中国科学: 信息科学 2024 年 第 54 卷 第 3 期: 638–652
SCIENTIA SINICA Informationis
论文

深度 ReLU 神经网络的万有一致性

刘霞[1], 王迪[2*]

1. 西安理工大学理学院, 西安 710048
2. 西安交通大学管理学院, 西安 710049
* 通信作者. E-mail: wang.di@xjtu.edu.cn

收稿日期: 2022–10–15; 修回日期: 2023–02–03; 接受日期: 2023–05–19; 网络出版日期: 2024–03–11

国家自然科学基金 (批准号: 12371514, 12271431, 12171388) 和陕西数理基础科学研究项目 (批准号: 22JSQ023) 资助

摘要 随着数据量爆炸式增长、计算资源愈加丰富, 浅层神经网络并不总能满足时代需求, 从而导致深度神经网络的出现. 深度神经网络的迅猛发展主要体现在应用领域, 其理论研究相对匮乏. 基于此, 本文聚焦研究深度 ReLU 神经网络的万有一致性, 具体内容包括: 首先, 是否存在一个具有统一结构的深度神经网络 (即深度、宽度、激活函数等均已确定) 使得该深度神经网络可以学习更多特征, 并具有万有逼近性; 其次, 针对已确定的深度神经网络模型, 证明其是强万有一致的; 最后, 从实验的角度验证理论结果的合理性.

关键词 深度神经网络, 万有一致性, 深度学习, ReLU 函数, 逼近性

1 引言

深度学习 [1~7] (deep learning, DL) 是机器学习中一种对数据进行特征学习的方法, 在人工智能的飞速发展中功不可没. 深度学习已广泛应用于图像处理、计算机视觉、语言识别、自然语言处理和机器人等研究领域. 深度学习的概念起源于对神经网络的研究, 含有多个隐层的神经网络 (深度神经网络) 就是一种深度学习结构, 也是一种典型的深度学习模型. 随着大数据时代的到来和信息技术的发展, 以深度学习为代表的机器学习理论及应用迎来了巨大的机遇与挑战. 深度学习在应用中的巨大成功表明深度学习所采用的深度模型 (深度神经网络) 具有强大的非线性特征提取能力, 然而其理论研究并未赶上应用的步伐. 解释深度神经网络的深度、宽度、连接方式及学习和逼近效果之间的联系是当下亟待解决的理论核心问题.

众多文献表明, 深度神经网络可以突破浅层网络的许多瓶颈. 具体地, Mhaskar [8] 证明了浅层神经网络能以最优的学习速率逼近光滑函数, 不足的是该网络的权重和偏差都很大, 从而导致神经网络空间的容量 [9] (用覆盖数 [10] 来衡量) 过大. 文献 [11] 表明深度神经网络可以克服浅层神经网络在函

引用格式: 刘霞, 王迪. 深度 ReLU 神经网络的万有一致性. 中国科学: 信息科学, 2024, 54: 638–652, doi: 10.1360/SSI-2022-0401
Liu X, Wang D. Universal consistency of deep ReLU neural networks (in Chinese). Sci Sin Inform, 2024, 54: 638–652, doi: 10.1360/SSI-2022-0401

数逼近中下界估计的不足. Lin 在文献 [12] 中指出: 浅层神经网络在逼近再生核 Hilbert 空间中的函数时, 均不能以很高的概率达到最优的下界估计. Chui 等 [13] 证明了具有 Heaviside 函数的浅层神经网络不具有局部逼近能力, 而深度神经网络却可以逼近更多的函数 [12,14~17]. 随后, 文献 [18] 进一步阐明: 浅层神经网络不具有旋转不变性, 但具有两个隐层的深度神经网络不仅可以逼近光滑函数, 还能够克服饱和性. 这里, 饱和性指的是即使函数的光滑度达到一定程度, 但逼近率却无法提高. Lin 在文献 [19] 提出: 与浅层神经网络相比, 具有两个隐层的深度神经网络可以达到最佳的学习速率且学习的函数更多. 此外, 文献 [20~22] 表明: 与浅层神经网络相比, 带有 ReLU 激活函数的深度神经网络在逼近光滑函数方面更有效, 并且针对许多学习任务其泛化性能更好.

虽然已有理论结果论证了深度神经网络的部分优势, 但目前仍普遍存在这样一个问题: 针对不同的先验信息 (学习任务), 若要达到理想的理论结果, 对应的深度神经网络模型却各不相同. 那么, 是否存在一个具有统一结构的深度神经网络可以学习不同的特征? 该深度神经网络是否具有万有逼近性和万有一致性? 这是本文聚焦的主要研究问题. 针对深度神经网络的万有一致性问题, Lin 在文献 [23] 中严格证明了深度卷积神经网络 (deep convolutional neural networks, DCNN) 的强万有一致性, 并从实验角度验证了 DCNN 的性能不比具有混合结构的深度神经网络的差, 这种混合结构包含收缩 (无零填充) 卷积层和全连接层. 基于此, 本文首先给出具有统一结构的深度 ReLU 神经网络模型, 其次着重研究该深度神经网络的万有一致性, 具体从以下 3 方面展开讨论:

(1) 给出一种具有统一结构的深度神经网络模型, 使得该网络具有万有逼近性, 即能以任意精度逼近连续函数.

(2) 在上述深度神经网络模型下, 针对不同的学习任务 (先验信息), 严格证明该神经网络具有强万有一致性.

(3) 从实验角度进一步验证理论结果.

本文的主要贡献可归纳为两个方面: (1) 在模型上, 本文给出了深度神经网络在深度、宽度、激活函数及网络连接方式中的选择方法, 从而获得一个具有统一结构 (而非特殊结构) 的深度神经网络模型; (2) 在理论上, 本文证明了上述深度神经网络模型在实现多重学习任务的同时不仅具有万有逼近性, 还具有强万有一致性. 综上, 本文在完善神经网络理论研究的同时, 为深入理解和认识复杂模型的本质和原理奠定基础.

本文剩余部分结构如下: 第 2 节介绍深度神经网络及部分理论结果. 第 3 节是本文的主要结论 —— 深度神经网络的万有一致性. 第 4 节是数值实验, 验证理论结果的合理性. 第 5 节是主要结论的证明, 具体包括容量的估计、深度神经网络的逼近性和万有一致性的证明. 第 6 节是全文总结.

2 深度神经网络

设 $d\in\mathbb{N}$, $X\subseteq\mathbb{R}^d$ 是输入空间, $Y\subseteq\mathbb{R}$ 是输出空间, 令 $Z:=X\times Y$, 且 $D=(x_i,y_i)_{i=1}^m$ 是 Z 上按某种分布 ρ 产生的样本. 设 $L\in\mathbb{N}$ 为神经网络的深度, $d_0=d$ 是输入数据的维度, $d_l\in\mathbb{N}$ 是第 l 层的宽度 $(l=1,2,\dots,L)$. $\sigma(t)=t_+:=\max\{t,0\}$ 是 ReLU 函数, 对于任意的 $x=(x^{(1)},\dots,x^{(d)})^{\mathrm{T}}\in\mathbb{R}^d$, 有

$$\sigma(x)=(\sigma(x^{(1)}),\dots,\sigma(x^{(d)}))^{\mathrm{T}}=(\max\{0,x^{(1)}\},\dots,\max\{0,x^{(d)}\})^{\mathrm{T}}.$$

令 $\mathcal{J}_l:\mathbb{R}^{d_{l-1}}\to\mathbb{R}^{d_l}$, $\mathcal{J}_l(x):=W_lx+\boldsymbol{b}_l$, 其中 W_l 是 $d_l\times d_{l-1}$ 的权矩阵, $\boldsymbol{b}_l\in\mathbb{R}^{d_l}$ 是阈值向量. 令 $\boldsymbol{a}\in\mathbb{R}^{d_L}$, 则含有 L 个隐层、激活函数是 ReLU 的深度网络定义为

$$N_{d_1,\dots,d_L}(x)=\boldsymbol{a}\cdot\sigma\circ\mathcal{J}_L\circ\sigma\circ\mathcal{J}_{L-1}\circ\cdots\circ\sigma\circ\mathcal{J}_1(x). \tag{1}$$

