Journal of Approximation Theory **104**, 183–203 (2000)
doi:10.1006/jath.1999.3441, available online at http://www.idealibrary.com on IDEAL®

A Better Approximation for Balls

Gerald H. L. Cheang

Division of Mathematics, School of Science, National Institute of Education, Nanyang Technological University, 469, Bukit Timah Road, Singapore 259756
E-mail: cheanghlg@nie.edu.sg

and

Andrew R. Barron

Department of Statistics, Yale University, P.O. Box 208290, New Haven, Connecticut 06520-8290
E-mail: Andrew.Barron@yale.edu

Communicated by Allan Pinkus

Received May 3, 1999; accepted in revised form October 22, 1999

Unexpectedly accurate and parsimonious approximations for balls in $\mathscr{R}^d$ and related functions are given using half-spaces. Instead of a polytope (an intersection of half-spaces) which would require exponentially many half-spaces (of order $(\frac{1}{\varepsilon})^d$) to have a relative accuracy ε, we use $T = c(d^2/\varepsilon^2)$ pairs of indicators of half-spaces and threshold a linear combination of them. In neural network terminology, we are using a single hidden layer perceptron approximation to the indicator of a ball. A special role in the analysis is played by probabilistic methods and approximation of Gaussian functions. The result is then applied to functions that have variation V_f with respect to a class of ellipsoids. Two hidden layer feedforward sigmoidal neural nets are used to approximate such functions. The approximation error is shown to be bounded by a constant times $V_f/T_1^{1/2} + V_f\, d/T_2^{1/4}$, where T_1 is the number of nodes in the outer layer and T_2 is the number of nodes in the inner layer of the approximation f_{T_1, T_2}.

1. INTRODUCTION

There already exists a rich literature on approximation of convex bodies with other sorts of convex bodies and polytopes. See, for example, Gruber [7], Fejes Tóth [6]. Like other convex bodies, a ball is an infinite interesection of tangent half-spaces. For a unit ball B in $\mathscr{R}^d$,

$$B = \bigcap_{a \in S^{d-1}} \{a \cdot x \leqslant 1\}, \tag{1}$$

0021-9045/00 $35.00

where S^{d-1} is the unit sphere in $\mathscr{R}^d$. If we approximate it with the intersection of T, $T \geqslant d+1$, of the half-spaces in (1), then we are approximating the ball with a T-faced polytope $\mathscr{P}_T$.

There are results that bound the approximation error between convex bodies and their polytope approximators. Dudley [5] has shown that for each convex body B, there exists a constant c such that for every T there is a polytope $\mathscr{P}_T$ achieving

$$\delta^H(B, \mathscr{P}_T) \leqslant \frac{c}{T^{2/(d-1)}}, \tag{2}$$

where δ^H is the Hausdorff metric. Results from Schneider and Wieacker [17] and Gruber and Kenderov [8] have shown that for a convex body with sufficiently smooth boundary such as the ball B, there exists a constant c such that for every polytope $\mathscr{P}_T$,

$$\delta(B, \mathscr{P}_T) \geqslant \frac{c}{T^{2/(d-1)}}, \tag{3}$$

where δ can be either the Hausdorff or the Lebesgue measure of the symmetric difference. Hence for an approximation error of ε, we would require a polytope with many faces of order $(\frac{1}{\varepsilon})^{(d-1)/2}$, which is exponential in d. To avoid this curse of dimensionality, we will use T half-spaces in the approximation in a different manner.

To illustrate the idea, consider the set of points in at least k out of n given half-spaces. For instance, if we were given the $T=9$ half-spaces determining the polygon approximation in Fig. 1, $k=9$ yields the nonagon

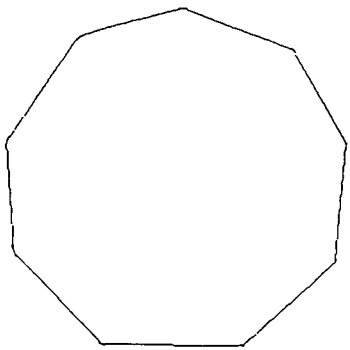

FIGURE 1

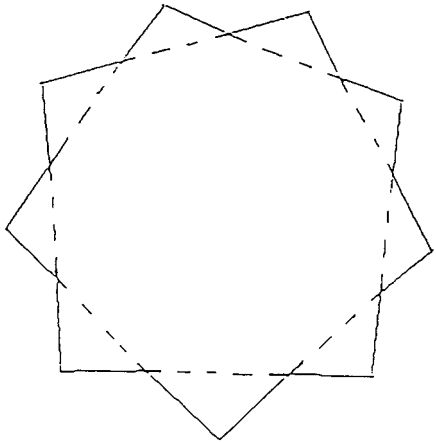

FIGURE 2

inscribed in the circle. In Fig. 2, we use $T=9$ half-spaces, but we set the threshold at $k=8$ to obtain the star-shaped approximation shown. In higher dimensions, our approximation will look somewhat like a jagged multi-faceted star-shaped object.

Here we can think of the T half-spaces as providing a test for membership in the set. Instead of requiring all T tests to be passed, we permit membership with at least k passed out of T. An extension of this idea is to weigh each test and determine membership by a weighted count exceeding a threshold.

While polygon approximation may appear superior in the low-dimensional example given in the figure, in high dimensions, polytopes have extremely poor accuracy as shown in (3). In contrast we show that the use of a weighted count to determine membership in a set permits accuracy that avoids the curse of dimensionality. Indeed, with $2T=cd^2/\varepsilon^2$ indicators of half-spaces, where c is a constant, we threshold a linear combination of them, in order to obtain accuracy ε. Note that the number of indicators of half-spaces needed is only quadratic in d and not exponential in d as in the classical method.

Our approximation to a ball takes the form

$$\mathcal{N}_{2T}=\left\{x\in\mathcal{R}^d:\sum_{i=1}^{2T}c_i 1_{\{a_i\cdot x\geqslant b_i\}}\geqslant k\right\}.$$

Let $\tilde{f}_{2T}=1_{\mathcal{N}_{2T}}$ be the indicator (characteristic) function of this set. In neural network terminology, we are using a two hidden layer perceptron approximation to the indicator of a ball. We show that there is a constant

c such that for every T and d, there is such an approximation $\mathcal{N}_{2T}$ that the Hausdorff distance between a ball B_R of radius R and $\mathcal{N}_{2T}$ satisfies

$$\delta^H(B_R, \mathcal{N}_{2T}) \leqslant c \cdot R \sqrt{\frac{d(d+1)}{T}},$$

where c is some real-valued constant. A special role in the analysis is played by probabilistic methods and approximation of Gaussian functions.

