A Sobolev-Type Upper Bound for Rates of Approximation by Linear Combinations of Plane Waves

Kainen, P.C.
2004

Institute of Computer Science
Academy of Sciences of the Czech Republic

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

Paul C. Kainen, Věra Kůrková, Andrew Vogt

Technical report No. 900

January 2004

Pod Vodárenskou věží 2, 182 07 Prague 8, phone: +420 266 053 111, fax: +420 286 585 789,
e-mail:ics@cs.cas.cz

Institute of Computer Science
Academy of Sciences of the Czech Republic

A Sobolev-type upper bound for rates of approximation by linear combinations of plane waves[1]

Paul C. Kainen[2], Věra Kůrková[3], Andrew Vogt[4]

Technical report No. 900

January 2004

Abstract:

Rates of approximation of smooth functions of d variables by linear combinations of n characteristic functions of half-spaces are investigated. It is shown that functions from balls in Sobolev seminorm $\|.\|_{d,1,\infty}$ can be approximated within $k_d/\sqrt{n}$, where k_d as a function of d is decreasing exponentially fast. The upper bound on rates of approximation is obtained from comparison of balls in Sobolev seminorm $\|.\|_{1,d,\infty}$ (defined as the maximum of the $\mathcal{L}_1$-norms of the d-th derivatives of f) and a norm called variation with respect to half-spaces. It is shown that if f is any function on $\Re^d$ (d odd) with all partial derivatives of order up to $d+1$ vanishing sufficiently rapidly at infinity, then variation with respect to half-spaces of f is at most k_d times $\|f\|_{1,d,\infty}$. The result are applied to the Gaussian function $\exp(\|x\|^2)$, its variation with respect to half-spaces is estimated from above by $2d$.

Keywords:
Rates of approximation by plane waves, characteristic functions of half-spaces, integral representation, variation with respect to half-spaces, Sobolev-type norm, neural networks, total variation, Radon transform, directional derivative, Gaussian function.

[1] V. K. was partially supported by GA ČR grant number 201/02/0428
[2] Department of Mathematics, Georgetown University, Washington, D.C. 20057-1233
[3] Institute of Computer Science, Academy of Sciences of the Czech Republic, Pod vodárenskou věží 2, 182 07 Prague 8, Czech Republic
[4] Department of Mathematics, Georgetown University, Washington, D.C. 20057-1233

1 Introduction

Upper bounds on rates of approximation, which depend on the number of variables of the functions to be approximated and on other characteristics of these functions, tell us how to compensate for an increase in the number of variables. In linear approximation of functions from balls in Sobolev spaces, there are tight estimates of worst-case errors of functions from balls in Sobolev spaces of the order of $\mathcal{O}(n^{-s/d})$, where d is the number of variables, s is the smoothness parameter for the Sobolev space, and n the dimension of the linear space of approximating functions [27, p. 232].

In nonlinear approximation of variable-basis type [20], balls in "variational" norms tailored to a type of a basis play a similar role as balls in Sobolev norm in the linear case. The Maurey-Jones-Barron theorem [28, 13, 3] states that all functions in the unit balls can be approximated within $\frac{1}{\sqrt{n}}$ by linear combinations of at most n functions from a given basis.

Thus to compare approximation capabilities of variable-basis approximation scheme with the linear one, one needs to investigate relationships between balls in Sobolev norms and balls in variational norms.

An important case of a family of functions belonging to the variable-basis approximation scheme is formed by linear combinations of n characteristic functions of half-spaces of $\Re^d$. Such functions can be computed by neural networks with n perceptrons with the Heaviside activation function.

The corresponding norm, introduced by Barron [2], is called *variation with respect to half-spaces*. Barron [3] described many types of functions with "small" variation while Kůrková, Savický and Hlaváčková [24] gave examples of classes of functions of d variables with variation depending either polynomially or exponentially on d.

The idea of using an integral formula to estimate variation from above for suitably smooth functions was introduced by Barron [3]. Further work by Girosi and Anzellotti [10] and by Kůrková, Kainen and Kreinovich [21] has extended this approach.

In this paper we utilize an integral formula from [18] which holds for functions that are sufficiently smooth and which decay sufficiently rapidly at infinity (called "polynomially vanishing") expressing them as a weighted combination of characteristic functions of closed half-spaces as in [16] and [21] rather than of trigonometric functions which were used in [3].

We show that any polynomially vanishing function on $\Re^d$, d odd, has finite variation with respect to half-spaces. Moreover, and the unit ball in variation with respect to half-spaces contains all polynomially vanishing functions which have all d-th order partials with $\mathcal{L}_1$-norm not exceeding $k_d^{-1} \sim (2\pi/e)^{d/2}$.

An outline of the paper follows: Section 2 covers variable basis approximation, the Maurey-Jones-Barron Theorem, variation with respect to a family of functions in a normed linear space, and the connection with neural networks. Section 3 gives upper bounds on variation with respect to the family of half-space characteristic functions, in both sup and $\mathcal{L}_2$-norms, establishing an upper bound in terms of the $\mathcal{L}_1$-norm of the weight function for functions which have suitable integral representations. The concept of polynomially vanishing functions on $\Re^d$ is introduced in Section 4 and an integral representation of a suitable type is given for them. The $\mathcal{L}_1$-norm of the weight function for this choice of integral representation for polynomially vanishing functions is shown to be bounded by an exponentially decreasing constant $k_d, k_d \sim \left(\frac{e}{2\pi}\right)^{d/2}$, times the maximum of all $\mathcal{L}_1$-norms of d-th order partials. Section 5 provides a bound on variation with respect to half-spaces using an integral average of the total variations of Radon transforms of iterated directional derivatives. The variation with respect to half spaces of the Gaussian is bounded in section 6, and in the last section, we point out where the results can be improved.

