arXiv:2305.06230v1 [stat.ML] 10 May 2023

弱依赖条件下一般损失函数下的惩罚深度神经网络估计量

2023年5月11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 电子邮件：*william.kengne@cyu.fr*；*modou.wade@cyu.fr*

摘要：本文针对弱依赖过程，采用一大类损失函数，构建稀疏惩罚深度神经网络预测器，以实现学习。我们研究了一个一般框架，其中包括回归估计、分类和时间序列预测。 $\cdots$本文考虑了 ψ-弱依赖结构；对于有界观测这一特例，也采用 θ_∞-系数。在 θ_∞-弱依赖情形下，本文给出了深度神经网络预测器类中的非渐近泛化界。对于学习 ψ 和 θ_∞-弱依赖过程，本文建立了稀疏惩罚深度神经网络估计量超额风险的预言不等式。当目标函数足够光滑时，这些超额风险的收敛速度接近 $\mathcal{O}(n^{-1/3})$。文中给出了一些模拟结果，并研究了对 Vit´oria 都市区颗粒物浓度的预测应用。

关键词：深度神经网络、稀疏性、弱依赖、超额风险、收敛速率。

1 引言

近年来，关于深度神经网络（DNNs）的研究在这类预测器的理论性质方面取得了显著进展；例如，[19], [23], [3], [24], [26], [14] [10] 给出了一些基于独立同分布（i.i.d.）观测所得的结果，而 [4], [16], [17], [18], [12], [11] 则研究了数据存在依赖或非独立同分布的情形。然而，深度神经网络的正则化仍备受关注。学习算法所引发的隐式正则化效应被称为隐式正则化，目前尚未得到完全理解。此外，文献中还研究了显式正则化方法，这些方法有助于提高深度神经网络预测器的准确性；例如， [20], [17], [13], 讨论了一种稀疏惩罚正则化方法。

考虑来自平稳遍历过程 $\{Z_t = (X_t, Y_t)\ t \in \mathbb{Z}\}$ 的观测值 $D_n := \{(X_1, Y_1) \cdots, (X_n, Y_n)\}$（训练样本）。该过程取值于 $\mathcal{Z} = \mathcal{X} \times \mathcal{Y} \subset \mathbb{R}^d \times \mathbb{R}$（其中 $d \in \mathbb{N}$），且 $\mathcal{X}$ 是

[1]本研究在 ANR 项目 RISK: ANR-17-CE26-0001-01 和 CY 卓越计划（“未来投资”资助项目 ANR-16-IDEX-0008）及“EcoDep”项目 PSI-AAP2020-0000000013 的支持下开展[2]并获 MME-DII 卓越中心（ANR-11-LABEX-0023-01）资助

输入空间， $\mathcal{Y}$ 是输出空间。令 $\mathcal{F}:=\mathcal{F}(\mathcal{X},\mathcal{Y})$ 为从 $\mathcal{X}$ 到 $\mathcal{Y}$ 的可测函数集合。下文中，我们考虑损失函数 $\ell:\mathbb{R}\times\mathcal{Y}\to[0,\infty)$，对于预测器 $h\in\mathcal{F}$，其风险定义为:

$$R(h)=\mathbb{E}_{Z_0}[\ell(h(X_0),Y_0)]. \tag{1.1}$$

我们将 h^*记为 $\mathcal{F}$中的目标预测器（假设其存在），其满足,

$$R(h^*)=\inf_{h\in\mathcal{F}}R(h). \tag{1.2}$$

对任意 $h\in\mathcal{F}$，定义超额风险为,

$$\mathcal{E}_{Z_0}(h)=R(h)-R(h^*),\ \text{with}\ Z_0=(X_0,Y_0). \tag{1.3}$$

The $\widehat{}$目标是构建一个 DNN 预测器 h_n，使其超额风险尽可能接近零。

我们关注一种带惩罚项的经验风险最小化方法。将稀疏惩罚 DNN（SPDNN）预测器定义为,

$$\widehat{h}_n=\underset{h\in\mathcal{H}_\sigma(L_n,N_n,B_n,F)}{\operatorname{argmin}}\left[\frac{1}{n}\sum_{i=1}^{n}\ell(h(X_i),Y_i)+J_n(h)\right], \tag{1.4}$$

