arXiv:2304.01880v1 [cs.LG] 4 Apr 2023

权重受限神经网络逼近的测度论结果

Vugar E. Ismailov*[1] 和 Ekrem Savas[2]

[1]阿塞拜疆巴库数学与力学研究所[2]土耳其伊斯坦布尔伊斯坦布尔商业大学数学系

摘要。本文研究权重沿有限多个方向变化、阈值取自某个开区间的单隐层神经网络的逼近性质。我们得到了此类网络在连续函数空间中稠密的一个充要测度论条件。此外，我们还证明了一个稠密性结果，其中神经网络采用特定构造的激活函数，且神经元数量固定。

2010 年数学学科分类：主要分类 41A30、41A63、92B20；次要分类 28A33、46E27。

关键词：神经网络；激活函数；稠密性；正交测度；博雷尔测度；弱* 收敛；闪电式；轨道

1. 引言

过去 30 年来，人工神经网络一直是一个充满活力的研究领域。如今，神经网络已成功应用于计算机科学、金融、医学、工程、物理等各种领域。神经网络最大的优势或许在于，它们可以用作任意函数的逼近机制。本文关注神经网络中单隐层感知器模型的稠密性问题（即以任意精度进行逼近）。隐层含有 r 个单元、输入 $\mathbf{x} = (x_1, ..., x_d)$ 的单隐层感知器模型计算如下形式的函数

$$\sum_{i=1}^{r} c_i \sigma(\mathbf{w}^i \cdot \mathbf{x} - \theta_i), \tag{1.1}$$

其中，权重 $\mathbf{w}^i$ 是 $\mathbb{R}^d$中的向量，阈值 θ_i 和系数 c_i 是实数，激活函数 σ 是单变量函数，本文假设其连续。对于各种激活函数 σ，许多论文已经证明，可以用形式如 (1.1) 的函数（r 不固定！）在 $\mathbb{R}^d$的任意紧子集上很好地逼近给定的连续函数 $d \geq 1$。换言之，集合

*通讯作者地址：阿塞拜疆巴库 AZ1141，B. Vahabzadeh 街 9 号，数学与力学研究所；电子邮箱：vugaris@mail.ru

$$\mathcal{M}(\sigma) = span\ \{\sigma(\mathbf{w}\cdot\mathbf{x}-\theta):\ \theta\in\mathbb{R},\ \mathbf{w}\in\mathbb{R}^d\}$$

在一致收敛于所有紧集的拓扑下，在空间 $C(\mathbb{R}^d)$ 中稠密（例如，参见 [2, 3, 4, 5, 9, 14, 32]）。这类更一般的结果归功于 Leshno、Lin、Pinkus 和 Schocken [20]。他们证明，任意连续激活函数具有稠密性这一性质的充要条件是它不是多项式。该结果表明，只要 σ 连续，单隐层感知器模型在激活函数 σ的所有可能选择下都有效。

我们回顾一下，网络能够在 d–维空间的任意紧子集上逼近任意多元函数，这一性质称为通用逼近性质。如上所述，具有非多项式激活函数的单隐层网络具有这一性质。已有若干结果（例如，见 [8, 14, 26, 32]）表明，权重集合受到适度限制的单隐层感知器仍具有通用逼近性质。但是，如果权重取自过于“狭窄”的集合，通用逼近性质通常就不再成立，因此产生了这样的问题：确定紧集 $X\subset\mathbb{R}^d$，使得所考虑的网络能够在 X上任意逼近任意给定的连续函数。在 [13], 中，第一作者研究了权重在有限条直线上变化、阈值取自全体实数的神经网络的逼近问题。他得到了这类网络实现良好逼近的一个充分条件和一个必要条件。结果证明，这两个条件并不等价；而获得一个既必要又充分的条件这一问题仍未解决。

本文的主要目的是，针对权重取自有限条直线集合、阈值取自某个开区间的单隐层网络，给出其稠密性的必要且充分条件。所得结果（见定理 2.4 和 2.5）基于所谓脊函数以及与这类函数正交的测度的性质。这些结果确定了所考察网络模型有效性的理论边界。也就是说，它们以射影测度为依据，刻画了紧集 $X\subset\mathbb{R}^d$，在这些集合上，该模型仍能保持其一般能力，对任意连续多元函数进行任意精度的逼近。我们还证明，存在某些光滑激活函数，即使将神经元数量固定，所得的稠密性仍然成立（见定理 2.6）。

2. 测度论结果

本节给出一个必要且充分的条件，用于判定能否通过权重和阈值取自特定有限集合的单隐层感知器进行逼近。更具体地说，我们允许权重沿有限个方向变化，并允许阈值在一个开区间内变化。

假设给定 k 个方向 $\mathbf{a}^j\in\mathbb{R}^d\backslash\{\mathbf{0}\}$, $j=1,...,k$。设 $L_j=\{t\mathbf{a}^j:\ t\in\mathbb{R}\}$,$j=1,...,k$，并且 $W=\cup_{j=1}^k L_j$. 设 Θ 是 $\mathbb{R}$ 中的一个开区间，且 $\sigma:\ \mathbb{R}\to\mathbb{R}$ 是一个连续

