Available online at www.sciencedirect.com

ScienceDirect

Journal of Approximation Theory 147 (2007) 1–10

Journal of Approximation Theory

www.elsevier.com/locate/jat

A Sobolev-type upper bound for rates of approximation by linear combinations of Heaviside plane waves☆

Paul C. Kainen[a,*], Věra Kůrková[b], Andrew Vogt[a]

[a] *Department of Mathematics, Georgetown University, Washington, DC 20057-1233, USA*

[b] *Institute of Computer Science, Academy of Sciences of the Czech Republic, Pod vodárenskou věží 2, 182 07 Prague 8, Czech Republic*

Received 13 August 2005; received in revised form 12 December 2006; accepted 15 December 2006

Communicated by Amos Ron
Available online 12 January 2007

Abstract

Quantitative bounds on rates of approximation by linear combinations of Heaviside plane waves are obtained for sufficiently differentiable functions f which vanish rapidly enough at infinity: for d odd and $f \in C^d(\Re^d)$, with lower-order partials vanishing at infinity and dth-order partials vanishing as $\|x\|^{-(d+1+\varepsilon)}$, $\varepsilon > 0$, on any domain $\Omega \subset \Re^d$ with unit Lebesgue measure, the $L_2(\Omega)$-error in approximating f by a linear combination of n Heaviside plane waves is bounded above by $k_d \|f\|_{d,1,\infty} n^{-1/2}$, where $k_d \sim (\pi d)^{1/2}(e/2\pi)^{d/2}$ and $\|f\|_{d,1,\infty}$ is the Sobolev seminorm determined by the largest of the L^1-norms of the dth-order partials of f on $\Re^d$. In particular, for d odd and $f(x) = \exp(-\|x\|^2)$, the $L_2(\Omega)$-approximation error is at most $(2\pi d)^{3/4} n^{-1/2}$ and the sup-norm approximation error on $\Re^d$ is at most $68\sqrt{2}(n-1)^{-1/2}(2\pi d)^{3/4} \sqrt{d+1}$, $n \geqslant 2$.

Keywords: Characteristic functions of closed half-spaces; Perceptron neural networks; Integral formulas; Variation with respect to half-spaces; Radon transform; Gaussian function; Rates of approximation

1. Introduction

Let $\mathcal{H}_d$ denote the set of all half-space characteristic functions on $\Re^d$, also called *Heaviside plane waves*. We approximate a given function f on $\Re^d$ by a finite linear combination of elements from $\mathcal{H}_d$.

☆ V.K. was partially supported by MŠMT ČR project number 1M684077004 (1M0567) (Center of Applied Cybernetics), P.C.K. and A.V. by a grant from Georgetown University's International Affairs Office.

* Corresponding author. Fax: +1 202 687 6067.
E-mail address: kainen@georgetown.edu (P.C. Kainen).

0021-9045/$ - see front matter © 2007 Published by Elsevier Inc.
doi:10.1016/j.jat.2006.12.009

Barron [5] studied the effect of smoothness on such approximation by plane waves (i.e., by perceptron neural networks). Maurey [31], Jones [16], and Barron showed that if f is in the $\mathcal{L}_2$-closure of the convex hull of $\mathcal{H}_d \cup -\mathcal{H}_d$ (restricted to a unit-volume domain of $\Re^d$), then f is within $\mathcal{L}_2$-distance of $n^{-1/2}$ of a subspace spanned by n members of $\mathcal{H}_d$. Applying this, Barron obtained an $O(n^{-1/2})$ estimate but the constant (with respect to n) depends on d in an unspecified fashion—see [5, p. 941].

In this paper we apply the Maurey–Jones–Barron upper bound and obtain an explicit constant for functions expressed directly as integral combinations of Heaviside plane waves. Similar integral representations have been used to prove density of linear spans generated by Heaviside plane waves, e.g. [6,13,15,26], and to bound rates of approximation by radial waves, e.g. [7,11,12,27]; see also the survey [30]. For other aspects, see [17,19].

We show that for d odd and $\Omega \subset \Re^d$ a subset of unit measure, the $\mathcal{L}_2(\Omega)$-error in approximating the Gaussian by a linear combination of n Heaviside plane waves is at most $(2\pi d)^{3/4}\ n^{-1/2}$. Thus, radial-basis-type approximations utilizing the Gaussian function can be replaced by perceptron-type approximations based on the Heaviside function with only a mild increase in the number of units needed to obtain a given accuracy.

An outline of the paper follows. Section 2 states the MJB Theorem in terms of variational norm. Section 3 bounds this norm above for functions representable by an integral formula involving Heavisides. Section 4 describes a class of functions f and a particular integral formula whose weight function permits evaluation of the bound. Section 5 relates these results to the Radon transform, and Section 6 determines the consequences for the Gaussian function.

2. Rates of approximation in Hilbert space

In this paper all normed linear spaces are over the reals $\Re$ and S^{d-1} denotes the unit sphere in $\Re^d$.

Let $(X, \|.\|)$ be a normed linear space with nonempty subset G. For $n \geqslant 1$, span G and span$_n$ G denote the set of all finite linear combinations (and all n-fold linear combinations, resp.) of elements from G, while conv G denotes the set of all finite linear combinations of elements from G using nonnegative scalars with sum equal to 1.

