arXiv:2303.02470v1 [math.ST] 4 Mar 2023

使用深度神经网络实现极小极大最优高维分类

Shuoyang Wang[a]、Zuofeng Shang[b,*]

[a]*Department Mathematics and Statistics, Auburn University*
[b]*Department of Mathematical Sciences, New Jersey Institute of Technology*

摘要
高维分类是高维数据分析中的一个基础性重要研究问题。本文推导了在特征维度随样本量呈指数级增长、且贝叶斯分类器具有复杂模块化结构时，极小极大超额误分类风险的非渐近收敛率。我们还证明，基于深度神经网络的分类器能够达到上述收敛率，因此是极小极大最优的。

关键词：高维分类、深度神经网络、极小极大超额误分类风险、模块化结构 2020年数学主题分类：62H30，62G99

1. 引言

高维设置下的分类是一项基础问题，在疾病分类、文档分类、图像识别等领域有着广泛应用；参见 [5, 6, 10], 及其中的参考文献。一个核心统计任务是在高维设置下构造极小极大最优分类器，并特别关注维数对极小极大超额误分类风险的影响。当 d 固定时，关于构造极小极大最优分类器已有丰富的研究文献。例如， [18, 23, 24, 13, 8, 7, 19, 11, 12]。在高维情形 $\log d = O(\mathrm{poly}(n))$ 中， n 表示训练样本数量，这一问题更具挑战性，目前的研究成果仅适用于特定的建模设定。例如，当总体密度为高斯分布且高斯均值向量之差是稀疏的时， [4, 3] 证明了判别分析方法是极小极大最优的。当数据是高维非高斯时，如何设计最优分类器仍不清楚；本文试图弥补这一空白。

我们考虑二元分类问题。令 Y 为由 $\{-1, 1\}$ 生成的二元变量，其先验概率 $\pi_p = \mathbb{P}(Y = 1)$、 $\pi_q = \mathbb{P}(Y = -1)$ 和 $\pi_p + \pi_q = 1$未知。令 $\boldsymbol{X} \in \mathcal{X} \equiv [0, 1]^d$ 为一个 d维随机特征向量，其类别标签 Y 满足 $\boldsymbol{X}|Y = 1 \sim p, \boldsymbol{X}|Y = -1 \sim q$，其中$p, q$是定义在 $\mathcal{X}$上的未知概率密度。令 $\theta = (p, q, \pi_p, \pi_q)$ 和 $\eta_\theta(\boldsymbol{x}) = \mathbb{P}_\theta(Y = 1|\boldsymbol{X} = \boldsymbol{x})$ $x \in \mathcal{X}$，后者是在 θ下将 $\boldsymbol{X} = \boldsymbol{x}$ 赋予标签 1 的条件概率。于是，最优贝叶斯分类器定义为：若 $\eta_\theta(\boldsymbol{x}) \geq \frac{1}{2}$，则 $C^*_\theta(\boldsymbol{x}) = 1$；否则为 $= -1$。该分类器在所有分类器 C: $\mathcal{X} \to \{-1, 1\}$中，使误分类风险 $R_\theta(C) = \mathbb{P}_\theta(C(\boldsymbol{X}) \neq Y)$ 最小。设 $(\boldsymbol{X}_1, Y_1) \ldots, (\boldsymbol{X}_n, Y_n)$ 为 iid 训练样本，其分布与 $(\boldsymbol{X}, Y)$ 相同。本文的具体目标是设计一个基于训练样本的分类器 C，使其在以下极小极大意义下逼近 C^*_θ：

$$\sup_{\theta \in \Theta} \mathcal{E}_\theta(\widehat{C}) \asymp \inf_{\widehat{C}} \sup_{\theta \in \Theta} \mathcal{E}_\theta(\widehat{C}), \tag{1}$$

其中 $\mathcal{E}_\theta(C) = \mathbb{E}[R_\theta(C) - R_\theta(C^*_\theta)]$ 是 C的超额误分类风险，右侧（1）中的下确界取遍所有基于训练样本的分类器，而 Θ 是 θ 的一个适当空间，稍后将对此进行描述。

*通讯作者。完整邮寄地址：323 小马丁·路德·金博士大道，新泽西州纽瓦克，邮编 07102。
Email addresses: szw0100@auburn.edu （Shuoyang Wang）， zshang@njit.edu （Zuofeng Shang）

^等价地说，我们旨在设计 C，使其达到极小极大超额误分类风险（MEMR），从而在“最坏”情形下取得“最佳”性能。

判别分析方法的成功取决于如下假设： p和 q 均服从高斯分布。值得注意的是，在高斯假设下，对数条件概率 $\log \eta_\theta$ 本质上是线性或二次多项式，因此可以通过线性或二次判别分析准确近似。当p和 q 是高维非高斯密度时， η_θ 通常具有复杂形式，因此判别分析不再适用。最近，[12] 提出了一种基于深度神经网络（DNN）的分类器，并证明其达到极小极大最优； [2] 提出了一种基于 DNN 的多分类方法并推导出风险上界，但两者都要求 d固定。目前尚不清楚，当 d 发散且 p和 q 为非高斯分布时，DNN 方法是否仍具有极小极大最优性。

我们的主要贡献是推导出（1）的非渐近收敛率，并证明深度神经网络分类器可以在对数阶误差范围内达到这一收敛率。具体而言，我们证明，当 Θ 由 θ 个满足 $\log \eta_\theta$（等价地，贝叶斯分类器）具有模块化结构的 $d = O$组成，并且 $\log n$（poly(n)）时，（1）的收敛率为 n^{-s_0}，其中 s_0 是一个通用常数，仅取决于噪声指数、内在维度以及模块成分的光滑性指数。此外，我们证明，深度神经网络分类器的误分类风险上确界有如下上界： n^{-s_0} 乘以 $\log n$的某个幂，因此其近乎极小极大最优。我们的结果表明，深度神经网络分类器有望处理复杂的高维数据。

本文其余部分的结构如下。在第2.1节中，我们回顾深度神经网络分类器。在第2.2节中，我们确立深度神经网络分类器在高维情形下的极小极大最优性。技术证明见第3节。

2. 深度神经网络分类器及其极小极大最优性

本节中，我们首先在第2.1节回顾深度神经网络方法，并在第2.2节确立其在高维非高斯情形下的极小极大最优性。

2.1. 深度神经网络分类器

下面详细回顾深度神经网络分类器。令 σ 表示修正线性单元（ReLU）激活函数，其定义为 $\sigma(x) = (x)_+$，其中 $x \in \mathbb{R}$。对于实向量 $\mathbf{v} = (v_1, \ldots, v_w)^\top$ 和 $\mathbf{y} = (y_1, \ldots, y_w)^\top$，定义移位激活函数 $\sigma_{\mathbf{v}}(\mathbf{y}) = (\sigma(y_1 - v_1) \ldots, \sigma(y_w - v_w))^\top$。对于 $L \geq 1$ 以及 $\mathbf{p} = (p_1, \ldots, p_L) \in \mathbb{N}^L$，令$\mathcal{F}(L, \mathbf{p})$表示具有 d 个输入、 L 个隐藏层，且对于 $l = 1, \ldots, L$，第 l 个隐藏层具有 p_l 个节点的前馈深度神经网络类。等价地，任意 $f \in \mathcal{F}(L, \mathbf{p})$可表示为

$$f(\mathbf{x}) = \mathbf{W}_L \sigma_{\mathbf{v}_L} \mathbf{W}_{L-1} \sigma_{\mathbf{v}_{L-1}} \ldots \mathbf{W}_1 \sigma_{\mathbf{v}_1} \mathbf{W}_0 \mathbf{x}, \quad \mathbf{x} \in \mathbb{R}^d, \tag{2}$$