激活函数。考虑以下权重取自 W 、阈值取自 Θ的单隐层感知器集合:

$$\mathcal{M}(\sigma; W, \Theta) = span\{\sigma(\mathbf{w} \cdot \mathbf{x} - \theta) : \ \mathbf{w} \in W, \ \theta \in \Theta\}. \tag{2.1}$$

注意， $\mathcal{M}(\sigma; W, \Theta)$中的元素是依赖于变量 $\mathbf{x} \in \mathbb{R}^d$的连续函数。我们提出并回答以下问题：对于哪些紧集 $X \subset \mathbb{R}^d$，来自 $\mathcal{M}(\sigma; W, \Theta)$ 的函数集（但视为定义在 X 上）在 $C(X)$ 中稠密？下文中，对于一族连续函数 $K \subset C(\mathbb{R}^d)$，记号 $\overline{K} = C(X)$（或表达式“K 在 $C(X)$ 中稠密”）表示将该函数族限制在 X 上后，在 $C(X)$ 中稠密。因此，我们要刻画满足 $\mathcal{M}(\sigma; W, \Theta) = C(X)$ 的紧集 X。下文将对此类集合给出测度论描述。

我们继续讨论有限符号测度 μ 及测度空间 $(U, \mathcal{A}, \mu)$ 的像这一概念。设 F是从集合 U 到集合 T. 的映射。若可测集 $B \in \mathcal{B}$ 是 T 的满足 $F^{-1}(B) \in \mathcal{A}$ 的子集，则测度空间 $(T, \mathcal{B}, \nu)$ 称为测度空间 $(U, \mathcal{A}, \mu)$ 的像，并且

$$\mu(F^{-1}(B)) = \nu(B), \text{ for all } B \in \mathcal{B}.$$

测度 ν 称为 μ 的像，记作 $F \circ \mu$。显然，

$$\|F \circ \mu\| \leq \|\mu\|,$$

由于在映射 F 下， μ 为正的那些集合与为负的那些集合的像可能会混淆。此外，注意，如果有界函数 $g: T \to \mathbb{R}$ 是 $F \circ \mu$-可测的，那么复合函数 $[g \circ F]: U \to \mathbb{R}$ 是 μ-可测的，并且

$$\int_U [g \circ F] d\mu = \int_T g d[F \circ \mu]. \tag{2.2}$$

众所周知，在泛函分析中，对于紧 Hausdorff 空间 X 以及连续函数空间 $C(X)$，子空间 $M \subset C(X)$ 在 $C(X)$ 中稠密，当且仅当与 M 正交的唯一 Borel 测度是零测度。我们将在进一步的分析中使用这一事实。在下文中，我们将取由所谓的脊函数的线性组合构成的集合作为子空间 M。脊函数最简单的形式是 $g(\mathbf{a} \cdot \mathbf{x})$，其中 g 是一元函数，$\mathbf{a} = (a_1, ..., a_d)$ 是 $\mathbb{R}^d \backslash \{\mathbf{0}\}$ 中的向量（方向）， $\mathbf{x} = (x_1, ..., x_d)$ 是变量，而 $\mathbf{a} \cdot \mathbf{x}$ 是内积。换言之，脊函数是一个一元函数与 $\mathbb{R}^d$. 上的线性泛函的复合。这些函数自然地出现在各个领域中。它们出现在偏微分方程中（在那里被称为平面波，参见例如 [15]）、计算机断层成像中（参见例如 [16, 21, 31]；脊函数这一名称由 Logan 和 Shepp 于 [21] 首次提出）、统计学中（尤其是在投影寻踪和投影回归理论中；参见例如 [1, 6, 10]）、逼近理论中（参见例如 [3, 11, 18, 19, 23, 25, 27]）以及神经网络理论中（参见例如 [3, 13, 22, 25, 26, 30, 33]）。关于脊函数的系统研究，参见 Pinkus 最近出版的专著 [28]。注意，这些函数

$\sigma(\mathbf{w}^i\cdot\mathbf{x}-\theta_i)$ 在 (1.1) 中是方向为 $\mathbf{w}^i$ 的脊函数。因此，神经网络理论中出现的许多逼近理论问题通常被归结为脊函数领域中的相应问题，这并不令人意外。

假设给定方向 k $\mathbf{a}^1$,..., $\mathbf{a}^k$. 考察脊函数的线性组合

$$\mathcal{R}\left(\mathbf{a}^1,...,\mathbf{a}^k\right)=\left\{\sum_{j=1}^{k}g_i\left(\mathbf{a}^i\cdot\mathbf{x}\right):\ g_i\in C(\mathbb{R}),\ i=1,...,k\right\}.$$

注意，我们固定方向 ^{i}，仅改变函数 g_i。因此 $\mathcal{R}\left(\mathbf{a}^1,...,\mathbf{a}^k\right)$ 是一个线性空间。这个空间通常出现在计算机断层摄影的数学问题中（例如，参见 [16, 21, 31]）。