表 1　不同先验信息对应的深度神经网络模型

Table 1　Deep neural network models based on different prior information

Features	Depth	Width	Structure	Activation function
Smoothness	2 / 3	$\mathcal{O}\left(m^{\frac{d}{2r+d}}\right)$	Full-structure / tree-structure	Sigmoidal and Heaviside [19] / sigmoidal [24]
Sparseness	2 / 3	$\mathcal{O}\left(\left(\frac{ms}{N^d}\right)^{\frac{d}{2r+d}}\right)$	Full-structure / tree-structure	Sigmoidal and Heaviside [19] / sigmoidal [24]
Radial	3	$\mathcal{O}\left(m^{\frac{1}{2r+1}}\right)$	Tree-structure	Univariate activation function [18]
Manifold	3	$\mathcal{O}\left(m^{\frac{1}{2r+d}}\right)$	Full-structure	Heaviside and square-rectifier [25]
Piecewise-smooth	3	$\mathcal{O}\left(m^{\frac{d}{2r+d}}\right)$	Full-structure	Sigmoidal [26]

显然, 深度神经网络 $N_{d_1,\ldots,d_L}(x)$ 的结构主要依赖于权矩阵 W_l, 阈值向量 $\boldsymbol{b}_l$ 和连接方式. 本文主要研究含有 L 个隐层、n 个非零参数、每个隐层宽度为 $d+1$、激活函数 σ 为 ReLU 且具有式 (1) 结构的深度全连接神经网络, 记作 $\mathcal{H}_{\sigma,L} := \mathcal{H}_{L,n,d+1,\sigma}$. 本文的假设空间选为 $\mathcal{H}_{\sigma,L}$, 则基于经验风险极小化的估计函数定义为

$$f_{D,L} := \arg\min_{f\in\mathcal{H}_{\sigma,L}} \mathcal{E}_D(f), \tag{2}$$

其中 $\mathcal{E}_D(f) = \frac{1}{m}\sum_{i=1}^{m}|f(x_i)-y_i|^2$.

已有结果表明不同的先验信息对应不同的深度神经网络模型, 具体参见表 1 [18,19,24~26]. 能否找到一个具有统一结构的深度神经网络模型, 使得该神经网络能学习具有不同先验信息的特征? 下述命题 1 回答了这个问题, 即具有两个隐层且宽度向量为 $(2d+1,d)^{\mathrm{T}}$ 的深度神经网络可以任意逼近连续函数.

命题 1　存在非多项式激活函数 $\sigma\in C(\mathbb{R})$ 满足: 对任意的 $f\in C(I^d)$ 及 $\varepsilon>0$, 存在 $b_j, b_{jk}, \nu_k\in\mathbb{R}$ 使得

$$\left|f(x)-\sum_{j=1}^{2d+1}\sigma\left(\sum_{k=1}^{d}\nu_k\sigma(x^{(k)}-b_{jk})-b_j\right)\right|<\varepsilon.$$

虽然具有两个隐层、有限个神经元的神经网络具有万有逼近性质, 遗憾的是下述命题 2 告诉我们该网络的容量太大、不可控.

命题 2　对任意的 $\Gamma>0$, 存在非多项式激活函数 $\sigma\in C(\mathbb{R})$, 使得对任意的 $\varepsilon>0$, 均有

$$\mathcal{N}(\varepsilon, N_{d,(2d+1)d,\sigma}, L_1(I^d)) \geqslant C\left(\frac{1}{\varepsilon}\right)^{\Gamma},$$

其中 C 是与 ε,Γ 无关的常数, $\mathcal{N}(\varepsilon, N_{d,(2d+1)d,\sigma}, L_1(I^d))$ 表示在 $L_1(I^d)$ 范数下 $N_{d,(2d+1)d,\sigma}$ 的覆盖数 (详细参见第 5 节中定义 2).

命题 1 中的深度神经网络虽然能够以任意精度逼近连续函数 (即具有万有逼近性), 但由于该神经网络容量不可控, 从而很难设计出有效的学习算法从 $N_{d,(2d+1)d,\sigma}$ 中寻找特定函数的逼近, 这也意味着万有一致性中一个很重要的条件 (假设空间容量可控) 不成立, 故该深度神经网络不具有万有一致性. 此外, 大容量的假设空间容易产生过拟合, 即所学的函数能很好地拟合训练数据, 但泛化能力可能会不理想. 因此, 命题 1 中所构造的深度神经网络虽然能回答引言中的问题 (1), 但并不满足问题 (2).

3 深度神经网络的万有一致性

对于任意的 $f: X \to Y$, 泛化误差 (也叫期望风险) 定义为

$$\mathcal{E}(f) = \int_Z (f(x) - y)^2 \mathrm{d}\rho. \tag{3}$$

由文献 [27] 知, 回归函数

$$f_\rho = \int_Y y \mathrm{d}\rho(y|x)$$

是泛化误差式 (3) 的极小化, 且对于任意的 $f \in L^2_{\rho_X}$ 均有

$$\mathcal{E}(f_\rho) \leqslant \mathcal{E}(f),$$

和

$$\mathcal{E}(f) - \mathcal{E}(f_\rho) = \|f - f_\rho\|^2_\rho, \tag{4}$$

这里 ρ_X 是 X 上的边缘分布, $L^2_{\rho_X}$ 是 X 上关于 ρ_X 的平方可积的希尔伯特 (Hilbert) 空间, 其范数记为 $\|\cdot\|^2_\rho$.

众所周知, 一致性描述的是随着样本数的增大, 泛化误差是否收敛于 0 的问题. 下面给出强万有一致的定义 [10, Definition 1.4].

定义1 对于任意满足 $\int_Y y^2 \mathrm{d}\rho(y|x) < \infty$ 的概率分布 ρ, 如果

$$\lim_{m\to\infty} \mathcal{E}(f_m) - \mathcal{E}(f_\rho) = 0$$

依概率 1 成立, 则称估计函数序列 $\{f_m\}_{m=1}^\infty$ 是强万有一致的.

不同于命题 1, 下述定理 1 给出了一个具有统一结构的深度神经网络, 即 L 个隐层、每层都是 $d+1$ 个神经元的 ReLU 全连接神经网络, 该深度神经网络模型不仅具有万有逼近性 (参见定理 1 的证明部分), 而且还具有万有一致性.

定理1 设 $\{f_{D,L_m}\}_{m=1}^\infty$ 如式 (2) 定义, σ 是 ReLU 激活函数. 当 $m \to \infty$ 时, 如果 $L := L_m$ 和 $M := M_m$ 满足

$$M_m \to \infty,\ L_m \to \infty,\ \frac{M_m^2}{m^\theta} \to 0, \tag{5}$$

和

$$\frac{C_0^* M_m^4 L_m^2 (d+1)(d+2)\log(L_m(d+1))\log(C_1^* M_m m^\theta)}{m^{1-2\theta}} \to 0, \tag{6}$$

则 $\pi_{M_m} f_{D,L_m}$ 是强万有一致的, 其中 $0 < \theta < 1/2$, C_0^* 和 C_1^* 是与 L_m, M_m, d, θ 无关的常数, $\pi_{M_m}(t) = \min\{M_m, |t|\}$ 是截断函数.

定理 1 的结论回答了引言中的问题 (1) 和 (2): 针对问题 (1), 定理 1 构造了一个具有统一结构的深度神经网络模型 (非表 1 中的特定模型), 即深度为 L_m 层、每层的宽度都是 $d+1$、激活函数是 ReLU 的全连接神经网络, 从函数逼近的角度证明了该深度神经网络具有万有逼近性. 针对问题 (2), 所构建的深度 ReLU 神经网络可以实现不同的学习任务 (即不受先验的限制). 同时, 定理 1 给出了该深度神经网络在满足强万有一致性时各参数之间的关系 (如式 (5) 和 (6) 所示).

