



神经网络的逼近理论: 从经典到高维

献给四川大学数学学科建设 130 周年

明平兵^{1,2*}, 于灏^{1,2}

1. 中国科学院数学与系统科学研究院, 数学科学全国重点实验室, 北京 100190;

2. 中国科学院大学数学科学学院, 北京 100049

E-mail: mpb@lsec.cc.ac.cn, yuhao@amss.ac.cn

收稿日期: 2026-03-31; 接受日期: 2026-05-12; 网络出版日期: 2026-08-20; * 通信作者

国家自然科学基金 (批准号: 12371438) 资助项目

摘要 神经网络在高维问题中的成功应用, 使其逼近理论成为近年来逼近论与机器学习交叉领域的重要研究方向. 本文系统综述神经网络逼近问题的最新理论进展. 一方面, 梳理神经网络的万有逼近定理以及在经典函数空间中的定量逼近估计等传统逼近理论; 另一方面, 聚焦于维数灾难问题, 深入分析高维逼近所面临的核心挑战, 并详细介绍神经网络在高维情形下逼近理论的最新突破与发展趋势.

关键词 神经网络 万有逼近 高维逼近 Barron 函数类 深度分离

MSC (2020) 主题分类 41-02, 41A46, 46E35, 65-02, 68T99

1 引言

逼近理论是机器学习理论研究的基础. 函数逼近论起源甚早, 发展已相对成熟, 其成果主要涵盖两个方面. 其一是正逼近定理与逆逼近定理: 正逼近定理 (Jackson 定理) 表明, 若目标函数具备某种光滑性, 则可用特定逼近工具 (如多项式) 对其进行逼近, 且逼近误差具有确定的衰减率; 一般而言, 函数光滑性越强, 逼近速度越快. 逆逼近定理 (Bernstein 定理) 则指出, 若逼近误差以某一速率趋于零, 则可反推目标函数具有相应的光滑性 (或至少存在一个几乎处处等于原函数的函数具有该光滑性). 其二是不同逼近算法的构造与误差估计. 经典函数逼近理论的一个核心结论是, 逼近算法的收敛速度取决于算法的精度与目标函数的光滑性.

与经典逼近论不同, 机器学习的研究重心聚焦于高维函数. 机器学习在计算机视觉等领域的突破, 本质上归功于对高维函数逼近问题的有效处理. 近年来, 深度学习算法已进一步拓展至科学计算领域, 如高维偏微分方程及变分问题的求解、复杂系统的控制问题以及金融期权定价估计, 并取得了令人瞩目的成果. 这些实践均印证了一个关键结论: 神经网络堪称逼近高维函数的利器. 然而, 从经典逼近论的视角审视, 基于神经网络的机器学习最令人困惑的问题在于: 为何神经网络逼近高维函数会如此有效? 经典逼近论的核心研究多集中于基于多项式的算法, 其分析所依托的函数空间主要是 Sobolev 空

英文引用格式: Ming P B, Yu H. Approximation theory of neural networks: From classical theory to high dimensions (in Chinese). Sci Sin Math, 2026, 56: 2019–2058, doi: 10.1360/SSM-2026-0097

间或 Besov 空间, 但这类理论结果均受困于“维数灾难”(curse of dimensionality)—收敛速率随维数增加而衰减, 典型表现为以 $1/d$ 的速度下降, 其中 d 为维数. 需注意, 此处的维数未必是物理空间的维数, 很可能是参数空间的维数. 因此, 以多项式、分片多项式、样条函数、小波函数等为基本逼近工具的传统方法, 在处理高维问题时往往力不从心. 这一现象具有普遍性, 也是科学与工程领域长期面临的基础性难题 (参见文献 [17]).

不少机器学习研究者试图用“万有逼近定理”(universal approximation theorem) 来解释神经网络的这种魔力. 所谓万有逼近定理是指, 如果神经元数量足够多, 网络结构合适, 则神经网络能够在紧集上逼近任意连续函数. 但这种解释并不合理, 因为多项式也有类似的万有逼近定理, 这正是经典的 Stone-Weierstrass 定理, 但多项式恰恰因“维数灾难”而无法有效逼近高维函数.

要解释神经网络逼近高维函数为何如此有效, 亟需构建一套专门针对高维函数的新型逼近理论. 其核心在于: 针对某一具体逼近模型 (如某类神经网络), 需首先明确其对应的函数空间, 即该模型能有效逼近的函数类别, 并在此基础上分析逼近误差与泛化误差. 然而, 寻找这样的函数空间极为困难, 目前尚无通用方法, 本质上这等同于逼近论中的反问题研究, 但这一探索又是不可或缺的.

本文旨在总结近年来发展迅速的神经网络逼近理论. 一般而言, 逼近理论需解决的关键问题是针对特定逼近方法 (如分片多项式或神经网络模型), 明确其对应的函数空间, 即同时满足正逼近定理与逆逼近定理成立的函数类. 对于深度神经网络模型, 我们尤其期望获得与维数无关的收敛阶, 而这有赖于建立全新的函数空间理论. 研究表明, 有别于以光滑性为核心的传统函数空间 (如 Hölder, Sobolev 及 Besov 空间), 此类新空间不再依赖于函数的正则性, 而是基于其在某种数学期望下的可表示性. 目前, 针对随机特征模型、两层神经网络及残差网络, 已建立起相对完整的逼近理论; 对于深层神经网络, 虽已取得初步成果, 但远未臻完善. 这些理论引出了如再生核 Hilbert 空间 (RKHS) 和 Barron 空间等新型函数空间, 但我们对其缺乏系统认识. 因此, 深入研究这些空间的性质以及它们与经典函数空间的关系, 成为一个重要课题 — 事实上, 这已被文献 [58] 列为一个公开问题. 对于前述其他问题, 目前所知甚少, 相关研究尚处起步阶段. 尤为值得一提的是, 谱 Barron 空间作为一类特例, 目前已获得相当深刻的理解.

另一方面, 目前仍有不少学者致力于研究基于传统光滑性假设的神经网络逼近理论, 这类结果通常受限于维数灾难, 但对于低维问题仍具有重要价值. 本文也将专辟一节对此进行讨论.

目前神经网络的逼近理论中, 一个让人困惑的问题是 Kolmogorov 定理^[87] 对神经网络的实际应用. 该定理断言, 任一连续函数均可由三层神经网络精确表示, 但其最外层激活函数异常复杂, 且目前缺乏可行的构造方法. 尽管该理论几经沉浮 (参见文献 [102, 131]), 但其在实际应用中的意义仍有待进一步揭示.

本文余下内容的结构安排如下. 第 2 节简要介绍神经网络的基本概念. 第 3 节讨论神经网络的古典逼近理论, 包括万有逼近定理、Kolmogorov 定理以及对于经典目标函数的逼近估计. 第 4 节介绍一些与神经网络相关的最优逼近与逼近下界, 揭示诸多情形下高维逼近面临的本质困难. 第 5 和 6 节探讨新型目标空间尤其是 Barron 型空间中函数的性质, 以及浅层、深层网络对 Barron 函数的逼近理论. 第 7 节介绍深度分离现象, 展示深度蕴藏的强大表示能力. 第 8 节讨论多种或特定激活函数实现指数收敛的结果, 展现极高效的信息组织能力. 第 9 节介绍 Transformer 的逼近问题. 最后一节对神经网络的逼近理论进行总结并提出了若干公开问题¹⁾.

后文中, $\mathbb{N}$, $\mathbb{N}_+$ 和 $\mathbb{R}$ 分别表示自然数集、正整数集和实数集. 自然数 d 表示维数. 对任何实数 s ,

1) 受篇幅所限, 本文未能尽述该领域的众多文献. 对于未能一一引用的研究成果, 我们深表歉意. 此外, 我们侧重评述已正式发表的工作, 对部分最新预印本仅作选择性介绍.

令 $s_+ = \max(s, 0)$. L^p 和 ℓ^p 分别表示 p 次 Lebesgue 可积函数空间、 p 次绝对可和的离散序列空间. 在渐近估计中, $b \lesssim a$ 和 $b \gtrsim a$ 分别表示存在与 a 无关的常数 $C, c > 0$, 使得 $|b| \leq Ca$ 和 $|b| \geq ca$; $b \simeq a$ 表示同时有 $b \lesssim a$ 和 $b \gtrsim a$; $b = \mathcal{O}(a)$ 等价于 $b \lesssim a$. 此外, 记 $\mathcal{S}(\mathbb{R}^d)$ 为 Schwartz 空间, $\mathcal{S}'(\mathbb{R}^d)$ 为 $\mathcal{S}(\mathbb{R}^d)$ 的拓扑对偶空间, 即缓增分布空间. 若 $f \in \mathcal{S}'(\mathbb{R}^d)$, 则 $\hat{f}$ 表示 f 在分布意义下的 Fourier 变换. $\mathcal{D}'(\Omega)$ 表示有界区域 Ω 的分布空间.

2 神经网络模型与基本定义

深度神经网络可视为由大量神经元按层级方式组合而成的参数化函数族. 单个神经元定义为形如 $x \mapsto \sigma(b \cdot x + c)$ 的函数, 其中 $x \in \mathbb{R}^d$ 为自变量, σ 为激活函数, $b \in \mathbb{R}^d$ 为权重向量, $c \in \mathbb{R}$ 为偏置. 该形式可追溯至 McCulloch 和 Pitts^[121] 的经典工作, 常被视为生物神经元的数学抽象. 作为直观例子, 若取 σ 为 Heaviside 函数 $\mathbf{1}_{[0, \infty)}$, 则当 $b \cdot x$ 超过阈值 $-c$ 时, 神经元被“激活”. 下面给出深层神经网络的一个严格定义.

定义 2.1 设 $L \in \mathbb{N}_+$, 维度序列 $\{d_\ell\}_{\ell=0}^{L+1} \subset \mathbb{N}_+$, 定义在实轴上的激活函数 σ . 令权重矩阵 $W^{(\ell)} \in \mathbb{R}^{d_{\ell+1} \times d_\ell}$ 和偏置向量 $b^{(\ell)} \in \mathbb{R}^{d_{\ell+1}}$, $0 \leq \ell \leq L$. 对任意输入 $x \in \mathbb{R}^{d_0}$, 定义层变量递推关系:

$$\begin{aligned} x^{(0)} &= x, \\ x^{(\ell)} &= \sigma(W^{(\ell-1)}x^{(\ell-1)} + b^{(\ell-1)}), \quad \ell = 1, \dots, L, \\ x^{(L+1)} &= W^{(L)}x^{(L)} + b^{(L)}, \end{aligned}$$

其中 σ 逐点作用于向量, 则称函数 $\Phi(x) = x^{(L+1)}$ 所定义的映射 $\Phi: \mathbb{R}^{d_0} \rightarrow \mathbb{R}^{d_{L+1}}$ 为一个前馈神经网络, 其中 L 为网络深度, $d_{\max} := \max_{1 \leq \ell \leq L} d_\ell$ 为网络宽度, 称 $(\sigma; d_0, \dots, d_{L+1})$ 为网络 Φ 的结构.

深层神经网络常被视为一个有向无环连接图, 其中节点是神经元, $x^{(0)}$ 为输入层, $x^{(L+1)}$ 为输出层, 其他层为中间隐藏层. 隐藏层的个数就是深度, 也是激活函数作用的次数. $L > 1$ 时网络被称为深层网络, 而 $L = 1$ 时为浅层网络. 本文主要考虑上述前馈网络, 需要指出的是, 神经网络还包含多种重要变体 (如循环神经网络等), 而在实际应用中网络结构也常作如下扩展.

(1) 异构激活函数: 不同层 (甚至不同神经元) 可采用不同的激活函数 (参见第 8 节). 为降低模型复杂度并提升实现与训练的统一性, 实践中常在同一网络内选用单一激活函数.

(2) 跨层连接: 在标准前馈神经网络中, 第 ℓ 层仅依赖第 $\ell - 1$ 层的输出; 而在残差网络等结构中, 跨层跳连使第 ℓ 层可利用更早层的信息, 这在训练深层模型时有助于缓解梯度消失等优化困难 (参见文献 [69]).

激活函数是神经网络的核心部分, 它为神经网络引入了非线性, 这对于神经网络的应用至关重要. 下面列举一些最常用的神经网络激活函数及其简要性质. 关于激活函数的一个比较详尽的综述可参见文献 [49]. 历史最悠久的激活函数之一是 Sigmoidal 函数 $\text{Sigmoid}(x) = 1/(1 + e^{-x})$, 它的函数值介于 0 和 1 之间且处处光滑, 其主要局限在于当 $|x|$ 较大时导数趋近于 0, 训练中容易出现梯度消失, 减缓深层网络的优化. 进一步地, Sigmoidal 激活函数常指一类满足在正无穷处极限为 1 且在负无穷处极限为 0 的有界可测函数. ReLU (rectified linear unit) 函数 $\sigma_1(x) = \max(x, 0)$ 是目前应用最广泛的激活函数之一, 其具有分片线性、计算代价低及缓解梯度消失等优势, 在工程实践中计算效率较高. 然而, 在实际训练中, ReLU 还存在“死神经元”现象, 即当输入为负时梯度为零, 导致部分隐藏单元在训练中无法接收到有效的误差信号, 从而停止更新. 尽管如此, ReLU 的引入相较于 Sigmoid 在训练效率上仍实现了显著提升, 尤其适合低精度计算场景, 因此它在许多实际应用中常被作为默认的激活函数.

然而, ReLU 并不适用于对精度要求较高的任务, 尤其是在科学与工程计算领域. 一个较为简单且有效的替代方案是采用 RePU 函数, 即 ReLU 的整数幂次推广 ReLU^k ²⁾:

$$\sigma_k(x) = \max(x, 0)^k, \quad k \in \mathbb{N}.$$

为解决 ReLU 存在的死神经元的问题, 人们引入了带泄漏的 (leaky) ReLU 激活函数 $\text{LReLU}(x) = \max(x, ax)$, 在应用中需要针对不同的问题调节参数 $a \in (0, 1)$. ReLU 的另一个局限是不光滑, 在 $x = 0$ 处没有古典导数, 为此, 人们引入了许多 ReLU 的光滑化版本, 如 SiLU 函数 $\text{SiLU}(x) = x/(1 + e^{-x})$ 、Softplus 函数 $\text{Softplus}(x) = \log(1 + e^x)$ 、Mish 函数 $\text{Mish}(x) = x \tanh(\log(1 + e^x))$ 和 Gauss 误差线性单元 (Gaussian error linear unit, GELU) 等. 以上介绍的是一些常规激活函数, 还有一些非常规的激活函数, 例如使用有理函数作为激活函数在某些任务中优于 ReLU (参见文献 [21, 170]).

值得指出的是, 神经网络逼近结果在某些情形下可在不同激活函数之间迁移. 鉴于 ReLU 激活函数的普及性或为了简化证明, 有关 ReLU 网络的逼近结果最为常见. 若能用某个一般激活函数的固定大小的网络在有限区间上以任意精度逼近 ReLU, 则 ReLU 网络在有界区域上的逼近上界可推广为该激活函数的逼近界. 文献 [192] 证明了一个宽度为 N 、深度为 L 的 ReLU 网络, 在有界集上可被一个宽度为 $3N$ 、深度为 $2L$ 的某类激活函数的网络以任意精度逼近, 该类包含了如 ReLU, LeakyReLU, ReLU^2 , ELU, CELU, SELU, Softplus, GELU, SiLU, Swish, Mish, Sigmoid, Tanh, Arctan, Softsign, dSiLU 和 SRS 等常用激活函数. 甚至对其中 ELU 到 Mish 这八种激活函数, 相同宽度和深度的网络即可做到任意精度逼近. 另外, 若激活函数为连续的分段多项式且不恒等于一个整体多项式, 如 RePU, 则利用有限差分即可实现任意精度逼近 ReLU. 当 $t \rightarrow \infty$ 时, 采用 Sigmoidal 激活的神经元 $\sigma(t(\omega \cdot x - b))$ 可在 $L^p(\Omega)$ 范数下以任意精度逼近采用 Heaviside 激活的神经元 $\sigma_0(\omega \cdot x - b)$, 其中 $p < \infty$. 后文中我们通常聚焦基于典型激活函数的逼近结果, 感兴趣的读者可以将它们推广到更一般的激活函数.

3 神经网络的古典逼近理论

3.1 万有逼近定理

本小节展示神经网络的万有逼近性, 即若神经元数量足够多, 网络结构合适, 则前馈神经网络可以在紧集上以任意精度一致逼近任意连续函数. 下面先给出万有逼近子的定义:

定义 3.1 设 $\mathcal{H}$ 是一个定义在 $\mathbb{R}^d$ 的实值函数集合. 对于任意 $\varepsilon > 0$, 任意紧集 $K \subset \mathbb{R}^d$, 以及任意 $f \in C(\mathbb{R}^d)$, 若存在 $\Phi \in \mathcal{H}$ 使得

$$\sup_{x \in K} |f(x) - \Phi(x)| < \varepsilon, \quad (3.1)$$

则称 $\mathcal{H}$ 是 $C(\mathbb{R}^d)$ 的一个万有逼近子 (universal approximator).

例 3.1 由 Stone-Weierstrass 定理可得, 全体多项式集合是 $C(\mathbb{R}^d)$ 的一个万有逼近子.

给定自然数 d, m, L 和 n , 令 σ 为定义在实轴上的激活函数, 记神经网络函数集合

$$\mathcal{N}_d^m(\sigma; L, n) := \{\Phi : \mathbb{R}^d \rightarrow \mathbb{R}^m \mid \Phi \text{ 如定义 2.1, 其深度不超过 } L, \text{ 宽度不超过 } n\},$$

并记 $\mathcal{N}_d^m(\sigma; L) = \bigcup_{n \in \mathbb{N}_+} \mathcal{N}_d^m(\sigma; L, n)$. 后文中, 在不引起误会的情况下, 有时会略去 $\mathcal{N}_d^m(\sigma; L)$ 和 $\mathcal{N}_d^m(\sigma; L, n)$ 的上下标.

2) 特别地, 约定 $\sigma_0(x) = \mathbf{1}_{\{x \geq 0\}}$, 即 Heaviside 阶跃函数.

定理 3.1 (浅层神经网络的万有逼近定理^[95]) 记

$$\mathcal{M} := \{\sigma \in L_{\text{loc}}^{\infty}(\mathbb{R}) \mid \text{存在实数轴的划分 } I_1, \dots, I_m \text{ 使得 } \sigma \in C(I_j), j = 1, \dots, m\}.$$

若 $\sigma \in \mathcal{M}$, 则 $\mathcal{N}_d^1(\sigma; 1)$ 是 $C(\mathbb{R}^d)$ 的一个万有逼近子当且仅当 σ 不是一个多项式.

万有逼近定理有诸多版本, 其差异主要体现在激活函数的假设, 最早的版本是 Vostretsov 和 Kreines^[175] 证明的. Cybenko^[36] 证明了 Sigmoidal 激活函数的万有逼近定理. 定理 3.1 属于 Leshno 等^[95]. 注意到集合 $\mathcal{M}$ 中的函数分段连续, 且允许存在有限多个间断点. 所以定理 3.1 基本包含了上一节介绍的所有激活函数, 如 ReLU, SiLU 和 Sigmoid, 甚至 Heaviside 这种典型间断激活函数. 以下定理表明深层网络同样具有万有逼近性:

定理 3.2 (深层神经网络的万有逼近定理) 设 $d, L \in \mathbb{N}_+$. 若 $\sigma \in \mathcal{M}$, 则 $\mathcal{N}_d^1(\sigma; L)$ 是 $C(\mathbb{R}^d)$ 的一个万有逼近子当且仅当 σ 不是一个多项式.

这个定理的证明依赖定理 3.1 以及恒等映射可被 $\mathcal{N}_d^d(\sigma; L-1)$ 中的函数逼近得充分好这一简单事实. 实际上, 数值逼近中使用的所有函数族, 如多项式、样条和小波等, 都能产生万有逼近子. 万有逼近性并不是神经网络在实践中如此成功的主要原因. 我们需要回答的问题是作为一种逼近工具, 神经网络究竟在哪些方面比其他方法更加有效?

上述两个万有逼近定理适用于宽度任意大、深度有界的神经网络. 事实上, 在宽度有界、深度任意大这一自然“对偶”情形下也有类似的万有逼近定理. 具体而言, 设输入和输出维数为 d 和 m , σ 为任意非仿射连续函数, 且在某点处具有非零的连续导数. 文献 [84] 证明了对任意紧集 $K \subset \mathbb{R}^d$, 所有深度任意大、宽度为 $d+m+2$ 、激活函数为 σ 的神经网络构成的函数类, 在空间 $C(K; \mathbb{R}^m)$ 中稠密. 这一结果不但涵盖了实践中使用的激活函数, 同时也包含多项式激活函数—这与上述经典版本的万有逼近定理 (定理 3.2) 截然不同. 这个定理还揭示了深层窄网络与浅层宽网络之间的一个本质差异. 关于其他范数意义下的万有逼近定理以及万有逼近定理成立的最小宽度的定量估计, 参见文献 [25, 108, 143, 173].

万有逼近定理表明神经网络可以在紧集上以任意精度逼近连续函数, 在应用中, 一个自然的问题是, 神经网络能否实现精确插值, 即给定激活函数 σ , 以及 n 个数据 $\{(x_j, y_j)\}_{j=1}^n \subset \mathbb{R}^d \times \mathbb{R}$, 是否存在 $\{a_i\}_{i=1}^m \subset \mathbb{R}$, $\{w_i\}_{i=1}^m \subset \mathbb{R}^d$ 和 $\{b_i\}_{i=1}^m \subset \mathbb{R}$ 使得

$$\sum_{i=1}^m a_i \sigma(w_i \cdot x_j + b_i) = y_j, \quad j = 1, \dots, n \quad (3.2)$$

成立? 此外, 样本数 n 与宽度 m 有没有关系? 对大量非多项式激活函数, 答案是肯定的, 证明也和万有逼近定理紧密相关. Ito 和 Saito^[76] 证明了对连续非减的 Sigmoidal 激活函数, 当 $n = m$ 时浅层网络能够实现插值, 且 $\{w_i\}_{i=1}^m \subset \mathbb{S}^{d-1}$. Huang 和 Babri^[73] 对有界连续且在无穷远处有极限的激活函数证明了类似的插值结果, 但不保证 $\{w_i\}_{i=1}^m$ 在单位球面上. 下面给出的插值结果本质上属于 Ito 和 Saito³⁾:

定理 3.3 若 $m = n$, 且 $\sigma \in C(\mathbb{R})$ 不是一个多项式. 则对任意 n 个数据 $\{(x_j, y_j)\}_{j=1}^n \subset \mathbb{R}^d \times \mathbb{R}$, 存在 $\{a_i\}_{i=1}^n, \{b_i\}_{i=1}^n \subset \mathbb{R}$ 和 $\{w_i\}_{i=1}^n \subset \mathbb{R}^d$ 使得插值条件 (3.2) 成立.

若上面定理中的 σ 是一个多项式, 则能不能插值取决于点集 $\{x_j\}_{j=1}^n$ 的选取以及多项式的次数. 若 σ 是一个 r 次多项式, 则

$$\text{span}\{\sigma(w \cdot x - b) \mid w \in \mathbb{S}^{d-1}, b \in \mathbb{R}\}$$

3) 为避免引入更多的概念, 我们没有给出保证 $\{w_i\}_{i=1}^m \subset \mathbb{S}^{d-1}$ 的条件.

是 r 次多项式集合. 古典意义下神经网络的插值问题可参见文献 [149, 第 11 和 12 节]. 而现代机器学习意义下插值方面的相关进展可参见综述文献 [15].

3.2 Kolmogorov 表示定理

我们在第 3.1 小节看到多种激活函数下的神经网络均具有万有逼近性, 本小节探讨激活函数对逼近精度的具体影响. 下面的命题表明, 存在一种特殊激活函数, 当输入为一维时, 单个神经元就能达到任意精度 (参见文献 [145, 命题 3.20]). 这种特殊激活函数的思想可追溯至文献 [113]. 一个相关的工作可参见文献 [64, 定理 2.1].

命题 3.1 存在激活函数 $\sigma \in C(\mathbb{R})$ 使得对任意紧集 $K \subset \mathbb{R}$ 、任意 $\varepsilon > 0$ 以及任意函数 $f \in C(K)$, 存在神经网络 $\Phi(x) = \sigma(wx + b) \in \mathcal{N}(\sigma; 1, 1)$ 满足一致逼近 (3.1).

为了将上述一维结果推广到高维, 下面介绍 Kolmogorov 定理 [87, 104].

定理 3.4 (Kolmogorov 定理) 设整数 $d \geq 2$. 则存在正数 $\{\lambda_j\}_{j=1}^d$ 满足 $\sum_{j=1}^d \lambda_j \leq 1$ 和严格单增的连续函数 $\phi_k : [0, 1] \rightarrow [0, 1]$, $1 \leq k \leq 2d + 1$, 使得对任意 $f \in C([0, 1]^d)$, 存在 $g \in C(\mathbb{R})$ 满足

$$f(x) = \sum_{k=1}^{2d+1} g(\lambda_1 \phi_k(x_1) + \cdots + \lambda_d \phi_k(x_d)), \quad x \in [0, 1]^d.$$

- 注 3.1** (i) 定理 3.4 中参数 $\{\lambda_j\}_{j=1}^d$ 和函数 $\{\phi_k\}_{k=1}^{2d+1}$ 均与目标函数 f 无关且具有显式表达式;
(ii) 内层激活函数 ϕ_k 的个数 $2d + 1$ 是最优的, 不能减少到 $2d$;
(iii) 外层激活函数 g 依赖 f 且目前 g 的显式表达式未知.

