g and direction

›sition theorem,
), 350–373.

ion, application
51.

problems from
1996), 269–287.

ing radial basis
7.

›erposition the-

es of functions,

ariate time se-

Real Variable,

Chapter 14

Rates of Approximation in a Feedforward Network Depend on the Type of Computational Unit

1 Introduction

The approximation capabilities of feedforward neural networks with a single hidden layer and with various activation functions has been widely studied ([19], [8], [1], [2], [13]). Mhaskar and Micchelli have shown in [22] that a network using any non-polynomial locally Riemann integrable activation can approximate any continuous function of any number of variables on a compact set to any desired degree of accuracy (i.e. it has the universal approximation property). This important result has advanced the investigation of the complexity problem: If one needs to approximate a function from a known class of functions within a prescribed accuracy, how many neurons are necessary to realize this approximation for all functions in the class? De Vore et al. ([3]) proved the following result: if one approximates continuously a class of functions of d variables with bounded partial derivatives on a compacta, in order to accomplish the order of approximation $\mathcal{O}(\frac{1}{n})$, it is necessary to use at least $\mathcal{O}(n^d)$ number of neurons, regardless of the activation function. In other words, when the class of functions being approximated is defined in terms of bounds on the partial derivatives, a dimension independent bound for the degree of approximation is not possible. Kůrková studied the relationship between approximation rates of one-hidden-layer neural networks with different types of hidden units. She showed in [14] that no sufficiently large class of functions can be approximated by one-hidden-layer networks with another type of unit than Heaviside perceptrons with a rate of approximation related to the rate of

approximation by perceptron networks.
We present an overview of some known rates of approximation of multivariable functions by feedforward neural networks. The paper is organized as follows: In chapter 2.1 we present the approximation rate for networks with spline activation functions by Mhaskar ([19]). Our rate for kernel basis function networks and radial basis function networks is in chapter 2.2. Chapter 2.3 examines networks with perceptron-type computational units: Barron's rate for sigmoidal networks ([1]), our rates for Heaviside activation functions and the class of real valued activation functions ([6]) and further Hornik's result on approximation rates on networks with activation functions with bounded partial derivatives ([9]). The last two sections (2.3.3 and 2.3.4) consist in Mhaskar and Micchelli's results ([22]) on approximation rates for networks with trigonometric polynomials and sigmoidals of order k and a general perceptron activation function. Chapter 3 discusses the approximation error of all the mentioned networks, their number of hidden units and the characteristics known about the function to be approximated.

2 Feedforward Networks with Various Computational Units

By $\mathcal{R}$, $\mathcal{N}$ and $\mathcal{Z}$ we denote the set of real numbers, positive integers, and integers respectively; For a bounded function $f : \mathcal{R}^d \rightarrow \mathcal{R}$ the uniform norm is defined by $\|f\|_\infty = \sup_{\mathbf{x} \in \mathcal{R}^d} |f(\mathbf{x})|$ and $\|f\|_A = \sup_{\mathbf{x} \in A} |f(\mathbf{x})|$ for some $A \subset \mathcal{R}^d$. Let A be a compact subset of $\mathcal{R}^d$. Let $[\mathrm{a}, \mathrm{b}] \subset \mathcal{R}^d$. Denote $\mathcal{C}(A)$ the space of continuous functions on $A \subset \mathcal{R}^d$ with the uniform norm and corresponding topology. In the paper, we deal only with feed-forward networks with one hidden layer.

2.1 Approximation by the Network with Spline Computational Units

Let $A = \prod_{j=1}^{d}[a_j, b_j]$. The *modulus of smoothness* $\omega_m^d(f, A)$ of a function $f : A \rightarrow \mathcal{R}$ is defined by
$\omega_m^d(f, A) = \inf \max_{\mathbf{x} \in A} |f(\mathbf{x}) - P(\mathbf{x})|$, where the infimum is taken over all polynomials P of degree at most $m-1$ in each of its d variables. *Modulus of $\delta-$ smoothness* is defined by $\omega_m^d(f, \delta, [0,1]^d) = \sup\{\omega_m^d(f, A) : A$ subcube of $[0,1]^d$, diam(A) $\leq \delta\}$.
Mhaskar dealt in [19] with the estimation of the error of approximation by multivariable spline functions with fixed knots. Let $d \geq 2$ be the number of input variables. The *tensor product quasi-interpolatory spline operator* is defined by $Q_n^d(f, \mathbf{x}) = \sum_{\mathbf{i}} \lambda_{\mathbf{i}} N_m^d(n\mathbf{x} - \mathbf{i})$, where $\mathbf{x} = (x_1, x_2, \ldots, x_d) \in \mathcal{R}^d$, $\mathbf{i} = (i_1, \ldots, i_d)$ and the *tensor product (cardinal) B-spline of order m*
$N_m^d(\mathbf{x}) = \prod_{j=1}^{d} N_m(x_j)$. Let $I = [0,1]^d$. We say that *interpolating points are properly spaced* if there are some interpolating ~~points~~ between any two

interpolating ~~points~~ knots. The corresponding neural network has the hidden units of the form $N_m^d(n\mathbf{x} - \mathbf{i})$.

Theorem 2.1 ([19]) *If $f : I \rightarrow \mathcal{R}$ is continuous and $m, n \geq 1$ are integers, then there exists a spline Q_n^d of order k with $(n+1)^d$ nodes so that if the interpolating points are properly spaced then*

$$\|f - Q_n^d\|_I \leq c\omega_m^d(f, \frac{1}{n}, I^d),$$

where c is a positive constant depending only on m and d.

The approximation error depends indirectly on the number of units in the hidden layer of the network. However, the modulus of smoothness does not provide an explicit formula showing the complexity of the rate and the number of nodes of the spline is exponential in d.

Williamson and Bartlett [27] studied the relationship between spline approximation, approximation with rational functions, and feedforward functions. They achieved a lower bound of approximation of rational functions of one variable by a feedforward network demonstrating that no great benefit in terms of the degree of approximation is to be obtained by using multiple hidden networks. Williamson and Bartlett investigated the rate of approximation of functions from $\mathcal{L}_p$ with a feedforward network with spline activation functions and n knots. They achieved the rate of complexity $\mathcal{O}(n^{-\alpha})$ where α is the number of continuous derivatives of f.

2.2 Approximation by the Network with a Radial and Kernel Basis Units

The following can be found in more detail in [5]. Let $f, g : \mathcal{R}^d \rightarrow \mathcal{R}$ are given functions and by $f * g = \int\limits_{\mathcal{R}^d} f(\mathbf{x})g(\mathbf{x} - \mathbf{y})dy$ we denote a *convolution* of f, g.

