Fixed design regression estimation based on real and artificial data

Dmytro Furer[1], Michael Kohler[1] and Adam Krzyżak[2,*]
[1] *Fachbereich Mathematik, Technische Universität Darmstadt, Schlossgartenstr. 7, 64289 Darmstadt, Germany, email: furer@mathematik.tu-darmstadt.de, kohler@mathematik.tu-darmstadt.de*
[2] *Department of Computer Science and Software Engineering, Concordia University, 1455 De Maisonneuve Blvd. West, Montreal, Quebec, Canada H3G 1M8, email: krzyzak@cs.concordia.ca*

May 21, 2012

Abstract

In this article we study fixed design regression estimation based on real and artificial data, where the artificial data comes from previously undertaken similar experiments. A least squares estimate is introduced which gives different weights to the real and the artificial data. It is investigated under which condition the rate of convergence of this estimate is better than the rate of convergence of an ordinary least squares estimate applied to the real data only. The results are illustrated using simulated and real data.

AMS classification: Primary 62G08; secondary 62G20.

Key words and phrases: Fixed design regression, nonparametric estimation, L_2 error, rate of convergence.

1 Introduction

In this article we study the fixed design regression estimation problem, where we are given data

$$(x_1, Y_1), \cdots, (x_n, Y_n) \tag{1}$$

satisfying $x_1, \ldots, x_n \in [0,1]$ and

$$Y_i = m(x_i) + \epsilon_i, \quad i = 1, \ldots, n$$

for so-called regression function $m : [0,1] \to \mathbb{R}$ and some independent random variables $\epsilon_1, \ldots, \epsilon_n$ with mean zero. The goal is to estimate m from the data (1).

There are two different approaches here: parametric regression where it is assumed that the structure of m is known and depends only on finitely many parameters, and the data (1) is used to construct estimates of these parameters, and nonparametric regression where there is no assumption on the structure of the regression function.

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

The principle of the least squares is a popular principle to construct regression estimates. One chooses for an estimate of the regression function a function which minimizes the so-called empirical L_2 risk over some given set $\mathcal{F}_n$ of functions $f : \mathbb{R}^d \to \mathbb{R}$, and defines the estimate by

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

For notational simplicity we assume here and in the sequel that the minimum exists. Depending on the structure of $\mathcal{F}_n$ this leads either to parametric or nonparametric regression estimation.

1.1 Fixed design regression using real and artificial data

Often one has the problem that the set $\mathcal{D}_n = \{(x_1, Y_1),...,(x_n, Y_n)\}$ contains only a few data points. One remedy is to generate N_n artificial data points and to add them to the existing data. The artificial data might come, e.g., from previously undertaken similar experiments (for details see Section 3 below). Then we have a new dataset

$$\mathcal{D}_N = \{(x_1, Y_1), \ldots, (x_n, Y_n), (x_{n+1}, \hat{Y}_{n+1}), \ldots, (x_N, \hat{Y}_N)\}$$

of size $N := n + N_n$. In the sequel we assume that the artificial data points have the property that

$$\frac{1}{N_n} \sum_{j=1}^{N_n} |\hat{Y}_{n+j} - m(x_{n+j})|^2 \quad \text{is “small”}. \tag{3}$$

We define least squares regression estimates by minimizing a weighted combination of two empirical squared risks:

$$\bar{m}_n(\cdot) = \arg \min_{f \in \mathcal{F}_n} \left(\sum_{i=1}^{n} w_i \cdot |f(x_i) - Y_i|^2 + \sum_{j=1}^{N_n} w_{n+j} \cdot |f(x_{n+j}) - \hat{Y}_{n+j}|^2 \right), \tag{4}$$

where

$$w_i = \frac{1}{n} \cdot w^{(n)} \quad (i \in \{1, \ldots, n\}) \quad \text{and} \quad w_{n+j} = \frac{1}{N_n} \cdot (1 - w^{(n)}) \quad (j \in \{1, \ldots, N_n\}) \tag{5}$$

for some $w^{(n)} \in [0, 1]$. Note that for this choice of w_i the sum of all weights is one.
Remark 1. For $w^{(n)} = 1$ we use only real data points for our estimate, and for $w^{(n)} = 0$ the estimate is based exclusively on artificial data. For $w^{(n)} = \frac{n}{N} = \frac{n}{n+N_n}$ we weigh all data points equally, i.e., in this case the estimate is given by

$$\bar{m}_n(\cdot) = \arg \min_{f \in \mathcal{F}_n} \frac{1}{N} \left(\sum_{i=1}^{n} |f(x_i) - Y_i|^2 + \sum_{j=1}^{N_n} |f(x_{n+j}) - \hat{Y}_{n+j}|^2 \right).$$

In this article we derive upper bounds on the L_2 error of the estimate (4). As it turns out, in view of the optimal rate of convergence it is not necessary to use simultaneously real and artificial data for the estimate: depending for which data the corresponding error of the regression estimate is smaller, it suffices to use only one kind of data in the estimate. However, we show with the help of simulated data that for a finite sample size this is not always the case. Here the estimate using simultaneously real and artificial data sometimes outperforms the estimate based only on one type of data. Finally, we apply the proposed estimate to fatigue analysis.

1.2 Discussion of related results

The fixed design regression estimation has been studied for a long time, for survey see, e.g., Gasser and Müller (1979) or Eubank (1999).

Our theoretical result is based on the empirical process theory as presented, e.g., in the monograph by van de Geer (2000). In particular we use in our proofs techniques introduced in Kohler (2006) in context of regression estimation with additional measurement errors in the dependent variable.

In application to fatigue analysis we use nonparametric regression with random design in order to generate the artificial data. The most popular estimates for random design regression include kernel regression estimate (cf., e.g., Nadaraya (1964, 1970), Watson (1964), Devroye and Wagner (1980), Stone (1977) or Devroye and Krzyżak (1989)), partitioning regression estimate (cf., e.g., Györfi (1981) or Beirlant and Györfi (1998)), nearest neighbor regression estimate (cf., e.g., Devroye (1982), Devroye, Györfi, Krzyżak and Lugosi (1994), Mack (1981) or Zhao (1987)), least squares estimates (cf., e.g., Lugosi and Zeger (1995)) or smoothing spline estimates (cf., e.g., Kohler and Krzyżak (2001)). The main theoretical results are summarized in the monograph by Györfi et al. (2002).

1.3 Outline

The main result is formulated in Section 2 and illustrated by applying the estimates to the simulated and real data in Section 3. Section 4 contains the proofs and an auxiliary result is proven in the Appendix.

2 Main result

We next describe the model. Set $N := n + N_n$ and let $x_1, \dots, x_N \in [0,1]$. Furthermore, set

$$Y_i = m(x_i) + W_i \quad (i = 1, \dots, n) \tag{6}$$

for some $m : [0,1] \to \mathbb{R}$ and some random variables $W_1, \dots, W_n$ which are independent and have mean zero. We assume that the W_i's are sub-Gaussian in the sense that

$$\max_{i=1,\dots,n} K^2 \mathbf{E}\{e^{W_i^2/K^2} - 1\} \leq \sigma_0^2 \tag{7}$$

for some $K, \sigma_0 > 0$. Our goal is to estimate m from the data

$$(x_1, Y_1), \dots, (x_n, Y_n), (x_{n+1}, \hat{Y}_{n+1}), \dots, (x_N, \hat{Y}_N),$$

where we assume that

$$\frac{1}{N_n} \sum_{j=1}^{N_n} |\hat{Y}_{n+j} - m(x_{n+j})|^2 \tag{8}$$

is "small". Let $\mathcal{F}_n$ be a set of functions $f : \mathbb{R}^d \to \mathbb{R}$. Consider the least squares estimate

$$\bar{m}_n(\cdot) = \arg\min_{f \in \mathcal{F}_n} \left(\sum_{i=1}^{n} w_i \cdot |f(x_i) - Y_i|^2 + \sum_{j=1}^{N_n} w_{n+j} \cdot |f(x_{n+j}) - \hat{Y}_{n+j}|^2 \right)$$

with weights w_i defined as in Section 1.

