# Estimation of regression functions by Lipschitz continuous functions

Gleb Beliakov[1] and Michael Kohler[2]
[1] *School of Information Technology, Deakin University, 221 Burwood Hwy, Burwood, 3125, Australia, email: gleb@deakin.edu.au*
[2] *Department of Mathematics, University of Stuttgart, Pfaffenwaldring 57, D-70569 Stuttgart, Germany, email: kohler@mathematik.uni-stuttgart.de*

January 21, 2005

**Abstract**

We consider the problem of estimation of a regression function given a sample of the underlying distribution. Least squares estimates are defined by minimization of the empirical $L_2$ risk over the set of all bounded Lipschitz continuous functions with Lipschitz constant bounded by some parameter $M_n$ of the estimate. Under suitable conditions on $M_n$ it is shown that the estimate is consistent with respect to all distributions where the predictor variable is bounded. In addition, in case that the regression function is Hoelder continuous, results concerning the rate of convergence of the estimate are derived. In particular we show that if the regression function is additive and the (univariate) functions in the sum representing the regression function are Hoelder continuous, then suitably defined estimates achieve the optimal rate of convergence for estimating univariate Hoelder continuous regression functions.

Based on new results on Lipschitz interpolation a practical algorithm for computing the estimate is derived. It is combined with methods for data-dependent choice of the parameter of the estimate and the resulting estimate is illustrated by applying it to simulated data.

**Key words:** Scattered data approximation, nonparametric regression, rate of convergence, regression estimation, Lipschitz approximation, smoothing.

**AMS Classification:** 62G05, 62G08, 65D10,41A29.

# 1 Introduction

This paper considers the problem of estimating a regression function given a sample of the underlying distribution. One way to construct a regression estimate is to use the principle of least squares to choose a function from a given class of functions. Usually in an application there is little or no a priori information concerning the regression function. This makes the proper choice of the class of functions, from which the estimate will be chosen, difficult. Therefore quite flexible classes of functions are necessary here, and often they are chosen such that they are getting more and more complex with increasing sample size (method of sieves). Popular classes of functions include piecewise constant functions, polynomial splines, neural networks and radial basis functions. Corresponding regression estimates are described, e.g., in Breiman et al. (1984), Friedman (1991), Lee, Bartlett and Williamson (1996) and Krzyżak and Linder (1998).

The choice of the class of functions is difficult because of the following three reasons: Firstly, the resulting estimate will be contained in this class of functions and can therefore approximate the unknown regression function not better than the best function in this class. As a consequence the class of functions should be as complex as possible. But secondly, the class of functions is not allowed to be too complex, because otherwise the least squares estimate will be adapted very well to the given sample but will not be able to predict well new data generated from the same distribution (problem of overfitting). And thirdly, the class of functions must be such that the least squares estimate can be computed in practice. If the class of functions is a finite dimensional linear vector space (where the dimension might depend on the sample size), then the second and third condition are satisfied. However, in practice such classes of functions are often too unflexible to approximate the unknown regression function well. So often the class of functions is chosen not as a finite dimensional linear vector space, which makes the computation of the estimate difficult. E.g., for neural networks the resulting minimization problem is solved only approximately using steepest decent, while for piecewise constant functions (which are used for so-called regression trees) the data is used to choose a linear subspace of the set of functions to which the least squares principle is then applied. Alternatively, a stepwise approach can be applied in which the underlying minimization problem is approximately solved, e.g. to compute multivariate regression splines as in Friedman (1991).

The above described function classes lead to estimates for which theoretical results (e.g. concerning the rate of convergence of the estimate) can be derived and of which simplified versions can be implemented and lead to nice results in applications. However, the problem is that there is a gap between theory and application, because the estimates which are succesfully applied to simulated data are not the same as the one analyzed theoretically.

In this paper we consider a rather complex and highly flexible class of functions, namely the class of all functions which are bounded and Lipschitz continuous with respect to some given bound on the Lipschitz constant. Here the bound on the Lipschitz constant enables us to control the complexity of the functions space and hence helps us to avoid overfitting. We show that the resulting least squares estimates are consistent (with respect to the $L_2$ error) for all distributions where the predictor variable is bounded. In case that the regression function is Hoelder continuous, we show that the $L_2$ error of the estimate achieves the corresponding optimal rate of convergence for dimension $d$ of the predictor variable equal to one and achieves for $d = 2$ an up to a logarithmic factor optimal rate of convergence. Furthermore we show that a suitably modified estimate achieves the optimal rate of convergence for estimating univariate Hoelder continuous functions for any dimension of $X$ provided the regression function is additive and the functions in the sum representing the regression function are Hoelder continuous. In the second part of the paper we describe how the estimate can be implemented by using Lipschitz interpolation. Here the estimate is also combined with methods to choose a data-dependent bound on the Lipschitz constant. We illustrate the estimate by applying it to several simulated data sets.

The theoretical results are presented in Section 2. In Section 3 we describe the implementation of the estimate. In Section 4 we analyze the case of estimates satisfying some additional restrictions, such as monotonicity and additivity. Application of the estimates to simulated data are contained in Section 5. The proofs are given in Section 6.

# 2 Definition and theoretical analysis of the estimate

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*

$$\mathbf{E}\{|f(X)-Y|^2\}. \tag{1}$$

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 $\mu$. Then

$$\begin{aligned}\mathbf{E}\{|f(X)-Y|^2\} &= \mathbf{E}\{((f(X)-m(X))+(m(X)-Y))^2\}\\ &= \mathbf{E}\{|f(X)-m(X)|^2\}+\mathbf{E}\{|m(X)-Y|^2\}\\ &= \mathbf{E}\{|m(X)-Y|^2\}+\int |f(x)-m(x)|^2\mu(dx). \end{aligned}\tag{2}$$

Here the second equation follows from

$$\begin{aligned}&\mathbf{E}\{(f(X)-m(X))\cdot(m(X)-Y)\}\\ &=\mathbf{E}\{(f(X)-m(X))\cdot\mathbf{E}\{(m(X)-Y)|X\}\}=0.\end{aligned}$$

Since the integral on the right-hand side of (2) is always nonnegative, (2) implies that the regression function is the optimal predictor in view of minimization of the $L_2$ risk:

$$\mathbf{E}\{|m(X)-Y|^2\}=\min_{f:\mathbb{R}^d\to\mathbb{R}}\mathbf{E}\{|f(X)-Y|^2\}. \tag{3}$$

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*

$$\int |f(x)-m(x)|^2\mu(dx) \tag{4}$$

is small. This motivates to measure the error caused by using a function $f$ instead of the regression function by the $L_2$ error (4).

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 $(X,Y)$, $(X_1,Y_1)$, $(X_2,Y_2)$, $\dots$ 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).

The regression function minimizes the $L_2$ risk (1) over the set of all measurable functions, so in principle it can be computed by solving a minimization problem. However, in an application the term to be minimized is unknown, because it depends on the unknown distribution of $(X, Y)$. For least squares estimates the given data is used to estimate the $L_2$ risk by the so-called empirical $L_2$ risk

$$\frac{1}{n} \sum_{i=1}^{n} |f(X_i) - Y_i|^2 \tag{5}$$

and the regression estimate is defined by minimizing (5). Minimization of (5) with respect to all measurable functions (as in (3)) leads to an estimate, which usually (at least if the values of $X_1, \ldots, X_n$ are distinct) interpolates the given data. Obviously, such an estimate is not a reasonable estimate for $m(x) = \mathbf{E}\{Y|X = x\}$. In order to avoid this so-called overfitting, for least squares estimates first a class $\mathcal{F}_n$ of functions $f : \mathbb{R}^d \to \mathbb{R}$ is chosen and then the estimate is defined by minimizing the empirical $L_2$ risk over $\mathcal{F}_n$, i.e.,

$$m_n(\cdot) = \arg \min_{f \in \mathcal{F}_n} \frac{1}{n} \sum_{i=1}^{n} |f(X_i) - Y_i|^2. \tag{6}$$

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

In the sequel we will choose $\mathcal{F}_n$ as a subset of the space of uniformly continuous functions with the supremum norm, $C(\mathcal{X}), \mathcal{X} \subset \mathbb{R}^d$. Here we assume that $\mathcal{X}$ is compact and that $X$ takes on with probability one only values in $\mathcal{X}$.

**Definition 1** *The function $f : \mathcal{X} \to \mathbb{R}$ is called Lipschitz-continuous in $\mathcal{X}$, if there exists a number $M$: $\forall x, y \in \mathcal{X} : |f(x) - f(y)| \leq M||x - y||$, where $\|x\|$ is the Euclidean norm of $x \in \mathbb{R}^d$. We call the smallest such number the (smallest) Lipschitz constant of $f$ in the norm $|| \cdot ||$.*

We denote the class of Lipschitz functions $f : \mathcal{X} \to \mathbb{R}$ with the smallest Lipschitz constant bounded by $M$ by $Lip(M)$, and set

$$\mathcal{F}_n = \{f \in Lip(M_n) \, : \, \|f\|_\infty \leq \beta_n\} \, ,$$

where $M_n, \beta_n \in \mathbb{R}_+$ are parameters of the estimate and $\|f\|_\infty = \sup_{x \in \mathcal{X}} |f(x)|$ is the supremum norm of $f$.

Our first theorem shows that under suitable conditions on $M_n$ and $\beta_n$ the estimate is consistent for all distributions of $(X, Y)$ with $X \in \mathcal{X}$ *a.s.* and $\mathbf{E}Y^2 < \infty$. Related, so-called universal consistency results, can be found, e.g., in Stone (1977), Devroye and Wagner (1980), Devroye et al. (1994), Lugosi and Zeger (1995), Lugosi and Nobel (1996) and Krzyżyk and Linder (1998).

**Theorem 1** *Let $m_n$ be the least squares estimate defined above. Assume that the parameters are chosen such that*

$$\begin{aligned}
&\beta_n \to \infty \quad (n \to \infty) \\
&\frac{\beta_n^4}{n^{1-\delta}} \to 0 \quad (n \to \infty) \quad \textit{for some } \delta > 0, \\
&M_n \to \infty \quad (n \to \infty), \\
&\frac{M_n^d \beta_n^{d+4} \log(n)}{n} \to 0 \quad (n \to \infty).
\end{aligned}$$

*Then the least squares estimate $m_n$ satisfies*

$$\int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx) \to 0 \quad a.s.$$

*for every distribution of $(X, Y)$ with $X \in \mathcal{X}$ a.s. for some compact set $\mathcal{X} \subset \mathbb{R}^d$ and $\mathbf{E}Y^2 < \infty$.*

Next we study the rate of convergence of the estimate. Here it is necessary to impose smoothness conditions on the regression function in order to get nontrivial results (cf., e.g., Chapter 3 in Györfi et al. (2002)).