Kolmogorov 定理否定地回答了 Hilbert 第 13 问题 [131]: 是否存在一个三元连续函数 $f : [0, 1]^3 \rightarrow \mathbb{R}$, 它不能被二元函数以函数复合和加法的运算来表示. 注意到 Kolmogorov 定理将 Hilbert 第 13 问题推广到了任意维数的情形. 关于该定理的证明以及 Hilbert 第 13 问题的历史渊源, 可参见文献 [131]. 作为 Kolmogorov 定理的一个应用, 我们把命题 3.1 推广到高维.

定理 3.5 设命题 3.1 中的激活函数为 σ . 则对任意紧集 $K \subset \mathbb{R}^d$ 、任意 $\varepsilon > 0$ 以及任意 $f \in C(K)$, 存在神经网络 $\Phi \in \mathcal{N}_d^1(\sigma; 2, 2d^2 + d)$ 满足 (3.1).

定理 3.5 可视为 Kolmogorov 定理的神经网络实现, 它表明确实存在一种“超级”激活函数, 利用它构建的神经网络可在宽度为 $\mathcal{O}(d^2)$ 的条件下, 实现对连续函数的任意精度逼近. 然而, 目前对这种神奇激活函数的认识依然十分有限, 文献 [131] 没有提供这种激活函数的显式表达式, 相关研究进展可参见文献 [186]. 定理 3.5 与常见逼近定理不同之处在于, 神经元个数仅依赖输入维数 d , 而与逼近精度 ε 无关. 其证明的关键之处在于 Kolmogorov 定理中的超级激活函数的存在性.

遗憾的是, 我们还是无法找到这样的激活函数的显式表达式. 这方面的进展可参见文献 [157, 158, 186]. 若略微放宽网络深度和宽度要求, 文献 [158] 构造了由三角波函数与 softsign 函数组合成的显式激活函数

$$\begin{aligned} \sigma(x) &:= \left| x - 2 \left\lfloor \frac{x+1}{2} \right\rfloor \right|, \quad \text{当 } x \geq 0, \\ \sigma(x) &:= \frac{x}{|x|+1}, \quad \text{当 } x < 0, \end{aligned} \tag{3.3}$$

其图示参见文献 [158, 图 1]; 并证明了 $\mathcal{N}_d^1(\sigma; 11, 36d(2d+1))$ 是 $C(\mathbb{R}^d)$ 的万有逼近子.

事实上, 我们也可以通过修改网络架构来实现“用极少的神经元实现万能逼近”. 文献 [100] 巧妙地利用了残差网络 (ResNet) 的恒等映射结构, 通过构建一个极深的残差网络, 即使每一层隐藏层只有

一个神经元 (使用 ReLU 激活函数), 整个网络也可以实现对任意 Lebesgue 可积函数的逼近. 这打破了传统全连接网络在每层宽度小于输入维度时无法成为万能逼近器的限制.

3.3 经典函数空间中神经网络的逼近理论

第 3.1 小节中万有逼近定理说明神经网络在紧集上能以任意精度一致逼近任意连续函数, 但并未刻画达到某一精度所需要的网络规模, 或给定网络规模时的误差上界, 即逼近阶. 在第 3.2 小节中, 我们看到对某些特定激活函数, $\mathcal{O}(d)$ 个神经元的网络即可实现任意精度逼近. 本小节介绍神经网络在经典函数空间中的逼近阶. 为此, 先定义三种常用的经典函数空间.

定义 3.2 设 $\Omega \subset \mathbb{R}^d$, $s > 0$, $k = \lfloor s \rfloor$, $\gamma = s - k$, 定义 Hölder 空间 $C^s(\Omega)$ 和相应范数分别为

$$C^s(\Omega) := \{f \in C^k(\Omega) \mid \|f\|_{C^s(\Omega)} < \infty\},$$

$$\|f\|_{C^s(\Omega)} := \sum_{|\alpha|_1 \leq k} \sup_{x \in \Omega} |\nabla^\alpha f(x)| + \sum_{|\alpha|_1 = k} \sup_{x \neq y} \frac{|\nabla^\alpha f(x) - \nabla^\alpha f(y)|}{|x - y|^\gamma} \mathbf{1}_{\{s \notin \mathbb{N}_+\}}.$$

定义 3.3 设 $\Omega \subset \mathbb{R}^d$, $s > 0$, $k = \lfloor s \rfloor$, $1 \leq p < \infty$. 定义 Sobolev 空间 $W^{s,p}(\Omega)$ 和相应范数

$$W^{s,p}(\Omega) := \{f \in L^p(\Omega) \mid \|f\|_{W^{s,p}(\Omega)} < \infty\},$$

$$\|f\|_{W^{s,p}(\Omega)} := \sum_{|\alpha|_1 \leq k} \|\nabla^\alpha f\|_{L^p(\Omega)} + \left(\sum_{|\alpha|_1 = k} \int_{\Omega} \int_{\Omega} \frac{|\nabla^\alpha f(x) - \nabla^\alpha f(y)|^p}{|x - y|^{d+(s-k)p}} dx dy \right)^{1/p} \mathbf{1}_{\{s \notin \mathbb{N}_+\}},$$

其中 $\nabla^\alpha f$ 为分布导数. 依照惯例记 $H^s(\Omega) = W^{s,2}(\Omega)$.

定义 3.4 设 $\{\varphi_j\}_{j=0}^\infty \subset \mathcal{S}(\mathbb{R}^d)$ 为一个单位分解, 满足 $\sum_{j=0}^\infty \varphi_j(x) \equiv 1$, $0 \leq \varphi_j \leq 1$,

$$\text{supp}(\varphi_j) \subset \{x \in \mathbb{R}^d \mid 2^{j-1} \mathbf{1}_{\{j \geq 1\}} \leq |x| \leq 2^{j+1}\}, \quad j \in \mathbb{N}.$$

并且对任意 $\alpha \in \mathbb{N}^d$, 存在常数 c_α 使得对任意 $j \in \mathbb{N}$ 和 $x \in \mathbb{R}^d$, 有 $2^{j|\alpha|} |\nabla^\alpha \varphi_j(x)| \leq c_\alpha$, 设 $s \in \mathbb{R}$, $1 \leq p, q \leq \infty$. 定义 Besov 空间

$$B_{p,q}^s(\mathbb{R}^d) := \{f \in \mathcal{S}'(\mathbb{R}^d) \mid \|f\|_{B_{p,q}^s(\mathbb{R}^d)} := \|2^{sj} \|(\varphi_j \hat{f})^\vee\|_{L^p(\mathbb{R}^d)}\|_{\ell^q(\mathbb{N})} < \infty\}.$$

有界区域 Ω 上的 Besov 空间可由限制定义 $B_{p,q}^s(\Omega) := \{f \in \mathcal{D}'(\Omega) \mid \|f\|_{B_{p,q}^s(\Omega)} < \infty\}$, 其中

$$\|f\|_{B_{p,q}^s(\Omega)} := \inf\{\|F\|_{B_{p,q}^s(\mathbb{R}^d)} : F \in B_{p,q}^s(\mathbb{R}^d), F|_\Omega = f\}.$$

Besov 空间 $B_{p,q}^s$ 是一类用于刻画函数光滑性、可积性和多尺度振荡行为的经典函数空间, 指标 s, p 和 q 分别表示光滑性、可积性和微调参数. 它与 Hölder 空间、Sobolev 空间联系紧密, 且提供了更细致的光滑性刻画. 当 Ω 是有界 Lipschitz 区域且 s 非整数时, Hölder 空间 C^s 等同于 $B_{\infty,\infty}^s$, 而 Sobolev 空间 $W^{s,p}$ 等同于 $B_{p,p}^s$. 有关 Besov 空间之间的嵌入关系参见文献 [171].

下面展示 B 样条函数在经典函数空间中的逼近.

定义 3.5 (B 样条函数) 设 $d, n \in \mathbb{N}_+$, $\ell \in \mathbb{N}$. 定义 n 阶一元 B 样条函数

$$B_n(x) := \frac{1}{(n-1)!} \sum_{k=0}^n (-1)^k \binom{n}{k} (x-k)_+^{n-1}, \quad x \in \mathbb{R}.$$

对 $t, x \in \mathbb{R}$, 记 $B_{\ell,t,n}(x) := B_n(2^\ell(x-t))$. 对 $t = (t_1, \dots, t_d) \in \mathbb{R}^d$, 定义 d 元张量积 B 样条函数

$$B_{\ell,t,n}^d(x) := \prod_{j=1}^d B_{\ell,t_j,n}(x_j), \quad x = (x_1, \dots, x_d) \in \mathbb{R}^d.$$

n 阶 B 样条字典 $\mathcal{B}^n := \{B_{\ell,t,n}^d \mid \ell \in \mathbb{N}, t \in \mathbb{R}^d\}$ 中函数的线性叠加可以逼近光滑函数, 这种逼近已有非常详尽的研究 (参见文献 [44, 135]). 以 Besov 函数为例, 收敛阶方面的经典结果如下:

定理 3.6 设 $d, n \in \mathbb{N}_+$, $\Omega = [0, 1]^d$, $0 < p < p' \leq \infty$, $0 < q, q' \leq \infty$, 且 $0 < s \leq n + \max(0, 1/p - 1)$ (其中等号仅当 $q = \infty$ 时允许成立), 以及 $0 < s' < \min(s - d(1/p - 1/p'), n + 1/p' - 1)$. 则对任意 $f \in B_{p,q}^s(\Omega)$ 和正整数 N , 存在 $\{B_i\}_{i=1}^N \subset \mathcal{B}^n$ 和常数 $\{c_i\}_{i=1}^N$ 使得

$$\left\| f - \sum_{i=1}^N c_i B_i \right\|_{B_{p',q'}^{s'}(\Omega)} \leq C N^{-(s-s')/d} \|f\|_{B_{p,q}^s(\Omega)},$$

其中常数 C 与 f 和 N 无关.

若 $s' = 0$, $p' = q'$, 则上述估计右端替换为 $C N^{-s/d} \|f\|_{B_{p,\infty}^s(\Omega)}$.

上述估计给出的 L^p -收敛阶为 $\mathcal{O}(N^{-s/d})$. 下一节将看到这已经是某种意义下 Besov 空间中的最优逼近率. 若想获得逼近精度 ε , 则需要逼近项数 N 达到 $\varepsilon^{-d/s}$ 量级, 随维数 d 指数增长, 这种现象常被称为维数灾难 (curse of dimensionality) [16]. 另外, 更高的光滑性有助于缓解维数灾难. 但是, 若希望逼近率在 d 增大时仍保持在常数量级, 则需要 s 随 d 至少线性增长. 在高维问题中, 这要求目标函数非常光滑, 且随着维数的增长越来越光滑.

下面介绍神经网络在经典函数空间中的一些典型逼近结果, 首先是 Mhaskar 和 Micchelli [126] 及 Mhaskar [123] 证明的 Sigmoidal 激活的深层网络逼近 Hölder 函数的收敛阶.

定理 3.7 设 $\sigma : \mathbb{R} \rightarrow \mathbb{R}$ 为 $k \geq 2$ 阶连续 Sigmoidal 函数, 即 $\sup_{x \in \mathbb{R}} |\sigma(x)| / (1 + |x|)^k < \infty$, $\lim_{x \rightarrow -\infty} \sigma(x)/|x|^k = 0$ 且 $\lim_{x \rightarrow \infty} \sigma(x)/x^k = 1$. 令 $d \in \mathbb{N}_+$, $s > 0$, 则对任意 $f \in C^s([0, 1]^d)$ 和 $N \in \mathbb{N}_+$, 存在深度不超过 $\max(1, \log_k s) + \lceil \log_k d \rceil$, 神经元个数不超过 N 的神经网络 Φ_N , 使得

$$\|f - \Phi_N\|_{L^\infty([0,1]^d)} \leq C N^{-s/d} \|f\|_{C^s([0,1]^d)},$$

其中常数 C 与 f 和 N 无关.

ReLU^k 是 k 阶 Sigmoidal 函数的一个简单例子. 上述定理表明神经网络逼近 Hölder 函数时能达到与样条方法同阶的精度, 所需参数规模量级也相当. 该结论的证明也与样条逼近理论密切相关: 首先, 由定理 3.6 可知, 任意 s 阶光滑函数都能被多元 B 样条以 $\mathcal{O}(N^{-s/d})$ 的精度逼近; 其次, 这些 B 样条又能被一定规模的神经网络精确表示 (或逼近). 结合这两步便得到定理 3.7. 这种思路后来常被用于借助样条或小波等逼近子来证明神经网络的逼近阶, 包括以下深层网络逼近 Besov 函数的结果:

定理 3.8 [167] 设 $s > 0$, $0 < p, q, r \leq \infty$ 且 $s > d(1/p - 1/r)_+$. 对充分大的 N , 令深度 $L = \mathcal{O}(\log N)$ 和宽度 $W = \mathcal{O}(N)$, 则存在与 N 无关的常数 C 使得

$$\sup_{\|f\|_{B_{p,q}^s([0,1]^d)} \leq 1} \inf_{\substack{\Phi_N \in \mathcal{N}(\text{ReLU}; L, W) \\ \text{其非零权重数 } \mathcal{O}(N \log N)}} \|f - \Phi_N\|_{L^r([0,1]^d)} \leq C N^{-s/d}.$$

定理 3.7 和 3.8 都是针对深层网络, 以下定理表明, 浅层 ReLU^k 网络能在某些 Sobolev 空间中实现 s/d 的收敛阶, 这个收敛阶几乎最优. 该定理的下界由典型的 Vapnik-Chervonenkis (VC) 维数论证得到, 而上界由 Sobolev 空间到变差空间的嵌入结果结合定理 6.1 给出, 与本节其他结果不同.

定理 3.9 [116, 146, 183] 设 Ω 为 $\mathbb{R}^d$ 中的单位球, $k \in \mathbb{N}$, $N \in \mathbb{N}_+$, $s > 0$, $1 \leq p, q \leq \infty$. 则存在与 N 无关的常数 $c, C > 0$ 使得

$$c(N \log N)^{-s/d} \leq \sup_{\|f\|_{W^{s,q}(\Omega)} \leq 1} \inf_{g \in \mathcal{N}(\text{ReLU}^k; 1, N)} \|f - g\|_{L^p(\Omega)} \leq C N^{-s/d},$$

其中第二个不等式在 $s < k + (d + 1)/2$ 且 $q \geq \max(p, 2)$, 或 $s = k + (d + 1)/2$ 且 $q \geq 2$ 时成立.

深层网络逼近经典函数已有较为系统的研究, 表 1 和 2 列举和比较了其中部分结果, 表 1 主要部分摘自文献 [105]. 不同于定理 3.7 和 3.8 给出的逼近阶 s/d , 对于 Hölder 函数, 文献 [159] 证明了 ReLU 网络宽度和深度各自变化时的最优收敛率; 对于 Besov 函数, 文献 [160] 证明了定宽 ReLU 网络关于深度的最优收敛率. 忽略对数因子, 最优收敛率均为 $2s/d$, 后文将介绍这些结果并说明其最优性. 对 Besov 函数, 对于深度和宽度没有限制的网络的最优逼近上界目前仍为公开问题.

更光滑的目标函数, 如 Gevrey 函数或解析函数, 能够被神经网络以次指数甚至指数速度逼近, Mhaskar^[124] 最早证明了浅层网络的谱收敛率, 近年神经网络谱收敛的进展可参见文献 [52, 133, 134] 以及表 3 中的结果.

定理 3.8 的结果是渐近的, 网络的深度和宽度随 N 同步增长. 与之不同, 下面定理 3.10 是非渐近的, 网络宽度和深度相互独立, 并且所有常数均显式给出.

定理 3.10^[105, 159] 设 $\Omega = [0, 1]^d$, N 和 L 为任意正整数, $p \in [1, \infty]$.

(1) 设 s 为自然数, 若 $f \in C^s(\Omega)$, 则存在深度为 $18s^2(L + 2)\log_2(4L) + 2d$ 、宽度为 $17s^{d+1}3^d d(N + 2)\log_2(8N)$ 的 ReLU 网络 $\Phi_{N,L}$, 使得

$$\|f - \Phi_{N,L}\|_{L^\infty([0,1]^d)} \leq 85(s+1)^d 8^s \|f\|_{C^s([0,1]^d)} (NL)^{-2s/d}.$$

表 1 深层 ReLU 网络逼近多项式及一般 $\text{Lip}(\Omega)$ 、 $C^s(\Omega)$ 函数的误差

参考文献	函数类	宽度	深度	精确度	误差范数指标	误差界是否紧	有效性
[184]	多项式	$\mathcal{O}(1)$	$\mathcal{O}(L)$	$\mathcal{O}(2^{-L})$	$p = \infty$	—	任意 $L \in \mathbb{N}_+$
[105]	多项式	$\mathcal{O}(N)$	$\mathcal{O}(L)$	$\mathcal{O}(N^{-L})$	$p = \infty$	—	任意 $N, L \in \mathbb{N}_+$
[155]	$\text{Lip}(\Omega)$	$\mathcal{O}(N)$	3	$\mathcal{O}(N^{-2/d})$	$p \in [1, \infty)$	对 N 几乎紧	任意 $N \in \mathbb{N}_+$
[185]	$\text{Lip}(\Omega)$	$2d + 10$	$\mathcal{O}(L)$	$\mathcal{O}(L^{-2/d})$	$p = \infty$	对 L 几乎紧	较大的 $L \in \mathbb{N}_+$
[156]	$\text{Lip}(\Omega)$	$\mathcal{O}(N)$	$\mathcal{O}(L)$	$\mathcal{O}(N^{-2/d} L^{-2/d})$	$p \in [1, \infty]$	对 N 和 L 几乎紧	任意 $N, L \in \mathbb{N}_+$
[159]	$\text{Lip}(\Omega)$	$\mathcal{O}(N)$	$\mathcal{O}(L)$	$\mathcal{O}((N^2 L^2 \ln N)^{-1/d})$	$p \in [1, \infty]$	对 N 和 L 紧	任意 $N, L \in \mathbb{N}_+$
[187]	$C^s(\Omega)$	$2d + 10$	$\mathcal{O}(L)$	$\mathcal{O}((L/\ln L)^{-2s/d})$	$p = \infty$	对 L 几乎紧	较大的 $L \in \mathbb{N}_+$
[105]	$C^s(\Omega)$	$\mathcal{O}(N \ln N)$	$\mathcal{O}(L \ln L)$	$\mathcal{O}(N^{-2s/d} L^{-2s/d})$	$p = \infty$	对 N 和 L 几乎紧	任意 $N, L \in \mathbb{N}_+$
[159]	$C^s(\Omega)$	$\mathcal{O}(N)$	$\mathcal{O}(L)$	$\mathcal{O}((N^2 L^2 \ln N)^{-s/d})$	$p \in [1, \infty]$	对 N 和 L 紧	任意 $N, L \in \mathbb{N}_+$

表 2 深层 ReLU 网络逼近 Sobolev 和 Besov 空间中的函数, $\Omega = [0, 1]^d$, 误差范数 $\|\cdot\|_{W^{\alpha,p}(\Omega)}$

参考文献	函数类	宽度	深度	非零参数规模	逼近误差	误差范数指标
[66]	$W^{s,p}(\Omega)$, $s \geq 2$	$\mathcal{O}(N)$	$\mathcal{O}(\ln N)$	N	$\mathcal{O}((\frac{N}{\ln N})^{-\frac{s-\alpha}{d}})$	$0 \leq \alpha \leq 1$, $1 \leq p \leq \infty$
[3]	$B_{\tau,q}^s(\Omega)$, $s \geq d(1/\tau - 1/p)$, $q \leq \tau \leq p$	$\mathcal{O}(N)$	$\mathcal{O}(\ln N)$	N	$\mathcal{O}(N^{-r})$, 对任意 $r < s/d$	$\alpha = 0$, $1 < p \leq \infty$
[167]	$B_{\tau,q}^s(\Omega)$, $s > d(1/\tau - 1/p)_+$	$\mathcal{O}(N)$	$\mathcal{O}(\ln N)$	$\mathcal{O}(N \ln N)$	$\mathcal{O}(N^{-s/d})$	$\alpha = 0$, $0 < p \leq \infty$
[63]	$B_{p,q}^s(\Omega)$, $0 < s < 1 + \min(1, 1/p)$	—	$1 + 2\lceil \log_2 d \rceil$	N	$\mathcal{O}(N^{-s/d})$	$\alpha = 0$, $0 < p \leq \infty$
[160]	$B_{\tau,q}^s(\Omega)$, $1 \leq \tau, q \leq \infty$, $s/d > 1/\tau - 1/p$	$25d + 31$	L	$\mathcal{O}(L)$	$\mathcal{O}(L^{-2s/d})$	$\alpha = 0$, $1 \leq p \leq \infty$

表 3 深层神经网络逼近解析函数和带边界点奇性加权 Gevrey 类 G^δ 的误差, 误差范数取 $\|\cdot\|_X$

参考文献	函数类	宽度	深度	非零参数规模	激活函数	逼近误差	X
[124]	在复邻域解析的函数	N	1	$\mathcal{O}(N)$	满足一定条件	$\mathcal{O}(\rho^{-N}), \rho > 1$	L^∞
[52]	解析类, 幂级数绝对收敛	$d+4$	L	$\mathcal{O}(L)$	ReLU	$\mathcal{O}(e^{-cL^{1/(2d)}})$	L^∞
[134]	解析类, 可全纯延拓到 Bernstein 多椭圆	—	—	N	ReLU	$\mathcal{O}(\exp(-bN^{1/(d+1)}))$	$W^{1,\infty}$
					$\text{ReLU}^r, r \geq 2$	$\mathcal{O}(\exp(-bN^{1/d}))$	$W^{k,\infty}$
[133]	Gevrey 类 $\delta \geq 1$	$\mathcal{O}(N)$	$\mathcal{O}(\ln N)$	N	ReLU 和 ReLU ²	$\mathcal{O}(\exp(-bN^{1/(1+\delta d)}))$	H^1
	Gevrey 类 $\delta < 1$				混合激活	$\mathcal{O}(\Gamma(N^{1/(1+\delta d)})^{-b(1-\delta)})$	

(2) 设 $\alpha \in (0, 1]$. 若 $f \in C^\alpha(\Omega)$ 且任意 $x, y \in \Omega$ 满足 $|f(x) - f(y)| \leq \lambda|x - y|^\alpha$, 则存在深度为 $11L + C_2$ 、宽度为 $C_1 \max(d\lfloor N^{1/d} \rfloor, N + 2)$ 的 ReLU 网络 $\Phi_{N,L}$, 使得

$$\|f - \Phi_{N,L}\|_{L^p(\Omega)} \leq 131\lambda\sqrt{d}(N^2L^2\log_3(N+2))^{-\alpha/d},$$

其中当 $p \in [1, \infty)$ 时, $C_1 = 16$ 且 $C_2 = 18$; 当 $p = \infty$ 时, $C_1 = 3^{d+3}$ 且 $C_2 = 18 + 2d$.

定理 3.10 中的逼近阶 $2s/d$ 显著优于传统逼近方法的最优收敛率 s/d , 该现象常被称为深层 ReLU 网络的超收敛 (参见文献 [41, 159, 160, 185]), 其证明方法是 Bartlett 等 [12, 13] 提出并发展的位提取 (bit-extraction) 技术, 该技术能用仅 $\mathcal{O}(\sqrt{N})$ 个参数的网络, 逼近定义在固定规则网格中 N 个单元上的分片多项式. 利用深层 ReLU 网络的 VC 维数, 文献 [12, 159, 160, 185] 证明了超收敛给出的逼近率是最优的, 定理 3.11(2) 展示了 Siegel [160] 证明的一般 L^p 误差范数下逼近 Besov 函数的下界. 由 $B_{\infty,1}^s(\Omega) \hookrightarrow C^s(\Omega)$, 该下界表明定理 3.10(2) 中关于宽度和深度的逼近率是最优的. 当宽度固定时, 逼近上下界收敛率均为 $L^{-2\alpha/d}$; 当深度固定时, 逼近上下界收敛率均为 $(N^2 \ln N)^{-\alpha/d}$. 同时改变宽度和深度, 除了区域 $\{(N, L) \in \mathbb{N}_+^2 : C_3 \leq N \leq L^{C_4}\}$ 外, 定理 3.10(2) 的上界都是紧的, 其中 C_3 和 C_4 是仅依赖 α 和 d 的正常数. 对于更一般的 Sobolev 和 Besov 函数, Siegel 在 $W^{s,q}(\Omega)$ 和 $B_{q,r}^s(\Omega)$ 紧嵌入 $L^p(\Omega)$ 的条件下证明了定宽 ReLU 网络关于深度的最优逼近率, 该结果在参数量方面也是最优的.

定理 3.11 [160] 令 $\Omega = [0, 1]^d$, $s > 0$ 以及 $1 \leq r, p, q \leq \infty$.

(1) (上界) 若 $1/q - s/d < 1/p$, 则存在常数 $C_1 := C_1(s, p, q, r, d)$ 使得

$$\inf_{f_L \in \mathcal{N}(\text{ReLU}; L, 25d+31)} \|f - f_L\|_{L^p(\Omega)} \leq C_1 \min(\|f\|_{W^{s,q}(\Omega)}, \|f\|_{B_{q,r}^s(\Omega)}) L^{-2s/d}.$$

(2) (下界) 设 $W, L \in \mathbb{N}_+$, 则存在函数 f 满足 $\|f\|_{W^{s,q}(\Omega)} \leq 1$ 且 $\|f\|_{B_{q,r}^s(\Omega)} \leq 1$, 使得

$$\inf_{f_{W,L} \in \mathcal{N}(\text{ReLU}; L, W)} \|f - f_{W,L}\|_{L^p(\Omega)} \geq C(p, d, s) \min\{W^2 L^2 \log(WL), W^3 L^2\}^{-s/d}. \quad (3.4)$$

上述两项分别来自两种 VC 维数上界. 前者适用于固定深度的大宽度情形, 后者适用于固定宽度的大深度情形.

