

Analysis of the rate of convergence of least squares neural network regression estimates in case of measurement errors

Michael Kohler and Jens Mehnert*

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

October 1, 2009

Abstract

Estimation of a regression function from data which consists of an independent and identically distributed sample of the underlying distribution with additional measurement errors in the independent variables is considered. It is allowed that the measurement errors are not independent and have nonzero mean. It is shown that the rate of convergence of suitably defined least squares neural network estimates applied to this data is similar to the rate of convergence of least squares neural network estimates applied to an independent and identically distributed sample of the underlying distribution as long as the measurement errors are small.

AMS classification: Primary 62G08; secondary 92B20.

Key words and phrases: least squares estimates, measurement error, neural networks, rate of convergence, regression estimates, L_2 error.

*Corresponding author. Tel: +49-6151-16-5288, Fax: +49-6151-16-6822

Running title: *Rate of convergence in case of measurement errors*

1 Introduction

Let $(X, Y), (X_1, Y_1), (X_2, Y_2), \dots$ be independent identically distributed $\mathbb{R}^d \times \mathbb{R}$ - valued random vectors with $\mathbf{E}Y^2 < \infty$. In regression analysis we want to estimate Y after having observed X , i.e., we want to determine a function f with $f(X)$ “close” to Y . If “closeness” is measured by the mean squared error, then one wants to find a function f^* minimizing the so-called L_2 -risk $\mathbf{E} \left\{ |f^*(X) - Y|^2 \right\}$, i.e., f^* should satisfy

$$\mathbf{E} \left\{ |f^*(X) - Y|^2 \right\} = \min_f \mathbf{E} \left\{ |f(X) - Y|^2 \right\}. \quad (1)$$

Let $m(x) := \mathbf{E}\{Y|X = x\}$ be the regression function. The well-known relation which holds for each measurable function f

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

implies that m is the solution of the minimization problem (1), $\mathbf{E}\{|m(X) - Y|^2\}$ is the minimum of (2) and for an arbitrary f , the L_2 error $\int |f(x) - m(x)|^2 \mathbf{P}_X(dx)$ is the difference between $\mathbf{E}\{|f(X) - Y|^2\}$ and $\mathbf{E}\{|m(X) - Y|^2\}$.

In the regression estimation problem the distribution of (X, Y) (and consequently m) is unknown. Given a sequence $\mathcal{D}_n = \{(X_1, Y_1), \dots, (X_n, Y_n)\}$ of independent observations of (X, Y) , the goal is to construct an estimate $m_n(x) = m_n(x, \mathcal{D}_n)$ of $m(x)$ such that the L_2 error $\int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx)$ is small. For a general introduction to regression estimation see, e.g., Györfi et al. (2002).

Sometimes it is possible to observe data from the underlying distribution only with measurement errors. In this context usually the problem is considered that the independent variable X can be observed only with additional random errors which have mean zero. More precisely, instead of X_i one observes $W_i = X_i + U_i$ for some random variables U_i which satisfy $\mathbf{E}\{U_i|X_i\} = 0$, and the problem is to estimate the regression function from $\{(W_1, Y_1), \dots, (W_n, Y_n)\}$. In the literature in this context often estimates of the distribution of U_i are constructed and estimates of the regression function are defined by using the estimated distribution of U_i (see, e.g., Fan and Truong (1993), Carroll, Maca and Ruppert (1999), Delaigle and Meister (2007), Delaigle, Fan and Carroll (2009) and the references therein).

In this paper we consider a setting, where basically nothing is assumed on the nature of the measurement errors. In particular, the measurement errors do not have to be independent or identically distributed, and they do not need to have expectation zero. The only assumption we are making is that these measurement errors are somehow “small”.

More precisely, we assume that we have given data

$$\bar{\mathcal{D}}_n = \{(\bar{X}_{1,n}, Y_1), \dots, (\bar{X}_{n,n}, Y_n)\},$$

where the only assumption on the random variables $\bar{X}_{1,n}, \dots, \bar{X}_{n,n}$ is that the average measurement error

$$\frac{1}{n} \sum_{i=1}^n \|X_i - \bar{X}_{i,n}\|_2 \tag{3}$$

is small, where $\|\cdot\|_2$ denotes the Euclidean norm. In particular, $\bar{\mathcal{D}}_n$ does not need to be independent or identically distributed, and $\mathbf{E}\{Y_1|X_{1,n} = x\}$ does not need to be equal to $m(x) = \mathbf{E}\{Y|X = x\}$. For notational simplicity we will suppress in the sequel a possible dependency of $\bar{X}_i = \bar{X}_{i,n}$ on the sample size n in our notation.

It is not clear how the L_2 error of an arbitrary regression estimate is influenced by additional measurement errors. Due to the fact that we assume nothing on the nature of these errors, in contrast to the classical setting described above there is now no chance to get rid of these errors, so these errors will necessarily increase the L_2 error of the estimate. Intuitively one can expect that measurement errors influence the error of the estimate not much as long as these measurement errors are small. In this article we show that this is indeed true for suitably defined least squares neural network estimates.

The basic idea behind the definition of our estimate is as follows: Since we assume that (3) is small, it is reasonable to estimate the L_2 risk of a Lipschitz continuous function f by the so-called empirical L_2 risk

$$\frac{1}{n} \sum_{i=1}^n |f(\bar{X}_i) - Y_i|^2$$

computed with the aid of the data with measurement error, and to define least squares estimates as if no measurement errors are present by

$$\bar{m}_n(\cdot) = \arg \min_{f \in \mathcal{F}_n} \frac{1}{n} \sum_{i=1}^n |f(\bar{X}_i) - Y_i|^2 \tag{4}$$

for some set \mathcal{F}_n of Lipschitz continuous functions $f : \mathbb{R}^d \rightarrow \mathbb{R}$. Here $z = \arg \min_{x \in A} G(x)$ is an abbreviation for $z \in A$ and $G(z) = \min_{x \in A} G(x)$ and we assume for simplicity that the minima in (4) exist, however we do not require them to be unique.