2 Rates of variable-basis approximation

Let $(X, \|.\|)$ be a normed linear space with a nonempty subset G. *Variable-basis approximation* (with respect to G) of $f \in X$ is the process of decreasing the distance $\|f - \mathrm{span}_n G\|$ from f to $\mathrm{span}_n G$ with n increasing. For $n \geq 1$, $\mathrm{span}_n G$ is the set of all n-fold linear combinations of elements from G. In the literature, this process is also called "projection pursuit" [13] and "approximation from a dictionary" [25].

The Maurey-Jones-Barron Theorem gives an upper bound on the rate of variable-basis approximation (see [28, p.V.2, Lemma 2]), [13, p. 611], and [3, p. 934]). This result was reformulated in [19] using a norm which extends the concept of variation with respect to half-spaces [2].

For a nonempty subset G of a normed linear space $(X, \|.\|)$, a norm, called G-*variation*, was defined in [19] as the Minkowski functional of the closed convex symmetric hull of G. Using a subscript to denote the dependence of the variational norm on G,

$$\|f\|_G = \inf\left\{c > 0 \,:\, c^{-1} f \in \mathrm{cl}\,\mathrm{conv}\,(G \cup -G)\right\},$$

where closure is with respect to the norm on X (when necessary, an additional subscript will be used to indicate which norm). See [23], [20], [22] for further properties and applications of this norm.

When X is a set of real-valued multivariable functions defined on $\Re^d$ or some subset, subject to $\mathcal{L}_p$- or sup-norm, the variational norm depends on d. If G consists of functions g of the form $g(x) = \psi(a \cdot x + b)$, where $a \in \Re^d$ and b is real with $\psi : \Re \to \Re$, then $\|f\|_G$ is the infimum of all constants c for which f is contained in the $\|\cdot\|_X$-closure of the convex hull of $c(G \cup -G)$, where as usual for S any subset of the normed linear space X, for c a real constant, $c\,S$ means the set of all multiples cs for $s \in S$ and $c\,\mathrm{cl}_X(U) = \mathrm{cl}_X(c\,U)$. Variable basis approximation in this context means that one may adjust the functions whose linear combination approximates f by tuning the parameters a and b.

For neural nets, G corresponds to the functions produced by a single hidden unit, with activation function ψ, and the parameters correspond to input weights and bias, respectively. See, e.g., [20].

The Maurey-Jones-Barron Theorem can now be stated as follows [**?**].

Theorem 2.1 *If $(X, \|.\|)$ is a Hilbert space, G a bounded subset and $s_G = \sup_{g\in G} \|g\|$, then for every $f \in X$ and every positive integer n,*

$$\|f - \mathrm{span}_n G\| \leq \frac{\sqrt{(s_G\,\|f\|_G)^2 - \|f\|^2}}{n^{1/2}}.$$

While the denominator is $n^{1/2}$, independent of the number d of variables, the numerator does depend on d. To effectively employ Theorem 2.1, we have to characterize functions with small G-variation. When G is the family of half-space characteristic functions, we provide such conditions below in Corollaries 3.4,4.3, and 5.2.

3 Upper bounds on variation with respect to half-spaces

Let $\Re^d$ denote the usual d-dimensional Euclidean space; we write S^{d-1} to denote the sphere of norm-1 vectors in $\Re^d$. In this section we derive upper bounds on variation with respect to the set of characteristic functions of closed half-spaces of $\Re^d$, introduced by Barron [2] as *variation with respect to half-spaces*.

We study variation with respect to half-spaces in two different normed linear spaces. The first one is the Hilbert space $(\mathcal{L}_2(\Omega), \|.\|_2)$ of square integrable functions with respect to Lebesgue measure λ. When Ω is a subset of $\Re^d$ of finite nonzero Lebesgue measure, then the set $H_d(\Omega)$ of characteristic functions of intersections of Ω with closed half-spaces of $\Re^d$ is a bounded subset of $\mathcal{L}_2(\Omega)$; we denote this variation by $\|.\|_{H_d(\Omega),\mathcal{L}_2}$.

The second type of normed linear space considered is $(\mathcal{M}(\Omega), \|.\|_{\sup})$ of bounded measurable functions on a subset $\Omega \subseteq \Re^d$ with the supremum norm; we denote variation here by $\|.\|_{H_d(\Omega),\sup}$.

It is easy to see that characteristic functions of half-spaces are compositions of affine functions with the *Heaviside* function $\vartheta : \Re \to \Re$ given by $\vartheta(t) = 0$ for $t < 0$ and $\vartheta(t) = 1$ for $t \geq 0$. Since $\vartheta(rt) = \vartheta(t)$ for $r > 0$, we have

$$H_d(\Omega) = \{\vartheta_{e,b}\,|e \in S^{d-1}, b \in \Re\}$$

where $\vartheta_{e,b} : \Omega \to \Re$ is given by

$$\vartheta_{e,b}(x) = \vartheta(e \cdot x + b).$$

To derive consequences of Theorem 2.1 for approximation by $\mathrm{span}_n H_d(\Omega)$ in the Hilbert space $\mathcal{L}_2(\Omega)$, we need upper bounds on $\|.\|_{H_d(\Omega),\mathcal{L}_2}$. The next two propositions show that it suffices to get upper bounds on $\|.\|_{H_d(\Re^d),\sup}$.

