Communicated by Halbert White

Universal Approximation Using Radial-Basis-Function Networks

J. Park
I. W. Sandberg
Department of Electrical and Computer Engineering,
University of Texas at Austin, Austin, Texas 78712 USA

There have been several recent studies concerning feedforward networks and the problem of approximating arbitrary functionals of a finite number of real variables. Some of these studies deal with cases in which the hidden-layer nonlinearity is not a sigmoid. This was motivated by successful applications of feedforward networks with nonsigmoidal hidden-layer units.

This paper reports on a related study of radial-basis-function (RBF) networks, and it is proved that RBF networks having one hidden layer are capable of universal approximation. Here the emphasis is on the case of typical RBF networks, and the results show that a certain class of RBF networks with the same smoothing factor in each kernel node is broad enough for universal approximation.

1 Introduction

There have been several recent studies concerning the capabilities of multilayered feedforward neural networks. Particularly pertinent to this paper are results that show that certain classes of neural networks are capable of providing arbitrarily good approximations to prescribed functionals of a finite number of real variables. From the theoretical point of view, these studies are important, because they address the question of whether a satisfactory solution is yielded by some member of a given class of networks. More specifically, suppose we have a problem that we want to solve using a certain type of neural network. Suppose also that there exists a decision function $f : \Re^r \rightarrow \Re^m$ whose implementation as a network plays a central role in the solution of the problem. Imagine that we have a family G of functions mapping $\Re^r$ to $\Re^m$ characterized by a certain structure and having certain elements (e.g., one might consider a set of multilayered perceptrons), and that we hope to solve the problem by implementing some satisfactory member of G. The first question we need to consider might be: Is this family G broad enough to contain f or a good approximation of f? Obviously, attempts to solve the problem without considering this question might be very time-consuming and might even be fruitless.

Several papers address this question for the case of multilayered perceptron models with sigmoidal nonlinearities, and affirmative answers have been obtained by showing that in a satisfactory sense the family G considered can actually approximate any decision function drawn from a certain large class (Cybenko 1989; Hornik *et al.* 1989).

At the present time, with the advantages and limitations of multilayered perceptron networks more transparent and with results containing comparative studies becoming available (e.g., Lippman 1989), research concerning different types of feedforward networks is very active. Among the various kinds of promising networks are the so-called radial-basis-function (RBF) networks (Lippman 1989). The block diagram of a version of an RBF classifier with one hidden layer is shown in Figure 1. Each unit in the hidden layer of this RBF network has its own centroid, and for each input $x = (x_1, x_2, \ldots, x_r)$, it computes the distance between x and its centroid. Its output (the output signal at one of the kernel nodes) is some nonlinear function of that distance. Thus, each kernel node in the RBF network computes an output that depends on a radially symmetric function, and usually the strongest output is obtained when the input is near the centroid of the node.

Assuming that there are r input nodes and m output nodes, the overall

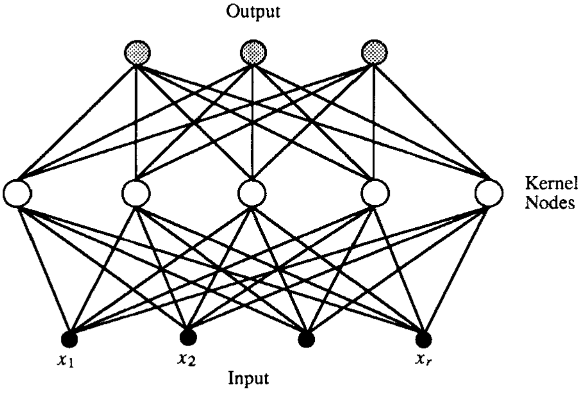

Figure 1: A radial-basis-function network.

response function without considering nonlinearity in an output node has the following form:

$$\sum_{i=1}^{M} W_i \cdot K\left(\frac{x - z_i}{\sigma_i}\right) = \sum_{i=1}^{M} W_i \cdot g\left(\frac{\| x - z_i \|}{\sigma_i}\right) \tag{1.1}$$

where $M \in \mathcal{N}$ the set of natural numbers is the number of kernel nodes in the hidden layer, $W_i \in \Re^m$ is the vector of weights from the ith kernel node to the output nodes, x is an input vector (an element of $\Re^r$), K is a radially symmetric kernel function of a unit in the hidden layer, z_i and σ_i are the centroid and smoothing factor (or width) of the ith kernel node, respectively, and $g : [0, \infty) \to \Re$ is a function called the activation function, which characterizes the kernel shape.