**Theorem 2** *Assume $X \in \mathcal{X}$ a.s. for some compact set $\mathcal{X} \subset \mathbb{R}^d$, $m$ bounded in absolute value by some constant $L > 0$, $m$ Hoelder continuous, i.e.,*

$$|m(x) - m(z)| \leq M \cdot \|x - z\|^p \quad (x, z \in \mathcal{X})$$

*for some $M > 0$ and $0 < p \leq 1$, and $Y - m(X)$ uniformly Sub-Gaussian, i.e.,*

$$K^2 \mathbf{E}\left\{e^{(Y-m(X))^2/K^2} - 1 \middle| X\right\} \leq \sigma_0^2 \quad \textit{almost surely}$$

*for some $K, \sigma_0 > 0$. Set $M_n = M^{(d+2)/(2p+d)} n^{(1-p)/(2p+d)}$, $\beta_n = L$, and define the estimate $m_n$ as above. Then we have for $d = 1$*

$$\mathbf{P}\left\{\int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx) > c_1 \cdot M^{2/(2p+1)} n^{-2p/(2p+1)}\right\} \to 0$$

*$(n \to \infty)$ for some constant $c_1$ which does not depend on $n$ or $M$, and for $d > 1$ we have*

$$\mathbf{P}\left\{\int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx) > c_2 \cdot (\log(n))^{\frac{3}{d}} M^{\frac{d+2}{2p+d}} n^{\frac{-2p}{2p+d} \cdot \frac{d+2}{2d}}\right\} \to 0$$

*$(n \to \infty)$ for some constant $c_2$ which does not depend on $n$ or $M$.*

It follows from Stone (1982) that under the conditions of Theorem 2 no estimate can achieve (in some Minimax sense) a rate of convergence which converges faster to zero than

$$M^{\frac{2d}{2p+d}} \cdot n^{-\frac{2p}{2p+d}}.$$

From this result we see that the rate of convergence in Theorem 2 is optimal for $d = 1$, up to a logarithmic factor optimal for $d = 2$, but is not optimal for $d > 2$.

For large $d$, the above rate of convergence results are not satisfactory, since then $2p/(2p + d)$ is close to zero. The only way to circumvent this so-called curse of dimensionality is to impose additional assumptions on the underlying distribution in order to be able to prove better rates of convergence. It was shown in Stone (1985) that suitably defined regression estimates achieve the univariate rate of convergence provided the regression function is additive, i.e.,

$$m(x_1, \dots, x_d) = m_1(x_1) + \cdots + m_d(x_d) \quad (x_1, \dots, x_d \in \mathbb{R})$$

for some univariate functions $m_1, \ldots, m_d : \mathbb{R} \to \mathbb{R}$. For related results, see, e.g., Chen (1991), Stone (1994) and Kohler (1998).

In the next theorem we show that our estimate has this property, too, provided it is defined by minimizing the empirical $L_2$ risk with respect to additive Lipschitz-continuous functions.

**Theorem 3** *Assume $X \in [0,1]^d$ a.s., $m$ is additive, i.e.,*

$$m(x_1, \ldots, x_d) = m_1(x_1) + \cdots + m_d(x_d) \quad (x_1, \ldots, x_d \in [0,1])$$

*for some univariate functions $m_1, \ldots, m_d : \mathbb{R} \to \mathbb{R}$, $m_j$ bounded in absolute value by some constant $L > 0$ $(j = 1, \ldots, d)$, $m_j$ Hoelder continuous, i.e.,*

$$|m_j(x) - m_j(z)| \leq M \cdot |x - z|^p \quad (x, z \in [0,1])$$

*for some $M > 0$ and $0 < p \leq 1$ $(j = 1, \ldots, d)$, and $Y - m(X)$ uniformly Sub-Gaussian, i.e.,*

$$K^2 \mathbf{E} \left\{ e^{(Y-m(X))^2/K^2} - 1 \middle| X \right\} \leq \sigma_0^2 \quad \textit{almost surely}$$

*for some $K, \sigma_0 > 0$. Set $M_n = M^{3/(2p+1)} n^{(1-p)/(2p+1)}$ and define the estimate $m_n$ by minimizing the empirical $L_2$ risk with respect to all functions $f : \mathbb{R}^d \to \mathbb{R}$ which satisfy*

$$f(x_1, \ldots, x_d) = f_1(x_1) + \cdots + f_d(x_d) \quad (x_1, \ldots, x_d \in \mathbb{R})$$

*for univariate functions $f_1, \ldots, f_d \in Lip(M_n)$ which are bounded in absolute value by $L$. Then we have for any $d \geq 1$*

$$\mathbf{P} \left\{ \int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx) > c_3 \cdot M^{2/(2p+1)} n^{-2p/(2p+1)} \right\} \to 0 \quad (n \to \infty)$$

*for some constant $c_3$ which does not depend on $n$ or $M$.*

**Remark 1.** It is well-known in the theory of random design regression, that least squares estimate need some kind of boundedness assumption in order to be consistent (cf., e.g., Chapter 10 in Györfi et al. (2002)). However we do not know whether for the above function class the bound on the supremum norm of the functions is really necessary in order to prove, e.g., consistency.
**Remark 2.** It follows from Theorem 7.1 in Györfi et al. (2002) (see also Hamers and Kohler (2003)) that if in the definition of the above estimates the constant $M_n$ is chosen data-dependent by splitting the sample then the corresponding estimates achieve the same rate of convergence as the estimates in Theorem 2 and 3 above.

# 3 Computation of the estimate

In this section we show how the problem (6) can be practically solved, and the estimate $m_n(\cdot)$ is computed. Consider first the case of a given Lipschitz constant $M_n$, i.e. a fixed class $Lip(M_n)$. We address the problem of estimating $M_n$ from the data $\mathcal{D}_n$ later in this section.

## 3.1 Fixed Lipschitz constant

We call the data set $\mathcal{D}_n$ compatible with the class $Lip(M_n)$ if there exists a function $f \in Lip(M_n)$ which interpolates the data.

**Proposition 1** *The data set $\mathcal{D}_n$ is compatible with the class $Lip(M_n)$ if and only if the following conditions hold*

$$\forall i, j \in \{1, \ldots, n\} : \; |Y_i - Y_j| \leq M_n ||X_i - X_j|| \tag{7}$$

**Proof.** Assume the above condition holds. Let us define the function

$$f(x) = \max_i (Y_i - M_n ||x - X_i||).$$

This function is a pointwise maximum of functions

$$f_i(x) = Y_i - M_n ||x - X_i||,$$

each $f_i \in Lip(M_n)$, and hence $f \in Lip(M_n)$. On the other hand,

$$f_i(X_i) = Y_i \geq Y_j - M_n ||X_j - X_i||$$

(which follows from (7)), and therefore $f(X_i) = f_i(X_i) = Y_i$. Hence $f$ interpolates the data.

Conversely, assume that for some $i, j$

$$|Y_i - Y_j| > M_n ||X_i - X_j||.$$

Then any interpolant $f$ will violate the Lipschitz condition implied in $Lip(M_n)$

$$|f(X_i) - f(X_j)| > M_n ||X_i - X_j||.$$

$\square$

If $\mathcal{D}_n$ is compatible with $Lip(M_n)$ we can construct the upper and the lower estimates as

$$\begin{aligned} m_n^{upper}(x) &= \min_i (Y_i + M_n ||x - X_i||), \\ m_n^{lower}(x) &= \max_i (Y_i - M_n ||x - X_i||). \end{aligned} \tag{8}$$

The two estimates may coincide, in which case $Lip(M_n)$ contains exactly one function interpolating the data. Otherwise there are many estimates $m_n$ interpolating the data, whose values are bounded by

$$m_n^{lower}(x) \le m_n(x) \le m_n^{upper}(x).$$

For instance a linear interpolating spline is one such estimate. The optimal estimate, which minimizes the largest possible pointwise error (assuming $m \in Lip(M_n)$), see Gaffney and Powell (1976), Micchelli and Rivlin (1984),

$$\max_{x \in \mathcal{X}} |m(x) - m_n(x)|,$$

is given as

$$m_n(x) = \frac{1}{2}(m_n^{lower}(x) + m_n^{upper}(x)), \forall x \in \mathcal{X}. \tag{9}$$

As we mentioned earlier, the case of the data set compatible with the class $Lip(M_n)$ is of little interest, as we want to avoid overfitting noisy data. However we will use equations (8) and (9) as a tool to compute an optimal estimate once the data have been smoothened.

Thus we look at the case of $\mathcal{D}_n$ being incompatible with the class $Lip(M_n)$. We approach the solution to (6) in two steps. First let us find the set of values

$$\hat{\mathcal{D}}_n = \left\{ (X_1, \hat{Y}_1), \ldots, (X_n, \hat{Y}_n) \right\}$$

compatible with $\mathcal{F}_n$, such that the expression

$$\frac{1}{n} \sum_{i=1}^{n} |\hat{Y}_i - Y_i|^2$$

is minimized. Then define the estimate as (9), where

$$\begin{aligned} m_n^{upper}(x) &= \min_i (\hat{Y}_i + M_n ||x - X_i||), \\ m_n^{lower}(x) &= \max_i (\hat{Y}_i - M_n ||x - X_i||). \end{aligned} \tag{10}$$

Note that $m_n, m_n^{upper}, m_n^{lower}$ all belong to $Lip(M_n)$. In order to impose the restriction $||f||_\infty \leq \beta_n$, which is part of the definition of $\mathcal{F}_n$, we will demonstrate the range property of $m_n$. It will be very handy when we impose the bounds on the estimate by restricting the values of $\hat{Y}_i$ in Eq. (12).

**Proposition 2** *The range of* $m_n(x)$ *is* $[\min \hat{Y}_i, \max \hat{Y}_i]$.

**Proof.** Consider $2m_n(x)$ given by

$$2m_n(x) = \min_i(\hat{Y}_i + M_n||x - X_i||) + \max_j(\hat{Y}_j - M_n||x - X_j||),$$

and suppose that the minimum and maximum are achieved at $i = p$ and $j = q$, i.e.,

$$2m_n(x) = \hat{Y}_p + M_n||x - X_p|| + \hat{Y}_q - M_n||x - X_q||.$$

If $p = q$, we have $m_n(x) = \hat{Y}_p$. Otherwise we have two cases.

