Neural Networks 10: 1061--1068, 1997

Estimates of the Number of Hidden Units and Variation with Respect to Half-spaces

Věra Kůrková
Institute of Computer Science, Czech Academy of Sciences, Prague, Czechia

Paul C. Kainen
Industrial Math, Washington, D.C., USA

Vladik Kreinovich
Department of Computer Science, University of Texas at El Paso
El Paso, Texas, USA

Acknowledgements: V. Kůrková was partially supported by GACR grants 201/93/0427 and 201/96/0917, P. C. Kainen acknowledges research support from the Department of Mathematics of Georgetown University, and V. Kreinovich was partially supported by NSF grant CDA-9015006, NASA research grant NAG 9-757 and NSF grant EEC-9322370. The first two authors wish to thank A. Vogt for his very helpful comments.

Requests for reprints should be sent to:
Věra Kůrková, Institute of Computer Science
Pod vodárenskou věží 2, 182 07 Prague 8, Czechia
tel. +420 266053231, fax +420 286585789, e-mail vera@uivt.cas.cz

Running title: Estimates of the number of hidden units

Estimates of the Number of Hidden Units and Variation with Respect to Half-spaces

Abstract

We estimate variation with respect to half-spaces in terms of "flows through hyperplanes". Our estimate is derived from an integral representation for smooth compactly supported multivariable functions proved using properties of the Heaviside and delta distributions. Consequently we obtain conditions which guarantee approximation error rate of order $\mathcal{O}(\frac{1}{\sqrt{n}})$ by one-hidden-layer networks with n sigmoidal perceptrons.

Keywords. approximation of functions, one-hidden-layer sigmoidal networks, estimates of the number of hidden units, variation with respect to half-spaces, integral representation

1 Introduction

Approximating functions from $\mathcal{R}^d$ to $\mathcal{R}^m$ by feedforward neural networks has been widely studied in recent years, and the existence of an arbitrarily close approximation, for any continuous or $\mathcal{L}_p$ function defined on a compact set, has been proven for one-hidden-layer networks with perceptron or radial-basis-function units with quite general activation functions (see e.g., Mhaskar & Micchelli, 1992, Park & Sandberg, 1993).

However, estimates of the number of hidden units that guarantee a given accuracy of approximation are less understood. Most upper estimates grow exponentially with the number of input units, i.e. with the number d of variables of the function f to be approximated (e.g., Mhaskar & Micchelli, 1992, Kůrková, 1992). A general result by deVore et al. (1989) confirms that there is no hope for a better estimate when the class of multivariable functions being approximated is defined in terms of the bounds of partial derivatives and parameters of approximating networks are chosen continuously. But in applications, functions of hundreds of variables are approximated sufficiently well by neural networks with only moderately many hidden units (e.g., Sejnowski & Yuhas, 1989).

Jones (1992) introduced a recursive construction of approximants with "dimension-independent" rates of convergence to elements in convex closures of bounded subsets of a Hilbert space and proposed to apply it to the space of functions achievable by a one-hidden-layer neural network (in fact, the idea of applying Jones' result to neural networks seems to have been a joint effort of Jones and Barron as both authors acknowledged in their papers – see Jones (1992) and Barron (1993)). Applying Jones' estimate Barron (1993) showed that it is possible to approximate any function satisfying a certain condition on its Fourier transform within $\mathcal{L}_2$ error of $\mathcal{O}(\frac{1}{\sqrt{n}})$ by a network whose hidden layer contains n perceptrons with a sigmoidal activation function.

Using a probabilistic argument Barron (1992) extended Jones' estimate also to supremum norm. His estimate holds for functions in the convex uniform closure of the set of characteristic functions of half-spaces multiplied by a real number less than or equal to B. He called the infimum of such B the *variation with respect to half-spaces* and noted that it could be defined for any class of characteristic functions.

In this paper, we prove two main results which are complementary. The first one (Theorem 3.5) bounds variation with respect to half-spaces for functions that can be represented by an integral equation corresponding metaphorically to a neural network with a weighted continuum of Heaviside perceptrons. The bound on

variation is the $\mathcal{L}_1$-norm of the weighting function. The second one (Theorem 4.1) provides, for compactly supported functions on $\mathcal{R}^d$, for d odd, with continuous d-th order partial derivatives, such a representation with weights corresponding to flows orthogonal to hyperplanes determined by the input weights and biases.

For these functions we derive an upper bound on their variation with respect to half spaces over a compact subset J of $\mathcal{R}^d$ equal to $\frac{1}{2}(2\pi)^{1-d}$ times the integral over the cylinder $S^{d-1} \times \mathcal{R}$ of the absolute value of the integral of the d-th directional derivative of f over the intersection of J with the cozero hyperplane determined by a point in the cylinder $S^{d-1} \times \mathcal{R}$ (corresponding to the affine functions determined by perceptron parameters: weight vector and bias).

Estimating these integrals we show that the variation with respect to half-spaces of a function f over a compact subset J of $\mathcal{R}^d$ is bounded above by the supremum of absolute values of integrals of directional derivatives of order d of f over orthogonal hyperplanes multiplied by a d-dimensional volume. For single variable functions our bound is identical to a well-known formula for total variation, which in the 1-dimensional case is the same as variation with respect to half-spaces.

Consequently, for d odd and f a compactly supported, real-valued function on $\mathcal{R}^d$ with continuous partial derivatives of order d, we can guarantee approximations for $\mathcal{L}_2$-norm with error rate at most $\mathcal{O}(\frac{1}{\sqrt{n}})$ by one-hidden-layer networks with n sigmoidal perceptrons for any bounded sigmoidal activation function.

Our proof is based on properties of the Heaviside and delta distributions. We use a representation of the d-dimensional delta distribution as an integral over the unit sphere S^{d-1} in $\mathcal{R}^d$ that is valid only for d odd. To obtain a representation for all positive integers d, one could extend functions f defined on $\mathcal{R}^d$ to $\mathcal{R}^{d+1}$ by composition with a projection from $\mathcal{R}^{d+1}$ to $\mathcal{R}^d$.

The remainder of the paper is organized as follows: Section 2 investigates functions in the convex closures of parameterized families of continuous functions and integral representations. Section 3 considers variation with respect to half-spaces, while section 4 gives an integral representation theorem and its consequence for a bound on variation. Section 5 is about rates of approximation and dimension independence. Section 6 is a brief discussion, while the proofs are given in section 7.

2 Approximation of functions in convex closures

Let $\mathcal{R}$, $\mathcal{N}$ denote the set of real and natural numbers, respectively.

Recall that a *convex combination* of elements $s_1, \ldots, s_m$ $(m \in \mathcal{N})$ in a linear space is a sum of the form $\sum_{i=1}^m a_i s_i$, where the a_i are all non-negative and $\sum_{i=1}^m a_i = 1$. A subset of a vector space is *convex* if it contains every convex combination of its elements; we denote the set of all convex combinations of elements of X by *conv* X, which is clearly a convex set, and call it the *convex hull* of X.

When we require measurability, it will be with respect to Lebesgue measure in some subset of $\mathcal{R}^k$. By $\lambda_k(A)$ is to denoted the *k-dimensional Lebesgue measure* of a set $A \subset \mathcal{R}^k$.

For a topological space X $\mathcal{C}(X)$ denotes the *set of all continuous real-valued functions on* X and $\|.\|_{\mathcal{C}}$ denotes the *supremum norm*. For a subset X of $\mathcal{R}^d$ and a positive integer d $\mathcal{C}^k(X)$ denotes the *set of all real-valued functions on X with continuous iterated partial derivatives of order k*; $\mathcal{C}^\infty(X)$ the *set of all functions with continuous partial derivatives of all orders*. For $p \in [1, \infty)$ and a subset X of $\mathcal{R}^d$ $\mathcal{L}_p(X)$ denotes the space of $\mathcal{L}_p$ functions and $\|.\|_p$ denote the *$\mathcal{L}_p$-norm*.

