Universal Approximation Bounds for Superpositions of a Sigmoidal Function

Andrew R. Barron, *Member, IEEE*

***Abstract*— Approximation properties of a class of artificial neural networks are established. It is shown that feedforward networks with one layer of sigmoidal nonlinearities achieve integrated squared error of order $O(1/n)$, where n is the number of nodes. The function approximated is assumed to have a bound on the first moment of the magnitude distribution of the Fourier transform. The nonlinear parameters associated with the sigmoidal nodes, as well as the parameters of linear combination, are adjusted in the approximation. In contrast, it is shown that for series expansions with n terms, in which only the parameters of linear combination are adjusted, the integrated squared approximation error cannot be made smaller than order $1/n^{2/d}$ uniformly for functions satisfying the same smoothness assumption, where d is the dimension of the input to the function. For the class of functions examined here, the approximation rate and the parsimony of the parameterization of the networks are surprisingly advantageous in high-dimensional settings.**

***Index Terms*— Artificial neural networks, approximation of functions, Fourier analysis, Kolmogorov n-widths.**

I. Introduction

APPROXIMATION bounds for a class of artificial neural networks are derived. Continuous functions on compact subsets of $\boldsymbol{R}^d$ can be uniformly well approximated by linear combinations of sigmoidal functions as independently shown by Cybenko [1] and Hornik, Stinchcombe, and White [2]. The purpose of this paper is to examine how the approximation error is related to the number of nodes in the network.

As in [1], we adopt the definition of a sigmoidal function $\phi(z)$ as a bounded measurable function on the real line for which $\phi(z) \to 1$ as $z \to \infty$ and $\phi(z) \to 0$ as $z \to -\infty$. Feedforward neural network models with one layer of sigmoidal units implement functions on $\boldsymbol{R}^d$ of the form

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

parameterized by $a_k \in \boldsymbol{R}^d$ and $b_k, c_k \in \boldsymbol{R}$, where $a \cdot x$ denotes the inner product of vectors in $\boldsymbol{R}^d$. The total number of parameters of the network is $(d+2)n+1$.

Manuscript received February 19, 1991. This work was supported by ONR under Contract N00014-89-J-1811. Material in this paper was presented at the IEEE International Symposium on Information Theory, Budapest, Hungary, June 1991.

The author was with the Department of Statistics, the Department of Electrical and Computer Engineering, the Coordinated Science Laboratory, and the Beckman Institute, University of Illinois at Urbana-Champaign. He is now with the Department of Statistics, Yale University, Box 2179, Yale Station, New Haven, CT 06520.

IEEE Log Number 9206966.

A smoothness property of the function to be approximated is expressed in terms of its Fourier representation. In particular, an average of the norm of the frequency vector weighted by the Fourier magnitude distribution is used to measure the extent to which the function oscillates. In this Introduction, the result is presented in the case that the Fourier distribution has a density that is integrable as well as having a finite first moment. Somewhat greater generality is permitted in the theorem stated and proven in Sections III and IV.

Consider the class of functions f on $\boldsymbol{R}^d$ for which there is a Fourier representation of the form

$$f(x) = \int_{\boldsymbol{R}^d} e^{i\omega \cdot x} \tilde{f}(\omega)\, d\omega, \tag{2}$$

for some complex-valued function $\tilde{f}(\omega)$ for which $\omega \tilde{f}(\omega)$ is integrable, and define

$$C_f = \int_{\boldsymbol{R}^d} |\omega| |\tilde{f}(\omega)|\, d\omega, \tag{3}$$

where $|\omega| = (\omega \cdot \omega)^{1/2}$. For each $C > 0$, let Γ_C be the set of functions f such that $C_f \le C$.

Functions with C_f finite are continuously differentiable on $\boldsymbol{R}^d$ and the gradient of f has the Fourier representation

$$\Delta f(x) = \int e^{i\omega \cdot x} \widetilde{\Delta f}(\omega)\, d\omega, \tag{4}$$

where $\widetilde{\Delta f}(\omega) = i\omega \tilde{f}(\omega)$. Thus, condition (3) may be interpreted as the integrability of the Fourier transform of the gradient of the function f. In Section III, functions are permitted to be defined on domains (such as Boolean functions on $\{0,1\}^d$) for which it does not make sense to refer to differentiability on that domain. Nevertheless, the conditions imposed imply that the function has an extension to $\boldsymbol{R}^d$ with a gradient that possesses an integrable Fourier representation.

The following approximation bound is representative of the results obtained in this paper for approximation by linear combinations of a sigmoidal function. The approximation error is measured by the integrated squared error with respect to an arbitrary probability measure μ on the ball $B_r = \{x: |x| \le r\}$ of radius $r > 0$. The function $\phi(z)$ is an arbitrary fixed sigmoidal function.

Proposition 1: For every function f with C_f finite, and every $n \ge 1$, there exists a linear combination of sigmoidal functions $f_n(x)$ of the form (1), such that

$$\int_{B_r} (f(x) - f_n(x))^2 \mu(dx) \le \frac{c_f'}{n}, \tag{5}$$

where $c_f' = (2rC_f)^2$. For functions in Γ_C, the coefficients of the linear combination in (1) may be restricted to satisfy $\sum_{k=1}^{n} |c_k| \leq 2rC$, and $c_0 = f(0)$.

Extensions of this result are also given to handle Fourier distributions that are not absolutely continuous, to bound the approximation error on arbitrary bounded sets, to restrict the parameters a_k and b_k to be bounded, to handle certain infinite-dimensional cases, and to treat iterative optimization of the network approximation. Examples of functions for which bounds can be obtained for C_f are given in Section IX.

A lower bound on the integrated squared error is given in Section X for approximations by linear combinations of fixed basis functions. For dimensions $d \geq 3$, the bounds demonstrate a striking advantage of adjustable basis functions (such as used in sigmoidal networks) when compared to fixed basis functions for the approximation of functions in Γ_C.

II. Discussion

The approximation bound shows that feedforward networks with one layer of sigmoidal nonlinearities achieve integrated squared error of order $O(1/n)$, where n is the number of nodes, uniformly for functions in the given smoothness class.

A surprising aspect of this result is that the approximation bound of order $O(1/n)$ is achieved using networks with a relatively small number of parameters compared to the exponential number of parameters required by traditional polynomial, spline, and trigonometric expansions. These traditional expansions take a linear combination of a set of fixed basis functions. It is shown in Section X that there is no choice of n fixed basis functions such that linear combinations of them achieve integrated squared approximation error of smaller order than $(1/n)^{(2/d)}$ uniformly for functions in Γ_C, in agreement with the theory of Kolmogorov n-widths for other similar classes of functions (see, e.g., [3, pp. 232–233]). This vanishingly small approximation rate ($2/d$ instead of 1 in the exponent of $1/n$) is a "curse of dimensionality" that does not apply to the methods of approximation advocated here for functions in the given class.

Roughly, the idea behind the proof of the lower bound result is that there are exponentially many orthonormal functions with the same magnitude of the frequency ω. Unless all of these orthonormal functions are used in the fixed basis, there will remain functions in Γ_C that are not well approximated. This problem is avoided by tuning or adapting the parameters of the basis functions to fit the target function as in the case of sigmoidal networks. The idea behind the proof of the upper bound result (Proposition 1) is that if the function has an integrable representation in terms of parameterized basis functions, then a random sample of the parameters of the basis functions from the right distribution leads to an accurate approximation.

Jones [4] has obtained similar approximation properties for linear combinations of sinusoidal functions, where the frequency variables are the nonlinear parameters. The class of functions he examines are those for which $\int |\tilde{f}(\omega)| d\omega$ is bounded, which places less of a restriction on the high-frequency components of the function (but more of a restriction on low-frequency components) than does the integrability of $|\omega||\tilde{f}(\omega)|$. In the course of our proof, it is seen that the integrability of $|\omega||\tilde{f}(\omega)|$ is also sufficient for a linear combination of sinusoidal functions to achieve the $1/n$ approximation rate. Siu and Brunk [5] have obtained similar approximation results for neural networks in the case of Boolean functions on $\{0, 1\}^d$. Independently, they developed similar probabilistic arguments for the existence of accurate approximations in their setting.

It is not surprising that sinusoidal functions are at least as well suited for approximation as are sigmoidal functions, given that the smoothness properties of the function are formulated in terms of the Fourier transform. The sigmoidal functions are studied here not because of any unique qualifications in achieving the desired approximation properties, but rather to answer the question as to what bounds can be obtained for this commonly used class of neural network models.

There are moderately good approximation rate properties in high dimensions for other classes of functions that involve a high degree of smoothness. In particular, for functions with $\int |\tilde{f}(\omega)|^2 |\omega|^{2s} d\omega$ bounded, the best approximation rate for the integrated squared error achievable by traditional basis function expansions using order m^d parameters is of order $O(1/m)^{2s}$ for $m = 1, 2, \cdots$, for instance, see [3] (for polynomial methods m is the degree, and for spline methods m is the number of knots per coordinate). If $s = d/2$ and n is of order m^d, then the approximation rates in the two settings match. However, the exponential number of parameters required for the series methods still prevent their direct use when d is large.

Unlike the condition $\int |\tilde{f}(\omega)|^2 |\omega|^{2s} d\omega < \infty$, which by Parseval's identity is equivalent to the square integrability of all partial derivatives of order s, the condition $\int |\tilde{f}(\omega)||\omega| d\omega < \infty$ is not directly related to a condition on derivatives of the function. It is necessary (but not sufficient) that all first-order partial derivatives be bounded. It is sufficient (but not necessary) that all partial derivatives of order less than or equal to s be square-integrable on R^d, where s is the least integer greater than $1 + d/2$, as shown in example 15 of Section IX. In the case of approximation on a ball of radius r, if the partial derivatives of order s are bounded on $B_{r'}$ for some $r' > r$, then there is a smooth extension of f for which the partial derivatives of order s are square integrable on R^d, thereby permitting the approximation bounds to be applied to this case.

Another class of functions with good approximation properties in moderately high dimensions is the set of functions with a bound on $\int (\partial^{sd} f(x)/\partial x_1^s \cdots \partial x_d^s)^2 dx$ (or equivalent, $\int |\omega_1|^{2s} \cdots |\omega_d|^{2s} |\tilde{f}(\omega)|^2 d\omega$). For this class, an approximation rate of order $O(1/n)^{2s}$ is achieved using $O(n(\log n)^{d-1})$ parameters, corresponding to a special subset of terms in a Fourier expansion (see Korobov [6] and Wahba [7, pp. 145–146]). Nevertheless, the $(\log n)^{d-1}$ factor still rules out practical use of these methods in dimensions of, say, 10 or more.

Thus far in the discussion, attention is focused on the comparison of the rate of convergence. In this respect, methods that adapt the basis functions (such as sigmoidal networks) are shown to be superior in dimensions $d \geq 3$ for the class Γ_C for

any value of C, no matter how large. Now it must be pointed out that the dimension d can also appear indirectly through the constant C_f. Dependence of the constant on d does not affect the convergence rate as an exponent of $1/n$. Nevertheless, if C_f is exponentially large in d, then an exponentially large value of n would be required for C_f^2/n to be small for approximation by sigmoidal networks. If C is exponentially large, then approximation by traditional expansions can be even worse. Indeed, since the lower bound developed in Section X is of the form $C^2/n^{(2/d)}$, a superexponentially large number of terms n would be necessary to obtain a small value of the integrated squared error for some functions in Γ_C.

The constant C_f involves a d-dimensional integral, and it is not surprising that often it can be exponentially large in d. Standard smoothness properties such as the existence of enough bounded derivatives guarantee that C_f is finite (as discussed above), but alone they are not enough to guarantee that C_f is not exponentially large. In Section IX, a large number of examples are provided for which C_f is only moderately large, e.g., $O(d^{1/2})$ or $O(d)$, together with certain closure properties for translation, scaling, linear combination, and composition of functions. Since in engineering and scientific contexts it is not unusual for functions to be built up in this way, the results suggest that Γ_C may be a suitable class for treating many functions that arise in such contexts.

Other classes of functions may ultimately provide better characterizations of the approximation capabilities of artificial neural networks. The class Γ_C is provided as a first step in the direction of identifying those classes of functions for which artificial neural networks provide accurate approximations.

Some improvements to the bound may be possible. Note that there can be more than one extension of a function outside of a bounded set B that possesses a gradient with an integrable transform. Each such extension provides an upper bound for the approximation error. An interesting open question is how to solve for the extension of a function outside of B_r that yields the smallest value for $\int |\omega||\tilde{f}(\omega)|d\omega$.

For small d, the bound $(2rC)^2/n$ on the integrated squared error in Proposition 1 is not the best possible. In particular, for $d = 1$, the best bound for approximation by step functions is $(rC_f/n)^2$ (which can be obtained by standard methods using the fact that, for functions in Γ_C, the absolute value of the derivative is bounded by C). For $d > 1$, it is recently shown in [20] that the rate for sigmoidal networks cannot be better than $(1/n)^{1+(2/d)}$ in the worse case for functions in Γ_C. Note that the gap between the upper and lower bounds on the rates vanishes in the limit of large dimension. Determination of the exact rate for each dimension is an open problem.

The bound in the proposition assumes that μ is a probability measure. More generally, if μ is a measure for which $\mu(B_r)$ is finite, it follows from Proposition 1 that

$$\int_B (f(x) - f_n(x))^2 \mu(dx) \le \frac{c_f'}{n} \mu(B_r). \tag{6}$$

In particular, with the choice of μ equal to Lebesgue measure, the bound is of order $O(1/n)$, which is independent of d, but the constant $\mu(B_r)$ is equal to the volume of the ball in d dimensions, which grows exponentially in d for $r > 1$.

