A Note on Error Bounds for Function Approximation Using Nonlinear Networks

Ajit T. Dingankar
IBM Corporation, Austin, TX 78758, U.S.A.
ajit@austin.ibm.com

Irwin W. Sandberg
The University of Texas at Austin, TX 78712, U.S.A.
sandberg@uts.cc.utexas.edu

ABSTRACT — **For a variety of problems concerning classification, compensation, adaptivity, identification or signal processing, results concerning the representation and approximation of nonlinear functions can be of particular interest to engineers. Here we consider a large class of functions f that map $\mathbb{R}^n$ into the set of real or complex numbers, and we give bounds on the number of parameters needed so that f is approximated to within a prescribed degree of accuracy using a certain approximation network. Related work in the neural networks literature is also described.**

I. Introduction

For a variety of problems concerning classification, compensation, adaptivity, identification or signal processing, results concerning the representation and approximation of nonlinear functions can be of particular interest to engineers. Especially pertinent is much of the progress that has been made in recent years in the neural networks area. For example, Cybenko [3] proved that any real-valued continuous map defined on a compact subset of $\mathbb{R}^n$ (n an arbitrary positive integer) can be uniformly approximated arbitrarily well using a neural network with a single hidden layer using any continuous sigmoidal activation function.[1] And subsequently Mhaskar and Micchelli [9] showed that even any nonpolynomial continuous function would do.

The important related problem of determining the number of hidden nodes needed to achieve a given degree of approximation was first addressed by Barron [1] who gives function approximation bounds for neural networks that employ the usual sigmoidal functions. Bounds of $O(1/\sqrt{m})$ for certain L_2 approximations are established, where m is the number of hidden nodes. A lemma due to Jones [8] is used, and a certain first-moment condition on the Fourier transform of the function to be approximated plays a central role. Barron also considers the determination of approximants that satisfy his approximation bounds. This involves the construction of certain approximations to a general element in the closure of the convex hull of a subset of an inner product space. We have observed [4] that there is a related algorithm that generates simpler approximants with somewhat less computational cost. These approximants are certain arithmetic averages (see also [5]). Ref [7] is a study related to [1] and [8] of the rates of convergence of function approximation using translates of a given function.

In this paper we give approximation bounds using an interesting different approximation structure, and which involve arithmetic average approximants. In place of Jones' lemma we develop and use a variant of a lemma due to Maurey [10]. Maurey's lemma asserts the existence of certain convex combinations of functions that serve as approximants. In our variant attention is focused on special combinations, arithmetic averages. This can lead to a reduction in the computation necessary to actually determine approximants, but the problem of computing approximations is not considered in this paper.

The family of functions for which we consider approximations is large, and is described in the next section. (It will become clear that it contains functions not addressed in [1].) Our main result is the theorem given in Section 3.

II. Preliminaries

In the next section we give an error bound for the approximation of certain functions defined on $\mathbb{R}^n$, where n is an arbitrary positive integer. We use S to denote the set of these functions, where

$$S = \{f\colon \mathbb{R}^n \to \mathbb{C} \mid \; f \in L_2(\mathbb{R}^n),\, \mathcal{F}f \in L_1(\mathbb{R}^n)\},$$

$\mathbb{C}$ is the set of complex numbers, $L_1(\mathbb{R}^n)$ and $L_2(\mathbb{R}^n)$ are the usual spaces of Lebesgue integrable

[1] See also [6].

and square integrable functions, respectively, and $\mathcal{F}f$ is the usual $L_2(\mathbb{R}^n)$ limit in the mean Fourier transform [2, pp. 120, 121] of f. As a simple example, consider $f\colon \mathbb{R} \to \mathbb{R}$ given by

$$f(t) = \exp(-|t|).$$

In this case

$$(\mathcal{F}f)(\omega) = 2/(1+\omega^2),$$

and we see that $f \in S$. However, the key requirement in [1] that the function $\omega \mapsto \omega \cdot (\mathcal{F}f)(\omega)$ belongs to $L_1(\mathbb{R})$ is not met, showing that we address a class of f's not treated in [1].

