Computational Learning Theory:
Proceedings of the Fourth Annual Workshop
Morgan Kaufman, 1991

Approximation and Estimation Bounds for Artificial Neural Networks

A. R. BARRON
Departments of Statistics and ECE
University of Illinois
725 S. Wright Street
Champaign, IL 61820
a-barron@uiuc.edu

Abstract

For a class of artificial neural networks, the mean integrated squared error between the estimated network and the target function is shown to be bounded by

$$O(1/n) + O(nd/N) \log N,$$

for target functions satisfying a given smoothness property, where n is the number of nodes, d is the input dimension of the function, and N is the number of training observations. The two contributions to this total risk are approximation error and estimation error. Approximation error refers to the distance between the target function and the closest neural network function of a given architecture and estimation error refers to the distance between this network function and an estimated network function. With $n = (N/(d \log N))^{1/2}$ nodes, the order of the bound on the mean integrated squared error is optimized to be $O((d/N) \log N)^{1/2}$. The bound demonstrates surprisingly favorable properties of network estimation compared to traditional series and nonparametric curve estimation techniques in the case that d is moderately large. Similar bounds are obtained when the number of nodes n is not preset as a function of N, but rather it is optimized by the use of a complexity regularization or minimum description length criterion. The analysis involves Fourier techniques for the approximation error, metric entropy considerations for the estimation error, and a calculation of the index of resolvability of minimum complexity estimation of the family of networks.

1 INTRODUCTION

With artificial neural networks or other methods of parametric estimation of functions, it is desirable to balance the objectives of small approximation error and small estimation error. The approximation error between the target function and the closest neural network function of a given network family can be made as small as desired by increasing the number of nodes, see [1,2,3]. However, a large number of nodes makes it more difficult to accurately estimate the parameters of this network for moderate sample sizes. In this paper we address the combined effect of the approximation and estimation error on the overall accuracy of a network as an estimate of the target function. The target function is not assumed to be known or even known to be a member of a finite dimensional family. Rather it is only assumed to satisfy a certain smoothness property expressed through the Fourier transform.

The theory of learning applied to neural networks, as in [4], has focussed on the estimation error component of the problem: that is, the difference in risks between an estimated network and the best network. The same may be said of much of the parametric statistical theory, as in [5], that could also be applied to artificial neural networks of a given architecture. In contrast, the nonparametric statistical theory of curve estimation and classification, which has been under development for the last 35 years, has shown that one can effectively deal with the total risk of estimation of functions (at least for functions of moderately small dimension), for target functions restricted only by general smoothness properties (see [6,7]).

In recent years, theory has been developed in which a parametric family is not restricted to a given size, but rather the dimension of the family is increased at a certain rate as a function of the sample size, so as to get the smallest possible total risk, uniformly over classes of smooth functions, see [8,9]. A surprising aspect of this work is that the same rates of convergence of the total risk that are achievable by nonparametric estimators can be achieved by sequences of parametric families. It is also possible in this context to allow the dimension of the family to grow, not as a deterministic function of the sample size, but rather as determined from data so as to optimize a model selection criterion, see [10–15]. We mention in particular [14,15], where a theory is developed that is applicable to classes of

artificial neural networks. Bounds on the total risk of network estimators are given there in terms of an index of resolvability. This index of resolvability expresses the bounds on the risk in terms of the approximation error, the complexity of the networks, and the sample size (see Theorem 2 below). However, at the time there were not yet available bounds on the approximation error that could be used to complete the application of that theory to artificial neural networks.

Very recently, a bound on the approximation error for feedforward networks with one layer of sigmoidal nodes has been developed. It is shown in [3] that for functions in a fairly general smoothness class, the integrated squared error of approximation is bounded by $O(1/n)$ where n is the number of nodes (see Theorem 1 below). Armed with this result we are here able to derived bounds on the total risk of network estimators. The mean squared error between the estimated network and the target function is shown to be bounded by $O(1/n) + O(nd/N)\log N$ where d is the dimension of the input, N is the sample size, and n is the number of nodes.

