Estimation of an improved surrogate model in uncertainty quantification by neural networks [*]

Benedict Götz[1], Sebastian Kersting[2,†] and Michael Kohler[2]

[1] *Fachgebiet Systemzuverlässigkeit, Adaptronik und Maschinenakustik SAM, Technische Universität Darmstadt, Magdalenenstr. 4, 64289 Darmstadt, Germany, email: goetz@szm.tu-darmstadt.de*

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

June 04, 2018

Abstract

Quantification of uncertainty of a technical system is often based on a surrogate model of a corresponding simulation model. In any application the simulation model will not describe the reality perfectly, and consequently the surrogate model will be imperfect. In this article we combine observed data from the technical system with simulated data from the imperfect simulation model in order to estimate an improved surrogate model consisting of multi-layer feedforward neural networks, and we show that under suitably assumptions this estimate is able to circumvent the curse of dimensionality. Based on this improved surrogate model we show a rate of convergence result for density estimates. The finite sample size performance of the estimates is illustrated by applying them to simulated data. The practical usefulness of the newly proposed estimates is demonstrated by using them to predict the uncertainty of a lateral vibration attenuation system with piezo–elastic supports.

AMS classification: Primary 62G07; secondary 62P30.

Key words and phrases: Curse of dimensionality, density estimation, imperfect models, L_1 error, neural networks, surrogate models, uncertainty quantification.

1 Introduction

1.1 An example

In this article we develop new methods for the statistical inference in connection with complex technical systems. As an example we consider the lateral vibration attenuation system with piezo–elastic supports described in Figure 1.

[*] Running title: *Uncertainty quantification by neural networks*

[†] Corresponding author. Tel: +49-6151-16-23374, Fax:+49-6151-16-23381

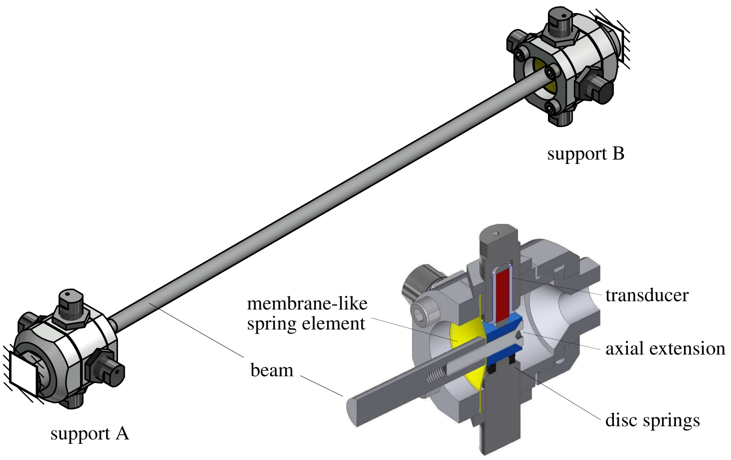

Figure 1: A CAD model of the lateral vibration attenuation system with piezo–elastic supports and a sectional view of one of the piezo–elastic supports, cf. Li et al (2017) .

This system consists of a beam with circular cross-section embedded in two piezo–elastic supports A and B where support A is used for lateral beam vibration excitation support and B is used for lateral beam vibration attenuation, as proposed in Götz, Platz and Melz (2016). The two piezo–elastic supports A and B are located at the beam's end and each consist of one elastic membrane-like spring element made of spring steel, two piezoelectric stack transducers arranged orthogonally to each other and mechanically prestressed with disc springs as well as the relatively stiff axial extension made of hardened steel that connects the piezoelectric transducers with the beam. For vibration attenuation in support B, optimally tuned electrical shunt circuits are connected to the piezoelectric transducers.

Our aim is to predict the maximal amplitude of the vibration occurring in an experiment with this attenuation system. If we construct such attenuation systems several times the constructed attenuation systems will be different due to variations in the parts used in the construction (e.g., the height or the stiffness of the used membrane) or in the construction process, and consequently the results which we measure in experiments with the systems will vary. E.g., building such systems ten times and measuring the maximal vibration amplitude in an experiment with each of the built systems, we got the following ten values in $[\frac{m}{s^2}/V]$:

$$\begin{aligned} &y_1 = 14.50,\ y_2 = 14.17,\ y_3 = 14.37,\ y_4 = 14.16,\ y_5 = 14.28, y_6 = 13.51, \\ &y_7 = 14.73, y_8 = 13.21,\ y_9 = 13.05,\ y_{10} = 16.26. \end{aligned} \tag{1}$$

We assume in the sequel that $y_1, \ldots, y_{10}$ are independent realizations of a real-valued random variable Y, and in order to get information about the distribution of Y we try to estimate the density $g : \mathbb{R} \to \mathbb{R}$ of Y with respect to the Lebesgue measure (which we assume to exist).

The classical statistical approach of doing this is to assume that Y is, e.g., normally distributed, to estimate its mean and its variance by maximum likelihood and to use the density of the corresponding normal distribution as an estimate of the density of Y. For the data in (1) this results in the blue curve in Figure 2. However the maximum vibration amplitudes represents extrem values of the lateral beam-column vibration transfere behaviour. According to Choi, Grandhi and Canfield (2007), the distribution of extreme values is characterized by a non-symmetric distribution about the most likely value. Thus this approach seems to be unpromising. The standard approach in modern statistics would be to use a nonparametric estimate of the density of Y, e.g. the classical kernel density estimate of Rosenblatt (1956) and Parzen (1962)

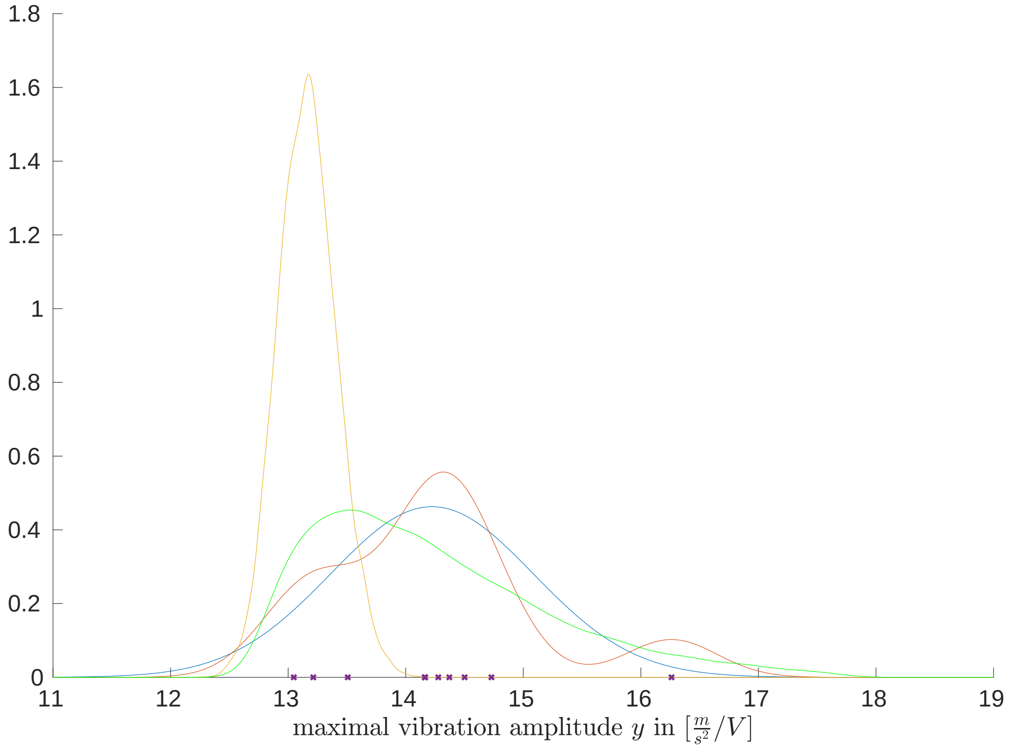

Figure 2: A parametric (blue line), a nonparametric (red line) estimate of the density of the data (1). A surrogate estimate on experimental data (5) (green line) and a surrogate estimate on computer simulated data (8) (green line). Additionally the data set (1) indicated on the x axis.

$$\hat{g}_{Y,n}(y) = \frac{1}{n \cdot h_n} \cdot \sum_{i=1}^{n} K\left(\frac{y - Y_i}{h_n}\right), \tag{2}$$

which we apply in the above formula to random variables $Y_1, \dots, Y_n$ which are independent and identically distributed as Y. Here $K : \mathbb{R}^d \to \mathbb{R}$ (so-called kernel, which is assumed to be a density) and $h_n > 0$ (so-called bandwidth) are parameters of the estimate. E.g., computing this kernel density estimate with the routine *ksdensity()* in *MATLAB* results in the red curve in Figure 2.

The obvious drawback of the first approach is that the error of this parametric estimate might be rather large in case that the true density of Y is not the density of a normal distribution, in particular if it cannot be approximated well by any such density. However, due to the small sample size in this example it is not clear that the second approach, i.e., the nonparametric density estimate, yields an estimate which is better than the parametric estimate. So in general neither of these two approaches will lead to satisfying results.

Unfortunately, it is not really possible to increase the sample size 10 of the data (1) in such a way that a nonparametric estimate seems promising, since experiments with the above attenuation system (in particular the construction and replacement of the membrane-like spring elements) are extremely time consuming. What we do instead in the sequel is to use some knowledge outside of the data (e.g., knowledge from engineering science about attenuation systems) in order to improve our estimation.

Often this is done in the framework of Bayesian statistics, where some kind of a priori distribution describing the system under consideration is assumed to be given, and under the assumption that this is indeed true, estimates are constructed which achieve good results even for very small sample sizes. However, this is an example of the saying 'We buy information with assumptions" (Coombs (1964)), which of course might lead to wrong informations in the case of wrong assumptions. And since it is not obvious how to transform the knowledge in engineering science into assumptions about an a priori distribution, we will not use this approach.

Instead, we will use the following knowledge in engineering science in order to construct an improved estimate: It is known that five parameters of the membrane in the attenuation system vary during the construction of the attenuation system and influence the maximal vibration amplitude: the lateral stiffness in direction of y ($k_{lat,y}$) and in direction of z ($k_{lat,z}$), the rotatory stiffness in direction of y ($k_{rot,y}$) and in direction of z ($k_{rot,z}$), and the height of the membrane (h_x). For given values of these five parameters it is possible to compute in a physical model of the attenuation system the corresponding maximal vibration amplitude. In order to generate values of these five parameters we need to determine their distributions. Therefore we measured the corresponding parameters for the ten built systems. As a result we got the data in Table 1. We assume that the four stiffness properties as well as the height property are multivariate normally distributed and estimate their distribution with the *mlest()* routine of the *mvnmle* package of the statistic software *R*. We obtain the expectation vector

$$\hat{\mu} = \begin{pmatrix} 124.9572 & 125.8931 & 33046576 & 32834749 & 0.00678 \end{pmatrix}$$

	1	2	3	4	5	6	7	8	9	10
$k_{rot,y} \times 10^2$	1.31	1.34	1.31	1.23	1.14	1.29	1.35	1.28	1.04	1.20
$k_{rot,z} \times 10^2$	1.31	1.28	1.43	1.25	1.30	1.34	1.22	1.16	1.18	1.11
$k_{lat,y} \times 10^7$	3.27	3.28	3.35	3.29	3.22	3.26	3.19	3.54	3.21	3.42
$k_{lat,z} \times 10^7$	3.07	3.22	3.29	3.25	3.30	3.18	3.16	3.51	3.37	3.44
$h_x \times 10^{-4}$	6.79	6.77	6.82	6.80	6.79	6.76	6.81	6.74	6.68	6.84
$y \times 10^1$	1.45	1.42	1.44	1.42	1.43	1.35	1.47	1.32	1.31	1.63

Table 1: Measured data for the ten built systems. The values of $k_{rot,y}$ and $k_{rot,z}$ are given in $[Nm/\,\mathrm{rad}]$, the values of $k_{lat,y}$ and $k_{lat,z}$ are given in $[N/m]$, the values of h_x are given in $[m]$ and the values of y are given in $[\frac{m}{s^2}/V]$.

and the covariance matrix

$$\hat{\Sigma} = \begin{pmatrix} 88.85741 & 32.74759 & 1595777 & -5647359 & 0.0001846703 \\ 32.74759 & 79.76893 & -2919445 & -6593387 & 0.0001762972 \\ 1595777 & -2919445 & 1.070764 \times 10^{12} & 884544431242 & -14.19626 \\ -5647359 & -6593387 & 8.845444 \times 10^{11} & 1.5991 \times 10^{12} & -32.52903 \\ 0.0001846703 & 0.0001762972 & -14.19626 & -32.52903 & 1.600001 \times 10^{-9} \end{pmatrix}.$$

By this assumption we have specified the distribution of a 5 dimensional random vector X, and our computer program computes a function $m : \mathbb{R}^5 \to \mathbb{R}$ such that the distribution of $m(X)$ is an approximation of the distribution of the maximal vibration amplitude Y occurring in experiments with our attenuation system.

