arXiv:2303.01406v1 [stat.ML] 2 Mar 2023

# 弱依赖下的稀疏惩罚深度神经网络估计量

2023年3月3日

William Kengne [1] 和 Modou Wade [2]

*THEMA，CY Cergy Paris Universit´e，法国 95011 Cergy-Pontoise Cedex，Boulevard du Port 33号 电子邮箱：william.kengne@cyu.fr；modou.wade@cyu.fr*

**摘要**：本文研究了 $\psi$-弱依赖过程的非参数回归和分类问题。这种弱依赖结构比混合、关联等条件更为一般。。。。本文采用一种针对稀疏深度神经网络的惩罚估计方法。对于非参数回归和二元分类问题，我们建立了稀疏惩罚深度神经网络估计量超额风险的预言机不等式，并推导了这些估计量超额风险的收敛速率。所展示的模拟结果表明，总体而言，所提出的估计量比未惩罚估计量表现更好。

关键词： 深度神经网络， $\psi$-弱依赖，稀疏性，惩罚，收敛速率。

## 1 引言

深度学习在文献中受到广泛关注，并在人工智能的若干应用领域取得了巨大成功，例如图像处理（参见（[15]））、语音识别（参见（[7]））
$\cdots$。近几十年来，许多研究者致力于理解采用稀疏正则化的深度神经网络（DNN）预测器的理论性质。例如，[18], [20], [1], [21],[22], [12] [9]（及其中的参考文献）给出了一些关于独立同分布（i.i.d.）观测的结果，而 [2], [14], [16], [17], [11], [10] 则给出了一些关于依赖观测或非 i.i.d. 观测的结果。除其他研究外， [19], [16] 研究了稀疏惩罚深度神经网络估计量。这些作者在混合条件下，为 i.i.d. 数据和依赖观测建立了一些预言机型不等式。

考虑观测值 $D_n \coloneqq \{(X_1,\ Y_1)\ \cdots,\ (X_n,\ Y_n)\}$（训练样本），这些观测值来自平稳遍历过程 $\{Z_t = (X_t,\ Y_t)\ t \in \mathbb{Z}\}$，其取值于 $\mathcal{Z} = (\mathcal{X} \times \mathcal{Y})$，其中 $\mathcal{X}$ 是输入空间， $\mathcal{Y}$ 是输出空间。下文考虑损失函数 $\ell$ : $\mathbb{R} \times \mathcal{Y} \to [0, \infty)$ 。

[1]本项目在 ANR-17-CE26-0001-01 和 CY卓越倡议（“未来投资计划”资助项目 ANR-16-IDEX-0008）及“EcoDep”项目 PSI-AAP2020-0000000013 的支持下开展[2]并获得 MME-DII卓越中心（ANR-11-LABEX-0023-01）的支持

我们基于惩罚经验风险最小化程序，针对非参数回归和分类任务构建稀疏深度神经网络预测器。估计量 $h_n \in \mathcal{H}_\sigma(L_n, N_n, B_n, F)$也称为稀疏惩罚深度神经网络（SPDNN）预测器，其定义为

$$\widehat{h}_n = \underset{h\in\mathcal{H}_\sigma(L_n,N_n,B_n,F)}{\operatorname{argmin}} \left[\frac{1}{n}\sum_{i=1}^{n}\ell(h(X_i),Y_i) + J_n(h)\right], \tag{1.1}$$

其中，$\mathcal{H}_\sigma(L_n, N_n, B_n, F)$是一个深度神经网络类（见 (3.3)），采用激活函数 $\sigma$，并具有适当选取的架构参数 $L_n, N_n, B_n, F$；$J_n(h)$是给定如下的稀疏惩罚：

$$J_n(h) \coloneqq J_{\lambda_n,\tau_n}(h) \coloneqq \lambda_n \|\theta(h)\|_{\mathrm{clip},\tau_n},$$

其中，调优参数为 $\lambda_n > 0$，$\tau_n > 0$，$\theta(h)$是 $h$的参数向量。$\|\cdot\|_{\mathrm{clip},\,\tau}$ 表示截断阈值为 $\tau > 0$ 的截断$L_1$ 范数，见 [24]，其定义如下：

$$\|\theta\|_{\mathrm{clip},\tau} = \sum_{j=1}^{p}\left(\frac{|\theta_j|}{\tau}\wedge 1\right),$$

其中，$\theta = (\theta_1, \cdots, \theta_p)'$是一个 $p$维向量，$'$表示转置。

对于弱依赖下的非参数回归和分类，我们重点研究网络类参数 $L_n, N_n, B_n, F$ 以及惩罚项参数 $\lambda_n$、$\tau_n$ 的校准，使 SPDNN 预测器 $h_n$ 具备预言机性质，并推导其收敛速率。 [19] 已在独立同分布情形下研究了这些问题。他们建立了超额风险的预言机不等式，并证明 SPDNN 能够自适应地达到极小极大最优性。 [16] 研究了 $\beta$-混合条件下的非参数时间序列回归 SPDNN 估计量。他们给出了 SPDNN 估计量的泛化误差界，并证明该估计量达到极小极大最优速率（相差一个多对数因子）。除了时间序列分类问题之外，本文考虑的 $\psi$ 弱依赖比混合条件更为一般（见 [3]）。

在这项新的研究中，我们考虑用于学习一个取值于 $Z = \mathcal{X}\times\mathcal{Y} \subseteq \mathbb{R}^d \times \mathbb{R}$ 的 $\psi$-弱依赖过程 $\{Z_t = (X_t, Y_t)\ t \in \mathbb{Z}\}$ 的 SPDNN 估计量，该估计量基于训练样本 $D_n = (X_1, Y_1) \cdots, (X_n, Y_n)$，并探讨以下问题。

(i) **非参数时间序列回归的预言机不等式**。我们给出关于 $L_n, N_n, B_n, F, \lambda_n, \tau_n$ 的条件，并建立 SPDNN 估计量 $L_2$ 误差的预言机不等式。(ii) **二元时间序列分类的预言机不等式**。我们给出关于 $L_n, N_n, B_n, F, \lambda_n, \tau_n$ 的条件，并建立 SPDNN 估计量超额风险的预言机不等式。(iii) **超额风险的收敛速率**。对于非参数回归和时间序列分类，我们推导出超额风险的收敛速率（该速率取决于 $L_n, N_n, B_n, F, \lambda_n, \tau_n$）。当回归问题中的真实回归函数和分类任务中的目标函数足够光滑时，这些速率接近于 $\mathcal{O}(n^{-1/2})$。

本文其余部分安排如下。第 2 节介绍一些记号和假设。第 3 节定义所考虑的深度神经网络类。第 4 节致力于非参数回归，而第 5 节则聚焦于二元时间序列分类。第 6 节给出了一些模拟结果，第 7 节则致力于证明主要结果。

# 2 符号与假设

对于分别配备范数 $\|\cdot\|_{E_1}$ 和 $\|\cdot\|_{E_2}$ 的两个可分巴拿赫空间 $E_1, E_2$，记 $\mathcal{F}(E_1, E_2)$ 为从 $E_1$ 到 $E_2$ 的可测函数集合。对于任意 $h \in \mathcal{F}(E_1, E_2)$ 和 $\epsilon > 0$，$B(h, \epsilon)$ 表示 $\mathcal{F}(E_1, E_2)$ 中以 $h$ 为中心、半径为 $\epsilon$ 的球，即，

$$B(h, \epsilon) = \big\{ f \in \mathcal{F}(E_1, E_2),\ \|f - h\|_\infty \leq \epsilon \big\},$$

其中，$\|\cdot\|_\infty$ 表示下文定义的上确界范数。令 $\mathcal{H} \subset \mathcal{F}(E_1, E_2)$，$\epsilon$-覆盖数 $\mathcal{N}(\mathcal{H}, \epsilon)$ 是覆盖 $\mathcal{H}$ 所需的半径为 $\epsilon$ 的球的最少数量；也就是说，

$$\mathcal{N}(\mathcal{H}, \epsilon) = \inf \Big\{ m \geq 1 \ : \exists h_1, \cdots, h_m \in \mathcal{H} \text{ such that } \mathcal{H} \subset \bigcup_{i=1}^{m} B(h_i, \epsilon) \Big\}.$$

对于函数 $h\colon\ E_1 \to E_2$ 和 $U \subseteq E_1$，定义为:

$$\|h\|_\infty = \sup_{x \in E_1} \|h(x)\|_{E_2},\ \|h\|_{\infty, U} = \sup_{x \in U} \|h(x)\|_{E_2} \text{ and}$$

$$\mathrm{Lip}_\alpha(h) \coloneqq \sup_{x_1, x_2 \in E_1,\ x_1 \neq x_2} \frac{\|h(x_1) - h(x_2)\|_{E_2}}{\|x_1 - x_2\|_{E_1}^\alpha} \text{ for any } \alpha \in [0, 1].$$

对于任意 $\mathcal{K}_\ell > 0$ 和 $\alpha \in [0, 1]$，$\Lambda_{\alpha, \mathcal{K}_\ell}(E_1, E_2)$（当 $E_2 \subseteq \mathbb{R}$ 时，简记为 $\Lambda_{\alpha, \mathcal{K}_\ell}(E_1)$）是满足 $\|h\|_\infty < \infty$ 且 $\mathrm{Lip}_\alpha(h) \leq \mathcal{K}_\ell$ 的函数集合 $h\colon\ E_1^u \to E_2$，其中某个 $u \in \mathbb{N}$。当 $\alpha = 1$ 时，我们令 $\mathrm{Lip}_1(h) = \mathrm{Lip}(h)$，且 $\Lambda_1(E_1) = \Lambda_{1,1}(E_1, \mathbb{R})$。

我们定 义弱依赖结构，见 [6] 和 [3]。设 $E$ 为一个可分巴拿赫空间 .

定义 2.1。 *An $E-$valued process $(Z_t)_{t \in \mathbb{Z}}$ is said to be $(\Lambda_1(E)\ \psi, \epsilon)-$weakly dependent if there exists a function $\psi\colon [0, \infty)^2 \times \mathbb{N}^2 \to [0, \infty)$ and a sequence $\epsilon = (\epsilon(r))_{r \in \mathbb{N}}$ decreasing to zero at infinity such that, for any $g_1,\ g_2 \in \Lambda_1(E)$ with $g_1\colon E^u \to \mathbb{R},\ g_2\colon E^v \to \mathbb{R}, (u, v \in \mathbb{N})$ and for any $u-$tuple $(s_1, \cdots, s_u)$ and any $v-$tuple $(t_1, \cdots, t_v)$ with $s_1 \leq \cdots \leq s_u \leq s_u + r \leq t_1 \leq \cdots \leq t_v$, the following inequality is fulfilled:*

$$|\mathrm{Cov}(g_1(Z_{s_1}, \cdots, Z_{s_u}), g_2(Z_{t_1}, \cdots, Z_{t_v}))| \leq \psi(\mathrm{Lip}(g_1), \mathrm{Lip}(g_2), u, v)\epsilon(r).$$

例如，选择 $\psi$ 可得到一些众所周知的弱依赖条件。

- $\psi(\mathrm{Lip}(g_1),\ \mathrm{Lip}(g_2)\ u,\ v) = v\mathrm{Lip}(g_2)$：$\theta$-弱依赖，记为 $\epsilon(r) = \theta(r)$；
- $\psi(\mathrm{Lip}(g_1),\ \mathrm{Lip}(g_2)\ u,\ v) = u\mathrm{Lip}(g_1) + v\mathrm{Lip}(g_2)$：$\eta$-弱依赖，记为 $\epsilon(r) = \eta(r)$；
- $\psi(\mathrm{Lip}(g_1),\ \mathrm{Lip}(g_2)\ u,\ v) = uv\mathrm{Lip}(g_1) \cdot \mathrm{Lip}(g_2)$：$\kappa$-弱依赖，记为 $\epsilon(r) = \kappa(r)$；
- $\psi(\mathrm{Lip}(g_1),\ \mathrm{Lip}(g_2)\ u,\ v) = u\mathrm{Lip}(g_1) + v\mathrm{Lip}(g_2) + uv\mathrm{Lip}(g_1) \cdot \mathrm{Lip}(g_2)$：$\lambda$-弱依赖，记为 e $\epsilon(r) = \lambda(r)$。

我们考虑取值于 $\mathcal{Z} = \mathcal{X} \times \mathcal{Y} \subset \mathbb{R}^d \times \mathbb{R}$ 的过程 $\{Z_t = (X_t, Y_t)\ t \in \mathbb{Z}\}$，损失函数$\ell\colon\ \mathbb{R} \times \mathcal{Y} \to [0, \infty)$、具有激活函数 $\sigma\colon\ \mathbb{R} \to \mathbb{R}$ 的深度神经网络类 $\mathcal{H}_\sigma(L_n, N_n, B_n, F)$，以及以下假设集。

**(A1)**：存在一个常数 $C_\sigma > 0$，使得激活函数 $\sigma \in \Lambda_{1,\ C_\sigma}(\mathbb{R})$。

**(A2)**：存在 $\mathcal{K}_\ell > 0$，使得损失函数 $\ell \in \Lambda_{1,\ \mathcal{K}_\ell}(\mathbb{R}\times\mathcal{Y})$ 以及 $M = \sup_{h\in\mathcal{H}_\sigma(L_n,\ N_n,\ B_n,\ F)}\sup_{z\in\mathcal{Z}} |\ell(h,\ z)| < \infty$。对于基于间隔的损失，假设这些条件适用于函数 $(u,\ y) \mapsto \ell(uy)$。

我们给出弱依赖假设。

**(A3)**：过程 $\{Z_t = (X_t,\ Y_t)\ t \in \mathbb{Z}\}$ 是平稳遍历的，并且具有 $(\Lambda_1(\mathcal{Z})\ \psi,\ \epsilon)$-弱依赖性，其中 $\epsilon_r = \mathcal{O}(r^{-\gamma})$ 对某个 $\gamma > 3$ 成立。

# 3 深度神经网络

具有 $(L,\ \mathbf{p})$ 网络架构的 DNN，其中 $L \in \mathbb{N}$ 是隐藏层的数量，且 $\mathbf{p} = (p_0,\ \cdots,\ p_{L+1}) \in \mathbb{N}^{L+2}$ 是宽度向量，是任意具有如下形式的函数，

$$h : \mathbb{R}^{p_0} \to \mathbb{R}^{p_{L+1}},\ x \mapsto h(x) = A_{L+1} \circ \sigma_L \circ A_L \circ \sigma_{L-1} \circ \cdots \circ \sigma_1 \circ A_1(x), \tag{3.1}$$

其中 $A_j$：$\mathbb{R}^{p_{j-1}} \to \mathbb{R}^{p_j}$ 是线性仿射映射，由 $A_j(x) := W_j x + \mathbf{b}j$ 定义，其中给定 $p_{j-1}\times p_j$ 权重矩阵$W_j$ 和偏移向量 $\mathbf{b}_j \in \mathbb{R}^{p_j}$；$\sigma_j$：$\mathbb{R}^{p_j} \to \mathbb{R}^{p_j}$ 是非线性逐元素激活映射，由 $\sigma_j(z) = (\sigma(z_1)\ \cdots,\ \sigma(z_{p_j}))'$ 定义。对于形如 (3.1) 的 DNN，记作，

$$\theta(h) := \left(vec(W_1)', \mathbf{b}_1', \cdots, vec(W_{L+1})', \mathbf{b}_{L+1}'\right)', \tag{3.2}$$

