16
Dimension-Independent Rates of Approximation by Neural Networks

Věra Kůrková[1]

ABSTRACT To characterize sets of functions that can be approximated by neural networks of various types with dimension-independent rates of approximation we introduce a new norm called variation with respect to a family of functions. We derive its basic properties and give upper estimates for functions satisfying certain integral equations. For a special case, variation with respect to characteristic functions of half-spaces, we give a characterization in terms of orthogonal flows throught layers corresponding to discretized hyperplanes. As a result we describe sets of functions that can be approximated with dimension-independent rates by sigmoidal perceptron networks.

KEY WORDS Approximation of multivariable functions, one-hidden-layer feedforward neural networks, variation with respect to a family of functions, upper bounds on rates of approximation.

16.1 Introduction

Approximation of multivariable functions by feedforward neural networks has been widely studied in recent years. It has been shown that any continuous or $\mathcal{L}_p$ function defined on a compact set has an arbitrarily close approximation by an input/output function of a one-hidden-layer network with either perceptrons or radial-basis-function units with quite general activation functions (see e.g. [1], [2], [3]). In neural network terminology, this approximation capability of a class of neural networks is called the *universal approximation property*. However, there is a price: as the accuracy of approximation increases one may require arbitrarily many computational units.

When neural networks are simulated on classical computers, the number of computational units is a critical limiting factor. Thus one is led to study

[1]This work was partially supported by GA AV ČR, grants A2030602 and A2075606.

M. Kárný et al. (eds.), *Computer-Intensive Methods in Control and Signal Processing*

the dependence of approximation error on the number of computational units which is called *rate of approximation*. Each class of neural networks having the universal aproximation property defines a hierarchy of complexity on function spaces measured by rates of approximation with respect to this class.

Upper estimates on rates of approximation derived from constructive proofs of the universal approximation property suffer by "curse of dimensionality" – they grow exponentially with the number of input units, i.e. with the number d of input variables of the function f to be approximated. A general result by deVore et al. [4] 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 on partial derivatives.

But in some successful applications, functions of hundreds of variables were approximated sufficiently well by neural networks with less than a dozen hidden units (see e.g. [5]). Jones [6] provided insight into properties of such functions. Inspired by projection pursuit methods, he introduced a recursive construction of approximants with "dimension-independent" rates of convergence to elements in convex closures of bounded subsets of a Hilbert space. Together with Barron [7] he proposed to apply it to spaces of functions achievable by one-hidden-layer neural networks. Mhaskar and Micchelli [8], by a different technique, constructed approximants that are finite linear combinations of a given orthonormal basis and also converge at dimension-independent rates. Darken et al. [9] extended these estimates to $\mathcal{L}_p$-spaces for $1 < p < \infty$ and exhibited counterexamples when $p = 1$ or $p = \infty$. Moreover they showed that, under special hypotheses, in these spaces as well dimension-independence is achievable. Also, Barron [10] and Girosi [11] obtained dimension-independent rates of convergence for function spaces equipped with the supremum norm using a theorem of Vapnik on VC dimension (a measure of capacity of sets of functions applicable to statistical learning theory).

The above mentioned theorems impose constraints on the functions to be approximated and on the family of approximating functions. In this paper, we show that such constraints can be mathematically studied in terms of certain norms that are tailored to a given neural network class. In particular, for the class of one-hidden-layer perceptron networks with the Heaviside activation function such a norm was called by Barron [10] *variation with respect to half-spaces* since in the case of functions of one variable it coincides with the notion of total variation. Kůrková et al. [12] gave an upper estimate on variation with respect to half-spaces for smooth functions in terms of orthogonal flows through hyperplanes.

In this paper, we introduce a concept of a *variation with respect to a set of functions* which includes variation with respect to half-spaces as a special case and apply it to sets of functions corresponding to various computational units. We derive basic properties of variation with respect to a set of functions and an elementary lower estimate. For functions satisfy-

ing certain integral equations corresponding to neural networks of a given type with "continuum of hidden units" we give an upper estimate on the variation with respect to the set of functions computable by computational units from a given network class. We show how this upper estimate can be combined with various integral representation theorems to obtain estimates of the approximation error for neural networks with various types of units (e.g. perceptrons having as the activation function the Heaviside function or any sigmoidal or trigonometric function). Further we give a general theorem describing relationship between such variations for certain types of function classes. Derivation of an integral representation corresponding to one type of hidden unit function then allows estimates of approximation error for more general hidden unit functions.