We say that $a_n = O_{\mathbf{P}}(b_n)$ if $\limsup_{n\to\infty} \mathbf{P}(a_n > c \cdot b_n) = 0$ for some finite constant c. Our main result is the following theorem, which bounds the L_2-error of $\bar{m}_n$

$$\int_0^1 |\bar{m}_n(x) - m(x)|^2 dx.$$

Theorem 1. *Let $\mathcal{F}_n$ be a set of functions which are Lipschitz-continuous with Lipschitz-constant bounded by $\log(n)$ and which are also bounded in absolute value by $\log(n)$ and assume that m is Lipschitz-continuous. There exists a constant $c_1 > 0$ which depends only on σ_0 and K such that for any $\delta_n > 0$ with*

$$\delta_n \to 0 \quad (n \to \infty) \quad \textit{and} \quad n \cdot \delta_n \to \infty \quad (n \to \infty)$$

and

$$\sqrt{n} \cdot \delta \geq c_1 \int_{\delta/(2^9\sigma_0)}^{\sqrt{\delta}} \left(\log \mathcal{N}_2 \left(u, \{f - g : f \in \mathcal{F}_n, \frac{1}{n}\sum_{i=1}^{n} |f(x_i) - g(x_i)|^2 \leq \delta\}, x_1^n \right)\right)^{1/2} du \quad (9)$$

for all $\delta \geq \delta_n$ and all $g \in \mathcal{F}_n$ we have

$$\int_0^1 |\bar{m}_n(x) - m(x)|^2 dx$$

$$= O_{\mathbf{P}} \Bigg(\log(n)^2 \cdot \left(\frac{w^{(n)}}{n} + \frac{(1 - w^{(n)})}{N_n} \right) + \frac{1 - w^{(n)}}{N_n} \cdot \sum_{j=1}^{N_n} |\hat{Y}_{n+j} - m(x_{n+j})|^2$$

$$+ w^{(n)} \cdot \delta_n + \min_{f \in \mathcal{F}_n} \sum_{i=1}^{n+N_n} w_i \cdot |f(x_i) - m(x_i)|^2 \Bigg).$$

Remark 2. a) We can interpret the above result as follows: our bound on the L_2 error depends on the sample size, the division between the real and artificial data, the quality of the artificial data, the complexity of the function space $\mathcal{F}_n$ (measured by δ_n and (9)), and the approximation error of the function space $\mathcal{F}_n$.
b) The upper bound in the last two lines in the above theorem is the sum of three terms which we can interpret as follows: the first summand bounds the difference between L_2 error and the empirical L_2 error and will be negligible with respect to the remaining terms. The second summand is $(1 - w^{(n)})$ times the average error of the artificial data. The remaining two summands are standard bounds on the empirical L_2 error of the least squares estimate based on the real data. As we will see below (see Corollary 2) they converge to zero at the same rate as the error of a least squares estimate applied to $\lceil n/w^{(n)} \rceil$ real data points.

c) In order to achieve the optimal rate of convergence, the choice of $w^{(n)}$ is obvious: if the average squared error of the artificial data converges faster (or slower) to 0, than the L_2 error of the least squares estimate applied to the real data, we should set $w^{(n)} = 0$ (or $w^{(n)} = 1$), respectively.

To illustrate the usefulness of our main result we show what happens if we apply it to the linear least squares estimates.

Corollary 1. *Let $\mathcal{F}_n$ be a set of functions which are Lipschitz-continuous with Lipschitz-constant bounded by* $\log(n)$ *and which are also bounded in absolute value by* $\log(n)$*, assume that $\mathcal{F}_n$ is a subset of a linear vector space of dimension D_n and assume that m is Lipschitz-continuous. Then*

$$\int_0^1 |\bar{m}_n(x) - m(x)|^2 dx$$

$$= O_{\mathbf{P}} \left(\log(n)^2 \cdot \left(\frac{w^{(n)}}{n} + \frac{\left(1 - w^{(n)}\right)}{N_n} \right) + \frac{1 - w^{(n)}}{N_n} \cdot \sum_{j=1}^{N_n} |\hat{Y}_{n+j} - m(x_{n+j})|^2 \right.$$

$$\left. + w^{(n)} \cdot \frac{D_n}{n} + \min_{f \in \mathcal{F}_n} \sum_{i=1}^{n+N_n} w_i \cdot |f(x_i) - m(x_i)|^2 \right).$$

Proof. The result follows immediately from Theorem 1 and the bound on the covering number of linear vector spaces given in Corollary 2.6 in van de Geer (2000), which implies that condition (9) is in the case of linear vector spaces satisfied for $\delta_n \geq c_2 \frac{D_n}{n}$ (cf., Example 9.3.1 in van de Geer (2000) or proof of Lemma 19.1 in Györfi et al. (2002)). □

For particular choices of sets $\mathcal{F}_n$ we derive bounds on the approximation error under appropriate smoothness assumptions on m

$$\min_{f \in \mathcal{F}_n} \sum_{i=1}^{n+N_n} w_i \cdot |f(x_i) - m(x_i)|^2 \leq \min_{f \in \mathcal{F}_n} \sup_{x \in [0,1]} |f(x) - m(x)|^2,$$

which together with Corollary 1 yield bounds on the rate of convergence of the estimate. Here we describe the smoothness of m as follows:

Definition 1. *Let $C > 0$ and $p = k + \beta$ for some $k \in \mathbb{N}_0$ and $0 < \beta \leq 1$. A function $m : [0,1] \to \mathbb{R}$ is called (p, C)–smooth if its k–th derivative $m^{(k)}$ exists and satisfies*

$$|m^{(k)}(x) - m^{(k)}(z)| \leq C|x - z|^\beta \tag{10}$$

for all $x, z \in [0,1]$.

If we choose $\mathcal{F}_n$ as set of piecewise polynomials, we get:

Corollary 2. *Let $L, C > 0$ and $p = k + \beta$ for some $k \in \mathbb{N}$ and $\beta \in (0,1]$. Assume $|m(x)| \leq L$ for some $L > 0$ and m (p, C)-smooth. Define $\mathcal{F}_n$ as the set of all piecewise polynomials of degree $M \geq k$ with respect to an equidistant partition of* $[0,1]$ *into*

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

equidistant intervals, where the coefficients of the piecewise polynomials are bounded in absolute value by $(\log n)/(M+1)^2$*, and where each piecewise polynomial is on* $[0,1]$ *Lipschitz-continuous with Lipschitz constant bounded by* $\log(n)$*. Let* $\bar{m}_n$ *be defined as in Section 1 for some* $w^{(n)} > 0$*. Then*

$$\int_0^1 |\bar{m}_n(x) - m(x)|^2 dx$$
$$= O_{\mathbf{P}} \left(\log(n)^2 \cdot \left(\frac{w^{(n)}}{n} + \frac{\left(1 - w^{(n)}\right)}{N_n} \right) + \frac{1 - w^{(n)}}{N_n} \cdot \sum_{j=1}^{N_n} |\hat{Y}_{n+j} - m(x_{n+j})|^2 \right.$$
$$\left. + \left(\frac{w^{(n)}}{n} \right)^{2p/(2p+1)} \right).$$

Proof. The proof follows from Corollary 1 and Lemma 11.1 in Györfi et al. (2002) (cf., proof of Corollary 19.1 in Györfi et al. (2002)). □

Remark 3. a) In Corollary 2 we estimate a smooth regression function. In this case smooth estimates are often used, which can be achieved in the situation above by choosing $\mathcal{F}_n$ as an appropriate spline space. It follows from the proof that in this case we can use, e.g., the spline space from Fromkorth and Kohler (2011) and the assertion of Corollary 2 still holds.

b) Any application of the above estimate to real data requires a data-driven choice of the parameters $(w^{(n)}, N_n, K_n, M)$ of the estimate. This can be done by, for instance, applying cross-validation to the real data.

3 Application to simulated and real data

In this section we illustrate the finite sample size performance of our newly proposed estimate by applying it to simulated and real data.