Proposition 3.1 *Let $d \geq 1$ and let $\Omega \subseteq \Re^d$ with $0 < \lambda(\Omega) < \infty$. Then for every $f \in \mathcal{L}_2(\Omega) \cap \mathcal{M}(\Omega)$*

$$\|f\|_{H_d(\Omega),\mathcal{L}_2} \leq \|f\|_{H_d(\Omega),\sup}.$$

Proof. Suppose $\|f\|_{H_d(\Omega),\sup} = t < \infty$. Let $\varepsilon > 0$ be given. Then there exist $(e_1, b_1), \ldots, (e_k, b_k) \in S^{d-1} \times \Re$ and $c_1, \ldots, c_k \in \Re$ such that

$$\|f - \sum_{i=1}^{k} c_i \vartheta_{e_i,b_i}\|_{\sup} < \varepsilon/\lambda(\Omega)^{1/2},$$

where $\sum |c_i| \leq t$. But then

$$\|f - \sum_{i=1}^{k} c_i \vartheta_{e_i,b_i}\|_{\mathcal{L}_2} < \varepsilon,$$

□

Proposition 3.2 *Let $d \geq 1$ and let $\Omega \subseteq \Re^d$. Then for every bounded $f \in \mathcal{C}(\Omega)$ and for every $\bar{f} \in \mathcal{M}(\Re^d)$ such that $\bar{f}|\Omega = f$,*

$$\|f\|_{H_d(\Omega),\sup} \leq \|\bar{f}\|_{H_d(\Re^d),\sup}$$

Proof. Sup norm on a subset cannot exceed that on the ambient space. Hence, the same argument as in the previous proof works with no need to change the epsilon. □

The next theorem gives an upper bound on variation with respect to half-spaces in $(\mathcal{M}(\Re^d), \|.\|_{\sup})$ for functions that can be expressed as integrals of plane waves.

Theorem 3.3 *If d is a positive integer and $f \in \mathcal{M}(\Re^d)$ can be expressed as $f(x) = \int_{S^{d-1}\times\Re} w(e,b)\vartheta(e \cdot x + b)dedb$, where w is continuous on $S^{d-1} \times \Re$ and $\int_{S^{d-1}\times\Re} |w(e,b)|dedb < \infty$, then*

$$\|f\|_{H_d(\Re^d),\sup} \leq \int_{S^{d-1}\times\Re} |w(e,b)|dedb.$$

Proof. To abbreviate notation, we denote by P the cylinder $S^{d-1} \times \Re$, so $p \in P$ means $p = (e, b)$. Also, we write ϑ_p for the function $\vartheta_p(x) = \vartheta(e \cdot x + b)$. Let λ_P denote the induced measure on P which is the product of the usual measure on the sphere S^{d-1} and Lebesgue measure on $\Re$. Thus, the induced measure on P is obtained by restricting a power of one-dimensional Lebesgue measure to P. Similarly, there is a metric on P induced by restricting the standard euclidean metric in $\Re^{d+1}$. Thus, for $P'' = S^{d-1} \times [0, 1]$, $\lambda_P(P'')$ is 2 when $d = 1$ and is 2π when $d = 2$, while the diameter of P'' is $\sqrt{5}$.

Let $\varepsilon > 0$ be arbitrary. As the $\mathcal{L}_1$-norm of w is finite, we can choose a closed interval $I' \subset \Re$ so that with $P' = S^{d-1} \times I'$, $\int_{P\setminus P'} |w(p)|d\lambda_P(p) < \varepsilon/3$.

Let $F(x) = \int_{P'} w(p)\vartheta_p(x)d\lambda_P(p)$. Then, since $\vartheta(t) \leq 1$, $\sup_{x\in\Re^d} |f(x)-F(x)| \leq \int_{P\setminus P'} |w(p)|d\lambda_P(p)$. Hence, $\|f - F\|_{\sup} < \varepsilon/3$.

Thus, to prove the theorem, it suffices by the definition of H_d-variation to show that within $2\varepsilon/3$ of F in supremum norm on $\Re^d$ there is a finite linear combination of characteristic functions of half-spaces where the sum of the absolute values of the coefficients does not exceed $\int_P |w_f| d\lambda_P(p)$. We shall obtain such characteristic functions and their coefficients from a sufficiently fine subdivision of P' into compact sets.

As w is continuous and P' is compact, both $\lambda_P(P')$ and $W' =: \sup\{|w(p)| : p \in P'\}$ are finite. Let δ_w denote the modulus of continuity of w on P', so if $|p - p'| < \delta_w(\varepsilon)$, then $|w(p) - w(p')| < \varepsilon$ for all $p, p' \in P'$.