当 $L \rightarrow \infty$ 时, 定理 3.11 中网络宽度 $W = 25d + 31$ 是固定的, 线性依赖于输入维数 d , 这种线性依赖是必要的, 因为若 $W \leq d$, 则深层 ReLU 网络集合在 $C(\Omega)$ 中并不稠密 (参见文献 [68]). 本节已观察到深度网络相较于传统逼近方法的一些优势: 超收敛. 然而, 定理 3.11(2) 表明即使 ReLU 网络实现了超收敛, 其逼近阶 $2s/d$ 仍无法克服维数灾难. 下一节介绍一般情形下神经网络在目标空间的逼近下界, 并从多个角度深入理解神经网络逼近高维 Besov 函数所面临的困难.

4 最优逼近率与逼近下界

逼近论中的一个经典问题, 是刻画特定方法在目标空间中的最优逼近率, 以界定其逼近能力的上限. 上一节表明, 样条方法与神经网络方法在经典光滑函数空间中的逼近率均为 s/d 量级, 其中 s 表示光滑性指标. 定理 3.11 的下界结果进一步指出, 该逼近率对 ReLU 网络而言是不可改进的, 因此其在 Besov 空间中的逼近将不可避免地遭受维数灾难. 本节将更系统地刻画不同逼近方法的最优逼近率, 并介绍一般情形下的逼近下界. 我们将看到, 包括神经网络在内的几类较为广泛的逼近方法在 Besov 空间中都会遭受维数灾难, 逼近阶不超过 s/d 量级.

首先, 对浅层网络, Maiorov 从 1999 年开始的系列工作证明了 n 个一般岭函数的线性组合在 Sobolev 空间中的最优逼近 (参见文献 [62, 110, 111]):

定理 4.1 设 $d \geq 2$. 记 $\mathbb{R}$ 上局部 L^2 可积的函数空间为 $L^2_{\text{loc}}(\mathbb{R})$. 给定 $N \in \mathbb{N}_+$, 考虑函数集合

$$M_N := \left\{ \sum_{k=1}^N g_k(a_k \cdot x) : a_k \in \mathbb{S}^{d-1}, g_k \in L^2_{\text{loc}}(\mathbb{R}) \right\},$$

其中 $\mathbb{S}^{d-1}$ 表示 $\mathbb{R}^d$ 中的单位球面. 设 $s > 0$, $1 \leq q \leq p \leq \infty$, Ω 为 $\mathbb{R}^d$ 中的单位球. 则

$$\sup_{f \in W^{s,p}(\Omega)} \inf_{g \in M_N} \|f - g\|_{L^q(\Omega)} \simeq N^{-\frac{s}{d-1}}, \quad (4.1)$$

其中隐含常数仅依赖于 d, s, p 和 q .

(4.1) 的下界表明, 即使我们有能力对每个目标函数都“精心”设计出每个神经元上最优的激活函数, 浅层网络关于宽度的逼近阶也不会超过 $s/(d-1)$. 通常使用的固定激活函数的网络集合 $\mathcal{N}(\sigma; 1, N)$ 在 $\sigma \in L^2_{\text{loc}}(\mathbb{R})$ 时是 M_N 的子集, 其逼近阶不会高于 $s/(d-1)$. 进一步地, Maiorov 等^[112] 在一族类似 Besov 空间的函数类上构造了均匀概率测度, 并证明了满足 (4.1) 下界的函数集合的测度趋近于 1, 即下界对应的“难逼近”的函数是普遍存在的, 并非孤例.

为了严谨、准确地讨论神经网络等方法在不同空间的最优逼近, 本文引入逼近论中最经典的几个概念. 考虑在 Banach 空间 X 中用一系列嵌套的集合 $\{\Sigma_n\}_{n=0}^\infty$ 以范数 $\|\cdot\|_X$ 逼近函数 f , 其中 $\Sigma_n \subset \Sigma_{n+1} \subset X$, $\forall n \in \mathbb{N}$. n 在某种意义上衡量了 Σ_n 的复杂性, 实践中 X 常使用 $L^p(\Omega)$ 空间. 给定逼近方法 $\Sigma := (\Sigma_n)_{n \geq 0}$ 和 $f \in X$, 定义用 Σ_n 中元素逼近 f 的误差

$$E_n(f)_X := E(f, \Sigma_n)_X := \inf_{g \in \Sigma_n} \|f - g\|_X. \quad (4.2)$$

对于某个目标函数类 $\mathcal{K} \subset X$, 它包含了被逼近函数 f 的某些潜在已知信息, 如正则性和稀疏性等, $\mathcal{K}$ 通常是 X 中的紧集, 人们更关心在整个目标函数类 $\mathcal{K}$ 上的逼近误差

$$E_n(\mathcal{K})_X := E(\mathcal{K}, \Sigma_n)_X := \sup_{f \in \mathcal{K}} E_n(f)_X.$$

在数值求解偏微分方程中, $\mathcal{K}$ 常由偏微分方程 (partial differential equation, PDE) 的正则性定理给出. 方法 Σ 在 $\mathcal{K}$ 上的逼近率同时依赖于 $\mathcal{K}$ 中函数和集合 Σ_n 的性质. 逼近方法分为两大类: 线性和非线性. 若每个集合 Σ_n 都是一个 n 维线性空间, 则称该方法为线性的⁴⁾, 如多项式空间、样条函数空间和小波空间. 所有其他逼近方法统称为非线性的, 对非线性方法, Σ_n 中元素的线性组合可能不再属于 Σ_n , 例如两个宽为 n 的浅层网络函数之和不一定能被宽为 n 的浅层网络表示. 给定两个逼近方法

4) 注意, “线性”一词并非指逼近结果如何依赖于 $f \in X$, 而仅指每个 Σ_n 的结构.

$\Sigma = (\Sigma_n)_{n \geq 0}$ 和 $\Sigma' = (\Sigma'_n)_{n \geq 0}$, 可以通过 $E(\mathcal{K}, \Sigma_n)_X$ 和 $E(\mathcal{K}, \Sigma'_n)_X$ 在 $n \rightarrow \infty$ 时的衰减来比较它们在 $\mathcal{K}$ 上的性能.

引入“宽度”这一概念旨在量化给定目标函数类 $\mathcal{K}$ 上逼近方法的最佳可能性能, 其中最著名的是 Kolmogorov n - 宽度, 它量化了线性逼近的最佳可能效果, 其定义为

$$d_0(\mathcal{K})_X := \sup_{f \in \mathcal{K}} \|f\|_X, \quad d_n(\mathcal{K})_X := \inf_{\dim Y = n} E(\mathcal{K}, Y)_X, \quad n \geq 1,$$

其中下确界对所有 n 维线性空间 $Y \subset X$ 取得. 为了估计下界 $d_n(\mathcal{K})_X$, 我们介绍一对在逼近论、统计、信息论中很重要的概念: 熵数与度量熵. 它们常用于度量最优性. 熵数 $\epsilon_n(\mathcal{K})_X$ 定义为使 $\mathcal{K}$ 可以被 2^n 个半径为 ϵ 的球覆盖的最小 ϵ ; 反之, 度量熵 $H_\epsilon(\mathcal{K})_X$ 定义为覆盖紧集 $\mathcal{K}$ 所需要的半径为 ϵ 的球的最小数量 $N_\epsilon(\mathcal{K})_X$ 的对数, 即 $H_\epsilon(\mathcal{K})_X := \log_2 N_\epsilon(\mathcal{K})_X$. 由定义可知, ϵ_n 提供了使用 n 个比特来编码的逼近方法能达到的最高精度. Carl^[28] 证明了, 若 $\mathcal{K}$ 能被线性空间以某个多项式速率逼近, 则其熵数衰减不慢于该速度, $\mathcal{K}$ 在度量意义下不能太复杂, 即对每个 $r > 0$, 都存在仅依赖 r 的常数 C_r 使得

$$\epsilon_n(\mathcal{K})_X \leq C_r (n+1)^{-r} \sup_{m \geq 0} (m+1)^r d_m(\mathcal{K})_X, \quad \text{对 } n \geq 0. \quad (4.3)$$

对神经网络在内的非线性逼近, 考虑 n 个参数的方法, 将其建模为两个映射的复合: 编码映射 $a_n : X \rightarrow \mathbb{R}^n$ 将目标函数 $f \in X$ 映射到 n 个参数, 重构映射 $M_n : \mathbb{R}^n \rightarrow X$ 构造 f 在参数流形中的逼近. 则对 f 的逼近误差和目标函数类 $\mathcal{K} \subset X$ 上的逼近误差分别为

$$E_{a_n, M_n}(f) := \|f - M_n(a_n(f))\|_X, \quad E_{a_n, M_n}(\mathcal{K})_X := \sup_{f \in \mathcal{K}} E_{a_n, M_n}(f).$$

为研究某类非线性逼近的最优收敛性, 人们需对 a_n 和 M_n 作适当的限制⁵⁾. 一个较弱的假设是取 a_n 和 M_n 均为连续映射, 由此可以定义 (连续) 流形 n - 宽度^[42]

$$\delta_n(\mathcal{K})_X := \inf_{a_n, M_n} \sup_{f \in \mathcal{K}} \|f - M_n(a_n(f))\|_X,$$

它刻画了所有通过连续参数选取实现的 n 参数方法的最优逼近误差. 一方面, 连续选参假设太弱, 实践中切实可行的方法常需要某种数值稳定性, 连续并不能保证这一点. 由此, 文献 [33] 对 a_n 和 M_n 引入了稳定性假设, 定义了若干稳定 (流形) n - 宽度, 并证明了 Carl 型不等式 (4.3) 或其带对数损失的变体成立. 粗略地说, 在文献 [33] 中最弱假设下, 即 a_n 有界且 M_n 满足一种 Hölder 型条件, 若忽略对数因子, 稳定逼近的最优多项式型收敛速率至多与目标空间熵数的衰减速率相当. 但另一方面, 在适当假设下, 文献 [82] 证明了任意从目标空间 $\mathcal{K}$ 到参数流形的连续映射都不是神经网络达到的最优甚至几乎最优逼近, 连续性会排除理论上达到最优或几乎最优的逼近, 使 $\delta_n(\mathcal{K})_X$ 可能低估神经网络的最佳逼近速度. 文献 [82] 的结果一定程度上解释了神经网络训练结果有时表现出关于目标函数的不稳定性. 研究适合神经网络的最优性度量仍是一个重要问题, 它促进人们更深入地理解神经网络逼近的内在机制. 后文还将介绍由网络伪维数给出的最优性.

流形 n - 宽度 $\delta_n(\mathcal{K})_X$ 可以用 Bernstein n - 宽度从下方估计. 记 $U(X)$ 为 X 的单位球, $\mathcal{K}$ 的 Bernstein n - 宽度定义为 $\mathcal{K}$ 所能包含的最大 $n+1$ 维球的半径, 即

$$b_n(\mathcal{K})_X := \sup_{X_{n+1}} \sup \{\rho : \rho U(X_{n+1}) \subset \mathcal{K}\},$$

5) 否则, 最优逼近误差将始终是 0. 因为由 $\mathcal{K}$ 的紧性, 即使 $n = 1$, 对任意精度, 都存在串联有限覆盖的一维参数化 a_n , 使得 $M_n \circ a_n$ 实现该逼近精度.

其中第一个上确界取遍 X 的所有 $n+1$ 维线性子空间 X_{n+1} . 文献 [42] 证明了 $\delta_n(\mathcal{K})_X \geq b_n(\mathcal{K})_X$.

由 Carl 型不等式和这个关系, 我们可利用 Kolmogorov 宽度、熵数和 Bernstein 宽度来量化线性和非线性方法的最优性. 表 4 整理了这些量在 Besov 空间、谱 Barron 空间和 ReLU^k 浅层网络变差空间的渐近收敛率或下界, 后两个空间是由浅层网络的高维逼近研究驱动定义的, 我们将在下一节介绍. 表 4 第一行说明, Besov 空间在高维非常大, 线性方法、 n 比特方法、连续选参方法或满足文献 [33] 中稳定性假设的方法, 其逼近阶都不高于 s/d ; 具体到神经网络, 若要求参数连续依赖于目标函数, 或实践中使用有限精度编码, 在 Besov 空间必有维数灾难. 事实上, 由定理 4.2 可知, 只要目标空间熵数的衰减率不超过 s/d 量级, 包括 Besov 为代表的经典光滑性空间, 以分片多项式、Sigmoidal 函数或一般 Pfaffian 函数^[83]、径向基函数^[154] 为激活函数的网络, 其逼近阶都在 s/d 量级, 与非线性 B 样条逼近 (定理 3.6) 相似.

至此可以看到, 神经网络在相当广泛的几类情形下都会在 Besov 空间遇到维数灾难, 如文献 [41] 中所言: “不能据此得出结论: 神经网络在逼近能力方面并未提供任何新的东西, 它们的逼近优势可能体现在能够处理一些非传统的目标函数类.” 文献 [38] 证明了两类函数能够被 ReLU 网络以指数速度逼近, 但很难被传统逼近子精确逼近. 其中一种具有自相似结构 (如 Takagi 函数), 另一种函数的定义类似三角级数.

为研究和解释神经网络在高维问题中取得的成功, 我们需要探索新的目标空间. 新空间应足够小, 使神经网络的逼近阶不受维数的制约; 新空间也应足够大, 包含人们真正感兴趣的高维函数. 我们将在后两节讨论部分相关工作. 从表 4 中可以看出, 线性方法在谱 Barron 空间和变差空间中会遭受维数灾难, 这两个空间并不太小; 而非线性方法的最佳逼近阶有可能达到 $1/2$, 不受维数影响.

以非线性宽度作为最优性度量, 可进一步评估第 3.3 小节中神经网络逼近结果的最优性. 为此, 通常需要将网络复杂度转化为非零参数个数, 并检验相应编码映射与重构映射是否满足连续性或稳定性条件. 一般而言, 从参数空间到网络函数类的重构映射连续性较易验证; 而编码映射的连续性往往依赖具体构造, 需检查原始证明. 下面举例说明. 定理 3.7 的证明采用固定架构、参数量为 $\mathcal{O}(N)$ 的网络 Φ_N , 且其编码、重构映射均 Lipschitz 连续, 故它在连续逼近和稳定逼近意义下均是最优的. 文献 [66] 中的构造同样满足连续性和稳定性假设, 故其结果在这两种意义下几乎最优, 仅差对数因子. 相比之下, 定理 3.8 所构造的网络虽连续依赖于目标函数, 但其编码映射未必连续, 不满足连续宽度的假设. 类似地, 定理 3.9 中的编码映射也未必连续, 其最优性借助 ReLU^k 网络的伪维数得到.

第 3 节还介绍了以定理 3.10 和 3.11 为代表的超收敛结果, 相关结论也见表 1. 这类结果通常借助 bit-extraction 技术, 将关于深度 L 的逼近指数提升至 $2s/d$. 由于深度为 L 、宽度为 N 的网络参数量为 $\mathcal{O}(LN^2)$, 该速率严格快于连续宽度和稳定宽度所允许的最优指数 s/d , 因此达到超收敛率的网络不满足连续或稳定选参条件. 事实上, 这些构造中至少部分网络参数会随深度增加或逼近误差减小而指数增长, 如编码精细几何结构的特殊权重^[185] 或子网络中实现二进制位提取所需要的斜率参数^[105, 156, 159]. 在实践中表示这类网络需要随精度提升而不断增加比特数, 这些额外要求对网络的数

表 4 Besov 空间、谱 Barron 空间和变差空间的 n - 宽度和熵数, $\Omega = [0, 1]^d$, $X = L^2(\Omega)$, $\mathcal{K}$ 取空间 Y 的单位球 $U(Y)$

目标空间 Y	Kolmogorov n - 宽度 d_n	熵数 ϵ_n	Bernstein n - 宽度 b_n
Besov 空间 $B_{p,q}^s(\Omega)$	$n^{-\frac{s}{d} + (\frac{1}{p} - \frac{1}{2}) +}$ [147]	$n^{-\frac{s}{d}}$ [43]	$n^{-\frac{s}{d}}$ [43]
谱 Barron 空间 $\mathcal{B}^s(\Omega)$	$\gtrsim n^{-\frac{s}{d}}$ [165]	$\gtrsim n^{-\frac{1}{2} - \frac{s}{d}}$ [144]	$\gtrsim n^{-\frac{1}{2} - \frac{s}{d}}$ [165]
变差空间 $\mathcal{V}(\mathbb{P}_k)$	$\gtrsim n^{-\frac{2k+1}{2d}}$ [165]	$n^{-\frac{1}{2} - \frac{2k+1}{2d}}$ [165]	$\gtrsim n^{-\frac{1}{2} - \frac{2k+1}{2d}}$ [165]

值稳定性、可训练性及实际应用均构成显著挑战. 值得一提的是, 文献 [23] 给出了一个基于参数范数约束刻画最优性的结果: 对于采用 Lipschitz 激活函数, 或导数增长由多项式控制的可微激活函数的网络, 若其非零权重个数不超过 N , 且权重幅值至多按 N 的某个多项式增长, 则该网络类对目标空间所能达到的关于 N 的最优逼近率, 不会快于该空间熵数的多项式衰减率.

本节最后, 介绍一个由目标空间度量熵和逼近集合伪维数给出的 L^p 范数下的一般下界 (参见文献 [1]). 设 $\mathcal{X} \subset \mathbb{R}^d$, 函数集合 $G \subseteq \mathbb{R}^{\mathcal{X}}$, 若对 $\mathcal{X}$ 的子集 $S = \{x_i\}_{i=1}^N$, 存在函数 $r: S \rightarrow \mathbb{R}$, 使得对任意 $E \subseteq S$, 均存在 $g_E \in G$ 满足 $g_E(x_i) \geq r(x_i)$ 当且仅当 $x_i \in E, i = 1, 2, \dots, N$, 则称 S 被 G 伪打散. 函数类 G 的伪维数 $\text{Pdim}(G)$ 定义为其在 $\mathcal{X}$ 上能够伪打散的子集基数的上确界.

定理 4.2 设 $1 \leq p < \infty$, $\mathcal{X} \subset \mathbb{R}^d$ 可测, 赋予了概率测度 μ . 令 F 为一个从 $\mathcal{X}$ 到 $[a, b]$ 的可测函数集合, $a < b$, 且存在 $c_0, \varepsilon_0, \alpha > 0$, 使得 $H_\varepsilon(F)_{L^p(\mu)} \geq c_0 \varepsilon^{-\alpha}, \forall \varepsilon \leq \varepsilon_0$ ⁶⁾.

(1) 若 $\text{Pdim}(G) < \infty$, 则存在仅依赖于 $b - a, p, c_0, \varepsilon_0$ 和 α 的常数 $c_1, \varepsilon_1 > 0$, 使得⁷⁾

$$E(F, G)_{L^p(\mu)} \geq \min\{\varepsilon_1, c_1 \text{Pdim}(G)^{-\frac{1}{\alpha}} \log^{-\frac{2}{\alpha}}(\text{Pdim}(G))\}.$$

(2) 设 σ 为分段多项式激活函数, 其在 $K \geq 2$ 个分段上的最大次数为 $\nu \in \mathbb{N}$. G 取深度为 L 、非零参数不超过 W 个的前馈网络函数集合. 则存在 $W_{\min} \in \mathbb{N}_+$ 和 $c_1, c_2, c_3 > 0$ 使得

$$E(F, G)_{L^p(\mu)} \geq \begin{cases} c_1 W^{-\frac{2}{\alpha}} \log^{-\frac{2}{\alpha}}(W), & \text{若 } \nu \geq 2, \\ c_2 (LW)^{-\frac{1}{\alpha}} \log^{-\frac{3}{\alpha}}(W), & \text{若 } \nu = 1, \\ c_3 W^{-\frac{1}{\alpha}} \log^{-\frac{3}{\alpha}}(W), & \text{若 } \nu = 0, \end{cases} \quad \text{对任意 } W \geq W_{\min} \text{ 和 } L \geq 1.$$

文献 [1] 中证明了更一般的结果. 与文献 [114, 187] 相似, 定理 4.2 的证明基于一个朴素但关键的思想: 若函数类 G 以误差 ε 逼近函数类 F , 则 G 在度量熵意义下至少“不小于” F , 即 $H_{2\varepsilon}(F)_{L^p} \leq H_\varepsilon(G)_{L^p}$. 此后, G 的度量熵可由其胖打散 (fat-shattering) 维数和伪维数控制. 类似地, 文献 [159, 160, 185] 基于网络的 VC 维数证明了在经典光滑性空间中的逼近下界 (如定理 3.11). 这些下界的特点是仅约束激活函数和网络规模, 无需对网络施加非线性宽度定义里的假设.

迄今为止, 我们尚未讨论激活函数对神经网络逼近精度的影响. 现有文献中的多数逼近结果主要针对少数典型激活函数, 如 Sigmoidal 函数和 ReLU; 关于更一般激活函数的系统研究仍相对有限. 文献 [39] 证明了, 具有两个隐藏层的 tanh 网络在逼近精度上可与深度很大的 ReLU 网络相媲美. 最近, 文献 [101] 进一步证明, 即使网络深度固定, 例如仅为 6 层或 7 层, 使用光滑激活函数的网络也可仅通过增加宽度达到逼近 Hölder 函数的最优速率. 光滑激活函数 (如 Swish, GELU 和 Tanh 等) 的一个显著优势在于其具备所谓的“光滑自适应”. 相比之下, ReLU 网络不具备该性质. 若需用 ReLU 网络逼近高度光滑的函数, 则必须增加网络深度; 目标函数越光滑, 逼近精度要求越高, 所需层数亦随之增加. 此外, ReLU 网络的逼近阶受到深度的严格限制, 存在所谓的“深度瓶颈”. 具体而言, 文献 [101] 为固定深度的 ReLU 网络建立了逼近下界, 表明固定深度 ReLU 网络无法对任意光滑度的目标函数达到极小极大最优逼近率; 其可达到的逼近速率本质上受限于网络深度. 这一结果在光滑与非光滑激活函数之间提供了一个可证明的分离性结论.

前两节总结了神经网络在经典函数空间中的逼近结果. 总体来看, 神经网络相较于多项式、小波和三角函数等传统逼近工具并未显示出显著优势. 作为一个较新的逼近子, 神经网络的优势究竟体现

6) 此处用到打包数 $M_\varepsilon(\mathcal{K})_X$ 与覆盖数 $N_\varepsilon(\mathcal{K})_X$ 的等价性: $M_{2\varepsilon}(\mathcal{K})_X \leq N_\varepsilon(\mathcal{K})_X \leq M_\varepsilon(\mathcal{K})_X$.

7) 此定理中约定: $0 \times \log^2(0) = 0$, 且当 $P = 1$ 时 $P^{-\frac{1}{\alpha}} \log^{-\frac{2}{\alpha}}(P) = +\infty$.

在哪里, 仍是一个值得深入研究的问题. 一方面, 可以考察其对新型目标函数空间的逼近性能, 尤其是对高维函数的逼近能力; 这也是下面两节的主要内容. 另一方面, 也可进一步研究其对特殊函数类的逼近能力, 如前述 Takagi 函数及奇异函数^[74]. 此外, 神经网络与传统逼近工具之间的关系, 以及如何推广神经网络以提升其对跨越超曲面奇性的逼近能力, 也是值得关注的方向. 众所周知, 神经网络在逼近跨越超平面奇性方面优于传统方法 (参见文献 [26]).

5 神经网络的新型目标函数空间

前几节中我们看到, 神经网络逼近具有经典光滑性的函数多半会遭受维数灾难. 因而在高维问题中表现优越的函数类, 应有别于经典光滑函数. 最近的研究表明, 神经网络在一些空间中确实能达到不受维数制约的逼近阶, 一个不完全的总结包括 Barron 函数^[7, 8, 81]、Radon BV 函数^[138]、变差空间函数^[10, 164]、带限函数^[29]、特殊结构函数^[125]、混合 Besov 函数^[167]、变指标 Besov 函数^[172] 和 Korobov 函数^[129, 182] 等, 其中后三类函数的逼近证明都用到稀疏网格相关技术进行展开, 部分证明思路和技术与第 3.3 小节相似.

本节介绍前三类函数以及深层网络目标空间. 这些空间从神经网络本身的表示结构自然导出. 浅层网络的基本原子是脊函数 $x \mapsto \sigma(w^\top x + b)$, 考虑宽度为 N 的浅层网络

$$f_N(x; \Theta) = \frac{1}{N} \sum_{k=1}^N a_k \sigma(w_k^\top x + b_k), \quad \text{其中参数集合 } \Theta := \{(a_k, w_k, b_k)\}_{k=1}^N. \quad (5.1)$$

若允许宽度趋于无穷, 直观上, 上述有限求和趋于关于参数 (w, b) 的积分表示

$$f(x) = \int_{\Xi} \sigma(w^\top x + b) d\mu(w, b), \quad x \in \Omega. \quad (5.2)$$

再将此积分表示用 N 个原子离散化, 就回到宽度为 N 的浅层网络, 例如 (5.1) 中的 f_N 可视为 (5.2) 中 f 的 Monte-Carlo 离散. 因而, 适合浅层网络的函数空间应刻画上述积分表示中测度 μ 的某种最小全变差; 这正是 Barron 空间、Radon-BV 空间和一般变差空间的共同出发点. Maurey-Jones-Barron 型引理^[8, 10, 81, 150] 说明, 只要这样的测度范数有限, 就能以 Monte-Carlo 平均或贪心选点的方式得到与维数无关的 $N^{-1/2}$ 型逼近; 反过来, 若一系列浅层网络在路径范数一致有界的意义下逼近某函数, 该函数也必然落入相应 Barron 类^[50, 51]. 这是上述空间作为神经网络目标函数空间的核心动机: 它们直接度量目标函数能否被神经元原子的凸包或测度凸包有效表示, 而不是仅度量传统意义下的局部光滑性. 从 Fourier 角度也有类似思想. 若 f 满足 Fourier 反演, 则它有关于平面波字典的积分表示. Barron^[7-9] 的原始谱条件要求 $\int |\xi| |\hat{f}(\xi)| d\xi$ 有限, 其控制积分中高频原子的加权总量, 从而把 Fourier 反演转化为浅层网络的离散化表示.

