On deep learning as a remedy for the curse of dimensionality in nonparametric regression *

Benedikt Bauer† and Michael Kohler
Fachbereich Mathematik, Technische Universität Darmstadt, Schlossgartenstr. 7, 64289 Darmstadt, Germany, email: bbauer@mathematik.tu-darmstadt.de, kohler@mathematik.tu-darmstadt.de

March 23, 2017

Abstract
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.

AMS classification: Primary 62G08; secondary 62G20.

Key words and phrases: curse of dimensionality, neural networks, nonparametric regression, rate of convergence.

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, i.e., 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

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

*Running title: *On nonparametric regression by deep learning*

†Corresponding author. Tel: +49-6151-16-22848 , Fax: +49-6151-16-23381

(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.
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),...,(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 Universal consistency

A sequence of estimates m_n is called *weakly universally consistent* if

$$\mathbf{E} \int_{\mathbb{R}^d} |m_n(x) - m(x)|^2 \mathbf{P}_X(dx) \to 0 \quad (n \to \infty)$$

for every distribution of (X, Y) with $\mathbf{E}Y^2 < \infty$. The sequence is called *strongly universally consistent* if

$$\int_{\mathbb{R}^d} |m_n(x) - m(x)|^2 \mathbf{P}_X(dx) \to 0 \quad a.s.$$

for every distribution of (X, Y) with $\mathbf{E}Y^2 < \infty$.

Stone (1977) showed for the first time that weakly universally consistent estimates exist. Later, this result, which was proven for nearest neighbor estimates, was extended by many additional results concerning weak and strong universal consistency of various estimates. Györfi et al. (2002) provide a list of references.

1.3 Slow rate

Universal consistency implies that the L_2 error of the estimate converges to zero for all distributions as the sample size tends to infinity. However, it says nothing about the rate of convergence of the L_2 error towards zero. In view of applications, where one has a

finite sample size, it would be very interesting to have results which imply that the L_2 error converges to zero with some given rate of convergence for all distributions.

Unfortunately, such results do not exist. Theorem 7.2 and Problem 7.2 in Devroye, Györfi and Lugosi (1996) and Section 3 in Devroye and Wagner (1980) imply the following slow rate of convergence result: Let $\{a_n\}$ be a sequence of positive numbers converging to zero with $1/64 \geq a_1 \geq a_2 \geq \cdots$. Then, for every sequence of regression estimates, a distribution of (X, Y) exists such that X is uniformly distributed, $Y = m(X)$ and $\mathbf{E} \int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx) \geq a_n$ for all n.

1.4 Rate of convergence

As we have seen above, one has to restrict the class of regression functions that one considers to obtain non-trivial 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 $\frac{\partial^q m}{\partial x_1^{\alpha_1} \ldots \partial x_d^{\alpha_d}}$ exists and satisfies*

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

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

Stone (1982) 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

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

1.5 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

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

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

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

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 $1 \leq i_1 < \cdots < i_{d^*} \leq d$ 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

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

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, i.e.,

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

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

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

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 (2016).

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*

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

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

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

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.*

This definition includes all the other types of structures of m mentioned earlier. 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 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 $r + 1$.

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

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

e.g., the so-called sigmoidal or logistic squasher

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

Most existing theoretical results concerning neural networks consider neural networks using only one hidden layer, i.e., functions of the form

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

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)).

1.7 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. 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. To prove our theoretical result, we derive new approximation results for neural networks with several hidden layers, bounded weights, and a bounded number of hidden neurons.

1.8 Notation

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 $\mathbb{R}$, 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

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

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 $f_1, \ldots, f_N : \mathbb{R}^d \to \mathbb{R}$ is called an ε–$\|\cdot\|_{\infty, A}$– cover of $\mathcal{F}$ if for any $f \in \mathcal{F}$ there exists $i \in \{1, \ldots, N\}$ 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.

1.9 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, N, d^*, d and level l as follows.

For $M \in \mathbb{N}$, $N \in \mathbb{N}_0$, $d \in \mathbb{N}$, $d^* \in \{1, \dots, d\}$ and $\alpha > 0$, we denote the set of all functions $f : \mathbb{R}^d \to \mathbb{R}$ that satisfy

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

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

$$|\mu_i| \leq \alpha, \quad |\lambda_{i,l}| \leq \alpha, \quad |\theta_{i,l,m}| \leq \alpha$$

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

$$\begin{aligned} W\left(\mathcal{F}_{M,N,d^*,d,\alpha}^{(neural\ networks)}\right) &:= \binom{d^*+N}{d^*} \cdot (N+1) \cdot (M+1)^{d^*} + 1 \\ &\quad + \binom{d^*+N}{d^*} \cdot (N+1) \cdot (M+1)^{d^*} \cdot (4d^*+1) \\ &\quad + \binom{d^*+N}{d^*} \cdot (N+1) \cdot (M+1)^{d^*} \cdot 4d^* \cdot (d+1) \\ &= \binom{d^*+N}{d^*} \cdot (N+1) \cdot (M+1)^{d^*} \cdot (4d^* \cdot (d+2)+2) + 1 \quad (5) \end{aligned}$$

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

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

For $l > 0$, we define recursively

$$\begin{aligned} \mathcal{H}^{(l)} \quad = \quad \Bigg\{ h : \mathbb{R}^d \to \mathbb{R} \quad : \quad & h(x) = \sum_{k=1}^{K} g_k \left(f_{1,k}(x), \dots, f_{d^*,k}(x) \right) \quad (x \in \mathbb{R}^d) \\ & \text{for some } g_k \in \mathcal{F}_{M,N,d^*,d^*,\alpha}^{(neural\ networks)} \text{ and } f_{j,k} \in \mathcal{H}^{(l-1)} \Bigg\}. \quad (6) \end{aligned}$$

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 $l > 0$ the class $\mathcal{H}^{(l)}$ is a set of neural networks with $2 \cdot l + 2$ hidden layers. Furthermore, let $N\left(\mathcal{H}^{(l)}\right)$ denote

the number of linked two-layered neural networks from $\mathcal{F}_{M,N,d^*,d,\alpha}^{(neural\ networks)}$ that define the functions from $\mathcal{H}^{(l)}$. Then the recursion

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

holds, yielding the solution

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

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

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

variable weights.

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

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

For our result we need to truncate this estimate. We define the truncation operator T_L as

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

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). 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 $N+1$ times continously 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), (X_2,Y_2),..., (X_n,Y_n) be independent and identically distributed random variables with values in $\mathbb{R}^d \times \mathbb{R}$ such that supp(X) is bounded and*

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

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 $N \geq q$. 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, i.e., assume that each such function f satisfies

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

and let all functions g_k be Lipschitz continuous with Lipschitz constant $L > 0$ (which follows from (11) if $q > 0$). Let $\eta_n = \log(n)^{\frac{3\cdot(N+3)}{N+q+3}} \cdot n^{-\frac{2\cdot(N+1)\cdot p+2d^*}{2p+d^*}}$. Let $\mathcal{H}^{(l)}$ be defined as in (6) with K, d, d^* as in the definition of m, $M = M_n = \left\lceil n^{\frac{1}{2p+d^*}} \right\rceil$, $\alpha = \log(n) \cdot \frac{M_n^{d^*+p\cdot(2N+3)+1}}{\eta_n}$, and using a 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).

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 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 $n_{train} = \lceil \frac{4}{5} \cdot n \rceil$ realizations are used to train the estimate several times with different choices for the parameters each time, whereas $n_{test} = n - n_{train}$ 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 (abbr. *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 $\{1, 2, 3\} \cup \left\{4, 8, 12, 16, \ldots, 4 \cdot \left\lfloor \frac{n_{train}}{4} \right\rfloor\right\}$ in our simulations.

The second competitive approach we examine is interpolation with radial basis functions (abbr. *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 chosen adaptively in our implementation, because doing so improved the RBF approach in the simulations.

The parameters l, K, d^*, N and M_n of our neural network estimate (abbr. *neural-x*) defined in Theorem 1 are selected in a data-dependent way as well. The selected values of these parameters to be tested include values up to 2 for l, up to 5 for K, up to d for d^*, and up to 50 for the outer summation bound in the definition of $\mathcal{F}_{M_n,N,d^*,d,\alpha}^{(neural\ networks)}$ (where N and M_n are involved), 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 (abbr. *neural-1*), three hidden layers that consist of 3, 6, 9, 12 or 15 neurons (abbr. *neural-3*), and six hidden layers that consist of 2, 4, 6, 8 or 10 neurons (abbr. *neural-6*).

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

$$
\begin{aligned}
m_1(x) &= \cot\left(\frac{\pi}{1 + \exp\left(x_1^2 + 2 \cdot x_2 + \sin\left(6 \cdot x_4^3\right) - 3\right)}\right) \\
&\quad + \exp\left(3 \cdot x_3 + 2 \cdot x_4 - 5 \cdot x_5 + \sqrt{x_6 + 0.9 \cdot x_7 + 0.1}\right) && \left(x \in \mathbb{R}^7\right), \\
m_2(x) &= \frac{2}{x_1 + 0.008} + 3 \cdot \log(x_2^7 \cdot x_3 + 0.1) \cdot x_4 && \left(x \in \mathbb{R}^7\right), \\
m_3(x) &= 2 \cdot \log(x_1 \cdot x_2 + 4 \cdot x_3 + |\tan(x_4)|) + x_3^4 \cdot x_5^2 \cdot x_6 - x_4 \cdot x_7 \\
&\quad + (3 \cdot x_8^2 + x_9 + 2)^{0.1 + 4 \cdot x_{10}^2} && \left(x \in \mathbb{R}^{10}\right), \\
m_4(x) &= x_1 + \tan(x_2) + x_3^3 + \log(x_4) + 3 \cdot x_5 + x_6 + \sqrt{x_7} && \left(x \in \mathbb{R}^7\right), \\
m_5(x) &= \exp(\|x\|) && \left(x \in \mathbb{R}^7\right).
\end{aligned}
$$

The examples m_1, m_2, and m_3 represent some ordinary general hierarchical interaction models (cf., Definition 2), whereas m_4 and m_5 carry the definition to the extremes, such that m_4 is just an additive model, i.e. $d^* = 1$, and m_5 is an interaction model with $d^* = d$. The n observations (for $n \in \{100, 200\}$) of the type (X, Y), which are available

for all estimates, are generated by

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

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 $\lambda_1 = 9.11$, $\lambda_2 = 5.68$, $\lambda_3 = 13.97$, $\lambda_4 = 1.94$, and $\lambda_5 = 1.64$ (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 risk, which is motivated by the desired properties of a regression estimate from Section 1.1 and Theorem 1. We define it as

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

where $X_1, X_2, \ldots, X_{\bar{N}}$ are independent realizations of the random variable X. 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.

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 seven to ten 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 and m_5, our approach is not always the best although it surprisingly performs well even here in some situations. For the additive model m_4, our estimate is better than the others in case of little noise and only slightly worse in case of heavy noise. However, the function m_5, which is rather densely connected

in the sense of interaction models because all components interact in only one function, is not perfectly imitated by our sparsely connected neural network estimate.

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 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.

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*

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

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

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

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 et al. (2009). A complete version of the proof is available from the authors on request. □

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

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

and the approximation error

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

m_1				
noise	5%		10%	
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	$n = 100$	$n = 200$	$n = 100$	$n = 200$
$\bar{\varepsilon}_{L_2,\bar{N}}(avg)$	5469.53	5423.18	5469.45	5422.36
approach	median (IQR)	median (IQR)	median (IQR)	median (IQR)
neural-1	0.6651 (0.6241)	0.4396 (0.6350)	0.7203 (0.7029)	0.3913 (0.9014)
neural-3	0.9326 (0.6135)	0.5123 (0.9669)	0.8983 (0.6746)	0.4068 (0.8176)
neural-6	1.0526 (0.6522)	0.8187 (0.7144)	1.0208 (0.5141)	0.7446 (0.7475)
neural-x	**0.4673 (1.0027)**	**0.0403 (0.0731)**	**0.2004 (0.7023)**	**0.0589 (0.5533)**
RBF	0.8177 (0.3828)	0.6546 (0.4650)	0.8872 (0.3648)	0.6713 (0.4398)
neighbor	0.8794 (0.0728)	0.8294 (0.0993)	0.8679 (0.1094)	0.8118 (0.1043)

Table 1: Median and interquartile range of the scaled empirical L_2 risk of estimates for m_1, m_2, and m_3

m_4				
noise	5%		20%	
sample size	$n = 100$	$n = 200$	$n = 100$	$n = 200$
$\bar{\varepsilon}_{L_2,\bar{N}}(avg)$	2.25	2.24	2.25	2.24
approach	median (IQR)	median (IQR)	median (IQR)	median (IQR)
neural-1	0.7969 (26.5904)	0.1653 (1.4011)	13.6267 (462.1001)	6.1392 (436.0140)
neural-3	0.2158 (0.4857)	0.1247 (0.2770)	1.0037 (2.7410)	0.2354 (0.5351)
neural-6	0.1942 (0.2214)	0.0772 (0.1128)	0.3004 (0.6797)	**0.1338 (0.2052)**
neural-x	**0.0837 (0.3036)**	**0.0313 (0.0764)**	0.3161 (0.9427)	0.2422 (0.5064)
RBF	0.1029 (0.0433)	0.0812 (0.0215)	**0.2207 (0.0583)**	0.2006 (0.0473)
neighbor	0.3820 (0.0692)	0.3072 (0.0395)	0.3757 (0.0818)	0.3092 (0.0565)
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)

Table 2: Median and interquartile range of the scaled empirical L_2 risk of estimates for m_4 and m_5

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. *Assume that the assumptions of Theorem 1 hold. Let* $\varepsilon_n \geq \frac{1}{n^{c_8}}$ *and let* $\frac{M_n}{\eta_n} \leq n^{c_9}$ *for large* n*. Then*

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

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 is available from the authors on request. □

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

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

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 prove the following result concerning the approximation of (p,C)–smooth function by multilayer feedforward neural networks with two hidden layers.

