arXiv:2301.04605v1 [cs.LG] 11 Jan 2023

探索乘法神经网络对光滑函数的逼近能力

Ido Ben-Shaul *ido.benshaul@gmail.com*
特拉维夫大学应用数学系，以色列；
eBay 研究院

Tomer Galanti *tomer22g@gmail.com*
麻省理工学院（*MIT*）剑桥，马萨诸塞州，美国

Shai Dekel *shaidekel6@gmail.com*
特拉维夫大学应用数学系，以色列

摘要

乘法层是多种重要神经网络模块的关键组成部分，包括自注意力和超网络层。本文研究了中间神经元通过简单乘法运算连接的深度神经网络的逼近能力。我们考虑两类目标函数：广义带限函数，这类函数常用于对具有有限带宽的现实世界信号进行建模；以及 Sobolev 型球，它们嵌入在 Sobolev 空间 $W^{r,2}$ 中。我们的结果表明，与标准 ReLU 神经网络相比，乘法神经网络能够以显著更少的层和神经元逼近这些函数，同时兼顾输入维度和逼近误差。这些发现表明，乘性门可以胜过标准前馈层，并有望改进神经网络设计。

1 引言

近年来，使用大型神经网络进行深度学习，在解决各类任务方面取得了巨大成功，包括图像分类（He et al., 2016; Dosovitskiy et al., 2021；Zhai et al., 2021）、语言处理（Vaswani et al., 2017；Devlin et al., 2019；Brown et al., 2020）、与开放式环境交互（Silver et al., 2016；Arulkumaran et al., 2019）以及代码合成（Chen et al., 2021）。

近期的实证研究表明，在中间神经元之间引入乘法运算的神经网络（Durbin & Rumelhart, 1989; Urban & van der Smagt, 2015; Trask et al., 2018 ），例如自注意力层（ Vaswani et al., 2017 ）和超网络（ Ha et al., 2016; Krueger et al., 2017; Littwin & Wolf, 2019），效果尤为出色。例如，自注意力层已广泛应用于计算机视觉（Dosovitskiy et al., 2021; Zhai et al., 2021）和语言处理（Cheng et al., 2016; Parikh et al., 2016; Paulus et al., 2017; Vaswani et al., 2017），并取得了显著成功。研究还表明，即使不使用非线性激活函数，Transformer 也能取得相当不错的表现（Levine et al., 2020）。此外，超网络通过乘法运算生成网络权重，似乎能够提升神经网络在各种元学习任务中的表现（von Oswald et al., 2020; Littwin & Wolf, 2019; Bensadoun et al., 2021）。然而，从理论角度来看，乘法层的优势尚未得到充分理解。

在这项工作中，我们研究带有乘法层的神经网络的表达能力。具体而言，我们希望评估在特定架构下，将给定函数近似到给定误差容限所需的神经元数量和层数。深度学习理论中的一项经典结果表明，只需一个隐藏层，神经网络就能近似任意光滑的目标函数，这一性质称为通用逼近性质（Cybenko, 1989; Hornik et al., 1989; ichi Funahashi, 1989; Leshno et al., 1991）。然而，这些论文并未具体说明，为达到给定的准确度，需要采用何种架构以及多少参数。这是一个关键问题，因为对这些资源的要求过高，可能会限制神经网络的通用性，并解释其在某些实际应用中的成效有限。

先前的研究表明，索博列夫空间中的函数可以用具有解析激活函数的单隐藏层神经网络进行逼近（Mhaskar，1996）。然而，要以至多为 ϵ 的误差在 L_∞ 范数下逼近这些函数，所需的神经元数量随$\epsilon^{-d/r}$按 O($\epsilon^{-d/r}$) 的量级增长，其中d 是输入维度， r 是目标函数的光滑度阶数， $\epsilon > 0$ 是误差率。这引出了一个问题：维数灾难——即模型复杂度随输入维度呈指数增长的现象——是否是神经网络固有的特性。

另一方面，DeVore 等人（1989）证明，任何能以 ϵ 的误差逼近所有阶数为 r 、维数为 d 的索оболев函数的连续函数逼近器，在L_∞ 范数下都至少需要 Ω($\epsilon^{-d/r}$) 个参数。这一结果与 Mhaskar（1996）的界相符，并证实，对于索博列夫空间中的函数，在 L_∞ 范数下进行逼近时，神经网络无法避免维数灾难。一个关键问题是，神经网络能否针对某些目标函数集合克服维数灾难，以及何种架构能够为逼近这些函数提供最佳保证。

为克服维数灾难，各类研究（Mhaskar 等，2017年；Poggio 等，2020年；Kohler & Krzyżak，2017年；Montanelli & Du，2019；Blanchard & Bennouna，2022b；Galanti & Wolf，2020）探讨了神经网络在表示其他类别函数时的逼近能力，并采用了较弱的距离概念，例如 L_2 距离。例如，Mhaskar 等（2017年）和 Poggio 等（2020年）证明，具有 r光滑度的光滑组合稀疏函数，可以用深度神经网络以 O($d\epsilon^{-2/r}$) 个神经元逼近，使 L_∞ 距离下的误差达到 ϵ 。其他结构约束也被应用于具有结构化输入空间的函数（Mhaskar，2010；Nakada & Imaizumi，2022；Schmidt-Hieber，2019）、函数复合（Kohler & Krzyżak，2017）以及分段光滑函数（Petersen & Voigtländer，2017；Imaizumi & Fukumizu，2018）。另一类研究则着眼于理解在正则性约束下，特定神经网络架构能够实现哪些类型的函数。例如，E 等（2021）证明，当权重大小受到限制时，2 层神经网络的空间等价于 Barron空间。他们进一步证明，Barron函数可以用 2 层网络以 O(ϵ^{-2}) 个神经元逼近至 ϵ 以内。另有研究考察函数空间上的谱条件，使函数能够表示为浅层网络的无限宽极限（Barron，1991；Klusowski & Barron，2016）。（Blanchard & Bennouna，2022b）研究了 Korobov 函数空间；这类函数在求解偏微分方程（PDEs）时具有实际用途。他们证明，任意 Korobov 函数均可由一个采用 ReLU激活函数的 2 层神经网络逼近：该网络含有 O(ϵ^{-1} log$(1/\epsilon)^{1.5(d-1)+1}$) 个神经元时， L_2 距离下的误差可达到 ϵ ；此外，深度为 O(log(d)) 的网络可使用 O($\epsilon^{-0.5}$ log$(1/\epsilon)^{1.5(d-1)+1}$) 个神经元。

在一篇近期论文中，Montanelli 等人（2021）为广义带限函数建立了逼近保证。这类函数通常用于建模频率范围有限的信号（例如波浪、视频和音频信号），这种特性称为有限带宽。物理学中许多偏微分方程的解都是带限函数，因为这些偏微分方程所描述的物理现象通常具有有限的频率范围。例如，描述波传播的波动方程，其解就是带限函数。Montanelli 等人（2021）证明，任意带限函数都可以使用深度为 O(log$^2(1/\epsilon)$)、含有 O(ϵ^{-2}log$^2(1/\epsilon)$) 个神经元的 ReLU 神经网络，在 L_2 距离下以误差 ϵ 逼近。

本文研究乘法神经网络架构在 L_2距离下的逼近能力。具体而言，我们证明，深度为 O(log($\frac{1}{\epsilon}$))、具有 O(ϵ^{-2} log($\frac{1}{\epsilon}$))

个神经元的乘法神经网络，可以将任意广义带限函数逼近到误差 ϵ 以内（常数取决于维数和频带）。此外，我们还研究了神经网络逼近阶数为 r 的 Sobolev 型球中函数时的逼近保证。我们证明，对于相同的误差容限 ϵ，乘法神经网络只需深度 O($d^2\epsilon^{-1/r}$) 和 O($d^2\epsilon^{-(1+1/r)}$) 个神经元即可逼近这些函数，而标准 ReLU 神经网络则需要深度 O($d^2\epsilon^{-2/r}$) 和 O($d^2\epsilon^{-(2+2/r)}$) 个神经元。这些结果展现了乘性门相较于标准全连接层的优越性能。表 1 对比了本文的新界与已有界，这些界涉及神经网络对 Sobolev 空间、带限函数以及 Sobolev 型球的逼近能力。

空间	模型	# 神经元	深度	参考文献
$\mathcal{W}^{r,p}$	C^∞，非多项式	O($\epsilon^{-d/r}$)	$\mathcal{O}(1)$	(Mhaskar，1996)
$\mathcal{W}^{r,\infty}$	ReLU	O($\epsilon^{-d/r}\log\frac{1}{\epsilon}$)	O($\log\frac{1}{\epsilon}$)	(Yarotsky，2017年)
带限函数	ReLU	O($\epsilon^{-2}\log^2\frac{1}{\epsilon}$)	O($\log^2\frac{1}{\epsilon}$)	(Montanelli 等人，2021年)
带限函数	乘法型	O($\epsilon^{-2}\log\frac{1}{\epsilon}$)	O($\log\frac{1}{\epsilon}$)	本文
$\mathcal{Z}_{2r,\,2}\subsetneq\mathcal{W}^{r,\,2}$	ReLU	O($d^2\epsilon^{-(2+2/r)}$)	O($d^2\epsilon^{-2/r}$)	本文
$\mathcal{B}_{2r,\,2}\subsetneq\mathcal{W}^{r,\,2}$	乘法型	O($d\epsilon^{-(2+1/r)}$)	O($d\epsilon^{-1/r}$)	本文
$\mathcal{B}_{1,1}$	S形	O($d\epsilon^{-2}$)	$\mathcal{O}(1)$	(Barron，1993年)

表 1：ReLU 和乘法神经网络对 Sobolev $W^{r,p}$、带限以及 $B_{2r,2}$ 函数的逼近结果。神经元数量和深度以 O 记号表示。

2 问题设置

我们关注的是，为了从理论上保证对未知目标函数f 的逼近误差不超过给定值 $\epsilon > 0$，模型应当具有多大的复杂度（即可训练参数、神经元和层的数量）。

形式上，我们考虑一个函数巴拿赫空间 V（例如，$L_p([0,1]^d)$），其配备范数 $\|\cdot\|_{\mathcal{V}}$（例如，$\|\cdot\|_{L_p([0,1]^d)}$），以及一组目标函数 U $\subseteq$ V。我们还考虑一组逼近器 H，并希望通过以下量来量化这些逼近器逼近 U 的能力

$$d_{\mathcal{V}}(\mathcal{H},\mathcal{U}) \;=\; \inf_{\hat{f}\in\mathcal{H}}\sup_{f\in\mathcal{U}}\|\hat{f}-f\|_{\mathcal{V}},$$

它衡量使用来自 H 的候选函数 $\hat{f}$ 逼近目标函数 $f \in$ U 时的最大逼近误差。通常，H 是一个函数参数化集合（例如，具有特定架构的神经网络），我们用 $\hat{f}_\theta \in$ H 表示由参数向量 $\theta \in \mathbb{R}^N$ 参数化的函数。为简洁起见，在上下文明确时，我们省略下标中的 θ。

