

Nonparametric regression based on hierarchical interaction models *

Michael Kohler¹ and Adam Krzyżak^{2,†}

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

² *Department of Computer Science and Software Engineering, Concordia University, 1455 De Maisonneuve Blvd. West, Montreal, Quebec, Canada H3G 1M8, email: krzyzak@cs.concordia.ca*

June 11, 2015

Abstract

In this paper we introduce so-called hierarchical interaction models where we assume that the computation of the value of a function $m : \mathbb{R}^d \rightarrow \mathbb{R}$ is done in several layers, where in each layer a function of at most d^* inputs computed by the previous layer is evaluated. We investigate two different regression estimates based on polynomial splines and on neural networks and show that if the regression function satisfies a hierarchical interaction model and all occurring functions in the model are smooth, the rate of convergence of these estimates depends on d^* (and not on d). Hence in this case the estimates can achieve good rate of convergence even for large d and are in this sense able to circumvent the so-called curse of dimensionality.

AMS classification: Primary 62G08; secondary 62G20.

Key words and phrases: Curse of dimensionality, dimension reduction, interaction models, L_2 error, nonparametric regression, projection pursuit, rate of convergence.

1. Introduction

In nonparametric regression a random vector (X, Y) with values in $\mathbb{R}^d \times \mathbb{R}$ satisfying $\mathbf{E}Y^2 < \infty$ is given and the goal is to predict the value of Y given the value of X . If the main aim of the analysis is minimization of the mean squared error or L_2 risk, then a function $m^* : \mathbb{R}^d \rightarrow \mathbb{R}$ is sought satisfying

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

Let $m : \mathbb{R}^d \rightarrow \mathbb{R}$, $m(x) = \mathbf{E}\{Y|X = x\}$ be the so-called regression function. Since

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

*Running title: *Hierarchical interaction models*

[†]Corresponding author. Tel: +1-514-848-2424 ext. 3007, Fax: +1-514-848-2830

(cf., e.g., Section 1.1 in Györfi et al. (2002)), the regression function is the optimal predictor $m^* = m$, and any function $f : \mathbb{R}^d \rightarrow \mathbb{R}$ is a good predictor in the sense that its L_2 risk is close to the optimal value if and only if the L_2 error

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

is small. In nonparametric regression a set of data

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

is given, where (X, Y) , (X_1, Y_1) , (X_2, Y_2) , \dots are independent and identically distributed random variables, and the aim is to construct a regression estimate $m_n(\cdot) = m_n(\cdot, \mathcal{D}_n)$ such that its L_2 error

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

is small. See Györfi et al. (2002) for a systematic and rigorous coverage of nonparametric regression estimation.

In order to derive nontrivial results on the rate of convergence of the expected L_2 error, it is necessary to impose smoothness assumptions on m (cf., e.g., Theorem 3.1 in Györfi et al. (2002)). It was shown in Stone (1982) that the optimal minimax rate of convergence for estimation of (p, C) -smooth regression function (where roughly speaking, see below for the exact definition, the regression function is p -times continuously differentiable) is

$$n^{-\frac{2p}{2p+d}}.$$

If d is large compared to p , then this rate of convergence is rather slow, which is a consequence of the fact, that high-dimensional regression problems are especially difficult to solve due to so-called curse of dimensionality. But unfortunately, most applications are high-dimensional problems and hence very hard to solve. The only way to circumvent this curse of dimensionality is to impose additional assumptions on the regression function in order to derive better rates of convergence.

Stone (1985) proposed to impose an additivity condition on the structure of the regression function. He assumed that

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

for (p, C) -smooth univariate functions $m_1, \dots, m_d : \mathbb{R} \rightarrow \mathbb{R}$, and he was able to show that in this case $n^{-2p/(2p+1)}$ is the optimal minimax rate of convergence. A generalization of this approach to so-called interaction models was presented in Stone (1994). Here it was assumed that for some $d^* \in \{1, \dots, d\}$ the regression function satisfies

$$m(x) = \sum_{I \subseteq \{1, \dots, d\}, |I|=d^*} m_I(x_I),$$

where $|I|$ denotes the cardinality of the set I , m_I are (p, C) -smooth functions defined on $\mathbb{R}^{|I|}$ and for $x = (x^{(1)}, \dots, x^{(d)})^T$ and $I = \{i_1, \dots, i_{d^*}\}$ with $1 \leq i_1 < \dots < i_{d^*} \leq d$ we

set $x_I = (x^{(i_1)}, \dots, x^{(i_{d^*})})^T$. In other words, it is assumed that the regression function is a sum of (p, C) -smooth functions where each function in the sum depends on at most d^* of the components of x . Under this assumption it was shown that $n^{-2p/(2p+d^*)}$ is the optimal minimax rate of convergence.

Other assumptions which yield good rates of convergence even for high dimensional data include single index models and projection pursuit. In single index models it is assumed that

$$m(x) = g(a^T x) \quad (x \in \mathbb{R}),$$

where $g : \mathbb{R} \rightarrow \mathbb{R}$ is an univariate function and $a \in \mathbb{R}^d$ is a d -dimensional vector (cf., e.g., Härdle, Hall and Ichimura (1993), Härdle and Stoker (1989), Yu and Ruppert (2002) and Kong and Xia (2007)). In projection pursuit the regression function is allowed to be a sum of functions of the above form, i.e.,

$$m(x) = \sum_{k=1}^K g_k(a_k^T x) \quad (x \in \mathbb{R}),$$

where $K \in \mathbb{N}$ is a natural number, $g_k : \mathbb{R} \rightarrow \mathbb{R}$ are univariate functions and $a_k \in \mathbb{R}^d$ are d -dimensional vectors (cf., e.g., Friedman and Stuetzle (1981)). If the univariate functions above are (p, C) -smooth, then corresponding regression estimates can achieve under these assumptions the corresponding univariate rates of convergence (cf., e.g., Chapter 22 in Györfi et al. (2002)).

A mixture of parametric and nonparametric approach is achieved in semiparametric models. Here it is assumed that for a part of the components of x the influence on the regression function is known and is described by a parametric model (e.g., a linear model), and only the remaining part is estimated nonparametrically (cf., e.g., Härdle et al. (2004)). Under this assumption the corresponding estimates are able to achieve rates of convergence corresponding to d^* dimensional problems, where d^* is the number of components of x for which a parametric model is not given.

In any application these estimates achieve good rates of convergence only if the imposed assumptions are satisfied. Our research in this paper is motivated by applications in connection with complex technical systems, which are constructed in a modular form (in particular a load bearing structure studied currently by the Collaborative Research Centre 805 at the Technische Universität Darmstadt). If such systems are constructed in a modular form, then it seems to be realistic to model the outcome of the system as a function of the results of the modular parts of it, where each of the modular parts computes a function depending only on a few of the components of the high-dimensional input. In the simplest case we formulate this by assuming that our regression function satisfies

$$m(x) = g(f_1(x_{I_1}), \dots, f_{d^*}(x_{I_{d^*}})) \quad (x \in \mathbb{R}^d),$$

where $d^* \in \{1, \dots, d\}$, $I_1, \dots, I_{d^*} \subseteq \{1, \dots, d\}$ are sets of cardinality d^* and g, f_1, \dots, f_{d^*} are (p, C) -smooth functions defined on \mathbb{R}^{d^*} . The corresponding general assumption can be found in Section 2. In the general assumption the above model is recursively applied, which is reasonable especially if we consider a complex technical system constructed in

a modular form, where each modular part may be again a complex system constructed in a similar modular form. Under the above assumption we show that suitable defined spline and neural network estimates achieve (up to some logarithmic factor) the rate of convergence $n^{-2p/(2p+d^*)}$.

Throughout the paper the following notation is used: The sets of natural numbers, natural numbers including 0, integers, non-negative real numbers and real numbers are denoted by \mathbb{N} , \mathbb{N}_0 , \mathbb{Z} , \mathbb{R}_+ and \mathbb{R} , resp. Let $D \subseteq \mathbb{R}^d$ and let $f : \mathbb{R}^d \rightarrow \mathbb{R}$ be a real-valued function defined on \mathbb{R}^d . We write $x = \arg \max_{z \in D} f(z)$ if $\max_{z \in D} f(z)$ exists and if x satisfies

$$x \in D \quad \text{and} \quad f(x) = \max_{z \in D} f(z).$$

The Euclidean and the supremum norms of $x \in \mathbb{R}^d$ are denoted by $\|x\|$ and $\|x\|_\infty$, resp. For $f : \mathbb{R}^d \rightarrow \mathbb{R}$

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

is its supremum norm, and the supremum norm of f on a set $A \subseteq \mathbb{R}^d$ is denoted by

$$\|f\|_{\infty, A} = \sup_{x \in A} |f(x)|.$$

The support of an \mathbb{R}^d -valued random variable X is abbreviated by

$$\text{supp}(X) = \left\{ x \in \mathbb{R}^d : \mathbf{P}_X(S_r(x)) > 0 \text{ for all } r > 0 \right\},$$

where $S_r(x)$ is the ball of radius r around x .