I. $\hat{Y}_q \geq \hat{Y}_p$. Since $\hat{Y}_p + M_n||x - X_p|| \leq \hat{Y}_q + M_n||x - X_q||$,

$$2m_n(x) \leq \hat{Y}_q + M_n||x - X_q|| + \hat{Y}_q - M_n||x - X_q|| = 2\hat{Y}_q.$$

II. $\hat{Y}_p \geq \hat{Y}_q$. Since $\hat{Y}_p + M_n||x - X_p|| \leq \hat{Y}_q + M_n||x - X_q||$,

$$M_n||x - X_p|| \leq M_n||x - X_q||.$$

Then $\hat{Y}_q - M_n||x - X_q|| \leq \hat{Y}_p - M_n||x - X_p||$, and hence

$$2m_n(x) \leq \hat{Y}_p + M_n||x - X_p|| + \hat{Y}_p - M_n||x - X_p|| = 2\hat{Y}_p.$$

In all cases $m_n(x) \leq \max\{\hat{Y}_p, \hat{Y}_q\}$. But this also means that $m_n(x) \leq \max_i \hat{Y}_i$. The second inequality $m_n(x) \geq \min_i \hat{Y}_i$ in proven in the same way. □

Using Proposition 1, we formulate the first step as the following quadratic programming problem. Define the residuals $R_i = \hat{Y}_i - Y_i$ and denote the pairwise distances by $d_{ij} = ||X_i - X_j||$. Then the system of inequalities

$$\forall i, j \in \{1, \ldots, n\},\ i \neq j : \ |Y_i - Y_j + R_i - R_j| \leq M_n d_{ij},$$

ensures that $\hat{\mathcal{D}}_n$ is compatible with $Lip(M_n)$. It then translates into

$$Y_j - Y_i - M_n d_{ij} \leq R_i - R_j \leq Y_j - Y_i + M_n d_{ij}. \tag{11}$$

Hence we solve a quadratic programming problem

$$\begin{aligned} & \min \tfrac{1}{n}\textstyle\sum_{i=1}^{n} R_i^2, \\ \text{s.t.} \quad & \forall i,j : R_i - R_j \le Y_j - Y_i + M_n d_{ij}, \\ & -\beta_n \le R_i + Y_i \le \beta_n, \end{aligned} \tag{12}$$

where the last set of inequalities ensures that the estimate is bounded, i.e., belongs to $\mathcal{F}_n$, as required by Theorems 1-2.

Solution of (12) yields the smoothened data set $\hat{\mathcal{D}}_n$, and the bounded estimate $m_n(\cdot)$ is computed using (9) and (10).

## 3.2 Sample splitting

Let us now address the question of estimating $M_n$ from the sample. We will employ two standard techniques, splitting the sample and cross-validation. Splitting the sample consists of using part of the data set $\mathcal{D}_n$ to construct the estimate $m_n$, and using the rest of the data to find an optimal value of a parameter used in the construction of $m_n$, in our case the Lipschitz constant $M_n$.

Subdivide the sample (randomly) into two non-overlapping subsets $\mathcal{D}_n^1 \cup \mathcal{D}_n^2 = \mathcal{D}_n$. We will use $\mathcal{I}_1, \mathcal{I}_2$ to denote the corresponding index sets, i.e., $\mathcal{D}_n^j = \{(X_i, Y_i) : i \in \mathcal{I}_j\}$ $(j \in \{1, 2\})$. We will use $\mathcal{D}_n^1$ to construct the estimate $m_n$ using (9) and (10), for any choice of $M$. We choose the optimal value $M_n$ based on how well $m_n$ predicts the values in the second subset $\mathcal{D}_n^2$. That is we solve the following optimization problem with respect to the variable $M > 0$

$$\min \sum_{i \in \mathcal{I}_2} (m_n(X_i) - Y_i)^2, \tag{13}$$

where $m_n$ is given by (9) and (10).

Notice that Eqs.(10) involve the smoothened values $\hat{Y}_i(M)$, found by solving problem (12) with a given $M$. Then we write (13) as the following bi-level optimization problem

$$\min_{M>0} \frac{1}{4} \sum_{i \in \mathcal{I}_2} \left[ \max_{j \in \mathcal{I}_1} (\hat{Y}_j(M) - M d_{ij}) + \min_{j \in \mathcal{I}_1} (\hat{Y}_j(M) + M d_{ij}) - 2Y_i \right]^2. \tag{14}$$

The values $\hat{Y}_j(M) = Y_j + R_j(M)$ are found as a solution to problem (12), which is now written as

$$\begin{aligned} & \min_R \textstyle\sum_{i\in\mathcal{I}_1} R_i^2, \\ \text{s.t.} \quad & R_i - R_j \le Y_j - Y_i + M d_{ij}, \\ & -\beta_n \le R_i + Y_i \le \beta_n, \\ & \forall i, j \in \mathcal{I}_1. \end{aligned} \tag{15}$$

At the outer level we solve (14) with respect to $M$, and at the inner level we solve (15), with $M$ supplied from the outer level. We note that the objective function in (14) is not smooth and may not be convex. Thus minimization of (14) should be performed by using non-smooth optimization techniques. In our study we used the Discrete Gradient method from Bagirov (1999) and Bagirov (2002), combined with simple grid search for a good choice of initial value of $M$.

## 3.3 Cross-validation

Another popular technique for estimating the value $M_n$ is cross-validation. Here we perform multiple splittings of the sample, and aggregate the results in the following way. Assume some value of $M$ and calculate the approximation to the data set not using the datum $(X_i, Y_i)$. Call this approximation $f_{(i,M)}(x)$. Then evaluate the error of approximation at $X_i$, i.e., $|f_{(i,M)}(X_i) - Y_i|$. Repeat for all $X_i$. In cross-validation $M_n$ is defined by

$$M_n = \arg\min_M \sum_{i=1}^{n} (f_{(i,M)}(X_i) - Y_i)^2, \tag{16}$$

where $f_{(i,M)} \in Lip(M)$ is the estimate with parameter $M$ applied to the data $\{(X_1, Y_1), \ldots, (X_{i-1}, Y_{i-1}), (X_{i+1}, Y_{i+1}), \ldots, (X_n, Y_n)\}$.

Let us now formulate an optimization problem resulting from the above criteria. From (9) and(10) we have

$$f_{(i,M)}(X_i) = \frac{1}{2}(\max_{j\neq i}(\hat{Y}_j^i(M) - M d_{ij}) + \min_{j\neq i}(\hat{Y}_j^i(M) + M d_{ij})),$$

where $\hat{Y}_j^i(M) = Y_j + R_j^i(M)$, and the superscript index $i$ underlines the dependence of the smoothened $\hat{Y}_j(M)$ on the excluded datum $i$.

The condition $f_{(i,M)} \in Lip(M)$ can be derived from Proposition 1 as

$$R_k^i - R_j^i \le Y_j - Y_k + Md_{kj},$$
$$\forall k, j \in \{1, \dots, n\} \setminus \{i\}.$$

Then minimization of the cross-validation criterion can be written as the following bi-level optimization problem. At the outer level we minimize with respect to $M$

$$\min_M \frac{1}{4} \sum_{i=1}^{n} \left[ \max_{j \ne i} (\hat{Y}_j^i(M) - Md_{ij}) + \min_{j \ne i} (\hat{Y}_j^i(M) + Md_{ij}) - 2Y_i \right]^2. \tag{17}$$

At the inner level, for every fixed $M$ we solve $n$ problems for all $i \in \{1, \dots, n\}$

$$\begin{aligned} &\min \textstyle\sum_{j \ne i} (R_j^i)^2 \\ \text{s.t.} \quad & R_k^i - R_j^i \le Y_j - Y_k + Md_{kj}, \\ & -\beta_n \le R_j^i + Y_j \le \beta_n, \\ & \forall k, j \in \{1, \dots, n\} \setminus \{i\}. \end{aligned} \tag{18}$$

As earlier, the objective function in (17) is not smooth, and we need to apply non-smooth optimization methods, such as Discrete Gradient, see Bagirov (1999), Bagirov (2002), combined with grid search for a good choice of initial value for $M$. We notice that for each fixed value of $M$ we need to solve $n$ quadratic programming problems (18), all having exactly the same (sparse) structure, but different right hand sides in the inequalities. Compared with sample splitting, there is an $n$-fold increase in the computational cost.

Once the optimal value $M_n$ is found, we substitute it in (12) and solve that problem to determine the set $\hat{\mathcal{D}}_n$. Then the estimate is computed from (9) and (10).

# 4 Estimation under additional conditions

In this section we consider the estimation problem subject to additional constraints on $\mathcal{F}_n$. In particular we will provide a method for estimating monotone and additive functions. Such conditions allow one to incorporate domain specific information about the function being estimated, known a priori. For example, monotonicity with respect to all or some variables is important in estimating aggregation functions in decision sciences, see Calvo et al. (2002), Beliakov (2003), and in empirical option pricing studies, Hutchison et al.(1994).

## 4.1 Monotone estimates

The additional information further limits the class of functions $\mathcal{F}_n$, and thus provides tighter estimates. We consider the class of monotone Lipschitz functions with the Lipschitz constant $M_n$,

$$\mathcal{MF}_n = \{f \in Lip(M_n) : f \text{ is monotone increasing}\}.$$

Monotonicity is assumed with respect to a subset of variables, denoted by $\mathcal{V} \subset \{1, \ldots, d\}$. That is for any such $\mathcal{V}$, $f$ is monotone increasing with respect to $\mathcal{V}$ on $\mathcal{X}$ if for all $x, y \in \mathcal{X}$, $x \succeq y$ implies $f(x) \geq f(y)$, where $x \succeq y$ means that $\forall i \in \mathcal{V} : x_i \geq y_i$, and $\forall j \notin \mathcal{V} : x_j = y_j$. This notion of monotonicity is equivalent to monotonicity of all univariate functions $f_i(x_i)$, $i \in \mathcal{V}$, obtained from $f(x)$ by fixing all but the $i$-th variable. Monotone decreasing functions are treated in the same way.

We call the data set $\mathcal{D}_n$ compatible with the class $\mathcal{MF}_n$ if there exists a monotone function $f \in \mathcal{MF}_n$ interpolating the data. As earlier, our interest is in the case when $\mathcal{D}_n$ is not compatible with $\mathcal{MF}_n$. We will build an estimate in two steps. First we will construct a data set $\hat{\mathcal{D}}_n$ compatible with $\mathcal{MF}_n$, and then construct an optimal estimate from this class.