In the case that the function is observed at sites $X_1, X_2, \cdots, X_N$ restricted to B_r, Proposition 1 provides a bound on the training error

$$\frac{1}{N}\sum_{i=1}^{N}(f(X_i) - \hat{f}_n(X_i))^2 \le \frac{c_f'}{n}, \tag{7}$$

where the estimate $\hat{f}_n = \hat{f}_{n,N}$ of the form (1) is chosen to minimize the sum of squared errors (or to achieve a somewhat simpler iterative minimization given in Section VIII). In this case, the integral in Proposition 1 is taken to be with respect to the empirical distribution.

The implications for the generalization capability of sigmoidal networks estimated from data are discussed briefly. There are contributions to the total mean squared error $\int_{B_r}(f - \hat{f}_n)^2 \, d\mu$ from the mean squared error of approximation $\int_{B_r}(f - f_n)^2 \, d\mu$ and the mean squared error of estimation $\int_{B_r}(f_n - \hat{f}_n)^2 \, d\mu$. An index of resolvability provides a bound to the total mean squared error in terms of the approximation error and the model complexity according to a theorem in [8] and [9] (see also [10] for related results). In [11], the approximation result obtained here is used to evaluate this index of resolvability for neural network estimates of functions in Γ, assuming a smoothness condition for the sigmoid. There it is concluded that statistically estimated sigmoidal networks achieve mean squared error bounded by a constant multiple of $C_f^2/n + (nd/N)\log N$. In particular, with $n \sim C_f(N/(d\log N))^{1/2}$, the bound on the mean squared error is a constant times $C_f((d/N)\log N)^{1/2}$. In the theory presented in [11], a bound of the same form is also obtained when the number of units n is not preset as a function of the sample size N, but rather it is optimized from the data by the use of a complexity regularization criterion.

Other relevant work on the statistical estimation of sigmoidal networks is in White [12] and Haussler [13] where metric entropy bounds play a key role in characterizing the estimation error. For these metric entropy calculations and for the complexity bounds in [11], it is assumed that domain bounds are imposed for the parameters of the sigmoidal network. In order that the approximation theory can be combined with such statistical results, the approximation bounds are refined in Section VI under constraints on the magnitudes of the parameter values. The size of the parameter domains for the sigmoids grows with n to preserve the same approximation rate as in the unbounded case.

For the practitioner, the theory provides the guidance to choose the number of variables d, the number of network nodes n, and the sample size N, such that $1/n$ and $(nd/N)\log N$ are small. But there are many other practical issues that must be addressed to successfully estimate network functions in high dimensions. Some of these issues include the iterative search for parameters, the selection of subsets of terms input to each node, the possible selection of higher order terms, and the automatic selection of the number of nodes on the basis of a suitable model selection criterion. See Barron and Barron [14] for an examination of some of these issues and the relationship between neural network methods and other

methods developed in statistics for the approximation and estimation of functions.

After the initial manuscript was distributed to colleagues, the methods and results of this paper have found application to approximation by hinged hyperplanes (Breiman [15]), slide functions (Tibshirani [16]), projection pursuit regression (Zhao [17]), radial basis functions (Girosi and Anzellotti [18]), and the convergence rate for neural net classification error (Farago and Lugosi [19]). Moreover, the results have been refined to give approximation bounds for network approximation in L_p norms, $1 < p < \infty$ (Darken *et al.* [31], in the L^∞ norm (Barron [20], Yukich [32]) and in Sobolev norms (Hornick *et al.* [21]).

Approximation rates for the sigmoidal networks have recently been developed in McGaffrey and Gallant [22], Mhaskar and Micchelli [23], and Kůrková [33] in the settings of more traditional smoothness classes that are subject to the curse of dimensionality. Reference [22] also gives implications for statistical convergence rates of neural networks in these settings. Jones [24] gives convergence rates and a set of "good weights" to use in the estimation of almost periodic functions. Zhao [17] gives conditions such that uniformly distributed weight directions are sufficient for accurate approximation.

A challenging problem for network estimation is the optimization of the parameters in high-dimensional settings. In Section VIII, a key lemma due to Jones [4] is presented that permits the parameters of the network to be optimized one node at a time, while still achieving the approximation bound of Proposition 1. This result considerably reduces the computational task of the parameter search. Nevertheless, it is not known whether there is a computational algorithm that can be proven to produce accurate estimates in polynomial time as a function of the number of variables for the class of functions studied here. We have avoided the effects of the curse of dimensionality in terms of the accuracy of approximation but not in terms of computational complexity.

III. Context and Statement of the Theorem

In this section, classes of functions are defined and then the main result is stated for the approximation of these functions by sigmoidal networks. The context of Fourier distribution permits both series and integral cases. A number of interesting examples make use of the Fourier distribution, as will be seen in Sections VII and IX.

The Fourier distribution of a function $f(x)$ on $\boldsymbol{R}^d$ is a unique complex-valued measure $\tilde{F}(d\omega) = e^{i\theta(\omega)}F(d\omega)$, where $F(d\omega)$ denotes the magnitude distribution and $\theta(\omega)$ denotes the phase at the frequency ω, such that

$$f(x) = \int e^{i\omega\cdot x}\tilde{F}(d\omega), \tag{8}$$

or, more generally,

$$f(x) = f(0) + \int (e^{i\omega\cdot x} - 1)\tilde{F}(d\omega), \tag{9}$$

for all $x \in \boldsymbol{R}^d$. If $\int F(d\omega)$ is finite, then both (8) and (9) are valid and (9) follows from (8). Assuming only that $\int |\omega|F(d\omega)$ is finite, (9) is used instead of (8) since then the required integrability follows from $|e^{i\omega\cdot x} - 1| \le 2|\omega\cdot x| \le 2|\omega||x|$. (See the Appendix for the characterization of the Fourier representation in this context.) The class of functions on $\boldsymbol{R}^d$ for which $C_f = \int |\omega|F(d\omega)$ is finite is denoted by Γ.

Functions are approximated on bounded measurable subsets of their domain in $\boldsymbol{R}^d$. Let B be a bounded set in $\boldsymbol{R}^d$ that contains the point $x = 0$, and let Γ_B be the set of functions f on B for which the representation (9) holds for $x \in B$ for some complex-valued measure $\tilde{F}(d\omega)$ for which $\int |\omega|F(d\omega)$ is finite, where F is the magnitude distribution corresponding to $\tilde{F}$. (The right side of (9) then defines an extension of the function f from B to $\boldsymbol{R}^d$ that is contained in Γ, and $\tilde{F}$ may be interpreted as the Fourier distribution of an a continuously differentiable extension of f from B to $\boldsymbol{R}^d$. Each such extension provides a possible Fourier representation of f on B.)

Linear functions $f(x) = a\cdot x$ and, more generally, the class of infinitely differentiable functions on $\boldsymbol{R}^d$ are not contained in Γ, but they are contained in Γ_B when restricted to any bounded set B (because such functions can be modified outside of the bounded set to produce a function in Γ; see Section IX). The set of functions on $\boldsymbol{R}^d$ with this property of containment in Γ_B for every bounded set of B is denoted for convenience by $\Gamma_* = \cap_B\Gamma_B$.

For each $C > 0$, let $\Gamma_{C,B}$ be the set of all functions f in Γ_B such that for some $\tilde{F}$ representing f on B,

$$\int |\omega|_B F(d\omega) \le C, \tag{10}$$

where $|\omega|_B = \sup_{x\in B}|\omega\cdot x|$. In the case of the ball $B_r = \{x: |x| \le r\}$, this norm simplifies to $|\omega|_{B_r} = r|\omega|$. (See Section V for the form of the bound for certain other domains such as cubes.)

The main theorem follows; a bound is given for the integrated squared error for approximation by linear combinations of a sigmoidal function.

Theorem 1: For every function f in $\Gamma_{B,C}$, every sigmoidal function ϕ, every probability measure μ, and every $n \ge 1$, there exists a linear combination of sigmoidal functions $f_n(x)$ of the form (1), such that

$$\int_B (f(x) - f_n(x))^2\mu(dx) \le \frac{(2C)^2}{n}. \tag{11}$$

The coefficients of the linear combination in (1) may be restricted to satisfy $\sum_{k=1}^n |c_k| \le 2C$, and $c_0 = f(0)$.

IV. Analysis

Denote the set of bounded multiples of a sigmoidal function composed with linear functions by

$$G_\phi = \{\gamma\phi(a\cdot x + b): |\gamma| \le 2C, \quad a \in \boldsymbol{R}^d, \quad b \in R\}. \tag{12}$$

For functions f in $\Gamma_{C,B}$, Theorem 1 bounds the error in the approximation of the function $\overline{f}(x) = f(x) - f(0)$ by convex combinations of functions in the set G_ϕ.

Proof of Theorem 1: The proof of Theorem 1 is based on the following fact about convex combinations in a Hilbert space, which is attributed to Maurey in Pisier [25]. We denote the norm of the Hilbert space by $\|\cdot\|$.

Lemma 1: If $\overline{f}$ is in the closure of the convex hull of a set G in a Hilbert space, with $\|g\| \le b$ for each $g \in G$, then for every $n \ge 1$, and every $c' > b^2 - \|\overline{f}\|^2$, there is an f_n in the convex hull of n points in G such that

$$\left\|\overline{f} - f_n\right\|^2 \le \frac{c'}{n}. \tag{13}$$

Proof: A proof of this lemma by use of an iterative approximation, in which the points of the convex combination are optimized one at a time, is due to Jones [4]. A slight refinement of his iterative Hilbert space approximation theorem is in Section VIII. The noniterative proof of Lemma 1 (credited to Maurey) is based on a law of large numbers bound as follows. Given $n \ge 1$ and $\delta > 0$, let f^* be a point in the convex hull of G with $\|\overline{f} - f^*\| \le \delta/n$. Thus, f^* is of the form $\sum_{k=1}^m \gamma_k g_k^*$ with $g_k^* \in G$, $\gamma_k \ge 0, \sum_{k=1}^m \gamma_k = 1$, for some sufficiently large m. Let g be randomly drawn from the set $\{g_1^*, \cdots, g_m^*\}$ with $P\{g = g_k^*\} = \gamma_k$; let $g_1, g_2, \cdots, g_n$ be independently drawn from the same distribution as g; and let $f_n = (1/n)\sum_{i=1}^n g_i$ be the sample average. Then $Ef_n = f^*$, and the expected value of the squared norm of the error is $E\|f_n - f^*\|^2 = (1/n)E\|g - f^*\|^2$, which equals $(1/n)(E\|g\|^2 - \|f^*\|^2)$ and is bounded by $(1/n)(b^2 - \|f^*\|^2)$. Since the expected value is bounded in this way, there must exist $g_1, g_2, \cdots, g_n$ for which $\|f_n - f^*\|^2 \le (1/n)(b^2 - \|f^*\|^2)$. Using the triangle inequality and $\|\overline{f} - f^*\| \le \delta/n$, the proof of Lemma 1 is completed by the choice of a sufficiently small δ. □

Fix a bounded measurable set B that contains the point $x = 0$ and a positive constant C. If it is shown that for functions in the class $\Gamma_{C,B}$, the function $\overline{f}(x) = f(x) - f(0)$ is in the closure of the convex hull of G_ϕ in $L_2(\mu, B)$, then it will follow by Lemma 1 that there exists a convex combination of n sigmoidal functions such that the square of the $L_2(\mu, B)$ norm is bounded by a constant divided by n. Therefore, the main task is to demonstrate the following theorem.

Theorem 2: For every function f in $\Gamma_{C,B}$, and every sigmoidal function ϕ, the function $f(x) - f(0)$ is in the closure of the convex hull of G_ϕ, where the closure is taken in $L_2(\mu, B)$).

The method used here to prove Theorem 2 is motivated by the techniques used in Jones [4] to prove convergence rate results for projection pursuit approximation, and in Jones [26] to prove the denseness property of sigmoidal networks in the space of continuous functions.

Proof: Let $\tilde{F}(d\omega) = e^{i\theta(\omega)}F(d\omega)$ denote the magnitude and phase decomposition in the Fourier representation of an extension of the function f on B for which $\int |\omega|_B F(d\omega) \le C$. Let $\Omega = \{\omega \in \boldsymbol{R}^d : \omega \ne 0\}$.

From the Fourier representation (9) and the fact that $f(x)$ is real-valued, it follows that

$$\begin{aligned} f(x) - f(0) &= \mathrm{Re}\int (e^{i\omega\cdot x} - 1)\tilde{F}(d\omega) \\ &= \mathrm{Re}\int_\Omega (e^{i\omega\cdot x} - 1)e^{i\theta(\omega)}F(d\omega) \\ &= \int_\Omega (\cos(\omega\cdot x + \theta(\omega)) - \cos(\theta(\omega)))F(d\omega) \\ &= \int_\Omega \frac{C_{f,B}}{|\omega|_B}(\cos(\omega\cdot x + \theta(\omega)) \\ &\quad - \cos(\theta(\omega)))\Lambda(d\omega) \\ &= \int_\Omega g(x,\omega)\Lambda(d\omega), \end{aligned} \tag{14}$$

for $x \in B$, where $C_{f,B} = \int |\omega|_B F(d\omega) \le C$ is the integral assumed to be bounded; $\Lambda(d\omega) = |\omega|_B F(d\omega)/C_{f,B}$ is a probability distribution; $|\omega|_B = \sup_{x\in B}|\omega\cdot x|$; and

$$g(x,\omega) = \frac{C_{f,B}}{|\omega|_B}(\cos(\omega\cdot x + \theta(\omega)) - \cos(\theta(\omega))). \tag{15}$$

Note that these functions are bounded by $|g(x,\omega)| \le C|\omega\cdot x|/|\omega|_B \le C$ for x in B and $\omega \ne 0$.