其参数向量，其中 $vec(W)$ 通过拼接矩阵 $W$ 的列向量，将其转换为相应的向量。令 $\mathcal{H}_{\sigma,\ p_0,\ p_{L+1}}$ 表示一类深度神经网络预测器，其输入维度为 $p_0$，输出维度为 $p_{L+1}$，并使用激活函数 $\sigma : \mathbb{R} \to \mathbb{R}$。在这里的设定中，$p_0 = d$ 和 $p_{L+1} = 1$。对于式 (3.1) 中的深度神经网络 $h$，分别以 $\text{depth}(h)$ 和 $\text{width}(h)$ 表示其深度和宽度；也就是说，$\text{depth}(h) = L$，且 $\text{width}(h) = \max_{1\le j\le L} p_j$。对于任意正常数 $L,\ N,\ B$ 和 $F$，我们定义

$$\mathcal{H}_\sigma(L, N, B) := \left\{h \in \mathcal{H}_{\sigma,q,1} : \text{depth}(h) \le L, \text{width}(h) \le N, \|\theta(h)\|_\infty \le B\right\},$$

and

$$\mathcal{H}_\sigma(L, N, B, F) := \left\{h : h \in H_\sigma(L, N, B), \|h\|_{\infty,\mathcal{X}} \le F\right\}. \tag{3.3}$$

稀疏度为 $S > 0$ 的一类稀疏性约束深度神经网络定义为

$$\mathcal{H}_\sigma(L, N, B, F, S) := \{h \in \mathcal{H}_\sigma(L, N, B, F)\ :\ \|\theta(h)\|_0 \le S\}, \tag{3.4}$$

其中，对于所有 $x = (x_1,\ .....,\ x_p)' \in \mathbb{R}^p (p \in \mathbb{N})$，有 $\|x\|_0 = \sum_{i=1}^p \mathbb{1}(x_i \ne 0)$。接下来，我们将在 $\psi$-弱依赖条件下，针对回归和分类问题建立稀疏惩罚深度神经网络估计量 (1.1) 的若干理论结果。

# 4 非参数回归

本节旨在研究非参数时间序列回归，其中输出 $Y_t \in \mathbb{R}$ 和输入$X_t \in \mathbb{R}^d$ 由以下模型生成，

$$Y_t = h^*(X_t) + \epsilon_t,\ X_0 \sim \mathrm{P}_{X_0}, \tag{4.1}$$

其中 $h^*:\mathbb{R}^d \to \mathbb{R}$ 是未知的回归函数， $\epsilon_t$ 是独立于输入变量 $X_t$的误差变量。我们对误差作如下次高斯假设，

$$\mathbb{E}[e^{t\epsilon_t}] \le e^{t^2\rho^2/2}, \tag{4.2}$$

对于任意 $t\in\mathbb{R}$，存在某个 $\rho>0$。因此，用 $\mathcal{H}_{\rho,H^*}$ 表示满足模型（**4.1** 的分布集合（$X_0,Y_0$） ）

$$\mathcal{H}_{\rho,H^*} \coloneqq \left\{\text{Model(4.1)} : \mathbb{E}[e^{t\epsilon_t}] \le e^{t^2\rho^2/2}, \forall t\in\mathbb{R}, \|h^*\|_\infty \le H^*\right\}. \tag{4.3}$$

我们将重点研究基于训练样本 $\{（X_i,Y_i）\}_{1\le i\le n}$ 对未知回归函数 $h^*$ 的估计，其中（$X_0,Y_0$） $\sim H$ ，且 $H\in\mathcal{H}_{\rho,H^*}$。我们采用 $L_2$ 误差来衡量估计量 $h_n$ 的性能，

$$\mathbb{E}\left[\|\widehat{h}_n - h^*\|^2_{2,P_{X_0}}\right],$$

其中

$$\|\widehat{h}_n - h^*\|^r_{r,P_{X_0}} \coloneqq \int |\widehat{h}_n(\mathrm{x}) - h^*(\mathrm{x})|^r dP_{X_0}(\mathrm{x}),$$

对于所有 $r\ge 1$。

下文中，我们采用如下记号： $a\vee b = \max(a,\ b)$， $a_n \lesssim b_n$或 $b_n \gtrsim a_n$，若存在一个正常数$C>0$ 使得对于所有 $n\in\mathbb{N}$都有 $a_n \le Cb_n$ ； 若 $a_n \lesssim b_n$ ， 则 $a_n \asymp b_n$ $a_n \gtrsim b_n$。 下面的定理给出了 SPDNN 估计量的 $L_2$ 误差的预言机不等式。

**定理 4.1**。 假设 ***(A1)-(A3)*** 成立，且$(X_0,Y_0)\sim H\in\mathcal{H}_{\rho,\ H^*}$。令 $F>0$, $L_n \lesssim \log n$, $N_n \lesssim n^{\nu_1}$, $1\le B_n \lesssim n^{\nu_2}$ ，其中 $\nu_1>0$, $\nu_2>0$。那么，由下式定义的稀疏惩罚深度神经网络估计量

$$\widehat{h}_n = \underset{h\in\mathcal{H}_\sigma(L_n,N_n,B_n,F)}{\operatorname{argmin}} \left[\frac{1}{n}\sum_{i=1}^n (Y_i - h(X_i))^2 + \lambda_n \|\theta(h)\|_{clip,\tau_n}\right], \tag{4.4}$$

对于 $\lambda_n \asymp$（对数 $n$）$^{\nu_3}/n^{\nu_4}$, $\nu_3>0$, $0<\nu_4<1$, $\tau_n \le \dfrac{\beta_n}{16K_n\ L_n+1}$（）（$(N_n+1)\ B_n$）$^{L_n+1}$对于 $\beta_n \coloneqq$ （对数$n$）$^{\nu_5}/n^{\nu_6}$, 满足

$$\begin{aligned}&\mathbb{E}\left[\|\widehat{h}_n - h^*\|^2_{2,P_{X_0}}\right]\\ &\le 2\left(\inf_{h\in\mathcal{H}_\sigma(L_n,N_n,B_n,F)}\{\|h-h^*\|^2_{2,P_{X_0}} + \lambda_n\|\theta(h)\|_{clip,\tau_n}\} \vee \frac{C_{\sigma,H^*}(\log n)^{\nu_5}}{n^{\nu_6}}\right);\end{aligned} \tag{4.5}$$

对于某个常数$C_{\rho,\ H^*}>0$ 其取决于 $\rho$ 和 $H^*$，且有 $\nu_5>0$, $\nu_6>0$, $\left(\nu_6<1/2, \nu_4+\nu_6<1\right)$ 或 $\left(\nu_6<1/2,\ \nu_4+\nu_6=1, \nu_5>1-\nu_3\right)$。

下一个定理为在多种情况下推导 SPDNN 估计量的收敛速率提供了一个有用工具。

**定理 4.2**。 假设满足定理 ***4.1*** 的条件，并对 $F$、 $L_n$、 $N_n$、 $B_n$、 $\lambda_n$、 $\tau_n$ 作类似选择。此外，令 $H^*>0$ 和 $\mathcal{H}^*$ 为定义在 $\mathbb{R}^d$ 上的一组实值函数，并假设存在常数 $\kappa>0$、 $r>0$、 $\epsilon_0>0$ 和 $C>0$ ，使得

$$\sup_{h^\diamond\in\mathcal{H}^*:\|h^\diamond\|_\infty\le H^*} \inf_{h\in\mathcal{H}_\sigma(L_n,N_n,B_n,F,S_{n,\epsilon})} \|h-h^\diamond\|_{2,P_{X_0}} \le \epsilon, \tag{4.6}$$

对于任意 $\epsilon \in (0、\epsilon_0)$ 以及 $n \in \mathbb{N}$，均有 $S_{n,\epsilon} := C\epsilon^{-\kappa}(\log n)^r$。那么，在（*4.4*）中定义的 *SPDNN* 估计量满足

$$\sup_{H\in\mathcal{H}_{\rho,H^*}:h^*\in\mathcal{H}^*} \mathbb{E}\Big[\|\widehat{h}_n - h^*\|^2_{2,P_{X_0}}\Big] \lesssim \frac{(\log n)^{r+\nu_3}}{n^{\frac{2\nu_4}{\kappa+2}}} \vee \frac{(\log n)^{\nu_5}}{n^{\nu_6}}, \tag{4.7}$$

对于某个常数 $\nu_3 > 0、\ 0 < \nu_4 < 1、\ \nu_5 > 0、\ \nu_6 > 0\ \Big(\nu_6 < 1/2, \nu_4 + \nu_6 < 1\Big)$ 或 $\Big(\nu_6 < 1/2, \nu_4 + \nu_6 = 1, \nu_5 > 1 - \nu_3\Big)$。

**注 4.3**。考虑光滑度为 $s > 0$、半径为 $\mathcal{K} > 0$ 的 *H¨older* 空间，其定义为：

$$\mathcal{C}^{s,\mathcal{K}}(\mathcal{X}) = \big\{h : \mathcal{X} \to \mathbb{R},\ \|h\|_{\mathcal{C}^s(\mathcal{X})} \le \mathcal{K}\big\},$$

其中 $\|\cdot\|_{\mathcal{C}^s(\mathcal{X})}$ 表示如下定义的 *H¨older* 范数：

$$\|h\|_{\mathcal{C}^s(\mathcal{X})} = \sum_{\beta\in\mathbb{N}_0^d,\ |\beta|\le[s]} \|\partial^\beta h\|_\infty + \sum_{\beta\in\mathbb{N}_0^d,\ |\beta|=[s]} Lip_{s-[s]}(\partial^\beta h),$$

其中，对于所有 $\beta = (\beta_1、\ldots、\beta_d)' \in \mathbb{N}_0^d$，$\partial^\beta$ 表示 $\beta$ 阶偏导数，$[x]$ 表示 $x$的整数部分。因此，若真实回归函数例如满足 $h^* \in \mathcal{C}^{s,\mathcal{K}}(\mathcal{X})$，对于某个 $\mathcal{K} > 0$，则条件 *(4.6)* 对各种激活函数类均成立，包括满足 $\kappa = d/s$ 和 $r = 1$ 的 *ReLU*（见 *[11]*）。在此情况下，若 $s \gg d$，则 *SPDNN* 估计量的收敛速率接近于 $\mathcal{O}(n^{-1/2})$。

# 5 二元时间序列分类

我们考虑过程 $\{Z_t = (X_t, Y_t)\ t \in \mathbb{Z}\}$的二元分类，其取值于 $\mathbb{R}^d \times \{-1, 1\}$。目标是构造一个函数 $h$：$\mathbb{R}^d \to \mathbb{R}$，使得 $h(X_t)$ 用于预测标签 $Y_t \in \{-1, 1\}$。我们将重点关注一种基于间隔的损失函数，通过 $h$评估预测性能；也就是说，损失函数的形式为：$(u, y) \mapsto \ell(uy)$。例如，带有 $\ell(uy) = \max(1 - uy, 0)$ 的合页损失。下文假设输入 $X_t \in \mathbb{R}^d$ 和输出 $Y_t \in \{-1, 1\}$ 由该模型生成

$$Y_t|X_t = \mathrm{x} \sim 2\mathcal{B}(\eta(\mathrm{x})) - 1, \qquad X_0 \sim \mathrm{P}_{X_0}, \tag{5.1}$$

其中，$\eta(\mathrm{x}) = P(Y_t = 1|X_t = \mathrm{x})$，且 $\mathcal{B}(\eta(\mathrm{x}))$ 是参数为 $\eta(\mathrm{x})$ 的伯努利分布。目标是构造一个分类器 $h$，使得由下式定义的 $h$ 的超额风险为，

$$\mathcal{E}_{Z_0}(h) := \mathbb{E}[\ell(Y_0 h(X_0))] - \mathbb{E}[\ell(Y_0 h^*_\ell(X_0))],\ \text{with}\ Z_0 := (X_0, Y_0) \tag{5.2}$$

接近于零，其中 $\ell$ 是给定的基于间隔的损失函数，$h^*_\ell = \operatorname{argmin}_{h\in\mathcal{F}} \mathbb{E}[\ell(Y_0 h(X_0))]$ 是目标函数，$\mathcal{F}$ 假设有界，即对于某个 $H^* > 0$ 和 $\mathcal{F}$，$\|h^*_\ell\|_\infty \le H^*$，$\mathcal{F}$ 是从 $\mathbb{R}^d$ 到 $\mathbb{R}$ 的可测函数集合。下文将满足此假设的分布集合定义为，

$$\mathcal{Q}_{H^*} := \{Model(5.1) : \|h^*_\ell\|_\infty \le H^*\}. \tag{5.3}$$

以下定理给出了基于损失函数 $\ell$的 SPDNN 估计量超额风险的预言机不等式；其中无需像 [19] 那样要求严格凸性。

**定理 5.1**。 假设 ***(A1)-(A3)*** 成立，且真实生成模型 $H$ 属于 $\mathcal{Q}_{H^*}$。令 $F>0, L_n \lesssim \log n, N_n \lesssim n^{\nu_1}$, $1 \leq B_n \lesssim n^{\nu_2}$，其中 $\nu_1>0$, $\nu_2>0$。那么，由下式定义的 ***SPDNN*** 估计量，

$$\widehat{h}_n = \underset{h \in \mathcal{H}_\sigma(L_n, N_n, B_n, F)}{\operatorname{argmin}} \left[ \frac{1}{n} \sum_{i=1}^{n} \ell(Y_i h(X_i)) + \lambda_n \|\theta(h)\|_{clip,\tau_n} \right], \tag{5.4}$$

其中 $\lambda_n \asymp$（对数 $n$）$^{\nu_3}/n$, $\tau_n \leq \frac{\beta_n}{4\mathcal{K}_\ell L_n + 1}$（）$((N_n+1) B_n)^{L_n+1}$ 对于 $\beta_n :=$（对数 $n$）$^{\nu_5}/n^{\nu_6}$, $\mathcal{K}_\ell>0$, 满足

$$\mathbb{E}\left[\mathcal{E}_{Z_0}(\widehat{h}_n)\right] \leq 2 \inf_{h \in \mathcal{H}_\sigma(L_n, N_n, B_n, F)} \{\mathcal{E}_{Z_0}(h) + \lambda_n \|\theta(h)\|_{clip,\tau_n}\} \vee \frac{C(\log n)^{\nu_5}}{n^{\nu_6}}, \tag{5.5}$$

对于某些普适常数 $C>0$, $\nu_5>0$, $\nu_6>0$, $\Big(\nu_6<1/2, \nu_4+\nu_6<1\Big)$ 或 $\Big(\nu_6<1/2, \nu_4+\nu_6=1, \nu_5>1-\nu_3\Big)$，其中期望是针对训练数据取的。