Our main result characterizes variation with respect to half-spaces of a function f of several variables in terms of the supremum of "discretized flows" of f through layers of partitions that in contrast to standard total variation are not restricted to boxes with faces parallel with the coordinate hyperplanes.

The paper is organized as follows. In section 2, we review general tools for obtaining dimension-independent rates of approximation. In section 3, we define variation with respect to a set of functions and derive its basic properties. In section 4, we give upper estimates on variation with respect to half-spaces and in section 5, we extend these estimates to variation with respect to more general activation functions. Section 6 is a brief conclusion.

16.2 Dimension-independent estimates of rates of approximation

Jones [6] estimated rates of approximation of functions from convex closures of bounded subsets of a Hilbert space; see also [7, p.934].

Theorem 16.2.1 (Jones) *Let $\mathcal{F}$ be a normed linear space, with a norm $\|.\|$ induced by an inner product, B a positive real number and $\mathcal{G}$ a subset of $\mathcal{F}$ 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*

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

Darken et al. [9] 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 [10] derived an analogous estimate of the uniform approximation error. Similar estimates were obtained by Girosi [11].

To use Jones' theorem to estimate the number of hidden units in neural networks, one takes $\mathcal{G}$ to be the set of bounded multiples of hidden-unit functions. Convex combinations of n such functions can be computed by a network with n hidden units and one linear output unit. For example, for perceptron networks with an activation function ψ we take $\mathcal{G}$ to be the set $\mathcal{P}_\psi(B) = \{w\psi(\mathbf{v} \cdot \mathbf{x} + b); \mathbf{v} \in \mathcal{R}^d, w, b \in \mathcal{R}, |w| \leq B\}$ for some bound B; for radial-basis-function networks with a radial function ψ we take $\mathcal{B}_\psi(B) = \{w\psi(b\|\mathbf{x} - \mathbf{v}\|); \mathbf{v} \in \mathcal{R}^d, w, b \in \mathcal{R}, |w| \leq B\}$ (where $\mathcal{R}$ denotes the set of real numbers).

Note that only for functions that are in convex closures of bounded subsets of $\mathcal{L}_2$ spaces for a *fixed* bound B rates of approximation guaranteed by Jones' theorem can be called dimension-independent. With an increasing number of variables this condition becomes more and more constraining. Each class of networks defines a measure of complexity on $\mathcal{L}_2$-spaces by determining a nested sequence of sets of functions satisfying such constraints with increasing bounds.

16.3 Variation with respect to a set of functions

To characterize functions in convex closures of bounded sets of functions we introduce a new type of a norm.

Let d be a positive integer and $(\mathcal{F}, \|.\|)$ be a normed linear space of real functions on a subset J of $\mathcal{R}^d$. By cl we denote the closure in the topology induced on $\mathcal{F}$ by the norm $\|.\|$ and by $conv$ the convex hull. For a subset $\mathcal{G}$ of $\mathcal{F}$ containing at least one non-zero function and a positive real number B we denote $\mathcal{G}(B) = \{wg; g \in \mathcal{G}, |w| \leq B\}$. By $\mathcal{R}_+$ is denoted the set of non-negative real numbers.

For a function $f \in \mathcal{F}$ define $V(f, \mathcal{G})$ *variation of* f *with respect to* $\mathcal{G}$ (or $\mathcal{G}$*-variation*) by

$$V(f, \mathcal{G}) = \inf\{B \in \mathcal{R}_+; f \in cl\ conv\ \mathcal{G}(B)\}.$$

Note that the concept of $\mathcal{G}$-variation depends on the choice of a normed linear space $(\mathcal{F}, \|.\|)$, but to simplify the notation we only write $V(f, \mathcal{G})$. Since the topology of uniform convergence is finer than any $\mathcal{L}_p$-topology, all upper estimates of $\mathcal{G}$-variation taken with respect to the supremum norm can be also applied to $\mathcal{G}$-variation with respect to any $\mathcal{L}_p$-norm.

The following proposition shows that the infimum in the definition of the variation with respect to $\mathcal{G}$ is always achieved.