2. SOME BACKGROUND AND THE GAUSSIAN FUNCTION

A single hidden-layer feedforward sigmoidal network is a family of real-valued functions $f_T(x)$ of the form

$$f_T(x) = \sum_{i=1}^{T} c_i \phi(a_i \cdot x + b_i) + k, \qquad x \in \mathcal{R}^d \tag{4}$$

parametrized by internal weight vectors a_i in $\mathcal{R}^d$, internal location parameters b_i in $\mathcal{R}$, external weights c_i and a constant term k (Cybenko [4] and Haykin [9]). By a sigmoidal function, we mean any nondecreasing function on $\mathcal{R}$ with distinct finite limits at $+\infty$ and $-\infty$. Such a network has d inputs, T hidden nodes and a linear output unit. It implements ridge-functions $\phi(a_i \cdot x + b_i)$ on the nodes in the hidden layer. Here we will exclusively use the Heaviside function $\phi(z) = 1_{\{z \geqslant 0\}}$, in which case (4) is a linear combination of indicators of half spaces. Such a network is also called a perceptron network (Rosenblatt [15, 16]). Thresholding the output of a single hidden-layer neural net at level k_1, we obtain $\tilde{f}_T(x) = \phi(f_T(x) - k_1)$ which equals

$$\tilde{f}_T(x) = \phi\left(\sum_{i=1}^{T} c_i \phi(a_i \cdot x + b_i) + k'\right). \tag{5}$$

For simplicity in the notation, we will often omit the parameters a_i, b_i, c_i, and k in the arguments of f_T and $\tilde{f}_T$.

To approximate a ball we first consider approximation of the Gaussian function $f(x) = \exp(-|x|^2/2)$ and then take level sets. A level set of a function f at level k is simply the set $\{x \in \mathcal{R}^d : f(x) \geqslant k\}$. Using the fact that the Gaussian is a positive definite function with Fourier transform $(2\pi)^{-d/2} \exp(-|\omega|^2/2)$, so that f has a representation in the convex hull of sinusoids, it is known that $f(x)$ can be expressed using the convex hull of indicators of half-spaces (see Barron [1, 2], Hornik *et al.* [11], Yukich *et al.* [19]). We take advantage of a similar representation here. We use $|\cdot|$ to denote the Euclidean $\mathcal{L}_2$ norm.

Let B_K be a ball of radius K large enough that it would contain B and $\mathcal{N}_{2T}$. As shown in Appendix A, on B_K the Gaussian function satisfies

$$f(x) = \int_{\mathcal{R}^d} \int_{-|a|\,K}^{|a|\,K} 1_{\{a\cdot x+b\geqslant 0\}} \sin(b) \frac{\exp(-|a|^2/2)}{(2\pi)^{d/2}}\, db\, da + \exp\left(-\frac{K^2}{2}\right). \tag{6}$$

Here $\exp(-K^2/2)$ is the value of the Gaussian evaluated on the surface of the ball B_K. As we will see later, we can arrange for the neural net level set $\mathcal{N}_{2T}$ to be entirely contained in B and hence take $K=1$.

Decomposing the integral representation of f into positive and negative parts, we have

$$f(x) - \exp\left(-\frac{K^2}{2}\right) = f_1(x) - f_2(x) \tag{7}$$

$$= \int_{\mathcal{R}^d} \int_{-|a|\,K}^{|a|\,K} 1_{\{a\cdot x+b\geqslant 0\}} \sin^+(b) \frac{\exp(-|a|^2/2)}{(2\pi)^{d/2}}\, db\, da$$

$$- \int_{\mathcal{R}^d} \int_{-|a|\,K}^{|a|\,K} 1_{\{a\cdot x+b\geqslant 0\}} \sin^-(b) \frac{\exp(-|a|^2/2)}{(2\pi)^{d/2}}\, db\, da$$

$$= v_1 \int 1_{\{a\cdot x+b\geqslant 0\}}\, dV_1 - v_2 \int 1_{\{a\cdot x+b\geqslant 0\}}\, dV_2, \tag{8}$$

where V_1 is the probability measure for (a, b) on $\mathcal{R}^d$ with density $1_{\{-|a|\,K<b<|a|\,K\}} \sin^+(b)(\exp(-|a|^2/2)/(2\pi)^{d/2}\, v_1)$ with normalizing constant

$$v_1 = \int_{\mathcal{R}^d} \int_{-|a|\,K}^{|a|\,K} \sin^+(b) \frac{\exp(-|a|^2/2)}{(2\pi)^{d/2}}\, db\, da$$

and similarly for V_2 and v_2 (with $\sin^-(b)$ in place of $\sin^+(b)$). We denote the positive part of $\sin(b)$ by $\sin^+(b)$ and the negative part of $\sin(b)$ by $\sin^-(b)$. The total variation of the measure used to represent f is

$$v = v_1 + v_2$$

$$= \int_{\mathcal{R}^d} \int_{-|a|\,K}^{|a|\,K} |\sin(b)| \frac{\exp(-|a|^2/2)}{(2\pi)^{d/2}}\, db\, da$$

$$< \int_{\mathcal{R}^d} \frac{2\,|a|\,K \exp(-|a|^2/2)}{(2\pi)^{d/2}}\, da \tag{9}$$

$$\leqslant 2K\sqrt{d}. \tag{10}$$

An integral representation of the Gaussian as an expected value invites Monte Carlo approximation by a sample average. In particular, both $f_1(x)$ and $f_2(x)$ in (7) are expected values of indicators of half-spaces in $\mathscr{R}^d$. Thus a $2T$-term neural net approximation to $f(x)$ is then

$$f_{2T}(x) - \exp\left(-\frac{K^2}{2}\right) = \frac{v_1}{T} \sum_{i=1}^{T} \phi(a_i \cdot x + b_i) - \frac{v_2}{T} \sum_{i=T+1}^{2T} \phi(a_i \cdot x + b_i), \tag{11}$$

where the parameters $(a_i, b_i)_{i=1}^{T}$ are drawn at random independently from the distribution V_1 and $(a_i, b_i)_{i=T+1}^{2T}$ from V_2. The sampling scheme is simple. For example, to obtain an approximation for $f_1(x)$, first draw a from a standard multi-variate normal distribution over $\mathscr{R}^d$, then draw b from $[-|a|\,K, |a|\,K]$ with density proportional to $\sin^+(b)$.

We now bound the $\mathscr{L}_\infty$ approximation error between $f(x)$ and $f_{2T}(x)$. We will draw on symmetrization techniques and the concept of Orlicz norms in empirical process theory (see, for example, Pollard [13]), and the theory of Vapnik–Červonenkis classes of sets (Vapnik and Červonenkis [18]). With the particular choice of $\Psi(x) = \frac{1}{5}\exp(x^2)$ used by Pollard [13], the Orlicz norm of a random variable Z is defined by

$$\|Z\|_\Psi = \inf\left\{C > 0 : E \exp\left(\frac{Z^2}{C^2}\right) \leqslant 5\right\}.$$

We examine the approximation error between $f_1(x)$ and $f_{1,T}(x)$, its T-term neural net approximation, first.

From empirical process theory, the following lemma is obtained. See Appendix B.