We start with the simulation using artificial data. Here we consider three different regression functions $m_i : [0,1] \to \mathbb{R}$, $i = 1,2,3$ defined by

$$m_1(x) = \sin(5x), \quad m_2(x) = e^{5x} - (5x)^3 \quad \text{and} \quad m_3(x) = \frac{1}{5x+1} + \sin(5x),$$

cf. Figures 1.

We define our real data by

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

where $x_i = i/n$ and $W_1, \dots, W_n$ are independent standard normal random variables. The artificial data will be generated by

$$Y_{n+j} = m(x_{n+j}) + \delta, \quad j = 1, \dots, N_n,$$

where $x_{n+j} = j/N_n$ and $\delta > 0$ is some fixed constant which takes different values in the simulations. In all our experiments we choose $n = 1000$ and $N_n = 1000$. The weights

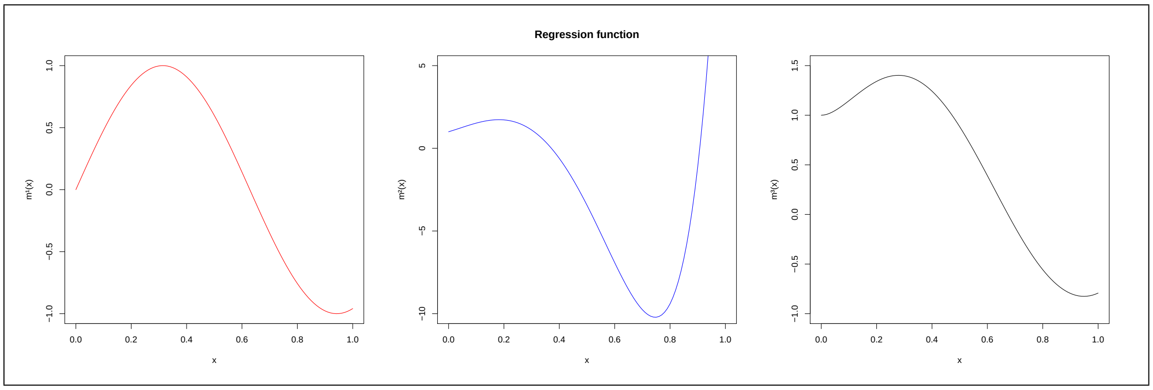

Figure 1: $m_1(x) = \sin(5x)$, $\quad m_2(x) = e^{5x} - (5x)^3 \quad$ and $\quad m_3(x) = \frac{1}{5x+1} + \sin(5x)$

of the least squares estimate are defined as in Section 1, and we consider three cases, namely $w^{(n)} = 1$ (we use only real data), $w^{(n)} = 0$ (we use only artificial data) and $w^{(n)} = n/(n + N_n)$ (we give all data points the same weight). For the function space we use polynomial splines of degree 2, where the number of equidistant knots is chosen from the set $\{1, \dots, 10\}$ by splitting of the sample (applied to the whole set of real and artificial data). For the the first regression function we do simulations for $\delta \in \{0.04, 0.1, 0.2\}$. For the second regression function we choose $\delta \in \{0.08, 0.15, 0.25\}$ and for the last regression function we consider $\delta \in \{0.04, 0.1, 0.2\}$. For each value of δ we generate independently 100 data sets, apply to each data set the three estimates corresponding to the above mentioned three values of $w^{(n)}$ and compute the square roots of the corresponding L_2 errors of the estimates. The results are presented in the boxplots in Figures 2, 3 and 4.

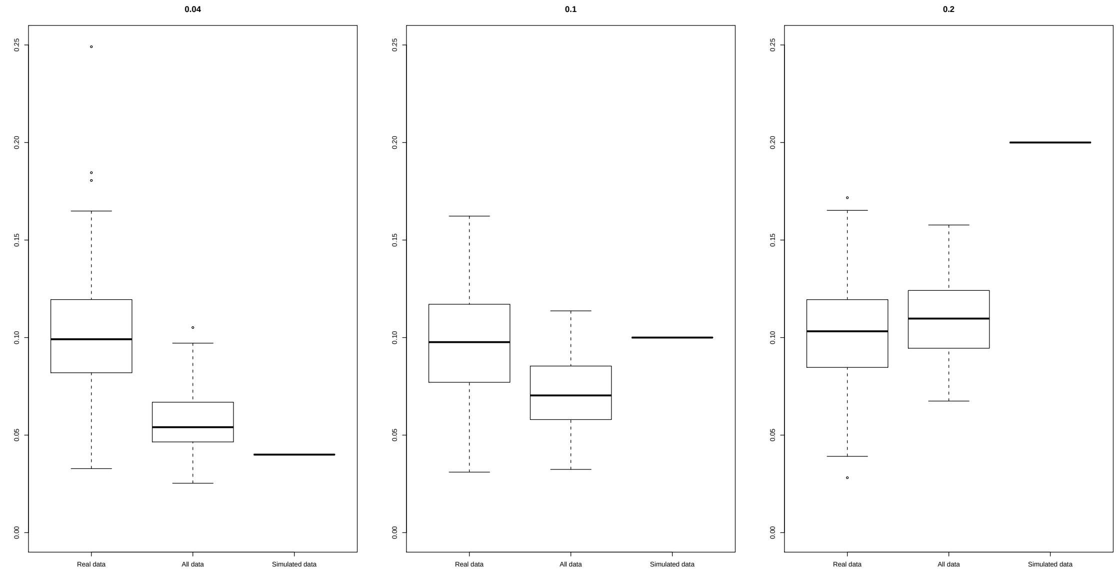

Figure 2: $m_1(x)$

For all three regression functions we see that for δ much larger (or much smaller) than

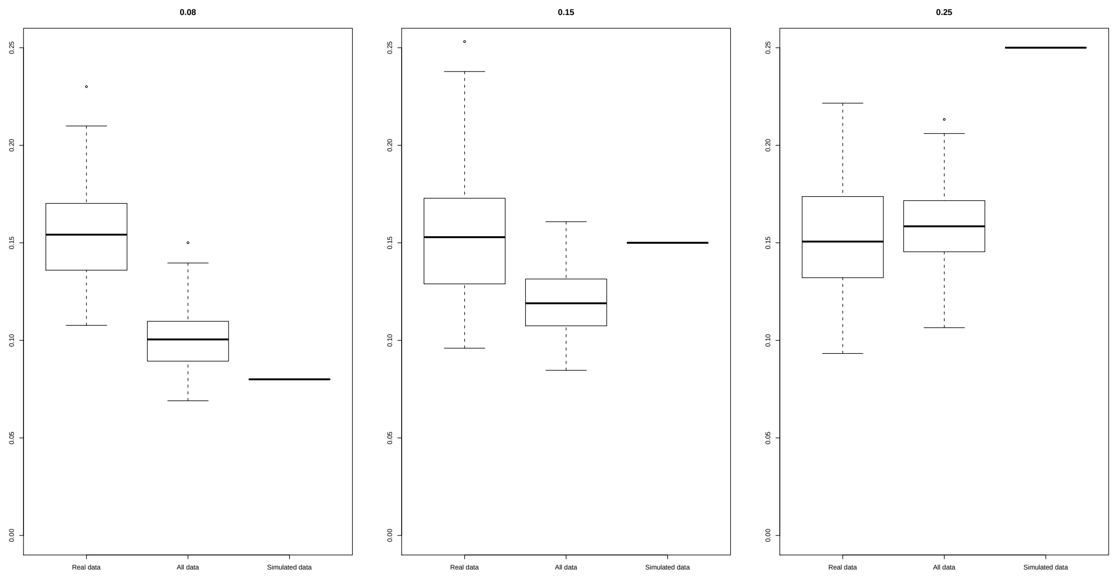

Figure 3: $m_2(x)$

the median error of the estimate based only on the real data, the estimate using the real data only (or the estimate using the artificial data only, respectively) is the best. This confirms our theoretical results of Section 2. However, when δ is approximately equal to the median error of the estimate using the real data only, the estimate giving equal weights to all data points performs better than the other two estimates. This shows that in the case of finite sample size it is sometimes beneficial to combine real and artificial data in the same estimate.

