近似指数型神经网络的本质逼近阶*

王建军 徐宗本**

(西安交通大学信息与系统科学研究所, 西安 710049)

摘要 针对一类比 Sigmoid 更为宽泛的指数型激活函数, 证明了三层前向神经网络的本质逼近阶. 特别地证明了对于定义在 R^d 中紧子集上的任意连续函数 f, 存在隐层单元数为 $m(n)=\min\limits_{B_d(f,n)<\varepsilon}(n+1)^d$ (其中 $B_d(f,n)=\frac{1}{2}\left(1+\frac{\pi^2}{2}\sqrt{d}\right)^2\omega_2\left(f,\frac{1}{n+2}\right)$, $\omega_2(f,\cdot)$ 为 f 的二阶连续模, n 为不小于 $1/\varepsilon$ 的任意正整数)的近似指数型神经网络 $R_n^\sigma(d)$ 使其逼近 f 的精度与速度满足

$$d_\infty(f,R_n^\sigma(d))\leqslant\frac{1}{2}\left(1+\frac{\pi^2}{2}\sqrt{d}\right)^2\omega_2\left(f,\frac{1}{n+2}\right).$$

同时, 当 f 属于 α-Lipschtz 函数类时, 网络达到其本质逼近阶 $n^{-\alpha}(0<\alpha\leqslant 2)$, 所获结果较完整地刻画了该类神经网络的逼近特征, 并揭示了该类神经网络逼近性态与网络拓扑之间的相依关系.

关键词 近似指数型神经网络 本质逼近阶 光滑模

近年来, 人工神经网络已广泛应用于科学与工程的各个领域. 之所以如此, 其主要原因是神经网络具有一定意义上的万有逼近性[1~9]. 所有这些研究的一个典型推论是: 定义在R^d上的任何一个连续函数可以通过一个三层前向神经网络任意逼近. 一个含d个输入, 1 个输出的三层前向神经网络数学上可表示为

$$\mathcal{N}(x)=\sum_{i=1}^{m}c_i\sigma\left(\sum_{j=1}^{d}w_{ij}x_j+\theta_i\right),\ x\in\mathcal{R}^d,\ d\geqslant 1, \tag{1}$$

其中$1\leqslant i\leqslant m$, $\theta_i\in\mathbb{R}$ 是阈值, $w_i=(w_{i1},w_{i2},\cdots,w_{id})^{\mathrm{T}}$ 是输入层与隐层第 i 个神经

收稿日期: 2005-10-24; 接受日期: 2006-02-05

* 国家自然科学基金(批准号: 10371097)和重点基金(批准号: 70531030)资助项目

** E-mail: zbxu@mail.xjtu.edu.cn

元的连接权，$c_i \in \mathbb{R}$ 是隐层与输出层之间的连接权，σ 是隐层节点的激活函数(传递函数). 通常情况下，激活函数 σ 取为 sigmoid 型，即满足 $\sigma(t) \to 1$ ($t \to +\infty$), $\sigma(t) \to 0$ ($t \to -\infty$). 用向量形式表示, (1)式可进一步表示为

$$\mathcal{N}(x) = \sum_{i=1}^{m} c_i \sigma(w_i \cdot x + \theta_i),\ x \in R^d.$$

多年来，人们一直致力于研究神经网络的可逼近性或稠密性. 文献[7,8,10~12]等先后证明：在 sigmoid 型激活函数等弱型条件下，对任何定义在 R^d 中紧子集 K 上的连续函数，存在一个形式如(1)式的三层前向神经网络，它以任意精度逼近该函数. 此后，许多学者建立了各种各样的有关(1)式逼近多元函数的稠密性和复杂性结果.

从应用角度看，人们更为关心神经网络逼近的定量性质，特别如希望知晓人工神经网络的拓扑结构(如网络隐层单元数)与网络的逼近能力及逼近速度之间有何种关系？文献[3,13~21]等对多种激活函数(包括 logistic, Heaviside 阶梯函数等)研究了这一问题. 他们构造性地给出了网络逼近的上界估计，即网络逼近连续函数的误差上界；同时他们还在固定误差容限下估计了神经网络所需的隐单元个数.

我们看到，已有的这些定量研究基本上都限于对网络逼近速度的上界估计. 然而，仅有上界估计对于刻画神经网络的逼近能力来说还是远远不够的，如它仍不能正确反映一个神经网络的本质逼近阶(即一个网络能够达到且仅能够达到的逼近程度[22]). 因此，为了准确地刻画所构造网络的逼近性能，应该不仅要对网络逼近速度的上界作出估计，而且更重要的是应对其下界也作出估计. 特别地，我们希望获得一个神经网络逼近速度的上下界具有相同阶的估计，从而精确刻画网络的本质逼近阶，以从根本上澄清网络的逼近性态与网络隐层单元数之间的关系. 文献[22]对一类满足sigmoid函数的神经网络给出了这样的本质逼近阶估计. 本文将推广上述工作到更广泛类型的神经网络. 我们注意到，在文献[29]中，Ritter引入了一类所谓的近似指数型函数的概念，一个实函数 $\sigma: R \to R$ 称为是近似指数型的，如果它在原点经过适当的仿射变换，可在负半轴上任意逼近指数型函数(确切定义见下一节). 这类函数包含通常的sigmoid型及指数型函数. 在文献[23]中，作者对近似指数型前向网络的逼近给出了一个上界估计，但没有给出下界. 考虑到这一类网络的广泛性，我们希望获得该网络逼近的下界估计，从而获得其本质逼近阶. 本文将构造性地证明该类神经网络对任意连续函数或可积函数逼近的上、下界及本质逼近阶估计. 所获结果不仅更为精确化了文献[23]的上界估计，给出了本质逼近阶估计，而且给出了网络逼近阶与网络隐层单元数之间的明确关系.