The approximating functions g we consider are described by

$$g(x) = \sum_{k=1}^{m} c_k \exp(i\,\langle x, p_k\rangle), \quad x \in \mathbb{R}^n \tag{1}$$

where m is a positive integer, the c_k belong to $\mathbb{C}$, $\langle\cdot,\cdot\rangle$ is the usual inner product in $\mathbb{R}^n$, and the p_k are drawn from $\mathbb{R}^n$. Our appproximating network thus takes the form shown in Figure 1.[2] For each m, let $A(m)$ denote the set of all g described by (1).

We gauge the degree to which an f is approximated by a g by

$$\int_{\mathbb{R}^n} |f(x) - g(x)|^2 \mu(dx),$$

where μ is a given positive Borel measure on $\mathbb{R}^n$ that satisfies the following conditions: $\mu(\mathbb{R}^n) = 1$, μ is absolutely continuous with respect to Lebesgue measure λ on $\mathbb{R}^n$, and the Radon-Nikodym derivative $[d\mu/d\lambda]$ is bounded (i.e., is a bounded function on $\mathbb{R}^n$).[3] An example of an important useful choice of μ is given by

$$\mu(A) = \lambda(A \cap B_r)/\lambda(B_r), \quad A \in \mathcal{B}(\mathbb{R}^n)$$

where $\mathcal{B}(\mathbb{R}^n)$ is the σ-algebra of Borel subsets of $\mathbb{R}^n$, r is a positive constant, and B_r is the open ball in $\mathbb{R}^n$ of radius r centered at the origin. Of course one might use this μ when the objective is to approximate f on B_r.

In the statement of the theorem in the next section, we use $||f||_\mu{}^2$ to denote

$$\int_{\mathbb{R}^n} |f(x)|^2 \mu(dx)$$

where $f \in S$. To see that this integral is finite, let M be an upper bound on the Radon-Nikodym derivative $[d\mu/d\lambda]$, and note that

$$\int_{\mathbb{R}^n} |f(x)|^2 \mu(dx) =$$

$$\int_{\mathbb{R}^n} |f(x)|^2 [d\mu/d\lambda] \lambda(dx) \le M||f||^2 < \infty,$$

where $||f||$ is the usual norm in $L_2(\mathbb{R}^n)$.

III. Approximation Theorem

Our main result is the following theorem which says, roughly speaking, that a network with $2m$ parameters achieves an error bound of $O(1/\sqrt{m})$ (the parameters in Figure 1 are the c_k and the p_k).

Theorem. Let f be an element of S, and let c satisfy $c > ||f||_\mu$ as well as

$$\frac{1}{(\sqrt{2\pi})^n} \int |(\mathcal{F}g)(\omega)| \lambda(d\omega) \le c.$$

Then for each m there exists an $f_m \in A(m)$ such that

$$\int_{\mathbb{R}^n} |f(x) - f_m(x)|^2 \mu(dx) \le \frac{c^2 - ||f||_\mu{}^2}{m}.$$

Proof.

We use a variant of Maurey's lemma [10]. Our lemma given below shows that functions that have a certain integral representation can be approximated, by certain arithmetic averages of functions, with an error bound along the lines of the conclusion of the theorem. While the proof (omitted here due to space limitations) uses ideas described in [1], the basic setting is somewhat different from that in [1] in that, for example, here inner product spaces do not play a central role.[4]

Lemma. Let X be a nonempty set, and let c be a positive constant. Let two positive measure

[2] We use the same structure as in [12], where nodes belonging to the single hidden layer implement exponentiation and the input nodes compute an inner product in $\mathbb{R}^n$. It is interesting that this network can be used for approximation both in the uniform norm for continuous functions on compact sets [12] and (as we show here) in an L_2 norm for functions defined on the whole space.

[3] $[d\mu/d\lambda]$, which exists because of the absolute continuity of μ, is measurable with respect to λ and nonnegative [11, p. 238].

[4] The lemma is a more general result than is needed here. It may be of interest in connection with other studies.