LEMMA 1. *Let $\xi = (a, b)$ and $g_x(\xi) = \phi(a \cdot x + b)$. If $h(x) = \int \phi(a \cdot x + b)\, P(da, db)$ for $x \in B_K$ for some probability measure P on (a, b), then there exist $\xi_1, \xi_2, \ldots, \xi_T$ such that*

$$\sup_{x \in B_K} \left| \frac{1}{T} \sum_{i=1}^{T} g_x(\xi_i) - h(x) \right| \leqslant 34 \sqrt{\frac{d+1}{T}}. \tag{12}$$

Recall that for the approximation of the Gaussian function, the approximation f_{2T} less $\exp(-K^2/2)$ can be split up into two parts, $f_{1,T}(x) = (v_1/T)\sum_{i=1}^{T} g_x(\xi_i)$ and $f_{2,T}(x) = (v_2/T)\sum_{i=T+1}^{2T} g_x(\xi_i)$, which

approximate the positive and negative parts f_1 and f_2 respectively. Using Lemma 1, we see that

$$\sup_{x \in B_K} \left| \frac{v_1}{T} \sum_{i=1}^{T} g_x(\xi_i) - f_1(x) \right| \leqslant 34 v_1 \sqrt{\frac{d+1}{T}} \tag{13}$$

and similarly,

$$\sup_{x \in B_K} \left| \frac{v_2}{T} \sum_{i=T+1}^{2T} g_x(\xi_i) - f_2(x) \right| \leqslant 34 v_2 \sqrt{\frac{d+1}{T}}. \tag{14}$$

Hence by the triangle inequality,

$$\begin{aligned} \sup_{x \in B_K} |f_{2T}(x) - f(x)| &\leqslant 34(v_1 + v_2) \sqrt{\frac{d+1}{T}} \\ &= 34 v \sqrt{\frac{d+1}{T}} \\ &\leqslant 68 K \sqrt{\frac{d(d+1)}{T}}. \end{aligned} \tag{15}$$

3. BOUNDING THE HAUSDORFF DISTANCE OF THE APPROXIMATION

The Hausdorff distance between two sets F and G is defined as

$$\delta^H(F, G) = \max\{ \sup_{x \in F} \inf_{y \in G} |x - y|, \sup_{y \in G} \inf_{x \in F} |x - y| \}.$$

The norm $|\cdot|$ is the usual Euclidean norm in $\mathscr{R}^d$. We bound the Hausdorff distance between the ball and its approximating set $\delta^H(B, \mathscr{N}_{2T})$ in this section. The ball is assumed to be centered at the origin. However, we apply the result later to other balls and ellipsoids that are not necessarily centered at the origin. Note that the unit ball B in $\mathscr{R}^d$ may be represented as

$$B = \left\{ x : \exp\left(-\frac{|x|^2}{2} \right) \geqslant \exp\left(-\frac{1}{2} \right) \right\}.$$

We define $\mathscr{N}_{2T}$ as

$$\mathscr{N}_{2T} = \{ x : f_{2T}(x) \geqslant \exp(-\tfrac{1}{2}) + K\varepsilon_T \}.$$

Let $f(x) = \exp(-|x|^2/2)$ and let $f_{2T}(x)$ be the approximation with T pairs of indicators. Here

$$\varepsilon_T := 68\sqrt{\frac{d(d+1)}{T}},$$

for which we have the $\mathscr{L}_\infty$ error between the Gaussian and its approximant bounded above by

$$\sup_{x \in B_K} |f_{2T}(x) - f(x)| \leqslant \varepsilon_T. \tag{16}$$

We are going to bound the Hausdorff distance between B and $\mathscr{N}_{2T}$, using this sup norm bound on the error between the functions f and f_{2T} which yield B and $\mathscr{N}_{2T}$ as level sets.

THEOREM 1. *Let B_R be a ball of radius R in $\mathscr{R}^d$ centered at the origin, and let $\mathscr{N}_{2T}$ be the level set of the neural net approximation. For sufficiently large T, such that $\varepsilon_T \leqslant \frac{1}{2}(\exp(-\frac{1}{4}) - \exp(-\frac{1}{2}))$,*

$$\delta^H(B_R, \mathscr{N}_{2T}) \leqslant 318R\sqrt{\frac{d(d+1)}{T}}.$$

Proof. The ball B coincides with the level set of f at the level $\exp(-\frac{1}{2})$. Let T be such that ε_T is less than $\frac{1}{2K}\exp(-\frac{1}{2})$. Choose r_0 such that $\exp(-r_0^2/2) = \exp(-\frac{1}{2}) + 2K\varepsilon_T$. Let B_{r_0} be the ball of radius r_0 centered at the origin. If $x \in \mathscr{N}_{2T}$, then $\exp(-\frac{1}{2}) \leqslant f_{2T}(x) - K\varepsilon_T \leqslant \exp(-|x|^2/2)$ which implies that $x \in B$. Similarly if $x \in B_{r_0}$, then $\exp(-\frac{1}{2}) + K\varepsilon_T \leqslant \exp(-|x|^2/2) - K\varepsilon_T \leqslant f_{2T}(x)$, which implies that $x \in \mathscr{N}_{2T}$. Thus

$$B_{r_0} \subset \mathscr{N}_{2T} \subset B.$$

Both B and its approximating set $\mathscr{N}_{2T}$ are sandwiched between B_{r_0} and B. Consequently

$$\delta^H(B, \mathscr{N}_{2T}) \leqslant 1 - r_0.$$

The function $g(r) = \exp(-r^2/2)$ has derivative $-rg(r)$ of magnitude which is largest at $r = 1$. Now

$$\begin{aligned} r_0 &= \sqrt{2\log(1/(e^{-1/2} + 2K\varepsilon_T))} \\ &= \sqrt{1 - 2\log(1 + 2K\varepsilon_T e^{1/2})}, \end{aligned}$$

which is close to 1. If T is large enough that ε_T is less than $\frac{1}{2K}(\exp(-\frac{1}{4}) - \exp(-\frac{1}{2}))$, then $r_0 g(r_0) \geqslant (1/\sqrt{2})\exp(-\frac{1}{2})$, and hence using the mean-value theorem

$$\begin{aligned}\delta^H(B, \mathcal{N}_{2T}) &\leqslant 1 - r_0 \\ &\leqslant \frac{g(r_0) - g(1)}{r_0\, g(r_0)} \\ &\leqslant 2\sqrt{2e}\, K\varepsilon_T \\ &\leqslant 136\sqrt{2e}\, K\sqrt{\frac{d(d+1)}{T}}. \end{aligned} \tag{17}$$

Now we set K. From Section 2, B_K need only be large enough to cover both the unit ball B and its approximation set $\mathcal{N}_{2T}$, which we have arranged to be contained in B. Thus we can take B_K to be B, whence $K = 1$. Again when $\varepsilon_T \leqslant \frac{1}{2}(\exp(-\frac{1}{4}) - \exp(-\frac{1}{2}))$, we have

$$\delta^H(B, \mathcal{N}_{2T}) \leqslant 136\sqrt{2e}\sqrt{\frac{d(d+1)}{T}} \leqslant 318\sqrt{\frac{d(d+1)}{T}}. \tag{18}$$