1　记号及主要结果

本文采用如下记号: $\mathbb{N},\mathbb{R},\mathbb{R}_+$ 分别表示非负整数, 实数和非负实数, R^d 表示 d 维实 Euclid 空间($d\geqslant 1$). 对任何

$$x=(x_1,x_2,\cdots,x_d)\in R^d,\quad y=(y_1,y_2,\cdots,y_d)\in R^d,$$

表示

$$e^x=(e^{x_1},e^{x_2},\cdots,e^{x_d}),\quad xy=(x_1y_1,x_2y_2,\cdots,x_dy_d).$$

若所有的 $x_i\geqslant 0$, $i=1,2,\cdots,d$, 进而表示

$$x^y=(x_1^{y_1},x_2^{y_2},\cdots,x_d^{y_d});$$

向量 x 与 y 的内积定义为

$$x\cdot y=\sum_{k=1}^{d}x_ky_k.$$

我们用 $P_n(d)$ 和 $T_n(d)$ 分别表示所有 d 元实的、次数不超过 n 的代数多项式和三角多项式, $P_n^E(d)$ 是形如 $\sum\limits_{\lambda\in l(0,1,\cdots,n)^d}a_\lambda e^{-\lambda\cdot x}$ 的所有 d 元实的、次数不超过 n 的指数型多项式; 对给定的激活函数 $\sigma:\mathbb{R}\to\mathbb{R}$, 所有形如 $\sum\limits_{\lambda\in l(0,1,\cdots,n)^d}a_\lambda\sigma(-\lambda\cdot x+b_\lambda)(l>0)$ 的函数全体记为 $R_n^\sigma(d)$. 用 $\|\cdot\|_\infty$ 表示 R^d 中的范数,即

$$\|f\|_\infty=\sup_{x\in R^d}|f(x)|;$$

对任何连续函数 f 和有界实函数集 S, 用 $d_\infty(f,S)$ 表示 f 与 S 的距离, 即

$$d_\infty(f,S)=\sup_{g\in S}\|f-g\|_\infty.$$

对任意定义在 R^d 上的光滑函数 f, 其$|m|$-阶偏导数表示为

$$D^{|m|}f(x)=\frac{\partial^{|m|}f}{\partial x^{|m|}}(x)=\frac{\partial^{|m|}f}{\partial x_1^{m_1}\partial x_2^{m_2}\cdots\partial x_d^{m_d}}(x).$$

连续模是刻画函数连续性和光滑性程度的一种度量, 在逼近论定量研究中有着基本的重要性, 它常被用于刻画某些重要函数类的数量特征, 假定 Q 是一个距离空间, d 是其上的距离, 我们用 $C(Q)$ 表示 Q 上连续函数的全体.

定义 1　设 $f\in C(Q)$, 函数 f 的连续模 $\omega(f,t)$ 是如下定义的由 $[0,\infty)\to[0,\infty)$ 的一个映射:

$$\omega(f,t)=\sup_{\|x_1-x_2\|\leqslant t}|f(x_1)-f(x_2)|.$$

设 e 是 d 维 Euclid 空间中的任一向量, 函数 f 在方向 e 上以 h 为步长的 r-阶对称

差分定义

$$\Delta_h^r f(x)=\sum_{i=0}^{r}(-1)^{r-i}\binom{r}{i}f\left(x+\left(\frac{r}{2}-i\right)he\right);$$

对任何函数 $f\in C(Q)$，f 的 r-阶连续模定义为

$$\omega_r(f,t)=\sup_{x\pm\frac{he}{2}\in Q,\|h\|\leqslant t}\left|\Delta_h^r f(x)\right|.$$

如果 $\omega_2(f,t)=O(t^\alpha)$, $\alpha\in(0,2]$, 则称 f 属于 α 阶 Lipschitz 函数类. 记作 $f\in Lip(\alpha)_2$.

定义 2[23] 一个函数 $\sigma:\mathbb{R}\to\mathbb{R}$ 称为是近似指数型的, 如果对任意 $\varepsilon>0$, 存在实数 β,γ,ρ,τ 使得 $\left|\gamma\sigma(\beta t+\tau)+\rho-e^t\right|<\varepsilon$ ($\forall t\leqslant 0$).

从几何上看, 定义 2 意味着: 一个近似指数型函数 σ 在 $\mathbb{R}$ 和原点经过适当的伸缩、旋转变换, 它在负半轴上可以任意逼近指数函数. 例如, 通过取 $\beta=1$, $\rho=0$, $\gamma=1$ 后, 可看出指数函数 e^t 是近似指数型函数. 而对于函数 $\sigma(t)=\dfrac{1}{1+e^{-t}}$, 取 $\beta=1$, $\rho=0$, $\gamma=\dfrac{1}{\sigma(\tau)}$, 我们观察到

$$\left|\frac{\sigma(t+\tau)}{\sigma(\tau)}-e^t\right|=\frac{e^{t+\tau}}{e^{t+\tau}+1}\left|1-e^t\right|\leqslant e^\tau\left|e^t-e^{2t}\right|. \tag{2}$$

而对任何 $t\leqslant 0$, (2)式收敛到 0 ($\tau\to-\infty$), 由此可见, 它也是一个近似指数型函数. 进一步, 容易验证, 大多数 sigmoid 型函数都是近似指数型函数.

