Uncertainty quantification based on (imperfect) simulation models with estimated input distributions [*]

Sebastian Kersting[†] and Michael Kohler
Fachbereich Mathematik, Technische Universität Darmstadt, Schlossgartenstr. 7, 64289 Darmstadt, Germany, email: kersting@mathematik.tu-darmstadt.de, kohler@mathematik.tu-darmstadt.de

August 06, 2019

Abstract
In this article we study uncertainty quantification of a technical system. We propose new density estimates which combine observed data of the technical system and simulated data from an (imperfect) simulation model based on estimated input distributions. We analyze the rate of convergence of these 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: Density estimation, estimated input distributions, L_1 error, simulation models, surrogate models, uncertainty quantification.

1 Introduction

We consider the problem of quantifying the uncertainty in an experiment with a technical system. This experiment is described by an $\mathbb{R}^d \times \mathbb{R}$-valued random variable (X, Y), where Y is the outcome of the experiment and the so-called input variable X describes "parameters" of the experiment. For example if one wants to analyze in an experiment the maximal relative compression Y of a spring damper component it is known that it is dependent on the free fall height and the spring stiffness which leads to a two dimensional input variable X.

We assume that Y has a density g with respect to the Lebesgue measure and our aim is to find an estimator $\hat{g}\colon \mathbb{R} \to \mathbb{R}$ such that the L_1 error

$$\int_{\mathbb{R}} |\hat{g}(x) - g(x)| dx$$

[*] Running title: *Uncertainty quantification with estimated input distributions*

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

is small. Since

$$\int_{\mathbb{R}} |\hat{g}(x) - g(x)| dx = 2 \cdot \sup_{B \in \mathcal{B}} \left| \int_B \hat{g}(x) dx - \int_B g(x) dx \right|$$

(cf. Theorem 5.1 in Devroye and Lugosi (2000)) such an approximation of g will allow us to estimate for each Borel set $B \subseteq \mathbb{R}$ the probability

$$\mathbf{P}\{Y \in B\} = \int_B g(x) dx \quad \text{by} \quad \int_B \hat{g}(x) dx$$

such that the maximal occurring error is small.

If an independent and identically distributed sample $Y_1, \dots, Y_n$ is available, one possibility to do this is to use the Rosenblatt-Parzen kernel density estimate

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

(c.f. Rosenblatt (1956) and Parzen (1962)). Here $K\colon \mathbb{R}^d \to \mathbb{R}$ (so-called kernel, which is assumed to be a density) and $h_n > 0$ (so-called bandwidth) are parameters of the estimate. But in many applications in engineering the sample size n is too small to apply such an estimate, because the experiments with technical systems are rather time consuming or expensive. Alternatively one could assume that the distribution of Y is an element of a known class of distributions which can be characterized by a parameter, i.e. $\mathbf{P}_Y \in \{w_\vartheta \colon \vartheta \in \Theta\}$, and estimate this parameter and thus the density of Y by a so-called maximum likelihood estimate (cf., e.g., Kalbfleisch (1979)). In any application the class of distributions of Y is usually not known. The standard approach would be to assume that Y is normally distributed, but for instance in the above example the maximal relative compression Y of a spring damper component is an extreme value and 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 it is not a normal distribution.

Our estimate will be based on the choice of a model for the input X described by a random variable $\bar{X}$ and a simulation model described by a function $m\colon \mathbb{R}^d \to \mathbb{R}$, both chosen such that $m(\bar{X})$ is in some sense a good approximation of Y. Here engineering knowledge is used to construct the simulation model $m\colon \mathbb{R}^d \to \mathbb{R}$, e.g. it could be the solution of a partial differential equation system. And the model for X is constructed on the basis of observed values of X.

We distinguish between two data models:

(i) In the first model we assume that our simulation model is perfect in sense that

$$Y = m(X) \tag{2}$$

holds, and that we have observed an independent and identically distributed sample

$$X_1, \dots, X_n \tag{3}$$

of X which we use to construct $\bar{X}$.

(ii) In our second model our simulation model is imperfect in the sense that we have

$$Y \neq m(X),$$

but we have observed an identically and independent distributed sample

$$(X_1, Y_1), \ldots, (X_n, Y_n) \tag{4}$$

of (X, Y). Furthermore we assume that there exisits a function $m^* \colon \mathbb{R}^d \to \mathbb{R}$ such that $Y = m^*(X)$ holds.

In the first data model we have no sample of Y available, but as in (1) we can use the simulation model and the input data and estimate the density of Y by

$$\hat{g}(y) = \frac{1}{n \cdot h_n} \sum_{i=1}^{n} K\left(\frac{y - m(X_i)}{h_n}\right).$$

In most applications the sample size n will be too small to achieve a good approximation of g. Alternatively we can use our sample of input values to construct a sample of $\bar{X}$. Then we can apply the estimate to a large independent and identically distributed sample

$$\bar{X}_1, \ldots, \bar{X}_{N_n} \tag{5}$$

and estimate g by

$$\hat{g}(y) = \frac{1}{N_n \cdot h_{N_n}} \sum_{i=1}^{N_n} K\left(\frac{y - m(\bar{X}_i)}{h_{N_n}}\right). \tag{6}$$

Usually, the simulation model is evaluated using a computer program. In most cases the evaluation of the simulation with a computer program is rather time consuming, so that it is not feasible to run the computer experiments with a large sample and consequently the density estimate (6) can not be applied with N_n large. Instead, one has to apply techniques which are able to quantify the uncertainty in the computer experiment using only a few evaluations of the computer program. There is a vast literature on the design and analysis of such computer experiments, cf., e.g., Santner, Williams and Notz (2003) or Fang, Li and Sudjianto (2010). There so-called surrogate models of the computer experiment are used. Thus we estimate a surrogate model $\hat{m}_n$ of m and use it to estimate the density of g by

$$\hat{g}(y) = \frac{1}{N_n \cdot h_{N_n}} \sum_{i=1}^{N_n} K\left(\frac{y - \hat{m}_n(\bar{X}_i)}{h_{N_n}}\right). \tag{7}$$

In the second data model a sample of output data Y is available. As described above the standard approach in modern statistics would be to use a nonparametric estimate of the density g of Y, e.g. the classical kernel density estimate, cf. (1). However in most applications the sample size n will be too small to achieve satisfying results. As in the first data model one could also use the simulation model or a surrogate model of it to estimate the density of Y on a sample of $\bar{X}$, as described by (6) and (7). Since the simulation model is imperfect in this data model, the surrogate model will also be imperfect and thus $\hat{m}_n(X)$ will possibly not be a good approximation of Y. Consequently a density estimate based on a surrogate model will not achieve good approximation results if the error of the surrogate model is large. In this article we circumvent this problem by using the data set (4) together with the simulation model m to construct an improved surrogate model and by estimating the density g of Y as in (7).

Our main results are as follows: In Theorem 1 below we present a general result on the expected L_1 error of our density estimate of g, which shows how the expected L_1 error depends on the error of the estimation of the distribution of X and of the error of the surrogate model $\hat{m}_n$. In case of normally distributed X we demonstrate how we can estimate its parameters such that the error of the resulting estimate of X achieves the parametric rate of convergence $n^{-1/2}$. We use both results to show in Corollary 1 that in the first data model and for normally distributed X our density estimate of g can achieve the parametric rate $n^{-1/2}$ in case of a general (smooth) density g. Furthermore we analyze the error of the density estimate (7) in the second data model. Here we show that in case that the error of our simulation model m (considered as an estimate of m^*) is small we get a rate of convergence of the density estimate, which depends on this error and on the smoothness of $m - m^*$, and which can (even in case of a large dimension d of X) be simultaneously smaller than the error of the density estimates (1) and (6). Hence in this case the combination of the observed values of the technical system together with the simulation model leads to an estimate which is better than the standard estimates using the observed values of the technical system or the simulation model alone.

1.1 Discussion of related results

Estimation of surrogate methods models have been introduced and investigated with the aid of the simulated and real data by several authors using a broad range of estimation techniques. First Bucher and Bourgund (1990), Kim and Na (1997) and Das and Zheng (2000). Later on Hurtado (2004), Deheeger and Lemaire (2010) and Bourinet, Deheeger and Lemaire (2011) investigated surrogate models in context of support vector machines and Papadrakakis and Lagaros (2002) concentrated on neural networks. Kaymaz (2005) and Bichon et al. (2007) used kriging. 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. (2018).

Kohler and Krzyżak (2017b) introduced a method to estimate an improved surrogate model and showed a result for smoothing spline estimates. The method uses only a very small sample of experimental data which is combined with a sample generated by computer experiments. Götz, Kersting and Kohler (2018) extended the method to least squares estimates and applied it to neural networks. Thus they where also able to apply it to high-dimensional settings, where smoothing spline estimates usually fail to deliver reasonable results because of the curse of dimensionality. In contrast to the results presented in our article these estimate need to assume that a large quantity of input values X is given or that they can be generated, i.e. their distribution is known, which is often not satisfied in an application.

1.2 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 $x \in \mathbb{R}^d$ we denote the i-th component of x by $x^{(i)}$. For a vector $v \in \mathbb{R}^d$

$$\|v\|_\infty = \max_{1 \leq i \leq d} |v^{(i)}|$$

is its supremum norm and $\|v\|$ is its Euclidean norm. For $f : \mathbb{R}^d \to \mathbb{R}$ and $B \subseteq \mathbb{R}^d$

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

is its supremum norm on B, where if $B = \mathbb{R}^d$ we write $\|f\|_{\infty,\mathbb{R}^d} = \|f\|_\infty$. For a matrix $A \in \mathbb{R}^{m \times n}$, where $A = (a_{ij})_{1 \leq i \leq m, 1 \leq j \leq n}$

$$\|A\|_\infty = \sqrt{m \cdot n} \cdot \max_{1 \leq i \leq m, 1 \leq j \leq n} |a_{ij}| \quad \text{and} \quad \|A\|_F = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} |a_{i,j}|^2}$$

is its supremum norm and its Frobenius norm, respectively.

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

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

If A is a set, then $\mathbb{1}_A$ is the indicator function corresponding to A, i.e. the function which takes on the value 1 on A and is zero elsewhere, and $\lambda(A)$ denotes its Lebesgue measure (in case $A \subseteq \mathbb{R}^d$).

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

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

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

Let $p = k + \beta$ for some $k \in \mathbb{N}_0$ and $0 < \beta \leq 1$, and let $C > 0$. We say that a function $f : \mathbb{R}^d \to \mathbb{R}$ is (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 $\frac{\partial^k f}{\partial x_1^{\alpha_1} \dots \partial x_d^{\alpha_d}}$ exists and satisfies

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

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

1.3 Outline

The outline of this paper is as follows: In Section 2 we show a general result for density estimates based on surrogate models and estimated input distribution. In Section 3 we introduce a method to generate a sample of input values based on an estimated distribution and show a result for its L_1 convergence rate. In Sections 4 and 5 we show results for a density estimate based on an (imperfect) simulation model and estimated input distributions. The finite sample size performance of our estimates is illustrated in Section 6 by applying the estimates to simulated data.

2 A general result

In the following we show a result for the general case, where we estimate the density g using a sample of $\bar{X}$ and a surrogate model $\hat{m}_n$ of m. Here we assume that we have available two data sets $\mathcal{D}_n^{(1)}$ and $\mathcal{D}_n^{(2)}$. In a first step we construct an estimate $\hat{f}_n$ of the density f by the data set $\mathcal{D}_n^{(1)}$. Then we generate an independent and identically distributed sample

$$\bar{X}_1, \dots, \bar{X}_{N_n} \tag{8}$$

of size N_n, such that $\hat{f}_n$ is its density. Next we construct a surrogate estimate $\hat{m}_n \colon \mathbb{R}^d \to \mathbb{R}$ of m by the sample $\mathcal{D}_n^{(2)}$. In this setting the following theorem concerning the L_1 rate of convergence of the density estimate

$$\hat{g}_{N_n}(y) = \frac{1}{N_n \cdot h_{N_n}} \sum_{i=1}^{N_n} K\left(\frac{y - \hat{m}_n(\bar{X}_i)}{h_{N_n}} \right). \tag{9}$$

of g holds, where $h_{N_n} > 0$ and $K \colon \mathbb{R} \to \mathbb{R}$.

Theorem 1. *Let $d, N_n \in \mathbb{N}$. Let $(X,Y),(X_1,Y_1),\ldots$ be independent and identically distributed $\mathbb{R}^d \times \mathbb{R}$ valued random variables. Let f be the density of X and g be the density of Y, and assume that g is (r,C)-smooth for some $r \in (0,1]$ and some $C > 0$. Let $S_n \subseteq \mathbb{R}$ be compact. Let $\hat{f}_n$ and the data* (8) *be defined as above, and assume that $\mathcal{D}_n^{(1)}$ and $\mathcal{D}_n^{(2)}$ are independent of* (8) *and of (X_{n+1},Y_{n+1}), (X_{n+2},Y_{n+2}), $\ldots$ Let $h_{N_n} > 0$ and let $K\colon \mathbb{R} \to \mathbb{R}$ be a symmetric and bounded density satisfying*

$$\int_{\mathbb{R}} K^2(u)\,du < \infty \quad \text{and} \quad \int_{\mathbb{R}} K(u)\cdot |u|^r\,du < \infty.$$

Define the estimate $\hat{g}_{N_n}$ of g by (9).

Then there exists $c_1, c_2, c_3 \in \mathbb{R}_+$ such that

$$\mathbf{E}\int_{\mathbb{R}} |\hat{g}_{N_n}(y) - g(y)|dy \leq 2\cdot\int_{S_n^c} g(y)dy + \frac{c_1\cdot\sqrt{\lambda(S_n)}}{\sqrt{N_n\cdot h_{N_n}}} + c_2\cdot\lambda(S_n)\cdot h_{N_n}^r$$
$$+\mathbf{E}\int |\hat{f}_n(x) - f(x)|\,dx + \frac{c_3}{h_{N_n}}\sqrt{\mathbf{E}\left\{|\hat{m}_n(X) - Y|^2\right\}}.$$

Remark 1. In the first data model with suitable assumptions on the tail probability of Y and S_n growing fast enough the first term on the right-hand side is neglectable. Also with suitable smoothness assumptions on m the last term on the right-hand side decreases for an increasing sample size of $\mathcal{D}_n^{(2)}$ and is insignificant for the rate. Finally if we choose N_n large enough and h_{N_n} small enough the second and third term on the right-hand side are also neglectable. Consequently the rate of convergence only depends on the rate of the density estimate $\hat{f}_n$.

