Proc. Seventh Yale Workshop on
Adaptive and Learning Systems
May 20-22, 1992

NEURAL NET APPROXIMATION

Andrew R. Barron

University of Illinois at Urbana-Champaign, and
Barron Associates, Inc., Stanardsville, Virginia

New address:
Dept. of Statistics
Yale University
Box 2179 - Yale Station
New Haven, CT 06520

ABSTRACT

Recent results on the accuracy of neural net approximations of functions are discussed and refined. The nets considered are feedforward artificial neural networks with one hidden layer of sigmoidal activation functions. Bounds on the maximum approximation error as well as the integrated squared error are given. Lower bounds on the approximation rate are developed that closely match the upper bounds when the dimension of the input vector is large. The surprising fact is that the limiting approximation rate is independent of the dimension. The functions approximated are assumed to satisfy a bound on their variation with respect to half-spaces, or, more restrictively, a bound on a spectral norm. Fourier analysis, empirical process theory, and the theory of nonparametric regression are used in the proofs of the approximation bounds.

INTRODUCTION

It is known from [1-2] that arbitrarily accurate approximations to continuous functions on bounded subsets of d variables is possible by the use of linear combinations of sigmoidal functions. Under additional restrictions on the functions to be approximated, bounds on the number of terms sufficient to obtain an accurate approximation are established by the author in [3]. There it is shown that if C_f is the first moment of the Fourier magnitude distribution of a function $f(x)$, then the L_2 norm of the approximation error by a T term sigmoidal network is bounded by $2C_f/T^{1/2}$. The surprising aspect of this approximation bound is that the rate $1/T^{1/2}$ is independent of the dimension d of the input vector. In the sigmoidal approximations, the location and orientation parameters internal to the nodes are adjusted in the approximation. This has the effect of nonlinear adjustment of the basis functions. In contrast, no linear combination of T fixed basis functions, as in traditional series expansions, can achieve approximation error uniformly smaller than order $C/T^{1/d}$. Consequently, for the class of functions studied, the nonlinear sigmoidal net approximations are considerably better for all dimensions $d > 2$.

Implications for the estimation of functions from a sample of N independent observations are given in [4]. In addition to the assumption of a spectral norm C_f that is not exponentially large in the dimension d, assumptions are also made regarding the sampling distributions of the observations of X, Y, with the target function given by $E[Y|X] = f(X)$. The function is assumed to be estimated by minimization of the sum of squares of errors of fit by a T term sigmoidal net, with a constraint imposed on the domains of the parameters (allowances are also made to incorporate Bayesian-type penalties on the parameters in the optimized criterion). The result is that the total mean squared error between the estimated function $\hat{f}(x)$ and the true function $f(x)$ is bounded by order $C_f^2/T + (Td/N)\log N$. This mean squared error quantifies the ability of the network to "generalize" to new data not observed in the data base (since the average in the definition of the mean squared error is taken over the distribution of possible values of x, and not just over the observed values). The two terms in this bound express the tradeoff between the accuracy of the best approximation (which requires T large) and the accuracy of the empirical fit to this theoretical approximation (which requires T/N small). Then either by setting the number of terms T to be of order $N^{1/2}$, or by estimating a number of terms $\hat{T}$ from the data by a complexity-based model selection criterion (related to Rissanen's MDL), it is shown in [4] that the mean squared error between $\hat{f}$ and f is bounded by order $1/N^{1/2}$ times a polynomial factor in d and a logarithmic factor in N. Thus exponentially large sample sizes are not required to get accurate estimates for the class of functions considered.

In this workshop paper several extensions to the approximation bounds are given that might be of some interest. Bounds of order $1/T^{1/2}$ are given for the maximum of the error of approximation, that is, the L_∞ norm, extending the results developed previously for the L_2 norm. The general condition for this approximation bound is stated in terms of a notion of bounded variation with respect to indicators of half-spaces. Functions with finite spectral norm then serve as a special case. An expression for functions with finite spectral norm is given that provides a integral representation as an infinite mixture of indicators of half-spaces, with a probability density function determined by the Fourier representation. The approximation bounds then follow from empirical process theory associated with samples from this density. Also lower bounds on the approximation rate for sigmoidal nets are given that closely match the upper bounds when the dimension d is large.

