathbf{b})} \cdot \max\{1, \max_{i,l}\{|x_i^{(l)}|\}\} \cdot |a_k| \cdot \exp\left(- \min_{i=1,\ldots,n} \left\{ \left| \sum_{j=1}^{d} b_{k,j} \cdot x_i^{(j)} + b_{k,0} \right| \right\} \right).$$

Proof. Using

$$|\sigma'(x)| = |\sigma(x) \cdot (1 - \sigma(x))| \leq \min\{|\sigma(x)|, |1 - \sigma(x)|\} \leq |\sigma(x) - 1_{[0,\infty)}(x)|$$

(where the first inequality holds due to $\sigma(x) \in [0,1]$) we can conclude from Lemma 4 that

$$\max_{i=1,\ldots,n} \left| \sigma' \left(\sum_{j=1}^{d} b_{k,j} \cdot x_i^{(j)} + b_{k,0} \right) \right|$$
$$\leq \max_{i=1,\ldots,n} \exp\left(- \left| \sum_{j=1}^{d} b_{k,j} \cdot x_i^{(j)} + b_{k,0} \right| \right)$$
$$= \exp\left(- \min_{i=1,\ldots,n} \left\{ \left| \sum_{j=1}^{d} b_{k,j} \cdot x_i^{(j)} + b_{k,0} \right| \right\} \right).$$

As a consequence, we get for $k \in \{1, \ldots, K\}$ and $j \in \{1, \ldots, d\}$ by the inequality of Cauchy-Schwarz

$$\left| \frac{\partial F}{\partial b_{k,j}}(\mathbf{a}, \mathbf{b}) \right|$$
$$= \left| \frac{2}{n} \sum_{i=1}^{n} (f_{net,(\mathbf{a},\mathbf{b})}(x_i) - y_i) \cdot a_k \cdot \sigma' \left(\sum_{j=1}^{d} b_{k,j} \cdot x_i^{(j)} + b_{k,0} \right) \cdot x_i^{(j)} \right|$$
$$\leq 2 \cdot |a_k| \cdot \frac{1}{n} \sum_{i=1}^{n} |f_{net,(\mathbf{a},\mathbf{b})}(x_i) - y_i| \cdot |x_i^{(j)}| \cdot \left| \sigma' \left(\sum_{j=1}^{d} b_{k,j} \cdot x_i^{(j)} + b_{k,0} \right) \right|$$

$$\leq 2 \cdot \sqrt{\frac{1}{n}\sum_{i=1}^{n} |f_{net,(\mathbf{a},\mathbf{b})}(x_i) - y_i|^2 \cdot (x_i^{(j)})^2} \cdot |a_k| \cdot \sqrt{\frac{1}{n}\sum_{i=1}^{n} |\sigma'(\sum_{j=1}^{d} b_{k,j} \cdot x_i^{(j)} + b_{k,0})|^2}$$

$$\leq 2 \cdot \sqrt{F(\mathbf{a},\mathbf{b})} \cdot \max_{i,l}\{|x_i^{(l)}|\} \cdot |a_k| \cdot \sqrt{\frac{1}{n}\sum_{i=1}^{n} |\sigma'(\sum_{l=1}^{d} b_{k,l} \cdot x_l^{(j)} + b_{k,0})|^2}$$

$$\leq 2 \cdot \sqrt{F(\mathbf{a},\mathbf{b})} \cdot \max_{i,l}\{|x_i^{(l)}|\} \cdot |a_k| \cdot \exp\left(- \min_{i=1,\ldots,n} \left\{ \left| \sum_{j=1}^{d} b_{k,j} \cdot x_i^{(j)} + b_{k,0} \right| \right\} \right).$$

Hence, we have shown

$$\begin{aligned} &|\bar{b}_{k,j} - b_{k,j}| \\ &= \lambda_n \cdot \left| \frac{\partial F}{\partial b_{k,j}}(\mathbf{a},\mathbf{b}) \right| \\ &\leq \lambda_n \cdot 2 \cdot \sqrt{F(\mathbf{a},\mathbf{b})} \cdot \max_{i,l}\{|x_i^{(l)}|\} \cdot |a_k| \cdot \exp\left(- \min_{i=1,\ldots,n} \left\{ \left| \sum_{j=1}^{d} b_{k,j} \cdot x_i^{(j)} + b_{k,0} \right| \right\} \right) \end{aligned}$$

for any $k \in \{1,\ldots,K\}$ and any $j \in \{1,\ldots,d\}$.

In case that $k \in \{1,\ldots,K\}$ and $j = 0$ we get in a similar fashion

$$\begin{aligned} |\bar{b}_{k,0} - b_{k,0}| &= \lambda_n \cdot \left| \frac{\partial F}{\partial b_{k,0}}(\mathbf{a},\mathbf{b}) \right| \\ &\leq \lambda_n \cdot 2 \cdot \sqrt{F(\mathbf{a},\mathbf{b})} \cdot 1 \cdot |a_k| \cdot \exp\left(- \min_{i=1,\ldots,n} \left\{ \left| \sum_{j=1}^{d} b_{k,j} \cdot x_i^{(j)} + b_{k,0} \right| \right\} \right), \end{aligned}$$

which implies the assertion. □

Lemma 7 *Define F by (15) and define $(\mathbf{a}^{(t)},\mathbf{b}^{(t)})$ by (16). Assume that $(\mathbf{a}^{(t)},\mathbf{b}^{(t)})$ satisfy for $t \in \{1,\ldots,t_n\}$*

$$F(\mathbf{a}^{(t)},\mathbf{b}^{(t)}) \leq c_5 < \infty, \tag{17}$$

$$\|\mathbf{a}^{(t)}\|^2 \leq c_6 \cdot n < \infty, \tag{18}$$

$$\min_{i=1,\ldots,n,k=1,\ldots,K} \left| \sum_{j=1}^{d} b_{k,j}^{(0)} \cdot x_i^{(j)} + b_{k,0}^{(0)} \right| \geq \delta_n > 0 \tag{19}$$

and

$$(d+1) \cdot t_n \cdot \lambda_n \cdot 2 \cdot \sqrt{c_5} \cdot \max\{1, \max_{i,l}\{|x_i^{(l)}|^2\}\} \cdot \sqrt{c_6 \cdot n} \cdot \exp\left(-\delta_n/2\right) \leq \frac{\delta_n}{2}. \tag{20}$$

Then we have for every $k \in \{1,\ldots,K\}$, any $j \in \{0,\ldots,d\}$ and any $t \in \{1,\ldots,t_n\}$:

$$|b_{k,j}^{(t)} - b_{k,j}^{(t-1)}| \leq \lambda_n \cdot 2 \cdot \sqrt{c_5} \cdot \max\{1, \max_{i,l}\{|x_i^{(l)}|\}\} \cdot \sqrt{c_6 \cdot n} \cdot \exp\left(-\delta_n/2\right). \tag{21}$$

Proof. We show (21) by induction on t. For $t = 1$ the assertion follows from Lemma 6 and (17)-(19). Now, we assume that (21) holds for all $t \in \{1, \ldots, s\}$, where $s \in \{1, \ldots, t_n - 1\}$. Then

$$|b_{k,j}^{(s)} - b_{k,j}^{(0)}| \leq t_n \cdot \lambda_n \cdot 2 \cdot \sqrt{c_5} \cdot \max\{1, \max_{i,l}\{|x_i^{(l)}|\}\} \cdot \sqrt{c_6 \cdot n} \cdot \exp\left(-\delta_n/2\right),$$

from which, together with assumption (19), we can conlcude that

$$\begin{aligned}
&\min_{i=1,\ldots,n,k=1,\ldots,K} \left| \sum_{j=1}^{d} b_{k,j}^{(s)} \cdot x_i^{(j)} + b_{k,0}^{(s)} \right| \\
&\geq \min_{i=1,\ldots,n,k=1,\ldots,K} \left| \sum_{j=1}^{d} b_{k,j}^{(0)} \cdot x_i^{(j)} + b_{k,0}^{(0)} \right| \\
&\quad - \max_{i=1,\ldots,n,k=1,\ldots,K} \left(\sum_{j=1}^{d} |b_{k,j}^{(s)} - b_{k,j}^{(0)}| \cdot |x_i^{(j)}| + |b_{k,0}^{(s)} - b_{k,0}^{(0)}| \right) \\
&\geq \delta_n - \max_{i=1,\ldots,n,k=1,\ldots,K} \left(\sum_{j=0}^{d} |b_{k,j}^{(s)} - b_{k,j}^{(0)}| \cdot \max\{1, \max_{i,l}\{|x_i^{(l)}|\}\} \right) \\
&\geq \delta_n - (d+1) \cdot t_n \cdot \lambda_n \cdot 2 \cdot \sqrt{c_5} \cdot \max\{1, \max_{i,l}\{|x_i^{(l)}|^2\}\} \cdot \sqrt{c_6 \cdot n} \cdot \exp\left(-\delta_n/2\right) \\
&\geq \frac{\delta_n}{2}, \qquad (22)
\end{aligned}$$

where the last inequality is implied by inequality (20). So, for the induction step, application of Lemma 6 together with (17) and (22) yields

$$\begin{aligned}
|b_{k,j}^{(s+1)} - b_{k,j}^{(s)}| &\leq \lambda_n \cdot 2 \cdot \sqrt{F(\mathbf{a}^{(s)}, \mathbf{b}^{(s)})} \cdot \max\{1, \max_{i,l}\{|x_i^{(l)}|\}\} \cdot |a_k^{(s)}| \\
&\qquad \cdot \exp\left(- \min_{i=1,\ldots,n} \left\{ \left| \sum_{j=1}^{d} b_{k,j}^{(s)} \cdot x_i^{(j)} + b_{k,0}^{(s)} \right| \right\} \right) \\
&\leq \lambda_n \cdot 2 \cdot \sqrt{c_5} \cdot \max\{1, \max_{i,l}\{|x_i^{(l)}|\}\} \cdot \sqrt{c_6 \cdot n} \cdot \exp\left(-\delta_n/2\right),
\end{aligned}$$

from which we conclude the assertion. □

Lemma 8 *Define F by (15), set*

$$\lambda_n = \frac{3}{K}$$

and define $(\mathbf{a}^{(t)}, \mathbf{b}^{(t)})$ by (16). Assume that $(\mathbf{a}^{(0)}, \mathbf{b}^{(0)})$ is chosen such that

$$F(\mathbf{a}^{(0)}, \mathbf{b}^{(0)}) \leq c_5 < \infty \qquad (23)$$

and

$$\min_{i=1,\ldots,n,k=1,\ldots,K} \left| \sum_{j=1}^{d} b_{k,j}^{(0)} \cdot x_i^{(j)} + b_{k,0}^{(0)} \right| \geq \delta_n \geq 1 \tag{24}$$

hold. Let $t_n \in \mathbb{N}$ *and assume* $2 \cdot c_1 \leq K \cdot n$,

$$36 \cdot \max\{1, \frac{c_5}{c_1}\} \cdot \max\{1, \frac{1}{c_1^2}\} \cdot \frac{(d+1)^2 \cdot n^2}{K} \cdot \max\{1, \max_{i,j} |x_i^{(j)}|^4\} \cdot \left(1 + c_5 + \frac{2}{n} \sum_{i=1}^{n} y_i^2\right)^4 \cdot t_n^2 \cdot \exp\left(-\delta_n/2\right) \leq 1 \tag{25}$$

and

$$3 \cdot t_n \cdot \exp(-\delta_n/4) \leq 1. \tag{26}$$

Then for any $t \in \{0, 1, \ldots, t_n - 1\}$ *we have*

$$\begin{aligned} & F(\mathbf{a}^{(t+1)}, \mathbf{b}^{(t+1)}) - \min_{\mathbf{a}} F(\mathbf{a}, \mathbf{b}^{(0)}) \\ & \leq \left(1 - \frac{c_1}{6 \cdot K \cdot n}\right)^{t+1} \cdot \left(F(\mathbf{a}^{(0)}, \mathbf{b}^{(0)}) - \min_{\mathbf{a}} F(\mathbf{a}, \mathbf{b}^{(0)})\right) + (2\sqrt{c_5} + 1) \cdot \exp\left(-\delta_n/4\right) \\ & \quad + \frac{6 \cdot K \cdot n}{c_1} \cdot 3 \cdot \exp\left(-\delta_n/4\right). \end{aligned}$$

