arXiv:2601.00455v1 [cs.LG] 1 Jan 2026

深度网络学习深层层次模型

Amit Daniely[1]

[1]耶路撒冷希伯来大学和 Google Research 特拉维夫

2026年1月5日

摘要

我们研究使用 n标签进行的监督学习，并证明，在残差网络上采用逐层随机梯度下降，可以高效地学习一类层次模型。该模型类别假设存在一个（未知的）标签层次结构 $L_1 \subseteq L_2 \subseteq \cdots \subseteq L_r = [n]$：其中，$L_1$ 中的标签是输入的简单函数；而对于$i > 1$， L_i 中的标签则是更简单标签的简单函数。

我们的类别超越了此前已被证明可由深度学习算法学习的模型，因为它达到了高效可学习性的深度极限。也就是说，该类别中存在需要多项式深度才能表达的模型，而此前的模型可以通过对数深度电路计算。

此外，我们认为，此类层次模型的可学习性最终可能成为理解深度学习的基础。除了它们天然适用于深度学习表现出色的定义域之外，我们还认为，人类“教师”的存在本身就支持这样一种假设：层次结构是内在可用的。通过提供细粒度标签，教师实际上揭示了大脑所使用的内部算法的“提示”或“片段”。我们将这一直觉形式化，表明在一个教师能够部分意识到其内部逻辑的简化模型中，会涌现出一种有助于高效学习的层次结构。

目录

1 引言 3 2 符号与预备知识 4 2.1 多项式阈值函数.4 2.2 强凸性.5 2.3 厄米特多项式.5 3 层次模型 6 3.1 “脑倾倒”层次结构.6 3.2 扩展到序列模型与集成模型.7 4 算法与主要结果 8 5 定理 4.3 的证明：残差网络的层次学习 9 6 结论与未来工作 14 7 更多预备知识 16 7.1 测度集中.16 7.2 杂项引理.17 7.3 一个泛化结果.17 7.4 核.18 7.5 随机特征方案.18 8 层级示例与定理 3.4 的证明 20 8.1 每个标签取决于 $O(1)$ 个更简单的标签.21 8.2 定理 3.4 的证明.22 9 随机神经元产生的核与引理 5.3 的证明 25

1 引言

深度学习理论的一个核心目标，是证明基于梯度的算法能够高效学习一类足够丰富、能够刻画现实的模型。这项研究始于十多年前，恰逢深度学习取得了毋庸置疑的实证成功。最初的理论结果证明，深度学习算法能够学习线性模型；随后，又有研究证明了简单非线性模型的可学习性。

这一进展非同寻常，尤其是考虑到直到最近，人们还不知道任何能够被深度学习算法证明可学习的模型。此外，该领域此前一直由困难性结果主导，这些结果表明神经网络的能力存在严重局限。然而，尽管取得了这些进展，学习线性模型或简单的非线性模型仍不足以解释深度学习在实践中的成功。

在本文中，我们通过证明深度学习算法——具体而言，在残差网络上进行逐层随机梯度下降——能够学习层次模型，推进了这项研究。我们考虑一个含有 n 个可能标签的监督学习场景，其中每个样本都与这些标签的一个子集相关联。令 $\mathbf{f}^*\colon \mathcal{X} \to \{\pm 1\}^n$ 为真实标注函数。我们假设存在一个未知的 标签层次结构 $L_1 \subseteq L_2 \subseteq \cdots \subseteq L_r = [n]$，其中，$L_1$ 中的标签是输入的简单函数（具体而言，是多项式阈值函数）；而对于 $i > 1$，L_i 中的任一标签都是较简单标签（即 L_{i-1}中的标签）的简单函数。

我们认为，层次模型的可学习性为理解深度学习提供了一个极具说服力的基础。首先，在神经网络表现出色的定义域中，层次模型十分自然。例如，在计算机视觉中，第一级标签可能是“这个像素是红色的”（即输入本身）；第二级标签可能是“曲线”或“暗色区域”；第三级标签则可能是“叶子”或“矩形”，以此类推。文本和语音处理中也存在类似的层次结构。事实上，这种层次结构促成了卷积网络和残差网络等成功架构的开发。

其次，我们甚至可以进一步主张，仅仅是人类“教师”的存在，就支持这样一种假设：层次标注确实存在，而且可以提供给算法。试想一个经典问题：识别图像中的汽车。早期人工智能方法（约在 20 世纪 70 至 80 年代）之所以失败，是因为它们试图手动编写人脑所使用的认知算法。后来，机器学习取代了这些方法，它根据输入-输出对来近似函数。尽管这种数据驱动的方法已经超越了人类的表现，但关于其成功的常见说法可能在某种程度上具有误导性。

我们认为，近期的突破并非仅仅源于“从头开始学习”，也得益于模型在包含大量细粒度标签的数据集上进行训练。这些标签介于显式编程和纯粹的输入-输出学习之间，为学习复杂概念提供了“提示”或中间步骤。尽管我们无法完全了解大脑内部的算法，但可以提供其逻辑的“片段”。通过识别低层特征——例如车窗、车轮或几何形状——我们实际上将任务分解成了一个层级。

从更宏观的角度来看，我们可以从以下视角思考大语言模型的诞生。从20世纪90年代至今，人类创造了互联网（网站、论坛、图像、视频等）。作为副产品，人类也无形中提供了大量标签和示例。由于这些标签数量庞大，涵盖从非常简单到极其复杂的内容，它们很可能具有层次结构。互联网诞生后，人们利用这些示例训练了大型模型；这些模型之所以取得成功，很大程度上要归功于这种结构（当然，也离不开海量数据和强大的计算能力）。从某种意义上说，互联网和现代大语言模型的演进，可以看作一项规模宏大的集体努力，旨在创造一个模拟人脑的电路：所有感兴趣的标签实际上都是这个电路与一个简单函数的复合。

我们对这一直觉提出了一种简化的形式化表述。我们将人脑建模为一个计算电路，其中每个标签（代表一个“脑片段”）对应于对大脑中某个神经元子集进行的多数表决。为了将这些标签既细粒度又多样化这一假设形式化，我们假设定义每个标签的具体神经元集合是在学习过程开始前随机选取的。我们证明，这种设定会产生一种层次结构，使残差网络能够高效学习。至关重要的是，残差网络架构和训练算法都不依赖于对这一内在标签层次结构的了解。

最后，我们指出，层次模型优于以往已证明可由随机梯度下降学习的模型类别。据我们所知，先前的结果仅限于可由对数深度电路实现的模型。相比之下，层次模型达到了高效可学习性的深度极限。对于任意多项式规模的电路，我们都可以构造一个相应的层次模型，使其能够通过残差网络上的随机梯度下降进行学习，从而有效地

将电路计算为其标签之一。

相关工作 线性模型或固定表示模型由固定的（通常为非线性的）特征映射及其后的可学习线性映射定义。这包括核方法、随机特征 [29], 等。过去十年中的多篇论文证明，神经网络能够学习各种线性模型，例如 [4, 15, 18, 11, 24, 9, 3, 12, 13]。一些研究考虑了超越固定表示的模型类别，但仍可通过神经网络上的基于梯度的方法高效学习。一类研究展示了在非均匀分布下学习奇偶函数的能力，或学习其他可由深度为二的神经网络直接表示的模型，例如 [25, 20, 32, 19, 14, 35, 33, 6, 8, 10]。与我们的方法更为接近的是 [1, 2, 16, 34] ，这些研究考虑了某些层次模型。如上所述，我们认为本研究是迈向能够捕捉现实的模型的又一步。从更形式化的角度来看，我们的工作优于以往研究，因为我们考虑的模型可以任意深。相比之下，所有提及的论文都考虑了可由对数深度网络实现的模型。事实上，除 [34] 考虑置换复合的研究之外，深度为二就足以表示上述所有模型。

另一条相关工作思路是 [23, 21, 27, 22] ，这些工作认为深度学习之所以成功，是因为层次结构。这一系列论文给出了一个可高效学习的层次模型示例，但推测该模型需要深层架构才能表达。其他试图论证层级对深度学习至关重要的工作包括 [28, 26, 7]

2 符号与预备知识

我们用粗体字母表示向量（例如，$\mathbf{x}$、$\mathbf{y}$、$\mathbf{z}$、$\mathbf{w}$、$\mathbf{v}$），并用普通字母表示其坐标。例如，x_i 表示 $\mathbf{x}$ 的第 i 个坐标。同样，我们用粗体字母（例如，$\mathbf{f}$、$\mathbf{g}$、$\mathbf{h}$、$\mathbf{p}$、$\mathbf{q}$、$\mathbf{r}$）表示向量值函数和多项式（即值域为 $\mathbb{R}^d$的函数和多项式），并用普通字母表示它们的第 i个坐标。我们将不加限制地使用广播操作。例如，若 $\bar{\mathbf{x}} = (\mathbf{x}1, \ldots, \mathbf{x}_n)$是 $\mathbb{R}^d$ 中的一列向量，且 g 是从 $\mathbb{R}^d$ 到某个集合 Y的函数，则 $g(\bar{\mathbf{x}})$表示序列（$g(\mathbf{x}_1) \ldots, g(\mathbf{x}_n)$）。类似地，对于矩阵$A \in M_q,\ _d$，我们记作 $A\bar{\mathbf{x}} = (A\mathbf{x}1, \ldots, A\mathbf{x}_n)$。

对于多项式 $p\colon \mathbb{R}^n \to \mathbb{R}$，我们用 $\|p\|_{\mathrm{co}}$ 表示 p的系数向量的欧几里得范数。我们将 $\|p\|_{\mathrm{co}}$ 称为 p的 系数范数 。对于 $\sigma\colon \mathbb{R} \to \mathbb{R}$，我们用 $\|\sigma\| = \sqrt{\mathbb{E}_{X\sim\mathcal{N}(0,1)}[\sigma^2(X)]}$ 表示相对于标准高斯测度的 ℓ^2 范数。我们用 $\|A\|_F = \sqrt{\sum_{i,j} A_{ij}^2}$表示矩阵 $A \in M_{n,m}$ 的弗罗贝尼乌斯范数，并用 $\|A\| = \max_{\|\mathbf{x}\|=1} \|A\mathbf{x}\|$表示谱范数。

我们用 $\mathbb{R}^{d,n}$ 表示 $\mathbb{R}^d$中由 n向量构成的序列空间。更一般地，对于集合 G，我们令$\mathbb{R}^{d,G} = \{\bar{\mathbf{x}} = (\mathbf{x}_g)_{g\in G}\colon \forall g \in G, \mathbf{x}_g \in \mathbb{R}^d\}$。我们用 $\mathbb{B}^d = \{\mathbf{x} \in \mathbb{R}^d\colon \|\mathbf{x}\| \le 1\}$表示欧几里得单位球。我们用 $\odot$ 表示向量和矩阵的逐点（Hadamard）乘法，并用 $(\mathbf{x}|\mathbf{y})$ 表示向量的拼接。对于 $\mathbf{x} \in \mathbb{R}^n$, $A \subseteq [n]$ 和 $\sigma \in \mathbb{Z}^n$，我们使用多重指标记号 $\mathbf{x}^A = \prod_{i\in A} x_i$ 和 $\mathbf{x}^\sigma = \prod_{i=1}^n x_i^{\sigma_i}$。对于 $\mathbf{f}\colon \mathcal{X} \to \mathbb{R}^n$ 和 $L \subseteq [n]$，我们用 $\mathbf{f}_L\colon \mathcal{X} \to \mathbb{R}^{|I|}$ 表示限制 $\mathbf{f}_L = (f_{i_1}, \ldots, f_{i_k})$，其中 $L = \{i_1, \ldots, i_k\}$ 满足 $i_1 < \ldots < i_k$。更一般地，对于 $\mathbf{f} = (\mathbf{f}_i)_{i\in[n]}\colon \mathcal{X} \to \mathbb{R}^{n,G}$，我们用 $\mathbf{f}_L\colon \mathcal{X} \to \mathbb{R}^{|I|,G}$ 表示限制 $\mathbf{f}_L = (\mathbf{f}_{i_1}, \ldots, \mathbf{f}_{i_k})$

2.1 多项式阈值函数

固定一个集合 $\mathcal{X} \subseteq [-1,1]^d$、一个函数 $f\colon \mathcal{X} \to \{\pm 1\}$、一个正整数 K 和 $M > 0$。如果存在一个次数为 $\le K$ 的多项式 $p\colon \mathbb{R}^d \to \mathbb{R}$，使得 $\|p\|_{\mathrm{co}} \le M$ 且对所有 $\forall \mathbf{x} \in \mathcal{X}$，$p(\mathbf{x})f(\mathbf{x}) \ge 1$，我们就称 f 为 (K, M)-多项式阈值函数（PTF）。更一般地，如果存在一个次数为 $\le K$ 的多项式 $p\colon \mathbb{R}^s \to \mathbb{R}$，使得 $\|p\|_{\mathrm{co}} \le M$ 且对所有 $\forall \mathbf{x} \in \mathcal{X}$，$p(\mathbf{h}(\mathbf{x}))f(\mathbf{x}) \ge 1$，我们就称 f 是 $\mathbf{h}\colon \mathcal{X} \to \mathbb{R}^s$ 的一个 (K, M)-PTF。我们将频繁使用的一个 $(K, 1)$-PTF 示例是一个函数 $f\colon \{\pm 1\}^d \to \{\pm 1\}$，它依赖于 K 个变量。事实上，关于 $\{\pm 1\}^d$ 的傅里叶分析告诉我们，f 是一个次数为 $\le K$ 的多项式 p 的限制，其中 $\|p\|_{\mathrm{co}} = 1$。对于这个多项式，我们有$\forall \mathbf{x} \in \mathcal{X}$，$p(\mathbf{x})f(\mathbf{x}) = 1$。

我们还需要对多项式阈值函数作更精细的定义，以便要求双侧不等式 $B \ge p(\mathbf{x})f(\mathbf{x}) \ge 1$，并要求其对 $\mathbf{x}$的扰动具有一定鲁棒性。为此，对于$\mathbf{x} \in [-1,1]^d$和 $r > 0$，我们定义

$$\mathcal{B}_r(\mathbf{x}) = \left\{\tilde{\mathbf{x}} \in [-1,1]^d : \|\mathbf{x} - \tilde{\mathbf{x}}\|_\infty \le r\right\} \tag{1}$$

固定 $B \geq 1$和 $1 \geq \xi > 0$。如果存在一个次数为 $\leq K$的多项式p: $\mathbb{R}^d \to \mathbb{R}$，使得 $\|p\|_{\text{co}} \leq M$并且，那么我们称 f为一个(K, M, B, ξ)-多项式阈值函数

$$\forall \mathbf{x} \in \mathcal{X}\ \forall \tilde{\mathbf{x}} \in \mathcal{B}_\xi(\mathbf{x}),\ \ B \geq p(\tilde{\mathbf{x}}) f(\mathbf{x}) \geq 1$$

同样地，我们称 f为一个 (K, M, B, ξ)-多项式阈值函数，作用于 $\mathbf{h} = (h_1, \ldots, h_s)$: $\mathcal{X} \to [-1, 1]$，如果存在一个次数为 $\leq K$的多项式 $p : \mathbb{R}^s \to \mathbb{R}$，使得 $\|p\|_{\text{co}} \leq M$ 且

$$\forall \mathbf{x} \in \mathcal{X}\ \forall \mathbf{y} \in \mathcal{B}_\xi(\mathbf{h}(\mathbf{x})),\ \ B \geq p(\mathbf{y}) f(\mathbf{x}) \geq 1$$

最后，我们称 f为一个 (K, M, B)-多项式阈值函数（相应地，为 $\mathbf{h}$ 的 (K, M, B)-多项式阈值函数），如果它是一个 $(K, M, B, 1)$-多项式阈值函数（相应地，是 $\mathbf{h}$ 的 $(K, M, B, 1)$-多项式阈值函数）。

2.2 强凸性

设 $W \subseteq \mathbb{R}^d$ 为凸函数。若一个可微函数 f: $W \to \mathbb{R}$ 满足：对于任意的 $\mathbf{x}$,$\mathbf{y} \in W$，都有

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \mathbf{y} - \mathbf{x}, \nabla f(\mathbf{x}) \rangle + \frac{\lambda}{2} \|\mathbf{y} - \mathbf{x}\|^2$$

注意，如果 f 是强凸的，那么当 $\mathbf{x} \in W$ 使 $\|\nabla f(\mathbf{x})\| \leq \epsilon$ 最小化时，其误差不超过加性误差 $\frac{\epsilon^2}{2\lambda}$。事实上，对于任意的 $\mathbf{y} \in W$，都有

$$\begin{aligned} f(\mathbf{x}) &\leq f(\mathbf{y}) - \frac{\lambda}{2}\|\mathbf{y} - \mathbf{x}\|^2 + \|\mathbf{y} - \mathbf{x}\| \cdot \|\nabla f(\mathbf{x})\| \\ &= f(\mathbf{y}) + \frac{\|\nabla f(\mathbf{x})\|^2}{2\lambda} - \frac{1}{2\lambda}\left(\|\nabla f(\mathbf{x})\| - \lambda\|\mathbf{y} - \mathbf{x}\|\right)^2 \\ &\leq f(\mathbf{y}) + \frac{\|\nabla f(\mathbf{x})\|^2}{2\lambda} \\ &\leq f(\mathbf{y}) + \frac{\epsilon^2}{2\lambda} \end{aligned} \tag{2}$$

2.3 厄米特多项式

接下来陈述的结果可见于 [5]。厄米特多项式 h_0，h_1，h_2，.....是与标准高斯测度相对应的正交归一多项式序列 μ，其定义在 $\mathbb{R}$上。也就是说，它们是通过对 1，x，x^2，x^3，$\ldots \in L^2(\mu)$进行 Gram–Schmidt过程得到的正交归一多项式序列。厄米特多项式满足以下递推关系

$$x h_n(x) = \sqrt{n+1} h_{n+1}(x) + \sqrt{n} h_{n-1}(x)\ \ , \quad h_0(x) = 1,\ h_1(x) = x \tag{3}$$

或者等价地

$$h_{n+1}(x) = \frac{x}{\sqrt{n+1}} h_n(x) - \sqrt{\frac{n}{n+1}} h_{n-1}(x)$$

厄米特多项式的生成函数为

$$e^{xt - \frac{t^2}{2}} = \sum_{n=0}^{\infty} \frac{h_n(x) t^n}{\sqrt{n!}} \tag{4}$$

我们还有

$$h_n' = \sqrt{n} h_{n-1} \tag{5}$$

同样，如果 X，$Y \sim \mathcal{N}\left(0, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right)$

$$\mathbb{E} h_i(X) h_j(Y) = \delta_{ij} \rho^i \tag{6}$$

3 层次模型

令 $\mathcal{X} \subseteq [-1,1]^d$ 为我们的实例空间。我们考虑多标签设置，其中每个实例可以有 0 到 n 个正标签，并且每个训练样本都附带一份包含其所有[1] 正标签的列表。因此，我们的目标是基于一个样本学习标注函数 $\mathbf{f}^*: \mathcal{X} \to \{\pm 1\}^n$

$$S = \{(\mathbf{x}^1, \mathbf{f}^*(\mathbf{x}^1), \ldots, (\mathbf{x}^m, \mathbf{f}^*(\mathbf{x}^m))\} \in (\mathcal{X} \times \{\pm 1\}^n)^m$$

