Machine Learning, 14, 115–133 (1994)

Approximation and Estimation Bounds for Artificial Neural Networks

ANDREW R. BARRON BARRON@BRANDY.STAT.YALE.EDU
Department of Statistics, Yale University, P.O. Box 208290, New Haven, CT 06520

Editors: Ming Li and Leslie Valiant

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

$$O\left(\frac{C_f^2}{n}\right) + O\left(\frac{nd}{N}\log N\right),$$

where n is the number of nodes, d is the input dimension of the function, N is the number of training observations, and C_f is the first absolute moment of the Fourier magnitude distribution of f. The two contributions to this total risk are the approximation error and the 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 ideal network function and an estimated network function. With $n \sim C_f(N/(d\log N))^{1/2}$ nodes, the order of the bound on the mean integrated squared error is optimized to be $O(C_f((d/N)\log N)^{1/2})$. 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 preselected as a function of C_f (which is generally not known *a priori*), but rather the number of nodes is optimized from the observed data 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.

Keywords: Neural nets, approximation theory, estimation theory, complexity regularization, statistical risk

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, for example, Cybenko, 1989, Hornik, et. al., 1989, and Barron, 1993). However, a large number of nodes makes it more difficult to estimate accurately 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. Previously, White (1990) showed that the overall statistical risk of an estimated neural network converges to zero as the sample size and number of nodes increases to infinity. Here we develop bounds that demonstrate the rate of convergence.

The theory of learning applied to neural networks, as in (Haussler, 1992), has focussed on the estimation error component of the problem: that is, the difference in risks between an estimated network and the best network of a given size and architecture. The same may be said of much of the parametric statistical theory (as, for example, in Seber and Wild, 1989) that could also be applied to artificial neural networks. In contrast, the nonparametric statistical theory of curve estimation and classification (which has seen extensive development for the last 35 years), has shown that one can deal effectively with the total risk of the estimation of functions, including both the approximation error (bias) and the estimation error (variance), at least for functions of moderately small dimension, for target functions restricted only by general smoothness properties (see, for example, Silverman (1986), Eubank (1988), Hardle (1990)).

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 Cox, 1988, Stone, 1990, Barron and Sheu, 1991). 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, for instance, Vapnik, 1982, Rissanen, 1983, Li, 1987, Barron and Cover, 1991). In this paper we build on past work of the author (Barron, 1989, 1990) where a theory of model selection is developed that is applicable to artificial neural networks and other nonlinear models. 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 Barron (1993) that the integrated squared error of approximation is bounded by $O(C_f^2/n)$, where n is the number of nodes and C_f quantifies the regularity of the function via an integral involving the Fourier transform, that is, $C_f = \int |\omega||\tilde{f}(\omega)|\, d\omega$ (see Theorem 1 below). Armed with this result we are here able to evaluate the index of resolvability and thereby to derive bounds on the total risk of network estimators. The mean squared error between the estimated network and the target function is shown (in

Theorems 3 and 4) to be bounded by $O(C_f^2/n) + O(nd/N)\log N$, where d is the dimension of the input, N is the sample size, and n is the number of nodes. Moreover, when the number of nodes is optimized (either by setting $n \sim C_f(N/(d\log N))^{1/2}$ or when C_f is unknown by using the complexity regularization criterion to select n) the mean squared error is shown to be bounded by $O(C_f((d/N)\log N)^{1/2})$.

A surprising aspect of this result is that, in terms of the first order behavior of the risk, the rate of convergence as a function of the sample size N is of order $(1/N)^{1/2}$ (times a logarithmic factor), where the exponent $1/2$ is independent of the dimension d. In contrast the minimax rate of convergence of the mean integrated squared error for functions in traditional smoothness classes (i.e., functions with bounded norms of the derivatives of order s for some $s > 0$) is of order $(1/N)^{2s/(2s+d)}$ as shown in the work of Ibragimov and Hasminskii (1980), Pinsker (1980), Stone (1982), and Nussbaum (1986). Characteristic of the traditional smoothness classes is the fact that linear combinations of fixed basis functions (e.g., sinusoids, polynomials, and splines) are asymptotically minimax. Even sigmoidal basis functions with fixed internal weights are nearly minimax as shown by McCaffrey and Gallant (1991). However, the use of fixed basis functions prevents the opportunity to provide more accurate estimators for interesting subclasses of functions. (Exceptions to the minimax optimality of linear estimators hold for certain non-Hilbertian classes in Nemirovskii (1985) and Nemirovskii, Polyak, and Tsybakov (1985).) The dependence of the rate on the dimension d in the denominator of the exponent is a curse of dimensionality that does not apply to the class of functions examined here.

Although the rate $(1/N)^{1/2}$ as a function of N is independent of the dimension d, it is possible for the constant C_f to be exponentially large in d for sequences of functions f of increasing dimensionality. Indeed, many radial functions have C_f exponentially large, so do many tensor products, that is, functions of the form $f(x) = g_1(x_1)g_2(x_2)\cdot\ldots\cdot g_d(x_d)$. (More elaborate networks than the single hidden layer networks considered here may be able to provide accurate estimates in some of these cases.) Nevertheless, there are a number of interesting examples for which C_f exhibits only moderate growth with the dimension, such as order $O(d)$, including positive definite functions that are continuously differentiable at the origin, and translation mixtures of such functions, see Barron (1993). Various closure properties of the class of functions are given there for sums, products, and certain compositions of functions with polynomially bounded spectral norms.

Key to the advantageous approximation and estimation properties of artificial neural networks is the fact that the model is not linear in all its parameters (activation weights). The adjustment of the scale, direction, and location parameters of the sigmoidal basis functions permits them to be adapted to the estimation of the target function. Nonlinear adjustment of sinusoidal, polynomial, spline, and wavelet basis functions is also possible, and it is anticipated that similar approximation and estimation bounds can be obtained in each of these cases by the same technique as used here for sigmoidal basis functions. (Indeed, approximation properties of nonlinearly adjusted sinusoidal expansions is at the heart of the analysis in Jones (1992) and Barron (1993) for projection pursuit and neural network ap-

proximation, respectively.) If attention were restricted to approximation by linear combinations of a fixed set of n basis functions, then by a result in (Barron 1993) there is no such basis for which the integrated squared approximation error is less than order $(C/d)(1/n)^{2/d}$ uniformly for all functions with $C_f \leq C$ for any $C > 0$. Consequently, it is seen that for the class of functions considered here, (adaptive) neural network estimation has approximation and estimation properties that are superior to traditional linear expansions for each dimension $d \geq 3$.