Proof. We have

$$\begin{aligned} & F(\mathbf{a}^{(t+1)}, \mathbf{b}^{(t+1)}) - \min_{\mathbf{a}} F(\mathbf{a}, \mathbf{b}^{(0)}) \\ & = \left(F(\mathbf{a}^{(t+1)}, \mathbf{b}^{(t+1)}) - F(\mathbf{a}^{(t+1)}, \mathbf{b}^{(t)})\right) + \left(F(\mathbf{a}^{(t+1)}, \mathbf{b}^{(t)}) - \min_{\mathbf{a}} F(\mathbf{a}, \mathbf{b}^{(t)})\right) \\ & \quad + \left(\min_{\mathbf{a}} F(\mathbf{a}, \mathbf{b}^{(t)}) - \min_{\mathbf{a}} F(\mathbf{a}, \mathbf{b}^{(0)})\right). \end{aligned}$$

We take a look at the second and the third term on the right-hand side of the above equality. Lemma 2 and $|\sigma(x)| \leq 1$, which give us

$$\left\|(\nabla_{\mathbf{a}} F)(\mathbf{a}_1, \mathbf{b}^{(t)}) - (\nabla_{\mathbf{a}} F)(\mathbf{a}_2, \mathbf{b}^{(t)})\right\| \leq \left(2 \cdot K + \frac{2 \cdot c_1}{n}\right) \cdot \|\mathbf{a}_1 - \mathbf{a}_2\| \leq 3 \cdot K \cdot \|\mathbf{a}_1 - \mathbf{a}_2\|,$$

together with Lemma 3 allow us to conclude from Lemma 1 that

$$\begin{aligned} & F(\mathbf{a}^{(t+1)}, \mathbf{b}^{(t)}) - \min_{\mathbf{a}} F(\mathbf{a}, \mathbf{b}^{(t)}) \\ & \leq \left(1 - \frac{c_1}{6 \cdot K \cdot n}\right) \cdot \left(F(\mathbf{a}^{(t)}, \mathbf{b}^{(t)}) - \min_{\mathbf{a}} F(\mathbf{a}, \mathbf{b}^{(t)})\right). \end{aligned}$$

As a consequence,

$$\begin{aligned} & \left(F(\mathbf{a}^{(t+1)}, \mathbf{b}^{(t)}) - \min_{\mathbf{a}} F(\mathbf{a}, \mathbf{b}^{(t)})\right) + \left(\min_{\mathbf{a}} F(\mathbf{a}, \mathbf{b}^{(t)}) - \min_{\mathbf{a}} F(\mathbf{a}, \mathbf{b}^{(0)})\right) \\ & \leq \left(1 - \frac{c_1}{6 \cdot K \cdot n}\right) \cdot \left(F(\mathbf{a}^{(t)}, \mathbf{b}^{(t)}) - \min_{\mathbf{a}} F(\mathbf{a}, \mathbf{b}^{(t)})\right) \end{aligned}$$

$$
\begin{aligned}
&+\min_{\mathbf{a}} F(\mathbf{a}, \mathbf{b}^{(t)}) - \min_{\mathbf{a}} F(\mathbf{a}, \mathbf{b}^{(0)}) \\
&= \left(1 - \frac{c_1}{6 \cdot K \cdot n}\right) \cdot \left(F(\mathbf{a}^{(t)}, \mathbf{b}^{(t)}) - \min_{\mathbf{a}} F(\mathbf{a}, \mathbf{b}^{(0)})\right) \\
&+\frac{c_1}{6 \cdot K \cdot n} \cdot \left(\min_{\mathbf{a}} F(\mathbf{a}, \mathbf{b}^{(t)}) - \min_{\mathbf{a}} F(\mathbf{a}, \mathbf{b}^{(0)})\right) \qquad (27)
\end{aligned}
$$

and so

$$
\begin{aligned}
&F(\mathbf{a}^{(t+1)}, \mathbf{b}^{(t+1)}) - \min_{\mathbf{a}} F(\mathbf{a}, \mathbf{b}^{(0)}) \\
&\leq \left(1 - \frac{c_1}{6 \cdot K \cdot n}\right) \cdot \left(F(\mathbf{a}^{(t)}, \mathbf{b}^{(t)}) - \min_{\mathbf{a}} F(\mathbf{a}, \mathbf{b}^{(0)})\right) \\
&\quad +\frac{c_1}{6 \cdot K \cdot n} \cdot \left(\min_{\mathbf{a}} F(\mathbf{a}, \mathbf{b}^{(t)}) - \min_{\mathbf{a}} F(\mathbf{a}, \mathbf{b}^{(0)})\right) \\
&\quad +F(\mathbf{a}^{(t+1)}, \mathbf{b}^{(t+1)}) - F(\mathbf{a}^{(t+1)}, \mathbf{b}^{(t)}). \qquad (28)
\end{aligned}
$$

In order to bound this term further, we introduce the following notation for simplicity

$$
\begin{aligned}
\gamma_t &= \left(F(\mathbf{a}^{(t)}, \mathbf{b}^{(t)}) - \min_{\mathbf{a}} F(\mathbf{a}, \mathbf{b}^{(0)})\right), \\
\alpha &= \frac{c_1}{6 \cdot K \cdot n}.
\end{aligned}
$$

In the sequel, we will derive upper bounds $\beta_1, \beta_2 > 0$ such that

$$
\begin{aligned}
\beta_1 &\geq \min_{\mathbf{a}} F(\mathbf{a}, \mathbf{b}^{(t)}) - \min_{\mathbf{a}} F(\mathbf{a}, \mathbf{b}^{(0)}), \\
\beta_2 &\geq F(\mathbf{a}^{(t+1)}, \mathbf{b}^{(t+1)}) - F(\mathbf{a}^{(t+1)}, \mathbf{b}^{(t)}).
\end{aligned}
$$

Then, the above results imply that we have

$$
\gamma_{t+1} \leq (1 - \alpha) \cdot \gamma_t + \alpha \cdot \beta_1 + \beta_2,
$$

and applying this relation recursively using standard techniques from the literature we get

$$
\begin{aligned}
\gamma_{t+1} &\leq (1 - \alpha) \cdot ((1 - \alpha) \cdot \gamma_{t-1} + \alpha \cdot \beta_1 + \beta_2) + \alpha \cdot \beta_1 + \beta_2 \\
&= (1 - \alpha)^2 \cdot \gamma_{t-1} + (1 - \alpha) \cdot \alpha \cdot \beta_1 + \alpha \cdot \beta_1 + (1 - \alpha) \cdot \beta_2 + \beta_2 \\
&\leq \dots \\
&\leq (1 - \alpha)^{t+1} \cdot \gamma_0 + \sum_{k=0}^{t} (1 - \alpha)^k \cdot \alpha \cdot \beta_1 + \sum_{k=0}^{t} (1 - \alpha)^k \cdot \beta_2 \\
&\leq (1 - \alpha)^{t+1} \cdot \gamma_0 + \sum_{k=0}^{\infty} (1 - \alpha)^k \cdot \alpha \cdot \beta_1 + \sum_{k=0}^{\infty} (1 - \alpha)^k \cdot \beta_2 \\
&= (1 - \alpha)^{t+1} \cdot \gamma_0 + \frac{\alpha \cdot \beta_1}{1 - (1 - \alpha)} + \frac{\beta_2}{1 - (1 - \alpha)} \\
&= (1 - \alpha)^{t+1} \cdot \gamma_0 + \beta_1 + \frac{\beta_2}{\alpha}.
\end{aligned}
$$

It remains to find β_1 and β_2. In order to derive and upper bound β_1, we will apply Lemma 7, for which we show the following claim for all $s \in \{0, 1, \ldots, t_n - 1\}$ by induction

$$\begin{aligned}
&\max\left\{F(\mathbf{a}^{(s+1)}, \mathbf{b}^{(s)}), F(\mathbf{a}^{(s+1)}, \mathbf{b}^{(s+1)})\right\} - \min_{\mathbf{a}} F(\mathbf{a}, \mathbf{b}^{(0)}) \\
&\leq c_5 + \frac{1}{n}\sum_{i=1}^{n} y_i^2 + 3 \cdot (s+1) \cdot \exp(-\delta_n/4). \qquad (29)
\end{aligned}$$

While doing so, we will be deriving an upper bound β_2 in the process. For $s = 0$ the inequality trivially holds by (27), (28), (23) and by the bound

$$F(\mathbf{a}^{(1)}, \mathbf{b}^{(1)}) - F(\mathbf{a}^{(1)}, \mathbf{b}^{(0)}) \leq 3 \cdot \exp\left(-\delta_n/4\right)$$

which will be proven below (cf., (32)).