For $f \in X$, let $\|f\|_{G,X} = \inf\ \{c > 0 : f/c \in \operatorname{cl}\operatorname{conv}(G \cup -G)\}$, where cl denotes closure with respect to the topology induced by the norm on X. This extended-real-valued functional is called G-variation; the subset G should be bounded, nonempty, and nonzero to avoid trivialities. For $c \geqslant 0$, $\|f\|_{G,X} \leqslant c$ if and only if for every $\varepsilon > 0$ there exists $n \geqslant 1$ such that for each j, $1 \leqslant j \leqslant n$, there exist $g_j \in G$ and $c_j \in \Re$ such that

$$\left\| f - \sum_{j=1}^{n} c_j g_j \right\| < \varepsilon, \qquad \sum_{j=1}^{n} |c_j| \leqslant c. \tag{1}$$

See [22] and for recent applications [23,25].

The Maurey–Jones–Barron Theorem [31, p. V.2, 16, p. 611, 5, p. 934] can be stated as follows:

Theorem 2.1. *If X is a Hilbert space with G a bounded, nonempty, and nonzero subset, then for every $f \in X$ and every positive integer n,*

$$\|f - \operatorname{span}_n G\|_X \leqslant n^{-1/2} \left(\sup_{g \in G} \|g\|_X \right) \|f\|_{G,X}.$$

3. Upper bounds on half-space variation

Let $\mathcal{H}_d$ be the set of characteristic functions of closed half-spaces. Then

$$\mathcal{H}_d = \{\vartheta_{e,b} \,|e \in S^{d-1}, b \in \Re\},$$

where $\vartheta_{e,b} : \Re^d \to \Re$ is given by $\vartheta_{e,b}(x) = \vartheta(e \cdot x + b)$, ϑ is the *Heaviside function* $\vartheta(t) = 0$ for $t < 0$ and $\vartheta(t) = 1$ for $t \geqslant 0$, and the parameters e, b belong, respectively, to S^{d-1} and $\Re$. Thus, $\mathcal{H}_d$ is the set of all compositions of affine functions with ϑ. The functions $\vartheta_{e,b}$ are called *Heaviside plane waves.*

Let $\mathcal{M}(\Omega)$ be the space of bounded measurable functions on a measurable subset Ω of $\Re^d$ under the sup norm. Note that $\mathcal{H}_d \subset \mathcal{M}(\Re^d)$.

Theorem 3.1. *Let $d \geqslant 1$. If $f \in \mathcal{M}(\Re^d)$ can be expressed as*

$$f(x) = \int_{S^{d-1} \times \Re} w(e, b)\vartheta(e \cdot x + b)\, de\, db,$$

where w is continuous on $S^{d-1} \times \Re$, then

$$\|f\|_{\mathcal{H}_d, \mathcal{M}(\Re^d)} \leqslant \int_{S^{d-1} \times \Re} |w(e, b)|\, de\, db.$$

Proof. Let $P = S^{d-1} \times \Re$. For $p = (e, b)$ in P and x in $\Re^d$, let $\vartheta_p(x) := \vartheta_{e,b}(x) = \vartheta(e \cdot x + b)$. Then $d\lambda(p) = de\, db$, where λ is the product of the usual measure μ_{d-1} on S^{d-1} and Lebesgue measure on $\Re$.

Let $\varepsilon > 0$ be arbitrary. Without loss of generality, we may assume that the $\mathcal{L}_1$-norm of w is finite, and choose a closed interval $[-M, M] \subset \Re$ so that with $\hat{P} = S^{d-1} \times [-M, M]$,

$$\int_{P \setminus \hat{P}} |w(p)|\, d\lambda(p) < \varepsilon.$$

As $\hat{P}$ is compact, $\lambda(\hat{P})$ is finite. Let δ_w denote the modulus of continuity of w on $\hat{P}$, so for $t \geqslant 0$, if $|p - p'| < \delta_w(t)$, then $|w(p) - w(p')| < t$ for all $p, p' \in \hat{P}$. Since w is continuous and $\hat{P}$ is compact, $\sup_{p \in \hat{P}} |w(p)| < \infty$.

Define $\hat{f}$ on $\Re^d$ by

$$\hat{f}(x) = \int_{\hat{P}} w(p)\vartheta_p(x)\, d\lambda(p).$$

Then, since $\vartheta(t) \leqslant 1$, $\|f - \hat{f}\|_{\mathcal{M}(\Re^d)} < \varepsilon$.

To prove the theorem, by the definition of $\mathcal{H}_d$-variation and the triangle inequality, it now suffices to show that there is a finite linear combination of characteristic functions of half-spaces within arbitrarily small sup-norm distance of $\hat{f}$ such that the sum of the absolute values of the coefficients does not exceed $\int_P |w|\, d\lambda(p)$. We obtain such characteristic functions and their coefficients from a sufficiently small-mesh tiling $\mathcal{R}$ of $\hat{P}$.

