单隐层神经网络的 L^p 同时逼近*

曹飞龙[1,2+]，李有梅[2,3]，徐宗本[2]

[1](绍兴文理学院 数理信息学院,浙江 绍兴 312000)

[2](西安交通大学 理学院信息与系统科学研究所,陕西 西安 710049)

[3](山西大学 计算机系,山西 太原 030006)

L^p Simultaneous Approximation by Neural Networks with One Hidden Layer

CAO Fei-Long[1,2+], LI You-Mei[2,3], XU Zong-Ben[2]

[1](School of Mathematical and Physical Information, Shaoxing College of Arts and Sciences, Shaoxing 312000, China)

[2](Institute for Information and System Sciences, Faculty Science, Xi'an Jiaotong University, Xi'an 710049, China)

[3](Department of Computer, Shanxi University, Taiyuan 030006, China)

+ Corresponding author: E-mail: flcao@263.net

Received 2002-06-19; Accepted 2003-03-04

Cao FL, Li YM, Xu ZB. L^p simultaneous approximation by neural networks with one hidden layer. *Journal of Software*, 2003,14(11):1869~1874.
http://www.jos.org.cn/1000-9825/14/1869.htm

Abstract: It is shown in this paper by a constructive method that for any Lebesgue integrable functions defined on a compact set in a multidimensional Euclidian space, the function and its derivatives can be simultaneously approximated by a neural network with one hidden layer. This approach naturally yields the design of the hidden layer and the convergence rate. The obtained results describe the relationship between the rate of convergence of networks and the numbers of units of the hidden layer, and generalize some known density results in uniform measure.

Key words: neural network; simultaneous approximation; hidden layer design; rate of convergence; Lebesgue measure

摘 要: 用构造性的方法证明对任何定义在多维欧氏空间紧集上的勒贝格可积函数以及它的导数可以用一个单隐层的神经网络同时逼近.这个方法自然地得到了网络的隐层设计和收敛速度的估计,所得到的结果描述了网络收敛速度与隐层神经元个数之间的关系,同时也推广了已有的关于一

* Supported by the National Natural Science Foundation of China under Grant No.69975016 (国家自然科学基金); the Foundation of Key Item of Science and Technology of the Ministry of Education of China under Grant No.03142 (国家教育部科学技术重点项目基金); the Foundation of Higher School of Ningxia Province of China under Grant No.JY2002107 (宁夏高校科研基金)

CAO Fei-Long was born in 1965. He is a professor. He received his Ph.D. degree from Xi'an Jiaotong University in 2003. His current research interests include neural network theory and approximation theory of function. **LI You-Mei** was born in 1965. She is a Ph.D. candidate in Institute for Information and System Sciences, Faculty of Science, Xi'an Jiaotong University. Her main research interest is neural network theory. **XU Zong-Ben** was born in 1955. He is a professor and doctoral supervisor at the Xi'an Jiaotong University. His research areas are artificial intelligence and its application.

致度量下的稠密性结果.
关键词: 神经网络;同时逼近;隐层设计;收敛速度;勒贝格尺度
中图法分类号: TP18 文献标识码: A

1 Introduction

Various problems concerning the applications of neural networks in many different disciplines can be converted into the problems of approximation multivariate functions by superposition of neural activation function of the networks. Typically, a neural network with one hidden layer is expressed mathematically as

$$N(x)=\sum_{j=1}^{m}c_j\varphi(<w_j\cdot x>+\theta_j),\quad x\in R^s,\quad s\ge 1 \tag{1.1}$$

where $\theta_j\in R$ is the thresholds, $\langle w_j\cdot x\rangle$ inner product of x and w_j, $w_j\in R^s$ connection weights, $c_j\in R$ coefficients, and φ is the neural activation function of the network. Approximating multivariate functions by networks in Eq.(1.1) has been widely studied in recent years with various results, concerning density or complexity, established for different situations, for instance, by Cybenko[1], Hornik[2], Anastassiou[3], Leshno *et al.*[4], Chui *et al.*[5], Chen[6,7], Li[8], Cao and Xu[9] and many others.