在文献[23]中, Ritter 利用连续模得到了如下关于近似指数型神经网络逼近的上界估计:

命题 1 对任何连续函数 $f\in C[0,1]^d$ 和任意 $n\in\mathbb{N}$, 存在隐层单元数大于 $m\geqslant\min\limits_{C_d(f,n)<\varepsilon}(n+1)^d$ (其中 $C_d(f,n)=\left(\dfrac{1}{2}+\dfrac{\pi^2}{4}\sqrt{d}\right)\omega\left(f,\dfrac{1}{n+2}\right)$)的形如(1)式的神经网络 $R_n^\sigma(d)$, 它满足

$$d_\infty(f,R_n^\sigma(d))\leqslant\left(\frac{1}{2}+\frac{\pi^2}{4}\sqrt{d}\right)\omega\left(f,\frac{1}{n+2}\right).$$

本文将应用高阶连续模从多个层面推广与深化上述结果. 首先, 我们用二阶连续模取代上述命题中的一阶连续模, 给出该类网络逼近的一个新的更为严格的逼近上界. 然后我们建立网络逼近的一个下界估计. 所给出的网络逼近上、下界估计具有相同阶, 因而我们进一步导出对应神经网络的本质逼近阶. 我们所获

得的主要结果如下:

定理 1　对于定义在 R^d 中紧子集 V 上的任何连续函数 f 和容许误差 ε, 存在隐层单元数大于或等于 $\min\limits_{B_d(f,n)<\varepsilon}(n+1)^d$ (其中 $B_d(f,n)=\frac{1}{2}\left(1+\frac{\pi^2}{2}\sqrt{d}\right)^2\omega_2\left(f,\frac{1}{n+2}\right)$, n 为不小于 $1/\varepsilon$ 的任意正整数)的三层近似指数型神经网络 $R_n^\sigma(d)$, 使满足

1) 逼近上界估计:

$$d_\infty(f,R_n^\sigma(d))\leqslant\frac{1}{2}\left(1+\frac{\pi^2}{2}\sqrt{d}\right)^2\omega_2\left(f,\frac{1}{n+2}\right),\tag{3}$$

2) 逼近下界估计:

$$\omega_2\left(f,\frac{1}{n+2}\right)\leqslant\frac{C}{n^2}\left\{\sum_{k=1}^{n}k\cdot d_\infty(f,R_k^\sigma(d))+\|f\|_\infty\right\},\tag{4}$$

3) 本质逼近阶估计; 当且仅当于 $f\in Lip(\alpha)_2$ 时, 有

$$d_\infty(f,R_n^\sigma(d))=O(n^{-\alpha}).\tag{5}$$

其中 C, C_1, C_2, C_3 均是与 n, f, x 无关的正常数, 但其值可能在上下文中不同.

上述定理中的(3)式是对网络逼近精度和速度的上界估计, 反映了网络逼近速度上界与隐层神经元个数之间的关系. 它说明网络逼近速度的上界受控于被逼近函数的二阶连续模及隐层神经元个数 $\min\limits_{B_d(f,n)<\varepsilon}(n+1)^d$. 显然当 $n\to\infty$ 时, $d_\infty(f,R_n^\sigma(d))\to 0$. 它说明对任意连续函数 f, 存在近似指数型三层前向神经网络对其任意逼近.

定理 1 中的(4)式是网络逼近精度与速度的平均下界估计, 它表明该类神经网络的逼近速度与精度随隐层元个数变化之平均值 $\frac{2}{n(n+1)}\sum\limits_{k=1}^{n}k\cdot d_\infty(f,R_k^\sigma(d))$ 不低于所被逼近函数的二阶连续模与一个无穷小量 $o\left(\frac{1}{n^2}\right)$ 之差(注意到, (4)式可重写为 $C\left(\frac{1}{2}+\frac{1}{n}\right)\left\{\frac{2}{n(n+1)}\sum\limits_{k=1}^{n}k\cdot d_\infty(f,R_k^\sigma(d))\right\}\geqslant\omega_2\left(f,\frac{1}{n+2}\right)-\frac{1}{n^2}\|f\|_\infty$). 由于

$$d_\infty(f,R_n^\sigma(d))\to 0,\quad n\to\infty,$$

且当 n 充分大(即, 隐层单元充分多)时, 有

$$C\omega_2\left(f,\frac{1}{n+2}\right)-C\frac{1}{n^2}\|f\|_\infty\leqslant d_\infty(f,R_n^\sigma(d))\leqslant\frac{1}{2}\left(1+\frac{\pi^2}{2}\sqrt{d}\right)^2\omega_2\left(f,\frac{1}{n+2}\right),$$

这说明该类神经网络能达到的最高逼近精度为 $\omega_2\left(f,\frac{1}{n+2}\right)$.

定理 1 中的(5)式刻画了神经网络的本质逼近阶. 它说明: 当被逼近函数 f 属于 α-Lipschitz 类时, 该类神经网络的本质逼近阶为 $n^{-\alpha}$. 该估计与(3), (4)式一起说明被逼近函数的光滑性越高, 网络逼近速度越快, 逼近精度也越高; 反之亦然. 我们将在下节证明一些预备结果, 而在第 3 节给出定理 1 的完整证明.