Choose a finite family $\mathcal{R}$ of subsets of P' such that:

(1) $P' = \bigcup \mathcal{R}$;
(2) All $R \in \mathcal{R}$ are compact and connected;
(3) $\lambda_P(R \cap R') = 0$ for all $R \neq R'$ in $\mathcal{R}$;
(4) For each $R \in \mathcal{R}$,

$$\operatorname{diam}(R) \leq \delta_w \left(\frac{\varepsilon}{3\lambda_P(P')} \right);$$

That is, we choose an essentially disjoint covering of P' by compact, connected sets with mesh not exceeding $\delta_w(\varepsilon/3\lambda_P(P'))$.

It is clear that there are many ways to do this. For any such family $\mathcal{R}$ and for each x in $\Re^d$, let $\mathcal{R}_x$ denote the set of all $R \in \mathcal{R}$ which contain some interior point $p = (e, b)$ for which $x \in H_{e,b}$, i.e., so that $e \cdot x + b = 0$. These are exactly those R on which $\vartheta_p(x)$ is not constant, viewed as a function of $p \in R$.

We shall choose $\mathcal{R}$ so that (1) through (4) hold and also:

(5) For all $x \in \Re^d$, $\sum_{R \in \mathcal{R}_x} \lambda_P(R) < \varepsilon/6W'$.

Let $\tau > 0$ be given and let $\mathcal{R}$ be the product tesselation of $S^{d-1} \times I'$ produced as follows: Consider the cubical tesselation of $[0, \pi] \times [0, \pi/2] \times \ldots \times [0, \pi/2]$ with all sides having length at most τ. Using spherical coordinates, this gives a tesselation of the sphere with mesh at most $\tau\sqrt{d}$. Subdivide I' into intervals of length τ. For the product tesselation $\mathcal{R}$ of $S^{d-1} \times I'$, the mesh is at most $\tau\sqrt{d+1}$.

For any $x \in \Re^d$, let $(x, 1)$ denote the element of $\Re^{d+1}$ with last coordinate 1 and projection to the first d coordinates equal to x. Then $(x, 1)^\perp = \{(e, b) : e \cdot x + b = 0\}$ is the orthogonal hyperplane. Let $A_x = (S^{d-1} \times \Re) \cap (x, 1)^\perp$. Then $A_x = \{z \in \Re^{d+1} : \sum_{j=1}^d z_j x_j = -z_{d+1}; \sum_{j=1}^d z_j^2 = 1\}$ is an algebraic variety of dimension $d - 1$. If $(u, c) \in (x, 1)^\perp$, then $c = -u \cdot x$ so if $u \in S^{d-1}$, then $|c| \leq \|x\|$ by the Cauchy-Schwartz inequality.

For all $x \in \Re^d$ put $A'_x = A_x \cap S^{d-1} \times I'$ and let $\mathcal{U}_x = \bigcup\{R \in \mathcal{R} : R \cap A'_x \neq \emptyset\}$, where $\mathcal{R}$ is the tesselation chosen following (5). Then $\mathcal{U}_x \subseteq \{(e, b) \in S^{d-1} \times I' : b \in [(-x \cdot e) - \tau, (-x \cdot e) + \tau] \cap I'$ so

$$\sum_{R \in \mathcal{R}_x} \lambda_P(R) \leq \lambda_P(\mathcal{U}_x) \leq \tau\omega_d.$$

Hence, if in the definition of $\mathcal{R}$ we require

$$\tau < \min\left(\frac{\varepsilon}{6W'\omega_d},\ \delta_w\left(\frac{\varepsilon}{3\lambda_P(P')\sqrt{d+1}} \right) \right),$$

then all five conditions hold.

As w is continuous and all R are compact, for each $R \in \mathcal{R}$ there exists $p_R \in R$ with $|w(p_R)| = \min\{|w(p)| : p \in R\}$. Set $c_R = w(p_R)\lambda_P(R)$.

We claim that $g = \sum_{R \in \mathcal{R}} c_R \vartheta_{p_R}$ satisfies $\sup_{x \in \Re^d} |g(x) - F(x)| < 2\varepsilon/3$. As $\sum_{R \in \mathcal{R}} |c_R| \leq \int_{P'} |w(p)| d\lambda_P(p) \leq \int_P |w(p)| d\lambda_P(p)$, this will prove the theorem.

It is easy to verify that for each x in $\Re^d$, the following inequality holds:

$$|g(x) - F(x)| \leq \left(\sum_{R \in \mathcal{R} \setminus \mathcal{R}_x} + \sum_{R \in \mathcal{R}_x} \right) \left(\int_R |w(p_R)\vartheta_{p_R}(x) - w(p)\vartheta_p(x)| d\lambda_P(p) \right).$$

The first sum is less than $\varepsilon/3$. Indeed, since $\vartheta_p(x)$ is constant on such R, each summand is at most $\int_R |w(p_R) - w(p)| d\lambda_P(p)$ and, hence, by (4) at most $\lambda_P(R)\varepsilon/3\lambda_P(P')$, which suffices. The second sum is also less than $\varepsilon/3$. For the summands are at most $\int_R 2W' d\lambda_P(p)$, so the second sum is at most $2W' \sum_{R\in\mathcal{R}_x} \lambda_P(R)$, which suffices by (5). Thus, the claim and theorem hold. □

By Propositions 3.1 and 3.2 we have the following consequence.

