The Annals of Statistics
2019, Vol. 47, No. 4, 2261–2285
https://doi.org/10.1214/18-AOS1747

ON DEEP LEARNING AS A REMEDY FOR THE CURSE OF DIMENSIONALITY IN NONPARAMETRIC REGRESSION

BY BENEDIKT BAUER AND MICHAEL KOHLER

Technische Universität Darmstadt

Assuming that a smoothness condition and a suitable restriction on the structure of the regression function hold, it is shown that least squares estimates based on multilayer feedforward neural networks are able to circumvent the curse of dimensionality in nonparametric regression. The proof is based on new approximation results concerning multilayer feedforward neural networks with bounded weights and a bounded number of hidden neurons. The estimates are compared with various other approaches by using simulated data.

1. Introduction.

1.1. *Nonparametric regression.* In regression analysis, a random vector (X, Y) with values in $\mathbb{R}^d \times \mathbb{R}$ satisfying $\mathbf{E}Y^2 < \infty$ is considered, and an estimation of the relation between X and Y is attempted, that is, it is tried to predict the value of the response variable Y from the value of the observation vector X. Usually, the aim is to minimize the mean squared error or L_2 risk. Thus, the construction of a (measurable) function $m^* : \mathbb{R}^d \to \mathbb{R}$, which satisfies

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

is of interest. In the following, let $m : \mathbb{R}^d \to \mathbb{R}$, $m(x) = \mathbf{E}\{Y|X = x\}$ denote the so-called regression function. Since m satisfies

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

(cf., e.g., Section 1.1 in Györfi et al. (2002)), it is the optimal predictor m^*. Moreover, a good estimate $f : \mathbb{R}^d \to \mathbb{R}$ (in the L_2 risk minimization sense) has to keep the so-called L_2 error

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

small.

Received November 2017; revised April 2018.
MSC2010 subject classifications. Primary 62G08; secondary 62G20.
Key words and phrases. Curse of dimensionality, neural networks, nonparametric regression, rate of convergence.

In applications, the distribution of (X, Y) and m are usually unknown, but a set of data

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

can often be observed, where $(X, Y), (X_1, Y_1), \ldots, (X_n, Y_n)$ are independent and identically distributed random variables. Given this data set, the aim is to construct regression estimates $m_n(\cdot) = m_n(\cdot, \mathcal{D}_n)$ such that their L_2 errors

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

are small. In contrast to parametric estimation, where a fixed structure of the regression function that depends only on finitely many parameters is assumed, in the nonparametric approach the regression function is not claimed to be describable by finitely many parameters and the whole function is estimated from the data. Györfi et al. (2002) provided a systematic overview of different approaches and nonparametric regression estimation results.

1.2. *Rate of convergence.* It is well known (see, e.g., Section 3.1 in Györfi et al. (2002)) that one has to restrict the class of regression functions that one considers to obtain nontrivial results for the rate of convergence. For that purpose, we introduce the following definition of (p, C)-smoothness.

DEFINITION 1. Let $p = q + s$ for some $q \in \mathbb{N}_0$ and $0 < s \leq 1$. A *function* $m : \mathbb{R}^d \to \mathbb{R}$ 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 $\frac{\partial^q m}{\partial x_1^{\alpha_1} \cdots \partial x_d^{\alpha_d}}$ exists and satisfies

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

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

Stone (1982) determined the optimal minimax rate of convergence in nonparametric regression for (p, C)-smooth functions. Here, a sequence of (eventually) positive numbers $(a_n)_{n \in \mathbb{N}}$ is called a *lower minimax rate of convergence* for the class of distributions $\mathcal{D}$ if

$$\liminf_{n \to \infty} \inf_{m_n} \sup_{(X,Y) \in \mathcal{D}} \frac{\mathbf{E} \int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx)}{a_n} = C_1 > 0.$$

The sequence is said to be an *achievable rate of convergence* for the class of distributions $\mathcal{D}$ if

$$\limsup_{n \to \infty} \sup_{(X,Y) \in \mathcal{D}} \frac{\mathbf{E} \int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx)}{a_n} = C_2 < \infty.$$

The sequence is called an *optimal minimax rate of convergence* if it is both a lower minimax and an achievable rate of convergence.

Stone (1982) showed that the optimal minimax rate of convergence for the estimation of a (p, C)-smooth regression function is

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

1.3. *Curse of dimensionality.* Despite the fact that it is optimal, the rate $n^{-\frac{2p}{2p+d}}$ suffers from a characteristic feature in case of high-dimensional functions: If d is relatively large compared with p, then this rate of convergence can be extremely slow. This phenomenon is well known and is often called the curse of dimensionality. Unfortunately, in many applications, the problems are high-dimensional, and hence very hard to solve. The only way to circumvent this curse of dimensionality is to impose additional assumptions on the regression function to derive better rates of convergence.

Stone (1985) assumed an additivity condition for the structure of the regression function, which said

$$m\bigl(x^{(1)}, \ldots, x^{(d)}\bigr) = m_1\bigl(x^{(1)}\bigr) + \cdots + m_d\bigl(x^{(d)}\bigr) \qquad \bigl[x = \bigl(x^{(1)}, \ldots, x^{(d)}\bigr)^T \in \mathbb{R}^d\bigr]$$

for (p, C)-smooth univariate functions $m_1, \ldots, m_d : \mathbb{R} \to \mathbb{R}$. Stone (1985) showed that in this case $n^{-2p/(2p+1)}$ is the optimal minimax rate of convergence. This approach has been generalized to so-called interaction models in Stone (1994). These models impose for some $d^* \in \{1, \ldots, d\}$ the structure

$$m(x) = \sum_{I \subseteq \{1,\ldots,d\}, |I| = d^*} m_I(x_I) \qquad \bigl[x = \bigl(x^{(1)}, \ldots, x^{(d)}\bigr)^T \in \mathbb{R}^d\bigr]$$

on the regression function, where all m_I are (p, C)-smooth functions defined on $\mathbb{R}^{|I|}$ and for $I = \{i_1, \ldots, i_{d^*}\}$ with $1 \le i_1 < \cdots < i_{d^*} \le d$ the abbreviation $x_I = (x^{(i_1)}, \ldots, x^{(i_{d^*})})^T$ is used. Then the optimal minimax rate of convergence becomes $n^{-2p/(2p+d^*)}$.

Another idea involves so-called single index models, in which

$$m(x) = g\bigl(a^T x\bigr) \qquad \bigl(x \in \mathbb{R}^d\bigr)$$

is assumed to hold, where $g : \mathbb{R} \to \mathbb{R}$ is a 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)). This concept is even extended in the so-called projection pursuit, where the regression function is assumed to be a sum of functions of the above form, that is,

$$m(x) = \sum_{k=1}^{K} g_k\bigl(a_k^T x\bigr) \qquad \bigl(x \in \mathbb{R}^d\bigr)$$

for $K \in \mathbb{N}$, $g_k : \mathbb{R} \to \mathbb{R}$ and $a_k \in \mathbb{R}^d$ (cf., e.g., Friedman and Stuetzle (1981)). If we assume that the univariate functions in these postulated structures are (p, C)-smooth, adequately chosen regression estimates can achieve the above univariate rates of convergence up to some logarithmic factor (cf., e.g., Chapter 22 in Györfi et al. (2002)).

Horowitz and Mammen (2007) studied the case of a regression function, which satisfies

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

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

These estimates achieve good rates of convergence only if the imposed assumptions are satisfied. Thus, it is useful to derive rates of convergence for more general types of functions, with which the regression functions in real applications comply more often (at least approximately) and ideally contain the simpler models as well. Our research is motivated by applications in connection with complex technical systems, which are constructed in a modular form. In this case, modeling the outcome of the system as a function of the results of its modular parts seems reasonable, where each modular part computes a function depending only on a few of the components of the high-dimensional input. The modularity of the system can be extremely complex and deep. Thus, a recursive application of the described relation makes sense and leads to the following assumption about the structure of m, which was introduced in Kohler and Krzyżak (2017).

DEFINITION 2. Let $d \in \mathbb{N}$, $d^* \in \{1, \ldots, d\}$ and $m : \mathbb{R}^d \to \mathbb{R}$.

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

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

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

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

(c) We say that the *generalized hierarchical interaction model* defined above *is (p, C)-smooth*, if all functions occurring in its definition are (p, C)-smooth according to Definition 1.

In order to enable the reader to better understand the above definition, we consider the additive model from the begin of this section as an example. Using the notation $\mathrm{id}\colon \mathbb{R} \to \mathbb{R}$ for the identity function and e_i for the ith unit vector, we can rewrite the additive model as

$$\sum_{i=1}^{d} m_i(x^{(i)}) = \sum_{i=1}^{d} m_i(\mathrm{id}(e_i^T x)) = \sum_{i=1}^{K} g_i(f_{1,i}(a_i^T x)),$$

where $K = d$, $g_i = m_i$, $f_{1,i} = \mathrm{id}$ and $a_i = e_i$. This structure corresponds to the definition of a generalized hierarchical interaction model of order 1 and level 1.