Denote $[\mathbf{a}, \mathbf{b}] = \prod_{j=1}^d [a_j, b_j]$ a given cube in $\mathcal{R}^d$. Define $U[\mathbf{a}, \mathbf{b}] = \{\mathbf{x};$ either $x_i = a_i$ or $x_i = b_i\}$ and let $\tau(\mathbf{x})$ denote the number of i so that $x_i = a_i$, where $\mathbf{x} = (x_1, \ldots, x_d)$. Denote $f|_{[\mathbf{a},\mathbf{b}]} = |\sum_{\mathbf{x} \in U[\mathbf{a},\mathbf{b}]} (-1)^{\tau(\mathbf{x})} f(\mathbf{x})|$. *Total variation* of f on $[\mathbf{a}, \mathbf{b}]$ is defined by $V(f) = V(f)|_{[\mathbf{a},\mathbf{b}]} = \sup_P \{\sum_{j=1}^k f|_{J_j}\}$, where $P = \{J_1, \ldots, J_k\}$ is a partition of $[\mathbf{a}, \mathbf{b}]$ so that $[\mathbf{a}, \mathbf{b}] = \cup_{j=1}^k J_j$ and $int(J_j) \cap int(J_l) = 0$ for all $j \neq l, = 1, \ldots, k$. ($int(A)$ denotes the interior of set A.) We say that *f is of bounded total variation* if $V(f)$ is finite.

A *radial basis function (RBF) unit* with d inputs is a computational unit computing a function of the form $\phi(\| \mathbf{x} - \mathbf{c} \| /b)$, where $\phi : \mathcal{R} \rightarrow \mathcal{R}$ is an even (radial) function, $\| . \|$ is a norm on $\mathcal{R}^d$, and $\mathbf{x}, \mathbf{c} \in \mathcal{R}^d$, $b \in \mathcal{R}$, $b > 0$. A *radial basis function (RBF) network* is a neural network with a single linear output unit, one hidden layer with RBF units with the same radial function ϕ and the same norm $\| . \|$ on $\mathcal{R}^d$, and d inputs. The most frequent radial function used in application is the Gaussian $\gamma(t) = \exp(-t^2)$.

Kernel basis function (KBF) unit with d inputs computes a function $\mathcal{R}^d \rightarrow \mathcal{R}$

of the form $k_n(\| \mathbf{x}-\mathbf{c} \|)$, where $\{k_n : \mathcal{R} \to \mathcal{R}\}$ is a sequence of functions, $\| \, . \, \|$ is a norm on $\mathcal{R}^d$, and $\mathbf{c} \in \mathcal{R}^d$, $n \in \mathcal{N}$ are parameters. A *kernel basis function (KBF) network* is a neural network with a single linear output unit, one hidden layer with KBF units with the same sequence of functions $\{k_n,\ n \in \mathcal{N}\}$ and the same norm $\| \, . \, \|$ on $\mathcal{R}^d$, and d inputs. By $\mathcal{K}_u(\{k_n,\ n \in \mathcal{N}\},\ \| \, . \, \|)$ we denote the set of functions computable by KBF networks with $\{k_n,\ n \in \mathcal{N}\}$ and $\| \, . \, \|$ with any number of hidden units and uniform k_n for all hidden units. In [13], we obtained the universal approximation property for the class $\mathcal{K}_u(\{k_n,\ n \in \mathcal{N}\},\ \| \, . \, \|)$ of continuous kernel functions on $\mathcal{C}(I^d)$ and every norm $\| \, . \, \|$ on $\mathcal{R}^d$ satisfying for every $n \in \mathcal{N}$ and every $\mathbf{x} \in \mathcal{R}^d$ $\int\limits_{\mathcal{R}^d} k_n(\| \mathbf{x}-\mathbf{y} \|)d\mathbf{y} = 1$ and for every $\delta > 0$ and every $\mathbf{x} \in \mathcal{R}^d$ $\lim\limits_{n \to \infty} \int\limits_{J_\delta(\mathbf{x})} k_n(\| \mathbf{x}-\mathbf{y} \|)d\mathbf{y} = 0$, where $J_\delta(\mathbf{x}) = \{\mathbf{y} |\ \mathbf{y} \in \mathcal{R}^d,\ \| \mathbf{x}-\mathbf{y} \| \geq \ \delta\}$.
The classical kernels such as the Féjer kernel, the Dirichlet kernel, the Jackson kernel, the Abel-Poisson kernel, the Weierstrass kernel, and the Landau kernel satisfy this assumption and thus KBF networks with any of these kernels are powerful enough to approximate continuous functions.

Theorem 2.2 **([5])** *Let $d \geq 0$ be a given integer. Let $f : \mathcal{R}^d \to \mathcal{R}$ be a continuous function, k_n a kernel function, $I = [0,1]^d$. Let $f * k_n$ be of a bounded total variation. Then for every $m \in \mathcal{N}$ there exists a KBF network with m hidden units computing a function $g \in \mathcal{K}_u(\{k_n\}, \|.\|)$ so that*

$$\|f-g\|_I \leq \epsilon_I(f,h) + \frac{d}{m}V(h),$$

where $h(\mathbf{x}) = f * k_n(\mathbf{x}) = \int_H f(\mathbf{t})k_n(\| \mathbf{x}-\mathbf{t} \|)d\mathbf{t}$ and $\epsilon_I(f,h) = \|f-h\|_I$.

For some of the above mentioned convolution kernels upper bounds on $\epsilon_I(\cdot,\cdot)$ are known. For example, we consider Jackson kernel on the interval
$P = [-\pi,\pi]^d$. Define the following operator:
$\int_P f(\mathbf{t})L_n(\mathbf{x}-\mathbf{t})d\mathbf{t} = \int_P f(\mathbf{x}+\mathbf{t})L_n(\mathbf{t})d\mathbf{t}$, where L_n is the *Jackson kernel* $L_n(\mathbf{t}) = \lambda_n^{-1}\left(\frac{\sin\|n.\mathbf{t}/2\|}{n\sin\|\mathbf{t}/2\|}\right)^4$, $\int_P L_n(\mathbf{t})d\mathbf{t} = 1$, and the last relation defines λ_n. It is convenient to normalize the operator in such a way as to obtain a trigonometric polynomial of degree n. For this purpose, we put $K_n(\mathbf{t}) = L_r(\mathbf{t})$, $r = [\frac{n}{2}]+1$. The operator $J_n(\mathbf{x}) = J_n(f,\mathbf{x}) = \int_P f(\mathbf{x}+\mathbf{t})K_n(\mathbf{t})d\mathbf{t}$ is called the (multivariable) *Jackson operator.* $\omega_P(f,\delta) = \max_{\mathbf{x},\mathbf{t},\|\mathbf{t}\|<\delta} |f(\mathbf{x}+\mathbf{t})-f(\mathbf{x})|$ is called *modulus of continuity of f.*

Theorem 2.3 **([5])** *There exists a constant $M \geq 0$ so that for every $f \in \mathcal{C}(\mathcal{P})$, for every n and for every $m \in \mathcal{N}$ and a function g computable by a KBF network with m with Jackson kernel hidden units J_n such that*

$$\|f-g\|_P \ \leq \ M\omega_P(f,\frac{1}{r}) + \frac{d}{m}V(J_n),$$

where $r = [\frac{n}{2}]+1$.

