On estimation of surrogate models for high-dimensional computer experiments [*]

Benedikt Bauer[1,†], Felix Heimrich[2], Michael Kohler[1] and Adam Krzyżak[3]

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

[2] *Fachbereich Maschinenbau, Technische Universität Darmstadt, Otto-Bernd-Str. 2, 64287 Darmstadt, Germany, email: heimrich@dik.tu-darmstadt.de*

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

December 14, 2016

Abstract

Estimation of surrogate models for computer experiments leads to nonparametric regression estimation problems without noise in the dependent variable. In this paper we propose an empirical maximal deviation minimization principle to construct estimates in this context, and analyze the rate of convergence of corresponding quantile estimates. As an application we consider estimation of high-dimensional computer experiments by neural networks, and show that here we can circumvent the so-called curse of dimensionality by imposing rather general assumptions on the structure of the regression function. The estimates are illustrated by applying them to simulated data and to a simulation model in mechanical engineering.

AMS classification: Primary 62G05; secondary 62G20.

Key words and phrases: Computer experiments, curse of dimensionality, neural networks, nonparametric regression without noise in the dependent variable, quantile estimates, rate of convergence, surrogate models.

1 Introduction

Physical phenomena are nowadays often described by mathematical models, which enables the use of so-called computer experiments instead of real experiments in order to analyze them.

In the simplest case the mathematical model is described by a function $m : \mathbb{R}^d \to \mathbb{R}$, which models the relation between d–dimensional input and a real-valued output. Due

[*] Running title: *On estimation of surrogate models*

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

to uncertainty in nature, it is often impossible to characterize the input exactly. Instead we can describe it by the input random variable X with a given distribution. So the mathematical model describes the outcome Y of a physical phenomenon by

$$Y = m(X), \tag{1}$$

where X is an $\mathbb{R}^d$-valued random variable and $m : \mathbb{R}^d \to \mathbb{R}$ is a real–valued function. Here this function can be, e.g., the solution of a partial differential equation system, where the value of X determines the values of parameters and initial conditions of this system. The aim of studying the physical phenomenon is to derive characteristics of the outcome Y. In the mathematical model Y is a real-valued random variable, and we are interested in the distribution of this random variable.

It is often possible to write a complex computer program which generates the values of function m. Using this computer model and independent copies $X_1, X_2, \dots$ of X, we can then carry out computer experiments: We evaluate the computer model for these random inputs, and get an (independent) sample

$$Y_1 = m(X_1), \dots, Y_n = m(X_n) \tag{2}$$

of the distribution of the outcome Y in our mathematical model. Using standard methods of (nonparametric) statistics, this distribution can be characterized, by, e.g., estimating its density or quantiles.

The mathematical models for problems that arise in practice and consequently the corresponding computer program are usually rather complex. Thus, computing of values of $m(X)$ is time-consuming, so it is not possible to generate the data in (2) for a large sample size n. One idea to circumvent this problem is to use so-called surrogate models for m. Here we begin by generating data

$$(x_1, m(x_1)), \dots, (x_n, m(x_n)) \tag{3}$$

by evaluating the computer model for suitably chosen input values $x_1, \dots, x_n \in \mathbb{R}^d$, and using this data to construct an estimate

$$m_n(\cdot) = m_n(\cdot, (x_1, m(x_1)), \dots, (x_n, m(x_n))) : \mathbb{R}^d \to \mathbb{R} \tag{4}$$

of m. The input values $x_1, \dots, x_n \in \mathbb{R}^d$ could be chosen as realizations of the random variables $X_1, \dots, X_n$, but in order to estimate (4) with small error we can choose them differently. Given the estimate (4), one can replace (1) by its surrogate model

$$Y = m_n(X) \tag{5}$$

and study the distribution of Y. In this estimated model a sample

$$Y_{n+1} = m_n(X_{n+1}), \dots, Y_{n+N_n} = m_n(X_{n+N_n}) \tag{6}$$

of Y can be usually computed for a sample size N_n which is much larger than n, and this sample can be used together with nonparametric statistics in order to estimate some aspects of the distribution of Y. In this paper, we will focus on quantile estimates based on surrogate models of Y and analyze their convergence.

The concept of surrogate models explained above has been introduced and investigated with the aid of simulated and real data by several authors using different estimation techniques. After Bucher and Burgund (1990), Kim and Na (1997) and Das and Zheng (2000) had relied on quadratic response surfaces, Hurtado (2004), Deheeger and Lemaire (2010) and Bourinet, Deheeger and Lemaire (2011) used support vector machines, whereas Papadrakakis and Lagaros (2002) concentrated on neural networks and Kaymaz (2005) and Bichon et al. (2008) made use of kriging. Theoretical results concerning the rate of convergence of corresponding quantile estimates have been derived in Enss et al. (2016).

For the purpose of constructing a surrogate m_n also any kind of nonparametric regression estimate could be used. For instance we could use kernel regression estimates (cf., e.g., Nadaraya (1964, 1970), Watson (1964), Devroye and Wagner (1980), Stone (1977, 1982) or Devroye and Krzyżak (1989)), partitioning regression estimates (cf., e.g., Györfi (1981) or Beirlant and Györfi (1998)), nearest neighbor regression estimates (cf., e.g., Devroye (1982) or Devroye, Györfi, Krzyżak and Lugosi (1994)), orthogonal series regression estimates (cf., e.g., Rafajłowicz (1987) or Greblicki and Pawlak (1985)), least squares estimates (cf., e.g., Lugosi and Zeger (1995) or Kohler (2000)) or smoothing spline estimates (cf., e.g., Wahba (1990) or Kohler and Krzyżak (2001)).

In this paper we will use neural network estimates with several hidden layers as surrogate models for our quantile estimates and derive novel results concerning the rates of convergence for them. For the idea of applying neural networks to nonlinear function estimation, classification and learning we refer the reader to the monographs Hertz, Krogh and Palmer (1991), Devroye, Györfi and Lugosi (1996), Ripley (2008), Anthony and Bartlett (1999), Györfi et al. (2002), Haykin (2008) and Hastie, Tibshirani and Friedman (2011). Consistency of nonparametric regression estimates using neural networks has been studied by Mielniczuk and Tyrcha (1993) and Lugosi and Zeger (1995). The rate of convergence of neural network regression estimates with one hidden layer has been analyzed by Barron (1991, 1993) and McCaffrey and Gallant (1994), and in connection with feedforward neural network with several hidden layers in Kohler and Krzyżak (2005, 2016).

It is well-known that one has to impose smoothness conditions on m in order to derive non-trivial rates of convergence. A usual assumption is the so-called (p, C)-smoothness, where (roughly speaking) m is p times continously differentiable. Considering the L_2-error, Stone (1982) showed that the optimal rate of convergence of an estimate of m is

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

Here the rate of convergence is not good for large dimension d. It was shown in Stone (1985, 1994) that this so-called curse of dimensionality can be avoided by imposing special assumptions on the structure of m. In particular, Stone (1994) showed that in case of a so-called interaction model, where m is a sum of (p, C)-smooth functions, such that each of them uses at most d^* input components of m, the optimal rate of convergence is

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

A similar rate of convergence was obtained in Kohler and Krzyżak (2016) for a more general set of functions using neural network estimates with several hidden layers.

All these rates assume that the observed values of m contain some noise. But due to the motivation by computer experiments, we can assume observations without additional errors in this paper. So it seems reasonable to examine, if better rates of convergence are possible in this case. Regarding the expected L_1-error Kohler and Krzyżak (2013) showed for some combinations of p and d that the rate of convergence

$$n^{-\frac{p}{d}} \tag{7}$$

is achievable. Kohler (2014) showed that in this context this rate cannot be improved, even in case of adaptively chosen values for the covariates.

In order to circumvent the curse of dimensionality, we will use the following assumption on the structure of m, which was introduced in Kohler and Krzyżak (2016). It should hold in particular if the outcome of a complex model is computed in several steps, where in each step only a few of the results of the previous step (maybe combined with some of the input parameters) are used.

Definition 1. *Let $d \in \mathbb{N}$, $d^* \in \{1, \dots, d\}$ 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 $a_1, \dots, a_{d^*} \in \mathbb{R}^d$ and $f : \mathbb{R}^{d^*} \to \mathbb{R}$ such that*

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

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

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

This definition includes other types of structures of m assumed in the literature, such as the additive model (cf., e.g., Stone (1985)), the interaction model (cf., e.g., Stone (1994)) or the projection pursuit (cf., e.g., Friedman and Stuetzle (1981)). Functions complying with one of the mentioned structures belong to the class of generalized hierarchical interaction models of order d^* and level $l = 1$, where $d^* = 1$ in case of additive functions or projection pursuit.

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

Definition 2. a) *Let $p = k + s$ for some $k \in \mathbb{N}_0$ and $0 < s \leq 1$. A* **function** $m : \mathbb{R}^d \to \mathbb{R}$ *is called* (p, C)**-smooth***, if for every $\alpha = (\alpha_1, \dots, \alpha_d) \in \mathbb{N}_0^d$ with $\sum_{j=1}^d \alpha_j = k$ the partial derivative $\frac{\partial^k m}{\partial x_1^{\alpha_1} \dots \partial x_d^{\alpha_d}}$ exists and satisfies*

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

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

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

In this paper we will derive some general results regarding the approximation quality of function estimates minimizing the empirical maximal deviation and we will analyze the rate of convergence of the corresponding surrogate model quantile estimates. We will use these results to analyze neural network estimate with several hidden layers. Given a function m, which satisfies a (p, C)–smooth generalized hierarchical interaction model of order d^* and level l with $p \in (0, 1]$ and some additional assumptions regarding the partial functions of the model, this estimate m_n has the property, that outside of an event, whose probability tends to zero for n tending to infinity, the following inequality holds:

$$\begin{aligned}&\mathbf{P}_X \left(\left\{ x \in \left[-\log(n)^2, \log(n)^2 \right]^d \, : \, |m_n(x) - m(x)| > c_1 \cdot \log(n)^{2p + \frac{p}{p+d^*}} \cdot n^{-\frac{p}{p+d^*}} \right\} \right) \\ &\leq c_2 \cdot \log(n)^{\frac{p}{p+d^*}} \cdot n^{-\frac{p}{p+d^*}}.\end{aligned}$$

The main improvement of this rate of convergence result over the rate in (7) is that here the rate of convergence depends on d^* rather than d in case that the function to be estimated satisfies a generalized hierarchical interaction model.

Using this estimate as a surrogate model for the construction of a quantile estimate, we obtain the rate of convergence of this quantile estimate of order

$$\log(n)^{2p + \frac{p}{p+d^*}} \cdot n^{-\frac{p}{p+d^*}} + \mathbf{P}_X \left(\mathbb{R}^d \setminus [-\log(n)^2, \log(n)^2]^d \right) + \frac{1}{\sqrt{N_n}},$$

where N_n is the size of the Monte Carlo sample. Here the rate of convergence is again independent of the dimension of X. By using simulated data we demonstrate in several high-dimensional settings that our newly proposed quantile estimate outperforms other quantile estimates for finite sample size.

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

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

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

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

For nonnegative random variables Z_n and Y_n, we write

$$Z_n = O_{\mathbf{P}}(Y_n)$$

if they satisfy $\lim_{c\to\infty} \limsup_{n\to\infty} \mathbf{P}\{Z_n > c \cdot Y_n\} = 0$. If not otherwise stated, any c_i with $i \in \mathbb{N}$ here and in the following symbolizes a real nonnegative constant, which is independent of the sample size n.

The outline of this paper is as follows: In Section 2 a general class of estimates is introduced, their rate of convergence is analyzed and it is shown how the results can be applied in order to estimate quantiles based on surrogate models. In Section 3 we describe the application of our general result to neural networks. Section 4 illustrates the finite sample size behaviour of the estimates by applying them to simulated and real data. The proofs are contained in Section 5.