2.1 目标函数空间

正如（Devroye 等，1996）中的定理7.2所示，使用标准神经网络通常无法逼近任意目标函数。因此，为了获得有意义的结果，我们常考虑满足特定光滑性假设的目标函数空间。本文具体关注满足以下光滑性假设的目标函数 U。本文将重点研究单位立方体 $B = [0,1]^d$ 上的函数逼近，其中 $|B| = 1$ 表示 B 的体积。

索博列夫空间。索博列夫空间是逼近理论中研究最为广泛的函数类之一（DeVore & Lorentz, 1993年；Yarotsky, 2017年；Liang & Srikant, 2017年）。这些空间由导数有界且阶数不超过某一特定阶数的光滑函数构成，在偏微分方程（偏微分方程）的研究中尤其有用。

我们首先定义给定函数 $f : \Omega \to \mathbb{R}$ 的 L_p 范数为 $\|f\|_{L_p(\Omega)} = (\int_\Omega |f(x)|^p \,\mathrm{d}x)^{1/p}$，其中 Ω 是配备了 σ-代数 Σ 和测度 μ的可测空间。若函数 f 满足 $\|f\|_{L_p(\Omega)} < \infty$，则称其属于 $L_p(\Omega)$。

设 $r \in \mathbb{N}$ 和 $p \in [1, \infty)$。Sobolev空间 $W^{r,p}(B)$由满足以下条件的函数 $f: B \to \mathbb{R}$ 组成：其 r-阶分布导数属于 L_p。Sobolev范数 $\|\cdot\|_{\mathcal{W}^{r,p}(B)}$ 定义为

$$\|f\|_{\mathcal{W}^{r,p}(B)} = \sum_{\alpha:|\alpha|_1 \le r} \|D^\alpha f\|_{L_p(B)},$$

其中 $\alpha = (\alpha_1, \ldots, \alpha_d) \in \{0, \ldots, r\}^d$, $|\alpha|_1 = \alpha_1 + \cdots + \alpha_d$, $D^\alpha f$是相应的分布导数。当 $p = \infty$ 时，使用本性上确界范数。我们还给出半范数：

$$|f|_{r,p} = \sum_{\alpha \,:\, |\alpha| = r} \|D^\alpha f(x)\|_{L_p(B)}.$$

文献中的经典结果表明，将 $W^{r,\infty}$ 中的函数近似到误差 ϵ 所需的参数数量下界为 $\Omega(\epsilon^{-d/r})$（DeVore等人，1989年）。这种对 d 的指数依赖被称为"维数灾难"。

广义带限函数。广义带限函数是指其频谱（即构成该函数的一组频率）被限制在某个特定频带或频率范围内的函数。这一性质使带限函数适用于某些应用，例如信号处理；在这些应用中，确保信号不含某个特定范围以外的频率非常重要。

广义带限函数通过其广义傅里叶变换来定义，该变换由解析函数 $K : \mathbb{R} \to \mathbb{C}$ （称为核）和标量 $M \ge 1$ （频带）给出。形式上，对于给定函数 $f: \mathbb{R}^d \to \mathbb{R}$，我们将所有满足以下条件的函数 $F: [-M, M]^d \to \mathbb{C}$ 的集合记为：它们利用核 K 从频率 $\omega \in [-M, M]^d$ 中还原出 f：

$$S_{f,K} = \left\{ F : [-M, M]^d \to \mathbb{R} \mid f(x) = \int_{[-M,M]^d} F(\omega) K(\omega \cdot x) \, \mathrm{d}\omega \right\}.$$

我们将 $F_f = \arg\min_{F \in S_{f,K}} \|F\|_{L^2([-M,M]^d)}$ 定义为 $S_{f,K}$ 中范数最小的 L_2 范数函数。例如，当 $K(x) = \exp(ix)$ 时，函数 F_f 对应于 f 的归一化标准傅里叶变换，其表达式为 $F = \frac{1}{(2\pi)^d}(\mathcal{F} f)(\omega) = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} f(x) \exp(-i\omega \cdot x) \, \mathrm{d}x$。

广义带限函数空间 $H_{K,M}(B)$ 是一个希尔伯特函数空间，其中的函数可表示为定义在有限域上的函数 K 的加权和。该空间配备了内积和范数，分别用于衡量这些函数的相似性和大小。我们将 $H_{K,M}(B)$ 定义为满足以下条件的函数 $f: B \to \mathbb{R}$

$$\mathcal{H}_{K,M}(B) = \left\{ \forall x \in B, f(x) = \int_{[-M,M]^d} F(\omega) K(\omega \cdot x) \, \mathrm{d}\omega \mid F : [-M, M]^d \to \mathbb{C} \text{ is in } L^2([-M, M]^d) \right\}.$$

该空间中的内积和范数定义如下：$\langle f, g \rangle_{\mathcal{H}_{K,M}(B)} = \int_{[-M,M]^d} F_f(\omega) \overline{F_g}(\omega) \, \mathrm{d}\omega$以及范数 $\|f\|_{\mathcal{H}_{K,M}(B)} = \|F_f\|_{L^2([-M,M]^d)}$。

广义带限函数的一个关键性质是，它们可以通过一组离散采样点完全重建。这被称为香农-奈奎斯特定理（Shannon，1984），是信号处理和通信领域的一项重要成果。该定理的一个有趣推论是，即使在高维空间中，可能出现看似不可预测的几何现象（例如 Blum 等人（2020），第 2 章），仍然可以根据函数在奈奎斯特频率处的取值将其完美重建。更多详情请参见附录 A。

Sobolev 型球。 Sobolev 型球是满足高阶导数光滑性约束的函数集合（Barron，1993；Jones，1992；Pinkus，1985；Wahba，1990；Blanchard & Bennouna，2022a）。其中一种约束是，函数的傅里叶变换的幅值 $|\mathcal{F} f(\omega)|$ 随着 $|\omega|$ 趋于无穷大而足够快地衰减。这些约束通过比较函数的傅里叶变换 $\mathcal{F} f(\omega)$ 的幅值与某个数 $r > 0$ 对应的 $|\omega|^r$ 的幅值来施加。本文定义了 Sobolev 型球的一种广义形式：

$$\mathcal{B}_{r,\rho} = \left\{ f : \mathbb{R}^d \to \mathbb{R}, f \in L_2(\mathbb{R}^d) : \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} |(\mathcal{F} f)(\omega)| \, \mathrm{d}\omega, \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} |\omega|^r |(\mathcal{F} f)(\omega)|^\rho \, \mathrm{d}\omega \le 1 \right\}.$$

在（Barron，1993）中，他们探讨了神经网络逼近空间 $P_1 = \{f \in L_2(\mathbb{R}^d) \mid \int_{\mathbb{R}^d} |\omega||(\mathrm{F}f)(\omega)| < \infty\}$ 中函数的能力。现在我们证明， P_1 中的任意函数在 $\mathrm{B}^{1,1}$ 中都有一个归一化表示。为此，我们证明，关于 P_1 的条件使我们能够对 $\|\mathrm{F}f\|_{L_1(\mathbb{R}^d)}$ 作出界定。具体而言，

$$\begin{aligned}\int_{\mathbb{R}^d} |(\mathcal{F}f)(\omega)| \, \mathrm{d}\omega &= \int_{[-1,1]^d} |(\mathcal{F}f)(\omega)| \, \mathrm{d}\omega + \int_{\mathbb{R}^d \setminus [-1,1]^d} |(\mathcal{F}f)(\omega)| \, \mathrm{d}\omega \\ &\leq C_1 \left(\int_{[-1,1]^d} |(\mathcal{F}f)(\omega)|^2 \, \mathrm{d}\omega \right)^{1/2} + \int_{\mathbb{R}^d \setminus [-1,1]^d} |\omega||(\mathcal{F}f)(\omega)| \, \mathrm{d}\omega \\ &\leq C_1 \|f\|_2 + \int_{\mathbb{R}^d} |\omega||(\mathcal{F}f)(\omega)| \, \mathrm{d}\omega,\end{aligned}$$

对于 $C_1 > 0$。因此，空间 P_1 中的任意函数都可以缩放为 $\mathrm{B}^{1,1}$中的一个函数。具体而言，他们证明，具有有界深度且包含 $O(d\epsilon^{-2})$ 个神经元的 S 形神经网络能够以至多 ϵ 的误差逼近这类函数。

本文将重点研究空间 $\mathrm{B}^{2r,2}$。我们在附录中的引理 2 证明，该空间是 $\mathrm{W}^{r,2}$的真子空间，其范数如下：

$$\|f\|_{\mathcal{B}_{2r,2}} = \left(\frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} (1 + |\omega|^{2r}) |(\mathcal{F}f)(\omega)|^2 \, \mathrm{d}\omega \right)^{1/2}. \tag{1}$$

值得一提的是，Pinkus（1985年）证明，利用传统基函数，空间 $P_2 = \{f \in L_2(\mathbb{R}^d) \mid \int_{\mathbb{R}^d} |\omega|^{2r} |(\mathrm{F}f)(\omega)|^2 < \infty\}$中的函数可以用 $O(\epsilon^{-d/2r})$ 个参数以至多 ϵ 的误差逼近。本文证明，通过对傅里叶系数的 L_1 范数施加额外约束，我们可以避免对维数 d的指数依赖。

2.2 神经网络架构

在上一节中，我们描述了一种设定，其中一类候选函数 H 用作一类目标函数 U 的逼近器。在本研究中，我们比较标准多层感知机与一类包含乘法运算的通用神经网络的逼近保证。我们的目标是了解乘法层是否能够为逼近带限函数以及 $\mathrm{B}^{2r,\,2}$ 的成员提供更好的保证。

多层感知机。 多层感知机是一种神经网络架构，由 L 层线性变换组成，层间穿插逐元素非线性激活函数（例如 ReLU 函数）。通常，最后一层不包含非线性激活。

定义 1 (多层感知机)。 多层感知机是一个函数 $f = y_{L,\,1}$： $\mathbb{R}^{p_0} \to \mathbb{R}$，由一组单变量函数 $\bigcup_{i=0}^{L} \{y_{i,\,j}\}_{j=1}^{p_i}$ 定义。每个函数 $y_{i,\,j}$： $\mathbb{R}^{p_0} \to \mathbb{R}$ （也称为神经元）按以下方式递归计算

$$\begin{aligned}y_{L,j}(x) &= \langle w_{L,j}, y_{L-1}(x) \rangle + b_{L,j} \\ y_{i,j}(x) &= \sigma(\langle w_{i,j}, y_{i-1}(x) \rangle + b_{i,j}) \\ y_{0,j}(x) &= x_j,\end{aligned}$$

其中， $i \in [L-1]$、 $j \in [p_i]$和$w_{i,\,j} \in \mathbb{R}^{p_{i-1}}$ 以及$b_{i,\,j} \in \mathbb{R}$ 是神经元$y_{i,\,j}$ 的权重和偏置，且$y_i = (y_1$、……、 $y_{p_i})$。函数 σ： $\mathbb{R} \to \mathbb{R}$ 是一个非线性激活函数。

