ESANN'1995 proceedings - European Symposium on Artificial Neural Networks
Brussels (Belgium), 19-20-21 April 1995, D-Facto public., ISBN 2-9600049-3-0, pp. 333-340

An upper estimate of the error of approximation of continuous multivariable functions by KBF networks

Kateřina Hlaváčková *

Institute of Computer Science,
Academy of Sciences of the Czech Republic,
Prague, Czech Republic
katka@uivt.cas.cz

Abstract. We present an estimate of approximation error of multivariable continuous functions by networks with kernel basis function (KBF) units. The estimate is a function of the number of hidden units and of the total variation of the convolution of the function to be approximated with a kernel basis function. We also present here known estimates of the error of approximation of continuous multivariable functions by networks with sigmoidal and spline activation functions. All the presented estimates of errors depend indirectly on the number of units in the hidden layer.

1. Introduction

In neural network theory, some estimates of the error of approximation of continuous functions by various feedforward neural networks are known. Unfortunately, they are mostly limited to one-dimensional functions.
We focus here on the approximation of multivariable functions by feedforward neural networks and deal with recent estimates for the most common neural networks: sigmoidal (Barron [1]), spline networks (Mhaskar [9]) and our kernel basis function network.
Section 2 contains the main definitions of the terms used in the paper and the estimates of the error for spline and sigmoidal networks. Section 3 deals with approximation by radial basis function networks (RBF) and by more general kernel basis networks (KBF). In section 4, we derive an estimate of rates of approximation by KBF networks based on approximation by convolutions. As an example, using Jackson's estimate, we give an upper bound on approximation error for KBF networks with the Jackson convolution kernel. Section 5 discusses the presented estimates.

*This work was supported by GACR under grant 201/93/0427.

ESANN'1995 proceedings - European Symposium on Artificial Neural Networks
Brussels (Belgium), 19-20-21 April 1995, D-Facto public., ISBN 2-9600049-3-0, pp. 333-340

2. Activation Functions, Functional Spaces and Neural Networks

2.1. Definitions

Let us recall the definitions of the terms used in the paper.
By $\mathcal{R}$ and $\mathcal{N}$ we denote the set of real numbers and positive integers, respectively; $I = [0,1]$ and $\mathcal{R}_+ = [0,\infty)$. For a bounded function $f : \mathcal{R}^d \to \mathcal{R}$ the uniform norm is defined by

$$\| f \|_\infty = \sup_{\mathbf{x}\in\mathcal{R}^d} |f(\mathbf{x})|.$$

A *convolution* of two functions $f, g : \mathcal{R}^d \to \mathcal{R}$ is $f * g = \int_{\mathcal{R}^d} f(\mathbf{x})g(\mathbf{x}-\mathbf{y})d\mathbf{y}$.

Let $f : \mathcal{R}^d \to \mathcal{R}$ be a function. Let $[\mathbf{a},\mathbf{b}] = \prod_{j=1}^d [a_j, b_j]$ be a given cube in $\mathcal{R}^d$. Let $int(A)$ denotes the interior of the set A.

The following definition is from [?]. Define $U[\mathbf{a},\mathbf{b}] = \{\mathbf{x}$; either $x_i = a_i$ or $x_i = b_i\}$ and let $\tau(\mathbf{x})$ denotes 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 a d-dimensional function f on the interval $[\mathbf{a},\mathbf{b}]$ 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 $j \neq l, = 1,\ldots,k$.

We say that *f is of bounded variation* if $V(f)$ is finite. The function

$$\omega(f,h) = \max_{\mathbf{x},\mathbf{t},\|\mathbf{t}\|\leq h} |f(\mathbf{x}+\mathbf{t}) - f(\mathbf{x})|$$

is called the *modulus of continuity of f*.
If $A = \prod_{j=1}^d [a_j, b_j]$, the *modulus of smoothness* of a function $f : A \to \mathcal{R}$ $\omega_m^d(f,A)$ may be 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.

