Estimation of a Regression Function by Maxima of Minima of Linear Functions

Adil M. Bagirov, Conny Clausen, and Michael Kohler

***Abstract*—In this paper, estimation of a regression function from independent and identically distributed random variables is considered. Estimates are defined by minimization of the empirical L_2 risk over a class of functions, which are defined as maxima of minima of linear functions. Results concerning the rate of convergence of the estimates are derived. In particular, it is shown that for smooth regression functions satisfying the assumption of single index models, the estimate is able to achieve (up to some logarithmic factor) the corresponding optimal one-dimensional rate of convergence. Hence, under these assumptions, the estimate is able to circumvent the so-called curse of dimensionality. The small sample behavior of the estimates is illustrated by applying them to simulated data.**

***Index Terms*—Adaptation, dimension reduction, L_2 error, nonparametric regression, rate of convergence, single index model.**

I. Introduction

THIS paper considers the problem of estimating a multivariate regression function given a sample of the underlying distribution. In applications, usually no *a priori* information about the regression function is known, therefore it is necessary to apply nonparametric methods for this estimation problem. There are two classes of established methods for this estimation problem. In the first class, knowledge about the structure of the regression function is assumed as in additive models (cf., e.g., [27] and [28]) or in semiparametric models (cf., e.g., [14]). The second class consists of completely nonparametric models, which do not explicitly assume any structure of the regression function in order to construct an estimate. Examples for established methods from this class include regression trees such as CART (cf., [9]), adaptive spline fitting such as MARS (cf., [12]), and least squares neural network estimates (cf., e.g., [16, Ch. 11]). All these methods minimize a kind of least squares risk of the regression estimate, either heuristically over a fixed and very complex function space as for neural networks or over a stepwise defined data-dependent space of piecewise constant functions or piecewise polynomials as for CART or MARS.

In this paper, we consider a rather complex function space consisting of maxima of minima of linear functions, over which we minimize a least squares risk. Since each maximum of minima of linear functions is in fact a continuous piecewise linear function, we fit a linear spline function with *free* knots to the data. This seems to be very promising since splines with free knots have very good approximation properties; see, e.g., [25]. But in contrast to MARS, we do not need heuristics to choose these free knots, but use instead advanced methods of optimization theory of nonlinear and nonconvex functions to compute our estimate approximately in an application.

A. Regression Estimation

In *regression analysis*, an $\mathbb{R}^d \times \mathbb{R}$-valued random vector (X, Y) with $\mathbf{E}Y^2 < \infty$ is considered and the dependency of Y on the value of X is of interest. More precisely, the goal is to find a function $f : \mathbb{R}^d \to \mathbb{R}$ such that $f(X)$ is a "good approximation" of Y. In this paper, we assume that the main aim of the analysis is minimization of the mean squared prediction error or L_2 *risk*

$$\mathbf{E}\{|f(X) - Y|^2\}.$$

In this case, the optimal function is the so-called *regression function*

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

Indeed, let $f : \mathbb{R}^d \to \mathbb{R}$ be an arbitrary (measurable) function and denote the distribution of X by μ. The well-known relation

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

(cf. e.g., [13, eq. (1.1)]) implies that the regression function is the optimal predictor in view of minimization of the L_2 risk

$$\mathbf{E}\{|m(X) - Y|^2\} = \min_{f:\mathbb{R}^d\to\mathbb{R}} \mathbf{E}\{|f(X) - Y|^2\}.$$

In addition, any function f is a good predictor in the sense that its L_2 risk is close to the optimal value, if and only if the so-called L_2 *error*

$$\int |f(x) - m(x)|^2\mu(dx) \tag{1}$$

is small. This motivates to measure the error caused by using a function f instead of the regression function by the L_2 error (1).

In applications, usually the distribution of (X, Y) (and hence also the regression function) is unknown. But often it

Manuscript received November 27, 2007; revised September 26, 2008. Current version published February 04, 2009.

A. Bagirov is with the School of Information Technology and Mathematical Sciences, University of Ballarat, Ballarat, Vic. 3353, Australia (e-mail: a.bagirov@ballarat.edu.au).

C. Clausen is with the Department of Mathematics, Universität des Saarlandes, D-66041 Saarbrücken, Germany (e-mail: clausen@math.uni-sb.de).

M. Kohler is with the Department of Mathematics, Technische Universität Darmstadt, D-64289 Darmstadt, Germany (e-mail: kohler@mathematik.tu-darmstadt.de).

Communicated by A. Krzyżak, Associate Editor for Pattern Recognition, Statistical Learning and Inference.

Color versions of Figures 1–4 in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIT.2008.2009835

is possible to observe a sample of the underlying distribution. This leads to the *regression estimation* problem. Here, $(X,Y),(X_1,Y_1),(X_2,Y_2),\ldots$ are independent and identically distributed random vectors. The set of data

$$\mathcal{D}_n = \{(X_1,Y_1),\ldots,(X_n,Y_n)\}$$

is given, and the goal is to construct an estimate

$$m_n(\cdot) = m_n(\cdot,\mathcal{D}_n) : \mathbb{R}^d \to \mathbb{R}$$

of the regression function such that the L_2 error

$$\int |m_n(x) - m(x)|^2 \mu(dx)$$

is small. For a detailed introduction to nonparametric regression, we refer the reader to the monograph [13].

B. Definition of the Estimate

In the sequel, we will use the principle of least squares to fit maxima of minima of linear functions to the data. More precisely, let $K_n \in \mathbb{N}$ and $L_{1,n},\ldots,L_{K_n,n} \in \mathbb{N}$ be parameters of the estimate and let $\mathcal{F}_n$ be the set of all functions

$$f : \mathbb{R}^d \to \mathbb{R}, f(x) = \max_{k=1,\ldots,K_n} \min_{l=1,\ldots,L_{k,n}} (a_{k,l}\cdot x + b_{k,l})$$

for some $a_{k,l} \in \mathbb{R}^d, b_{k,l} \in \mathbb{R}$, where

$$a_{k,l}\cdot x = a_{k,l}^{(1)}\cdot x^{(1)} + \ldots + a_{k,l}^{(d)}\cdot x^{(d)}$$

denotes the scalar product between $a_{k,l} = (a_{k,l}^{(1)},\ldots,a_{k,l}^{(d)})^T$ and $x = (x^{(1)},\ldots,x^{(d)})^T$. For this class of functions, the estimate $\tilde{m}_n$ is defined by

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

Here we assume that the minimum exists, however we do not require it to be unique.

In Section II, we will analyze the rate of convergence of a truncated version of this least squares estimate defined by

$$m_n(\cdot) = T_{\beta_n}(\tilde{m}_n(\cdot))$$

where

$$T_{\beta_n}(z) = \begin{cases} \beta_n, & z > \beta_n \\ z, & -\beta_n \le z \le \beta_n \\ -\beta_n, & z < -\beta_n \end{cases}$$

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

C. Main Results

Under a sub-Gaussian condition on the distribution of Y and for bounded support of the distribution of X, we show that the L_2 error of the estimate achieves for (p,C)-smooth regression function with $p \le 2$ (where roughly speaking all partial derivates of the regression function of order p exist) the corresponding optimal rate of convergence

$$n^{-2p/(2p+d)}$$

up to some logarithmic factor. For single index models, where the regression function m satisfies in addition

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

for some univariate function $\overline{m}$ and some vector $\beta \in \mathbb{R}^d$, we show furthermore that our estimate achieves (up to some logarithmic factor) the one-dimensional rate of convergence

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

Hence, under these assumptions, the estimate is able to circumvent the so-called curse of dimensionality.

D. Discussion of Related Results

In multivariate nonparametric regression function estimation, there is a gap between theory and practice. The established estimates such as CART, MARS, or least squares neural networks are based on several heuristics for computing the estimates, which makes it basically impossible to analyze their rate of convergence theoretically. However, if one defines them without these heuristics, their rate of convergence can be analyzed (and this has been done for neural networks, e.g., in [5] and [6] and for CART in [18]), but in this form, the estimates cannot be computed in an application. For our estimate, a similar phenomenon occurs since we need heuristics to compute it approximately in an application. The difference between our approach and the above established estimates is that we use heuristics from advanced optimization theory, in particular, from the optimization theory of nonlinear and nonconvex optimization (cf., e.g., [1], [2], and [4]) instead of complicated heuristics from statistics for stepwise computation as for CART or MARS, or a simple gradient descent as for least squares neural networks.

It follows from [26] that the rates of convergence, which we derive, are optimal (in some minimax sense) up to a logarithmic factor. The idea of imposing additional restrictions on the structure of the regression function (such as additivity or the assumption in the single index model) and to derive under these assumption better rates of convergence is due to [27] and [28].

We use a theorem of [22] to derive our rate of convergence results. This approach is described in detail in [13, Sec. 11.3]. Below we extend this approach to unbounded data (which satisfies a sub-Gaussian condition) by introducing new truncation arguments. In this way, we are able to derive the results under similar general assumptions on the distribution of Y as with alternative methods from empirical process theory; see, e.g., the monograph [29] or [19] and [20].

Maxima of minima of linear functions have been used in regression estimation previously in [7]. The least squares estimates there are derived by minimizing the empirical L_2 risk over classes of functions consisting of Lipschitz smooth functions where a bound on the Lipschitz constant is given. It is shown that the resulting estimate is in fact a maximum of minima of linear functions, where the number of minima occurring in the maximum is equal to the sample size. Additional restrictions (e.g., on the linear functions in the minima) ensure that there will be no overfitting. In contrast, the number of linear functions that we consider in this paper is much smaller and restrictions on

these linear functions are therefore not necessary. This seems to be promising, because we do not fit too many parameters to the data.