We now apply our methodology to estimation of fatigue behavior of steel under cyclic loading. Our data is obtained in a series of seven experiments where for seven values of the total strain amplitude ϵ the corresponding number of cycles N_f till failure and the corresponding stress amplitude σ are determined. The observed values are given in Table 1.

ϵ	0.003	0.0035	0.004	0.004	0.0045	0.005	0.005
N_f	28572	8077	7878	2919	2950	1865	4015
σ	402.9	437.2	426.1	434.3	456.6	475.3	447.1

Table 1: Observed values in experimental fatigue tests.

Our least squares estimate is based on the Manson-Coffin-Basquin relation (cf., e.g., Manson (1965))

$$\epsilon = \frac{\sigma'_f}{E} \cdot (2N_f)^b + \epsilon'_f \cdot (2N_f)^c, \tag{11}$$

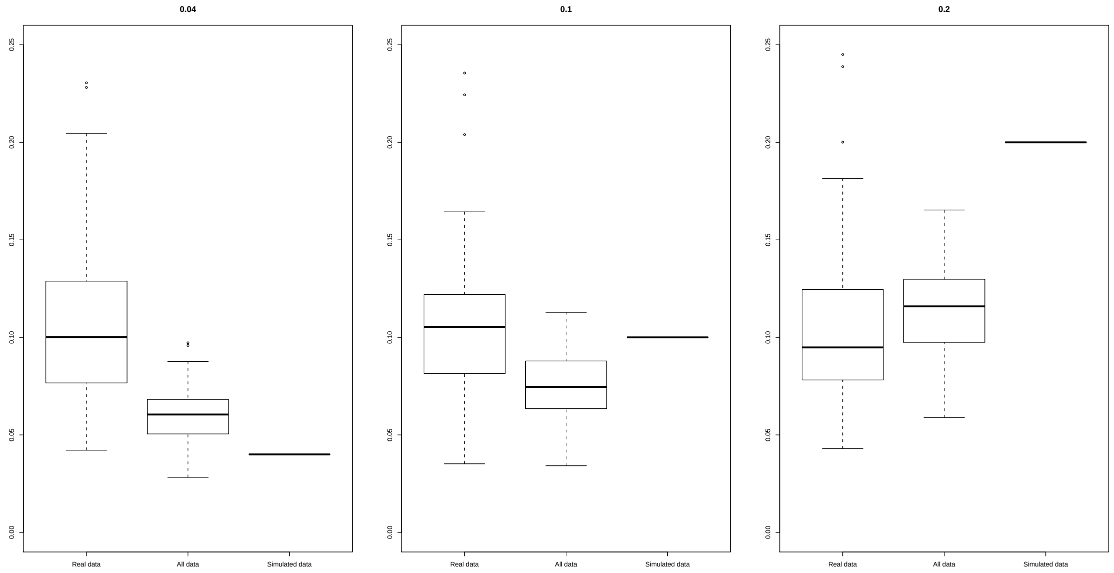

Figure 4: $m_3(x)$

and the Ramberg-Osgood equation (cf., e.g., Ramberg and Osgood (1943))

$$\epsilon = \epsilon_e + \epsilon_p = \frac{\sigma}{E} + \left(\frac{\sigma}{K'}\right)^{\frac{1}{n'}}. \tag{12}$$

Here σ'_f,ϵ'_f, b, c, K'_f, n' and E are parameters describing the tested material. E denotes the modulus of elasticity which is known from a previously performed static tensile material test.

We fit this nonlinear model for the inverse relations between ϵ and N and between ϵ and σ to the data by minimizing the least squares criterion with gradient descent based on the computation of the gradient using the implicit function theorem. Figure 5 shows the real data points together with the estimate of the relations between the strain amplitudes and the number of cycles till failure. As estimates for the parameters for the relation between the strain amplitudes and the number of cycles till failure we get $\sigma'_f = 651.178$, $\epsilon'_f = 0.0406$, $b = -0.0431$ and $c = -0.3331$.

Since the above experiments are extremely time consuming we augment our measured data by artificially generated data. To do this we use experiments for related materials (where the relation to our steel is measured by using so called static material parameters like yield limit for 0.2% residual elongation, temperature, modulus of elasticity and sensitivity of static stress strain curve), determine from the relation (11) and (12) for chosen values of the strain amplitude the corresponding values of N_f and σ, and use nonparametric regression to estimate the corresponding values for our material on the basis of static parameters of our steel. In this way we generate $N_n = 100$ additional data points containing values of number of cycles N_f till failure and stress amplitude σ. We choose the weight of our combined least squares estimate using cross-validation applied to our seven real data points from the set $\{0, 0.01, \ldots, 1\}$.

Figure 6 shows the real data points, the artificial data points and the estimate of the relation between the strain amplitudes and number of cycles till failure based on the combination of both data points. Here our data driven choice of the weight yields $w = 0.94$. As estimates for the parameters for the relation between the strain amplitudes and the number of cycles till failure we get this time $\sigma'_f = 680.0259$, $\epsilon'_f = 0.0279$, $b = -0.054$ and $c = -0.2866$.

The two estimates are compared in Figure 7.

4 Proofs

4.1 A deterministic lemma

In this subsection we formulate a deterministic lemma which we will need in the next section in order to bound the empirical L_2 error of our least squares estimate.

Lemma 1. *Let $t > 0$, $N := n + N_n$, $w_1, \dots, w_N \in \mathbb{R}_+$, $x_1, \dots, x_N \in \mathbb{R}^d$, $y_1, \dots, y_n \in \mathbb{R}$ and $\hat{y}_{n+1}, \dots, \hat{y}_N \in \mathbb{R}$. Let m be a function $m : \mathbb{R}^d \to \mathbb{R}$ and let $\mathcal{F}$ be a set of functions $f : \mathbb{R}^d \to \mathbb{R}$. Set*

$$\bar{m}_n(\cdot) = \arg\min_{f \in \mathcal{F}} \left(\sum_{i=1}^{n} w_i \cdot |f(x_i) - y_i|^2 + \sum_{j=1}^{N_n} w_{n+j} \cdot |f(x_{n+j}) - \hat{y}_{n+j}|^2 \right)$$

and

$$m_n^*(\cdot) = \arg\min_{f \in \mathcal{F}} \sum_{i=1}^{n+N_n} w_i \cdot |f(x_i) - m(x_i)|^2$$

and assume that both minima exist. Then

$$\begin{aligned} &\sum_{i=1}^{n+N_n} w_i \cdot |\bar{m}_n(x_i) - m(x_i)|^2 \\ &> t + 512 \sum_{j=1}^{N_n} w_{n+j} \cdot |m(x_{n+j}) - \hat{y}_{n+j}|^2 + 18 \min_{f \in \mathcal{F}} \sum_{i=1}^{n+N_n} w_i \cdot |f(x_i) - m(x_i)|^2 \end{aligned}$$

implies

$$\frac{t}{2} < \sum_{i=1}^{n+N_n} w_i \cdot |\bar{m}_n(x_i) - m_n^*(x_i)|^2 \leq 16 \sum_{i=1}^{n} w_i \cdot (\bar{m}_n(x_i) - m_n^*(x_i)) \cdot (y_i - m(x_i)).$$

Lemma 1 follows immediately from a straightforward generalization of Lemma 1 in Kohler (2006). For the sake of completeness we present a complete proof in the Appendix.