2 Empirical maximal deviation minimization

2.1 Definition of the estimates

In order to define our estimates, we choose a set $\mathcal{F}_n$ of functions $f : \mathbb{R}^d \to \mathbb{R}$ and select from this set a function, which fits our data within a given set $B_n \subseteq \mathbb{R}^d$ well. Often this is done using the principle of the least squares, where the function is chosen such that the average squared error on the observed data is minimal. We will not apply this for fitting a surrogate model to a computer experiment, since there we observe the function to be estimated without additional random errors. Instead we propose to minimize the empirical maximal deviation of the estimate on B_n, i.e., we propose to minimize the maximal absolute error on the observed data contained in B_n. Formally, the resulting least empirical deviation estimate is defined by

$$m_n(\cdot) = \arg \min_{f \in \mathcal{F}_n} \max_{\substack{i=1,\dots,n,\\ X_i \in B_n}} |f(X_i) - m(X_i)| . \tag{8}$$

Here we assume for simplicity, that the above minimum exists, however we do not require it to be unique. In case that it does not exist, one can define the estimate by

$$m_n(\cdot) \in \mathcal{F}_n \quad \text{and} \quad \max_{\substack{i=1,\dots,n,\\ X_i \in B_n}} |m_n(X_i) - m(X_i)| \leq \inf_{f \in \mathcal{F}_n} \max_{\substack{i=1,\dots,n,\\ X_i \in B_n}} |f(X_i) - m(X_i)| + \frac{1}{n},$$

in which case the theoretical results below also hold.
Computation of the estimate (8) can be done using nonlinear programming, e.g., gradient descent or quasi-Newton methods.

2.2 General error bounds

Theorem 1. *Let X, X_1, X_2, ... be independent and identically distributed $\mathbb{R}^d$–valued random variables and let $m : \mathbb{R}^d \to \mathbb{R}$ be a (measurable) function. For $n \in \mathbb{N}$ let $\epsilon_n \geq 0$, let $\mathcal{F}_n$ be a set of (measurable) functions and let $B_n \subseteq \mathbb{R}^d$ be measurable. Assume that* *the size of the smallest $\varepsilon_n - \|\cdot\|_{\infty,B_n}$–cover of $\mathcal{F}_n$ (abbreviated by $\mathcal{N}(\varepsilon_n, \mathcal{F}_n, \|\cdot\|_{\infty,B_n})$) can be bounded by*

$$\mathcal{N}(\varepsilon_n, \mathcal{F}_n, \|\cdot\|_{\infty,B_n}) \leq \mathcal{N}_{\infty,B_n}(\varepsilon_n) \tag{9}$$

for all $n \in \mathbb{N}$, where $\mathcal{N}_{\infty,B_n}(\varepsilon_n) \to \infty$ for $n \to \infty$.
Then outside of an event, whose probability tends to zero for $n \to \infty$, the estimate m_n of m defined by (8) satisfies

$$\mathbf{P}_X\left(\left\{x \in B_n \,:\, |m_n(x) - m(x)| > 3 \cdot \epsilon_n + 2 \cdot \inf_{f \in \mathcal{F}_n} \|f - m\|_{\infty,B_n}\right\}\right)$$
$$\leq 2 \cdot \frac{\log\left(\mathcal{N}_{\infty,B_n}(\varepsilon_n)\right)}{n}.$$

2.3 Quantile estimation based on surrogate models

In the sequel we use Theorem 1 to define estimates of quantiles of $Y = m(X)$. Let

$$G_{m(X)}(y) = \mathbf{P}\{m(X) \leq y\} \tag{10}$$

be the cumulative distribution function of $m(X)$, and for $\alpha \in (0,1)$ let

$$q_{m(X),\alpha} = \inf\{y \in \mathbb{R} \,:\, G_{m(X)}(y) \geq \alpha\} \tag{11}$$

be the α–quantile of $m(X)$. In order to estimate $q_{m(X),\alpha}$, we use the surrogate m_n of Theorem 1 together with N_n additional values $X_{n+1}, \ldots, X_{n+N_n}$ of X and define an estimate

$$\hat{G}_{m_n(X),N_n}(y) = \frac{1}{N_n}\sum_{i=1}^{N_n} I_{(-\infty,y]}(m_n(X_{n+i})) \tag{12}$$

of G. Then we estimate $q_{m(X),\alpha}$ by the corresponding plug–in quantile estimate

$$\hat{q}_{m_n(X),N_n,\alpha} = \inf\{y \in \mathbb{R} \,:\, \hat{G}_{m_n(X),N_n}(y) \geq \alpha\}. \tag{13}$$

In the next theorem we present the rate of convergence result for (13).

Theorem 2. *Let X, X_1, X_2, …be independent and identically distributed $\mathbb{R}^d$–valued random variables, let $m : \mathbb{R}^d \to \mathbb{R}$ be a (measurable) function, let $B_n \subseteq \mathbb{R}^d$ such that $\lim_{n\to\infty} \mathbf{P}_X\left(\mathbb{R}^d \setminus B_n\right) = 0$ and let $\alpha \in (0,1)$. Assume that $m(X)$ has a density with respect to the Lebesgue measure, which is continuous on $\mathbb{R}$ and positive at $q_{m(X),\alpha}$. Let the surrogate m_n satisfy*

$$\mathbf{P}_X\left(\{x \in B_n \,:\, |m_n(x) - m(x)| > \delta_n\}\right) \leq \zeta_n \tag{14}$$

for some $\delta_n, \zeta_n \geq 0$, where $\zeta_n \to 0$ for $n \to \infty$. Assume $N_n \to \infty$ $(n \to \infty)$. Then the quantile estimate $\hat{q}_{m_n(X),N_n,\alpha}$ defined above satisfies

$$|\hat{q}_{m_n(X),N_n,\alpha} - q_{m(X),\alpha}| = O_{\mathbf{P}}\left(\frac{1}{\sqrt{N_n}} + \zeta_n + \mathbf{P}_X\left(\mathbb{R}^d \setminus B_n\right) + \delta_n\right).$$

Corollary 1. *Let the assumptions of Theorem 1 hold for* $\varepsilon_n = \frac{1}{n}$ *and let the surrogate* m_n *be defined as therein. If the assumptions of Theorem 2 also hold, then the corresponding quantile estimate* $\hat{q}_{m_n(X),N_n,\alpha}$ *defined as in (13) satisfies*

$$\begin{aligned}&|\hat{q}_{m_n(X),N_n,\alpha} - q_{m(X),\alpha}|\\ &= O_{\mathbf{P}}\left(\frac{1}{\sqrt{N_n}} + \frac{1}{n} + \frac{\log\left(\mathcal{N}_{\infty,B_n}\left(\frac{1}{n}\right)\right)}{n} + \mathbf{P}_X\left(\mathbb{R}^d \setminus B_n\right) + \inf_{f \in \mathcal{F}_n} \|f - m\|_{\infty,B_n}\right).\end{aligned}$$

Proof. Since the requirements of Theorem 1 are met, its assertion holds. Consequently, (14) holds for $\delta_n = 3 \cdot \epsilon_n + 2 \cdot \inf_{f \in \mathcal{F}_n} \|f - m\|_{\infty,B_n}$ and $\zeta_n = 2 \cdot \frac{\log(\mathcal{N}_{\infty,B_n}(\varepsilon_n))}{n}$. Application of Theorem 2 implies the assertion. □

3 Neural network surrogate models

Neural network estimates often use a type of activation function $\sigma : \mathbb{R} \to [0,1]$ called squashing activation function $\sigma : \mathbb{R} \to [0,1]$ that is nondecreasing and satisfying

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

Examples of squashing functions include the sigmoidal squasher

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

and the piecewise linear squashing function

$$\sigma(x) = \begin{cases} 1 & \text{for } x \geq \frac{1}{2} \\ x + \frac{1}{2} & \text{for } x \in \left(-\frac{1}{2}, \frac{1}{2}\right) \\ 0 & \text{for } x \leq -\frac{1}{2}. \end{cases}$$

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

$$f(x) = \sum_{i=1}^{K_l} c_i^{(l)} \cdot f_i^{(l)}(x) + c_0^{(l)}, \tag{15}$$

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

$$f_i^{(r)}(x) = \sigma\left(\sum_{j=1}^{K_{r-1}} c_{i,j}^{(r-1)} \cdot f_j^{(r-1)}(x) + c_{i,0}^{(r-1)}\right) \tag{16}$$

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

$$f_i^{(1)}(x) = \sigma\left(\sum_{j=1}^{d} c_{i,j}^{(0)} \cdot x^{(i)} + c_{i,0}^{(0)}\right) \tag{17}$$

for some $c_{i,0}^{(0)}, \dots, c_{i,d}^{(0)} \in \mathbb{R}$. This means the layers of the network are numbered from inside to outside, i.e., the innermost sum corresponds to the first layer and so on.

We define so-called spaces of hierarchical neural networks with parameters K, M, d^*, d and level l as follows (see Kohler and Krzyżak (2016)). For $M \in \mathbb{N}$, $d \in \mathbb{N}$, $d^* \in \{1, \dots, d\}$ and $\alpha, \beta, \gamma > 0$ we denote the sets of all functions $f : \mathbb{R}^d \to \mathbb{R}$ which satisfy

$$f(x) = \sum_{i=1}^{(M+1)^{d^*}} d_i \cdot \sigma\left(\sum_{j=1}^{d^*} b_{i,j} \cdot \sigma\left(\sum_{m=1}^{d} a_{i,j,m} \cdot x^{(m)} + a_{i,j,0}\right) + b_{i,0}\right) + d_0 \quad (x \in \mathbb{R}^d) \tag{18}$$

for some $a_{i,j,m}, b_{i,j}, d_i \in \mathbb{R}$, where

$$|a_{i,j,m}| \le \alpha, \quad |b_{i,j}| \le \beta \quad \text{and} \quad |d_i| \le \gamma$$

for all $i \in \{0, 1, \dots, (M+1)^{d^*}\}$, $j \in \{0, 1, \dots, d^*\}$ and $m \in \{0, 1, \dots, d\}$, by $\mathcal{F}_{M,d^*,d,\alpha,\beta,\gamma}^{(neural\ networks)}$. Here in the first and in the second hidden layer we are using $d^* \cdot (M+1)^{d^*}$ and $(M+1)^{d^*}$ neurons, respectively. However, the neural network has only

$$\begin{aligned} W\left(\mathcal{F}_{M,d^*,d,\alpha,\beta,\gamma}^{(neural\ networks)}\right) &:= (M+1)^{d^*} + 1 + (M+1)^{d^*} \cdot (d^*+1) + (M+1)^{d^*} \cdot d^* \cdot (d+1) \\ &= (M+1)^{d^*} \cdot (d^* \cdot (d+2) + 2) + 1 \end{aligned} \tag{19}$$

weights. This is due to the fact, that the two hidden layers of the neural network are not fully connected. Instead, each neuron in the second hidden layer is connected with d^* neurons in the first hidden layer, and this is done in such a way that each neuron in the first hidden layer is connected with exactly one neuron in the second hidden layer. The exemplary network in Figure 1 gives a good idea of how the sparse connection works.

In case $l = 0$ we define our space of hierarchical neural networks by

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

And for $l > 0$ we define

$$\begin{aligned} \mathcal{H}^{(l)} \quad = \quad \Bigg\{ h : \mathbb{R}^d \to \mathbb{R} \quad : \quad & h(x) = \sum_{k=1}^{K} g_k\left(f_{1,k}(x), \dots, f_{d^*,k}(x)\right) \quad (x \in \mathbb{R}^d) \\ & \text{for some } g_k \in \mathcal{F}_{M,d^*,d^*,\alpha,\beta,\gamma}^{(neural\ networks)} \text{ and } f_{j,k} \in \mathcal{H}^{(l-1)} \Bigg\}. \end{aligned} \tag{20}$$

The class $\mathcal{H}^{(0)}$ is a set of neural networks with two hidden layers and a number of weights given by (19). From this one can conclude recursively that for $l > 0$ the class $\mathcal{H}^{(l)}$ is a

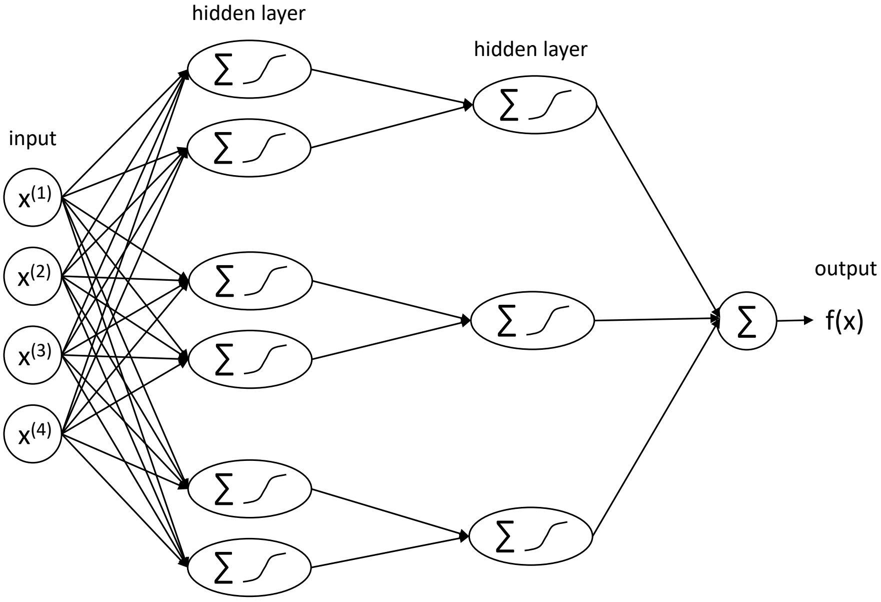

Figure 1: A not completely connected neural network of the type $f(x) = \sum_{i=1}^{3} d_i \cdot \sigma\left(\sum_{j=1}^{2} b_{i,j} \cdot \sigma\left(\sum_{m=1}^{4} a_{i,j,m} \cdot x^{(m)} + a_{i,j,0}\right) + b_{i,0}\right) + d_0$, where all weights with an index containing 0 are neglected in the diagram.