Simultaneous approximation of multivariate continuous functions and their partial derivatives on compact sets was also studied in Refs.[8,10,11]. In particular, Li[8] used a constructive method based on the multivariate Bernstein operator and gave a density result on uniform simultaneous approximation of multivariate functions and their derivatives by neural network. All the results for simultaneous approximation, however, only concerned about the density in uniform measure. These results therefore contribute almost nothing to answering such important question as how many hidden-layer units are needed to approximate specific functions within certain specified error. On the other hand, it is well known that the L^p norm to measure the quality of approximation has a penetrating background in engineering, physics, etc.

In this paper, we will address the investigation for simultaneous approximation by neural networks with one hidden layer in Lebesgue measure. Our main result will describe the relationship between the rate of convergence of networks and the number of hidden layer nodes when the approximated functions are the Lipshcitz functions, which is different from the results of Refs.[8,10,11]. Furthermore, our approach is constructive, mainly based on the elementary Taylor expansion and the multivariate Bernstein-Durrmeyer operator (a modification of the Bernstein operator), which is much more realizable in computations.

2 Simultaneous Approximation by Univariate Polynomials

To facilitate the following discussion, we assume that r is a fixed integer, and let

$$H_r(x)=a_0+a_1x+\ldots+a_rx^r \tag{2.1}$$

denote the univariate polynomial of degree r defined on $[a,b]$, where $a_i\in R$. In this part, we will study the simultaneous approximation of H_r and its derivatives by networks.

Theorem 2.1. Let S be a compact subset of $R^s, s\ge 1$. Also, let $\varphi^{(m+1)}\in C(R), m\in N, \varphi^{(i)}$ bounded on R, and there exists a $\theta\in R$ such that $\varphi^{(i)}(\theta)\ne 0$ for $1\le i\le m+1$. $H_r(x)$ is a univariate polynomial given in (2.1), and $0\le m\le r$. Then for any $\varepsilon>0$, we can construct a neural network with one hidden layer, one input and $(r-m+1)$ units in the hidden layer: $N_n(x)=\sum_{i=1}^{n}c_i\varphi(\omega_i x+\theta),\ \ c_i,\omega_i\in R,\ \ n=r-m+1,$ such that

$$\left\|N_n^{(m)}-H_r^{(m)}\right\|_p<\varepsilon,\quad 1\le p\le\infty.$$

Proof. Let $N_n(x)=\sum_{j=1}^{n}c_j\varphi(\omega_j x+\theta),\ c_j,\omega_j\in R$, then we can choose some appropriate c_i, and ω_i, $i=1,2,\ldots,n$, such that for $N_n^{(m)}(0)+N_n^{(m+1)}(0)x+\ldots+\frac{1}{(r-m)!}N_n^{(r)}(0)x^{r-m}=\frac{r!}{(r-m)!}a_r x^{r-m}+\ldots+m!a_m=H_r^{(m)}(x)$. In fact, by $N_n^{(s)}(x)=\sum_{i=1}^{n}c_i\omega_i^s\varphi^{(s)}(\omega_i x+\theta)$, it is not difficult to obtain that

$$BC=A \tag{2.2}$$

where

$$B=\begin{pmatrix}\varphi^{(m)}(\theta)\omega_1^m & \varphi^{(m)}(\theta)\omega_2^m & \cdots & \varphi^{(m)}(\theta)\omega_n^m\\ \varphi^{(m+1)}(\theta)\omega_1^{m+1} & \varphi^{(m+1)}(\theta)\omega_2^{m+1} & \cdots & \varphi^{(m+1)}(\theta)\omega_n^{m+1}\\ \cdots & \cdots & \cdots & \cdots\\ \varphi^{(r)}(\theta)\omega_1^r & \varphi^{(r)}(\theta)\omega_2^r & \cdots & \varphi^{(r)}(\theta)\omega_n^r\end{pmatrix},\quad C=(c_1,c_2,\ldots,c_n)^T,\ A=(m!a_m,(m+1)!a_{m+1},\ldots,r!a_r)^T.$$