Proposition 16.3.1 *Let d be a positive integer, $J \subseteq \mathcal{R}^d$, $(\mathcal{F}, \|.\|)$ be a normed linear space of functions on J. Then for every $f \in \mathcal{F}$*

$$f \in cl\ conv\ \mathcal{G}(V(f, \mathcal{G})).$$

It is straightforward to verify that $\mathcal{G}$-variation is a norm and to give an elementary lower bound.

Proposition 16.3.2 *For every positive integer d, for every $J \subseteq \mathcal{R}^d$, for every normed linear space $(\mathcal{F}, \|.\|)$ of functions from J to $\mathcal{R}$ and for every subset $\mathcal{G}$ of $\mathcal{F}$*
(i) the set of functions $\mathcal{B}(\mathcal{G}) = \{f \in \mathcal{F}; V(f, \mathcal{G}) < \infty\}$ is a linear subspace of $\mathcal{F}$,
(ii) $V(., \mathcal{G})$ is a norm on the factor space $\mathcal{B}(\mathcal{G})/\sim$, where the equivalence $\sim$ is defined by $f \sim g$ when $\|f - g\| = 0$,
(iii) for every $f \in \mathcal{F}$ $V(f, \mathcal{G}) \geq \|f\| / \sup\{\|g\|; g \in \mathcal{G}\}$.

Thus we can characterize functions that can be approximated with dimension-independent rates by networks containing hidden units that compute functions from $\mathcal{G}$ as functions with $\mathcal{G}$-variation bounded by a fixed bound.

The following complementary theorem gives an upper bound on $\mathcal{G}$-variation for functions that can be represented by an integral equation corresponding metaphorically to a neural network with a continuum of hidden units. It is a reformulation of a theorem proved by Kůrková et al. [12, Theorem 2.2]. By cl_C is denoted the closure with respect to the topology of uniform convergence.

Theorem 16.3.3 *Let d, p be positive integers, $J \subseteq \mathcal{R}^d$ and $f \in \mathcal{C}(J)$ be any function which can be represented as $f(\mathbf{x}) = \int_Y w(\mathbf{y}) g(\mathbf{x}, \mathbf{y}) d\mathbf{y}$ where $Y \subseteq \mathcal{R}^p$, $w \in \mathcal{C}(Y)$ is compactly supported and let $\mathcal{G} = \{g(., \mathbf{y}) : J \to \mathcal{R}; \mathbf{y} \in Y\}$. Then $f \in cl_C\ conv\ \mathcal{G}(B)$, where $B = \int_Y |w(\mathbf{y})| d\mathbf{y}$; that is, $V(f, \mathcal{G}) \leq \int_Y |w(\mathbf{y})| d\mathbf{y}$.*

Thus for functions satisfying the hypotheses of this theorem, the $\mathcal{G}$-variation with respect to the topology of uniform convergence is bounded above by the $\mathcal{L}_1$-norm of the weighting function w.

16.4 Variation with respect to half-spaces

One of the most popular type of a computational unit is a perceptron with sigmoidal activation function. The most simple sigmoidal function is the discontinuous threshold Heaviside function denoted here by ϑ and defined by $\vartheta(x) = 1$ for $x \geq 0$ and $\vartheta(x) = 0$ for $x < 0$. Consider the set $\mathcal{P}_\vartheta$ of functions on J computable by Heaviside perceptrons, i.e. the set of

functions of the form $\{\vartheta(\mathbf{e}\cdot\mathbf{x}+b);\mathbf{e}\in S^{d-1}, b\in\mathcal{R}\}$, where S^{d-1} denotes the unit sphere in $\mathcal{R}^d$. Note that this set is equal to the set of characteristic functions of half-spaces.

The concept of variation with respect to a family of functions is a generalization of the notion of variation with respect to half-spaces introduced by Barron [10] corresponding to $\mathcal{P}_\vartheta$-variation (or in other words variation with respect to characteristic functions of half-spaces) in the space $\mathcal{C}(J)$ of all continuous functions on a subset J of $\mathcal{R}^d$ with the supremum norm and induced topology of uniform convergence.

