Estimation of a regression function by maxima of minima of linear functions *

Adil M. Bagirov[1], Conny Clausen[2] and Michael Kohler[3]

[1] School of Information Technology and Mathematical Sciences, University of Ballarat, PO Box 663, Ballarat Victoria 3353 Australia, email: a.bagirov@ballarat.edu.au

[2] Department of Mathematics, Universität des Saarlandes, Postfach 151150, D-66041 Saarbrücken Germany, email: clausen@math.uni-sb.de

[3] Department of Mathematics, Technische Universität Darmstadt, Schloßgartenstr. 7, D-64289 Darmstadt, Germany, email: kohler@mathematik.tu-darmstadt.de

Abstract

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 behaviour of the estimates is illustrated by applying them to simulated data.

Key words and phrases: Adaptation, dimension reduction, nonparametric regression,

*Running title: *Estimation of a regression function*

Please send correspondence and proofs to: Conny Clausen, Department of Mathematics, Universität des Saarlandes, Postfach 151150, D-66041 Saarbrücken Germany, email: clausen@math.uni-sb.de

rate of convergence, single index model, L_2 error.

MSC 2000 subject classifications: Primary 62G08; secondary 62G05

1 Introduction

1.1 Scope of this paper. 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 several established methods for nonparametric regression, including regression trees like CART (cf., Breiman et al. (1984)), adaptive spline fitting like MARS (cf., Friedman (1991)) and least squares neural network estimates (cf., e.g., Chapter 11 in Hastie, Tibshirani and Friedmann (2001)). 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 maxima 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. 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.

1.2 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 the sequel 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

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

(cf. e.g., Györfi et al. (2002), 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 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), ... are independent and identically distributed random vectors. The set of data

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

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 monography Györfi et al. (2002).

1.3 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 $L_{1,n}, \dots, L_{K_n,n} \in \mathbb{N}$ be parameters of the estimate and set

$$\mathcal{F}_n = \Big\{ f : \mathbb{R}^d \to \mathbb{R} \, : \, f(x) = \max_{k=1,\dots,K_n} \min_{l=1,\dots,L_{k,n}} (a_{k,l} \cdot x + b_{k,l}) \quad (x \in \mathbb{R}^d) \\ \text{for some } a_{k,l} \in \mathbb{R}^d, b_{k,l} \in \mathbb{R} \Big\}$$

where

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

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

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

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

In Section 2 we will analyze the rate of convergence of a truncated version of this least squares estimate defined by

$$m_n(\cdot) = T_{\beta_n}(\tilde{m}_n(\cdot)), \text{ where } T_{\beta_n}(z) = \begin{cases} \beta_n & z > \beta_n, \\ z & -\beta_n \leq z \leq \beta_n, \\ -\beta_n & z < -\beta_n \end{cases}$$

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

1.4 Main results. Under a Sub-Gaussian condition on the distribution of Y and for bounded distribution of X we show that the L_2 error of the estimate achieves for (p, C)-smooth regression function with $p \leq 2$ (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 satifies in addition

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

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 curse of dimensionality.

1.5 Discussion of related results. In multivariate nonparametric regression function estimation there is a gap between theory and practice: The established estimates like 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 Barron (1993, 1994) and for CART in Kohler (1999)), but in this form the estimates cannot be computed in an application. For our estimate, a similar phenomen occurs since we need heuristics to compute it approximately in an application. The difference to 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., Bagirov (1999, 2002) and Bagirov and Udon (2006)) 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 Stone (1982) 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 (like additivity or like the assumption in the single index model) and to derive under these assumption better rates of convergence is due to Stone (1985, 1994).

We use a theorem of Lee, Bartlett and Williamson (1996) to derive our rate of convergence results. This approach is described in detail in Section 11.3 of Györfi at al. (2002). 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 monography van de Geer (2000) or Kohler (2000, 2006).

Maxima of minima of linear functions have been used in regression estimation already in Beliakov and Kohler (2005). There least squares estimates 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 maxima of minima of linear functions, where the number of minima occurring in the maxima 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 which we consider in this article 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.

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. Similar results are shown in Section 22.2 of Györfi et al. (2002). However, in contrast to the estimate defined there our newly proposed estimate can be computed in an application (which we will demonstrate in Section 3). So the main result here is to derive this good rate of convergence for an estimate which can be computed in an application.

1.6 Notations. The sets of natural numbers, natural numbers including zero, real numbers and non-negative 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 scalarproduct 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.

1.7 Outline of the paper The main theoretical result is formulated in Section 2 and proven in Section 4. In Section 3 the estimate is illustrated by applying it to simulated data.