2 TECHNICAL SUMMARY

Functions $f(x)$ on $\mathbf{R}^d$ are approximated using feedforward neural network models with one layer of sigmoidal nonlinearities, see [1,2,3]. These networks implement functions of the form

$$f_n(x) = f_n(x,\theta) = \sum_{k=1}^{n} c_k \phi(a_k^T x + b_k) + c_0 \qquad (1)$$
$$= \sum_{k=1}^{n} c_k \phi(s_k(\alpha_k^T + \beta_k)) + c_0,$$

which is parameterized by the vector θ, consisting of $a_k \in \mathbf{R}^d$, b_k, $c_k \in \mathbf{R}$, for $k = 1, 2, \ldots, n$, and $c_0 \in \mathbf{R}$. Here $s_k = |a_k|$, $\alpha_k = a_k/s_k$, and $\beta_k = b_k/s_k$ are the scale, direction, and location parameters, respectively, of the kth node. The function $\phi(z)$ is assumed to be a given sigmoidal function, that is, it is a bounded function on the real line satisfying $\phi(z) \to 1$ as $z \to \infty$ and $\phi(z) \to 0$ as $z \to -\infty$.

Given a positive constant c, let Γ_c be the class of real-valued functions $f(x)$ on $\mathbf{R}^d$ represented in terms of a Fourier transform $\tilde{f}(\omega)$ satisfying

$$\int_{\mathbf{R}^d} |\omega||\tilde{f}(\omega)|d\omega \le c. \qquad (2)$$

We measure the accuracy of an approximation $f_n(x)$ to the target function $f(x)$ in terms of the $L_2(\mu, B_r)$ norm

$$\|f - f_n\|^2 = \int_{B_r} |f(x) - f_n(x)|^2 \mu(dx) \qquad (3)$$

for an arbitrary probability measure μ with support in $B_r = \{x \in \mathbf{R}^d : |x| \le r\}$ for some given $r > 0$. In the case of a neural network function $\hat{f}_n$ estimated from data, the norm $\|f - \hat{f}_n\|$ measures the ability of the network function to generalize to new data drawn with distribution μ. In contrast the empirical risk $(1/N)\sum_{i=1}^{N}(f(X_i) - f_n(X_i))^2$ only measures the accuracy at the observed data points X_i, $i = 1, 2, \ldots, N$.

We shall make use of the following special case of a recent result in [3].

Theorem 1: *Given an arbitrary sigmoidal function ϕ, and probability measure μ on B_r, and $r > 0$, then for every f in Γ_c and every $n \ge 1$, there exists an artificial neural network of the form (1) such that*

$$\|f - f_n\|^2 \le \frac{c'}{n}, \qquad (4)$$

where $c' = (2rc)^2$. The parameters in (1) may be restricted to satisfy $\sum_{k=1}^{n} |c_k| \le 2rc$, $c_0 = f(0)$, and $|\beta_k| \le r$. If also we impose the restriction that the scale parameter $|a_k|$ is not larger that some positive value τ_n, and if the sigmoidal function ϕ is nondecreasing and satisfies the symmetry property $\phi(-z) = 1 - \phi(z)$, then there exists an artificial neural network of the form (1) with

$$\|f - f_n\| \le \frac{2rc}{\sqrt{n}} + 4rc\delta_{r\tau_n} \qquad (5)$$

where

$$\delta_\tau = \inf_{\varepsilon > 0}(\varepsilon + 2\phi(-\tau\varepsilon)). \qquad (6)$$

In particular, if $\phi(z) = 1/(1 + e^{-z})$ is the logistic sigmoid, if $\tau_n \ge \sqrt{n}\ln n$, and if $f \in \Gamma_c$, then

$$\delta_{r\tau_n} \le O\left(\frac{1}{\sqrt{n}}\right), \qquad (7)$$

so there is an artificial neural network of the form (1) with $|a_k| \le \tau_n$ and

$$\|f - f_n\|^2 \le O\left(\frac{1}{n}\right). \qquad (8)$$

Now suppose that (X_i, Y_i), $i = 1, 2, \ldots, N$ are independently drawn from a distribution $P_{X,Y}$ with conditional mean $f(x) = E(Y \mid X = x)$ and marginal distribution $P_X = \mu$. We assume that the support of

Y is in a known interval I with length bounded some $b > 0$. (It is also possible to develop a theory for some distributions for Y that have unbounded support, such as the case that the conditional distribution of Y given x is normal with mean $f(x)$ and variance σ^2). Here are two important cases:

(a) The function f is observed without error at randomly selected sites, that is, $Y_i = f(X_i)$, for $i = 1, 2, \ldots, N$, and the range of f is in a known interval of length bounded by b.