set of neural networks with $2 \cdot l + 2$ hidden layers. Let $N\left(\mathcal{H}^{(l)}\right)$ denote the number of linked two-layered neural networks from $\mathcal{F}_{M,d^*,d,\alpha,\beta,\gamma}^{(neural\ networks)}$ that define the functions from $\mathcal{H}^{(l)}$. Then the recursion

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

holds, yielding the solution

$$N\left(\mathcal{H}^{(l)}\right) = \sum_{t=1}^{l} d^{*t-1} \cdot K^t + \left(d^* \cdot K\right)^l. \tag{21}$$

Furthermore, the weights of a function from $\mathcal{H}^{(l)}$ can be parameterized by at most

$$N\left(\mathcal{H}^{(l)}\right) \cdot W\left(\mathcal{F}_{M,d^*,d,\alpha,\beta,\gamma}^{(neural\ networks)}\right) \tag{22}$$

parameters.

Next we choose in our least maximal deviation estimate (8) the set $\mathcal{F}_n$ as the set $\mathcal{H}^{(l)}$, with parameters specified in the following theorem.

Theorem 3. *Let $X, X_1, X_2, \ldots, X_n$ be independent and identically distributed $\mathbb{R}^d$–valued random variables, which comply with $\mathbf{E}\left\{\exp\left(\bar{c} \cdot \|X\|_\infty\right)\right\} < \infty$ for some $\bar{c} > 0$, and let $m : \mathbb{R}^d \to \mathbb{R}$ be a (measurable) function, which satisfies a (p, C)–smooth generalized hierarchical interaction model of order d^* and finite level l with $p \in (0, 1]$. Respecting Definition 1 b), let all functions $g_k, f_{j,k}$ be bounded and let all functions g_k be Lipschitz continuous. For $n \in \mathbb{N}$ let $\mathcal{F}_n$ be $\mathcal{H}^{(l)}$ defined as in (20) with K, d, d^* as in the definition of m, $M = M_n = \left\lfloor \left(\frac{n}{\log(n)}\right)^{1/(p+d^*)} \right\rfloor$, $\alpha = \log(n) \cdot n^2 \cdot M_n$, $\beta = \log(n) \cdot M_n^{d^*}$, $\gamma = \log(n)$, and a Lipschitz continuous squashing function σ that satisfies*

$$|\sigma(x) - 1| \leq \frac{1}{x} \quad \text{if} \quad x > 0 \quad \text{and} \quad |\sigma(x)| \leq \frac{1}{|x|} \quad \text{if} \quad x < 0. \tag{23}$$

Define the estimate m_n of m by (8). Then outside of an event, whose probability tends to zero for n tending to infinity,

$$\begin{aligned} &\mathbf{P}_X \left(\left\{ x \in \left[-\log(n)^2, \log(n)^2\right]^d \ : \ |m_n(x) - m(x)| > c_1 \cdot \log(n)^{2p + \frac{p}{p+d^*}} \cdot n^{-\frac{p}{p+d^*}} \right\} \right) \\ &\quad \leq c_2 \cdot \log(n)^{\frac{p}{p+d^*}} \cdot n^{-\frac{p}{p+d^*}} \end{aligned}$$

holds for $c_1, c_2 > 0$ independent of n.

Remark 1. The requirement $\mathbf{E}\left\{\exp\left(\bar{c} \cdot \|X\|_\infty\right)\right\} < \infty$ for X is satisfied by many commonly used distributions, e.g., all normal distributions and uniform distributions on bounded sets.

Remark 2. It is easy to see that assumption (23) is satisfied in case of the sigmoidal squasher or the piecewise linear activation function defined above.

Keeping the result of Theorem 3 in mind, we can deduce the following corollary from Theorem 2:

Corollary 2. *Let the assumptions of Theorems 2 and 3 hold and let the surrogate m_n be defined as therein. Then the corresponding quantile estimate $\hat{q}_{m_n(X),N_n,\alpha}$ defined as in (13) satisfies*

$$\begin{aligned} &|\hat{q}_{m_n(X),N_n,\alpha} - q_{m(X),\alpha}| \\ &\quad = O_{\mathbf{P}} \left(\frac{1}{\sqrt{N_n}} + \mathbf{P}_X \left(\mathbb{R}^d \setminus \left[-\log(n)^2, \log(n)^2\right]^d \right) + \log(n)^{2p + \frac{p}{p+d^*}} \cdot n^{-\frac{p}{p+d^*}} \right). \end{aligned}$$

Proof. Since the requirements of Theorem 3 are met, its assertion holds. Using this in assumption (14) of Theorem 2 implies the assertion of the corollary. □

Remark 3. The rate of convergence in Corollary 2 does not depend on the dimension d of X and hence circumvents the curse of dimensionality. It should be noted that a simple order statistics estimate, which does not use the values of X at all, also achieves a rate of convergence independent of d. Here the rate of convergence is $1/\sqrt{n}$, which converges in general even faster to zero than our complicated estimate in Corollary 2 (since $p/(p+d^*) \leq 1/2$ for $p \leq 1$). However, we conjecture that in case of (p,C)–smooth functions with $p > 1$ our newly proposed estimate achieves a rate of convergence faster than the one in Corollary 2. This will be illustrated in the next section using simulated data.

4 Application to simulated and real data

In order to illustrate how the quantile estimate developed in the previous sections behaves in case of finite sample sizes, we apply it to simulated and real data and compare the results with the conventional estimates using the software *MATLAB*.

The first alternative approach is a simple order statistic estimate $\hat{q}_{m(X),n,\alpha}$ (abbr. *order*), where we estimate $q_{m(X),\alpha}$ by the $\lceil \alpha \cdot n \rceil$-smallest value of n given data points.

The second competitive approach we consider works with an estimate based on a surrogate model. It approximates m by interpolating it with radial basis functions (abbr. *RBF*). Regarding the variety of modifications of this approach known in the literature, we focus on the version in Lazzaro and Montefusco (2002), where the authors use Wendland's compactly supported radial basis function $\phi(r) = (1-r)_+^6 \cdot (35r^2 + 18r + 3)$. The radius which scales the basis functions is chosen adaptively in our implementation, because this improved the RBF approach in the simulations.

The parameters l, K, d^*, and M_n of our neural network estimate (abbr. *neural*) defined in Theorem 3 and the corresponding quantile estimate in Corollary 2 are also chosen adaptively by splitting of the sample, using $n_{train} = \lfloor \frac{4}{5} \cdot n \rfloor$ realizations as the training set and $n_{test} = n - n_{train}$ realizations as the test set. The tested choices of these parameters were $\{0,1,2\}$ for l, $\{1,\ldots,5\}$ for K, $\{1,\ldots,d\}$ for d^*, and $\{0,\ldots,12\}$ for M_n, although the set of possible choices was reduced for some settings if several test runs showed that the whole range of choices is not needed. Since the (generally) non-linear and non-convex optimization problem in (8) which defines our estimate m_n cannot be easily solved exactly, we use the quasi-Newton method of the function *fminunc* in *MATLAB* to approximate its solution. At first it runs with the least squares objective function in order to reduce the deviations at all realizations simultaneously during these initializing steps, then the objective function is replaced by the maximal deviation criterion from (8) for the remaining approximative steps.

The illustrative simulated settings we use to compare the different approaches are listed in Table 1. These settings cover different distributions of the input random variable X and different dimensions of the regression function m. For the latter, we try

$$\begin{aligned} m_1(x) = \cot & \left(\frac{\pi}{1 + \exp\left(x_1^2 + 2 \cdot x_2 + \sin\left(6 \cdot x_4^3\right) - 3\right)} \right) \\ & + \exp\left(3 \cdot x_3 + 2 \cdot x_4 - 5 \cdot x_5 + \sqrt{x_6 + 0.9 \cdot x_7 + 0.1}\right) \end{aligned}$$

$$
\begin{aligned}
m_2(x) &= 5 \cdot x_1 \cdot x_2^2 - \frac{3}{2 \cdot x_3 + x_4 - 5 \cdot x_5 + 7} + \sin\left(x_2 - x_6 + \sqrt{x_7 + 1}\right) \\
&\quad - \exp\left(x_5 + 3 \cdot x_8 \cdot x_9^3\right) \\
m_3(x) &= \log(0.014 \cdot \left|2 \cdot x_1 - x_2 + 0.5 \cdot x_3 - x_4^2\right| + 0.95) - \frac{91}{5 \cdot x_4 - x_5 - 2 \cdot x_6 - 8} \\
&\quad + \frac{x_7}{131 \cdot \sqrt{|x_2|} + 1} - x_8 \cdot \min\left\{|x_9|, \frac{1}{76}\right\}.
\end{aligned}
$$

In all of these settings we approximate different quantiles $q_{m(X),\alpha}$ with levels $\alpha \in \{0.9, 0.95, 0.99\}$, where we use $N_n = 10^5$ artificially generated samples by the surrogate model m_n in case of *RBF* and *neural* approaches. We assess these quantile estimates by comparing them with $\hat{q}_{m(X),\tilde{N},\alpha}$ using sample size $\tilde{N} = 10^8$, because the exact $q_{m(X),\alpha}$ are not easily identifiable. The considered error measure is

Table 1: Test settings

no.	function m	d	n	distribution of X per component
I	m_1	7	180	$\mathcal{U}[0,1]$
II				$0.5 \cdot Z + 0.1, \quad Z \sim \exp(2)$
III	m_2	9	250	$\mathcal{U}[0,1]$
IV				$\min\{0.7 \cdot \lvert Z\rvert, 1\}, \quad Z \sim \mathcal{N}(0,1)$
V	m_3			$\mathcal{U}[0,1]$
VI				$\mathcal{N}\left(0.5, 0.2^2\right)$

$$
\varepsilon_\alpha\left(m_n\right) = \frac{\left|\hat{q}_{m_n(X),N_n,\alpha} - \hat{q}_{m(X),\tilde{N},\alpha}\right|}{\hat{q}_{m(X),\tilde{N},\alpha} - \hat{q}_{m(X),\tilde{N},0.01}},
$$

where the scaling in the denominator conveys a better idea of the relative extent of the error.

In view of the fact that simulation results depend on the randomly chosen data points, we compute the estimates repeatedly (50 times in the case of the surrogate model approaches and 10^4 times in case of the order statistic, which can be computed much faster) for repeatedly generated realizations of X and examine the median (plus interquartile range IQR) of $\varepsilon_\alpha\left(m_n\right)$.

Examination of the results in Table 2 shows that the less extreme quantiles (e.g. $\alpha = 0.9$) are mostly approximated significantly better than the extreme examples (e.g. $\alpha = 0.99$). This is not surprising thanks to the comparatively small number n of given observations of m, which cannot adequately represent every extreme behavior of m. Furthermore, in the considered test settings our newly proposed neural network estimate clearly outperforms the order statistic estimate and the competitive RBF approach.

