PERGAMON

AVAILABLE AT
www.ComputerScienceWeb.com
POWERED BY SCIENCE DIRECT®

Neural Networks 16 (2003) 1019–1028

Neural Networks

www.elsevier.com/locate/neunet

Relaxed conditions for radial-basis function networks to be universal approximators

Yi Liao*, Shu-Cherng Fang, Henry L.W. Nuttle

Operations Research and Industrial Engineering, North Carolina State University, Raleigh, NC 27695-7906, USA

Received 10 April 2002; accepted 10 October 2002

Abstract

In this paper, we investigate the universal approximation property of Radial Basis Function (RBF) networks. We show that RBFs are not required to be integrable for the REF networks to be universal approximators. Instead, RBF networks can uniformly approximate any continuous function on a compact set provided that the radial basis activation function is continuous almost everywhere, locally essentially bounded, and not a polynomial. The approximation in $L^p(\mu)(1 \leq p < \infty)$ space is also discussed. Some experimental results are reported to illustrate our findings.

© 2003 Elsevier Science Ltd. All rights reserved.

Keywords: Universal approximation; Radial-basis function networks

1. Introduction

Universal approximation by feedforward neural networks has been studied by many authors (Chen & Chen, 1995; Cybenko, 1989; Hornik, 1990, 1993; Leshno, Lin, Pinkus, & Shochen, 1993; Mhaskar & Micchelli, 1992; Park & Sandberg, 1991, 1993). Under very mild assumptions on the activation functions used in the hidden layer, it has been shown that a three-layered feedforward neural network is capable of approximating a large class of functions including the continuous functions and integrable functions.

The known results in literature are mainly built upon the three-layered neural networks with one linear output node. In this paper, we adopt this standard setting. The class of functions realized by a three-layered feedforward neural network has the following form

$$\sum_{i=1}^{N} c_i g(x, \theta_i, b_i),$$

where N is the number of hidden nodes, $x \in R^n$ is a variable, $c_i \in R$, $\theta_i \in R^n$, $b_i \in R$ are parameters, and $g(x, \theta_i, b_i)$ is the activation function used in the hidden layer. Notice that most activation functions used in the hidden layer can be categorized into two classes:the ridge functions and radial-basis functions. A ridge function has the following form

$$g(x, \theta, b) = \sigma(\theta^{\mathrm{T}} x + b),$$

where σ is a mapping from R into R, $x \in R^n$ is a variable, $\theta \in R^n$ is a 'direction vector', and $b \in R$ is a 'threshold'. The commonly used sigmoid function $g(x) = 1/(1 + \exp(-(\theta^{\mathrm{T}} x + b)))$ is an example. A radial-basis function has the following form

$$g(x, \theta, b) = \phi\left(\frac{x - \theta}{b}\right),$$

where ϕ maps from R^n into R, $x \in R^n$ is a variable, $\theta \in R^n$ is a 'center vector', and $b \in R$ is a 'spread parameter'. The Gaussian function

$$g(x) = \exp\left(-\frac{\|x-\theta\|}{b}\right),$$

is a typical example.

The research on ridge activation functions is extensive (Cybenko, 1989; Hornik, 1990, 1993; Leshno et al., 1993). The work of Leshno et al. (1993) could be one of the most general results. They showed that if the ridge activation function used in the hidden layer is continuous almost everywhere, locally essentially bounded, and not a polynomial, then a three-layered neural network can

* Corresponding author.
E-mail address: yliao2@eos.ncsu.edu (Y. Liao).

0893-6080/03/$ - see front matter © 2003 Elsevier Science Ltd. All rights reserved.
PII: S0893-6080(02)00227-7

approximate any continuous function with respect to the uniform norm.

Compared with ridge activation functions, the research on radial-basis activation functions is less extensive (Chen & Chen, 1995; Mhaskar & Micchelli, 1992; Park & Sandberg, 1991, 1993). The most well-known result is due to Park and Sandberg (1991, 1993). They showed that if the radial-basis activation function used in the hidden layer is continuous almost everywhere, bounded and integrable on R^n, and the integration is not zero, then a three-layered neural network can approximate any function in $L^p(R^n)$ with respect to the L^p norm with $1 \leq p < \infty$.

In this paper, we extend Park and Sandberg's result by showing that the integrability assumption is not necessary for the radial-basis function networks to be universal approximators. Instead, the relaxed conditions are more like the results of Leshno. More specifically, we show that, if the radial-basis activation function used in the hidden layer is continuous almost everywhere, locally essentially bounded, and not a polynomial, then the three-layered radial-basis function network can approximate any continuous function with respect to the uniform norm. Moreover, Radial Basis Function (RBF) networks can approximate any function in $L^p(\mu)$, where $1 \leq p < \infty$ and μ is any finite measure, if the radial-basis activation function used in the hidden layer is essentially bounded and not a polynomial.

This paper is organized as follows. Basic definitions and notations are introduced in Section 2. The main results are presented in Section 3. Some numerical examples are given in Section 4, and conclusions are drawn in Section 5.

2. Basic definitions and notations

Throughout this paper, R^n denotes the n-dimensional Euclidean space, K is a compact set in R^n, $C(K)$ is the set of all continuous functions defined on K, with the uniform norm $\|f\|_{C(K)} = \max_{x\in K}|f(x)|$. Moreover, $C^\infty(R^n)$ denotes the set of all infinitely differentiable functions defined on R^n, and $C_c^\infty(R^n)$ the set of all infinitely differentiable functions with compact support in R^n.

The essential supremum of a given function $f(x)$ is defined by

$$\operatorname*{ess\,sup}_{x\in R^n} f(x) = \inf\{\lambda | \mu\{x : |f(x)| \geq \lambda\} = 0\},$$

where μ is a measure. We also denote the essential supremum as

$$\|f\|_{L^\infty(R^n)} = \operatorname*{ess\,sup}_{x\in R^n} f(x).$$

Moreover, for a finite measure, μ and $1 \leq p < \infty$,

$$\|f\|_{L^p(\mu)} = \left(\int_{R^n} |f(x)|^p d\mu(x)\right)^{1/p}.$$

We denote the set of all functions f for which $\|f\|_{L^\infty(R^n)}$ is finite by $L^\infty(R^n)$ and the set of all functions f for which $\|f\|_{L^p(\mu)}$ is finite by $L^p(\mu)$, and we call these functions 'essentially bounded' with respect to $\|\cdot\|_{L^\infty(R^n)}$ norm or $\|\cdot\|_{L^p(R^n)}$ norm, respectively. Similarly, for a compact set $K \subset R^n$, we have the $\|\cdot\|_{L^\infty(K)}$ norm with $L^\infty(K)$ space and the $\|\cdot\|_{L^p(K)}$ norm with $L^p(K)$ space. A function f is locally essentially bounded with respect to the $\|\cdot\|_{L^\infty(K)}$ norm (or $\|\cdot\|_{L^p(K)}$ norm), if $\|f\|_{L^\infty(K)}$ (or $\|f\|_{L^p(K)}$) is finite for every compact set $K \subset R^n$. The set of all locally essentially bounded functions is denoted by $L^\infty_{\mathrm{loc}}(R^d)$ or $L^p_{\mathrm{loc}}(R^d)$, depending on the norm used.

