Kersten, D., 1990. Statistical limits to image understanding. In *Vision: Coding and Efficiency*, C. Blakemore, ed. Cambridge University Press, Cambridge.

Kohonen, T., 1984. *Self Organization and Associative Memory*, Springer-Verlag, Berlin.

Kullback, S., 1959. *Information Theory and Statistics*. Wiley, New York.

Linsker, R. 1989. An application of the principle of maximum information preservation to linear systems. In *Advances in Neural Information Processing Systems*, D. S. Touretzky, ed., Vol. 1, pp. 186–194. Morgan Kaufmann, San Mateo, CA.

Redlich, A. N. 1992. Supervised factorial learning. Preprint.

Shannon, C. E., and Weaver, W. 1949. *The Mathematical Theory of Communication*. The University of Illinois Press, Urbana.

Singh, J., 1966. *Great Ideas in Information Theory, Language and Cybernetics*, Chap. 16, Dover, New York.

Uttley, A. M., 1979. *Information Transmission in the Nervous System*. Academic Press, London.

von der Malsburg, C. 1973. Self-organization of orientation sensitive cells in the striate cortex. *Kybernetik* **14**, 85–100.

Received 12 December 1991; accepted 29 September 1992.

Communicated by Halbert White

Approximation and Radial-Basis-Function Networks

Jooyoung Park
Irwin W. Sandberg
Department of Electrical and Computer Engineering,
The University of Texas at Austin, Austin, TX 78712 USA

This paper concerns conditions for the approximation of functions in certain general spaces using radial-basis-function networks. It has been shown in recent papers that certain classes of radial-basis-function networks are broad enough for universal approximation. In this paper these results are considerably extended and sharpened.

1 Introduction

This paper concerns the approximation capabilities of radial-basis-function (RBF) networks. It has been shown in recent papers that certain classes of RBF networks are broad enough for universal approximation (Park and Sandberg 1991; Cybenko 1989). In this paper these results are considerably extended and sharpened.

Throughout this paper, we use the following definitions and notation, in which $\mathcal{N}$ and $\Re$ denote the natural numbers and the set of real numbers, respectively, and, for any positive integer r, $\Re^r$ denotes the normed linear space of real r-vectors with norm $\|\cdot\|$. $\langle\cdot,\cdot\rangle$ denotes the standard inner product in $\Re^r$. $L^p(\Re^r)$, $L^\infty(\Re^r)$, 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, and continuous with compact support. With $W \subset \Re^r$, $C(W)$ denotes the space of continuous $\Re$-valued maps defined on W. The usual L^p and uniform norms are denoted by $\|\cdot\|_p$ and $\|\cdot\|_\infty$, respectively. The characteristic function of a Lebesgue measurable subset A of $\Re^r$ is denoted by 1_A. The convolution operation is denoted by "$*$," and the Fourier transform (Stein and Weiss 1971) of a Fourier-transformable function f is written as $\hat{f}$. By a *cone* in $\Re^r$ we mean a set $C \subset \Re^r$ such that $x \in C$ implies that $\alpha x \in C$ for all $\alpha \geq 0$. By a *proper cone* we mean a cone that is neither empty nor the singleton $\{0\}$.

The block diagram of a typical RBF network 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

Neural Computation 5, 305–316 (1993) © 1993 Massachusetts Institute of Technology

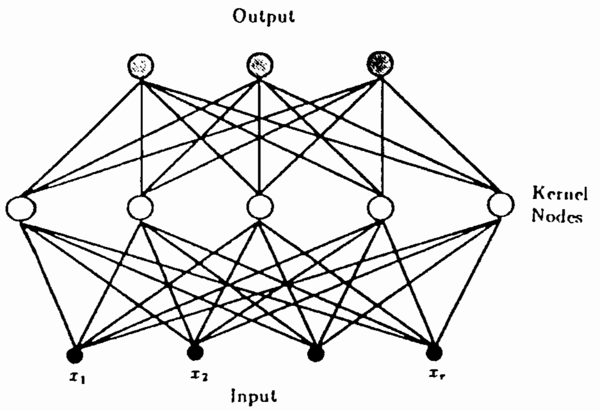

Figure 1: A radial-basis-function network.

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 at the centroid of the node. Each output node gives a weighted summation of the outputs of kernel nodes.