与非参数回归中的定理 4.2 类似，以下定理为推导 SPDNN 估计量的收敛速率提供了有用工具。

**定理 5.2**。 假设满足定理 *5.1* 的条件，并对 $F, L_n, N_n, B_n, \lambda_n, \tau_n$ 作出类似选择。令$H^*>0$，并令 $\mathcal{H}^*$ 为定义在 $\mathbb{R}^d$ 上的实值函数集合，且假设存在常数 $\kappa>0$, $r>0, \epsilon_0>0$ 和 $C>0$，使得，

$$\sup_{h^\diamond \in \mathcal{H}^*: \|h^\diamond\|_\infty \leq H^*} \inf_{h \in \mathcal{H}_\sigma(L_n, N_n, B_n, F, S_{n,\epsilon})} \|h - h^\diamond\|_{1, P_{X_0}} \leq \epsilon, \tag{5.6}$$

其中 $S_{n,\ \epsilon} := C\epsilon^{-\kappa}(\log n)^r$ 对于任意 $\epsilon \in (0,\ \epsilon_0)$ 以及 $n \in \mathbb{N}$。那么，在 ***(5.4)*** 中定义的 ***SPDNN*** 估计量满足，

$$\sup_{H \in \mathcal{Q}_{H^*}: h_\ell^* \in \mathcal{H}^*} \mathbb{E}\left[\mathcal{E}_{Z_0}(\widehat{h}_n)\right] \lesssim \frac{(\log n)^{r+\nu_3}}{n^{\frac{\nu_4}{\kappa+1}}} \vee \frac{(\log n)^{\nu_5}}{n^{\nu_6}}. \tag{5.7}$$

正如注 4.3 中强调的，这一结果可用于在多种情形下推导 SPDNN 估计量的超额风险收敛速率。

# 6 数值结果

本节使用 SPDNN 对自回归模型和二元时间序列进行预测。

## 6.1 自回归模型的预测

考虑一个带有外生协变量的非线性自回归过程 $(Y_t,\ \mathcal{X}_t)_{t\in\mathbb{Z}}$，其取值在 $\mathbb{R}\times\mathbb{R}$ 中，并满足：

$$Y_t = f(Y_{t-1}, \ldots, Y_{t-p}; \mathcal{X}_{t-1}) + \epsilon_t, \tag{6.1}$$

对于某个可测函数 $f: \mathbb{R}^{p+1} \to \mathbb{R}$ $(p \in \mathbb{N})$，其中 $(\epsilon_t)_{t\in\mathbb{Z}}$ 是由标准化均匀分布 $\mathcal{U}[-2$生成的独立同分布序列；$2]$ 而 $(\mathcal{X}_t)_{t\in\mathbb{Z}}$ 是由标准化的 $\mathcal{U}[-2$生成的一阶自回归过程；$2]$ 创新项。式（6.1）中的过程 $(Y_t,\ \mathcal{X}_t)_{t\in\mathbb{Z}}$ 是 [4] 所研究的带外生协变量的仿射因果模型的一个具体例子。下文中，我们令 $X_t = (Y_{t-1},\ \cdots,\ Y_{t-p},\ \mathcal{X}_{t-1})$。因此可以看出，这个非线性自回归过程是模型（4.1）的一个特例。如果 $f \in \Lambda_1(\mathbb{R}^{p+1})$（也就是说，$f$ 是利普希茨的），并且

在 $f$ 的利普希茨型系数满足某些经典条件时（例如， $Lip(f) < 1/p$），那么，式（6.1）存在一个解 $(Y_t, X_t)_{t\in\mathbb{Z}}$，满足上述假设 **(A3)**；详情见 [5]。

设 $(Y_1, \mathcal{X}_1) \cdots, (Y_n, \mathcal{X}_n)$ 是过程 $(Y_t, \mathcal{X}_t)_{t\in\mathbb{Z}}$ 的一条轨迹。我们旨在根据该训练样本预测 $Y_{n+1}$。我们使用上文构建的 SPDNN 预测器进行学习理论研究，其中输入变量为 $X_t = (Y_{t-1}, \mathcal{X}_{t-1})$，输入空间为 $\mathcal{X} \subset \mathbb{R}^p \times \mathbb{R}$，输出空间为 $\mathcal{Y} \subset \mathbb{R}$。我们考虑 (6.1) 中的以下情形:

$$\begin{aligned}
&\text{DGP1}: \quad Y_t = 1 - 0.2Y_{t-1} + 0.3Y_{t-2} + 0.25Y_{t-3} - 0.6\frac{1}{1+\mathcal{X}_{t-1}^2} + \epsilon_t;\\
&\text{DGP2}: \quad Y_t = 0.5 + \big(-0.4 + 0.25e^{-2Y_{t-1}^2}\big)Y_{t-1} + 1.5\mathcal{X}_{t-1} + \epsilon_t.
\end{aligned}$$

DGP1 是一种经典的带非线性协变量的自回归模型，而 DGP2 是一种带协变量的指数自回归模型。

对于每个数据生成过程，我们采用具有 2 个隐藏层的网络架构，每层包含 100 个隐藏节点。隐藏层和输出层分别使用 ReLU 激活函数和线性激活函数。使用算法 Adam ([13])、学习率 $10^{-3}$ 和 32 的小批量大小，通过 R软件中的 Keras 软件包训练网络权重。当均方误差 (MSE) 连续 30 个训练轮次未得到改善时，便停止训练。

对于 $n = 250$，500 和 1000，依据真实数据生成过程生成一条轨迹 $((Y_1, \mathcal{X}_1), (Y_2, \mathcal{X}_2) \ldots, (Y_n, \mathcal{X}_n))$。预测器 $\widehat{h}_n$ 由 (4.4) 得出，其调优参数形式为 $\lambda_n = 10^{-i}\log(n)/n$ 和$\tau_n = 10^{-j}/\log(n)$，其中 $i, j = 0, 1, \ldots,$ 10 通过最小化验证数据集$\widehat{\ }((Y_1', \mathcal{X}_1'), (Y_2', \mathcal{X}_2') \ldots, (Y_n', \mathcal{X}_n'))$ 上的均方误差进行校准。随后，基于一条新的轨迹 $((Y_1'', \mathcal{X}_1''), (Y_2'', \mathcal{X}_2'') \ldots, (Y_m'', \mathcal{X}_m''))$ 计算 $h_n$ 的经验 $L_2$ 误差，其中 $m = 10^4$。

图 1 展示了在 100 次重复试验中，SPDNN 和非惩罚深度神经网络（NPDNN，由式 (4.4) 在 $\lambda_n = 0$ 条件下得到）预测器的经验 $L_2$ 误差箱线图。可以看出，在 DGP1 中，稀疏惩罚深度神经网络估计量的表现优于非惩罚深度神经网络估计量。在数据生成过程 2 中，稀疏惩罚深度神经网络估计量的表现略优于非惩罚深度神经网络估计量。这些发现表明，与非惩罚深度神经网络估计量相比，稀疏惩罚深度神经网络估计量能够改善$L_2$ 误差。

## 6.2 二元时间序列的预测

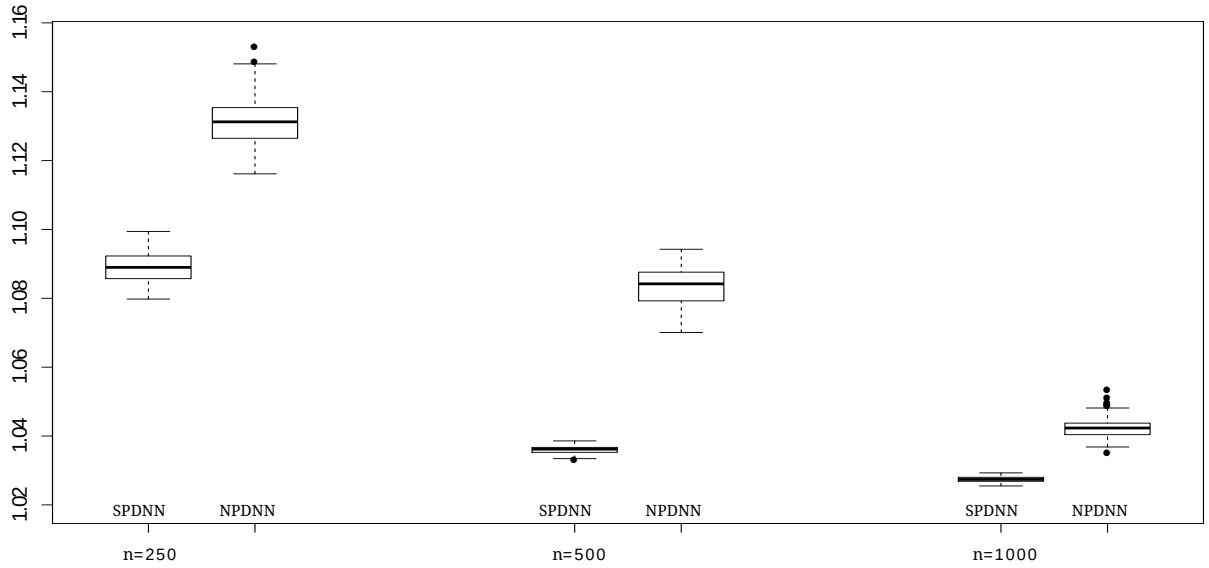
考虑一个带有外生协变量的二元自回归过程（$Y_t$， $\mathcal{X}_t$），$_{t\in\mathbb{Z}}$其取值属于 $\{-1, 1\} \times \mathbb{R}$，并满足:

$$Y_t|\mathcal{F}_{t-1} \sim 2\mathcal{B}(p_t) - 1 \text{ with } 2p_t - 1 = \mathbb{E}[Y_t|\mathcal{F}_{t-1}] = f(Y_{t-1}, \cdots, Y_{t-p}, ; \mathcal{X}_{t-1}), \tag{6.2}$$

其中 $\mathcal{F}_{t-1} = \sigma\{Y_{t-1}, \cdots; \ \mathcal{X}_{t-1}, \ \cdots\}$ 对于某个可测函数 $f$: $\mathbb{R}^{p+1} \to [-1, 1]$ $(p \in \mathbb{N})$，$(\mathcal{X}_t)_{t\in\mathbb{Z}}$ 是一阶自回归过程，且 $\mathcal{B}(p_t)$ 表示参数为 $p_t$ 的伯努利分布。令 $X_t = (Y_{t-1}, \ \cdots, \ Y_{t-p}, \ \mathcal{X}_{t-1})$。因此，可以看出，二元模型 (6.2) 是 (5.1) 的一个特例。在 $f$ 满足某个经典的利普希茨型条件时，过程 $(Y_t, \ \mathcal{X}_t)_{t\in\mathbb{Z}}$ 满足弱依赖假设 **(A3)**，见 [11]。

令 $(Y_1, \ \mathcal{X}_1) \cdots, (Y_n, \ \mathcal{X}_n)$ 为过程 $(Y_t, \ \mathcal{X}_t)_{t\in\mathbb{Z}}$ 的一条轨迹，目标是根据该训练样本预测 $Y_{n+1}$。我们使用本文提出的 SPDNN 预测器和 $p = 1, \quad X_t =$进行学习理论分析

（a）DGP1 中经验 $L_2$ 误差的箱线图




（b）DGP2 中经验 $L_2$ 误差的箱线图

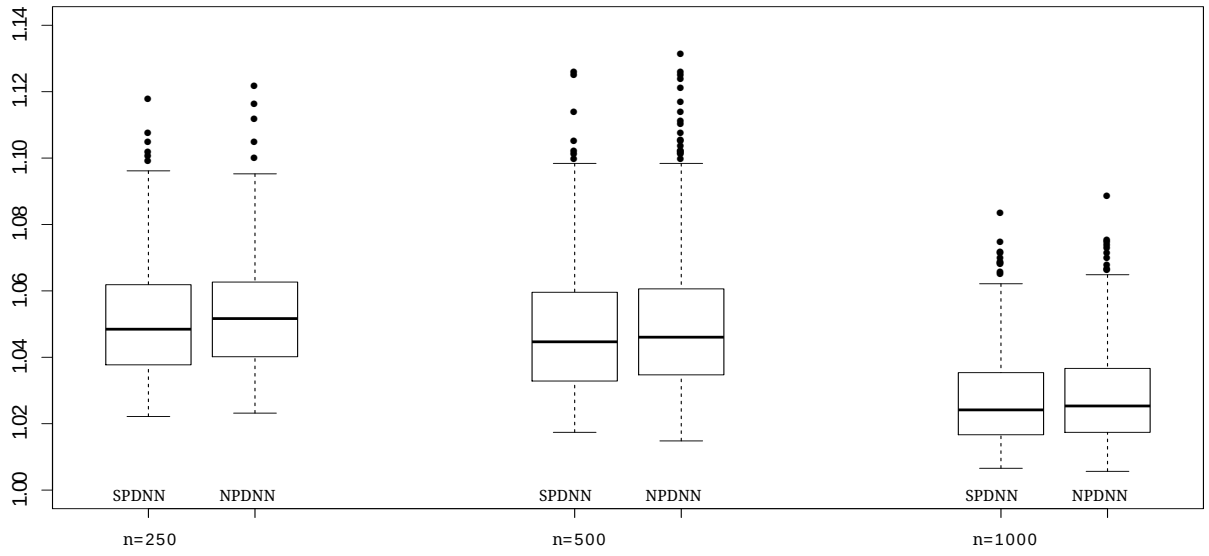



图 1： *SPDNN* 和 *NPDNN* 预测器在 *DGP1*（*a*）和 *DGP2*（*b*）中 $n = 250$、500 和 *1000* 时的经验 $L_2$ 误差箱线图。

在 DGP3 中为 $(Y_{t-1},\ \mathcal{X}_{t-1})$，在 DGP4 中为 $p = 2,\quad X_t = (Y_{t-1},\ Y_{t-2},\ \mathcal{X}_{t-1})$。我们研究 (6.2) 的以下情形：

$$\begin{aligned}
&\text{DGP3}: \quad f(Y_{t-1};\mathcal{X}_{t-1}) = -0.15 + \left(0.1 - 0.2e^{-0.5Y_{t-1}^2}\right)Y_{t-1} + 0.25\frac{1}{1+\mathcal{X}_{t-1}^2}\\
&\text{DGP4}: \quad f(Y_{t-1},Y_{t-2};\mathcal{X}_{t-1}) = 0.1 + 0.15Y_{t-1} - 0.25Y_{t-2} - 0.2e^{-\mathcal{X}_{t-1}^2}.
\end{aligned}$$

以下考虑合页损失函数（$\ell(z) = \max(1-z,0)$），并定义

$$h_\ell^*(X_t) = 2\mathbb{1}_{\{f(X_t)\geq 0\}} - 1 \text{ for all } t \in \mathbb{Z}. \tag{6.3}$$

可以看出， $h_\ell^*$ 是关于此损失的贝叶斯分类器；也就是说

$$\mathcal{E}_{Z_0}(h_\ell^*) = \inf_{h\in\mathcal{F}(\mathcal{X},\mathcal{Y})} \mathcal{E}_{Z_0}(h),$$