We say a function is continuous almost everywhere (with respect to a measure), if the measure of the set of all discontinuous points of the function is zero. A set S of functions is said to be dense in $C(K)$ (or $L^p(\mu)$), if for any $\varepsilon > 0$ and $f \in C(K)$ (or $f \in L^p(\mu)$), there is a function $g \in S$, such that $\|g - f\|_{L^\infty(K)} < \varepsilon$ (or $\|g - f\|_{L^p(\mu)} < \varepsilon$).

The convolution of two functions is defined as $f * g(x) = \int f(x-t)g(t)\mathrm{d}t$. The Fourier transform of a Fourier transformable function f is denoted as $\hat{f}$. The support of function f is denoted by $\operatorname{supp} f$. An n-tuple $\alpha = (\alpha_1, \ldots, \alpha_n)$ of non-negative integers is called a multi-index. We define $|\alpha| = \alpha_1 + \cdots + \alpha_n$ and $\alpha! = \alpha_1! \cdots \alpha_n!$. The differential operator D^α is defined as

$$D^\alpha = \left(\frac{\partial}{\partial x_1}\right)^{\alpha_1} \cdots \left(\frac{\partial}{\partial x_n}\right)^{\alpha_n}.$$

For a function $\phi(ax+\theta)$, where $x \in R^n$, $a \in R$ and $\theta \in R^n$, $\operatorname{span}\{\phi(ax+\theta) : a \in R, \theta \in R^n\}$ denotes the set of all functions on R^n of the form

$$x \to \sum_{i=1}^{N} \beta_i \phi(a_i x + \theta_i),$$

where $\beta_i \in R^n$ and N is a given positive integer. Related terminologies and properties of functional analysis can be found in Rudin (1987).

3. Main results

The original radial-basis function is of the form $\phi((x-\theta)/b)$, where $x \in R^n$, $\theta \in R^n$ and $b \in R$. In this paper, for convenience, we write it as $\phi(ax+\theta)$, where $x \in R^n$, $a \in R$ and $\theta \in R^n$.

First, we have the following result.

Theorem 1. *Let ϕ be a mapping from R^n to R. If $\phi \in C^\infty(R^n)$ and is not a polynomial, then for any compact set $K \subset R^n$, $\Phi = \operatorname{span}\{\phi(ax+\theta) : a \in R, \theta \in R^n\}$ is dense in $C(K)$ with respect to the uniform norm, i.e., given any $f \in C(K)$ and any $\varepsilon > 0$, there exists a $g \in \Phi$ such that $|f(x) - g(x)| \leq \varepsilon$ for all $x \in K$.*

Proof. Assume that Φ is not dense in $C(K)$. By the dual space argument of Cybenko (1989), there exists a non-zero

signed finite measure λ on K, such that

$$\int_K \phi(ax+\theta)\mathrm{d}\lambda(x)=0,$$

for all $a\in R$ and $\theta\in R^n$.

Since $\phi\in C^\infty(R^n)$, using the multivariate Taylor expansion, we have

$$\begin{aligned}\phi(ax+\theta)&=\sum_{|\alpha|=0}^{\infty}\frac{1}{\alpha!}(D^\alpha\phi)(\theta)(ax)^\alpha=\sum_{|\alpha|=0}^{\infty}\frac{a^{|\alpha|}}{\alpha!}(D^\alpha\phi)(\theta)x^\alpha\\&=\phi(\theta)+a\sum_{|\alpha|=1}\frac{1}{\alpha!}(D^\alpha\phi)(\theta)x^\alpha\\&\quad+a^2\sum_{|\alpha|=2}\frac{1}{\alpha!}(D^\alpha\phi)(\theta)x^\alpha+\cdots.\end{aligned}$$

Let $H(a)=\int_K\phi(ax+\theta)\mathrm{d}\lambda(x)$. Since $H(a)=0$ for every $a\in R$ and $\theta\in R^n$, the kth derivative of H with respect to a becomes

$$\begin{aligned}\frac{\mathrm{d}^kH}{\mathrm{d}a^k}&=\int_K\left[\sum_{|\alpha|=k}\frac{k!}{\alpha!}(D^\alpha\phi)(\theta)x^\alpha\right.\\&\quad\left.+a\sum_{|\alpha|=k+1}\frac{(k+1)!}{\alpha!}(D^\alpha\phi)(\theta)x^\alpha+\cdots\right]\mathrm{d}\lambda(x)=0,\end{aligned}$$

for all $\theta\in R^n$. If we set $a=0$, then

$$\begin{aligned}\left.\frac{\mathrm{d}^kH}{\mathrm{d}a^k}\right|_{a=0}&=\int_K\left[\sum_{|\alpha|=k}\frac{k!}{\alpha!}(D^\alpha\phi)(\theta)x^\alpha\right]\mathrm{d}\lambda(x)\\&=\sum_{|\alpha|=k}\left[\frac{k!}{\alpha!}(D^\alpha\phi)(\theta)\int_{R^n}x^\alpha\mathrm{d}\lambda(x)\right]=0,\end{aligned}$$

for all $\theta\in R^n$. Equivalently, we have

$$\sum_{i=1}^{r(k)}c_i(\theta)t_i=0,$$

for all $\theta\in R^n$, where $c_i(\theta)=k!(D^\alpha\phi)(\theta)$, $t_i=1/\alpha!\int_K x^\alpha\mathrm{d}\lambda(x)$, and $r(k)$ is the number of αs such that $|\alpha|=k$.

Since $\phi\in C^\infty(R^n)$ and is not a polynomial, $c_i(\theta)$ is continuous and not a constant. Therefore, $c_i(\theta)$ can have infinitely many values for different θs. Hence, there exist at least $(r(k)+1)\theta$s, such that the above linear system is overdetermined. Therefore, the only solution for the above linear system is $t_i=0$ for all i. That is, $\int_K x^\alpha\mathrm{d}\lambda(x)=0$ for all multi-index $\alpha\geq 0$. This means that the Fourier transform (of λ) $\hat{\lambda}(t)=\int_K \mathrm{e}^{-\mathrm{i}t^{\mathrm{T}}x}\mathrm{d}\lambda(x)=0$ for all $t\in R^n$. By Rudin (1987, Theorem 1.3.7.b), we have $\lambda=0$. But this is impossible and, hence, the proof is complete. □

The above theorem says that if the activation function used in the hidden layer is infinitely differentiable and not a polynomial, then the three-layered radial-basis function network is a universal approximator. The requirement of infinite differentiability is very strong in theory. But since neural networks are often trained with back-propagation algorithms, which usually assume the activation function used in the hidden layer to be differentiable, this requirement does not cause too much problem in practice. Fortunately, we can relax this requirement in the following derivation.