2. Technical Summary

Functions $f(x)$ with bounded domain in $\mathbf{R^d}$ are approximated using feedforward neural network models with one layer of sigmoidal nonlinearities. These networks implement functions of the form

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

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 $c_0 \in \mathbf{R}$, where $n \geq 1$ is the number of nonlinear terms (also called nodes or hidden units). 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 $\phi(z) \to -1$ as $z \to -\infty$. This property of a sigmoidal function implies that for large τ the scaled sigmoidal function $\phi_\tau(z) = \phi(\tau z)$ is close to the signum function $\mathrm{sgn}(z)$, which equals $+1$ for z positive and -1 for z negative.

For functions $f(x)$ on $\mathbf{R}^d$ with a Fourier representation of the form $f(x) = \int_{\mathbf{R}^d} e^{i\omega^T x} \tilde{f}(\omega) d\omega$, let

$$C_f = \int |\omega|_1 |\tilde{f}(\omega)| d\omega, \tag{2}$$

where $|\omega|_1 = \sum_{j=1}^d |\omega_j|$ is the ℓ_1 norm of ω in $\mathbf{R}^d$. More generally, if $f(x) = \int_{\mathbf{R}^d} e^{i\omega^T x} \tilde{F}(d\omega)$, for some complex-valued Fourier distribution $\tilde{F}$, we define $C_f = \int |\omega|_1 F(d\omega)$ where $F = |\tilde{F}|$ is the Fourier magnitude distribution of f. For approximation on a bounded set B, it is required only that the representation $f(x) = \int_{\mathbf{R}^d} e^{i\omega^T x} \tilde{F}(d\omega)$ holds for x in B for some $\tilde{F}$ with $\int |\omega|_1 |\tilde{F}(d\omega)|$ finite.

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

$$\|f - f_n\|^2 = \int_B |f(x) - f_n(x)|^2 \mu(dx), \tag{3}$$

for an arbitrary probability measure μ with support B assumed to be contained in the cube $[-1, 1]^d$. (Other bounded domains may be rescaled to be contained in this cube; see Barron (1993) for the form of the approximation bound for arbitrary bounded domains.) For approximation of a Boolean function the measure is

restricted to the set $\{0, 1\}^d$. For simplicity it is assumed that the vector $x = 0$ is included in the domain of f.

In the case that $\hat{f}_n$ is a neural network function estimated from data, the norm $\|f - \hat{f}_n\|$ measures the ability of the neural net 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, $i = 1, 2, \ldots, N$. The first step in obtaining a bound on the generalization error (statistical risk) $\|f - \hat{f}_n\|$ is to bound the approximation error $\|f - f_n\|$ of the best neural network of size n.

We shall make use of the following special case of a recent result in Barron (1993).

Theorem 1 *Given an arbitrary sigmoidal function ϕ, an arbitrary target function f with C_f finite, and a probability measure μ on a domain in $[-1, 1]^d$, then for every $n \geq 1$, there exists an artificial neural network of the form (1) such that*

$$\|f - f_n\| \leq \frac{C_f}{\sqrt{n}}, \tag{4}$$

For functions f with $C_f \leq C$, the parameters in (1) may be restricted to satisfy $\sum_{k=1}^n |c_k| \leq C$, $|c_0 - f(0)| \leq C$, and $|b_k| \leq |a_k|_1$.

Corollary 1 *If we constrain $|a_k|_1$ to be not larger than τ_n, then under the same restrictions as in Theorem 1, there exists an artificial neural network of the form (1) such that*

$$\|f - f_n\| \leq \frac{C_f}{\sqrt{n}} + C_f \, \mathrm{dist}(\phi_{\tau_n}, \mathrm{sgn}),$$

where $\mathrm{dist}(\phi_\tau, \mathrm{sgn})$ denotes the distance between the scaled sigmoidal function and the signum function given by

$$\mathrm{dist}(\phi_\tau, \mathrm{sgn}) = \inf_{0 < \epsilon \leq 1/2} (2\epsilon + \sup_{|z| \geq \epsilon} |\phi(\tau z) - \mathrm{sgn}(z)|).$$

In particular, assume τ_n is chosen such that

$$\mathrm{dist}(\phi_{\tau_n}, \mathrm{sgn}) \leq 1/\sqrt{n}. \tag{5}$$

Then

$$\|f - f_n\| \leq \frac{2C_f}{\sqrt{n}}. \tag{6}$$

The constraints on the magnitudes of the parameters are convenient for obtaining the statistical risk bound in Theorem 3 below. Later in this section, for Theorem reftheorem4, we drop these constraints and use penalty terms in the performance criteria to permit the automatic determination of magnitudes for the $|a_k|_1$ and $\sum_{k=1}^n |c_k|$ and the network size n that best resolves the function.

The choice of τ_n is based on the rate at which $\phi(z)$ approaches its limits. If $\phi(z)$ is equal to ± 1 outside a finite interval then τ_n may be set to be of order

$\sqrt{n}$, if $\phi(z)$ approaches its limits exponentially fast (as in the case of the standard sigmoid $(1-e^{-z})/(1+e^{-z})$) then τ_n may be set to be of order $\sqrt{n}\log n$, and if $\phi(z)$ approaches its limits polynomially fast (in the sense that $(-1-\phi(z))/|z|^p$ and $(1-\phi(z))/|z|^p$ remain bounded as $z \to -\infty$ and $z \to +\infty$, respectively, for some $p > 0$) then τ_n may be set to be of order $n^{(p+1)/2p}$. Henceforth, we restrict attention to sigmoidal functions ϕ for which τ_n is bounded by a polynomial function of n, i.e., $\tau_n \leq r_0 n^{r_1}$ for some $r_0, r_1 > 0$. For convenience, we will also restrict τ_n to be greater than or equal to some positive value τ_0 to be specified later in the treatment.

Throughout this paper logarithms are taken with base e. Expressions of the form $O(g(C_f, n, d, N))$ refer to quantities bounded by a constant times $g(C_f, n, d, N)$, for all permitted values of C_f, n, d, and N, where the constant is independent of f, n, d, and N. However, the constants may depend on the choice of sigmoid ϕ (e.g., through the quantities r_0 and r_1) and they may depend on the assumed range of the response variable (through the quantities λ and b introduced below); for simplicity, that dependence is not always made explicit.

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 $P_X = \mu$ with support contained in $[-1, 1]^d$, and with sample size $N \geq 2$. We assume that the support of Y is in a known interval I_0 with length bounded by some $b > 0$. (These distributional assumptions are a special case of somewhat more general assumptions in Barron (1990) that can also be used to handle some distributions for Y that have unbounded support and satisfy Bernstein's moment conditions, 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 $I_0 = [0, 1]$.

