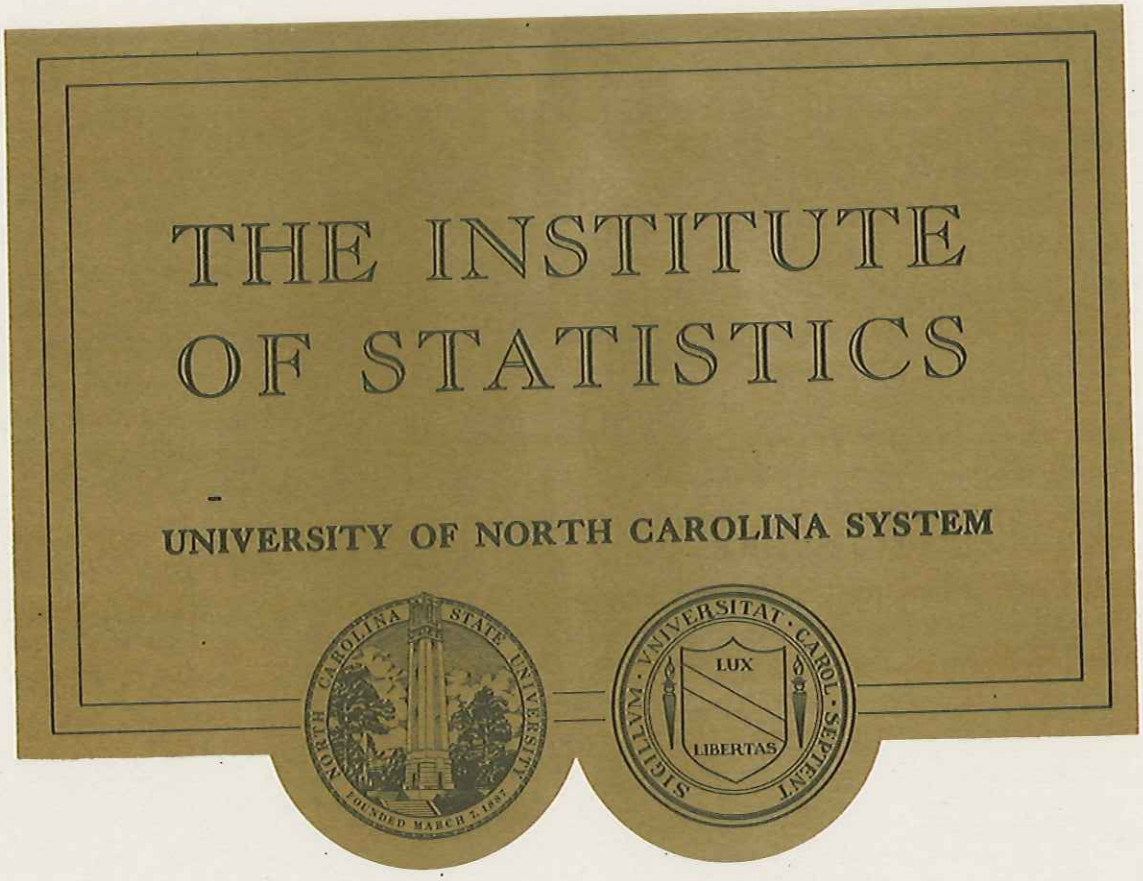
File copy



CONVERGENCE RATES FOR SINGLE HIDDEN LAYER FEEDFORWARD NETWORKS

by

Daniel F. McCaffrey and A. Ronald Gallant

Institute of Statistics Mimeograph Series No. 2207

November 1991

NORTH CAROLINA STATE UNIVERSITY
Raleigh, North Carolina

Convergence Rates For Single Hidden Layer Feedforward Networks

by

Daniel F. McCaffrey
RAND
1700 Main Street
Santa Monica, CA 90407

A. Ronald Gallant
Department of Statistics
North Carolina State University
Raleigh, NC 27695-8203

November 1991

* This research was supported by National Science Foundation Grants SES-8808015. We thank Stephen Ellner and Douglas Nychka for helpful comments throughout. Reprint requests: Daniel F. McCaffrey, RAND, 1700 Main Street, Santa Monica, CA 90407, (310)393-0411.

Running title: Convergence Rates for Networks

Convergence Rates for Feedforward Single Hidden Layer Networks

ABSTRACT

By allowing the training set to become arbitrarily large, appropriately trained and configured single hidden layer feedforward networks converge in probability to the smooth function which they were trained to estimate. A bound on the probabilistic rate of convergence of these network estimates is given. The convergence rate is calculated as a function of the sample size n. If the function being estimated has square integrable mth order partial derivatives then the L_2-norm estimation error approaches $O_p(n^{-1/2})$ for large m. Two steps are required for determining these bounds. A bound on the rate of convergence of approximations to an unknown smooth function by members of a special class of single hidden layer feedforward networks is determined. The class of networks considered can embed Fourier series. Using this fact and results on the approximation properties of Fourier series yields a bound on L_2-norm approximation error. This bound is less than $O(q^{-1/2})$ for approximating a smooth function by networks with q hidden units. A modification of Barron's (1991) results for bounding estimation error provides a general theorem for calculating estimation error convergence rates. Combining this result with the bound on approximation rates yields the final convergence rates.

Key words: Nonparametric Regression, Fourier Series, Embedding

1. Introduction.

Many authors have investigated the universal approximation properties of neural networks (Gallant and White, 1988; Cybenko, 1989; Funahashi, 1989; Hecht-Nielson, 1988; Hornick et al., 1989, 1990; and Stinchcombe and White, 19898). Using a wide variety of proof strategies all have demonstrated that, under general regularity conditions, a sufficiently complex single hidden layer feedforward network can approximate any member of a class of functions to any desired degree of accuracy. The complexity of a single hidden layer feedforward network is measured by the number of hidden units in the hidden layer.

More recently, Barron (1991a) using a result of Jones (1991) has calculated the convergence rate for approximation error. Approximation error is the L_2-norm of the difference between a function and a single hidden layer feedforward network configured to approximate the function. Barron gives his rate in terms of the number of hidden units in the network. If q denotes the number of hidden units, then the approximation error is $O(q^{-1/2})$. This bound applies to all functions satisfying a smoothness condition stated in terms of the integrability of the Fourier transforms of the functions.

Barron (1991b) also determines estimation convergence rates. A network estimator of an unknown function is a single hidden layer feedforward network trained using a finite training set to learn the unknown function. Implicitly, each output of the input-output pairs of the training set is assumed to be smooth function of its corresponding input plus an additional noise term. One unknown function is assumed to have generated all the outputs. White (1989) demonstrated that if network complexity increases with the sample size, n, then

the probability of making a large estimation error can be made arbitrarily small by allowing for an arbitrarily large sample size. Barron furthers White's result by showing that the expected value of the square of the L_2-norm of the difference between an unknown regression function and the network estimate is $O\{n^{-1/2}(\log n)^{1/2}\}$.

To ensure his conversion rate, Barron requires estimation over a lattice of possible connection strengths and assumes that the unknown function belongs to the class of smooth functions discussed above. This paper offers an alternative approach for calculating estimation convergence rates. This alternative approach provides rates which hold for all functions in certain Sobolev function spaces and does not require estimation over a lattice of networks.

First an alternative bound on approximation error is given. This bound holds for approximating functions with square integrable mth order partial derivatives. When m is large this bound is less than $O(q^{-1/2})$. A direct extension of Barron's results for bounding estimation error provides a general theorem for calculating rates when estimating over an entire class of networks rather than a grid or lattice. Combining this result with the alternative bound on approximation rates yields an estimation error rate which approaches $O_p(n^{-1/2})$ when m is large.

The next section provides a detailed description of the networks under consideration and a preliminary discussion of the approach to determining error rates. Sections 3 contains the development of the bound for approximation error. Section 4 provides theoretical results to verify the hypothesis of the basic theorem stated in Section 2. Combining these results with the rate given in

Section 3 yields estimation error convergence rates. Finally, Section 5 contains a discussion of the new results and possible avenues for further research. Proofs of the technical results are contained in a mathematical appendix.

2. Background.

In the most general terms, a neural network uncovers the underlying relationship between an input vector x and a scaler y. For example, in a classification or pattern recognition problem, one uses a vector of descriptive characteristics of an object, i.e. a pattern, and train the network to achieve the correct classification, y. For forecasting problems, the network forecasts the variable y based on the the values of the elements of the vector x. In each example, the trained network provides a mechanism to relate the input x to a target y.

The appropriate statistical model to describe the relationship between x and y is the standard regression model,

$$y = f(x) + e,$$

where f is an unknown function and e is random error. The error term, e, allows for situations where the relationship between x and y is not precise; for example when y is measured with error. In practice, a finite set of x's and y's which are generated according to this model are observed. This data is denoted $\{(y_t,x_t)\}_{t=1}^{n}$ with $x_t \in \mathbb{R}^d$ for a fixed integer d and $y_t \in \mathbb{R}^1$. We will refer to this data either as our sample or because the data is used to train the network, as a training set.

Let $\mathfrak{S}$ denote the range of the x_t. The regression function f is assumed to be an element of the set of all functions, h, which are m-times differentiable with $h \in L_2(\mathfrak{S})$ and all m^{th} order partial derivatives of h are also in $L_2(\mathfrak{S})$. The space $L_2(\mathfrak{S})$ is the set of all functions $h:\mathfrak{S}\rightarrow\mathbb{R}^1$ such that $\int_{\mathfrak{S}} h^2(x)dx < \infty$, coupled with

the norm $\|\cdot\|_{2,\mathfrak{G}}$, where $\|h\|_{2,\mathfrak{G}} = \left\{\int_{\mathfrak{G}} h^2(x)dx\right\}^{1/2}$. This assumption requires f to be smooth.

The following distributional assumptions concerning the x_t's and the e_t's complete the statistical model. The x_t and e_t are realizations of random variables defined on a probability space $(\Omega,\mathcal{F},P)$. The x_t's are independent and identically distributed according to a common distribution function $\mu(x)$ having support $\mathfrak{G}$, which is the open set $\times_{i=1}^{d}(\epsilon, 2\pi-\epsilon) = (\epsilon, 2\pi-\epsilon)^d$, for some $\epsilon > 0$. Furthermore, $\mu(x)$ has a continuous, bounded density function. The e_t's are independent of the x_t's and are symmetric independent identically distributed $\mathcal{P}(e)$ with support $\mathcal{E}$. Each e_t has zero mean and finite pth absolute moment, $p = (2m+d+5)/2$. For details concerning probability spaces, distributions and densities see Ash (1972).

The assumptions on the x_t's are not as prohibitive as they may first appear. Any bounded random variable can be rescaled so that its support is $(\epsilon, 2\pi-\epsilon)^d$ for some $\epsilon > 0$. Therefore, the assumption simply requires bounded x_t's.