后文将从更严谨的数学角度, 介绍这些空间的定义和性质.

5.1 谱 Barron 空间 (spectral Barron space)

Barron 函数类作为神经网络逼近的目标函数类最早由 Barron^[8] 和 Jones^[81] 引进. 我们简称其为谱 Barron 类. 他们证明了宽度为 N 的浅层网络逼近谱 Barron 函数的收敛阶为 $\mathcal{O}(N^{-1/2})$. 与之前的结果相比, 最引人注目的是, 收敛阶与维数无关. 文献中还有很多与谱 Barron 类相近的函数类, 我们将在之后的小节中逐一简要介绍. 首先给出谱 Barron 空间的严格定义 (参见文献 [99]) 并研究其性质.

定义 5.1 (谱 Barron 空间) 设 $f \in \mathcal{S}'(\mathbb{R}^d)$, 对任意 $s \in \mathbb{R}$, 定义谱半范数

$$v_{f,s} := \int_{\mathbb{R}^d} |\xi|^s |\hat{f}(\xi)| d\xi,$$

其中 $|\xi|$ 是向量 ξ 的 ℓ_2 范数. 定义谱 Barron 范数 $\|f\|_{\mathcal{B}^s} := v_{f,0} + v_{f,s}$ 和 s 阶谱 Barron 空间

$$\mathcal{B}^s(\mathbb{R}^d) := \{f \in \mathcal{S}'(\mathbb{R}^d) \mid \|f\|_{\mathcal{B}^s} < \infty\}.$$

事实上, 谱 Barron 空间的概念可追溯至 Hörmander^[72] 于 20 世纪 60 年代提出的一个函数空间; 该空间后来被称为 Hörmander 空间^[127], 主要用于研究微分算子以及拟微分算子的谱理论.

需要指出的是, 谱范数定义中通常采用的 ℓ_2 范数, 也可以替换成其他范数, 如 ℓ_1 范数. 在某些区域内, 采用 ℓ_1 范数得到的逼近误差可以比 ℓ_2 范数获得更小的常数因子^[8, 27].

在 Barron 和 Jones 等的原始文献中, 谱 Barron 空间并未被严格定义, 主要问题在于未能精确刻画函数 f 所属的可容许函数类. 这些作者仅简单地将具有 Fourier 变换的函数类视为谱 Barron 空间的可容许集合. 然而, 这种定义存在缺陷: 某些显而易见的谱 Barron 函数类并不构成完备的度量空间, 因而不是 Banach 空间. 例如, 文献 [93] 定义的如下函数类, 并不是 Banach 空间:

$$\tilde{\mathcal{B}}^s(\mathbb{R}^d) := \{f \in L^1(\mathbb{R}^d) \mid \|f\|_{\tilde{\mathcal{B}}^s} := v_{f,0} + v_{f,s} < \infty\}, \quad (5.3)$$

$\tilde{\mathcal{B}}^s(\mathbb{R}^d)$ 关于范数 $\|\cdot\|_{\tilde{\mathcal{B}}^s}$ 不完备性的证明参见文献 [99, 引理 2.3].

有界区域 Ω 上的谱 Barron 空间 $\mathcal{B}^s(\Omega)$ 定义如下:

定义 5.2 设 $s \in \mathbb{R}$, f 是有界区域 $\Omega \subset \mathbb{R}^d$ 上的函数, $x_0 \in \Omega$. 定义谱半范数和范数分别为

$$\begin{aligned} v_{f,s,\Omega} &:= \inf_{f_e|_{\Omega}=f} \int_{\mathbb{R}^d} |\xi|_{\Omega,x_0}^s |\hat{f}_e(\xi)| d\xi, \\ \|f\|_{\mathcal{B}^s(\Omega)} &:= \inf_{f_e|_{\Omega}=f} \int_{\mathbb{R}^d} (1 + |\xi|_{\Omega,x_0}^s) |\hat{f}_e(\xi)| d\xi, \end{aligned}$$

其中下确界关于所有 f 在 $\mathbb{R}^d$ 上的延拓 $f_e \in \mathcal{S}'(\mathbb{R}^d)$ 取, 向量对偶范数⁸⁾

$$|\xi|_{\Omega,x_0} := \sup_{x \in \Omega} |(\xi, x - x_0)|. \quad (5.4)$$

谱 Barron 空间 $\mathcal{B}^s(\Omega)$ 定义为 $\mathcal{B}^s(\Omega) := \{f : \Omega \rightarrow \mathbb{R} \mid \|f\|_{\mathcal{B}^s(\Omega)} < \infty\}$.

以下命题整理了谱 Barron 空间的基本性质, 其中性质 (i) 可显著简化文献 [8, 第 7 节] 中判别一个函数是否属于谱 Barron 空间的验证过程. 性质 (i) 在 $\Omega = \mathbb{R}^d$ 情形的证明非常标准, 参见文献 [72]; 性质 (ii) 的证明较初等, 参见文献 [99]; 性质 (iii) 可在文献 [72, 定理 2.2.2] 中找到, 我们在此给出了具体的嵌入常数.

命题 5.1 (i) 设 $\Omega = \mathbb{R}^d$ 或 $\Omega \subset \mathbb{R}^d$ 为有界区域, 则对任意 $s \in \mathbb{R}$, $\mathcal{B}^s(\Omega)$ 是 Banach 空间. 进一步地, 对任意 $s \geq 0$, $\mathcal{B}^s(\Omega)$ 在如下等价范数下构成 Banach 代数: $\|f\|_{\mathcal{B}_*^s(\mathbb{R}^d)} := \|(1 + |\cdot|)^s \hat{f}\|_{L^1(\mathbb{R}^d)}$; 以及当 Ω 有界时, $\|f\|_{\mathcal{B}_*^s(\Omega)} := \inf_{f_e|_{\Omega}=f} \|f_e\|_{\mathcal{B}_*^s(\mathbb{R}^d)}$.

(ii) 谱范数 $v_{f,s}$ 关于 s 为 $\log$ 凸函数. 具体地, 对任意的 $0 \leq s_1 \leq s \leq s_2$ 及 $0 \leq \alpha \leq 1$, 设 $s = \alpha s_1 + (1 - \alpha)s_2$, 则

$$v_{f,s} \leq v_{f,s_1}^\alpha v_{f,s_2}^{1-\alpha} \quad \text{且} \quad \|f\|_{\mathcal{B}^s(\mathbb{R}^d)} \leq \|f\|_{\mathcal{B}^{s_1}(\mathbb{R}^d)}^\alpha \|f\|_{\mathcal{B}^{s_2}(\mathbb{R}^d)}^{1-\alpha}.$$

(iii) 若 $0 \leq s_1 \leq s_2$, 则 $\mathcal{B}^{s_2}(\mathbb{R}^d) \hookrightarrow \mathcal{B}^{s_1}(\mathbb{R}^d)$, 且 $\|f\|_{\mathcal{B}^{s_1}(\mathbb{R}^d)} \leq (2 - s_1/s_2) \|f\|_{\mathcal{B}^{s_2}(\mathbb{R}^d)}$.

8) 注意到, $|\xi|_{\Omega,x_0}$ 中基点 x_0 的不同选取是等价的, 因为对任意 $x_0, x_1 \in \Omega$, 有

$$\sup_{x \in \Omega} |(\xi, x - x_0)| \leq \sup_{x \in \Omega} |(\xi, x - x_1)| + |(\xi, x_1 - x_0)| \leq 2 \sup_{x \in \Omega} |(\xi, x - x_1)|.$$

证明 我们仅证明 Ω 为有界区域的情形, 此时向量范数 $|\xi|_{\Omega, x_0}$ 和 $|\xi|$ 等价. 我们仅需在等价范数 $\|f\|_{\mathcal{B}_*^s(\mathbb{R}^d)}$ 和 $\|f\|_{\mathcal{B}_*^s(\Omega)}$ 下证明. 定义限制映射 $R_\Omega: \mathcal{B}^s(\mathbb{R}^d) \rightarrow \mathcal{D}'(\Omega)$, $R_\Omega F = F|_\Omega$. 任取 $\phi \in C_0^\infty(\Omega)$, 视其为 $C_0^\infty(\mathbb{R}^d)$ 中的函数, 则

$$|\langle R_\Omega F, \phi \rangle| = |\langle F, \phi \rangle| \leq \int_{\mathbb{R}^d} (1 + |\xi|)^s |\widehat{F}(\xi)| (1 + |\xi|)^{-s} |\widehat{\phi}(-\xi)| d\xi.$$

由于 $\widehat{\phi} \in \mathcal{S}(\mathbb{R}^d)$ 是 Schwartz 函数, 所以可得到 $F \mapsto \langle R_\Omega F, \phi \rangle$ 是 $\mathcal{B}^s(\mathbb{R}^d)$ 上的连续线性泛函, 则

$$N_\Omega^s := \ker R_\Omega = \{F \in \mathcal{B}^s(\mathbb{R}^d) : F|_\Omega = 0\}$$

是 $\mathcal{B}^s(\mathbb{R}^d)$ 的闭线性子空间. 按定义, 有 $\mathcal{B}^s(\Omega) = R_\Omega(\mathcal{B}^s(\mathbb{R}^d))$. 由商范数的定义知, 映射

$$\mathcal{B}^s(\mathbb{R}^d)/N_\Omega^s \rightarrow \mathcal{B}^s(\Omega), \quad F + N_\Omega^s \mapsto F|_\Omega$$

是等距同构. 由文献 [35, 第 III 章, 定理 4.2(b)] 知 $\mathcal{B}^s(\Omega)$ 是 Banach 空间. 当 $s \geq 0$, N_Ω^s 还是 Banach 代数 $\mathcal{B}^s(\mathbb{R}^d)$ 中的闭理想: 若 $F|_\Omega = 0$, 则对任意 $G \in \mathcal{B}^s(\mathbb{R}^d)$ 有 $(FG)|_\Omega = 0$. 由文献 [35, 第 VII 章, 定理 2.6] 可知, $\mathcal{B}^s(\Omega)$ 在等价范数下构成 Banach 代数. $\square$

当 $s \geq 0$ 时, 谱 Barron 函数满足点态 Fourier 反演

$$(\widehat{f^\vee})(x) = f(x), \quad \text{a.e. } x \in \mathbb{R}^d. \quad (5.5)$$

此性质在神经网络逼近定理的证明中非常关键, 它常用于推导目标函数的某种积分表示, 如 (5.2). 但人们通常假设点态 Fourier 反演自动成立. 该结果的准确表述及证明参见文献 [99, 引理 2.1].

关于 $\mathcal{B}^s(\mathbb{R}^d)$ 中函数的运算性质, 最著名的是下面的 Wiener-Lévy 定理 [96, 177]:

定理 5.1 设 $s \geq 0$, $f \in \mathcal{B}^s(\mathbb{R}^d)$, F 是一个满足 $F(0) = 0$ 的实解析函数, 则 $F \circ f \in \mathcal{B}^s(\mathbb{R}^d)$.

上述定理中, 一个典型的例子是 $F(x) = e^x - 1$. 关于谱 Barron 空间的更多性质, 特别是代数性质, 在专著 [72] 中有详尽的讨论. 然而, 谱 Barron 空间和标准函数空间 (如 Hölder 空间、Sobolev 空间以及 Besov 空间) 的嵌入关系很长一段时间以来并不明确, 文献 [58] 将其列为一个公开问题. 文献 [99] 给出了如下谱 Barron 空间和 Besov 空间的双边嵌入关系.

定理 5.2 记 ν_d 是 d 维单位球的体积. 设 $s \geq 0$, 则有如下嵌入关系和相应的范数控制:

$$B_{2,1}^{s+d/2}(\mathbb{R}^d) \hookrightarrow \mathcal{B}^s(\mathbb{R}^d) \hookrightarrow B_{\infty,1}^s(\mathbb{R}^d), \quad (5.6)$$

$$2^{-s} \|f\|_{B_{\infty,1}^s(\mathbb{R}^d)} \leq \|f\|_{\mathcal{B}^s(\mathbb{R}^d)} \leq 2^{s+1+d/2} \sqrt{\nu_d} \|f\|_{B_{2,1}^{s+d/2}(\mathbb{R}^d)}. \quad (5.7)$$

嵌入关系 (5.6) 在下面的意义是最优的:

$$\begin{aligned} B_{p,q}^\alpha(\mathbb{R}^d) &\hookrightarrow \mathcal{B}^s(\mathbb{R}^d) \text{ 当且仅当 } B_{p,q}^\alpha(\mathbb{R}^d) \hookrightarrow B_{2,1}^{s+d/2}(\mathbb{R}^d), \\ \mathcal{B}^s(\mathbb{R}^d) &\hookrightarrow B_{p,q}^\alpha(\mathbb{R}^d) \text{ 当且仅当 } B_{\infty,1}^s(\mathbb{R}^d) \hookrightarrow B_{p,q}^\alpha(\mathbb{R}^d). \end{aligned}$$

嵌入关系 (5.6) 的一些较弱的版本散见于之前的文献. 例如, 文献 [8] 叙述了当 $\alpha > 1 + d/2$ 时,

$$H^\alpha(\mathbb{R}^d) \hookrightarrow \mathcal{B}^1(\mathbb{R}^d).$$

事实上, 注意到嵌入关系 $H^\alpha(\mathbb{R}^d) = B_{2,2}^\alpha(\mathbb{R}^d) \hookrightarrow B_{2,1}^{1+d/2}(\mathbb{R}^d)$, 上述结果就是 (5.6) 中第一个嵌入关系的直接推论. 再如, Pinkus [148] 给出了在单位球上的双边嵌入关系

$$W^{\lceil d/2 \rceil + 1, \infty} \hookrightarrow \mathcal{B}^1 \hookrightarrow W^{1, \infty} \hookrightarrow W^{1, 2}.$$

它是定理 5.2 的直接推论, 感兴趣的读者可自行验证. 嵌入不等式 (5.7) 右端的常数虽含维数因子, 但该因子随维数增大而趋于 0, 故不随维数增长, 因为

$$2^{s+1+d/2}\sqrt{\nu_d} \sim \frac{2^{s+1}}{(\pi d)^{1/4}} \left(\frac{8\pi e}{d} \right)^{d/4} \rightarrow 0, \quad \text{当 } d \rightarrow \infty \text{ 时}.$$

下面给出定理 5.2 中嵌入关系的证明. 嵌入关系最优性的证明比较复杂, 参见文献 [99].

定理 5.2 的证明 我们只需要证明范数估计 (5.7). 对任意 $f \in \mathcal{B}^s(\mathbb{R}^d)$, 有

$$\begin{aligned} \|f\|_{\mathcal{B}^s(\mathbb{R}^d)} &= \sum_{j=0}^{\infty} \int_{\mathbb{R}^d} (1 + |\xi|^s) \varphi_j(\xi) |\widehat{f}(\xi)| d\xi \\ &\leq \sum_{j=0}^{\infty} \left(\int_{\text{supp } \varphi_j} (1 + |\xi|^s)^2 d\xi \right)^{1/2} \|\varphi_j \widehat{f}\|_{L^2(\mathbb{R}^d)}. \end{aligned}$$

注意到对于任意的 $j \in \mathbb{N}$, 有

$$\begin{aligned} \int_{\text{supp } \varphi_j} (1 + |\xi|^s)^2 d\xi &\leq \int_{0 \leq |\xi| \leq 2^{j+1}} (1 + |\xi|^s)^2 d\xi \\ &\leq 2\omega_{d-1} \int_0^{2^{j+1}} (1 + r^{2s}) r^{d-1} dr \leq 4\nu_d 2^{(j+1)(2s+d)}. \end{aligned}$$

因此, 综合上面两个估计, 利用 Plancherel 定理可得

$$\|f\|_{\mathcal{B}^s(\mathbb{R}^d)} \leq 2^{s+1+d/2} \sqrt{\nu_d} \sum_{j=0}^{\infty} 2^{j(s+d/2)} \|(\varphi_j \widehat{f})^\vee\|_{L^2(\mathbb{R}^d)}. \quad (5.8)$$

(5.7) 中第二个不等式得证. 其次, 对 $f \in \mathcal{B}^s(\mathbb{R}^d)$, 有 $\varphi_j \widehat{f} \in L^1(\mathbb{R}^d)$. 由 Hausdorff-Young 不等式, 可得

$$\begin{aligned} \|f\|_{B_{\infty,1}^s(\mathbb{R}^d)} &\leq \sum_{j=0}^{\infty} 2^{sj} \|\varphi_j \widehat{f}\|_{L^1(\mathbb{R}^d)} \\ &\leq \|\varphi_0 \widehat{f}\|_{L^1(\mathbb{R}^d)} + 2^s \sum_{j=1}^{\infty} \int_{\mathbb{R}^d} \varphi_j(\xi) |\xi|^s |\widehat{f}(\xi)| d\xi. \end{aligned}$$

因此, $\|f\|_{B_{\infty,1}^s(\mathbb{R}^d)} \leq v_{f,0} + 2^s v_{f,s}$. 嵌入不等式 (5.7) 得证. $\square$

从定理 5.2 出发, 容易给出谱 Barron 空间与 Sobolev 空间、Hölder 空间的嵌入关系.

推论 5.1 设 $s \geq 0$, 以下三个嵌入关系成立:

- (i) 若 $1 \leq p \leq 2$ 且 $\alpha > s + d/p$, 则 $W^{\alpha,p}(\mathbb{R}^d) \hookrightarrow \mathcal{B}^s(\mathbb{R}^d)$;
- (ii) 若 s 不是正整数, 或 s 是正整数且 $d \geq 2$, 则 $W^{s+d,1}(\mathbb{R}^d) \hookrightarrow \mathcal{B}^s(\mathbb{R}^d)$;
- (iii) $\mathcal{B}^s(\mathbb{R}^d) \hookrightarrow W^{s,\infty}(\mathbb{R}^d)$ 且 $\mathcal{B}^s(\mathbb{R}^d) \hookrightarrow C^s(\mathbb{R}^d)$.

推论 5.1 的性质 (i) 和 (iii) 可直接由定理 5.2 得到, 性质 (ii) 参见文献 [99, 推论 2.13]. 直观上, 可将 s 视为谱 Barron 空间的光滑性指标, 这是 Barron 函数类一个较为普适的性质, 如下一小节定义的 Barron 空间以及文献 [122] 引入的空间都有类似的性质. 下面对性质 (iii) 中 $\mathcal{B}^s(\mathbb{R}^d) \hookrightarrow C^s(\mathbb{R}^d)$ 在 $0 \leq s \leq 1$ 的情形给出一个初等证明.

推论 5.1 的证明 设 $0 \leq s \leq 1$. 由 Fourier 反演 (5.5) 得 $\sup_{x \in \mathbb{R}^d} |f(x)| \leq v_{f,0}$. 对任意 $x, y \in \mathbb{R}^d$, 有

$$f(x) - f(y) = \int_{\mathbb{R}^d} \widehat{f}(\xi) (e^{2\pi i x \cdot \xi} - e^{2\pi i y \cdot \xi}) d\xi.$$

直接计算可知 $|e^{2\pi i x \cdot \xi} - e^{2\pi i y \cdot \xi}| \leq 2$ 且 $|e^{2\pi i x \cdot \xi} - e^{2\pi i y \cdot \xi}| \leq 2\pi |\xi| |x - y|$. 因此, 对任意 $s \in [0, 1]$, 有

$$|e^{2\pi i x \cdot \xi} - e^{2\pi i y \cdot \xi}| \leq 2^{1-s} (2\pi |\xi| |x - y|)^s = 2\pi^s |\xi|^s |x - y|^s.$$

综合以上两个不等式可得

$$|f(x) - f(y)| \leq \int_{\mathbb{R}^d} |\widehat{f}(\xi)| |e^{2\pi i x \cdot \xi} - e^{2\pi i y \cdot \xi}| d\xi \leq 2\pi^s v_{f,s} |x - y|^s.$$

故当 $0 \leq s \leq 1$ 时, $\mathcal{B}^s(\mathbb{R}^d) \hookrightarrow C^s(\mathbb{R}^d)$ 且 $\|f\|_{C^s(\mathbb{R}^d)} \leq 2\pi^s \|f\|_{\mathcal{B}^s(\mathbb{R}^d)}$, 嵌入常数与 d 无关. $\square$

上述推论可以从熟知的嵌入关系直接推导出. 上述证明可大幅简化文献 [122, 定理 4.4] 的证明.

定理 5.2 和推论 5.1 基本廓清了谱 Barron 空间、Besov 空间、Hölder 空间以及 Sobolev 空间之间的嵌入关系. 谱 Barron 空间和其他函数空间如位势空间的嵌入关系也可以从上述结论直接推导出.

5.2 Barron 空间

本小节介绍 E 等 [50] 引进的 Barron 空间并研究其性质, 该空间是 Barron 函数类的一个变种, 其中的函数本质上是具有无穷宽度的浅层神经网络. 首先介绍文献 [70] 中定义的基于一般激活函数 σ 的 Barron 空间. Radon 测度是具有有界全变差的正则符号 Borel 测度, 记局部紧 Hausdorff 空间 $\Xi \subset \mathbb{R}^d \times \mathbb{R}$ 上的 Radon 测度空间为 $\mathcal{M}(\Xi)$, 符号测度 $\mu \in \mathcal{M}(\Xi)$ 的全变差测度记为 $|\mu|$. 考虑具有积分表示 (5.2) 的函数. 由于可能存在多个不同的测度 μ 能通过 (5.2) 表示同一个函数 f , 我们将它们归类为集合 $\mathbb{G}_{\sigma, f|\Xi} = \{\mu \in \mathcal{M}(\Xi) \mid (5.2) \text{ 成立}\}$.

定义 5.3 (Barron 空间) 设 $\sigma = \text{ReLU}^s$ 对某个 $s \geq 0$ 或 σ 是 Lipschitz 连续的. 若 $\Omega \subset \mathbb{R}^d$ 有界, 则令向量范数 $|\cdot|_* = |\cdot|_{\Omega, 0}$ 取对偶范数 (5.4), 否则 $|\cdot|_* = |\cdot|_2$. 定义 Barron 范数

$$\|f\|_{\mathcal{B}_\sigma(\Omega)|\Xi} = \begin{cases} \inf_{\mu \in \mathbb{G}_{\sigma, f|\Xi}} \int_{\Xi} (|w|_* + |b|)^s d|\mu|(w, b), & \text{若 } \sigma = \text{ReLU}^s \text{ 且 } s \geq 0, \\ \inf_{\mu \in \mathbb{G}_{\sigma, f|\Xi}} \int_{\Xi} (1 + |w|_* + |b|) d|\mu|(w, b), & \text{其他情形.} \end{cases}$$

则基于激活函数 σ 的 Barron 空间定义为

$$\mathcal{B}_\sigma(\Omega)|\Xi = \{f : \Omega \rightarrow \mathbb{R} \mid \|f\|_{\mathcal{B}_\sigma(\Omega)|\Xi} < \infty\}.$$

当 $\Xi = \mathbb{R}^{d+1}$ 时, 我们在上面定义的记号中省略 $|\Xi$. 另外, 我们简记 $\mathcal{B}_{\text{ReLU}^s}(\Omega)|\Xi$ 为 $\mathcal{B}_s(\Omega)|\Xi$.

Barron 函数关于浅层神经网络有 Jackson 定理和 Bernstein 定理:

定理 5.3 (参见文献 [51, 定理 1 和 2]) 设 $\Omega = [0, 1]^d$ 且 $\sigma = \text{ReLU}$. 记路径范数不超过 Q 的浅层神经网络集合

$$\mathcal{N}_Q := \left\{ f_N(\cdot; \Theta) \text{ 如 (5.1)} \mid \|\Theta\|_{\mathcal{P}} := \frac{1}{N} \sum_{k=1}^N |a_k| (|w_k|_{\Omega, 0} + |b_k|) \leq Q, N \in \mathbb{N}_+ \right\}.$$

(1) 对任意自然数 N 以及 $f \in \mathcal{B}_\sigma(\Omega)$, 存在两层神经网络 $f_N(\cdot; \Theta)$ 使得

$$\|f - f_N\|_{L^2(\Omega)} \leq \sqrt{3} \|f\|_{\mathcal{B}_\sigma(\Omega)} N^{-1/2} \quad \text{且} \quad \|\Theta\|_{\mathcal{P}} \leq 2 \|f\|_{\mathcal{B}_\sigma(\Omega)}.$$

(2) 设 $f^* : \Omega \rightarrow \mathbb{R}$ 连续. 若存在常数 Q 以及一个函数序列 $\{f_m\} \subset \mathcal{N}_Q$ 使得 $f_m \rightarrow f^*$ 在 Ω 上点成立, 则 $f^* \in \mathcal{B}_\sigma(\Omega)$ 且 $\|f^*\|_{\mathcal{B}_\sigma(\Omega)} \leq Q$.

基于定理 5.3, 我们有理由相信 Barron 空间是浅层网络一个合适的目标函数类. 若激活函数满足

$$\int_{\mathbb{R}} |\sigma''(x)| (1 + |x|) dx < \infty,$$

则类似的 Jackson 逼近定理同样成立. 最早被提出且研究最多的是由 ReLU 激活函数定义的 Barron 空间 $\mathcal{B}_1(\mathbb{R}^d)$. 基于文献 [54] 中的结果, 我们将其核心性质总结如下.

定理 5.4 (Barron 函数的结构^[54]) (1) 嵌入关系: 设 Ω 是有界区域. 当 $s > d/2 + 2$ 时, $H^s(\Omega) \hookrightarrow \mathcal{B}_1(\Omega) \hookrightarrow C^{0,1}(\Omega)$, 其中 $C^{0,1}$ 是 Lipschitz 连续函数空间.

