AISM (2006) 58: 131–151
DOI 10.1007/s10463-005-0005-9

Michael Hamers · Michael Kohler

Nonasymptotic bounds on the L_2 error of neural network regression estimates

Received: 20 November 2003 / Revised: 24 November 2004 / Published online: 15 March 2006

Abstract The estimation of multivariate regression functions from bounded i.i.d. data is considered. The L_2 error with integration with respect to the design measure is used as an error criterion. The distribution of the design is assumed to be concentrated on a finite set. Neural network estimates are defined by minimizing the empirical L_2 risk over various sets of feedforward neural networks. Nonasymptotic bounds on the L_2 error of these estimates are presented. The results imply that neural networks are able to adapt to additive regression functions and to regression functions which are a sum of ridge functions, and hence are able to circumvent the curse of dimensionality in these cases.

Keywords Neural networks · Nonparametric regression · Dimension reduction · Additive models · Curse of dimensionality

1 Introduction

Neural networks are frequently implemented in applications. They are motivated by the desire to model human brain by computer. The original biological motivation for such networks stems from McCulloch and Pitts (1943) who modeled a neuron by a binary thresholding device in discrete time. This so-called *perceptron* applies a threshold element to a linear combination of its d inputs:

$$g(x) = \sigma\left(a^{\mathrm{T}}x + b\right),$$

M. Hamers (✉)
Fachbereich Mathematik,
Universität Stuttgart, Pfaffenwaldring 57, D-70569 Stuttgart, Germany
E-mail: hamers@mathematik.uni-stuttgart.de

M. Kohler
Fachrichtung Mathematik,
Universität des Saarlandes, Postfach 151150, D-66041 Saarbrücken, Germany
E-mail: kohler@math.uni-sb.de

where $x \in \mathbb{R}^d$ is an input vector, $a = (a_1, \ldots, a_d)^{\mathrm{T}} \in \mathbb{R}^d$, $b \in \mathbb{R}$ are the weights and $\sigma(x) = I_{\{x \in [0,\infty)\}}$ is the threshold element.

As shown by Minsky and Papert (1969), the class of functions which can be approximated well by perceptrons is very limited. These limitations can be obviated by adding additional layers of neurons called hidden layers which leads to multilayer perceptron neural networks. One example of such networks is *feedforward neural networks with one hidden layer* defined by

$$f(x) = \sum_{i=1}^{k} c_i \sigma\left(a_i^{\mathrm{T}} x + b_i\right) + c_0 \quad (x \in \mathbb{R}^d)$$

where $k \in \mathbb{N}$ is the number of hidden neurons, $\sigma : \mathbb{R} \to [0, 1]$ is called sigmoid function and $a_1, \ldots, a_k \in \mathbb{R}^d$, $b_1, \ldots, b_k, c_0, \ldots, c_k \in \mathbb{R}$ are the weights of the network. The so-called squashing functions $\sigma : \mathbb{R} \to [0, 1]$, i.e. nondecreasing functions which satisfy

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

are often used as sigmoid functions. The class of feedforward neural networks with one hidden layer is very powerful. For example, as shown independently by Cybenko (1989), Funahashi (1989) and Hornik, Stinchcombe and White (1989), any continuous function on $\mathbb{R}^d$ can be arbitrarily approximated closely by such networks in supremum norm on compact sets. General introductions to neural networks can be found, e.g., in the monographs (Anthony and Bartlett 1999; Devroye et al. 1996; 1996; Györfi, Kohler, Krzyźak & Walk 2002; Hertz, Krogh & Palimir 1991; Ripley 1996).

In this article we use neural networks to estimate a regression function from observed data. To describe the regression estimation problem precisely, let (X, Y), (X_1, Y_1), (X_2, Y_2), ... be independent identically distributed $\mathbb{R}^d \times \mathbb{R}$ - valued random vectors with $\mathbf{E}Y^2 < \infty$. In regression analysis you want to estimate Y after having observed X, i.e., you have to determine a function f with $f(X)$ "close" to Y. If "closeness" is measured by the mean squared error, then you have to find a function f^* such that

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

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

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

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

In the regression estimation problem the distribution of (X, Y) (and consequently m) is unknown. Given a sequence $\mathcal{D}_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ of independent observations of (X, Y), the aim is to construct an estimate $m_n(x) = m_n(x, \mathcal{D}_n)$ of $m(x)$ such that the L_2 error $\int |m_n(x) - m(x)|^2 \mu(dx)$ is small.

A very powerful principle to construct regression estimates is the principle of least squares. Here, the L_2 risk

$$\mathbf{E}\{|f(X) - Y|^2\} \tag{3}$$

of a function $f : \mathbb{R}^d \to \mathbb{R}$ is estimated by the so-called empirical L_2 risk

$$\frac{1}{n} \sum_{i=1}^{n} |f(X_i) - Y_i|^2, \tag{4}$$

and an estimate of the regression function (i.e., the function which minimizes Eq. (3)) is constructed by minimizing the empirical L_2 risk Eq. (4). Minimizing Eq. (4) over all functions would result (at least if the $X_1, \ldots, X_n$ are distinct) in an estimate which interpolates the data. Obviously, this is not a reasonable estimate of the regression function. Therefore, sets $\mathcal{F}_n$ of functions $f : \mathbb{R}^d \to \mathbb{R}$ are defined which depend on the sample size n and get more and more complex for n tending to infinity. Equation (4) is then minimized only over $\mathcal{F}_n$, i.e., least squares estimates m_n are defined by

$$m_n(\cdot) = \arg \min_{f \in \mathcal{F}_n} \frac{1}{n} \sum_{i=1}^{n} |f(X_i) - Y_i|^2. \tag{5}$$

Here, and in the sequel, we assume for simplicity that the minima in Eq. (5) exist, but we do not assume them to be unique.

The crucial point in the definition of least squares estimates is the choice of the set $\mathcal{F}_n$ of functions over which the empirical L_2 risk is minimized. On the one hand, it should not be too "complex" in order to guarantee that the error introduced by minimizing the empirical L_2 risk instead of the L_2 risk is small. On the other hand, it must be chosen in such a way that the regression function can be approximated well by functions from this set.

In this article we study least squares estimates using neural networks, where $\mathcal{F}_n$ is chosen to be as a class of feedforward neural networks with one hidden layer. Such estimates have been investigated in many articles, e.g., concerning L_2 consistency in White (1990) and Lugosi and Zeger (1995), and concerning rate of convergence of the L_2 error in Barron (1991, 1994) and McGaffrey and Gallant (1994).

The derivation of rate-of-convergence results requires approximation results concerning L_2 norm or supremum norm. Barron (1994) imposed conditions on the Fourier transform to derive such approximation results for neural networks. These conditions do not fit the usual statistical framework for analyzing regression estimates, where it is common to impose conditions such as Lipschitz continuity on derivatives of the regression function like in Stone (1982), together with assumptions on the structure of the regression function such as additivity which make it possible to derive good rates of convergence even for high-dimensional data and hence to circumvent the so-called curse of dimensionality (cf. Stone 1985, 1994). More precisely, it was shown in Stone (1985) that if the regression function is a sum of univariate functions of its d components where these univariate functions are p-times continuously differentiable, then the L_2 error of suitably defined estimates converges to zero with the rate $n^{-2p/(2p+1)}$. In contrast, if one does not assume anything about the structure of the regression function, the optimal rate of convergence for estimation of p-times continuously differentiable functions is $n^{-2p/(2p+d)}$

(cf. Stone 1982), which converges to zero rather slowly provided the dimension d of X is large. For additional results on additive models and related models see, e.g., Andrews and Whang 1990; Bickel et al 1993; Breiman 1993; Breiman and Freiman 1985; Breiman 1993; Burman 1990; Chen 1991, Hastie and Tibshirani 1990; Huang 1998; Kohler (1998); Linton 1997; Linton and Härdle 1996; Linton 1997; Linton and Nielsen 1995; Newey 1994; Stone 1994; Wahba et al 1995; and the literature cited therein.

One can conclude from Barron (1994) that the rate of convergence of suitably defined neural network regression estimates is "good" even if the dimension of X is large, provided that the smoothness of the regression function increases with increasing dimension of X. In the sequel we try to avoid such a condition and impose instead of conditions on the structure of the regression function (such as additivity) in order to circumvent the curse of dimensionality. Unfortunately, it seems very hard to derive sharp approximation results for neural networks under these assumptions.