The outline of this paper is as follows: The assumption on the structure of the regression function is described in Section 2, in Section 3 we introduce estimates based on polynomial splines and present a result concerning their rates of convergence, Section 4 does the same for neural networks, and Section 5 contains the proofs.

2. Hierarchical interaction models

In this section we formalize the assumption that a function value is computed in several layers where in each layer a function of at most d^* inputs is computed and where the inputs are outputs of the previous layer (or components of the input variable, in case that there is no previous layer). We do this in the following recursive definition.

Definition 1. Let $d \in \mathbb{N}$, $d^* \in \{1, \dots, d\}$ and $m : \mathbb{R}^d \rightarrow \mathbb{R}$.

a) We say that m satisfies a **hierarchical interaction model of order d^* and level 0**, if there exist $I \subseteq \{1, \dots, d\}$ with $|I| = d^*$ and $f : \mathbb{R}^{d^*} \rightarrow \mathbb{R}$ such that

$$m(x) = f(x_I) \quad \text{for all } x \in \mathbb{R}^d.$$

b) We say that m satisfies a **hierarchical interaction model of order d^* and level $l + 1$** , if there exist $g : \mathbb{R}^{d^*} \rightarrow \mathbb{R}$ and $f_1, \dots, f_{d^*} : \mathbb{R}^d \rightarrow \mathbb{R}$ such that f_1, \dots, f_{d^*} satisfy a hierarchical interaction model of order d^* and level l and such that

$$m(x) = g(f_1(x), \dots, f_{d^*}(x)) \quad \text{for all } x \in \mathbb{R}^d.$$

The class of functions satisfying a hierarchical interaction model of order 1 neither includes all additive functions nor all functions satisfying the assumption of projection pursuit. But after a slight extension of the definition, which we present next, all such functions are included.

Definition 2. Let $d \in \mathbb{N}$, $d^* \in \{1, \dots, d\}$ and $m : \mathbb{R}^d \rightarrow \mathbb{R}$.

a) We say that m satisfies a **generalized hierarchical interaction model of order d^* and level 0**, if there exist $a_1, \dots, a_{d^*} \in \mathbb{R}^{d^*}$ and $f : \mathbb{R}^{d^*} \rightarrow \mathbb{R}$ such that

$$m(x) = f(a_1^T x, \dots, a_{d^*}^T x) \quad \text{for all } x \in \mathbb{R}^d.$$

b) We say that m satisfies a **generalized hierarchical interaction model of order d^* and level $l + 1$** , if there exist $K \in \mathbb{N}$, $g_k : \mathbb{R}^{d^*} \rightarrow \mathbb{R}$ ($k = 1, \dots, K$) and $f_{1,k}, \dots, f_{d^*,k} : \mathbb{R}^d \rightarrow \mathbb{R}$ ($k = 1, \dots, K$) such that $f_{1,k}, \dots, f_{d^*,k}$ ($k = 1, \dots, K$) satisfy a generalized hierarchical interaction model of order d^* and level l and such that

$$m(x) = \sum_{k=1}^K g_k(f_{1,k}(x), \dots, f_{d^*,k}(x)) \quad \text{for all } x \in \mathbb{R}^d.$$

Obviously, each hierarchical interaction model is also a generalized hierarchical interaction model of the same order and same level (because we can choose a'_k s as unit vectors and $K = 1$). Furthermore, additive functions, all functions satisfying the assumption of projection pursuit or of interaction model belong to the class of generalized hierarchical interaction model of order d^* and level 1, where $d^* = 1$ in case of additive functions or projection pursuit.

Our smoothness assumptions imposed on the functions occurring in a hierarchical interaction model are formalized in the next definition.

Definition 3. a) Let $p = k + s$ for some $k \in \mathbb{N}_0$ and $0 < s \leq 1$. A function $m : \mathbb{R}^d \rightarrow \mathbb{R}$ is called **(p, C) -smooth**, if for every $\alpha = (\alpha_1, \dots, \alpha_d) \in \mathbb{N}_0^d$ with $\sum_{j=1}^d \alpha_j = k$ the partial derivative $\frac{\partial^k m}{\partial x_1^{\alpha_1} \dots \partial x_d^{\alpha_d}}$ exists and satisfies

$$\left| \frac{\partial^k m}{\partial x_1^{\alpha_1} \dots \partial x_d^{\alpha_d}}(x) - \frac{\partial^k m}{\partial x_1^{\alpha_1} \dots \partial x_d^{\alpha_d}}(z) \right| \leq C \cdot \|x - z\|^s$$

for all $x, z \in \mathbb{R}^d$.

b) We say that the **(generalized) hierarchical interaction model** in Definition 1 (Definition 2) is **(p, C) -smooth**, if all functions occurring in its definition are (p, C) -smooth according to part a) of this definition.

Remark 1. a) If

$$m(x) = g(f_1(x), \dots, f_{d^*}(x)) \quad (x \in \mathbb{R}^d)$$

for some (p, C) -smooth functions $g : \mathbb{R}^{d^*} \rightarrow \mathbb{R}$ and $f_1, \dots, f_{d^*} : \mathbb{R}^d \rightarrow \mathbb{R}$, then we get in case $p \leq 1$

$$|m(x) - m(z)| = |g(f_1(x), \dots, f_{d^*}(x)) - g(f_1(z), \dots, f_{d^*}(z))|$$

$$\leq C \cdot \sqrt{d} \cdot \max_{j=1, \dots, d^*} |f_j(x) - f_j(z)|^p.$$

Using this and

$$\|x_I - z_I\| \leq \|x - z\|$$

for $I \subseteq \{1, \dots, d\}$ and $x, z \in \mathbb{R}^d$ we see that for any $p \leq 1$ any function which satisfies a hierarchical interaction model of level l which is (p, C) -smooth according to Definition 3 b) is (p^{l+1}, \bar{C}) -smooth according to Definition 3 a).

b) In the definition of (generalized) hierarchical interaction model it is possible to choose some of the occurring functions as projections on some component of their input, which are always (p, C) -smooth functions. Consequently, (p, C) -smooth functions depending on at most d^* components of its input variable belong to the class of functions satisfying (p, C) -smooth (generalized) hierarchical interaction models of order d^* and any fixed level. Therefore we can conclude from Stone (1982) that the minimax rate of convergence of estimation of (p, C) -smooth (generalized) hierarchical interaction models of order d^* is lower bounded by $n^{-2p/(2p+d^*)}$. In the next two sections we show that suitably defined spline and neural network estimates achieve this rate of convergence up to some logarithmic factor. In order to simplify the notation the result for splines is derived only for hierarchical interaction models, however the result for neural networks considers also generalized hierarchical interaction models.

3. Estimates based on polynomial splines

In the next two sections we study least squares estimates defined by

$$m_n(\cdot) = \arg \min_{h \in \mathcal{H}_n} \frac{1}{n} \sum_{i=1}^n |Y_i - h(X_i)|^2, \quad (1)$$

where \mathcal{H}_n is a set of functions $h : \mathbb{R}^d \rightarrow \mathbb{R}$. For simplicity we assume here and in the sequel that the minimum above indeed exists. When this is not the case our theoretical results also hold for any estimate which minimizes the above empirical L_2 risk up to a small additional term (e.g., $1/n$).

In this section we will define \mathcal{H}_n by using tensor products of polynomial splines, i.e., tensor products of piecewise polynomials satisfying a global smoothness condition. Concerning applications of tensor products of polynomial splines in nonparametric regression we refer to Friedman (1991), Stone (1994), Stone et al. (1997), Kohler (2000) and the literature cited therein.

In the sequel we introduce spaces of tensor product B-splines defined on \mathbb{R}^d and then compose them according to the definition of hierarchical interaction models. Our function spaces will depend on parameters $\alpha > 0$ (controlling the supremum norm of the functions), $\beta > 0$ (controlling the support of the functions), $\gamma > 0$ (controlling the Lipschitz constant of the functions), $M_0 \in \mathbb{N}$ (the degree of the splines), $K \in \mathbb{N}$ (controlling the degrees of freedom) and d (the dimension of \mathbb{R}^d).