Next we illustrate, how our newly proposed method can be used to estimate quantiles using the data produced by computer experiments, which simulate the behavior of suspension struts such as aircraft landing gears. The experimental setup is the digital twin

Table 2: Median (IQR) of the scaled error $\varepsilon_\alpha(m_n)$

no.	α	*0.9*	*0.95*	*0.99*
I	*order*	0.1182 (0.1417)	0.1392 (0.1638)	0.1993 (0.245)
	RBF	0.1791 (0.1031)	0.0895 (0.1076)	0.1839 (0.1371)
	neural	**0.01 (0.014)**	**0.0124 (0.0179)**	**0.0171 (0.0282)**
II	*order*	0.102 (0.1241)	0.1554 (0.1777)	0.3565 (0.4468)
	RBF	0.1406 (0.119)	0.1419 (0.231)	0.2104 (0.2911)
	neural	**0.0187 (0.0186)**	**0.0246 (0.025)**	**0.0624 (0.0984)**
III	*order*	0.0082 (0.0101)	0.0095 (0.0118)	0.0127 (0.0157)
	RBF	0.0171 (0.0141)	0.0157 (0.0177)	0.0231 (0.0294)
	neural	**0.0024 (0.0035)**	**0.0059 (0.0081)**	**0.0089 (0.0185)**
IV	*order*	0.0052 (0.0063)	0.0063 (0.0079)	0.006 (0.0072)
	RBF	0.0268 (0.0132)	0.0303 (0.0187)	0.0418 (0.0315)
	neural	**0.0019 (0.0018)**	**0.0031 (0.0048)**	**0.0035 (0.0087)**
V	*order*	0.0363 (0.045)	0.0395 (0.0474)	0.0489 (0.0599)
	RBF	0.0092 (0.0062)	0.0054 (0.005)	0.0381 (0.0125)
	neural	**0.002 (0.0028)**	**0.0023 (0.0027)**	**0.0039 (0.0041)**
VI	*order*	0.0345 (0.0419)	0.0411 (0.0503)	0.0704 (0.0858)
	RBF	0.0084 (0.0057)	0.0046 (0.0077)	0.021 (0.018)
	neural	**0.0021 (0.0028)**	**0.0022 (0.0029)**	**0.0054 (0.0069)**

of a real demonstrator model designed at the Collaborative Research Centre 805 at TU Darmstadt that may help to control uncertainty in load-carrying structures, by testing different approaches and methods developed in CRC 805. Figure 2 shows the detailed structure of this virtual demonstrator. In this case, a modular active spring damper system is guided on a frame and falls down on the base of the frame. Virtual sensors allow the measurement of different parameters such as acceleration, absolute position of the modular active spring damper system or the forces at the point of impact. In the considered setup the effect of nine normally distributed input variables (parameters of the system, see Table 3) on the computed output variable (the maximum force at the point of impact), was examined, which can be interpreted as a function $m : \mathbb{R}^9 \to \mathbb{R}$ in the sense of Section 1. Especially the quantiles of this force are of interest to the prediction of the stress and its deviation in the later product usage phase in order to define the correct load capacity of the product in the earlier product development phase. The computation of a single output value, during which a differential-algebraic equation systems must be solved by the procedure *RecurDyn* of the software *Siemens NX*, takes about three minutes in this setup. Based on $n = 250$ generated realizations of the nine-dimensional input distribution and the corresponding observed outputs, we applied our neural network quantile estimate as described above, where we increased N_n to 10^6. Table 4 summarizes the results.

Table 3: Parameters of the $\mathcal{N}(\mu, \sigma^2)$-distributions of the input variables

system property	μ	σ
spring stiffness $[N/m]$	27000	1200
damping constant $[N/sm]$	140	7
mass of spring support $[kg]$	20.35	0.25
mass of sphere in lower truss structure $[kg]$	0.76	0.03
mass of sphere in upper truss structure $[kg]$	0.76	0.03
mass of crosslink in upper truss structure $[kg]$	13.74	0.5
mass of threaded rod in truss structure $[kg]$	0.363	0.015
mass of joint middle part $[kg]$	0.9236	0.05
mass of joint arm $[kg]$	1.46	0.075

Table 4: Estimated quantiles for the maximum force at the point of impact

α	*0.5*	*0.9*	*0.95*	*0.99*
quantile estimate $[N]$	38148.75	38507.09	38600.71	38765.67

5 Proofs

5.1 Proof of Theorem 1.

Lemma 1. *Assume that the assumptions of Theorem 1 hold. Then we have outside of an event, whose probability tends to zero for $n \to \infty$,*

$$\mathbf{P}_X \left(\left\{ x \in B_n \, : \, |m_n(x) - m(x)| > 3 \cdot \epsilon_n + 2 \cdot \max_{\substack{i=1,\dots,n, \\ X_i \in B_n}} |m_n(X_i) - m(X_i)| \right\} \right) \leq 2 \cdot \frac{\log\left(\mathcal{N}_{\infty, B_n}\left(\varepsilon_n\right)\right)}{n}.$$

Proof. Let $\mathcal{G}_n$ be an ϵ_n–$\| \cdot \|_{\infty, B_n}$–cover of $\mathcal{F}_n$ of minimal size $\mathcal{N}\left(\epsilon_n, \mathcal{F}_n, \| \cdot \|_{\infty, B_n}\right)$. Choose $\bar{m}_n \in \mathcal{G}_n$ such that

$$\|m_n - \bar{m}_n\|_{\infty, B_n} \leq \epsilon_n.$$

Then

$$\begin{aligned}
&\mathbf{P}_X \left(\left\{ x \in B_n \, : \, |m_n(x) - m(x)| > 3 \cdot \epsilon_n + 2 \cdot \max_{\substack{i=1,\dots,n, \\ X_i \in B_n}} |m_n(X_i) - m(X_i)| \right\} \right) \\
&\leq \mathbf{P}_X \left(\left\{ x \in B_n \, : \, |\bar{m}_n(x) - m(x)| > 2 \cdot \max_{\substack{i=1,\dots,n, \\ X_i \in B_n}} |\bar{m}_n(X_i) - m(X_i)| \right\} \right) \\
&\leq \max_{g \in \mathcal{G}_n} \mathbf{P}_X \left(\left\{ x \in B_n \, : \, |g(x) - m(x)| > 2 \cdot \max_{\substack{i=1,\dots,n, \\ X_i \in B_n}} |g(X_i) - m(X_i)| \right\} \right). \qquad (24)
\end{aligned}$$

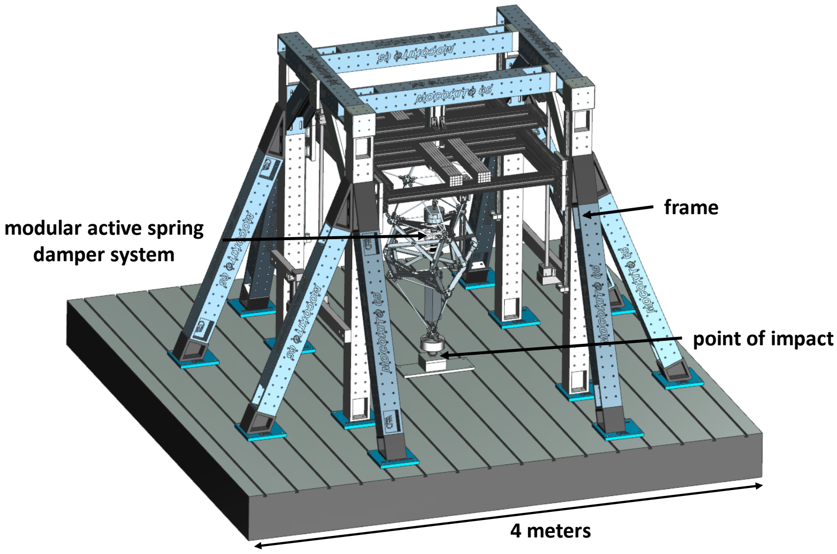

Figure 2: The suspension strut demonstrator

Let $\mathcal{G}_n^* \subseteq \mathcal{G}_n$ consist of all $g \in \mathcal{G}_n$, which satisfy

$$\mathbf{P}_X\left(\{x \in B_n \,:\, |g(x) - m(x)| > 0\}\right) > 2 \cdot \frac{\log\left(\mathcal{N}_{\infty,B_n}\left(\varepsilon_n\right)\right)}{n}. \tag{25}$$

Respecting this definition in combination with (24), it suffices to show

$$\max_{g \in \mathcal{G}_n^*} \mathbf{P}_X\left(\left\{x \in B_n \,:\, |g(x) - m(x)| > 2 \cdot \max_{\substack{i=1,\ldots,n,\\ X_i \in B_n}} |g(X_i) - m(X_i)|\right\}\right) \\ \leq 2 \cdot \frac{\log\left(\mathcal{N}_{\infty,B_n}\left(\varepsilon_n\right)\right)}{n} \tag{26}$$

in order to prove the assertion of the lemma.

Next, we choose a $\lambda_g > 0$ for every $g \in \mathcal{G}_n^*$, which fulfills

$$\mathbf{P}_X\left(\{x \in B_n \,:\, |g(x) - m(x)| > 2 \cdot \lambda_g\}\right) \leq 2 \cdot \frac{\log\left(\mathcal{N}_{\infty,B_n}\left(\varepsilon_n\right)\right)}{n} \\ \leq \mathbf{P}_X\left(\{x \in B_n \,:\, |g(x) - m(x)| > \lambda_g\}\right). \tag{27}$$

A value $\lambda_g > 0$ of this type exists, because the function

$$p : \mathbb{R}_+ \to [0,1], \qquad p(\lambda) = \mathbf{P}_X\left(\{x \in B_n \,:\, |g(x) - m(x)| > \lambda\}\right)$$

is monotonically decreasing and has the properties $\lim_{\lambda \to 0} p(\lambda) > 2 \cdot \frac{\log\left(\mathcal{N}_{\infty,B_n}(\varepsilon_n)\right)}{n}$ (due to (25)) and $\lim_{\lambda \to \infty} p(\lambda) = 0$ for all $g \in \mathcal{G}_n^*$. Thus, if we consider the recursion $\lambda_{g,i+1} =$

$2 \cdot \lambda_{g,i}$ for $i \in \mathbb{N}_0$ starting at a value $\lambda_{g,0} > 0$ that satisfies $p(\lambda_{g,0}) > 2 \cdot \frac{\log\left(\mathcal{N}_{\infty,B_n}(\varepsilon_n)\right)}{n}$, we will find two consecutive values complying with (27). Set

$$D_g = \{x \in B_n \, : \, |g(x) - m(x)| > \lambda_g\} \quad (g \in \mathcal{G}_n^*).$$

Then $X_j \in D_g$ for some $g \in \mathcal{G}_n^*$ and $j \in \{1, \dots, n\}$ implies

$$\max_{\substack{i=1,\dots,n,\\ X_i \in B_n}} |g(X_i) - m(X_i)| \geq |g(X_j) - m(X_j)| > \lambda_g.$$

So we can conclude thanks to (27)

$$\begin{aligned}
&\mathbf{P}_X \left(\left\{ x \in B_n \, : \, |g(x) - m(x)| > 2 \cdot \max_{\substack{i=1,\dots,n,\\ X_i \in B_n}} |g(X_i) - m(X_i)| \right\} \right) \\
&\leq \mathbf{P}_X \left(\{ x \in B_n \, : \, |g(x) - m(x)| > 2 \cdot \lambda_g \} \right) \\
&\leq 2 \cdot \frac{\log\left(\mathcal{N}_{\infty,B_n}\left(\varepsilon_n\right)\right)}{n}.
\end{aligned}$$

Therefore, on the event

$$A_n = \bigcap_{g \in \mathcal{G}_n^*} \{\exists i \in \{1, \dots, n\} : X_i \in D_g\}$$

the condition (26) is valid, and it remains to show

$$\mathbf{P}\left(A_n^c\right) \to 0 \quad (n \to 0).$$

For this purpose, we use the union bound to deduce

$$\begin{aligned}
\mathbf{P}\left(A_n^c\right) &= \mathbf{P}\left(\bigcup_{g \in \mathcal{G}_n^*} \{\forall i \in \{1, \dots, n\} : X_i \notin D_g\} \right) \\
&\leq \sum_{g \in \mathcal{G}_n^*} \mathbf{P}\left\{\forall i \in \{1, \dots, n\} : X_i \notin D_g\right\} \\
&\leq |\mathcal{G}_n| \cdot \max_{g \in \mathcal{G}_n^*} \mathbf{P}\left\{\forall i \in \{1, \dots, n\} : X_i \notin D_g\right\} \\
&= |\mathcal{G}_n| \cdot \max_{g \in \mathcal{G}_n^*} \left(1 - \mathbf{P}_X(D_g)\right)^n \\
&\leq |\mathcal{G}_n| \cdot \left(1 - 2 \cdot \frac{\log\left(\mathcal{N}_{\infty,B_n}\left(\varepsilon_n\right)\right)}{n}\right)^n,
\end{aligned}$$

where the second last line works, because the random variables X_i are independent and identically distributed, and (27) is used to justify the last inequality.
Finally, the well-known relation $1 + x \leq \exp(x)$ $(x \in \mathbb{R})$ allows to bound the last expression by

$$|\mathcal{G}_n| \cdot \left(\exp\left(-2 \cdot \frac{\log\left(\mathcal{N}_{\infty,B_n}\left(\varepsilon_n\right)\right)}{n}\right)\right)^n = \frac{|\mathcal{G}_n|}{\mathcal{N}_{\infty,B_n}\left(\varepsilon_n\right)^2} \leq \frac{1}{\mathcal{N}_{\infty,B_n}\left(\varepsilon_n\right)},$$

which tends to zero for n tending to infinity due to the properties of $\mathcal{N}_{\infty,B_n}(\varepsilon_n)$. $\square$