For any topological space X with a topology τ, we write $cl_\tau A$ for the *closure* of a subset A of X (smallest closed subset containing A). So $cl_{\mathcal{C}}$ denotes the closure in the topology of uniform convergence and $cl_{\mathcal{L}_p}$ the closure with respect to $\mathcal{L}_p$-topology. Closure of the convex hull is called the *convex closure*. For a function $f : X \to \mathcal{R}$

the *support* of f denoted by $supp(f)$ is defined by $supp(f) = cl_\tau\{x \in X; f(x) \neq 0\}$. For $f : X \to \mathcal{R}$ and $A \subset X$, $f|_A$ denotes the restriction of f to A; when it is clear from context, we omit the subscript.

Jones (1992) estimated rates of approximation of functions from convex closures of bounded subsets of a Hilbert space; see also Barron (1993), p.934.

Theorem 2.1 *Let $\mathcal{H}$ be a Hilbert space with a norm $\|.\|$, B be a positive real number and $\mathcal{G}$ a subset of $\mathcal{H}$ such that for every $g \in \mathcal{G}$ $\|g\| \leq B$. Then for every $f \in$ cl conv $\mathcal{G}$, for every $c > B^2 - \|f\|^2$ and for every natural number n there exists f_n that is a convex combination of n elements of $\mathcal{G}$ such that*

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

To use this theorem to estimate the number of hidden units in neural networks, one takes $\mathcal{G}$ to be the set of functions computable by single-hidden-unit networks for various types of computational units. Convex combinations of n such functions can be computed by a network with n hidden units and one linear output unit.

Several authors have derived characterizations of such sets of functions from integral representations (e.g., Barron, 1993, used Fourier representation, Girosi and Anzellotti, 1993, convolutions with signed measures). Here we formulate a general characterization of this type for parameterized families of functions.

For X, Y topological spaces, a function $\phi : X \times Y \to \mathcal{R}$, a positive real number B and a subset $J \subseteq X$ define $\mathcal{G}(\phi, B, J) = \{f : J \to \mathcal{R}; f(x) = w\phi(x, y); w \in \mathcal{R}, |w| \leq B, y \in Y\}$. So $\mathcal{G}(\phi, B, J)$ consists of a family of real-valued functions on J parameterized by $y \in Y$ and then scaled by a constant at most B in absolute value.

Theorem 2.2 *Let d be any positive integer, J be a compact subset of $\mathcal{R}^d$ and let $f \in \mathcal{C}(J)$ be any function that can be represented as $f(\mathbf{x}) = \int_Y w(\mathbf{y})\phi(\mathbf{x}, \mathbf{y})d\mathbf{y}$, where $Y \subseteq \mathcal{R}^k$ for some positive integer k, $w \in \mathcal{C}(Y)$ is compactly supported and $\phi \in \mathcal{C}(\mathcal{R}^d \times Y)$. Then $f \in cl_{\mathcal{C}}conv\ \mathcal{G}(\phi, B, J)$, with $B = \int_{J_\phi} |w(\mathbf{y})|d\mathbf{y}$, where $J_\phi = \{\mathbf{y} \in Y; (\exists \mathbf{x} \in J)(w(\mathbf{y})\phi(\mathbf{x}, \mathbf{y}) \neq 0)\}$.*

Notice that *cl* J_ϕ is compact and when ϕ is never 0, then $J_\phi = supp(w)$.

To apply this theorem to perceptron type networks with an activation function $\psi : \mathcal{R} \to \mathcal{R}$ put $Y = \mathcal{R}^d \times \mathcal{R}$ and define $P_\psi(\mathbf{x}, \mathbf{v}, b) = \psi(\mathbf{v} \cdot \mathbf{x} + b)$. So $\mathcal{G}(P_\psi, B, J)$ denotes the set of functions computable by a network with d inputs, one hidden perceptron with an activation function ψ and one linear output unit with weight bounded by B in absolute value. Typically, ψ is *sigmoidal*, i.e. it satisfies $\lim_{t\to\infty} \psi(t) = 1$ and $\lim_{t\to-\infty} \psi(t) = 0$. Note that many authors add to definition of a sigmoidal function an additional assumption that the function is monotonically increasing. However, for our results this weaker definition is sufficient.

Corollary 2.3 *Let $\psi : \mathcal{R} \to \mathcal{R}$ be a continuous activation function, $\phi = P_\psi$, d be any positive integer, J be a compact subset of $\mathcal{R}^d$ and $f \in \mathcal{C}(J)$ be any function that can be represented as $f(\mathbf{x}) = \int_K w(\mathbf{v}, b)\psi(\mathbf{v} \cdot \mathbf{x} + b)d(\mathbf{v}, b)$, where $K \subseteq \mathcal{R}^{d+1}$, $w \in \mathcal{C}(\mathcal{R}^d \times \mathcal{R})$ is compactly supported. Then $f \in cl_{\mathcal{C}}conv\ \mathcal{G}(P_\psi, B, J)$, with $B = \int_{J_\phi} |w(\mathbf{v}, b)|d(\mathbf{v}, b)$, where $J_\phi = \{(\mathbf{v}, b) \in \mathcal{R}^d \times \mathcal{R}; (\exists \mathbf{x} \in J)(w(\mathbf{v}, b)\psi(\mathbf{v}\cdot\mathbf{x}+b) \neq 0)\}$.*

So for functions computable by perceptron networks with a "continuum" of hidden units, we can find a suitable bound B for Jones' theorem by taking $B = \int_{J_\phi} |w(\mathbf{v}, b)|d\mathbf{v}db$.

3 Variation with respect to half-spaces

Let ϑ denote the Heaviside function ($\vartheta(x) = 0$ for $x < 0$ and $\vartheta(x) = 1$ for $x \geq 0$). It is easy to see that for J a proper subset of $\mathcal{R}^d$ $\mathcal{G}(P_\vartheta, B, J) = \{g : J \to \mathcal{R}; g(\mathbf{x}) = w\vartheta(\mathbf{e} \cdot \mathbf{x} + b), \mathbf{e} \in S^{d-1}, w, b \in \mathcal{R}, |w| \leq B\}$, where S^{d-1} denotes the unit sphere in $\mathcal{R}^d$ (for $J = \mathcal{R}^d$ these two sets only differ by constant functions of norm at most B).

Let $J \subset \mathcal{R}^d$ and let $\mathcal{F}(J)$ be a linear space of functions from J to $\mathcal{R}$ and τ be a topology on $\mathcal{F}(J)$. For $f \in \mathcal{F}(J)$ put

$$V(f, \tau, J) = \inf\{B \in \mathcal{R}; f \in cl_\tau conv\ \mathcal{G}(P_\vartheta, B, J)\}$$

and call $V(f, \tau, J)$ the *variation of f on J with respect to half-spaces and topology τ*. For $f : \mathcal{R}^d \to \mathcal{R}$, if $f|_J \in \mathcal{F}(J)$, then we write $V(f, \tau, J)$ instead of $V(f|_J, \tau, J)$. The following proposition shows that when the topology τ is induced by a norm, this infimum is achieved.

Proposition 3.1 *Let d be a positive integer, $J \subset \mathcal{R}^d$, $\mathcal{F}(J)$ be a linear space of functions on J with a topology τ induced by a norm $\|.\|$. Then for every $f \in \mathcal{F}(J)$ $f \in cl_\tau\ conv\ \mathcal{G}(P_\vartheta, V(f, \tau, J), J)$.*

When J is fixed, variation with respect to half-spaces is a norm on function spaces as described in the next proposition.

Proposition 3.2 *For every positive integer d, for every $J \subset \mathcal{R}^d$ and for every topology τ induced by a norm $\|.\|$ on a linear space $\mathcal{F}(J)$ of functions from J to $\mathcal{R}$*
(i) the set of functions $\mathcal{B}(J) = \{f \in \mathcal{F}(J); V(f, \tau, J) < \infty\}$ is a linear subspace of $\mathcal{F}(J)$,
(ii) $V(., \tau, J)$ is a norm on its factor space $\mathcal{B}(J)/\sim$, where the equivalence $\sim$ is defined by $f \sim g$ when $\|f - g\| = 0$,
(iii) for every $f \in \mathcal{F}(J)$ $\|f\| \leq V(f, \tau, J) \sup\{\|\vartheta(\mathbf{e} \cdot \mathbf{x} + b)\|; \mathbf{e} \in S^{d-1}, b \in \mathcal{R}\}$.