We start by introducing univariate space of polynomial spline functions and a corresponding B-spline basis consisting of basis functions with compact support as follows: For $K \in \mathbb{N}$ and $M \in \mathbb{N}_0$ set $u_k = k \cdot \beta / K$ ($k \in \mathbb{Z}$). For $k \in \mathbb{Z}$ let $B_{k,M} : \mathbb{R} \rightarrow \mathbb{R}$ be the univariate B-spline of degree M with knot sequence $(u_l)_{l \in \mathbb{Z}}$ and support $\text{supp}(B_{k,M}) = [u_k, u_{k+M+1}]$. In case $M = 0$ this means that $B_{k,0}$ is the indicator function of the interval $[u_k, u_{k+1})$, and for $M = 1$ we have

$$B_{k,1}(x) = \begin{cases} \frac{x-u_k}{u_{k+1}-u_k} & , u_k \leq x \leq u_{k+1}, \\ \frac{u_{k+2}-x}{u_{k+2}-u_{k+1}} & , u_{k+1} < x \leq u_{k+2}, \\ 0 & , \text{else,} \end{cases}$$

(so-called hat-function). The general definition of $B_{k,M}$ can be found, e.g., in de Boor (1978), or in Section 14.1 of Görfi et al. (2002). These B-splines are basis functions of sets of univariate piecewise polynomials of degree M , where the piecewise polynomials are globally $(M-1)$ -times continuously differentiable and where the M -th derivative of the functions have jump points only at the knots u_l ($l \in \mathbb{Z}$).

For $\mathbf{k} = (k_1, \dots, k_d) \in \mathbb{Z}^d$ we define the tensor product B-spline $B_{\mathbf{k},M} : \mathbb{R}^d \rightarrow \mathbb{R}$ by

$$B_{\mathbf{k},M}(x^{(1)}, \dots, x^{(d)}) = B_{k_1,M}(x^{(1)}) \cdots B_{k_d,M}(x^{(d)}) \quad (x^{(1)}, \dots, x^{(d)} \in \mathbb{R}).$$

And we define $\mathcal{S}_{K,M}$ as the set of all linear combinations of all those of the above tensor product B-splines, where the support has nonempty intersection with $(-\beta, \beta)^d$, i.e., we set

$$\mathcal{S}_{K,M} = \left\{ \sum_{\mathbf{k} \in \{-K-M, -K-M+1, \dots, K-1\}^d} a_{\mathbf{k}} \cdot B_{\mathbf{k},M} : a_{\mathbf{k}} \in \mathbb{R} \right\}.$$

It can be shown by using standard arguments from spline theory, that the functions in $\mathcal{S}_{K,M}$ are in each component $(M-1)$ -times continuously differentiable, that they are equal to a (multivariate) polynomial of degree less than or equal to M (in each component) on each rectangle

$$[u_{k_1}, u_{k_1+1}) \times \cdots \times [u_{k_d}, u_{k_d+1}) \quad (\mathbf{k} = (k_1, \dots, k_d) \in \mathbb{Z}^d),$$

and that they vanish outside of the set

$$\left[-\beta - M \cdot \frac{\beta}{K}, \beta + M \cdot \frac{\beta}{K} \right]^d.$$

For our estimate we need to impose bounds on the supremum norm and the Lipschitz constant of our functions. We do this by restricting the coefficients in the spline space as follows: Let e_i be the i -th unit vector in \mathbb{R}^d ($i = 1, \dots, d$). Then we set

$$\mathcal{S}_{K,M,\alpha,\beta,\gamma,d} = \left\{ \sum_{\mathbf{k} \in \mathbb{Z}^d} a_{\mathbf{k}} \cdot B_{\mathbf{k},M} : |a_{\mathbf{k}}| \leq \alpha, |a_{\mathbf{k}} - a_{\mathbf{k}-e_i}| \leq \frac{\beta \cdot \gamma}{\sqrt{d} \cdot K} \quad (i = 1, \dots, d), \right.$$

$$a_{\mathbf{k}} = 0 \text{ if } \text{supp}(B_{\mathbf{k},M}) \cap (-\beta, \beta)^d = \emptyset \Big\}.$$

The definition of the B-splines implies that $\mathcal{S}_{K,M,\alpha,\beta,\gamma,d}$ is a subset of a linear vector space of dimension $(2 \cdot K + M)^d$. Furthermore, by standard results on B-splines and their derivatives (cf., e.g., Lemmas 14.4 and 14.6 in Györfi et al. (2002)) it can be shown that the functions in $\mathcal{S}_{K,M,\alpha,\beta,\gamma,d}$ are bounded in absolute value by α and are for $M > 0$ Lipschitz continuous with Lipschitz constant bounded by γ (since all partial derivatives of order one are bounded in absolute value by γ/\sqrt{d}).

Now we assume that we have given a hierarchical interaction model of order d^* and that we know all subsets I occurring in its definition. We use them to define similarly a composition of our spline spaces as follows:

For level 0 we define $\mathcal{H}^{(0)}$ by choosing $I \subseteq \{1, \dots, d\}$ with $|I| = d^*$ and by setting

$$\mathcal{H}^{(0)} = \left\{ h : \mathbb{R}^d \rightarrow \mathbb{R} \quad : \quad h(x) = f(x_I) \quad (x \in \mathbb{R}^d) \quad \text{for some } f \in \mathcal{S}_{K,M,\alpha,\beta,\gamma,d^*} \right\}.$$

For level $l+1$ we define $\mathcal{H}_1^{(l)}, \dots, \mathcal{H}_{d^*}^{(l)}$ according to the functions chosen in the definition of our hierarchical interaction model of level l and set

$$\begin{aligned} \mathcal{H}^{(l+1)} = \left\{ h : \mathbb{R}^d \rightarrow \mathbb{R} \quad : \quad h(x) = g(f_1(x), \dots, f_{d^*}(x)) \quad (x \in \mathbb{R}^d) \quad \text{for some} \right. \\ \left. g \in \mathcal{S}_{K,M,\alpha,\beta,\gamma,d^*}, f_1 \in \mathcal{H}_1^{(l)}, \dots, f_{d^*} \in \mathcal{H}_{d^*}^{(l)} \right\}. \end{aligned}$$

If we choose this function space in our estimate (1), we get the following result.

Theorem 1. *Let $(X, Y), (X_1, Y_1), (X_2, Y_2), \dots$ be independent and identically distributed random variables with values in $\mathbb{R}^d \times \mathbb{R}$ such that*

$$\mathbf{E} \exp(c_1 \cdot Y^2) < \infty$$

for some constant $c_1 > 0$. Let m be the corresponding regression function and assume that m satisfies a hierarchical interaction model of order d^ and level $l \in \mathbb{N}_0$, which is (p, C) -smooth according to Definition 3 for some $p \in \mathbb{N}$ and $C > 0$. Furthermore assume that $\text{supp}(X)$ is bounded.*

Let m_n be the least squares estimate defined by (1), where the function space is chosen as above using tensor product spline functions and where the construction is done accordingly to the hierarchical interaction model for m with parameters

$$K = K_n = \left\lceil n^{1/(2p+d^*)} \right\rceil, \alpha = \alpha_n = \log n, \beta = \beta_n = \log n \text{ and } \gamma = \gamma_n = \log n$$

and degree $M > p - 1$. Then we have for n sufficiently large

$$\mathbf{E} \int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx) \leq c_2 \cdot \log^{\max\{3, 2p\}}(n) \cdot n^{-2p/(2p+d^*)}.$$

Remark 1. In the definition of the estimate in Theorem 1 we have to choose parameters depending on the smoothness and the structure of the assumed hierarchical interaction model, which is not possible in an application since there the smoothness of the regression function will be usually unknown. But there are standard data-driven methods to choose the parameters of a regression estimate, e.g., splitting of the sample (cf., e.g., Chapter 7 in Györfi et al. (2002)). If we apply splitting of the sample, then the result of Theorem 1 can be shown also for an estimate whose definition does not depend on the smoothness of the regression function.