The estimate considered in this paper is a continuous piecewise linear function. In [8], a sum of maxima or minima of two linear functions was fitted to data, which also produces continuous piecewise linear estimates. As it is shown in [24], the fitting procedure proposed in [8] can be considered as the Newton algorithm for function minimization applied to a sum of squared error criterion. In contrast, our estimate is able to vary locally much more than a sum of maxima or minima of two linear functions, and the fitting procedure we use is different and is based on much more advanced techniques from optimization theory.

In Corollary 2, we show that even for large dimension of X the L_2 error of our estimate converges to zero quickly if the regression function satisfies the structural assumption of single index models. More general results in this respect have been proven in [17], where a detailed discussion of related results in the literature can also be found. A problem with this kind of results is always the implementation of the estimate. In [17], a backfitting procedure is proposed, where each of the steps leads to a relatively simple minimization problem. However, there is no guarantee that the combination of these steps leads to an algorithm really solving the considered global minimization problem. In contrast, in this paper, we try to solve the global optimization problem in one step. So the main result here is to derive this good rate of convergence for an estimate for which an algorithm exists, which really tries to solve the optimization problem that occurs by using advanced techniques from optimization theory. This algorithm is described in detail in [3].

The independence of the data assumed in this paper could possibly be weakened to permit martingale difference or mixing sequences of data. Since this would complicate the technical analysis and produce a less transparent treatment, we did not try to do this. However, the derived approximation results and bounds on covering numbers can be used, e.g., together with techniques introduced in [11] to derive results under weaker conditions.

E. Notations

The sets of natural numbers, natural numbers including zero, real numbers, and nonnegative real numbers are denoted by $\mathbb{N}$, $\mathbb{N}_0$, $\mathbb{R}$, and $\mathbb{R}_+$, respectively. For vectors $x \in \mathbb{R}^n$, we denote by $||x||$ the Euclidian norm of x and by $x \cdot y$ the scalar product between x and y. The least integer greater or equal to a real number x will be denoted by $\lceil x \rceil$. $\log(x)$ denotes the natural logarithm of $x > 0$. For a function $f : \mathbb{R}^d \to \mathbb{R}$

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

denotes the supremum norm.

F. Outline of the Paper

The main theoretical result is formulated in Section II and proven in Section IV. In Section III, the estimate is illustrated by applying it to simulated data.

II. Analysis of the Rate of Convergence of the Estimate

Our first theorem gives an upper bound for the expected L_2 error of our estimate.

Theorem 1: Let $K_n, L_{1,n}, \ldots, L_{K_n,n} \in \mathbb{N}$, with

$$K_n \cdot \max\{L_{1,n}, \ldots, L_{K_n,n}\} \leq n^2$$

and set $\beta_n = c_1 \cdot \log(n)$ for some constant $c_1 > 0$. Assume that the distribution of (X, Y) satisfies

$$\mathbf{E}\left(e^{c_2 \cdot |Y|^2}\right) < \infty \tag{3}$$

for some constant $c_2 > 0$ and that the regression function m is bounded in absolute value. Then, for the estimate m_n defined as in Section I-C

$$\begin{aligned}\mathbf{E} \int |m_n(x) - m(x)|^2 \mu(dx) &\leq \frac{c_3 \cdot \log(n)^3 \cdot \sum_{k=1}^{K_n} L_{k,n}}{n} \\ &\quad + \mathbf{E}\left(2 \inf_{f \in \mathcal{F}_n} \left(\frac{1}{n} \sum_{i=1}^{n} |f(X_i) - Y_i|^2 \right.\right. \\ &\quad \left.\left. - \frac{1}{n} \sum_{i=1}^{n} |m(X_i) - Y_i|^2\right)\right) \end{aligned} \tag{4}$$

for some constant $c_3 > 0$ and hence also

$$\begin{aligned}\mathbf{E} \int |m_n(x) - m(x)|^2 \mu(dx) &\leq \frac{c_3 \cdot \log(n)^3 \cdot \sum_{k=1}^{K_n} L_{k,n}}{n} \\ &\quad + 2 \cdot \inf_{f \in \mathcal{F}_n} \int |f(x) - m(x)|^2 \mu(dx)\end{aligned}$$

where c_3 does not depend on n, β_n or the parameters of the estimate.

The condition (3) is a modified sub-Gaussian condition and it is particularly satisfied, if $\mathbf{P}_{Y|X=x}$ is the normal distribution $\mathcal{N}_{(m(x),\sigma^2)}$ and the regression function m is bounded. This condition allows us to consider an unbounded support of the conditional distribution of Y.

Together with an approximation result this theorem implies the next corollary, which considers the rate of convergence of the estimate. Here it is necessary to impose smoothness conditions on the regression function.

Definition 1: Let $p = k + \beta$ for some $k \in \mathbb{N}_0$ and $0 < \beta \leq 1$ and let $C > 0$. A function $m : [a, b]^d \to \mathbb{R}$ is called (p, C)-smooth if for every $\alpha = (\alpha_1, \ldots, \alpha_d), \alpha_i \in \mathbb{N}_0, \sum_{j=1}^{d} \alpha_j = k$, the partial derivative

$$\frac{\partial^k f}{\partial x_1^{\alpha_1} \ldots \partial x_d^{\alpha_d}}$$

exists and satisfies

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

for all $x, z \in [a, b]^d$.

Corollary 1: Assume that the distribution of (X,Y) satisfies that $X \in [a,b]^d$ a.s. for some $a,b \in \mathbb{R}$, that the modified sub-Gaussian condition $\mathbf{E}(\exp(c_2 \cdot |Y|^2)) < \infty$ is fulfilled for some constant $c_2 > 0$ and that m is (p,C)-smooth for some $0 < p \leq 2$ and $C \geq 1$. Set $\beta_n = c_1 \cdot \log(n)$ for some $c_1 > 0$

$$K_n = \left\lceil C^{\frac{2d}{2p+d}} \cdot \left(\frac{n}{\log(n)^3} \right)^{d/(2p+d)} \right\rceil$$

and $L_{k,n} = L_k = 2d+1$ $(k = 1, \ldots, K_n)$. Then, we have for the estimate m_n defined as above

$$\mathbf{E} \int |m_n(x) - m(x)|^2 \mu(dx) \leq \text{const} \cdot C^{\frac{2d}{2p+d}} \cdot \left(\frac{\log(n)^3}{n} \right)^{\frac{2p}{2p+d}}.$$

The above rate of convergence is slow in case of large dimension d of the predictor variable X (so-called curse of dimensionality). Next we present a result that shows that under structural assumptions on the regression function (more precisely, for single index models) our estimate is able to circumvent the so-called curse of dimensionality.

Corollary 2: Assume that the distribution of (X,Y) satisfies that $X \in [a,b]^d$ a.s. for some $a,b \in \mathbb{R}$ and that the modified sub-Gaussian condition $\mathbf{E}(\exp(c_2 \cdot |Y|^2)) < \infty$ is fulfilled for some constant $c_2 > 0$. Furthermore assume that the regression function m satisfies

$$m(x) = \overline{m}(\alpha \cdot x), \qquad (x \in \mathbb{R}^d)$$

for a function $\overline{m} : \mathbb{R} \to \mathbb{R}$ and some $\alpha \in \mathbb{R}^d$, and assume that $\bar{m}$ is (p,C)-smooth for some $0 < p \leq 2$ and $C \geq 1$. Then, for the estimate m_n as above and with the setting $\beta_n = c_1 \cdot \log(n)$ for some $c_1 > 0$

$$K_n = \left\lceil C^{\frac{2}{2p+1}} \cdot \left(\frac{n}{\log(n)^3} \right)^{1/(2p+1)} \right\rceil$$

and $L_{k,n} = L_k = 3$ $(k = 1, \ldots, K_n)$, we get

$$\mathbf{E} \int |m_n(x) - m(x)|^2 \mu(dx) \leq \text{const} \cdot C^{\frac{2}{2p+1}} \cdot \left(\frac{\log(n)^3}{n} \right)^{\frac{2p}{2p+1}}.$$

Remark 1: It follows from [26] that under the conditions of Corollary 1 no estimate can achieve (in some Minimax sense) a rate of convergence that converges faster to zero than

$$n^{-2p/(2p+d)}$$

(cf., e.g., [13, Ch. 3]). Hence, Corollary 1 implies that our estimate has an optimal rate of convergence up to the logarithmic factor.

Remark 2: In any application, the smoothness of the regression function [measured by (p,C)] is not known in advance and hence the parameters of the estimate have to be chosen in a data-dependent way. This can be done, e.g., by *splitting of* the sample, where the estimate is computed for various values of the parameters on a learning sample (consisting, e.g., of the first half of the data points) and the parameters are chosen such that the empirical L_2 risk on a testing sample (consisting, e.g., of the second half of the data points) is minimized (cf., e.g., [13, Ch. 7]).

Theoretical results concerning splitting of the sample can be found in [15] and [13, Ch. 7].

Remark 3: The assumption on the boundedness of the support of X in Corollary 1 can be replaced by the weaker assumption

$$\mathbf{E}(\|X\|^\beta) < \infty$$

for some $\beta > 2p$. To prove Corollary 1 under this weaker assumption, one replaces the partition Π in the proof of Corollary 1 by the partition used in [21, Sec. II], where the diameter of the cubes depends on the distance to the origin. Arguing then as in the proof of Theorem 1 in [21], one gets the assertion.

In the same way, the assumption of boundedness of the support of X in Corollary 2 can be replaced by

$$\mathbf{E}(|\alpha \cdot X|^\beta) < \infty$$

for some $\beta > 2p$.