We need an analogue of the Proposition 1.

**Proposition 3** *The data set $\mathcal{D}_n$ is compatible with the class $\mathcal{MF}_n$ if and only if the following conditions hold*

$$\forall i, j \in \{1, \ldots, n\} : Y_i - Y_j \leq M_n ||(X_i - X_j)_{\mathcal{V}+}||, \tag{19}$$

*where $Z_{\mathcal{V}+}$ denotes the positive part of vector $Z$ with respect to subset of components $\mathcal{V}$: $Z_{\mathcal{V}+} = (\bar{Z}_1, \ldots, \bar{Z}_d)$, with*

$$\bar{Z}_i = \begin{cases} \max\{Z_i, 0\}, & \text{if } i \in \mathcal{V}, \\ Z_i & \text{otherwise.} \end{cases}$$

To prove Proposition 3 we need the following auxiliary result from convex analysis, see Rockafellar (1970).

**Proposition 4** *Let $f, g$ be convex functions $\mathbb{R}^n \to \mathbb{R}$, and $f(x) \le g(x)$ for all $x \in \mathbb{R}^n$. If $g \in Lip(M)$, i.e., Lipschitz with Lipschitz constant $M$, then so is $f$.*

**Proof.** Every convex function can be represented as an upper envelop of its affine minorants

$$f = \sup\{l(x) \in U_f \subset \mathcal{A}\}, \quad g = \sup\{l(x) \in U_g \subset \mathcal{A}\},$$

where $U_f, U_g$ are nonempty subsets of affine minorants of $f$ and $g$ respectively (Rockafellar (1970)), and

$$\mathcal{A} = \{l(x) : l(x) = a \cdot x + b, a \in \mathbb{R}^n, b \in \mathbb{R}\}.$$

The Lipschitz constant of $f$ is the supremum of the Lipschitz constants of functions from $U_f$. Since $l(x) \le f(x) \le g(x)$ for all $l \in U_f$, then $U_f \subset U_g$. For each $l \in U_g$, its Lipschitz constant is no greater than that of $g$. Since the Lipschitz constant of $g$ is $M$, Lipschitz constants of all $l \in U_f$ are bounded by $M$, and so is their supremum, hence $f \in Lip(M)$. $\square$

**Proof of Proposition 3.** Assume the condition (19) holds. Let us define the function

$$f(x) = \max_i (Y_i - M_n ||(X_i - x)_{\mathcal{V}+}||),$$

This function is a pointwise maximum of functions

$$f_i(x) = Y_i - M_n ||(X_i - x)_{\mathcal{V}+}||.$$

We shall prove that each $f_i \in \mathcal{MF}_n$.

Recall that monotonicity as defined earlier means monotonicity with respect to each variable in $\mathcal{V}$. To prove monotonicity of $f_i(x)$ with respect to the $j$-th variable, $j \in \mathcal{V}$, fix the other components of $x$. Let $x, z \in \mathbb{R}^n$, $z_m = x_m, j \ne m$ and $z_j \ge x_j$. Then $z \succeq_{\mathcal{V}} x$.

$$\forall j : \max\{X_{ij} - x_j, 0\} \ge \max\{X_{ij} - z_j, 0\},$$

and

$$||(X_i - x)_{\mathcal{V}+}|| \ge ||(X_i - z)_{\mathcal{V}+}||.$$

Consequently

$$f_i(x) = Y_i - M_n ||(X_i - x)_{\mathcal{V}+}|| \le Y_i - M_n ||(X_i - z)_{\mathcal{V}+}|| = f_i(z).$$

To prove Lipschitz continuity, note that functions

$$\rho_1(x) = M_n||(X_i - x)_{\mathcal{V}+}|| \text{ and } \rho_2(x) = M_n||X_i - x||$$

are convex on $\mathbb{R}^n$, satisfy $\rho_1(x) \le \rho_2(x)$, and $\rho_2 \in Lip(M_n)$. By Proposition 4, $\rho_1 \in Lip(M_n)$, and consequently $f_i(x) = Y_i - \rho_1(x)$ is in $Lip(M_n)$.

Hence each $f_i \in \mathcal{MF}_n$, and their pointwise maximum $f \in \mathcal{MF}_n$. On the other hand, $f_i(X_i) = Y_i \ge Y_j - M_n||(X_j - X_i)_{\mathcal{V}+}||$ (which follows from (19)), and therefore $f(X_i) = f_i(X_i) = Y_i$. Hence $f$ interpolates the data.

Now, conversely, assume that for some $i, j$

$$Y_i - Y_j > M_n||(X_i - X_j)_{\mathcal{V}+}||.$$

Then any interpolant $f$ must satisfy

$$f(X_i) - f(X_j) > M_n||(X_i - X_j)_{\mathcal{V}+}||.$$

This violates the monotonicity and Lipschitz conditions implied in $\mathcal{MF}_n$. To show this consider three cases.

Case I. $X_i \succeq X_j$. Then $||(X_i - X_j)_{\mathcal{V}+}|| = ||(X_i - X_j)||$, and therefore the Lipschitz condition is violated

$$f(X_i) - f(X_j) > M_n||(X_i - X_j)||.$$

Case II. $X_i \preceq X_j$. Then $||(X_i - X_j)_{\mathcal{V}+}|| = 0$, and monotonicity is violated

$$f(X_i) - f(X_j) > 0, \text{ but } X_i \preceq X_j.$$

Case III. Neither of the above holds. Define vector $z$ as

$$z_k = \begin{cases} \max\{X_{ik}, X_{jk}\}, & \text{if } k \in \mathcal{V}, \\ X_{ik} & \text{otherwise.} \end{cases}$$

Because of monotonicity of $f$, $f(z) \ge f(X_i) > f(X_j)$. Also

$$||(X_i - X_j)_{\mathcal{V}+}|| = ||(z - X_j)_{\mathcal{V}+}|| = ||z - X_j||.$$

But then the Lipschitz condition is violated

$$f(z) - f(X_j) \ge f(X_i) - f(X_j) > M_n||(X_i - X_j)_{\mathcal{V}+}|| = M_n||z - X_j||. \quad \square$$

Next, to find the smoothened data set $\hat{\mathcal{D}}_n$ we solve the following quadratic programming problem, derived analogously to (12)

$$\begin{aligned} & \min \tfrac{1}{n}\textstyle\sum_{i=1}^{n} R_i^2, \\ \text{s.t.} \quad & R_i - R_j \le Y_j - Y_i + M_n d_{ij}, \\ & -\beta_n \le R_i + Y_i \le \beta_n, \end{aligned} \tag{20}$$

where as earlier $R_i = \hat{Y}_i - Y_i$, and $d_{ij} = ||(X_i - X_j)_{\mathcal{V}+}||$. Note that $d_{ij} \neq d_{ji}$.

Unlike the estimate $m_n$ in Section 3, monotone estimate does not have the range property $m_n(x) \in [\min_i \hat{Y}_i, \max_i \hat{Y}_i]$ on the whole $\mathbb{R}^n$, but on a subset

$$\{x \in \mathbb{R}^n : X_p \preceq x \preceq X_q\},$$

where $p, q$ are found from $\hat{Y}_p = \min_i \hat{Y}_i$, $\hat{Y}_q = \max_i \hat{Y}_i$. This easily follows from monotonicity of $m_n$. To ensure the range property on $\mathbb{R}^n$, we truncate the estimate by using

$$\begin{aligned} m_n^{upper}(x) &= \min\{\min_i(\hat{Y}_i + M_n||(x - X_i)_{\mathcal{V}+}||), \hat{Y}_q\} \\ m_n^{lower}(x) &= \max\{\max_i(\hat{Y}_i - M_n||(X_i - x)_{\mathcal{V}+}||), \hat{Y}_p\}. \end{aligned} \tag{21}$$

The last set of inequalities in (20) ensures that the estimate $m_n$ is bounded $||m_n||_\infty \le \beta_n$. In addition, the functions in (21) are monotone and Lipschitz continuous, with Lipschitz constant bounded by $M_n$.

Once the smoothened data set $\hat{\mathcal{D}}_n$ is found, we compute the optimal estimate by

$$m_n(x) = \frac{1}{2}(m_n^{lower}(x) + m_n^{upper}(x)), \tag{22}$$

where the upper and lower bounds are given in (21).

We note that simple compatibility of the data set with $\mathcal{MF}_n$ is not sufficient for monotonicity of the estimate. For instance, in more than two variables, interpolating linear splines may not be monotone, even if the data set is. (Consider the data generated by $f = (xy)^{1/4}$ with four data points at $(0,1), (1,0), (1,1), (0.1, 0.1)$. No linear spline interpolating these data is monotone.) It is therefore important to ensure monotonicity of the estimate after the data set $\hat{\mathcal{D}}_n$ is computed. The estimate $m_n$ given in (21) and (22) does guarantee monotonicity.

Analogously to Section 3, we can formulate the problem of estimating $M_n$ from the data via sample splitting or cross-validation. The changes will relate to a new definition of the values $d_{ij}$ above.

## 4.2 Additive estimates

Consider now additive functions $f(x) = f_1(x_1) + f_2(x_2) + \ldots + f_d(x_d)$. We will assume that each of the component functions $f_i \in Lip(M_i), i = 1, \ldots, d$. Without loss of generality, we will use the same Lipschitz constant with respect to all variables $M_1 = M_2 = \ldots = M_d = M$. We can always scale the coordinate axes so that such condition holds.

In turn this implies that $f \in Lip(\tilde{M})$, $\tilde{M} \le \sqrt{d}M$. However the individual Lipschitz conditions for the component functions are stronger, and thus we can obtain a better estimate by limiting the class of allowable estimates. Note that the values of the component functions are unknown, we observe only values of $f(X_i) = f_1(X_{i1}) + f_2(X_{i2}) + \ldots + f_d(X_{id})$ with additional random errors.