With Kainen and Kreinovich we characterized in [12] variation with respect to half-spaces for smooth functions using the following integral representation theorem. By $H_{\mathbf{e}b}$ is denoted the cozero hyperplane of the affine function $\mathbf{e}\cdot\mathbf{x}+b$ and $D_{\mathbf{e}}^{(d)}$ denotes the directional derivative of order d in the direction of $\mathbf{e}$. Recall [13] 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 the 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 $D_{\mathbf{e}}^{(k)}f(\mathbf{y}) = D_{\mathbf{e}}(D_{\mathbf{e}}^{(k-1)}f)(\mathbf{y})$. By $\mathcal{C}^d(\mathcal{R}^d)$ is denoted the set of functions on $\mathcal{R}^d$ with continuous partial derivatives of order d.

Theorem 16.4.1 *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 $a_d = -1^{\frac{d-1}{2}}/(2(2\pi)^{d-1})$.

The proof of this theorem uses properties of the Heaviside and delta distributions. The theorem provides for compactly supported functions on $\mathcal{R}^d$ with continuous d-th order partial derivatives (for d odd), a representation by a network with a "continuum of hidden units" with output weights corresponding to flows orthogonal to hyperplanes determined by the input weights and biases.

Combining this integral representation with an extension of Theorem 3.3 that also allows a discontinuous hidden unit function, namely P_ϑ, we derived in [12] the following upper bound on the variation with respect to half-spaces.

Corollary 16.4.2 *For every odd positive integer d and for every compactly supported function $f\in\mathcal{C}^d(\mathcal{R}^d)$*

$$V(f,\mathcal{P}_\vartheta)\le|a_d|\int_{S^{d-1}}\int_{\mathcal{R}}|w_f(\mathbf{e},b)|dbd\mathbf{e},$$

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

Thus for a smooth compactly supported function f its variation with respect to half-spaces is 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 cozero hyperplane determined by a point in the cylinder $S^{d-1} \times \mathcal{R}$ (corresponding to the affine function determined by perceptron parameters: weight vector and bias).

Recall that for a function $f : J \to \mathcal{R}$ where $J = [s,t] \subset \mathcal{R}$ the *total variation* of f is defined as $T(f,J) = \sup\{\sum_{i=1}^{k} |f(x_{i+1} - f(x_i)|; s = x_1 < x_2 \ldots < x_k = t\}$. Note that for $d = 1$ the concept of variation with respect to half-spaces (half-lines) coincides with the notion of total variation. Thus Corollary 4.2 extends a well-known characterization of the total variation of a differentible function f by the equality $T(f,J) = \int_J |f'(x)|dx$.

However, Corollary 4.2 can be only used for smooth functions. To obtain a bound on variation with respect to half-spaces valid for all functions we have to extend the approach based on supremum over subdivisions. Instead of the limit of sums of absolute values of differences that is used in the definition of total variation, to estimate variation with respect to half-spaces we have to consider the limit of sums of absolute values of certain characteristics of the function over lattice layers corresponding to discretized hyperplanes.

First, we introduce some notation. Let $\mathbf{e} \in S^{d-1}$ and $\varepsilon > 0$. Choose a finite partition $\mathcal{U}(\varepsilon)$ of S^{d-1} such that for every $U \in \mathcal{U}(\varepsilon)$ $diam(U) \leq \varepsilon$. For each $U \in \mathcal{U}(\varepsilon)$ choose a unit vector $\mathbf{e}_U \in U$. Denote by $E(\varepsilon) = \{\mathbf{e}_U; U \in \mathcal{U}(\varepsilon)\}$. Let $\mathbf{u}_1(\mathbf{e}), \ldots, \mathbf{u}_d(\mathbf{e})$ be a fixed orthonormal base of $\mathcal{R}^d$ such that $\mathbf{u}_1(\mathbf{e}) = \mathbf{e}$. We call the set $\mathcal{L}(\mathbf{e},\varepsilon) = \{\mathbf{x} \in \mathcal{R}^d; \mathbf{x} = \sum_{j=1}^{d} b_j \varepsilon \mathbf{u}_j(\mathbf{e}), (\forall j = 1,\ldots,d)(b_j \in \mathcal{Z})\}$ (where $\mathcal{Z}$ denotes the set of all integers) the $\mathbf{e}$-ε-lattice. For $b \in \mathcal{Z}$ we call the set $L(\mathbf{e},\varepsilon,b) = \{\mathbf{x} \in \mathcal{R}^d; \mathbf{x} = \sum_{j=1}^{d} b_j \varepsilon \mathbf{u}_j(\mathbf{e}); b_1 = b, (\forall j = 2\ldots,d)(b_j \in \mathcal{Z})\}$ the b-th layer of the lattice $\mathcal{L}(\mathbf{e},\varepsilon)$. Such layer approximates the hyperplane $H_{\mathbf{e},b}$. For $J \subseteq \mathcal{R}^d$ let $\mathcal{L}(\mathbf{e},\varepsilon,J) = \mathcal{L}(\mathbf{e},\varepsilon) \cap J$, $L(\mathbf{e},\varepsilon,b,J) = L(\mathbf{e},\varepsilon,b) \cap J$ and $K(\mathbf{e},\varepsilon,J) = \{b \in \mathcal{Z}; L(\mathbf{e},\varepsilon,b,J) \neq \emptyset\}$.