In this article we will use for \mathcal{F}_n suitably defined sets of neural networks. Our main result is that if we restrict the weights of the neural networks such that the resulting functions are Lipschitz continuous with respect to some Lipschitz constant depending on the sample size, then the L_2 error of the corresponding least squares neural network regression estimates applied to data with additional measurement errors in the independent variables is basically the sum of the usually error bound for such an estimate applied to data without measurement errors and the product of the measurement error (3) and the Lipschitz constant.

1.1 Notation

The sets of natural, real numbers and d -dimensional real numbers are denoted by \mathbb{N} , \mathbb{R} and \mathbb{R}^d , respectively. For $x \in \mathbb{R}^d$ we denote by $\|x\|_2$ the Euclidian norm of x . The least integer greater than or equal to a real number x will be denoted by $\lceil x \rceil$. For a function $f : \mathbb{R}^d \rightarrow \mathbb{R}$

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

denotes the supremum norm. I_A is the indicator function of a set A , and $|\mathcal{Q}|$ is the cardinality of a set \mathcal{Q} . For $z \in \mathbb{R}$ and $\beta > 0$ we define

$$T_\beta z = \min\{\max\{z, -\beta\}, \beta\}.$$

1.2 Outline

The definition of the estimate is given in Section 2, the main result is formulated in Section 3. Section 4 contains the proofs.

2 Definition of the least squares neural network regression estimates

A feedforward neural network with one hidden layer and k hidden neurons is a real-valued function on \mathbb{R}^d of the form

$$f(x) = \sum_{i=1}^k c_i \cdot \sigma(a_i^T x + b_i) + c_0 \quad (5)$$

where $\sigma : \mathbb{R} \rightarrow [0, 1]$ is called a sigmoidal function and $a_1, \dots, a_k \in \mathbb{R}^d$, b_1, \dots, b_k , $c_0, c_1, \dots, c_k \in \mathbb{R}$ are the parameters that specify the network. For the sigmoidal function σ one often uses so-called squashing functions, i.e. a function which is non-decreasing and satisfies

$$\lim_{x \rightarrow -\infty} \sigma(x) = 0 \quad \text{and} \quad \lim_{x \rightarrow \infty} \sigma(x) = 1.$$

It is well-known that feedforward neural networks with one hidden layer are dense on compact sets with respect to the supremum norm in the set of continuous functions. In other words, every continuous function on \mathbb{R}^d can be approximated arbitrarily close uniformly over any compact set by functions realized by neural networks, see, e.g., Cybenko (1989), Hornik, Stinchcombe and White (1989), and Funahashi (1989). For a survey of such denseness results we refer the reader to Barron (1989) and Hornik (1993).

Motivated by these approximation results neural networks have been applied to various estimation problems, see, e.g., the monographs Hertz, Krogh and Palmer (1991), Devroye, Györfi and Lugosi (1996), Ripley (1996), Anthony and Bartlett (1999) and Györfi et al. (2002). The papers Barron (1991, 1993), Miłniczuk and Trycha (1993), McCaffrey and Gallant (1994), Lugosi and Zeger (1995), Kohler and Krzyżak (2005) and Hamers and Kohler (2006) contain various theoretical results concerning regression estimation with neural networks in case that measurement errors do not occur. In particular it follows from Barron (1993) that in case of regression functions for which the Fourier transform has a finite first moment the L_2 error of suitably defined neural network estimates converges (up to some logarithmic factor) to zero with rate $n^{-1/2}$ (cf., e.g., Section 16.3 in Györfi et al. (2002)).

The aim of this paper is to show a similar result in case of “small” measurement errors.

To do this we will use as sigmoidal function the so-called logistic squasher

$$\sigma(x) = \frac{1}{1 + \exp(-x)} \quad (x \in \mathbb{R}).$$

We have

$$\sigma'(x) = \frac{\exp(-x)}{(1 + \exp(-x))^2} = \frac{1}{\exp(x) + 2 + \exp(-x)} \in [0, 1],$$

hence (5) satisfies for any $x, z \in \mathbb{R}^d$

$$\begin{aligned} |f(x) - f(z)| &= \left| \sum_{i=1}^k c_i \cdot (\sigma(a_i^T x + b_i) - \sigma(a_i^T z + b_i)) \right| \\ &\leq \sum_{i=1}^k |c_i| \cdot |a_i^T x + b_i - (a_i^T z + b_i)| \\ &\leq \sum_{i=1}^k |c_i| \cdot \max_{j=1, \dots, k} \|a_j\|_2 \cdot \|x - z\|_2. \end{aligned}$$

Choose $\alpha_n, \beta_n > 0$ and define for $k \in \mathbb{N}$

$$\mathcal{F}_{k,n} = \left\{ \sum_{i=1}^k c_i \cdot \sigma(a_i^T x + b_i) + c_0 : a_i \in \mathbb{R}^d, b_i, c_i \in \mathbb{R}, \max_{j=1, \dots, k} \|a_j\|_2 \leq \alpha_n, \sum_{i=1}^k |c_i| \leq \beta_n \right\}.$$

Then the functions in $\mathcal{F}_{k,n}$ are all Lipschitz continuous with Lipschitz constant bounded by $\alpha_n \cdot \beta_n$. Furthermore, since σ is bounded in absolute value by 1, the functions in $\mathcal{F}_{k,n}$ are bounded in absolute value by β_n .