In this stochastic model of our attenuation system we can generate independent data $X_{n+1}, \ldots, X_{n+L_n}$, compute $m(X_{n+1}), \ldots, m(X_{n+L_n})$ and define a kernel density estimate by

$$\hat{g}_{L_n}(y) = \frac{1}{L_n \cdot h_{L_n}} \cdot \sum_{i=1}^{L_n} K\left(\frac{y - m(Y_{n+i})}{h_{L_n}}\right).$$

However, the evaluation of the computer program for our technical system will often be rather time consuming and consequently L_n (although much larger than n) might not be really large. One possibility to circumvent this problem is to define an estimate of g on the basis of the data

$$(X_{n+1}, m(X_{n+1})), \ldots, (X_{n+L_n}, m(X_{n+L_n})), X_{n+L_n+1}, \ldots, X_{n+L_n+N_n} \tag{3}$$

by estimating in a first step a surrogate

$$\begin{aligned} & m_{(X,m(X)),L_n}(\cdot) = \\ & m_{(X,m(X)),L_n}(\cdot, (X_{n+1}, m(X_{n+1})), \ldots, (X_{n+L_n}, m(X_{n+L_n}))) : \mathbb{R}^d \to \mathbb{R} \end{aligned} \tag{4}$$

of m and by defining in a second step the corresponding surrogate density estimate via

$$\hat{g}_{(X,m(X)),L_n} = \frac{1}{N_n \cdot h_{N_n}} \cdot \sum_{i=1}^{N_n} K\left(\frac{y - m_{(X,m(X)),L_n}(X_{n+L_n+i})}{h_{N_n}}\right). \tag{5}$$

Computing such an surrogate density estimate results in the yellow line in Figure 2. Alternatively, one can also ignore the simulation model completely, and can use instead the data

$$(X_1, Y_1), \dots, (X_n, Y_n), X_{n+L_n+1}, \dots, X_{n+L_n+N_n} \tag{6}$$

in order to construct an estimate

$$m_{(X,Y),n}(\cdot) = m_{(X,Y),n}(\cdot, (X_1, Y_1), \dots, (X_n, Y_n)) : \mathbb{R}^d \to \mathbb{R} \tag{7}$$

of $m^*(x) = \mathbf{E}\{Y|X = x\}$, and can define the corresponding surrogate density estimate by

$$\hat{g}_{(X,Y),n} = \frac{1}{N_n \cdot h_{N_n}} \cdot \sum_{i=1}^{N_n} K\left(\frac{y - m_{(X,Y),n}(X_{n+L_n+i})}{h_{N_n}}\right). \tag{8}$$

The main question which we want to investigate theoretically in this paper is whether there exist situations in which suitably defined estimates based on the complete data

$$\begin{aligned} &(X_1, Y_1), \dots, (X_n, Y_n), (X_{n+1}, m(X_{n+1})), \dots, (X_{n+L_n}, m(X_{n+L_n})), \\ &X_{n+L_n+1}, \dots, X_{n+L_n+N_n} \end{aligned} \tag{9}$$

(where $n, L_n, N_n \in \mathbb{N}$) achieve simultaneously better rate of convergence results than the estimates (2), (5) and (8).

1.2 Mathematical setting

The mathematical setting which we consider is as follows: Let (X, Y), (X_1, Y_1), (X_2, Y_2), ... be independent and identically distributed random variables with values in $\mathbb{R}^d \times \mathbb{R}$, and let $m : \mathbb{R}^d \to \mathbb{R}$ be a measurable function. Here Y describes the outcome of an experiment with the technical system, and our aim is to predict the density g of Y (w.r.t. the Lebesgue measure), which we assume to exist. The random vector X and the measurable function m describe our stochastic model of the technical system, and in this model we use $m(X)$ as an approximation of Y. Let $m^*(x) = \mathbf{E}\{Y|X = x\}$ be the regression function of (X, Y). In the sequel we will assume that

$$\mathbf{E}\left\{|Y - m^*(X)|^2\right\}$$

is small, so that it is reasonable to try to approximate Y by some $\hat{m}_n(X)$. Given the data (9) our goal is to construct an estimate of g.

1.3 Definition of a class of neural networks

In order to construct such an estimate, we proceed as follows: Let $\sigma : \mathbb{R} \to \mathbb{R}$ be a so-called squashing function, i.e., assume that σ is monotonically increasing and satisfies $\lim_{x\to-\infty} \sigma(x) = 0$ and $\lim_{x\to\infty} \sigma(x) = 1$. In our applications in Section 3 we will use the so–called logistic squasher $\sigma(x) = 1/(1 + \exp(-x))$ $(x \in \mathbb{R})$.

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

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

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

$$|\mu_i| \leq B_n, \quad |\lambda_{i,j}| \leq B_n, \quad |\theta_{i,j,v}| \leq B_n$$

for all $i \in \{0, 1, \dots, M\}$, $j \in \{0, \dots, 4d^*\}$ and $v \in \{0, \dots, d\}$, by $\mathcal{F}_{M,d,d^*,B_n}^{(\text{neural networks})}$.

We will impose the following assumption (which was introduced in Kohler and Krzyżak (2017a) as an assumption which is realistic in connection with complex technical systems which are build in a modular way) on the functions which we want to approximate by neural networks:

Definition 1 *Let $d \in \mathbb{N}$, $d^* \in \{1, \dots, d\}$ and $m \colon \mathbb{R}^d \to \mathbb{R}$.*
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 $f \colon \mathbb{R}^{d^*} \to \mathbb{R}$ 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 $K \in \mathbb{N}$, $g_k \colon \mathbb{R}^{d^*} \to \mathbb{R}$ $(k = 1, \dots, K)$ and $f_{1,k}, \dots, f_{d^*,k} \colon \mathbb{R}^d \to \mathbb{R}$ $(k = 1, \dots, K)$ 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*

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

Definition 2 *Let $p = k + \beta$ for some $k \in \mathbb{N}_0$ and $0 < \beta \leq 1$, and let $C > 0$.*
a) *We say that 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*

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

for all $x, z \in \mathbb{R}^d$.
b) *We say that a* **generalized hierarchical interaction model** *is* (p, C)**-smooth***, if all functions occurring in its definition are* (p, C)**-smooth**.

We will use the following recursively defined classes of neural networks (with parameters K, M, d, $d^* \in \mathbb{N}$ and $B_n > 0$) in order to approximate functions which satisfy a generalized hierarchical interaction model: For $l = 0$, we define our space of hierarchical neural networks by

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

For $l > 0$, we define recursively

$$\mathcal{H}^{(l)}_{K,M,d,d^*,B_n} = \Bigg\{ h \colon \mathbb{R}^d \to \mathbb{R},\ h(x) = \sum_{k=1}^{K} g_k(f_{1,k}(x), \dots, f_{d^*,k}(x)) \quad (x \in \mathbb{R}^d) \\ \text{for some } g_k \in \mathcal{F}^{(\text{neural networks})}_{M,d^*,d^*,B_n} \text{ and } f_{j,k} \in \mathcal{H}^{(l-1)}_{K,M,d,d^*,B_n} \Bigg\}. \tag{10}$$

1.4 Definition of the estimate

Give the data (9) we want to estimate the density g of Y. We start with defining a surrogate estimate

$$m_{L_n}(\cdot) = m_{L_n}(\cdot, (X_{n+1}, m(X_{n+1})), \dots, (X_{n+L_n}, m(X_{n+L_n}))) : \mathbb{R}^d \to \mathbb{R} \tag{11}$$

of the function m. For this we use a least squares estimate defined by

$$\tilde{m}_{L_n}(\cdot) = \arg \min_{h \in \mathcal{H}^{(l)}_{K_1, M_{1,n}, d, d^*, B_{1,n}}} \left(\frac{1}{L_n} \sum_{i=n+1}^{n+L_n} |h(X_i) - m(X_i)|^2 \right), \tag{12}$$

where $K_1, M_{1,n}, d^* \in \mathbb{N}$ and $B_{1,n} > 0$ are parameters of the estimate. For simplicity we assume here and in the sequel that the minimum above indeed exists. When this is not the case our theoretical results also hold for any estimate which minimizes the above empirical L_2 risk up to a sufficiently small additional term (e.g., $1/n$). In order to be able to analyze the rate of the convergence of this estimate for an arbitrary distribution of X we truncate this estimate at some height $\beta > 0$, i.e., we define

$$m_{L_n}(x) = T_\beta(\tilde{m}_{L_n}(x)) \quad (x \in \mathbb{R}^d) \tag{13}$$

where

$$T_\beta(z) = \begin{cases} \text{sign}(z) \cdot \beta & |z| > \beta \\ z & \text{otherwise} \end{cases}$$

for $z \in \mathbb{R}$. (Here we will assume later that $|m(x)| \leq \beta$ $(x \in \mathbb{R}^d)$ holds.) Next we define an estimate of $m^* - m_{L_n}$ on the basis of the residuals

$$\hat{\epsilon}_i = Y_i - m_{L_n}(X_i) \quad (i = 1, \dots, n). \tag{14}$$

To do this we define

$$\tilde{m}^{\hat{\epsilon}}_n(\cdot) = \arg \min_{h \in \mathcal{H}^{(l)}_{K_2, M_{2,n}, d, d^*, B_{2,n}}} \left(\frac{w^{(n)}}{n} \sum_{i=1}^{n} (\hat{\epsilon}_i - h(X_i))^2 + \frac{1 - w^{(n)}}{N_{1,n}} \sum_{i=1}^{N_{1,n}} (0 - h(X_{n+L_n+i}))^2 \right) \tag{15}$$

for some weight $w^{(n)} \in [0, 1]$ and parameters $K_2, M_{2,n}, d^* \in \mathbb{N}$ and $B_{2,n} > 0$, and set

$$\hat{m}^{\hat{\epsilon}}_n(x) = T_{c_1 \cdot \alpha_n} \tilde{m}^{\hat{\epsilon}}_n(x) \quad (x \in \mathbb{R}^d), \tag{16}$$

where $c_1 \geq 1$ and $\alpha_n > 0$.

We define our final surrogate model $(X, \hat{m}_n(X))$ for (X,Y) by

$$\hat{m}_n(x) = m_{L_n}(x) + \hat{m}_n^{\hat{\epsilon}}(x) \quad (x \in \mathbb{R}^d), \tag{17}$$

and estimate the density g of Y by applying a kernel density estimate to a sample of $\hat{m}_n(X)$. Therefore we choose a kernel $K \colon \mathbb{R} \to \mathbb{R}$ and a bandwidth $h_{N_{2,n}} > 0$ and set

$$\hat{g}_{N_{2,n}}(y) = \frac{1}{N_{2,n} \cdot h_{N_{2,n}}} \cdot \sum_{i=1}^{N_{2,n}} K\left(\frac{y - \hat{m}_n(X_{n+L_n+i})}{h_{N_{2,n}}}\right). \tag{18}$$

1.5 Main results

Our main assumptions in our theoretical result are the following: We assume for some $\alpha_n \geq \alpha_n^* > 0$ that

$$\mathbf{E}\left\{|Y - m^*(X)|^2\right\} \leq (\alpha_n^*)^2 \quad \text{and} \quad \sup_{x \in \mathbb{R}^d} |m(x) - m^*(x)| \leq \alpha_n,$$

and that $m : \mathbb{R}^d \to \mathbb{R}$ and the function

$$x \mapsto \mathbf{E}\left\{\frac{1}{\alpha_n}(Y - m(X))\Big| X = x\right\} = \frac{1}{\alpha_n}(m^* - m)(x)$$

both satisfy 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 $s \in (0, 1]$. Under some minor additional assumptions and with properly chosen parameters we are then able to show that our improved surrogate estimate satisfies

$$\mathbf{E}\left\{|Y - \hat{m}_n(X)|^2\right\} \leq c_2 \cdot \max\left\{(\alpha_n^*)^2, \alpha_n^2 \cdot (\log n)^3 \cdot n^{-\frac{2p}{2p+d^*}}, (\log L_n)^3 \cdot L_n^{-\frac{2p}{2p+d^*}}\right\}.$$

From this we are able to conclude for α_n^* sufficiently small and L_n sufficiently large that the L_1 error of our density estimate satisfies in case of a (r, C)–smooth density g

$$\mathbf{E} \int_{\mathbb{R}} |\hat{g}_{N_{2,n}}(y) - g(y)| \, dy \leq c_3 \cdot \left(\alpha_n \cdot (\log n)^{3/2} \cdot n^{-\frac{p}{2p+d^*}}\right)^{\frac{r}{r+1}}.$$

In case $\alpha_n \cdot (\log n)^{3/2} \to 0$ $(n \to \infty)$ sufficiently fast this rate of convergence converges faster to zero than any of the rate of convergences

$$n^{-\frac{r}{2r+1}}, \quad \alpha_n^{\frac{r}{r+1}} \quad \text{and} \quad \left(n^{-\frac{p}{2p+d^*}}\right)^{\frac{r}{r+1}} \tag{19}$$

which we would expect for the estimates (2), (5) and (8), resp.

The finite sample size behaviour of our estimates is illustrated by using simulated data, and we illustrate the usefulness of our newly proposed estimates for uncertainty quantification by applying them in the application above.

1.6 Discussion of related results

Neural networks belong since many years to the most promising approaches in nonparametric statistics in view of multivariate statistical applications, in particular in pattern recognition and in nonparametric regression (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)). New theoretical results in nonparametric regression show that neural networks with many hidden layer are able to circumvent under proper assumptions the so–called curse of dimensionality and achieve therefore good rate of convergence results in high-dimensional estimation problems (cf., Kohler and Krzyżak (2017a), Bauer and Kohler (2017) and Schmidt-Hieber (2017)). Our results in this article demonstrate that the techniques introduced in these papers also lead to good theoretical results in uncertainty quantification.

Estimation of surrogate methods for uncertainty quantification based on neural networks has been proposed in Papadrakakis and Lagaros (2002), but theoretical results for the proposed estimates have not been developped there. Other ways to estimate surrogate models have been introduced and investigated with the aid of the simulated and real data in connection with the quadratic response surfaces in Bucher and Burgund (1990), Kim and Na (1997) and Das and Zheng (2000), in context of support vector machines in Hurtado (2004), Deheeger and Lemaire (2010) and Bourinet, Deheeger and Lemaire (2011), and in context of kriging in Kaymaz (2005) and Bichon et al. (2008). See also Santner, Williams, and Notz (2003) and the literature cited therein for additional literature on the design and analysis of computer experiments.

Consistency and rate of convergence of density estimates based on surrogate models have been studied in Devroye, Felber and Kohler (2013), Bott, Felber and Kohler (2015) and Felber, Kohler and Krzyżak (2015a). A method for the adaptive choice of the smoothing parameter of such estimates has been presented in Felber, Kohler and Krzyżak (2015b).

In Bayesian analysis of computer experiments, Kennedy and O'Hagan (2001), Bayarri et al. (2007), Goh et al. (2013), Han, Santner and Rawlinson (2009), Higdon et al. (2013) and Wang, Chen and Tsui (2009) model the discrepancy between the computer experiments and the outcome of the technical system by a Gaussian process. Tuo and Wu (2015) pointed out that this approach might fail in case of an imperfect computer model, for which there exists no values of the parameters which fit the technical system perfectly, and suggested and analyzed non-Bayesian methods for the choice of the parameters of such models. Related methods for the calibration of computer models have been considered in Wong, Storlie and Lee (2017). There the error of the resulting model was estimated by using bootstrap. Confidence intervals for quantiles based on data from imperfect simulation models have been derived in Kohler et al. (2016).