(b) The response $Y_i \in \{0, 1\}$ is a class label for a binary classification problem with overlapping class boundaries and $f(x) = P\{Y = 1 \mid X = x\}$ is the optimal discriminant function based on X. In this case $b = 1$ and $I = [0, 1]$.

Because the function $f(x)$ is known to have range in a given interval $I = [i_1, i_2]$, we improve the fit by replacing $f_n(x, \theta)$ with

$$\bar{f}_n(x, \theta) = \text{clip}(f_n(x, \theta)),$$

where clip$(y) = y$ for $y \in I$, $= i_1$ for $y < i_1$, and $= i_2$ for $y > i_2$. Note that clip() is a sigmoidal function with range I. Thus the composition of clip() with functions of the form (1) forms a two layer sigmoidal network. By clipping the candidate functions to a range of a given length b, we satisfy one of the requirements for the application of a theorem from [15], see Theorem 2 below.

For each number of nodes n and sample size N, let Θ_n be a net of parameter vectors θ, and let $C_{n,N}(\theta)$ be nonnegative numbers satisfying, $C_{n,N}(\theta) \geq l$ for some constant $l > 0$, and

$$\sum_{\theta \in \Theta_n} e^{-C_{n,N}(\theta)} \leq 1. \tag{9}$$

The numbers $C_{n,N}(\theta) \ln 2$, rounded up to the nearest integer, may be interpreted as binary codelengths, then (9) becomes the Kraft-McMillan condition for the existence of uniquely decodable codes of these lengths. Another interpretation is that $e^{-C_{n,N}(\theta)}$ is a prior probability for $\theta \in \Theta_{n,\varepsilon}$. If $C_{n,N}(\theta) = C_{n,N}$ is constant and $\Theta_n = \Theta_{n,\varepsilon}$ is an ε-net of points such that every θ is within ε of a point in $\Theta_{n,\varepsilon}$, then (9) implies that $C_{n,N}$ is a bound on the Kolmogorov ε-entropy (or metric entropy) of the set of possible θ's. As in [13], we blend these complexity, prior probability, and metric entropy interpretations.

For a given n and N, we define the *index of resolvability*, as in [13,14,15], to be

$$R_{n,N}(f) = \min_{\theta \in \Theta_n} \left(\|f - \bar{f}_n(\cdot, \theta)\|^2 + \lambda \frac{C_{n,N}(\theta)}{N} \right), \tag{10}$$

where λ is a given positive constant. Equation (10) gives the resolvability for a neural network family with a given number of nodes n. Let $C(n)$ be numbers satisfying $\sum_{n=1}^{\infty} e^{-C(n)} \leq 1$. For the collection of networks indexed by $n = 1, 2, \ldots$, the index of resolvability is

$$R_N(f) = \min_{n \geq 1} \left(R_{n,N}(f) + \lambda \frac{C(n)}{N} \right). \tag{11}$$

It will be seen that we may restrict the minimization in (11) to $n \leq N/d$, without affecting our bounds on the resolvability. Indeed, it is advisable to restrict n such that the number of parameters is of smaller order than the sample size N. The minimization in (11) determines the n that yields the best resolvability.

The minimum complexity estimator (or complexity regularization estimator) of a neural network of a given size n is

$$\hat{f}_{n,N}(x) = \bar{f}_n(x, \hat{\theta}_{n,N}) \tag{12}$$

where

$$\hat{\theta}_{n,N} = \underset{\theta \in \Theta_n}{\text{argmin}} \left(\frac{1}{N} \sum_{i=1}^{N} (Y_i - \bar{f}_n(X_i, \theta))^2 + \lambda \frac{C_{n,N}(\theta)}{N} \right). \tag{13}$$

Thus $\hat{f}_{n,N}$ is the least squares estimator with a complexity penalty. The minimum complexity estimator, with both n and θ estimated, is

$$\hat{f}_N = \hat{f}_{\hat{n},N} \tag{14}$$

where

$$\hat{n} = \underset{n}{\text{argmin}} \left(\frac{1}{N} \sum_{i=1}^{N} (Y_i - \hat{f}_{n,N}(X_i))^2 + \lambda \frac{C_{n,N}(\hat{\theta}_{n,N})}{N} + \lambda \frac{C(n)}{N} \right). \tag{15}$$

Complexity regularization (as defined in (13) and (15)) is closely related to Vapnik's method of structural risk minimization [10] and Rissanen's minimum description-length criterion [11].