其中，$\mathbf{W}_l \in \mathbb{R}^{p_{l+1} \times p_l}$（ $l = 0$、……、 L）为权重矩阵，$\mathbf{v}_l \in \mathbb{R}^{p_l}$（ $l = 1$、……、 L）为平移向量。这里我们采用约定：$p_0 = d$ 和 $p_{L+1} = 1$。

为避免过拟合，我们采用一种常见策略，按照[21]对网络参数进行稀疏化。具体而言，考虑以下稀疏深度神经网络类:

$$\begin{aligned} \mathcal{F}(L, s, B, \mathbf{p}) \quad = \quad & \Big\{ f \in \mathcal{F}(L, \boldsymbol{p}) : \max_{0 \leq l \leq L} \|\mathbf{W}_l\|_0 \leq s, \max_{1 \leq l \leq L} \|\mathbf{v}_l\|_0 \leq s, \\ & \max_{0 \leq l \leq L} \|\mathbf{W}_l\|_\infty \leq B, \max_{1 \leq l \leq L} \|\mathbf{v}_l\|_\infty \leq B, \|f\|_\infty \leq B \Big\}, \end{aligned} \tag{3}$$

其中， $\|\cdot\|_\infty$ 表示矩阵/向量的最大元素范数或函数的上确界范数， $\|\cdot\|_0$ 表示向量或矩阵中的非零元素数量， $s > 0$ 控制非零参数的数量， $B > 0$ 控制深度神经网络的最大权重、偏移量及上确界范数。令 $\phi(x) = (1 - x)_+$，~ ~$x \in \mathbb{R}$ 为合页损失。按照 [12]，，我们基于训练^样本构造分类器 $C_{\phi,\ n}(\boldsymbol{X}) := \operatorname{sgn}(f_{\phi,\ n}(\boldsymbol{X}))$，其中 $f_{\phi,\ n}$ 是通过最小化以下经验损失得到的深度神经网络:

$$\widehat{f}_{\phi,n} = \operatorname*{arg\,min}_{f \in \mathcal{F}(L,s,B,\mathbf{p})} \frac{1}{n} \sum_{i=1}^{n} \phi(Y_i f(\boldsymbol{X}_i)). \tag{4}$$

深度神经网络有一些防止过拟合的技术，例如 l_1 正则化和随机失活。在实践中，这些技术很容易在 R 和 Python 中实现。例如，应用随机失活时，只需控制随机失活率 ϱ，即可轻松实现深度神经网络的稀疏性 s。

对于训练过程，我们建议采用以下数据划分方法来选择 ϱ:

- 步骤 1. 将全样本（$\boldsymbol{X}_i$, Y_i）随机分配到两个子集，分别以 $\mathcal{I}_{train}$ 和 $\mathcal{I}_{test}$为索引，其规模约为 $|\mathcal{I}_{train}| = 0.7n$ 和 $|\mathcal{I}_{test}| = 0.3n$。
- 步骤 2. 对于每个 ϱ，我们训练一个深度神经网络 f_ϱ，使用（4）基于子集 $\mathcal{I}_{train}$，然后计算测试误差，基于子集 $\mathcal{I}_{test}$ 作为

$$\mathrm{err}(\varrho) = \frac{1}{|\mathcal{I}_{test}|} \sum_{i \in \mathcal{I}_{test}} I(\widehat{f}_\varrho(\boldsymbol{X}_i Y_i < 0). \tag{5}$$

- 步骤 3. 选择 ϱ（可能从预先选定的集合中选择），以最小化 err(ϱ)。

[12] 表明，上述 $C_{\phi,\ n}$ 在d 固定时可以达到由 [1] 建立的 MEMR 尖锐速率。当 d 发散时，该尖锐速率要么过慢，要么甚至无法收敛到零。在下一节中，我们将重构经典极小极大理论以纳入高维性，并建立 $C_{\phi,\ n}$ 在高维情形下的极小极大最优性。

2.2. 高维中的极小极大最优性

在本节中，我们建立深度神经网络分类器（4）在高维情形下的极小极大最优性。我们首先为 θ引入一个适当的参数空间，并在此基础上建立我们的结果。

对于 $t \geq 1$、K、$\beta > 0$ 以及一个可测子集 $D \subset \mathbb{R}^t$，定义半径为K 的 β-H¨older 函数球为

$$\mathcal{C}_t^\beta(D, K) = \left\{ f : D \mapsto \mathbb{R} : \sum_{\boldsymbol{\alpha}:|\boldsymbol{\alpha}|<\beta} \|\partial^{\boldsymbol{\alpha}} f\|_\infty + \sum_{\boldsymbol{\alpha}:|\boldsymbol{\alpha}|=\lfloor\beta\rfloor} \sup_{\boldsymbol{x},\boldsymbol{x}' \in D, \boldsymbol{x} \neq \boldsymbol{x}'} \frac{|\partial^{\boldsymbol{\alpha}} f(\boldsymbol{x}) - \partial^{\boldsymbol{\alpha}} f(\boldsymbol{x}')|}{\|\boldsymbol{x} - \boldsymbol{x}'\|_\infty^{\beta - \lfloor\beta\rfloor}} \leq K \right\},$$

其中，$\partial^\alpha = \partial^{\alpha_1} \ldots \partial^{\alpha_t}$ 表示具有多重指标 $\alpha = (\alpha_1, \ldots, \alpha_t) \in \mathbb{N}^t$的偏微分算子，且$|\alpha| = \alpha_1 + \cdots + \alpha_t$。对于 $q \geq 1$，$\boldsymbol{d} = (d_1, \ldots, d_q) \in \mathbb{N}^q$，$\boldsymbol{t} = (t_0, \ldots, t_q) \in \mathbb{N}^{q+1}$，其中 $t_0 \leq d$，且对于$u = 1, \ldots, q$，$t_u \leq d_u$；$\boldsymbol{\beta} = (\beta_0, \ldots, \beta_q) \in \mathbb{R}_+^{q+1}$，$\boldsymbol{a} = (a_1, \ldots, a_{q+1}) \in \mathbb{R}^{q+1}$，$\boldsymbol{b} = (b_1, \ldots, b_{q+1}) \in \mathbb{R}^{q+1}$，其中对于 $u = 1, \ldots, q+1$，$a_u < b_u$；$\boldsymbol{K} = (K_0, \ldots, K_q) \in \mathbb{R}_+^{q+1}$，定义 $\mathcal{G} := \mathcal{G}(q, \boldsymbol{d}, \boldsymbol{t}, \boldsymbol{\beta}, \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{K})$ 为函数类g，其具有一个模块化表达式

$$g(\boldsymbol{x}) = g_q \circ \cdots \circ g_0(\boldsymbol{x}), \quad \boldsymbol{x} \in \mathcal{X}, \tag{6}$$