Single hidden layer feedforward networks are considered. Given an input vector, x, the output of such a network can be represented by

$$g(x) = \beta_0 + \sum_{i=1}^{q} \beta_i \psi(\gamma_i' x + \gamma_{0i})$$

where ψ is the activation function and the β's and γ's are the connection strengths. For more details on such networks see Rumelhart and McClelland (1986). The activation function is assumed to be the cosine squasher originally defined by Gallant and White (1988):

$$\psi(u) = \begin{cases} 0 & -\infty < u \leq -\frac{\pi}{2} \\ [\cos(u+\frac{3\pi}{2})+1]/2 & -\frac{\pi}{2} \leq u \leq \frac{\pi}{2} \\ 1 & \frac{\pi}{2} \leq u < \infty. \end{cases}$$

Training the network consists of selecting the connection strengths based on the training set and according to a learning rule. If $\hat{\beta}$'s and $\hat{\gamma}$'s represent the value of the connection strengths selected according to the learning rule, then $\hat{g}_n = \hat{\beta}_0 + \sum_{i=1}^{q} \hat{\beta}_i \psi(\hat{\gamma}_i' x + \hat{\gamma}_{0i})$ is the single hidden layer feedforward network estimate of the true regression function f. White (1989) demonstrated that if the learning rule is correctly specified, then the "error" associated with $\hat{g}_n$ can in a probabilistic sense be made arbitrarily small by allowing for an arbitrarily large training set.

In this paper, we specify a learning rule which is similar to White's and determine a probabilistic bound for the rate at which error converges to zero as n grows large. A network trained according to our rule, using the training set $\{(y_t,x_t)\}$, will have connection weights set at the values that minimize $\frac{1}{n}\sum_{t=1}^{n}[y_t - g(x_t)]^2$, where $g(\cdot)$ is the output of a single hidden layer feedforward network. The minimization is taken over all network configurations such that $\sum_{i=0}^{q_n} |\beta_i| < \Delta_n$ and $\sum_{i=1}^{q_n}\sum_{j=0}^{d} |\gamma_{ij}| < \Delta_n q_n$. The set of all such network configurations is denoted by Λ_n. Thus, $\hat{g}_n$ solves the minimization problem:

$$\underset{g \in \Lambda_n}{\text{minimize}}\ E_n s(y,x;g)$$

where $y_t = Y(x_t,e_t) = f(x_t) + e_t$ and $s: \mathcal{S} \times \mathcal{X} \times \Lambda_n \to \mathbb{R}^1$ has the form $s[Y(x,e),x;g]$ with $s(y,x;g) = [y - g(x)]^2$. (For any measurable function $u: \mathcal{S} \times \mathcal{X} \to \mathbb{R}^1$, $E_n u(e,x)$

$= \frac{1}{n}\sum_{t=1}^{d} u(e_t,x_t).$) The function $s(\cdot,\cdot,\cdot)$ is referred as the objective function and the class of functions containing f is denoted by Λ. That is, $\Lambda = \{u:\mathfrak{S}\to\mathbb{R}^1$: $u \in L_2(\mathfrak{S})$ and all m^{th} order partials of u are also elements of $L_2(\mathfrak{S})\}$. White (1989) referred to $\{\Lambda_n\}$ as sieves and $\hat{g}_n$ as a connectionist sieve estimator.

Because $\hat{g}_n$ approximates an unknown function and we desire to provide accurate estimates of f(x) for every $x \in \mathfrak{S}$, we use a global measure of error. Specifically, we use the metric associated with the weighted L_2-norm, $\|\cdot\|_{2,\mu}$. That is the distance between any two functions u and v is defined as $\|u-v\|_{2,\mu} = \left\{\int_{\mathfrak{S}} |u(x)-v(x)|^2 d\mu(x)\right\}^{1/2}$. Thus, $\|f-\hat{g}_n\|_{2,\mu}$ is the estimation error generated by estimating a regression function using a single hidden layer network trained with a finite training set of n observations to minimize $E_n s(y,x;g)$ over all networks in Λ_n.

Let g_n denote any element of Λ_n such that $\|f-g_n\|_{2,\mu} \le \|f-g\|_{2,\mu}$ for all $g \in \Lambda_n$. Approximation error refers to $\|f-g_n\|_{2,\mu}$. Also, we assume that the regression function uniquely solves the minimization problem:

$$\underset{g\in\Lambda}{\text{minimize}}\ E_0 s(y,x;g),$$

where for each measurable $u:\mathcal{S}\times\mathfrak{S}\to\mathbb{R}^1$, $E_0 u(e,x) = \int_{\mathfrak{S}}\int_{\mathcal{S}} u(e,x)d\mathcal{P}(e)d\mu(x)$. If $u:\mathfrak{S}\to\mathbb{R}^1$ is measurable, then $E_n u(x)$ and $E_0 u(x)$ denote $\frac{1}{n}\sum_{t=1}^{n} u(x_t)$ and $\int_{\mathfrak{S}} u(x)d\mu(x)$ respectively.

Calculating the convergence rate requires determining the metric entropy of each estimation space Λ_n. Let A be a subset of a metric space Ξ with metric ρ, and let $\epsilon>0$ be given. A family $U_1,\ldots,U_l$, of subsets of Ξ is called an ϵ-covering of A if the diameter of each U_j does not exceed 2ϵ and if $A \subset \bigcup_{j=1}^{l} U_j$. Let

$N_\epsilon(A)$ denote the minimum value of l such that there exists a family $U_1, \ldots, U_l$ which is an ϵ- covering. For a given positive ϵ, the metric entropy of the set A is denoted by $H(\epsilon,A,\rho)$ and defined as the natural logarithm of $N_\epsilon(A)$. A finite set of points $\xi_1, \ldots, \xi_m$ from Ξ is an ϵ-net for A if

$$\sup_{\xi \in A} \min_{1 \le j \le m} \rho(\xi,\xi_j) < \epsilon.$$

Let m_0 be the smallest value of m for which there exists a set $\xi_1, \ldots, \xi_m$ which is an ϵ-net. Under general conditions (Kolmogorov and Tihomirov (1961), Theorem 5)

$$H(\epsilon,A,\rho) = \log m_0.$$

We will assume that any metric of interest satisfies these conditions and consider the two statements as equivalent definitions of metric entropy.

We have now provided the notation necessary to pose our problem in precise terms. We determine a bound on the rate at which the estimation error of a network trained with the training set $\{(y_t,x_t)\}_{t=1}^n$ converges to zero in probability as n tends towards infinity. The network will be trained to minimize $E_n s(y,x;g)$ over all network configurations which satisfy the constraints of Λ_n and q_n and Δ_n will grow appropriately large along with n. The following theorem establishes our basic result.

Theorem 0. Consider the following conditions:

i. The sample of data $\{y_{tn},x_t\}_{t=1}^n$, satisfies $y_{tn} = f(x_t) + e_{tn}$ for some unknown regression function f. The x_t's are iid random variables with distribution μ. The support of μ is bounded and f is continuous. The e_{tn}'s are a

triangular array of are iid random variables with $E_{0n}e_{1n} = 0$ and $E_{0n}e_{1n}^2 = \sigma_n^2$ and are distributed independently of the x_t's. For any function ξ, define $s(\xi) = s(y,x;\xi) = [y - \xi(x)]^2$, and $r_n(\xi) = E_n s(\xi) - \frac{1}{n}\sum_{t=1}^{n} e_{tn}^2$ and $r_{0n}(\xi) = E_{0n}s(\xi) - \sigma_n^2$. For some estimation space Ξ_n, $\hat{\xi}_n = \underset{\xi \in \Xi_n}{\operatorname{argmin}} E_n s(\xi)$ and $\xi_n = \underset{\xi \in \Xi_n}{\operatorname{argmin}} E_{0n} s(\xi)$.

ii. The support of e_{tn} t=1, . . .,n, and the range of all ξ in Ξ_n are contained in a known interval of length b_n, where $b_n \to \infty$ as $n \to \infty$.

iii. For any sequence of constants $\{\nu_n\}$, there exists a sequences of constants $\{\epsilon_n(\nu_n)\}$ and bounds k_1 and k_2, such that for any $\xi_1, \xi_2 \in \Xi_n$ satisfying $\rho(\xi_1, \xi_2) < \epsilon_n$, implies $|r_{0n}(\xi_1) - r_{0n}(\xi_2)| \le k_2\nu_n$ and $|r_n(\xi_1) - r_n(\xi_2)| \le k_1\nu_n$ a.s..

Under Conditions i-iii:

Given a sequence $\{\nu_n\}$, if $H_n = H(\epsilon_n(\nu_n), \Xi_n, \rho)$, then

$$E_{0n} \| f - \hat{\xi}_n \|_{2,\mu}^2 = O\{ \| f - \xi_n \|_{2,\mu}^2 + \frac{b_n^2 H_n}{n} + \nu_n \}. \qquad \square$$

Note that Theorem 0 is a general result and is not restricted to network estimation. The notation $a_n = O_p(b_n)$ indicates that the ratio of a_n to b_n is bounded in probability. That is given any $\epsilon > 0$ there exists an M_ϵ such that $Pr(|a_n/b_n| > M_\epsilon) < \epsilon$. The notation $a_n = o_p(b_n)$ indicates that the ratio of a_n to b_n converges in probability to zero. That is given any $\epsilon, \delta > 0$ there exists an $n_0(\epsilon,\delta)$ such that $Pr(|a_n/b_n| > \delta) < \epsilon$ for all $n > n_0$.

3. A bound on approximation error.

To use Theorem 0, we begin by studying approximation error. To find a bound for approximation error, we use the results of Gallant and White (1988) which relate our single hidden layer feedforward network to a Fourier series. Using results concerning the best approximation available by a Fourier series we then obtain bounds for approximation error. This technique, however, requires that Λ_n be sufficiently large so as to include all the possible Fourier series we wish to consider. Such restrictions on Λ_n translate to requirements on the relative size of q_n and Δ_n.

This section begins with an introduction into the additional notation necessary to write a multivariate Fourier series. A multi-index k is a vector in $\mathbb{R}^d$ with integral components. Let $\{k_\alpha\}_{\alpha=0}^{\infty}$ be a sequence of multi-indices ordered so that $|k_\alpha| = \sum_{i=1}^{d} |k_{i\alpha}|$ is nondecreasing in α and complete in the sense that all of the multi-indices with $|k_\alpha| = 0,1, \ldots,$ are present in the sequence. A Fourier series of degree K is written as

$$h(x) = a_0 + \sum_{|k_\alpha| \le K} a_\alpha \cos(k_\alpha' x) + b_\alpha \sin(k_\alpha' x).$$

Define $w_i = k_{[i/2]}$, where [t] denotes the integer part of t and $b_i = i \bmod 2$. Because $\sin(k_\alpha' x) = \cos(k_\alpha' x + \frac{\pi}{2})$, we can express a Fourier series of degree K as

$$h(x) = \theta_0 + \sum_{i=1}^{l} \theta_i \cos(w_i' x + b_i \tfrac{\pi}{2}),$$

where l is a function of K and $l \approx K^d$ (Edmunds and Moscatelli, 1977). ($l \approx K^d$ means that there exists constants c_1 and c_2 such that $c_1 l \le K^d \le c_2 l$ or equivalently constants c_1 and c_2 such that $c_1 K^d \le l \le c_2 K^d$.)

