Accepted for publication by IEEE Transactions on Neural Networks.

Universal Approximation to Nonlinear Operators by Neural Networks with Arbitrary Activation Functions and Its Application to Dynamical Systems

Tianping Chen[1] and Robert Chen[2]

Abstract

The purpose of this paper is to investigate neural network capability systematically. The main results are: (1) Every Tauber-Wiener function is qualified as an activation function in the hidden layer of a three-layered neural network; (2) For a continuous function to be a Tauber-Wiener function, the necessary and sufficient condition is that it is not a polynomial; (3) The capability of approximating nonlinear functionals defined on some Banach space and nonlinear *operators* has been shown, which implies that (4) we can use neural network computation to approximate the output as a whole (*not* at a fixed point) of a dynamical system.

Key words: Approximation theory, neural networks, dynamical systems, compact set, functional, operator.

[1]The author is with the Department of Mathematics, Fudan University, Shanghai, P.R.China.

[2]The author was with the Department of Electrical Engineering, University of Notre Dame, Notre Dame, Indiana 46556, USA. He is now with VLSI Libraries, Inc., 1836 Cabrillo Ave., Santa Clara, CA 95050.

1 Introduction

There have been many papers related to approximation to a continuous function of several variables. In 1987, Wieland and Leighton [1] dealt with the capability of networks consisting of one or two hidden layers. Miyake and Irie [2] obtained an integral representation formula with an integral kernel fixed beforehand. This representation formula is a kind of integrals, which could be realized by a three-layered neural network. In 1989, several papers related to this topic appeared. They all claimed that a three-layered neural network with sigmoid units on the hidden layer can approximate continuous or other kinds of functions defined on compact set in $\mathbf{R}^n$. They used different methods. Carrol and Dickinson [4] used inverse Radon transform. Cybenko [3] used Hahn-Banach theorem and Riesz representation theorem. Funahashi [5] approximated Irie and Miyake's integral representation by a finite sum, using a kernel which can be expressed as a difference of two sigmoidal functions. Hornik et al. [6] applied Stone-Weierstrass theorem, using trigonometric functions.

However, in all these papers, sigmoidal functions must be assumed to be continuous or monotone. Recently [9], we pointed out that the boundedness of the sigmoidal function plays an essential role for its being an activation function in the hidden layer.

In addition to sigmoidal functions, many other functions can be used as activation functions in the hidden layer. For example, Hornik [6] proved that any bounded non-constant continuous function is qualified to be an activation function. Mhaskar and Micchelli [11] showed that under some restriction on the amplitude of a continuous function near infinity, any non-polynomial function is qualified to be an activation function.

It is clear that all the aforementioned works are concerned with approximation to a continuous function defined on a compact set in $\mathbf{R}^n$ (a space of *finite* dimensions). However, in engineering problems such as computing the output of dynamic systems or designing neural system identifiers, we often encounter the problem of approximating nonlinear functionals defined on some function space, even nonlinear operators from one function space (a space of *infinite* dimensions) to another function space (another space of *infinite* dimensions). In [10], Sandberg gave an interesting theorem on approximating nonlinear functionals by superposition and composition of several linear functionals and a continuous function of one variable. Yet, two problems remain open: 1. Can we give those linear functionals explicitly? 2. Can we approximate nonlinear operators rather than functionals? Problem 1 is essential in application, since otherwise we are not able to construct real networks. Problem 2 is important in computing dynamic systems, for a dynamic system is in fact an operator. In [12], we discussed in detail the problem of approximating nonlinear functionals defined on some compact set in $C[a,b]$ or $L^p[a,b]$ and obtained some explicit results. However, the problem of neural network's capability in approximating nonlinear operators with its related application in computing the output as a whole of a dynamic system still remains open. Moreover, a unified and systematic treatment of neural network approximation to continuous functions, functionals and operators is much needed but nevertheless also remains to be an open problem.

Specifically, it is quite natural to raise the following issues: (1) What is the characteristic property for a continuous function in the hidden layer of a neural network? (2) To give a neural network model to approximate nonlinear functionals defined on some compact set in $C(K)$, where K is some compact set in some Banach space. (3) To give a neural network model, which can be used to approximate the output of some dynamic system as a whole (not merely at a special point, cf. [10][12]), thus to identify the dynamic system.

In this paper, we systematically give strong results for these issues.

The paper is organized as follows. In section 2, we review some definitions and notations. In section 3, we show that the necessary and sufficient condition for a continuous function in $S'(\mathbf{R}^1)$ (tempered distributions in $\mathbf{R}^1$) to be a Tauber-Wiener function (for definitions, see section 2) is that it is not a polynomial; and any Tauber-Wiener function can be used as an activation function, i.e., any non-polynomial continuous function in $S'(\mathbf{R}^1)$ is an activation function. What is more interesting is that we show the approximation is equiuniform on any compact set in $C(K)$, which is crucial in discussing approximation to continuous operators by neural networks. In section 4, we show the capability of neural networks to approximate continuous functionals defined on some compact set in $C(K)$, where K is a compact set in some Banach space; and through which we establish the capability of neural networks to approximate continuous operators from $C(K_1)$ to $C(K_2)$. The main results in section 4 has a direct application to computing output of dynamic systems thus identifying the systems, which is discussed in section 5.