Corollary 3.4 *For $d \geq 1$ and $\Omega \subseteq \Re^d$ with $0 < \lambda(\Omega) < \infty$, let $f \in \mathcal{L}_2(\Omega) \cap \mathcal{M}(\Omega)$ and let $\bar{f} \in \mathcal{M}(\Re^d)$ satisfy $\bar{f}|\Omega = f$. If $\bar{f}(x) = \int_{S^{d-1}\times\Re} w(e,b)\vartheta(e\cdot x + b) dedb$ for some continuous function w on $S^{d-1}\times\Re$ with $\int_{S^{d-1}\times\Re} |w(e,b)| dedb < \infty$, then*

$$\|f\|_{H_d(\Omega),\mathcal{L}_2} \leq \int_{S^{d-1}\times\Re} |w(e,b)| dedb.$$

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

To take advantage of Corollary 3.4, we use an integral representation in terms of Heaviside plane waves.

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

$$D_e^{(r)} = \sum_{|\alpha|=r} \binom{r}{\alpha} e^\alpha (D^\alpha) \tag{4.1}$$

where α denotes a multi-index, that is, a length-d vector of nonnegative integers $(\alpha_1, \ldots, \alpha_d)$, $|\alpha|$ is the sum of the coordinates (the *degree* of the multi-index), and $\binom{r}{\alpha}$ is the multinomial coefficient with value $r!/\alpha_1! \cdots \alpha_d!$; for a vector e, v^α denotes $v_1^{\alpha_1} \cdots v_d^{\alpha_d}$; and $D^\alpha f$ is the partial derivative with differentiation of order α_j with respect to x_j, $1 \leq j \leq d$ (see, e.g., [8, p. 130]). Also, we write $|e|^\alpha$ for $|e_1|^{\alpha_1} \cdots |e_d|^{\alpha_d}$.

For $f \in \mathcal{C}^d(\Re^d)$, $\|f\|_{d,1}$ denotes the Sobolev norm $\sum_{|\alpha|\leq d} \|D^\alpha f\|_{\mathcal{L}_1(\Re^d)}$ [1, p. 59]. Our estimates, however, use a *mixed Sobolev seminorm* (cf. [1, pp. 50,59,101])

$$\|f\|_{d,1,\infty} = \max_{|\alpha|=d} \|D^\alpha f\|_{\mathcal{L}_1(\Re^d)}. \tag{4.2}$$

The reader can easily check that this functional is, indeed, a seminorm. Clearly, $\|f\|_{d,1,\infty} \leq \|f\|_{d,1}$.

To estimate variation with respect to half-spaces in terms of the mixed Sobolev seminorm, we use an integral representation [18] (see also [16]) valid for the class of polynomially vanishing functions. A function f on $\Re^d$ is called *polynomially vanishing at infinity* (or shortly *polynomially vanishing*) if $f \in \mathcal{C}^{d+1}(\Re^d)$ and there exists $\varepsilon > 0$ such that for each multi-index α with $|\alpha| \leq d+1$

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

where $\|.\|$ denotes l_2-norm on $\Re^d$.

Clearly, polynomially vanishing functions include the Gaussian function $\gamma(x) = \exp(-\|x\|^2)$ and all other "rapidly decreasing" [1, p. 251] smooth functions in the Schwartz class $\mathcal{S}$ consisting of all $f : \Re^d \to \Re$ such that $\forall\alpha, \forall\beta)(\sup_{x\in\Re^d} x^\alpha D^\beta f(x) < \infty)$. In particular, the polynomially vanishing functions include all sufficiently smooth functions of compact support, and the integral formula for polynomially vanishing functions extends the integral formula [21] for functions of compact support. Polynomially vanishing functions on $\Re^d$ satisfy $\|D^\alpha f\|_{\mathcal{L}_1} < \infty$ if $|\alpha| \geq d$ [18].

By $H_{e,b}$ we denote the hyperplane determined by $e \in S^{d-1}$ and $b \in \Re^d$, i.e., $H_{e,b} = \{x \in \Re^d : e\cdot x + b = 0\}$.

The following integral representation was derived in [21], [16], and [18]. The latter two references give an integral representation which holds when d is even, but the weight function for that case requires an additional logarithmic factor. In this paper, we only consider the case d odd.

Theorem 4.1 *If $d > 0$ is an odd integer and f is a polynomially vanishing function on $\Re^d$, then*

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

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

Note that for f polynomially vanishing, w_f is continuous. Indeed, ... some justification here - future problem: modulus of continuity of w_f. We next give a bound on $\|f\|_{\mathcal{L}_1}$ for polynomially vanishing f.

Let $\omega_d = \lambda(S^{d-1})$. It is well-known (e.g., [6, p.303]) that $\omega_d = 2\pi^{d/2}/\Gamma(\frac{d}{2})$, where for $x > 0$, $\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}dt$. Stirling's approximation (e.g., [11, p. 165]) is $\Gamma(x+1) \sim \sqrt{2\pi x}(x/e)^x$, where $r(x) \sim s(x)$ means $\lim_{x\to\infty} r(x)/s(x) = 1$. Although asymptotic, Stirling's approximation is within one percent of equality even for $d = 9$ (see, e.g., [11, p. 170]).

Theorem 4.2 *If $d > 0$ is an odd integer and $f \in \mathcal{C}^{d+1}(\Re^d)$ is polynomially vanishing, then*

$$\int_{S^{d-1}} \int_{\Re} |w_f(e,b)|dedb \le k_d \|f\|_{d,1,\infty},$$

where $k_d = |a_d|\omega_d d^{d/2} \sim \left(\frac{4\pi}{d}\right)^{1/2}\left(\frac{e}{2\pi}\right)^{d/2}$.