Let $Q = [0, 2\pi]^d$. Because $\mathfrak{G} = (\epsilon, 2\pi - \epsilon)^d$ for some $\epsilon > 0$, we have $\bar{\mathfrak{G}} \subset Q$ ($\bar{\mathfrak{G}}$ is the closure of $\mathfrak{G}$). For any $h \in \Lambda$, h can be extended to a function $\bar{h}: Q \to \mathbb{R}$ such that $\bar{h}$ is periodic in each coordinate and $\bar{h}(x) = h(x)$ for every $x \in \mathfrak{G}$. For any $h \in \Lambda$ the Fourier series expansion of $\bar{h}$ is denoted by $T_\infty \bar{h}$ and defined as

$$T_\infty \bar{h} = a_0 + \sum_{\alpha=1}^{\infty} a_\alpha \cos(k_\alpha' x) + b_\alpha \sin(k_\alpha' x)$$

where

$$a_\alpha = (2\pi)^{-\frac{d}{2}} \int_{\mathfrak{G}} \bar{h}(x) \cos(k_\alpha' x) dx$$

and

$$b_\alpha = (2\pi)^{-\frac{d}{2}} \int_{\mathfrak{G}} \bar{h}(x) \sin(k_\alpha' x) dx.$$

Furthermore,

$$T_K \bar{h} = a_0 + \sum_{|k_\alpha| \le K} a_\alpha \cos(k_\alpha' x) + b_\alpha \sin(k_\alpha' x).$$

denotes the partial sum of order K of the Fourier series expansion of $\bar{h}$.

For each $x \in \mathfrak{G}$ let

$$S_\infty h(x) = T_\infty \bar{h}(x)$$

and for any K

$$S_K h(x) = T_K \bar{h}(x).$$

We refer to $S_\infty h$ as the Fourier series expansion of h and $S_K h$ as the Fourier series approximation or expansion of h of degree K. From the above discussion on $T_K h$,

$$S_K h(x) = a_0 + \sum_{|k_\alpha| \le K} a_\alpha \cos(k_\alpha' x) + b_\alpha \sin(k_\alpha' x)$$

or

$$S_K h(x) = \theta_0 + \sum_{i=1}^{l} \theta_i \cos(w_i' x + b_i \frac{\pi}{2})$$

These two equivalent expansions will be used interchangeably throughout the remainder of this paper. See Edmunds and Moscatelli (1977) for details on Fourier series approximation.

The next result provides a rate of convergence for a Fourier series approximation to any function in the function space Λ.

Theorem 1. (Edmunds and Moscatelli, 1977 Theorem 1) If $h \in \Lambda$ then for any $\epsilon \geq 0$,

$$\| h - S_K h \|_{2,\mathfrak{B}} = o(K^{-m+\epsilon}). \qquad \square$$

This theorem indicates that there exists coefficients, θ_i, such that $\sum_{i=0}^{l} \theta_i \cos(w_i' x + b_i \frac{3\pi}{2})$ converges to any function in Λ at a rate which depends only on the order of approximation. The convergence is in terms of $\| \cdot \|_{2,\mathfrak{B}}$ rather than $\| \cdot \|_{2,\mu}$, the norm we are considering. However, the assumptions on the distribution function μ provide a bound for $\| \cdot \|_{2,\mu}$ in terms of $\| \cdot \|_{2,\mathfrak{B}}$. Let ν denote the density of μ, note that by assumption ν is bounded. Let $\mathfrak{K} = \{\sup_{x \in \mathfrak{B}} | \nu(x) | \}^{1/2}$. For any $h \in L_2(\mathfrak{B})$,

$$\begin{aligned} \| h \|_{2,\mu} &= \left\{ \int_{\mathfrak{B}} | h(x) |^2 d\mu(x) \right\}^{\frac{1}{2}} \\ &= \left\{ \int_{\mathfrak{B}} | h(x) |^2 \nu(x) dx \right\}^{\frac{1}{2}} \\ &\leq \mathfrak{K} \left\{ \int_{\mathfrak{B}} | h(x) |^2 dx \right\}^{\frac{1}{2}} \end{aligned}$$

$$\leq \mathcal{K} \| h \|_{2,\mathcal{G}}.$$

Thus, any bounds on $\| \cdot \|_{2,\mathcal{G}}$ can be directly applied to $\| \cdot \|_{2,\mu}$. Therefore, Theorem 1 indicates that for any element of Λ, h, there exist Fourier series of degree K which converge to h in the $\| \cdot \|_{2,\mu}$ at rate of $a_K = o(K^{-m+\epsilon})$ for any $\epsilon > 0$.

The following results demonstrate that, for large sample sizes, the estimation space embeds Fourier series approximations of functions in the parameter space. Combining these embedding result with the convergence rates for Fourier series approximation yields the approximation rates for networks. The first result is the theorem of Gallant and White which establishes that a network with cosine squashing can embed a Fourier series. Their technique is then used to develop a correspondence between single hidden layer network with cosine squashing and Fourier series to compare network complexity to the degree of the Fourier series.

Theorem 2. (Gallant and White, 1988). Let $\psi(\cdot)$ denote the cosine squasher, then for any $h \in \Lambda$ and finite K, there exists a vector of weights $(\beta_0, \beta_1, \gamma_1', \gamma_{10}, \ldots, \beta_q, \gamma_q', \gamma_{q0})'$, $q = q(K)$, such that

$$S_K h(x) = \beta_0 + \sum_{i=1}^{q} \beta_i \psi(\gamma_i' x + \gamma_{i0}),$$

for all $x \in \mathcal{G}$. □

We say the network embeds the Fourier series expansion. The following lemma quantifies the relationship between K and q(K).

Lemma 1. Let $h \in \Lambda$.

a. Let $K > 0$ be finite. For a feedforward single hidden layer network to embed $S_K h$ requires $q = O(K^{d+1})$ hidden units.

b. There exists a constant C, such that if $K = Cq^{1/(d+1)}$ then there exists a feedforward single hidden layer network with q hidden units which embeds $S_K h$. □

Theorem 3 and Lemma 1 derive from the fact that for each term in a Fourier series expansion, $\theta_i \cos(w_i'x + b_i\frac{\pi}{2})$, there exists weights such that $\sum_{j=1}^{m} \beta_j \psi(\gamma_j'x + \gamma_{j0}) = \theta_i \cos(w_i'x + b_i\frac{\pi}{2}) + \text{const.}$, for $m \le 4|w_i|$ and for all $x \in \mathfrak{X}$. Each $\beta_j = 2\theta_i$, each $\gamma_j = w_i$ and each γ_{j0} is a function of $|w_i|$ and b_i. For details see the proofs of Theorem 3 and Lemma 1 in the appendix.

Lemma 1 b. implies that networks in the estimation space Λ_n have sufficiently many hidden units to embed Fourier series expansions of at least degree $K_n = \text{const.}\, q_n^{1/(d+1)}$ for some constant. The correspondence discussed in the previous paragraph, however, demonstrates that embedding a Fourier series in a network may require weights which exceed the bounds of the estimation space, $\sum_{i=0}^{q_n} |\beta_i| \le \Delta_n$ and $\sum_{i=1}^{q_n} \sum_{j=0}^{d} |\gamma_{ij}| \le \Delta_n q_n$. This problem can be overcome by selecting Δ_n judiciously.

First we investigate the requirements on Δ_n so that $\sum_{i=0}^{q_n} |\beta_i| \le \Delta_n$ is not restrictive. By a well known fact, in a Fourier series expansion of a function with m derivatives, a term with $|w_i| = j$ has $\theta_i = o(j^{-m})$. A Fourier series of degree $q_n^{1/(d+1)}$ has terms with $|w_i| = 1,2,\ldots,[q_n^{1/(d+1)}]$. Each term with

$|w_i| = j$, satisfies $|w_i'x| \in [-2\pi j, 2\pi(j+1)]$ and $4j+2$ hidden units are required to construct this term of the Fourier series. Let $l(j)$ denote the number of multi-indices in $\mathbb{R}^d$ which satisfy $|w| \leq j$. As discussed above, there exists a constant c_1 such that $l(j) \leq c_1 j^d$ for all j and therefore there exist less than $c_1 j^d$ units with $|w_i| = j$. (Note that $c_1 \geq 2d+1$ because there exist $2d+1$ indicies w with $|w| \leq 1$.) The previous discussion provides that,

$$\sum_{i=1}^{q_n} |\beta_i| \leq \sum_{j=1}^{[q_n^{1/(d+1)}]} (4j+2)c_1 j^d j^{-m}$$

$$\leq \text{const.} \sum_{j=1}^{[q_n^{1/(d+1)}]} j^{d-m+1}.$$

Also,

$$|\beta_0| = |\theta_0 + \sum_{i=1}^{l} [\theta_i - \theta_i 2(2M_i+1)]|$$

$$= |\theta_0 - \sum_{i=1}^{l} [\theta_i(4M_i+1)]|.$$

Because less than $c_1 j^d$ of the M_i equal j, $j = 1,2,\ldots,[q_n^{1/(d+1)}]$,

$$|\beta_0| \leq |\theta_0| + \text{const.} \sum_{j=1}^{[q_n^{1/(d+1)}]} j^{d-m+1}.$$

Thus, if $m \geq d+3$ then Δ_n converging to infinity at any rate would be sufficient and $\Delta_n > \text{const.}\, q_n^{2/(d+1)}$, for a fixed positive constant, is sufficient if $m \geq d$.

Similar arguments determine a lower bound for Δ_n so that the constraint $\sum_{i=1}^{q_n} \sum_{j=0}^{d} |\gamma_{ij}| \leq \Delta_n q_n$ does not place a restriction on the selection of the coefficients of an embedded Fourier series. In a network configured to embed a Fourier series each γ_i is equal to w_ν for some ν. Reconstructing a term of a Fourier series expansion with $\cos(w_\nu'x + \beta_\nu)$ requires $4|w_\nu| + 2$ hidden units and

fewer than $c_1 j^d$ of the w_ν satisfy $|w_\nu| = j$. Therefore, a network configured to embed a Fourier series of degree $q_n^{1/(d+1)}$ has less than $c_1(4j+2)j^d$ hidden units with $|\gamma_i| = j$, $j=1,2,\ldots,[q_n^{1/(d+1)}]$. This implies that

$$\begin{aligned}
\sum_{i=1}^{q_n}\sum_{j=1}^{d}|\gamma_{ij}| &= \sum_{i=1}^{q_n}|\gamma_i| \\
&= \sum_{j=1}^{[q_n^{1/(d+1)}]}\sum_{|\gamma_i|=j}|\gamma_i| \\
&\le \sum_{j=1}^{[q_n^{1/(d+1)}]} c_1(4j+2)j^d j \\
&\le \sum_{j=1}^{[q_n^{1/(d+1)}]} \text{const.} j^{d+2} \\
&\le \text{const.}\, q_n^{(d+3)/(d+1)}.
\end{aligned}$$

Furthermore, if $\gamma_i = w_\nu$, $|w_\nu| = j$, then $\gamma_{i0} = \frac{\pi}{2} - m_i 2\pi$ or $\gamma_{i0} = -\frac{3\pi}{2} + m_i 2\pi$ for some integer $m_i \in [-j, j]$. Therefore, $|\gamma_{i0}| \le \frac{3\pi}{2} + |m_i| 2\pi$ and

$$\begin{aligned}
\sum_{i=1}^{q_n}|\gamma_{i0}| &\le \sum_{j=1}^{[q_n^{1/(d+1)}]} c_1 j^d \sum_{m=j}^{j}\left(\frac{3\pi}{2} + |m| 2\pi\right) \\
&\le \sum_{j=1}^{[q_n^{1/(d+1)}]} \text{const.} j^{d+2} \\
&\le \text{const.}\, q_n^{(d+3)/(d+1)}.
\end{aligned}$$

This implies that

$$\sum_{i=1}^{q_n}\sum_{j=0}^{d}|\gamma_{ij}| \le \text{const.}\, q_n(q_n^{2/(d+1)}), \text{ for some fixed constant.}$$

Thus, if $m \ge d$, then there exists some constant C such that $\Delta_n > C q_n^{2/(d+1)}$ is sufficiently large so that for any $h \in \Lambda$ the weights of the q_n hidden unit network

configured to embed $S_{K_n}h$, $K_n = \text{const.}\, q_n^{1/(d+1)}$, satisfy the constraints $\sum_{i=0}^{q_n} |\beta_i| \le \Delta_n$ and $\sum_{i=1}^{q_n}\sum_{j=0}^{d} |\gamma_{ij}| \le \Delta_n q_n$ for all large n.