Our approach is again to obtain a data set $\hat{\mathcal{D}}_n$ compatible with the class

$$\mathcal{F}_n = \{f : f \text{ is additive}, f_i \in Lip(M), ||f_i||_\infty \le \beta_n, i = 1, \ldots, d\}.$$

Let us denote the (unknown) values of the individual estimates $m_{ij} = m_j(X_{ij}), j = 1, \ldots, d$, so that $\hat{Y}_i = \sum_{j=1}^d m_j(X_{ij})$. The residues are expressed as

$$R_i = \hat{Y}_i - Y_i = \sum_{j=1}^{d} m_{ij} - Y_i,$$

and the Lipschitz conditions are written as

$$\forall j = \{1, \ldots, d\}, \forall i, k \in \{1, \ldots, n\} : m_{ij} - m_{kj} \le M|X_{ij} - X_{kj}|.$$

Then we formulate a quadratic programming problem to minimize the norm of the residuals

$$\begin{aligned} & \min \tfrac{1}{n} \textstyle\sum_{i=1}^n R_i^2, \\ \text{s.t.} \quad & m_{ij} - m_{kj} \le M|X_{ij} - X_{kj}|, \\ & \textstyle\sum_{j=1}^d m_{ij} - R_i = Y_i, \\ & -\beta_n \le m_{ij} \le \beta_n, \\ & j = 1, \ldots, d,\ i, k = 1, \ldots, n. \end{aligned} \tag{23}$$

The smoothened data values are computed from $\hat{Y}_i = Y_i + R_i$. Then the optimal estimate is obtained from

$$m_n(x) = \frac{1}{2}(m_n^{lower}(x) + m_n^{upper}(x)), \tag{24}$$

where

$$m_n^{upper}(x) = \sum_{j=1}^{d} \min_i (m_{ij} + M|x_j - X_{ij}|), \tag{25}$$

$$m_n^{lower}(x) = \sum_{j=1}^{d} \max_i (m_{ij} - M|x_j - X_{ij}|).$$

Compared to Problem (12), there are $d$ times more inequalities in (23), which apparently make the numerical solution less efficient. However note that many constraints are redundant, and can be eliminated from (23). For this for each $j \in \{1, \ldots, d\}$ sort $X_{ij}$ in increasing order. Denote by $X_{(i)j}$ the sorted components and by $m_{(i)j}$ the corresponding sorted values. Then

$$\begin{aligned} m_{(i)j} - m_{(i+1)j} &\leq M|X_{(i)j} - X_{(i+1)j}| \text{ and} \\ m_{(i+1)j} - m_{(i+2)j} &\leq M|X_{(i+1)j} - X_{(i+2)j}| \text{ imply} \\ m_{(i)j} - m_{(i+2)j} &\leq M|X_{(i)j} - X_{(i+2)j}|, \end{aligned}$$

hence the last inequality is redundant. In fact the only non-redundant inequalities in (23) are those involving the indices $(i)$ and $(i+1)$, of which there are $2d \times (n-1)$. By eliminating the redundant constraints from the problem formulation, we make the solution algorithm more efficient. For comparison, problem (12) involves a significantly larger number $n \times (n-1)$ of inequality constraints.

In complete analogy to Section 3, we can formulate the problem of estimating $M$ from the data by using sample splitting or cross-validation.

# 5 Application to simulated data

To study the performance and applicability of the estimation methods discussed in Sections 3-4, we generated data sets using model distributions, involving a regression function $m(x)$ and distributions of $X$ and $Y$. For $X$ we used uniform distribution on some closed interval (or hyperrectangle in the multivariate case), as for $Y$ we used Gaussian distribution with the mean $m(X)$ and standard deviation $\sigma$. Our goal is to find an estimate $m_n(x)$ from the class $\mathcal{F}_n$, and evaluate its $L_2$ error defined by (4).

To compare with other estimates, we used two popular methods: Neural Networks and Regression Trees. Implementations of these methods in the statistical package R were used.

For the model distributions, we used the following regression functions.

$$
\begin{aligned}
m_A(x) &= \max\{-x^2 + x + 2, 1 - \frac{x}{2}, \min\{1 + \frac{x}{2}, 4 - x\}\}, x \in [-2, 3],\\
m_B(x) &= x \sin(x^2), x \in [-2, 2],\\
m_C(x) &= x_1 \sin(x_1^2) - x_2 \sin(x_2^2), x \in [-2, 2]^2,\\
m_D(x) &= x_1 \sin(x_1^2) - x_2 \sin(x_2^2) + x_3 \sin(x_3^2), x \in [-2, 2]^3,\\
m_E(x) &= x_1 \sin(x_1^2) - x_2 \sin(x_2^2) + x_3 \sin(x_3^2) - x_4 \sin(x_4^2), x \in [-2, 2]^4,
\end{aligned}
$$

For each of these functions, we randomly generated data sets $\mathcal{D}_n$ of various sizes. Then we used these data sets to find estimates $m_n$ by one of the indicated methods, and then evaluated numerically the $L_2$ error of the estimates defined in (4). Since this error is a random variable itself, we repeated the experiment 100 times, each time with a different sample with the same distribution of $(X, Y)$, and reported the mean $L_2$ error and its standard deviation.

We used sample splitting technique with $|\mathcal{I}_1| = |\mathcal{I}_2| = n/2$ to find the optimal parameters in each estimation method. For the Lipschitz approximation, the parameter is the Lipschitz constant of the estimate $M_n$. For neural network estimate, the parameter is the size of the hidden layer (i.e., we used networks with $1 - H - 1$ architecture, with the best $H$ determined by sample splitting. Typically $H$ was around 10 in our experiments). For regression trees, the optimal partition was determined automatically by the `rpart` module in the R software.

Table 1 summarizes the results. The values of $L_2$ error were estimated by using out of sample prediction for a large number of $X$ uniformly distributed in the domain of $m$. Our simulations show that the Lipschitz estimate is as good as the other two well established estimates for univariate functions, and outperformed them for multivariate functions.

We notice that the additivity of the estimate (if this property holds and is taken into account), helps improving the accuracy of the estimate, as well as reduces computing time. In addition, our technique allows one to construct monotone estimates, which are required in some applications. In contrast, it is not clear how one can construct monotone estimates with neural networks or regression trees.

Figures 1-5 illustrate the computed estimate in one and two variables.

| Regression function $m(x)$ | $d$ | $n$ | $\sigma$ (noise) | Lipschitz estimate | ANN | R-trees |
|---|---|---|---|---|---|---|
| $m_A(x)$ | 1 | 500 | 1 | **0.026** (0.015) | 0.029 (0.017) | 0.05 (0.02) |
| $m_A(x)$ | 1 | 3000 | 1 | **0.0052** (0.005) | 0.0058 (0.003) | 0.023 (0.01) |
| $m_A(x)$ | 1 | 500 | 0.2 | 0.0024 (0.003) | **0.0018**(0.005) | 0.009 (0.004) |
| $m_A(x)$ | 1 | 3000 | 0.2 | 0.0029 (0.03) | **0.0021** (0.02) | 0.008 (0.003) |
| $m_B(x)$ | 1 | 500 | 0.2 | 0.0021 (0.012) | **0.0019** (0.01) | 0.05 (0.015) |
| $m_B(x)$ | 1 | 3000 | 0.2 | 0.0022 (0.02) | **0.0021** (0.02) | 0.04 (0.01) |
| $m_C(x)$ | 2 | 500 | 0.2 | **0.084** (0.011) | 0.096 (0.05) | 0.27 (0.01) |
| $m_C(x)$[1] | 2 | 500 | 0.2 | **0.012** (0.003) | 0.096 (0.05) | 0.27 (0.01) |
| $m_C(x)$[1] | 2 | 3000 | 0.2 | **0.006** (0.001) | 0.084 (0.06) | 0.25 (0.01) |
| $m_D(x)$ | 3 | 500 | 0.2 | **0.46** (0.03) | 0.53 (0.06) | 0.88 (0.02) |
| $m_D(x)$[1] | 3 | 500 | 0.2 | **0.02** (0.013) | 0.53 (0.06) | 0.88 (0.02) |
| $m_D(x)$[1] | 3 | 3000 | 0.2 | **0.04** (0.01) | 0.46 (0.05) | 0.79 (0.01) |
| $m_E(x)$ | 4 | 500 | 0.2 | **0.71** (0.1) | 0.86 (0.5) | 1.1 (0.01) |
| $m_E(x)$[1] | 4 | 500 | 0.2 | **0.11** (0.1) | 0.86 (0.5) | 1.1 (0.01) |
| $m_E(x)$[1] | 4 | 3000 | 0.2 | **0.07** (0.02) | 0.51 (0.4) | 0.95 (0.01) |

Table 1. The mean $L_2$ error for the Lipschitz estimate, and its standard deviation (in brackets), compared to $L_2$ error of Neural Networks and Regression Trees.
[1] Additivity property was used for the Lipschitz estimate.

# 6 Proofs

In the proofs we need the notion of covering numbers.

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

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

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

For sets of Lipschitz continuous functions we prove the following bounds on the covering number:

**Lemma 1** *Let $\mathcal{F}$ be a set of functions $f : [0,1]^d \to \mathbb{R}$, which are Lipschitz continuous with Lipschitz constant $M > 0$ and bounded in absolute value by $\beta > 0$. Let $0 < \epsilon < \beta$ and let $x_1, \dots, x_n \in \mathbb{R}^d$. Then*

$$\mathcal{N}_1(\epsilon, \mathcal{F}, x_1^n) \le \left( c_4 \frac{\beta^2}{\epsilon^2} \right)^{(\sqrt{d}\frac{M}{\epsilon}+2)^d} \tag{26}$$

*for some constant $c_4 > 0$, and for any $\delta > 0$ and $g : \mathbb{R}^d \to \mathbb{R}$*

$$\begin{aligned} &\mathcal{N}_2 \left( \epsilon, \left\{ f - g : f \in \mathcal{F}, \frac{1}{n} \sum_{i=1}^{n} |f(x_i) - g(x_i)|^2 \le \delta^2 \right\}, x_1^n \right) \\ &\le \left( \frac{8\delta + 5\epsilon}{\epsilon} \right)^{\left(\sqrt{d}\frac{M}{\epsilon}+2\right)^d} . \end{aligned} \tag{27}$$