Therefore, we analyze neural network regression estimates in the framework proposed by Hamers and Kohler (2004), which enables us to avoid difficult approximation problems. We assume that, as often in applications, the distribution of the design (i.e., μ) is concentrated on a finite set. As a consequence, we need approximation results for neural networks only concerning supremum norm on the finite support of μ, which are rather easy to derive.

We give bounds on the expected L_2 error of neural network estimates for general regression functions, for additive regression functions and for regression functions which are a sum of ridge functions. The results imply that neural networks are able to adapt to additive regression functions and to regression functions which are a sum of ridge functions, and hence are able to circumvent the curse of dimensionality in these cases.

1.1 Notation

$\mathbb{N}$ and $\mathbb{R}$ are the sets of natural and real numbers, respectively. I_A denotes the indicator function, $\text{card}(A)$ the cardinality of a set A. The natural logarithm is denoted by $\log(\cdot)$, the distribution of X is denoted by μ and $\text{supp}(X)$ is the support of the distribution of the random variable X. The Euclidean norm of $x \in \mathbb{R}^d$ is denoted by $\|x\|$, and the components of x are denoted by $x^{(1)}, \ldots, x^{(d)}$.

1.2 Outline

The main results are stated in Sect. 2. In Sect. 3, we derive bounds on expected maximal deviations of sample averages from their means which we use in the proofs of the main results. Section 4 contains a general bound on the expected L_2 error of least squares estimates. Approximation properties of neural networks are derived in Sect. 5. The proofs of the main results are given in Sect. 6.

2 Main results

In the sequel we assume that Y is bounded in absolute value by some constant L almost surely (a.s.) and that the distribution μ of X is concentrated on a finite set.

Let K denotes the cardinality of the support of X. Define the class of neural networks

$$\mathcal{F}_n := \left\{ f(x) = c_0 + \sum_{i=1}^{K-1} c_i \sigma(a_i^T x + b_i) \ : \ a_i \in \mathbb{R}^d, b_i, c_i \in \mathbb{R}, |c_i| \leq 2L \right\},$$

where σ is an arbitrary squashing function. Choose $\tilde{m}_n(\cdot) \in \mathcal{F}_n$ such that

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

and define $m_n(\cdot)$ by truncating $\tilde{m}_n(\cdot)$ at $\pm L$. Then analogously to Theorem 1 in Hamers and Kohler (2004) the following theorem holds (which is actually weaker than the bound in Hamers and Kohler (2004), cf. Remark 1 below):

Theorem 2.1 *Let $L \geq 1$, $d \in \mathbb{N}$ and $K \in \mathbb{N}$. Assume that the distribution of (X, Y) satisfies $(X, Y) \in \mathbb{R}^d \times [-L, L]$ a.s. and* $\text{card}(\text{supp}(X)) = K$. *Let the neural network estimate m_n be defined as above. Then for all $n \in \mathbb{N}$*

$$\mathbf{E} \int |m_n(x) - m(x)|^2 \mu(\mathrm{d}x) \leq c_n \cdot \frac{K}{n}$$

where

$$c_n = 101 L^2 \cdot (2d + 5) \cdot \log(48 e L^2 n^3) = O(\log(n)).$$

The upper bound on the L_2 error in Theorem 2.1 is of the parametric form const $\cdot \frac{K \cdot \log n}{n}$ and hence much smaller than the minimax bound $n^{-2p/(2p+d)}$ which is usually derived in case of p-times differentiable regression functions. This is not surprising: As we assume the distribution of X to be concentrated on a set of cardinality K, we only need to estimate the value of the regression function at K points which should be possible with a parametric rate of convergence of order K/n. It is shown in Hamers and Kohler (2004) that the minimax estimation error of a regression function under the assumption $\text{card}(\text{supp}(X)) = K$ is indeed the form of const $\cdot K/n$, hence the upper bound in Theorem 2.1 is optimal up to a logarithmic factor.

The upper bound in Theorem 2.1 is not satisfying in case the dimension d of X is large. Even if each component of X takes on only two values, the cardinality of the support of X can be 2^d which might be rather large compared to sample sizes n occuring in applications. The only possibility to circumvent this so-called curse of dimensionality is to impose additional assumptions on the structure of the regression function and to use estimates which are able to adapt to these assumptions. Usual assumptions are additivities of the regression function (cf. Stone 1985), i.e.,

$$m(x) = m_1(x^{(1)}) + \cdots + m_d(x^{(d)}) \quad (x = (x^{(1)}, \ldots, x^{(d)})^T \in \mathbb{R}^d) \tag{6}$$

for some univariate functions $m_1, \ldots, m_d : \mathbb{R} \to \mathbb{R}$, or the assumption that the regression function is a sum of ridge functions (cf. Friedman and Stuetzle 1981), i.e.,

$$m(x) = m_1(\beta_1^T x) + \cdots + m_{d^*}(\beta_{d^*}^T x) \quad (x \in \mathbb{R}^d) \tag{7}$$

for some $d^* \in \mathbb{N}$, $\beta_1, \ldots, \beta_{d^*} \in \mathbb{R}^d$, $m_1, \ldots, m_{d^*} : \mathbb{R} \to \mathbb{R}$. Note that Eq. (6) is a special case of Eq. (7) (choose $d^* = d$ and let $\beta_1, \ldots, \beta_d$ be the unit vectors). We show in the sequel that neural networks are able to adapt to both kinds of assumptions.

Theorem 2.2 covers the case of an additive regression function. Let $K_j := \text{card}(\text{supp}(X^{(j)}))$ define the class of neural networks as

$$\mathcal{F}_n := \left\{ f(x) = c_0 + \sum_{i=1}^{K_1+\cdots+K_d-d} c_i \sigma(a_i^{\mathrm{T}} x + b_i) : \ a_i \in \mathbb{R}^d, b_i, c_i \in \mathbb{R}, \right.$$
$$\left. |c_0| \leq 2Ld, |c_i| \leq 2L \ (i \geq 1) \right\}$$

and let $\tilde{m}_n(\cdot)$ and $m_n(\cdot)$ be defined as in the general case.

Theorem 2.2 *Let $L \geq 1$, $d \in \mathbb{N}$ and $K_1, \ldots, K_d \in \mathbb{N}$. Assume that the distribution of (X, Y) satisfies $(X, Y) \in \mathbb{R}^d \times [-L, L]$ a.s., $\text{card}(\text{supp}(X^{(j)})) = K_j$ $(j = 1, \ldots, d)$ and*

$$m(x) = m_1(x^{(1)}) + \cdots + m_d(x^{(d)}) \quad (x = (x^{(1)}, \ldots, x^{(d)})^T \in \mathbb{R}^d)$$

for some functions $m_1, \ldots, m_d : \mathbb{R} \to [-L, L]$. Let the neural network estimate m_n be defined as above. Then for all $n \in \mathbb{N}$

$$\mathbf{E} \int |m_n(x) - m(x)|^2 \mu(dx) \leq c_n \cdot \frac{K_1 + \cdots + K_d}{n}$$

where

$$c_n = 101L^2 \cdot (2d + 5) \cdot \log(48eL^2 n^3) = O(\log(n)).$$

Theorem 2.2 can be generalized elegantly to the case that the regression function is a sum of ridge functions, i.e., Eq. (7) holds. With $K_j^* := \text{card}(\text{supp}(\beta_j^T X))$ and

$$\mathcal{F}_n := \left\{ f(x) = c_0 + \sum_{i=1}^{K_1^*+\cdots+K_{d^*}^*-d^*} c_i \sigma(a_i^T x + b_i) : a_i \in \mathbb{R}^d, b_i \in \mathbb{R}, \right.$$
$$\left. |c_0| \leq 2Ld^*, |c_i| \leq 2L \ (i \geq 1) \right\},$$

we get

Theorem 2.3 *Let $L \geq 1$, $d, d^* \in \mathbb{N}$ and $K_1^*, \ldots, K_{d^*}^* \in \mathbb{N}$. Assume that the distribution of (X, Y) satisfies $(X, Y) \in \mathbb{R}^d \times [-L, L]$ a.s., $K_j^* = \text{card}(\text{supp}(\beta_j^T X))$ $(j = 1, \ldots, d^*)$ and*

$$m(x) = m_1(\beta_1^T x) + \cdots + m_{d^*}(\beta_{d^*}^T x) \quad (x \in \mathbb{R}^d)$$

for some $\beta_1, \ldots, \beta_{d^*} \in \mathbb{R}^d$ *and some functions* $m_1, \ldots, m_{d^*} : \mathbb{R} \to [-L, L]$. *Let the neural network estimate* m_n *be defined as above. Then for all* $n \in \mathbb{N}$