A gaussian function is often used as an activation function, and the smoothing factors of kernel nodes may be the same or may vary across nodes.

In this paper, RBF networks having the representation 1.1 are studied. Strong results are obtained to the effect that, under certain mild conditions on the kernel function K (or the activation function g), RBF networks represented by 1.1 with the *same* σ_i in each kernel node have the capability of universal approximation. Cybenko (1989) also considers feedforward networks with a single hidden layer of kernel functions. However, only L^1 approximation is considered in the corresponding part of Cybenko (1989), and only the case in which the smoothing factors can vary across nodes is addressed. A detailed comparison is given in Section 3.

This paper is organized as follows: In Section 2 our main results are presented, and in Section 3 a discussion of our results is given.

2 Main Results

In this section, we consider the approximation of a function by some element of a specific family of RBF networks.

Throughout the paper, we use the following notation and definitions, in which $\mathcal{N}$, $\Re$ and $\Re^r$ denote the set of natural numbers, the set of real numbers, and the set of real r-vectors, respectively. Let $L^p(\Re^r)$, $L^\infty(\Re^r)$, $C(\Re^r)$, and $C_c(\Re^r)$, respectively, denote the usual spaces of $\Re$-valued maps f defined on $\Re^r$ such that f is pth power integrable, essentially bounded, continuous, and continuous with compact support. The usual L^p and L^∞ norms are denoted by $\| \cdot \|_p$ and $\| \cdot \|_\infty$, respectively. The integral of $f \in L^1(\Re^r)$ over a Lebesgue measurable set A in $\Re^r$ is written as $\int_A f(x)dx$ or, if f is a function of several variables and, say, $f(a, \cdot) \in L^1(\Re^r)$ we write $\int_A f(a, x)dx$ to denote the integral of $f(a, \cdot)$ over A. The convolution operation is denoted by " $*$," and the characteristic function of a Lebesgue measurable subset A of $\Re^r$ is written as 1_A.

The family of RBF networks considered here consists of functions $q : \Re^r \rightarrow \Re$ represented by

$$q(x) = \sum_{i=1}^{M} w_i \cdot K\left(\frac{x - z_i}{\sigma}\right) \tag{2.1}$$

where $M \in \mathcal{N}$, $\sigma > 0$, $w_i \in \Re$, and $z_i \in \Re^r$ for $i = 1.\ldots.M$. We call this family S_K.

Note that 2.1 is the same as 1.1, with the exception that the smoothing factors in all kernel nodes are same, and the output space is $\Re$ instead of $\Re^m$. It will become clear that the extension of our results to multidimensional output spaces is trivial, and so we consider only a one-dimensional output space.

We will use the following result, which is a slight modification of a theorem in (Bochner and Chandrasekharan 1949, p. 101).

Lemma 1. *Let $f \in L^p(\Re^r)$, $p \in [1, \infty)$, and let $\phi : \Re^r \rightarrow \Re$ be an integrable function such that $\int_{\Re^r} \phi(x)dx = 1$. Define $\phi_\epsilon : \Re^r \rightarrow \Re$ by $\phi_\epsilon(x) = (1/\epsilon^r)\phi(x/\epsilon)$ for $\epsilon > 0$. Then $\| \phi_\epsilon * f - f \|_p \rightarrow 0$ as $\epsilon \rightarrow 0$.*

Proof. Note that $\phi_\epsilon \in L^1(\Re^r)$. By a direct extension from $\Re$ to $\Re^r$ of a standard theorem in analysis (Bochner and Chandrasekharan 1949, p. 99), one has $\phi_\epsilon * f \in L^p(\Re^r)$, which is used below.