4. Estimates based on neural networks

In this section we assume that the function space \mathcal{H}_n in the definition of our least squares estimate (1) consists of multilayer feedforward neural networks. The starting point in defining such neural networks is the choice of a so-called sigmoidal function $\sigma : \mathbb{R} \rightarrow [0, 1]$. Usually one uses here so-called squashing functions which are defined as functions $\sigma : \mathbb{R} \rightarrow [0, 1]$ which are nondecreasing and satisfy

$$\lim_{x \rightarrow -\infty} \sigma(x) = 0 \quad \text{and} \quad \lim_{x \rightarrow \infty} \sigma(x) = 1.$$

Examples of sigmoidal functions which are squashing functions include the logistic squasher

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

or the Gaussian squasher

$$\sigma(x) = \frac{1}{\sqrt{2 \cdot \pi}} \int_{-\infty}^x \exp(-u^2/2) du.$$

Multilayer feedforward neural networks with sigmoidal functions can be defined recursively as follows: A multilayer feedforward neural network with l hidden layers, $K_1, \dots, K_l \in \mathbb{N}$ neurons in the first, second, \dots , l -th layer, respectively, and sigmoidal function σ is a real-valued function defined on \mathbb{R}^d of the form

$$f(x) = \sum_{i=1}^{K_l} c_i^{(l)} \cdot f_i^{(l)}(x) + c_0^{(l)}, \quad (2)$$

for some $c_0^{(l)}, \dots, c_{K_l}^{(l)} \in \mathbb{R}$ and for $f_i^{(l)}$'s recursively defined by

$$f_i^{(r)}(x) = \sigma \left(\sum_{j=1}^{K_{r-1}} c_{i,j}^{(r-1)} \cdot f_j^{(r-1)}(x) + c_{i,0}^{(r-1)} \right) \quad (3)$$

for some $c_{i,0}^{(r-1)}, \dots, c_{i,K_{r-1}}^{(r-1)} \in \mathbb{R}$ and

$$f_i^{(1)}(x) = \sigma \left(\sum_{j=1}^d c_{i,j}^{(0)} \cdot x^{(j)} + c_{i,0}^{(0)} \right) \quad (4)$$

for some $c_{i,0}^{(0)}, \dots, c_{i,d}^{(0)} \in \mathbb{R}$.

For applications of neural networks to nonlinear function estimation, classification and learning we refer the reader to the monographs Hertz, Krogh and Palmer (1991), Devroye, Györfi and Lugosi (1996), Ripley (2008), Anthony and Bartlett (1999), Györfi et al. (2002), Haykin (2008) and Hastie, Tibshirani and Friedman (2011). Consistency of nonparametric regression estimates using neural networks has been studied by Mielniczuk and Tyrcha (1993) and Lugosi and Zeger (1995). The rate of convergence of neural network regression estimates with one hidden layer has been analyzed by Barron (1991; 1993) and McCaffrey and Gallant (1994), and in connection with feedforward neural network with two hidden layers in Kohler and Krzyżak (2005).

Our choice of the set of neural networks suitable for estimation of generalized hierarchical interaction models is motivated by the following approximation result presented in Mhaskar (1993): Let $m : \mathbb{R}^d \rightarrow \mathbb{R}$ be a (p, C) -smooth function, where $0 < p \leq 1$, let $N \in \mathbb{N}$ and let A be a compact subset of \mathbb{R}^d . Then there exists a neural network

$$t(x) = \sum_{i=1}^{N^d} c_i \cdot \sigma \left(\sum_{j=1}^d b_{i,j} \cdot \sigma \left(\sum_{k=1}^d a_{i,j,k} \cdot x^{(k)} + a_{i,j,0} \right) + b_{i,0} \right) + c_0$$

with two hidden layers such that

$$|t(x) - m(x)| \leq c_3 \cdot C \cdot \frac{1}{N^p}$$

for "nearly" all $x \in A$ (see Lemma 6 below for details).

Now assume that m satisfies a generalized hierarchical interaction model of order d^* and level 0, which is (p, C) -smooth, i.e.,

$$m(x) = f(a_1^T x, \dots, a_{d^*}^T x) \quad \text{for all } x \in \mathbb{R}^d$$

for some $a_1, \dots, a_{d^*} \in \mathbb{R}^{d^*}$ and some (p, C) -smooth function $f : \mathbb{R}^{d^*} \rightarrow \mathbb{R}$. Approximating f by the above feedforward neural network with two hidden layers defined on \mathbb{R}^{d^*} we see that we can approximate m by the following feedforward neural network with two hidden layers defined on \mathbb{R}^d :

$$f(x) = \sum_{i=1}^{N^{d^*}} c_i \cdot \sigma \left(\sum_{j=1}^{d^*} b_{i,j} \cdot \sigma \left(\sum_{k=1}^d a_{i,j,k} \cdot x^{(k)} + a_{i,j,0} \right) + b_{i,0} \right) + c_0 \quad (x \in \mathbb{R}^d). \quad (5)$$

Here in the first and in the second hidden layer we are using $d^* \cdot N^{d^*}$ and N^{d^*} neurons, respectively. However, the neural network has only

$$N^{d^*} + 1 + N^{d^*} \cdot (d^* + 1) + N^{d^*} \cdot d^* \cdot (d + 1) = N^{d^*} \cdot (d^* \cdot d + 2 \cdot d^* + 1) + 1 \quad (6)$$

weights. This is due to the fact, that the two hidden layers of the neural network are not completely connected. Instead, each neuron in the second hidden layer is connected with

d^* neurons in the first hidden layer, and this is done in such a way that each neuron in the first hidden layer is connected with exactly one neuron in the second hidden layer.

For $N \in \mathbb{N}$, $d \in \mathbb{N}$, $d^* \in \{1, \dots, d\}$ and $\alpha > 0$ we denote the sets of all functions $f : \mathbb{R}^d \rightarrow \mathbb{R}$ which satisfy (5) for some $a_{i,j,k}, b_{i,j}, c_i \in \mathbb{R}$, where

$$|a_{i,j,k}| \leq \alpha, \quad |b_{i,j}| \leq \alpha \quad \text{and} \quad |c_i| \leq \alpha$$

for all $i \in \{0, 1, \dots, N^{d^*}\}$, $j \in \{0, 1, \dots, d^*\}$ and $k \in \{0, 1, \dots, d\}$, by $\mathcal{F}_{N,d^*,d,\alpha}^{(\text{neural networks})}$. Motivated by the definition of a generalized hierarchical interaction model, we define so-called spaces of hierarchical neural networks with parameters K, N, d^*, d and level l as follows. In case $l = 0$ we set

$$\mathcal{H}^{(0)} = \mathcal{F}_{N,d^*,d,\alpha}^{(\text{neural networks})}.$$

And for $l > 0$ we define

$$\mathcal{H}^{(l)} = \left\{ h : \mathbb{R}^d \rightarrow \mathbb{R} \quad : \quad h(x) = \sum_{k=1}^K g_k(f_{1,k}(x), \dots, f_{d^*,k}(x)) \quad (x \in \mathbb{R}^d) \right. \\ \left. \text{for some } g_k \in \mathcal{F}_{N,d^*,d^*,\alpha}^{(\text{neural networks})} \text{ and } f_{j,k} \in \mathcal{H}^{(l-1)} \right\}.$$

The class $\mathcal{H}^{(0)}$ is a set neural networks with two hidden layers and number of weights given by (6). From this one can conclude recursively that for $l > 0$ the class $\mathcal{H}^{(l)}$ is a set neural networks with $2 \cdot l$ hidden layers, where the weights can be parameterized by

$$(K+1)^l \cdot \left(N^{d^*} \cdot (d+1)^2 + 1 \right)$$

many parameters (there are in fact much more weights in the neural networks, however, the are related to each other (in the sense that they are products of weights $a_{i,j,k}$ of the network at level $2r+1$ and of weights c_i of the network at level $2r$) and can therefore be parameterized by the above number of parameters).

Next we choose in our least squares estimate (1) the set \mathcal{H}_n as the set $\mathcal{H}^{(l)}$, with parameters $K = K_{max}$, $N = K_n$, d^* , d and level l , where d^* and l are the values from the definition of the generalized hierarchical interaction model for m . Then the following result holds.

Theorem 2. *Let $(X, Y), (X_1, Y_1), (X_2, Y_2), \dots$ be independent and identically distributed random variables with values in $\mathbb{R}^d \times \mathbb{R}$ such that*

$$\mathbf{E} \exp(c_1 \cdot Y^2) < \infty$$

for some constant $c_1 > 0$ and such that $\text{supp}(X)$ is bounded. Let m be the corresponding regression function and assume that m satisfies a generalized hierarchical interaction model of order d^ which is (p, C) -smooth according to Definition 3 for some $0 < p \leq 1$*

and $C > 0$ and where all functions occurring in Definition 2 b) are Lipschitz continuous. Let K_{\max} be the maximal number of summands in the different levels in Definition 2 b).