现对定理 1 中的条件 (5) 和 (6) 作以下具体说明: 在式 (5) 中, 条件 $L := L_m \to \infty$ 要求假设空间 $\mathcal{H}_{\sigma,L}$ 在连续函数空间中是稠密的, 然而条件 (6) 要求 $\mathcal{H}_{\sigma,L}$ 不能太大, 从而可以给出它的容量估计.

同时, 条件 (6) 还给出了深度神经网络的层数 L_m 的约束条件. 由于一致性刻画的是当样本数趋于无穷时, 估计函数和回归函数之间误差的极限形式, 而估计函数 f_{D,L_m} 是在假设空间 $\mathcal{H}_{\sigma,L}$ 中通过最小二乘的优化策略获得的, 其不具有有界性, 所以假设输出函数有界是不合适的. 但在理论分析中, 我们不得不给出估计函数的一些结构信息, 比如有界性, 因此这里用截断算子来截断估计函数 f_{D,L_m} 是必要的. 此外, 由于我们很难给出估计函数的严格结构, 而且用有界的估计函数来预测无界的样本本身也不合理, 因此需要对估计函数给出一定的限制, 即当 $m \to \infty$ 时, 要求 $M_m \to \infty$. 条件 $\frac{M_m^2}{m^\theta} \to 0$ 在描述截断算子截断水平的同时, 控制输出样本的有界性, 并表明 M_m 随着 m 增长的速度要慢于 m^θ. 总之, 增加条件 (5) 和 (6) 是为了控制估计函数的复杂性, 也是为了控制假设空间的容量. 在无界参数的条件下证明假设空间的容量可控是定理 1 的证明过程中的一个关键问题, 也是本文证明过程中的一个难点.

4 数值实验

为了验证定理 1 的有效性, 本节分别在人工数据集和真实数据集上进行了数值实验. 在实验中, 我们采用全连接神经网络和 ReLU 激活函数, 且每个隐藏层的神经元个数 (即网络宽度 d_l) 相同. 对于参数选择问题, 将神经网络的深度 L 和训练迭代轮次 T 作为算法实施的关键参数, 通过网格搜索方法选取, 其他实验设置如表 2 所示.

4.1 人工数据

人工数据集的生成是按照如下方式进行的. 训练样本集的输入 $\{x_i\}_{i=1}^N$ 通过对超立方体 $[0,1]^d$ 上均匀分布的独立采样获得, 其对应的输出 $\{y_i\}_{i=1}^N$ 根据回归模型 $y_i = g(x_i) + \epsilon_i$ 生成, 其中 ϵ_i 是独立的高斯 (Gaussian) 噪声 $\mathcal{N}(0,\sigma^2)$, 定义函数

$$g(x) = \begin{cases} (1-\|x\|_2)^6(35\|x\|_2^2 + 18\|x\|_2 + 3), & \text{if } 0 < \|x\|_2 \leqslant 1, \\ 0, & \text{if } \|x\|_2 > 1. \end{cases}$$

测试样本集的输入 $\{x_i'\}_{i=1}^{N'}$ 同样通过对超立方体 $[0,1]^d$ 上均匀分布的独立采样获得, 而对应的输出 $\{y_i'\}_{i=1}^{N'}$ 通过 $y_i' = g(x_i')$ 生成.

为了验证深度神经网络泛化误差随着训练样本个数的变化情况, 我们取训练样本个数 $N \in \{500, 1000, \ldots, 10000\}$, 并固定测试样本个数 $N' = 1000$. 在实验中, 设置数据维度为 $d = 3$; 生成 3 种幅度的高斯噪声, 即 $\sigma = 0.1$, $\sigma = 0.5$ 和 $\sigma = \sqrt{0.5}$; 采用 3 种宽度的神经网络, 即 $d_l = d$, $d_l = d+1$ 和 $d_l = 10d$. 神经网络的深度 L 和训练迭代轮次 T 分别从集合 $\{3,4,\ldots,15\}$ 和 $\{1,2,\ldots,2000\}$ 中选取. 针对每一组 (N,σ,d_l), 重复执行 50 次实验, 并记录每次实验在最优参数下的均方根误差 (root mean square error, RMSE), 将 50 次实验 RMSE 的平均值作为泛化误差的度量.

不同宽度的深度神经网络的泛化误差随着训练样本个数的变化情况如图 1 所示, 其中图 1(a)~(c) 分别表示在高斯噪声为 $\mathcal{N}(0,0.1^2)$, $\mathcal{N}(0,0.5^2)$ 和 $\mathcal{N}(0,0.5)$ 的结果. 从以上结果可以获得以下结论: (1) 对于不同宽度的深度神经网络和不同噪声幅度的训练数据而言, 深度神经网络的泛化误差均随着训练样本个数的增加而递减. (2) 在 3 种高斯噪声的训练数据集上, 虽然网络宽度的增加使得其泛化性能获得了改进, 但是宽度为 $d+1$ 的神经网络的泛化误差对宽度为 d 的神经网络的泛化误差的改进效果显著优于宽度 $10\,d$ (远远大于 $d+1$) 对 $d+1$ 的改进, 这种现象在噪声较大的数据上更加明显, 这

表 2 深度神经网络训练的实验设置

Table 2 Experimental setup for deep nets training

Initialization	Optimizer	Learning rate & decay				Batch size	Weight decay	Loss
		Toy	Real	Toy	Real			
MSRA [28]	SGD	0.01	0.1	1	0.95	50	0.001	MSELoss

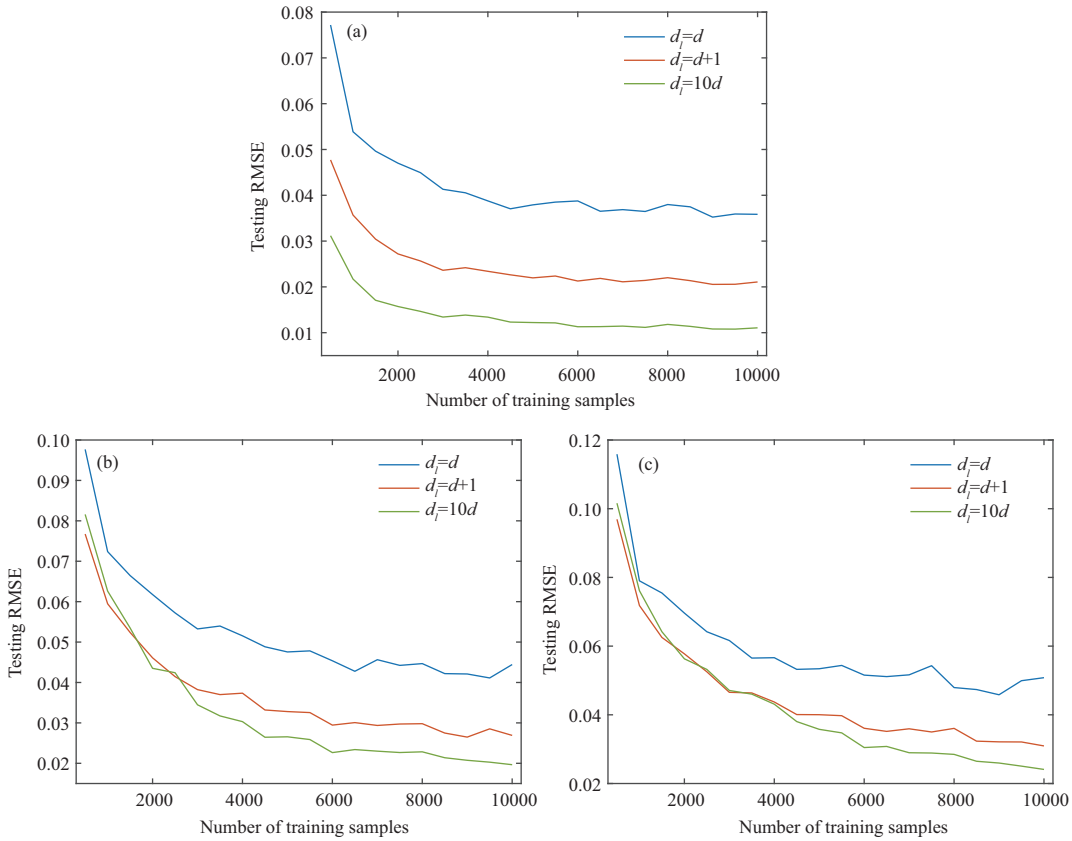

图 1 (网络版彩图) 不同宽度的深度神经网络的泛化误差随着训练样本个数增加的变化情况比较

Figure 1 (Color online) Performance comparison of deep nets with different widths as the number of training samples increases. (a) Gaussian noise $\mathcal{N}(0, 0.1^2)$; (b) Gaussian noise $\mathcal{N}(0, 0.5^2)$; (c) Gaussian noise $\mathcal{N}(0, 0.5)$