2.3 Approximation by the Network with Perceptron-Type Computational Units

2.3.1 Sigmoidal and Heaviside Activation Functions

Let $\sigma : \mathcal{R} \rightarrow \mathcal{R}$ be a bounded measurable function on the real line for which $\lim_{x\rightarrow-\infty} \sigma(x) = 0$, $\lim_{x\rightarrow\infty} \sigma(x) = 1$. We call this function *sigmoidal.* Feedforward neural network models with one layer of sigmoidal units implement functions on $\mathcal{R}^d$ of the form

$$f_n(\mathbf{x}) = \sum_{k=1}^{n} c_k\sigma(\mathbf{a}_k.\mathbf{x} + b_k) + c_0 \tag{1}$$

parametrized by $\mathbf{a}_k \in \mathcal{R}^d$ and $b_k, c_k \in \mathcal{R}$, where $\mathbf{a}_k.\mathbf{x}$ denotes the inner product of vectors in $\mathcal{R}^d$.
Let $\mathcal{X}$ be a real vector space with a norm $||.||_2$ which is generated by an inner product $f \cdot g$ for any two functions $f, g \in \mathcal{X}$. *cl conv* $\mathcal{G}$ means the closure of the convex hull of $\mathcal{G}$, where $\mathcal{G}$ is a subset of $\mathcal{X}$. The closure is taken with respect to the topology generated by the norm $||.||_2$ ($||.||_2 = \sqrt{f \cdot f}$). $\mathcal{N}$ denotes the set of positive integers.

Theorem 2.4 (Jones-Barron) *Let $\mathcal{X}$ be a real vector space with a norm $||.||_2$ generated by an inner product on $\mathcal{X}$, B be a positive real number and $\mathcal{G}$ be a subset of $\mathcal{X}$ such that for every $g \in \mathcal{G}$ $||g||_2 \leq B$. Then for every $f \in$ cl conv $\mathcal{G}$, for every real number c such that $c > B^2 - ||f||_2^2$ and for every $n \in \mathcal{N}$, there exists f_n which is a convex combination of n elements of $\mathcal{G}$ such that $||f - f_n||_2 \leq \sqrt{\frac{c}{n}}$.*

Here we mention a result by Barron ([1]) on approximation by sigmoidal functions which is a corollary of Theorem 2.4. Barron showed that it is possible to approximate any function satisfying certain conditions on its Fourier transform within an $\mathcal{L}_2$ error of $\mathcal{O}(\frac{1}{\sqrt{n}})$ using a feedforward neural network with one hidden layer comprising of n neurons, each with a sigmoidal activation function. The approximation error is measured by the integrated squared error with respect to an arbitrary probability measure μ on the ball $B_r = \{\mathbf{x} : |\mathbf{x}| \leq r\}$ of radius $r \geq 0$. The function σ is an arbitrary fixed sigmoidal function.
Consider the class of functions f on $\mathcal{R}^d$ for which there is a Fourier representation of the form $f(\mathbf{x}) = \int_{\mathcal{R}^d} e^{i\omega.\mathbf{x}}\hat{f}(\omega)d\omega$ for some complex-valued function $\hat{f}(\omega)$ for which $\omega\hat{f}(\omega)$ is integrable, and define $C_f = \int_{\mathcal{R}^d} |\omega||\hat{f}(\omega)|d\omega$, where $|\omega| = (\omega.\omega)^{1/2}$. For each $C > 0$, let Γ_C be the set of functions f such that $C_f \leq C$. Let $||g||_{\mathcal{L}_2(B_r)} = \sqrt{\int_{B_r} g(\omega)^2 d\omega}$ denotes the $\mathcal{L}_2$ norm of g on B_r.

Theorem 2.5 ([1]) *For every function f with C_f finite, and every $n \geq 1$, there exists a linear combination of sigmoidal functions $f_n(\mathbf{x})$ of the form* (1), *so that*

$$||f - f_n||_{\mathcal{L}_2(B_r)} \leq \frac{2rC_f}{\sqrt{n}}.$$

For functions in Γ_C, the coefficients of the linear combination in (1) *may be restricted to satisfy $\sum_{k=1}^{n} |c_k| \leq 2rC$ and $c_0 = f(0)$.*

Kûrková et al. achieved an $\mathcal{L}_2$ error rate of the order $\mathcal{O}(\frac{1}{\sqrt{n}})$ by one hidden layer networks with n sigmoidals in [12]. They use an integral representation of smooth functions of d variables and express the rate of approximation in terms of the variation with respect to half spaces, but they do not use a Fourier representation of a function to be approximated.

In [6] we investigated subclasses of the so called *real-valued boolean functions,* i.e. functions $f : \{0,1\}^d \rightarrow \mathcal{R}$. Real-valued functions with multiple Boolean variables are exactly representable by one-hidden-layer Heaviside perceptron networks with an exponential number of hidden units. We derived upper bounds on the approximation error of the form $\frac{c}{\sqrt{n}}$ where c depends on certain norms of the function being approximated and n is the number of hidden units. We gave examples of functions for which these norms grow polynomially and exponentially with increasing input dimension.

