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. 321-326

Approximation of functions by Gaussian RBF networks with bounded number of hidden units

Věra Kůrková *

Institute of Computer Science, Czech Academy of Sciences, Prague, Czechia
e-mail: vera@uivt.cas.cz

Abstract. We characterize functions that are in the uniform closures of sets of functions computable by Gaussian RBF networks with a bounded number of hidden units. We show how description of such functions can be applied to comparison of rates of approximation by networks with perceptrons and with Gaussian RBF units.

1 Introduction

In recent years, approximation of functions between finite dimensional Euclidean spaces by feedforward neural networks has received wide research interest. Capabilities of many classes of neural networks to approximate arbitrarily well continuous or measurable functions were proven. But to approximate within a specific accuracy some given function, it is not known which type of network can be least complex in the sense of requiring the fewest number of hidden units. For example, perceptrons and radial-basis-function (RBF) units are geometrically opposite: perceptrons apply a sigmoidal activation function to a weighted sum of inputs plus a bias and so respond to *non-localized* regions of the input space by partitioning it with fuzzy hyperplanes (or sharp ones if the sigmoid is Heaviside's step-function), while RBF units calculate the distance between an input vector and a centroid, multiply by a width, and then apply a kernel function - hence, respond to *localized* regions. So perceptron type networks and RBF networks may be efficient in approximating different types of functions.

Suppose for instance that the rates of approximations of functions from a set $\mathcal{S}$ by perceptron networks are related to the rates of approximation by RBF networks in such way that there exists a mapping $r : \mathcal{N} \to \mathcal{N}$ ($\mathcal{N}$ denotes the set of natural numbers) such that for every $f \in \mathcal{S}$ and $\varepsilon > 0$ if f can be approximated within ε by a perceptron network with k hidden units, then f can be also approximated within ε by a RBF network with $r(k)$ hidden units. If $\mathcal{S}$ contains a function computable by a single perceptron, then for every ε there exists a RBF network with $r(1)$ hidden units approximating it within ε. In mathematical terms this means that the function computable by a single

* This work was partially supported by GACR grant 201/93/0427 and Warsaw University of Technology Project PATIA 503/901/2.

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. 321-326

perceptron is in the closure (with respect to the norm in which the error is measured) of the set of functions computable by RBF networks with $r(1)$ hidden units.

In this paper we characterize functions that are in the uniform closures of sets of functions computable by Gaussian RBF networks with fixed number of hidden units. We show that no ridge function, in particular no function computable by a single perceptron with any activation function, could be contained in these closures. The opposite question is whether the rates of approximations by RBF networks are related to the rates of approximations by perceptron networks. We answered it in [2] by showing that sets of functions computable by networks with bounded number of Heaviside perceptrons are closed. Our work was inspired by Chui et al. [1] who compared approximation capabilities of one and two-hidden-layer perceptron networks with Heaviside activation function.

The paper is organized as follows. In section 2 we recall basic concepts and results, in section 3 we present our characterization of closures of spaces of functions computable by Gaussian RBF networks with bounded number of hidden units, and in section 4 we apply these results to comparison of rates of approximation. A sketch of the proof of our main theorem is given in the Appendix.

2 The universal and the best approximation property

In this paper we consider the problem of approximating continuous functions by a one-hidden-layer radial-basis-function (RBF) network with a single linear output unit. Since in any practical application, values of the inputs can vary only within certain limits, we can suppose that input vectors are within the unit cube I^d, where $I = [0,1]$ and d is the number of inputs. For an even function $\psi : \mathcal{R} \to \mathcal{R}$ we denote by $\mathcal{F}_d(\psi)$ the set of functions computable by networks with d inputs and any finite number of RBF hidden units with a radial function ψ and Euclidean norm $\|.\|$ on $\mathcal{R}^d$, i.e. $\mathcal{F}_d(\psi)$ is the set of all functions from $\mathcal{R}^d$ to $\mathcal{R}$ of the form $\sum_{i=1}^{k} w_i \psi\left(\frac{\|\mathbf{x}-\mathbf{c}_i\|}{b_i}\right)$, where $w_i, b_i \in \mathcal{R}$, $b_i > 0$ and $\mathbf{c}_i \in \mathcal{R}^d$. By $\mathcal{F}_d(\psi, k)$ we denote the subset of $\mathcal{F}_d(\psi)$ containing only functions computable by networks with at most k hidden units and by $\mathcal{F}_d(\psi, k, B)$ the subset of $\mathcal{F}_d(\psi, k)$ containing functions computable by networks with parameters bounded by B, i.e. satisfying $|w_i| \leq B$, $|\frac{1}{b_i}| \leq B$. The standard choice of a radial function is Gaussian that we denote by γ, i.e. $\gamma(x) = e^{-x^2}$.