So, for the induction hypothesis, assume that (29) holds for $s = t - 1$ for arbitrary $t \in \{1, \ldots, t_n - 1\}$. Trivially we have

$$\min_{\mathbf{a}} F(\mathbf{a}, \mathbf{b}^{(t)}) - \min_{\mathbf{a}} F(\mathbf{a}, \mathbf{b}^{(0)}) \leq F(\mathbf{0}, \mathbf{b}^{(t)}) = \frac{1}{n}\sum_{i=1}^{n} y_i^2,$$

hence by (27) and by the induction assumption we get

$$\begin{aligned}
&F(\mathbf{a}^{(t+1)}, \mathbf{b}^{(t)}) - \min_{\mathbf{a}} F(\mathbf{a}, \mathbf{b}^{(0)}) \\
&\leq \left(1 - \frac{c_1}{6 \cdot K \cdot n}\right) \cdot \left(F(\mathbf{a}^{(t)}, \mathbf{b}^{(t)}) - \min_{\mathbf{a}} F(\mathbf{a}, \mathbf{b}^{(0)})\right) \\
&\quad + \frac{c_1}{6 \cdot K \cdot n} \cdot \left(\min_{\mathbf{a}} F(\mathbf{a}, \mathbf{b}^{(t)}) - \min_{\mathbf{a}} F(\mathbf{a}, \mathbf{b}^{(0)})\right) \\
&\leq \left(1 - \frac{c_1}{6 \cdot K \cdot n}\right) \cdot \left(c_5 + \frac{1}{n}\sum_{i=1}^{n} y_i^2 + 3 \cdot t \cdot \exp(-\delta_n/4)\right) \\
&\quad + \frac{c_1}{6 \cdot K \cdot n} \cdot \frac{1}{n}\sum_{i=1}^{n} y_i^2 \\
&\leq c_5 + \frac{1}{n}\sum_{i=1}^{n} y_i^2 + 3 \cdot t \cdot \exp(-\delta_n/4). \qquad (30)
\end{aligned}$$

Next, by (28) and by the induction hypothesis we get

$$\begin{aligned}
&F(\mathbf{a}^{(t+1)}, \mathbf{b}^{(t+1)}) - \min_{\mathbf{a}} F(\mathbf{a}, \mathbf{b}^{(0)}) \\
&\leq c_5 + \frac{1}{n}\sum_{i=1}^{n} y_i^2 + 3 \cdot t \cdot \exp(-\delta_n/4) + F(\mathbf{a}^{(t+1)}, \mathbf{b}^{(t+1)}) - F(\mathbf{a}^{(t+1)}, \mathbf{b}^{(t)}). \qquad (31)
\end{aligned}$$

Further, we have

$$F(\mathbf{a}^{(t+1)}, \mathbf{b}^{(t+1)}) - F(\mathbf{a}^{(t+1)}, \mathbf{b}^{(t)})$$

$$
\begin{aligned}
&= \frac{1}{n}\sum_{i=1}^{n}(f_{net,(\mathbf{a}^{(t+1)},\mathbf{b}^{(t+1)})}(x_i) + f_{net,\mathbf{a}^{(t+1)},\mathbf{b}^{(t)})}(x_i) - 2y_i) \\
&\qquad \cdot(f_{net,(\mathbf{a}^{(t+1)},\mathbf{b}^{(t+1)})}(x_i) - f_{net,(\mathbf{a}^{(t+1)},\mathbf{b}^{(t)})}(x_i)) \\
&= \frac{1}{n}\sum_{i=1}^{n}(2f_{net,(\mathbf{a}^{(t+1)},\mathbf{b}^{(t)})}(x_i) - 2y_i) \cdot (f_{net,(\mathbf{a}^{(t+1)},\mathbf{b}^{(t+1)})}(x_i) - f_{net,(\mathbf{a}^{(t+1)},\mathbf{b}^{(t)})}(x_i)) \\
&\quad + \frac{1}{n}\sum_{i=1}^{n}(f_{net,(\mathbf{a}^{(t+1)},\mathbf{b}^{(t+1)})}(x_i) - f_{net,(\mathbf{a}^{(t+1)},\mathbf{b}^{(t)})}(x_i))^2 \\
&\leq 2 \cdot \sqrt{F(\mathbf{a}^{(t+1)},\mathbf{b}^{(t)})} \cdot \sqrt{\frac{1}{n}\sum_{i=1}^{n}(f_{net,(\mathbf{a}^{(t+1)},\mathbf{b}^{(t+1)})}(x_i) - f_{net,(\mathbf{a}^{(t+1)},\mathbf{b}^{(t)})}(x_i))^2} \\
&\quad + \frac{1}{n}\sum_{i=1}^{n}(f_{net,(\mathbf{a}^{(t+1)},\mathbf{b}^{(t+1)})}(x_i) - f_{net,(\mathbf{a}^{(t+1)},\mathbf{b}^{(t)})}(x_i))^2.
\end{aligned}
$$

Since σ is Lipschitz continuous applying the inequality of Cauchy-Schwarz a second time yields

$$
\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}(f_{net,(\mathbf{a}^{(t+1)},\mathbf{b}^{(t+1)})}(x_i) - f_{net,(\mathbf{a}^{(t+1)},\mathbf{b}^{(t)})}(x_i))^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{k=1}^{K}(\mathbf{a}^{(t+1)})_k \cdot \left(\sigma\left(\sum_{j=1}^{d} b_{k,j}^{(t+1)} \cdot x_i^{(j)} + b_{k,0}^{(t+1)}\right) - \sigma\left(\sum_{j=1}^{d} b_{k,j}^{(t)} \cdot x_i^{(j)} + b_{k,0}^{(t)}\right)\right)\right)^2 \\
&\leq \sum_{k=1}^{K}(\mathbf{a}^{(t+1)})_k^2 \cdot \max\{1, \max_{i,j}|x_i^{(j)}|^2\} \cdot (d+1) \cdot \sum_{k=1}^{K}\sum_{j=0}^{d}|b_{k,j}^{(t+1)} - b_{k,j}^{(t)}|^2 \\
&\leq \frac{n}{c_1} \cdot F(\mathbf{a}^{(t+1)},\mathbf{b}^{(t)}) \cdot \max\{1, \max_{i,j}|x_i^{(j)}|^2\} \cdot (d+1) \cdot \sum_{k=1}^{K}\sum_{j=0}^{d}|b_{k,j}^{(t+1)} - b_{k,j}^{(t)}|^2.
\end{aligned}
$$

By Lemma 7 (where (17) and (18) are true because of the fact that the induction hypothesis implies that we have

$$
F(\mathbf{a}^{(t)},\mathbf{b}^{(t)}) \leq c_5 + \frac{1}{n}\sum_{i=1}^{n} y_i^2 + 1 + F(\mathbf{0},\mathbf{b}^{(0)}) \leq 1 + c_5 + \frac{2}{n}\sum_{i=1}^{n} y_i^2,
$$

from which (together with the defnition of F) we can conclude that (17) and (18) hold if we replace there c_5 and c_6 by

$$
1 + c_5 + \frac{2}{n}\sum_{i=1}^{n} y_i^2 \quad \text{and} \quad \left(1 + c_5 + \frac{2}{n}\sum_{i=1}^{n} y_i^2\right) \cdot \frac{1}{c_1}, \text{ respectively,}
$$

and where (20) holds because of (25)) and because of (24) we know that for any $k \in \{1,\ldots,K\}$ and any $j \in \{0,\ldots,d\}$ we have

$$
|b_{k,j}^{(t+1)} - b_{k,j}^{(t)}|
$$

$$\leq \lambda_n \cdot 2 \cdot \left(1 + c_5 + \frac{2}{n}\sum_{i=1}^{n} y_i^2\right) \cdot \max\{1, \max_{i,l}\{|x_i^{(l)}|\}\} \cdot \sqrt{n} \cdot \exp\left(-\delta_n/2\right)/\sqrt{c_1}.$$

Together with

$$\begin{aligned} F(\mathbf{a}^{(t+1)}, \mathbf{b}^{(t)}) &\leq F(\mathbf{a}^{(t)}, \mathbf{b}^{(t)}) \\ &\leq \min_{\mathbf{a}} F(\mathbf{a}, \mathbf{b}^{(0)}) + c_5 + \frac{1}{n}\sum_{i=1}^{n} y_i^2 + 3 \cdot t \cdot \exp(-\delta_n/4) \\ &\leq 1 + c_5 + \frac{2}{n}\sum_{i=1}^{n} y_i^2 \end{aligned}$$

(where the first inequality follows trivially from Lemma 1, and where we have used the induction assumption for the second inequality), this implies

$$\begin{aligned} &\frac{1}{n}\sum_{i=1}^{n}(f_{net,(\mathbf{a}^{(t+1)},\mathbf{b}^{(t+1)})}(x_i) - f_{net,(\mathbf{a}^{(t+1)},\mathbf{b}^{(t)})}(x_i))^2 \\ &\leq 4 \cdot \frac{n^2}{c_1^2} \cdot \lambda_n^2 \cdot \left(1 + c_5 + \frac{2}{n}\sum_{i=1}^{n} y_i^2\right)^3 \cdot \max\{1, \max_{i,j}|x_i^{(j)}|^4\} \cdot (d+1)^2 \cdot K \cdot \exp\left(-\delta_n\right) \\ &\leq 36 \cdot \frac{(d+1)^2 \cdot n^2}{c_1^2 \cdot K} \cdot \max\{1, \max_{i,j}|x_i^{(j)}|^4\} \cdot \left(1 + c_5 + \frac{2}{n}\sum_{i=1}^{n} y_i^2\right)^4 \cdot \exp\left(-\delta_n/2\right) \\ &\quad \cdot \min\left\{1, (F(\mathbf{a}^{(t+1)}, \mathbf{b}^{(t)}))^{-1}\right\} \cdot \exp\left(-\delta_n/2\right) \\ &\leq \min\left\{1, (F(\mathbf{a}^{(t+1)}, \mathbf{b}^{(t)}))^{-1}\right\} \cdot \exp\left(-\delta_n/2\right). \end{aligned}$$

(Here the last inequality follows from (25)). Summarizing the above results we get

$$F(\mathbf{a}^{(t+1)}, \mathbf{b}^{(t+1)}) - F(\mathbf{a}^{(t+1)}, \mathbf{b}^{(t)}) \leq 3 \cdot \exp\left(-\delta_n/4\right) = \beta_2. \tag{32}$$

By combining this inequality with the results above we get (29) for $s = t + 1$. This concludes the proof of (29).

Now, we see that the conditions of Lemma 7 are met, since we can conclude from

$$\min_{\mathbf{a}} F(\mathbf{a}, \mathbf{b}^{(0)}) \leq F(\mathbf{0}, \mathbf{b}^{(0)}) = \frac{1}{n}\sum_{i=1}^{n} y_i^2$$

and from inequalities (29) and (26) that also (17) and (because of the defintion of F) (18) hold where c_5 and c_6 are replaced by

$$1 + c_5 + \frac{2}{n}\sum_{i=1}^{n} y_i^2 \quad \text{and} \quad \left(1 + c_5 + \frac{2}{n}\sum_{i=1}^{n} y_i^2\right) \cdot \frac{1}{c_1}, \text{ respectively.}$$

As above we also see that (20) holds.