We make use of a theorem from [14,15], specialized here to neural network estimators.

Theorem 2: *Let a neural network be estimated by least squares with a complexity penalty as in (13), (15), where the range of Y and each candidate function is restricted to a known interval of length b, then for $\lambda > 5b^2/3$, for all $n \geq 1$, and all $N \geq 1$,*

$$E\|f - \hat{f}_{n,N}\|^2 \leq \gamma R_{n,N}(f) + \frac{2\gamma\lambda}{N}, \tag{16}$$

and

$$E\|f - \hat{f}_N\|^2 \le \gamma R_N(f) + \frac{2\gamma\lambda}{N}, \tag{17}$$

where $\gamma = (3\lambda + b^2)/(3\lambda - 5b^2)$. *Thus*

$$E\|f - \hat{f}_N\|^2 \le O(R_N(f)). \tag{18}$$

The index of resolvability automatically captures the effect of the approximation error $\|f - f_n\|^2$ and the estimation error, $E\|f_n - \hat{f}_{n,N}\|^2$.

Theorem 2, as proved in [15], is based on the idea that by controlling these sources of error, it is possible to obtain bounds on the total mean squared error. Indeed, by the triangle inequality,

$$\|f - \hat{f}_{n,N}\| \le \|f - f_n\| + \|f_n - \hat{f}_{n,N}\|. \tag{19}$$

Armed with the recent results summarized in Theorems 1 and 2, we are now in position to obtain a specific rate of convergence for the mean squared error and the index of resolvability for functions in the smoothness class Γ_c. This leads to Theorems 3 and 4 below, which are the main results of this paper.

The following assumptions are made of the sigmoidal function $\phi(z)$ for Theorems 3 and 4. We assume that it is nondecreasing, satisfies the symmetry property $\phi(-z) = 1 - \phi(z)$, and has a bounded derivative $\phi'(z)$ on R^d. Suppose also that ϕ has a tail which converges to its limits at least polynomially fast, that is, $\limsup \phi(z)/|z|^p < \infty$ as $z \to -\infty$, for some $p > 0$.

We restrict attention to scale parameters of the sigmoids that are bounded by τ_n, where τ_n is chosen such that $\delta_{r\tau_n} = O(1/\sqrt{n})$. In particular, if $\phi(z)$ equals zero and one outside of a finite interval then we may set $\tau_n = \sqrt{n}$, if $\phi(z)$ approaches its limits exponentially fast we may set $\tau_n = \sqrt{n}\ln n$, and if $\phi(z)$ approaches its limits polynomially fast we may set $\tau_n = n^{(p+1)/2p}$.

The sigmoidal networks are now parameterized as

$$f_n(x,\theta) = \sum_{k=1}^{n} c_k \phi(\tau_n(\alpha_k^T x + \beta_k)) + c_0,$$

and

$$\bar{f}_n(x,\theta) = \mathrm{clip}(f_n(x,\theta)), \tag{20}$$

where the vector θ consists of $\alpha_k \in \mathbf{R}^d$, β_k, $c_k \in \mathbf{R}$, for $k = 1, 2, \ldots, n$, and c_0. Given $n \ge 1$, the parameter space $\Theta_{n,r,c} \subset \mathbf{R}^{n(d+2)+1}$ is taken to be the set of all such θ for which $|\alpha_k| \le 1$, $|\beta_k| \le r$, $\sum_k |c_k| \le 2rc$, and $c_0 \in I$, where I is the given range of the target function.

Next we control the precisions with which the coordinates of the parameter vectors are allowed to be represented. Let $\Theta_n = \Theta_{n,\varepsilon}$ be a net of parameter points in $\mathbf{R}^{n(d+2)+1}$ that ε-covers $\Theta_{n,r,c}$ in the sense that for every θ in $\Theta_{n,r,c}$ there is a θ^* in $\Theta_{n,\varepsilon}$ such that

$$|\alpha_k - \alpha_k^*| \le \sqrt{d}\,\varepsilon_1,$$

$$|\beta_k - \beta_k^*| \le \varepsilon_1, \tag{21}$$

for $k = 1, 2, \ldots, n$, and

$$|c_k - c_k^*| \le \varepsilon_2, \tag{22}$$

for $k = 0, 1, 2, \ldots, n$. In particular, we may take $\Theta_{n,\varepsilon}$ to be a rectangular grid on $\mathbf{R}^{n(d+2)+1}$ spaced at width ε_1 for the α_k and β_k coordinates and spaced at width ε_2 for the c_k coordinates.