说明深度神经网络的宽度应该至少为 $d+1$, 也进一步验证了定理 1 中的深度 ReLU 神经网络的万有一致性, 从而引言中的问题 (3) 得以证实.

4.2 真实数据

我们在 5 个标准数据集上验证了定理 1 的正确性. 这 5 个标准数据集分别来自于 UCI 机器学习库的 Wine Quality 数据集[1)], 波尔图大学 (University of Porto) 人工智能与计算机科学实验室的 Delta_ailerons, Delta_elevators 和 Bank8FM 数据集[2)]. 其中, Wine Quality 包括与来自葡萄牙北部的红葡萄酒和白葡萄酒样本相关的两个数据集, 我们分别记为 Wine_red 和 Wine_white, 其目标是基于

1) http://archive.ics.uci.edu/ml/datasets.php.
2) https://www.dcc.fc.up.pt/%7eltorgo/Regression/.

表 3　不同宽度的神经网络在真实数据集上的泛化误差比较

Table 3　Performance comparison of deep nets with different widths on real-world data

Dataset	Neuron number in hidden layers				
	d	$d+1$	1.5 d	3 d	5 d
Wine_red	0.6374	0.6186	0.6181	0.6180	0.6338
	(0.0497)	(0.0176)	(0.0173)	(0.0183)	(0.0484)
Wine_white	0.7921	0.7613	0.7457	0.7545	0.7363
	(0.0690)	(0.0360)	(0.0278)	(0.0524)	(0.0366)
Delta_ailerons	3.0029	2.9559	2.8490	2.9062	2.9765
($\times 10^{-4}$)	(0.1608)	(0.0961)	(0.2950)	(0.2688)	(0.2937)
Delta_elevators	2.2700	2.1472	2.0574	2.2118	2.2480
($\times 10^{-3}$)	(0.1557)	(0.3150)	(0.3859)	(0.2644)	(0.2434)
Bank8FM	4.6527	3.5052	3.5966	3.8827	3.8416
($\times 10^{-2}$)	(3.5837)	(0.2346)	(0.1927)	(0.2637)	(0.1718)

一些物理化学指标对葡萄酒质量进行建模. Wine_red 和 Wine_white 各自包含了 1599 和 4898 个样本, 每个样本具有 11 个自变量维度. 数据集 Delta_ailerons 和 Delta_elevators 是从 F16 飞机的副翼控制和升降舵控制任务中获得, 分别包含拥有 5 个自变量维度的 7129 个样本和拥有 6 个自变量维度的 9517 个样本. 数据集 Bank8FM 用来预测由于所有开放的出纳员都排满了队而被银行拒之门外的银行客户比例, 其包含 4499 个样本, 每个样本具有 8 个自变量维度.

实验设置是按照如下方式进行. 针对每个数据集, 随机选择 80% 的数据作为训练集, 剩余 20% 的数据作为测试集. 采用 5 种宽度的神经网络, 即 $d_l = d, d+1, 1.5d, 3d, 5d$. 在优化求解过程中, 运用固定步长衰减的方式设置学习率, 其中初始学习率设置为 0.1, 学习率衰减设置为 0.95, 衰减步长设置为 10. 神经网络的深度 L 和训练迭代轮次 T 分别从集合 $\{3,4,5,6,7\}$ 和 $\{1,2,\ldots,2000\}$ 中选取. 在每个数据集上重复执行 10 次实验, 将 RMSE 的平均值作为泛化误差的度量.

不同宽度的神经网络在真实数据集上的泛化误差如表 3 所示, 其中括号内的数值代表标准差. 可以看出, 网络宽度由 d 变为 $d+1$, 泛化误差均会发生本质性的改变, 并接近在最优宽度下的泛化误差, 这种现象在数据集 Wine_red, Wine_white 和 Bank8FM 上尤为明显. 这进一步说明了神经网络的宽度应该至少为 $d+1$, 同时也验证了本文所建立的深度 ReLU 神经网络的万有一致性理论在真实数据集的实用性.

5　主要结论的证明

首先给出覆盖数 (定义 2) 和填充数 (定义 3) 的定义 (具体参见文献 [10, Definition 9.3 和 Definition 9.4]).

定义2　设 $\varepsilon > 0$, $\mathcal{G}: \mathbb{R}^d \to \mathbb{R}$, $1 \leqslant p < \infty$, ϑ 是 $\mathbb{R}^d$ 上的概率测度. 令 $f: \mathbb{R}^d \to \mathbb{R}$ 且

$$\|f\|_{L_p(\vartheta)} := \left\{ \int |f(z)|^p \mathrm{d}\vartheta \right\}^{\frac{1}{p}}. \tag{7}$$

(1) 设 $\{g_1, \ldots, g_N\}$ 是 $\mathcal{G}$ 中有限函数集, 对任意的 $g \in \mathcal{G}$, 总存在一个 $j = j(g) \in \{1, \ldots, N\}$ 满足

$$\|g - g_j\|_{L_p(\vartheta)} \leqslant \varepsilon,$$

则称 $\{g_1,\dots,g_N\}$ 为 $\mathcal{G}$ 在 $\|\cdot\|_{L_p(\vartheta)}$ 范数下的 ε 覆盖.

(2) $\mathcal{G}$ 在 $\|\cdot\|_{L_p(\vartheta)}$ 范数下的最小 ε 覆盖的元素数量称为 $\mathcal{G}$ 在 $\|\cdot\|_{L_p(\vartheta)}$ 范数下的覆盖数, 记为 $\mathcal{N}(\varepsilon,\mathcal{G},\|\cdot\|_{L_p(\vartheta)})$. 如果 $\mathcal{G}$ 的 ε 覆盖不存在, 则 $\mathcal{N}(\varepsilon,\mathcal{G},\|\cdot\|_{L_p(\vartheta)})=\infty$.

(3) 设 $z_1^n=(z_1,\dots,z_n)$ 是 $\mathbb{R}^d$ 中 n 个固定点, ϑ_n 是 $\mathbb{R}^d$ 上的经验测量, 且

$$\|f\|_{L_p(\vartheta_n)}:=\left\{\frac{1}{n}\sum_{i=1}^{n}|f(z_i)|^p\right\}^{\frac{1}{p}}. \tag{8}$$

$\mathcal{G}$ 在 $\|\cdot\|_{L_p(\vartheta_n)}$ 范数下的覆盖数 $\mathcal{N}_p(\varepsilon,\mathcal{G},z_1^n)$ 是满足下述性质的最小 k: 存在函数 $g_1,g_2,\dots,g_k:\mathbb{R}^d\to\mathbb{R}$ 使得对于任意的 $g\in\mathcal{G}$, 总存在一个 $j=j(g)\in\{1,\dots,k\}$ 满足

$$\left\{\frac{1}{n}\sum_{i=1}^{n}|g(z_i)-g_j(z_i)|^p\right\}^{\frac{1}{p}}<\varepsilon.$$

定义3 设 $\varepsilon>0$, $\mathcal{G}:\mathbb{R}^d\to\mathbb{R}$, $1\leqslant p<\infty$, ϑ 是 $\mathbb{R}^d$ 上的概率测度.