So we have an elementary lower bound on the variation with respect to half-spaces. In particular, for any compact $J \subset \mathcal{R}^d$ and $f \in \mathcal{C}(\mathcal{R}^d)$ $\|f\|_\mathcal{C} \leq V(f, \mathcal{C}, J)$, while for $f \in \mathcal{L}_p(J)$ for some $p \in [1, \infty)$ $\lambda_d(J)^{-1/p}\|f\|_p \leq V(f, \mathcal{L}_p, J)$, where λ_d denotes the Lebesgue measure on $\mathcal{R}^d$.

Let $p \in [1, \infty]$. For every $X \subseteq \mathcal{L}_p(J)$ we have $cl_\mathcal{C} X \subseteq cl_{\mathcal{L}_p} X$; hence, $V(f, \mathcal{L}_p, J) \leq V(f, \mathcal{C}, J)$. In addition to changing the topology, one could also change the function ψ generating P_ψ. The first one of the following two estimates is obtained using an approximation of the Heaviside function ϑ in the $\mathcal{L}_1$-norm by a sequence of sigmoidal functions with increasing steepness. The second one follows from a possibility to approximate uniformly any continuous non-decreasing sigmoid by a "staircase" function.

Proposition 3.3 *Let $\sigma : \mathcal{R} \to \mathcal{R}$ be a bounded sigmoidal function. Then for every positive integer d, for every compact $J \subset \mathcal{R}^d$ and for every $p \in [1, \infty)$*

$$cl_{\mathcal{L}_p} conv\ \mathcal{G}(P_\vartheta, B, J) \subseteq cl_{\mathcal{L}_p} conv\ \mathcal{G}(P_\sigma, B, J).$$

Proposition 3.4 *Let $\sigma : \mathcal{R} \to \mathcal{R}$ be a non-decreasing continuous sigmoidal function. Then for every positive integer d, for every compact $J \subset \mathcal{R}^d$ and for every $B \geq 0$,*

$$cl_\mathcal{C} conv\ \mathcal{G}(P_\sigma, B, J) \subseteq cl_\mathcal{C} conv\ \mathcal{G}(P_\vartheta, B, J)$$

and for every $p \in [1, \infty)$,

$$cl_{\mathcal{L}_p} conv\ \mathcal{G}(P_\sigma, B, J) \subseteq cl_{\mathcal{L}_p} conv\ \mathcal{G}(P_\vartheta, B, J).$$

So in $\mathcal{L}_p$-topology variation with respect to half-spaces is unchanged if the Heaviside function is replaced by any continuous non-decreasing sigmoidal function.

Recall that for a function $f : \mathcal{R} \to \mathcal{R}$ and an interval $[s,t] \subset \mathcal{R}$ *total variation* of f on $[s,t]$ denoted by $T(f,[s,t])$ is defined to be $\sup\{\sum_{i=1}^{k} |f(t_{i+1}) - f(t_i)|; s = t_1 < \ldots < t_k = t, k \in \mathcal{N}\}$ (see e.g., McShane, 1944). For functions of one variable satisfying $f(s) = 0$, the concept of total variation on $[s,t]$ coincides with the concept of variation with respect to half-spaces (half-lines) and the topology of uniform convergence, since $T(f,[s,t]) = V(f,\mathcal{C},[s,t])$ (this follows from Lemma 7.1 or see Barron, 1992, also Darken et al., 1993, Theorem 6). In contrast to variation with respect to half-spaces total variation is only a semi-norm (see Hewitt & Stromberg, 1965, p.271).

When generalizing to functions of several variables, there is no unique way to extend the notion of total variation since we lose the linear ordering property. One well-known method divides d-dimensional cubes into boxes with faces parallel to the coordinate hyperplanes. One defines
$T(f,J) = \sup_{\mathcal{J}} \sum_{i=1}^{k} |f(J_i)|$, where $\mathcal{J}$ is the set of all subdivisons $\{J_i; i = 1,\ldots,k\}$ of J into boxes and $f(J_i) = \sum_{j=1}^{2^d} (-1)^{\nu(j)} f(\mathbf{x}_{ij})$, where $\{\mathbf{x}_{ij}; j = 1,\ldots,2^d\}$ are the corner points of J_i and $\nu(j) = \pm 1$ is a parity (see McShane, 1944). For $d \geq 2$ this concept is different from Barron's variation with respect to half-spaces. For example, the characteristic function χ of the set $\{(x_1,x_2) \in [0,1]^2; x_1 \geq x_2\}$ has the variation w.r.t. half-spaces and any topology equal to 1, while the total variation $T(\chi,[0,1]^2)$ is infinite.

For a differentiable function, total variation can be characterized as an integral of the absolute value of its derivative. Formally, if $J \subset \mathcal{R}$ is an interval and $f' \in \mathcal{L}_1(J)$ then $T(f,J) = \int_J |f'(x)|dx$ (see McShane, 1944, p.242). The following result extends this bound to variation with respect to half-spaces.

Theorem 3.5 *Let d be a positive integer, $K \subseteq S^{d-1} \times \mathcal{R}$, J be a compact subset of $\mathcal{R}^d$ and $f \in \mathcal{C}(J)$ be any function which can be represented as $f(\mathbf{x}) = \int_K w(\mathbf{e},b)\vartheta(\mathbf{e}\cdot\mathbf{x}+b)d(\mathbf{e},b)$, where $w \in \mathcal{C}(S^{d-1} \times \mathcal{R})$ is compactly supported and supp$(w) \subseteq K$. Then $f \in cl_{\mathcal{C}} conv\ \mathcal{G}(P_\vartheta,B,J)$, where $B = \int_K |w(\mathbf{e},b)|d(\mathbf{e},b)$; that is, $V(f,\mathcal{C},J) \leq \int_K |w(\mathbf{e},b)|d(b,\mathbf{e})$.*

4 Integral representation theorem

To estimate variation with respect to half-spaces using Corollary 3.5 we need an integral representation theorem of the form of a neural network with continuum of Heaviside perceptrons $\{\vartheta(\mathbf{e}\cdot\mathbf{x}+b); \mathbf{e} \in S^{d-1}, b \in \mathcal{R}\}$. The following theorem provides such a representation with output weights $w(\mathbf{e},b)$ corresponding to orthogonal "flows of order d" of f through cozero hyperplanes $H_{\mathbf{e}b} = \{\mathbf{y} \in \mathcal{R}^d; \mathbf{e}\cdot\mathbf{y}+b = 0\}$.

Recall (see e.g., Rudin, 1964) that for $\mathbf{e}$ a unit vector in $\mathcal{R}^d$ and f a real-valued function defined on $\mathcal{R}^d$, the *directional derivative* of f in direction $\mathbf{e}$ is defined by $D_{\mathbf{e}}f(\mathbf{y}) = \lim_{t\to 0} \frac{f(\mathbf{y}+t\mathbf{e})-f(\mathbf{y})}{t}$ and the *k-th directional derivative* is inductively defined by $D_{\mathbf{e}}^{(k)}f(\mathbf{y}) = D_{\mathbf{e}}(D_{\mathbf{e}}^{(k-1)}f)(\mathbf{y})$. It is well-known (see e.g., Rudin, 1964, p.222) that $D_{\mathbf{e}}f(\mathbf{y}) = \bigtriangledown f(\mathbf{y})\cdot\mathbf{e}$. More generally, the k-th order directional derivative is a weighted sum of the corresponding k-th order partial derivatives, where the weights are polynomials in the coordinates of $\mathbf{e}$ multiplied by multinomials (see e.g., Edwards, 1994, p.130). Hence existence and continuity of partial derivatives implies existence and continuity of directional derivatives.