$$\mathbf{E} \int |m_n(x) - m(x)|^2 \mu(dx) \leq c_n \cdot \frac{K_1^* + \cdots + K_{d^*}^*}{n}$$

where

$$c_n = 101L^2 \cdot (2d + 5) \cdot \log(48eL^2n^3) = O(\log(n)).$$

Remark 1 Clearly, the constants in the theorems above are not optimal, but we believe that this is due to the mathematical difficulty in the (worst-case-) analysis of the neural network estimator and especially in the difficulty involved in bounding the covering number of sets of neural networks, and that the estimator really behaves better in practice. This can actually be seen in the setting of Theorem 2.1, where the neural network estimator will take on the same values as the estimator defined in Theorem 1 of Hamers and Kohler (2004) for the values of X which occur in the sample (because both are solutions to the minimization of the empirical L_2 risk), and makes an error of at most $2L$ for the values of X for which there are no observations in the sample. So, from looking at the proof of Theorem 1 in Hamers and Kohler (2004), it is clear that, under the assumptions of Theorem 2.1, the following error bound for the neural network estimator also holds:

$$\mathbf{E} \int |m_n(x) - m(x)|^2 \mu(\mathrm{d}x) \leq \left(\frac{(2L)^2}{e} + 2L^2 \right) \cdot \frac{K}{n}.$$

Remark 2 The error bounds in Theorems 2.2 and 2.3 are useful also if the dimension of the predictor variable X is high, provided the regression function has a structure which fits Theorem 2.2 or 2.3. Note that for high-dimensional data, the sums of the cardinalities of the supports of the components of X, i.e., $K_1 + \cdots + K_d$ or $K_1^* + \cdots + K_{d^*}^*$, respectively, can be much smaller than K, the cardinality of the entire support of X, for example, suppose $d = 20$ and X taking on two different values in each component, resulting in $K_1 + \cdots + K_{20} = 20 \cdot 2 = 40$ compared to the possible maximal cardinality of the entire support $K = 2^{20} \approx 10^6$.

Remark 3 In the above theorems the number of hidden neurons depends on the structure of the regression function and the distribution of X. Of course, this is not possible in an application because the distribution of (X, Y) (and in particular the regression function) is unknown. What can then be done is to consider the number of hidden neurons as a parameter of the estimate and to choose this parameter in a data-dependent way. A very popular method for doing this is splitting the data, into the so-called learning data and testing data, where the learning data is used to define estimates with various numbers of hidden neurons, then the empirical L_2 risk on the testing data is computed for each of these estimates, and finally the estimate with minimal empirical L_2 risk on the testing data is chosen. It follows from Theorem 2 in Hamers and Kohler (2003) that for neural network estimates obtained in this way similar bounds as in Theorems 2.1 to 2.3 hold.

Remark 4 In applications it is usually impossible to minimize the empirical L_2 risk over sets of neural networks because this leads to nonlinear optimization

problems. Instead, a steepest descent algorithm [the so-called- backpropagation, cf. Rummelhart and McClelland (1986)] is used to iteratively minimize the empirical L_2 risk. This algorithm usually leads to a local minimum of the L_2 risk, which is not guaranteed to be a global minimum.

3 A bound on the expected maximal deviations of sample averages from their means

A crucial step in the analysis of least squares estimates is to bound the difference between the L_2 risk and the empirical L_2 risk of the estimate (i.e., the difference between a mean and a sample average). In this section we introduce tools from empirical process theory which are helpful for this purpose. The approach we will use is similar to the one in Lee, Barlett and Williamson (1996), but we consider expectations instead of tail probabilities which enable us to reduce the constants in Theorem 4.1 below.

We first state a result proven in Hamers and Kohler (2003).

Lemma 3.1 *Let $L > 0$ and let $\mathcal{H}$ be a class of functions $h : \mathbb{R}^d \to \mathbb{R}$ bounded in absolute value by L. Let $X_1, \ldots, X_n$ be independent $\mathbb{R}^d$-valued random variables.*

Then, for all $c_1 > 0$

$$\mathbf{E}\left\{\max_{h\in\mathcal{H}}\left(\mathbf{E}\left\{\frac{1}{n}\sum_{i=1}^{n} h(X_i)\right\} - \frac{1}{n}\sum_{i=1}^{n} h(X_i) - c_1 \cdot \frac{1}{n}\sum_{i=1}^{n} \mathbf{Var}\{h(X_i)\}\right)\right\}$$
$$\leq \left(\frac{2L}{3} + \frac{1}{2c_1}\right)\frac{\log \operatorname{card}(\mathcal{H})}{n}.$$

Next, we extend this lemma to the case that $\mathcal{H}$ contains infinitely many functions. In order to avoid measurability problems in the case of uncountable collections of functions, we assume throughout this paper that the class of functions considered is permissible in the sense of Pollard (1984, Appendix C). This mild measurability condition is satisfied for most classes of functions used in the applications, including the classes of neural networks used in this paper.

We measure the "complexity" of a set $\mathcal{F}$ of functions $f : \mathbb{R}^d \to \mathbb{R}$ by the so-called covering numbers: Let $z_1, \ldots, z_n \in \mathbb{R}^d$ and set $z_1^n = (z_1, \ldots, z_n)$. Define the distance $d_1(f, g)$ between $f, g : \mathbb{R}^d \to \mathbb{R}$ by

$$d_1(f, g) = \frac{1}{n}\sum_{i=1}^{n} |f(z_i) - g(z_i)|.$$

An L_1-ϵ-cover of $\mathcal{F}$ on z_1^n is a set of functions $f_1, \ldots, f_k : \mathbb{R}^d \to \mathbb{R}$ with the property

$$\min_{1\leq j\leq k} d_1(f, f_j) < \epsilon \quad \text{for all } f \in \mathcal{F}.$$

Let $\mathcal{N}_1(\epsilon, \mathcal{F}, z_1^n)$ denote the cardinality k of the smallest L_1-ϵ-cover of $\mathcal{F}$ on z_1^n, and set $\mathcal{N}_1(\epsilon, \mathcal{F}, z_1^n) = \infty$ if there does not exist any L_1-ϵ-cover of finite cardinality of $\mathcal{F}$ on z_1^n.

With this notation, we can generalize the above lemma as follows.

Theorem 3.1 *Let $R > 0$ and let $\mathcal{G}$ be a class of functions $g : \mathbb{R}^d \to \mathbb{R}$ bounded in absolute value by R. Let $Z, Z_1, \dots, Z_{2n}$ be independent identically distributed $\mathbb{R}^d$-valued random variables. Then, for all $c_2 > 0, \epsilon > 0$,*

$$\mathbf{E}\left\{\sup_{g\in\mathcal{G}}\left\{\mathbf{E}\{g(Z)\} - \frac{1}{n}\sum_{i=1}^{n} g(Z_i) - c_2\mathbf{E}\{g^2(Z)\}\right\}\right\}$$
$$\leq \left(6c_2R^2 + \frac{4}{3}R + \frac{12}{5c_2}\right)\cdot\frac{\mathbf{E}\{\log\left(\mathcal{N}_1(\epsilon,\mathcal{G},Z_1^{2n})\right)\}}{n} + (6c_2R+2)\epsilon.$$

Proof The basic ideas of the proof follow the usual proof of Vapnik–Chervonenkis' Theorem. The proof will be divided into four steps.