**Proof.** Partition $[0,1]^d$ into $K = \left( \lceil \sqrt{d} M / \epsilon \rceil \right)^d$ equidistant cubes $A_1, \dots, A_K$ of sidelength $\delta = 1 / \lceil \sqrt{d} M / \epsilon \rceil$, and let $\mathcal{G}$ be the set of all functions which are piecewise constant with respect to that partition. For $k \in \{1, \dots, K\}$ let $z_k$ be the center of $A_k$, and define for $f : [0,1]^d \to \mathbb{R}$ a function $g_f : [0,1]^d \to \mathbb{R}$ by

$$g_f = \sum_{k=1}^{K} f(z_k) \cdot 1_{A_k}.$$

By Lipschitz continuity we have for any $f \in \mathcal{F}$:

$$\sup_{x \in [0,1]^d} |f(x) - g_f(x)| \leq M \max_{k=1,\dots,K} \sup_{x \in A_k} \|x - z_k\| \leq M \cdot \sqrt{d} \cdot \frac{\delta}{2} \leq \frac{\epsilon}{2},$$

from which we can conclude

$$\mathcal{N}_1(\epsilon, \mathcal{F}, x_1^n) \leq \mathcal{N}_1(\frac{\epsilon}{2}, \{g_f : f \in \mathcal{F}\}, x_1^n).$$

Since $\{g_f : f \in \mathcal{F}\}$ is a subset of a linear vector space of dimension $K$ consisting of functions bounded in absolute value by $\beta$, we can conclude from this together with Lemma 9.2, Theorem 9.4 and Theorem 9.5 in Györfi et al. (2002)

$$\mathcal{N}_1(\epsilon, \mathcal{F}, x_1^n) \leq 3 \cdot \left( \frac{2e(2\beta)}{\epsilon/2} \cdot \log \frac{3e(2\beta)}{\epsilon/2} \right)^{(\sqrt{d} \cdot \frac{M}{\epsilon} + 1)^d + 1}.$$

Next we prove (27). Set

$$\mathcal{H} = \left\{ g_f - g \,:\, f \in \mathcal{F}, \frac{1}{n} \sum_{i=1}^n |f(x_i) - g(x_i)|^2 \leq \delta^2 \right\}.$$

For each $f \in \mathcal{F}$ we have

$$\sup_{x \in [0,1]^d} |(f(x) - g(x)) - (g_f(x) - g(x))| = \sup_{x \in [0,1]^d} |f(x) - g_f(x)| \leq \frac{\epsilon}{2},$$

so each $\epsilon/2$–cover of $\mathcal{H}$ on $x_1^n$ is also a $\epsilon$-cover of

$$\left\{ f - g \,:\, f \in \mathcal{F}, \frac{1}{n} \sum_{i=1}^n |f(x_i) - g(x_i)|^2 \leq \delta^2 \right\}$$

on $x_1^n$, from which we can conclude

$$\mathcal{N}_2 \left( \epsilon, \left\{ f - g \,:\, f \in \mathcal{F}, \frac{1}{n} \sum_{i=1}^n |f(x_i) - g(x_i)|^2 \leq \delta^2 \right\}, x_1^n \right) \leq \mathcal{N}_2 \left( \frac{\epsilon}{2}, \mathcal{H}, x_1^n \right). \tag{28}$$

$\mathcal{H}$ is a subset of a linear vector space of dimension $K+1$, and each $h = g_f - g \in \mathcal{H}$ satisfies

$$\begin{aligned}\sqrt{\frac{1}{n}\sum_{i=1}^{n} h(x_i)^2} &\leq \sqrt{\frac{1}{n}\sum_{i=1}^{n} |f(x_i) - g_f(x_i)|^2} + \sqrt{\frac{1}{n}\sum_{i=1}^{n} |f(x_i) - g(x_i)|^2} \\ &\leq \frac{\epsilon}{2} + \delta.\end{aligned}$$

By Corollary 2.6 in van de Geer (2000) we get

$$\mathcal{N}_2\left(\frac{\epsilon}{2}, \mathcal{H}, x_1^n\right) \leq \left(\frac{4(\epsilon/2+\delta)+\epsilon/2}{\epsilon/2}\right)^{K+1} \leq \left(\frac{8\delta+5\epsilon}{\epsilon}\right)^{\left(\sqrt{d}\frac{M}{\epsilon}+2\right)^d}. \quad (29)$$

The assertion follows from (28) and (29). $\square$

**Proof of Theorem 1.** W.l.o.g. $\mathcal{X} = [0,1]^d$. For $z \in \mathbb{R}$ set $T_L z = \min\{L, \max\{-L, z\}\}$. By Theorem 10.2 in Györfi et al. (2002), it suffices to show

$$\lim_{n\to\infty} \inf_{f\in\mathcal{F}_n} \int |f(x) - m(x)|^2 \mathbf{P}_X(dx) = 0 \quad (30)$$

and

$$\sup_{f\in\mathcal{F}_n} \left| \frac{1}{n}\sum_{i=1}^{n} |f(X_i) - T_L Y_i|^2 - \mathbf{E}\{|f(X) - T_L Y|^2\} \right| \to 0 \quad (31)$$

$(n\to\infty)$ $a.s.$ for all $L > 0$ and all distributions of $(X, Y)$ with $X \in [0,1]^d$ $a.s.$ and $\mathbf{E}Y^2 < \infty$.

Since $m$ is in $L_2(\mathbf{P}_X)$, we can find for any $\epsilon > 0$ an infinitely often differentiable function $g : [0,1]^d \to \mathbb{R}$ such that

$$\int |g(x) - m(x)|^2 \mathbf{P}_X(dx) < \epsilon.$$

Since $[0,1]^d$ is compact, $g$ is bounded and Lipschitz constant with respect to some finite Lipschitz bound $M > 0$. Because of $M_n \to \infty$ and $\beta_n \to \infty$ $(n \to \infty)$ we have $g \in Lip(M_n)$ and $\|g\|_\infty \leq \beta_n$ for $n$ sufficiently large, which implies (30).

Next we prove (31). Let $0 < \epsilon < L < \beta_n$. Arguing similar to the proof of Theorem 10.3 in Györfi et al. (2002) we can conclude from Theorem 9.1 in

Györfi et al. (2002) and Lemma 1 above

$$\mathbf{P}\left[\sup_{f\in\mathcal{F}_n}\left|\frac{1}{n}\sum_{i=1}^{n}|f(X_i)-T_LY_i|^2-\mathbf{E}\{|f(X)-T_LY|^2\}\right|>\epsilon\right]$$
$$\leq 8\cdot\left(c_5\frac{\beta_n^4}{\epsilon^2}\right)^{(c_6\frac{\beta_n M_n}{\epsilon})^d} e^{-c_7\frac{n\cdot\epsilon^2}{\beta_n^4}}.$$

By the assupmtions on $M_n$ and $\beta_n$ we can bound the right-hand side above for $n$ sufficiently large by

$$24\cdot e^{-c_7\frac{n\cdot\epsilon^2}{2\beta_n^4}},$$

which is summable for each $\epsilon>0$. An application of the lemma of Borel-Cantelli yields (31). □

In the proof of Theorem 2 we need the following lemma from Kohler (2002).

**Lemma 2** *(Lemma 5 in Kohler (2002)). Let $L\geq 1$, let $m:\mathbb{R}^d\to[-L,L]$ and let $\mathcal{F}$ be a class of functions $f:\mathbb{R}^d\to[-L,L]$. Let $0<\epsilon<1$ and $\alpha>0$. Assume that*

$$\sqrt{n}\epsilon\sqrt{\alpha}\geq 1152L$$

*and that, for all $x_1,\dots,x_n\in\mathbb{R}^d$ and all $\delta\geq 2L^2\alpha$,*

$$\frac{\sqrt{n}\epsilon\delta}{768\sqrt{2}L^2}\geq$$
$$\int_{\frac{\epsilon\delta}{128L^2}}^{\sqrt{\delta}}\left(\log\mathcal{N}_2\left(\frac{u}{4L},\left\{f-m\ :\ f\in\mathcal{F},\frac{1}{n}\sum_{i=1}^{n}|f(x_i)-m(x_i)|^2\leq\frac{\delta}{L^2}\right\},x_1^n\right)\right)^{1/2}du.$$

*Then*

$$\mathbf{P}\left\{\sup_{f\in\mathcal{F}}\frac{\left|\mathbf{E}\{|f(X)-m(X)|^2\}-\frac{1}{n}\sum_{i=1}^{n}|f(X_i)-m(X_i)|^2\right|}{\alpha+\mathbf{E}\{|f(X)-m(X)|^2\}+\frac{1}{n}\sum_{i=1}^{n}|f(X_i)-m(X_i)|^2}>\epsilon\right\}$$
$$\leq 15\exp\left(-\frac{n\alpha\epsilon^2}{512\cdot 2304L^2}\right).$$

**Proof of Theorem 2.** W.l.o.g. we assume $\mathcal{X}=[0,1]^d$. Set

$$\alpha_n=c_8M_n^{\frac{2}{2p+1}}n^{-\frac{2p}{2p+1}}$$

for $d = 1$ and

$$\alpha_n = c_9 (\log(n))^{\frac{3}{d}} M^{\frac{d+2}{2p+d}} n^{\frac{-2p}{2p+d} \cdot \frac{d+2}{2d}}$$

for $d > 1$. We have

$$\mathbf{P}\left\{ \int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx) > \alpha_n \right\} \leq P_{1,n} + P_{2,n},$$

where

$$P_{1,n} = \mathbf{P}\Bigg\{ \int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx) \\ -2 \cdot \frac{1}{n} \sum_{i=1}^{n} |m_n(X_i) - m(X_i)|^2 > \frac{\alpha_n}{2} \Bigg\}$$

and

$$P_{2,n} = \mathbf{P}\left\{ \frac{1}{n} \sum_{i=1}^{n} |m_n(X_i) - m(X_i)|^2 > \frac{\alpha_n}{4} \right\}.$$

In the first part of the proof we bound $P_{2,n}$. This can be done by results from van de Geer (2000), in particular by applying Theorem 10.11 in van de Geer (2000). We use here a version of this theorem, which was proven in Lemma 2 in Kohler (2002). From this we get

$$\mathbf{P}\left\{ \frac{1}{n} \sum_{i=1}^{n} |m_n(X_i) - m(X_i)|^2 > c_{10} \left( \delta_n + \min_{f \in \mathcal{F}_n} \frac{1}{n} \sum_{i=1}^{n} |f(X_i) - m(X_i)|^2 \right) \right\} \to 0$$