Capabilities of networks to approximate functions are studied mathematically in terms of closures and dense subspaces; see, e.g. [7] for the basic definitions and theorems. By $cl(X)$ we denote the closure of a subset X in the space $\mathcal{C}(I^d)$ of all continuous functions on I^d with the topology of uniform convergence (i.e. topology induced by the supremum norm). So, $cl(X) = \{f : I^d \to \mathcal{R}; (\forall \varepsilon > 0)(\exists g \in X)(\sup\{|f(\mathbf{x}) - g(\mathbf{x})|; \mathbf{x} \in I^d\} < \varepsilon)\}$.

For any continuous function ψ with finite non-zero integral, the sets $\mathcal{F}_d(\psi)$ are known to be dense in $\mathcal{C}(I^d)$, i.e. $cl(\mathcal{F}_d(\psi)) = \mathcal{C}(I^d)$ ([5], [6], [3]). In neural

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. 321-326

networks terminology this capability is called the *universal approximation property*. However, one may require an arbitrarily large number of hidden units as well as size of their parameters as the accuracy of the approximation increases.

In practical situations, the number of hidden units is bounded by some fixed positive integer. In addition, the parameters are also bounded. Under these conditions, we showed in [2] that for many types of feedforward networks including RBF, given any continuous function, there is a choice of network parameterization (not necessarily unique) producing an approximation with the minimum error. We call this the *best approximation property*. In fact we showed that such spaces are compact, which in particular implies that $\mathcal{F}_d(\psi, k, B)$ is closed for any bounded ψ and thus no function that is not already contained in $\mathcal{F}_d(\psi, k, B)$ can be approximated with any accuracy by RBF networks with bounds on both the size of parameters and the number of hidden units. A major question which is not yet fully understood is how quickly such best approximation error decreases with the growth of the number of hidden units.

3 Closures of spaces of functions computable by Gaussian RBF networks with bounded number of hidden units

In contrast to $\mathcal{F}_d(\psi, k, B)$, the sets $\mathcal{F}_d(\psi, k)$ containing functions computable by RBF networks with only the number of hidden units being constrained need not be closed as subspaces of $\mathcal{C}(I^d)$. For example, the function $-2x^2\gamma(x)$ can be approximated with any accuracy by a Gaussian RBF network with only two hiden units. Indeed,

$$-2x^2\gamma(x) = -x\gamma'(x) = \frac{\partial\gamma(\frac{x-c}{b})}{\partial b}\Big|_{b=1,c=0} = \lim_{n\to\infty} n\left(\gamma\left(\frac{x-c}{b+\frac{1}{n}}\right) - \gamma\left(\frac{x-c}{b}\right)\right).$$

We will show that for no other functions than linear combinations of partial derivatives of $\gamma\left(\frac{\|\mathbf{x}-\mathbf{c}\|}{b}\right)$ with respect to $c_1, \ldots, c_d$ and b does there exists a bound on the number of hidden units needed for an arbitrarily close approximation. Denote by $\Phi_d(\gamma, k)$ the set of all functions of the form

$$\sum_{i=1}^{m}\left(\sum_{j=1}^{m_i}\left(v_{ij0}\frac{\partial^{k_j}\gamma\left(\frac{\|\mathbf{x}-\mathbf{c}_j\|}{b_j}\right)}{\partial {b_j}^{k_j}} + \sum_{l=1}^{d} v_{ijl}\frac{\partial^{k_j}\gamma\left(\frac{\|\mathbf{x}-\mathbf{c}_j\|}{b_j}\right)}{\partial {c_{jl}}^{k_j}}\right)\right) + a,$$

where $\sum_{i=1}^{m}\sum_{j=1}^{m_i} m_i k_j \leq k$ and $a, v_{ijl} \in \mathcal{R}$, $i = 1, \ldots, m$, $j = 1, \ldots, m_i$, $l = 0, \ldots, d$.