其中 $\mathcal{H}_\sigma(L_n,\ N_n,\ B_n,\ F)$是一个 DNN 类（见（3.3））， 其激活函数为 σ，并采用适当选取的架构参数 L_n, $N_n,\ B_n,\ F>0$; 而 $J_n(h)$表示一个稀疏惩罚项，定义为,

$$J_n(h)\coloneqq J_{\lambda_n,\tau_n}(h)\coloneqq\lambda_n\|\theta(h)\|_{\mathrm{clip},\tau_n}, \tag{1.5}$$

对于调节参数 $\lambda_n>0,\ \tau_n>0$，且 $\theta(h)$ 是与 h 相关的网络参数（见 (3.2)）。$\|\cdot\|_{\mathrm{clip},\tau}$表示截断的 L_1 范数, 截断阈值为 $\tau>0$，见 [27]，其定义为

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

其中， $\theta=(\theta_1,\ \cdots,\ \theta_p)^T$是一个 p维向量， ^{T}表示转置。下文中，我们还考虑非惩罚预测量，其定义为,

$$\widehat{h}_{n,NP}=\underset{h\in\mathcal{H}_\sigma(L,N,B,F,S)}{\operatorname{argmin}}\left[\frac{1}{n}\sum_{i=1}^{n}\ell(h(X_i),Y_i)\right], \tag{1.6}$$

对于某种网络架构 $L,\ N,\ B,\ F,\ S>0$, 其中 $\mathcal{H}_\sigma(L,\ N,\ B,\ F,\ S)$ 表示一类稀疏性约束深度神经网络，其定义见 (3.4)。

$\widehat{}$预测器 $h_{n,NP}$基于稀疏性约束 S，具有良好的理论性质。对于观测值独立同分布（i.i.d.）的回归和分类问题，它能够达到最优收敛速率，参见例如 [24], [14]。在观测值弱相关的一般框架下，该 DNN 估计量可以达到不高于或接近于 $\mathcal{O}(n^{-1/3})$的学习速率，参见 [12] 以及下文的定理 4.2。然而，$\widehat{}$由于稀疏性约束具有离散性质，计算 $h_{n,NP}$ 较为困难。为克服这一困难，已有研究提出了稀疏惩罚正则化等方法。 [20]已将 SPDNN 预测器用于独立同分布观测值的非参数回归和分类。他们建立了超额风险的预言不等式，并推导出收敛速率。 [17]提出了在 β-混合条件下用于非参数时间序列回归的 SPDNN 估计量，并给出了一个泛化误差界。

并且他们已经证明，该估计量能够达到极小极大意义下的最优收敛速率，且仅相差一个多对数因子。[13]已经将 SPDNN预测器应用于 ψ-弱相关过程的非参数回归和分类问题。他们建立了超额风险的预言机不等式，并给出了收敛速率，该速率在渐近意义下可以接近 $\mathcal{O}(n^{-1/2})$。

然而，上述研究主要关注非参数回归这一特定问题和/或二元分类的某种特定情形。尽管 [13]处理了一类一般的弱依赖过程，但该论文仍存在一些值得强调的局限性：(i) 与 [20] 和 [17], 一样，他们关注的是采用 L_2 损失函数的非参数回归；(ii) 与 [20]和 [17], 一样，他们对其模型施加了次高斯假设；(iii) 他们的模型具有较强的限制性，例如，无法应用于非线性 AR-ARCH 类型的过程。