ESANN'1995 proceedings - European Symposium on Artificial Neural Networks
Brussels (Belgium), 19-20-21 April 1995, D-Facto public., ISBN 2-9600049-3-0, pp. 333-340

We may also define the *modulus of δ-smoothness*

$$\omega_m^d(f, \delta, [0,1]^d) = \sup\{\omega_m^d(f, A) : A \text{ subcube of } [0,1]^d, \text{ diam(A)} \leq \delta\}.$$

Let $[\mathbf{a}, \mathbf{b}] \subset \mathcal{R}^d$. Denote $\mathcal{C}[\mathbf{a}, \mathbf{b}]$ the space of continuous functions on $[\mathbf{a}, \mathbf{b}]$ with the uniform norm and corresponding topology.

2.2. Some Known Estimates of the Error of Approximation

There are various activation functions used in the hidden layer of feedforward networks for approximation of functions, namely sigmoidal functions, polynomials, spline functions, radial basis functions and kernel basis functions. We will present here some known results on estimates of the error of approximation by the networks with spline and sigmoidal activation functions.

Denote $\mathcal{S}_m(\Delta)$ - a space of splines of order m with the partition $\Delta = \{\mathbf{x}_i\}_0^{k+1}$ in the interval $[\mathbf{a}, \mathbf{b}] \subset \mathcal{R}^d$. The members of a basis of the space $\mathcal{S}^m(\Delta)$ are called *B-spline functions.*
The corresponding neural network has spline activation functions assigned to the neurons in the hidden layer.
Mhaskar dealt in [9] with the estimation of the error of approximation by multivariable spline functions with fixed knots. His result is based on the generalization of the results in Schumaker [12] considering tensor product.
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*

$$N_m^d(\mathbf{x}) = \prod_{j=1}^{d} N_m(x_j).$$

Theorem 1 ((Mhaskar)) *If $f : [0,1]^d \to \mathcal{R}$ 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 such that if the interpolating points are properly spaced, we have*

$$\max_{\mathbf{x} \in [0,1]^d} |f(\mathbf{x}) - Q_n^d(\mathbf{x})| \leq c\omega_m^d(f, \frac{1}{n}, [0,1]^d),$$

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

In other words, the approximation error depends indirectly on the number of units in the hidden layer of the network. The more units are in the hidden

layer, the better approximation is achieved.

Let $k \geq 0$ be an integer, $\sigma : \mathcal{R} \to \mathcal{R}$ be a bounded measurable function on the real line for which $\lim_{x \to -\infty} \sigma(x) = 0$, $\lim_{x \to \infty} \sigma(x) = 1$. 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 us mention here an interesting result by Barron [1] on approximation by sigmoidal functions. 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 $C_f \leq C$.

Theorem 2 ((Barron)) *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), such that*

$$\int_{B_r} (f(\mathbf{x}) - f_n(\mathbf{x}))^2 \mu(d\mathbf{x}) \leq \frac{c_f'}{n},$$

where $c_f' = (2rC_f)^2$. 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 $c_0 = f(0)$.

It is also shown in [1] that for $f(\mathbf{x}) = e^{-\|\mathbf{x}^2\|/2}$ (Gaussian function) $C_f \leq d^{1/2}$, so $\frac{c_f'}{n} = \frac{4r^2 d}{n}$. The more units are in the hidden layer, the better approximation is achieved.

3. RBF and KBF Networks

Radial basis function (RBF) networks have been recently studied by many authors (for example, by Broomhead and Lowe [2], Moody and Darken [10], Girosi and Poggio [5], and Park and Sandberg [11]). Let us recall here the definitions of RBF and KBF networks.

A *radial basis function (RBF) unit* with d inputs is a computational unit that computes a function from $\mathcal{R}^d$ to $\mathcal{R}$ of the form $\phi(\| \mathbf{x}-\mathbf{c} \| /b)$, where $\phi : \mathcal{R} \to \mathcal{R}$ is an even (radial) function, $\| . \|$ is a norm on $\mathcal{R}^d$, and $\mathbf{c} \in \mathcal{R}^d$, $b \in \mathcal{R}$, $b > 0$.
A *radial basis function (RBF) network* is a neural network with a single linear