Proof of Theorem 1. Since

$$\max_{\substack{i=1,\dots,n,\\ X_i\in B_n}} |m_n(X_i)-m(X_i)| = \min_{f\in\mathcal{F}_n} \max_{\substack{i=1,\dots,n,\\ X_i\in B_n}} |f(X_i)-m(X_i)| \le \inf_{f\in\mathcal{F}_n} \|f-m\|_{\infty,B_n},$$

we have

$$\begin{aligned}&\mathbf{P}_X\left(\left\{x\in B_n\,:\,|m_n(x)-m(x)|>3\cdot\varepsilon_n+2\cdot\inf_{f\in\mathcal{F}_n}\|f-m\|_{\infty,B_n}\right\}\right)\\ &\le \mathbf{P}_X\left(\left\{x\in B_n\,:\,|m_n(x)-m(x)|>3\cdot\varepsilon_n+2\cdot\max_{\substack{i=1,\dots,n,\\ X_i\in B_n}}|m_n(X_i)-m(X_i)|\right\}\right).\end{aligned}$$

The application of Lemma 1 to this expression yields the assertion of the theorem. $\square$

5.2 Proof of Theorem 2

Lemma 2. *Let $X, X_1, \dots, X_n, \bar{X}, \bar{X}_1, \dots, \bar{X}_n$ be real valued random variables. Define*

$$\hat{q}_{X,n,\tilde{\alpha}} = \inf\left\{y\in\mathbb{R} : \frac{1}{n}\sum_{i=1}^{n} I_{(-\infty,y]}(X_i)\ge\tilde{\alpha}\right\}$$

for $\tilde{\alpha}\in\mathbb{R}$ and define $\hat{q}_{\bar{X},n,\tilde{\alpha}}$ analogously. Let $a>0$ be a (possibly random) finite constant and set

$$\gamma_n = \frac{1}{n}\sum_{i=1}^{n} I_{\left\{|X_i-\bar{X}_{i,n}|>a\right\}}.$$

Then it holds for $\alpha\in\mathbb{R}$ and the plug-in estimates defined as in (13) that

$$\hat{q}_{X,n,\alpha-\gamma_n}-a \le \hat{q}_{\bar{X},n,\alpha} \le \hat{q}_{X,n,\alpha+\gamma_n}+a.$$

Proof. The result corresponds to Lemma 1 in Hansmann and Kohler (2016). A proof can be found therein. $\square$

Proof of Theorem 2. Set

$$\gamma_n = \frac{1}{N_n}\sum_{i=1}^{N_n} I_{\{|m(X_{n+i})-m_n(X_{n+i})|>\delta_n\}}.$$

Using this auxiliary expression, we will divide the proof into five steps.
Step 1: At first, we show that it suffices to prove

$$\lim_{c\to\infty}\limsup_{n\to\infty}\mathbf{P}\left\{|\hat{q}_{m(X),N_n,\alpha+\gamma_n}-q_{m(X),\alpha}|>c\cdot\left(\mathbf{P}_X\left\{B_n^C\right\}+\zeta_n+\frac{1}{\sqrt{N_n}}\right)\right\}=0. \quad (28)$$

Keeping in mind that Lemma 2 implies

$$\hat{q}_{m_n(X),N_n,\alpha} \in \left[\hat{q}_{m(X),N_n,\alpha-\gamma_n} - \delta_n, \hat{q}_{m(X),N_n,\alpha+\gamma_n} + \delta_n\right], \tag{29}$$

we can conclude

$$\begin{aligned}
&\lim_{c\to\infty} \limsup_{n\to\infty} \mathbf{P}\left\{|\hat{q}_{m_n(X),N_n,\alpha} - q_{m(X),\alpha}| > c \cdot \left(\mathbf{P}_X\left\{B_n^C\right\} + \zeta_n + \frac{1}{\sqrt{N_n}} + \delta_n\right)\right\} \\
&\leq \lim_{c\to\infty} \limsup_{n\to\infty} \mathbf{P}\Bigg\{ \max\left\{|\hat{q}_{m(X),N_n,\alpha+\gamma_n} + \delta_n - q_{m(X),\alpha}|, |\hat{q}_{m(X),N_n,\alpha-\gamma_n} - \delta_n - q_{m(X),\alpha}|\right\} \\
&\qquad\qquad > c \cdot \left(\mathbf{P}_X\left\{B_n^C\right\} + \zeta_n + \frac{1}{\sqrt{N_n}} + \delta_n\right)\Bigg\} \\
&\leq \lim_{c\to\infty} \limsup_{n\to\infty} \mathbf{P}\Bigg\{ \max\left\{|\hat{q}_{m(X),N_n,\alpha+\gamma_n} - q_{m(X),\alpha}|, |\hat{q}_{m(X),N_n,\alpha-\gamma_n} - q_{m(X),\alpha}|\right\} \\
&\qquad\qquad > c \cdot \left(\mathbf{P}_X\left\{B_n^C\right\} + \zeta_n + \frac{1}{\sqrt{N_n}}\right)\Bigg\}, \\
&\leq \lim_{c\to\infty} \limsup_{n\to\infty} \mathbf{P}\left\{|\hat{q}_{m(X),N_n,\alpha+\gamma_n} - q_{m(X),\alpha}| > c \cdot \left(\mathbf{P}_X\left\{B_n^C\right\} + \zeta_n + \frac{1}{\sqrt{N_n}}\right)\right\} \\
&\quad + \lim_{c\to\infty} \limsup_{n\to\infty} \mathbf{P}\left\{|\hat{q}_{m(X),N_n,\alpha-\gamma_n} - q_{m(X),\alpha}| > c \cdot \left(\mathbf{P}_X\left\{B_n^C\right\} + \zeta_n + \frac{1}{\sqrt{N_n}}\right)\right\}.
\end{aligned}$$

Since all the following steps work completely analogously for the second summand in this expression, it suffices to show (28).

Step 2: Now we use the triangle inequality

$$|\hat{q}_{m(X),N_n,\alpha+\gamma_n} - q_{m(X),\alpha}| \leq |\hat{q}_{m(X),N_n,\alpha+\gamma_n} - q_{m(X),\alpha+\gamma_n}| + |q_{m(X),\alpha+\gamma_n} - q_{m(X),\alpha}|,$$

and bound (28) from above by

$$\begin{aligned}
&\lim_{c\to\infty} \limsup_{n\to\infty} \mathbf{P}\left\{|\hat{q}_{m(X),N_n,\alpha+\gamma_n} - q_{m(X),\alpha}| > c \cdot \left(\mathbf{P}_X\left\{B_n^C\right\} + \zeta_n + \frac{1}{\sqrt{N_n}}\right)\right\} \\
&\leq \lim_{c\to\infty} \limsup_{n\to\infty} \mathbf{P}\left\{|q_{m(X),\alpha+\gamma_n} - q_{m(X),\alpha}| > c \cdot \left(\mathbf{P}_X\left\{B_n^C\right\} + \zeta_n + \frac{1}{2} \cdot \frac{1}{\sqrt{N_n}}\right)\right\} \\
&\quad + \lim_{c\to\infty} \limsup_{n\to\infty} \mathbf{P}\left\{|\hat{q}_{m(X),N_n,\alpha+\gamma_n} - q_{m(X),\alpha+\gamma_n}| > c \cdot \frac{1}{2} \cdot \frac{1}{\sqrt{N_n}}\right\},
\end{aligned} \tag{30}$$

so that we can show separately that both of these summands are equal to zero, which will be done in the fourth and fifth step.

Step 3: Next, we will show that

$$\lim_{c\to\infty} \limsup_{n\to\infty} \mathbf{P}\left\{c_3 \cdot \gamma_n > c \cdot \left(\mathbf{P}_X\left\{B_n^C\right\} + \zeta_n + \frac{1}{2} \cdot \frac{1}{\sqrt{N_n}}\right)\right\} = 0 \tag{31}$$

holds for arbitrary $c_3 > 0$. For this purpose, we define $A_n = \{x \in B_n : |m(x) - m_n(x)| > \delta_n\}$. From

$$\left\{x \in \mathbb{R}^d : |m(x) - m_n(x)| > \delta_n\right\} = \left\{x \in \mathbb{R}^d : |m(x) - m_n(x)| \leq \delta_n\right\}^C$$
$$\subseteq \{x \in B_n : |m(x) - m_n(x)| \leq \delta_n\}^C = (B_n \setminus A_n)^C$$

and (14) we can conclude

$$\lim_{c\to\infty} \limsup_{n\to\infty} \mathbf{P}\left\{c_3 \cdot \gamma_n > c \cdot \left(\mathbf{P}_X\left\{B_n^C\right\} + \zeta_n + \frac{1}{2} \cdot \frac{1}{\sqrt{N_n}}\right)\right\}$$
$$\leq \lim_{c\to\infty} \limsup_{n\to\infty} \mathbf{P}\left\{\gamma_n > \frac{c}{c_3} \cdot \left(\mathbf{P}_X\left\{B_n^C\right\} + \mathbf{P}_X\left\{A_n\right\} + \frac{1}{2} \cdot \frac{1}{\sqrt{N_n}}\right)\right\}$$
$$\leq \lim_{c\to\infty} \limsup_{n\to\infty} \mathbf{P}\left\{\gamma_n > \mathbf{P}_X\left\{B_n^C\right\} + \mathbf{P}_X\left\{A_n\right\} + c \cdot \frac{1}{2} \cdot \frac{1}{\sqrt{N_n}}\right\}$$
$$\leq \lim_{c\to\infty} \limsup_{n\to\infty} \mathbf{P}\left\{\frac{1}{N_n}\sum_{i=1}^{N_n} I_{(B_n\setminus A_n)^C}\left(X_i\right) > \mathbf{P}_X\left\{B_n^C\right\} + \mathbf{P}_X\left\{A_n\right\} + c \cdot \frac{1}{2} \cdot \frac{1}{\sqrt{N_n}}\right\}.$$

Using Hoeffding's inequality we can bound the probability in the last expression by

$$\leq \mathbf{P}\left\{\frac{1}{N_n}\sum_{i=1}^{N_n} I_{(B_n\setminus A_n)^C}\left(X_i\right) - \mathbf{P}_X\left\{(B_n \setminus A_n)^C\right\} > c \cdot \frac{1}{2} \cdot \frac{1}{\sqrt{N_n}}\right\}$$
$$\leq \exp\left(-N_n \cdot \frac{c^2}{2 \cdot N_n}\right) = \exp\left(-c^2/2\right).$$

This tends to zero for c tending to infinity, which proves (31).
Step 4: In this step we show that the first summand of the right-hand side in (30) is zero. Respecting (31), we observe

$$\lim_{c\to\infty} \limsup_{n\to\infty} \mathbf{P}\left\{|q_{m(X),\alpha+\gamma_n} - q_{m(X),\alpha}| > c \cdot \left(\mathbf{P}_X\left\{B_n^C\right\} + \zeta_n + \frac{1}{2} \cdot \frac{1}{\sqrt{N_n}}\right)\right\}$$
$$\leq \lim_{c\to\infty} \limsup_{n\to\infty} \mathbf{P}\left\{|q_{m(X),\alpha+\gamma_n} - q_{m(X),\alpha}| > c \cdot \left(\mathbf{P}_X\left\{B_n^C\right\} + \zeta_n + \frac{1}{2} \cdot \frac{1}{\sqrt{N_n}}\right),\right.$$
$$\left.\gamma_n \leq c \cdot \left(\mathbf{P}_X\left\{B_n^C\right\} + \zeta_n + \frac{1}{2} \cdot \frac{1}{\sqrt{N_n}}\right)\right\}. \tag{32}$$