Theorem 2. *Let $a \geq 1$ 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*

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

Let ν be an arbitrary probability measure on $\mathbb{R}^d$. Let $N \in \mathbb{N}_0$ be chosen such that $N \geq q$ 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 $a \leq M$ must hold), a neural network of the type

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

exists such that

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

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

$$\begin{aligned} |\mu_i| &\leq c_{14}\cdot a^q \cdot M^{N\cdot p} \\ |\lambda_{i,l}| &\leq M^{d+p\cdot(N+2)} \\ |\theta_{i,l,m}| &\leq 6\cdot d\cdot \frac{1}{\eta}\cdot M^{d+p\cdot(2N+3)+1} \end{aligned}$$

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

In the proof of Theorem 2, we will need several auxiliary results, which we formulate and prove next.

For our first auxiliary result we need to introduce the following notations: Let $N \in \mathbb{N}$ and $d \in \mathbb{N}$. For

$$\alpha = \left(\alpha^{(0)}, \ldots, \alpha^{(d)}\right) \in \mathbb{R}^{d+1}$$

set

$$f_\alpha(x) = \left(\alpha^{(0)} + \sum_{k=1}^{d} \alpha^{(k)} \cdot x^{(k)}\right)^N \qquad (x \in \mathbb{R}^d).$$

Obviously we have

$$f_\alpha(x) = \sum_{\substack{r_0,\ldots,r_d \in \mathbb{N}_0, \\ r_0+\cdots+r_d=N}} \binom{N}{r_0, \ldots, r_d} \cdot \prod_{k=0}^{d} \left(\alpha^{(k)}\right)^{r_k} \cdot \prod_{k=1}^{d} \left(x^{(k)}\right)^{r_k}. \tag{16}$$

Let $\mathcal{P}_N$ be the linear span of all monomials of the form

$$\prod_{k=1}^{d} \left(x^{(k)}\right)^{r_k} \tag{17}$$

for some $r_1, \ldots, r_d \in \mathbb{N}_0$, $r_1 + \cdots + r_d \leq N$. Then, $\mathcal{P}_N$ is a linear vector space of functions of dimension

$$dim\, \mathcal{P}_N = \left|\left\{(r_0, \ldots, r_d) \in \mathbb{N}_0^{d+1} \,:\, r_0 + \cdots + r_d = N\right\}\right| = \binom{d+N}{d},$$

and we have $f_\alpha \in \mathcal{P}_N$ for all $\alpha \in \mathbb{R}^{d+1}$.

Lemma 3. *Set $K = dim\, \mathcal{P}_N$. For almost all $\alpha_1, \ldots, \alpha_K \in \mathbb{R}^{d+1}$ (with respect to the Lebesgue measure in $\mathbb{R}^{(d+1)\cdot K}$) we have that f_{α_1}, ..., f_{α_K} is a basis of the linear vector space $\mathcal{P}_N$.*

Proof. It suffices to show that f_{α_1}, ..., f_{α_K} are linearly independent. To do this, let $\beta_1, \ldots, \beta_K \in \mathbb{R}$ be such that

$$\sum_{k=1}^{K} \beta_k \cdot f_{\alpha_k} = 0. \tag{18}$$

The monomials (17) are linearly independent. Thus, equation (18) implies

$$\sum_{k=1}^{K} \beta_k \cdot \prod_{j=0}^{d} \left(\alpha_k^{(j)}\right)^{r_j} = 0 \quad \text{for all } \mathbf{r} = (r_0, \ldots, r_d) \in R, \tag{19}$$

where

$$R = \left\{(r_0, \ldots, r_d) \in \mathbb{N}_0^{d+1} \quad : \quad r_0 + \cdots + r_d = N\right\}.$$

It suffices to show that the matrix

$$A = \left(\prod_{j=0}^{d} \left(\alpha_k^{(j)}\right)^{r_j}\right)_{\mathbf{r}\in R, k\in\{1,\ldots,K\}}$$

is regular, which is equivalent to the assertion that the matrix

$$A^T = \left(\prod_{j=0}^{d} \left(\alpha_k^{(j)} \right)^{r_j} \right)_{k \in \{1,\ldots,K\}, \mathbf{r} \in R}$$

is regular. To prove this, it suffices to show that for arbitrary $\gamma_{\mathbf{r}} \in \mathbb{R}$ ($\mathbf{r} \in R$), we have that

$$\sum_{\mathbf{r} \in R} \gamma_{\mathbf{r}} \cdot \prod_{j=0}^{d} \left(\alpha_k^{(j)} \right)^{r_j} = 0 \quad \text{for } k = 1, \ldots, K \tag{20}$$

implies

$$\gamma_{\mathbf{r}} = 0 \quad \text{for all } \mathbf{r} \in R. \tag{21}$$

Thus, let $\gamma_{\mathbf{r}} \in \mathbb{R}$ ($\mathbf{r} \in R$) be arbitrary and assume that (20) holds. Then the polynomial

$$p(x) = \sum_{\mathbf{r} \in R} \gamma_{\mathbf{r}} \cdot \prod_{k=1}^{d} \left(x^{(k)} \right)^{r_k},$$

which is contained in $\mathcal{P}_N$, satisfies

$$p(\alpha_k) = 0 \quad \text{for } k = 1, \ldots, K. \tag{22}$$

Proposition 4 in Sauer (2006) implies that the condition (22) has the only solution $p = 0$ in $\mathcal{P}_N$ for Lebesgue almost all $\alpha_1, \ldots \alpha_K \in \mathbb{R}^{d+1}$, which in turn implies (21). The proof is complete. □

Lemma 4. *Let $\sigma : \mathbb{R} \to [0,1]$ satisfy the properties (i) and (ii) of Definition 3. Then for any $R > 0$ coefficients $\gamma_1, \ldots, \gamma_{N+1} \in \mathbb{R}$ and $\beta_1, \ldots, \beta_{N+1}$ with*

$$|\gamma_k| \leq c_{15} \cdot R^N \quad \textit{and} \quad |\beta_k| \leq \frac{N}{R} \tag{23}$$

for all $1 \leq k \leq N+1$ exist, such that for all $x \in [-a, a]$

$$\left| \sum_{k=1}^{N+1} \gamma_k \cdot \sigma\left(\beta_k \cdot x + t_\sigma \right) - x^N \right| \leq c_{16} \cdot \frac{a^{N+1}}{R}$$

holds, where c_{15} and c_{16} depend on N but not on a and R.

Proof. The rather technical proof of this lemma is mainly based on a Taylor series expansion and follows from the proof of Theorem 2 in Scarselli and Tsoi (1998). A complete version is available from the authors on request. □

Lemma 5. *Let $p \in \mathcal{P}_N$ for $N \in \mathbb{N}_0$. Let $m_1, \ldots, m_{\binom{d+N}{d}}$ denote all monomials in $\mathcal{P}_N$. Define $r_i \in \mathbb{R}$ $(i = 1, \ldots, \binom{d+N}{d})$ by*

$$p(x) = \sum_{i=1}^{\binom{d+N}{d}} r_i \cdot m_i(x) \qquad (x \in \mathbb{R}^d), \tag{24}$$

and set $\bar{r}(p) = \max_{i=1,\ldots,\binom{d+N}{d}} |r_i|$. *Let* $\sigma : \mathbb{R} \to [0,1]$ *be* N*-admissible according to Definition 3. Then for any* $R, a > 0$ *a neural network of the type*

$$s(x) = \sum_{j=1}^{\binom{d+N}{d}} d_j \sum_{k=1}^{N+1} b_k \cdot \sigma \left(\sum_{m=1}^{d} a_{j,k,m} \cdot x^{(m)} + a_{j,k,0} \right)$$

exists such that

$$|s(x) - p(x)| \leq c_{17} \cdot \bar{r}(p) \cdot \frac{a^{N+1}}{R}$$

holds for all $x \in [-a,a]^d$*, and the coefficients of this neural network satisfy*

$$\begin{aligned} &|d_j| \leq c_{18} \cdot \bar{r}(p), \\ &|b_k| \leq c_{15} \cdot R^N, \\ &|a_{j,k,m}| \leq \frac{N \cdot \max\{1,a\}}{R \cdot (d+1)} + |t_\sigma| \end{aligned}$$

for $j \in \left\{1, \ldots, \binom{d+N}{d}\right\}$, $k \in \{1, \ldots, N+1\}$, *and* $m \in \{0, \ldots, d\}$, *where* c_{18} *and* c_{15} *depend on* N *and* d *but not on* a, R *and* p.

Proof. It follows from Lemma 3 that we can reconstruct all of the m_i by a sum of the form

$$m_i(x) = \sum_{j=1}^{\binom{d+N}{d}} \bar{d}_{i,j} \cdot \left(\sum_{m=1}^{d} \alpha_j^{(m)} \cdot x^{(m)} + \alpha_j^{(0)} \right)^N, \tag{25}$$

where we can choose $\alpha_j^{(m)} \in \left[-\frac{1}{d+1}, \frac{1}{d+1}\right]$ and $\alpha_j^{(0)} \in \left[-\frac{a}{d+1}, \frac{a}{d+1}\right]$ for $j \in \left\{1, \ldots, \binom{d+N}{d}\right\}$ and $m \in \{1, \ldots, d\}$, because these alternatives form a set with positive Lebesgue measure. Then $\sum_{m=1}^{d} \alpha_j^{(m)} \cdot x^{(m)} + \alpha_j^{(0)} \in [-a,a]$ holds for all $j \in \left\{1, \ldots, \binom{d+N}{d}\right\}$. After selecting a possible assignment of this type for all these inner coefficients, we can bound the maximum absolute value of the coefficients $\bar{d}_{i,j}$ by a constant

$$\max_{i,j \in \left\{1,\ldots,\binom{d+N}{d}\right\}} \bar{d}_{i,j} = c_{19}. \tag{26}$$

If we replace the subfunction $g(z) = z^N$ in (25) by its neural network approximation from Lemma 4, we obtain

$$\begin{aligned} &\left| m_i(x) - \sum_{j=1}^{\binom{d+N}{d}} \bar{d}_{i,j} \cdot \left(\sum_{k=1}^{N+1} \gamma_k \cdot \sigma \left(\beta_k \left(\sum_{m=1}^{d} \alpha_j^{(m)} \cdot x^{(m)} + \alpha_j^{(0)} \right) + t_\sigma \right) \right) \right| \\ &\leq \sum_{j=1}^{\binom{d+N}{d}} |\bar{d}_{i,j}| \cdot \left| \left(\sum_{m=1}^{d} \alpha_j^{(m)} \cdot x^{(m)} + \alpha_j^{(0)} \right)^N \right. \end{aligned}$$

$$- \left(\sum_{k=1}^{N+1} \gamma_k \cdot \sigma \left(\sum_{m=1}^{d} \beta_k \cdot \alpha_j^{(m)} \cdot x^{(m)} + \beta_k \cdot \alpha_j^{(0)} + t_\sigma \right) \right) \Bigg|$$
$$\leq \binom{d+N}{d} \cdot c_{19} \cdot c_{16} \cdot \frac{a^{N+1}}{R}.$$

By using representation (24), we conclude

$$\begin{aligned}
&\left| p(x) - \sum_{i=1}^{\binom{d+N}{d}} r_i \sum_{j=1}^{\binom{d+N}{d}} \bar{d}_{i,j} \cdot \left(\sum_{k=1}^{N+1} \gamma_k \cdot \sigma \left(\sum_{m=1}^{d} \beta_k \cdot \alpha_j^{(m)} \cdot x^{(m)} + \beta_k \cdot \alpha_j^{(0)} + t_\sigma \right) \right) \right| \\
&\leq \sum_{i=1}^{\binom{d+N}{d}} |r_i| \cdot \left| m_i(x) - \sum_{j=1}^{\binom{d+N}{d}} \bar{d}_{i,j} \cdot \left(\sum_{k=1}^{N+1} \gamma_k \cdot \sigma \left(\sum_{m=1}^{d} \beta_k \cdot \alpha_j^{(m)} \cdot x^{(m)} + \beta_k \cdot \alpha_j^{(0)} + t_\sigma \right) \right) \right| \\
&\leq \binom{d+N}{d} \cdot \bar{r}(p) \cdot \binom{d+N}{d} \cdot c_{19} \cdot c_{16} \cdot \frac{a^{N+1}}{R}.
\end{aligned}$$

Thus, $d_j = \sum_{i=1}^{\binom{d+N}{d}} \bar{d}_{i,j} \cdot r_i$, $b_k = \gamma_k$, $a_{j,k,m} = \beta_k \cdot \alpha_j^{(m)}$ and $a_{j,k,0} = \beta_k \cdot \alpha_j^{(0)} + t_\sigma$ satisfy the assertion of the lemma, because they are bounded in the required way due to (26), (23), and the choice of the coefficients subsequent to (25). □

Remark 2. We notice that we can rewrite $s(x)$ in Lemma 5 as

$$s(x) = \sum_{l=1}^{\binom{d+N}{d} \cdot (N+1)} \tilde{b}_l \cdot \sigma \left(\sum_{m=1}^{d} \tilde{a}_{l,m} \cdot x^{(m)} + \tilde{a}_{l,0} \right),$$

if we define $\tilde{b}_{(j-1)\cdot(N+1)+k} = d_j \cdot b_k$ and $\tilde{a}_{(j-1)\cdot(N+1)+k,m} = a_{j,k,m}$ for $j = 1, \ldots, \binom{d+N}{d}$ and $k = 1, \ldots, N+1$. This allows us to bound these coefficients by

$$\begin{aligned}
&|\tilde{b}_l| \leq c_{20} \cdot \bar{r}(p) \cdot R^N, \\
&|\tilde{a}_{l,m}| \leq \frac{N \cdot \max\{1, a\}}{R \cdot (d+1)} + |t_\sigma| = c_{21} \cdot \frac{\max\{1, a\}}{R} + |t_\sigma|
\end{aligned}$$

for all $l = 1, \ldots, \binom{d+N}{d} \cdot (N+1)$ and $m = 0, \ldots, d$.