The integral in (14) represents $\overline{f}$ as an infinite convex combination of functions in the class

$$G_{\cos} = \left\{\frac{\gamma}{|\omega|_B}(\cos(\omega\cdot x + b) - \cos(b)) : \omega \ne 0, \quad |\gamma| \le C, \quad b \in \boldsymbol{R}\right\}. \tag{16}$$

It follows that $\overline{f}$ is in the closure of the convex hull of $G_{\cos}$. This can be seen by Riemann–Stieltjes integration theory in the case that F has a continuous density function on $\boldsymbol{R}^d$. More generally, it follows form an L_2 law of large numbers. Indeed, if $\omega_1, \omega_2, \cdots, \omega_n$ is a random sample of n points, independently drawn from the distribution Λ, then by Fubini's Theorem the expected square of the $L_2(\mu, B_r)$ norm is

$$\begin{aligned} &E\int_{B_r}\left(f(x) - \frac{1}{n}\sum_{i=1}^n g(x,\omega_i)\right)^2\mu(dx) \\ &= \int_{B_r} E\left(f(x) - \frac{1}{n}\sum_{i=1}^n g(x,\omega_i)\right)^2\mu(dx) \\ &= \frac{1}{n}\int_{B_r}\mathrm{var}(g(x,\omega))\mu(dx) \\ &\le \frac{C^2}{n}. \end{aligned} \tag{17}$$

Thus, the mean value of the squared $L_2(\mu, B)$ norm of a convex combination of n points in $G_{\cos}$ converges to zero as $n \to \infty$. (Note that it converges at rate $O(1/n)$ in accordance with Lemma 1.) Therefore, there exists a sequence of convex combinations of points in $G_{\cos}$ that converges to $\overline{f}$ in $L_2(\mu, B)$. We have proven the following.

Lemma 2: For each f in $L_{C,B}$, the function $f(x) - f(0)$ is in the closure of the convex hull of $G_{\cos}$.

Next it is shown that functions in $G_{\cos}$ are in the closure of the convex hull of G_ϕ. The case that ϕ is the unit step function is treated first.

Each function in $G_{\cos}$ is the composition of a one-dimensional sinusoidal function $g(z) = \gamma/|\omega|_B(\cos(|\omega|_B z + b) - \cos(b))$ and a linear function $z = \alpha\cdot x$, where $\alpha = \omega/|\omega|_B$ for some $\omega \ne 0$. For x in B, the variable $z = \alpha\cdot x$ takes values in a subset of $[-1, 1]$. Therefore, it suffices to examine the

approximation of the sinusoidal function g on $[-1, 1]$. Note that g has derivative bounded by $|\gamma| \leq C$. Now since g is uniformly continuous on $[-1, 1]$, it follows that it is uniformly well approximated by piecewise constant functions, for any sequence of partitions of $[-1, 1]$ into intervals of maximum width tending to zero. Such piecewise constant functions may be represented as linear combinations of unit step functions. Moreover, it can be arranged that the sum of the absolute values of the coefficients of the linear combination are bounded by $2C$.

In particular, consider first the function $g(z)$ restricted to $0 \leq z \leq 1$, and note that $g = (0) = 0$. For a partition $0 = t_0 < t_1 < \cdots < t_k = 1$, define

$$g_{k,+}(z) = \sum_{i=1}^{k-1} (g(t_i) - g(t_{i-1})) 1_{\{z \geq t_i\}}. \tag{18}$$

This piecewise constant function interpolates the function g at the points t_i for $i \leq k-1$. Note that $g_{k,+}$ is a linear combination of step functions. Now since the derivative of g is bounded by C on $[0, 1]$, it follows that the sum of the absolute values of the coefficients $\sum_i |g(t_i) - g(t_{i-1})|$ is bounded by C. In a similar way, define $g_{k,-}(z) = \sum_{i=1}^{k-1} (g(-t_i) - g(-t_{i-1})) 1_{\{z \leq -t_i\}}$. Adding these components $g_{n,-}(z) + g_{n,+}(z)$ yields a sequence of piecewise constant functions on $[-1, 1]$ that are uniformly close to $g(z)$ (as the maximum interval width tends to zero), and each of these approximating functions is a linear combination of step functions with the sum of the absolute values of the coefficients bounded by $2C$. It follows that the functions $g(z)$ are in the closure of the convex hull of the set of functions γ step $(z-t)$ and γ step $(-z-t)$ with $|\gamma| \leq 2C$ and $|t| \leq 1$, where step $(z) = 1_{\{z \geq 0\}}$ denotes the unit step function. Defining

$$G_{\text{step}} = \{\gamma \,\text{step}\,(\alpha \cdot x - t) : |\gamma| \leq 2C, |\alpha|_B = 1, \quad |t| \leq 1\}, \tag{19}$$

the following lemma has been demonstrated, where the closure property holds with the supremum norm on B, and hence with respect to $L_2(\mu, B)$.

Lemma 3: $G_{\cos}$ is in the closure of the convex hull of G_{step}.

It can be seen that Lemma 3 continues to work if, for each α, the parameter t is restricted to a subset T_α that is dense in $[-1, 1]$. In particular, restrict t to the continuity points of the distribution of $z = \alpha \cdot x$ induced by the measure μ on $\boldsymbol{R}^d$. Let G^{μ}_{step} be the subset of step functions in G_{step} with locations t restricted in this way. Then the following result holds.

Lemma 3′: $G_{\cos}$ is in the closure of the convex hull of G^{μ}_{step}.

Functions in G^{μ}_{step} are in the closure of the class of sigmoidal functions, taking the closure in $L_2(\mu, B)$. This follows by taking the sequence of sigmoidal functions $\phi(|a|(\alpha \cdot x - t))$ with $|a| \to \infty$. This sequence has pointwise limit equal to step $(\alpha \cdot x - t)$ (except possibly for x in the set with $\alpha \cdot x - t = 0$, which has μ measure zero by the restriction imposed on t). Consequently, by the dominated convergence theorem, the limit also holds in $L_2(\mu, B)$. Thus the desired closure property holds.

Lemma 4: G^{μ}_{step} is in the closure of G_ϕ.

Together, Lemmas 2, 3′, and 4 show that in $L_2(\mu, B)$,

$$\Gamma^0_C \subset \overline{c}oG_{\cos} \subset \overline{c}oG^{\mu}_{\text{step}} \subset \overline{c}oG_{\phi}, \tag{20}$$

where $\overline{c}oG$ denotes the closure in $L_2(\mu, B)$ of the convex hull of G, and Γ^0_C is the set of functions in Γ_C with $f(0) = 0$. Here the fact is used that the closure of a convex set is convex, so that it is not necessary to take the convex hull operation twice when combining the Lemmas. This completes the proof of Theorem 2. □

The proof of Theorem 1 is completed by using Lemma 1 and Theorem 2. To see that the constant in Theorem 1 can be taken to equal $(2C)^2$ for functions f in $\Gamma_{C,B}$, proceed as follows. The approximation bound is trivially true if $\|\overline{f}\| = 0$, where $\overline{f}(x) = f(0)$, for then $f(x)$ is equal to a constant μ-almost everywhere on B. So now suppose that $\|\overline{f}\| > 0$. If the sigmoidal function is bounded by one, then the functions in G_ϕ are bounded by $b = 2C$. Consequently, any c' greater than $(2C)^2 - \|\overline{f}\|^2$ is a valid choice for the application of Lemma 1. The conclusion is that there is a convex combination of n functions in G_ϕ for which the square of the $L_2(\mu, B)$ norm of the approximation error is bounded by c'/n.

If the sigmoidal function $\phi(x)$ is not bounded by one, first use Lemma 1 and the conclusion that $\Gamma^0_C \subset \overline{c}oG^{\mu}_{\text{step}}$ to obtain a convex combination of n functions in G^{μ}_{step} for which the squared $L_2(\mu, B)$ norm of the approximation error is bounded by $(2C)^2 - (1/2)\|\overline{f}\|^2$ divided by n. Then, using Lemma 4, by a suitable choice of scale of the sigmoidal function, sufficiently accurate replacements to the step functions can be obtained such that the resulting convex combination of n functions in G_ϕ yields a square $L_2(\mu, B)$ norm bounded by $(2C)^2/n$. This completes the proof of Theorem 1. □

Note that the argument given above simplifies slightly in the case that the distribution of $\alpha \cdot x$ is continuous for every α, for then G_{step} can be used in place of G^{μ}_{step} and there would be no need for Lemma 3′.

A variant of the theory just developed is to replace the function step (z) with the function $\text{step}_\phi(z)$, which is the same as the unit step function, except at $z = 0$ where it is set to equal $\phi(0)$. By a modification of the proof of Lemma 3, it can be shown that $G_{\cos}$ is in the closure of the convex hull of G_{step_ϕ}. The advantage of this variant is that G_{step_ϕ} is in the closure of G_ϕ, without any restriction on the location of the points t in $[-1, 1]$. But if $|\phi(0)| > 1$, then an additional argument would still be needed (as above, where t is restricted to the continuity points of $\alpha \cdot x$) in order to show that the constant c' can be taken to be not larger than $(2C)^2$.

V. Approximation on Other Bounded Domains in $\boldsymbol{R}^d$

In this brief section, the form of the constant $C_{f,B} = \int |\omega|_B |\tilde{f}(\omega)| d\omega$ in the approximation bound is determined for various choices of the bounded set B other than the Euclidean ball of radius r mentioned in the Introduction.

Recall that, by definition, $|\omega|_B = \sup_{x \in B} |\omega \cdot x|$. The interpretation is that $|\omega|_B$ bounds the domain of the trigono-

metric component $e^{i\omega\cdot x}$ that has frequency ω in the Fourier representation of the function restricted to x in B.

Clearly, if B is contained in a ball of radius r, that is, if $|x| \le r$ for x in B, then, by the Cauchy–Schwarz inequality, $|\omega|_B \le r|\omega|$. Thus,

$$C_{f,B} \le r \int |\omega| \left|\tilde{f}(\omega)\right| d\omega. \tag{21}$$

However, for some natural sets B, a fairly large radius ball would be required for application of the bound in that form. It is better to determine $|\omega|_B$ directly in some cases. If B is a multiple of a unit ball with respect to some norm on $\boldsymbol{R}^d$, then $|\omega|_B$ is determined by the dual norm.

In particular, if $B = B_{\infty,r} = \{x: |x|_\infty \le r\}$ is the l_∞ ball of radius r (the cube centered at $x = 0$ with sidelength $2r$), then $|\omega|_B = r|\omega|_1$ where $|\omega|_1$ is the l_1 norm and

$$C_{f,B_{\infty,r}} = r \int |\omega|_1 \left|\tilde{f}(\omega)\right| d\omega. \tag{22}$$

More generally, if $B = B_{p,r} = \{x: |x|_p \le r\}$ is the l_p ball of radius r, then $|\omega|_B = r|\omega|_q$ where $1/p + 1/q = 1$ (as can be seen by a standard application of Hölder's inequality). Here the l_p norm is given by $|x|_p = (\sum_{i=1}^d |x_i|^p)^{1/p}$ for $1 \le p < \infty$, and $|x|_\infty = \max_i |x_i|$ for $p = \infty$. Thus,

$$C_{f,B_{p,r}} = r \int |\omega|_q \left|\tilde{f}(\omega)\right| d\omega. \tag{23}$$

The approximation bound becomes

$$\int_{B_{p,r}} (f(x) - f_n(x))^2 \mu(dx) \le \frac{(2r)^2}{n} \left(\int |\omega|_q \left|\tilde{f}(\omega)\right| d\omega \right)^2, \tag{24}$$

for some network f_n of the form (1).

Note also that the center of the domain of integration may be taken to be any point x_0 in B not necessarily equal to 0. (This follows by a simple argument, since the magnitude of the Fourier transform is unchanged by translation.) In particular, for any cube C of side length s, the result becomes

$$\int_C (f(x) - f_n(x))^2 \mu(dx) \le \frac{s^2}{n} \left(\int |\omega|_1 \left|\tilde{f}(\omega)\right| d\omega \right)^2, \tag{25}$$

for some network f_n of the form (1). In like manner, a scaling argument shows that if Rect is any rectangle with side lengths $s_1, s_2, \cdots, s_d$, then there is a network f_n such that

$$\int_{\text{Rect}} (f(x) - f_n(x))^2 \mu(dx) \le \frac{1}{n} \left(\sum_{i=1}^d s_i \int |\omega_i| \left|\tilde{f}(\omega)\right| d\omega \right)^2. \tag{26}$$

In general, for a bounded set B, the point x_0 to take for the centering that would lead to the smallest approximation bound is one such that $C_{f,B,x_0} = \int |\omega|_{B,x_0} |\tilde{f}(\omega)| d\omega$ is minimized where $|\omega|_{B,x_0} = \sup_{x\in B} |\omega \cdot (x - x_0)|$. In this context, the representation (4) would become

$$f(x) = f(x_0) + \int (e^{i\omega\cdot x} - e^{i\omega\cdot x_0}) \tilde{f}(\omega)\, d\omega.$$

VI. Refinement

In the above analysis, the approximation results were proved by allowing the magnitude of the parameters a_k to be arbitrarily large. The absence of restrictions on $|a_k|$ yields a difficult problem of searching an unbounded domain. Large values of $|a_k|$ contribute to large gradients of the sigmoidal function which can also lead to difficulties of computation. In this section, we control the growth of $|a_k|$ and bound the effect on the approximation error. Knowledge of the relationship between the magnitude of the parameters and the accuracy of the network makes it possible to bound the index of resolvability of sigmoidal networks as in [11]. Bounds on the parameters are also required in the metric entropy computations as in White [12, Lemma 4.3] and Haussler [13].

Given $\tau > 0$, $C > 0$ and a bounded set B, let

$$G_{\phi,\tau} = \{\gamma\phi(\tau(\alpha \cdot x + b)): |\gamma| \le 2C, \ |\alpha|_B \le 1, \quad |b| \le 1\}. \tag{27}$$

This is the class of bounded multiples of a sigmoidal function, with the scale parameter of the sigmoid not larger than τ. We desire to bound the approximation error achievable by convex combinations of n functions in $G_{\phi,\tau}$.