2 代数、三角及指数多项式对连续函数的一致逼近

逼近论中的Jackson's定理[9]是利用连续模刻画定义在区间上的连续函数被多项式逼近速度的一种估计. 文献[24~26]等将其推广到多元情况. 特别地, 文献[27]通过引进新的连续模(即现在为人们熟知的 Ditzian-Totik 模)研究了相关问题. 文献[24,26]曾得到如下重要的命题, 它构成了文献[23]的基础.

命题 2 设对任何连续函数 $f:[0,1]^d\to\mathbb{R}$, $n\in\mathbb{N}$, 有

$$d_\infty(f,P_n(d))\leqslant\left(\frac{1}{2}+\frac{\pi^2}{4}\sqrt{d}\right)\omega\left(f,\frac{1}{n+2}\right). \tag{6}$$

我们将通过应用更为精确的二阶连续模, 深化上述定理而获得一个更为精确的结果:

定理 2 对任何连续函数 $f:[a,b]^d\to\mathbb{R}$, $n\in\mathbb{N}$, 有

$$d_\infty(f,P_n(d))\leqslant\frac{1}{2}\left(1+\frac{\pi^2}{2}\sqrt{d}\right)^2\omega_2\left(f,\frac{1}{n+2}\right). \tag{7}$$

我们将在本文附录中给出该定理的证明.

定理 3 对任何连续函数 $f:[a,b]^d\to\mathbb{R}$, $n\in\mathbb{N}$, 有

$$d_\infty(f,P_n^E(d))\leqslant\frac{1}{2}\left(1+\frac{\pi^2}{2}\sqrt{d}\right)^2\omega_2\left(f,\frac{1}{n+2}\right). \tag{8}$$

证 我们假设 $[a,b]^d=[0,1]^d$ (一般情况利用简单的线性变换即可证明), 由命题 2, 对任何 $f:[a,b]^d\to\mathbb{R}$, 存在代数多项式

$$p(x)=\sum_{\lambda\in(0,1,2,\cdots,n)^d}a_\lambda\prod_{k=1}^d x_k^{\lambda_k},$$

使得

$$\|f-p\|_\infty\leqslant\frac{1}{2}\left(1+\frac{\pi^2}{2}\sqrt{d}\right)^2\omega_2\left(f,\frac{1}{n+2}\right)+\varepsilon. \tag{9}$$

对 $\xi>0$, 我们定义函数

$$\rho_\xi(x)=\left(\frac{1-e^{-\xi_1 x_1}}{1-e^{-\xi_1}},\frac{1-e^{-\xi_2 x_2}}{1-e^{-\xi_2}},\cdots,\frac{1-e^{-\xi_d x_d}}{1-e^{-\xi_d}}\right),\ x\in[0,1]^d.$$

则显然当 $\xi\to 0$ 时, $\rho_\xi(x)$ 收敛于恒等函数 $\rho_0(x)\equiv x=(x_1,x_2,\cdots,x_d)$. 因此, 通过对 ξ 做合适的选择, 可以使得

$$\left\|p(\rho_\xi)-p(\rho_0)\right\|_\infty\leqslant\varepsilon. \tag{10}$$

注意到 $p(\rho_\xi)\in P_n^E(d)$, 由(9)和(10)式, 我们获得估计

$$\begin{aligned}\left\|p(\rho_\xi)-f\right\|_\infty&=\left\|p(\rho_\xi)-f(\rho_0)\right\|_\infty\\&\leqslant\left\|p(\rho_\xi)-p(\rho_0)\right\|_\infty+\left\|p(\rho_0)-f(\rho_0)\right\|_\infty\\&\leqslant\frac{1}{2}\left(1+\frac{\pi^2}{2}\sqrt{d}\right)^2\omega_2\left(f,\frac{1}{n+2}\right)+2\varepsilon.\end{aligned}$$

由 ε 的任意性, 即知定理 3 成立.

我们将在下节应用定理 2 与 3 来证明我们的主要结论.

3　定理 1 的证明

我们先证明(3)式. 让 τ 表示 $[0,1]^d\to V\subseteq R^d$. f 是定义在紧集 V 上的多元函数, 于是 $f(\tau)$ 是 f 在 V 上的连续延伸, 且它们具有相同的连续模. 由定理 3, 存在一个指数多项式

$$p(x)=\sum_{\lambda\in(0,1,2,\cdots,n)^d}a_\lambda e^{\lambda\cdot x},$$

使得

$$\left\|f-p|_C\right\|_\infty\leqslant\left\|f(\tau)-p|_{[0,1]^d}\right\|_\infty\leqslant\frac{1}{2}\left(1+\frac{\pi^2}{2}\sqrt{d}\right)^2\omega_2\left(f,\frac{1}{n+2}\right)+\varepsilon.$$

由指数型激活函数 σ 的定义, 指数函数 e^x 可以被 $\gamma\sigma(\beta x+\tau)+\rho$ 在负半平面所逼近, 且误差不超过 $\varepsilon/\sum|a_\lambda|$. 于是和式

$$\sum_{\lambda\in(0,1,2,\cdots,n)^d\setminus\{0\}}a_\lambda\gamma\sigma(\beta x+\tau)+\left(\gamma a_0\sigma(\tau)+\rho\sum_{\lambda\in(0,1,2,\cdots,n)^d\setminus\{0\}}a_\lambda\right)$$

属于 $R_n^\sigma(d)$. 因此它与 f 的距离不超过 $\frac{1}{2}\left(1+\frac{\pi^2}{2}\sqrt{d}\right)^2\omega_2\left(f,\frac{1}{n+2}\right)+2\varepsilon$. 由 ε 的任意性, (3)式得证.