(1) 对任意的 $1\leqslant j<k\leqslant N$, 若每个有限函数集 $\{g_1,\dots,g_N\}\subseteq\mathcal{G}$ 满足

$$\|g_j-g_k\|_{L_p(\vartheta)}\geqslant\varepsilon,$$

则称 $\{g_1,\dots,g_N\}$ 为 $\mathcal{G}$ 在 $\|\cdot\|_{L_p(\vartheta)}$ 范数下的 ε 填充, 这里 $\|\cdot\|_{L_p(\vartheta)}$ 如式 (7) 定义.

(2) $\mathcal{G}$ 在 $\|\cdot\|_{L_p(\vartheta)}$ 范数下的最大 ε 填充的元素数量称为 $\mathcal{G}$ 在 $\|\cdot\|_{L_p(\vartheta)}$ 范数下的填充数, 记为 $\mathcal{M}_p(\varepsilon,\mathcal{G},\|_{L_p(\vartheta)})$. 如果对于任意的 $N\in\mathbb{N}$, 在 $\|\cdot\|_{L_p(\vartheta)}$ 范数下存在 $\mathcal{G}$ 的一个大小为 N 的 ε 填充, 则 $\mathcal{M}_p(\varepsilon,\mathcal{G},\|_{L_p(\vartheta)})=\infty$.

(3) 设 $z_1^n=(z_1,\dots,z_n)$ 是 $\mathbb{R}^d$ 中 n 个固定点, ϑ_n 是 $\mathbb{R}^d$ 上的经验测量, $\mathcal{G}$ 在 $\|\cdot\|_{L_p(\vartheta_n)}$ 范数下的填充数 $\mathcal{M}_p(\varepsilon,\mathcal{G},z_1^n)$ 是满足下述性质的最大的 N: 存在函数 $g_1,g_2,\dots,g_N:\mathbb{R}^d\to\mathbb{R}$, 对任意的 $1\leqslant j<k\leqslant N$ 使得

$$\left\{\frac{1}{n}\sum_{i=1}^{n}|g_j(z_i)-g_k(z_i)|^p\right\}^{\frac{1}{p}}\geqslant\varepsilon.$$

下述引理 1 表明: 存在一个激活函数, 使得只有一个神经元的浅层神经网络仍具有万有逼近性.

引理1 对于任意的 $\varepsilon>0$, $f\in C([0,1])$, 存在非多项式激活函数 $\sigma\in C(\mathbb{R})$ 和 $b\in\mathbb{R}$ 使得

$$|f(t)-\sigma(t-b)|<\varepsilon.$$

证明 由于 $C([0,1])$ 是可分的 Banach 空间, 从而存在可数的稠密子集 $\{\nu_k\}_{k=1}^{\infty}$, 即对任意的 $\varepsilon>0$, 存在 $k_0\in\mathbb{N}$ 使得

$$|f(t)-\nu_{k_0}(t)|<\varepsilon. \tag{9}$$

对任意的 $m\in\mathbb{N}$, 定义

$$\sigma(t)=\begin{cases}\nu_m(t-(2m-2)), & t\in[2m-2,2m-1],\\ \sigma(2m-1)+(\sigma(2m)-\sigma(2m-1))(t-2m+1), & t\in(2m-1,2m),\\ \nu_1(0), & t\in(-\infty,0).\end{cases} \tag{10}$$

显然 $\sigma \in C(\mathbb{R})$ 且不是多项式, 由式 (10) 易知,

$$\nu_m(t) = \sigma(t + 2m - 2), t \in [0, 1].$$

再结合式 (9) 知, 对任意的 $k > k_0$ 和 $\varepsilon > 0$, 下式成立

$$|f(t) - \sigma(t + 2k - 2)| = |f(t) - \nu_k(t)| < \varepsilon.$$

引理 2 [29] 表明: 任意一个多维连续函数均可以表示成有限个一维函数的复合函数.

引理2 任意一个定义在 d 维超立方体 I^d $(d \geqslant 2)$ 上的连续函数 $f(x) = f(x^{(1)}, \ldots, x^{(d)})$ 均可表示成

$$f(x^{(1)}, \ldots, x^{(d)}) = \sum_{j=1}^{2d+1} g\left(\sum_{k=1}^{d} \nu_k \psi_j(x^{(k)})\right),$$

其中 $\nu_k \in \mathbb{R}$ $(k = 1, \ldots, d)$ 满足 $\nu_k \geqslant 0$ 且 $\sum_{k=1}^{d} \nu_k \leqslant 1$, $g \in C[0, 1]$, $\psi_j : [0, 1] \to [0, 1]$ $(j = 1, \ldots, 2d+1)$.

5.1 命题 1 的证明

证明 由引理 1 知, 存在非多项式激活函数 $\sigma \in C(\mathbb{R})$, b_j 和 b_{jk}, 对引理 2 中的任意g, ψ_j $(j = 1, 2, \ldots, 2d + 1)$, 以及 $\nu > 0$ 均满足

$$|g(t) - \sigma(t - b_j)| < \nu, \ \forall t \in \mathbb{R},$$

以及

$$|\psi_j(x^{(k)}) - \sigma(x^{(k)} - b_{jk})| \leqslant \nu, \quad \forall j = 1, \ldots, 2d + 1, \quad k = 1, \ldots, d,$$

从而有

$$\begin{aligned}
&\left|\sum_{j=1}^{2d+1} g\left(\sum_{k=1}^{d} \nu_k \psi_j(x^{(k)})\right) - \sum_{j=1}^{2d+1} \sigma\left(\sum_{k=1}^{d} \nu_k \sigma(x^{(k)} - b_{jk}) - b_j\right)\right| \\
&\quad \leqslant \left|\sum_{j=1}^{2d+1} g\left(\sum_{k=1}^{d} \nu_k \psi_j(x^{(k)})\right) - \sum_{j=1}^{2d+1} g\left(\sum_{k=1}^{d} \nu_k \sigma(x^{(k)} - b_{jk})\right)\right| \\
&\qquad + \left|\sum_{j=1}^{2d+1} g\left(\sum_{k=1}^{d} \nu_k \sigma(x^{(k)} - b_{jk})\right) - \sum_{j=1}^{2d+1} \sigma\left(\sum_{k=1}^{d} \nu_k \sigma(x^{(k)} - b_{jk}) - b_j\right)\right|.
\end{aligned}$$

由于 g 在 $C[0, 1]$ 上连续, 故对于任意小的 $\delta > 0$, 存在 $\eta > 0$, 使得当 $|t - t'| \leqslant \eta$ 时, 有 $|g(t) - g(t')| \leqslant \delta$. 令 $\nu < \eta$, 则由 $\nu_k \geqslant 0$, $\sum_{k=1}^{d} \nu_k \leqslant 1$ 可得

$$\left|\sum_{k=1}^{d} \nu_k \psi_j(x^{(k)}) - \sum_{k=1}^{d} \nu_k \sigma(x^{(k)} - b_{jk})\right| \leqslant \sum_{k=1}^{d} \nu_k \nu \leqslant \nu.$$

从而有

$$\left|\sum_{j=1}^{2d+1} g\left(\sum_{k=1}^{d} \nu_k \psi_j(x^{(k)})\right) - \sum_{j=1}^{2d+1} g\left(\sum_{k=1}^{d} \nu_k \sigma(x^{(k)} - b_{jk})\right)\right| \leqslant \delta.$$

又因为

$$\left|\sum_{j=1}^{2d+1} g\left(\sum_{k=1}^{d}\nu_k\sigma(x^{(k)}-b_{jk})\right)-\sum_{j=1}^{2d+1}\sigma\left(\sum_{k=1}^{d}\nu_k\sigma(x^{(k)}-b_{jk})-b_j\right)\right|\leqslant(2d+1)\nu,$$

故有

$$\left|\sum_{j=1}^{2d+1} g\left(\sum_{k=1}^{d}\nu_k\psi_j(x^{(k)})\right)-\sum_{j=1}^{2d+1}\sigma\left(\sum_{k=1}^{d}\nu_k\sigma(x^{(k)}-b_{jk})-b_j\right)\right|\leqslant(2d+1)\nu+\delta.$$

令 $(2d+1)\nu+\delta=\varepsilon$, 则命题 1 得证.