Theorem 3: For every $f \in \Gamma_{C,B}$, $\tau > 0$, $n \ge 1$, every probability measure μ, and every sigmoidal function ϕ with $0 \le \phi(x) \le 1$, there is a function f_n in the convex hull of n functions in $G_{\phi,\tau}$ such that

$$\|\overline{f} - f_n\| \le 2C\left(\frac{1}{n^{1/2}} + \delta_\tau\right) \tag{28}$$

where $\|\cdot\|$ denotes the $L_2(\mu, B)$ norm, $\overline{f}(x) = f(x) - f(0)$, and

$$\delta_\tau = \inf_{0<\epsilon\le 1/2} \left\{ 2\epsilon + \sup_{|z|\ge\epsilon} \left|\phi(\tau z) - 1_{\{z>0\}}\right| \right\}. \tag{29}$$

Here, δ_τ is a distance between the unit step function and the scaled sigmoidal function. Note that $\delta_\tau \to 0$ as $\tau \to \infty$.

If ϕ is the unit step function, then $\delta_\tau = 0$ for all $\tau > 0$, and Theorem 3 reduces to Theorem 1.

If ϕ is the logistic sigmoidal function

$$\phi(z) = \frac{1}{1 + e^{-z}}, \tag{30}$$

then $|\phi(\tau z) + 1_{\{z>0\}}| \le e^{-\tau\epsilon}$ for $|z| \ge \epsilon$. Setting $\epsilon = (\ln\tau)/\tau$ yields

$$\delta_\tau \le \frac{1 + 2\ln\tau}{\tau}. \tag{31}$$

Therefore, if we set $\tau \ge n^{1/2}\ln n$, then from Theorem 3, for functions f in $\Gamma_{C,B}$,

$$\|\overline{f} - f_n\| \le O\left(\frac{1}{n^{1/2}}\right). \tag{32}$$

Similar conclusions hold for other sigmoidal functions. The size of τ_n required to preserve the order $(1/n)^{1/2}$ approximation depends on the rate at which the sigmoid approach the limits of 0 and 1 as $x \to \pm\infty$.

The proof of Theorem 3 is based on the following result for the univariate case. Let $g(z)$ be a function with a bounded derivative on $[-1, 1]$. Assume that 0 is in the range of the function g. Let g_τ denote a function in the convex hull of $G_{\phi,\tau}$.

Lemma 5: If g is a function on $[-1, 1]$ with derivative bounded by a constant C, then for every $\tau > 0$,

$$\inf_{g_\tau \in co G_{\phi,\tau}} \sup_{|z|\le\tau} |g(z) - g_\tau(z)| \le 2C\,\delta_\tau. \tag{33}$$

Proof: The proof of Lemma 5 is as follows. Given $0 < \epsilon \le 1/2$, let k be an integer satisfying $(1/\epsilon) - 1 < k \le 1/\epsilon$. Partition $[-1, 1]$ into k intervals of width $2/k$. Then approximate the function g by a linear combination of k unit step functions as in the proof of Lemma 3. Since the derivative is bounded by C, the error of the approximation is bounded by $2C/k$ for all z in $[-1, 1]$. Replacing each step function $1_{\{z\ge t_i\}}$ or $1_{\{z\le t_i\}}$ by the sigmoidal function $\phi(\tau(z - t_i))$ or $\phi(-\tau(z - t_i))$, respectively, one obtains a function g_τ in the convex hull of k functions in $G_{\phi,\tau}$, which has error bounded by

$$|g(z) - g_\tau(z)| \le \frac{2C}{k} + 2C \sup_{|y|\ge 1/k} |\phi(\tau y) - 1_{\{y>0\}}|, \tag{34}$$

for every z in $[-1, 1]$, where $2C/k$ bounds the contribution to the error from the replacement of the unit step function by the sigmoidal function centered at the point t_i closest to z, and $2C \sup_{|y|\ge 1/k} |\phi(\tau y) - 1_{\{y>0\}}|$ bounds the cumulative errors from the other sigmoidal functions (each of which is centered at distance at least $1/k$ from the point z). Since $\epsilon \le 1/k \le \epsilon/(1 - \epsilon)$, it follows that

$$|g(z) - g_\tau(z)| \le 2C\frac{\epsilon}{1-\epsilon} + 2C \sup_{|y|\ge\tau\epsilon} |\phi(\tau y) - 1_{\{y>0\}}|. \tag{35}$$

Using $\epsilon/(1-\epsilon) \le 2\epsilon$ for $0 < \epsilon \le 1/2$, and taking the infimum, completes the proof of Lemma 5. □

Proof of Theorem 3: The proof of Theorem 3 is as follows. From Lemma 2, $\overline{f}$ is in the closure of the convex hull of functions in $G_{\cos}$. The functions in $G_{\cos}$ are univariate functions $g(z)$ evaluated at a linear combination $z = \alpha \cdot x$ with $|g'(z)| \le C$ and $|z| \le 1$. Each such function g is approximated by a function g_τ as in Lemma 5 with supremum error bounded by a quantity arbitrarily close to $2C\delta_\tau$. It follows that there is a function f_τ in the closure of the convex hull of $G_{\phi,\tau}$ such that $\|\overline{f} - f_\tau\| \le 2C\delta_\tau$. From Lemma 1 there is an f_n in the convex hull of n points in $G_{\phi,\tau}$ such that $\|f_\tau - f_n\| \le 2C/n^{1/2}$. By the triangle inequality, this completes the proof of Theorem 3. □

VII. Extension

An extension of the theory is to replace $\boldsymbol{R}^d$ by a (possibly infinite dimensional) Hilbert space H, where now $\omega \cdot x$ denotes the Hilbert space inner product, and $|\cdot|$ denotes the Hilbert space norm. For instance, H may be the space L_2 of square integrable signals $(x(t), 0 \le t \le 1)$ with inner product $w \cdot x = \int_0^1 w(t)x(t)\,dt$. A real-valued function $f(x)$ of the signal $x \in H$ is to be approximated. The Fourier representation we require is that there is a complex-valued measure $\tilde{F}(d\omega) = e^{i\theta(\omega)}F(d\omega)$ on H such that $f(x) = \int_H e^{i\omega\cdot x}\tilde{F}(d\omega)$ or $f(x) = f(0) + \int_H (e^{i\omega\cdot x} - 1)\tilde{F}(d\omega)$.

Theorem 4: Let $f(x), x \in H$ be a function on a Hilbert space H with $C_f = \int_H |\omega| F(d\omega) < \infty$; then for every $r > 0$, every sigmoidal function ϕ on $\boldsymbol{R}^1$, every probability measure μ on H, and every $n \ge 1$, there is a linear combination of sigmoidal functions $f_n(x) = \sum_{k=1}^n c_k\phi(a_k \cdot x + b_k) + c_0$, such that $\int_{B_r} (f(x) - f_n(x))^2\mu(dx) \le (2rC_f)^2/n$, where $B_r = \{x \in H: |x| \le r\}$ is the Hilbert space ball of radius r.

The parameters a_k take values in the Hilbert space, while the other parameters are real-valued. With modification to the approximation bound, the norms of the parameters may be restricted in the same way as in Theorem 3.

For an interesting class of examples in this Hilbert space setting, let $R(t, s)$ be a positive definite function on $[0, 1]^2$ (a valid covariance function for a Gaussian process on $[0, 1]$, and suppose for simplicity that $R(t, t) = 1$. Let f be the function defined by

$$f(x) = \exp\left\{-\int_0^1\int_0^1 x(s)x(t)R(s, t)\,ds\,dt/2\right\} \tag{36}$$

for square-integrable x on $[0, 1]$. Note that $f(x)$ is the characteristic function of the Gaussian process $(\omega(t), 0 \le t \le 1)$ with mean zero and covariance $R(s, t) = E(\omega(s)\omega(t))$: that is, if F is the Gaussian measure on ω,

$$\begin{aligned}\int e^{i\omega\cdot x}F(d\omega) &= E[e^{i\omega\cdot x}] \\ &= \exp\left\{-\int_0^1\int_0^1 x(s)x(t)R(s, t)\,ds\,dt/2\right\} \\ &= f(x).\end{aligned} \tag{37}$$

Now, from the identity $E(\omega^2(t)) = R(t, t) = 1$, it follows that $E|\omega|^2 = \int_0^1 E\omega^2(t)\,dt = 1$. Therefore, the constant C_f in the approximation bound satisfies

$$\begin{aligned}C_f &= \int |\omega| F(d\omega) = E|\omega| \\ &\le (E|\omega|^2)^{1/2} \\ &= 1.\end{aligned} \tag{38}$$

Thus, for any probability measure μ on x and for any sigmoidal function ϕ on $\boldsymbol{R}^1$, it follows from the theorem that for this infinite-dimensional example there exists $f_n(x)$ such that

$$\int_{\{|x|\le 1\}} (f(x) - f_n(x))^2\mu(dx) \le \frac{4}{n}. \tag{39}$$

An even more general context may be treated in which the nonlinear functions are defined on a normed linear space. Let B be a bounded subset of a normed linear space X, and let ω take values in a set of bounded linear operators on X (the dual space of X). Now $\omega \cdot x$ denotes the operator ω applied to x and $|\omega|_B = \sup_{x\in B} |\omega \cdot x|$ denotes the norm of the operator restricted to B. If there is a measurable set of ω's and some complex-valued measure $\tilde{F}(d\omega)$ on this set, such that the function f has the representation $f(x) = \int e^{i\omega\cdot x}\tilde{F}(d\omega)$ or $f(x) = f(0) + \int (e^{i\omega\cdot x} - 1)\tilde{F}(d\omega)$ with

$C_{f,B} = \int |\omega|_B |\tilde{F}(d\omega)|$ finite, then for every $n > 1$, every probability measure μ on X, and every sigmoidal function ϕ, there will exist $f_n = \sum_{k=1}^{n} c_k \phi(a_k \cdot x + b_k) + c_0$ such that

$$\int_B (f(x) - f_n(x))^2 \mu(dx) \leq \frac{(2C_{f,B})^2}{n}, \tag{40}$$

where now the a_k's take values in the dual space of X.

One context for this more general result is the case that X is the set of bounded signals $(x(t), 0 \leq t \leq 1)$, $B = \{x: \sup_t |x(t)| \leq r\}$, and $\omega \cdot x = \int_0^1 \omega(t)x(t)\,dt$, where the bounded linear operators ω are identified with integrable functions on $[0, 1]$. Then $|\omega|_B = r\int_0^1 |\omega(t)|\,dt$, the Fourier distribution $\tilde{F}(d\omega)$ would be a measure supported on the set of ω in L_1, and the a_k's would be integrable functions on $[0, 1]$.

VIII. Iterative Approximation

In this section, it is seen that the bounds in Theorems 1, 3, and 4 can be achieved by an iterative sequence of approximations taking the form

$$f_n(x) = \alpha_n f_{n-1}(x) + c_n \phi(a_n \cdot x + b_n). \tag{41}$$

The optimization is restricted to the parameters α_n, γ_n, a_n, and b_n of the nth node, with the parameter values from earlier nodes held fixed. This iterative formulation considerably reduces the complexity of the surface to be optimized at each step.

This reduction in the complexity of the surface is particularly useful in the case that the function f is only observed at sites $X_1, X_2, \cdots, X_N$ in a bounded set B. The iterative approximation theory shows that to find an estimate with average squared error bounded by $(1/N)\sum_{i=1}^{N}(f(X_i) - f_n(X_i))^2 \leq c'/n$, it suffices to optimize the parameters of the network one node at a time. Avoiding global optimization has computational benefits. The error surface is still multimodal as a function of the parameters of the nth node, but there is a reduction in the dimensionality of the search problem by optimizing one node at a time.

A recent result of Jones [4] on iterative approximation in a Hilbert space is the key to the iterative approximation bound in the neural network case. As in the noniterative case, the applicability of Jones' Theorem is based on our demonstration that functions in Γ_C^0 are in the closure of the convex hull of G_ϕ.

To avoid cluttering the notation in this section, the notation f (instead of $\overline{f}$) is used to denote the point to be approximated by elements of the convex hull. As before, for the application to the approximation by sigmoidal networks of functions in Γ_C, one subtracts off the value of the function at $x = 0$ to obtain the function in Γ_C^0 which is approximated by functions in the convex hull of G_ϕ.

Let G be a subset of a Hilbert space. Let f_n be a sequence of approximations to an element f that take the form

$$f_n = \alpha_n f_{n-1} + \overline{\alpha}_n g_n \tag{42}$$

where $\overline{\alpha}_n = (1 - \alpha_n)$ and $0 \leq \alpha_n \leq 1$ and $g_n \in G$. Here α_n and g_n are chosen to achieve a nearly minimal value for $\|\alpha f_{n-1} + \overline{\alpha} g - f\|$. The iterations (42) are initialized with $\alpha_1 = 0$, so that f_1 is a point g_1 in G that achieves a nearly minimal value for $\|g_1 - f\|$. Note that f_n, as defined in (42), is in the convex hull of the points $g_1, \cdots, g_n$.

Jones [4] showed that if f is in the closure of the convex hull of G, then $\|f_n - f\|^2 \leq O(1/n)$, for the sequence of approximations defined as in (42). Here Jones' theorem and proof are presented with a minor refinement. The constant in the approximation bound is improved so as to agree with the constant in the noniterative version (Lemma 1). As noted by Jones, the error $\|\alpha f_{n-1} + \overline{\alpha} g - f\|$ need not be exactly minimized; here it is shown that it is enough to achieve a value within $O(1/n)^2$ of the infimum on each iteration.