ESANN'1995 proceedings - European Symposium on Artificial Neural Networks
Brussels (Belgium), 19-20-21 April 1995, D-Facto public., ISBN 2-9600049-3-0, pp. 333-340

output unit, one hidden layer with RBF units that have the same radial function ϕ and the same norm $\| \cdot \|$ on $\mathcal{R}^d$, and d inputs.
By $\mathcal{F}(\phi, \| \cdot \|)$ we denote the set of real-valued functions on I^d computable by RBF networks with the radial function ϕ and the norm $\| \cdot \|$ with any number of hidden units. The most popular radial function currently used in applications is the Gaussian $\gamma(t) = \exp(-t^2)$ (see [10]).
Approximating a convolution by an appropriate Riemann sum, we obtained in [6] that for every positive integer d and for every continuous function $\phi : \mathcal{R} \rightarrow \mathcal{R}_+$ with a finite non-zero integral and every norm $\| \cdot \|$ on $\mathcal{R}^d$, $\mathcal{F}(\phi, \| \cdot \|)$ is dense in $\mathcal{C}(I^d)$. In other words, the class of single hidden layer RBF networks with uniform width has the universal approximation property.
In [7], we built on these results and derived estimates of rates of approximation for one-dimensional continuous functions.

In this paper, we build on these results to derive estimates of rates of approximation for multivariable continuous functions with bounded total variation. We show here that for any of a number of classical kernel functions the rate of approximation is bounded above by terms depending on the total variation of the convolution of the approximated function and a given kernel function.
Recall here the definition of general kernel functions as we introduced them in [6] into neural networks by defining *kernel basis function (KBF) units.*

A KBF unit with d inputs computes a function $\mathcal{R}^d \rightarrow \mathcal{R}$ of the form $k_n(\| \mathbf{x} - \mathbf{c} \|)$, where $\{k_n : \mathcal{R} \rightarrow \mathcal{R}\}$ 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. We call n *sharpness.*
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 $\| \cdot \|$ on $\mathcal{R}^d$, and d inputs.
By $\mathcal{K}(\{k_n, n \in \mathcal{N}\}, \| \cdot \|)$ we denote the set of functions computable by KBF networks with $\{k_n, n \in \mathcal{N}\}$ and $\| \cdot \|$ with any number of hidden units. By $\mathcal{K}_u(\{k_n, n \in \mathcal{N}\}, \| \cdot \|)$ we denote the set of functions computable by KBF networks with $\{k_n, n \in \mathcal{N}\}$ and $\| \cdot \|$ with any number of hidden units and uniform k_n for all hidden units.
Similarly, we obtained the universal approximation property for quite general KBF networks in [6] which is presented in the following theorem:

Theorem 3 *For every positive integer d and for every sequence of continuous functions $\{k_n : \mathcal{R} \rightarrow \mathcal{R}_+, n \in \mathcal{N}\}$ and every norm $\| \cdot \|$ on $\mathcal{R}^d$ satisfying for every $n \in \mathcal{N}$ and every $\mathbf{x} \in \mathcal{R}^d$ $\int_{\mathcal{R}^d} k_n(\| \mathbf{x} - \mathbf{y} \|) d\mathbf{y} = 1$ and for every $\delta > 0$ and every $\mathbf{x} \in \mathcal{R}^d$ $\lim_{n \rightarrow \infty} \int_{J_\delta(\mathbf{x})} k_n(\| \mathbf{x} - \mathbf{y} \|) d\mathbf{y} = 0$, where $J_\delta(\mathbf{x}) = \{\mathbf{y} | \, \mathbf{y} \in \mathcal{R}^d, \| \mathbf{x} - \mathbf{y} \| \geq \delta\}$, the class $\mathcal{K}_u(\{k_n, n \in \mathcal{N}\}, \| \cdot \|)$ is dense in $\mathcal{C}(I^d)$.*

Note that 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 (for the definitions see [6]) satisfy the assumptions of Theorem

ESANN'1995 proceedings - European Symposium on Artificial Neural Networks
Brussels (Belgium), 19-20-21 April 1995, D-Facto public., ISBN 2-9600049-3-0, pp. 333-340

3 and so KBF networks with any of these kernels are powerful enough to approximate continuous functions (of course, to achieve arbitrary accuracy, one must increase the number of hidden units).