The definition of our improved surrogate model is motivated by Kohler and Krzyżak (2017b), where a result for smoothing spline estimates is shown. In this article we extend this result from smoothing spline to least squares estimates, and apply it to neural networks. The main advantage of our new results is that we are able to apply our method also successfully to high-dimensional settings, where smoothing spline estimates usually fail to deliver reasonable results because of the curse of dimensionality.

1.7 Notation

Throughout this paper we use the following notation: $\mathbb{N}$, $\mathbb{N}_0$, $\mathbb{R}$ and $\mathbb{R}_+$ are the sets of positive integers, nonnegative integers, real numbers, and nonnegative real numbers, respectively. For $z \in \mathbb{R}$, we denote the smallest integer greater than or equal to z by $\lceil z \rceil$. For $f : \mathbb{R}^d \to \mathbb{R}$

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

is its supremum norm.

If X is a random variable, then $\mathbf{P}_X$ is the corresponding distribution, i.e., the measure associated with the random variable.

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

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

For $\epsilon > 0$, $x_1^n = (x_1, \dots, x_n) \in (\mathbb{R}^d)^n$ and a set $\mathcal{F}$ of functions $f \colon \mathbb{R}^d \to \mathbb{R}$ we define the L_2 covering number $\mathcal{N}_2(\epsilon, \mathcal{F}, x_1^n)$ as the minimal number $l \in \mathbb{N}$ of functions $g_1, \dots, g_l \colon \mathbb{R}^d \to \mathbb{R}$ which have the property

$$\left(\min_{j=1,\dots,l} \frac{1}{n} \sum_{i=1}^n |f(x_i) - g_j(x_i)|^2 \right)^{1/2} \le \epsilon$$

for each $f \in \mathcal{F}$.

1.8 Outline

The outline of this paper is as follows: The main results are presented in Section 2 and proven in Section 4. The finite sample size performance of our estimates is illustrated in Section 3 by applying it to simulated and real data.

2 Main results

In order to formulate our main result on the rate of convergence of our improved surrogate estimate we need the following definition.

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

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

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

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

It is easy to see that the logistic squasher $\sigma(x) = 1/(1 + \exp(-x))$ is N–admissible for any $N \in \mathbb{N}$ (cf., Bauer and Kohler (2017)).

Theorem 1 *Let $d, n, L_n \in \mathbb{N}$ with $2 \leq n \leq L_n$ and with $n^{c_4} \leq L_n \leq n^{c_5}$ for some $c_4, c_5 > 0$. Let (X, Y), (X_1, Y_1), ... be independent and identically distributed $\mathbb{R}^d \times \mathbb{R}$–valued random variables with $\mathbf{E}\{|Y|\} < \infty$ and with supp(X) bounded. Let $m^*(\cdot) = \mathbf{E}\{Y|X = \cdot\}$ be the regression function of (X, Y). Let $C > 0$ and let $p = q + s$ for some $q \in \mathbb{N}_0$ and $s \in (0, 1]$. Let $m\colon \mathbb{R}^d \to \mathbb{R}$ be a measurable function, which satisfies a (p, C)-smooth generalized hierarchical interaction model of order d^* and finite level l, and assume that in Definition 1 b) all partial derivates of order less than or equal to q of the functions $g_k, f_{j,k}$ of this generalized hierarchical interaction model are bounded, i.e., assume that each such function f satisfies*

$$\max_{\substack{j_1,\dots,j_d \in \{0,1,\dots,q\} \\ j_1+\dots+j_d \leq q}} \left\| \frac{\partial^{j_1+\dots+j_d} f}{\partial^{j_1} x^{(1)} \dots \partial^{j_d} x^{(d)}} \right\|_\infty \leq c_6, \tag{20}$$

and let all functions g_k be Lipschitz continuous with Lipschitz constant $L > 0$. Assume that for some $1 \leq \beta \leq n + L_n$

$$|m(x)| \leq \beta \quad (x \in \mathbb{R}^d). \tag{21}$$

Let $\alpha_n > \alpha_n^* \geq 0$ and assume

$$\mathbf{E}\left\{|Y - m^*(X)|^2\right\} \leq (\alpha_n^*)^2 \quad \textit{and} \quad \mathbf{E}\left\{|Y - m^*(X)|^3\right\} \leq (\alpha_n^*)^3, \tag{22}$$

that there exists $K, \sigma_0 > 0$ such that

$$K^2 \cdot \left(\mathbf{E}\left\{ \exp\left(\frac{(Y - m^*(X))^2}{\alpha_n \cdot K} \right) \Big| X \right\} - 1 \right) \leq \sigma_0 \quad \textit{a.s.}, \tag{23}$$

and that the regression function $\mathbf{E}\{\frac{1}{\alpha_n}(Y - m(X))|X = x\} = \frac{1}{\alpha_n}(m^* - m)(x)$ satisfies a (p, C)-smooth generalized hierarchical interaction model of order d^* and finite level l. Furthermore assume that in Definition 1 b) all partial derivates of order less than or equal to q of the functions $g_k, f_{j,k}$ of this generalized hierarchical interaction model are bounded, i.e., assume that each such function f satisfies (20), and let all functions g_k be Lipschitz continuous with Lipschitz constant $L > 0$. Assume

$$\sup_{x \in \mathbb{R}^d} |m^*(x) - m(x)| \leq \alpha_n \tag{24}$$

and

$$\left((\log L_n)^3 \cdot L_n^{-\frac{2p}{2p+d^*}} \right)^{1/3} \leq \alpha_n. \tag{25}$$

Define the estimate $\hat{m}_n$ as in Subsection 1.4, where we choose K_1, d, and d^* as in the definition of the generalized hierarchical interaction model for m and

$$M_{1,n} = \lceil c_7 \cdot L_n^{\frac{d^*}{2p+d^*}} \rceil$$

and $B_{1,n} = L_n^{c_8}$, where we choose K_2, d, and d^* as in the definition of the generalized hierarchical interaction model for $(m^* - m)/\alpha_n$, $N_{1,n}, M_{2,n} \in \mathbb{N}$ with $M_{2,n} \leq N_{1,n}/\log(N_{1,n})$ and $B_{2,n} = n^{c_8}$, where $\sigma \colon \mathbb{R} \to [0,1]$ is N-admissible according to Definition 3 for some $N \geq q$, and where we use some weight $w^{(n)} \in [0,1]$. Then there exists constants $c_9, \ldots, c_{14} \in \mathbb{R}_+$ such that

$$\begin{aligned}
&\mathbf{E}\left\{|Y - \hat{m}_n(X)|^2\right\} \\
&\leq c_9 \cdot (\alpha_n^*)^2 + c_{10} \cdot \alpha_n^2 \cdot (\log n)^3 \cdot M_{2,n}^{-\frac{2p}{d^*}} + c_{11} \cdot w^{(n)} \cdot \alpha_n^2 \cdot (\log n)^3 \cdot \frac{M_{2,n}}{n} \\
&\quad + c_{12} \cdot (1 - w^{(n)}) \cdot \alpha_n^2 + c_{13} \cdot (\log L_n)^3 \cdot L_n^{-\frac{2p}{2p+d^*}} + c_{14} \cdot \frac{\alpha_n^2}{n},
\end{aligned}$$

for n sufficiently large.

In particular, in case $w^{(n)} = 1$ and $M_{2,n} = \lceil c_{15} \cdot n^{\frac{d^*}{2p+d^*}} \rceil$ we get

$$\begin{aligned}
&\mathbf{E}\left\{|Y - \hat{m}_n(X)|^2\right\} \\
&\leq c_{16} \cdot \max\left\{(\alpha_n^*)^2, \alpha_n^2 \cdot (\log n)^3 \cdot n^{-\frac{2p}{2p+d^*}}, (\log L_n)^3 \cdot L_n^{-\frac{2p}{2p+d^*}}\right\}
\end{aligned}$$

for some $c_{16} \in \mathbb{R}_+$ and n sufficiently large.

Theorem 1 implies the following corollary concerning the L_1 error of the density estimate (18):

Corollary 1 *Assume that the density g of Y is (r, C)–smooth for some $r \in (0,1]$ and that its support is compact. Let $K : \mathbb{R} \to \mathbb{R}$ be a symmetric and bounded density which decreases monotonically on $\mathbb{R}_+$ and define the estimate $\hat{g}_{N_{2,n}}$ as in Subsection 1.4, where $\hat{m}_n$ is defined as in the end of Theorem 1. Assume that the assumptions of Theorem 1 are satsified, and that, in addition,*

$$\max\left\{(\alpha_n^*)^2, (\log L_n)^3 \cdot L_n^{-\frac{2p}{2p+d^*}}\right\} \leq \alpha_n^2 \cdot (\log n)^3 \cdot n^{-\frac{2p}{2p+d^*}}$$

holds. Set

$$h_{N_{n,2}} = c_{17} \cdot \left(\alpha_n \cdot (\log n)^{3/2} \cdot n^{-\frac{p}{2p+d^*}}\right)^{\frac{1}{r+1}}$$

and assume

$$N_{2,n} \geq \left(\frac{n^{\frac{p}{2p+d^*}}}{\alpha_n \cdot (\log n)^{3/2}}\right)^{\frac{2r+1}{r+1}}$$

Then we have for some $c_{18} \in \mathbb{R}_+$

$$\mathbf{E} \int_{\mathbb{R}} |\hat{g}_{N_{2,n}}(y) - g(y)| \, dy \leq c_{18} \cdot \left(\alpha_n \cdot (\log n)^{3/2} \cdot n^{-\frac{p}{2p+d^*}} \right)^{\frac{r}{r+1}} ,$$

for n sufficiently large.

Proof. Lemma 1 in Bott, Felber and Kohler (2015) implies that for any $z_1, z_2 \in \mathbb{R}$ we have

$$\int \left| K\left(\frac{y - z_1}{h_n}\right) - K\left(\frac{y - z_2}{h_n}\right) \right| \, dy \leq 2 \cdot K(0) \cdot |z_1 - z_2|.$$

Consequently,

$$\hat{g}_{Y,N_{2,n}}(y) = \frac{1}{N_{2,n} \cdot h_{N_{2,n}}} \cdot \sum_{i=1}^{N_{2,n}} K\left(\frac{y - Y_{n+L_n+i}}{h_{N_{2,n}}}\right)$$

satisfies

$$\int |\hat{g}_{N_{2,n}}(y) - \hat{g}_{Y,N_{2,n}}(y)| \, dy \leq \frac{1}{N_{2,n} \cdot h_{N_{2,n}}} \cdot \sum_{i=1}^{N_{2,n}} 2 \cdot K(0) \cdot |\hat{m}_n(X_{n+L_n+i}) - Y_{n+L_n+i}|.$$

From this and standard bounds on the L_1 error of kernel density estimates (cf., e.g., proof of Theorem 1 in Felber, Kohler and Krzyżak (2015a)) we conclude

$$\begin{aligned}
&\mathbf{E} \int_{\mathbb{R}} |\hat{g}_{N_{2,n}}(y) - g(y)| \, dy \\
&\leq \mathbf{E} \int_{\mathbb{R}} |\hat{g}_{N_{2,n}}(y) - \hat{g}_{Y,N_{2,n}}(y)| \, dy + \mathbf{E} \int_{\mathbb{R}} |\hat{g}_{Y,N_{2,n}}(y) - g(y)| \, dy \\
&\leq \frac{2 \cdot K(0)}{h_{N_{2,n}}} \cdot \mathbf{E}\left\{|m_n(X) - Y|\right\} + \frac{c_{19}}{\sqrt{N_{2,n} \cdot h_{N_{2,n}}}} + c_{20} \cdot h_{N_{2,n}}^r \\
&\leq \frac{2 \cdot K(0)}{h_{N_{2,n}}} \cdot \sqrt{\mathbf{E}\left\{|m_n(X) - Y|^2\right\}} + \frac{c_{19}}{\sqrt{N_{2,n} \cdot h_{N_{2,n}}}} + c_{20} \cdot h_{N_{2,n}}^r.
\end{aligned}$$

Application of Theorem 1 yields the assertion. □

Remark 1. As already mentioned in Subsection 1.5, we have that for $\alpha_n \to 0$ $(n \to \infty)$ sufficiently fast the nonasymptotic error bound in Corollary 1 converges faster to zero than any of the rate of convergences in (19) which we would expect for the estimates (2), (5) and (8), resp.

Remark 2. Since the rate of convergence in Corollary 1 does not depend on the dimension d of X, our newly proposed estimate is able to circumvent the curse of dimensionality under suitably assumptions on the structure of m.

Remark 3. The parameters of the estimate in Corollary 1 depend on the distribution of (X, Y) and on m. In the next subsection we propose data–dependent choices for these parameters and investigate the finite sample size performance of the resulting estimate with the aid of simulated data.

3 Application to simulated and real data

In this section we want to describe the implementation of our introduced surrogate estimation method and analyze the performance of the estimate by applying it to simulated and real data.

The surrogate estimate is defined by combining the least squares neural network estimates m_{L_n} and $\hat{m}_n^{\hat{\epsilon}}$ as described in Subsection 1.4. In both cases we use multi-layer feedforward neural networks, however the network parameters are chosen differently. For the estimate m_{L_n} we choose the parameter from the sets $l \in \{0, 1, 2\}$, $K_1 \in \{1, 2\}$, $d^* \in \{1, \dots, d\}$ and $M_{1,n} \in \{1, \dots, 5, 6, 16, \dots, 46\}$. For the estimate m_{L_n} the parameter selection is done data-dependent by a splitting of the sample, where we use $\lceil \frac{2}{3} \cdot L_n \rceil$ train data and $L_n - \lceil \frac{2}{3} \cdot L_n \rceil$ test data and we consider the parameter combination with the smallest empirical L_2 risk evaluated on the test data. Since the data set $(X_1, Y_1), \dots, (X_n, Y_n)$ is considered rather small we reduce the sets of possible parameters for $\hat{m}_n^{\epsilon}$ to $l \in \{0\}$, $K_2 \in \{1\}$, $d^* \in \{1, 2, 4\}$, $M_{2,n} \in \{1, 3, 5\}$ and the additional weighting parameter w is chosen from $\{0, 0.25, \dots, 1\}$. For the residual estimate we select the parameter with a 5-fold cross-validation. To solve the least squares problems in (12) and (15), we use the Levenberg-Marquardt algorithm implemented in the *MATLAB* function *lsqnonlin()* to approximate their solution. For our density estimate we use a sample of size $N_{2,n}$ of $\hat{m}_n(X)$ and apply a standard kernel density estimate implemented in the *MATLAB* function *ksdensity()*.