Due to $\left\{y \in \mathbb{R} : G_{m(X)}(y) \geq \alpha\right\} \supseteq \left\{y \in \mathbb{R} : G_{m(X)}(y) \geq \alpha + \gamma_n\right\}$ the definition of quantiles in (11) implies $q_{m(X),\alpha} \leq q_{m(X),\alpha+\gamma_n}$. Since $m(X)$ has a density $g_{m(X)}$, $G_{m(X)}$ is continous and $G_{m(X)}\left(q_{m(X),\alpha}\right) = \alpha$ and $G_{m(X)}\left(q_{m(X),\alpha+\gamma_n}\right) = \alpha + \gamma_n$ hold. So we can conclude

$$\gamma_n = (\alpha + \gamma_n) - \alpha = G_{m(X)}\left(q_{m(X),\alpha+\gamma_n}\right) - G_{m(X)}\left(q_{m(X),\alpha}\right) = \int_{q_{m(X),\alpha}}^{q_{m(X),\alpha+\gamma_n}} g_{m(X)}(y)\mathrm{d}y.$$

Since $g_{m(X)}$ is continuous on $\mathbb{R}$ and positive at $q_{m(X),\alpha}$ this implies that in case of $\gamma_n \to 0$ $(n \to \infty)$ we also have $q_{m(X),\alpha+\gamma_n} \to q_{m(X),\alpha}$ $(n \to \infty)$. Hence, if γ_n is sufficiently small, which is guaranteed by the second inequality in (32) for increasing n, $g_{m(X)} \geq \frac{1}{c_3}$ holds on $\left[q_{m(X),\alpha}, q_{m(X),\alpha+\gamma_n}\right]$ for a certain $c_3 > 0$ thanks to the continuity of the density. This leads to

$$\int_{q_{m(X),\alpha}}^{q_{m(X),\alpha+\gamma_n}} g_{m(X)}(y)\mathrm{d}y \geq \frac{1}{c_3} \cdot \left(q_{m(X),\alpha+\gamma_n} - q_{m(X),\alpha}\right),$$

which implies $|q_{m(X),\alpha+\gamma_n} - q_{m(X),\alpha}| \leq c_3 \cdot \gamma_n$, so that (32) can be bounded from above by

$$\lim_{c\to\infty} \limsup_{n\to\infty} \mathbf{P}\left\{c_3 \cdot \gamma_n > c \cdot \left(\mathbf{P}_X\left\{B_n^C\right\} + \zeta_n + \frac{1}{2} \cdot \frac{1}{\sqrt{N_n}}\right)\right\},$$

which is zero due to (31).

Step 5: In order to prove that also the second summand in (30) is zero, we consider the complementary event of the event therein. This leads to

$$\begin{aligned}
&\mathbf{P}\left\{|\hat{q}_{m(X),N_n,\alpha+\gamma_n} - q_{m(X),\alpha+\gamma_n}| \leq \frac{c}{2\sqrt{N_n}}\right\} \\
&= \mathbf{P}\left\{q_{m(X),\alpha+\gamma_n} - \frac{c}{2\sqrt{N_n}} \leq \hat{q}_{m(X),N_n,\alpha+\gamma_n} \leq q_{m(X),\alpha+\gamma_n} + \frac{c}{2\sqrt{N_n}}\right\} \\
&\geq \mathbf{P}\left\{\hat{G}_{m(X),N_n}\left(q_{m(X),\alpha+\gamma_n} - \frac{c}{2\sqrt{N_n}}\right) < \alpha + \gamma_n < \hat{G}_{m(X),N_n}\left(q_{m(X),\alpha+\gamma_n} + \frac{c}{2\sqrt{N_n}}\right)\right\}.
\end{aligned}$$

Arguing in the same way as in (32) we know that

$$\inf_{x\in\left[q_{m(X),\alpha+\gamma_n} - \frac{c}{2\sqrt{N_n}}, q_{m(X),\alpha+\gamma_n} + \frac{c}{2\sqrt{N_n}}\right]} g_{m(X)}(x) \geq c_4$$

holds for n sufficiently large, because γ_n becomes sufficiently small. Then

$$\begin{aligned}
G_{m(X)}\left(q_{m(X),\alpha+\gamma_n} - \frac{c}{2\sqrt{N_n}}\right) &\leq \alpha + \gamma_n - c_4 \cdot \frac{c}{2\sqrt{N_n}} \\
\leq \alpha + \gamma_n + c_4 \cdot \frac{c}{2\sqrt{N_n}} &\leq G_{m(X)}\left(q_{m(X),\alpha+\gamma_n} + \frac{c}{2\sqrt{N_n}}\right)
\end{aligned}$$

is valid. Since the probability above contains $\hat{G}_{m(X),N_n}$ instead of $G_{m(X)}$, we have to show

$$\lim_{c\to\infty} \limsup_{n\to\infty} \mathbf{P}\left\{\sup_{y\in\mathbb{R}}\left|G_{m(X)}(y) - \hat{G}_{m(X),N_n}(y)\right| \leq c_5 \cdot \frac{c}{2\sqrt{N_n}}\right\} = 1$$

for a $c_5 < c_4$, in order to prove that the second expression in (30) is zero as well. This follows immediately from the Dvoretzky-Kiefer-Wolfowitz inequality (cf., e.g., Massart (1990)). So the assertion of the theorem holds. □

5.3 Proof of Theorem 3

In the proof we will use the following auxiliary results.

Lemma 3. *Let $m : \mathbb{R}^d \to \mathbb{R}$ be a (p, C)–smooth function, where $0 < p \leq 1$, let $M \in \mathbb{N}$, let $a > 0$, let $\eta \in (0, 1]$ and let ν be a probability measure on $\mathbb{R}^d$. Let $\sigma : \mathbb{R} \to [0, 1]$ be a squashing function. Then there exists a neural network*

$$t(x) = \sum_{i=1}^{(M+1)^d} d_i \cdot \sigma \left(\sum_{j=1}^{d} b_{i,j} \cdot \sigma \left(\sum_{k=1}^{d} a_{i,j,k} \cdot x^{(k)} + a_{i,j,0} \right) + b_{i,0} \right) + d_0$$

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

$$|t(x) - m(x)| \leq c_6 \cdot \frac{a^p}{M^p}$$

with a constant $c_6 > 0$ independent of a and M. In case that σ satisfies

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

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

$$|d_i| \leq 2^d \cdot \|m\|_\infty, \quad |b_{i,j}| \leq 4 \cdot d \cdot (M+1)^d \quad \textit{and} \quad |a_{i,j,k}| \leq 8 \cdot d^2 \cdot \max\left\{3, \frac{1}{a}\right\} \cdot \frac{M}{\eta}$$

hold for all indices in t.

Proof. The result can be proven similar to Lemma 6 in Kohler and Krzyżak (2016) by modifying the proof of Theorem 3.4 in Mhaskar (1993). A complete proof of this result is available from the authors on request. □

From Lemma 3 we conclude

Lemma 4. *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 $p \in (0, 1]$. Respecting Definition 1 b), let all functions $g_k, f_{j,k}$ be bounded and let all functions g_k be Lipschitz continuous with Lipschitz constant $L > 0$. Let $M_n \in \mathbb{N}$, let $a_n > 0$ be increasing, such that $a_n \leq M_n$ for large n, and let $\eta_n \in (0, 1]$. Let $\mathcal{H}^{(l)}$ be defined as in (20) with K, d, d^* as in the definition of m, $M = M_n$, $\alpha = \log(n) \cdot \frac{1}{\eta_n} \cdot M_n$, $\beta = \log(n) \cdot M_n^{d^*}$, $\gamma = \log(n)$. Then for arbitrary $c > 0$ and all n greater than a certain $n_0(c) \in \mathbb{N}$, there exists a $t(x) \in \mathcal{H}^{(l)}$ such that outside of a set of $\mathbf{P}_X$-measure less than or equal to $c \cdot \eta_n$ we have*

$$|t(x) - m(x)| \leq c_7 \cdot \frac{a_n^p}{{M_n}^p}$$

for all $x \in [-a_n, a_n]^d$ and $c_7 > 0$ independent of a_n and M_n (but depending on c).

Proof. We will proof 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 Lemma 3 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 $x \in [-d \cdot \bar{a}_{\max} \cdot a_n, d \cdot \bar{a}_{\max} \cdot a_n]^{d^*}$ with an error of

$$\left|\hat{f}(x) - f(x)\right| \leq c_6 \cdot \frac{(d \cdot \bar{a}_{\max} \cdot a_n)^p}{{M_n}^p} = c_7 \cdot \frac{a_n^p}{{M_n}^p}$$

except for a set $\tilde{D}_0$ of $\mathbf{P}_{s(X)}$–measure less than or equal to $c \cdot \eta_n > 0$. If we plug $s(x)$ into that approximation and condense the inner coefficients per summand, this leads (using the notation of Lemma 3) to the approximant $t(x) = \hat{f}(s(x))$ of the form

$$\begin{aligned}
t(x) &= \sum_{i=1}^{M_n^{d^*}} d_i \cdot \sigma\left(\sum_{j=1}^{d^*} b_{i,j} \cdot \sigma\left(\sum_{k=1}^{d^*} a_{i,j,k} \cdot a_k^T x + a_{i,j,0}\right) + b_{i,0}\right) + d_0 \\
&= \sum_{i=1}^{M_n^{d^*}} d_i \cdot \sigma\left(\sum_{j=1}^{d^*} b_{i,j} \cdot \sigma\left(\sum_{k=1}^{d^*} a_{i,j,k} \cdot \sum_{m=1}^{d} a_k^{(m)} \cdot x^{(m)} + a_{i,j,0}\right) + b_{i,0}\right) + d_0 \\
&= \sum_{i=1}^{M_n^{d^*}} d_i \cdot \sigma\left(\sum_{j=1}^{d^*} b_{i,j} \cdot \sigma\left(\sum_{m=1}^{d} \left(\sum_{k=1}^{d^*} a_{i,j,k} \cdot a_k^{(m)}\right) \cdot x^{(m)} + a_{i,j,0}\right) + b_{i,0}\right) + d_0 \\
&=: \sum_{i=1}^{M_n^{d^*}} d_i \cdot \sigma\left(\sum_{j=1}^{d^*} b_{i,j} \cdot \sigma\left(\sum_{m=1}^{d} \tilde{a}_{i,j,m} \cdot x^{(m)} + \tilde{a}_{i,j,0}\right) + b_{i,0}\right) + d_0,
\end{aligned}$$

where

$$|d_i| \leq 2^{d^*} \cdot \|f\|_\infty \leq \gamma, \qquad |b_{i,j}| \leq 4 \cdot d^* \cdot (M_n + 1)^{d^*} \leq \beta,$$

and

$$|\tilde{a}_{i,j,m}| \leq d^* \cdot \bar{a}_{\max} \cdot \max_{k=1,\dots,d^*} |a_{i,j,k}| \leq 8 \cdot (d^*)^3 \cdot \bar{a}_{\max} \cdot \max\left\{3, \frac{1}{d \cdot \bar{a}_{\max} \cdot a_n}\right\} \cdot \frac{M_n}{c \cdot \eta_n} \leq \alpha$$

are satisfied for n sufficiently large, 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 $s\left([-a_n, a_n]^d\right) \subseteq [-d \cdot \bar{a}_{\max} \cdot a_n, d \cdot \bar{a}_{\max} \cdot a_n]^{d^*}$,

$$|t(x) - m(x)| \leq c_7 \cdot \frac{a_n^p}{{M_n}^p}$$

holds for all $x \in [-a_n, a_n]^d$ outside of a 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 assertion for $l = 0$.