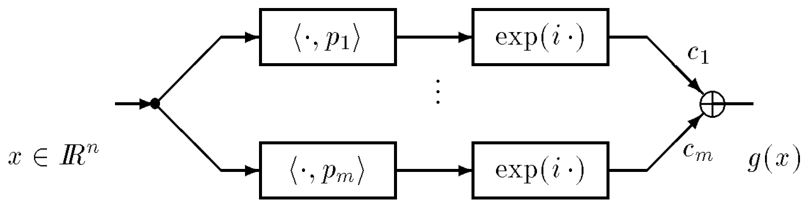

Figure 1: Approximating network.

spaces (X, Σ_1, ν_1) and (X, Σ_2, ν_2) be given such that $\nu_1(X) = \nu_2(X) = 1$. Let $d: X \times X \to \mathbb{C}$ be such that $|d(x, \omega)| \le c$ for all x and almost all ω, and assume that d is measurable with respect to $\Sigma_1 \times \Sigma_2$. Let $f \in L_2(X, \nu_1)$ (the space of square-integrable $\mathbb{C}$-valued functions on X with respect to the measure ν_1, with the usual norm which we denote by $||\cdot||_{\nu_1}$) be a function with the integral representation

$$f(x) = \int_X d(x, \omega) \nu_2(d\omega) \tag{2}$$

for almost all $x \in X$ (with respect to ν_1). Then for each m there exists an arithmetic average f_m of m elements of $\{d(\cdot, \omega) \mid \omega \in X\}$ such that

$$\int_X |f(x) - f_m(x)|^2 \nu_1(dx) \le \frac{c^2 - ||f||_{\nu_1}{}^2}{m}.$$

We now prove our theorem. We can certainly approximate f as required if f is the zero element of $L_2(\mathbb{R}^n)$. Thus assume that f is not the zero element of $L_2(\mathbb{R}^n)$.

Since $f \in S$, f is the inverse Fourier transform of $\mathcal{F}f$. This inverse Fourier transform is an $L_2(\mathbb{R}^n)$ limit in the mean. But since $\mathcal{F}f \in L_1(\mathbb{R}^n)$, it follows that the inverse transform coincides almost everywhere with the function

$$\frac{1}{(\sqrt{2\pi})^n} \int_{\mathbb{R}^n} \exp(i\langle \omega, \cdot \rangle)(\mathcal{F}f)(\omega)\lambda(d\omega).^5$$

[5] To see this, recall that f is the limit in $L_2(\mathbb{R}^n)$ of the usual sequence $f_1, f_2, \ldots$ of functions [2, pp. 120, 121]. It follows that this sequence has a subsequence $f_{k_1}, f_{k_2}, \ldots$ that converges almost everywhere to f. And since $\mathcal{F}f \in L_1(\mathbb{R}^n)$ the subsequence converges everywhere to integral given above.

Thus,

$$f(x) = \frac{1}{(\sqrt{2\pi})^n} \int_{\mathbb{R}^n} \exp(i\langle \omega, x \rangle)(\mathcal{F}f)(\omega)\lambda(d\omega)$$

for almost all x.

Writing

$$(\mathcal{F}f)(\omega) = |(\mathcal{F}f)(\omega)| \exp(i\theta(\omega)),$$

we have

$$f(x) = \int_{\mathbb{R}^n} d(x, \omega)\nu_2(d\omega) \quad \text{(a.e.)}$$

where

$$d(x, \omega) = c_f \exp(i[\langle \omega, x \rangle + \theta(\omega)]),$$

$$c_f = \frac{1}{(\sqrt{2\pi})^n} \int |(\mathcal{F}f)(\omega)|\lambda(d\omega)\ ^6$$

and

$$\nu_2(d\omega) = \frac{|(\mathcal{F}f)(\omega)|\lambda(d\omega)}{(\sqrt{2\pi})^n c_f}.$$

Note that ν_2 is a positive measure such that $\nu_2(\mathbb{R}^n) = 1$.

We have already observed (in Section 2) that $f \in L_2(\mathbb{R}^n, \mu)$ (notation of the lemma). Note that $|d(x, \omega)|$ is bounded by c for all x and almost all ω. Thus, by our lemma (with $X = \mathbb{R}^n$ and $\nu_1 = \mu$), for each m there exists an arithmetic average f_m of m points of $\{c_f \exp(i[\langle \omega, x \rangle + \theta(\omega)]), \omega \in \mathbb{R}^n\}$ such that

$$\int_{\mathbb{R}^n} |f(x) - f_m(x)|^2 \mu(dx) \le \frac{c^2 - ||f||_{\mu}{}^2}{m}.$$

[6] Since here f is not the zero element of $L_2(\mathbb{R}^n)$, it follows that c_f is not zero.