2 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 $K_n, L_{1,n}, ..., L_{K_n,n} \in \mathbb{N}$, with $K_n \cdot \max\{L_{1,n}, ..., L_{K_n,n}\} \leq n^2$, and set $\beta_n = c_1 \cdot \log(n)$ for some constant $c_1 > 0$. Assume that the distribution of (X, Y) satifies*

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

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 Subsection 1.3

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

for some constant $c_3 > 0$ and hence also

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

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 particulary 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 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 $\alpha = (\alpha_1, ..., \alpha_d), \alpha_i \in \mathbb{N}_0, \sum_{j=1}^d \alpha_j = k$ the partial derivative*

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

exists and satisfies

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

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

Corollary 1. *Assume that the distribution of (X,Y) satifies, 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 fullfiled for some constant $c_2 > 0$ and that m is (p,C)-smooth for some $0 < p \leq 2$ and $C > 1$. Set $\beta_n = c_1 \cdot \log(n)$ for some $c_1 > 0$,*

$$K_n = \left\lceil C^{\frac{2d}{2p+d}} \cdot \left(\frac{n}{\log(n)^3} \right)^{d/(2p+d)} \right\rceil \quad \textit{and} \quad L_{k,n} = L_k = 2d + 1 \quad (k = 1, ..., K_n).$$

Then we have for the estimate m_n defined as above

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

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

Corollary 2. *Assume that the distribution of* (X, Y) *satifies, 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 fullfiled 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 some (p, C)*-smooth function* $\overline{m} : \mathbb{R} \to \mathbb{R}$ *and some* $\alpha \in \mathbb{R}^d$. *Then for the estimate* m_n *as above and with the setting* $\beta_n = c_1 \cdot \log(n)$ *for some* $c_1 > 0$,

$$K_n = \left\lceil C^{\frac{2}{2p+1}} \cdot \left(\frac{n}{\log(n)^3} \right)^{1/(2p+1)} \right\rceil \quad \text{and} \quad L_{k,n} = L_k = 3 \quad (k = 1, ..., K_n)$$

we get

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

Remark 1. It follows from Stone (1982) that under the conditions of Corollary 1 no estimate can achieve (in some Minimax sense) a rate of convergence which converges faster to zero than

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

(cf., e.g., Chapter 3 in Györfi et al. (2002)). 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 data-dependent. 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., Chapter 7 in Györfi et al. (2002)).

Theoretical results concerning splitting of the sample can be found in Hamers and Kohler (2003) and Chapter 7 in Györfi et al. (2002).

3 Application to simulated data

In our applications we choose the number of linear functions considered in the maxima and the minima 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 number of linear functions K an estimate $\tilde{m}_{n_l,K}$, 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 $\tilde{m}_{n_l,K}$. Then we choose K by minimizing this estimate with respect to K. 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

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

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

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

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 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 which is near 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 minimzation 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., Mifflin (1977)), moreover it is a smooth composition of so-called quasidifferentiable functions (see, Demyanov and Rubinov (1995) for the definition of quasidifferentiable functions). Therefore we can use the discrete gradient method from Bagirov (2002) to solve it. Furthermore, it is piecewise partially separable (see Bagirov and Ugon (2006) for the definition of such functions). We use the version of the discrete gradient method described in Bagirov and Ugon (2006) 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 Bagirov, Clausen and Kohler (2007). An implementation of the estimate in Fortran is available from the authors by request.

In Bagirov, Clausen and Kohler (2007) the estimate is also compared to various other nonparametric regression estimates. In the sequel we will only illustrate it 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 Figures 1 to 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 figures shows the true regression function together with its formula, a corresponding sample of size $n = 500$ and our estimate applied to this sample.