Remark 2. In our the second data model we will see in Corollary 2 below that for normally distributed X and an appropriate choice of h_{N_n} and N_n the expected L_1 error of $\hat{g}_{N_n}$ is bounded by some constant times

$$\max\left\{n^{-1/2}, (\log n)\cdot\left(\mathbf{E}\left\{|\hat{m}_n(X) - Y|^2\right\}\right)^{\frac{r}{2r+2}}\right\}.$$

3 Estimation of the input distribution

If the distribution of X is an element of a parametric class of distributions, then it is possible to estimate its parameters (e.g., by maximum likelihood), and to use a technique especially designed for this parametric class to generate a sample of the corresponding distribution (cf., e.g., Devroye (1986)). In the sequel we demonstrate how this can be done in case of a multivariate normal distribution. Here we estimate the mean μ and variance Σ of X given the sample (3) by

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{10}$$

and

$$\hat{\Sigma} = \left(\frac{1}{n}\sum_{k=1}^{n}(X_k^{(i)} - \hat{\mu}^{(i)})(X_k^{(j)} - \hat{\mu}^{(j)})\right)_{1\leq i,j\leq d}. \tag{11}$$

In order to generate a sample

$$\bar{X}_1, \dots, \bar{X}_{N_n} \tag{12}$$

of size $N_n \in \mathbb{N}$, which is independent and normally distributed with mean $\hat{\mu}$ and covariance matrix $\hat{\Sigma}$, we consider the eigendecomposition

$$\hat{\Sigma} = \hat{O}\hat{\Lambda}\hat{O}^T$$

of $\hat{\Sigma}$. Here $\hat{\Lambda} = \text{diag}(\hat{\lambda}_1, \dots, \hat{\lambda}_d)$ is a diagonal matrix consisting of eigenvalues of $\hat{\Sigma}$ and $\hat{O}$ is a orthogonal matrix whose columns are eigenvectors of $\hat{\Sigma}$. Then we generate an independent sample $Z_1, \dots, Z_{N_n}$ of d-dimensional vectors, where for each vector the components are independent and standard normally distributed, and set for every $i = 1, \dots, N_n$

$$\bar{X}_i = \hat{O}\hat{\Lambda}^{1/2} Z_i + \hat{\mu}. \tag{13}$$

It is easy to see that $\bar{X}_1, \dots, \bar{X}_{N_n}$ are independent and multivariate normally distributed with mean $\hat{\mu}$ and covariance $\hat{\Sigma}$. We denote the density of $\bar{X}_1$ by $\hat{f}_n$. For this estimate the following lemma concerning the L_1 rate of convergence holds:

Lemma 1. *Let $d, n \in \mathbb{N}$. Let $X, X_1, \dots$ independent and multivariate normally distributed with mean vector μ and positive definite covariance matrix Σ. Let f be the density of X. Estimate $\hat{\mu}$ by* (10) *and $\hat{\Sigma}$ by* (11) *and let $\hat{f}_n$ the density of $\bar{X}_1$ defined as above. Then there exists a constant $c_4 \in \mathbb{R}_+$ such that*

$$\mathbf{E} \int_{\mathbb{R}} |\hat{f}_n(x) - f(x)| \, dx \leq c_4 \cdot n^{-1/2}$$

holds.

4 Uncertainty quantification in case of perfect simulation models

In this section we consider uncertainty quantification in our first data model. Here we want to estimate the density of real valued random variable Y which depends on an $\mathbb{R}^d$-valued random variable X. We have available a perfect simulation model $m \colon \mathbb{R}^d \to \mathbb{R}$, satisfying $Y = m(X)$ and an independent and identically distributed sample

$$X_1, \dots, X_n \tag{14}$$

of X. We will use this sample to estimate the density $f \colon \mathbb{R}^d \to \mathbb{R}$ of X and based on this estimate $\hat{f}_n$ we will generate an independent and identically distributed sample

$$\bar{X}_1, \dots, \bar{X}_{N_n} \tag{15}$$

as described in Section 3. Based on this sample and a surrogate model for m we will then estimate the density of Y by (9).

Our estimate uses a neural network as a surrogate for the simulation model. To construct this neural network 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 theoretical results and applications below 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 $\gamma > 0$, 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

$$|\mu_i| \le \gamma, \quad |\lambda_{i,j}| \le \gamma, \quad |\theta_{i,j,v}| \le \gamma$$

for all $i \in \{0, 1, \dots, M\}$, $j \in \{0, \dots, 4d^*\}$ and $v \in \{0, \dots, d\}$, by $\mathcal{F}^{(\text{neural networks})}_{M,d,d^*,\gamma}$. We will use the following recursively defined classes of neural networks (with parameters I, M, d, $d^* \in \mathbb{N}$ and $\gamma > 0$): For $l = 0$, we define our space of hierarchical neural networks by

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

For $l > 0$, we define recursively

$$\begin{aligned} \mathcal{H}^{(l)}_{I,M,d,d^*,\gamma} \quad = \quad & \Bigg\{ h \colon \mathbb{R}^d \to \mathbb{R},\, h(x) = \sum_{k=1}^{I} g_k(f_{1,k}(x), \dots, f_{d^*,k}(x)) \quad (x \in \mathbb{R}^d) \\ & \qquad \text{for some } g_k \in \mathcal{F}^{(\text{neural networks})}_{M,d^*,d^*,\gamma} \text{ and } f_{j,k} \in \mathcal{H}^{(l-1)}_{I,M,d,d^*,\gamma} \Bigg\}. \end{aligned} \tag{16}$$

We start constructing the estimate by defining a surrogate estimate of our simulation model m. To do this we generate a sample of size $L_n \in \mathbb{N}$ consisting of independent and uniformly on $B_n := [-c_5 \cdot (\log L_n), c_5 \cdot (\log L_n)]^d$ distributed random variables $U_{1,n}, \dots, U_{L_n,n}$, which are independent of all other random variables mentioned before. Next we define our surrogate estimate

$$\hat{m}_{L_n}(\cdot) = \hat{m}_{L_n}(\cdot, (U_{1,n}, m(U_{1,n})), \dots, (U_{L_n,n}, m(U_{L_n,n}))) \colon \mathbb{R}^d \to \mathbb{R}$$

of the simulation model m by a least squares neural network estimate given by

$$\tilde{m}_{L_n}(\cdot) = \arg \min_{f \in \mathcal{H}^{(l)}_{I_1, M_{L_n}, d, d^*, \gamma_{L_n}}} \frac{1}{L_n} \sum_{i=1}^{L_n} |f(U_{i,n}) - m(U_{i,n})|^2, \tag{17}$$

where $I_1, M_{L_n}, d^* \in \mathbb{N}$ and $\gamma_{L_n} > 0$ 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 convergence of this estimate we need to truncate the estimate at some height $\beta_n > 0$, i.e., we define

$$\hat{m}_{L_n}(x) = T_{\beta_n}(\tilde{m}_{L_n}(x)) \quad (x \in \mathbb{R}^d), \tag{18}$$

where $T_{\beta_n}(z) = \operatorname{sign}(z) \cdot \min\{|z|, \beta_n\}$ for $z \in \mathbb{R}$.

Next we define our density estimate $\hat{g}_{N_n} \colon \mathbb{R} \to \mathbb{R}$ of g by applying a kernel density estimate on the sample $\hat{m}_{L_n}(\bar{X}_1), \dots, \hat{m}_{L_n}(\bar{X}_{N_n})$. Therefore we choose a kernel $K \colon \mathbb{R} \to \mathbb{R}$ and a bandwidth $h_{N_n} > 0$ and define $\hat{g}_{N_n}$ by (9) with $\hat{m}_n$ replaced by $\hat{m}_{L_n}$.

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 $I \in \mathbb{N}$, $g_k \colon \mathbb{R}^{d^*} \to \mathbb{R}$ $(k = 1, \dots, I)$ and $f_{1,k}, \dots, f_{d^*,k} \colon \mathbb{R}^d \to \mathbb{R}$ $(k = 1, \dots, I)$ such that $f_{1,k}, \dots, f_{d^*,k}$ $(k = 1, \dots, I)$ satisfy a generalized hierarchical interaction model of order d^* and level l and*

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

c) *We say that a* **generalized hierarchical interaction model** *is* **(p, C)-smooth**, *if all functions f and g_k occurring in its definition are* **(p, C)-smooth**.

In order to prove our main result of this section we will make the following assumptions:

(A1) The random variable X has a density $f : \mathbb{R}^d \to \mathbb{R}$ (with respect to the Lebesgue measure) which is bounded by some constant, i.e., which satisfies

$$\|f\|_\infty \leq c_6 \tag{19}$$

for some $c_6 \in \mathbb{R}_+$.

(A2) The random variable Y satisfies $Y = m(X)$ for some measurable function $m : \mathbb{R}^d \to \mathbb{R}$ and has a density $g : \mathbb{R} \to \mathbb{R}$ which is (r, C)-smooth for some $r \in (0, 1]$ and some $C > 0$.

(A3) The function $m : \mathbb{R}^d \to \mathbb{R}$ in (A2) satisfies a (p, C)-smooth generalized hierarchical interaction model of order d^* and finite level l with $p = q + s$, where $q \in \mathbb{N}_0$ and $s \in (0, 1]$. Here in the definition of this generalized hierarchical interaction model all partial derivates of order less than or equal to q of the functions g_k, f of this generalized hierarchical interaction model are bounded, i.e., each such function f satisfies

$$\max_{\substack{j_1,\dots,j_d \in \{0,1,\dots,q\} \\ j_1+\dots+j_d \le q}} \left\| \frac{\partial^{j_1+\dots+j_d} f}{\partial^{j_1} x^{(1)} \dots \partial^{j_d} x^{(d)}} \right\|_\infty \le c_7, \tag{20}$$

and all functions g_k are Lipschitz continuous with Lipschitz constant $\tilde{L} > 0$.

(A4) The function $m : \mathbb{R}^d \to \mathbb{R}$ satisfies

$$\|m\|_{\infty, B_n} \le \beta_n, \tag{21}$$

where $B_n = [-c_5 \cdot \log(L_n), c_5 \cdot \log(L_n)]^d$ and $1 \le \beta_n \le L_n^{c_8}$ for some constant $c_8 \in (0, 1]$.

Here assumptions $(A1)$ and $(A4)$ enable us to estimate the surrogate model based on observations of the simulation model at x-values uniformly distributed on B_n, assumption $(A2)$ is our smoothness assumption on the density of $Y = m(X)$, and assumption $(A3)$ is the main smoothness assumption on the simulation model.

Theorem 2. *Let $d, n, L_n, N_n \in \mathbb{N}$. Let $X, X_1, \dots$ be independent and identically distributed $\mathbb{R}^d$-valued random variables, let $m : \mathbb{R}^d \to \mathbb{R}$ and assume that (A1)–(A4) hold.*

Let $\hat{f}_n$ be an estimate of f based on the sample (14) *and generate the sample* (15) *such that its density is $\hat{f}_n$. Let $\sigma \colon \mathbb{R} \to [0, 1]$ be the logistic squasher $\sigma(x) = 1/(1 + \exp(-x))$ $(x \in \mathbb{R})$. Let $U_{1,n}, \dots, U_{L_n,n}$ be independent and uniformly distributed on B_n and define the surrogate estimate $\hat{m}_{L_n}$ by* (17) *and* (18)*, where we choose I_1, d and d^* as in the definition of the generalized hierarchical interaction model for m and set $M_{L_n} = \left\lceil L_n^{\frac{d^*}{2p+d^*}} \right\rceil$ and $\gamma_{L_n} = L_n^{c_9}$.*

Assume that $K \colon \mathbb{R} \to \mathbb{R}$ is a symmetric and bounded density satisfying

$$\int_{\mathbb{R}} K^2(u)\, du < \infty \quad \text{and} \quad \int_{\mathbb{R}} K(u) \cdot |u|^r \, du < \infty,$$

and define the estimate $\hat{g}_{N_n}$ of g by (9) with $\hat{m}_n$ replaced by $\hat{m}_{L_n}$.

Then there exists some constants $c_{10}, c_{11}, c_{12} \in \mathbb{R}_+$ such that

$$\begin{aligned} &\mathbf{E} \int_{\mathbb{R}} |\hat{g}_{N_n}(y) - g(y)| dy \\ &\le 2 \cdot \int_{S_n^c} g(y) dy + \frac{c_{10} \cdot \sqrt{\lambda(S_n)}}{\sqrt{N_n \cdot h_{N_n}}} + c_{11} \cdot \lambda(S_n) \cdot h_{N_n}^r + \mathbf{E} \int |\hat{f}_n(x) - f(x)| dx \\ &\quad + \frac{c_{12}}{h_{N_n}} \left(\beta_n^4 \cdot \lambda(B_n) \cdot (\log L_n)^{4p+6} \cdot L_n^{-\frac{2p}{2p+d^*}} + \beta_n^2 \cdot \int_{\mathbb{R}^d \setminus B_n} f(x)\, dx \right. \end{aligned}$$

$$+\int_{\mathbb{R}^d \setminus B_n} m(x)^2 \, \mathbf{P}_X(dx) \Bigg)^{1/2}$$

holds for L_n sufficiently large.

In case X multivariate normally distributed the following corollary holds:

Corollary 1. *Let $X, X_1, \dots, X_n$ be independent and multivariate normally distributed with expectation $\mu \in \mathbb{R}^d$ and positive definite covariance matrix $\Sigma \in \mathbb{R}^{d \times d}$. Estimate $\hat{\mu}$ by* (10) *and $\hat{\Sigma}$ by* (11)*, and let $\hat{f}_n$ be the multivariate normal density with mean vector $\hat{\mu}$ and covariance matrix $\hat{\Sigma}$. Assume that the assumptions of Theorem 2 are satisfied and furthermore that $\mathbf{E}\{|Y|\} < \infty$ holds. Set $B_n = [-c_5 \cdot \log(L_n), c_5 \cdot \log(L_n)]^d$ and $S_n = [-n^{1/2}, n^{1/2}]$. Set*

$$h_{N_n} = n^{-\frac{1}{r}} \quad \text{and} \quad \beta_n = c_{13} \cdot \log(L_n).$$

Assume that $L_n, N_n \in \mathbb{N}$ are chosen such that the following inequalities hold:

$$L_n \geq \left((\log L_n)^{4p+d+10} \cdot n^{\frac{2+r}{r}} \right)^{\frac{2p+d^*}{2p}}, \quad N_n \geq n^{\frac{3r+2}{2r}}$$

and

$$\int_{\mathbb{R}^d \setminus B_n} m(x)^2 \, \mathbf{P}_X(dx) \leq \lambda(B_n) \cdot (\log L_n)^{4p+10} \cdot L_n^{-\frac{2p}{2p+d^*}}.$$

Then for some constant $c_{14} \in \mathbb{R}_+$

$$\mathbf{E} \int_{\mathbb{R}} |\hat{g}_{N_n}(y) - g(y)| dy \leq c_{14} \cdot n^{-1/2}$$

holds for L_n sufficiently large.

Remark 3. Corollary 1 shows that in case of a perfect simulation model a parametric assumption on the density of X leads to the parametric rate $n^{-1/2}$ for the estimation of the density g of Y, even if this density is not contained in a parametric class of densities.

5 Uncertainty quantification in case of imperfect simulation models

In this section we consider uncertainty quantification in the second data model. I.e. we want to estimate the density of a real valued random variable Y where we know that there exists a functional relationship such that for an $\mathbb{R}^d$-valued random variable X and some measurable function $m^* : \mathbb{R}^d \to \mathbb{R}$

$$Y = m^*(X) \tag{22}$$

holds. We have available an imperfect simulation model $m_{sim,n} : \mathbb{R}^d \to \mathbb{R}$ with

$$Y \neq m_{sim,n}(X)$$

and an independent and identically distributed sample

$$(X_1, Y_1), \dots, (X_n, Y_n) \tag{23}$$

of (X, Y). We will use this sample to estimate the density $f\colon \mathbb{R}^d \to \mathbb{R}$ of X and based on this estimate $\hat{f}_n$ we will generate an independent and identically distributed sample

$$\bar{X}_1, \dots, \bar{X}_{N_n} \tag{24}$$

as described in Section 3. Based on our imperfect simulation model $m_{sim,n}$ and the sample (23) we will estimate an improved surrogate model, which we will evaluate on the sample (24) in order to estimate the density of Y.