We first consider RBF networks represented by functions $q : \Re^r \rightarrow \Re$ of the form

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

where $M \in \mathcal{N}$ is the number of kernel nodes in the hidden layer, $w_i \in \Re$ is the weight from the ith kernel node to the output node, x is an input vector (an element of $\Re^r$), and K is the common radially symmetric kernel function of the units in the hidden layer. Here $z_i \in \Re^r$ and $\sigma > 0$ are the centroid and smoothing factor (or width) of the ith kernel node, respectively. We call this family $S_0(K)$. Note that the networks in this family have the same positive smoothing factor in each kernel node.

Families with a translation-invariant vector space structure are also often important. For example, networks are widely used in which the smoothing factors are positive real numbers as in $S_0(K)$, but can have different values across kernel nodes. This family is the smallest vector space among those containing $S_0(K)$ as a subset. We call this vector space $S_1(K)$. Its general element $q : \Re^r \rightarrow \Re$ is represented by

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

where $M \in \mathcal{N}$, $\sigma_i > 0$, $w_i \in \Re$, and $z_i \in \Re^r$ for $i = 1, 2, \ldots, M$.

For the sake of clarity and convenience, we consider only a one-dimensional output space instead of outputs represented by multiple nodes as in Figure 1. The extension of our results to multidimensional output spaces is trivial. Notice that the kernel function K characterizes the families $S_0(K)$ and $S_1(K)$, and that each kernel node has its output derived from K indexed by two parameters (the centroid and smoothing factor), one for position and the other for scale. Ordinarily K is radially symmetric with respect to the norm $\|\cdot\|$ in the sense that $\|x\| = \|y\|$ implies $K(x) = K(y)$. However, as we shall see in the next section, radial symmetry of the kernel function $K : \Re^r \rightarrow \Re$ is needed in the development of only one of the approximation results in this study. Except where indicated to the contrary, radial symmetry of the kernel function K is not assumed.

In Park and Sandberg (1991) it is shown that $S_0(K)$ is dense in $L^p(\Re^r)$, $p \in [1, \infty)$ if K is an integrable bounded function such that K is continuous almost everywhere and $\int_{\Re^r} K(x)\,dx \neq 0$. In Cybenko (1989) it is pointed out that a consequence of a generalization of a theorem due to Wiener is that the elements of a vector space related to $S_1(K)$ are capable of approximating functions in $L_1(\Re^r)$. The purpose of this paper is to report on a substantial sharpening of the results in Park and Sandberg (1991) and Cybenko (1989).

2 Approximation Results

As mentioned above, in Park and Sandberg (1991) it is shown that $S_0(K)$ is dense in $L^p(\Re^r)$, $p \in [1, \infty)$ if K is an integrable bounded function such that K is continuous almost everywhere and $\int_{\Re^r} K(x)\,dx \neq 0$. Our first theorem concerns the $p = 1$ case; a necessary and sufficient condition is given for approximation with $S_0(K)$.

Theorem 1. *Assuming that $K : \Re^r \rightarrow \Re$ is integrable, $S_0(K)$ is dense in $L^1(\Re^r)$ if and only if $\int_{\Re^r} K(x)\,dx \neq 0$.*

Proof. Suppose first that $\int_{\Re^r} K(x)\,dx \neq 0$, and define $J = |\int_{\Re^r} K(x)\,dx|$. Let $f \in L^1(\Re^r)$ and $\epsilon > 0$ be given. Since $C_c(\Re^r)$ is dense in $L^1(\Re^r)$ (Rudin 1987), we can choose a nonzero $f_c \in C_c(\Re^r)$ such that

$$\|f - f_c\|_1 < \epsilon/4 \tag{1}$$

Since f_c has a compact support, there exists a positive T such that

$$\operatorname{supp} f_c \subset [-T, T]^r$$

Choose a function $K_c \in C_c(\Re^r)$ such that

$$\|K - K_c\|_1 < \min\left[\frac{J}{2(2T)^r\|f_c\|_\infty}\frac{\epsilon}{4}, J/2\right] \tag{2}$$

Note that 2 implies

$$\left|\int_{\Re^r} K_c(x)\,dx\right| > J/2$$

because

$$\begin{aligned}\left|\int_{\Re^r} K_c(x)\,dx\right| &= \left|\int_{\Re^r} K(x)\,dx - \int_{\Re^r}[K(x) - K_c(x)]\,dx\right| \\ &\geq \left|\int_{\Re^r} K(x)\,dx\right| - \left|\int_{\Re^r}[K(x) - K_c(x)]\,dx\right| \\ &\geq J - \|K - K_c\|_1 > J/2\end{aligned}$$