The following theorem summarizes these results.

Theorem 3. Let $m \ge d$, and $q_n, \Delta_n \to \infty$ as $n \to \infty$, such that $q_n^{2/(d+1)} = o(\Delta_n)$. Then there exists a constant c_1 such that if

$$K_n = c_1 q_n^{1/(d+1)}$$

then for any $h \in \Lambda$,

$$S_{K_n} h \in \Lambda_n,$$

for all large n. □

In particular, Theorem 2 holds for the regression function f. Combining Theorem 2 with Theorem 1 yields an upper bound for approximation error.

Theorem 4. Given an unknown function $f \in \Lambda$. If g_n satisfies

$$g_n = \underset{\Lambda_n}{\operatorname{argmin}} \| f - g_n \|_{2,\mu}$$

and if $q_n, \Delta_n \to \infty$ as $n \to \infty$ with

$$q_n^{2/(d+1)} = o(\Delta_n)$$

then for any $\epsilon > 0$,

$$\| f - g_n \|_{2,\mu} \le o(q_n^{-m/(d+1)+\epsilon}) \text{ as } n \to \infty. \quad \square$$

4. Convergence in L_2-Norm

Theorem 0 is applied to $\Xi_n = \Lambda_n$, $\rho(g_1,g_2) = \sup_{x \in \mathcal{S}} |g_1(x) - g_2(x)|$ and training sets $\{(y_t,x_t)\}_{t=1}^n$ to determine a convergence rate for network estimation error. Direct application of Theorem 0 to the training sets described in Section 2 is impossible because the error terms do not satisfy Condition ii. of the theorem. That is the e_t, $t = 1, \ldots,n$, in $y_t = f(x_t) + e_t$ have support $\mathcal{S}$ which may not be contained in an interval known length b_n.

Theorem 0 is therefore is used to find convergence rates for a surrogate network $g_n^{\#}$. The convergence rate for $\hat{g}_n$ is then found by comparison of $\hat{g}_n$ to $g_n^{\#}$. The network $g_n^{\#}$ satisfies

$$g_n^{\#} = \underset{\Lambda_n}{\operatorname{argmin}} \frac{1}{n} \sum_{t=1}^{n} [f(x_t) + e_{tn} - g(x_t)]^2,$$

where e_{tn}, $t = 1, \ldots,n$, is a truncation of e_t. In other words, $e_{tn} = e_t I\{e_t \in I_n\}$ where $I\{e_t \in I_n\}$ denotes the indicator function for the set $\{e_t \in I_n\}$. That is, $I\{e_t \in I_n\} = 1$ if $e_t \in I_n$ and zero otherwise. The interval $I_n = [-\Delta_n, \Delta_n]$ is chosen so that $\{e_{tn}\}_{t=1}^n$ satisfy Conditions i. and ii. of Theorem 0.

Recall that Condition i. requires that for each n: 1) the $e_{tn} = e_t I\{e_t \in I_n\}$ $t = 1, \ldots,n$, be iid; 2) $E_{0n}e_{1n} = 0$, where E_{0n} denotes expectation taken with respect to the distribution induced by truncation; and 3) $E_{0n}e_{1n}^2$ equal a finite constant σ_n^2. Furthermore, Condition ii. requires that the support of e_{tn}, $t = 1, \ldots,n$, be a subset of the interval containing the range of functions in the estimation space Λ_n.

For each $g \in \Lambda_n$ g satisfies

$$\sup_{\mathcal{S}_n} |g(x)| \leq \sum_{i=0}^{q_n} |\beta_i| \leq \Delta_n.$$

Thus, $I_n = [-\Delta_n, \Delta_n]$ satisfies Condition ii. and the value of the associated b_n is $2\Delta_n$. Because the e_t, $t = 1, \ldots, n$, are iid the e_{tn} will be also. By symmetry, $E_0 e_1 I\{e_1 \in I_n\} = 0$ and because $E_{0n} e_{1n}^2 = E_0[e_1 I\{e_1 \in I_n\}]^2 \leq E_0 e_n^2$, the variance condition is immediate. Thus, if $I_n = [-\Delta_n, \Delta_n]$ then $\{e_{tn}\}_{t=1}^n$ satisfy Conditions i. and ii. of Theorem 0.

Let $r_n(g) = \frac{1}{n}\sum_{t=1}^{n}[f(x_t) + e_{tn} - g(x_t)]^2 - \frac{1}{n}\sum_{t=1}^{n} e_{tn}^2$ and $r_{on}(g) = E_{on}[f(x) + e - g(x)]^2 - \sigma_n^2$ where e_{tn}'s are the truncations defined above and E_{on} defines expectation taken with respect to the distribution induced by the truncations. Using Theorem 0 to determine the rate at which $g_n^{\#}$ converges to f requires that r_n and r_{on} satisfy Condition iii. of the theorem.

The following lemma establishes the absolute continuity of r_n and r_{on} with respect to the sup-norm metric.

Lemma 2. If $g_1, g_2 \in \Lambda_n$ with $\sup_{\mathfrak{X}} |g_1(X) - g_2(X)| < \epsilon_n$, then,

$$|r_{on}(g_1) - r_{on}(g_2)| \leq 5\Delta_n\epsilon_n \quad \text{and}$$

$$|r_n(g_1) - r_n(g_2)| \leq 7\Delta_n\epsilon_n \text{ a.s..} \qquad \square$$

Thus, the bounds k_1 and k_2 of Theorem 0 equal 7 and 5 respectively and given any sequence $\{\nu_n\}$ the corresponding sequence $\{\epsilon_n(\nu_n)\}$ satisfies $\epsilon_n(\nu_n) = \nu_n/\Delta_n$, $n = 1,2, \ldots$.

Given a sequence $\{\nu_n\}$, the truncated errors satisfy all conditions of Theorem 0 and therefore,

$$E_{on} \| f - g_n^{\#} \|_{2,\mu}^2 = O\{ \| f - g_n \|_{2,\mu}^2 + \frac{4\Delta_n^2 H[\epsilon_n(\nu_n), \Lambda_n, \rho]}{n} + \nu_n \}.$$

Approximation error is independent of the training set. Thus for any $\epsilon > 0$, $\| f - g_n \|_{2,\mu}^2 = o(q_n^{-2m/(d+1)+\epsilon})$ provided $q_n^{2/(d+1)} = o(\Delta_n)$. The convergence rate for $g_n^{\#}$ is determined by selecting values for ν_n which balance the approximation error against the term involving the metric entropy. The first step to establishing this balance is to calculate a bound for the metric entropy of Λ_n in terms of q_n and Δ_n.

Lemma 3. For any $\epsilon > 0$, the metric entropy for Λ_n (with the metric induced by the sup norm) satisfies

$$H(\epsilon, \Lambda_n, \rho) \leq \log 4 + [q_n(2+d)+1][\log \tfrac{2}{\epsilon} + \log \Delta_n(1+2\pi\Delta_n)] + q_n(d+1)\log q_n. \quad \square$$

Let $\Delta_n = q_n^{2/(d+1)} \log n$, then $q_n^{2/(d+1)}/\Delta_n \to 0$ as $n \to \infty$ and Δ_n satisfies the conditions of Theorem 4. Therefore, $\| f - g_n \|_{2,\mu}^2 = o(q_n^{-2m/(d+1)+\epsilon})$ for any $\epsilon > 0$. Setting $\nu_n = q_n^{-2m/(d+1)}$ implies that $\epsilon_n(\nu_n) = q_n^{-2m/(d+1)}/\Delta_n$. Substituting these values of Δ_n, ν_n and $\epsilon_n(\nu_n)$ into Lemma 4 yields a bound on $H[\epsilon_n(\nu_n), \Lambda_n, \rho]$.

Lemma 4. If $\epsilon_n = q_n^{-2m/(d+1)}/\Delta_n$, $\Delta_n = q_n^{2/(d+1)} \log n$ and $q_n \approx n^{\alpha}$ for some constant α, then

$$H(\epsilon_n, \Lambda_n, \rho) \leq O(q_n \log n). \quad \square$$

A bound on the convergence rate for $g_n^{\#}$ follows immediately from this lemma.

Theorem 5. Let $m \geq d$, $\Delta_n = q_n^{2/(d+1)} \log n$, $q_n = n^{(d+1)/(2m+d+5)}$ and

$$g_n^{\#} = \underset{\Lambda_n}{\operatorname{argmin}} \frac{1}{n} \sum_{t=1}^{n} [f(x_t) + e_{tn} - g(x_t)]^2,$$

where $e_{tn} = e_t I\{e_t \in I_n\}$, $t = 1, \ldots, n$, with the $\{e_t\}_{t=1}^n$ defined as in Section 2 and with the intervals $I_n = [-\Delta_n, \Delta_n]$, then for any $\epsilon > 0$,

$$\| f - g_n^{\#} \|_{2,\mu} = O_p(n^{-m/(2m+d+5)+\epsilon}). \qquad \square$$

The value for $q_n = n^{(d+1)/(2m+d+5)}$ derives from equating the order of the approximation error to the order of the bound for the metric entropy term. That is q_n solves

$$q_n^{-2m/(d+1)} = \frac{\Delta_n^2 q_n \log n}{n} = \frac{q_n^{(d+5)/(d+1)} (\log n)^3}{n}.$$

Because the convergence rate will be dominated by the larger of $q_n^{-2m/(d+1)}$ or $\Delta_n^2 q_n \log n$ the best rate obtains from equating these two expressions. The convergence rate for network estimation error follows as a corollary to Theorem 5.

Corollary 1. Let $m \geq d$, $\Delta_n = q_n^{2/(d+1)} \log n$ and $q_n = n^{(d+1)/(2m+d+5)}$, then for any $\epsilon > 0$

$$\| f - \hat{g}_n \|_{2,\mu}^2 = O_p(n^{-2m/(2m+d+5)+\epsilon}). \qquad \square$$

4. Summary and Directions for Further Research.

The optimal rate for nonparametric regression estimates given by Stone (1982) is $O_p(n^{-2m/(2m+d)})$. Thus, n^{ϵ} term, this approach to calculating convergence rates for neural networks gives a rate which differs from the optimal only by a constant in the denominator of the exponent.