其中，对于 $u = 0, \ldots, q$，$g_u = (g_{u1}, \ldots, g_{ud_{u+1}}) : [a_u, b_u]^{d_u} \mapsto [a_{u+1}, b_{u+1}]^{d_{u+1}}$，且 $g_{uv} \in \mathcal{C}_{t_u}^{\beta_u}\left([a_u, b_u]^{t_u}, K_u\right)$。注意，每个组成部分 g_{uv} 仅依赖于 $t_u (\leq d_u)$ 个变量，这意味着 g_u 在内在上依赖于 d_u 个变量的小型局部簇，即 g_u 呈现出一种低维结构。因此，t_u可被视为 g_u的内在维度。结构（6）已被 [21, 20, 25, 14, 17] 用于基于深度学习的多元回归，以应对“维度灾难”。（6）的例子包括广义加性模型 [9, 16], 张量积空间方差分析模型 [15], 等。具体而言，前者对应于 $g_0(\boldsymbol{x}) = (f_j(x_j))_{j \in \mathcal{J}}$，其中 $f_1, \ldots, f_d$ 为一元光滑函数，$\mathcal{J} \subset \{1, \ldots, d\}$ 为一组 d_1 个索引，且 $g_1(z_1, \ldots, z_{d_1}) = z_1 + \cdots + z_{d_1}$，使得 $d_0 = d, d_2 = 1, t_0 = 1, t_1 = d_1$；后者对应于 $g_0(\boldsymbol{x}) = (f_{j_1}(x_{j_1}) \cdots f_{j_r}(x_{j_r}))_{(j_1,\ldots,j_r) \in \mathcal{J}}$，其中 $\mathcal{J} \subset \{(j_1, \ldots, j_r) : 1 \leq j_1 < \cdots < j_r \leq d\}$为一组 d_1 r元组，$1 \leq r < d$为交互阶数，且 $g_1(z_1, \ldots, z_{d_1}) = z_1 + \cdots + z_{d_1}$，使得 $d_0 = d, d_2 = 1, t_0 = r, t_1 = d_1$。除上述可加性和乘法性之外，（6）还允许局部变量之间存在更一般类型的交互。

此外，我们对 η_θ 作出如下假设。

假设1。存在 $\alpha \geq 0$ 和 $C_d = O(\log d)$，使得

$$\mathbb{P}\left(\left|\eta_\theta(\boldsymbol{X}) - \frac{1}{2}\right| \leq t\right) \leq C_d t^\alpha, \qquad \forall t > 0. \tag{7}$$

假设1 被称为噪声条件，它刻画了 η_θ 与 1/2（随机猜测）之间的差异。当 d 固定时， C_d也固定，假设 1 便退化为[18, 23, 1]中考虑的经典噪声条件。当 d 发散时，我们允许 C_d 以 $\log d$ 的速率发散，这允许 η_θ 与 1/2 之间可能存在更小的差距。

令 Θ 为如下定义的 θ 空间:

$$\begin{aligned}\Theta &\equiv \Theta\left(q, \boldsymbol{d}, \boldsymbol{t}, \boldsymbol{\beta}, \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{K}, C_d, \alpha, c\right) \\ &= \{\theta=(p, q, \pi_p, \pi_q): p, q \text{ are probability densities on } \mathcal{X} \text{ such that} \\ &\quad \eta_\theta \in \mathcal{G} \text{ and } \eta_\theta \text{ satisfies Assumption } 1, c \leq \pi_p, \pi_q \leq 1-c, \pi_p+\pi_q=1\},\end{aligned}$$

其中 $c \in (0, 1/2]$ 是一个给定的常数。令

$$s_0=\min_{0\leq u\leq q} \frac{\beta_u^*(\alpha+1)}{\beta_u^*(\alpha+2)+t_u}, s_1=\max_{0\leq u\leq q} \frac{t_u}{\beta_u^*(\alpha+2)+t_u},$$

其中 $\beta_u^*=\beta_u \prod_{k=u+1}^q(\beta_k \wedge 1)$。以下结果给出了 MEMR 的非渐近下界。

定理 2.1。存在一个正常数 D_1，其取决于 $q, \boldsymbol{t}, \boldsymbol{\beta}, \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{K}, \alpha, c$，使得 t

$$\inf_{\widehat{C}} \sup_{\theta\in\Theta} \mathcal{E}_\theta(\widehat{C}) \geq D_1 C_d\left(\frac{\log d}{n}\right)^{s_0} \wedge 1,$$

其中，下确界取遍所有基于训练样本的分类器。

下面的定理进一步推导出超额误分类风险 $\widehat{}$的$C_{\phi,n}$非渐近上界。

定理 2.2. 假设 $C_d^{1/s_0} \log d=O\left(n \log^{-2} n\right)$，且网络类 $\mathcal{F}(L, s, B, \boldsymbol{p})$ 满足

(i) $L \asymp \log n$;

(ii) $\max_{0\leq \ell\leq L} p_\ell \asymp \max\left\{d, \left(\frac{n}{\log^2 n(\log n+\log d)}\right)^{s_1}\right\}$;

(iii) $s \asymp \left(\frac{n}{\log^2 n(\log n+\log d)}\right)^{s_1} \log n$;

(iv) $B \asymp \left(\frac{n}{\log^2 n(\log n+\log d)}\right)^{\frac{s_0}{\alpha+1}}$.

那么存在一个常数 D_2，它依赖于 $q, \boldsymbol{t}, \boldsymbol{\beta}, \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{K}, \alpha, c$，使得

$$\sup_{\theta\in\Theta} \mathcal{E}_\theta(\widehat{C}_{\phi,n}) \leq D_2 C_d\left(\frac{\log^3 n+\log^2 n \log d}{n}\right)^{s_0}.$$

结合定理 2.1 和2.2，我们得到了 MEMR$\widehat{}$的非渐近（近乎）尖锐速率，以及$C_{\phi,n}$的极小极大最优性。特别地，若 $\log d \asymp n^\varepsilon$ 对 $0<\varepsilon<s_0/(s_0+1)$ 成立，且 $C_d \asymp \log d$，则有

$$D_1 C_d n^{-(1-\varepsilon)s_0} \leq \inf_{\widehat{C}} \sup_{\theta\in\Theta} \mathcal{E}_\theta(\widehat{C}) \leq D_2 C_d n^{-(1-\varepsilon)s_0} \log^{2s_0} n, \tag{8}$$

$\widehat{}$ $\widehat{}$这可通过 $C=C_{\phi,n}$ 实现，具体做法是适当选择网络架构，即满足条件(i)-(iv)。当 $\beta_u^* \equiv \beta$ 且 $t_u \equiv t$时，我们得到 $s_0=$ $^{\beta}_{(\alpha+1)\ \beta(\alpha+2)+t}$。尽管 d 很大，但速率（8）足够快，因为 s_0 仅涉及内在维度 t （而非 d）。

本节最后，我们讨论参数空间 Θ、假设1的可行性，以及关于网络架构选择的条件(i)-(iv) 。注意

$$\eta_\theta(\boldsymbol{x})=\frac{p(\boldsymbol{x})\pi_p}{p(\boldsymbol{x})\pi_p+q(\boldsymbol{x})\pi_q}=\frac{1}{1+\frac{\pi_q}{\pi_p}\cdot\frac{q(\boldsymbol{x})}{p(\boldsymbol{x})}}, \boldsymbol{x}\in\mathcal{X}.$$