设 X 是 $\mathbb{R}^d$. 的一个紧子集。用 $C^*(X)$ 表示定义在 X. 的 Borel子集上的正则测度类。当测度 $\mu\in C^*(X)$ 与 $\mathcal{R}\left(\mathbf{a}^1,...,\mathbf{a}^k\right)$正交? 也就是说，何时

$$\int_X f d\mu=0,$$

对每个函数 $f\in\mathcal{R}\left(\mathbf{a}^1,...,\mathbf{a}^k\right)$? 由 (2.2) 立即可得，$\mu$与 $\mathcal{R}\left(\mathbf{a}^1,...,\mathbf{a}^k\right)$ 正交，当且仅当射影测度

$$\mu_{\mathbf{a}^i}\overset{def}{=}\pi_i\circ\mu\equiv 0,\ i=1,...,k,$$

其中 π_i是沿方向 $^i\mathbf{a}$ 的自然投影。也就是说，$\pi_i(\mathbf{x})=\mathbf{a}^i\cdot\mathbf{x}$，$i=1,...,k$.由此得到以下引理。

引理2.1。子空间 $\mathcal{R}\left(\mathbf{a}^1,...,\mathbf{a}^k\right)$ 在 $C(X)$ 中稠密，当且仅当对于任意满足 $\mu_{\mathbf{a}^i}\equiv 0$ 的正则Borel测度 μ，都有 $\mu\equiv 0$。

让我们更仔细地考察一下引理2.1的假设。考虑一个有限集$\{\mathbf{x}^1,\ldots,\mathbf{x}^n\}\subset X$。对于每个 $i=1,...,k$，令集合 $\{\mathbf{a}^i\cdot\mathbf{x}^j,\ j=1,...,n\}$ 具有 s_i 个不同的值，记为 $\gamma_1^i,\gamma_2^i,...,\gamma_{s_i}^i$。对于每个 $i=1,...,k$，以及每个 $m=1,...,s_i$，考虑线性方程

$$\sum_j\lambda_j=0, \tag{2.3}$$

其中，求和遍历所有满足 $\mathbf{a}^i\cdot\mathbf{x}^j=\gamma_m^i$. 的指标 $j\in\{1,...,n\}$。合并这些方程，得到关于 $\lambda_1,...,\lambda_n$ 的 $\sum_{i=1}^n s_i$ 个齐次线性方程组成的方程组。注意，这些方程的系数是整数 0 和 1。如果方程组 (2.3) 有解 $\{\lambda_1,...,\lambda_n\}$，且 $\lambda_j\neq 0$, $j=1,...,n$，那么有限集$\{\mathbf{x}^1,\ldots,\mathbf{x}^n\}$ 称为关于方向 $\mathbf{a}^1,...,\mathbf{a}^k$ 的闭路径（见 [12]）。闭路径的存在意味着子空间 $\mathcal{R}\left(\mathbf{a}^1,...,\mathbf{a}^k\right)$ 在 $C(X)$ 中不稠密。事实上，对于闭路径 $l=\{\mathbf{x}^1,\ldots,\mathbf{x}^n\}$，我们可以构造一个测度

$$\mu_l = \sum_{j=1}^{n} \lambda_j \delta_{\mathbf{x}^j}, \tag{2.4}$$

其中 $\delta_{\mathbf{x}^j}$ 是 $\mathbf{x}^j$, $j = 1, ..., n$ 处的点质量。注意，该测度非零，且与子空间 $\mathcal{R}\left(\mathbf{a}^1, ..., \mathbf{a}^k\right)$ 正交。因此，若要使 $\mathcal{R}(\mathbf{a}^1, ..., \mathbf{a}^k)$ 在 X 中稠密，集合 X 必须不含闭路径。例如，设 $k = 3$, $\mathbf{a}^i, i = 1, 2, 3$ 与坐标方向重合，且 ε 为一个充分小的正数。若 X 包含集合$l = \{(0, 0, 0), (0, 0, \varepsilon), (0, \varepsilon, 0), (\varepsilon, 0, 0), (\varepsilon, \varepsilon, \varepsilon)\}$（或其平移），那么对于 (2.4) 中系数 λ_j 分别等于 $-2, 1, 1, 1$ 的非平凡测度 μ，我们有$\mu_{\mathbf{a}^i} = 0, i = 1, 2, 3$。也就是说，$\mu \in \mathcal{R}^{\perp}(\mathbf{a}^1, \mathbf{a}^2, \mathbf{a}^3)$，但 $\mu \neq 0$；因此 $\mathcal{R}(\mathbf{a}^1, \mathbf{a}^2, \mathbf{a}^3) \neq C(X)$。上述分析表明，根据引理2.1，只有没有内点的特定集合 X 才可能满足稠密性。

对于情形 $k = 2$，“对任意满足 $\mu_{\mathbf{a}^i} \equiv 0$ 的正则Borel测度 μ，都有 $\mu \equiv 0$”这一条件在几何上已有清晰的理解。在这种情形下，我们可以用由闪电链诱导的测度来改述这一条件。相对于两个方向 $\mathbf{a}^1$ 和 $\mathbf{a}^2$，闪电链（简称“链”）是 $\mathbb{R}^d$ 中有限或无限的有序点集 $(\mathbf{p}^1, \mathbf{p}^2, ...)$，其中 $\mathbf{p}^i \neq \mathbf{p}^{i+1}$ 与其单位向量 $\mathbf{p}^{i+1} - \mathbf{p}^i$ 交替地垂直于方向 $\mathbf{a}^1$ 和 $\mathbf{a}^2$（见 [17, 24]）。若有限闪电链 $(\mathbf{p}^1, ..., \mathbf{p}^n)$ 中 n 为偶数，且点集 $(\mathbf{p}^1, ..., \mathbf{p}^n, \mathbf{p}^1)$ 也构成一条闪电链，则称该闪电链为闭合闪电链。闪电链是几何上明确的对象，可有效用于研究稠密性~~问题。容易~~看出，若 $\mathcal{R}(\mathbf{a}^1, \mathbf{a}^2) = C(X)$，则 X 不含闭合闪电链。事实上，在一条闭合闪电链 $(\mathbf{p}^1, ..., \mathbf{p}^{2n})$ 上，可以构造非平凡测度

