

Estimation of a density using an improved surrogate model

Michael Kohler

*Fachbereich Mathematik
Technische Universität Darmstadt
Schlossgartenstr. 7, 64289 Darmstadt, Germany
e-mail: kohler@mathematik.tu-darmstadt.de*

and

Adam Krzyżak*

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

Abstract: Quantification of uncertainty of a technical system is often based on a surrogate model of a corresponding simulation model. In any application the simulation model will not describe the reality perfectly, and consequently also the surrogate model will be imperfect. In this article we show how observed data of the real technical system can be used to improve such a surrogate model, and we analyze the rate of convergence of density estimates based on the improved surrogate model. The results are illustrated by applying the estimates to simulated and real data.

MSC2020 subject classifications: Primary 62G07; secondary 62P30.

Keywords and phrases: Density estimation, imperfect models, L_1 error, surrogate models, uncertainty quantification.

Received September 2019.

Contents

1	Introduction	651
2	A new method for improving an imperfect surrogate model by real data	655
3	Main results	657
4	Application to simulated and real data	662
5	Proofs	669
5.1	Auxiliary results	669
5.1.1	A deterministic lemma	669
5.1.2	A bound on a covering number	670
5.1.3	A bound on the error for smoothing spline estimates for fixed design regression	670
5.1.4	A bound on the deviation between the L_2 error and the empirical L_2 error for smoothing splines	671

*Corresponding author. Tel: +1-514-848-2424 ext. 3007, Fax: +1-514-848-2830.

5.2	A general result on penalized least squares estimates	672
5.3	Proof of Theorem 1	680
	Acknowledgment	682
	References	683
	Appendix	685

1. Introduction

Any design of complex technical systems by engineers nowadays is based on some sort of mathematical model of the technical system. Such models are never able to describe the reality perfectly, therefore their analysis has to take into account some kind of uncertainty. This uncertainty might occur, e.g., because some of the parameters of the model are not exactly known, because of the use of an imperfect mathematical model of the technical system during the design process which does not really describe all aspects of the underlying technical system, or because of lack of knowledge about future use. A good quantification of uncertainty of the system is essential in order to avoid oversizing and to conserve resources.

In this article we quantify uncertainty of the technical system by estimating a density of a real-valued random variable representing the outcome of an experiment with the technical system. The starting point for our estimation problem is a physical model of the technical system with uncertain parameters. This physical model has parameters which are chosen randomly because their exact values are uncertain and consequently unknown, and it computes the outcome of the technical system by computing the value of a function depending on concrete values of these parameters. In case the distribution of the parameters is known (which we will assume from now on) and the function, which has to be computed, is given, Monte Carlo can be used to estimate either quantiles or the density of the output of the technical system.

Usually, the distribution of X is estimated by random sampling using a computer program, and computer experiments can be used to generate values for the Monte Carlo estimates. However, it often happens that generation of the values is rather time consuming, so that standard Monte Carlo estimates cannot be used. 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 vast literature on the design and analysis of such computer experiments, cf., e.g., Santner, Williams and Notz (2003) and the literature cited therein. Often, so-called surrogate models of the computer experiment are used in order to analyze computer experiments. Surrogate models have been introduced and investigated with the aid of the simulated and real data in connection with the quadratic response surfaces in Bucher and Bourgund (1990), Kim and Na (1997) and Das and Zheng (2000), in context of support vector machines in Hurtado (2004), Deheeger and Lemaire (2010) and Bourinet, Deheeger and Lemaire (2011), in connection with neural networks in Papadrakakis and Lagaros (2002), and in context of kriging in Sacks et al.

(1989), Kaymaz (2005) and Bichon et al. (2008). 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).

Computer simulation usually allows to estimate parameters of the computer model to ensure that it closely matches the performance of the technical system. In a Bayesian context, a corresponding approach towards parameter estimation (so-called calibration) was described in Kennedy and O’Hagan (2001). Applications of this approach in various fields can be found, e.g., in Bayarri et al. (2007), Goh et al. (2013), Han, Santner and Rawlinson (2009), Higdon et al. (2013) and Wang, Chen and Tsui (2009). Tuo and Wu (2015) pointed out that this approach might lead to estimates, which are not consistent (in the sense introduced by them) in case of an imperfect computer model, for which there exist no values of the parameters which fit the technical system perfectly. They suggested and analyzed non-Bayesian methods for the choice of parameters of such models.

It is clear that the mathematical/computer model cannot perfectly mimick the performance of the technical system and thus the question arises about the characterization of this mismatch in terms of uncertainty quantification. The standard approach in science is to make some assumptions about the reality, and to try to quantify the uncertainty under these assumptions. E.g., in Bayesian analysis of computer experiments, Kennedy and O’Hagan (2001) and the papers cited above, which applied their methods, suggested to model the discrepancy between the computer experiments and the outcome of the technical system by a Gaussian process. Under the assumption that the reality is described by this Gaussian process perfectly, this enables to compute various estimates and to compute bounds on their error even for small sample sizes. But of course the derived bounds on the error of the estimates do only hold if the assumptions about the reality are true, which illustrates the saying “We buy information with assumptions” (Coombs (1964)).

Due to the fact that erroneous specification of uncertainty in the technical system (e.g., inadequate physical model of the maximum relative compression of the suspension strut of the airplane wheel assembly – see Section 4 and Fig. 5) might result in catastrophic failure of the technical system during its operation (e.g., malfunction of the aircraft’s front wheel during touchdown), it is very important to avoid arbitrary assumptions in uncertainty quantification as much as possible. In this paper we are interested in a non-Bayesian approach towards uncertainty quantification in case of imperfect models. Recently, such an approach was proposed in Wong, Storlie and Lee (2017), where standard nonparametric regression estimates have been applied in order to estimate the discrepancy function between the model and the real data by using these methods to smooth the residuals of the model. The authors used bootstrapping for uncertainty quantification by means of confidence regions and they analysed it using an empirical norm.

There are some similarities between our approach and gradient boosting introduced by Friedman (2001; 2002). As in our approach the residuals are also smoothed in gradient boosting, however in gradient boosting the residuals of original data are smoothed while in our approach the residuals of the new (and better) data are smoothed.

In this paper we use a similar approach as in Wong, Storlie and Lee (2017). We assume that we have available an additional (small) sample of the real technical system, and we consider the problem of estimation from this sample together with the imperfect simulation model an improved surrogate model. But in order to be able to apply our method also for small samples sizes (in our application below we have just $n = 10$ real data points available) we use a weighted penalized least squares estimates where we augment the $n = 10$ residuals of the model by artificial data points. Furthermore, unlike Wong, Storlie and Lee (2017) we analyze the error of our improved surrogate model in the theoretical (and not the empirical) L_2 norm which allows us to derive error bounds for density estimates based on our improved surrogate model.

The mathematical setting which we consider is as follows: Let $(X, Y), (X_1, Y_1), (X_2, Y_2), \dots$ be independent and identically distributed random variables with values in $\mathbb{R}^d \times \mathbb{R}$, and let $m : \mathbb{R}^d \rightarrow \mathbb{R}$ be a measurable function (known and given). Here Y describes the outcome of an experiment with our technical system, and our aim is to estimate the density g of Y (w.r.t. the Lebesgue measure), which we assume to exist. The random vector X with known and given distribution and the measurable function m describe our physical model with uncertain parameters, and in this model we use $m(X)$ as an approximation of Y . Given the data

$$\begin{aligned} &(X_1, Y_1), \dots, (X_n, Y_n), (X_{n+1}, m(X_{n+1})), \dots, (X_{n+L_n}, m(X_{n+L_n})), \\ &X_{n+L_n+1}, \dots, X_{n+L_n+N_n} \end{aligned} \tag{1.1}$$

(where $L_n, N_n \in \mathbb{N}$) our goal is to construct an estimate for g .

The simplest way of solving this problem is to ignore X and m and to use only the data

$$Y_1, \dots, Y_n \tag{1.2}$$

to estimate g by the kernel density estimate

$$\hat{g}_{Y,n}(y) = \frac{1}{n \cdot h_n} \cdot \sum_{i=1}^n K\left(\frac{y - Y_i}{h_n}\right). \tag{1.3}$$

Here $K : \mathbb{R} \rightarrow \mathbb{R}$ (so-called kernel, which is assumed to be a density) and $h_n > 0$ (so-called bandwidth) are parameters of the estimate.

In the sequel we assume that for $m^* : \mathbb{R}^d \rightarrow \mathbb{R}$ defined by $m^*(x) = \mathbf{E}\{Y|X = x\}$ the expected squared error,

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

is small. In this case an alternative way to estimate g is to use the data

$$(X_1, Y_1), \dots, (X_n, Y_n), X_{n+L_n+1}, \dots, X_{n+L_n+N_n} \tag{1.4}$$

to construct an estimate

$$\hat{m}_{(X,Y),n}(\cdot) = \hat{m}_{(X,Y),n}(\cdot, (X_1, Y_1), \dots, (X_n, Y_n)) : \mathbb{R}^d \rightarrow \mathbb{R} \quad (1.5)$$

of m^* , and to define the corresponding surrogate density estimate

$$\hat{g}_{(X,Y),n}(y) = \frac{1}{N_n \cdot h_{N_n}} \cdot \sum_{i=1}^{N_n} K \left(\frac{y - \hat{m}_{(X,Y),n}(X_{n+L_n+i})}{h_n} \right). \quad (1.6)$$

In this article we are interested in situations, where the sample size n is rather small (in our application in Section 4 we will have $n = 10$), since the collection of the real data (1.2) is rather expensive owing to the complex nature of the physical system. Consequently, it might also be useful to use data from our model to estimate g . One possibility of doing this is to define an estimate of g on the basis of the data

$$(X_{n+1}, m(X_{n+1})), \dots, (X_{n+L_n}, m(X_{n+L_n})), X_{n+L_n+1}, \dots, X_{n+L_n+N_n} \quad (1.7)$$

by first estimating a surrogate (often called an emulator in the literature)

$$\begin{aligned} \hat{m}_{(X,m(X)),L_n}(\cdot) = & \quad (1.8) \\ \hat{m}_{(X,m(X)),L_n}(\cdot, (X_{n+1}, m(X_{n+1})), \dots, (X_{n+L_n}, m(X_{n+L_n}))) : & \mathbb{R}^d \rightarrow \mathbb{R} \end{aligned}$$

of m and by subsequently defining the corresponding surrogate density estimate via

$$\hat{g}_{(X,m(X)),L_n}(y) = \frac{1}{N_n \cdot h_{N_n}} \cdot \sum_{i=1}^{N_n} K \left(\frac{y - \hat{m}_{(X,m(X)),L_n}(X_{n+L_n+i})}{h_n} \right). \quad (1.9)$$

The main question which we want to investigate is whether there exist situations in which suitably defined estimates based on the complete data (1.1) achieve simultaneously better rate of convergence than the estimates (1.3), (1.6) and (1.9). Due to the fact that our sample size n is rather small, we will consider only those cases where the model m is rather good, and we will show that in this case we can improve a density estimate based on this model by real data.

To do this, we propose in the next section a novel method for improving the surrogate models (1.5) and (1.8) by using a combination of the real data (1.4) and the simulation data (1.7). Our main result is that the rate of convergence of the corresponding surrogate density estimate is at least as good as the rates of convergence of the density estimates (1.3), (1.6) and (1.9), and is, in special situations, better than any of the above rates of convergence. The finite sample size behaviour of our estimates is validated on simulated data (see Section 4). The usefulness of our newly proposed estimates for uncertainty quantification is demonstrated by using it to analyze the uncertainty occurring in experiments with a suspension strut.

Throughout this paper we use the following notation: \mathbb{N} , \mathbb{N}_0 and \mathbb{R} are the sets of positive integers, nonnegative integers and real numbers, respectively.

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

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

for all $x, z \in \mathbb{R}^d$. 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 $f : \mathbb{R}^d \rightarrow \mathbb{R}$ be a real-valued function defined on \mathbb{R}^d .

The outline of this paper is as follows: In Section 2 the construction of the improved surrogate model is explained. The main results are presented in Section 3 and proven in Section 5. The finite sample size performance of our estimates is illustrated in Section 4 by applying it to simulated and real data.

2. A new method for improving an imperfect surrogate model by real data

In this section we describe our ideas behind the construction of the improved surrogate model.

In order to construct density estimates on the basis of the data (1.1) we proceed as follows: We start by defining a surrogate model

$$\hat{m}_{L_n}(\cdot) = \hat{m}_{L_n}(\cdot, (X_{n+1}, m(X_{n+1})), \dots, (X_{n+L_n}, m(X_{n+L_n}))) : \mathbb{R}^d \rightarrow \mathbb{R} \quad (2.1)$$

of m . The reason behind using this surrogate model instead of m is that it might be much more complicated to evaluate m (because, e.g., we may have to solve a partial differential equation in order to compute a function value) than to evaluate the surrogate model. In principle, any nonparametric regression estimate can be used to construct the surrogate model. In Section 4 we will use the penalized least squares estimate defined by

$$\tilde{m}_{L_n, (k, \lambda_{L_n})}(\cdot) = \arg \min_{f \in W^k(\mathbb{R}^d)} \left(\frac{1}{L_n} \sum_{i=n+1}^{n+L_n} |f(X_i) - m(X_i)|^2 + \lambda_{L_n} \cdot J_k^2(f) \right), \quad (2.2)$$

where $k \in \mathbb{N}$ with $2k > d$ and

$$J_k^2(f) = \sum_{\alpha_1, \dots, \alpha_d \in \mathbb{N}, \alpha_1 + \dots + \alpha_d = k} \frac{k!}{\alpha_1! \dots \alpha_d!} \int_{\mathbb{R}^d} \left| \frac{\partial^k f}{\partial x_1^{\alpha_1} \dots \partial x_d^{\alpha_d}}(x) \right|^2 dx$$

is a penalty penalizing the roughness of the estimate and where $W^k(\mathbb{R}^d)$ denotes the Sobolev space

$$\left\{ f : \frac{\partial^k f}{\partial x_1^{\alpha_1} \dots \partial x_d^{\alpha_d}} \in L_2(\mathbb{R}^d) \text{ for all } \alpha_1, \dots, \alpha_d \in \mathbb{N} \text{ with } \alpha_1 + \dots + \alpha_d = k \right\}.$$

The condition $2k > d$ implies that the functions in $W^k(\mathbb{R}^d)$ are continuous and hence the value of a function at a point is well defined. In order to be able to analyze the rate of convergence of this estimate for an arbitrary distribution of X and dimension $d > 1$ we will truncate this estimate at some level $\beta > 0$, i.e., we define

$$\hat{m}_{L_n}(x) = T_\beta(\tilde{m}_{L_n, (k, \lambda_{L_n})}(x)) \quad (x \in \mathbb{R}^d), \quad (2.3)$$

where

$$T_\beta z = \begin{cases} \beta, & z > \beta \\ z, & -\beta \leq z \leq \beta \\ -\beta, & z < -\beta \end{cases}$$

for $z \in \mathbb{R}$. If we know the bound on the estimated function in an application we can truncate our estimate at the same level.