Our next lemma is a modification of Proposition 3.8 in Mhaskar (1993).

Lemma 6. *Let $K \subseteq \mathbb{R}^d$ be a polytope bounded by hyperplanes $v_j \cdot x + w_j \leq 0$ ($j = 1, \ldots, H$), where $v_1, \ldots, v_H \in \mathbb{R}^d$ and $w_1, \ldots, w_H \in \mathbb{R}$. For $\delta > 0$ set*

$$K_\delta^0 := \left\{ x \in \mathbb{R}^d \,:\, v_j \cdot x + w_j \leq -\delta \quad \textit{for all } j \in \{1, \ldots, H\} \right\}$$

and

$$K_\delta^c := \left\{ x \in \mathbb{R}^d \,:\, v_j \cdot x + w_j \geq \delta \quad \textit{for some } j \in \{1, \ldots, H\} \right\}.$$

Let $\sigma : \mathbb{R} \to [0,1]$ be a sigmoidal function, which satisfies

$$|\sigma(y) - 1| \leq \frac{1}{y} \quad \textit{if} \quad y > 0 \quad \textit{and} \quad |\sigma(y)| \leq \frac{1}{|y|} \quad \textit{if} \quad y < 0. \tag{27}$$

Let $\varepsilon, \delta \in (0,1]$ be arbitrary. Then a neural network of the form

$$f(x) = \sum_{j=1}^{H} \sigma \left(\sum_{k=1}^{d} a_{j,k} \cdot x^{(k)} + a_{j,0} \right)$$

exists, satisfying

$$\begin{aligned} &|f(x)| \leq H \textit{ for } x \in \mathbb{R}^d, \\ &|f(x)| \leq H \cdot \varepsilon \textit{ for } x \in K_\delta^0, \\ &f(x) \geq 1 - \varepsilon \textit{ for } x \in K_\delta^c. \end{aligned} \tag{28}$$

The weights above can be chosen such that

$$|a_{j,k}| \leq \frac{1}{\varepsilon \cdot \delta} \cdot \max\{\|v_1\|_\infty, |w_1|, \ldots, \|v_H\|_\infty, |w_H|\} \quad \textit{for all } j = 1, \ldots, H; k = 0, \ldots, d.$$

Proof. We set

$$a_{j,k} = \frac{1}{\varepsilon \cdot \delta} \cdot v_j^{(k)} \quad \text{and} \quad a_{j,0} = \frac{1}{\varepsilon \cdot \delta} \cdot w_j \quad \text{for all } j = 1, \ldots, H; k = 1, \ldots, d.$$

So for $x \in K_\delta^0$

$$\sum_{k=1}^{d} a_{j,k} \cdot x^{(k)} + a_{j,0} \leq -\frac{1}{\varepsilon} \quad \text{for all } j = 1, \ldots, H,$$

which implies

$$\left| \sum_{j=1}^{H} \sigma \left(\sum_{k=1}^{d} a_{j,k} \cdot x^{(k)} + a_{j,0} \right) \right| \leq \sum_{j=1}^{H} \left| \sigma \left(\sum_{k=1}^{d} a_{j,k} \cdot x^{(k)} + a_{j,0} \right) \right| \leq H \cdot \varepsilon$$

due to (27). For $x \in K_\delta^c$ we know that there is a $j^* \in \{1, \ldots, H\}$, which satisfies

$$\sum_{k=1}^{d} a_{j^*,k} \cdot x^{(k)} + a_{j^*,0} \geq \frac{1}{\varepsilon}.$$

This leads to

$$\sum_{j=1}^{H} \sigma \left(\sum_{k=1}^{d} a_{j,k} \cdot x^{(k)} + a_{j,0} \right) \geq \sigma \left(\sum_{k=1}^{d} a_{j^*,k} \cdot x^{(k)} + a_{j^*,0} \right) \geq 1 - \varepsilon$$

because of (27) and $\sigma(y) \geq 0$ for all $y \in \mathbb{R}$. Furthermore, $\|\sigma\|_\infty \leq 1$ implies $|f(x)| \leq H$ and the announced bound for the coefficients follows immediately from their definition above. □

Lemma 7. *Let $K \subseteq \mathbb{R}^d$ be a polytope bounded by hyperplanes $v_j \cdot x + w_j \leq 0$ $(j = 1, \ldots, H)$, where $v_1, \ldots, v_H \in \mathbb{R}^d$ and $w_1, \ldots, w_H \in \mathbb{R}$, and let $a \geq 1$. Let $M \in \mathbb{N}$ be sufficiently large (independent of the size of a, but $a \leq M$ must hold). For $\delta > 0$ define K_δ^0 and K_δ^c as in Lemma 6. Let $p : \mathbb{R}^d \to \mathbb{R}$ be a polynomial from $\mathcal{P}_N$ with $\bar{r}(p)$ defined as in Lemma 5 and let $\sigma : \mathbb{R} \to [0,1]$ be N-admissible according to Definition 3. Then a function*

$$t(x) = \sum_{j=1}^{\binom{d+N}{d}\cdot(N+1)} \mu_j \cdot \sigma\left(\sum_{l=1}^{2d+H} \lambda_{j,l} \cdot \sigma\left(\sum_{m=1}^{d} \theta_{l,m} \cdot x^{(m)} + \theta_{l,0}\right) + \lambda_{j,0}\right)$$

exists, such that

$$\begin{aligned} |t(x) - p(x)| &\leq c_{22} \cdot \bar{r}(p) \cdot a^{N+3} \cdot M^{-p} && \text{for } x \in K_\delta^0 \cap [-a,a]^d, \\ |t(x)| &\leq c_{23} \cdot \bar{r}(p) \cdot M^{-d-2p} && \text{for } x \in K_\delta^c \cap [-a,a]^d, \\ |t(x)| &\leq c_{24} \cdot \bar{r}(p) \cdot M^{N \cdot p} && \text{for } x \in \mathbb{R}^d \end{aligned}$$

hold. Here the coefficients can be chosen such that they satisfy

$$\begin{aligned} |\mu_j| &\leq c_{20} \cdot \bar{r}(p) \cdot M^{N \cdot p}, \\ |\lambda_{j,l}| &\leq M^{d+p\cdot(N+2)}, \\ |\theta_{l,m}| &\leq \max\left\{|t_\sigma|, \frac{M^{d+p\cdot(2N+3)}}{\delta} \cdot \max\{\|v_1\|_\infty, |w_1|, \ldots, \|v_H\|_\infty, |w_H|\}\right\}, \end{aligned}$$

for $j \in \left\{1, \ldots, \binom{d+N}{d} \cdot (N+1)\right\}$, $l \in \{0, \ldots, 2d+H\}$, and $m \in \{0, \ldots, d\}$.

Proof. Let t_σ, c_{21} and c_{15} be defined as in Remark 2 and Lemma 4 and set $R = M^p$, $\tilde{R} = M^{p\cdot(N+1)}$, $B = M^{d+p\cdot(N+2)}$, $\varepsilon = M^{-d-p\cdot(2N+3)}$. For a sufficiently large $M \in \mathbb{N}$, we have

$$\left(c_{21} \cdot \frac{a}{M^p} + |t_\sigma|\right) \cdot \left(2 \cdot c_{15} \cdot M^{p\cdot(N+1)} + 1\right) \leq M^{d+p\cdot(N+2)} \cdot \left(\frac{3}{4} - M^{-d-p\cdot(2N+3)}\right),$$

because $a \leq M$. Consequently R, $\tilde{R}$, B, and ε satisfy

$$\left(c_{21} \cdot \frac{a}{R} + |t_\sigma|\right) \cdot \left(2 \cdot c_{15} \cdot \tilde{R} + 1\right) \leq B \cdot \left(\frac{3}{4} - \varepsilon\right). \tag{29}$$

Let

$$s(x) = \sum_{j=1}^{\binom{d+N}{d}\cdot(N+1)} \tilde{b}_j \cdot \sigma\left(\sum_{l=1}^{d} \tilde{a}_{j,l} \cdot x^{(l)} + \tilde{a}_{j,0}\right),$$

be chosen as the approximation in Remark 2 using the above R. At first, we replace the terms $x^{(l)}$ by their approximation from Lemma 4 using $N = 1$ and $\tilde{R}$ therein and insert an additional term of the type $f(x)$ in Lemma 6 multiplied by $-B$. This leads to

$$t(x) = \sum_{j=1}^{\binom{d+N}{d}\cdot(N+1)} \tilde{b}_j \cdot \sigma \left(\sum_{l=1}^{d} \tilde{a}_{j,l} \cdot \sum_{k=1}^{2} \gamma_k \cdot \sigma \left(\beta_k \cdot x^{(l)} + t_\sigma \right) - B \cdot \sum_{l=1}^{H} \sigma \left(\sum_{k=1}^{d} a_{l,k} \cdot x^{(k)} + a_{l,0} \right) + \tilde{a}_{j,0} \right). \tag{30}$$

Since the properties of σ entail Lipschitz continuity with a Lipschitz constant $L > 0$, Lemma 4, Lemma 6 respecting the above ε, and Remark 2 imply for $x \in K_\delta^0 \cap [-a,a]^d$

$$\begin{aligned}
|t(x) - p(x)| &\leq \left| t(x) - \sum_{j=1}^{\binom{d+N}{d}\cdot(N+1)} \tilde{b}_j \cdot \sigma \left(\sum_{l=1}^{d} \tilde{a}_{j,l} \cdot \sum_{k=1}^{2} \gamma_k \cdot \sigma \left(\beta_k \cdot x^{(l)} + t_\sigma \right) + \tilde{a}_{j,0} \right) \right| \\
&\quad + \left| \sum_{j=1}^{\binom{d+N}{d}\cdot(N+1)} \tilde{b}_j \cdot \sigma \left(\sum_{l=1}^{d} \tilde{a}_{j,l} \cdot \sum_{k=1}^{2} \gamma_k \cdot \sigma \left(\beta_k \cdot x^{(l)} + t_\sigma \right) + \tilde{a}_{j,0} \right) - s(x) \right| \\
&\quad + |s(x) - p(x)| \\
&\leq \sum_{j=1}^{\binom{d+N}{d}\cdot(N+1)} \left| \tilde{b}_j \right| \cdot L \cdot B \cdot \left| \sum_{l=1}^{H} \sigma \left(\sum_{k=1}^{d} a_{l,k} \cdot x^{(k)} + a_{l,0} \right) \right| \\
&\quad + \sum_{j=1}^{\binom{d+N}{d}\cdot(N+1)} \left| \tilde{b}_j \right| \cdot L \cdot \sum_{l=1}^{d} \left| \tilde{a}_{j,l} \right| \cdot \left| \sum_{k=1}^{2} \gamma_k \cdot \sigma \left(\beta_k \cdot x^{(l)} + t_\sigma \right) - x^{(l)} \right| \\
&\quad + |s(x) - p(x)| \\
&\leq \binom{d+N}{d} \cdot (N+1) \cdot c_{20} \cdot \bar{r}(p) \cdot R^N \cdot L \cdot B \cdot H \cdot \varepsilon \\
&\quad + \binom{d+N}{d} \cdot (N+1) \cdot c_{20} \cdot \bar{r}(p) \cdot R^N \cdot L \cdot d \cdot \left(c_{21} \cdot \frac{a}{R} + |t_\sigma| \right) \cdot c_{16} \cdot \frac{a^2}{\tilde{R}} \\
&\quad + c_{17} \cdot \bar{r}(p) \cdot \frac{a^{N+1}}{R} \\
&\leq c_{25} \cdot \bar{r}(p) \cdot \left(R^N \cdot B \cdot \varepsilon + \left(\frac{a}{R} + |t_\sigma| \right) \cdot R^N \cdot \frac{a^2}{\tilde{R}} + \frac{a^{N+1}}{R} \right) \\
&\leq c_{22} \cdot \bar{r}(p_{\mathbf{i}}) \cdot a^{N+3} \cdot M^{-p}.
\end{aligned}$$