来自分布 $\mathcal{D}$ 在 $\mathcal{X}$上的独立同分布带标签样本。具体而言，我们的目标是找到一个预测器 $\hat{\mathbf{f}}: \mathcal{X} \to \mathbb{R}^n$，使其误差 $\mathrm{Err}_{\mathcal{D}}(\hat{\mathbf{f}}) = \Pr_{\mathbf{x}\sim\mathcal{D}}\left(\mathrm{sign}(\hat{\mathbf{f}}(\mathbf{x})) \neq \mathbf{f}^*(\mathbf{x})\right)$ 较小。我们假设存在一个标签层次结构（算法未知），并约定

- 层次结构的第一层由一些简单的标签组成（即 = 易于学习的输入函数）。具体而言，每个此类标签都是输入的多项式阈值函数（PTF）。
- 层次结构第 i 层的任何标签，都是由层次结构较低层级的标签构成的简单函数（同样是多项式阈值函数）。

接下来，我们给出层次结构的形式化定义。

定义 3.1 (层次结构)。 设 $\mathcal{L} = \{L_1, \ldots, L_r\}$ 为一个集合族，满足 $L_1 \subseteq L_2 \subset \ldots \subseteq L_r = [n]$。我们称 $\mathcal{L}$ 是 $\mathbf{f}^*: \mathcal{X} \to \{\pm 1\}^n$ 的一个复杂度为 (r, K, M) 的层次结构（简称为 (r, K, M)-层次结构），如果对于任意 $j \in L_1$，函数 f_j^* 都是一个 (K, M)-多项式阈值函数；并且对于 $i \geq 2$，以及 $j \in L_i$，我们有 $\mathbf{f}_j^* = f_j \circ \mathbf{f}_{L_{i-1}}^*$，其中 (K, M)-多项式阈值函数 $\tilde{f}_j: \{\pm 1\}^{|L_{i-1}|} \to \{\pm 1\}$。

例 3.2。 固定 $\mathcal{L} = \{L_1, \ldots, L_r\}$ 如定义 *3.1*中所述，并回忆一下，依赖于 K 个坐标的布尔函数是一个 $(K,1)$-多项式阈值函数。因此，如果对于任意 $i \geq 2$，任意标签 $j \in L_i$ 至多依赖于 K 个来自 L_{i-1}的标签，并且任意标签 $j \in L_1$ 都是输入的一个 $(K,1)$-多项式阈值函数，那么 $\mathcal{L}$ 就是一个 $(r, K, 1)$-层级。

假设 K 为常数，我们的主要结果将表明：给定 poly$(n, d, M, 1/\epsilon)$ 个样本，一个在规模为 poly$(n, d, M, 1/\epsilon)$ 的残差网络上运行的多项式时间随机梯度下降算法，可以学习任意函数 $\mathbf{f}^*: \mathcal{X} \to \{\pm 1\}^n$，其误差为 ϵ，前提是 $\mathbf{f}^*$ 具有复杂度为 (r, K, M) 的层级（算法和网络不依赖于该层级，而只依赖于 r, K, M）。

该结果证明中的一个步骤是证明，任意定义在 K， M 的子集上的 $[-1,1]^n$-PTF 必然是一个 K， $2M$， B， ξ-PTF，其中 $\xi = \frac{1}{2^{n+1}}\binom{K+1}{2}^{K+1} KM$，且 $B = 2(\max(n, d) + 1)^{K/2} M$（见引理 8.2）。这足以建立上文非正式描述的主要结果。然而，在某些值得关注的情况下，我们可以有大得多的 ξ 和小得多的 B。在这种情况下，我们可以用更小的网络、更少的样本和更短的运行时间来保证可学习性。因此，接下来我们通过添加 B 和 ξ 作为参数来细化层级的定义。

定义 *3.3*（层级）。令 $\mathcal{L} = \{L_1, \ldots, L_r\}$ 为满足 $L_1 \subseteq L_2 \subset \ldots \subseteq L_r = [n]$的集合族。若对于任意 $j \in L_1$，函数 f_j^*是一个 (K, M, B)-多项式阈值函数，并且对于 $i \geq 2$和 $j \in L_i$，我们有 $f_j^* = \tilde{f}_j \circ \mathbf{f}_{L_{i-1}}^*$对应于一个 (K, M, B, ξ)-多项式阈值函数 $\tilde{f}_j: \{\pm 1\}^{|L_{i-1}|} \to \{\pm 1\}$，则称 $\mathcal{L}$ 是 $\mathbf{f}^*$ 的一个复杂度为 (r, K, M, B, ξ) 的层级（或简写为 (r, K, M, B, ξ)-层级）。

3.1 “BrainDump”层级

固定一个定义域 $\mathcal{X} \subseteq \{\pm 1\}^d$ 和一个函数序列 $G^i: \{\pm 1\}^d \to \{\pm 1\}^d$，其中 $1 \leq i \leq r$。我们假设$G^0(\mathbf{x}) = \mathbf{x}$，并且对于任意深度 $i \in [r]$ 和坐标 $j \in [d]$，有

$$\forall \mathbf{x} \in \mathcal{X}, \quad G_j^i(\mathbf{x}) = h_j^i(G^{i-1}(\mathbf{x})),$$

其中 $h_j^i: \{\pm 1\}^d \to \{\pm 1\}$ 是一个依赖于 K 个坐标的函数。我们将序列 G^1、.....、G^r视为一个计算电路，或一个“大脑”模型。

[1]我们注意到，在实践中，一个样本通常可能具有多个正标签（例如，“狗”和“动物”）。然而，每个训练样本通常只附带其中一个正标签。我们希望未来的工作能够处理这种更符合现实的监督类型。

假设我们希望在能够访问带标签样本（$\mathbf{x}$ $f^*(\mathbf{x})$）的情况下，学习形式为 $f^* = h \circ G^r$ 的函数，其中 h：$\{\pm 1\}^d \to \{\pm 1\}$ 也只依赖于 K 个输入。函数 f^* 可能极其复杂。例如，G 可能计算一个密码学函数。在这种情况下，仅凭带标签样本（$\mathbf{x}$ $f^*(\mathbf{x})$）来学习 f^* 很可能难以实现；如果我们对 f^* 的访问仅限于上述黑盒情形，那么这项任务似乎是不可能的。另一方面，如果我们能够对白盒访问 f^*——也就是说，获得电路 G的完整描述——学习问题就会变得轻而易举。然而，如果 G 确实模拟了人脑，那么这种透明访问是不现实的。

考虑介于黑盒和白盒情形之间的一种折中方案。假设我们可以向标注者（其大脑由 G 建模）询问更多信息。例如，如果 f^* 是一个能够识别图像中汽车的函数，我们可以询问标注者，不仅要判断图像中是否有汽车，还要指出具体特征：车轮、车窗、暗色区域、曲线，以及任何他认为相关的内容。每个此类附加标签都代表在电路 G 上计算的另一个简单函数。我们将这些辅助标签建模为随机选取的 G_j^i 的随机多数函数。我们将证明，只要有足够多的此类标签，所得问题就具有低复杂度层级，因此可以高效学习。

形式化地，固定一个整数 q。我们假设，对于每个深度 $i \in [r]$，都有 q 个辅助标签 $f_{i,\ j}^*$，其中 $1 \le j \le q$，每个标签都是由 G^i 的奇数个分量构成的带符号多数函数。此外，我们假设这些函数是随机的。具体而言，在学习之前，标注者会独立采样 qr 个函数，使得对于任意 $i \in [r]$ 和 $j \in [q]$，

$$f_{i,j}^*(\mathbf{x}) = \operatorname{sign}\left(\sum_{l=1}^{d} w_l^{i,j} G_l^i(\mathbf{x})\right),$$

其中，权重向量$\mathbf{w}^{i,j} \in \mathbb{R}^d$ 是从以下分布中独立选取的均匀向量

$$\mathcal{W}_{d,k} := \left\{\mathbf{w} \in \{-1,0,1\}^d : \sum_{l=1}^{d} |w_l| = k\right\}$$

对于某个奇数 k。

定理 **3.4.** 若 $q = \tilde{\omega}\left(k^2 d \log(|\mathcal{X}|)\right)$，则$\mathbf{f}^*$ 以概率 $1 - o(1$具有 $\left(r, K, O\left(kd^K\right), 2k+1\right)$-层级 $)$

3.2 对序列模型和集成模型的扩展

接下来，我们将层级的概念扩展到一种常见情形：学习得到的函数的输入和输出都是向量集。设 G 为某个集合。我们将 G 中的元素称为位置。在图像情境下，一个自然的选择是 $G = [T_1] \times [T_2]$，其中 $T_1 \times T_2$ 是输入图像的最大尺寸。在语言情境下，一个自然的选择是 $G = [T]$，其中 T 是输入中的词元最大数量。我们用 $\vec{\mathbf{x}} = (\mathbf{x}_g)_{g \in G}$ 表示向量集，并令 $\mathbb{R}^{d,G} = \{\vec{\mathbf{x}} = (\mathbf{x}_g)_{g \in G} : \forall g \in G,\ \mathbf{x}_g \in \mathbb{R}^d\}$。

固定 $\mathcal{X} \subseteq [-1,1]^d$，并令 $\mathcal{X}^G$ 为我们的实例空间。假设共有 n 个标签。我们考虑这样一种设定：每个位置上的每个实例都可以具有 0 到 n 个正标签。鉴于此，我们的目标是基于一个样本学习标注函数 $\mathbf{f}^*$：$\mathcal{X}^G \to \{\pm 1\}^{n,G}$

$$S = \{(\vec{\mathbf{x}}^1, \mathbf{f}^*(\vec{\mathbf{x}}^1)), \ldots, (\vec{\mathbf{x}}^m, \mathbf{f}^*(\vec{\mathbf{x}}^m))\} \in \left(\mathcal{X}^G \times \{\pm 1\}^{n,G}\right)^m$$

该样本由来自 $\mathcal{X}^G$ 上某个分布 $\mathcal{D}$ 的独立同分布带标签样本构成。我们假设存在一个标签层次结构（算法未知），并约定

- 层次结构的第一层由简单（= 易于学习）的输入函数构成。具体而言，位置 g 上的每个此类标签都是输入在附近 g的一个多项式阈值函数。
- 层次结构第 i 层中的任何标签，都是较低层级标签的简单函数。具体而言，位置 g 上的每个此类标签，都是位置 g 附近的较低层级标签的多项式阈值函数。

我们将通过邻近映射来刻画 G中位置的 G邻近性，该映射为任意元素 $g \in G$指定 w 附近的位置。我们始终将 g 自身视为 g附近的一个点。这可由以下定义表示

定义 3.5（邻近映射）。一个宽度为 w 的邻近映射是一个映射 $\mathbf{e} = (e_1, \ldots, e_w) : G \to G^w$ 满足 $e_1(g) = g$ 对任意 g。

例如，如果 $G = [T]$，自然可以选择 $\mathbf{e}$：$G \to G^{2w+1}$，使得 $\{e_1(g) \ldots, e_{2w+1}(g)\} = \{g' \in T : |g' - g| \le w\}$。同样，如果 $G = [T] \times [T]$，自然可以选择 $\mathbf{e}$：$G \to G^{(2w+1)^2}$，使得 $\{e_1(g_1, g_2) \ldots, e_{(2w+1)^2}(g_1, g_2)\} = \{(g_1', g_2') \in T \times T : |g_1' - g_1| \le w \text{ 和 } |g_2' - g_2| \le w\}$。给定一个邻近映射 $\mathbf{e}$ 和 $\vec{\mathbf{x}} \in \mathbb{R}^d$，^{G}，我们将 $E_g(\vec{\mathbf{x}})$ 定义为所有满足以下条件的向量 $\mathbf{x}_{g'}$ 的拼接：根据 $\mathbf{e}$，g' 与 g 接近。形式化地，

定义 *3.6*。给定一个邻近映射 $\boldsymbol{e}$：$G \to G^w$，$g \in G$ 和 $\vec{x} \in \mathbb{R}^{d,G}$，我们定义 $E_g(\vec{x}) = (x_{e_1(g)} | \cdots | x_{e_w(g)}) \in \mathbb{R}^{dw}$。同样，我们令 $E(\vec{x}) \in \mathbb{R}^{dw,G}$ 为 $E(\vec{x}) = (E_g(\vec{x}))_{g \in G}$。

接下来，我们扩展多项式阈值函数的定义，使其适用于集成设定。

定义 3.7 (层级)。设 $\mathcal{L} = \{L_1, \ldots, L_r\}$ 是一个集合族，且满足 $L_1 \subseteq L_2 \subset \ldots \subseteq L_r = [n]$。设 $\mathbf{e} : G \to G^w$ 是一个邻近函数。若满足以下条件，我们称 $(\mathcal{L}, \mathbf{e})$ 是 $\mathbf{f}^* : \mathcal{X}^G \to \{\pm 1\}^{n,G}$ 的一个复杂度为 (r, K, M, B, ξ) 的层级（简称为 (r, K, M, B, ξ)-层级）：

- 对于任意 $j \in L_1$，存在一个 (K, M, B, ξ)-多项式阈值函数 $f_j : \tilde{\mathcal{X}}^w \to \{\pm 1\}$，使得 $f_{j,g}(\mathbf{x}) = f(E_g(\mathbf{x}))$ 对任意 $\mathbf{x} \in \tilde{\mathcal{X}}^G$ 和 $g \in G$ 均成立

- 对于 $i \ge 2$，且对于 $j \in L_i$，存在一个 (K, M, B, ξ)-多项式阈值函数 $\tilde{f}_j : \{\pm 1\}^{|L_1|w} \to \{\pm 1\}$，使得 $f_{j,g}(\mathbf{x}) = \tilde{f}(E_g(\mathbf{f}^*_{L_{i-1}}(\mathbf{x})))$ 对任意 $\mathbf{x} \in \mathcal{X}^G$ 和 $g \in G$ 均成立

注意，先前对层级的定义（即定义3.1和3.3）是特殊情况 $w = |G| = 1$。

4 算法与主要结果

固定 $\mathcal{X} \subseteq [-1, 1]^d$、位置集 G、宽度为 w 的邻近映射 $e : G \times N \to G$、某个常数整数 $K \ge 1$，以及一个满足利普希茨条件、有界且非恒定的激活函数 $\sigma : \mathbb{R} \to \mathbb{R}$。我们将 σ 和 K 视为固定量，并允许使用大-O 记号来隐藏依赖于 σ 和 K 的常数。

我们首先描述要考虑的残差网络架构。令 $\mathcal{X}^G$ 为我们的实例空间。网络的第一层（实际上是两层，但将其视为一层会更方便）将计算函数

$$\Psi_1(\vec{\mathbf{x}}) = W_2^1 \sigma(W_1^1 E(\vec{\mathbf{x}}) + \mathbf{b}^1)$$

我们假设 $W_2^1 \in \mathbb{R}^{n \times q}$ 初始化为 0，而 $(W_1^1, \mathbf{b}^1) \in \mathbb{R}^{q \times wd} \times \mathbb{R}^q$ 则使用下文定义的 β-*Xavier* 初始化进行初始化。

定义 4.1 (Xavier 初始化)。固定 $1 \ge \beta \ge 0$。若随机对 $(W, \mathbf{b}) \in \mathbb{R}^{q \times d} \times \mathbb{R}^q$ 具有 β-Xavier 分布，则 W 的各项服从方差为 $\frac{1-\beta^2}{d}$ 的独立同分布中心高斯分布，并且 $\mathbf{b}$ 与 W 独立，且其各项服从方差为 β^2 的独立同分布中心高斯分布。

其余层的形式如下

$$\Psi_k(\vec{\mathbf{x}}) = \vec{\mathbf{x}} + W_2^k \sigma(W_1^k E(\vec{\mathbf{x}}) + \mathbf{b}^k)$$

其中 $(W_1^k, \mathbf{b}^k) \in \mathbb{R}^{q \times (wn)} \times \mathbb{R}^q$ 使用 β-Xavier 初始化进行初始化，而 $W_2^k \in \mathbb{R}^{n \times q}$ 初始化为 0。最后一层计算

$$\Psi_D(\vec{\mathbf{x}}) = W^D \vec{\mathbf{x}}$$

对于正交矩阵 $W^D \in \mathbb{R}^{n \times n}$。我们将权重矩阵的集合记为 $\vec{W}$，将网络计算出的ˆ函数记为 $\mathbf{f}_{\vec{W}}$。固定一个凸损失函数 $\ell : \mathbb{R} \to [0, \infty)$，我们通过取平均将其扩展为损失 $\ell : \mathbb{R}^G \times \{\pm 1\}^G \to [0, \infty)$：

$$\ell(\hat{\mathbf{y}}, \mathbf{y}) = \frac{1}{|G|} \sum_{g \in G} \ell(\hat{y}_g \cdot y_g)$$

同样，对于函数 $\hat{\mathbf{f}}: \mathcal{X}^G \to \mathbb{R}^{n,\ G}$ 和 $j \in [n]$，我们定义

$$\ell_{S,j}(\hat{\mathbf{f}}) = \ell_{S,j}(\hat{\mathbf{f}}_j) = \frac{1}{m}\sum_{t=1}^{m} \ell\left(\hat{\mathbf{f}}_j(\vec{\mathbf{x}}^t), \mathbf{y}_j^t\right)$$

最后，令

$$\ell_S(\hat{\mathbf{f}}) = \sum_{j=1}^{n} \ell_{S,j}(\hat{\mathbf{f}}) \qquad \text{and} \qquad \ell_S(\vec{W}) = \ell_S\left(\hat{\mathbf{f}}_{\vec{W}}\right)$$

我们将考虑以下算法

算法 4.2. 在每一步 $k = 1$、...、$D-1$，对 $\ell_S(\vec{W}) + \frac{\epsilon_{\text{opt}}}{2}\|W_2^k\|^2$ 关于 W_2^k 进行优化，直到梯度大小达到 $\leq \epsilon_{\text{opt}}$。（由于第 k 步的目标函数是 ϵ_{opt}-强凸的，该算法能找到它的一个 $\frac{\epsilon_{\text{opt}}}{2}$-极小化器。）

我们将考虑以下损失函数。

$$\ell = \ell_{1/(2B)} + \frac{1}{4m|G|}\ell_{1-\xi/2} \qquad \text{for} \qquad \ell_\eta(z) = \begin{cases} 1 - \frac{z}{\eta} & 0 \leq z \leq \eta \\ 0 & \eta \leq z \leq 1 \\ \infty & \text{otherwise} \end{cases} \tag{7}$$

现在我们可以陈述主要结果了。

定理 4.3 (Mai n). 假设 $\mathbf{f}^*$ 具有 (r, K, M, B, ξ)-层级，并令 $\gamma = \frac{1}{32}$ 最小 $\left(\frac{1}{B}, \xi\right)$。假设 th t

- $D > r \cdot \left(\left\lceil \frac{\ln(8m|G|/\xi)}{\gamma} \right\rceil + 1\right)$
- $\epsilon_{\text{opt}} \leq \frac{(1-e^{-\gamma})\xi}{16m^2|G|^2}$

那么，可以选择 β 和 $q = \tilde{O}\left(^{(M+1)^4}(\ _{\gamma^{4+2K}}^{wn\ ^{2K}}\ )\right)$，使得算法 4.2 能够学习到一个期望误差至多为 $\tilde{O}\left(^{D^2}(\ _{\gamma^{4+2K}m}^{M+1)\ \ ^4\ wn\ ^{2K+1}}\ )\right)$() 的分类器。