By a change of variable,

$$(\phi_\epsilon * f)(\alpha) = \int_{\Re^r} f(\alpha - x)\phi_\epsilon(x)dx = \int_{\Re^r} f(\alpha - \epsilon x)\phi(x)dx$$

Thus,

$$| (\phi_\epsilon * f)(\alpha) - f(\alpha) | \quad = \quad | \int_{\Re^r} [f(\alpha - \epsilon x) - f(\alpha)]\phi(x)dx |$$

With q defined by $1/p + 1/q = 1$,

$$\begin{aligned} \| \phi_\epsilon * f - f \|_p \leq & \sup_{h \in L^q(\Re^r), \|h\|_q = 1} \int_{\Re^r} | h(\alpha) | \int_{\Re^r} | f(\alpha - \epsilon x) - f(\alpha) | \\ & \cdot | \phi(x) | \, dx d\alpha \\ = & \sup_{h \in L^q(\Re^r), \|h\|_q = 1} \int_{\Re^r} | \phi(x) | \int_{\Re^r} | h(\alpha) | \\ & \cdot | f(\alpha - \epsilon x) - f(\alpha) | \, d\alpha dx \\ \leq & \int_{\Re^r} | \phi(x) | \cdot \| f(\cdot - \epsilon x) - f(\cdot) \|_p \, dx \end{aligned}$$

by Fubini's theorem and Holder's inequality.

Since $\| f(\cdot - \epsilon x) - f(\cdot) \|_p \leq 2 \| f \|_p$ and translation is continuous in $L^p(\Re^r)$ (see Bochner and Chandrasekharan 1949, p. 98, and consider its direct extension to $\Re^r$) we have

$$\| \phi_\epsilon * f - f \|_p \rightarrow 0 \quad \text{as} \quad \epsilon \rightarrow 0$$

by Lebesgue's dominated convergence theorem. This proves the lemma.

Our Theorem 1 (below) establishes that, under certain mild conditions on the kernel function K, RBF networks represented by 2.1 are capable of approximating arbitrarily well any function in $L^p(\Re^r)$.

Theorem 1. *Let $K : \Re^r \rightarrow \Re$ be an integrable bounded function such that K is continuous almost everywhere and $\int_{\Re^r} K(x)dx \neq 0$. Then the family S_K is dense in $L^p(\Re^r)$ for every $p \in [1, \infty)$.*

Proof. Let $p \in [1, \infty), f \in L^p(\Re^r)$, and $\epsilon > 0$.

Since $C_c(\Re^r)$ is dense in $L^p(\Re^r)$ (Rudin 1986, p. 69), there exists an $f_c \in C_c(\Re^r)$ such that $\| f_c - f \|_p < \epsilon/3$. We will assume below that f_c is nonzero. Notice that this involves no loss of generality.

Let $\phi : \Re^r \rightarrow \Re$ be defined by $\phi(x) = (1/\int_{\Re^r K(\alpha)d\alpha}) \cdot K(x)$, for $x \in \Re^r$. Then ϕ satisfies the conditions on ϕ in Lemma 1. Thus, by defining $\phi_\sigma : \Re^r \rightarrow \Re$ as in Lemma 1, we obtain $\| \phi_\sigma * f_c - f_c \|_p \rightarrow 0$ as $\sigma \rightarrow 0$. Therefore, there is a positive σ such that $\| \phi_\sigma * f_c - f_c \|_p < \epsilon/3$.

Since f_c has compact support, there exists a positive T such that $\text{supp} f_c \subset [-T, T]^r$. Note that $\phi_\sigma(\alpha - \cdot)f_c(\cdot)$ is Riemann integrable on $[-T, T]^r$, because it is continuous almost everywhere and is bounded by $\| \phi_\sigma \|_\infty \cdot \| f_c \|_\infty$.

Define $v_n : \Re^r \rightarrow \Re$ by

$$v_n(\alpha) = \sum_{i=1}^{n^r} \phi_\sigma(\alpha - \alpha_i) f_c(\alpha_i) \left(\frac{2T}{n} \right)^r$$

where the set $\{\alpha_i \in \Re^r : i = 1, 2, \ldots, n^r\}$ consists of all points in $[-T, T]^r$ of the form $[-T + (2i_1T/n), \ldots, -T + (2i_rT/n)]$, $i_1, i_2, \ldots, i_r = 1, 2, \ldots, n$. Note that $v_n(\alpha)$ is a Riemann sum for $\int_{[-T,T]^r} \phi_\sigma(\alpha - x)f_c(x)dx$, and $\int_{[-T,T]^r} \phi_\sigma(\alpha - x)f_c(x)dx = \int_{\Re^r} \phi_\sigma(\alpha - x)f_c(x)dx = (\phi_\sigma * f_c)(\alpha)$. Thus, for any $\alpha \in \Re^r, v_n(\alpha) \rightarrow (\phi_\sigma * f_c)(\alpha)$ as $n \rightarrow \infty$. Since $\phi_\sigma * f_c \in L^p(\Re^r)$, there is a positive T_1 such that

$$\int_{\Re^r \setminus [-T_1, T_1]^r} | (\phi_\sigma * f_c)(\alpha) |^p \, d\alpha < (\epsilon/9)^p$$