Note, in particular, that the commonly studied setting of Boolean functions that are observed without error at randomly selected inputs is covered by both of these cases. Our framework allows for the possibility that the response is subject to error, that is, $Y_i = f(X_i) + \epsilon_i$ where $E(\epsilon_i | X_i) = 0$; however, the assumption regarding the support of Y requires that both $f(x)$ and ϵ are bounded.

Because the function $f(x)$ is known to have range in a given interval $I_0 = [i_1, i_2]$, we improve the fit by replacing $f_n(x, \theta)$ with

$$\bar{f}_n(x, \theta) = \text{clip}(f_n(x, \theta)), \tag{7}$$

where $\text{clip}(y) = y$ for $y \in I_0$, $= i_1$ for $y < i_1$, and $= i_2$ for $y > i_2$. Note that *clip*() is a sigmoidal function with range I_0. 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 Barron (1990), see Theorem 2 below.

Our task is to incorporate the approximation bound into an analysis of the actual risk $\|f - \hat{f}_{n,N}\|$ for the network $\hat{f}_{n,N}$ that minimizes the empirical risk $(1/N)\sum_{i=1}^{N}(Y_i - f_n(X_i, \theta))^2$ (or a penalized version of this empirical risk). First we recall some machinery from Barron (1990) that permits us to carry out this task.

For each number of nodes n and sample size N, let $\Theta_n = \Theta_{n,N}$ be a discrete set of parameter vectors θ, and let $L_{n,N}(\theta)$ be nonnegative numbers satisfying, $L_{n,N}(\theta) \geq l$ for some constant $l > 0$, and

$$\sum_{\theta \in \Theta_n} e^{-L_{n,N}(\theta)} \leq 1. \tag{8}$$

The numbers $L_{n,N}(\theta) \ln 2$, rounded up to the nearest integer, may be interpreted as the lengths of a binary code for $\theta \in \Theta_n$, for then (8) becomes the Kraft-McMillan condition for the existence of uniquely decodable codes of these lengths (see Cover and Thomas (1991)). Another interpretation is that $e^{-L_{n,N}(\theta)}$ is a prior probability for $\theta \in \Theta_n$, for then (8) becomes the condition that these prior probabilities sum to not more than one. If $L_{n,N}(\theta) = L_{n,N}$ is constant and $\Theta_n = \Theta_{n,\varepsilon}$ is a finite set of points such that every θ is within distance ε of a point in $\Theta_{n,\varepsilon}$, then (8) implies that $L_{n,N}$ is a bound on the Kolmogorov ε-entropy (or metric entropy) of the set of possible θ's. Indeed, in that case (8) reduces to $\log \#\Theta_{n,\varepsilon} \leq L_{n,N}$, where $\#\Theta_{n,\varepsilon}$ denotes the cardinality of the set. (There is considerable freedom in the choice of metric to use in defining the ε-entropy: the requirement for us is that the metric on the parameter permits the distance between network functions to be controlled, see condition (19) and Lemmas 1 and 2). As in Barron and Cover (1991), we blend these complexity, prior probability, and metric entropy interpretations. The notation $L_{n,N}(\theta)$ is used here in place of the notation $C_N(\theta)$ from Barron (1990,1991) to avoid confusion with the quantity C_f defined above. We call $L_{n,N}(\theta)$ the complexity assigned to the parameter value θ in Θ_n, for a given n and a given sample size N.

For a given n and N, we define the *index of resolvability* (as in Barron 1990) to be

$$R_{n,N}(f) = \min_{\theta \in \Theta_n} \left(\|f - \bar{f}_n(\cdot, \theta)\|^2 + \lambda \frac{L_{n,N}(\theta)}{N} \right), \tag{9}$$

where λ is a given positive constant that will be restricted to a certain range of values in Theorem 2 (one valid choice is $\lambda = 2b^2$). Equation (9) gives the resolvability for a neural network family with a given number of nodes n. Let $L(n)$ be numbers satisfying $\sum_{n=1}^{\infty} e^{-L(n)} \leq 1$. For the collection of networks indexed by $n = 1, 2, \ldots$, the index of resolvability is

$$R_N(f) = \min_{n \geq 1} \left(R_{n,N}(f) + \lambda \frac{L(n)}{N} \right). \tag{10}$$

It will be seen that we may restrict the minimization in (10) to $n \leq N/d$, without affecting our bounds on the resolvability. This restriction has the effect of obeying the statistical rule that the number of parameters be less than the sample size. In fact the best n is typically much smaller. The minimization in (10) 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

$$\hat{f}_{n,N}(x) = \bar{f}_n(x, \hat{\theta}_{n,N}) \tag{11}$$

where

$$\hat{\theta}_{n,N} = \underset{\theta \in \Theta_n}{\operatorname{argmin}} \left(\frac{1}{N} \sum_{i=1}^{N} (Y_i - \bar{f}_n(X_i, \theta))^2 + \lambda \frac{L_{n,N}(\theta)}{N} \right). \tag{12}$$

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

$$\hat{f}_N = \hat{f}_{\hat{n},N} \tag{13}$$

where

$$\hat{n} = \underset{n}{\operatorname{argmin}} \left(\frac{1}{N} \sum_{i=1}^{N} (Y_i - \hat{f}_{n,N}(X_i))^2 + \lambda \frac{L_{n,N}(\hat{\theta}_{n,N})}{N} + \lambda \frac{L(n)}{N} \right). \tag{14}$$

Complexity regularization (as defined in (12) and (14)) is closely related to Vapnik's method of structural risk minimization (Vapnik 1982) and Rissanen's minimum description-length criterion (Rissanen, 1983, Barron and Cover, 1991).

We make use of a theorem from Barron (1989,1990), specialized here to neural network estimators.

Theorem 2 *Let a neural network be estimated by least squares with a complexity penalty as in (12), (14), 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 $N \geq 1$,*

$$E\|f - \hat{f}_{n,N}\|^2 \leq \gamma R_{n,N}(f) + \frac{2\gamma\lambda}{N}, \tag{15}$$

and

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

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

$$E\|f - \hat{f}_N\|^2 \leq O(R_N(f)). \tag{17}$$

To illustrate the small magnitude of the constant γ in the bound on the risk, note that $\gamma = 7$ and $\gamma = 2.5$ for the choices $\lambda = 2b^2$ and $\lambda = 3b^2$, respectively,

in the complexity regularization criterion. (Smaller values of λ than the $5b^2/3$ assumed in the theorem are sometimes advocated: for instance, the choice $\lambda = 2\sigma^2$ is motivated by the minimum description-length criterion using a Gaussian error model with variance σ^2 and Barron and Cover (1991) give analogous resolvability bounds for that case. It may be possible to use $\lambda = 2\sigma^2$ even in non-Gaussian cases, but it is not proven whether the theorem is valid with such a choice of λ.)