Theorem 5: Suppose f is in the closure of the convex hull of a set G in a Hilbert space, with $\|g\| \leq b$ for each $g \in G$. Set $b_f^2 = b^2 - \|f\|^2$. Suppose that f_1 is chosen to satisfy $\|f_1 - f\|^2 \leq \inf_{g \in G} \|g - f\|^2 + \epsilon_1$ and, iteratively, f_n is chosen to satisfy

$$\|f_n - f\|^2 \leq \inf_{0 \leq \alpha \leq 1} \inf_{g \in G} \|\alpha f_{n-1} + \overline{\alpha} g - f\|^2 + \epsilon_n \tag{43}$$

where $\overline{\alpha} = 1 - \alpha$, $c' \geq b_f^2$, $\rho = c'/b_f^2 - 1$, and

$$\epsilon_n \leq \frac{\rho c'}{n(n + \rho)}. \tag{44}$$

Then for every $n \geq 1$,

$$\|f - f_n\|^2 \leq \frac{c'}{n}. \tag{45}$$

Proof: The proof of Theorem 5 is as follows. We show that if f is in the closure of the convex hull of G, then, for any given f_{n-1} and $0 \leq \alpha \leq 1$,

$$\begin{aligned} &\inf_{g \in G} \|\alpha f_{n-1} + \overline{\alpha} g - f\|^2 \\ &= \inf_{g \in G} \|\alpha(f_{n-1} - f) + \overline{\alpha}(g - f)\|^2 \\ &\leq \alpha^2 \|f_{n-1} - f\|^2 + (1 - \alpha)^2 b_f^2. \end{aligned} \tag{46}$$

The proof is then completed by setting $\alpha = b_f^2/(b_f^2 + \|f_{n-1} - f\|^2)$ to get

$$\|f_n - f\|^2 \leq \frac{b_f^2 \|f_{n-1} - f\|^2}{b_f^2 + \|f_{n-1} - f\|^2} + \epsilon_n, \tag{47}$$

or, equivalently,

$$\frac{1}{\|f_n - f\|^2 - \epsilon_n} \geq \frac{1}{\|f_{n-1} - f\|^2} + \frac{1}{b_f^2}. \tag{48}$$

Equation (48) provides what is needed to verify that $\|f_n - f\|^2 \leq c'/n$ by an induction argument. Indeed, (46) with $\alpha = 0$ shows that the desired inequality is true for $n = 1$. Suppose $\|f_{n-1} - f\|^2 \leq c'/(n-1)$, then plugging this into (48) and using $c' = b_f^2(1 + \rho)$ yields

$$\begin{aligned} \frac{1}{\|f_n - f\|^2 - \epsilon_n} &\geq \frac{1}{\|f_{n-1} - f\|^2} + \frac{1}{b_f^2} \\ &\geq \frac{n-1}{c'} + \frac{1}{b_f^2} \\ &= \frac{n + \rho}{c'}. \end{aligned} \tag{49}$$

Reciprocating and using the assumed bound on ϵ_n yields

$$\begin{aligned}\|f_n - f\|^2 &\le \frac{c'}{n+\rho} + \epsilon_n \\ &= \frac{c'}{n} - \frac{c'\rho}{n(n+\rho)} + \epsilon_n \\ &\le \frac{c'}{n}, \end{aligned} \tag{50}$$

as desired.

Thus, it remains to verify (46). Given $\delta > 0$, let f^* be a point in the convex hull of G with $\|f - f^*\| \le \delta$. Thus f^* is of the form $\sum_{k=1}^m \gamma_k g_k^*$ with $g_k^* \in G, \gamma_k \ge 0, \sum_{k=1}^m \gamma_k = 1$, for some sufficiently large m. Then,

$$\begin{aligned}&\|\alpha(f_{n-1} - f) + \bar{\alpha}(g - f)\| \\ &\qquad \le \|\alpha(f_{n-1} - f) + \bar{\alpha}(g - f^*)\| + \delta. \end{aligned} \tag{51}$$

Expanding the square yields

$$\begin{aligned}&\|\alpha(f_{n-1} - f) + \bar{\alpha}(g - f^*)\|^2 \\ &= \alpha^2\|f_{n-1} - f\|^2 + \bar{\alpha}^2\|g - f^*\|^2 \\ &\quad + 2\alpha\bar{\alpha}(f_{n-1} - f, g - f^*), \end{aligned} \tag{52}$$

where $(\cdot,\cdot)$ denotes the inner product. Now the average value of the last two terms is, for $g \in \{g_1^*, \cdots, g_m^*\}$,

$$\begin{aligned}&\sum_{k=1}^m \gamma_k(\bar{\alpha}^2\|g_k^* - f^*\|^2 + 2\alpha\bar{\alpha}(f_{n-1} - f, g_k^* - f^*)) \\ &= \bar{\alpha}^2\sum_{k=1}^m \gamma_k\|g_k^* - f^*\|^2 + 0 \\ &= \bar{\alpha}^2\left(\sum_{k=1}^m \gamma_k\|g_k^*\|^2 - \|f^*\|^2\right) \\ &\le \bar{\alpha}^2(b^2 - \|f^*\|^2). \end{aligned} \tag{53}$$

Since the average value is bounded in this way, there must exist $g \in \{g_1^*, \cdots, g_m^*\}$, such that

$$\begin{aligned}&\|\alpha(f_{n-1} - f) + \bar{\alpha}(g - f^*)\|^2 \\ &\qquad \le \alpha^2\|f_{n-1} - f\|^2 + \bar{\alpha}^2(b^2 - \|f^*\|^2). \end{aligned} \tag{54}$$

Now by the triangle inequality, $\|f^*\| > \|f\| - \delta$. So using (51) and letting $\delta \to 0$, it follows that

$$\begin{aligned}&\inf_{g\in G}\|\alpha(f_{n-1} - f) + \bar{\alpha}(g - f)\|^2 \\ &\qquad \le \alpha^2\|f_{n-1} - f\|^2 + \bar{\alpha}^2(b^2 - \|f\|^2), \end{aligned} \tag{55}$$

as desired. This completes the proof of Theorem 5. □

Inspection of the proof shows an alternative optimization that may provide further simplification in some cases. Instead of minimizing the sum of squares $\|\alpha f_{n-1} + \bar{\alpha}g - f\|^2$ at each iteration, one may instead choose $g \in G$ to maximize the inner product $(f - f_{n-1}, g)$. (In this case, one can derive the bound $\|f - f_n\| \le (2b)^2/n$.) For sigmoids, the search task reduces to finding the parameters a_n and b_n such that the inner product of $\phi(a \cdot x + b)$ and $f - f_{n-1}$ is maximized. The function f_n depends linearly on the other parameters α_n and c_n in (41), so they may be determined by ordinary least squares.

IX. Properties and Examples of Functions in Γ

In this section, sevcral properties and examples of functions $f(x)$ are presented for which the Fourier integral $C_f = \int |\omega|F(d\omega)$ is evaluated or bounded, where $F(d\omega)$ is the magnitude distribution of the Fourier transform. Note that $C_f = C_{f,B}$ in the case that B is the unit ball centered at zero. Examples are also given of classes of functions in Γ_*, that is, functions on $\boldsymbol{R}^2$ that are contained in Γ_B when restricted to any bounded set B. The simpler facts are stated without proof.

1) *Translation:* If $f(x) \in \Gamma_C$, then $f(x + b) \in \Gamma_C$.
2) *Scaling:* If $f(x) \in \Gamma_C$, then $f(ax) \in \Gamma_{|a|C}$.
3) *Combination:* If $f_i(x) \in \Gamma_{C_i}$, then $\sum \beta_i f_i \in \Gamma_{\sum |\beta_i| C_i}$.
4) *Gaussian:* If $f(x) = e^{-|x|^2/2}$, then $C_f \le d^{1/2}$. Indeed, $\tilde{f}(\omega) = (2\pi)^{-d/2}e^{-|\omega|^2/2}$ and $C_f = \int |\omega|\tilde{f}(\omega)\,d\omega$ which is bounded by $(\int |\omega|^2\tilde{f}(\omega)\,d\omega)^{1/2} = d^{1/2}$.
5) *Positive Definite Functions:* $C_f \le (-f(0)\nabla^2 f(0))^{1/2}$. A positive definite function $f(x)$ is one such that $\sum_{i,j} x_i x_j f(x_i - x_j)$ is nonnegative for all $x_1, x_2, \cdots, x_k$ in R^d. Positive definite functions arise as covariance functions for random fields and as characteristic functions for probability distributions on $\boldsymbol{R}^d$. The essential property (due to Bochner) is that continuous positive definite functions are characterized as functions that have a Fourier representation $f(x) = \int e^{i\omega\cdot x} F(d\omega)$ in terms a positive real-valued measure F. If f is a twice continuously differentiable positive definite function, then by the Cauchy–Schwarz inequality $\int |\omega|F(d\omega) \le (\int F(d\omega) \int |\omega|^2 F(d\omega))^{1/2} = (-f(0)\nabla^2 f(0))^{1/2}$, where $\nabla^2 f(x) = \sum_{i=1}^d \partial^2 f(x)/\partial x_i^2$. (Positive definite functions have a maximum at $x = 0$, so $\nabla^2 f(0) \le 0$.) What is noteworthy about this class of functions for our approximation purposes is that the C_f is bounded in terms of behavior at a single point $x = 0$. Moreover, since $\nabla^2 f(0)$ is a sum of d terms, it is plausible to model a moderate behavior of the constant C_f, such as order $d^{1/2}$, for positive definite functions of many variables.
6) *Integral Representations:* Suppose $f(x) = \int K(a(x + b))G(da, db)$ for some location and scale mixture of a function $K(x)$ in Γ, for $a > 0$ and $b \in R^d$. (For instance, $K(x)$ may be a Gaussian density or other positive definite kernel on R^d.) Then $C_f \le C_K \int |a||G|(da, db)$. In the same way, if the function has a representative $f(x) = \int K(a \cdot x + b)G(da, db)$, for $a \in \boldsymbol{R}^d$ and $b \in R^d$, for some $K(z)$ on $\boldsymbol{R}^1$ and some signed measure $G(da, db)$, then $C_f \le C_K \int |a||G|(da, db)$.
7) *Ridge Functions:* If $f(x) = g(a \cdot x)$ for some direction $a \in \boldsymbol{R}^d$ with $|a| = 1$ and some univariate function $g(z)$ with integrable Fourier transform $\tilde{g}$ on $\boldsymbol{R}^1$, then f has a Fourier representation in terms of a singular distribution $\tilde{F}(d\omega)$ concentrated on the set of ω in $\boldsymbol{R}^d$ in the direction a, that is, $f(x) = \int e^{it a\cdot x}\tilde{g}(t)\,dt$. In this case, $C_f = C_g = \int_{\boldsymbol{R}^1} |t||\tilde{g}(t)|\,dt$. If $f(x) = g(a \cdot x)$ for some $a \in \boldsymbol{R}^d$ with $|a| = 1$ and the derivative of g is a continuous positive definite function on $\boldsymbol{R}^1$, then

$C_f = C_g = g'(0)$. Note that C_f is independent of the dimension d.

8) *Sigmoidal Functions on R^d:* These are ridge functions of the form $f(x) = \phi(a \cdot x + b)$ for some a and b in R^d, for some sigmoidal function $\phi(z)$ on R^1. Generally, such sigmoidal functions do not have an integrable Fourier transform. Nevertheless, typical choices of smooth sigmoidal functions $\phi(x)$ have a derivative $\phi'(z)$ with an integrable transform $\tilde{\phi}'(t)$. In this case, f is in Γ with $C_f = |a| \int |\tilde{\phi}'(t)|\, dt$. Using the closure properties for linear combinations and compositions, it is seen that certain multiple-layer sigmoidal networks are also in Γ.

9) *Radial Functions:* Suppose $f(x) = g(|x|)$ is a function that depends on x only through the magnitude $|x|$ (i.e., the angular components of $f(x)$ are constant), and that f has a Fourier representation $f(x) = \int e^{i\omega \cdot x} \tilde{f}(\omega)\, d\omega$. Then $\tilde{f}(\omega)$ is also a radial function; that is, $\tilde{f}(\omega) = \tilde{g}(|\omega|)$ for some function $\tilde{g}$ on R^1. Integrating $|\omega||\tilde{F}(\omega)$ using polar coordinates yields $C_f = S_d \int_0^\infty r^d |\tilde{g}(r)|\, dr$, where S_d is the $d-1$-dimensional volume of the unit sphere in R^d. The factor r^d in the integrand suggests that C_f is typically exponentially large in d; for an exception, the Gaussian function in example (4) is a radial function with $C_f \leq d^{1/2}$.

10) *Sigmoidal Approximation with an Augmented Input Vector:* For a d-dimensional input x, let x' in R^{2d} consist of the coordinates of x and the squares of these coordinates. Then with the unit step function for ϕ, the terms in the approximation $f_n(x') = \sum c_k \phi(a_k \cdot x' + b_k)$ with $a_k, b_k \in R^{2d}$ consist of indicators of ellipsoidal regions. Functions $f(x)$ on R^d can have a significantly smaller value for C_f when represented as a function of x' on R^{2d}. In particular, consider the functions of the form $f(x) = g(\sum a_i x_i^2)$ on R^d with $\sum a_i^2 = 1$. These functions include the radial functions and may be interpreted as a ridge function in the squared components. In this case, if g has a Fourier transform $\tilde{g}$ on R^1, then there is a representation of the function f as a function on R^{2d}, with C_f given by the one-dimensional integral $\int_{R^1} |t||\tilde{g}(t)|\, dt$. This potential improvement in the constant in the approximation bound helps justify the common practice of including the squares of the inputs in the sigmoidal network.

For the following examples, let $\Gamma(a, c) \subset \Gamma_c$ be the class of functions $f(x)$ on R^d for which there is a Fourier representation $f(x) = \int e^{i\omega \cdot x} \tilde{F}(d\omega)$ with magnitude distribution $F(d\omega)$ satisfying $\int F(d\omega) \leq a$ and $\int |\omega| F(d\omega) \leq c$.

11) *Products of Functions in Γ:* If $f_1 \in \Gamma(a_1, c_1)$ and $f_2 \in \Gamma(a_2, c_2)$, then the product $f_1(x) f_2(x)$ is a function in $\Gamma(a_1 a_2, a_1 c_2 + a_2 c_1)$. This follows by applying Young's convolution inequality to the Fourier representations of the product $f(x)g(x)$ and its gradient $f(x)\nabla g(x) + g(x)\nabla f(x)$. Together with property (3), this shows that the class of functions in Γ for which both $\int F(d\omega)$ and $\int |\omega| F(d\omega)$ are finite forms an algebra of functions closed under linear combinations and products.