We define our regression estimate as a truncated version of the corresponding least squares estimate, where the number k of neurons is chosen by splitting of the sample. More precisely, set

$$\mathcal{P}_n = \{1, 2, \dots, n\}.$$

We subdivide the given data in a learning sample of size $n_l = \lceil n/2 \rceil$ and a testing sample of size $n_t = n - n_l$ and define for a given $k \in \mathcal{P}_n = \{1, \dots, n\}$ our regression estimate by

$$m_{n_l, k}(\cdot) = \arg \min_{f \in \mathcal{F}_{k,n}} \left(\frac{1}{n_l} \sum_{i=1}^{n_l} |f(\bar{X}_i) - Y_i|^2 \right). \quad (6)$$

Then we minimize the empirical L_2 risk on the testing sample in order to choose the value of parameter k . So we choose

$$\hat{k} = \arg \min_{k \in \mathcal{P}_n} \frac{1}{n_t} \sum_{i=n_l+1}^n |m_{n_l, k}(\bar{X}_i) - Y_i|^2 \quad (7)$$

and define our final neural networks regression estimate by

$$m_n(x) = m_{n, \hat{k}}(x) \quad (x \in \mathbb{R}^d). \quad (8)$$

3 Main results

Our main result is the following theorem.

Theorem 1 *Set $\beta_n = c_1 \cdot \log(n)$ for some $c_1 > 0$ and define the estimate m_n as in Section 2. Assume that $\bar{\mathcal{D}}_{n_l}$ is independent of $(X, Y), (X_{n_l+1}, Y_{n_l+1}), \dots, (X_n, Y_n)$, that Y is sub-Gaussian in the sense that*

$$\mathbf{E} \left\{ e^{c_2 Y^2} \right\} < \infty \quad (9)$$

for some $c_2 > 0$ and that the regression function is bounded in absolute value by some $0 \leq L \leq \beta_n$. Then

$$\begin{aligned} \mathbf{E} \int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx) \leq c_3 \cdot & \left(\alpha_n \cdot \log(n)^2 \cdot \mathbf{E} \left\{ \frac{1}{n} \sum_{i=1}^n \|X_i - \bar{X}_{i,n}\|_2 \right\} \right. \\ & \left. + \min_{k \in \mathcal{P}_n} \left(\frac{k \cdot \log(n)^5}{n} + \inf_{f \in \mathcal{F}_{k,n}} \int |f(x) - m(x)|^2 \mathbf{P}_X(dx) \right) \right) \end{aligned}$$

for some constant $c_3 > 0$.

Remark 1. The sub-Gaussian condition (9) is in particular satisfied if

$$Y = m(X) + \epsilon,$$

where $m : \mathbb{R}^d \rightarrow \mathbb{R}$ is bounded, X and ϵ are independent and ϵ is normally distributed with mean zero.

Theorem 1 implies a rate of convergence result as soon as we impose some smoothness condition on the regression function. For neural networks usually such smoothness conditions are defined by imposing conditions on the Fourier transform of the regression function. The Fourier transform \tilde{F} of a function $f \in L_1(\mathbb{R}^d)$ is defined by

$$\tilde{F}(\omega) = \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} e^{-i \cdot \omega^T x} f(x) dx \quad (\omega \in \mathbb{R}^d).$$

If $\tilde{F} \in L_1(\mathbb{R}^d)$ then the inverse formula

$$f(x) = \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} e^{i \cdot \omega^T x} \tilde{F}(\omega) d\omega \quad (10)$$

holds almost everywhere with respect to the Lebesgue measure. Let $0 < C < \infty$ and consider the class of functions \mathcal{F}_C for which (10) holds on \mathbb{R}^d and, in addition,

$$\int_{\mathbb{R}^d} \|\omega\|_2 F(\omega) d\omega \leq C. \quad (11)$$

A class of functions satisfying (11) is a subclass of functions with Fourier transform having first absolute moment finite, i.e., $\int_{\mathbb{R}^d} \|\omega\|_2 F(\omega) d\omega < \infty$ (these functions are continuously differentiable on \mathbb{R}^d). The next corollary provides the rate of convergence of the estimate.

Corollary 1 *Assume that $\|X\|_2$ is bounded almost surely, that Y is sub-Gaussian in the sense that (9) holds, that $m \in \mathcal{F}_C$ for some $C > 0$ and that*

$$\|m\|_\infty = \sup_{x \in \mathbb{R}^d} |m(x)| \leq L < \infty$$

for some $L > 0$. Define the estimate m_n as in Section 2 with

$$\alpha_n = c_4 \cdot n^{1/4} \quad \text{and} \quad \beta_n = c_5 \cdot \log(n).$$

Assume that the measurement error satisfies

$$\mathbf{E} \left\{ \frac{1}{n} \sum_{i=1}^n \|X_i - \bar{X}_{i,n}\|_2 \right\} \leq c_6 \cdot n^{-3/4}. \quad (12)$$

Then

$$\mathbf{E} \int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx) \leq c_7 \sqrt{\frac{\log(n)^5}{n}}$$

for some constant $c_7 > 0$ for n sufficiently large.

Proof. Application of Theorem 1 yields

$$\begin{aligned} \mathbf{E} \int |m_n(x) - m(x)|^2 P_X(dx) &\leq c_8 \cdot \left(\frac{\log(n)^2}{n^{1/2}} \right. \\ &\quad \left. + \frac{k \cdot \log(n)^5}{n} + \inf_{f \in \mathcal{F}_{k,n}} \int |f(x) - m(x)|^2 \mathbf{P}_X(dx) \right) \end{aligned}$$

for any $k \in \mathcal{P}_n$ and for sufficiently large constant $c_8 > 0$. Let $r > 0$ such that $\|X\|_2 \leq r$ a.s. Theorem 3 of Baron (1993) gives us that there exists $f \in \mathcal{F}_{k,n}$ such that

$$\begin{aligned} \int (f(x) - m(x))^2 \mu(dx) &\leq 8C^2 r^2 \left(\frac{1}{k} + \frac{(1 + 2 \log(\alpha_n))^2}{\alpha_n^2} \right) \\ &= 8C^2 r^2 \left(\frac{1}{k} + \frac{(1 + 2 \log(c_4 \cdot n^{1/4}))^2}{c_4^2 \cdot n^{1/2}} \right). \end{aligned}$$