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 Barron (1990), 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,

$$\|f - \hat{f}_{n,N}\| \leq \|f - f_n\| + \|f_n - \hat{f}_{n,N}\|. \tag{18}$$

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 with finite C_f. This leads to Theorem 3 below, which is the new result of this paper.

For these estimation results, more regularity of the sigmoid in our analysis is needed than for the approximation results. One natural assumption is that the sigmoidal function $\phi(z)$ has a bounded derivative on $\mathbf{R}$. More generally, it is assumed that ϕ satisfies a Lipshitz condition, that is, $|\phi(z) - \phi(z^*)| \leq v_1|z - z^*|$ for some $v_1 > 0$. We let $v_0 \geq 1$ denote a bound on $|\phi(z)|$ and set $v = \max\{v_0, v_1\}$. (A special role is played by the ramp sigmoid that equals $v_1 z$ for $|z| < 1/v_1$ and equals -1 and $+1$, respectively, for $z \leq -1/v_1$ and $z \geq 1/v_1$. For Lipshitz sigmoidal functions it is seen that the quantity τ_n in (5) satisfies $\tau_n \geq \frac{2\sqrt{n}}{v_1}$ with equality when ϕ is a ramp sigmoid with slope v_1.) We continue to assume that $\phi(z)$ approaches its limits at least polynomially fast (as $z \to +\infty$ and $z \to -\infty$, respectively) so that, as mentioned above, the quantity τ_n can be bounded above by a polynomial function of n.

Recall that the parameter vector θ in the sigmoidal network (1) consists of the weights a_k, b_k, c_k for $k = 1, 2, \ldots, n$, and c_0, where each a_k is $d-$dimensional vector. For $\tau > 0$, a continuous parameter space $\Theta_{n,\tau}$ contained in $\mathbf{R}^{n(d+2)+1}$ is taken to be the set of all such θ for which $|a_k|_1 \leq \tau$ and $|b_k| \leq \tau$. For any $C > 0$, let $\Theta_{n,\tau,C} \subset \Theta_{n,\tau}$ be the subset of parameters with $\sum_{k=1}^{n} |c_k| \leq C$ and $c_0 \in I_C$, where $I_C = [i_1 - C, i_2 + C]$ and $I_0 = [i_1, i_2]$ is the given range of the target function. Theorem 1 and its corollary guarantees the existence of an accurate approximation to the target function $f(x)$ by a network $f_n(x, \theta)$ with $\theta \in \Theta_{n,\tau_n,C}$ provided $C \geq C_f$ and provided $\tau = \tau_n$ is chosen to satisfy (5). Note that the parameter space Θ_{n,τ_n} may be realized as the union of the compact sets $\Theta_{n,\tau_n,C}$ for $C = 1, 2, \ldots$: this fact enables us to provide estimates of f in the case that a bound on C_f is not known.

Next we control the precisions with which the coordinates of the parameter vectors are allowed to be represented. For each $\varepsilon > 0$ and $C \geq 1$, let $\Theta_{n,\varepsilon,\tau,C}$ be a discrete set of parameter points in $\mathbf{R}^{n(d+2)+1}$ that ε-covers $\Theta_{n,\tau,C}$ in the sense that for every

θ in $\Theta_{n,\tau,C}$ there is a θ^* in $\Theta_{n,\varepsilon,\tau,C}$ such that for $k = 1, 2, \ldots, n$,

$$|a_k - a_k^*|_1 \le \varepsilon$$

$$|b_k - b_k^*| \le \varepsilon$$

$$\sum_{k=1}^{n} |c_k - c_k^*| \le C\varepsilon \tag{19}$$

and

$$|c_0 - c_0^*| \le C\varepsilon.$$

In (19) one may use a distance of ε instead of $C\varepsilon$ for the c_k's; however, the broader spacing permits a somewhat smaller cardinality set with the same order of accuracy in covering the space of functions $\{f_n(\cdot,\theta) : \theta \in \Theta_{n,\tau,C}\}$. The accuracy of this coverage is indicated in the following lemma. In the lemma the choice of τ is arbitrary, whereas the choice $\tau = \tau_n$ satisfying (5) is needed for the corollary.

Lemma 1 *If (19) holds, then for each θ in the continuous parameter set $\Theta_{n,\tau,C}$ there is a θ^* in the discrete set $\Theta_{n,\varepsilon,\tau,C}$ such that uniformly for $x \in B$*

$$|f_n(x,\theta) - f_n(x,\theta^*)| \le 4vC\varepsilon$$

and hence

$$||f_n(\cdot,\theta) - f_n(\cdot,\theta^*)|| \le 4vC\varepsilon,$$

where $f_n(x,\theta)$ is the family of sigmoidal networks of the form (1).

Corollary 2 *For functions f with $C_f \le C$, there exists a neural net approximation f_n of the form (1) with parameter restricted to the discrete set $\Theta_{n,\varepsilon,\tau_n,C}$ such that*

$$||f - f_n|| \le \frac{2C_f}{\sqrt{n}} + 4vC\varepsilon.$$

Consequently, if a sequence ε_n is chosen to be of order $O(1/\sqrt{n})$, then the approximation error remains of the same order as in Theorem 1, that is,

$$||f - f_n|| = O(\frac{C}{\sqrt{n}}).$$

The proof of Lemma 1 is in Section 3. The corollary follows from Theorem 1 and Lemma 1 by application of the triangle inequality. It will be seen that a somewhat different choice of ε_n can achieve the best tradeoff between approximation error and complexity for a given n, N, and C.

Now we consider a choice for the finite set $\Theta_{n,\varepsilon,\tau,C}$ which satisfies (19) and examine the cardinality. We may take $\Theta_{n,\varepsilon,\tau,C}$ to be a rectangular grid spaced at width ε/d for the coordinates of a_k, width ε for b_k, width $C\varepsilon$ for c_0, and width $C\varepsilon/n$ for c_k for $k = 1, 2, \ldots, n$. Intersecting the grid with $\Theta_{n,\tau,C}$ yields a set $\Theta_{n,\varepsilon,\tau,C}$ that

satisfies the requirements of (19). The following lemma bounds the cardinality of this set. We have expressed both the grid spacing and the range for the c_k parameters as proportional to C, so the bound on the cardinality is independent of C; nevertheless, dependence of the cardinality on C may occur implicitly through choices of n and ε to be stated later.

Lemma 2 *For each $\varepsilon > 0$ and $C \ge 1$, there is a set $\Theta_{n,\varepsilon,\tau,C}$ that satisfies (19) and has cardinality bounded by*

$$\#\Theta_{n,\varepsilon,\tau,C} \le \left(\frac{2e(\tau+\varepsilon)}{\varepsilon}\right)^{n(d+1)} \left(\frac{2(1+\varepsilon)}{\varepsilon}\right)^{n} \left(\frac{b+2+\varepsilon}{\varepsilon}\right).$$

The proof of Lemma 2 is also given in Section 3. The factors in the bound arise naturally from a count of the number of choices for a_k, b_k, and c_k.