Now, let $n=r-m+1$, then B is a square matrix. We choose $\omega_i=i\delta$, $0<\delta<1$, $i=1,2,\ldots,n$, thus, $\omega_i\ne\omega_j$ for $i\ne j$. Recalling that $\varphi^{(s)}(\theta)\ne 0$ gives $|B|=\varphi^{(m)}(\theta)\varphi^{(m+1)}(\theta)\ldots\varphi^{(r)}(\theta)\omega_1^m\omega_2^m\ldots\omega_n^m\prod_{1\le j<i\le r-m+1}(\omega_i-\omega_j)\ne 0$, which implies that the Eq.(2.2) has a unique solution:

$$C=B^{-1}A. \tag{2.3}$$

For $n=r-m+1,\omega_j=j\delta, j=1,2,\ldots,n$, we have

$$B=\begin{pmatrix}\varphi^{(m)}(\theta)\delta^m & 0 & \cdots & 0\\ 0 & \varphi^{(m+1)}(\theta)\delta^{m+1} & \cdots & 0\\ \cdots & \cdots & \cdots & \cdots\\ 0 & 0 & \cdots & \varphi^{(r)}(\theta)\delta^r\end{pmatrix}\begin{pmatrix}1 & 2^m & \cdots & n^m\\ 1 & 2^{m+1} & \cdots & n^{m+1}\\ \cdots & \cdots & \cdots & \cdots\\ 1 & 2^r & \cdots & n^r\end{pmatrix}.$$

Set

$$D=\begin{pmatrix}1 & 2^m & \cdots & n^m\\ 1 & 2^{m+1} & \cdots & n^{m+1}\\ \cdots & \cdots & \cdots & \cdots\\ 1 & 2^r & \cdots & n^r\end{pmatrix},\quad E=\begin{pmatrix}\frac{1}{\varphi^{(m)}(\theta)\delta^m} & \cdots & 0\\ \cdots & \ddots & \cdots\\ 0 & \cdots & \frac{1}{\varphi^{(r)}(\theta)\delta^r}\end{pmatrix},$$

then $B^{-1}=D^{-1}E$, and D^{-1} is a known nonsingular square matrix dependent only on n, m and r. Let M_2 be the maximum of absolute value of the elements in D^{-1}, and $M_3=\max\left\{(|\varphi^{(i)}(\theta)|)^{-1}, i=m,m+1,\ldots,r\right\}$. So, for given θ,r and m, then M_2 and M_3 are the positive constants. By (2.3), we have

$$\left|c_j\right|\le M_2M_3\sum_{i=m}^{r}i!\left|a_i\right|\delta^{-i},\quad j=1,2,\ldots,n. \tag{2.4}$$

Using Taylor expansion gives $N_n^{(m)}(x)=N_n^{(m)}(0)+N_n^{(m+1)}(0)x+\ldots+\frac{1}{(r-m)!}N_n^{(r)}(0)x^{r-m}+R_r(x)$, where $R_r(x)=\frac{1}{(r-m+1)!}N_n^{(r+1)}(\xi)x^{r-m+1},\ \xi\in[a,b]$. Let $\left|\varphi^{(r+1)}(x)\right|\le M_0$, then $R_r(x)\le\frac{M_1^{r-m+1}}{(r-m+1)!}\sum_{j=1}^{n}\left|c_j\right|j^{r+1}\delta^{r+1}M_0\le M\delta$, here we used (2.4) and the constant $M=\frac{M_1^{r-m+1}}{(r-m+1)!}M_2M_3M_0\left(\sum_{i=m}^{r}i!\left|a_i\right|\right)\left(\sum_{j=1}^{r-m+1}j^{r+1}\right)$. Finally, let $\delta<\frac{\varepsilon}{M+1}$, then $\left|R_r(x)\right|<\varepsilon$. The proof of Theorem 2.1 is complete.