By setting $k = \left\lceil \frac{n^{1/2}}{\log(n)^{2.5}} \right\rceil$ we get the assertion. \square

Remark 2. In Corollary 1 we show for “small” measurement errors up to a logarithmic factor the same rate of convergence as follows from the approximation result in Barron (1993) for regression estimation from data without measurement errors (cf., e.g., Section 16.3 in Györfi et al. (2002)). It is clear from the proof, that the rate of convergence will change as soon as the measurement error will be larger than in (12). In this situation it makes sense to change the definition of α_n in order to optimize the resulting rate of convergence. It is an open problem how to choose this parameter in a data-dependent way such that it achieves the best possible rate of convergence in view of the magnitude of the measurement errors.

4 Proofs

The following lemma is an extension of Lemma 1 in Bagirov, Clausen and Kohler (2008) to data with measurement errors. It is about bounding the L_2 error of estimates, which are defined by splitting of the sample. Let $n = n_l + n_t$, let \mathcal{Q}_n be a finite set of parameters and assume that for each parameter $h \in \mathcal{Q}_n$ an estimate

$$m_{n_l}^{(h)}(\cdot) = m_{n_l}^{(h)}(\cdot, \bar{\mathcal{D}}_{n_l})$$

is given, which depends only on the training data $\bar{\mathcal{D}}_{n_l} = \{(\bar{X}_1, Y_1), \dots, (\bar{X}_{n_l}, Y_{n_l})\}$, and which is Lipschitz continuous with Lipschitz constant L_n . Then we define

$$m_n(x) = m_{n_l}^{(H)}(x) \quad \text{for all } x \in \mathbb{R}^d, \quad (13)$$

where $H \in \mathcal{Q}_n$ is chosen such that

$$\frac{1}{n_t} \sum_{i=n_l+1}^n |m_{n_l}^{(H)}(\bar{X}_i) - Y_i|^2 = \min_{h \in \mathcal{Q}_n} \frac{1}{n_t} \sum_{i=n_l+1}^n |m_{n_l}^{(h)}(\bar{X}_i) - Y_i|^2. \quad (14)$$

Lemma 1 *Assume that $\bar{\mathcal{D}}_{n_l}$ is independent of $(X, Y), (X_{n_l+1}, Y_{n_l+1}), \dots, (X_n, Y_n)$. Let $\beta_n = c_1 \cdot \log(n)$ for some constant $c_1 > 0$ and assume that the estimates $m_{n_l}^{(h)}$ are bounded in absolute value by β_n for $h \in \mathcal{Q}_n$. Assume furthermore that the distribution of (X, Y) satisfies the Sub-Gaussian condition (9) for some constant $c_2 > 0$, and that the regression*

function fulfils $\|m\|_\infty \leq L$ for some $L \in \mathbb{R}_+$, with $L \leq \beta_n$. Then, for every estimate m_n defined by (13) and (14) and any $\delta > 0$,

$$\begin{aligned} & \mathbf{E} \int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx) \\ & \leq (1 + \delta) \min_{h \in \mathcal{Q}} \mathbf{E} \int |m_{n_l}^{(h)}(x) - m(x)|^2 \mathbf{P}_X(dx) + c_9 \cdot \beta_n^2 \cdot \frac{1 + \log |\mathcal{Q}_n|}{n_t} + c_{10} \frac{\log(n)}{n} \\ & \quad + 8\beta_n \cdot (1 + \delta) \cdot L_n \cdot \mathbf{E} \left\{ \frac{1}{n_t} \sum_{i=n_1+1}^n \|X_i - \bar{X}_{i,n}\|_2 \right\} \end{aligned}$$

holds, with $c_9 = 16/\delta + 35 + 19\delta$ and a sufficiently large constant $c_{10} > 0$.

In the proof we will need the following lemma, which follows from the proof of Lemma 1 in Bagirov, Clausen and Kohler (2008).

Lemma 2 *Let $\beta_n = c_1 \cdot \log(n)$ for some constant $c_1 > 0$, let (X, Y) and Z be random variables and assume that Y satisfies*

$$\mathbf{E}\{\exp(c_2|Y|^2)\} < \infty$$

and that $|Z| \leq \beta_n$ a.s. Set

$$m(X) = \mathbf{E}\{Y|X\} \quad \text{and} \quad m_{\beta_n}(X) = \mathbf{E}\{T_{\beta_n}Y|X\},$$

where

$$T_{\beta_n}Y = \min\{\beta_n, \max\{-\beta_n, Y\}\},$$

and assume $|m(X)| \leq L$ a.s. for some $0 \leq L \leq \beta_n$. Then

$$|\mathbf{E}(|Z - Y|^2) - \mathbf{E}(|Z - T_{\beta_n}Y|^2)| \leq c_{11} \cdot \frac{\log(n)}{n}$$

and

$$|\mathbf{E}(|m(X) - Y|^2) - \mathbf{E}(|m_{\beta_n}(X) - T_{\beta_n}Y|^2)| \leq c_{11} \cdot \frac{\log(n)}{n}$$

for some sufficiently large constant $c_{11} > 0$.

For the sake of completeness we give the proof of Lemma 2 in the appendix.