The following characterization of $cl(\mathcal{F}_d(\gamma, k))$ is based on the properties of total differential and a proof technique that we developed in our previous paper [4] to verify essential uniqueness of a Gaussian RBF network parameterization.

Theorem 3.1 *For all positive integers d, k and for every $f \in cl(\mathcal{F}_d(\gamma, k))$ there exist $h \in \mathcal{F}_d(\gamma, k)$, and $\phi \in \Phi_d(\gamma, k)$ such that $f = h + \phi$.*

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. 321-326

The following lemma is easy to verify by induction.

Lemma 3.2 *For every positive integer k there exist polynomials $p_k, q_k : \mathcal{R}^2 \to \mathcal{R}$ such that for every positive integer d* $\frac{\partial^k \gamma\left(\frac{\|\mathbf{x}-\mathbf{c}\|}{b}\right)}{\partial b^k} = \frac{p_k(\|\mathbf{x}-\mathbf{c}\|,b)}{b^{3k}}\,\gamma\left(\frac{\|\mathbf{x}-\mathbf{c}\|}{b}\right)$ *and for every $l = 1,\ldots,d$* $\frac{\partial^k \gamma\left(\frac{\|\mathbf{x}-\mathbf{c}\|}{b}\right)}{\partial {c_l}^k} = \frac{q_k(x_l-c_l,b)}{b^{2k}}\,\gamma\left(\frac{\|\mathbf{x}-\mathbf{c}\|}{b}\right)$.

Theorem 3.1 together with Lemma 3.2 imply the following characterization of $cl(\mathcal{F}_d(\gamma,k))$.

Corollary 3.3 *There exist polynomials $p_n, q_n : \mathcal{R}^2 \to \mathcal{R}$, $n \in \mathcal{N}$, such that for every positive integers d, k every $f \in cl(\mathcal{F}_d(\gamma,k))$ can be represented as*

$$\sum_{i=1}^{m} \gamma\left(\frac{\|\mathbf{x}-\mathbf{c}_i\|}{b_i}\right)\left(w_i + \sum_{j=1}^{m_i}\left(\frac{u_{ij}p_{k_j}(\|\mathbf{x}-\mathbf{c}_i\|,b_i)}{{b_i}^{3k_j}} + \sum_{l=1}^{d}\frac{v_{ijl}q_{k_j}(x_l-c_l,b_i)}{{b_i}^{2^{k_j}}}\right)\right)+a,$$

where $m, m_i, k_j \in \mathcal{N}$ $(i = 1,\ldots,m,\ j = 1,\ldots,m_i)$ satisfy $\sum_{i=1}^{m}\sum_{j=1}^{m_i} m_i k_j \leq k$ and $a, w_i, v_{ijl}, u_{ij} \in \mathcal{R}$ $(i = 1,\ldots,m,\ j = 1,\ldots,m_i,\ l = 1,\ldots,d)$.

4 Comparison of rates of approximation

Characterization of closures of spaces of functions computable by networks with a bounded number of hidden units might be useful for comparison of rates of approximation of functions by networks with different types of units. Let $\mathcal{F}$ and $\mathcal{G}$ be sets of functions computable by two different types of feedforward networks. By $\mathcal{F}(k)$ and $\mathcal{G}(k)$ denote subsets of $\mathcal{F}$ and $\mathcal{G}$, resp., containing functions computable by networks with at most k hidden units. We say that *the rate of approximation of functions from a set $\mathcal{S}$ by $\mathcal{F}$ is related to the rate of approximation of functions from $\mathcal{S}$ by $\mathcal{G}$* if there exists a mapping $r : \mathcal{N} \to \mathcal{N}$ such that for every $f \in \mathcal{S}$ and $\varepsilon > 0$ if f can be approximated uniformly on I^d within ε by a function from $\mathcal{F}(k)$, then it can be uniformly approximated within ε by a function from $\mathcal{G}(r(k))$. If $\mathcal{S}$ is large enought to contain a function from $\mathcal{F}(1)$, i.e. a function f computable a single-hidden-unit network of the first type, then there exists k $(k = r(1))$ such that $f \in cl(\mathcal{G}(k))$.