在本文中，我们关注采用 ReLU 激活的神经网络，其定义为 $\sigma(x) = \max(0,\ x)$。不过，值得注意的是，文献中还提出了其他激活函数，例如满足 $\eta(x) \to 0$ （当 $x \to -\infty$ 时）且 $\eta(x) \to 1$ （当$x \to \infty$时）的可测函数 η ，即 S形函数。

乘法神经网络。在这项工作中，我们旨在比较标准 ReLU 网络与引入乘法层的神经网络的逼近能力

（也称为乘积单元（Durbin & Rumelhart，1989））。为了全面了解乘法门带来的额外优势，我们提出以下问题：神经元之间的乘法是否足以替代多层感知机中的非线性激活？

定义 2（乘法网络）。 乘法神经网络是一个函数 $f = y_{L,\,1} : \mathbb{R}^{p_0} \to \mathbb{R}$，由一组单变量函数 $\bigcup_{i=1}^{L}\{y_{i,\,j}\}_{j=1}^{p_i}$ 定义。每个函数 $y_{i,\,j}:\ \mathbb{R}^{p_0} \to \mathbb{R}$ （也称为神经元）按以下方式递归计算

$$\begin{aligned} y_{i,j}(x) &= \langle w_{i,j}, y_{i-1}(x)\rangle + a_{i,j}\, y_{i-1,j_1}(x)\, y_{i-1,j_2}(x) + b_{i,j} \\ y_{0,j}(x) &= x_j, \end{aligned}$$

其中 $a_{i,\,j}, b_{i,\,j} \in \mathbb{R}$, $w_{i,\,j} \in \mathbb{R}^{p_{i-1}}$ 是可训练参数，而 $y_i = (y_1, \ldots, y_{p_i})$。

乘法网络与标准多层感知机有两个主要区别。首先，我们不在中间层之间加入非线性（ReLU 激活）。此外，每个神经元都在前两个神经元之间加入一个乘法门。乘法层中的可训练参数总数为 $p_{i+1}p_i + 2p_i$，比全连接层多 p_i 个参数。遵循 Yarotsky (2017年)；Blanchard & Bennouna (2022a)，我们使用神经网络的深度 L 及其包含的神经元数量 $G = \sum_{i=1}^{L-1} p_i$ 来衡量其复杂度。

自注意力与乘法层。 下面介绍原始 Transformer（Vaswani 等人，2017年）中的单头自注意力操作。深度为 L 的 Transformer 编码器的每一层 $i \in [L]$ 定义如下。第 i 层的输入是由 N 个词元组成的序列，记作 $x_i = \{x_{i,j}\}_{j=1}^{N}$，其中每个 $x_{i,j} \in \mathbb{R}^{d_x}$ 表示第 i 层的第 j 个词元。要计算第 i 层在特定位置 $e \in [N]$ 的输出，我们使用以下公式:

$$f_{\mathrm{SA}}^{i,e}(x_i) \;=\; \sum_{j=1}^{N} \mathrm{softmax}_j\left(\tfrac{1}{\sqrt{d_a}}\langle W^{Q,i}x_{i,e}, W^{K,i}x_{i,e}\rangle\right) W^{V,i}x_{i,j},$$

其中，$\mathrm{softmax}_j(f(x)) = \exp(f(x)_j)/\sum_{j'}\exp(f(x)_{j'})$ 是 softmax 运算符，而可训练权重矩阵 $W^{K,i}, W^{Q,i}, W^{V,i} \in \mathbb{R}^{d_a \times d_x}$ 将表示从其维数 d_x 映射到注意力维度 $d_a = d_x$，分别生成“键”、“查询”和“值”表示。可以看出，自注意力层在计算以下内积 $\langle W^{Q,i}x_{i,e}, W^{K,i}x_{i,e}\rangle$ 时使用乘法连接。该操作会对同一词元 $x_{i,e}$ 的变换结果各坐标进行乘法运算。换言之，可以将其视为对输入向量 $x_{i,e}$ 应用乘法层。顺带一提，除了自注意力层，Transformer 还包含全连接层、残差连接和归一化层等常用层；这些并非本文关注的重点。

3 乘法神经网络的表示能力

本节将探讨具有乘法层的神经网络的表达能力。我们首先证明，这些网络能够轻松表示多项式函数，并进一步利用这一点近似带限函数。由此，我们能够近似空间 $\mathrm{B}_{2r,\,2}$ 中的函数，而不受维数灾难的影响。具体而言，我们证明以下引理:

引理 1。 对于任意次数为 n、形式为 $p_n(x) = \sum_{k=0}^{n} c_k x^k$ 的多项式 $p_n:\ \mathbb{R} \to \mathbb{R}$，存在一个深度为 $L_n = \mathcal{O}(n)$、具有 $G_n = \mathcal{O}(n)$ 个神经元的乘法神经网络 $f_n^{\mathrm{POL}}:\ \mathbb{R}^3 \to \mathbb{R}$，使得 $f_n^{\mathrm{POL}}(x, x, c_0) = p_n(x)$ 对 $c_0 \in \mathbb{R}$ 成立，且 $x \in \mathbb{R}$。

有了这个引理，我们便可以展示如何利用乘法网络表示多项式的能力来逼近解析核函数 $K : \mathbb{R} \to \mathbb{C}$ 通过逼近某一类多项式，我们可以说明如何使用这些网络逼近解析核。具备逼近解析核的能力后，我们可以利用 Maurey 定理（见引理 3）将任意函数 $f \in \mathrm{H}_M$ 表示为有限个核叠加之和，并构造一个逼近该和式的网络。

实值解析函数的逼近。 在本节中，我们将展示如何逼近某些实值解析函数。为此，我们回顾伯恩斯坦 s-椭圆的概念。这是一种定义在复平面上的几何形状，在逼近理论中非常有用。

定义 3. 设 $M \geq 1$, $s > 1$ 为两个标量。$[-M, M]$ 上的伯恩斯坦 *s*-椭圆定义如下

$$E_s^M \;=\; \left\{ x + iy \in \mathbb{C} : \tfrac{x^2}{(a_s^M)^2} + \tfrac{y^2}{(b_s^M)^2} = 1 \right\},$$

其半长轴和半短轴分别为 $a_s^M = M\frac{s+s^{-1}}{2}$ 和 $b_s^M = M\frac{s-s^{-1}}{2}$。

参数 s 控制椭圆的形状，参数 M 决定其大小。例如，当 $s = 2$时，该椭圆是一个以原点为圆心、半径为 M 的圆。随着 s 增大，椭圆会变得更加狭长，其半短轴则会减小。椭圆的半长轴和半短轴分别决定其最大值和最小值。在陈述定理 2 的结果之前，我们先回顾 Trefethen (2019) 的以下定理：

定理 1 (Trefethen (2019) 的定理 8.2)。 设 $M > 0, s > 2$ 为标量，且 $K\colon [-M, M] \to \mathbb{R}$ 为一个解析函数，它可解析延拓至椭圆 E_s^M，并且在该椭圆上满足 上确界$_{x \in E_s^M} |K(x)| \leq C_K$，其中常数为 $C_K > 0$。对于任意 $n \in \mathbb{N}$，存在一个多项式 $h_n\colon \mathbb{R} \to \mathbb{C}$，其次数为 n，使得

$$\|h_n - K\|_{L^\infty([-M,M])} \;\leq\; \frac{2C_K s^{-n}}{s-1}.$$

定理 1 表明，任何在伯恩斯坦 s-椭圆 E_s^M 上有界的解析函数 K，都可以通过次数为 n 的多项式 h_n 进行精确逼近，并且误差会随着次数 n 的增加而呈指数级减小。

如下文所示，在某些情况下，在神经网络中使用乘法层可以提高函数逼近的效率。下面的定理表明，深度乘法网络可以在有界区间上逼近实值解析函数。

定理 2。 设 $M \geq 1$、 $s > 2$、 $C_K > 0$ 和 $\epsilon \in (0, 1)$ 为标量。那么，对于任意实值解析函数 $K\colon [-M, M] \to \mathbb{R}$，如果该函数可以解析延拓到椭圆 E_s^M，其中 $|K(x)| \leq C_K$，则存在一个深度乘法网络 $f^{\mathrm{MA}}\colon [-M, M]^3 \to \mathbb{R}$（*MA* 代表“乘法解析”），其深度为 $L_\epsilon = \mathcal{O}(\;{}^{1}\log_2 s \log_2 \frac{C_K}{\epsilon})$，具有 $G_\epsilon = \mathcal{O}(\;{}^{1}\log_2 s \log_2 \frac{C_K}{\epsilon})$ 个神经元，并满足

$$\|f^{\mathrm{MA}}(x, x, x) - K(x)\|_{L^\infty([-M,M])} \;\leq\; \epsilon.$$

定理 2 证明，具有二次乘法的深度乘法网络可以逼近任何在伯恩斯坦 s-椭圆 E_s^M上有界的实值解析函数，且误差上界随网络深度增加而指数级减小。

在（Montanelli 等人，2021）中，作者表明，使用深度 L_ϵ =O($\;{}^{1}$对数$_2^2 s$ 对数$_2^2 \frac{C_K}{\epsilon}$) 和 G_ϵ = O($\epsilon^2\;{}^{1}$对数$_2^2 s$ ~~对数$_2^2 \frac{C_K}{\epsilon}$~~) 个神经元可以逼近~~该核~~。相比之下，我们的方法在深度上提升了 O(log$(1/\epsilon)$)，在神经元数量上提升了 O(log$(1/\epsilon^2)$)，从而能以快得多的指数速度收敛到目标函数。

带限函数的逼近。接下来，我们研究神经网络逼近带限函数的能力。首先，我们说明如何使用解析核逼近带限函数。为此，我们将使用 Maurey 定理（Pisier，1980—1981年；Vitali D. Milman，1986年）。该定理指出，对于范数有界的希尔伯特子空间，凸包中的任意函数都可以利用该子空间中的点轻松逼近，且逼近速率取决于所用点的数量。

以下定理表明，深度乘法网络可以用相对较少的神经元和较浅的深度，将 $B = [0, 1]^d$ 中的带限函数近似到给定的误差容限内。

定理 3。 令 $\epsilon \in (0, 1)$、 $M > 1$、 $d \geq 2$ 和 $K\colon \mathbb{R} \to \mathbb{C}$ 为一个解析核，其关于 $s > 2$、$C_K > 0$ 满足定理 *2* 的假设，并且在 $[-dM, dM]$ 上由常数 $D_K \in (0, 1]$ 所界。令 f

为 $H_{K,M}(B)$ 中的一个实值函数。此外，令 $F:\ [-M,M]^d \to \mathbb{C}$ 为一个平方可积函数，使得 $f(x) = \int_{[-M,M]^d} F(\omega)K(\omega\cdot x)\,d\omega$。我们定义 $C_F = \int_{\mathbb{R}^d}|F(\omega)|\,d\omega = \int_{[-M,M]^d}|F(\omega)|\,d\omega$。那么，存在一个深度为 $L_\epsilon = O\left(\log_2 s \log_2 \frac{C_F C_K}{\epsilon}\right)$、具有 $G_\epsilon = O\left(\epsilon^{-2} C_F^2 \log_2 s \log_2 \frac{C_F C_K}{\epsilon}\right)$ 个神经元的深度乘法网络 $f^{\mathrm{MBL}}:\ B \to \mathbb{R}$（MBL 代表“乘法带限”），满足 $\|f - f^{\mathrm{MBL}}\|_{L^2(B)} \le \epsilon$。