5 P定理 4.3：残差网络的层次学习证明 s

为了证明定理 4.3，只需证明下面的定理 5.1，该定理表明可以选择 β 和 $q = \tilde{O}\left(^{(M+1)^4}(\ _{\gamma^{4+2K}}^{wn\ ^{2K}}\ )\right)$，使得算法 4.2 能够以概率 $\frac{1}{m}$ 学习到经验大间隔误差为 0 的分类器。也就是说，我们定义

$$\mathrm{Err}_{S,\gamma}(\hat{\mathbf{f}}) = \frac{1}{m}\sum_{t=1}^{m} 1\left[\exists (i,g) \in [n] \times G \text{ s.t. } \hat{f}_{i,g}(\vec{\mathbf{x}}^t) \cdot f^*_{i,g}(\vec{\mathbf{x}}^t) < \gamma\right] \tag{8}$$

并证明算法 4.2 将以概率 $\frac{1}{m}$ 学得一个分类器 $\hat{\mathbf{f}}$，使得 $\mathrm{Err}_{S,1/2}(\hat{\mathbf{f}}) = 0$。我们称这样的算法为 $(1/m)$-一致算法。根据这一保证，定理 4.3 可由标准的参数计数论证推出：训练参数的数量为 $p = Dqn$，且由于 ℓ^2 正则化项，其大小有 $\frac{2n}{\epsilon_{\text{opt}}} + 1$ 的上界。同样，排除某个初始权重的大小为 $\geq \ln(Dq(n+d)wm)$ 这一小概率事件（由于所有 $Dq(n+d)w$ 初始权重均为方差为 ≤ 1 的中心高斯分布，该事件发生的概率为 $\ll \frac{1}{m}$）后，不难验证，作为 $2D$ 层的复合，网络输出相对于训练参数是 L-Lipchitz 的，其中 $L = 2^{O(D)}$。因此，任意 $(1/m)$-一致算法的期望误差为 $\tilde{O}\left(^{p}\log(\ _{m}^{L}\ )\right) = \tilde{O}\left(\frac{Dp}{m}\right) = \tilde{O}\left(\frac{D^2qn}{m}\right)$。（精确表述见引理 7.7。）

定理 5.1（主要结果（重述））。令 $\gamma = \frac{1}{32}\min\left(\frac{1}{B}, \xi\right)$。假设

- $\mathbf{f}^*$ *has* (r, K, M, B, ξ)*-hierarchy* $(\mathcal{L}, e)$
- $D > r \cdot \left(\left\lceil \frac{\ln(8m|G|/\xi)}{\gamma} \right\rceil + 1 \right)$
- $\epsilon_{\text{opt}} \leq \frac{(1-e^{-\gamma})\xi}{16m^2|G|^2}$

存在一种 β 选择，使得以概率 $1 - 2nmD|G|\exp\left(-\Omega\left(q \cdot \frac{\gamma^{2K+4}}{wn^{2K}()(M+1)^4}\right)\right)$，对于权重的初始选择，算法 4.2 将学习得到一个分类器 $\hat{\mathbf{f}}: \mathcal{X}^G \to \mathbb{R}^{n,G}$ 其 $\text{Err}_{S,1/2}(\hat{\mathbf{f}}) = 0$。

对于 $1 \leq k \leq D$，设 $\hat{\mathbf{f}}^k: \mathcal{X}^G \to \mathbb{R}^{n,G}$ 为网络在第 k 层训练完成后计算的函数。此外，设 $\Gamma^k: \mathcal{X}^G \to \mathbb{R}^{n,G}$ 为第 k 层训练完成后，由第 1 层至第 k 层计算的函数。对于 $k = 0$，我们用 $\hat{\mathbf{f}}^0 = \Gamma^0$ 表示从 $\mathcal{X}^G$ 到 $\mathbb{R}^{d,G}$ 的恒等映射。注意，当算法 4.2 训练第 k 层时，对于任意 $k' > k$，都有 $W_2^{k'} = 0$。因此，

$$\Psi_{k'}(\vec{\mathbf{x}}) = \vec{\mathbf{x}} + W_2^{k'}\sigma(W_1^{k'}E(\vec{\mathbf{x}}) + \mathbf{b}^{k'}) = \vec{\mathbf{x}}$$

因此，当训练第 k 层时，对于任意 $k' > k$，第 k' 层只是恒等函数。由此可得 $\hat{\mathbf{f}}^k(\mathbf{x}) = W^D\Gamma^k(\mathbf{x})$。

在定理 5.1 的证明中，我们首先注意到，算法 4.2 的第 k 步（即从 $\hat{\mathbf{f}}^{k-1}$ 得到 $\hat{\mathbf{f}}^k$）本质上等同于在该数据表示 $\vec{\mathbf{x}} \mapsto \hat{\mathbf{f}}^{k-1}(\vec{\mathbf{x}})$ 的随机特征扩展上学习一个线性分类器。具体而言，定义一个输入空间嵌入 $\Phi^{k-1}: \mathcal{X}^G \to \mathbb{R}^{q,G}$，使得

$$\Phi^{k-1}(\vec{\mathbf{x}}) = \sigma(W_1^k E(\Gamma^{k-1}(\vec{\mathbf{x}})) + \mathbf{b}^k) = \sigma(W_1^k E((W^D)^{-1}\hat{\mathbf{f}}^k(\mathbf{x})) + \mathbf{b}^k) =$$

对于 $\mathbf{w} \in \mathbb{R}^q$，我们定义

$$\hat{\mathbf{f}}_{j,\mathbf{w}}^k(\vec{\mathbf{x}}) = \hat{\mathbf{f}}_j^{k-1}(\vec{\mathbf{x}}) + \mathbf{w}^\top\Phi^{k-1}(\vec{\mathbf{x}})$$

我们有

引理 5.2。 对于任意 $D - 1 \geq k \geq 1$ $\hat{\mathbf{f}}_j^k = \hat{\mathbf{f}}_{j,\mathbf{w}}^k$ 其中 $\mathbf{w}$ 是凸目标函数的 $\frac{\epsilon_{\text{opt}}}{2}$-极小化器

$$\ell_{S,j}^k(\mathbf{w}) \quad = \quad \ell_{S,j}\left(\hat{\mathbf{f}}_{j,\mathbf{w}}^k\right) + \frac{\epsilon_{\text{opt}}}{2}\|\mathbf{w}\|^2$$

在 $\mathbf{w} \in \mathbb{R}^q$ 上。此外，

$$\ell_{S,j}(\hat{\mathbf{f}}^k) \leq \ell_{S,j}^k(\mathbf{w}^*) + \frac{\epsilon_{\text{opt}}}{2}\|\mathbf{w}^*\|^2 + \frac{\epsilon_{\text{opt}}}{2}$$

证明。当训练第 k 层时，由于在此训练阶段所有更深的层都是恒等函数，网络的输出作为 W_2^k（在第 k 步训练的参数）的函数为

$$G(W_2^k, \vec{\mathbf{x}}) = W^D\left(\Gamma_{k-1}(\vec{\mathbf{x}}) + W_2^k\Phi^{k-1}(\vec{\mathbf{x}})\right) = \hat{\mathbf{f}}^{k-1}(\vec{\mathbf{x}}) + W^D W_2^k\Phi^{k-1}(\vec{\mathbf{x}})$$

特别地，如果我们将 $\hat{}$ 表示为 W_2^k 在 W_2^k 经过第 k 层训练后的值，那么有 $\hat{\mathbf{f}}_j^k = \hat{\mathbf{f}}_{j,\mathbf{w}}^k$ 其中 $\mathbf{w}$ j 是矩阵 $W = W^D$ 的第 $W_2^k\hat{}W_2^k$ 行。因此，仍需证明 $\mathbf{w}$ 使 ℓ_S^k 最小化，j。为此，我们注意到，在第 k 步，算法 4.2 找到一个 $\frac{\epsilon_{\text{opt}}}{2}$-极小化器

$$L(W_2^k) = \frac{\epsilon_{\text{opt}}}{2}\|W_2^k\|^2 + \frac{1}{m}\sum_{t=1}^{m}\sum_{j=1}^{n}\ell(\hat{\mathbf{f}}^{k-1}(\vec{\mathbf{x}}) + W^D W_2^k\Phi^{k-1}(\vec{\mathbf{x}}), \mathbf{y}_j^t)$$

因此，$\hat{W} := W^D \hat{W}_2^k$是一个$\frac{\epsilon_{\text{opt}}}{2}$-极小化器

$$
\begin{aligned}
L'(W) = L((W^D)^{-1}W) &= \frac{\epsilon_{\text{opt}}}{2}\|(W^D)^{-1}W\|^2 + \frac{1}{m}\sum_{t=1}^{m}\sum_{j=1}^{n}\ell(\hat{\mathbf{f}}_j^{k-1}(\vec{\mathbf{x}}) + W^d(W^D)^{-1}W\Phi^{k-1}(\vec{\mathbf{x}}), \mathbf{y}_j^t) \\
&\overset{W^D \text{ is orthogonal}}{=} \frac{\epsilon_{\text{opt}}}{2}\|W\|^2 + \frac{1}{m}\sum_{t=1}^{m}\sum_{j=1}^{n}\ell(\hat{\mathbf{f}}_j^{k-1}(\vec{\mathbf{x}}) + W\Phi^{k-1}(\vec{\mathbf{x}}), \mathbf{y}_j^t) \\
&= \sum_{j=1}^{n}\left(\frac{\epsilon_{\text{opt}}}{2}\|W_{j\cdot}\|^2 + \frac{1}{m}\sum_{t=1}^{m}\ell(\hat{\mathbf{f}}_j^{k-1}(\vec{\mathbf{x}}) + W_{j\cdot}\Phi^{k-1}(\vec{\mathbf{x}}), \mathbf{y}_j^t)\right) \\
&= \sum_{j=1}^{n}\ell_{S,j}^k(W_{j\cdot})
\end{aligned}
$$

特别地，$\mathbf{w} = \hat{W}_{j\cdot}$必须是 $\frac{\epsilon_{\text{opt}}}{2}$-极小化器，满足$\ell_{S,\,j}^k$最后，由于$\ell_{S,\,j}^k$是 ϵ_{opt}-强凸的，方程（2）意味着对于任意 $\mathbf{w}^* \in \mathbb{R}^q$,

$$
\ell_{S,j}(\hat{\mathbf{f}}^k) \le \ell_{S,j}^k(\mathbf{w}^*) + \frac{\epsilon_{\text{opt}}}{2}\|\mathbf{w}^*\|^2 + \frac{\epsilon_{\text{opt}}}{2}
$$

□

有了引理 5.2 ，我们可以介绍证明策略。由于 L_1 中的标签是输入的多项式阈值函数（PTF），因此当第一层训练完成时，我们就能学会这些标签。也就是说，$\hat{\mathbf{f}}^1$ 将正确预测 L_1 中的标签。原因是，粗略地说，通过在随机特征嵌入之上训练线性分类器，可以高效地学习多项式阈值函数。

由于，$\hat{\mathbf{f}}^1$能正确预测 L_1中的标签， L_2中的标签就成为 $\hat{\mathbf{f}}^1$的简单函数。具体而言，它们是 $\text{sign}(\hat{\mathbf{f}}^1)$ 的多项式阈值函数（PTF）。因此，人们很容易想到沿用上述推理来证明：训练下一层后，我们将学会 L_2 中的标签；更一般地，训练完 r 层后，网络将能正确预测所有标签。然而，这种方法并不那么顺利：$\text{sign}(\hat{\mathbf{f}}^1)$ 的多项式阈值函数（PTF）未必能通过在 $\hat{\mathbf{f}}^1$之上训练线性分类器并使用随机特征嵌入来学习。为了解决这一问题，我们证明，在网络正确预测某个标签 j后，继续训练更多层会使该标签的损失持续降低，因此，再训练额外的 $O(B + 1/\xi)$ 层后，损失将足够小，从而保证 L_2 中的标签是 $\hat{\mathbf{f}}^1$ 的多项式阈值函数（而不只是 $\text{sign}(\hat{\mathbf{f}}^1)$ 的多项式阈值函数）。因此，训练完 $O(B + 1/\xi)$ 层后，网络将能正确预测 L_2 中的标签；更一般地，训练完 $O(rB + r/\xi)$ 层后，网络将能正确预测所有标签。

证明的过程如下

1. 我们先介绍引理 5.4，它表明，如果标签 j 是 $\hat{\mathbf{f}}^k$ 的一个较大多项式阈值函数（PTF），那么 $\hat{\mathbf{f}}^{k+1}$ 就能正确预测它。更准确地说，我们证明：如果对于某个多项式 p，$\ell_{S,j}(p \circ E \circ \hat{\mathbf{f}}^k)$ 的鲁棒版本较小，那么 $\ell_{S,j}(\hat{\mathbf{f}}^{k+1})$ 也较小。2. 接着，我们介绍引理 5.5，该引理利用引理 5.4 证明：(i) 对任意$j \in L_1$，$\ell_{S,j}(\hat{\mathbf{f}}^1)$ 都较小；(ii) 对任意 $j \in [n]$，如果 $\ell_{S,j}(\hat{\mathbf{f}}^k)$ 较小，那么随着我们训练更深的层，它会呈指数级缩小；(iii) 如果对任意 $j \in L_{i-1}$，$\ell_{S,j}(\hat{\mathbf{f}}^k)$ 都非常小，那么对任意 $j \in L_i$，$\ell_{S,j}(\hat{\mathbf{f}}^{k+1})$ 也较小。3. 基于引理 5.5，我们将证明定理 5.1。

为了完成第一步，我们需要一些记号。首先，我们将 ϵ 鲁棒版本 ℓ 定义为

$$
\ell^{\text{rob},\epsilon}(z) = \max(\ell(z), \ell(z-\epsilon)) = \max_{0 \le t \le \epsilon}\ell(z-t) \tag{9}
$$

注意，对于 $z \le 1$，我们有 $\ell^{\text{rob},\,\epsilon}(z) = \ell(z-\epsilon)$；而对于 $z < 0$，我们有 $\ell^{\text{rob},\,\epsilon}(z) = \ell(z) = \infty$。将 σ 的厄米特展开记为

$$
\sigma = \sum_{s=0}^{\infty} a_s h_s \tag{10}
$$

令 K' 为满足以下条件的最小整数 $K' \geq K$：$a_{K'} \neq 0$（这样的 K' 必然存在，否则 σ 是一个多项式，这与其有界且非常数的假设相矛盾）。对于 $\epsilon > 0$，定义 $\beta(\epsilon) = \beta_{\sigma, K', K}(\epsilon) < 1$ 为大于 [3 4] 的最小正数，使得若 $\beta_{\sigma, K', K}(\epsilon) \leq \beta < 1$，则

$$\frac{\|\sigma\|}{a_{K'}} 2^{(K'+2)/2} \frac{1-\beta^2}{\sqrt{1-2(1-\beta^2)^2}} \leq \frac{\epsilon}{2}$$

注意，$\beta(\epsilon)$ 定义良好，因为 $h(\beta) := \frac{1-\beta^2}{\sqrt{1-2(1-\beta^2)^2}}$ 在 $\beta = 1$ 附近连续，且在 $\beta = 1$ 处等于 0。事实上，由于 h 在 $\beta = 1$ 附近可微，我们有 $1 - \beta(\epsilon) = \Omega\left(\epsilon 2^{-K'} \frac{a_{K'}}{\|\sigma\|}\right)$。特别地，对于固定的 σ，K'，K，我们有 $1 - \beta(\epsilon) = \Omega(\epsilon)$。另定义

$$\delta(\epsilon, \beta, q, M, n) = \delta_{\sigma,K',K}(\epsilon, \beta, q, M, n) = \begin{cases} 1 & \frac{4\|\sigma\|_\infty}{\epsilon\sqrt{q}} \cdot \frac{1}{a_{K'}^2 \beta^{2K'-2K}} \left(\frac{n}{1-\beta^2}\right)^K M^2 > 1 \\ 2\exp\left(-q \cdot \frac{a_{K'}^4 \beta^{4K'-4K}(1-\beta^2)^{2K}\epsilon^4}{512 n^{2K} M^4 \|\sigma\|_\infty^4}\right) & \text{otherwise} \end{cases}$$

注意，对于固定的 σ，K'，K 以及 $1 - \beta = \Omega(\epsilon)$，我们有

$$\delta(\epsilon, \beta, q, M, n) = \exp\left(-\Omega\left(q \cdot \frac{\epsilon^{2K+4}}{n^{2K} M^4}\right)\right) \tag{11}$$

我们需要以下引理，该引理在第 9 节末尾证明，并表明可以通过复合一个随机层和一个线性函数来近似一个多项式。

引理 5.3。固定 $\mathcal{X} \subset [-1,1]^n$、次数为 K 的多项式 $p: \mathcal{X} \to [-1,1]$、$K' \geq K$ 和 $\epsilon > 0$。令 $(W、\mathbf{b}) \in \mathbb{R}^{q \times n} \times \mathbb{R}^q$ 为关于 $1 > \beta \geq \beta_{\sigma, K', K}(\epsilon)$ 的 β-*Xavier* 对。那么存在一个向量 $\mathbf{w} = \mathbf{w}(W、\mathbf{b}) \in \mathbb{B}^q$，使得

$$\forall \mathbf{x} \in \mathcal{X},\ \Pr\left(|\langle \mathbf{w}, \sigma(W\mathbf{x} + \mathbf{b})\rangle - p(\mathbf{x})| \geq \epsilon\right) \leq \delta_{\sigma,K',K}(\epsilon, \beta, q, \|p\|_{\text{co}}, n)$$

现在我们可以证明：若存在一个多项式 $p: \mathbb{R}^{wn} \to \mathbb{R}$ 使得 $\ell^{\text{rob}}_{S, \epsilon_1, j}(p \circ E_g \circ \hat{\mathbf{f}}^k)$ 足够小，那么以高概率，$\ell_{S,j}(\hat{\mathbf{f}}^{k+1})$ 也会足够小。

引理 5.4。固定 $\epsilon_1 > 0$, $1 > \beta > \beta(\epsilon_1/2)$ 以及一个多项式 $p: \mathbb{R}^{wn} \to \mathbb{R}$。已知 $\ell^{\text{rob}}_{S, \epsilon_1, j}(p \circ E_g \circ \hat{\mathbf{f}}^k) \leq \epsilon$，则以概率 $1 - m|G|\delta(\epsilon_1/2, \beta, q, \|p\|_{\text{co}} + 1, wn)$，有 $\ell_{S,j}(\hat{\mathbf{f}}^{k+1}) \leq \epsilon + \epsilon_{\text{opt}}$。

证明。由引理 5.2 可得 $\ell_{S,j}(\hat{\mathbf{f}}^{k+1}) \leq \ell_{S,j}\left(\hat{\mathbf{f}}_j^{k+1}, \mathbf{w}^*\right) + \frac{\epsilon_{\text{opt}}}{2}\|\mathbf{w}^*\|^2 + \frac{\epsilon_{\text{opt}}}{2}$ 对任意 $\mathbf{w}^* \in \mathbb{R}^q$ 均成立。因此，只需证明：以概率 $1 - m|G|\delta(\epsilon_1/2, \beta, q, \|p\|_{\text{co}} + 1, wn) =: 1 - \delta$，在 W_1^k 的选取中存在 $\mathbf{w}^* \in \mathbb{B}^d$ 使得 $\ell_{S,j}\left(\hat{\mathbf{f}}_j^{k+1}, \mathbf{w}^*\right) \leq \epsilon$。根据 $\ell^{\text{rob}}_{S, \epsilon_1, j}$ 的定义，只需证明以概率 $1 - \delta$ 存在 $\mathbf{w}^* \in \mathbb{B}^d$ 使得

$$y_{j,g}^t \cdot p \circ E_g \circ \hat{\mathbf{f}}^k(\vec{\mathbf{x}}^t) - \epsilon_1 \leq y_{j,g}^t \cdot \hat{f}_{j,g,\mathbf{w}^*}^{k+1}(\vec{\mathbf{x}}^t) \leq y_{j,g}^t \cdot p \circ E_g \circ \hat{\mathbf{f}}^k(\vec{\mathbf{x}}^t) \tag{12}$$