In the application on simulated data we consider the following setting. We choose the independent random variable X as uniformly distributed on $[0, 1]^d$ and an error term ϵ uniformly distributed on $[-1, 1]$ such that X and ϵ are independent. The dependent variable Y is defined by

$$Y = m^*(X) + \sigma^* \cdot \lambda^* \cdot \epsilon$$

for some $m^* \colon \mathbb{R}^d \to \mathbb{R}$, a noise factor $\sigma^* \in \{0.05, 0.2\}$ and $\lambda^* > 0$ selected as the empirical interquartile range of $m^*(X)$. We set

$$m(x) = m^*(x) + \sigma_m \cdot \lambda^* \quad (x \in \mathbb{R}^d)$$

where $\sigma_m \in \{0.1, 0.2, 0.5\}$.

Let $(X, Y), (X_1, Y_1), (X_2, Y_2) \dots$ be independent and identically distributed and random variables. Our estimate gets

$$(X_1, Y_1), \dots, (X_n, Y_n)$$

as data from our real technical system,

$$(X_{n+1}, m(X_{n+1})), \dots, (X_{n+L_n}, m(X_{n+L_n}))$$

as data from our (imperfect) model and the additional X-values

$$X_{n+L_n+1}, \dots, X_{n+L_n+N_n}.$$

We consider five different models with a constant deviation in the computer model. In each model we use sample sizes $n = 10$, $L_n = 200$, $N_{1,n} = 200$ and $N_{2,n} = 10^5$. The different functions used as m^* are listed below.

$$\begin{aligned}
m_1^*(x) =& \cot\left(\frac{\pi}{1+\exp\left(x_1^2 + 2\cdot x_2 + sin(6\cdot x_4^3) - 3\right)}\right) \\
&+ \exp\left(3\cdot x_3 + 2\cdot x_4 - 5\cdot x_5 + \sqrt{x_6 + 0.9\cdot x_7 + 0.1}\right) && (x \in [0,1]^7) \\
m_2^*(x) =& \frac{2}{x_1 + 0.008} + 3\cdot\log(x_2^7\cdot x_3 + 0.1)\cdot x_4 && (x \in [0,1]^7) \\
m_3^*(x) =& 2\cdot\log(x_1\cdot x_2 + 4\cdot x_3 + |\tan(x_4)| + 0.1) + x_3^4\cdot x_5^2\cdot x_6 \\
&- x_4\cdot x_7 + (3\cdot x_8^2 + x_9 + 2)^{0.1+4\cdot x_{10}^2} && (x \in [0,1]^{10}) \\
m_4^*(x) =& x_1 + \tan(x_2) + x_3^3 + \log(x_4 + 0.1) + 3\cdot x_5 + x_6 + \sqrt{x_7 + 0.1} && (x \in [0,1]^7) \\
m_5^*(x) =& \exp(\|x\|) && (x \in [0,1]^7)
\end{aligned}$$

As mentioned before, the parameter λ^* is chosen as the empirical interquartile range of $m^*(X)$ calculated on 10^5 realizations of X. The used values are $\lambda_1^* = 9.11$, $\lambda_2^* = 5.68$, $\lambda_3^* = 13.97$, $\lambda_4^* = 1.77$ and $\lambda_5^* = 1.64$.

The density of Y is the convolution of the density of $m^*(X)$ and a uniform density. We do not try to compute its exact form, instead we compute it approximately by a kernel density estimate (as implemented in the *MATLAB* routine *ksdensity())* applied to a sample of size 10^6. In order to evaluate the performance of our density estimates the result is treated as if it is the real density.

We compare our estimate (est. 4) with three other density estimates. The first one (est. 1) is a standard kernel density estimate applied to a sample of size n of Y, cf. (2). The estimates 2 and 3 are surrogate density estimates where the kernel density estimate of *MATLAB* is applied to a sample of size $N_{2,n}$ of the surrogate model. For the second estimate (est. 2) the surrogate model is chosen as a neural network trained on L_n realizations of $(X, m(X))$, cf. (5). For the third estimate (est. 3) the surrogate model is chosen as a neural network trained on n realizations of (X, Y), cf. (8).

The estimates are compared by their L_1 error. Therefore we approximate the integral by a Riemann sum defined on an equidistant partition consisting of 10^4 subintervals. Since we need to take the randomness of the L_1 error into account, we repeat each simulation 50 times and report in Table 2 and Table 3 the median (and in brackets the interquartile range) of the 50 L_1 errors.

Our newly proposed estimate outperforms the other three estimates in 22 of 30 cases. In all cases if σ_m is sufficiently small our estimate yields a smaller L_1 error than estimates 1 and 3, where the biggest difference is in model four where it is eight times smaller. In any simulation except one it is able to reduce the L_1 error compared to the surrogate estimate on computer model data (est. 2). The resulting L_1 error of estimate 3 is in any simulation higher than the error of the other three used estimates. We assume this is due to the complexity of the used functions m^* and the small sample size of 10.

We apply the four different estimates on the lateral vibration attenuation system data and illustrate the results in Figure 3. The number of experimental data is equal to 10.

	σ^*	5%		
	σ_m	0.1	0.2	0.5
m_1^*	est. 1	0.704 (0.168)	0.704 (0.168)	0.704 (0.168)
	est. 2	0.271 (0.043)	0.503 (0.077)	0.954 (0.085)
	est. 3	0.998 (0.345)	0.998 (0.345)	0.998 (0.345)
	est. 4	**0.162 (0.134)**	**0.218 (0.136)**	**0.191 (0.166)**
m_2^*	est. 1	0.525 (0.183)	0.525 (0.183)	**0.525 (0.183)**
	est. 2	**0.240 (0.919)**	0.330 (0.820)	0.811 (0.782)
	est. 3	1.086 (0.459)	1.086 (0.459)	1.086 (0.459)
	est. 4	0.284 (0.957)	**0.290 (0.866)**	0.644 (0.984)
m_3^*	est. 1	0.786 (0.163)	**0.786 (0.163)**	**0.786 (0.163)**
	est. 2	0.616 (0.460)	0.935 (0.124)	1.233 (0.263)
	est. 3	1.472 (0.847)	1.472 (0.847)	1.472 (0.847)
	est. 4	**0.562 (0.606)**	0.835 (0.595)	0.999 (0.590)
m_4^*	est. 1	0.329 (0.175)	0.329 (0.175)	0.329 (0.175)
	est. 2	0.102 (0.016)	0.208 (0.015)	0.516 (0.015)
	est. 3	0.878 (1.328)	0.878 (1.328)	0.878 (1.328)
	est. 4	**0.040 (0.029)**	**0.035 (0.018)**	**0.036 (0.022)**
m_5^*	est. 1	0.317 (0.183)	0.317 (0.183)	0.317 (0.183)
	est. 2	0.107 (0.035)	0.212 (0.032)	0.522 (0.031)
	est. 3	0.836 (1.422)	0.836 (1.422)	0.836 (1.422)
	est. 4	**0.064 (0.031)**	**0.068 (0.050)**	**0.067 (0.050)**

Table 2: Median (and interquartile range) of the L_1 error of the four different estimates for the five different models with a constant error in the computer model and five percent noise

To improve the stability of our estimate we increase the sample sizes L_n and $N_{1,n}$ to 500. As discussed in the introduction, we assume that the distribution of the maximal vibration amplitude is characterized by a non-symmetric distribution about the most likely value. This characteristic is described by the estimate 2 and our estimate 4, whereas the estimate 4 predicts higher values. If one considers the experimental data this is a plausible correction by the residual estimate $\hat{m}_n^{\hat{\epsilon}}$. Since we only have 10 real data, it is unclear how reliable the estimates 1 and 3 are.

4 Proofs

4.1 A general result on weighted generalized penalized least squares estimates

In the proof of Theorem 1 we will use an error bound for weighted generalized penalized least squares estimates, which will enable us to generalize the results in Kohler and Krzyżak (2017b) from smoothing spline estimates to least squares estimates.

	σ^*	20 %		
	σ_m	0.1	0.2	0.5
m_1^*	est. 1	0.697 (0.241)	0.697 (0.241)	0.697 (0.241)
	est. 2	0.272 (0.105)	0.470 (0.098)	0.934 (0.089)
	est. 3	1.185 (0.604)	1.185 (0.604)	1.185 (0.604)
	est. 4	**0.245 (0.131)**	**0.272 (0.157)**	**0.216 (0.162)**
m_2^*	est. 1	0.547 (0.181)	0.547 (0.181)	**0.547 (0.181)**
	est. 2	**0.233 (0.926)**	0.315 (0.966)	0.694 (0.764)
	est. 3	1.140 (0.401)	1.140 (0.401)	1.140 (0.401)
	est. 4	0.272 (0.951)	**0.296 (1.038)**	0.625 (1.018)
m_3^*	est. 1	0.666 (0.217)	**0.666 (0.217)**	**0.666 (0.217)**
	est. 2	0.579 (0.480)	0.844 (0.229)	1.212 (0.252)
	est. 3	1.263 (0.832)	1.263 (0.832)	1.263 (0.832)
	est. 4	**0.573 (0.543)**	0.776 (0.499)	0.999 (0.499)
m_4^*	est. 1	0.348 (0.219)	0.348 (0.219)	0.348 (0.219)
	est. 2	0.105 (0.015)	0.209 (0.016)	0.513 (0.015)
	est. 3	1.006 (1.057)	1.006 (1.057)	1.006 (1.057)
	est. 4	**0.055 (0.054)**	**0.055 (0.045)**	**0.049 (0.038)**
m_5^*	est. 1	0.372 (0.196)	0.372 (0.196)	0.372 (0.196)
	est. 2	0.110 (0.034)	0.207 (0.033)	0.518 (0.03)
	est. 3	1.003 (1.062)	1.003 (1.062)	1.003 (1.062)
	est. 4	**0.079 (0.045)**	**0.085 (0.085)**	**0.082 (0.057)**

Table 3: Median (and interquartile range) of the L_1 error of the four different estimates for the five different models with a constant error in the computer model and twenty percent noise

Theorem 2 *Let $d, n, L_n \in \mathbb{N}$, $w^{(n)} \in [0,1]$ with $2 \leq n \leq L_n$ and $1 \leq \beta \leq \beta_n = n + L_n$. Let (X,Y), (X_1,Y_1), ... be independent and identically distributed $\mathbb{R}^d \times \mathbb{R}$–valued random variables with $\mathbf{E}\{|Y|\} < \infty$. Set $m(x) = \mathbf{E}\{Y|X=x\}$ and assume*

$$|m(x)| \leq \beta \quad (x \in \mathbb{R}^d). \tag{26}$$

Let $\bar{Y}_{1,n}$, ..., $\bar{Y}_{n+L_n,n}$ be arbitrary $\mathbb{R}$–valued random variables satisfying

$$\max_{i=1,\dots,n+L_n} \mathbf{E}\left\{|\bar{Y}_{i,n}|^3\right\} \leq c_{21} < \infty. \tag{27}$$

Let $\mathcal{F}_n$ be a set of functions and

$$pen_n^2(f) \geq 0$$

be a penalty term for each $f \in \mathcal{F}_n$. Define the estimate m_n by

$$\tilde{m}_n(\cdot) = \arg\min_{f \in \mathcal{F}_n} \left(\sum_{i=1}^{n+L_n} w_i \cdot |f(X_i) - \bar{Y}_{i,n}|^2 + pen_n^2(f) \right)$$

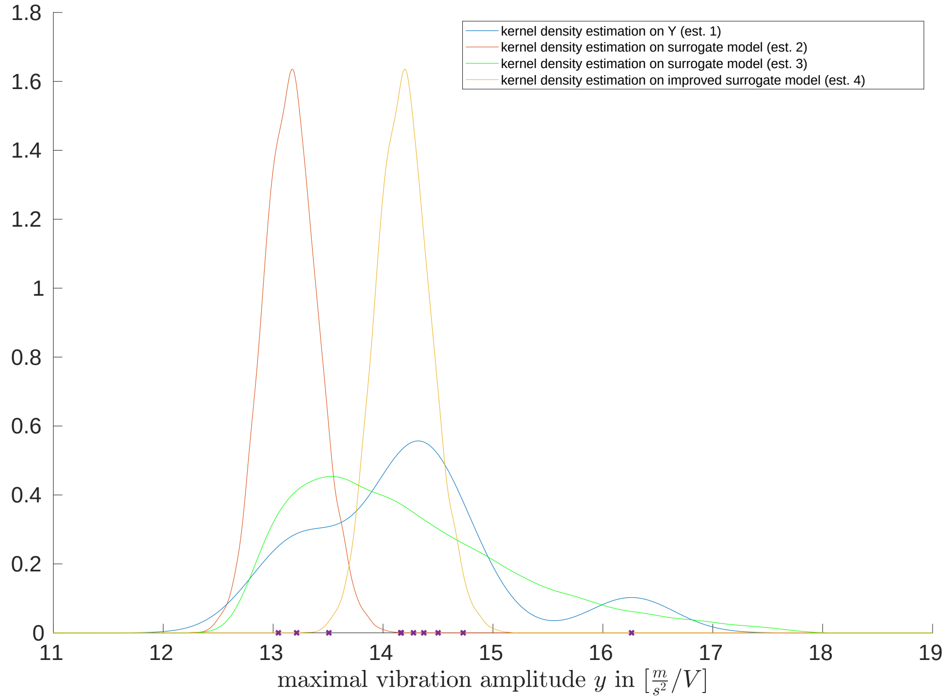

Figure 3: Four different density estimates and as reference the data set (1) indicated on the x axis.