Define $\phi : \Re^r \to \Re$ and $\phi_\sigma : \Re^r \to \Re$ for $\sigma > 0$ by

$$\phi(x) = \frac{1}{\int_{\Re^r} K_c(x)\,dx} K_c(x)$$

and

$$\phi_\sigma(x) = \frac{1}{\sigma^r} \cdot \phi\left(\frac{x}{\sigma}\right)$$

By Lemma 1 (in the appendix), we have

$$\|f_c - \phi_\sigma * f_c\|_1 \to 0 \qquad \text{as} \qquad \sigma \to 0$$

Choose $\sigma > 0$ such that

$$\|f_c - \phi_\sigma * f_c\|_1 < \epsilon/4 \tag{3}$$

Note that $\phi_\sigma(\alpha - \cdot)f_c(\cdot)$ is Riemann integrable on $[-T, T]^r$, since ϕ_σ and f_c are each continuous and bounded.

Define $v_n : \Re^r \to \Re$ 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

$$\left[-T + \frac{2i_1T}{n}, \ldots, -T + \frac{2i_rT}{n}\right], \qquad i_1, i_2, \ldots, i_r = 1, 2, \ldots, n$$

Note that $v_n(\alpha)$ is a Riemann sum for $\int_{[-T,T]^r} \phi_\sigma(\alpha - x) f_c(x)\,dx$, and that

$$(\phi_\sigma * f_c)(\alpha) = \int_{\Re^r} \phi_\sigma(\alpha - x) f_c(x)\,dx = \int_{[-T,T]^r} \phi_\sigma(\alpha - x) f_c(x)\,dx$$

Thus, for any $\alpha \in \Re^r$,

$$v_n(\alpha) \to (\phi_\sigma * f_c)(\alpha) \qquad \text{as} \qquad n \to \infty$$

Since $(\phi_\sigma * f_c)$ and the v_n are dominated by an integrable bounded function with compact support, by the dominated convergence theorem

$$\int_{\Re} |(\phi_\sigma * f_c)(\alpha) - v_n(\alpha)|\,d\alpha \to 0 \qquad \text{as} \qquad n \to \infty$$

Thus, there is an $N \in \mathcal{N}$ for which

$$\|\phi_\sigma * f_c - v_N\|_1 < \epsilon/4 \tag{4}$$

Note that

$$v_N(\alpha) = \frac{1}{N^r} \sum_{i=1}^{N^r} \frac{f_c(\alpha_i)(2T)^r}{\int_{\Re^r} K_c(\alpha)\,d\alpha} \frac{1}{\sigma^r} K_c\left(\frac{\alpha - \alpha_i}{\sigma}\right)$$

Since

$$\left\|\frac{1}{\sigma^r} K_c\left(\frac{\cdot - \alpha_i}{\sigma}\right) - \frac{1}{\sigma^r} K\left(\frac{\cdot - \alpha_i}{\sigma}\right)\right\|_1 = \|K_c - K\|_1$$

$\tilde{v}_N : \Re^r \to \Re$ defined by

$$\tilde{v}_N(\alpha) = \frac{1}{N^r} \sum_{i=1}^{N^r} \frac{f_c(\alpha_i)(2T)^r}{\int_{\Re^r} K_c(\alpha)\,d\alpha} \frac{1}{\sigma^r} K\left(\frac{\alpha - \alpha_i}{\sigma}\right)$$

has the property that

$$\|v_N - \tilde{v}_N\|_1 < \frac{2(2T)^r\|f_c\|_\infty}{J}\|K - K_c\|_1 \leq \frac{\epsilon}{4} \tag{5}$$

By equations 1, 3, 4, and 5,

$$\|f - \tilde{v}_N\|_1 < \epsilon$$

Since $\tilde{v}_N(\cdot) = \sum_{i=1}^{N^r} w_i \cdot K\frac{\cdot - \alpha_i}{\sigma} \in S_0(K)$ with

$$w_i = \frac{1}{\sigma^r} f_c(\alpha_i) \left(\frac{2T}{N}\right)^r \frac{1}{\int_{\Re^r} K_c(\alpha)\,d\alpha}$$

$S_0(K)$ is dense in $L^1(\Re^r)$.