对任意 t 和 g。由于 $y_{j,g}^t \cdot p \circ E_g \circ \hat{\mathbf{f}}^k(\tilde{\mathbf{x}}^t) \geq \epsilon_1$（否则我们会有 $\ell^{\text{rob}}_{S, \epsilon_1, j}(p \circ E_g \circ \hat{\mathbf{f}}^k) = \infty$），因此只需证明，以概率 $1 - \delta$ 存在 $\tilde{\mathbf{w}}^* \in \mathbb{B}^d$ 使得

$$\left| p \circ E_g \circ \hat{\mathbf{f}}^k(\vec{\mathbf{x}}^t) - \hat{f}_{j,g,\tilde{\mathbf{w}}^*}^{k+1}(\vec{\mathbf{x}}^t) \right| \leq \frac{\epsilon_1}{2}$$

对任意 t 和 g。事实上，在这种情况下，方程（12）对 $\mathbf{w}^* = \tilde{\mathbf{w}}^*_{1+\epsilon_1/2}$ 成立。最后，由于

$$\hat{f}_{j,g,\mathbf{w}^*}^{k+1}(\vec{\mathbf{x}}) = \hat{f}_{j,g}^k(\vec{\mathbf{x}}) + \left\langle \mathbf{w}^*, \sigma(W_1^{k+1} E_g \circ \hat{\mathbf{f}}^k(\vec{\mathbf{x}}) + \mathbf{b}^{k+1}) \right\rangle$$

只需证明，以概率 $1 - \delta$ 存在 $\tilde{\mathbf{w}}^* \in \mathbb{B}^d$ 使得

$$\left| \tilde{p} \circ E_g \circ \hat{\mathbf{f}}^k(\vec{\mathbf{x}}^t) - \left\langle \mathbf{w}^*, \sigma(W_1^{k+1} E_g \circ \hat{\mathbf{f}}^k(\vec{\mathbf{x}}^t) + \mathbf{b}^{k+1}) \right\rangle \right| \leq \frac{\epsilon_1}{2}$$

对于多项式 $\tilde{p}(\mathbf{x}^1 | \ldots | \mathbf{x}^w) = p(\mathbf{x}^1 | \ldots | \mathbf{x}^w) - x_j^1$（注意，$\tilde{p}(E(\hat{\mathbf{f}}^k(\tilde{\mathbf{x}}))) = p(E(\hat{\mathbf{f}}^k(\tilde{\mathbf{x}}))) - \hat{\mathbf{f}}_j^k(\tilde{\mathbf{x}})$ 且 $\|\tilde{p}\|_{\text{co}} \leq \|p\|_{\text{co}} + 1$），以及任意 t 和 g。根据引理 5.3 并对 $X = \{E_g \circ \hat{\mathbf{f}}^k(\tilde{\mathbf{x}}^t) : g \in G, t \in [m]\}$ 应用并联合界，可知此类 $\mathbf{w}^*$ w.p. $1 - \delta$ 的存在性成立。 □

接下来我们给出以下引理，它定量描述了随着训练越来越深的层，不同标签的损失如何改善。

引理 5.5. 令 $\gamma = \frac{1}{32}\min\left(\frac{1}{B}, \xi\right)$ 假设 $1 > \beta \geq \beta(\gamma/2)$，并令 $\delta = m|G|\delta(\gamma/2, \beta, q, \|p\|_{\text{co}} + 5, wn)$，则：

- 对于任意 $j \in L_1$，以概率至少 $1-\delta$，$\ell_{S,\,j}(\mathbf{f}^1) \leq \frac{1}{4m|G|} + \epsilon_{\text{opt}}$
- 给定 $\ell_{S,j}(\mathbf{f}^k) \leq \frac{1}{2m|G|}$，我们有 $\ell_{S,j}(\mathbf{f}^{k+1}) \leq e^{-\gamma}\ell_{S,j}(\mathbf{f}^k) + \epsilon_{opt}$，且其成立概率至少为 $1-\delta$。此外，如果 $\epsilon_{opt} \leq \frac{1-e^{-\gamma}}{2m|G|}$，那么以至少 $1-t\delta$ 的概率，我们有 $\ell_{S,j}(\mathbf{f}^{k+t}) \leq e^{-\gamma t}\ell_{S,j}(\mathbf{f}^k) + \frac{1-e^{-\gamma t}}{1-e^{-\gamma}}\epsilon_{opt}$。
- 由于 $\ell_{S,j'}(\mathbf{f}^k) \leq \frac{\xi}{8m^2|G|^2}$ 对于任意 $j' \in L_{i-1}$，我们有 $\ell_{S,j}(\mathbf{f}^{k+1}) \leq \frac{1}{4m|G|} + \epsilon_{\text{最优}}$ 对于任意 $j \in L_i$，以概率 $1 - |L_i|\delta$

在证明引理 5.5 蕴含定理 5.1之前，我们先说明它确实蕴含定理。

证明。（定理 5.1的证明）取 $\beta = \beta(\gamma/2)$（更一般地，$1 > \beta \geq \beta(\gamma/2)$，使得 $1-\beta = \Omega(\gamma)$）。记 $\delta = m|G|\delta(\gamma/2, \beta, q, M+5, wn)$，并注意到根据方程（11）我们有

$$\delta = m|G|\exp\left(-\Omega\left(q \cdot \frac{\gamma^{2K+4}}{(wn)^{2K}(M+1)^4}\right)\right)$$

由于 $\epsilon_{\text{opt}} \leq \frac{(1-e^{-\gamma})\xi}{16m^2|G|^2}$，我们有：若 $\ell_{S,j}(\mathbf{f}^k) \leq \frac{1}{2m|G|}$，则以概率 $1-t\delta$

$$\ell_{S,j}(\mathbf{f}^{k+t}) \leq e^{-\gamma t}\ell_{S,j}(\mathbf{f}^k) + \frac{1}{1-e^{-\gamma}}\epsilon_{\text{opt}} \leq \frac{e^{-\gamma t}}{2m|G|} + \frac{\xi}{16m^2|G|^2}$$

取 $t_0 = \left\lceil \frac{\ln(8m|G|/\xi)}{\gamma} \right\rceil$，我们得到

$$\ell_{S,j}(\mathbf{f}^{k+t_0}) \leq \frac{\xi}{8m^2|G|^2}$$

以 $1 - t_0\delta$ 的概率。因此，通过对 $1 \leq i \leq r$ 进行归纳法，不难验证：对于任意 $j \in L_i$，若 $k \geq i(t_0+1)$，则

$$\ell_{S,j}(\mathbf{f}^k) \leq \frac{\xi}{8m^2|G|^2}$$

以 $1 - nk\delta$ 的概率 □

为证明引理 5.5 我们将使用以下事实，它是损失定义的直接推论。

事实 5.6. - 若 $\ell_{S,\,j}(\hat{\mathbf{f}}) \leq \frac{\epsilon}{m|G|}$ 则对于任意 $t \in [m]$ 和 $g \in G$，我们有 $1 \geq \hat{f}_{j,\,g}(\vec{\mathbf{x}}^t) \cdot f^*_{j,\,g}(\vec{\mathbf{x}}^t) \geq (1-\epsilon)\frac{2}{B}$
- 若 $\ell_{S,\,j}(\hat{\mathbf{f}}) \leq \frac{\epsilon}{4m^2|G|^2}$ 则对于任意 $t \in [m]$ 和 $g \in G$，我们有 $1 \geq \hat{f}_{j,\,g}(\vec{\mathbf{x}}^t) \cdot f^*_{j,\,g}(\vec{\mathbf{x}}^t) \geq (1-\epsilon)(1-\xi/2)$
- 如果对于任意 $t \in [m]$ 和 $g \in G$，都有 $1 \geq \hat{f}_{j,g}(\vec{\mathbf{x}}^t) \cdot f^*_{j,g}(\vec{\mathbf{x}}^t) \geq \frac{1}{B}$ 那么 $\ell_{\text{rob}S}{}^{1/2B}{}_{,j}(\hat{\mathbf{f}}) \leq \frac{1}{4m|G|}$

接下来我们证明引理 5.5。

证明。（引理 5.5）设 $p_1, \ldots p_n$ 是多项式，它们证明 $(\mathcal{L}, e)$ 是 $\mathbf{f}^*$ 的一个 (r, K, M, B, ξ)-层级。我们从第一项开始。根据层级的定义，对于任意 $t \in [m]$ 和 $g \in G$，都有 $B \geq p_j(E_p(\mathbf{f}^0(\vec{\mathbf{x}}^t)))f_{j,g}(\vec{\mathbf{x}}^t) \geq 1$。事实 5.6 表明，对于 $\tilde{p}_j = \frac{1}{B}p_j$，有 $\ell^{\text{rob}}{}_{,S}{}^{\gamma}{}_{,j}(\tilde{p}_j \circ \hat{\mathbf{f}}^0) \leq \ell^{\text{rob}}{}_{,S}{}^{1/2B}{}_{,j}(\tilde{p}_j \circ \hat{\mathbf{f}}^0) \leq \frac{1}{4m|G|}$。因此，第一项由引理 5.4 得证。

第三项的证明类似。若 $\ell_{S,j'}(\mathbf{f}^k) \leq \frac{\xi}{8m^2|G|^2}$ 对任意 $j' \in L_{i-1}$，则事实 5.6 意味着对任意 $j' \in L_{i-1}$，$t \in [m]$ 且 $g \in G$ 我们有

$$1 \geq y^t_{j',g}\hat{f}^k_{j',g}(\vec{\mathbf{x}}^t) \geq (1-\xi/2)(1-\xi/2) \geq 1-\xi$$

因此，根据层级的定义，对于任意 $t \in [m]$ 和 $g \in G$, $B \geq p_j(E_p(\hat{\mathbf{f}}^k(\bar{\mathbf{x}}^t)))f_{j,g}(\bar{\mathbf{x}}^t) \geq 1$。事实 5.6 现在意味着，对于 $\tilde{p}_j = \frac{1}{B}p_j$，我们有 $\ell^{\text{rob}}{}_{,S}{}^{\gamma}{}_{,j}(\tilde{p}_j \circ \hat{\mathbf{f}}^k) \leq \ell^{\text{rob}}{}_{,S}{}^{1/2B}{}_{,j}(\tilde{p}_j \circ \hat{\mathbf{f}}^k) \leq \frac{1}{4m|G|}$。由此，第三项可由引理 5.4 得出。

还需证明 第二项。定义 $q: \mathbb{R}^n \to \mathbb{R}$ 由 $q(\mathbf{x}) = 1.5x_j - 0.5x_j^3$。根据引理 5.4，可知 s 只需证明

$$\ell_{S,j}^{\text{rob},\gamma}(q \circ \hat{\mathbf{f}}^k) \leq e^{-\gamma}\ell_{S,j}(\hat{\mathbf{f}}^k) \tag{13}$$

为此，注意到由于 $\ell_{S,j}(\hat{\mathbf{f}}^k) \leq \frac{1}{2m|G|}$，事实 5.6 表明 $\forall t, g,\ y_{j,g}^t \hat{f}_{j,g}^k(\bar{\mathbf{x}}^t) \geq 1/(4B)$。现在，由于 q 是奇数，我们有

$$\ell\left(y_{j,g}^t q\left(\hat{f}_{j,g}^k(\vec{\mathbf{x}}^t)\right)\right) = \ell\left(q\left(y_{j,g}^t \cdot \hat{f}_{j,g}^k(\vec{\mathbf{x}}^t)\right)\right)$$

因此，方程 (13) 可由以下断言推出

断言 1. 设 $\tilde{q}(x) = 1.5x - 0.5x^3$。那么，对于任意 $\frac{1}{4B} \leq x \leq 1$，都有 $\ell^{\text{rob},\gamma}(\tilde{q}(x)) = \ell(\tilde{q}(x) - \gamma) \leq e^{-\gamma}\ell(x)$。

证明。 记 $x' = \min(x,\ 1 - \xi/2)$，并注意到 $\ell(x) = \ell(x')$，且

$$\tilde{q}(x') - x' = \frac{1}{2}x'(1 - x'^2) = \frac{1}{2}x'(1 - x')(1 + x') \geq \frac{1}{2}x'(1 - x') \geq \frac{1}{4}\min\left(1/4B, 1/2\xi\right) \geq 2\gamma \tag{14}$$

现在，我们有

$$\begin{aligned}
\ell(\tilde{q}(x) - \gamma) &\overset{x' \leq x}{\leq} \ell(\tilde{q}(x') - \gamma) \\
&\overset{\text{Eq. (14)}}{\leq} \ell(x' + \gamma) \\
&= \ell\left(\frac{1 - x' - \gamma}{1 - x'}x' + \frac{\gamma}{1 - x'}\right) \\
&\overset{\text{Convexity and } \frac{\gamma}{1-x'} \leq 1}{\leq} \frac{1 - x' - \gamma}{1 - x'}\ell(x') + \frac{\gamma}{1 - x'}\ell(1) \\
&\overset{\ell(1)=0 \text{ and } \ell(x')=\ell(x)}{=} \frac{1 - x' - \gamma}{1 - x'}\ell(x) \\
&\leq e^{-\gamma}\ell(x)
\end{aligned}$$

□

□

6 结论与未来工作

在这项工作中，我们论证了广泛且细粒度的标签的可用性表明，现代深度学习中的目标函数本质上具有层次结构；我们还展示了深度学习——特别是残差网络上的随机梯度下降——能够利用这种层次结构。我们的证明建立在学习过程的逐层机制之上，其中每一层同时充当表征学习器和预测器，迭代地细化前一层的输出。我们的结果引出了几个值得探讨的方向，概述如下：

- **监督学习本质上是易处理的。** 不同于最坏情况的困难性结果，教师的存在（以及由此产生的层级）意味着，在获得正确监督的情况下，该问题可以在多项式时间内学习。

- **极深模型已被证明是可学习的。** 与以往的理论研究不同，我们证明残差网络能够学习那些只有极深电路才能实现的模型。

- **软件工程与学习之间的折中方案**。现代深度学习可以被视为对软件工程的放宽 和对经典学习的强化。我们既不手动“编写大脑的算法”（传统人工智能），也不从输入-输出对中盲目学习（经典机器学习），而是通过相关标签提供大脑逻辑的片段。这种方法使学习任务变得可行，而无需全面了解底层电路。

- **学习理论的新叙述**。历史上，尤其是从计算视角来看，学习理论的主导叙述如下：（i）不可能学习所有函数。（ii）仔细考察后，我们发现只对能够高效计算的函数感兴趣。（iii）这一函数类可以通过多项式数量的样本来学习。（iv）遗憾的是，学习它需要指数时间。（v）尽管如此，一些简单的函数类仍然可以学习。

  然而，上述叙述与实践并不相符。我们的工作表明，或许可以将第（v）项替换为以下内容：“（v）重新审视我们的研究范围后，我们主要关注人类能够高效计算的函数。(vi) 我们有充分理由相信，这些函数具有层次结构。（vii）因此，可以用多项式时间和多项式数量的样本来学习它们。”

我们的工作建议以层次模型为基础来理解神经网络。要推进这一方向，未来还需要开展大量工作。首先，在理论方面，拓展层次模型的适用范围将会有所裨益。为此，可以考虑：

- 从层次模型的视角分析注意力机制。

- 扩展层次模型，使其能够刻画“单函数层次结构”。这指的是这样一种情形：函数 f 具有易于学习的“简单版本”，掌握这些版本后，f 函数本身也就容易学习。这与此前通过基于梯度的算法研究非线性模型可学习性的工作相一致（例如，[1]），因为这些研究中有许多都假设（通常是隐含地假设）目标模型具有这样的层次结构。

- 通过推广定理 3.4，进一步阐明层次模型的内在合理性。也就是说，定义教师对其内部逻辑“部分知晓”的形式化模型，并证明这类教师能够提供有助于高效学习的层次标注。换言之，证明层次函数的“通用非线性投影”本身也是层次结构。

- 为已知算法找出低复杂度层次结构。这可能催生新的层次架构，甚至有助于揭示人类如何发现这些算法，并促进其教学。

其次，在实证方面，以下工作将很有价值：

- 基于层次学习视角，更直接地运用有理论依据的优化方法，构建实用的学习算法。

- 通过实证检验这一假设：只要标签足够多，真实世界数据就呈现出层次结构。在这方面，发现这种显式的层次结构，可以视为对学习得到的模型的一种解释。

最后，我们会指出本研究结果的具体局限性，因为这些结果依赖于若干假设。下面概述其中最突出的一些，并希望未来的工作能够放宽这些限制。

我们先讨论技术假设。未来工作的一个明确方向是改进我们的定量界限；尽管这些界限是多项式级别的，但很可能远非最优。其他技术限制包括假设输出矩阵是正交的，以及标签数量等于隐藏层的维度。更自然的做法是考虑任意数量的标签，并将输出矩阵初始化为 Xavier 矩阵（不过，我们指出，Xavier 矩阵“几乎正交”）。最后，我们的分析所用的损失函数并非常见形式。

接下来，我们讨论一些更为内在的局限性。首先，我们假设了极强的监督条件：每个样本都带有其拥有的全部正标签。但在实践中，通常每个样本只能获得一个正

标签。我们注意到，尽管可以轻易证明，在这种标准监督下层次模型能够被高效学习，但要证明神经网络上的基于梯度的算法在这一设置下能够成功，仍然是一个悬而未决的问题。

另一个局限在于，我们假设采用逐层训练，而实际上通常会对所有层进行联合训练。虽然这使数学分析更加复杂，但联合训练可能有几个方面的优势。首先，从经验上看，它是标准方法。其次，如果训练较低层的目标仅仅是学习表征，那么为了让损失略有改善而耗费大量数据并无多大意义。事实上，为了确保数据效率，最好在特征足够好时就加以利用（也就是说，这些特征的梯度较大时）。

7 更多预备知识

在下文中，我们用 $(\mathbb{R}^n)^{\otimes t}$ 表示所有轴的维度均为 n 的 t 阶实张量空间。我们为其配备内积 $\langle A, B\rangle = \sum_{1\le i_1,\ \ldots,\ i_t\le n} A_{i_1,\ \ldots,\ i_t} B_{i_1,\ \ldots,\ i_t}$。对于 $\mathbf{x}\in\mathbb{R}^d$，我们用 $\mathbf{x}^{\otimes t}\in(\mathbb{R}^n)^{\otimes t}$ 表示一个张量，其 $(i_1,\ \ldots,\ i_t)$ 项为 $\prod_{j=1}^t x_{i_j}$。注意，$\langle \mathbf{x}^{\otimes t},\ \mathbf{y}^{\otimes t}\rangle = \langle \mathbf{x},\ \mathbf{y}\rangle^t$。

7.1 集中不等式

我们将使用切尔诺夫不等式和霍夫丁不等式：

引理 7.1 (霍　　夫丁)。 设 $X_1, \ldots, X_q \in [-B, B]$ 独立同分布，均值为 μ。那么，对于任意 $\epsilon > 0$，我们有

$$\Pr\left(\left|\frac{1}{q}\sum_{i=1}^{q} X_i - \mu\right| \ge \epsilon\right) \le 2e^{-\frac{q\epsilon^2}{2B^2}}$$