Theorem 4.1 *For every odd positive integer d every compactly supported function*

$f \in \mathcal{C}^d(\mathcal{R}^d)$ can be represented as

$$f(\mathbf{x}) = -a_d \int_{S^{d-1}} \int_{\mathcal{R}} \left(\int_{H_{\mathbf{e}b}} D_{\mathbf{e}}^{(d)} f(\mathbf{y}) d\mathbf{y} \right) \vartheta(\mathbf{e} \cdot \mathbf{x} + b) db d\mathbf{e},$$

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

Our proof of Theorem 4.1 makes use of the theory of distributions. For a positive integer k, denote by δ_k the *delta distribution* operating by convolution as the identity on the linear space $\mathcal{D}(\mathcal{R}^k)$ of all *test functions* (i. e. the subspace of $\mathcal{C}^\infty(\mathcal{R}^k)$ containing compactly supported functions). For d odd, one can represent the delta distribution δ_d as an integral over the unit sphere $\delta_d(\mathbf{x}) = a_d \int_{S^{d-1}} {\delta_1}^{(d-1)}(\mathbf{e} \cdot \mathbf{x}) d\mathbf{e}$ (see Courant & Hilbert, 1992, p.680) (by ${\delta_1}^{(d-1)}$ is denoted the d-1-st distributional derivative of δ_1). We also utilize the fact that δ_1 is the first distributional derivative of ϑ.

Extension to all compactly supported functions with continuous partial derivatives of order d follows from a basic result of distribution theory: each continuous compactly supported function can be uniformly approximated on $\mathcal{R}^d$ by a sequence of test functions (see e.g., Zemanian, 1987, p.3).

Integral representation 4.1 together with Theorem 3.5 gives an estimate of variation with respect to half-spaces toward which we have been aiming.

Theorem 4.2 *If d is an odd positive integer, $J \subset \mathcal{R}^d$ is compact, and $f \in \mathcal{C}^d(\mathcal{R}^d)$ is compactly supported, then*

$$V(f, \mathcal{C}, J) \leq |a_d| \int_{S^{d-1}} \int_{\mathcal{R}} |w_f(\mathbf{e}, b)| db d\mathbf{e},$$

where $w_f(\mathbf{e}, b) = \int_{H_{\mathbf{e}b}} D_{\mathbf{e}}^{(d)} f(\mathbf{y}) d\mathbf{y}$ and $|a_d| = (1/2)(2\pi)^{1-d}$.

It is easy to verify that when $d = 1$ Corollary 4.2 gives the estimate $V(f, \mathcal{C}, J) \leq \int_{\mathcal{R}} |f'(b)| db$ which, for f with $supp(f) = J$, agrees with the above mentioned characterization of total variation for functions of one variable; $a_1 = 1/2$ and the sphere S^0 consists of two points so the constant comes out correctly.

In Corollary 4.2, instead of integrating over $S^{d-1} \times \mathcal{R}$ we can restrict to integration only over $J^* = \{(\mathbf{e}, b) \in S^{d-1} \times \mathcal{R}; (\exists \mathbf{x} \in J)(\int_{H_{\mathbf{e}b}} D_{\mathbf{e}}^{(d)} f(\mathbf{y}) d\mathbf{y}) \vartheta(\mathbf{e} \cdot \mathbf{x} + b) \neq 0\}$. Note that $J^* \subseteq \{(\mathbf{e}, b) \in S^{d-1} \times \mathcal{R}; H_{\mathbf{e}b} \cap supp(f) \neq \emptyset \ \& \ H_{\mathbf{e}b}^+ \cap J \neq \emptyset\}$, where $H_{\mathbf{e}b}^+ = \{\mathbf{x} \in \mathcal{R}^d; \mathbf{e} \cdot \mathbf{x} + b \geq 0\}$. It is easier to compute a slightly larger integral over $J_f = \{(\mathbf{e}, b) \in S^{d-1} \times \mathcal{R}; H_{\mathbf{e}b} \cap supp(f) \neq \emptyset\}$.

Bounding the integral in Corollary 4.2 and applying Theorem 3.5 we get the following corollary.

Corollary 4.3 *If d is an odd positive integer, $J \subset \mathcal{R}^d$ is compact, and $f \in \mathcal{C}^d(\mathcal{R}^d)$ is compactly supported, then*

$$V(f, \mathcal{C}, J) \leq |a_d| \lambda_d(J_f) W(f),$$

where $|a_d| = (1/2)(2\pi)^{1-d}$, $J_f = \{(\mathbf{e}, b) \in S^{d-1} \times \mathcal{R}; H_{\mathbf{e}b} \cap supp(f) \neq \emptyset\}$, $W(f) = \sup\{|w_f(\mathbf{e}, b)|; (\mathbf{e}, b) \in J_f\}$ and $w_f(\mathbf{e}, b) = \int_{H_{\mathbf{e}b}} D_{\mathbf{e}}^{(d)} f(\mathbf{y}) d\mathbf{y}$.

Note that the first factor in this bound, a_d, as a function of d is converging to zero exponentially fast.

The second factor, $\lambda_d(J_f)$, is the measure of the subset of $S^{d-1} \times \mathcal{R}$ of parameters of all hyperplanes intersecting $supp(f)$ non-trivially. This measure can be much smaller than the measure of $supp(f)$ itself. For instance, when $supp(f)$ is the ball

of radius r centered at the origin, then one can easily check that J_f is an annulus consisting of $S^{d-1} \times [-r, +r]$. The volume of the ball $supp(f)$ is a constant times r^d, but the volume of J_f only involves a linear term in r.

The third factor $W(f)$ is bounded above by $X(f)Y(f)$, where $X(f)$ is the largest cross-sectional area,

$$X(f) = \sup\{\lambda_{d-1}(supp(f) \cap H_{\mathbf{e}b}); (\mathbf{e}, b) \in J_f\};$$

and $Y(f)$ is the supremum of the absolute value of the average flow per unit area across one of the hyperplane sections of $supp(f)$,

$$Y(f) = \sup\left\{\frac{|w_f(\mathbf{e}, b)|}{\lambda_{d-1}(H_{\mathbf{e}b} \cap supp(f))}\right\}; (\mathbf{e}, b) \in J_f\}.$$

Getting a bound on $X(f)$ is a geometric problem; results are known for two of the simplest choices for $supp(f)$: d-dimensional ball and cube.

The largest cross-section of any d-dimensional ball by a hyperplane is a $d-1$-dimensional ball. Hence for f with $supp(f) = \{\mathbf{x} \in \mathcal{R}^d; \|\mathbf{x}\| \leq r\}$ we have $X(f) = \lambda_{d-1}(\{\mathbf{x} \in \mathcal{R}^d; \|\mathbf{x}\| \leq r, x_1 = 0\})$. The volume of a d-dimensional ball of radius r is $r^d \pi^{d/2} (\Gamma(d - 2/2))^{-1}$, where Γ is the Gamma function – so its volume as a function of the dimension d converges to zero unless the radius is a function of d that is growing fast enough (see Hamming, 1986, p.78).

For d-dimensional cubes, the maximal cross-section problem seems to be non-trivial. Ball (1986) has shown that the largest cross-sectional area of a hyperplane section of a d-dimensional cube is $\sqrt{2}$. Thus, for f with $supp(f) = [0, 1]^d$, $X(f) = \sqrt{2}$.

A weak upper bound on $Y(f)$ is $\sup\{|D^{(d)}_{\mathbf{e}} f(\mathbf{y})|; \mathbf{y} \in supp(f)\}$ since the average flow cannot exceed the maximum flow. In applications, one might often find that there is a natural bound $F(f)$ on the maximum of orthogonal directional derivative flow of f across any hyperplane. Then $Y(f) \leq F(f)/X(f)$ and hence $W(f) \leq F(f)$.

5 Dimension-independent rates of approximation by neural networks

Since ϑ can be approximated in $\mathcal{L}_p$-norm ($p \in [1, \infty)$) by a sequence of steep sigmoidals, estimates of variation with respect to half-spaces can be used to bound approximation error achievable by one-hidden-layer neural networks with σ perceptrons for any bounded sigmoidal activation function σ.

Let $f \in \mathcal{C}^d(\mathcal{R}^d)$ be a compactly supported function and $J \subset \mathcal{R}^d$ be compact. Denote by B_f the product of the upper bound on $V(f, \mathcal{C}, J)$ given by Corollary 4.2 with $\lambda_d(J)^{\frac{1}{2}}$ (an upper bound on $\mathcal{L}_2$-norms of characteristic functions of all half-spaces of J), i.e. $B_f = |a_d| \lambda_d(J)^{\frac{1}{2}} \int_{J^*} \left| \int_{H_{\mathbf{e}b}} D^{(d)}_{\mathbf{e}} f(\mathbf{y}) d\mathbf{y} \right| d(\mathbf{e}, b)$.