因此， $\eta_\theta \in \mathcal{G}$ 等价于 $q/p \in \mathcal{G}$，这意味着两个总体密度的似然比取决于低维结构。换言之， q/p 仅涉及局部变量之间有限多个低阶交互作用。从概念上说，这类似于[4, 3] 在高维高斯情形下提出的稀疏性假设。为说明这一点，假设 p 和 q 是 d元高斯分布，其均值向量为 μ_p， μ_q 协方差矩阵为 Σ_p， Σ_q。那么， $q/p \in \mathcal{G}$意味着 $\|\Sigma_p^{-1} - \Sigma_q^{-1}\|_0$和 $\|\Sigma_p^{-1}(\mu_p - \mu_q)\|_0$都被一个有限数值所界定，即均值向量之差和精度矩阵之差都是稀疏的。

以下结果给出了假设 1 的一个充分条件。

命题 1。假设存在 $0 < \delta < 1$、 $0 \le \tau < 1$ 和 $M_d = O(\log d)$，使得 t

$$\sup_{r:|r-\pi_p/\pi_q|\le\delta} f_{q/p}(r)|r - \pi_p/\pi_q|^\tau \le M_d, \tag{9}$$

其中， $f_{q/p}$是 $q(\boldsymbol{x})/p(\boldsymbol{x})$的概率密度函数。则假设 *1* 成立，其中 $\alpha = 1 - \tau$和$C_d = \max\{C^{(1)}_{\tau,c,\delta} M_d, C^{(2)}_{\tau,c,\delta}\}$， $C^{(1)}_{\tau\,c\,,,\delta} = \frac{2^{2-\tau}}{1-\tau}(\frac{2(1-c)+c\delta}{c})^{1-\tau}$、 $C^{(2)}_{\tau\,c\,,,\delta} = (\frac{4-2c}{c\delta})^{1-\tau}$， 以及 $c = \min\{\pi_p, 1-\pi_p\}$。

由于 $\{\boldsymbol{x}: \ q(\boldsymbol{x})/p(\boldsymbol{x}) = \pi_p/\pi_q\}$ 定义了决策边界，9式中 $f_{q/p}(r)$ 当 r 属于 δ-邻域时， π_p/π_q 刻画了决策边界附近的最大密度。直观地说，当这一最大密度变大时，分类任务会变得更具挑战性。命题1表明，只要这一最大密度受到适当的上界约束，假设 1 就成立。该上界条件非常宽松，因为在 π_p/π_q 附近，它甚至可以以适当的速度发散至无穷。因此， （9）是关于 q/p 附近决策边界处最大密度的一个温和正则性条件，且已证明该条件可推出假设 1。

条件 (i)-(iv) 对网络架构施加了精确的阶数限制。需要指出的是，通过更繁琐的论证，可以将这些条件推广为架构参数的适当取值范围，同时仍能保证定理 2.2. 的相同结论。

3. 定理 **2.1**和 **2.2** 的证明

3.1. 定理*2.1* 的证明

证明。令 $u^* = \arg\min_{u=0,\ldots,q} \frac{\beta^*_u}{(\alpha+1)\beta^*_u(\alpha+2)+t_u}$， $\beta^* = \beta_{u^*}$， $t^* = t_{u^*}$， $\beta^{**} = \beta^*_{u^*}$和 $\widetilde{\beta} = \prod_{k=u^*+1}^{q}(\beta_k \wedge 1)$。不失一般性，假设 u^* 是唯一的。

对于整数 $\nu \ge 1$，定义 $\mathbb{R}^{t^*}$ 上的正则网格如下：

$$G_\nu = \left\{ \left(\frac{2k_1+1}{2\nu}, \ldots, \frac{2k_{t^*}+1}{2\nu} \right) : k_\ell \in \{0, \ldots, \nu-1\}, \ell = 1, \ldots, t^* \right\}.$$

令 $n_\nu(\boldsymbol{x}) \in G_\nu$ 为 G_ν 中距离 $\boldsymbol{x} \in \mathbb{R}^{t^*}$ 最近的点。令 $\mathcal{X}_\ell$、 $\ell = 0$、...、 m 为 $\mathbb{R}^{t^*}$ 的划分，该划分在[1], 中定理 4.1 的证明里定义，其中 $m \le \nu^{t^*}$。

设 $\gamma: \mathbb{R}^+ \to \mathbb{R}^+$ 是一个非增的无穷可微函数，使得 $\gamma = 1$ 在 $[0, 1/4]$ 上且 $\gamma = 0$ 在 $[1/2, \infty$ 上）。例如， γ 可以按[1]中的方法构造： $\gamma(x) = \left(\int_{1/4}^{1/2} \gamma_1(s)ds\right)^{-1} \int_x^\infty \gamma_1(t)dt$，其中

$$\gamma_1(x) = \begin{cases} \exp\left\{\frac{1}{(x-1/4)(x-1/2)}\right\}, & x \in (1/4, 1/2), \\ 0, & \text{otherwise.} \end{cases}$$

设 h_{u^*}: $\mathbb{R}^{t^*} \to \mathbb{R}^+$ 是一个定义为 $h_{u^*}($ $\boldsymbol{x}$ $\boldsymbol{x} - n_\nu\ \boldsymbol{x}$ $\boldsymbol{x} =$ $\nu^{|s|-\beta^*} C_\gamma D^s \gamma\ \|\nu\ \boldsymbol{x} - n_\nu\ \boldsymbol{x}\|$ $s \in \mathbb{N}^{t^*}$ $|s| \le \lceil\beta^*\rceil$ C_γ $) = \nu^{-\beta^*} C_\gamma \gamma(\|\nu(\,(\,)\,)\|)$ 的函数，其中对于任意满足条件的， $D^s h_{u^*}(\,)\,(\,(\,(\,)\,)\,)$，且 是一个足够小的常数，以确保 $h_{u^*} \in \mathcal{C}^{\beta^*}_{t^*}(\mathbb{R}^{t^*}, K^*)$ 对某个常数 $K^* > 0$ 成立。这里，我们要求 C_γ 足够小，使得 h_{u^*} 的利普希茨常数为 K^*。

下面，我们基于 h_{u^*} 构造一个特殊的复合函数。对于 $\sigma := (\sigma_1, \ldots, \sigma_m) \in \{-1, 1\}^m$， 令 $h(\boldsymbol{x}) = \sum_{j=1}^m \sigma_j h_j(\boldsymbol{x})$，使得 $h_j(\boldsymbol{x}) = h_{u^*}\mathbb{I}(\boldsymbol{x} \in \mathcal{X}_j)$。容易验证， $h \in \mathcal{C}^{\beta^*}_{t^*}(\mathbb{R}^{t^*}, 2K^*)$。定义以下函数

$$\begin{cases} g_u(x_1, \ldots, x_{d_u}) = (x_1, \ldots, x_{d_u}), & u < u^*, \\ g_u(x_1, \ldots, x_{d_u}) = (h(x_1, \ldots, x_{t^*}), 0, \ldots, 0), & u = u^*, \\ g_u(x_1, \ldots, x_{d_u}) = (x_1^{\beta_u \wedge 1}, 0, \ldots, 0), & u > u^*. \end{cases}$$

