NEURAL NET APPROXIMATION
Andrew R. Barron
University of Illinois at Urbana-Champaign, and
Barron Associates, Inc., Stauntonville, Virginia

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 proof 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 $C_1$ 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 $C_1/2^{T/2}$. The surprising aspect of this approximation bound is that it is the rate $1/2^T$ 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 $1/2^T$. 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 $C_1$ that is not exponentially large in the dimension $d$, assumptions are also made regarding the sampling distributions of the observations of $X_1$, 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 norms 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 $f_s(x)$ and the true function $f(x)$ is bounded by order $C_1/2^T + (T/d)\log N$. 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 $T$ from the data by a complexity-based model selection criterion (related to Shannon's MLI), it is shown in [4] that the mean squared error between $f_s$ 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 $1/2^{T/2}$ are given for the maximum of the error of approximation. That is, the $L_2$ 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 traditional uniform convergence 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 concept involves real-valued functions $f(x)$ of $d$ variables. The input vector $x$ takes values in a bounded set $B$ in $\mathbb{R}^d$, specified for convenience to include the point 0. For instance $B$ may be the cube $[-1,1]^d$. Let $S$ be a class of subsets of $B$. A function $f$ is said to have bounded variation with respect to $S$ if $f$ is in the closure of the set of linear combinations of indicator functions $1_{S}(x)$ for $S \in S$, with the sum of the absolute values of the coefficients of linear combinations not greater than some finite number $V$. The infimum of such $V$ is called the variation of the function $f$ with respect to $S$ and is denoted $\text{Var}_S(f)$. The closure is taken with respect to uniform convergence. Particular interest is given to the case that $S = S_d$, the class of half-spaces $\{ x \cdot \mathbf{a} + \mathbf{b} > 0 \}$ or $\{ x \cdot \mathbf{a} + \mathbf{b} < 0 \}$ 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. [Note that if a function $f(x)$ is continuously differentiable except at a discrete set of jump...
points, \( V_j \) equals \( f_j (\mathbf{x} (t)) \) plus the sum of the jump heights. For \( d > 1, \) \( V_j \) is one of the possible extensions of the notion of bounded variation (another extension would be to use the \( d \)-norm of the \( d \times d \)-matrix spaces. Unlike the one-dimensional case, \( V_j \) is not equiv.

to the \( L_1 \) norm of the gradient for continuously diffeo-

enrable 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 rule is to be played by

the sets of parameters that provide convex of each \( x \) in \( X \). \( S \) is a class of subsets \( S \subseteq \mathbb{R}^d \), parameterized by a vector \( \alpha \) of dimensionality \( d \). Then we have a dual class \( S' \) of subsets parameterized by \( \alpha \) in \( \mathbb{R}^d \). Thus, \( \alpha \in S' \) if and only if \( x \in S_\alpha \), Note that for the n-dimensional subspaces \( X = \{ x : \| x \|_2 \leq 1 \}, \) parameterized by \( u \). \( u \in S \), the dual sets are halfspaces. The same is true for

ellipsoids and hyperbolic regions in which the linear functions \( a \cdot x + b \) are replaced by quadratic polynomials: the dual sets are halfspaces but with a larger dimension \( \ell \). The Yaglik-

Cherkovenski condition for a dual class \( S' \) restricted to \( x \in \mathcal{B} \) is the requirement that for all \( u_1, u_2, \ldots, u_n \) the number of subsets of \( \{ u_1, u_2, \ldots, u_n \} \) obtained by intersecting with \( S_\alpha \) for \( x \in \mathcal{B} \) is strictly less than \( 2^\ell \) for some \( 

\mathcal{T} \). The first \( T \) for which the condition holds is the \( V_0 \)-class dimension \( D \). For halfspaces, \( D = \ell \).

A desired approximation property for functions with bounded variation with respect classes of subsets \( S \) is that given a function \( f \), there exists parameter values \( u_1, u_2, \ldots, u_{\ell} \) such that the approximation \( \sum_{i=1}^{\ell} \chi_{[u_i]}(x) \) has approximation error \( \| f - \sum_{i=1}^{\ell} \chi_{[u_i]}(x) \|_\alpha \) bounded by order \( 1/2 \) uniformly on \( \mathcal{B} \). Thus, the parameterizability property can be proven by a probabilistic argument for classes of subsets for which the dual \( S' \) satisfies the \( V_0 \)-condition.