Moreover, Definition 2 includes all the other types of structures of m mentioned earlier what can be shown in a similar way. Functions complying with the single index model belong to the class of generalized hierarchical interaction models of the order 1 and level 0, the additive model (see above) and projection pursuit correspond to order 1 and level 1. In addition, the interaction model is in conformity with order d^* and level 1, whereas the assumptions of Horowitz and Mammen (2007) are consistent with order 1 and level $r + 1$.

1.4. *Neural networks*. For many years, the use of neural networks has been one of the most promising approaches in connection with applications related to approximation and estimation of multivariate functions (see, e.g., the monographs Hertz, Krogh and Palmer (1991), Devroye, Györfi and Lugosi (1996), Anthony and Bartlett (1999), Györfi et al. (2002), Haykin (2008) and Ripley (2008)). Recently, the focus is on multilayer neural networks, which use many hidden layers, and the corresponding techniques are called deep learning (cf., e.g., Schmidhuber (2015) and the literature cited therein).

Multilayer feedforward neural networks with sigmoidal function $\sigma : \mathbb{R} \to [0, 1]$ can be defined recursively as follows: A multilayer feedforward neural network with l hidden layers, which has $K_1, \ldots, K_l \in \mathbb{N}$ neurons in the first, second,..., lth hidden layer, respectively, and uses the activation function σ, is a real-valued function defined on $\mathbb{R}^d$ of the form

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

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

$$f_i^{(r)}(x) = \sigma\left(\sum_{j=1}^{K_{r-1}} c_{i,j}^{(r-1)} \cdot f_j^{(r-1)}(x) + c_{i,0}^{(r-1)}\right) \tag{2}$$

for some $c_{i,0}^{(r-1)}, \ldots, c_{i,K_{r-1}}^{(r-1)} \in \mathbb{R}$ and $r = 2, \ldots, l$ and

$$f_i^{(1)}(x) = \sigma\left(\sum_{j=1}^{d} c_{i,j}^{(0)} \cdot x^{(i)} + c_{i,0}^{(0)}\right) \tag{3}$$

for some $c_{i,0}^{(0)}, \dots, c_{i,d}^{(0)} \in \mathbb{R}$. Neural network estimates often use an activation function $\sigma : \mathbb{R} \to [0, 1]$ that is nondecreasing and satisfies

$$\lim_{z \to -\infty} \sigma(z) = 0 \quad \text{and} \quad \lim_{z \to \infty} \sigma(z) = 1,$$

for example, the so-called sigmoidal or logistic squasher

$$\sigma(z) = \frac{1}{1 + \exp(-z)} \qquad (z \in \mathbb{R}).$$

Most existing theoretical results concerning neural networks consider neural networks using only one hidden layer, that is, functions of the form

$$f(x) = \sum_{j=1}^{K} c_j \cdot \sigma\left(\sum_{k=1}^{d} c_{j,k} \cdot x^{(k)} + c_{j,0} \right) + c_0. \tag{4}$$

Consistency of neural network regression estimates has been studied by Mielniczuk and Tyrcha (1993) and Lugosi and Zeger (1995). The rate of convergence has been analyzed by Barron (1991, 1993, 1994), McCaffrey and Gallant (1994) and Kohler and Krzyżak (2005, 2017). For the L_2 error of a single hidden layer neural network, Barron (1994) proves a dimensionless rate of $n^{-1/2}$ (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.

The rate of convergence of neural network regression estimates based on two layer neural networks has been analyzed in Kohler and Krzyżak (2005). Therein, interaction models were studied, and for (p, C)-smooth interaction models with $p \leq 1$ it was shown that suitable neural network estimates achieve a rate of convergence of $n^{-2p/(2p+d^*)}$ (up to some logarithmic factor), which is again a convergence rate independent of d. In Kohler and Krzyżak (2017), this result was extended to (p, C)-smooth generalized hierarchical interaction models of the order d^*. It was shown that for such models suitably defined multilayer neural networks (in which the number of hidden layers depends on the level of the generalized interaction model) achieve the rate of convergence $n^{-2p/(2p+d^*)}$ (up to some logarithmic factor) in case $p \leq 1$. Nevertheless, this result cannot generate extremely good rates of convergence, because, even in case of $p = 1$ and a value of $d^* = 5$ (for a modular technical system not large), it leads to $n^{-\frac{2}{7}}$.

Given the successful application of multilayer feedforward neural networks, the current focus in the theoretical analysis of approximation properties of neural networks is also on a possible theoretical advantage of multilayer feedforward neural networks in contrast to neural networks with only one hidden layer (cf., e.g., Eldan and Shamir (2015) and Mhaskar and Poggio (2016)).

1.5. *Main results in this article.* In this article, we analyze the rate of convergence of suitable multilayer neural network regression estimates when the regression function satisfies a (p, C)-smooth generalized hierarchical interaction model of given order d^* and given level l. Here, $p > 0$ might be arbitrarily large. Thus, unlike Kohler and Krzyżak (2005, 2017), we also allow the case $p > 1$; this leads to far better rates of convergence. We define sets of multilayer feedforward neural networks that correspond to such a generalized hierarchical interaction model and define our regression estimates as least squares estimates based on this class of neural networks. Our main finding is that the L_2 errors of these least squares neural network regression estimates achieve the rate of convergence

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

(up to some logarithmic factor), which does not depend on d. Similar rates have already been obtained in the literature but with much more stringent assumptions on the functional class the regression function belongs to. So this article considerably generalizes previous results in this regard.

In order to achieve the mentioned rate, completely new approximation results for neural networks with several hidden layers were needed. We present such results in Theorems 2 and 3 in the proof section and the main result in Theorem 1 relies on them.

Furthermore, by applying our estimate to simulated data we demonstrate that these estimates outperform other nonparametric regression estimates for a large d, provided the regression function satisfies a generalized hierarchical interaction model.

After the original version of this paper, a relating arXiv article was uploaded by Schmidt-Hieber (2017). Therein a similar result is proven using a particular unbounded activation function in the neural networks.

1.6. *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 $\lceil z \rceil$, and $\lfloor z \rfloor$ denotes the largest integer that is less than or equal to z. Let $D \subseteq \mathbb{R}^d$ and let $f : \mathbb{R}^d \to \mathbb{R}$ be a real-valued function defined on $\mathbb{R}^d$. We write $x = \arg\min_{z \in D} f(z)$ if $\min_{z \in \mathcal{D}} f(z)$ exists and if x satisfies $x \in D$ and $f(x) = \min_{z \in \mathcal{D}} f(z)$. The Euclidean and the supremum norms of $x \in \mathbb{R}^d$ are denoted by $\|x\|$ and $\|x\|_\infty$, respectively. For $f : \mathbb{R}^d \to \mathbb{R}$,

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

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

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

Let $A \subseteq \mathbb{R}^d$, let $\mathcal{F}$ be a set of functions $f : \mathbb{R}^d \to \mathbb{R}$ and let $\varepsilon > 0$. A finite collection $f_1, \dots, f_N : \mathbb{R}^d \to \mathbb{R}$ is called an ε–$\|\cdot\|_{\infty,A}$–cover of $\mathcal{F}$ if for any $f \in \mathcal{F}$ there exists $i \in \{1, \dots, N\}$ such that

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

The ε–$\|\cdot\|_{\infty,A}$–covering number of $\mathcal{F}$ is the size N of the smallest ε–$\|\cdot\|_{\infty,A}$–cover of $\mathcal{F}$ and is denoted by $\mathcal{N}(\varepsilon, \mathcal{F}, \|\cdot\|_{\infty,A})$.

If not otherwise stated, then any c_i with $i \in \mathbb{N}$ symbolizes a real nonnegative constant, which is independent of the sample size n.

1.7. *Outline*. In Section 2, we present our main result on the rate of convergence of nonparametric regression estimates using special types of multilayer feedforward neural networks in the case of generalized hierarchical interaction models. The finite sample size behavior of these estimates is analyzed by applying the estimates to simulated data in Section 3. Section 4 contains the proofs.