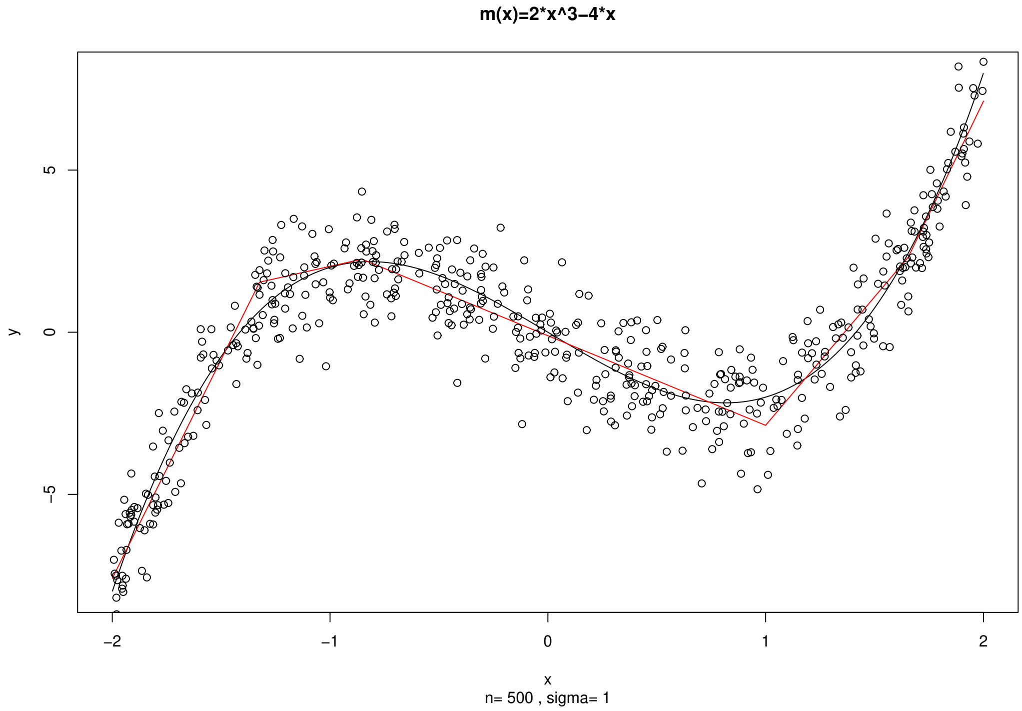

Figure 1: Simulation with the first univariate regression function.

Here the first two examples show how the maxmin-estimate looks like for 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

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

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

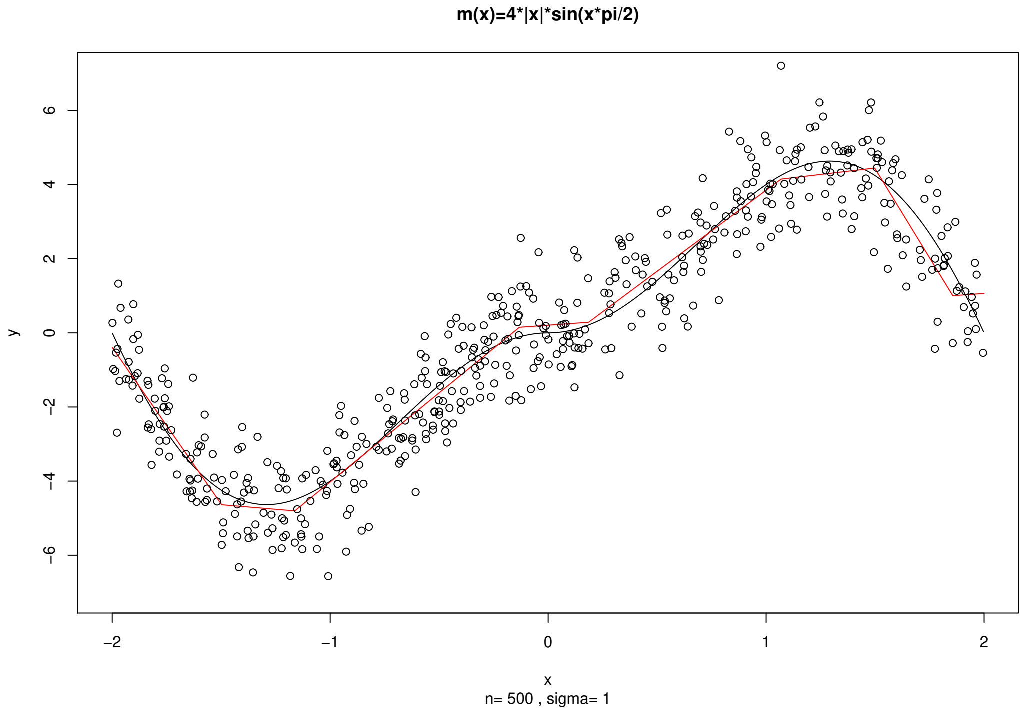

Figure 2: Simulation with the second univariate regression function.

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 $\sigma = 0.2$. Figures 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 $\sigma = 0.2$. Figures 7 show 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 behaviour of the regression function.