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

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

Since 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 lemma, we can choose the approximations $\hat{f}_{j,k} \in \mathcal{H}^{(l-1)}$ according to the induction hypothesis with $\frac{c}{2 \cdot d^* \cdot K} \cdot \eta_n$. Then each of the terms $|f_{j,k}(x) - \hat{f}_{j,k}(x)|$ can be bounded by $c_8 \cdot \frac{a_n^p}{{M_n}^p}$ for all n sufficiently large and $x \in \left[-a_n, a_n\right]^d$ 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 $\hat{f}_k(x) = \left(\hat{f}_{1,k}(x),\dots,\hat{f}_{d^*,k}(x)\right)^T$ and let $\bar{f}_{k,\max}$ denote $\max_{j=1,\dots,d^*} \|f_{j,k}\|_\infty$ for all $k = 1,\dots,K$. Since $\frac{a_n}{M_n} \leq 1$ thanks to the assumptions of this lemma, $\hat{f}_k(x)$ falls into

$$\hat{F}_k = \left[-\bar{f}_{k,\max} - c_8, \bar{f}_{k,\max} + c_8\right]^{d^*}$$

for all $x \in \left[-a_n, a_n\right]^d$ outside of the union of the sets $D_{j,k}$ $(j = 1,\dots,d^*,\ k = 1,\dots,K)$ and n sufficiently large. Applying Lemma 3 with $\frac{c \cdot \eta_n}{2 \cdot K}$, 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 $c_6 \cdot \left(\bar{f}_{k,\max} + c_8\right)^p / {M_n}^p \leq c_9 \cdot \frac{1}{M_n^p}$ on $\hat{F}_k$ outside of a set $\tilde{D}_k$ with a probability measure $\mathbf{P}_{\hat{f}_k(X)}$ less than or equal to $\frac{\eta_n}{2 \cdot K}$. For n sufficiently large, the weights according to the notation of Lemma 3 satisfy

$$|d_i| \leq 2^{d^*} \cdot \|g_k\|_\infty \leq \gamma, \qquad |b_{i,j}| \leq 4 \cdot d^* \cdot \left(M_n + 1\right)^{d^*} \leq \beta$$

and

$$|a_{i,j,k}| \leq 8 \cdot {d^*}^2 \cdot \max\left\{3, \frac{1}{\bar{f}_{k,\max} + c_8}\right\} \cdot \frac{M_n}{\frac{c \cdot \eta_n}{2K}} \leq \alpha,$$

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

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

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 $\hat{g}_k \in \mathcal{F}^{(neural\ networks)}_{M_n,d^*,d^*,\alpha,\beta,\gamma}$ 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

$$|t(x) - m(x)| \leq K \cdot L \cdot d^* \cdot c_8 \cdot \frac{a_n^p}{M_n^p} + K \cdot c_9 \cdot \frac{1}{M_n^p} \leq c_7 \cdot \frac{a_n^p}{M_n^p}$$

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

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

which proves the assertion of the lemma. □

Lemma 5. *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*

$$|\sigma_r(x)| \leq L \cdot \max\{|x|, 1\} \quad (x \in \mathbb{R}). \tag{33}$$

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

$$\left|f_1^{(l+1)}(x) - \bar{f}_1^{(l+1)}(x)\right|$$

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

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

Proof. At first, we notice that (33) implies

$$\left| \bar{f}_i^{(r)}(x) \right| \leq L \cdot (K_{r-1}+1) \cdot \overline{C} \cdot \max_{j=1,\dots,K_{r-1}} \left\{ \left| \bar{f}_j^{(r-1)}(x) \right|, 1 \right\}$$

for $r = 1, \dots, l$ and $i = 1, \dots, K_r$, from which we can conclude

$$\left| \bar{f}_i^{(r)}(x) \right| \leq L^r \cdot \prod_{\tilde{r}=1}^{r} (K_{\tilde{r}-1}+1) \cdot \overline{C}^r \cdot \max\{\|x\|_\infty, 1\}. \tag{34}$$

Using the Lipschitz continuity of σ_r and the triangle inequality in combination with (34) we get

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

for all $r = 1, \dots, l+1$. Now we start with the above inequality for $r = l+1$ and plug it in repeatedly for decreasing r in the expression $\left| f_j^{(r-1)}(x) - \bar{f}_j^{(r-1)}(x) \right|$ on the right-hand side of the inequality. Finally, the summand containing $\left| f_j^{(0)}(x) - \bar{f}_j^{(0)}(x) \right|$ vanishes in the case of $r = 1$, which implies the assertion. □

From Lemma 5 we conclude

Lemma 6. *Assume that the assumptions of Theorem 3 hold and $\varepsilon_n = \frac{a_n^p}{M_n^p}$ for an $a_n > 0$ which satisfies $a_n \leq M_n$ for large n. Then*

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

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

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

$$\begin{aligned}\mathcal{H}^{(l)} \quad = \quad & \Bigg\{ h : \mathbb{R}^d \to \mathbb{R} \ : \ h(x) = \sum_{k=1}^{K} \sigma_{id}\left(g_k\left(\sigma_{id}\left(f_{1,k}(x)\right), \ldots, \sigma_{id}\left(f_{d^*,k}(x)\right)\right)\right) \quad (x \in \mathbb{R}^d) \\ & \qquad\qquad\qquad \text{for some } g_k \in \mathcal{F}^{(neural\ networks)}_{M_n,d^*,d^*,\alpha,\beta,\gamma} \text{ and } f_{j,k} \in \mathcal{H}^{(l-1)} \Bigg\},\end{aligned}$$

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

$$\begin{aligned} g(x) &= \sum_{i=1}^{(M_n+1)^{d^*}} d_i \cdot \sigma\left(\sum_{j=1}^{d^*} b_{i,j} \cdot \sigma\left(\sum_{m=1}^{d^*} a_{i,j,m} \cdot x^{(m)} + a_{i,j,0}\right) + b_{i,0}\right) + d_0 \\ &= \sum_{i=1}^{(M_n+1)^{d^*}} d_i \cdot \sigma\left(\sum_{\substack{j=1,\ldots,d^*,\\ \bar{i}=1,\ldots,M_n^{d^*}}} b_{i,\bar{i},j} \cdot \sigma\left(\sum_{m=1}^{d^*} a_{\bar{i},j,m} \cdot x^{(m)} + a_{\bar{i},j,0}\right) + b_{i,\bar{i},0}\right) + d_0, \end{aligned}$$

where the new coefficients are defined by

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

for all $i, \bar{i} \in \left\{1, \ldots, (M_n+1)^{d^*}\right\}$ and $j \in \{1, \ldots, d^*\}$ (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 5, 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 the squashing function σ from Theorem 3. Thus, the property (33) is satisfied due to $\|\sigma\|_\infty \leq 1$, $L \geq 1$, and $|\sigma_{id}(x)| = |x|$. The parameter l in Lemma 5 is $4l + 2$ (regarding the l in $\mathcal{H}^{(l)}$ above) and the parameters K_r with $r = 0, \ldots, l$ take repeatedly the values $\tilde{d}, d^* \cdot (M_n+1)^{d^*}, (M_n+1)^{d^*}, K$ 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 $r = 0, \ldots, l$, $i = 1, \ldots, K_{r+1}$, $j = 1, \ldots, K_r$ (using $K_{l+1} = 1$ again) are 0, 1, or one of the $a_{i,j,m}, b_{i,j}, d_i$ in the definition of $\mathcal{F}^{(neural\ networks)}_{M_n,d^*,d,\alpha,\beta,\gamma}$, we can use $\overline{C} = \max\{\alpha, \beta, \gamma\}$ for n sufficiently large.

Let h and $\bar{h}$ be functions in $\mathcal{F}_n$. Since they comply with the structure of the functions in Lemma 5 according to the above argumentation, we can conclude

$$\begin{aligned}
&\|h-\bar{h}\|_{\infty,[-a_n,a_n]^d} \\
&\leq (4l+3)\cdot L^{4l+3}\cdot\left(d^*\cdot(M_n+1)^{d*}+1\right)^{4l+3}\cdot\max\{\alpha,\beta,\gamma\}^{4l+2} \\
&\qquad\qquad\qquad\cdot\max\{a_n,1\}\cdot\max_{\substack{r=0,\dots,\tilde{l},\ i=1,\dots,K_{r+1},\\ j=0,\dots,K_r}}\left|c_{i,j}^{(r)}-\bar{c}_{i,j}^{(r)}\right| \\
&\leq a_n\cdot n^{c_{11}}\cdot\max_{\substack{r=0,\dots,\tilde{l},\ i=1,\dots,K_{r+1},\\ j=0,\dots,K_r}}\left|c_{i,j}^{(r)}-\bar{c}_{i,j}^{(r)}\right|
\end{aligned}$$

for n sufficiently large and an adequately chosen $c_{11}>0$. 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

$$\left|c_{i,j}^{(r)}-\bar{c}_{i,j}^{(r)}\right|\leq\frac{\varepsilon_n}{a_n\cdot n^{c_{11}}} \tag{35}$$

holds for all possible indices, in order to satisfy $\|h(x)-\bar{h}(x)\|_{\infty,[-a_n,a_n]^d}\leq\varepsilon_n$. For n sufficiently large, which is assumed permanently in the following, the coefficients $c_{i,j}^{(r)}$ have to take values in $[-\max\{\alpha,\beta,\gamma\},\max\{\alpha,\beta,\gamma\}]$ and the relations $\max\{\alpha,\beta,\gamma\}\leq\log(n)\cdot n^2\cdot M_n^{d^*}\leq n^4$ and $a_n\leq M_n\leq n$ hold. Then due to $\varepsilon_n=\frac{a_n^p}{M_n^p}$ a number of

$$\left\lceil\frac{2\cdot\max\{\alpha,\beta,\gamma\}\cdot a_n\cdot n^{c_{11}}}{2\cdot\varepsilon_n}\right\rceil\leq n^{c_{12}}$$

different $\bar{c}_{i,j}^{(r)}$ suffices to guarantee, that at least one of them satisfies the relation (35) 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 $a_{i,j,m},b_{i,j},d_i$ in the definition of $\mathcal{F}_{M_n,d^*,d,\alpha,\beta,\gamma}^{(neural\ networks)}$. Using (22), their number can be bounded by $c_{13}\cdot M_n^{d^*}$. So the logarithm of the covering number $\mathcal{N}(\varepsilon_n,\mathcal{F}_n,\|\cdot\|_{\infty,[-a_n,a_n]^d})$ can be bounded by

$$\log\left(\mathcal{N}(\varepsilon_n,\mathcal{F}_n,\|\cdot\|_{\infty,[-a_n,a_n]^d})\right)\leq\log\left((n^{c_{12}})^{c_{13}\cdot M_n^{d^*}}\right)\leq c_{10}\cdot\log(n)\cdot M_n^{d^*},$$

which proves the assertion. □

Proof of Theorem 3. Choose the maximum $\mathbf{P}_X$-measure of the exceptional set in Lemma 4 (called D_n in the following) as $\eta_n=\frac{1}{n^2}$, set $a_n=\log(n)^2$, $c_1=2\cdot c_7+3$ and $\varepsilon_n=\frac{a_n^p}{M_n^p}$. Then Lemma 1, Lemma 6, the union bound and Markov's inequality imply

$$\begin{aligned}
&\mathbf{P}_X\left(\left\{x\in[-a_n,a_n]^d\ :\ |m_n(x)-m(x)|>c_1\cdot\log(n)^{2p+\frac{p}{p+d^*}}\cdot n^{-\frac{p}{p+d^*}}\right\}\right) \\
&\leq\mathbf{P}_X\left(\left\{x\in[-a_n,a_n]^d\ :\ |m_n(x)-m(x)|>c_1\cdot\varepsilon_n\right\}\right)
\end{aligned}$$

$$
\begin{aligned}
&\leq \mathbf{P}_X\left(\left\{x \in [-a_n, a_n]^d \,:\, |m_n(x) - m(x)| > 3 \cdot \varepsilon_n + 2 \cdot \max_{\substack{i=1,\dots,n,\\ X_i \in B_n}} |m_n(X_i) - m(X_i)|\right\}\right) \\
&\quad + \mathbf{P}\left\{\exists i \in \{1, \dots, n\} : X_i \in D_n\right\} + \mathbf{P}\left\{\exists i \in \{1, \dots, n\} : X_i \notin [-a_n, a_n]^d\right\} \\
&\leq c_2 \cdot \log(n)^{\frac{p}{p+d^*}} \cdot n^{-\frac{p}{p+d^*}} + n \cdot \frac{1}{n^2} + n \cdot \frac{\mathbf{E}\left\{\exp\left(\bar{c} \cdot \|X\|_\infty\right)\right\}}{n^{\bar{c} \cdot \log(n)}}.
\end{aligned}
$$

The proof is complete. □

6 Acknowledgment

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

References

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

[2] Barron, A. R. (1991). Complexity regularization with application to artificial neural networks. In G. Roussas (ed.), *Nonparametric Functional Estimation and Related Topics*, pp. 5621–576, NATO ASI Series, Kluwer Academic Publishers, Dordrecht, Netherlands.

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

[4] Beirlant, J. and Györfi, L. (1998). On the asymptotic L_2-error in partitioning regression estimation. *Journal of Statistical Planning and Inference*, **71**, pp. 93–107.