Remark 4: By using [13, Th. A.1], together with a result concerning approximation of smooth functions by continuous linear functions, it is easy to see that Theorem 1 implies the following consistency result: If $K_n, L_{1,n}, \ldots, L_{K_n,n} \in \mathbb{N}$ satisfy $L_{i,n} \geq 2d+1$ $(i = 1, \ldots, K_n)$, $K_n \cdot \max\{L_{1,n}, \ldots, L_{K_n,n}\} \leq n^2$

$$K_n \to \infty, \qquad (n \to \infty)$$

and

$$\frac{\log(n)^3 \cdot \sum_{k=1}^{K_n} L_{k,n}}{n} \to 0, \qquad (n \to \infty)$$

then

$$\mathbf{E} \int |m_n(x) - m(x)|^2 \mu(dx) \to 0, \qquad (n \to \infty)$$

for all distributions of (X,Y), which satisfy (3). Here condition (3) can be relaxed to $\mathbf{E}Y^2 < \infty$ by using [13, Th. 10.3] together with the results derived in the proof of Theorem 1.

III. Application to Simulated Data

In our applications, we choose the number K of minima and the number $L = L_1 = \ldots = L_K$ of linear functions in each minimum in a data-dependent way by splitting of the sample. We split the sample of size n in a learning sample of size $n_l < n$ and a testing sample of size $n_t = n - n_l$. We use the learning sample to define for a fixed numbers K and L an estimate $\tilde{m}_{n_l,(K,L)}$, and compute the empirical L_2 risk of this estimate on the testing sample. Since the testing sample is independent of the learning sample, this gives us an unbiased estimate of the L_2 risk of $\tilde{m}_{n_l,(K,L)}$. Then, we choose (K,L) by minimizing this estimate with respect to (K,L). In the sequel, we use $n \in \{500, 3000\}$ and $n_t = n_l = n/2$.

To compute the estimate for given numbers of linear functions, we have to minimize

$$\frac{1}{n} \sum_{i=1}^{n} \left| \left(\max_{k=1,\ldots,K} \min_{l=1,\ldots,L_k} (a_{k,l} \cdot x_i + b_{k,l}) \right) - y_i \right|^2$$

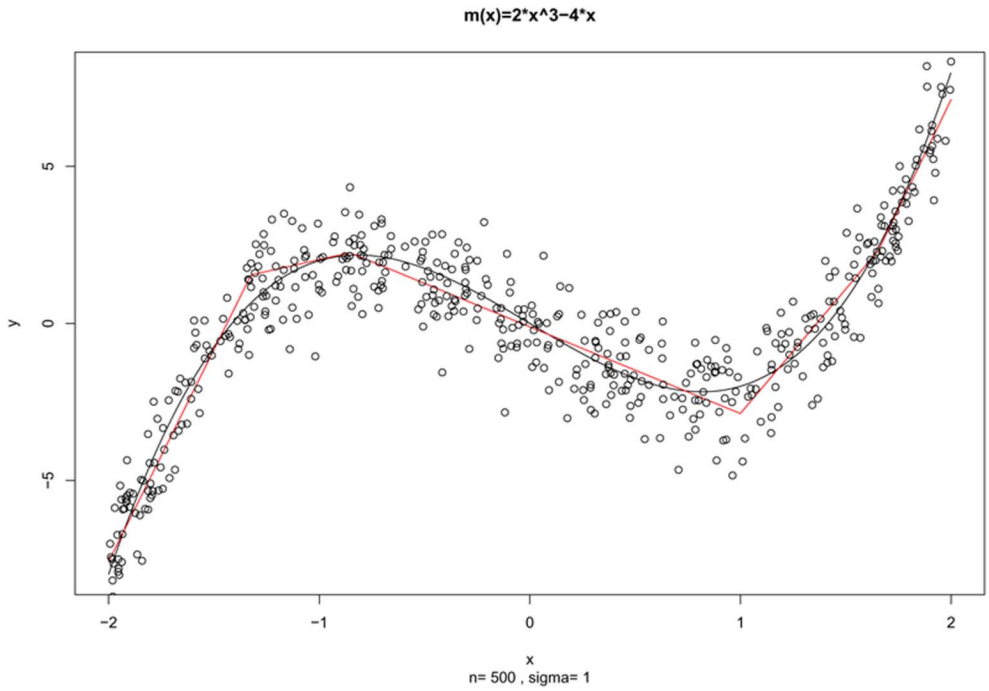

Fig. 1. Simulation with the first univariate regression function.

for given (fixed) $x_1, \ldots, x_n \in \mathbb{R}^d$, $y_1, \ldots, y_n \in \mathbb{R}$ with respect to

$$a_{k,l} \in \mathbb{R}^d \quad b_{k,l} \in \mathbb{R}, \qquad (k = 1, \ldots, K, l = 1, \ldots, L_k).$$

Unfortunately, we cannot solve this minimization problem exactly in general. The reason is that the function to be minimized is nonsmooth and nonconvex. Depending on K and L_k, it may have a large number of variables (more than a hundred even in the case of univariate data). The function has many local minima and their number increases drastically as the number of maxima and minima functions increases. Most of the local minimizers do not provide a good approximation to the data and therefore one is interested to find either a global minimizer or a minimizer that is close to a global one. Conventional methods of global optimization are not effective for minimizing of such functions, since they are very time consuming and cannot solve this problem in a reasonable time. Furthermore, the function to be minimized is a very complicated nonsmooth function and the calculation even of only one subgradient of such a function is a difficult task. Therefore, subgradient-based methods of nonsmooth optimization are not effective here.

Even though we cannot solve this minimization problem exactly, we are able to compute the estimate approximately. For this, we use the following properties of the function to be minimized: It is a semismooth function (cf., [23]); moreover, it is a smooth composition of so-called quasi-differentiable functions (see [10] for the definition of quasi-differentiable functions). Therefore, we can use the discrete gradient method from [2] to solve it. Furthermore, it is piecewise partially separable (see [4] for the definition of such functions). We use the version of the discrete gradient method described in [4] for minimizing piecewise partially separable functions to solve it. The discrete gradient method is a derivative-free method and it is especially effective for minimization of nonsmooth and nonconvex function when the subgradient is not available or it is difficult to calculate the subgradient.

A detailed description of the algorithm used to compute the estimate is given in [3]. An implementation of the estimate in Fortran is available from the authors by request.

In [3], the estimate is also compared to various other nonparametric regression estimates. In the sequel, we will illustrate it only by applying it to a few simulated data sets. Here, we define (X, Y) by

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

where X is uniformly distributed on $[-2, 2]^d$, ϵ is standard normally distributed and independent of X, and $\sigma \geq 0$. In Figs. 1–4, we choose $d = 1$ and $\sigma = 1$, and use four different univariate regression functions in order to define four different data sets of size $n = 500$. Each figure shows the true regression function together with its formula, a corresponding sample of size $n = 500$ and our estimate applied to this sample.

Here the first two examples show how the max–min estimate looks like rather simple regression estimates, while in the third and fourth example, the regression function has some local irregularity. Here it can be seen that our newly proposed estimate is able to adapt locally to such irregularities in the regression function.

Next we consider the case $d = 2$. In our fifth example, we choose

$$m(x^{(1)}, x^{(2)}) = x^{(1)} \cdot \sin((x^{(1)})^2) - x^{(2)} \cdot \sin((x^{(2)})^2)$$

and $n = 5000$ and $\sigma = 0.2$. Fig. 5 shows the regression function and our estimate applied to a corresponding data set of sample size 5000.

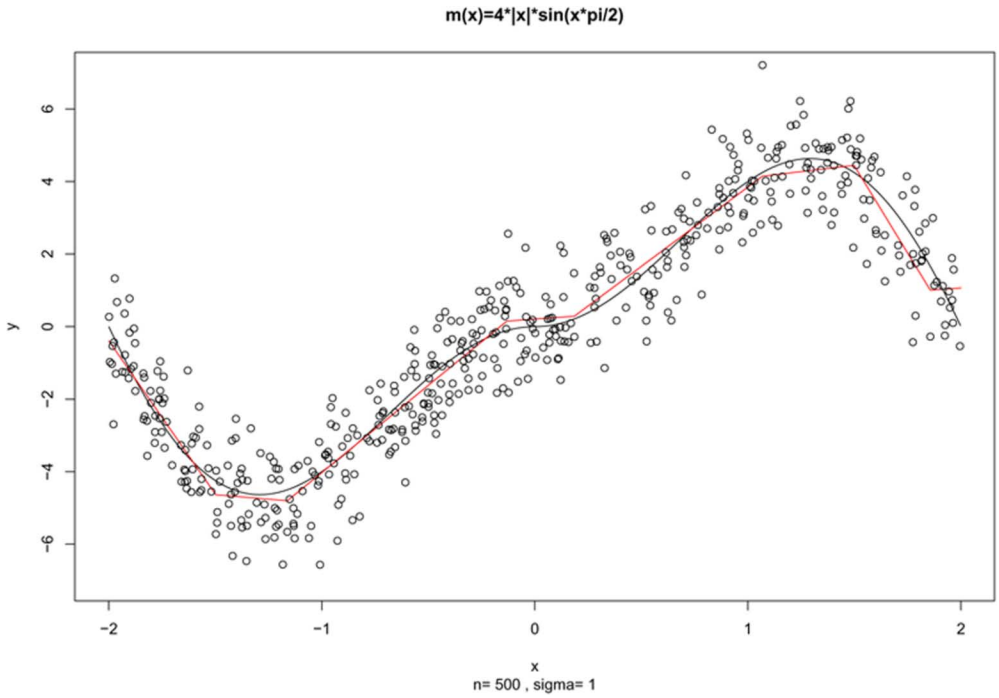

Fig. 2. Simulation with the second univariate regression function.