(2) 无穷远处的渐近分解: 任意 $f \in \mathcal{B}_1(\mathbb{R}^d)$ 均可唯一分解为两部分之和, 即 $f = f_0 + f_\infty$, 其中 f_0 是 $\mathbb{R}^d$ 上的有界函数, f_∞ 是正一次齐次函数, 即所有 $\lambda > 0$ 满足 $f_\infty(\lambda x) = \lambda f_\infty(x)$.

(3) 奇异集的几何: 设 $f \in \mathcal{B}_1(\mathbb{R}^d)$, 则存在一个集合 $S \subset \mathbb{R}^d$, 它是可数个维数至多为 $d-1$ 的仿射子空间的并集, 使得 f 在 $\mathbb{R}^d \setminus S$ 上是连续可微的.

(4) 空间的不变性: 设 $\Phi : \mathbb{R}^d \rightarrow \mathbb{R}^d$ 为一个 C^1 微分同胚. 若复合算子 $\mathcal{C}_\Phi : f \mapsto f \circ \Phi$ 将 Barron 空间映射到其自身 (即对所有 $f \in \mathcal{B}_1(\mathbb{R}^d)$, 都有 $f \circ \Phi \in \mathcal{B}_1(\mathbb{R}^d)$), 则 Φ 必须是仿射变换.

(5) 衰减性: 对任意 $k \geq 0$, 存在一个非零的 $f \in \mathcal{B}_1(\mathbb{R}^d)$ 以及一个仅依赖 k 的常数 $C > 0$, 使得

$$\sup_{|x| \geq r} |f(x)| \leq C |f(0)| r^{-2k-1}, \quad \text{对任意 } r \geq 2^{-1/2}.$$

上述定理中, 性质 (1) 的第一个嵌入关系可由定理 5.6 和推论 5.1 给出; 性质 (2) 表明 $\mathcal{B}_1(\mathbb{R}^d)$ 函数在无穷远处渐近线性增长; 性质 (4) 表明 Barron 空间对于坐标变换较为敏感; 事实上, 对于谱 Barron 空间, 性质 (4) 对应著名的 Beurling-Helson 定理^[18]. 尽管 ReLU 网络函数通常在无穷远处发散, 但性质 (5) 说明 Barron 空间包含非平凡的快速衰减函数, 进一步地, 文献 [178] 构造了具有有限 Barron 半范数的径向紧支函数. 在下面的定理中, 我们转向刻画不同激活函数 (如 ReLU, Sigmoid 以及 GELU 和 Swish 等) 所定义的 Barron 空间之间的嵌入关系 (参见文献 [70]). 它从 Barron 空间的视角说明激活函数的选取很大程度上影响浅层网络的逼近性质.

定理 5.5 设 $\Omega = [0, 1]^d$, ψ 和 ϕ 为 Lipschitz 激活函数. 若存在 Radon 测度 $\gamma \in \mathcal{M}(\mathbb{R}^2)$ 使得

$$\phi(x) = \int_{\mathbb{R}^2} \psi(wx + b) d\gamma(w, b)$$

对所有 $x \in \mathbb{R}$ 成立且 $\int_{\mathbb{R}^2} (1 + |w| + |b|) d\gamma(w, b) < \infty$, 则 $\mathcal{B}_\phi(\Omega) \hookrightarrow \mathcal{B}_\psi(\Omega)$. 此外,

(1) 对于 $0 \leq s \leq t$, 有 $\mathcal{B}_t(\Omega) \hookrightarrow \mathcal{B}_s(\Omega)$;

(2) 当 $\psi \in C^1(\mathbb{R})$ 且 $\psi'' \in L^1(\mathbb{R})$ 时, $\mathcal{B}_\psi(\Omega) \hookrightarrow \mathcal{B}_1(\Omega)$;

(3) 当 $s \geq 1$, $\psi \in C^{\lceil s \rceil}(\mathbb{R})$, 且右 Caputo 导数 $\partial_+^{s+1} \psi \in L^1((0, \infty))$ 以及左 Caputo 导数 $\partial_-^{s+1} \psi \in L^1((-\infty, 0))$, 并且 Ξ 有界时, $\mathcal{B}_\psi(\Omega) | \Xi \hookrightarrow \mathcal{B}_s(\Omega) | \Xi$.

尽管文献 [50, 51, 54] 对 Barron 函数类进行了较为细致的研究, 但仍有不少问题未得到解决, 如 Barron 函数类与标准函数空间的关系以及基于残差神经网络的 Barron 函数类的性质等. 值得注意的是, 文献 [70, 179] 证明了 Barron 空间与谱 Barron 空间的双边嵌入关系:

定理 5.6 ^[70, 179] 设 $\Omega \subset \mathbb{R}^d$ 为紧集. 对任意自然数 s 和 $\varepsilon \in (0, 1)$, 有

$$\mathcal{B}^{s+1}(\Omega) \hookrightarrow \mathcal{B}_s(\Omega) \hookrightarrow \mathcal{B}^{s-\varepsilon}(\Omega),$$

且对于任意 $f \in \mathcal{B}_s(\Omega)$, 存在仅依赖 s 的常数 C_1 和 C_2 使得

$$C_1 \varepsilon \|f\|_{\mathcal{B}^{s-\varepsilon}(\Omega)} \leq \|f\|_{\mathcal{B}_s(\Omega)} \leq C_2 \|f\|_{\mathcal{B}^{s+1}(\Omega)},$$

其中左侧的嵌入和上式右侧的不等式对 $s = 0$ 也成立.

结合上述嵌入关系和定理 5.2, 可得 $\mathcal{B}_s(\mathbb{R}^d)$ 和 Besov 空间的双边嵌入关系

$$B_{2,1}^{s+1+d/2}(\Omega) \hookrightarrow \mathcal{B}_s(\Omega) \hookrightarrow B_{\infty,1}^{s-\varepsilon}(\Omega),$$

但这个嵌入关系不是最优的, 因为这两个空间的微分维数 (differential dimension) [171] 不相等:

$$s+1+\frac{d}{2}-\frac{d}{2}=s+1>s-\varepsilon=s-\varepsilon-\frac{d}{\infty}.$$

事实上, 谱 Barron 空间、(5.3) 中 $\tilde{\mathcal{B}}^s(\mathbb{R}^d)$ 和上述 Barron 类均可视为 Barron 原始定义的不同实现形式. 此外, 文献 [122] 通过引入函数可积性条件, 提出了一种新的 Barron 类实现, 称之为 Barron 谱空间, 并证明了该空间确为一个 Banach 空间, 且具备与谱 Barron 空间相似的嵌入性质. 由于引入了可积性条件, 该空间中的径向函数具备额外的衰减性质与更好的可积性. 通过引入对数光滑性, 文献 [193] 推广了 Barron 谱空间, 并研究了该空间的嵌入性质与浅层神经网络逼近性质.

5.3 Radon BV 空间

Radon BV 空间最早在文献 [139] 中被研究, Parhi 和 Nowak 为了从连续域变分角度刻画神经网络的训练, 并获得类似样条理论中的表示定理, 想要建立一种能够产生稀疏原子解的正则化框架. 经典 Hilbert 正则化 (如 Sobolev 型 L^2 半范数) 所对应的估计子本质上属于线性方法, 其先验偏好通常表现为各向同性的整体光滑. 当目标函数存在局部奇异性或空间非均匀结构时, 这类正则化往往缺乏自适应性, 从而在统计效率与结构刻画上受限.

相比之下, 测度的全变差 (TV) 范数正则化属于 Banach 空间框架, 采用这类正则项的凸变分问题, 其解集在适当条件下构成弱 * 紧凸集, 极点解通常对应于由有限个 Dirac 原子组成的测度, 从而在连续域呈现原子型稀疏结构. 例如, 在一维, 若以 $\|D^k f\|_{\mathcal{M}}$ 作为正则化范数, 则最优解具有局部自适应样条结构, 可表示为“有限个 Green 函数的平移叠加与有限维零空间多项式之和”.

浅层神经网络的基本单元是岭函数 $x \mapsto \sigma(w^\top x - b)$, 其非光滑性主要沿少数方向集中. Radon 变换提供了天然的按方向降维: $\mathcal{R}f(\alpha, t)$ 将 f 表示为方向 $\alpha \in \mathbb{S}^{d-1}$ 上关于位移 $t \in \mathbb{R}$ 的一维函数. 具体地, 对于 Schwartz 函数 $\varphi \in \mathcal{S}(\mathbb{R}^d)$, 其 Radon 变换定义为

$$\mathcal{R}\{\varphi\}(\alpha, t) := \int_{\mathbb{R}^d} \varphi(x) \delta(\alpha^\top x - t) dx, \quad (\alpha, t) \in \mathbb{S}^{d-1} \times \mathbb{R},$$

其中 δ 为一维 Dirac 测度. $\mathcal{R}\{\varphi\}(\alpha, t)$ 是 φ 在超平面 $\{x \in \mathbb{R}^d : \alpha^\top x = t\}$ 上的积分. 令 ∂_t 表示对 t 的导数, $\mathcal{H}_t$ 表示关于变量 t 的 Hilbert 变换, 定义 Radon 域滤波算子

$$K^{d-1} := c_d (-\partial_t^2)^{\frac{d-1}{2}} = \begin{cases} c_d (-1)^{\frac{d-1}{2}} \partial_t^{d-1}, & d \text{ 为奇数,} \\ c_d (-1)^{\frac{d-2}{2}} \mathcal{H}_t \partial_t^{d-1}, & d \text{ 为偶数,} \end{cases}$$

其中 $c_d := (2\pi)^{1-d}/2$. 基于此, 文献 [138] 定义了所谓的 Radon 化导数 $D_{\mathcal{R}}^k := \partial_t^k K^{d-1} \mathcal{R}$, 并以测度范数 $\mathcal{R}TV^k(f) := \|D_{\mathcal{R}}^k f\|_{\mathcal{M}}$ 作为与岭函数结构相容的稀疏正则化项. 为确保 $D_{\mathcal{R}}^k$ 的零空间为有限维多项式空间, 还需对函数施加增长约束. 于是, Radon 域有界变差 (BV) 空间定义为 (参见文献 [138])

$$\mathcal{R}BV^k(\mathbb{R}^d) := \left\{ f \in \mathcal{S}'(\mathbb{R}^d) : \|D_{\mathcal{R}}^k f\|_{\mathcal{M}} < \infty, \operatorname{ess\,sup}_{x \in \mathbb{R}^d} |f(x)|(1+|x|)^{1-k} < \infty \right\}.$$

该定义将“Radon 域沿 t 方向的一维高阶 BV 正则”与“ α 方向上聚合的测度范数”结合起来. 它不但继承了 BV/TV 的保稀疏性质, 而且适应于神经元的岭结构, 为建立以下稀疏岭样条/神经网络的表示定理奠定了理论基础.

定理 5.7 (表示定理) 考虑如下情形, 假设满足

- (1) 非负损失函数 $\ell(\cdot, \cdot) : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}$ 在其第二个自变量上是凸、强制且下半连续的;
- (2) 线性测量泛函 $h_m : \mathcal{B}BV^k(\mathbb{R}^d) \rightarrow \mathbb{R}$, $f \mapsto \langle h_m, f \rangle$, $m = 1, \dots, M$, 两两线性无关且关于弱 $*$ 拓扑是连续的, 观测个数 M 严格大于算子 $D_{\mathcal{H}}^k$ 的零空间的维数;
- (3) 正则化超参数 $\lambda > 0$ 为固定常数.

则对任意固定的观测向量 $y \in \mathbb{R}^M$, 如下数据拟合变分问题的解集

$$\mathcal{V} := \arg \min_{f \in \mathcal{B}BV^k(\mathbb{R}^d)} \sum_{m=1}^M \ell(y_m, \langle h_m, f \rangle) + \lambda \mathcal{R}TV^k(f)$$

非空、凸且弱 $*$ 紧. 进一步地, 若 $\ell(\cdot, \cdot)$ 是严格凸的, 或损失函数施加了硬约束 $\{y_m = \langle h_m, f \rangle\}_{m=1}^M$, 则解集 $\mathcal{V}$ 等于其极点凸包的弱 $*$ 闭包, 且这些极点解都可以表示为

$$f_{\text{ridge}}(x) = \sum_{n=1}^{N_0} v_n \text{ReLU}^{k-1}(w_n^\top x - b_n) + c(x),$$

其中 $\{v_n\}_{n=1}^{N_0} \subset \mathbb{R} \setminus \{0\}$, $\{w_n\}_{n=1}^{N_0} \subset \mathbb{S}^{d-1}$, $\{b_n\}_{n=1}^{N_0} \subset \mathbb{R}$, $c(\cdot)$ 是次数不超过 $k-1$ 的多项式, 并且 $N_0 < M$. 此外, 所有解的正则化开销都相同, 为

$$\mathcal{R}TV^k(f_{\text{ridge}}) = \sum_{n=1}^{N_0} |v_n|.$$

定理 5.7 的核心结论是在 $\mathcal{B}BV^k(\mathbb{R}^d)$ 上的数据拟合变分问题, 其解集可以被完全刻画为“浅层神经网络 + 多项式”所能实现的函数, 且该表示所需的神经元数严格少于观测数 M . 神经元数严格小于测量数这一事实, 体现了 $\mathcal{R}TV^k$ 正则所带来的稀疏化效应.

在有界区域 Ω 上, $\mathcal{B}BV^k$ 空间可以由标准的限制做法定义

$$\mathcal{B}BV^k(\Omega) := \{f \in \mathcal{D}'(\Omega) : \exists g \in \mathcal{B}BV^k(\mathbb{R}^d) \text{ s.t. } g|_{\Omega} = f\}.$$

当赋予如下范数时, $\mathcal{B}BV^k(\Omega)$ 构成一个 Banach 空间

$$\|f\|_{\mathcal{B}BV^k(\Omega)} := \inf_{g \in \mathcal{B}BV^k(\mathbb{R}^d)} \|g\|_{\mathcal{B}BV^k(\mathbb{R}^d)} \quad \text{s.t. } g|_{\Omega} = f.$$

与定理 5.7 完全类似的表示定理同样成立, 其中网络内层参数 $\{(w_n, b_n)\}_{n=1}^{N_0} \subset Z_{\Omega}$, Z_{Ω} 为所有与区域 Ω 有非空交的超平面的参数集合

$$Z_{\Omega} := \{z = (\alpha, t) \in \mathbb{S}^{d-1} \times \mathbb{R} : \{x \in \mathbb{R}^d : \alpha^\top x = t\} \cap \Omega \neq \emptyset\}.$$

5.4 变差空间 (variation space)

前三小节分别从 Fourier 衰减、无穷宽网络的连续化、连续域上数据拟合变分问题三个角度引入了与浅层网络逼近紧密相关的谱 Barron 空间、Barron 空间和 Radon BV 空间. 本小节引入变差空间,

并从变差空间的视角揭示这些空间之间的紧密联系. 对给定的可分 Hilbert 空间 H 中一个字典 $\mathbb{D} \subset H$, 假设 $\sup_{d \in \mathbb{D}} \|d\|_H = K_{\mathbb{D}} < \infty$, 考虑相对字典 $\mathbb{D}$ 的 Minkowski 泛函

$$\|f\|_{\mathbb{D}} := \inf\{c > 0 \mid f/c \in \overline{\text{conv}(\pm\mathbb{D})}\},$$

其中字典 $\mathbb{D}$ 的对称凸包

$$\overline{\text{conv}(\pm\mathbb{D})} = \overline{\left\{ \sum_{j=1}^n a_j h_j : n \in \mathbb{N}_+, h_j \in \mathbb{D}, \sum_{j=1}^n |a_j| \leq 1 \right\}}$$

也称 $\|f\|_{\mathbb{D}}$ 为变差范数. 变差空间 $\mathcal{X}(\mathbb{D})$ 定义为 $\mathcal{X}(\mathbb{D}) := \{f \in H \mid \|f\|_{\mathbb{D}} < \infty\}$. 在范数 $\|\cdot\|_{\mathbb{D}}$ 下, $\mathcal{X}(\mathbb{D})$ 构成一个 Banach 空间, 其单位球就是 $\overline{\text{conv}(\pm\mathbb{D})}$.

有界区域 $\Omega \subset \mathbb{R}^d$ 上, 谱 Barron 空间和 Barron 空间可以视为变差空间在不同字典下的实现. 对任意 $s \in \mathbb{R}$ 和 $k \in \mathbb{N}$, 考虑两种不同的字典, 即谱字典 $\mathbb{F}_s$ 和 ReLU k 字典 $\mathbb{P}_k$, 分别定义为

$$\begin{aligned} \mathbb{F}_s &:= \{(1 + |\xi|)^{-s} e^{2\pi i \xi \cdot x} \mid \xi \in \mathbb{R}^d\}, \\ \mathbb{P}_k &:= \{\sigma_k(\omega \cdot x + b) \mid \omega \in \mathbb{S}^{d-1}, b \in [c_1, c_2]\}, \end{aligned}$$

其中常数 c_1 和 c_2 满足 $c_1 < \inf\{\omega \cdot x \mid \omega \in \mathbb{S}^{d-1}, x \in \Omega\} < \sup\{\omega \cdot x \mid \omega \in \mathbb{S}^{d-1}, x \in \Omega\} < c_2$.

定理 5.8 (范数等价性结果^[164]) 设 Ω 为有界区域, 则

- (1) 谱字典的变差范数与等价谱 Barron 范数相等, 即当 $s \geq 0$ 时, $\|f\|_{\mathbb{F}_s} = \|f\|_{\mathcal{B}_s^*(\Omega)}$;
- (2) $\mathbb{P}_1$ -变差范数与 Barron 范数的等价性: $\sqrt{d}\|f\|_{\mathbb{P}_1} \simeq \|f\|_{\mathcal{B}_1(\Omega)}$, 此处隐去的常数仅依赖于在定义字典 $\mathbb{P}_1$ 时所选取的常数 c_1 和 c_2 ;
- (3) 模去多项式, $\mathbb{P}_k$ 变差范数与 Radon BV 半范数等价: $\inf_{p \in \mathcal{P}_k} \|f + p\|_{\mathbb{P}_k} = \frac{1}{k!} \mathcal{R}TV_{\Omega}^k(f)$, 其中 $\mathcal{P}_k$ 表示次数不超过 k 的多项式所构成的线性空间.

上述定理为浅层网络相关函数空间提供了一个统一刻画, 建立了变差空间理论与 Barron 和 Radon 型表征之间的严格联系, 统一了浅层网络的部分逼近研究. 定理 5.8 中, 区域有界性和条件 $b \in [c_1, c_2]$ 是关键, 它们保证了 $\mathbb{F}_s$ 和 $\mathbb{P}_k$ 是 $L^2(\Omega)$ 中的有界集. 定理 5.8 证明的核心之一是文献 [164, 引理 3], 它表明紧字典上的变差范数与积分表示中的最小测度全变差完全等价. 事实上, 沿着文献 [164, 命题 4] 相同的思路, 可证明 $\mathbb{P}_k$ -变差范数和 $\mathcal{B}_k(\Omega)$ 范数在 $k \in \mathbb{N}$ 时都等价. 此外, 由定理 5.8 和 5.6, 人们常将谱 Barron 空间 $\mathcal{B}^{k+1}(\Omega)$ 和变差空间 $\mathcal{X}(\mathbb{P}_k)$ 紧密联系, 但由表 4, 这两个空间的大小实质上可能不同, 其熵数衰减率在忽略对数因子下相差 $1/(2d)$.

5.5 深层网络空间

前四小节介绍了由浅层网络高维逼近研究驱动定义的几个空间, 第 5.4 小节将它们一定程度地统一起来. 我们将在第 6 节中看到, 浅层网络已能在这些空间中实现与维数无关的逼近阶. 人们更关心的是, 深层网络在什么空间中能克服维数灾难? 它的逼近空间应如何刻画, 能否建立相应的 Jackson 定理和 Bernstein 定理? 这些空间如何依赖于网络的结构和规模? 这方面的工作可分为两类, 相比于浅层网络, 还不够系统. 一类是构造性地推广浅层网络空间, 如深层 Radon BV 空间^[140, 142]、树状 Barron 空间^[53]、流诱导的空间^[51]、具有组合结构的 Sobolev 函数空间^[125] 等, 其优点是将浅层网络的一些结果推广至深层 (如正逼近定理), 反映了深层网络结构的某些特点, 但缺少一般性和统一性. 深度分离现象 (见第 7 节) 让人们认识到网络深度的增加可能显著增强网络的表达能力, 但从函数空间角度刻画深度增加时, 网络能高效逼近的函数类究竟变大了多少, 目前尚不明确. 另一类是文献 [63], 该工

作定义了深层网络的逼近空间, 研究某个网络架构和规模下, 网络能达到某一逼近阶的全体目标函数构成的空间, 刻画这些空间之间的关系以及它们与经典函数空间之间的关系. 下面介绍该类逼近空间中的部分核心内容.

考虑 X 的任意子集族 $\Sigma = (\Sigma_n)_{n \in \mathbb{N}}$, 其中 $\Sigma_n \subset X$. 对 $f \in X$, $\alpha \in (0, \infty)$ 和 $q \in (0, \infty]$, 定义

$$\|f\|_{A_q^\alpha(X, \Sigma)} := \begin{cases} \left(\sum_{n=1}^{\infty} |n^\alpha E(f, \Sigma_{n-1})_X|^q \frac{1}{n} \right)^{1/q}, & \text{若 } 0 < q < \infty, \\ \sup_{n \geq 1} n^\alpha E(f, \Sigma_{n-1})_X, & \text{若 } q = \infty, \end{cases}$$

其中 $E(f, \Sigma_{n-1})_X$ 如 (4.2) 中定义, $\|f\|_{A_q^\alpha(X, \Sigma)} \in [0, \infty]$. 关联的逼近类定义为

$$A_q^\alpha(X, \Sigma) := \{f \in X : \|f\|_{A_q^\alpha(X, \Sigma)} < \infty\}.$$

当 $q = \infty$ 时, 该类恰好是满足 $E(f, \Sigma_n)_X = \mathcal{O}(n^{-\alpha})$ 的全体 $f \in X$; 而对应 $0 < q < \infty$ 的类则是该类的精细变体. 本小节中, 令 $d, k \in \mathbb{N}_+$ 任取, 区域 $\Omega \subseteq \mathbb{R}^d$. 当 $p \in (0, \infty)$ 时, $X := L^p(\Omega; \mathbb{R}^k)$; 当 $p = \infty$ 时, X 为定义在 Ω 上取值于 $\mathbb{R}^k$ 在无穷远处消失的一致连续函数空间, 并赋予上确界范数.

对于神经网络, 设 σ 为激活函数, $\mathcal{L} : \mathbb{N}_+ \rightarrow \mathbb{N}_+ \cup \{\infty\}$ 为一个深度增长函数, 它单调不减. 当复杂度为 n 时, 记 $SW_n(X, \sigma, \mathcal{L})$ 和 $SR_n(X, \sigma, \mathcal{L})$ 分别表示 $\mathcal{N}_d^k(\sigma; \mathcal{L}(n)) \cap X$ 中非零权重数不超过 n 和神经元总数不超过 n 的网络函数的集合. 定义相应的逼近空间

$$SW_q^\alpha(X, \sigma, \mathcal{L}) := A_q^\alpha(X, (SW_n(X, \sigma, \mathcal{L}))_{n \in \mathbb{N}_+})$$

和

$$SR_q^\alpha(X, \sigma, \mathcal{L}) := A_q^\alpha(X, (SR_n(X, \sigma, \mathcal{L}))_{n \in \mathbb{N}_+}).$$

文献 [63] 将由上述两类网络定义的逼近空间称为严格 σ - 网络逼近空间. 若允许某些神经元的激活函数取恒等映射, 则称其为广义 σ - 网络, 可类似地定义 $\mathcal{W}_n$, $\mathcal{R}_n$, $\mathcal{W}_q^\alpha$ 和 $\mathcal{R}_q^\alpha$. 广义网络涵盖多种重要网络结构, 如多层稀疏线性变换 [92], ResNet [69] 和 U-net [153].

以下定理建立了深层网络生成的逼近空间构成结构良好的拟 Banach 空间的充分条件.

定理 5.9 (逼近空间的完备性与结构) 设逼近率 $\alpha > 0$ 及 $q \in (0, \infty]$ 任取. 记网络的最大深度上界 $L := \sup_{n \in \mathbb{N}_+} \mathcal{L}(n)$. 假设以下两个性质之一成立:

- (1) σ 是连续且非多项式函数, $L \geq 1$, 且区域 Ω 是有界的;
- (2) 网络函数集 $\mathcal{N}_d^1(\sigma; L) \cap L^p(\mathbb{R}^d)$ 中包含某个非零的、紧支集的、有界的非负函数 $g \neq 0$.

则逼近空间 $(\mathcal{W}_q^\alpha(X, \sigma, \mathcal{L}), \|\cdot\|_{\mathcal{W}_q^\alpha})$ 和 $(\mathcal{R}_q^\alpha(X, \sigma, \mathcal{L}), \|\cdot\|_{\mathcal{R}_q^\alpha})$ 均为拟 Banach 空间.

一个有趣的结果是, 在适当条件下, 严格与广义 σ - 网络诱导的逼近空间相同, 此时跳跃连接并不改变逼近空间的大小, 即不提升该意义下的逼近表达能力. 具体结论见定理 5.10.

定理 5.10 (严格与广义网络的逼近空间) 设可测集 Ω 测度非零, 且以下两个条件之一成立:

- (1) Ω 有界, σ 连续, σ 在某个 $x_0 \in \mathbb{R}$ 处可微且 $\sigma'(x_0) \neq 0$;
- (2) 可以用 $m \in \mathbb{N}_+$ 项 σ 表示 $\mathbb{R}$ 上的恒等映射.