Step 1. In the first step, we replace $\mathbf{E}\{g(Z)\}$ by a mean taken over a ghost sample and split the error into two terms. Let c_3 with $0 < c_3 < c_2/2$ be arbitrary. Then

$$\mathbf{E}\left\{\sup_{g\in\mathcal{G}}\left\{\mathbf{E}\{g(Z)\} - \frac{1}{n}\sum_{i=1}^{n} g(Z_i) - c_2\mathbf{E}\{g(Z)^2\}\right\}\right\}$$
$$= \mathbf{E}\left\{\sup_{g\in\mathcal{G}}\mathbf{E}\left\{\frac{1}{n}\sum_{i=1}^{n}(g(Z_{n+i}) - g(Z_i)) - c_2\mathbf{E}\{g(Z)^2\}\,\middle|\,Z_1,\dots,Z_n\right\}\right\}$$
$$\leq \mathbf{E}\left\{\sup_{g\in\mathcal{G}}\left\{\frac{1}{n}\sum_{i=1}^{n}(g(Z_{n+i}) - g(Z_i)) - c_2\mathbf{E}\{g(Z)^2\}\right\}\right\}$$
$$\text{(since } \sup\mathbf{E} \leq \mathbf{E}\sup\text{)}$$
$$= \mathbf{E}\left\{\sup_{g\in\mathcal{G}}\left\{\frac{1}{n}\sum_{i=1}^{n}(g(Z_{n+i}) - g(Z_i)) - \frac{c_3}{n}\sum_{i=1}^{n}(g(Z_{n+i})^2 + g(Z_i)^2)\right.\right.$$
$$\left.\left. + \frac{c_3}{n}\sum_{i=1}^{n}(g(Z_{n+i})^2 + g(Z_i)^2) - c_2\mathbf{E}\{g(Z)^2\}\right\}\right\}$$
$$\leq \mathbf{E}\left\{\sup_{g\in\mathcal{G}}\left\{\frac{1}{n}\sum_{i=1}^{n}(g(Z_{n+i}) - g(Z_i)) - \frac{c_3}{n}\sum_{i=1}^{n}(g(Z_{n+i})^2 + g(Z_i)^2)\right\}\right\}$$
$$+ 2\mathbf{E}\left\{\sup_{g\in\mathcal{G}}\left\{\frac{c_3}{n}\sum_{i=1}^{n} g(Z_i)^2 - \frac{c_2}{2}\mathbf{E}\{g(Z)^2\}\right\}\right\}. \tag{8}$$

Step 2. To bound the first term of the right-hand side of Eq. (8), we first introduce random signs: Since $Z_1, \dots, Z_{2n}$ are i.i.d. random variables, the value of the term considered remains unchanged if Z_i values are interchanged, especially if, for some $i \in \{1, \dots, n\}$, $g(Z_{n+i}) - g(Z_i)$ is replaced by $g(Z_i) - g(Z_{n+i})$ (and $g(Z_{n+i})^2 + g(Z_i)^2$ by $g(Z_i)^2 + g(Z_{n+i})^2$). We do this at random: Let U_i $(i = 1, \dots, n)$ be independent random variables independent of $Z_1, \dots, Z_{2n}$ such

that $\mathbf{P}\{U_i = 1\} = \mathbf{P}\{U_i = -1\} = 1/2$. Then

$$\mathbf{E}\left\{\sup_{g\in\mathcal{G}}\left\{\frac{1}{n}\sum_{i=1}^{n}(g(Z_{n+i}) - g(Z_i)) - \frac{c_3}{n}\sum_{i=1}^{n}(g(Z_{n+i})^2 + g(Z_i)^2)\right\}\right\}$$
$$= \mathbf{E}\left\{\sup_{g\in\mathcal{G}}\left\{\frac{1}{n}\sum_{i=1}^{n}U_i\cdot(g(Z_{n+i}) - g(Z_i))\right.\right.$$
$$\left.\left. -\frac{c_3}{n}\sum_{i=1}^{n}(g(Z_{n+i})^2 + g(Z_i)^2)\right\}\right\}.$$

Step 3. In this step, we introduce a finite covering of $\mathcal{G}$ so that we can apply Lemma 3.1. Since

$$\mathbf{E}\left\{\sup_{g\in\mathcal{G}}\left\{\frac{1}{n}\sum_{i=1}^{n}U_i\cdot(g(Z_{n+i}) - g(Z_i)) - \frac{c_3}{n}\sum_{i=1}^{n}(g(Z_{n+i})^2 + g(Z_i)^2)\right\}\right\}$$
$$= \mathbf{E}\left\{\mathbf{E}\left\{\sup_{g\in\mathcal{G}}\left\{\frac{1}{n}\sum_{i=1}^{n}U_i\cdot(g(Z_{n+i}) - g(Z_i))\right.\right.\right.$$
$$\left.\left.\left. -\frac{c_3}{n}\sum_{i=1}^{n}(g(Z_{n+i})^2 + g(Z_i)^2)\right\}\,\middle|\, Z_1,\dots,Z_{2n}\right\}\right\},$$

let us first consider

$$\mathbf{E}\left\{\sup_{g\in\mathcal{G}}\left\{\frac{1}{n}\sum_{i=1}^{n}U_i\cdot(g(z_{n+i}) - g(z_i)) - \frac{c_3}{n}\sum_{i=1}^{n}(g(z_{n+i})^2 + g(z_i)^2)\right\}\right\}$$

for fixed $z_i \in \mathbb{R}^d (i = 1,\dots,2n)$.
Let $\mathcal{G}^*$ be a L_1-ϵ-cover of minimal cardinality of $\mathcal{G}$ on $z_1,\dots,z_{2n}$, i.e., for each $g \in \mathcal{G}$ there is a function $g^* \in \mathcal{G}^*$ such that

$$\frac{1}{2n}\sum_{i=1}^{2n}|g(z_i) - g^*(z_i)| < \epsilon,$$

and $\mathrm{card}(\mathcal{G}^*) = \mathcal{N}_1(\epsilon, \mathcal{G}, z_1^{2n})$.
W.l.o.g. we can choose $\mathcal{G}^*$ such that $|g^*(x)| \le R$ for all $x \in \mathbb{R}^d$, g^* in $\mathcal{G}^*$. (It is understood that $\mathcal{G}^*$ may depend on z_1^{2n}, even if we did not show this in the notation.)
The following calculations bound the error which may arise from replacing any $g \in \mathcal{G}$ by its corresponding g^*, the function in $\mathcal{G}^*$ closest to it (in the empirical L_1-norm):

$$
\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n} U_i(g(z_{i+n}) - g(z_i)) \\
&\quad = \frac{1}{n}\sum_{i=1}^{n} U_i(g(z_{i+n}) - g^*(z_{i+n}) + g^*(z_{i+n}) - g^*(z_i) + g^*(z_i) - g(z_i)) \\
&\quad \le \frac{1}{n}\sum_{i=1}^{n} U_i(g^*(z_{i+n}) - g^*(z_i)) + \frac{1}{n}\sum_{i=1}^{2n} |g(z_i) - g^*(z_i)| \\
&\quad \le \frac{1}{n}\sum_{i=1}^{n} U_i(g^*(z_{i+n}) - g^*(z_i)) + 2\epsilon
\end{aligned}
$$

and

$$
\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n} (g(z_i)^2 + g(z_{n+i})^2) \\
&\quad = \frac{1}{n}\sum_{i=1}^{n} (g^*(z_i)^2 + g^*(z_{n+i})^2) - \frac{1}{n}\sum_{i=1}^{2n} (g^*(z_i)^2 - g(z_i)^2) \\
&\quad = \frac{1}{n}\sum_{i=1}^{n} (g^*(z_i)^2 + g^*(z_{n+i})^2) - \frac{1}{n}\sum_{i=1}^{2n} (g^*(z_i) - g(z_i))(g^*(z_i) + g(z_i)) \\
&\quad \ge \frac{1}{n}\sum_{i=1}^{n} (g^*(z_i)^2 + g^*(z_{n+i})^2) - \frac{1}{n}\sum_{i=1}^{2n} |g^*(z_i) - g(z_i)||g^*(z_i) + g(z_i)| \\
&\quad \ge \frac{1}{n}\sum_{i=1}^{n} (g^*(z_i)^2 + g^*(z_{n+i})^2) - 4R\epsilon.
\end{aligned}
$$