Lemma 1. *Let σ be a mapping from R^n to R. If $\sigma\in L^\infty_{\mathrm{loc}}(R^n)$ and σ is not a polynomial, then there exists at least one $\omega\in C_c^\infty(R^n)$, such that $\sigma*\omega(x)=\int_{R^n}\sigma(x-t)\omega(t)\mathrm{d}t$ is not a polynomial.*

Proof. Since $\sigma\in L^\infty_{\mathrm{loc}}(R^n)$, $\sigma*\omega$ is well-defined and $\sigma*\omega\in C^\infty(R^n)$. Suppose that $\sigma*\omega(x)$ is a polynomial for every $\omega\in C_c^\infty(R^n)$. Then for any multi-index α such that $|\alpha|=\infty$, we have

$$D^\alpha\sigma*\omega(x)=0,$$

for all $\omega\in C_c^\infty(R^n)$ and $x\in R^n$. But according to Friedman (1963, pp. 57–59), σ is a polynomial of degree $<|\alpha|=\infty$, which causes a contradiction, and the proof is complete. □

Lemma 2. *Let σ be a mapping from R^n to R. If $\sigma\in L^\infty_{\mathrm{loc}}(R^n)$ and σ is continuous almost everywhere, then for each $\omega\in C_c^\infty(R^n)$, $\sigma*\omega(x)$ can be uniformly approximated by $\Sigma=\mathrm{span}\{\sigma(ax+\theta):a\in R,\theta\in R^n\}$.*

Proof. Recall that

$$\sigma*\omega(x)=\int_{R^n}\sigma(x-t)\omega(t)\mathrm{d}t,$$

is well-defined.

Suppose that the supp $\omega\subseteq[-T,T]^n$. We show that $\sigma*\omega(x)$ can be uniformly approximated on $[-T,T]^n$ by

$$\sum_{i=1}^{m^n}\sigma(x-t_i)\omega(t_i)\left(\frac{2T}{m}\right)^n,$$

where $\{t_i\in R^n:i=1,\ldots,m^n\}$ is a set consisting of all points in $[-T,T]^n$ of the form

$$\left[-T+\frac{2i_1T}{m},\ldots,-T+\frac{2i_nT}{m}\right]^T,\quad i_1,\ldots,i_n=1,2,\ldots,m.$$

Set $\Delta_i=[t_{i-1},t_i]$. By the triangular inequality, we have

$$\begin{aligned}&\left|\int_{R^n}\sigma(x-t)\omega(t)\mathrm{d}t-\sum_{i=1}^{m^n}\sigma(x-t_i)\omega(t_i)\left(\frac{2T}{m}\right)^n\right|\\&\quad=\left|\int_{R^n}\sigma(x-t)\omega(t)\mathrm{d}t-\sum_{i=1}^{m^n}\int_{\Delta_i}\sigma(x-t_i)\omega(t)\mathrm{d}t\right.\\&\qquad\left.+\sum_{i=1}^{m^n}\int_{\Delta_i}\sigma(x-t_i)\omega(t)\mathrm{d}t-\sum_{i=1}^{m^n}\sigma(x-t_i)\omega(t_i)\left(\frac{2T}{m}\right)^n\right|\\&\quad\leq\left|\int_{R^n}\sigma(x-t)\omega(t)\mathrm{d}t-\sum_{i=1}^{m^n}\int_{\Delta_i}\sigma(x-t_i)\omega(t)\mathrm{d}t\right|\\&\qquad+\left|\sum_{i=1}^{m^n}\int_{\Delta_i}\sigma(x-t_i)\omega(t)\mathrm{d}t-\sum_{i=1}^{m^n}\sigma(x-t_i)\omega(t_i)\left(\frac{2T}{m}\right)^n\right|.\end{aligned}\tag{1}$$

Notice that in the second term of Eq. (1), $\int_{\Delta_i} \mathrm{d}t = (2T/m)^n$. Therefore, we have

$$\left|\sum_{i=1}^{m^n} \int_{\Delta_i} \sigma(x-t_i)\omega(t)\mathrm{d}t - \sum_{i=1}^{m^n} \sigma(x-t_i)\omega(t_i)\left(\frac{2T}{m}\right)^n\right|$$
$$= \left|\sum_{i=1}^{m^n} \int_{\Delta_i} \sigma(x-t_i)[\omega(t)-\omega(t_i)]\mathrm{d}t\right|$$
$$\leq \sum_{i=1}^{m^n} \int_{\Delta_i} |\sigma(x-t_i)||\omega(t)-\omega(t_i)|\mathrm{d}t.$$

Since ω is continuous, it is uniformly continuous on $[-T,T]^n$. Therefore, we may choose m to be sufficiently large such that

$$\sum_{i=1}^{m^n} \int_{\Delta_i} |\sigma(x-t_i)||\omega(t)-\omega(t_i)|\mathrm{d}t \leq \varepsilon.$$

For the first term of Eq. (1), we have

$$\left|\int_{R^n} \sigma(x-t)\omega(t)\mathrm{d}t - \sum_{i=1}^{m^n} \int_{\Delta_i} \sigma(x-t_i)\omega(t)\mathrm{d}t\right|$$
$$= \left|\sum_{i=1}^{m^n} \int_{\Delta_i} [\sigma(x-t)-\sigma(x-t_i)]\omega(t)\mathrm{d}t\right|$$
$$\leq \sum_{i=1}^{m^n} \int_{\Delta_i} |\sigma(x-t)-\sigma(x-t_i)||\omega(t)|\mathrm{d}t.$$

Since σ is continuous almost everywhere, the measure of the set of its discontinuous points is zero. Therefore, given any small number $\delta > 0$, we can find a countable number of intervals, whose union is of measure δ, such that σ is uniformly continuous on $[-T,T]^n \backslash t$. For any Δ_i, since $\Delta_i = (\Delta_i \backslash t) \cup (\Delta_i \cap t)$, the above equation can be written as

$$\sum_{i=1}^{m^n} \int_{\Delta_i} |\sigma(x-t)-\sigma(x-t_i)||\omega(t)|\mathrm{d}t$$
$$= \sum_{i=1}^{m^n} \int_{\Delta_i \backslash t} |\sigma(x-t)-\sigma(x-t_i)||\omega(t)|\mathrm{d}t$$
$$+ \sum_{i=1}^{m^n} \int_{\Delta_i \cap t} |\sigma(x-t)-\sigma(x-t_i)||\omega(t)|\mathrm{d}t. \quad (2)$$

Now, for the first term of Eq. (2), since σ is uniformly continuous on $[-T,T]^n \backslash t$, we may choose m to be sufficiently large so that

$$\sum_{i=1}^{m^n} \int_{\Delta_i \backslash t} |\sigma(x-t)-\sigma(x-t_i)||\omega(t)|\mathrm{d}t \leq \varepsilon.$$