STATEMENT OF THE BOUNDS

First we introduce a convenient class of functions for studying network approximation. The context involves real-valued functions $f(x)$ of d variables. The input vector x takes values in a bounded set B in R^d, assumed for convenience to include the point 0. For instance B may be the cube $[-1,1]^d$. Let $\mathcal{S}$ be a class of subsets of R^d. A function f is said to have bounded variation with respect to $\mathcal{S}$ if it is in the closure of the set of linear combinations of indicator functions $1_S(x)$ for $S \in \mathcal{S}$, with the sum of the absolute values of the coefficients of linear combination not greater than some finite number V. The infimum of such V is called the variation of the function f with respect to $\mathcal{S}$ and is denoted $V_f = V_{f,\mathcal{S},B}$. The closure is taken with respect to uniform convergence on B. Particular interest is given to the case that $\mathcal{S} = \mathcal{S}_{h_s}$ is the class of half-spaces $\{x : a \cdot x + b > 0\}$ or $\{x : a \cdot x + b \geq 0\}$. When $d = 1$ and the value 0 is in the range of the function f, the above notion agrees with the classic definition of bounded variation. [Recall that if a function if $f(x)$ is continuously differentiable except at a discrete set of jump points, then V_f equals $\int_B |f'(x)|dx$ plus the sum of the jump heights.] For $d > 1$, $V_{f,\mathcal{S}_{hs},B}$ is one of the possible extensions of the notion of bounded variation (another extension would be to use the regions $\{x : x_1 \le a_1, \ldots, x_d \le a_d\}$, $a \in R^d$ in place of the half-spaces). Unlike the one-dimensional case, $V_{f,\mathcal{S}_{hs},B}$ does not equate to the L_1 norm of the gradient for continously differentiable functions. Nevertheless, it can be bounded in terms of the L_1 norm of the Fourier transform of the gradient as will be seen below.

For parameterized classes of subsets, a role is to be played by the sets of parameters that provide coverage of each x in B. Let $\mathcal{S}$ be a class of subsets S_u of R^d, parameterized by a vector u of dimension, say d'. Then we have a dual class $\mathcal{S}'$ of subsets of $R^{d'}$ given by $S'_x = \{u \in R^d : x \in S_u\}$, which is parameterized by x in R^d. Thus $u \in S'_x$ if and only if $x \in S_u$. Note that for the half-spaces $S_u = \{x : a \cdot x + b > 0\}$, parameterized by $u = (a, b)$, the dual sets are half-spaces. The same is true for elipsoidal and hyperbolic regions in which the linear functions $a \cdot x + b$ are replaced by quadratic polynomials: the dual sets are half-spaces but with a larger dimension d'. The Vapnik-Chervonenkis condition for a dual class $\mathcal{S}'$ (restricted to $x \in B$) is the requirement that for all $u_1, u_2, \ldots u_T$ the number of subsets of $\{u_1, u_2, \ldots u_T\}$ obtained by intersecting with S'_x for $x \in B$ is strictly less than 2^T for some T. The first T for which the condition holds is the V-C dimension D. For half-spaces, $D = d'$.

A desired approximation property for functions with bounded variation with respect classes of subsets $\mathcal{S}$ is this: given a function f, there exists parameter values $u_1, \ldots, u_T$ and $c_1, \ldots, c_T$ such that the approximation $f_T(x) = \sum_{k=1}^{T} c_k 1_{S_{u_k}}(x)$ has approximation error $|f(x) - f_T(x)|$ bounded by order $V/T^{1/2}$ uniformly on B. This deterministic approximation property can be proven by a probabilistic argument for classes of subsets for which the dual $\mathcal{S}'$ satsifies the V-C condition.