With this, we obtain

$$
\begin{aligned}
&\mathbf{E}\left\{\sup_{g\in\mathcal{G}}\left\{\frac{1}{n}\sum_{i=1}^{n} U_i \cdot (g(z_{n+i}) - g(z_i)) - \frac{c_3}{n}\sum_{i=1}^{n} (g(z_{n+i})^2 + g(z_i)^2)\right\}\right\} \\
&\quad \le \mathbf{E}\left\{\max_{g\in\mathcal{G}^*}\left\{\frac{1}{n}\sum_{i=1}^{n} U_i \cdot (g(z_{n+i}) - g(z_i))\right.\right. \\
&\qquad \left.\left. -\frac{c_3}{n}\sum_{i=1}^{n} (g(z_{n+i})^2 + g(z_i)^2) + (4Rc_3 + 2)\epsilon\right\}\right\} \\
&\quad \le \mathbf{E}\left\{\max_{g\in\mathcal{G}^*}\left\{\frac{1}{n}\sum_{i=1}^{n} U_i \cdot (g(z_{n+i}) - g(z_i)) - \frac{c_3}{2n}\sum_{i=1}^{n} (g(z_{n+i}) - g(z_i))^2\right\}\right\} \\
&\qquad + (4Rc_3 + 2)\,\epsilon \\
&\qquad\qquad\qquad\qquad (\text{since } a^2 + b^2 \ge \tfrac{(a-b)^2}{2}), \\
&\quad \le \left(\frac{4R}{3} + \frac{1}{c_3}\right) \cdot \frac{\log(\mathcal{N}_1(\epsilon, \mathcal{G}, z_1^{2n}))}{n} + (4Rc_3 + 2)\epsilon,
\end{aligned}
$$

where we have applied Lemma 3.1 (with $c_1 = c_3/2$ and $L = 2R$, $X_i = i \cdot U_i$, $h(j) = -\text{sign}(j) \cdot (g(z_{n+|j|}) - g(z_{|j|}))$) in the last step. (Note that $\mathbf{E}\{\frac{1}{n}h(X_i)\} = 0$ since we have chosen $\mathbf{P}\{U_i = 1\} = \mathbf{P}\{U_i = -1\} = 1/2$.)
Taking the expectation, we obtain the following bound for the first term of the right-hand side of Eq. (8):

$$\mathbf{E}\left\{\sup_{g\in\mathcal{G}}\left\{\frac{1}{n}\sum_{i=1}^{n} g(Z_{n+i}) - g(Z_i) - \frac{c_3}{n}\sum_{i=1}^{n}(g(Z_{n+i})^2 + g(Z_i)^2)\right\}\right\}$$
$$\leq \left(\frac{4R}{3} + \frac{1}{c_3}\right) \cdot \frac{\mathbf{E}\{\log(\mathcal{N}_1(\epsilon, \mathcal{G}, Z_1^{2n}))\}}{n} + (4Rc_3 + 2)\epsilon.$$

Step 4. By applying the techniques of the Steps 1–3 (with g replaced by g^2) to the second term of the right-hand side of Eq. (8), one gets, after a long but straightforward calculation, the following bound:

$$2\mathbf{E}\left\{\sup_{g\in\mathcal{G}}\left\{\frac{c_3}{n}\sum_{i=1}^{n} g(Z_i)^2 - \frac{c_2}{2}\mathbf{E}\{g(Z)^2\}\right\}\right\}$$
$$\leq \frac{c_2 + 2c_3}{2} \cdot \left(\frac{2}{3} + \frac{(c_2 + 2c_3)}{2(c_2 - 2c_3)}\right) R^2 \cdot \frac{\mathbf{E}\{\log(\mathcal{N}_1(\epsilon, \mathcal{G}, Z_1^{2n}))\}}{n}$$
$$+(6c_2 - 4c_3)R\epsilon.$$

For our purposes, $c_3 = 0.42 \cdot c_2$ is a good choice. The above term can then be bounded from above by

$$6R^2 \cdot c_2 \cdot \frac{\mathbf{E}\{\log(\mathcal{N}_1(\epsilon, \mathcal{G}, Z_1^{2n}))\}}{n} + (6c_2 - 4c_3)R\epsilon,$$

and the term obtained in Step 3 by

$$\left(\frac{4R}{3} + \frac{12}{5c_2}\right) \cdot \frac{\mathbf{E}\{\log(\mathcal{N}_1(\epsilon, \mathcal{G}, Z_1^{2n}))\}}{n} + (4Rc_3 + 2)\epsilon.$$

Summing up these bounds, the proof is complete. □

4 A general result on least squares regression estimates

Let $\mathcal{F}_n$ be a set of functions $f : \mathbb{R}^d \to \mathbb{R}$. Choose $\tilde{m}_n \in \mathcal{F}_n$ such that

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

(For better readability and ease of notation, we assume here that the infimum is taken for some $f \in \mathcal{F}_n$ and omit the arbitrarily small ϵ which had to be added throughout otherwise.)

Define $m_n(\cdot)$ by truncating $\tilde{m}_n(\cdot)$ at $\pm L$. Then the following theorem holds:

Theorem 4.1 *Let $n \in \mathbb{N}$ and $1 \leq L < \infty$. Assume $|Y| \leq L$ a.s. Then, for m_n defined as above,*

$$\mathbf{E}\left\{\int |m_n(x) - m(x)|^2 \mu(\mathrm{d}x)\right\} \leq 100L^2 \cdot \frac{\mathbf{E}\left\{\log\left(\mathcal{N}_1\left(\frac{1}{4Ln}, \mathcal{F}_n, X_1^{2n}\right)\right)\right\}}{n} + \frac{7}{n} + 2 \inf_{f \in \mathcal{F}_n} \left\{\int |f(x) - m(x)|^2 \mu(\mathrm{d}x)\right\}.$$

Proof We use the error decomposition

$$\begin{aligned}
&\int |m_n(x) - m(x)|^2 \mu(\mathrm{d}x) \\
&= \mathbf{E}\{|m_n(X) - Y|^2 | \mathcal{D}_n\} - \mathbf{E}\{|m(X) - Y|^2\} \\
&= \mathbf{E}\{|m_n(X) - Y|^2 | \mathcal{D}_n\} - \mathbf{E}\{|m(X) - Y|^2\} \\
&\quad -2\left(\frac{1}{n}\sum_{i=1}^n |m_n(X_i) - Y_i|^2 - \frac{1}{n}\sum_{i=1}^n |m(X_i) - Y_i|^2\right) \\
&\quad +2\left(\frac{1}{n}\sum_{i=1}^n |m_n(X_i) - Y_i|^2 - \frac{1}{n}\sum_{i=1}^n |m(X_i) - Y_i|^2\right).
\end{aligned}$$

First, consider

$$\begin{aligned}
&\mathbf{E}\left\{\frac{1}{n}\sum_{i=1}^n |m_n(X_i) - Y_i|^2 - \frac{1}{n}\sum_{i=1}^n |m(X_i) - Y_i|^2\right\} : \\
&\mathbf{E}\left\{\frac{1}{n}\sum_{i=1}^n |m_n(X_i) - Y_i|^2 - \frac{1}{n}\sum_{i=1}^n |m(X_i) - Y_i|^2\right\} \\
&\leq \mathbf{E}\left\{\frac{1}{n}\sum_{i=1}^n |\tilde{m}_n(X_i) - Y_i|^2 - \frac{1}{n}\sum_{i=1}^n |m(X_i) - Y_i|^2\right\} \\
&\qquad \text{(since } |Y_i| \leq L \text{ a.s.)} \\
&= \mathbf{E}\left\{\inf_{f \in \mathcal{F}_n}\left\{\frac{1}{n}\sum_{i=1}^n |f(X_i) - Y_i|^2 - \frac{1}{n}\sum_{i=1}^n |m(X_i) - Y_i|^2\right\}\right\} \\
&\leq \inf_{f \in \mathcal{F}_n} \mathbf{E}\left\{\frac{1}{n}\sum_{i=1}^n |f(X_i) - Y_i|^2 - \frac{1}{n}\sum_{i=1}^n |m(X_i) - Y_i|^2\right\} \\
&= \inf_{f \in \mathcal{F}_n} \mathbf{E}\left\{|f(X) - Y|^2 - |m(X) - Y|^2\right\} \\
&= \inf_{f \in \mathcal{F}_n} \int |f(x) - m(x)|^2 \mu(\mathrm{d}x).
\end{aligned}$$

It remains to show that

$$\mathbf{E}\Bigg\{\mathbf{E}\{|m_n(X)-Y|^2|\mathcal{D}_n\}-\mathbf{E}\{|m(X)-Y|^2\} \\ -2\bigg(\frac{1}{n}\sum_{i=1}^n |m_n(X_i)-Y_i|^2-\frac{1}{n}\sum_{i=1}^n |m(X_i)-Y_i|^2\bigg)\Bigg\} \\ \leq 100L^2\cdot\frac{\mathbf{E}\{\log\left(\mathcal{N}_1(\frac{1}{4Ln},\mathcal{F}_n,X_1^{2n})\right)\}}{n}+\frac{7}{n}.$$

Set $Z=(X,Y)$, $Z_i=(X_i,Y_i)$ and $g(z)=g(x,y)=(m_n(x)-y)^2-(m(x)-y)^2$ for $|y|\leq L$, $g(z)=0$ else. Then, $|g(x,y)|\leq 4L^2$ for all $(x,y)\in\mathbb{R}^d\times\mathbb{R}$ and

$$\mathbf{E}\{g(Z)^2|\mathcal{D}_n\}\leq 8L^2\mathbf{E}\{g(Z)|\mathcal{D}_n\}. \tag{9}$$

(cf. Barron 1991, Eq. (39)).