上述定理表明，使用深度为 $L_\epsilon = \mathrm{O}(\log(1/\epsilon))$ 的乘法神经网络，并使用 $G_\epsilon = \mathrm{O}\left(\epsilon^{-2}\log(1/\epsilon)\right)$ 个神经元，可以将带限函数逼近到误差 ϵ。相比之下，Montanelli 等人（2021年）证明，使用深度为 $L_\epsilon = \mathrm{O}\left(\log^2(1/\epsilon)\right)$、含有 $G_\epsilon = \mathrm{O}\left(\epsilon^{-2}\log^2(1/\epsilon)\right)$ 个神经元的标准 ReLU 网络，也可以将带限函数逼近到相同的精度。这一结果展示了乘法神经网络相较于标准 ReLU 网络固有的参数效率优势。

光滑函数的逼近。现在我们转而展示 Sobolev 型函数的结果。我们利用定理 3 中关于带限函数的结果，来逼近 $\mathrm{B}_{2r,2} \subsetneq \mathrm{W}^{r,2}$ 中的函数。我们将证明，在稍强的假设下，使用乘法神经网络（即不含非线性激活），可以获得与 Barron（1993年）所示结果（其中网络是浅层 sigmoid 网络模型）相当的逼近率。此外，由于这些函数实际上属于 $\mathrm{W}^{r,2}$，我们可以更好地刻画这类函数。

定理 4. 设 $d \ge 2, r \in \mathbb{N}$，$f \in \mathcal{B}_{2r,2}$ 和 $\epsilon > 0$。存在一个深度为 $L_\epsilon = \mathcal{O}(d^2\epsilon^{-2/r})$ 且包含 $G_\epsilon = \mathcal{O}(d^2\epsilon^{-(2+2/r)})$ 个神经元的深度 *ReLU* 网络 f^{RS}（代表“*ReLU Sobolev*”），使得 $\|f^{\mathrm{RS}} - f\|_{L_2(B)} \le \epsilon$。

定理 5. 设 $d \ge 2, r \in \mathbb{N}$，$f \in \mathcal{B}_{2r,2}$ 和 $\epsilon > 0$。存在一个深度为 $L_\epsilon = \mathcal{O}(d\epsilon^{-1/r})$ 且包含 $G_\epsilon = \mathcal{O}(d\epsilon^{-(2+(1/r))})$ 个神经元的深度 *ReLU* 网络 f^{MS}（代表“乘法型 *Sobolev*”），使得 $\|f^{\mathrm{MS}} - f\|_{L_2(B)} \le \epsilon$。

证明。令 $f \in \mathrm{B}_{2r,\,2}$。我们希望先用带限函数 f_M 近似 f，然后用乘法神经网络 $f^{\mathrm{BL}} = f^{\mathrm{MS}}$ 近似 f_M。令 $M > 1$ 为一个频带。

回顾由下式给出的傅里叶逆变换：$(\mathrm{F}^{-1}g)(x) = \frac{1}{(2\pi)^d}\int_{\mathbb{R}^d} g(\omega)\exp(i\omega\cdot x)\,\mathrm{d}\omega$。我们将 f 的带限化定义为：$\mathbb{R}^d \to \mathbb{R}$，即 $f_M = \mathrm{F}^{-1}(\mathrm{F} f_{\mathbb{I}_{[-M,M]^d}})$，使得 $f_M \in \mathrm{H}_{K,\,M}(B)$ 对于 $K(x) = \exp(ix)$，且 $F = \frac{1}{(2\pi)^d}\mathrm{F} f$。我们有

$$\begin{aligned}\|f - f_M\|_{L_2(B)} &\le \left(\frac{1}{(2\pi)^d}\left\|\mathcal{F}f - \mathcal{F}f_{\mathbb{I}_{[-M,M]^d}}\right\|^2_{L_2(\mathbb{R}^d)}\right)^{1/2}\\ &= \left(\frac{1}{(2\pi)^d}\int_{\mathbb{R}^d\setminus[-M,M]^d}|\mathcal{F}f(\omega)|^2\,\mathrm{d}\omega\right)^{1/2}.\end{aligned}$$

对于任意 $\omega \in \mathbb{R}^d\setminus[-M,M]^d$，我们有 $|M^{-1}\omega|^{2r} \ge 1$。因此，

$$\begin{aligned}\left(\frac{1}{(2\pi)^d}\int_{\mathbb{R}^d\setminus[-M,M]^d}|\mathcal{F}f(\omega)|^2\,\mathrm{d}\omega\right)^{1/2} &\le \left(\frac{1}{(2\pi)^d}\int_{\mathbb{R}^d\setminus[-M,M]^d}\left|M^{-1}\omega\right|^{2r}|\mathcal{F}f(\omega)|^2\,\mathrm{d}\omega\right)^{1/2}\\ &\le M^{-r}\left(\frac{1}{(2\pi)^d}\int_{\mathbb{R}^d}|\omega|^{2r}|\mathcal{F}f(\omega)|^2\,\mathrm{d}\omega\right)^{1/2}\\ &\le M^{-r},\end{aligned}$$

其中，最后一个不等式由 $f \in \mathrm{B}_{2r,\,2}$ 得出。对于任意 $\epsilon > 0$，我们设 $M = (2/\epsilon)^{1/r}$。我们将构造一个神经网络 f^{MS}，用于近似带限函数 f_M，使得

$$\|f_M - f^{\mathrm{MS}}\|_{L_2(B)} \le \epsilon/2.$$

假设我们已经构造了这样的 f^{MS}，那么根据三角不等式，我们可以得到

$$\|f - f^{\mathrm{MS}}\|_{L_2(B)} \;\le\; \|f - f_M\|_{L_2(B)} + \|f_M - f^{\mathrm{MS}}\|_{L_2(B)} \;\le\; M^{-r} + \epsilon/2 \;\le\; \epsilon.$$

我们定义函数 $F\colon\ [-M,M]^d \to \mathbb{C}$ 如下

$$F(\omega) \;=\; \frac{1}{(2\pi)^d}(\mathcal{F}f)(\omega).$$

于是我们得到以下恒等式

$$f_M(x) \;=\; \int_{[-M,M]^d} F(\omega)K(\omega\cdot x)\,\mathrm{d}\omega \;=\; \int_{[-M,M]^d} \tfrac{1}{(2\pi)^d}\mathcal{F}f(\omega)\exp(i\omega\cdot x)\,\mathrm{d}\omega.$$

现在，我们可以在定理 3 的条件下进行推导。我们考虑 $C_F = \int_{[-M,M]^d}|F(\omega)|\,\mathrm{d}\omega = (2\pi)^d \int_{[-M,M]^d}|\mathrm{F}f(\omega)|\,\mathrm{d}\omega \le 1$，其中不等式由 $\mathrm{B}_{2r,2}$ 的定义得出。核 $K(t) = \exp(it)$ 的输入为 $t = \omega\cdot x$，其中 $x \in B$和 $\omega \in [-M,M]^d$。因此，$t \in [-dM, dM]$，且 $K\colon\ [-dM,dM] \to \mathbb{R}$。我们注意到，$K$ 可延拓至 Bernstein 四椭圆 E_4^{dM}。我们注意到，$a_4^{dM} = dM(4+4^{-1})/2$ 是较长轴，因此 K 在 E_s^{dM}上的最大范数为 $K(a_4^{dM})$:

$$C_K \;=\; \max_t |K(t)| \;\le\; \exp(dM\tfrac{4+4^{-1}}{2}).$$

此外，对于任意 $t \in \mathbb{R}$，我们有 $|K(t)| \le 1 = D_K$。

现在我们用乘法网络近似 f_M。利用定理 3 的结果，存在一个深度乘法神经网络 f^{MS}，它在 $L_2(B)$ 中近似带限函数 f_M，其误差不超过 $\epsilon/2$，且深度

$$\begin{aligned} L_\epsilon \;&\le\; C_1\frac{1}{\log_2 4}\log_2\frac{2C_KC_F}{\epsilon} \\ &\le\; C_2\frac{1}{\log_2 4}\log_2\frac{\exp(dM\frac{4+4^{-1}}{2})}{\epsilon} \\ &\le\; C_2\Big(3d\epsilon^{-1/r} + \log_2(1/\epsilon)\Big) \\ &\le\; C_3\cdot d\epsilon^{-1/r}, \end{aligned} \tag{2}$$

其中 $C_1, C_2, C_3 > 0$ 为一些常数。此外，神经元数量可界定为

$$\begin{aligned} G_\epsilon \;&\le\; C_1'C_F^2\cdot\epsilon^{-2}\log_2\frac{2C_KC_F}{\epsilon} \\ &\le\; C_2'\cdot\epsilon^{-2}\log_2\frac{\exp(dM\frac{4+4^{-1}}{2})}{\epsilon} \\ &\le\; C_2'\cdot\epsilon^{-2}(3d(2/\epsilon)^{1/r} + \log_2(1/\epsilon)) \\ &\le\; C_3'\cdot d\epsilon^{-(2+1/r)}, \end{aligned} \tag{3}$$

其中 $C_1', C_2', C_3' > 0$ 为一些常数。 □

定理4-5揭示了 Sobolev型球 $\mathrm{B}_{2r,\,2}$的若干有趣性质。首先，我们看到，逼近这些函数时不会遭遇逼近整个 Sobolev空间时出现的维数灾难。其次，对于 ReLU 网络和乘法网络，随着光滑度 r 提高，误差界会变得更紧，这是逼近误差界所期望的性质。事实上，当 r 趋于无穷时，该误差界趋近于 (Barron, 1993) 中提出的误差界。最后，我们证明，对于相同的误差容限 ϵ，乘法神经网络可以用深度为 $O(d^2\epsilon^{-1/r})$、包含 $O(d^2\epsilon^{-(1+1/r)})$ 个神经元的网络逼近目标函数$f \in \mathrm{B}_{2r,\,2}$；而标准 ReLU 神经网络则需要深度为 $O(d^2\epsilon^{-2/r})$、包含 $O(d^2\epsilon^{-(2+2/r)})$ 个神经元的网络。这一结果表明，在逼近 Sobolev型空间中的函数时，乘法神经网络具有更强的逼近保证。

4 结论

先前的论文研究了标准全连接神经网络对 Barron空间 $B^{1,1}$ (Barron，1993)、带限函数空间（Montanelli 等人，2021）以及 Korobov空间（Blanchard & Bennouna，2022a）中函数的逼近保证。这些研究表明，全连接网络能够逼近广泛的光滑函数而不遭遇维数灾难，并揭示了神经网络在学习特定类型函数时宽度与深度之间的权衡。然而，这些结果仅适用于全连接网络的若干变体，并未提供关于其他类型架构的信息。