The idea of the proof is as follows. Given any $\delta > 0$, there is a linear combination of indicators of sets in $\mathcal{S}$, with the sum of the absolute values of the coefficients not more than V, such that the maximum of the approximation error is less than δ. We take such a linear combination with $\delta = \gamma V/T^{1/2}$, where γ is an arbitrary positive constant. (The assumption of bounded variation with respect to $\mathcal{S}$ guarentees that this can be done, but it does not as yet place a restriction on the number of terms in this sum.) We partition this approximation into two sets of terms depending on whether the sign of the coefficients is positive or negative. Let V^+ and V^- be the sum of the coefficient values for the two sets, repectively. With $T' = T/2$, we draw $u_1, u_2, \ldots, u_{T'}$ at random (with replacement) from the set of positive terms in the linear combination with probabilities proportional to the coefficients in this combination. If the class $\mathcal{S}'$ satisfies the Vapnik-Chervonenkis condition, then by the central limit theorem for empirical processes (due to Dudley [5]), the probability that the maximum for $x \in B$ of the difference between the sample average $f_T^+(x) = (1/T')\sum_{k=1}^{T'}(V^+)1_{S_{u_k}}(x)$ and its expected value is greater than the amount $(\gamma V^+)/T^{1/2}$ converges as $T -> \infty$ to a probability that is strictly between 0 and 1. This implies that there exists choices for $u_1, u_2, \ldots, u_{T'}$ for each large T for which the maximum difference between the sample average and its expectation is less than $(\gamma V^+)/T^{1/2}$. Doing the same for the negative part and setting $f_T(x) = f_T^+(x) - f_T^-(x)$, we find by the triangle inequality that $\sup_{x\in B} |f_T(x) - f(x)| < 2\gamma V/T^{1/2}$. Since in particular, the class of half-spaces is a Vapnik-Chervonenkis class, this provides an approximation theorem for artificial neural networks with unit step activation functions. For completeness we state also an L_2 bound proved by the method of [3] that does not require that $\mathcal{S}'$ be a V-C class, and which holds for all T.

Theorem 1 (Upper bound on the approximation rate). For each function f with bounded variation on B with respect to a class of sets $\mathcal{S}$, there is an approximation f_T which is a linear combination of T indicators of sets in $\mathcal{S}$, such that for all $T \ge 1$,

$$\|f - f_T\|_2 \le \frac{V_{f,\mathcal{S},B}}{T^{1/2}}, \tag{1}$$

where $\|f - f_T\|_2$ is the L_2 approximation error with respect to any given probability measure μ on B (that is, $\|f - f_T\|_2^2 = \int_B (f(x) - f_T(x))^2 \mu(dx)$) and $V_{f,\mathcal{S},B}$ is the variation of f on B with respect to the class of sets $\mathcal{S}$.

If also $\mathcal{S}$ is a parameterized class of sets for which the dual is a V-C class (such as the class of half-spaces) of dimension D, then for every $\gamma > 0$, there exists $T(\gamma, D)$ such that for all $T \ge T(\gamma, D)$,

$$\sup_{x\in B} |f(x) - f_T(x)| \le \gamma \frac{V_{f,\mathcal{S},B}}{T^{1/2}}. \tag{2}$$

Consequently, $\|f - f_T\|_\infty = o(1/T^{1/2})$ where $\|f - f_T\|_\infty$ denotes the L_∞ norm on B. Moreover, there exists a constant γ_D such that (2) holds for all $T \ge 1$, with $\gamma = \gamma_D$.

Calculations using empirical process bounds from Pollard [6] for the nonasymptotic case, show that we can take γ_D not larger than $60D$. More refined bounds on the constant for the L_∞ bound are being sought.

The degree of generality of the assumption of bounded variation with respect to half-spaces is revealed in part by the following Theorem. Here it is shown that for functions with an integrable Fourier representation, the variation with respect to $\mathcal{S}_{hs}$ is related to a spectral norm.