Theorem 2.1 together with Corollary 4.2 and Proposition 3.4 imply the following estimate of rates of approximation by one-hidden-layer networks with sigmoidal perceptrons.

Theorem 5.1 *Let d be an odd positive integer, $f \in \mathcal{C}^d(\mathcal{R}^d)$ be compactly supported and $\sigma : \mathcal{R} \to \mathcal{R}$ be a bounded sigmoidal function. If $c > B_f^2 - \|f\|_2^2$ then for every $n \in \mathcal{N}$ there exists a function f_n computable by a neural network with a linear output unit and n σ-perceptrons in the hidden layer such that $\|f - f_n\|_2 \leq \sqrt{\frac{c}{n}}$.*

6 Discussion

Using the inverse Radon transform, Ito (1991) obtained an integral representation similar to Theorem 4.1. Our proof of Theorem 4.1 uses a different approach and describes the coefficients $w_f(\mathbf{e}, b)$ in terms of the integral over the hyperplane parameterized by $(\mathbf{e}, b)$ of the iterated directional derivative.

Darken et al. (1993) extended Jones' theorem to $\mathcal{L}_p$ spaces for $p \in (1, \infty)$ with a slightly worse rate of approximation – of order only $\mathcal{O}(n^{-\frac{1}{q}})$, where $q = \max(p, \frac{p}{p-1})$. They also showed that in the case of $\mathcal{L}_1$ and $\mathcal{L}_\infty$ the construction used by Jones does not guarantee convergence to all functions in convex closures of bounded subsets. However, for certain bounded subsets, including sets of functions computable by perceptron networks, Barron (1992) derived an estimate of the uniform approximation error of the form $\frac{V(f,\mathcal{C},J)}{\sqrt{n}}$. Similar estimates were obtained by Girosi (1995).

Since Corollary 4.2 bounds the variation with respect to half-spaces with respect to the topology of uniform convergence, we can combine it with all of these results to estimate both uniform and any $\mathcal{L}_p$ error in approximation by perceptron type networks.

Note that our estimate of $\mathcal{L}_2$ approximation error, given in Theorem 5.1, is different than the estimate obtained by Barron (1993) since we use the method of "plane waves" while Barron used Fourier representation and then approximated sigmoidal activation function by a linear combination of sines. Our bound also involves the factor $a_d = \pm\frac{1}{2}(2\pi)^{1-d}$.

The issue of the size of the numerator in estimates derived from Jones' theorem was raised by Barron (1993), p.932. He notes that when the numerator is derived from an integral representation it may grow as a function of d exponentially fast since it is an integral over a d-dimensional volume. However, our analysis following Corollary 4.3 shows that all of the constants involved may, in fact, be small in reasonable situations. In particular, the volume of the d-dimensional ball of radius r is going as a function of d asymptotically to zero. So Barron's estimates are even somewhat better than were claimed in Barron (1993), p.932.

A result of DeVore et al. (1989) shows that an upper bound on partial derivatives is not sufficient to guarantee dimension-independent rates of approximation by one-hidden-layer neural networks with parameters depending continuously on an approximated function. Our results show that it is sufficient that the product of the volume of J_f times the largest cross-sectional area of a hyperplane section of $supp(f)$ times the largest average d-th iterate of the directional derivative orthogonal to a hyperplane (averaged over the intersection with $supp(f)$) is not growing faster than $(2\pi)^{d-1}$.

7 Proofs

First, we prove several technical lemmas.

Lemma 7.1 *Let X, Y be sets, $J \subseteq X$, $\phi : X \times Y \to \mathcal{R}$ be a function and B be a positive real number. Then conv $\mathcal{G}(\phi, B, J) = \{f : J \to \mathcal{R}; f(x) = \sum_{i=1}^{m} w_i \phi(x, y_i); y_i \in Y; w_i \in \mathcal{R}, \sum_{i=1}^{m} |w_i| \leq B\}$*

Proof.
It is easy to verify once it is recalled that any convex combination of elements, each of norm not exceeding B, also is bounded in norm by B. □

Lemma 7.2 *Let $(\mathcal{F}(X), \|.\|)$ be a normed linear space of real-valued functions on X, $f : X \to \mathcal{R}$, $\{f_i : X \to \mathcal{R}; i \in \mathcal{N}\}$ be a sequence of functions such that $\lim_{i\to\infty} f_i = f$ in $\|.\|$. Let $\phi : X \times Y \to \mathcal{R}$ be such that $\sup_{y\in Y} \|\phi(x, y)\| < \infty$.*

Let $\{B_i; i \in \mathcal{N}\}$ *be a sequence of real numbers such that* $\lim_{i\to\infty} B_i = B$ *and let for every* $i \in \mathcal{N}$ $f_i \in cl\ conv\ \mathcal{G}(\phi, B_i, X)$, *where cl denotes the closure in the topology induced by* $\|.\|$. *If* $\lim_{i\to\infty} f_i = f$ *in* $\|.\|$, *then* $f \in cl\ conv\ \mathcal{G}(\phi, B, X)$.

Proof.
Set $c = \sup_{y\in Y} \|\phi(x,y)\|$. For every $\varepsilon > 0$ choose $i_\varepsilon \in \mathcal{N}$ such that for every $i > i_\varepsilon$ $|B - B_i| < \frac{\varepsilon}{3}$ and $\|f - f_i\| < \frac{\varepsilon}{3}$. Since $f_i \in cl\ conv\ \mathcal{G}(\phi, B_i, X)$ there exists $g_i \in conv\ \mathcal{G}(\phi, B_i, X)$ such that $\|f_i - g_i\| < \frac{\varepsilon}{3}$. So $g_i(x) = \sum_{j=1}^{m_i} w_{ij}\phi(x, y_{ij})$, where $\sum_{j=1}^{m_i} |w_{ij}| \le B_i$. Put $\hat{w}_{ij} = w_{ij} - \frac{\varepsilon}{2cm_i}$ for $w_{ij} > 0$ and $\hat{w}_{ij} = w_{ij} + \frac{\varepsilon}{2cm_i}$ for $w_{ij} < 0$. Put $\hat{g}_i(\mathbf{x}) = \sum_{j=1}^{m_i} \hat{w}_{ij}\phi(\mathbf{x}, \mathbf{y}_{ij})$. Since for all $i \in \mathcal{N}$ $\sum_{j=1}^{m_i} |\hat{w}_{ij}| \le B$ we have $\hat{g}_i \in conv\ \mathcal{G}(\phi, B, X)$. For every $i \ge i_\varepsilon$ $\|f - \hat{g}_i\| \le \|f - f_i\| + \|f_i - g_i\| + \|g_i - \hat{g}_i\| \le \frac{2\varepsilon}{3} + \sum_{j=1}^{m_i} \frac{\varepsilon}{3cm_i}\|\phi(x, y_{ij})\| < \varepsilon$. So, $f \in cl\ conv\ \mathcal{G}(\phi, B, X)$. □

Proof of Theorem 2.2.
Let $\{\mathcal{P}_i; i \in \mathcal{N}\}$ be a sequence of partitions of J_ϕ such that for every $i \in \mathcal{N}$ $\mathcal{P}_{i+1}$ is refining $\mathcal{P}_i$ and diameters of all sets from $\mathcal{P}_i$ are smaller than η_i, where $\lim_{i\to\infty} \eta_i = 0$. Let $\mathcal{P}_i = \{P_{ij}; j \in I_i\}$ and choose basepoints $\mathbf{y}_{ij} \in P_{ij}$. For $\mathbf{x} \in J$, put $f_i(\mathbf{x}) = \sum_{j\in I_i} w(\mathbf{y}_{ij})\phi(\mathbf{x}, \mathbf{y}_{ij})\lambda(P_{ij})$ and let $B_i = \sum_{j\in I_i} |w(\mathbf{y}_{ij})|\lambda(P_{ij})$. By Lemma 7.1, for every $i \in \mathcal{N}$ $f_i \in conv\ \mathcal{G}(\phi, B_i, J)$. Note that B_i and f_i depend on the partition including choice of basepoints.