For the second term of Eq. (2), since $\int_t \mathrm{d}t = \delta$, we can choose δ to be sufficiently small such that

$$\sum_{i=1}^{m^n} \int_{\Delta_i \cap t} |\sigma(x-t)-\sigma(x-t_i)||\omega(t)|\mathrm{d}t$$
$$\leq 2\|\sigma\|_{L^\infty[-T,T]}\|\omega\|_{L^\infty}\delta \leq \varepsilon.$$

Therefore,

$$\sum_{i=1}^{m^n} \int_{\Delta_i} |\sigma(x-t)-\sigma(x-t_i)||\omega(t)|\mathrm{d}t \leq 2\varepsilon.$$

Consequently,

$$\left|\int_{R^n} \sigma(x-t)\omega(t)\mathrm{d}t - \sum_{i=1}^{m^n} \sigma(x-t_i)\omega(t_i)\left(\frac{2T}{m}\right)^n\right| \leq 3\varepsilon,$$

for all $x \in [-T,T]$. The proof is complete. □

By combining the above lemmas and theorem, we have the following main result:

Theorem 2. *Let σ be a mapping from R^n to R. If σ is continuous almost everywhere, locally essentially bounded, and not a polynomial, then for any compact set $K \subset R^n$, $\Sigma = span\{\sigma(ax+\theta) : a \in R, \theta \in R^n\}$ is dense in $C(K)$ with respect to the uniform norm, i.e., given any $f \in C(K)$ and any $\varepsilon > 0$, there exists a $g \in \Sigma$, such that $\|f(x) - g(x)\|_{L^\infty(K)} \leq \varepsilon$ for all $x \in K$.*

Proof. From Lemma 1, we know that there exists some $\omega \in C_c^\infty(R^n)$, such that $\sigma * \omega(x)$ is not a polynomial. Since $\sigma * \omega(x) \in C^\infty(R^n)$, by Theorem 1, we know span$\{\sigma * \omega(ax+\theta)\}$ is dense in $C(K)$. From Lemma 2, $\sigma * \omega$ can be uniformly approximated by Σ. It follows that span$\{\sigma * \omega(ax+\theta)\}$ can be uniformly approximated by Σ. Thus Σ is dense in $C(K)$. □

The above result says that if the activation function used in the hidden layer is continuous almost everywhere, locally essentially bounded, and not a polynomial, then the three-layered radial-basis function network is a universal approximator. This significantly extends Park and Sandberg's results. In particular, the activation function is no longer required to be integrable. Notice that if the activation function used in the hidden layer is a polynomial, the neural network can only produce a polynomial of a certain degree. Therefore, the requirement of being 'not a polynomial' is also a necessary condition. We do not know if the requirement of being 'continuous almost everywhere and locally essentially bounded' is also a necessary condition.

Besides the approximation on $C(K)$, we have the following result for universal approximation in the $L^p(\mu)$ space, where $1 \leq p < \infty$ and μ is a finite measure.

Theorem 3. *Let σ be a mapping from R^n to R. For any finite measure μ, if $\sigma \in L^\infty(\mu)$ and is not a polynomial, then*

$\Sigma = span\{\sigma(ax+\theta) : a \in R, \theta \in R^n\}$ *is dense in* $L^p(\mu)$, *for* $1 \le p < \infty$.

Proof. Assume that Σ is not dense in $L^p(\mu)$. Since $\sigma \in L^\infty(\mu)$, Σ is a subspace of $L^p(\mu)$. By the dual space argument of Cybenko (1989), there exists a non-zero signed finite measure λ such that

$$\int_{R^n} \sigma(ax+\theta)\mathrm{d}\lambda(x) = 0,$$

for all $a \in R$ and $\theta \in R^n$. From Lemma 1, we can choose a $\omega \in C_c^\infty(R^n)$ such that $\sigma * \omega$ is not a polynomial.

Now consider the integral

$$\int_{R^n} \sigma * \omega(ax+\theta)\mathrm{d}\lambda(x) = \int_{R^n}\int_{R^n} \sigma(ax+\theta-t)\omega(t)\mathrm{d}t\mathrm{d}\lambda(x).$$

Since

$$\int_{R^n}\int_{R^n} |\sigma(ax+\theta-t)\omega(t)|\mathrm{d}t\mathrm{d}|\lambda(x)|$$

$$\le \|\sigma(ax+\theta-t)\|_{L^\infty}\|\omega\|_{L^1}\lambda(R^n) < \infty,$$

by Fubini's Theorem, we have

$$\begin{aligned}\int_{R^n} \sigma * \omega(ax+\theta)\mathrm{d}\lambda(x) &= \int_{R^n}\int_{R^n} \sigma(ax+\theta-t)\omega(t)\mathrm{d}t\mathrm{d}\lambda(x)\\ &= \int_{R^n}\left[\int_{R^n} \sigma(ax+\theta-t)\mathrm{d}\lambda(x)\right]\omega(t)\mathrm{d}t = 0.\end{aligned}$$

Since $\sigma * \omega \in C^\infty(R^n)$ and it is not a polynomial, the rest of the proof is identical to that of Theorem 1.

4. Numerical Experiments

In this section, we present some numerical experiments to demonstrate the validity of the obtained results. We show that even when the activation function used in the hidden layer is not integrable, a radial-basis function network may still be a universal approximator as long as the activation function meets our (much relaxed) conditions.

In our experiments, we use two different activation functions in the hidden layer. One is the traditional Gaussian function

$$g(x) = \exp\left(-\frac{\|x-\theta\|}{b}\right),$$

where $x \in R^n$, $\theta \in R^n$ is a center vector, and $b \in R$ is a spread parameter. The Gaussian function is integrable and meets Park and Sandberg's requirements. The other activation function is

$$g(x) = \exp\left(\frac{\|x-\theta\|}{b}\right),$$

where $x \in R^n$, $\theta \in R^n$ is a center vector, and $b \in R$ is a spread parameter. This function is not integrable, but it meets the relaxed conditions proposed in this paper.

Corresponding radial-basis function networks are constructed to approximate a set of N input–output pairs $(x_i, y_i), i = 1, \ldots, N$. We show that both radial-basis function networks are capable of performing universal approximation.

In our experiments, the Mean Squared Error between the targets and actual outputs is used as the performance measure. The radial-basis function networks are trained according to the following steps:

Step 1: set the number of hidden nodes to be 1;
Step 2: choose the spread parameter and the center vectors;
Step 3: optimize the weights on the links between the hidden layer and the output layer. Feed the inputs into the network and get the outputs;
Step 4: increase the number of hidden nodes by 1. If the number of hidden nodes is less than or equal to the number of samples, go to Step 2; otherwise, stop.

The optimization of weights in Step 3 is straightforward, since it only involves solving a linear system. We use two approaches to select the center vectors, θ. One is the Random Selection method, which randomly generates a set of centers. The other one is the Orthogonal Least Squares method (Chen, Cowan, & Grant, 1991), which considers the complete set of training samples to be candidates for centers, and selects the one that reduces the mean squared error the most. In the following, we denote the Random Selection method as RS, and the Orthogonal Least Squares method as OLS.