本文通过探究乘法神经网络和标准全连接网络在逼近带限函数以及 Sobolev 型球 $B^{2r,2}$中的元素时的逼近保证，拓展了这些结果。我们的结果表明，与标准 ReLU 网络相比，乘法神经网络能够实现更强的逼近保证。此外，我们还证明，与 Barron 空间和带限函数空间不同，$B^{2r,2}$ 是 Sobolev空间 $W^{r,2}$的子集。因此，我们的结果表明，对于 Sobolev空间中的广大子集，可以避免维数灾难。

致谢

Tomer Galanti 获得了 Center for Minds, Brains and Machines (CBMM) 的资助，该中心由美国国家科学基金会（NSF）STC 项目 CCF-1231216 和美国国家科学基金会（NSF）项目 213418 提供经费支持。

参考文献

David W. Albrecht、Xuan Thinh Duong 和 Alan Mcintosh。《算子理论与调和分析》。*Springer* 数学与统计学会议录，2021年。

Kai Arulkumaran、Antoine Cully 和 Julian Togelius。Alphastar：进化计算视角。遗传与进化计算会议伴会，2019。

A. R. Barron。人工神经网络的近似与估计界。载于 第四届计算学习理论年度研讨会论文集，COLT '91，第 243–249 页，美国加利福尼亚州旧金山，1991 年。Morgan Kaufmann 出版公司。ISBN 1558602135。

Andrew R. Barron。S形函数叠加的通用近似界。*IEEE* 信息论汇刊，39：930–945，1993年。

Raphael Bensadoun、Shir Gur、Tomer Galanti 和 Lior Wolf。元内部学习。载于 M. Ran- zato、A. Beygelzimer、Y. Dauphin、P.S. Liang 和 J. WortmanVaughan（编），神经信息处理系统进展，第 34 卷，第 20645–20656 页。Curran Associates, Inc.，2021年。URL https://proceedings.neurips.cc/paper/2021/file/ac796a52db3f16bbdb6557d3d89d1c5a-Paper.pdf。

Moise Blanchard 和 Mohammed Amine Bennouna。浅层网络和深度网络是 Korobov 函数的近最优逼近器。发表于*ICLR*，2022a。

Moise Blanchard 和 Mohammed Amine Bennouna。浅层网络和深度网络是 Korobov 函数的近最优逼近器。发表于国际学习表征会议，2022b。URLhttps://openreview.net/forum?id=AV8FPoMTTa。

Avrim Blum、John Hopcroft 和 Ravindran Kannan。数据科学基础。剑桥大学出版社，2020年。doi: 10.1017/9781108755528。

Tom Brown、Benjamin Mann、Nick Ryder、Melanie Subbiah、JaredD Kaplan、Prafulla Dhariwal、Arvind Neelakantan、Pranav Shyam、Girish Sastry、Amanda Askell、Sandhini Agarwal、Ariel Herbert-Voss、Gretchen Krueger、Tom Henighan、Rewon Child、Aditya Ramesh、Daniel Ziegler、Jeffrey Wu、Clemens Winter、Chris Hesse、Mark Chen、Eric Sigler、Mateusz Litwin、Scott Gray、Benjamin Chess、Jack Clark、Christopher Berner、Sam McCandlish、Alec Radford、Ilya Sutskever 和 Dario Amodei。语言模型是少样本学习者。发表于神经信息处理系统进展，第 33 卷，第 1877–1901 页。Curran Associates, Inc.，2020年。

Mark Chen、Jerry Tworek、Heewoo Jun、Qiming Yuan、Henrique Ponde de Oliveira Pinto、Jared Kaplan、Harri Edwards、Yuri Burda、Nicholas Joseph、Greg Brockman、Alex Ray、Raul Puri、Gretchen Krueger、Michael Petrov、Heidy Khlaaf、Girish Sastry、Pamela Mishkin、Brooke Chan、Scott Gray、Nick Ryder、Mikhail Pavlov、Alethea Power、Lukasz Kaiser、Mohammad Bavarian、Clemens Winter、Philippe Tillet、Fe- lipe Petroski Such、Dave Cummings、Matthias Plappert、Fotios Chantzis、Elizabeth Barnes、Ariel Herbert- Voss、William Hebgen Guss、Alex Nichol、Alex Paino、Nikolas Tezak、Jie Tang、Igor Babuschkin、Suchir Balaji、Shantanu Jain、William Saunders、Christopher Hesse、Andrew N. Carr、Jan Leike、Josh Achiam、Vedant Misra、Evan Morikawa、Alec Radford、Matthew Knight、Miles Brundage、Mira Murati、Katie Mayer、Peter Welinder、Bob McGrew、Dario Amodei、Sam McCandlish、Ilya Sutskever 和 Wojciech Zaremba。《评估在代码上训练的大语言模型》，2021年。

Jianpeng Cheng、Li Dong 和 Mirella Lapata。用于机器阅读的长短期记忆网络。载于自然语言处理实证方法会议，2016。

George V. Cybenko。通过 S 形函数的叠加进行逼近。控制、信号与系统数学，2：303–314，1989。

Jacob Devlin、Ming-Wei Chang、Kenton Lee 和 Kristina Toutanova。BERT：用于语言理解的深度双向 Transformer 预训练。载于*2019*年计算语言学协会北美分会年会论文集：人类语言技术，第*1*卷（长篇与短篇论文）。计算语言学协会，2019年6月。

Ronald A. DeVore 和 George G. Lorentz。构造逼近。载于数学科学基础丛书，1993年。Ronald A. DeVore、Ralph Howard 和 Charles A. Micchelli。最优非线性逼近。*manuscriptamathematica*，63：469–478，1989年。Luc Devroye、László Györfi 和 Gábor Lugosi。模式识别概率理论。载于随机建模与应用概率，1996年。Alexey Dosovitskiy、Lucas Beyer、Alexander Kolesnikov、Dirk Weissenborn、Xiaohua Zhai、Thomas Un- terthiner、Mostafa Dehghani、Matthias Minderer、Georg Heigold、Sylvain Gelly、Jakob Uszkoreit 和 Neil Houlsby。一张图像胜过16×16个词：用于大规模图像识别的Transformer。载于国际学习表征会议，2021年。URL https://openreview.net/forum?id=YicbFdNTTy。Richard Durbin 和 David E. Rumelhart。乘积单元：一种计算能力强且符合生物学原理的反向传播网络扩展。*Neural Comput.*，1(1)：133–142，1989年3月。ISSN 0899-7667。doi：10.1162/neco.1989.1.1.133。URL https://doi.org/10.1162/neco.1989.1.1.133。鄂维南、马超和吴磊。神经网络模型的 Barron 空间与流诱导函数空间。构造逼近，55：369–406，2021年。Tomer Galanti 和 Lior Wolf。论超网络的模块化。载于神经信息处理系统进展第*33*卷。Curran Associates, Inc.，2020年。David Ha、Andrew Dai 和 Quoc V Le。超网络。*arXiv*预印本 *arXiv:1609.09106*，2016年。何恺明、张祥雨、任少卿和孙剑。用于图像识别的深度残差学习。载于*2016*年*IEEE*计算机视觉与模式识别会议论文集，CVPR '16，第770–778页。IEEE，2016年6月。doi：10.1109/CVPR.2016.90。URL http://ieeexplore.ieee.org/document/7780459。Kurt Hornik、Maxwell B. Stinchcombe 和 Halbert L. White。多层前馈网络是通用逼近器。神经网络，2：359–366，1989年。船桥贤一。神经网络对连续映射的近似实现。神经网络，2：183–192，1989年。Masaaki Imaizumi 和 Kenji Fukumizu。深度神经网络能够有效地学习非光滑函数。载于国际人工智能与统计会议，2018年。Lee K. Jones。关于希尔伯特空间中贪心逼近的一个简单引理，以及投影寻踪回归和神经网络训练的收敛速率。统计年刊，20：608–613，1992年。Jason M. Klusowski 和 Andrew R. Barron。高维脊函数组合（包括神经网络）的风险界。*arXiv*：统计理论，2016年。Michael Kohler 和 Adam Krzyżak。基于层次交互模型的非参数回归。*IEEE* 信息论汇刊，63(3)：1620–1630，2017年。doi：10.1109/TIT.2016.2634401。David Krueger、Chin-Wei Huang、Riashat Islam、Ryan Turner、Alexandre Lacoste 和 Aaron Courville。贝叶斯超网络。*arXiv*预印本 *arXiv:1710.04759*，2017年。Moshe Leshno、Vladimir Ya. Lin、Allan Pinkus 和 Shimon Schocken。具有非多项式激活函数的多层前馈网络可以逼近任意函数。纽约大学斯特恩商学院研究论文系列，1991年。Yoav Levine、Noam Wies、Or Sharir、Hofit Bata 和 Amnon Shashua。自注意力深度效率的局限。载于 H. Larochelle、M. Ranzato、R. Hadsell、M.F. Balcan 和 H. Lin（编），神经信息处理系统进展，第33卷，第22640–22651页。Curran Associates, Inc.，2020年。URL https://proceedings.neurips.cc/paper/2020/file/ff4dfdf5904e920ce52b48c1cef97829-Paper.pd

f.

Shiyu Liang 和 Rayadurgam Srikant。为什么深度神经网络适用于函数逼近？发表于*ICLR*，2017年。

Gidi Littwin 和 Lior Wolf。用于形状表示的深度元泛函。发表于*IEEE*国际计算机视觉会议 *(ICCV)*，2019。

H.N. Mhaskar。用于流形上函数逼近的特征网络。应用与计算调和分析，29(1):63–87，2010年。ISSN 1063-5203。doi: https://doi.org/10.1016/j.acha.2009.08.006。URL https://www.sciencedirect.com/science/article/pii/S106352030900089X。

Hrushikesh Mhaskar、Qianli Liao 和 Tomaso Poggio。深度网络何时以及为何优于浅层网络？发表于第三十一届*AAAI*人工智能会议论文集，AAAI'17，第2343–2349页。AAAI出版社，2017年。

Hrushikesh Narhar Mhaskar. 用神经网络对光滑函数和解析函数进行最优逼近。神经计算，8:164–177，1996。

Hadrien Montanelli 和 Qiang Du. 使用稀疏网格为深度 ReLU 网络提供新的误差界。*SIAM*数据科学数学期刊，1(1):78–92，2019。doi: 10.1137/18M1189336。URLhttps://doi.org/10.1137/18M1189336。

Hadrien Montanelli、Haizhao Yang 和 Qiang Du. 深度 ReLU 网络克服带限函数的维数灾难。计算数学杂志，39:801–815，2021年。

Ryumei Nakada 和 Masaaki Imaizumi. 具有内在维数的深度神经网络的自适应逼近与泛化。机器学习研究杂志，21(1)，2022年6月。ISSN 1532-4435。

Ankur Parikh、Oscar Täckström、Dipanjan Das 和 Jakob Uszkoreit。《用于自然语言推理的可分解注意力模型》。载于*2016*年自然语言处理实证方法会议论文集，第 2249–2255 页，得克萨斯州奥斯汀，2016年11月。计算语言学协会。doi：10.18653/v1/D16-1244。URL https://aclanthology.org/D16-1244。