Next we derive an upper bound on

$$\min_{\mathbf{a}} F(\mathbf{a}, \mathbf{b}^{(t)}) - \min_{\mathbf{a}} F(\mathbf{a}, \mathbf{b}^{(0)}).$$

Choose $\bar{\mathbf{a}}$ such that

$$F(\bar{\mathbf{a}}, \mathbf{b}^{(0)}) = \min_{\mathbf{a}} F(\mathbf{a}, \mathbf{b}^{(0)}).$$

Then

$$\frac{c_1}{n} \cdot \sum_{k=0}^{n} \bar{\mathbf{a}}_k^2 \leq F(\bar{\mathbf{a}}, \mathbf{b}^{(0)}) \leq F(\mathbf{a}^{(0)}, \mathbf{b}^{(0)}) \leq c_5,$$

hence

$$\sum_{k=0}^{K} \bar{\mathbf{a}}_k^2 \leq \frac{c_5 \cdot n}{c_1}.$$

We have

$$\begin{aligned}
&\min_{\mathbf{a}} F(\mathbf{a}, \mathbf{b}^{(t)}) - \min_{\mathbf{a}} F(\mathbf{a}, \mathbf{b}^{(0)}) = \min_{\mathbf{a}} F(\mathbf{a}, \mathbf{b}^{(t)}) - F(\bar{\mathbf{a}}, \mathbf{b}^{(0)}) \\
&\leq F(\bar{\mathbf{a}}, \mathbf{b}^{(t)}) - F(\bar{\mathbf{a}}, \mathbf{b}^{(0)}) \\
&= \frac{1}{n} \sum_{i=1}^{n} (f_{net,(\bar{\mathbf{a}}, \mathbf{b}^{(t)})}(x_i) + f_{net,(\bar{\mathbf{a}}, \mathbf{b}^{(0)})}(x_i) - 2y_i) \cdot (f_{net,(\bar{\mathbf{a}}, \mathbf{b}^{(t)})}(x_i) - f_{net,(\bar{\mathbf{a}}, \mathbf{b}^{(0)})}(x_i)) \\
&= \frac{1}{n} \sum_{i=1}^{n} (2 f_{net,(\bar{\mathbf{a}}, \mathbf{b}^{(0)})}(x_i) - 2y_i) \cdot (f_{net,(\bar{\mathbf{a}}, \mathbf{b}^{(t)})}(x_i) - f_{net,(\bar{\mathbf{a}}, \mathbf{b}^{(0)})}(x_i)) \\
&\quad + \frac{1}{n} \sum_{i=1}^{n} (f_{net,(\bar{\mathbf{a}}, \mathbf{b}^{(t)})}(x_i) - f_{net,(\bar{\mathbf{a}}, \mathbf{b}^{(0)})}(x_i))^2 \\
&\leq 2 \cdot \sqrt{F(\bar{\mathbf{a}}, \mathbf{b}^{(0)})} \cdot \sqrt{\frac{1}{n} \sum_{i=1}^{n} (f_{net,(\bar{\mathbf{a}}, \mathbf{b}^{(t)})}(x_i) - f_{net,(\bar{\mathbf{a}}, \mathbf{b}^{(0)})}(x_i))^2} \\
&\quad + \frac{1}{n} \sum_{i=1}^{n} (f_{net,(\bar{\mathbf{a}}, \mathbf{b}^{(t)})}(x_i) - f_{net,(\bar{\mathbf{a}}, \mathbf{b}^{(0)})}(x_i))^2.
\end{aligned}$$

Applying the inequality of Cauchy-Schwarz a second time and since σ is Lipschitz continuous we get

$$\begin{aligned}
&\frac{1}{n} \sum_{i=1}^{n} (f_{net,(\bar{\mathbf{a}}, \mathbf{b}^{(t)})}(x_i) - f_{net,(\bar{\mathbf{a}}, \mathbf{b}^{(0)})}(x_i))^2 \\
&= \frac{1}{n} \sum_{i=1}^{n} \left(\sum_{k=1}^{K} \bar{a}_k \cdot \left(\sigma \left(\sum_{j=1}^{d} b_{k,j}^{(t)} \cdot x_i^{(j)} + b_{k,0}^{(t)} \right) - \sigma \left(\sum_{j=1}^{d} b_{k,j}^{(0)} \cdot x_i^{(j)} + b_{k,0}^{(0)} \right) \right) \right)^2 \\
&\leq \sum_{k=1}^{K} \bar{a}_k^2 \cdot \max\{1, \max_{i,j} |x_i^{(j)}|^2\} \cdot (d+1) \cdot \sum_{k=1}^{K} \sum_{j=0}^{d} |b_{k,j}^{(t)} - b_{k,j}^{(0)}|^2.
\end{aligned}$$

By Lemma 7 we know that for any $k \in \{1, \dots, K\}$ and any $j \in \{0, \dots, d\}$

$$
\begin{aligned}
&|b_{k,j}^{(t)} - b_{k,j}^{(0)}| \\
&\leq |b_{k,j}^{(t)} - b_{k,j}^{(t-1)}| + |b_{k,j}^{(t-1)} - b_{k,j}^{(t-2)}| + \cdots + |b_{k,j}^{(1)} - b_{k,j}^{(0)}| \\
&\leq t \cdot \lambda_n \cdot 2 \cdot \left(1 + c_5 + \frac{2}{n}\sum_{i=1}^{n} y_i^2\right) \cdot \max\{1, \frac{1}{c_1}\} \cdot \max\{1, \max_{i,l}\{|x_i^{(l)}|\}\} \cdot \sqrt{n} \cdot \exp\left(-\delta_n/2\right).
\end{aligned}
$$

From this we conclude that

$$
\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}(f_{net,(\bar{\mathbf{a}},\mathbf{b}^{(t)})}(x_i) - f_{net,(\bar{\mathbf{a}},\mathbf{b}^{(0)})}(x_i))^2 \\
&\leq \sum_{k=1}^{K} \bar{a}_k^2 \cdot \max\{1, \max_{i,j}|x_i^{(j)}|^2\} \cdot (d+1) \cdot K \cdot (d+1) \\
&\quad \cdot \left(t \cdot \lambda_n \cdot \left(1 + c_5 + \frac{2}{n}\sum_{i=1}^{n} y_i^2\right) \cdot \max\{1, \frac{1}{c_1}\} \cdot \max\{1, \max_{i,l}\{|x_i^{(l)}|\}\} \cdot \sqrt{n} \cdot \exp\left(-\delta_n/2\right)\right)^2 \\
&\leq \frac{c_5}{c_1} \cdot \max\{1, \frac{1}{c_1}\} \cdot \left(1 + c_5 + \frac{2}{n}\sum_{i=1}^{n} y_i^2\right)^2 \cdot n^2 \cdot \max\{1, \max_{i,j}|x_i^{(j)}|^4\} \cdot (d+1)^2 \\
&\qquad\qquad \cdot K \cdot t^2 \cdot \lambda_n^2 \cdot \exp\left(-\delta_n\right) \\
&\leq \exp\left(-\delta_n/2\right),
\end{aligned}
$$

where the last inequality follows from (25). Hence,

$$
\begin{aligned}
&\min_{\mathbf{a}} F(\mathbf{a}, \mathbf{b}^{(t)}) - \min_{\mathbf{a}} F(\mathbf{a}, \mathbf{b}^{(0)}) \\
&\leq 2 \cdot \sqrt{F(\bar{\mathbf{a}}, \mathbf{b}^{(0)})} \cdot \sqrt{\frac{1}{n}\sum_{i=1}^{n}(f_{net,(\bar{\mathbf{a}},\mathbf{b}^{(t)})}(x_i) - f_{net,(\bar{\mathbf{a}},\mathbf{b}^{(0)})}(x_i))^2} \\
&\quad + \frac{1}{n}\sum_{i=1}^{n}(f_{net,(\bar{\mathbf{a}},\mathbf{b}^{(t)})}(x_i) - f_{net,(\bar{\mathbf{a}},\mathbf{b}^{(0)})}(x_i))^2 \\
&\leq (2 \cdot \sqrt{F(\bar{\mathbf{a}}, \mathbf{b}^{(0)})} + 1) \cdot \sqrt{\frac{1}{n}\sum_{i=1}^{n}(f_{net,(\bar{\mathbf{a}},\mathbf{b}^{(t)})}(x_i) - f_{net,(\bar{\mathbf{a}},\mathbf{b}^{(0)})}(x_i))^2} \\
&\leq (2 \cdot \sqrt{c_5} + 1) \cdot \exp\left(-\delta_n/4\right) = \beta_1.
\end{aligned}
$$

Plugging in the above results yields

$$
\begin{aligned}
&F(\mathbf{a}^{(t+1)}, \mathbf{b}^{(t+1)}) - \min_{\mathbf{a}} F(\mathbf{a}, \mathbf{b}^{(0)}) \\
&\leq (1-\alpha)^{t+1} \cdot \gamma_0 + \beta_1 + \frac{\beta_2}{\alpha}. \\
&\leq \left(1 - \frac{c_1}{6 \cdot K \cdot n}\right)^{t+1} \cdot \left(F(\mathbf{a}^{(0)}, \mathbf{b}^{(0)}) - \min_{\mathbf{a}} F(\mathbf{a}, \mathbf{b}^{(0)})\right) + (2 \cdot \sqrt{c_5} + 1) \cdot \exp\left(-\delta_n/4\right)
\end{aligned}
$$

$$+\frac{6 \cdot K \cdot n}{c_1} \cdot 3 \cdot \exp\left(-\delta_n/4\right),$$

which concludes the proof.

□

5.3. Two auxiliary results from empirical process theory

Lemma 9 *Let $\beta_n = c_3 \cdot \log(n)$ for some suitably large constant $c_3 > 0$. Assume that the distribution of (X, Y) satisfies (7) for some constant $c_2 > 0$ and that the regression function m is bounded in absolute value. Let $\mathcal{F}_n$ be a set of functions $f : \mathbb{R}^d \to \mathbb{R}$ and assume that the estimate m_n satisfies*

$$m_n = T_{\beta_n} \tilde{m}_n,$$

$$\tilde{m}_n(\cdot) = \tilde{m}_n(\cdot, (X_1, Y_1), \ldots, (X_n, Y_n)) \in \mathcal{F}_n$$

and

$$\frac{1}{n}\sum_{i=1}^{n} |Y_i - \tilde{m}_n(X_i)|^2 \cdot I_{\{|Y_i| \le \beta_n \text{ for all } i \in \{1,\ldots,n\}\}}$$
$$\le \min_{l \in \Theta_n} \left(\frac{1}{n}\sum_{i=1}^{n} |Y_i - g_{n,l}(X_i)|^2 + pen_n(g_{n,l}) + \epsilon_{n,l} \right)$$

for some random functions $g_{n,l} : \mathbb{R}^d \to \mathbb{R}$, some nonempty parameter set Θ_n and some random penalty terms $pen_n(g_{n,l}) \ge 0$, and some additional deterministic term $\epsilon_{n,l}$, where $g_{n,l}$ and $pen_n(g_{n,l})$ are independent of the data (X_1, Y_1), …, (X_n, Y_n).

Then m_n satisfies

$$\mathbf{E} \int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx) \le \frac{c_7 \cdot (\log n)^2 \cdot \left(\log\left(\sup_{x_1^n} \mathcal{N}_1\left(\frac{1}{n \cdot \beta_n}, \mathcal{F}_n, x_1^n \right) \right) + 1 \right)}{n}$$
$$+ 2 \cdot \mathbf{E}\left\{ \inf_{l \in \Theta_n} \left(\int |g_{n,l}(x) - m(x)|^2 \mathbf{P}_X(dx) + pen_n(g_{n,l}) + \epsilon_{n,l} \right) \right\}$$

for $n > 1$ and some constant $c_7 > 0$, which does not depend on n, β_n or the parameters of the estimate.

Proof. This lemma follows in a straightforward way from the proof of Theorem 1 in Bagirov et al. (2009). A complete version of the proof is given in the Supplement. □

In order to bound the covering number $\mathcal{N}_1\left(\frac{1}{n \cdot \beta_n}, \mathcal{F}_n, x_1^n \right)$ we will use the following lemma.