Since ϕ_σ is bounded and $\phi_\sigma \in L^1(\Re^r)$, we have $\phi_\sigma \in L^p(\Re^r)$. Thus, there exists $T_2 > 0$ such that

$$\int_{\Re^r \setminus [-T_2, T_2]^r} | \phi_\sigma(\alpha) |^p \, d\alpha < \left(\frac{\epsilon}{9 \| f_c \|_\infty (2T)^r} \right)^p$$

Note that $\mid v_n(\alpha) \mid \leq \| f_c \|_\infty (2T)^r(1/n^r)\sum_{i=1}^{n^r} \mid \phi_\sigma(\alpha-\alpha_i) \mid$. By Jensen's inequality (Rudin 1986, p. 62),

$$\left[\frac{1}{n^r}\sum_{i=1}^{n^r} \mid \phi_\sigma(\alpha-\alpha_i) \mid\right]^p \leq \frac{1}{n^r}\sum_{i=1}^{n^r} \mid \phi_\sigma(\alpha-\alpha_i) \mid^p$$

Therefore,

$$\mid v_n(\alpha) \mid^p \leq [\| f_c \|_\infty (2T)^r]^p \frac{1}{n^r}\sum_{i=1}^{n^r} \mid \phi_\sigma(\alpha-\alpha_i) \mid^p$$

Define $T_0 = \max(T_1, T_2 + T)$. Using $\mid \alpha_{ij} \mid \leq T$ for all $j \in \{1, 2, \ldots, r\}$,

$$\int_{\Re^r\setminus[-T_0,T_0]^r} \mid \phi_\sigma(\alpha-\alpha_i) \mid^p d\alpha \leq \int_{\Re^r\setminus[-T_2,T_2]^r} \mid \phi_\sigma(\alpha) \mid^p d\alpha$$

and so

$$\int_{\Re^r\setminus[-T_0,T_0]^r} \mid v_n(\alpha) \mid^p d\alpha < (\epsilon/9)^p \tag{2.2}$$

Also,

$$\int_{\Re^r\setminus[-T_0,T_0]^r} \mid (\phi_\sigma * f_c)(\alpha) \mid^p d\alpha < (\epsilon/9)^p \tag{2.3}$$

because $T_0 \geq T_1$. Since $\phi_\sigma * f_c \in L^p(\Re^r)$ and $\| v_n \|_\infty \leq \| \phi_\sigma \|_\infty \| f_c \|_\infty (2T)^r$, one has

$$\int_{[-T_0,T_0]^r} \mid (\phi_\sigma * f_c)(\alpha) - v_n(\alpha) \mid^p d\alpha \to 0 \text{ as } n \to \infty$$

by the dominated convergence theorem. Thus, there is an $N \in \mathcal{N}$ for which

$$\int_{[-T_0,T_0]^r} \mid (\phi_\sigma * f_c)(\alpha) - v_N(\alpha) \mid^p d\alpha < (\epsilon/9)^p$$

Therefore, using 2.2 and 2.3,

$$\begin{aligned}\| v_N - \phi_\sigma * f_c \|_p &\leq \| v_N \cdot 1_{\Re^r\setminus[-T_0,T_0]^r} \|_p + \| (v_N - \phi_\sigma * f_c) \\ &\quad \cdot 1_{[-T_0,T_0]^r} \|_p + \| (\phi_\sigma * f_c) \cdot 1_{\Re^r\setminus[-T_0,T_0]^r} \|_p \\ &< \epsilon/9 + \epsilon/9 + \epsilon/9 = \epsilon/3\end{aligned}$$

From the above, $\| v_N - f \|_p < \epsilon$. Since

$$v_N(\cdot) = \sum_{i=1}^{N^r} \phi_\sigma(\cdot - \alpha_i) f_c(\alpha_i) \left(\frac{2T}{N}\right)^r = \sum_{i=1}^{N^r} w_i K\left(\frac{\cdot - \alpha_i}{\sigma}\right) \in S_K$$

with

$$w_i = \frac{1}{\sigma^r} f_c(\alpha_i) (\frac{2T}{N})^r \frac{1}{\int_{\Re^r} K(x)dx}$$

the proof is complete.