2. Nonparametric regression estimation by multilayer feedforward neural networks. Motivated by the generalized hierarchical interaction models, we define so-called spaces of hierarchical neural networks with parameters K, M^*, d^*, d and level l as follows. The parameter M^* is introduced for technical reasons and originates from the composition of several smaller networks in the later proof of our approximation result. It controls the accuracy of the approximation and its ideal value will depend on certain properties of the estimated function. For $M^* \in \mathbb{N}$, $d \in \mathbb{N}$, $d^* \in \{1, \dots, d\}$ and $\alpha > 0$, we denote the set of all functions $f : \mathbb{R}^d \to \mathbb{R}$ that satisfy

$$f(x) = \sum_{i=1}^{M^*} \mu_i \cdot \sigma\left(\sum_{j=1}^{4d^*} \lambda_{i,j} \cdot \sigma\left(\sum_{v=1}^{d} \theta_{i,j,v} \cdot x^{(v)} + \theta_{i,j,0}\right) + \lambda_{i,0}\right) + \mu_0$$

($x \in \mathbb{R}^d$) for some $\mu_i, \lambda_{i,j}, \theta_{i,j,v} \in \mathbb{R}$, where

$$|\mu_i| \le \alpha, \qquad |\lambda_{i,j}| \le \alpha, \qquad |\theta_{i,j,v}| \le \alpha$$

for all $i \in \{0, 1, \dots, M^*\}$, $j \in \{0, \dots, 4d^*\}$, $v \in \{0, \dots, d\}$, by $\mathcal{F}_{M^*,d^*,d,\alpha}^{(\text{neural networks})}$. In the first and the second hidden layer, we use $4 \cdot d^* \cdot M^*$ and M^* neurons, respectively. However, the neural network has only

$$\begin{aligned} &W\big(\mathcal{F}_{M^*,d^*,d,\alpha}^{(\text{neural networks})}\big) \\ &\quad = M^* + 1 + M^* \cdot \big(4d^* + 1\big) + M^* \cdot 4d^* \cdot (d+1) \\ &\quad = M^* \cdot \big(4d^* \cdot (d+2) + 2\big) + 1 \end{aligned} \tag{5}$$

weights, because the first and the second hidden layer of the neural network are not fully connected. Instead, each neuron in the second hidden layer is connected

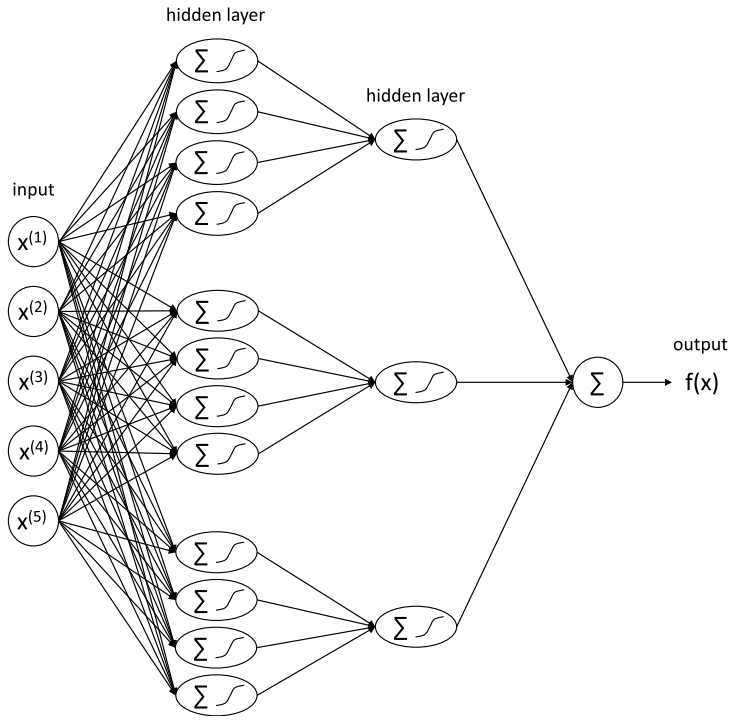

FIG. 1. *A not completely connected neural network* $f : \mathbb{R}^5 \to \mathbb{R}$ *from* $\mathcal{F}_{3,1,5,\alpha}^{(\text{neural networks})}$ *with the structure* $f(x) = \sum_{i=1}^{3} \mu_i \cdot \sigma(\sum_{j=1}^{4} \lambda_{i,j} \cdot \sigma(\sum_{v=1}^{5} \theta_{i,j,v} \cdot x^{(v)}))$ (*all weights with an index including zero neglected for a clear illustration*).

with $4d^*$ 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. The exemplary network in Figure 1 gives a good idea of how the sparse connection works.

For $l = 0$, we define our space of hierarchical neural networks by

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

For $l > 0$, we define recursively

$$\mathcal{H}^{(l)} = \Bigg\{ h : \mathbb{R}^d \to \mathbb{R} : h(x) = \sum_{k=1}^{K} g_k\big(f_{1,k}(x), \ldots, f_{d^*,k}(x)\big) \text{ for some } g_k \in \mathcal{F}_{M^*,d^*,d^*,\alpha}^{(\text{neural networks})} \text{ and } f_{j,k} \in \mathcal{H}^{(l-1)} \Bigg\}. \tag{6}$$

The class $\mathcal{H}^{(0)}$ is a set of neural networks with two hidden layers and a number of weights given by (5). From this, one can conclude (again recursively) that for

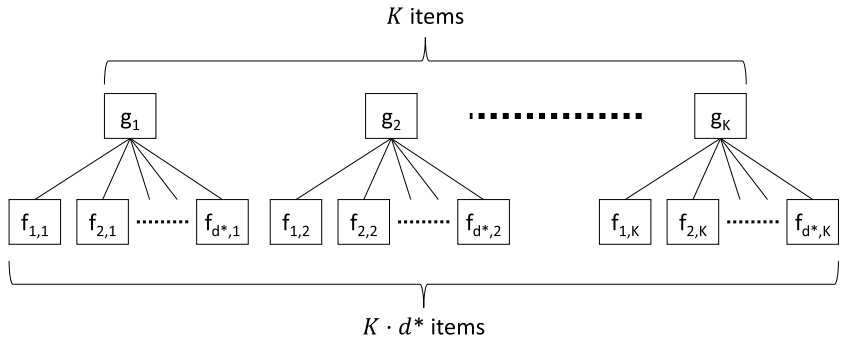

FIG. 2. *Illustration of the components of a function from* $\mathcal{H}^{(l)}$.

$l > 0$ the class $\mathcal{H}^{(l)}$ is a set of neural networks with $2 \cdot l + 2$ hidden layers. Furthermore, let $N(\mathcal{H}^{(l)})$ denote the number of linked two-layered neural networks from $\mathcal{F}^{(\text{neural networks})}_{M^*,d^*,d,\alpha}$ that define the functions from $\mathcal{H}^{(l)}$. Then the recursion

$$N\big(\mathcal{H}^{(0)}\big) = 1,$$
$$N\big(\mathcal{H}^{(l)}\big) = K + K \cdot d^* \cdot N\big(\mathcal{H}^{(l-1)}\big) \qquad (l \in \mathbb{N})$$

holds, which can be easily retraced in Figure 2. The above functions $g_1, \ldots, g_K$ therein correspond to K networks from $\mathcal{F}^{(\text{neural networks})}_{M^*,d^*,d,\alpha}$ and the $K \cdot d^*$ inner functions $f_{1,1}, \ldots, f_{d^*,K}$ originate from $\mathcal{H}^{(l-1)}$ per definition, which leads to $K \cdot d^* \cdot N(\mathcal{H}^{(l-1)})$ additional networks.

This recursive consideration yields the solution

$$N\big(\mathcal{H}^{(l)}\big) = \sum_{t=1}^{l} d^{*t-1} \cdot K^t + \big(d^* \cdot K\big)^l. \tag{7}$$

Consequently, a function from $\mathcal{H}^{(l)}$ has at most

$$N\big(\mathcal{H}^{(l)}\big) \cdot W\big(\mathcal{F}^{(\text{neural networks})}_{M^*,d^*,d,\alpha}\big) \tag{8}$$

variable weights. Although this number of weights is exponential in the number of layers l, it can be controlled because a typical example of the technical systems which motivated Definition 2 has only a moderate finite l. As explained after Definition 2, all typical assumptions for the regression function in the literature also correspond to a small l.

We define $\tilde{m}_n$ as the least squares estimate

$$\tilde{m}_n(\cdot) = \arg\min_{h \in \mathcal{H}^{(l)}} \frac{1}{n} \sum_{i=1}^{n} \big|Y_i - h(X_i)\big|^2. \tag{9}$$

For our result we need to truncate this estimate. We define the truncation operator T_β with level $\beta > 0$ as

$$T_\beta u = \begin{cases} u & \text{if } |u| \le \beta, \\ \beta \cdot \operatorname{sign}(u) & \text{otherwise.} \end{cases}$$

Regarding the sigmoidal function σ within the neural networks, our results require a few additional properties, which are satisfied by several common activation functions [e.g., the sigmoidal squasher, for which they can be straightforwardly checked with arbitrary $N \in \mathbb{N}_0$; see Supplement A in Bauer and Kohler (2019)]. We summarize them in the next definition.

DEFINITION 3. A nondecreasing and Lipschitz continous function $\sigma : \mathbb{R} \to [0, 1]$ is called *N-admissible*, if the following three conditions are satisfied:

(i) The function σ is at least $N + 1$ times continuously differentiable with bounded derivatives.

(ii) A point $t_\sigma \in \mathbb{R}$ exists, where all derivatives up to the order N of σ are different from zero.

(iii) If $y > 0$, the relation $|\sigma(y) - 1| \le \frac{1}{y}$ holds. If $y < 0$, the relation $|\sigma(y)| \le \frac{1}{|y|}$ holds.

Our main result is the following theorem.