本文研究了由 DNNs 对 ψ-弱依赖过程 $\{Z_t = (X_t, Y_t) \quad t \in \mathbb{Z}\}$ 的学习问题，该过程的取值为 $\mathcal{Z} = \mathcal{X} \times \mathcal{Y} \subset \mathbb{R}^d \times \mathbb{R}$（其中 $d \in \mathbb{N}$）。我们处理了一个包含回归估计、分类、时间序列预测、 $\cdots$ 以及广泛损失函数类别的一般框架。推导了 θ_∞-弱依赖条件下的泛化界。对于 ψ 和θ_∞-弱依赖过程，均采用 SPDNN估计量。更确切地说，我们将讨论以下问题。

(i) **针对 θ_∞-弱依赖过程的泛化界**。我们考虑在函数类 $\mathcal{H}_\sigma(L, N, B, F, S)$（见（3.4））上，对所有 $L, N, B, F, S > 0$，采用经验风险最小化（ERM）原理。我们得到了一个收敛速率为 $\mathcal{O}(n^{-1/4})$ 的非渐近界。对于 ψ-弱依赖下的非渐近泛化界，该收敛速率优于 [12] 中得到的速率。

(ii) **在 ψ 和 θ_∞-弱依赖下 SPDNN预测器的超额风险预言不等式**。考虑不受稀疏性约束的深度神经网络预测器类 $\mathcal{H}_\sigma(L_n, N_n, B_n, F)$（见（3.3））， 其中 $L_n, N_n, B_n, F > 0$。我们给出关于 L_n, N_n, B_n, F 以及 λ_n, τ_n （见（1.5））的条件，使得对于$\widehat{}$基于一般损失函数类的 SPDNN预测器 h_n （见（1.4）），无论是 ψ 过程还是 θ_∞-弱依赖过程，都可建立超额风险预言不等式。

(iii) **SPDNN预测器超额风险的收敛速率**。当目标函数属于一类 H¨older 光滑函数，且满足 θ_∞-弱依赖时，我们建立了 SPDNN预测器 $\widehat{h}_n$超额风险的一个非渐近界，其速率为 $\mathcal{O}(n^{-1/3})$。

(iv) **仿射因果模型的应用及在 Vit PM_{10} oria 大都市区´的预测**。我们考虑一类具有外生协变量的仿射因果模型，并聚焦于非参数回归。对于 L_1 和 L_2 损失函数，结果表明，(i)、(ii)、(iii)中的结果适用于此类模型。在 Vit´oria 大都市区对 PM_{10} （直径小于 $10\mu m$ 的颗粒物）的应用表明，SPDNN预测器的表现略优于文献 ([6]) 中提出的模型。

本文其余部分安排如下。第 2 节介绍一些记号和假设，并回顾 ψ 与 θ_∞-弱依赖的定义。第 3 节定义所考虑的深度神经网络类别。第 4 节推导深度神经网络学习 θ_∞-弱依赖过程的泛化界。第 5 节给出预言机不等式以及 SPDNN 预测器的超额风险界。第 6 节探讨其在仿射因果模型非参数回归中的应用，第 7 节则聚焦于 PM_{10}预测的应用。第 8 节专门证明主要结果。

2 记号与假设

下文中，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 $\epsilon > 0$，令 $B(h, \epsilon)$ 表示以 h为中心、半径为 ϵ 的球，即

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

其中，$\|\cdot\|_\infty$ 表示（2.1）中定义的上确界范数。令 $\mathcal{H} \subset \mathcal{F}(E_1, E_2)$，$\epsilon$-覆盖数 $\mathcal{N}(\mathcal{H}, \epsilon)$ 表示覆盖 $\mathcal{H}$所需的半径为 ϵ 的球的最小数量；也就是说，

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

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

$$\|h\|_\infty = \sup_{x \in E_1} \|h(x)\|_{E_2},\ \|h\|_{\infty, U} = \sup_{x \in U} \|h(x)\|_{E_2}, \tag{2.1}$$

and

$$\mathrm{Lip}_\alpha(h) \coloneqq \sup_{x_1, x_2 \in E_1,\ x_1 \ne x_2} \frac{\|h(x_1) - h(x_2)\|_{E_2}}{\|x_1 - x_2\|_{E_1}^\alpha} \text{ for any } \alpha \in [0, 1]. \tag{2.2}$$