Next we compute the residuals on the data set (1.4) of the estimate (2.1), i.e., we compute

$$\hat{\epsilon}_i = Y_i - \hat{m}_{L_n}(X_i) \quad (i = 1, \dots, n). \quad (2.4)$$

Then we define an estimate

$$\hat{m}_n^{\hat{\epsilon}}(\cdot) : \mathbb{R}^d \rightarrow \mathbb{R} \quad (2.5)$$

which smoothes these residuals (see below) and define our final surrogate model $(X, \hat{m}_n(X))$ (where \hat{m}_n is an estimate of $m^*(\cdot) = \mathbf{E}\{Y|X = \cdot\}$) by setting

$$\hat{m}_n(x) = \hat{m}_{L_n}(x) + \hat{m}_n^{\hat{\epsilon}}(x) \quad (x \in \mathbb{R}^d). \quad (2.6)$$

In order to define the estimate (2.5) we use two data sets: The first data set corresponding to the residuals of \hat{m}_{L_n} on X_1, \dots, X_n , i.e., the data set

$$\{(X_1, \hat{\epsilon}_1), \dots, (X_n, \hat{\epsilon}_n)\} = \{(X_1, Y_1 - \hat{m}_{L_n}(X_1)), \dots, (X_n, Y_n - \hat{m}_{L_n}(X_n))\}. \quad (2.7)$$

And the second data set corresponding to the residuals of \hat{m}_{L_n} on the artificial sample with measurement errors

$$\{(X_{n+L_n+1}, \hat{m}_{L_n}(X_{n+L_n+1})), \dots, (X_{n+L_n+N_n}, \hat{m}_{L_n}(X_{n+L_n+N_n}))\} \quad (2.8)$$

of (X, Y) , i.e., the data set

$$\{(X_{n+L_n+1}, 0), \dots, (X_{n+L_n+N_n}, 0)\}. \quad (2.9)$$

The second data set will be, in particular, useful in case when \hat{m}_{L_n} is already very close to

$$m^* : \mathbb{R}^d \rightarrow \mathbb{R}, m^*(x) = \mathbf{E}\{Y|X = x\},$$

since the sample size n of the data set (2.7) might be too small to detect that $0 \approx m^* - \hat{m}_{L_n}$ is the optimal choice for $\hat{m}_n^{\hat{\epsilon}}$.

Since both data sets are not equally trustworthy, we weigh them by some weight $w^{(n)} \in [0, 1]$, and set

$$\begin{aligned} \tilde{m}_n^{\hat{\epsilon}}(\cdot) &= \tilde{m}_n^{\hat{\epsilon}}(\cdot, (X_{n+1}, m(X_{n+1})), \dots, (X_{n+L_n}, m(X_{n+L_n})), \\ &\quad X_{n+L_n+1}, \dots, X_{n+L_n+N_n}) \\ &= \arg \min_{f \in W^k(\mathbb{R}^d)} \left(\frac{w^{(n)}}{n} \sum_{i=1}^n (\hat{\epsilon}_i - f(X_i))^2 + \frac{1-w^{(n)}}{N_n} \sum_{i=1}^{N_n} (0 - f(X_{n+L_n+i}))^2 \right. \\ &\quad \left. + \lambda_n \cdot J_k^2(f) \right) \end{aligned} \tag{2.10}$$

and

$$\hat{m}_n^{\hat{\epsilon}}(x) = T_{c_1, \alpha_n}(\tilde{m}_n^{\hat{\epsilon}}(x)) \quad (x \in \mathbb{R}^d), \tag{2.11}$$

where $c \geq 1$ and $\alpha_n > 0$. Finally, we use $(X, \hat{m}_n(X))$ as a surrogate model for (X, Y) and estimate the density g of Y by applying a kernel density estimate to a sample of $\hat{m}_n(X)$. To do this, we choose a kernel $K : \mathbb{R} \rightarrow \mathbb{R}$ and a bandwidth $h_{N_n} > 0$ and define

$$\hat{g}_{N_n}(y) = \frac{1}{N_n \cdot h_{N_n}} \cdot \sum_{i=1}^{N_n} K\left(\frac{y - \hat{m}_n(X_{n+L_n+i})}{h_{N_n}}\right). \tag{2.12}$$

The procedure is summarized in Algorithm 1.

3. Main results

In the next theorem we present bounds on the rate of convergence of our surrogate estimate, which we will use to derive bounds on the rate of convergence of our density estimate. In principle, all of our error bounds are also valid for finite n . In order to simplify the presentation, we consider the case $n \rightarrow \infty$ and assume that the distribution of (X, Y) and also the physical model with uncertain parameters $(X, m(X))$ change for increasing n such that $Y - m^*(X)$ and the error $m(X) - m^*(X)$ get smaller with increasing n . In order to simplify the notation we write (X, Y) and m instead of $(X^{(n)}, Y^{(n)})$ and $m^{(n)}$, resp.

In order to derive our main result we will use the following assumptions:

- (A1) $(X, Y), (X_1, Y_1), \dots$ are independent and identically distributed $\mathbb{R}^d \times \mathbb{R}$ -valued random variables such that $\text{supp}(X)$ is bounded and $\mathbf{E}\{Y^2\} < \infty$.
- (A2) For some $\alpha_n^* \geq 0$ and $m^*(x) = \mathbf{E}\{Y|X = x\}$ we have

$$\mathbf{E}\{|Y - m^*(X)|^2\} \leq (\alpha_n^*)^2 \quad \text{and} \quad \mathbf{E}\{|Y - m^*(X)|^3\} \leq (\alpha_n^*)^3. \tag{3.1}$$

- (A3) We are given some measurable function $m : \mathbb{R}^d \rightarrow \mathbb{R}$ satisfying for some $c_2 \in \mathbb{R}_+$ and $1 \leq \beta_n \leq n + L_n$

$$|m(x)| \leq \beta_n \quad (x \in \mathbb{R}^d) \quad \text{and} \quad J_k^2(m) \leq c_2 < \infty. \tag{3.2}$$

Collect real data $(X_1, Y_1), \dots, (X_n, Y_n)$ from experiment with the physical system; generate additional data $(X_{n+1}, m(X_{n+1})), \dots, (X_{n+L_n}, m(X_{n+L_n}))$ using known computer model $m(x)$;
 generate additional data $X_{n+L_n+1}, \dots, X_{n+L_n+N_n}$ from the distribution of X ;
 Compute the penalized least squares regression estimate
 $\tilde{m}_{L_n, (k, \lambda_{L_n})}(\cdot) = \arg \min_{f \in W^k(\mathbb{R}^d)} \left(\frac{1}{L_n} \sum_{i=n+1}^{n+L_n} |f(X_i) - m(X_i)|^2 + \lambda_{L_n} \cdot J_k^2(f) \right)$,
 Truncate $\tilde{m}_{L_n, (k, \lambda_{L_n})}(\cdot)$ at level $\pm\beta$ yielding
 $\hat{m}_{L_n}(x) = T_\beta(\tilde{m}_{L_n, (k, \lambda_{L_n})}(x)) \quad (x \in \mathbb{R}^d)$
 Compute residuals for the real data
 $\hat{\epsilon}_i = Y_i - \hat{m}_{L_n}(X_i) \quad (i = 1, \dots, n)$
 Compute the penalized least squares regression estimate weighing residuals corresponding to real and artificial data

$$\tilde{m}_n^\epsilon(\cdot) = \arg \min_{f \in W^k(\mathbb{R}^d)} \left(\frac{w^{(n)}}{n} \sum_{i=1}^n (\hat{\epsilon}_i - f(X_i))^2 + \frac{1 - w^{(n)}}{N_n} \sum_{i=1}^{N_n} (0 - f(X_{n+L_n+i}))^2 + \lambda_n \cdot J_k^2(f) \right)$$

Truncate estimate \tilde{m}_n^ϵ yielding
 $\hat{m}_n^\epsilon(x) = T_{c_1 \cdot \alpha_n}(\tilde{m}_n^\epsilon(x)) \quad (x \in \mathbb{R}^d)$,
 Define final surrogate estimate by
 $\hat{m}_n(x) = \hat{m}_{L_n}(x) + \hat{m}_n^\epsilon(x)$
 Use $(X, \hat{m}_n(X))$ as a surrogate model for (X, Y) and estimate the density g by
 $\hat{g}_{N_n}(y) = \frac{1}{N_n \cdot h_{N_n}} \cdot \sum_{i=1}^{N_n} K \left(\frac{y - \hat{m}_n(X_{n+L_n+i})}{h_{N_n}} \right)$. Set as **result** the value $\hat{g}_{N_n}(y)$
 return **result**;

Algorithm 1: Proposed estimate of the density g of Y using a surrogate model for (X, Y) .

(A4) For some $\alpha_n > \alpha_n^*$ (where α_n^* is the value from (A2)) and for some $K, \sigma_0 > 0$ we have

$$K^2 \cdot \left(\mathbf{E} \left\{ \exp \left(\frac{(Y - m^*(X))^2}{\alpha_n \cdot K} \right) \middle| X \right\} - 1 \right) \leq \sigma_0 \quad a.s., \quad (3.3)$$

and the regression function $\mathbf{E}\{Y - m(X) | X = x\} = (m^* - m)(x)$ satisfies

$$\sup_{x \in \mathbb{R}^d} |m^*(x) - m(x)| \leq \alpha_n \quad (3.4)$$

and

$$J_k^2(m^* - m) \leq (\alpha_n)^2. \quad (3.5)$$

Here (A1) imposes some mild condition on the data $(X_1, Y_1), \dots, (X_n, Y_n)$ from our experiments with the technical system. (A2) describes how well Y can be approximated by a function of X , and α_n^* is an upper bound on the error of this approximation. The simplest (but most restrictive) case is that (A2) is satisfied for $\alpha_n^* = 0$ which means that $Y = m^*(X)$ holds. In (A3) the function $m : \mathbb{R}^d \rightarrow \mathbb{R}$ describes the simulation model which we use. Here we impose

smoothness assumptions on this function. Our main assumption is (A4): Here α_n is a bound on the error of our model m (when we compare it with m^*), and we impose a smoothness assumption on $m^* - m$.

Theorem 1. *Let $d, k \in \mathbb{N}$ with $2k > d$. Let $n \in \mathbb{N}$ with $n \geq 2$ and let $L_n, N_n \in \mathbb{N}$ with $n \leq L_n \leq N_n$ and let $\alpha_n > \alpha_n^* \geq 0$. Assume that (A1), (A2), (A3) and (A4) hold. Furthermore, assume that*

$$\alpha_n \geq \frac{1}{n^l} \quad \text{for some } l \in \mathbb{N}. \tag{3.6}$$

Define $m_{L_n, (k, \lambda)}$ by (2.2) and (2.3), where

$$\lambda_{L_n} = c_3 \cdot \left(\frac{\log L_n}{L_n} \right)^{\frac{2k}{2k+d}},$$

and assume that

$$\alpha_n \geq \left(\frac{\log L_n}{L_n} \right)^{\frac{2k}{3(2k+d)}}.$$

Define $\hat{m}_n^{\hat{\epsilon}}$ by (2.10) and (2.11) for some N_n satisfying $N_n \leq c_4 \cdot n^l$ for some $l \in \mathbb{N}$, choose $\lambda_n > 0$ such that

$$\frac{\log n}{n} \leq \lambda_n \leq \left(\frac{\log n}{n} \right)^{\frac{2k}{d}},$$

let $w^{(n)} \in [0, 1]$ and define \hat{m}_n by (2.6). Then there exist constants $c_5, \dots, c_{10} \in \mathbb{R}_+$ such that

$$\begin{aligned} & \mathbf{E} \{ |Y - \hat{m}_n(X)|^2 \} \\ & \leq c_5 \cdot (\alpha_n^*)^2 + c_6 \cdot \alpha_n^2 \cdot \lambda_n + c_7 \cdot w^{(n)} \cdot \alpha_n^2 \cdot \frac{\log n}{n \cdot \lambda_n^{d/(2k)}} + c_8 \cdot \left(\frac{\log L_n}{L_n} \right)^{\frac{2k}{2k+d}} \\ & \quad + c_9 \cdot (1 - w^{(n)}) \cdot \alpha_n^2 \left(1 + \frac{\log N_n}{N_n \cdot \lambda_n^{d/(2k)}} \right) + \frac{c_{10} \cdot \alpha_n^2}{\min\{n, N_n\}} + \frac{c_{10}}{L_n}. \end{aligned}$$

In particular, in case $w^{(n)} = 1$ and $\lambda_n = c_{11} \cdot ((\log n)/n)^{2 \cdot k/(2 \cdot k+d)}$ we get

$$\mathbf{E} \{ |Y - \hat{m}_n(X)|^2 \} \leq c_{12} \cdot \max \left\{ (\alpha_n^*)^2, \alpha_n^2 \cdot \left(\frac{\log n}{n} \right)^{\frac{2k}{2k+d}}, \left(\frac{\log L_n}{L_n} \right)^{\frac{2k}{2k+d}} \right\}.$$

Remark 1. In any application of the estimate in Theorem 1 we have to choose the parameters depending on the data. In Section 4 we will use k -fold cross-validation applied to the data $(X_1, Y_1), \dots, (X_n, Y_n)$ in order to choose $w^{(n)}$ and λ_n , and we choose λ_{L_n} by generalized cross-validation applied the data $(X_{n+i}, m(X_{n+i}))$ ($i = 1, \dots, L_n$).