则对任意 $\alpha > 0$, $q \in (0, \infty]$, 存在常数 C 使得

$$\begin{aligned} SW_q^\alpha(X, \sigma, \mathcal{L}) &= \mathcal{W}_q^\alpha(X, \sigma, \mathcal{L}), \quad \|\cdot\|_{\mathcal{W}_q^\alpha} \leq \|\cdot\|_{SW_q^\alpha} \leq C \|\cdot\|_{\mathcal{W}_q^\alpha}, \\ SR_q^\alpha(X, \sigma, \mathcal{L}) &= \mathcal{R}_q^\alpha(X, \sigma, \mathcal{L}), \quad \|\cdot\|_{\mathcal{R}_q^\alpha} \leq \|\cdot\|_{SR_q^\alpha} \leq C \|\cdot\|_{\mathcal{R}_q^\alpha}. \end{aligned} \tag{5.9}$$

ResNet 是广义网络的特例, 在逼近空间意义下, 其表达能力并不显著优于严格网络. 要理解残差结构有助于深层网络训练这一经验事实^[69], 需进一步研究具体架构所诱导的逼近类^[63].

以下定理阐明了 ReLU^r 激活函数下严格与广义网络逼近空间的等价性和完备性.

定理 5.11 设任意 $\alpha > 0$, $q \in (0, \infty]$, $r \in \mathbb{N}_+$. 令 $\sigma = \text{ReLU}^r$. 则 (5.9) 成立, 逼近空间等价.

对空间完备性, 若 $L := \sup_{n \in \mathbb{N}_+} \mathcal{L}(n)$ 满足最小深度条件: $L \geq 2$, 或在 Ω 有界或 $d = 1$ 时, $L \geq 1$, 则逼近空间 $(\mathcal{W}_q^\alpha(X, \sigma, \mathcal{L}), \|\cdot\|_{\mathcal{W}_q^\alpha})$ 和 $(\mathcal{R}_q^\alpha(X, \sigma, \mathcal{L}), \|\cdot\|_{\mathcal{R}_q^\alpha})$ 均为拟 Banach 空间.

研究神经网络逼近空间为理解网络的表达能力提供了新视角, 也为将逼近论各种成果应用于深度网络奠定了基础, 然而相关进展仍远未成熟. 第 3.3 小节中的结果提供了经典光滑性空间与神经网络逼近空间之间的部分嵌入关系, 但定理 5.2 所刻画的最优双边嵌入关系才是我们最希望建立的. 除了谱 Barron 空间之外, 其他 Barron 类与经典函数空间之间的双边嵌入关系也仍有待深入研究. 这些嵌入关系及其空间性质不仅是研究高维偏微分方程适定性的理论基础, 近期相关工作可参见文献 [31, 57, 60, 65, 128], 同时也被用于研究偏微分方程中的一些经典问题, 如 Hele-Shaw^[34]、Navier-Stokes 方程^[94] 和 Boltzmann 方程^[48] 等.

6 Barron 型空间上神经网络的高维逼近: 最优速率与深度增益

6.1 浅层网络在 Barron 空间、Radon BV 空间和变差空间中的逼近理论

上一节介绍了神经网络高维逼近研究中定义的若干新型目标函数空间, 它们的共同点是神经网络在其中都能实现与维数无关的逼近阶. 本节重点讨论神经网络在 Barron 型空间上的各类逼近结果, 它们都是在 Barron^[7-9] 和 Jones^[81] 的奠基性工作上发展起来的, 相关的工作甚多, 基于第 5 节的准备, 我们可以把这些结果统一起来. 基于定理 5.8, 对于 Barron 空间、Radon BV 空间以及字典 $\mathbb{P}_k$ 下的变差空间, 在范数等价意义下, 逼近结果可以相互迁移; 基于定理 5.6, $\mathcal{B}^{s+1}(\Omega)$ 中谱 Barron 函数的逼近结果可由变差空间中的逼近结果给出; 基于定理 5.2, 我们还能将这些结果与经典光滑性空间中的结果相比对. 因此, 本小节重点介绍变差空间中的逼近结果, 第 6.2 和 6.3 小节专门介绍谱 Barron 空间中的逼近结果.

对第 5.2 小节中定义的 Barron 空间, 定理 5.3 给出了其正逼近定理和逆逼近定理. 进一步地, 文献 [97] 构建并刻画了浅层 ReLU^k 网络适配的扩展 Barron 空间 $\mathcal{B}_1^k$, 证明了浅层 ReLU^k 网络在 $H^m(\Omega)$ 范数下可实现 $\mathcal{O}(N^{-1/2})$ 的逼近率, 其中 $0 \leq m \leq k$, 当 $m < k$ 时逼近率可提升为 $\mathcal{O}(N^{-1/2-1/d})$.

下面给出浅层网络在变差空间的最优逼近结果 (参见文献 [161, 165]). 在范数等价意义下, 文献 [6, 141, 188] 的结果与定理 6.1 紧密相关或是它的特殊情形.

定理 6.1 设 Ω 为 d 维单位球, $k \in \mathbb{N}$ 且 $0 \leq m \leq k$. 记外层系数 ℓ^1 -模有界的浅层网络集合 $\Sigma_n^1(\mathbb{P}_k) := \{\sum_{i=1}^n a_i d_i, a_i \in \mathbb{R}, d_i \in \mathbb{P}_k, \sum_{i=1}^n |a_i| \leq 1\}$. 则存在仅依赖 d 和 k 的常数 C_1 和 C_2 使得

$$\sup_{\|f\|_{\mathbb{P}_k} \leq 1} \inf_{f_n \in \Sigma_n^1(\mathbb{P}_k)} \|f - f_n\|_{W^{m,\infty}(\Omega)} \leq C_1 n^{-\frac{1}{2} - \frac{2(k-m)+1}{2d}},$$

$$\sup_{\|f\|_{\mathbb{P}_k} \leq 1} \inf_{f_n \in \mathcal{N}(\text{ReLU}^k; 1, n)} \|f - f_n\|_{L_2(\Omega)} \geq C_2 (n \log n)^{-\frac{1}{2} - \frac{2k+1}{2d}}.$$

由文献 [114, 定理 3] 和 [165, 推论 3], 上述逼近上下界在 $k = 0$ 时同样适用于一般的具有有界变差的 Sigmoidal 激活函数. 定理 6.1 的上界是比较强的, 误差范数是 L^∞ 型的, 且对网络外层系数有 ℓ^1 控制. 定理 3.9 的上界是嵌入关系 $H^{(d+2k+1)/2}(\Omega) \subset \mathcal{K}_1(\mathbb{P}_k)$ 和上述定理的直接推论. 定理 6.1 的证明依赖于几何偏差理论^[118-120], 在文献 [161] 之前, 文献 [6, 109] 分别使用文献 [118, 119] 的结果, 证明

了 $k = 1$ 和 $k = 0$ 情形下的 L^∞ 范数逼近, 而文献 [161] 改进并推广了这些结果. 进一步地, 文献 [45] 定义了依赖超平面切割与边界位置的加权变差空间, 并证明了该空间虽严格大于无权变差空间, 但仍具有与定理 6.1 相同的最优 L^2 逼近率, 揭示了区域几何是刻画浅层 ReLU 可逼近函数类的一个本质因素.

6.2 浅层网络在谱 Barron 空间中的逼近理论

浅层网络逼近谱 Barron 函数可追溯到 Barron 的原始文献 [8, 定理 1], 我们将结果叙述如下:

定理 6.2 设 $\Omega \subset \mathbb{R}^d$ 有界, $f \in \mathcal{B}^1(\Omega)$. 则对任意 Sigmoidal 函数 σ , Ω 上概率分布 μ 和 $N \in \mathbb{N}_+$, 存在 $\Phi_N = \sum_{k=1}^N c_k \sigma(a_k^\top \cdot + b_k) + c_0$, $\sum_{k=1}^N |c_k| \leq 2v_{f,1,\Omega}$, $c_0 = f(0)$, 使得

$$\|f - \Phi_N\|_{L^2(\Omega; d\mu)} \leq 2v_{f,1,\Omega} N^{-1/2}.$$

Barron [8] 假设目标函数类形如 $\tilde{\mathcal{B}}^1(\mathbb{R}^d)$ 且更加宽泛. 定理 6.2 最值得关注的是其 $1/2$ 阶收敛率. 有趣的是, 由嵌入关系 (5.6), $\mathcal{B}^s(\mathbb{R}^d)$ 包含 Besov 空间 $B_{2,1}^{s+d/2}(\mathbb{R}^d)$, 若用 N 个 B 样条来逼近 $B_{2,1}^{s+d/2}(\mathbb{R}^d)$ 中的函数, 则收敛率也大致是 $N^{-1/2}$, 这是逼近论中的经典结论 (参见文献 [40, 135]). 与标准样条逼近不同, 除谱范数外, 定理 6.2 估计中的常数均为显式且与维数无关的. 但谱范数本身通常依赖于维数, 例如 f 是一个 Gauss 函数, $v_{f,1} \simeq d^{1/2}$, 其他例子可参见文献 [8, 93]. 同时对一般的函数, 精确计算其谱范数并非易事, 文献 [8, 93] 中有相关的初步讨论; 事实上, 判定一个函数是否属于某个谱 Barron 空间也非平凡, 文献 [93] 更将其列为一个公开问题. 另外需指出, $\mathcal{B}^s(\mathbb{R}^d)$ 本质上属于 Donoho 定义的混合变差空间 [47], 该空间的光滑性指标 s 与维数无关, 但该空间中的函数的光滑性确实依赖维数.

定理 6.2 的证明可基于如下引理, 也被称为 Maurey-Jones-Barron 引理.

引理 6.1 (2-型 Banach 空间下的 Maurey-Pisier 引理 [150]) 设 $(X, \|\cdot\|)$ 为一个 2-型 Banach 空间, 即存在常数 $T_2(X) > 0$, 使得对于 X 中任意有限个独立零均值的随机变量序列 $\{Z_i\}_{i=1}^n$ 有

$$\mathbb{E} \left\| \sum_{i=1}^n Z_i \right\|^2 \leq T_2^2(X) \sum_{i=1}^n \mathbb{E} \|Z_i\|^2.$$

设 $G \subset X$ 为有界子集, 满足 $\sup_{g \in G} \|g\| \leq B$, 并设 $f \in \overline{\text{conv}}(G)$ 属于 G 的闭凸包. 则对任意正整数 n , 存在至少一组 $\{g_i\}_{i=1}^n \subset G$ 使得

$$\left\| f - \frac{1}{n} \sum_{i=1}^n g_i \right\| \leq \frac{T_2(X)B}{\sqrt{n}}.$$

引理 6.1 一度构成了神经网络逼近理论的基石. 任一 Hilbert 空间是 $T_2 = 1$ 的 2-型 Banach 空间; 当 $2 \leq p < \infty$ 时, 概率空间上的 L^p 空间亦是 2-型 Banach 空间, 其 T_2 为 Khintchine 不等式中的最优常数 $C_p = \sqrt{2\pi^{-1/p}\Gamma((p+1)/2)^{1/p}}$ [67], 其中 Γ 表示 Gamma 函数, C_p 量级为 $\sqrt{p}$. 不同于 Jones 在 Hilbert 空间中使用的构造性贪心算法, Banach 空间版本的证明常依赖于概率方法, 即证明随机抽样的均值在期望上满足误差界, 从而推导出存在性.

有关浅层网络逼近谱 Barron 函数的进一步研究, 一个重要问题是探求最优收敛率. 这方面, Makovoz [114, 115] 首先观察到对某些特定的字典, 半阶收敛率可以改进:

引理 6.2 (参见文献 [114, 定理 1]) 设 $\Phi := \{\phi_1, \phi_2, \dots\}$ 为 Hilbert 空间 H 中的有界序列, 定义 $\varepsilon_n(\Phi)$ 为使 Φ 可被 n 个半径为 ε 的球覆盖的最小 ε . 则对任意 $f \in \mathcal{K}(\Phi)$ 和 $n \in \mathbb{N}_+$, 存在 $g = \sum_i a_i \phi_i$ 具有至多 n 个非零系数 a_i , 满足

$$\|f - g\|_H \leq 2\varepsilon_n(\Phi)n^{-1/2}\|f\|_\Phi \quad \text{且} \quad \sum_i |a_i| \leq \|f\|_\Phi.$$

当 $n \rightarrow \infty$ 时, 若 $\varepsilon_n(\Phi) \rightarrow 0$, 上述引理得到的收敛率优于引理 6.1. Makovoz 利用引理 6.2 得到了某些 Sigmoidal 浅层网络逼近其变差空间函数的最优收敛率 $1/2 + 1/(2d)$, 而后可迁移为对 $\mathcal{B}^1$ 的逼近率. 针对其他字典的改进速率可参见后续工作 [6, 86, 163, 165, 180]. 改进的收敛率一般形如 $1/2 + \mathcal{O}(s/d)$, 其中 s 表示目标函数的某种光滑性指标. 有关浅层网络在其变差空间中的最优逼近已有结果, 如定理 6.1, 但对浅层网络逼近不同光滑性的谱 Barron 函数的最优速率, 尚未完全解决. 下面展示来自文献 [163] 的上界, 它给出目前在中、高指标谱 Barron 光滑性下最好的逼近率.

定理 6.3 ^[163] 设 $\Omega = [0, 1]^d$, $k \in \mathbb{N}$, $0 \leq m \leq s - 1/2$ 且 $m < k + 1/2$. 则对任意 $n \geq 2$, 有

$$\sup_{\|f\|_{\mathcal{B}^s(\Omega)} \leq 1} \inf_{f_n \in \mathcal{N}(\text{ReLU}^k; 1, n)} \|f - f_n\|_{H^m(\Omega)} \lesssim n^{-t} \log(n)^q,$$

其中隐含常数不依赖于 n , 指数 t 和 q 分别为

$$t = \frac{1}{2} + \min \left(\frac{2(s-m)-1}{2(d+1)}, k-m+\frac{1}{2} \right),$$

$$q = 1_{\{s \geq (d+1)(k-m+\frac{1}{2})+m+\frac{1}{2}\}} + \left(k-m+\frac{1}{2} \right) 1_{\{s=(d+1)(k-m+\frac{1}{2})+m+\frac{1}{2}\}}.$$

在定理 6.3 中, 被省略的常数可能指数地依赖于维数. 取 $k = m = 0$, 定理 6.3 保证了在 $\mathcal{B}^{1/2}$ 中浅层网络即可达到半阶收敛, 由定理 6.6 知此情形下的收敛率是最优的. 另外, 定理 6.3 说明, 当 s 很大时, 逼近阶会出现饱和现象; 由定理 6.5 知, 在达到饱和时定理 6.3 在相差一个对数因子的意义下也是最优的. 其他文献中的结果 [86, 180] 大多可通过结合嵌入定理 (定理 5.6) 和定理 6.1 得到. 例如, 取 $s = k + 1$, $m = 0$, 结合定理 5.6 和 6.1 可得收敛率 $\frac{1}{2} + \frac{2s-1}{2d}$, 此情形下定理 6.1 的结果仍可改进.

定理 6.2 和 6.3 中用于逼近的网络的内外层参数可能无界. 若要对这些参数加以更多的控制, 则要求目标函数 f 更加光滑, 例如, $f \in \mathcal{B}^s(\mathbb{R}^d)$, $s \geq 2$ ^[86], 目前并不清楚这些结果中关于目标函数光滑性的要求是否最优. 对 Barron 函数类在其他范数下的逼近结果, 如无穷范数, 参见文献 [7, 27]. 相较于文献 [7], 以下结果仅在 $\mathcal{B}^1$ 正则性下就得到极大模误差的半阶收敛率, 且对网络权重和偏差有所控制.

定理 6.4 (参见文献 [27, 命题 2.2]) 设 $\Omega = [0, 1]^d$, $f \in \mathcal{B}^1(\Omega)$, 存在 $\Phi_N \in \mathcal{N}(\text{ReLU}; 1, 8N)$ 使得

$$\|f - \Phi_N\|_{L^\infty(\Omega)} \leq C\sqrt{d}v_{f,1,\Omega}N^{-1/2},$$

其中 C 为通用常数. 此外 Φ_N 的所有权重和偏置均不超过 $6\sqrt{v_{f,1,\Omega}}$.

本节最后, 考察 ReLU^k 网络的逼近下界. 下述定理证实了目标函数非常光滑时的饱和现象, 即无论光滑度 s 多大, 定理 6.3 中 $k - m + 1$ 这一最大速率均无法进一步提升.

定理 6.5 (参见文献 [163, 定理 7]) 设 $\Omega = [0, 1]^d$, $k \in \mathbb{N}$ 且 $0 \leq m \leq s$. 则存在 $f \in \mathcal{B}^s(\Omega)$ 使得

$$\inf_{f_n \in \mathcal{N}(\text{ReLU}^k; 1, n)} \|f - f_n\|_{H^m(\Omega)} \gtrsim n^{m-k-1},$$

其中隐含常数与 n 无关.

事实上, 基于定理 4.2 和 $\mathcal{B}^s(\Omega)$ 的熵数估计 ^[144, 165], 我们能得到另一个一般 s 时的逼近下界. 例如, 对浅层 ReLU 网络, 收敛率不会超过 $1/2 + s/d$. 综合已介绍的结果和定理 6.6, 以 ReLU 激活为例, 图 1 清晰展示了浅层网络逼近光滑性低、中、高的谱 Barron 函数所具有的不同表现和逼近率. 当 $s \leq 1/2$ 时, 最优逼近率与 s 成正比; 当 s 很大时, 逼近率饱和, 受限于激活函数; 而在中间区域, 若 d 较大, 逼近率增长速度较慢, 现有的逼近率上下限已很接近, 但阴影部分表示的理论间隙尚待弥合.

此外, 文献 [162] 证明了几类具有特定性质的一般激活函数, 其浅层网络逼近中高光滑性的谱 Barron 函数也能实现与维数无关的收敛阶.

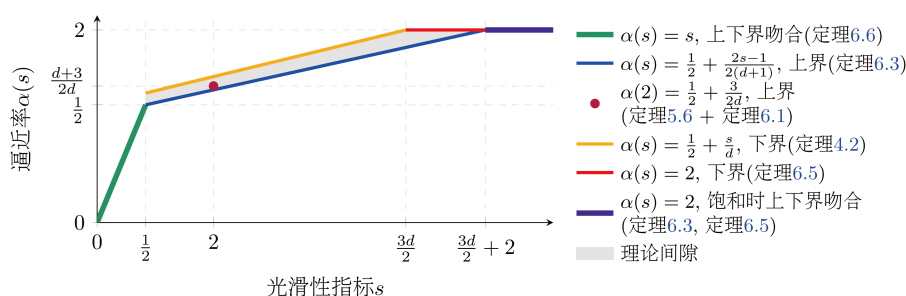


图 1 (网络版彩图) 浅层 ReLU 网络在谱 Barron 空间 $\mathcal{B}^s(\Omega)$ 中的逼近率 (忽略对数因子)

注 6.1 除了深度和宽度, 参数范数也是刻画神经网络表达能力与逼近复杂度的重要因素. 其中最常用且研究最充分的是 ℓ^0 范数, 即非零参数个数; 它与网络连接的稀疏性密切相关, 前文已作详细的讨论. 在用凸组合形式的浅层网络逼近 Barron 型函数时, 经典结果表明, 网络外层参数的 ℓ^1 范数可由目标函数的 Barron 范数或变差范数控制 (参见文献 [6, 50, 51, 165]). 文献 [86] 进一步刻画了 ReLU^k 浅层网络逼近 Barron 函数时, 逼近率与 ℓ^1 型系数控制、 ℓ^0 型稀疏控制之间的直接联系; 定理 6.4 则利用参数受控网络证明了 L^∞ 逼近结果. 如前所述, 这类参数受控结果通常要求目标函数具有更高正则性. 对于深层 ReLU 网络, 由 ReLU 的正齐次性, 路径范数、谱范数乘积型复杂度以及由多层结构诱导的 Banach 空间范数, 常用于刻画其表示能力与逼近复杂度 (参见文献 [11, 53, 59, 132]). 相比之下, 关于参数 ℓ^∞ 型范数如何影响逼近精度的结果仍较少. 文献 [80] 给出了在同时限制深度、宽度和权重矩阵 ℓ^∞ 范数条件下, 网络逼近 Hölder 函数的上下界估计.

从最优性的角度看, 第 4 节已介绍文献 [23] 中关于参数至多多项式增长时的最优逼近率结果. 进一步地, 文献 [136] 给出了深度为 L 、宽度为 W 且参数幅值不超过 B 的全连接 ReLU 网络类的度量熵估计: $\log N_\epsilon(\mathcal{R}(d, W, L, B))_{L^p([0,1]^d)} \asymp W^2 L \log((W+1)^L B^L \epsilon^{-1})$, 其上下界仅相差绝对常数. 将该估计与具有多项式熵增长的目标函数类的度量熵相比较, 可得相应的逼近下界; 其中 B 的影响仅以对数形式出现. 总体来说, 误差对参数范数的精确依赖以及参数受控网络的逼近理论, 在现有文献中仍缺乏系统刻画; 目前结果多限于特殊构造和零星下界估计. 这一问题对于理解神经网络的逼近能力、泛化机制及其实际应用具有重要意义, 值得进一步研究.

6.3 深层网络在谱 Barron 空间中的逼近率与深度增益

定理 6.3 表明半阶收敛可在 $\mathcal{B}^{1/2}$ 中达到. 从空间插值的视角, 我们猜想浅层网络可在 $\mathcal{B}^s$ 中达到 s 阶收敛, s 很小时, 逼近率趋于 0. 另外, 当 s 很小时, $\mathcal{B}^s$ 函数可能含有较大的高频组分, 使正则性变差且难以逼近. 是否能利用深层网络更强的表达能力, 提升对这些函数的逼近率? 以下定理给出了肯定的回答, 展示了深度神经网络逼近光滑性较差的谱 Barron 函数的最优速率:

定理 6.6 ^[99] 设 $\Omega = [0, 1]^d$, $s > 0$, L 和 N 为任意正整数且 $0 < sL \leq 1/2$.

(1) 若 $f \in \mathcal{B}^s(\mathbb{R}^d)$, 则存在 $f_N \in \mathcal{N}(\text{ReLU}; L, 2N+4)$ 使得

$$\|f - f_N\|_{L^2(\Omega)} \leq 29 v_{f,s,\Omega} N^{-sL}.$$

(2) 任取 $0 < \varepsilon \leq 1/2$, 存在 $f \in \mathcal{B}^s(\mathbb{R}^d)$ 满足 $v_{f,s,\Omega} \leq 1 + \varepsilon$, 使得对任意 $1 \leq p \leq \infty$, 有

$$\inf_{f_N \in \mathcal{N}(\text{ReLU}; L, N)} \|f - f_N\|_{L^p(\Omega)} \geq 18^{-1} (1 - \varepsilon) N^{-sL}.$$

定理 6.6 说明, 当 s 很小时, 深度为 L 的网络对 $\mathcal{B}^s$ 函数的逼近率为 sL , 是浅层网络的 L 倍. 深度为 $\lceil 1/(2s) \rceil$ 的网络能达到半阶收敛. 该结果彰显了网络深度在逼近非光滑函数时补偿函数光滑性的重要作用. 相较于定理 6.3, 上述估计右端的常数因子是一个绝对常数, 与维数无关. 此外, 定理 6.6 适用于更一般的 (L, N) -网络 (参见文献 [99, 注 3.10]), (2) 中的下界仅在文献 [99] 的基础上进行了简单的改进.

深层网络函数在空间上的振荡频率可随层数的增加而指数增长 (参见文献 [130, 169]). 下面说明定理 6.6(1) 证明的核心部分, 从而展示深层网络在逼近高频函数时的强大能力. 对任意定义在 $[0, 1]$ 上且关于 $x = 1/2$ 对称的函数 g , 用符号 g_n 表示函数 g 在 $[0, 1]$ 区间内重复 n 次后的形式, 即

$$g_n(t) = g(nt - j), \quad j = 0, \dots, n-1, \quad 0 \leq nt - j \leq 1. \quad (6.1)$$

定义帽函数 $\beta(t) = \text{ReLU}(2t) - 2\text{ReLU}(2t - 1) + \text{ReLU}(2t - 2)$. 根据 (6.1), 可知 β_n 表示一个具有 n 个峰值的三角波函数, 且可由 $3n$ 个 ReLU 单元表示:

$$\beta_n(t) = \sum_{j=0}^{n-1} \beta(nt - j), \quad 0 \leq t \leq 1.$$

引理 6.3 设 g 为定义在 $[0, 1]$ 上且关于 $x = 1/2$ 对称的函数, 则在 $[0, 1]$ 上有 $g_{n_2} \circ \beta_{n_1} = g_{2n_1 n_2}$.

上述引理来自文献 [169, 命题 5.1], 它说明周期变化的高频函数可以由函数复合高效表示, 其几何解释可参见文献 [55, 图 3], 一个关键例子是 $\cos(2\pi n_2 \beta_{n_1}(t)) = \cos(4\pi n_1 n_2 t)$. 为证明定理 6.6, 我们将推导一个谱 Barron 函数关于深层网络函数的期望表示. 为此, 对 $r \in (0, 1)$, 令 χ 为区间上的特征函数且

$$\alpha(t, r) = \chi_{[0, \infty)}(t - r/2) - \chi_{[0, \infty)}(t - (1 - r)/2).$$

则 $\text{supp}(\alpha(\cdot, r)) \subset (0, 1/2)$, 且对每个固定的 r , $\alpha(t, r)$ 关于 $t = 1/4$ 对称. 令

$$\gamma(t, r) = \alpha(t + 1/4, r) - \alpha(t - 1/4, r) + \alpha(t - 3/4, r).$$

则 $\gamma(t, r)$ 关于 $t = 1/2$ 对称. 由 (6.1) 可知, $\gamma_n(\cdot, r)$ 在 $[0, 1]$ 上良定且可用 $4n$ 个 Heaviside 函数 $\chi_{[0, \infty)}$ 表示 (注意到 $\alpha(nt + 1/4, r)$ 和 $\alpha(nt - n + 1/4, r)$ 各只需要一个 Heaviside 函数):

$$\gamma_n(t, r) = \sum_{j=0}^n \alpha(nt - j + 1/4, r) - \sum_{j=0}^{n-1} \alpha(nt - j - 1/4, r).$$

上述构造的一个直接推论是以下引理 6.4. 由 Fourier 反演, 谱 Barron 函数有关于三角函数的期望表示; 利用引理 6.4 以及 γ_n 可由深层网络高效表示, 我们将得到其关于深层网络函数的精确期望表示.