The idea of the proof is as follows. Given any \( t > 0 \), there is a linear combination of indicators of sets \( S \) with the sum of the absolute values of the coefficients not more than \( V_0 \), such that the maximum of the approximation error is less than \( t \). We take such a linear combination with \( \varepsilon = \sqrt{d} \), where \( \varepsilon \) is an arbitrary positive constant. (The assumption of bounded variation with respect to \( S \) guarantees that this can be done, but it does not as yet place a restriction on the choice of \( S \) or \( \alpha \) 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, respectively. With \( \varepsilon = \sqrt{d} \), we draw \( u_1, u_2, \ldots, u_{\ell} \) at random (with replacement) from the set of \( \alpha \)-positive terms in the linear combination with probabilities proportional to the coeffi-

c\n
cients in this combination. If the class \( S' \) satisfies the Yaglik-

Cherkovenski condition, then by the central limit theorem for empirical processes (due to Dudley [3]), the probability that the maximum for \( x \in \mathcal{B} \) of the difference between the sample av-

\[ f(x) = \frac{1}{\sqrt{d}} \sum_{i=1}^{d} \alpha_i \chi_{[u_i]}(x) \] 

and its expected value is greater than the amount \( \varepsilon^2/2 \) converges as \( \ell \to \infty \) to a probability that is strictly between 0 and 1. This implies that there exists choices for \( u_1, u_2, \ldots, u_{\ell} \) for each large \( \ell \) for

which the maximum difference between the sample average and

its expectation is less than \( (\varepsilon^2/2)^{1/2} \). Doing the same for the negative part and setting \( f_+(x) = f(x) - f_-(x) \), we find by the triangle inequality that \( \mathrm{sup}_{x \in \mathcal{B}} |f(x) - f_+(x)| \leq 2\varepsilon\). Since in particular, the class of halfspaces is a Yaglik-Cherkovenski class, this provides an approximation theorem for artificial neural networks with unit step activation functions. For completeness, we state also as \( L_1 \) bound proved by the method of \( \alpha \) \([5]\) that does not require that \( \mathcal{S} \) be a \( V_0 \)-class, and which holds for all \( \ell \).

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

\[ \| f - f_j \|_\alpha \leq \frac{\varepsilon}{2\sqrt{d}}. \]

(1) \n
where \( \| f - f_j \|_\alpha \) is the \( L_1 \) approximation error with respect to any given probability measure \( \mu \) on \( \mathcal{B} \) (that is, \( \| f - f_j \|_\alpha = \int f(x) - f_j(x) \cdot \mu (dx) \)). \n
(2) \n
(3) \n
(4) \n
Suppose that \( f_{j+1} \) is \( \| f - f_j \|_\alpha \)-uniformly bounded by \( \varepsilon/2 \) uniformly on \( \mathcal{B} \).

Calculations using empirical process bounds from Pollard [5] for

the nonasymptotic case, show that we can take \( \gamma \), not larger

than \( O(d) \). More refined bounds on the constant for the \( L_1 \)

bound are being sought.

The degree of generality of the assumption of bounded variation

with respect to halfspaces 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 \( S \) is related to a spectral norm.

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

\[ f(x) = \frac{1}{\sqrt{d}} \int_{\mathbb{R}^d} e^{i\omega \cdot x} \left( \int_0^1 e^{i\omega \cdot x} e^{-i\omega \cdot x} \right) \, d\omega \]

(3)

(4)

where \( \omega \) denotes the orientation of the frequency vector, \( \omega(x) = \int_0^1 e^{i\omega \cdot x} \, d\omega \) denotes the Fourier magnitude and phase decomposition, and \( |\omega(x)| \) is the \( l_1 \)-norm of \( \omega(x) \) (which equals the \( l_2 \)-norm of \( \omega(x) \) if \( \mathcal{B} = [-1, 1]^d \)). It follows from (3) that\( f(x) = f(x) - f_0 \) is expressed as an infinite convex combination of signed indicators of halfspaces times a constant.

\[ f(x) = \frac{1}{\sqrt{d}} \int_{\mathbb{R}^d} e^{i\omega \cdot x} \left( \int_0^1 e^{i\omega \cdot x} e^{-i\omega \cdot x} \right) \, d\omega \]