下面我们证明(4)式. 先证明一个引理:

引理 1 对任意 $n\in\mathbb{N}$，若序列 $\{a_n\},\{b_n\}$ 满足 $(p>0)$

$$a_n \leqslant \left(\frac{k}{n}\right)^p a_k + b_k \quad (1\leqslant k\leqslant n), \tag{11}$$

则有

$$a_n \leqslant C_p n^{-p}\left\{\sum_{k=1}^{n} k^{p-1} b_k + a_1\right\}.$$

证 对于 $n\geqslant 2$，选择 N 使得 $2^N\leqslant n<2^{N+1}$，则存在 $m_k\in\mathbb{N}$, $1\leqslant k\leqslant N$ 且 $\frac{n}{2^{k+1}}\leqslant m_k\leqslant\frac{n}{2^k}$，满足 $a_{m_k}\leqslant a_j\left(\frac{n}{2^{k+1}}\leqslant j\leqslant\frac{n}{2^k}\right)$，置 $m_{N+1}=1$，则由(11)式可得

$$\begin{aligned}
a_n &\leqslant \left(\frac{m_0}{n}\right)^p a_{m_0} + b_{m_0} \\
&= n^{-p}\sum_{k=0}^{N} m_k^p\left(a_{m_k} - \left(\frac{m_{k+1}}{m_k}\right)^p a_{m_{k+1}}\right) + b_{m_0} + n^{-p}a_1 \\
&\leqslant \sum_{k=0}^{N} 2^{-pk} b_{m_{k+1}} + b_{m_0} + n^{-p}a_1 \\
&= \sum_{k=1}^{N+1} 2^{-p(k-1)} b_{m_k} + b_{m_0} + n^{-p}a_1 \\
&\leqslant 2^p\sum_{k=0}^{N+1} 2^{-pk} b_{m_k} + n^{-p}a_1 \\
&\leqslant 2^p\sum_{k=0}^{N+1} 2^{-pk}\frac{2^{k+1}}{n}\sum_{\frac{n}{2^{k+1}}\leqslant j\leqslant\frac{n}{2^k}} b_j + n^{-p}a_1 \\
&\leqslant 2^{p+1}\sum_{k=0}^{N+1} n^{-p}\sum_{\frac{n}{2^{k+1}}\leqslant j\leqslant\frac{n}{2^k}} j^{p-1} b_j + n^{-p}a_1 \\
&\leqslant C_p n^{-p}\left\{\sum_{k=1}^{n} k^{p-1} b_k + a_1\right\}.
\end{aligned}$$

引理 1 得证.

我们现在证明(4)式. 对任意 $1\leqslant q\leqslant\infty$ 和卷积算子

$$L(v_n, f) = (2\pi)^{-d}\int_{[-\pi,\pi]^d} f(x-t)v_n(t)\mathrm{d}t, \quad x\in R^d.$$

显然,

$$D^\nu L(v_n, f) = (2\pi)^{-d}\int_{[-\pi,\pi]^d} D^\nu f(x-t)v_n(t)\mathrm{d}t, \quad x\in R^d.$$

由于 $(2\pi)^{-d}\int_{[-\pi,\pi]^d} v_n(t)\mathrm{d}t = 1$ 且 $\|L(v_n,f)\|_q \leqslant \|f\|_q \|v_n\|_1$，故有

$$\left\|D^{|\nu|}L(v_n,f)\right\|_q \leqslant \left\|D^{\nu}f\right\|_q, \tag{12}$$

进一步由 Bernstein 不等式，我们可导出

$$\left\|D^{|\nu|}L(v_n,f)\right\|_q \leqslant Cn^{|\nu|}\left\|D^{\nu}f\right\|_q. \tag{13}$$

令 $a_n = \frac{1}{n^2}\left\|D^{|\nu|}L(v_n,f)\right\|_q$，$|\nu| = 2$，$b_n = \|L(v_n,f)-f\|_q$. 于是由(12), (13)式，我们得出

$$\begin{aligned} a_n &\leqslant \frac{1}{n^2}\left\|D^{|\nu|}L(v_n,L(v_k,f))\right\|_q + \frac{1}{n^2}\left\|D^{|\nu|}L(v_n,f-L(v_k,f))\right\|_q \\ &\leqslant \frac{1}{n^2}\left\|D^{|\nu|}L(v_k,f)\right\|_q + C\|f-L(v_k,f)\|_q \\ &= \left(\frac{k}{n}\right)^2 a_k + Cb_k, \end{aligned}$$

根据引理 1

$$a_n \leqslant C_2 n^{-2}\left\{\sum_{k=1}^{n} kb_k + a_1\right\},$$

有

$$\sup_{|\nu|=2}\left\|D^{|\nu|}L(v_n,f)\right\|_q \leqslant C_2\left\{\sum_{k=1}^{n} k\|L(v_k,f)-f\|_q + \|f\|_q\right\}. \tag{14}$$

再者，对 $n \geqslant 2$，存在 $m \in N$，使 $\frac{n}{2} \leqslant m \leqslant n$ 且

$$\|L(v_m,f)-f\|_q \leqslant \|L(v_k,f)-f\|_q,\quad \frac{n}{2} \leqslant k \leqslant n. \tag{15}$$