For a ball B_R of radius R, the Hausdorff distance between it and its approximation set is simply $318R\sqrt{\frac{d(d+1)}{T}}$. This concludes the proof of Theorem 1. ∎

4. AN $\mathscr{L}_1$ BOUND

Let B_R be a ball of radius R, $\mathcal{N}_{2T}$ the level set induced by the approximation as explained in Section 1, μ is the Lebesgue measure, and δ is the Hausdorff distance between B_R and $\mathcal{N}_{2T}$ as obtained above. Since the symmetric difference $B_R \triangle \mathcal{N}_{2T}$ is included in the shell $B_{R+\delta} \setminus B_{R-\delta}$, one has

$$\begin{aligned}\int |1_{B_R} - 1_{\mathcal{N}_{2T}}| \frac{\mu(dx)}{\mu(B_R)} &= \frac{\mu(B_R \triangle \mathcal{N}_{2T})}{\mu(B_R)} \\ &\leqslant \frac{\mu(B_R) - \mu(B_{R-\delta})}{\mu(B_R)} \\ &\leqslant 1 - \left(1 - \frac{\delta}{R}\right)^d \\ &\leqslant d\frac{\delta}{R} \\ &\leqslant 318d\sqrt{\frac{d(d+1)}{T}}, \end{aligned} \tag{19}$$

thus the following theorem is established.

THEOREM 2. *The relative Lebesgue measure of the symmetric difference* $\mu(B_R \triangle \mathcal{N}_{2T})/\mu(B_R)$ *between* B_R *and its approximation set* $\mathcal{N}_{2T}$ *is bounded above by*

$$\frac{\mu(B_R \triangle \mathcal{N}_{2T})}{\mu(B_R)} \leqslant 318d\sqrt{\frac{d(d+1)}{T}}.$$

5. ELLIPSOID APPROXIMATION

Consider an ellipsoid $E = \{x : x'Mx \leqslant 1\}$ centered at the origin with $M = A'A$ strictly positive definite with a $d \times d$ positive definite square root A. Equivalently $E = \{x : \exp(-x'A'Ax/2) \geqslant \exp(-1/2)\}$ is the level set of a Gaussian surface. In a similar manner to the ball, it can also be accurately and parsimoniously approximated by a single hidden-layer neural net. Let the eigenvalues of A be $r_1 \leqslant r_2 \leqslant \cdots \leqslant r_d$ with the corresponding eigenvectors $\{\mathbf{r_1}, \mathbf{r_2}, ..., \mathbf{r_d}\}$. If the approximating set for the unit ball takes the form $\{x : \sum_{i=1}^{2T} c_i 1_{\{a_i \cdot x \geqslant b_i\}} \geqslant k\}$, then the one for the ellipsoid E is

$$E_{2T} = \left\{x : \sum_{i=1}^{2T} c_i 1_{\{a_i \cdot Ax \geqslant b_i\}} \geqslant k\right\}.$$

We are interested in bounding $\delta^H(E, E_{2T})$, the Hausdorff distance between the ellipsoid and its approximating set.

THEOREM 3. *The Hausdorff distance between the ellipsoid* E *and its approximating set* E_{2T} *is bounded above by*

$$318r_d\sqrt{\frac{d(d+1)}{T}}. \tag{20}$$

Proof. The matrix transformation A transforms the unit ball to an ellipsoid by stretching the unit radius to length r_i in the $\mathbf{r}_i$ direction and the approximating set $\mathcal{N}_T$ is similarly stretched in the same way to E_{2T}. For the ball B_{r_0} (as defined in the proof of Theorem 1), the matrix transformation A transforms it to an ellipsoid E' by stretching its radius to length $r_i r_0$ in the $\mathbf{r}_i$ direction. Thus the order of inclusivity is still preserved after the transformation and

$$E' \subset E_{2T} \subset E.$$

Note that the ellipsoids E and E' are similar, centered at the origin and aligned along the same axes. The only difference is in the scale.

The two extreme parts of the ellipsoids are along the direction $\mathbf{r}_1$ and $\mathbf{r}_d$. The ellipsoid is contained in a ball of radius r_d. Thus $\delta^H(E, E_{2T})$ is bounded above by the greatest distance between E and E', and this occurs along the direction of $\mathbf{r}_d$, and hence is bounded above by the Hausdorff distance between that of a ball of radius r_d (containing the ellipsoid) and a ball of radius $r_d r_0$, and that is in turn bounded above by

$$318 r_d \sqrt{\frac{d(d+1)}{T}}. \quad \blacksquare$$

The error is the same as for approximation of a ball except that the radius of the ball is replaced by the maximal eigenvalue (length of major axis).

Now consider an ellipsoid E with axial lengths $r_1 \leqslant \cdots \leqslant r_{d-1} \leqslant r_d = R$ and its approximating set E_{2T}. The ellipsoid $E^\delta = (1 - \frac{\delta}{R})\, E$ is a scaled down version of E and it has axial lengths $r_1(1 - \frac{\delta}{R}) \leqslant \cdots \leqslant r_{d-1}(1 - \frac{\delta}{R}) \leqslant r_d(1 - \frac{\delta}{R}) = R - \delta$. Recall that the approximation set E_{2T} is obtained by scaling $\mathcal{N}_{2T}$ (the approximation set for the unit ball) by a factor of r_i along the ith axis of the ellipsoid E. The Hausdorff distance between E and E_{2T} is δ which is bounded by $318R\sqrt{\frac{d(d+1)}{T}}$ from Theorem 3.

COROLLARY 1. *The measure of the symmetric difference $\mu(E \triangle E_{2T})$ between E and its approximation set E_{2T} is bounded above by*

$$\mu(E \triangle E_{2T}) \leqslant 318 \mu(E)\, d \sqrt{\frac{d(d+1)}{T}}.$$

Proof. Since the difference $E \triangle E_{2T}$ is included in the shell $E \backslash E^\delta$, we obtain

$$\begin{aligned}
\int |1_E - 1_{E_{2T}}|\, \mu(dx) &= \mu(E) - \mu(E_{2T}) \\
&\leqslant \mu(E) - \mu(E^\delta) \\
&= \mu(E) - \left(1 - \frac{\delta}{R}\right)^d \mu(E) \\
&= \mu(E) \left(1 - \left(1 - \frac{\delta}{R}\right)^d\right) \\
&\leqslant \mu(E)\, d \frac{\delta}{R} \\
&\leqslant 318 \mu(E)\, d \sqrt{\frac{d(d+1)}{T}}. \quad \blacksquare
\end{aligned} \tag{21}$$

6. REMARKS

In earlier work of one of the authors (Cheang [3]), an $\mathscr{L}_2$ approximation bound of order $T^{-1/6}$ was obtained with T pairs of half-spaces in a neural net with a ramp sigmoid applied to the output. Here our order $T^{-1/2}$ bound gives an improved rate.