A *tiling* of a compact subset Y of a finite-dimensional Euclidean space is a finite collection $\mathcal{R}$ of subsets $R \subset Y$ such that the union of $\mathcal{R}$ is Y, each R in $\mathcal{R}$ is compact and connected, and the pairwise intersections of elements in $\mathcal{R}$ have measure zero. The *mesh* $\tau(\mathcal{R})$ of a tiling $\mathcal{R}$ is the maximum of the diameters of its elements. Clearly, for any $M \geqslant 0$, $S^{d-1} \times [-M, M]$ has tilings with arbitrarily small mesh.

For any compact connected $R \subset \hat{P}$ let $c_R := \min\{|w(p)| : p \in R\}$ and let $p_R \in R$ satisfy $|w(p_R)| = c_R$.

Let $\eta > 0$. We claim that there exists $\tau_0 > 0$ such that for all tilings $\mathcal{R}$ of $\hat{P}$ with $\tau(\mathcal{R}) \leqslant \tau_0$, if $g_{\mathcal{R}} := \sum_{R \in \mathcal{R}} c_R \lambda(R) \vartheta_{p_R}$, then $\sup_{x \in \Re^d} |g_{\mathcal{R}}(x) - \hat{f}(x)| < 2\eta$. Since $\sum_{R \in \mathcal{R}} |c_R \lambda(R)| \leqslant \int_{\hat{P}} |w(p)| d\lambda(p) \leqslant \int_P |w(p)|\, d\lambda(p)$, this will prove the theorem.

Let $\mathcal{R}$ be any tiling of $\hat{P}$. For each x in $\Re^d$, $\mathcal{R}_x$ denotes the set of all $R \in \mathcal{R}$ that contain some interior point $p = (e, b)$ for which $x \in H_{e,b}$, i.e., so that $e \cdot x + b = 0$. Then R is in $\mathcal{R}_x$ if and only if $\vartheta_p(x)$ takes on two distinct values on subsets of R with positive measure.

For each x in $\Re^d$, the following inequality holds:

$$|g_{\mathcal{R}}(x) - \hat{f}(x)| \leqslant \left(\sum_{R \in \mathcal{R} \setminus \mathcal{R}_x} + \sum_{R \in \mathcal{R}_x} \right) \left(\int_R |w(p_R)\vartheta_{p_R}(x) - w(p)\vartheta_p(x)|\, d\lambda(p) \right).$$

The first sum is less than η if the mesh $\tau(\mathcal{R})$ is sufficiently small. Indeed, since $\vartheta_p(x)$ is constant (up to a set of measure zero) on such R, each summand is at most $\int_R |w(p_R) - w(p)|\, d\lambda(p) \leqslant \eta\lambda(R)/\lambda(\hat{P})$ when $\tau(\mathcal{R}) \leqslant \delta_w(\eta/\lambda(\hat{P}))$.

The second sum cannot exceed $2 \sup_{p \in \hat{P}} |w(p)| \sum_{R \in \mathcal{R}_x} \lambda(R)$. Thus, the theorem holds provided that we can make the sum of the measures of the tiles in $\mathcal{R}_x$ as small as we please simultaneously for all x in $\Re^d$.

Let $x \in \Re^d$ and let $\hat{A}_x := \{(e, b) \in \hat{P} : e \cdot x + b = 0\} = (S^{d-1} \times [-M, M]) \cap (x, 1)^{\perp}$. Let $A_{x,\tau}$ denote the subset of $\hat{P}$ consisting of all points $p' \in \hat{P}$ within distance τ of $\hat{A}_x$. Then $\bigcup\{R : R \in \mathcal{R}_x\} \subset A_{x,\tau}$ if $\tau(\mathcal{R}) \leqslant \tau$.

Moreover, $\lambda(A_{x,\tau})$ can be made arbitrarily small by taking τ sufficiently small. Indeed, when $\|x\| \leqslant M$, then $\hat{A}_x$ is a generalized ellipsoid and

$$\lambda(A_{x,\tau}) \leqslant 2\omega_d\, \tau\sqrt{1 + M^2}.$$

When $\|x\| > M$, then for each b in $[-M, M]$, $(x, 1)^{\perp}$ intersects $S^{d-1} \times \{b\}$ in a copy of S^{d-2} with radius $\sqrt{1 - b^2/\|x\|^2}$. For τ sufficiently small and $d \geqslant 3$, it can be shown that

$$\lambda(A_{x,\tau}) \leqslant 4M\omega_{d-1}\, \tau\sqrt{1 + 1/\|x\|^2} < 4\omega_{d-1}\, \tau\sqrt{M^2 + 1}$$

for all x with $\|x\| > M$. For $d = 2$, the upper bound above is replaced by a similar expression involving $\delta(2\tau\sqrt{1 + 1/M^2}\,)$, where δ is the modulus of continuity of the inverse cosine. When $d = 1$, $A_{x,\tau}$ is empty.

Thus, $\|\hat{f} - g_{\mathcal{R}}\|_{\mathcal{M}(\Re^d)}$ is arbitrarily small when the mesh $\tau(\mathcal{R})$ is sufficiently small, and the theorem is proved. □

For f in $\mathcal{M}(\Re^d)$ and $\Omega \subset \Re^d$ with finite nonzero measure, let $\|f\|_{\mathcal{H}_d, \mathcal{L}_2(\Omega)}$ denote the variation of $f|_\Omega$ with respect to $\mathcal{L}_2(\Omega)$ and its subset $\{h|_\Omega : h \in \mathcal{H}_d\}$.