Proof of Lemma 1. We use the following error decomposition

$$\begin{aligned}
& \mathbf{E} \left(\int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx) \middle| \bar{\mathcal{D}}_{n_l} \right) \\
&= \mathbf{E} \left(\int |m_{n_l}^{(H)}(x) - m(x)|^2 \mathbf{P}_X(dx) \middle| \bar{\mathcal{D}}_{n_l} \right) \\
&= \left[\mathbf{E} \left(|m_{n_l}^{(H)}(X) - Y|^2 \middle| \bar{\mathcal{D}}_{n_l} \right) - \mathbf{E} (|m(X) - Y|^2) \right. \\
&\quad \left. - \mathbf{E} \left(|m_{n_l}^{(H)}(X) - T_{\beta_n} Y|^2 \middle| \bar{\mathcal{D}}_{n_l} \right) - \mathbf{E} (|m_{\beta_n}(X) - T_{\beta_n} Y|^2) \right] \\
&+ \left[\mathbf{E} \left(|m_{n_l}^{(H)}(X) - T_{\beta_n} Y|^2 \middle| \bar{\mathcal{D}}_{n_l} \right) - \mathbf{E} (|m_{\beta_n}(X) - T_{\beta_n} Y|^2) \right. \\
&\quad \left. - (1 + \delta) \cdot \frac{1}{n_t} \sum_{i=n_l+1}^n \left(|m_{n_l}^{(H)}(X_i) - T_{\beta_n} Y_i|^2 - |m_{\beta_n}(X_i) - T_{\beta_n} Y_i|^2 \right) \right] \\
&+ \left[(1 + \delta) \cdot \frac{1}{n_t} \sum_{i=n_l+1}^n \left(|m_{n_l}^{(H)}(X_i) - T_{\beta_n} Y_i|^2 - |m_{\beta_n}(X_i) - T_{\beta_n} Y_i|^2 \right) \right. \\
&\quad \left. - (1 + \delta) \cdot \frac{1}{n_t} \sum_{i=n_l+1}^n \left(|m_{n_l}^{(H)}(\bar{X}_i) - T_{\beta_n} Y_i|^2 - |m_{\beta_n}(X_i) - T_{\beta_n} Y_i|^2 \right) \right] \\
&+ \left[(1 + \delta) \cdot \frac{1}{n_t} \sum_{i=n_l+1}^n \left(|m_{n_l}^{(H)}(\bar{X}_i) - T_{\beta_n} Y_i|^2 - |m_{\beta_n}(X_i) - T_{\beta_n} Y_i|^2 \right) \right. \\
&\quad \left. - (1 + \delta) \cdot \frac{1}{n_t} \sum_{i=n_l+1}^n \left(|m_{n_l}^{(H)}(\bar{X}_i) - Y_i|^2 - |m(X_i) - Y_i|^2 \right) \right] \\
&+ \left[(1 + \delta) \cdot \frac{1}{n_t} \sum_{i=n_l+1}^n \left(|m_{n_l}^{(H)}(\bar{X}_i) - Y_i|^2 - |m(X_i) - Y_i|^2 \right) \right] = \sum_{i=1}^5 T_{i,n},
\end{aligned}$$

where $T_{\beta_n} Y$ denotes the truncated version of Y and $m_{\beta_n}(x) = \mathbf{E} \{T_{\beta_n} Y | X = x\}$.

Since the estimates are Lipschitz continuous with Lipschitz constant L_n and bounded by β_n we get

$$\begin{aligned}
& T_{3,n} \\
&= (1 + \delta) \cdot \frac{1}{n_t} \sum_{i=n_l+1}^n \left(|m_{n_l}^{(H)}(X_i) - T_{\beta_n} Y_i|^2 - |m_{n_l}^{(H)}(\bar{X}_i) - T_{\beta_n} Y_i|^2 \right) \\
&\leq (1 + \delta) \cdot \frac{1}{n_t} \sum_{i=n_l+1}^n |m_{n_l}^{(H)}(X_i) - m_{n_l}^{(H)}(\bar{X}_i)| \cdot |m_{n_l}^{(H)}(X_i) - T_{\beta_n} Y_i + m_{n_l}^{(H)}(\bar{X}_i) - T_{\beta_n} Y_i| \\
&\leq (1 + \delta) \cdot 4\beta_n \cdot L_n \cdot \frac{1}{n_t} \sum_{i=n_l+1}^n \|X_i - \bar{X}_{i,n}\|_2. \tag{15}
\end{aligned}$$

By Lemma 2 we get furthermore

$$T_{1,n} + \mathbf{E}\{T_{4,n}|\bar{\mathcal{D}}_{n_l}\} \leq c_{11} \frac{\log(n)}{n}.$$

And by bounding $T_{2,n}$ as in the proof of Theorem 7.1 in Györfi et al. (2002) we get

$$\mathbf{E}\{T_{2,n}|\bar{\mathcal{D}}_{n_l}\} \leq c_9 \cdot \beta_n^2 \cdot \frac{1 + \log|\mathcal{Q}_n|}{n_t}.$$

So it remains to bound $T_{5,n}$. By definition of the estimate we have for any $h \in \mathcal{Q}_n$