对于 $\boldsymbol{z} \in \mathbb{R}^d$ 和 $\boldsymbol{x} \in \mathbb{R}^{t^*}$，将 $z_g(\boldsymbol{z})$ 定义为 $g_{u^*} \circ g_{u^*-1} \circ \ldots \circ g_0(\boldsymbol{z})$ 的第一个元素，并令 $h(\boldsymbol{x}) = z_g(\boldsymbol{z})$。令

$$\eta_{\theta,\sigma}(\boldsymbol{z}) \quad = \quad \frac{1}{2} + \frac{1}{2} g_q \circ \ldots \circ g_{u^*+1} \circ g_{u^*} \circ g_{u^*-1} \circ \ldots \circ g_0(\boldsymbol{z}) = \frac{1}{2} + \frac{1}{2} \sum_{j=0}^{m} \sigma_j h_{u^*}^{\widetilde{\beta}}(\boldsymbol{x}) \mathbb{I}(\boldsymbol{x} \in \mathcal{X}_j).$$

此外，令 $\eta^*_{\theta,\sigma}(\boldsymbol{x}) = \frac{1}{2} + \frac{1}{2} \sum_{j=0}^{m} \sigma_j h_{u^*}^{\beta}(\boldsymbol{x}) \mathbb{I}(\boldsymbol{x} \in \mathcal{X}_j)$，且 $\eta^*_{\theta,\sigma} \in \mathcal{C}_{t^*}^{\beta^{**}}\left(\mathbb{R}^{t^*}, 2K^*\right)$。

令 $\nu = \nu_0 (\log d)^{\frac{1}{(2+\alpha)\beta^*+t^*}}$，$\nu_0 = \lceil C n^{\frac{2+\alpha}{(2+\alpha)\beta^*+t^*}} (\log d)^{\frac{\beta^*}{(2+\alpha)\beta^*+t^*}} \rceil$，$w = C' \nu^{2\beta^*} n^{-1}$，$m = C_d \nu_0^d (\log d)^{\frac{t^*}{(2+\alpha)\beta^*+t^*}} (2+\alpha)^{\frac{\beta^*}{(2+\alpha)\beta^*+t^*}}$，其中 d满足 $\log d = O(n)$，C和 C' 是正的绝对常数。我们假设 $C_d \leq \log d$，并忽略通用常数项。根据 [1], 的证明，我们可以验证

$$m = C_d (\log d)^{-1} q^d \leq q^d, \quad mw = O\left(\nu_0^{-\alpha\beta^*}\right) (\log d)^{\frac{-\alpha\beta^*}{(2+\alpha)\beta^*+t^*}} = O\left(\nu^{-\alpha\beta^*}\right).$$

因此，边际条件得到满足。其余证明直接沿用 Audibert 和 Tsybakov 对定理 4.1 的证明，其中我们有 $\nu^{-\beta^*}\sqrt{nw} = C'$ and

$$mw\nu^{-\beta^*} = O\left(C_d \nu^{-(1+\alpha)\beta^*}\right) = O\left(C_d (\log d/n)^{\frac{(1+\alpha)\beta^*}{(2+\alpha)\beta^*+t^*}}\right).$$

选择适当的常数 D_1，其取决于 q、t β、a、b、K α and c，证明即告完成。

□

3.2. 定理2.2的证明

在证明定理 2.2之前，我们先给出一些预备引理。

引理 1。（[21]中的引理 3）设 h_{uv} 为如下定义的函数:

$$h_0 = \frac{g_0}{2K_0} + \frac{1}{2}, \quad h_i = \frac{g_i(2K_{i-1} \cdot -K_{i-1})}{2K_i} + \frac{1}{2}, \quad h_q = g_q(2K_{q-1} \cdot -K_{q-1}),$$

其中，对于任意 $\boldsymbol{x} \in [0,1]^{d_u}$、 $u \in 0$、.....、 q， $2K_{u-1}\boldsymbol{x} - K_{u-1}$ 等价于对每个 $u = 1$、.....、 d_u的 $2K_{u-1}x_u - K_{u-1}$的变换。于是，对于满足 $\widetilde{h}_{uv}: [0,1]^{t_u} \to [0,1]$ 的任意函数 $\widetilde{h}_u = \left(\widetilde{h}_{uv}\right)^{\intercal}$，

$$\|h_q \circ \ldots \circ h_0 - \widetilde{h}_q \circ \ldots \circ \widetilde{h}_0\|_\infty \leq K_q \prod_{\ell_0}^{q-1} (2K_\ell)^{\beta_{\ell+1}} \sum_{u=0}^{q} \| |h_u - \widetilde{h}_u|_\infty \|_\infty^{\prod_{\ell=u+1}^{q} \min\{\beta_\ell, 1\}}.$$

引理 2。（[21]中的定理 5）对于任意函数 $f \in \mathcal{C}^\beta([0,1]^r, K)$ 以及任意整数 $m \geq 1$ 和 $N \geq \max\{(\beta+1)^r, (K+1)e^r\}$，存在一个网络

$$\widetilde{f} \in \mathcal{F}(L, r, (r, 6(r+\lceil\beta\rceil)N, \ldots, 6(r+\lceil\beta\rceil)N, 1), s, \infty)$$

其深度为

$$L = 8 + (m+5)(1 + \lceil \log_2(\max\{r, \beta\}) \rceil)$$

且参数数量为

$$s \leq 141(r+\beta+1)^{3+r} N(m+6),$$

使得

$$\|\widetilde{f} - f\|_\infty \leq (2K+1)(1+r^2+\beta^2) 6^r N 2^{-m} + K 3^\beta N^{-\beta/r}.$$

下文中，不失一般性，我们考虑 $\pi_p = \pi_q = 1/2$ with $c = 1/2$。结果很容易推广到一般的 π_p, π_q。

~引理 3.（回归函数逼近）对于任意 $m, N \in \mathbb{N}^+$，存在一个 $f \in \mathcal{F}(L, d, \boldsymbol{p}, s, B)$,~ ~$C = sign(f)$ 和一个常数 C_1，满足

$$\sup_{(p,q,\pi_p,\pi_q)\in\Pi} \mathbb{E}\left[R(\widetilde{C}) - R(C^*)\right] \leq 2^{\alpha+3} C_1 C_d \left(\sum_{u=0}^{q} A_u C(t_u)(N2^{-m})^{\beta_u^{**}} + \sum_{u=0}^{q} B_u N^{-\beta_u^*/t_u}\right)^{\alpha+1},$$

使得

$$L = 3q - 1 + \sum_{u=0}^{q} \left[8 + (m+5)(1 + \lceil \log_2(t_u \vee \beta_u) \rceil)\right], \quad \boldsymbol{p} = (d, 12rN, \ldots, 12rN, 3, 3, 1),$$

$$s = 2\sum_{u=0}^{q} d_{u+1}\left[141(t_u + \beta_u + 1)^{3+t_u} N(m+6) + 4\right] + 7,$$

$$B = \left(\sum_{u=0}^{q} A_u C(t_u)(N2^{-m})^{\beta_u^{**}} + \sum_{u=0}^{q} B_u N^{-\beta_u^*/t_u}\right)^{-1},$$