4. An Estimate of the Error of Approximation for KBF Networks

Let us present here the main result of this paper. For some of the above mentioned convolution kernels upper bounds on convolution approximation are known. Our following theorem derives an estimate of the rate of approximation by KBF networks depending on the error of approximation $E(f,k_n) = \| f - f * k_n \|_H = \sup_{\mathbf{x}\in H} |f(\mathbf{x}) - f * k_n(\mathbf{x})|$, $H \subset \mathcal{R}^d$ and the total variation of $f * k_n$.

Theorem 4 *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, $I^d = [0,1]^d$, $E(f,k_n) = \| f(\mathbf{x}) - \int_{I^d} f(\mathbf{t}) k_n(\| \mathbf{x} - \mathbf{t} \|) d\mathbf{t} \|_{I^d}$. Let $f * k_n$ be of a bounded variation. Then for every $m \in \mathcal{N}$ there exists a KBF network with m hidden units computing a function $g \in \mathcal{K}_u(\{k_n\}, \|.\|)$ such that for every $\mathbf{x} \in I^d$*

$$|f(\mathbf{x}) - g(\mathbf{x})| \leq E(f,k_n) + \frac{d}{m} V(h),$$

where

$$h(\mathbf{x}) = f * k_n(\mathbf{x}) = \int_H f(\mathbf{t}) k_n(\| \mathbf{x} - \mathbf{t} \|) d\mathbf{t}.$$

As in the previous theorems, the greater the number of the units in the hidden layer the smaller the approximation error.

Proof:

By assumption, $|f(\mathbf{x}) - \int_{I^d} f(\mathbf{x}) k_n(\| \mathbf{x} - \mathbf{t} \|) d\mathbf{t}| \leq E(f,k_n)$ for every $\mathbf{x} \in I^d$. We estimate the function $\int_{I^d} f(\mathbf{t}) k_n(\| \mathbf{x} - \mathbf{t} \|) d\mathbf{t}$ by multivariable Riemann sums s_m (see [4], $p.129$). Under the condition that $V(f * k)$ is bounded, we have

$$\left| \int_{I^d} f(\mathbf{t}) k_n(\| (\mathbf{x} - \mathbf{t}) \|) d\mathbf{t} - s_m(\mathbf{x}) \right| \leq \frac{d}{m} V(f * k),$$

and

$$s_m(\mathbf{x}) = \frac{1}{m^d} \sum_{i_1,\ldots i_d = 0}^{m-1} f(\frac{i_1}{m}, \ldots \frac{i_d}{m}) k_n(\| (x_1 - \frac{i_1}{m}, \ldots, x_d - \frac{i_d}{m}) \|),$$

where $i = (i_1, \ldots, i_d)$.

Then $|\int_{I^d} f(\mathbf{x}) d\mathbf{x} - s_m(\mathbf{x})| \leq |\int_{I^d} f(\mathbf{x}) d\mathbf{x} - \int_{I^d} f(\mathbf{x}) k_n(\| \mathbf{x} - \mathbf{t} \|) d\mathbf{t}| +$
$+ |\int_H f(\mathbf{x}) k_n((\| \mathbf{x} - \mathbf{t} \|) d\mathbf{t} - s_m(\mathbf{x})| \leq E(f,k_n) + \frac{d}{m} V(f * k_n)$. □

As an example, we use this theorem to estimate the error of approximation

ESANN'1995 proceedings - European Symposium on Artificial Neural Networks
Brussels (Belgium), 19-20-21 April 1995, D-Facto public., ISBN 2-9600049-3-0, pp. 333-340

for the KBF networks based on the Jackson kernel with inputs in the interval $P = [-\pi, \pi]^d$. Consider the following operator:

$$\int_P f(\mathbf{t}) L_n(\mathbf{x} - \mathbf{t}) d\mathbf{t} = \int_P f(\mathbf{x} + \mathbf{t}) L_n(\mathbf{t}) d\mathbf{t}, \tag{2}$$