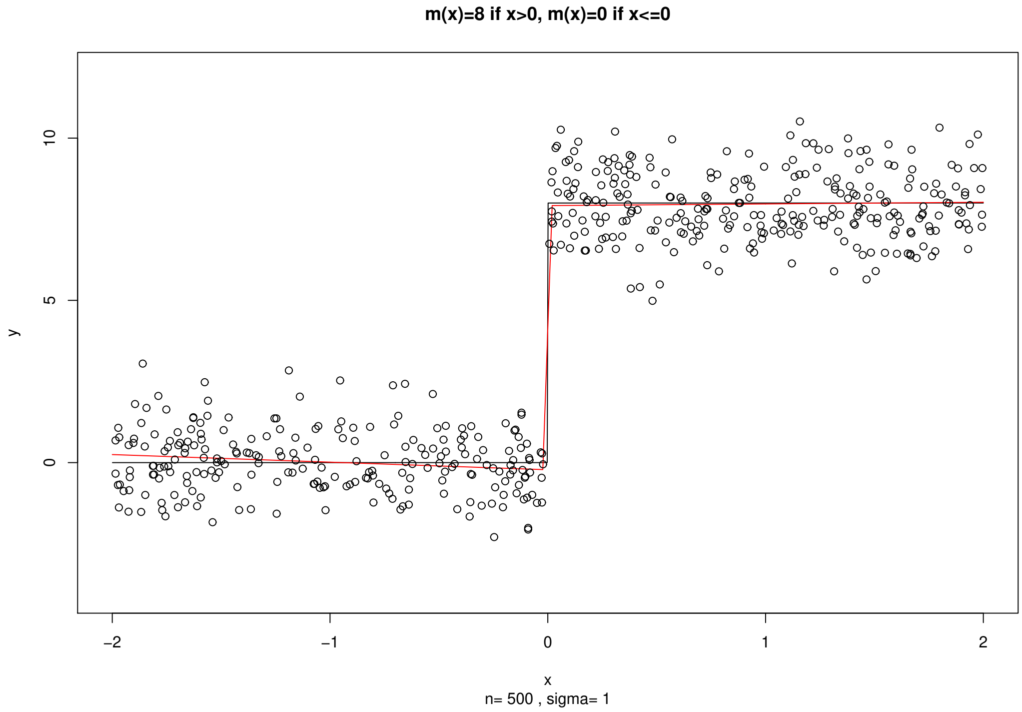

Figure 3: Simulation with the third univariate regression function.

4 Proofs

In the proofs we need the notation of covering numbers.

Definition 2. *Let $x_1, ..., x_n \in \mathbb{R}^d$ and set $x_1^n = (x_1, ..., x_n)$. Let $\mathcal{F}$ be a set of functions $f : \mathbb{R}^d \to \mathbb{R}$. A L_p-ϵ-cover of $\mathcal{F}$ on x_1^n is a finite set of functions $f_1, ..., f_k : \mathbb{R}^d \to \mathbb{R}$ 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 exist 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