$$\mu = \sum_{j=1}^{2n} (-1)^{j+1} \delta_{\mathbf{p}^j},$$

对于其中 $\mu_{\mathbf{a}^1} \equiv 0$ 和 $\mu_{\mathbf{a}^2} \equiv 0$。不存在闭螺栓是密度 $\mathcal{R}(\mathbf{a}^1, \mathbf{a}^2) = C(X)$ 的一个非常简单且自然的必要条件。然而，这一条件并不充分。在一些非常复杂的情形下，X 不包含闭螺栓，但~~同时 $\mathcal{R}$~~$(\mathbf{a}^1, \mathbf{a}^2) \neq C(X)$。例如，假设 X 具有以下性质：

(1) X 是两个互不相交的闭集 X_1 和 X_2 的并，使得$\mathbf{a}^i \cdot X_1 = \mathbf{a}^i \cdot X_2, i = 1, 2$.

(2) X 不包含闭合螺栓。

请注意，存在具有这些性质的紧集（例如参见 [17]）。由性质 (1)-(2) 可知，X 包含一个无限螺栓，我们将其记为 $l =(\mathbf{p}^1, \mathbf{p}^2, ...)$。我们将这一螺栓与如下测度序列相关联。

$$\mu_n = \frac{1}{n} \sum_{j=1}^{n} (-1)^{j+1} \delta_{\mathbf{p}^j}, \ n = 1, 2, ... \tag{2.5}$$

显然，对于每个 $n.$，$\|\mu_n\| = 1$。由巴拿赫–阿劳格鲁定理（见 [29, 第 66] 页），序列 $\{\mu_n\}_{n=1}^{\infty}$ 存在一个弱* 极限点 $\mu^*.$。我们断言 $\mu^* \in \mathcal{R}^{\perp}(\mathbf{a}^1, \mathbf{a}^2)$ 且 $\mu^* \neq 0$。事实上，根据 bolt 的定义和 (2.5)，可得

$$\left|\int_X (g_1+g_2)d\mu_n\right| \le \frac{2}{n}(\|g_1\|+\|g_2\|),$$

对于任意 $g_1+g_2 = g_1(\mathbf{a}^1\cdot\mathbf{x})+g_2(\mathbf{a}^2\cdot\mathbf{x})$ 属于 $\mathcal{R}(\mathbf{a}^1,\mathbf{a}^2)$。令 n 趋于无穷大，得到

$$\int_X (g_1+g_2)d\mu^* = 0,$$

并且由于 g_1+g_2 是 $\mathcal{R}(\mathbf{a}^1,\mathbf{a}^2)$ 中的任意和，因此我们得出 $\mu^* \in \mathcal{R}^\perp(\mathbf{a}^1,\mathbf{a}^2)$。断言 $\mu^* \neq 0$ 显然成立，因为对于满足以下性质的连续函数 $f\colon X\to\mathbb{R}$：对所有 $x\in X_1$，都有 $f(x)=1$；对所有 $x\in X_2$，都有 $f(x)=-1$，我们有

$$\left|\int_X f d\mu_n\right| = 1$$

因此

$$\left|\int_X f d\mu^*\right| = 1.$$

在某些（值得考虑的）情况下，存在无限多个螺栓并不违反引理2.1的稠密性条件。首先注意，当 $\mathbf{x}$ 和 $\mathbf{y}$ 属于给定紧集 $X\subset\mathbb{R}^d$ 中的某个螺栓时，关系 $\mathbf{x}\sim\mathbf{y}$ 定义了一个等价关系。等价类称为轨道。如果 $X\subset\mathbb{R}^d$ 的轨道是拓扑闭集，那么 $\mathcal{R}(\mathbf{a}^1,\mathbf{a}^2)=C(X)$ 等价于 X 中不存在闭螺栓（见 [24]）。例如，设 $k=2$、$\mathbf{a}^1=(1,1)$、$\mathbf{a}^2=(1,-1)$ 和 X 为集合

$$\left\{(0,0),(1,-1),(0,-2),(-1\frac{1}{2},-\frac{1}{2}),(0,1),(\frac{3}{4},\frac{1}{4}),(0,-\frac{1}{2}),(-\frac{3}{8},-\frac{1}{8}),(0,\frac{1}{4}),(\frac{3}{16},\frac{1}{16}),...\right\}.$$

注意，X 是一个形成单一轨道的无限螺栓。由于该轨道是闭集，我们得到 $\mathcal{M}(\sigma;W,\Theta)=C(X)$。

以下命题是 Marshall 和 O'Farrell 针对两个代数之和所建立的已知一般结果的一个特例 [24]。

命题2.2。假设 X 是 $\mathbb{R}^d$的一个紧子集。空间 $\mathcal{R}(\mathbf{a}^1,\mathbf{a}^2)$在 $C(X)$中稠密，当且仅当 X 不包含闭螺栓，并且对于每个无限螺栓 $(\mathbf{p}^1,\mathbf{p}^2,...)$，$\mu_n$ 弱* 收敛到零 $\subset X$.

