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

Kůrková, Věra
1995

INSTITUTE OF COMPUTER SCIENCE

ACADEMY OF SCIENCES OF THE CZECH REPUBLIC

Estimates of the number of hidden units and variation with respect to half-spaces

Věra Kůrková
Institute of Computer Science,
Academy of Sciences of the Czech Republik,
P.O. Box 5, 182 07 Prague 8, Czech Republik
Paul C. Kainen
Industrial Math, 3044 N St., N.W.,
Washington, D.C. 20007
Vladik Kreinovich
Department of Computer Science,
University of Texas at El Paso,
El Paso, TX 79968

Technical report No. 645

Institute of Computer Science, Academy of Sciences of the Czech Republic
Pod vodárenskou věží 2, 182 07 Prague 8, Czech Republic
phone: (+422) 66414244 fax: (+422) 8585789
e-mail: vera@uivt.cas.cz

INSTITUTE OF COMPUTER SCIENCE

ACADEMY OF SCIENCES OF THE CZECH REPUBLIC

Estimates of the number of hidden units and variation with respect to half-spaces

Věra Kůrková[1]
Institute of Computer Science,
Academy of Sciences of the Czech Republik,
P.O. Box 5, 182 07 Prague 8, Czech Republik
Paul C. Kainen
Industrial Math, 3044 N St., N.W.,
Washington, D.C. 20007
Vladik Kreinovich [2]
Department of Computer Science,
University of Texas at El Paso,
El Paso, TX 79968

Technical report No. 645

Abstract

Utilizing an integral representation of smooth functions of d variables proved using properties of delta and Heaviside distributions we estimate variation with respect to half-spaces in terms of "flows through hyperplanes". Consequently we obtain conditions which guarantee $\mathcal{L}_2$ 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]V. K. was partially supported by GACR Grant 201/93/0427.

[2]V.K. was partially supported by NSF Grant No. CDA-9015006 and NASA Research Grant No. NAG 9-757.

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 d-dimensional box, has been proven for one-hidden-layer networks with perceptron or radial-basis-function units with quite general activation functions (see, e.g. Mhaskar and Micchelli [14], Park and Sandberg [15]).

However, estimates of the number of hidden units that guarantee a given accuracy of an approximation are less understood. Most upper estimates grow exponentially with the number of input units, i.e. with the number d of input variables of the function f to be approximated (e.g., Mhaskar and Micchelli [14], Kůrková [12]). A general result by deVore et al. [7] 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. But in applications, functions of hundreds of variables are approximated sufficiently well by neural networks with only moderately many hidden units (e.g., Sejnowski and Yuhas [18]).

Jones [10] 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 together with Barron proposed to apply it to the space of functions achievable by a one-hidden-layer neural network. Applying Jones' estimate Barron [1] 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 [2] 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 (3.2) bounds variation with respect to half-spaces for functions represented by a "neural network with a continuum of Heaviside perceptrons". Our second result (4.1) gives such a representation with output weights corresponding to flows orthogonal to hyperplanes determined by the input weights and biases. As a result, we show that the variation with respect to half spaces of a sufficiently smooth, compactly supported function f defined on $\mathcal{R}^d$, for d odd, is bounded above by a constant of order $\mathcal{O}(2\pi)^{1-d}$ times the integral over parameters of all perceptrons of an integrand which is the absolute value of the integral of the d-th directional derivative of f over the cozero hyperplane of the affine functions determined by perceptron parameters (weight vector and bias). So variation with respect to half-spaces is bounded above by the supremum of absolute values of integrals of directional derivatives of order d over orthogonal hyperplanes multiplied by a d-dimensional volume. For single variable functions our bound is identical with a well-known bound on 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 also use a representation of the d-dimensional delta distribution as an integral over the unit sphere 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.

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.

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 $\mathcal{C}^d(X)$ denotes the *set of all real-valued functions on* X *with continuous 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 [10] estimated rates of approximation of functions from convex closures of bounded subsets of a Hilbert space. The following is a slight reformulation of his result.