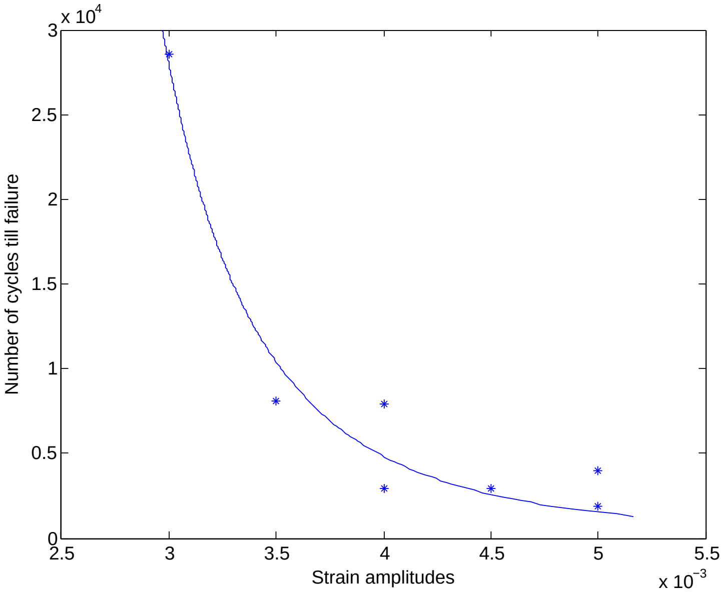

Figure 5: Observed real data points and the corresponding estimate

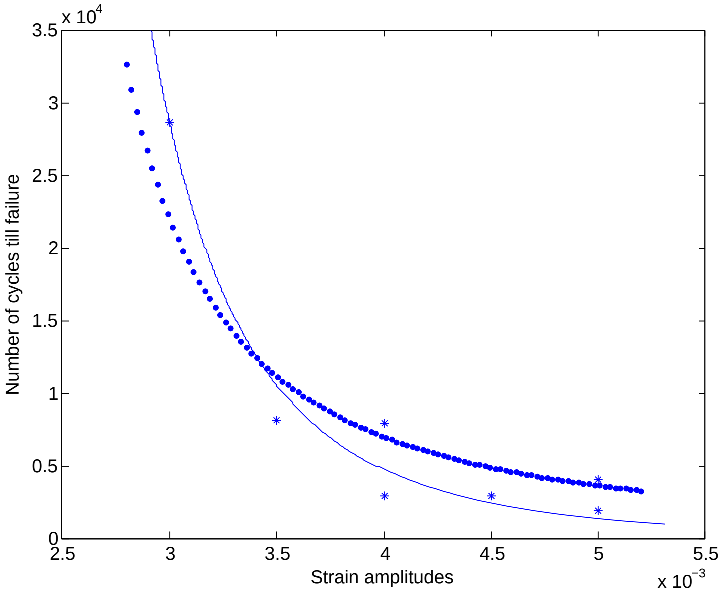

Figure 6: Observed real data points (*), artificial data points (.) and the corresponding estimate

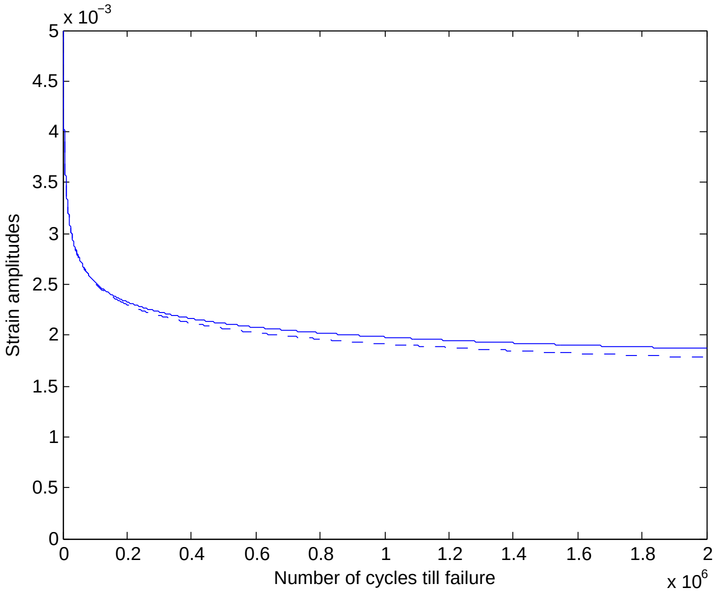

Figure 7: Comparison of the two estimates. Solid line is the estimate based only on real data points, dashed line shows the estimate based on real and artificial data.

4.2 Results for fixed design regression

In this subsection we bound the empirical L_2 error of $\bar{m}_n$

$$\sum_{i=1}^{n+N_n} w_i \cdot |\bar{m}_n(x_i) - m(x_i)|^2$$

by the sum of the approximation error

$$\min_{f \in \mathcal{F}_n} \left(\sum_{i=1}^{n+N_n} w_i \cdot |f(x_i) - m(x_i)|^2 \right),$$

a term which depends on the complexity of $\mathcal{F}_n$ measured by covering numbers and the error of the artificial data described by (8).

Lemma 2. *Let $\bar{m}_n$ be the estimate defined by (4) and (5). There exist constants $c_3, c_4 > 0$ which depend only on σ_0 and K such that for any $\delta_n > 0$ with*

$$\delta_n \to 0 \quad (n \to \infty) \quad \textit{and} \quad n \cdot \delta_n \to \infty \quad (n \to \infty)$$

and

$$\sqrt{n} \cdot \delta \geq c_3 \int_{\delta/(2^9 \sigma_0)}^{\sqrt{\delta}} \left(\log \mathcal{N}_2 \left(u, \{ f - g : f \in \mathcal{F}_n, \frac{1}{n} \sum_{i=1}^n |f(x_i) - g(x_i)|^2 \leq \delta \}, x_1^n \right) \right)^{1/2} du \quad (13)$$

for all $\delta \geq \delta_n$ and all $g \in \mathcal{F}_n$ we have

$$\begin{aligned} &\mathbf{P} \Bigg\{ \sum_{i=1}^{n+N_n} w_i \cdot |\bar{m}_n(x_i) - m(x_i)|^2 \\ &\qquad > c_4 \Bigg(\sum_{j=1}^{N_n} w_{n+j} \cdot |\hat{Y}_{n+j} - m(x_{n+j})|^2 + w^{(n)} \cdot \delta_n + \min_{f \in \mathcal{F}_n} \sum_{i=1}^{n+N_n} w_i \cdot |f(x_i) - m(x_i)|^2 \Bigg) \Bigg\} \\ &\to 0 \quad (n \to \infty). \end{aligned}$$

Proof. Set

$$m_n^*(\cdot) = \arg \min_{f \in \mathcal{F}_n} \left(\sum_{i=1}^{n+N_n} w_i \cdot |f(x_i) - m(x_i)|^2 \right)$$

By Lemma 1,

$$\begin{aligned} &\mathbf{P} \Bigg\{ \sum_{i=1}^{n+N_n} w_i \cdot |\bar{m}_n(x_i) - m(x_i)|^2 > w^{(n)} \cdot \delta_n + 512 \sum_{j=1}^{N_n} w_{n+j} \cdot |m(x_{n+j}) - \hat{Y}_{n+j}|^2 \\ &\qquad\qquad + 18 \min_{f \in \mathcal{F}_n} \sum_{i=1}^{n+N_n} w_i \cdot |f(x_i) - m(x_i)|^2 \Bigg\} \\ &\leq \mathbf{P} \left\{ w^{(n)} \cdot \frac{\delta_n}{2} < \sum_{i=1}^{n+N_n} w_i \cdot |\bar{m}_n(x_i) - m_n^*(x_i)|^2 \leq 16 \sum_{i=1}^{n} w_i \cdot (\bar{m}_n(x_i) - m_n^*(x_i)) \cdot W_i \right\} \\ &\leq P_1 + P_2, \end{aligned}$$

where

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

and

$$\begin{aligned} P_2 = \quad & \mathbf{P}\left\{ \frac{1}{n}\sum_{i=1}^n W_i^2 \leq 2\sigma_0^2,\, w^{(n)} \cdot \frac{\delta_n}{2} < \sum_{i=1}^{n+N_n} w_i \cdot |\bar{m}_n(x_i) - m_n^*(x_i)|^2 \leq \right. \\ & \left. 16 \sum_{i=1}^n w_i \cdot (\bar{m}_n(x_i) - m_n^*(x_i)) \cdot W_i \right\}. \end{aligned}$$