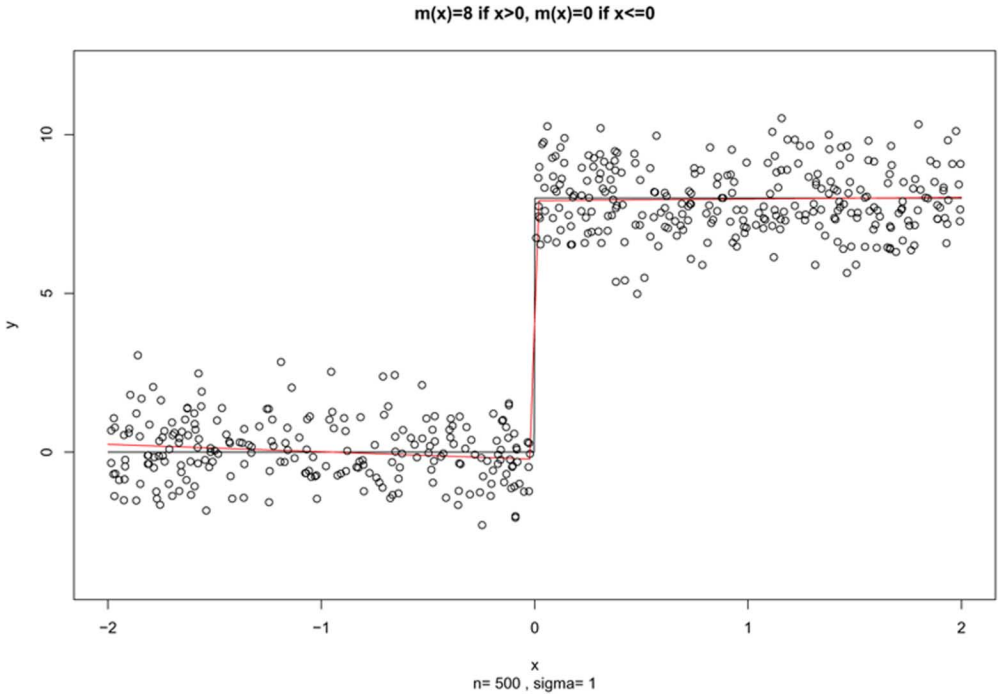

Fig. 3. Simulation with the third univariate regression function.

In our sixth example, we choose

$$m(x^{(1)}, x^{(2)}) = \frac{4}{1 + 4 * (x^{(1)})^2 + 4 * (x^{(2)})^2}$$

and again $n = 5000$ and $\sigma = 0.2$. Fig. 6 shows the regression function and our estimate applied to a corresponding data set of sample size 5000.

In our seventh (and final) example, we choose

$$m(x^{(1)}, x^{(2)}) = 6 - 2 * \min(3, 4 * (x^{(1)})^2 + 4 * |x^{(2)}|)$$

and again $n = 5000$ and $\sigma = 0.2$. Fig. 7 shows the regression function and our estimate applied to a corresponding data set of sample size 5000.

From the last simulation, we see again that our estimate is able to adapt to the local behavior of the regression function.

IV. Proofs

In the proofs, we need the notation of covering numbers.

Definition 2: Let $x_1, \ldots, x_n \in \mathbb{R}^d$ and set $x_1^n = (x_1, \ldots, x_n)$. Let $\mathcal{F}$ be a set of functions $f : \mathbb{R}^d \to \mathbb{R}$.

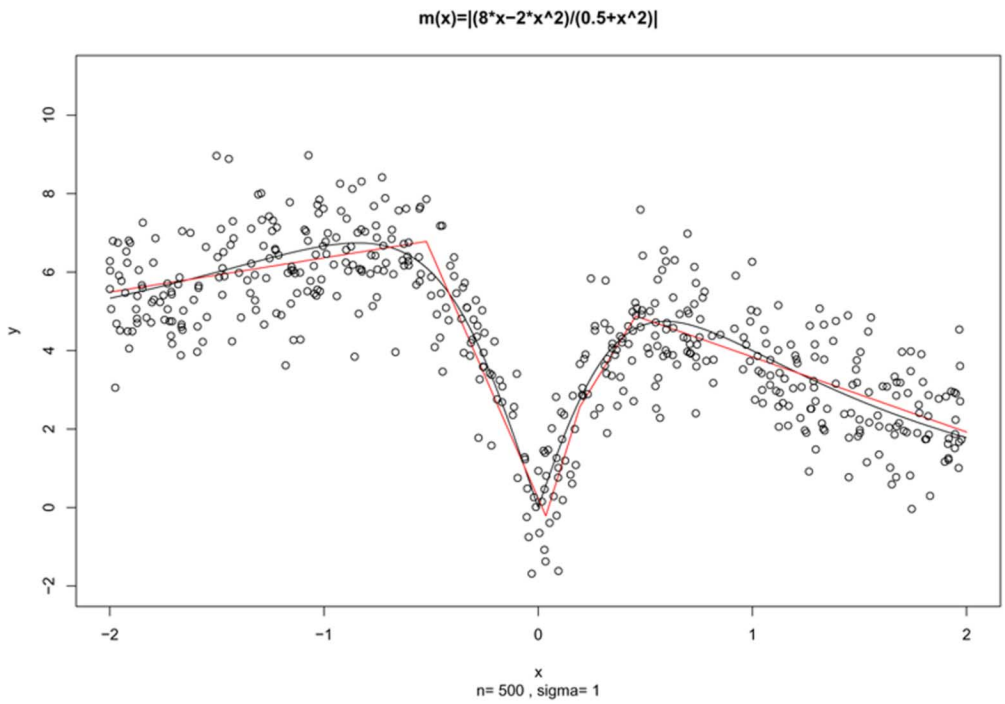

Fig. 4. Simulation with the fourth univariate regression function.

An L_p-ϵ-cover of $\mathcal{F}$ on x_1^n is a finite set of functions $f_1, \ldots, f_k : \mathbb{R}^d \to \mathbb{R}$ with the property

$$\min_{1\le j\le k}\left(\frac{1}{n}\sum_{i=1}^{n}|f(x_i)-f_j(x_i)|^p\right)^{1/p}<\epsilon, \qquad \text{for all } f\in\mathcal{F}. \tag{5}$$

The L_p-ϵ-covering number $\mathcal{N}_p(\epsilon,\mathcal{F},x_1^n)$ of $\mathcal{F}$ on x_1^n is the minimal size of a L_p-ϵ-cover of $\mathcal{F}$ on x_1^n. In case that there exists no finite L_p-ϵ-cover of $\mathcal{F}$, the L_p-ϵ-covering number of $\mathcal{F}$ on x_1^n is defined by $\mathcal{N}_p(\epsilon,\mathcal{F},x_1^n)=\infty$.

To get bounds for covering numbers of sets of maxima of minima of linear functions, we first show the connection between the L_p-ϵ-covering numbers of sets $\mathcal{G}_1,\mathcal{G}_2,\ldots$ and the L_p-ϵ-covering number of their maximum

$$\begin{aligned}&\max\{\mathcal{G}_1,\ldots,\mathcal{G}_l\}\\ &\quad=\Big\{f:\mathbb{R}^d\to\mathbb{R}: f(x)=\max\{g_1(x),\ldots,g_l(x)\},\\ &\qquad\qquad \text{for some } g_1\in\mathcal{G}_1,\ldots,g_l\in\mathcal{G}_l\Big\}\end{aligned}$$

and minimum (defined analogously), respectively.

Lemma 1: Let $\mathcal{G}_1,\mathcal{G}_2,\ldots,\mathcal{G}_l$ be l sets of functions from $\mathbb{R}^d$ to $\mathbb{R}$ and let $x_1^n=(x_1,\ldots,x_n)\in\mathbb{R}^d\times\cdots\times\mathbb{R}^d$ be n fixed points in $\mathbb{R}^d$. Then

$$\mathcal{N}_p\left(\epsilon,\max\left\{\mathcal{G}_1,\ldots,\mathcal{G}_l\right\},x_1^n\right)\le\prod_{i=1}^{l}\mathcal{N}_p\left(\frac{\epsilon}{l^{1/p}},\mathcal{G}_i,x_1^n\right) \tag{6}$$

and

$$\mathcal{N}_p\left(\epsilon,\min\left\{\mathcal{G}_1,\ldots,\mathcal{G}_l\right\},x_1^n\right)\le\prod_{i=1}^{l}\mathcal{N}_p\left(\frac{\epsilon}{l^{1/p}},\mathcal{G}_i,x_1^n\right). \tag{7}$$

Proof: Inequality (6) follows from

$$\begin{aligned}&\left(\frac{1}{n}\sum_{k=1}^{n}\left|\max_{i=1,\ldots,l}g_i(x_k)-\max_{i=1,\ldots,l}g_i^{j_i}(x_k)\right|^p\right)^{1/p}\\ &\quad\le\left(\frac{1}{n}\sum_{k=1}^{n}\max_{i=1,\ldots,l}\left|g_i(x_k)-g_i^{j_i}(x_k)\right|^p\right)^{1/p}\\ &\quad\le\left(\frac{1}{n}\sum_{k=1}^{n}\sum_{i=1}^{l}\left|g_i(x_k)-g_i^{j_i}(x_k)\right|^p\right)^{1/p}\\ &\quad\le l^{1/p}\cdot\max_{i=1,\ldots,l}\left(\frac{1}{n}\sum_{k=1}^{n}|g_i(x_k)-g_i^{j_i}(x_k)|^p\right)^{1/p}.\end{aligned}$$

Inequality (7) follows directly from (6) with $\min\{\mathcal{G}_1,\ldots,\mathcal{G}_l\}=-\max\{-\mathcal{G}_1,\ldots,-\mathcal{G}_l\}$. □

In the next lemma, we bound the L_p-ϵ-covering number of a truncated version of our class $\mathcal{F}_n$ of functions.