5.2 命题 2 的证明

证明 由命题 1 可知, 存在非多项式激活函数 $\sigma\in C(\mathbb{R})$ 满足: 对任意的 $f\in C(I^d)$ 及 $\nu>0$, 均有

$$\mathrm{dist}(f,N_{d,(2d+1)d,\sigma},L_1(I^d))<\nu,$$

其中 $\mathrm{dist}(f,N_{d,(2d+1)d,\sigma},L_1(I^d))$ 表示在 $L_1(I^d)$ 范数下 f 与 $N_{d,(2d+1)d}$ 的距离. 而命题 1 的逆否命题表明, 若有

$$\mathrm{dist}(C(I^d),N_{d,(2d+1)d,\sigma},L_1(I^d))\geqslant C'(n\log_2(n+1))^{-1}\geqslant C'n^{-2},$$

则必存在与 ν,ε 无关的实数 $\beta,\tilde{C}_1,\tilde{C}_2$ 使得

$$\mathcal{N}(\varepsilon,N_{d,(2d+1)d,\sigma},L_1(I^d))\geqslant\tilde{C}_1\left(\frac{\tilde{C}_2n^{\beta}}{\varepsilon}\right)^{n}.$$

令 $\nu=C'n^{-2}$, 即 $n=(C')^{-1}\nu^{-2}$, 则有

$$\mathcal{N}(\varepsilon,N_{d,(2d+1)d,\sigma},L_1(I^d))\geqslant\tilde{C}_1\left(\frac{\tilde{C}_2(C')^{-\beta}\nu^{-2\beta}}{\varepsilon}\right)^{(C')^{-1}\nu^{-2}}.$$

由于 $\nu>0$ 可以任意小, 不妨令 $(C')^{-1}\nu^{-2}\geqslant\Gamma$, 则有

$$\mathcal{N}(\varepsilon,N_{d,(2d+1)d,\sigma},L_1(I^d))\geqslant C\left(\frac{\Gamma^{\beta}}{\varepsilon}\right)^{\Gamma}.$$

命题 2 证毕.

5.3 定理 1 的证明

定理 1 的证明主要分为 3 部分: 首先对假设空间进行容量估计, 其次讨论深度神经网络的逼近性, 最后将泛化误差分解为 8 项, 并逐一进行误差估计.

5.3.1 容量估计

我们利用覆盖数给出假设空间容量大小的估计, 引理 3 给出了覆盖数和填充数之间的关系, 可参见文献 [10, Lemma 9.2].

引理3 设 $\varepsilon>0$, $p\geqslant1$, $\mathcal{G}$ 是 $\mathbb{R}^d$ 上的某类函数集, ϑ 是 $\mathbb{R}^d$ 上的概率测度, 则

$$\mathcal{M}(2\varepsilon,\mathcal{G},\|\cdot\|_{L_p(\vartheta)})\leqslant\mathcal{N}(\varepsilon,\mathcal{G},\|\cdot\|_{L_p(\vartheta)})\leqslant\mathcal{M}(\varepsilon,\mathcal{G},\|\cdot\|_{L_p(\vartheta)}).$$

特别地, 对所有的 $z_1,\dots,z_n\in\mathbb{R}^d$ 有

$$\mathcal{M}_p(2\varepsilon,\mathcal{G},z_1^n)\leqslant\mathcal{N}_p(\varepsilon,\mathcal{G},z_1^n)\leqslant\mathcal{M}_p(\varepsilon,\mathcal{G},z_1^n).$$

下面主要利用填充数和伪维数的关系 (参见文献 [23, Lemma 2 和 Lemma 3]), 给出填充数的上界估计, 即

$$\mathcal{M}(\varepsilon,\pi_M\mathcal{H}_{\sigma,L},\|\cdot\|_{L_1(\vartheta)})\leqslant 2\left(\frac{2eM}{\varepsilon}\right)^{2C_0Ln\log\tilde{d}},\tag{11}$$

其中 $C_0>0$ 是一个常数, n 和 $\tilde{d}$ 分别是深度全连接神经网络空间 $\mathcal{H}_{\sigma,L}$ 中的非零参数和神经元个数. 由于 $\mathcal{H}_{\sigma,L}:=\mathcal{H}_{L,n,d+1,\sigma}$ 含有 L 个隐层、每个隐层的宽度 (神经元个数) 为 d+1、输入维数是 d、输出维数是 1, 结合式 (1) 中神经网络的结构可知, L 个隐层中神经元的总个数为 $\tilde{d}=L(d$+1) 个, 第 1 个隐层内权数和阈值数共有 $d(d$+1)+d+1 个, 第 2∼L 层每层的内权数和阈值数共有 $(d$+1)(d+2) 个, 第 L 层的输出权数为 d+1 个, 故该神经网络中参数的总个数为 $n=(d$+1)^2+(L−1)(d+1)(d+2)+d+1 = $L(d$+1)(d+2) 个. 再结合引理 3 和式 (11) 的结果, 可以直接得到本文中假设空间 $\mathcal{H}_{\sigma,L}$ 的覆盖数的上界估计.

引理4 设 $0<\varepsilon\leqslant M$, 则有

$$\mathcal{N}_p(\varepsilon,\mathcal{H}_{\sigma,L},z_1^n)\leqslant 2\left(\frac{2eM}{\varepsilon}\right)^{2C_0L^2(d+1)(d+2)\log(L(d+1))},$$

即

$$\log\mathcal{N}_p(\varepsilon,\mathcal{H}_{\sigma,L},z_1^n)\leqslant C_1L^2(d+1)(d+2)\log(L(d+1))\log\left(\frac{M}{\varepsilon}\right),$$

其中 C_1 是与 L, M, d, ε 无关的常数.

5.3.2 深度神经网络的逼近性

基于文献 [30,31], 给出深度神经网络 $\mathcal{H}_{\sigma,L}$ 的逼近结果.

引理5 设 $f:[0,1]^d\to\mathbb{R}_+$, $f\in C[0,1]^d$ 且 $\|f\|_C=1$. $\mathcal{H}_{\sigma,L}$ 是一个含有 L 个隐层的深度全连接神经网络, 每个隐层宽度为 $d+1$, 激活函数 σ 是 ReLU 函数, 输入维数是 d, 输出维数是 1, 则对于任意的 $f_H\in\mathcal{H}_{\sigma,L}$ 有

$$\lim_{L\to\infty}\|f-f_H\|_C=0,$$

其中 $\|f\|_C:=\sup_{x\in[0,1]^d}|f(x)|$.

下述引理 6 是集中不等式 [10, Theorem 11.4], 它是泛化误差分析的主要工具之一.

引理6 设 $B\geqslant 1$, $\mathcal{F}$ 是由函数 $f:X\to\mathbb{R}$ 所构成的集合且满足 $|f(x)|\leqslant B$. 则对于任意的 $f\in\mathcal{F}$,

$$\mathcal{E}(f)-\mathcal{E}(f_\rho)-(\mathcal{E}_D(f)-\mathcal{E}_D(f_\rho))\leqslant\varepsilon(\alpha+\beta+\mathcal{E}(f)-\mathcal{E}(f_\rho))$$

依概率

$$1-14\max_{x_1^m\in X^m}\mathcal{N}_1\left(\frac{\beta\varepsilon}{20B},\mathcal{F},x_1^m\right)\exp\left(-\frac{\varepsilon^2(1-\varepsilon)\alpha m}{214(1+\varepsilon)B^4}\right)$$

成立, 这里 $m\geqslant 1$, $\alpha,\beta>0$, $0<\varepsilon\leqslant 1/2$.