To show the "only if" part, we prove the contrapositive: Assume that $\int_{\Re^r} K(x)\,dx = 0$. Then for any $f \in L^1(\Re^r)$ such that $\int_{\Re^r} f(x)\,dx \triangleq J > 0$, there is no $g \in S_0(K)$ satisfying $\|f - g\|_1 < J/2$, because

$$\|f - g\|_1 \geq \int_{\Re^r} |f(x) - g(x)]\,dx = \int_{\Re^r} f(x)\,dx = J$$

for $g \in S_0(K)$. Thus $S_0(K)$ is now not dense in $L^1(\Re^r)$, which completes the proof. □

Since the family $S_0(K)$ is a proper subset of $S_1(K)$, the "if" part of this theorem holds also with $S_0(K)$ replaced by $S_1(K)$. A family similar to $S_1(K)$ is considered in Cybenko (1989) with regard to the approximation of functions in $L^1(\Re^r)$; it is noted there that if $K \in L^1(\Re^r)$ and $\int_{\Re^r} K(x)\,dx \neq$

0, then the family $S_2(K)$ consisting of functions $q : \Re^r \to \Re$ of the following form is dense in $L^1(\Re^r)$:

$$q(x) = \sum_{i=1}^{M} w_i \cdot K(t_i x + y_i)$$

where $M \in \mathcal{N}$, $t_i \in \Re$, and $y_i \in \Re^r$ for $i = 1, \ldots, M$. The proof of this follows immediately from a generalization (Rudin 1973, Theorem 9.4) of a theorem due to Wiener. For the readers' convenience we state the theorem in the appendix. Here we make some pertinent observations:

1. $S_1(K)$ defined above is a proper subset of $S_2(K)$, and even $S_1(K)$ is dense in $L^1(\Re^r)$ under the condition $\int_{\Re^r} K(x)\,dx \neq 0$. This can be easily shown: Assume to get a contradiction that there is an $s_0 \in \Re^r$ such that $\hat{f}(s_0) = 0$ for all $f \in S_1(K)$. Then $\hat{K}(\sigma s_0) = 0$ for all $\sigma > 0$, since

$$\left[K\left(\frac{\cdot - z}{\sigma}\right)\right]^{\wedge}(s_0) = \sigma^r \exp(-2\pi i \langle z, s_0\rangle)\hat{K}(\sigma s_0)$$

   Since $\hat{K}$ is continuous, the above implies that

$$\hat{K}(0) = \int_{\Re^r} K(x)\,dx = 0$$

   which contradicts the nonzero-integral condition. The main difference between $S_1(K)$ and $S_2(K)$ is the set from which the smoothing factors are drawn. In this connection, it is easy to see that our conclusion here, and also in Theorem 1, can be strengthened in that they hold if our $\sigma_i > 0$ and $\sigma > 0$ conditions are replaced by the conditions that $\sigma_i \in S$ and $\sigma \in S$, where S is any subset of $(0, \infty)$ such that zero is a cluster point of S. Also, note that the denseness of $S_1(K)$ in $L^1(\Re^r)$ is a corollary of Theorem 1.

2. When $K : \Re^r \to \Re$ is integrable, $S_1(K)$ is dense in $L^1(\Re^r)$ only if $\int_{\Re^r} K(x)\,dx \neq 0$. The "only if" part of the proof of Theorem 1 shows this.

The above observations give the following theorem:

Theorem 2. *Assuming that $K : \Re^r \to \Re$ is integrable, $S_1(K)$ is dense in $L^1(\Re^r)$ if and only if $\int_{\Re^r} K(x)\,dx \neq 0$.*

Up to this point our results concern the approximation of functions in $L^1(\Re^r)$ under the condition that $\int_{\Re^r} K(x)\,dx \neq 0$. As shown above, this condition is necessary for approximation with $S_0(K)$ or $S_1(K)$. A natural question that arises is whether the nonzero-integral condition is necessary for approximation in $L^p(\Re^r)$, $p \in (1, \infty)$. We will see below that it is not necessary for $p = 2$.

In the following theorem, attention is focused on kernel functions $K : \Re^r \to \Re$ with the property that for all $M \subset \Re^r$ with positive measure there is a $\sigma > 0$ such that $\hat{K}(\sigma \cdot) \neq 0$ almost everywhere on some positive measure subset of M. We call such K *pointable*. We shall use the fact that the negation of this condition on K is that for some M of positive measure, $\hat{K}(\sigma \cdot) = 0$ almost everywhere on M for all $\sigma > 0$.