Theorem 1 implies the following corollary concerning the L_1 error of the density estimate (2.12):

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

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

holds. Set

$$h_{N_n} = c_{13} \cdot \left(\alpha_{N_n} \cdot \left(\frac{\log N_n}{N_n} \right)^{\frac{k}{2k+d}} \right)^{\frac{1}{r+1}}$$

and assume

$$N_n \geq \frac{1}{\alpha_n^{(2r+1)/(r+1)}} \cdot \left(\frac{n}{\log n} \right)^{\frac{k}{2k+d} \cdot \frac{2r+1}{r+1}}.$$

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

$$\mathbf{E} \int_{\mathbb{R}} |\hat{g}_{N_n}(y) - g(y)| dy \leq c_{14} \cdot \left(\alpha_n \cdot \left(\frac{\log n}{n} \right)^{\frac{k}{2k+d}} \right)^{\frac{r}{r+1}}$$

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

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

Consequently,

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

satisfies

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

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

$$\begin{aligned} & \mathbf{E} \int_{\mathbb{R}} |\hat{g}_{N_n}(y) - g(y)| dy \\ & \leq \mathbf{E} \int_{\mathbb{R}} |\hat{g}_{N_n}(y) - \hat{g}_{Y, N_n}(y)| dy + \mathbf{E} \int_{\mathbb{R}} |\hat{g}_{Y, N_n}(y) - g(y)| dy \\ & \leq \frac{2 \cdot K(0)}{h_{N_n}} \cdot \mathbf{E} \{ |\hat{m}_n(X) - Y| \} + \frac{c_{15}}{\sqrt{N_n \cdot h_{N_n}}} + c_{16} \cdot h_{N_n}^r \end{aligned}$$

$$\leq \frac{2 \cdot K(0)}{h_{N_n}} \cdot \sqrt{\mathbf{E}\{|\hat{m}_n(X) - Y|^2\}} + \frac{c_{15}}{\sqrt{N_n \cdot h_{N_n}}} + c_{16} \cdot h_{N_n}^r.$$

Application of Theorem 1 yields the assertion. □

Remark 2. It is well-known that the L_1 error of the standard kernel density applied to the data (1.2) achieves under the assumptions of Corollary 1 the (optimal) rate of convergence

$$n^{-r/(2r+1)}.$$

It follows from the proof of Corollary 1 (together with standard error bounds on the L_2 error of smoothing spline estimates, cf., e.g., Chapter 21 in Györfi et al. (2002)), that the L_1 errors of the surrogate density estimates defined in (1.6) and (1.9) achieve under the assumptions of Corollary 1 the rates of convergence

$$\left(\alpha_n^* + \left(\frac{\log n}{n}\right)^{\frac{k}{2k+d}}\right)^{\frac{r}{r+1}} \quad \text{and} \quad (\alpha_n)^{\frac{r}{r+1}}, \text{ respectively.}$$

For α_n suitably small the bound on the rate of convergence in Corollary 1 converges faster to zero than any of the above rates of convergence, which proves, that there exist situations in which our estimate theoretically outperforms the estimates defined in (1.3), (1.6) and (1.9). So our main result is that in case that our model m is really good, we can improve the rate of convergence of a density estimate based on this model by incorporating real data. In the next section we demonstrate with simulated data that this is also the case for finite sample sizes.

Remark 3. In case that the error α_n of our simulation model does not converge to zero quickly enough, it might happen that the rate of convergence to zero in Corollary 1 is slower than the rate $n^{-r/(2r+1)}$ of the standard kernel density applied to the data (1.2). For large n it is easy to modify our estimate such that it always achieves at least the rate of convergence of this standard kernel density estimate: To do this we can split the data (1.2) in two parts of approximately equal size, define our estimate in Corollary 1 and the standard kernel density estimate by using only the first part of this data, and use the second part to choose via the combinatorial method in density estimation of Devroye and Lugosi (2001) the better of this estimate. By using the results presented in Devroye and Lugosi (2001) it is easy to see that the L_1 error of the resulting density estimate converges to zero with the rate

$$\min \left\{ \left(\alpha_n \cdot \left(\frac{\log n}{n} \right)^{\frac{k}{2k+d}} \right)^{\frac{r}{r+1}}, n^{-\frac{r}{2r+1}} \right\}$$

under the assumptions of Corollary 1. Since in this paper we are interested in applications, where the sample size is rather small (we have $n = 10$ in our real data application in the next section), we are interested in situations where the standard kernel density estimate cannot produce satisfying results, and therefore

we do not incorporate this splitting of the sample in our simulations in the next section.

Remark 4. The rate of convergence in Corollary 1 gets worse in case that the dimension d of X increases. This is a consequence of the well-known curse of dimensionality in nonparametric regression, which states that estimation of functions in high-dimensions is particular difficult. In our method we use smoothing splines in order to estimate a surrogate model of our model m and the difference between m and m^* . It is well-known that this kind of estimate usually requires $d \in \{1, 2, 3\}$ in order to produce satisfying results. For larger d one has to use other techniques from function estimation in our procedure, and to impose structural assumptions on the functions to be estimated in order to get satisfying results. How this can be achieved with neural networks in the context of the newly proposed density estimate is described in Götz, Kersting and Kohler (2020).

4. Application to simulated and real data

In this section we illustrate the finite sample size performance of our estimates by applying them to simulated and real data.

We begin with the simulated data, which we use to illustrate how the size of the error of the model influences the performance of our estimate. To do this, we choose X with d -dimensional standard normal distribution and ϵ uniformly distributed on $[0, 1]$ such that X and ϵ are independent and set

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

for some $m : \mathbb{R}^d \rightarrow \mathbb{R}$ defined below and $\sigma \in \{0.1, 0.5, 1\}$, and let $(X_1, Y_1), (X_2, Y_2), \dots$ be independent and identically distributed random variables. Our estimate uses the data

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

from the real technical system as well as data

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

from the (imperfect) computer model (where σ controls the maximal error occurring in this model), and the additional X -values

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

If we compare this setting with Theorem 1 we see that following the notation of Theorem 1 we have $m^*(x) = \mathbf{E}\{Y|X = x\} = m(x) + \frac{1}{2} \cdot \sigma$ and it is easy to see that in this setting the assumptions of Theorem 1 are satisfied if we choose $\alpha_n^* = \sigma/(32)^{1/3}$ and $\alpha_n = \sigma/2$.

In all of our applications we choose $n \in \{10, 20, 40\}$ and $L_n = 500$. As surrogate estimate we use a thin plate spline as implemented in the routine *Tps()* in the statistics software *R*, where we use 5-fold cross-validation (applied

to the data \mathcal{D}_n) to choose the degree of freedom df of the fitted spline from the set $\{4, 8, 16, \dots, 256\}$. In the same way we also choose $w^{(n)}$ from the set $\{0, 0.1, \dots, 1\}$, i.e., we choose simultaneously the degree of freedom df and the weight $w^{(n)}$ by 5-fold cross-validation. During the 5-fold cross-validation we compute an estimate of the L_1 error of our smoothing spline estimate.

For our newly proposed density estimate we use a sample of size $N_n = 500,000$ of $\hat{m}_n(X)$ (where \hat{m}_n is the estimate introduced in Section 2) and construct a kernel density estimate based on this sample with L_2 cross-validation as implemented in the routine `density()` in the statistics package *R*. Here the bandwidth is chosen as the maximum of the bandwidth chosen by L_2 cross-validation for the kernel density estimate (as implemented in *R*), and two times the estimated absolute L_1 error of our smoothing spline estimates.

The density of Y is the convolution of the density of $m(X)$ and a uniform density. We do not try to compute its exact form, instead we compute it approximately by applying a kernel density estimate (as implemented in the routine `density()` in *R*) with a sample of size 1,000,000 of Y . In order to judge the quality of our density estimates the resulting density is treated in our simulations as if it were the real density.

We compare our estimate (*est. 4*) with four other density estimates. The first one (*est. 1*) is the standard kernel density estimate as implemented in *R* applied to the sample of size n of Y , cf., (1.3). The estimates two and three are surrogate density estimates, where the kernel density estimate of *R* is used with a sample of size $N_n = 500,000$ of the surrogate model. For *est. 2* the surrogate model is chosen by applying a thin plate spline (as implemented in *R*) to the sample of size n of (X, Y) , cf., (1.6). And for *est. 3* the surrogate model is computed in the same way, but using this time the sample of size $L_n = 500$ of our model $(X, m(X))$, cf., (1.9). *est. 5* is our newly proposed estimate, but this time with weight $w^{(n)} = 1$ chosen independent of the data.

We consider three different models. In the first model we choose $d = 2$ and define it by

$$m(x_1, x_2) = 2 \cdot x_1 + x_2 + 2.$$

Figure 1 shows plots of the reference density and of the five different density estimates for a data set of model 1, where we use $n = 20$, $\sigma = 0.5$, $L_n = 500$ and $N_n = 500,000$.

In the second model we choose again $d = 2$, but define m this time by

$$m(x_1, x_2) = x_1^2 + x_2^2.$$

Figure 2 shows plots of the reference density and of the five different density estimates for a data set of model 2, where we use $n = 20$, $\sigma = 0.5$, $L_n = 500$ and $N_n = 500,000$.

In the third model we choose $d = 1$ and define m by

$$m(x) = \exp(x).$$

Figure 3 shows plots of the reference density and of the five different density estimates for a data set of model 3, where we use $n = 20$, $\sigma = 0.5$, $L_n = 500$ and $N_n = 500,000$.

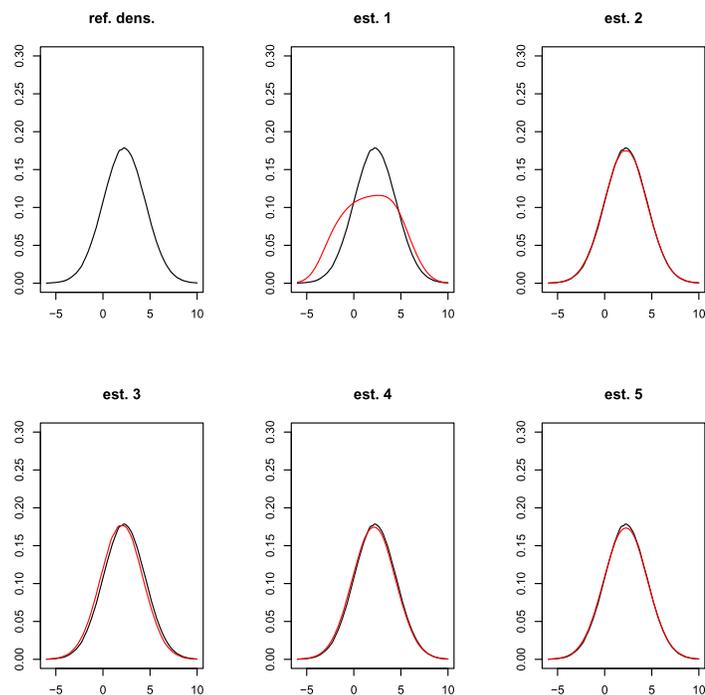


FIG 1. The reference density and the five different density estimates (in red) in simulation from model 1 with parameters $n = 20$, $\sigma = 0.5$, $L_n = 500$ and $N_n = 500,000$.

We compare the L_1 errors of our five different estimates. To do this, we approximate the integral by a Riemann sum defined on an equidistant partition consisting of 8192 subintervals of the interval $[-6, 10]$ (in model 1) or the interval $[0, 10]$ (in models 2 and 3). Since this L_1 error is random, we repeat each simulation 100 times and report in Table 1 the median (and in brackets the interquartile range) of the 100 L_1 errors for each of our four estimates.

From Table 1 we see that our estimate outperforms the estimates 1 through 3 in 19 out of 27 settings, and in these cases often its error is between 2 to 3 times smaller than the errors of all other estimates. And in the five cases where it does not achieve the smallest error compared with estimates 1 through 3, its error is approximately of the same size as the smallest error (and at most 30 percent larger). These larger errors occur only in model 1, where the function m is a linear function which can be easily estimated even from a small sample of observation, and where therefore the surrogate density estimate based on an estimated surrogate is rather good.

By comparing the performance of estimates 4 and 5 in Table 1, we can furthermore get an idea about the usefulness of our choice for the data-dependent weight. Here we see that estimate 4 is better than estimate 5 in 11 cases, and estimate 5 is better than estimate 4 in 11 cases. However, this occurs for different

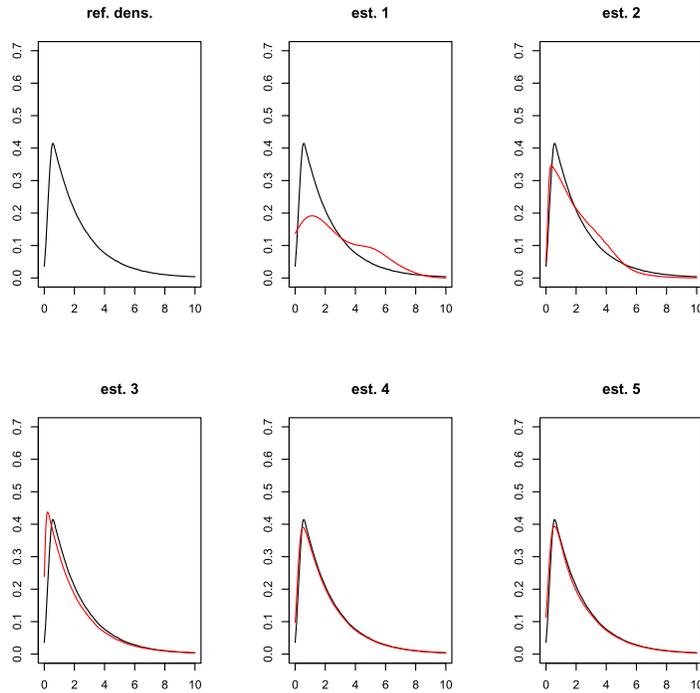


FIG 2. The reference density and the five different density estimates (in red) in simulation from model 1 with parameters $n = 20$, $\sigma = 0.5$, $L_n = 500$ and $N_n = 500,000$.

sample sizes: for $n = 10$ estimate 4 with the data-dependent weight is 6 times better (and only 1 time worse) than estimate 5 (with weight $w^{(n)} = 1$). But for $n = 40$ estimate 4 is only 2 times better (and 5 times worse) than estimate 5. From this we see that our data-dependent choice of the weight is useful for very small sample size (which is the most important case in practice).

The merits of the new uncertainty quantification approach proposed in this paper are reinforced by applying it to the real data obtained in the experimental study of the suspension strut illustrated in Figure 4. This assembly has been used to study uncertainty in load distributions and feasibility to control vibrations, stability and load paths in suspension struts such as aircraft landing gears. The structure of the strut assembly is shown in the left panel of Figure 5. The right panel presents the simplified physical model of the suspension strut. This suspension strut consists of an upper structure and lower structure. The lower structure contains a spring-damper component, which serves as mediating agent transmitting axial forces between the upper structure and the foot of the suspension strut. Our goal is study the impact of the randomly selected free-fall height h_f on the maximum relative compression of the spring-damper component. To that end we assume that the free-fall heights are independent normally distributed random variables with mean 0.05 meter and standard deviation 0.0057 meter.

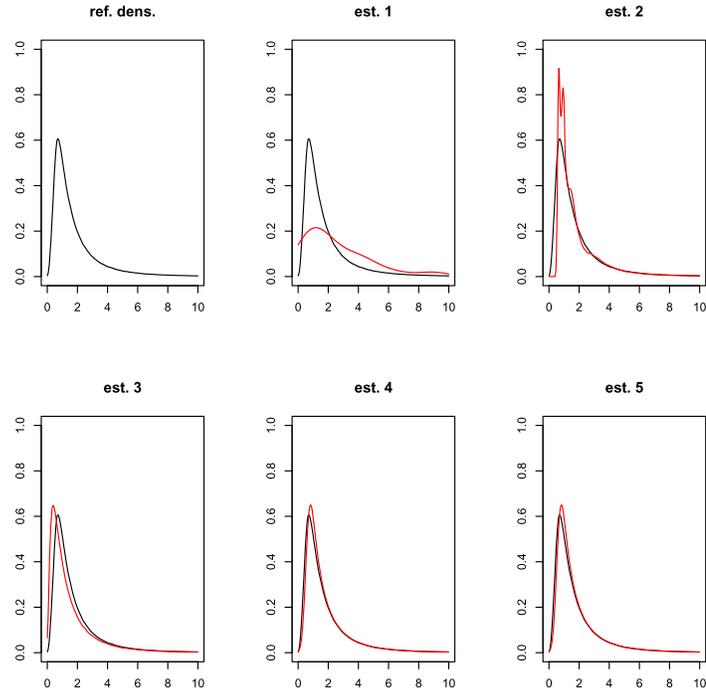


FIG 3. The reference density and the five different density estimates (in red) in simulation from model 3 with parameters $n = 20$, $\sigma = 0.5$, $L_n = 500$ and $N_n = 500,000$.