THEOREM 1. *Let $(X, Y), (X_1, Y_1), \ldots, (X_n, Y_n)$ be independent and identically distributed random variables with values in $\mathbb{R}^d \times \mathbb{R}$ such that* $\operatorname{supp}(X)$ *is bounded and*

$$\mathbf{E} \exp(c_1 \cdot Y^2) < \infty \tag{10}$$

for some constant $c_1 > 0$. Let m be the corresponding regression function, which satisfies a (p, C)-smooth generalized hierarchical interaction model of order d^* and finite level l with $p = q + s$ for some $q \in \mathbb{N}_0$ and $s \in (0, 1]$. Let $N \in \mathbb{N}_0$ with $N \ge q$. Furthermore, assume that in Definition 2(b) all partial derivatives of order less than or equal to q of the functions g_k, $f_{j,k}$ are bounded, that is, assume that each such function f satisfies

$$\max_{\substack{j_1,\ldots,j_d \in \{0,1,\ldots,q\}, \\ j_1+\cdots+j_d \le q}} \left\| \frac{\partial^{j_1+\cdots+j_d} f}{\partial^{j_1} x^{(1)} \cdots \partial^{j_d} x^{(d)}} \right\|_\infty \le c_2, \tag{11}$$

and let all functions g_k be Lipschitz continuous with Lipschitz constant $L > 0$ [which follows from (11) if $q > 0$]. Let $\mathcal{H}^{(l)}$ be defined as in (6) with K, d, d^* as in the definition of m, $M^* = \lceil c_{56} \cdot n^{\frac{d^*}{2p+d^*}} \rceil$, $\alpha = n^{c_{57}}$ for sufficiently large constants $c_{56}, c_{57} > 0$ and using an N-admissible $\sigma : \mathbb{R} \to [0, 1]$ according to Definition 3.

Let $\tilde{m}_n$ be the least squares estimate defined by (9) and define $m_n = T_{c_3 \cdot \log(n)} \tilde{m}_n$. Then

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

holds for sufficiently large n.

REMARK 1. For $p \geq 1$ and $C \geq 1$, the class of (p, C)-smooth generalized hierarchical interaction models of order d^* satisfying the assumptions of Theorem 1 contains all (p, C)-smooth functions, which depend at the most on d^* of its input components. This is because in the definition of generalized hierarchical interaction models all functions that occur in Definition 2 might be chosen as projections. Consequently, the rate of convergence in Theorem 1 is optimal up to some logarithmic factor according to Stone (1982).

REMARK 2. Some parameters of the estimate m_n considered in Theorem 1 (like l, K or d^*) can be unknown in practice. Then they have to be chosen in a data-dependent way. Several adaptive choices of parameters and their effects have been studied in the literature. We refer to Chapters 7 and 8 in Györfi et al. (2002), for example.

REMARK 3. Condition (10) in Theorem 1 prevents heavy tails and ensures that the distribution of Y is sufficiently concentrated in order to allow good estimates. It is satisfied by many common distributions like the normal distribution.

COROLLARY 1. *Let $(X, Y), (X_1, Y_1), \ldots, (X_n, Y_n)$ be independent and identically distributed random variables with values in $\mathbb{R}^d \times \mathbb{R}$ such that* $\operatorname{supp}(X)$ *is bounded and $\mathbf{E} \exp(c_1 \cdot Y^2) < \infty$ for some constant $c_1 > 0$. Let m be the corresponding regression function, which satisfies a $(2, C)$-smooth generalized hierarchical interaction model of order 2 and finite level 0. Furthermore, assume that in Definition 2*(b) *all partial derivatives of order less than or equal to 1 of the functions g_k, $f_{j,k}$ are bounded. Let $\mathcal{H}^{(0)} = \mathcal{F}^{(\text{neural networks})}_{M^*,2,d,\alpha}$ be defined as before with $M^* = \lceil c_{56} \cdot n^{\frac{1}{3}} \rceil$, $\alpha = n^{c_{57}}$, and using $\sigma(z) = \frac{1}{1+\exp(-z)}$ $(z \in \mathbb{R})$. Let $\tilde{m}_n$ be the least squares estimate defined by (9) and define $m_n = T_{c_3 \cdot \log(n)} \tilde{m}_n$. Then*

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

holds for sufficiently large n.

PROOF. Using the notation from Theorem 1, we can choose $N = q = 1$ and the sigmoidal squasher σ is 1-admissible (see Supplement A in Bauer and Kohler (2019) for an explanation). Then the application of Theorem 1 implies the corollary. □

3. Application to simulated data. To illustrate how the introduced nonparametric regression estimate based on our special type of multilayer feedforward neural networks behaves in case of finite sample sizes, we apply it to simulated data and compare the results with conventional estimates using the software *MATLAB*. Particularly, in connection with small sample sizes, the number of different approaches for the estimation of high-dimensional functions is rather limited. All the examined approaches, including ours, contain some parameters (specified subsequently) that have an influence on their behavior. In the following, we choose these parameters in a data-dependent way by splitting of the sample. This means that $n_{\text{train}} = \lceil \frac{4}{5} \cdot n \rceil$ realizations are used to train the estimate several times with different choices for the parameters each time, whereas $n_{test} = n - n_{\text{train}}$ realizations are used to test by comparison of the empirical L_2 risk on this set, which parameter assignment leads to the best estimate according to this criterion.

The first alternative approach we consider is a simple nearest neighbor estimate (*neighbor*). This means that the function value at a given point x is approximated by the average of the values $Y_1, \ldots, Y_{k_n}$ observed for the data points $X_1, \ldots, X_{k_n}$, which are closest to x with respect to the Euclidean norm (choosing the smallest index in case of ties). The parameter $k_n \in \mathbb{N}$, which denotes the number of involved neighbors, is chosen adaptively from $\{1, 2, 3\} \cup \{4, 8, 12, 16, \ldots, 4 \cdot \lfloor \frac{n_{\text{train}}}{4} \rfloor\}$ in our simulations.

The second competitive approach we examine is interpolation with radial basis functions (*RBF*). With regard to the variety of modifications of this approach known in the literature, we focus on the version in Lazzaro and Montefusco (2002), where Wendland's compactly supported radial basis function $\phi(r) = (1 - r)_+^6 \cdot (35r^2 + 18r + 3)$ is used. The radius that scales the basis functions is also selected adaptively from the set $\{0.1, 0.5, 1, 5, 30, 60, 100\}$ in our implementation, because doing so improved the RBF approach in the simulations.

The parameters l, K, d^*, M^* of our neural network estimate (*neural-x*) defined in Theorem 1 are chosen in a data-dependent way as well. The selected values of these parameters to be tested were $\{0, 1, 2\}$ for l, $\{1, \ldots, 5\}$ for K, $\{1, \ldots, d\}$ for d^*, and $\{1, \ldots, 5, 6, 11, 16, 21, \ldots, 46\}$ for M^*, although the set of possible choices is reduced for some settings if several test runs show that the whole range of choices is not needed. To solve the least squares problem in (9), we use the quasi-Newton method of the function *fminunc* in *MATLAB* to approximate its solution.

Furthermore, we compare our neural network estimate, which is characterized by the data-dependent choice of its structure and not completely connected neurons, to more ordinary fully connected neural networks with predefined numbers of layers but adaptively chosen numbers of neurons per layer. In this context, we examine structures with one hidden layer that consists of 5, 10, 25, 50 or 75 neurons (*neural-1*), three hidden layers that consist of 3, 6, 9, 12 or 15 neurons (*neural-3*), and six hidden layers that consist of 2, 4, 6, 8 or 10 neurons (*neural-6*).

The functions we use in the illustrative simulated settings to compare the different approaches are listed below:

$$m_1(x) = \cot\left(\frac{\pi}{1+\exp(x_1^2 + 2\cdot x_2 + \sin(6\cdot x_4^3) - 3)}\right) + \exp(3\cdot x_3 + 2\cdot x_4 - 5\cdot x_5 + \sqrt{x_6 + 0.9\cdot x_7 + 0.1}) \qquad (x \in [0,1]^7),$$

$$m_2(x) = \frac{2}{x_1 + 0.008} + 3\cdot\log(x_2^7\cdot x_3 + 0.1)\cdot x_4 \qquad (x \in [0,1]^7),$$

$$m_3(x) = 2\cdot\log(x_1\cdot x_2 + 4\cdot x_3 + |\tan(x_4)| + 0.1) + x_3^4\cdot x_5^2\cdot x_6 - x_4\cdot x_7 + (3\cdot x_8^2 + x_9 + 2)^{0.1+4\cdot x_{10}^2} \qquad (x \in [0,1]^{10}),$$

$$m_4(x) = x_1 + \tan(x_2) + x_3^3 + \log(x_4 + 0.1) + 3\cdot x_5 + x_6 + \sqrt{x_7 + 0.1} \qquad (x \in [0,1]^7),$$

$$m_5(x) = \exp(\|x\|) \qquad (x \in [0,1]^7),$$

$$m_6(x) = m_1\left(\frac{1}{2}\cdot|O|\cdot x\right) \qquad (x \in [0,1]^7).$$

The examples m_1, m_2 and m_3 represent some ordinary general hierarchical interaction models (cf. Definition 2), whereas m_4, m_5 and m_6 carry the definition to the extremes, such that m_4 is just an additive model, that is, $d^* = 1$, and m_5 is an interaction model with $d^* = d$. Function m_6 was added due to the interest of a referee in a modified and "less sparse" version of our existing examples. The matrix $O \in [-1,1]^{7\times 7}$ therein is a randomly generated (but fixed) dense orthogonal matrix (see Supplement B in Bauer and Kohler (2019) for details) which also corresponds to $d^* = d$.