Let m_n be the least squares estimate defined by (1) with \mathcal{H}_n defined as above with

$$K_n = \left\lceil \left(\frac{n}{\log(n)} \right)^{1/(2p+d^*)} \right\rceil \quad \text{and} \quad \alpha_n = n^4.$$

Assume that the sigmoidal function $\sigma : \mathbb{R} \rightarrow [0, 1]$ is a Lipschitz continuous squashing function which satisfies

$$|\sigma(y) - 1| \leq \frac{1}{y} \quad \text{if } y > 0 \quad \text{and} \quad |\sigma(y)| \leq \frac{1}{|y|} \quad \text{if } y < 0.$$

Then

$$\mathbf{E} \int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx) \leq c_4 \cdot \log^3(n) \cdot n^{-2p/(2p+d^*)}.$$

Remark 2. The class of (p, C) -smooth generalized hierarchical interaction models of order d^* , where all functions occurring in Definition 2 b) are Lipschitz continuous, contains all (p, C) -smooth functions which depend on at most d^* of its input components, since in the definition of generalized hierarchical interaction models all functions occurring in Definition 2 might be chosen as projections. Consequently the rate of convergence in Theorem 2 is optimal up to some logarithmic factor according to Stone (1982).

Remark 3. As in Remark 1 the parameters of our neural network estimate can be chosen in a data-dependent way by splitting of the sample.

5. Proofs

5.1. A general result on least squares estimates

The estimates in Theorems 1 and 2 are least squares estimates. The L_2 error of such estimates depends on the approximation properties and the complexity of the used functions spaces. The latter one can be measured by so-called covering numbers, which we introduce next.

Definition 4. Let $\epsilon > 0$, let \mathcal{G} be a set of functions $g : \mathbb{R}^d \rightarrow \mathbb{R}$, let $x_1, \dots, x_n \in \mathbb{R}^d$ and set $x_1^n = (x_1, \dots, x_n)$.

a) Every finite collection of functions $g_1, \dots, g_N : \mathbb{R}^d \rightarrow \mathbb{R}$ with the property that for every $g \in \mathcal{G}$ there exists a $j = j(g) \in \{1, \dots, N\}$ such that

$$\|g - g_j\|_\infty < \epsilon$$

is called a supremum norm ϵ -cover of \mathcal{G} . The size of the smallest supremum norm ϵ -cover of \mathcal{G} is called supremum norm ϵ -covering number of \mathcal{G} and is denoted by $\mathcal{N}_\infty(\epsilon, \mathcal{G})$. Here we set $\mathcal{N}_\infty(\epsilon, \mathcal{G}) = \infty$ in case that there exists no finite supremum norm ϵ -cover of \mathcal{G} .

b) Every finite collection of functions $g_1, \dots, g_N : \mathbb{R}^d \rightarrow \mathbb{R}$ with the property that for every $g \in \mathcal{G}$ there exists a $j = j(g) \in \{1, \dots, N\}$ such that

$$\frac{1}{n} \sum_{i=1}^n |g(x_i) - g_j(x_i)| < \epsilon$$

is called a L_1 norm ϵ -cover of \mathcal{G} on x_1^n . The size of the smallest L_1 norm ϵ -cover of \mathcal{G} on x_1^n is called L_1 norm ϵ -covering number of \mathcal{G} on x_1^n and is denoted by $\mathcal{N}_1(\epsilon, \mathcal{G}, x_1^n)$. Here we set $\mathcal{N}_1(\epsilon, \mathcal{G}, x_1^n) = \infty$ in case that there exists no finite L_1 norm ϵ -cover of \mathcal{G} on x_1^n .

Using the notion of covering numbers we can formulate the following general result on the least squares estimates.

Lemma 1. Let $(X, Y), (X_1, Y_1), (X_2, Y_2), \dots$ be independent and identically distributed random variables with values in $\mathbb{R}^d \times \mathbb{R}$ such that

$$\mathbf{E} \exp(c_1 \cdot Y^2) < \infty$$

for some constant $c_1 > 0$. Let m be the corresponding regression function and let m_n be the least squares estimate defined by (1). Assume that the function space \mathcal{H}_n consists of functions bounded in absolute value by $c_5 \cdot \log(n)$ and that its covering number satisfies

$$\sup_{x_1, \dots, x_n \in \mathbb{R}^d} \mathcal{N}_1\left(\frac{1}{n}, \mathcal{H}_n, x_1^n\right) \leq \mathcal{N}_1\left(\frac{1}{n}, \mathcal{H}_n\right)$$

for some $\mathcal{N}_1\left(\frac{1}{n}, \mathcal{H}_n\right) \geq 3$. Then

$$\begin{aligned} & \mathbf{E} \int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx) \\ & \leq c_5 \cdot \log(n)^2 \cdot \frac{\log(\mathcal{N}_1\left(\frac{1}{n}, \mathcal{H}_n\right))}{n} + 2 \inf_{h \in \mathcal{H}_n} \int |h(x) - m(x)|^2 \mathbf{P}_X(dx). \end{aligned}$$

Proof. The result is a consequence of the standard error bounds on least squares estimates derived by using results from the empirical process theory, cf., e.g., proof of Theorem 11.5 in Györfi et al. (2002) and proof of Theorem 1 in Bagirov, Clausen and Kohler (2009). \square

5.2. Proof of Theorem 1

In the proof of Theorem 1 we will apply Lemma 1. In order to bound the covering number and the approximation error (i.e., $\inf_{h \in \mathcal{H}_n} \int |h(x) - m(x)|^2 \mathbf{P}_X(dx)$), we will need the following auxiliary results.

Lemma 2. Let $g, \bar{g} : \mathbb{R}^d \rightarrow \mathbb{R}$, $f_1, \bar{f}_1, \dots, f_d, \bar{f}_d : \mathbb{R}^d \rightarrow \mathbb{R}$ and define m and \bar{m} by

$$m(x) = g(f_1(x), \dots, f_d(x)) \quad (x \in \mathbb{R}^d) \quad \text{and} \quad \bar{m}(x) = \bar{g}(\bar{f}_1(x), \dots, \bar{f}_d(x)) \quad (x \in \mathbb{R}^d).$$

If g is Lipschitz continuous with Lipschitz constant $C > 0$, then we have for any $x \in \mathbb{R}^d$

$$|m(x) - \bar{m}(x)| \leq C \cdot \sqrt{d} \cdot \sum_{j=1}^d |f_j(x) - \bar{f}_j(x)| + \|g - \bar{g}\|_\infty.$$

Proof. The result follows from the triangle inequality, the Lipschitz continuity of g and a bound on the L_2 norm by the L_1 norm:

$$\begin{aligned} |m(x) - \bar{m}(x)| &\leq |g(f_1(x), \dots, f_d(x)) - g(\bar{f}_1(x), \dots, \bar{f}_d(x))| \\ &\quad + |g(\bar{f}_1(x), \dots, \bar{f}_d(x)) - \bar{g}(\bar{f}_1(x), \dots, \bar{f}_d(x))| \\ &\leq C \cdot \left\| (f_1(x) - \bar{f}_1(x), \dots, f_d(x) - \bar{f}_d(x)) \right\|^T + \|g - \bar{g}\|_\infty \\ &\leq C \cdot \sqrt{d} \cdot \sum_{j=1}^d |f_j(x) - \bar{f}_j(x)| + \|g - \bar{g}\|_\infty. \end{aligned}$$

□

Let $\mathcal{G}, \mathcal{F}_1, \dots, \mathcal{F}_d$ be sets of functions $f : \mathbb{R}^d \rightarrow \mathbb{R}$ and define

$$\begin{aligned} \mathcal{H} &= \mathcal{G}(\mathcal{F}_1, \dots, \mathcal{F}_d) \\ &= \left\{ h : \mathbb{R}^d \rightarrow \mathbb{R} \quad : \quad h(x) = g(f_1(x), \dots, f_d(x)) \quad (x \in \mathbb{R}^d) \quad \text{for some} \right. \\ &\quad \left. g \in \mathcal{G}, f_1 \in \mathcal{F}_1, \dots, f_d \in \mathcal{F}_d \right\}. \end{aligned}$$