By K radially symmetric, we mean that $\| x \|_2 = \| y \|_2$ implies $K(x) = K(y)$. In this case, the activation function $g : [0, \infty) \to \Re$ is obtained by defining $g(d) = K(z)$, where z is any element of $\Re^r$ such that $\| z \|_2 = d$. Therefore, in the case of radial symmetry, 2.1 can be written as

$$q(x) = \sum_{i=1}^{M} w_i \cdot K\left(\frac{x - z_i}{\sigma}\right) = \sum_{i=1}^{M} w_i \cdot g\left(\frac{\| x - z_i \|_2}{\sigma}\right)$$

Note that there is no requirement of radial symmetry of the kernel function K in the above theorem. Thus, the theorem is stronger than necessary for RBF networks, and might be useful for other purposes. Similarly, in the following theorem and corollaries, radial symmetry of the kernel function K is not assumed, even though we are interested primarily in radial-basis-function networks.

If we interpret the term "radially symmetric" more generally than literally, then we may say that K is radially symmetric with respect to $\| \cdot \|$ if $\| x \| = \| y \|$ implies $K(x) = K(y)$, where $\| \cdot \|$ is some norm defined on $\Re^r$. With this generalization in mind, we sometimes use $\| x - z_i \|$ for the distance between x and z_i instead of $\| x - z_i \|_2$.

A slight modification of Theorem 1 given below addresses the case in which the function f we wish to approximate with an RBF network is not an element of $L^p(\Re^r)$, but an element of $L^p_{\text{loc}}(\Re^r)$ for some $p \in [1, \infty)$. Here the locally-L^p space $L^p_{\text{loc}}(\Re^r)$, $1 \leq p < \infty$ is defined as the set of all measurable $f : \Re^r \to \Re$ such that $f \cdot 1_{[-N,N]^r} \in L^p(\Re^r)$ for every $N \in \mathcal{N}$. One way to define a metric on $L^p_{\text{loc}}(\Re^r)$ is by

$$\rho_{\text{loc}}(f, g) = \sum_{n=1}^{\infty} 2^{-n} \cdot \frac{\| (f - g) \cdot 1_{[-n,n]^r} \|_p}{1 + \| (f - g) \cdot 1_{[-n,n]^r} \|_p}$$

The following is direct corollary of Theorem 1.

Corollary 1. Let $K : \Re^r \to \Re$ be an integrable bounded function such that K is continuous almost everywhere and $\int_{\Re^r} K(x)dx \neq 0$. Then the family S_K is dense in $L^p_{\text{loc}}(\Re^r)$ for every $p \in [1, \infty)$.

Proof. Let $p \in [1, \infty)$, $f \in L^p_{\text{loc}}(\Re^r)$, and $\epsilon > 0$. Choose $m \in \mathcal{N}$ such that $\sum_{n=m+1}^{\infty} 2^{-n} < \epsilon/2$.

Since $f \cdot 1_{[-m,m]^r} \in L^p(\Re^r)$, by Theorem 1 there is a $v \in S_K$ such that $\| f \cdot 1_{[-m,m]^r} - v \|_p < \epsilon/2$. Thus,

$$\begin{aligned} \rho_{\mathrm{loc}}(f,v) &\leq \sum_{n=m+1}^{\infty} 2^{-n} + \sum_{n=1}^{m} 2^{-n} \| (f-v)\cdot 1_{[-n,n]^r} \|_p \\ &< \epsilon/2 + \| (f-v)\cdot 1_{[-m,m]^r} \|_p < \epsilon \end{aligned}$$

which establishes the corollary.

Theorem 1 and Corollary 1 concern approximation with respect to the L^p metric or a metric induced by L^p metric. We next give a theorem concerning the approximation of continuous functions with respect to a metric induced by the uniform metric.