Lemma 2 shows that the logarithm of the cardinality of $\Theta_{n,\varepsilon,\tau_n,C}$ is bounded by

$$n(d+1)\log\left(2e(1+\frac{\tau_n}{\varepsilon})\right) + n\log\left(2(1+\frac{1}{\varepsilon})\right) + \log\left(1+\frac{b+2}{\varepsilon}\right).$$

Here we choose $\tau_0 \ge 1$ to be large enough that $2e(1+\tau_0) \ge b+3$. Then for $\tau_n \ge \tau_0$ and ε restricted to $0 < \varepsilon \le 1$ we have the convenient (but weaker) bound

$$\log \#\Theta_{n,\varepsilon,\tau_n,C} \le m_n \log\left(\frac{2e(1+\tau_n)}{\varepsilon}\right).$$

Here m_n is the total number of parameters in the network which equals

$$m_n = n(d+2)+1.$$

The metric entropy of the class of neural networks $\{f_n(\cdot,\theta) : \theta \in \Theta_{\tau,C}\}$ is the logarithm of the cardinality of the smallest set covering the class with error $||f_n(\cdot,\theta) - f_n(\cdot,\theta)||$ bounded by a given $\delta > 0$. Equating $\delta = 4vC\varepsilon$, it follows from Lemma 1 and Lemma 2 that the metric entropy is bounded by $m_n(\log(8vCe(1+\tau)/\delta)$, for $0 < \delta \le 4vC$.

ACCURACY OF THE CONSTRAINED LEAST SQUARES NEURAL NET ESTIMATOR

For each n, N, and C, let $\hat{f}_{n,N,C}(x)$ be defined as the least squares estimator with θ restricted to $\Theta_{n,\varepsilon,\tau_n,C}$, that is,

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

where

$$\hat{\theta}_{n,N,C} = \underset{\theta\in\Theta_{n,\varepsilon,\tau_n,C}}{\operatorname{argmin}} \left(\frac{1}{N}\sum_{i=1}^{N}(Y_i - \bar{f}_n(X_i,\theta))^2\right). \tag{20}$$

Note that this neural net estimator is the same as a complexity regularization estimator with constant complexity penalty $\lambda L_{n,N}(\theta)/N$ for θ in $\Theta_{n,\varepsilon,\tau_n,C}$, where $L_{n,N}(\theta)$ is set to equal the log-cardinality of $\Theta_{n,\varepsilon,\tau_n,C}$. Using the corollary to Lemma 1 and Lemma 2 we can bound the index of resolvability (defined in (9)) as follows

$$\begin{aligned} R_{n,N}(f) &\leq \|f - f_n\|^2 + \tfrac{\lambda}{N} \log \#\Theta_{n,\varepsilon,\tau_n,C} \\ &\leq 2\left(\tfrac{(2C_f)^2}{n} + (4v C\varepsilon)^2\right) + \tfrac{\lambda}{N} m_n \log\left(\tfrac{2e(1+\tau_n)}{\varepsilon}\right), \end{aligned} \tag{21}$$

assuming $0 < \varepsilon \leq 1$. By calculus it is seen that the choice for ε that optimizes this bound is

$$\varepsilon = \frac{1}{8vC}\left(\lambda \frac{m_n}{N}\right)^{1/2} \tag{22}$$

provided this choice is not greater than 1. With this choice for ε, the resolvability bound becomes

$$R_{n,N}(f) \leq \frac{8C_f^2}{n} + \frac{\lambda}{N}\frac{m_n}{2} \log\left((16vC(1+\tau_n))^2 e^3 \frac{N}{\lambda m_n}\right).$$

We recognize that this bound is of order

$$R_{n,N}(f) \leq O\left(\frac{C^2}{n}\right) + O\left(\frac{nd}{N} \log \frac{\tau_n^2 N}{nd}\right).$$

The implication for the constrained least squares estimator is summarized in the following Theorem. Using (21) it is seen that bounds of the desired order hold for a broad range of choices of ε. The precision $\varepsilon = \varepsilon(n, d, N, C)$ may be allowed to depend on $n \geq 1$, $d \geq 1$, $C \geq 1$, and $N \geq 2$. It is assumed to be not smaller than the reciprocal of a polynomial in N, n, d and C, that is, for some $p \geq 1$,

$$\varepsilon \geq \left(\frac{1}{CndN}\right)^p. \tag{23}$$

Also it is assumed either that

$$\varepsilon \leq O(\frac{1}{\sqrt{n}}) \tag{24}$$

or that

$$\varepsilon \leq O\left(\frac{1}{C}\left(\frac{nd}{N}\log N\right)^{1/2}\right). \tag{25}$$

The constraint in (24) or (25) is imposed so that in the additional error due to discretization of the parameters of the network approximation is not large compared to either the approximation term or the complexity term, respectively, of the index of resolvability. The constraint in (23) is imposed to prevent excessive complexity penalties that would result from too fine a precision. A consequence of (23) (and the assumption that τ_n is bounded by a polynomial in n) is that the log-cardinality of $\Theta_{n,\varepsilon,\tau_n,C}$ satisfies

$$\log \#\Theta_{n,\varepsilon,\tau_n,C} \leq O\left(nd \log(CndN)\right). \tag{26}$$

Theorem 3 *Let $\hat{f}_{n,N,C}(x)$ be the least squares neural net estimator defined above with parameter vector restricted to a set $\Theta_{n,\varepsilon,\tau_n,C}$, satisfying (19), (26), and either (24) or (25), with τ_n satisfying (5). If the target function f satisfies $C_f \leq C$, then the global accuracy of the estimator, as measured by the mean integrated squared error, satisfies*

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

which is of order $O(C((d/N)\log N)^{1/2})$ for $n \sim C(N/(d\log N))^{1/2}$.

The proof of Theorem 3 follows from plugging the bound on ε from either (24) or (25) into the approximation error bound of the corollary to Lemma 1 and then adding the complexity penalty bound from (26) to get that the index of resolvability satisfies

$$R_{n,N}(f) \leq O\left(\frac{C^2}{n}\right) + O\left(\frac{nd}{N}\log(CndN)\right).$$

Consequently, by Theorem 2,

$$E\|f - \hat{f}_{n,N,C}\|^2 \leq O\left(\frac{C^2}{n}\right) + O\left(\frac{nd}{N}\log(CndN)\right). \tag{28}$$

To complete the proof note that $\|f - \hat{f}_{n,N,C}\|^2$ is bounded by a constant $(2b)^2$. Therefore, the desired bound (27) trivially holds for those C, n, d, and N for which either $C^2 \geq n$ or $nd \geq N/\log N$. It remains to restrict attention the those cases in which $C^2 < n$ and $nd < N/\log N$, but then the desired bound follows from (28).