Note that we use different precisions for the parameters at different layers of the sigmoidal network. This is because large slopes in the sigmoidal function (scaled by τ_n) lead to requirements of greater precision for the α_k and β_k parameters than for the c_k parameters in our analysis.

As we scale the problem, by allowing arbitrarily large $n \ge 1$, $N \ge 2$, and $d \ge 1$, subject to $n \le N$, the range of permitted precisions ε_1 and ε_2 are constrained as follows: The upper bounds are

$$\varepsilon_1 \le O\left(\left(\frac{1}{\tau_n}\right)\left(\frac{n}{N}\right)^{1/2}(\log N)^{1/2}\right)$$

$$\varepsilon_2 \le O\left(\frac{\log N}{nN}\right)^{1/2} \tag{23}$$

and

$$\varepsilon_2 \le O\left(\frac{1}{n}\right);$$

the lower bounds are that ε_1 and ε_2 be not smaller than the reciprocal of a polynomial in N,

$$\varepsilon_1 \ge \left(\frac{1}{N}\right)^p \tag{24}$$

and

$$\varepsilon_2 \ge \left(\frac{1}{N}\right)^p$$

for some $p \ge 1$.

The constraints in (23) are imposed so that in the analysis below accurate network approximations, which are

know (by Theorem 1) to exist for continuous-valued parameters, can be shown to be realized by discretized parameter values. The constraints in (24) are imposed to prevent excessive complexity penalties that would result from too fine a precision.

It will be seen in the proof that the choice of ε_1 and ε_2 of the following form optimizes the order of the bounds on the resolvability:

$$\varepsilon_1 = \varepsilon_{1,n,N} = C_1 \left(\frac{1}{\tau_n} \left(\frac{n}{N} \right)^{1/2} \right) \tag{25}$$

and

$$\varepsilon_2 = \varepsilon_{2,n,N} = C_2 \left(\frac{1}{nN} \right)^{1/2}, \tag{26}$$

where C_1 and C_2 are positive constants.

First we state the main result in the case of a constant complexity penalty, $C_{n,N}(\theta) = C_{n,N}$, equal to the logarithm of the number of points in a net $\Theta_{n,\varepsilon}$ that ε-covers $\Theta_{n,r,c}$. Thus $C_{n,N}$ may be interpreted as a bound on the metric entropy.

One simple choice for $\Theta_{n,\varepsilon}$ is a rectangular grid that covers the set $\{\theta : -1 \le \alpha_{n,j} \le 1, -r \le \beta_k \le r, -2rc \le c_k \le 2rc, c_0 \in I, k = 1, 2, \ldots, n, j = 1, 2, \ldots, d\}$ that is somewhat larger than $\Theta_{n,r,c}$. In this case, the number of points in $\Theta_{n,\varepsilon}$ is $(2/\varepsilon_1)^{nd}(2r/\varepsilon_1)^n(4rc/\varepsilon_2)^n(b/\varepsilon_2)$. Using (24) we have that the logarithm of the number of points in $\Theta_{n,\varepsilon}$ satisfies $C_{n,N} = O(nd \log N)$.

Note that for a constant complexity penalty, the estimator that acheives the minimum in (13) is the same as the least squares estimator restricted to $\Theta_{n,\varepsilon}$.

Theorem 3: *Let $\hat{f}_{n,N}$ be a neural network of the form (20) chosen to achieve the minimum sum of squared errors, subject to the constraint that $\theta \in \Theta_{n,\varepsilon}$ (where $\Theta_{n,\varepsilon}$ covers $\Theta_{n,r,c}$ in the sense of (21) and (22), ε_1 and ε_2 satisfy (23), and the logarithm of the cardinality of $\Theta_{n,\varepsilon}$ satisfies $C_{n,N} = O(nd \log N)$). If the target function f is in Γ_c, then*

$$E\|f - \hat{f}_{n,N}\|^2 \le O\left(\frac{1}{n}\right) + O\left(\frac{nd}{N} \log N\right), \tag{27}$$

for all d, n, and N with $n \le N$. In particular, if $n = O(N/(d \log N))^{1/2}$, so as to optimize the order of the bound in (21), then

$$E\|f - \hat{f}_{n,N}\|^2 \le O\left(\frac{d}{N} \log N\right)^{1/2}. \tag{28}$$

Here the expressions of the form $O(nd/N) \log N$ and $O(1/n)$, refer to quantities bounded by a constant times $(nd/N) \log N$ and a constant times $(1/n)$, respectively, where the constants are independent of d, n, and N, subject to the contraints that $n \le N$, $d \ge 1$, $n \ge 1$, and $N \ge 2$.