Since $\lim_{i\to\infty} \eta_i = 0$, the sequence $\{f_i; i \in \mathcal{N}\}$ converges to f on J pointwise. Since w is continuous and compactly supported, the integral $\int_{J_\phi} |w(\mathbf{y})|d\mathbf{y} = B$ exists and $\lim_{i\to\infty} B_i = B$. So by Lemma 7.2 it is sufficient to verify that $\{f_i; i \in \mathcal{N}\}$ converges to f uniformly on J.

It is well-known (see e.g., Kelley, 1955, p. 232) that an equicontinuous family of functions converging pointwise on a compact set converges uniformly. For some $\eta > 0$ choose i_0 such that for every $i \ge i_0$ $\frac{B_i}{B} < 1+\eta$. We will show that continuity of ϕ implies equicontinuity of $\{f_i; i \ge i_0, i \in \mathcal{N}\}$. Indeed, for $\varepsilon > 0$ put $\varepsilon' = \frac{\varepsilon}{1+\eta}$. Since J and $supp(w)$ are compact, ϕ is uniformly continuous on $J \times supp(w)$. Hence there exists ν such that if $|\mathbf{x} - \mathbf{x}'| < \nu$ then for every $\mathbf{y} \in Y$ $|w(\mathbf{y})\phi(\mathbf{x}, \mathbf{y}) - w(\mathbf{y})\phi(\mathbf{x}', \mathbf{y})| < \frac{\varepsilon'}{B}$. Hence for every $i \ge i_0$ $|f_i(\mathbf{x}) - f_i(\mathbf{x}')| = \sum_{j\in J_i} |w(\mathbf{y}_{ij})|\ \lambda(P_{ij})\ |\phi(\mathbf{x}, \mathbf{y}_{ij}) - \phi(\mathbf{x}', \mathbf{y}_{ij})| < \frac{\varepsilon' B_i}{B} < \varepsilon$. □

Proof of Proposition 3.1.
By definition of $V(f, \tau, J)$ for every $n \in \mathcal{N}$ there exists $B_n \in \mathcal{R}$ such that $V(f, \tau, J) \le B_n < V(f, \tau, J) + \frac{1}{n}$ and $f \in conv\ \mathcal{G}(P_\vartheta, B_n, J)$. Since $\lim_{n\to\infty} B_n = V(f, \tau, J)$ and putting $f_n = f$ we get by Lemma 7.2 $f \in cl_\tau \mathcal{G}(P_\vartheta, V(f, \tau, J), J)$. □

Proof of Proposition 3.2.
(i) It is easy to verify that for every $f, g \in \mathcal{B}(J)$, $V(f + g, \tau, J) \le V(f, \tau, J) + V(g, \tau, J)$ and for every $a \in \mathcal{R}$, $V(af, \tau, J) = |a|V(f, \tau, J)$. In particular, $V(f + c, \tau, J) = V(f, \tau, J) + |c|$ for every constant c. Thus, $\mathcal{B}(J)$ is a linear subspace of $\mathcal{F}(J)$ and $V(f, \tau, J)$ is a pseudo-norm on $\mathcal{B}(J)$.
(ii) Since τ is generated by a norm, by Proposition 3.1, $V(f, \tau, J) = \min\{B; f \in cl_\tau conv\ \mathcal{G}(P_\vartheta, B, J)\}$. If $V(f, \tau, J) = 0$ then there exists a sequence $\{f_i; i \in \mathcal{N}\}$ such that $f = \lim_{i\to\infty} f_i$ in τ and for every $i \in \mathcal{N}$ $f_i \in conv\ \mathcal{G}(P_\vartheta, 0, J)$. Since for all $i \in \mathcal{N}$ f_i is a constant equal to 0, we have $\|f\| = 0$. Thus, $V(f, \tau, J)$ is a norm.
(iii) For any $g \in conv\ \mathcal{G}(P_\vartheta, V(f, \tau, J), J)$ $\|g\| \le V(f, \tau, J) \sup\{\|\vartheta(\mathbf{e} \cdot \mathbf{x} + b)\|; \mathbf{e} \in S^{d-1}, b \in \mathcal{R}\}$. So the statement follows by continuity with respect to τ. □

Proof of Proposition 3.3.
When $B = 0$ both $\mathcal{G}(P_\vartheta, B, J)$ and $\mathcal{G}(P_\sigma, B, J)$ consist only of the zero function and so they are equal. Assume that $B > 0$ and let $f \in cl_{\mathcal{L}_p}\ conv\ \mathcal{G}(P_\vartheta, B, J)$.

Then for every $\varepsilon > 0$ $\|f - \sum_{i=1}^{k} w_i \vartheta(\mathbf{e}_i \cdot \mathbf{x} + b_i)\|_p < \frac{\varepsilon}{2}$, where $\sum_{i=1}^{k} |w_i| \leq B$. By boundedness of σ, for every $i \in \{1, \ldots, k\}$ $\lim_{m\to\infty} \sigma(m(\mathbf{e}_i \cdot \mathbf{x} + b_i)) = \vartheta(\mathbf{e}_i \cdot \mathbf{x} + b_i)$ in $\mathcal{L}_p(J)$. So there exists $m_0 \in \mathcal{N}$ such that for every $m \geq m_0$ and for all $i = 1, \ldots, k$ $\|\sigma(m(\mathbf{e}_i \cdot \mathbf{x} + b_i)) - \vartheta(\mathbf{e}_i \cdot \mathbf{x} + b_i)\|_p < \frac{\varepsilon}{2B}$. Hence $\|f - \sum_{i=1}^{k} w_i \sigma(m(\mathbf{v}_i \cdot \mathbf{x} + b_i))\|_p < \|f - \sum_{i=1}^{k} w_i \vartheta(\mathbf{e}_i \cdot \mathbf{x} + b_i)\|_p + \|\sum_{i=1}^{k} w_i(\vartheta(\mathbf{e}_i \cdot \mathbf{x} + b_i) - \sigma(m(\mathbf{e}_i \cdot \mathbf{x} + b_i)))\|_p < \varepsilon$. □

Proof of Proposition 3.4.
Similarly as in the proof of Proposition 3.3 we can assume that $B > 0$. Let $f \in conv\ \mathcal{G}(P_\sigma, B, J)$, i.e. $f(\mathbf{x}) = \sum_{i=1}^{m} w_i \sigma(\mathbf{v}_i \cdot \mathbf{x} + b_i)$ with $\sum_{i=1}^{m} |w_i| \leq B$. Let $\varepsilon > 0$. Since σ is a non-decreasing continuous function with $\lim_{t\to-\infty} \sigma(t) = 0$ and $\lim_{t\to+\infty} \sigma(t) = 1$, there exists $k \in \mathcal{N}$ and $a_j, c_j \in \mathcal{R}$ $(j = 1, \ldots, k)$ such that for every $t \in \mathcal{R}$ $\left|\sigma(t) - \sum_{j=1}^{k} a_j \vartheta(t + c_j)\right| < \frac{\varepsilon}{B}$ and for every $j = 1, \ldots, k$ $a_j \geq 0$ and $\sum_{j=1}^{k} a_j = 1$. Define $f_\varepsilon(\mathbf{x}) = \sum_{i=1}^{m} \sum_{j=1}^{k} w_i a_j \vartheta(\mathbf{v}_i \cdot \mathbf{x} + b_i + c_j)$. Then $f_\varepsilon \in conv\ \mathcal{G}(P_\vartheta, B, J)$ and for every $\mathbf{x} \in J$ $|f(\mathbf{x}) - f_\varepsilon(\mathbf{x})| \leq \sum_{i=1}^{m} |w_i| \left|\sigma(\mathbf{v}_i \cdot \mathbf{x} + b_i) - \sum_{j=1}^{m} \vartheta(\mathbf{v}_i \cdot \mathbf{x} + b_i + c_j)\right| \leq \sum_{i=1}^{m} |w_i| \frac{\varepsilon}{B} \leq \varepsilon$. Hence $conv\ \mathcal{G}(P_\sigma, B, J) \subseteq cl_{\mathcal{C}} conv\ \mathcal{G}(P_\vartheta, B, J)$ and so also $cl_{\mathcal{C}} conv\ \mathcal{G}(P_\sigma, B, J) \subseteq cl_{\mathcal{C}} conv\ \mathcal{G}(P_\vartheta, B, J)$. Similarly $cl_{\mathcal{L}_p}\ conv\ \mathcal{G}(P_\sigma, B, J) \subseteq cl_{\mathcal{L}_p}\ conv\ \mathcal{G}(P_\vartheta, B, J)$. □