(4)
where \( x = s(x,t) = x + 1 \) and \(-1\), respectively, for positive and negative values of the function \( \sin\{\theta(x) + \theta_0\} \). Here the probability density function is given by

\[
\tau(x,t) = \frac{1}{2} \mathbb{E}[\omega(x_t^c) \sin(\omega(x_t + \theta_0))] \mathbb{I}(\omega),
\]

where the constant is

\[
u = \int_0^1 \mathbb{E}[\omega(x_t^c) \sin(\omega(x_t + \theta_0))] \mathbb{I}(\omega) \, dx.
\] (5)

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

\[
Vf|_{x_t^c} \leq 2v \leq 2C_{x,t}B,
\] (6)

where \( C_{x,t} \) is the spectral norm defined by

\[
C_{x,t} = \int |\omega| \mathbb{E}[|\omega|] \, d\omega.
\] (7)

It follows that artificial neural networks of the form

\[
f_{\theta}(x) = \sum_{k=1}^{\infty} \alpha_k \phi_{\theta}(x_k - b_k) + b_0
\] (8)

satisfy the following approximation bounds for all \( T \geq 1 \)

\[
\|f - f_{\theta}\|_2 \leq \frac{2C_{x,t}B}{T^{\frac{1}{2}}},
\] (9)

and

\[
\|f - f_{\theta}\|_{\infty} \leq \frac{C_{x,t}B}{T^{\frac{1}{2}}},
\] (10)

for some constant \( \gamma \).

The nodes of the network (or terms of the network function) in (8) are assumed to be of the form \( \phi(x - z) + b \) where \( \phi(x) \) is a fixed bounded function with limits equal to \( 0 \) and \( 1 \) as \( x \to -\infty \) and \( x \to \infty \), 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 \( \phi(x) = 1_{x > 0} \) for which \( \phi(x - z) + b \) 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 sigmoidals by taking the magnitudes of \( z \) and \( b \) to be large. The proof of Theorem 2 proceeds from the Fourier representation by noting that \( \mathbb{E}[f(x) - f(0)] = \mathbb{E}[\cos(x - t) - 1]f(0) \, dx + \mathbb{E}[\sin(x - t)]f(0) \, dx \) and that \( e^{itx} = 1 + i\int_0^{\infty} e^{ix\omega} \, d\omega \) when \( 0 \leq x \leq \tau \) and equals \(-i\int_0^{\infty} e^{i\omega x} \, d\omega \) when \( -\tau \leq x \leq 0 \). Note that only one of these two expressions is positive depending on the sign of \( z \), so it follows that \( e^{itx} - 1 = i\int_0^{\infty} (1_{x > 0} - 1_{x < 0}) e^{i\omega x} \, d\omega \). Plugging in \( x = u + v \) and \( v = 0 \) to \( e^{itx} \) and integrating yields

\[
f(x) - f(0) = \int_0^{\infty} \int_0^{\infty} \left( 1_{u > v} - 1_{u < v} \right) e^{i\omega x} \, d\omega \, dx.
\] (11)

Taking the real part of both sides, changing variables with \( u = \lambda(v) \) for \( 0 \leq \lambda \leq 1 \), and applying Poisson'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 combinations of signed indicators of half-spaces times the constant \( \nu \) as given in (4) and (5). A sampling argument as in the proof of Theorem 1, but now drawing the parameter from the density \( \rho(x,t) \), shows that \( f \) is in the closure of the set of finite linear combinations with a sum of absolute values of coefficients not greater \( \nu \). This shows that the function is of bounded variation with respect to half-spaces with \( Vf|_{x_t^c} \leq \nu \).

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 higher order of smoothness that are contained among the functions with a bound on the spectral norm \( C_{x,t} \). For simplicity we now take \( B \) to be the unit ball in \( \mathbb{R}^d \). The function \( \phi \) is taken to be any continuously differentiable sigmoidal activation function for which the difference between \( \phi(x) \) and its limit 0 and 1 is bounded by a polynomial function of \( |x| \) as \( x \to -\infty \) and \( x \to \infty \), respectively (This includes all the commonly used cases). Given any \( C > 0 \), let \( \mathcal{W}_{C} \) be the class of functions with variation with respect to half-spaces bounded by \( C \). For each \( f \) in this class, let \( f_{\theta} \) be a best \( T \) term sigmoidal network approximation of the \( f \) in (8) in the sense that the norm \( \|f - f_{\theta}\|_2 \) is minimized. The following bound shows that no approximation rate better than \( O(T^{-\frac{1}{2}}) \) is possible uniformly over the class of functions.