The n observations (for $n \in \{100, 200\}$) of the type (X, Y), which are available for all estimates, are generated by

$$Y = m_i(X) + \sigma_j\cdot\lambda_i\cdot\varepsilon \qquad (i \in \{1,2,3,4,5,6\},\ j \in \{1,2\})$$

for $\sigma_j \geq 0$ and $\lambda_i \geq 0$, where X is uniformly distributed on $[0,1]^d$ (here an additional index i at d, X, and Y is neglected) and ε is standard normally distributed and independent of X. For reasons of comparability, we choose λ_i in a way that respects the range covered by m_i in the most common situations based on the distribution of X. This range is determined empirically as the interquartile range of 10^5 independent realizations of $m_i(X)$ (and stabilized by taking the median of a hundred repetitions of this procedure), which leads to $\lambda_1 = 9.11$, $\lambda_2 = 5.68$, $\lambda_3 = 13.97$, $\lambda_4 = 1.77$, $\lambda_5 = 1.64$ and $\lambda_6 = 2.47$ (rounded to two decimal places). The parameters scaling the noise are fixed as $\sigma_1 = 5\%$ and $\sigma_2 = 20\%$.

To examine the quality of an estimate $m_{n,i}$ for a correct function m_i in one of the above settings, we consider an empirical L_2 error, which is motivated by the

desired properties of a regression estimate from Section 1.1 and Theorem 1. We define it as

$$\varepsilon_{L_2,\bar{N}}(m_{n,i}) = \frac{1}{\bar{N}} \sum_{k=1}^{\bar{N}} \big(m_{n,i}(X_k) - m_i(X_k)\big)^2,$$

where $X_1, X_2, \ldots, X_{\bar{N}}$ are completely new independent realizations of the random variable X (different from the first n given data points for the estimate). Here, we choose $\bar{N} = 10^5$. Since this error strongly depends on the behavior of the correct function m_i, we consider it in relation to the error of the simplest estimate for m_i 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 you obtain if you plug 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.

In view of the fact that simulation results depend on the randomly chosen data points, we compute the estimates 50 times for repeatedly generated realizations of X and examine the median (plus interquartile range IQR) of $\varepsilon_{L_2,\bar{N}}(m_{n,i})/\bar{\varepsilon}_{L_2,\bar{N}}(avg)$. The results can be found in Tables 1 and 2.

In these simulations, the occurring chosen values for the parameters of our estimate vary between 0 and 2 for l, between 1 and 2 for K and d^* and between 1 and 46 for M^*. For the latter one the average occurring value in the simulations is reported in Table 3.

We observe that our estimate outperforms the other approaches in the three typical examples for generalized hierarchical interaction models m_1, m_2, and m_3. Especially in the nested case with the highest dimension, m_3, the error of our estimate is roughly six to seven times smaller than the error of the second best approach for $n = 200$. A remarkable fact is that in these cases, the relative improvement of our estimate with an increasing sample size is often much larger than the improvement of the other approaches. This result is a plausible indicator of a better rate of convergence.

With regard to the extreme cases of m_4, m_5 and m_6, our approach is not always the best although it surprisingly performs well even here in some situations. However, neither the additive model m_4 nor the function m_5, which is rather densely connected in the sense of interaction models because all components interact in only one function, are perfectly imitated by our sparsely connected neural network estimate. In case of m_6, we can observe that our estimate very often performs just as well as the best other approach or provides the best results itself.

Furthermore, it makes sense that in some of the examined test settings where our estimate leads to good approximations, one of the fully connected neural network approaches is reasonably good as well. This happens because some of our

TABLE 1
Median and IQR of the scaled empirical L_2 error of estimates for m_1, m_2 and m_3

	m_1			
Noise	**5%**		**20%**	
Sample size	$n = 100$	$n = 200$	$n = 100$	$n = 200$
$\bar{\varepsilon}_{L_2,\bar{N}}(avg)$	**596.52**	**597.61**	**596.51**	**597.63**
Approach	**Median (IQR)**	**Median (IQR)**	**Median (IQR)**	**Median (IQR)**
neural-1	0.2622 (2.7248)	0.1064 (0.3507)	0.3004 (2.1813)	0.1709 (3.8163)
neural-3	0.1981 (0.4732)	0.0609 (0.1507)	0.2784 (0.4962)	0.0848 (0.1239
neural-6	0.2953 (0.9293)	0.1207 (0.1672)	0.2663 (0.5703	0.1106 (0.2412)
neural-x	**0.0497 (0.2838)**	**0.0376 (0.2387)**	**0.0596 (0.2460)**	**0.0200 (0.1914)**
RBF	0.3095 (0.4696)	0.1423 (0.0473)	0.3182 (0.5628)	0.1644 (0.0639)
neighbor	0.6243 (0.1529)	0.5398 (0.1469)	0.6303 (0.1014)	0.5455 (0.1562)

	m_2			
Noise	**5%**		**20%**	
Sample size	$n = 100$	$n = 200$	$n = 100$	$n = 200$
$\bar{\varepsilon}_{L_2,\bar{N}}(avg)$	**407.56**	**408.34**	**407.45**	**408.47**
Approach	**Median (IQR)**	**Median (IQR)**	**Median (IQR)**	**Median (IQR)**
neural-1	0.9135 (4.6170)	0.3644 (1.4536)	0.7563 (0.9990)	0.6935 (2.8923)
neural-3	0.7010 (0.8556)	**0.1000 (0.1471)**	0.6871 (0.6646)	0.3456 (0.4573)
neural-6	0.5809 (1.0208)	0.1468 (0.5747)	0.8678 (1.2043)	0.3128 (0.4199)
neural-x	**0.4838 (1.0463)**	0.1049 (0.1574)	**0.5271 (1.4364)**	**0.1682 (0.2816)**
RBF	0.9993 (0.1301)	0.9232 (0.2180)	0.9823 (0.2503)	0.8873 (0.2316)
neighbor	0.8681 (0.0646)	0.8299 (0.0640)	0.8807 (0.0682)	0.8519 (0.0611)

	m_3			
Noise	**5%**		**20%**	
Sample size	$n = 100$	$n = 200$	$n = 100$	$n = 200$
$\bar{\varepsilon}_{L_2,\bar{N}}(avg)$	**5492.87**	**5461.66**	**5477.62**	**5476.46**
Approach	**Median (IQR)**	**Median (IQR)**	**Median (IQR)**	**Median (IQR)**
neural-1	0.5300 (0.2200)	0.2390 (0.1890)	0.6370 (0.2580)	0.2460 (0.1320)
neural-3	0.7180 (0.1340)	0.4280 (0.0940)	0.7190 (0.0910)	0.3980 (0.1100)
neural-6	0.9520 (0.1120)	0.5470 (0.0960)	0.9670 (0.1020)	0.5620 (0.1220)
neural-x	**0.1277 (0.2609)**	**0.0336 (0.0728)**	**0.1616 (0.9936)**	**0.0420 (0.2148)**
RBF	0.8249 (0.3896)	0.6661 (0.4597)	1.0020 (0.3357)	0.6676 (0.4433)
neighbor	0.8772 (0.0936)	0.7935 (0.0903)	0.8675 (0.0920)	0.8237 (0.0967)

TABLE 2
Median and IQR of the scaled empirical L_2 error of estimates for m_4, m_5 and m_6

	m_4			
Noise	**5%**		**20%**	
Sample size	$n = 100$	$n = 200$	$n = 100$	$n = 200$
$\bar{\varepsilon}_{L_2,\bar{N}}(avg)$	**1.60**	**1.59**	**1.61**	**1.61**
Approach	**Median (IQR)**	**Median (IQR)**	**Median (IQR)**	**Median (IQR)**
neural-1	**0.0140** (0.0040)	**0.0050** (0.0020)	**0.0370** (0.0150)	**0.0240** (0.0090)
neural-3	0.0160 (0.0060)	0.0080 (0.0020)	0.0450 (0.0110)	**0.0240** (0.0050)
neural-6	0.0210 (0.0080)	0.0090 (0.0030)	0.0530 (0.0130)	0.0290 (0.0090)
neural-x	0.0311 (0.1026)	0.0085 (0.0205)	0.2623 (1.5689)	0.1042 (0.2296)
RBF	0.0188 (0.0084)	0.0148 (0.0030)	0.1594 (0.0589)	0.1386 (0.0299)
neighbor	0.3024 (0.07565)	0.2033 (0.0321)	0.2868 (0.0952)	0.2211 (0.0355)

	m_5			
Noise	5%		20%	
Sample size	$n = 100$	$n = 200$	$n = 100$	$n = 200$
$\bar{\varepsilon}_{L_2,\bar{N}}(avg)$	**1.49**	**1.49**	**1.49**	**1.49**
Approach	**Median (IQR)**	**Median (IQR)**	**Median (IQR)**	**Median (IQR)**
neural-1	0.7246 (9.3962)	0.0648 (0.0879)	2.0865 (75.4682)	0.6659 (26.0015)
neural-3	0.3954 (0.9887)	0.1087 (0.1909)	1.5671 (7.0394)	0.2370 (1.4065)
neural-6	0.1023 (0.3572)	0.0716 (0.0760)	0.2482 (0.6611)	**0.0836 (0.1646)**
neural-x	0.1386 (0.4205)	0.0637 (0.0499)	0.3699 (1.3039)	0.1854 (0.3660)
RBF	**0.0127 (0.0044)**	**0.0112 (0.0033)**	**0.1445 (0.0671)**	0.1352 (0.0298)
neighbor	0.3263 (0.0842)	0.2471 (0.0381)	0.3360 (0.0707)	0.2620 (0.0464)