Lemma 2: Let $x_1^n\in\mathbb{R}^d\times\cdots\times\mathbb{R}^d$ and set $L_n:=\max\{L_{1,n},\ldots,L_{K_n,n}\}$. Then, for $0<\epsilon<\beta/2$

$$\mathcal{N}_1\left(\epsilon,T_\beta\mathcal{F}_n,x_1^n\right)\le 3\left(\frac{6e\beta}{\epsilon}\cdot K_nL_n\right)^{2(d+2)\left(\sum_{k=1}^{K_n}L_{k,n}\right)}.$$

Proof: In the first step of the proof, we show that we can involve the truncation operator into the class of functions, i.e., we show that $T_\beta\mathcal{F}_n$ is the class of all functions of the form

$$f:\mathbb{R}^d\to\mathbb{R}, f(x)=\max_{1\le k\le K_n}\min_{1\le l\le L_{k,n}}T_\beta\left(a_{k,l}\cdot x+b_{k,l}\right)$$

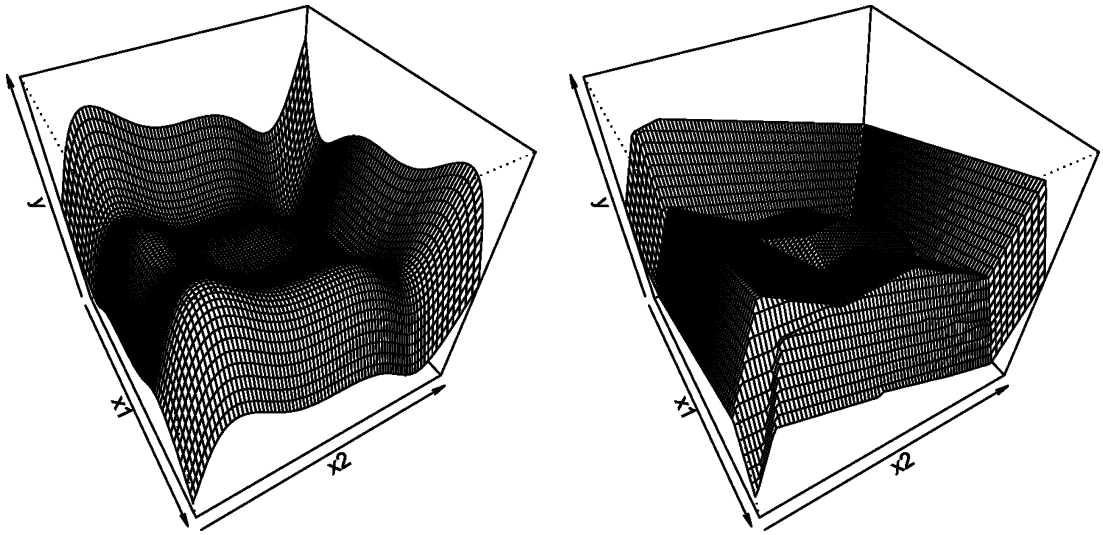

Fig. 5. Bivariate regression function together with our max–min estimate in the fifth example.

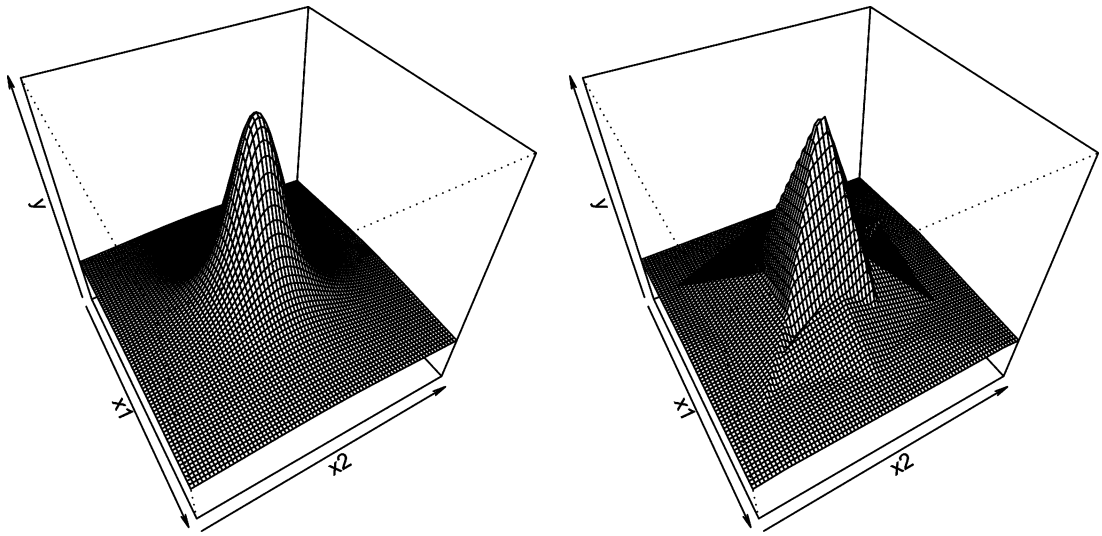

Fig. 6. Bivariate regression function together with our max–min estimate in the sixth example.

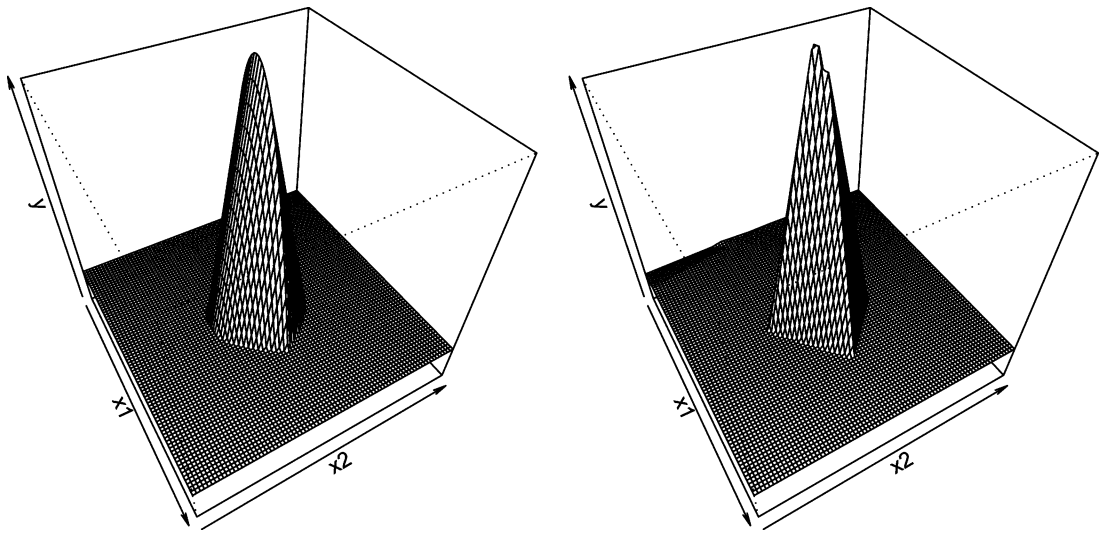

Fig. 7. Bivariate regression function together with our max–min estimate in the seventh example.

for some $a_{k,l} \in \mathbb{R}^d, b_{k,l} \in \mathbb{R}$. At the beginning, we observe that by monotonicity of the mapping $x \mapsto T_\beta x$, the equality

$$T_\beta \max_{1\le i\le n} z_i = \max_{1\le i\le n} T_\beta z_i \tag{8}$$

holds for real numbers $z_i \in \mathbb{R}, (i = 1,\dots,n)$. With $\min_{1\le i\le n} z_i = -\max_{1\le i\le n}(-z_i)$ and $T_\beta(-z) = -T_\beta(z)$, we get also

$$T_\beta \min_{1\le i\le n} z_i = \min_{1\le i\le n} T_\beta z_i$$

which implies the assertion of the first step. Set

$$\mathcal{G} = \Big\{ g : \mathbb{R}^d \to \mathbb{R} : g(x) = a_{k,l}\cdot x + b_{k,l}, \text{ for some } a_{k,l} \in \mathbb{R}^d, b_{k,l} \in \mathbb{R} \Big\}.$$

From Theorem 9.4, Theorem 9.5, and inequality (10.23) in [13], we get

$$\mathcal{N}_1(\epsilon, T_\beta \mathcal{G}, x_1^n) \le 3\left(\frac{4e\beta}{\epsilon}\cdot\log\frac{6e\beta}{\epsilon}\right)^{(d+1)+1}.$$

By applying Lemma 1, we get the desired result. □

With this bound of the covering number of $T_\beta \mathcal{F}_n$, we can now start with the proof of Theorem 1.