Application of Chernoff's exponential bounding technique (cf., Chernoff (1952)) together with (7) yield

$$\begin{aligned} P_1 &= \mathbf{P}\left\{ \frac{1}{n}\sum_{i=1}^n W_i^2/K^2 > 2\sigma_0^2/K^2 \right\} \\ &\leq \mathbf{P}\left\{ \exp\left(\sum_{i=1}^n W_i^2/K^2 \right) > \exp\left(2n \cdot \sigma_0^2/K^2 \right) \right\} \\ &\leq \exp\left(-2n \cdot \sigma_0^2/K^2 \right) \cdot \mathbf{E}\left\{ \exp\left(\sum_{i=1}^n W_i^2/K^2 \right) \right\} \\ &\leq \exp\left(-2n \cdot \sigma_0^2/K^2 \right) \cdot \left(1 + \sigma_0^2/K^2 \right)^n \\ &\leq \exp\left(-2n \cdot \sigma_0^2/K^2 \right) \cdot \exp\left(n \cdot \sigma_0^2/K^2 \right) \\ &= \exp\left((-n) \cdot \sigma_0^2/K^2 \right) \to 0 \quad (n \to \infty). \end{aligned}$$

From the definition of w_i we conclude

$$\begin{aligned} P_2 \leq \quad & \mathbf{P}\left\{ \frac{1}{n}\sum_{i=1}^n W_i^2 \leq 2\sigma_0^2,\, w^{(n)} \cdot \frac{\delta_n}{2} < 16 \sum_{i=1}^n w_i \cdot (\bar{m}_n(x_i) - m_n^*(x_i)) \cdot W_i, \right. \\ & \left. \sum_{i=1}^n w_i \cdot |\bar{m}_n(x_i) - m_n^*(x_i)|^2 \leq 16 \sum_{i=1}^n w_i \cdot (\bar{m}_n(x_i) - m_n^*(x_i)) \cdot W_i \right\} \\ & \leq \mathbf{P}\left\{ \frac{1}{n}\sum_{i=1}^n W_i^2 \leq 2\sigma_0^2, \frac{\delta_n}{2} < 16 \cdot \frac{1}{n}\sum_{i=1}^n (\bar{m}_n(x_i) - m_n^*(x_i)) \cdot W_i, \right. \\ & \left. \frac{1}{n}\sum_{i=1}^n |\bar{m}_n(x_i) - m_n^*(x_i)|^2 \leq 16 \cdot \frac{1}{n}\sum_{i=1}^n (\bar{m}_n(x_i) - m_n^*(x_i)) \cdot W_i \right\}. \end{aligned} \qquad (14)$$

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

$$\begin{aligned} & 16 \cdot \frac{1}{n}\sum_{i=1}^n (\bar{m}_n(x_i) - m_n^*(x_i)) \cdot W_i \\ \leq \quad & 16 \cdot \sqrt{\frac{1}{n}\sum_{i=1}^n (\bar{m}_n(x_i) - m_n^*(x_i))^2} \cdot \sqrt{\frac{1}{n}\sum_{i=1}^n W_i^2} \end{aligned}$$

$$\leq 16 \cdot \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\bar{m}_n(x_i) - m_n^*(x_i))^2} \cdot \sqrt{2\sigma_0^2},$$

hence inside probability (14) we have

$$\frac{1}{n}\sum_{i=1}^{n}(\bar{m}_n(x_i) - m_n^*(x_i))^2 \leq 512\sigma_0^2.$$

Set

$$S = \min\{s \in \mathbb{N}_0 \, : \, 2^s\delta_n \geq 512\sigma_0^2\}.$$

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

$$\begin{aligned} P_2 \leq \; & \sum_{s=0}^{S} \mathbf{P}\Bigg\{\frac{1}{n}\sum_{i=1}^{n} W_i^2 \leq 2\sigma_0^2, 2^{s-1}\delta_n \cdot I_{\{s>0\}} < \frac{1}{n}\sum_{i=1}^{n}|\bar{m}_n(x_i) - m_n^*(x_i)|^2 \leq 2^s\delta_n, \\ & \qquad \max\left\{\frac{\delta_n}{2}, \frac{1}{n}\sum_{i=1}^{n}|\bar{m}_n(x_i) - m_n^*(x_i)|^2\right\} \leq 16\frac{1}{n}\sum_{i=1}^{n}(\bar{m}_n(x_i) - m_n^*(x_i)) \cdot W_i\Bigg\} \\ \leq \; & \sum_{s=0}^{S} \mathbf{P}\Bigg\{\frac{1}{n}\sum_{i=1}^{n} W_i^2 \leq 2\sigma_0^2, \frac{1}{n}\sum_{i=1}^{n}|\bar{m}_n(x_i) - m_n^*(x_i)|^2 \leq 2^s\delta_n, \\ & \qquad \frac{1}{n}\sum_{i=1}^{n}(\bar{m}_n(x_i) - m_n^*(x_i)) \cdot W_i > \frac{2^s\delta_n}{32}\Bigg\}. \end{aligned}$$

The probabilities in the above sum can be bounded using Corollary 8.3 in van de Geer (2000) (use there $R = \sqrt{2^s\delta_n}, \delta = \frac{2^s\delta_n}{32}, \sigma = \sqrt{2\sigma_0}$), which yields

$$\begin{aligned} P_2 &\leq \sum_{s=0}^{S} c_5 \exp\left(-\frac{n \cdot (2^s\delta_n/32)^2}{4c_5 \cdot 2^s\delta_n}\right) = \sum_{s=0}^{S} c_5 \exp\left(-\frac{n \cdot 2^s \cdot \delta_n}{4 \cdot 32^2 \cdot c_5}\right) \\ &\leq c_6 \exp\left(-\frac{n \cdot \delta_n}{c_6}\right) \to 0 \end{aligned}$$

for $n \to \infty$. □

4.3 Approximating integrals by averages

In this subsection we state the following auxiliary result which we will use to bound the difference between L_2 error and the empirical L_2 error.

Lemma 3. *Let $f : [0,1] \to \mathbb{R}$ be Lipschitz-continuous with Lipschitz constant L and let $N \in \mathbb{N}$. Then*

$$\left|\int_0^1 f(x)\,dx - \frac{1}{N}\sum_{i=1}^{N} f(i/N)\right| \leq \frac{1}{2} \cdot L \cdot \frac{1}{N}.$$

Proof. Since f is Lipschitz-continuous we have

$$\begin{aligned}\left|\int_0^1 f(x)\,dx - \frac{1}{N}\sum_{i=1}^N f(i/N)\right| &\leq \sum_{i=1}^N \left|\int_{(i-1)/N}^{i/N} \left(f(x)-f(i/N)\right)\,dx\right| \\ &\leq \sum_{i=1}^N \int_{(i-1)/N}^{i/N} \left|f(x)-f(i/N)\right|\,dx \\ &\leq L\cdot\sum_{i=1}^N \int_{(i-1)/N}^{i/N} \left(\frac{i}{N}-x\right)\,dx = \frac{1}{2}\cdot L\cdot\frac{1}{N}.\end{aligned}$$

□.