If f is a linear but not convex combination of functions from $\mathcal{G}$, then $\mathcal{G}$ in Theorem 2.4 can be replaced by real multiples of functions from $\mathcal{G}$ bounded by a constant. This leads to the term of variation, first introduced by Barron for a set of characteristic functions of half-spaces. For a normed vector space $(\mathcal{X}, ||.||)$ consisting of real functions on $J \subset \mathcal{R}^d$ for an integer d, let the *variation of a function $f \in \mathcal{X}$ with respect to a subset $\mathcal{G}$* of $\mathcal{X}$ be $V(f,\mathcal{G}) = \inf\{B \geq 0; f \in \mathit{cl\ conv}\ \mathcal{G}(B)\}$, where the closure is taken with respect to the topology generated by the norm $||.||$ and $\mathcal{G}(B) = \{wg; g \in \mathcal{G}, w \in \mathcal{R}, |w| \leq B\}$. This definition was introduced by Kûrková in [16] and is a generalization of Barron's definition of variation with respect to half-spaces. The following theorem is a corollary of the Jones-Barron theorem formulated by means of variation. Since in our applications set $\mathcal{G}$ is finite, we use a stronger formulation of the theorem for compact sets $\mathcal{G}$.

Theorem 2.6 **([6])** *Let $(\mathcal{X}, ||.||)$ be a real vector space with the norm $||.||$ generated by an inner product and $\mathcal{G}$ be a compact subset of $\mathcal{X}$. Then for every $f \in \mathcal{X}$ such that $V(f,\mathcal{G}) < \infty$ and for every $n = 1, \ldots, card\ \mathcal{G}$ there exists f_n which is a linear combination of n elements of $\mathcal{G}$ such that $||f - f_n||_2 \leq \sqrt{\frac{B^2 - ||f||_2^2}{n}}$, where $B = V(f,\mathcal{G}) \sup_{g \in \mathcal{G}} ||g||_2$.*

If $\mathcal{G}$ is an orthonormal basis, we can prove a stronger estimate improving the Mhaskar and Micchelli's result from [24] by a factor of two. For any orthonormal basis let A of $\mathcal{X}$ denote by $||.||_{1,A}$ the l_1-norm with respect to A, i.e. for every $f \in \mathcal{X}$ $\;||f||_{1,A} = \sum_{g \in A} |f \cdot g|$.

Theorem 2.7 **([6])** *Let $\mathcal{X}$ be a finite dimensional real vector space with a norm $||.||_2$ generated by an inner product and let A be its orthonormal basis. Then for every $f \in \mathcal{X}$ and for every $n = 1, \ldots, \dim\ \mathcal{X}$ there exists f_n which is a linear combination of n elements of A such that $||f - f_n||_2 \leq \frac{||f||_{1,A}}{2\sqrt{n}}$.*

If $||f||_2$ is also known, then the bound from Theorem 2.7 can be improved.

Theorem 2.8 ([6]) *Let $\mathcal{X}$ be a finite dimensional real vector space with an inner product, let A be its orthonormal basis, let $f \in \mathcal{X}$ and let $1 \leq n \leq \dim \mathcal{X}$. Then, there exists a function g expressible as a linear combination of at most n functions from A satisfying*

$$||f-g||_2 \leq \frac{||f||^2_{1,A} - ||f||^2_2}{2||f||_{1,A}\sqrt{n-1}}.$$

If both $||f||_{1,A}$ and $||f||_2$ are known, then Theorem 2.8 yields a good bound only if $4n \geq ||f||^2_{1,A}/||f||^2_2$. Otherwise, the trivial bound $||f||_2$ for the error of the approximation by the zero function is better. In fact, these two bounds together, i.e. the minimum of $||f||_2$ and the bound from Theorem 2.8, yield a bound that differs from the best possible bound based only on $||f||_{1,A}$ and $||f||_2$ by a constant factor.

The linear space of all real functions of d Boolean variables (where d is a positive integer) is denoted by $\mathcal{F}(\{0,1\}^d)$. For any $f, g \in \mathcal{F}(\{0,1\}^d)$, the standard Euclidean inner product is $f \cdot g = \sum\limits_{x\in\{0,1\}^d} f(x)g(x)$. Here we study representations and approximations of functions in $\mathcal{F}(\{0,1\}^d)$ by functions computable by networks with one linear output unit and one hidden layer with the *Heaviside function* ϑ defined by $\vartheta(t) = 0$ for $t < 0$ and $\vartheta(t) = 1$ for $t \geq 0$. The set of functions expressible by such networks with a bounded number of hidden units can be denoted by:
$\mathcal{P}_d(n) = \{f \in \mathcal{F}(\{0,1\}^d); f(x) = \sum_{i=1}^n w_i\vartheta(v_i \cdot x + b_i);\ w_i, b_i \in \mathcal{R}, v_i \in \mathcal{R}^d\}$.

Denote by $E = \{e_u; u \in \{0,1\}^d\}$ the *Euclidean orthonormal basis* of $\mathcal{F}(\{0,1\}^d)$, i.e. $e_u(u) = 1$ and $e_u(x) = 0$ for $x \neq u$. It is easy to verify that e_u can be computed by one Heaviside perceptron, i.e. $e_u \in \mathcal{P}_d(1)$. Together with the representation of any function $f \in \mathcal{F}(\{0,1\}^d)$ as $f(x) = \sum_{u\in\{0,1\}^d} f(u)e_u$, this yields that $\mathcal{F}(0,1^d) = \mathcal{P}_d(2^d)$.

A representation of a different type can be obtained from the *orthonormal Fourier basis* $F = \{\frac{1}{\sqrt{2^d}}\cos(\pi u \cdot x); u \in \{0,1\}^d\}$ of $\mathcal{F}(\{0,1\}^d)$. Since in our context both x and u are Boolean vectors, we have $\cos(\pi u \cdot x) = (-1)^{u\cdot x}$. Thus every function $f \in \mathcal{F}(\{0,1\}^d)$ can be represented as
$f(x) = \frac{1}{\sqrt{2^d}} \sum\limits_{u\in\{0,1\}^d} \tilde{f}(u)(-1)^{u\cdot x}$, where the Fourier coefficients $\tilde{f}(u)$ are given by the formula
$\tilde{f}(u) = \frac{1}{\sqrt{2^d}} \sum\limits_{x\in\{0,1\}^d} f(x)(-1)^{u\cdot x}$. Note that for any $f \in \mathcal{F}(\{0,1\}^d)$ $||f||_{1,F} = ||\tilde{f}||_1 = \sum_{u\in\{0,1\}^d} |\tilde{f}(u)|$. Furthermore, all functions from the Fourier basis are computable by Heaviside perceptron networks. In contrast to the Euclidean basis, where one hidden unit was sufficient for one basis function, $d+1$ hidden units are needed for the members in the Fourier basis. Thus we have a representation of any $f \in \mathcal{F}(\{0,1\}^d)$ as an element of $\mathcal{P}_d((d+1)2^d)$ if we replace $(-1)^{u\cdot x}$ by ϑ in the Fourier representation.