Proof. Note that $k_d = 2^{1-d}\pi^{1-d/2}d^{d/2}/\Gamma(\frac{d}{2}) = 2\pi(\frac{d}{4\pi})^{d/2}/\Gamma(\frac{d}{2}) \sim (\frac{4\pi}{d})^{1/2}(\frac{e}{2\pi})^{d/2}$ so k_d is exponentially decreasing as a function of d.

By Theorem 4.1 we have $\int_{S^{d-1}} \int_{\Re} |w_f(e,b)|dedb \le |a_d| \int_{S^{d-1}} \int_{\Re} |D_e^d(f)|d_H y$. Using standard properties of the integral, the definition of the Sobolev seminorm, and the multinomial theorem, we have

$$\int_{S^{d-1}} \int_{\Re} \left| \int_{H_{e,b}} D_e^d(f)d_H y \right| \le \int_{S^{d-1}} \int_{\Re} \int_{H_{e,b}} \sum_{|\alpha|=d} \binom{d}{\alpha} \left| e^\alpha (D^\alpha f)(y) \right| d_H y dbde =$$

$$\int_{S^{d-1}} \int_{\Re^d} \sum_{|\alpha|=d} \binom{d}{\alpha} \left| e^\alpha (D^\alpha f)(y) \right| dyde = \int_{S^{d-1}} \sum_{|\alpha|=d} \binom{d}{\alpha} |e^\alpha| \int_{\Re^d} \left| (D^\alpha f)(y) \right| dyde$$

$$\le \int_{S^{d-1}} \sum_{|\alpha|=d} \binom{d}{\alpha} |e|^\alpha \|f\|_{d,1,\infty} de = \|f\|_{d,1,\infty} \int_{S^{d-1}} (\sum_{i=1}^d |e_i|)^d de.$$

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

By Theorem 2.1 and 4.2, we have the following bound on rate of approximation.

Corollary 4.3 *If $d > 0$ is an odd integer, $\Omega \subseteq \Re^d$ with $0 < \lambda(\Omega) < \infty$, $f \in \mathcal{L}_2(\Omega)$, and $\bar{f} \in \mathcal{C}^{d+1}(\Re^d)$ is polynomially vanishing with $\bar{f}|\Omega = f$, then for every positive integer n*

$$\|f - \text{span}_n H_d\|_{\mathcal{L}_2(\Omega)} \le \frac{k_d \|\bar{f}\|_{d,1,\infty}}{\sqrt{n}}.$$

where $k_d \sim \left(\frac{4\pi}{d}\right)^{1/2}\left(\frac{e}{2\pi}\right)^{d/2}$.

Hence, as the number of variables d increases, the unit ball in variational norm contains a ball in the mixed Sobolev seminorm of an exponentially increasing radius.

Neural networks with n Heaviside perceptrons and a single linear output unit compute functions from the set $span_n H_d(\Omega)$. The corollary shows that functions on $\Re^d$, d odd, with exponentially large Sobolev seminorm can be approximated with rates less than or equal to $\frac{1}{\sqrt{n}}$ by neural nets.

5 Variation with respect to half spaces and total variation

Let $-\infty < a < b < \infty$. 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 $\sum_{j=1}^{k} |h(a_j) - h(a_{j+1}|$ (e.g., Natanson [26, p. 215]). One says 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 hence absolutely continuous [26, pp. 216, 244, 259].

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 of bounded variation when its total variation on the line is finite (Hewitt in [26, p. 238]).

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

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

For a function ψ on $\Re^d$, the *Radon transform* $\mathcal{R}(\psi)$ [12] is the function on $S^{d-1} \times \Re$ defined by

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

where $H_{e,b}$ is the hyperplane $\{y \in \Re^d : e \cdot y + b = 0\}$. Clearly, $\phi_{f,e}(b) = \mathcal{R}(D_e^{d-1} f)(e,b)$.

For almost all $e \in S^{d-1}$ the integral $J_{f,e} = \int_{\Re} |\phi'_{f,e}(b)|db$ is finite by Theorem 4.2, where "almost all" means "except on a set of measure zero".

Proposition 5.1 *If $d \geq 0$, e in S^{d-1} and f is polynomially vanishing on $\Re^d$, and $J_{f,e} < \infty$, then $T(\phi_{f,e}) = J_{f,e}$.*

Proof. As $\phi_{f,e}$ is continuously differentiable on $\Re$, on any finite interval $[a,b]$, $T_{[a,b]}(\phi_{f,e}) = \int_a^b |\phi'_{f,e}(t)|dt$. Take the sup over all $[a,b]$ of both sides, using the assumption that $J_{f,e}$ is finite. □

It is convenient to think of $T(\phi_{f,e})$ as the variation of f in direction e. The following corollary of Theorem 3.3 estimates the variation with respect to half-spaces of a function f as a multiple of its average directional variations.

Corollary 5.2 *Let d be odd. If f is polynomially vanishing on $\Re^d$, then*

$$\|f\|_{H_d(\Re^d),\sup} \leq (1/2)(2\pi)^{1-d} \int_{S^{d-1}} T(\phi_{f,e}) de.$$

Thus, variation with respect to half-spaces of suitable functions f is bounded above by a multiple of the spherical average of the total variation of the Radon transform of the iterated directional derivatives.