12) *Composition with Polynomials:* If $g \in \Gamma(a, c)$, then $(g(x))^k$ is in $\Gamma(a^k, ka^{k-1}c)$. It follows that if $f(z)$ is a polynomial function of one variable and $g \in \Gamma(a, c)$, then the composition $f(g(x))$ is in $\Gamma(f_{\text{abs}}(a), c f'_{\text{abs}}(a))$, where f_{abs} is the polynomial obtained from f by replacing each coefficient with its absolute value. A similar statement can be made for the composition of a polynomial function of severable variables with functions in $\Gamma(a, c)$.

13) *Composition with Analytic Functions:* If $g \in \Gamma(a, c)$ and $f(x)$ is an analytic function represented by a power series $f(z) = \sum_{k=0}^\infty a_k z^k$ with radius of absolute convergence $r > a$, then the composition $f(g(x))$ is in $\Gamma(f_{\text{abs}}(a), c f'_{\text{abs}}(a))$ where $f_{\text{abs}}(z) = \sum_k |a_k| z^k$.

The next examples concern functions in Γ_*—that is, functions which can be modified outside of bounded sets B to produce functions in Γ. For f in Γ_*, let $C^*_{f,B} = \inf_g C_{g,B}$, where the infimum is over g in Γ that agree with f on the set B, $C_{g,B} = \int |\omega|_B G(d\omega)$, and G is the magnitude distribution in the Fourier representation of g on R^d. For functions in Γ_*, the approximation error $\int_B (f(x) - f_n(x))^2 \mu(dx)$ is bounded by $(2C^*_{f,B})^2/n$, for some sigmoidal network f_n of the form (1).

14) *Linear Functions and Other Polynomials:* If $f(x) = a \cdot x$, then f is in Γ_*. Moreover, $C^*_{f,B} \leq |a| r$, for every set B contained in the ball $\{x: |x| \leq r\}$, for every radius $r > 0$. This is shown first in the case that $d = 1$ and $f(x) = x$ on $[-r, r]$, by considering certain extrapolations $h(x)$. In particular, let $h(x) = h_b(x)$ have derivative $h'(x)$ that is equal to 1 for $|x| \leq r$, equal to 0 for $|x| \geq r + b$, and interpolates linearly between 1 and 0 for $r \leq |x| \leq r + b$, for some $b > 0$. Then $h'(x)$ has a Fourier transform that can be calculated to be $2 \sin(\omega b/2) \sin(\omega(r + b/2))/(\omega^2 \pi b)$. By a change of variables ($t = \omega b/2$) and an application of the dominated convergence theorem, it is seen that as b tends to infinity, $C_{h_b} = \int |\omega||\tilde{h}_b(\omega)\, d\omega = \int |\sin(t) \sin(t(1 + 2r/b))|/(t^2 \pi)\, dt$ converges to $\int (\sin(t))^2/(t^2 \pi)\, dt = 1$. (This matches the intuition that as b tends to infinity, $h_b(x)$ approaches $f(x)$ which has a constant derivative equal to one, so the Fourier transform of $h'_b(x)$ ought to approximate a "delta function" and the integral of $|\tilde{h}'_b|$ should be close to one.)[1] Consequently, $C^*_{f,[-r,r]} \leq \lim_{b \to \infty} r C_{h_b} = r$. For $f(x) = a \cdot x$ on R^d, let $g_b(x) = |a| h_b(\alpha \cdot x)$ where $\alpha = a/|a|$. Then for sets B in the ball B_r of radius r, $C^*_{f,B} \leq |a| \inf_b C_{h_b,[-r,r]} \leq |a| r$. Other extrapolations of $f(x) = x$ on $[-r, r]$ can be constructed for which $\tilde{g}(\omega)$, as well as $\omega \tilde{g}(\omega)$, are integrable. Together with property (12), this implies that polynomial functions (in one or several variables) are contained in Γ_*. Manageable bounds on the constraints $C_{f,B}$ can be obtained in the case of sparse polynomials and in the

[1] For an alternative treatment in which the Fourier representation is generalized to allow for functions with a linear component on R^d, see the remarks in the Appendix.

case of multiple-layer polynomials networks, which are polynomials defined in terms of a restricted number of elementary compositions (sums and products).

15) *Functions with Derivatives of Sufficiently High Order:* If the partial derivatives of $f(x)$ of order $s = \lfloor d/2 \rfloor + 2$ are continuous on $\boldsymbol{R}^d$, then f is in Γ_*. Consider first the case that the partial derivatives of order less than or equal to s are square-integrable on $\boldsymbol{R}^d$. In this case, f is in Γ. Indeed, write $|\tilde{f}(\omega)||\omega| = a(\omega)b(\omega)$ with $a(\omega) = (1 + |\omega|^{2k})^{-1/2}$ and $b(\omega) = |\tilde{f}(\omega)||\omega|(1 + |\omega|^{2k})^{1/2}$, where $k = s - 1$. By the Cauchy–Schwarz inequality, $C_f = \int a(\omega)b(\omega)\,d\omega$ is bounded by the product of $(\int a^2(\omega)\,d\omega)^{1/2}$ and $(\int b^2(\omega)\,d\omega)^{1/2}$. Now the integral $\int a^2(\omega)\,d\omega = \int (1 + |\omega|^{2k})^{-1}\,d\omega$ is finite for $2k > d$ and, by Parseval's identity, the integral $\int b^2(\omega)\,d\omega = \int |\tilde{f}(\omega)|^2(|\omega|^2 + |\omega|^{2s})\,d\omega$ is finite when the partial derivatives of order s and of order 1 are square-integrable on $\boldsymbol{R}^d$. This demonstrates that f is in Γ. Now suppose that the partial derivatives of order s are continuous, but not necessarily square-integrable on $\boldsymbol{R}^d$. Given $r > 0$, let $\rho(x)$ be an s-times continuously differentiable function that is equal to 1 on $B_r = \{x: |x| \leq r\}$ and equal to 0 for $|x| \geq r'$ for some $r' > r$. (In particular, we can take $\rho(x) = \rho_1(|x|)$ where $\rho_1(z)$ equals 1 for $z \leq r$, 0 for $z \geq r'$, and interpolates by a (piecewise polynomial) spline of order s for $r \leq z \leq r'$.) Consider the function $f_r(x) = f(x)\rho(x)$, which agrees with $f(x)$ on B_r. It has continuous partial derivatives for order s which are equal to zero for $|x| > r'$, and hence are integrable on $\boldsymbol{R}^d$. Consequently, for each $r > 0$, the function $f_r(x)$ is in Γ. It follows that f is in Γ_*. A disadvantage of characterizing approximation capabilities in terms of high-order differentiability properties is that the bound on the constant C_f can be quite large. Indeed, the integral $\int |\tilde{f}(\omega)|^2|\omega|^{2s}\,d\omega$, as characterized by Parseval's identity, involves a sum of exponentially many terms—one for each partial derivative of order s.

The last three examples involve functions in Γ which have a discrete domain for either of the variables x or ω.

16) *Absolutely Convergent Fourier Series:* Suppose f is a continuous function on $[0, 1]^d$; let $\tilde{f}_k = (2\pi)^{-d} \int_{[0, 1]^d} e^{-i2\pi k\cdot x} f(x)\,dx$ be the Fourier series coefficients of f, and suppose that f_k and $k\tilde{f}_k$ are absolutely summable sequences for $k \in \boldsymbol{Z}^d$, where $\boldsymbol{Z}^d$ is the set of vectors with integer coordinates. Then f is a function in Γ with the Fourier series representation $f(x) = \sum_k e^{i2\pi k\cdot x}\tilde{f}_k$ and $C_f = 2\pi \sum_k |k||\tilde{f}_k|$. Thus, f has a Fourier distribution $\tilde{F}$ restricted to the lattice of points ω of the form $2\pi k$, with $\tilde{F}(\{2\pi k\}) = \tilde{f}_k$. In this case, for $\tilde{f}_k$ and $k\tilde{f}_k$ to be absolutely summable, it is necessary that the function $f(x)$ possess a continuously differentiable periodic extension to $\boldsymbol{R}^d$. It is a simple matter to periodically extend a function defined on $[0, 1]^d$ such that $f(x + k) = f(x)$ for all vectors of integers k; however, for functions that arise in practice, it is rare for this periodic extension to be continuous or to possess a continuous derivative on the boundary points of $[0, 1]^d$. (It is for this reason that in this paper the Fourier distribution has not been restricted exclusively to such a lattice; allowing the coordinates of ω to take arbitrary values relaxes the boundary constraints.) In some cases, it is possible to take a function defined on a subset of $[0, 1]^d$ and extend it in a continuously differentiable way to a function that is zero and has zero gradient on the boundary of $[0, 1]^d$, so that the requirement of a continuously differentiable periodic extension is satisfied. Examples of this are similar to those given in 14) and 15).

17) *Functions of Integers:* Functions defined on $\boldsymbol{Z}^d$ are in Γ_B for any finite subset B of $\boldsymbol{Z}^d$. Indeed, given such a set B, set $\tilde{f}_B(\omega)$ to equal $(1/2\pi)^d \sum_{x\in B} e^{-i\omega\cdot x} f(x)$ for ω in $[-\pi, \pi]^d$ and to equal 0 otherwise.[2] Then the Fourier representation $f(x) = \int_{[-\pi, \pi]^d} e^{i\omega\cdot x}\tilde{f}_B(\omega)\,d\omega$ holds for x in B. Now $\tilde{f}_B(\omega)$ is a continuous function on the set $[-\pi, \pi]^d$, which implies that $|\tilde{f}_B(\omega)|$ is bounded and $C_{f,B} = \int_{[-\pi, \pi]^d} |\omega|_B|\tilde{f}_B(\omega)|\,d\omega$ is finite. Consequently, f is in Γ_B. Moreover, $C_{f,B}$ may be bounded in terms of the L_1 norm of the Fourier transform: that is $C_{f,B} \leq \pi s d \int_{[-\pi, \pi]^d} |\tilde{f}_B(\omega)|\,d\omega$, where $s = \max_{x\in B}|x|_\infty$. As a practical matter, if d is large, then additional structure needs to be assumed of the function to have a moderate value of $C_{f,B}$.

18) *Boolean Functions:* Here, we consider functions defined on $B = \{0, 1\}^d$ and note that, for Boolean functions, $\Gamma_{C,B}$ is related to a class of functions recently examined by Siu and Bruck [5] and Kushilevitz and Mansour [27]. This leads to a number of additional interesting examples of functions with not excessively large values of $C_{f,B}$. For functions $f(x)$ on $\{0, 1\}^d$, the Fourier distribution may be restricted to the set of ω of the form πk with $k \in \{0, 1\}^d$ (for then the functions $e^{i\pi k\cdot x}$, $k \in \{0, 1\}^d$ are orthogonal functions that span the 2^d-dimensional linear space of real-valued functions on $\{0, 1\}^d$). Consequently, $f(x) = \sum_{k\in\{0, 1\}^d} e^{i\pi k\cdot x}\tilde{f}_k$, where $\tilde{f}_k = 2^{-d}\sum_{x\in\{0, 1\}^d} e^{-i\pi k\cdot x} f(x)$. Here, $C_{f,B} = \pi \sum_{k\in\{0, 1\}^d} |k|_1|\tilde{f}_k|$ which is bounded above by $\pi d L(f)$, where $L(f) = \sum_{k\in\{0, 1\}^d} |\tilde{f}_k|$ is the spectral norm. Now the class PL of Boolean functions, for which $L(f)$ is bounded by a polynomial function of d (that is, $L(f) \leq d^c$ for some fixed $c \geq 1$), is examined in [5] and [27].[3] In particular, Siu and Bruck [5] show, among other examples, that the Boolean function on $\{0, 1\}^{2d}$ defined by the comparison of two d-bit

[2] If it happens that $f(x)$ is an absolutely summable function n $\boldsymbol{Z}^d$, then the transform $\tilde{f}(\omega) = (1/2\pi)^d \sum_x e^{-i\omega\cdot x} f(x)$, $\omega \in [-\pi, \pi]^d$, may be used in place of $\tilde{f}_B(\omega)$.

[3] The cited references express the Fourier transform in terms of a polynomial basis that turns out to be identical to the Fourier basis used here. Indeed, for x restricted to $\{0, 1\}^d$, the Fourier basis functions $e^{i\pi k\cdot x} = \prod_{j=1}^d (e^{i\pi x_j})^{k_j}$ may be expressed in the polynomial form $\prod_{j=1}^d \bar{x}_j^{k_j}$, where $\bar{x}_j = 1 - 2x_j$ equals 1, -1 for x_j equal to 0, 1, respectively (in agreement with the values assigned by $e^{i\pi x_j}$).

integers and the functions defined by (each bit of) the addition of two such integers are functions in PL with $L(f) = d+1$. It follows that $C_{f,B} \leq 2\pi d(d+1)$ for the *comparison* and *addition* functions. On the other hand, they show that the *majority* function $1_{\{\sum_{j=1}^{d} x_j - d/2\}}$ (which has a simple network representation) is not in the class PL. Kushilevitz and Mansour [27] show that a class of binary *decision trees* represent Boolean functions satisfying $L(f) \leq m$, where m is the number of nodes of the tree. It follows that $C_{f,B} \leq \pi m d$ for such decision trees. Bellare [28] generalizes the results of [27] by allowing decision trees with more general PL functions implemented at the nodes of the tree. He gives bounds on $L(f)$ in terms of spectral norms of the node functions, from which bounds on $C_{f,B}$ follow for the classes of decision trees he considers. The implication of polynomial bounds on $C_{f,B}$, as a consequence of the bound $2C_{f,B}/\sqrt{n}$ from Theorem 1, is that a polynomial rather than an exponential number of nodes n is sufficient for accurate approximation by sigmoidal networks.