引理 7.2 (切尔诺夫)。 设 $X_1, \ldots, X_q \in \{0,1\}$ 独立同分布，均值为 μ。那么，对于任意 $0 \le \epsilon \le \mu$，我们有

$$\Pr\left(\left|\frac{1}{q}\sum_{i=1}^{q} X_i - \mu\right| \ge \epsilon\right) \le 2e^{-\frac{q\epsilon^2}{3\mu}}$$

我们还需要以下形式的切尔诺夫界。

引理 7.3。 设 $X_1, \ldots, X_q \in \{-1, 1, 0\}$ 为均值为 μ 的独立同分布随机变量。那么对于 $\epsilon \le \min(\Pr(X_i=1), \Pr(X_i=-1))\, 2|\mu|$，$\Pr\left(\left|\frac{1}{q|\mu|}\sum_{i=1}^n X_i - \frac{\mu}{|\mu|}\right| \ge \epsilon\right) \le 4e^{-\frac{q\epsilon^2|\mu|^2}{12\Pr(X_i\neq 0)}}$

证明。(引理 7.3 的证明) 设 $X_i^+ = \max(X_i, 0)$，并设 $\mu_+ = \mathbb{E}X_i^+ = \Pr(X_i = 1)$。类似地，设 $X_i^- = \max(-X_i, 0)$，且 $\mu_- = \mathbb{E}X_i^- = \Pr(X_i = -1)$。根据切尔诺夫界（引理 7.2），对于 $0 \le \delta \le 1$

$$\Pr\left(\left|\frac{1}{q}\sum_{i=1}^{n} X_i^+ - \mu_+\right| \ge \delta\mu_+\right) \le 2e^{-\frac{q\delta^2\mu_+}{3}}$$

因此，

$$\Pr\left(\left|\frac{1}{q|\mu|}\sum_{i=1}^{n} X_i^+ - \frac{\mu_+}{|\mu|}\right| \ge \delta\frac{\mu_+}{|\mu|}\right) \le 2e^{-\frac{q\delta^2\mu_+}{3}}$$

定义 $\epsilon = \delta\frac{\mu_+}{|\mu|}$，则对于 $\epsilon \le \frac{\mu_+}{|\mu|}$ 有

$$\Pr\left(\left|\frac{1}{q|\mu|}\sum_{i=1}^{n} X_i^+ - \frac{\mu_+}{|\mu|}\right| \ge \epsilon\right) \le 2e^{-\frac{q\epsilon^2|\mu|^2}{3\mu_+}} \le 2e^{-\frac{q\epsilon^2|\mu|^2}{3\Pr(X_i\neq 0)}}$$

类似的论证意味着，对于 $\epsilon \le \frac{\mu_-}{|\mu|}$，我们有

$$\Pr\left(\left|\frac{1}{q|\mu|}\sum_{i=1}^{n} X_i^- - \frac{\mu_-}{|\mu|}\right| \ge \epsilon\right) \le 2e^{-\frac{q\epsilon^2|\mu|^2}{3\Pr(X_i\neq 0)}}$$

因此，对于 $\epsilon \leq \min(\mu_+, \mu_-)\, 2|\mu|$，我们有

$$\begin{aligned}\Pr\left(\left|\frac{1}{q|\mu|}\sum_{i=1}^{n} X_i - \frac{\mu}{|\mu|}\right| \geq \epsilon\right) &\leq \Pr\left(\left|\frac{1}{q|\mu|}\sum_{i=1}^{n} X_i^+ - \frac{\mu_+}{|\mu|}\right| \geq \frac{\epsilon}{2}\right) + \Pr\left(\left|\frac{1}{q|\mu|}\sum_{i=1}^{n} X_i^- - \frac{\mu_-}{|\mu|}\right| \geq \frac{\epsilon}{2}\right) \\ &\leq 4e^{-\frac{q\epsilon^2|\mu|^2}{12\Pr(X_i\neq 0)}}\end{aligned}$$

□

7.2 杂项引理

我们将使用以下关于二项式系数的渐近结果，该结果由斯特林近似推出

引理 7.4。我们有 $\binom{2k}{k} 2^{2k} \sim \frac{1}{\sqrt{\pi k}}$

我们还需要使用以下利用多项式近似符号函数的结果。

引理 7.5。设 $0 < \xi < 1$ 和 $\epsilon > 0$。存在一个多项式 $p\colon \mathbb{R} \to \mathbb{R}$ 使得

- $p([-1,1]) \subseteq [-1,1]$
- 对于任意 $x \in [-1,1] \setminus [-\xi, \xi]$，都有 $|p(\mathbf{x}) - \text{符号函数}(\mathbf{x})| \leq \epsilon$。
- 次数$(p) = O\left(\frac{\log(1/\epsilon)}{\xi}\right)$
- p的系数均有界于 $2^{O(\log(1/\epsilon)/\xi)}$ ——

满足前三个性质的多项式的存在性在 [17]中得到了证明。系数的界（最后一项）由 [31] 中的引理 2.8 得出（另见此处）。最后，我们将使用以下关于多项式与线性函数复合的系数范数的界。

引理 7.6。固定一个次数为 K 的多项式 $p\colon \mathbb{R}^n \to \mathbb{R}$ 以及 $A \in M_{n,m}$ 其行的欧几里得范数至多为 R。定义 $q(\mathbf{x}) = p(A\mathbf{x})$。则，$\|q\|_{\text{co}} \leq \|p\|_{\text{co}} R^K (n+1)^{K/2}$

证明。令$\mathbf{a}_i$ 为 A的第 i行。记 $p(\mathbf{x}) = \sum_{\alpha\in\{0,\ldots,K\}^n,\ \|\alpha\|_1\leq K} b_\alpha \mathbf{x}^\alpha$ 以及 $e_\alpha(\mathbf{x}) = \prod_{i=1}^n \langle \mathbf{a}_i,\ \mathbf{x}\rangle^{\sigma_i}$。我们有 $q = \sum_{\alpha\in\{0,\ldots,K\}^n,\ \|\alpha\|_1\leq K} b_\alpha e_\alpha$。因此，

$$\|q\|_{\text{co}} \leq \sum_{\alpha\in\{0,\ldots,K\}^n, \|\alpha\|_1\leq K} |b_\alpha| \cdot \|e_\alpha\| \overset{\text{C.S.}}{\leq} \|p\|_{\text{co}} \cdot \sqrt{\sum_{\alpha\in\{0,\ldots,K\}^n, \|\alpha\|_1\leq K} \|e_\alpha\|^2}$$

最后

$$\|e_\alpha\|^2 = \left\|\mathbf{a}_1^{\otimes\sigma_1} \otimes \ldots \otimes \mathbf{a}_n^{\otimes\sigma_n}\right\|^2 = \prod_{i=1}^n \|\mathbf{a}_1\|^{2\sigma_i} \leq R^{2K}$$

□

7.3 泛化结果

众所周知，对于由参数向量定义函数的“行为良好”函数类，其样本复杂度与参数数量成正比。例如，对于形如 $\mathcal{F} = \{\mathbf{x} \mapsto F(\mathbf{w}, \mathbf{x}) : \mathbf{w} \in [-B, B]^p\}$ 的函数类，其中函数 F关于第一个参数是 L-利普希茨的，其可实现的大间隔样本复杂度为 $\tilde{O}\left(\frac{p}{\epsilon}\right)$。更准确地说，如果 $\mathcal{F}$ 中存在一个 γ-误差为 0 的函数，那么任何能保证返回经验 γ-误差为 0 的函数的算法，都能享有上述样本复杂度保证。接下来，我们将稍微扩展这一事实，允许 F 是随机的，并允许算法以较小的概率失败。

引理 7.7. 假设 $\mathcal{F} \subset (\mathbb{R}^n)^{\mathcal{X}}$ 是一个随机函数类，使得

- 存在一个随机函数 F: $[-B,B]^p \times \mathcal{X} \to \mathbb{R}^n$ 使得 $\mathcal{F} = \{\mathbf{x} \mapsto F(\mathbf{w}、\mathbf{x})\colon \mathbf{w} \in [-B,B]^p\}$
- 以 $1-\delta_1$ 的概率，对于任意 $\mathbf{x} \in \mathcal{X}$，$\mathbf{w} \mapsto F(\mathbf{w}, \mathbf{x})$ 相对于 ℓ^∞ 范数是 L-利普希茨的。

设 $\mathcal{A}$ 是一个算法，并假设对于某个 $\mathbf{f}^*\colon \mathcal{X} \to \{\pm 1\}^n$，$\mathcal{A}$ 具有如下性质：对于任何由 $\mathbf{f}^*$ 标记的 m 点ˆ ˆ 样本 S，它都以 $1-\delta_2$ 的概率返回一个 $\mathbf{f} \in \mathcal{F}$，满足 $\mathrm{Err}_{S,\gamma}(\mathbf{f}) = 0$ （概率是针对 F 的随机性和 $\mathcal{A}$ 的内部随机性而言的）。那么，如果 S 是由 $\mathbf{f}^*$ 标记的独立同分布样本，我们有

- $\mathrm{Err}_{\mathcal{D}}(\hat{\mathbf{f}}) \le \epsilon$ $w.p.$ $(LB/\gamma)^{O(p)}(1-\epsilon)^m + \delta_1 + \delta_2$
- $\mathbb{E}_S \mathrm{Err}_{\mathcal{D}}(\hat{\mathbf{f}}) \le O\left(\frac{p\ln(LB/\gamma)+\ln(m)}{m}\right) + \delta_1 + \delta_2$

ˆ证明。（概要）对于 $\mathbf{f}\colon \mathcal{X} \to \mathbb{R}^n$，我们定义

$$\mathrm{Err}_{\mathcal{D},\gamma}(\hat{\mathbf{f}}) = \Pr_{\mathbf{x}\sim\mathcal{D}}\left(\exists i \in [n] \text{ s.t. } \hat{f}_i(\mathbf{x}) \cdot f_i(\mathbf{x}) < \gamma\right)$$

不难看出，以 $1-\delta_1$ 的概率，存在一个大小为 $N = (LB/\gamma)^{O(p)}$ 的 $\tilde{\mathcal{F}} \subseteq \mathcal{F}$，使得对于任意 $\mathbf{g} \in \mathcal{F}$，都存在˜ 个 ˜$\mathbf{g} \in \mathcal{F}$，使得

$$\forall \mathbf{x} \in \mathcal{X},\ \|\mathbf{g}(\mathbf{x}) - \tilde{\mathbf{g}}(\mathbf{x})\|_\infty \le \frac{\gamma}{2}$$

˜ ˆ令 A 表示这样的事件：存在这样的 $\mathcal{F}$，使得 $\mathcal{A}$ 返回一个属于 $\mathcal{F}$ 的函数，且 $\mathrm{Err}_{S,\gamma}(\mathbf{f}) = 0$；并且对于任意满足 $\mathrm{Err}_{\mathcal{D},\gamma/2}(\tilde{\mathbf{g}}) \ge \epsilon$ 的 ˜$\mathbf{g} \in \mathcal{F}$，都有 $\mathrm{Err}_{S,\gamma/2}(\tilde{\mathbf{g}}) > 0$。事件 A 发生的概率至少为$1-\delta_1-\delta_2-N(1-\epsilon)^m$。给定 A，对于任意$\mathbf{g} \in \mathcal{F}$，我们有：

$$\mathrm{Err}_{\mathcal{D}}(\mathbf{g}) \ge \epsilon \Rightarrow \mathrm{Err}_{\mathcal{D},\gamma/2}(\tilde{\mathbf{g}}) \ge \epsilon \Rightarrow \mathrm{Err}_{S,\gamma/2}(\tilde{\mathbf{g}}) > 0 \Rightarrow \mathrm{Err}_{S,\gamma}(\mathbf{g}) > 0$$

因此，$\mathcal{A}$ 返回误差为 $\ge \epsilon$ 的函数的概率至多为 $N(1-\epsilon)^m + \delta_1 + \delta_2$，这就证明了引理的第一部分。至于第二部分，我们注意到：

$$\mathbb{E}_S \mathrm{Err}_{\mathcal{D}}(\hat{\mathbf{f}}) \le \mathbb{E}_S[\mathrm{Err}_{\mathcal{D}}(\hat{\mathbf{f}})|A] + \Pr(A^{\complement}) \le \epsilon + N(1-\epsilon)^m + \delta_1 + \delta_2$$

对 ϵ 进行优化，得到 $\mathbb{E}_S \mathrm{Err}_{\mathcal{D}}(\hat{\mathbf{f}}) \le \frac{\ln(Nm)}{m} + \delta_1 + \delta_2$，这就证明了第二部分 □

7.4 核

接下来陈述的结果可见于 Sch¨olkopf 和 Smola [30]的第2章。设 $\mathcal{X}$ 是一个集合。核 是一个函数 k: $\mathcal{X} \times \mathcal{X} \to \mathbb{R}$，满足对于每个 $x_1, \ldots, x_m \in \mathcal{X}$，矩阵 $\{k(x_i, x_j)\}_{i,j}$ 都是半正定的。核空间 是一个希尔伯特空间 $\mathcal{H}$，其元素为从 $\mathcal{X}$ 到 $\mathbb{R}$ 的函数，并且对于每个 $x \in \mathcal{X}$，线性泛函 $f \in \mathcal{H} \mapsto f(x)$ 都是有界的。下面的定理描述了核与核空间之间的一一对应关系。

定理7.8。对于每个核 k，都存在唯一的核空间 $\mathcal{H}_k$，使得对于每个 $x, x' \in \mathcal{X}$,$k(x, x') = \langle k(\cdot, x)\ k(\cdot, x') \rangle_{\mathcal{H}_k}$。同样，对于每个核空间 $\mathcal{H}$，都存在一个核 k，使得 $\mathcal{H} = \mathcal{H}_k$。

我们用 $\|\cdot\|_k$ 和 $\langle \cdot, \cdot \rangle_k$ 表示 $\mathcal{H}_k$ 中的范数和内积。下面的定理描述了核与将 $\mathcal{X}$ 嵌入希尔伯特空间之间的紧密联系。

定理 7.9。函数 $k\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ 是核，当且仅当存在一个映射 $\Psi\colon \mathcal{X} \to \mathcal{H}$ 到某个希尔伯特空间，使得 $k(x, x') = \langle \Psi(x)\ \Psi(x') \rangle_{\mathcal{H}}$。在这种情况下，$\mathcal{H}_k = \{f_{\Psi,\mathbf{v}} \mid \mathbf{v} \in \mathcal{H}\}$，其中$f_{\Psi,\mathbf{v}}(x) = \langle \mathbf{v}\ \Psi(x) \rangle_{\mathcal{H}}$。此外，$\|f\|_k = \min\{\|\mathbf{v}\|_{\mathcal{H}}\colon f_{\Psi,\mathbf{v}}\}$，且极小化器是唯一的。

7.5 随机特征方案

设 $\mathcal{X}$ 为一个可测空间，$k\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ 为一个核。k 的一个随机特征方案 (RFS) 是一对 (ψ, μ)，其中 μ 是定义在可测空间 Ω 上的概率测度，且 $\psi\colon \Omega \times \mathcal{X} \to \mathbb{R}$ 是可测函数，使得

$$\forall \mathbf{x}, \mathbf{x}' \in \mathcal{X}, \quad k(\mathbf{x}, \mathbf{x}') = \mathbb{E}_{\omega\sim\mu}\psi(\omega, \mathbf{x})\psi(\omega, \mathbf{x}')\,. \tag{15}$$

我们通常将 ψ（而非 (ψ,μ)）称为 RFS。我们定义 $\|\psi\|_\infty = \sup_{\mathbf{x}} \|\psi(\cdot,\mathbf{x})\|_\infty$，并称 ψ 是 C-有界的，若 $\|\psi\|_\infty \le C$。由 ψ 生成的随机 q-嵌入 是随机映射

$$\Psi_{\boldsymbol{\omega}}(\mathbf{x}) := (\psi(\omega_1,\mathbf{x}),\ldots,\psi(\omega_q,\mathbf{x}))^\top,$$

其中 $\omega_1,\ldots,\omega_q \sim \mu$ 相互独立同分布。与 $\Psi_{\boldsymbol{\omega}}$ 对应的随机 q-核 为 $k_{\boldsymbol{\omega}}(\mathbf{x},\mathbf{x}') = \langle \Psi_{\boldsymbol{\omega}}(\mathbf{x}), \Psi_{\boldsymbol{\omega}}(\mathbf{x}') \rangle$。同样，与 $\frac{1}{\sqrt{q}}\Psi_{\boldsymbol{\omega}}$ 对应的随机 q-核空间 为 $\mathcal{H}_{k_{\boldsymbol{\omega}}}$。接下来，我们讨论用 $\mathcal{H}_{k_{\boldsymbol{\omega}}}$ 中的函数近似 $\mathcal{H}_k$ 中的函数。考虑嵌入将会很有帮助

$$\mathbf{x} \mapsto \Psi^{\mathbf{x}} \quad \text{where} \quad \Psi^{\mathbf{x}} := \psi(\cdot,\mathbf{x}) \in L^2(\Omega). \tag{16}$$

由 (15) 可知，对于任意 $\mathbf{x},\mathbf{x}' \in \mathcal{X}$, $k(\mathbf{x},\mathbf{x}') = \left\langle \Psi^{\mathbf{x}}, \Psi^{\mathbf{x}'} \right\rangle_{L^2(\Omega)}$。特别地，由定理 7.9可知，对于每个 $f \in \mathcal{H}_k$，存在唯一的函数 $\check{f} \in L^2(\Omega)$ 使得

$$\|\check{f}\|_{L^2(\Omega)} = \|f\|_k \tag{17}$$

并且对于每个 $\mathbf{x} \in \mathcal{X}$,

$$f(\mathbf{x}) = \left\langle \check{f}, \Psi^{\mathbf{x}} \right\rangle_{L^2(\Omega)} = \mathbb{E}_{\omega\sim\mu} \check{f}(\omega)\psi(\omega,\mathbf{x}). \tag{18}$$

记 $f_{\boldsymbol{\omega}}(\mathbf{x}) = \frac{1}{q}\sum_{i=1}^q \left\langle \check{f}(\omega_i)\ \psi(\cdot,\mathbf{x}) \right\rangle$。由 (18) 可知，$\mathbb{E}_{\boldsymbol{\omega}}[f_{\boldsymbol{\omega}}(\mathbf{x})] = f(\mathbf{x})$。此外，对于每个$\mathbf{x}$，$f_{\boldsymbol{\omega}}(\mathbf{x})$ 的方差至多为

$$\begin{aligned}\frac{1}{q}\mathbb{E}_{\omega\sim\mu}\left|\check{f}(\omega)\psi(\omega,\mathbf{x})\right|^2 &\le \frac{\|\psi\|_\infty^2}{q}\mathbb{E}_{\omega\sim\mu}\left|\check{f}(\omega)\right|^2 \\ &= \frac{\|\psi\|_\infty^2\|f\|_k^2}{q}.\end{aligned}$$

直接得到的结果是以下推论。

推论 7.10（函数近似）。对于所有 $\mathbf{x} \in \mathcal{X}$，$\mathbb{E}_{\boldsymbol{\omega}}|f(\mathbf{x}) - f_{\boldsymbol{\omega}}(\mathbf{x})|^2 \le \frac{\|\psi\|_\infty^2\|f\|_k^2}{q}$。