Suppose that $\mathcal{S}$ contains a function computable by a single perceptron network with an activation function ψ, i.e. $\mathcal{S}$ contains a function of the form $\psi(\mathbf{v}\cdot\mathbf{x}+b)$. This function is constant on every hyperplane parallel with the cozero hyperplane of the affine function $\mathbf{v}\cdot\mathbf{x}+b$. It follows from Corollary 3.3 that for any natural number k this function cannot be contained in $cl(\mathcal{F}_d(\gamma,k))$ since functions described in 3.3 are not constant on such hyperplanes. So the rate of approximation of functions from $\mathcal{S}$ by one-hidden-layer networks with perceptrons with activation function ψ is not related to the rate of approximation of functions from $\mathcal{S}$ by RBF networks with Gaussian radial function. In other

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. 321-326

words, we cannot approximate arbitrarily well a function computable by a single perceptron network by RBF networks using only finitely many hidden units even if the weights could be arbitrarily large. This result can be extended to sets $\mathcal{S}$ containing a function computable by a perceptron network, i.e. a function of the form $\sum_{i=1}^{m} w_i \psi(\mathbf{v} \cdot \mathbf{x} + b)$.

5 Appendix

Unless specified otherwise all limits involving functions are pointwise.

Sketch of proof of Theorem 3.1
Let $f \in cl(\mathcal{F}_d(\gamma, k))$. Then there exists a sequence $\{f_n, n \in \mathcal{N}\} \subseteq \mathcal{F}_d(\psi, k)$ such that $f = \lim_{n\to\infty} f_n$. Since we can choose from every sequence of real numbers either a converging or a diverging one we can assume that for every $n \in \mathcal{N}$ $f_n(x) = \sum_{i=1}^{m} w_{in}\gamma\left(\frac{\|\mathbf{x}-\mathbf{c}_{in}\|}{b_{in}}\right)$, where $m \leq k$ and all the sequences $\{w_{in}; n \in \mathcal{N}\}$, $\{b_{in}; n \in \mathcal{N}\}$, $\{\|\mathbf{c}_{in}\|; n \in \mathcal{N}\}$ and $\{\frac{\|\mathbf{c}_{in}\|}{b_{in}}; n \in \mathcal{N}\}$ are either converging or diverging. In the case of convergence put $w_i = \lim_{n\to\infty} w_{in}$, $b_i = \lim_{n\to\infty} b_{in}$ and $\mathbf{c}_i = \lim_{n\to\mathcal{N}} \|\mathbf{c}_{in}\|$, otherwise put $w_i = \infty$ or $w_i = -\infty$ and $b_i = +\infty$ and $\mathbf{c}_i = +\infty$. Denote by J_{wbc}, J_{bc}, J_{wc}, J_{wb}, J_w, J_b, J_c, J the subsets of the set $\{1, \ldots, k\}$ such that the indexes w, b, c, resp., indicate that i is in J with such an index, if the sequence $\{w_{in}; n \in \mathcal{N}\}$, $\{b_{in}; n \in \mathcal{N}\}$, $\{\mathbf{c}_{in}; n \in \mathcal{N}\}$, resp., is converging. Since γ is continuous and bounded $h_1(\mathbf{x}) = \sum_{i\in