我们已经讨论了引理2.1的假设。也就是说，我们分析了满足 $\mu_{\mathbf{a}^i}\equiv 0$ 的任意正则Borel测度 μ 为零测度的情形。我们的分析总结如下列推论。

推论2.3。假设给定一个紧集 $X\subset\mathbb{R}^d$和方向$\mathbf{a}^1,...,\mathbf{a}^k\in\mathbb{R}^d\backslash\{\mathbf{0}\}$.

1) 设 $k>2$。如果对于任何满足 $\mu_{\mathbf{a}^i}\equiv 0$ 的 $C^*(X)$ 中的正则Borel测度 μ，都有$\mu\equiv 0$，那么集合 X 必然不包含闭路径。

2) 对于任何满足 $\mu_{\mathbf{a}^i} \equiv 0$ 的 $C^*(X)$ 中的正则Borel测度 μ，我们有$\mu \equiv 0$ 当且仅当集合 X 不包含闭螺栓，并且对于任何无限螺栓 $(\mathbf{p}^1, \mathbf{p}^2, ...)$，若存在，则 (2.5) 中的测度序列 μ_n 弱*收敛于零。若 X 的轨道是闭的，则 μ_n 总是弱* 收敛于零。

注。请注意，推论 2.3 仅在 $k = 2$ 情形下是完整的。尝试将（未闭合的）闪电形推广到 $k > 2$ 情形时，我们会遇到巨大的组合困难。我们不知道如何合理地描述点序列 $(\mathbf{p}^1, \mathbf{p}^2, ...)$ 和加权点质量测度 μ_n （类似于 (2.5)），使得 μ_n 的任意弱* 极限点都与 $\mathcal{R}\left(\mathbf{a}^1, ..., \mathbf{a}^k\right)$ 正交。对于 $k > 2$，与 $\mathcal{R}\left(\mathbf{a}^1, ..., \mathbf{a}^k\right)$ 正交的测度的完整描述似乎超出了本文所讨论方法的范围。即使在最简单的情形下，当 $k = d$ 和 $\mathbf{a}^i$ 是 $\mathbb{R}^d$ 中的坐标方向时，这个问题仍未解决（例如参见 [17, 24]）。

现在，我们将上述内容应用于神经网络$\mathcal{M}(\sigma; W, \Theta)$ 的逼近问题（见 (2.1)）。下面的定理成立。

定理2.4. 假设 $\sigma : \mathbb{R} \to \mathbb{R}$ 是一个连续的非多项式激活函数，且 X是 $\mathbb{R}^d$的一个紧子集。那么，当且仅当对于满足 $\mu_{\mathbf{a}^i} \equiv 0$的任意测度 $\mu \in C^*(X)$，都有 $\mu \equiv 0$时，集合 $\mathcal{M}(\sigma; W, \Theta)$ 在 $C(X)$ 中稠密。

证明。充分性。设对于满足 $\mu_{\mathbf{a}^i} \equiv 0$的任意测度 $\mu \in C^*(X)$，都有 $\mu \equiv 0$。由引理 2.1可知，集合 $\mathcal{R}\left(\mathbf{a}^1, ..., \mathbf{a}^k\right)$ 在 $C(X)$ 中稠密。这意味着，对于任意函数 $f \in C(X)$ 和任意正实数 ε，存在连续的一元函数 $g_i, i = 1, ..., k$，使得

$$\left| f(\mathbf{x}) - \sum_{i=1}^{k} g_i\left(\mathbf{a}^i \cdot \mathbf{x}\right) \right| < \frac{\varepsilon}{k+1} \tag{2.6}$$

对所有 $\mathbf{x} \in X$ 都成立。由于 X是紧的，集合 $Y_i = \{\mathbf{a}^i \cdot \mathbf{x} : \mathbf{x} \in X\}$, $i = 1, 2, ..., k$也都是紧集。注意，若 σ 不是多项式，则集合

$$span\left\{\sigma(ty - \theta) : \ t \in \mathbb{R}, \ \theta \in \Theta\right\}$$

在一致收敛拓扑下稠密于 $C(\mathbb{R})$（见 [26]）。这一稠密性结果意味着，对于给定的 ε，存在数值 c_{ij}、$t_{ij} \in \mathbb{R}$、$\theta_{ij} \in \Theta$、$i = 1$、2、...、k、$j = 1$、...、m_i，使得

$$\left| g_i(y) - \sum_{j=1}^{m_i} c_{ij}\sigma(t_{ij}y - \theta_{ij}) \right| < \frac{\varepsilon}{k+1} \tag{2.7}$$

对于所有 $y \in Y_i$、$i = 1$、2、...、$k.$，由（2.6）和（2.7）可得

$$\left\| f(\mathbf{x}) - \sum_{i=1}^{k} \sum_{j=1}^{m_i} c_{ij}\sigma(t_{ij}\mathbf{a}^i \cdot \mathbf{x} - \theta_{ij}) \right\|_{C(X)} < \varepsilon. \tag{2.8}$$