Therefore we will next present a method to estimate an improved surrogate model. We generate an independent and uniformly on $B_n := [-c_5 \cdot \log(L_n), c_5 \cdot \log(L_n)]^d$ distributed sample

$$U_{1,n}, \dots, U_{L_n,n} \tag{25}$$

of size L_n independent of all other random variables mentioned before, and define our surrogate estimate $\hat{m}_{L_n}$ by

$$\tilde{m}_{L_n}(\cdot) = \arg \min_{f \in \mathcal{H}^{(l)}_{I_1, M_{L_n}, d, d^*, \gamma_{L_n}}} \frac{1}{L_n} \sum_{i=1}^{L_n} |f(U_{i,n}) - m_{sim,n}(U_{i,n})|^2 \tag{26}$$

and

$$\hat{m}_{L_n}(x) = T_{\beta_n}(\tilde{m}_{L_n}(x)) \quad (x \in \mathbb{R}^d). \tag{27}$$

Next we define an estimate on basis of the residuals

$$\epsilon_i = Y_i - \hat{m}_{L_n}(X_i) \quad (i = 1, \dots, n), \tag{28}$$

by a least squares neural network estimate

$$\tilde{m}^{\epsilon}_n(\cdot) = \arg \min_{f \in \mathcal{H}^{(l)}_{I_2, M_n, d, d^*, \gamma_n}} \frac{1}{n} \sum_{i=1}^{n} |f(X_i) - \epsilon_i|^2, \tag{29}$$

where $I_2, M_n, d^* \in \mathbb{N}$ and $\gamma_n > 0$ are parameters of the estimate. We set

$$\hat{m}^{\epsilon}_n(x) = T_{c_{15} \cdot \alpha_n}(\tilde{m}^{\epsilon}_n(x)) \quad (x \in \mathbb{R}^d), \tag{30}$$

where $c_{15} \geq 1$ and $\alpha_n > 0$. We define our final surrogate model $(X, \hat{m}_n(X))$ for (X, Y) by

$$\hat{m}_n(x) = \hat{m}_{L_n}(x) + \hat{m}^{\epsilon}_n(x) \quad (x \in \mathbb{R}^d), \tag{31}$$

and estimate the density g of Y by applying a kernel density estimate to a sample of $\hat{m}_n(\bar{X})$. Therefore we choose a kernel $K\colon \mathbb{R} \to \mathbb{R}$ and a bandwidth $h_{N_n} > 0$ and define $\hat{g}_{N_n}$ by (9).

To formulate the main theorem of this section we need assumption $(A1)$, the following modifications of $(A2)$, $(A3)$ and $(A4)$ and the additional assumption $(A5)$.

(A2*) The random variable Y satisfies $Y = m^*(X)$ for some measurable function $m^* : \mathbb{R}^d \to \mathbb{R}$ and has a density $g : \mathbb{R} \to \mathbb{R}$ which is (r, C)-smooth for some $r \in (0, 1]$ and some $C > 0$.

(A3*) The function $m_{sim,n} : \mathbb{R}^d \to \mathbb{R}$ satisfies a (p, C)-smooth generalized hierarchical interaction model of order d^* and finite level l with $p = q + s$, where $q \in \mathbb{N}_0$ and $s \in (0, 1]$. Here in the definition of this generalized hierarchical interaction model all partial derivates of order less than or equal to q of the functions g_k, f of this generalized hierarchical interaction model are bounded, and all functions g_k are Lipschitz continuous with Lipschitz constant $\tilde{L} > 0$.

(A4*) The function $m_{sim,n} : \mathbb{R}^d \to \mathbb{R}$ satisfies

$$\|m_{sim,n}\|_{\infty,B_n} \leq \beta_n, \tag{32}$$

where $B_n = [-c_5 \cdot \log(L_n), c_5 \cdot \log(L_n)]^d$ and $1 \leq \beta_n \leq L_n^{c_8}$ for some constant $c_x \in (0, 1]$.

(A5) Let $0 < \alpha_n \leq 1$ and assume that

$$\|m^* - m_{sim,n}\|_\infty \leq \alpha_n. \tag{33}$$

Furthermore assume that $\frac{1}{\alpha_n}(m^* - m_{sim,n}) \colon \mathbb{R}^d \to \mathbb{R}$ satisfies a (p, C)-smooth generalized hierarchical interaction model of order d^* and finite level l with $p = q + s$, where $q \in \mathbb{N}_0$ and $s \in (0, 1]$. Assume that in Definition 1 b) all partial derivates of order less than or equal to q of the functions g_k, f of this generalized hierarchical interaction model are bounded, and let all functions g_k be Lipschitz continuous with Lipschitz constant $\tilde{L} > 0$.

Theorem 3. *Let $d, n, L_n, N_n \in \mathbb{N}$ with $2 \leq n \leq L_n$. Let $(X, Y), (X_1, Y_1), \ldots$ be independent and identically distributed $\mathbb{R}^d \times \mathbb{R}$ valued random variables. Assume that assumptions (A1), (A2*), (A3*), (A4*) and (A5) hold. Generate the sample (24) such that its density is $\hat{f}_n$. Assume that $\mathbf{E}\{|Y|\} < \infty$.*

Let $\sigma \colon \mathbb{R} \to [0, 1]$ be the logistic squasher $\sigma(x) = 1/(1 + \exp(-x))$ $(x \in \mathbb{R})$. Let $U_{1,n}, \ldots, U_{L_n,n}$ be independent and uniformly distributed on

$$B_n := [-c_5 \cdot \log(L_n), c_5 \cdot \log(L_n)]^d$$

and define the surrogate estimate $\hat{m}_{L_n}$ by (26) and (27), where we choose I_1, d and d^* as in the definition of the generalized hierarchical interaction model for $m_{sim,n}$ (and assume that these values are independent of n) and set $M_{L_n} = \left\lceil L_n^{\frac{d^*}{2p+d^*}} \right\rceil$ and $\gamma_{L_n} = L_n^{c_{16}}$.

Assume that

$$c_{17} \cdot \Bigg(\beta_n^2 \cdot \lambda(B_n) \cdot (\log L_n)^{4p+6} L_n^{-\frac{2p}{2p+d^*}} + \beta_n^2 \cdot \int_{\mathbb{R}^d \setminus B_n} f(x)\, dx$$

$$+ \int_{\mathbb{R}^d \setminus B_n} m_{sim,n}(x)^2 \, \mathbf{P}_X(dx) \Bigg) \leq \frac{\alpha_n^3}{\beta_n}, \tag{34}$$

$$\int_{\mathbb{R}^d \setminus B_n} |m_{sim,n}(x)|^3 \, \mathbf{P}_X(dx) \leq c_{18} \cdot \alpha_n^3 \tag{35}$$

and

$$\int_{\mathbb{R}^d \setminus B_n} f(x)\, dx \leq c_{19} \cdot \frac{\beta_n^3}{\alpha_n^3} \tag{36}$$

holds.

Define the estimate of the residuals $\hat{m}_n^\epsilon$ *by* (29) *and* (30)*, where we choose* I_2*,* d *and* d^* *as in the hierarchical interaction model for* $(m^* - m_{sim,n})/\alpha_n$ *(and assume that these values are independent of* n*) and set* $M_n = \left\lceil n^{\frac{d^*}{2p+d^*}} \right\rceil$ *and* $\gamma_n = n_n^{c_{20}}$*. Furthermore define the improved surrogate estimate by*

$$\hat{m}_n(x) = \hat{m}_{L_n}(x) + \hat{m}_n^\epsilon(x) \quad (x \in \mathbb{R}^d). \tag{37}$$

Let $S_n \subseteq \mathbb{R}$*, let* $h_{N_n} > 0$ *and define the estimate* $\hat{g}_{N_n}$ *of* g *by* (9)*.*

Then there exists constants $c_{21}, c_{22}, c_{23} \in \mathbb{R}_+$ *such that*

$$\begin{aligned}
&\mathbf{E} \int_{\mathbb{R}} |\hat{g}_{N_n}(y) - g(y)| dy \\
&\leq 2 \cdot \int_{S_n^c} g(y) dy + \frac{c_{21} \cdot \sqrt{\lambda(S_n)}}{\sqrt{N_n \cdot h_{N_n}}} + c_{22} \cdot \lambda(S_n) \cdot h_{N_n}^r + \mathbf{E} \int |\hat{f}_n(x) - f(x)| dx \\
&\quad + \frac{c_{23}}{h_{N_n}} \Bigg(\alpha_n^2 \cdot (\log n)^{4p+6} \cdot n^{-\frac{2p}{2p+d^*}} + \frac{\alpha_n^2}{n} + (\alpha_n^2 \cdot n + \beta_n^2) \cdot \int_{\mathbb{R}^d \setminus B_n} f(x)\, dx \\
&\qquad + \int_{\mathbb{R}^d \setminus B_n} m_{sim,n}(x)^2 \mathbf{P}_X(dx) + \beta_n^2 \cdot \lambda(B_n) \cdot (\log L_n)^{4p+6} \cdot L_n^{-\frac{2p}{2p+d^*}} \Bigg)^{1/2}
\end{aligned}$$

holds for n *sufficiently large.*

In case of X multivariate normally distributed we get the following corollary.

Corollary 2. *Let* $X, X_1, \dots, X_n$ *be independent and multivariate normally distributed with expectation* $\mu \in \mathbb{R}^d$ *and positive definite covariance matrix* $\Sigma \in \mathbb{R}^{d \times d}$*. Estimate* $\hat{\mu}$ *by* (10) *and* $\hat{\Sigma}$ *by* (11)*. Assume that the assumptions of Theorem 3 are satisfied and that in addition*

$$\mathbf{E}\{\exp(c_{24} \cdot |Y|)\} < \infty$$

holds. Assume furthermore that $\alpha_n \leq \beta_n$*. Set* $B_n := [-c_5 \cdot \log(L_n), c_5 \cdot \log(L_n)]^d$ *and* $S_n = [-c_{25} \cdot \log(n), c_{25} \cdot \log(n)]$*. Set* $\beta_n = c_{13} \cdot \log(L_n)$ *and*

$$h_{N_n} = \left(\alpha_n \cdot (\log n)^{4p+6} \cdot n^{-\frac{p}{2p+d^*}} \right)^{\frac{1}{r+1}}$$

where $c_{13} \in \mathbb{R}_+$. *Furthermore assume that*

$$\max\left\{\lambda(B_n)\cdot(\log L_n)^{4p+8}\cdot L_n^{-\frac{2p}{2p+d^*}}, \int_{\mathbb{R}^d\setminus B_n} m_{sim,n}(x)^2\mathbf{P}_X(dx)\right\} \leq \alpha_n^2\cdot(\log n)^{4p+6}\cdot n^{-\frac{2p}{2p+d^*}}$$

and

$$N_n \geq n^{c_{26}}\cdot\left(\alpha_n\cdot(\log n)^{4p+6}\cdot n^{-\frac{p}{2p+d^*}}\right)^{-\frac{1}{r+1}}$$

holds. Then for some constant $c_{27} \in \mathbb{R}_+$

$$\mathbf{E}\int_{\mathbb{R}} |\hat{g}_{N_n}(y) - g(y)|dy \leq c_{27}\cdot\max\left\{n^{-1/2}, (\log n)\cdot\left(\alpha_n\cdot(\log n)^{4p+6}\cdot n^{-\frac{p}{2p+d^*}}\right)^{\frac{r}{r+1}}\right\}$$

holds for n *sufficiently large.*

6 Application to simulated data

In the following a simulation study considering the second data model of Section 11 is realized. The implementation of the density estimate introduced in Section 5 which is based on an improved surrogate model is described and its performance is analyzed by applying it to simulated data. In the supplementary material we apply our method on a real world example.

In the simulation study we consider the following setting. We choose the dimension d as 5 and X multivariate standard normally distributed. The dependent variable Y is defined by

$$Y = m^*(X)$$

for some $m^*\colon \mathbb{R}^5 \to \mathbb{R}$. We set

$$m(x) = m^*(x) + \sigma_m\cdot\lambda^*,$$

where $\sigma_m \in \{0.1, 0.2, 0.5\}$ and $\lambda^* > 0$ is selected as the empirical interquartile range of $m^*(X)$.

We consider four different functions for $m^*\colon \mathbb{R}^5 \to \mathbb{R}$. In each case we use sample sizes $n = 10$, $L_n = 200$ and $N_n = N_{1,n} + N_{2,n}$, where $N_{1,n} = 200$ and $N_{2,n} = 10^4$. The different functions used as m^* are the following:

$$\begin{aligned}
m_1^*(x) =& 2\cdot\log(|x_1\cdot x_2| + 4\cdot\sin(x_3)^2 + |\tan(x_4)| + 0.1) + \cos(\sqrt{|x_3|}\cdot x_5^2 - x_1\cdot x_3)\\
m_2^*(x) =& x_1 + \frac{\cot(|x_2| + 0.002) + x_3^3 + \log(|x_4| + 0.1)}{9\pi} + 3\cdot x_5\\
m_3^*(x) =& \frac{2}{|x_1| + 0.1} + 3\cdot\log(x_2^6 + 0.2)\cdot x_4 + \frac{x_5}{|x_1| + 0.1}\\
m_4^*(x) =& \frac{10}{(1 + x_1^2)} + 5\cdot\sin(x_3\cdot x_4) + 2\cdot x_5 + \exp(x_1) + x_2^2 + \sin(x_3\cdot x_4)^2 - 10
\end{aligned}$$

As mentioned before, the parameter λ^* is chosen as the empirical interquartile range of $m^*(X)$ calculated on 10^7 realizations of X. The used values are $\lambda_1^* = 1.65$, $\lambda_2^* = 4.32$, $\lambda_3^* = 7.27$ and $\lambda_4^* = 5.86$.

We estimate $\hat{\mu}$ by (10) and $\hat{\Sigma}$ by (11). Based on these estimates we generate the sample (12) by the *MATLAB* function *mvnrnd()*.

Our improved surrogate estimate is defined by combining two least squares neural network estimates $\hat{m}_{L_n}$ and $\hat{m}_n^\epsilon$. For reasons of simplicity we will neglect the truncation of the estimates in the implementation. To improve the performance of the estimate we will use the following generalization of the least squares estimate $\hat{m}_n^\epsilon$. We split the sample (12) in a sample of size $N_{n,1} \in \mathbb{N}$ and $N_{n,2} = N_n - N_{n,1}$ and use the following weighted least squares estimate

$$\hat{m}_n^\epsilon(\cdot) = \arg \min_{f \in \mathcal{H}^{(l)}_{I_2, M_n, d, d^*, \gamma_n}} \left(\frac{w^{(n)}}{n} \sum_{i=1}^{n} |f(X_i) - \epsilon_i|^2 + \frac{(1 - w^{(n)})}{N_{n,1}} \sum_{i=1}^{N_{n,1}} |f(\bar{X}_i) - 0|^2 \right), \qquad (38)$$

where $w^{(n)} \in [0, 1]$. Here the additional function values of $\bar{X}_1, \dots, \bar{X}_{N_{n,1}}$ are compared with 0, which can be seen as a form of regularization, based on the assumption that the surrogate estimate $\hat{m}_{L_n}$ is almost perfect. In the case that $w^{(n)} = 1$ this estimate coincides with the estimate introduced in Section 5. For both cases we use the in Section 4 introduced class of neural networks, however the network parameters are chosen differently. For both estimates we neglect the bounds on the weights (,i.e $\gamma_n = \infty$ and $\gamma_{L_n} = \infty$). For $\hat{m}_{L_n}$ we choose the parameters 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. We calculate the least squares estimate by solving (17) approximately using the Levenberg-Marquard algorithm implemented in the *MATLAB* routine *lsqnonlin()*. Then we consider the parameter combination with the smallest occurring L_2 risk evaluated on the test data. The parameters are chosen from the sets $l \in \{0, 1, 2\}$, $I_1 \in \{1, 2\}$, $d^* \in \{1, \dots, d\}$ and $M_{L_n} \in \{1, \dots, 5, 6, 16, \dots, 46\}$.