4.1. One-dimensional example

The first experiment is a one-dimensional example, which is used as an illustrative example for function approximation in Matlab 5.3. We are given 21 input–output pairs $(x_i, y_i), i = 1, \ldots, 21$, which are depicted in Fig. 1.

The approximation results for different activation functions and center selection methods are depicted in Figs. 2–5, respectively. As we can see from the figures, in all cases, the Mean Squared Errors between the targets and the actual outputs decreases to zero as the number of hidden nodes increases. This demonstrates that the activation function used in the hidden layer needs not to be integrable in order to achieve universal approximation.

4.2. Multi-dimensional example

In this example, we used the well-known Cleveland heart disease data as the data samples (Murphy & Aha, 1992). This data set was originally used for pattern classification to diagnose heart disease. It contains 303 points, each point

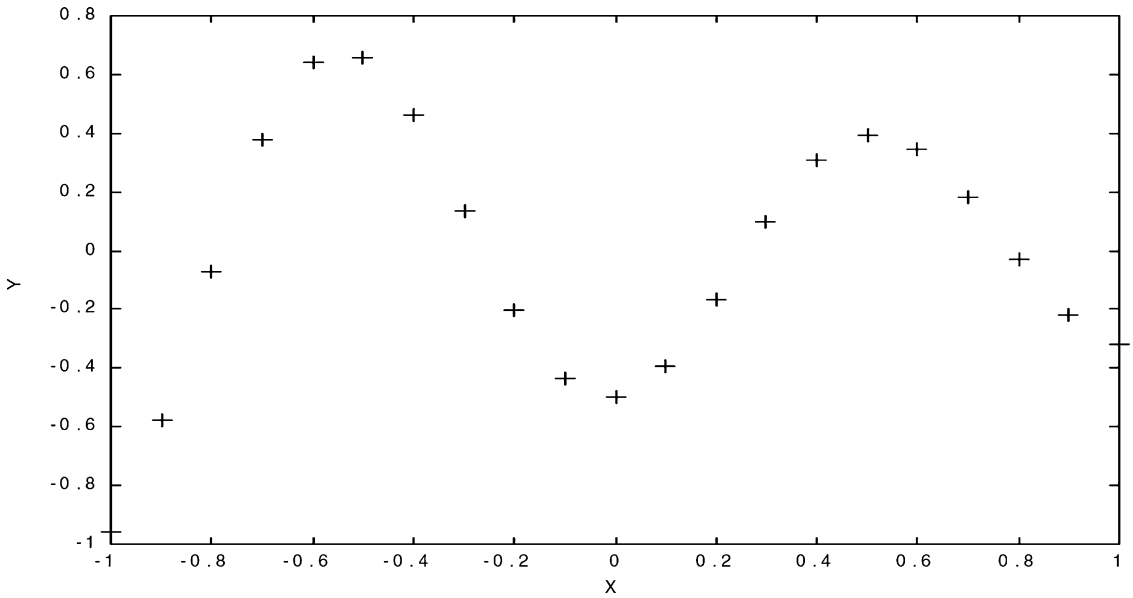

Fig. 1. The function to be approximated.

consisting of 13 features. All features are continuous variables. The 'positive' class (heart disease) contains 164 points and the negative class (no heart disease) contains 139 points. In the data set, the 'positive' class is denoted as 1, the 'negative' class is denoted as 0. Therefore, the input space is of dimension 13, and the output is either 1 or 0. Since the outputs are discrete, this data set is more difficult to be approximated.

The approximation results for different activation functions and center selection methods are depicted in Figs. 6–9, respectively. The results are basically the same as before, i.e., the Mean Squared Errors decreases to zero as

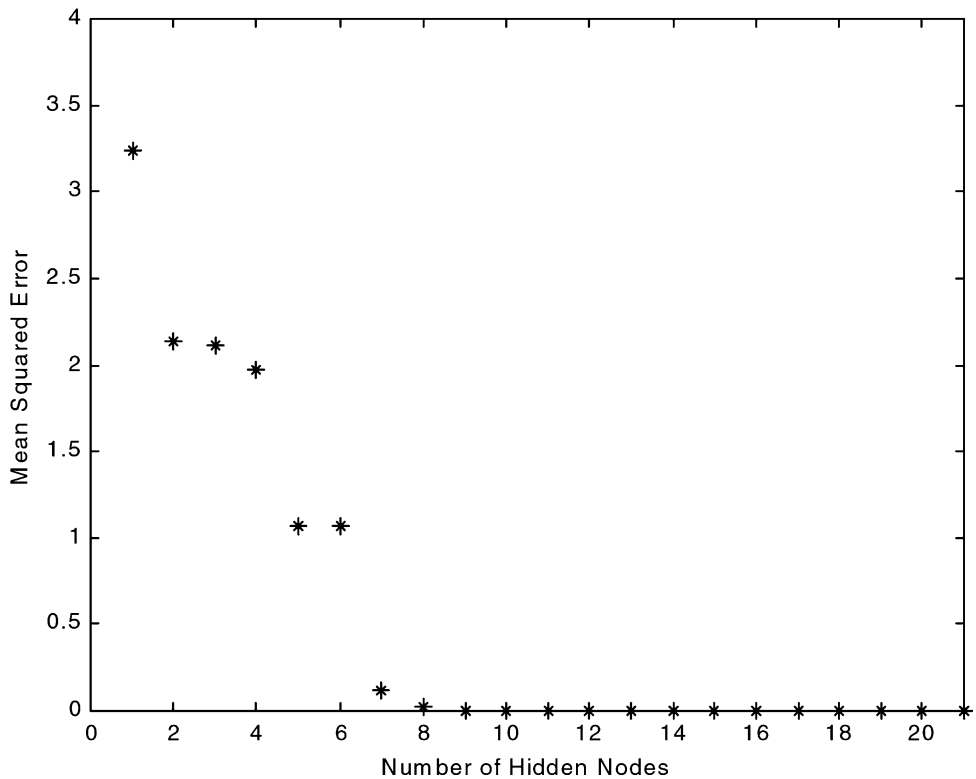

Fig. 2. The approximation result for the one-dimensional example with activation function $g(x) = \exp(\frac{\|x-\theta\|}{b})$, trained with RS.