Romain Paulus、Caiming Xiong 和 Richard Socher。《用于抽象式摘要的深度强化模型》，2017年。URL https://arxiv.org/abs/1705.04304。

Philipp Christian Petersen 和 Felix Voigtländer。《使用深度ReLU神经网络对分段光滑函数进行最优逼近》。神经网络：国际神经网络学会官方期刊，108：296–330，2017年。

Allan Pinkus。《逼近理论中的 *N*-宽度》。施普林格出版社，1985年。

Gilles Pisier。关于 B. Maurey 一项未发表结果的评注，1980—1981年。网址 http://www.numdam.org/item/SAF_1980-1981____A5_0/。

Tomaso Poggio、Andrzej Banburski 和 Qianli Liao。《深度网络中的理论问题》。《美国国家科学院院刊》，117(48):30039–30045，2020年。doi：10.1073/pnas.1907369117。URL https://www.pnas.org/doi/abs/10.1073/pnas.1907369117。John J. F. Fournier、Robert A. Adams。《索博列夫空间》。纯粹与应用数学。学术出版社，第2版，2003年。ISBN 9780120441433,9781435608108,0120441438。URL http://gen.lib.rus.ec/book/index.php?md5=72db1ed30d242bd3ebcb73bcd099a6b4。Johannes Schmidt-Hieber。流形上函数的深度 ReLU 网络逼近，2019。URLhttps://arxiv.org/abs/1908.00695。Claude E. Shannon。噪声环境下的通信。《*IEEE* 会刊》，72:1192–1201，1984年。David Silver、Aja Huang、Chris J. Maddison、Arthur Guez、Laurent Sifre、George van den Driessche、Julian Schrittwieser、Ioannis Antonoglou、Veda Panneershelvam、Marc Lanctot、Sander Dieleman、Dominik Grewe、John Nham、Nal Kalchbrenner、Ilya Sutskever、Timothy Lillicrap、Madeleine Leach、Koray Kavukcuoglu、Thore Graepel 和 Demis Hassabis。利用深度神经网络和树搜索掌握围棋。《自然》，529:484–489，2016。ISSN 0028-0836。doi：10.1038/nature16961。

Andrew Trask、Felix Hill、Scott Reed、Jack Rae、Chris Dyer 和 Phil Blunsom。神经算术逻辑单元，2018年。URL https://arxiv.org/abs/1808.00508。Lloyd N. Trefethen。《《逼近理论与逼近实践（扩展版）》》。工业与应用数学学会，宾夕法尼亚州费城，2019。doi：10.1137/1.9781611975949。URLhttps://epubs.siam.org/doi/abs/10.1137/1.9781611975949。Sebastian Urban 和 Patrick van der Smagt。一种用于在加性交互与乘性交互之间平滑且可微地转换的神经传递函数，2015年。URL https://arxiv.org/abs/1503.05724。Ashish Vaswani、Noam Shazeer、Niki Parmar、Jakob Uszkoreit、Llion Jones、Aidan N Gomez、Ł ukasz Kaiser 和 Illia Polosukhin。注意力就是你所需要的一切。载于神经信息处理系统进展，第30卷。Curran Associates, Inc.，2017年。Gideon Schechtman、Vitali D. Milman。《*Maurey-Pisier*定理》，第85–97页。施普林格，柏林，海德堡，柏林，海德堡，1986年。ISBN 978-3-540-38822-7。doi：10.1007/978-3-540-38822-7_13。URL https://doi.org/10.1007/978-3-540-38822-7_13。Johannes von Oswald、Christian Henning、João Sacramento 和 Benjamin F. Grewe。利用超网络进行持续学习。载于国际学习表征会议，2020年。Grace Wahba。《《观测数据的样条模型》》。工业与应用数学学会，1990年。doi：10.1137/1.9781611970128。URL https://epubs.siam.org/doi/abs/10.1137/1.9781611970128。Dmitry Yarotsky。深度 ReLU 网络逼近的误差界。《神经网络：国际神经网络学会官方期刊》，94:103–114，2017年。Xiaohua Zhai、Alexander Kolesnikov、Neil Houlsby 和 Lucas Beyer。视觉Transformer的扩展。*2022*年*IEEE/CVF*计算机视觉与模式识别会议（*CVPR*），第1204–1213页，2021年。

A 示例

示例 1. 令 $f \in L_2(\mathbb{R}^d)$ 满足条件，使得 f 是带宽为 π 的带限函数（即 $\operatorname{supp}(\mathcal{F}f) \subset [-\pi,\pi]^d$）。令 $\phi(x) = \operatorname{sinc}(x) = \sin(\pi x)/\pi x$，并令 $\phi_k = \phi(\cdot - k)$ 对所有 $k \in \mathbb{Z}^d$ 成立。根据 *Shannon-Nyquist* 定理（*Shannon，1984*），我们有：

$$f(x) = \sum_{k\in\mathbb{Z}^d} \langle f, \phi_k\rangle \phi_k(x) = \sum_{k\in\mathbb{Z}^d} f(k)\phi(x-k).$$

这意味着，每个带限函数都可以通过一组离散的整数采样点完全确定。对于高维函数（d 很大）而言，这一结果尤其令人惊讶，因为点 x 与采样点之间的最大距离会随 d增大。例如，$\mathbb{R}^d$中的单位立方体顶点与其中心的距离为 $\sqrt{d}/2$。尽管如此，即使距离按 $\sqrt{d}$的比例增大，我们仍然可以从整数顶点处恢复采样值。使用神经网络恢复高维函数时，这一性质很有用。参见图*1*。

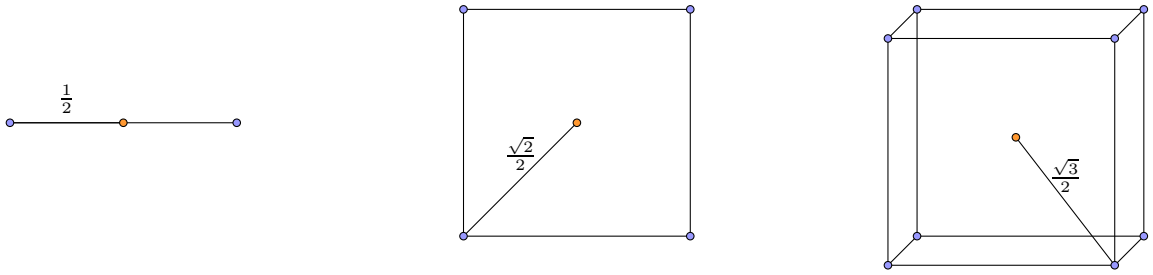

图1：例 1 的示意图。带限函数 f 可以通过一组离散的值重构，尽管每个立方体中的距离会以 $O(\sqrt{d})$ 的速度增长。

B 证明

定义 4. 对于两个巴拿赫空间 V_1 和 V_2，如果存在某个常数 C，使得对于 $u \in V_1$，$\|u\|_{V_2} \le C\|u\|_{V_1}$，则称 V_1 嵌入于 V_2。

引理 2. 设 $r > 0$ 和 $d \ge 2$。则，$\mathcal{B}_{2r,2}$ 作为真子空间嵌入于$\mathcal{W}^{r,}$ ^{2}.

证明。由于 $f \in B_{2r,2}$，我们有 $f \in L_2(\mathbb{R}^d)$，因此，$\mathbf{k}f\mathbf{k}_2 < \infty$。此外，根据 Robert A. Adams (2003)（定理 5.2），存在一个常数 $C_1 > 0$，使得

$$\|f\|_{\mathcal{W}^{r,2}} \le C_1(\|f\|_2 + |f|_{r,2}),$$

其中 C_1 取决于 r 和维数 d。因此，剩下要证明的是 $|f|_{r,2}$ 有界。为此，我们使用多元函数空间的一些性质。

我们先给出以下结论，以便后续使用。设 $\omega \in \mathbb{R}^d, d \ge 2$，则有

$$|\omega|^{2r} = \|\omega\|_{l_2}^{2r} = \left(\sum_{m=1}^{d} \omega_m^2\right)^r = \sum_{\|\alpha\|_1=r} \binom{r}{\alpha_1,\ldots,\alpha_d} \cdot (\omega^\alpha)^2, \tag{4}$$

其中 $\omega^\alpha = \prod_{i=1}^d \omega^{\alpha_i}$ 满足 $\alpha \in \mathbb{R}^d$。

根据詹森不等式，

$$|f|_{r,2}^2 = \left(\sum_{\alpha:|\alpha|=r} \|D^\alpha f\|_2\right)^2 \le d^r \sum_{\alpha:|\alpha|=r} \|D^\alpha f\|_2^2$$

根据 $\mathbb{R}^d$ 中的帕塞瓦尔恒等式（Albrecht 等人，2021）：

$$\|D^\alpha f\|_2^2 \;=\; \frac{1}{(2\pi)^d}\int_{\mathbb{R}^d}|\mathcal{F}(D^\alpha f)(\omega)|^2\,\mathrm{d}\omega \;=\; \frac{1}{(2\pi)^d}\int_{\mathbb{R}^d}|(\mathcal{F}f)(\omega)|^2|\omega^\alpha|^2\,\mathrm{d}\omega.$$

因此，由（4）

$$\begin{aligned}\sum_{\alpha:|\alpha|=r}\|D^\alpha f\|_2^2 \;&=\; \sum_{\alpha:|\alpha|=r}\frac{1}{(2\pi)^d}\int_{\mathbb{R}^d}|(\mathcal{F}f)(\omega)|^2|\omega^\alpha|^2\,\mathrm{d}\omega\\ &\le\; \frac{1}{(2\pi)^d}\int_{\mathbb{R}^d}|(\mathcal{F}f)(\omega)|^2\left(\sum_{\alpha:|\alpha|=r}\binom{r}{\alpha_1,\ldots,\alpha_d}\cdot|\omega^\alpha|^2\right)\mathrm{d}\omega\\ &=\; \frac{1}{(2\pi)^d}\int_{\mathbb{R}^d}|\mathcal{F}f(\omega)|^2|\omega|^{2r}\,\mathrm{d}\omega.\end{aligned}$$

其中，不等式由以下事实得出：$\binom{r}{\alpha_1\ldots,,\alpha_d}\ge 1$。最后，我们得出结论

t

$$\begin{aligned}\|f\|_{\mathcal{W}^{r,2}} \;\le\; C_1(\|f\|_2+|f|_{r,2}) \;&\le\; C_1\Big(\|f\|_2+\tfrac{d^r}{(2\pi)^{d/2}}\left(\textstyle\int_{\mathbb{R}^d}|(\mathcal{F}f)(\omega)|^2|\omega|^{2r}\,\mathrm{d}\omega\right)^{1/2}\Big)\\ &\le\; C_2\|f\|_{\mathcal{B}_{2r,2}},\end{aligned}$$

对于某个 $C_1,C_2>0$，其中第二个不等式源于方程 1 中给出的 $\mathrm{k}\cdot\mathrm{k}_{\mathcal{B}_{2r,2}}$ 的定义。