For $x \in K_\delta^c \cap [-a,a]^d$ we know for the same reason and from the monotonicity of σ that

$$|t(x)| \leq \sum_{j=1}^{\binom{d+N}{d}\cdot(N+1)} \left| \tilde{b}_j \right| \cdot \sigma \left(\sum_{l=1}^{d} \tilde{a}_{j,l} \cdot \sum_{k=1}^{2} \gamma_k \cdot \sigma \left(\beta_k \cdot x^{(l)} + t_\sigma \right) \right.$$

$$
\begin{aligned}
&\qquad\qquad -B \cdot \sum_{l=1}^{H} \sigma\left(\sum_{k=1}^{d} a_{l,k} \cdot x^{(k)} + a_{l,0}\right) + \tilde{a}_{j,0}\Bigg) \\
&\leq \sum_{j=1}^{\binom{d+N}{d}\cdot(N+1)} \left|\tilde{b}_j\right| \cdot \sigma\left(\left(c_{21} \cdot \frac{a}{R} + |t_\sigma|\right) \cdot 2 \cdot c_{15} \cdot \tilde{R} \cdot 1\right. \\
&\qquad\qquad \left. -B \cdot (1-\varepsilon) + c_{21} \cdot \frac{a}{R} + |t_\sigma|\right) \\
&\leq \sum_{j=1}^{\binom{d+N}{d}\cdot(N+1)} \left|\tilde{b}_j\right| \cdot \sigma\left(-\frac{B}{4}\right) \\
&\leq \binom{d+N}{d} \cdot (N+1) \cdot c_{20} \cdot \bar{r}(p) \cdot R^N \cdot \frac{4}{B} \\
&\leq c_{23} \cdot \bar{r}(p) \cdot \frac{R^N}{B} \leq c_{23} \cdot \bar{r}(p) \cdot M^{-d-2p},
\end{aligned}
$$

where (29) and property (iii) in Definition 3 were used in the third and fourth inequality, respectively. Moreover, the property $\|\sigma\|_\infty \leq 1$ implies

$$
\begin{aligned}
|t(x)| &\leq \sum_{j=1}^{\binom{d+N}{d}\cdot(N+1)} \left|\tilde{b}_j\right| \cdot 1 \leq \binom{d+N}{d} \cdot (N+1) \cdot c_{20} \cdot \bar{r}(p) \cdot R^N \\
&\leq c_{24} \cdot \bar{r}(p) \cdot R^N \leq c_{24} \cdot \bar{r}(p) \cdot M^{N \cdot p}
\end{aligned}
$$

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

Next we observe that we can condense the representation of $t(x)$ in (30) into

$$
\begin{aligned}
t(x) &= \sum_{j=1}^{\binom{d+N}{d}\cdot(N+1)} \mu_j \cdot \sigma\left(\sum_{l=1}^{2d} \lambda_{j,l} \cdot \sigma\left(\sum_{m=1}^{d} \theta_{l,m} \cdot x^{(m)} + \theta_{l,0}\right)\right. \\
&\qquad\qquad \left. + \sum_{l=2d+1}^{2d+H} \lambda_{j,l} \cdot \sigma\left(\sum_{m=1}^{d} \theta_{l,m} \cdot x^{(m)} + \theta_{l,0}\right) + \lambda_{j,0}\right) \\
&= \sum_{j=1}^{\binom{d+N}{d}\cdot(N+1)} \mu_j \cdot \sigma\left(\sum_{l=1}^{2d+H} \lambda_{j,l} \cdot \sigma\left(\sum_{m=1}^{d} \theta_{l,m} \cdot x^{(m)} + \theta_{l,0}\right) + \lambda_{j,0}\right),
\end{aligned}
$$

if we choose

$$
\begin{aligned}
\mu_j &= \tilde{b}_j \\
\lambda_{j,l} &= \begin{cases} \tilde{a}_{j,0} & \text{if } l = 0 \\ \tilde{a}_{j,\lceil \frac{l}{2} \rceil} \cdot \gamma_{2-l+2\cdot\lfloor \frac{l}{2} \rfloor} & \text{if } l \in \{1, \ldots, 2d\} \\ -B & \text{if } l \in \{2d+1, \ldots, 2d+H\} \end{cases}
\end{aligned}
$$

$$\theta_{l,m} = \begin{cases} t_\sigma & \text{if } l \in \{1,\dots,2d\}, m = 0 \\ \beta_{2-l+2\cdot\lfloor \frac{l}{2} \rfloor} \cdot 1_{\{\lceil \frac{l}{2} \rceil = m\}} & \text{if } l \in \{1,\dots,2d\}, m \in \{1,\dots,d\} \\ a_{l-2d,m} & \text{if } l \in \{2d+1,\dots,2d+H\} \end{cases}.$$

For sufficiently large M, this leads to

$$\begin{aligned} |\mu_j| &\leq c_{20} \cdot \bar{r}(p) \cdot R^N = c_{20} \cdot \bar{r}(p) \cdot M^{N \cdot p}, \\ |\lambda_{j,l}| &\leq \max\left\{ c_{21} \cdot \frac{a}{R} + |t_\sigma|, \left(c_{21} \cdot \frac{a}{R} + |t_\sigma| \right) \cdot c_{15} \cdot \tilde{R}, B \right\} = M^{d+p\cdot(N+2)}, \\ |\theta_{l,m}| &\leq \max\left\{ |t_\sigma|, \frac{1}{\tilde{R}}, \frac{1}{\varepsilon \cdot \delta} \cdot \max\{\|v_1\|_\infty, |w_1|, \dots, \|v_H\|_\infty, |w_H|\} \right\} \\ &= \max\left\{ |t_\sigma|, \frac{M^{d+p\cdot(2N+3)}}{\delta} \cdot \max\{\|v_1\|_\infty, |w_1|, \dots, \|v_H\|_\infty, |w_H|\} \right\}, \end{aligned}$$

which completes the proof. □

Lemma 8. *Let $p = q + s$ for some $q \in \mathbb{N}_0$ and $s \in (0,1]$, and let $C > 0$. Let $f : \mathbb{R}^d \to \mathbb{R}$ be a (p,C)-smooth function, let $x_0 \in \mathbb{R}^d$ and let p_q be the Taylor polynomial of total degree q around x_0, i.e.,*

$$p_q(x) = \sum_{\substack{j_1,\dots,j_d \in \{0,1,\dots,q\}, \\ j_1+\dots+j_d \leq q}} \frac{1}{j_1! \cdots j_d!} \cdot \frac{\partial^{j_1+\dots+j_d} f}{\partial^{j_1} x^{(1)} \cdots \partial^{j_d} x^{(d)}}(x_0) \cdot \left(x^{(1)} - x_0^{(1)}\right)^{j_1} \cdots \left(x^{(d)} - x_0^{(d)}\right)^{j_d}.$$

Then for any $x \in \mathbb{R}^d$

$$|f(x) - p_q(x)| \leq c_{26} \cdot C \cdot \|x - x_0\|^p$$

holds for some constant $c_{26} > 0$ depending only on q and d.

Proof. See Lemma 1 in Kohler (2014). □

Proof of Theorem 2. We subdivide $\left[-a - \frac{2a}{M}, a\right]^d$ into $(M+1)^d$ cubes of side length $2a/M$ and for comprehensibility, we number these cubes $C_{\mathbf{i}}$ by $\mathbf{i} \in \{1,\dots,M+1\}^d$, such that index $\mathbf{i} = (i_1,\dots,i_d)$ corresponds to the cube

$$\left[-a + (i_1 - 2) \cdot \frac{2a}{M}, -a + (i_1 - 1) \cdot \frac{2a}{M}\right] \times \cdots \times \left[-a + (i_d - 2) \cdot \frac{2a}{M}, -a + (i_d - 1) \cdot \frac{2a}{M}\right].$$

Moreover, we denote the corners of these cubes by $x_{\mathbf{i}}$ for $\mathbf{i} \in \{1,\dots,M+2\}^d$ in the same way, such that for all $C_{\mathbf{i}}$ the point $x_{\mathbf{i}}$ means the "bottom left" corner of this cube and the additional indices result from the right border of the whole grid. Therefore, each cube $C_{\mathbf{i}}$ can be written as a polytope defined by

$$-x^{(j)} + x_{\mathbf{i}}^{(j)} \leq 0 \quad \text{and} \quad x^{(j)} - x_{\mathbf{i}}^{(j)} - \frac{2a}{M} = x^{(j)} - x_{\mathbf{i+1}}^{(j)} \leq 0 \quad (j = 1,\dots,d), \tag{31}$$

where $\mathbf{i}+\mathbf{1}$ means that each component of $\mathbf{i}$ is increased by 1.
Let $p_{\mathbf{i}}$ denote the Taylor polynomial of m with order q around the center of $C_{\mathbf{i}}$. For each $\mathbf{i} \in \{1,\dots,M+1\}^d$, we treat $C_{\mathbf{i}}$ as K in Lemma 7. This implies $H = 2d$ therein and we choose $N \in \mathbb{N}_0$ with $N \geq q$ and $\delta = a \cdot \eta/(2 \cdot d \cdot M)$. Lemma 7 says that for a sufficiently large M neural networks $t_{\mathbf{i}}(x)$ of the type

$$t_{\mathbf{i}}(x) = \sum_{j=1}^{\binom{d+N}{d}\cdot(N+1)} (\mu_j)_{\mathbf{i}} \cdot \sigma\left(\sum_{l=1}^{4d} (\lambda_{j,l})_{\mathbf{i}} \cdot \sigma\left(\sum_{m=1}^{d} (\theta_{l,m})_{\mathbf{i}} \cdot x^{(m)} + (\theta_{l,0})_{\mathbf{i}}\right) + (\lambda_{j,0})_{\mathbf{i}}\right)$$

exist, with coefficients bounded as therein, such that

$$\begin{aligned} |t_{\mathbf{i}}(x) - p_{\mathbf{i}}(x)| &\leq c_{22} \cdot \bar{r}(p_{\mathbf{i}}) \cdot a^{N+3} \cdot M^{-p} && \text{for } x \in (C_{\mathbf{i}})_\delta^0 \cap [-a,a]^d, \\ |t_{\mathbf{i}}(x)| &\leq c_{23} \cdot \bar{r}(p_{\mathbf{i}}) \cdot M^{-d-2p} && \text{for } x \in (C_{\mathbf{i}})_\delta^c \cap [-a,a]^d, \\ |t_{\mathbf{i}}(x)| &\leq c_{24} \cdot \bar{r}(p_{\mathbf{i}}) \cdot M^{N \cdot p} && \text{for } x \in \mathbb{R}^d \end{aligned}$$

hold for the corresponding cube with index $\mathbf{i} \in \{1,\dots,M+1\}^d$. Assumption (14) and a transformation of the Taylor polynomial

$$p_{\mathbf{i}}(x) = \sum_{\substack{j_1,\dots,j_d \in \{0,1,\dots,q\}, \\ j_1+\cdots+j_d \leq q}} \frac{1}{j_1! \cdots j_d!} \cdot \frac{\partial^{j_1+\cdots+j_d} m}{\partial^{j_1} x^{(1)} \cdots \partial^{j_d} x^{(d)}}(x_{\mathbf{i}}) \cdot \left(x^{(1)} - x_{\mathbf{i}}^{(1)}\right)^{j_1} \cdots \left(x^{(d)} - x_{\mathbf{i}}^{(d)}\right)^{j_d}$$

(cf., Lemma 8) into a representation with monomials allow us to bound all values $\bar{r}(p_{\mathbf{i}})$ by

$$\max_{\mathbf{i} \in \{1,\dots,M+1\}^d} \bar{r}(p_{\mathbf{i}}) \leq c_{27} \cdot a^q \tag{32}$$

for a constant $c_{27} > 0$ which depends on q but not on a. Set

$$t(x) = \sum_{\mathbf{i} \in \{1,\dots,M+1\}^d} t_{\mathbf{i}}(x).$$

Then, we obtain (from the inequalities above and Lemma 8) for $x \in (C_{\mathbf{i}})_\delta^0 \cap [-a,a]^d$

$$\begin{aligned} |t(x) - m(x)| &\leq |t_{\mathbf{i}}(x) - p_{\mathbf{i}}(x)| + |p_{\mathbf{i}}(x) - m(x)| + \left|\sum_{\mathbf{j} \in \{1,\dots,M+1\}^d \setminus \{\mathbf{i}\}} t_{\mathbf{j}}(x)\right| \\ &\leq c_{22} \cdot \bar{r}(p_{\mathbf{i}}) \cdot a^{N+3} \cdot M^{-p} + c_{26} \cdot C \cdot d^{\frac{p}{2}} \cdot \left(\frac{a}{M}\right)^p \\ &\quad + \left((M+1)^d - 1\right) \cdot c_{23} \cdot \bar{r}(p_{\mathbf{i}}) \cdot M^{-d-2p} \\ &\leq c_{13} \cdot a^{N+q+3} \cdot M^{-p}. \end{aligned} \tag{33}$$