因此 $\overline{\mathcal{M}(\sigma; W, \Theta)} = C(X)$。

必要性。设 X 是 $\mathbb{R}^d$ 的紧子集，且集合 $\mathcal{M}(\sigma; W, \Theta)$ 在 $C(X)$ 中稠密。那么，对于任意正实数 ε，不等式（2.8）对某些系数 c_{ij}、t_{ij} 以及 θ_{ij}、$i = 1$、2、$j = 1$、...、m_i. 成立。由于对于每个 $i = 1$、2、...、k，函数 $\sum_{j=1}^{m_i} c_{ij}\sigma(t_{ij}\mathbf{a}^i \cdot \mathbf{x} - \theta_{ij})$ 具有 $g_i(\mathbf{a}^i \cdot \mathbf{x})$ 的形式，子空间 $\mathcal{R}\left(\mathbf{a}^1、...、\mathbf{a}^k\right)$ 在 $C(X)$ 中稠密。于是，根据引理2.1，对于满足 $\mu_{\mathbf{a}^i} \equiv 0$ 的任意测度 $\mu \in C^*(X)$，我们有 $\mu \equiv 0$。定理2.4得证。 □

对于情形 形 $k = 2$，应用命题 2.2 可得以下定理 .

定理 2.5。假设 σ： $\mathbb{R} \to \mathbb{R}$ 是连续的非多项式激活函数，X 是 $\mathbb{R}^d$ 的紧子集，且 $k = 2$. 那么，如果且仅当 X 不含闭螺栓，并且对于每个无限螺栓 $(\mathbf{p}^1, \mathbf{p}^2, ...) \subset X$，式 (2.5) 中的测度 μ_n 弱收敛* 到零时，集合 $\mathcal{M}(\sigma; W, \Theta)$ 在 $C(X)$ 中稠密。

例子。

(1) 设 $X \subset \mathbb{R}^d$ 是一个含有内点的紧集。则在该点的一个开邻域（包含于 X）内，我们可以构造一个非平凡的形如 (2.4) 的测度。因此，由定理 2.4， $\mathcal{M}(\sigma; W, \Theta) \neq C(X)$。

(2) 设给定方向 $\{\mathbf{a}^j\}_{j=1}^k$ 以及 γ中的一条紧曲线 $\mathbb{R}^d$，使得对于任意 $c \in \mathbb{R}$， γ与每个超平面 $\mathbf{a}^j \cdot \mathbf{x} = c$、$j = 1$、...、$r$.至多有一个公共点。那么，显然，对于任意测度 $\mu \in C^*(X)$，条件 $\mu_{\mathbf{a}^i} \equiv 0$ 蕴含 $\mu \equiv 0$。因此，$\mathcal{M}(\sigma; W, \Theta)$ 在 $C(X)$ 中稠密。

(3) 设 $k = 2$ and X 为两条平行线段的并集，且它们不垂直于给定方向 $\mathbf{a}^1$ and $\mathbf{a}^2$。那么 X 不包含闭合螺栓和无限螺栓。因此，根据定理 2.5.，$\mathcal{M}(\sigma; W, \Theta) = C(X)$。

上述定理意味着，为了得到稠密性结果，我们有很多选择：必须从“非多项式”集合中选取激活函数 σ。尽管我们能够对权重和阈值作出一些限制，但单隐层中的神经元数量可能会随着逼近精度的要求而增加。也就是说，如果定理 2.4 的条件得到满足，那么对于给定的 $f \in C(X)$ 和 $\varepsilon > 0$，存在一个网络$h \in \mathcal{M}(\sigma; W, \Theta)$，使得

$$\|f - h\| < \varepsilon.$$

但这个不等式并不意味着，对于所有 f 和 ε， h 中的神经元数量都一致有界。需要注意的是，在实际应用中，必须确定隐藏层中应当设置多少个神经元。神经元数量越多，网络给出精确结果的可能性就越大。但遗憾的是，随着隐藏层中神经元数量的增加，实用性会降低。有若干针对这一问题的研究结果表明，即使隐藏层中的神经元数量一致有界，具有两个或更多隐藏层的神经网络仍能以任意高的精度逼近任何给定的连续多元函数（例如，参见 [7, 22, 26]）。

以下结果使我们不仅可以限制权重集合，还可以限制近似单隐层网络中的神经元数量。事实证明，在非多项式函数中，存在光滑激活函数 σ ，即使单隐层中的神经元数量有限，稠密性仍然成立。

我们用 $\mathcal{M}_r(\sigma; W; \mathbb{R})$表示 $\mathcal{M}(\sigma; W; \mathbb{R})$中隐藏层恰有 r个神经元的网络集合。也就是说，

$$\mathcal{M}_r(\sigma; W; \mathbb{R}) = \left\{ \sum_{i=1}^{r} c_i \sigma(\mathbf{w}^i \cdot \mathbf{x} - \theta_i) : \ c_i \in \mathbb{R}, \ \mathbf{w}^i \in W, \ \theta_i \in \mathbb{R}, \ i = 1, ..., r \right\}.$$