Lemma 10 *Let $\max\{K, \beta_n, \gamma_n\} \le n^{c_8}$ and define $\mathcal{F}$ by*

$$\mathcal{F} \;=\; \left\{ f : \mathbb{R}^d \to \mathbb{R} \,:\, f(x) = \sum_{k=0}^{K} a_k \cdot \sigma\left(\sum_{j=1}^{d} b_{k,j} \cdot x^{(j)} + b_{k,0} \right) \quad (x \in \mathbb{R}^d) \right.$$

$$\text{for some } a_k, b_{k,j} \in \mathbb{R} \text{ satisfying } \sum_{k=0}^{K} a_k^2 \leq \gamma_n. \Bigg\}$$

Then we have for any $x_1^n \in (\mathbb{R}^d)^n$*:*

$$\log\left(\mathcal{N}_1\left(\frac{1}{n\cdot\beta_n}, \mathcal{F}, x_1^n\right)\right) \leq c_9 \cdot \log n \cdot K.$$

Proof. Using that

$$\sum_{k=0}^{K} |a_k|^2 \leq \gamma_n$$

implies

$$\sum_{k=0}^{K} |a_k| \leq \sqrt{K+1} \cdot \sqrt{\sum_{k=0}^{K} |a_k|^2} \leq \sqrt{(K+1)\cdot\gamma_n},$$

we can conclude from Lemma 16.6 in Györfi et al. (2002) that we have

$$\begin{aligned}
&\mathcal{N}_1\left(\frac{1}{n\cdot\beta_n}, \mathcal{F}, x_1^n\right) \\
&\leq \left(\frac{e(\sqrt{(K+1)\gamma_n} + 1/(n\cdot\beta_n))}{1/(2\cdot n\cdot\beta_n)}\right)^{K+1} \cdot \left(\mathcal{N}_1\left(\frac{1/(2\cdot n\cdot\beta_n)}{\sqrt{(K+1)\gamma_n} + 1/(n\cdot\beta_n)}, \mathcal{G}, x_1^n\right)\right)^{K+1},
\end{aligned}$$

where

$$\begin{aligned}
\mathcal{G} \quad = \quad \Bigg\{ & g : \mathbb{R}^d \to \mathbb{R} \,:\, g(x) = \sigma\left(\sum_{j=1}^{d} b_j \cdot x^{(j)} + b_0\right) \quad (x \in \mathbb{R}^d) \\
& \text{for some } b_0, \ldots, b_d \in \mathbb{R} \Bigg\}.
\end{aligned}$$

By Lemma 16.3, Theorem 9.5 and Theorem 9.4 in Györfi et al. (2002) we get

$$\begin{aligned}
&\mathcal{N}_1\left(\frac{1/(2\cdot n\cdot\beta_n)}{\sqrt{(K+1)\gamma_n} + 1/(n\cdot\beta_n)}, \mathcal{G}, x_1^n\right) \\
&\leq 3 \cdot \Bigg(2e \cdot (2\cdot n\cdot\sqrt{K+1}\cdot\beta_n\cdot\sqrt{\gamma_n} + 2) \\
&\qquad \cdot \log\Big(3e\cdot(2\cdot n\cdot\sqrt{K+1}\cdot\beta_n\cdot\sqrt{\gamma_n}+2)\Big)\Bigg)^{d+2},
\end{aligned}$$

which implies the assertion. □

5.4. Proof of Theorem 1

From the definition of the estimate and from Lemma 8 we get

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}|Y_i-\tilde{m}_n(X_i)|^2\cdot I_{\{|Y_i|\le\beta_n \text{ for all } i\in\{1,\dots,n\}\}}\\
&\le \min_{\mathbf{a}\in\mathbb{R}^{K+1},l=1,\dots,L_n}\left(\frac{1}{n}\sum_{i=1}^{n}|Y_i-f_{net,(\mathbf{a},(\mathbf{b}^{(l)})^{(0)})}(X_i)|^2+\frac{c_1}{n}\sum_{k=0}^{K\cdot r}a_k^2+\epsilon_n\right)
\end{aligned}$$

where

$$\begin{aligned}
\epsilon_n &= \left(1-\frac{c_1}{6\cdot K\cdot n}\right)^{t_n}\cdot\beta_n^2+(2\cdot\beta_n+1)\cdot\exp\left(-\frac{\sqrt{d}\cdot A\cdot\rho_n}{4\cdot(n+1)\cdot(K-1)}\right)\\
&\qquad+\frac{6\cdot K\cdot n}{c_1}\cdot 3\cdot\exp\left(-\frac{\sqrt{d}\cdot A\cdot\rho_n}{4(n+1)\cdot(K-1)}\right)\\
&\le \exp\left(-\frac{c_1}{6}\cdot(\log n)^2\right)\cdot\beta_n^2+(2\cdot\beta_n+1)\cdot\exp\left(-\frac{\sqrt{d}\cdot A\cdot n}{8}\right)\\
&\qquad+\frac{6\cdot K\cdot n}{c_1}\cdot 3\cdot\exp\left(-\frac{\sqrt{d}\cdot A}{8}\cdot n\right).
\end{aligned}$$

Application of Lemma 9 and of Lemma 10 yields

$$\begin{aligned}
&\mathbf{E}\int|m_n(x)-m(x)|^2\mathbf{P}_X(dx)\\
&\le c_{10}\cdot\frac{(\log n)^3\cdot K}{n}\\
&\quad+2\cdot\mathbf{E}\left\{\min_{\mathbf{a}\in\mathbb{R}^{K+1},l=1,\dots,L_n}\left(\int|f_{net,(\mathbf{a},(\mathbf{b}^{(l)})^{(0)})}(x)-m(x)|^2\mathbf{P}_X(dx)+\frac{c_1}{n}\sum_{k=0}^{K\cdot r}a_k^2\right)\right\}\\
&\quad+2\cdot\exp\left(-\frac{c_1}{6}\cdot(\log n)^2\right)\cdot\beta_n^2+2\cdot(2\cdot\beta_n+1)\cdot\exp\left(-\frac{\sqrt{d}\cdot A\cdot n}{8}\right)\\
&\quad+2\cdot\frac{6\cdot K\cdot n}{c_1}\cdot 3\cdot\exp\left(-\frac{\sqrt{d}\cdot A}{8}\cdot n\right).
\end{aligned}$$

The (p,C)-smoothness of the f_k implies

$$\begin{aligned}
|m(x)-\sum_{s=1}^{r}f_s(\bar{\mathbf{c}}_s^Tx)| &= |\sum_{s=1}^{r}f_s(\mathbf{c}_s^Tx)-\sum_{s=1}^{r}f_s(\bar{\mathbf{c}}_s^Tx)|\\
&\le \sum_{s=1}^{r}C\cdot|\mathbf{c}_s^Tx-\bar{\mathbf{c}}_s^Tx|^p
\end{aligned}$$

$$\leq \quad r \cdot C \cdot \|x\|^p \cdot \max_{s=1,\dots,r} \|\mathbf{c}_s - \bar{\mathbf{c}}_s\|^p.$$

Together with Lemma 5 this implies

$$\begin{aligned}
&\mathbf{E} \int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx) \\
&\leq c_{11} \cdot \frac{(\log n)^3 \cdot K}{n} + c_{12} \cdot r^2 \cdot C^2 \cdot \frac{1}{K^{2p}} + c_{13} \cdot \mathbf{E} \left\{ \min_{l=1,\dots,L_n} \max_{s=1,\dots,r} \|\mathbf{c}_s - \bar{\mathbf{c}}_s^{(l)}\|^{2p} \right\}.
\end{aligned}$$

The definition of K implies

$$c_{11} \cdot \frac{(\log n)^3 \cdot K}{n} + c_{12} \cdot r^2 \cdot C^2 \cdot \frac{1}{K^{2p}} \leq c_{14} \cdot \left(\frac{(\log n)^3}{n} \right)^{\frac{2p}{2p+1}},$$

hence it remains to show that we also have

$$\mathbf{E} \left\{ \min_{l=1,\dots,L_n} \max_{s=1,\dots,r} \|\mathbf{c}_s - \bar{\mathbf{c}}_s^{(l)}\|^{2p} \right\} \leq c_{15} \cdot \left(\frac{(\log n)^3}{n} \right)^{\frac{2p}{2p+1}}.$$

By the random choice of the $\mathbf{c}_s^{(l)}$ we know for any $t \in (0, 1]$

$$\begin{aligned}
\mathbf{P} \left\{ \min_{l=1,\dots,L_n} \max_{s=1,\dots,r} \|\mathbf{c}_s - \bar{\mathbf{c}}_s^{(l)}\| > t \right\} &= \prod_{i=1}^{L_n} \left(1 - \mathbf{P} \left\{ \max_{s=1,\dots,r} \|\mathbf{c}_s - \bar{\mathbf{c}}_s^{(i)}\| \leq t \right\} \right)) \\
&\leq \left(1 - \left(\frac{t}{2} \right)^{r \cdot d} \right)^{L_n}
\end{aligned}$$

from which we conclude

$$\begin{aligned}
&\mathbf{E} \left\{ \min_{l=1,\dots,L_n} \max_{s=1,\dots,r} \|\mathbf{c}_s - \bar{\mathbf{c}}_s^{(l)}\|^{2p} \right\} \\
&\leq \left(\frac{(\log n)^3}{n} \right)^{\frac{2p}{2p+1}} + (2d)^{2p} \cdot \mathbf{P} \left\{ \min_{l=1,\dots,L_n} \max_{k=1,\dots,r} \|\mathbf{c}_k - \bar{\mathbf{c}}_k^{(l)}\| > \left(\frac{(\log n)^3}{n} \right)^{\frac{1}{2p+1}} \right\} \\
&\leq \left(\frac{(\log n)^3}{n} \right)^{\frac{2p}{2p+1}} + (2d)^{2p} \cdot \left(1 - \frac{1}{2^{r \cdot d}} \cdot \left(\frac{(\log n)^3}{n} \right)^{\frac{r \cdot d}{2p+1}} \right)^{L_n} \\
&\leq c_{15} \cdot \left(\frac{(\log n)^3}{n} \right)^{\frac{2p}{2p+1}},
\end{aligned}$$

where the last inequality follows from

$$L_n \geq (\log n)^2 \cdot \left(\frac{n}{(\log n)^3} \right)^{\frac{r \cdot d}{2p+1}}.$$

Summarizing the above results we get the assertion. □

References

[1] Bagirov, A. M., Clausen, C., and Kohler, M. (2009). Estimation of a regression function by maxima of minima of linear functions. *IEEE Transactions on Information Theory*, **55**, pp. 833-845.

[2] Barron, A. R. (1993). Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information Theory*, **39**, pp. 930-944.

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

[4] Bauer, B., and Kohler, M. (2017). On deep learning as a remedy for the curse of dimensionality in nonparametric regression. To appear in *Annals of Statistics*.

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

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

[7] Dippon, J. (1998). Globally Convergent Stochastic Optimization with Optimal Asymptotic Distribution. *Journal of Applied Probability* **35**, pp. 395-406.

[8] Dippon, J., and Fabian, V. (1994). Stochastic approximation of global minimum points. *Journal of Statistical Planning and Inference* **41**, pp. 327-347.

[9] Eckle, K., and Schmidt-Hieber, J. (2018). A comparison of deep networks with ReLU activation function and linear spline-type methods. *arXiv:1804.02253v2*.

[10] Fabian, V. (1994). Comment on White (1989). *Journal of the American Statistical Association*, **89**, p. 1571.

[11] Friedman, J. H., and Stuetzle, W. (1981). Projection pursuit regression. *Journal of the American Statistical Association*, **76**, pp. 817-823.