Lemma 3. Let $\mathcal{G}, \mathcal{F}_1, \dots, \mathcal{F}_d$ be sets of functions $f : \mathbb{R}^d \rightarrow \mathbb{R}$ and define $\mathcal{H} = \mathcal{G}(\mathcal{F}_1, \dots, \mathcal{F}_d)$ as above. If the functions in \mathcal{G} are Lipschitz continuous with Lipschitz constant $C > 0$, then we have for any $x_1^n \in (\mathbb{R}^d)^n$ and any $\epsilon > 0$:

$$\mathcal{N}_1(\epsilon, \mathcal{H}, x_1^n) \leq \mathcal{N}_\infty\left(\frac{\epsilon}{2}, \mathcal{G}\right) \cdot \prod_{j=1}^d \mathcal{N}_1\left(\frac{\epsilon}{2 \cdot \sqrt{d} \cdot d \cdot C}, \mathcal{F}_j, x_1^n\right).$$

Proof. Follows directly from Lemma 2. □

Lemma 4. Let $\mathcal{G}, \mathcal{F}_1, \dots, \mathcal{F}_d$ be sets of functions $\mathbb{R}^d \rightarrow \mathbb{R}$ and define $\mathcal{H} = \mathcal{G}(\mathcal{F}_1, \dots, \mathcal{F}_d)$ as above. Let $g, f_1, \dots, f_d : \mathbb{R}^d \rightarrow \mathbb{R}$ and define m by

$$m(x) = g(f_1(x), \dots, f_d(x)) \quad (x \in \mathbb{R}^d).$$

If g is Lipschitz continuous with Lipschitz constant $C > 0$, then

$$\inf_{h \in \mathcal{H}} \|m - h\|_\infty \leq \sqrt{d} \cdot C \cdot \sum_{j=1}^d \inf_{\bar{f} \in \mathcal{F}_j} \|f_j - \bar{f}\|_\infty + \inf_{\bar{g} \in \mathcal{G}} \|g - \bar{g}\|_\infty.$$

Proof. Follows directly from Lemma 2. \square

The following lemma describes a bound on the approximation error of the tensor product spline spaces introduced in Section 3.

Lemma 5. *Let $p \in \mathbb{N}$ and $C > 0$ and let $m : \mathbb{R}^d \rightarrow \mathbb{R}$ be a (p, C) -smooth function with compact support. Let $M > p - 1$, $K \in \mathbb{N}$ and set*

$$\alpha = \alpha_n = \log n, \beta = \beta_n = \log n \text{ and } \gamma = \gamma_n = \log n$$

and define the tensor product spline space $\mathcal{S}_{K,M,\alpha,\beta,\gamma,d}$ as in Section 3. Then we have for n sufficiently large

$$\inf_{f \in \mathcal{S}_{K,M,\alpha,\beta,\gamma,d}} \|m - f\|_\infty \leq c_6 \cdot \left(\frac{\log(n)}{K} \right)^p.$$

Proof. Follows from Lemma 15.2, Theorem 15.1 and the proof of Theorem 15.2 in Györfi et al. (2002). Here we use the fact that the coefficients of the spline approximand constructed in the proof of Theorem 15.2 in Györfi et al. (2002) satisfy

$$|a_{\mathbf{k}}| \leq \log n \quad \text{and} \quad |a_{\mathbf{k}} - a_{\mathbf{k}-e_i}| \leq \frac{(\log n)^2}{\sqrt{d} \cdot K}$$

for n sufficiently large, since $a_{\mathbf{k}}$ is a linear combination of point evaluations of the bounded function m and since $a_{\mathbf{k}} - a_{\mathbf{k}-e_i}$ is a linear combinations of differences of point evaluations of the Lipschitz continuous function m at points which have a supremum norm distance less than or equal to $(2M + 2) \cdot \beta/K$. \square

Proof of Theorem 1: An easy discretization of the (bounded) coefficients in the definition of the spline space $\mathcal{S}_{K,M,\alpha,\beta,\gamma,d^*}$ together with Lemma 15.2 in Györfi et al. (2002) shows that

$$\mathcal{N}_\infty(\epsilon, \mathcal{S}_{K,M,\alpha,\beta,\gamma,d^*}) \leq \left(\frac{2 \cdot \log n}{\epsilon} \right)^{(2 \cdot K + M)^{d^*}}.$$

From this we get by a (w.r.t. l) recursive application of Lemma 3

$$\sup_{x_1, \dots, x_n \in \mathbb{R}^d} \mathcal{N}_1 \left(\frac{1}{n}, \mathcal{H}_n^{(l)}, x_1^n \right) \leq n^{c_7 \cdot K^{d^*}}$$

for n sufficiently large (for some constant c_7 which depends on l). Furthermore, recursive application of Lemma 4 again together with Lemma 5 and the Lipschitz smoothness of all functions g occurring in Definition 1 b) of the hierarchical interaction model from m (which follows from the (p, C) -smoothness of the model and $p \geq 1$) implies

$$\inf_{h \in \mathcal{H}_n} \int |h(x) - m(x)|^2 \mathbf{P}_X(dx) \leq \inf_{h \in \mathcal{H}_n} \|h(x) - m(x)\|_{\infty, \text{supp}(X)}^2 \leq c_8 \cdot \left(\frac{\log(n)}{K} \right)^{2p}$$

for n sufficiently large. Using these two bounds we get the assertion by an application of Lemma 1 and the definition of K . \square

5.3. Proof of Theorem 2

In the proof we will use the following auxiliary results.

Lemma 6. *Let $m : \mathbb{R}^d \rightarrow \mathbb{R}$ be a (p, C) -smooth function, where $0 < p \leq 1$, let $N \in \mathbb{N}$, let $A \supseteq [0, 1]^d$ be a compact subset of \mathbb{R}^d , let $\eta \in (0, 1]$ and let ν be a probability measure on \mathbb{R}^d . Let $\sigma : \mathbb{R} \rightarrow [0, 1]$ be a squashing function. Then there exists a neural network*

$$t(x) = \sum_{i=1}^{N^d} c_i \cdot \sigma \left(\sum_{j=1}^d b_{i,j} \cdot \sigma \left(\sum_{k=1}^d a_{i,j,k} \cdot x^{(k)} + a_{i,j,0} \right) + b_{0,j} \right) + c_0$$

with two hidden layers such that outside of a set of ν -measure less than or equal to η we have for all $x \in A$

$$|t(x) - m(x)| \leq c_9 \cdot \frac{1}{N^p}.$$

In case that σ satisfies

$$|\sigma(y) - 1| \leq \frac{1}{y} \quad \text{if } y > 0 \quad \text{and} \quad |\sigma(y)| \leq \frac{1}{|y|} \quad \text{if } y < 0$$

the weights in the neural network above can be chosen such that

$$|c_i| \leq 2^{d+1} \cdot \|m\|_\infty, \quad |b_{i,j}| \leq 4 \cdot d \cdot N^d \quad \text{and} \quad |a_{i,j,k}| \leq 24 \cdot d^2 \cdot \left(\max_{z \in A} \|z\|_\infty + 1 \right) \cdot \frac{N}{\eta}$$

$$(i \in \{1, \dots, N^d\}, j, k \in \{1, \dots, d\}).$$

Proof. The result can be proven by modifying the proof of Theorem 3.4 in Mhaskar (1993). For the sake of completeness we present a complete proof of this result in the appendix. \square

Lemma 7. *Let $\sigma : \mathbb{R} \rightarrow [0, 1]$ be a sigmoidal function which is Lipschitz continuous with Lipschitz constant $C \geq 1$. Define $f, \bar{f} : \mathbb{R}^d \rightarrow \mathbb{R}$ recursively by*

$$f(x) = \sum_{i=1}^{K_l} c_i^{(l)} \cdot f_i^{(l)}(x) + c_0^{(l)} \quad \text{and} \quad \bar{f}(x) = \sum_{i=1}^{K_l} \bar{c}_i^{(l)} \cdot \bar{f}_i^{(l)}(x) + \bar{c}_0^{(l)}$$

for some $c_0^{(l)}, \bar{c}_0^{(l)}, \dots, c_{K_l}^{(l)}, \bar{c}_{K_l}^{(l)} \in \mathbb{R}$ and for $f_i^{(l)}, \bar{f}_i^{(l)}$'s recursively defined by

$$f_i^{(r)}(x) = \sigma \left(\sum_{j=1}^{K_{r-1}} c_{i,j}^{(r-1)} \cdot f_j^{(r-1)}(x) + c_{i,0}^{(r-1)} \right)$$

and

$$\bar{f}_i^{(r)}(x) = \sigma \left(\sum_{j=1}^{K_{r-1}} \bar{c}_{i,j}^{(r-1)} \cdot \bar{f}_j^{(r-1)}(x) + \bar{c}_{i,0}^{(r-1)} \right)$$