其中 $r = \max_u d_{u+1}(t_u + \lceil \beta_u \rceil)$, $A_u = \left\{K_q \prod_{\ell=0}^{q-1} (2K_\ell)^{\beta_{\ell+1}}\right\}(2Q_u + 1)$, $B_u = \left\{K_q \prod_{\ell=0}^{q-1} (2K_\ell)^{\beta_{\ell+1}}\right\}(Q_u 3^{\beta_u})^{\beta_u^{**}}$，

$C(t_u) = 2t_u^2 6^{t_u \beta_u^{**}}$, $\beta_u^{**} = \prod_{\ell=u+1}^{q} \beta_\ell \wedge 1$, Q_u 是仅依赖于 β_u, K_u 的一些绝对常数。

证明。记条件概率函数类 $\mathcal{G}(q, \boldsymbol{d}, \boldsymbol{t}, \boldsymbol{\beta}, \boldsymbol{K})$。令 m 为深度参数，且 N 为宽度参数。根据引理 1 和引理 2，我们有

$$\inf_{\eta^* \in \mathcal{F}(L,\boldsymbol{p},s,B)} \|\eta^* - \eta\|_\infty \leq \|h_q \circ \ldots \circ h_0 - \widetilde{h}_q \circ \ldots \circ \widetilde{h}_0\|_\infty \leq \sum_{u=0}^{q} A_u C(t_u)(N2^{-m})^{\beta_u^{**}} + \sum_{u=0}^{q} B_u N^{-\beta_u^*/t_u},$$

相应的网络结构在特定近似误差下为

$$L = 3(q-1) + \sum_{u=0}^{q} \left[8 + (m+5)(1 + \lceil \log_2(t_u \vee \beta_u) \rceil)\right], \quad \boldsymbol{p} = (d, 6rN, \ldots, 6rN, 1),$$

$$s = \sum_{u=0}^{q} d_{u+1}\left[141(t_u + \beta_u + 1)^{3+t_u} N(m+6) + 4\right],$$

且 $B = 1$，其中 $r = \max_u d_{u+1}(t_u + \lceil \beta_u \rceil)$。定义

$$\epsilon = \sum_{u=0}^{q} A_u C(t_u)(N2^{-m})^{\beta_u^{**}} + \sum_{u=0}^{q} B_u N^{-\beta_u^*/t_u}, \quad \widetilde{\eta} = \operatorname*{arg\,min}_{\eta^* \in \mathcal{F}(L,\boldsymbol{p},s,B)} \|\eta^* - \eta_\theta\|_\infty.$$

现在我们构造

$$\widetilde{f} = 2\left(\sigma\left(\epsilon^{-1}\left(\widetilde{\eta} - \frac{1}{2}\right)\right) - \sigma\left(\epsilon^{-1}\left(\widetilde{\eta} - \frac{1}{2}\right) - 1\right)\right) - 1.$$

~我们还需要从 $\widetilde{\eta}$ 到 f的两层，这可以通过以下方式获得

$$\widetilde{\eta} \mapsto \sigma\left(\epsilon^{-1}\left(\widetilde{\eta} - \frac{1}{2}\right)\right), \quad \widetilde{\eta} \mapsto \sigma\left(\epsilon^{-1}\left(\widetilde{\eta} - \frac{1}{2}\right) - 1\right)$$

权重的最大值由 ϵ^{-1} 上界控制。由于减法部分乘以 2，我们需要将宽度加倍，并在最后一层加上带偏置项 -1 的加性结构。定义

$$\boldsymbol{A} = \left\{\boldsymbol{x} \in [0,1]^d : |\eta_\theta(\boldsymbol{x}) - \frac{1}{2}| > 2\epsilon\right\},$$

~则当 $\boldsymbol{x} \in \boldsymbol{A}$ 时， $f(\boldsymbol{x}) = f_\phi^*(\boldsymbol{x})$，其中 $f_\phi^* = \arg\min_f \mathbb{E}\left[\phi(Yf(\boldsymbol{x}))\right]$，对 $[0,1]^d$ 上所有可测实值函数均成立。这是因为当 $\eta_\theta(\boldsymbol{x}) - \frac{1}{2} > 2\epsilon$ 时，我们有

$$\widetilde{\eta}(\boldsymbol{x}) - \frac{1}{2} = \widetilde{\eta}(\boldsymbol{x}) - \eta_\theta(\boldsymbol{x}) + \eta_\theta(\boldsymbol{x}) - \frac{1}{2} > -\epsilon + 2\epsilon = \epsilon$$

当 $\eta_\theta(\boldsymbol{x}) - \frac{1}{2} < -2\epsilon$ 时，我们有

$$\widetilde{\eta}(\boldsymbol{x}) - \frac{1}{2} = \widetilde{\eta}(\boldsymbol{x}) - \eta_\theta(\boldsymbol{x}) + \eta_\theta(\boldsymbol{x}) - \frac{1}{2} < \epsilon - 2\epsilon = -\epsilon.$$

对于适当选取的 m 和 N， ϵ 在 0 附近可以足够小。注意， ϕ 是 Fisher 一致的，即 $sign(f_\phi^*) = C^*$，并且根据 [22], 的定理 2.31，存在一个常数 C_1，使得

$$\sup_{(p,q,\pi_p,\pi_q)\in\Pi} \mathbb{E}\left[R(\widetilde{C}) - R(C^*)\right] \leq C_1 \sup_{(p,q,\pi_p,\pi_q)\in\Pi} \mathbb{E}\left[\phi\left(Y\widetilde{f}\right) - \phi\left(YC^*\right)\right].$$

因此,

$$\begin{aligned}
\mathbb{E}\left[\phi\left(Y\widetilde{f}(\boldsymbol{x})\right) - \phi\left(YC^*(\boldsymbol{x})\right)\right] =& \int \left|\widetilde{f}(\boldsymbol{x}) - C^*(\boldsymbol{x})\right| |2\eta_\theta(\boldsymbol{x}) - 1| \, dP(\boldsymbol{x}) \\
=& 2\int_{\boldsymbol{A}^c} \left|\widetilde{f}(\boldsymbol{x}) - C^*(\boldsymbol{x})\right| \left|\eta_\theta(\boldsymbol{x}) - \frac{1}{2}\right| dP(\boldsymbol{x}) \\
\leq& 4\int_{\boldsymbol{A}^c} \left|\eta_\theta(\boldsymbol{x}) - \frac{1}{2}\right| dP(\boldsymbol{x}) \leq 8\epsilon\mathbb{P}\left(\left|\eta_\theta(\boldsymbol{x}) - \frac{1}{2}\right| \leq 2\epsilon\right) \leq 2^{\alpha+3} C_d \epsilon^{\alpha+1},
\end{aligned}$$

证明完毕。 □

定义 3.1（覆盖数）设 $\upsilon > 0$和 $\|f\|_\infty = \sup_{\boldsymbol{z}\in\mathcal{C}} |f(\boldsymbol{z})|$。若对任意 $f \in \mathcal{F}$，均存在一个 f_k 使得 $\|f_k - f\|_\infty \leq \upsilon$，则称子集 $\{f_k \in \mathcal{F}\}_{k\geq 1}$为关于 $\|f\|_\infty$ 的 υ-覆盖集。 $\mathcal{F}$ 关于 $\|f\|_\infty$ 的 υ-覆盖数定义为

$$\mathcal{N}(\upsilon, \mathcal{F}, \|\cdot\|_\infty) = \inf\left\{N \in \mathbb{N} : \exists f_1, \ldots, f_N, s.t. \ \mathcal{F} \subset \bigcup_{k=1}^{N} \{f \in \mathcal{F} : \|f_k - f\|_\infty \leq \upsilon\}\right\}.$$