Theorem 3. *Assuming that $K : \Re^r \to \Re$ is a square integrable function, $S_1(K)$ is dense in $L^2(\Re^r)$ if and only if K is pointable.*

Proof. We make use of the following characterization of closed translation-invariant subspaces of $L^2(\Re^r)$, which is an easy modification of (Rudin, 1987, Theorem 9.17). □

Lemma 2. *Associate to each measurable set $E \subset \Re^r$ the linear space M_E of all $f \in L^2(\Re^r)$ such that $\hat{f} = 0$ almost everywhere on E. Then each M_E is a closed translation-invariant subspace of $L^2(\Re^r)$, and every closed translation-invariant subspace of $L^2(\Re^r)$ is M_E for some E.*

Consider any K satisfying the indicated conditions, and suppose that the closure of $S_1(K)$ is not $L^2(\Re^r)$. Then, since this closure is translation-invariant, by Lemma 2 there is a measurable subset E of $\Re^r$ having positive measure such that

$$\hat{f} = 0 \qquad \text{almost everywhere on } E$$

for any f in the closure of $S_1(K)$. In particular,

$$\sigma^r \exp(-2\pi i \langle z, \cdot\rangle)\hat{K}(\sigma \cdot) = 0 \qquad \text{almost everywhere on } E$$

for any $z \in \Re^r$ and $\sigma > 0$. Thus, $\hat{K}(\sigma \cdot) = 0$ almost everywhere on E for all $\sigma > 0$, which contradicts our supposition.

To show the "only if" part, we prove the contrapositive: Assume that there is a measurable set $M \subset \Re^r$ with positive measure such that

$$\hat{K}(\sigma \cdot) = 0 \qquad \text{almost everywhere on } M$$

for all $\sigma > 0$. Then for any $f \in L^2(\Re^r)$ with $J \triangleq \|\hat{f} 1_M\|_2 > 0$,[1] there is no $g \in S_1(K)$ satisfying $\|f - g\|_2 < J/2$, because

$$\begin{aligned}\left\| f - \sum_i w_i K\left(\frac{\cdot - z_i}{\sigma_i}\right)\right\|_2 &= \left\| \hat{f} - \sum_i w_i \sigma_i^r \exp(-2\pi i \langle z_i, \cdot\rangle)\hat{K}(\sigma_i \cdot)\right\|_2 \\ &\geq \|\hat{f} 1_M\|_2 = J\end{aligned}$$

This completes the proof. □

[1] Here we use $\|\cdot\|_2$ to denote also the usual norm on the space of *complex-valued* square-integrable functionals.

A large class of kernel functions satisfies the conditions of pointability. For example, kernel functions K such that $\hat{K} \neq 0$ almost everywhere on some ball centered at the origin are pointable. Note that this class includes functions K with $\int_{\Re^r} K(x)\,dx = 0$.

A result for the general $L^p(\Re^r)$ case along the lines of the "if part" of Theorem 2 is:

Proposition 1. *With $p \in (1,\infty)$, let $K : \Re^r \to \Re$ be an integrable function such that*

$$\int_{\Re^r} K(x)\,dx \neq 0$$

and

$$\int_{\Re^r} |K(x)|^p\,dx < \infty$$

Then $S_1(K)$ is dense in $L^p(\Re^r)$.

Proof. Suppose that $S_1(K)$ is not dense in $L^p(\Re^r)$. Then by the Hahn–Banach theorem (Rudin 1987), there exists a bounded linear functional Λ on $L^p(\Re^r)$ such that

$$\Lambda[\text{the closure of } S_1(K)] = \{0\} \tag{6}$$

but

$$\Lambda(L^p(\Re^r)) \neq \{0\}$$

By the Riesz representation theorem (Rudin 1987), $\Lambda : L^p(\Re^r) \to \Re$ can be represented by

$$\Lambda(f) = \int_{\Re^r} f(x) g_\Lambda(x)\,dx$$

for some function g_Λ in $L^q(\Re^r)$,[2] where q is the conjugate exponent of p defined by $1/p + 1/q = 1$. In particular, from equation 6

$$\int_{\Re^r} \frac{1}{\sigma^r} K\left(\frac{x-z}{\sigma}\right) g_\Lambda(x)\,dx = 0$$

for any $z \in \Re^r$ and $\sigma > 0$.