	m_6			
Noise	**5%**		**20%**	
Sample size	$n = 100$	$n = 200$	$n = 100$	$n = 200$
$\bar{\varepsilon}_{L_2,\bar{N}}(avg)$	**4.43**	**4.41**	**4.42**	**4.41**
Approach	**Median (IQR)**	**Median (IQR)**	**Median (IQR)**	**Median (IQR)**
neural-1	1.7664 (379.8016)	0.0556 (24.5402)	0.4763 (39.8014)	18.6135 (462.5238)
neural-3	0.0374 (0.1204)	0.0161 (0.0317)	0.8293 (6.3857)	0.1421 (0.8679)
neural-6	0.0656 (0.1071)	0.0165 (0.0596)	0.1870 (0.5070)	**0.0585 (0.0851)**
neural-x	0.0275 (0.1406)	0.0388 (0.1488)	**0.1212 (0.2715)**	0.0719 (0.4423)
RBF	**0.0240 (0.0090)**	**0.0123 (0.0038)**	0.1271 (0.0376)	0.1045 (0.0233)
neighbor	0.3343 (0.0675)	0.2328 (0.0587)	0.3466 (0.0818)	0.2467 (0.0519)

TABLE 3
Average occurring value in the simulations for M^*

Noise	**5%**		**20%**	
Sample size	$n = 100$	$n = 200$	$n = 100$	$n = 200$
m_1	4.24	4.00	2.88	4.46
m_2	4.68	4.62	3.66	5.00
m_3	3.38	2.56	2.60	2.96
m_4	4.62	4.36	3.28	5.42
m_5	4.22	12.9	4.08	11.06
m_6	3.38	2.80	3.38	3.40

sparse networks can be expressed by fully connected networks (e.g., by fixing the weights of unnecessary connections to zero), but the data-dependent adjustment of a smaller number of weights, as in the case of our estimate, is statistically easier.

As remarked by one of our referees, many practitioners prefer nonsmooth activation functions like $\sigma(z) = \max\{0, z\}$ instead of the sigmoid squasher for computational reasons. Therefore, we tested all neural networks estimates using such an activation function again. The results, which can be found in Supplement B (Bauer and Kohler (2019)), are quite good in some settings but almost always significantly worse than the results of our estimate with the sigmoid function.

As pointed out by the associate editor and by one of our referees, it could be helpful to consider the decrease of the estimation error for an increasing sample size in detail. Therefore, we computed the estimates from the simulation section once for each $n \in \{100, 120, 140, \ldots, 560\}$ in case of m_1. These errors are plotted in Figure 3 using logarithmically scaled axes and adding a regression line. Table 4 provides the exact coefficients of these lines. Our new estimate shows the steepest negative slope, which at least suggests the best rate of convergence because the slope in a logarithmically scaled representation corresponds to the exponent of n in the usual error term. Surprisingly, we get a rate of convergence faster than $\frac{1}{n}$ for the sample sizes which we consider. We believe that this is due to the fact that finite sample sizes do not always show the asymptotic behavior of the error.

4. Proofs.

4.1. *Outline of the proof of Theorem 1.* In the proof of Theorem 1, we will use the following bound on the expected L_2 error of least squares estimates.

LEMMA 1. *Let $\beta_n = c_5 \cdot \log(n)$ for some constant $c_5 > 0$. Assume that the distribution of (X, Y) satisfies*

$$\mathbf{E}\left(e^{c_6 \cdot |Y|^2}\right) < \infty \tag{12}$$

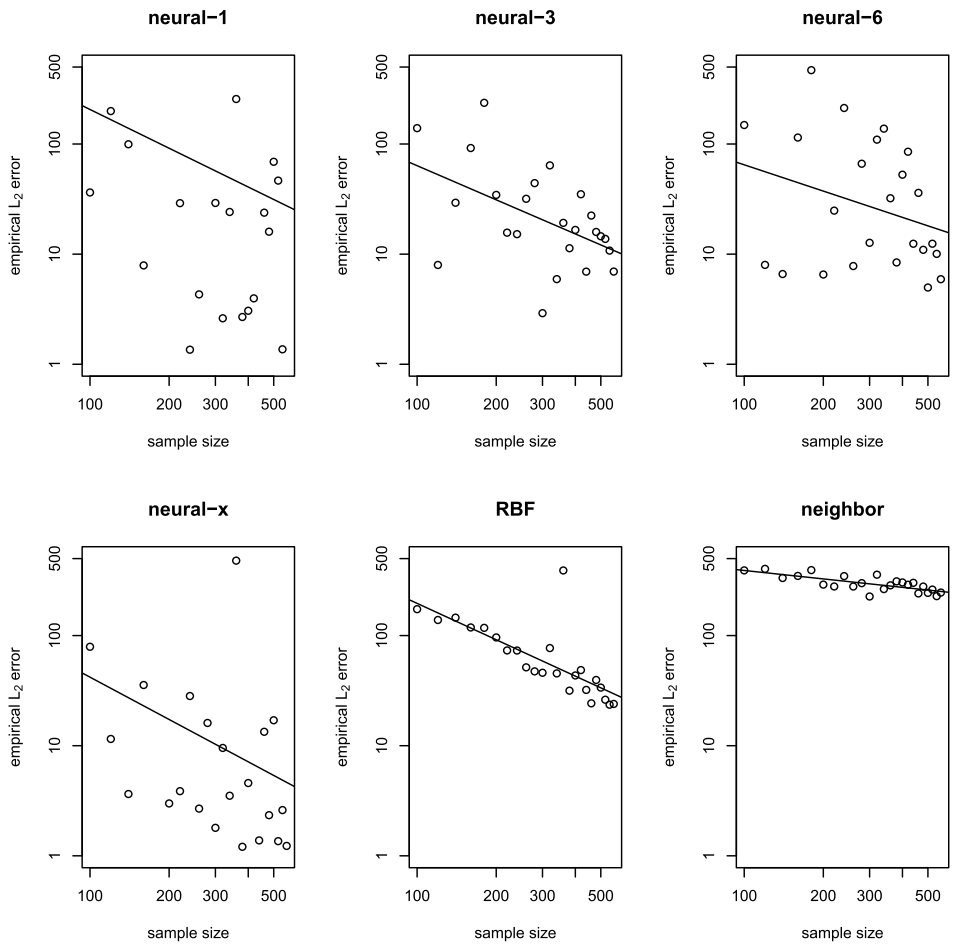

FIG. 3. *Empirical L_2 error of different estimation approaches for an increasing sample size in case of m_1.*

TABLE 4
Coefficients of the regression lines corresponding to Figure 3

Approach	Intercept	Slope
neural-1	10.711546	−1.1691606
neural-3	8.892746	−1.0294270
neural-6	7.822762	−0.7930447
neural-x	9.614176	−1.2765884
RBF	10.332239	−1.0974016
neighbor	7.138497	−0.2545437

for some constant $c_6 > 0$ and that the regression function m is bounded in absolute value. Let $\tilde{m}_n$ be the least squares estimate

$$\tilde{m}_n(\cdot) = \arg\min_{f \in \mathcal{F}_n} \frac{1}{n} \sum_{i=1}^{n} |Y_i - f(X_i)|^2$$

based on some function space $\mathcal{F}_n$ and define $m_n = T_{\beta_n} \tilde{m}_n$ using the truncation operator defined prior to Theorem 1. Then m_n satisfies

$$\begin{aligned} &\mathbf{E} \int |m_n(x) - m(x)|^2 \mu(dx) \\ &\quad \leq \frac{c_7 \cdot \log(n)^2 \cdot (\log(\mathcal{N}(\frac{1}{n \cdot \beta_n}, \mathcal{F}_n, \|\cdot\|_{\infty, \mathrm{supp}(X)})) + 1)}{n} \\ &\qquad + 2 \cdot \inf_{f \in \mathcal{F}_n} \int |f(x) - m(x)|^2 \mathbf{P}_X(dx) \end{aligned}$$

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, Clausen and Kohler (2009). A complete version of the proof can be found in Supplement A (Bauer and Kohler (2019)). □

From Lemma 1, we see that we need to bound the covering number

$$\mathcal{N}\left(\frac{1}{n \cdot \beta_n}, \mathcal{H}^{(l)}, \|\cdot\|_{\infty, \mathrm{supp}(X)}\right)$$

and the approximation error

$$\inf_{f \in \mathcal{H}^{(l)}} \int |f(x) - m(x)|^2 \mathbf{P}_X(dx) \tag{13}$$

for our class of hierarchical neural networks $\mathcal{H}^{(l)}$. Given that we assume that our sigmoidal function is Lipschitz continuous, deriving a bound on the covering number is easy. The next lemma summarizes the result.