Theorem 2.1 *Let $\mathcal{H}$ be a Hilbert space, with a norm $\|.\|$, B 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 $f \in cl(conv(\mathcal{G}))$ and for every natural number n there exists f_n that is a convex combination of n*

elements of $\mathcal{G}$ *such that*

$$\|f - f_n\| \leq \frac{\|f\| + B}{\sqrt{n}} \leq \frac{2B}{\sqrt{n}}.$$

To use this theorem to estimate the number of hidden units in neural networks, we need to investigate the convex closures of sets 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 [1] used Fourier representation, Girosi and Anzellotti [8] 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 and let $f \in \mathcal{C}(\mathcal{R}^d)$ 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, $w \in \mathcal{C}(Y)$ compactly supported and $\phi \in \mathcal{C}(\mathcal{R}^d \times Y)$. Then for every compact subset $J \subset \mathcal{R}^d$ $f \in cl_{\mathcal{C}}(conv(\mathcal{G}(\phi, B, J)))$, with $B = \int_{J^*} |w(\mathbf{y})|d\mathbf{y}$ where $J^* = \{y \in Y; (\exists x \in J)(w(y)\phi(x, y) \neq 0)\}$.*

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 $\phi(\mathbf{x}, \mathbf{v}, b) = \psi(\mathbf{v} \cdot \mathbf{x} + b)$. Let $\mathcal{E}_d(\psi, B, J) = \mathcal{G}(\phi, B, J)$. So $\mathcal{E}_d(\psi, B, J)$ 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. Typically, ψ is *sigmoidal*, i.e. it satisfies $\lim_{t\to\infty} \psi(t) = 1$ and $\lim_{t\to-\infty} \psi(t) = 0$.

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

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 $B = \int_{J^*} |w(\mathbf{v}, b)|d(\mathbf{v}, b)$.

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 the non-constant functions in $\mathcal{E}_d(\vartheta, B, J)$ are exactly the set $\{g : J \to \mathcal{R};\ g(\mathbf{x}) = w\vartheta(\mathbf{e} \cdot \mathbf{x} + b), \mathbf{e} \in S^{d-1}, w, b \in \mathcal{R}, |w| \leq B\}$, where S^{d-1} denotes the unit sphere in $\mathcal{R}^d$.

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

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

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 $V(f|_J,\tau,J)$.

It is easy to verify that when the topology τ is induced by a norm, this infimum is achieved, i.e., $f \in cl_\tau(conv(\mathcal{E}_d(\vartheta,V(f,\tau,J),J)))\}$. Also, for every $f,g \in \mathcal{F}(J)$, $V(f+g,\tau,J) \le V(f,\tau,J)+V(g,\tau,J)$ 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.

Let $p \in [1,\infty]$. Since for every $X \subseteq \mathcal{L}_p(J)$ we have $cl_{\mathcal{C}}(X) \subseteq cl_{\mathcal{L}_p}(X)$, clearly $V(f,\mathcal{L}_p,J) \le V(f,\mathcal{C},J)$.

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 by $T(f,[s,t]) = \sup\{\sum_{i=1}^k |f(t_{i+1})-f(t_i)|; s = t_1 < \ldots < t_k = t, k \in \mathcal{N}\}$ (see e.g. [13]). 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 $T(f,[s,t]) = V(f,\mathcal{C},[s,t])$ (see Barron [2], also Darken et al. [5, Theorem 6]).

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 $T(f,J) = \sup\{\sum_{i=1}^k |f(J_i)|$, where $\{J_i; i = 1,\ldots,k\}$ is a subdivision of J into boxes $\}$, $f(J_i) = \sum_{j=1}^{2^d}(-1)^{\nu(j)}f(\mathbf{x}_{ij})$, $\{\mathbf{x}_{ij}; j = 1,\ldots,2^d\}$ are the corner points of J_i and $\nu(j) = \pm 1$ is a parity [13]. For $d \ge 2$ this concept is different from Barron's variation with respect to half-spaces. For example, the characteristic function χ of the set $\{(x_1,x_2) \in [0,1]^2; x_1 \ge x_2\}$ 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 $T(f,J) = \int_J |f'(x)|dx$ [13, p.242]. The corollary below extends this to variation with respect to half-spaces.