Next we relax the requirement that the net of parameter points is restricted to a compact set. For the estimation of functions in Γ_c, $\Theta_{n,\varepsilon}$ need only satisfy (21) and (22) for $\theta \in \Theta_{n,r,c}$. However, since c may not be known in advance, it is desirable that the net satisfy the indicated properties for all positive c. It is more than enough that (21) and (22) be satisfied for every $\theta \in \mathbf{R}^{n(d+2)+1}$. In particular, we may take $\Theta_{n,\varepsilon}$ to be a rectangular grid on all of $\mathbf{R}^{n(d+2)+1}$ spaced at width ε_1 for the coordinates of the α_k's and β_k's, and width ε_2 for the c_k's.

We now allow for complexity penalties $C_{n,N}(\theta)$ that depend on θ, subject to the summability requirement (9), and subject to the requirement that

$$C_{n,N}(\theta) = O(nd \log N) \tag{29}$$

uniformly for θ in $\Theta_{n,r,c} \cap \Theta_{n,\varepsilon}$, for every $c > 0$.

The complexity term $C_{n,N}(\theta)$ may be based on the total number of bits in the binary representation of the coordinates of θ, to the prescribed accuracies. For coordinates that take values outside of $[0, 1)$, the contribution to the complexity will depend on the logarithm of the magnitude of the integer part.

Other choices for $C_{n,N}(\theta)$ may be used that relax the requirement of prior constraints on the magnitudes of the parameters, while satisfying the conditions (9) and (29). One way to do this in the case of a rectangular grid is to use a continuous and positive prior probability density function $p(\theta)$ on $\mathbf{R}^{n(d+2)+1}$. We set $C_{n,N}(\theta)$ to be equal to minus the logarithm of the prior probability of the rectangle in the grid that includes the point θ. The prior probability of these small rectangles is approximated by the volume of the rectangle, which is $(\varepsilon_1)^{n(d+1)}(\varepsilon_2)^{n+1}$, times the prior density at θ. On compact subsets of $\mathbf{R}^{n(d+2)+1}$ this approximation is uniformly accurate, which permits verification of the requirement (29). In particular, we have in this case that

$$C_{n,N}(\theta) = n(d+1) \log \frac{1}{\varepsilon_1} + (n+1) \log \frac{1}{\varepsilon_2} + \log \frac{1}{p(\theta)} + O(1). \tag{30}$$

For reasonable choices of the prior density $p(\theta)$ it can be verified that $|\log 1/p(\theta)| = O(nd)$ (that is, it is of order not larger than the number of parameters), uniformly on $\Theta_{n,r,c}$ for each $c > 0$. In such cases, the $\log 1/p(\theta)$ term is of smaller order than the first two

terms on the right side of (30). Then (24) implies that $C_{n,N}(\theta) = O(nd\log N)$ uniformly in $\Theta_{n,r,c}$ for each $c > 0$ as required by (29).

Theorem 4: *Let a neural network of the form (20) be estimated by least squares with a complexity penalty as in (13), (15), with $\lambda > 5b^2/3$, and $\Theta_{n,\varepsilon}$ and $C_{n,N}(\theta)$ satisfying (21-22), (23), and (29). If the target function f is in Γ_c then, subject to the constraint that $n \le N$,*

$$E\|f - \hat{f}_{n,N}\|^2 \le O(R_{n,N}(f)) \le O\left(\frac{1}{n}\right) + O\left(\frac{nd}{N}\log N\right), \tag{31}$$

and if also $C(n) \le O(n)$, then

$$E\|f - \hat{f}_N\|^2 \le O(R_N(f)) \le O\left(\frac{d}{N}\log N\right)^{1/2}. \tag{32}$$

3 PROOFS

As stated above Theorems 1 and 2 are recently proven in [3] and [15], respectively. The conclusions of these results serve as the tools in proving the main results of this paper which are Theorems 3 and 4. Here we give the proof of Theorem 4. Theorem 3 follows as a special case by taking $C_{n,N}(\theta)$ to be a constant equal to the logarithm of the number of points in $\Theta_{n,\varepsilon}$.