Now, we can rewrite

$$\mathbf{E}\{|m_n(X)-Y|^2|\mathcal{D}_n\}-\mathbf{E}\{|m(X)-Y|^2\} \\ -2\bigg(\frac{1}{n}\sum_{i=1}^n |m_n(X_i)-Y_i|^2-\frac{1}{n}\sum_{i=1}^n |m(X_i)-Y_i|^2\bigg)$$

as

$$\mathbf{E}\{g(Z)|\mathcal{D}_n\}-\frac{2}{n}\sum_{i=1}^n g(Z_i) \\ =2\left(\mathbf{E}\{g(Z)|\mathcal{D}_n\}-\frac{1}{n}\sum_{i=1}^n g(Z_i)-\frac{1}{2}\mathbf{E}\{g(Z)|\mathcal{D}_n\}\right) \\ \leq 2\left(\mathbf{E}\{g(Z)|\mathcal{D}_n\}-\frac{1}{n}\sum_{i=1}^n g(Z_i)-\frac{1}{16L^2}\mathbf{E}\{g(Z)^2|\mathcal{D}_n\}\right),$$

where we have used Eq. (9) in the last step. Applying Theorem 3.1 (with $c_2=\frac{1}{16L^2}$, $R=4L^2$, $\epsilon=\frac{1}{n}$) to the last term, we obtain

$$\mathbf{E}\Bigg\{\mathbf{E}\{|m_n(X)-Y|^2|\mathcal{D}_n\}-\mathbf{E}\{|m(X)-Y|^2\} \\ -2\bigg(\frac{1}{n}\sum_{i=1}^n |m_n(X_i)-Y|^2-\frac{1}{n}\sum_{i=1}^n |m(X_i)-Y_i|^2\bigg)\Bigg\} \\ \leq 2\left(\left(6L^2+\frac{16}{3}L^2+\frac{12\cdot 16L^2}{5}\right)\cdot\frac{\mathbf{E}\left\{\log\left(\mathcal{N}_1\left(\frac{1}{n},\mathcal{G}_n,Z_1^{2n}\right)\right)\right\}}{n}+\frac{\frac{3}{2}+2}{n}\right),$$

where $\mathcal{G}_n$ is the class of functions

$$\Bigg\{g:\mathbb{R}^d\times\mathbb{R}\to\mathbb{R}: g(x,y)=(f(x)-y)^2-(m(x)-y)^2\ ((x,y)\in\mathbb{R}^d\times\mathbb{R}) \\ \text{for some } f\in\bar{\mathcal{F}}_n\Bigg\},$$

and $\bar{\mathcal{F}}_n$ is the class of all functions which can be obtained from functions contained in $\mathcal{F}_n$ by truncation at $\pm L$. Since, for any $g_1, g_2 \in \mathcal{G}_n$ and $z_1^{2n} \subseteq \mathbb{R}^d \times [-L, L]$

$$\begin{aligned}
&\frac{1}{2n}\sum_{i=1}^{2n} |g_1(z_i) - g_2(z_i)| \\
&= \frac{1}{2n}\sum_{i=1}^{2n} |(f_1(x_i) - y_i)^2 - (m(x_i) - y_i)^2 \\
&\quad -((f_2(x_i) - y_i)^2 - (m(x_i) - y_i)^2)| \\
&= \frac{1}{2n}\sum_{i=1}^{2n} |(f_1(x_i) - y_i)^2 - (f_2(x_i) - y_i)^2| \\
&= \frac{1}{2n}\sum_{i=1}^{2n} |(f_1(x_i) + f_2(x_i) - 2y_i)(f_1(x_i) - f_2(x_i))| \\
&\leq 4L \cdot \frac{1}{2n}\sum_{i=1}^{2n} |f_1(x_i) - f_2(x_i)|,
\end{aligned}$$

we can construct an L_1-$4L\epsilon$-cover of $\mathcal{G}_n$ on z_1^{2n} from an L_1-ϵ-cover of $\bar{\mathcal{F}}_n$ on x_1^{2n}, which implies

$$\mathcal{N}_1\left(\frac{1}{n}, \mathcal{G}_n, z_1^{2n}\right) \leq \mathcal{N}_1\left(\frac{1}{4Ln}, \bar{\mathcal{F}}_n, x_1^{2n}\right) \leq \mathcal{N}_1\left(\frac{1}{4Ln}, \mathcal{F}_n, x_1^{2n}\right),$$

and the proof is complete. □

5 Approximation properties of neural networks

In this section, we derive three approximation results to bound

$$\inf_{f \in \mathcal{F}_n} \int |f(x) - m(x)|^2 \mu(dx)$$

for general regression functions, for additive regression functions and for regression functions which are a sum of ridge functions. We start with the general case.

Lemma 5.1 *Let $K \in \mathbb{N}$ and let $m : \mathbb{R}^d \to \mathbb{R}$ be an arbitrary function bounded in absolute value by L. Suppose that the distribution μ of X is concentrated on a subset of $\mathbb{R}^d$ of cardinality K.*

Then, for arbitrary $\delta > 0$ and arbitrary-squashing function σ, there is a neural network of the form $f(x) = c_0 + \sum_{i=1}^{K-1} c_i \sigma(a_i^{\mathrm{T}} x + b_i)$, where $a_i \in \mathbb{R}^d$, $b_i, c_i \in \mathbb{R}$ and $|c_i| \leq 2L$, such that

$$\int |f(x) - m(x)|^2 \mu(\mathrm{d}x) < \delta.$$

Proof W.l.o.g., assume $\sqrt{\delta} < K^2 L$. Set $\text{supp}(X) =: \{x_1, \dots, x_K\}$. Since for each of the $\frac{K(K-1)}{2}$ subsets of the form $\{x_i, x_j\}$ $(i \neq j)$, the set of all vectors $v \in \mathbb{R}^d$ for which $v^{\mathrm{T}} x_i = v^{\mathrm{T}} x_j$ (i.e., the set of all $v \in \mathbb{R}^d$ such that $v^{\mathrm{T}}(x_i - x_j) = 0$) is a hyperplane in $\mathbb{R}^d$, the set of all vectors $v \in \mathbb{R}^d$ for which there is at least one pair (x_i, x_j) $(i \neq j)$ such that $v^T x_i = v^T x_j$ is the union of at most $\frac{K(K-1)}{2}$ hyperplanes in $\mathbb{R}^d$, and thus a proper subset of $\mathbb{R}^d$. So, there is a vector $\tilde{a} \in \mathbb{R}^d$ such that $\tilde{a}^{\mathrm{T}} x_i \neq \tilde{a}^{\mathrm{T}} x_j$ for all $1 \leq i < j \leq K$. Define $t_1, \dots, t_K \in \mathbb{R}^d$ such that $\{t_1, \dots, t_K\} = \{x_1, \dots, x_K\}$ and $\tilde{a}^{\mathrm{T}} t_1 < \tilde{a}^{\mathrm{T}} t_2 < \cdots < \tilde{a}^T t_K$.
Since σ is a squashing function, there are $z_1 < z_2 \in \mathbb{R}$ such that $\sigma(z_1) < \frac{\sqrt{\delta}}{2K^2 L}$ and $\sigma(z_2) > 1 - \frac{\sqrt{\delta}}{2K^2 L}$. Set

$$a = \frac{z_2 - z_1}{\min_{i \in \{1,\dots,K-1\}} \{\tilde{a}^T t_{i+1} - \tilde{a}^T t_i\}} \cdot \tilde{a} \text{ and } b_j = -a^T t_j + z_1 \quad \text{for } j = 1, \dots, K-1.$$

Then, by monotonicity of σ and the choice of z_1, z_2,

$$\sigma(a^T t_i + b_j) \leq \sigma(a^T t_j + b_j) = \sigma(a^T t_j - a^T t_j + z_1) < \frac{\sqrt{\delta}}{2K^2 L} \quad \text{for } i \leq j$$

and

$$\begin{aligned}
\sigma(a^{\mathrm{T}} t_i + b_j) &\geq \sigma(a^{\mathrm{T}}(t_{j+1} - t_j) + a^{\mathrm{T}} t_j + b_j) \\
&= \sigma\left(\frac{z_2 - z_1}{\min_{i \in \{1,\dots,K-1\}} \{\tilde{a}^T t_{i+1} - \tilde{a}^{\mathrm{T}} t_i\}} \cdot \tilde{a}^{\mathrm{T}}(t_{j+1} - t_j) + z_1 \right) \\
&\geq \sigma(z_2 - z_1 + z_1) \\
&> 1 - \frac{\sqrt{\delta}}{2K^2 L}, \quad \text{for } i > j.
\end{aligned}$$

With $a_i = a$, $b_i = -a^{\mathrm{T}} t_i + z_1$, $c_0 = m(t_1)$ and $c_i = m(t_{i+1}) - m(t_i)$ $(i = 1, \dots, K-1)$, set

$$f(x) = c_0 + \sum_{i=1}^{K-1} c_i \sigma(a_i^{\mathrm{T}} x + b_i).$$

Obviously, f satisfies the conditions imposed on the a_i's, b_i's and c_i's.