Since the data set $(X_1, Y_1), \dots, (X_n, Y_n)$ is quite small we consider as network parameters for $\hat{m}_n^\epsilon$ only the sets $l \in \{0\}$, $I_2 \in \{1\}$, $d^* \in \{1, 2, 4\}$, $M_n \in \{1, 3, 5\}$ and the additional weighting parameter $w^{(n)}$ is chosen also data dependent from $\{0, 0.25, \dots, 1\}$. For the parameter selection we use a 5-fold cross validation. Again we calculate the least squares estimate by solving (38) approximately by the Levenberg-Marquard algorithm. To calculate the density estimate g_{N_n} we use the remaining part of data set (12) of size $N_{n,2}$. Consequently we denote the density estimate by $g_{N_{n,2}}$ and our density estimate of the density of Y is defined by

$$\hat{g}_{N_{n,2}}(y) = \frac{1}{N_{n,2} \cdot h_{N_{n,2}}} \sum_{i=N_{n,1}+1}^{N_{n,1}+N_{n,2}} K\left(\frac{y - \hat{m}_n(\bar{X}_i)}{h_{N_{n,2}}} \right). \qquad (39)$$

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. (1). 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) a surrogate model of the simulation model m as defined in (17) is used. For the third estimate (est. 3) the surrogate model is chosen as a least squares neural network estimate trained on n realizations of (X, Y), i.e.

$$\hat{m}_n^{(\text{est. } 3)}(\cdot) = \arg \min_{f \in \mathcal{H}^{(l)}_{I_1, M_{L_n}, d, d^*, \gamma_{L_n}}} \frac{1}{n} \sum_{i=1}^{n} |f(X_i) - Y_i|^2. \tag{40}$$

The estimates are compared by their L_1 error. Therefore it is necessary that the real density of Y is available. 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 were the real density. To calculate the L_1 error 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 1 the median (and in brackets the interquartile range) of the 50 L_1 errors.

	σ_m	0.1	0.2	0.5
m_1^*	est. 1	0.422 (0.232)	0.407 (0.219)	0.456 (0.227)
	est. 2	0.408 (0.188)	0.469 (0.280)	0.688 (0.269)
	est. 3	0.691 (0.340)	0.685 (0.407)	0.649 (0.344)
	est. 4	**0.387 (0.194)**	**0.387 (0.186)**	**0.454 (0.221)**
m_2^*	est. 1	0.362 (0.154)	0.399 (0.263)	**0.306 (0.169)**
	est. 2	0.318 (0.213)	0.391 (0.233)	0.612 (0.247)
	est. 3	0.564 (0.292)	0.556 (0.316)	0.506 (0.252)
	est. 4	**0.314 (0.202)**	**0.348 (0.244)**	0.356 (0.205)
m_3^*	est. 1	0.456 (0.246)	0.439 (0.221)	0.409 (0.157)
	est. 2	0.313 (0.214)	0.443 (0.225)	0.643 (0.277)
	est. 3	0.658 (0.259)	0.642 (0.217)	0.660 (0.309)
	est. 4	**0.296 (0.186)**	**0.383 (0.199)**	**0.384 (0.321)**
m_4^*	est. 1	0.302 (0.238)	0.425 (0.195)	0.328 (0.231)
	est. 2	0.250 (0.177)	0.312 (0.206)	0.571 (0.239)
	est. 3	0.539 (0.237)	0.597 (0.410)	0.572 (0.311)
	est. 4	**0.228 (0.163)**	**0.298 (0.228)**	**0.279 (0.197)**

Table 1: Median (and interquartile range) of the L_1 error of the four different estimates for the four different models with a constant error in the computer model and five percent noise

Our newly proposed estimate outperforms the other three estimates in 11 of 12 cases and it always outperforms the other surrogate models (est. 2) and (est. 3). The resulting L_1 error of (est. 3) is in any simulation higher than the error of the other estimates. We assume this is due to the complexity of the used functions m^* and the small sample size of 10.

7 Supplementary Material

The Supplementary Material contains an application of the in Section 5 introduced method on a real world example and all proofs.

8 Acknowledgment

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - Projektnummer 57157498 - SFB 805.

References

Bauer, B. and Kohler, M. (2019). On deep learning as a remedy for the curse of dimensionality in nonparametric regression. *The Annals of Statistics* **47**, pp. 654–694.

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.

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.

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.

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.

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

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

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.

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

Devroye, L. (1986). *Non-Uniform Random Variate Generation.* Springer-Verlag, New York.

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.

Devroye, L. and Lugosi, G. (2000). Combinatorial Methods in Density Estimation. *Springer-Verlag*, New York.

Devroye, L., Mehrabian, A. and Reddad, T. (2019). The total variation distance between high-dimensional Gaussians. *arXiv:1810.08693 [math.ST]*

Fang, K.-T., Li, R. and Sudjianto, A. (2010). *Design and modeling for computer experiments.* Boca Raton: Chapman & Hall.

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.

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

Gänssler, P, and Stute, W. (1977). *Probability theory* (in german). Springer, New York.

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 multifidelity simulators. *Technometrics* **55**, pp. 501–512.

Götz, B., Kersting, S. and Kohler, M. (2018). Estimation of an improved surrogate model in uncertainty quantification by neural networks. *Submitted for publication.*

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.

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

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.

Hurtado, J. (2004). *Structural Reliability Statistical Learning Perspectives.* Vol. 17 of lecture notes in applied and computational mechanics. Springer.0

Kalbfleisch, J. G. (1979). *Probability and statistical inference. II.* Universitext, Springer-Verlag, New York-Heidelberg.

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

Kim, S.-H. and Na, S.-W. (1997). Response surface method using vector projected sampling points. *Structural Safety* **19**, pp. 319.

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.

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

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

Kohler, M., Krzyzak, A., Mallapur, S., and Platz, R. (2018). Uncertainty Quantification in Case of Imperfect Models: A Non-Bayesian Approach. *Scandinavian Journal of Statistics* **45**, pp. 729–752.

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.

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.

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

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

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

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

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

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

Application to real data

As a real world example we consider the lateral vibration attenuation system with piezo–elastic supports described in Figure 1. 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 and B support is used for lateral beam vibration attenuation, as proposed in Götz et al. (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.

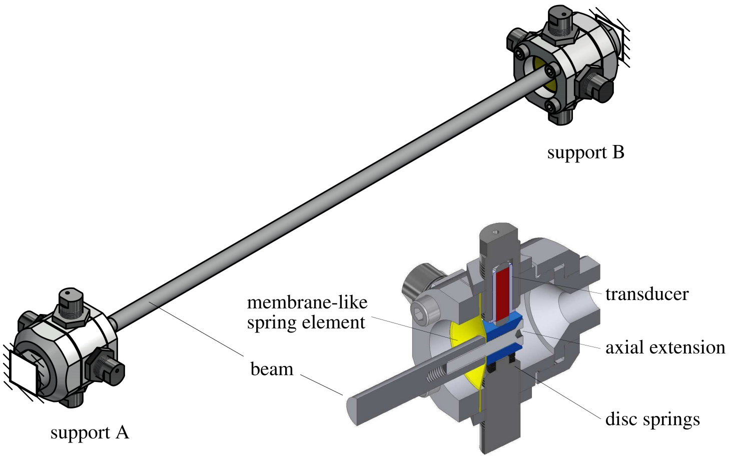

Figure 1: A CAD model of the lateral vibration attenuation system with piezo–elastic supports and a sectional view of one of the piezo–elastic supports, cf. Li et al. (2017).

Our aim is to predict the maximal amplitude of the vibration occurring in an experiment with this attenuation system. 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). A physical computer model is available with which we can compute the maximal vibration amplitude to a corresponding input value. To apply our estimate we measured the corresponding parameters for the ten built systems. As a result we got the data in Table 2.

Since the parameters vary in scale, it does not make sense to estimate the surrogate model $\hat{m}_{L_n}$ on $U_{i,n} \sim U([-c_5 \cdot \log(L_n), c_5 \cdot \log(L_n)]^d)$. Instead we rescale the components

	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 2: 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 $[\frac{m}{s^2}/V]$.

of $U_{i,n}$ such that for each component $U_{i,n}^{(j)} \sim U([\hat{\mu}^{(j)} - 2 \cdot \sqrt{\hat{\sigma}_{jj}}, \hat{\mu}^{(j)} + 2 \cdot \sqrt{\hat{\sigma}_{jj}}])$ holds.

We apply the four estimates described in Section 6 to the given data and obtain as an result Figure 2.

As discussed in the introduction, the distribution of extreme values is characterized by a non-symmetric distribution about the most likely value. This characteristic is described by the (est. 2) and our (est. 4), whereas the (est. 4) predicts higher values. If one considers the experimental data this is a plausible correction by the residual estimate $\hat{m}_n^{\epsilon}$.

Proof of Theorem 1

Scheffés Lemma implies that

$$\begin{aligned}\mathbf{E} \int_{\mathbb{R}} |\hat{g}_{N_n}(y) - g(y)| dy &\leq 2 \cdot \mathbf{E} \int_{S_n} (g(y) - \hat{g}_{N_n}(y))_+ \, dy + 2 \cdot \int_{S_n^c} g(y)\, dy. \\ &\leq 2 \cdot \mathbf{E} \int_{S_n} |g(y) - \hat{g}_{N_n}(y)| \, dy + 2 \cdot \int_{S_n^c} g(y)\, dy.\end{aligned}$$

Set

$$\hat{g}_{\hat{m}_n(X),N_n}(y) = \frac{1}{N_n \cdot h_{N_n}} \cdot \sum_{i=n+1}^{n+N_n} K\left(\frac{y - \hat{m}_n(X_i)}{h_{N_n}}\right)$$

and

$$\mathcal{D}_n = \mathcal{D}_n^{(1)} \cup \mathcal{D}_n^{(2)}.$$

By the triangle inequality

$$\begin{aligned}&\mathbf{E} \int_{S_n} |\hat{g}_{N_n}(y) - g(y)| dy \\ &\leq \mathbf{E} \int_{S_n} |\hat{g}_{N_n}(y) - \mathbf{E}\left\{\hat{g}_{N_n}(y) \,\middle|\mathcal{D}_n\right\}| dy + \mathbf{E} \int_{S_n} |\mathbf{E}\left\{\hat{g}_{N_n}(y) \,\middle|\mathcal{D}_n\right\} \\ &\quad -\mathbf{E}\left\{\hat{g}_{\hat{m}_n(X),N_n}(y)\middle|\mathcal{D}_n\right\}| dy + \mathbf{E} \int_{S_n} |\mathbf{E}\left\{\hat{g}_{\hat{m}_n(X),N_n}(y)\middle|\mathcal{D}_n\right\} - g(y)| dy.\end{aligned}$$

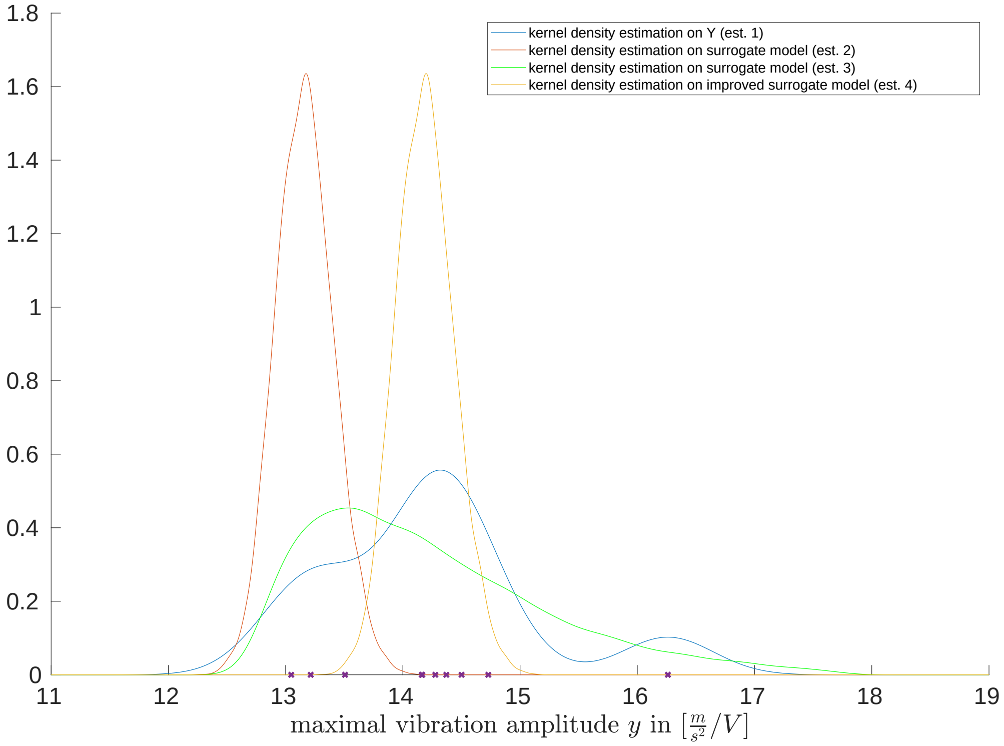

Figure 2: Four different density estimates and as reference the data $Y_1, \ldots, Y_n$ indicated on the x axis.