Barron (1991a) calculates a rate of $O_p(n^{-1/2}(\log n)^{1/2})$. In fact, Barron's results can be extended by using Theorem 0 to give this rate even if estimation is not restricted to a grid. Barron's rate hinges on his approximation rate $\| f - g_n \|_{2,\mu}^2 = O(q_n^{-1})$ (Barron, 1991c). This rate holds only when f is in the set of all function with $\int_{\mathbb{R}^d} |\omega| \, |\tilde{f}(\omega)| \, d\omega < C$ for some constant C, where $\tilde{f}$ is the Fourier transform of f. If $m \geq d+1$ is a reasonable assumption, then using the Fourier series bound given in Theorem 3 provides a faster approximation rate. If $2m \geq d+5$ then Theorem 5 yields faster rates and allows the regression function to be in a more general function space. The assumption $2m \geq d+5$ requires that the noise terms, $\{e_t\}$, have finite $3(d+5)/2$ moments. Normal random variable satisfy this constraint. Another special case which satisfies the moment assumption is the case where the support of e_1 is contained in a bounded interval.

Assuming this special case and that the parameter space, Λ, containing f satisfies $\sup_{\Lambda} \sup_{\mathfrak{X}} |g(X)| < \mathfrak{B}$ for some fixed constant $\mathfrak{B}$, then the coefficients of the Fourier series expansion of f will also be bounded. Under this assumption, the set Λ_n of all network satisfying $\sum_{i=0}^{q_n} |\beta_i| < \Delta_n$ and $\sum_{i=1}^{q_n} \sum_{j=0}^{d} |\gamma_{ij}| < \Upsilon_n q_n$ where $\Delta_n \approx \log n$ and $\Upsilon_n \approx q_n^{2/(d+1)}$ would be sufficient to achieve $q_n^{-2m/(d+1)}$ approximation rates. The rates given by Theorem 4 would then be on the order

of $n^{-2m/(2m+d+1)}$ (the n^{ϵ} term). This very special case is the case considered by Barron. Thus with minimal smoothness or relatively small dimension using the Fourier series bound improves Barron's dimensionless bound for finding estimation rates.

When d is large the rates given by Theorem 4 may be slower than $n^{-1/2}$. Under the special case discussed in the preceding paragraph, however, this upper bound cannot be exceed. This demonstrates the true strength of the dimensionless bound. The value of m is unknown, and therefore high dimensional estimation is risky even with traditional nonparametric regression estimators which achieve the optimal rate of $n^{-2m/(2m+d)}$. With neural networks the rate is guaranteed to be at least as fast as $n^{-1/2}$ even for large dimensions. Thus, the result given in Section 4 provide a compliment to Barron's results. Our results ensure fast convergence for smooth functions and relatively small dimensions. Barron's work guarantees reasonable rates for large dimensional problems.

Both Barron's work and Theorem 4 require strong restrictions on the noise terms. These assumption can be relaxed at the price of slower rate. Assuming only a finite eighth moment for the noise terms, and using the uniform convergence results of Severini and Wong (Severini and Wong, 1987) based on Pollard's (Pollard, 1984 Chapter 2) techniques in conjunction with the Fourier series bound McCaffrey (1991) arrives at a convergence rate of $n^{-m/(2m+d+1)}$ (ignoring log terms).

One direction for further research is to improve the currently available bound on approximation error. The obvious goal of such research is to derive a

rate which accounts for both the smoothness of the functions in Λ and the adaptive projective nature of networks. Such a convergence rate would increase with the smoothness of the function without paying a penalty for larger dimension.

One short coming with Theorem 0 is that this theorem only provides rates for error defined terms of the L_2 norm. Convergence rates for stronger norms which ensure convergence of derivatives are also important, see McCaffrey et al. (1990). Therefore, another area for further research is to develop rates for stronger norms either by expanding uniform convergence results or by developing new techniques for determining convergence rates for networks.

Mathematical Appendix.

Theorem 0 is a direct extension of Barron's Convergence Theorem for Complexity Regularization (1991b) and the proof uses several of his arguments.

Proof of Theorem 0. Let Ξ_n' be an ϵ_n-net for Ξ_n with respect to the metric ρ (i.e. Ξ_n' is a finite set and for each $\xi \in \Xi_n$ there exists a $\xi' \in \Xi_n'$ such that $\rho(\xi,\xi') < \epsilon_n$). We use a preliminary result from Craig's derivation of the Berstein Inequality (Craig, 1933). For independent random variables u_i i=1, . . . ,n, which satisfy $|u_i - Eu_i| \le 3h$, $P\{\bar{u} - E\bar{u} \ge \frac{\tau}{n\epsilon} + \frac{n\epsilon \operatorname{var}(\bar{u})}{2(1-c)}\} \le \exp\{-\tau\}$, where $\tau > 0$ and $0 < \epsilon h \le c < 1$.

Let I_n denote the interval of length b_n which contains the support of the e_{tn}, $t = 1, \ldots, n$. Because e_{tn}'s are mean zero random variables contained in an interval of length b_n, $|e_{tn}| \le b_n$, $t = 1, \ldots, n$. Otherwise, the support of each e_{tn} (the interval I_n) would consist of either entirely positive or negative numbers and therefore e_{tn} could not be mean zero. Thus, I_n straddles zero and is of length b_n where $b_n \to \infty$. Because $\mathfrak{X}$ is bounded and the regression function f is continuous, $f(X) \in I_n$ for all $X \in \mathfrak{X}$ and all large n.

$$r_n(\xi) = \frac{1}{n}\sum_{t=1}^{n}\left\{[y_{tn} - \xi(X_t)]^2 - e_{tn}^2\right\} = -\frac{1}{n}\sum_{t=1}^{n} u_t = -\bar{u}, \qquad \text{where} \qquad u_t = e_{tn}^2 - [y_{tn} - \xi(X_t)]^2.$$

$E_{0n} r_n(\xi) = r_{0n}(\xi)$ and the variance of $r_n(\xi)$, $\operatorname{Var}_{0n}[r_n(\xi)]$, satisfies

$$n\mathrm{Var}_{on}[r_n(\xi)] \le \frac{1}{n}\sum_{t=1}^{n} E_{on}\{[y_{tn}-\xi(X_t)]^2 - e_{tn}^2\}^2$$

$$\le \frac{1}{n}\sum_{t=1}^{n} E_{on}\{[f(X_t)-\xi(X_t)]^2 - 2e_{tn}[f(X_t)-\xi(X_t)]\}^2$$

$$\le \frac{1}{n}\sum_{t=1}^{n}\{E_{on}[f(X_t)-\xi(X_t)]^4 - 4E_{on}e_{tn}^2[f(X_t)-\xi(X_t)]^2\}.$$

Note that $[f(X_t)-\xi(X_t)]^2 \le b_n^2$ and $r_{on}(\xi) = E_{on}[y_{tn}-\xi(X_t)]^2 - \sigma_n^2 = E_{on}[f(X_t)-\xi(X_t)]^2$. Thus,

$$n\mathrm{Var}_{on}[r_n(\xi)] \le \frac{1}{n}\sum_{t=1}^{n}\{b_n^2\, E_{on}[f(X_t)-\xi(X_t)]^2 + 4\sigma_n^2 E_{on}[f(X_t)-\xi(X_t)]^2\}.$$

The variance of all random variables restricted to the interval I_n is less than $b_n^2/4$, the variance of a random variable which assumes either of the endpoints of I_n with probability 1/2. Thus, σ_n^2 is less than $b_n^2/4$ and

$$n\mathrm{Var}_{on}[r_n(\xi)] \le 2b_n^2 r_{on}(\xi).$$

Furthermore, $u_t = -2[f(X_t)-\xi(X_t)]e_{tn} - [f(X_t)-\xi(X_t)]^2 \le -3b_n^2$. Thus, $-E_{on}\bar{u} = r_{on}(\xi)$, $n\,\mathrm{Var}_{on}(\bar{u}) \le 2b_n^2 r_{on}(\xi)$ and $|u_i - \bar{u}| \le 3b_n^2 = 3h$. Let $\epsilon = 1/3b_n^2$ and $c = \epsilon h = 1/3$, then for each $\xi \in \Xi_n'$

$$P\{r_{on}(\xi) - r_n(\xi) \ge \frac{\tau}{n\epsilon} + \frac{\epsilon b_n^2 r_{on}(\xi)}{(1-c)}\} \le P\{\bar{u} - E_o\bar{u} \ge \frac{\tau}{n\epsilon} + \frac{n\epsilon\,\mathrm{Var}_{on}(\bar{u})}{2(1-c)}\} \le e^{-\tau}.$$

Let $\tau = H_n + \log(1/\delta)$ and $\alpha = \epsilon b_n^2/(1-c) = 1/2$, then

$$P\{r_{on}(\xi) - r_n(\xi) \ge \frac{H_n 3b_n^2}{n} + \alpha r_{on}(\xi) + \frac{3b_n^2\log(1/\delta)}{n}, \text{ for all } \xi \in \Xi_n'\}$$

$$\le \sum_{\xi\in\Xi_n} P\{r_{on}(\xi) - r_n(\xi) \ge \frac{H_n 3b_n^2}{n} + \alpha r_{on}(\xi) + \frac{3b_n^2\log(1/\delta)}{n}\}$$

$$\le \exp\{-H_n + \log\delta + H_n\}$$

$$\leq \delta.$$

Note that the values of ϵ and c were chosen so that both c and α are less than 1. Any values which satisfy this constraint will suffice.

For any $\xi \in \Xi_n$, there exists $\xi' \in \Xi_n'$ such that $\rho(\xi, \xi') < \epsilon_n$ and therefore by Condition iii, $|r_n(\xi_1) - r_n(\xi_2)| < k_1\nu_n$ a.s. and $|r_{on}(\xi_1) - r_{on}(\xi_2)| < k_2\nu_n$. This implies that

$$\begin{aligned} r_{on}(\xi) - r_n(\xi) &= r_{on}(\xi) - r_{on}(\xi') + r_{on}(\xi') - r_n(\xi') + r_n(\xi') - r_n(\xi) \\ &\leq |r_{on}(\xi) - r_{on}(\xi')| + r_{on}(\xi') - r_n(\xi') + |r_n(\xi') - r_n(\xi)| \\ &< k_2\nu_n + \frac{H_n 3b_n^2}{n} + \alpha r_{on}(\xi') + \frac{3b_n^2 \log(1/\delta)}{n} + k_1\nu_n, \end{aligned}$$

except on a set with probability less than δ. Note that $r_{on}(\xi') \leq r_{on}(\xi) + k_2\nu_n$, so

$$r_{on}(\xi) - r_n(\xi) < k_2\nu_n + \frac{H_n 3b_n^2}{n} + \alpha k_2\nu_n + \alpha r_{on}(\xi) + \frac{3b_n^2 \log(1/\delta)}{n} + k_1\nu_n.$$