Arguing in the same way for all $\mathbf{i} \in \{1,\dots,M+1\}^d$ we can conclude that this bound holds for all $x \in [-a,a]^d$ which are not contained in

$$\bigcup_{j=1,\dots,d} \bigcup_{\mathbf{i} \in \{1,\dots,M+2\}^d} \left\{x \in \mathbb{R}^d \quad : \quad |x^{(j)} - x_{\mathbf{i}}^{(j)}| < \delta\right\}. \tag{34}$$

By slightly shifting the whole grid of cubes along the jth component (i.e. modifying all $x_{\mathbf{i}}^{(j)}$ by the same additional summand which is less than $\frac{2a}{M}$) for fixed $j \in \{1,\dots,d\}$ we can construct

$$\left\lfloor \frac{2a/M}{2\delta} \right\rfloor = \left\lfloor \frac{2a}{M} \cdot \frac{2 \cdot d \cdot M}{2 \cdot a \cdot \eta} \right\rfloor = \left\lfloor \frac{2 \cdot d}{\eta} \right\rfloor \geq d/\eta$$

different versions of t, that still satisfy (33) for all $x \in [-a,a]^d$ up to corresponding disjoint versions of

$$\bigcup_{\mathbf{i} \in \{1,\dots,M+2\}^d} \left\{ x \in \mathbb{R}^d \quad : \quad |x^{(j)} - x_{\mathbf{i}}^{(j)}| < \delta \right\},$$

within (34), and because 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 $x_{\mathbf{i}}$ such that (34) has ν–measure less than or equal to η. This finding implies the first assertion of the theorem.
Furthermore, we can bound the coefficients of $t(x)$, if we use the bounds provided by Lemma 7 and observe that due to (31) in this case $v_j \in \{-\mathbf{e}_m, \mathbf{e}_m : m \in \{1,\dots,d\}\}$ (where $\mathbf{e}_m$ denotes the mth unit vector) and $w_j \in \left\{ x_{\mathbf{i}}^{(m)}, -x_{\mathbf{i+1}}^{(m)} : m \in \{1,\dots,d\} \right\}$ hold for each $C_{\mathbf{i}}$ and $j = 1,\dots,2d$. From (32) and the fact that M is sufficiently large, this leads to

$$\begin{aligned}
|\,(\mu_j)_{\mathbf{i}}\,| &\leq c_{20} \cdot \bar{r}(p) \cdot M^{N \cdot p} \leq c_{20} \cdot c_{27} \cdot a^q \cdot M^{N \cdot p} \\
|\,(\lambda_{j,l})_{\mathbf{i}}\,| &\leq M^{d+p \cdot (N+2)} \\
|\,(\theta_{l,m})_{\mathbf{i}}\,| &\leq \max\left\{ |t_\sigma|, \frac{M^{d+p\cdot(2N+3)}}{\delta} \cdot \max\{\|v_1\|_\infty, |w_1|, \dots, \|v_H\|_\infty, |w_H|\} \right\} \\
&\leq \max\left\{ |t_\sigma|, \frac{2 \cdot d \cdot M^{d+p\cdot(2N+3)+1}}{a \cdot \eta} \cdot \max\left\{1, a + \frac{2a}{M}\right\} \right\} \\
&\leq \max\left\{ |t_\sigma|, 2 \cdot d \cdot M^{d+p\cdot(2N+3)+1} \cdot \frac{1}{\eta} \cdot \max\left\{\frac{1}{a}, 3\right\} \right\} \\
&\leq 6 \cdot d \cdot \frac{1}{\eta} \cdot M^{d+p\cdot(2N+3)+1}
\end{aligned}$$

for all $\mathbf{i} \in \{1,\dots,M+1\}^d$, $j \in \left\{1,\dots,\binom{d+N}{d} \cdot (N+1)\right\}$, $l \in \{0,\dots,4d\}$, and $m \in \{0,\dots,d\}$, which completes the proof. □

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*

$$\max_{\substack{j_1,\dots,j_d \in \{0,1,\dots,q\},\\ j_1+\cdots+j_d \leq q}} \left\| \frac{\partial^{j_1+\cdots+j_d} f}{\partial^{j_1} x^{(1)} \dots \partial^{j_d} x^{(d)}} \right\|_\infty \leq c_{28}, \tag{35}$$

and let all functions g_k be Lipschitz continuous with Lipschitz constant $L > 0$ (which follows from (35) 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, $M = M_n$, $\alpha = \log(n) \cdot \frac{M_n^{d^*+p\cdot(2N+3)+1}}{\eta_n}$, 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

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

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

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

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

Proof. We will prove the result by induction and ignore the case $c \cdot \eta_n \geq 1$, which is trivially true. For a function $m(x) = f(a_1^T x, \dots, a_{d^*}^T x)$, which satisfies a generalized hierarchical interaction model of order d^* and level $l = 0$, let $s : \mathbb{R}^d \to \mathbb{R}^{d^*}$ be characterized by $s(x) = (a_1^T x, \dots, a_{d^*}^T x)^T$ and let $\bar{a}_{\max}$ denote $\max_{k=1,\dots,d^*} \|a_k\|_\infty$. Applying Theorem 2 (which is possible because of the assumptions of this theorem) for the probability measure $\mathbf{P}_{s(X)}$, the function $f : \mathbb{R}^{d^*} \to \mathbb{R}$ in m can be approximated by a two-layered neural network $\hat{f}$ for all $x \in [-d \cdot \bar{a}_{\max} \cdot a_n, d \cdot \bar{a}_{\max} \cdot a_n]^{d^*}$, except for a set $\tilde{D}_0$ of $\mathbf{P}_{s(X)}$–measure less than or equal to $c \cdot \eta_n > 0$, with an error of

$$\left|\hat{f}(x) - f(x)\right| \leq c_{13} \cdot (d \cdot \bar{a}_{\max} \cdot a_n)^{N+q+3} \cdot M_n^{-p} \leq c_{29} \cdot a_n^{N+q+3} \cdot M_n^{-p}.$$

If we plug $s(x)$ into that approximation and condense the inner coefficients per summand, this leads (using the notation of Theorem 2) to the approximation $t(x) = \hat{f}(s(x))$ of the form

$$t(x) = \sum_{i=1}^{\binom{d^*+N}{d^*}\cdot(N+1)\cdot(M_n+1)^{d^*}} \mu_i \cdot \sigma\left(\sum_{l=1}^{4d^*} \lambda_{i,l} \cdot \sigma\left(\sum_{k=1}^{d^*} \theta_{i,l,k} \cdot a_k^T x + \theta_{i,l,0}\right) + \lambda_{i,0}\right)$$

$$
\begin{aligned}
&= \sum_{i=1}^{\binom{d^*+N}{d^*}\cdot(N+1)\cdot(M_n+1)^{d^*}} \mu_i \cdot \sigma\left(\sum_{l=1}^{4d^*} \lambda_{i,l} \cdot \sigma\left(\sum_{m=1}^{d}\sum_{k=1}^{d^*} a_k^{(m)} \cdot \theta_{i,l,k} \cdot x^{(m)} + \theta_{i,l,0}\right) + \lambda_{i,0}\right) \\
&=: \sum_{i=1}^{\binom{d^*+N}{d^*}\cdot(N+1)\cdot(M_n+1)^{d^*}} \mu_i \cdot \sigma\left(\sum_{l=1}^{4d^*} \lambda_{i,l} \cdot \sigma\left(\sum_{m=1}^{d} \tilde{\theta}_{i,l,m} \cdot x^{(m)} + \theta_{i,l,0}\right) + \lambda_{i,0}\right),
\end{aligned}
$$

where

$$
\begin{aligned}
|\mu_i| &\leq c_{14} \cdot (d \cdot \bar{a}_{\max} \cdot a_n)^q \cdot M_n^{N \cdot p} \leq \alpha, \\
|\lambda_{i,l}| &\leq M_n^{d^*+p\cdot(N+2)} \leq \alpha, \\
|\tilde{\theta}_{i,l,m}| &\leq d^* \cdot \bar{a}_{\max} \cdot 6 \cdot d^* \cdot \frac{1}{\eta_n} \cdot M_n^{d^*+p\cdot(2N+3)+1} \leq \alpha
\end{aligned}
$$

are satisfied for a sufficiently large n, such that $t \in \mathcal{H}^{(0)}$ is valid. Since $\mathbf{P}_{s(X)}\left\{\tilde{D}_0\right\} = \mathbf{P}_X\left\{s^{-1}\left(\tilde{D}_0\right)\right\}$ and $s\left([-a_n, a_n]^d\right) \subseteq [-d \cdot \bar{a}_{\max} \cdot a_n, d \cdot \bar{a}_{\max} \cdot a_n]^{d^*}$,

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

holds for all $x \in [-a_n, a_n]^d$ outside of the set $D_0 = s^{-1}\left(\tilde{D}_0\right)$ of $\mathbf{P}_X$–measure less than or equal to $c \cdot \eta_n$, which proves the first part of the assertion for $l = 0$. Furthermore, since $\|\sigma\|_\infty \leq 1$ holds according to our assumptions, we know that

$$
\begin{aligned}
|t(x)| &\leq \binom{d^*+N}{d^*} \cdot (N+1) \cdot (M_n+1)^{d^*} \cdot \max_{i=1,\ldots,\binom{d^*+N}{d^*}\cdot(N+1)\cdot(M_n+1)^{d^*}} |\mu_i| \\
&\leq c_{31} \cdot a_n^q \cdot M_n^{d^*+N\cdot p}
\end{aligned}
$$

is valid for all $x \in \mathbb{R}^d$.
When $l > 0$, we consider the following bound of the difference between $m(x) = \sum_{k=1}^K g_k\left(f_{1,k}(x), \ldots, f_{d^*,k}(x)\right)$ and an estimate $\hat{m}(x) = \sum_{k=1}^K \hat{g}_k\left(\hat{f}_{1,k}(x), \ldots, \hat{f}_{d^*,k}(x)\right)$ at a point $x \in [-a_n, a_n]^d$:

$$
\begin{aligned}
|m(x) - \hat{m}(x)| \leq{}& \left|\sum_{k=1}^K g_k\left(f_{1,k}(x), \ldots, f_{d^*,k}(x)\right) - \sum_{k=1}^K g_k\left(\hat{f}_{1,k}(x), \ldots, \hat{f}_{d^*,k}(x)\right)\right| \\
&+ \left|\sum_{k=1}^K g_k\left(\hat{f}_{1,k}(x), \ldots, \hat{f}_{d^*,k}(x)\right) - \sum_{k=1}^K \hat{g}_k\left(\hat{f}_{1,k}(x), \ldots, \hat{f}_{d^*,k}(x)\right)\right| \\
\leq{}& \sum_{k=1}^K L \cdot \sum_{j=1}^{d^*} |f_{j,k}(x) - \hat{f}_{j,k}(x)| \\
&+ \sum_{k=1}^K \left|g_k\left(\hat{f}_{1,k}(x), \ldots, \hat{f}_{d^*,k}(x)\right) - \hat{g}_k\left(\hat{f}_{1,k}(x), \ldots, \hat{f}_{d^*,k}(x)\right)\right|.
\end{aligned}
$$

All the $f_{j,k}$ satisfy a (p, C)-smooth generalized hierarchical interaction model of order d^* and level $l-1$ and respect the requirements of this theorem. Thus, we can choose the approximations $\hat{f}_{j,k} \in \mathcal{H}^{(l-1)}$ according to the induction hypothesis with η_n replaced by $\frac{\eta_n}{2 \cdot d^* \cdot K}$. Then each of the terms $|f_{j,k}(x) - \hat{f}_{j,k}(x)|$ can be bounded by $c_{32} \cdot a_n^{N+q+3} \cdot M_n^{-p}$ for all n sufficiently large and $x \in [-a_n, a_n]^d$ outside of a set $D_{j,k}$ of $\mathbf{P}_X$–measure less than or equal to $\frac{c}{2 \cdot d^* \cdot K} \cdot \eta_n$.