and

$$m_n(x) = T_\beta(\tilde{m}_n(x)) \quad (x \in \mathbb{R}^d),$$

where

$$w_i = \frac{w^{(n)}}{n} \quad \textit{for } i = 1, \dots, n$$

and

$$w_i = \frac{1 - w^{(n)}}{L_n} \quad \textit{for } i = n+1, \dots, n+L_n.$$

Assume

$$K^2 \cdot \left(\mathbf{E} \left\{ \exp\left(\frac{(Y - m(X))^2}{K^2} \right) \big| X \right\} - 1 \right) \leq \sigma_0^2 \quad a.s. \tag{28}$$

for some $K, \sigma_0 > 0$. Choose $\delta_k > 0$ with $\delta_k \to 0$ $(k \to \infty)$ and $\delta_n \geq \delta_{L_n}$, such that for all $k \geq n$ we have

$$\delta_k > c_{22} \cdot \frac{\beta^2}{k}, \tag{29}$$

$$\sqrt{k}\delta \geq c_{23} \int_{\delta/(12\sigma_0)}^{\sqrt{48\delta}} \Bigg(\log \mathcal{N}_2\Bigg(u, \{T_{\beta_n} f - g : f \in \mathcal{F}_n, \tag{30}$$

$$\frac{1}{k}\sum_{i=1}^{k} |T_{\beta_n} f(x_i) - g(x_i)|^2 + pen_n^2(f) \leq 48 \cdot \delta\}, x_1^k \Bigg)\Bigg)^{1/2} du$$

for all $\delta \geq \delta_k/6$, all $g \in \mathcal{F}_n$, and

$$\frac{\sqrt{k}\delta}{\beta^2} \geq c_{23} \int_{\delta/(c_{24}\cdot\beta^2)}^{\sqrt{\delta}} \Bigg(\log \mathcal{N}_2\Bigg(u, \{(T_\beta f - m)^2 : f \in \mathcal{F}_n, \tag{31}$$

$$\frac{1}{k}\sum_{i=1}^{k} |T_\beta f(x_i) - m(x_i)|^2 \leq \frac{\delta}{\beta^2}, pen_n^2(f) \leq \delta\}, x_1^k \Bigg)\Bigg)^{1/2} du$$

for all $\delta \geq \delta_k$ and all $x_1, \dots, x_k \in \mathbb{R}^d$. Then there exists constants $c_{25}, c_{26}, c_{27} \in \mathbb{R}_+$ such that

$$\begin{aligned}
&\mathbf{E} \int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx) \\
&\leq 9 \cdot \inf_{f \in \mathcal{F}_n} \left(pen_n^2(f) + \int |f(x) - m(x)|^2 \mathbf{P}_X(dx) \right) \\
&\quad + c_{25} \cdot w^{(n)} \cdot \left(\delta_n + \mathbf{E}\left\{ \frac{1}{n} \cdot \sum_{i=1}^{n} |\bar{Y}_{1,n} - Y_i|^2 \right\} \right) \\
&\quad + c_{26} \cdot (1 - w^{(n)}) \cdot \left(\delta_{L_n} + \mathbf{E}\left\{ \frac{1}{L_n} \cdot \sum_{i=n+1}^{n+L_n} |\bar{Y}_{1,n} - Y_i|^2 \right\} \right) + \frac{c_{27}}{n}.
\end{aligned}$$

Proof. The proof follows by a generalization of the proof of Theorem 2 in Kohler and Krzyżak (2017b). A complete proof is available from the authors on request.

4.1.1 Application to neural networks

In the following subsection we want to introduce a corollary of Theorem 2, where we choose our function space as hierarchical neural networks as defined in Subsection 1.3.

Corollary 2 *Let $d, n, L_n \in \mathbb{N}$, $w^{(n)} \in [0,1]$ with $2 \leq n \leq L_n$ and $1 \leq \beta \leq n + L_n$. Let (X, Y), (X_1, Y_1), ... be independent and identically distributed $\mathbb{R}^d \times \mathbb{R}$–valued random variables with $\mathbf{E}\{|Y|\} < \infty$ and with supp(X) bounded. Let $m(\cdot) = \mathbf{E}\{Y|X = \cdot\}$ be the regression function, which satisfies a generalized hierarchical interaction model of order d^* and finite level l and assume*

$$|m(x)| \leq \beta \quad (x \in \mathbb{R}^d). \tag{32}$$

Let $\bar{Y}_{1,n}$, ..., $\bar{Y}_{n+L_n,n}$ be arbitrary $\mathbb{R}$–valued random variables satisfying

$$\max_{i=1,\dots,n+L_n} \mathbf{E}\left\{ |\bar{Y}_{i,n}|^3 \right\} \leq c_{28} < \infty. \tag{33}$$

Assume

$$K^2 \cdot \left(\mathbf{E}\left\{ \exp\left(\frac{(Y-m(X))^2}{K^2} \right) \big| X \right\} - 1 \right) \leq \sigma_0^2 \quad a.s. \tag{34}$$

for some $K, \sigma_0 > 0$. *Let* $N \in \mathbb{N}_0$ *and* $H^{(l)}_{K,M_n,d,d^*,B_n}$ *be the set of hierarchical neural networks introduced in Subsection 1.3, where* K, d, d^* *are chosen as in the definition of the generalized hierarchical interaction model for* m, *and where* $M_n \leq n^{c_{29}}$, $B_n = n^{c_{30}}$, *and where* $\sigma\colon \mathbb{R} \to \mathbb{R}$ *is a Lipschitz continuous function with Lipschitz constant* L, *which satisfy*

$$|\sigma(x)| \leq L \cdot \max\{|x|, 1\} \quad (x \in \mathbb{R}). \tag{35}$$

Define the estimate m_n *by*

$$\tilde{m}_n(\cdot) = \arg \min_{h \in \mathcal{H}^{(l)}_{K,M_n,d,d^*,B_n}} \left(\sum_{i=1}^{n+L_n} w_i \cdot |h(X_i) - \bar{Y}_{i,n}|^2 \right)$$

and

$$m_n(x) = T_\beta(\tilde{m}_n(x)) \quad (x \in \mathbb{R}^d),$$

where

$$w_i = \frac{w^{(n)}}{n} \quad \textit{for } i = 1, \dots, n$$

and

$$w_i = \frac{1 - w^{(n)}}{L_n} \quad \textit{for } i = n+1, \dots, n+L_n.$$

Then there exists constants $c_{31}, c_{32}, c_{33} \in \mathbb{R}_+$ *such that*

$$\begin{aligned}
&\mathbf{E} \int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx) \\
&\leq 9 \cdot \inf_{h \in \mathcal{H}^{(l)}_{K,M_n,d,d^*,B_n}} \left(\int |h(x) - m(x)|^2 \mathbf{P}_X(dx) \right) \\
&\quad + c_{31} \cdot w^{(n)} \cdot \left(\frac{\log(n)}{n} \cdot M_n + \mathbf{E}\left\{ \frac{1}{n} \cdot \sum_{i=1}^{n} |\bar{Y}_{1,n} - Y_i|^2 \right\} \right) \\
&\quad + c_{32} \cdot (1 - w^{(n)}) \cdot \left(\frac{\log(L_n)}{L_n} \cdot M_n + \mathbf{E}\left\{ \frac{1}{L_n} \cdot \sum_{i=n+1}^{n+L_n} |\bar{Y}_{1,n} - Y_i|^2 \right\} \right) + \frac{c_{33}}{n},
\end{aligned}$$

for n *sufficiently large.*

Proof. Set $pen_n^2(f) = 0$ and

$$\delta_k = c_{34} \cdot \frac{\log(k)}{k} \cdot M_n.$$

We show that Theorem 2 is applicable by the assumptions of Corollary 2 and the choice of δ_k. First we observe that

$$\delta_k > c_{35} \cdot \frac{\beta^2}{k}$$

and

$$\delta_n = c_{36} \cdot \frac{\log(n)}{n} \cdot M_n \geq c_{36} \cdot \frac{\log(L_n)}{L_n} \cdot M_n = \delta_{L_n},$$

since $2 \leq n \leq L_n$. In order to be able to apply Theorem 2 it suffices to show that (30) and (31) are fulfilled. First we show that (31) holds. Since the values of the estimate on $supp(X)$ will not change in case that we replace $\mathcal{H}^{(l)}_{K,M_n,d,d^*,B_n}$ by

$$\left\{ h \cdot I_{supp(X)} \, : \, h \in \mathcal{H}^{(l)}_{K,M_n,d,d^*,B_n} \right\}$$

in the definition of $\tilde{m}_n$, it suffices to show that (31) holds for $x_1, \dots, x_k \in supp(X)$. Next we observe that using $|a^2 - b^2|^2 \leq (|a| + |b|)^2 \cdot |a - b|^2$ $(a, b \in \mathbb{R})$ (which we apply with $a = (T_\beta f - m)(x_i)$ and $b = g(x_i)$, where g is approximating $T_\beta f - m$) and $|m(x)| \leq \beta$ $(x \in \mathbb{R}^d)$ we get

$$\begin{aligned}
&\left(\frac{1}{k} \sum_{i=1}^k |(T_\beta f - m)^2(x_i) - g^2(x_i)|^2 \right)^{1/2} \\
&\leq \left(\frac{1}{k} \sum_{i=1}^k \Big(|(T_\beta f - m)(x_i) - g(x_i)|^2 \cdot (|(T_\beta f - m)(x_i)| + |g(x_i)|)^2 \Big) \right)^{1/2} \\
&\leq 4 \cdot \beta \cdot \left(\frac{1}{k} \sum_{i=1}^k |(T_\beta f - m)(x_i) - g(x_i)|^2 \right)^{1/2}
\end{aligned}$$

for any $x_1, \dots, x_k \in supp(X)$, which implies

$$\begin{aligned}
&\mathcal{N}_2 \Big(u, \left\{ (T_\beta f - m)^2 : f \in \mathcal{H}^{(l)}_{K,M_n,d,d^*,B_n} \right\}, x_1^k \Big) \\
&\leq \mathcal{N}_2 \Big(\frac{u}{4\beta}, \left\{ T_\beta f - m : f \in \mathcal{H}^{(l)}_{K,M_n,d,d^*,B_n} \right\}, x_1^k \Big).
\end{aligned}$$

Using this we see that for any $\delta \geq \delta_k$

$$\begin{aligned}
&\int_{\delta/(c_{37} \cdot \beta^2)}^{\sqrt{\delta}} \Big(\log \mathcal{N}_2 \Big(u, \{ (T_\beta f - m)^2 : f \in \mathcal{H}^{(l)}_{K,M_n,d,d^*,B_n} \}, x_1^k \Big) \Big)^{1/2} du \\
&\leq \int_{\delta/(c_{37} \cdot \beta^2)}^{\sqrt{\delta}} \left(\log \mathcal{N}_2 \left(\frac{u}{4\beta}, \{ T_\beta f - m : f \in \mathcal{H}^{(l)}_{K,M_n,d,d^*,B_n} \}, x_1^k \right) \right)^{1/2} du
\end{aligned}$$

which is bounded by

$$\sqrt{\delta} \cdot \Big(\log \mathcal{N}_2 \Big(\frac{c_{38}}{k}, \{ T_\beta f - m : f \in \mathcal{H}^{(l)}_{K,M_n,d,d^*,B_n} \}, x_1^k \Big) \Big)^{1/2}$$

since

$$\frac{u}{4\beta} \geq \frac{c_{38}}{k} \quad \text{for} \quad u \geq \frac{\delta}{c_{37} \cdot \beta^2} \geq \frac{\delta_k}{c_{37} \cdot \beta^2} \geq \frac{c_{39}}{c_{37} \cdot k}.$$

Set $a_k = k^{c_{40}}$. Applying Lemma 2 from Bauer and Kohler (2017) yields for any $x_1, \dots, x_k \in [-a_k, a_k]^d$

$$\log\left(\mathcal{N}_2\left(\frac{c_{41}}{k}, \{T_\beta f - m : f \in \mathcal{F}_n\}, x_1^k\right)\right) \leq c_{42} \cdot \log(k) \cdot M_n.$$

Since $supp(X)$ is bounded the relationship $supp(X) \subseteq [-a_k, a_k]^d$ holds for k sufficiently large. Combing the above results we see that (31) is implied by

$$\frac{\sqrt{k} \cdot \delta}{\beta^2} \geq \sqrt{\delta} \cdot (c_{42} \cdot \log(k) \cdot M_n)^{1/2}$$

which in turn follows from $\delta \geq \delta_k$.

By the choice of δ_k we have for any $\delta \geq \delta_k/6$

$$\frac{\delta}{12\sigma_0} > \frac{c_{43}}{k}.$$

Arguing as above, this implies that (30) holds. Consequently Theorem 2 is applicable which yields the assertion.