Lemma 7.3 *Let $K \subseteq S^{d-1} \times \mathcal{R}$ and w be a measurable and bounded function on K. For every $m \in \mathcal{N}$ let σ_m be the function on $\mathcal{R}$ defined by $\sigma_m(t) = (1 + e^{-mt})^{-1}$. Then*
$\lim_{m\to\infty} \int_K w(\mathbf{e}, b)\sigma_m(\mathbf{e} \cdot \mathbf{x} + b)d(\mathbf{e}, b) = \int_K w(\mathbf{e}, b)\vartheta(\mathbf{e} \cdot \mathbf{x} + b)d(\mathbf{e}, b)$ *uniformly on $\mathcal{R}^d$.*

Proof.
Because w is bounded, we only need to verify that $\lim_{m\to\infty} \int_K |(\vartheta - \sigma_m)(\mathbf{e} \cdot \mathbf{x} + b)|d(\mathbf{e}, b) = 0$ uniformly on $\mathcal{R}^d$. For every $\mathbf{x} \in \mathcal{R}^d$ and for every $m \in \mathcal{N}$ $\int_K |(\vartheta - \sigma_m)(\mathbf{e} \cdot \mathbf{x} + b)|d(\mathbf{e}, b) \leq \lambda_{d-1}(S^{d-1}) \int_{\mathcal{R}} |(\vartheta - \sigma_m)(t)|dt$. Since $\vartheta = \lim_{m\to\infty} \sigma_m$ in $\mathcal{L}_1$ the proof is complete. □

Proof of Theorem 3.5.
Let $\{\sigma_m; m \in \mathcal{N}\}$ be the sequences of functions defined in Lemma 7.3. For every $m \in \mathcal{N}$ define $f_m(\mathbf{x}) = \int_K w(\mathbf{e}, b)\sigma_m(\mathbf{e} \cdot \mathbf{x} + b)d(\mathbf{e}, b)$, i.e. f_m is obtained from the integral representation of f by replacing the Heaviside function ϑ with σ_m.

Let $B = \int_K |w(\mathbf{e}, b)|d(\mathbf{e}, b)$. For every m, if $\phi = P_{\sigma_m}$, then $J_\phi = \{(\mathbf{e}, b) \in S^{d-1} \times \mathcal{R}; (\exists \mathbf{x} \in J)(w(\mathbf{e}, b)\sigma_m(\mathbf{e} \cdot \mathbf{x} + b) \neq 0)\} = supp(w)$. Since $supp(w) \subseteq K$ we have $\int_{J_\phi} |w(\mathbf{e}, b)|d(\mathbf{e}, b) = B$. By Corollary 2.3 and the remark following Theorem 2.2 for every $m \in \mathcal{N}$ $f_m \in cl_{\mathcal{C}}\ conv\ \mathcal{G}(P_{\sigma_m}, B, J)$.

By definition for all $m \in \mathcal{N}$ $\mathcal{G}(P_{\sigma_m}, B, J) = \mathcal{G}(P_\sigma, B, J)$. By Proposition 3.4 $cl_{\mathcal{C}}\ conv\ \mathcal{G}(P_\sigma, B, J) \subseteq cl_{\mathcal{C}} conv\ \mathcal{G}(P_\vartheta, B, J)$. Since by Lemma 7.3 $f = \lim_{m\to\infty} f_m$ uniformly on J we can apply Lemma 7.2 to get $f \in cl_{\mathcal{C}} conv\ \mathcal{G}(P_\vartheta, B, J)$. □

To prove Theorem 4.1 we need two technical lemmas. The first one can be found in Courant & Hilbert (1992), p.680.

Lemma 7.4 *For every odd positive integer d*

$$\delta_d(\mathbf{x}) = a_d \int_{S^{d-1}} \delta_1^{(d-1)}(\mathbf{e} \cdot \mathbf{x})d\mathbf{e},$$

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

Lemma 7.5 *For all positive integers d,k, for every function* $f \in \mathcal{C}^d(\mathcal{R}^d)$ *and for every unit vector* $\mathbf{e} \in \mathcal{R}^d$ *and for every* $b \in \mathcal{R}$ $\frac{\partial^k}{\partial b^k} \int_{H_{\mathbf{e}b}} f(\mathbf{y})d\mathbf{y} = \int_{H_{\mathbf{e}b}} D_{\mathbf{e}}{}^{(k)} f(\mathbf{y})d\mathbf{y}$.

Proof.
First, we will verify that the statement is true for $k = 1$:

$$\frac{\partial}{\partial b}\int_{H_{\mathbf{e}b}} f(\mathbf{y})d\mathbf{y} = \lim_{t\to 0} t^{-1}\left(\int_{H_{\mathbf{e}b}} f(\mathbf{y})d\mathbf{y} - \int_{H_{\mathbf{e}b+t}} f(\mathbf{y})d\mathbf{y}\right) =$$

$$\lim_{t\to 0} t^{-1}\int_{H_{\mathbf{e}b}} (f(\mathbf{y}+t\mathbf{e})-f(\mathbf{y}))d\mathbf{y} = \int_{H_{\mathbf{e}b}} \lim_{t\to 0} t^{-1}(f(\mathbf{y}+t\mathbf{e})-f(\mathbf{y})) = \int_{H_{\mathbf{e}b}} D_{\mathbf{e}} f(\mathbf{y})d\mathbf{y}.$$

Suppose that the statement is true for $k-1$. Then

$$\frac{\partial^k}{\partial b^k}\int_{H_{\mathbf{e}b}} f(\mathbf{y})d\mathbf{y} = \lim_{t\to 0} t^{-1}\left(\int_{H_{\mathbf{e}b}} D_{\mathbf{e}}^{(k-1)} f(\mathbf{y})d\mathbf{y} - \int_{H_{\mathbf{e}b+t}} D_{\mathbf{e}}^{(k-1)} f(\mathbf{y})d\mathbf{y}\right) =$$

$$\lim_{t\to 0} t^{-1}\int_{H_{\mathbf{e}b}} \left(D_{\mathbf{e}}^{(k-1)} f(\mathbf{y}+t\mathbf{e}) - D_{\mathbf{e}}^{(k-1)} f(\mathbf{y})\right) d\mathbf{y} =$$

$$\int_{H_{\mathbf{e}b}} \lim_{t\to 0} t^{-1}\left(D_{\mathbf{e}}^{(k-1)} f(\mathbf{y}+t\mathbf{e}) - D_{\mathbf{e}}^{(k-1)} f(\mathbf{y})\right) = \int_{H_{\mathbf{e}b}} D_{\mathbf{e}}^{(k)} f(\mathbf{y})d\mathbf{y}.\square$$

Proof of Theorem 4.1.
We first prove the theorem for test functions. For $f \in \mathcal{D}(\mathcal{R}^d)$ by definition of the delta distribution we have $f(\mathbf{x}) = (f * \delta_d)(\mathbf{x}) = \int_{\mathcal{R}^d} f(\mathbf{z})\delta_d(\mathbf{x}-\mathbf{z})d\mathbf{z}$ (see e.g., Zemanian, 1987). By Lemma 7.3 $\delta_d(\mathbf{x}-\mathbf{z}) = a_d \int_{S^{d-1}} {\delta_1}^{(d-1)}(\mathbf{e}\cdot\mathbf{x} - \mathbf{e}\cdot\mathbf{z})d\mathbf{e}$. Thus, $f(\mathbf{x}) = a_d \int_{S^{d-1}} \int_{\mathcal{R}^d} f(\mathbf{z}){\delta_1}^{(d-1)}(\mathbf{x}\cdot\mathbf{e} - \mathbf{z}\cdot\mathbf{e})d\mathbf{z}d\mathbf{e}$. So rearranging the inner integration, we have
$f(\mathbf{x}) = a_d \int_{S^{d-1}} \int_{\mathcal{R}} \int_{H_{\mathbf{e}b}} f(\mathbf{y}){\delta_1}^{(d-1)}(\mathbf{x}\cdot\mathbf{e}+b)d\mathbf{y}dbd\mathbf{e}$, where $H_{\mathbf{e}b} = \{\mathbf{y} \in \mathcal{R}; \mathbf{y}\cdot\mathbf{e} = -b\}$. Let $u(\mathbf{e},b) = a_d \int_{H_{\mathbf{e}b}} f(\mathbf{y})d\mathbf{y}$, so $f(\mathbf{x}) = \int_{S^{d-1}} \int_{\mathcal{R}} u(\mathbf{e},b){\delta_1}^{(d-1)}(\mathbf{x}\cdot\mathbf{e}+b)dbd\mathbf{e}$.