定义 3.2（括号熵）若 $\|f_k^L - f_k^U\|_\infty \leq \upsilon$ 且对任意 $f \in \mathcal{F}$，均存在一对 $\left(f_k^L, f_k^U\right)$ 使得 $f_k^L \leq f \leq f_k^U$，则称一对集合 $\left\{\left(f_k^L, f_k^U\right) \in \mathcal{F}\times\mathcal{F}\right\}_{k\geq 1}$为关于 $\|f\|_\infty$的 υ-括号集。关于 $\|f\|_\infty$ 的最小 υ-括号集的基数称为 υ-括号数，记作 $\mathcal{N}_B(\upsilon, \mathcal{F}, \|\cdot\|_\infty)$。定义 υ-括号熵为 $H_B(\upsilon, \mathcal{F}, \|\cdot\|_\infty) = \log \mathcal{N}_B(\upsilon, \mathcal{F}, \|\cdot\|_\infty)$。对于任意 $\upsilon > 0$，已知

$$\log\mathcal{N}(\upsilon, \mathcal{F}, \|\cdot\|_\infty) \leq H_B(\upsilon, \mathcal{F}, \|\cdot\|_\infty) \leq \log\mathcal{N}(\upsilon/2, \mathcal{F}, \|\cdot\|_\infty).$$

引理 4. 给定稀疏神经网络类 $\mathcal{F}(L, d, \boldsymbol{p}, s, B)$ 以及任意 $\upsilon > 0$，我们有

$$\log\mathcal{N}(\upsilon, \mathcal{F}(L, d, \boldsymbol{p}, s, B), \|\cdot\|_\infty) \leq 2L(s+1)\log\left\{\upsilon^{-1}(L+1)\left(\max\{\|\boldsymbol{p}\|_\infty, d\}+1\right)(B\vee 1)\right\}.$$

Le引理 5。 对于引理 $\mathcal{F}(L$、 d、 $\boldsymbol{p}$、 s、 $B)$ 中的上述函数类 3，当取 e

$$m = \max_u \left((\beta_u^{**})^{-1}\log_2(A_u C(t_u)/B_u) + (1 + \beta_u/t_u)\log_2 N\right),$$

存在一个常数 C_2，使得

对数 $\mathcal{N}(\upsilon$,$(L, d, \boldsymbol{p}, s, B), \|\cdot\|_\infty) \leq C_2 \log_2^2(1/\epsilon)\left(\epsilon^{-1}\right)^{\max_u t_u/\beta_u^*}\left(\log_2\left(\upsilon^{-1}\right) + \log_2 d + \log_2(1/\epsilon)\right)$

证明。 当 $m = \text{最大值}_u \left((\beta_u^{**})^{-1} \text{对数}_2(A_u C(t_u)/B_u) + (1+\beta_u/t_u) \text{对数}_2 N\right)$，根据引理 3，我们有 $\sum_{i=0}^{q} A_i C(t_u)(N2^{-m})^{\beta_i^{**}} \leq \sum_{i=0}^{q} B_i N^{-\beta_i^*/t_u}$，并且 $\sum_{u=0}^{q} B_u N^{-\beta_u^*/t_u}$ 支配近似误差，例如 ϵ。

对于较大的 $N \gg \text{最大值}_u B_u$，我们有 $N^{-\text{最小值}^u \beta_u^*/t_u} \asymp \epsilon$。因此，引理 3中的上述网络类具有

$$L \asymp \log_2(\epsilon^{-1}),\ \|\boldsymbol{p}\|_\infty \asymp \max\left\{d, \left(\epsilon^{-1}\right)^{\max_u t_u/\beta_u^*}\right\},\ s \asymp \log_2(1/\epsilon)\left(\epsilon^{-1}\right)^{\max_u t_u/\beta_u^*},$$

且 $B \asymp \epsilon^{-1}$。根据引理 4，我们有

$$\log \mathcal{N}(\upsilon, \mathcal{F}(L,d,\boldsymbol{p},s,B), \|\cdot\|_\infty) \leq C_2 \log_2^2(1/\epsilon)\left(\epsilon^{-1}\right)^{\max_u t_u/\beta_u^*}\left(\log_2\left(\upsilon^{-1}\right) + \log_2 d + \log_2(1/\epsilon)\right),$$

其中 C_2 取决于 q、t β、a、b、K 和 c。 □

Le引理 6。 (*[12]*中的定理 A.2) 在假设 *1* 及某些正则性条件下：

(i) 对于一个正数列 $\{\delta_n^*\}_{n\geq 1}$，存在函数类序列 $\{\mathcal{F}_n\}_{n\geq 1}$ 和 $f_n \in \mathcal{F}_n$，使得 *that*

$$\mathbb{E}\left[\phi\left(Y f_n(\boldsymbol{x})\right) - \phi\left(Y C^*(\boldsymbol{x})\right)\right] \leq \delta_n^*.$$

(ii) 存在一个序列 $\{\delta_n\}_{n\geq 1}$，使得 $H_B(\delta_n, \mathcal{F}_n, \|\cdot\|_2) \leq C_2 n \delta_n^{(\alpha+2)/(\alpha+1)}$，其中某个常数为$C_2 > 0$ 且 $\{\mathcal{F}_n\}_{n\geq 1}$ 在 *(i)*中。

若对任意足够小的 $a > 0$，$n(\max\{\delta_n^*, \delta_n\})^{2(\alpha+2)/(\alpha+1)} \gtrsim \log_2^{1+a} n$，则 $\widehat{f}_{\phi,\ n}$ 满足

$$\mathbb{E}\left[R(\widehat{C}^{FDNN}) - R(C^*)\right] \lesssim \left(\max\{\delta_n^*, \delta_n\}\right)^2.$$

现在我们准备好证明定理 2.2。

证明。 对于上述 ϵ，令 $\delta^2 = \epsilon^{\alpha+1}$，其对应于引理 δ_n^* 中的6(i)。根据引理 5，我们有

$$\log \mathcal{N}(\delta^2, \mathcal{F}(L,d,\boldsymbol{p},s,B), \|\cdot\|_\infty) \leq C_2\left(\delta^2\right)^{-\max_u \frac{t_u}{(\alpha+1)\beta_u^*}} \log^2\left(\delta^{-1}\right)\left(\log\left(\delta^{-1}\right) + \log d\right).$$

根据引理 6(ii)，我们有 $\left(\delta^2\right)^{\max_u \frac{t_u}{\alpha+1}\left(\frac{1}{\beta_u^*} + \frac{\alpha+2}{\alpha+1}\right)} \gtrsim n^{-1} \log^2\left(\delta^{-1}\right)\left(\log\left(\delta^{-1}\right) + \log d\right)$，由此得到 $C_d\delta^2 \gtrsim C_d\left(\frac{\log^2 n (\log n + d)}{n}\right)^{\min_u \frac{\beta_u^*}{\beta_u^* \log}\left(\alpha+1\right)\left(\alpha+2\right)^{+t_u}}$。因此，应用引理 6 即可完成证明，其中涉及的 C_d 与引理 3 中相同。令 $D_2 = 2^{\alpha+3} C_1 C_2$，则可由 $\epsilon^{-1} = \delta^{-2/(1+\alpha)}$ 得到 L、d、$\|\boldsymbol{p}\|_\infty$、$s$ 和 B 的渐近阶。 □