Furthermore, let $\hat{f}_k : \mathbb{R}^d \to \mathbb{R}^{d^*}$ be characterized by $\hat{f}_k(x) = \left(\hat{f}_{1,k}(x), \ldots, \hat{f}_{d^*,k}(x)\right)^T$ and set $\bar{f}_{k,\max} = \max_{j=1,\ldots,d^*} \|f_{j,k}\|_\infty$ for all $k = 1, \ldots, K$. Given that $c_{32} \cdot a_n^{N+q+3} \cdot M_n^{-p} \leq c_{32}$ for all sufficiently large n because of the assumptions of Theorem 3, $\hat{f}_k(x)$ falls into

$$\hat{F}_k = \left[-\bar{f}_{k,\max} - c_{32}, \bar{f}_{k,\max} + c_{32}\right]^{d^*}$$

for all $x \in [-a_n, a_n]^d$ outside of the union of the sets $D_{j,k}$ $(j = 1, \ldots, d^*,\ k = 1, \ldots, K)$ and n sufficiently large. Applying Theorem 2 (if the condition $\bar{f}_{k,\max} + c_{32} \geq 1$ is not satisfied, modify c_{32} adequately) with $\eta = \frac{c \cdot \eta_n}{2 \cdot K}$, it is possible to choose a neural network $\hat{g}_k$ for every g_k in the second sum with a maximum approximation error of

$$c_{13} \cdot \left(\bar{f}_{k,\max} + c_{32}\right)^{N+q+3} \cdot M_n^{-p} \leq c_{33} \cdot M_n^{-p}$$

on $\hat{F}_k$ outside of a set $\tilde{D}_k$ that satisfies $\mathbf{P}_{\hat{f}_k(X)}(\tilde{D}_k) \leq \frac{\eta_n}{2 \cdot K}$. For n sufficiently large, the weights of $\hat{g}_k$ according to the notation of Theorem 2 satisfy

$$\begin{aligned}
|\mu_i| &\leq c_{14} \cdot \left(\bar{f}_{k,\max} + c_{32}\right)^q \cdot M_n^{N \cdot p} \leq \alpha, \\
|\lambda_{i,l}| &\leq M_n^{d^* + p \cdot (N+2)} \leq \alpha, \\
|\theta_{i,l,m}| &\leq 6 \cdot d \cdot \frac{1}{\eta_n} \cdot M_n^{d^* + p \cdot (2N+3)+1} \leq \alpha,
\end{aligned}$$

which implies $\hat{g}_k \in \mathcal{F}_{M_n,N,d^*,d^*,\alpha}^{(neural\ networks)}$. Since $\mathbf{P}_{\hat{f}_k(X)}\left(\tilde{D}_k\right) = \mathbf{P}_X\left(\hat{f}_k^{-1}\left(\tilde{D}_k\right)\right)$, $\hat{g}_k\left(\hat{f}_k(x)\right)$ approximates $g_k\left(\hat{f}_k(x)\right)$ with the above maximum error for all

$$x \in [-a_n, a_n]^d \setminus \bigcup_{j=1,\ldots,d^*} D_{j,k}$$

outside of a set $D_k = \hat{f}_k^{-1}\left(\tilde{D}_k\right)$ of $\mathbf{P}_X$–measure less than or equal to $\frac{c \cdot \eta_n}{2 \cdot K}$. Choosing $t(x) = \hat{m}(x) = \sum_{k=1}^K \hat{g}_k\left(\hat{f}_{1,k}(x), \ldots, \hat{f}_{d^*,k}(x)\right)$ as described, we can conclude from $\hat{g}_k \in \mathcal{F}_{M_n,N,d^*,d^*,\alpha}^{(neural\ networks)}$ and $\hat{f}_{j,k} \in \mathcal{H}^{(l-1)}$ for all $j = 1, \ldots, d^*$ and $k = 1, \ldots, K$ that $t \in \mathcal{H}^{(l)}$ is valid and that for a sufficiently large n

$$|t(x) - m(x)| \leq K \cdot L \cdot d^* \cdot c_{32} \cdot a_n^{N+q+3} \cdot M_n^{-p} + K \cdot c_{33} \cdot M_n^{-p} \leq c_{29} \cdot a_n^{N+q+3} \cdot M_n^{-p}$$

holds for all $x \in [-a_n, a_n]^d$ outside of the union of all exceptional sets so far. The $\mathbf{P}_X$-measure of this union satisfies

$$\begin{aligned}\mathbf{P}_X\left(\bigcup_{\substack{j=1,\dots,d^*,\\ k=1,\dots,K}} D_{j,k} \cup \bigcup_{k=1,\dots,K} D_k\right) &\leq \sum_{\substack{j=1,\dots,d^*,\\ k=1,\dots,K}} \mathbf{P}_X\left(D_{j,k}\right) + \sum_{k=1,\dots,K} \mathbf{P}_X\left(D_k\right) \\ &\leq \sum_{\substack{j=1,\dots,d^*,\\ k=1,\dots,K}} \frac{c \cdot \eta_n}{2 \cdot d^* \cdot K} + \sum_{k=1,\dots,K} \frac{c \cdot \eta_n}{2 \cdot K} \\ &= c \cdot \eta_n,\end{aligned}$$

which proves the first assertion of the theorem when $l > 0$. The second assertion can be shown analogously to the case of $l = 0$ by

$$\begin{aligned}|t(x)| &\leq K \cdot \binom{d^* + N}{d^*} \cdot (N+1) \cdot (M_n + 1)^{d^*} \cdot \max_{k=1,\dots,K} c_{14} \cdot \left(\bar{f}_{k,\max} + c_{32}\right)^q \cdot M_n^{N \cdot p} \\ &\leq c_{34} \cdot M_n^{d^* + N \cdot p}\end{aligned}$$

for all $x \in \mathbb{R}^d$, which is an even stronger bound than the announced. □

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 $supp(X) \subseteq [-a_n, a_n]^d$ holds, which implies $\mathcal{N}\left(\delta, \mathcal{G}, \|\cdot\|_{\infty, supp(X)}\right) \leq \mathcal{N}\left(\delta, \mathcal{G}, \|\cdot\|_{\infty, [-a_n, a_n]^d}\right)$ for an arbitrary function space $\mathcal{G}$ and $\delta > 0$. Then applying Lemma 1 leads to

$$\begin{aligned}&\mathbf{E} \int |T_{c_3 \cdot \log(n)} m_n(x) - m(x)|^2 \mathbf{P}_X(dx) \\ &\leq \frac{c_7 \cdot \log(n)^2 \cdot \log\left(\mathcal{N}\left(\frac{1}{n \cdot c_3 \cdot \log(n)}, \mathcal{H}^{(l)}, \|\cdot\|_{\infty, supp(X)}\right)\right)}{n} \\ &\qquad + 2 \cdot \inf_{h \in \mathcal{H}^{(l)}} \int |h(x) - m(x)|^2 \mathbf{P}_X(dx).\end{aligned}$$

Given that $\frac{1}{n \cdot c_3 \cdot \log(n)} \geq \frac{1}{n^{c_8}}$ and $\frac{M_n}{\eta_n} \leq n^{c_9}$ hold, Lemma 2 allows us to bound the first summand by

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

for a sufficiently large n. If we choose a $h^* \in \mathcal{H}^{(l)}$, which satisfies the approximation properties of Theorem 3 using the above a_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

$$\int |h^*(x) - m(x)|^2 \cdot 1_{D_n^C} \mathbf{P}_X(dx) + \int |h^*(x) - m(x)|^2 \cdot 1_{D_n} \mathbf{P}_X(dx)$$

$$
\begin{aligned}
&\leq \left(c_{29} \cdot a_n^{(N+q+3)} \cdot M_n^{-p}\right)^2 + \left(2 \cdot c_{30} \cdot a_n^q \cdot M_n^{d^*+N\cdot p}\right)^2 \cdot \eta_n \\
&\leq c_{36} \cdot \log(n)^3 \cdot n^{-\frac{2p}{2p+d^*}} + c_{37} \cdot \log(n)^{\frac{3q}{N+q+3}} \cdot n^{\frac{2d^*+2N\cdot p}{2p+d^*}} \cdot \log(n)^{\frac{3\cdot(N+3)}{N+q+3}} \cdot n^{-\frac{2\cdot(N+1)\cdot p+2d^*}{2p+d^*}} \\
&\leq c_{11} \cdot \log(n)^3 \cdot n^{-\frac{2p}{2p+d^*}},
\end{aligned}
$$

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

5 Acknowledgment

The authors would like to thank the German Research Foundation (DFG) for funding this project within the Collaborative Research Centre 805.

References

[1] Abramovitz, M., and Stegun, I. A. (1972). *Handbook of mathematical functions.* Dover Publications, New York, US.

[2] Anthony, M., and Bartlett, P. L. (1999). *Neural Networks and Learning: Theoretical Foundations.* Cambridge University Press, Cambridge, UK.

[3] Bauer, B., Heimrich, F., Kohler, M., and Krzyżak, A. (2017). *On estimation of surrogate models for high-dimensional computer experiments.* Manuscript submitted for publication.

[4] 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.

[5] Barron, A. R. (1991). Complexity regularization with application to artificial neural networks. In: G. Roussas (ed.), *Nonparametric Functional Estimation and Related Topics*, pp. 561-576, NATO ASI Series, Kluwer Academic Publishers, Dordrecht, Netherlands.

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

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

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

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

[10] Eldan, R., and Shamir, O. (2015). The power of depth for feedforward neural networks. *arXiv preprint.*

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

[12] Györfi, L., Kohler, M., Krzyżak, A. and Walk, H. (2002). *A Distribution-Free Theory of Nonparametric Regression.* Springer Series in Statistics, Springer-Verlag, New York.

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

[14] 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.

[15] Haykin, S. O. (2008). *Neural Networks and Learning Machines.* 3rd ed. Prentice-Hall, New York, US.

[16] Hertz, J., Krogh, A., and Palmer, R. G. (1991). *Introduction to the Theory of Neural Computation.* Addison-Wesley, Redwood City, California, US.

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

[18] Kohler, M. (2014). Optimal global rates of convergence for noiseless regression estimation problems with adaptively chosen design. *Journal of Multivariate Analysis*, **132**, pp. 197-208.

[19] Kohler, M., and Krzyżak, A. (2005). Adaptive regression estimation with multilayer feedforward neural networks. *Journal of Nonparametric Statistics*, **17**, pp. 891-913.

[20] Kohler, M., and Krzyżak, A. (2016). Nonparametric regression based on hierarchical interaction models. *IEEE Transaction on Information Theory*, **63**, pp. 1620-1630.

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

[22] Lazzaro, D., and Montefusco, L. (2002). Radial Basis Functions for the Multivariate Interpolation of Large Scattered Data Sets. *Journal of Computational and Applied Mathematics*, **140**, pp. 521-536.

[23] Lugosi, G., and Zeger, K. (1995). Nonparametric estimation via empirical risk minimization. *IEEE Transactions on Information Theory*, **41**, pp. 677-687.

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

[25] Mhaskar, H. N. (1993). Approximation properties of a multilayer feedforward artificial neural network. *Advances in Computational Mathematics*, **1**, pp. 61-80.

[26] Mhaskar, H. N., and Poggio, T. (2016). Deep vs. shallow networks: An approximation theory perspective. *Analysis and Applications*, **14**, pp. 829-848.

[27] Mielniczuk, J., and Tyrcha, J. (1993). Consistency of multilayer perceptron regression estimators. *Neural Networks*, **6**, pp. 1019-1022.

[28] Ripley, B. D. (2008). *Pattern Recognition and Neural Networks*. Cambridge University Press, Cambridge, UK.

[29] Sauer, T. (2006). Polynomial interpolation in several variables: lattices, differences, and ideals. In: Jetter, K., Buhmann, M. D., Haussmann, W., Schaback, R., and Stöckler, J. (eds.), Studies in Computational Mathematics, *Topics in Multivariate Approximation and Interpolation*, **12**, pp. 191-230.

[30] Scarselli, F., and Tsoi, A. C. (1998). Universal Approximation Using Feedforward Neural Networks: A Survey of Some Existing Methods, and Some New Results. *Neural Networks*, **11**, pp. 15-37.

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

[32] Stone, C. J. (1977). Consistent nonparametric regression. *Annals of Statistics*, **5**, pp. 595-645.

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

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

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

[36] 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.