The integral representation to the Gaussian on B_K may also be written

$$\exp\left(-\frac{|x|^2}{2}\right) = \int_{\mathscr{R}^d} \int_{-|a|\,K}^{|a|\,K} 1_{\{\mathrm{sgn}(b)\, a\cdot x + b\,\mathrm{sgn}(b) \geqslant 0\}}\, |\sin(b)|\, \frac{\exp(-|a|^2/2)}{(2\pi)^{d/2}}\, db\, da$$
$$-\frac{1}{2} \int_{\mathscr{R}^d} \int_{-|a|\,K}^{|a|\,K} \sin^-(b)\, \frac{\exp(-|a|^2/2)}{(2\pi)^{d/2}}\, db\, da$$
$$+\exp\left(-\frac{K^2}{2}\right). \tag{22}$$

Sampling from the distribution V proportional to $|\sin(b)| \exp(-\frac{|a|^2}{2})$, the approximation to the ball takes the form

$$\mathscr{N}_T = \left\{ x \in \mathscr{R}^d : \sum_{i=1}^{T} 1_{\{a_i \cdot x \geqslant b_i\}} \geqslant k \right\},$$

that is, x is in $\mathscr{N}_T$ if it is in at least k of the half-spaces. This approximation achieves

$$\delta^H(B, \mathscr{N}_T) \leqslant 318 \sqrt{\frac{d(d+1)}{T}}. \tag{23}$$

In particular, when $2T$ sigmoids are used in the approximation,

$$\delta^H(B, \mathscr{N}_{2T}) \leqslant 318 \sqrt{\frac{d(d+1)}{2T}}$$

when the representation (22) is used, reducing the constant by a factor of $1/\sqrt{2}$ from the bound in Theorem 1.

It may be possible to extend our results to neural network approximation of other classes of closed convex sets with smooth boundaries, for example, to classes of sets of the form $\mathscr{D} = \{x \in \mathscr{R}^d : f(x) \cdot f(x) \leqslant 1\}$, where $f: \mathscr{R}^d \to \mathscr{R}^d$ has a strictly positive definite derivative. If this is achieved, the results pertaining to functions which have total finite variation with respect to a class of ellipsoids (in the following section) could be extended to those for a class of convex sets with some suitable smoothness properties.

7. APPROXIMATION BOUNDS FOR TWO LAYER NETS

The second (outer) layer of a two layer net takes a linear combination of level sets H of functions represented by linear combinations on the first (inner) layer. The class of sets represented by level sets of combinations of first layer nodes include half-spaces and rectangles, and (as we have seen) approximations to ellipsoids.

A function f is said to have variation $V_{f,\mathscr{H}}$ with respect to a class of sets $\mathscr{H}$ if $V_{f,\mathscr{H}}$ is the infimum of numbers V such that f/V is in the closure of the convex hull of signed indicators of sets in $\mathscr{H}$, where the closure is taken in $\mathscr{L}_2(P_X)$. A special case of finite variation is the case we call total variation with respect to a class of sets. Suppose $f(x)$ defined over a bounded region S in $\mathscr{R}^d$. We say that f has total variation V with respect to a class of sets $\mathscr{H} = \{H_\xi : \xi \in \Xi\}$ if there exist some signed measure v over the measurable space Ξ and

$$f(x) = \int_\Xi 1_{H_\xi}(x)\, v(d\xi) \qquad \text{for} \quad x \in S, \tag{24}$$

and if v has finite total variation V. The sets H_ξ are parametrized by ξ in Ξ. In our context, the H_ξ are half spaces in $\mathscr{R}^d$ where the ξ consist of the location and orientation parameters. In the event that the representation (24) is not unique, we take the measure v that yields the smallest total variation V.

The function class $\mathscr{F}_{V,\mathscr{H}}$ of functions with variation $V_{f,\mathscr{H}}$ bounded by V arises naturally when thinking of the functions obtained by linear combinations on a layer of a network where the sum of absolute values of the coefficients of linear combination are bounded by V and the level sets from the preceding layer yield the sets in $\mathscr{H}$. In our analysis of the two layer case we will take advantage of both $\mathscr{L}_\infty$ approximations bounds (used to yield approximations to the indicators of ellipsoids in the inner layer) and $\mathscr{L}_2$ approximation bounds for convex hulls of indicators of ellipsoids (essentially achieved by the outer layer of the network). First we state a simple $\mathscr{L}_2$ approximation bound, which is a counterpart to the $\mathscr{L}_\infty$ bound of Lemma 1, but with smaller constants (and without requirement of integral representation).

LEMMA 2. *If f has variation V_f with respect to a class of sets $\mathscr{H}$ then for each T there exists $H_1, ..., H_T$ and $c_1, ..., c_T$ with $\sum_{i=1}^T |c_i| \leqslant V_f$ such that the approximation $f_T(x) = \sum_{i=1}^T c_i 1_{H_i}(x)$ achieves*

$$\|f - f_T\|_2 \leqslant \frac{V_f}{\sqrt{T}}. \tag{25}$$

This lemma as a tool of approximation theory and two probabilistic proofs (one based on probabilistic sampling and one based on a greedy algorithm) are in Barron [2] and (with a somewhat larger constant) an earlier form of the greedy algorithm proof of the approximation result is in Jones [12]. The probabilistic sampling bound on $\mathscr{L}_2$ norms of averages used in the proof is classical Hilbert theory.

Proof. The proof is based on the Monte Carlo sampling idea as in Section 2. First fix T and suppose that f is not identically constant. (Equality occurs in (25) only if f is identically constant.) Since f is in the closure of the convex hull of $G = \{\pm V_f 1_H : H \in \mathscr{H}\}$, one takes a $\tilde{f}$ that is a (potentially very large) finite convex combination with $\|f - \tilde{f}\|_2 < \delta$. In particular we take $\delta = \varepsilon/\sqrt{T}$ and ε small, say $\varepsilon < V_f - \sqrt{V_f^2 - \|f\|^2/4}$, which is less than $\frac{\|f\|}{2}$.

By the triangle inequality,

$$\begin{aligned} \|f - f_T\|_2 &\leqslant \|f - \tilde{f}\|_2 + \|\tilde{f} - f_T\|_2 \\ &\leqslant \frac{\varepsilon}{\sqrt{T}} + \|\tilde{f} - f_T\|_2. \end{aligned} \tag{26}$$

Suppose $\tilde{f} = \sum_i p_i g_i$ with g_i in G, and $p_i > 0$ with $\sum_i p_i = 1$. Since $\tilde{f}$ is an expectation, we apply the Monte Carlo sampling technique. Draw indices $i_1, \ldots, i_T$ independently according to the distribution p_i in the representation of $\tilde{f}$ and let $f_T = \frac{1}{T}\sum_{j=1}^{T} g_{i_j}$. Then

$$\begin{aligned} E_i \|\tilde{f} - f_T\|_2^2 &= \frac{E_i \|g_i\|^2 - \|\tilde{f}\|^2}{T} \\ &\leqslant \frac{V_f^2 - \|\tilde{f}\|^2}{T} \\ &\leqslant \frac{V_f^2 - \|f\|^2/4}{T}, \end{aligned} \tag{27}$$

and so there exists a choice of such an f_T with

$$\|\tilde{f} - f_T\|_2^2 < \frac{V_f^2 - \|f\|^2/4}{T}.$$

That is,

$$\|\tilde{f} - f_T\|_2 < \frac{\sqrt{V_f^2 - \|f\|^2/4}}{\sqrt{T}}. \tag{28}$$

Substituting this bound back into (26) completes the proof. ∎

As a consequence of the lemma above, we have the following corollary involving approximation with a class of ellipsoids. Let ξ be the parameters that define the ellipsoids, and $1_{E_\xi}(x)$ the indicator of the ellipsoid.