3 Bernstein-Durrmeyer Operator Defined on Simplex

For description, we introduce some notations used in the sequel. Let S be a compact subset in $R^s, s\ge 1$.

Denote by Z_+^s the set of all non-negative multi-integers in R^S. For $x=(x_1,x_2,...,x_s)\in R^s$, and $m=(m_1,m_2,...,m_s)\in Z_+^s$, let $|x|=\sum_{i=1}^s x_i$, $|m|=\sum_{i=1}^s m_i$, $x^m=x_1^{m_1}x_2^{m_2}...x_s^{m_s}$, and $m!=m_1!m_2!...m_s!$. We say that $x\le y$, where $x,y\in R^s$, if $x_i\le y_i, 1\le i\le s$. By $L^p(s), 1\le p<+\infty$, we denote the space of Lebesgue measurable function on S for which the norm $\|f\|_p^p=\int_S|f(u)|^p du$ is finite. $L^\infty(S)=C(S)$ denotes the space of continuous functions on S equipped with the maximum norm. For a smooth function f on R^s, let $D^{|m|}f(x)$, $|m|=\sum_{i=1}^s m_i$, be the $|m|$-th order partial derivative of f. For the compact set S of R^s, by $S_p(S)$, we mean that there is an open Ω (depending on f) such that $S\in\Omega$, and $D^{|m|}f\in L^p(\Omega), 1\le p\le+\infty$. For $f\in L^p(S), 1\le p\le+\infty$, the modulus of smoothness of f is defined by $\omega(f,\delta)_p=\sup_{0<h\le\delta}\|f(\cdot+h)-f(\cdot)\|_p$. If $\omega(f,\delta)_p=O(\delta^\alpha)$, $0<\alpha\le1$, then we say that f is the Lipschitz functions and write $f\in Lip(\alpha)$.

To study the simultaneous approximation of multivariate functions and their derivatives by neural networks in Lebesgue measure, we will use a simultaneous approximation result for the Bernstein-Durrmeyer operator as an intermediate step. Let T be the simplex in R^s, the Bernstein-Durrmeyer operator is defined by $D_nf(x)=\sum_{|k|\le n}P_{n,k}(x)\Phi_{n,k}(f)$, where $P_{n,k}(x)=\frac{n!}{k!(n-|k|)!}x^k(1-|x|)^{n-|k|}$, $\Phi_{n,k}(f)=\frac{(n+s)!}{n!}\int_T P_{n,k}(u)f(u)\mathrm{d}u$. The polynomial operator, which was first introduced and studied by Derriennic[12] in 1985, is an integral modification of the well-known Bernstein operator. It is also a positive linear bounded operator form $L^p(T)$ to itself. Here we give a result on simultaneous approximation of the operator in Lebesgue measure, which will be key in the proof of our main result, and its proof is omitted.

Theorem 3.1. Suppose $f\in S_p(T), 1\le p\le\infty$, and $D^{|m|}f\in Lip(\alpha), 0<\alpha\le1$, $m\in Z_+^s, |m|\le n$, then

$$\left\|D^{|m|}D_nf-D^{|m|}f\right\|_p\le Cn^{-\alpha/2},$$

where and in the following C denotes a positive constant independent of n and f.