Theorem 3.1 *Let d be any positive integer and let $f \in \mathcal{C}(\mathcal{R}^d)$ be any function that can be represented as $f(\mathbf{x}) = \int_{S^{d-1}}\int_{\mathcal{R}} w(\mathbf{e},b)\vartheta(\mathbf{e}\cdot\mathbf{x}+b)dbd\mathbf{e}$, where $w \in \mathcal{C}(S^{d-1}\times\mathcal{R})$ is compactly supported. Then for every compact subset J in $\mathcal{R}^d$ and every $p \in [1,\infty)$ $f \in cl_{\mathcal{L}_p}(conv(\mathcal{E}_d(\vartheta,B,J)))$, where $B = \int_{J^*}|w(\mathbf{e},b)|d(\mathbf{e},b)$, where $J^* = \{(\mathbf{e},b) \in S^{d-1}\times\mathcal{R}; (\exists\mathbf{x} \in J)(w(\mathbf{e},b)\vartheta(\mathbf{e}\cdot\mathbf{x}+b) \ne 0)\}$.*

Corollary 3.2 *Let d be any positive integer and let $f \in \mathcal{C}(\mathcal{R}^d)$ be any function that can be represented as $f(\mathbf{x}) = \int_{S^{d-1}}\int_{\mathcal{R}} w(\mathbf{e},b)\vartheta(\mathbf{e}\cdot\mathbf{x}+b)dbd\mathbf{e}$, where $w \in \mathcal{C}(S^{d-1}\times\mathcal{R})$ is compactly supported. Then for every compact subset $J \subset \mathcal{R}^d$ and for every $p \in [1,\infty)$ $V(f,\mathcal{L}_p,J) \le \int_{J^*}|w(\mathbf{e},b)|d(\mathbf{e},b)$.*

4 Integral representation theorem

To estimate variation with respect to half-spaces using Corollary 3.2 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 [17] that the *directional derivative* $D_{\mathbf{e}}f(\mathbf{y})$ 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., [17, 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 [6, 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*

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

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

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 $\delta_d(\mathbf{x})=a_d\int_{S^{d-1}}\delta_1{}^{(d-1)}(\mathbf{e}\cdot\mathbf{x})d\mathbf{e}$ [4, p.680] (by $\delta_1{}^{(d-1)}$ is denoted the d-1-st distributational 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 [19, p.3].

Integral representation 4.1 together with Corollary 3.2 give the following estimate of variation with respect to half-spaces.

Theorem 4.2 *For every odd positive integer d, a compact subset $J\subset\mathcal{R}^d$, $f\in\mathcal{C}^d(\mathcal{R}^d)$ and for every $p\in[1,\infty)$*

$$V(f,\mathcal{L}_p,J)\le|a_d|\int_{J^*}\left|\int_{H_{\mathbf{e}b}}D_{\mathbf{e}}{}^{(d)}f(\mathbf{y})d\mathbf{y}\right|d(\mathbf{e},b)$$

where $|a_d|=(1/2)(2\pi)^{1-d}$ *and* $J^*=\{(\mathbf{e},b)\in S^{d-1}\times\mathcal{R};(\exists\mathbf{x}\in J)(w(\mathbf{e},b)\vartheta(\mathbf{e}\cdot\mathbf{x}+b)\neq 0)\}$.

It is easy to verify that when $d = 1$ Theorem 4.2 gives estimate $V(f, \mathcal{C}, J) \leq \int_J |f'(b)|db$ which agrees with the above mentioned characterization of total variation for functions of one variable.

Estimating the integrals in Theorem 4.2 we get the following corollary. We write λ_k to denote the Lebesgue measure on $\mathcal{R}^k$.

Corollary 4.3 *For every odd positive integer d, a compact $J \subset \mathcal{R}^d$, $f \in \mathcal{C}^d(\mathcal{R}^d)$ with $supp(f)$ a d-dimensional cube and for every $p \in [1, \infty)$*

$$V(f, \mathcal{L}_p, J) \leq \sqrt{2}|a_d|\lambda_d(J^*)\lambda_d(supp(f)) \sup\{\left|D_{\mathbf{e}}{}^{(d)}(\mathbf{y})\right| ; \mathbf{y} \in supp(f), \mathbf{e} \in S^{d-1}\}.$$

Using the Radon transform, Ito [9] obtained an integral representation as in Corollary 3.2. Our proof of Theorem 4.1 uses a different approach and describes coefficients $w(\mathbf{e}, b)$ in terms of directional derivatives.

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.

Lemma 5.1 *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 $p \in [1, \infty)$*

$$cl_{\mathcal{L}_p}(conv(\mathcal{E}_d(\vartheta, B, J))) \subset cl_{\mathcal{L}_p}(conv(\mathcal{E}_d(\sigma, B, J))).$$

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 estimate of $V(f, \mathcal{L}_p, J)$ given by Theorem 4.2, i.e. $B_f = |a_d| \int_{J^*} \left| \int_{H_{\mathbf{e}b}} D_{\mathbf{e}}{}^{(d)} f(\mathbf{y}) d\mathbf{y} \right| d(\mathbf{e}, b)$.