[12] Györfi, L., Kohler, M., Krzyżak, A., and Walk, H. (2002). *A Distribution–Free Theory of Nonparametric Regression*. Springer.

[13] Härdle, W. and Stoker, T. M. (1989). Investigating smooth multiple regression by the method of average derivatives. *Journal of the American Statistical Association*, **84**, pp. 986-995.

[14] Härdle, W., Hall, P. and Ichimura, H. (1993). Optimal smoothing in single-index models. *Annals of Statistics*, **21**, pp. 157-178.

[15] Horowitz, J. L., and Mammen, E. (2007). Rate-optimal estimation for a general class of nonparametric regression models with unknown link functions. *Annals of Statistics*, **35**, pp. 2589-2619.

[16] Huber, P. J. (1985). Projection pursuit. *Annals of Statistics*, **13**, pp. 435-475.

[17] Imaizumi, M., and Fukamizu, K. (2018). Deep Neural Networks Learn Non-Smooth Functions Effectively. *arXiv:1802.04474*.

[18] Karimi, H., Nutinie, J., and Schmidt, M. (2018). Linear Convergence of Gradient and Proximal-Gradient Methods Under the Polyak-Lojasiewicz Condition. arXiv: 1608.04636v3.

[19] Kohler, M., and Krzyżak, A. (2017). Nonparametric regression based on hierarchical interaction models. *IEEE Transaction on Information Theory*, **63**, pp. 1620-1630.

[20] Kohler, M., Krzyżak, A., and Langer, S. (2019). Deep learning and MARS: a connection. Submitted for publication.

[21] Kohler, M., and Langer, S. (2018). Deep versus deeper learning in nonparametric regression. Submitted for publication.

[22] Kong, E. and Xia, Y. (2007). Variable selection for the single-index model *Biometrika*, **94**, pp. 217-229.

[23] Kushner, H., and Yin, G. (2003). *Stochastic Approximation and Recursive Algorithms and Applications (2nd ed.)*. Springer, New York, USA.

[24] Lepski, O and Serdyukova, O. (2014). Adaptive estimation under single-index constraint in a regression model. *Annals of Statistics*, **42**, pp. 1-28.

[25] Luenberger, D., and Ye, Y. (2016). *Linear and Nonlinear Programming (4th ed.)*. Springer Science + Business Media, New York, USA.

[26] McCaffrey, D. F., and Gallant, A. R. (1994). Convergence rates for single hidden layer feedforward networks. *Neural Networks*, **7**, pp. 147-158.

[27] Poljak, B. T. (1981). Iterative algorithms for singular minimization problems. *Nonlinear Programming*, **4**, pp. 147-166.

[28] Schmidhuber, J. (2015). Deep learning in neural networks: an overview. *Neural Networks*, **61**, pp. 85-117.

[29] Schmidt-Hieber, J. (2017). Nonparametric regression using deep neural networks with ReLU activation function. *arXiv:1708.06633v2*.

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

[31] Stone, C. J. (1985). Additive regression and other nonparametric models. *Annals of Statistics*, **13**, pp. 689-705.

[32] Stone, C. J. (1994). The use of polynomial splines and their tensor products in multivariate function estimation. *Annals of Statistics*, **22**, pp. 118-184.

[33] White, H. (1989). Some asymptotic results for learning in single hidden-layer feed-forward network models. *Journal of the American Statistical association*, **84**, pp. 1003-1013. Correction ibid. 87, p. 1252.

[34] Yu, Y., and Ruppert, D. (2002). Penalized Spline Estimation for Partially Linear Single-Index Models. *Journal of the American Statistical Association*, **97**, pp. 1042-1054.

A. Supplementary material

A.1. Proof of Lemma 1.

a) For $s \in [0,1]$ set

$$H(s) = F(\mathbf{a}^{(t)} + s \cdot (\mathbf{a}^{(t+1)} - \mathbf{a}^{(t)})).$$

Then the fundamental theorem of calculus, the chain rule, the Cauchy-Schwarz inequality and assumption (12) imply

$$\begin{aligned}
&F(\mathbf{a}^{(t+1)}) - F(\mathbf{a}^{(t)}) = H(1) - H(0) = \int_0^1 H'(s)\, ds \\
&= \int_0^1 (\nabla_{\mathbf{a}} F)(\mathbf{a}^{(t)} + s \cdot (\mathbf{a}^{(t+1)} - \mathbf{a}^{(t)})) \cdot (\mathbf{a}^{(t+1)} - \mathbf{a}^{(t)})\, ds \\
&= \int_0^1 \Big((\nabla_{\mathbf{a}} F)(\mathbf{a}^{(t)} + s \cdot (\mathbf{a}^{(t+1)} - \mathbf{a}^{(t)})) - (\nabla_{\mathbf{a}} F)(\mathbf{a}^{(t)}) \Big) \cdot (\mathbf{a}^{(t+1)} - \mathbf{a}^{(t)})\, ds \\
&\quad + \int_0^1 (\nabla_{\mathbf{a}} F)(\mathbf{a}^{(t)}) \cdot (\mathbf{a}^{(t+1)} - \mathbf{a}^{(t)})\, ds \\
&\leq \int_0^1 L_n \cdot \|s \cdot (\mathbf{a}^{(t+1)} - \mathbf{a}^{(t)})\| \cdot \|\mathbf{a}^{(t+1)} - \mathbf{a}^{(t)}\|\, ds \\
&\quad + (\nabla_{\mathbf{a}} F)(\mathbf{a}^{(t)}) \cdot (\mathbf{a}^{(t+1)} - \mathbf{a}^{(t)}) \\
&= \frac{L_n}{2} \cdot \|\mathbf{a}^{(t+1)} - \mathbf{a}^{(t)}\|^2 + (\nabla_{\mathbf{a}} F)(\mathbf{a}^{(t)}) \cdot (\mathbf{a}^{(t+1)} - \mathbf{a}^{(t)}).
\end{aligned}$$

Using (10) and (11) we get

$$\begin{aligned}
F(\mathbf{a}^{(t+1)}) - F(\mathbf{a}^{(t)}) &\leq \frac{L_n}{2} \cdot \lambda_n^2 \cdot \|(\nabla_{\mathbf{a}} F)(\mathbf{a}^{(t)})\|^2 - \lambda_n \|(\nabla_{\mathbf{a}} F)(\mathbf{a}^{(t)})\|^2 \\
&= -\frac{1}{2 \cdot L_n} \cdot \|(\nabla_{\mathbf{a}} F)(\mathbf{a}^{(t)})\|^2.
\end{aligned}$$

b) From **a)** and (13) we get

$$\begin{aligned}
&F(\mathbf{a}^{(t+1)}) - F(\mathbf{a}_{opt}) \\
&\leq F(\mathbf{a}^{(t)}) - F(\mathbf{a}_{opt}) - \frac{1}{2 \cdot L_n} \cdot \|(\nabla_{\mathbf{a}} F)(\mathbf{a}^{(t)})\|^2 \\
&\leq F(\mathbf{a}^{(t)}) - F(\mathbf{a}_{opt}) - \frac{1}{2 \cdot L_n} \cdot \rho_n \cdot (F(\mathbf{a}^{(t)}) - F(\mathbf{a}_{opt})) \\
&= \left(1 - \frac{\rho_n}{2 \cdot L_n}\right) \cdot (F(\mathbf{a}^{(t)}) - F(\mathbf{a}_{opt})).
\end{aligned}$$

□

A.2. Proof of Lemma 9

In the proof we use the following error decomposition:

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

$$
\begin{aligned}
&= \Big[\mathbf{E}\Big\{|m_n(X)-Y|^2|\mathcal{D}_n\Big\} - \mathbf{E}\Big\{|m(X)-Y|^2\Big\} \\
&\qquad -\Big(\mathbf{E}\Big\{|m_n(X)-T_{\beta_n}Y|^2|\mathcal{D}_n\Big\} - \mathbf{E}\Big\{|m_{\beta_n}(X)-T_{\beta_n}Y|^2\Big\}\Big)\Big] \\
&\quad + \Bigg[\mathbf{E}\Big\{|m_n(X)-T_{\beta_n}Y|^2|\mathcal{D}_n\Big\} - \mathbf{E}\Big\{|m_{\beta_n}(X)-T_{\beta_n}Y|^2\Big\} \\
&\qquad -2\cdot\frac{1}{n}\sum_{i=1}^{n}\Big(|m_n(X_i)-T_{\beta_n}Y_i|^2 - |m_{\beta_n}(X_i)-T_{\beta_n}Y_i|^2\Big)\Bigg] \\
&\quad + \Bigg[2\cdot\frac{1}{n}\sum_{i=1}^{n}|m_n(X_i)-T_{\beta_n}Y_i|^2 - 2\cdot\frac{1}{n}\sum_{i=1}^{n}|m_{\beta_n}(X_i)-T_{\beta_n}Y_i|^2 \\
&\qquad -\left(2\cdot\frac{1}{n}\sum_{i=1}^{n}|m_n(X_i)-Y_i|^2 - 2\cdot\frac{1}{n}\sum_{i=1}^{n}|m(X_i)-Y_i|^2\right)\Bigg] \\
&\quad + \left[2\left(\frac{1}{n}\sum_{i=1}^{n}|m_n(X_i)-Y_i|^2 - \frac{1}{n}\sum_{i=1}^{n}|m(X_i)-Y_i|^2\right)\right] \\
&= \sum_{i=1}^{4} T_{i,n},
\end{aligned}
$$

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

$$
m_{\beta_n}(x) = \mathbf{E}\Big\{T_{\beta_n}Y|X=x\Big\}.
$$

We start with bounding $T_{1,n}$. By using $a^2-b^2=(a-b)(a+b)$ we get

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

With the Cauchy-Schwarz inequality and

$$
I_{\{|Y|>\beta_n\}} \le \frac{\exp(c_2/2\cdot|Y|^2)}{\exp(c_2/2\cdot\beta_n^2)} \tag{33}
$$

we conclude

$$
\begin{aligned}
|T_{5,n}| &\le \sqrt{\mathbf{E}\big\{|T_{\beta_n}Y-Y|^2\big\}}\cdot\sqrt{\mathbf{E}\big\{|2m_n(X)-Y-T_{\beta_n}Y|^2|\mathcal{D}_n\big\}} \\
&\le \sqrt{\mathbf{E}\big\{|Y|^2\cdot I_{\{|Y|>\beta_n\}}\big\}}\cdot\sqrt{\mathbf{E}\big\{2\cdot|2m_n(X)-T_{\beta_n}Y|^2+2\cdot|Y|^2|\mathcal{D}_n\big\}}
\end{aligned}
$$

$$
\begin{aligned}
&\leq \sqrt{\mathbf{E}\Big\{|Y|^2 \cdot \frac{\exp(c_2/2 \cdot |Y|^2)}{\exp(c_2/2 \cdot \beta_n^2)}\Big\}} \\
&\qquad \cdot \sqrt{\mathbf{E}\big\{2 \cdot |2m_n(X) - T_{\beta_n} Y|^2 \big| \mathcal{D}_n\big\} + 2\mathbf{E}\big\{|Y|^2\big\}} \\
&\leq \sqrt{\mathbf{E}\Big\{|Y|^2 \cdot \exp(c_2/2 \cdot |Y|^2)\Big\}} \cdot \exp\left(-\frac{c_2 \cdot \beta_n^2}{4}\right) \cdot \sqrt{2(3\beta_n)^2 + 2\mathbf{E}\big\{|Y|^2\big\}}.
\end{aligned}
$$