利用引理 4 和 6 可得下述结论.

引理7 若存在 $\theta \in (0,1/2)$ 满足 $\frac{M^2}{m^\theta} \to 0$ 和式 (6), 则

$$\lim_{m\to\infty} \mathcal{E}_{\pi_M}(\pi_M f_{D,L}) - \mathcal{E}_{\pi_M,D}(\pi_M f_{D,L}) = 0$$

几乎处处成立.

证明 为方便起见, 记 $y_M = \pi_M y, y_{i,M} = \pi_M y_i$, 并定义

$$\mathcal{E}_{\pi_M}(f) = \int_Z (f(x) - y_M)^2 \mathrm{d}\rho,$$

$$\mathcal{E}_{\pi_M,D}(f) = \frac{1}{m}\sum_{i=1}^m (f(x_i) - y_{i,M})^2.$$

因为 $|\pi_M f_{D,L}|, |y_M|, |y_{i,M}| \leqslant M$, 所以

$$|\mathcal{E}_{\pi_M}(\pi_M f_{D,L}) - \mathcal{E}_{\pi_M}(f_\rho)| \leqslant 8M^2.$$

由引理 6 (令 $\alpha = \beta = 1, \varepsilon = m^{-\theta}$) 可得

$$\mathcal{E}_{\pi_M}(\pi_M f_{D,L}) - \mathcal{E}_{\pi_M}(f_\rho) - (\mathcal{E}_{\pi_M,D}(\pi_M f_{D,L}) - \mathcal{E}_{\pi_M,D}(f_\rho)) \leqslant (8M^2+2)m^{-\theta}$$

依概率

$$1 - 14 \max_{x_1^m \in X^m} \mathcal{N}_1\left(\frac{1}{20Mm^\theta}, \pi_M\mathcal{H}_{\sigma,L}, x_1^m\right) \exp\left(-\frac{m^{1-2\theta}}{428M^4}\right)$$

成立.

再结合引理 4 可得

$$\begin{aligned}
&\max_{x_1^m \in X^m} \mathcal{N}_1\left(\frac{1}{20Mm^\theta}, \pi_M\mathcal{H}_{\sigma,L}, x_1^m\right) \exp\left(-\frac{m^{1-2\theta}}{428M^4}\right) \\
&\quad\leqslant \exp\left(C_0^* L^2(d+1)(d+2)\log(L(d+1))\log(C_1^* M m^\theta) - \frac{m^{1-2\theta}}{428M^4}\right) \\
&\quad= \exp\left\{-\frac{m^{1-2\theta}}{M^4}\left(\frac{1}{428} - \frac{M^4(C_0^* L^2(d+1)(d+2)\log(L(d+1))\log(C_1^* M m^\theta))}{m^{1-2\theta}}\right)\right\}. \qquad (12)
\end{aligned}$$

从而由 $0 < \theta < \frac{1}{2}$, 式 (6) 和 (12), 可得

$$\lim_{m\to\infty} \max_{x_1^m \in X^m} \mathcal{N}_1\left(\frac{1}{20Mm^\theta}, \pi_M\mathcal{H}_{\sigma,L}, x_1^m\right) \exp\left(-\frac{m^{1-2\theta}}{428M^4}\right) = 0. \qquad (13)$$

因此, 由大数定理、式 (13) 及条件 $\frac{M^2}{m^\theta} \to 0$ 可知, 当 $m \to \infty$ 时, 下式几乎处处成立:

$$\mathcal{E}_{\pi_M}(\pi_M f_{D,L}) - \mathcal{E}_{\pi_M,D}(\pi_M f_{D,L}) \leqslant \frac{8M^2+2}{m^\theta} \to 0.$$

引理 5 证毕.

5.3.3 泛化误差分析

证明 [定理 1 的证明] 因为 $\int_Y y^2 \mathrm{d}\rho(y|x) < \infty$, 所以 $f_\rho \in L^2_{\rho_X} \subseteq C(X)$, 从而由引理 5 知, 对于任意的 $\varepsilon > 0$, 存在 $f_\varepsilon \in \mathcal{H}_{\sigma,L}$ 使得

$$\|f_\rho - f_\varepsilon\|^2_{L^2_{\rho_X}} \leqslant \varepsilon. \tag{14}$$

现将泛化误差分解为

$$\begin{aligned}
&\mathcal{E}(\pi_M f_{D,L_m}) - \mathcal{E}(f_\rho) \\
&\quad \leqslant \mathcal{E}(\pi_M f_{D,L_m}) - (1+\varepsilon)\mathcal{E}_{\pi_M}(\pi_M f_{D,L_m}) \\
&\qquad +(1+\varepsilon)\left(\mathcal{E}_{\pi_M}(\pi_M f_{D,L_m}) - \mathcal{E}_{\pi_M,D}(\pi_M f_{D,L_m})\right) \\
&\qquad +(1+\varepsilon)(\mathcal{E}_{\pi_M,D}(\pi_{M_m} f_{D,L_m}) - \mathcal{E}_{\pi_M,D}(f_{D,L_m})) \\
&\qquad +(1+\varepsilon)\mathcal{E}_{\pi_M,D}(f_{D,L_m}) - (1+\varepsilon)^2\mathcal{E}_D(f_{D,L_m}) \\
&\qquad +(1+\varepsilon)^2(\mathcal{E}_D(f_{D,L_m}) - \mathcal{E}_D(f_\varepsilon)) \\
&\qquad +(1+\varepsilon)^2(\mathcal{E}_D(f_\varepsilon) - \mathcal{E}(f_\varepsilon)) \\
&\qquad +(1+\varepsilon)^2(\mathcal{E}(f_\varepsilon) - \mathcal{E}(f_\rho)) \\
&\qquad +((1+\varepsilon)^2 - 1)\mathcal{E}(f_\rho) \\
&\quad =: \sum_{j=1}^{8} B_j.
\end{aligned} \tag{15}$$

对于 $\forall a, b > 0$, 由于 $(a+b)^2 \leqslant (1+\varepsilon)a^2 + (1+\frac{1}{\varepsilon})b^2$, 则根据强大数定理可得

$$\begin{aligned}
B_1 &= \mathcal{E}(\pi_M f_{D,L_m}) - (1+\varepsilon)\mathcal{E}_{\pi_M}(\pi_M f_{D,L_m}) \\
&= \int_Z |(\pi_M f_{D,L_m}(x) - y_M) + (y_M - y)|^2 \mathrm{d}\rho - (1+\varepsilon)\int_Z (\pi_M f_{D,L_m}(x) - y_M)^2 \mathrm{d}\rho \\
&\leqslant \left(1 + \frac{1}{\varepsilon}\right)\int_Z |y - y_M|^2 \mathrm{d}\rho,
\end{aligned}$$

$$\begin{aligned}
B_4 &= (1+\varepsilon)\frac{1}{m}\sum_{i=1}^{m} |(f_{D,L_m}(x_i) - y_i) + (y_i - y_{i,M})|^2 - (1+\varepsilon)^2 \frac{1}{m}\sum_{i=1}^{m} |f_{D,L_m}(x_i) - y_i|^2 \\
&\leqslant (1+\varepsilon)\left(1 + \frac{1}{\varepsilon}\right)\frac{1}{m}\sum_{i=1}^{m} |y_i - y_{i,M}|^2 \\
&\to (1+\varepsilon)\left(1 + \frac{1}{\varepsilon}\right)\int_Z |y - y_M|^2 \mathrm{d}\rho \qquad (m \to \infty).
\end{aligned}$$

由引理 7 可知, 当 $m \to \infty$ 时 $B_2 \to 0$ 几乎处处成立.