现在，如果 $\mathcal{D}$ 是 $\mathcal{X}$ 上的一个分布，我们得到

$$\mathbb{E}_{\boldsymbol{\omega}}\|f - f_{\boldsymbol{\omega}}\|_{2,\mathcal{D}} \overset{\text{Jensen}}{\le} \sqrt{\mathbb{E}_{\boldsymbol{\omega}}\|f - f_{\boldsymbol{\omega}}\|_{2,\mathcal{D}}^2} = \sqrt{\mathbb{E}_{\boldsymbol{\omega}}\mathbb{E}_{\mathbf{x}\sim\mathcal{D}}|f(\mathbf{x}) - f_{\boldsymbol{\omega}}(\mathbf{x})|^2} = \sqrt{\mathbb{E}_{\mathbf{x}}\mathbb{E}_{\boldsymbol{\omega}}|f(\mathbf{x}) - f_{\boldsymbol{\omega}}(\mathbf{x})|^2} \le \frac{\|\psi\|_\infty\|f\|_k}{\sqrt{q}}$$

因此，$O\left(\frac{\|f\|_k^2}{\epsilon^2}\right)$ 个随机特征足以保证$\mathbb{E}_{\boldsymbol{\omega}}\|f - f_{\boldsymbol{\omega}}\|_{2,\mathcal{D}} \le \epsilon$。在本文中，这样的 ℓ^2保证并不足够，我们需要关于更强的ℓ^∞ 范数，用 $\mathcal{H}_{k_{\boldsymbol{\omega}}}$ 中的函数来近似 $\mathcal{H}_k$ 中的函数。接下来我们将展示如何实现这一点，但遗憾的是，所需特征数量将呈二次增长。对于 $z \in \mathbb{R}$，我们定义 $\langle z\rangle_B = \begin{cases} z & |z| \le B \\ 0 \end{cases}$，否则。我们将考虑 $f_{\boldsymbol{\omega}}$ 的以下截断版本

$$f_{\boldsymbol{\omega},B}(\mathbf{x}) = \frac{1}{q}\sum_{i=1}^q \left\langle \check{f}(\omega_i)\right\rangle_B \cdot \psi(\omega_i,\mathbf{x})$$

现在，若 ψ 是 C-有界的，则 $f_{\boldsymbol{\omega},B}(\mathbf{x})$ 是 q 个独立同分布的 CB-有界随机变量的平均值。根据霍夫丁不等式，我们有

$$\Pr\left(|f_{\boldsymbol{\omega},B}(\mathbf{x}) - \mathbb{E}_{\boldsymbol{\omega}'} f_{\boldsymbol{\omega}',B}(x)| > \epsilon/2\right) \le 2e^{-\frac{q\epsilon^2}{8B^2C^2}} \tag{19}$$

同理，我们有

$$\begin{aligned}
|f(x)-\mathbb{E}_{\boldsymbol{\omega}'}f_{\boldsymbol{\omega}',B}(x)| &= |\mathbb{E}\left(f_{\boldsymbol{\omega}}(x)-f_{\boldsymbol{\omega},B}(x)\right)| \\
&= \left|\mathbb{E}\left(\check{f}(\omega)-\left\langle\check{f}(\omega)\right\rangle_B\right)\cdot\psi(\omega,\mathbf{x})\right| \\
&= \left|\mathbb{E}1_{|\check{f}(\omega)|>B}\check{f}(\omega)\psi(\omega,\mathbf{x})\right| \\
&\leq \sqrt{\Pr(|\check{f}(\omega)|>B)\mathbb{E}\left(\check{f}(\omega)\psi(\omega,\mathbf{x})\right)^2} \\
&\leq \|\psi\|_\infty\sqrt{\Pr(|\check{f}(\omega)|>B)\mathbb{E}\left(\check{f}(\omega)\right)^2} \\
&= \frac{\|\psi\|_\infty\|f\|_k^2}{B}
\end{aligned}$$

我们得到

引理 7。 **11.** 设 $f\in\mathcal{H}_k$，其中 $\|f\|_k\leq M$，并假设 $\|\psi\|_\infty\leq C$。对于 $B=\frac{2CM^2}{\epsilon}$，我们有

$$\Pr\left(|f_{\boldsymbol{\omega},B}(\mathbf{x})-f(\mathbf{x})|>\epsilon\right)\leq 2e^{-\frac{q\epsilon^4}{32M^4C^4}}$$

此外，定义 $f_{\boldsymbol{\omega},B}$ 的权重向量的范数，即 $\mathbf{w}=\frac{1}{q}\left(\left\langle\check{f}(\omega_1)\right\rangle_B,\ldots,\left\langle\check{f}(\omega_q)\right\rangle_B\right)$，满足 s

$$\|\mathbf{w}\|\leq\frac{2CM^2}{\epsilon\sqrt{q}}$$

8 层次结构示例与证明定理3.4

固定 $\mathcal{X}\subset[-1,1]^n$、邻近映射 $\mathsf{e}:G\to G^w$，以及一个集合族 $\mathcal{L}=\{L_1,\ldots,L_r\}$，使得 $L_1\subseteq L_2\subset\ldots\subseteq L_r=[n]$。到目前为止，我们只见过一个层级的正式示例：在非集成设定下（即 $w=|G|=1$），例 3.2 表明，如果任一标签依赖于 K 个更简单的标签，并且第一级的标签是输入的 $(K,1)$-多项式阈值函数，那么 $\mathcal{L}$ 就是一个 $(r,K,1)$-层级。本节将扩展示例的范围。我们首先证明（引理 8.1），如果 $(\mathcal{L},\mathsf{e})$ 是一个 (r,K,M)-层级，那么对于适当的 B 和 ξ，它也是一个 $(r,K,2M,B,\xi)$-层级。接着，在第 8.1 节中，我们将更详细地考察每个标签依赖于少数几个更简单标签、且这些标签位于少数几个位置的情形，并证明此时可以改进引理 8.1 得到的参数。最后，在第 8.2 节中，我们证明定理 3.4：如果所有标签都是来自某个给定电路的“随机片段”，且数量足够多，那么目标函数就具有一个低复杂度层级。

引理 8.1。 任何(r,K,M)-层级的$\mathbf{f}^*:\ \mathcal{X}^G\to\{\pm1\}^{n,\ G}$也是一个$(r,K,2M,B,\xi)$-层级，适用于 $\xi=\frac{1}{2\ wn+1\ \left(\right)^{\frac{K+1}{2}}KM}$和 $B=2(w\max(n,d)+1)^{K/2}M$

引理8.1 由层级的定义和以下引理可直接得出

引理 8.2。 任何(K,M)-多项式阈值函数 $f:\ \mathcal{X}\to\{\pm1\}$相对于 $\xi=\frac{1}{2\ n+1\ \left(\right)^{\frac{K+1}{2}}KM}$和$B=2(n+1)^{k/2}M$，都是一个$(K,2M,B,\xi)$-多项式阈值函数

引理 8.2 可由引理 8.3 和8.4推出

引理 8.3。 设 $p:\ \mathbb{R}^n\to\mathbb{R}$是一个次数为 K 的多项式。那么 p关于 $\|\cdot\|_\infty$范数，在 $[-1,1]^n$上是$((n+1)^{\frac{K+1}{2}}K\|p\|_{\mathrm{co}})$-利普希茨函数，并且对任意 $\mathbf{x}\in[-1,1]^n$，满足 $|p(\mathbf{x})|\leq(n+1)^{k/2}\|p\|_{\mathrm{co}}$。

证明。 记 $p(\mathbf{x})$ 为 $=\sum_{\alpha\in\{0,\ldots,K\}^n,\ \|\alpha\|_1\leq K}a_\alpha\mathbf{x}^\alpha$。我们有

$$\frac{\partial p}{\partial x_i}(\mathbf{x})=\sum_{\alpha\in\{0,\ldots,K-1\}^n,\|\alpha\|_1\leq K-1}a_{\alpha+\mathbf{e}_i}\cdot(\alpha_i+1)\cdot\mathbf{x}^\alpha$$

这意味着对于任意 $\mathbf{x} \in [-1,1]^n$，我们有

$$\begin{aligned}\left|\frac{\partial p}{\partial x_i}(\mathbf{x})\right| &\le \sum_{\alpha\in\{0,\dots,K-1\}^n,\|\alpha\|_1\le K-1} |a_{\alpha+\mathbf{e}_i}\cdot(\alpha_i+1)\cdot\mathbf{x}^\alpha| \\ &\le K \sum_{\alpha\in\{0,\dots,K-1\}^n,\|\alpha\|_1\le K-1} |a_{\alpha+\mathbf{e}_i}| \\ &\le K\sqrt{(n+1)^{K-1}}\|p\|_{\mathrm{co}}\end{aligned}$$

因此，$\|\nabla p(\mathbf{x})\|_1 \le nK\sqrt{(n+1)^{K-1}}\|p\|_{\mathrm{co}} \le K\sqrt{(n+1)^{K+1}}\|p\|_{\mathrm{co}}$。证明 p 关于 $[-1,1]^n$ 上的$\|\cdot\|_\infty$范数是 $((n+1)^{\frac{K+1}{2}}K\|p\|_{\mathrm{co}})$-利普希茨的。同样，对于任意 $\mathbf{x}\in[-1,1]^n$，我们有

$$p(\mathbf{x}) \le \sum_{\alpha\in\{0,\dots,K\}^n,\|\alpha\|_1\le K} |a_\alpha| \le 2(n+1)^{K/2}\|p\|_{\mathrm{co}}$$

□

引理 8.4. 假设 $f\colon \mathcal{X}\to\{\pm1\}$ 是关于.....的 (K,M)-多项式阈值函数，由多项式 $p\colon \mathbb{R}^n\to\mathbb{R}$ 见证，且该多项式关于 $\|\cdot\|_\infty$ 是 L-利普希茨的。

- 如果 p 被 B 有界，则 $\cup_{\mathbf{x}\in\mathcal{X}}\mathcal{B}_{1/(2L)}(\mathbf{x})$。那么， f 是由 $2p$ 见证的 $(K,2M,2B,\frac{1}{2L})$-多项式阈值函数
- 如果 p由 B界定，则为 $\mathcal{X}$。那么， f是由 $2p$见证的 $(K、2M、2B\pm1、\frac{1}{2L})$-多项式阈值函数

证明。我们首先注意到，第二项由第一项推出。事实上，如果 p 在 $\mathcal{X}$ 中由 B 界定，那么 p 由 $B+1/2$ 界定，即为 $\cup_{\mathbf{x}\in\mathcal{X}}\mathcal{B}_{1/(2L)}(\mathbf{x})$。为了证明第一项，我们需要证明，对于任意 $\mathbf{x}\in\mathcal{X}$和$\tilde{}\mathbf{x}\in\mathcal{B}_\xi(\mathbf{x})$，都有

$$2B \ge 2p(\tilde{\mathbf{x}})f(\mathbf{x}) \ge 1$$

左侧不等式显然成立。对于右侧不等式，我们假设 $f(\mathbf{x})=1$（另一种情况类似）。由于 $\|\mathbf{x}-\tilde{}\mathbf{x}\|_\infty \le \frac{1}{2L}$，我们有 —

$$\begin{aligned}p(\tilde{\mathbf{x}}) &\ge p(\mathbf{x}) - |p(\tilde{\mathbf{x}})-p(\mathbf{x})| \\ &\ge p(\mathbf{x}) - L\cdot\|\mathbf{x}-\tilde{\mathbf{x}}\|_\infty \\ &\ge 1-\frac{L}{2L} \\ &= \frac{1}{2}\end{aligned}$$

□

8.1 每个标签都依赖于 $O(1)$ 个更简单的标签

现在假设 $\mathcal{X}\subseteq\{\pm1\}^d$，并且任意标签 $j\in L_i$ 至多依赖于来自 L_{i-1} K 的 K 个标签，且这些标签至多位于 K 个输入位置中的位置（如果 $i=1$）。也就是说，对于任意 $j\in\tilde{L_i}$，存在一个函数 $f_j\colon\{\pm1\}^{wn}\to\{\pm1\}$（或者如果 $i=1$，则存在$\tilde{f_j}\colon\{\pm1\}^{dw}\to\{\pm1\}$），它至多依赖于来自 $\{kn+l\colon 0\le k\le w-1, l\in L_{i-1}\}\tilde{}$（如果 $i=1$，则来自 $[dw]$）的 K 个坐标，并且满足以下条件。对于任意 $g\in G$, $f^*_{j,g}$（$\tilde{}\mathbf{x}$），都有 $=f_j$（E_g(f *($\tilde{}\mathbf{x}$))）（或者如果 $i=1$，则有 $f^*_{j,g}$（$\tilde{}\mathbf{x}$）$=f_j$（$E_g(\tilde{}\mathbf{x})$）。

如例3.2所示，由于任何依赖于 K 个变量的布尔函数都是一个 $(K,1)$-多项式阈值函数，因此$\tilde{}$函数 f_j 是 $(K,1)$- 多项式阈值函数，这意味着 $(\mathcal{L},\mathbf{e})$ 构成一个 $(r,K,1)$- 层级。引理 8.1 表明，$(\mathcal{L},\mathbf{e})$ 对于 $\xi = \frac{1}{2K\ wn+1\ ()^{(K+1)/2}}$ 和 $B=2(w\max(n,d)+1)^{K/2}$，构成一个 $(r,K,2,B,\xi)$- 层级。下面的引理表明，这一点可以得到大幅改进。

引理 8.5. 任何依赖于 K 个坐标的布尔函数都是一个$(K,2,3,\xi)$-多项式阈值函数，其中 $\xi=\frac{1}{K2^{(K+2)/2}\ o}$ 因此，$(\mathcal{L},\mathbf{e})$ 是一个$(r,K,2,3,\xi)$-层级。

引理 8.5 由以下引理与引理 8.4 推出

引理 8.6. 令 $f\colon \{\pm1\}^K \to \{\pm1\}$，并令 $F(\mathbf{x}) = \sum_{A\subseteq[K]} a_A \mathbf{x}^A$ 为其标准多线性延拓。则，F 关于 $[-1,1]^K$ 范数是 $(K2^{K/2})$-利普希茨的。

证明。对于 $\mathbf{x} \in [-1,1]^K$，我们有

$$\left|\frac{\partial F}{\partial x_i}\right| = \left|\sum_{i\in A\subseteq[K]} a_A \mathbf{x}^A\right| \le \sum_{i\in A\subseteq[K]} |a_A| \le \sum_{i\in A\subseteq[K]} |a_A|$$

因此，

$$\|\nabla F(\mathbf{x})\|_1 \le \sum_{A\subseteq[K]} |A|\,|a_A| \overset{\text{Cauchy Schwartz}}{\le} K2^{K/2}$$

□

下面的引理表明，可以进一步改进 ξ 和 B，代价是次数和系数范数。

引理 8.7。 对于任意 $0 < \xi < 1 < B$，任何依赖于 K 个坐标的布尔函数都是一个(K', M, B, ξ)-多项式阈值函数，其参数为 *or* $K' = O\left(\frac{K^{2+K}\log((B+1)/(B-1))}{1-\xi}\right)$()()和 $M = 2^{O\left(\frac{K^{2+K}}{\log((B+1)/(B-1))}\right)}$()()。因此，$(\mathcal{L}, \mathbf{e})$ 是 (r, K', M, B, ξ)- 层级

证明。固定 $f\colon \{\pm1\}^K \to \{\pm1\}$。我们需要证明 f 是一个（K', M, B, ξ）-多项式阈值函数。令 $\epsilon = \frac{B-1}{B+1}$。根据引理 7.5，存在一个次数为 $O\left(\frac{K+\log(1/\epsilon)}{1-\xi}\right)$ 的单变量多项式 q，使得对于任意 $y \in [-1,1] \setminus [-1+\xi, 1-\xi]$，我们有 $q([-1,1]) \subseteq [-1,1]$ 且 $|q(y) - \mathrm{sign}(y)| \le \frac{\epsilon}{K2^{K/2}}$，且 q 的系数均有界于 $2^{O(K+\log(1/\epsilon))(1-\xi)}$。现在考虑多项式 $\tilde{p}(\mathbf{x}) = F(q(\mathbf{x}))$，其中 F 是 f 上的多线性扩展。不难验证，$\deg(\tilde{p}) \le \deg(q)K = O\left(\frac{K^{2+K}\log(1/\epsilon)}{1-\xi}\right)$，且 $\|\tilde{p}\|_{\mathrm{co}} \le 2^{O\left(\frac{K^{2+K}}{1-\xi}\log(1/\epsilon)\right)}$。最后，固定 $\mathbf{x} \in \{\pm1\}^K$ 并 $\tilde{\mathbf{x}} \in \mathcal{B}_\xi(\mathbf{x})$。注意，$\mathbf{x} = \mathrm{sign}(\tilde{\mathbf{x}})$。由于相对于$[-1,1]^K$ 中的 $\|\cdot\|_\infty$ 范数，F 是 $K2^{k/2}$-利普希茨的（引理 8.6），我们有

$$|\tilde{p}(\tilde{\mathbf{x}}) - f(\mathbf{x})| = |\tilde{p}(\tilde{\mathbf{x}}) - f(\mathrm{sign}(\tilde{\mathbf{x}}))| = |F(q(\tilde{\mathbf{x}})) - F(\mathrm{sign}(\tilde{\mathbf{x}}))| \le \|q(\tilde{\mathbf{x}}) - \mathrm{sign}(\tilde{\mathbf{x}})\|_\infty \le \epsilon$$

由于 $f(\mathbf{x}) \in \{\pm1\}$，这意味着

$$1 + \epsilon \ge \tilde{p}(\tilde{\mathbf{x}}) f(\mathbf{x}) \ge 1 - \epsilon$$

取 $p(x) = \frac{1}{1-\epsilon}\tilde{p}(x)$，并注意到 $B = \frac{1+\epsilon}{1-\epsilon}$，我们得到

$$B \ge p(\tilde{\mathbf{x}}) f(\mathbf{x}) \ge 1$$

这意味着 f 是一个 (K', M, B, ξ)-多项式阈值函数。 □

8.2 定理 3.4 的证明

本节将证明定理 3.4（稍作扩展的版本）。我们先回顾并稍作扩展相关设定。固定一个定义域 $\mathcal{X} \subseteq \{\pm1\}^d$ 以及一个函数序列 $G^i : \{\pm1\}^d \to \{\pm1\}^d$ 其中 $1 \le i \le r$。我们假设 $G^0(\mathbf{x}) = \mathbf{x}$，并且对于任意深度 $i \in [r]$ 和坐标 $j \in [d]$，都有

$$\forall \mathbf{x} \in \mathcal{X}, \quad G^i_j(\mathbf{x}) = p^i_j(G^{i-1}(\mathbf{x})), \tag{20}$$

其中 $p^i_j\colon \{\pm1\}^d \to \{\pm1\}$ 是一个函数，其多重线性扩展是次数至多为 K 的多项式。此外，我们假设该扩展关于 $[-1,1]^d$，相对于 ℓ_∞ 范数是 L-利普希茨的（如果 p^i_j 依赖于 K 个坐标，如3.1节中的问题描述，那么引理 8.6 蕴含该性质成立，且$L = K2^{K/2}$）。固定一个整数 q。我们假设，对于每个深度 $i \in [r]$，都有 q 个辅助标签 $f^*_{i,\,j}$，对应于$1 \le j \le q$，其中每个都是由 G^i 的奇数个分量构成的带符号多数函数。此外，我们假设

这些函数是随机的。具体而言，在学习之前，标注者会独立地抽取 qr 个函数，使得对于任意 $i \in [r]$ 和 $j \in [q]$,

$$f^*_{i,j}(\mathbf{x}) = \operatorname{sign}\left(\sum_{l=1}^{d} w_l^{i,j} G_l^i(\mathbf{x})\right), \tag{21}$$