Theorem 2: If a function $f(x)$ has a Fourier representation $f(x) = \int e^{i\omega\cdot x}\tilde{f}(\omega)d\omega$ valid for $x \in B$, and if $\omega\tilde{f}(\omega)$ is integrable, then the following integral representation holds,

$$f(x) = f(0) + \int_{R^d}\int_0^1 (1_{\{\alpha\cdot x < -t\}} - 1_{\{\alpha\cdot x > t\}})|\omega|_B \sin(t|\omega|_B + \theta_\omega)|\tilde{f}(\omega)|d\omega dt \tag{3}$$

where $\alpha = \omega/|\omega|_B$ denotes the orientation of the frequency vector, $\tilde{f}(\omega) = e^{i\theta_\omega}|\tilde{f}(\omega)|$ denotes the Fourier magnitude and phase decomposition, and $|\omega|_B = \sup_{x\in B}|\omega \cdot x|$ (which equals the ℓ_1 norm $|\omega|_1$ when $B = [-1, 1]^d$). It follows from (3) that $\bar{f}(x) = f(x) - f(0)$ is expressed as an infinite convex combination of signed indicators of half-spaces times a constant,

$$\bar{f}(x) = v\int_{R^d}\int_0^1 (1_{\{\alpha\cdot x < -t\}} - 1_{\{\alpha\cdot x > t\}})s(\omega, t)p(\omega, t)d\omega dt, \tag{4}$$

where $s = s(\omega, t)$ is $+1$ and -1, respectively, for positive and negative values of the function $\sin(t|\omega|_B + \theta_\omega)$. Here the probability density function is given by

$$p(\omega, t) = \frac{1}{v}|\omega|_B |\sin(t|\omega|_B + \theta_\omega)||\tilde{f}(\omega)|,$$

where the constant is

$$v = \int_{R^d}\int_0^1 |\omega|_B|\sin(t|\omega|_B + \theta_\omega)||\tilde{f}(\omega)|d\omega dt. \quad (5)$$

Consequently, f has bounded variation with respect to half-spaces and

$$V_{\tilde{f}, S_{h,s}, B} \leq 2v \leq 2C_{f,B} \quad (6)$$

where $C_{f,B}$ is the spectral norm defined by

$$C_{f,B} = \int |\omega|_B|\tilde{f}(\omega)|d\omega. \quad (7)$$

It follows that artificial neural networks of the form

$$f_T(x) = \sum_{k=0}^{T} c_k\phi(a_k \cdot x + b_k) + c_0 \quad (8)$$

satisfy the following approximation bounds for all $T \geq 1$

$$||f - f_T||_2 \leq \frac{2C_{f,B}}{T^{1/2}} \quad (9)$$

and

$$||f - f_T||_\infty \leq \gamma_d \frac{C_{f,B}}{T^{1/2}} \quad (10)$$

for some constant γ_d.

The nodes of the network (or terms of the network function) in (8) are assumed to be of the form $\phi(a \cdot x + b)$ where $\phi(z)$ is a fixed bounded function with limits equal to 0 and 1 as $z \to -\infty$ and $z \to \infty$, respectively (taken to be the definition of a sigmoidal function in [1] and [3]). Of particular interest is the choice of a unit step function $\phi(z) = 1_{\{z>0\}}$ for which $\phi(a \cdot x + b)$ becomes the indicator of a half-space. Using the fact that the functions f in Theorem 2 are uniformly continuous, the bounds in (9) and (10) are proven first for the unit step function and then extended to arbitrary sigmoids by taking the magnitudes of a and b to be large.