Theorem 2. *Let $K : \Re^r \to \Re$ be an integrable bounded function such that K is continuous and $\int_{\Re^r} K(x)dx \neq 0$. Then the family S_K is dense in $C(\Re^r)$ with respect to the metric d defined by*

$$d(f,g) = \sum_{n=1}^{\infty} 2^{-n} \frac{\| (f-g)\cdot 1_{[-n,n]^r} \|_\infty}{1 + \| (f-g)\cdot 1_{[-n,n]^r} \|_\infty}$$

Proof. Let $f : \Re^r \to \Re$ be any continuous function, and $\epsilon > 0$. Define $\phi : \Re^r \to \Re$ by normalizing K, and define $\phi_\sigma : \Re^r \to \Re$ for $\sigma > 0$ as in the proof of Theorem 1.

Pick a natural number m such that $2^{-m} < \epsilon/3$, and then choose a positive T such that $T > m$.

Since f is continuous on the compact set $[-m,m]^r$, we can obtain a nonzero continuous function $\tilde{f} : \Re^r \to \Re$ with the property that $\tilde{f}(x) = f(x)$ for $x \in [-m,m]^r$, and $\tilde{f}(x) = 0$ for $x \in \Re^r \setminus [-T,T]^r$. Note that $\tilde{f}$ is bounded and uniformly continuous.

Using $\phi \in L^1(\Re^r)$, pick a positive T_0 such that

$$\int_{\Re^r \setminus [-T_0,T_0]^r} | \phi(x) | \, dx < \frac{\epsilon}{12 \| \tilde{f} \|_\infty} \tag{2.4}$$

Since $\tilde{f}$ is uniformly continuous, there is a $\delta > 0$ for which $\| x - y \|_2 < \delta$ implies

$$| \tilde{f}(x) - \tilde{f}(y) | < \frac{\epsilon}{6 \| \phi \|_1} \tag{2.5}$$

Choose $\sigma > 0$ such that $\| \sigma x \|_2 < \delta$ for all $x \in [-T_0,T_0]^r$. Let $\alpha \in [-m,m]^r$. Then using 2.4 and 2.5,

$$\begin{aligned} |(\phi_\sigma * \tilde{f})(\alpha) - \tilde{f}(\alpha)| &\leq \int_{\Re^r} | \tilde{f}(\alpha - \sigma x) - \tilde{f}(\alpha) | \cdot | \phi(x) | \, dx \\ &\leq \int_{[-T_0,T_0]^r} | \tilde{f}(\alpha - \sigma x) - \tilde{f}(\alpha) | \cdot | \phi(x) | \, dx \\ &\quad + \int_{\Re^r \setminus [-T_0,T_0]^r} 2 \| \tilde{f} \|_\infty \cdot |\phi(x)| \, dx < \epsilon/3 \end{aligned} \tag{2.6}$$

Note that $(\phi_\sigma * \tilde{f})(\alpha) = \int_{[-T,T]^r} \phi_\sigma(\alpha - x)\tilde{f}(x)dx$. Define $v_n : \Re^r \to \Re$ by

$$v_n(\alpha) = \sum_{i=1}^{n^r} \phi_\sigma(\alpha - \alpha_i)\tilde{f}(\alpha_i)\left(\frac{2T}{n}\right)^r$$

where the set $\{\alpha_i \in \Re^r : i = 1, 2, \ldots, n^r\}$ consists of all points in $[-T, T]^r$ of the form $[-T + (2i_1T/n), \ldots, -T + (2i_rT/n)]$, $i_1, \ldots, i_r = 1, \ldots, n$.

Since the map $(s, x) \mapsto \phi_\sigma(s - x)\tilde{f}(x)$ is uniformly continuous on $[-m, m]^r \times [-T, T]^r$, there is a $\delta_0 > 0$ such that $s \in [-m, m]^r$, $x, y \in [-T, T]^r$ with $\| x - y \|_2 < \delta_0$ implies $| \phi_\sigma(s - x)\tilde{f}(x) - \phi_\sigma(s - y)\tilde{f}(y) | < \epsilon/3(2T)^r$. It easily follows that for $n > 2\sqrt{r}T/\delta_0$,

$$| v_n(\alpha) - \int_{[-T,T]^r} \phi_\sigma(\alpha - x)\tilde{f}(x)dx | < \epsilon/3 \tag{2.7}$$