As we have seen it is possible to give explicit bounds that involve rather messy expressions, if instead of treating general sets, we use the specific choice of $\Theta_{n,\varepsilon,\tau_n,C}$ that has the cardinality bound of Lemma 2. It is of some interest to give a detailed bound in the case of the ramp sigmoid (for which $\phi(z)$ equals z for $|z| \leq 1$ and equals -1, $+1$ for $z \leq -1$, $z \geq 1$, respectively), because the ramp sigmoid has the smallest τ_n, equal to $2\sqrt{n}$, when $v = 1$. With this choice and with ε chosen as in (22), the ratio τ_n/ε is not greater than $16C(N/\lambda(d+2))^{1/2}$ (uniformly in the size of the network n) and the index of resolvability is bounded by

$$R_{n,N}(f) \leq \frac{8C^2}{n} + \frac{\lambda}{N} m_n \log\left(2e^{3/2}\left(1 + 16C\left(\frac{N}{\lambda(d+2)}\right)^{1/2}\right)\right). \tag{29}$$

This bound is optimized with a number of network nodes equal to

$$n = C\left(8N/\left(\lambda(d+2)\log\left(2e^{3/2}\left(1+16C\left(\frac{N}{\lambda(d+2)}\right)^{1/2}\right)\right)\right)\right)^{1/2}$$

rounded to an integer, yielding a resolvability that satisfies

$$R_{n,N}(f) \leq 4C\left(\frac{2\lambda(d+3)}{N}\log\left(2e^{3/2}\left(1+16C\left(\frac{N}{\lambda(d+2)}\right)^{1/2}\right)\right)\right)^{1/2}. \tag{30}$$

Consequently, we have an accurate neural network estimator if $C(d/N)^{1/2}$ is small.

ACCURATE NEURAL NET ESTIMATOR WHEN C IS UNKNOWN

If a bound on C_f is not known, then the constrained least squares estimators given above would not necessarily yield accurate estimates because we would not necessarily have C_f less than or equal to a prescribed C. One fix is to try an increasing sequence of bounds C, impose a suitable penalty, and chose a $\hat{C}$ that yields an accurate estimate.

Let $\hat{f}_{n,N}(x)$ be the complexity regularization estimator of the function for θ in $\Theta_{n,\varepsilon,\tau_n} = \bigcup_{C=1}^{\infty} \Theta_{n,\varepsilon,\tau_n,C}$ given by

$$\hat{f}_{n,N}(x) = \hat{f}_{n,N,\hat{C}}(x), \tag{31}$$

where

$$\hat{C} = \underset{C}{\operatorname{argmin}} \left(\frac{1}{N} \sum_{i=1}^{N} (Y_i - \hat{f}_{n,N,C}(X_i))^2 + \frac{\lambda}{N} \left(\log \#\Theta_{n,\varepsilon,\tau_n,C} + 2\log(1+C) \right) \right), \tag{32}$$

and $\hat{f}_{n,N,C}$ is the least squares estimate given in (20). The minimization in (32) is restricted to positive integer values of C. The term $2\log(1+C)$ is a convenient choice for which the summability condition (8) is satisfied by the complexity penalty. In place of $2\log(1+C)$ we may use $\log 1/p(C)$ for any probability mass function $p(C)$ for which $\log 1/p(C) = O(\log C)$.

The index of resolvability can be bounded in the same manner as before, using the fact that the parameter space now includes integers C that are greater than C_f. We find that it is bounded by

$$R_{n,N}(f) \le O\left(\frac{C_f^2}{n}\right) + O\left(\frac{nd}{N} \log C_f ndN\right), \tag{33}$$

and consequently that the mean integrated squared error satisfies

$$E\|f - \hat{f}_{n,N}\|^2 \le O\left(\frac{C_f^2}{n}\right) + O\left(\frac{nd}{N} \log N\right). \tag{34}$$

NEURAL NET ESTIMATES USING A COMPLEXITY PENALTY BASED ON PRIOR PROBABILITY

The use of a prior for C is a special case of a more general class of complexity penalties. The parameter space does not have to be a union of discrete sets restricted to the spaces $\Theta_{n,\tau_n,C}$, with complexity penalties that are constant on each. These choices were partially a matter of convenience. We are free to use larger parameter spaces that do not restrict the magnitude of the a_k b_k, and c_k and to use non-constant penalties subject to the summability constraint (8). The essence

of what is needed for the above reasoning to work is that there is are finite subsets of the parameter space Θ_n that accurately cover the spaces $\Theta_{n,\tau_n,C}$ and that the complexities assigned to the elements of at least one such approximate cover is of order not larger than $nd\log(CndN)$.

One way to accomplish this is to take a prior probability distribution with density function $p_n(\theta)$ that is continuous and positive on all of $\mathbf{R}^{m_n}$ where $m_n = n(d+2)+1$. Then partition $\mathbf{R}^{m_n}$ into disjoint rectangles of width $\varepsilon/(d+1)$ for the coordinates of a_k and b_k and width $\varepsilon/(n+1)$ for the c_k, where $\varepsilon = 1/\sqrt{n}$. With these choices conditions (19) and (24) are satisfied. By the mean value theorem, within each of these rectangles there is a point θ (called a representer of the rectangle) such that the prior probability of the rectangle is equal to the volume times the value of the density at θ. We take the complexity penalty to be the minus logarithm of the prior probability of the rectangles (for each of the disjoint rectangles in $\mathbf{R}^{m_n}$) and we take the discrete parameter space Θ_n to be these representers θ. Then the summability condition (8) is satisfied as is the requirement (19) for $C \ge 1$. The complexity penalties satisfy

$$L_{n,N}(\theta) = n(d+1)\log\frac{(d+1)}{\varepsilon} + (n+1)\log\frac{n}{\varepsilon} + \log\frac{1}{p_n(\theta)}. \tag{35}$$

The first two terms play an important role in complexity regularization. Compare with maximum posterior likelihood estimators, in which only the $\log 1/p_n(\theta)$ term would appear. For a given size model, complexity regularization and maximum posterior likelihood estimators can be essentially the same; however, complexity regularization adds additional terms that provide the possibility of accurate selection among different size models.