Supplementary material for the referees

Proof of Lemma 1. In the proof we use the following error decomposition:

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

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

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

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

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

With the Cauchy-Schwarz inequality and

$$I_{\{|Y| > \beta_n\}} \leq \frac{\exp(c_1/2 \cdot |Y|^2)}{\exp(c_1/2 \cdot \beta_n^2)} \tag{36}$$

we conclude

$$\begin{aligned}
|T_{5,n}| &\leq \sqrt{\mathbf{E}\big\{|T_{\beta_n}Y - Y|^2\big\}} \cdot \sqrt{\mathbf{E}\big\{|2m_n(X) - Y - T_{\beta_n}Y|^2 \big| \mathcal{D}_n\big\}} \\
&\leq \sqrt{\mathbf{E}\big\{|Y|^2 \cdot I_{\{|Y|>\beta_n\}}\big\}} \cdot \sqrt{\mathbf{E}\big\{2 \cdot |2m_n(X) - T_{\beta_n}Y|^2 + 2 \cdot |Y|^2 \big| \mathcal{D}_n\big\}} \\
&\leq \sqrt{\mathbf{E}\left\{|Y|^2 \cdot \frac{\exp(c_1/2 \cdot |Y|^2)}{\exp(c_1/2 \cdot \beta_n^2)}\right\}} \\
&\qquad \cdot \sqrt{\mathbf{E}\big\{2 \cdot |2m_n(X) - T_{\beta_n}Y|^2 \big| \mathcal{D}_n\big\} + 2\mathbf{E}\big\{|Y|^2\big\}} \\
&\leq \sqrt{\mathbf{E}\Big\{|Y|^2 \cdot \exp(c_1/2 \cdot |Y|^2)\Big\}} \cdot \exp\left(-\frac{c_1 \cdot \beta_n^2}{4}\right) \cdot \sqrt{2(3\beta_n)^2 + 2\mathbf{E}\big\{|Y|^2\big\}}.
\end{aligned}$$

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

$$|Y|^2 \leq \frac{2}{c_1} \cdot \exp\left(\frac{c_1}{2} \cdot |Y|^2\right)$$

and hence $\sqrt{\mathbf{E}\Big\{|Y|^2 \cdot \exp(c_1/2 \cdot |Y|^2)\Big\}}$ is bounded by

$$\mathbf{E}\left(\frac{2}{c_1} \cdot \exp\left(c_1/2 \cdot |Y|^2\right) \cdot \exp(c_1/2 \cdot |Y|^2)\right) \leq \mathbf{E}\left(\frac{2}{c_1} \cdot \exp\left(c_1 \cdot |Y|^2\right)\right) \leq c_{38}$$

which is less than infinity by the assumptions of the lemma. Furthermore the third term is bounded by $\sqrt{18\beta_n^2 + c_{39}}$ because

$$\mathbf{E}(|Y|^2) \leq \mathbf{E}(1/c_1 \cdot \exp(c_1 \cdot |Y|^2) \leq c_{39} < \infty, \tag{37}$$

which follows again as above. With the setting $\beta_n = c_{26} \cdot \log(n)$ it follows for some constants $c_{40}, c_{41} > 0$ that

$$|T_{5,n}| \leq \sqrt{c_{38}} \cdot \exp\left(-c_{40} \cdot \log(n)^2\right) \cdot \sqrt{(18 \cdot c_{26} \cdot \log(n)^2 + c_{39})} \leq c_{41} \cdot \frac{\log(n)}{n}.$$

From the Cauchy-Schwarz inequality we get

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

where we can bound the second factor on the right-hand side in the above inequality in the same way we have bounded the second factor from $T_{5,n}$, because by assumption

$||m||_\infty$ is bounded and furthermore m_{β_n} is bounded by β_n. Thus we get for some constant $c_{42} > 0$

$$\sqrt{\mathbf{E}\left\{\left|m(X) + m_{\beta_n}(X) - Y - T_{\beta_n}Y\right|^2\right\}} \leq c_{42} \cdot \log(n).$$

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

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

Hence we get

$$T_{6,n} \quad \leq \quad \sqrt{4 \cdot \mathbf{E}\left\{|Y - T_{\beta_n}Y|^2\right\}} \cdot c_{42} \cdot \log(n)$$

and therefore with the calculations from $T_{5,n}$ it follows that $T_{6,n} \leq c_{43} \cdot \log(n)/n$ for some constant $c_{43} > 0$. Altogether we get

$$T_{1,n} \leq c_{44} \cdot \frac{\log(n)}{n}$$

for some constant $c_{44} > 0$.
Next we consider $T_{2,n}$ and conclude for $t > 0$

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

where $T_{\beta_n, supp(X)}\mathcal{F}_n$ is defined as $\left\{T_{\beta_n} f \cdot 1_{supp(X)} : f \in \mathcal{F}_n\right\}$. Theorem 11.4 in Györfi et al. (2002) and the relation $\mathcal{N}\left(\delta, \left\{\frac{1}{\beta_n} g : g \in \mathcal{G}\right\}, \|\cdot\|_{\infty, supp(X)}\right) \leq \mathcal{N}\left(\delta \cdot \beta_n, \mathcal{G}, \|\cdot\|_{\infty, supp(X)}\right)$ for an arbitrary function space $\mathcal{G}$ and $\delta > 0$ lead to

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

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

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

Choosing

$$\varepsilon_n = \frac{5136 \cdot \beta_n^2}{n} \cdot \log\left(14 \cdot \mathcal{N}\left(\frac{1}{n \cdot \beta_n}, \mathcal{F}_n, \|\cdot\|_{\infty, supp(X)}\right)\right)$$

(which satisfies the necessary condition $\varepsilon_n \geq \frac{80}{n}$ if the constant c_5 in the definition of β_n is not too small) minimizes the right-hand side and implies

$$\mathbf{E}(T_{2,n}) \leq \frac{c_7 \cdot \log(n)^2 \cdot \log\left(\mathcal{N}\left(\frac{1}{n \cdot \beta_n}, \mathcal{F}_n, \|\cdot\|_{\infty, supp(X)}\right)\right)}{n}.$$

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

$$\mathbf{E}(T_{3,n}) \quad \leq \quad c_{45} \cdot \frac{\log(n)}{n}$$

for some large enough constant $c_{45} > 0$ and hence we get in total

$$\mathbf{E}\left(\sum_{i=1}^{3} T_{i,n}\right) \quad \leq \quad \frac{c_{46} \cdot \log(n)^2 \cdot \log\left(\mathcal{N}\left(\frac{1}{n \cdot \beta_n}, \mathcal{F}_n, \|\cdot\|_{\infty, supp(X)}\right)\right)}{n}$$

for some sufficient large constant $c_{46} > 0$.

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

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

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

$$\begin{aligned}
\frac{1}{2} \cdot T_{7,n} &\leq \sqrt{\mathbf{E}\left(\left(|m_n(X_1) - Y_1|^2\right)^2\right)} \cdot \sqrt{\mathbf{P}(A_n)} \\
&\leq \sqrt{\mathbf{E}\left(\left(2|m_n(X_1)|^2 + 2|Y_1|^2\right)^2\right)} \cdot \sqrt{n \cdot \mathbf{P}\{|Y_1| > \beta_n\}} \\
&\leq \sqrt{\mathbf{E}\left(8|m_n(X_1)|^4 + 8|Y_1|^4\right)} \cdot \sqrt{n \cdot \frac{\mathbf{E}\left(\exp(c_1 \cdot |Y_1|^2)\right)}{\exp(c_1 \cdot \beta_n^2)}},
\end{aligned}$$

where the last inequality follows from inequality (36). With $x \leq \exp(x)$ for $x \in \mathbb{R}$ we get

$$\mathbf{E}\left(|Y|^4\right) \quad = \quad \mathbf{E}\left(|Y|^2 \cdot |Y|^2\right) \leq \mathbf{E}\left(\frac{2}{c_1} \cdot \exp\left(\frac{c_1}{2} \cdot |Y|^2\right) \cdot \frac{2}{c_1} \cdot \exp\left(\frac{c_1}{2} \cdot |Y|^2\right)\right)$$

$$= \frac{4}{c_1^2} \cdot \mathbf{E}\left(\exp\left(c_1 \cdot |Y|^2\right)\right),$$

which is less than infinity by condition (12) of the theorem. Furthermore $||m_n||_\infty$ is bounded by β_n and therefore the first factor is bounded by

$$c_{47} \cdot \beta_n^2 = c_{48} \cdot \log(n)^2$$

for some constant $c_{48} > 0$. The second factor is bounded by $1/n$, because by the assumptions of the theorem $\mathbf{E}\left(\exp\left(c_1 \cdot |Y_1|^2\right)\right)$ is bounded by some constant $c_{49} < \infty$ and hence we get

$$\sqrt{n \cdot \frac{\mathbf{E}\left(\exp(c_1 \cdot |Y_1|^2)\right)}{\exp(c_1 \cdot \beta_n^2)}} \leq \sqrt{n} \cdot \frac{\sqrt{c_{49}}}{\sqrt{\exp(c_1 \cdot \beta_n^2)}} \leq \frac{\sqrt{n} \cdot \sqrt{c_{49}}}{\exp((c_1 \cdot c_{26}^2 \cdot \log(n)^2)/2)}.$$

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

$$T_{7,n} \leq c_{50} \cdot \frac{\log(n)^2 \sqrt{n}}{n^2} \leq c_{51} \cdot \frac{\log(n)^2}{n}.$$

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

$$\begin{aligned}
T_{8,n} &\leq 2 \cdot \mathbf{E}\left(\frac{1}{n}\sum_{i=1}^n |\tilde{m}_n(X_i) - Y_i|^2 \cdot I_{A_n^c} - \frac{1}{n}\sum_{i=1}^n |m(X_i) - Y_i|^2\right) \\
&\leq 2 \cdot \mathbf{E}\left(\frac{1}{n}\sum_{i=1}^n |\tilde{m}_n(X_i) - Y_i|^2 - \frac{1}{n}\sum_{i=1}^n |m(X_i) - Y_i|^2\right) \\
&\leq 2 \cdot \mathbf{E}\left(\inf_{f \in \mathcal{F}_n} \frac{1}{n}\sum_{i=1}^n |f(X_i) - Y_i|^2 - \frac{1}{n}\sum_{i=1}^n |m(X_i) - Y_i|^2\right),
\end{aligned}$$

because $|T_\beta z - y| \leq |z - y|$ holds for $|y| \leq \beta$. Hence

$$\mathbf{E}(T_{4,n}) \leq c_{51} \cdot \frac{\log(n)^2}{n} + 2 \cdot \mathbf{E}\left(\inf_{f \in \mathcal{F}_n} \frac{1}{n}\sum_{i=1}^n |f(X_i) - Y_i|^2 - \frac{1}{n}\sum_{i=1}^n |m(X_i) - Y_i|^2\right)$$

holds. If we choose an $f^* \in \mathcal{F}_n$ which satisfies

$$\int |f^*(x) - m(x)|^2 \mathbf{P}_X(dx) \leq \inf_{f \in \mathcal{F}_n} \int |f(x) - m(x)|^2 \mathbf{P}_X(dx) + \frac{1}{n},$$

we can conclude

$$\begin{aligned}
&\mathbf{E}\left(\inf_{f \in \mathcal{F}_n} \frac{1}{n}\sum_{i=1}^n |f(X_i) - Y_i|^2 - \frac{1}{n}\sum_{i=1}^n |m(X_i) - Y_i|^2\right) \\
&\leq \mathbf{E}\left\{\frac{1}{n}\sum_{i=1}^n |f^*(X_i) - Y_i|^2 - \frac{1}{n}\sum_{i=1}^n |m(X_i) - Y_i|^2\right\}
\end{aligned}$$

$$
\begin{aligned}
&= \mathbf{E}\left\{|f^*(X) - Y|^2\right\} - \mathbf{E}\left\{|m(X) - Y|^2\right\} \\
&= \mathbf{E}\left\{|f^*(X) - m(X)|^2\right\} + \mathbf{E}\left\{|m(X) - Y|^2\right\} - \mathbf{E}\left\{|m(X) - Y|^2\right\} \\
&\leq \inf_{f \in \mathcal{F}_n} \int |f(x) - m(x)|^2 \mathbf{P}_X(dx) + \frac{1}{n}.
\end{aligned}
$$

In combination with the other considerations in the proof this implies the assertion of Lemma 1. □

In order to prove Lemma 2, we will apply the following lemma.