Choose $N \in \mathcal{N}$ such that $N > 2\sqrt{r}T/\delta_0$. Then using 2.6 and 2.7,

$$| v_N(\alpha) - \tilde{f}(\alpha) | < 2\epsilon/3$$

in which $\alpha \in [-m, m]^r$ is arbitrary. Since $f(x) = \tilde{f}(x)$ for $x \in [-m, m]^r$,

$$\begin{aligned} d(v_N, f) &= \sum_{n=1}^{\infty} 2^{-n} \frac{\| (f - v_N)1_{[-n,n]^r} \|_\infty}{1 + \| (f - v_N)1_{[-n,n]^r} \|_\infty} \\ &\leq \| (\tilde{f} - v_N)1_{[-m,m]^r} \|_\infty + \sum_{n=m+1}^{\infty} 2^{-n} < \epsilon \end{aligned}$$

which finishes the proof.

The statement in Theorem 2 is equivalent to the statement that S_K is uniformly dense on compacta in $C(\Re^r)$ under the indicated conditions on K. That is, under the conditions on K of Theorem 2, for any continuous function $f : \Re^r \to \Re$, for any $\epsilon > 0$, and for any compact subset $C \subset \Re^r$, there exists a $q \in S_K$ such that $\| (q - f) \cdot 1_C \|_\infty < \epsilon$. Thus, by a useful relationship between uniform convergence on compacta and convergence in measure (Hornik *et al.* 1989, lemma 2.2), we have the following corollary:

Corollary 2. Let μ be a finite measure on $\Re^r$. Then under the conditions on K of Theorem 2, the family S_K is dense in $C(\Re^r)$ with respect to the metric ρ_μ defined by $\rho_\mu(f, g) = \inf\{\epsilon > 0 : \mu\{x \in \Re^r : | f(x) - g(x) | > \epsilon\} < \epsilon\}$.

3 Conclusions and Discussion

The results in Section 2 establish that under certain mild conditions on the kernel function, radial-basis-function networks having one hidden layer and the same smoothing factor in each kernel are broad enough for universal approximation. This provides an analytical basis for the design of neural networks using radial basis functions.

To the extent that the results of this paper bear on the approximation of a function in $L^1(\Re^r)$ with a finite sum $\sum_{i=1}^M w_i \cdot K(\cdot - z_i/\sigma_i)$ of kernel functions, there is some overlap of a part of Cybenko (1989) and this study. Using a theorem due to Wiener (Rudin 1973, p. 210) and the pertinent argument used in Cybenko (1989), it can be shown that the set $\{\sum_{i=1}^M w_i \cdot K(\cdot - z_i/\sigma_i) : M \in \mathcal{N}, w_i \in \Re, z_i \in \Re^r, \sigma_i \neq 0\}$ is dense in $L^1(\Re^r)$, under the condition that $K \in L^1(\Re^r)$ and $\int_{\Re^r} K(x)dx \neq 0$. This certainly shows the capability of certain RBF networks with respect to approximating an arbitrary L^1 function. However, note that here the smoothing factor σ_i in each kernel node has a full degree of freedom, that is, the σ_is can have different values across the kernel nodes. Thus, the major differences between this L^1 approximation and the results given in Section 2 concern the class of RBF networks considered as well as the metrics used.[1]

From the theoretical point of view, this condition concerning the same smoothing factor is often very important, because many studies are concerned with approximation using the functions $\sum_{i=1}^M w_i \cdot h(\| \cdot - z_i \|)$ (Broomhead and Lowe 1988; Powell 1985; Sun 1989), and radial basis functions with the same smoothing factor in each kernel node are often used in real applications (Broomhead and Lowe 1988). In connection with studies of approximation using radial basis functions, the recent results concerning the solvability of radial-function interpolation (Powell 1985; Sun 1989) are interesting, because they are directly applicable to the training of neural networks of the type we have focused attention on. These studies (Powell 1985; Sun 1989) are concerned with the interpolation of data by the m functions $h(\| \cdot - z_i \|)$, $i = 1, \ldots, m$, when the data (z_i, y_i) with $z_i \in \Re^r, y_i \in \Re, i = 1, \ldots, m$ are given. More precisely, the existence of a unique interpolant $\sum_{i=1}^m w_i \cdot h(\| \cdot - z_i \|)$ for distinct data (z_i, y_i) with $z_i \in \Re^r, y_i \in \Re, i = 1, \ldots, m$ has been shown for a certain class of pairs of h and $\| \cdot \|$. This existence leads us to an interesting observation: Suppose that training data $(z_i, y_i), i = 1, \ldots, m$ are given, where $z_i \in \Re^r, y_i = 1$ if $z_i \in A, y_i = -1$ if $z_i \in B$, and $A, B \subseteq \Re^r$ with $A \cap B = \emptyset$. From the given data, construct a new data set $z_i^* \in \Re^m, i = 1, \ldots, m$, by defining