For application of the theory in the case of variable length complexity penalties, we check the analogue of (26), namely that

$$L_{n,N}(\theta) = O(nd\log(CndN)) \tag{36}$$

uniformly in the intersection of Θ_n with the bounded set $\Theta_{n,\tau_n,C}$, for each $C \ge 1$. With the given choice of $\varepsilon = 1/\sqrt{n}$, it is enough to check that

$$\log\frac{1}{p_n(\theta)} = O(nd\log Cnd) \tag{37}$$

uniformly on $\Theta_{n,\tau_n,C}$ for each $C \ge 1$. For example, a prior density that makes the coordinates of the parameter vector independent standard Cauchy random variables satisfies the condition (37) as do other densities with polynomial tails, while a prior density with independent standard normal components does not satisfy this condition. It is possible to formulate the reasonable bounds for a normal prior, provided care is taken to chose prior variances for the a_k and b_k components to be proportional to τ_n^2. Then with unit prior variance for the c_k components, the normal prior satisfies $\log\frac{1}{p_n(\theta)} = O(C^2) + O(nd\log Cnd)$ uniformly on $\Theta_{\tau_n,C}$ for each $C \ge 1$. The order C^2 term in the complexity does not adversely affect the

order of the bounds, since it contributes order C^2/N to the resolvability, which is dominated by the sum of the other terms in the bound.

We conclude the main part of the paper by stating a theorem that encompasses the different cases in which we have obtained the rate of convergence stated in the abstract. It is assumed that there is a discrete parameter space $\Theta_{n,N}$ and complexity penalties satisfying (8). It is assumed that for each $C \geq 1$, there is a subset of $\Theta_{n,N} \cap \Theta_{n,\tau_n,C}$ denoted $\Theta_{n,\varepsilon,\tau_n,C}$ satisfying (19) for some ε satisfying (24) or (25). The complexity penalties assigned to the parameter values in this subset are assumed to satisfy

$$L_{n,N}(\theta) \leq O(C^2) + O(nd \log CndN) \tag{38}$$

uniformly for θ in $\Theta_{n,\varepsilon,\tau_n,C}$ for each $C \geq 1$.

We let $\hat{f}_{n,N}$ be the neural network estimated by complexity regularization as in (11), (12) and we let $\hat{f}_N(x) = \hat{f}_{\hat{n},N}(x)$ be the complexity regularization estimator with $\hat{n}$ determined as in (14) with the penalty $L(n)$ chosen to satisfy $\sum_{n=1}^{\infty} e^{-L(n)} \leq 1$. (One such convenient choice is $L(n) = 2\log(1+n)$.) It is assumed that $L(n) \leq O(n)$.

Theorem 4 *Let $\hat{f}_{n,N}$ and $\hat{f}_N$ be the neural network estimators defined above, where the complexity penalties satisfy the summability condition (8) over all candidate values of the parameters. It is assumed for each $C \geq 1$ there is a subset of these parameters that satisfies conditions (19) and (38) for some precision satisfying either (24) or (25). Then the statistical risk satisfies*

$$E\|f - \hat{f}_{n,N}\|^2 \leq O\left(\frac{C_f^2}{n}\right) + O\left(\frac{nd}{N} \log N\right) \tag{39}$$

and

$$E\|f - \hat{f}_N\|^2 \leq O\left(C_f \left(\frac{d}{N} \log N\right)^{1/2}\right), \tag{40}$$

respectively.

Theorem 4 is proven in the same manner as Theorem 3 above, using the lemmas and appealing to the results of Theorems 1 and 2.

3. Proofs of Lemmas

Here we prove Lemmas 1 and 2.

For the proof of Lemma 1, let θ and θ^*, respectively, be parameter vectors in $\Theta_{n,\tau,C}$ and $\Theta_{n,\varepsilon,\tau,C}$ for which (19) holds. Consider the difference between the values of the corresponding network functions

$$f_n(x,\theta) - f_n(x,\theta^*) = \sum_{k=1}^{n} c_k \phi(z_k) - c_k^* \phi(z_k^*) + (c_0 - c_0^*)$$

$$= \sum_{k=1}^{n} c_k \left(\phi(z_k) - \phi(z_k^*)\right) + \sum_{k=1}^{n} (c_k - c_k^*)\phi(z_k^*) + (c_0 - c_0^*). \tag{41}$$

where $z_k = a_k^T x + b_k$ and $z_k^* = a_k^{*T} x + b_k^*$. By the triangle inequality and the Lipshitz condition this yields

$$|f_n(x,\theta) - f_n(x,\theta^*)| \leq \sum_{k=1}^{n} |c_k||\phi(z_k) - \phi(z_k^*)| + \sum_{k=1}^{n} |c_k - c_k^*||\phi(z_k^*)| + |c_0 - c_0^*|$$

$$\leq v_1 \sum_{k=1}^{n} |c_k||z_k - z_k^*| + v_0 \sum_{k=1}^{n} |c_k - c_k^*| + |c_0 - c_0^*|. \tag{42}$$

Now $|z_k - z_k^*|$ is bounded by 2ε for x in B and $v = \max(v_1, v_0)$ is at least 1, so it follows that

$$|f_n(x,\theta) - f_n(x,\theta^*)| \leq 2v\varepsilon \sum_{k=1}^{n} |c_k| + v \sum_{k=1}^{n} |c_k - c_k^*| + |c_0 - c_0^*|$$

$$\leq 4vC\varepsilon. \tag{43}$$

Here we have used (19) and the bound $\sum_{k=1}^{n} |c_k| \leq C$ for θ in $\Theta_{n,\tau,C}$. This completes the proof of Lemma 1.

Finally we give the proof of Lemma 2. Consider a rectangular grid spaced at width ε/d for the coordinates of a_k, width ε for b_k width $C\varepsilon$ for c_0, and width $C\varepsilon/n$ for c_k, for $k = 1, 2, \ldots, n$. Intersecting the grid with $\Theta_{n,\tau,C}$ yields the desired set $\Theta_{n,\varepsilon,\tau,C}$ satisfying the requirements of (19) and we bound its cardinality. Now $\Theta_{n,\tau,C} = \{\theta : |a_k|_1 \leq \tau, |b_k| \leq \tau, c_0 \in I_C, \sum_{k=1}^{n} |c_k| \leq C\}$ is a cartesian product of constraint sets for the a's b's and c's so the desired cardinality is obtained as a product of the corresponding counts. First we bound the number of grid points in the simplex $S_\tau = \{a \in R^d : |a|_1 \leq \tau\}$, where the grid points are spaced at width ε/d in each coordinate. The union of the small cubes that intersect S_τ is contained in $S_{(\tau+\varepsilon)}$. (Indeed any point a in this union has ℓ_1 distance less than ε from a point a' in S_τ, whence $|a|_1 \leq |a'|_1 + |a - a'|_1 \leq \tau + \varepsilon$, so a is in $S_{\tau+\varepsilon}$.) Trivially, the volume of this union of cubes is the product of the number of cubes and $(\varepsilon/d)^d$. Also, the volume of the covering simplex is $(2(\tau+\varepsilon))^d/d!$. So the number of cubes that intersect S_τ is not greater $(2d(\tau+\varepsilon)/\varepsilon)^d/d! \leq (2e(\tau+\varepsilon)/\varepsilon)^d$. For n such parameter vectors a_k, the total count is bounded by $(2e(\tau+\varepsilon)/\varepsilon)^{nd}$. In the same way the count for the b's is not more than $(2(\tau+\varepsilon)/\varepsilon)^n$, the count for c_0 is not more than $(b + 2C + C\varepsilon)/(C\varepsilon) \leq (b+2+\varepsilon)/(\varepsilon)$ for $C \geq 1$, and the count for the rest of the c_k's is not more than $(2e(1+\varepsilon)/\varepsilon)^n$. Taking the product of the counts yields