COROLLARY 3. *If f has variation $V_f = V_{f,\mathscr{E}}$ with respect to the class $\mathscr{E}$ of ellipsoids then there is a choice of ellipsoids $E_1, ..., E_{T_1}$ and $s_1, ..., s_{T_1} \in \{-1, +1\}$, and $c_i = V_f s_i / T_1$ such that*

$$f_{T_1}(x) = \sum_{i=1}^{T_1} c_i 1_{E_i} \tag{29}$$

satisfies

$$\|f_{T_1} - f\|_2 \leqslant \frac{V_f}{\sqrt{T_1}}. \tag{30}$$

The indicators of ellipsoids have two layer sigmoidal network approximations consisting of a single outer node and a single hidden inner layer. These approximations to 1_{E_i} may be substituted into the approximation in (29) to yield a two hidden layer approximation to f.

Let $\mathscr{E} = \{E_\xi : \xi \in \Xi\}$ be the set of ellipsoids with $\mu(E_\xi) \leqslant \mu(\mathscr{S})$ where μ is the Lebesgue measure. Let P_X be the uniform probability measure over $\mathscr{S}$, and let E_{2T_2} be the neural net level set with $2T_2$ sigmoids that is used to approximate E. Using the bound in Corollary 1, for each $E \in \mathscr{E}$,

$$\begin{aligned}\int_{\mathscr{S}} |1_E(x) - 1_{E_{2T_2}}(x)|^2 P_X(dx) &= \frac{\mu((E - E_{2T_2}) \cap \mathscr{S})}{\mu(\mathscr{S})} \\ &\leqslant \frac{\mu(E)}{\mu(\mathscr{S})} 318d \sqrt{\frac{d(d+1)}{2}} \\ &\leqslant 318d \sqrt{\frac{d(d+1)}{T_2}}. \end{aligned} \tag{31}$$

After replacing the indicators of the ellipsoids in (29) with their neural net approximations, we obtain

$$f_{T_1, 2T_2} = \sum_{i=1}^{T_1} c_i \phi \left(\sum_{j=1}^{T_2} \omega_{ij} \phi(a_{ij} \cdot x - b_{ij}) - d_i \right). \tag{32}$$

The following theorem bounds the mean-squared approximation error. An ellipsoid in $\mathscr{E}$ is denoted by E.

THEOREM 4. *If f has variation V_f with respect to the class of ellipsoids $\mathscr{E}$, with $\mu(E) \leqslant \mu(\mathscr{S})$ and P_X is the uniform probability measure over $\mathscr{S}$, then*

there exist a choice of parameters $(a_{ij}, b_{ij}, c_i, d_i, \omega_{ij})$ *such that a two hidden layer net with step activation function achieves approximation error bounded by*

$$\|f - f_{T_1, 2T_2}\|_2 \leqslant \frac{V_f}{\sqrt{T_1}} + V_f \left(318d \sqrt{\frac{d(d+1)}{T_2}}\right)^{1/2}, \tag{33}$$

and

$$\|f - f_{T_1, 2T_2}\|_1 \leqslant \frac{V_f}{\sqrt{T_1}} + V_f 318d \sqrt{\frac{d(d+1)}{T_2}},$$

where $\|\cdot\|_p$ *denotes the* $\mathscr{L}_p(P_X)$ *norm; provided that* T_2 *is large enough that* $68\sqrt{d(d+1)/T_2} \leqslant \frac{1}{2}(\exp(-\frac{1}{4}) - \exp(-\frac{1}{2}))$.

Proof. By the triangle inequality,

$$\|f - f_{T_1, 2T_2}\|_2 \leqslant \|f - f_{T_1}\|_2 + \|f_{T_1} - f_{T_1, 2T_2}\|_2. \tag{34}$$

Now

$$\|f - f_{T_1}\|_2 \leqslant \frac{V_f}{\sqrt{T_1}}$$

from Corollary 1. The other term on the right hand side of (34) is bounded as follows. Let $\tilde{E}_i$ be the neural net level set of the approximation to E_i from Section 5. Then

$$\begin{aligned} \|f_{T_1} - f_{T_1, 2T_2}\|_2 &= \left\| \sum_{i=1}^{T_1} c_i (1_{E_i} - 1_{\tilde{E}_i}) \right\|_2 \\ &\leqslant \frac{1}{T_1} \sum_{i=1}^{T_1} |c_i| \, \|1_{E_i} - 1_{\tilde{E}_i}\|_2 \\ &\leqslant V_f \left(318d \sqrt{\frac{d(d+1)}{T_2}}\right)^{1/2}, \end{aligned} \tag{35}$$

where (31) bounds the last inequality (35). ∎

The proof of the $\mathscr{L}_1$ bound is similar (using $\|f - f_{T_1}\|_1 \leqslant \|f - f_{T_1}\|_2 \leqslant V_f/\sqrt{T_1}$) except that the square root in (35) is not used in bounding $\|1_{E_i} - 1_{\tilde{E}_i}\|_1$.

We conclude with two examples of functions with variation with respect to a class of balls (ellipsoids).

EXAMPLE 1. Convex Combination of Balls. Let $B(a, b)$ denote a ball centered at a with radius b. In $\mathscr{R}^3$, the function

$$f(x) = \frac{4\pi}{3} - \sqrt{2}\,\pi(x_1^2 + x_2^2 + x_3^2)^{1/2} + \frac{\pi}{3\sqrt{2}}(x_1^2 + x_2^2 + x_3^2)^{3/2}$$

$$= \int 1_{B(\theta, 1)}(x)\, 1_{B(0, 1)}(\theta)\, d\theta \tag{36}$$

is a convex combination of indicators of balls. Thus

$$f_{T_1}(x) = \frac{4\pi}{3T_1} \sum_{i=1}^{T_1} 1_{B(\theta_i, 1)}(x) \tag{37}$$

is an approximation to $f(x)$ where the θ_i's are sampled from the uniform distribution in a unit ball. We then approximate each ball $1_{B(\theta_i, 1)}(x)$ with the form (5).