6 Variation with respect to half spaces of the Gaussian

We now consider the Gaussian function γ_d on $\Re^d$ defined as $\gamma_d(x) = \exp(-\|x\|^2))$; we write γ for γ_1. The next theorem shows that for d odd, the variation with respect to half-spaces of γ_d grows at most linearly with d.

To prove the theorem, we take advantage of a result by Sonin, as extended by Polya (see [31, p. 166] and [4]) on decrease of local maxima. When a function $y = y(x)$ satisfies the second-order differential equation

$$(k(x)y(x)')' + \phi(x)y(x) = 0,$$

with both k and ϕ positive and continuously differentiable on an open interval (a,b) and $k\phi$ increasing, then the successive relative maxima of $|y|$ form a decreasing sequence. Indeed, the envelope of y,

$u = y^2 + \frac{(ky')^2}{k\phi}$, is decreasing as its derivative $u' = -(y'/\phi)^2(k\phi)'$ is negative. As the relative maxima of $|y|$ are zeroes of y', the values of $|y|$ at its maxima are the square roots of u, and so they also form a decreasing sequence. In the next proof we apply Sonin's result to $d-1$-st derivative of the Gaussian.

Theorem 6.1 *Let $d > 0$ be an odd integer. Then $\|\gamma_d\|_{H_d(\Re),\sup} \leq 2d$.*

Proof. For d odd, using Theorem 5.2 and the fact that γ depends only on $\|x\|$, $\|\gamma_d\|_{\mathcal{L}_1(\Re^d)} = \|\gamma\|^d_{\mathcal{L}_1(\Re)} = \pi^{d/2}$, letting e_1 denote the standard unit vector along the x_1-axis, one has

$$\|\gamma_d\|_{H_d(\Re^d)} \leq (1/2)(2\pi)^{1-d} \int_{S^{d-1}} T(\phi_{\gamma_d,e})de = |a_d|\omega_d \int_{\Re} \left| \int_{H_{e_1,b}} (D^d_{e_1}\ \gamma_d)(y) d_H y \right| db$$

$$= |a_d|\omega_d \int_{\Re} |\gamma^{(d)}(b)|db \int_{\Re^{d-1}} \gamma_{d-1}(y)dy = l_d T(\gamma^{(d-1)}),$$

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

Thus it remains to estimate the total variation of the $d-1$-st derivative of the one-dimensional Gaussian for d odd.

Observe that for any differentiable function h of bounded variation on $\Re$ which is asymptotically zero at both ∞ and $-\infty$, if h has a finite number m of local extrema and s is the maximum of the absolute values of the extrema, then the total variation of h cannot exceed $2sm$.

We first prove that $\gamma^{(d-1)}$ has at most d extrema and for d odd it achieves a maximum or a minimum at zero. This follows from expression of derivatives of the Gaussian in terms of Hermite polynomials: the r-th derivative of the Gaussian is, up to a sign, the Hermite polynomial of degree r (denoted here by P_r) multiplied by the Gaussian, i.e.,

$$\gamma^{(r-1)}(t) = (-1)^r P_r(t)\exp(-t^2)$$

[7, pp.91-92]. Since the extrema of the $d-1$-st derivative of the Gaussian are zeros of its d-th derivative, $\gamma^{(d)}$ has at most d extremes. For d odd, the last term in the expression of d-th order Hermite polynomial $P_d(t)$ is

$$(-1)^{(d-1)/2}\frac{d!2t}{((d-1)/2)!}.$$

Thus $P_d(0) = 0$ as well as $\gamma^{(d)}(0) = 0$. The absolute value of $\gamma^{(d-1)}$ at zero is equal to the absolute value of the last term in the expression of P_{d-1}, which is

$$\frac{(d-1)!}{((d-1)/2)!}$$

as $d-1$ is even.

Using the above mentioned result by Sonin, we prove that $|\gamma^{(d-1)}(0)|$ is the largest value among the extremes of $|\gamma^{(d-1)}|$. Multiplying by $\exp(-t^2)$ all three terms of the recursion for the Hermite polynomials (e.g., [7, p.92]) $P_{d+1} - 2xP_d + 2dP_{d-1} = 0$, we get for $y = \gamma^{(d-1)}$.

$$y'' + 2xy' + (2r+2)y = 0.$$

Setting $k(t) = \exp(t^2)$ and $\phi(t) = 2dk(t)$, we get

$$(k(x)y(x)')' + \phi(x)y(x) = 0$$

and so the maxima of $|\gamma^{(d-1)}|$ are decreasing. Thus the total variation of $\gamma^{(d-1)}$ is at most $2d|\gamma^{(d-1)}(0)| = 2d\frac{(d-1)!}{((d-1)/2)!}$.

Hence $\|\gamma_d\|_{H_d(\Re^d),\sup} \leq |a_d|\omega_d 2d\frac{(d-1)!}{((d-1)/2)!}$.

7 Discussion

Our upper bound on the Gaussian's variation with respect to half-spaces can be improved by a factor not exceeding d. Theorem 6.1 actually shows that $\|\gamma_d\|_{H_d(\Re),\sup} \leq l_d T(\gamma^{(d-1)}) \leq 2d$. But the upper bound on total variation is only an equality if all local maxima are equal (all peaks of equal height), while the local maxima of $|\gamma^{(d-1)}|$ decay rapidly away from zero. For instance, for $d = 3$, the upper bound on $T(\gamma^{(2)})$ is 12 while a direct calculation shows that $T(\gamma^{(2)}) \approx 7.9$.