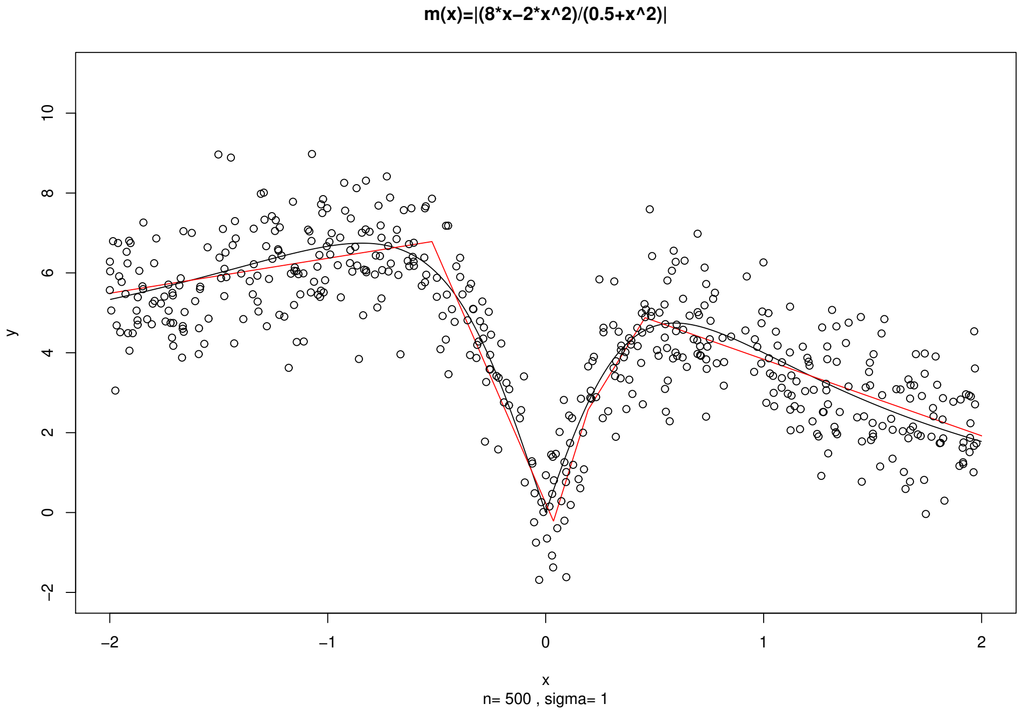

Figure 4: Simulation with the fourth univariate regression function.

$\mathcal{G}_1, \mathcal{G}_2, \ldots$ and the L_p-ϵ-covering number of their maximum

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

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*

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

and

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

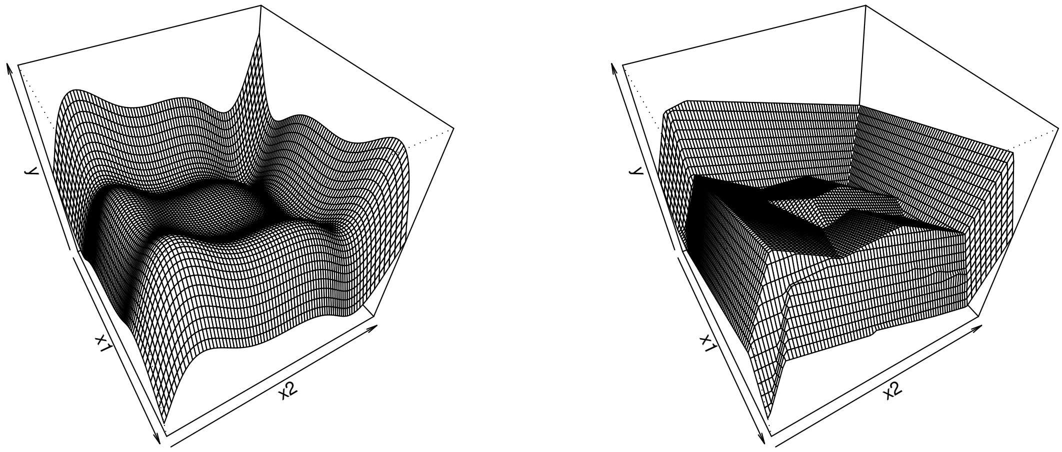

Figure 5: The bivariate regression function together with our max-min-estimate in the fifth example.

Proof. Inequaltity (6) follows from

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

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

□

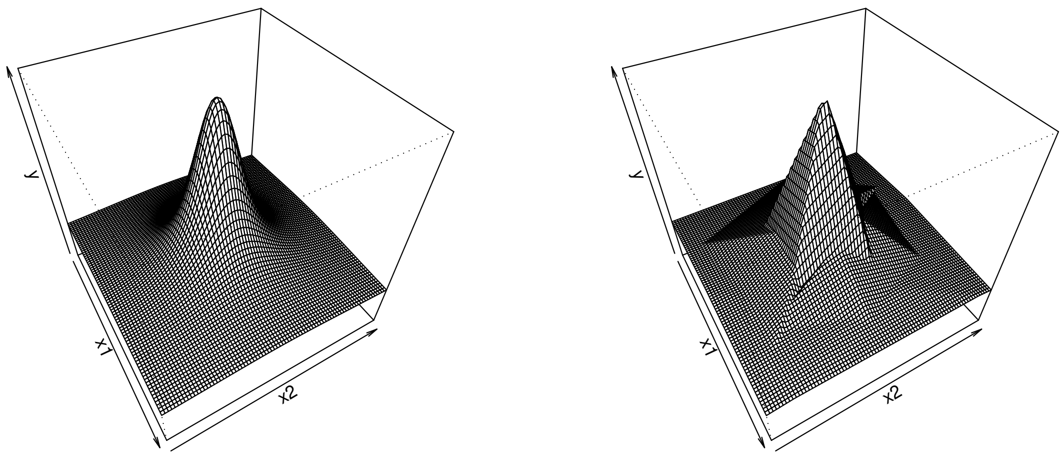

Figure 6: The bivariate regression function together with our max-min-estimate in the sixth example.