其中，权重向量 $\mathbf{w}^{i,\,j} \in \mathbb{R}^d$ 是从以下集合中独立均匀选取的向量

$$\mathcal{W}_{d,k} := \left\{\mathbf{w} \in \{-1,0,1\}^d : \sum_{l=1}^{d} |w_l| = k\right\}$$

对于某个奇整数 k。以下定理略微扩展了定理 3.4，它表明，如果 $q \gg dL^2 \log(|\mathcal{X}|$，那么对于 $\mathbf{f}^*$ 的选取，以高概率，目标函数 $\mathbf{f}^*$ 具有一个 $\left(r,\ K,\ O\left(kd^K\right),\ 2k+1\right)$-层级。

定理 **8.8.** 以概率 $1 - 4drq|\mathcal{X}|e^{-\Omega\left(\frac{q}{L^2k^2d}\right)}$，函数 $\mathbf{f}^*$ 具有 $\left(r, K, O\left(kd^K\right), 2k+1\right)$-层级

为了证明定理 8.8，只需证明对于任意 $i \in [r]$ 和 $j \in [q]$, $f^*_{i,\,j}$ 是一个 $\left(K,\ O\left(kd^K\right),\ 2k+1\right)$-多项式阈值函数，定义在

$$\Psi_{i-1}(\mathbf{x}) = \left(f^*_{i-1,1}(\mathbf{x}), \ldots, f^*_{i-1,q}(\mathbf{x})\right)$$

由方程（21）和（20）可得

$$f^*_{i,j}(\mathbf{x}) = \operatorname{sign}\left(\sum_{l=1}^{d} w_l^{i,j} p_l^i(G^{i-1}(\mathbf{x}))\right) =: \operatorname{sign}\left(q(G^{i-1}(\mathbf{x}))\right)$$

因此，$f^*_{i,j}$ 是 G^{i-1}的（K,k）-多项式阈值函数，其见证函数为 q（注意，由于 $q(G^{i-1}(\mathbf{x}))$是 $\{\pm 1\}$ 中的 k 个数之和且 k 为奇数，因此 $1 \le |q(G^{i-1}(\mathbf{x}))| \le k$。同样地，$\|q\|_{\mathrm{co}} \le \sum_{l=1}^{d} |w^{i,\,j}_{l}| \cdot \|p_l^i\|_{\mathrm{co}} \overset{\|p_l^i\|_{\mathrm{co}} \le 1}{\le} \sum_{l=1}^{d} |w^{i,\,j}_{l}| = k$）。由于 q 是 (kL)-利普希茨函数且有界于 k，引理 8.4表明，$f^*_{i,j}$ 是G^{i-1} 的（$K, k, 2k+1, 1/(2kL)$）-多项式阈值函数。因此，定理 8.8可由下述引理以及对 rq 个不同的$f^*_{i,j}$应用并联合界得到。

引理 8.9。 令 $f\colon\ \mathcal{X} \to \{\pm 1\}$ 为一个 (K, M, B, ξ)-多项式阈值函数，令 $\mathbf{w}^1$、……、$\mathbf{w}^q \in \mathcal{W}_{d,\,k}$ 相互独立且服从均匀分布。定义 $\psi_i(\mathbf{x})$ =符号函数$(\langle \mathbf{w}^i,\ \mathbf{x}\rangle)$。则，以概率 $1 - 4d|\mathcal{X}|e^{-\Omega\left(\frac{\xi^2 q}{d}\right)}$ f，$\left(K, O\left(Md^K\right), B\right)$-多项式阈值函数为$\Psi = (\psi_1,\ \ldots\ldots,\ \psi_q)$。

证明。令 $W = [\mathbf{w}_1 \cdots \mathbf{w}_q] \in M_{d,\,q}$我们首先证明，以高概率，$W$能够从 $\Psi(\mathbf{x})$近似重构$\mathbf{x}$

断言 2。 令 $\alpha_{d,\,k} = \frac{k}{d} \cdot \frac{k-1}{(k-1)/2} 2^{k-1}$。对于任意 $\mathbf{x} \in \{\pm 1\}^d$ 以及 $\frac{1}{4} \ge \epsilon > 0$，有 概率 $\left(\left\|\frac{1}{q\alpha_{d,\,k}} W\Psi(\mathbf{x}) - \mathbf{x}\right\|_\infty \ge \epsilon\right) \le 4de^{-\Omega\left(\frac{\epsilon^2 q}{d}\right)}$

在证明该命题之前，我们先说明它蕴含引理。事实上，它意味着以 $1-4d|\mathcal{X}|e^{-\Omega\left(\frac{\xi^2 q}{d}\right)}$ 的概率，我们有 $\left\|\frac{1}{q\alpha_{d,\,k}} W\Psi(\mathbf{x}) - \mathbf{x}\right\|_\infty \le \frac{\xi}{2}$，对任意 $\mathbf{x} \in \mathcal{X}$ 均成立。在此事件下，我们有

$$1 - \xi \le \frac{1 - \xi/2}{q\alpha_{d,k}} \left(W\Psi(\mathbf{x}) \odot \mathbf{x}\right)_j \le 1$$

对任意 $\mathbf{x} \in \mathcal{X}$ 和 $j \in [d]$ 均成立。因此，若 $p\colon\ \mathcal{X} \to \mathbb{R}$ 是一个多项式帽见证，且 f 是 $(K,\ M,\ B,\ \xi)$-多项式阈值函数，则我们有

$$B \ge p\left(\frac{1 - \xi/2}{q\alpha_{d,k}} W\Psi(\mathbf{x})\right) \cdot f(\mathbf{x}) \ge 1$$

因此，对于 $q(\mathbf{y}) := p\left(\frac{1-\xi/2}{q\alpha_{d,k}} W\mathbf{y}\right)$，我们有 f 是 Ψ 的 $(K, \|q\|_{\mathrm{co}}, B)$-多项式阈值函数。根据引理 7.6 以及每一行的范数至多为 $\frac{1}{\sqrt{q}\alpha_{d,k}}$ 的事实（因为 W 的元素都属于 $\{-1, 1, 0\}$），我们有

$$\|q\|_{\mathrm{co}} \le \|p\|_{\mathrm{co}} \cdot \left(\frac{\sqrt{q+1}}{\sqrt{q}\alpha_{d,k}}\right)^K$$

这就得出了该引理，因为 $\alpha_{d,k} = \Theta\left(\sqrt{\frac{k}{d}}\right)$ 根据引理 7.4。

证明。（命题 2）固定一个坐标 $j \in [d]$。只需证明 $\Pr\left(\left|\frac{1}{q\alpha_{d,k}} \left(W\Psi\left(\mathbf{x}\right)\right)_j - x_j\right| \geq \epsilon\right) \leq 4e^{-\Omega\left(\epsilon^2 q / d\right)}$。注意到

$$\frac{1}{q\alpha_{d,k}} \left(W\Psi(\mathbf{x})\right)_j = \frac{1}{q} \sum_{i=1}^{q} \frac{w_j^i \text{sign}(\langle \mathbf{w}^i, \mathbf{x} \rangle)}{\alpha_{d,k}}$$

记 $X_i = w_j^i$符号函数$(\langle \mathbf{w}^i, \mathbf{x} \rangle)$。注意到 $X_1, \ldots, X_q$是独立同分布的。我们有

$$\begin{aligned}
\Pr(X_i = x_j) &= \frac{k}{2d}\left[\Pr(\text{sign}(\langle \mathbf{w}^i, \mathbf{x} \rangle) = 1 | w_j = x_j) + \Pr(\text{sign}(\langle \mathbf{w}^i, \mathbf{x} \rangle) = -1 | w_j = -x_j)\right] \\
&= \frac{k}{2d2^{k-1}}\left[\binom{k-1}{\geq (k-1)/2} + \binom{k-1}{\geq (k-1)/2}\right] \\
&= \frac{k}{2d}\left[1 + \frac{\binom{k-1}{(k-1)/2}}{2^{k-1}}\right]
\end{aligned}$$

类似地,

$$\begin{aligned}
\Pr(X_i = -x_j) &= \frac{k}{2d}\left[\Pr(\text{sign}(\langle \mathbf{w}^i, \mathbf{x} \rangle) - 1 | w_j = x_j) + \Pr(\text{sign}(\langle \mathbf{w}^i, \mathbf{x} \rangle) = 1 | w_j = -x_j)\right] \\
&= \frac{k}{2d2^{k-1}}\left[\binom{k-1}{> (k-1)/2} + \binom{k-1}{> (k-1)/2}\right] \\
&= \frac{k}{2d}\left[1 - \frac{\binom{k-1}{(k-1)/2}}{2^{k-1}}\right]
\end{aligned}$$

因此

$$\mathbb{E}X_i = \left(\Pr(X_i = x_j) - \Pr(X_i = -x_j)\right) x_j = \alpha_{d,k} \cdot x_j$$

And,

$$\Pr(X_i \neq 0) = \Pr(X_i = x_j) + \Pr(X_i = -x_j) = \frac{k}{d}$$

这意味着

$$\frac{\min\left(\Pr(X_i = 1), \Pr(X_i = -1)\right)}{|\mathbb{E}X_i|} = \frac{k}{2d\alpha_{d,k}}\left[1 - \frac{\binom{k-1}{(k-1)/2}}{2^{k-1}}\right] \geq \frac{k}{\alpha_{d,k} 4d} \geq \frac{1}{2}$$

并且

$$\frac{|\mathbb{E}X_i|^2}{\Pr(\text{sign}(\langle \mathbf{w}, \mathbf{x} \rangle) w_i \neq 0)} = \frac{k}{d}\left(\frac{\binom{k-1}{(k-1)/2}}{2^{k-1}}\right)^2 \overset{\text{Lemma 7.4}}{=} \Theta\left(\frac{1}{d}\right)$$

根据引理 7.3可得

$$\Pr\left(\left|\frac{1}{q\alpha_{d,k}} \left(W\Psi(\mathbf{x})\right)_j - x_j\right| \geq \epsilon\right) \leq 4e^{-\Omega\left(\frac{\epsilon^2 q}{d}\right)}$$

□

□

9 Ker来自随机神经元的特征及引理的证明5.3

取一个有界激活函数 σ： $\mathbb{R} \to \mathbb{R}$。给定 $0 \le \beta \le 1$，称为偏置幅度，我们在 $\mathbb{R}^n$上定义一个核:

$$k_{\sigma,\beta,n}(\mathbf{x},\mathbf{y}) = \mathbb{E}[\sigma(\mathbf{w}^\top \mathbf{x} + b)\sigma(\mathbf{w}^\top \mathbf{y} + b)]\,, \quad b \sim \mathcal{N}(0,\beta^2),\ \mathbf{w} \sim \mathcal{N}\left(0, \frac{1-\beta^2}{n} I_n\right) \tag{22}$$

注意， $\psi(\mathbf{w}\ b),\ \mathbf{x}) = \sigma(\mathbf{w}^\top \mathbf{x} + b)$ 是 $k_{\sigma,\ \beta,\ n}$ 的随机特征映射。接下来，我们分析相应核空间 $\mathcal{H}_{\sigma,\ \beta,\ n}$ 中的函数。为此，我们将使用 σ 的厄米特展开，以求得 $k_{\sigma,\ \beta,\ n}$ 的显式表达式，以及一个显式嵌入 $\Psi_{\sigma,\ \beta,\ n}$：
$\mathbb{R}^n \to \bigoplus_{s=0}^{\infty} \left(\mathbb{R}^{n+1}\right)^{\otimes s}$，其核为 $k_{\sigma,\ \beta,\ n}$。令

$$\sigma = \sum_{s=0}^{\infty} a_s h_s \tag{23}$$

为 σ 的厄米特展开。对于 $r \ge 1$，记

$$a_s(r) = \sum_{j=0}^{\infty} a_{s+2j} \sqrt{\frac{(s+2j)!}{s!}} \frac{(r^2-1)^j}{j!2^j} \tag{24}$$

注意， $a_s\mathbf{(1)} = a_s$

引理 9.1。我们有

$$k_{\sigma,\beta,n}(\mathbf{x},\mathbf{y}) = \sum_{s=0}^{\infty} a_s\left(\sqrt{\frac{1-\beta^2}{n}\|\mathbf{x}\|^2 + \beta^2}\right) a_s\left(\sqrt{\frac{1-\beta^2}{n}\|\mathbf{y}\|^2 + \beta^2}\right)\left(\frac{1-\beta^2}{n}\langle \mathbf{x},\mathbf{y}\rangle + \beta^2\right)^s$$

同样，$k_{\sigma,\ \beta,\ n}$ 是嵌入 $\Psi_{\sigma,\ \beta,\ n}$： $\mathbb{R}^n \to \bigoplus_{s=0}^{\infty} \left(\mathbb{R}^{n+1}\right)^{\otimes s}$ 的核，其定义为

$$\Psi_{\sigma,\beta,n}(\mathbf{x}) = \left(a_s\left(\sqrt{\frac{1-\beta^2}{n}\|\mathbf{x}\|^2 + \beta^2}\right) \cdot \begin{bmatrix} \sqrt{\frac{1-\beta^2}{n}}\mathbf{x} \\ \beta \end{bmatrix}^{\otimes s}\right)_{s=0}^{\infty}$$

为证明引理9.1 我们将使用以下引理。

引理 9.2. 我们有 $h_s(ax) = \sum_{j=0}^{\lfloor s/2 \rfloor} \sqrt{\frac{s!}{(s-2j)!}} \frac{a^{s-2j}\left(a^2-1\right)^j}{j!2^j} h_{s-2j}(x)$

证明。 根据公式（4）可得

$$\begin{aligned}
\sum_{s=0}^{\infty} \frac{h_s(ax)t^s}{\sqrt{s!}} &= e^{xat - \frac{t^2}{2}} \\
&= e^{xat - \frac{(at)^2}{2} + \frac{(at)^2}{2} - \frac{t^2}{2}} \\
&\overset{\text{Eq. (4)}}{=} e^{\frac{(at)^2}{2} - \frac{t^2}{2}} \left(\sum_{s=0}^{\infty} \frac{h_s(x)a^s t^s}{\sqrt{s!}}\right) \\
&= e^{(a^2-1)\frac{t^2}{2}} \left(\sum_{s=0}^{\infty} \frac{h_s(x)a^s t^s}{\sqrt{s!}}\right) \\
&= \left(\sum_{s=0}^{\infty} \frac{(a^2-1)^s}{s!2^s} t^{2s}\right)\left(\sum_{s=0}^{\infty} \frac{h_s(x)a^s}{\sqrt{s!}} t^s\right) \\
&= \sum_{s=0}^{\infty} \left(\sum_{j=0}^{\lfloor \frac{s}{2} \rfloor} \frac{(a^2-1)^j}{j!2^j} \frac{h_{s-2j}(x)a^{s-2j}}{\sqrt{(s-2j)!}}\right) t^s
\end{aligned}$$

因此，

$$\frac{h_s(ax)}{\sqrt{s!}} = \sum_{j=0}^{\lfloor \frac{s}{2} \rfloor} \frac{(a^2-1)^j}{j!2^j} \frac{a^{s-2j}}{\sqrt{(s-2j)!}} h_{s-2j}(x)$$

□

证明。（引理 9.1 的证明）我们将证明 $k_{\sigma,\beta,n}$ 的公式。不难验证，根据事实 $\langle \mathbf{x}^{\otimes s}, \mathbf{y}^{\otimes s} \rangle = \langle \mathbf{x}, \mathbf{y} \rangle^s$，它意味着 $k_{\sigma,\beta,n}$ 是 $\Psi_{\sigma,\beta,n}$ 的核。根据定义，$k_{\sigma,\beta,n}(\mathbf{x}, \mathbf{y}) = \mathbb{E}[\sigma(\mathbf{w}^\top \mathbf{x} + b)\sigma(\mathbf{w}^\top \mathbf{y} + b)]$，其中 $b \sim \mathcal{N}(0, \beta^2)$ 且 $\mathbf{w} \sim \mathcal{N}\left(0, \frac{1-\beta^2}{n} I_n\right)$。令 $X = \mathbf{w}^\top \mathbf{x} + b$，且 $Y = \mathbf{w}^\top \mathbf{y} + b$。注意，$(X, Y)$ 是一个中心化高斯向量，其相关矩阵为 $\begin{pmatrix} \frac{1-\beta^2}{n}\|\mathbf{x}\|^2 + \beta^2 & \frac{1-\beta^2}{n}\langle \mathbf{x}, \mathbf{y} \rangle + \beta^2 \\ \frac{1-\beta^2}{n}\langle \mathbf{x}, \mathbf{y} \rangle + \beta^2 & \frac{1-\beta^2}{n}\|\mathbf{y}\|^2 + \beta^2 \end{pmatrix}$。记 $r_{\mathbf{x}} = \sqrt{\frac{1-\beta^2}{n}\|\mathbf{x}\|^2 + \beta^2}$ 和 $r_{\mathbf{y}} = \sqrt{\frac{1-\beta^2}{n}\|\mathbf{y}\|^2 + \beta^2}$。类似地，令 $\tilde{X} = \frac{1}{r_{\mathbf{x}}} X$ 和 $\tilde{Y} = \frac{1}{r_{\mathbf{y}}} Y$。注意，对于 $\rho = \frac{\frac{1-\beta^2}{n}\langle \mathbf{x}, \mathbf{y} \rangle + \beta^2}{r_{\mathbf{x}} r_{\mathbf{y}}}$，$(X, Y)$ 是一个中心化高斯向量，其相关矩阵为 $\begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$。现在，根据引理 9.2，我们有

$$\begin{aligned}
\sigma(rx) &= \sum_{s=0}^{\infty} h_s(rx) \\
&= \sum_{s=0}^{\infty} \left(\sum_{j=0}^{\infty} a_{s+2j} \sqrt{\frac{(s+2j)!}{s!}} \frac{(r^2-1)^j}{j!2^j} \right) r^s h_s(x) \\
&= : \sum_{s=0}^{\infty} a_s(r) r^s h_s(x)
\end{aligned}$$

因此，

$$\begin{aligned}
k_{\sigma,\beta,n}(\mathbf{x}, \mathbf{y}) &= \mathbb{E}\sigma(r_{\mathbf{x}}\tilde{X})\sigma(r_{\mathbf{y}}\tilde{Y}) \\
&= \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} a_i(r_{\mathbf{x}}) r_{\mathbf{x}}^i a_j(r_{\mathbf{y}}) r_{\mathbf{x}}^j \mathbb{E} h_i(\tilde{X}) h_j(\tilde{Y}) \\
&\overset{\text{Eq. (6)}}{=} \sum_{s=0}^{\infty} a_s(r_{\mathbf{x}}) r_{\mathbf{x}}^s a_s(r_{\mathbf{y}}) r_{\mathbf{y}}^s \rho^s \\
&= \sum_{s=0}^{\infty} a_s(r_{\mathbf{x}}) a_s(r_{\mathbf{y}}) \left(\frac{1-\beta^2}{n} \langle \mathbf{x}, \mathbf{y} \rangle + \beta^2 \right)^s
\end{aligned}$$

□

引理 9.3。设 $r > 0$ 满足 $|1 - r^2| =: \epsilon < \frac{1}{2}$。我们有

$$|a_s(r) - a_s(1)| \le \|\sigma\| 2^{(s+2)/2} \frac{\epsilon}{\sqrt{1-2\epsilon^2}}$$