EXAMPLE 2. A Radial Function Let $\mu \geqslant 2$,

$$f(x) = \begin{cases} \frac{1}{2}(|x| - \mu + 2), & \mu - 2 < |x| \leqslant \mu \\ \frac{1}{2}(\mu + 2 - |x|), & \mu < |x| \leqslant \mu + 2 \\ 0, & \text{otherwise.} \end{cases} \tag{38}$$

Then

$$f(x) = \int_{\mathscr{R}} 1_{(\theta - 1, \theta + 1)}(|x|)\, \tfrac{1}{2} 1_{[-1, 1]}(\theta - \mu)\, d\theta \tag{39}$$

and thus $f(x)$ can be approximated by

$$f_{T_1}(x) = \frac{1}{T_1} \sum_{i=1}^{T_1} \{1_{B(0, \theta_i + 1)}(x) - 1_{B(0, \theta_i - 1)}(x)\}, \tag{40}$$

where $\theta_i \sim$ iid Uniform $(\mu - 1, \mu + 1)$.

APPENDIX A

Starting with the right hand side of (6) and recalling that $|a \cdot x| \leqslant |a|\, K$ for all $x \in B_K$, we obtain

$$\int_{\mathscr{R}^d} \int_{-|a|\,K}^{|a|\,K} 1_{\{a\cdot x+b\geqslant 0\}} \sin(b) \frac{\exp(-|a|^2/2)}{(2\pi)^{d/2}}\, db\, da \tag{41}$$

$$= -\operatorname{Im} \int_{\mathscr{R}^d} \int_{-|a|\,K}^{|a|\,K} 1_{\{a\cdot x+b\geqslant 0\}} \exp(-ib) \frac{\exp(-|a|^2/2)}{(2\pi)^{d/2}}\, db\, da$$

$$= -\operatorname{Im} \int_{\mathscr{R}^d} \left[\int_{a\cdot x-|a|\,K}^{a\cdot x+|a|\,K} 1_{\{s\geqslant 0\}} \exp(-is)\, ds \right] \frac{\exp(ia\cdot x)\exp(-|a|^2/2)}{(2\pi)^{d/2}}\, da$$

$$= -\operatorname{Im} \int_{\mathscr{R}^d} \left[\int_{0}^{a\cdot x+|a|\,K} \exp(-is)\, ds \right] \frac{\exp(ia\cdot x)\exp(-|a|^2/2)}{(2\pi)^{d/2}}\, da$$

$$= \operatorname{Im} i \int_{\mathscr{R}^d} [1-\exp(-ia\cdot x)\exp(-i\,|a|\,K)] \frac{\exp(ia\cdot x)\exp(-|a|^2/2)}{(2\pi)^{d/2}}\, da$$

$$= \exp\left(-\frac{|x|^2}{2}\right) - \int_{\mathscr{R}^d} \frac{\exp(-i\,|a|\,K)\exp(-|a|^2/2)}{(2\pi)^{d/2}}\, da \tag{42}$$

$$= f(x) - \exp\left(-\frac{K^2}{2}\right). \tag{43}$$

In (41), we did a substitution $s = a\cdot x + b$.

APPENDIX B

We prove a more general version of Lemma 1. Let a parameterized class of sets $\mathscr{H} = \{H_\xi : \xi \in \Xi\}$ in $\mathscr{R}^d$ be given where Ξ is a measurable space. Let $\tilde{\mathscr{H}} = \{\tilde{H}_x : x \in \mathscr{R}^d\}$ with $\tilde{H}_x = \{\xi : x \in H_\xi\}$ be the dual class of sets in Ξ parametrized by x.

First we define some terms that will be used in the lemma. Let $\mathscr{G}$ be a class of functions mapping from $\mathscr{X}$ to $\mathscr{R}$ and let $x_1, ..., x_N \in \mathscr{X}$. We say that $x_1, ..., x_N$ are shattered by $\mathscr{G}$ if there exists $r \in \mathscr{R}^N$ such that for each $b = (b_1, ..., b_N) \in \{0, 1\}^N$, there is an $g \in \mathscr{G}$ such that for each i,

$$g(x_i) \begin{cases} \geqslant r_i & \text{if} \quad b_i = 1 \\ < r_i & \text{if} \quad b_i = 0. \end{cases}$$

The pseudo-dimension is defined as

$$\dim_P(\mathscr{G}) = \max\{N : \exists x_1, ..., x_N, \mathscr{G} \text{ shatters } x_1, ..., x_N\} \tag{44}$$

if such a maximum exists, and ∞ otherwise. For the class of unit step functions $\phi(a\cdot x+b)$, the pseudo-dimension and the VC-dimension D coincide and is $d+1$. The ε-packing number $D_T(\varepsilon, \mathscr{L}_p)$ for a subset of a metric space

is defined as the largest number m for which there exist points $t_1, ..., t_m$ in the subset of the metric space with $d_p(t_i, t_j) > \varepsilon$ for $i \neq j$, where d_p is the $\mathscr{L}_p$ metric.

LEMMA 4. *If $\tilde{\mathscr{H}}$ has VC-dimension D and if h is a function in the convex hull of the indicators of sets in $\mathscr{H}$ which possesses an integral representation*

$$h(x) = \int 1_{H_\xi}(x)\, P(d\xi) \qquad \text{for} \quad x \in B_K,$$

then there is a choice of $\xi_1, \xi_2, ..., \xi_T$ such that the approximation $h_T(x) = \frac{1}{T}\sum_{i=1}^{T} 1_{H_{\xi_i}}(x)$ satisfies

$$\sup_{x \in B_K} |h_T(x) - h(x)| \leqslant 34 \sqrt{\frac{D}{T}}. \tag{45}$$

Remark. Such a uniform approximation bound holds over any subset of $\mathscr{R}^d$ in which the integral representation holds. In our application we use B_K, the ball of radius K.

Proof. Let $g_x(\xi) = 1_{\tilde{H}_x}(\xi) = 1_{H_\xi}(x)$ and let σ_i be independent random variables taking the values ± 1 with probablity $\frac{1}{2}$. Define $\underline{\xi} = (\xi_1, \xi_2, ..., \xi_T)$, where the ξ_i are independently and identically distributed with respect to $P(\cdot)$, and $\underline{\sigma} = (\sigma_1, \sigma_2, ..., \sigma_T)$. By symmetrization, using Jensen's inequality as in Pollard [13, p. 7], for $C > 0$, we have

$$E\Psi\left(\frac{\sup_{x \in B_K} |\sum_{i=1}^{T} g_x(\xi_i) - Th(x)|}{C}\right)$$

$$\leqslant E_{\underline{\xi}} E_{\underline{\sigma}} \Psi\left(\frac{2 \sup_{x \in B_K} |\sum_{i=1}^{T} \sigma_i g_x(\xi_i)|}{C}\right). \tag{46}$$

Conditioning on ξ, we need to find an upper bound to $E_{\underline{\sigma}} \Psi(2 \sup_{x \in B_K} |\sum_{i=1}^{T} \sigma_i g_x(\xi_i)|)$. This involves bounding the Orlicz norm $\|2 \sup_{x \in B_K} |\sum_{i=1}^{T} \sigma_i g_x(\xi_i)|\|_\Psi$ with $\underline{\xi}$ fixed. Using a result in Pollard [13, pp. 35–37],

$$\left\| 2 \sup_{x \in B_K} \left| \sum_{i=1}^{T} \sigma_i g_x(\xi_i) \right| \right\|_\Psi \leqslant 18 \sqrt{T} \int_0^1 \sqrt{\log D_T(\varepsilon, \mathscr{L}_2)}\, d\varepsilon, \tag{47}$$

where $D_T(\varepsilon, \mathscr{L}_2)$ is the $\mathscr{L}_2$ ε-packing number for $\tilde{\mathscr{H}}$, where the $\mathscr{L}_2$ norm on Ξ is taken with respect to the empirical probability measure on $\xi_1, \xi_2, ..., \xi_T$.