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

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

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

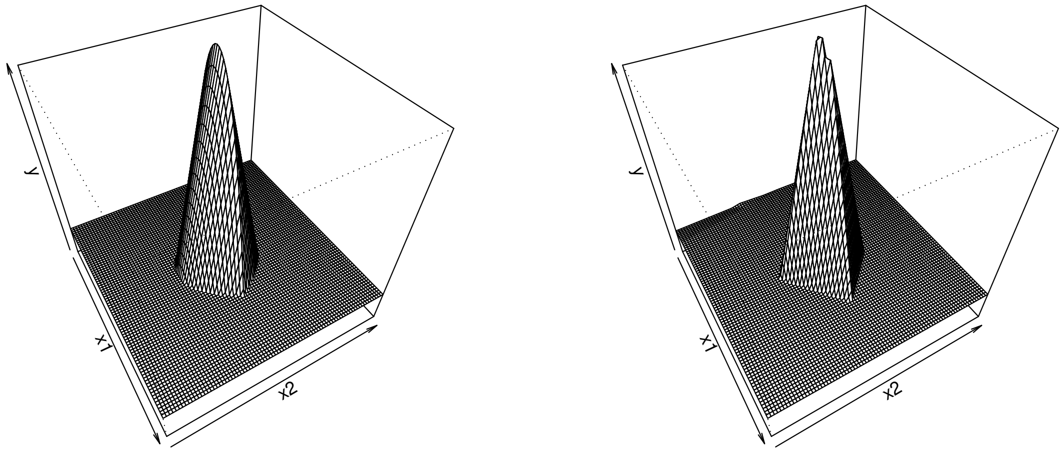

Figure 7: The bivariate regression function together with our max-min-estimate in the seventh example.

$$T_\beta \mathcal{F}_n = \Bigg\{ f : \mathbb{R}^d \to \mathbb{R} \, : \, f(x) = \max_{1 \le k \le K_n} \min_{1 \le l \le L_{k,n}} T_\beta \left(a_{k,l} \cdot x + b_{k,l} \right) \qquad \text{for some } a_{k,l} \in \mathbb{R}^d, b_{k,l} \in \mathbb{R} \Bigg\} \tag{8}$$

At the beginning we observe, that by monotonicity of the mapping $x \mapsto T_\beta x$ the equality

$$T_\beta \max_{1 \le i \le n} z_i = \max_{1 \le i \le n} T_\beta z_i \tag{9}$$

holds for real numbers $z_i \in \mathbb{R}$ $(i = 1, ..., n)$. 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 (8). 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 Györfi et al. (2002) we get

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

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:

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

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

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

With the Cauchy-Schwarz inequality and

$$I_{\{|Y|>\beta_n\}} \leq \frac{\exp(c_2/2 \cdot |Y|^2)}{\exp(c_2/2 \cdot \beta_n^2)} \tag{10}$$

it follows

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

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

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

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

$$\mathbf{E}\left(\frac{2}{c_2} \cdot \exp\left(c_2/2 \cdot |Y|^2\right) \cdot \exp(c_2/2 \cdot |Y|^2)\right) \leq \mathbf{E}\left(\frac{2}{c_2} \cdot \exp\left(c_2 \cdot |Y|^2\right)\right) \leq c_4$$

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

$$\mathbf{E}(|Y|^2) \leq \mathbf{E}(1/c_2 \cdot \exp(c_2 \cdot |Y|^2) \leq c_5 < \infty \tag{11}$$

which follows again as above. With the setting $\beta_n = c_1 \cdot \log(n)$ it follows for some constants $c_6, c_7 > 0$

$$|T_{5,n}| \quad \leq \quad \sqrt{c_4} \cdot \exp\left(-c_6 \cdot \log(n)^2\right) \cdot \sqrt{(18 \cdot c_1 \cdot \log(n)^2 + c_5)} \leq c_7 \cdot \frac{\log(n)}{n}.$$

From the Cauchy-Schwarz inequality we get

$$\begin{aligned} T_{6,n} \quad &\leq \quad \sqrt{2\mathbf{E}\Big\{|(m(X) - m_{\beta_n}(X))|^2\Big\} + 2\mathbf{E}\Big\{|(T_{\beta_n}Y - Y)|^2\Big\}} \\ &\quad \cdot \sqrt{\mathbf{E}\Big\{\Big|m(X) + m_{\beta_n}(X) - Y - T_{\beta_n}Y\Big|^2\Big\}}, \end{aligned}$$

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 $c_8 > 0$

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

Next we consider the first term. With the inequality from Jensen it follows

$$\mathbf{E}\Big\{|m(X) - m_{\beta_n}(X)|^2\Big\} \quad \leq \quad \mathbf{E}\left\{\mathbf{E}\Big(|Y - T_{\beta_n}Y|^2\Big|X\Big)\right\} = \mathbf{E}\Big\{|Y - T_{\beta_n}Y|^2\Big\}.$$

Hence we get

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

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

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

Thus with Theorem 11.4 in Györfi et al. (2002) 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 $x_1^n = (x_1, ..., x_n) \in \mathbb{R}^d \times ... \times \mathbb{R}^d$

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

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

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

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 $\epsilon \geq 1/n$

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

and this expression is minimized for

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

Alltogether we get

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

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 $c_{14} > 0$.