Theorems 2.1, 4.2 and Lemma 5.1 imply the following estimate of rates of approximation by one-hidden-layer networks with sigmoidal perceptrons.

Theorem 5.2 *Let d be an odd positive integer, $f \in \mathcal{C}^d(\mathcal{R}^d)$ compactly supported and $\sigma : \mathcal{R} \to \mathcal{R}$ be a bounded sigmoidal function. 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 $\|f - f_n\|_2 \leq \frac{B_f + \|f\|_2}{\sqrt{n}}$.*

6 Discussion

A result of DeVore et al. [7] shows that an upper bound on partial derivatives is not sufficient to guarantee dimension-independent rates of approximation by one-hidden-layer neural networks. Our results show that it *is* sufficient to bound the d-th directional derivatives multiplied by the product of the d-dimensional volume of the support of the

function and the d-dimensional volume of J^*. Since d-dimensional volume can grow exponentially with increasing dimension, to keep $V(f, \mathcal{L}_2, J)$ bounded by the same bound B, the flows of order d must be decreasing with increasing d.

Thus, the dimension-independent rates of approximation must be interpreted and used carefully. The constant factor $2B_f$ can at realistic scales, dominate the $\frac{1}{\sqrt{n}}$ factor. The size of spaces of functions that can be approximated with rates of approximation $\mathcal{O}(\frac{1}{\sqrt{n}})$ is decreasing with increasing input dimension d.

7 Proofs

First, we prove several technical lemmas.

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

Proof. It is easy to verify once you recall 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. Put $c = \sup_{y\in Y} \|\phi(x, y)\|$. 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 $g_i(x) = \sum_{j=1}^{m_i} a_{ij} u_{ij} \phi(x, y_{ij})$, where a_{ij} are coefficients of convex combination and $|u_{ij}| \leq B_i$. Put $\hat{u}_{ij} = u_{ij} - \frac{\varepsilon}{2c}$ for $u_{ij} > 0$ and $\hat{u}_{ij} = u_{ij} + \frac{\varepsilon}{2c}$ for $u_{ij} < 0$. Put $\hat{g}_i(\mathbf{x}) = \sum_{j=1}^{m_i} a_{ij} \hat{u}_{ij} \phi(\mathbf{x}, \mathbf{y}_{ij})$. Since for all $i \in \mathcal{N}$ and $j \in \{1, \ldots, m_i\}$ $|\hat{u}_{ij}| \leq B$ we have $\hat{g}_i \in conv(\mathcal{G}(\phi, B, X))$. For every $i \geq i_\varepsilon$ $\|f - \hat{g}_i| \leq \|f - f_i\| + \|f_i - g_i\| + \|g_i - \hat{g}_i\| \leq \frac{2\varepsilon}{3} + \sum_{j=1}^{m_i} a_{ij} \frac{\varepsilon}{3c} \|\phi(x, y_{ij})\| < \varepsilon$. 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^* 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 $\mathbf{y}_{ij} \in P_{ij}$. 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 $i \in \mathcal{N}$ $f_i \in conv(\mathcal{G}(\phi, B_i, J))$.

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 $\int_{J^*} |w(\mathbf{y})| d\mathbf{y} = B$ 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. [11, 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 $i \geq i_0$ $\frac{B_i}{B} < 1 + \eta$. We will show that continuity of $w\phi$ implies equicontinuity of $\{f_i; i \geq i_0, i \in \mathcal{N}\}$. Indeed, for $\varepsilon > 0$ put $\varepsilon' = \frac{\varepsilon}{1+\eta}$. Since J is compact, $w\phi$ is uniformly continuous on J. 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 $i \geq i_0$ $|f_i(\mathbf{x}) - f(\mathbf{x}')| = \sum_{j \in J_i} |w(\mathbf{y}_{ij})|\lambda(P_{ij})|\phi(\mathbf{x}, \mathbf{y}_{ij}) - \phi(\mathbf{x}', \mathbf{y}_{ij})| < \frac{\varepsilon' B_i}{B} < \varepsilon$. $\square$

Proof of Theorem 3.1. Let $\sigma : \mathcal{R} \to \mathcal{R}$ be the logistic sigmoidal function, i.e. $\sigma(t) = \frac{1}{1+e^{-t}}$. For every $m \in \mathcal{N}$ put $f_m(\mathbf{x}) = \int_{S^{d-1}} \int_{\mathcal{R}} w(\mathbf{e}, b)\sigma(m(\mathbf{e} \cdot \mathbf{x} + b))dbd\mathbf{e}$.