4.4 Proof of Theorem 1

Using the definition of w_i we get

$$\begin{aligned}&\int_0^1 |\bar{m}_n(x)-m(x)|^2dx \\ &= \int_0^1 |\bar{m}_n(x)-m(x)|^2dx - \sum_{i=1}^N w_i\cdot|\bar{m}_n(x_i)-m(x_i)|^2 + \sum_{i=1}^N w_i\cdot|\bar{m}_n(x_i)-m(x_i)|^2 \\ &= w^{(n)}\cdot\left(\int_0^1 |\bar{m}_n(x)-m(x)|^2dx - \frac{1}{n}\sum_{i=1}^n |\bar{m}_n(x_i)-m(x_i)|^2\right) \\ &\quad +(1-w^{(n)})\cdot\left(\int_0^1 |\bar{m}_n(x)-m(x)|^2dx - \frac{1}{N_n}\sum_{i=1}^{N_n} |\bar{m}_n(x_i)-m(x_i)|^2\right) \\ &\quad +\sum_{i=1}^N w_i\cdot|\bar{m}_n(x_i)-m(x_i)|^2 \\ &=: T_{1,n}+T_{2,n}+T_{3,n}.\end{aligned}$$

W.l.o.g. m is Lipschitz-continuous on $[0,1]$ with Lipschitz constant bounded by $\log n$ and m is bounded in absolute value by $\log n$. Consequently

$$\begin{aligned}&\left||\bar{m}_n(x)-m(x)|^2-|\bar{m}_n(z)-m(z)|^2\right| \\ &= |\bar{m}_n(x)-m(x)-\bar{m}_n(z)+m(z)|\cdot|\bar{m}_n(x)-m(x)+\bar{m}_n(z)-m(z)| \\ &\leq 2\cdot\log n\cdot|x-z|\cdot(2\cdot\|\bar{m}_n\|_\infty+2\cdot\|m\|_\infty),\end{aligned}$$

hence $g_n(x)=|\bar{m}_n(x)-m(x)|^2$ is Lipschitz-continuous with Lipschitz constant bounded by $8\log^2 n$. By Lemma 3 we conclude

$$T_{1,n}\leq 4\cdot\log^2(n)\cdot\frac{w^{(n)}}{n}$$

and

$$T_{2,n}\leq 4\cdot\log^2(n)\cdot\frac{1-w^{(n)}}{N_n}.$$

Application of Lemma 2 to $T_{3,n}$ yields the assertion. □.

Acknowledgement.

The authors would like to thank the German Research Foundation (DFG) for founding this project within the Collaborative Research Center 666 and Natural Sciences and Engineering Research Council of Canada for providing an additional research support.

References

[1] Beirlant, J. and Györfi, L. (1998). On the asymptotic L_2-error in partitioning regression estimation. *Journal of Statistical Planning and Inference*, **71**, pp. 93–107.

[2] Chernoff, H. (1952). A measure of asymptotic efficiency of tests of a hypothesis based on the sum of observations. *Annals of the Mathematical Statistics*, **23**, pp. 493–507.

[3] Devroye, L. (1982). Necessary and sufficient conditions for the almost everywhere convergence of nearest neighbor regression function estimates. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, **61**, pp. 467–481.

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

[5] Devroye, L. and Krzyżak, A. (1989). An equivalence theorem for L_1 convergence of the kernel regression estimate. *Journal of Statistical Planning and Inference*, **23**, pp. 71–82.

[6] Devroye, L. and Wagner, T. J. (1980). Distribution-free consistency results in nonparametric discrimination and regression function estimation. *Annals of Statistics*, **8**, pp. 231–239.

[7] Eubank, R. L. (1999). *Nonparametric Regression and Spline Smoothing.* 2nd edition, Marcel Dekker, New York.

[8] Fromkorth, A. and Kohler, M. (2011). Analysis of least squares regression estimates in case of additional errors in the variables. *Journal of Statistical Planning and Inference*, **141**, pp. 172-188.

[9] Gasser, T. and Müller, M.-H. (1979). Kernel estimation of regression functions. In *Smoothing Techniques for Curve Estimation*, Gasser, T. and Rosenblatt, M., Eds., pp. 23-68. Lecture Notes in Mathematics 757, Springer-Verlag, Heidelberg.

[10] Györfi, L. (1981). Recent results on nonparametric regression estimate and multiple classification. *Problems of Control and Information Theory*, **10**, pp. 43–52.

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

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

[13] Kohler, M. and Krzyżak, A. (2001). Nonparametric regression estimation using penalized least squares. *IEEE Transactions on Information Theory*, **47**, pp. 3054–3058.

[14] Lugosi, G. and Zeger, K. (1995). Nonparametric estimation via empirical risk minimization. *IEEE Transactions on Information Theory*, **41**, pp. 677-687.

[15] Mack, Y. P. (1981). Local properties of k–nearest neighbor regression estimates. *SIAM Journal on Algebraic and Discrete Methods*, **2**, pp. 311–323.

[16] Manson, S. S. (1965). Fatigue: A complex subject - some simple approximation. *Experimental Mechanics*, **5**, pp. 193-226.

[17] Nadaraya, E. A. (1964) On estimating regression. *Theory of Probability and its Applications*, **9**, pp. 141–142.

[18] Nadaraya, E. A. (1970). Remarks on nonparametric estimates for density functions and regression curves. *Theory of Probability and its Applications*, **15**, pp. 134–137.

[19] Ramberg, W. and Osgood, W.R. (1943). Description of stress-strain curves by three parameters. *Technical Note*, **902**, *National Advisory Committee for Aeronautics,Washington DC*, pp. 1-28.

[20] Stone, C. J. (1977). Consistent nonparametric regression. *Annals of Statistics*, **5**, pp. 595–645.

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

[22] Watson, G. S. (1964). Smooth regression analysis. *Sankhya Series A*, **26**, pp. 359–372.

[23] Zhao, L. C. (1987). Exponential bounds of mean error for the nearest neighbor estimates of regression functions. *Journal of Multivariate Analysis*, **21**, pp. 168–178.

Appendix: Proof of Lemma 1

Lemma 1 follows immediately from the following generalization of Lemma 1 in Kohler (2006).

Lemma 4. *Let $t > 0$, $w_1, \dots, w_n \in \mathbb{R}_+$, $x_1, \dots, x_N \in \mathbb{R}^d$ and $y_1, \bar{y}_1, \dots, y_N, \bar{y}_N \in \mathbb{R}$. Let m be a function $m : \mathbb{R}^d \to \mathbb{R}$ and let $\mathcal{F}$ be a set of functions $f : \mathbb{R}^d \to \mathbb{R}$. Set*

$$\bar{m}_n = \arg\min_{f \in \mathcal{F}} \sum_{i=1}^N w_i \cdot |f(x_i) - \bar{y}_i|^2$$

and

$$m_n^* = \arg\min_{f \in \mathcal{F}} \sum_{i=1}^N w_i \cdot |f(x_i) - m(x_i)|^2$$

and assume that both minima exist. Then

$$\sum_{i=1}^N w_i \cdot |\bar{m}_n(x_i) - m(x_i)|^2 > t + 512 \sum_{i=1}^N w_i \cdot |y_i - \bar{y}_i|^2 + 18 \min_{f \in \mathcal{F}} \sum_{i=1}^N w_i \cdot |f(x_i) - m(x_i)|^2 \quad (15)$$

implies

$$\frac{t}{2} < \sum_{i=1}^{N} w_i \cdot |\bar{m}_n(x_i) - m_n^*(x_i)|^2 \leq 16 \sum_{i=1}^{N} w_i \cdot (\bar{m}_n(x_i) - m_n^*(x_i)) \cdot (y_i - m(x_i)). \qquad (16)$$

For the sake of completeness we present next a complete proof of Lemma 4.

Proof of Lemma 4. The proof is divided into four steps. *In the first step of the proof* we show that (15) implies

$$\sum_{i=1}^{N} w_i \cdot |\bar{m}_n(x_i) - m_n^*(x_i)|^2 > \frac{t}{2} + 256 \sum_{i=1}^{N} w_i \cdot |y_i - \bar{y}_i|^2 + 8 \sum_{i=1}^{N} w_i \cdot |m_n^*(x_i) - m(x_i)|^2. \quad (17)$$

Indeed, by definition of m_n^* we have

$$\sum_{i=1}^{N} w_i \cdot |m_n^*(x_i) - m(x_i)|^2 = \min_{f \in \mathcal{F}} \sum_{i=1}^{N} w_i \cdot |f(x_i) - m(x_i)|^2,$$

which together with (15) and

$$\sum_{i=1}^{N} w_i \cdot |\bar{m}_n(x_i) - m(x_i)|^2 \leq 2 \sum_{i=1}^{N} w_i \cdot |\bar{m}_n(x_i) - m_n^*(x_i)|^2 + 2 \sum_{i=1}^{N} w_i \cdot |m_n^*(x_i) - m(x_i)|^2$$

implies

$$2 \sum_{i=1}^{N} w_i \cdot |\bar{m}_n(x_i) - m_n^*(x_i)|^2 > t + 512 \sum_{i=1}^{N} w_i \cdot |y_i - \bar{y}_i|^2 + 16 \sum_{i=1}^{N} w_i \cdot |m_n^*(x_i) - m(x_i)|^2.$$

This is equivalent to (17).