$$z_i^* = \left[g\left(\frac{\| z_i - z_1 \|}{\sigma} \right), \ldots, g\left(\frac{\| z_i - z_m \|}{\sigma} \right) \right]$$

Note that $z_i^* \in \Re^m$, while $z_i \in \Re^r$. Then by the above existence property,

[1] In this connection, Wiener's theorem referred to above can also be used to give a direct proof that L^1 approximations can be achieved with linear combinations of translates of any element of $L^1(\Re^r)$ whose Fourier transform never vanishes. The gaussians $\exp(-\alpha \| \cdot \|_2^2)$ are examples of such functions.

for certain classes of g and $\| \cdot \|$, there exist $\lambda_j \in \Re, j = 1, \ldots, m$ such that for each $i \in \{1, 2, \ldots, m\}$.

$$y_i = \sum_{j=1}^{m} \lambda_j \cdot g\left(\frac{\| z_i - z_j \|}{\sigma}\right)$$

Thus, with $\Lambda = (\lambda_1, \lambda_2, \ldots, \lambda_m)^T$, $z_i^* \Lambda > 0$ if $z_i \in A$, and $z_i^* \Lambda < 0$ if $z_i \in B$. In other words, $\{(z_i^*, y_i) : i = 1, 2, \ldots, m\}$ is linearly separable in this case. Therefore, the perceptron learning rule suffices for the training of this network.

Additional related papers are (Hartman *et al.* 1990; Sandberg 1991). The work of Hartman *et al.* (1990), which appeared after this work was completed, considers gaussian functions and approximations on compact subsets of $\Re^r$ that are convex. It is shown there that networks with a single layer of gaussian units are universal approximators. In Sandberg (1991) more general results for gaussian functions are given as a special case of propositions concerning the uniform approximation of functionals defined on compact subsets of spaces that need not be finite dimensional. Also, it is observed in Sandberg (1991) that (what might be called) "function-space feedforward neural networks" with an input layer of bounded linear functionals and just one hidden nonlinear layer are universal approximators of real continuous functionals on compact subsets of a normed linear space.

Acknowledgments

This work was supported in part by the National Science Foundation under Grant MIP-8915335.

References

Bochner, S., and Chandrasekharan, K. 1949. *Fourier Transform.* Princeton University Press, Princeton, NJ.

Broomhead, D. S., and Lowe, D. 1988. Multi-variable functional interpolation and adaptive networks. *Complex Syst.* **2**, 321–355.

Cybenko, G. 1989. Approximation by superpositions of a sigmoidal function. *Math. Control, Signals, Syst.* **2**, 303–314.

Hartman, E. J., Keeler, J. D., and Kowalski, J. M. 1990. Layered neural networks with gaussian hidden units as universal approximations. *Neural Comp.* **2**, 210–215.

Hornik, K. M., Stinchcombe, M., and White, H. 1989. Multilayer feedforward networks are universal approximators. *Neural Networks* **2**, 359–366.

Lippman, R. P. 1989. Pattern classification using neural networks. *IEEE Commun. Mag.* **27**, 47–64.

Powell, M. J. D. 1985. Radial basis functions for multi-variable interpolation: A review. IMA Conference on Algorithms for the Approximation of Functions and Data, RMCS Shrivenham, UK.

Rudin, W. 1973. *Functional Analysis*. McGraw-Hill, New York.

Rudin, W. 1986. *Real and Abstract Analysis*, 3rd ed. McGraw-Hill, New York.

Sandberg, I. W. 1991. Gaussian basis functions and approximations for nonlinear systems. *Proceedings of the Ninth Kobe International Symposium on Electronics and Information Sciences*, Kobe, Japan.

Sun, X. 1989. On the solvability of radial function interpolation. *Approximation Theory VI* **2**, 643–646.

Received 17 September 1990; accepted 25 January 1991.