By definition of distributional derivative $\int_{\mathcal{R}} u(\mathbf{e},b){\delta_1}^{(d-1)}(\mathbf{e}\cdot\mathbf{x}+b)db =$
$(-1)^{d-1}\int_{\mathcal{R}} \frac{\partial^{d-1}u(\mathbf{e},b)}{\partial b^{d-1}}\delta_1(\mathbf{e}\cdot\mathbf{x}+b)db$ for every $\mathbf{e} \in S^{d-1}$ and $\mathbf{x} \in \mathcal{R}^d$. Since d is odd, we have $\int_{\mathcal{R}} u(\mathbf{e},b){\delta_1}^{(d-1)}(\mathbf{e}\cdot\mathbf{x}+b)db = \int_{\mathcal{R}} \frac{\partial^{d-1}u(\mathbf{e},b)}{\partial b^{d-1}}\delta_1(\mathbf{e}\cdot\mathbf{x}+b)db$.

Since the first distributional derivative of the Heaviside function is the delta distribution (see e.g., Zemanian, 1987, p.47), it follows that for every $\mathbf{e} \in S^{d-1}$ and $\mathbf{x} \in \mathcal{R}^d$ $\int_{\mathcal{R}} u(\mathbf{e},b){\delta_1}^{(d-1)}(\mathbf{e}\cdot\mathbf{x}+b)db = -\int_{\mathcal{R}} \frac{\partial^d u(\mathbf{e},b)}{\partial b^d}\vartheta(\mathbf{e}\cdot\mathbf{x}+b)db$.

By Lemma 7.4 $\frac{\partial^d u(\mathbf{e},b)}{\partial b^d} = \frac{\partial^d}{\partial b^d}\int_{H_{\mathbf{e}b}} f(\mathbf{y})d\mathbf{y} = \int_{H_{\mathbf{e}b}} D_{\mathbf{e}}^{(d)} f(\mathbf{y})d\mathbf{y}$. Hence, $f(\mathbf{x}) = -a_d \int_{S^{d-1}} \int_{\mathcal{R}} \left(\int_{H_{\mathbf{e}b}} D_{\mathbf{e}}^{(d)} f(\mathbf{y})d\mathbf{y}\right)\vartheta(\mathbf{e}\cdot\mathbf{x}+b)dbd\mathbf{e}$.

Now let $f \in \mathcal{C}^d(\mathcal{R}^d)$ be compactly supported. Then there exists a sequence $\{f_i; i \in \mathcal{N}\}$ of test functions converging to f uniformly on $\mathcal{R}^d$ (see e.g., Zemanian, 1987, p.3). It is easy to check that for every $\mathbf{e} \in S^{d-1}$ $\{D_{\mathbf{e}}^{(d)} f_i; i \in \mathcal{N}\}$ converges uniformly on $\mathcal{R}^d$ to $D_{\mathbf{e}}^{(d)} f$. Hence we can interchange limit and integration (see e.g., Edwards, 1994, p.233). So $\lim_{i\to\infty} \int_{H_{\mathbf{e}b}} D_{\mathbf{e}}^{(d)} f_i(\mathbf{y})d\mathbf{y} = \int_{H_{\mathbf{e}b}} D_{\mathbf{e}}^{(d)} f(\mathbf{y})d\mathbf{y}$. Put $g_i(\mathbf{x},\mathbf{e},b) = \int_{H_{\mathbf{e}b}} \left(D_{\mathbf{e}}^{(d)} f_i(\mathbf{y})d\mathbf{y}\right)\vartheta(\mathbf{e}\cdot\mathbf{x}+b)$ and $g(\mathbf{x},\mathbf{e},b) = \int_{H_{\mathbf{e}b}} \left(D_{\mathbf{e}}^{(d)} f(\mathbf{y})d\mathbf{y}\right)\vartheta(\mathbf{e}\cdot\mathbf{x}+b)$. It is easy to see that for every $\mathbf{x} \in \mathcal{R}^d$
$\lim_{i\to\infty} g_i(\mathbf{x},\mathbf{e},b) = g(\mathbf{x},\mathbf{e},b)$ uniformly on $S^{d-1}\times\mathcal{R}$. Hence for every $\mathbf{x} \in \mathcal{R}^d$
$f(\mathbf{x}) = \lim_{i\to\infty} \int_{S^{d-1}} \int_{\mathcal{R}} g_i(\mathbf{x},\mathbf{e},b)dbd\mathbf{e} = \int_{S^{d-1}} \int_{\mathcal{R}} g(\mathbf{x},\mathbf{e},b)dbd\mathbf{e} =$

$\int_{S^{d-1}} \int_{\mathcal{R}} \left(\int_{H_{\mathbf{e}b}} D_{\mathbf{e}}^{(d)} f(\mathbf{y}) d\mathbf{y} \right) \vartheta(\mathbf{e} \cdot \mathbf{x} + b) dbd\mathbf{e}$ (using again interchangebility of integration and limit for a sequence of functions converging uniformly). □

References

Ball, K. (1986). Cube slicing in R^d. *Proceedings of AMS*, **99**, 1-10.

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

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

Courant, R., Hilbert, D. (1992). *Methods of Mathematical Physics, vol.II.* New York: Interscience.

Darken, C., Donahue, M., Gurvits, L., & Sontag, E. (1993). Rate of approximation results motivated by robust neural network learning. In *Proceedings of the 6th Annual ACM Conference on Computational Learning Theory* (pp. 303-309). New York: ACM.

DeVore, R., Howard, R., & Micchelli, C. (1989). Optimal nonlinear approximation. *Manuscripta Mathematica*, **63**, 469-478.

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

Girosi, F. (1995). Approximation error bounds that use VC-bounds. In *Proceedings of ICANN'95* (pp. 295-302). Paris:EC2 & Cie.

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

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

Hewitt, E., & Stromberg, K. (1965). *Real and Abstract Analysis.* New York: Springer.

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

Jones, L. K. (1992). A simple lemma on greedy approximation in Hilbert space and convergence rates for projection pursuit regression and neural network training. *The Annals of Statistics*, **20**, 601-613.

Kelley, J. L. (1955). *General Topology.* Princeton: Van Nostrand.

Kůrková, V. (1992). Kolmogorov's theorem and multilayer neural networks. *Neural Networks*, **5**, 501-506.

McShane, E. J. (1944). *Integration.* Princeton: Princeton University Press.

Mhaskar, H. N., & Micchelli, C. A. (1992). Approximation by superposition of sigmoidal and radial basis functions. *Advances in Applied Mathematics*, **13**, 350-373.

Park, J., & Sandberg, I. W. (1993). Approximation and radial-basis-function networks. *Neural Computation*, **5**, 305-316.

Rudin, W. (1964). *Principles of Mathematical Analysis.* New York: McGraw-Hill.

Rudin, W. (1973). *Functional Analysis.* New York: McGraw-Hill.

Sejnowski, T. J., & Yuhas, B. P. (1989). Mapping between high-dimensional representations of acoustic and speech signal. In *Computation and Cognition* (pp. 52-68). Philadelphia: Siam.

Zemanian, A. H. (1987). *Distribution Theory and Transform Analysis.* New York: Dover.