Lemma 9. *Let $l \in \mathbb{N}_0$ and let $\sigma_r : \mathbb{R} \to \mathbb{R}$ for $r = 1, \ldots, l+1$ be Lipschitz continuous functions with Lipschitz constant $L \geq 1$, which satisfy*

$$
|\sigma_r(x)| \leq L \cdot \max\{|x|, 1\} \quad (x \in \mathbb{R}). \tag{38}
$$

Let $K_0 = d$, $K_r \in \mathbb{N}$ for $r \in \{1, \ldots, l\}$ and $K_{l+1} = 1$. For $r \in \{1, \ldots, l+1\}$ and $i \in \{1, \ldots, K_r\}$ define recursively

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

and

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

where $c_{i,0}^{(r-1)}, \bar{c}_{i,0}^{(r-1)}, \ldots, c_{i,K_{r-1}}^{(r-1)}, \bar{c}_{i,K_{r-1}}^{(r-1)} \in \mathbb{R}$, and $f_j^{(0)}(x) = \bar{f}_j^{(0)}(x) = x^{(j)}$. Furthermore, set

$$
\overline{C} = \max_{\substack{r=0,\ldots,l, i=1,\ldots,K_{r+1}, \\ j=1,\ldots,K_r}} \max\left\{\left|c_{i,j}^{(r)}\right|, \left|\bar{c}_{i,j}^{(r)}\right|, 1\right\}.
$$

Then

$$
\begin{aligned}
&\left|f_1^{(l+1)}(x) - \bar{f}_1^{(l+1)}(x)\right| \\
&\leq (l+1) \cdot L^{l+1} \cdot \prod_{r=0}^{l} (K_r + 1) \cdot \overline{C}^l \cdot \max\{\|x\|_\infty, 1\} \cdot \max_{\substack{r=0,\ldots,l, i=1,\ldots,K_{r+1}, \\ j=0,\ldots,K_r}} \left|c_{i,j}^{(r)} - \bar{c}_{i,j}^{(r)}\right|
\end{aligned}
$$

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

Proof. See Lemma 5 in Bauer et al. (2017). □

Proof of Lemma 2. At first, we notice the space $\mathcal{H}^{(l)}$ (with $l > 0$) can be expressed as

$$
\mathcal{H}^{(l)} = \left\{ h : \mathbb{R}^d \to \mathbb{R} \ : \ h(x) = \sum_{k=1}^{K} \sigma_{id}\left(g_k\left(\sigma_{id}\left(f_{1,k}(x)\right), \ldots, \sigma_{id}\left(f_{d^*,k}(x)\right)\right)\right) \quad (x \in \mathbb{R}^d) \right.
$$

$$\text{for some } g_k \in \mathcal{F}^{(neural\ networks)}_{M_n,N,d^*,d^*,\alpha,\beta,\gamma} \text{ and } f_{j,k} \in \mathcal{H}^{(l-1)} \Bigg\},$$

where $\sigma_{id} : \mathbb{R} \to \mathbb{R}$ is the identity $\sigma_{id}(x) = x$ for all $x \in \mathbb{R}$. Furthermore, all $g \in \mathcal{F}^{(neural\ networks)}_{M_n,N,d^*,d^*,\alpha}$ can be written as

$$\begin{aligned} g(x) &= \sum_{i=1}^{\binom{d^*+N}{d^*}\cdot(N+1)\cdot(M+1)^{d^*}} \mu_i \cdot \sigma\left(\sum_{l=1}^{4d^*} \lambda_{i,l} \cdot \sigma\left(\sum_{m=1}^{d^*} \theta_{i,l,m} \cdot x^{(m)} + \theta_{i,l,0}\right) + \lambda_{i,0}\right) \\ &= \sum_{i=1}^{\binom{d^*+N}{d^*}\cdot(N+1)\cdot(M+1)^{d^*}} \mu_i \cdot \sigma\Bigg(\sum_{\substack{l=1,\dots,4d^*,\\ \bar{i}=1,\dots,\binom{d^*+N}{d^*}\cdot(N+1)\cdot(M+1)^{d^*}}} \lambda_{i,\bar{i},l} \\ &\qquad \cdot\sigma\left(\sum_{m=1}^{d^*} \theta_{\bar{i},l,m} \cdot x^{(m)} + \theta_{\bar{i},l,0}\right) + \lambda_{i,\bar{i},0}\Bigg) \end{aligned}$$

where the new coefficients are defined by

$$\lambda_{i,\bar{i},l} := \begin{cases} \lambda_{i,l} & \text{if } \bar{i} = i \\ 0 & \text{otherwise} \end{cases}$$

for all $i, \bar{i} \in \left\{1, \dots, \binom{d^*+N}{d^*} \cdot (N+1) \cdot (M+1)^{d^*}\right\}$ and $l \in \{0, \dots, 4d^*\}$ (which works analogously for $h \in \mathcal{H}^{(0)}$). Respecting the above representations, all the functions $\sigma_{id}(h) = h$ for $h \in \mathcal{H}^{(l)}$ comply with the structure of the functions $f_1^{(l+1)}$ in Lemma 9, if we use the following specifications of the parameters in that lemma: The Lipschitz constant L is chosen as the maximum of the Lipschitz constants of σ_{id} (which is obviously 1) and of the N-admissible sigmoidal function σ. Thus, property (38) is satisfied due to $\|\sigma\|_\infty \leq 1$, $L \geq 1$, and $|\sigma_{id}(x)| = |x|$. The parameter l in this lemma is $4l+2$ (regarding the l in $\mathcal{H}^{(l)}$ above) and the parameters K_r with $r = 0, \dots, l$ take repeatedly the values $\tilde{d}, 4 \cdot d^* \cdot \binom{d^*+N}{d^*} \cdot (N+1) \cdot (M_n+1)^{d^*}, \binom{d^*+N}{d^*} \cdot (N+1) \cdot (M_n+1)^{d^*}$ and K one after another, where $\tilde{d}$ is equal to d^* except for K_0, where it is d. Since all the coefficients $c_{i,j}^{(r)}$ with $r = 0, \dots, l$, $i = 1, \dots, K_{r+1}$, $j = 1, \dots, K_r$ (using $K_{l+1} = 1$ again) are 0, 1, or one of the $\mu_i, \lambda_{i,l}, \theta_{i,l,m}$ in the definition of $\mathcal{F}^{(neural\ networks)}_{M_n,N,d^*,d,\alpha}$, we can use $\overline{C} = \alpha$ for n sufficiently large.

Let h and $\bar{h}$ be functions in $\mathcal{H}^{(l)}$. Since they comply with the structure of the functions in Lemma 9 according to the above argumentation, we can conclude

$$\begin{aligned} &\|h - \bar{h}\|_{\infty,[-a_n,a_n]^d} \\ &\leq (4l+3) \cdot L^{4l+3} \cdot \left(4 \cdot d^* \cdot \binom{d^*+N}{d^*} \cdot (N+1) \cdot (M_n+1)^{d^*} + 1\right)^{4l+3} \cdot \alpha^{4l+2} \\ &\qquad \cdot \max\{a_n, 1\} \cdot \max_{\substack{r=0,\dots,\tilde{l},\ i=1,\dots,K_{r+1},\\ j=0,\dots,K_r}} \left|c_{i,j}^{(r)} - \bar{c}_{i,j}^{(r)}\right| \end{aligned}$$

$$\leq n^{c_{52}} \cdot \max_{\substack{r=0,\ldots,\tilde{l},\ i=1,\ldots,K_{r+1},\\ j=0,\ldots,K_r}} \left| c_{i,j}^{(r)} - \bar{c}_{i,j}^{(r)} \right|$$

for n sufficiently large and an adequately chosen $c_{52} > 0$ thanks to $a_n \leq M_n \leq \frac{M_n}{\eta_n} \leq n^{c_9}$. Thus, if we consider an arbitrary $h \in \mathcal{H}^{(l)}$, it suffices to choose the coefficients $\bar{c}_{i,j}^{(r)}$ of a function $\bar{h} \in \mathcal{H}^{(l)}$ such that

$$\left| c_{i,j}^{(r)} - \bar{c}_{i,j}^{(r)} \right| \leq \frac{\varepsilon_n}{n^{c_{52}}} \tag{39}$$

holds for all possible indices, in order to satisfy $\|h - \bar{h}\|_{\infty,[-a_n,a_n]^d} \leq \varepsilon_n$. These coefficients $c_{i,j}^{(r)}$ have to take values in $[-\alpha, \alpha]$, and for n sufficiently large, which is assumed in the following, the relation $\alpha = \log(n) \cdot \frac{M_n^{d^*+p\cdot(2N+3)+1}}{\eta_n} \leq n^{c_{53}}$ holds. Then due to $\varepsilon_n \geq \frac{1}{n^{c_8}}$ a number of

$$\left\lceil \frac{2 \cdot \alpha \cdot n^{c_{52}}}{2 \cdot \varepsilon_n} \right\rceil \leq n^{c_{54}}$$

different $\bar{c}_{i,j}^{(r)}$ suffices to guarantee, that at least one of them satisfies the relation (39) for any $c_{i,j}^{(r)}$ with fixed indices. Furthermore, the coefficients $c_{i,j}^{(r)}$, which can actually differ regarding different $h \in \mathcal{H}^{(l)}$, are the ones originating from the coefficients $\mu_i, \lambda_{i,l}, \theta_{i,l,m}$ in the definition of $\mathcal{F}_{M_n,N,d^*,d,\alpha}^{(neural\ networks)}$. Using (8), their number can be bounded by $c_{55} \cdot M_n^{d^*}$. So the logarithm of the covering number $\mathcal{N}(\varepsilon_n, \mathcal{H}^{(l)}, \|\cdot\|_{\infty,[-a_n,a_n]^d})$ can be bounded by

$$\log\left(\mathcal{N}(\varepsilon_n, \mathcal{H}^{(l)}, \|\cdot\|_{\infty,[-a_n,a_n]^d})\right) \leq \log\left((n^{c_{54}})^{c_{55}\cdot M_n^{d^*}}\right) \leq c_{10} \cdot \log(n) \cdot M_n^{d^*},$$

which proves the assertion. □

Proof of Lemma 4. Since due to Definition 3 there is a point $t_\sigma \in \mathbb{R}$, such that none of the derivatives up to the order N is zero in t_σ, the one-layered neural network described in the assertion of this lemma can be formulated explicitly as

$$\sum_{k=1}^{N+1} (-1)^{N+k-1} \cdot \frac{R^N}{\sigma^{(N)}(t_\sigma)} \cdot \binom{N}{k-1} \cdot \sigma\left(\frac{k-1}{R} \cdot x + t_\sigma\right) \tag{40}$$

$$= \sum_{k=0}^{N} (-1)^{N+k} \cdot \frac{R^N}{\sigma^{(N)}(t_\sigma)} \cdot \binom{N}{k} \cdot \sigma\left(\frac{k}{R} \cdot x + t_\sigma\right)$$

$$= (-1)^N \cdot \frac{R^N}{\sigma^{(N)}(t_\sigma)} \cdot \sum_{k=0}^{N} (-1)^k \cdot \binom{N}{k} \cdot \sigma\left(\frac{k}{R} \cdot x + t_\sigma\right) \tag{41}$$

Since Definition 3 implies, that σ is $N+1$ times continously differentiable, it can be expanded in a Taylor series with Lagrange remainder around t_σ up to order N and we

can conclude (defining $0^0 = 1$)

$$\begin{aligned}
&\sum_{k=0}^{N} (-1)^k \cdot \binom{N}{k} \cdot \sigma\left(\frac{k}{R} \cdot x + t_\sigma\right) \\
&= \sum_{k=0}^{N} (-1)^k \cdot \binom{N}{k} \cdot \left(\sum_{j=0}^{N} \frac{\sigma^{(j)}(t_\sigma) \cdot (xk)^j}{R^j \cdot j!} + \frac{\sigma^{(N+1)}(\xi_k) \cdot (xk)^{N+1}}{R^{N+1} \cdot (N+1)!}\right) \\
&= \sum_{j=0}^{N} \frac{\sigma^{(j)}(t_\sigma) \cdot x^j}{R^j \cdot j!} \cdot \sum_{k=0}^{N} (-1)^k \cdot k^j \cdot \binom{N}{k} \\
&\qquad + \frac{x^{N+1}}{R^{N+1} \cdot (N+1)!} \cdot \sum_{k=0}^{N} (-1)^k \cdot k^{N+1} \cdot \sigma^{(N+1)}(\xi_k) \cdot \binom{N}{k},
\end{aligned}$$

where $\xi_k \in \left[t_\sigma - \frac{k}{R} \cdot |x|, t_\sigma + \frac{k}{R} \cdot |x|\right]$ for all $0 \leq k \leq N$. Next, we notice that

$$\sum_{k=0}^{N} (-1)^k \cdot k^j \cdot \binom{N}{k} = N! \cdot (-1)^N \cdot S_j^{(N)}$$

holds, where $S_j^{(N)}$ is the well-known Stirling number of the second kind, which describes the number of options to split a set of j elements into N non-empty subsets and which is equal to zero for $0 \leq j < N$ and equal to one for $j = N$ (cf., e.g., the recurrence relation on page 825 in Abramovitz and Stegun (1972), which actually holds for all combinations of $j, N \in \mathbb{N}_0$ and implies the mentioned values in connection with the binomial theorem). This simplifies the above sum to

$$\frac{\sigma^{(N)}(t_\sigma) \cdot x^N}{R^N} \cdot (-1)^N + \frac{x^{N+1}}{R^{N+1} \cdot (N+1)!} \cdot \sum_{k=0}^{N} (-1)^k \cdot k^{N+1} \cdot \sigma^{(N+1)}(\xi_k) \cdot \binom{N}{k}.$$

Plugging this into the representation of (41) leads to

$$\begin{aligned}
&(-1)^N \cdot \frac{R^N}{\sigma^{(N)}(t_\sigma)} \cdot \left(\frac{\sigma^{(N)}(t_\sigma) \cdot x^N}{R^N} \cdot (-1)^N \right. \\
&\qquad \left. + \frac{x^{N+1}}{R^{N+1} \cdot (N+1)!} \cdot \sum_{k=0}^{N} (-1)^k \cdot k^{N+1} \cdot \sigma^{(N+1)}(\xi_k) \cdot \binom{N}{k}\right) \\
&= x^N + \frac{(-1)^N \cdot x^{N+1}}{R \cdot \sigma^{(N)}(t_\sigma) \cdot (N+1)!} \cdot \sum_{k=0}^{N} (-1)^k \cdot k^{N+1} \cdot \sigma^{(N+1)}(\xi_k) \cdot \binom{N}{k}.
\end{aligned}$$

This implies the assertion of the lemma, since the derivative of order $N+1$ is bounded on $\mathbb{R}$ (cf., Definition 3) and the weights chosen in (40) satisfy the announced bounds. □