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

$$\begin{aligned}
\mathbf{E}(T_{4,n}) &\leq 2 \cdot \mathbf{E}\left(\frac{1}{n}\sum_{i=1}^{n} |m_n(X_i) - Y_i|^2 \cdot I_{A_n}\right) \\
&\quad +2 \cdot \mathbf{E}\left(\frac{1}{n}\sum_{i=1}^{n} |m_n(X_i) - Y_i|^2 \cdot I_{A_n^c} - \frac{1}{n}\sum_{i=1}^{n} |m(X_i) - Y_i|^2\right) \\
&= 2 \cdot \mathbf{E}\left(|m_n(X_1) - Y_1|^2 \cdot I_{A_n}\right) \\
&\quad +2 \cdot \mathbf{E}\left(\frac{1}{n}\sum_{i=1}^{n} |m_n(X_i) - Y_i|^2 \cdot I_{A_n^c} - \frac{1}{n}\sum_{i=1}^{n} |m(X_i) - Y_i|^2\right) \\
&= T_{7,n} + T_{8,n}.
\end{aligned}$$

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

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

where the last inequality follows from inequality (10). With $x \leq \exp(x)$ for $x \in \mathbb{R}$ we get

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

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

$$\sqrt{n \cdot \frac{\mathbf{E}\left(\exp(c_2 \cdot |Y_1|^2)\right)}{\exp(c_2 \cdot \beta_n^2)}} \quad \leq \quad \sqrt{n} \cdot \frac{\sqrt{c_{17}}}{\sqrt{\exp(c_2 \cdot \beta_n^2)}} \leq \frac{\sqrt{n} \cdot \sqrt{c_{17}}}{\exp((c_2 \cdot c_1^2 \cdot \log(n)^2)/2)}.$$

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

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

With the definition of A_n^c and $\tilde{m}_n$ defined as in (2) it follows

$$\begin{aligned}
T_{8,n} \quad &\leq \quad 2 \cdot \mathbf{E}\left(\frac{1}{n}\sum_{i=1}^n |\tilde{m}_n(X_i) - Y_i|^2 \cdot I_{A_n^c} - \frac{1}{n}\sum_{i=1}^n |m(X_i) - Y_i|^2\right) \\
&\leq \quad 2 \cdot \mathbf{E}\left(\frac{1}{n}\sum_{i=1}^n |\tilde{m}_n(X_i) - Y_i|^2 - \frac{1}{n}\sum_{i=1}^n |m(X_i) - Y_i|^2\right) \\
&\leq \quad 2 \cdot \mathbf{E}\left(\inf_{f \in \mathcal{F}_n} \frac{1}{n}\sum_{i=1}^n |f(X_i) - Y_i|^2 - \frac{1}{n}\sum_{i=1}^n |m(X_i) - Y_i|^2\right),
\end{aligned}$$

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

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

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

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

such that

$$f^{lin}(z) = \max_{i=1,...,K_n} \min_{k=0,..,2d} f_{i,k}(z) \qquad \text{for all } z \in \{x_1, ..., x_n\}.$$

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

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

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

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

for some univariate intervals $I_i^{(j)}$ ($i = 1, \ldots, K_n$). We denote the left and the right endpoint 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 $i \in \{1, ..., K_n\}$

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

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

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

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

$$f_{i,2j-1}(z) \leq f^{lin}(z) \text{ for all } z = (z^{(1)}, ..., z^{(d)}) \in \{x_1, ..., x_n\} \text{ satisfying } z^{(j)} < a_{i,j}$$

and

$$f_{i,2j-1}(z) \geq f^{lin}(z) \text{ for all } z = (z^{(1)}, ..., z^{(d)}) \in \{x_1, ..., x_n\} \text{ satisfying } z^{(j)} > a_{i,j}.$$