4 Main Result and Its Proof

In this section, we give and prove the main result of this paper. First, we give some notations that can be found in Section 4 of Ref.[8]. Note that there are $N_p=C_{p+s-1}^{s-1}$ multi-integers i in Z_+^s that satisfy $i_1+i_2+...+i_s=p$, and there are $I_p=C_{p+s-2}^{s-2}$ multi-integers j' in Z_+^{s-1} that satisfy $j_1'+j_2'+...+j_{s-1}'=p$. Clearly $N_p=N_{p-1}+I_p$. Denote by $j'_{N_{p-1+l}}$, $1\le l\le I_p$ the multi-integers j' in Z_+^{s-1} that satisfy $j_1'+j_2'+...+j_{s-1}'=p$. Hence, $\{j_l':1\le l\le N_p\}$ is a set of multi-integers j' in Z_+^{s-1} that satisfy $j_1'+j_2'+...+j_{s-1}'\le p$. For each p, let $i_l^{(p)}=(p-|j_l'|,j_l'), 1\le l\le N_p$. Hence, each $i_l^{(p)}$ is a multi-integer in Z_+^s with $|i_l^{(p)}|=p, 1\le l\le N_p$. Define $p_l=(1,j_l')$, $1\le l$, then for each fixed integer $p\ge0$, by setting $p_l^{(p)}=\frac{1}{2(1+p)}p_l, 1\le l\le N_p$, we have $|p_l^{(p)}|\le1/2$ for $1\le l\le N_p$.

Secondly, we use the expression $\Phi_{n,k}(f)=\frac{(n+s)!}{n!}\int_T P_{n,k}(u)f(u)\mathrm{d}u$ in the operator D_nf to replace the expression $f\left(\frac{k}{n}\right)$ in the Bernstein operator $B_nf(x)=\sum_{|k|\le n}P_{n,k}(x)f\left(\frac{k}{n}\right)$, then from Proposition 4.1 in Ref.[8] or Theorem 3.1 and Proposition 3.1 in [5],it follows that $D_nf(x)=\sum_{p=0}^n\sum_{l=1}^{N_p}d_l^{(p)}\left\langle x\cdot p_l^{(n)}\right\rangle^p$, where $d_l^{(p)}$ are uniquely determined by

$$\begin{pmatrix} (j_1')^{j_1'} & (j_2')^{j_2'} & \cdots & (j_{N_p}')^{j_1'} \\ \vdots & \vdots & \vdots & \vdots \\ (j_1')^{j_{N_P}'} & (j_2')^{j_{N_P}'} & \cdots & (j_{N_p}')^{j_{N_P}'} \end{pmatrix} \begin{pmatrix} d_1^{(p)} \\ \vdots \\ d_{N_p}^{(p)} \end{pmatrix} = \frac{(2(1+n))^p}{p!} \begin{pmatrix} i_1^{(p)}! c_1^{(p)}(f) \\ \vdots \\ i_{N_p}^{(p)}! c_{N_p}^{(p)}(f) \end{pmatrix} \tag{4.1}$$

with $c_l^{(p)}(f) = \dfrac{n!}{(n-p)!} \sum_{q \le i_l^{(p)}} \Phi_n, q(f) \dfrac{(-1)^{\left|i_l^{(p)}-q\right|}}{q!(i_l^{(p)}-q)!}$, $1 \le l \le N_p$.

Our main result can now be stated as follows.

Theorem 4.1. Let S be a compact subset of $R^s, s \ge 1$. Also, let $\varphi \in \in C(R), \varphi^{(|m|+1)} \in C(R), \varphi^{(i)}$ bounded on R, and there exists a $\theta \in R$ such that $\varphi^{(i)}(\theta) \ne 0$ for $1 \le i \le |m|+1, m \in Z_+^s$. Then for any $f \in S_p(S)$ and $D^{|m|} f \in Lip(\alpha), 0 < \alpha \le 1$, there is a neural network $N(x)$ with one hidden layer, $(n+1)(n-|m|+1)$ units in the hidden layer and activation function φ, such that $\left\|D^{|m|}N - D^{|m|}f\right\|_p \le Cn^{-\alpha/2}$, $|m| \le n$.