3.3. 命题的证明 *1*

证明。对于 $0 < t < \frac{c\delta}{4\ 1-c\ ()+2c\delta}$，我们有如下结果

$$\begin{aligned}
\mathbb{P}\left(\left|\eta_\theta(\boldsymbol{X}) - \frac{1}{2}\right| \leq t\right) &= \mathbb{P}\left(\left|\frac{p(\boldsymbol{X})\pi_p}{p(\boldsymbol{X})\pi_p + q(\boldsymbol{X})\pi_q} - \frac{1}{2}\right| \leq t\right) \\
&\leq \mathbb{P}\left(\frac{-4t}{1-2t}\cdot\frac{\pi_p}{\pi_q} \leq q/p(\boldsymbol{X}) - \pi_p/\pi_q \leq \frac{4t}{1-2t}\cdot\frac{\pi_q}{\pi_p}\right) \\
&\leq \int_{|y-\pi_p/\pi_q| \leq \frac{4t}{1-2t}\cdot\frac{1-c}{c}} f_{q/p}(y)dy \\
&\leq M_d \int_{|y-\pi_p/\pi_q| \leq \frac{4t}{1-2t}\cdot\frac{1-c}{c}} |y-\pi_p/\pi_q|^{-\tau} dy \\
&= 2M_d \int_0^{\frac{4t}{1-2t}\cdot\frac{1-c}{c}} z^{-\tau} dz \\
&= \frac{2M_d}{1-\tau}\cdot\left(\frac{1-c}{c}\right)^{1-\tau}\left(\frac{4t}{1-2t}\right)^{1-\tau} \\
&\leq C_{\tau,c,\delta}^{(1)} M_d t^{1-\tau},
\end{aligned}$$

其中 $C^{(1)}_{\tau\, c\, ,,\delta} = \frac{2^{2-\tau}}{1-\tau} \cdot \left(\frac{2(1-c)+c\delta}{c}\right)^{1-\tau}$，这是由于 $(1-2t)^{-1} \leq (2(1-c) + c\delta)\,(2(1-c))^{-1}$。

对于 $t \geq \frac{c\delta}{4\,1-c\,()+2c\delta}$，我们有 $t \geq c(4-2c)^{-1}\delta$，因此，

$$\mathbb{P}\left(|\eta_\theta(\boldsymbol{X}) - \frac{1}{2}| \leq t\right) \leq \delta^{\tau-1}\delta^{1-\tau} \leq C^{(2)}_{\tau,c,\delta}\, t^{1-\tau},$$

其中 $C^{(2)}_{\tau\, c\, ,,\delta} = \left(\frac{4-2c}{c\delta}\right)^{1-\tau}$。因此，假设 1 成立，其中 $C_d = \max\left\{C^{(1)}_{\tau\, c\, ,,\delta} M_d, C^{(2)}_{\tau\, c\, ,,\delta}\right\}$ 和 $\alpha = 1-\tau$.

□

致谢

Zuofeng Shan g 感谢美国国家科学基金会（NSF）资助的 DMS 1764280 和 DMS 1821157 项目对本研究的支持。

参考文献

[1] Jean-Yves Audibert 和 Alexandre B. Tsybakov。插件分类器的快速学习速率。《统计学年鉴》，35：608–633，2007。[2] Thijs Bos 和 Johannes Schmidt-Hieber。深度 ReLU 网络用于多分类的收敛速率。arXiv:2108.00969，2021。[3] T. Tony Cai 和 Linjun Zhang。高维稀疏二次判别分析的凸优化方法。arXiv:1912.02872，2019。[4] T. Tony Cai 和 Linjun Zhang。高维线性判别分析：最优性、自适应算法与缺失数据。《皇家统计学会期刊》B辑：统计方法论，81(4)：675–705，2019。[5] Jianqing Fan 和 Yingying Fan。利用特征退火独立规则进行高维分类。《统计学年鉴》，36：2605–2637，2008。[6] Jianqing Fan、Yingying Fan 和 Yichao Wu。《统计学前沿：高维数据分析》。世界科学出版社，2010。[7] Farzan Farnia 和 David Tse。一种监督学习的极小极大方法。载于第30届神经信息处理系统国际会议论文集，NIPS'16，第4240–4248页，美国纽约州红钩，2016。[8] Pedro Galeano、Esdras Joseph 和 Rosa E. Lillo。函数型数据的马氏距离及其在分类中的应用。《技术计量学》，57(2)：281–291，2015。[9] Travor J. Hastie 和 Robert J. Tibshirani。《广义加性模型》。Chapman & Hall/CRC，1990。[10] Trevor Hastie、Robert Tibshirani 和 Jerome Friedman。《统计学习要素》。施普林格统计学丛书。施普林格纽约公司，美国纽约州纽约，2001。[11] Tianyang Hu、Zuofeng Shang 和 Guang Cheng。教师-学生设定下深度神经网络分类器的尖锐收敛速率。arXiv:2001.06892，2020。[12] Yongdai Kim、Ilsang Ohn 和 Dongha Kim。深度神经网络用于分类的快速收敛速率。《神经网络》，138：179–197，2021。[13] Guillaume Lecu´e。对具有稀疏表示的贝叶斯规则类进行分类的极小极大快速速率。《电子统计学期刊》，2：741–773，2008。[14] Kexuan Li、Fangfang Wang、Ruiqi Liu、Fan Yang 和 Zuofeng Shang。通过深度神经网络从含噪数据中校准多维复常微分方程。arXiv:2106.03591，2021。

[15] Yi Lin。张量积空间方差分析模型。《统计学年鉴》，28：734–755，2000。[16] Ruiqi Liu、Ben Boukai 和 Zuofeng Shang。通过深度神经网络实现最优非参数推断。《数学分析与应用杂志》，505：125561，2022。[17]Ruiqi Liu、Zuofeng Shang 和 Guang Cheng。关于深度工具变量估计。arXiv:2004.14954，2021。[18] Enno Mammen 和 Alexandre B. Tsybakov。光滑判别分析。《统计学年鉴》，27：1808–1829，1999。[19] Santiago Mazuelas、Andrea Zanoni 和 Aritz Perez。具有0-1损失和性能保证的极小极大分类。arXiv:2010.07964，2020。[20] Nicholas Polson 和 Veronika Rˇockov´a。稀疏深度学习的后验集中。第32届神经信息处理系统年度会议（NeurIPS），2018。[21] J. Schmidt-Hieber。使用ReLU激活函数的深度神经网络进行非参数回归。《统计学年鉴》，48(4)：1875–1897，2020。[22] Ingo Steinwart 和 Andreas Christmann。支持向量机。施普林格科学与商业媒体，2008。[23] Alexandre B. Tsybakov。统计学习中分类器的最优聚合。《统计学年[23] 鉴》，32：135–166，2004。[24]Alexandre B. Tsybakov。《非参数估计导论》。施普林格统计学丛书。施普林格，纽约，2009。[25] Shuoyang Wang、Zuofeng Shang、Guanqun Cao 和 S. Jun Liu。函数型数据的最优分类。已提交， 2021+。