This implies that except on a set with probability less than δ

$$r_{on}(\hat{\xi}_n) - r_n(\hat{\xi}_n) < k_2\nu_n + \frac{H_n 3b_n^2}{n} + \alpha k_2\nu_n + \alpha r_{on}(\hat{\xi}_n) + \frac{3b_n^2 \log(1/\delta)}{n} + k_1\nu_n,$$

or

$$(1\text{-}\alpha) r_{on}(\hat{\xi}_n) < r_n(\hat{\xi}_n) + k_2\nu_n + \frac{H_n 3b_n^2}{n} + \alpha k_2\nu_n + \frac{3b_n^2 \log(1/\delta)}{n} + k_1\nu_n$$

and

$$(1\text{-}\alpha) r_{on}(\hat{\xi}_n) < r_n(\xi_n) + k_2\nu_n + \frac{H_n 3b_n^2}{n} + \alpha k_2\nu_n + \frac{3b_n^2 \log(1/\delta)}{n} + k_1\nu_n$$

where $\xi_n = \underset{\xi \in \Xi_n}{\operatorname{argmin}} \| f - \xi \|_{2,\mu}$. This last inequality holds because

$$\hat{\xi}_n = \underset{\xi \in \Xi_n}{\operatorname{argmin}}\, r_n s(\xi).$$

Define $u_t = [y_{tn} - \xi(X_t)]^2 - e_{tn}^2$. Then $r_n(\xi) = \bar{u}$ and $E_{on}\bar{u} = r_{on}$. As with u_t and $\bar{u}$ defined above, $|u_t - \bar{u}| \le 3b_n^2$ and $n\,Var_{on}(\bar{u}) \le 2b_n^2 r_{on}(\xi)$. However, $\bar{u} - E_{on}\bar{u} = r_n(\xi) - r_{on}(\xi)$. Applying the modified Berstein inequality to these u_t's with $\tau = \log(1/\delta)$, we have

$$P\{r_n(\xi_n) - r_{on}(\xi_n) \ge \frac{3b_n^2 \log(1/\delta)}{n} + \alpha r_{on}(\xi_n)\} \le \delta.$$

Thus, except on a set with probability less than 2δ,

$$(1\text{-}\alpha) r_{on}(\hat{\xi}_n) < (1+\alpha) r_{on}(\xi_n) + k_2\nu_n + \frac{H_n 3b_n^2}{n} + \alpha k_2\nu_n + \frac{6b_n^2 \log(1/\delta)}{n} + k_1\nu_n.$$

Let $v = (1\text{-}\alpha) r_{on}(\hat{\xi}_n) - (1+\alpha)[r_{on}(\xi_n) + k_2\nu_n] - \frac{H_n 3b_n^2}{n} - k_1\nu_n$, then

$$P\{v \ge \frac{6b_n^2 \log(1/\delta)}{n}\} \le 2\delta.$$

Let $v' = \max(v, 0)$. Then $P\{v' \ge \frac{6b_n^2 \log(1/\delta)}{n}\} = P\{v \ge \frac{6b_n^2 \log(1/\delta)}{n}\}$ and $E_{on}v \le E_{on}v'$. Let $\delta = \exp\{-nt/6b_n^2\}$, then by an application of Fubinni's Theorem (Ash, 1970 p. 101),

$$E_{on}v \le E_{on}v' \le \int_0^\infty P\{v' \ge t\}dt \le \int_0^\infty 2\exp\{-nt/6b_n^2\}dt = \frac{12b_n^2}{n}.$$

Thus,

$$E_o r_{on}(\hat{\xi}_n) \le (\frac{1+\alpha}{1-\alpha})[r_{on}(\xi_n) + k_2\nu_n] + (\frac{1}{1-\alpha})[\frac{H_n 3bq_n^2}{n} + k_1\nu_n + \frac{12b_n^2}{n}]$$

$$\le \mathcal{K}\{r_{on}(\xi_n) + \nu_n + \frac{H_n b_n^2}{n}\}$$

for some constant $\mathcal{K}$ and all large n. The result follows because $r_{on}(\xi) = \| f - \xi \|_{2,\mu}^2$ for all functions ξ. □

Proof of Theorem 2. Let $h \in \Lambda$ be arbitrary and $K > 0$ be finite. As noted in Section 2,

$$S_K h(x) = \theta_0 + \sum_{i=1}^{l} \theta_i \cos(w_i'x + b_i\tfrac{\pi}{2}),$$

for some coefficients θ_i, $i = 0,1, \ldots, l$ and where $|w_i| \le K$, $b_i = i \bmod 2$ and $l \approx K^d$. The thrust of this proof is to note that with ψ as the cosine squasher

$$\sum_{m=-M}^{M} 2[\psi(-t + \tfrac{\pi}{2} - m2\pi) + \psi(t - \tfrac{3\pi}{2} + m2\pi) - 1]$$

$$= \begin{cases} 0 & -\infty < t < -2\pi M \\ \cos(t) - 1 & -2\pi M \le t \le 2\pi(M+1) \\ 0 & 2\pi(M+1) < t < \infty \end{cases}.$$

See Gallant and White (1988) for details. For any $x \in \mathcal{X}$, $x \in [\epsilon, 2\pi - \epsilon]^d$ and for all $i = 1,2, \ldots, l$ $w_i'x + b_i\tfrac{\pi}{2} \in [-2\pi M_i, 2\pi(M_i+1)]$ for some $M_i \le K$. Therefore

$$\sum_{m=-M_i}^{M_i} 2\theta_i\{\psi[-(w_i'x + b_i\tfrac{\pi}{2}) + \tfrac{\pi}{2} - m2\pi] + \psi[(w_i'x + b_i\tfrac{\pi}{2}) - \tfrac{3\pi}{2} + m2\pi]\}$$

$$= \theta_i \cos(w_i'x + b\tfrac{\pi}{2}) - \theta_i + \theta_i 2(2M_i + 2)$$

or

$$\sum_{k=1}^{4M_i+2} \beta_{ik} \psi(\gamma_{ik}'x + \gamma_{iko}) = \theta_i \cos(w_i'x + b\tfrac{\pi}{2}) - \theta_i + \theta_i 2(2M_i + 2)$$

where $\beta_{ik} = 2\theta_i$, $\gamma_{ik} = w_i(-1)^{k \bmod 2}$ and

$\gamma_{iko} = (b_i + 1)\tfrac{\pi}{2} + 2\pi(-1)^{k \bmod 2}\{-M_i + [\tfrac{k-1}{2}] - (j-1) \bmod 2\}$. For any real number,

t, [t] denotes the integer part of t. Furthermore, if $\beta_0 = \theta_0 + \sum_{i=1}^{l}[\theta_i - \theta_i 2(2M_i+2)]$, then for any $x \in \mathfrak{S}$,

$$S_K h(x) = \beta_0 + \sum_{i=1}^{l}\sum_{k=1}^{4M_i+2} \beta_{ik}\psi(\gamma_{ik}'x + \gamma_{ik0}).$$

Let $q = \sum_{i=1}^{l}(4M_i+2)$ and the result follows. □

Proof of Lemma 1.

a. Theorem 3 provides that there exists weights such that

$$S_K h(x) = \beta_0 + \sum_{i=1}^{l}\sum_{k=1}^{4M_i+2} \beta_{ik}\psi(\gamma_{ik}'x + \gamma_{ik0})$$

where $M_i \le K$ and $l \approx K^d$. Thus, embedding $S_K h$ requires q hidden units, $q = \sum_{i=1}^{l}(4M_i+2) \le 4Kl + 2l = O(K^{d+1})$ for large K.

b. Recall that there exist $l(K)$ terms in a degree K Fourier series and $l(K)$ satisfies $c_1 K \le l(K) \le c_2 K$ for some constants $c_1, c_2 > 0$. The constants c_1 and c_2 do not depend on K and without loss of generality we can assume $c_2 > 1$. Let $\mathcal{C} = \frac{1}{5c_2}$ and $K = \mathcal{C}q^{1/(d+1)}$. As show in the proof of Theorem 3, for each term in $S_K h$, $\cos(w_i'x + b_i\frac{\pi}{2})$, i=1, . . . $l(K)$, $|w_i| \le \frac{1}{5c_2}q^{1/(d+1)}$ and for x restricted to $\mathfrak{S}$, the function $\cos(w_i'x + b_i\frac{\pi}{2})$ can be constructed with no more than $2(\frac{2}{5c_2}q^{1/(d+1)} + 1)$ hidden units. For large q, $2 \le \frac{1}{5c_2}q^{1/(d+1)}$ and

$2(2\frac{1}{5c_2}q^{1/(d+1)}+1) \le \frac{1}{c_2}q^{1/(d+1)}$. Because $K = cq^{1/(d+1)}$, there are at most $c_2(\frac{1}{5c_2}q^{1/(d+1)})^d$ terms in $S_K h$. Thus, embedding $S_K h$ by a network requires less than $c_2(\frac{1}{5c_2}q^{1/(d+1)})^d \times \frac{1}{c_2}q^{1/(d+1)} \le q$ hidden units. In other words, if $K = cq^{1/(d+1)}$ then there exists a vector of weights $(\beta_0,\beta_1,\gamma_1',\gamma_{10}, \cdots ,\beta_q,\gamma_q',\gamma_{q0})'$, such that $S_K h(x) = \beta_0 + \sum_{i=1}^{q} \beta_i \psi(\gamma_i' x + \gamma_{i0})$, for all $x \in \mathfrak{X}$.

Proof of Theorem 3. By Lemma 1b there exists a constant c_1 such that if $K_n = c_1 q_n^{1/(d+1)}$ then there exists a vector of weights $(\beta_0,\beta_1,\gamma_1',\gamma_{10}, \cdots ,\beta_{q_n},\gamma_{q_n}',\gamma_{q_n 0})'$, such that $S_{K_n} h(x) = \beta_0 + \sum_{i=1}^{q_n} \beta_i \psi(\gamma_i' x + \gamma_{i0})$, for all $x \in \mathfrak{X}$. By the discussion following Lemma 1 there also exists a constant c_2 such that if $\Delta_n > c_2 q_n^{2/(d+1)}$ then these weights satisfy the constraints $\sum_{i=0}^{q_n} |\beta_i| \le \Delta_n$ and $\sum_{i=1}^{q_n}\sum_{j=0}^{d} |\gamma_{ij}| \le \Delta_n q_n$ for all large n. Because $q_n^{2/(d+1)} = o(\Delta_n)$, $q_n^{2/(d+1)}/\Delta_n \to 0$ and $c_2 q_n^{2/(d+1)} < \Delta_n$ for all large n. Thus the weights of the network which embeds $S_{K_n} h$ satisfy the constraints of Λ_n for all large n.

Proof of Theorem 4. By Theorem 3 there exists $K_n = O(q_n^{1/(d+1)})$ such that $S_{K_n} f \in \Lambda_n$ for all large n. Because $g_n = \underset{\Lambda_n}{\operatorname{argmin}} \| f - g_n \|_{2,\mu}$ and $S_{K_n} f \in \Lambda_n$,

$$\begin{aligned} \| f - g \|_{2,\mu} &\le \| f - S_{K_n} f \|_{2,\mu} \\ &= o(K_n^{-m+\epsilon}) \end{aligned}$$

$$= O(q_n^{-m/(d+1)+\epsilon})$$

for all $\epsilon > 0$. □

Proof of Lemma 2.

$$|r_{on}(g_1) - r_{on}(g_2)| = |E_{on}[f(X) - g_1(X) + e_{1n}]^2 - E_{on}[f(X) - g_2(X) + e_{1n}]^2|$$

$$= |E_{on}\{[f(X) - g_2(X) + e_{1n}] + [g_1(X) - g_2(X)]\}^2$$

$$- E_{on}[f(X) - g_2(X) + e_{1n}]^2|$$

$$= |2E_{on}[f(X) - g_2(X) + e_{1n}][g_1(X) - g_2(X)]$$

$$- E_{on}[g_1(X) - g_2(X)]^2|$$

$$\leq 2\epsilon_n E_{on}|f(X) - g_2(X)| + \epsilon_n^2.$$

For sufficiently large n, $\sup_{\mathcal{S}}|f(X)| \leq \Delta_n$ and $\epsilon_n \leq \Delta_n$. Furthermore, for any $g \in \Lambda_n$ $\sup_{\mathcal{S}}|g(X)| \leq |\beta_0| + \sum_{i=1}^{q_n} \beta_i \sup_{\mathcal{S}}|\psi(\gamma_i'X + \gamma_{io})| \leq \Delta_n$. This implies that $|r_{on}(g_1) - r_{on}(g_2)| \leq 5\Delta_n\epsilon_n$.