LEMMA 2. *Let $\varepsilon_n \geq \frac{1}{n^{c_8}}$ and let $\mathcal{H}^{(l)}$ be defined as in (6) with $\max\{a_n, \alpha, M^*\} \leq n^{c_9}$ for large n and certain constants $c_8, c_9 > 0$. Then*

$$\log\left(\mathcal{N}\left(\varepsilon_n, \mathcal{H}^{(l)}, \|\cdot\|_{\infty, [-a_n, a_n]^d}\right)\right) \leq c_{10} \cdot \log(n) \cdot M^*$$

holds for sufficiently large n and a constant $c_{10} > 0$ independent of n.

PROOF. The assertion follows by a straightforward modification of the proof of Lemma 8 in Kohler and Krzyżak (2017). A complete proof can be found in Supplement A (Bauer and Kohler (2019)). □

The main difficulty in the proof is to bound the approximation error (13). Here, we will show that under the assumptions of Theorem 1 we have

$$\inf_{f\in\mathcal{H}^{(l)}} \int |f(x)-m(x)|^2 \mathbf{P}_X(dx) \le c_{11}\cdot \log(n)^3 \cdot n^{-\frac{2p}{2p+d^*}}.$$

For this purpose, we derive a new result concerning the approximation of (p,C)-smooth functions by multilayer feedforward neural networks with two hidden layers in Theorem 2 below.

4.2. *Approximation of smooth functions by multilayer feedforward neural networks.* The aim of this subsection is to present the following result concerning the approximation of (p,C)-smooth function by multilayer feedforward neural networks with two hidden layers.

THEOREM 2. *Let $a \ge 1$ and $p = q+s$ for some $q \in \mathbb{N}_0$ and $s \in (0,1]$, and let $C > 0$. Let $m : \mathbb{R}^d \to \mathbb{R}$ be a (p,C)-smooth function, which satisfies*

$$\max_{\substack{j_1,\dots,j_d\in\{0,1,\dots,q\},\\ j_1+\cdots+j_d\le q}} \left\| \frac{\partial^{j_1+\cdots+j_d} m}{\partial^{j_1}x^{(1)}\cdots\partial^{j_d}x^{(d)}} \right\|_{\infty,[-2a,2a]^d} \le c_{12}. \tag{14}$$

Let ν be an arbitrary probability measure on $\mathbb{R}^d$. Let $N \in \mathbb{N}_0$ be chosen such that $N \ge q$ and let $\sigma : \mathbb{R} \to [0,1]$ be N-admissible according to Definition 3. Then, for any $\eta \in (0,1)$ and $M \in \mathbb{N}$ sufficiently large (independent of the size of a and η, but $a \le M$ must hold), a neural network of the type

$$t(x) = \sum_{i=1}^{\binom{d+N}{d}\cdot(N+1)\cdot(M+1)^d} \mu_i \cdot \sigma\left(\sum_{l=1}^{4d} \lambda_{i,l}\cdot\sigma\left(\sum_{v=1}^{d}\theta_{i,l,v}\cdot x^{(v)} + \theta_{i,l,0}\right)+\lambda_{i,0}\right) \tag{15}$$

exists such that

$$|t(x)-m(x)| \le c_{13}\cdot a^{N+q+3}\cdot M^{-p}$$

holds for all $x \in [-a,a]^d$ up to a set of ν-measure less than or equal to η. The coefficients of $t(x)$ can be bounded by

$$|\mu_i| \le c_{14}\cdot a^q \cdot M^{N\cdot p},$$

$$|\lambda_{i,l}| \le M^{d+p\cdot(N+2)},$$

$$|\theta_{i,l,v}| \le 6\cdot d\cdot \frac{1}{\eta}\cdot M^{d+p\cdot(2N+3)+1}$$

for all $i \in \{1,\dots,\binom{d+N}{d}\cdot(N+1)\cdot(M+1)^d\}$, $l \in \{0,\dots,4d\}$, and $v \in \{0,\dots,d\}$.

PROOF. The overall idea of this proof is not complicated. It is based on two fundamental properties of our neural networks, which we show in several lemmas. On the one hand, neural networks with adequately chosen weights can roughly behave like multivariate polynomials. On the other hand, it is possible that neural networks almost vanish outside of a polytope. Then the constructive proof works as follows: The considered area, where the function m shall be estimated, is divided into a grid of equally large cubes. For each of these cubes, a small neural network is constructed, which roughly behaves like the taylor polynomial of m around the center of that cube and vanishes outside of it. The sum of all such networks forms a larger neural network, which leads to a good approximation on the whole considered area. Unfortunately, there are certain transition zones on the boundary of these cubes, where the required properties cannot be guaranteed. Finally, a shifting argument for the whole grid of cubes leads to the conclusion that this exception set can be bounded. Despite this easily comprehensible idea, the technical details require many laborious lemmas. Therefore, the detailed proof is outsourced to Supplement A in Bauer and Kohler (2019). □

4.3. *Approximation of smooth generalized hierarchical interaction models by multilayer feedforward neural networks.* In this subsection, we use Theorem 2 to derive the following result concerning the approximation of (p, C)-smooth generalized hierarchical interaction models by multilayer feedforward neural networks.

THEOREM 3. *Let X be a $\mathbb{R}^d$-valued random variable and let $m : \mathbb{R}^d \to \mathbb{R}$ satisfy a (p, C)-smooth generalized hierarchical interaction model of order d^* and finite level l with $p = q + s$, where $q \in \mathbb{N}_0$ and $s \in (0, 1]$. Let $N \in \mathbb{N}_0$ with $N \geq q$. Assume that in Definition 2(b), all partial derivatives of the order less than or equal to q of the functions g_k, $f_{j,k}$ are bounded, that is, let us assume that each such function f satisfies*

$$\max_{\substack{j_1,\dots,j_d \in \{0,1,\dots,q\},\\ j_1+\cdots+j_d \leq q}} \left\| \frac{\partial^{j_1+\cdots+j_d} f}{\partial^{j_1} x^{(1)} \cdots \partial^{j_d} x^{(d)}} \right\|_\infty \leq c_{28}, \tag{16}$$

and let all functions g_k be Lipschitz continuous with Lipschitz constant $L > 0$ [which follows from (16) if $q > 0$]. Let $M_n \in \mathbb{N}$ and let $1 \leq a_n \leq M_n$ be increasing such that $a_n^{N+q+3} \leq M_n^p$ is satisfied for n sufficiently large. Let $\eta_n \in (0, 1]$. Let $\mathcal{H}^{(l)}$ be defined as in (6) with K, d, d^* as in the definition of m, $M^* = \binom{d^*+N}{d^*} \cdot (N+1) \cdot (M_n+1)^{d^*}$, $\alpha = \log(n) \cdot \frac{M_n^{d^*+p\cdot(2N+3)+1}}{\eta_n}$, and using an N-admissible $\sigma : \mathbb{R} \to [0, 1]$ according to Definition 3. Then, for arbitrary $c > 0$ and all n greater than a certain $n_0(c) \in \mathbb{N}$, $t \in \mathcal{H}^{(l)}$ exists such that outside of a set of $\mathbf{P}_X$-measure less than or equal to $c \cdot \eta_n$ we have

$$\left|t(x) - m(x)\right| \leq c_{29} \cdot a_n^{N+q+3} \cdot M_n^{-p}$$

for all $x \in [-a_n, a_n]^d$ *and with* c_{29} *independent of the other factors on the right-hand side (that are variable by* n*), but depending on fixed values (like* c, d, d^**). Furthermore, this* t *can be chosen in such a way, that*

$$|t(x)| \le c_{30} \cdot a_n^q \cdot M_n^{d^*+N\cdot p}$$

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

PROOF. This theorem follows by induction from Theorem 2. A complete proof can be found in Supplement A (Bauer and Kohler (2019)). □

4.4. *Proof of Theorem 1.* Let $a_n = \log(n)^{\frac{3}{2\cdot(N+q+3)}}$. For a sufficiently large n, the relation $\operatorname{supp}(X) \subseteq [-a_n, a_n]^d$ holds, which implies $\mathcal{N}(\delta, \mathcal{G}, \|\cdot\|_{\infty,\operatorname{supp}(X)}) \le \mathcal{N}(\delta, \mathcal{G}, \|\cdot\|_{\infty,[-a_n,a_n]^d})$ for an arbitrary function space $\mathcal{G}$ and $\delta > 0$. Then applying Lemma 1 leads to

$$\begin{aligned}
&\mathbf{E}\int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx) \\
&\quad \le \frac{c_7 \cdot \log(n)^2 \cdot (\log(\mathcal{N}(\frac{1}{n\cdot c_3\cdot\log(n)}, \mathcal{H}^{(l)}, \|\cdot\|_{\infty,[-a_n,a_n]^d})) + 1)}{n} \\
&\qquad + 2 \cdot \inf_{h\in\mathcal{H}^{(l)}} \int |h(x) - m(x)|^2 \mathbf{P}_X(dx).
\end{aligned}$$

Due to the fact that $\frac{1}{n\cdot c_3\cdot\log(n)} \ge \frac{1}{n^{c_8}}$ and $\max\{a_n, \alpha, M^*\} \le n^{c_9}$ hold for certain constants $c_8, c_9 > 0$, Lemma 2 allows us to bound the first summand by

$$c_7 \cdot \log(n)^2 \cdot \frac{c_{10}\cdot\log(n)\cdot M^*}{n} \le c_{35} \cdot \log(n)^3 \cdot n^{-\frac{2p}{2p+d^*}}$$

for a sufficiently large n.