Now we prove Theorem 4.1. By the discussion in Section 2 of Ref.[8], it is sufficient to prove Theorem 4.1 in the case $S=T$. Let $L = \sum_{p=0}^{n} \sum_{l=1}^{N_p} \left|d_l^{(p)}\right|$. According to Theorem 2.1, for $H_r(x) = a_0 + a_1 x + \ldots + a_r x^r$, there is $N(x) = \sum_{i=1}^{r-m+1} c_i \varphi(\omega_i x + \theta), c_i, \omega_i \in R,$ such that for j, $0 \le j \le r$, and $\forall \varepsilon > 0$, $\left\|N^{(m)} - H_r^{(m)}\right\|_p < \varepsilon$. Set $N(x) = \sum_{p=0}^{n} \sum_{l=1}^{N_p} d_l^{(p)} \sum_{i=1}^{n-|m|+1} a_{p,i} \varphi(b_{p,i} \left\langle x \cdot p_l^{(n)} \right\rangle + \theta)$, then this is a neural network with one hidden layer, and its number of units in hidden layer is $(n+1)(n-|m|+1)$. Set $x \cdot p_l^{(n)} = u$, then $\left\langle x \cdot p_l^{(n)} \right\rangle^p$ can be written as a polynomial of n degree. By noting that $\left|\left\langle x \cdot p_l^{(n)} \right\rangle\right| \le 1$ for $x \in T$, $\left|p_l^{(n)}\right| \le 1/2$, and setting $\varepsilon = n^{-\alpha/2}, 0 < \alpha \le 1$, it follows from Theorem 3.1 and Eq.(4.1) that $\left\|D^{|m|}N - D^{|m|}f\right\|_p \le \left\|D^{|m|}N - D^{|m|}D_n f\right\|_p + \left\|D^{|m|}D_n - D^{|m|}f\right\|_p \le Cn^{-\alpha/2}$. This completes the proof of Theorem 4.1.

References:

[1] Cybenko G. Approximation by superpositions of sigmoidal function. Mathematics of Control Signals, and System, 1989, 2:303~314.

[2] Hornik K. Approximation capabilities of multilayer feedforward networks. Neural Networks, 1991,4:251~257.

[3] Anastassiou GA. Rate of convergence of some multivariate neural network operators to the unit. Computer and Mathematics with Applications, 2000,40:1~19.

[4] Leshno M, Lin VY, Pinkus A, Schocken S. Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. Neural Networks, 1993,6:861~867.

[5] Chui CK, Li X. Approximation by ridge functions and neural networks with one hidden layer. Journal of Approximation Theory, 1992,70:131~141.

[6] Chen TP, Chen H. Universal approximation to nonlinear operators by neural networks with arbitrary activation functions and its application to dynamical systems. IEEE Transactions on Neural Networks, 1995,6:911~917.

[7] Chen TP. Approximation problems in system identification with neural networks. Science in China (Series A), 1994,24(1):1~7 (in Chinese with English abstract).

[8] Li X. Simultaneous approximations of multivariate functions and their derivatives by neural networks with one hidden layer. Neurocomputing, 1996,12:327~343.

[9] Cao FL, Xu ZB. Neural network approximation for multivariate periodic functions: Estimates on approximation order. Chinese Journal of Computers, 2001,24(9):903~908 (in Chinese with English abstract).

[10] Hornik K, Stinchombe M, White H. Universal approximation of an unknown mapping and its derivatives using multilayer feedforward networks. Neural Networks, 1990,3:551~560.

[11] Gallant AR, White H. On learning the derivatives of an unknown mapping with multilayer feedforward networks. Neural Networks, 1992,5:129~138.

[12] Derriennic MM. On multivariate approximation by Bernstein-type polynomials. Journal of Approximation Theory, 1985,45: 155~166.

附中文参考文献:

[7] 陈天平.神经网络及其在系统识别中的逼近问题.中国科学(A辑),1994:24(1):1~7.

[9] 曹飞龙,徐宗本.多变元周期函数的神经网络逼近:逼近阶估计.计算机学报,2001,24(9):903~908.