We analyze the uncertainty in the maximum relative compression in our suspension strut using a simplified mathematical/physical model of the suspension strut (cf., Figure 5 (right panel)). In the model the upper and the lower structures of the suspension strut are represented by the lump masses m and m_1 , respectively, the spring-damper component is characterized by the stiffness and damping coefficients k and b , respectively, and elasticity of the foot component is described by the stiffness parameter k_{ef} . Assuming linear stiffness and an axiomatic damping it is possible to compute the maximum relative compression by solving an ordinary differential equation using Runge-Kutta algorithm (cf., model a) in Mallapur and Platz (2017)). We use $L_n = 500$ samples generated in computer experiments to construct a surrogate estimate \hat{n}_{L_n} described above.

In Figure 6 we see in the upper left panel data from $L_n = 500$ computer experiments together with the corresponding surrogate model (solid line), and in the upper right panel the corresponding surrogate density estimate. In the lower left panel we see again the surrogate model (dashed-dotted) based on the data from the computer experiments together with $n = 10$ real data points from the experiment. Clearly, our (dashed-dotted) surrogate model based only on the computer experiment is imperfect since it does not really fit the real data. We can improve this imperfect surrogate model by using the techniques

TABLE 1
Simulation results in the three different models.

model	d	σ	n	est. 1	est. 2	est. 3	est. 4	est. 5
1	2	0.1	10	0.373 (0.205)	0.010 (0.004)	0.019 (0.003)	0.010 (0.005)	0.010 (0.006)
1	2	0.1	20	0.272 (0.125)	0.009 (0.003)	0.019 (0.003)	0.009 (0.003)	0.009 (0.003)
1	2	0.1	40	0.199 (0.111)	0.008 (0.002)	0.019 (0.003)	0.008 (0.003)	0.008 (0.003)
1	2	0.5	10	0.356 (0.220)	0.035 (0.031)	0.090 (0.003)	0.038 (0.030)	0.043 (0.036)
1	2	0.5	20	0.303 (0.126)	0.022 (0.017)	0.090 (0.003)	0.025 (0.017)	0.023 (0.015)
1	2	0.5	40	0.180 (0.095)	0.014 (0.011)	0.090 (0.003)	0.017 (0.014)	0.014 (0.011)
1	2	1	10	0.365 (0.188)	0.067 (0.047)	0.178 (0.003)	0.065 (0.057)	0.065 (0.059)
1	2	1	20	0.299 (0.146)	0.045 (0.043)	0.178 (0.003)	0.048 (0.035)	0.045 (0.037)
1	2	1	40	0.209 (0.095)	0.028 (0.026)	0.178 (0.003)	0.036 (0.026)	0.031 (0.019)
2	2	0.1	10	0.432 (0.203)	0.308 (0.181)	0.034 (0.002)	0.015 (0.013)	0.014 (0.013)
2	2	0.1	20	0.326 (0.149)	0.140 (0.060)	0.034 (0.002)	0.014 (0.008)	0.012 (0.008)
2	2	0.1	40	0.269 (0.082)	0.063 (0.034)	0.034 (0.002)	0.012 (0.006)	0.011 (0.006)
2	2	0.5	10	0.400 (0.208)	0.333 (0.143)	0.188 (0.003)	0.066 (0.041)	0.074 (0.037)
2	2	0.5	20	0.303 (0.115)	0.200 (0.091)	0.188 (0.003)	0.047 (0.024)	0.043 (0.018)
2	2	0.5	40	0.257 (0.098)	0.112 (0.047)	0.188 (0.003)	0.040 (0.021)	0.040 (0.017)
2	2	1	10	0.429 (0.203)	0.375 (0.226)	0.351 (0.003)	0.117 (0.058)	0.121 (0.057)
2	2	1	20	0.295 (0.141)	0.232 (0.087)	0.352 (0.003)	0.087 (0.043)	0.088 (0.048)
2	2	1	40	0.252 (0.101)	0.173 (0.058)	0.352 (0.003)	0.074 (0.023)	0.071 (0.023)
3	1	0.1	10	0.425 (0.187)	0.135 (0.066)	0.064 (0.003)	0.043 (0.039)	0.058 (0.075)
3	1	0.1	20	0.380 (0.139)	0.095 (0.070)	0.064 (0.002)	0.033 (0.026)	0.028 (0.024)
3	1	0.1	40	0.299 (0.094)	0.065 (0.030)	0.064 (0.003)	0.022 (0.019)	0.019 (0.012)
3	1	0.5	10	0.432 (0.171)	0.333 (0.197)	0.304 (0.003)	0.136 (0.065)	0.168 (0.080)
3	1	0.5	20	0.353 (0.136)	0.236 (0.136)	0.304 (0.003)	0.108 (0.052)	0.112 (0.056)
3	1	0.5	40	0.304 (0.082)	0.193 (0.077)	0.304 (0.003)	0.075 (0.042)	0.079 (0.040)
3	1	1	10	0.439 (0.188)	0.441 (0.211)	0.528 (0.003)	0.214 (0.105)	0.265 (0.106)
3	1	1	20	0.350 (0.115)	0.370 (0.195)	0.529 (0.003)	0.156 (0.059)	0.159 (0.060)
3	1	1	40	0.256 (0.097)	0.310 (0.145)	0.528 (0.003)	0.135 (0.051)	0.137 (0.052)

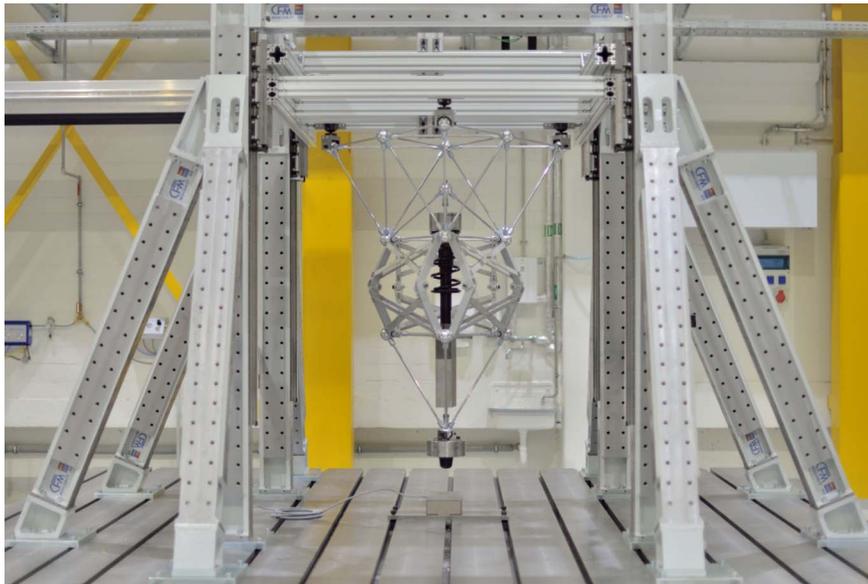


FIG 4. A photo of the demonstrator of a suspension strut and its experimental test setup.

introduced in this paper and effectively obtain the solid line in the lower left panel of Figure 6. The corresponding surrogate density estimate is shown in the lower right panel of Figure 6. We observe that the use of $n = 10$ addi-

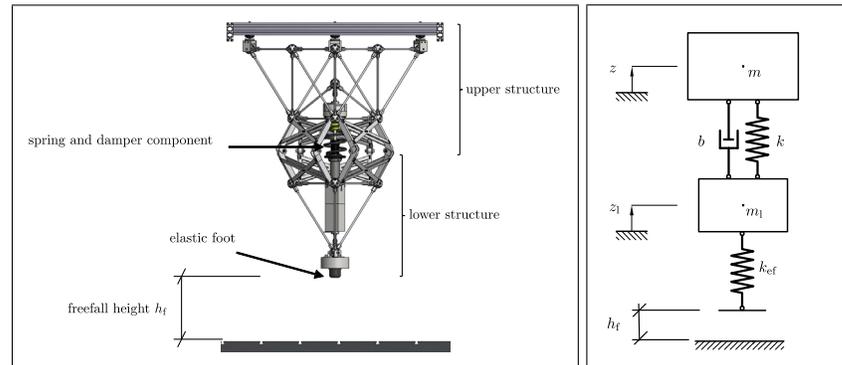


FIG 5. A CAD illustration of the suspension strut (left panel) and simplified physical model of the suspension strut (right panel).

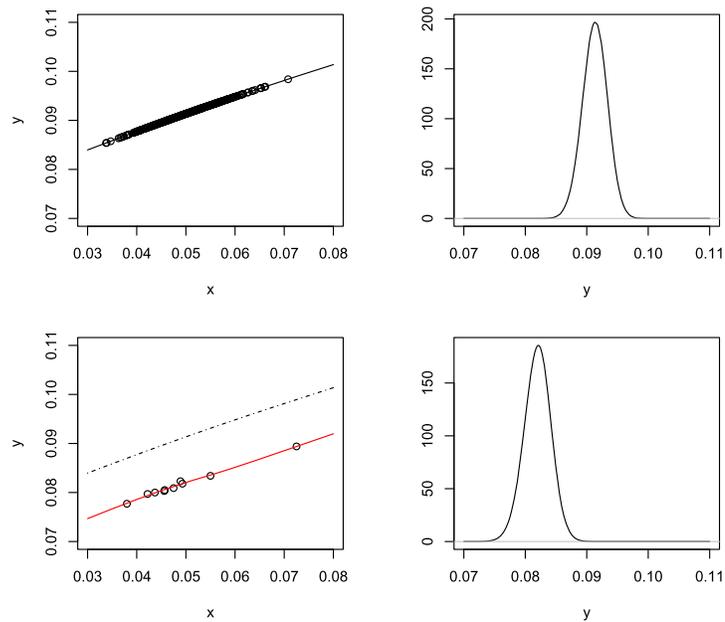


FIG 6. Data from $L_n = 500$ computer experiments together with a corresponding surrogate model (upper left panel), the corresponding density estimate (upper right panel), the surrogate model based only on the data from the computer experiments (dashed-dotted) together with $n = 10$ experimental data points and the corresponding surrogate model proposed in this paper (solid line) (lower left panel) and the corresponding density estimate proposed in this paper (lower right panel).

tional data points clearly leads in this example to a different density estimate than the estimate based only on the model data in the upper right panel of Figure 6.

5. Proofs

5.1. Auxiliary results

In this subsection we present various auxiliary results on smoothing spline estimates, which we use in the next subsection in order to derive a new error bound on smoothing splines applied to weighted data with additional measurement errors in the dependent variable, cf., Theorem 2 below. This result will be used to proof Theorem 1.

5.1.1. A deterministic lemma

Lemma 1. *Let $d, N \in \mathbb{N}$, $t > 0$, $w_1, \dots, w_N \in \mathbb{R}_+$, $x_1, \dots, x_N \in \mathbb{R}^d$, $\beta_N \geq L > 0$, $z_1, \dots, z_N \in \mathbb{R}$ and $\bar{z}_1, \dots, \bar{z}_N \in [-L, L]$. Let $m : \mathbb{R}^d \rightarrow \mathbb{R}$ be a function. Let \mathcal{F}_N be a set of functions $f : \mathbb{R}^d \rightarrow \mathbb{R}$ and for $f \in \mathcal{F}_N$ let*

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

be a penalty term. Define

$$\tilde{m}_N = \arg \min_{f \in \mathcal{F}_N} \left(\sum_{i=1}^N w_i \cdot |f(x_i) - \bar{z}_i|^2 + pen^2(f) \right)$$

(where we tacitly assume that the above minimum exists), and

$$\hat{m}_N(x) = T_{\beta_N}(\tilde{m}_N(x)) \quad (x \in \mathbb{R}^d)$$

and let $\hat{m}_N^* \in \mathcal{F}_N$ be arbitrary. Then

$$\begin{aligned} & \sum_{i=1}^N w_i \cdot |\hat{m}_N(x_i) - m(x_i)|^2 + pen^2(\tilde{m}_N) \\ & \geq 3 \left(\sum_{i=1}^N w_i \cdot |\hat{m}_N^*(x_i) - m(x_i)|^2 + pen^2(\hat{m}_N^*) \right) + 128 \sum_{i=1}^N w_i \cdot |z_i - \bar{z}_i|^2 + t \end{aligned} \tag{5.1}$$

implies

$$\begin{aligned} & \sum_{i=1}^N w_i \cdot (\hat{m}_N(x_i) - \hat{m}_N^*(x_i)) \cdot (z_i - m(x_i)) \\ & \geq \frac{1}{24} \left(\sum_{i=1}^N w_i \cdot |\hat{m}_N(x_i) - \hat{m}_N^*(x_i)|^2 + pen^2(\tilde{m}_N) \right) + \frac{t}{6}. \end{aligned} \tag{5.2}$$

Proof. Using

$$\sum_{i=1}^N w_i \cdot |\hat{m}_N(x_i) - \bar{z}_i|^2 \leq \sum_{i=1}^N w_i \cdot |\tilde{m}_N(x_i) - \bar{z}_i|^2$$

the assertion follows as in the proof of Lemma 5 in Furer and Kohler (2015). A complete proof is included in the Appendix. \square

5.1.2. A bound on a covering number

Definition 1. Let $l \in \mathbb{N}$ and let \mathcal{F} be a class of functions $f : \mathbb{R}^l \rightarrow \mathbb{R}$. The covering number $\mathcal{N}_2(\epsilon, \mathcal{F}, x_1^n)$ is defined for any $\epsilon > 0$ and $x_1^n = (x_1, \dots, x_n) \in (\mathbb{R}^l)^n$ as the smallest integer k such that there exist functions $g_1, \dots, g_k : \mathbb{R}^l \rightarrow \mathbb{R}$ with

$$\min_{1 \leq i \leq k} \left(\frac{1}{n} \sum_{j=1}^n |f(x_j) - g_i(x_j)|^2 \right)^{1/2} \leq \epsilon$$

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

Lemma 2. Let $L, A, c > 0$ and set

$$\mathcal{F} = \{T_L f : f \in W^k(\mathbb{R}^d) \text{ and } J_k^2(f) \leq c\}.$$

Then there exist constants $c_{17}, c_{18}, c_{19} \in \mathbb{R}_+$ depending only on A, k and d such that for any $\epsilon > 0$ and all $x_1, \dots, x_n \in [-A, A]^d$

$$\log \mathcal{N}_2(\epsilon, \mathcal{F}, x_1^n) \leq \left(c_{17} \cdot \left(\frac{\sqrt{c}}{\epsilon} \right)^{d/k} + c_{18} \right) \cdot \log \left(c_{19} \cdot \frac{L^2 n}{\epsilon^2} \right). \quad (5.3)$$