引理 6.4 对于 $t \in [0, 1]$, 有下式成立:

$$\frac{\pi}{2} \int_0^1 \cos(\pi r) \gamma_n(t, r) dr = \cos(2\pi nt).$$

证明 对于任意 $t \in [0, 1/2]$, 有

$$\frac{\pi}{2} \int_0^1 \cos(\pi r) \alpha(t, r) dr = \pi \int_0^{2t} \cos(\pi r) dr = \sin(2\pi t).$$

固定 $t \in [0, 1]$. 若存在整数 j 满足 $0 \leq j \leq n$ 且 $|nt - j| \leq 1/4$, 则 $\gamma_n(t, r) = \alpha(nt - j + 1/4, r)$ 且

$$\frac{\pi}{2} \int_0^1 \cos(\pi r) \alpha(nt - j + 1/4, r) dr = \sin(2\pi(nt - j + 1/4)) = \cos(2\pi nt).$$

否则, 存在整数 j 满足 $0 \leq j \leq n - 1$ 且 $1/4 \leq nt - j \leq 3/4$. 此时 $\gamma_n(t, r) = -\alpha(nt - j - 1/4, r)$ 且

$$-\frac{\pi}{2} \int_0^1 \cos(\pi r) \alpha(nt - j - 1/4, r) dr = -\sin(2\pi(nt - j - 1/4)) = \cos(2\pi nt).$$

引理得证. □

基于 $\mathcal{B}^s$ 函数关于深层网络的期望表示和独立无偏采样, 同时估计抽取的随机网络规模与逼近误差, 可证明定理 6.6(1) 成立. 定理 6.6 下界的证明依赖于深层网络函数在空间上的振荡频率随层数至多指数增长 (参见文献 [169]). 利用改进的技术, 可证明定理 6.6 情形下 L^∞ 逼近的 $\mathcal{O}(N^{-sL} \sqrt{\ln N})$ 收敛率.

我们指出, 定理 6.6 中 $sL \leq 1/2$ 是现有证明方法所引入的技术性假设, $sL > 1/2$ 时的最优逼近率, 以及增加网络深度是否仍能带来额外的逼近增益, 有待进一步研究. 另外, 从最优性角度, 由定理 4.2, 即使对全连接深层 ReLU 网络, 关于宽度的 sL 阶收敛率至多存在于 $sL \leq 1 + 2s/d$ 的范围内, 而定理 6.6 中构造的网络具有特殊结构, 其权重矩阵是分块秩一的, 这不一定能充分发挥全连接网络在 $1/2 < sL \leq 1 + 2s/d$ 情况下的表达能力.

本节对 Barron 型函数类上的神经网络逼近结果作了较为系统的梳理. 对于若干典型 Barron 目标空间, 相应逼近阶已经达到最优, 且主要估计与维数无关. 另外, 当对网络权重参数施加额外限制时, 现有逼近理论仍尚不完整; 这一问题与泛化误差估计密切相关, 值得进一步研究. 此外, 能否将本节结果推广到更一般的激活函数, 也是值得研究的问题.

7 神经网络逼近中的深度分离现象

前文我们主要讨论了什么样的函数能被浅层或深层网络逼近以及逼近精度的问题, 但我们并未专门区分浅层网络和深层网络的逼近能力, 尽管定理 6.6 证明了网络深度能补偿浅层网络对非光滑函数逼近精度不足的问题. 事实上, 深度网络有时可用比浅层网络少得多的神经元来表达给定的函数, 正如 Montúfar^[130] 指出, 网络函数将输入空间划分成很多区域, 这些区域的个数在网络深度上是指数级的, 而在网络宽度上是多项式级的, 从而使用浅层网络表示相同的函数需要指数量级的参数. Telgarsky^[168] 给出了一个简洁优美的构造: 对任意 k , 存在一个深度为 k 宽度为 $\mathcal{O}(1)$ 的 ReLU 网络, 可以表示一个定义在区间 $[0, 1]$ 上振荡次数为 $\mathcal{O}(2^k)$ 的锯齿函数, 但是, 这个函数无法被宽度 $\text{poly}(k)$ 且深度仅为 $\mathcal{O}(k/\ln k)$ 的 ReLU 网络有效逼近. 文献 [169] 将该结论推广到更一般的半代数门模型, 涵盖 ReLU、max、分段多项式、卷积网络、和 - 积网络、提升决策树模型等, 证明了存在深度为 $\mathcal{O}(k^3)$ 宽度为 $\mathcal{O}(1)$ 的网络, 其函数若由仅 $\mathcal{O}(k)$ 层网络逼近, 则必须具有 k 指数级的参数规模. 文献 [168, 169] 的核心思想均在于: 函数复合能快速增加振荡复杂度, 而浅层结构在相同规模下无法产生足够复杂的振荡模式.

Eldan 和 Shamir^[56] 构造了一个更具体的体现三层和两层网络⁹⁾ 深度分离现象的例子, 即能被 $\text{poly}(d)$ 个神经元的三层网络表示, 但至少需要 d 指数量级宽度的两层网络才能逼近的函数. 他们的主要结果可表述如下:

9) 为沿用原文献的表述, 本节中三层和两层网络对应于定义 2.1 中深度为 2 和 1 的网络.

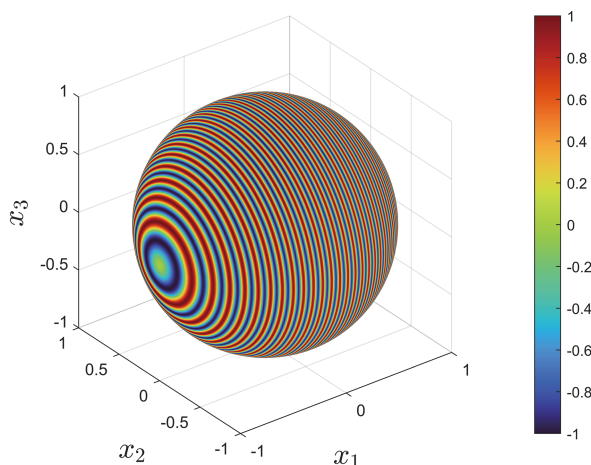


图 2 (网络版彩图) (7.1) 中函数 g 关于单个变量的高频振荡结构, g 的取值仅依赖于 x 和 x' 的相对位置, 因此取 $d = 3$, $x' = (1, 0, 0)$, 则 $g(x, x') = \sin(27\pi x_1)$

定理 7.1 [56] 若对激活函数 σ 作适当技术假设, 则存在普适正常数 c 和 C , 使得对任意维数 $d > C$, 存在 $\mathbb{R}^d$ 上的概率测度 μ 和函数 $g: \mathbb{R}^d \rightarrow \mathbb{R}$, 满足:

(1) g 的值域包含于 $[-2, 2]$, 其支集包含于 $\{x: |x| \leq C\sqrt{d}\}$, 且 g 可由一个宽度为 $Cc_\sigma d^{19/4}$ 的三层神经网络表示, c_σ 是仅依赖 σ 的常数;

(2) 任意可由宽度至多为 ce^{cd} 的两层网络表示的函数 f , 都满足 $\mathbb{E}_{x \sim \mu} (f(x) - g(x))^2 \geq c$.

常见的 ReLU 和单增的 Sigmoidal 激活函数均满足上述定理, 文献 [56] 中构造的 g 和 μ 都是径向的, 相关图解参见文献 [56, 图 1]. 类似的结果还有文献 [37, 117], 文献 [117] 中 g 是模 2 内积, μ 是 $\{0, 1\}^d \times \{0, 1\}^d$ 上的均匀分布; 文献 [37] 中 μ 为 $\mathbb{S}^{d-1} \times \mathbb{S}^{d-1}$ 上的均匀分布. 特别地, Daniely [37] 构造了函数

$$g(x, x') = \sin(\pi d^3 \langle x, x' \rangle), \quad x, x' \in \mathbb{S}^{d-1}, \quad (7.1)$$

并证明了 g 能被宽 $16\pi d^5/\epsilon$ 、权重 $\mathcal{O}(d^3)$ 的三层 ReLU 网络以 ϵ 精度逼近, 但任何权重不超过 2^d 的两层 ReLU 网络, 若要以常数精度逼近 g , 其宽度至少是 2 的 $d \ln d$ 指数量级, 比 2^d 量级更大. 其本质原因是, 该函数的一维分量 $\sin(\pi d^3 x)$ 无法被低阶多项式有效逼近, 因而两层网络难以表示其高频振荡结构, 而三层网络则能通过近似内积再近似一维非线性的复合方式高效逼近. 图 2 给出了 g 关于单个变量的高频振荡结构, 便于读者直观理解. 需要指出的是, 定理 7.1 对权重大小没有任何约束, 而文献 [37, 117] 中要求两层网络的权重不超过指数大小. 总体来看, 深度分离现象的研究, 让人们从理论上认识到网络深度增加可能带来的表达能力的跃升, 并挖掘和利用蕴藏于深度中的强大能力.

8 多种或特定激活函数实现指数收敛

我们在第 4 节看到很多情形下, 特别是当目标函数具有通常光滑性时, 维数灾难似乎不可避免. 但是 Kolmogorov 定理告诉我们, 如果能找到一种超级激活函数, 维数灾难不是问题. Yarotsky 等在文献 [187, 表 1] 中完整刻画了 ReLU 网络中逼近率为 s/d 的连续相, 以及逼近率介于 s/d 与 $2s/d$ 之间的深层不连续相, 最富想象力的是其构造 Lipschitz 周期函数和 ReLU 两种激活函数的网络, 实现极高效的查表、二分式解码机制, 将 Hölder 函数的主要编码信息压缩到单个编码权重中, 进而证明 W 个

权重、深度为 $\sqrt{W}$ 的网络能达到逼近精度 $\mathcal{O}(e^{-c\sqrt{W}})$, 即根指数收敛, 其中 c 是依赖维数和光滑性的常数. 进一步地, 文献 [157] 使用取整函数、指数函数 2^x 和阶跃函数为激活函数的深度为 3、宽度为 $\max(d, N)$ 的网络, 证明了对任意 $f \in C([0, 1]^d)$ 有逼近误差 $2\omega_f(2\sqrt{d})2^{-N} + \omega_f(2\sqrt{d}2^{-N})$, 其中 ω_f 为 f 的连续模. 该网络深度极浅, 且逼近误差对维数 d 的依赖较弱, 但激活函数是间断的. 文献 [78] 构造了一个以 ReLU, sin 和指数函数 2^x 为激活函数的深度为 6 或 $2d+6$ 的网络, 并证明了类似的指数收敛阶, 从而收敛率与维数无关. 该网络函数是广义可微的, 可基于一阶优化方法训练. 理论上, 文献 [158] 证明了一个更强的结果: 构造了三角波函数与 softsign 函数组合而成的显式激活函数 (3.3), 并证明了 $\mathcal{N}(\sigma; 11, 36d(2d+1))$ 是 $C(\mathbb{R}^d)$ 的万有逼近子, 有限深度和宽度的网络即可做到任意精度逼近. 综上所述, 在这些结果中, 作者们通过引入周期性、指数性、离散门控或三角波式结构, 显著提升了网络对高维信息的表达能力. 这些理论结果一方面展示了激活函数更灵活, 信息编码更加紧凑, 中小网络也具有强大的表示能力; 但另一方面, 这些网络在实践中仍面临对编码精度和存储要求过高、难以训练等诸多挑战, 相关问题有待进一步研究.

9 Transformer 的逼近问题

Transformer 是一类由多层多头注意力模块与逐点前馈网络交替堆叠而成的新型神经网络架构, 其中注意力机制是其核心组成部分 (参见文献 [174]).

Transformer 同样具有万有逼近性质. 文献 [189] 证明了, 对定义在紧集上的任意连续序列到序列映射, 存在 Transformer 网络, 甚至单层 Transformer 网络, 可在任意精度下逼近该映射. 这一结果依赖于位置编码的引入. 若无位置编码, Transformer 只能逼近置换等变映射, 即输出随输入顺序的置换作相应置换; 而加入位置编码后, 它能逼近更一般的、对输入顺序敏感的序列映射. 这从理论上保证了 Transformer 拥有足够强大的表达能力来处理复杂的序列任务. 最近文献 [32] 为 Transformer 类架构的万有逼近性质建立了一个统一的理论框架, 将此前依赖具体架构显式构造的证明方式, 转变为基于两个可验证条件的通用判定准则: 其一, 逐点前馈层满足非线性与仿射不变性; 其二, Token 混合层, 即注意力机制, 满足 Token 可区分性. 文中验证了多种现有 Transformer 架构 (如 Softmax 注意力、径向基函数 (radial basis function, RBF) 核注意力、Performer、Linformer、SkyFormer 和 BigBird 等) 均满足条件, 从而具备万有逼近性. 其中, 部分结果 (如 RBF 核注意力) 首次被证明. 该框架不仅能验证现有模型, 还可用于设计具有万有逼近保证的新架构. 作为示例, 文献 [32] 构造了分别具备 D_n 对称性 (正多边形) 和 C_n 对称性 (循环群) 的注意力机制, 证明了它们能实现对应群作用下的万有逼近.

文献 [77] 针对 Hölder 连续目标函数, 显式估计了逼近误差随网络规模, 如层数、宽度和注意力头数, 增大而衰减的速率. 这些速率为比较 Transformer 与循环神经网络 (recurrent neural network, RNN) 等架构在不同任务类别中的理论效率提供了定量依据.

最近, 文献 [79] 证明了 Transformer 在特定条件下可以克服“维数灾难”. 他们证明了对于定义在高维空间中的 Hölder 连续函数, 存在一种 Transformer 网络, 其达到指定精度所需的网络规模 (参数数量) 仅与维数的对数或多项式相关, 而非传统方法的指数级相关. 这为 Transformer 在自然语言处理、图像处理等超高维问题上的成功提供了一个强有力的理论解释.

10 结论与展望

本文概述了神经网络逼近理论的基本结果, 重点梳理了近年来发展较快的高维函数神经网络逼近理论. 针对目标函数属于 Barron 函数类这一特定情形, 我们进行了深入分析, 追溯了该类函数的起源并探讨了其基本性质. 值得注意的是, 该类函数在现有文献中常以不同名称及相似定义出现, 尤其

在偏微分方程研究领域较为常见, 相关综述可参见文献 [122]. 关于经典神经网络逼近理论, 已有综述 [41, 148] 与专著 [149] 进行了详尽论述, 本文不再赘述. 值得一提的是, 我们特别探讨了 Kolmogorov 定理在神经网络逼近中的应用, 这一内容在一般文献中较少涉及, 而该方向目前仍是神经网络逼近理论中一个值得深入探索的问题. 受篇幅限制, 本文未能系统展开神经网络在概率分布学习与算子学习中的逼近理论, 仅列出若干代表性文献供读者进一步参考. 概率分布学习方面, 相关研究主要包括潜变量生成模型、生成对抗网络、归一化流、Wasserstein 生成对抗网络 (Wasserstein generative adversarial network, WGAN)、基于得分函数的生成模型及扩散生成模型等, 其逼近机制可从推前测度、可逆变换表示、积分概率度量和随机过程反演等角度理解; 代表性工作参见文献 [4, 46, 61, 71, 85, 93, 107, 152, 166], 相关综述可参见文献 [20, 137, 181]. 在算子学习方面, 神经网络逼近非线性算子的思想可追溯到早期算子万有逼近理论 [30], 近年来, DeepONet、神经算子和 Fourier 神经算子将该方向推进到函数空间映射和 PDE 解算子学习, 并发展出相应的误差估计、复杂度界和模型降阶分析, 代表性工作可参见文献 [19, 88–91, 98, 106, 176], 综述性介绍可参见文献 [5, 22]. 此外, 本文未涉及基于微分方程的逼近问题, 强化学习的逼近理论 [24, 103, 151], 不变及等变函数的逼近理论 [14, 190, 191] 以及动力系统机器学习的逼近理论, 尤其是大模型的逼近问题. 总体而言, 这些领域的研究尚不完善, 特别缺乏定量结果, 有待进一步发展. 本文也没有探讨深度神经网络在实际问题上的表现以及相应的算法实现, 事实上, 这方面的初步研究 [2] 表明, 深度神经网络的理论逼近极限能力与实际场景表现有较大差距, 亟需发展有效的算法来弥补这一差距.

以下列出若干值得进一步研究的理论问题, 其中多数为本领域尚未解决的开放问题, 也包括我们在研究过程中注意到的若干具体问题.

- (1) 古典部分: 在非渐近意义下, Besov 函数的神经网络最优逼近率.
- (2) 对于给定函数, 如何判定其为 Barron 函数以及如何精确计算其谱范数.
- (3) Transformer 的一般逼近理论与收敛率估计.
- (4) 具有不变或等变性质目标函数的神经网络逼近理论.
- (5) Kolmogorov 定理的构造性证明.

致谢 人工智能驱动的实验、模拟和模型训练是在中国科学院机器人 AI 科学家平台上进行的. 作者衷心感谢审稿人提出的宝贵意见, 这些意见对完善本文的内容和表述均有重要帮助.

参考文献

- 1 Achour E M, Foucault A, Gerchinovitz S, et al. A general approximation lower bound in L^p norm, with applications to feed-forward neural networks. In: *Advances in Neural Information Processing Systems*, vol. 35. Red Hook: Curran Associates, 2022, 22396–22408
- 2 Adcock B, Dexter N. The gap between theory and practice in function approximation with deep neural networks. *SIAM J Math Data Sci*, 2021, 3: 624–655
- 3 Ali M, Nouy A. Approximation of smoothness classes by deep rectifier networks. *SIAM J Numer Anal*, 2021, 59: 3032–3051
- 4 Arjovsky M, Chintala S, Bottou L. Wasserstein generative adversarial networks. In: *Proceedings of the 34th International Conference on Machine Learning. Proceedings of Machine Learning Research*, vol. 70. Ann Arbor: PMLR, 2017, 214–223
- 5 Azizzadenesheli K, Kovachki N, Li Z, et al. Neural operators for accelerating scientific simulations and design. *Nat Rev Phys*, 2024, 6: 320–328
- 6 Bach F. Breaking the curse of dimensionality with convex neural networks. *J Mach Learn Res*, 2017, 18: 1–53
- 7 Barron A R. Neural net approximation. In: *Proceedings of the 7th Yale Workshop on Adaptive and Learning Systems*, vol. 1. New Haven: Yale University, 1992, 69–72
- 8 Barron A R. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Trans Inform Theory*, 1993, 39: 930–945

- 9 Barron A R. Approximation and estimation bounds for artificial neural networks. *Mach Learn*, 1994, 14: 115–133
- 10 Barron A R, Cohen A, Dahmen W, et al. Approximation and learning by greedy algorithms. *Ann Statist*, 2008, 36: 64–94
- 11 Bartlett P L, Foster D J, Telgarsky M J. Spectrally-normalized margin bounds for neural networks. In: *Advances in Neural Information Processing Systems*, vol. 30. Red Hook: Curran Associates, 2017, 6240–6249
- 12 Bartlett P L, Harvey N, Liaw C, et al. Nearly-tight VC-dimension and pseudodimension bounds for piecewise linear neural networks. *J Mach Learn Res*, 2019, 20: 1–17
- 13 Bartlett P L, Maiorov V, Meir R. Almost linear VC-dimension bounds for piecewise polynomial networks. *Neural Comput*, 1998, 10: 2159–2173
- 14 Behler J, Parrinello M. Generalized neural-network representation of high-dimensional potential-energy surfaces. *Phys Rev Lett*, 2007, 98: 146401
- 15 Belkin M. Fit without fear: Remarkable mathematical phenomena of deep learning through the prism of interpolation. *Acta Numer*, 2021, 30: 203–248
- 16 Bellman R. On the theory of dynamic programming. *Proc Natl Acad Sci USA*, 1952, 38: 716–719
- 17 Bellman R. *Dynamic Programming*. Princeton: Princeton Univ Press, 1957
- 18 Beurling A, Helson H. Fourier-Stieltjes transforms with bounded powers. *Math Scand*, 1953, 1: 120–126
- 19 Bhattacharya K, Hosseini B, Kovachki N B, et al. Model reduction and neural networks for parametric PDEs. *SMAI J Comput Math*, 2021, 7: 121–157
- 20 Bond-Taylor S, Leach A, Long Y, et al. Deep generative modelling: A comparative review of VAEs, GANs, normalizing flows, energy-based and autoregressive models. *IEEE Trans Pattern Anal Mach Intell*, 2022, 44: 7327–7347
- 21 Boullé N, Nakatsukasa Y, Townsend A. Rational neural networks. In: *Advances in Neural Information Processing Systems*, vol. 33. Red Hook: Curran Associates, 2020, 14243–14253
- 22 Boullé N, Townsend A. A mathematical guide to operator learning. In: *Numerical Analysis Meets Machine Learning. Handbook of Numerical Analysis*. Amsterdam: Elsevier, 2024, 83–125
- 23 Bölcskei H, Grohs P, Kutyniok G, et al. Optimal approximation with sparsely connected deep neural networks. *SIAM J Math Data Sci*, 2019, 1: 8–45
- 24 Buşoniu L, de Bruin T, Tolić D, et al. Reinforcement learning for control: Performance, stability, and deep approximators. *Annu Rev Control*, 2018, 46: 8–28
- 25 Cai Y. Achieve the minimum width of neural networks for universal approximation. In: *Proceedings of the International Conference on Learning Representations*, 2023, <https://openreview.net/forum?id=hfUJ4ShyDEU>
- 26 Candès E J. *Ridgelets: Theory and applications*. PhD Thesis. Stanford: Stanford University, 1998
- 27 Caragea A, Petersen P, Voigtlaender F. Neural network approximation and estimation of classifiers with classification boundary in a Barron class. *Ann Appl Probab*, 2023, 33: 3039–3079
- 28 Carl B. Entropy numbers, s -numbers, and eigenvalue problems. *J Funct Anal*, 1981, 41: 290–306
- 29 Chen L, Wu C. A note on the expressive power of deep rectified linear unit networks in high-dimensional spaces. *Math Methods Appl Sci*, 2019, 42: 3400–3404
- 30 Chen T P, Chen H. Universal approximation to nonlinear operators by neural networks with arbitrary activation functions and its application to dynamical systems. *IEEE Trans Neural Netw*, 1995, 6: 911–917
- 31 Chen Z A, Lu J F, Lu Y L, et al. A regularity theory for static Schrödinger equations on $\mathbb{R}^d$ in spectral Barron spaces. *SIAM J Math Anal*, 2023, 55: 557–570
- 32 Cheng J P, Lin T, Shen Z W, et al. A unified framework for establishing the universal approximation of transformer-type architectures. In: *Advances in Neural Information Processing Systems*, vol. 38. Red Hook: Curran Associates, 2025, 129131–129163
- 33 Cohen A, DeVore R, Petrova G, et al. Optimal stable nonlinear approximation. *Found Comput Math*, 2022, 22: 607–648
- 34 Constantin P, Pugh M. Global solutions for small data to the Hele-Shaw problem. *Nonlinearity*, 1993, 6: 393–415
- 35 Conway J B. *A Course in Functional Analysis*. New York: Springer, 1990
- 36 Cybenko G. Approximation by superpositions of a sigmoidal function. *Math Control Signals Systems*, 1989, 2: 303–314
- 37 Daniely A. Depth separation for neural networks. In: *Proceedings of the 2017 Conference on Learning Theory. Proceedings of Machine Learning Research*, vol. 65. Ann Arbor: PMLR, 2017, 690–696
- 38 Daubechies I, DeVore R, Foucart S, et al. Nonlinear approximation and (deep) ReLU networks. *Constr Approx*, 2022, 55: 127–172
- 39 De Ryck T, Lanthaler S, Mishra S. On the approximation of functions by tanh neural networks. *Neural Networks*, 2021, 143: 732–750
- 40 DeVore R. Nonlinear approximation. In: *Acta Numerica*, 1998. *Acta Numer*, vol. 7. Cambridge: Cambridge Univ