Note that all norms on $\mathcal{R}^{2^d}$ are topologically equivalent, in particular for every $f \in \mathcal{F}(\{0,1\}^d)$ $||f||_2 \leq ||f||_1 \leq \sqrt{2^d}||f||_2$ and $||f||_2 \leq ||\tilde{f}||_1 \leq \sqrt{2^d}||f||_2$. Since each of these inequalities is tight, the differences between the norms may be exponential in dimension d.

Theorem 2.9 **([6])** *Let d be a positive integer and $f \in \mathcal{F}(\{0,1\}^d)$ and $n \geq 2$. Then*
(i) there exists a function $f_n \in \mathcal{P}_d(n)$ such that

$$\|f - f_n\|_2 \leq \frac{\|f\|_1^2 - \|f\|_2^2}{2\|f\|_1\sqrt{n-1}};$$

(ii) there exists a function $f_n \in \mathcal{P}_d((d+1)n)$ such that

$$\|f - f_n\|_2 \leq \frac{\|\tilde{f}\|_1^2 - \|f\|_2^2}{2\|\tilde{f}\|_1\sqrt{n-1}}.$$

We present two examples of functions for which the upper bounds on the approximation error from Theorem 2.9 yield a feasible approximation. The bounds are compared with the approximation error by general half-spaces.
It can be shown that Theorem 2.9 implies $\|f - f_n\|_2 \ll \|f\|_2$ for a feasible n only if $\min\{\|f\|_1, \|\tilde{f}\|_1\}$ is not much larger than $\|f\|_2$. In fact, if the equality holds, the bound implies an exact representation. For every orthonormal basis A, $\|f\|_{1,A} = \|f\|_2$ is satisfied if and only if f is a multiple of just one of the elements of the basis. The functions for which this situation occurs for the Fourier basis are functions represented by $f(x) = (-1)^{u \cdot x}$, $u \in \{0,1\}^d$. These functions correspond to the Boolean functions called parity functions, since the value of $(-1)^{u \cdot x}$ depends on the parity of the sum $\sum_{i \in I} x_i$, where $I = \{i; u_i = 1\}$.
Let f be a function represented by a decision tree of polynomial size and let the ratio $(\max_x |f(x)|)/(\min_x |f(x)|)$ be defined and polynomially bounded. Using the method of [17], it can be proven that $\|\tilde{f}\|_1/\|f\|_2$ is polynomially bounded. Using Theorem 2.9, this implies that f can be approximated by a polynomial number of hidden units.
We now turn to the functions for which our two bases do not yield a good approximation. A function from $\mathcal{F}(\{0,1\}^d)$ is called *bent*, if for every $x, u \in \{0,1\}^d$ $|f(x)| = 1$ and $|\tilde{f}(u)| = 1$. Bent functions were introduced by Rothaus [25]. Recall that a bent function of d variables exists if and only if d is even. For every bent function, $\|f\|_1 = \|\tilde{f}\|_1 = \sqrt{2^d}\|f\|_2$. Thus, Theorem 2.9 does not imply a good approximation error. Moreover, it is possible to prove that for any bent function, the approximation error cannot be small if we only take approximations in the two bases. For any bent function f and any function f_n, which is a linear combination of at most n elements of the Euclidean basis or a linear combination of at most n elements of the Fourier basis,
$\|f - f_n\|_2 \geq \sqrt{2^d - n}$ holds.
In the rest of this section, we deal with the approximation where the only constraint is a limited number of perceptrons.

For every even d, let the function $\phi_d : \{0,1\}^d \to \{-1,1\}$ be defined by

$$\phi_d(x) = \left\{ \begin{array}{ll} -1 & \text{if } |x| \equiv 0 \text{ (mod 4) or } |x| \equiv 1 \text{ (mod 4)} \\ 1 & \text{otherwise} \end{array} \right\}$$

where $|x|$ denotes the number of ones in a vector $x \in \{0,1\}^d$. This function is symmetric, i.e. it does not depend on the order of input variables. In other words, it depends only on the number of ones in the input vector x. We can easily show that every such function is a linear combination of functions $g_j(x) = \vartheta(\sum_{i=1}^d x_i - j)$ for $j = 0, 1, \ldots, d$. Hence, ϕ_d is easily expressible by perceptrons, in particular $\phi_d \in \mathcal{P}_d(d+1)$, but as it is bent, Theorem 2.9 does not provide a good approximation.

2.3.2 Activation Function with Bounded Partial Derivatives

Hornik at al. [9] extended Barron's result (Theorem 2.5) to feedforward networks with possibly nonsigmoid activation functions approximating mappings and their derivatives simultaneously. The error is of order $\mathcal{O}(\frac{1}{\sqrt{n}})$ and the numerator of the error is a product of constants dependent on the dimension d, norm of the activation function and a signed measure on $\mathcal{R}^d \times \mathcal{R}$.

Let $\mathcal{B}_d^m$ denote the space of all bounded measurable functions on $\mathcal{R}^d$ that have continuous and uniformly bounded (partial) derivatives up through order $m, m \geq 1$. Denote μ a probability measure on $\mathcal{R}^d$ representing the relative frequency with which input patterns occur. For all $f \in \mathcal{B}_d^m$, define the norm $||f||_{\mathcal{B}_d^m} = \max_{0 \leq |\alpha| \leq m} \sup_{x \in \mathcal{R}^d} |D^\alpha f(x)| < \infty$. Denote by $\mathcal{B}_d^m = \mathcal{B}_d^m(\mu)$ the Hilbert space that completes $(\mathcal{B}_d^m, ||.||_{m,\mu})$ where $||.||_{m,\mu}$ denotes the norm induced by the inner product ($\rho_{m,\mu}$ denotes the associated metric). Denote $|a| = \sqrt{(a^T a)}$. Let $l(a) = \max(|a|, 1)$ and for functions $\psi : \mathcal{R} \rightarrow \mathcal{R}$ let $\mathcal{G}_d^m(\psi, B, n) = \{g : \mathcal{R}^d \rightarrow \mathcal{R}; g(x) = \sum_{i=1}^n \beta_i l_i(a_i)^{-m} \psi(a_i^T x + \theta_i), \theta_i, \beta_i \in \mathcal{R}, \sum_{i=1}^n |\beta_i| \leq B\}$. We consider the class $\mathcal{F}_d^m$ of all real valued functions on $\mathcal{R}^d$ represented as $f(x) = \int_{\mathcal{R}^d} e^{ia^T x} d\rho_f(a)$, where ρ_f is a complex measure on $\mathcal{R}^d$ satisfying $||\rho_f||_m = \int_{\mathcal{R}^d} l(a)^m d|\rho_f|(a) < \infty$. Denote $\rho_{m,\mu}(f, G) = \inf_{g \in G} \rho_{m,\mu}(f, g)$.