For $j = 1, \dots, K-1$,

$$\begin{aligned}
&|m(t_{j+1}) - m(t_j) - (f(t_{j+1}) - f(t_j))| \\
&= |c_j - \sum_{i=1}^{K-1} c_i (\sigma(a^{\mathrm{T}} t_{j+1} + b_i) - \sigma(a^{\mathrm{T}} t_j + b_i))| \\
&\leq |c_j (1 - (\sigma(a^{\mathrm{T}} t_{j+1} + b_j) - \sigma(a^{\mathrm{T}} t_j + b_j)))| \\
&\quad + |\sum_{\substack{i=1 \\ i \neq j}}^{K-1} c_i (\sigma(a^{\mathrm{T}} t_{j+1} + b_i) - \sigma(a^{\mathrm{T}} t_j + b_i))|
\end{aligned}$$

$$< 2L \cdot \frac{\sqrt{\delta}}{K^2 L} + (K-2)2L \cdot \frac{\sqrt{\delta}}{2K^2 L}$$
$$= \frac{\sqrt{\delta}}{K}.$$

Thus, for $j = 1, \dots, K$,

$$|m(t_j) - f(t_j)|$$
$$= |\sum_{i=1}^{j-1} \{m(t_{i+1}) - m(t_i) - (f(t_{i+1}) - f(t_i))\} + m(t_1) - f(t_1)|$$
$$\leq (j-1) \cdot \frac{\sqrt{\delta}}{K} + |c_0 - c_0 - \sum_{i=1}^{K-1} c_i \sigma(a^T t_1 + b_i)|$$
$$\leq (j-1) \cdot \frac{\sqrt{\delta}}{K} + (K-1) \cdot 2L \cdot \frac{\sqrt{\delta}}{2K^2 L} < \frac{j\sqrt{\delta}}{K} \leq \sqrt{\delta},$$

which together with $\text{supp}(X) = \{x_1, \dots, x_K\} = \{t_1, \dots t_K\}$ implies

$$\int |f(x) - m(x)|^2 \mu(\mathrm{d}x) \leq \sup_{x \in \text{supp}(X)} |f(x) - m(x)|^2 < \delta.$$

□

The next lemma covers the case of an additive regression function.

Lemma 5.2 *Let $K_1, \dots, K_d \in \mathbb{N}$ and let $m : \mathbb{R}^d \to \mathbb{R}$ be a function of the form $m(x) = m_1(x^{(1)}) + \cdots + m_d(x^{(d)})$, where the m_i are univariate functions bounded in absolute value by L. Suppose that the distribution of the jth component $X^{(j)}$ of X is concentrated on a subset of $\mathbb{R}$ of cardinality K_j. Then, for arbitrary $\delta > 0$ and arbitrary-squashing function σ, there is a neural network of the form $f(x) = c_0 + \sum_{i=1}^{K_1 + \cdots + K_d - d} c_i \sigma(a_i^{\mathrm{T}} x + b_i)$, where $a_i \in \mathbb{R}^d$, $b_i, c_i \in \mathbb{R}$, $|c_0| \leq 2Ld$ and $|c_i| \leq 2L (i \geq 1)$, such that*

$$\int |f(x) - m(x)|^2 \mu(\mathrm{d}x) < \delta. \tag{10}$$

Proof From the proof of Lemma 5.1, it is easily concluded that there are neural networks $f_1, \dots, f_d$ of the form

$$f_j(t) = c_{j,0} + \sum_{i=1}^{K_j - 1} c_{j,i} \cdot \sigma(a_j^* t + b_{j,i}) \quad (t \in \mathbb{R}, a_j^*, b_{j,i}, c_{j,i} \in \mathbb{R})$$

with

$$|m_j(t) - f_j(t)| < \frac{\sqrt{\delta}}{d} \text{ for } t \in \text{supp}(X^{(j)}). \tag{11}$$

Set $a_j = a_j^* \cdot e_j$ (e_j denoting the jth d-dimensional unit vector), $c_0 = \sum_{j=1}^d c_{j,0}$ and $f(x) = c_0 + \sum_{j=1}^d \sum_{i=1}^{K_j-1} c_{j,i} \cdot \sigma(a_j^{\mathrm{T}} x + b_{j,i})$. Then for any $x \in \mathrm{supp}(X)$ we have $x^{(j)} \in \mathrm{supp}(X^{(j)})$ and hence

$$\begin{aligned}
&|m(x) - f(x)| \\
&= |\sum_{j=1}^d m_j(x^{(j)}) - \sum_{j=1}^d \{c_{j,0} + \sum_{i=1}^{K_j-1} c_{j,i} \cdot \sigma(a_j^T x + b_{j,i})\}| \\
&= |\sum_{j=1}^d \{m_j(x^{(j)}) - c_{j,0} - \sum_{i=1}^{K_j-1} c_{j,i} \cdot \sigma(a_j^* x^{(j)} + b_{j,i})\}| \\
&< d \cdot \frac{\sqrt{\delta}}{d} = \sqrt{\delta}.
\end{aligned}$$

From this the assertion follows as in the proof of Lemma 5.1. □

Additive models can be generalized by replacing the components of X by projections of X onto vectors $\beta_j \in \mathbb{R}^d$, and by assuming that $m(x)$ is a sum of univariate functions m_j, where each of these univariate functions is applied to one of the very projections of X:

$$m(x) = \sum_{j=1}^{d^*} m_j(\beta_j^{\mathrm{T}} x) \qquad \text{for some } \beta_1, \ldots, \beta_{d^*} \in \mathbb{R}^d.$$

Choosing the $f_1, \ldots, f_{d^*}$ such that Eq. (11) in the proof of Lemma 5.2 holds for $t \in \mathrm{supp}(\beta_j^{\mathrm{T}} X)$, and replacing a_j by $a_j^* \cdot \beta_j$, the proof of Lemma 5.3 is exactly along the lines of the proof of Lemma 5.2.

Lemma 5.3 *Let $d^* \in \mathbb{N}$, $\beta_1, \ldots, \beta_{d^*} \in \mathbb{R}^d$ and let $m : \mathbb{R}^d \to \mathbb{R}$ be a function of the form*

$$m(x) = m_1(\beta_1^{\mathrm{T}} x) + \cdots + m_{d^*}(\beta_{d^*}^{\mathrm{T}} x) \quad (x \in \mathbb{R}^d),$$

where the m_i are univariate functions bounded in absolute value by L. Set $K_j^* = \mathrm{card}(\mathrm{supp}(\beta_j^{\mathrm{T}} X))$. Then, for arbitrary $\delta > 0$ and arbitrary-squashing function σ, there is a neural network of the form

$$f(x) = c_0 + \sum_{i=1}^{K_1^* + \cdots + K_{d^*}^* - d^*} c_i \sigma(a_i^{\mathrm{T}} x + b_i),$$

where $a_i \in \mathbb{R}^d$, $b_i, c_i \in \mathbb{R}$, $|c_0| \leq 2Ld^*$ and $|c_i| \leq 2L$ ($i \geq 1$), such that

$$\int |f(x) - m(x)|^2 \mu(\mathrm{d}x) < \delta.$$

6 Proof of Theorems 2.1–2.3

Let $\mathcal{F}_n$ be the set of neural networks considered in Theorem 2.1, 2.2 or 2.3. Then, according to Lemma 5.1, 5.2 or 5.3, respectively, the assumptions on X and on the structure of m imply

$$\inf_{f \in \mathcal{F}_n} \int |f(x) - m(x)|^2 \mu(\mathrm{d}x) = 0.$$

Thus, the application of Theorem 4.1 yields

$$\mathbf{E}\left\{\int |m_n(x) - m(x)|^2 \mu(\mathrm{d}x)\right\} \leq 100L^2 \cdot \frac{\mathbf{E}\left\{\log\left(\mathcal{N}_1\left(\frac{1}{4Ln}, \mathcal{F}_n, X_1^{2n}\right)\right)\right\}}{n} + \frac{7}{n}.$$

So all we need in the sequel is a bound on the covering number.