Regarding the second summand, we would like to use Theorem 3. Therefore, we set $M_n = \lceil n^{\frac{1}{2p+d^*}} \rceil$ and $\eta_n = \log(n)^{\frac{3\cdot(N+3)}{N+q+3}} \cdot n^{-\frac{2\cdot(N+1)\cdot p+2d^*}{2p+d^*}}$. The resulting values for M^* and α therein (which are defined depending on M_n and η_n) are consistent with the specifications $M^* = \lceil c_{56} \cdot n^{\frac{d^*}{2p+d^*}} \rceil$ and $\alpha = n^{c_{57}}$ in Theorem 1 for sufficiently large constants $c_{56}, c_{57} > 0$. Even if the specification $\alpha = n^{c_{57}}$ in Theorem 1 leads to a larger value than in Theorem 3, the corresponding version of $\mathcal{H}^{(l)}$ contains the approximation from Theorem 3 all the more. If we choose a $h^* \in \mathcal{H}^{(l)}$ such that it satisfies the approximation properties of Theorem 3 using the above a_n and M_n, and denote the exception set with measure η_n therein by D_n, we can bound $\inf_{h\in\mathcal{H}^{(l)}} \int |h(x) - m(x)|^2 \mathbf{P}_X(dx)$ by

$$\begin{aligned}
&\int |h^*(x) - m(x)|^2 \cdot 1_{D_n^C} \mathbf{P}_X(dx) + \int |h^*(x) - m(x)|^2 \cdot 1_{D_n} \mathbf{P}_X(dx) \\
&\quad \le \left(c_{29} \cdot a_n^{(N+q+3)} \cdot M_n^{-p}\right)^2 + \left(2 \cdot c_{30} \cdot a_n^q \cdot M_n^{d^*+N\cdot p}\right)^2 \cdot \eta_n
\end{aligned}$$

$$\leq c_{36} \cdot \log(n)^3 \cdot n^{-\frac{2p}{2p+d^*}}$$
$$+ c_{37} \cdot \log(n)^{\frac{3q}{N+q+3}} \cdot n^{\frac{2d^*+2N\cdot p}{2p+d^*}} \cdot \log(n)^{\frac{3\cdot(N+3)}{N+q+3}} \cdot n^{-\frac{2\cdot(N+1)\cdot p+2d^*}{2p+d^*}}$$
$$\leq c_{11} \cdot \log(n)^3 \cdot n^{-\frac{2p}{2p+d^*}},$$

where we assumed $m(x) \leq c_{30} \cdot a_n^q \cdot M_n^{d^*+N\cdot p}$ on $\operatorname{supp}(X)$ in the second integral, which is true for a sufficiently large n because of the assumptions of the theorem. This proves the theorem.

Acknowledgments. The authors would like to thank the German Research Foundation (DFG) for funding this project within the Collaborative Research Centre 805. Furthermore, the authors would like to thank an anonymous Associate Editor and three anonymous referees for their invaluable comments improving an early version of this manuscript.

SUPPLEMENTARY MATERIAL

Supplement A: Further proofs (DOI: 10.1214/18-AOS1747SUPPA; .pdf). This supplementary file contains the rather technical proofs of several lemmas and assertions in this article.

Supplement B: Further simulation results (DOI: 10.1214/18-AOS1747SUPPB; .pdf). This file contains the results of some experiments with another activation function in the neural network estimates.

REFERENCES

ANTHONY, M. and BARTLETT, P. L. (1999). *Neural Network Learning*: *Theoretical Foundations*. Cambridge Univ. Press, Cambridge. MR1741038

BAGIROV, A. M., CLAUSEN, C. and KOHLER, M. (2009). Estimation of a regression function by maxima of minima of linear functions. *IEEE Trans. Inform. Theory* **55** 833–845. MR2597271

BARRON, A. R. (1991). Complexity regularization with application to artificial neural networks. In *Nonparametric Functional Estimation and Related Topics* (*Spetses*, 1990). *NATO Adv. Sci. Inst. Ser. C Math. Phys. Sci.* **335** 561–576. Kluwer Academic, Dordrecht. MR1154352

BARRON, A. R. (1993). Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Trans. Inform. Theory* **39** 930–945. MR1237720

BARRON, A. R. (1994). Approximation and estimation bounds for artificial neural networks. *Mach. Learn.* **14** 115–133.

BAUER, B. and KOHLER, M. (2019). Supplement to "On deep learning as a remedy for the curse of dimensionality in nonparametric regression." DOI:10.1214/18-AOS1747SUPPA, DOI:10.1214/18-AOS1747SUPPB.

DEVROYE, L., GYÖRFI, L. and LUGOSI, G. (1996). *A Probabilistic Theory of Pattern Recognition*. *Applications of Mathematics* (*New York*) **31**. Springer, New York. MR1383093

ELDAN, R. and SHAMIR, O. (2015). The power of depth for feedforward neural networks. Arxiv preprint.

FRIEDMAN, J. H. and STUETZLE, W. (1981). Projection pursuit regression. *J. Amer. Statist. Assoc.* **76** 817–823. MR0650892

GYÖRFI, L., KOHLER, M., KRZYŻAK, A. and WALK, H. (2002). *A Distribution-Free Theory of Nonparametric Regression*. Springer, New York. MR1920390

HÄRDLE, W., HALL, P. and ICHIMURA, H. (1993). Optimal smoothing in single-index models. *Ann. Statist.* **21** 157–178. MR1212171

HÄRDLE, W. and STOKER, T. M. (1989). Investigating smooth multiple regression by the method of average derivatives. *J. Amer. Statist. Assoc.* **84** 986–995. MR1134488

HAYKIN, S. O. (2008). *Neural Networks and Learning Machines*, 3rd ed. Prentice Hall, New York.

HERTZ, J., KROGH, A. and PALMER, R. G. (1991). *Introduction to the Theory of Neural Computation*. Addison-Wesley, Redwood City, CA. MR1096298

HOROWITZ, J. L. and MAMMEN, E. (2007). Rate-optimal estimation for a general class of nonparametric regression models with unknown link functions. *Ann. Statist.* **35** 2589–2619. MR2382659

KOHLER, M. and KRZYŻAK, A. (2005). Adaptive regression estimation with multilayer feedforward neural networks. *J. Nonparametr. Stat.* **17** 891–913. MR2192165

KOHLER, M. and KRZYŻAK, A. (2017). Nonparametric regression based on hierarchical interaction models. *IEEE Trans. Inform. Theory* **63** 1620–1630. MR3625984

KONG, E. and XIA, Y. (2007). Variable selection for the single-index model. *Biometrika* **94** 217–229. MR2367831

LAZZARO, D. and MONTEFUSCO, L. B. (2002). Radial basis functions for the multivariate interpolation of large scattered data sets. *J. Comput. Appl. Math.* **140** 521–536.

LUGOSI, G. and ZEGER, K. (1995). Nonparametric estimation via empirical risk minimization. *IEEE Trans. Inform. Theory* **41** 677–687. MR1331260

MCCAFFREY, D. F. and GALLANT, A. R. (1994). Convergence rates for single hidden layer feedforward networks. *Neural Netw.* **7** 147–158.

MHASKAR, H. N. and POGGIO, T. (2016). Deep vs. shallow networks: An approximation theory perspective. *Anal. Appl. (Singap.)* **14** 829–848. MR3564936

MIELNICZUK, J. and TYRCHA, J. (1993). Consistency of multilayer perceptron regression estimators. *Neural Netw.* **6** 1019–1022.

RIPLEY, B. D. (2008). *Pattern Recognition and Neural Networks*. Cambridge Univ. Press, Cambridge. Reprint of the 1996 original. MR2451352

SCHMIDHUBER, J. (2015). Deep learning in neural networks: An overview. *Neural Netw.* **61** 85–117.

SCHMIDT-HIEBER, J. (2017). Nonparametric regression using deep neural networks with ReLU activation function. Available at arXiv:1708.06633v2.

STONE, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *Ann. Statist.* **10** 1040–1053. MR0673642

STONE, C. J. (1985). Additive regression and other nonparametric models. *Ann. Statist.* **13** 689–705. MR0790566

STONE, C. J. (1994). The use of polynomial splines and their tensor products in multivariate function estimation. *Ann. Statist.* **22** 118–184. MR1272079

YU, Y. and RUPPERT, D. (2002). Penalized spline estimation for partially linear single-index models. *J. Amer. Statist. Assoc.* **97** 1042–1054. MR1951258

FACHBEREICH MATHEMATIK
TU DARMSTADT
SCHLOSSGARTENSTR. 7
64289 DARMSTADT
GERMANY
E-MAIL: bbauer@mathematik.tu-darmstadt.de
kohler@mathematik.tu-darmstadt.de