Note that for large \( T \), the lower bound rate closely matches the upper bound which is

\[
\sup_{f \in \mathcal{W}, \theta \in [0,1]} \|f - f_{\theta}\|_2 \leq C \left( \frac{1}{T} \right).
\] (12)

\[
\sup_{f \in \mathcal{W}, \theta \in [0,1]} \|f - f_{\theta}\|_2 \geq \gamma C \left( \frac{1}{T} \right).\] (13)

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

\[
\sup_{f \in \mathcal{W}, \theta \in [0,1]} \|f - f_{\theta}\|_2 \geq \gamma C \left( \frac{1}{T} \right).
\] (14)

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 \( \|f - f_{\theta}\|_2 \leq \gamma C(1/T)^p \) for some positive \( \gamma \), uniformly over the class of functions. Then by results in [4] the mean squared error of statistical estimates of the function \( f \) can be bounded. Indeed, let \( \mathcal{P}_K \) be a probability distribution with \( P_{\mathcal{P}_K} = \mu \) concentrated on \( B \) with conditional mean \( E[\hat{f}(X) | X] = f(X) \), and with the range of \( Y \) bounded. Let \( \{X_i\}_{i=1}^{N_{\mathcal{P}_K}} \) be a random sample independently drawn from \( \mathcal{P}_K \). Then a sigmoidal network estimator \( f_{\mathcal{P}_K} \) is defined in [4] such that, uniformly over the class of
functions,

\[ E[|f - \hat{f}_{u,v}||^2] \leq 2E[|f - \hat{f}_u||^2] + C_1 \frac{T_d}{N} \log N \]

\[ \leq 2\gamma^2 C^2 \left( \frac{1}{T} \right)^{\gamma} + C_2 \frac{T_d}{N} \log N, \]

(14)

for some constant \( C_2 \), where \( \| \cdot \| \) denotes the \( L_2(s,B) \) norm. Setting \( T = C(N/(d \log N))^{1/(d+1)} \) to achieve the best order in the bound yields, for some constant \( \gamma_2 \) depending on \( d \),

\[ E[|f - \hat{f}_{u,v}||^2] \leq \gamma_2 C^2 \left( \frac{\log N}{N} \right)^{2/(1+4d)}, \]

(15)

uniformly over the class of functions \( BV_{p,d} \). Now as shown in Theorem 2, included among these functions are those with Fourier transform satisfying \( \int |\omega|/|\hat{f}(\omega)||d\omega \leq 2C \). 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 \( \int |\omega|^p |\hat{f}(\omega)|^2 + |\omega|^{2p} |d\omega \leq \gamma C_p \), where \( \gamma = 4/2 + 1 + \epsilon \), and \( \gamma_2 \) is a positive constant depending on \( \epsilon \) and \( d \). But lower bounds on the maximum of the mean squared error for arbitrary estimators \( s \) 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 \( \gamma_2 \) depending on \( s \) and \( d \),

\[ \sup_{f \in BV_{p,d}} E[|f - \hat{f}||^2] \geq \sup_{f \in W} E[|f - \hat{f}||^2] \geq \gamma \gamma_2 C^2 \left( \frac{1}{N} \right)^{2/(1+4d)}, \]

(16)

where \( r = s/d \). Comparing (15) and (16) we conclude that \( p \) cannot exceed \( r \), which in the present case equals \( 1/2 + l/d + c/d \).

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 \( (1/T)^{1/2(1+c/d)} \). This rate, which is quite reasonable in high dimensions, is to be contrasted with the disastrous rate \( (1/T)^{1/4} \) that is best possible for linear subspace (traditional series type) approximations (see [2]).

We conclude that when \( s \) is large, \( (1/T)^{1/2} \) characterizes the best approximation rate bound in \( L_2 \) and \( L_\infty \) 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