Define $\tilde{K} : \Re^r \to \Re$ and $\tilde{K}_\sigma : \Re^r \to \Re$ for $\sigma > 0$ by

$$\tilde{K}(x) = \frac{1}{\int_{\Re^r} K(x)\,dx} K(-x)$$

and

$$\tilde{K}_\sigma(x) = \frac{1}{\sigma^r} \tilde{K}\left(\frac{x}{\sigma}\right)$$

Note that for any $\sigma > 0$ and z in $\Re^r$,

$$(\tilde{K}_\sigma * g_\Lambda)(z) = \frac{1}{\int_{\Re^r} K(x)\,dx} \int_{\Re^r} \frac{1}{\sigma^r} K\left(\frac{x-z}{\sigma}\right) g_\Lambda(x)\,dx = 0 \tag{7}$$

Since $\tilde{K} \in L^1(\Re^r)$ and $\int_{\Re^r} \tilde{K}(x)\,dx = 1$, by Lemma 1 (in the appendix),

$$\|\tilde{K}_0 * g_\Lambda - g_\Lambda\|_q \to 0 \qquad \text{as} \qquad \sigma \to 0 \tag{8}$$

By 7 and 8, we conclude that g_Λ is zero almost everywhere. This implies that Λ is the zero functional, which contradicts our supposition.

Our focus has been on L^p approximation. We next give a theorem concerning the uniform approximation of continuous functions on compact subsets of $\Re^r$.

Theorem 4. *Let $K : \Re^r \to \Re$ be an integrable function such that K is continuous and such that $\hat{K}^{-1}(0)$ includes no proper cone.[3] Then $S_1(K)$ is dense in $C(W)$ with respect to the norm $\|\cdot\|_\infty$ for any compact subset W of $\Re^n$.*

Proof. Consider any compact subset W of $\Re^r$. Suppose that $S_1(K)$ is not dense in $C(W)$. Then proceeding as in the proof of Proposition 1, we see that there is a nonzero finite signed measure μ that is concentrated on W and that satisfies

$$\int_{\Re^r} K\left(\frac{x-z}{\sigma}\right) d\mu(x) = \int_W K\left(\frac{x-z}{\sigma}\right) d\mu(x) = 0 \tag{9}$$

for any $z \in \Re^r$ and $\sigma > 0$.

With $z \in \Re^r$, $\sigma > 0$, and any function $h \in L^1(\Re^r) \cap L^\infty(\Re^r)$ whose Fourier transform has no zeros[4] (e.g., the gaussian function $\exp(-\alpha\|\cdot\|_2^2)$ with $\alpha > 0$), consider the integral

$$\int_{\Re^r} \int_{\Re^r} K\left(\frac{y+x-z}{\sigma}\right) h(x)\,dx\,d\mu(y)$$

Note that

$$\int_{\Re^r} \int_{\Re^r} \left|K\left(\frac{y+x-z}{\sigma}\right) h(x)\right| dx\,d|\mu|(y) \leq \sigma^r \|K\|_1 \|h\|_\infty |\mu|(\Re^r) < \infty$$

where $|\mu|$ is the total variation of μ.

By equation 9 and Fubini's theorem (see, e.g., Rudin 1962), we have

$$\begin{aligned} 0 &= \int_{\Re^r} \left[\int_{\Re^r} K\left[\frac{y-(z-x)}{\sigma}\right] d\mu(y)\right] h(x)\,dx \\ &= \int_{\Re^r} \left[\int_{\Re^r} K\left(\frac{x+y-z}{\sigma}\right) h(x)\,dx\right] d\mu(y) \end{aligned} \tag{10}$$

[2] The strategy of using the Hahn–Banach theorem together with representations of linear functionals was first used in the neural-networks literature in Cybenko (1989).

[3] Since $\hat{K}(-w)$ equals the conjugate of $\hat{K}(w)$ for any w in $\Re^r$, this condition can be stated in terms of *subspaces* instead of *cones*.

[4] Here we use a strategy along the lines of Hornick (1991, proof of Theorem 5).