分别由截断算子和估计函数的定义, 可知 B_3 和 B_5 都是非正的. 当 $m \to \infty$ 时, 再次利用强大数定理可得 $B_6 \to 0$ 几乎处处成立.

最后, 由式 (14) 可得

$$B_7 = (1+\varepsilon)^2 \int_Z |f_\varepsilon(x) - f_\rho(x)|^2 \mathrm{d}\rho \leqslant (1+\varepsilon)^2 \varepsilon.$$

综上所有估计, 可推出

$$\begin{aligned}&\limsup_{m\to\infty}\mathcal{E}(\pi_M f_{D,L_m})-\mathcal{E}(f_\rho)\\&\quad\leqslant(2+\varepsilon)\left(1+\frac{1}{\varepsilon}\right)\int_Z|y-y_M|^2\mathrm{d}\rho+(1+\varepsilon)^2\varepsilon+(2\varepsilon+\varepsilon^2)\mathcal{E}(f_\rho)\end{aligned}$$

几乎处处成立. 由式 (5) 可知当 $m\to\infty$ 时, $M:=M_m\to\infty$, 即 $y_M\to y$, 故 $\limsup_{m\to\infty}\mathcal{E}(\pi_M f_{D,L_m})-\mathcal{E}(f_\rho)=0$. 换言之, $\{\pi_M f_{D,L_m}\}_{m=1}^{\infty}$ 是强万有一致的. 定理 1 证毕.

6 总结

浅层神经网络的优越性及瓶颈已被广泛研究, 加深神经网络一定程度上可以克服浅层神经网络的不足, 但已有理论结果只是部分性地证明深度神经网络的优势. 因此, 如何从理论的角度全面解释深度神经网络的深度、宽度、激活函数、连接方式等选择 (统称为结构选择) 及学习和逼近效果是一个核心的理论问题. 基于此, 本文首先给出一种具有统一结构的深度神经网络模型, 统一的结构是为了实现该模型不受先验信息的影响, 从而可以学习不同特征; 其次, 从逼近的角度, 论证该模型具有万有逼近性; 最后, 从学习理论的角度证明存在很多学习任务, 使得该模型具有良好的理论结果 —— 万有一致性. 上述 3 个论断分别从神经网络的结构选择、逼近、学习 3 个角度阐明了深度神经网络的优势.

参考文献

1 Zhou Z H. Machine Learning. Beijing: Tsinghua University Press, 2016 [周志华. 机器学习. 北京: 清华大学出版社, 2016]
2 Qiu X P. Neural Networks and Deep Learning. Beijing: China Machine Press, 2020 [邱锡鹏. 神经网络与深度学习. 北京: 机械工业出版社, 2020]
3 Deng L. Deep learning: methods and applications. FNT Signal Process, 2013, 7: 197–387
4 Schmidhuber J. Deep learning in neural networks: an overview. Neural Netw, 2015, 61: 85–117
5 Bengio Y, LeCun Y, Hinton G E. Deep learning. Nature, 2015, 521: 436–444
6 Jiao L C, Zhao J, Yang S Y, et al. Deep Learning, Optimization and Recognition. Beijing: Tsinghua University Press, 2017 [焦李成, 赵进, 杨淑媛, 等. 深度学习、优化与识别. 北京: 清华大学出版社, 2017]
7 Goodfellow I, Bengio Y, Courville A, et al. Deep Learning. Cambridge: MIT Press, 2016
8 Mhaskar H N. Neural networks for optimal approximation of smooth and analytic functions. Neural Comput, 1996, 8: 164–177
9 Guo Z C, Shi L, Lin S B. Realizing data features by deep nets. IEEE Trans Neural Netw Learn Syst, 2020, 31: 4036–4048
10 Györfi L, Kohler M, Krzyżak A, et al. A Distribution-Free Theory of Nonparametric Regression. Berlin: Springer, 2002
11 Chui C K, Li X, Mhaskar H N. Limitations of the approximation capabilities of neural networks with one hidden layer. Adv Comput Math, 1996, 5: 233–243
12 Lin S B. Limitations of shallow nets approximation. Neural Netw, 2017, 94: 96–102
13 Chui C K, Li X, Mhaskar H N. Neural networks for localized approximation. Math Comput, 1994, 63: 607
14 Eldan R, Shamir O. The power of depth for feedforward neural networks. In: Proceedings of the Conference on Learning Theory, 2016. 907–940
15 Mhaskar H N, Poggio T. Deep vs. shallow networks: an approximation theory perspective. Anal Appl, 2016, 14: 829–848
16 Raghu M, Poole B, Kleinberg J, et al. On the expressive power of deep neural networks. In: Proceedings of the 34th International Conference on Machine Learning, 2017. 70: 2847–2854

17 Telgarsky M. Benefits of depth in neural networks. In: Proceedings of the 29th Annual Conference on Learning Theory, 2016. 49: 1–23
18 Chui C K, Lin S B, Zhou D X. Deep neural networks for rotation-invariance approximation and learning. Anal Appl, 2019, 17: 737–772
19 Lin S B. Generalization and expressivity for deep Nets. IEEE Trans Neural Netw Learn Syst, 2019, 30: 1392–1406
20 Han Z, Yu S, Lin S B, et al. Depth selection for deep ReLU Nets in feature extraction and generalization. IEEE Trans Pattern Anal Mach Intell, 2022, 44: 1853–1868
21 Petersen P, Voigtlaender F. Optimal approximation of piecewise smooth functions using deep ReLU neural networks. Neural Netw, 2018, 108: 296–330
22 Yarotsky D. Error bounds for approximations with deep ReLU networks. Neural Netw, 2017, 94: 103–114
23 Lin S B, Wang K, Wang Y, et al. Universal consistency of deep convolutional neural networks. IEEE Trans Inform Theor, 2022, 68: 4610–4617
24 Liu X. Approximating smooth and sparse function by deep neural networks: optimal approximation rates and saturation. 2020. ArXiv:2001.04114
25 Chui C K, Lin S B, Zhou D X. Construction of neural networks for realization of localized deep learning. Front Appl Math Stat, 2018, 4: 1–11
26 Liu X. Learning and approximating piecewise smooth functions by deep sigmoid neural networks. Math Found Comput, 2023, doi: 10.3934/mfc.2023039
27 Cucker F, Smale S. On the mathematical foundations of learning. Bull Amer Math Soc, 2002, 39: 1–49
28 He K M, Zhang X Y, Ren S Q, et al. Delving deep into rectifiers: surpassing human-level performance on imagenet classification. In: Proceedings of the International Conference on Computer Vision (ICCV), Las Condes, 2015. 1026–1034
29 Lorentz G G, Golistschek M V, Makovoz Y. Constructive Approximation: Advanced Problems. Berlin: Springer, 1996
30 Hanin B. Universal function approximation by deep neural Nets with bounded width and ReLU activations. Mathematics, 2019, 7: 992
31 Hanin B, Sellke M. Approximating continuous functions by ReLU Nets of minimal width. 2017. ArXiv:1710.11278

Universal consistency of deep ReLU neural networks

Xia LIU[1] & Di WANG[2*]

1. *School of Science, Xi'an University of Technology, Xi'an* 710048*, China*;
2. *School of Management, Xi'an Jiaotong University, Xi'an* 710049*, China*
* Corresponding author. E-mail: wang.di@xjtu.edu.cn

Abstract With the explosive growth of data and richer computing resources, shallow neural networks can not always meet the requirements of the times, resulting in the emergence of deep neural networks. The rapid development of deep neural networks is mainly reflected in applications, and the theoretical research is relatively scarce. This paper focuses on the universal consistency of deep ReLU neural networks. The contents include: firstly, whether there is a deep neural network with a unified structure (i.e., the depth, width, and activation function have been determined) that can learn more features and has universal approximation; secondly, the determined deep neural network model has the property of universal consistency; finally, we verify the theoretical results from the perspective of experiments.

Keywords deep neural networks, universal consistency, deep learning, ReLU function, approximation