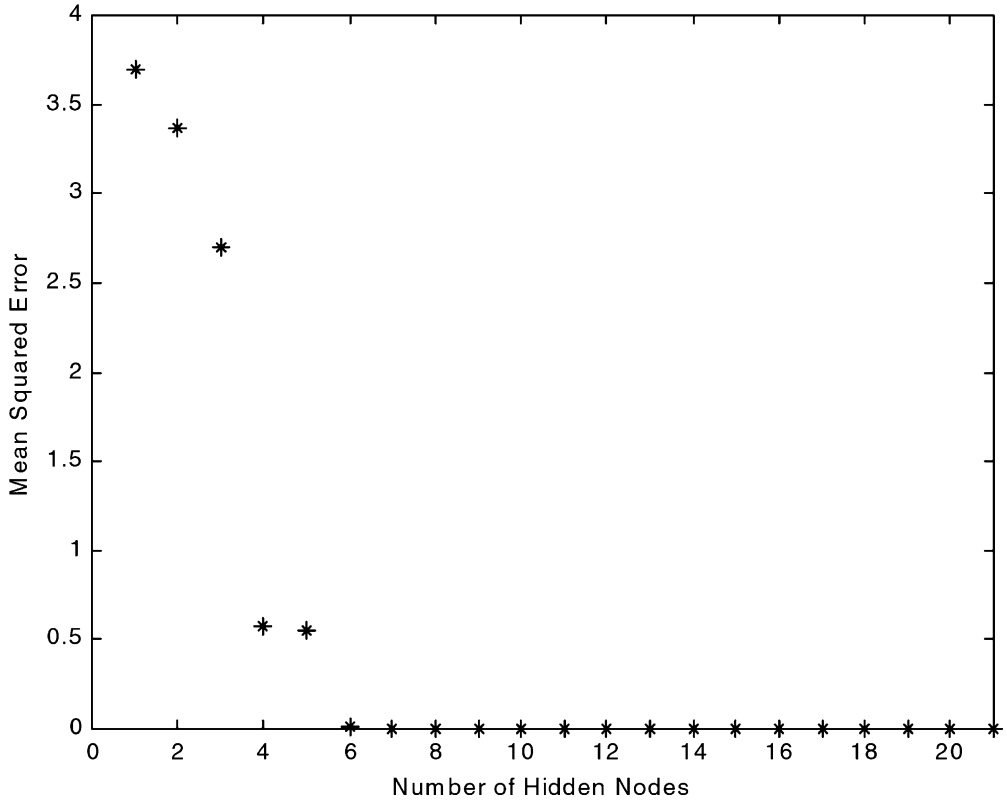

Fig. 3. The approximation result for the one-dimensional example with activation function $g(x) = \exp(-\frac{\|x-\theta\|}{b})$, trained with RS.

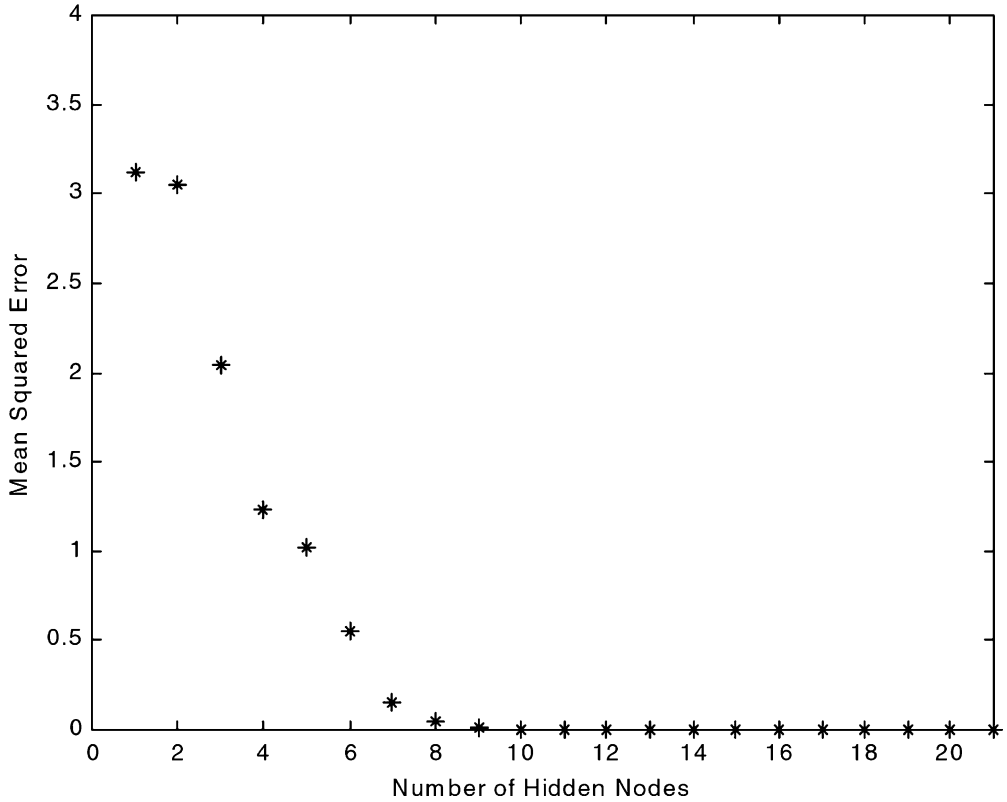

Fig. 4. The approximation result for the one-dimensional example with activation function $g(x) = \exp(\frac{\|x-\theta\|}{b})$, trained with OLS.

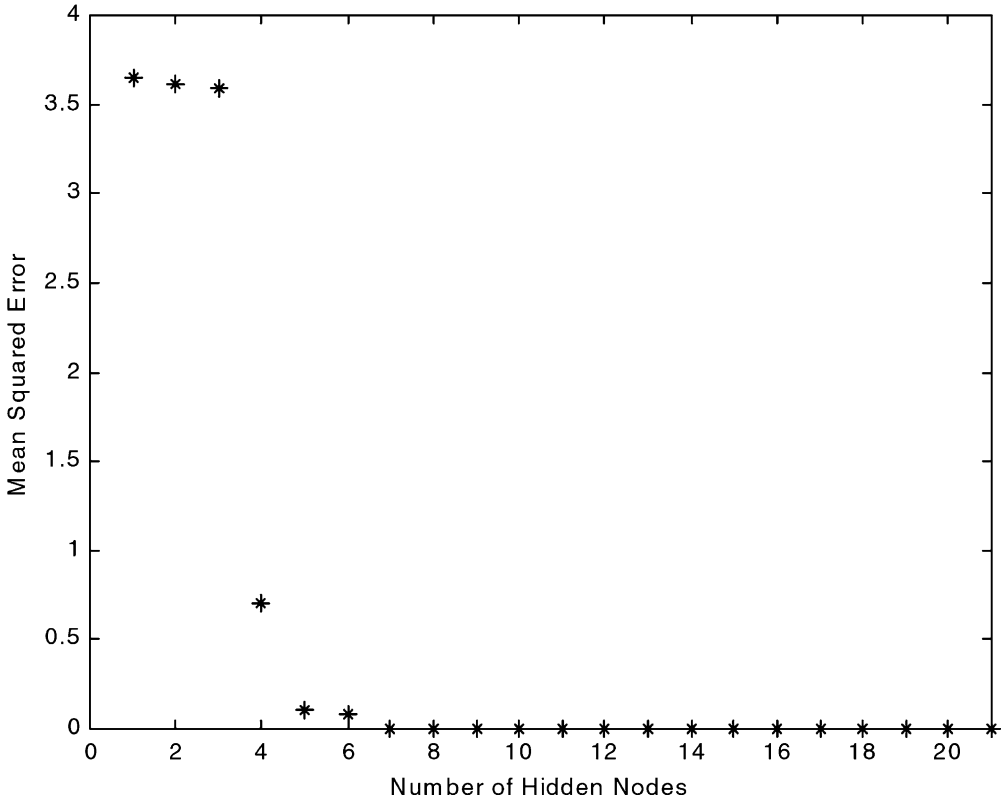

Fig. 5. The approximation result for the one-dimensional example with activation function $g(x) = \exp(-\frac{\|x-\theta\|}{b})$, trained with OLS.