This completes the proof.

Comments

An example of peripheral results that add to the usability of the theorem is the following. Let Ω be a positive number and let f belong to $L_2(\mathbb{R}^n)$ and be bandlimited to band $[-\Omega, \Omega]^n$ (i.e., and be such that $(\mathcal{F}f)(\omega) = 0$ for almost all $\omega \notin [-\Omega, \Omega]^n$). Then [by $\mathcal{F}f \in L_2(\mathbb{R}^n)$ and the Schwarz inequality] $f \in S$ and, with regard to the inequality

$$\frac{1}{(\sqrt{2\pi})^n} \int |(\mathcal{F}g)(\omega)| \lambda(d\omega) \leq c,$$

of the theorem one has

$$\frac{1}{(\sqrt{2\pi})^n} \int |(\mathcal{F}g)(\omega)| \lambda(d\omega) \leq (\Omega/\pi)^{n/2} \; ||f||,$$

which is provable using Schwarz's inequality and Plancherel's identity.[7]

For related function bounds in the setting of convex hulls of bounded subsets of an inner product space, see [4].

Real-valued f's

With regard to the conclusion of the theorem, for the case in which $f \in S$ is real valued the approximating functions f_m can be taken to be given instead by

$$f_m(x) = \sum_{k=1}^{m} a_k \cos(\langle x, p_k \rangle + b_k), \quad x \in \mathbb{R}^n \qquad (3)$$

in which the a_k and b_k are real constants and the p_k belong to $\mathbb{R}^n$. This follows at once from the fact that the real part of $c_k \exp(i \langle x, p_k \rangle)$ can be written in the form $a_k \cos(\langle x, p_k \rangle + b_k)$.

REFERENCES

[1] A. R. Barron, Universal approximation bounds for superpositions of a sigmoidal function, *IEEE Transactions on Information Theory*, vol. 39, no. 3, pp. 930–945, May 1993.

[2] S. Bochner and K. Chandrasekharan, *Fourier Transforms*, Princeton University Press, 1949.

[3] G. Cybenko, Approximation by superposition of a single function, *Mathematics of Control, Signals and Systems*, vol. 2, pp. 303–314, 1989.

[4] A. T. Dingankar and I. W. Sandberg, A note on error bounds for approximation in inner product spaces, *Circuits, Systems and Signal Processing*, vol. 15, no. 4, pp. 519–522, 1996.

[5] D. Docampo, C. T. Abdallah and D. Hush, Error bounds in constructive approximation, *Proceedings of the Fourth Bayona Workshop on Intelligent Methods in Signal Processing and Communications*, Bayona, Spain, June 1996.

[6] K. Funahashi, On the approximate realization of continuous mappings by neural networks, *IEEE Transactions on Neural Networks*, vol. 2, pp. 183–192, 1989.

[7] F. Girosi and G. Anzellotti, Rates of convergence for radial basis functions and neural networks, *Artificial Neural Networks with Applications in Speech and Vision*, Chapman & Hall, 1993.

[8] L. K. Jones, A simple lemma on greedy approximation in Hilbert space and convergence rates for projection pursuit regression and neural network training, *Annals of Statistics*, vol. 20, no. 1, pp. 608–613, March 1992.

[9] H. N. Mhaskar and C. A. Micchelli, Approximation by superposition of sigmoidal and radial basis functions, *Advances in Applied Mathematics*, vol. 3, pp. 350–373, 1992.

[10] G. Pisier, Remarques sur un resultat non publié de B. Maurey, *Sem. d'Analyse Fonctionelle*, vol. 1, no. 12, Ecole Poly. Centre de Math., Palaiseau, 1980–1981.

[11] H. L. Royden, *Real Analysis*, Macmillan, New York, 1968.

[12] I. W. Sandberg, Approximations for nonlinear functionals, *IEEE Transactions on Circuits and Systems-I. Fundamental Theory and Applications*, vol. 39, no. 1, pp. 65–67, January 1992.

[7] This inequality is well known, at least for $n = 1$.