With standard techniques from Vapnik–Chervonenkis-theory, it can be shown that for

$$\mathcal{F}_n = \left\{ f(x) = c_0 + \sum_{i=1}^{k_n} c_i \sigma(a_i^T x + b_i) : a_i \in \mathbb{R}^d, b_i, c_i \in \mathbb{R}, \sum_{i=0}^{k_n} |c_i| \leq \gamma_n \right\},$$

the L_1-covering number can be upper-bounded by

$$\mathcal{N}_1\left(\epsilon, \mathcal{F}_n, X_1^n\right) \leq \left(\frac{6e\gamma_n(k_n+1)}{\epsilon}\right)^{(2d+5)k_n+1}$$

(see Györfi et al. 2002, proof of Theorem 16.1).

So, for $\mathcal{F}_n$ as in Theorem 2.1, we get

$$\mathcal{N}_1\left(\frac{1}{4Ln}, \mathcal{F}_n, X_1^{2n}\right) \leq \left(\frac{6e \cdot 2LK \cdot K}{\frac{1}{4Ln}}\right)^{(2d+5)(K-1)+1},$$

giving

$$\mathbf{E}\left\{\log\left(\mathcal{N}_1\left(\frac{1}{4Ln}, \mathcal{F}_n, X_1^{2n}\right)\right)\right\} \leq (2d+5)K \log(48eL^2K^2n),$$

for $\mathcal{F}_n$ as in Theorem 2.2

$$\mathcal{N}_1\left(\frac{1}{4Ln}, \mathcal{F}_n, X_1^{2n}\right) \leq \left(\frac{6e \cdot 2L \cdot (K_1 + \cdots + K_d)^2}{\frac{1}{4Ln}}\right)^{(2d+5)(K_1+\cdots+K_d-d)+1},$$

and for $\mathcal{F}_n$ as in Theorem 2.3

$$\mathcal{N}_1\left(\frac{1}{4Ln}, \mathcal{F}_n, X_1^{2n}\right) \leq \left(\frac{6e \cdot 2L \cdot (K_1^* + \cdots + K_{d^*}^*)^2}{\frac{1}{4Ln}}\right)^{(2d+5)(K_1^*+\cdots+K_{d^*}^*-d^*)+1},$$

yielding similar bounds on $\mathbf{E}\left\{\log\left(\mathcal{N}_1\left(\frac{1}{4Ln}, \mathcal{F}_n, X_1^{2n}\right)\right)\right\}$ and thus completing the proofs of these three results. □

Acknowledgements The authors wish to thank an anonymous referee for many detailed and helpful comments.

References

Andrews, D.W.K., Whang, Y.J. (1990). Additive interactive regression models: Circumvention of the curse of dimensionality. *Econometric Theory*, *6*, 466–479.

Anthony, M., Bartlett, P. (1999). *Neural network learning: theoretical foundations*, Cambridge: Cambridge University Press.

Barron, A.R. (1991). Complexity regularization with application to artificial neural networks, In G. Roussas (Ed.) *Nonparametric functional estimation and related topics*, NATO ASI Series, (pp. 561–576) Dodrecht: Kluwer.

Barron, A.R. (1994). Approximation and estimation bounds for artificial neural networks, *Machine Learning*, *14*, 115–133.

Bickel, P.J., Klaasen, C.A.J., Ritov, Y., Wellner, J.A. (1993). *Efficent and adaptive estimation for semiparametric models*, Baltimore: The John Hopkins University Press.

Breiman, L. (1993). Fitting additive models to regression data. *Computational Statistics and Data Analysis*, *15*, 13–46.

Breiman, L., Freiman, J.H. (1985). Estimating optimal transformations for multiple regression and correlation. *Journal of the American Statistical Association*, *80*, 580–598.

Burman, P. (1990). Estimation of generalized additive models. *Journal of Multivariate Analysis*, *32*, 230–255.

Chen, Z. (1991). Interaction spline models and their convergence rates. *Annals of Statistics*, *19*, 1855–1868.

Cybenko, G. (1989). Approximations by superpositions of sigmoidal functions. *Mathematic Control, Signals, Systems*, *2*, 303–314.

Devroye, L., Györfi, L., Lugosi, G. (1996). *A Probabilistic Theory of Pattern Recognition*. Berlin Heidelberg New York: Springer.

Friedman, J.H., Stuetzle, W. (1981). Projection pursuit regression. *Journal of the American Statistical Association*, *76*, 817–823.

Funahashi, K. (1989). On the approximate realization of continuous mappings by neural networks. *Neural Networks*, *2*, 183–192.

Györfi, L., Kohler, M., Krzyżak, A., Walk, H. (2002). *A distribution-free theory of nonparametric regression*. Berlin Heidelberg New York: Springer.

Hamers, M., Kohler, M. (2003). A bound on the expected maximal deviation of averages from their means. *Statistics & Probability Letters*, *62*, 137–144.

Hamers, M., Kohler, M. (2004). How well can a regression function be estimated if the distribution of the (random) design is concentrated on a finite set? *Journal of Statistical Planning and Inference*, *123*, 377–394.

Hastie, T., Tibshirani, R.J. (1990). *Generalized additive models*. London: Chapman and Hall.

Hertz, J., Krogh, A., Palmer, R.G. (1991). *Introduction to the Theory of Neural Computation*. Redwood City: Addison-Wesley.

Hornik, K., Stinchcombe, M., White, H. (1989). Multi-layer feedforward networks are universal approximators. *Neural Networks*, *2*, 359–366.

Huang, J. (1998). Projection estimation in multiple regression with applications to functional anova models. *Annals of Statistics*, *26*, 242–272.

Kohler, M. (1998). Nonparametric regression function estimation using interaction least squares splines and complexity regularization. *Metrika*, *47*, 147–163.

Lee, W.S., Bartlett, P.L., Williamson, R.C. (1996). Efficient agnostic learning of neural networks with bounded fan–in. *IEEE Transactions on Information Theory*, *42*, 2118–2132.

Linton, O.B. (1997). Efficient estimation of additive nonparametric regression models. *Biometrika*, *84*, 469–474.

Linton, O.B., Härdle, W. (1996). Estimating additive regression models with known links. *Biometrika*, *83*, 529–540.

Linton, O.B., Nielsen, J.B. (1995). A kernel method of estimating structured nonparametric regression based on marginal integration. *Biometrika*, *82*, 93–100.

Lugosi, G., Zeger, K. (1995). Nonparametric estimation via empirical risk minimization. *IEEE Transactions on Information Theory*, *41*, 677–687.

McCulloch, W.S., Pitts, W. (1943). A logical calculus of ideas immanent in neural activity. *Bulletin of Mathematical Biophysics*, *5*, 115–133.
McGaffrey, D.F., Gallant, A.R. (1994). Convergence rates for single hidden layer feedforward networks. *Neural Networks*, *7*, 147–158.
Minsky, M.L., Papert, S. (1969). *Perceptrons: An introduction to computational geometry*. Cambridge: MIT Press.
Newey, W.K. (1994). Kernel estimation of partial means and a general variance estimator. *Econometric Theory*, *10*, 233–253.
Pollard, D. (1984). *Convergence of stochastic processes*. Berlin Heidelberg New York: Springer.
Ripley, B.D. (1996). *Pattern recognition and neural networks*. Cambridge: Cambridge University Press.
Rummelhart, D.E., McClelland, J.L. (1986). *Parallel distributed processing: explorations in microstructure of cognition. Vol. 1. Foundations*, Cambridge: MIT Press.
Stone, C.J. (1982). Optimal global rates of convergence for nonparametric regression. *Annals of Statistics*, *10*, 1040–1053.
Stone, C.J. (1985). Additive regression and other nonparametric models. *Annals of Statistics*, *13*, 689–705.
Stone, C.J. (1994). The use of polynomial splines and their tensor products in multivariate function estimation. *Annals of Statistics*, *22*, 118–184.
Wahba, G., Wang, Y., Gu, C., Klein, R., Klein, B. (1995). Smoothing spline anova for exponential families, with application to the wisconsin epidemiological study of diabetic retinopathy. *Annals of Statistics*, *23*, 1865–1895.
White, H. (1990). Connectionist nonparametric regression: multilayer feedforward networks can learn arbitrary mappings. *Neural Networks*, *3*, 535–549.