2 Notations and Definitions

Definition 1. A function $\sigma : \mathbf{R}^1 \to \mathbf{R}^1$ is called a sigmoidal function, if it satisfies

$$\begin{cases} \lim_{x \to -\infty} \sigma(x) = 0 \ , \\ \lim_{x \to \infty} \sigma(x) = 1 \ . \end{cases}$$

$\square$

Definition 2. If a function $g : \mathbf{R} \to \mathbf{R}$ (continuous or discontinuous) satisfies that all the linear combinations $\sum_{i=1}^{N} c_i g(\lambda_i x + \theta_i)$, $\lambda_i \in \mathbf{R}$, $\theta_i \in \mathbf{R}$, $c_i \in \mathbf{R}$, $i = 1, 2, \ldots, N$, are dense in every $C[a, b]$, then g is called a Tauber-Wiener function, or simply (TW) function. $\square$

Definition 3. Suppose that X is a Banach space, $V \subseteq X$ is called a compact set in X, if for every sequence $\{x_n\}_{n=1}^{\infty}$ with all $x_n \in V$, there is a subsequence $\{x_{n_k}\}$, which converges to some element $x \in V$. $\square$

It is well known that if $V \subseteq X$ is a compact set in X, then for any $\delta > 0$, there is a δ-net $N(\delta) = \{x_1, \ldots, x_{n(\delta)}\}$, with all $x_i \in V$, $i = 1, \ldots, n(\delta)$, *i.e.* for every $x \in X$, there is some $x_i \in N(\delta)$ such that $\|x_i - x\|_X < \delta$.

In the sequel, we will often use the following notations.

X: some Banach space with norm $\|\cdot\|_X$.

$\mathbf{R}^n$: Euclidean space of dimension n.

K: some compact set in a Banach space.

$C(K)$: Banach space of all continuous functions defined on K, with norm $\|f\|_{C(K)} =$

$\max_{x \in K} |f(x)|$.

(TW): All the Tauber-Wiener functions.

$S(\mathbf{R}^n)$: Schwartz functions in tempted distribution theory, *i.e.* rapidly decreasing and infinitely differentiable functions.

$S'(\mathbf{R}^n)$: Tempered distributions, *i.e.* linear continuous functionals defined on $S(\mathbf{R}^n)$.

$C^\infty(\mathbf{R}^n)$: Infinitely differentiable functions.

$C_c^\infty(\mathbf{R}^n)$: Infinitely differentiable functions with compact support in $\mathbf{R}^n$.

$C_p[-1,1]^n$: All 2-periodic functions with period 2 with respect to every variable x_i, $i = 1, \ldots, n$.

3 Characteristics of Activation Functions

In this Section, we prove three theorems.

Theorem 1 *Suppose that g is a continuous function, and $g \in S'(\mathbf{R}^1)$, then $g \in (TW)$, if and only if g is not a polynomial.* $\square$

Theorem 2 *If σ is a bounded sigmoidal function, then $\sigma \in (TW)$.* $\square$

Theorem 3 *Suppose that K is a compact set in $\mathbf{R}^n$, U is a compact set in $C(K)$, $g \in (TW)$, then for any $\epsilon > 0$, there exist a positive integer N, real numbers θ_i,* *vectors* $\omega_i \in \mathbf{R}^n$, $i = 1, \ldots, N$, *which are independent of* $f \in C(K)$ *and constants* $c_i(f)$, $i = 1, \ldots, N$ *depending on* f, *such that*

$$|f(x) - \sum_{i=1}^{N} c_i(f) g(\omega_i \cdot x + \theta_i)| < \epsilon \tag{1}$$

holds for all $x \in K$ *and* $f \in U$. *Moreover, each* $c_i(f)$ *is a linear continuous functional defined on* U. $\square$

Remark 1. Theorem 3 shows that for a function (continuous or discontinuous) to be qualified as an activation function, a sufficient condition is that it belongs to (TW) class. Therefore, in order to prove that a neural network is capable of approximating any continuous function of n variables, all we need to do is to deal with the case $n = 1$, thus we have reduced the complexity of the problem in terms of its dimensionality. Moreover, by examining the approximated function $f(x_1, \ldots, x_n) = f(x_1, 0, \ldots, 0) = f^*(x_1)$, where $f^*(x_1)$ is a continuous function of one variable, it is straightforward to see that the condition is also a necessary one.

Remark 2. The equiuniform convergence property in Theorem 3 will play a crucial role in approximating nonlinear operators by neural networks.

Remark 3. When a sigmoidal function is used as an activation in a neural network, Theorem 2 shows that the only necessary condition imposed on is its boundedness. In contrast, in almost all other papers ([1], [2], [3], [4], [5], [6], [7], [8]), sigmoidal functions must be assumed to be either continuous or monotone.

Remark 4. In [11], some result similar to Theorem 1 was obtained under more restrictions imposed on g, *i.e.*, there are positive integer N and a constant C_N, such that $|(1+|x|)^{-N}g(x)| \leq C$ for all $x \in \mathbf{R}^1$. This restriction is essential for [11], for the proof in [11] depends heavily on a variation of Paley-Wiener Theorem. However, in Theorem 1, we only assume that $g \in C(\mathbf{R}^1) \cap S'(\mathbf{R}^n)$, which is weaker than the assumptions used in [11].

Proof of Theorem 1. We will prove by contradiction. If all the linear combinations $\sum_{i=1}^n c_i g(\lambda_i x + \theta_i)$ are not dense in $C[a,b]$, then Hahn-Banach extension theorem and Riesz representation of linear continuous functionals show that there is a signed Borel measure $d\mu$ with $\mathrm{supp}(d\mu) \subseteq [a,b]$ and

$$\int_{\mathbf{R}^1} g(\lambda x + \theta)\, d\mu(x) = 0 \tag{2}$$

for all $\lambda \neq 0$ and $\theta \in \mathbf{R}^1$. Take any $w \in S(\mathbf{R}^1)$, then

$$\int_{\mathbf{R}^1} w(\theta)\, d\theta \int_{\mathbf{R}^1} g(\lambda x + \theta)\, d\mu(x) = 0. \tag{3}$$