Proof. See Lemma 20.6 and Problem 20.9 in Györfi et al. (2002), or Lemma 3 in Kohler, Krzyżak and Schäfer (2002). \square

5.1.3. A bound on the error for smoothing spline estimates for fixed design regression

Let $L \geq 0$ and

$$Y_i = m(x_i) + W_i \quad (i = 1, \dots, n)$$

for some $x_1, \dots, x_n \in \mathbb{R}^d$, $m : \mathbb{R}^d \rightarrow \mathbb{R}$ and some random variables W_1, \dots, W_n which are independent and have expectation zero. We assume that the W_i 's are sub-Gaussian in the sense that

$$\max_{i=1, \dots, n} K^2 \mathbf{E}\{e^{W_i^2/K^2} - 1\} \leq \sigma_0^2 \quad (5.4)$$

for some $K, \sigma_0 > 0$. Our goal is to estimate m from $(x_1, \bar{Y}_{1,n}), \dots, (x_n, \bar{Y}_{n,n})$, where $\bar{Y}_{1,n}, \dots, \bar{Y}_{n,n} \in [-L, L]$ are arbitrary (bounded) random variables with the property that the average squared measurement error

$$\frac{1}{n} \sum_{i=1}^n |Y_i - \bar{Y}_{i,n}|^2$$

is “small”. Let \mathcal{F}_n be a set of functions $f : \mathbb{R}^d \rightarrow \mathbb{R}$ and consider the least squares estimate with complexity penalty

$$\tilde{m}_n(\cdot) = \arg \min_{f \in \mathcal{F}_n} \left(\frac{1}{n} \sum_{i=1}^n |f(x_i) - \bar{Y}_{i,n}|^2 + \text{pen}_n^2(f) \right) \quad \text{and} \quad \hat{m}_n = T_{\beta_n} \tilde{m}_n, \quad (5.5)$$

where for $f \in \mathcal{F}_n$

$$\text{pen}_n^2(f) \geq 0$$

is a penalty term penalizing the complexity of f and where $\beta_n \geq L$. Set

$$\|f\|_n^2 = \frac{1}{n} \sum_{i=1}^n |f(x_i)|^2.$$

Lemma 3. Assume that the sub-Gaussian condition (5.4) holds. Let \mathcal{F}_n be a set of functions: $\mathbb{R}^d \rightarrow \mathbb{R}$, let

$$\tilde{m}_n(\cdot) = \tilde{m}_n(\cdot, (x_1, \bar{Y}_{1,n}), \dots, (x_n, \bar{Y}_{n,n})) \in \mathcal{F}_n$$

and set $\hat{m}_n = T_{\beta_n} \tilde{m}_n$. Then there exist constants $c_{20}, c_{21}, c_{22} > 0$ which depend only on σ_0 and K such that for any $\delta_n > c_{20}/n$ with

$$\begin{aligned} \sqrt{n} \cdot \delta \geq c_{21} \int_{\delta/(12\sigma_0)}^{\sqrt{48\delta}} & \left(\log \mathcal{N}_2 \left(u, \{T_{\beta_n} f - g : f \in \mathcal{F}_n, \right. \right. & (5.6) \\ & \left. \left. \frac{1}{n} \sum_{i=1}^n |T_{\beta_n} f(x_i) - g(x_i)|^2 + \text{pen}_n^2(f) \leq 48 \cdot \delta\}, x_1^n \right) \right)^{1/2} du \end{aligned}$$

for all $\delta \geq \delta_n/6$ and all $g \in \mathcal{F}_n$ we have for any $m_n^* \in \mathcal{F}_n$

$$\begin{aligned} \mathbf{P} \left\{ \|\hat{m}_n - \hat{m}_n^*\|_n^2 + \text{pen}_n^2(\tilde{m}_n) + 4 \cdot \delta_n \leq \frac{24}{n} \cdot \sum_{i=1}^n (\hat{m}_n(x_i) - \hat{m}_n^*(x_i)) \cdot W_i \right\} \\ \leq c_{22} \cdot \exp \left(-\frac{n \cdot \min\{\delta_n, \sigma_0^2\}}{c_{22}} \right). \end{aligned}$$

Proof. The result follows from the proof of Lemma 2 in Kohler and Krzyżak (2012). A detailed proof is included in the Appendix. \square

5.1.4. A bound on the deviation between the L_2 error and the empirical L_2 error for smoothing splines

Let $(X, Y), (X_1, Y_1), \dots$ be independent and identically distributed $\mathbb{R}^d \times \mathbb{R}$ valued random variables with $\mathbf{E}Y^2 < \infty$. Let $m(x) = \mathbf{E}\{Y|X = x\}$ be the corresponding regression function. Let $\bar{Y}_{1,n}, \dots, \bar{Y}_{n,n}$ be \mathbb{R} -valued random variables and define the estimate \tilde{m}_n by

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

where \mathcal{F}_n is a set of functions $f : \mathbb{R}^d \rightarrow \mathbb{R}$ and for $f \in \mathcal{F}_n$

$$\text{pen}_n^2(f) \geq 0$$

is a penalty term penalizing the complexity of f . Set

$$\hat{m}_n = T_{\beta_n} \tilde{m}_n$$

for some $\beta_n > 0$. Then the following result holds.

Lemma 4. *Let $\beta_n \geq L \geq 1$ and assume that the m is bounded in absolute value by L . Let \mathcal{F}_n be a set of functions $f : \mathbb{R}^d \rightarrow \mathbb{R}$, let*

$$\tilde{m}_n(\cdot) = \tilde{m}_n(\cdot, (X_1, \bar{Y}_{1,n}), \dots, (X_n, \bar{Y}_{n,n})) \in \mathcal{F}_n$$

and set $\hat{m}_n = T_{\beta_n} \tilde{m}_n$. Then there exist constants $c_{23}, c_{24}, c_{25}, c_{26} > 0$ such that for any $\delta_n > 0$ which satisfies

$$\delta_n > c_{23} \cdot \frac{\beta_n^2}{n}$$

and

$$c_{24} \frac{\sqrt{n}\delta}{\beta_n^2} \geq \int_{c_{25}\delta/\beta_n^2}^{\sqrt{\delta}} \left(\log \mathcal{N}_2 \left(u, \{(T_{\beta_n} f - m)^2 : f \in \mathcal{F}_n, \frac{1}{n} \sum_{i=1}^n |T_{\beta_n} f(x_i) - m(x_i)|^2 \leq \frac{\delta}{\beta_n^2}, \text{pen}_n^2(f) \leq \delta\}, x_1^n \right) \right)^{1/2} du$$

for all $\delta \geq \delta_n$ and all $x_1, \dots, x_n \in \mathbb{R}^d$, we have for $n \in \mathbb{N}$

$$\begin{aligned} & \mathbf{P} \left\{ \int |\hat{m}_n(x) - m(x)|^2 \mathbf{P}_X(dx) > \delta_n + 3 \cdot \text{pen}_n^2(\tilde{m}_n) \right. \\ & \qquad \qquad \qquad \left. + 3 \frac{1}{n} \sum_{i=1}^n |\hat{m}_n(X_i) - m(X_i)|^2 \right\} \\ & \leq c_{26} \cdot \exp \left(- \frac{n \cdot \delta_n}{c_{26} \beta_n^2} \right). \end{aligned}$$

Proof. The result follows from the bound on $P_{1,n}$ presented in the proof of Lemma 3 in Kohler and Krzyżak (2012). A detailed proof is included in the Appendix. \square

5.2. A general result on penalized least squares estimates

Theorem 2. *Let $d, k, n, L_n \in \mathbb{N}$, $w^{(n)} \in [0, 1]$ with $n \geq 2$ and $1 \leq \beta \leq n + L_n$. Let $(X, Y), (X_1, Y_1), \dots$ be independent and identically distributed $\mathbb{R}^d \times \mathbb{R}$ -valued random variables with $\mathbf{E}\{Y^2\} < \infty$ and with $\text{supp}(X)$ bounded. Set $m(x) = \mathbf{E}\{Y|X = x\}$. Let $\bar{Y}_{1,n}, \dots, \bar{Y}_{n+L_n,n}$ be arbitrary \mathbb{R} -valued random variables satisfying*

$$\max_{i=1, \dots, n+L_n} \mathbf{E} \{ |\bar{Y}_{i,n}|^3 \} \leq c_{27} < \infty. \tag{5.7}$$

Set

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

and

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

Assume $2 \cdot k > d$ and define the estimate \hat{m}_n by

$$\tilde{m}_n(\cdot) = \arg \min_{f \in W^k(\mathbb{R}^d)} \left(\sum_{i=1}^{n+L_n} w_i \cdot |f(X_i) - \bar{Y}_{i,n}|^2 + \lambda_n \cdot J_k^2(f) \right)$$

and

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

Assume

$$K^2 \cdot \left(\mathbf{E} \left\{ \exp \left(\frac{(Y - m(X))^2}{K^2} \right) \middle| X \right\} - 1 \right) \leq \sigma_0^2 \quad \text{a.s.} \quad (5.8)$$

for some $K, \sigma_0 > 0$,

$$|m(x)| \leq \beta \quad (x \in \mathbb{R}^d) \quad (5.9)$$

and

$$J_k^2(m) < \infty. \quad (5.10)$$

Choose $\lambda_n \in \mathbb{R}_+$ such that

$$\frac{\log n}{n} \leq \lambda_n \leq \left(\frac{1}{\log L_n} \right)^{\frac{2k}{d}}. \quad (5.11)$$

Assume furthermore

$$n \leq L_n \leq n^l \quad (5.12)$$

for some $l \in \mathbb{N}$. Then there exist constants $c_{28}, c_{29}, c_{30}, c_{31} \in \mathbb{R}_+$ such that

$$\begin{aligned} & \mathbf{E} \int |\hat{m}_n(x) - m(x)|^2 \mathbf{P}_X(dx) \\ & \leq c_{28} \cdot \lambda_n \cdot J_k^2(m) + c_{29} \cdot w^{(n)} \cdot \left(\frac{\log n}{n \cdot \lambda_n^{d/2k}} + \mathbf{E} \left\{ \frac{1}{n} \cdot \sum_{i=1}^n |\bar{Y}_{i,n} - Y_i|^2 \right\} \right) \\ & \quad + c_{30} \cdot (1 - w^{(n)}) \cdot \left(\frac{\log L_n}{L_n \cdot \lambda_n^{d/2k}} + \mathbf{E} \left\{ \frac{1}{L_n} \cdot \sum_{i=n+1}^{n+L_n} |\bar{Y}_{i,n} - Y_i|^2 \right\} \right) + \frac{c_{31}}{n}. \end{aligned}$$

Proof. Set $\beta_n = n + L_n$.

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

$$\bar{Y}_{i,n} \in [-\beta_n, \beta_n] \quad \text{for all } i = 1, \dots, n + L_n. \quad (5.13)$$

To do this, we let

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

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

$$\begin{aligned} \mathbf{P}(A_n^c) &\leq (n + L_n) \cdot \max_{i=1, \dots, n+L_n} \mathbf{P}\{|\bar{Y}_{i,n}| > \beta_n\} \\ &\leq (n + L_n) \cdot \frac{\max_{i=1, \dots, n+L_n} \mathbf{E}\{|\bar{Y}_{i,n}|^3\}}{\beta_n^3} \\ &\leq \frac{c_{27}}{n}. \end{aligned}$$

On the event A_n the estimate \hat{m}_n coincides with the estimate $\hat{m}_n^{(trunc)}$ defined by

$$\tilde{m}_n^{(trunc)}(\cdot) = \arg \min_{f \in W^k(\mathbb{R}^d)} \left(\sum_{i=1}^{n+L_n} w_i \cdot |f(X_i) - T_{\beta_n} \bar{Y}_{i,n}|^2 + \lambda_n \cdot J_k^2(f) \right)$$

and

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

From this we can conclude that

$$\begin{aligned} &\mathbf{E} \int |\hat{m}_n(x) - m(x)|^2 \mathbf{P}_X(dx) \\ &\leq \mathbf{E} \left\{ \int |\hat{m}_n(x) - m(x)|^2 \mathbf{P}_X(dx) \cdot I_{A_n} \right\} + 4 \cdot \beta^2 \cdot \mathbf{P}(A_n^c) \\ &= \mathbf{E} \left\{ \int |\hat{m}_n^{(trunc)}(x) - m(x)|^2 \mathbf{P}_X(dx) \cdot I_{A_n} \right\} + 4 \cdot \beta^2 \cdot \mathbf{P}(A_n^c) \\ &\leq \mathbf{E} \int |\hat{m}_n^{(trunc)}(x) - m(x)|^2 \mathbf{P}_X(dx) + 4 \cdot \beta^2 \cdot \frac{c_{27}}{n}, \end{aligned}$$

which completes the first step of the proof.

So from now on we assume that (5.13) holds. Set

$$\begin{aligned} \delta_n &= c_{32} \cdot \frac{\log n}{n \cdot \lambda_n^{d/(2k)}}, \quad \delta_{L_n} = c_{32} \cdot \frac{\log L_n}{L_n \cdot \lambda_n^{d/(2k)}}, \\ \gamma_n &= w^{(n)} \cdot \delta_n + (1 - w^{(n)}) \cdot \delta_{L_n} \end{aligned}$$

and

$$\begin{aligned} T_n &= \int |\hat{m}_n(x) - m(x)|^2 \mathbf{P}_X(dx) \\ &\quad - \left(9 \cdot \lambda_n \cdot J_k^2(m) + 384 \cdot \sum_{i=1}^{n+L_n} w_i \cdot |Y_i - \bar{Y}_{i,n}|^2 \right). \end{aligned}$$

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

$$\int_{36 \cdot \gamma_n}^{\infty} \mathbf{P}\{T_n > t\} dt \leq \frac{c_{33}}{n}.$$

To do this, we observe

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

$$\begin{aligned} &\leq \mathbf{E} \left\{ \int |\hat{m}_n(x) - m(x)|^2 \mathbf{P}_X(dx) \right. \\ &\quad \left. - \left(9 \cdot \lambda_n \cdot J_k^2(m) + 384 \cdot \sum_{i=1}^{n+L_n} w_i \cdot |Y_i - \bar{Y}_{i,n}|^2 \right) \right\} \\ &\quad + 9 \cdot \lambda_n \cdot J_k^2(m) + 384 \cdot \mathbf{E} \left\{ \sum_{i=1}^{n+L_n} w_i \cdot |Y_i - \bar{Y}_{i,n}|^2 \right\} \\ &\leq 36 \cdot \gamma_n + \int_{36 \cdot \gamma_n}^{\infty} \mathbf{P}\{T_n > t\} dt + 9 \cdot \lambda_n \cdot J_k^2(m) \\ &\quad + 384 \cdot \mathbf{E} \left\{ \sum_{i=1}^{n+L_n} w_i \cdot |Y_i - \bar{Y}_{i,n}|^2 \right\}. \end{aligned}$$

The definition of γ_n and of the weights implies the assertion of step 2.