Corollary 3.2. *Let f in $\mathcal{M}(\Re^d)$ satisfy $f(x) = \int_{S^{d-1} \times \Re} w(e, b)\vartheta(e \cdot x + b)\, de\, db$ for some continuous function w on $S^{d-1} \times \Re$. Then for $\Omega \subset \Re^d$ with finite nonzero measure*

$$\|f\|_{\mathcal{H}_d, \mathcal{L}_2(\Omega)} \leqslant \int_{S^{d-1} \times \Re} |w(e, b)|\, de\, db.$$

Proof. By (1) and an easy argument, $\|f\|_{\mathcal{H}_d,\mathcal{L}_2(\Omega)} \leqslant \|f\|_{\mathcal{H}_d,\mathcal{M}(\Re^d)}$. □

4. An upper bound on the $\mathcal{L}_1$-norm of a weight function

In this section we give an explicit integral formula for sufficiently smooth functions on $\Re^d$, d odd, that vanish together with their derivatives sufficiently rapidly at infinity. We also show that the $\mathcal{L}_1$-norm of the weight function is bounded by the product of a Sobolev seminorm of f and a factor approximately equal to $\sqrt{\pi d}\ (\frac{e}{2\pi})^{d/2}$. The integral representation for such functions f is in terms of Heaviside plane waves and the weight function w_f involves iterated directional derivatives.

Recall that the *rth iterated directional derivative* $D_e^{(r)} f(x)$ of a function f on $\Re^d$ at the point $x \in \Re^d$ for the unit vector $e \in S^{d-1}$ is defined recursively as $D_e^{(0)} f(x) = f(x)$ and $D_e^{(r+1)} f(x) = \nabla(D_e^{(r)} f(x)) \cdot e$, where ∇ denotes the gradient vector $(\partial/\partial x_1, \ldots, \partial/\partial x_d)$. It is convenient to expand the directional derivative in the following operator equation (e.g., [10, p. 130]):

$$D_e^{(r)} = \sum_{|\alpha|=r} \binom{r}{\alpha} e^\alpha D^\alpha, \tag{2}$$

where $\alpha = (\alpha_1, \ldots, \alpha_d)$ is a multi-index with nonnegative integer components, $|\alpha| = \alpha_1 + \cdots + \alpha_d$, $\binom{r}{\alpha} = r!/\alpha_1! \cdots \alpha_d!$, $v^\alpha = v_1^{\alpha_1} \cdots v_d^{\alpha_d}$ for $v = (v_1, \ldots, v_d)$ in $\Re^d$, and $D^\alpha = \Pi_{i=1}^d (\partial/\partial x_i)^{\alpha_i}$. Also, we write $|v|^\alpha$ for $|v_1|^{\alpha_1} \cdots |v_d|^{\alpha_d}$.

A function f is called *of weakly controlled decay* [20, Proposition 3] if

(i) $f \in \mathcal{C}^d(\Re^d)$ for d odd,
(ii) for every α, $0 \leqslant |\alpha| < d$, $\lim_{\|x\| \to \infty} (D^\alpha f)(x) = 0$, and
(iii) there exists $\varepsilon > 0$ such that for each multi-index α with $|\alpha| = d$

$$\lim_{\|x\| \to \infty} (D^\alpha f)(x) \|x\|^{d+1+\varepsilon} = 0.$$

Examples of functions of weakly controlled decay on $\Re^d$ are the Gaussian function $\gamma_d(x) = \exp(-\|x\|^2)$, the other members of the Schwartz class $\mathcal{S}(\Re^d)$ [2], and all d-times continuously differentiable functions of compact support.

The following integral representation was derived in [24,18,20] by methods from [9], cf. [14]. Let $d_H y$ denote the volume element of the hyperplane $\mathcal{H}_{e.b} := \{y \in \Re^d : e \cdot y + b = 0\}$.

Theorem 4.1 (*Kainen et al. [20, Theorem 1]*). *If f is of weakly controlled decay on $\Re^d$, for d odd, then*

$$f(x) = \int_{S^{d-1}} \int_{\Re} w_f(e, b) \vartheta(e \cdot x + b)\, db\, de,$$

where $w_f(e, b) := a_d \int_{H_{e,b}} (D_e^{(d)}(f))(y)\, d_H y$ *with* $a_d = (-1)^{(d-1)/2}(1/2)(2\pi)^{1-d}$.

A related representation was derived in [18,20] for the case d even but the weight function for that case requires an additional logarithmic factor.

In the theorem, w_f is continuous. To evaluate its $\mathcal{L}_1$-norm, we need the following: let $\omega_d = \mu_{d-1}(S^{d-1})$ be the usual measure of the unit sphere in d-space. Then (e.g., [8, p. 303]) $\omega_d = 2\pi^{d/2}/\Gamma(\frac{d}{2})$, where Γ is the gamma function. By Stirling's Formula (e.g., [1, 6.1.38, p. 257]) $\Gamma(x) \sim \sqrt{2\pi} x^{x-1/2} \exp(-x)$ where $r(x) \sim s(x)$ means $\lim_{x\to\infty} r(x)/s(x) = 1$.