J_{wbc}} w_i\psi\left(\frac{\|\mathbf{x}-\mathbf{c}_i\|}{b_i}\right) = \lim_{n\to\infty} \sum_{i\in J_{wbc}} w_{in}\gamma\left(\frac{\|\mathbf{x}-\mathbf{c}_{in}\|}{b_{in}}\right)$ uniformly. Since γ is asymptotically zero, there exists $a_0 \in \mathcal{R}$ such that $a_0 = \lim_{n\to\infty} \sum_{i\in J_{wc}\cup J_{wb}\cup J_w} w_{in}\gamma\left(\frac{\|\mathbf{x}-\mathbf{c}_{in}\|}{b_{in}}\right)$. So, $f(\mathbf{x}) - h_1(\mathbf{x}) - a_0 = \lim_{n\to\infty} \sum_{i\in J_{bc}\cup J_b\cup J_c\cup J} w_{in}\gamma\left(\frac{\|\mathbf{x}-\mathbf{c}_{in}\|}{b_{in}}\right)$. Let H be such a subset of $J_{bc}\cup J_b\cup J_c\cup J$ that all the pairs $\{(b_i, \mathbf{c}_i); i \in H\}$ are mutually different and for every $j \in J_{bc}\cup J_b\cup J_c\cup J$ there exists $i \in H$ such that $(b_j, \mathbf{c}_j) = (b_i, \mathbf{c}_i)$. For each $i \in H$ put $K_i = \{j \in J_{bc} \cup J_b \cup J_c \cup J; (b_j, \mathbf{c}_j) = (b_i, \mathbf{c}_i)\}$ and for each $n \in \mathcal{N}$ put $\hat{w}_{in} = \sum_{j\in K_i} w_{jn}$. Put $L_{bc} = H \cap J_{bc}$, $L_b = H \cap J_b$, $L_c = H \cap J_c$ and $L = H \cap J$. For each $i \in L_{bc}$ and $j \in K_i$ and $n \in \mathcal{N}$ put $\phi_{jn}(\mathbf{x}) = \gamma\left(\frac{\|\mathbf{x}-\mathbf{c}_{jn}\|}{b_{jn}}\right) - \gamma\left(\frac{\|\mathbf{x}-\mathbf{c}_i\|}{b_i}\right)$. For every $i \in L_b \cup L_c \cup L$ and for every $n \in \mathcal{N}$ put $a_i = \lim_{n\to\infty} \gamma\left(\frac{\|\mathbf{x}-\mathbf{c}_{in}\|}{b_{in}}\right)$. Then $f(\mathbf{x}) - h_1(\mathbf{x}) - a_0 = \lim_{n\to\infty}\left(\sum_{i\in L_{bc}}\left(\hat{w}_{in}\gamma\left(\frac{\|\mathbf{x}-\mathbf{c}_i\|}{b_i}\right) + \sum_{j\in K_i} w_{jn}\phi_{jn}(\mathbf{x})\right) + \sum_{i\in L_b\cup L_c\cup L} w_{in}a_i\right)$. Put $P = \{i \in L_{bc}; \{\hat{w}_{in}; n \in \mathcal{N}\}$ *is converging*$\}$ and $Q = \{i \in L_{bc}; \{\hat{w}_{in}; n \in \mathcal{N}\}$ *is diverging*$\}$. Put $h_2(\mathbf{x}) = \sum_{i\in P} \hat{w}_i\gamma\left(\frac{\|\mathbf{x}-\mathbf{c}_i\|}{b_i}\right)$ and $h = h_1 + h_2$. Put $v_n = \max\{|\hat{w}_{in}|; i \in L_{bc}\}$. Let $\mathcal{M}$ be an infinite subset of $\mathcal{N}$ such that there exists $i_0 \in L_{bc}$ such that for every $n \in \mathcal{M}$ $v_n = |\hat{w}_{i_0n}|$. Put $v_{in} = \frac{\hat{w}_{in}}{v_n}$ and $v_i =$

$\lim_{\substack{n\to\infty \\ n\in\mathcal{M}}} v_{in}$. Then $\lim_{\substack{n\to\infty \\ n\in\mathcal{M}}} \frac{f(\mathbf{x})-h(\mathbf{x})-a}{v_n} = \lim_{\substack{n\to\infty \\ n\in\mathcal{M}}} \left(\sum_{i\in Q} v_{ni}\gamma\left(\frac{\|\mathbf{x}-\mathbf{c}_i\|}{b_i}\right) + \sum_{i\in L_{bc}}\sum_{j\in K_i}\frac{w_{jn}}{v_n}\phi_{jn}(\mathbf{x}) + \sum_{i\in L_b\cup L_c\cup L}\frac{w_{in}a_i}{v_n}\right)$. So,