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

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

where

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

and

$$\begin{aligned} P_{2,n}(t) &= \mathbf{P} \left\{ 3 \cdot \sum_{i=1}^{n+L_n} w_i \cdot |\hat{m}_n(X_i) - m(X_i)|^2 + 3 \cdot \lambda_n \cdot J_k^2(\tilde{m}_n) \right. \\ &\quad \left. > \frac{t}{2} + 9 \cdot \left(\sum_{i=1}^{n+L_n} w_i \cdot |m(X_i) - m(X_i)|^2 + \lambda_n J_k^2(m) \right) \right. \\ &\quad \left. + 384 \cdot \sum_{i=1}^{n+L_n} w_i \cdot |Y_i - \bar{Y}_{i,n}|^2 \right\}. \end{aligned}$$

Using

$$\begin{aligned} T_n &= \left(\int |\hat{m}_n(x) - m(x)|^2 \mathbf{P}_X(dx) - 3 \cdot \lambda_n \cdot J_k^2(\tilde{m}_n) \right. \\ &\quad \left. - 3 \cdot \sum_{i=1}^{n+L_n} w_i \cdot |\hat{m}_n(X_i) - m(X_i)|^2 \right) \\ &\quad + \left(3 \cdot \sum_{i=1}^{n+L_n} w_i \cdot |\hat{m}_n(X_i) - m(X_i)|^2 + 3 \cdot \lambda_n \cdot J_k^2(\tilde{m}_n) \right) \end{aligned}$$

$$\begin{aligned}
 & - \left(9 \cdot \left(\sum_{i=1}^{n+L_n} w_i \cdot |m(X_i) - m(X_i)|^2 + \lambda_n J_k^2(m) \right) \right. \\
 & \qquad \qquad \qquad \left. + 384 \cdot \sum_{i=1}^{n+L_n} w_i \cdot |Y_i - \bar{Y}_{i,n}|^2 \right) \\
 & =: T_{1,n} + T_{2,n}
 \end{aligned}$$

this immediately follows from

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

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

$$\int_{36 \cdot \gamma_n}^{\infty} P_{1,n}(t) dt.$$

Let $t \geq 36 \cdot \gamma_n$. The definition of the weights together with

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

implies that we have

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

We show next that Lemma 4 is applicable to the two different probabilities, where both times β_n is replaced by β and where we use sample sizes n and L_n , resp. Since $t \geq 36 \cdot \gamma_n \geq 2 \cdot \gamma_n$ implies

$$\frac{\delta_n}{w^{(n)} \cdot \delta_n + (1 - w^{(n)}) \cdot \delta_{L_n}} \cdot \frac{t}{2} \geq \delta_n$$

and

$$\frac{\delta_{L_n}}{w^{(n)} \cdot \delta_n + (1 - w^{(n)}) \cdot \delta_{L_n}} \cdot \frac{t}{2} \geq \delta_{L_n},$$

in order to show that Lemma 4 is applicable to the first probability, it suffices to show

$$\delta_n > c_{34} \cdot \frac{\beta^2}{n}$$

and

$$c_{35} \frac{\sqrt{n} \delta}{\beta^2} \geq \int_{c_{36} \delta / \beta^2}^{\sqrt{\delta}} \left(\log \mathcal{N}_2 \left(u, \left\{ (T_\beta f - m)^2 : f \in W_k(\mathbb{R}^d), J_k^2(f) \leq \frac{\delta}{\lambda_n} \right\}, x_1^n \right) \right)^{1/2} du$$

for all $\delta \geq \delta_n$ and all $x_1, \dots, x_n \in \mathbb{R}^d$. Using $|a^2 - b^2|^2 \leq (|a| + |b|)^2 \cdot |a - b|^2$ ($a, b \in \mathbb{R}$) (which we apply with $a = T_\beta f(x_i) - m(x_i)$ and $b = g(x_i)$, where g is approximating $T_\beta f - m$), we see that we have

$$\begin{aligned} & \mathcal{N}_2 \left(u, \left\{ (T_\beta f - m)^2 : f \in W_k(\mathbb{R}^d), J_k^2(f) \leq \frac{\delta}{\lambda_n} \right\}, x_1^n \right) \\ & \leq \mathcal{N}_2 \left(\frac{u}{16\beta^2}, \left\{ T_{\beta_n} f - m : f \in W_k(\mathbb{R}^d), J_k^2(f) \leq \frac{\delta}{\lambda_n} \right\}, x_1^n \right). \end{aligned}$$

Using this together with Lemma 2 we see that Lemma 4 is applicable to the first probability, if $\delta_n > c_{34} \cdot \frac{\beta^2}{n}$ and the following inequality hold:

$$\frac{\sqrt{n} \cdot \delta}{\beta^2} \geq c_{37} \cdot \int_0^{\sqrt{\delta}} \left(\left(\left(\frac{\sqrt{\delta/\lambda_n}}{u/16\beta^2} \right)^{d/k} + 1 \right) \cdot \log(c_{38} \cdot \beta^2 \cdot n^3) \right)^{1/2} du.$$

The last condition is implied by

$$\sqrt{n} \cdot \delta \geq c_{39} \cdot \sqrt{\log(c_{38} \cdot \beta^2 \cdot n^3)} \cdot \left(\left(\frac{\delta}{\lambda_n} \right)^{d/(4k)} \cdot \delta^{\frac{1}{2} - \frac{d}{4k}} + \sqrt{\delta} \right),$$

which in turn follows from

$$\delta \geq c_{40} \cdot \frac{\log n}{n \cdot \lambda_n^{d/(2k)}} \quad \text{and} \quad \delta \geq c_{40} \cdot \frac{\log n}{n}.$$

In case that

$$\lambda_n \leq 1$$

the last two conditions hold for all $\delta \geq \delta_n$, provided c_{32} is chosen large enough.

In the same way one can show that Lemma 4 is also applicable to the second probability above.

By applying Lemma 4 to the two different probabilities we get

$$\begin{aligned} P_{1,n}(t) \leq & c_{41} \cdot \exp\left(-c_{42} \cdot n \cdot \frac{\delta_n}{w^{(n)} \cdot \delta_n + (1 - w^{(n)}) \cdot \delta_{L_n}} \cdot \frac{t}{2}\right) \\ & + c_{41} \cdot \exp\left(-c_{42} \cdot L_n \cdot \frac{\delta_{L_n}}{w^{(n)} \cdot \delta_n + (1 - w^{(n)}) \cdot \delta_{L_n}} \cdot \frac{t}{2}\right), \end{aligned}$$

which implies

$$\begin{aligned} & \int_{36 \cdot \gamma_n}^{\infty} P_{1,n}(t) dt \\ & \leq \frac{c_{43}}{n} \cdot \frac{w^{(n)} \cdot \delta_n + (1 - w^{(n)}) \cdot \delta_{L_n}}{\delta_n} \cdot \exp(-c_{44} \cdot n \cdot \delta_n) \\ & \quad + \frac{c_{43}}{L_n} \cdot \frac{w^{(n)} \cdot \delta_n + (1 - w^{(n)}) \cdot \delta_{L_n}}{\delta_{L_n}} \cdot \exp(-c_{44} \cdot L_n \cdot \delta_{L_n}) \leq \frac{c_{45}}{n}. \end{aligned}$$

Here the last inequality follows from the assumptions (5.11) and (5.12), from which we can conclude

$$\begin{aligned} n \cdot \delta_n & \geq c_{32} \cdot \log^2(n), & L_n \cdot \delta_{L_n} & \geq c_{32} \cdot \log^2(n), \\ \frac{w^{(n)} \cdot \delta_n + (1 - w^{(n)}) \cdot \delta_{L_n}}{n \cdot \delta_n} & \leq n^s & \text{and} & \quad \frac{w^{(n)} \cdot \delta_n + (1 - w^{(n)}) \cdot \delta_{L_n}}{L_n \cdot \delta_{L_n}} \leq n^s \end{aligned}$$

for some $s > 0$.

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

$$\int_{36 \cdot \gamma_n}^{\infty} P_{2,n}(t) dt.$$

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

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

which will allow us to increase the truncation level in the following. To do this, we observe that the above inequality together with (5.13) and Lemma 1 (applied with $\hat{m}_n^* = m$) implies

$$\begin{aligned} & P_{2,n}(t) \\ & \leq \mathbf{P} \left\{ 3 \cdot \sum_{i=1}^{n+L_n} w_i \cdot |T_{\beta_n}(\tilde{m}_n)(X_i) - m(X_i)|^2 + 3 \cdot \lambda_n \cdot J_k^2(\tilde{m}_n) \right\} \end{aligned}$$

$$\begin{aligned}
 &> \frac{t}{2} + 9 \cdot \left(\sum_{i=1}^{n+L_n} w_i \cdot |m(X_i) - \tilde{m}(X_i)|^2 + \lambda_n J_k^2(\tilde{m}) \right) \\
 &\quad + 384 \cdot \sum_{i=1}^{n+L_n} w_i \cdot |Y_i - \bar{Y}_i|^2 \Big\} \\
 &\leq \mathbf{P} \left\{ \sum_{i=1}^{n+L_n} w_i \cdot (T_{\beta_n}(\tilde{m}_n)(X_i) - m(X_i)) \cdot (Y_i - m(X_i)) \geq \right. \\
 &\quad \left. \frac{1}{24} \left(\sum_{i=1}^{n+L_n} w_i \cdot |T_{\beta_n}(\tilde{m}_n)(X_i) - m(X_i)|^2 + \lambda_n \cdot J_k^2(\tilde{m}_n) \right) + \frac{t}{36} \right\}.
 \end{aligned}$$

Proceeding as in the proof of step 4 we can conclude from the definition of the weights that the last probability is bounded by

$$\begin{aligned}
 &\mathbf{P} \left\{ \frac{1}{n} \cdot \sum_{i=1}^n (T_{\beta_n}(\tilde{m}_n)(X_i) - m(X_i)) \cdot (Y_i - m(X_i)) \geq \right. \\
 &\quad \left. \frac{1}{n} \cdot \frac{1}{24} \left(\sum_{i=1}^n |T_{\beta_n}(\tilde{m}_n)(X_i) - m(X_i)|^2 + \lambda_n \cdot J_k^2(\tilde{m}_n) \right) \right. \\
 &\quad \left. + \frac{\delta_n}{w^{(n)} \cdot \delta_n + (1 - w^{(n)}) \cdot \delta_{L_n}} \cdot \frac{t}{36} \right\} \\
 &+ \mathbf{P} \left\{ \frac{1}{L_n} \cdot \sum_{i=n+1}^{n+L_n} (T_{\beta_n}(\tilde{m}_n)(X_i) - m(X_i)) \cdot (Y_i - m(X_i)) \geq \right. \\
 &\quad \left. \frac{1}{L_n} \cdot \frac{1}{24} \left(\sum_{i=n+1}^{n+L_n} |T_{\beta_n}(\tilde{m}_n)(X_i) - m(X_i)|^2 + \lambda_n \cdot J_k^2(\tilde{m}_n) \right) \right. \\
 &\quad \left. + \frac{\delta_{L_n}}{w^{(n)} \cdot \delta_n + (1 - w^{(n)}) \cdot \delta_{L_n}} \cdot \frac{t}{36} \right\},
 \end{aligned}$$

and that Lemma 3 can be applied to both probabilities. From this we can conclude that the above probabilities are bounded by

$$\begin{aligned}
 &c_{46} \cdot \exp \left(-c_{47} \cdot n \cdot \frac{\delta_n}{w^{(n)} \cdot \delta_n + (1 - w^{(n)}) \cdot \delta_{L_n}} \cdot \frac{t}{36} \right) \\
 &+ c_{46} \cdot \exp \left(-c_{47} \cdot L_n \cdot \frac{\delta_{L_n}}{w^{(n)} \cdot \delta_n + (1 - w^{(n)}) \cdot \delta_{L_n}} \cdot \frac{t}{36} \right),
 \end{aligned}$$

which implies as above

$$\begin{aligned}
 &\int_{36 \cdot \gamma_n}^{\infty} P_{2,n}(t) dt \\
 &\leq \frac{c_{48}}{n} \cdot \frac{w^{(n)} \cdot \delta_n + (1 - w^{(n)}) \cdot \delta_{L_n}}{\delta_n} \cdot \exp(-c_{49} \cdot n \cdot \delta_n)
 \end{aligned}$$

$$\begin{aligned}
& + \frac{c_{48}}{L_n} \cdot \frac{w^{(n)} \cdot \delta_n + (1 - w^{(n)}) \cdot \delta_{L_n}}{\delta_{L_n}} \cdot \exp(-c_{49} \cdot L_n \cdot \delta_{L_n}) \\
& \leq \frac{c_{50}}{n}.
\end{aligned}$$

Summarizing the above results we get the assertion. \square

5.3. Proof of Theorem 1

Using the definition of \hat{m}_n , $(a + b + c)^2 \leq 3a^2 + 3b^2 + 3c^2$ ($a, b, c \in \mathbb{R}$), (3.1), the definition of \hat{m}_n and (3.2) we get

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

Hence in order to prove the assertion it suffices to show

$$\mathbf{E} \int |\hat{m}_{L_n}(x) - m(x)|^2 \mathbf{P}_X(dx) \leq c_{51} \cdot \lambda_{L_n} \cdot J_k^2(m) + c_{52} \cdot \frac{\log L_n}{L_n \cdot \lambda_{L_n}^{d/(2k)}} + \frac{c_{53}}{L_n} \quad (5.14)$$

and

$$\begin{aligned}
& \mathbf{E} \int |\hat{m}_n^\varepsilon(x) - (m^* - m)(x)|^2 \mathbf{P}_X(dx) \\
& \leq c_{54} \cdot \alpha_n^2 \cdot \lambda_n + c_{55} \cdot w^{(n)} \cdot \alpha_n^2 \cdot \frac{\log n}{n \cdot \lambda_n^{d/(2k)}} + c_{56} \cdot \left(\frac{\log L_n}{L_n} \right)^{\frac{2k}{2k+d}} \\
& \quad + c_{57} \cdot (1 - w^{(n)}) \cdot \alpha_n^2 \cdot \left(1 + \frac{\log N_n}{N_n \cdot \lambda_n^{d/(2k)}} \right) + \frac{c_{58} \cdot \alpha_n^2}{\min\{n, N_n\}}. \quad (5.15)
\end{aligned}$$

Inequality (5.14) follows from Theorem 2 applied with $(X, Y) = (X, m(X))$, $n = L_n$, $w^{(n)} = 1$ and $\bar{Y}_{i, L_n + \bar{L}_n} = Y_i = m(X_{n+i})$ ($i = 1, \dots, L_n$) and suitably chosen $\bar{Y}_{L_n+1, L_n + \bar{L}_n}, \dots, \bar{Y}_{L_n + \bar{L}_n, L_n + \bar{L}_n}$.