□

4.2 Proof of Theorem 1

Using the definition of $\hat{m}_n$, $(a+b+c)^2 \leq 3a^2 + 3b^2 + 3c^2$ $(a, b, c \in \mathbb{R})$ and (22) we get

$$\begin{aligned}
&\mathbf{E}\left\{|Y - \hat{m}_n(X)|^2\right\} \\
&= \mathbf{E}\left\{\left|(Y - m^*(X)) + (m^*(X) - m(X) - \hat{m}_n^{\hat{\epsilon}}(X)) + (m(X) - m_{L_n}(X))\right|^2\right\} \\
&\leq 3 \cdot \mathbf{E}\left\{|Y - m^*(X)|^2\right\} + 3 \cdot \mathbf{E}\left\{\left|m^*(X) - m(X) - \hat{m}_n^{\hat{\epsilon}}(X)\right|^2\right\} \\
&\quad +3 \cdot \mathbf{E}\left\{|m(X) - m_{L_n}(X)|^2\right\} \\
&\leq 3(\alpha_n^*)^2 + 3 \cdot \mathbf{E}\int \left|\hat{m}_n^{\hat{\epsilon}}(x) - (m^* - m)(x)\right|^2 \mathbf{P}_X(dx) \\
&\quad +3 \cdot \mathbf{E}\int |m_{L_n}(x) - m(x)|^2 \mathbf{P}_X(dx).
\end{aligned}$$

Hence in order to prove the assertion it suffices to show

$$\mathbf{E}\int |m_{L_n}(x) - m(x)|^2 \mathbf{P}_X(dx) \leq c_{44} \cdot \log(L_n)^3 \cdot L_n^{-\frac{2p}{2p+d^*}} \tag{36}$$

and

$$\begin{aligned}
&\mathbf{E}\int \left|\hat{m}_n^{\hat{\epsilon}}(x) - (m^* - m)(x)\right|^2 \mathbf{P}_X(dx) \\
&\leq c_{45} \cdot \alpha_n^2 \cdot (\log n)^3 \cdot M_{2,n}^{-\frac{2p}{d^*}} + c_{46} \cdot w^{(n)} \cdot \alpha_n^2 \cdot \log(n) \cdot \frac{M_{2,n}}{n}
\end{aligned}$$

$$+c_{47} \cdot (1 - w^{(n)}) \cdot \alpha_n^2 + c_{48} \cdot (\log L_n)^3 \cdot L_n^{-\frac{2p}{2p+d^*}} + c_{49} \cdot \frac{\alpha_n^2}{n} \tag{37}$$

To proof inequality (36) we apply Corollary 2 with $(X, Y) = (X, m(X))$, $n = L_n$, $w^{(n)} = 1$ and $\bar{Y}_{i,L_n+\bar{L}_n} = Y_i = m(X_{n+i})$ $(i = 1, \dots, L_n)$ and suitably chosen $\bar{Y}_{L_n+1,L_n+\bar{L}_n}$, $\dots$,$\bar{Y}_{L_n+\bar{L}_n,L_n+\bar{L}_n}$ and observe

$$\begin{aligned}
&\mathbf{E} \int |m_{L_n}(x) - m(x)|^2 \, \mathbf{P}_X(dx) \\
&\leq 9 \cdot \inf_{h \in \mathcal{H}^{(l)}_{K_1,M_{1,n},d,d^*,B_{1,n}}} \left(\int |h(x) - m(x)|^2 \mathbf{P}_X(dx) \right) + c_{50} \cdot \frac{\log(L_n)}{L_n} \cdot M_{1,n} + \frac{c_{51}}{L_n} \\
&\leq 9 \cdot \inf_{h \in \mathcal{H}^{(l)}_{K_1,M_{1,n},d,d^*,B_{1,n}}} \left(\int |h(x) - m(x)|^2 \mathbf{P}_X(dx) \right) + c_{52} \cdot \log(L_n) \cdot L_n^{\frac{-2p}{2p+d^*}},
\end{aligned}$$

for sufficiently large n. Next we want to derive a bound on the approximation error. Set $a_{L_n} = (\log L_n)^{\frac{3}{2 \cdot (N+q+3)}}$ and $\eta_{L_n} = (\log L_n)^{\frac{3 \cdot (N+3)}{N+q+3}} \cdot L_n^{-\frac{2 \cdot (N+1) \cdot p + 2d^*}{2p+d^*}}$ and assume w.l.o.g. that $supp(X) \subseteq [-a_{L_n}, a_{L_n}]^d$. Using Theorem 3 in Bauer and Kohler (2017) we see that there exists a $h^* \in \mathcal{H}^{(l)}_{K_1,M_{1,n},d^*,d,B_{1,n}}$ and an exception set D_{L_n} with $\mathbf{P}_X$-measure of η_{L_n} such that

$$\begin{aligned}
&\int |h^*(x) - m(x)|^2 \cdot I_{D^c_{L_n}}(x) \, \mathbf{P}_X(dx) + \int |h^*(x) - m(x)|^2 \cdot I_{D_{L_n}}(x) \, \mathbf{P}_X(dx) \\
&\leq \left(c_{53} \cdot a_{L_n}^{(N+q+3)} \cdot M_{1,n}^{-p/d^*} \right)^2 + \left(2 \cdot c_{54} \cdot a_{L_n}^q \cdot M_{1,n}^{(d^*+N \cdot p)/d^*} \right)^2 \cdot \eta_{L_n} \\
&\leq c_{55} \cdot (\log L_n)^3 \cdot L_n^{-\frac{2p}{2p+d^*}} + c_{56} \cdot (\log L_n)^{\frac{3q}{N+q+3}} \cdot L_n^{\frac{2d^*+2N \cdot p}{2p+d^*}} \cdot (\log L_n)^{\frac{3 \cdot (N+3)}{N+q+3}} \cdot L_n^{-\frac{2 \cdot (N+1) \cdot p + 2d^*}{2p+d^*}} \\
&\leq c_{57} \cdot (\log L_n)^3 \cdot L_n^{-\frac{2p}{2p+d^*}},
\end{aligned}$$

where we have used that $|m(x)| \leq \beta \leq c_{58} \cdot a_{L_n}^q \cdot M_{1,n}^{(d^*+N \cdot p)/d^*}$.

In order to prove (37) we first observe that

$$\mathbf{E}\{Y - m(X) | X = x\} = m^*(x) - m(x),$$

hence $m^* - m$ is the regression function to $(X, Y - m(X))$, and $(m^* - m)/\alpha_n$ is the regression function to $(X, (Y - m(X))/\alpha_n)$. Clearly,

$$\int \left| \hat{m}_n^{\hat{\epsilon}}(x) - (m^* - m)(x) \right|^2 \mathbf{P}_X(dx) = \alpha_n^2 \cdot \int \left| \frac{1}{\alpha_n} \cdot \hat{m}_n^{\hat{\epsilon}}(x) - \frac{1}{\alpha_n} \cdot (m^* - m)(x) \right|^2 \mathbf{P}_X(dx).$$

It is easy to see that the definition of $\hat{m}_n^{\hat{\epsilon}}$ implies

$$\frac{1}{\alpha_n} \cdot \hat{m}_n^{\hat{\epsilon}}(x) = \frac{1}{\alpha_n} \cdot T_{c_1 \cdot \alpha_n}(\tilde{m}_n^{\hat{\epsilon}}(x)) = T_{c_1} \left(\frac{1}{\alpha_n} \cdot \tilde{m}_n^{\hat{\epsilon}}(x) \right) \quad (x \in \mathbb{R}^d),$$

and

$$\frac{1}{\alpha_n} \cdot \tilde{m}_n^{\hat{\epsilon}}(\cdot) \quad = \quad \arg \min_{h \in \frac{1}{\alpha_n} \mathcal{H}^{(l)}_{K_2, M_{2,n}, d^*, d, B_{2,n}}} \Bigg(\frac{w^{(n)}}{n} \sum_{i=1}^{n} \left(\frac{1}{\alpha_n} \cdot \hat{\epsilon}_i - h(X_i) \right)^2 + \frac{1 - w^{(n)}}{N_{1,n}} \sum_{i=1}^{N_{1,n}} (0 - h(X_{n+L_n+i}))^2 \Bigg),$$

where

$$\frac{1}{\alpha_n} \mathcal{H}^{(l)}_{K_2, M_{2,n}, d, d^*, B_{2,n}} = \left\{ h/\alpha_n \ : \ h \in \mathcal{H}^{(l)}_{K_2, M_{2,n}, d^*, d, B_{2,n}} \right\}.$$

The assumptions in Theorem 1 together with (36) imply that we have

$$\sup_{x \in \mathbb{R}^d} |m^*(x) - m(x)| \leq \alpha_n$$

and

$$\begin{aligned}
&\max_{i=1,\dots,n} \mathbf{E} \left\{ \left| \frac{Y_i - m_{L_n}(X_i)}{\alpha_n} \right|^3 \right\} \\
&\leq \frac{9}{\alpha_n^3} \cdot \left(\mathbf{E} \left\{ |Y - m^*(X)|^3 \right\} + \mathbf{E} \left\{ |m^*(X) - m(X)|^3 \right\} + \mathbf{E} \left\{ |m(X) - m_{L_n}(X)|^3 \right\} \right) \\
&\leq 9 \cdot \left(\frac{(\alpha_n^*)^3}{\alpha_n^3} + 1 + \frac{c_{59} \cdot \left((\log L_n)^3 \cdot L_n^{-\frac{2p}{2p+d^*}} \right)}{\alpha_n^3} \right) \leq 18 + c_{59}
\end{aligned}$$

We consider

$$\frac{1}{\alpha_n} \cdot \hat{\epsilon}_i = \frac{1}{\alpha_n} \cdot (Y_i - m_{L_n}(X_i)) = \frac{1}{\alpha_n} \cdot (Y_i - m(X_i)) + \frac{1}{\alpha_n} \cdot (m(X_i) - m_{L_n}(X_i))$$

as an observation of $(Y_i - m(X_i))/\alpha_n$ with an additional measurement error

$$\frac{1}{\alpha_n} \cdot (m(X_i) - m_{L_n}(X_i))$$

$(i = 1, \dots, n)$. And we consider

$$0 = \frac{1}{\alpha_n} \cdot (Y_{n+L_n+i} - m(X_{n+L_n+i})) - \frac{1}{\alpha_n} \cdot (Y_{n+L_n+i} - m(X_{n+L_n+i}))$$

as an observation of $\frac{1}{\alpha_n} \cdot (Y_{n+L_n+i} - m(X_{n+L_n+i}))$ with an additional measurement error

$$(-1) \cdot \frac{1}{\alpha_n} \cdot (Y_{n+L_n+i} - m(X_{n+L_n+i}))$$

$(i = 1, \dots, N_{1,n})$.

From inequality (36) we can conclude

$$\mathbf{E}\left\{\frac{1}{n}\sum_{i=1}^{n}\left|\frac{1}{\alpha_n}\cdot(m(X_i)-m_{L_n}(X_i))\right|^2\right\} \leq \frac{1}{\alpha_n^2}\cdot\mathbf{E}\int|m(x)-m_{L_n}(x)|^2\,\mathbf{P}_X(dx)$$
$$\leq \frac{1}{\alpha_n^2}\cdot c_{60}\cdot(\log L_n)^3\cdot L_n^{-\frac{2p}{2p+d^*}},$$

and the assumptions in Theorem 1 imply

$$\mathbf{E}\left\{\frac{1}{N_{1,n}}\sum_{i=1}^{N_{1,n}}\left|\frac{1}{\alpha_n}\cdot(Y_{n+L_n+i}-m(X_{n+L_n+i}))\right|^2\right\}$$
$$\leq 2\cdot\mathbf{E}\left\{\frac{1}{N_{1,n}}\sum_{i=1}^{N_{1,n}}\left|\frac{1}{\alpha_n}\cdot(Y_{n+L_n+i}-m^*(X_{n+L_n+i}))\right|^2\right\}$$
$$+2\cdot\mathbf{E}\left\{\frac{1}{N_{1,n}}\sum_{i=1}^{N_{1,n}}\left|\frac{1}{\alpha_n}\cdot(m^*(X_{n+L_n+i})-m(X_{n+L_n+i}))\right|^2\right\}$$
$$\leq 2\cdot\frac{(\alpha_n^*)^2}{\alpha_n^2}+2\leq 4.$$

We observe that by dividing the function space $\mathcal{H}^{(l)}_{K_2,M_{2,n},d,d^*,B_{2,n}}$ by α_n we change the μ_i in the last level of the hierarchical neural network. Since $\alpha_n \geq \left((\log L_n)^3\cdot L_n^{\frac{-2p}{2p+d^*}}\right)^{1/3}$ and $L_n \leq n^{c_4}$ the μ_i are bounded by

$$\frac{1}{\alpha_n}\cdot B_{2,n}\leq n^{c_{60}}.$$

Thus in the proof of Corollary 2 the bound on the covering number holds. Application of Corollary 2 yields

$$\mathbf{E}\int\left|\frac{1}{\alpha_n}\cdot\hat{m}_n^{\hat{\epsilon}}(x)-\frac{1}{\alpha_n}\cdot(m^*-m)(x)\right|^2\mathbf{P}_X(dx)$$
$$\leq 9\cdot\inf_{h\in\frac{1}{\alpha_n}\cdot\mathcal{H}^{(l)}_{K_2,M_{2,n},d,d^*,B_{2,n}}}\left(\int\left|h(x)-\frac{1}{\alpha_n}\cdot(m^*-m)(x)\right|^2\mathbf{P}_X(dx)\right)$$
$$+c_{61}\cdot w^{(n)}\cdot\left(\log(n)\cdot\frac{M_{2,n}}{n}+\frac{1}{\alpha_n^2}\cdot c_{62}\cdot(\log L_n)^3\cdot L_n^{-\frac{2p}{2p+d^*}}\right)$$
$$+c_{63}\cdot(1-w^{(n)})\cdot\left(\log(N_{1,n})\cdot\frac{M_{2,n}}{N_{1,n}}+4\right)+\frac{c_{64}}{n}.$$

Analogously as before we can bound the approximation error by using Theorem 3 in Bauer and Kohler (2017) and can conclude

$$\mathbf{E}\int\left|\frac{1}{\alpha_n}\cdot\hat{m}_n^{\hat{\epsilon}}(x)-\frac{1}{\alpha_n}\cdot(m^*-m)(x)\right|^2\mathbf{P}_X(dx)$$

$$
\begin{aligned}
&\leq c_{65} \cdot (\log n)^3 \cdot M_{2,n}^{-\frac{2p}{d^*}} + c_{66} \cdot w^{(n)} \cdot \left(\log(n) \cdot \frac{M_{2,n}}{n} + \frac{1}{\alpha_n^2} \cdot (\log L_n)^3 \cdot L_n^{-\frac{2p}{2p+d^*}} \right) \\
&\quad + c_{67} \cdot (1 - w^{(n)}) + \frac{c_{68}}{n}.
\end{aligned}
$$

The above results implies (37) which implies the assertion. □

5 Acknowledgment

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

References

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

[2] Bauer, B. and Kohler, M. (2017). On deep learning as a remedy for the curse of dimensionality in nonparametric regression. Submitted for publication.

[3] Bayarri, M. J., Berger, J. O., Cafeo, J., Garcia-Donato, G., Liu, F., Palomo, J., Parthasarathy, R. J., Paulo, R., Sacks, J., and Walsh, D. (2007). Computer model validation with functional output. *The Annals of Statistics*, **35**, pp. 1874–1906.