其中， $\mathcal{E}_{Z_0}(\cdot)$在（5.2）中定义，而 $\mathcal{F}(\mathcal{X},\mathcal{Y})$是从 $\mathcal{X}$到 $\mathcal{Y}$的可测函数类。此外，对于这些数据生成过程中的每一个，我们都采用与第（6.1）小节相同的网络架构。

对于 $n = 250$，500和1000，从真实数据生成过程中生成一条轨迹（$(Y_1,\mathcal{X}_1)\ldots,(Y_n,\mathcal{X}_n)$)。预测器 $h_n$ 由（5.4）得到，其中调优参数 $\lambda_n$, $\tau_n$ 按照第（6.1）小节的方法，基于验证数据集（$(Y_1',\mathcal{X}_1'),(Y_2',\mathcal{X}_2')\ldots,(Y_n',\mathcal{X}_n')$)进行选择。 $h_n$的经验超额风险根据一条新的轨迹（$(Y_1'',\mathcal{X}_1''),(Y_2'',\mathcal{X}_2'')\ldots,(Y_m'',\mathcal{X}_m'')$)计算，其中 $m = 10^4$。

图2展示了在100次重复试验中，SPDNN预测器和NPDNN预测器的经验超额风险箱线图。可以看出，SPDNN的整体表现优于NPDNN。这再次表明，与NPDNN估计量相比，SPDNN估计量能够提高预测准确性。

# 7 主要结果的证明

## 7.1 定理4.1的证明

在证明中，若存在一个仅依赖于 $\rho$ 和 $H^*$ 的常数$C_{\rho,H^*} > 0$ ，使得$a_n \leq C_{\rho,H^*} b_n$ 对任意 $n \in \mathbb{N}$均成立，则记为$a_n \lesssim_{\rho,H^*} b_n$ 。令 $K_n \coloneqq (\sqrt{32}\rho^2(\log n)^{1/2}) \vee \overline{H}^*$。令 $Y_0^\perp \coloneqq \text{sign}(Y_0)(|Y_0| \wedge K_n)$，这是 $Y_0$ 的截断版本；并令 $h^\perp$ 为$Y_0^\perp$的回归函数，即

$$h^\perp(x) \coloneqq \mathbb{E}(Y_0^\perp | X_0 = x).$$

为方便起见，我们在记号 $Y_0^\perp$ 和 $h^\perp$ 中省略对 $n$ 的依赖。与 [19], 中一样，考虑以下分解

$$\|\widehat{h}_n - h^*\|_{2,P_{X_0}}^2 = \mathbb{E}\Big[\big(Y_0 - \widehat{h}_n(X_0)\big)^2\Big] - \mathbb{E}\Big[(Y_0 - h^*(X_0))^2\Big] = \sum_{i=1}^4 A_{i,n}, \tag{7.1}$$

其中,

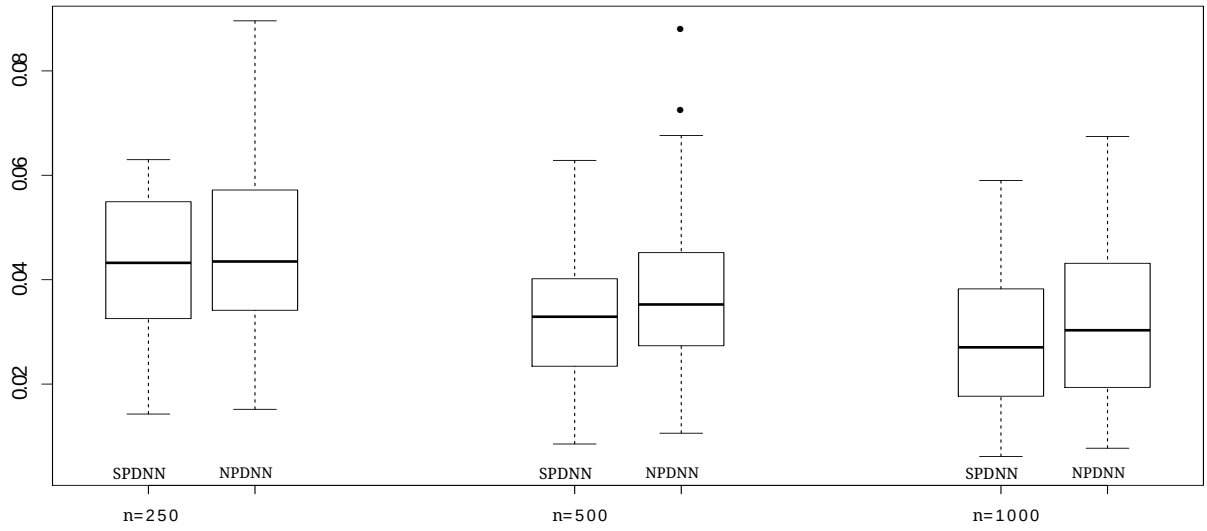
$$A_{1,n} \coloneqq \Big[\mathbb{E}\big(Y_0 - \widehat{h}_n(X_0)\big)^2 - \mathbb{E}\,(Y_0 - h^*(X_0))^2\Big] - \Big[\mathbb{E}\big(Y_0^\perp - \widehat{h}_n(X_0)\big)^2 - \mathbb{E}\,\big(Y_0^\perp - h^\perp(X_0)\big)^2\Big];$$

$$A_{2,n} \coloneqq \Big[\mathbb{E}\big(Y_0^\perp - \widehat{h}_n(X_0)\big)^2 - \mathbb{E}\,\big(Y_0^\perp - h^\perp(X_0)\big)^2\Big] - 2\Big[\frac{1}{n}\sum_{i=1}^n \big(Y_i^\perp - \widehat{h}_n(X_i)\big)^2 - \frac{1}{n}\sum_{i=1}^n \big(Y_i^\perp - h^\perp(X_i)\big)^2\Big] - 2J_{\lambda_n,\tau_n}(\widehat{h}_n);$$

$$A_{3,n} \coloneqq 2\Big[\frac{1}{n}\sum_{i=1}^n \big(Y_i^\perp - \widehat{h}_n(X_i)\big)^2 - \frac{1}{n}\sum_{i=1}^n \big(Y_i^\perp - h^\perp(X_i)\big)^2\Big] - 2\Big[\frac{1}{n}\sum_{i=1}^n \big(Y_i - \widehat{h}_n(X_i)\big)^2 - \frac{1}{n}\sum_{i=1}^n (Y_i - h^*(X_i))^2\Big];$$

$$A_{4,n} \coloneqq 2\left[\frac{1}{n}\sum_{i=1}^n \big(Y_i - \widehat{h}_n(X_i)\big)^2 - \frac{1}{n}\sum_{i=1}^n (Y_i - h^*(X_i))^2\right] + 2J_{\lambda_n,\tau_n}(\widehat{h}_n).$$

（a）DGP3 中经验超额风险的箱线图




(b) DGP4 中经验超额风险的箱线图

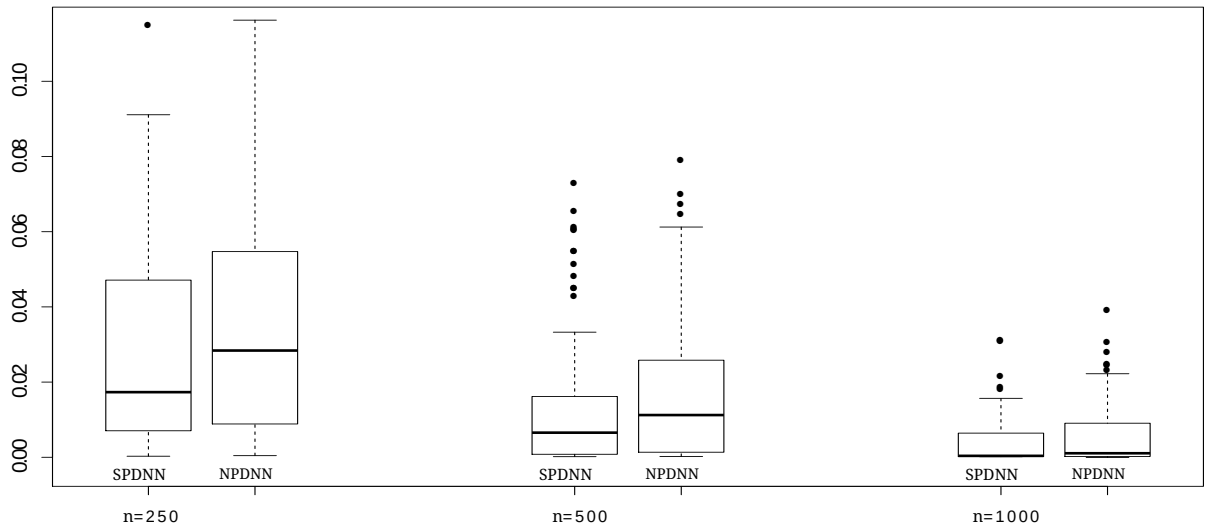



图2： *DGP3*（*a*）和 *DGP4*（*b*）中，*SPDNN* 和非惩罚深度神经网络预测器在 $n = 250$、$500$ 和 *1000* 时的经验超额风险箱线图。

（7.1）中的第一个等式利用了 $X_0$ 与 $\epsilon_0$的独立性。为了界定 $A_{1,n}$，我们回顾一下次高斯变量 $\mathbb{E}(\epsilon_0) = 0$ 和 $\mathbb{E}(e^{\epsilon_0^2/4\rho^2}) \leq \sqrt{2}$的性质，参见 [23]中的定理 2.6。例如，令

$$A_{1,1,n} \coloneqq \mathbb{E}\left((Y_0^{\perp} - Y_0)(2\widehat{h}_n(X_0) - Y_0 - Y_0^{\perp})\right);$$
$$A_{1,2,n} \coloneqq \mathbb{E}\left[(Y_0^{\perp} - h^{\perp}(X_0) - Y_0 + h^*(X_0))(Y_0^{\perp} - h^{\perp}(X_0) + Y_0 - h^*(X_0))\right].$$

可以看出，$A_{1,\ n} = A_{1,\ 1,\ n} + A_{1,\ 2,\ n}$。利用柯西-施瓦茨不等式可得

$$|A_{1,1,n}| \leq \sqrt{E(Y_0^{\perp} - Y_0)^2}\sqrt{\mathbb{E}(2\widehat{h}_n(X_0) - Y_0 - Y_0^{\perp})^2}.$$

我们有，

$$\begin{aligned}(Y_0^{\perp} - Y_0)^2 &= (|Y_0| \wedge K_n)^2 - 2|Y_0|(|Y_0| \wedge K_n) + |Y_0|^2 \\ &= (|Y_0| - K_n)^2 \mathbb{1}(|Y_0| > K_n).\end{aligned}$$

因此，

$$\begin{aligned}\mathbb{E}(Y_0^{\perp} - Y_0)^2 &= \mathbb{E}\left[(|Y_0| - K_n)^2 \mathbb{1}(|Y_0| > K_n)\right] \\ &\leq \mathbb{E}\left[|Y_0|^2 \mathbb{1}(|Y_0| > K_n)\right].\end{aligned} \tag{7.2}$$

根据假设（4.3）以及 $X_0$ 与 $\epsilon_0$的独立性，我们得到，$\mathbb{E}(e^{Y_0^2/(8\rho^2)}) \leq e^{(H^*)^2/(4\rho^2)}\mathbb{E}e^{\epsilon_0^2/(4\rho)} \leq \sqrt{2}e^{(H^*)^2/(4\rho^2)}$。此外，容易看出，$Y_0^2 \leq 16\rho^2 e^{Y_0^2/16\rho^2}$ 以及$\mathbb{1}(|Y_0| > K_n) \leq e^{(Y_0^2 - K_n)/16\rho^2}$。因此，

$$\begin{aligned}\mathbb{E}(Y_0^{\perp} - Y_0)^2 &\leq \mathbb{E}\left[16\rho^2 e^{Y_0^2/(16\rho^2)} e^{Y_0^2/(16\rho^2) - K_n^2/(16\rho^2)}\right] \\ &\leq 16\sqrt{2}\rho^2 e^{(H^*)^2/(4\rho^2)} e^{-2\log n} \leq 16\sqrt{2}\rho^2 e^{(H^*)^2/(4\rho^2)} n^{-2}.\end{aligned} \tag{7.3}$$

我们还有，

$$\mathbb{E}(2\widehat{h}_n(X_0) - Y_0 - Y_0^{\perp})^2 \leq 2\mathbb{E}(Y_0^2) + 2\mathbb{E}(2\widehat{h}_n(X_0) - Y_0^{\perp})^2,$$

and,

$$\begin{aligned}|(2\widehat{h}_n(X_0) - Y_0^{\perp})^2| &\leq 4|\widehat{h}_n(X_0)|^2 + 4|\widehat{h}_n(X_0)||Y_0^{\perp}| + |Y_0^{\perp}|^2 \\ &\leq 4K_n^2 + 4K_n^2 + K_n^2.\end{aligned}$$

由于存在一个常数 $C$，使得 $K_n^2 \leq C\log n$，因此我们有

$$\begin{aligned}\mathbb{E}(2\widehat{h}_n(X_0) - Y_0 - Y_0^{\perp})^2 &\leq 16\rho^2\mathbb{E}(e^{Y^2/(8\rho^2)}) + 18K_n^2 \\ &\lesssim C_{\rho,H^*}\log n.\end{aligned} \tag{7.4}$$

因此，$|A_{1,\ 1,\ n}| \lesssim C_{\rho,\ H^*}\log n/n$。现在来处理 $A_{1,\ 2,\ n}$。利用柯西-施瓦茨不等式，我们得到，

$$|A_{1,2,n}| \leq \sqrt{2\mathbb{E}(Y_0^{\perp} - Y_0)^2 + 2\mathbb{E}(h^{\perp}(X_0) - h^*(X_0))^2} \times \sqrt{\mathbb{E}(Y_0 + Y_0^{\perp} - h^{\perp}(X_0) - h^*(X_0))^2}.$$

以与 (7.4) 类似的方式，我们得到，$\mathbb{E}(Y_0 + Y_0^{\perp} - h^{\perp}(X_0) - h^*(X_0))^2 \lesssim_{C_{\rho,\ H^*}} \log n$。由于 $\mathbb{E}(\epsilon_0) = 0$，根据詹森不等式，可以很容易地得到，

$$\mathbb{E}\left[(h^{\perp}(X_0) - h^*(X_0))^2\right] = \mathbb{E}\left(\mathbb{E}(Y_0^{\perp}|X_0) - \mathbb{E}(Y_0|X_0)\right)^2 \leq \mathbb{E}(Y_0^{\perp} - Y_0)^2.$$

利用 (7.2)，可得$|A_{1,\ 2,\ n}| \lesssim C_{\rho,\ H^*}\log n/n$。现在，我们有：

$$\mathbb{E}[A_{3,n}] = -\frac{2}{n}\sum_{i=1}^{n}\left\{\left[\mathbb{E}[(Y_i - \widehat{h}_n(X_i))^2] - \mathbb{E}[(Y_i - h^*(X_i))^2]\right] - \left[\mathbb{E}[(Y_i^{\perp} - \widehat{h}_n(X_i))^2] - \mathbb{E}[(Y_i^{\perp} - h^{\perp}(X_i))^2]\right]\right\}.$$