where L_n is the *Jackson kernel*

$$L_n(\mathbf{t}) = \lambda_n^{-1} \left(\frac{\sin \| \, n.\mathbf{t}/2 \, \|}{n \sin \| \, \mathbf{t}/2 \, \|} \right)^4, \quad \int_P L_n(\mathbf{t}) d\mathbf{t} = 1,$$

where the last relation defines λ_n. It is convenient to normalize the operator (2) in such a way as to obtain a trigonometric polynomial of degree n. For this purpose, we put

$$K_n(\mathbf{t}) = L_r(\mathbf{t}), \qquad r = [\frac{n}{2}] + 1$$

The operator $J_n(\mathbf{x}) = J_n(f, \mathbf{x}) = \int_P f(\mathbf{x} + \mathbf{t}) K_n(\mathbf{t}) dt$ is called the (multivariable) *Jackson operator.*

Theorem 5 ((Jackson)) *There exists a constant M such that, for each function $f \in C(P)$, where $P = [-\pi, \pi]^d$ and for every $n \in \mathcal{N}$, $|f(\mathbf{x}) - J_n(\mathbf{x})| \leq M \omega_P(f, \frac{1}{n})$.*

The proof is a multivariable generalization of the theorem in [8], p.56.

Theorem 6 *There exists a constant M such that for every $f \in C(\mathcal{P})$, $P = [-\pi, \pi]^d$, for every n (sharpness of the Jackson kernel) and for every $m \in \mathcal{N}$ and a function g computable by a Jackson KBF network with m hidden units and with sharpness n such that for every $\mathbf{x} \in P$*

$$|f(\mathbf{x}) - g(\mathbf{x})| \leq M \omega_P(f, \frac{1}{r}) + \frac{d}{m} V(J_n), \tag{3}$$

where $r = [\frac{n}{2}] + 1$ and $P^* = [-2\pi, 2\pi]$.

Proof. ¿From Theorems 6 and 7, where $E(f, k_n) \leq M \omega_P(f, \frac{1}{n})$. □

5. Discussion

In this paper, we have presented some estimates of error of approximation of a multivariable continuous function on a compact set by the most common neural networks: sigmoidal networks (by Barron), spline networks (by Mhaskar) and our estimate for kernel basis function networks. All the presented estimates of errors depend indirectly on the number of units in the hidden layer.
It is difficult to decide which estimate is the best for a general continuous function. Each result was achieved for a function under different conditions. For example, for some functions it can be demanding to compute a Fourier

transform as it is required in Theorem 2. On the other hand, the modulus of δ–smoothness or the total variation of the function can be gained more easily as it is required in Theorem 1. The quality of approximation in Theorem 6 depends on the kernel function k_n by $E(f, k_n)$ and the total variation of the convolution of f and k_n.

References

[1] A.R. Barron: Universal Bounds for Superpositions of a Sigmoidal Function, IEEE Transactions on Information Theory, vol.39, no.3, 930-945, May (1993).

[2] D.S. Broomhead, D. Lowe: Multivariable Functional Interpolation and Adaptive Networks. Complex Systems 2, 321-355 (1988).

[3] C.K. Chui, X. Li, H.N. Mhaskar: Neural Networks For Localized Approximation, in preparation.

[4] P. J. Davis, P. Rabinowitz: Numerical Integration, Blaisdell Publishing Company, MA, USA, (1967).

[5] F. Girosi, T. Poggio: Networks and the Best Approximation Property. Biological Cybernetics 63, 169-176 (1990).

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

[7] V. Kůrková, K. Hlaváčková: Approximation of Continuous Functions by RBF and KBF Networks, Proceedings of ESANN'94, Brussels, 167-174 (1994).

[8] G.G. Lorentz: Approximation of Functions, Holt, Rinehart and Winston, New York (1966).

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

[10] J. Moody, Ch.J. Darken: Learning with Localized Receptive Fields. Proceedings of the 1988 Connectionist Models Summer School, San Mateo, CA (1989).

[11] J. Park, I.W. Sandberg: Approximation and Radial-Basis-Function networks. Neural Computation 5, 305-316 (1993).

[12] L.L. Schumaker: Spline Functions:Basic Theory. John Wiley and Sons, New York (1981).