对于任意 $\mathcal{K}_\ell > 0$和 $\alpha \in [0, 1]$, $\Lambda_{\alpha, \mathcal{K}_\ell}(E_1, E_2)$（或当 $E_2 \subseteq \mathbb{R}$时，简记为 $\Lambda_{\alpha, \mathcal{K}_\ell}(E_1)$）是满足以下条件的函数 h 集合：对于某个 $u \in \mathbb{N}$，$E_1^u \to E_2$，满足 $\|h\|_\infty < \infty$ 且 $\mathrm{Lip}_\alpha(h) \le \mathcal{K}_\ell$。当 $\alpha = 1$时，我们令 $\mathrm{Lip}_1(h) = \mathrm{Lip}(h)$ 且 $\Lambda_1(E_1) = \Lambda_{1,1}(E_1, \mathbb{R})$。

我们来定义弱依赖结构，参见 [8] 和 [5]。设 E 是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 ψ: $[0,\infty)^2\times\mathbb{N}^2 \to [0,\infty)$ 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: $E^u \to \mathbb{R}$, g_2: $E^v \to \mathbb{R}$, $(u, v \in \mathbb{N})$ and for any $u-$tuple $(s_1, \cdots, s_u)$ and any $v-$tuple $(t_1, \cdots, t_v)$ with $s_1 \le \cdots \le s_u \le s_u + r \le t_1 \le \cdots \le t_v$, the following inequality is fulfilled:*

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

例　如，以下对 ψ 的选择会导出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，记作 $\epsilon(r) = \theta(r)$；
- $\psi(\mathrm{Lip}(g_1),\ \mathrm{Lip}(g_2)\ u,\ v) = u\mathrm{Lip}(g_1) + v\mathrm{Lip}(g_2)$：$\eta$-弱依赖，记作 $\epsilon(r) = \eta(r)$；
- $\psi(\mathrm{Lip}(g_1),\ \mathrm{Lip}(g_2)\ u,\ v) = uv\mathrm{Lip}(g_1) \cdot \mathrm{Lip}(g_2)$：$\kappa$-弱依赖，记作 $\epsilon(r) = \kappa(r)$；
- $\psi(\mathrm{Lip}(g_1),\ \mathrm{Lip}(g_2)\ u,\ v) = u\mathrm{Lip}(g_1) + v\mathrm{Lip}(g_2) + uv\mathrm{Lip}(g_1)\cdot\mathrm{Lip}(g_2)$：$\lambda$-弱依赖，则记 $\epsilon(r) = \lambda(r)$。

下文中，对上述 ψ 的四种选择，分别令

$$\Psi(u, v) = 2v,\ \Psi(u, v) = u + v,\ \Psi(u, v) = uv,\ \Psi(u, v) = (u + v + uv)/2. \tag{2.3}$$

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$\ell$：$\mathbb{R} \times \mathcal{Y} \to [0, \infty)$、具有激活函数$\sigma : \mathbb{R} \to \mathbb{R}$ 的深度神经网络类 $\mathcal{H}_\sigma(L_n, N_n, B_n, F)$ $(L_n, N_n, B_n, F > 0)$，以及以下一组假设。

(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 $\ell \in \Lambda_{1,\mathcal{K}_\ell}(\mathbb{R}\times\mathcal{Y})$ 和 $M_n = \sup_{h\in\mathcal{H}_\sigma(L_n,N_n,B_n,F)}\sup_{z\in\mathcal{Z}} |\ell(h,z)| < \infty$。

在**(A2)**条件下，我们得到，

$$G_n := \sup_{h_1,h_2\in\mathcal{H}_\sigma(L_n,N_n,B_n,F),h_1\neq h_2}\sup_{z\in\mathcal{Z}} \frac{|\ell(h_1,z)-\ell(h_2,z)|}{\|h_1-h_2\|_\infty} < \infty. \tag{2.4}$$