由(12)~(15)式，我们构造如下 K-泛函:

$$K_2(f,t^2) = \inf_{g\in A.C.\mathrm{loc}}\left\{\|f-g\|_q + t^2\sup_{|m|=2}\left\|D^{|m|}g\right\|_q\right\}.$$

由文献[28]，函数 f 的连续模与 K-泛函有如下等价关系(其中 C'是正常数):

$$C'^{-1}K_2(f,t^2) \leqslant \omega_2(f,t) \leqslant C'K_2(f,t^2). \tag{16}$$

于是

$$K_2\left(f,\frac{1}{(n+2)^2}\right)$$

$$
\begin{aligned}
&\leqslant \left\|L(v_m,f)-f\right\|_q+\frac{1}{(n+2)^2}\sup_{|v|=2}\left\|D^{|v|}L(v_n,f)\right\|_q\\
&\leqslant \frac{4}{n^2}\sum_{\frac{n}{2}\leqslant k\leqslant n}k\left\|L(v_k,f)-f\right\|_q+\frac{C_2}{(n+2)^2}\left\{\sum_{k=1}^{n}k\left\|L(v_k,f)-f\right\|_q+\left\|f\right\|_q\right\}\\
&\leqslant \frac{C_3}{n^2}\left\{\sum_{k=1}^{n}k\left\|L(v_k,f)-f\right\|_q+\left\|f\right\|_q\right\}\\
&\leqslant \frac{C_3}{n^2}\left\{\sum_{k=1}^{n}kd_\infty(f,P_k(d))+\left\|f\right\|_\infty\right\}.
\end{aligned}
$$

由(16)式及类似于定理 3 和(3)式的证明, 我们得到

$$
\begin{aligned}
&\omega_2\left(f,\frac{1}{n+2}\right)\\
&\leqslant \frac{C}{n^2}\left\{\sum_{k=1}^{n}k(d_\infty(f,R_k^\sigma(d))+\varepsilon)+\left\|f\right\|_\infty\right\}\\
&\leqslant \frac{C}{n^2}\left\{\sum_{k=1}^{n}kd_\infty(f,R_k^\sigma(d))+\left\|f\right\|_\infty\right\}+C\varepsilon.
\end{aligned}
$$

由 ε 的任意性, (4)式得证.

定理 1 的 3)可由 1)和 2)直接推出.

为了获得神经网络隐层节点数的估计. 我们令 $n\geqslant\left\lceil\frac{1}{\varepsilon}\right\rceil$(即大于 $\frac{1}{\varepsilon}$ 的最小正整数), 则由定理 1 中的(3)式得

$$
m=\min_{B_d(f,n)<\varepsilon}(n+1)^d,
$$

$$
B_d(f,n)=\frac{1}{2}\left(1+\frac{\pi^2}{2}\sqrt{d}\right)^2\omega_2\left(f,\frac{1}{n+2}\right),
$$

于是定理 1 得证.

4 评注

有关神经网络的定量研究, 近年来受到广泛关注, 但已有研究基本上仅限于网络逼近的上界估计, 从而不能完全刻画所对应网络的本质逼近阶及揭示网络逼近精度与网络拓扑之间的关系. 本文从理论上证明了, 对任意给定的精度 ε 和任何定义在紧集上的多元连续函数 f, 存在隐层单元个数为 $\min_{B_d(f,n)<\varepsilon}(n+1)^d$ (其中 n 为不小于 $1/\varepsilon$ 的任意正整数, $B_d(f,n)=\frac{1}{2}\left(1+\frac{\pi^2}{2}\sqrt{d}\right)^2\omega_2\left(f,\frac{1}{n+2}\right)$)的三层近似指数型

前向神经网络, 使它能在给定精度下逼近该函数. 我们不仅给出了该类神经网络逼近速度与精度的上界和下界估计, 而且揭示了该类神经网络的本质逼近阶. 我们也具体地阐明了网络逼近阶与网络隐层单元个数之间的关系. 所获结果对于澄清近似指数型神经网络的本质逼近性态与拓扑构造有重要意义与参考价值.

参　考　文　献

1 Attali J G, Pages G. Approximation of functions by a multilayer perceptron: A new approach. Neural Netw, 1997, 10(6): 1069－1081[DOI]

2 Cardaliaguet P, Euvrard G. Approximation of a function and its derivatives with a neural networks. Neural Netw, 1992, 5(2): 207－220[DOI]

3 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

4 Chui C K, Li X. Approximation by ridge functions and neural networks with one hidden layer. J Approx Theory, 1992, 70(1): 131－141[DOI]

5 Chui C K, Li X. Neural networks with one hidden layer. In: Jetter K, Utreras F I, eds Multivariate Approximation: From CAGD to Wavelets. Singapore:World Scientific Press, 1993. 77－89

6 Chen T P, Chen H. Universal approximation to nonlinear operators by neural networks with arbitrary activation functions and its application to a dynamic system. IEEE Trans Neural Netw, 1995, 6(4): 911－917[DOI]

7 陈天平. 神经网络及其在系统识别应用中的逼近问题. 中国科学, A 辑, 1994, 24(1): 1－7

8 Hornik K, Stinchombe M, White H. Universal approximation of an unknown mapping and its derivatives using multilayer feedforward networks. Neural Netw, 1990, 3(4): 551－560[DOI]

9 Jackson D. On the approximation by trigonometric sums and polynomials. Trans Amer Math Soc, 1912, 13: 491－515

10 Cybenko G. Approximation by superpositions of sigmoidal function. Math Control Signal Syst, 1989, 2(4): 303－314

11 Funahashi K I. On the approximate realization of continuous mappings by neural networks. Neural Netw, 1989, 2(2): 183－192[DOI]