With $x \leq \exp(x)$ for $x \in \mathbb{R}$ we get

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

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

$$
\mathbf{E}\left(\frac{2}{c_2} \cdot \exp\left(c_2/2 \cdot |Y|^2\right) \cdot \exp(c_2/2 \cdot |Y|^2)\right) \leq \mathbf{E}\left(\frac{2}{c_2} \cdot \exp\left(c_2 \cdot |Y|^2\right)\right) \leq c_{16}
$$

which is less than infinity by the assumptions of the lemma. Furthermore the third term is bounded by $\sqrt{18\beta_n^2 + c_{17}}$ because

$$
\mathbf{E}(|Y|^2) \leq \mathbf{E}(1/c_2 \cdot \exp(c_2 \cdot |Y|^2) \leq c_{18} < \infty, \tag{34}
$$

which follows again as above. With the setting $\beta_n = c_3 \cdot \log(n)$ it follows for some constants $c_{19}, c_{20} > 0$ that

$$
|T_{5,n}| \leq \sqrt{c_{16}} \cdot \exp\left(-c_{19} \cdot \log(n)^2\right) \cdot \sqrt{(18 \cdot c_3^2 \cdot (\log n)^2 + c_{18})} \leq c_{20} \cdot \frac{\log(n)}{n}.
$$

From the Cauchy-Schwarz inequality we get

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

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 furthermore m_{β_n} is bounded by β_n. Thus we get for some constant $c_{21} > 0$

$$
\sqrt{\mathbf{E}\left\{\left|m(X) + m_{\beta_n}(X) - Y - T_{\beta_n} Y\right|^2\right\}} \leq c_{21} \cdot \log(n).
$$

Next we consider the first term. With Jensen's inequality it follows that

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

Hence we get

$$T_{6,n} \leq \sqrt{4 \cdot \mathbf{E}\left\{|Y - T_{\beta_n} Y|^2\right\}} \cdot c_{22} \cdot \log(n)$$

and therefore with the calculations from $T_{5,n}$ it follows that $T_{6,n} \leq c_{23} \cdot \log(n)/n$ for some constant $c_{23} > 0$. Altogether we get

$$T_{1,n} \leq c_{24} \cdot \frac{\log(n)}{n}$$

for some constant $c_{24} > 0$.
Next we consider $T_{2,n}$ and conclude for $t > 0$

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

where $T_{\beta_n, supp(X)} \mathcal{F}_n$ is defined as $\left\{T_{\beta_n} f \cdot 1_{supp(X)} : f \in \mathcal{F}_n\right\}$. Theorem 11.4 in Györfi et al. (2002) and the relation

$$\mathcal{N}_1\left(\delta, \left\{\frac{1}{\beta_n} g : g \in \mathcal{G}\right\}, x_1^n\right) \leq \mathcal{N}_1\left(\delta \cdot \beta_n, \mathcal{G}, \|\cdot\|_{\infty, supp(X)}\right)$$

for an arbitrary function space $\mathcal{G}$ and $\delta > 0$ lead to

$$\mathbf{P}\{T_{2,n} > t\} \leq 14 \cdot \sup_{x_1^n} \mathcal{N}_1\left(\frac{t}{80 \cdot \beta_n}, \mathcal{F}_n, \|\cdot\|_{\infty, supp(X)}\right) \cdot \exp\left(-\frac{n}{5136 \cdot \beta_n^2} \cdot t\right).$$

Since the covering number is decreasing in t, we can conclude for $\varepsilon_n \geq \frac{80}{n}$

$$\begin{aligned}
\mathbf{E}(T_{2,n}) & \leq \varepsilon_n + \int_{\varepsilon_n}^{\infty} \mathbf{P}\{T_{2,n} > t\} dt \\
& \leq \varepsilon_n + 14 \cdot \sup_{x_1^n} \mathcal{N}_1\left(\frac{1}{n \cdot \beta_n}, \mathcal{F}_n, x_1^n\right) \cdot \exp\left(-\frac{n}{5136 \cdot \beta_n^2} \cdot \varepsilon_n\right) \cdot \frac{5136 \cdot \beta_n^2}{n}.
\end{aligned}$$

Choosing

$$\varepsilon_n = \frac{5136 \cdot \beta_n^2}{n} \cdot \log\left(14 \cdot \sup_{x_1^n} \mathcal{N}_1\left(\frac{1}{n \cdot \beta_n}, \mathcal{F}_n, x_1^n\right)\right)$$

(which satisfies the necessary condition $\varepsilon_n \geq \frac{80}{n}$ if the constant c_3 in the definition of β_n is not too small) minimizes the right-hand side and implies

$$\mathbf{E}(T_{2,n}) \leq \frac{c_{25} \cdot \log(n)^2 \cdot \log\left(\sup_{x_1^n} \mathcal{N}_1\left(\frac{1}{n \cdot \beta_n}, \mathcal{F}_n, x_1^n\right)\right)}{n}.$$

By bounding $T_{3,n}$ similarly to $T_{1,n}$ we get

$$\mathbf{E}(T_{3,n}) \quad \leq \quad c_{26} \cdot \frac{\log(n)}{n}$$

for some large enough constant $c_{26} > 0$ and hence we get in total

$$\mathbf{E}\left(\sum_{i=1}^{3} T_{i,n}\right) \quad \leq \quad \frac{c_{27} \cdot (\log n)^2 \cdot \left(\log\left(\mathcal{N}_1\left(\frac{1}{n \cdot \beta_n}, \mathcal{F}_n, \|\cdot\|_{\infty, supp(X)}\right)\right) + 1\right)}{n}$$

for some sufficient large constant $c_{27} > 0$.

We finish the proof by bounding $T_{4,n}$. Let A_n be the event, that there exists $i \in \{1, ..., n\}$ such that $|Y_i| > \beta_n$ and let I_{A_n} be the indicator function of A_n. Then we get

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

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

$$\begin{aligned}
\frac{1}{2} \cdot T_{7,n} &\leq \sqrt{\mathbf{E}\left(\left(|m_n(X_1) - Y_1|^2\right)^2\right)} \cdot \sqrt{\mathbf{P}(A_n)} \\
&\leq \sqrt{\mathbf{E}\left(\left(2|m_n(X_1)|^2 + 2|Y_1|^2\right)^2\right)} \cdot \sqrt{n \cdot \mathbf{P}\{|Y_1| > \beta_n\}} \\
&\leq \sqrt{\mathbf{E}\left(8|m_n(X_1)|^4 + 8|Y_1|^4\right)} \cdot \sqrt{n \cdot \frac{\mathbf{E}\left(\exp(c_2 \cdot |Y_1|^2)\right)}{\exp(c_2 \cdot \beta_n^2)}},
\end{aligned}$$

where the last inequality follows as in the proof of inequality (33). With $x \leq \exp(x)$ for $x \in \mathbb{R}$ 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 assumption (7) of the lemma. Furthermore $||m_n||_\infty$ is bounded by β_n and therefore the first factor is bounded by

$$c_{28} \cdot \beta_n^2 = c_{29} \cdot (\log n)^2$$

for some constant $c_{29} > 0$. The second factor is bounded by $1/n$, because by the assumptions of the lemma $\mathbf{E}\left(\exp\left(c_2 \cdot |Y_1|^2\right)\right)$ is bounded by some constant $c_{30} < \infty$ and hence we get

$$\sqrt{n \cdot \frac{\mathbf{E}\left(\exp(c_2 \cdot |Y_1|^2)\right)}{\exp(c_2 \cdot \beta_n^2)}} \quad \leq \quad \sqrt{n} \cdot \frac{\sqrt{c_{30}}}{\sqrt{\exp(c_2 \cdot \beta_n^2)}} \leq \frac{\sqrt{n} \cdot \sqrt{c_{30}}}{\exp((c_2 \cdot c_3^2 \cdot (\log n)^2)/2)}.$$

Since $\exp(-c \cdot \log(n)^2) = O(n^{-2})$ for any $c > 0$, we get altogether

$$T_{7,n} \quad \leq \quad c_{31} \cdot \frac{(\log n)^2 \sqrt{n}}{n^2} \leq c_{32} \cdot \frac{(\log n)^2}{n}.$$

With the definition of A_n^c and $\tilde{m}_n$ defined as in the assumptions of this lemma we conclude

$$\begin{aligned} T_{8,n} &\leq 2 \cdot \mathbf{E}\left(\frac{1}{n}\sum_{i=1}^{n} |\tilde{m}_n(X_i) - Y_i|^2 \cdot I_{A_n^c} - \frac{1}{n}\sum_{i=1}^{n} |m(X_i) - Y_i|^2\right) \\ &\leq 2 \cdot \mathbf{E}\left(\min_{l \in \Theta_n} \frac{1}{n}\sum_{i=1}^{n} |g_{l,n}(X_i) - Y_i|^2 + pen_n(g_{n,l}) + \epsilon_{n,l} - \frac{1}{n}\sum_{i=1}^{n} |m(X_i) - Y_i|^2\right) \end{aligned}$$

because $|T_\beta z - y| \leq |z - y|$ holds for $|y| \leq \beta$. Hence by the independence of $g_{n,l}$ and $pen_n(g_{n,l})$ of $(X_1, Y_1), \ldots, (X_n, Y_n)$ we get that

$$\begin{aligned} &\mathbf{E}(T_{4,n}) \\ &\leq c_{32} \cdot \frac{(\log n)^2}{n} \\ &\quad + 2 \cdot \mathbf{E}\left(\min_{l \in \Theta_n} \frac{1}{n}\sum_{i=1}^{n} |g_{l,n}(X_i) - Y_i|^2 + pen_n(g_{n,l}) + \epsilon_{n,l} - \frac{1}{n}\sum_{i=1}^{n} |m(X_i) - Y_i|^2\right) \\ &\leq c_{32} \cdot \frac{(\log n)^2}{n} + 2 \cdot \mathbf{E}\left(\min_{l \in \Theta_n}\left(\int |g_{l,n}(x) - m(x)|^2 \mathbf{P}_X(dx) + 2 \cdot pen_n(g_{n,l}) + \epsilon_{n,l}\right)\right) \end{aligned}$$

holds. In combination with the other considerations in the proof this implies the assertion of Lemma 6. □

Aline Braun
Fachbereich Mathematik, Technische Universitt Darmstadt
E-Mail: braun@mathematik.tu-darmstadt.de

Michael Kohler
Fachbereich Mathematik, Technische Universitt Darmstadt
E-Mail: kohler@mathematik.tu-darmstadt.de

Harro Walk
Fachbereich Mathematik, Universität Stuttgart, Stuttgart, Germany
E-Mail: walk@mathematik.uni-stuttgart.de

Erschienene Preprints ab Nummer 2012-001

Komplette Liste: `http://www.mathematik.uni-stuttgart.de/preprints`

2019-003 *Braun, A.; Kohler, M.; Walk, H.:* On the rate of convergence of a neural network regression estimate learned by gradient descent

2019-002 *Györfi, L.; Henze, N.; Walk, H.:* The limit distribution of the maximum probability nearest neighbor ball