Let $\lambda x + \theta = u$ and change order of integration, we have

$$\int_{\mathbf{R}^1} g(u) \int_{\mathbf{R}^1} w(\theta)\, d\mu(\frac{u-\theta}{\lambda}) = 0 \tag{4}$$

which is equivalent to

$$\hat{g}(\hat{w}(\cdot)\hat{d\mu}(\lambda\cdot)) = 0 \tag{5}$$

where $\hat{g}$ represents Fourier transform of g in the sense of tempered distribution, and (5) is also understood in the sense of distribution (see [13]). In order that the left hand side of (5) makes sense, we have to show that $\hat{w}(t)\widehat{d\mu}(\lambda t) \in S(\mathbf{R}^1)$. Since $\mathrm{supp}(d\mu) \subseteq [a,b]$, it is straightforward to show that $\widehat{d\mu}(t) \in C^\infty(\mathbf{R}^1)$ and for each $k = 1, 2, \ldots$, there is a constant c_k such that

$$|\frac{\partial^k}{\partial t^k}\widehat{d\mu}(t)| \leq c_k\ . \tag{6}$$

Consequently, $\hat{w}(t)\widehat{d\mu}(t) \in S(\mathbf{R}^1)$.

Since $d\mu \not\equiv 0$ and $\widehat{d\mu}(t) \in C^\infty(\mathbf{R}^1)$, hence there exists some $t_0 \neq 0$ with some neighborhood $(t_0-\delta, t_0+\delta)$ such that $\widehat{d\mu}(t) \neq 0$ for all $t \in (t_0-\delta, t_0+\delta)$. Now, if $t_1 \neq 0$, let $\lambda = \frac{t_0}{t_1}$, then $\widehat{d\mu}(\lambda t) \neq 0$ for all $t \in (t_1-\frac{\delta}{\lambda}, t_1+\frac{\delta}{\lambda})$. Take any $\hat{w} \in C_c^\infty(t_0-\frac{\delta}{2\lambda}, t_0+\frac{\delta}{2\lambda})$, then $\hat{w}(t)/\widehat{d\mu}(\lambda t) \in S(\mathbf{R}^1)$, and by (5)

$$\hat{g}(\hat{w}(\cdot)) = \hat{g}(\frac{\hat{w}(\cdot)}{\widehat{d\mu}(\lambda\cdot)}\widehat{d\mu}(\lambda\cdot)) = 0 \tag{7}$$

Previous argument shows that for any fixed point t^*, there is a neighborhood $[t^*-\eta, t^*+\eta]$ such that $\hat{g}(\hat{w}(\cdot)) = 0$ for all $\hat{w}$ with compact support $[t^*-\eta, t^*+\eta]$, *i.e.* $\text{supp}(\hat{g}) \subseteq \{0\}$. By the distribution theory, $\hat{g}$ is some linear combination of δ-Dirac function and its derivatives, which is equivalent to that g is a polynomial. Theorem 1 is proved. $\square$

Proof of Theorem 2 can be found in [9]. Here we only give a brief proof for the completeness of this paper.

Proof of Theorem 2. Without loss of generality, we can assume that $[a,b] = [0,1]$. Since f is continuous on $[-1,1]$ for any $\epsilon > 0$, there is an integer $M > 0$, such that $|f(x') - f(x'')| < \epsilon/4$, provided that $x', x'' \in [-1,1]$ and $|x' - x''| < 1/M$.

Divide $[-1,1]$ into $2M$ equal segments, each has length of $1/M$. Let

$$-1 = x_0 < x_1 < \ldots < x_M = 0 < x_{M+1} < \ldots < x_{2M} = 1 \tag{8}$$

and $t_i = \frac{1}{2}(x_i + x_{i+1})$, $t_{-1} = -1 - \frac{1}{2M}$. From the assumption, there exists $W > 0$, such that if $u > W$, then $|\sigma(u) - 1| < \frac{1}{M^2}$; if $u < -W$, then $|\sigma(u)| < \frac{1}{M^2}$. Let $K > 0$

be such that $K \cdot \frac{1}{2M} > W$. Construct

$$g(x) = f(-M)\sigma(K(x - t_{-1})) + \sum_{i=1}^{N}[f(x_i) - f(x_{i-1})]\sigma(K(x - t_{i-1})) \tag{9}$$

then we can prove

$$|g(x) - f(x)| < \epsilon \qquad \text{for all } x \in [-1, 1] \tag{10}$$

Theorem 2 is thus proved. □

Prior to proving Theorem 3, we need to establish the following lemmas.

Lemma 1 *Suppose that K is a compact set in $\mathbf{R}^n$, $f \in C(K)$, then there is a continuous function $E(f) \in C(\mathbf{R}^n)$, such that (1) $f(x) = E(f)(x)$ for all $x \in K$; (2) $\sup_{x\in\mathbf{R}^n} |E(f)(x)| \leq \sup_{x\in K} |f(x)|$; (3) there is a constant c such that*

$$\sup_{|x'-x''|<\delta} |E(f)(x') - E(f)(x'')| \leq c \sup_{\substack{x'-x''<\delta \\ x',x''\in K}} |f(x') - f(x'')| \tag{11}$$

□

Proof. The proof of Lemma 1 can be found in [14] (p. 175). □

Lemma 2 *[15] V is a compact set in $C(K)$, if and only if*

1. *V is a closed set in $C(K)$.*

2. *There is a constant M, such that $\|f(x)\|_{C(K)} \leq M$ for all $f \in V$.*

3. *V is equicontinuous,* i.e. *for any $\epsilon > 0$, there is a $\delta > 0$ such that $|f(x') - f(x'')| < \epsilon$ for all $f \in V$, provided that x', $x'' \in K$ and $\|x' - x''\|_K < \delta$.* □

Lemma 3 *Suppose that K is a compact set in $\mathbf{I}^n = [0,1]^n$, V is a compact set in $C(K)$, then V can be extended to a compact set in $C_p[-1,1]^n$.* $\square$