By the change of variable $x+y \rightarrow x$ and Theorem 1:4.5 of Petersen (1983), equation 10 is equivalent to

$$\begin{aligned} 0 &= \int_{\Re^r} \left[\int_{\Re^r} K\left(\frac{x-z}{\sigma}\right) h(x-y)\, dx \right] d\mu(y) \\ &= \int_{\Re^r} K\left(\frac{x-z}{\sigma}\right) \left[\int_{\Re^r} h(x-y)\, d\mu(y) \right] dx \\ &= \int_{\Re^r} K\left(\frac{x-z}{\sigma}\right) (h*\mu)(x)\, dx. \qquad (11) \end{aligned}$$

Note that $h*\mu$ is integrable (by Theorem 1:4.5 of Petersen 1983). It is also essentially bounded, because

$$\begin{aligned} |(h*\mu)(x)| &\leq \int_{\Re^r} |h(x-y)|\, d|\mu|(y) \\ &\leq \|h\|_\infty |\mu|(\Re^r) \end{aligned}$$

for almost all $x \in \Re^r$.

Consider the closed translation-invariant subspace I of $L^1(\Re^r)$ defined as the L^1-closure of $S_1(K)$. By equation 11 and the essential boundedness of $h*\mu$, it easily follows that

$$\int_{\Re^r} f(x)(h*\mu)(x)\, dx = 0 \qquad (12)$$

for any f in I. Following the notation in Rudin (1962), define the zero set $Z(I)$ of I to be the set of w where the Fourier transforms of all functions in I vanish. We claim that a nonzero element in $\Re^r$ cannot be a member of $Z(I)$ when $\hat{K}^{-1}(0)$ includes no proper cone. Assume to get a contradiction that $w \neq 0$ and $w \in Z(I)$. Then, using the definition of $Z(I)$,

$$K\left(\frac{\cdot - z}{\sigma}\right)^{\wedge} (w) = \sigma^r \exp(-2\pi i \langle z, w\rangle) \hat{K}(\sigma w) = 0$$

for any $z \in \Re^r$ and $\sigma > 0$. This implies that

$$\hat{K}(\sigma w) = 0 \qquad \text{for all } \sigma > 0$$

Since $\hat{K}$ is continuous, this means that $\hat{K}^{-1}(0)$ includes the cone $\{\sigma w \in \Re^r : \sigma \geq 0\}$, which contradicts the cone condition. Thus, $Z(I)$ is either the empty set or $\{0\}$. In either case, by Theorems 7.1.2 and 7.2.4 of Rudin (1962), any integrable function from $\Re^r$ to $\Re$ with zero integral is a member of I. Thus, equation 12 gives

$$\int_{\Re^r} f(x)(h*\mu)(x)\, dx = 0 \qquad (13)$$

for any f in $L^1(\Re^r)$ wth $\int_{\Re^r} f(x)\, dx = 0$.

Note that the property 13 can hold only for $h*\mu$ in the class of almost everywhere constant functions. But since $h*\mu \in L^1(\Re^r)$ and zero is the only constant function in $L^1(\Re^r)$, we have

$$h*\mu = 0 \qquad \text{almost everywhere.} \qquad (14)$$

Since $\hat{h}$ has no zeros, by Theorem 2:2.2 of Petersen (1983) and Theorem 1.3.6 of Rudin (1962), equation 14 implies $\mu = 0$. This contradicts our supposition, and thus proves the theorem. □

A corollary of this theorem is that $S_1(K)$ is dense in $C(W)$ for any compact subset W of $\Re^r$ when the kernel $K : \Re^r \rightarrow \Re$ is integrable, continuous and satisfies $\int_{\Re^r} K(x)\, dx \neq 0$.

Finally, when $K : \Re^r \rightarrow \Re$ is integrable and radially symmetric with respect to the Euclidean norm, $\hat{K}$ is also radially symmetric with respect to the Euclidean norm (Bochner and Chandrasekharan 1949, p. 69). In this setting, every K not equivalent to the zero element of $L^1(\Re^r)$ satisfies the cone condition of Theorem 4. This observation gives the following:

Theorem 5. *Let $K : \Re^r \rightarrow \Re$ be a nonzero integrable function such that K is continuous and radially symmetric with respect to the Euclidean norm. Then $S_1(K)$ is dense in $C(W)$ with respect to the norm $\|\cdot\|_\infty$ for any compact subset W of $\Re^r$.*

3 Concluding Remarks

The results in this paper significantly improve previous results. In particular, we have given sharp conditions on the kernel function under which radial-basis-function networks having one hidden layer are capable of universal approximation on $\Re^r$ or on compact subsets of $\Re^r$. A related result concerning uniform approximation using the elements of $S_0(K)$ with integrable K is given in Park and Sandberg (1991, p. 254).