$$\#\Theta_{n,\varepsilon,\tau,C} \leq \left(\frac{2e(\tau+\varepsilon)}{\varepsilon}\right)^{nd} \left(\frac{2(\tau+\varepsilon)}{\varepsilon}\right)^{n} \left(\frac{2e(1+\varepsilon)}{\varepsilon}\right)^{n} \left(\frac{b+2+\varepsilon}{\varepsilon}\right).$$

This completes the proof of Lemma 2.

We note that some reduction in the count is possible by restricting to positive c_k coefficients (which does not restrict the class of functions represented by formula (1) in the case of sigmoids that satisfy the symmetry property $\phi(-z) = -\phi(z)$). Also, the a_k vectors may be restricted to have magnitude equal to τ_n (instead of $|a_k|_1 \leq \tau_n$). The effect of the equality constraint would be to reduce the dimension of the parameter space from $n(d+2)+1$ to $n(d+1)+1$. For simplicity we have not taken advantage of the slight improvements in the constants that would result from such reductions.

Remark

After this paper was completed for the *1991 Workshop on Computational Learning Theory*, McCaffrey and Gallant (1991) obtained an extension to Theorem 2 that uses a continuity argument based on the proof in Barron (1990) to show that the network optimization need not be restricted to a grid of parameter points. For instance, using their extension it is permitted to use the least squares estimator over the continuum of values of θ in $\Theta_{n,\tau_n,C}$, to get the bound $O(C^2/n)+O(nd/N)\log N$ for the risk in the case of estimation of functions with $C_f \leq C$. McCaffrey and Gallant use their result to establish rates of convergence for the statistical risk of sigmoidal networks for functions in the traditional Sobolev smoothness classes. It may also be possible to adapt their argument to determine conditions that permit a suitable penalty term that depends continuously on θ without constraint on the magnitudes of the parameters.

Acknowledgements

This work was supported by ONR contracts N00014-89-J-1811 and N00014-93-1-0085.

References

Barron, A. R. (1989). Statistical properties of artificial neural networks. *Proceedings of the IEEE International Conference on Decision and Control*, (pp. 280–285). New York: IEEE.

Barron, A. R. (1990). Complexity regularization with applications to artificial neural networks. In G. Roussas (ed.) *Nonparametric Functional Estimation*, (pp. 561–576). Boston, MA and Dordrecht, the Netherlands: Kluwer Academic Publishers.

Barron, A. R. (1991). Approximation and estimation bounds for artificial neural networks. *Proceedings of the Fourth Workshop on Computational Learning Theory*, (pp.243–249). San Mateo, CA: Morgan Kaufmann Publishers. (Prelimary version of the present paper).

Barron, A. R. (1992). Neural net approximation. *Proceedings of the Seventh Yale Workshop on Adaptive and Learning Systems*, (pp. 69–72). K. S. Narendra (ed.), Yale University.

Barron, A. R. (1993). Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information Theory, 39*, 930–945.

Barron, A. R. & Cover, T. M. (1991). Minimum complexity density estimation. *IEEE Transactions on Information Theory, 37*, 1034–1054.

Barron, A. R. & Sheu, C.-H. (1991). Approximation of density functions by sequences of exponential families. *Annals of Statistics, 19*, 1347–1369.

Cover, T. M. & Thomas, J. (1991). *Elements of Information Theory*, New York: Wiley.

Cox, D. D. (1988). Approximation of least squares regression on nested subspaces. *Annals of Statistics, 16*, 713–732.

Cybenko, G. (1989). Approximations by superpositions of a sigmoidal function. *Mathematics of Control, Signals, and Systems, 2*, 303–314.

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

Hardle, W. (1990). *Applied Nonparametric Regression*, Cambridge, U.K. and New York: Cambridge University Press.

Haussler, D. (1992). Decision theoretic generalizations of the PAC model for neural net and other learning applications. *Information and Computation, 100*, 78-150.

Hornik, K., Stinchcombe, M., & White, H. (1988). Multi-layer feedforward networks are universal approximators. *Neural Networks, 2*, 359-366.

Ibragimov, I. A., and Hasminskii, R. Z. (1980). On nonparametric estimation of regression. *Doklady Acad. Nauk SSSR, 252*, 780-784.

Jones, L. K. (1992). A simple lemma on greedy approximation in Hilbert space and convergence rates for projection pursuit regression and neural network training. *Annals of Statistics, 20*, 608-613.

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

McCaffrey, D. F. & Gallant, A. R. (1991). Convergence rates for single hidden layer feedforward networks. Rand Corporation working paper, Santa Monica, California and Institute of Statistics Mimeograph Series, Number 2207, North Carolina State University.

Nemirovskii, A. S. (1985). Nonparametric estimation of smooth regression functions. *Soviet Journal of Computer and Systems Science, 23*, 1-11.

Nemirovskii, A. S., Polyak, B. T. & Tsybakov, A. B. (1985). Rate of convergence of nonparametric estimators of maximum-likelihood type. *Problems of Information Transmission, 21*, 258-272.

Nussbaum, M. (1986). On nonparametric estimation of a regression function that is smooth in a domain on R^k. *Theory of Probability and its Applications, 31*, 118-125.

Pinsker, M. S. (1980). Optimal filtering of square-integrable signals on a background of Gaussian noise. *Problems in Information Transmission, 16*.

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

Seber, G. A. F. & Wild, C. M. (1989). *Nonlinear Regression*, New York: Wiley.

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

Stone, C. J. (1982). Optimal global rates of convergence for nonparametric estimators. *Annals of Statistics, 10*, 1040-1053.

Stone, C. J. (1990). Large-sample inference for log-spline models. *Annals of Statistics, 18*, 717–741.

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

White, H. (1990). Connectionist nonparametric regression: multilayer feedforward networks can learn arbitrary mappings. *Neural Networks, 3*, 535-550.