12 Yoshifusa I. Approximation of functions on a compact set by finite sums of sigmoid function without scaling. Neural Netw, 1991, 4(5): 817－826[DOI]

13 Kůrkova V, Kainen P C, Kreinovich V. Estimates for the number of hidden units and variation with respect to half-space. Neural Netw, 1997, 10(6): 1068－1078

14 Mhaskar H N, Micchelli C A. Approximation by superposition of a sigmoidal function. Adv Appl Math, 1992, 13(3): 350－373[DOI]

15 Mhaskar H N. Neural networks for optimal approximation for smooth and analytic functions. Neural comput, 1996, 8(2): 164－177

16 Mhaskar H N, Micchelli C A. Degree of approximation by neural networks with a single hidden layer. Adv Appl Math, 1995, 16(2): 151－183[DOI]

17 Mhaskar H N, Khachikyan L. Neural networks for functions approximation. In: Makhoul J, Manolakos E, Wilson E, eds. Neural Networks for Signal Processing, Proc. 1995 IEEE Workshop, V, Girosi F (Cambridge, MA). New York: IEEE Press, 1995. 21－29

18 Mhaskar H N, Micchelli C A. Dimension-independent bounds on the degree of approximation by neural networks. IBM J Res Develop, 1994, 38(3): 277－284

19 Maiorov V, Meir R S. Approximation bounds for smooth functions in $C(R^d)$ by neural and mixture networks. IEEE Trans Neural Netw, 1998, 9(5): 969－978[DOI]

20 Suzuki S. Constructive function approximation by three-layer artificial neural networks. Neural Netw, 1998,

11(6): 1049－1058[DOI]
21 Ritter G. Jackson's theorems and the number of hidden units in neural networks for uniform approximation. Univ Passau, Fak Math Inform, Technical Report, MIP9415, 1994
22 Xu Z B, Cao F L. The essential order of approximation for neural networks. Sci China, Ser F Inf Sci, 2004, 47(1): 97－112
23 Ritter G. Efficient estimation of neural weights by polynomial approximation. IEEE Trans Inf Theory, 1999, 45(5): 1541－1550[DOI]
24 Feinerman R P, Newman D J. Polynomial Approximation. BAltimore, MD: Williams & Wilkins, 1974
25 Nikol'skii S M. Approximation of Functions of Several Variables and Imbedding Theorems. Berlin: Springer, 1975
26 Soardi P M. Serie di fouroer in piú variabili. Quad dell'Unione Mat Italiana, 26. Bologna, Italy: Pitagora Editrice, 1984
27 Ditzian Z, Totik V. Moduli of Smoothness. NewYork: Springer-Verlag, 1987
28 Johnen H, Scherer K. On the equivalence of the K-functional and modulus of continuity and some applications. In: Schempp W, Zeller K, eds. Constructive Theory of Functions of Several Variable. Berlin: Springer-Verlag, 1977. 119－140
29 Hornik K, Stinchombe M, White H. Multilayer feedforward networks are universal approximation. Neural Netw, 1989, 2(3): 359－366[DOI]
30 Leshno M, Lin V Y, Pinks A. et al. Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. Neural Netw, 1993, 6(7): 861－867[DOI]

附录 A

在本附录中我们给出定理 2 的证明，我们先做一些预备工作.

对 $n \geqslant 1$, $n \in \mathbb{N}$，令

$$u_n(x) = \frac{\left(1-\cos^2\frac{\pi}{n+2}\right)\left(t_{n+2}(\cos x) - t_{n+2}(x)\right)}{(n+2)\left(\cos x - \cos\frac{\pi}{n+2}\right)^2}$$

是 Korovkin 核，它是一元次数不超过 n 的三角多项式，其中 $t_n(x)$ 是 Chebyshev 多项式，即 $t_n(x) = \arccos(nx)$, $u_n(x) \in T_n(1)$, $u_n \geqslant 0$ 且 $\frac{1}{2\pi}\int_{-\pi}^{\pi} u_n(x)\mathrm{d}x = 1$. 根据文献[21]，$u_n(x)$ 也可表示为

$$u_n(x) = \frac{2}{n+2}\left\{\sum_{-n/2}^{n/2}\cos\frac{k\pi}{n+2}\cos kx\right\}^2. \tag{A1}$$

多元情形的 Korovkin 核可通过 d 个一元的乘积来定义，即

$$v_n(x_1, x_2, \cdots, x_d) = \overbrace{u_n(x)\times u_n(x)\times\cdots u_n(x)}^{d} \in T_n(d)$$

显然，$v_n \geqslant 0$,

$$(2\pi)^{-d}\int_{[-\pi,\pi]^d} v_n(x)\mathrm{d}x = 1.$$

引理 A1　设函数 $f:R^d\to\mathbb{R}$ 是一个以 2π 为周期的函数. 则对任意 $n\in\mathbb{N}$, 我们有

$$d_\infty\left(f,T_n(d)\right)\leqslant\frac{1}{2}\left(\frac{\sqrt{d}\pi^2}{2}+1\right)^2\omega_2\left(f,\frac{1}{n+2}\right). \tag{A2}$$