Proof of Theorem 1: In the proof, we use the following error decomposition:

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

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

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

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

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

With the Cauchy–Schwarz inequality and

$$I_{\{|Y|>\beta_n\}} \le \frac{\exp(c_2/2\cdot|Y|^2)}{\exp(c_2/2\cdot\beta_n^2)} \tag{9}$$

it follows that

$$\begin{aligned}
&|T_{5,n}| \\
&\quad \le \sqrt{\mathbf{E}\{|T_{\beta_n} Y - Y|^2\}} \\
&\qquad \cdot \sqrt{\mathbf{E}\{|2m_n(X) - Y - T_{\beta_n} Y|^2 | \mathcal{D}_n\}} \\
&\quad \le \sqrt{\mathbf{E}\{|Y|^2 \cdot I_{\{|Y|>\beta_n\}}\}} \\
&\qquad \cdot \sqrt{\mathbf{E}\{2\cdot|2m_n(X) - T_{\beta_n} Y|^2 + 2\cdot|Y|^2 | \mathcal{D}_n\}} \\
&\quad \le \sqrt{\mathbf{E}\left\{|Y|^2 \cdot \frac{\exp(c_2/2\cdot|Y|^2)}{\exp(c_2/2\cdot\beta_n^2)}\right\}} \\
&\qquad \cdot \sqrt{\mathbf{E}\{2\cdot|2m_n(X) - T_{\beta_n} Y|^2 | \mathcal{D}_n\} + 2\mathbf{E}\{|Y|^2\}} \\
&\quad \le \sqrt{\mathbf{E}\Big\{|Y|^2 \cdot \exp(c_2/2\cdot|Y|^2)\Big\}} \cdot \exp\left(-\frac{c_2\cdot\beta_n^2}{4}\right) \\
&\qquad \cdot \sqrt{2(3\beta_n)^2 + 2\mathbf{E}\{|Y|^2\}}.
\end{aligned}$$

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

$$|Y|^2 \le \frac{2}{c_2}\cdot\exp\left(\frac{c_2}{2}|Y|^2\right)$$

and hence $\sqrt{\mathbf{E}\Big\{|Y|^2\cdot\exp(c_2/2\cdot|Y|^2)\Big\}}$ is bounded by

$$\begin{aligned}
&\mathbf{E}\left(\frac{2}{c_2}\cdot\exp\left(c_2/2\cdot|Y|^2\right)\cdot\exp(c_2/2\cdot|Y|^2)\right) \\
&\quad \le \mathbf{E}\left(\frac{2}{c_2}\cdot\exp\left(c_2\cdot|Y|^2\right)\right) \le c_4
\end{aligned}$$

which is less than infinity by the assumptions of the theorem. Furthermore, the third term is bounded by $\sqrt{18\beta_n^2 + c_5}$ because

$$\mathbf{E}(|Y|^2) \le \mathbf{E}(1/c_2\cdot\exp(c_2\cdot|Y|^2) \le c_5 < \infty \tag{10}$$

which follows again as above. With the setting $\beta_n = c_1 \cdot \log(n)$, it follows that for some constants $c_6, c_7 > 0$ that $|T_{5,n}|$ is bounded by

$$\sqrt{c_4} \cdot \exp\left(-c_6 \cdot \log(n)^2\right) \cdot \sqrt{(18 \cdot c_1 \cdot \log(n)^2 + c_5)} \leq c_7 \cdot \frac{\log(n)}{n}.$$

From the Cauchy–Schwarz inequality, we get that $T_{6,n}$ is bounded by

$$\sqrt{2\mathbf{E}\left\{|(m(X) - m_{\beta_n}(X))|^2\right\} + 2\mathbf{E}\left\{|(T_{\beta_n}Y - Y)|^2\right\}} \cdot \sqrt{\mathbf{E}\left\{\left|m(X) + m_{\beta_n}(X) - Y - T_{\beta_n}Y\right|^2\right\}}$$

where we can bound the second factor on the right-hand side in the above inequality in the same way we have bounded the second factor from $T_{5,n}$, because by assumption, $\|m\|_\infty$ is bounded and furthermore m_{β_n} is bounded by β_n. Thus, we get for some constant $c_8 > 0$

$$\sqrt{\mathbf{E}\left\{\left|m(X) + m_{\beta_n}(X) - Y - T_{\beta_n}Y\right|^2\right\}} \leq c_8 \cdot \log(n).$$

Next we consider the first term. With the inequality of Jensen, it follows that

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

Hence, we get

$$T_{6,n} \leq \sqrt{4\mathbf{E}\left\{|Y - T_{\beta_n}Y|^2\right\}} \cdot c_8 \cdot \log(n)$$

and therefore with the calculations from $T_{5,n}$, it follows that $T_{6,n} \leq c_9 \cdot \log(n)/n$ for some constant $c_9 > 0$. Altogether, we get

$$T_{1,n} \leq c_{10} \cdot \frac{\log(n)}{n}$$

for some constant $c_{10} > 0$.

Next we consider $T_{2,n}$. Let $t > 1/n$ be arbitrary. Then, $\mathbf{P}\{T_{2,n} > t\}$ can be bounded by as shown in the equation at the bottom of the page. Thus, with [13, Th. 11.4] and

$$\mathcal{N}_1\left(\delta, \left\{\frac{1}{\beta_n} f : f \in \mathcal{F}\right\}, x_1^n\right) \leq \mathcal{N}_1\left(\delta \cdot \beta_n, \mathcal{F}, x_1^n\right)$$

we get for $x_1^n = (x_1, \ldots, x_n) \in \mathbb{R}^d \times \cdots \times \mathbb{R}^d$

$$\mathbf{P}\{T_{2,n} > t\} \leq 14 \sup_{x_1^n} \mathcal{N}_1\left(\frac{t}{80\beta_n}, T_{\beta_n}\mathcal{F}_n, x_1^n\right) \exp\left(-\frac{n}{5136 \cdot \beta_n^2} t\right).$$

From Lemma 2, we know that with $L_n := \max\{L_{1,n}, \ldots, L_{K_n,n}\}$ for $1/n < t < 40\beta_n$

$$\begin{aligned}&\mathcal{N}_1\left(\frac{t}{80\beta_n}, T_{\beta_n}\mathcal{F}_n, x_1^n\right) \\ &\quad\leq 3\left(\frac{6e\beta_n \cdot 80\beta_n \cdot K_n L_n}{t}\right)^{2(d+2)\left(\sum_{k=1}^{K_n} L_{k,n}\right)} \\ &\quad\leq n^{c_{11} \cdot \sum_{k=1}^{K_n} L_{k,n}}\end{aligned}$$

for some sufficient large $c_{11} > 0$. (This inequality holds also for $t \geq 40\beta_n$, since the right-hand side above does not depend on t and the covering number is decreasing in t.) Using this, we get for arbitrary $\epsilon \geq 1/n$

$$\begin{aligned}&\mathbf{E}(T_{2,n}) \\ &\quad\leq \epsilon + \int_\epsilon^\infty \mathbf{P}\{T_{2,n} > t\} dt \\ &\quad= \epsilon + 14 \cdot n^{c_{11}\left(\sum_{k=1}^{K_n} L_{k,n}\right)} \frac{5136\beta_n^2}{n} \cdot \exp\left(-\frac{n}{5136\beta_n^2}\epsilon\right)\end{aligned}$$

and this expression is minimized for

$$\epsilon = \frac{5136 \cdot \beta_n^2}{n} \log\left(14 \cdot n^{c_{11}\left(\sum_{k=1}^{K_n} L_{k,n}\right)}\right).$$

Altogether, we get

$$\mathbf{E}(T_{2,n}) \leq \frac{c_{12} \cdot \log(n)^3 \cdot \sum_{k=1}^{K_n} L_{k,n}}{n}$$

for some sufficient large constant $c_{12} > 0$, which does not depend on n, β_n, or the parameters of the estimate.

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

$$\mathbf{E}(T_{3,n}) \leq c_{13} \cdot \frac{\log(n)}{n}$$

for some large enough constant $c_{13} > 0$, and hence, we get over all

$$\mathbf{E}\left(\sum_{i=1}^{3} T_{i,n}\right) \leq \frac{c_{14} \cdot \log(n)^3 \cdot \sum_{k=1}^{K_n} L_{k,n}}{n}$$

for some sufficient large constant $c_{14} > 0$.

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

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

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

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

$$\begin{aligned}
\frac{1}{2}\cdot T_{7,n}&\le\sqrt{\mathbf{E}\left((|m_n(X_1)-Y_1|^2)^2\right)}\cdot\sqrt{\mathbf{P}(A_n)}\\
&\le\sqrt{\mathbf{E}\left((2|m_n(X_1)|^2+2|Y_1|^2)^2\right)}\\
&\quad\cdot\sqrt{n\cdot\mathbf{P}\{|Y_1|>\beta_n\}}\\
&\le\sqrt{\mathbf{E}\left(8|m_n(X_1)|^4+8|Y_1|^4\right)}\\
&\quad\cdot\sqrt{n\cdot\frac{\mathbf{E}\left(\exp(c_2\cdot|Y_1|^2)\right)}{\exp(c_2\cdot\beta_n^2)}}
\end{aligned}$$

where the last inequality follows from inequality (9). With $x\le\exp(x)$ for $x\in\mathbb{R}$, we get

$$\begin{aligned}
&\mathbf{E}\left(|Y|^4\right)\\
&\quad=\mathbf{E}\left(|Y|^2\cdot|Y|^2\right)\\
&\quad\le\mathbf{E}\left(\frac{2}{c_2}\cdot\exp\left(\frac{c_2}{2}\cdot|Y|^2\right)\cdot\frac{2}{c_2}\cdot\exp\left(\frac{c_2}{2}\cdot|Y|^2\right)\right)\\
&\quad=\frac{4}{c_2^2}\cdot\mathbf{E}\left(\exp\left(c_2\cdot|Y|^2\right)\right)
\end{aligned}$$

which is less than infinity by condition (3) of the theorem. Furthermore, $\|m_n\|_\infty$ is bounded by β_n, and therefore, the first factor is bounded by

$$c_{15}\cdot\beta_n^2=c_{16}\cdot\log(n)^2$$

for some constant $c_{16}>0$. The second factor is bounded by $1/n$, because by the assumptions of the theorem $\mathbf{E}\left(\exp\left(c_2\cdot|Y_1|^2\right)\right)$ is bounded by some constant $c_{17}<\infty$, and hence, we get

$$\begin{aligned}
\sqrt{n\cdot\frac{\mathbf{E}\left(\exp(c_2\cdot|Y_1|^2)\right)}{\exp(c_2\cdot\beta_n^2)}}&\le\sqrt{n}\cdot\frac{\sqrt{c_{17}}}{\sqrt{\exp(c_2\cdot\beta_n^2)}}\\
&\le\frac{\sqrt{n}\cdot\sqrt{c_{17}}}{\exp((c_2\cdot c_1^2\cdot\log(n)^2)/2)}.
\end{aligned}$$