$(n \to \infty)$ provided $\delta_n > 0$ satisfies

$$\delta_n \to 0 \quad (n \to \infty), \quad n \cdot \delta_n \to \infty \quad (n \to \infty)$$

and

$$\sqrt{n}\delta \geq c_{11} \int_{\delta/(2^9 \sigma_0)}^{\sqrt{\delta}} \Bigg( \log \mathcal{N}_2(u, \{f - g : f \in \mathcal{F}_n, \\ \frac{1}{n} \sum_{i=1}^{n} |f(x_i) - g(x_i)|^2 \leq \delta\}, x_1^n \Bigg)^{1/2} du \quad (32)$$

for all $\delta \geq \delta_n$ and all $g \in \mathcal{F}_n$. By Lemma 1 we can bound the covering number above by

$$\mathcal{N}_2\left(u, \{f-g : f \in \mathcal{F}_n, \frac{1}{n}\sum_{i=1}^n |f(x_i)-g(x_i)|^2 \leq \delta\}, x_1^n\right)$$
$$\leq \left(\frac{8\sqrt{\delta}+5u}{u}\right)^{(\sqrt{d}M_n/u+2)^d},$$

which implies that the right-hand side of (32) is bounded from above by

$$c_{11}\int_{\delta/(2^9\sigma_0)}^{\sqrt{\delta}} (\sqrt{d}M_n/u+2)^{d/2}\left(\log\frac{8\sqrt{\delta}+5u}{u}\right)^{1/2} du$$
$$= c_{11}\int_1^{c_{12}/\sqrt{\delta}} \left(\sqrt{d}\frac{M_n}{\sqrt{\delta}}v+2\right)^{d/2} (\log(8v+5))^{1/2}\frac{\sqrt{\delta}}{v^2}dv.$$

In case $d=1$ this in turn is bounded by

$$c_{13}M_n^{d/2}\delta^{1/2-d/4} = c_{13}M_n^{1/2}\delta^{1/4},$$

from which we get that (32) is implied by

$$\sqrt{n}\delta \geq c_{14}M_n^{1/2}\delta^{1/4}$$

for all $\delta \geq \delta_n$, which is equivalent to

$$\delta_n \geq c_{15}M_n^{2/3}n^{-2/3} = c_{15}M^{2/(2p+1)}n^{-2p/(2p+1)}.$$

In case $d>1$ we assume $\delta_n \geq 1/n$. From this we conclude $\sqrt{\log(8v+5)} \leq c_{16}\sqrt{\log(n)}$ for $v \leq c_{12}/\sqrt{\delta}$ and $\delta \geq \delta_n$, and bound the above integral by

$$c_{17}\frac{M_n^{d/2}}{\delta^{d/4}}\cdot\delta^{1/2}\int_1^{c_{12}/\sqrt{\delta}} \sqrt{\log(8v+5)}v^{d/2-2}dv \leq c_{18}M^{\frac{d(d+2)}{4p+2d}}n^{\frac{d(1-p)}{4p+2d}}(\log(n))^{3/2}\delta^{1-d/2},$$

from which we can conclude that (32) is implied by

$$\delta_n \geq c_{19}(\log(n))^{\frac{3}{d}}M^{\frac{d+2}{2p+d}}n^{\frac{-2p}{2p+d}\cdot\frac{d+2}{2d}}.$$

Set $K = \lceil M^{2/(2p+d)} n^{1/(2p+d)} / d^{1/(1-p)} \rceil$ and $z_j = (j_1/K, \ldots, j_d/K)^T$ for $j = (j_1, \ldots, j_d) \in \{0, 1, \ldots, K\}^d$. Since for any $z_j$ and $z_k$ with $\sum_{l=1}^d |j_l - k_l| = 1$

$$\frac{|m(z_j) - m(z_k)|}{\|z_j - z_k\|} \leq M \cdot \|z_j - z_k\|^{p-1} = \frac{M}{K^{p-1}} \leq \frac{M_n}{d},$$

we have for arbitrary $j, k \in \{0, 1, \ldots, K\}^d$

$$|m(z_j) - m(z_k)| \leq M_n \|z_j - z_k\|.$$

According to Propositions 1 and 2 we can find $g \in \mathcal{F}_n$ with

$$g(z_j) = m(z_j) \quad (j \in \{0, 1, \ldots, K\}^d).$$

For $x \in [0,1]^d$ choose $j \in \{0, 1, \ldots, K\}^d$ with

$$\|x - z_j\| \leq \frac{\sqrt{d}}{2K}.$$

Then

$$\begin{aligned}
|g(x) - m(x)| &\leq |g(x) - g(z_j)| + |g(z_j) - m(z_j)| + |m(z_j) - m(x)| \\
&\leq M_n \|x - z_j\| + 0 + M \|z_j - x\|^p \\
&\leq \frac{\sqrt{d} \cdot M_n}{2K} + \frac{d^{p/2} M}{2^p K^p} \\
&\leq c_{20} M^{d/(2p+d)} n^{-p/(2p+d)},
\end{aligned}$$

from which we can conclude

$$\begin{aligned}
\min_{f \in \mathcal{F}_n} \frac{1}{n} \sum_{i=1}^n |f(X_i) - m(X_i)|^2 &\leq \frac{1}{n} \sum_{i=1}^n |g(X_i) - m(X_i)|^2 \\
&\leq \left( \max_{x \in [0,1]^d} |g(x) - m(x)| \right)^2 \\
&\leq c_{21} M^{2d/(2p+d)} n^{-2p/(2p+d)}.
\end{aligned}$$

Summarizing the above results we get

$$P_{2,n} \leq \mathbf{P} \left\{ \frac{1}{n} \sum_{i=1}^n |m_n(X_i) - m(X_i)|^2 > \frac{\alpha_n}{4} \right\} \to 0 \quad (n \to \infty).$$

In the second part of the proof we bound $P_{1,n}$. We have

$$\begin{aligned}
&P_{1,n}\\
&= \mathbf{P}\left\{\frac{\int |m_n(x)-m(x)|^2\mathbf{P}_X(dx) - \frac{1}{n}\sum_{i=1}^n |m_n(X_i)-m(X_i)|^2}{\alpha_n/2 + \int |m_n(x)-m(x)|^2\mathbf{P}_X(dx)} > \frac{1}{2}\right\}\\
&\leq \mathbf{P}\Bigg\{\exists f \in Lip(M_n), \|f\|_\infty \leq L :\\
&\qquad\qquad \frac{\mathbf{E}\{|f(X)-m(X)|^2\} - \frac{1}{n}\sum_{i=1}^n |f(X_i)-m(X_i)|^2}{\alpha_n/2 + \mathbf{E}\{|f(X)-m(X)|^2\}} > \frac{1}{2}\Bigg\}.
\end{aligned}$$

As in the proof of (32) for $\delta_n = \alpha_n$ we get that the assumptions of Lemma 2 are satisfied, and by applying it to the last probability we get

$$P_{1,n} \leq 15 \cdot \exp\left(-\frac{n \cdot \frac{\alpha_n}{2} \cdot \frac{1}{4}}{512 \cdot 2304 \cdot L^2}\right) \to 0 \quad (n \to \infty)$$

□

**Proof of Theorem 3.** For an additive function $f : [0,1]^d \to \mathbb{R}$ with

$$f(x_1,\ldots,x_d) = f_1(x_1) + \cdots + f_d(x_d) \quad (x_1,\ldots,x_d \in [0,1])$$

define $g_f : [0,1]^d \to \mathbb{R}$ by

$$g_f(x_1,\ldots,x_d) = g_{f_1}(x_1) + \cdots + g_{f_d}(x_d),$$

where $g_{f_j}$ is defined as in the proof of Lemma 1 (with $d=1$). If $f$ is given as above with $f_1,\ldots,f_d \in Lip(M)$, then we get from the results in the proof of Lemma 1

$$\sup_{x\in[0,1]^d} |f(x)-g_f(x)| \leq \sup_{x\in[0,1]} |f_1(x)-g_{f_1}(x)| + \cdots + \sup_{x\in[0,1]} |f_d(x)-g_{f_d}(x)| \leq d\cdot\frac{\epsilon}{2}.$$

As in the proof of Lemma 1 this implies that we have for any $g : [0,1]^d \to \mathbb{R}$

$$\begin{aligned}
&\mathcal{N}_2\Bigg(\frac{d+1}{2}\epsilon, \Big\{f-g \,:\, f = f_1+\cdots+f_d \text{ with } f_j \in Lip(M),\\
&\qquad\qquad \frac{1}{n}\sum_{i=1}^n |f(x_i)-g(x_i)|^2 \leq \delta^2\Big\}, x_1^m\Bigg)\\
&\leq \left(\frac{8\delta + (4d+1)\cdot\epsilon}{\epsilon}\right)^{d\cdot(M/\epsilon+2)}.
\end{aligned}$$

From this we get the assertion as in the proof of Theorem 2. □

# 7 Conclusion

In this paper we considered the problem of estimation of a regression function from a given sample of the underlying distribution of $(X, Y)$, by using Lipschitz continuous estimates. We have shown that the least squares Lipschitz estimates are consistent with respect to $L_2$ error, for all distributions, where the predictor variable is bounded. Further, for Hoelder-continuous regression functions, we established the rate of convergence of the estimate, which turns out to be optimal for dimension $d = 1$, and achieves up to a logarithmic factor optimal rate of convergence for $d = 2$. We also showed that for additive regression functions, the optimal rate of convergence is achieved for any dimension.

This paper provides a number of practical methods to compute the optimal Lipschitz estimate, formulated as a quadratic programming problem. We developed methods to estimate the unknown Lipschitz constant from the noisy data, by using sample splitting and cross-validation. We also constructed the estimates with additional restrictions, such as monotone and additive estimates. In our numerical simulations, the Lipschitz estimate outperformed two other popular non-parametric estimation methods (neural networks and regression trees) for multivariate estimates for a number of distributions.

# References

[1] Bagirov, A. (1999). Derivative-free methods for unconstrained nonsmooth optimization and its numerical analysis. *Journal Investigacao Operacional* **19**, pp. 75–93.

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

[3] Beliakov, G. (2003). How to build aggregation operators from data? *Int. J. Intelligent Systems* **18**, pp. 903–923.

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

[5] Calvo, T., Kolesarova, A., Komornikova, M., and Mesiar, R. (2002). Aggregation operators: properties, classes and construction methods. In

Calvo, T., Mayor, G., and Mesiar, R., editors, *Aggregation operators. New trends and applications*, pages 3–104. Physica-Verlag, Heidelberg, New York.

[6] Chen, Z. (1991). Interaction spline models and their convergence rates. *Ann. Statist.* **19**, pp. 1855–1868.

[7] Devroye, L., Györfi, L., Krzyżak, A. and Lugosi, G. (1994). On the strong universal consistency of nearest neighbor regression function estimates. *Ann. Statist.* **22**, pp. 1371–1385.