Theorem 2.10 ([9]) *Suppose that $f \in \mathcal{F}_d^{m+1}$, that $\psi \in \mathcal{B}_1^m$ and all the derivatives of ψ up to order m are integrable. Let $\omega \neq 0$ be chosen in a way that $\tilde{\phi}(\omega) \neq 0$. Let μ and ψ are compactly supported. Then there exist constants B_0 and C depending only on m, d, μ, and ψ such that*

$$\rho_{m,\mu}[f, \mathcal{G}_d^m(\psi, B, n)] < \frac{C||\rho_f||_{m+1}}{\sqrt{n}},$$

provided that $B \geq B_0||\rho_f||_{m+1}$.

2.3.3 Trigonometric Polynomial and Sigmoidal of Order k As Activation Functions

The following theory is adapted from Mhaskar and Micchelli [22]. As pointed out by Hecht-Nielsen in [4], the problem of approximating any function on a compact set can be reduced to one in which the function being approximated is 2π-periodic in each of its variables.

Denote C^d the class of all continuous functions on $[-1,1]^d$ and C^{d*} the class of all 2π- periodic functions. Let $\Pi_{n,l,d,\psi}$ denote the set of all possible outputs of feedforward networks consisting of n neurons arranged in l hidden layers and each neuron evaluating an activation function ψ, where the input of the network is from $\mathcal{R}^d$. Let f have continuous derivatives of order $r \geq 1$ and let the sum of the norms of all the partial derivatives up to the order r be bounded. Without loss of generality, we can assume that the function to be approximated is normalized. Denote Y_r^d (Y_r^{d*} for periodic functions) the class of all functions satisfying this condition. We deal with the classes of functions that satisfy the universal approximation property. We want to estimate

$$\sup_{f \in Y_r^d} E_{n,l,d,\psi}(f), \text{ where } E_{n,l,d,\psi}(f) = \inf_{P \in \Pi_{n,l,d,\psi}} ||f - P||.$$

$E_{n,l,d,\psi}(f)$ measures the theoretically possible best order of approximation of a function f by a network with n neurons. Or we can have an equivalent dual formulation

$$\tilde{E}_{n,l,d,\psi}(Y_r^d) = min\{m \in \mathcal{Z}; \sup E_{m,l,d,\psi}(f) \leq \frac{1}{n}\}.$$

This quantity measures the minimum number of neurons required to obtain accuracy of $\frac{1}{n}$ for all functions in Y_r^d (analogically for Y_r^{d*}).
Let T_n^d denote the class of all d-variable trigonometric polynomials of the order at most n and for a continuous function f, $2\pi-$ periodic in each of its d variables,

$$E_n^d(f) = \min_{P \in T_n^d} ||f - P||.$$

The class T_n^d can be thought of as a subclass of all outputs of networks with a single layer consisting of at most $(2n+1)^d$ neurons, each evaluating the activation function $\sin x$. It is well known that $\sup_{f \in Y_r^{d*}} E_n^d(f) \leq cn^{-r}$. The dual formulation of this estimate gives $\tilde{E}_{n,1,d,\sin}(Y_r^{d*}) = \mathcal{O}(n^{\frac{d}{r}})$. De Vore et al. proved in [3] that any "reasonable" approximation process that aims to approximate all functions in Y_r^{d*} up to an order of accuracy $\frac{1}{n}$ must necessarily depend on at least $\mathcal{O}(n^{\frac{d}{r}})$ parameters. Thus the activation function $\sin x$ provides optimal convergence rates for the class Y_r^{d*}.
Mhaskar introduced the following generalization of the sigmoidal function.
Let $k \geq 0$. We say that a function $\sigma : \mathcal{R} \rightarrow \mathcal{R}$ is *sigmoidal of order k* if $\lim_{x \to \infty} \frac{\sigma(x)}{x^k} = 1$, $\lim_{x \to -\infty} \frac{\sigma(x)}{x^k} = 0$ and $|\sigma(x)| \leq c(1+|x|)^k$, $x \in \mathcal{R}$. A sigmoidal function of order 0 is the customary bounded sigmoidal function. It was proved in [21] that for any integer $r \geq 1$ and any sigmoidal function σ of order $r-1$,

$$\tilde{E}_{n,1,1,\sigma}(Y_r^1) = \mathcal{O}(n^{\frac{1}{r}}) \text{ and } \tilde{E}_{n,1,d,\sigma}(Y_r^d) = \mathcal{O}(n^{\frac{d}{r}+\frac{(d+2r)}{r^2}}) \text{ for } d \geq 2.$$

Mhaskar showed in [20] that if σ is a sigmoidal function of order $k \geq 2$ and $r \geq 1$, then with $l = \mathcal{O}(\frac{\log r}{\log k})$, $\tilde{E}_{n,l,d,\sigma}(Y_r^d) = \mathcal{O}(n^{\frac{d}{r}})$. Thus an optimal network can be constructed using a sigmoidal function of a higher order.

, 1]^d and C^{d*} the
set of all possible
anged in l hidden
, where the input
es of order $r \geq 1$
to the order r be
he function to be
nctions) the class
asses of functions
to estimate

$- P||$.

of approximation
ve an equivalent

$\}$.

quired to obtain

omials of the or-
in each of its d

f networks with
aluating the ac-
$) \leq cn^{-r}$. The
$\frac{d}{r}$). De Vore et
ss that aims to
must necessar-
n function $\sin x$

ial function.
al of order k if
$x \in \mathcal{R}$. A sig-
al function. It
l function σ of

or $d \geq 2$.

der $k \geq 2$ and
otimal network

2.3.4 A General Perceptron Type Activation Function

The following results are from Mhaskar and Micchelli [22]. They investigated the degree of approximation of periodic functions using periodic activation functions. Their general formulation also includes the case of radial basis functions and customary sigmoidal neural networks. The approximation of functions in $\mathcal{C}^{d*}$ is considered by linear combinations of the form $\phi(\mathbf{A}\mathbf{x} + \mathbf{t})$ where $\mathbf{A}$ is a $s \times d$ matrix, $d \geq s \geq 1$, $\phi \in \mathcal{C}^{s*}$ and $\mathbf{t} \in \mathcal{R}^s$. When $d = s$, $\mathbf{A}$ is an identity matrix and ϕ is a radial function, then a linear combination of n such quantities represents the output of a RBF network with n hidden neurons.
We define the Fourier coefficients of ϕ by the formula

$$\hat{\phi}(\mathbf{m}) = \frac{1}{(2\pi)^d} \int_{[-\pi,\pi]^s} \phi(\mathbf{t}) e^{-i\mathbf{m}.\mathbf{t}} dt, \ \mathbf{m} \in \mathcal{Z}^s.$$