In accordance with the requirements imposed on the sigmoidal function, suppose $\phi(z)$ and its derivative are bounded by $|\phi(z)| \le v$ and $|\phi'(z)| \le v$ for some $v \ge 1$. Given n, let θ be a parameter vector in the set $\Theta_{n,r,c} = \{\theta : |\alpha_k| \le 1, |\beta_k| \le r, \sum_{k=1}^n |c_k| \le 2rc, \text{ and } c_0 \in I\}$, and let θ^* in $\Theta_{n,\varepsilon}$ be chosen to satisfy (21) and (22). By first order Taylor expansion, for any $x \in R^d$,

$$\begin{aligned} f_n(x,\theta) - f_n(x,\theta^*) &= (\theta - \theta^*)^T \nabla f_n(x,\theta) \\ &= \tau_n \sum_{k=1}^n \tilde{c}_k (\alpha_k - \alpha_k^*)^T x \phi'(\tilde{z}_k) \\ &\quad + \tau_n \sum_{k=1}^n \tilde{c}_k (\beta_k - \beta_k^*) \phi'(\tilde{z}_k) \\ &\quad + \sum_{k=1}^n (c_k - c_k^*)\phi(\tilde{z}_k) \\ &\quad + (c_0 - c_0^*), \end{aligned} \tag{33}$$

where $\tilde{\theta} = \theta^* + t(\theta - \theta^*)$, with $0 \le t \le 1$, is some parameter point in between θ and θ^*, and $\tilde{z}_k = \tau_n(\tilde{\alpha}_k^T x + \tilde{\beta}_k)$.

This leads to the following bound, which holds uniformly for $x \in B_r$.

$$\begin{aligned} &|f_n(x,\theta) - f_n(x,\theta^*)| \\ &\quad \le (d^{1/2}+1) r v \tau_n \varepsilon_1 \sum_{k=1}^n |\tilde{c}_k| + (vn+1)\varepsilon_2 \\ &\quad \le (d^{1/2}+1) r v \tau_n \varepsilon_1 (2rc + n\varepsilon_2) + (vn+1)\varepsilon_2, \end{aligned} \tag{34}$$

where we have used the fact that $\sum_k |\tilde{c}_k| \le \sum_k |c_k| + \sum_k |c_k - \tilde{c}_k|$, which is not greater than $2rc + n\varepsilon_2$. It follows then that

$$\|f_n(\cdot,\theta) - f_n(\cdot,\theta^*)\| \le O(d^{1/2}\tau_n\varepsilon_1) + O(n\varepsilon_2)^2. \tag{35}$$

Now by Theorem 1, there is a θ in $\Theta_{n,r,c}$ such that $\|f - f_n(\cdot,\theta)\| = O(1/n)^{1/2}$. So applying the triangle inequality, there is a θ^* in $\Theta_{n,\varepsilon}$ such that

$$\|f - f_n(\cdot,\theta^*)\|^2 \le O\left(\frac{1}{n}\right) + O(d\tau_n^2\varepsilon_1^2) + O(n\varepsilon_2)^2. \tag{36}$$

Therefore, by (29), (35), (36), and the definition of the index of resolvability, we have

$$\begin{aligned} R_{n,N}(f) &= \min_{\theta^* \in \Theta_{n,\varepsilon}} \left(\|f - \bar{f}_n(\cdot,\theta^*)\|^2 + \lambda \frac{C_{n,N}(\theta^*)}{N} \right) \\ &\le \min_{\theta^* \in \Theta_{n,\varepsilon}} \left(\|f - f_n(\cdot,\theta^*)\|^2 + \lambda \frac{C_{n,N}(\theta^*)}{N} \right) \\ &\le O\left(\frac{1}{n}\right) + O(d\tau_n^2\varepsilon_1^2) + O(n\varepsilon_2)^2 \\ &\quad + O\left(\frac{nd\log N}{N}\right). \end{aligned} \tag{37}$$