for some $c_{i,0}^{(r-1)}, \bar{c}_{i,0}^{(r-1)}, \dots, c_{i,K_{r-1}}^{(r-1)}, \bar{c}_{i,K_{r-1}}^{(r-1)} \in \mathbb{R}$ ($r \in \{2, \dots, l\}$, $i \in \{1, \dots, K_r\}$) and

$$f_i^{(1)}(x) = \sigma \left(\sum_{j=1}^d c_{i,j}^{(0)} \cdot x^{(j)} + c_{i,0}^{(0)} \right) \quad \text{and} \quad \bar{f}_i^{(1)}(x) = \sigma \left(\sum_{j=1}^d \bar{c}_{i,j}^{(0)} \cdot x^{(j)} + \bar{c}_{i,0}^{(0)} \right)$$

for some $c_{i,0}^{(0)}, \dots, c_{i,d}^{(0)}, \bar{c}_{i,0}^{(0)}, \dots, \bar{c}_{i,d}^{(0)} \in \mathbb{R}$ ($i \in \{1, \dots, K_1\}$). Then

$$\begin{aligned} |f(x) - \bar{f}(x)| &\leq \max\{\|x\|_\infty, 1\} \cdot (d+1) \cdot (l+1) \cdot \left(1 + \max_{r=0, \dots, l, i=1, \dots, K_{r+1}, j=1, \dots, K_r} |c_{i,j}^{(r)}| \right)^{l+1} \\ &\quad \cdot C^l \cdot \prod_{r=0}^l (K_r + 1) \cdot \max_{r=0, \dots, l, i=1, \dots, K_{r+1}, j=0, \dots, K_r} |c_{i,j}^{(r)} - \bar{c}_{i,j}^{(r)}| \end{aligned}$$

for any $x \in \mathbb{R}^d$, where we have set $K_0 = d$, $K_{l+1} = 1$ and $c_{1,i}^{(l)} = c_i^{(l)}$.

Proof. By triangle inequality and $\|\sigma\|_\infty \leq 1$ we get

$$\begin{aligned} &|f(x) - \bar{f}(x)| \\ &\leq \sum_{i=1}^{K_l} |c_i^{(l)}| \cdot |f_i^{(l)}(x) - \bar{f}_i^{(l)}(x)| + \sum_{i=1}^{K_l} |c_i^{(l)} - \bar{c}_i^{(l)}| \cdot |\bar{f}_i^{(l)}(x)| + |c_0^{(l)} - \bar{c}_0^{(l)}| \\ &\leq K_l \cdot \max_{i=1, \dots, K_l} |c_i^{(l)}| \cdot \max_{i=1, \dots, K_l} |f_i^{(l)}(x) - \bar{f}_i^{(l)}(x)| + (K_l + 1) \cdot \max_{i=0, \dots, K_l} |c_i^{(l)} - \bar{c}_i^{(l)}|. \end{aligned}$$

Using the Lipschitz continuity of σ and again the triangle inequality we get furthermore

$$\begin{aligned} &|f_i^{(r)}(x) - \bar{f}_i^{(r)}(x)| \\ &\leq C \cdot \left| \sum_{j=1}^{K_{r-1}} c_{i,j}^{(r-1)} \cdot f_j^{(r-1)}(x) + c_{i,0}^{(r-1)} - \sum_{j=1}^{K_{r-1}} \bar{c}_{i,j}^{(r-1)} \cdot \bar{f}_j^{(r-1)}(x) - \bar{c}_{i,0}^{(r-1)} \right| \\ &\leq C \cdot K_{r-1} \cdot \max_{j=1, \dots, K_{r-1}} |c_{i,j}^{(r-1)}| \cdot \max_{j=1, \dots, K_{r-1}} |f_j^{(r-1)}(x) - \bar{f}_j^{(r-1)}(x)| \\ &\quad + C \cdot (K_{r-1} + 1) \cdot \max_{j=0, \dots, K_{r-1}} |c_{i,j}^{(r-1)} - \bar{c}_{i,j}^{(r-1)}|. \end{aligned}$$

Finally, in the same way we see

$$|f_i^{(1)}(x) - \bar{f}_i^{(1)}(x)| \leq C \cdot \max\{\|x\|_\infty, 1\} \cdot (d+1) \cdot \max_{j=0, \dots, d} |c_{i,j}^{(0)} - \bar{c}_{i,j}^{(0)}|,$$

which implies the assertion. \square

Lemma 8. Let $\sigma : \mathbb{R} \rightarrow [0, 1]$ be a sigmoidal function which is Lipschitz continuous with Lipschitz constant $C \geq 1$ and let A be a compact subset of \mathbb{R}^d . Then for any $l \in \mathbb{N}$, $x_1^n \in A^n$, $d^* \in \{1, \dots, d\}$, $K_n \geq 2$ and $\alpha_n \geq 2$ we have

$$\mathcal{N}_1(\epsilon, \mathcal{H}^{(l)}, x_1^n) \leq c_{10} \cdot \left(\frac{\alpha_n \cdot K_n}{\epsilon} \right)^{c_{10} \cdot K_n^{d^*}}$$

for some constant c_{10} , which depends on l , d and d^* .

Proof. The neural networks in $\mathcal{H}^{(l)}$ can be parameterized using

$$(K_{max} + 1)^l \cdot \left(N^{d^*} \cdot (d + 1)^2 + 1 \right)$$

many parameters. Discretizing them using a grid of size $\delta > 0$ results in a set of functions of size

$$\left(\frac{2 \cdot \alpha_n}{\delta} \right)^{c_{11} \cdot K_n^{d^*}},$$

which has according to Lemma 7 the property that for each $f \in \mathcal{H}^{(l)}$ there exists a function \bar{f} in this set satisfying

$$|f(x) - \bar{f}(x)| \leq c_{12} \cdot (K_n \cdot \alpha_n)^{c_{12}} \cdot \delta.$$

Here we have used that some of the weights of the neural network are products of the above parameters, and that for such products we have

$$\begin{aligned} |a \cdot b - \bar{a} \cdot \bar{b}| &\leq |a - \bar{a}| \cdot |b| + |b - \bar{b}| \cdot |\bar{a}| \\ &\leq 2 \cdot \max\{|\bar{a}|, |b|\} \cdot \max\{|a - \bar{a}|, |b - \bar{b}|\}. \end{aligned}$$

The result follows by setting

$$\delta = \frac{\epsilon}{c_{12} \cdot (K_n \cdot \alpha_n)^{c_{12}}}.$$

□

Proof of Theorem 2. Repeated application of Lemma 4, which is possible because of the Lipschitz continuity of the functions occurring in Definition 2 b), together with Lemma 6 implies

$$\inf_{h \in \mathcal{H}_n} \int |h(x) - m(x)|^2 \mathbf{P}_X(dx) \leq c_{13} \cdot \left(\frac{1}{K_n} \right)^{2p}$$

for n sufficiently large. Here we use that for a generalized hierarchical interaction model the bound on the approximation error in Lemma 6 holds simultaneously for all occurring functions outside of an event of \mathbf{P}_X -measure at most c_{14}/n^2 . And on this event the integrand in the above integral is bounded in absolute value by $c_{15} \cdot K_n^{d^*} \leq c_{16} \cdot n$.

Furthermore, by Lemma 8 we can bound the covering number by

$$\mathcal{N}_1\left(\frac{1}{n}, \mathcal{H}_n, x_1^n\right) \leq c_{14} \cdot n^{c_{15} \cdot K_n^{d^*}}$$

for any $x_1, \dots, x_n \in \text{supp}(X)$. Using these two bounds we get the assertion by an application of Lemma 1 and the definition of K_n . □

6. Acknowledgment

The authors would like to thank the German Research Foundation (DFG) for funding this project within the Collaborative Research Centre 805 and Natural Sciences and Engineering Research Council of Canada for additional support.