对于 $i=1,\ \cdots,\ n$，令：

$$\begin{aligned}
A_{3,n,i} &= \left[\mathbb{E}[(Y_i-\widehat{h}_n(X_i))^2]-\mathbb{E}[(Y_i-h^*(X_i))^2]\right]-\left[\mathbb{E}[(Y_i^{\perp}-\widehat{h}_n(X_i))^2]-\mathbb{E}[(Y_i^{\perp}-h^{\perp}(X_i))^2]\right]\\
A_{3,1,n,i} &= \mathbb{E}\left[(Y_i^{\perp}-Y_i)(2\widehat{h}_n(X_i)-Y_i-Y_i^{\perp})\right]\\
A_{3,2,n,i} &= \mathbb{E}\left[(Y_i^{\perp}-h^{\perp}(X_i)-Y_i+h^*(X_i))(Y_i^{\perp}-h^{\perp}(X_i)+Y_i-h^*(X_i))\right].
\end{aligned}$$

我们有，对于 $i=1,\ \cdots,\ n$，

$$|A_{3,1,n,i}|\le\sqrt{\mathbb{E}\left[(Y_i^{\perp}-Y_i)^2\right]}\sqrt{\mathbb{E}\left[(2\widehat{h}_n(X_i)-Y_i-Y_i^{\perp})^2\right]}.$$

通过采用与 $A_{1,\ 1,\ n}$ 中类似的论证，对于 $i=1,\ \cdots,\ n$，可得

$$|A_{3,1,n,i}|\lesssim C_{\rho,H^*}\log n/n.$$

此外，按照 $A_{1,\ 2,\ n}$ 中的方法，对于 $i=1,\ \cdots,\ n$，有

$$|A_{3,2,n,i}|\le\sqrt{\mathbb{E}\left[(Y_i^{\perp}-h^{\perp}(X_i)-Y_i+h^*(X_i))^2\right]}\sqrt{\mathbb{E}\left[(Y_i^{\perp}-h^{\perp}(X_i)+Y_i-h^*(X_i))^2\right]},$$

并且还可以得到$|A_{3,\ 2,\ n,\ i}|\lesssim C_{\rho,\ H^*}\log\ n/n$。因此，对于 $i=1,\ \cdots,\ n$，

$$|A_{3,n,i}|\lesssim C_{\rho,H^*}\log n/n.$$

因此，

$$|\mathbb{E}(A_{3,n})|=|-\frac{2}{n}\sum_{i=1}^{n}A_{3,n,i}|\le\frac{2}{n}\sum_{i=1}^{n}|A_{3,n,i}|\lesssim C_{\rho,H^*}\log n/n.$$

对于 $A_{2,n}$，定义 $\Delta h(Z_0):=(Y_0^{\perp}-h(X_0))^2-(Y_0^{\perp}-h^{\perp}(X_0))^2$，其中 $Z_0:=(X_0,Y_0)$，适用于 $h\in\mathcal{H}_\sigma(L_n,N_n,B_n,F)$。令 $\alpha>0$，则可写为

$$\begin{aligned}
P(A_{2,n}>\alpha) &\le P\left(\sup_{h\in\mathcal{H}_\sigma(L_n,N_n,B_n,F)}\frac{\mathbb{E}(\Delta h(Z_0))-\frac{1}{n}\sum_{i=1}^{n}\Delta h(Z_i)}{\alpha+2J_{\lambda_n,\tau_n}+\mathbb{E}(\Delta h(Z_0))}\ge\frac{1}{2}\right)\\
&\le\sum_{j=0}^{\infty}P\left(\sup_{h\in\mathcal{H}_{n,j,\alpha}}\frac{\mathbb{E}(\Delta h(Z_0))-\frac{1}{n}\sum_{1=1}^{n}\Delta h(Z_i)}{2^j\alpha+\mathbb{E}(\Delta h(Z_0))}\ge\frac{1}{2}\right),
\end{aligned}$$

其中

$$\mathcal{H}_{n,j,\alpha}:=\left\{h\in\mathcal{H}_\sigma(L_n,N_n,B_n,F):2^{j-1}\mathbb{1}(j\ne0)\alpha\le J_{\lambda_n,\tau_n}(h)\le2^j\alpha\right\}.\tag{7.5}$$

事实上,

$$
\begin{aligned}
&P(A_{2,n}>\alpha)\\
&=P\Bigg(\left[\mathbb{E}\left(Y_0^{\perp}-\widehat{h}_n(X_0)\right)^2-\mathbb{E}\left(Y_0^{\perp}-h^{\perp}(X_0)\right)^2\right]-2\left[\frac{1}{n}\sum_{i=1}^{n}\left(Y_i^{\perp}-\widehat{h}_n(X_i)\right)^2-\frac{1}{n}\sum_{i=1}^{n}\left(Y_i^{\perp}-h^{\perp}(X_i)\right)^2\right]\\
&\qquad\qquad -2J_{\lambda_n,\tau_n}(\widehat{h}_n)>\alpha\Bigg)\\
&\le P\Bigg(2\left[\mathbb{E}\left(Y_0^{\perp}-\widehat{h}_n(X_0)\right)^2-\mathbb{E}\left(Y_0^{\perp}-h^{\perp}(X_0)\right)^2\right]-\left[\frac{1}{n}\sum_{i=1}^{n}\left(Y_i^{\perp}-\widehat{h}_n(X_i)\right)^2-\frac{1}{n}\sum_{i=1}^{n}\left(Y_i^{\perp}-h^{\perp}(X_i)\right)^2\right]\\
&\qquad\qquad >\alpha+2J_{\lambda_n,\tau_n}(\widehat{h}_n)+\mathbb{E}\left(Y_0^{\perp}-\widehat{h}_n(X_0)\right)^2-\mathbb{E}\left(Y_0^{\perp}-h^{\perp}(X_0)\right)^2\Bigg)\\
&\le P\Bigg(\exists h\in\mathcal{H}_\sigma(L_n,N_n,B_n,F):\left[\mathbb{E}\left(Y_0^{\perp}-h(X_0)\right)^2-\mathbb{E}\left(Y_0^{\perp}-h^{\perp}(X_0)\right)^2\right]\\
&\qquad\qquad -\left[\frac{1}{n}\sum_{i=1}^{n}\left(Y_i^{\perp}-h(X_i)\right)^2-\frac{1}{n}\sum_{i=1}^{n}\left(Y_i^{\perp}-h^{\perp}(X_i)\right)^2\right]\\
&\qquad\qquad >\frac{1}{2}\left(\alpha+2J_{\lambda_n,\tau_n}(h(X_0))+\mathbb{E}\left(Y_0^{\perp}-h(X_0)\right)^2-\mathbb{E}\left(Y_0^{\perp}-h^{\perp}(X_0)\right)^2\right)\Bigg)\\
&\le P\Bigg(\exists h\in\mathcal{H}_\sigma(L_n,\ \ N_n,\ \ B_n,\\
&\ F):\\
&\qquad\qquad \frac{\left[\mathbb{E}\left(Y_0^{\perp}-h(X_0)\right)^2-\mathbb{E}\left(Y_0^{\perp}-h^{\perp}(X_0)\right)^2\right]-\left[\frac{1}{n}\sum_{i=1}^{n}\left(Y_i^{\perp}-h(X_i)\right)^2-\frac{1}{n}\sum_{i=1}^{n}\left(Y_i^{\perp}-h^{\perp}(X_i)\right)^2\right]}{\text{上确界}_{\,h\in\mathcal{H}_\sigma(L_n,}\left(\alpha+2J_{\lambda_n,\tau_n}(h(X_0))+\mathbb{E}\left(Y_0^{\perp}-h(X_0)\right)^2-\mathbb{E}\left(Y_0^{\perp}-h^{\perp}(X_0)\right)^2\right)}\ge\frac{1}{2}\Bigg)\\
&\le\sum_{j=0}^{\infty}P\Bigg(\\
&\ N_n,\ B_n,\ F)\\
&\qquad\qquad \frac{\left[\mathbb{E}\left(Y_0^{\perp}-h(X_0)\right)^2-\mathbb{E}\left(Y_0^{\perp}-h^{\perp}(X_0)\right)^2\right]-\left[\frac{1}{n}\sum_{i=1}^{n}\left(Y_i^{\perp}-h(X_i)\right)^2-\frac{1}{n}\sum_{i=1}^{n}\left(Y_i^{\perp}-h^{\perp}(X_i)\right)^2\right]}{\left(2^j\alpha+2J_{\lambda_n,\tau_n}(h(X_0))+\mathbb{E}\left(Y_0^{\perp}-h(X_0)\right)^2-\mathbb{E}\left(Y_0^{\perp}-h^{\perp}(X_0)\right)^2\right)}\ge\frac{1}{2}\Bigg)\\
&\le\sum_{j=0}^{\infty}P\left(\sup_{h\in\mathcal{H}_{n,j,\alpha}}\frac{\mathbb{E}\Delta h(X_0)-\frac{1}{n}\sum_{i=1}^{n}\Delta h(X_i)}{(2^j\alpha+2J_{\lambda_n,\tau_n}(h(X_0))+\mathbb{E}\Delta h(X_0))}\ge\frac{1}{2}\right).
\end{aligned}
$$

对于 $h\in\mathcal{H}_{n,\ j,\ \alpha}$, 我们有

$$
P\left(\sup_{h\in\mathcal{H}_\sigma(L_n,N_n,B_n,F)}\frac{\mathbb{E}\Delta h(Z_0)-\frac{1}{n}\sum_{i=1}^{n}\Delta h(Z_i)}{(\alpha+2J_{\lambda_n,\tau_n}(h(X_0))+\mathbb{E}\Delta h(Z_0))}\ge\frac{1}{2}\right)\le\sum_{j=0}^{\infty}P\left(\sup_{h\in\mathcal{H}_{n,j,\alpha}}\frac{\mathbb{E}\Delta h(Z_0)-\frac{1}{n}\sum_{i=1}^{n}\Delta h(Z_i)}{(2^j\alpha+\mathbb{E}\Delta h(Z_0))}\ge\frac{1}{2}\right).
$$

容易证明,

$$
\begin{aligned}
\mathbb{E}(\Delta h(Z_0)]&=\mathbb{E}[(h(X_0)-h^{\perp}(X_0))^2]-\mathbb{E}[(h^{\perp}(X_0))^2]+2\mathbb{E}[(h^{\perp}(X_0))^2]-\mathbb{E}[(h^{\perp})^2(X_0)]\\
&=\mathbb{E}[(h(X_0)-h^{\perp}(X_0))^2]\ge 0.
\end{aligned}
$$

因此，

$$\sum_{j=1}^{\infty} P\left(\sup_{h\in\mathcal{H}_{n,j,\alpha}} \frac{\mathbb{E}\Delta h(Z_0)-\frac{1}{n}\sum_{i=1}^{n}\Delta h(Z_i)}{(2^j\alpha+\mathbb{E}\Delta h(Z_0))}\geq\frac{1}{2}\right)\leq\sum_{j=1}^{\infty} P\left(\sup_{h\in\mathcal{H}_{n,j,\alpha}} \frac{\mathbb{E}\Delta h(Z_0)-\frac{1}{n}\sum_{i=1}^{n}\Delta h(Z_i)}{2^j\alpha}\geq\frac{1}{2}\right)$$
$$\leq\sum_{j=1}^{\infty} P\left(\sup_{h\in\mathcal{H}_{n,j,\alpha}}\left\{\mathbb{E}\Delta h(Z_0)-\frac{1}{n}\sum_{i=1}^{n}\Delta h(Z_i)\right\}\geq\frac{2^j\alpha}{2}\right)\leq\sum_{j=1}^{\infty} P\left(\sup_{g\in\mathcal{G}_{n,j,\alpha}}\left\{\mathbb{E}[g(Z_0)]-\frac{1}{n}\sum_{i=1}^{n}g\right\}\geq\frac{2^j\alpha}{2}\right),$$

其中，

$$\mathcal{G}_{n,j,\alpha}:=\left\{\Delta(h):\mathbb{R}^d\times\mathbb{R}\to\mathbb{R}, h\in\mathcal{H}_{n,j,\alpha}\right\}. \tag{7.6}$$

设 $h\in\mathcal{H}_\sigma(L_n,\ N_n,\ B_n,\ F)$。考虑函数 $g(x,\ y)=\Delta h(x,\ y)$。容易证明，$g$是利普希茨函数，其Lipschitz系数为 $\max\left(2FLip(h)+2K_nLip(h)\ 6K_n+2F\right)$。因此，可以得到，过程 $(g(Z_t))_{t\in\mathbb{Z}}$ 也是 $\psi$-弱依赖的。因此，由 [8] （另见 [5]），我们有

$$\begin{aligned}P\left\{\mathbb{E}[g(Z_0)]-\frac{1}{n}\sum_{i=1}^{n}g(Z_i)>\varepsilon\right\}&=P\left\{\sum_{i=1}^{n}\left(\mathbb{E}[g(X_0,Y_0)]-g(X_i,Y_i)\right)\geq n\varepsilon\right\}\\&\leq P\left\{\left|\sum_{i=1}^{n}\left(\mathbb{E}[g(X_0,Y_0)]-g(X_i,Y_i)\right)\right|\geq n\varepsilon\right\}\\&\leq C_3\log n\exp\left(-\frac{n^2\varepsilon^2}{A_n'+B_n'(n\varepsilon)^\nu}\right)\\&\leq C_3\exp\left(\log\log n-\frac{n^2\varepsilon^2}{A_n'+B_n'(n\varepsilon)^\nu}\right).\end{aligned} \tag{7.7}$$

对于某个常数 $C_3>0$，任意满足 $A_n'\geq\mathbb{E}\hat{\imath}(\sum_{i=1}^{n}(g(X_i,\ Y_i)-\mathbb{E}[g(X_0,\ Y_0)])^2\acute{o}$ 和 $B_n=\ ^{n^{3/4}}$对数$_{A_n'}\ ^n$ 的序列 $(A_n')_{n\in\mathbb{N}\circ}$ 令 $l=\mathcal{N}(\varepsilon,\ \mathcal{G}_{n,\ j,\ \alpha},\ \|\cdot\|_\infty)$。当 n 足够大时，我们有

$$\begin{aligned}P\left\{\sup_{g\in\mathcal{G}_{n,j,\alpha}}\left[\mathbb{E}[g(Z_0)]-\frac{1}{n}\sum_{i=1}^{n}g(Z_i)\right]>\varepsilon\right\}&\leq C_3\sum_{i=1}^{l}\exp\left(\log\log n-\frac{n^2\varepsilon^2/4}{A_n'+B_n'(n\varepsilon/2)^\nu}\right)\\&\leq C_3\cdot l\exp\left(\log\log n-\frac{n^2\varepsilon^2/4}{A_n'+B_n'(n\varepsilon/2)^\nu}\right)\\&\leq C_3\mathcal{N}(\varepsilon,\mathcal{G}_{n,j,\alpha})\exp\left(\log\log n-\frac{n^2\varepsilon^2/4}{A_n'+B_n'(n\varepsilon/2)^\nu}\right).\end{aligned}$$