Proof. By Lemmas 1 and 2, V can be extended to be a compact set V_1 in $C[0,1]^n$. Now, for every $f \in V_1$, define an even extension of f as follows

$$f^*(x_1, \dots, x_k, \dots, x_n) = f(x_1, \dots, -x_k, \dots, x_n) \tag{12}$$

then $U = \{f^* : f \in V_1\}$ is the required compact set in $C_p[-1,1]^n$. $\square$

Lemma 4 *Suppose that U is a compact set in $C_p[-1,1]$,*

$$B_R(f;x) = \sum_{|m|\le R} (1 - \frac{|m|^2}{R^2})^\alpha c_m(f) e^{i\pi m\cdot x} \tag{13}$$

is the Bochner-Riesz means of Fourier series of f, where $m = (m_1, \dots, m_n)$, $|m|^2 = \sum_{i=1}^n |m_i|^2$, $c_m(f)$ are Fourier coefficients of f, then for any $\epsilon > 0$, there is $R > 0$ such that

$$|B_R(f;x) - f(x)| < \epsilon \tag{14}$$

for every $f \in U$ and $x \in [-1,1]^n$, provided that $\alpha > (n-1)/2$. $\square$

Proof. The proof of Lemma 4 can be found in [13]. $\square$

Proof of Theorem 3. Without loss of generality, we can assume that $K \subseteq [0,1]^n$. By Lemma 3, we can assume that $K = [-1,1]^n$ and $U \subseteq C_p[-1,1]^n$. By Lemma 4, for any $\epsilon > 0$, there exists $R > 0$, such that for any $x = (x_1, \dots, x_n) \in [-1,1]^n$ and $f \in U$, there holds

$$|\sum_{|m|\le R} (1 - \frac{|m|^2}{R^2})^\alpha c_{m_1\cdots m_n}(f^*) exp(i\pi(m_1x_1 + \cdots + m_nx_n)) - f^*(x_1, \dots, x_n)| < \frac{\epsilon}{2} \tag{15}$$

By the definition of the Fourier coefficients and evenness of $f^*(x)$, we can rewrite (15) as

$$|\sum_{|m|\leq R} d_{m_1\cdots m_n}\cos(\pi(m_1x_1+\cdots+m_nx_n)-f^*(x_1,\ldots,x_n)|<\frac{\epsilon}{2} \tag{16}$$

where $d_{m_1\cdots m_n}$ are real numbers. It is obvious that for every $x\in[-1,1]^n$, there is a unique $u\in[-\sqrt{n}\pi R,\sqrt{n}\pi R]$, such that

$$u=\pi m\cdot x=\pi(m_1x_1+\ldots+m_nx_n)\ . \tag{17}$$

where $m=(m_1,\ldots,m_n)$. Since $\cos(u)$ is a continuous function in $[-\sqrt{n}\pi R,\sqrt{n}\pi R]$ and $g\in(TW)$, we can find an integer M, real numbers s_j, η_j and ξ_j, $j=1,\ldots,M$, such that

$$|\sum_{j=1}^{M} s_jg(\xi_ju+\eta_j)-\cos(u)|<\frac{\epsilon}{2L} \tag{18}$$

holds uniformly for all $u\in[-\sqrt{n}\pi R,\sqrt{n}\pi R]$, where $L=\sum_{|m|\leq R}|d_{m_1,\ldots,m_n}|$. Thus

$$|\sum_{j=1}^{M} s_jg(\xi_j\pi(m\cdot x)+\eta_j)-\cos(\pi m\cdot x)|<\frac{\epsilon}{2L} \tag{19}$$

holds for $x\in[-1,1]^n$. Substituting (19) into (16), we conclude that there exist N, c_i, $\theta_i\in\mathbf{R}$, $\omega_i\in\mathbf{R}^n$, $i=1,\ldots,N$, such that

$$|f^*(x)-\sum_{i=1}^{N} c_i(f^*)f^*(\omega_i\cdot x+\theta_i)|<\epsilon \tag{20}$$

is true for all $x\in[-1,1]^n$ and $g\in U$. Thus

$$|f(x)-\sum_{i=1}^{N} c_i(f)g(\omega_i\cdot x+\theta_i)|<\epsilon$$

is true for all $x\in[0,1]^n$ and $f\in V$.

It is obvious that for each fixed $m = (m_1, \dots, m_n)$, the Fourier coefficient

$$\int_{-1}^{1} \cdots \int_{-1}^{1} e^{-i\pi(m_1 x_1 + \cdots + m_n x_n)} g(x_1, \dots, x_n)\, dx_1 \cdots dx_n$$

is a continuous functional defined on U (and also a continuous functional defined on V), and $c_i(f)$, being a finite linear combination of the Fourier coefficients of f^*, is surely a continuous functional defined on V. The proof of Theorem 3 is completed.

□

4 Approximation to Nonlinear Continuous Functionals and Maps

In this section, we will discuss the problem of approximating nonlinear continuous functionals and operators by neural network computation. The main results are as follows.