Let $S_\phi \subset \{\mathbf{m} \in \mathcal{Z}^s : \hat{\phi}(\mathbf{m}) \neq 0\}$ and assume that there is a set J containing $s \times d$ matrices with integer entries such that $\mathcal{Z}^d = \{\mathbf{A}^T\mathbf{m} : \mathbf{m} \in S_\phi, \mathbf{A} \in J\}$. If $s = d$ and ϕ is a function with none of its Fourier coefficients equal zero (the RBF case) then we may choose $S_\phi = \mathcal{Z}^d$ and $J = \{I_{d\times d}\}$. For $\mathbf{m} \in \mathbf{Z}^d$, we let $\mathbf{k_m}$ be the multi-integer with minimum magnitude such that $\mathbf{m} = \mathbf{A}^T\mathbf{k_m}$ for some $\mathbf{A} = \mathbf{A_m} \in J$. Denote $m_n := \min\{|\hat{\phi}(\mathbf{k_m})| : -2n \leq \mathbf{m} \leq 2n\}$ and $N_n := \max\{|\mathbf{k_m}| : -2n \leq \mathbf{m} \leq 2n\}$ where $|\mathbf{k_m}|$ is the maximum absolute value of the components of $\mathbf{k_m}$. In the neural network case, we have $m_n = |\hat{\phi}(1)|$ and $N_n = 1$. In the radial basis case, $N_n = 2n$. Denote $P = [-\pi, \pi]^s$.

Theorem 2.11 ([22]) *Let $d \geq s \geq 1$, $n \geq 1$ and $N \geq N_n$ be integers, $f \in C^{d*}$, $\phi \in C^{s*}$. It is possible to construct a network $G_{n,N,\phi}(f;\mathbf{x}) := \sum_{\mathbf{j}} d_{\mathbf{j}}\phi(A_{\mathbf{j}}\mathbf{x}+t_{\mathbf{j}})$ such that*

$$||f - G_{n,N,\phi}(f)||_P \leq c\left\{E_n^d(f) + \frac{E_N^s(\phi)n^{d/2}}{m_n}||f||\right\}$$

where the constant c depends on r, d, ψ but not on f and n. In $G_{n,N,\phi}(f;\mathbf{x})$, the sum contains at most $\mathcal{O}(n^d N^s)$ terms, $A_{\mathbf{j}} \in J$, $t_{\mathbf{j}} \in \mathbf{R}^s$, and $d_{\mathbf{j}}$ are linear functionals of f, depending upon n, N, ϕ.

The rate of approximation relates the degree of approximation of f by neural networks explicitly in terms of the degree of approximation of f and ϕ by trigonometric polynomials. Well known estimates from the approximation theory, such as $\sup_{f\in Y_r^{d*}} E_n^d(f) \leq cn^{-r}$ provide close connections between the smoothness of the functions involved and their degree of trigonometric polynomial approximation. In particular, the rate achieved in Theorem 2.11 indicates that the smoother the function ϕ the better the degree of approximation will be The explicit constructions of $G_{n,N,\phi}$ is given in [23]. The network can be trained in a very simple manner, given the Fourier coefficients of the target function. The weights and thresholds (or the centers for RBF) are determined universally for all functions being approximated. Only the coefficients at the output layer depend on the function. They are given as linear combinations of Fourier coefficients of the target functions. It is shown in [23] that $G_{n,N,\phi}$ for a RBF network contains only $\mathcal{O}(n+N)^d$ summands. The generality of this method, however, affects the number of hidden units which is exponential in d. If the activation function σ is not periodic, but satisfied certain decay conditions near ∞, it is still possible to construct a periodic function for which the general theorem can be applied (see [22]). This process was applied on a number of functions σ and the results are summarized in the following table [22].

Table: Order of magnitude of $\tilde{E}_{n,l,s,\sigma}(Y_r^d)$ for different σ's

Function σ	$\tilde{E}_{n,l,s,\sigma}$	Remarks
Sigmoidal, order $r-1$	$n^{1/r}$	$d=s=1,\ l=1$
Sigmoidal, order $r-1$	$n^{d/r+(d+2r)/r^2}$	$d\geq 2,\ s=1,\ l=1$
x^k, if $x\geq 0$, 0, if $x<0$.	$n^{d/r+(2r+d)/2rk}$	$k\geq 2,\ d\geq 2,\ s=1,\ l=1$
$(1+e^{-x})^{-1}$	$n^{d/r}(\log n)^2$	$d\geq 2,\ s=1,\ l=1$
Sigmoidal, order k	$n^{d/r}$	$k\geq 2,\ d\geq 1,\ s=1,$ $l=\mathcal{O}(\log r/\log k)$
$\exp(-\lvert\mathbf{x}\rvert^2/2)$	$n^{2d/r}$	$d=s\geq 2,\ l=1$
$\lvert\mathbf{x}\rvert^k(\log\lvert\mathbf{x}\rvert)^\delta$	$n^{2d/r)(2+(3d+2r)/k)}$	$d=s\geq 2,\ k>0,\ k+d$ even $\delta=0$ if d odd, 1 if d even, $l=1$