证　对任意两个以 2π为周期的函数 f, $g:R^d\to\mathbb{R}$, 它们的卷积为

$$L(f,g)=(2\pi)^{-d}\int_{[-\pi,\pi]^d}f(x-t)g(t)\mathrm{d}t.\ x\in R^d.$$

由于 $u_n(x)$ 是偶函数, $(2\pi)^{-d}\int_{[-\pi,\pi]^d}v_n(x)\mathrm{d}x=1$. 从而应用 Cauchy-Shwartz 不等式, 我们获得

$$\begin{aligned}
&\left|L(v_n,f)(x)-f(x)\right|\\
&=\left|(2\pi)^{-d}\int_{[-\pi,\pi]^d}f(x-t)v_n(t)\mathrm{d}t-(2\pi)^{-d}\int_{[-\pi,\pi]}f(x)v_n(t)\mathrm{d}t\right|\\
&=\left|(2\pi)^{-d}\int_{[-\pi,\pi]^d}\left(f(x-t)-f(x)\right)v_n(t)\mathrm{d}t\right|\\
&=\left|(2\pi)^{-d}\int_{[-\pi,\pi]^d}\frac{f(x-t)-2f(x)+f(x+t)}{2}v_n(t)\mathrm{d}t\right|\\
&\leqslant\frac{1}{2(2\pi)^d}\int_{[-\pi,\pi]^d}\left|\frac{f(x-t)-2f(x)+f(x+t)}{2}\right|v_n(t)\mathrm{d}t\\
&\leqslant\frac{1}{2(2\pi)^d}\int_{[-\pi,\pi]^d}\omega_2(f,\|t\|)v_n(t)\mathrm{d}t\\
&\leqslant\frac{1}{2(2\pi)^d}\omega_2\left(f,\frac{1}{n+2}\right)\int_{[-\pi,\pi]^d}(1+(n+2)\|t\|)^2v_n(t)\mathrm{d}t\\
&\leqslant\omega_2\left(f,\frac{1}{n+2}\right)\frac{1}{2(2\pi)^d}\left\{(n+2)^2\int_{[-\pi,\pi]^d}\|t\|^2v_n(t)\mathrm{d}t\right.\\
&\quad\left.+2(n+2)\left(\sum_{k=1}^d(2\pi)^{-d}\int_{[-\pi,\pi]^d}t_k^2u_n(t)\mathrm{d}t\right)^{\frac{1}{2}}+1\right\}\\
&=\omega_2\left(f,\frac{1}{n+2}\right)\frac{1}{2}\left\{(n+2)^2\frac{d}{2\pi}\int_{[-\pi,\pi]}t^2u_n(t)\mathrm{d}t\right.\\
&\quad\left.+2(n+2)\left(\frac{d}{2\pi}\int_{[-\pi,\pi]}t^2u_n(t)\mathrm{d}t\right)^{\frac{1}{2}}+1\right\}.
\end{aligned}$$

将 $u_n(x)$ 写成三角形式　$u_n(x)=1+2\sum_{k=1}^n c_k\cos kx$. 则由(A1)式我们可得 $c_1=\cos\frac{\pi}{n+2}$. 因此由不等式　$\sin x\geqslant\frac{2}{\pi}x,\ x\in\left[-\frac{\pi}{2},\frac{\pi}{2}\right]$, 可推出

$$
\begin{aligned}
&\frac{1}{2\pi}\int_{[-\pi,\pi]} t^2 u_n(t)\mathrm{d}t \\
&\leqslant \frac{\pi}{2}\int_{[-\pi,\pi]} \sin^2\frac{t}{2} u_n(t)\mathrm{d}t \\
&= \frac{(\pi)^2}{2}(1-c_1) \\
&= \frac{(\pi)^2}{2} 2\sin^2\frac{\pi}{2(n+2)} \\
&\leqslant \frac{(\pi)^4}{4(n+2)^2}.
\end{aligned}
$$

于是

$$
\left|L(v_n,f)(x)-f(x)\right| \leqslant \frac{1}{2}\left(\frac{\sqrt{d}\pi^2}{2}+1\right)^2 \omega_2\left(f,\frac{1}{n+2}\right).
$$

由于 $L(v_n,f)(x)\in T_n(d)$，引理得证.

现在我们来证明定理 2. 不妨设 $a=-1$, $b=1$ (一般情况可类似证明). 记 $\cos x=(\cos x_1, \cos x_2,\cdots,\cos x_d)$. 设 $f\in C[-1,1]^d$, 则 $h(x)=f(\cos x)$ 是周期为 2π 的连续函数，由 Weierstrass 定理，存在三角多项式 $T(x)\in P_n(d)$ 使得

$$
\|h-T\|_\infty<\varepsilon, \quad \forall\varepsilon>0,
$$

我们可写 $T(x)$为

$$
\sum_{\lambda\in(0,1,2,\cdots,n)^d} a_\lambda \prod_{i=1}^{d}\cos\lambda_i x_i,
$$

由 Chebychev 多项式 $t_k(x)$，我们构造如下代数多项式:

$$
p(y)=\sum_{\lambda\in(0,1,2,\cdots,n)^d} a_\lambda \prod_{i=1}^{d} t_{\lambda_i}(y_i), \quad y=(y_1,\cdots,y_d)\in[-1,1]^d.
$$

则显然有 $\|p-f\|_\infty=\|T-h\|_\infty\leqslant\varepsilon$，因此

$$
d_\infty(f,P_n(d))=d_\infty(h,T_n(d)).
$$

由于 $|\cos x-\cos y|\leqslant|x-y|$，故 $\omega_2(h,\delta)\leqslant\omega_2(f,\delta)$. 于是定理 2 得证.