The proof of Theorem 2 proceeds from the Fourier representation by noting that $f(x) - f(0) = \int (e^{i\omega\cdot x} - 1)\tilde{f}(\omega)d\omega$ and that $(e^{iz} - 1) = i\int_0^z e^{iu}du$ which equals $i\int_0^c 1_{\{z>u\}}e^{iu}du$ when $0 \leq z \leq c$ and equals $-i\int_0^c 1_{\{z<-u\}}e^{iu}du$ when $-c \leq z \leq 0$. Note that only one of these two expressions is positive depending on the sign of z, so it follows that $e^{iz} - 1 = i\int_0^c (1_{\{z>u\}} - 1_{\{z<-u\}})e^{iu}du$. Plugging in $z = \omega \cdot x$ and $c = \sup_B |\omega \cdot x| = |\omega|_B$ and integrating yields

$$f(x) - f(0) = i\int_{R^d}\left(\int_0^{|\omega|_B}(1_{\{\omega\cdot x>u\}} - 1_{\{\omega\cdot x<-u\}})e^{iu}du\right)\tilde{f}(\omega)d\omega. \quad (11)$$

Taking the real part of both sides, changing variables with $u = |\omega|_B t$ for $0 \leq t \leq 1$, and applying Fubini's theorem to exchange the order of the integrals completes the proof of the integral representation (3).

The integral representation shows that the function is in an infinite convex combination of signed indicators of half-spaces times the constant v as given in (4) and (5). A sampling argument as in the proof of Theorem 1, but now drawing the parameters from the density $p(\omega, t)$, shows that $\tilde{f}$ is in the closure of the set of finite linear combinations with a sum of absolute values of coefficients not greater $2v$. This shows that the function is of bounded variation with respect to half-spaces with $V_{\tilde{f},S,B} \leq 2v$. The remaining conclusions of Theorem 2 follow by application of Theorem 1.

We conclude this paper by stating a lower bound on the approximation rate of sigmoidal networks, in the worst case for the classes of functions considered here. As the proof shows, the bounds hold even for functions of a high order of smoothness that are contained among the functions with a bound on the spectral norm $C_{f,B}$. For simplicity we now take B to be the unit ball in R^d. The function ϕ is taken to be any continuously differentiable sigmoid activation function for which the difference between $\phi(z)$ and its limits 0 and 1 is bounded by a polynomial function of $1/|z|$ as $z -> -\infty$ and $z -> \infty$, respectively. (This includes all the commonly used cases.) Given any $C > 0$, let $BV_{S,C}$ be the class of functions with variation with respect to half-spaces bounded by C. For each f in this class, let f_T be a best T term sigmoidal network approximation of the form (8) in the sense that the norm $||f - f_T||_2$ is minimized. The following bound shows that no approximation rate better that $(1/T)^{(1/2)+(1/d)}$ is possible uniformly over the class of functions. Note that for large d this lower bound rate closely matches the upper bound which is

$$\sup_{f\in BV_{S,C}} ||f - f_T||_2 \leq C\left(\frac{1}{T}\right)^{1/2}. \quad (12)$$

Theorem 3 (Lower bound on the sigmoidal network approximation rate.) For each positive ϵ, there is a positive constant $\gamma = \gamma(\epsilon, d)$ such that

$$\sup_{f\in BV_{S,C}} ||f - f_T||_2 \geq \gamma C\left(\frac{1}{T}\right)^{(1/2)+(1/d)+\epsilon}. \quad (13)$$

The proof of Theorem 3 is outlined as follows. Once again the deterministic conclusion is established by using probabilistic reasoning. Let $p > 0$ be an achievable approximation rate, that is $||f - f_T||_2 \leq \gamma C(1/T)^p$ for some positive γ, uniformly over the class of functions. Then by results in [4] the mean squared error of statistical estimates of the function can be bounded. Indeed, let $P_{X,Y}$ be a probability distribution with $P_X = \mu$ concentrated on B, with conditional mean $E(Y|X) = f(X)$, and with the range of Y bounded. Let $(X_i, Y_i)_{i=1}^N$ be a random sample independently drawn from $P_{X,Y}$. Then a sigmoidal net estimator $\hat{f}_{T,N}$ is defined in [4] such that, uniformly over the class of functions,