定理 2.6. 设 X 是 $\mathbb{R}^d$的一个紧子集。此外，设$\mathbf{a}^j \in \mathbb{R}^d \backslash \{\mathbf{0}\}$, $j = 1, ..., k$，且 $W = \cup_{j=1}^{k} \{t \mathbf{a}^j : \ t \in \mathbb{R}\}$. 。那么存在一个无穷可微的激活函数$\sigma$：$\mathbb{R} \to \mathbb{R}$，使得以下结论成立：集合 $\mathcal{M}_k(\sigma; W, \mathbb{R})$在$C(X)$中稠密，当且仅当对于满足$\mu_{\mathbf{a}^i} \equiv 0$的任意测度 $\mu \in C^*(X)$，都有 $\mu \equiv 0$。

证明。充分性。设对于满足$\mu_{\mathbf{a}^i} \equiv 0$的任意测度 $\mu \in C^*(X)$，都有 $\mu \equiv 0$。那么根据引理2.1，集合 $\mathcal{R}\left(\mathbf{a}^1, ..., \mathbf{a}^k\right)$ 在 $C(X)$中稠密。这意味着，对于任意函数 $f \in C(X)$和任意正实数 ε ，都存在连续单变量函数 g_i, $i = 1, ..., k$，使不等式(2.6)成立。由于 X 是紧的，集合 $Y_i = \{\mathbf{a}^i \cdot \mathbf{x} : \mathbf{x} \in X\}$, $i = 1, ..., k$也都是紧的。假设所有集合 Y_i 都包含在线段 $[-l, l]$.中

令 α 为任意数。将区间 $[\alpha, +\infty)$ 分成若干段 $[\alpha, 2\alpha], [2\alpha, 3\alpha]$, ... 我们分两个阶段构造满足定理 2.5 假设的 σ。令 $\{p_m(t)\}_{m=1}^{\infty}$ 为定义在 $[-l, l]$. 上的有理系数多项式序列。首先，我们在闭区间 $[(2m-1)\alpha, 2m\alpha]$, $m = 1, 2, ...$ 上定义 σ为如下函数

$$\sigma(t) = p_m(\frac{2l}{\alpha} t - 4ml + l), \ t \in [(2m-1)\alpha, 2m\alpha],$$

或者等价地，

$$\sigma(\frac{\alpha}{2l} t + 2m\alpha - \frac{\alpha}{2}) = p_m(t), \ t \in [-l, l]. \tag{2.9}$$

在第二阶段，我们将 σ 延拓到区间 $(2m\alpha, (2m+1)\alpha)$ $m = 1, 2, ...$, 和 $(-\infty, \alpha)$ 上，同时保持 C^∞ 性质。

对于任意单变量函数 $g \in C(\mathbb{R})$ 和任意 $\varepsilon > 0$ ，存在一个有理系数多项式 $p(t)$，使得

$$|g(t) - p(t)| < \frac{\varepsilon}{k},$$

对所有 $t \in [-l, l]$. ，这与 (2.9) 一起意味着

$$\left| g(t) - \sigma(\frac{\alpha}{2l} t - r) \right| < \frac{\varepsilon}{k},$$

对于某个 $r \in \mathbb{R}$ 以及所有 $t \in [-l, l]$. 。因此，对于式 (2.6) 中的函数 g_i ，我们可以写成

$$\left|g_i(\mathbf{a}^i\cdot\mathbf{x}) - \sigma(\lambda\mathbf{a}^i\cdot\mathbf{x} - r_i)\right| < \frac{\varepsilon}{k}, \tag{2.10}$$

其中 $\lambda = \frac{\alpha}{2l}$. 。由式 (2.6) 和 (2.10) 可得

$$\left|f(\mathbf{x}) - \sum_{i=1}^{k}\sigma(\lambda\mathbf{a}^i\cdot\mathbf{x} - r_i)\right| < \varepsilon \tag{2.11}$$

不等式(2.11)可以写成如下形式

$$\left|f(\mathbf{x}) - \sum_{i=1}^{k}c_i\sigma(\mathbf{w}^i\cdot\mathbf{x} - \theta_i)\right| < \varepsilon, \tag{2.12}$$

其中，$\mathbf{w}^i \in W$、 $c_i = 1$和 θ_i是一些实数。这意味着集合 $\mathcal{M}_k(\sigma; W, \mathbb{R})$在 $C(X)$中稠密。

必要性。设 X是 $\mathbb{R}^d$的一个紧子集，并假设集合 $\mathcal{M}_k(\sigma; W, \mathbb{R})$在 $C(X)$中稠密。那么，对于任意正实数 ε，不等式(2.12)都成立，其中某些系数为 c_i、 $\theta_i \in \mathbb{R}$和$\mathbf{w}^i \in W$，且 $i = 1$，...，k. 。由于函数 $\sum_{i=1}^{k} c_i\sigma(\mathbf{w}^i\cdot\mathbf{x} - \theta_i)$是$\sum_{i=1}^{k} g_i(\mathbf{a}^i\cdot\mathbf{x})$形式的函数，因此子空间 $\mathcal{R}\left(\mathbf{a}^1, ..., \mathbf{a}^k\right)$在 $C(X)$中稠密。于是，由引理2.1，对于任意满足 $\mu_{\mathbf{a}^i} \equiv 0$的测度 $\mu \in C^*(X)$，有 $\mu \equiv 0$。 □