For $f \in \mathcal{C}^d(\Re^d)$, we consider the Sobolev seminorm (cf. [2, p. 101])

$$\|f\|_{d,1,\infty} := \max_{|\alpha|=d} \|D^\alpha f\|_{\mathcal{L}_1(\Re^d)}. \tag{3}$$

Then $\|f\|_{d,1,\infty}$ is not larger (and usually is much smaller) than the ordinary Sobolev norm $\|f\|_{d,1}$, which is the sum of the $\mathcal{L}_1$ norms of a larger set of partials. Also, $\|f\|_{d,1,\infty}$ is finite for *f* of weakly controlled decay on $\Re^d$.

Theorem 4.2. *If f is of weakly controlled decay on* $\Re^d$, *d odd, then*

$$\int_{S^{d-1}\times\Re} |w_f(e,b)|\, de\, db \leqslant k_d \|f\|_{d,1,\infty},$$

where $k_d = |a_d|\omega_d d^{d/2} = 2^{1-d}\pi^{1-d/2}d^{d/2}/\Gamma(\frac{d}{2}) \sim (\pi d)^{1/2}\left(\frac{e}{2\pi}\right)^{d/2}$.

Proof. By definition of w_f, using standard properties of the integral, the definition of the Sobolev seminorm, and the multinomial theorem, we have

$$\begin{aligned}\int_{S^{d-1}}\int_{\Re}\left|\int_{H_{e,b}} D_e^{(d)}(f)\, d_H y\right| db\, de &\leqslant \int_{S^{d-1}}\int_{\Re}\int_{H_{e,b}} \sum_{|\alpha|=d}\binom{d}{\alpha}\left|e^\alpha (D^\alpha f)(y)\right| d_H y\, db\, de \\ &= \int_{S^{d-1}}\int_{\Re^d} \sum_{|\alpha|=d}\binom{d}{\alpha}\left|e^\alpha (D^\alpha f)(y)\right| dy\, de \\ &= \int_{S^{d-1}} \sum_{|\alpha|=d}\binom{d}{\alpha}|e|^\alpha \int_{\Re^d}\left|(D^\alpha f)(y)\right| dy\, de \\ &\leqslant \int_{S^{d-1}} \sum_{|\alpha|=d}\binom{d}{\alpha}|e|^\alpha \|f\|_{d,1,\infty}\, de \\ &= \|f\|_{d,1,\infty}\int_{S^{d-1}}\left(\sum_{i=1}^d |e_i|\right)^d de.\end{aligned}$$

As $\sum_{i=1}^d |e_i|$ is maximized on S^{d-1} when $|e_i| = d^{-1/2}$ for $i = 1,\ldots,d$, we have $\int_{S^{d-1}}(\sum_{i=1}^d |e_i|)^d de \leqslant \omega_d d^{d/2}$. □

Using Theorems 3.3, 4.1, and 4.2, we get the following inequality:

Corollary 4.3. *If f is of weakly controlled decay on* $\Re^d$, *d odd, then*

$$\|f\|_{\mathcal{H}_d,\mathcal{M}(\Re^d)} \leqslant k_d \|f\|_{d,1,\infty}.$$

By Theorem 2.1 (with $X = \mathcal{L}_2(\Omega)$ and $G = \mathcal{H}_d$, so that $\sup_{g\in G}\|g\| = \lambda(\Omega)^{1/2}$), Corollary 3.2, and Theorem 4.2, we have the following bound on rate of approximation.

Corollary 4.4. *Let d be odd and let* $\Omega \subset \Re^d$ *be measurable with finite nonzero measure. If f is of weakly controlled decay on* $\Re^d$, *then for* $n \geqslant 1$

$$\|f|_\Omega - \mathrm{span}_n\, \mathcal{H}_d\|_{\mathcal{L}_2(\Omega)} \leqslant n^{-1/2}\lambda(\Omega)^{1/2}k_d\|f\|_{d,1,\infty},$$

where $k_d = 2^{1-d}\pi^{1-d/2}d^{d/2}/\Gamma(\frac{d}{2})$.

This implies, for instance, that functions of weakly controlled decay with Sobolev seminorm exponentially large in d (e.g., $2^{d/2}$) can be approximated in $\mathcal{L}_2(\Omega)$ with rates less than or equal to $n^{-1/2}$ by neural nets with n Heaviside perceptrons and a single linear output unit when Ω has unit measure.

Other bounds have been obtained for d both even and odd. Barron [4] estimated rates of approximation by $\mathrm{span}_n\, \mathcal{H}_d$ for functions of bounded half-space variation in the supremum norm on a bounded subset of $\Re^d$ and showed they are $O(n^{-1/2})$. Makovoz [26] showed that on the unit ball in $\Re^d$ the sup-norm distance between $\mathrm{span}_n\, \mathcal{H}_d$ and a function having a suitable integral representation in Heaviside plane waves is at most $Cn^{-1/2-1/2d}\sqrt{\log n}$.