In order to prove (5.15) we first observe that

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

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

$$\int |\hat{m}_n^\varepsilon(x) - (m^* - m)(x)|^2 \mathbf{P}_X(dx)$$

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

By definition of $\hat{m}_n^{\hat{\epsilon}}$ we have

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

where

$$\begin{aligned} \frac{1}{\alpha_n} \cdot \tilde{m}_n^{\hat{\epsilon}}(\cdot) = \arg \min_{f \in W^k(\mathbb{R}^d)} & \left(\frac{w^{(n)}}{n} \sum_{i=1}^n \left(\frac{1}{\alpha_n} \cdot \hat{\epsilon}_i - f(X_i) \right)^2 \right. \\ & \left. + \frac{1 - w^{(n)}}{N_n} \sum_{i=1}^{N_n} (0 - f(X_{n+L_n+i}))^2 + \lambda_n \cdot J_k^2(f) \right). \end{aligned}$$

The assumptions in Theorem 1 together with (5.14) imply that we have

$$\sup_{x \in \mathbb{R}^d} \left| \frac{1}{\alpha_n} \cdot (m^* - m)(x) \right| \leq 1 \leq c_1$$

and

$$\begin{aligned} & \max_{i=1, \dots, n} \mathbf{E} \left\{ \left| \frac{Y_i - \hat{m}_{L_n}(X_i)}{\alpha_n} \right|^3 \right\} \\ & \leq \frac{27}{\alpha_n^3} \cdot (\mathbf{E} \{|Y - m^*(X)|^3\}) \\ & \quad + \mathbf{E} \{|m^*(X) - m(X)|^3\} + \mathbf{E} \{|m(X) - \hat{m}_{L_n}(X)|^3\} \\ & \leq \frac{(\alpha_n^*)^3}{\alpha_n^3} + 1 + \frac{c_{59} \cdot \left(\frac{\log L_n}{L_n} \right)^{\frac{2k}{2k+d}}}{\alpha_n^3} \leq 2 + c_{59}. \end{aligned}$$

We consider

$$\begin{aligned} \frac{1}{\alpha_n} \cdot \hat{\epsilon}_i &= \frac{1}{\alpha_n} \cdot (Y_i - \hat{m}_{L_n}(X_i)) \\ &= \frac{1}{\alpha_n} \cdot (Y_i - m(X_i)) + \frac{1}{\alpha_n} \cdot (m(X_i) - \hat{m}_{L_n}(X_i)) \end{aligned}$$

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

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

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

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

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

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

($i = 1, \dots, N_n$).

From inequality (5.14) we can conclude

$$\begin{aligned} & \mathbf{E} \left\{ \frac{1}{n} \sum_{i=1}^n \left| \frac{1}{\alpha_n} \cdot (m(X_i) - \hat{m}_{L_n}(X_i)) \right|^2 \right\} \\ & \leq \frac{1}{\alpha_n^2} \cdot \left(c_{51} \cdot \lambda_{L_n} \cdot J_k^2(m) + c_{52} \cdot \frac{\log L_n}{L_n \cdot \lambda_{L_n}^{d/(2k)}} + \frac{c_{53}}{L_n} \right), \end{aligned}$$

and the assumptions of Theorem 1 imply

$$\begin{aligned} & \mathbf{E} \left\{ \frac{1}{N_n} \sum_{i=1}^{N_n} \left| \frac{1}{\alpha_n} \cdot (Y_{n+L_n+i} - m(X_{n+L_n+i})) \right|^2 \right\} \\ & \leq 2 \cdot \mathbf{E} \left\{ \frac{1}{N_n} \sum_{i=1}^{N_n} \left| \frac{1}{\alpha_n} \cdot (Y_{n+L_n+i} - m^*(X_{n+L_n+i})) \right|^2 \right\} \\ & \quad + 2 \cdot \mathbf{E} \left\{ \frac{1}{N_n} \sum_{i=1}^{N_n} \left| \frac{1}{\alpha_n} \cdot (m^*(X_{n+L_n+i}) - m(X_{n+L_n+i})) \right|^2 \right\} \\ & \leq 2 \cdot \frac{(\alpha_n^*)^2}{\alpha_n^2} + 2 \leq 4. \end{aligned}$$

Application of Theorem 2 yields

$$\begin{aligned} & \mathbf{E} \int \left| \frac{1}{\alpha_n} \cdot \hat{m}_n^\varepsilon(x) - \frac{1}{\alpha_n} \cdot (m^* - m)(x) \right|^2 \mathbf{P}_X(dx) \\ & \leq c_{28} \cdot \lambda_n \cdot J_k^2 \left(\frac{1}{\alpha_n} \cdot (m^* - m) \right) \\ & \quad + c_{29} \cdot w^{(n)} \cdot \left(\frac{\log n}{n \cdot \lambda_n^{d/(2k)}} + \frac{1}{\alpha_n^2} \cdot (c_{51} \cdot \lambda_{L_n} \cdot J_k^2(m) \right. \\ & \quad \left. + c_{52} \cdot \frac{\log L_n}{L_n \cdot \lambda_{L_n}^{d/(2k)}} + \frac{c_{53}}{L_n} \right) \\ & \quad + c_{30} \cdot (1 - w^{(n)}) \cdot \left(\frac{\log N_n}{N_n \cdot \lambda_n^{d/(2k)}} + 4 \right) + \frac{c_{31}}{\min\{n, N_n\}}, \end{aligned}$$

which implies (5.15). □

Acknowledgment

The first author would like to thank the German Research Foundation (DFG) for funding this project (Projektnummer 57157498 – SFB 805) The second author

would like to acknowledge the support from the Natural Sciences and Engineering Research Council of Canada under Grant RGPIN-2020-06793.

References

- 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. [MR2363956](#)
- Bichon, B., Eldred, M., Swiler, M., Mahadevan, S., and McFarland, J. (2008). Efficient global reliability analysis for nonlinear implicit performance functions. *American Institute of Aeronautics and Astronautics 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. [MR3375660](#)
- 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.
- Chernoff, H. (1952). A measure of asymptotic efficiency of tests of a hypothesis based on the sum of observations. *The Annals of Mathematical Statistics*, **23**, pp. 493–507.
- Coombs, Clyde H. (1964). *A Theory of Data*. John Wiley & Sons, New York.
- 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 efficient 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., 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. [MR3030761](#)
- Devroye, L., and Lugosi, G. (2001). *Combinatorial Methods in Density Estimation*. Springer-Verlag, New York. [MR1843146](#)
- 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. [MR3304357](#)
- Felber, T., Kohler, M., and Krzyżak, A. (2015b). Density estimation with small measurement errors. *IEEE Transactions on Information Theory*, **61**, pp. 3446–3456. [MR3352508](#)
- Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *The Annals of Statistics*, **29**, pp. 1189–1232. [MR1873328](#)

- Friedman, J. H. (2002). Stochastic gradient boosting. *Computational Statistics & Data Analysis*, **38**, pp. 367–378. [MR1884869](#)
- Furer, D., and Kohler, M. (2015). Smoothing spline regression estimation based on real and artificial data. *Metrika*, **78**, pp. 711–746. [MR3369340](#)
- 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. [MR3176554](#)
- Götz, B., Kersting, S., and Kohler, M. (2020). Estimation of an improved surrogate model in uncertainty quantification by neural networks. *Annals of the Institute of Statistical Mathematics*, to appear.
- Györfi, L., Kohler, M., Krzyżak, A., and Walk, H. (2002). *A Distribution-Free Theory of Nonparametric Regression*. Springer-Verlag, New York. [MR1920390](#)
- 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. [MR2756481](#)
- 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. [MR3176553](#)
- Hurtado, J. (2004). *Structural Reliability – Statistical Learning Perspectives*. Vol. 17 of Lecture Notes in Applied and Computational Mechanics. Springer. [MR2104067](#)
- Kaymaz, I. (2005). Application of Kriging method to structural reliability problems. *Structural Safety*, **27**, pp. 133–151.
- 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. [MR1858398](#)
- Kim, S.-H., and Na, S.-W. (1997). Response surface method using vector projected sampling points. *Structural Safety*, **19**, pp. 3–19.
- Kohler, M., Krzyżak, A., and Schäfer, D. (2002). Application of structural risk minimization to multivariate smoothing spline regression estimates. *Bernoulli*, **8**, pp. 475–489. [MR1914699](#)
- Kohler, M., and Krzyżak, A. (2012). Pricing of American options in discrete time using least squares estimates with complexity penalties. *Journal of Statistical Planning and Inference*, **142**, pp. 2289–2307. [MR2911845](#)
- Mallapur, S., and Platz, R. (2017). Quantification and Evaluation of Uncertainty in the Mathematical Modelling of a Suspension Strut using Bayesian Model Validation Approach. In: *Proceedings of the International Modal Analysis Conference IMAC-XXXV*, Garden Grove, California, USA, Paper 117, 30. Jan–2. Feb., 2017. [MR3674425](#)
- 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.
- Sacks, J., Welch, J. W., Mitchell, T. J., and Wynn, P. (1989). Design and analysis of computer experiments. *Statistical Science*, pp. 409–423. [MR1041765](#)
- Santner, T. J., Williams, B. J., and Notz, W. I. (2003). *The Design and*

Analysis of Computer Experiments, 2nd Edition. Springer-Verlag, New York. [MR3887662](#)

Tuo, R., and Wu, C. F. J. (2015). Efficient calibration for imperfect computer models. *The Annals of Statistics*, **43**, pp. 2331–2352. [MR3405596](#)

van de Geer, S. (2000). *Empirical Processes in M-estimation*. Cambridge University Press, Cambridge.

Wang, S., Chen, W., and Tsui, K. L. (2009). Bayesian validation of computer models. *Technometrics*, **51**, pp. 439–451. [MR2756479](#)

Wong, R. K. W., Storlie, C. B., and Lee, T. C. (2017). A frequentist approach to computer model calibration. *Journal of the Royal Statistical Society: Series B*, **79**, pp. 635–648. [MR3611763](#)

Appendix

Proof of Lemma 1. Since $|\bar{z}_i| \leq L \leq \beta_n$ ($i = 1, \dots, N$) and $w_i \geq 0$ ($i = 1, \dots, N$) we have

$$\begin{aligned} \sum_{i=1}^N w_i \cdot |\bar{z}_i - \hat{m}_N(x_i)|^2 &= \sum_{i=1}^N w_i \cdot |\bar{z}_i - T_{\beta_N}(\tilde{m}_N(x_i))|^2 \\ &\leq \sum_{i=1}^N w_i \cdot |\bar{z}_i - \tilde{m}_N(x_i)|^2. \end{aligned}$$

This together with the definition of the estimate implies

$$\sum_{i=1}^N w_i \cdot |\bar{z}_i - \hat{m}_N(x_i)|^2 + pen^2(\tilde{m}_N) \leq \sum_{i=1}^N w_i \cdot |\bar{z}_i - \hat{m}_N^*(x_i)|^2 + pen^2(\hat{m}_N^*),$$

hence

$$\begin{aligned} &\sum_{i=1}^N w_i \cdot |\bar{z}_i - m(x_i)|^2 + 2 \sum_{i=1}^N w_i \cdot (m(x_i) - \hat{m}_N(x_i)) \cdot (\bar{z}_i - m(x_i)) \\ &\quad + \sum_{i=1}^N w_i \cdot |m(x_i) - \hat{m}_N(x_i)|^2 + pen^2(\tilde{m}_N) \\ &\leq \sum_{i=1}^N w_i \cdot |\bar{z}_i - m(x_i)|^2 + 2 \sum_{i=1}^N w_i \cdot (m(x_i) - m_N^*(x_i)) \cdot (\bar{z}_i - m(x_i)) \\ &\quad + \sum_{i=1}^N w_i \cdot |m(x_i) - m_N^*(x_i)|^2 + pen^2(m_N^*), \end{aligned}$$

which implies

$$\sum_{i=1}^N w_i \cdot |m(x_i) - \hat{m}_N(x_i)|^2 + pen^2(\tilde{m}_N)$$

$$\begin{aligned}
& 1 - \sum_{i=1}^N w_i \cdot |m(x_i) - m_N^*(x_i)|^2 - \text{pen}^2(m_N^*) \\
& \leq 2 \sum_{i=1}^N w_i \cdot (\bar{z}_i - m(x_i)) \cdot (\hat{m}_N(x_i) - m_N^*(x_i)) \\
& = 2 \sum_{i=1}^N w_i \cdot (\bar{z}_i - z_i) \cdot (\hat{m}_N(x_i) - m_N^*(x_i)) \\
& \quad + 2 \sum_{i=1}^N w_i \cdot (z_i - m(x_i)) \cdot (\hat{m}_N(x_i) - m_N^*(x_i)) \\
& =: T_1 + T_2.
\end{aligned}$$

We show next that $T_1 \leq T_2$. Assume to the contrary that this is not true. Then

$$\begin{aligned}
& \sum_{i=1}^N w_i \cdot |m(x_i) - \hat{m}_N(x_i)|^2 + \text{pen}^2(\tilde{m}_N) \\
& \quad - \sum_{i=1}^N w_i \cdot |m(x_i) - m_N^*(x_i)|^2 - \text{pen}^2(m_N^*) \\
& < 4 \sum_{i=1}^N w_i \cdot (\bar{z}_i - z_i) \cdot (\hat{m}_N(x_i) - m_N^*(x_i)) \\
& \leq 4 \cdot \sqrt{\sum_{i=1}^N w_i \cdot (\bar{z}_i - z_i)^2} \cdot \sqrt{\sum_{i=1}^N w_i \cdot (\hat{m}_N(x_i) - m_N^*(x_i))^2} \\
& \leq 4 \cdot \sqrt{\sum_{i=1}^N w_i \cdot (\bar{z}_i - z_i)^2} \\
& \quad \cdot \left(2 \sum_{i=1}^N w_i \cdot |\hat{m}_N(x_i) - m(x_i)|^2 + 2\text{pen}^2(\tilde{m}_N) \right. \\
& \quad \left. + 2 \sum_{i=1}^N w_i \cdot |m_N^*(x_i) - m(x_i)|^2 + 2\text{pen}^2(m_N^*) \right)^{1/2}.
\end{aligned}$$