With Fubini's theorem and the Cauchy-Schwarz inequality the first term is bounded by

$$\begin{aligned}
&\mathbf{E} \int_{S_n} |\hat{g}_{N_n}(y) - \mathbf{E}\left\{\hat{g}_{N_n}(y)\,\middle|\mathcal{D}_n\right\}|dy \\
&= \int_{S_n} \mathbf{E}\{\mathbf{E}\{|\hat{g}_{N_n}(y) - \mathbf{E}\left\{\hat{g}_{N_n}(y)\,\middle|\mathcal{D}_n\right\}|\,\middle|\mathcal{D}_n\right\}\}dy \\
&\leq \int_{S_n} \mathbf{E}\left\{\sqrt{\mathbf{V}\left\{\hat{g}_{N_n}(y)\middle|\mathcal{D}_n\right\}}\right\} dy \\
&= \mathbf{E}\left\{\int_{S_n} \sqrt{\mathbf{V}\left\{\hat{g}_{N_n}(y)\middle|\mathcal{D}_n\right\}} dy\right\} \\
&\leq \sqrt{\lambda(S_n)} \cdot \mathbf{E}\left\{\left(\int_{S_n} \mathbf{V}\left\{\hat{g}_{N_n}(y)\middle|\mathcal{D}_n\right\} dy\right)^{1/2}\right\}.
\end{aligned}$$

Next we observe

$$\begin{aligned}\int \mathbf{V}\left\{\hat{g}_{N_n}(y)\,\big|\mathcal{D}_n\right\} dy =& \int \frac{1}{N_n \cdot h_{N_n}} \cdot \frac{1}{h_{N_n}} \cdot \mathbf{V}\left\{K\left(\frac{y-\hat{m}_n(\bar{X}_1)}{h_{N_n}}\right)\big|\mathcal{D}_n\right\} dy\\ \leq& \int \frac{1}{N_n \cdot h_{N_n}} \cdot \frac{1}{h_{N_n}} \cdot \mathbf{E}\left\{K^2\left(\frac{y-\hat{m}_n(\bar{X}_1)}{h_{N_n}}\right)\big|\mathcal{D}_n\right\} dy\\ =& \frac{1}{N_n \cdot h_{N_n}} \cdot \int\int \frac{1}{h_{N_n}} \cdot K^2\left(\frac{y-\hat{m}_n(x)}{h_{N_n}}\right)\mathbf{P}_{\bar{X}}(dx)\,dy\\ =& \frac{1}{N_n \cdot h_{N_n}} \cdot \int\int \frac{1}{h_{N_n}} \cdot K^2\left(\frac{y-\hat{m}_n(x)}{h_{N_n}}\right)\,dy\,\mathbf{P}_{\bar{X}}(dx)\\ =& \frac{1}{N_n \cdot h_{N_n}} \cdot \int K^2(u)\;du \leq \frac{c_{28}}{N_n \cdot h_{N_n}}.\end{aligned}$$

Thus we can bound the variance term by

$$\sqrt{\lambda(S_n)} \cdot \mathbf{E}\left\{\left(\int_{S_n} \mathbf{V}\left\{\hat{g}_{N_n}(y)\big|\mathcal{D}_n\right\} dy\right)^{1/2}\right\} \leq \frac{\sqrt{c_{28} \cdot \lambda(S_n)}}{\sqrt{N_n \cdot h_{N_n}}}.$$

Next we observe

$$\begin{aligned}&\mathbf{E}\int |\mathbf{E}\left\{\hat{g}_{N_n}(y)\big|\mathcal{D}_n\right\} - \mathbf{E}\left\{\hat{g}_{\hat{m}_n(X),N_n}(y)\big|\mathcal{D}_n\right\}|dy\\ &= \mathbf{E}\int \left|\mathbf{E}\left\{\frac{1}{h_{N_n}} \cdot K\left(\frac{y-\hat{m}_n(\bar{X}_1)}{h_{N_n}}\right)\big|\mathcal{D}_n\right\} - \mathbf{E}\left\{\frac{1}{h_{N_n}} \cdot K\left(\frac{y-\hat{m}_n(X_{n+1})}{h_{N_n}}\right)\big|\mathcal{D}_n\right\}\right|dy\\ &= \mathbf{E}\int \left|\int \frac{1}{h_{N_n}} \cdot K\left(\frac{y-\hat{m}_n(x)}{h_{N_n}}\right) \cdot \hat{f}_n(x)\,dx - \int \frac{1}{h_{N_n}} \cdot K\left(\frac{y-\hat{m}_n(x)}{h_{N_n}}\right) \cdot f(x)\,dx\right|dy\\ &\leq \mathbf{E}\int\int \frac{1}{h_{N_n}} \cdot K\left(\frac{y-\hat{m}_n(x)}{h_{N_n}}\right) \cdot |\hat{f}_n(x) - f(x)|\,dx\,dy\\ &= \mathbf{E}\int\int \frac{1}{h_{N_n}} \cdot K\left(\frac{y-\hat{m}_n(x)}{h_{N_n}}\right)\,dy \cdot |\hat{f}_n(x) - f(x)|\,dx\\ &\leq \mathbf{E}\int |\hat{f}_n(x) - f(x)|\,dx.\end{aligned}$$

Set

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

To bound the last term we observe

$$\begin{aligned}&\mathbf{E}\int_{S_n} |\mathbf{E}\left\{\hat{g}_{\hat{m}_n(X),N_n}(y)\big|\mathcal{D}_n\right\} - g(y)|dy\\ &\leq \mathbf{E}\int_{S_n} |\mathbf{E}\left\{\hat{g}_{\hat{m}_n(X),N_n}(y)\big|\mathcal{D}_n\right\} - \mathbf{E}\left\{\hat{g}_{Y,N_n}(y)\big|\mathcal{D}_n\right\}|dy\\ &\quad + \int_{S_n} |\mathbf{E}\left\{\hat{g}_{Y,N_n}(y)\big|\mathcal{D}_n\right\} - g(y)|dy\end{aligned}$$

By the assumptions on g we have

$$\begin{aligned}\int_{S_n} |\mathbf{E}\left\{\hat{g}_{Y,N_n}(y)\big|\mathcal{D}_n\right\} - g(y)|dy =& \int_{S_n} \left| \int \frac{1}{h_{N_n}} \cdot K\left(\frac{y-x}{h_{N_n}}\right) \cdot g(x)dx - g(y)\right|dy \\ \leq& \int_{S_n} \int \frac{1}{h_{N_n}} \cdot K\left(\frac{y-x}{h_{N_n}}\right) \cdot |g(x)-g(y)|dx\,dy \\ \leq& \int_{S_n} \int \frac{1}{h_{N_n}} \cdot K\left(\frac{y-x}{h_{N_n}}\right) \cdot C \cdot |x-y|^r dx\,dy \\ \leq& c_{29} \cdot h_{N_n}^r \cdot \int_{S_n} \int K(u) \cdot |u|^r du\,dy \\ =& c_{29} \cdot h_{N_n}^r \cdot \lambda(S_n) \cdot \int K(u) \cdot |u|^r du \\ \leq& c_{30} \cdot h_{N_n}^r \cdot \lambda(S_n).\end{aligned}$$

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

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

Thus

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

From this we conclude

$$\begin{aligned}&\mathbf{E} \int_{S_n} |\mathbf{E}\left\{\hat{g}_{\hat{m}_n(X),N_n}(y)\big|\mathcal{D}_n\right\} - \mathbf{E}\left\{\hat{g}_{Y,N_n}(y)\big|\mathcal{D}_n\right\}|dy \\ &\leq \int_{S_n} \mathbf{E}\left\{|\hat{g}_{\hat{m}_n(X),N_n}(y) - \hat{g}_{Y,N_n}(y)|\right\} dy \\ &\leq \mathbf{E} \int_{\mathbb{R}} |\hat{g}_{\hat{m}_n(X),N_n}(y) - \hat{g}_{Y,N_n}(y)|\, dy \\ &\leq \frac{2 \cdot K(0)}{h_{N_n}} \cdot \mathbf{E}\left\{|\hat{m}_n(X) - Y|\right\} \\ &\leq \frac{2 \cdot K(0)}{h_{N_n}} \cdot \sqrt{\mathbf{E}\left\{|\hat{m}_n(X) - Y|^2\right\}}.\end{aligned}$$

Combining the above results yields the assertion. □

Proof of Lemma 1

In order to prove Lemma 1 we need the following auxiliary lemma:

Lemma 2. *Let $d, n \in \mathbb{N}$. Let $X, X_1, \ldots$ independent and multivariate normally distributed with mean $\mu \in \mathbb{R}^d$ and positive definite covariance $\Sigma \in \mathbb{R}^{d \times d}$. Estimate $\hat{\mu}$ by* (10) *and $\hat{\Sigma}$ by* (11)*. Then there exists constants $c_{56}, c_{57} \in \mathbb{R}_+$ such that*

$$\mathbf{E}\left\{\|\hat{\mu} - \mu\|_\infty\right\} \leq \frac{c_{56}}{\sqrt{n}}$$

and

$$\mathbf{E}\left\{\|\hat{\Sigma} - \Sigma\|_\infty\right\} \leq \frac{c_{57}}{\sqrt{n}}.$$

Proof. If Z, Z_1, ..., Z_n are independent and identically distributed real-valued random variables with $\mathbf{E}\{Z^2\} < \infty$, then

$$\mathbf{E}\left\{\left|\frac{1}{n}\sum_{i=1}^n Z_i - \mathbf{E}Z\right|\right\} \leq \sqrt{\mathbf{V}\left(\frac{1}{n}\sum_{i=1}^n Z_i\right)} = \sqrt{\frac{\mathbf{V}(Z)}{n}},$$

which implies the first inequality.

The second inequality follows similarly using

$$\begin{aligned}
&\mathbf{E}\Bigg\{\Bigg|\frac{1}{n}\sum_{k=1}^n \Big(X_k^{(i)} - \frac{1}{n}\sum_{l=1}^n X_l^{(i)}\Big)\Big(X_k^{(j)} - \frac{1}{n}\sum_{l=1}^n X_l^{(j)}\Big) \\
&\qquad\qquad\qquad - (\mathbf{E}\{X^{(i)}X^{(j)}\} - \mathbf{E}\{X^{(i)}\}\mathbf{E}\{X^{(j)}\})\Bigg|\Bigg\} \\
&= \mathbf{E}\left\{\left|\frac{1}{n}\sum_{k=1}^n X_k^{(i)}X_k^{(j)} - \frac{1}{n}\sum_{k=1}^n X_k^{(i)} \cdot \frac{1}{n}\sum_{k=1}^n X_k^{(j)} - (\mathbf{E}\{X^{(i)}X^{(j)}\} - \mathbf{E}\{X^{(i)}\}\mathbf{E}\{X^{(j)}\})\right|\right\} \\
&\leq \mathbf{E}\left\{\left|\frac{1}{n}\sum_{k=1}^n X_k^{(i)}X_k^{(j)} - \mathbf{E}\{X^{(i)}X^{(j)}\}\right|\right\} \\
&\quad + \mathbf{E}\left\{\left|\frac{1}{n}\sum_{k=1}^n X_k^{(i)} \cdot \mathbf{E}\{X^{(j)}\} - \mathbf{E}\{X^{(i)}\}\mathbf{E}\{X^{(j)}\}\right|\right\} \\
&\quad + \mathbf{E}\left\{\left|\left(\frac{1}{n}\sum_{k=1}^n X_k^{(i)} - \mathbf{E}\{X^{(i)}\}\right) \cdot \left(\frac{1}{n}\sum_{l=1}^n X_l^{(j)} - \mathbf{E}\{X^{(j)}\}\right)\right|\right\} \\
&\quad + \mathbf{E}\left\{\left|\mathbf{E}\{X^{(i)}\} \cdot \left(\frac{1}{n}\sum_{l=1}^n X_l^{(j)} - \mathbf{E}\{X^{(j)}\}\right)\right|\right\}
\end{aligned}$$

and

$$\begin{aligned}
&\mathbf{E}\left\{\left|\left(\frac{1}{n}\sum_{k=1}^n X_k^{(i)} - \mathbf{E}\{X^{(i)}\}\right) \cdot \left(\frac{1}{n}\sum_{l=1}^n X_l^{(j)} - \mathbf{E}\{X^{(j)}\}\right)\right|\right\} \\
&\leq \sqrt{\mathbf{E}\left\{\left|\frac{1}{n}\sum_{k=1}^n X_k^{(i)} - \mathbf{E}\{X^{(i)}\}\right|^2\right\}} \cdot \sqrt{\mathbf{E}\left\{\left|\frac{1}{n}\sum_{l=1}^n X_l^{(j)} - \mathbf{E}\{X^{(j)}\}\right|^2\right\}}.
\end{aligned}$$

□

Proof of Lemma 1. Scheffés Lemma implies that

$$\mathbf{E} \int_{\mathbb{R}} |\hat{f}_n(x) - f(x)| \, dx = 2 \cdot \mathbf{E} \left\{ \sup_{A \in \mathcal{B}^d} |\mathbf{P}_{\bar{X}}(A) - \mathbf{P}_X(A)| \right\}.$$

Since $\hat{\mu}$ is normally distributed with expectation μ we have

$$\mathbf{P}\left\{\mu = \hat{\mu}\right\} = 0, \tag{1}$$

thus w.l.o.g. we can assume that

$$|(\mu - \hat{\mu})^{(i)}| > 0 \tag{2}$$

for some $i \in \{1, \ldots, d\}$. Using Theorem 1.2 from Devroye, Mehrabian and Reddad (2019) we have

$$\begin{aligned} &\sup_{A \in \mathcal{B}^d} |\mathbf{P}_{\bar{X}}(A) - \mathbf{P}_X(A)| \\ &\leq \frac{9}{2} \cdot \max \left\{ \frac{|(\mu - \hat{\mu})^T (\Sigma - \hat{\Sigma})(\mu - \hat{\mu})|}{(\mu - \hat{\mu})^T \Sigma (\mu - \hat{\mu})}, \frac{(\mu - \hat{\mu})^T (\mu - \hat{\mu})}{\sqrt{(\mu - \hat{\mu})^T \Sigma (\mu - \hat{\mu})}}, \right. \\ &\qquad \left. \left\| (\Pi^T \Sigma \Pi)^{-1} \Pi^T \hat{\Sigma} \Pi - I_{d-1} \right\|_F \right\}, \end{aligned}$$

where Π is a $d \times d-1$ orthogonal matrix whose columns form a basis for the subspace orthogonal to $\mu - \hat{\mu}$ and I_{d-1} is the $d-1$ dimensional identity matrix. Since Π only needs to be orthogonal to $\mu - \hat{\mu}$, we choose Π to be orthonormal, thus we have

$$\|\Pi\|_\infty \leq c_{58}. \tag{3}$$

Since Σ is symmetric and positive definite we have

$$(\mu - \hat{\mu})^T \Sigma (\mu - \hat{\mu}) \geq c_{59} \cdot \|\mu - \hat{\mu}\|_\infty^2. \tag{4}$$

We observe by (4) that

$$\begin{aligned} \frac{|(\mu - \hat{\mu})^T (\Sigma - \hat{\Sigma})(\mu - \hat{\mu})|}{(\mu - \hat{\mu})^T \Sigma (\mu - \hat{\mu})} &\leq c_{60} \cdot \frac{\|\mu - \hat{\mu}\|_\infty^2 \cdot \|\Sigma - \hat{\Sigma}\|_\infty}{\|\mu - \hat{\mu}\|_\infty^2} \\ &\leq c_{60} \cdot \|\Sigma - \hat{\Sigma}\|_\infty \end{aligned}$$

and

$$\begin{aligned} \frac{(\mu - \hat{\mu})^T (\mu - \hat{\mu})}{\sqrt{(\mu - \hat{\mu})^T \Sigma (\mu - \hat{\mu})}} &\leq c_{61} \cdot \frac{\|\mu - \hat{\mu}\|_\infty^2}{\|\mu - \hat{\mu}\|_\infty} \\ &= c_{61} \cdot \|\mu - \hat{\mu}\|_\infty. \end{aligned}$$

Let $\Sigma = O^T \Lambda O$ be the eigendecomposition of Σ where O is orthonormal. Using

$$\|A \cdot B\|_F \leq \|A\|_F \cdot \|B\|_F,$$

and

$$\|C\|_F \leq \|C\|_\infty$$

for matrices $A \in \mathbb{R}^{d_1 \times d_2}$, $B \in \mathbb{R}^{d_2 \times d_3}$ and $C \in \mathbb{R}^{d_1 \times d_1}$, with $d_1, d_2, d_3 \in \mathbb{N}$, we see that

$$\begin{aligned}
\left\|(\Pi^T \Sigma \Pi)^{-1} \Pi^T \hat{\Sigma} \Pi - I_{d-1}\right\|_F &= \|(\Pi^T \Sigma \Pi)^{-1} \cdot (\Pi^T (\hat{\Sigma} - \Sigma)\Pi)\|_F \\
&\leq \|(\Pi^T \Sigma \Pi)^{-1}\|_F \cdot \|\Pi^T (\hat{\Sigma} - \Sigma)\Pi\|_F \\
&= \|(\Pi^T O^T \Lambda O \Pi)^{-1}\|_F \cdot \|\Pi^T (\hat{\Sigma} - \Sigma)\Pi\|_F \\
&= \|(O\Pi)^T \Lambda^{-1} (O\Pi)\|_F \cdot \|\Pi^T (\hat{\Sigma} - \Sigma)\Pi\|_F \\
&\leq c_{62} \cdot \|\Pi^T (\hat{\Sigma} - \Sigma)\Pi\|_F \\
&\leq c_{63} \cdot \|\hat{\Sigma} - \Sigma\|_\infty,
\end{aligned}$$

where the last two steps are implied since Σ is symmetric and positive definite, thus all its eigenvalues are greater than zero and since Π and O are orthonormal, thus their entries are bounded.