X. Lower Bounds for Approximation by Linear Subspaces

The purpose of this section is to present and derive a lower bound on the best approximation error for linear combinations of any fixed basis functions for functions in Γ_C. These results, taken together with Theorem 1, show that fixed basis function expansion must have a worst-case performance that is much worse that that which is proven to be achievable by certain adaptable basis function methods (such as neural nets).

Let μ be the uniform probability distribution on the unit cube $B = [0,1]^d$, and let $d(f,g) = (\int_{[0,1]^d}(f(x) - g(x))^2\,dx)^{1/2}$ be the distance between functions in $L_2(\mu, B)$. For a function f and a set of functions G, let $d(f,G) = \inf_{g\in G} d(f,g)$. For a collection of basis functions $h_1, h_2, \cdots, h_n$

$$d(f, H_n) = d(f, \operatorname{span}(h_1, h_2, \cdots, h_n)) \tag{56}$$

denotes the error in the approximation of a function f by the best linear combination of the functions $h_1, h_2, \cdots, h_n$, where $H_n = \operatorname{span}(h_1, h_2, \cdots, h_n)$. The behavior of this approximation error for functions in $\Gamma_C = \Gamma_{C,B}$ may be characterized (in the worse case) by

$$\sup_{f\in\Gamma_C} d(f, H_n). \tag{57}$$

Here, a lower bound to this approximation error is determined that holds uniformly over all choices of fixed basis functions. In this formulation, the functions h_i are not allowed to depend on f (in contrast, sigmoidal basis functions have nonlinear parameters that are allowed to be adjusted in the fit to f). Let

$$W_n = \inf_{h_1,\cdots,h_n} \sup_{f\in\Gamma_C} d(f, \operatorname{span}(h_1, h_2, \cdots, h_n)). \tag{58}$$

This is the Kolmogorov n-width of the class of functions Γ_C.

Theorem 6: For every choice of fixed basis functions $h_1, h_2, \cdots, h_n$,

$$\sup_{f\in\Gamma_C} d(f, \operatorname{span}(h_1, h_2, \cdots, h_n)) \geq \kappa \frac{C}{d}\left(\frac{1}{n}\right)^{1/d}, \tag{59}$$

where κ is a universal constant not smaller than $1/(8\pi e^{\pi-1})$. Thus, the Kolmogorov n-width of the class of functions Γ_C satisfies

$$W_n \geq \kappa \frac{C}{d}\left(\frac{1}{n}\right)^{1/d}. \tag{60}$$

The proof of Theorem 6 is based on the following Lemma.

Lemma 6: No linear subspace of dimension n can have squared distance less than $1/2$ from every basis function in an orthonormal basis of a $2n$-dimensional space.

Proof: For the proof of Lemma 6, it is to be shown that if $e_1, \cdots, e_{2n}$ is an orthonormal basis and $G_n = \operatorname{span}\{g_1, \cdots, g_n\}$ is a linear subspace of dimension n, then there is an e_j such that the squared distance $d^2(e_j, G_n) \geq 1/2$. Indeed, let P denote projection onto G_n. Then, $d^2(e_j, G_n) = \|e_j - Pe_j\|^2 = \|e_j\|^2 - \|Pe_j\|^2 = 1 - \|Pe_j\|^2$. Thus it is equivalent to show that there is an e_j such that the norm squared of the projection satisfies $\|Pe_j\|^2 \leq 1/2$. Without loss of generality, take $g_1, \cdots, g_n$ to be an orthonormal basis of G_n. Then the projection Pe_j takes the form $\sum_{i=1}^{n}(e_j, g_i)g_i$. So the norm squared of the projection satisfies $\|Pe_j\|^2 = \sum_{i=1}^{n}(e_j, g_i)^2$. Taking the average for $j = 1, \cdots, 2n$, exchanging the order of summation, and using $\|g_i\| = 1$, yields

$$\begin{aligned}\frac{1}{2n}\sum_{j=1}^{2n}\|Pe_j\|^2 &= \frac{1}{2n}\sum_{j=1}^{2n}\sum_{i=1}^{n}(e_j, g_i)^2 \\ &= \frac{1}{2n}\sum_{i=1}^{n}\sum_{j=1}^{2n}(e_j, g_i)^2 \\ &= \frac{1}{2n}\sum_{i=1}^{n}\|g_i\|^2 \\ &= \frac{n}{2n} = \frac{1}{2}.\end{aligned} \tag{61}$$

Since the average value of the norm squared of the projection $\|Pe_j\|^2$ is equal to 1/2, there must exist a choice of the basis function e_j for some $1 \leq j \leq 2n$ for which $\|Pe_j\|^2 \geq 1/2$. This completes the proof of Lemma 6. □

Proof of Theorem 6: Let $h_1^*, h_2^*, \cdots$ be the functions $\cos(\omega \cdot x)$ for $\omega = 2\pi k$ for $k \in \{0, 1, \cdots\}^d$ ordered in terms of increasing l_1 norm $|k|_1 = \sum_{i=1}^{d}|k_i|$. Let H_{2n}^* denote the span of $h_1^*, \cdots, h_{2n}^*$. We proceed as follows. First reduce the supremum over Γ_C by restricting to functions in H_{2n}^*, then lower bound further by replacing the arbitrary basis functions $h_1, \cdots, h_n$ with their projections onto H_{2n}^*, which we denote by $g_1, \cdots, g_n$. Then $g_1, \cdots, g_n$ span an n-dimensional linear subspace of H_{2n}^* and a lower bound is obtained by taking the infimum over all n-dimensional linear subspaces G_n. The supremum is then restricted to multiples of the orthogonal functions h_j^* that belong to Γ_C, which permits application of

the lemma. Thus, putting it all together,

$$\begin{aligned} W_n &= \inf_{h_1,\cdots,h_n} \sup_{f\in\Gamma_C} d(f, \operatorname{span}(h_1, h_2, \cdots, h_n)) \\ &\geq \inf_{h_1,\cdots,h_n} \sup_{f\in H_{2n}^* \bigcap \Gamma_C} d(f, \operatorname{span}(h_1, h_2, \cdots, h_n)) \\ &\geq \inf_{h_1,\cdots,h_n} \sup_{f\in H_{2n}^* \bigcap \Gamma_C} d(f, \operatorname{span}(g_1, g_2, \cdots, g_n)) \\ &\geq \inf_{G_n} \sup_{f\in H_{2n}^* \bigcap \Gamma_C} d(f, G_n) \\ &\geq \inf_{G_n} \sup_{f\in\{(C/|\omega_j|)\cos(\omega_j\cdot x),\, j=1,\cdots,2n\}} d(f, G_n) \\ &\geq \min_{j=1,\cdots,2n} \left(\frac{C}{2\pi|k_j|}\right) \\ &\quad \cdot \left(\inf_{G_n} \sup_{f\in\{\cos(\omega_j\cdot x),\, j=1,\cdots,2n\}} d(f, G_n)\right) \\ &\geq \min_{j=1,\cdots,2n} \left(\frac{C}{2\pi|k_j|}\right)\frac{1}{2} \\ &\geq \frac{C}{4\pi m}, \end{aligned} \tag{62}$$

for m satisfying $\binom{m+d}{d} \geq 2n$ (such that the number of multiindices with norm $|k| \leq m$ is at least $2n$). A bound from Stirling's formula yields $\binom{m+d}{d} \geq (m/\tau d)^d$ for a universal constant $\tau \geq e^{\pi-1}$. Setting $m = \lceil \tau d n^{1/d} \rceil$ and adjusting the constant to account for the rounding of m to an integer, the desired bound is obtained, namely,

$$W_n \geq \frac{C}{8\pi\tau d}\left(\frac{1}{n}\right)^{1/d} \tag{63}$$

This completes the proof of Theorem 6. □

XI. Conclusion

The error in the approximation of functions by artificial neural networks is bounded. For an artificial neural network with one layer of n sigmoidal nodes, the integrated squared error of approximation, integrating on a bounded subset of d variables, is bounded by c_f'/n, where c_f' depends on a norm of the Fourier transform of the function being approximated. This rate of approximation is achieved under growth constraints on the magnitudes of the parameters of the network. The optimization of a network to achieve these bounds may proceed one node at a time. Because of the economy of number of parameters, order nd instead of n^d, these approximation rates permit satisfactory estimators of functions using artificial neural networks even in moderately high-dimensional problems.

Appendix

In this appendix, equivalent characterizations of the class of functions Γ are given in the context of general Fourier distributions on $\boldsymbol{R}^d$. This appendix is not needed for the proofs of the theorems in the paper. It is intended to supplement the understanding of the class of functions for which the approximation bounds are obtained.

Recall that Γ is defined (in Section III) as the class of functions f on $\boldsymbol{R}^d$ such that $f(x) = f(0) + \int_{\boldsymbol{R}^d}(e^{i\omega\cdot x} - 1)\tilde{F}(d\omega)$ for some complex-valued measure $\tilde{F}(d\omega)$ for which $\int_{\boldsymbol{R}^d} |\omega||\tilde{F}(d\omega)|$. Complex-valued measures take the form $e^{i\theta(\omega)}F(d\omega)$, for some real-valued measure $F(d\omega) = |\tilde{F}(d\omega)|$ called the magnitude distribution and some function $\theta(\omega)$ called the phase (see, for instance, Rudin [29, theorem 6.12]). A complex vector-valued measure $G(d\omega)$ on $\boldsymbol{R}^d$ is a vector of complex-valued measures $(G_1(d\omega), \cdots, G_d(d\omega))$. Let $|G(d\omega)|_1 = \sum_{k=1}^d |G_k(d\omega)|$ denote the sum of the magnitude distributions of the coordinate measures.

Proposition: The following are equivalent for a function f on $\boldsymbol{R}^d$.

a) The gradient of f has the Fourier representation $\nabla f(x) = \int e^{i\omega\cdot x} G(d\omega)$ for some complex vector-valued measure G with $\int |G(d\omega)| < \infty$ and $G(\{0\}) = 0$ (in which case it follows that $G(d\omega) = i\omega\tilde{F}(d\omega)$ for some complex scalar-valued measure $\tilde{F}$).

b) The function f has the representation $f(x) = f(0) + \int (e^{i\omega\cdot x} - 1)\tilde{F}(d\omega)$ for $x \in \boldsymbol{R}^d$, for some complex-valued measure $\tilde{F}$ with $\int |\omega||\tilde{F}(d\omega)| < \infty$.

c) The increments of the function f of the form $f_h(x) = f(x+h) - f(x)$ have a Fourier representation $f_h(x) = \int e^{i\omega\cdot x}(e^{i\omega\cdot h} - 1)\tilde{F}(d\omega)$, $x \in \boldsymbol{R}^d$, for each $h \in \boldsymbol{R}^d$, for some complex-valued measure $\tilde{F}$ with $\int |\omega|\tilde{F}(d\omega)| < \infty$.

If any one of a), b), or c) is satisfied for some $\tilde{F}$, then the other two representations hold with the same $\tilde{F}$.

Proof: The proof of this proposition is as follows. First, recall that $|e^{i\omega\cdot h} - 1|$ is bounded by $2h|\omega|$, so $\int |\omega||\tilde{F}(d\omega)| < \infty$ implies the absolute integrability of the representations in b) and c). Now, b) implies c) since the difference of the integrands at x and $x + h$ is integrable, and c) implies b) by taking a specific choice of x and h; consequently, b) and c) are equivalent. Next, a) follows from c) by the dominated convergence theorem; c) follows from a) by plugging the Fourier representation of the gradient into $f(x+h) - f(x) = \int_0^1 h \cdot \nabla f(x + th)\, dt$ and applying Fubini's theorem.

It remains to show that in a), if the gradient of f has an absolutely integrable Fourier representation $\nabla f(x) = \int e^{i\omega\cdot x} G(d\omega)$, and if G assigns no mass to the point $\omega = 0$, then $G(d\omega)$ is proportional to ω (that is, the measures $(1/\omega_k)G_k(d\omega)$ are the same for $k = 1, 2, \cdots, d$). Now, if the gradient of f has an absolutely integrable Fourier representation, then so do the increments f_h. Indeed, $f_h(x) = \int_0^1 h \cdot \nabla f(x + th)\, dt = \int_0^1 h \cdot \int_{\boldsymbol{R}^d} e^{i\omega\cdot(x+th)} G(d\omega)\, dt$, and integrating first with respect to t yields $f_h(x) = \int_{\boldsymbol{R}^d} e^{i\omega\cdot x}((e^{i\omega\cdot h} - 1)/i\omega\cdot h) h \cdot G(d\omega)$ (the exchange in order of integration is valid by Fubini's theorem since the integral of $e^{it\omega\cdot h}$ is $(e^{i\omega\cdot h} - 1)/i\omega\cdot h$, which has magnitude bounded by 2). Thus, f_h has a Fourier distribution

$$\tilde{F}_h(d\omega) = (e^{i\omega\cdot h} - 1)\frac{h\cdot G(d\omega)}{h\cdot\omega}. \tag{64}$$

It is argued that the factor $h \cdot G(d\omega)/h \cdot \omega$ determines a measure that does not depend on h (from which it follows that $G(d\omega)$ is proportional to ω). Now, the increments of f satisfy $f_h(x + y) = f_{y+h}(x) - f_y(x)$, so it follows that their