Define a mapping $\Delta f : J \times S^{d-1} \times \mathcal{R}_+ \to \mathcal{R}$ by

$$\Delta f(\mathbf{x},\mathbf{e},\varepsilon) = \sum_{j=0}^{d} (-1)^{d-j} \binom{d}{j} f(\mathbf{x} + j\varepsilon\mathbf{e}).$$

Note that $\lim_{\varepsilon \to 0} \frac{\Delta f(\mathbf{x},\mathbf{e},\varepsilon)}{\varepsilon^d} = D_{\mathbf{e}}^{(d)}(\mathbf{x})$. Thus when both $\mathbf{e}$ and ε are fixed, the sum of $\Delta f(\mathbf{x},\mathbf{e},\varepsilon)$ over the layer $L(\mathbf{e},\varepsilon,b,J)$ approximates an orthogonal flow of order d of f through the hyperplane $H_{\mathbf{e}b}$. Denote this sum by $W_f(\mathbf{e},\varepsilon,b,J) = \sum_{\mathbf{x} \in L(\mathbf{e},\varepsilon,b,J)} \Delta f(\mathbf{x},\mathbf{e},\varepsilon)$. Thus $W_f(\mathbf{e},\varepsilon,b,J)$ approximates

For any function $f : J \to \mathcal{R}$ define

$$V^*(f,\mathcal{P}_\vartheta) = \lim_{\varepsilon \to 0} \sum_{\mathbf{e} \in E(\varepsilon)} \sum_{b \in K(\mathbf{e},\varepsilon,J)} |W_f(\mathbf{e},\varepsilon,b,J)|.$$

Approximating integrals by sums and directional derivatives by differences

$\Delta f(\mathbf{x}, \mathbf{e}, \varepsilon)$ we obtain the following theorem.

Theorem 16.4.3 *For every odd positive integer d there exists a constant c_d such that for every compact $J \subset \mathcal{R}^d$ and for every $f \in \mathcal{C}^d(J)$*

$$V^*(f, \mathcal{P}_\vartheta) = c_d \int_{S^{d-1}} \int_{\mathcal{R}} \left| \int_{H_{\mathbf{e}b} \cap J} D_{\mathbf{e}}^{(d)} f(\mathbf{y}) d\mathbf{y} \right| db d\mathbf{e}.$$

Thus variation with respect to half-spaces can be estimated by computing absolute values of sums of difference Δf over layers approximating hyperplanes.

16.5 Variation with respect to sets of functions computable by perceptrons with various activation functions

From upper bounds on variation with respect to half-spaces we can derive estimates of $\mathcal{P}_\psi$-variation for more general activation functions ψ, namely for all bounded sigmoidals. Recall that a function $\sigma : \mathcal{R} \to [0, 1]$ is called *sigmoidal* if $\lim_{x \to -\infty} \sigma(x) = 0$ and $\lim_{x \to \infty} \sigma(x) = 1$.

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 by approximating uniformly any continuous function with a "staircase-like" function that is a linear combination of translations of the Heaviside function. Define $T(\psi, \mathcal{R}) = \sup\{T(\psi, I); I = [s, t], s, t \in \mathcal{R}\}$.