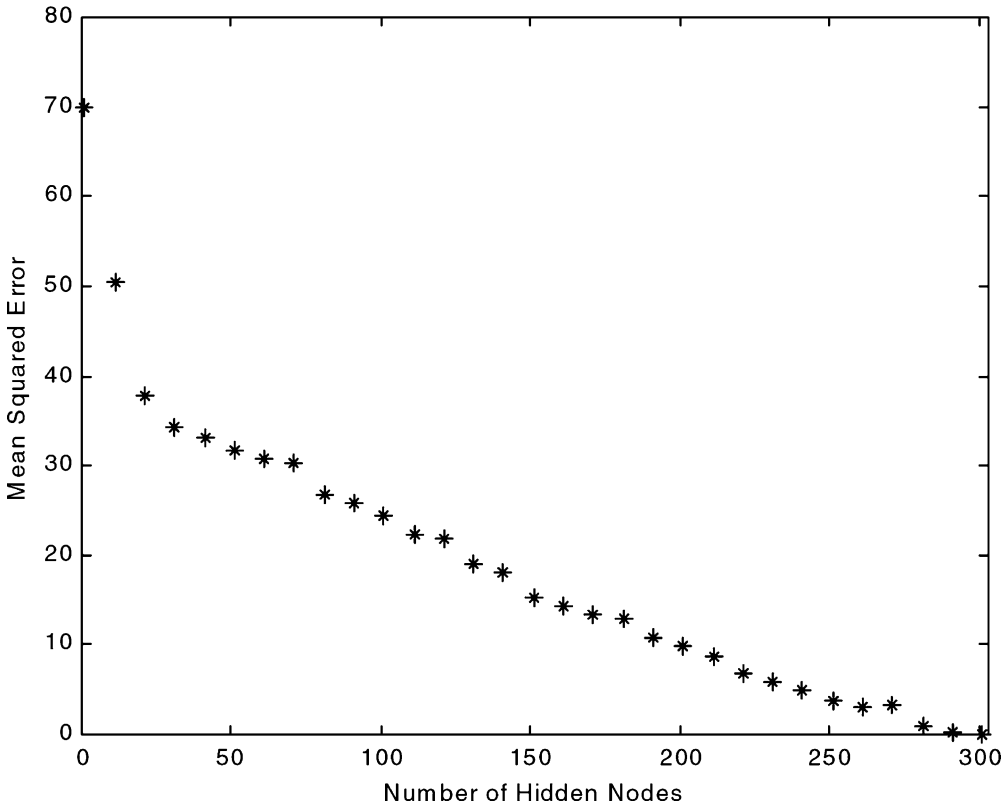

Fig. 6. The approximation result for the multi-dimensional example with activation function $g(x) = \exp(\frac{\|x-\theta\|}{b})$, trained with RS.

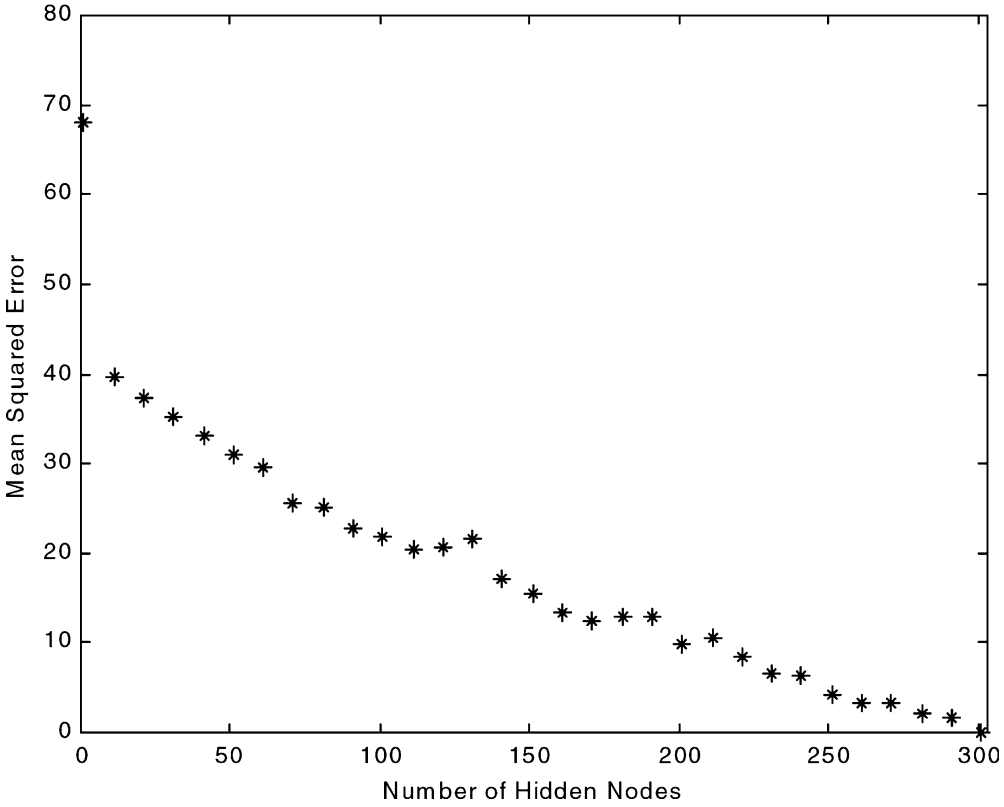

Fig. 7. The approximation result for the multi-dimensional example with activation function $g(x) = \exp(-\frac{\|x-\theta\|}{b})$, trained with RS.