It is easy to verify that $\lim_{m\to\infty} f_m = f$ in $\mathcal{L}_p(J)$. Let $\{\mathcal{P}_i; i \in \mathcal{N}\}$ be a sequence of partitions of $supp(w)$ as in the proof of Theorem 2.2. and let $\mathcal{P}_i = \{P_{ij}; j \in I_i\}$ and let $(\mathbf{e}_{ij}, b_{ij}) \in P_{ij}$. Put $B_i = \sum_{j \in I_i} |w(\mathbf{e}_{ij}, b_{ij})|\lambda(P_{ij})$ and $f_{mi}(\mathbf{x}) = \sum_{j \in I_i} w(\mathbf{e}_{ij}, b_{ij})\sigma(m(\mathbf{e}_{ij} \cdot \mathbf{x} + b_{ij})\lambda(P_{ij})$.

Since J is compact we have as in the proof of Theorem 2.2 for each $m \in \mathcal{M}$ $\lim_{i\to\infty} f_{mi} = f_m$ uniformly on J. Put $c_m = \sup\{\|\sigma(m(\mathbf{e} \cdot \mathbf{x} + b)) - \vartheta(\mathbf{e} \cdot \mathbf{x} + b)\|_p; \mathbf{e} \in S^{d-1}, b \in \mathcal{R}\}$ where $\|.\|_p$ denotes the $\mathcal{L}_p$-norm on J. Since $\lim_{m\to\infty} f_m = f$ uniformly, $\lim_{i\to\infty} f_{mi} = f_m 0$ in $\mathcal{L}_p$, $\lim_{i\to\infty} B_i = B$ and $\lim_{m\to\infty} c_m = 0$, we can construct recursively two strictly increasing sequences of natural numbers $\{m_n; n \in \mathcal{N}\}$ and $\{i_n; n \in \mathcal{N}\}$ such that for all $n \in \mathcal{N}$ $\|f - f_{m_n}\|_p < \frac{1}{3n}$, $\|f_{m_n} - f_{m_n i_n}\|_c < \frac{1}{3n}$ and $c_{m_n} B_{i_n} < \frac{1}{3n}$.

For every $n \in \mathcal{N}$ put $h_n(\mathbf{x}) = \sum_{j \in J_{i_n}} w(\mathbf{e}_{i_n j}, b_{i_n j})\vartheta(\mathbf{e}_{i_n j} \cdot \mathbf{x} + b_{i_n j})\lambda(P_{i_n j})$. Since for all $n \in \mathcal{N}$ $h_n \in conv(\mathcal{E}_d(\vartheta, B_{i_n}, J)$ by Lemma 7.2 it is sufficient to verify that $\lim_{n\to\infty} h_n = f$ in $\mathcal{L}_p(J)$. Indeed, for every $n \in \mathcal{N}$ we have $\|f - h_n\|_p \leq \|f - f_{m_n}\|_p + \|f_{m_n} - f_{m_n i_n}\|_p + \|f_{m_n i_n} - h_n\|_p < \frac{1}{n}$. $\square$

To prove Theorem 4.1 we need two technical lemmas. The first one can be found in [4, p.680].

Lemma 7.3 *For every odd positive integer d*

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

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

Lemma 7.4 *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 $b \in \mathcal{R}$* $\frac{\partial^k}{\partial b^k} \int_{H_{\mathbf{e}b}} f(\mathbf{y})d\mathbf{y} = \int_{H_{\mathbf{e}b}} \left(D_{\mathbf{e}}{}^{(k)} f(\mathbf{y})\right) d\mathbf{y}$.

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

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

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

$$\int_{H_{\mathbf{e}b}} D_{\mathbf{e}}{}^{(k)}f(\mathbf{y})d\mathbf{y}.$$

□

Proof of Theorem 4.1. We first prove the theorem for test functions. For $f \in \mathcal{D}(\mathcal{R}^d)$ 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 [19]). 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 [19, 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, $f(\mathbf{x}) = -a_d\int_{S^{d-1}}\int_{\mathcal{R}}\left(\int_{H_{\mathbf{e}b}} D_{\mathbf{e}}{}^{(d)}f(\mathbf{y})d\mathbf{y}\right)\vartheta(\mathbf{e}\cdot\mathbf{x}+b)dbd\mathbf{e}$.