[4] Bichon, B., Eldred, M., Swiler, M., Mahadevan, S. and McFarland, J. (2008). Efficient global reliability analysis for nonlinear implicit performance functions. *AIAA Journal*, **46**, pp. 2459–2468.

[5] Bott, A. K., Felber, T., and Kohler, M. (2015). Estimation of a density in a simulation model. *Journal of Nonparametric Statistics*, **27**, pp. 271-285.

[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] Choi, S.-K., Grandhi, R. V. and Canfield, R. A. (2007). *Reliability-based Structural Design.* Springer-Verlag London Limited.

[9] Coombs, Clyde H. (1964). *A Theory of Data.* New York, John Wiley & Sons.

[10] 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.

[11] 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.

[12] Devroye, L., Felber, T., and Kohler, M. (2013). Estimation of a density using real and artificial data. *IEEE Transactions on Information Theory*, **59**, No. 3, pp. 1917-1928.

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

[14] Felber, T., Kohler, M., and Krzyżak, A. (2015a). Adaptive density estimation based on real and artificial data. *Journal of Nonparametric Statistics*, **27**, pp. 1-18.

[15] Felber, T., Kohler, M., and Krzyżak, A. (2015b). Density estimation with small measurement errors. *IEEE Transactions on Information Theory*, **61**, pp. 3446-3456.

[16] Goh, J., Bingham, D., Holloway, J. P., Grosskopf, M. J., Kuranz, C. C., and Rutter, E. (2013). Prediction and computer model calibration using outputs from mulitfidelity simulators. *Technometrics*, **55**, pp. 501-512.

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

[18] Götz, B., Schaeffner, M., Platz, R. and Melz, T. (2016). Lateral vibration attenuation of a beam with circular cross-section by a support with integrated piezoelectric transducers shunted to negative capacitances. *Smart Materials and Structures*, **25.9**, pp. 1–10.

[19] Han, G., Santner, T. J., Rawlinson, J. J. (2009). Simultaneous determination of tuning and calibration parameters for computer experiments. *Technometrics*, **51**, pp. 464-474.

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

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

[22] Higdon, D., Gattiker, J., Lawrence, E., Jackson, C., Tobis, M., Pratola, M., Habib, S., Heitmann, K., and Price, S. (2013). Computer model calibration using the ensemble kalman filter. *Technometrics*, **55**, pp. 488–500.

[23] Hurtado, J. (2004). *Structural Reliability – Statistical Learning Perspectives.* Vol. 17 of lecture notes in applied and computational mechanics. Springer.

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

[25] Kennedy, M. C., and O'Hagan, A. (2001). Bayesian calibration of computer models (with discussion). *Journal of the Royal Statistical Society: Series B*, **63**, pp. 425-464.

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

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

[28] Kohler, M., and Krzyżak, A. (2017b). Improving a surrogate model in uncertainty quantification by real data. Submitted for publication.

[29] Kohler, M., Krzyżak, A., Mallapur, S., and Platz, R. (2016). Uncertainty Quantification in Case of Imperfect Models: A Non-Bayesian Approach. To appear in *Scandinavian Journal of Statistics.*

[30] Li, S., Götz, B., Schaeffner, M. and Platz, R. (2017). Approach to prove the efficiency of the monte carlo method combined with the elementary effect method to quantify uncertainty of a beam structure with piezo–elastic supports. *Proceedings of the 2nd International Conference on Uncertainty Quantification in Computational Sciences and Engineering (UNCECOMP 2017)*, pp. 441–455.

[31] Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. *Annals of Mathematical Statistics*, **27**, pp. 832–837.

[32] 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.

[33] Parzen, E. (1962). On the estimation of a probability density function and the mode. *Annals of Mathematical Statistics*, **33**, pp. 1065–1076.

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

[35] Santner, T. J., Williams, B. J. and Notz, W. I. (2003). *The Design and Analysis of Computer Experiments.* Springer-Verlag, New York.

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

[37] Tuo, R., and Wu, C. F. J. (2015). Efficient calibration for imperfect computer models. *Annals of Statistics* **43**, pp. 2331-2352.

[38] Wang, S., Chen, W., and Tsui, K. L. (2009). Bayesian validation of computer models. *Technometrics*, **51**, pp. 439-451.

[39] Wong, R. K. W., Storlie, C. B., and Lee, T. C. M. (2017). A frequentist approach to computer model calibration. *Journal of the Royal Statistical Society, Series B*, **79**, pp. 635–648.

Supplementary material for the referees

Proof of Theorem 2. *In the first step of the proof* we show that we can assume w.l.o.g.

$$\bar{Y}_{i,n} \in [-\beta_n, \beta_n] \quad \text{for all} \quad i = 1, \dots, n + L_n. \tag{38}$$

To do this, we let

$$A_n = \left\{ |\bar{Y}_{i,n}| \leq \beta_n \quad \text{for all } i = 1, \dots, n + L_n \right\}$$

be the event that all $\bar{Y}_{i,n}$ be bounded in absolutely value by β_n. The union bound together with Markov inequality and (27) implies

$$\begin{aligned} \mathbf{P}(A_n^c) &\leq (n + L_n) \cdot \max_{i=1,\dots,n+L_n} \mathbf{P}\{|\bar{Y}_{i,n}| > \beta_n\} \leq (n + L_n) \cdot \frac{\max_{i=1,\dots,n+L_n} \mathbf{E}\left\{|\bar{Y}_{i,n}|^3\right\}}{\beta_n^3} \\ &\leq \frac{c_{69}}{n}. \end{aligned}$$

On the event A_n the estimate m_n coincides with the estimate $m_n^{(trunc)}$ defined by

$$\tilde{m}_n^{(trunc)}(\cdot) = \arg \min_{f \in \mathcal{F}_n} \left(\sum_{i=1}^{n+L_n} w_i \cdot |f(X_i) - T_{\beta_n} \bar{Y}_{i,n}|^2 + pen_n^2(f) \right)$$

and

$$m_n^{(trunc)}(x) = T_\beta \tilde{m}_n^{(trunc)}(x) \quad (x \in \mathbb{R}^d).$$

From this we can conclude that

$$\begin{aligned} &\mathbf{E} \int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx) \\ &\leq \mathbf{E} \left\{ \int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx) \cdot I_{A_n} \right\} + 4 \cdot \beta^2 \cdot \mathbf{P}(A_n^c) \\ &= \mathbf{E} \left\{ \int |m_n^{(trunc)}(x) - m(x)|^2 \mathbf{P}_X(dx) \cdot I_{A_n} \right\} + 4 \cdot \beta^2 \cdot \mathbf{P}(A_n^c) \\ &\leq \mathbf{E} \int |m_n^{(trunc)}(x) - m(x)|^2 \mathbf{P}_X(dx) + 4 \cdot \beta^2 \cdot \frac{c_{69}}{n}, \end{aligned}$$

which completes the first step of the proof.

So from now on we assume that (38) holds. Set

$$\gamma_n = w^{(n)} \cdot \delta_n + (1 - w^{(n)}) \cdot \delta_{L_n}$$

and

$$\begin{aligned} T_n &= \int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx) \\ &\quad - \left(9 \cdot \inf_{f \in \mathcal{F}_n} \left(pen_n^2(f) + \sum_{i=1}^{n+L_n} w_i \cdot |f(X_i) - m(X_i)|^2 \right) + 384 \cdot \sum_{i=1}^{n+L_n} w_i \cdot |Y_i - \bar{Y}_{i,n}|^2 \right). \end{aligned}$$

In the second step of the proof we show that the assertion follows from

$$\int_{12\cdot\gamma_n}^{4\cdot\beta^2} \mathbf{P}\{T_n > t\}\, dt \leq \frac{c_{70}}{n} + c_{71} \cdot \left(w^{(n)} \cdot \delta_n + (1-w^{(n)}) \cdot \delta_{L_n} \right).$$

To do this, we observe

$$\begin{aligned}
&\mathbf{E} \int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx) \\
&\leq \mathbf{E}\Bigg\{ \int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx) \\
&\qquad -\left(9 \cdot \inf_{f \in \mathcal{F}_n} \left(pen_n^2(f) + \sum_{i=1}^{n+L_n} w_i \cdot |f(X_i) - m(X_i)|^2 \right) + 384 \cdot \sum_{i=1}^{n+L_n} w_i \cdot |Y_i - \bar{Y}_{i,n}|^2 \right) \Bigg\} \\
&\qquad +\mathbf{E}\left\{ 9 \cdot \inf_{f \in \mathcal{F}_n} \left(pen_n^2(f) + \sum_{i=1}^{n+L_n} w_i \cdot |f(X_i) - m(X_i)|^2 \right) + 384 \cdot \sum_{i=1}^{n+L_n} w_i \cdot |Y_i - \bar{Y}_{i,n}|^2 \right\} \\
&\leq 12 \cdot \gamma_n + \int_{12\cdot\gamma_n}^{4\cdot\beta^2} \mathbf{P}\{T_n > t\} dt + 384 \cdot \mathbf{E}\left\{ \sum_{i=1}^{n+L_n} w_i \cdot |Y_i - \bar{Y}_{i,n}|^2 \right\} \\
&\qquad +9 \cdot \inf_{f \in \mathcal{F}_n} \left(pen_n^2(f) + \int |f(x) - m(x)|^2 \mathbf{P}_X(dx) \right) \\
&= 12 \cdot \gamma_n + \int_{12\cdot\gamma_n}^{4\cdot\beta^2} \mathbf{P}\{T_n > t\} dt + 384 \cdot \mathbf{E}\left\{ \sum_{i=1}^{n+L_n} w_i \cdot |Y_i - \bar{Y}_{i,n}|^2 \right\} \\
&\qquad +9 \cdot \inf_{f \in \mathcal{F}_n} \left(pen_n^2(f) + \int |f(x) - m(x)|^2 \mathbf{P}_X(dx) \right),
\end{aligned}$$

where the last equation holds since $T_n \leq 4\beta^2$. The definition of γ_n and of the weights implies the assertion of step 2.

In the third step of the proof we show that we have for $t > 0$

$$\mathbf{P}\{T_n > t\} \leq P_{1,n}(t) + P_{2,n}(t),$$

where

$$\begin{aligned}
P_{1,n}(t) \quad &= \quad \mathbf{P}\Bigg\{ \int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx) \\
&\qquad > \frac{t}{2} + 3 \cdot pen_n^2(\tilde{m}_n) + 3 \cdot \sum_{i=1}^{n+L_n} w_i \cdot |m_n(X_i) - m(X_i)|^2 \Bigg\}
\end{aligned}$$

and

$$P_{2,n}(t) \quad = \quad \mathbf{P}\Bigg\{ 3 \cdot \sum_{i=1}^{n+L_n} w_i \cdot |m_n(X_i) - m(X_i)|^2 + 3 \cdot pen_n^2(\tilde{m}_n)$$

$$
> \frac{t}{2} + 9 \cdot \inf_{f \in \mathcal{F}_n} \left(\sum_{i=1}^{n+L_n} w_i \cdot |f(X_i) - m(X_i)|^2 + pen_n^2(f) \right)
+384 \cdot \sum_{i=1}^{n+L_n} w_i \cdot |Y_i - \bar{Y}_{i,n}|^2 \Bigg\}.
$$

Using

$$
\begin{aligned}
&T_n \\
&= \left(\int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx) - 3 \cdot pen_n^2(\tilde{m}_n) - 3 \cdot \sum_{i=1}^{n+L_n} w_i \cdot |m_n(X_i) - m(X_i)|^2 \right) \\
&\quad + \Bigg(3 \cdot \sum_{i=1}^{n+L_n} w_i \cdot |m_n(X_i) - m(X_i)|^2 + 3 \cdot pen_n^2(\tilde{m}_n) \\
&\quad - \left(9 \cdot \inf_{f \in \mathcal{F}_n} \left(\sum_{i=1}^{n+L_n} w_i \cdot |f(X_i) - m(X_i)|^2 + pen_n^2(f) \right) + 384 \cdot \sum_{i=1}^{n+L_n} w_i \cdot |Y_i - \bar{Y}_{i,n}|^2 \right) \Bigg) \\
&=: T_{1,n} + T_{2,n}
\end{aligned}
$$

this immediately follows from

$$
\mathbf{P}\{T_n > t\} = \mathbf{P}\{T_{1,n} + T_{2,n} > t\} \leq \mathbf{P}\{T_{1,n} > t/2\} + \mathbf{P}\{T_{2,n} > t/2\}.
$$

In the fourth step of the proof we derive a upper bound on

$$
\int_{12 \cdot \gamma_n}^{4 \cdot \beta^2} P_{1,n}(t) \, dt.
$$

Let $12 \cdot \gamma_n \leq t \leq 4 \cdot \beta^2$. The definition of the weights together with

$$
a + b > c + d \quad \Rightarrow \quad (a > c \text{ or } b > d)
$$

implies that we have

$$
\begin{aligned}
P_{1,n}(t) \quad \leq \quad & \mathbf{P}\Bigg\{ w^{(n)} \cdot \int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx) > \frac{w^{(n)} \cdot \delta_n}{w^{(n)} \cdot \delta_n + (1 - w^{(n)}) \cdot \delta_{L_n}} \cdot \frac{t}{2} \\
& \qquad + w^{(n)} \cdot 3 \cdot pen_n^2(\tilde{m}_n) + w^{(n)} \cdot 3 \cdot \frac{1}{n} \cdot \sum_{i=1}^{n} |m_n(X_i) - m(X_i)|^2 \Bigg\} \\
& + \mathbf{P}\Bigg\{ (1 - w^{(n)}) \cdot \int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx) \\
& \qquad > \frac{(1 - w^{(n)}) \cdot \delta_{L_n}}{w^{(n)} \cdot \delta_n + (1 - w^{(n)}) \cdot \delta_{L_n}} \cdot \frac{t}{2} + (1 - w^{(n)}) \cdot 3 \cdot pen_n^2(\tilde{m}_n) \\
& \qquad + (1 - w^{(n)}) \cdot 3 \cdot \frac{1}{L_n} \cdot \sum_{i=n+1}^{n+L_n} |m_n(X_i) - m(X_i)|^2 \Bigg\}
\end{aligned}
$$

$$
\begin{aligned}
&\le \mathbf{P}\Bigg\{ \int |m_n(x)-m(x)|^2 \mathbf{P}_X(dx) > \frac{\delta_n}{w^{(n)}\cdot\delta_n + (1-w^{(n)})\cdot\delta_{L_n}} \cdot \frac{t}{2} \\
&\qquad\qquad +3\cdot pen_n^2(\tilde{m}_n) + 3\cdot\frac{1}{n}\cdot\sum_{i=1}^{n}|m_n(X_i)-m(X_i)|^2 \Bigg\} \\
&\quad +\mathbf{P}\Bigg\{ \int |m_n(x)-m(x)|^2 \mathbf{P}_X(dx) > \frac{\delta_{L_n}}{w^{(n)}\cdot\delta_n + (1-w^{(n)})\cdot\delta_{L_n}} \cdot \frac{t}{2} \\
&\qquad\qquad +3\cdot pen_n^2(\tilde{m}_n) + 3\cdot\frac{1}{L_n}\cdot\sum_{i=n+1}^{n+L_n}|m_n(X_i)-m(X_i)|^2 \Bigg\} \\
&= P_{1,n}^{(1)}(t) + P_{1,n}^{(2)}(t).
\end{aligned}
$$