Theorem 4 *Suppose that $g \in (TW)$, X is a Banach Space, $K \subseteq X$ is a compact set, V is a compact set in $C(K)$, f is a continuous functional defined on V, then for any $\epsilon > 0$, there are an positive integer N, m points $x_1, \dots, x_m \in K$, and real constants c_i, θ_i, ξ_{ij}, $i = 1, \dots, N$, $j = 1, \dots, m$, such that*

$$|f(u) - \sum_{i=1}^{N} c_i g(\sum_{j=1}^{m} \xi_{ij} u(x_j) + \theta_i)| < \epsilon \tag{21}$$

holds for all $u \in V$. □

Theorem 5 *Suppose that $g \in (TW)$, X is a Banach Space, $K_1 \subseteq X$, $K_2 \subseteq \mathbf{R}^n$ are two compact sets in X and $\mathbf{R}^n$ respectively, V is a compact set in $C(K_1)$, G is a* *nonlinear continuous operator, which maps V into $C(K_2)$, then for any $\epsilon > 0$, there are positive integers M, N, m, constants c_i^k, ζ_k, $\xi_{ij}^k \in \mathbf{R}$, points $\omega_k \in \mathbf{R}^n$, $x_j \in K_1$, $i = 1, \ldots, M$, $k = 1, \ldots, N$, $j = 1, \ldots, m$, such that*

$$|G(u)(y) - \sum_{k=1}^{N}\sum_{i=1}^{M} c_i^k g(\sum_{j=1}^{m} \xi_{ij}^k u(x_j) + \theta_i^k) g(\omega_k \cdot y + \zeta_k)| < \epsilon \tag{22}$$

holds for all $u \in V$ and $y \in K_2$. □

Following lemmas are well known and will be used in the proof of Theorems 4 and 5.

Lemma 5 *Let X be a Banach Space and $K \subseteq X$, then K is a compact set if and only if the following two conditions are satisfied simultaneously: (1) K is a closed set in X; (2) for any $\delta > 0$, there is a δ-net $N(\delta) = \{x_1, \ldots, x_{n(\delta)}\}$,* i.e. *for any $x \in K$, there constitute an $x_k \in N(\delta)$ such that $\|x - x_k\|_X < \delta$.* □

Lemma 6 *If $V \subseteq C(K)$ is a compact set in $C(K)$, then it is uniformly bounded and equicontinuous,* i.e. *(1) There is $A > 0$ such that $\|u(x)\|_{C(K)} \leq A$ for all $u \in V$ and (2) for any $\epsilon > 0$, there is $\delta > 0$ such that $|u(x') - u(x'')| < \epsilon$ for all $u \in V$, provided that $\|x' - x''\|_X < \delta$.* □

Now pick a sequence $\epsilon_1 > \epsilon_2 > \cdots > \epsilon_n \to 0$, then we can find another sequence $\delta_1 > \delta_2 > \cdots > \delta_n \to 0$, such that $|f(u) - f(v)| < \epsilon_k$ for all $f \in V$, provided that $u, v \in V$ and $\|u - v\|_{C(K)} < 2\delta_k$, for f is a continuous functional defined on a compact set V.

By Lemma 6, we can also find $\eta_1 > \eta_2 > \cdots \eta_n \to 0$ such that $|u(x') - u(x'')| < \delta_k$ for all $u \in V$, whenever $x', x'' \in K$ and $\|x' - x''\|_K < \eta_k$.

By induction and rearrangement, we can find a sequence $\{x_i\}_{i=1}^{\infty}$ with each $x_i \in K$ and a sequence of positive integers $n(\eta_1) < n(\eta_2) < \ldots < n(\eta_k) \to \infty$, such that the first $n(\eta_k)$ elements $N(\eta_k) = \{x_1, \ldots, x_{n(\eta_k)}\}$ is an η_k-net in K.

For each η_k-net, define functions

$$T^*_{\eta_k,j}(x) = \begin{cases} 1 - \frac{\|x - x_j\|_k}{\eta_k} & \text{if } \|x - x_j\|_X \le \eta_k \\ 0 & \text{otherwise} \end{cases} \tag{23}$$

and

$$T_{\eta_k,j}(x) = \frac{T^*_{\eta_k,j}(x)}{\sum_{j=1}^{n(\eta_k)} T^*_{\eta_k,j}(x)} \tag{24}$$

for $j = 1, \ldots, n(\eta_k)$. It is easy to verify that $\{T_{\eta_k,j}(x)\}$ is a partition of unity, *i.e.*

$$0 \le T_{\eta_k,j}(x) \le 1 \tag{25}$$

$$\sum_{j=1}^{n(\eta_k)} T_{\eta_k,j}(x) \equiv 1 \tag{26}$$

$$T_{\eta_k,j}(x) = 0 \qquad \text{if } \|x - x_j\|_X > \eta_k \ . \tag{27}$$

For each $u \in V$, define a function

$$u_{\eta_k}(x) = \sum_{j=1}^{n(\eta_k)} u(x_j) T_{\eta_k,j}(x) \tag{28}$$

and sets $V_{\eta_k} = \{u_{\eta_k} : u \in V\}$ and $V^* = V \cup (\cup_{k=1}^{\infty} V_{\eta_k})$. We then have the following result.

Lemma 7

1. *For each fixed k, V_{η_k} is a compact set in a subspace of dimension $n(\eta_k)$ in $C(K)$.*

2. *For every $u \in V$, there holds*

$$\|u - u_{\eta_k}\|_{C(K)} < \delta_k \tag{29}$$

3. *V^* is a compact set in $C(K)$.* $\square$

Proof. We will prove the three propositions individually as follows.

1. For a fixed k, let $u_{\eta_k}^{(i)}$, $i = 1, 2, \ldots$, be a sequence in V_{η_k} and $u^{(i)}$ be a sequence in V, such that

$$u_{\eta_k}^{(i)} = \sum_{j=1}^{n(\eta_k)} u^{(i)}(x_j) T_{\eta_k,j}(x) \ . \tag{30}$$

Since V is a compact set, there is a subsequence $u^{(i_l)}(x)$, which converges to some $u \in V$, then it is obvious that $u_{\eta_k}^{(i_l)}(x)$ converges to $u_{\eta_k}(x) \in V_{\eta_k}$, *i.e.* V_{η_k} is a compact subset in $C(K)$.