Combining the above results we have

$$\begin{aligned}
\mathbf{E} \int_{\mathbb{R}} |\hat{f}_n(x) - f(x)| \, dx &\leq c_{64} \cdot \mathbf{E}\left\{\max\left\{\|\mu - \hat{\mu}\|_\infty, \|\Sigma - \hat{\Sigma}\|_\infty\right\}\right\} \\
&\leq c_{64} \cdot \left(\mathbf{E}\left\{\|\mu - \hat{\mu}\|_\infty, \right\} + \mathbf{E}\left\{\|\Sigma - \hat{\Sigma}\|_\infty\right\}\right).
\end{aligned}$$

Application of Lemma 2 yields the assertion. □

Proof of Theorem 2

In this section we prove Theorem 2. Therefore we need an auxiliary result concerning a surrogate model of a simulation model $m \colon \mathbb{R}^d \to \mathbb{R}$ based on a general class of functions. We estimate the surrogate model using a uniformly on $B_n \subseteq \mathbb{R}^d$ distributed sample

$$U_{1,n}, \ldots, U_{L_n,n}$$

by a penalized least squares estimate. We define the estimate $\hat{m}_{L_n}$ of m by

$$\tilde{m}_{L_n}(\cdot) = \arg \min_{f \in \mathcal{F}_{L_n}} \frac{1}{L_n} \sum_{i=1}^{L_n} |f(U_{i,n}) - m(U_{i,n})|^2 + pen_n^2(f), \tag{1}$$

where $\mathcal{F}_{L_n}$ is a set of functions $f : \mathbb{R}^d \to \mathbb{R}$ and $pen_n^2(f) \geq 0$ is a penalty term for each $f \in \mathcal{F}_{L_n}$, and

$$\hat{m}_{L_n}(x) = T_{\beta_n}(\tilde{m}_{L_n}(x)) \quad (x \in \mathbb{R}) \tag{2}$$

for some $\beta_n > 0$.

Theorem 4. *Let $d, n, L_n \in \mathbb{N}$ with $2 \leq L_n$. Let X be a $\mathbb{R}^d$ valued random variable. Let*

$$U_{1,n}, \dots, U_{L_n,n}$$

be independent and uniformly distributed on $B_n \subseteq \mathbb{R}^d$.

Let f be the density of X and assume that

$$\|f\|_\infty \leq c_{31}. \tag{3}$$

Let $m \colon \mathbb{R}^d \to \mathbb{R}$ be a measurable function and assume that for some $1 \leq \beta_n \leq L_n$

$$\|m\|_{\infty, B_n} \leq \beta_n. \tag{4}$$

Estimate the surrogate model $\hat{m}_{L_n}(\cdot) \colon \mathbb{R}^d \to \mathbb{R}$ by (1) and (2), where $\mathcal{F}_{L_n}$ is a set of functions and $pen_n^2(g) \geq 0$ is a penalty term for every $g \in \mathcal{F}_{L_n}$.

Choose $\delta_{L_n} > 0$ such that

$$\delta_{L_n} > c_{32} \cdot \frac{\beta_n^2}{L_n},$$

$$\frac{\sqrt{L_n} \cdot \delta}{\beta_n} \geq c_{33} \int_{\delta/(c_{34} \cdot \beta_n)}^{\sqrt{48\delta}} \Bigg(\log \mathcal{N}_2 \Bigg(\frac{u}{4\beta_n}, \{ T_{\beta_n} f - g : f \in \mathcal{F}_{L_n}, \tag{5}$$
$$\frac{1}{L_n} \sum_{i=1}^{L_n} |T_{\beta_n} f(x_i) - g(x_i)|^2 + pen_n^2(f) \leq 48 \cdot \delta \}, x_1^{L_n} \Bigg) \Bigg)^{1/2} du$$

for all $\delta \geq \delta_{L_n}$, $g \in \{m\} \cup \mathcal{F}_n$ and all $x_1, \dots, x_{L_n} \in B_n$.

Then we have for some constant $c_{35} \in \mathbb{R}_+$

$$\begin{aligned}
&\mathbf{E}\{|\hat{m}_{L_n}(X) - m(X)|^2\} \\
&\leq c_{35} \cdot \lambda(B_n) \cdot \left(\inf_{f \in \mathcal{F}_{L_n}} \left(\int |f(x) - m(x)|^2 \mathbf{P}_{U_{1,n}}(dx) + pen_n^2(f) \right) + \beta_n^2 \cdot \delta_{L_n} + \frac{\beta_n^2}{L_n} \right) \\
&\quad + 2\beta_n^2 \cdot \int_{\mathbb{R}^d \setminus B_n} f(x)\, dx + 2 \cdot \int_{\mathbb{R}^d \setminus B_n} |m(x)|^2 \, \mathbf{P}_X(dx).
\end{aligned}$$

Proof. First we observe

$$\begin{aligned}
&\mathbf{E}\left\{ |\hat{m}_{L_n}(X) - m(X)|^2 \right\} \\
&= \mathbf{E} \int |\hat{m}_{L_n}(x) - m(x)|^2 \cdot f(x)\, dx \\
&= \mathbf{E} \int_{B_n} |\hat{m}_{L_n}(x) - m(x)|^2 \cdot f(x)\, dx + \mathbf{E} \int_{\mathbb{R}^d \setminus B_n} |\hat{m}_{L_n}(x) - m(x)|^2 \cdot f(x)\, dx.
\end{aligned}$$

Using $(a+b)^2 \leq 2a^2 + 2b^2$ and since by assumption $\hat{m}_{L_n}(\cdot)$ is bounded in absolute value by β_n we have

$$\mathbf{E} \int_{\mathbb{R}^d \setminus B_n} |\hat{m}_{L_n}(x) - m(x)|^2 \cdot f(x)\, dx \leq 2\beta_n^2 \cdot \int_{\mathbb{R}^d \setminus B_n} f(x)\, dx + 2 \cdot \int_{\mathbb{R}^d \setminus B_n} |m(x)|^2 \, \mathbf{P}_X(dx).$$

Using (3) and that the density of $U_{1,n}$ has a constant value $1/\lambda(B_n)$ on B_n we have

$$\begin{aligned}\mathbf{E} \int_{B_n} |\hat{m}_{L_n}(x) - m(x)|^2 \cdot f(x)\, dx &\leq c_{36} \cdot \mathbf{E} \int_{B_n} |\hat{m}_{L_n}(x) - m(x)|^2 \, dx \\ &= c_{36} \cdot \lambda(B_n) \cdot \mathbf{E} \int |\hat{m}_{L_n}(x) - m(x)|^2 \, \mathbf{P}_{U_{1,n}}(dx).\end{aligned}$$

Next we apply Theorem 2 from Götz, Kersting and Kohler (2018) with $\beta = \beta_n$, $(X_i, Y_i) = (U_{i,n}, m(U_{i,n}))$, $w^{(n)} = 1$, $n = L_n$, $\bar{Y}_{i,L_n+\bar{L}_n} = Y_i = m(U_{i,n})$ $(i = 1, \ldots, L_n)$ and suitably chosen $\bar{Y}_{L_n+1,L_n+\bar{L}_n}, \ldots, \bar{Y}_{L_n+\bar{L}_n,L_n+\bar{L}_n}$, and obtain

$$\begin{aligned}&\mathbf{E} \int |\hat{m}_{L_n}(x) - m(x)|^2 \, \mathbf{P}_{U_{1,n}}(dx) \\ &\leq c_{37} \cdot \beta_n^2 \cdot \delta_{L_n} + \frac{c_{38} \cdot \beta_n^2}{L_n} + 9 \cdot \inf_{f \in \mathcal{F}_{L_n}} \left(\int |f(x) - m(x)|^2 \mathbf{P}_{U_{1,n}}(dx) + pen_n^2(f) \right).\end{aligned}$$

Combining the above results we get the assertion. □

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

$$\delta_{L_n} = c_{39} \cdot \beta_n^2 \cdot \frac{\log(L_n)}{L_n} \cdot M_{L_n}.$$

First we show that Theorem 4 is applicable by the assumptions of Theorem 2 and the choice of δ_{L_n}. For

$$\delta \geq \delta_{L_n} > c_{39} \cdot \frac{\beta_n^2}{L_n}$$

and $x_1^{L_n} \in B_n$ we have

$$\begin{aligned}&\int_{\delta/(c_{40} \cdot \beta_n)}^{\sqrt{48\delta}} \left(\log \mathcal{N}_2 \left(\frac{u}{4\beta_n}, \{T_{\beta_n} h - g : h \in \mathcal{H}^{(l)}_{I_1, M_{L_n}, d, d^*, \gamma_{L_n}}\}, x_1^{L_n} \right) \right)^{1/2} du \\ &\leq \sqrt{48\delta} \cdot \left(\log \mathcal{N}_2 \left(\frac{c_{41}}{L_n}, \{T_{\beta_n} h - g : h \in \mathcal{H}^{(l)}_{I_1, M_{L_n}, d, d^*, \gamma_{L_n}}\}, x_1^{L_n} \right) \right)^{1/2}.\end{aligned}$$

Set $a_{L_n} = c_5 \cdot \log(L_n)$, then we have $B_n \subseteq [-a_{L_n}, a_{L_n}]^d$. Since $\max\{a_{L_n}, \gamma_{L_n}, M_{L_n}\} \leq L_n^{c_{42}}$ holds we can apply Lemma 2 from Bauer and Kohler (2019) to bound the above covering number by

$$\log \left(\mathcal{N}_2 \left(\frac{c_{41}}{L_n}, \{T_{\beta_n} h - g : h \in \mathcal{H}^{(l)}_{I_1, M_{L_n}, d, d^*, \gamma_{L_n}}\}, x_1^{L_n} \right) \right) \leq c_{42} \cdot \log(L_n) \cdot M_{L_n},$$

for L_n sufficiently large. Combing the above results we see that (5) is implied by

$$\frac{\sqrt{L_n} \cdot \delta}{\beta_n} \geq c_{43} \cdot \sqrt{48\delta} \cdot (c_{42} \cdot \log(L_n) \cdot M_{L_n})^{1/2}$$

which in turn follows from $\delta \geq \delta_{L_n}$, for a suitably chosen $c_{39} \in \mathbb{R}_+$.

Applying Theorem 1 and Theorem 4 yields

$$
\begin{aligned}
&\mathbf{E} \int_{\mathbb{R}} |\hat{g}_{N_n}(y) - g(y)| dy \\
&\leq 2 \cdot \int_{S_n^c} g(y) dy + \frac{c_{44} \cdot \sqrt{\lambda(S_n)}}{\sqrt{N_n \cdot h_{N_n}}} + c_{45} \cdot \lambda(S_n) \cdot h_{N_n}^r + \mathbf{E} \int |\hat{f}_n(x) - f(x)| dx \\
&\quad + \frac{c_{46}}{h_{N_n}} \Bigg(\lambda(B_n) \Bigg(\inf_{h \in \mathcal{H}^{(l)}_{I_1, M_{L_n}, d, d^*, \gamma_{L_n}}} \left(\left(\int |f(x) - m(x)|^2 \mathbf{P}_{U_{1,n}}(dx) \right) + \beta_n^2 \cdot \delta_{L_n} + \frac{\beta_n^2}{L_n} \right) \\
&\qquad\qquad + \beta_n^2 \cdot \int_{\mathbb{R}^d \setminus B_n} f(x)\, dx + \int_{\mathbb{R}^d \setminus B_n} |m(x)|^2 \, \mathbf{P}_X(dx) \Bigg)^{1/2} .
\end{aligned}
$$

To derive a bound on the approximation error we first observe since $U_{1,n}$ is uniformly distributed on B_n

$$
\int |h(x) - m(x)|^2 \mathbf{P}_{U_{1,n}}(dx) = \int_{B_n} |h(x) - m(x)|^2 \mathbf{P}_{U_{1,n}}(dx) \tag{6}
$$

holds for an arbitrary $h \in \mathcal{H}^{(l)}_{I_1, M_{L_n}, d, d^*, \gamma_{L_n}}$. We set $\eta_{L_n} = (\log L_n)^{4p+6-2q} \cdot L_n^{-\frac{2 \cdot (q+1) \cdot p + 2d^*}{2p+d^*}}$. Using Theorem 3 in Bauer and Kohler (2019) we see that there exists a $h^* \in \mathcal{H}^{(l)}_{I_1, M_{L_n}, d^*, d, \gamma_{L_n}}$ and an exception set D_{L_n} with $\mathbf{P}_X$-measure of η_{L_n} such that

$$
\begin{aligned}
&\int_{B_n} |h^*(x) - m(x)|^2 \cdot I_{D_{L_n}^c}(x) \, \mathbf{P}_{U_{1,n}}(dx) + \int_{B_n} |h^*(x) - m(x)|^2 \cdot I_{D_{L_n}}(x) \, \mathbf{P}_{U_{1,n}}(dx) \\
&\leq \left(c_{47} \cdot a_{L_n}^{(2q+3)} \cdot M_{L_n}^{-p/d^*} \right)^2 + \left(2 \cdot c_{48} \cdot a_{L_n}^q \cdot M_{L_n}^{(d^* + q \cdot p)/d^*} \right)^2 \cdot \eta_{L_n} \\
&\leq c_{49} \cdot (\log L_n)^{4p+6} \cdot L_n^{-\frac{2p}{2p+d^*}} + c_{50} \cdot (\log L_n)^{2q} \cdot L_n^{\frac{2d^* + 2q \cdot p}{2p+d^*}} \cdot (\log L_n)^{4p+6-2q} \cdot L_n^{-\frac{2 \cdot (q+1) \cdot p + 2d^*}{2p+d^*}} \\
&\leq c_{51} \cdot (\log L_n)^{4p+6} \cdot L_n^{-\frac{2p}{2p+d^*}},
\end{aligned}
$$

where we have used that $\|m\|_{\infty, B_n} \leq \beta_n \leq c_{48} \cdot a_{L_n}^q \cdot M_{L_n}^{(d^* + q \cdot p)/d^*}$.

To conclude by the choice of δ_{L_n} we have that

$$
\begin{aligned}
&\mathbf{E} \int_{\mathbb{R}} |\hat{g}_{N_n}(y) - g(y)| dy \\
&\leq 2 \cdot \int_{S_n^c} g(y) dy + \frac{c_{44} \cdot \sqrt{\lambda(S_n)}}{\sqrt{N_n \cdot h_{N_n}}} + c_{45} \cdot \lambda(S_n) \cdot h_{N_n}^r + \mathbf{E} \int |\hat{f}_n(x) - f(x)| dx \\
&\quad + \frac{c_{52}}{h_{N_n}} \Bigg(\beta_n^4 \cdot \lambda(B_n) \cdot (\log L_n)^{4p+6} \cdot L_n^{-\frac{2p}{2p+d^*}} + \beta_n^2 \cdot \int_{\mathbb{R}^d \setminus B_n} f(x)\, dx \\
&\qquad\qquad + \int_{\mathbb{R}^d \setminus B_n} |m(x)|^2 \, \mathbf{P}_X(dx) \Bigg)^{1/2}
\end{aligned}
$$

holds for L_n sufficiently large.