为了证明 $\mathrm{B}_{2r,2}$ 是 $\mathrm{W}^{r,2}$的真子空间，我们以函数 $f\in\mathrm{W}^{r,2}\big(\mathbb{R}^d\big)$为例，通过其傅里叶变换来定义它：

$$\mathcal{F}f(\omega)=\begin{cases}1 & |\omega|\le 1\\ |\omega|^{-(r+(d+\epsilon)/2)} & |\omega|>1\end{cases}$$

对于任意 $\epsilon>0$。

事实上，$f\in\mathrm{W}^{r,2}\big(\mathbb{R}^d\big)$，因为：

$$\begin{aligned}\int_{\mathbb{R}^d}|\omega|^{2r}|\mathcal{F}f(\omega)|^2\,\mathrm{d}\omega \;&=\; \int_{|\omega|\le 1}|\omega|^{2r}|\mathcal{F}f(\omega)|^2d\omega+\int_{|\omega|>1}|\omega|^{2r}|\mathcal{F}f(\omega)|^2\,\mathrm{d}\omega\\ &\le\; 1+\int_{|\omega|>1}|\omega|^{-(d+\epsilon)}\,\mathrm{d}\omega<\infty.\end{aligned}$$

然而，对于这个例子，当维数相对于光滑度指数较高时，$\mathrm{F}f(\omega)\notin L_1$。也就是说，只要

$$r+(d+\epsilon)/2\le d\Leftrightarrow 2r+\epsilon\le d.$$

这意味着 f 不满足 $\mathrm{B}_{2r,\,2}$所要求的条件 ^{1} $(2\pi)^{-d}\int_{\mathbb{R}^d}|\mathrm{F}f(\omega)|d\omega\le 1$。我们看到，对 $\mathrm{B}_{2r,\,r}$ 中的函数的傅里叶变换所施加的 L_1 条件，使我们在 Sobolev 空间中进行逼近时能够规避维数灾难。

□

引理 1. 对于任意形式为 $p_n(x)=\sum_{k=0}^n c_k x^k$、次数为 n 的多项式 $p_n:\ \mathbb{R}\to\mathbb{R}$，存在一个乘法神经网络 $f_n^{\mathrm{POL}}:\ \mathbb{R}^3\to\mathbb{R}$，其深度为 $L_n=\mathcal{O}(n)$，包含 $G_n=\mathcal{O}(n)$ 个神经元，并满足$f_n^{\mathrm{POL}}(x,x,c_0)=p_n(x)$，其中 $c_0\in\mathbb{R}$，且 $x\in\mathbb{R}$。

证明。令 $p_n(x)=\sum_{k=0}^n c_k x^k$ 为一个次数为 n 的多项式，且 $p_{n,\,i}=\sum_{k=0}^i c_k x^k$ 其截至第i项的部分和。我们构造一个网络 f_n^{POL}，使其第 i 层满足：

$$(f_n^{\mathrm{POL}})_i(x,x,c_0)=(x,x^{i+1},p_{n,i}(x)).$$

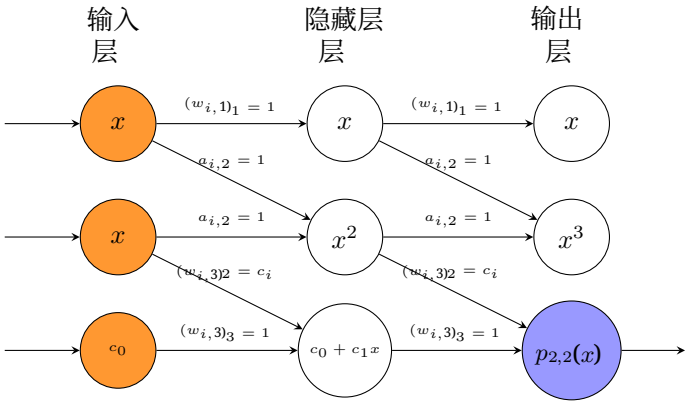

图 2：引理 1 的证明中乘法网络的示意图，其中多项式次数为 $n = 2$。

在第 i 层，三个神经元的模型权重定义如下：

$$w_{i,1} = (1, 0, 0), \quad a_{i,2} = 1 \text{ for } (j_1, j_2) = (1, 2) , \quad w_{i,3} = (0, c_i, 1)$$

其余权重均为零。

我们论证，在任意层 $i \geq 0$中，神经元 1 包含 x，神经元 2 包含 x^{i+1} ，神经元 3 包含$p_{n,i}(x)$。设 $y_{i+1,1}$, $y_{i+1,2}$,$y_{i+1,3}$ 为层 $i+1$中的三个神经元。根据假设，影响第一个神经元的唯一非零权重位于 $w_{i+1,1}$, $y_{i+1,1} = y_{i,1} = x$ 。影响第二个神经元的唯一非零权重出现在 $a_{i=1,2}$，因此 $y_{i+1,2} = y_{i,1}y_{i,2} = x^{i+1}$。最后，$y_{i+1,3}$ 仅取决于$w_{i+1,3}$，因此 $y_{i+1,3} = c_{i+1}y_{i,2} + y_{i,3} = p_{n,i}(x) + c_{i+1}x^{i+1} := p_{n,i+1}(x)$。利用$p_{n,n}(x) = p_n(x)$ 这一事实，我们得出结论。 □

定理 2。 设 $M \geq 1$、 $s > 2$、 $C_K > 0$ 和 $\epsilon \in (0, 1)$ 为标量。则对于任意实值解析函数 K： $[-M, M] \to \mathbb{R}$，若其可解析延拓至椭圆 E_s^M，其中 $|K(x)| \leq C_K$，则存在一个深度乘法网络 f^{MA}： $[-M, M]^3 \to \mathbb{R}$ （*MA* 代表“乘法解析”），其深度为 $L_\epsilon = \mathcal{O}(\ ^1\log_2 s \log_2 \frac{C_K}{\epsilon})$，包含 $G_\epsilon = \mathcal{O}(\ ^1\log_2 s \log_2 \frac{C_K}{\epsilon})$ 个神经元，并满足

$$\|f^{\text{MA}}(x, x, x) - K(x)\|_{L^\infty([-M,M])} \leq \epsilon.$$

证明。 设 $M \geq 1$、 $s > 2$、 $C_K > 0$、 $\epsilon \in (0、1)$，并设 K 为满足所需条件的解析函数。首先，我们使用一个次数为 $n \geq 2$ （稍后定义）的多项式 h_n 来逼近 K 。然后，根据引理 1，使用深度乘法网络实现 h_n 。我们回顾定理 1 中描述的多项式 h_n 。对于任意整数 $n \geq 2$，有

$$\|h_n - K\|_{L^\infty([-M,M])} \leq \frac{C_K s^{-n}}{s-1} = \mathcal{O}(C_K s^{-n}).$$

通过选择 $n = \ ^1\log_2 s \log_2 \frac{C_K}{\epsilon(s-1)} \leq \ ^1\log_2 s \log_2 \frac{C_K}{\epsilon}$，我们得到以下关系：

$$\begin{aligned} \frac{\epsilon(s-1)}{C_K} &= 2^{\log_2(\frac{\epsilon(s-1)}{C_K})} \\ &= 2^{-\frac{1}{\log_2 s}(\log_2(\frac{C_K}{\epsilon(s-1)}))\log_2 s} \\ &= s^{-\frac{1}{\log_2 s}(\log_2(\frac{C_K}{\epsilon(s-1)}))} = s^{-n}. \end{aligned} \tag{5}$$

特别地,

$$\|h_n - K\|_{L^\infty([-M,M])} \leq \frac{C_K s^{-n}}{s-1} \leq \epsilon,$$

对于 $s \geq 1$。鉴于 h_n 是一个次数为 n 的多项式，根据引理 1，存在 f^{POL} 和 $c_0 = h_{n,\ 0}$ ，使得 $h_n(x) = f^{\text{POL}}(x,\ x,\ c_0)$，从而得到所需结果。 □

现在让我们回顾一下 Maurey 定理（Pisier，1980—1981年；Vitali D. Milman，1986年），它将有助于我们证明定理 3。

引理 3 (Maurey 定理)。 设 $\mathcal{V}$ 是一个范数为 $\|\cdot\|_{\mathcal{V}}$ 的希尔伯特空间。假设存在 $Q \subset \mathcal{V}$，使得对每个 $q \in Q$，都有 $\|q\|_{\mathcal{V}} \leq b$，其中某个 $b > 0$。那么，对凸包 Q 中的任意 f 以及任意整数 $n \geq 1$，都存在一个属于 Q 中 n 个点的凸包的 f_n，以及一个常数 $c > b^2 - \|f\|_{\mathcal{V}}^2$，使得$\|f_n - f\|_{\mathcal{V}}^2 \leq \frac{c}{n}$。

定理 *3.* 设 $\epsilon \in (0,1)$、 $M > 1$、 $d \geq 2$ 和 K： $\mathbb{R} \to \mathbb{C}$ 是一个解析核，它满足关于 $s > 2$、 $C_K > 0$ 的定理 *2* 的假设，并且在 $[-dM, dM]$ 上由常数 $D_K \in (0, 1]$）界定。设 f 是 $H_{K,M}(B)$ 中的一个实值函数。此外，设 F: $[-M, M]^d \to \mathbb{C}$ 是一个平方可积函数，满足 $f(x) = \int_{[-M,M]^d} F(\omega)K(\omega \cdot x)\, d\omega$。我们定义 $C_F = \int_{\mathbb{R}^d} |F(\omega)|\, d\omega = \int_{[-M,M]^d} |F(\omega)|\, d\omega$。那么，存在一个深度乘法网络 f^{MBL}： $B \to \mathbb{R}$ （MBL 表示“乘法带限”），其深度为 $L_\epsilon = O\Big(\underline{\quad}^{1} log_2 s \log_2 \frac{C_F C_K}{\epsilon}\Big)$，包含 $G_\epsilon = O\Big(\underline{\quad}_{\epsilon^2}^{C_F^2} log_2 s \log_2 \frac{C_F C_K}{\epsilon}\Big)$ 个神经元，并满足 $\|f - f^{MBL}\|_{L^2(B)} \leq \epsilon$。

证明。 令 $f \in \mathrm{H}_{K,M}(B)$. 此外，令 $F(\omega) = |F(\omega)| \cdot \exp(i\theta(\omega))$，即 $F(\omega)$ 的极坐标表示。以下结论成立：

$$\begin{aligned} f(x) &= \mathrm{Re}\left(\int_{[-M,M]^d} F(\omega) K(\omega \cdot x)\, \mathrm{d}\omega \right) \\ &= \mathrm{Re}\left(\int_{[-M,M]^d} C_F \exp(i\theta(\omega)) K(\omega \cdot x) \frac{|F(\omega)|}{C_F}\, \mathrm{d}\omega \right) \\ &= \int_{[-M,M]^d} C_F [\cos(\theta(\omega)) K_R(\omega \cdot x) - \sin(\theta(\omega)) K_I(\omega \cdot x)] \frac{|F(\omega)|}{C_F}\, \mathrm{d}\omega, \end{aligned} \tag{6}$$