参考文献

[1] E. J. Cand`es, 《脊小波：利用脊函数进行估计》, 《统计年刊》31（2003）, 1561-1599。[2] T. Chen 和 H. Chen, 《利用神经网络逼近连续泛函及其在动态系统中的应用》, 《IEEE 神经网络汇刊》4（1993）, 910-918。[3] C. K. Chui 和 X. Li, 《利用脊函数和单隐层神经网络进行逼近》, 《逼近理论杂志》70（1992）, 131-141。[4] D. Costarelli 和 R. Spigler, 《通过叠加 S 型函数进行构造性逼近》, 《分析理论与应用》29（2013）, 169–196。[5] G. Cybenko, 《通过 S 型函数的叠加进行逼近》, 《数学、控制、信号与系统》2（1989）, 303-314。[6] J. H. Friedman、W. Stuetzle, 《投影寻踪回归》, 《美国统计学会杂志》76（1981）, 817-823。[7] G. Gripenberg, 《利用每层节点数有界的神经网络进行逼近》, 《逼近理论杂志》122（2003）, 260-266。

[8] B. I. Hong, N. Hahm，具有正整数权重的神经网络构造性逼近，Korean J. Math. 23 (2015), 327-336.[9] K. Hornik，多层前馈网络的逼近能力，Neural Networks 4 (1991), 251-257.[10] P. J. Huber，投影追踪，Ann. Statist. 13 (1985), 435-475.[11] V. E. Ismailov，脊函数极值和的刻画，J. Comp. Appl. Math. 205 (2007), 105-115.[12] V. E. Ismailov，关于线性叠加表示，J. Approx. Theory 151 (2008), 113-125.[13] V. E. Ismailov，在有限个方向上权重变化的神经网络逼近，J. Math. Anal. Appl. 389 (2012), no. 1, 72-83.[14] Y. Ito，通过带有或不带缩放的 sigmoid 函数平移旋转的线性组合逼近 $\mathbb{R}^d$ 上的连续函数，Neural Networks 5 (1992), 105-115.[15] F. John，《应用于偏微分方程的平面波与球面平均》，Interscience, New York, 1955.[16] I. Kazantsev，使用脊函数从任意方向进行断层重建，Inverse Problems 14 (1998), 635-645.[17] S.Ya.Khavinson，通过线性叠加实现最佳逼近（近似诺莫图法），由 D. Khavinson 根据俄文手稿译出。数学专著译丛，第159卷。American Mathematical Society, Providence, RI, 1997, 175页.[18] A. Kro´o，关于脊函数逼近，Constr. Approx. 13 (1997), 447-460.[19] V. Ya Lin, A. Pinkus，脊函数的基本性，J. Approx. Theory 75 (1993), 295-311.[20] M. Leshno, V. Ya. Lin, A. Pinkus and S. Schocken，具有非多项式激活函数的多层前馈网络能够逼近任意函数，Neural Networks 6 (1993), 861-867.[21] B. F. Logan and L. A. Shepp，根据函数投影进行最优重建，Duke Math. J. 42 (1975), 645-659.[22] V. Maiorov and A. Pinkus，MLP 神经网络逼近的下界，Neurocomputing 25 (1999), 81-91.

[23] V. E. Maiorov, 《关于脊函数的最佳逼近》, 《逼近理论杂志》99（1999）, 68-94。[24] D. E. Marshall 和 A. G. O'Farrell, 《用两个代数之和进行逼近：闪电原理》, 《泛函分析杂志》52（1983）, 353-368。[25] P. P. Petrushev, 《脊函数与神经网络的逼近》, 《SIAM 数学分析杂志》30（1998）, 155-189。[26] A. Pinkus, 《神经网络中 MLP 模型的逼近理论》, 《数值学报》8（1999）, 143-195。[27] A. Pinkus, 《用脊函数进行逼近》, 载于《曲面拟合与多分辨率方法》（A. Le M´ehaut´e、C. Rabut 和 L. L. Schumaker 编）, 范德比尔特大学出版社（纳什维尔）, 1997, 279-292。[28] A. Pinkus, 《脊函数》, 剑桥数学论丛第205卷, 剑桥大学出版社, 剑桥, 2015, 218页。[29] W. Rudin, 《泛函分析》, 麦格劳-希尔高等数学丛书。麦格劳-希尔图书公司, 1973, 397页。[30] M. Sanguineti, 《脊计算模型与神经网络的通用逼近：综述》, 《开放应用数学杂志》2（2008）, 31–58。[31] L. A. Shepp、J. B. Kruskal, 《计算机断层扫描：新的医学 X 射线技术》, 《美国数学月刊》85（1978）, 420–439。[32] M. Stinchcombe 和 H. White, 《使用权重有界的多层前馈网络逼近和学习未知映射》, 载于《1990年 IEEE 国际神经网络联合会议论文集》, 1990, 第3卷, IEEE, 纽约, 7-16。[33] T. F. Xie、F. L. Cao, 《多项式的脊函数表示及其在神经网络中的应用》, 《数学学报（英文版）》27（2011）, 2169–2176。