$$\begin{aligned} T_{5,n} &\leq (1 + \delta) \cdot \frac{1}{n_t} \sum_{i=n_l+1}^n \left(|m_{n_l}^{(h)}(\bar{X}_i) - Y_i|^2 - |m(X_i) - Y_i|^2 \right) \\ &= \left[(1 + \delta) \cdot \frac{1}{n_t} \sum_{i=n_l+1}^n \left(|m_{n_l}^{(h)}(\bar{X}_i) - Y_i|^2 - |m(X_i) - Y_i|^2 \right) \right. \\ &\quad \left. - (1 + \delta) \cdot \frac{1}{n_t} \sum_{i=n_l+1}^n \left(|m_{n_l}^{(h)}(\bar{X}_i) - T_{\beta_n} Y_i|^2 - |m(X_i) - Y_i|^2 \right) \right] \\ &\quad + \left[(1 + \delta) \cdot \frac{1}{n_t} \sum_{i=n_l+1}^n \left(|m_{n_l}^{(h)}(\bar{X}_i) - T_{\beta_n} Y_i|^2 - |m(X_i) - Y_i|^2 \right) \right. \\ &\quad \left. - (1 + \delta) \cdot \frac{1}{n_t} \sum_{i=n_l+1}^n \left(|m_{n_l}^{(h)}(X_i) - T_{\beta_n} Y_i|^2 - |m(X_i) - Y_i|^2 \right) \right] \\ &\quad + \left[(1 + \delta) \cdot \frac{1}{n_t} \sum_{i=n_l+1}^n \left(|m_{n_l}^{(h)}(X_i) - T_{\beta_n} Y_i|^2 - |m(X_i) - Y_i|^2 \right) \right. \\ &\quad \left. - (1 + \delta) \cdot \frac{1}{n_t} \sum_{i=n_l+1}^n \left(|m_{n_l}^{(h)}(X_i) - Y_i|^2 - |m(X_i) - Y_i|^2 \right) \right] \\ &\quad + \left[(1 + \delta) \cdot \frac{1}{n_t} \sum_{i=n_l+1}^n \left(|m_{n_l}^{(h)}(X_i) - Y_i|^2 - |m(X_i) - Y_i|^2 \right) \right] \\ &= \sum_{i=6}^9 T_{i,n}. \end{aligned}$$

As in (15) we get

$$T_{7,n} \leq (1 + \delta) \cdot 4\beta_n \cdot L_n \cdot \frac{1}{n_t} \sum_{i=n_l+1}^n \|X_i - \bar{X}_{i,n}\|_2.$$

Bounding $T_{6,n}$ and $T_{8,n}$ by Lemma 2 we get

$$\mathbf{E}\{T_{6,n}|\bar{\mathcal{D}}_{n_l}\} + \mathbf{E}\{T_{8,n}|\bar{\mathcal{D}}_{n_l}\} \leq c_{11} \frac{\log(n)}{n}.$$

Finally we get

$$\mathbf{E}\{T_{9,n}|\bar{\mathcal{D}}_{n_l}\} = (1 + \delta)\mathbf{E} \int |m_{n_l}^{(h)}(x) - m(x)|^2 \mathbf{P}_X(dx).$$

Summarizing the above results we get the assertion. \square

Proof of Theorem 1. By Lemma 1 we get

$$\begin{aligned} \mathbf{E} \int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx) &\leq 2 \min_{k \in \mathcal{P}_n} \mathbf{E} \int |m_{n_l,k}(x) - m(x)|^2 \mathbf{P}_X(dx) + c_{12} \frac{(\log n)^3}{n_t} \\ &\quad + 16 \cdot \alpha_n \cdot \beta_n^2 \cdot \mathbf{E} \left\{ \frac{1}{n_t} \sum_{i=n_1+1}^n \|X_i - \bar{X}_{i,n}\|_2 \right\}. \end{aligned}$$

Hence it suffices to show:

$$\begin{aligned} \mathbf{E} \int |m_{n_l,k}(x) - m(x)|^2 \mathbf{P}_X(dx) &\leq c_{13} \cdot \left(\alpha_n \cdot \beta_n^2 \cdot \mathbf{E} \left\{ \frac{1}{n_l} \sum_{i=1}^{n_l} \|X_i - \bar{X}_{i,n}\|_2 \right\} \right. \\ &\quad \left. + \frac{k \cdot \log(n)^5}{n} + \inf_{f \in \mathcal{F}_{k,n}} \int |f(x) - m(x)|^2 \mathbf{P}_X(dx) \right). \end{aligned} \quad (16)$$

In order to prove (16), set

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

choose an arbitrary $f \in \mathcal{F}_{k,n}$ and consider the error decomposition

$$\begin{aligned} \mathbf{E} \int |m_{n_l,k}(x) - m(x)|^2 \mathbf{P}_X(dx) &= \mathbf{E}\{|m_{n_l,k}(X) - Y|^2 | \bar{\mathcal{D}}_{n_l}\} - \mathbf{E}\{|m(X) - Y|^2\} \\ &= \sum_{i=1}^9 T_{i,n} \end{aligned}$$

where

$$\begin{aligned} T_{1,n} &= \mathbf{E}\{|m_{n_l,k}(X) - Y|^2 | \bar{\mathcal{D}}_{n_l}\} - \mathbf{E}\{|m(X) - Y|^2\} \\ &\quad - (\mathbf{E}\{|m_{n_l,k}(X) - T_{\beta_n} Y|^2 | \bar{\mathcal{D}}_{n_l}\} - \mathbf{E}\{|m_{\beta_n}(X) - T_{\beta_n} Y|^2\}), \end{aligned}$$

$$\begin{aligned} T_{2,n} &= \mathbf{E}\{|m_{n_l,k}(X) - T_{\beta_n} Y|^2 | \bar{\mathcal{D}}_{n_l}\} - \mathbf{E}\{|m_{\beta_n}(X) - T_{\beta_n} Y|^2\} \\ &\quad - \frac{2}{n_l} \sum_{i=1}^{n_l} (|m_{n_l,k}(X_i) - T_{\beta_n} Y_i|^2 - |m_{\beta_n}(X_i) - T_{\beta_n} Y_i|^2) \end{aligned}$$

$$\begin{aligned} T_{3,n} &= \frac{2}{n_l} \sum_{i=1}^{n_l} (|m_{n_l,k}(X_i) - T_{\beta_n} Y_i|^2 - |m_{\beta_n}(X_i) - T_{\beta_n} Y_i|^2) \\ &\quad - \frac{2}{n_l} \sum_{i=1}^{n_l} (|m_{n_l,k}(\bar{X}_i) - T_{\beta_n} Y_i|^2 - |m_{\beta_n}(X_i) - T_{\beta_n} Y_i|^2) \end{aligned}$$