- Press, 1998, 51–150
- 41 DeVore R, Hanin B, Petrova G. Neural network approximation. *Acta Numer*, 2021, 30: 327–444
 - 42 DeVore R, Howard R, Micchelli C. Optimal nonlinear approximation. *Manuscripta Math*, 1989, 63: 469–478
 - 43 DeVore R, Kyriazis G, Leviatan D, et al. Wavelet compression and nonlinear n -widths. *Adv Comput Math*, 1993, 1: 197–214
 - 44 DeVore R, Lorentz G G. *Constructive Approximation*. Berlin: Springer, 1993
 - 45 DeVore R, Nowak R D, Parhi R, et al. Weighted variation spaces and approximation by shallow ReLU networks. *Appl Comput Harmon Anal*, 2025, 74: 101713
 - 46 Dinh L, Krueger D, Bengio Y. NICE: Non-linear independent components estimation. In: *Proceedings of the International Conference on Learning Representations*, 2015, <https://arxiv.org/abs/1410.8516>
 - 47 Donoho D. High-dimensional data analysis: The curses and blessings of dimensionality. In: *AMS Math Challenges Lecture*. Los Angeles, 2000, 1–32
 - 48 Duan R, Liu S, Sakamoto S, et al. Global mild solutions of the Landau and non-cutoff Boltzmann equations. *Comm Pure Appl Math*, 2021, 74: 932–1020
 - 49 Dubey S R, Singh S K, Chaudhuri B B. Activation functions in deep learning: A comprehensive survey and benchmark. *Neurocomputing*, 2022, 503: 92–108
 - 50 E W, Ma C, Wu L. A priori estimates of the population risk for two-layer neural networks. *Commun Math Sci*, 2019, 17: 1407–1425
 - 51 E W, Ma C, Wu L. The Barron space and the flow-induced function spaces for neural network models. *Constr Approx*, 2022, 55: 369–406
 - 52 E W, Wang Q. Exponential convergence of the deep neural network approximation for analytic functions. *Sci China Math*, 2018, 61: 1733–1740
 - 53 E W, Wojtowytsch S. On the Banach spaces associated with multi-layer ReLU networks: Function representation, approximation theory and gradient descent dynamics. *CSIAM Trans Appl Math*, 2020, 1: 387–440
 - 54 E W, Wojtowytsch S. Representation formulas and pointwise properties for Barron functions. *Calc Var Partial Differential Equations*, 2022, 61: 46
 - 55 Elbrachter D, Perekrestenko D, Grohs P, et al. Deep neural network approximation theory. *IEEE Trans Inform Theory*, 2021, 67: 2581–2623
 - 56 Eldan R, Shamir O. The power of depth for feedforward neural networks. In: *Proceedings of the 29th Annual Conference on Learning Theory*. Ann Arbor: PMLR, 2016, 907–940
 - 57 Feng Y, Lu J F. Solution theory of Hamilton-Jacobi-Bellman equations in spectral Barron spaces. *SIAM J Math Anal*, 2026, 58: 636–660
 - 58 Girosi F, Anzellotti G. Rates of convergence for radial basis functions and neural networks. In: *Artificial Neural Networks for Speech and Vision*. London: Chapman & Hall, 1993, 96–113
 - 59 Golowich N, Rakhlin A, Shamir O. Size-independent sample complexity of neural networks. In: *Proceedings of the 31st Conference on Learning Theory*. *Proceedings of Machine Learning Research*, vol. 75. Ann Arbor: PMLR, 2018, 297–299
 - 60 Golse F, Jin S, Paul T. The random batch method for N -body quantum dynamics. *J Comput Math*, 2021, 39: 897–922
 - 61 Goodfellow I J, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets. In: *Advances in Neural Information Processing Systems*, vol. 27. Red Hook: Curran Associates, 2014, 2672–2680
 - 62 Gordon Y, Maiorov V, Meyer M, et al. On the best approximation by ridge functions in the uniform norm. *Constr Approx*, 2001, 18: 61–85
 - 63 Gribonval R, Kutyniok G, Nielsen M, et al. Approximation spaces of deep neural networks. *Constr Approx*, 2022, 55: 259–367
 - 64 Guliyev N J, Ismailov V E. A single hidden layer feedforward network with only one neuron in the hidden layer can approximate any univariate function. *Neural Comput*, 2016, 28: 1289–1304
 - 65 Guo Y, Ming P, Yu H. Generalization error estimates of a machine learning method for solving high-dimensional Schrödinger eigenvalue problems. *Sci China Math*, 2026, 69: 1063–1114
 - 66 Gühring I, Kutyniok G, Petersen P. Error bounds for approximations with deep ReLU neural networks in $W^{s,p}$ norms. *Anal Appl (Singap)*, 2020, 18: 803–859
 - 67 Haagerup U. The best constants in the Khintchine inequality. *Studia Math*, 1981, 70: 231–283
 - 68 Hanin B. Universal function approximation by deep neural nets with bounded width and ReLU activations. *Mathematics*, 2019, 7: 992
 - 69 He K, Zhang X, Ren S, et al. Deep residual learning for image recognition. In: *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Las Vegas: IEEE, 2016, 770–778

- 70 Heeringa T J, Spek L, Schwenninger F L, et al. Embeddings between Barron spaces with higher-order activation functions. *Appl Comput Harmon Anal*, 2024, 73: 101691
- 71 Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models. In: *Advances in Neural Information Processing Systems*, vol. 33. Red Hook: Curran Associates, 2020, 6840–6851
- 72 Hörmander L. *Linear Partial Differential Operators*. Berlin: Springer-Verlag, 1963
- 73 Huang G B, Babri H A. Upper bounds on the number of hidden neurons in feedforward networks with arbitrary bounded nonlinear activation functions. *IEEE Trans Neural Netw*, 1998, 9: 224–229
- 74 Huybrechs D, Trefethen L N. Sigmoid functions, multiscale resolution of singularities, and hp -mesh refinement. *SIAM Rev*, 2024, 66: 683–693
- 75 Ismailov V E. On the approximation by neural networks with bounded number of neurons in hidden layers. *J Math Anal Appl*, 2014, 417: 963–969
- 76 Ito Y, Saito K. Superposition of linearly independent functions and finite mappings by neural networks. *Math Sci*, 1996, 21: 27–33
- 77 Jiang H T, Li Q X. Approximation rate of the transformer architecture for sequence modeling. In: *Advances in Neural Information Processing Systems*, vol. 37. Red Hook: Curran Associates, 2024, 68926–68955
- 78 Jiao Y L, Lai Y M, Lu X L, et al. Deep neural networks with ReLU-Sine-exponential activations break curse of dimensionality in approximation on Hölder class. *SIAM J Math Anal*, 2023, 55: 3635–3649
- 79 Jiao Y L, Lai Y M, Wang Y, et al. Transformers can overcome the curse of dimensionality: A theoretical study from an approximation perspective. *J Mach Learn Res*, 2026, 27: 1–34
- 80 Jiao Y L, Wang Y, Yang Y F. Approximation bounds for norm constrained neural networks with applications to regression and GANs. *Appl Comput Harmon Anal*, 2023, 65: 249–278
- 81 Jones L K. A simple lemma on greedy approximation in Hilbert space and convergence rates for projection pursuit regression and neural network training. *Ann Statist*, 1992, 20: 608–613
- 82 Kainen P C, Kůrková V, Vogt A. Approximation by neural networks is not continuous. *Neurocomputing*, 1999, 29: 47–56
- 83 Karpinski M, Macintyre A. Polynomial bounds for VC dimension of sigmoidal and general Pfaffian neural networks. *J Comput System Sci*, 1997, 54: 169–176
- 84 Kidger P, Lyons T. Universal approximation with deep narrow networks. In: *Proceedings of the Thirty-Third Conference on Learning Theory*. Ann Arbor: PMLR, 2020, 2306–2327
- 85 Kingma D P, Welling M. Auto-encoding variational Bayes. In: *Proceedings of Machine Learning Research*, vol. 125. Ann Arbor: PMLR, 2020, 2306–2327
- 86 Klusowski J M, Barron A R. Approximation by combinations of ReLU and squared ReLU ridge functions with ℓ^1 and ℓ^0 controls. *IEEE Trans Inform Theory*, 2018, 64: 7649–7656
- 87 Kolmogorov A N. O predstavlenii nepreryvnykh funktsii mnogikh peremennykh superpozitsiyami nepreryvnykh funktsii odnogo peremennogo i slozheniem. *Dokl Akad Nauk SSSR*, 1957, 114: 953–956
- 88 Kovachki N, Lanthaler S, Mishra S. On universal approximation and error bounds for Fourier neural operators. *J Mach Learn Res*, 2021, 22: 1–76
- 89 Kovachki N, Li Z, Liu B, et al. Neural operator: Learning maps between function spaces with applications to PDEs. *J Mach Learn Res*, 2023, 24: 1–97
- 90 Lanthaler S. Operator learning with PCA-Net: Upper and lower complexity bounds. *J Mach Learn Res*, 2023, 24: 1–67
- 91 Lanthaler S, Mishra S, Karniadakis G E. Error estimates for DeepONets: A deep learning framework in infinite dimensions. *Trans Math Appl*, 2022, 6: tnac001
- 92 Le Magoarou L, Gribonval R. Flexible multilayer sparse approximations of matrices and applications. *IEEE J Sel Top Signal Process*, 2016, 10: 688–700
- 93 Lee H, Ge R, Ma T, et al. On the ability of neural nets to express distributions. In: *Proceedings of the 2017 Conference on Learning Theory*. *Proceedings of Machine Learning Research*, vol. 65. Ann Arbor: PMLR, 2017, 1271–1296
- 94 Lei Z, Lin F H. Global mild solutions of Navier-Stokes equations. *Comm Pure Appl Math*, 2011, 64: 1297–1304
- 95 Leshno M, Lin V Y, Pinkus A, et al. Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. *Neural Networks*, 1993, 6: 861–867
- 96 Lévy P. Sur la convergence absolue des séries de Fourier. *Compos Math*, 1935, 1: 1–14
- 97 Li Y, Lu S, Mathé P, et al. Two-layer networks with the ReLU^k activation function: Barron spaces and derivative approximation. *Numer Math*, 2024, 156: 319–344
- 98 Li Z, Kovachki N, Azzadenesheli K, et al. Fourier neural operator for parametric partial differential equations. In: *Proceedings of the International Conference on Learning Representations*, 2021, <https://openreview>.

- net/forum?id=c8P9NQVtmnO
- 99 Liao Y, Ming P. Spectral Barron space for deep neural network approximation. *SIAM J Math Data Sci*, 2025, 7: 1053–1076
 - 100 Lin H, Jegelka S. ResNet with one-neuron hidden layers is a universal approximator. In: *Advances in Neural Information Processing Systems*, vol. 31. Red Hook: Curran Associates, 2018, 6169–6178
 - 101 Liu Y H, Wang Z L, Wu L, et al. Smoothness adaptivity in constant-depth neural networks: Optimal rates via smooth activations. *arXiv:2602.19691v2*, 2026
 - 102 Liu Z, Wang Y, Vaidya S, et al. KAN: Kolmogorov-Arnold networks. In: *Proceedings of the Thirteenth International Conference on Learning Representations*, 2025, <https://openreview.net/forum?id=Ozo7qJ5vZi>
 - 103 Long J, Han J. Reinforcement learning with function approximation: From linear to nonlinear. *J Mach Learn*, 2023, 2: 161–193
 - 104 Lorentz G G, Golitschek M V, Makovoz Y. *Constructive Approximation: Advanced Problems*. New York: Springer, 1996
 - 105 Lu J F, Shen Z W, Yang H Z, et al. Deep network approximation for smooth functions. *SIAM J Math Anal*, 2021, 53: 5465–5506
 - 106 Lu L, Jin P, Pang G, et al. Learning nonlinear operators via DeepONet based on the universal approximation theorem of operators. *Nat Mach Intell*, 2021, 3: 218–229
 - 107 Lu Y L, Lu J F. A universal approximation theorem of deep neural networks for expressing probability distributions. In: *Advances in Neural Information Processing Systems*, vol. 33. Red Hook: Curran Associates, 2020, 3094–3105
 - 108 Lu Z, Pu H, Wang F, Hu Z, Wang L. The expressive power of neural networks: A view from the width. In: *Advances in Neural Information Processing Systems*, vol. 30. Red Hook: Curran Associates, 2017, 6231–6239
 - 109 Ma L, Siegel J W, Xu J. Uniform approximation rates and metric entropy of shallow neural networks. *Res Math Sci*, 2022, 9: 46
 - 110 Maiorov V E. On best approximation by ridge functions. *J Approx Theory*, 1999, 99: 68–94
 - 111 Maiorov V E. Best approximation by ridge functions in L_p -spaces. *Ukrain Mat Zh*, 2010, 62: 452–466
 - 112 Maiorov V, Meir R, Ratsaby J. On the approximation of functional classes equipped with a uniform measure using ridge functions. *J Approx Theory*, 1999, 99: 95–111
 - 113 Maiorov V, Pinkus A. Lower bounds for approximation by MLP neural networks. *Neurocomputing*, 1999, 25: 81–91
 - 114 Makovoz Y. Random approximants and neural networks. *J Approx Theory*, 1996, 85: 98–109
 - 115 Makovoz Y. Uniform approximation by neural networks. *J Approx Theory*, 1998, 95: 215–228
 - 116 Mao T, Siegel J W, Xu J. Approximation rates for shallow ReLU^k neural networks on Sobolev spaces via the Radon transform. *SIAM J Math Anal*, 2026, 58: 1171–1186
 - 117 Martens J, Chattopadhyaya A, Pitassi T, et al. On the representational efficiency of restricted Boltzmann machines. In: *Advances in Neural Information Processing Systems*, vol. 26. Red Hook: Curran Associates, 2013, 2877–2885
 - 118 Matoušek J. Tight upper bounds for the discrepancy of half-spaces. *Discrete Comput Geom*, 1995, 13: 593–601
 - 119 Matoušek J. Improved upper bounds for approximation by zonotopes. *Acta Math*, 1996, 177: 55–73
 - 120 Matoušek J. *Geometric Discrepancy*. Berlin: Springer-Verlag, 1999
 - 121 McCulloch W S, Pitts W. A logical calculus of the ideas immanent in nervous activity. *Bull Math Biophys*, 1943, 5: 115–133
 - 122 Meng Y, Ming P. A new function space from Barron class and application to neural network approximation. *Commun Comput Phys*, 2022, 32: 1361–1400
 - 123 Mhaskar H N. Approximation properties of a multilayered feedforward artificial neural network. *Adv Comput Math*, 1993, 1: 61–80
 - 124 Mhaskar H N. Neural networks for optimal approximation of smooth and analytic functions. *Neural Comput*, 1996, 8: 164–177
 - 125 Mhaskar H, Liao Q, Poggio T. When and why are deep networks better than shallow ones? In: *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*. Palo Alto: AAAI Press, 2017, 2343–2349
 - 126 Mhaskar H, Micchelli C A. Approximation by superposition of sigmoidal and radial basis functions. *Adv Appl Math*, 1992, 13: 350–373
 - 127 Mikhailets V A, Murach A A. *Hörmander Spaces, Interpolation, and Elliptic Problems*. Berlin-Boston: Walter de Gruyter, 2014
 - 128 Ming P, Yu H. Barron regularity of many particle Schrödinger eigenfunctions. *arXiv:2508.17722*, 2025
 - 129 Montanelli H, Du Q. New error bounds for deep ReLU networks using sparse grids. *SIAM J Math Data Sci*, 2019, 1: 78–92
 - 130 Montúfar G, Pascanu R, Cho K, et al. On the number of linear regions of deep neural networks. In: *Advances in Neural Information Processing Systems*, vol. 27. Red Hook: Curran Associates, 2014, 2924–2932

- 131 Morris S A. Hilbert 13: Are there any genuine continuous multivariate real-valued functions? *Bull Amer Math Soc (NS)*, 2021, 58: 107–118
- 132 Neyshabur B, Tomioka R, Srebro N. Norm-based capacity control in neural networks. In: *Proceedings of the 28th Conference on Learning Theory. Proceedings of Machine Learning Research*, vol. 40. Ann Arbor: PMLR, 2015, 1376–1401
- 133 Opschoor J A A, Schwab C. Exponential expressivity of ReLU^k neural networks on Gevrey classes with point singularities. *Appl Math*, 2024, 69: 695–724
- 134 Opschoor J A A, Schwab C, Zech J. Exponential ReLU DNN expression of holomorphic maps in high dimension. *Constr Approx*, 2022, 55: 537–582
- 135 Oswald P. On the degree of nonlinear spline approximation in Besov-Sobolev spaces. *J Approx Theory*, 1990, 61: 131–157
- 136 Ou W T, Bölcskei H. Covering numbers for deep ReLU networks with applications to function approximation and nonparametric regression. *Found Comput Math*, 2026, in press
- 137 Papamakarios G, Nalisnick E, Rezende D, et al. Normalizing flows for probabilistic modeling and inference. *J Mach Learn Res*, 2021, 22: 1–64
- 138 Parhi R. On ridge splines, neural networks, and variational problems in Radon-domain BV spaces. PhD Thesis. Madison: University of Wisconsin-Madison, 2022
- 139 Parhi R, Nowak R D. Banach space representer theorems for neural networks and ridge splines. *J Mach Learn Res*, 2021, 22: 43
- 140 Parhi R, Nowak R D. What kinds of functions do deep neural networks learn? Insights from variational spline theory. *SIAM J Math Data Sci*, 2022, 4: 464–489
- 141 Parhi R, Nowak R D. Near-minimax optimal estimation with shallow ReLU neural networks. *IEEE Trans Inform Theory*, 2023, 69: 1125–1140
- 142 Parhi R, Nowak R D. Compositional function spaces for deep learning. *SIAM Rev*, 2026, 68: 127–149
- 143 Park S, Yun C, Lee J, Shin J. Minimum width for universal approximation. In: *International Conference on Learning Representations*, 2021, <https://openreview.net/forum?id=O-XJwyoIF-k>
- 144 Petersen P, Voigtlaender F. Optimal learning of high-dimensional classification problems using deep neural networks. *arXiv:2112.12555*, 2021
- 145 Petersen P, Zech J. Mathematical theory of deep learning. *arXiv:2407.18384v4*, 2026
- 146 Petrushev P P. Approximation by ridge functions and neural networks. *SIAM J Math Anal*, 1998, 30: 155–189
- 147 Pinkus A. n -widths in Approximation Theory. Berlin: Springer-Verlag, 1985
- 148 Pinkus A. Approximation theory of the MLP model in neural networks. *Acta Numer*, 1999, 8: 143–195
- 149 Pinkus A. Ridge Functions. Cambridge: Cambridge Univ Press, 2015
- 150 Pisier G. Remarques sur un résultat non publié de B. Maurey. In: *Seminar on Functional Analysis*. Palaiseau: École Polytechnique, Centre de Mathématiques, 1981, 13
- 151 Powell W B. Approximate Dynamic Programming: Solving the Curses of Dimensionality, 2nd ed. Hoboken: Wiley, 2011
- 152 Rezende D, Mohamed S. Variational inference with normalizing flows. In: *Proceedings of the 32nd International Conference on Machine Learning. Proceedings of Machine Learning Research*, vol. 37. Ann Arbor: PMLR, 2015, 1530–1538
- 153 Ronneberger O, Fischer P, Brox T. U-Net: Convolutional networks for biomedical image segmentation. In: *Medical Image Computing and Computer-Assisted Intervention-MICCAI 2015. Lecture Notes in Computer Science*, vol. 9351. Cham: Springer, 2015, 234–241
- 154 Schmitt M. RBF neural networks and Descartes’ rule of signs. In: *Algorithmic Learning Theory. Lecture Notes in Computer Science*, vol. 2533. Berlin: Springer, 2002, 321–335
- 155 Shen Z W, Yang H Z, Zhang S J. Nonlinear approximation via compositions. *Neural Networks*, 2019, 119: 74–84
- 156 Shen Z W, Yang H Z, Zhang S J. Deep network approximation characterized by number of neurons. *Commun Comput Phys*, 2020, 28: 1768–1811
- 157 Shen Z W, Yang H Z, Zhang S J. Neural network approximation: Three hidden layers are enough. *Neural Networks*, 2021, 141: 160–173
- 158 Shen Z W, Yang H Z, Zhang S J. Deep network approximation: Achieving arbitrary accuracy with fixed number of neurons. *J Mach Learn Res*, 2022, 23: 1–60
- 159 Shen Z W, Yang H Z, Zhang S J. Optimal approximation rate of ReLU networks in terms of width and depth. *J Math Pures Appl (9)*, 2022, 157: 101–135
- 160 Siegel J W. Optimal approximation rates for deep ReLU neural networks on Sobolev and Besov spaces. *J Mach Learn Res*, 2023, 24: 1–52

- 161 Siegel J W. Optimal approximation of zonoids and uniform approximation by shallow neural networks. *Constr Approx*, 2025, 62: 441–469
- 162 Siegel J W, Xu J. Approximation rates for neural networks with general activation functions. *Neural Networks*, 2020, 128: 313–321
- 163 Siegel J W, Xu J. High-order approximation rates for shallow neural networks with cosine and ReLU activation functions. *Appl Comput Harmon Anal*, 2022, 58: 1–26
- 164 Siegel J W, Xu J. Characterization of the variation spaces corresponding to shallow neural networks. *Constr Approx*, 2023, 57: 1109–1132
- 165 Siegel J W, Xu J. Sharp bounds on the approximation rates, metric entropy, and n -widths of shallow neural networks. *Found Comput Math*, 2024, 24: 481–537
- 166 Song Y, Sohl-Dickstein J, Kingma D P, et al. Score-based generative modeling through stochastic differential equations. In: *Proceedings of the International Conference on Learning Representations*, 2021, <https://openreview.net/forum?id=PXTIG12RRHS>
- 167 Suzuki T. Adaptivity of deep ReLU network for learning in Besov and mixed smooth Besov spaces: Optimal rate and curse of dimensionality. In: *International Conference on Learning Representations*, 2019, <https://arxiv.org/abs/1810.08033>
- 168 Telgarsky M. Representation benefits of deep feedforward networks. *arXiv:1509.08101*, 2015
- 169 Telgarsky M. Benefits of depth in neural networks. In: *29th Annual Conference on Learning Theory. Proceedings of Machine Learning Research*, vol. 49. Ann Arbor: PMLR, 2016, 1517–1539
- 170 Telgarsky M. Neural networks and rational functions. In: *Proceedings of the 34th International Conference on Machine Learning. Proceedings of Machine Learning Research*, vol. 70. Ann Arbor: PMLR, 2017, 3387–3393
- 171 Triebel H. *Theory of Function Spaces*. Basel: Birkhäuser, 1983
- 172 Tsuji K, Suzuki T. Estimation error analysis of deep learning on the regression problem on the variable exponent Besov space. *Electron J Stat*, 2021, 15: 1869–1908
- 173 Vardi G, Yehudai G, Shamir O. Width is less important than depth in ReLU neural networks. In: *Proceedings of the Thirty-Fifth Conference on Learning Theory. Proceedings of Machine Learning Research*, vol. 178. Ann Arbor: PMLR, 2022, 1249–1281
- 174 Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. Red Hook: Curran Associates, 2017, 6000–6010
- 175 Vostretsov B A, Kreines M A. Approximation of continuous functions by superpositions of plane waves. *Dokl Akad Nauk SSSR*, 1961, 140: 1237–1240
- 176 Wang S, Wang H, Perdikaris P. Learning the solution operator of parametric partial differential equations with physics-informed DeepONets. *Sci Adv*, 2021, 7: eabi8605
- 177 Wiener N. Tauberian theorems. *Ann of Math (2)*, 1932, 33: 1–100
- 178 Wojtowytsch S. Optimal bump functions for shallow ReLU networks: Weight decay, depth separation, curse of dimensionality. *J Mach Learn Res*, 2024, 25: 1–49
- 179 Wu L. Embedding inequalities for Barron-type spaces. *J Mach Learn*, 2023, 2: 259–270
- 180 Xu J. Finite neuron method and convergence analysis. *Commun Comput Phys*, 2020, 28: 1707–1745
- 181 Yang L, Zhang Z, Song Y, et al. Diffusion models: A comprehensive survey of methods and applications. *ACM Comput Surv*, 2024, 56: 1–39
- 182 Yang Y F, Lu Y L. Near-optimal deep neural network approximation for Korobov functions with respect to L^p and H^1 norms. *Neural Networks*, 2024, 180: 106702
- 183 Yang Y F, Zhou D X. Optimal rates of approximation by shallow ReLU^k neural networks and applications to nonparametric regression. *Constr Approx*, 2025, 62: 329–360
- 184 Yarotsky D. Error bounds for approximations with deep ReLU networks. *Neural Networks*, 2017, 94: 103–114
- 185 Yarotsky D. Optimal approximation of continuous functions by very deep ReLU networks. In: *Proceedings of the 31st Conference on Learning Theory. Proceedings of Machine Learning Research*, vol. 75. Ann Arbor: PMLR, 2018, 639–649
- 186 Yarotsky D. Elementary superexpressive activations. In: *Proceedings of the 38th International Conference on Machine Learning. Proceedings of Machine Learning Research*, vol. 139. Ann Arbor: PMLR, 2021, 11932–11940
- 187 Yarotsky D, Zhevnerchuk A. The phase diagram of approximation rates for deep neural networks. In: *Advances in Neural Information Processing Systems*, vol. 33. Red Hook: Curran Associates, 2020, 13005–13015
- 188 Yukich J E, Stinchcombe M B, White H. Sup-norm approximation bounds for networks through probabilistic methods. *IEEE Trans Inform Theory*, 1995, 41: 1021–1027
- 189 Yun C, Chang Y-W, Bhojanapalli S, et al. $O(n)$ connections are expressive enough: Universal approximability of sparse transformers. In: *Advances in Neural Information Processing Systems*, vol. 33. Red Hook: Curran Associates,

- 2020, 13783–13794
- 190 Zaheer M, Kottur S, Ravanbakhsh S, et al. Deep sets. In: Advances in Neural Information Processing Systems, vol. 30. Red Hook: Curran Associates, 2017, 3391–3401
- 191 Zhang L F, Han J, Wang H, et al. End-to-end symmetry preserving inter-atomic potential energy model for finite and extended systems. In: Advances in Neural Information Processing Systems, vol. 31. Red Hook: Curran Associates, 2018, 4436–4446
- 192 Zhang S J, Lu J F, Zhao H K. Deep network approximation: Beyond ReLU to diverse activation functions. J Mach Learn Res, 2024, 25: 1–39
- 193 Zhao R G, Tang C Q, Wu S Q, et al. Logarithmic Barron spectrum spaces and their applications. Anal Appl (Singap), 2026, in press

Approximation theory of neural networks: From classical theory to high dimensions

Pingbing Ming & Hao Yu

Abstract The success of neural networks in high-dimensional problems has made their approximation theory a key intersection of approximation theory and machine learning. In this paper, we review recent theoretical progress in neural network approximation. We cover classical results, including the universal approximation theorem and quantitative estimates in standard function spaces, and then address the curse of dimensionality by discussing the challenges and recent advances in high-dimensional approximation.

Keywords neural network, universal approximation, high-dimensional approximation, Barron function class, depth separation

MSC(2020) 41-02, 41A46, 46E35, 65-02, 68T99

doi: 10.1360/SSM-2026-0097