[8] Devroye, L. P. and Wagner, T. J. (1980). Distribution–free consistency results in nonparametric discrimination and regression function estimation. *Ann. Statist.* **8**, pp. 231–239.

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

[10] Gaffney, P. and Powell, M. (1976). Optimal interpolation. In Watson, G., editor, *Numerical Analysis*, volume 506 of *Lecture Notes in Mathematics*, pages 90–99. Springer, Heidelberg.

[11] Györfi, L., Kohler, M., Krzyżyak, A., Walk,H. (2002). *A distribution–free theory of nonparametric regression.* Springer–Verlag, New York.

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

[13] Hutchison, J., Lo, A., and Poggio, T. (1994). A nonparametric approach to pricing and hedging derivative securities via learning networks. *J. of Finance* **49**, pp. 851–889.

[14] Kohler, M. (1998). Nonparametric Regression Function Estimation Using Interaction Least Squares Splines and Complexity Regularization. *Metrika* **47**, pp. 147-163.

[15] Kohler, M. (2002). Nonparametric regression with additional measurements errors in the dependent variable. Preprint 2002-14, University of Stuttgart. Available from http://www.isa.uni-stuttgart.de/LstStoch/Kohler/artikel20.ps

[16] Krzyżak, A. and Linder, T. (1998). Radial basis function networks and complexity regularization in function learning. *IEEE Transaction on Neural Networks* **9**, pp. 247–256.

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

[18] Lugosi, G. and Nobel, A. (1996). Consistency of data-driven histogram methods for density estimation and classification. *Ann. Statist.* **24**, pp. 687–706.

[19] Lugosi, G. and Zeger, K. (1995). Nonparametric estimation via empirical risk minimization. *IEEE Trans. Inform. Theory* **41**, pp. 677-687.

[20] Micchelli, C. and Rivlin, T. (1984). Lectures on optimal recovery. In Turner, P., editor, *Numerical Analysis*, volume 1129 of *Lecture Notes in Mathematics*, pages 12–93. Springer, Heidelberg.

[21] Rockafellar, R. (1970). *Convex Analysis.* Princeton University Press, Princeton.

[22] Stone, C. J. (1977). Consistent nonparametric regression. *Ann. Statist.* **5**, pp. 595–645.

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

[24] Stone, C. J. (1985). Additive regression and other nonparametric models. *Ann. Statist.* **13**, pp. 689–705.

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

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

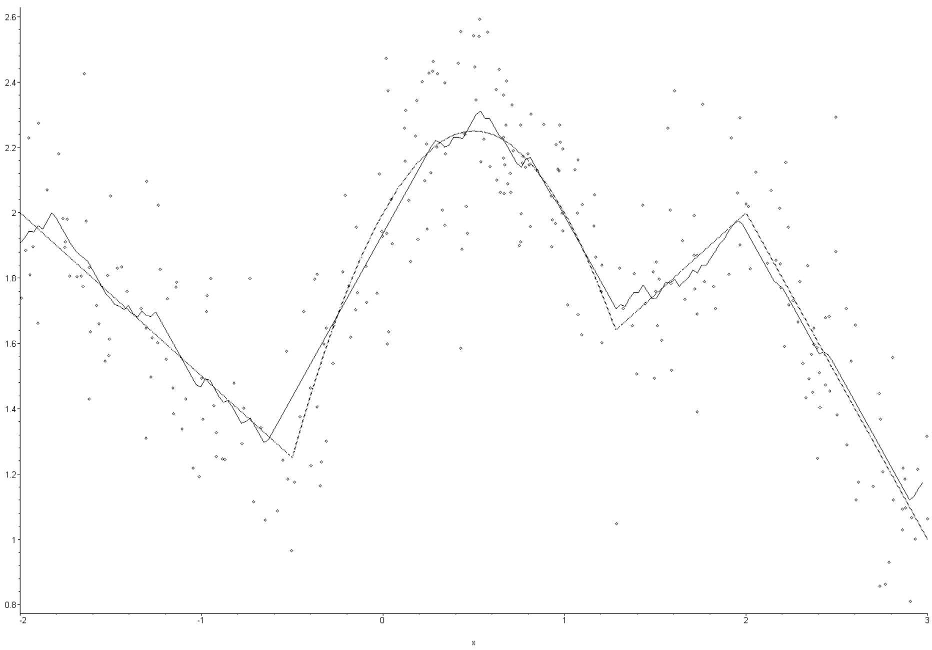


Figure 1: Lipschitz estimate of test function $m_A(x)$ using 300 noisy values ($\sigma = 0.2$). The Lipschitz constant was estimated by using sample splitting.

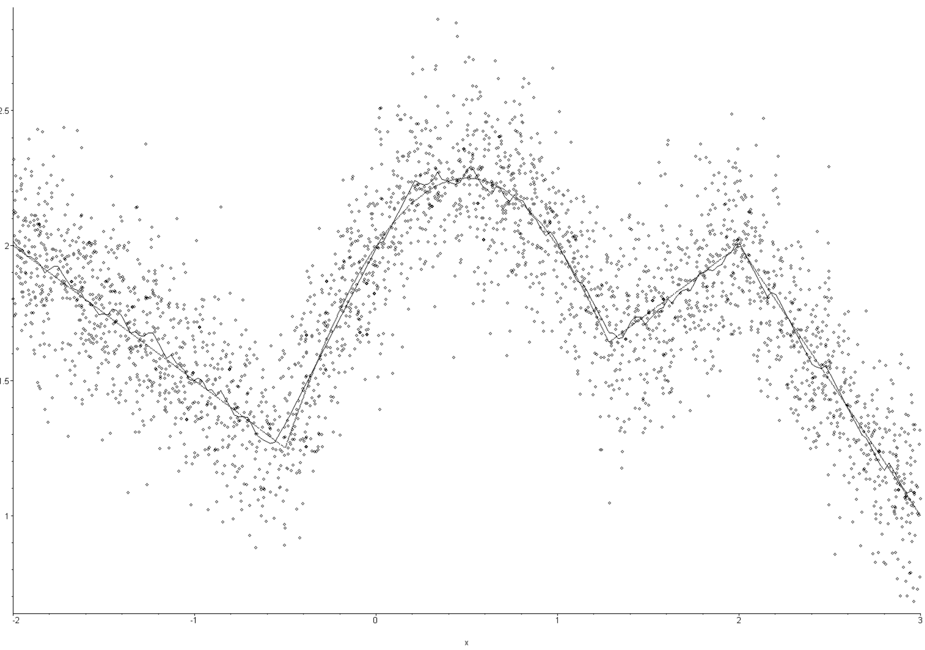


Figure 2: Lipschitz estimate of test function $m_A(x)$ using 3000 noisy values ($\sigma = 0.2$). The Lipschitz constant was estimated by using sample splitting.

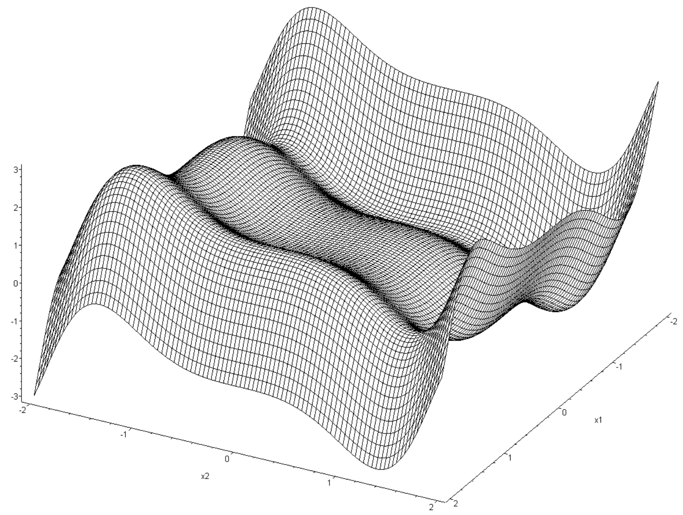


Figure 3: The model function $m_C(x)$ in the case of two variables.

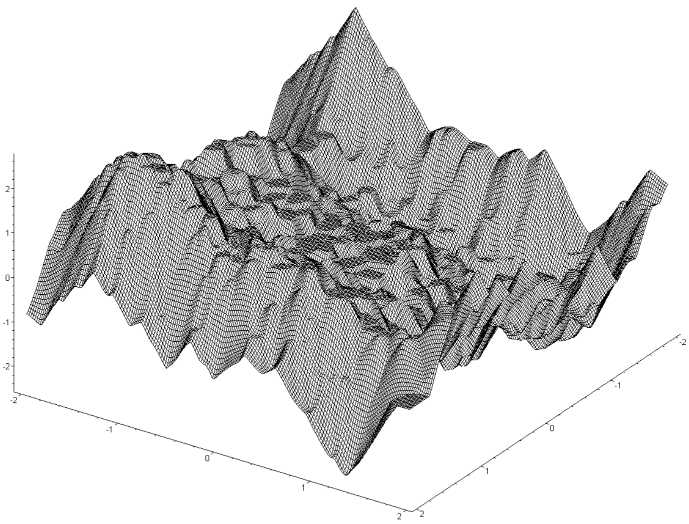


Figure 4: Lipschitz estimate of test function $m_C(x)$ using 500 noisy values ($\sigma = 0.2$). No assumption of additivity was used.

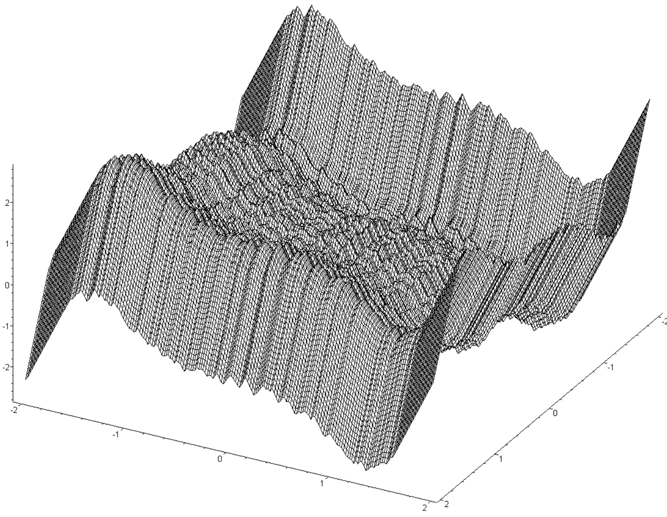


Figure 5: Additive Lipschitz estimate of test function $m_C(x)$ using 500 noisy values ($\sigma = 0.2$).