References

- [1] Anthony, M. and Bartlett, P. L. (1999). *Neural Networks and Learning: Theoretical Foundations*. Cambridge University Press, Cambridge.
- [2] de Boor, C. (1978). *A Practical Guide to Splines*. Springer, New York.
- [3] 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.
- [4] Barron, A. R. (1991). Complexity regularization with application to artificial neural networks. In G. Roussas (ed.), *Nonparametric Functional Estimation and Related Topics*, pp. 5621–576, NATO ASI Series, Kluwer Academic Publishers, Dordrecht, Netherlands.
- [5] Barron, A. R. (1993). Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information Theory*, **39**, 930–944.
- [6] Devroye, L., Györfi, L., and Lugosi, G. (1996). *A Probabilistic Theory of Pattern Recognition*. Springer, New York.
- [7] Friedman, J. H. (1991). Multivariate adaptive regression splines (with discussion). *Annals of Statistics*, **19**, pp. 1–141.
- [8] Friedman, J. H., and Stuetzle, W. (1981). Projection pursuit regression. *Journal of the American Statistical Association*, **76**, pp. 817–823.
- [9] Györfi, L., Kohler, M., Krzyżak, A. and Walk, H. (2002). *A Distribution-Free Theory of Nonparametric Regression*. Springer-Verlag, New York.
- [10] Härdle, W., Hall, P. and Ichimura, H. (1993). Optimal smoothing in single-index models. *The Annals of Statistics*, **21**, pp. 157–178.
- [11] Härdle, W., Müller, M., Sperlich, S. and Werwatz, A. (2004). *Nonparametric and Semiparametric Models*. Springer-Verlag, New York.
- [12] 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.
- [13] Hastie, T., Tibshirani, R., and Friedman, J. (2011). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed., Springer, New York.
- [14] Haykin, S. O. (2008). *Neural Networks and Learning Machines*. 3rd ed. Prentice-Hall, New York.
- [15] Hertz, J., Krogh, A., and Palmer, R.G. (1991). *Introduction to the Theory of Neural Computation*. Addison-Wesley, Redwood City, CA.

- [16] Kohler, M. (2000). Inequalities for uniform deviations of averages from expectations with applications to nonparametric regression. *Journal of Statistical Planning and Inference*, **89**, pp. 1-23.
- [17] Kohler, M., and Krzyzak, A. (2005). Adaptive regression estimation with multilayer feedforward neural networks. *Journal of Nonparametric Statistics*, **17**, pp. 891-913, 2005.
- [18] Kong, E., and Xia, Y. (2007). Variable selection for the single index model. *Biometrika*, **94**, pp. 217-229.
- [19] Lugosi, G., and Zeger, K. (1995). Nonparametric estimation via empirical risk minimization. *IEEE Transactions on Information Theory*, **41**, 677-687.
- [20] McCaffrey, D.F., and Gallant, A.R. (1994). Convergence rates for single hidden layer feedforward networks. *Neural Networks*, **7**, 147-158.
- [21] Mhaskar, H. N. (1993). Approximation properties of multilayer feedforward artificial neural network. *Advances in Computational Mathematics*, **1**, pp. 61-80.
- [22] Mielniczuk, J., and Tyrcha, J. (1993). Consistency of multilayer perceptron regression estimators. *Neural Networks*, **6**, 1019-1022.
- [23] Ripley, B. D. (2008). *Pattern Recognition and Neural Networks*. Cambridge University Press, Cambridge, UK.
- [24] Stone, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *Annals of Statistics*, **10**, pp. 1040–1053.
- [25] Stone, C. J. (1985). Additive regression and other nonparametric models. *Annals of Statistics*, **13**, pp. 689–705.
- [26] Stone, C. J. (1994). The use of splines and their tensor products in multivariate function estimation. *Annals of Statistics*, **22**, pp. 118–184.
- [27] Stone, C. J., Hansen, M. H., Kooperberg, C., and Truong, Y. K. (1997). Polynomial splines and their tensor product in extended linear modelling. *Annals of Statistics*, **25**, pp. 1371-1410.
- [28] 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. Proof of Lemma 6.

In the proof we will use Proposition 3.8 in Mhaskar (1993), which we reformulate here (in a slightly different form) as Lemma 9.

Lemma 9. Let $K \subseteq \mathbb{R}^d$ be a polytope bounded by hyperplanes $v_i \cdot x + b_i \geq 0$ ($i = 1, \dots, L$), where $v_1, \dots, v_L \in \mathbb{R}^d$ and $b_1, \dots, b_L \in \mathbb{R}$. For $\delta > 0$ set

$$K_\delta^0 := \left\{ x \in \mathbb{R}^d : v_i \cdot x + b_i \geq \delta \text{ for all } i \in \{1, \dots, L\} \right\}$$

and

$$K_\delta^c := \left\{ x \in \mathbb{R}^d : v_i \cdot x + b_i \leq -\delta \text{ for some } i \in \{1, \dots, L\} \right\}.$$

Let $\sigma : \mathbb{R} \rightarrow [0, 1]$ be a squashing function. Let $\epsilon, \delta \in (0, 1]$ be arbitrary. Then there exists a neural network of the form

$$f(x) = \sigma \left(\sum_{i=1}^L b_i \cdot \sigma \left(\sum_{j=1}^d c_{i,j} \cdot x^{(j)} + c_{i,0} \right) + b_0 \right)$$

satisfying

$$\begin{aligned} |f(x)| &\leq 1 \text{ for } x \in \mathbb{R}^d, \\ |f(x) - 1| &\leq \epsilon \text{ for } x \in K_\delta^0, \\ |f(x)| &\leq \epsilon \text{ for } x \in K_\delta^c. \end{aligned} \tag{7}$$

In case that the squashing function satisfies

$$|\sigma(y) - 1| \leq \frac{1}{y} \text{ if } y > 0 \quad \text{and} \quad |\sigma(y)| \leq \frac{1}{|y|} \text{ if } y < 0,$$

the weights above can be chosen such that

$$|b_i| \leq \max \left\{ \frac{4}{\epsilon}, 2 \cdot L \right\} \quad \text{and} \quad |c_{i,j}| \leq \frac{4 \cdot L}{\delta} \cdot \max\{\|v_1\|_\infty, |b_1|, \dots, \|v_L\|_\infty, |b_L|\}$$

($i = 0, \dots, L, j = 0, \dots, d$).

Proof. Follows from the proof of Proposition 3.8 in Mhaskar (1993). \square

Proof of Lemma 6. W.l.o.g we assume that A is a cube. We partition this cube into N^d equivolume cubes of side length c_{16}/N (where $c_{16} \geq 1$ since $[0, 1]^d \subseteq A$). Approximating m by a piecewise constant approximand with respect to this partition yields (since m is (p, C) -smooth) a function S satisfying

$$\|S - m\|_{\infty, A} \leq c_{16} \cdot N^{-p}. \tag{8}$$

S can be expressed in the form

$$S(x) = m(x_0) + \sum_{j \in \{1, \dots, N\}^d} d_j \cdot \prod_{i=1}^d (x^{(i)} - x_j^{(i)})_+^0,$$

where x_j are the corners of the rectangles comprising the above partition and d_j are constants satisfying

$$|d_j| \leq c_{17} \cdot N^{-p}$$

(constructed by using differences of function values of m at the corners of the above partition) and $x_+ = \max\{x, 0\}$. Let K_j be the polytope defined by $x^{(i)} - x_j^{(i)} \geq 0$ ($i = 1, \dots, d$). Set $\epsilon = N^{-d}$, $\delta = \eta/(6 \cdot d \cdot N)$ and apply Lemma 9 for each K_j (i.e., with $L = d$, $v_i = \mathbf{e}_i$ and $b_i = -x_j^{(i)}$, where \mathbf{e}_i denotes the i -th unit vector) to obtain $f_j(x)$ satisfying (7) with K_j instead of K . Let

$$P(x) = m(x_0) + \sum_{j \in \{1, \dots, N\}^d} d_j \cdot f_j(x).$$

Then we can conclude from (7)

$$|P(x) - S(x)| \leq c_{17} \cdot N^{-p}$$

for all $x \in A$ which are not contained in

$$\cup_{i=1, \dots, d} \cup_{j \in \{1, \dots, N\}^d} \left\{ x \in \mathbb{R}^d \quad : \quad |x^{(i)} - x_j^{(i)}| < \eta/(6 \cdot d \cdot N) \right\}. \quad (9)$$

By shifting the positions of the x_j 's in the i -th component we can construct $\lceil d/\eta \rceil$ disjoint versions of

$$\cup_{j \in \{1, \dots, N\}^d} \left\{ x \in \mathbb{R}^d \quad : \quad |x^{(i)} - x_j^{(i)}| < \eta/(6 \cdot d \cdot N) \right\},$$

and since the sum of the ν -measures of these sets is less than or equal to one, at least one of them must have measure less than or equal to η/d . Consequently we can shift the x_j 's such that (9) has ν -measure less than η . This together with (8) implies the assertion. \square