[5] Bichon, B. J., Eldred, M. S., Swiler, L. P., Mahadevan, S. and McFarland, J. M. (2008). Efficient Global Reliability Analysis for Nonlinear Implicit Performance Functions. *AIAA Journal*, **46**, pp. 2459–2468.

[6] Bourinet, J.-M., Deheeger, F. and Lemaire, M. (2011). Assessing small failure probabilities by combined subset simulation and support vector machines. *Structural Safety*, **33**, pp. 343–353.

[7] Bucher, C. and Bourgund, U. (1990). A fast and efficient response surface approach for structural reliability problems. *Structural Safety*, **7**, pp. 57-66.

[8] Das, P.-K. and Zheng, Y. (2000). Cumulative formation of response surface and its use in reliability analysis. *Probabilistic Engineering Mechanics*, **15**, pp. 309-315.

[9] Deheeger, F. and Lemaire, M. (2010). Support vector machines for efficient subset simulations: ^{2}SMART method. In: *Proceedings of the 10th International Conference on Applications of Statistics and Probability in Civil Engineering (ICASP10)*, Tokyo, Japan.

[10] Devroye, L. (1982). Necessary and sufficient conditions for the almost everywhere convergence of nearest neighbor regression function estimates. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, **61**, pp. 467–481.

[11] Devroye, L., Györfi, L., Krzyżak, A., and Lugosi, G. (1994). On the strong universal consistency of nearest neighbor regression function estimates. *Annals of Statistics*, **22**, pp. 1371–1385.

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

[13] Devroye, L. and Krzyżak, A. (1989). An equivalence theorem for L_1 convergence of the kernel regression estimate. *Journal of Statistical Planning and Inference*, **23**, pp. 71–82.

[14] 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.

[15] Dubourg, V., Sudret, B. and Deheeger, F. (2013). Metamodel-based importance sampling for structural reliability analysis. *Probabilistic Engineering Mechanics*, **33**, pp. 47–57.

[16] Enss, G., Kohler, M., Krzyżak, A., and Platz, R. (2016). Nonparametric quantile estimation based on surrogate models. *IEEE Transactions on Information Theory*, **62**, pp. 5727–5739.

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

[18] Greblicki, W. and Pawlak, M. (1985). Fourier and Hermite series estimates of regression functions. *Annals of the Institute of Statistical Mathematics*, **37**, pp. 443-454.

[19] Györfi, L. (1981). Recent results on nonparametric regression estimate and multiple classification. *Problems of Control and Information Theory*, **10**, pp. 43–52.

[20] 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.

[21] Hansmann, M. and Kohler, M. (2016). Estimation of quantiles from data with additional measurement errors. To appear in *Statistica Sinica* (2017).

[22] Hastie, T., Tibshirani, R., and Friedman, J. (2011). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed., Springer, New York.

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

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

[25] Hurtado, J. (2004). *Structural reliability – Statistical learning perspectives.* Vol. 17 of lecture notes in applied and computational mechanics. Springer.

[26] Kaymaz, I. (2005). Application of Kriging method to structural reliability problems. *Strutural Safety*, **27**, pp. 133–151.

[27] Kim, S.-H. and Na, S.-W. (1997). Response surface method using vector projected sampling points. *Structural Safety*, **19**, pp. 3–19.

[28] Kohler, M. (2000). Inequalities for uniform deviations of averages from expectations with applications to nonparametric regression. *Journal of Statistical Planning and Inference*, **89**, pp. 1–23.

[29] 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.

[30] Kohler, M. and Krzyżak, A. (2001). Nonparametric regression estimation using penalized least squares. *IEEE Transactions on Information Theory*, **47**, pp. 3054–3058.

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

[32] Kohler, M., and Krzyżak, A. (2013). Optimal global rates of convergence for interpolation problemswith random design. *Statistics and Probability Letters*, **83**, pp. 1871–1879.

[33] Kohler, M., and Krzyżak, A. (2016). Nonparametric regression based on hierarchical interaction models. To appear in *IEEE Transaction on Information Theory* (2017).

[34] Kohler, M., and Tent, R. (2015). Nonparametric Quantile Estimation Using Surrogate Models and Importance Sampling. Submitted for publication.

[35] 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, 521-536.

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

[37] Massart, P. (1990). The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality. *Annals of Probability* **18**, pp. 1269–1283.

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

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

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

[41] Nadaraya, E. A. (1964). On estimating regression. *Theory of Probability and its Applications*, **9**, pp. 141–142.

[42] Nadaraya, E. A. (1970). Remarks on nonparametric estimates for density functions and regression curves. *Theory of Probability and its Applications*, **15**, pp. 134–137.

[43] Papadrakakis, M. and Lagaros, N. (2002). Reliability–based structural optimization using neural networks and Monte Carlo simulation. *Computer Methods in Applied Mechanics and Engineering*, **191**, pp. 3491–3507.

[44] Rafajłowicz, E. (1987). Nonparametric orthogonal series estimators of regression: A class attaining the optimal convergence rate in L2. *Statistics and Probability Letters*, **5**, pp. 219-224.

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

[46] Stone, C. J. (1977). Consistent nonparametric regression. *Annals of Statististics*, **5**, pp. 595–645.

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

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

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

[50] Watson, G. S. (1964). Smooth regression analysis. *Sankhya Series A*, **26**, pp. 359–372.

[51] Wahba, G. (1990). *Spline Models for Observational Data.* SIAM, Philadelphia, PA.

Supplementary material for the referees

Proof of Lemma 3.

In the proof we will use Proposition 3.8 in Mhaskar (1993), which we reformulate here (in a slightly different form) as Lemma 7.

Lemma 7. *Let $K \subseteq \mathbb{R}^d$ be a polytope bounded by hyperplanes $v_j \cdot x + w_j \geq 0$ $(j = 1, \ldots, L)$, where $v_1, \ldots, v_L \in \mathbb{R}^d$ and $w_1, \ldots, w_L \in \mathbb{R}$. For $\delta > 0$ set*

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

and

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

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

$$f(x) = \sigma \left(\sum_{j=1}^{L} b_j \cdot \sigma \left(\sum_{k=1}^{d} a_{j,k} \cdot x^{(k)} + a_{j,0} \right) + b_0 \right)$$

satisfying

$$\begin{aligned} &|f(x)| \leq 1 \textit{ for } x \in \mathbb{R}^d, \\ &|f(x) - 1| \leq \varepsilon \textit{ for } x \in K_\delta^0, \\ &|f(x)| \leq \varepsilon \textit{ for } x \in K_\delta^c. \end{aligned} \tag{36}$$

In case that the squashing function satisfies

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

the weights above can be chosen such that

$$\begin{aligned} &|b_j| \leq \frac{4L}{\varepsilon} \quad \textit{for all } j = 0, \ldots, L \\ &|a_{j,k}| \leq \frac{4L}{\delta} \cdot \max\{\|v_1\|_\infty, |w_1|, \ldots, \|v_L\|_\infty, |w_L|\} \quad \textit{for all } j = 1, \ldots, L, k = 0, \ldots, d. \end{aligned}$$

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

Proof of Lemma 3. We partition $\left[-a - \frac{2a}{M}, a\right]^d$ into $(M+1)^d$ equivolume cubes of side length $2a/M$. Approximating m by a piecewise constant approximant with respect to this partition yields (since m is (p, C)–smooth with $p \leq 1$) a function S satisfying

$$\|S - m\|_{\infty, [-a,a]^d} \leq \sqrt{d} \cdot C \cdot \left(\frac{2a}{M} \right)^p. \tag{37}$$

If we we choose S suitably, it can be expressed in the form

$$S(x) = m(x_{(1,\dots,1)}) + \sum_{i \in \{1,\dots,M+1\}^d \setminus \{(1,\dots,1)\}} d_i \cdot \prod_{j=1}^{d} \left(x^{(j)} - x_i^{(j)} \right)_+^0 ,$$

where x_i are the corners of the cubes forming the above partition (indexed in ascending order per component), $0^0 := 0$, $x_+ := \max\{x, 0\}$, and d_i for $i = (i_1, \dots, i_d)$ as above are constants satisfying

$$d_i = \sum_{J \subseteq \{1,\dots,d\} \setminus \{k \,:\, i_k = 1\}} (-1)^{|J|} \cdot m\left(x_{i-J}\right), \tag{38}$$

where $i - J$ symbolizes the index i with i_j replaced by $i_j - 1$ for all $j \in J$. Since for a fixed set with $n > 0$ elements the number of subsets with even and uneven cardinality is 2^{n-1}, respectively, and the corners used in the above expression have a distance of at most $\sqrt{d} \cdot \frac{2a}{M}$, we can conclude from the (p, C)–smoothness of m

$$|d_i| \leq 2^{d - |\{k \,:\, i_k = 1\}| - 1} \cdot C \cdot d^{\frac{p}{2}} \cdot \left(\frac{2a}{M} \right)^p \leq c_{14} \cdot \left(\frac{2a}{M} \right)^p . \tag{39}$$

Let K_i be the polytope defined by $x^{(j)} - x_i^{(j)} \geq 0$ $(j = 1, \dots, d)$. Set $\varepsilon = (M+1)^{-d}$, $\delta = a \cdot \eta / (2 \cdot d \cdot M)$ and apply Lemma 7 for each K_j (i.e., with $L = d$, $v_j = \mathbf{e}_j$ and $w_j = -x_i^{(j)}$, where $\mathbf{e}_j$ denotes the j-th unit vector) to obtain $f_i(x)$ satisfying (36) with K_i instead of K. Let

$$P(x) = m(x_{(1,\dots,1)}) + \sum_{i \in \{1,\dots,M+1\}^d \setminus \{(1,\dots,1)\}} d_i \cdot f_i(x).$$

Then we can conclude from (36) and (39)

$$\begin{aligned} |P(x) - S(x)| &\leq \sum_{i \in \{1,\dots,M+1\}^d \setminus \{(1,\dots,1)\}} |d_i| \cdot \left| f_i(x) - \prod_{j=1}^{d} \left(x^{(j)} - x_i^{(j)} \right)_+^0 \right| \\ &\leq \sum_{i \in \{1,\dots,M+1\}^d \setminus \{(1,\dots,1)\}} |d_i| \cdot (M+1)^{-d} \\ &\leq c_{14} \cdot \left(\frac{2a}{M} \right)^p \end{aligned} \tag{40}$$

for all $x \in [-a, a]^d$ which are not contained in

$$\bigcup_{j=1,\dots,d} \bigcup_{i \in \{1,\dots,M+1\}^d} \left\{ x \in \mathbb{R}^d \quad : \quad |x^{(j)} - x_i^{(j)}| < a \cdot \eta / (2 \cdot d \cdot M) \right\}. \tag{41}$$

By shifting the positions of the x_i in the jth component slightly to the right (in the sense of increasing values) 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 P, that still satisfy (37) and (40) for all $x \in [-a,a]^d$, and corresponding disjoint versions of

$$\bigcup_{i \in \{1,\dots,M+1\}^d} \left\{ x \in \mathbb{R}^d \quad : \quad |x^{(j)} - x_i^{(j)}| < a \cdot \eta / (2 \cdot d \cdot M) \right\},$$

and since the sum of the ν–measures of these sets is less than or equal to one, at least one of them must have measure less than or equal to η/d. Consequently we can shift the x_i such that (41) has ν–measure less than or equal to η. This together with (37) and (40) implies the first assertion of the lemma, because $P(x)$ complies with the structure of the postulated neural network $t(x)$.

In case that σ satisfies the conditions specified in the second part of the lemma, Lemma 7 allows to bound the coefficients of the neural network $t(x) := P(x)$ respecting the values of the parameters we used during the application of this lemma above. This leads to

$$|a_{i,j,k}| \leq \frac{4 \cdot d \cdot 2 \cdot d \cdot M}{a \cdot \eta} \cdot \max\left\{1, a + \frac{2a}{M}\right\} \leq 8 \cdot d^2 \cdot \frac{M}{\eta} \cdot \max\left\{\frac{1}{a}, 3\right\}$$

for all $i \in \left\{1, \dots, (M+1)^d\right\}$, $j \in \{1, \dots, d\}$, $k \in \{0, \dots, d\}$ and

$$|b_{i,j}| \leq 4 \cdot d \cdot (M+1)^d$$

for all $i \in \left\{1, \dots, (M+1)^d\right\}$, $j \in \{0, \dots, d\}$. Furthermore, the definition of P and (38) imply

$$|d_i| \leq 2^d \cdot \|m\|_\infty$$

for all $i \in \left\{0, \dots, (M+1)^d\right\}$, which leads to the second assertion of the lemma. □