Cheang and Barron [7, pp. 188, 201] state that if Ω is a measurable subset of $\Re^d$ and h is a function on Ω with a Heaviside integral representation with respect to a probability measure on $S^{d-1} \times \Re$, then $\|h - \mathrm{span}_n\, H_d\|_{\mathcal{M}(\Omega)} \leqslant 34 n^{-1/2}\sqrt{d+1}$. It follows that if f has an integral representation on $\Re^d$ with weight function w, then for $n \geqslant 1$

$$\|f - \mathrm{span}_{2n}\, H_d\|_{\mathcal{M}(\Omega)} \leqslant 68 n^{-1/2}\sqrt{d+1}\|w\|_{\mathcal{L}_1(S^{d-1}\times\Re)}. \tag{4}$$

Indeed, if $\|w\|_{\mathcal{L}_1} = \infty$ or 0, the result is trivial. Otherwise, as in [7, pp. 187–189], write $w = w_+ - w_-$, let f_+ and f_- be the corresponding functions, normalize, and apply the probability measure.

By 4.2 and (4) we have

Corollary 4.5. *If f is of weakly controlled decay on $\Re^d$, d odd, then for $n \geqslant 1$*

$$\|f - \mathrm{span}_{2n}\, \mathcal{H}_d\|_{\mathcal{M}(\Re^d)} \leqslant 68 n^{-1/2}\sqrt{d+1}\; k_d \|f\|_{d,1,\infty}.$$

For neural networks with Gaussian radial-basis function units, Girosi [11, Proposition 3.2] derived an analogous expression for sup-norm approximation on $\Re^d$ when the function to be approximated is the convolution of a Bessel potential and an integrable function; also see [21] for some extensions of Girosi's results.

5. Variation with respect to half-spaces and total variation

In this section we establish a formula expressing half-space variation in terms of the total variation of a related one-dimensional function. We also point out a connection with the Radon transform.

Let $a < b$ be finite. For $h : [a, b] \to \Re$ the *total variation* $T_{[a,b]}(h)$ of h on the interval $[a, b]$ is the supremum over all finite partitions $a = a_1 < \cdots < a_k = b$ of the sum $\sum_{j=1}^{k} |h(a_j) - h(a_{j+1})|$. We say that h has *bounded variation* on the interval $[a, b]$ when $T_{[a,b]}(h) < \infty$. Every continuously differentiable function h on $[a, b]$ has bounded variation and $T_{[a,b]}(h) = \int_{[a,b]} |h'(t)|\, dt$ since h is Lipschitz and absolutely continuous. For $h : \Re \to \Re$, the *total variation* $T(h)$ is defined to be the supremum over all finite intervals $[a, b]$ of $T_{[a,b]}(h|_{[a,b]})$, and h is said to be of bounded variation (on $\Re$) when its total variation is finite [28, pp. 215–259].

Given a function f of weakly controlled decay on $\Re^d$ and a unit vector $e \in S^{d-1}$, define $\phi_{f,e}$ on $\Re$ by

$$\phi_{f,e}(b) = \int_{H_{e,b}} D_e^{(d-1)} f(y)\, d_H y. \tag{5}$$

Proposition 5.1. *If f is of weakly controlled decay, d odd, then*

$$\|w_f\|_{\mathcal{L}_1(S^{d-1}\times\Re)} = (1/2)(2\pi)^{1-d}\int_{S^{d-1}} T(\phi_{f,e})\,de.$$

Proof. Since $\int_{H_{e,b}} D_e^{(d)} f(y)\,d_H y = -(\partial/\partial t)(\int_{H_{e,t}} D_e^{(d-1)} f(y)\,d_H y)|_{t=b}$,

$$\int_{S^{d-1}\times\Re} |\int_{H_{e,b}} D_e^{(d)} f(y)\,d_H y\,|\,de\,db = \int_{S^{d-1}}\int_{\Re} |\phi'_{f,e}(b)|\,db\,de = \int_{S^{d-1}} T(\phi_{f,e})\,de.$$

Hence, 5.1 follows from 4.1. □

For a function ψ in $\mathcal{C}(\Re^d)$ with $\psi(x) = O(1/\|x\|^a)$ for $a > d-1$, the *Radon transform* $\mathcal{R}(\psi)$ is defined to be the function on $S^{d-1}\times\Re$ given by

$$\mathcal{R}(\psi)(e,b) = \int_{H_{e,b}} \psi(y)\,d_H y.$$

For d odd and f on $\Re^d$ of weakly controlled decay, the iterated directional derivative in (5) can be replaced by an iterated Laplacian (cf. [20, Proposition 2])), and

$$\phi_{f,e}(b) = \mathcal{R}(\Delta^{\frac{d-1}{2}} f)(e,b).$$

Since $\Delta^{\frac{d-1}{2}} f$ vanishes at infinity to order greater than $d-1$, the right-hand side above is finite.

Proposition 5.1 shows that if f satisfies the conditions of Theorem 4.1, then the half-space variation of f is bounded above by a multiple of the spherical average of the total variation of the Radon transform of an iterated Laplacian of f.

6. Half-space variation of the Gaussian

Consider the Gaussian function γ_d on $\Re^d$ defined as $\gamma_d(x) = \exp(-\|x\|^2))$. It is shown that for d odd, the half-space variation of γ_d is at most $(2\pi d)^{3/4}$. Consequences are given for rate of approximation of the Gaussian by Heaviside perceptron networks in the $\mathcal{L}_2$ and sup norms.