Similarly,

$$|r_n(g_1) - r_n(g_2)| = |\tfrac{1}{n}\sum_{t=1}^{n}\{2[f(X_t) - g_2(X_t) + e_{tn}][g_1(X_t) - g_2(X_t)]$$

$$- [g_1(X) - g_2(X)]^2\}|$$

$$\leq \tfrac{2\epsilon_n}{n}\sum_{t=1}^{n}|f(X_t) - g_2(X_t)| + \tfrac{2\epsilon_n}{n}\sum_{t=1}^{n}|e_{tn}| + \epsilon_n^2.$$

By definition $|e_{tn}| \leq \Delta_n$ a.s. for $t = 1, \ldots, n$, so $\frac{1}{n}\sum_{t=1}^{n} |e_{tn}| \leq \Delta_n$ a.s..

Therefore,

$$|r_n(g_1) - r_n(g_2)| \leq 4\Delta_n\epsilon_n + 2\Delta_n\epsilon_n + \epsilon_n^2 \text{ a.s.}$$

$$\leq 7\Delta_n\epsilon_n \text{ a.s.} \qquad \square$$

Lemma 4 is similar to White's Lemma 4.3 (1990) and many of the details of this proof have been taken from White's proof of that lemma.

Proof of Lemma 3. From the definition of metric entropy, to calculate the metric entropy of Λ_n, $H(\epsilon, \Lambda_n, \rho)$, requires finding an ϵ-net for Λ_n and setting $H(\epsilon, \Lambda_n, \rho)$ equal to natural log of the number elements in this net. Recall that

$$\Lambda_n = \{g:\mathbb{R}^d \to \mathbb{R}^1: \; g(x) = \beta_0 + \sum_{i=1}^{q_n} \beta_i \psi(\gamma_i' x + \gamma_{0i}) \text{ with } \sum_{i=0}^{q_n} |\beta_i| \leq \Delta_n$$
$$\text{and } \sum_{i=1}^{q_n}\sum_{j=0}^{d} |\gamma_{ij}| \leq \Delta_n q_n\}$$

Let

$$B = \{\beta \in \mathbb{R}^{q_n+1}: \; \sum_{i=0}^{q_n} |\beta_i| \leq \Delta_n \},$$

$$G = \{\gamma \in \mathbb{R}^{q_n(d+1)}: \; \sum_{i=1}^{q_n}\sum_{j=0}^{d} |\gamma_{ij}| \leq \Delta_n q_n\} \text{ and}$$

$$D = B \times G,$$

then

$$\Lambda_n = \{g:\mathbb{R}^d \to \mathbb{R}^1: \; g(x) = \beta_0 + \sum_{i=1}^{q_n} \beta_i \psi(\gamma_i' x + \gamma_{i0}) \text{ with}$$

$$\delta = (\beta_0,\beta_1,\gamma_1', \ldots ,\beta_{q_n},\gamma_{q_n}')' \text{ and } \delta \in D\}.$$

Let B_η be an η-net for B in terms of the metric induced by the L_1-norm. If $u \in \mathbb{R}^\nu$ then the L_1-norm of u is $\| u \| = \sum_{i=1}^{\nu} | u_i |$. Similarly let G_η be an η-net for G and $D_\eta = B_\eta \times G_\eta$. D_η is an η-net for D. Let

$$\Lambda_\eta = \{g:\mathbb{R}^d \to \mathbb{R}^1:\ g(x) = \beta_0 + \sum_{i=1}^{q_n} \beta_i \psi(\gamma_i' x + \gamma_{i0}) \text{ with } \delta \in D_\eta\}.$$

We now show that by selecting η correctly Λ_η is an ϵ-net for Λ_n.

Let $\epsilon > 0$ be given and $g \in \Lambda_n$ be arbitrary with $g(x) = \beta_0 + \sum_{i=1}^{q_n} \beta_i \psi(\gamma_i' x + \gamma_{i0})$. Because B_η and G_η are η-nets for B and G respectively there exists β^* and γ^* such that $\sum_{i=0}^{q_n} | \beta_i - \beta_i^* | \le \eta$ and $\sum_{i=1}^{q_n} \sum_{j=0}^{d} | \gamma_{ij} - \gamma_{ij}^* | \le \eta$. Let $\delta^* = (\beta_0^*,\beta_1^*,\gamma_1^{*\prime}, \ldots ,\beta_{q_n}^*,\gamma_{q_n}^*)'$ and $g^*(x) = \beta_0^* + \sum_{i=1}^{q_n} \beta_i^* \psi(\gamma_i^{*\prime} x + \gamma_{i0}^*)$ then $\delta^* \in D_\eta$ and $g^* \in \Lambda_\eta$.

$$\begin{aligned}
\sup_{x \in \mathfrak{X}} | g(x) - g^*(x) | &= \sup_{x \in \mathfrak{X}} | \beta_0 + \sum_{i=1}^{q_n} \beta_i \psi(\gamma_i' x + \gamma_{i0}) - \\
&\qquad \beta_0^* + \sum_{i=1}^{q_n} \beta_i^* \psi(\gamma_i^{*\prime} x + \gamma_{i0}^*) \ | \\
&\le \sup_{x \in \mathfrak{X}} | \beta_0 - \beta_0^* + \sum_{i=1}^{q_n} (\beta_i - \beta_i^*) \psi(\gamma_i' x + \gamma_{i0}) | \\
&\qquad + \sup_{x \in \mathfrak{X}} | \sum_{i=1}^{q_n} \beta_i^* [\psi(\gamma_i' x + \gamma_{i0}) - \psi(\gamma_i^{*\prime} x + \gamma_{i0}^*)] | \\
&\le \sum_{i=0}^{q_n} | \beta_i - \beta_i^* | \\
&\qquad + \sup_{x \in \mathfrak{X}} \sum_{i=1}^{q_n} | \beta_i^* | \, | \psi(\gamma_i' x + \gamma_{i0}) - \psi(\gamma_i^{*\prime} x + \gamma_{i0}^*) | .
\end{aligned}$$

The Mean Value Theorem provides that for any $u, u' \in \mathbb{R}^1$, $u < u'$,

$$|\psi(u) - \psi(u')| \leq \psi'(u^*)\,|u - u'|,$$

where $u^* \in (u, u')$. However, $0 \leq \psi'(u^*) \leq 1$ for all $u^* \in \mathbb{R}^1$ and thus

$$|\psi(u) - \psi(u')| \leq |u - u'|.$$

Therefore,

$$\begin{aligned}
\sup_{x \in \mathfrak{S}} |g(x) - g^*(x)| &\leq \eta + \sup_{x \in \mathfrak{S}} \sum_{i=1}^{q_n} |\beta_i^*|\,|(\gamma_i' x + \gamma_{i0}) - (\gamma_i^{*\prime} x + \gamma_{i0}^*)| \\
&\leq \eta + \Delta_n [\sum_{i=1}^{q_n} |\gamma_{i0} - \gamma_{i0}^*| + \sum_{i=1}^{q_n} \sup_{x \in \mathfrak{S}} |\gamma_i' x - \gamma_i^{*\prime} x|] \\
&\leq \eta + \Delta_n [\sum_{i=1}^{q_n} |\gamma_{i0} - \gamma_{i0}^*| + \sum_{i=1}^{q_n} \sum_{j=1}^{d} \sup_{x \in \mathfrak{S}} |\gamma_{ij} x_j - \gamma_{ij}^* x_j|] \\
&\leq \eta + 2\pi\Delta_n \sum_{i=1}^{q_n} \sum_{j=0}^{d} |\gamma_{ij} - \gamma_{ij}^*| \\
&\leq \eta + 2\pi\Delta_n \eta.
\end{aligned}$$

If $\eta = \epsilon/(1 + 2\pi\Delta_n)$ then $\rho(g, g^*) \leq \epsilon$ and because g was arbitrary Λ_η is an ϵ-net for Λ_n.

To find $H(\epsilon, \Lambda_n, \rho)$ we must determine the number of elements in Λ_η, $\eta = \epsilon/(1 + 2\pi\Delta_n)$. We can determine this value by calculating the number of elements in D_η. Because $D_\eta = B_\eta \times G_\eta$ we have $\#D_\eta = \#B_\eta \cdot \#G_\eta$, where # is the cardinality operator, i.e. for any set A, #A = the number elements in A. Using the results of Kolmogorov and Tihomirov (Theorems V, IX and X, 1961) we have $\#B_\eta \leq 2(2\Delta_n/\eta)^{q_n+1}$ and $\#G_\eta \leq 2(2\Delta_n q_n/\eta)^{q_n(d+1)}$. Thus

$$\log \#D_\eta \le \log[4(2\Delta_n/\eta)^{q_n+1}(2\Delta_n q_n/\eta)^{q_n(d+1)}]$$

$$\le \log 4 + [q_n(d+2)+1]\log(2\Delta_n/\eta) + q_n(d+1)\log q_n$$

$$\le \log 4 + [q_n(d+2)+1][\log(\tfrac{2}{\epsilon}) + \log(\Delta_n + 2\pi\Delta_n^2)] + q_n(d+1)\log q_n.$$

The lemma follows because of the one-to-one correspondence between D_η and Λ_η.□

Proof of Lemma 4. From Lemma 3, if $\beta = d+2$ and $\nu = m/(d+1)$ then

$$H(\frac{\Delta_n}{q_n^{2\nu}}, \Lambda_n, \rho) \le \log 4 + [q_n\beta+1][\log q_n^{2\nu} + \log(1+2\pi\Delta_n)] + q_n(d+1)\log q_n$$

Because $\Delta_n \to \infty$ there exists an n_0 such that $\Delta_n > (1+2\pi)$ for all $n > n_0$. Thus,

$$H(\frac{\Delta_n}{q_n^{2\nu}}, \Lambda_n, \rho) \le \log 4 + [q_n\beta+1][\log q_n^{2\nu} + \log\Delta_n^2] + q_n(d+1)\log q_n$$

for all $n > n_0$. Because $q_n \to \infty$, $\log q_n > \log 4 > 0$ for all large n, so that $\log 4 + q_n(d+1)\log q_n \le (q_n\beta+1)\log q_n$ and $q_n\beta > 1$. Without a loss of generality we can assume this holds for all $n > n_0$. Thus for all $n > n_0$,

$$H(\frac{\Delta_n}{q_n^{2\nu}}, \Lambda_n, \rho) \le [q_n\beta+1][\log\Delta_n^2 + \log q_n^{2\nu} + \log q_n],$$

$$\le 2q_n\beta[\log\Delta_n^2 q_n^{1+2\nu}].$$

The result follows because by the assumption that $\Delta_n = q_n^{2/(d+1)}\log n$ and $q_n \le n^\alpha$ for some constant α.□