在 [19], 中，我们有

$$\mathcal{N}(\varepsilon,\mathcal{G}_{n,j,\alpha},\|\cdot\|_\infty)\leq\mathcal{N}(\frac{\varepsilon}{4K_n},\mathcal{H}_{n,j,\alpha},\|\cdot\|_\infty). \tag{7.8}$$

容易看出，

$$\mathcal{H}_{n,j,\alpha}\subset\left\{h\in\mathcal{H}_\sigma(L_n,N_n,B_n,F,\frac{2^j\alpha}{\lambda_n}):\|\theta(h)\|_{\text{clip},\tau_n}\leq\frac{2^j\alpha}{\lambda_n}\right\}. \tag{7.9}$$

因此，在 [19], 中，我们有以下不等式

$$\begin{aligned}\mathcal{N}(\varepsilon,\mathcal{G}_{n,j,\alpha},\|\cdot\|_\infty)&\leq\mathcal{N}(\frac{\varepsilon}{4K_n},\mathcal{H}_{n,j,\alpha},\|\cdot\|_\infty)\\&\leq\mathcal{N}(\frac{\varepsilon}{4K_n},\mathcal{H}_\sigma(L_n,N_n,B_n,F,\frac{2^j\alpha}{\lambda_n}),\|\cdot\|_\infty)\\&\leq\exp\left(2\frac{2^j\alpha}{\lambda_n}(L_n+1)\log\left(\frac{(L_n+1)(N_n+1)B_n}{\frac{\varepsilon}{4K_n}-\tau_n(L_n+1)((N_n+1)B_n)^{L_n+1}}\right)\right).\end{aligned} \tag{7.10}$$

我们有,

$$P\Big\{\sup_{g\in\mathcal{G}_{n,j,\alpha}}\left[\mathbb{E}[g(Z_0)]-\frac{1}{n}\sum_{i=1}^{n}g(Z_i)\right]>\varepsilon\Big\}$$
$$\leq C_3\exp\left(2\frac{2^j\alpha}{\lambda_n}(L_n+1)\log\left(\frac{(L_n+1)(N_n+1)B_n}{\frac{\varepsilon}{4K_n}-\tau_n(L_n+1)((N_n+1)B_n)^{L_n+1}}\right)+\log\log n-\frac{n^2\varepsilon^2/4}{A'_n+B'_n(n\varepsilon/2)^\nu}\right). \tag{7.11}$$

因此,

$$\sum_{j=1}^{\infty}P\left\{\sup_{g\in\mathcal{G}_{n,j,\alpha}}\left[\mathbb{E}[g(Z_0)]-\frac{1}{n}\sum_{i=1}^{n}g(Z_i)\right]>\varepsilon\right\}$$
$$\leq C_3\sum_{j=1}^{\infty}\exp\left(2\frac{2^j\alpha}{\lambda_n}(L_n+1)\log\left(\frac{(L_n+1)(N_n+1)B_n}{\frac{\varepsilon}{4K_n}-\tau_n(L_n+1)((N_n+1)B_n)^{L_n+1}}\right)+\log\log n-\frac{n^2\varepsilon^2/4}{A'_n+B'_n(n\varepsilon/2)^\nu}\right). \tag{7.12}$$

令 $\varepsilon=\frac{2^j\alpha}{2}$，我们有

$$\sum_{j=1}^{\infty}P\Big\{\sup_{g\in\mathcal{G}_{n,j,\alpha}}\left[\mathbb{E}[g(Z_0)]-\frac{1}{n}\sum_{i=1}^{n}g(Z_i)\right]>\frac{2^j\alpha}{2}\Big\}$$
$$\leq C_3\sum_{j=1}^{\infty}\exp\left(2\frac{2^j\alpha}{\lambda_n}(L_n+1)\log\left(\frac{(L_n+1)(N_n+1)B_n}{\frac{2^j\alpha}{8K_n}-\tau_n(L_n+1)((N_n+1)B_n)^{L_n+1}}\right)+\log\log n-\frac{(n2^j\alpha)^2/16}{A'_n+B'_n(n2^j\alpha/4)^\nu}\right)$$
$$\leq C_3\exp(\log\log n)\sum_{j=1}^{\infty}\exp\left(2\frac{2^j\alpha}{\lambda_n}(L_n+1)\log\left(\frac{(L_n+1)(N_n+1)B_n}{\frac{2^j\alpha}{8K_n}-\tau_n(L_n+1)((N_n+1)B_n)^{L_n+1}}\right)-\frac{(n2^j\alpha)^2/16}{A'_n+B'_n(n2^j\alpha/4)^\nu}\right). \tag{7.13}$$

通过选择 $A'_n=nC$ 和 $B'_n=\log(\frac{n}{n^{1/4}C})$，我们有

$$\sum_{j=1}^{\infty}P\Big\{\sup_{g\in\mathcal{G}_{n,j,\alpha}}\left[\mathbb{E}[g(Z_0)]-\frac{1}{n}\sum_{i=1}^{n}g(Z_i)\right]>\frac{2^j\alpha}{2}\Big\}$$
$$\leq C_3\exp(\log\log\ n)$$
$$\times\sum_{j=1}^{\infty}\exp\left(2\frac{2^j\alpha}{\lambda_n}(L_n+1)\log\left(\frac{(L_n+1)(N_n+1)B_n}{\frac{2^j\alpha}{8K_n}-\tau_n(L_n+1)((N_n+1)B_n)^{L_n+1}}\right)-\frac{(n2^j\alpha)^2/16}{nC+\log(\frac{n}{n^{1/4}C})(n2^j\alpha/4)^\nu}\right). \tag{7.14}$$

Let,

$$\log(\frac{n}{n^{1/4}C})(n2^j\alpha/8)^\nu>nC\Longrightarrow\alpha>\frac{8}{n}\left(\frac{nC}{\log(\frac{n}{n^{1/4}C})}\right)^{1/\nu}:=\alpha_n. \tag{7.15}$$

由（7.15）可轻易看出

$$-\frac{(n2^j\alpha)^2/16}{nC+\log(\frac{n}{n^{1/4}C})(n2^j\alpha/8)^\nu}\leq-\frac{(n2^j\alpha)^2/16}{2\log(\frac{n}{n^{1/4}C})(n2^j\alpha/4)^\nu}.$$

我们还可以看到，当渐近地 $\alpha_n > 1$。
**步骤 1**： $\alpha > \alpha_n$ 在该假设下

$$\tau_n \leq \frac{1}{16K_n(L_n+1)((N_n+1)B_n)^{L_n+1}},$$

并且对于 $2^j\alpha > 1$，我们有

$$\sum_{j=1}^{\infty} \exp\left( 2\frac{2^j\alpha}{\lambda_n}(L_n+1)\log\left( \frac{(L_n+1)(N_n+1)B_n}{\frac{2^j\alpha}{8K_n} - \tau_n(L_n+1)((N_n+1)B_n)^{L_n+1}} \right) - \frac{(n2^j\alpha)^2/16}{nC + \log(\frac{n}{n^{1/4}C})(n2^j\alpha/4)^{\nu}} \right)$$
$$\leq \exp\left( 2\frac{2^j\alpha}{\lambda_n}(L_n+1)\log\left(16K_n(L_n+1)(N_n+1)B_n\right) - \frac{(n2^j\alpha)^{2-\nu}/16}{2\log(\frac{n}{n^{1/4}C})/4^{\nu}} \right). \tag{7.16}$$

对于 $\alpha > \alpha_n$，我们有 $2^j\alpha < ((2\alpha)^{2-\nu})^j$ 因此，

$$\sum_{j=1}^{\infty} P\Big\{ \sup_{g\in\mathcal{G}_{n,j,\alpha}} \left[ \mathbb{E}[g(Z_0)] - \frac{1}{n}\sum_{i=1}^{n} g(Z_i) \right] > \frac{2^j\alpha}{2} \Big\}$$
$$\leq \sum_{j=1}^{\infty} \exp\left( 2^j\alpha \left( \frac{2}{\lambda_n}(L_n+1)\log\left(16K_n(L_n+1)(N_n+1)B_n\right) - \frac{n^{2-\nu}/32}{\log(\frac{n}{n^{1/4}C})/4^{\nu}} \right) \right). \tag{7.17}$$

我们可以看到，当 n 足够大时，

$$\delta_{n,1} := \frac{2}{\lambda_n}(L_n+1)\log\left(16K_n(L_n+1)(N_n+1)B_n\right) \leq \frac{n^{2-\nu}/32}{2\log(\frac{n}{n^{1/4}C})/4^{\nu}}.$$

因此，

$$\sum_{j=1}^{\infty} P\Big\{ \sup_{g\in\mathcal{G}_{n,j,\alpha}} \left[ \mathbb{E}[g(Z_0)] - \frac{1}{n}\sum_{i=1}^{n} g(Z_i) \right] > \frac{2^j\alpha}{2} \Big\} \leq C_3\log n \sum_{j=1}^{\infty} \exp\left( -2^j \frac{n^{2-\nu}}{64\log(\frac{n}{n^{1/4}C})/4^{\nu}}\alpha \right).$$

Let

$$\beta_n := \frac{n^{2-\nu}}{64\log\left(\frac{n}{n^{1/4}C}\right)/4^{\nu}}\alpha,$$

我们有

$$\sum_{j=0}^{\infty} \exp\left(-\beta_n(2^j\right) \leq \exp\left(-\beta_n\right) \sum_{j=0}^{\infty} \left( 2^{-\frac{\beta_n}{\log 2}} \right)^j \leq \exp\left(-\beta_n\right) \left( \frac{1}{1-2^{-\frac{\beta_n}{\log 2}}} \right) \leq \exp\left(-\beta_n\right) \tag{7.18}$$

应用（7.18），我们有

$$\sum_{j=1}^{\infty} \exp\left( -2^j \frac{n^{2-\nu}}{64\log(\frac{n}{n^{1/4}C})/4^{\nu}}\alpha \right) \leq \exp\left( -\frac{n^{2-\nu}}{64\log(\frac{n}{n^{1/4}C})/4^{\nu}}\alpha \right). \tag{7.19}$$

因此

$$P(A_{2,n} > \alpha) \lesssim C_3\log n \sum_{j=1}^{\infty} \exp\left( -2^j \frac{n^{2-\nu}}{64\log(\frac{n}{n^{1/4}C})/4^{\nu}}\alpha \right) \lesssim C_3\log n \exp\left( -\frac{n^{2-\nu}}{64\log(\frac{n}{n^{1/4}C})/4^{\nu}}\alpha \right), \tag{7.20}$$

对于 $\alpha \geq \alpha_{n^\circ}$

$$\begin{aligned}\int_{\alpha_n}^{\infty} P(A_{2,n} > \alpha)d\alpha &\leq \int_{\alpha_n}^{\infty} C_3 \log n \exp\left(-\frac{n^{2-\nu}}{64\log(\frac{n}{n^{1/4}C})/4^\nu}\alpha\right) d\alpha \\ &\leq 64C_3 \log n \frac{\log(\frac{n}{n^{1/4}C})}{n^{2-\nu}} \exp\left(-\frac{n^{2-\nu}}{64\log(\frac{n}{n^{1/4}C})/4^\nu} \times \frac{8}{n}\left(\frac{nC}{\log\frac{n}{n^{1/4}C}}\right)\right).\end{aligned}$$

Let

$$nC < \log(\frac{n}{n^{1/4}C})(n2^j\alpha/4)^\nu \Longrightarrow \alpha < \frac{8}{n}\left(\frac{nC}{\log(\frac{n}{n^{1/4}C})^\nu}\right)^{1/\nu} \coloneqq \alpha_n, \tag{7.21}$$

因此,

$$-\frac{(n2^j\alpha)^2/16}{nC + \log(\frac{n}{n^{1/4}C})(n2^j\alpha/4)^\nu} \leq -\frac{(n2^j\alpha)^2/16}{2nC}.$$

我们可以看到，当渐近地 $\alpha_n > 1,\quad 2^j > 1_\circ$ 因此,

**步骤2**: $1 \leq \alpha \leq \alpha_n$

因此我们有

$$\begin{aligned}&\sum_{j=1}^{\infty} \exp\left(2\frac{2^j\alpha}{\lambda_n}(L_n+1)\log\left(\frac{(L_n+1)(N_n+1)B_n}{\frac{2^j\alpha}{8K_n} - \tau_n(L_n+1)((N_n+1)B_n)^{L_n+1}}\right) - \frac{(n2^j\alpha)^2/16}{nC + \log(\frac{n}{n^{1/4}C})(n2^j\alpha/4)^\nu}\right) \\ &\leq \sum_{j=1}^{\infty} \exp\left(2\frac{2^j\alpha}{\lambda_n}(L_n+1)\log\left(16K_n(L_n+1)(N_n+1)B_n\right) - \frac{(n2^j\alpha)^2/16}{2nC}\right).\end{aligned} \tag{7.22}$$

对于 $1 \leq \alpha \leq \alpha_n$，我们有 $2^j\alpha \leq (2^j\alpha)^2$

$$\begin{aligned}&\sum_{j=1}^{\infty} P\left\{\sup_{g\in\mathcal{G}_{n,j,\alpha}}\left[\mathbb{E}[g(Z_0)] - \frac{1}{n}\sum_{i=1}^{n} g(Z_i)\right] > \frac{2^j\alpha}{2}\right\} \\ &\leq C_3 \log n \sum_{j=1}^{\infty} \exp\left(2^j\alpha\left(\frac{2}{\lambda_n}(L_n+1)\log\left(16K_n(L_n+1)(N_n+1)B_n\right) - \frac{n^2}{32nC}\right)\right).\end{aligned} \tag{7.23}$$

我们可以看到，当 n 足够大时

$$\delta_{n,2} \coloneqq \frac{2}{\lambda_n}(L_n+1)\log\left(16K_n(L_n+1)(N_n+1)B_n\right) \leq \frac{n^2/16}{4nC}.$$

因此,

$$\sum_{j=1}^{\infty} P\Big\{\sup_{g\in\mathcal{G}_{n,j,\alpha}}\left[\mathbb{E}[g(Z_0)] - \frac{1}{n}\sum_{i=1}^{n} g(Z_i)\right] > \frac{2^j\alpha}{2}\Big\} \leq C_3 \log n \sum_{j=1}^{\infty} \exp\left(-2^j\frac{n}{64C}\alpha\right). \tag{7.24}$$

令 $\beta_n^{'} = \dfrac{n}{64C}$，运用与 (7.18) 类似的论证可得

$$P(A_{2,n} > \alpha) \lesssim C_3 \log n \sum_{j=1}^{\infty} \exp\left(-2^j\frac{n}{64C}\alpha\right) \lesssim C_3 \log n \exp\left(-\frac{n}{64C}\alpha\right),$$

