FEEDFORWARD NETWORKS WITH ONE HIDDEN LAYER AND THEIR RATES OF APPROXIMATION

Kateřina Hlaváčková
Institute of Computer Science, Academy of Sciences of the Czech Republic
P.O. Box 5, 182 07 Prague 8, Czech Republic

Abstract

We present an overview of some rates of approximation with respect to various computational units in the hidden layer of a one-layered feed-forward neural network. The problem of estimating the number of units in the hidden layer is examined to ensure a given degree of approximation for a given function class. The results are discussed in terms of their complexity and functional characteristics.

Keywords: feed-forward network, rates of approximation, activation function, polynomial and exponential complexity

1 Introduction

Feedforward neural networks with a single hidden layer and with various activation functions has been recently widely studied ([11], [7], [1], [2], [9]). It is a well known 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), Mhaskar and Micchelli, [13]. This result presents another question: 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.

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 ([11]). 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]) and our rates for Heaviside activation functions and the class of real valued activation functions ([6]). The last two sections (2.3.2 and 2.3.3) consist in Mhaskar and Micchelli's results ([13]) on approximation rates for networks with trigonometric polynomials and sigmoidals of order k and a general perceptron activation function. Chapter 3 is discussion of these approximation errors.

2 Feedforward Networks with One Hidden Layer

For a bounded function $f : \mathcal{R}^d \to \mathcal{R}$ 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 $A \subset \mathcal{R}^d$. Denote $\mathcal{C}(A)$ the space of continuous functions on compact $A \subset \mathcal{R}^d$ with the uniform norm and corresponding topology. In the paper, we deal only with feedforward 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 $f : A \to \mathcal{R}$ 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 δ–smoothness* is defined by $\omega_m^d(f, \delta, [0,1]^d) = \sup\{\omega_m^d(f, A) : A \text{ subcube of } [0,1]^d, \text{ diam(A)} \le \delta\}$. The estimation of the error of approximation by multivariable spline functions with fixed knots is by Mhaskar [11]. Let $d \ge 2$ 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 knots.

The hidden units of the corresponding network are of the form $N_m^d(n\mathbf{x}-\mathbf{i})$.

Theorem 2.1 ([11]) *If $f: I \to \mathcal{R}$ 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 modulus of smoothness is a constant depending on the dimension d exponentially. Moreover, the number of nodes is exponential in d.

2.2 KBF and RBF Networks

The more detailed work in in [5]. Let $f, g: \mathcal{R}^d \to \mathcal{R}$ are given functions and by $f * g$ 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 unit computing a function of the form $\phi(\| \mathbf{x}-\mathbf{c} \| /b)$, where $\phi: \mathcal{R} \to \mathcal{R}$ is an even function, $\| \cdot \|$ 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 norm $\| \cdot \|$ 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 $\mathcal{R}^d \to \mathcal{R}$ of the form $k_n(\| \mathbf{x}-\mathbf{c} \|)$, where $\{k_n : \mathcal{R} \to \mathcal{R}\}$ is a sequence of functions, $\| \cdot \|$ 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 norm $\| \cdot \|$ on $\mathcal{R}^d$, and d inputs. By $\mathcal{K}_u(\{k_n, n \in \mathcal{N}\}, \| \cdot \|)$ we denote the set of functions computable by KBF networks with uniform k_n for all hidden units. In [9], we obtained the universal approximation property for the class $\mathcal{K}_u(\{k_n, n \in \mathcal{N}\}, \| \cdot \|)$ of continuous kernel functions on $\mathcal{C}(I^d)$ satisfying a slight condition and every norm $\| \cdot \|$ on $\mathcal{R}^d$.

The classical kernels (i.e. the Féjer kernel, the Dirichlet kernel, the Jackson kernel, the Abel-Poisson kernel, the Weierstrass kernel, and the Landau kernel) satisfy the condition 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 an 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 $h = f * k_n$ and $\epsilon_I(f, h) = \|f - h\|_I$.

Upper bounds on ϵ_I are known for some of the above mentioned convolution kernels.

2.3 Networks with Perceptron Hidden Units

2.3.1 Sigmoidal and Heaviside Activation Functions

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

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

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 $\|.\|_2$ generated by an inner product $f \cdot g$ for any $f, g \in \mathcal{X}$. *cl conv* $\mathcal{G}$ means the closure of the convex hull of $\mathcal{G}$, where $\mathcal{G} \subset \mathcal{X}$. The closure is taken with respect to the topology generated by the norm $\|.\|_2$. $\mathcal{N}$ denotes the set of positive integers.