□

Proof of Corollary 1

Since $\frac{|y|}{\sqrt{n}} \geq 1$ for every $y \in S_n^c$ and $\mathbf{E}\{|Y|\} < \infty$ we have

$$\int_{S_n^c} g(y)\,dy \quad \leq \quad \int_{S_n^c} \frac{|y|}{\sqrt{n}} \cdot g(y)\,dy \leq c_{53} \cdot n^{-1/2}.$$

Next we see since Σ is positive definite, we have $\sigma_{ii} > 0$ for all $i \in \{1, \ldots, d\}$. Furthermore we observe that for each component of X it holds $X^{(i)} \sim \mathcal{N}(\mu^{(i)}, \sigma_{ii}^2)$. Thus

$$\begin{aligned}
&\int_{\mathbb{R}^d \setminus B_n} f(x)\,dx \\
&\leq \sum_{i=1}^{d} \mathbf{P}\{|X^{(i)}| \geq c_5 \cdot (\log L_n)\} \\
&\leq \sum_{i=1}^{d} \mathbf{P}\left\{ \frac{X^{(i)} - \mu_i}{\sigma_{ii}} \geq \frac{c_5 \cdot (\log L_n) - \mu_i}{\sigma_{ii}} \right\} + \sum_{i=1}^{d} \mathbf{P}\left\{ \frac{X^{(i)} - \mu_i}{\sigma_{ii}} \leq - \frac{c_5 \cdot (\log L_n) + \mu_i}{\sigma_{ii}} \right\} \\
&\leq 2 \cdot \sum_{i=1}^{d} \mathbf{P}\left\{ \frac{X^{(i)} - \mu_i}{\sigma_{ii}} \geq \frac{c_5 \cdot (\log L_n) - \mu_i}{\sigma_{ii}} \right\}.
\end{aligned}$$

Using Lemma 1.19.2 from Gänssler and Stute (1977) and we have for every $i \in \{1, \ldots, d\}$

$$\begin{aligned}
&\mathbf{P}\left\{ \frac{X^{(i)} - \mu_i}{\sigma_{ii}} \geq \frac{c_5 \cdot (\log L_n) - \mu_i}{\sigma_{ii}} \right\} \\
&\leq \frac{\sigma_{ii}}{(c_5 \cdot (\log L_n) - \mu_i) \cdot \sqrt{2\pi}} \cdot \exp\left(-\frac{1}{2} \cdot \left(\frac{c_5 \cdot (\log L_n) - \mu_i}{\sigma_{ii}} \right)^2 \right),
\end{aligned}$$

which is smaller than $\lambda(B_n) \cdot \log(L_n)^{4p+6} \cdot L_n^{-\frac{2p}{2p+d^*}}$ for L_n sufficiently large. Application of Theorem 2 and Lemma 1 together with the assumptions yields the assertion. □

8.1 Proof of Theorem 3

In this section we prove Theorem 3. Therefore we will show an auxiliary result concerning the rate of convergence of an improved surrogate model for an imperfect simulation model $m \colon \mathbb{R}^d \to \mathbb{R}$. I.e. we we consider the second data model where $m(X) \neq Y = m^*(X)$ and we have an observed independent and identically distributed sample

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

of (X, Y). To estimate the simulation model we generate an independent and uniformly on $B_n := [-c_5 \cdot \log(L_n), c_5 \cdot \log(L_n)]^d$ distributed sample

$$U_{1,n}, \dots, U_{L_n,n}$$

and define the estimate $\hat{m}_{L_n}$ of m by (1) and (2). Next we define an estimate of $m^* - \hat{m}_{L_n}$ on basis of the residuals

$$\epsilon_i = Y_i - \hat{m}_{L_n}(X_i) \quad (i = 1, \dots, n), \tag{1}$$

by a penalized least squares estimate

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

for a set of functions $\mathcal{F}_n$ and a penalty term $pen_n^2(f) \geq 0$ for each $f \in \mathcal{F}_n$, where we assume that the penalty term satisfies $pen_n^2(\alpha \cdot f) = \alpha^2 \cdot pen_n^2(f)$ for $\alpha \in \mathbb{R}$ and $f \in \mathcal{F}_n$ with $\alpha \cdot f \in \mathcal{F}_n$. We set

$$\hat{m}_n^\epsilon(x) = T_{c_{65} \cdot \alpha_n}(\tilde{m}_n^\epsilon(x)) \quad (x \in \mathbb{R}^d), \tag{3}$$

where $c_{65} \geq 1$ and $\alpha_n > 0$. We define our final surrogate model $(X, \hat{m}_n(X))$ for (X, Y) by

$$\hat{m}_n(x) = \hat{m}_{L_n}(x) + \hat{m}_n^\epsilon(x) \quad (x \in \mathbb{R}^d). \tag{4}$$

Theorem 5. *Let $d, n, L_n, N_n \in \mathbb{N}$ with $2 \leq n \leq L_n$. Let $(X, Y), (X_1, Y_1), \dots$ be independent and identically distributed $\mathbb{R}^d \times \mathbb{R}$ valued random variables. Let $f \colon \mathbb{R}^d \to \mathbb{R}$ be the density of X w.r.t. the Lebesgue measure which we assume to exist. Assume that*

$$\|f\|_\infty \leq c_{66} \tag{5}$$

for some $c_{66} \in \mathbb{R}_+$. Assume that $\mathbf{E}\{|Y|\} < \infty$.

Let $m \colon \mathbb{R}^d \to \mathbb{R}$ be a measurable function and assume that for some $1 \leq \beta_n \leq L_n$

$$\|m\|_{\infty, B_n} \leq \beta_n, \tag{6}$$

where

$$B_n := [-c_5 \cdot \log(L_n), c_5 \cdot \log(L_n)]^d$$

for some $c_5 \in \mathbb{R}_+$. Let $U_{1,n}, \dots, U_{L_n,n}$ be independent and uniformly distributed on B_n and define the surrogate estimate $\hat{m}_{L_n}$ by (1) and (2).

Assume that there exists a (measurable) function $m^* \colon \mathbb{R}^d \to \mathbb{R}$ such that $m^*(X) = Y$. Let

$$c_{67} \cdot \lambda(B_n) \cdot \left(\beta_n^2 \delta_{L_n} + \frac{\beta_n^2}{L_n} + \inf_{f \in \mathcal{F}_n} \left(\int |f(x) - m(x)|^2 \mathbf{P}_{U_{1,n}}(dx) + pen_n^2(f) \right) \right)$$
$$+ 2\beta_n^2 \int_{\mathbb{R}^d \setminus B_n} f(x)\, dx + \int_{\mathbb{R}^d \setminus B_n} m(x)^2\, \mathbf{P}_X(dx) \leq \frac{\alpha_n^3}{\beta_n}, \tag{7}$$

$$\int_{\mathbb{R}^d \setminus B_n} |m(x)|^3 \mathbf{P}_X(dx) \le c_{68} \cdot \alpha_n^3 \tag{8}$$

and

$$\int_{\mathbb{R}^d \setminus B_n} f(x)\, dx \le \frac{\alpha_n^3}{\beta_n^3}. \tag{9}$$

Assume that

$$\|m^* - m\|_\infty \le \alpha_n \tag{10}$$

and set

$$\frac{1}{\alpha_n} \mathcal{F}_n = \{ f/\alpha_n \colon f \in \mathcal{F}_n \}.$$

Define the estimate of the residuals $\hat{m}_n^\epsilon$ *by* (2) *and* (3) *and the improved surrogate estimate by*

$$\hat{m}_n(x) = \hat{m}_{L_n}(x) + \hat{m}_n^\epsilon(x) \quad (x \in \mathbb{R}^d). \tag{11}$$

Choose $\delta_k > 0$ *such that for all* $k \ge n$ *we have*

$$\delta_k > c_{69} \cdot \frac{\beta_n^2}{k},$$

$$\frac{\sqrt{L_n}\delta}{\beta_n} \ge c_{70} \int_{\delta/(c_{71} \cdot \beta_n)}^{\sqrt{48\delta}} \Bigg(\log \mathcal{N}_2 \Bigg(\frac{u}{4\beta_n}, \{ T_{L_n} f - g : f \in \mathcal{F}_{L_n}, \tag{12}$$
$$\frac{1}{k} \sum_{i=1}^k |T_{L_n} f(x_i) - g(x_i)|^2 + pen_n^2(f) \le 48 \cdot \delta \}, x_1^{L_n} \Bigg) \Bigg)^{1/2} du$$

for all $\delta \ge \delta_{L_n}$, $g \in \{m\} \cup \mathcal{F}_{L_n}$ *and all* $x_1, \dots, x_{L_n} \in B_n$ *and*

$$\frac{\sqrt{n}\delta}{\beta_n} \ge c_{72} \int_{\delta/(c_{73} \cdot \beta_n)}^{\sqrt{48\delta}} \Bigg(\log \mathcal{N}_2 \Bigg(\frac{u}{4\beta_n}, \{ T_n f - g : f \in \frac{1}{\alpha_n} \mathcal{F}_n, \tag{13}$$
$$\frac{1}{k} \sum_{i=1}^k |T_n f(x_i) - g(x_i)|^2 + pen_n^2(f) \le 48 \cdot \delta \}, x_1^n \Bigg) \Bigg)^{1/2} du$$

for all $\delta \ge \delta_n$, $g \in \{m^*\} \cup \frac{1}{\alpha_n}\mathcal{F}_n$ *and all* $x_1, \dots, x_n \in B_n$.

Then there exists constants $c_{74}, \dots, c_{77}$ *such that*

$$\begin{aligned}
&\mathbf{E}\{|Y - \hat{m}_n(X)|^2\} \\
&\le c_{74} \cdot \alpha_n^2 \cdot \delta_n + \frac{c_{75} \cdot \alpha_n^2}{n} + c_{76} \cdot (\alpha_n^2 \cdot n + \beta_n^2) \cdot \int_{\mathbb{R}^d \setminus B_n} f(x)\, dx + 2 \cdot \int_{\mathbb{R}^d \setminus B_n} m(x)^2 \mathbf{P}_X(dx) \\
&\quad + 9 \cdot \alpha_n^2 \cdot \inf_{f \in \frac{1}{\alpha_n}\mathcal{F}_n} \left(\int |f(x) - \frac{1}{\alpha_n}(m^* - m)(x)|^2 \mathbf{P}_X(dx) + pen_n^2(f) \right) \\
&\quad + c_{77} \cdot \lambda(B_n) \cdot \left(\beta_n^2 \cdot \delta_{L_n} + \frac{\beta_n^2}{L_n} + \inf_{f \in \mathcal{F}_n} \left(\int |f(x) - m(x)|^2 \mathbf{P}_{U_{1,n}}(dx) + pen_n^2(f) \right) \right).
\end{aligned}$$

Proof. Using the definition of $\hat{m}_n$ and $(a+b)^2 \leq 2a^2 + 2b^2$ $(a, b \in \mathbb{R})$ we have

$$\begin{aligned}
&\mathbf{E}\left\{|Y - \hat{m}_n(X)|^2\right\} = \mathbf{E}\left\{|m^*(X) - \hat{m}_n(X)|^2\right\} \\
&= \mathbf{E}\left\{|(m^*(X) - m(X) - \hat{m}_n^\epsilon(X)) + (m(X) - \hat{m}_{L_n}(X))|^2\right\} \\
&\leq 2 \cdot \mathbf{E}\left\{|m^*(X) - m(X) - \hat{m}_n^\epsilon(X)|^2\right\} + 2 \cdot \mathbf{E}\left\{|m(X) - \hat{m}_{L_n}(X)|^2\right\}.
\end{aligned}$$

Application of Theorem 4 yields

$$\begin{aligned}
&\mathbf{E}\left\{|\hat{m}_{L_n}(X) - m(X)|^2\right\} \\
&\leq c_{78} \cdot \lambda(B_n) \cdot \left(\inf_{f \in \mathcal{F}_{L_n}} \left(\int |f(x) - m(x)|^2 \mathbf{P}_{U_{1,n}}(dx) + pen_n^2(f) \right) + \beta_n^2 \cdot \delta_{L_n} + \frac{\beta_n^2}{L_n} \right) \\
&\quad + 2\beta_n^2 \cdot \int_{\mathbb{R}^d \setminus B_n} f(x)\, dx + 2 \cdot \int_{\mathbb{R}^d \setminus B_n} m(x)^2 \mathbf{P}_X(dx). \qquad (14)
\end{aligned}$$

Hence in order to prove the assertion it suffices to show

$$\begin{aligned}
&\mathbf{E} \int |\hat{m}_n^\epsilon(x) - (m^* - m)(x)|^2\, \mathbf{P}_X(dx) \qquad (15) \\
&\leq 9 \cdot \alpha_n^2 \cdot \inf_{f \in \frac{1}{\alpha_n}\mathcal{F}_n} \left(\int |f(x) - \frac{1}{\alpha_n}(m^* - m)(x)|^2 \mathbf{P}_X(dx) + pen_n^2(f) \right) \\
&\quad + c_{79} \cdot \alpha_n^2 \cdot \delta_n + c_{80} \cdot \lambda(B_n) \cdot \left(\inf_{f \in \mathcal{F}_{L_n}} \left(\int |f(x) - m(x)|^2 \mathbf{P}_{U_{1,n}}(dx) + pen_n^2(f) \right) \right. \\
&\quad\quad \left. + \beta_n^2 \cdot \delta_{L_n} + \frac{\beta_n^2}{L_n} \right) + (c_{81} \cdot \alpha_n^2 \cdot n + 4\beta_n^2) \cdot \int_{\mathbb{R}^d \setminus B_n} f(x)\, dx + 2 \cdot \int_{\mathbb{R}^d \setminus B_n} |m(x)|^2 \mathbf{P}_X(dx).
\end{aligned}$$

In order to prove (15) we first observe that

$$\begin{aligned}
&\int |\hat{m}_n^\epsilon(x) - (m^* - m)(x)|^2\, \mathbf{P}_X(dx) \\
&= \int_{B_n} |\hat{m}_n^\epsilon(x) - (m^* - m)(x)|^2\, \mathbf{P}_X(dx) + \int_{\mathbb{R}^d \setminus B_n} |\hat{m}_n^\epsilon(x) - (m^* - m)(x)|^2\, \mathbf{P}_X(dx) \\
&\leq \int_{B_n} |\hat{m}_n^\epsilon(x) - (m^* - m)(x)|^2\, \mathbf{P}_X(dx) + c_{82} \cdot \alpha_n^2 \cdot \int_{\mathbb{R}^d \setminus B_n} f(x)\, dx.
\end{aligned}$$

Next we see that

$$\int_{B_n} |\hat{m}_n^\epsilon(x) - (m^* - m)(x)|^2\, \mathbf{P}_X(dx) = \alpha_n^2 \cdot \int_{B_n} \left| \frac{1}{\alpha_n} \cdot \hat{m}_n^\epsilon(x) - \frac{1}{\alpha_n} \cdot (m^* - m)(x) \right|^2 \mathbf{P}_X(dx). \qquad (16)$$

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

$$\frac{1}{\alpha_n} \cdot \hat{m}_n^\epsilon(x) = \frac{1}{\alpha_n} \cdot T_{c_{65} \cdot \alpha_n}(\tilde{m}_n(x)) = T_{c_{65}}\left(\frac{1}{\alpha_n} \cdot \tilde{m}_n(x) \right) \quad (x \in \mathbb{R}^d),$$

and that by the definition of the estimate $\tilde{m}_n$

$$\frac{1}{\alpha_n}\tilde{m}_n(\cdot) = \arg \min_{f \in \frac{1}{\alpha_n}\mathcal{F}_n} \left(\frac{1}{n}\sum_{i=1}^n \left| f(X_i) - \frac{\epsilon_i}{\alpha_n} \right|^2 + pen_n^2(f) \right)$$

holds.