其中， K_R, K_I 分别是 K 的实部和虚部。鉴于该积分将 f 表示为其中函数的无限凸组合

$$Q_{K,M} = \left\{ \gamma[\cos(\beta) K_R(\omega \cdot x) - \sin(\beta) K_I(\omega \cdot x)] \;\middle|\; |\gamma| \leq C_F, \beta \in \mathbb{R}, \omega \in [-M, M]^d \right\},$$

那么 f 属于 $Q_{K,M}$ 的凸包的闭包。由于 $x \in [0,1]^d$ 和 $\omega \in [-M, M]^d$,$t = \omega \cdot x \in [-dM, dM]$。根据 D_K 的定义，$Q_{K,M}$ 中的函数在 $L^2(B)$- 范数下有界，界为 $2C_F D_K \leq 2C_F$。利用引理 3，存在实系数 $\{b_j\}$ 和 $\{\beta_j\}$，以及向量 $\omega_j \in [-M, M]^d$ $(1 \leq j \leq \lceil 1/\epsilon_0^2 \rceil)$，使得：

$$f_{\epsilon_0}(x) = \sum_{j=1}^{\lceil 1/\epsilon_0^2 \rceil} b_j [\cos(\beta_j) K_R(\omega_j \cdot x) - \sin(\beta_j) K_I(\omega_j \cdot x)], \quad \sum_{j=1}^{\lceil 1/\epsilon_0^2 \rceil} |b_j| \leq C_F,$$

使 $0 < \epsilon_0 < 1$ 可在稍后定义，从而

$$\|f_{\epsilon_0}(x) - f(x)\|_{L^2(B)} \leq 2C_F \epsilon_0.$$

现在，我们可以使用定义在 B 上的深度乘法神经网络 f^{MBL} 来逼近 $f_{\epsilon_0}(x)$。我们注意到， $K_R(x)$ 和 $K_I(x)$ 是满足定理 2 假设的解析核。因此，可以使用网络 f^{RMA}、 f^{IMA} 将它们逼近到 ϵ_0 误差，其深度和神经元数量为

$$G_{\epsilon_0} \sim L_{\epsilon_0} = \mathcal{O}\left(\frac{1}{\log_2 s} \log_2 \frac{C_K}{\epsilon_0}\right),$$

其中，RMA 表示“实乘法解析”，IMA 表示“虚乘法解析”。我们将乘法网络 $f^{\mathrm{MBL}}(x)$ 定义为

$$f^{\mathrm{MBL}}(x) = \sum_{j=1}^{\lceil 1/\epsilon_0^2 \rceil} b_j \left[\cos(\beta_j) f^{\mathrm{RMA}}(\omega_j \cdot x) - \sin(\beta_j) f^{\mathrm{IMA}}(\omega_j \cdot x) \right].$$

该网络的深度为 $L_{\epsilon_0}=\mathrm{O}(\frac{1}{\log_2 s}\log_2\frac{C_K}{\epsilon_0})$，并含有 $G_{\epsilon_0}=\mathrm{O}(\frac{1}{\epsilon_0^2}\frac{1}{\log_2 s}\log_2\frac{C_K}{\epsilon_0})$ 个神经元。

$$\begin{aligned}\|f^{\mathrm{MBL}}(x)-f_{\epsilon_0}(x)\|_{L^\infty(B)} &\le \sum_{j=1}^{\lceil 1/\epsilon_0^2\rceil}|b_j|\|f^{\mathrm{RMA}}(\omega_j\cdot x)-K_R(\omega_j\cdot x)\|_{L^\infty(B)}\\ &\quad+\sum_{j=1}^{\lceil 1/\epsilon_0^2\rceil}|b_j|\|f^{\mathrm{IMA}}(\omega_j\cdot x)-K_I(\omega_j\cdot x)\|_{L^\infty(B)}\\ &\le 2C_F\epsilon_0,\end{aligned}$$

这意味着

$$\|f^{\mathrm{MBL}}(x)-f(x)\|_{L^2(B)}\le\|f^{\mathrm{MBL}}(x)-f_{\epsilon_0}(x)\|_{L^2(B)}+\|f_{\epsilon_0}(x)-f(x)\|_{L^2(B)}\le 4C_F\epsilon_0,$$

其中最后一个不等式是因为

$$\|g\|_{L^2(B)}=\left(\int_B|g|^2\right)^{1/2}\le\|g\|_{L^\infty(B)}=\|g\|_{L^\infty(B)}.$$

取 $\epsilon_0=\epsilon/(4C_F)$即可得到所需结果。 □

我们进一步考察定理 3 中的常数 C_K 随 M 和 s 变化的情况。令 $K(x)=\exp(ix)$ 为一个示例核， $x\in[-M,M]^d$。我们注意到，对于 $s>2$， $a_s^M=M(s+s^{-1})/2$ 是较长轴，因此 K 在 E_s^M 上的最大范数为 $K(a_s^M)$:

$$\max_x|K(x)|\le\exp(M\tfrac{s+s^{-1}}{2})=C_K(s,M).$$

在我们的设定中，输入位于区间 $t=\omega\cdot x\in[-dM,dM]$ 内，因此我们可以使用界常数

$$\max_t|K(t)|\ \le\ \exp(dM\tfrac{s+s^{-1}}{2})=C_K(s,dM).$$

由此得到的网络 f^{MBL} 的深度为

$$L_\epsilon\ =\ \mathcal{O}\left(\frac{1}{\log_2 s}\left(dM\frac{s+s^{-1}}{2}+\log_2\frac{C_F}{\epsilon}\right)\right)$$

and

$$G_\epsilon=\mathcal{O}\left(\frac{C_F^2}{\epsilon^2\log_2 s}\left(dM\frac{s+s^{-1}}{2}+\log_2\frac{C_F}{\epsilon}\right)\right)$$

神经元。在这种情况下，我们看到，较大的频带 M 和较大的维数 d 都会对第一项产生线性影响。

定理 4. 设 $d\ge 2, r\in\mathbb{N}$, $f\in\mathcal{B}_{2r,2}$和 $\epsilon>0$。存在一个深度 *ReLU* 网络 f^{RS}（简称“*ReLU Sobolev*”），其深度为 $L_\epsilon=\mathcal{O}(d^2\epsilon^{-2/r})$，包含 $G_\epsilon=\mathcal{O}(d^2\epsilon^{-(2+2/r)})$ 个神经元，使得$\left\|f^{\mathrm{RS}}-f\right\|_{L_2(B)}\le\epsilon$。

证明。 设 $f\in\mathrm{B}_{2r,2}$和 $M>1$。与定理 5 的证明类似，我们定义f 的带限化： $\mathbb{R}^d\to\mathbb{R}$ 如下

$$f_M\ =\ \mathcal{F}^{-1}(\mathcal{F}f_{\mathbb{I}^{[-M,M]^d}}).$$

定义 F： $[-M,M]^d\to\mathbb{C}$:

$$F(\omega)\ =\ \frac{1}{(2\pi)^d}(\mathcal{F}f)(\omega),$$

以及 $K(x) = \exp(ix)$。由此可得以下恒等式：

$$f_M(x) = \int_{[-M,M]^d} F(\omega)K(\omega \cdot x)\,\mathrm{d}\omega = \int_{[-M,M]^d} \tfrac{1}{(2\pi)^d}\mathcal{F}f(\omega)\exp(i\omega \cdot x)\,\mathrm{d}\omega.$$

因此很容易看出 $f_M \in \mathrm{H}_{K,\,M}(B)$。我们选择 $M = (2/\epsilon)^{1/r}$。按照定理 5 证明中的相同推导，我们寻求用网络 f^{RS} 逼近 f_M，使得

$$\|f_M - f^{\mathrm{RS}}\|_{L_2(B)} \le \epsilon/2.$$

于是我们得到

$$\begin{aligned}\|f - f^{\mathrm{RS}}\|_{L_2(B)} &\le \|f - f_M\|_{L_2(B)} + \|f_M - f^{\mathrm{RS}}\|_{L_2(B)} \\ &\le M^{-r} + \epsilon/2 \le \epsilon.\end{aligned} \tag{7}$$

我们考虑 $C_F = \int_{[-M,M]^d} |F(\omega)|\,\mathrm{d}\omega = \frac{1}{(2\pi)^d}\int_{[-M,M]^d} |\mathcal{F}f(\omega)|\,\mathrm{d}\omega \le 1$，其中该不等式由 $\mathrm{B}_{2r,\,2}$ 的定义得出。核 $K(t) = \exp(it)$ 的输入为 $t = \omega \cdot x$，其中 $x \in B$ 和 $\omega \in [-M,M]^d$。因此，$t \in [-dM, dM]$，且 $K\colon [-dM,dM] \to \mathbb{R}$。我们注意到 K 可以解析延拓至 Bernstein 四椭圆 E_4^{dM}。我们注意到 $a_4^{dM} = dM(4 + 4^{-1})/2$ 是较长轴，因此 K 在 E_s^{dM} 上的最大范数为 $K(a_4^{dM})$：

$$C_K = \max_t |K(t)| \le \exp(dM\tfrac{4+4^{-1}}{2}).$$

此外，对于任意 $t \in \mathbb{R}$，我们有 $|K(t)| \le 1 = D_K$。

使用定理 3。 2，根据 Montanelli 等人，2021 的结果，我们可以构造一个深度 ReLU 网络 f^{RS}，使得

$$\|f_M - f^{\mathrm{RS}}\|_{L_2(B)} \le \epsilon/2$$

其深度为

$$\begin{aligned}L_\epsilon &\le C_1 \frac{1}{\log_2^2 4}\log_2^2 \frac{2C_F C_K}{\epsilon} \\ &\le C_2 \frac{1}{\log_2^2 4}\left(\log_2 \frac{\exp(dM\frac{4+4^{-1}}{2})}{\epsilon}\right)^2 \\ &\le 12C_2\left(d\left(\frac{2}{\epsilon}\right)^{\frac{1}{r}} - \log_2\frac{1}{\epsilon}\right)^2 \\ &\le C_3 d^2 \epsilon^{-\frac{2}{r}},\end{aligned} \tag{8}$$

for 某些常数 $C_1, C_2, C_3 > 0$。此外，神经元的数量可以界定为

$$\begin{aligned}G_\epsilon &\le C_1' C_F^2 \frac{1}{\epsilon^2 \log_2^2 4}\log_2^2 \frac{2C_K C_F}{\epsilon} \\ &\le C_2' \frac{1}{\epsilon^2 \log_2^2 4}\left(\log_2 \frac{\exp(dM\frac{4+4^{-1}}{2})}{\epsilon}\right)^2 \\ &\le \frac{12C_2'}{\epsilon^2}\left(d\left(\frac{2}{\epsilon}\right)^{\frac{1}{r}} - \log_2\frac{1}{\epsilon}\right)^2 \\ &\le C_3' d^2 \epsilon^{-(2+\frac{2}{r})},\end{aligned} \tag{9}$$

对于某些常数 $C_1', C_2', C_3' > 0$。 □