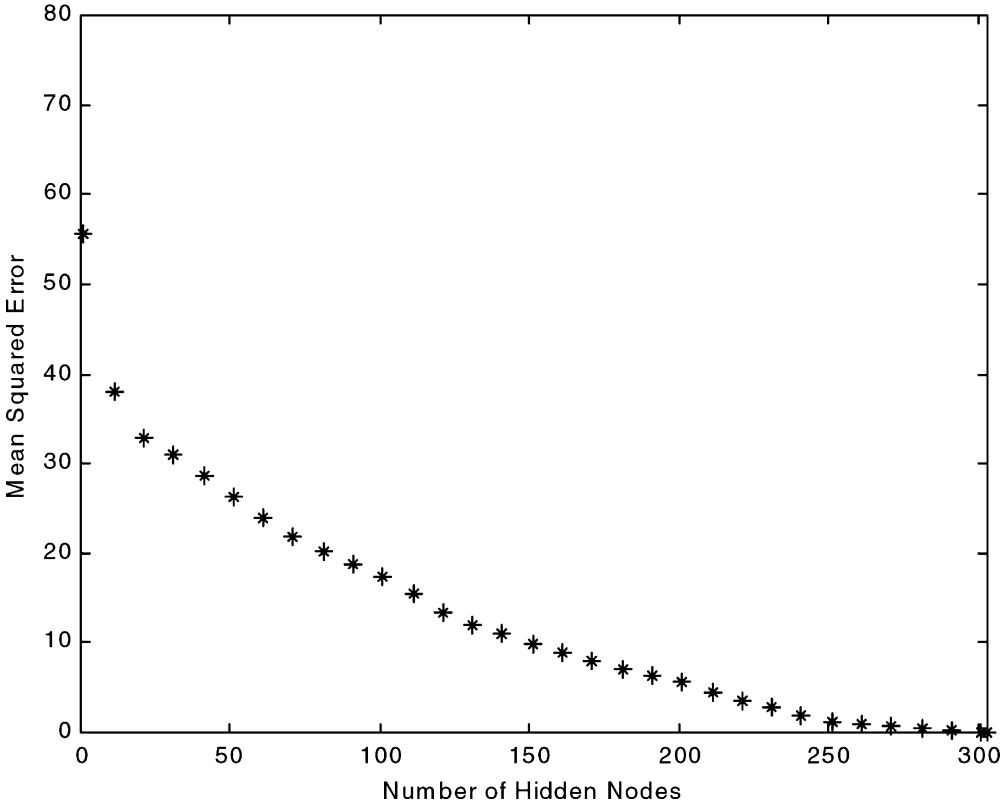

Fig. 8. The approximation result for the multi-dimensional example with activation function $g(x) = \exp(\frac{\|x-\theta\|}{b})$, trained with OLS.

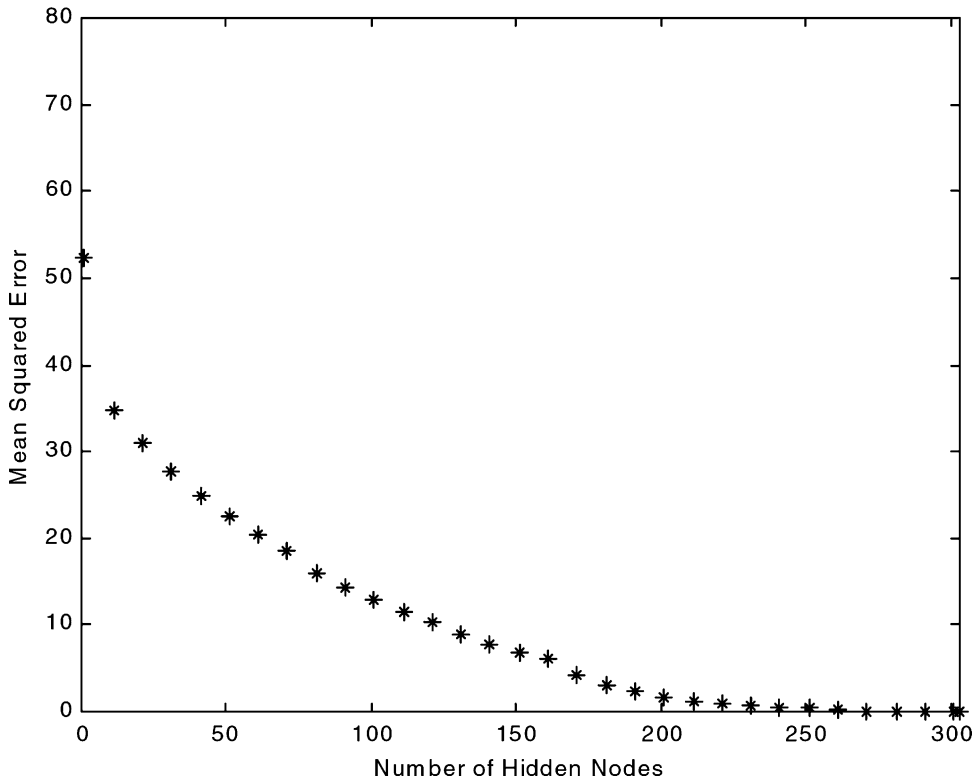

Fig. 9. The approximation result for the multi-dimensional example with activation function $g(x) = \exp(-\frac{\|x-\theta\|}{b})$, trained with OLS.

the number of hidden nodes increases. And as we can see from the figures, the speeds of convergence of these two different RBF networks are also about the same, even though the one whose RBF is not integrable. This further demonstrates the validity of our results.

Similar results have been obtained for other examples, including the data sets for Ionosphere (Murphy & Aha, 1992) and Wisconsin Breast Cancer (Murphy & Aha, 1992). Due to the space limit, we do not present all of them in here.

5. Conclusion

In this paper, we have studied the universal approximation property of three-layered radial-basis function networks. We have shown that the integrability property usually required of the activation functions is not necessary. We have shown that a RBF network can be a universal approximator in the continuous function space, if the activation function used in the hidden layer is continuous almost everywhere, locally essentially bounded, and not a polynomial. Moreover, for the universal approximation in $L^p(\mu)$ space, with $1 \leq p < \infty$ and μ being a finite measure, we only need the activation function used in he hidden layer to be essentially bounded and not a polynomial. The experimental results support our theoretical findings.

References

Chen, T., & Chen, H. (1995). Universal approximation to nonlinear operators by neural networks with arbitrary activation functions and its application to dynamical systems. *IEEE Transactions on Neural Networks*, *6*(4), 911–917.

Chen, S., Cowan, C. F. N., & Grant, P. M. (1991). Orthogonal least squares learning algorithm for radial basis function networks. *IEEE Transactions on Neural Networks*, *2*(2), 302–309.

Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals, and Systems*, *3*, 303–314.

Friedman, A. (1963). *Generalized functions and partial differential equations*. Englewood Cliffs, NJ: Prentice-Hall.

Hornik, K. (1990). Approximation capabilities of multilayer feedforward neural networks. *Neural Networks*, *4*, 251–257.

Hornik, K. (1993). Some new results on neural network approximation. *Neural Networks*, *6*, 1069–1072.

Leshno, M., Lin, V., Pinkus, A., & Shochen, S. (1993). Multilayer feedforward networks with a polynomial activation function can approximate any function. *Neural Networks*, *6*, 861–867.

Mhaskar, H., & Micchelli, C. (1992). Approximation by superposition of sigmoidal and radial basis functions. *Advances in Applied Mathematics*, *13*, 350–373.

Murphy, P. M., Aha, D. W. (1992). UCI repository of machine learning databases. URL: http://www.ics.uci.edu/~mlearn/MLRepository.html.

Park, J., & Sandberg, I. W. (1991). Universal approximation using radial-basis-function networks. *Neural Computation*, *3*(2), 246–257.

Park, J., & Sandberg, I. W. (1993). Approximation and radial-basis-function networks. *Neural Computation*, *5*, 305–316.

Rudin, W. (1987). *Real and complex analysis* (3rd ed.). New York: McGraw-Hill.