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

$$T_{7,n}\le c_{18}\cdot\frac{\log(n)^2\sqrt{n}}{n^2}\le c_{19}\cdot\frac{\log(n)^2}{n}.$$

With the definition of A_n^c and $\tilde{m}_n$ defined as in (2), it follows that $T_{8,n}$ is bounded by

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

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

$$\begin{aligned}
\mathbf{E}(T_{4,n})&\le c_{19}\cdot\frac{\log(n)^2}{n}\\
&\quad+2\mathbf{E}\left(\inf_{f\in\mathcal{F}_n}\frac{1}{n}\sum_{i=1}^{n}|f(X_i)-Y_i|^2-\frac{1}{n}\sum_{i=1}^{n}|m(X_i)-Y_i|^2\right)
\end{aligned}$$

which completes the proof. □

In the sequel, we will bound

$$\inf_{f\in\mathcal{F}_n}\frac{1}{n}\sum_{i=1}^{n}|f(X_i)-Y_i|^2-\frac{1}{n}\sum_{i=1}^{n}|m(X_i)-Y_i|^2.$$

Therefore, we will use the following lemma.

Lemma 3: Let $K_n\in\mathbb{N}$ and let Π be a partition of $[a,b]^d$ consisting of K_n rectangles. Assume that $f^{\text{lin}}:[a,b]^d\to\mathbb{R}$ is a piecewise polynomial of degree $M=1$ (in each coordinate) with respect to Π and assume that f is continuous. Furthermore, let $x_1,\dots,x_n\in\mathbb{R}^d$ be n fixed points in $\mathbb{R}^d$. Then, there exist linear functions

$$f_{1,0},\dots,f_{1,2d},\dots,f_{K_n,0},\dots,f_{K_n,2d}:\mathbb{R}^d\to\mathbb{R}$$

such that

$$f^{\text{lin}}(z)=\max_{i=1,\dots,K_n}\min_{k=0,\dots,2d}f_{i,k}(z)$$

for all $z\in\{x_1,\dots,x_n\}$.

Proof: Since f^{lin} is a piecewise polynomial of degree 1, it is of the shape

$$\begin{aligned}
f^{\text{lin}}(z)&=\sum_{i=1}^{K_n}f_i^{\text{lin}}(z)\cdot I_{A_i}\\
&=\sum_{i=1}^{K_n}\left(\sum_{j=1}^{d}\alpha_{i,j}\cdot z^{(j)}+\alpha_{i,0}\right)\cdot I_{A_i}
\end{aligned}$$

for some constants $\alpha_{i,j}\in\mathbb{R}$ $(i=1,\dots,K_n, j=0,\dots,d)$, where $\Pi=\{A_1,\dots,A_{K_n}\}$ is a partition of $[a,b]^d$ and

$$A_i=I_i^{(1)}\times\ldots\times I_i^{(d)}$$

for some univariate intervals $I_i^{(j)}$ $(i=1,\dots,K_n)$. We denote the left and right endpoints of $I_i^{(j)}$ by $a_{i,j}$ and $b_{i,j}$, resp., i.e.,

$$I_i^{(j)}=[a_{i,j},b_{i,j})\quad\text{or}\quad I_i^{(j)}=[a_{i,j},b_{i,j}].$$

This choice is without restriction of any kind because f^{lin} is continuous. Now we choose for every $i \in \{1,\dots,K_n\}$

$$f_{i,0}(x) = f_i^{\text{lin}}(x) = \sum_{j=1}^{d} \alpha_{i,j} \cdot x^{(j)} + \alpha_{i,0}.$$

This implies that $f_{i,0}$ and the given piecewise polynomial f^{lin} matches on A_i for every $i = 1,\dots,K_n$. Furthermore, for $i = 1,\dots,K_n$ and $j = 1,\dots,d$, we define

$$f_{i,2j-1}(x) = f_i^{\text{lin}}(x) + (x^{(j)} - a_{i,j}) \cdot \beta_{i,j}$$

where $\beta_{i,j} \geq 0$ is such that

$$f_{i,2j-1}(z) \leq f^{\text{lin}}(z)$$

for all $z = (z^{(1)},\dots,z^{(d)}) \in \{x_1,\dots,x_n\}$ satisfying $z^{(j)} < a_{i,j}$ and

$$f_{i,2j-1}(z) \geq f^{\text{lin}}(z)$$

for all $z = (z^{(1)},\dots,z^{(d)}) \in \{x_1,\dots,x_n\}$ satisfying $z^{(j)} > a_{i,j}$. The above conditions are satisfied, if

$$\beta_{i,j} \geq \max_{k=1,\dots,n; x_k^{(j)} \neq a_{i,j}} \frac{f^{\text{lin}}(x_k) - f_i^{\text{lin}}(x_k)}{x_k^{(j)} - a_{i,j}}.$$

For $z^{(j)} = a_{i,j}$ obviously $f_{i,2j-1}(z) = f_i^{\text{lin}}(z)$.

Analogously, we choose

$$f_{i,2j}(x) = f_i^{\text{lin}}(x) - (x^{(j)} - b_{i,j}) \cdot \gamma_{i,j}$$

where $\gamma_{i,j} \geq 0$ is such that

$$f_{i,2j}(z) \geq f^{\text{lin}}(z)$$

for all $z = (z^{(1)},\dots,z^{(d)}) \in \{x_1,\dots,x_n\}$ satisfying $z^{(j)} < b_{i,j}$ and

$$f_{i,2j}(z) \leq f^{\text{lin}}(z)$$

for all $z = (z^{(1)},\dots,z^{(d)}) \in \{x_1,\dots,x_n\}$ satisfying $z^{(j)} > b_{i,j}$. In this case, the conditions from above are satisfied, if

$$\gamma_{i,j} \geq \max_{k=1,\dots,n; x_k^{(j)} \neq a_{i,j}} \frac{f_i^{\text{lin}}(x_k) - f^{\text{lin}}(x_k)}{x_k^{(j)} - b_{i,j}}.$$

From this choice of functions $f_{i,j}$ $(i = 1,\dots,K_n)$, $(j = 0,\dots,2d)$ results directly that

$$\min_{k=0,..,2d} f_{i,k}(z) = f_i^{\text{lin}}(z) = f^{\text{lin}}(z)$$

for $z \in A_i \cap \{x_1,\dots,x_n\}$ and

$$\min_{k=0,..,2d} f_{i,k}(z) \leq f^{\text{lin}}(z) \text{ for } z \in \{x_1,\dots,x_n\}$$

holds for all $i = 1,\dots,K_n$, which implies the assertion. □

Proof of Corollary 1: Lemma 3 yields

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

where $\mathcal{G}$ is the set of functions, which contains all continuous piecewise polynomials of degree 1 with respect to an arbitrary partition Π consisting of K_n rectangles. Next we increase the right-hand side above by choosing Π such that it consists of equivolume cubes. Now we can apply approximation results from spline theory; see, e.g., [25, Th. 12.8, (13.62)]. From this, the $(p, C)-$ smoothness of m and Theorem 1, we conclude for some sufficient large constant $c_{20} > 0$

$$\begin{aligned}
&\mathbf{E}\int |m_n(x)-m(x)|^2 \mu(dx) \\
&\quad \leq c_3 \cdot \frac{K_n \cdot (2d+1) \cdot \log(n)^3}{n} + c_{20} \cdot C^2 \cdot K_n^{-\frac{2p}{d}} \\
&\quad \leq c_{20} \cdot C^{\frac{2d}{2p+d}} \cdot \left(\frac{\log(n)^3}{n}\right)^{\frac{2p}{2p+d}}
\end{aligned}$$

where the last inequality results from the choice of K_n. □

Proof of Corollary 2: With the assumptions on the regression function m, the second term on the right-hand side of inequality (4) equals

$$\mathbf{E}\left(2 \inf_{f\in\mathcal{F}_n}\left(\frac{1}{n}\sum_{i=1}^{n}|f(X_i)-Y_i|^2 - \frac{1}{n}\sum_{i=1}^{n}|\overline{m}(\alpha\cdot X_i)-Y_i|^2\right)\right)$$

and with $\mathcal{F}_n^1 := \{\max_{k=1,\dots,K_n} \min_{l=1,\dots,L_k} a_{k,l} \cdot x + b_{k,l}$, for some $a_{k,l}, b_{k,l} \in \mathbb{R}\}$, this expected value is less than or equal to

$$\mathbf{E}\left(2 \inf_{h\in\mathcal{F}_n^1}\left(\frac{1}{n}\sum_{i=1}^{n}|h(\alpha\cdot X_i)-Y_i|^2 - \frac{1}{n}\sum_{i=1}^{n}|\overline{m}(\alpha\cdot X_i)-Y_i|^2\right)\right)$$

because for every function $h \in \mathcal{F}_n^1$ and every vector $\alpha \in \mathbb{R}^d$