2. By the definition and the property of unity partition, we have

$$\begin{aligned} u(x) - u_{\eta_k}(x) &= \sum_{j=1}^{n(\eta_k)} [u(x) - u(x_j)] T_{\eta_k,j}(x) \\ &= \sum_{\|x - x_j\|_X \le \eta_k} [u(x) - u(x_j)] T_{\eta_k,j}(x) \ . \end{aligned} \tag{31}$$

Consequently,

$$\|u(x) - u_{\eta_k}(x)\|_X \le \delta_k \sum_{j=1}^{n(\eta_k)} T_{\eta_k,j}(x) = \delta_k \qquad \text{for all } u \in V \ . \tag{32}$$

3. Suppose $\{u^i\}_{j=1}^{\infty}$ is a sequence in V^*. If there is a subsequence $\{u^{i_l}\}_{l=1}^{\infty}$ of $\{u^i\}_{j=1}^{\infty}$ with all $u^{i_l} \in V$, $l = 1, \ldots$, then by the fact that V is compact, there is

a subsequence of $\{u^{i_l}\}_{l=1}^{\infty}$, which converges to some $u \in V$. Otherwise, to each u^i, there corresponds a positive integer $k(i)$ such that $u^i = v_{\eta_{k(i)}}$. There are two possibilities: (i) We can find infinite i_l and a fixed k_0 such that $\eta_{k(i_1)} = \eta_{k(i_2)} = \cdots = \eta_{k(i_l)} = \cdots = \eta_{k_0}$, i.e. $u^{i_l} \in V_{\eta_{k_0}}$ for all i_l. By proposition 1. of this lemma, $V_{n(\eta_{k_0})}$ is a compact set, there is a subsequence of $\{v^i_{n(\eta_{k(i)})}\}$, which converges to some $v \in V_{n(\eta_{k_0})}$, *i.e.* there is a subsequence of $\{u^i\}$ converging to $v \in V_{n(\eta_{k_0})}$. (ii) There are sequences $i_1 < i_2 < \ldots \to \infty$ and $k(i_1) < k(i_2) < \ldots \to \infty$ such that $u^{i_l} \in V_{n(\eta_{k(i_l)})}$. Let $v^{i_l} \in V$ be such that

$$u^{i_l}(x) = \sum_{j=1}^{n(\eta_{k(i_l)})} v^{i_l}(x_j) T_{\eta_{k(i_l)},j}(x) \ . \tag{33}$$

Since $V^{i_l} \in V$ and V is compact, we see that there is a subsequence of $\{v^{i_l}\}_{l=1}^{\infty}$, which converges to some $v \in V$. By the proposition 2. of this lemma, the corresponding subsequence of $\{u^{i_l}\}_{l=1}^{\infty}$ also converges to v. Thus the compactness of V^* is proved. □

Proof of Theorem 4. By Tietze Extension Theorem, we can define a continuous functional on V^* such that

$$f^*(x) = f(x) \qquad \text{if } x \in V \tag{34}$$

Because f^* is a continuous functional defined on the compact set V^*, therefore for any $\epsilon > 0$, we can find a $\delta > 0$ such that $|f^*(u) - f^*(v)| < \epsilon/2$ provided that $u, v \in V^*$ and $\|u - v\|_{C(K)} < \delta$.

Let k be fixed such that $\delta_k < \delta$, then by (29) for every $u \in V$,

$$\|u - u_{\eta_k}\|_X < \delta_k \tag{35}$$

which implies

$$|f^*(u) - f^*(u_{\eta_k})| < \epsilon/2 \tag{36}$$

for all $u \in V$.

By proposition 1. of Lemma 7, we see that $f^*(u_{\eta_k})$ is a continuous functional defined on the compact set V_{η_k} in $\mathbf{R}^{n(\eta_k)}$. By Theorem 3, we can find N, c_i, ξ_{ij}, θ_i, $i = 1, \ldots, N$, $j = 1, \ldots, n(\eta_k)$, such that

$$|f^*(u_{\eta_k}) - \sum_{i=1}^{N} c_i g(\sum_{j=1}^{n(\eta_k)} \xi_{ij} u(x_j) + \theta_i)| < \epsilon/2 \ . \tag{37}$$

Combining it with (36), we conclude that

$$|f(u) - \sum_{i=1}^{N} c_i g(\sum_{j=1}^{m} \xi_{ij} u(x_j) + \theta_i)| < \epsilon \tag{38}$$

where $m = n(\eta_k)$. Thus, Theorem 4 is proved. □

Proof of Theorem 5. From the assumption that G is a continuous operator which maps a compact set V in $C(K_1)$ into $C(K_2)$, it is straightforward to prove that the range $G(V) = \{G(u) : u \in V\}$ is also a compact set in $C(K_2)$. By Theorem 3, for any $\epsilon > 0$, there are a positive integer N, real numbers $c_k(G(u))$ and ζ_k, vectors $\omega_k \in \mathbf{R}^n$, $k = 1, \ldots, N$, such that

$$|G(u)(y) - \sum_{k=1}^{N} c_k(G(u)) g(\omega_k \cdot y + \zeta_k)| < \epsilon/2 \tag{39}$$

holds for all $y \in K_2$ and $u \in V$.

Since G is a continuous operator, combining with the last proposition of Theorem 3, we conclude that for each $k = 1, \ldots, N$, $c_k(G(u))$ is a continuous functional defined on V. Repeatedly applying Theorem 4, for each $k = 1, \ldots, N$, we can find positive integers N_k, m_k, constants c_i^k, ξ_{ij}^k, $\theta_i^k \in \mathbf{R}$ and $x_j \in K_1$, $i = 1, \ldots, N_k$, $j = 1, \ldots, m_k$, such that

$$|c_k(G(u)) - \sum_{i=1}^{N_k} c_i^k g(\sum_{j=1}^{m_k} \xi_{ij}^k u(x_j) + \theta_i^k)| < \frac{\epsilon}{2L} \tag{40}$$

holds for all $k = 1, \ldots, N$ and $u \in V$, where

$$L = \sum_{k=1}^{N} \sup_{y \in K_2} |g(\omega_k \cdot y + \zeta_k)| \, . \tag{41}$$

Substituting (40) into (39), we obtain that

$$|G(u)(y) - \sum_{k=1}^{N} \sum_{i=1}^{N_k} c_i^k g(\sum_{j=1}^{m_k} \xi_{ij}^k u(x_j) + \theta_i^k) g(\omega_k \cdot y + \zeta_k)| < \epsilon \tag{42}$$

holds for all $u \in V$ and $y \in K_2$.