证明。我们有

$$\begin{aligned}
|a_s(r)-a_s(1)| &= \left|\sum_{j=1}^{\infty} a_{s+2j}\sqrt{\frac{(s+2j)!}{s!}}\frac{(r^2-1)^j}{j!2^j}\right| \\
&\overset{\text{Cauchy-Schwartz and } \|\sigma\|=\sqrt{\sum_{i=0}^{\infty}a_i^2}}{\leq} \|\sigma\|\sqrt{\sum_{j=1}^{\infty}\frac{(s+2j)!}{s!}\frac{(r^2-1)^{2j}}{(j!)^2 2^{2j}}} \\
&\overset{(2j)!\leq (j!2^j)^2}{\leq} \|\sigma\|\sqrt{\sum_{j=1}^{\infty}\frac{(s+2j)!}{s!(2j)!}(r^2-1)^{2j}} \\
&= \|\sigma\|\sqrt{\sum_{j=1}^{\infty}\binom{s+2j}{s}(r^2-1)^{2j}} \\
&\leq \|\sigma\|\sqrt{\sum_{j=1}^{\infty}2^{s+2j}(r^2-1)^{2j}} \\
&= \|\sigma\|2^{s/2}\sqrt{\sum_{j=1}^{\infty}(2r^2-2)^{2j}} \\
&= \|\sigma\|2^{s/2}|2r^2-2|\frac{1}{\sqrt{1-(2r^2-2)^2}}
\end{aligned}$$

□

引理 9.4。假设 $1-\beta^2<\frac{1}{2}$ 对于 $\beta>0$。设 $\mathcal{X}\subseteq[-1,1]^n$。设 $p:\ \mathcal{X}\to\mathbb{R}$ 是一个次数为 K 的多项式。设 $K'\geq K$。存在 $g\in\mathcal{H}_{\sigma,\,\beta,\,n}\ (\mathcal{X})$ 使得

1. $g(\mathbf{x})=\frac{a_{K'}\left(\sqrt{\frac{1-\beta^2}{n}\|\mathbf{x}\|^2+\beta^2}\right)}{a_{K'}}p(\mathbf{x})$
2. $\|g\|_{\sigma,\beta,n}\leq\frac{1}{a_{K'}\beta^{K'-K}}\left(\frac{n}{1-\beta^2}\right)^{K/2}\|p\|_{\text{co}}$
3. $\|g-p\|_{\infty}\leq\|p\|_{\infty}\frac{\|\sigma\|}{a_{K'}}2^{(K'+2)/2}\frac{1-\beta^2}{\sqrt{1-2(1-\beta^2)^2}}$

证明。设 $p(\mathbf{x})=\sum_{\alpha\in\{0,\ldots,K\}^n,\ \|\alpha\|_1\leq K}b_\alpha\mathbf{x}^\alpha$。对于 $\alpha\in\{0,\ldots,K\}^n,\ \|\alpha\|_1\leq K$ 我们令 $\tilde{\alpha}\in[n+1]^{K'}$ 为一个序列，使得对于任意 $i\in[n]$，恰有 α_i 个指标 $j\in[K']$ 满足 $\tilde{\alpha}_j=i$，其余 $K'-\|\alpha\|_1$ 个指标 $\tilde{\alpha}_j=n+1$ 满足 $A\in(\mathbb{R}^{n+1})^{\otimes K'}\subseteq\bigoplus_{s=0}^{\infty}(\mathbb{R}^{n+1})^{\otimes s}$ 为张量

$$A_\gamma=\begin{cases}\frac{1}{a_{K'}\beta^{K'-\|\alpha\|_1}}\left(\frac{n}{1-\beta^2}\right)^{\|\alpha\|_1/2}b_\alpha & \gamma=\tilde{\alpha}\text{ for some }\alpha\\ 0 & \text{otherwise}\end{cases}$$

并令

$$g(\mathbf{x})=\langle A,\Psi_{\sigma,\beta,n}(\mathbf{x})\rangle$$

不难验证，$g(\mathbf{x})=\frac{a_{K'}\left(\sqrt{\frac{1-\beta^2}{n}\|\mathbf{x}\|^2+\beta^2}\right)}{a_{K'}}p(\mathbf{x})$。由定理 7。9 $g\in\mathcal{H}_{\sigma,\,\beta,\,n}$ 并且满足$\|g\|_{\sigma,\,\beta,\,n}\leq\|A\|$。最后，由于 $\frac{1}{\beta^{K'-\|\alpha\|_1}}\left(\frac{n}{1-\beta^2}\right)^{\|\alpha\|_1/2}\leq\frac{1}{\beta^{K'-K}}\left(\frac{n}{1-\beta^2}\right)^{K/2}$ 我们有$\|A\|\leq\frac{1}{a_{K'}\beta^{K'-K}}\left(\frac{n}{1-\beta^2}\right)^{K/2}\|p\|_{\text{co}}$。因此，我们证明了第一项和第二项。为了证明最后一项，我们注意到，对于任意 $\mathbf{x}\in\mathcal{X}$，我们

have

$$\begin{aligned}|g(\mathbf{x})-p(\mathbf{x})| &= |p(\mathbf{x})|\cdot\left|\frac{a_{K'}\left(\sqrt{\frac{1-\beta^2}{n}\|\mathbf{x}\|^2+\beta^2}\right)}{a_{K'}}-1\right| \\ &= \frac{|p(\mathbf{x})|}{a_{K'}}\left|a_{K'}\left(\sqrt{\frac{1-\beta^2}{n}\|\mathbf{x}\|^2+\beta^2}\right)-a_{K'}\right|\end{aligned}$$

定义 $r=\sqrt{\frac{\|\mathbf{x}\|^2}{n}(1-\beta^2)+\beta^2}$，并注意到由于 $0\le\|\mathbf{x}\|^2\le n$，我们有

$$\beta^2\le r^2\le 1\Rightarrow\epsilon:=|1-r^2|\le 1-\beta^2<\frac{1}{2}$$

因此，根据引理9.3 我们有

$$|g(\mathbf{x})-p(\mathbf{x})|\le\frac{|p(\mathbf{x})|}{a_{K'}}\|\sigma\|_2^{(K'+2)/2}\frac{1-\beta^2}{\sqrt{1-2(1-\beta^2)^2}}$$

这证明了最后一项 □

结合引理 9.4 和引理 7.11 可得

引理 9.5。 假设 $1-\beta^2<\frac{1}{2}$ *for* $\beta>0$。令 $\mathcal{X}\subset[-1,1]^n$。固定一个次数为 K 的多项式 $p:\ \mathcal{X}\to[-1,1]$ *and* $K'\ge K$。令 $(W,\mathbf{b})\in\mathbb{R}^{q\times n}\times\mathbb{R}^q$ 为 β-*Xavier* 对。则存在一个向量 $\mathbf{w}=\mathbf{w}(W,\mathbf{b})\in\mathbb{R}^q$ 使得

$$\forall\mathbf{x}\in\mathcal{X},\ \Pr\left(|\langle\mathbf{w},\sigma(W\mathbf{x}+\mathbf{b})\rangle-p(\mathbf{x})|\ge\epsilon+\frac{\|\sigma\|}{a_{K'}}2^{(K'+2)/2}\frac{1-\beta^2}{\sqrt{1-2(1-\beta^2)^2}}\right)\le\delta$$

for

$$\delta=2\exp\left(-q\cdot\frac{a_{K'}^4\beta^{4K'-4K}(1-\beta^2)^{2K}\epsilon^4}{32n^{2K}\|p\|_{\text{co}}^4\|\sigma\|_\infty^4}\right)$$

此外

$$\|\mathbf{w}\|\le\frac{2\|\sigma\|_\infty}{\epsilon\sqrt{q}}\cdot\frac{1}{a_{K'}^2\beta^{2K'-2K}}\left(\frac{n}{1-\beta^2}\right)^K\|p\|_{\text{co}}^2$$

接下来，我们针对本文的需要对引理 9.5 作具体化，并说明它如何推出引理 5.3。回顾一下，对于 $\epsilon>0$，我们将 $\frac{3}{4}\le\beta_{\sigma,K',K}(\epsilon)<1$ 定义为满足以下条件的最小数：若 $\beta_{\sigma,K',K}(\epsilon)\le\beta<1$，则

$$\frac{\|\sigma\|}{a_{K'}}2^{(K'+2)/2}\frac{1-\beta^2}{\sqrt{1-2(1-\beta^2)^2}}\le\frac{\epsilon}{2}$$

我们还定义了

$$\delta_{\sigma,K',K}(\epsilon,\beta,q,M,n)=\begin{cases}1 & \frac{4\|\sigma\|_\infty}{\epsilon\sqrt{q}}\cdot\frac{1}{a_{K'}^2\beta^{2K'-2K}}\left(\frac{n}{1-\beta^2}\right)^K M^2>1\\ 2\exp\left(-q\cdot\frac{a_{K'}^4\beta^{4K'-4K}(1-\beta^2)^{2K}\epsilon^4}{512n^{2K}M^4\|\sigma\|_\infty^4}\right) & \text{otherwise}\end{cases}$$

现在我们可以证明引理5.3 ， 下面将重新表述该引理。

引理 9.6. （引理 5.3重述）固定 $\mathcal{X}\subset[-1,1]^n$，一个次数为 K 的多项式 $p:\ \mathcal{X}\to[-1,1]$、$K'\ge K$和 $\epsilon>0$。设 $(W,\mathbf{b})\in\mathbb{R}^{q\times n}\times\mathbb{R}^q$ 是 β-*Xavier* 对，对应于 $1>\beta\ge\beta_{\sigma,K',K}(\epsilon)$。那么存在一个向量$\mathbf{w}=\mathbf{w}(W,\mathbf{b})\in\mathbb{B}^q$，使得

$$\forall\mathbf{x}\in\mathcal{X},\ \Pr\left(|\langle\mathbf{w},\sigma(W\mathbf{x}+\mathbf{b})\rangle-p(\mathbf{x})|\ge\epsilon\right)\le\delta_{\sigma,K',K}(\epsilon,\beta,q,\|p\|_{\text{co}},n)$$

证明。 固定 $\mathbf{x} \in \mathcal{X}$。由引理 9.5 可知，存在一个向量 $\mathbf{v} \in \mathbb{R}^q$ 使得

$$\Pr\left(|\langle \mathbf{v}, \sigma(W\mathbf{x}+\mathbf{b})\rangle - p(\mathbf{x})| \geq \epsilon\right) \leq \Pr\left(|\langle \mathbf{v}, \sigma(W\mathbf{x}+\mathbf{b})\rangle - p(\mathbf{x})| \geq \frac{\epsilon}{2} + \frac{\|\sigma\|}{a_{K'}} 2^{(K'+2)/2} \frac{1-\beta^2}{\sqrt{1-2(1-\beta^2)^2}}\right) \leq \delta \quad (25)$$

for

$$\delta = 2\exp\left(-q \cdot \frac{a_{K'}^4 \beta^{4K'-4K}(1-\beta^2)^{2K}\epsilon^4}{512 n^{2K}\|p\|_{\mathrm{co}}^4 \|\sigma\|_\infty^4}\right)$$

此外

$$\|\mathbf{v}\| \leq \frac{4\|\sigma\|_\infty}{\epsilon\sqrt{q}} \cdot \frac{1}{a_{K'}^2 \beta^{2K'-2K}} \left(\frac{n}{1-\beta^2}\right)^K \|p\|_{\mathrm{co}}^2$$

定义$\mathbf{w}$ 为$\mathbf{v}$ 在 $\mathbb{B}^d$上的投影。现在分情况讨论。若 $\frac{4\|\sigma\|_\infty}{\epsilon\sqrt{q}} \cdot \frac{1}{a_{K'}^2 \beta^{2K'-2K}} \left(\frac{n}{1-\beta^2}\right)^K \|p\|_{\mathrm{co}}^2 \leq 1$则$\mathbf{v} = \mathbf{w}$ 且$\delta = \delta_{\sigma, K', K}(\epsilon, \beta, q, \|p\|_{\mathrm{co}}, n)$，因此由方程（25）可得引理。否则，我们有$\delta_{\sigma, K', K}(\epsilon, \beta, q, \|p\|_{\mathrm{co}}, n) = 1$，引理显然成立。 □

致谢

本文所述研究由欧洲研究委员会（ERC）通过欧盟“地平线2022年”研究与创新计划（资助协议编号：101041711）以及西蒙斯基金会（作为深度学习数学与科学基础合作项目的一部分）资助。作者感谢 Elchanan Mossel 和 Mariano Schain 提供有益的意见。

参考文献

[1] Emmanuel Abbe、Enric Boix-Adsera、Matthew S Brennan、Guy Bresler 和 Dheeraj Nagaraj。《阶梯性质：层次结构如何指导深度学习》。《神经信息处理系统进展》，34：26989–27002，2021年。4，15[2] Zeyuan Allen-Zhu 和 Yuanzhi Li。《反向特征校正：深度学习如何执行深度学习》。《arXiv预印本》arXiv:2001.04413，2020年。4[3] Zeyuan Allen-Zhu、Yuanzhi Li 和 Yingyu Liang。《过参数化神经网络中的学习与泛化：超越两层网络》。《神经信息处理系统进展》，32，2019年。4[4] A. Andoni、R. Panigrahy、G. Valiant 和 L. Zhang。《用神经网络学习多项式》。《第31届国际机器学习会议论文集》，第1908–1916页，2014年。4[5] George E. Andrews、Richard Askey 和 Ranjan Roy。《特殊函数》，《数学及其应用百科全书》第71卷。剑桥大学出版社，剑桥，1999年。5[6] Alberto Bietti、Joan Bruna、Clayton Sanford 和 Min Jae Song。《用浅层神经网络学习单指标模型》。《神经信息处理系统进展》，35：9768–9783，2022年。4[7] Joan Bruna 和 St´ephane Mallat。《不变散射卷积网络》。《IEEE模式分析与机器智能汇刊》，35(8)：1872–1886，2013年。4[8] Joan Bruna、Loucas Pillaud-Vivien 和 Aaron Zweig。《高斯数据之外的单指标模型》。《arXiv预印本》arXiv:2307.15804，2023年。4[9] Yuan Cao 和 Quanquan Gu。《随机梯度下降在宽而深的神经网络上的泛化界》。《arXiv预印本》arXiv:1905.13210，2019年。4[10] Elisabetta Cornacchia 和 Elchanan Mossel。《奇偶函数课程学习的数学模型》。《国际机器学习会议》，第6402–6423页。机器学习研究论文集，2023年。4

[11] Amit Daniely。《随机梯度下降学习网络的共轭核类》。《神经信息处理系统进展》，第2422–2430页，2017年。4[12] Amit Daniely。《几乎无需过参数化的神经网络学习与记忆》。《神经信息处理系统大会》，2020年。4[13] Amit Daniely。《通过对神经网络进行梯度下降，在无需过参数化的情况下记忆高斯分布》。《arXiv预印本》arXiv:2003.12895，2020年。4[14] Amit Daniely 和 Eran Malach。《用神经网络学习奇偶函数》。《神经信息处理系统进展》，33：20356–20365，2020年。4[15] Amit Daniely、Roy Frostig 和 Yoram Singer。《深入理解神经网络：初始化的力量与表达能力的对偶视角》。《神经信息处理系统进展》，29，2016年。4[16] Amit Daniely、Mariano Schain 和 Gilad Yehudai。《Redex：通过凸优化超越固定表示方法》。《国际算法学习理论会议》，第518–543页。机器学习研究论文集，2024年。4[17] Ilias Diakonikolas、Parikshit Gopalan、Ragesh Jaiswal、Rocco A Servedio 和 Emanuele Viola。《有界独立性可以骗过半空间》。《SIAM计算杂志》，39(8)：3441–3462，2010年。17[18] Simon S Du、Jason D Lee、Yuandong Tian、Barnabas Poczos 和 Aarti Singh。《梯度下降能够学习单隐层卷积神经网络：不必害怕伪局部极小值》。《arXiv预印本》arXiv:1712.00779，2017年。4[19] Spencer Frei、Yuan Cao 和 Quanquan Gu。《用梯度下降无先验地学习单个神经元》。《神经信息处理系统进展》，33：5417–5428，2020年。4[20] Rong Ge、Jason D Lee 和 Tengyu Ma。《通过景观设计学习单隐层神经网络》。《arXiv预印本》arXiv:1711.00501，2017年。4[21] Han Huang 和 Elchanan Mossel。《树上广播问题的最优低次数困难性》。《arXiv预印本》arXiv:2502.04861，2025年。URL https://arxiv.org/abs/2502.04861。4[22]Frederic Koehler 和 Elchanan Mossel。《树上的重构与低次数多项式》。《神经信息处理系统进展》，第35卷，第18942–18954页，2022年。URL https://proceedings.neurips.cc/paper_files/paper/2022/file/78107952a1380965e6344d68239257e8-Paper-Conference.pdf。4[23] Rupert Li 和 Elchanan Mossel。《层次函数的噪声敏感性与学习下界》。《arXiv预印本》arXiv:2502.05073，2025年。URL https://arxiv.org/abs/2502.05073。4[24] Yuanzhi Li 和 Yingyu Liang。《通过对结构化数据进行随机梯度下降来学习过参数化神经网络》。《神经信息处理系统进展》，31，2018年。4[25] Roi Livni、Shai Shalev-Shwartz 和 Ohad Shamir。《神经网络训练的计算效率》。《神经信息处理系统进展》，第855–863页，2014年。4[26] Hrushikesh Mhaskar、Qianli Liao 和 Tomaso Poggio。《深度网络何时以及为何优于浅层网络？》。《AAAI人工智能会议论文集》，第31卷，2017年。URL https://ojs.aaai.org/index.php/AAAI/article/view/10913。4[27] Elchanan Mossel。《深度学习与层次生成模型》。《arXiv预印本》arXiv:1612.09057，2016年。URL https://arxiv.org/abs/1612.09057。4[28] Ankit B Patel、Tan Nguyen 和 Richard G Baraniuk。《深度学习的概率框架》。《神经信息处理系统进展》，第29卷，2016年。URL https://proceedings.neurips.cc/paper/2016/file/b341e2bfc451b73927ee1b5f270d216b-Paper.pdf。4

[29] A. Rahimi 和 B. Recht。《大规模核机器的随机特征》。载于 *NIPS*，2007年。 4[30] BernhardSch¨ olkopf 和 Alexander J Smola。《核学习：支持向量机、正则化、优化及其他》。麻省理工学院出版社，马萨诸塞州剑桥，2002年。ISBN 978-0262194754。18[31] Alexander A Sherstov。算法多项式。载于 第*50*届 *ACM SIGACT*计算理论研讨会论文集，第311–324页，2018年。 17[32] Yuandong Tian。双层ReLU网络总体梯度的解析公式及其在收敛和临界点分析中的应用。载于第*34*届国际机器学习会议论文集（第*70*卷），第3404–3413页。JMLR.org，2017年。 4[33] Gal Vardi、Gilad Yehudai 和 Ohad Shamir。使用梯度下降学习带偏置的单个神经元。神经信息处理系统进展，34：28690–28700，2021年。 4[34] Zixuan Wang、Eshaan Nichani、Alberto Bietti、Alex Damian、Daniel Hsu、Jason D Lee 和 Denny Wu。利用Transformer从由易到难数据中学习复合函数。 *arXiv*预印本 *arXiv:2505.23683*，2025年。 4[35] Gilad Yehudai 和 Ohad Shamir。使用梯度方法学习单个神经元。 *arXiv*预印本 *arXiv:2001.05205*，2020年。 4