$$f(x) = h(\alpha \cdot x) \quad (x \in \mathbb{R}^d)$$

is contained in $\mathcal{F}_n$. Together with Lemma 3, this yields

$$\begin{aligned}
&\mathbf{E}\left(2 \inf_{f\in\mathcal{F}_n}\left(\frac{1}{n}\sum_{i=1}^{n}|f(X_i)-Y_i|^2 - \frac{1}{n}\sum_{i=1}^{n}|m(X_i)-Y_i|^2\right)\right) \\
&\quad \leq \mathbf{E}\left(2 \inf_{h\in\mathcal{G}}\left(\frac{1}{n}\sum_{i=1}^{n}|h(\alpha\cdot X_i)-Y_i|^2 - \frac{1}{n}\sum_{i=1}^{n}|\overline{m}(\alpha\cdot X_i)-Y_i|^2\right)\right) \\
&\quad \leq 2 \cdot \inf_{h\in\mathcal{G}} \int |h(\alpha\cdot x)-\overline{m}(\alpha\cdot X)|^2 \mu(dx) \\
&\quad \leq 2 \cdot \inf_{h\in\mathcal{G}}\left(\max_{x\in[a,b]^d} |h(\alpha\cdot x)-\overline{m}(\alpha\cdot x)|^2\right) \\
&\quad \leq 2 \cdot \inf_{h\in\mathcal{G}}\left(\max_{x\in[\hat{a},\hat{b}]} |h(x)-\overline{m}(x)|^2\right)
\end{aligned}$$

where $\mathcal{G}$ is the set of functions from $\mathbb{R}$ to $\mathbb{R}$, which contains all piecewise polynomials of degree one with respect to a partition of $[\hat{a},\hat{b}]$ consisting of K_n intervals. Here $[\hat{a},\hat{b}]$ is chosen

such that $\alpha \cdot x \in [\hat{a}, \hat{b}]$ for $x \in [a,b]^d$. Hence, again with the approximation result from spline theory, we get as in the proof of Corollary 1 for some sufficiently large constant c_{21}

$$\mathbf{E}\left(2 \inf_{f \in \mathcal{F}_n}\left(\frac{1}{n}\sum_{i=1}^{n}|f(X_i)-Y_i|^2 - \frac{1}{n}\sum_{i=1}^{n}|m(X_i)-Y_i|^2\right)\right) \leq c_{21} \cdot C^2 \cdot K_n^{-2p}.$$

Summarizing the above results, we get by Theorem 1

$$\begin{aligned}&\mathbf{E}\int |m_n(x)-m(x)|^2 \mu(dx)\\ &\quad \leq \frac{c_3 \cdot \log(n)^3 \cdot \sum_{k=1}^{K_n} L_{k,n}}{n} + c_{21} \cdot C^2 \cdot K_n^{-2p}\\ &\quad \leq c_{22} \cdot C^{2/(2p+1)} \cdot \left(\frac{\log(n)^3}{n}\right)^{\frac{2p}{2p+1}}. \qquad \square\end{aligned}$$

ACKNOWLEDGMENT

The authors would like to thank two anonymous referees and the associate editor for several helpful comments.

REFERENCES

[1] A. M. Bagirov, "Minimization methods for one class of nonsmooth functions and calculation of semi-equilibrium prices," in *Progress in Optimization: Contribution From Australia*, A. Eberhard, Ed. *et al.* Norwell, MA: Kluwer, 1999, pp. 147–175.

[2] A. M. Bagirov, "A method for minimization of quasidifferentiable functions," *Optim. Methods Software*, vol. 17, pp. 31–60, 2002.

[3] A. M. Bagirov, C. Clausen, and M. Kohler, "An algorithm for the estimation of a regression function by continuous piecewise linear functions," *Comput. Optim. Appl.*, 2008, to be published.

[4] A. M. Bagirov and J. Ugon, "Piecewise partially separable functions and a derivative-free method for large-scale nonsmooth optimization," *J. Global Optim.*, vol. 35, pp. 163–195, 2006.

[5] A. R. Barron, "Universal approximation bounds for superpositions of a sigmoidal function," *IEEE Trans. Inf. Theory*, vol. 39, no. 3, pp. 930–944, May 1993.

[6] A. R. Barron, "Approximation and estimation bounds for neural networks," *Neural Netw.*, vol. 14, pp. 115–133, 1994.

[7] G. Beliakov and M. Kohler, "Estimation of regression functions by Lipschitz continuous functions," 2005, unpublished.

[8] L. Breiman, "Hinging hyperplanes for regression, classification and function approximation," *IEEE Trans. Inf. Theory*, vol. 39, no. 3, pp. 999–1013, May 1993.

[9] L. Breiman, J. H. Friedman, R. H. Olshen, and C. J. Stone, *Classification and Regression Trees*. Belmont, CA: Wadsworth, 1984.

[10] V. F. Demyanov and A. M. Rubinov, *Constructive Nonsmooth Analysis*. Frankfurt am Main, Germany: Peter Lang, 1995.

[11] J. Franke and M. Diagne, "Estimating market risk with neural networks," *Statist. Decision*, vol. 24, pp. 233–253, 2006.

[12] J. H. Friedman, "Multivariate adaptive regression splines (with discussion)," *Ann. Statist.*, vol. 19, pp. 1–141, 1991.

[13] L. Györfi, M. Kohler, A. Krzyżak, and H. Walk, "A distribution-free theory of nonparametric regression," in *Springer Series in Statistics*. Berlin, Germany: Springer-Verlag, 2002.

[14] W. Härdle, M. Müller, S. Sperlich, and A. Werwatz, *Nonparametric and Semiparametric Models*. New York: Springer-Verlag, 2004.

[15] M. Hamers and M. Kohler, "A bound on the expected maximal deviations of sample averages from their means," *Statist. Probab. Lett.*, vol. 62, pp. 137–144, 2003.

[16] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*. New York: Springer-Verlag, 2001.

[17] J. L. Horowitz and E. Mammen, "Rate-optimal estimation for a general class of nonparametric regression models with unknown link functions," *Ann. Statist.*, vol. 35, pp. 2589–2619, 2007.

[18] M. Kohler, "Nonparametric estimation of piecewise smooth regression functions," *Statist. Probab. Lett.*, vol. 43, pp. 49–55, 1999.

[19] M. Kohler, "Inequalities for uniform deviations of averages from expectations with applications to nonparametric regression," *J. Statist. Planning Inference*, vol. 89, pp. 1–23, 2000.

[20] M. Kohler, "Nonparametric regression with additional measurements errors in the dependent variable," *J. Statist. Planning Inference*, vol. 136, pp. 3339–3361, 2006.

[21] M. Kohler, A. Krzyżak, and H. Walk, "Rates of convergence for partitioning and nearest neighbor regression estimates with unbounded data," *J. Multivariate Anal.*, vol. 97, pp. 311–323, 2006.

[22] W. S. Lee, P. L. Bartlett, and R. C. Williamson, "Efficient agnostic learning of neural networks with bounded fan-in," *IEEE Trans. Inf. Theory*, vol. 42, no. 6, pt. 2, pp. 2118–2132, Nov. 1996.

[23] R. Mifflin, "Semismooth and semiconvex functions in constrained optimization," *SIAM J. Control Optim.*, vol. 15, pp. 957–972, 1977.

[24] P. Pucar and J. Sjöberg, "On the hinge-finding algorithm for hinging hyperplanes," *IEEE Trans. Inf. Theory*, vol. 44, no. 3, pp. 1310–1319, May 1998.

[25] L. Schumaker, *Spline Functions: Basic Theory*. New York: Wiley, 1981.

[26] C. J. Stone, "Optimal global rates of convergence for nonparametric regression," *Ann. Statist.*, vol. 10, pp. 1040–1053, 1982.

[27] C. J. Stone, "Additive regression and other nonparametric models," *Ann. Statist.*, vol. 13, pp. 689–705, 1985.

[28] C. J. Stone, "The use of polynomial splines and their tensor products in multivariate function estimation," *Ann. Statist.*, vol. 22, pp. 118–184, 1994.

[29] S. van de Geer, *Empirical Processes in M-Estimation*. Cambridge, U.K.: Cambridge Univ. Press, 2000.

Adil M. Bagirov was born on January 7, 1960, in Bilesuvar, Azerbaijan. He received the M.S. degree in applied mathematics from the Baku State University, Azerbaijan, in 1983, the Ph.D. degree in mathematical cybernetics from the Institute of Cybernetics, Azerbaijan National Academy of Sciences, Azerbaijan, in 1989, and the Ph.D. degree in optimization from the University of Ballarat, Ballarat, Vic., Australia, in 2001.

From 2001 to 2005, he was a Research Fellow at the University of Ballarat. Since 2006, he has been an Australian Research Council Research Fellow at the University of Ballarat. His main research interests are in the area of nonsmooth and global optimization and their applications in data mining and regression analysis.

Conny Clausen was born on June 8, in 1980, in Flensburg, Germany. She received a degree in mathematics and the Ph.D. degree in mathematics from Saarland University, Saarbrücken, Germany, in 2005 and 2008, respectively.

Since 2008, she has been working as IT-Consultant at beck et al. projects GmbH in Munich, Germany.

Michael Kohler was born on July 17, 1969, in Esslingen, Germany. He received degrees in computer science and mathematics and the Ph.D. degree in mathematics from the University of Stuttgart, Stuttgart, Germany, in 1995 and 1997, respectively.

In 1998, he was a Visiting Scientist at the Stanford University, Stanford, CA. From 2005 to 2007, he was the Professor of Applied Mathematics at the University of Saarbrücken, Saarbrücken, Germany. Since 2007, he has been the Professor of Mathematical Statistics at the University of Darmstadt, Darmstadt, Germany. He coauthored (with L. Györfi, A. Krzyżak, and H. Walk) the book *A Distribution-Free Theory of Nonparametric Regression* (New York: Springer-Verlag, 2002). His main research interest are in the area of nonparametric statistics, especially curve estimation and applications in mathematical finance.