$$\sum_{i\in Q} v_i\gamma\left(\frac{\|\mathbf{x}-\mathbf{c}_i\|}{b_i}\right) = -\lim_{\substack{n\to\infty \\ n\in\mathcal{M}}}\left(\sum_{i\in L_{bc}}\sum_{j\in K_i}\frac{w_{jn}}{v_n}\phi_{jn}(\mathbf{x}) + \sum_{i\in L_b\cup L_c\cup L}\frac{w_{in}}{v_n}a_i\right). \tag{1}$$

It follows from the properties of total differential that either the limit on the right side of (1) is not a finite function or there exists $u\in\mathcal{R}$ and $\mathbf{u}_j = (u_{j0},\ldots,u_{jd})\in\mathcal{R}^{d+1}$ such that the limit is equal to

$$\sum_{i\in L_{bc}}\sum_{j\in K_i} u_{j0}\left(\frac{\partial^{k_j}\gamma\left(\frac{\|\mathbf{x}-\mathbf{c}_i\|}{b_i}\right)}{\partial {b_i}^{k_j}} + \sum_{l=1}^{d} u_{jl}\frac{\partial^{k_j}\gamma\left(\frac{\|\mathbf{x}-\mathbf{c}_i\|}{b_i}\right)}{\partial {c_{il}}^{k_j}}\right) + u,$$

where $\sum_{i\in L_{bc}}\sum_{j\in\hat{K}_i}\le m$, where $\hat{K}_i = \{j\in K_i; \mathbf{u}_j\neq\mathbf{0}\}$. So, by Lemma 3.2 $\sum_{i\in Q}\gamma\left(\frac{\|\mathbf{x}-\mathbf{c}_i\|}{b_i}\right)\left(v_i + \sum_{j\in K_i}\left(\frac{u_{j0}p_{k_j}(\|\mathbf{x}-\mathbf{c}_i\|,b_i)}{b^{3k_j}} + \sum_{l=1}^{d}\frac{u_{jl}q_{k_j}(x_l-c_l,b_i)}{b^{2k_j}}\right)\right) = u$.

Since all the pairs $(b_i,\mathbf{c}_i)$ are mutually different, an argument similar to the one we used in [4] in the proof of Theorem 2.3, based on Schwartz's inequality and asymptotic properties of $\exp(x)$, shows that this can only happen when either all $v_i = 0$ or $Q = \emptyset$. Since there exists i_0 such that for every $n\in\mathcal{M}$ $|v_{i_0n}| = 1$, we have $Q = \emptyset$. Putting $\phi(\mathbf{x}) = a_0 + \lim_{\substack{n\to\infty \\ n\in\mathcal{M}}}\left(\sum_{i\in P}\sum_{j\in K_i} w_{jn}\phi_{jn}(\mathbf{x}) + \sum_{i\in L_b\cup L_c\cup L} w_{in}a_i\right)$, we get $f = h+\phi$. It follows from properties of total differential that $\phi\in\Phi_d(\gamma,k)$.

References

1. C. K. Chui, X. Li, H. N. Mhaskar: Neural networks for localized approximation. *Mathematics of Computation* (in press).
2. V. Kůrková: Approximation of functions by perceptron networks with bounded number of hidden units. *Neural Networks* (in press).
3. V. Kůrková, K. Hlaváčková: Approximation of continuous functions by RBF and KBF networks. In *Proceedings of ESANN'94* (pp. 167-174). D facto: Brussels (1994).
4. V. Kůrková, R. Neruda: Uniqueness of functional representations by Gaussian basis function networks. In *Proceedings of ICANN'94* (pp. 471-474). Springer: London (1994).
5. J. Park, I.W. Sandberg: Universal approximation using radial-basis-function networks. *Neural Computation*, 3, 246-257 (1991).
6. J. Park, I.W. Sandberg: Approximation and radial-basis-function networks. *Neural Computation*, 5, 305-316 (1993).
7. F.G. Simmons: *Introduction to Topology and Modern Analysis*. New York: McGraw-Hill (1963).