Proof of Theorem 5. Under the conditions of the theorem, Theorem 0 holds and

$$E_{on}\|f-g_n^{\#}\|_{2,\mu}^2 = O\{\|f-g_n\|_{2,\mu}^2 + \frac{4\Delta_n^2 H[\epsilon_n(\nu_n),\Lambda_n,\rho]}{n} + \nu_n\}$$

for any sequence of constants $\{\nu_n\}$. Furthermore, $\Delta_n = q_n^{2/(d+1)}\log n$ implies that $q_n^{2/(d+1)} = o(\Delta_n)$. Therefore, Theorem 4 holds and $\|f-g_n\|_{2,\mu}^2 = o(q_n^{-2m/(d+1)+\epsilon})$ for any $\epsilon > 0$. Let $\nu_n = q_n^{-2m/(d+1)}$. Then $q_n = n^{(d+1)/(2m+d+5)}$ and Lemma 5 provide that $H[\epsilon_n(\nu_n),\Lambda_n,\rho] \le O(q_n \log n)$ and $4\Delta_n^2 H[\epsilon_n(\nu_n),\Lambda_n,\rho]/n = O(q_n^{(d+5)/(d+1)}\log n/n)$. Thus,

$$E_{on}\|f-g_n^{\#}\|_{2,\mu}^2 = O\{\max(q_n^{-2m/(d+1)+\epsilon}, \frac{q_n^{(d+5)/(d+1)}\log n}{n})\}.$$

$q_n = n^{(d+1)/(2m+d+5)}$ implies that $q_n^{-2m/(d+1)} \approx q_n^{(d+5)/(d+1)}/n$ and

$$E_{on}\|f-g_n^{\#}\|_{2,\mu}^2 = O\{q_n^{-2m/(d+1)+\epsilon}\} = O\{n^{-2m/(2m+d+5)+\epsilon}\}.$$

Finally, by Chebyshev's inequality $E_{on}\|f-g_n^{\#}\|_{2,\mu}^2 = O\{n^{-2m/(2m+d+5)+\epsilon}\}$ implies that $\|f-g_n^{\#}\|_{2,\mu}^2 = O_p\{n^{-2m/(2m+d+5)+\epsilon}\}$ or $\|f-g_n^{\#}\|_{2,\mu} = O_p\{n^{-m/(2m+d+5)+\epsilon}\}$. □

The following technical lemma is used in the proof of Corollary 1.

Lemma A.1. Under the conditions of Theorem 5, $\Pr\{g_n^{\#} \ne \hat{g}_n\} \to 0$ as $n \to \infty$.

Proof of Lemma A.1. If all the $e_t \in I_n$ then $g_n^{\#} = \hat{g}_n$. Thus, $\bigcap_{t=1}^{n}\{e_t \in I_n\} \subset \{g_n^{\#} = \hat{g}_n\}$ and $\{g_n^{\#} \ne \hat{g}_n\} \subset \bigcup_{t=1}^{n}\{e_t \notin I_n\}$. Because the e_t are iid the $\Pr\{e_t \notin I_n\} = \Pr\{e_1 \notin I_n\} = \Pr\{|e_1| \ge \Delta_n\}$ for $t = 1, \ldots, n$ and $\Pr\{g_n^{\#} \ne \hat{g}_n\} \le$

$\sum_{t=1}^{n} \Pr\{e_t \notin I_n\} = n\Pr\{e_1 \notin I_n\} = n\Pr\{|e_1| \geq \Delta_n\}$. Let $p = (2m+d+5)/2$. Using the definition of q_n and Δ_n given in Theorem 5 yields $\Delta_n = n^{1/p}\log n$ or $n \leq \Delta_n^p$. This yields $nP\{|e_1| \geq \Delta_n\} \leq \Delta_n^p P\{|e_1| \geq \Delta_n\}$. For every P-measurable set B, let $\lambda(B) = \int_B |e_1|^p dP$. By the assumptions on e_1 given in Section 2, λ is a finite measure. Let $B_n = \{|e_1| \geq \Delta_n\}$; because $\Delta_n \to \infty$, $B_n \to B^*$ with $P(B^*) = 0$. Thus, $\lambda(B_n) \to 0$ and

$$0 \geq \lim_{n\to\infty} \int_{B_n} |e_1|^p dP \geq \lim_{n\to\infty} \int_{B_n} \Delta_n^p dP = \lim_{n\to\infty} \Delta_n^p P\{|e_1| \geq \Delta_n\} \geq 0$$

and

$$0 \leq \lim_{n\to\infty} \Pr\{g_n^{\#} \neq \hat{g}_n\} \leq \lim_{n\to\infty} nP\{|e_1| \geq \Delta_n\} \leq \lim_{n\to\infty} \Delta_n^p P\{|e_1| \geq \Delta_n\} = 0$$

□

Proof of Corollary 1. For any $\epsilon > 0$, let $a_n = n^{-m/(2m+d+5)}$. Given any $\delta > 0$, by Theorem 5 there exists an M_δ, such that

$$\Pr\{\|f - g_n^{\#}\|_{2,\mu} > a_n M_\delta\} < \frac{\delta}{2},$$

for all large n. If $\|f - g_n^{\#}\|_{2,\mu} \leq a_n M_\delta$ and $g_n^{\#} = \hat{g}_n$, then $\|f - \hat{g}_n\|_{2,\mu} \leq a_n M_\delta$. Thus, $\{(\|f - g_n^{\#}\|_{2,\mu} \leq a_n M_\delta) \cap (g_n^{\#} = \hat{g}_n)\} \subset \{\|f - g_n^{\#}\|_{2,\mu} \leq a_n M_\delta\}$ and

$$\Pr\{\|f - g_n^{\#}\|_{2,\mu} \leq a_n M_\delta\} \geq \Pr\{(\|f - g_n^{\#}\|_{2,\mu} \leq a_n M_\delta) \cap (g_n^{\#} = \hat{g}_n)\}$$

$$= 1 - \Pr\{(\|f - g_n^{\#}\|_{2,\mu} > a_n M_\delta) \cup (g_n^{\#} \neq \hat{g}_n)\}$$

$$\geq 1 - \Pr\{\|f - g_n^{\#}\|_{2,\mu} > a_n M_\delta\} - \Pr\{g_n^{\#} \neq \hat{g}_n\}.$$

By Theorem 5 $\Pr\{\|f - g_n^{\#}\|_{2,\mu} > a_n M_\delta\} \leq \delta/2$ for all large n and by Lemma A.1 $\Pr\{g_n^{\#} \neq \hat{g}_n\} \leq \delta/2$ for all large n also. $\Pr\{\|f - g_n^{\#}\|_{2,\mu} \leq a_n M_\delta\} \to 1$ as $n \to \infty$ and $\|f - g_n^{\#}\|_{2,\mu} = O_p(a_n)$ by definition. □

5. References

Ash, Robert B. (1972). *Real analysis and probability.* New York: Academic Press, Inc.

Barron, Andrew R. (1991a). Approximation and estimations bounds for artificial neural networks. *Proceedings of the fourth annual workshop on computational learning, University of California, Santa Cruz, August 5-7, 1991.* San Mateo, CA: Morgan Kaufman Publishers, 243-249.

Barron, Andrew R. (1991b). Complexity regularization with application to artificial neural networks. In G. Roussas (Ed.) *Nonparametric functional estimation and related topics* (pp. 561-576). Dordrecht, Netherlands: Klewer Academic Publishers.

Barron, Andrew R. (1991c). Universal approximation bounds for superpositions of a sigmoidal function. University of Illinois at Urbana-Champaign, Department of Statistics, Technical Report #58.

Craig, Cecil C. (1933). On the Tchebychef inequality of Berstein. *The Annals of Mathematical Statistics*, 4, 94-102.

Cybenko, G. (1989). Approximation by superposition of a sigmoidal function. *Mathematics of Control, Signals, and Systems*, 2, 303-314.

Edmunds, D.E. and V.B. Moscatelli (1977). Fourier approximation and

embeddings of Sobolev spaces. *Dissertationes Mathematicae* CXLV.

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

Gallant, A. Ronald, and Halbert White (1988). There exists a neural network that does not make avoidable mistakes. *Proceedings of the second annual IEEE conference on neural networks, San Diego, California* (pp. I.657-I.664). New York: IEEE Press.

Gallant, A. Ronald, and Halbert White (1989). "On Learning the Derivatives of an Unknown Mapping with Multilayer Feedforward Networks," North Carolina State University, Institute of Statistics Mimeograph Series No. 1964.

Hecht-Nielsen, Robert (1989). Theory of the backpropagation neural network. *Proceedings of the international joint conference on neural networks, Washington D.C* (pp. I.593-I.605). New York: IEEE Press.

Hornick, Kurt, Maxwell Stinchcombe, and Halbert White (1989). Multilayer feedforward networks are universal approximators. *Neural Networks*, 2, 359-366.

Hornick, Kurt, Maxwell Stinchcombe, and Halbert White (1990). Universal approximation of an unknown mapping and its derivatives using multilayer

feedforward networks. *Neural Networks*, 3, 551-560.

Jones, Lee K. (1991). A simple lemma on greedy approximation in hilbert space and convergence rates for projection pursuit regression and neural network training. *The Annals of Statistics*, forthcoming.

Kolmogorov, A.N. and V.M. Tihomirov (1961). ϵ-entropy and ϵ-capacity of sets in functional spaces. *American Mathematical Society Translations*, 2, 17, 277-364.

McCaffrey, Daniel F., Stephen P. Ellner, A. Ronald Gallant and Douglas W. Nychka (1992). Estimating the Lyapunov exponent of a chaotic system with nonparametric regression. *Journal of The American Statistical Association*, forthcoming.

McCaffrey, Daniel F. (1991). Estimating lyapunov exponents with nonparametric regression and convergence rates for feedforward single hidden layer networks. Ph.D. Dissertation, North Carolina State University, Department of Statistics.

Pollard, David (1984). *Convergence of stochastic processes.* New York: Springer-Verlag.

Rumelhart, D.E. and J.L. McClelland (1986). *Parallel distributed processing: explorations in the microstructure of cognition*, Vol 1. Cambridge MA: MIT

Press.

Severini, Thomas A. and Wing Hung Wong (1987). Convergence rates of maximum likelihood and related estimation in general parameter spaces. The University of Chicago, Department of Statistics Technical Report No. 207.

Stinchcombe, Maxwell, and Halbert White (1989). Universal approximation using feedforward networks with non-sigmoid hidden layer activation functions. *Proceedings of the international joint conference on neural networks, Washington D.C.* (pp. I.613-I.617). New York: IEEE Press.

Stone, Charles J. (1982). Optimal global rates of convergence for nonparametric regression. *The Annals of Statistics*, 4, 1040-1053.

White, Halbert (1990). Multilayer feedforward networks can learn arbitrary mappings: connectionist nonparametric regression with automatic and semi-automatic determination of network complexity. *Neural Networks*, 3, 535-550.