Set

$$
\bar{\delta}_n := \frac{\delta_n}{w^{(n)}\cdot\delta_n + (1-w^{(n)})\cdot\delta_{L_n}} \cdot \frac{t}{12} \quad \text{and} \quad \bar{\delta}_{L_n} := \frac{\delta_{L_n}}{w^{(n)}\cdot\delta_n + (1-w^{(n)})\cdot\delta_{L_n}} \cdot \frac{t}{12}.
$$

We have

$$
\begin{aligned}
&P_{1,n}^{(1)}(t) \\
&= \mathbf{P}\left\{ \int |m_n(x)-m(x)|^2 \mathbf{P}_X(dx) > 6\cdot\bar{\delta}_n + 3\cdot pen_n^2(\tilde{m}_n) + 3\frac{1}{n}\sum_{i=1}^{n}|m_n(X_i)-m(X_i)|^2 \right\} \\
&\le \mathbf{P}\left\{ \int |m_n(x)-m(x)|^2 \mathbf{P}_X(dx) > \bar{\delta}_n + 3\cdot pen_n^2(\tilde{m}_n) + 3\frac{1}{n}\sum_{i=1}^{n}|m_n(X_i)-m(X_i)|^2 \right\}.
\end{aligned}
$$

Next we want to use Lemma 4 from Kohler and Krzyżak (2017b) on the above probability, where we replace β_n in the notation of Lemma 4 by β. The assumptions of the lemma are satisfied since (29) and (31) hold for every $k \ge n$. Thus

$$
\begin{aligned}
&\mathbf{P}\left\{ \int |m_n(x)-m(x)|^2 \mathbf{P}_X(dx) > \bar{\delta}_n + 3\cdot pen_n^2(\tilde{m}_n) + 3\frac{1}{n}\sum_{i=1}^{n}|m_n(X_i)-m(X_i)|^2 \right\} \\
&\le c_{72}\cdot\exp\left(-\frac{n\cdot\bar{\delta}_n}{c_{72}\cdot\beta^2}\right)
\end{aligned}
$$

holds. For $P_{1,n}^{(2)}$ we use an analogous transformation, apply Lemma 4, use the sample size L_n instead of n and replace again β_n by β and obtain

$$
\begin{aligned}
P_{1,n}^{(2)} \quad \le \quad & \mathbf{P}\Bigg\{ \int |m_n(x)-m(x)|^2 \mathbf{P}_X(dx) > \bar{\delta}_{L_n} + 3\cdot pen_n^2(\tilde{m}_n) \\
&\qquad\qquad +3\frac{1}{L_n}\sum_{i=n+1}^{n+L_n}|m_n(X_i)-m(X_i)|^2 \Bigg\}.
\end{aligned}
$$

$$\leq \quad c_{72} \cdot \exp\left(-\frac{L_n \cdot \bar{\delta}_{L_n}}{c_{72} \cdot \beta^2}\right).$$

The results for $P_{1,n}^{(1)}$ and $P_{1,n}^{(2)}$ are implying

$$\begin{aligned}
\int_{12\cdot\gamma_n}^{4\cdot\beta^2} P_{1,n}(t)dt \quad &\leq \quad \frac{c_{73} \cdot \beta^2}{n} \cdot \frac{w^{(n)} \cdot \delta_n + (1-w^{(n)}) \cdot \delta_{L_n}}{\delta_n} \cdot \exp\left(-\frac{n}{c_{73} \cdot \beta^2} \cdot \delta_n\right) \\
&\quad + \frac{c_{73} \cdot \beta^2}{L_n} \cdot \frac{w^{(n)} \cdot \delta_n + (1-w^{(n)}) \cdot \delta_{L_n}}{\delta_{L_n}} \cdot \exp\left(-\frac{L_n}{c_{73} \cdot \beta^2} \cdot \delta_{L_n}\right) \\
&\leq \quad \frac{c_{73}}{n} \cdot \frac{\delta_n}{\delta_n} \cdot \exp\left(-\frac{c_{22} \cdot \beta^2}{c_{73} \cdot \beta^2}\right) \\
&\quad + \frac{c_{73} \cdot \beta^2 \cdot \left(w^{(n)} \cdot \delta_n + (1-w^{(n)}) \cdot \delta_{L_n}\right)}{c_{22} \cdot \beta^2} \cdot \exp\left(-\frac{c_{22} \cdot \beta^2}{c_{73} \cdot \beta^2}\right) \\
&\leq \quad \frac{c_{74}}{n} + c_{75} \cdot \left(w^{(n)} \cdot \delta_n + (1-w^{(n)}) \cdot \delta_{L_n}\right)
\end{aligned}$$

where we have used $\delta_n \geq \delta_{L_n} > 0$, and that (29) implies

$$\delta_n \cdot n > c_{22} \cdot \beta^2 \quad \text{and} \quad \delta_{L_n} \cdot L_n > c_{22} \cdot \beta^2.$$

In the fifth step of the proof we derive a upper bound on

$$\int_{12\cdot\gamma_n}^{4\cdot\beta^2} P_{2,n}(t)\, dt.$$

Since $|m(x)| \leq \beta \leq \beta_n$ ($x \in \mathbb{R}^d$) and $w_i \geq 0$ ($i \in \{1, \dots, n + L_n\}$) we have

$$\sum_{i=1}^{n+L_n} w_i \cdot |m_n(X_i) - m(X_i)|^2 \leq \sum_{i=1}^{n+L_n} w_i \cdot |T_{\beta_n}\tilde{m}_n(X_i) - m(X_i)|^2$$

which implies

$$\begin{aligned}
P_{2,n}(t) \quad &\leq \quad \mathbf{P}\Bigg\{ \sum_{i=1}^{n+L_n} w_i \cdot |T_{\beta_n}\tilde{m}_n(X_i) - m(X_i)|^2 + pen_n^2(\tilde{m}_n) \\
&\qquad > \frac{t}{6} + 3 \cdot \inf_{f \in \mathcal{F}_n} \left(\sum_{i=1}^{n+L_n} w_i \cdot |f(X_i) - m(X_i)|^2 + pen_n^2(f) \right) \\
&\qquad + 128 \cdot \sum_{i=1}^{n+L_n} w_i \cdot |Y_i - \bar{Y}_{i,n}|^2 \Bigg\}.
\end{aligned}$$

Choose $m_n^* \in \mathcal{F}_n$ such that

$$3 \cdot \left(\sum_{i=1}^{n+L_n} w_i \cdot |m_n^*(X_i) - m(X_i)|^2 + pen_n^2(m_n^*) \right)$$

$$\leq 3 \cdot \inf_{f \in \mathcal{F}_n} \left(\sum_{i=1}^{n+L_n} w_i \cdot |f(X_i) - m(X_i)|^2 + pen_n^2(f) \right) + \frac{t}{12}$$

Then we can conclude by Lemma 1 from Kohler and Krzyżak (2017b) that the above probability is bounded by

$$\begin{aligned}
&\mathbf{P}\Bigg\{ \sum_{i=1}^{n+L_n} w_i \cdot |T_{\beta_n}\tilde{m}_n(X_i) - m(X_i)|^2 + pen_n^2(\tilde{m}_n) \\
&\qquad \geq \frac{t}{12} + 3 \cdot \left(\sum_{i=1}^{n+L_n} w_i \cdot |m_n^*(X_i) - m(X_i)|^2 + pen_n^2(m_n^*) \right) + 128 \cdot \sum_{i=1}^{n+L_n} w_i \cdot |Y_i - \bar{Y}_{i,n}|^2 \Bigg\} \\
&\leq \mathbf{P}\Bigg\{ \sum_{i=1}^{n+L_n} w_i \cdot (T_{\beta_n}\tilde{m}_n(X_i) - m_n^*(X_i)) \cdot (Y_i - m(X_i)) \\
&\qquad \geq \frac{1}{24} \cdot \left(\sum_{i=1}^{n+L_n} w_i \cdot |T_{\beta_n}\tilde{m}_n(X_i) - m_n^*(X_i)|^2 + pen_n^2(\tilde{m}_n) \right) + \frac{t}{72} \Bigg\}
\end{aligned}$$

The definition of the weights together with

$$a + b > c + d \quad \Rightarrow \quad (a > c \text{ or } b > d)$$

implies

$$\begin{aligned}
&\mathbf{P}\Bigg\{ \sum_{i=1}^{n+L_n} w_i \cdot (T_{\beta_n}\tilde{m}_n(X_i) - m_n^*(X_i)) \cdot (Y_i - m(X_i)) \\
&\qquad \geq \frac{1}{24} \cdot \left(\sum_{i=1}^{n+L_n} w_i \cdot |T_{\beta_n}\tilde{m}_n(X_i) - m_n^*(X_i)|^2 + pen_n^2(\tilde{m}_n) \right) + \frac{t}{72} \Bigg\} \\
&\leq \mathbf{P}\Bigg\{ \frac{w^{(n)}}{n} \sum_{i=1}^{n} (T_{\beta_n}\tilde{m}_n(X_i) - m_n^*(X_i)) \cdot (Y_i - m(X_i)) \\
&\qquad \geq \frac{w^{(n)}}{24} \cdot \left(\frac{1}{n} \sum_{i=1}^{n} |T_{\beta_n}\tilde{m}_n(X_i) - m_n^*(X_i)|^2 + pen_n^2(\tilde{m}_n) \right) + w^{(n)} \cdot \frac{\bar{\delta}_n}{6} \Bigg\} \\
&\quad + \mathbf{P}\Bigg\{ \frac{(1 - w^{(n)})}{L_n} \sum_{i=n+1}^{n+L_n} (T_{\beta_n}\tilde{m}_n(X_i) - m_n^*(X_i)) \cdot (Y_i - m(X_i)) \\
&\qquad \geq \frac{(1 - w^{(n)})}{24} \cdot \left(\frac{1}{L_n} \sum_{i=n+1}^{n+L_n} |T_{\beta_n}\tilde{m}_n(X_i) - m_n^*(X_i)|^2 + pen_n^2(\tilde{m}_n) \right) + (1 - w^{(n)}) \cdot \frac{\bar{\delta}_{L_n}}{6} \Bigg\} \\
&\leq \mathbf{P}\Bigg\{ \frac{1}{n} \sum_{i=1}^{n} (T_{\beta_n}\tilde{m}_n(X_i) - m_n^*(X_i)) \cdot (Y_i - m(X_i)) \\
&\qquad \geq \frac{1}{24} \cdot \left(\frac{1}{n} \sum_{i=1}^{n} |T_{\beta_n}\tilde{m}_n(X_i) - m_n^*(X_i)|^2 + pen_n^2(\tilde{m}_n) \right) + \frac{1}{6} \cdot \bar{\delta}_n \Bigg\}
\end{aligned}$$

$$+ \mathbf{P}\Bigg\{ \frac{1}{L_n} \sum_{i=n+1}^{n+L_n} (T_{\beta_n}\tilde{m}_n(X_i) - m_n^*(X_i)) \cdot (Y_i - m(X_i))$$

$$\geq \frac{1}{24} \cdot \left(\frac{1}{L_n} \sum_{i=n+1}^{n+L_n} |T_{\beta_n}\tilde{m}_n(X_i) - m_n^*(X_i)|^2 + pen_n^2(\tilde{m}_n) \right) + \frac{1}{6} \cdot \bar{\delta}_{L_n} \Bigg\}$$

$$= P_{2,n}^{(1)}(t) + P_{2,n}^{(2)}(t).$$

Next we want to use Lemma 3 from Kohler and Krzyżak (2017b) in order to bound $P_{2,n}^{(1)}(t)$. The assumptions of the Lemma are satisfied since (29) and (30) hold for every $k \geq n$. Thus

$$P_{2,n}^{(1)}(t) \leq c_{76} \cdot \exp\left(-\frac{n \cdot \min\{\bar{\delta}_n, \sigma_0^2\}}{c_{76}} \right)$$

and because of $t \leq 4\beta^2$ we can w.l.o.g. assume that $\sigma_0^2 \geq \bar{\delta}_n$ holds. Thus

$$P_{2,n}^{(1)}(t) \leq c_{76} \exp\left(-\frac{n\bar{\delta}_n}{c_{76}} \right)$$

and by the same arguments we can apply Lemma 3 from Kohler and Krzyżak (2017) and obtain

$$P_{2,n}^{(2)}(t) \leq c_{77} \exp\left(-\frac{L_n\bar{\delta}_{L_n}}{c_{77}} \right).$$

Analogously as before $\delta_n \geq \delta_{L_n} > 0$ and

$$\delta_n \cdot n > c_{22} \quad \text{and} \quad \delta_{L_n} \cdot L_n > c_{22}$$

implies

$$\int_{12\cdot\gamma_n}^{4\cdot\beta^2} P_{2,n}(t)dt \leq \frac{c_{78}}{n} + c_{79} \cdot \left(w^{(n)} \cdot \delta_n + (1 - w^{(n)}) \cdot \delta_{L_n} \right).$$

Summarizing the above results we get the assertion. □