我们来设定弱依赖性假设。

(A3)： 令 Ψ： $[0,\infty)^2 \times \mathbb{N}^2 \to [0,\infty)$ 为 **(2.3)** 中的一种选择。过程 $\{Z_t = (X_t, Y_t)\ t \in \mathbb{Z}\}$ 是平稳遍历的，且具有 $(\Lambda_1(\mathcal{Z})\ \psi, \epsilon)$-弱依赖性，使得存在满足以下条件的 $L_1,\ L_2,\ \mu \geq 0$。

$$\sum_{j\geq 0}(j+1)^k \epsilon_j \leq L_1 L_2^k (k!)^\mu \quad \text{for all } k \geq 0. \tag{2.5}$$

在下文中，我们还使用 $\theta_{\infty,k}(1)$ 系数，它是 [5], 的一种 γ-混合版本，参见 [21], ，另见 [22]。

定义 2.2. 当 $\{Z_t = (X_t, Y_t)\ t \in \mathbb{Z}\}$ 是一个有界平稳遍历过程时，对任意$k > 0$，定义$\theta_{\infty,\ k}(1)$-弱依赖系数如下：

$$\theta_{\infty,k}(1) := \sup_{h\in\Lambda_1(\mathcal{Z}),0<j_1<\cdots<j_k} \Big\| |\mathbb{E}[h(Z_{j_1},\cdots,Z_{j_k})|Z_t, t\leq 0] - \mathbb{E}[h(Z_{j_1},\cdots,Z_{j_k})]| \Big\|_\infty, \tag{2.6}$$

其中， $\|V\|_\infty$ 表示随机变量 $\|V\|$的本质上确界；对于随机向量 V， $\Lambda_1(\mathcal{Z})$的定义见第 *2* 节。

(A4) 过程 $\{Z_t = (X_t, Y_t)\ t \in \mathbb{Z}\}$ 有界、平稳且遍历，其 $\theta_{\infty,k}(1)$-弱依赖系数满足对所有 $k > 0, \theta_{\infty,k}(1) \leq \mathcal{C}$, 其中某个 $\mathcal{C} > 0$。

3 深度神经网络

回顾一下，DNN 架构 $(L,\mathbf{p})$ 表示一个正整数 L，称为隐藏层数或深度，以及一个宽度向量 $\mathbf{p} = (p_0, p_1, \cdots, p_{L+1}) \in \mathbb{N}^{L+2}$。因此，具有网络架构 $(L, \mathbf{p})$ 的 DNN 是任意具有以下形式的函数，

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

其中， A_j: $\mathbb{R}^{p_{j-1}} \to \mathbb{R}^{p_j}$是一个线性仿射映射，定义为 $A_j(x) := W_j x + \mathbf{b}^j$，其中给定了 $p_{j-1}\times p_j$权重矩阵 W_j 和偏移向量 $\mathbf{b}_j \in \mathbb{R}^{p_j}$； σ_j： $\mathbb{R}^{p_j} \to \mathbb{R}^{p_j}$ 是一个逐元素非线性激活映射，对所有 $z = (z_1, \cdots, z_{p_j})^T$，其定义为 $\sigma_j(z) = (\sigma(z_1)\ \cdots, \sigma(z_{p_j}))^T$，而 T表示转置。对于形式为 **(3.1)** 的 DNN，记 $\theta(h)$ 为其参数向量，即

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

其中， $vec(W)$表示通过拼接矩阵 W的列向量得到的向量。下文中，我们将讨论一个激活函数 $\sigma : \mathbb{R} \to \mathbb{R}$，并用$\mathcal{H}_{\sigma,\ p_0,\ p_{L+1}}$表示深度神经网络类。