Theorem 2.3 ([1]) *[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 $\|f - f_n\|_2 \leq \sqrt{\frac{c}{n}}$.*

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. The following result on approximation by sigmoidal functions is a corollary of Theorem 2.3.

Theorem 2.4 ([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 $c_0 = f(0)$.

Kůrková 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 [8]. 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 without using Fourier representations.

If f is a linear but not convex combination of functions from $\mathcal{G}$, then $\mathcal{G}$ in Theorem 2.3 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 $V(f, \mathcal{G}) = \inf\{B \geq 0; f \in cl\ conv\ \mathcal{G}(B)\}$, 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 Kůrková in [10] 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 2.5 ([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 Michelli's result from [15] 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 2.6 ([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 2.6 can be improved.

Theorem 2.7 ([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*

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

If both $\|f\|_{1,A}$ and $\|f\|_2$ are known, then Theorem 2.7 yields a good bound only if $4n \geq \|f\|_{1,A}^2 / \|f\|_2^2$. 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 2.7, yield a bound that differs from the best possible bound based only on $\|f\|_{1,A}$ and $\|f\|_2$ by a constant factor.

In [6] we investigated subclasses of *real-valued boolean functions*, i.e. functions $f : \{0,1\}^d \to \mathcal{R}$. 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.

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 $f \cdot g = \sum_{x \in \{0,1\}^d} f(x) g(x)$. 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
$f(x) = \frac{1}{\sqrt{2^d}} \sum_{u\in\{0,1\}^d} \tilde{f}(u)(-1)^{u\cdot x}$, where the Fourier coefficients $\tilde{f}(u)$ are given by the formula
$\tilde{f}(u) = \frac{1}{\sqrt{2^d}} \sum_{x\in\{0,1\}^d} f(x)(-1)^{u\cdot x}$. 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 $\hat{\vartheta}$ 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 2.8 ([6]) *Let d be a positive integer and $f \in \mathcal{F}(\{0,1\}^d)$. Then for every integer $n \geq 2$ there exists a function $f_n \in \mathcal{P}_d((d+1)n)$ such that $\|f - f_n\|_2 \leq \frac{\|\tilde{f}\|_1}{\sqrt{n}}(1 - \frac{\|f\|_2^2}{\|f\|_1^2})$.*

In [6] we presented examples of functions for which the upper bounds on the approximation error from Theorem 2.8 yield a feasible approximation. Easy examples for which the relative error of approximation is less or equal to 1 are linear combinations of "small" number of generalized parities. An example for which Theorem 2.8 gives a non-trivial estimate are functions which are representable by decision trees of polynomial size with the ratio of the maximum and minimum value of $|f(x)|$ bounded by a polynomial in d.
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 [16]. 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 2.8 does not imply a good approximation error.

2.3.2 Trigonometric Polynomial and Sigmoidal of Order k In the Hidden Layer

The following theory is adapted from Mhaskar and Micchelli [13]. 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 $\sup_{f\in Y_r^d} E_{n,l,d,\psi}(f)$, 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 $\sigma : \mathcal{R} \to \mathcal{R}$ 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 [?] that for any integer $r \geq 1$ and any sigmoidal function σ of order $r-1$,

$$\tilde{E}_{n,1,1,\sigma}(Y_r^1) = \mathcal{O}(n^{\frac{1}{r}})$$

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 [12] 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.

2.3.3 A General Perceptron Type Hidden Layer

The following results are from Mhaskar and Micchelli [13] where the degree of approximation of periodic functions using periodic activation functions is investigated. 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

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

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 2.9 ([13]) *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 2.9 indicates that the smoother the function ϕ the better the degree of approximation. The explicit constructions of $G_{n,N,\phi}$ is given in [14]. 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 [14] 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 [13]).

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. 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 knowing some constants which are derived from function f. Their computation can be however exponential in d. When such constants are a priori given (modulus of continuity, certain norms, constants derived from Fourier representation of the function, etc.) then the approximation is guaranteed to be polynomial or quadratical.

In other words, knowledge of such constants enables us to avoid the exponential complexity of the approximation (so called "curse of dimensionality"). The general theorem by Mhaskar confirms that the error of approximation will be otherwise bounded by a function a number exponential in dimension d.

Acknowledgements:

This paper was supported by GACR grant No. 201/96/0917 and GA AVCR grant No. A2030606.

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

[8] 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.

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

[10] 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.

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

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

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

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

[15] 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.

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