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 $\mathcal{R}^d$ [19, 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 [6, 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 $\mathbf{x}\in\mathcal{R}^d$ $f(\mathbf{x}) = \lim_{i\to\infty}\int_{S^{d-1}}\int_{\mathcal{R}} g_i(\mathbf{x},bfe,b)dbd\mathbf{e} = \int_{S^{d-1}}\int_{\mathcal{R}} g(\mathbf{x},\mathbf{e},b)dbd\mathbf{e} = \int_{S^{d-1}}\int_{\mathcal{R}}\left(\int_{H_{\mathbf{e}b}} D_{\mathbf{e}}{}^{(d)}f(\mathbf{y})d\mathbf{y}\right)\vartheta(bfe\cdot\mathbf{x}+b)dbd\mathbf{e}$ (using again interchangebility of integration and limit for a sequence of functions converging uniformly). □

Proof of Corollary 4.3. For any hyperplane $H\subset\mathcal{R}^d$, $\lambda_{d-1}(X\cap H)\le c_d(X)\lambda_d(X)$, where $c_d(X)$ is the geometric constant that describes the ratio of the largest possible $\lambda_{d-1}(X\cap H)$ divided by the smallest $\lambda_{d-1}(X')$, where X' is a face of X. Ball [3] proved that for every d and for every d-dimensional cube X $c_d(X) = \sqrt{2}$. □

Proof of Lemma 5.1. Let $f\in cl_{\mathcal{L}_p}(conv(\mathcal{E}_d(\vartheta,B,J)))$. Then for every $\varepsilon>0$ $\|f - \sum_{i=1}^k a_i u_i\vartheta(\mathbf{e}_i\cdot\mathbf{x}+b_i)\|_p < \frac{\varepsilon}{2}$ where a_i are coefficients of convex combination and all $|u_i|\le B$. By boundedness of σ, for every $i\in\{1,\dots,k\}$ $\lim_{m\to\infty}\sigma(m(\mathbf{e}_i\cdot\mathbf{x}+b_i)) = \vartheta(\mathbf{e}_i\cdot\mathbf{x}+b_i)$

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 $\|f - \sum_{i=1}^k a_i u_i \sigma(m(\mathbf{v}_i \cdot \mathbf{x} + b_i))\|_p < \|f - \sum_{i=1}^k a_i u_i \vartheta(\mathbf{e}_i \cdot \mathbf{x} + b_i)\|_p + \|\sum_{i=1}^k a_i u_i (\vartheta(\mathbf{e}_i \cdot \mathbf{x} + b_i) - \sigma(m(\mathbf{e}_i \cdot \mathbf{x} + b_i)\|_p < \varepsilon$. $\square$

Bibliography

[1] A. R. Barron: Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information Theory* **39**, 930-945, 1993.

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

[3] K. Ball: Cube slicing in R^d. *Proceedings of AMS*, **99** 1-10, 1986.

[4] R. Courant, D. Hilbert: *Methods of Mathematical Physics, vol.II.* New York: Interscience, 1992.

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

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

[7] R. DeVore, R. Howard, C. A. Micchelli: Optimal nonlinear approximation. *Manuscripta Mathematica* **63**, 469-478, 1989.

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

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

[10] L. K. Jones: 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, 1992.

[11] J. L. Kelley: *General Topology.* Princeton: Van Nostrand, 1955.

[12] V. Kůrková: Kolmogorov's theorem and multilayer neural networks. *Neural Networks* **5**, 501-506, 1992.

[13] E. J. McShane: *Integration.* Princeton: Princeton University Press, 1944.

[14] H. N. Mhaskar, C. A. Micchelli: Approximation by superposition of a sigmoidal function. *Advances in Applied Mathematics* **13**, 350-373, 1992.

[15] J. Park, I. W. Sandberg: Approximation and radial-basis-function networks. *Neural Computation* **5**, 305-316, 1993.

[16] W. Rudin: *Principles of Mathematical Analysis*. New York: McGraw-Hill, 1964.

[17] W. Rudin: *Functional Analysis*. New York: McGraw-Hill, 1973.

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

[19] A. H. Zemanian: *Distribution Theory and Transform Analysis*. New York: Dover, 1987.