In particular for $C_{n,N}(\theta)$ of the form (30) the bound becomes

$$\begin{aligned} R_{n,N}(f) &\le O\left(\frac{1}{n}\right) + O(d\tau_n^2\varepsilon_1^2) + O(n\varepsilon_2)^2 \\ &\quad + \frac{n(d+1)}{N}\log\frac{1}{\varepsilon_1} + \frac{n+1}{N}\log\frac{1}{\varepsilon_2} \\ &\quad + O\left(\frac{nd}{N}\right). \end{aligned} \tag{38}$$

Optimizing this bound yields $\varepsilon_1 = C_1((1/\tau_n)(n/N)^{1/2})$ and $\varepsilon_2 = C_2(1/(nN)^{1/2})$ as in (25), (26). The corresponding complexity is

$$\begin{aligned} C_{n,N}(\theta) &= O(nd\log\tau_n(N/n)^{1/2}) + O(n\log(nN)^{1/2}) \\ &= O(nd\log N), \end{aligned} \tag{39}$$

for $n \le N$ and $\tau_n \le O(n^{(p+1)/2p})$.

From (23) and (37) we have that the index of resolvability is bounded as follows

$$R_{n,N}(f) \leq O\left(\frac{1}{n}\right) + O\left(\frac{nd}{N}\log N\right). \tag{40}$$

The choice $n = O(N/(d\log N))^{1/2}$ optimizes the order of the bound in (40), yielding

$$R_{n,N}(f) \leq O((d/N)\log N)^{1/2}.$$

It follows then from (11) and $C(n) \leq O(n)$, that

$$\begin{aligned} R_N(f) &= \min_n \left(R_{n,N}(f) + \lambda\frac{C(n)}{N}\right) \\ &\leq O\left(\frac{d}{N}\log N\right)^{1/2}. \end{aligned} \tag{41}$$

Applying Theorem 2 shows that the index of resolvability bounds the mean squared error of the neural network estimator. This completes the proof of Theorem 4.

Acknowledgement

This work was supported by ONR Contract N00014-89-J-1811.

References

[1] Cybenko, G. (1989) "Approximations by superpositions of a sigmoidal function." *Math. Control, Signals, Systems*, vol. 2, p. 303-314.

[2] Hornik, K., Stinchcombe, M., and White, H. (1988) "Multi-layer feedforward networks are universal approximators." Department of Economics Technical Report, University of California, San Diego. (To appear in *Neural Networks*).

[3] Barron, A. R. (1991) "Universal approximation bounds for superpositions of a sigmoidal function." Submitted to the *IEEE Transactions on Information Theory.*

[4] Haussler, D. (1991) "Decision theoretic generalizations of the PAC model for neural net and other learning applications." Computer Research Laboratory Technical Report UCSC-CRL-91-02. University of California, Santa Cruz.

[5] Seber, G. A. F. and Wild, C. M. (1989) *Nonlinear Regression*, Wiley, New York.

[6] Silverman, B. W. (1986) *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, New York.

[7] Eubank, R. (1988) *Spline Smoothing and Nonparametric Regression*, Marcel Dekker, New York.

[8] Cox, D. D. (1988) "Approximation of least squares regression on nested subspaces," *Annals of Statistics*, vol. 16, pp. 713–732.

[9] Barron, A. R. (1991) "Approximation of density functions by sequences of exponential families," *Annals of Statistics*, vol. 19, no. 3.

[10] Vapnik, V. (1982) *Estimation of Dependences Based on Empirical Data*, Springer-Verlag, New York.

[11] Rissanen, J. (1983) "A universal prior for integers and estimation by minimum description length." *Annals of Statistics.* vol. 11, pp. 416–431.

[12] Li, K. C. (1987) "Asymptotic optimality for C_p, C_L, cross-validation, and generalized cross-validation: discrete index set." *Annals of Statistics*, vol. 15, pp. 958–975.

[13] Barron, A. R. and Cover, T. M. (1991) "Minimum complexity density estimation." *IEEE Transactions on Information Theory*, vol. 37, no. 4.

[14] Barron, A. R. (1989) "Statistical properties of artificial neural networks. *Proceeding of the IEEE International Conference on Decision and Control*, Tampa, Florida, Dec. 13–15, p. 280–285.

[15] Barron, A. R. (1990) "Complexity regularization with applications to artificial neural networks." *Proceedings of the NATO ASI on Nonparametric Functional Estimation*, Spetses, Greece, Aug. 1–10, G. Roussas, editor, Kluwer Academic Publishers, Dordrecht, The Netherlands.