From Pollard [13, p. 14],

$$D_T(\varepsilon, \mathscr{L}_2) \leqslant \left(\frac{3}{\varepsilon}\right)^D, \tag{48}$$

uniformly over all $\xi_1, \xi_2, ..., \xi_T$. We now work out an upper bound to $\int_0^1 \sqrt{\log D_T(\varepsilon, \mathscr{L}_2)}\, d\varepsilon$. From the Cauchy–Schwartz inequality,

$$\begin{aligned}\int_0^1 \sqrt{\log D_T(\varepsilon, \mathscr{L}_2)}\, d\varepsilon &\leqslant \sqrt{\int_0^1 \log D_T(\varepsilon, \mathscr{L}_2)\, d\varepsilon} \\ &= \sqrt{D \log 3 - D \int_0^1 \log \varepsilon\, d\varepsilon} \\ &\leqslant \sqrt{(1+\log 3)\, D}. \end{aligned} \tag{49}$$

Substituting (49) into (47), we see that

$$\left\| 2 \sup_{x \in B_K} \left| \sum_{i=1}^n \sigma_i g_x(\xi_i) \right| \right\|_{\Psi} \leqslant 18 \sqrt{(1+\log 3)\, TD}. \tag{50}$$

From the definition of the Orlicz norm, the choice of $C_0 = 18\sqrt{(1+\log 3)\, TD}$ ensures that

$$E_\sigma \Psi\left(\frac{2 \sup_{x \in B_K} |\sum_{i=1}^T \sigma_i g_x(\xi_i)|}{C_0}\right) \leqslant 1,$$

and hence,

$$E\Psi\left(\frac{\sup_{x \in B_K} |\sum_{i=1}^T g_x(\xi_i) - Th(x)|}{C_0}\right) \leqslant 1. \tag{51}$$

Thus we conclude that there exists $\xi_1, \xi_2, ..., \xi_T$ such that

$$\Psi\left(\frac{\sup_{x \in B_K} |\sum_{i=1}^T g_x(\xi_i) - Th(x)|}{C_0}\right) \leqslant 1,$$

whence

$$\begin{aligned}\sup_{x \in B_K} \left| \frac{1}{T} \sum_{i=1}^T g_x(\xi_i) - h(x) \right| &\leqslant \frac{18\sqrt{D(1+\log 3)\log 5}}{\sqrt{T}} \\ &\leqslant 34 \sqrt{\frac{D}{T}}. \quad \blacksquare \end{aligned} \tag{52}$$

In our case, $\xi = (a, b)$ and $g_x(\xi) = 1_{H_\xi}(x) = 1_{\{a \cdot x + b \geqslant 0\}}$. The dual class of sets in Ξ are $\tilde{H}_x = \{\xi : g_x(\xi) = 1\} = \{(a, b) : a \cdot x + b \geqslant 0\}$. Since $(a, b) \in \mathscr{R}^d \times \mathscr{R}$, which is a vector space of dimension $d+1$, the class of sets $\tilde{\mathscr{H}} = \{\tilde{H}_x : x \in \mathscr{R}^d\}$ has VC-dimension $D = d+1$ (Pollard [12, p. 20], Haussler [10]). Thus Lemma 1.

REFERENCES

1. A. R. Barron, Neural net approximation, *in* "Proc. of the 7th Yale Workshop on Adaptive and Learning Systems," 1992.
2. A. R. Barron, Universal approximation bounds for superpositions of a sigmoidal function, *IEEE Trans. Inform. Theory* **39** (1993), 930–945.
3. G. H. L. Cheang, Neural network approximation and estimation of functions, *in* "Proc. of the 1994 IEEE-IMS Workshop on Info. Theory and Stat.," 1994.
4. G. Cybenko, Approximations by superpositions of a sigmoidal function, *Math. Control Signals Systems* **2** (1989), 303–314.
5. R. Dudley, Metric entropy of some classes of sets with differentiable boundaries, *J. Approx. Theory* **10** (1974), 227–236; Correction, *J. Approx. Theory* **26** (1979), 192–193.
6. L. Fejes Tóth, "Lagerungen in der Ebene, auf der Kugel und im Raum," Springer-Verlag, Berlin, 1953, 1972.
7. P. M. Gruber, Approximation of convex bodies, *in* "Convexity and its Appplications" (P. M. Gruber and J. M. Wills, Eds.), pp. 131–162, Birkhäuser, Basel, 1983.
8. P. M. Gruber and P. Kenderov, Approximation of convex bodies by polytopes, *Rend. Circ. Mat. Palermo* **31** (1982), 195–225.
9. S. S. Haykin, "Neural Networks: a Comprehensive Foundation," Macmillan, New York, 1994.
10. D. Haussler, Decision theoretic generalizations of the PAC model for neural net and other learning applications, *Inform. Comp.* **100**, No. 1 (1992), 78–150.
11. K. Hornik, M. B. Stinchcombe, and H. White, Multi-layer feedforward networks are universal approximators, *Neural Networks* **2** (1988), 359–336.
12. 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.* **20** (1990), 608–613.
13. D. Pollard, "Convergence of Stochastic Processes," Springer-Verlag, Berlin, 1984.
14. D. Pollard, "Empirical Processes: Theory and Applications," NSF-CBMS Regional Conf. Ser. Probab. Statist., Vol. 2, SIAM, Philadelphia, 1990.
15. F. Rosenblatt, The perceptron: A probabilistic model for information storage and organization in the brain, *Psych. Rev.* **65** (1958), 386–408.
16. F. Rosenblatt, "Principles of Neurodynamics," Spartan, Washington, DC, 1962.
17. R. Schneider and J. A. Wieacker, Approximation of convex bodies by polytopes, *Bull. London Math. Soc.* **13** (1981), 149–156.
18. V. N. Vapnik and A. Ya. Červonenkis, On the uniform convergence of relative frequencies of events to their probabilities, *Theory Probab. Appl.* **16**, No. 2 (1971), 264–280.
19. J. E. Yukich, M. B. Stinchcombe, and H. White, Sup-norm approximation bounds for networks through probabilistic methods, *IEEE Trans. Inform. Theory* **41** (1995), 1021–1027.