Fourier distributions satisfy

$$e^{i\omega\cdot y}\tilde{F}_h(d\omega) = \tilde{F}_{y+h}(d\omega) + \tilde{F}_y(d\omega) \tag{65}$$

for all y, $h \in \boldsymbol{R}^d$. Examination of this identity suggests that $\tilde{F}_h(d\omega)$ must be of the form $(e^{i\omega\cdot h} - 1)\tilde{F}(d\omega)$ for some measure $\tilde{F}$ which does not depend on h. Indeed, by (64), the measures $\tilde{F}_h$ are dominated by $|G|_1$ for all h, so (64) and (65) may be reexpressed as identities involving the densities of these measures with respect to $|G|_1$. Consequently,

$$\begin{aligned} e^{i\omega\cdot y}(e^{i\omega\cdot h} - 1)\frac{h\cdot g(\omega)}{h\cdot\omega} \\ = (e^{i\omega\cdot(y-h)} - 1)\frac{(y+h)\cdot g(\omega)}{(y+h)\cdot\omega} \\ + (e^{i\omega\cdot y} - 1)\frac{y\cdot g(\omega)}{y\cdot\omega}, \end{aligned} \tag{66}$$

where $g(\omega)$ is a complex vector-valued function such that $G(d\omega) = g(\omega)|G(d\omega)|_1$. (For each y and h in $\boldsymbol{R}^d$, (66) holds—except possibly for a set of ω of measure zero with respect to $|G|_1$—so if y and h are restricted to a countable dense set, then there is one $|G|_1$-null set outside of which (66) holds for all such y and h.) Now take a derivative in (66), replacing h with th, dividing both sides by t, and letting $t \to 0$ (along a countable sequence of values with th restricted to the dense set). The identity that results from this derivative calculation, after a rearrangement of the terms, is

$$\omega\cdot h\left(\frac{e^{i\omega\cdot y} - 1}{\omega\cdot y} - ie^{i\omega\cdot y}\right)\left(\frac{h\cdot g(\omega)}{h\cdot\omega} - \frac{y\cdot g(\omega)}{y\cdot\omega}\right) = 0. \tag{67}$$

Therefore, $h\cdot g(\omega)/h\cdot\omega = y\cdot g(\omega)/y\cdot\omega$, whenever $h\cdot\omega$ and $y\cdot\omega$ are not equal to zero (for y and h in the countable dense set and for almost every ω). Let $\rho(\omega) = y\cdot g(\omega)/y\cdot\omega$ denote the common value of this ratio for all such y (for ω outside of the null set). Then, $y\cdot(g(\omega) - \omega\rho(\omega)) = 0$; so taking d points y which span $\boldsymbol{R}^d$, it follows that $g(\omega) = \omega\rho(\omega)$ for almost every ω. Consequently, $G(d\omega) = \omega\rho(\omega)|G(d\omega)|_1$, which may be expressed in the form $G(d\omega) = i\omega\tilde{F}(d\omega)$ for some complex-valued measure $\tilde{F}$ on $\boldsymbol{R}^d$. This completes the proof of the proposition. □

The usefulness of the above proposition is that it provides a Fourier characterization of $\tilde{F}$ for functions in Γ in the case that $\int|\tilde{F}'(d\omega)|$ is not necessarily finite. It is the unique complex-valued measure such that $G(d\omega) = i\omega\tilde{F}(d\omega)$, where G is the Fourier distribution of the gradient of the function. For several of the examples in Section IX of functions in Γ, including sigmoidal functions, the function f does not possess an integrable Fourier representation (in the traditional form $f(x) = \int e^{i\omega\cdot x}\tilde{F}(d\omega)$), but the gradient of f does possess an integrable Fourier representation, and in this way $\tilde{F}$ is determined for the modified Fourier representation $f(x) = f(0) + \int(e^{i\omega\cdot x} - 1)\tilde{F}(d\omega)$.

A Remark Concerning Functions with a Linear Component: If the Fourier distribution G of the gradient of the function f has $G(\{0\}) \neq 0$, that is, if the gradient has a nonzero constant component, then (strictly speaking) the function f is not in Γ. Nevertheless, it is possible to treat this more general situation by using the representation $f(x) = f(0) + a\cdot x + \int(e^{i\omega\cdot x} - 1)\tilde{F}(d\omega)$, where $a = G(\{0\})$, and $\tilde{F}(d\omega)$ is characterized by $G(d\omega) = i\omega\tilde{F}(d\omega)$ on $\boldsymbol{R}^d - \{0\}$. The component $a\cdot x$ is approximated by linear combinations of sigmoidal functions in the same way as the sinusoidal components as in the proof of Theorem 1. Now let $C_{f,B} = \int|G(d\omega)|_B$, where $|G|_B$ is the measure that assigns mass $|G(\{0\})|_B = |a|_B$ at $\omega = 0$, and that equals $|G(d\omega)|_B = |\omega|_B||\tilde{F}(d\omega)|$ when restricted to $\boldsymbol{R}^d - \{0\}$ (recall that, by definition, $|a|_B = \sup_{x\in B}|a\cdot x|$). It can be shown in this context that there is a linear combination of n sigmoidal functions $f_n(x)$ of the form (1), such that the $L_2(\mu, B)$ norm of the error $f - f_n$ is bounded by $2C_{f,B}/\sqrt{n}$. The same bound can also be obtained by the extrapolation method in example (14).

Additional Remarks: In the case that the distribution $\tilde{F}$ has an integrable Fourier density $\tilde{f}(\omega)$, there is a forward transform characterization in terms of Gaussian summability, that is,

$$\tilde{f}(\omega) = \lim_{\epsilon\to 0}(2\pi)^{-d}\int_{\boldsymbol{R}^d} e^{-i\omega\cdot x}f(x)e^{-(\epsilon|x|)^2}\,dx \tag{68}$$

for almost every ω (see, for instance, Stein and Weiss [30]). In the same way, $i\omega\tilde{f}(\omega)$ is determined as the Gauss–Fourier transform of $\nabla f(x)$ for functions in Γ in the case that Fourier distribution of the gradient is absolutely continuous. If $f(x)$ or $\nabla f(x)$, respectively, is integrable on $\boldsymbol{R}^d$, then $\tilde{f}(\omega)$ is determined by an ordinary forward transform, that is, $\tilde{f}(\omega) = (2\pi)^{-d}\int e^{-i\omega\cdot x}f(x)\,dx$ or $i\omega\tilde{f}(\omega) = (2\pi)^{-d}\int e^{-i\omega\cdot x}\nabla f(x)\,dx$ for almost every ω.

Note Added in Proof: A factor of two improvement in the constant in the approximation bound can be obtained.Indeed, if $\phi(z)$ is a given sigmoid with range in [0,1], then subtracting a constant of $1/2$ yields a sigmoid with range in $[-1/2, 1/2$. Allowing for a change in the additive constant c_o, the class of functions represented by superpositions of this new sigmoid is the same as represented by superpositions of the original sigmoid. Therefore, an approximation bound using the new one also is achievable by the original sigmoid. Now the new sigmoid has norm bounded by $1/2$ instead of1. Applying this fact in the proof of Theorem 1 yields the existence of a network function $f_n(x)$ of the form (1) such that

$$\int_B (f(x) - f_n(x))^2\mu(dx) \leq \frac{C_{f,B}^2}{n}. \tag{69}$$

Other scales for the magnitude of the sigmoid are also permitted, including popular choices with for which $\phi(z)$ has limits ± 1 as $z \to \pm\infty$. In that case, the bound (69) holds with the constraint on the coefficients of f_n that $\sum_{k=1}^n |c_k| \leq C$, provided the spectral norm satifies $C_{f,B} \leq C$.

ACKNOWLEDGMENT

The author is grateful for the collaboration of L. Jones. As he points out in [4], together we observed that his theorem applies to artificial neural networks, provided it could be shown that the function to be approximated is in the closure of the convex hull of bounded multiples of sigmoidal functions. This motivated the work presented here.

R. Dudley pointed out [25] where the result of Lemma 1 is attributed to Maurey. E. Sontag noted the necessity of the restriction to continuity points of the distribution for the validity of Lemma 4. B. Hajek suggested the extension to functions on a Hilbert space. H. White encouraged the development of the lower bounds as in Theorem 6, and D. Haussler and J. Zinn contributed to the proof of Lemma 6.

REFERENCES

[1] G. Cybenko, "Approximations by superpositions of a sigmoidal function," *Math. Contr. Signals, Syst.*, vol. 2, pp. 303–314, 1989.

[2] K. Hornik, M. Stinchcombe, and H. White, "Multi-layer feedforward networks are universal approximators," *Neural Networks*, vol. 2, pp. 359–366, 1989.

[3] A. Pinkus, *n-Widths in Approximation Theory*. New York: Springer-Verlag, 1985.

[4] L. K. Jones, "A simple lemma on greedy approximation in Hilbert space and convergence rates for projection pursuit regression and neural network training," *Ann. Statist.*, vol. 20, pp. 608–613, Mar. 1992.

[5] K.-Y. Siu and J. Brunk, "On the power threshold circuits with small weights," *SIAM J. Discrete Math.*, 1991.

[6] N. M. Korobov, *Number Theoretic Methods of Approximate Analysis*. Moscow: Fiznatgiz, 1963.

[7] G. Wahba, *Spline Models for Observational Data*. Philadelphia, PA: SIAM, 1990.

[8] A. R. Barron, "Statistical properties of artificial neural networks," in *Proc. IEEE Int. Conf. Decision Contr.*, Tampa, FL, Dec. 13–15, 1989, pp. 280–285.

[9] ——, "Complexity regularization with applications to artificial neural networks," in *Nonparametric Functional Estimation*, G. Roussas, Ed. Dordrecht, The Netherlands: Kluwer Academic, 1991, pp. 561–576.

[10] A. R. Barron and T. M. Cover, "Minimum complexity density estimation," *IEEE Trans. Inform. Theory*, vol. IT-37, pp. 1034–1053, 1991.

[11] A. R. Barron, "Approximation and estimation bounds for artificial neural networks," in *Proc. 4th Annu. Workshop Computation. Learning Theory*, San Mateo: Morgan Kauffman, Aug. 1991, pp. 243–249. (Also to appear in *Machine Learning*.)

[12] H. White, "Connectionists nonparametric regression: Multilayer feedforward networks can learn arbitrary mappings," *Neural Networks*, vol. 3, no. 5, pp. 535–550, 1990.

[13] D. Haussler, "Decision theoretic generalizations of the PAC model for neural net and other learning applications," Computer Res. Lab. Tech. Rep. UCSC-CRL-91-02. Univ. of California, Santa Cruz, 1991.

[14] A. R. Barron and R. L. Barron, "Statistical learning networks: A unifying view," in *Computing Science and Statistics: Proc. 21st Symp. Interface*, Alexandria: American Statistical Assoc., 1988, pp. 192–203.

[15] L. Breiman, "Hinging hyperplanes for regression, classification, and function approximation," *IEEE Trans. Inform. Theory*, vol. 39, pp. 999–1013, May 1993.

[16] R. Tibshirani, "Slide functions for projection pursuit regression and neural networks," Dep. Stat. Tech. Rep. 9205, Univ. Toronto, 1992.

[17] Y. Zhao, "On projection pursuit learning," Ph.D. dissertation, Dept. Math. Art. Intell. Lab., M.I.T., 1992.

[18] F. Girosi and G. Anzellotti, "Convergence rates of approximation by translates," Art. Intell. Lab. Tech. Rep. 1288, Mass. Inst. Technol., 1992.

[19] A. Farago and G. Lugosi, "Strong universal consistency of neural network classifiers," Dep. Elect. Eng. Tech. Rep., Tech. Univ. Budapest, 1992. (To appear in the *IEEE Trans. Inform. Theory*.)

[20] A. R. Barron, "Neural net approximation," in *Proc. Yale Workshop Adaptive Learning Syst.*, K. Narendra, Ed., Yale Univ., May 1992.

[21] K. Hornik, M. Stinchcombe, H. White, and P. Auer, "Degree of approximation results for feedforward networks approximating unknown mappings and their derivatives," preprint, 1992.

[22] D. F. McCaffrey and A. R. Gallant, "Convergence rates for single hidden layer feedforward networks," Tech. Rep. RAND Corp., Santa Monica, CA, and Dept. Statist., North Carolina State Univ., 1992.

[23] H. N. Mhaskar and C. A. Micchelli, "Approximation by superposition of a sigmoidal function," *Advances in Appl. Math.*, vol. 13, pp. 350–373, 1992.

[24] L. K. Jones, "Good weights and hyperbolic kernels for neural networks, projection pursuit, and pattern classification: Fourier strategies for extracting information from high-dimensional data," Tech. Rep., Dept. Math. Sci., Univ. of Massachusetts, Lowell, 1991. (To appear in *IEEE Trans. Inform. Theory*.)

[25] G. Pisier, "Remarques sur un resultat non publie de B. Maurey," presented at the Seminaire d' analyse fonctionelle 1980–1981, Ecole Polytechnique, Centre de Mathematiques, Palaiseau.

[26] L. K. Jones, "Constructive approximations for neural networks by sigmoidal functions," *Proc. IEEE: Special Issue on Neural Networks*, vol. 78, pp. 1586–1589, 1991.

[27] E. Kushilevitz and Y. Mansour, "Learning decision trees using the Fourier spectrum," in *Proc. 23rd ACM Symp Theory Comput.*, 1991, pp. 455–464.

[28] M. Bellare, "The spectral norm of finite functions," MIT-LCS Tech. Rep. TR-465, 1991.

[29] W. Rudin, *Real and Complex Analysis*. New York: McGraw-Hill, 1984.

[30] E. M. Stein and G. Weiss, *Introduction to Fourier Analysis on Euclidean Spaces*. Princeton, NJ: Princeton Univ. Press, 1971.

[31] C. Darken, M. Donahue, L. Gurvits, and E. Sontag, "Rate of approximation results motivated b robust neural network learning," to appear in *Proc. Sixth ACM Wrokshop on Computat. Learning Theory*.

[32] J. Yukich, "Sup norm approximation bounds for networks via Vapnik–Chervonenkis classes," Notes, Dept. of Math., Lehigh Univ., Bethlehem, PA, Jan. 1993.

[33] V. Kůrková, "Kolmogorov's theorem and multilayer neural networks," *Neural Net.*, vol. 5, pp. 501–506, 1992.