$$T_{4,n} = \frac{2}{n_l} \sum_{i=1}^{n_l} (|m_{n_l,k}(\bar{X}_i) - T_{\beta_n} Y_i|^2 - |m_{\beta_n}(X_i) - T_{\beta_n} Y_i|^2) \\ - \frac{2}{n_l} \sum_{i=1}^{n_l} (|m_{n_l,k}(\bar{X}_i) - Y_i|^2 - |m(X_i) - Y_i|^2)$$

$$T_{5,n} = \frac{2}{n_l} \sum_{i=1}^{n_l} (|m_{n_l,k}(\bar{X}_i) - Y_i|^2 - |m(X_i) - Y_i|^2) \\ - \frac{2}{n_l} \sum_{i=1}^{n_l} (|f(\bar{X}_i) - Y_i|^2 - |m(X_i) - Y_i|^2)$$

$$T_{6,n} = \frac{2}{n_l} \sum_{i=1}^{n_l} (|f(\bar{X}_i) - Y_i|^2 - |m(X_i) - Y_i|^2) \\ - \frac{2}{n_l} \sum_{i=1}^{n_l} (|f(\bar{X}_i) - T_{\beta_n} Y_i|^2 - |m(X_i) - Y_i|^2)$$

$$T_{7,n} = \frac{2}{n_l} \sum_{i=1}^{n_l} (|f(\bar{X}_i) - T_{\beta_n} Y_i|^2 - |m(X_i) - Y_i|^2) \\ - \frac{2}{n_l} \sum_{i=1}^{n_l} (|f(X_i) - T_{\beta_n} Y_i|^2 - |m(X_i) - Y_i|^2)$$

$$T_{8,n} = \frac{2}{n_l} \sum_{i=1}^{n_l} (|f(X_i) - T_{\beta_n} Y_i|^2 - |m(X_i) - Y_i|^2) \\ - \frac{2}{n_l} \sum_{i=1}^{n_l} (|f(X_i) - Y_i|^2 - |m(X_i) - Y_i|^2)$$

$$T_{9,n} = \frac{2}{n_l} \sum_{i=1}^{n_l} (|f(X_i) - Y_i|^2 - |m(X_i) - Y_i|^2).$$

By Lemma 2 we get

$$\mathbf{E}\{T_{i,n}\} \leq c_{11} \cdot \frac{\log(n)}{n}$$

for $i \in \{1, 4, 6, 8\}$. Furthermore we get by Lipschitz continuity of the functions in $\mathcal{F}_{k,n}$

$$\mathbf{E}\{T_{j,n}\} \leq c_{13} \cdot \alpha_n \cdot \beta_n^2 \cdot \mathbf{E} \left\{ \frac{1}{n_l} \sum_{i=1}^{n_l} \|X_i - \bar{X}_{i,n}\|_2 \right\}$$

for $j \in \{3, 7\}$ (cf. proof of (15)), and $f \in \mathcal{F}_{k,n}$ and the definition of the estimate implies

$$T_{5,n} \leq 0.$$

Thus it remains to bound $T_{2,n}$ and $T_{9,n}$. For $T_{9,n}$ we get

$$\mathbf{E}\{T_{9,n}\} = 2 \cdot \int |f(x) - m(x)|^2 \mathbf{P}_X(dx).$$

In order to bound $T_{2,n}$, choose $s > 0$ and consider

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

The last probability can be bounded by Theorem 11.4 in Györfi et al. (2002). With the usual bounds for covering numbers of classes of neural networks (cf. Lemma 16.6 in Györfi et al. (2002)) we get

$$\mathbf{P}\{T_{2,n} > s\} \leq (c_{14} n_l)^{c_{15} k n} \cdot \exp\left(-\frac{s \cdot n_l}{c_{16} \log(n)^4}\right).$$

Using

$$\mathbf{E}T_{2,n} \leq u + \int_u^\infty \mathbf{P}\{T_{2,n} > s\} ds$$

and minimizing the above bound with respect to $u > 0$ we get after some tedious calculations

$$\mathbf{E}T_{2,n} \leq \frac{k \cdot \log(n)^5}{n}.$$

Summarizing the above results, the proof is complete. \square

A Proof of Lemma 2

By using $a^2 - b^2 = (a - b)(a + b)$ we get

$$|T_{1,n}| = \left| \mathbf{E}\left(|Z - Y|^2 - |Z - T_{\beta_n} Y|^2\right) \right| = \left| \mathbf{E}\left((T_{\beta_n} Y - Y)(2Z - Y - T_{\beta_n} Y)\right) \right|.$$

With the Cauchy-Schwarz inequality and

$$I_{\{|Y| > \beta_n\}} \leq \frac{\exp(c_2/2 \cdot |Y|^2)}{\exp(c_2/2 \cdot \beta_n^2)}, \quad (17)$$

the last term can be bounded by

$$\begin{aligned}
& \sqrt{\mathbf{E}(|T_{\beta_n} Y - Y|^2)} \cdot \sqrt{\mathbf{E}(|2Z - Y - T_{\beta_n} Y|^2)} \\
& \leq \sqrt{\mathbf{E}(|Y|^2 \cdot I_{\{|Y| > \beta_n\}})} \cdot \sqrt{\mathbf{E}(2 \cdot |2Z - T_{\beta_n} Y|^2 + 2 \cdot |Y|^2)} \\
& \leq \sqrt{\mathbf{E}\left(|Y|^2 \cdot \frac{\exp(c_2/2 \cdot |Y|^2)}{\exp(c_2/2 \cdot \beta_n^2)}\right)} \cdot \sqrt{2(3\beta_n)^2 + 2\mathbf{E}(|Y|^2)} \\
& \leq \sqrt{\mathbf{E}\left(|Y|^2 \exp(c_2/2 \cdot |Y|^2)\right)} \exp\left(-\frac{c_2 \cdot \beta_n^2}{4}\right) \sqrt{2(3\beta_n)^2 + 2\mathbf{E}(|Y|^2)},
\end{aligned}$$

owing to the boundedness of Z . With $x \leq \exp(x)$ for $x \in \mathbb{R}$ we get