Let $M = \max_k\{N_k\}$, $m = \max_k\{m_k\}$ and for all $N_k < i \le M$, let $c_i^k = 0$. For all $m_k < j \le m$, let $\xi_{ij}^k = 0$. Thus (42) can be rewritten as

$$|G(u)(y) - \sum_{k=1}^{N} \sum_{i=1}^{M} c_i^k g(\sum_{j=1}^{m} \xi_{ij}^k u(x_j) + \theta_i^k) g(\omega_k \cdot y + \zeta_k)| < \epsilon \tag{43}$$

holds for all $u \in V$ and $y \in K_2$. This completes the proof of Theorem 5. $\square$

A graphical representation of Theorem 5 is shown in Fig. 1.

5 Application to Nonlinear Dynamical Systems

In [12], we discussed the problem of approximating the output of a dynamical system at a fixed point (or time) by neural networks. As a direct application of Theorem 5, we can use neural networks to approximate the output *as a whole* of a nonlinear dynamical system. Indeed, built upon the several keystone theorems proved earlier in Section 4, our result on this topic follows naturally.

The significance of the previous results lies in that we can use neural networks to identify a system (linear or nonlinear). The procedure is as follows:

Let a system be $V = KU$, where U is the input, V is the output and K is the system to be identified.

Suppose that according to some prior knowledge or experiments, we know several input-output relationships $V_1 = KU_1, \ldots, V_n = KU_n$. Generally, they can be expressed by discrete data sets $\{u_s(x_j), s = 1, \ldots, n, j = 1, \ldots, m\}$, $\{v_s(y_l), s = 1, \ldots, n, j = 1, \ldots, L\}$. Using these data, and by Theorem 5, we can construct a functional

$$E = \sum_{l=1}^{L} \sum_{s=1}^{n} |V_s(y_l) - \sum_{k=1}^{N} \sum_{i=1}^{M} C_i^k g(\sum_{j=1}^{m} \xi_{i,j}^k u_s(x_j) + \theta_i^k) g(\omega_k \cdot y_l + \zeta_k)|^2 \tag{44}$$

Parameters C_i^k, $\xi_{i,j}^k$, θ_i^k, ω_k, ζ can be determined by minimizing E (for example, by using back-propagation algorithm). Then the equation

$$v(y) = \sum_{k=1}^{N} \sum_{i=1}^{M} C_i^k g(\sum_{j=1}^{m} \xi_{i,j}^k u(x_j) + \theta_i^k) g(\omega_k \cdot y + \zeta_k) \tag{45}$$

can be viewed as an approximant of $V(y) = (KU)(y)$, and so identifies the system K.

If the system is linear, then E, $V(y)$ can be simplified as

$$E = \sum_{l=1}^{L} \sum_{s=1}^{n} |V_s(y_l) - \sum_{k=1}^{N} \sum_{i=1}^{M} \sum_{j=1}^{m} \xi_{i,j}^k u(x_j) g(\omega_k \cdot y_l + \zeta_k)|^2 \tag{46}$$

$$v(y) = \sum_{k=1}^{N} \sum_{i=1}^{M} \sum_{j=1}^{m} \xi_{i,j}^{k} u(x_j) g(\omega_k \cdot y + \zeta_k). \tag{47}$$

The larger the values of n, L, m are, the better accuracy we will obtain for this approximation.

Therefore, we have pointed to a way of constructing neural network models for identifying dynamic systems.

Acknowledgements. The authors wish to express their gratefulness to the reviewers for their valuable comments and suggestions on revising this paper.

6 Conclusion

In this paper, the problem of approximating functions of several variables, functionals and nonlinear operators are thoroughly studied. The necessary and sufficient condition for a continuous function in $S'(\mathbf{R}^1)$ to be qualified for an activation function is given, which is a broad generalization of previous results ([1]- [8], especially [11]). It is also pointed out that to prove neural network approximation capability, one needs only to treat the one dimensional case. As applications, we show how to construct neural networks to approximate the output of a dynamical system *as a whole*, not merely at a fixed point, thus show the capability of neural network in identifying dynamic systems. Moreover, we point out that using existing algorithms in literatures (for example, back-propagation algorithm), we can determine those parameters in the network, *i.e.* identify the system.

References

[1] A. Wieland and R. Leighten, "Geometric Analysis of Neural Network Capacity," in *IEEE First ICNN. 1*, pp. 385-392 (1987).

[2] B. Irie and S. Miyake, "Capacity of Three-layered Perceptrons," in *IEEE ICNN 1*, pp. 641-648, (1988).

[3] G. Cybenko, "Approximation by Superpositions of a Sigmoidal Function," in *Math. of Control, Signals and Systems*, Vol. 2, No. 4, pp. 303-314 (1989).

[4] S. M. Carroll and B. W. Dickinson, "Construction of Neural Nets using Radon Transform," in *IJCNN Proc. I*, pp. 607-611 (1989).

[5] K. Funahashi, "On the Approximate Realization of Continuous Mappings by Neural Networks," *Neural Networks*, pp. 183-192, Vol. 2, (1989).