3 Discussion

In this paper, we presented some estimates of the approximation error of a multivariable continuous function on a compact set by neural networks with various activation functions in the hidden units. Each of the results expresses the dependence of the number of hidden units in the neural network on various characteristics known about the function to be approximated. For the spline activation function the approximation error was derived by means of the modulus of smoothness of the function f and for $(n+1)^d$ hidden units. In the case of kernel basis functions, the error depends on the total variation of the convolution of the functions f and the kernel functions and on the distance between them and the number of hidden units is only n. Barron's approach for sigmoidal activation functions achieves the error in the form $\frac{C_f}{\sqrt{n}}$, where the constant C_f depends on the Fourier coefficients of the function f. If continuous real functions on a compact set are not examined, but the subclasses of real-valued boolean functions functions are examined, the network with Heaviside activation function and n hidden units achieves the error of the form $\frac{C_f}{\sqrt{n-1}}$, where C_f is a constant depending on l_1 and l_2 norms of function f. If function f is approximated by a network with $n(d+1)$ hidden units with Heaviside activation functions then the approximation error is of the form $\frac{\tilde{C}_f}{\sqrt{n-1}}$ where $\tilde{C}_f$ is a constant depending on l_1 and l_2 norms of the Fourier transform of function f. These two errors are polynomial in dimension d for some functions (for example functions represented by a decision tree of polynomial size) and exponential for bent functions. A network having n hidden units and activation functions with bounded partial derivatives and having Fourier representation approximates a function f within the error of order $\mathcal{O}(\frac{1}{\sqrt{n}})$ and the numerator of the error is a product of constants dependent on the dimension d, norm of the activation function and signed measure on $\mathcal{R}^d \times \mathcal{R}$. Network with trigonometric activation functions with $(2n+1)^d$ units have the error of the form $\mathcal{O}(n^{-r})$, where r is the number of bounded partial derivatives bounded. If the activation function is a sigmoidal of order $k \geq 2$ and $(2n+1)^d$ hidden units, the error is of order $\mathcal{O}(n^{-d/r})$ for $r \geq 1$. A network with a general perceptron type function representable by Fourier transform with n^dN^s (N is defined in section 2.3.4.) achieves the error exponential in d. Keeping in mind the de Vore et al.'s result ([3]), we can see from the above mentioned results that the polynomial or quadratical approximation errors with respect to dimension d were achieved either by increasing the number of hidden units to an exponential number in d or by breaking continuity of the approximation or by using some constants which are derived from function f and their computation can be exponential in d. When such constants are a priori given, then the approximation is polynomial or quadratical. To avoid the exponential complexity of the approximation ("curse of dimensionality") can be by knowing such characteristics about the function or classes of functions to be approximated.

Acknowledgement

This paper was supported by GACR grant No. 201/96/0917.

References

[1] Barron A.R.: Universal Bounds for Superpositions of a Sigmoidal Function, *IEEE Transactions on Information Theory* 1993; vol.39; 3: 930-945

[2] Cybenko G.: Approximation by Superposition of Sigmoidal Functions, *Mathematics of Control, Signals and Systems* 1989, 2; 4: 303-314

[3] DeVore R.H., Micchelli C.A.: Optimal Nonlinear Approximation, *Manuscripta Mathematica* 1989; 63: 469-478

[4] Hecht-Nielsen R.: Theory of Backpropagation Neural Network. In Proceedings of IEEE International Conference on Neural Networks 1, pp 593-605, 1988

[5] Hlaváčková K: An Upper Estimate of the Error of Approximation of continuous multivariable functions by KBF networks. In Proceedings of ESANN'95, Brussels, pp 333-340, 1995

[6] Hlaváčková K., Kůrková V., Savický P.: Upper Bounds on the Approximation Rates of Real-valued Boolean Functions by Neural Networks. In Proceedings of ICANNGA'97, Norwich, England, in print, 1997

[7] Hlaváčková, K., Kůrková, V.: Rates of approximation of real -valued Boolean functions by neural networks. In Proceedings of ESANN'96, Bruges, Belgium, pp 167-172, 1996

[8] Hornik K., Stinchcombe M., White H.: Multilayer Feedforward networks Are Universal Approximators. *Neural Networks* 1989; 2: 359-366

[9] Hornik K., Stinchcombe M., White H., Auer P.: Degree of Approximation Results for Feedforward Networks Approximating Unknown Mappings and Their Derivatives. *Neural Computation* 1994; 6: 1265-1275

[10] Ito, Y.: Finite mapping by neural networks and truth functions. *Math. Scientist* 1992; vol 17: 69-77

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

[12] Kůrková V., Kainen P.C., Kreinovich V.: Dimension-independent Rates of Approximation by Neural Networks and Variation with Respect to Half-spaces. In Proceedings of WCNN'95. INNS Press, vol 1, pp 54-57, 1995

[13] Kůrková V., Hlaváčková K.: Uniform Approximation by KBF Networks. In Proceedings of NEURONET'93. Prague, 1-7, 1993

[14] Kůrková V.: Approximation of Functions by Perceptron Networks with Bounded Number of Hidden Units. *Neural Networks* 1995, vol 8; 5: 745-750

[15] Kůrková V., Kainen, P.C., Kreinovich, V.: Estimates of the number of hidden units and variation with respect to half-spaces. *Neural Networks* 1997 (in press)

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

Approximation,

ɜtwork. In Pro-
rorks 1, pp 593-

proximation of
Proceedings of

ɔn the Approx-
ıl Networks. In
1997

of real -valued
of ESANN'96,

ward networks
ʖ9-366

Approximation
Mappings and

nctions. *Math.*

Hilbert space
ıeural network

ndent Rates of
spect to Half-
54-57, 1995

BF Networks.

Ietworks with
vol 8; 5: 745-

he number of
ural Networks

on by neural
ıl Processing:
ırkhauser, pp

[17] Kushilevitz E., Mansour Y.: Learning decision trees using the Fourier spectrum. In Proceedings of 23rd STOC, pp 455-464, 1991

[18] Lorentz G.G.: Approximation of Functions. Holt, Rienhart and Winston, New York, 1966

[19] Mhaskar H.N.: Noniterative Training Algorithms for Mapping Networks, Research Report, California State University, USA, 1993, Version 183601281993

[20] Mhaskar H.N.: Approximation Properties of a Multilayered Feedforward Artificial Neural Network. *Advances in Computational Mathematics* 1993; 1: 61-80

[21] Mhaskar H.N., Micchelli C.A.: Approximation by Superposition of a Sigmoidal Function and Radial Basis functions. *Advances in Applied Mathematics* 1992; 13: 350-373

[22] Mhaskar H.N., Micchelli C.A.: How to Choose an Activation Function. Manuscript, 1994

[23] Mhaskar H.N., Micchelli C.A.: Degree of Approximation by Superposition of a Fixed Function. In preparation.

[24] Mhaskar H.N., Micchelli C.A.: Dimension-independent bounds on the degree of approximation by neural networks. *IBM J. Res Develop.* 1994; vol 38, 3: 277-283

[25] Rothaus O.S.: On "Bent" Functions. *Journal of Combin. Theory* 1976; Ser. A; 20: 300–305

[26] Weaver, H.J.: Applications of discrete and continuous Fourier analysis. John Wiley, New York, 1983

[27] Williamson R.C., Bartlett P.L.: Splines, Rational Functions and Neural Networks. Touretzky Edition, CA, 1992

PERSPECTIVES IN NEURAL COMPUTING

M. Kárný, K. Warwick and V. Kůrková (Eds)

Dealing with Complexity

A NEURAL NETWORKS APPROACH

Springer