In the second step of the proof we show

$$\sum_{i=1}^{N} w_i \cdot |\bar{m}_n(x_i) - m_n^*(x_i)|^2$$
$$\leq 4 \sum_{i=1}^{N} w_i \cdot |m_n^*(x_i) - m(x_i)|^2 + 4 \sum_{i=1}^{N} w_i \cdot (\bar{m}_n(x_i) - m_n^*(x_i)) \cdot (\bar{y}_i - m(x_i)). \quad (18)$$

We have

$$\sum_{i=1}^{N} w_i \cdot |\bar{m}_n(x_i) - m_n^*(x_i)|^2 \leq 2 \sum_{i=1}^{N} w_i \cdot |\bar{m}_n(x_i) - m(x_i)|^2 + 2 \sum_{i=1}^{N} w_i \cdot |m_n^*(x_i) - m(x_i)|^2 \quad (19)$$

and

$$\sum_{i=1}^{N} w_i \cdot |\bar{m}_n(x_i) - \bar{y}_i|^2$$
$$= \sum_{i=1}^{N} w_i \cdot |\bar{m}_n(x_i) - m(x_i)|^2 + \sum_{i=1}^{N} w_i \cdot |m(x_i) - \bar{y}_i|^2 + 2 \sum_{i=1}^{N} w_i \cdot (\bar{m}_n(x_i) - m(x_i)) \cdot (m(x_i) - \bar{y}_i),$$

which implies

$$\begin{aligned}
&\sum_{i=1}^N w_i \cdot |\bar{m}_n(x_i) - m_n^*(x_i)|^2 \\
&\leq 2\left(\sum_{i=1}^N w_i \cdot |\bar{m}_n(x_i) - \bar{y}_i|^2 - \sum_{i=1}^N w_i \cdot |m(x_i) - \bar{y}_i|^2 - 2\sum_{i=1}^N w_i \cdot (\bar{m}_n(x_i) - m(x_i)) \cdot (m(x_i) - \bar{y}_i)\right) \\
&\quad +2\sum_{i=1}^N w_i \cdot |m_n^*(x_i) - m(x_i)|^2.
\end{aligned}$$

By definition of $\bar{m}_n$

$$\begin{aligned}
\sum_{i=1}^N w_i \cdot |\bar{m}_n(x_i) - \bar{y}_i|^2 &\leq \sum_{i=1}^N w_i \cdot |m_n^*(x_i) - \bar{y}_i|^2 \\
&= \sum_{i=1}^N w_i \cdot |m_n^*(x_i) - m(x_i)|^2 + \sum_{i=1}^N w_i \cdot |m(x_i) - \bar{y}_i|^2 \\
&\quad +2\sum_{i=1}^N w_i \cdot (m_n^*(x_i) - m(x_i)) \cdot (m(x_i) - \bar{y}_i).
\end{aligned}$$

This together with the previous inequality and (19) yields (18).

In the third step of the proof we show that (17) implies

$$\sum_{i=1}^N w_i \cdot |m_n^*(x_i) - m(x_i)|^2 \leq \sum_{i=1}^N w_i \cdot (\bar{m}_n(x_i) - m_n^*(x_i)) \cdot (\bar{y}_i - m(x_i)). \tag{20}$$

To do this, we assume that (20) does not hold and show that (17) does not hold. Indeed, if (20) does not hold then we can conclude from (18)

$$\sum_{i=1}^N w_i \cdot |\bar{m}_n(x_i) - m_n^*(x_i)|^2 \leq 8\sum_{i=1}^N w_i \cdot |m_n^*(x_i) - m(x_i)|^2,$$

which implies that (17) does not hold.

In the fourth step of the proof we show that (17) and (20) imply (16). To do this, we conclude from (18) and (20)

$$\sum_{i=1}^N w_i \cdot |\bar{m}_n(x_i) - m_n^*(x_i)|^2 \leq 8\sum_{i=1}^N w_i \cdot (\bar{m}_n(x_i) - m_n^*(x_i)) \cdot (\bar{y}_i - m(x_i)),$$

which together with (17) yields

$$\begin{aligned}
\frac{t}{2} + 256\sum_{i=1}^N w_i \cdot |y_i - \bar{y}_i|^2 &< \sum_{i=1}^N w_i \cdot |\bar{m}_n(x_i) - m_n^*(x_i)|^2 \\
&\leq 8\sum_{i=1}^N w_i \cdot (\bar{m}_n(x_i) - m_n^*(x_i)) \cdot (\bar{y}_i - m(x_i)). \qquad (21)
\end{aligned}$$

If

$$\sum_{i=1}^{N} w_i \cdot (\bar{m}_n(x_i) - m_n^*(x_i)) \cdot (\bar{y}_i - y_i) \leq \sum_{i=1}^{N} w_i \cdot (\bar{m}_n(x_i) - m_n^*(x_i)) \cdot (y_i - m(x_i)) \qquad (22)$$

holds, then we have

$$\begin{aligned}
&\sum_{i=1}^{N} w_i \cdot (\bar{m}_n(x_i) - m_n^*(x_i)) \cdot (\bar{y}_i - m(x_i)) \\
&= \sum_{i=1}^{N} w_i \cdot (\bar{m}_n(x_i) - m_n^*(x_i)) \cdot (\bar{y}_i - y_i) + \sum_{i=1}^{N} w_i \cdot (\bar{m}_n(x_i) - m_n^*(x_i)) \cdot (y_i - m(x_i)) \\
&\leq 2 \sum_{i=1}^{N} w_i \cdot (\bar{m}_n(x_i) - m_n^*(x_i)) \cdot (y_i - m(x_i))
\end{aligned}$$

which together with (21) implies (16). We conclude the proof by showing that we get a contradiction, if (22) doesn't hold. So assume, that (22) does not hold. Then (21) together with the Cauchy-Schwarz inequality implies

$$\begin{aligned}
&\sum_{i=1}^{N} w_i \cdot |\bar{m}_n(x_i) - m_n^*(x_i)|^2 \\
&\leq 8 \sum_{i=1}^{N} w_i \cdot (\bar{m}_n(x_i) - m_n^*(x_i)) \cdot (\bar{y}_i - y_i) + 8 \sum_{i=1}^{N} w_i \cdot (\bar{m}_n(x_i) - m_n^*(x_i)) \cdot (y_i - m(x_i)) \\
&\leq 16 \sum_{i=1}^{N} w_i \cdot (\bar{m}_n(x_i) - m_n^*(x_i)) \cdot (\bar{y}_i - y_i) \\
&\leq 16 \sqrt{\sum_{i=1}^{N} w_i \cdot (\bar{m}_n(x_i) - m_n^*(x_i))^2} \cdot \sqrt{\sum_{i=1}^{N} w_i \cdot (\bar{y}_i - y_i)^2},
\end{aligned}$$

which in turn implies

$$\sqrt{\sum_{i=1}^{N} w_i \cdot (\bar{m}_n(x_i) - m_n^*(x_i))^2} \leq 16 \sqrt{\sum_{i=1}^{N} w_i \cdot (\bar{y}_i - y_i)^2}.$$

From this together with (21) it follows

$$\frac{t}{2} + 256 \sum_{i=1}^{N} w_i \cdot |y_i - \bar{y}_i|^2 < 256 \sum_{i=1}^{N} w_i \cdot |y_i - \bar{y}_i|^2,$$

which is the desired contradiction. □

Proof of Lemma 1. Set $N = n + N_n$, and for $i \in \{1, \ldots, N\}$ choose

$$\bar{y}_i = y_i \quad \text{for} \quad i \leq n \quad \text{and} \quad \bar{y}_i = \hat{y}_i \quad \text{for} \quad i > n$$

and

$$y_i = y_i \quad \text{for} \quad i \leq n \quad \text{and} \quad y_i = m(x_i) \quad \text{for} \quad i > n$$

in Lemma 4. Then we immediately get the assertion of Lemma 1. □