The results in Section 2 concern the approximation of real-valued functions. Approximations for complex-valued functions are also of interest. In this connection, it is a straightforward exercise to verify that Theorems 1–5 and Proposition 1 remain true if "$K : \Re^r \rightarrow \Re$" is replaced with the condition that K maps $\Re^r$ into the set C of complex numbers, $L^p(\Re^r)$ denotes instead the corresponding space of C-valued functions, the elements of $C(W)$ are taken to be C-valued, and $S_0(K)$ and $S_1(K)$ refer instead to the corresponding sets in which the weights w_i are drawn from C.

An important problem we have not addressed is that of determining the network parameters so that a prescribed degree of approximation is achieved.

Appendix

Lemma 1. [5] *Let $f \in L^p(\Re^r)$, $p \in [1, \infty)$, and let $\phi : \Re^r \to \Re$ be an integrable function such that*

$$\int_{\Re^r} \phi(x)\,dx = 1$$

Define $\phi_\epsilon : \Re^r \to \Re$ by

$$\phi_\epsilon(x) = (1/\epsilon^r)\phi(x/\epsilon)$$

for $\epsilon > 0$. Then $\|\phi_\epsilon * f - f\|_p \to 0$ as $\epsilon \to 0$.

Theorem 9.4 of Rudin (1973). *If Y is a closed translation-invariant subspace of $L^1(\Re^r)$, and if*

$$Z(Y) = \cap_{f \in Y}\{s \in \Re^r : \hat{f}(s) = 0\}$$

is empty, then $Y = L^1(\Re^r)$.

References

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

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

Hornik, K. 1991. Approximation capabilities of multilayer feedforward networks. *Neural Networks* 4, 251–257.

Park, J., and Sandberg, I. W. 1991. Universal approximation using radial-basis-function networks. *Neural Comp.* 3, 246–257.

Petersen, B. E. 1983. *Introduction to the Fourier Transform and Pseudo-Differential Operators.* Pitman, Marshfield, MA.

Rudin, W. 1962. *Fourier Analysis on Groups.* Interscience Publishers, New York.

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

Rudin, W. 1987. *Real and Complex Analysis,* 3rd ed. McGraw-Hill, New York.

Stein, E. M., and Weiss, G. 1971. *Introduction to Fourier Analysis on Euclidean Spaces.* Princeton University Press, Princeton.

Received 13 April 1992; accepted 3 September 1992.

[5] This lemma is used in Park and Sandberg (1991) where it is observed to be a slight modification of a theorem in Bochner and Chandrasekharan (1949). We have since found earlier proofs of the lemma (e.g., Petersen 1983, p. 72).

Communicated by Richard Lippmann

A Polynomial Time Algorithm for Generating Neural Networks for Pattern Classification: Its Stability Properties and Some Test Results

Somnath Mukhopadhyay
Asim Roy
Lark Sang Kim
Sandeep Govil
Department of Decision and Information Systems,
Arizona State University, Tempe, AZ 85287 USA

Polynomial time training and network design are two major issues for the neural network community. A new algorithm has been developed that can learn in polynomial time and also design an appropriate network. The algorithm is for classification problems and uses linear programming models to design and train the network. This paper summarizes the new algorithm, proves its stability properties, and provides some computational results to demonstrate its potential.

1 Introduction

One of the critical issues in the field of neural networks is the development of polynomial time algorithms for neural network training. With the advent of polynomial time methods (Karmarkar 1984; Khachian 1979 and others), linear programming has drawn increased attention for its potential for training neural networks in polynomial time (Glover 1990; Mangasarian *et al.* 1990; Bennett *et al.* 1992; Roy and Mukhopadhyay 1991; Roy *et al.* 1992). This paper presents the method of Roy and Mukhopadhyay (1991) and Roy *et al.* (1992) in summary form and proves its stability properties under translation and rotation of data points. Application of the method to some well-known learning problems is also shown.

2 A Linear Programming Method for Neural Network Generation

The following notation is used henceforth. An input pattern is represented by the N-dimensional vector x, $x = (X_1, X_2, \ldots, X_N)$. The pattern space, which is the set of all possible values that x may assume, is represented by Ω_x. K denotes the total number of classes. The method is for supervised learning where the training set $x_1, x_2, \ldots, x_n$ is a set of sample patterns with known classification.

Neural Computation 5, 317–330 (1993) © 1993 Massachusetts Institute of Technology