Proposition 16.5.1 *Let $\sigma : \mathcal{R} \to \mathcal{R}$ be a bounded sigmoidal function. Then for every positive integer d, for every compact $J \subset \mathcal{R}^d$, for every $p \in [1, \infty)$ and for every $f \in \mathcal{L}_p(J)$*

$$V(f, P_\sigma) \leq V(f, P_\vartheta)$$

with respect to the $\mathcal{L}_p$-norm.

Proposition 16.5.2 *Let $\psi : \mathcal{R} \to \mathcal{R}$ be a function such that $T(\psi, \mathcal{R})$ is finite. Then for every positive integer d, for every compact $J \subset \mathcal{R}^d$ and for every $f \in \mathcal{C}(J)$ or $f \in \mathcal{L}_p(J)$ for $p \in [1, \infty)$*

$$V(f, P_\vartheta) \leq T(\psi, \mathcal{R}) V(f, P_\psi)$$

with respect to the supremum or $\mathcal{L}_p$-norm, respectively.

Since the total variation $T(\sigma)$ of any continuous non-decreasing sigmoidal function σ over any interval is bounded by 1, variation with respect to

half-spaces remains unchanged if the Heaviside function is replaced by this sigmoidal function, i.e. $\mathcal{P}_\vartheta$-variation = $\mathcal{P}_\sigma$-variation. Moreover, Proposition 5.2 shows that any integral representation corresponding to a network with a continuum of ψ-perceptrons can be used to estimate variation with respect to half-spaces. This result generalizes a method used by Barron [7] who combined Fourier integral representation with an approximation of the cosine activation function by sigmoidals.

The theory of $\mathcal{G}$-variation can also be used to derive upper bounds on variation with respect to radial-basis-functions generalizing a method proposed by Girosi and Anzellotti [14] based on a convolution with a radial function.

16.6 Conclusion

A result of DeVore et al. [4] shows that an upper bound on partial derivatives is not sufficient to guarantee dimension-independent rates of approximation by one-hidden-layer neural networks. We have proposed a new concept of a norm tailored to a given class of neural networks in such a way that any multivariable function f with this norm bounded by a fixed bound B can be approximated by networks from this class with n hidden units within the error $\sqrt{\frac{c}{n}}$ where $c > B^2 - \|f\|^2$.

16.7 References

[1] H.N. Mhaskar and C.A. Micchelli, "Approximation by superposition of sigmoidal and radial basis functions", *Advances in Applied Mathematics*, vol. 13, pp. 350–373, 1992.

[2] M. Leshno, V. Lin, A. Pinkus, and S. Schocken, "Multilayer feedforward networks with a non-polynomial activation function can approximate any function", *Neural Networks*, vol. 6, pp. 861–867, 1993.

[3] J. Park and I.W. Sandberg, "Approximation and radial-basis-function networks", *Neural Computation*, vol. 5, pp. 305–316, 1993.

[4] R. DeVore, R. Howard, and C. Micchelli, "Optimal nonlinear approximation", *Manuscripta Mathematica*, vol. 63, pp. 469–478, 1989.

[5] T.J. Sejnowski and C. Rosenberg, "Parallel networks that learn to pronounce English text", *Complex Systems*, vol. 1, pp. 145–168, 1987.

[6] L.K. Jones, "A simple lemma on greedy approximation in Hilbert space and convergence rates for projection pursuit regression and neural network training", *Annals of Statistics*, vol. 20, pp. 608–613, 1992.

[7] A.R. Barron, "Universal approximation bounds for superposition of a sigmoidal function", *IEEE Transactions on Information Theory*, vol. 39, no. 3, pp. 930–945, 1993.

[8] H.N. Mhaskar and C. A. Micchelli, "Dimension-independent bounds on the degree of approximation by neural networks", *IBM Journal of Research and Development*, vol. 38, pp. 277–284, 1994.

[9] C. Darken, M. Donahue, L. Gurvits, and E. Sontag, "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. ACM, New York, 1993.

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

[11] F. Girosi, "Approximation error bounds that use VC-bounds", in *Proceedings of ICANN'96*, pp. 295–302. EC & Cie, Paris, 1995.

[12] V. Kůrková, P.C. Kainen, and V. Kreinovich, "Estimates of the number of hidden units and variation with respect to half-spaces", *Neural Networks*, in press.

[13] C.H. Edwards, *Advanced calculus of several variables*, Dover, New York, 1994.

[14] 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. Chapman & Hall, London, 1993.