To bound (16) we use a straightforward modification of Theorem 2 from Götz, Kersting and Kohler (2018), where we replace $\int |\cdot|^2 \mathbf{P}_X(dx)$ by $\int_{B_n} |\cdot|^2 \mathbf{P}_X(dx)$. We will apply this theorem with $w^{(n)} = 1$, $L_n = 0$, $(X,Y) = (X, (Y - m(X))/\alpha_n)$, $\bar{Y}_{i,n} = (Y_i - \hat{m}_{L_n}(X_i))/\alpha_n$ $(i = 1, \dots, n)$ and $m = (m^* - m)/\alpha_n$. Therefore we first need to show that

$$\max_{i=1,\dots,n} \mathbf{E}\left\{ \left| \frac{Y_i - \hat{m}_{L_n}(X_i)}{\alpha_n} \right|^3 \right\} < \infty.$$

We observe by (7), (8), (9), (10), and (14) that we have

$$\begin{aligned}
&\max_{i=1,\dots,n} \mathbf{E}\left\{ \left| \frac{Y_i - \hat{m}_{L_n}(X_i)}{\alpha_n} \right|^3 \right\} = \frac{1}{\alpha_n^3} \cdot \mathbf{E}\left\{ |m^*(X) - \hat{m}_{L_n}(X)|^3 \right\} \\
&\leq \frac{8}{\alpha_n^3} \cdot \Bigg(\mathbf{E}\left\{ |m^*(X) - m(X)|^3 \right\} + \int_{B_n} |m(x) - \hat{m}_{L_n}(x)|^3 \mathbf{P}_X(dx) \\
&\qquad\qquad + \int_{\mathbb{R}^d \setminus B_n} |m(x) - \hat{m}_{L_n}(x)|^3 \mathbf{P}_X(dx) \Bigg) \\
&\leq \frac{8}{\alpha_n^3} \cdot \left(\alpha_n^3 + 2\beta_n \cdot \frac{\alpha_n^3}{\beta_n} + \int_{\mathbb{R}^d \setminus B_n} |m(x)|^3 \mathbf{P}_X(dx) + \int_{\mathbb{R}^d \setminus B_n} |\hat{m}_{L_n}(x)|^3 \mathbf{P}_X(dx) \right) \leq c_{83}.
\end{aligned}$$

By application of Theorem 2 from Götz, Kersting and Kohler (2018) we observe

$$\begin{aligned}
&\mathbf{E} \int_{B_n} \left| \frac{1}{\alpha_n}\hat{m}_n^{\epsilon}(x) - \frac{1}{\alpha_n}(m^* - m)(x) \right|^2 \mathbf{P}_X(dx) \\
&\leq \frac{c_{84}}{n} + c_{85} \cdot \left(\delta_n + n \cdot \int_{\mathbb{R}^d \setminus B_n} f(x)\,dx + \int_{\mathbb{R}^d \setminus B_n} \left| \frac{1}{\alpha_n}(m^* - m)(x) \right|^2 \mathbf{P}_X(dx) \right) \\
&\quad + 9 \cdot \inf_{f \in \frac{1}{\alpha}\mathcal{F}_n} \left(\int |f(x) - \frac{1}{\alpha_n}(m^* - m)(x)| \mathbf{P}_X(dx) + pen_n^2(f) \right) \\
&\quad + \mathbf{E}\left\{ \frac{1}{n}\sum_{i=1}^n \left| \frac{m(X_i) - \hat{m}_{L_n}(X_i)}{\alpha_n} \right|^2 \right\}.
\end{aligned}$$

From (14) we can conclude

$$\mathbf{E}\left\{ \frac{1}{n}\sum_{i=1}^n \left| \frac{m(X_i) - \hat{m}_{L_n}(X_i)}{\alpha_n} \right|^2 \right\}$$

$$
\begin{aligned}
&= \frac{1}{\alpha_n^2} \cdot \mathbf{E}\left\{|m(X) - \hat{m}_{L_n}(X)|^2\right\} \\
&\leq \frac{1}{\alpha_n^2}\left(c_{86} \cdot \lambda(B_n) \cdot \left(\inf_{f \in \mathcal{F}_{L_n}} \left(\int |f(x) - m(x)|^2 \mathbf{P}_{U_{1,n}}(dx) + pen_n^2(f) \right) + \beta_n^2 \cdot \delta_{L_n} + \frac{\beta_n^2}{L_n} \right) \right. \\
&\qquad\qquad \left. + 4\beta_n^2 \cdot \int_{\mathbb{R}^d \setminus B_n} f(x)\, dx + 2 \int_{\mathbb{R}^d \setminus B_n} |m(x)|^2 \mathbf{P}_X(dx) \right).
\end{aligned}
$$

Summarizing the above results we get the assertion.

□

Proof of Theorem 3. Set $pen_n^2(f) = 0$, $a_k = c_5 \cdot \log(k)$ and

$$\delta_k = c_{87} \cdot \beta_n^2 \cdot \frac{\log(k)}{k} \cdot M_k.$$

First we show that Theorem 5 is applicable by the assumptions of Theorem 3 and the choice of δ_k. We observe as in the proof of Theorem 2 that (12) holds.

For

$$\delta \geq \delta_n > c_{87} \cdot \frac{\beta_n^2}{n}$$

and $x_1^n \in B_n$ we have

$$
\begin{aligned}
&\int_{\delta/(c_{88} \cdot \beta_n)}^{\sqrt{48\delta}} \left(\log \mathcal{N}_2 \left(\frac{u}{4\beta_n}, \{T_n h - m_{sim,n} : h \in \frac{1}{\alpha_n} \mathcal{H}^{(l)}_{I_2, M_n, d, d^*, \gamma_n}\}, x_1^n \right) \right)^{1/2} du \\
&\leq \sqrt{48\delta} \cdot \left(\log \mathcal{N}_2 \left(\frac{c_{89}}{n}, \{T_n h - m_{sim,n} : h \in \frac{1}{\alpha_n} \mathcal{H}^{(l)}_{I_2, M_n, d, d^*, \gamma_n}\}, x_1^n \right) \right)^{1/2}.
\end{aligned}
$$

Since $\max\{a_n, \gamma_n/\alpha_n, M_n\} \leq n^{c_{90}}$ holds we can apply Lemma 2 from Bauer and Kohler to bound the above covering number by

$$\log \left(\mathcal{N}_2 \left(\frac{c_{89}}{n}, \{T_n h - m_{sim,n} : h \in \frac{1}{\alpha_n} \mathcal{H}^{(l)}_{I_2, M_n, d, d^*, \gamma_n}\}, x_1^n \right) \right) \leq c_{91} \cdot \log(n) \cdot M_n,$$

for n sufficiently large. Combing the above results we see that (13) is implied by

$$\frac{\sqrt{n} \cdot \delta}{\beta_n} \geq \sqrt{48\delta} \cdot (c_{91} \cdot \log(n) \cdot M_n)^{1/2}$$

which in turn follows from $\delta \geq \delta_n$, for a suitably chosen $c_{87} \in \mathbb{R}_+$.

Applying Theorem 1 and Theorem 5 yields

$$
\begin{aligned}
&\mathbf{E} \int_{\mathbb{R}} |\hat{g}_{N_n}(y) - g(y)| dy \\
&\leq 2 \cdot \int_{S_n^c} g(y) dy + \frac{c_{92} \cdot \sqrt{\lambda(S_n)}}{\sqrt{N_n \cdot h_{N_n}}} + c_{93} \cdot \lambda(S_n) \cdot h_{N_n}^r + \mathbf{E} \int |\hat{f}_n(x) - f(x)| dx
\end{aligned}
$$

$$
+ \frac{c_{94}}{h_{N_n}} \Bigg(\alpha_n^2 \cdot \delta_n + \frac{\alpha_n^2}{n} + (\alpha_n^2 \cdot n + \beta_n^2) \cdot \int_{\mathbb{R}^d \setminus B_n} f(x)\,dx + 2 \cdot \int_{\mathbb{R}^d \setminus B_n} |m_{sim,n}(x)|^2 \mathbf{P}(dx)
$$

$$
+ 9 \cdot \alpha_n^2 \cdot \inf_{h \in \frac{1}{\alpha_n} \mathcal{H}^{(l)}_{I_2, M_n, d, d^*, \gamma_n}} \int |h(x) - \frac{1}{\alpha_n}(m^* - m_{sim,n})(x)|^2 \mathbf{P}_X(dx)
$$

$$
+ \lambda(B_n) \cdot \Bigg(\beta_n^2 \cdot \delta_{L_n} + \frac{\beta_n^2}{L_n}
$$

$$
+ \inf_{h \in \mathcal{H}^{(l)}_{I_1, M_{L_n}, d, d^*, \gamma_{L_n}}} \int |h(x) - m_{sim,n}(x)|^2 \mathbf{P}_{U_{1,n}}(dx) \Bigg) \Bigg)^{1/2}.
$$

Analogous as in the proof of Theorem 2 using Theorem 3 from Bauer and Kohler (2019) we get

$$
\inf_{h \in \mathcal{H}^{(l)}_{I_1, M_{L_n}, d, d^*, \gamma_{L_n}}} \left(\int |h(x) - m_{sim,n}(x)|^2 P_{U_{1,n}}(dx) \right) \leq c_{95} \cdot (\log L_n)^{4p+6} \cdot L_n^{-\frac{2p}{2p+d^*}}.
$$

We observe that by definition for every $h \in \frac{1}{\alpha_n} \mathcal{H}^{(l)}_{I_2, M_n, d, d^*, \gamma_n}$

$$
\|h\|_\infty \leq \frac{1}{\alpha_n} \cdot I_2 \cdot (M_n + 1) \cdot \gamma_n \leq c_{96} \cdot \frac{M_n \cdot \gamma_n}{\alpha_n} \tag{17}
$$

holds. Using furthermore that $\|\frac{1}{\alpha_n}(m^* - m_{sim,n})\|_\infty \leq 1$ holds by assumption, we have

$$
\begin{aligned}
&\int |h(x) - \frac{1}{\alpha_n}(m^* - m_{sim,n})(x)|^2 \mathbf{P}_X(dx) \\
&\leq \int_{B_n} |h(x) - \frac{1}{\alpha_n}(m^* - m_{sim,n})(x)|^2 \mathbf{P}_X(dx) + c_{96} \cdot \left(\frac{M_n \cdot \gamma_n}{\alpha_n} \right)^2 \cdot \int_{\mathbb{R}^d \setminus B_n} f(x)\,dx,
\end{aligned}
$$

for every $h \in \frac{1}{\alpha_n} \mathcal{H}^{(l)}_{I_2, M_n, d, d^*, \gamma_n}$.

We set $\eta_n = (\log n)^{4p+6-2q} \cdot n^{-\frac{2 \cdot (q+1) \cdot p + 2d^*}{2p+d^*}}$. Using Theorem 3 in Bauer and Kohler (2019) we see that there exists a $h^* \in \frac{1}{\alpha_n} \mathcal{H}^{(l)}_{I_2, M_n, d, d^*, \gamma_n}$ and an exception set D_n with $\mathbf{P}_X$-measure of η_n such that

$$
\begin{aligned}
&\int_{B_n} |h^*(x) - \frac{1}{\alpha_n}(m^* - m_{sim,n})(x)|^2 \cdot I_{D_n^c}(x)\, \mathbf{P}_X(dx) \\
&\quad + \int_{B_n} |h^*(x) - \frac{1}{\alpha_n}(m^* - m_{sim,n})(x)|^2 \cdot I_{D_n}(x)\, \mathbf{P}_X(dx) \\
&\leq \left(c_{97} \cdot a_n^{(2q+3)} \cdot M_n^{-p/d^*} \right)^2 + \left(2 \cdot c_{97} \cdot a_n^q \cdot M_n^{(d^* + q \cdot p)/d^*} \right)^2 \cdot \eta_{L_n} \\
&\leq c_{97} \cdot (\log n)^{4p+6} \cdot n^{-\frac{2p}{2p+d^*}} + c_{97} \cdot (\log n)^{2q} \cdot n^{\frac{2d^* + 2q \cdot p}{2p+d^*}} \cdot (\log n)^{4p+6-2q} \cdot n^{-\frac{2 \cdot (q+1) \cdot p + 2d^*}{2p+d^*}}
\end{aligned}
$$

$$\leq c_{97} \cdot (\log n)^{4p+6} \cdot n^{-\frac{2p}{2p+d^*}},$$

where we have used that $\|\frac{1}{\alpha_n}(m^* - m_{sim,n})\|_{\infty,B_n} \leq 1 \leq c_{97} \cdot a_n^q \cdot M_n^{(d^*+q\cdot p)/d^*}$.

Thus by the choice of δ_k we have that

$$\begin{aligned}
&\mathbf{E} \int_{\mathbb{R}} |\hat{g}_{N_n}(y) - g(y)| dy \\
&\leq 2 \cdot \int_{S_n^c} g(y) dy + \frac{c_{98} \cdot \sqrt{\lambda(S_n)}}{\sqrt{N_n \cdot h_{N_n}}} + c_{99} \cdot \lambda(S_n) \cdot h_{N_n}^r + \mathbf{E} \int |\hat{f}_n(x) - f(x)| dx \\
&\quad + \frac{c_{97}}{h_{N_n}} \Bigg(\alpha_n^2 \cdot (\log n)^{4p+6} \cdot n^{-\frac{2p}{2p+d^*}} + \frac{\alpha_n^2}{n} + (\alpha_n^2 \cdot n + \beta_n^2 + \left(\frac{M_n \gamma_n}{\alpha_n}\right)^2) \cdot \int_{\mathbb{R}^d \setminus B_n} f(x)\, dx \\
&\qquad + \int_{\mathbb{R}^d \setminus B_n} m_{sim,n}(x)^2 \mathbf{P}_X(dx) + \beta_n^4 \cdot \lambda(B_n) \cdot (\log L_n)^{4p+6} \cdot L_n^{-\frac{2p}{2p+d^*}} + \frac{\beta_n^2}{L_n} \Bigg)^{1/2}
\end{aligned}$$

holds for n sufficiently large.

□

Proof of Corollary 2

Analogous to the proof of Corollary 1 one can show that

$$(\alpha_n^2 \cdot n + \beta_n^2 + \left(\frac{M_n \gamma_n}{\alpha_n}\right)^2) \cdot \int_{\mathbb{R}^d \setminus B_n} f(x)\, dx \leq c_{54} \cdot \alpha_n^2 \cdot (\log n)^{4p+6} \cdot n^{-\frac{2p}{2p+d^*}}$$

holds for n sufficiently large and that

$$\int_{S_n^c} g(y)\, dy \leq c_{55} \cdot n^{-1/2}$$

holds. Application of Theorem 3 and Lemma 1 together with the assumptions yields the assertion.

□