对于 $1 \le \alpha \le \alpha_{n\circ}$ 这意味着

$$\int_1^{\alpha_n} P(A_{2,n} > \alpha) \le C_3 \log n \int_1^{\alpha_n} \exp\Big(-\frac{n}{64C}\alpha\Big)\, d\alpha$$
$$\le \frac{64CC_3 \log n}{n} \exp\left(-\frac{n}{64C} \times \frac{8}{n}\left(\frac{nC}{\log(\frac{n}{n^{1/4}C})}\right)^{1/\nu}\right) - \frac{64CC_3 \log n}{n} \exp\Big(-\frac{n}{64C}\Big).$$

**第3**: $0 < \alpha \le 1 < \alpha_{n\circ}$

回顾该假设

$$\tau_n \le \frac{\beta_n}{16K_n(L_n+1)((N_n+1)B_n)^{L_n+1}},$$

满足条件 $\beta_n := (\log\ n)^{\nu_5}/n^{\nu_6}$，其中某个 $\nu_5,\ \nu_6 > 0$ 且

$$\Big(\nu_6 < 1/2, \nu_4 + \nu_6 < 1\Big) \text{ or } \Big(\nu_6 < 1/2, \nu_4 + \nu_6 = 1, \nu_5 > 1 - \nu_3\Big). \tag{7.25}$$

我们对所有 $\alpha \in (\beta_n,\ 1)$ 都有，

$$\sum_{j=1}^{\infty} \exp\left(2\frac{2^j\alpha}{\lambda_n}(L_n+1)\log\left(\frac{(L_n+1)(N_n+1)B_n}{\frac{2^j\alpha}{8K_n} - \tau_n(L_n+1)((N_n+1)B_n)^{L_n+1}}\right) - \frac{(n2^j\alpha)^2/16}{nC + \log(\frac{n}{n^{1/4}C})(n2^j\alpha/8)^{\nu}}\right)$$
$$\le \sum_{j=1}^{\infty} \exp\Big(\frac{2(L_n+1)2^j\alpha}{\lambda_n}\log\Big(\frac{16K_n(L_n+1)(N_n+1)B_n}{\beta_n}\Big) - \frac{(n2^j\alpha)^2}{32nC}\Big)$$
$$\le \sum_{j=1}^{\infty} \exp\Big(\frac{2(L_n+1)(2^j)^2\alpha}{\lambda_n}\log\Big(\frac{16K_n(L_n+1)(N_n+1)B_n}{\beta_n}\Big) - \frac{(n2^j\alpha)^2}{32nC}\Big)$$
$$\le \sum_{j=1}^{\infty} \exp\left[-(2^j)^2\left(\frac{n\alpha^2}{32C} - \frac{2(L_n+1)\alpha}{\lambda_n}\log\Big(\frac{16K_n(L_n+1)(N_n+1)B_n}{\beta_n}\Big)\right)\right]$$
$$\le \sum_{j=1}^{\infty} \exp\big(-4^j\phi_n(\alpha)\big) \tag{7.26}$$

with

$$\phi_n(\alpha) := \frac{n\alpha^2}{32C} - \frac{2(L_n+1)\alpha}{\lambda_n}\log\Big(\frac{16K_n(L_n+1)(N_n+1)B_n}{\beta_n}\Big) \quad \text{for all } \beta_n < \alpha \le 1 < \alpha_n. \tag{7.27}$$

由于 $K_n,\ L_n,\ N_n,\ B_n \ge 1$，我们有 $16K_n(L_n+1)(N_n+1)B_n/\beta_n > 1$，因此，

$$\phi_n(\alpha) \ge \frac{n\beta_n^2}{32C} - \frac{2(L_n+1)\beta_n}{\lambda_n}\log\Big(\frac{16K_n(L_n+1)(N_n+1)B_n}{\beta_n}\Big) = \phi_n(\beta_n), \quad \text{for all } \beta_n < \alpha \le 1 < \alpha_n. \tag{7.28}$$

在假设 $\lambda_n \asymp (\log n)^{\nu_3}/n^{\nu_4}$ 对某个 $0 < \nu_4 < 1$、$\nu_6 < 1/2$ 以及 **(7.25)** 成立的条件下，我们得到：

$$\varphi_n := \phi_n(\beta_n) \underset{n\to\infty}{\longrightarrow} \infty. \tag{7.29}$$

可以看出，

$$\exp(-\beta 4^j) \le \exp(-\beta) \times \exp(-\beta j), \text{for all } \beta \ge 0,\ j \in \mathbb{N}.$$

因此，由（7.26）、（7.28）和（7.29）可知，对于所有 $\alpha \in (\beta_n, 1)$，当 $n$ 足够大时，

$$\sum_{j=1}^{\infty} \exp\left(2\frac{2^j\alpha}{\lambda_n}(L_n+1)\log\left(\frac{16K_n(L_n+1)(N_n+1)B_n}{2^j\alpha}\right) - \frac{(n2^j\alpha)^2/16}{nC+\log(\frac{n}{n^{1/4}C})(n2^j\alpha/8)^{\nu}}\right)$$
$$\leq \sum_{j=1}^{\infty} \exp\big(-4^j\varphi_n\big) \leq \sum_{j=1}^{\infty} \exp(-\varphi_n)\exp(-\varphi_n j) \leq \frac{\exp(-2\varphi_n)}{1-\exp(-\varphi_n)} \leq 2\exp(-2\varphi_n).$$

因此，对于充分大的 $n$,

$$\int_0^1 P(A_{2,n} > \alpha)d\alpha \leq \beta_n + \int_{\beta_n}^1 P(A_{2,n} > \alpha)d\alpha \leq \beta_n + 2\exp(-2\varphi_n) \leq 2\beta_n,$$

其中 $\beta_n$ 如上所定义，并满足 (7.25)。因此，对于 $\nu_5 > 0$, $\nu_6 < 1/2$，我们有

$$\begin{aligned}
\mathbb{E}[A_{2,n}] &= \int_0^{\infty} P(A_{2,n} > \alpha)d\alpha \\
&\leq \int_0^{\beta_n} P(A_{2,n} > \alpha)d\alpha + \int_{\beta_n}^1 P(A_{2,n} > \alpha)d\alpha + \int_1^{\alpha_n} P(A_{2,n} > \alpha)d\alpha + \int_{\alpha_n}^{\infty} P(A_{2,n} > \alpha)d\alpha \\
&\leq 2\beta_n + \frac{64CC_3\log n}{n}\exp\left(-\frac{n}{64C}\times\frac{8}{n}\left(\frac{nC}{\log(\frac{n}{n^{1/4}C})}\right)^{1/\nu}\right) - \frac{64CC_3\log n}{n}\exp\left(-\frac{n}{64C}\right) \\
&+ 32C_3\log n\frac{\log(\frac{n}{n^{1/4}C})}{n^{2-\nu}}\exp\left(-\frac{n^{2-\nu}}{32\log(\frac{n}{n^{1/4}C})/8^{\nu}}\times\frac{8}{n}\left(\frac{nC}{\log\frac{n}{n^{1/4}C}}\right)\right) \\
&\lesssim 2\beta_n. \qquad (7.30)
\end{aligned}$$

对于 $A_{4,\ n}$，我们选择一个神经网络函数 $h_n^0 \in \mathcal{H}_\sigma(L_n,\ N_n,\ B_n,\ F)$，使得

$$\|h_n^0 - h^*\|_{2,P_{X_0}}^2 + J_{\lambda_n,\tau_n}(h_n^0) \leq \inf_{h\in\mathcal{H}_\sigma(L_n,N_n,B_n,F)}\Big[\|h-h^*\|_{2,P_{X_0}}^2 + J_{\lambda_n,\tau_n}(h)\Big] + n^{-1}.$$

根据基本不等式

$$\frac{1}{n}\sum_{i=1}^n\Big(Y_i - \widehat{h}_n(X_i)\Big)^2 + J_{\lambda_n,\tau_n}(\widehat{h}_n) \leq \frac{1}{n}\sum_{i=1}^n(Y_i - h(X_i))^2 + J_{\lambda_n,\tau_n}(h),$$

对于任意 $h \in \mathcal{H}_\sigma(L_n,\ N_n,\ B_n,\ F)$，我们有

$$\begin{aligned}
A_{4,n} &= 2\Big[\frac{1}{n}\sum_{i=1}^n(Y_i - \widehat{h}_n(X_i))^2 - \frac{1}{n}\sum_{i=1}^n(Y_i - h_n^0(X_i))^2\Big] + 2J_{\lambda_n,\tau_n}(\widehat{h}_n) \\
&+ 2\Big[\frac{1}{n}\sum_{i=1}^n(Y_i - h_n^0(X_i))^2 - \frac{1}{n}\sum_{i=1}^n(Y_i - h^*(X_i))^2\Big] \qquad (7.31) \\
&\leq 2J_{\lambda_n,\tau_n}(h_n^0) + 2\Big[\frac{1}{n}\sum_{i=1}^n(Y_i - h_n^0(X_i))^2 - \frac{1}{n}\sum_{i=1}^n(Y_i - h^*(X_i))^2\Big]
\end{aligned}$$

并且我们有

$$\mathbb{E}[A_{4,n}] \leq 2J_{\lambda_n,\tau_n}(h_n^0) + \frac{2}{n}\sum_{i=1}^n\Big[\mathbb{E}(Y_i - h_n^0(X_i))^2 - \mathbb{E}(Y_i - h^*(X_i))^2\Big],$$

可以看出，对于 $i = 1, \ldots, n$,

$$\mathbb{E}[(Y_i - h_n^0(X_i))^2] - \mathbb{E}[(Y_i - h^*(X_i))^2] \leq \|h_n^0 - h^*\|_{2,P_{X_0}}^2.$$

因此，

$$\mathbb{E}[A_{4,n}] \leq 2J_{\lambda_n,\tau_n}(h_n^0) + 2\|h_n^0 - h^*\|_{2,P_{X_0}}^2 \quad \leq 2\inf_{h\in\mathcal{H}_\sigma(L_n,N_n,B_n,F)}\Big[\|h-h^*\|_{2,P_{X_0}}^2 + J_{\lambda_n,\tau_n}(h)\Big] + 2n^{-1}.$$

因此，

$$\mathbb{E}\Big[\|\widehat{h}_n - h^*\|_{2,P_{X_0}}^2\Big] = \mathbb{E}\left[\sum_{i=1}^4 A_{i,n}\right] \lesssim 3\frac{\log n}{n} + 2\frac{(\log n)^{\nu_5}}{n^{\nu_6}} + 2\inf_{h\in\mathcal{H}_\sigma(L_n,N_n,B_n,F)}\Big[\|h-h^*\|_{2,P_{X_0}}^2\Big] + 2n^{-1}$$

$$\lesssim 2\left(\inf_{h\in\mathcal{H}_\sigma(L_n,N_n,B_n,F)}\{\|h-h^*\|_{2,P_{X_0}}^2 + \lambda_n\|\theta(h)\|_{\mathrm{clip},\tau_n}\} \vee \frac{(\log n)^{\nu_5}}{n^{\nu_6}}\right). \tag{7.32}$$

定理得证。

■

## 7.2 定理 4.2 的证明

令 $\epsilon_n = n^{-\nu_4/(\kappa+2)}$。根据定理（4.1）、假设（4.6）以及对于任意 $\tau > 0$, $\|\theta(h)\|_{\mathrm{clip},\tau} \leq \|\theta(h)\|_0$这一事实，对于任意 $h^* \in \mathcal{H}^*$，有

$$\begin{aligned}
&\sup_{h\in\mathcal{H}_{\rho,H^*}:h^*\in\mathcal{H}^*}\mathbb{E}\Big[\|\widehat{h}_n - h^*\|_{2,P_{X_0}}^2\Big] \\
&\lesssim \sup_{h\in\mathcal{H}_{\rho,H^*}:h^*\in\mathcal{H}^*}\ \inf_{h\in\mathcal{H}_\sigma(L_n,N_n,B_n,F,C\epsilon_n^{-\kappa}(\log n)^r)}\Big\{\|h-h^*\|_{2,P_{X_0}}^2 + \lambda_n\|\theta(h)\|_{\mathrm{clip},\tau_n}\Big\} \vee \frac{(\log n)^{\nu_5}}{n^{\nu_6}} \\
&\lesssim \sup_{h\in\mathcal{H}_{\rho,H^*}:h^*\in\mathcal{H}^*}\Big[\inf_{h\in\mathcal{H}_\sigma(L_n,N_n,B_n,F,C\epsilon_n^{-\kappa}(\log n)^r)}\|h-h^*\|_{2,P_{X_0}}^2 + \lambda_n\epsilon_n^{-\kappa}(\log n)^r\Big] \vee \frac{(\log n)^{\nu_5}}{n^{\nu_6}} \\
&\lesssim \sup_{h\in\mathcal{H}_{\rho,H^*}:h^*\in\mathcal{H}^*}\Big[\epsilon_n^2 + \frac{(\log n)^{\nu_3}}{n^{\nu_4}}\epsilon_n^{-\kappa}(\log n)^r\Big] \vee \frac{(\log n)^{\nu_5}}{n^{\nu_6}} \\
&\lesssim \frac{(\log n)^{r+\nu_3}}{n^{2\nu_4/(\kappa+2)}} \vee \frac{(\log n)^{\nu_5}}{n^{\nu_6}}.
\end{aligned} \tag{7.33}$$

因此，定理得证。 ■

## 7.3 定理 5.1 的证明

我们将 $\mathcal{E}_{Z_0}(h_n)$分解如下。

$$\mathcal{E}_{Z_0}(\widehat{h}_n) = \mathbb{E}[\ell(Y_0\widehat{h}_n(X_0))] - \mathbb{E}[\ell(Y_0h_\ell^*(X_0))] \coloneqq B_{1,n} + B_{2,n}, \tag{7.34}$$

其中

$$B_{1,n} = \Big[\mathbb{E}[\ell(Y_0\widehat{h}_n(X_0))] - \mathbb{E}[\ell(Y_0h_\ell^*(X_0))]\Big] - \frac{2}{n}\sum_{i=1}^n\Big[\ell(Y_i\widehat{h}_n(X_i)) - \ell(Y_ih_\ell^*(X_i))\Big] - 2J_{\lambda_n,\tau_n}(\widehat{h}_n);$$

$$B_{2,n} = \frac{2}{n}\sum_{i=1}^n\Big[\ell(Y_i\widehat{h}_n(X_i)) - \ell(Y_ih_\ell^*(X_i))\Big] + 2J_{\lambda_n,\tau_n}(\widehat{h}_n).$$

可以按照定理 4.1 证明中对 $A_{2,n}$ 的相同方式，对 $B_{1,n}$ 进行界定。令 $\Delta(h)(Z_0) \coloneqq \ell(Y_0h(X_0)) - \ell(Y_0h_\ell^*(X_0))$, 其中 $Z_0 \coloneqq (X_0, Y_0)$。设

$$\mathcal{H}_{n,j,\alpha} \coloneqq \{h \in \mathcal{H}_\sigma(L_n, N_n, B_n, F) : 2^{j-1}\mathbb{1}(j \neq 0)\alpha \leq J_{\lambda_n,\tau_n}(h) \leq 2^j\alpha\},$$

and

$$\mathcal{G}_{n,j,\alpha} := \Big\{\Delta(h) : \mathbb{R}^d \times \{-1,1\} \mapsto \mathbb{R} : h \in \mathcal{H}_{n,j,\alpha}\Big\}. \tag{7.35}$$