$$
\begin{aligned}
E\|f-\hat{f}_{T,N}\|^2 &\le 2\|f-f_T\|^2 + C_2\frac{Td}{N}\log N \\
&\le 2\gamma^2 C^2\left(\frac{1}{T}\right)^{2p} + C_2\frac{Td}{N}\log N,
\end{aligned} \tag{14}
$$

for some constant C_2, where $\|\cdot\|$ denotes the $L_2(\mu, B)$ norm. Setting $T = C(N/(d\log N))^{1/(p+1)}$ to achieve the best order in the bound yields, for some constant γ_2 depending on d,

$$
E\|f-\hat{f}_{T,N}\|^2 \le \gamma_2 C^2 \left(\frac{\log N}{N}\right)^{2p/(2p+1)}, \tag{15}
$$

uniformly over the class of functions $BV_{S,C}$. Now as shown in Theorem 2, included among these functions are those with Fourier transform satisfying $\int |\omega||\tilde{f}(\omega)|d\omega \le 2C$. Furthermore, using the reasoning in [3], property (15) (based on the Cauchy-Schwarz inequality), this includes the functions in the Sobolev space W of functions with $\int |\tilde{f}(\omega)|^2(|\omega|^2 + |\omega|^{2s})d\omega \le \gamma_3 C^2$, where $s = d/2 + 1 + \epsilon$, and γ_3 is a positive constant depending on ϵ and d. But lower bounds on the maximum of the mean squared error for arbitrary estimators in such a Sobolev space are known from the theory of nonparametric regression (bounds of the desired type were first obtained by Pinsker [7] and Stone [8], see also, Eubank [9] and Wahba [10]). It follows that, for some positive γ_4 (depending on s and d),

$$
\sup_{f\in BV_{S,C}} E\|f-\hat{f}\|^2 \ge \sup_{f\in W} E\|f-\hat{f}\|^2 \ge \gamma_4 C^2\left(\frac{1}{N}\right)^{2r/(2r+1)} \tag{16}
$$

where $r = s/d$. Comparing (15) and (16) we conclude that p cannot exceed r, which in the present case equals $1/2+1/d+\epsilon/d$. This completes the proof of Theorem 3.

Thus the best sigmoidal net approximation bound for the class of functions has rate between $(1/T)^{1/2}$ and $(1/T)^{(1/2)+(1/d)}$. This rate, which is quite reasonable in high dimensions, is to be contrasted with the disasterous rate $(1/T)^{1/d}$ that is best possible for linear subspace (traditional series type) approximations (see [3]).

We conclude that when d is large, $(1/T)^{1/2}$ characterizes the best approximation rate bound in L_2 and L_∞ for sigmoidal networks with T terms, for the class of functions with bounded variation with respect to half-spaces and for the class of functions with a bound on the spectral norm.

REFERENCES

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

[2] Hornik, K., Stinchcombe, M., and White, H. (1988). Multilayer feedforward networks are universal approximators. *Neural Networks*, **2**, 359-366.

[3] Barron, A. R. (1992). Universal approximation bounds for superpositions of a sigmoidal function. To appear in the *IEEE Transactions on Information Theory.*

[4] Barron, A. R. (1992). Approximation and estimation bounds for artificial neural networks. Invited paper, to appear, special issue of *Machine Learning.* (See also, *Proceedings of the Fourth Workshop on Computational Learning Theory*, 243-249. Morgan Kaufman, San Mateo, California).

[5] Dudley, R. M. (1978). Central limit theorems for empirical measures. *Annals of Probability* **6**, 899-929.

[6] Pollard, D. (1990). *Empirical Process Theory and Applications.* NSF-CBMS Series, **2**, Institute of Mathematical Statistics, Hayward, California.

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

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

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

[10] Wahba, G. (1990). *Spline Models for Observational Data*, SIAM, Philidelphia.