Theorem 6.1. *Let d be odd. Then* $\|w_{\gamma_d}\|_{\mathcal{L}_1(S^{d-1}\times\Re)} \leqslant (2\pi d)^{3/4}$.

Proof. Let e_1 denote the standard unit vector along the x_1-axis and write γ for γ_1. Then $\|\gamma_d\|_{\mathcal{L}_1(\Re^d)} = \|\gamma\|^d_{\mathcal{L}_1(\Re)} = \pi^{d/2}$, and by Proposition 5.1 we have

$$\begin{aligned}\|w_{\gamma_d}\|_{\mathcal{L}_1} &= (1/2)(2\pi)^{1-d}\int_{S^{d-1}} T(\phi_{\gamma_d,e})\,de\\ &= |a_d|\omega_d \int_{\Re}\left|\int_{H_{e_1,b}} (D_{e_1}^{(d)}\gamma_d)(y)\,d_H y\right| db\\ &= |a_d|\omega_d \int_{\Re} |\gamma^{(d)}(b)|\,db \int_{\Re^{d-1}} \gamma_{d-1}(y)\,dy = l_d T(\gamma^{(d-1)}),\end{aligned}$$

where $l_d = |a_d|\omega_d\pi^{(d-1)/2} = (1/2)(2\pi)^{1-d}\frac{2\pi^{d/2}}{\Gamma(d/2)}\pi^{(d-1)/2} = \frac{2^{1-d}\sqrt{\pi}}{\Gamma(d/2)} = \frac{2^{(1-d)/2}}{(d-2)(d-4)\cdots 1}$.

Szász [32] proved the following result: for $d \geqslant 0$ and all real x

$$|H_d(x)|e^{-x^2/2} \leqslant (2^d d!)^{1/2}.$$

It follows that

$$T(\gamma^{(d-1)}) = \int_{\mathcal{R}} |H_d(x)|e^{-x^2}\, dx \leqslant (2^d d!)^{1/2}\sqrt{2\pi}.$$

However, using Stirling's approximation [1, p. 257], for d an odd integer $\geqslant 3$,

$$(2^d d!)^{1/2}\sqrt{2\pi}\, l_d \leqslant (2\pi d)^{3/4} \left(\frac{d-1}{d}\right)^{3/4} \exp\left(\frac{1}{6(d-1)}\right) \leqslant (2\pi d)^{3/4}.$$

For $d = 1$, the left-hand side equals $\sqrt{4\pi} < (2\pi)^{3/4}$. Hence, $\|w_{\gamma_d}\|_{\mathcal{L}_1} \leqslant (2\pi d)^{3/4}$. □

Combining 2.1 (with $\sup_{g \in G} \|g\| = \lambda(\Omega)^{1/2}$), 3.2, and 6.1, we obtain

Corollary 6.2. *Let d be odd and $n \geqslant 1$. If $\Omega \subset \Re^d$ has finite nonzero measure, then*

$$\|\gamma_d - \mathrm{span}_n\, \mathcal{H}_d\|_{\mathcal{L}_2(\Omega)} \leqslant (2\pi d)^{3/4}\, \lambda(\Omega)^{1/2}\, n^{-1/2}.$$

By (4) and Theorem 6.1, for d odd

$$\|\gamma_d - \mathrm{span}_{2n}\, \mathcal{H}_d\|_{\mathcal{M}(\Re^d)} \leqslant 68(2\pi d)^{3/4}\, n^{-1/2}\sqrt{d+1}. \tag{6}$$

By a change of variables, for all $n \geqslant 2$

$$\|\gamma_d - \mathrm{span}_n\, \mathcal{H}_d\|_{\mathcal{M}(\Re^d)} \leqslant 68(2\pi d)^{3/4}\sqrt{2}\,(n-1)^{-1/2}\sqrt{d+1}. \tag{7}$$

In contrast, in [7, p. 189] Cheang and Barron find that for d even and odd,

$$\|\gamma_d(\cdot/\sqrt{2}) - \mathrm{span}_{2n}\, \mathcal{H}_d\|_{\mathcal{M}(B_r)} \leqslant 68\; r\; n^{-1/2}\sqrt{d(d+1)}, \tag{8}$$

where B_r is the ball of radius r in $\Re^d$. By a change of variables, (8) implies

$$\|\gamma_d - \mathrm{span}_{2n}\, \mathcal{H}_d\|_{\mathcal{M}(B_r)} \leqslant 68\sqrt{2}\; r\; n^{-1/2}\sqrt{d(d+1)}. \tag{9}$$

Thus (6) gives a better sup-norm bound on B_r than (9) when $r > (2\pi d)^{3/8}$ and d is odd.

We thank one of the referees for pointing out Szász's result and its citation by Szegő [33, p. 190]; see also [3, p. 340]. One can prove a different result, which replaces $(2\pi d)^{3/4}$ by $2d$ (again for d odd); for $d < \pi^3/2$, the $2d$ bound is better. The $2d$ bound estimates the total variation of a derivative of the one-dimensional Gaussian as twice the number of local extrema multiplied by the largest local extremum which must occur at zero by a theorem of Sonin and Polya (see also Butlewski) [33, p. 166].