Using (5.1) we see that

$$\begin{aligned}
& \sum_{i=1}^N w_i \cdot |\hat{m}_N(x_i) - m(x_i)|^2 + \text{pen}^2(\tilde{m}_N) \\
& \quad - \sum_{i=1}^N w_i \cdot |m_N^*(x_i) - m(x_i)|^2 - \text{pen}^2(m_N^*) \\
& \geq \frac{1}{2} \cdot \left(\sum_{i=1}^N w_i \cdot |\hat{m}_N(x_i) - m(x_i)|^2 + \text{pen}^2(\tilde{m}_N) \right) \\
& \quad + \frac{1}{2} \cdot \left(3 \left(\sum_{i=1}^N w_i \cdot |m_N^*(x_i) - m(x_i)|^2 + \text{pen}^2(m_N^*) \right) \right. \\
& \quad \quad \left. + 128 \cdot \sum_{i=1}^N w_i \cdot |z_i - \bar{z}_i|^2 + t \right) \\
& \quad - \sum_{i=1}^N w_i \cdot |m_N^*(x_i) - m(x_i)|^2 - \text{pen}^2(m_N^*)
\end{aligned}$$

$$\geq \frac{1}{2} \cdot \left(\sum_{i=1}^N w_i \cdot |\hat{m}_N(x_i) - m(x_i)|^2 + pen^2(\tilde{m}_N) + \sum_{i=1}^N w_i \cdot |m_N^*(x_i) - m(x_i)|^2 + pen^2(m_N^*) \right),$$

which implies

$$\begin{aligned} & \frac{1}{2} \cdot \left(\sum_{i=1}^N w_i \cdot |\hat{m}_N(x_i) - m(x_i)|^2 + pen^2(\tilde{m}_N) + \sum_{i=1}^N w_i \cdot |m_N^*(x_i) - m(x_i)|^2 + pen^2(m_N^*) \right)^{1/2} \\ & < 4 \cdot \sqrt{2} \cdot \sqrt{\sum_{i=1}^N w_i \cdot |z_i - \bar{z}_i|^2} \end{aligned}$$

i.e.,

$$\begin{aligned} & \sum_{i=1}^N w_i \cdot |\hat{m}_N(x_i) - m(x_i)|^2 + pen^2(\tilde{m}_N) + \sum_{i=1}^N w_i \cdot |m_N^*(x_i) - m(x_i)|^2 + pen^2(m_N^*) \\ & < 128 \cdot \sum_{i=1}^N w_i \cdot |z_i - \bar{z}_i|^2. \end{aligned}$$

But this is a contradiction to (5.1), so we have indeed proved $T_1 \leq T_2$. As a consequence we can conclude from (5.1)

$$\begin{aligned} & 4 \sum_{i=1}^N w_i \cdot (\hat{m}_N(x_i) - \hat{m}_N^*(x_i)) \cdot (z_i - m(x_i)) \\ & \geq \sum_{i=1}^N w_i \cdot |\hat{m}_N(x_i) - m(x_i)|^2 + pen^2(\tilde{m}_N) - \sum_{i=1}^N w_i \cdot |m_N^*(x_i) - m(x_i)|^2 - pen^2(m_N^*) \\ & \geq \frac{1}{3} \left(\sum_{i=1}^N w_i \cdot |\hat{m}_N(x_i) - m(x_i)|^2 + pen^2(\tilde{m}_N) \right) + \frac{2}{3} \left(2 \sum_{i=1}^N w_i \cdot |m_N^*(x_i) - m(x_i)|^2 + 2pen^2(m_N^*) + t \right) \end{aligned}$$

$$\begin{aligned}
& - \sum_{i=1}^N w_i \cdot |m_N^*(x_i) - m(x_i)|^2 - \text{pen}^2(m_N^*) \\
& = \frac{1}{3} \sum_{i=1}^N w_i \cdot |\hat{m}_N(x_i) - m(x_i)|^2 + \frac{1}{3} \text{pen}^2(\tilde{m}_N) \\
& \quad + \frac{1}{3} \sum_{i=1}^N w_i \cdot |m_N^*(x_i) - m(x_i)|^2 + \frac{1}{3} \text{pen}^2(m_N^*) + \frac{2}{3}t \\
& = \frac{1}{3} \sum_{i=1}^N w_i \cdot |(\hat{m}_N(x_i) - m_N^*(x_i)) - (m(x_i) - m_N^*(x_i))|^2 \\
& \quad + \frac{1}{3} \text{pen}^2(\tilde{m}_N) + \frac{1}{3} \sum_{i=1}^N w_i \cdot |m_N^*(x_i) - m(x_i)|^2 + \frac{1}{3} \text{pen}^2(m_N^*) + \frac{2}{3}t \\
& \geq \frac{1}{6} \sum_{i=1}^N w_i \cdot |\hat{m}_N(x_i) - m_N^*(x_i)|^2 - \frac{1}{3} \sum_{i=1}^N w_i \cdot |m(x_i) - m_N^*(x_i)|^2 \\
& \quad + \frac{1}{3} \text{pen}^2(\tilde{m}_N) + \frac{1}{3} \sum_{i=1}^N w_i \cdot |m_N^*(x_i) - m(x_i)|^2 + \frac{1}{3} \text{pen}^2(m_N^*) + \frac{2}{3}t \\
& \geq \frac{1}{6} \left(\sum_{i=1}^N w_i \cdot |\hat{m}_N(x_i) - m_N^*(x_i)|^2 + \text{pen}^2(\tilde{m}_N) \right) + \frac{2}{3}t.
\end{aligned}$$

In the next to last inequality we have used, that $a^2/2 - b^2 \leq (a - b)^2$ ($a, b \in \mathbb{R}$) with $a = \hat{m}_N(x_i) - m_N^*(x_i)$ and $b = m(x_i) - m_N^*(x_i)$. \square

Proof of Lemma 3. We have

$$\begin{aligned}
& \mathbf{P} \left\{ \|\hat{m}_n - \hat{m}_n^*\|_n^2 + \text{pen}_n^2(\tilde{m}_n) + 4\delta_n \leq \frac{24}{n} \sum_{i=1}^n (\hat{m}_n(x_i) - \hat{m}_n^*(x_i)) \cdot W_i \right\} \\
& \leq P_1 + P_2
\end{aligned}$$

where

$$P_1 = \mathbf{P} \left\{ \frac{1}{n} \sum_{i=1}^n W_i^2 > 2\sigma_0^2 \right\}$$

and

$$\begin{aligned}
P_2 & = \mathbf{P} \left\{ \frac{1}{n} \sum_{i=1}^n W_i^2 \leq 2\sigma_0^2, \|\hat{m}_n - \hat{m}_n^*\|_n^2 + \text{pen}_n^2(\tilde{m}_n) + 4\delta_n \right. \\
& \quad \left. \leq \frac{24}{n} \sum_{i=1}^n (\hat{m}_n(x_i) - \hat{m}_n^*(x_i)) \cdot W_i \right\}.
\end{aligned}$$

Application of Chernoff's exponential bounding method (cf. Chernoff (1952)) together with (5.4) yields

$$P_1 = \mathbf{P} \left\{ \sum_{i=1}^n W_i^2 / K^2 > 2n\sigma_0^2 / K^2 \right\}$$

$$\begin{aligned}
 &\leq \mathbf{P} \left\{ \exp \left(\sum_{i=1}^n W_i^2 / K^2 \right) > \exp (2n\sigma_0^2 / K^2) \right\} \\
 &\leq \exp (-2n\sigma_0^2 / K^2) \cdot \mathbf{E} \left\{ \exp \left(\sum_{i=1}^n W_i^2 / K^2 \right) \right\} \\
 &\leq \exp (-2n\sigma_0^2 / K^2) \cdot (1 + \sigma_0^2 / K^2)^n \\
 &\leq \exp (-2n\sigma_0^2 / K^2) \cdot \exp (n \cdot \sigma_0^2 / K^2) = \exp (-n\sigma_0^2 / K^2).
 \end{aligned}$$

To bound P_2 , we observe first that $1/n \sum_{i=1}^n W_i^2 \leq 2\sigma_0^2$ together with the Cauchy-Schwarz inequality implies

$$\begin{aligned}
 \frac{24}{n} \sum_{i=1}^n (\hat{m}_n(x_i) - \hat{m}_n^*(x_i)) \cdot W_i &\leq 24 \cdot \sqrt{\frac{1}{n} \sum_{i=1}^n (\hat{m}_n(x_i) - \hat{m}_n^*(x_i))^2} \cdot \sqrt{2\sigma_0^2} \\
 &\leq 24 \cdot \sqrt{\frac{1}{n} \sum_{i=1}^n (\hat{m}_n(x_i) - \hat{m}_n^*(x_i))^2 + \text{pen}_n^2(\tilde{m}_n)} \cdot \sqrt{2\sigma_0^2}
 \end{aligned}$$

hence inside of P_2 we have

$$\frac{1}{n} \sum_{i=1}^n (\hat{m}_n(x_i) - \hat{m}_n^*(x_i))^2 + \text{pen}_n^2(\tilde{m}_n) \leq 1152\sigma_0^2.$$

Set

$$S = \min\{s \in \mathbb{N}_0 : 4 \cdot 2^s \delta_n > 1152\sigma_0^2\}.$$

Application of the peeling device (cf. Section 5.3 in van de Geer (2000)) yields

$$\begin{aligned}
 P_2 &= \sum_{s=1}^S \mathbf{P} \left\{ \frac{1}{n} \sum_{i=1}^n W_i^2 \leq 2\sigma_0^2, 4 \cdot 2^{s-1} \delta_n \cdot I_{\{s \neq 1\}} \right. \\
 &\quad \left. \leq \|\hat{m}_n - \hat{m}_n^*\|_n^2 + \text{pen}_n^2(\tilde{m}_n) < 4 \cdot 2^s \delta_n, \right. \\
 &\quad \left. \|\hat{m}_n - \hat{m}_n^*\|_n^2 + \text{pen}_n^2(\tilde{m}_n) + 4\delta_n \leq \frac{24}{n} \sum_{i=1}^n (\hat{m}_n(x_i) - \hat{m}_n^*(x_i)) \cdot W_i \right\} \\
 &\leq \sum_{s=1}^S \mathbf{P} \left\{ \frac{1}{n} \sum_{i=1}^n W_i^2 \leq 2\sigma_0^2, \|\hat{m}_n - \hat{m}_n^*\|_n^2 + \text{pen}_n^2(\tilde{m}_n) < 4 \cdot 2^s \delta_n, \right. \\
 &\quad \left. \frac{1}{12} \cdot 2^s \delta_n \leq \frac{1}{n} \sum_{i=1}^n (\hat{m}_n(x_i) - \hat{m}_n^*(x_i)) \cdot W_i \right\}
 \end{aligned}$$

The probabilities in the above sum can be bounded by Corollary 8.3 in van de Geer (2000) (use there $R = \sqrt{4 \cdot 2^s \delta_n}$, $\delta = \frac{1}{12} \cdot 2^s \delta_n$ and $\sigma = \sqrt{2}\sigma_0$). This yields

$$\begin{aligned}
 P_2 &\leq \sum_{s=1}^{\infty} c_{60} \exp \left(-\frac{n \cdot (\frac{1}{12} \cdot 2^s \delta_n)^2}{4c_{60} \cdot 4 \cdot 2^s \delta_n} \right) = \sum_{s=1}^{\infty} c_{60} \exp \left(-\frac{n \cdot 2^s \cdot \delta_n}{c_{61}} \right) \\
 &\leq \sum_{s=1}^{\infty} c_{60} \exp \left(-\frac{n \cdot (s+1) \cdot \delta_n}{c_{60}} \right) \leq c_{62} \exp \left(-\frac{n\delta_n}{c_{62}} \right). \quad \square
 \end{aligned}$$

Proof of Lemma 4. For $f : \mathbb{R}^d \rightarrow \mathbb{R}$ set

$$\|f\|_n^2 = \frac{1}{n} \sum_{i=1}^n |f(X_i)|^2.$$

We have

$$\begin{aligned} & \mathbf{P} \left\{ \int |\hat{m}_n(x) - m(x)|^2 \mathbf{P}_X(dx) \right. \\ & \qquad \qquad \qquad \left. > \delta_n + 3 \cdot \text{pen}_n^2(\tilde{m}_n) + 3 \frac{1}{n} \sum_{i=1}^n |\hat{m}_n(X_i) - m(X_i)|^2 \right\} \\ &= \mathbf{P} \left\{ 2 \int |\hat{m}_n(x) - m(x)|^2 \mathbf{P}_X(dx) - 2 \|\hat{m}_n - m\|_n^2 \right. \\ & \qquad \qquad \qquad \left. > \delta_n + 3 \cdot \text{pen}_n^2(\tilde{m}_n) + \int |\hat{m}_n(x) - m(x)|^2 \mathbf{P}_X(dx) + \|\hat{m}_n - m\|_n^2 \right\} \\ &\leq \mathbf{P} \left\{ \exists f \in \mathcal{F}_n : \right. \\ & \qquad \qquad \qquad \left. \frac{|\int |T_{\beta_n} f(x) - m(x)|^2 \mathbf{P}_X(dx) - \|T_{\beta_n} f - m\|_n^2|}{\delta_n + 3 \cdot \text{pen}_n^2(f) + \int |T_{\beta_n} f(x) - m(x)|^2 \mathbf{P}_X(dx) + \|T_{\beta_n} f - m\|_n^2} > \frac{1}{2} \right\} \\ &\leq \sum_{s=1}^{\infty} \mathbf{P} \left\{ \exists f \in \mathcal{F}_n : I_{\{s \neq 0\}} \cdot 2^{s-1} \cdot \delta_n \leq \text{pen}_n^2(f) \leq 2^s \delta_n, \right. \\ & \qquad \qquad \qquad \left. \frac{|\int |T_{\beta_n} f(x) - m(x)|^2 \mathbf{P}_X(dx) - \|T_{\beta_n} f - m\|_n^2|}{\delta_n + 3 \cdot \text{pen}_n^2(f) + \int |T_{\beta_n} f(x) - m(x)|^2 \mathbf{P}_X(dx) + \|T_{\beta_n} f - m\|_n^2} > \frac{1}{2} \right\} \\ &\leq \sum_{s=1}^{\infty} \mathbf{P} \left\{ \exists f \in \mathcal{F}_n : \text{pen}_n^2(f) \leq 2^s \delta_n, \right. \\ & \qquad \qquad \qquad \left. \frac{|\int |T_{\beta_n} f(x) - m(x)|^2 \mathbf{P}_X(dx) - \|T_{\beta_n} f - m\|_n^2|}{2^{s-1} \delta_n + \int |T_{\beta_n} f(x) - m(x)|^2 \mathbf{P}_X(dx) + \|T_{\beta_n} f - m\|_n^2} > \frac{1}{2} \right\}. \end{aligned}$$

The probabilities in the above sum can be bounded by Theorem 19.2 in Györfi et al. (2002) (which we apply with

$$\mathcal{F} = \{(T_{\beta_n} f - m)^2 : f \in \mathcal{F}_n, \text{pen}_n^2(f) \leq 2^s \delta_n\},$$

$K = 4\beta_n^2$, $\epsilon = 1/2$, and $\alpha = 2^{s-1} \delta_n$. Here in the integral of the covering number we use the fact that for $\delta \geq \alpha \cdot K/2 \geq 2 \cdot \alpha = 2^s \cdot \delta_n$ the condition $\text{pen}_n^2(f) \leq 2^s \delta_n$ inside \mathcal{F} implies $\text{pen}_n^2(f) \leq \delta$.) This yields

$$P_{1,n} \leq \sum_{s=1}^{\infty} 15 \cdot \exp\left(-\frac{n \cdot 2^s \cdot \delta_n}{c_{63} \cdot \beta_n^2}\right) \leq c_{64} \cdot \exp\left(-\frac{n \cdot \delta_n}{c_{64} \cdot \beta_n^2}\right). \quad \square$$