In the proof of Theorem 4.2, $\omega_d d^{d/2}$ is not necessarily a good upper bound on $\int_{S^{d-1}} (\sum_{i=1}^{d} e_i)^d de$. Though for $d = 1$, both sides are equal (to 2), for $d \geq 2$ the upper bound is strictly bigger. An explicit calculation gives $2\pi + 4$ for the integral and 4π for the upper bound. Using computer algebra (Pari-GP) to calculate the ratios for $d = 3, 5, 7, 9$, one obtains $\cong .67, .45, .30, .20$, respectively.

The Sobolev seminorm used in our estimates is in general very much smaller than the Sobolev norm. The set S of multi-indices of length d and degree d has cardinality $C(2d-1, d-1)$ [9, p. 38], where $C(a, b)$ denotes the binomial coefficient. By Stirling's formula, S has about 2^{2d} elements and the seminorm takes its maximum over S. But there are in fact exactly twice as many multi-indices over which we must *sum* for the Sobolev norm (see, e.g., [29, pp. 42–44]). The interesting sequence of numbers which arises $1, 3, 10, 35, 126, \ldots$ can be found in [30].

Finally, it is tempting to conjecture that Theorem 5.2 holds under the weaker hypothesis that for almost all $e \in S^{d-1}$, the function $\phi_{f,e}$ is of bounded variation.

Acknowledgement

The authors thank Michael Somos for a remark simplifying the proof of Theorem 4.2 and other helpful comments. He also provided the Pari calculations.

Bibliography

[1] Adams, R. A., & Fournier, J. J. F. (2003). *Sobolev Spaces*. Second edition. Amsterdam: Academic Press.

[2] Barron, A. R. (1992). Neural net approximation. In *Proceedings of the 7th Yale Workshop on Adaptive and Learning Systems* (pp. 69–72).

[3] Barron, A. R. (1993). Universal approximation bounds for superposition of a sigmoidal function. *IEEE Transactions on Information Theory*, **39**, 930–945.

[4] Butlewski, Z. (1936). Sur les integrales d'une équation differentielle du second ordre, *Mathematica* (Cluj) **12**, 36-48.

[5] Cheang, G. H. L. and Barron, A. R. (2000). A better approximation for balls. *Journal of Approximation Theory*, **104**, 183–203.

[6] Courant, R. (1960). *Calculus*, vol. 2. New York: Wiley.

[7] Courant, R., Hilbert, D. (1989). Methods of Mathematical Physics. Vol.I. New York: John Wiley & Sons.

[8] Edwards, C. H. (1994). *Advanced Calculus of Several Variables*. New York: Dover.

[9] Feller, W. F. (1968). *An introduction to probability theory and its applications, vol I.* New York: Wiley.

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

[11] Hamming, R. W. (1986). *Coding and Information Theory*. Prentice-Hall:Englewood Cliffs, NJ.

[12] Helgason, S. (1980). *The Radon Transform* Boston: Birkhäuser.

[13] Jones, L. K. (1992). A simple lemma on greedy approximation in Hilbert space and convergence rates for projection pursuit regression and neural network training. *Ann. of Statistics*, **20**:608-613.

[14] Kainen, P. C., Kůrková, V. & Sanguineti, M. (2003). Minimization of error functionals over variable-basis functions. *SIAM J. of Optimization*, **14**, 732–742.

[15] Kainen, P. C., Kůrková, V. & Vogt, A. (2000). Geometry and topology of continuous best and near best approximations. *Journal of Approximation Theory*, **105**, 252–262.

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

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

[18] Kainen, P. C., Kůrková, V. & Vogt, A. (2004). Integral combinations of Heavisides, manuscript.

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

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

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

[22] Kůrková, V. & Sanguineti, M. (2001). Bounds on rates of variable-basis and neural network approximation. *IEEE Transactions on Information Theory*, **47**, 2659–2665.

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

[24] Kůrková, V., Savický, P. & Hlaváčková, K. (1998). Representations and rates of approximation of real-valued Boolean functions by neural networks. *Neural Networks*, **11**, 651–659.

[25] Levitan, D. & Temlyakov, V. N., Simultaneous approximation by greedy algorithms, www.math.tau.ac.il/ leviatan/greedy.pdf

[26] Natanson, I. P. (1961). *Theory of functions of a real variable*. Vol. I. (Transl. L. F. Boron with editorial annotations by E. Hewitt). Ungar: New York.

[27] Pinkus, A. (1986). *n-Width in Approximation Theory*. Berlin: Springer.

[28] Pisier, G. (1980). Remarques sur un résultat non publié de B. Maurey. In *Séminaire d'Analyse Fonctionnelle 1980-81, Exposé* no. V, pp. V.1-V.12, École Polytechnique, Centre de Mathématiques, Palaiseau, France.

[29] Reimer, M. (2003). Multivariate polynomial approximation, Birkhauser, Basel.

[30] Sloane, N., Ed.(2004). Encyclopedia of Integer Sequences, sequence 1700, addendum by P. C. Kainen and M. Somos. http://www.research.att.com/ njas/sequences/

[31] Szegő, G. (1975). *Orthogonal Polynomials*. Providence: American Mathematical Society Colloquium Series XXIII.