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

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

$$\mathbf{E}\left(\frac{2}{c_2} \cdot \exp(c_2/2 \cdot |Y|^2) \cdot \exp(c_2/2 \cdot |Y|^2)\right) \leq \mathbf{E}\left(\frac{2}{c_2} \cdot \exp(c_2 \cdot |Y|^2)\right) \leq c_{17},$$

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

$$\mathbf{E}(|Y|^2) \leq \mathbf{E}(1/c_2 \cdot \exp(c_2 \cdot |Y|^2)) \leq c_{18} < \infty,$$

which follows again as above. With the setting $\beta_n = c_1 \cdot \log(n)$ it follows for some constants $c_{19}, c_{20}, c_{21}, c_{22} > 0$

$$|T_{1,n}| \leq \sqrt{c_{19}} \cdot \exp(-c_{20} \cdot \log(n)^2) \cdot \sqrt{(18 \cdot c_1^2 \cdot \log(n)^2 + 2c_{21})} \leq c_{22} \cdot \frac{\log(n)}{n}.$$

Arguing in the same way we get from the Cauchy-Schwarz inequality

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

where we can bound the second factor on the right hand-side in the above inequality in the same way we have bounded the second factor from $T_{1,n}$, because by assumption $|m(X)|$ is

bounded a.s., and $|m_{\beta_n}(X)|$ is clearly also bounded, namely by β_n . Thus, we get for some constant $c_{23} > 0$,

$$\sqrt{\mathbf{E}\left(\left|m(X) + m_{\beta_n}(X) - Y - T_{\beta_n}Y\right|^2\right)} \leq c_{23} \cdot \log(n).$$

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

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

Hence we get,

$$T_{2,n} \leq \sqrt{4\mathbf{E}\left(\left|Y - T_{\beta_n}Y\right|^2\right) \cdot c_{23} \cdot \log(n)},$$

and therefore the calculations from $T_{1,n}$ imply $T_{2,n} \leq c_{24} \cdot \log(n)/n$, which completes the proof. \square

References

- [1] Anthony, M. and Bartlett, P. L. (1999). *Neural Networks and Learning: Theoretical Foundations*. Cambridge University Press, Cambridge.
- [2] Bagirov, A.M., Clausen, C., and Kohler, M. (2006). An L_2 -boosting algorithm for estimation of a regression function. Submitted for publication.
- [3] Barron, A. R. (1991). Complexity regularization with application to artificial neural networks. In G. Roussas (ed.), *Nonparametric Functional Estimation and Related Topics*, pp. 5621–576, NATO ASI Series, Kluwer Academic Publishers, Dordrecht, Netherlands.
- [4] Barron, A. R. (1993). Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information Theory* **39**, pp. 930–944.
- [5] Carroll, R. J., Maca, J. D. and Ruppert, D. (1999). Nonparametric regression in the presence of measurement error. *Biometrika* **86**, pp. 541-554.
- [6] Cybenko, G. (1989). Approximations by superpositions of sigmoidal functions. *Mathematics of Control, Signals, and Systems* **2**, pp. 303-314.

- [7] Delaigle, A. and Meister, A. (2007). Nonparametric regression estimation in the heteroscedastic errors-in-variables problem. *Journal of the American Statistical Association* **102**, pp. 1416-1426.
- [8] Delaigle, A., Fan, J. and Carroll, R.J. (2009). A design-adaptive local polynomial estimator for the errors-in-variables problem. *Journal of the American Statistical Association* **104**, pp. 348-359.
- [9] Devroye, L., Györfi, L. and Lugosi, G. (1996). *A Probabilistic Theory of Pattern Recognition*. Springer-Verlag, New York.
- [10] Fan, J. and Truong, Y.K. (1993). Nonparametric regression with errors in variables. *Annals of Statistics* **21**, pp. 1900-1925.
- [11] Funahashi, K. (1989). On the approximate realization of continuous mappings by neural networks. *Neural Networks* **2**, pp. 183–192.
- [12] Györfi, L., Kohler, M., Krzyżak, A. and Walk, H. (2002). *A distribution-free theory of nonparametric regression*. Springer Series in Statistics, Springer.
- [13] Hamers, M and Kohler, M. (2006). Nonasymptotic bounds on the L_2 error of neural network regression estimates. *Annals of the Institute of Statistical Mathematics* **58**, pp. 131-151.
- [14] Hertz, J., Krogh, A., and Palmer, R.G. (1991). *Introduction to the Theory of Neural Computation*. Addison-Wesley, CA.
- [15] Hornik, K. (1993). Some new results on neural network approximation. *Neural Networks* **6**, pp. 1069-1072.
- [16] Hornik, K., Stinchcombe, M., and White, H. (1989). Multi-layer feedforward networks are universal approximators. *Neural Networks* **2**, pp. 359–366.
- [17] Kohler, M. and Krzyżak, A. (2005). Adaptive regression estimation with multilayer feedforward neural networks. *Journal of Nonparametric Statistics* **17**, pp. 891-913.
- [18] Lugosi, G. and Zeger, K. (1995). Nonparametric estimation via empirical risk minimization. *IEEE Trans. Inform. Theory* **41**, 677-687.

- [19] McCaffrey, D.F. and Gallant, A.R. (1994). Convergence rates for single hidden layer feedforward networks. *Neural Networks* **7**, pp. 147-158.
- [20] Mielniczuk, J. and Tyrcha, J. (1993). Consistency of multilayer perceptron regression estimators. *Neural Networks* **6**, pp. 1019-1022.
- [21] Ripley, B. D. (1996). *Pattern Recognition and Neural Networks*. Cambridge University Press, Cambridge, UK.