References

[1] M. Abramowitz, I.A. Stegun, Handbook of Mathematical Functions, 8th Dover Printing (10th of GPO), New York, 1973.

[2] R.A. Adams, J.J.F. Fournier, Sobolev Spaces, second ed., Academic Press, Amsterdam, 2003.

[3] G.E. Andrews, R. Askey, R. Roy, Special Functions, Cambridge University Press, Cambridge, 1999.

[4] A.R. Barron, Neural net approximation, in: Proceedings of the Seventh Yale Workshop on Adaptive and Learning Systems, 1992, pp. 69–72.

[5] A.R. Barron, Universal approximation bounds for superposition of a sigmoidal function, IEEE Trans. Inform. Theory 39 (1993) 930–945.

[6] S.M. Carroll, B.W. Dickinson, Construction of neural nets using the Radon transform, in: Proceedings of IJCNN, IEEE Press, New York, 1989, pp. 607–611.

[7] G.H.L. Cheang, A.R. Barron, A better approximation for balls, J. Approx. Theory 104 (2000) 183–203.

[8] R. Courant, Differential and Integral Calculus, vol. 2, Wiley, New York, 1936 (1964 ed., transl. E.J. McShane).

[9] R. Courant, D. Hilbert, Methods of Mathematical Physics, vol. 1, Wiley, New York, 1989.

[10] C.H. Edwards, Advanced Calculus of Several Variables, Dover, New York, 1994.

[11] F. Girosi, Approximation error bounds that use VC-bounds, in: Proceedings of the International Conference on Artificial Neural Networks (ICANN'95), EC2 & Cie, Paris, 1995, pp. 295–302.

[12] F. Girosi, G. Anzellotti, Rates of convergence for radial basis function and neural networks, in: Artificial Neural Networks for Speech and Vision, Chapman & Hall, London, 1993, pp. 97–113.

[13] L. Gurvits, P. Koiran, Approximation and learning of convex superpositions, J. Comput. System Sci. 55 (1997) 161–170.

[14] S. Helgason, The Radon Transform, Birkhäuser, Boston, 1980.

[15] Y. Ito, Representation of functions by superpositions of a step or sigmoid function and their applications to neural network theory, Neural Networks 4 (1991) 385–394.

[16] 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 (1992) 608–613.

[17] P.C. Kainen, V. Kůrková, A. Vogt, Geometry and topology of continuous best and near best approximations, J. Approx. Theory 105 (2000) 252–262.

[18] P.C. Kainen, V. Kůrková, A. Vogt, An integral formula for Heaviside neural networks, Neural Network World 3 (2000) 313–319.

[19] P.C. Kainen, V. Kůrková, A. Vogt, Best approximation by Heaviside perceptron networks, Neural Networks 13 (2000) 695–697.

[20] P.C. Kainen, V. Kůrková, A. Vogt, Integral combinations of Heavisides, 2005, submitted for publication (see ftp.cs.cas.cz/pub/reports/v968-06.pdf).

[21] M.A. Kon, L.A. Raphael, D.A. Williams, Extending Girosi's approximation estimates for functions in Sobolev spaces via statistical learning theory, J. Anal. Appl. 3 (2005) 67–90.

[22] V. Kůrková, Dimension-independent rates of approximation by neural networks, in: K. Warwick, M. Kárný (Eds.), Computer-Intensive Methods in Control and Signal Processing: Curse of Dimensionality, Birkhauser, Boston, 1997, pp. 261–270.

[23] V. Kůrková, High-dimensional approximation and optimization by neural networks, in: J. Suykens et al. (Eds.), Advances in Learning Theory: Methods, Models and Applications, IOS Press, Amsterdam, 2003, pp. 69–88, (Chapter 4).

[24] V. Kůrková, P.C. Kainen, V. Kreinovich, Estimates of the number of hidden units and variation with respect to half-spaces, Neural Networks 10 (1997) 1061–1068.

[25] V. Kůrková, M. Sanguineti, Comparison of worst-case errors in linear and neural network approximation, IEEE Trans. Inform. Theory 48 (2002) 264–275.

[26] Y. Makovoz, Uniform approximation by neural networks, J. Approx. Theory 95 (1998) 215–228.

[27] H.N. Mhaskar, C.A. Micchelli, Dimension independent bounds on the degree of approximation by neural networks, IBM J. Res. Develop. 38 (1994) 277–284.

[28] I.P. Natanson, Theory of Functions of a Real Variable, vol. I, Ungar, New York, 1961 (Transl. L.F. Boron with editorial annotations by E. Hewitt).

[30] A. Pinkus, Approximation theory of the MLP model in neural networks, Acta Numer., 1999, 143–195.

[31] G. Pisier, Remarques sur un résultat non publié de B. Maurey, in: Séminaire d'Analyse Fonctionnelle 1980–81, Exposé, no. V, École Polytechnique, Centre de Mathématiques, Palaiseau, France, 1980, pp. V.1–V.12.

[32] O. Szász, On the relative extrema of the Hermite orthogonal functions, J. Indian Math. Soc. 25 (1951) 129–134.

[33] G. Szegő, Orthogonal Polynomials, American Mathematical Society Colloquium Series XXIII, Providence, RI, 1975.