2019-001 *Gaspoz, F.D.; Heine, C.-J.; Siebert, K.G.:* An Alternative Proof of H^1-Stability of the L_2-Projection on Graded Meshes

2018-003 *Kollross, A.:* Octonions, triality, the exceptional Lie algebra F4, and polar actions on the Cayley hyperbolic plane

2018-002 *Diaź-Ramos, J.C.; Domínguez-Vázquez, M.; Kollross, A.:* On homogeneous manifolds whose isotropy actions are polar

2018-001 *Grundhöfer, T.; Stroppel, M.; Van Maldeghem, H.:* Embeddings of hermitian unitals into pappian projective planes

2017-011 *Hansmann, M.; Kohler, M.; Walk, H.:* On the strong universal consistency of local averaging regression estimates

2017-010 *Devroye, L.; Györfi, L.; Lugosi, G.; Walk, H.:* A nearest neighbor estimate of a regression functional

2017-009 *Steinke, G.; Stroppel, M.:* On elation Laguerre planes with a two-transitive orbit on the set of generators

2017-008 *Steinke, G.; Stroppel, M.:* Laguerre planes and shift planes

2017-007 *Blunck, A.; Knarr, N.; Stroppel, B.; Stroppel, M.:* Transitive groups of similitudes generated by octonions

2017-006 *Blunck, A.; Knarr, N.; Stroppel, B.; Stroppel, M.:* Clifford parallelisms defined by octonions

2017-005 *Knarr, N.; Stroppel, M.:* Subforms of Norm Forms of Octonion Fields

2017-004 *Apprich, C.; Dieterich, A.; Höllig, K.; Nava-Yazdani, E.:* Cubic Spline Approximation of a Circle with Maximal Smoothness and Accuracy

2017-003 *Fischer, S.; Steinwart, I.:* Sobolev Norm Learning Rates for Regularized Least-Squares Algorithm

2017-002 *Farooq, M.; Steinwart, I.:* Learning Rates for Kernel-Based Expectile Regression

2017-001 *Bauer, B.; Devroye, L; Kohler, M.; Krzyzak, A.; Walk, H.:* Nonparametric Estimation of a Function From Noiseless Observations at Random Points

2016-006 *Devroye, L.; Györfi, L.; Lugosi, G.; Walk, H.:* On the measure of Voronoi cells

2016-005 *Kohls, C.; Kreuzer, C.; Rösch, A.; Siebert, K.G.:* Convergence of Adaptive Finite Elements for Optimal Control Problems with Control Constraints

2016-004 *Blaschzyk, I.; Steinwart, I.:* Improved Classification Rates under Refined Margin Conditions

2016-003 *Feistauer, M.; Roskovec, F.; Sändig, AM.:* Discontinuous Galerkin Method for an Elliptic Problem with Nonlinear Newton Boundary Conditions in a Polygon

2016-002 *Steinwart, I.:* A Short Note on the Comparison of Interpolation Widths, Entropy Numbers, and Kolmogorov Widths

2016-001 *Köster, I.:* Sylow Numbers in Spectral Tables

2015-016 *Hang, H.; Steinwart, I.:* A Bernstein-type Inequality for Some Mixing Processes and Dynamical Systems with an Application to Learning

2015-015 *Steinwart, I.:* Representation of Quasi-Monotone Functionals by Families of Separating Hyperplanes

2015-014 *Muhammad, F.; Steinwart, I.:* An SVM-like Approach for Expectile Regression

2015-013 *Nava-Yazdani, E.:* Splines and geometric mean for data in geodesic spaces

2015-012 *Kimmerle, W.; Köster, I.:* Sylow Numbers from Character Tables and Group Rings

2015-011 *Györfi, L.; Walk, H.:* On the asymptotic normality of an estimate of a regression functional

2015-010 *Gorodski, C, Kollross, A.:* Some remarks on polar actions

2015-009 *Apprich, C.; Höllig, K.; Hörner, J.; Reif, U.:* Collocation with WEB–Splines

2015-008 *Kabil, B.; Rodrigues, M.:* Spectral Validation of the Whitham Equations for Periodic Waves of Lattice Dynamical Systems

2015-007 *Kollross, A.:* Hyperpolar actions on reducible symmetric spaces

2015-006 *Schmid, J.; Griesemer, M.:* Well-posedness of Non-autonomous Linear Evolution Equations in Uniformly Convex Spaces

2015-005 *Hinrichs, A.; Markhasin, L.; Oettershagen, J.; Ullrich, T.:* Optimal quasi-Monte Carlo rules on higher order digital nets for the numerical integration of multivariate periodic functions

2015-004 *Kutter, M.; Rohde, C.; Sändig, A.-M.:* Well-Posedness of a Two Scale Model for Liquid Phase Epitaxy with Elasticity

2015-003 *Rossi, E.; Schleper, V.:* Convergence of a numerical scheme for a mixed hyperbolic-parabolic system in two space dimensions

2015-002 *Döring, M.; Györfi, L.; Walk, H.:* Exact rate of convergence of kernel-based classification rule

2015-001 *Kohler, M.; Müller, F.; Walk, H.:* Estimation of a regression function corresponding to latent variables

2014-021 *Neusser, J.; Rohde, C.; Schleper, V.:* Relaxed Navier-Stokes-Korteweg Equations for Compressible Two-Phase Flow with Phase Transition

2014-020 *Kabil, B.; Rohde, C.:* Persistence of undercompressive phase boundaries for isothermal Euler equations including configurational forces and surface tension

2014-019 *Bilyk, D.; Markhasin, L.:* BMO and exponential Orlicz space estimates of the discrepancy function in arbitrary dimension

2014-018 *Schmid, J.:* Well-posedness of non-autonomous linear evolution equations for generators whose commutators are scalar

2014-017 *Margolis, L.:* A Sylow theorem for the integral group ring of $PSL(2, q)$

2014-016 *Rybak, I.; Magiera, J.; Helmig, R.; Rohde, C.:* Multirate time integration for coupled saturated/unsaturated porous medium and free flow systems

2014-015 *Gaspoz, F.D.; Heine, C.-J.; Siebert, K.G.:* Optimal Grading of the Newest Vertex Bisection and H^1-Stability of the L_2-Projection

2014-014 *Kohler, M.; Krzyżak, A.; Walk, H.:* Nonparametric recursive quantile estimation

2014-013 *Kohler, M.; Krzyżak, A.; Tent, R.; Walk, H.:* Nonparametric quantile estimation using importance sampling

2014-012 *Györfi, L.; Ottucsák, G.; Walk, H.:* The growth optimal investment strategy is secure, too.

2014-011 *Györfi, L.; Walk, H.:* Strongly consistent detection for nonparametric hypotheses

2014-010 *Köster, I.:* Finite Groups with Sylow numbers $\{q^x, a, b\}$

2014-009 *Kahnert, D.:* Hausdorff Dimension of Rings

2014-008 *Steinwart, I.:* Measuring the Capacity of Sets of Functions in the Analysis of ERM

2014-007 *Steinwart, I.:* Convergence Types and Rates in Generic Karhunen-Loève Expansions with Applications to Sample Path Properties

2014-006 *Steinwart, I.; Pasin, C.; Williamson, R.; Zhang, S.:* Elicitation and Identification of Properties

2014-005 *Schmid, J.; Griesemer, M.:* Integration of Non-Autonomous Linear Evolution Equations

2014-004 *Markhasin, L.:* L_2- and $S^r_{p,q}B$-discrepancy of (order 2) digital nets

2014-003 *Markhasin, L.:* Discrepancy and integration in function spaces with dominating mixed smoothness

2014-002 *Eberts, M.; Steinwart, I.:* Optimal Learning Rates for Localized SVMs

2014-001 *Giesselmann, J.:* A relative entropy approach to convergence of a low order approximation to a nonlinear elasticity model with viscosity and capillarity

2013-016 *Steinwart, I.:* Fully Adaptive Density-Based Clustering

2013-015 *Steinwart, I.:* Some Remarks on the Statistical Analysis of SVMs and Related Methods

2013-014 *Rohde, C.; Zeiler, C.:* A Relaxation Riemann Solver for Compressible Two-Phase Flow with Phase Transition and Surface Tension

2013-013 *Moroianu, A.; Semmelmann, U.:* Generalized Killling spinors on Einstein manifolds

2013-012 *Moroianu, A.; Semmelmann, U.:* Generalized Killing Spinors on Spheres

2013-011 *Kohls, K; Rösch, A.; Siebert, K.G.:* Convergence of Adaptive Finite Elements for Control Constrained Optimal Control Problems

2013-010 *Corli, A.; Rohde, C.; Schleper, V.:* Parabolic Approximations of Diffusive-Dispersive Equations

2013-009 *Nava-Yazdani, E.; Polthier, K.:* De Casteljau's Algorithm on Manifolds

2013-008 *Bächle, A.; Margolis, L.:* Rational conjugacy of torsion units in integral group rings of non-solvable groups

2013-007 *Knarr, N.; Stroppel, M.J.:* Heisenberg groups over composition algebras

2013-006 *Knarr, N.; Stroppel, M.J.:* Heisenberg groups, semifields, and translation planes

2013-005 *Eck, C.; Kutter, M.; Sändig, A.-M.; Rohde, C.:* A Two Scale Model for Liquid Phase Epitaxy with Elasticity: An Iterative Procedure

2013-004 *Griesemer, M.; Wellig, D.:* The Strong-Coupling Polaron in Electromagnetic Fields

2013-003 *Kabil, B.; Rohde, C.:* The Influence of Surface Tension and Configurational Forces on the Stability of Liquid-Vapor Interfaces

2013-002 *Devroye, L.; Ferrario, P.G.; Györfi, L.; Walk, H.:* Strong universal consistent estimate of the minimum mean squared error

2013-001 *Kohls, K.; Rösch, A.; Siebert, K.G.:* A Posteriori Error Analysis of Optimal Control Problems with Control Constraints

2012-013 *Diaz Ramos, J.C.; Dominguez Vázquez, M.; Kollross, A.:* Polar actions on complex hyperbolic spaces

2012-012 *Moroianu; A.; Semmelmann, U.:* Weakly complex homogeneous spaces

2012-011 *Moroianu; A.; Semmelmann, U.:* Invariant four-forms and symmetric pairs

2012-010 *Hamilton, M.J.D.:* The closure of the symplectic cone of elliptic surfaces

2012-009 *Hamilton, M.J.D.:* Iterated fibre sums of algebraic Lefschetz fibrations

2012-008 *Hamilton, M.J.D.:* The minimal genus problem for elliptic surfaces

2012-007 *Ferrario, P.:* Partitioning estimation of local variance based on nearest neighbors under censoring

2012-006 *Stroppel, M.:* Buttons, Holes and Loops of String: Lacing the Doily

2012-005 *Hantsch, F.:* Existence of Minimizers in Restricted Hartree-Fock Theory

2012-004 *Grundhöfer, T.; Stroppel, M.; Van Maldeghem, H.:* Unitals admitting all translations

2012-003 *Hamilton, M.J.D.:* Representing homology classes by symplectic surfaces

2012-002 *Hamilton, M.J.D.:* On certain exotic 4-manifolds of Akhmedov and Park

2012-001 *Jentsch, T.:* Parallel submanifolds of the real 2-Grassmannian