[6] K. Hornik, M. Stinchcombe and H. White, "Multi-layer Feedforward Networks are Universal Approximators," *Neural Networks*, Vol. 2, pp. 359-366 (1989).

[7] K. Hornik, "Approximation Capabilities of Multilayer Feedforward Networks," *Neural Networks*, Vol. 4, pp. 251-257 (1991).

[8] V. Y. Kreinovich, "Arbitrary Nonlinearity is Sufficient to Represent All Functions by Neural Networks: a Theorem," *Neural Networks*, Vol. 4, pp. 381-383 (1991).

[9] Tianping Chen, Hong Chen and Ruey-wen Liu, "A Constructive Proof of Cybenko's Approximation Theorem and Its Extensions," pp. 163 - 168 in *Computing Science and Statistics* (editors LePage and Page), Proc. of the 22nd Symposium on the Interface (East Lansing, Michigan, May 1990), Springer-Verlag, ISBN 0-387-97719-8. Also submitted for publication.

[10] I. W. Sandberg, "Approximations for Nonlinear Functionals," *IEEE Trans. on Circuits and Systems*, Vol. 39, No. 1, pp. 65-67, Jan. (1992).

[11] H.N. Mhaskar and C.A. Micchelli, "Approximation by Superposition of Sigmoidal and Radial Basis Functions," *Advances in Applied Mathematics*, Vol. 13, pp. 350-373, (1992).

[12] Tianping Chen and Hong Chen, "Approximation to Continuous Functionals by Neural Networks with Application to Dynamical Systems," accepted by *IEEE Trans. on Neural Networks*, to appear.

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

[14] E. M. Stein, *Singular Integrals and Differentiability Properties of Functions*, Princeton University Press, (1970).

[15] J. Diedonne, *Foundation of Modern Analysis*, Academic Press : New York and London (1969), p. 142.

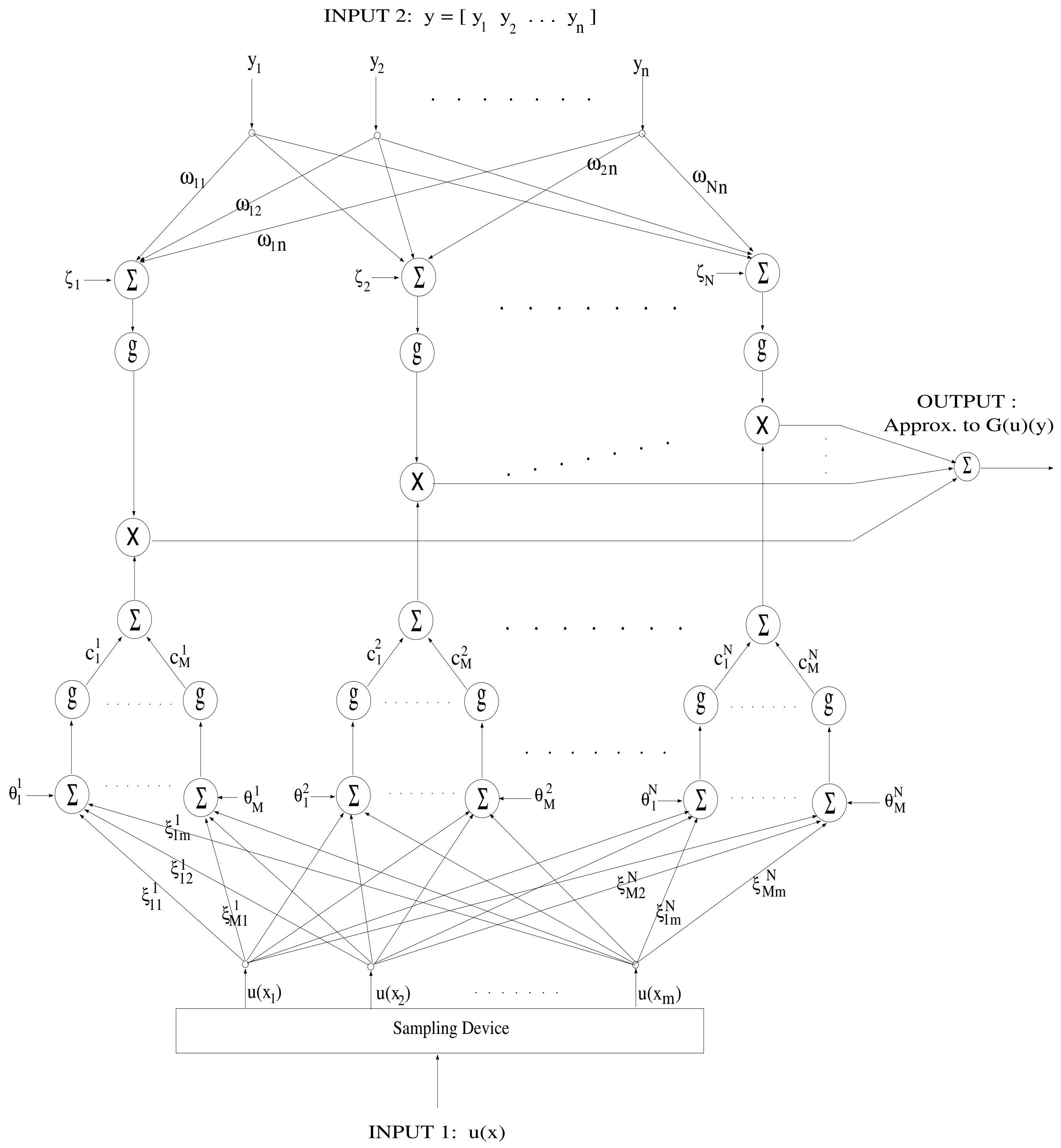

Figure 1: A neural network architecture of approximation to nonlinear operator $G(u)(y)$