The above conditions are satisfied, if

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

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

Analogously we choose

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

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

$$f_{i,2j}(z) \geq f^{lin}(z) \text{ for all } z = (z^{(1)}, ..., z^{(d)}) \in \{x_1, ..., x_n\} \text{ satisfying } z^{(j)} < b_{i,j}$$

and

$$f_{i,2j}(z) \leq f^{lin}(z) \text{ for all } z = (z^{(1)}, ..., z^{(d)}) \in \{x_1, ..., x_n\} \text{ satisfying } z^{(j)} > b_{i,j}.$$

In this case the conditions from above are satisfied, if

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

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

$$\min_{k=0,..,2d} f_{i,k}(z) \begin{cases} = f_i^{lin}(z) = f^{lin}(z) & \text{for } z \in A_i \cap \{x_1, ..., x_n\} \\ \leq f^{lin}(z) & \text{for } z \in \{x_1, ..., x_n\} \end{cases}$$

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

Proof of Corollary 1. Lemma 3 yields

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

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

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 rectangulars. 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., Schumaker (1981), Theorem 12.8 and (13.62). From this, the $(p,C)-$ smoothness of m and Theorem 1 we conclude for some sufficient large constant $c_{20}>0$

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

where the last inequaltity results from the choice of K_n. $\square$

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 $\mathcal{F}_n^1:=\{\max_{k=1,\dots,K_n}\min_{l=1,\dots,L_k}a_{k,l}\cdot x+b_{k,l},\text{ for some }a_{k,l},b_{k,l}\in\mathbb{R}\}$ 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 $\alpha\in\mathbb{R}^d$

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

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

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

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

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}

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

Summarizing the above results we get by Theorem 1

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

□

References

[1] Bagirov, A. M. (1999). Minimization methods for one class of nonsmooth functions and calculation of semi-equilibrium prices. In: A. Eberhard et al. (eds.) *Progress in Optimization: Contribution from Australia*, Kluwer Academic Publishers, 1999, pp. 147-175.

[2] Bagirov, A. M. (2002). A method for minimization of quasidifferentiable functions. *Optimization Methods and Software* **17**, pp. 31–60.

[3] Bagirov, A. M., Clausen, C., and Kohler, M. (2007). An algorithm for the estimation of a regression function by continuous piecewise linear functions. Submitted for publication.

[4] Bagirov, A. M., and Ugon, J. (2006). Piecewise partially separable functions and a derivative-free method for large-scale nonsmooth optimization. *Journal of Global Optimization* **35**, pp. 163-195.

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

[6] Barron, A. R. (1994). Approximation and estimation bounds for neural networks. *Neural Networks* **14**, pp. 115-133.

[7] Beliakov, G., and Kohler, M. (2005). Estimation of regression functions by Lipschitz continuous functions. Submitted for publication.

[8] Breiman, L., Friedman, J. H., Olshen, R. H. and Stone, C. J. (1984). *Classification and regression trees.* Wadsworth, Belmont, CA.

[9] Demyanov, V.F., and Rubinov, A.M. (1995). *Constructive Nonsmooth Analysis.* Peter Lang, Frankfurt am Main, 1995.

[10] Friedman, J. H. (1991). Multivariate adaptive regression splines (with discussion). Annals of Statistics **19**, pp. 1-141.

[11] Györfi, L., Kohler, M., Krzyżak, A., and Walk, H. (2002). *A Distribution-Free Theory of Nonparametric Regression.* Springer Series in Statistics, Springer.

[12] Hamers, M. and Kohler, M. (2003). A bound on the expected maximal deviations of sample averages from their means. *Statistics & Probability Letters* **62**, pp. 137–144.

[13] Hastie, T., Tibshirani, R. and Friedman, J. (2001). *The elements of statistical learning.* Springer-Verlag, New York.

[14] Kohler, M. (1999). Nonparametric estimation of piecewise smooth regression functions. *Statistics & Probability Letters* **43**, pp. 49–55.

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

[16] Kohler, M. (2006). Nonparametric regression with additional measurements errors in the dependent variable. *Journal of Statistical Planning and Inference* **136**, pp. 3339-3361.

[17] Lee, W. S., Bartlett, P. L., Williamson, R. C. (1996). Efficient agnostic learning of neural networks with bounded fan–in. *IEEE Trans. Inform. Theory* **42**, pp. 2118–2132.

[18] Mifflin, R. (1977). Semismooth and semiconvex functions in constrained optimization. *SIAM Journal on Control and Optimization* **15**, pp. 957-972.

[19] Schumaker, L., 1981. *Spline functions: Basic Theory.* Wiley, New York.

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

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

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

[23] van de Geer, S. (2000). *Empirical Processes in M–estimation.* Cambridge University Press.