根据 $h_\ell^*$ 的定义可知， $\mathbb{E}\Delta(h)(Z_0) \geq 0$。对于所有 $\alpha > 0$，有

$$\begin{aligned}
P(B_{1,n} > \alpha) &\leq P\Big(\sup_{h\in\mathcal{H}_\sigma(L_n,N_n,B_n,F)} \frac{\mathbb{E}\Delta(h)(Z_0) - \frac{1}{n}\sum_{i=1}^n \Delta(h)(Z_i)}{\alpha + 2J_{\lambda_n,\tau_n}(h) + \mathbb{E}\Delta(h)(Z_0)} \geq \frac{1}{2}\Big) \\
&\leq \sum_{j=0}^{\infty} P\Big(\sup_{h\in\mathcal{H}_{n,j,\alpha}} \frac{\mathbb{E}\Delta(h)(Z_0) - \frac{1}{n}\sum_{i=1}^n \Delta(h)(Z_i)}{2^j\alpha + \mathbb{E}\Delta(h)(Z_0)} \geq \frac{1}{2}\Big) \\
&\leq \sum_{j=0}^{\infty} P\Big(\sup_{g\in\mathcal{G}_{n,j,\alpha}} \frac{\mathbb{E}g(Z_0) - \frac{1}{n}\sum_{i=1}^n g(Z_i)}{2^j\alpha} \geq \frac{1}{2}\Big) \leq \sum_{j=0}^{\infty} P\Big(\sup_{g\in\mathcal{G}_{n,j,\alpha}} \mathbb{E}g(Z_0) - \frac{1}{n}\sum_{i=1}^n g(Z_i) \geq \frac{2^j\alpha}{2}\Big) \\
&\leq \sum_{j=0}^{\infty} P\Big(\sup_{g\in\mathcal{G}_{n,j,\alpha}} \sum_{i=1}^n \Big(\mathbb{E}g(Z_0) - g(Z_i)\Big) \geq \frac{2^j\alpha n}{2}\Big).
\end{aligned}$$

令 $\Delta(h_1)$ $\Delta(h_2) \in \mathcal{G}_{n,\ j,\ \alpha}$ 和 $(x,\ y) \in \mathbb{R}^d \times \{-1,1\}$，我们有

$$\|\Delta(h_1)(x,y) - \Delta(h_2)(x,y)\|_\infty \leq \|\ell(yh_1(x)) - \ell(yh_2(x))\|_\infty \leq \mathcal{K}_\ell |h_1(x) - h_2(x)|.$$

因此，

$$\mathcal{N}\Big(\varepsilon, \mathcal{G}_{n,j,\alpha}, \|\cdot\|_\infty\Big) \leq \mathcal{N}\Big(\frac{\varepsilon}{\mathcal{K}_\ell}, \mathcal{H}_{n,j,\alpha}, \|\cdot\|_\infty\Big).$$

正如定理 4.1 的证明中所强调的，过程 $\Big(g(Z_t) := \ell(h(X_t)\ Y_t)\Big)_{t\in\mathbb{Z}}$ 也是 $\psi$-弱依赖的。由 (7.9) 和 (7.10)，我们有，

$$\begin{aligned}
\mathcal{N}(\varepsilon, \mathcal{G}_{n,j,\alpha}, \|\cdot\|_\infty) &\leq \mathcal{N}\Big(\frac{\varepsilon}{\mathcal{K}_\ell}, \mathcal{H}_{n,j,\alpha}, \|\cdot\|_\infty\Big) \\
&\leq \mathcal{N}\Big(\frac{\varepsilon}{\mathcal{K}_\ell}, \mathcal{H}_\sigma(L_n, N_n, B_n, F, \frac{2^j\alpha}{\lambda_n}), \|\cdot\|_\infty\Big) \\
&\leq \exp\Big(2\frac{2^j\alpha}{\lambda_n}(L_n+1)\log\Big(\frac{(L_n+1)(N_n+1)B_n}{\varepsilon/\mathcal{K}_\ell - \tau_n(L_n+1)((N_n+1)B_n)^{L_n+1}}\Big)\Big).
\end{aligned} \tag{7.36}$$

令 $\varepsilon := \frac{2^j\alpha}{2}$。通过使用与 (7.14) 中类似的论证，我们有

$$\begin{aligned}
&\sum_{j=1}^{\infty} P\Big\{\sup_{g\in\mathcal{G}_{n,j,\alpha}} \Big[\mathbb{E}[g(Z_0)] - \frac{1}{n}\sum_{i=1}^n g(Z_i)\Big] > \frac{2^j\alpha}{2}\Big\} \\
&\leq C_3 \exp(\log\log\ n) \\
&\times \sum_{j=1}^{\infty} \exp\left(2\frac{2^j\alpha}{\lambda_n}(L_n+1)\log\left(\frac{(L_n+1)(N_n+1)B_n}{\frac{2^j\alpha}{2\mathcal{K}_\ell} - \tau_n(L_n+1)((N_n+1)B_n)^{L_n+1}}\right) - \frac{(n2^j\alpha)^2/16}{nC + \log(\frac{n}{n^{1/4}C})(n2^j\alpha/8)^\nu}\right).
\end{aligned}$$

在假设 $L_n \lesssim \log\ n,\ N_n \lesssim n^{\nu_1},\ B_n \lesssim n^{\nu_2},\ \tau_n \leq \frac{\beta_n}{4\mathcal{K}_\ell\ L_n+1}$ 下；在满足关于 $\nu_1 > 0,\ \nu_2 > 0,\ \beta_n, \nu_5,\ \nu_6$ 的条件下，按照 $A_{2,n}$ 中的方法（参见定理 4.1 的证明），可得，

$$\mathbb{E}[B_{1,n}] \lesssim 2\frac{(\log n)^{\nu_5}}{n^{\nu_6}}.$$

现在来处理 $B_{2,n}$。我们选取一个神经网络函数 $h_n^\circ \in \mathcal{H}_\sigma(L_n, N_n, B_n, F)$，使得

$$\mathcal{E}_{Z_0}(h_n^\circ) + J_{\lambda_n,\tau_n}(h_n^\circ) \leq \inf_{h\in\mathcal{H}_\sigma(L_n,N_n,B_n,F)} \Big[\mathcal{E}_{Z_0}(h) + J_{\lambda_n,\tau_n}(h)\Big] + \frac{1}{n}. \tag{7.37}$$

于是，由基本不等式可得，

$$\frac{1}{n}\sum_{i=1}^n \ell(Y_i\widehat{h}_n(X_i)) + J_{\lambda_n,\tau_n}(\widehat{h}_n) \leq \frac{1}{n}\sum_{i=1}^n \ell(Y_i h(X_i)) + J_{\lambda_n,\tau_n}(h),$$

我们有，

$$\begin{aligned}
B_{2,n} &= \frac{2}{n}\sum_{i=1}^n \Big[\ell(Y_i\widehat{h}_n(X_i)) - \ell(Y_i h_n^\circ(X_i))\Big] + 2J_{\lambda_n,\tau_n}(\widehat{h}_n)) + \frac{2}{n}\Big[\ell(Y_i h_n^\circ(X_i)) - \ell(Y_i h_\ell^*(X_i))\Big] \\
&\leq 2J_{\lambda_n,\tau_n}(h_n^\circ) + \frac{2}{n}\Big[\ell(Y_i h_n^\circ(X_i)) - \ell(Y_i h_\ell^*(X_i))\Big] \\
&\leq 2\inf_{h\in\mathcal{H}_\sigma(L_n,N_n,B_n,F)} \Big[\mathcal{E}_{Z_0}(h) + J_{\lambda_n,\tau_n}(h)\Big] + \frac{1}{n}.
\end{aligned}$$

因此

$$\mathbb{E}\Big[\mathcal{E}_{Z_0}(\widehat{h}_n)\Big] \leq 2\inf_{h\in\mathcal{H}_\sigma(L_n,N_n,B_n,F)} \Big\{\mathcal{E}_{Z_0}(h) + \lambda_n\|\theta(h)\|_{\text{clip},\tau_n}\Big\} \vee \frac{(\log n)^{\nu_5}}{n^{\nu_6}}.$$

这就证明了该定理。 ■

## 7.4 定理 5.2 的证明

在假设 $\ell$ 是 $\mathcal{K}_\ell$ –利普希茨的条件下，我们有

$$\begin{aligned}
|\mathcal{E}_{z_0}(h)| &= |\mathbb{E}[\ell(Y_0 h(X_0))] - \mathbb{E}[\ell(Y_0 h_\ell^*(X_0))]| \\
&\leq \mathbb{E}[|\ell(h(X_0)) - \ell(h_\ell^*(X_0))|] \\
&\leq K_\ell \mathbb{E}[|h(X_0) - h_\ell^*(X_0)|]
\end{aligned}$$

因此，

$$\mathcal{E}_{Z_0}(h) \leq \mathcal{K}_\ell \|h - h_\ell^*\|_{1,P_{X_0}}.$$

令 $\epsilon_n = n^{-\frac{\nu_4}{\kappa+1}}$，根据定理 5.1、条件 (5.6)，以及 $\|\theta(h)\|_{\text{clip},\,\tau} \leq \|\theta\|_0$ 对任意 $\tau > 0$ 均成立这一事实，我们有：对任意 $h_\ell^* \in \mathcal{H}^*$，

$$\begin{aligned}
\mathbb{E}\Big[\mathcal{E}_{Z_0}(\widehat{h}_n)\Big] &\lesssim \inf_{h\in\mathcal{H}_\sigma(L_n,N_n,B_n,F,C\epsilon_n^{-\kappa}(\log n)^r)} \Big\{\mathcal{E}_{Z_0}(h) + \lambda_n\|\theta(h)\|_{\text{clip},\tau_n}\Big\} \vee \frac{(\log n)^{\nu_5}}{n^{\nu_6}} \\
&\lesssim \Big[\inf_{h\in\mathcal{H}_\sigma(L_n,N_n,B_n,F,C\epsilon_n^{-\kappa}(\log n)^r)} \|h - h_\ell^*\|_{1,P_{X_0}} + \lambda_n\epsilon_n^{-\kappa}(\log n)^r\Big] \vee \frac{(\log n)^{\nu_5}}{n^{\nu_6}} \\
&\lesssim \frac{(\log n)^{r+\nu_3}}{n^{\frac{\nu_4}{\kappa+1}}} \vee \frac{(\log n)^{\nu_5}}{n^{\nu_6}}.
\end{aligned}$$

因此，

$$\sup_{H\in\mathcal{Q}_H^*:h_\ell^*\in\mathcal{H}^*} \mathbb{E}\Big[\mathcal{E}_{Z_0}(\widehat{h}_n)\Big] \lesssim \frac{(\log n)^{r+\nu_3}}{n^{\frac{\nu_4}{\kappa+1}}} \vee \frac{(\log n)^{\nu_5}}{n^{\nu_6}}. \tag{7.38}$$

■

# 参考文献

[1] Bauer, B., and Kohler, M. 论深度学习如何应对非参数回归中的维数灾难。[2] Chen, J., Du, Y., Liu, L., Zhang, P., and Zhang, W. 基于样本熵和深度神经网络的 Bbs帖子时间序列分析。Entropy 21, 1（2019年），57。[3] Dedecker, J., Doukhan, P., Lang, G., José Rafael, L. R., Louhichi, S., and Prieur, C. 弱依赖。载于《弱依赖：示例与应用》。Springer，2007年，第9–20页。[4] Diop, M. L., and Kengne, W. 具有外生协变量的一般因果时间序列中的推断与模型选择。Electronic Journal of Statistics 16, 1（2022年），116–157。[5] Diop, M. L., and Kengne, W. 针对 $\psi$-弱依赖过程的统计学习。arXiv预印本 arXiv:2210.00088（2022年）。[6] Doukhan, P., and Louhichi, S. 一种新的弱依赖条件及其在矩不等式中的应用。Stochastic Processes and their Applications 84, 2（1999年），313–342。[7] Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A.-r., Jaitly, N., Senior, A., Van-houcke, V., Nguyen, P., Sainath, T. N., et al. 语音识别中的声学建模深度神经网络：四个研究小组的共同观点。IEEE Signal Processing Magazine 29, 6（2012年），82–97。[8] Hwang, E., and Shin, D. W. 关于 $\psi$-弱依赖序列的指数不等式。Communications for Statistical Applications and Methods 21, 3（2014年），245–251。[9] Imaizumi, M., and Fukumizu, K. 深度神经网络在估计超曲面上具有奇异性的函数方面的优势。Journal of Machine Learning Research 23（2022年），1–54。[10] Kengne, W. 弱依赖下深度学习的超额风险界。arXiv预印本 arXiv:2302.07503（2023年）。[11] Kengne, W., and Modou, W. 针对 $psi$-弱依赖过程的深度学习。arXiv预印本 arXiv:2302.00333（2023年）。[12] Kim, Y., Ohn, I., and Kim, D. 用于分类的深度神经网络的快速收敛速率。Neural Networks 138（2021年），179–197。[13] Kingma, D. P., and Ba, J. Adam：一种随机优化方法。arXiv预印本 arXiv:1412.6980（2014年）。[14] Kohler, M., and Krzyzak, A. 相关数据下回归问题中深度循环神经网络估计量的收敛速率。arXiv预印本 arXiv:2011.00328（2020年）。[15] Krizhevsky, A., Sutskever, I., and Hinton, G. E. 使用深度卷积神经网络进行 ImageNet 分类。Communications of the ACM 60, 6（2017年），84–90。

[16] Kurisu, D., Fukami, R., and Koike, Y. 用于非参数时间序列回归的自适应深度学习。*arXiv* 预印本 *arXiv:2207.02546*（2022年）。[17] Ma, M., and Safikhani, A. 深度神经网络在时间依赖观测下的理论分析。*arXiv* 预印本 *arXiv:2210.11530*（2022年）。[18] Ohn, I., and Kim, Y. 一般激活函数下深度神经网络的光滑函数逼近。*Entropy 21*, 7（2019年）, 627。[19] Ohn, I., and Kim, Y. 深度神经网络的非凸稀疏正则化及其最优性。*Neural Computation 34*, 2（2022年）, 476–517。[20] Schmidt-Hieber, J. 流形上函数的深度ReLU网络逼近。*arXiv* 预印本 *arXiv:1908.00695*（2019年）。[21] Schmidt-Hieber, J. 使用ReLU激活函数的深度神经网络进行非参数回归。*The Annals of Statistics 48*, 4（2020年）, 1875–1897。[22] Tsuji, K., and Suzuki, T. 变指数Besov空间上回归问题的深度学习估计误差分析。*Electronic Journal of Statistics 15*, 1（2021年）, 1869–1908。[23] Wainwright, M. J. 高维统计：非渐近视角，第48卷。剑桥大学出版社，2019年。[24] Zhang, T. 稀疏正则化多阶段凸松弛的分析。*Journal of MachineLearning Research 11*, 3（2010年）。