具有 p_0 维输入和 p_{L+1} 维输出的预测变量。在此处考虑的框架下， $p_0 = d$ 和 $p_{L+1} = 1$ 。对于如 (3.1) 中的 DNN h ，分别用 depth(h) 和 width(h) 表示其深度和宽度；即，depth(h)$= L$，且 width(h) $=\max_{1\le j\le L} p_j$。对于任意正常数 L, N, B 和 F，令

$$\mathcal{H}_\sigma(L,N,B) \coloneqq \big\{h \in \mathcal{H}_{\sigma,q,1} : \mathrm{depth}(h) \le L, \mathrm{width}(h) \le N, \|\theta(h)\|_\infty \le B\big\},$$

and

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

还定义了稀疏度水平为 $S > 0$ 的稀疏约束深度神经网络类别:

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

其中，对于所有 $x = (x_1, \dots, x_p)^T \in \mathbb{R}^p (p \in \mathbb{N})$， $\|x\|_0 = \sum_{i=1}^p \mathbb{1}(x_i \neq 0)$ $\|x\| = \max_{1\le i\le p} |x_i|$。

4 弱 θ_∞-弱依赖下的泛化界

本节聚焦于 θ_∞-弱依赖过程上，针对一类深度神经网络预测器 $\mathcal{H}_\sigma(L,N,B,F,S)$ 的 ERM 算法的泛化界，其中对所有 $L, N, B, F, S > 0$ 均成立。下述定理给出了函数类 $\mathcal{H}_\sigma(L,N,B,F,S)$ 上风险与其经验版本之间的一致指数集中不等式。

定理 4.1. 假设*(A1)*、*(A2)*（其中包含 $L_n = L$、 $N_n = N$、 $B_n = B$ 和 $M_n = M$）以及*(A4)*均成立。对于所有 $n \in \mathbb{N}$、 $\varepsilon > 0$ 和 $\nu_0 \in (0,1)$，我们有，

$$\begin{aligned} &P\left\{\sup_{h\in\mathcal{H}_\sigma(L,N,B,F,S)} \left[R(h) - \widehat{R}_n(h)\right] > \varepsilon\right\} \\ &\le \mathcal{N}\left(\mathcal{H}_\sigma(L,N,B,F,S), \frac{\varepsilon}{4G}\right) \exp\left(-n^{\nu_0}\varepsilon + \frac{n^{2\nu_0-1}(M+\theta_{\infty,n}(1))^2}{2}\right), \end{aligned} \tag{4.1}$$

其中 $\theta_{\infty,\ n}(1)$ 在 *(2.6)* 中定义，且 $G = G_n$ 由 *(2.4)* 给出，其中 $L_n = L, N_n = N, B_n = B$.

考虑通过最小化经验风险得到的预测器 $h_{n,NP}$ （见 (1.6)）。设 $h_{\mathcal{H}_\sigma}$ 为目标神经网络（假定其存在），定义为

$$h_{\mathcal{H}_\sigma} = \operatorname*{argmin}_{h\in\mathcal{H}_\sigma(L,N,B,F,S)} R(h). \tag{4.2}$$

下面的定理给出了一个泛化界。

定理 4.2. 假设定理 *4.1* 的条件成立。令 $\eta, \nu_0 \in (0,1)$。假设

$$n > \max\left(n_0(L,S,M,\nu_0,\theta_{\infty,n}(1)), \left(\frac{1}{M}\left(C_1 - 2L(S+1)\log(2M) - \log(\eta)\right)_+\right)^{1/2\nu_0}\right),$$

其中

$$C_1 = 2L(S+1)\log\left(4GC_\sigma L(N+1)(B\vee 1)\right), \ n_0(L,S,M,\nu_0,\theta_{\infty,n}(1)) \in \mathbb{N} \textit{ defined at (8.12).}$$

(i) 以至少 $1-\eta$ 的概率，我们有

$$R\left(\widehat{h}_{n,NP}\right)-\widehat{R}_n\left(\widehat{h}_{n,NP}\right)\leq\varepsilon\left(n,\eta,\nu_0\right), \tag{4.3}$$

其中， $\varepsilon(n$、 η、 $\nu_0)\leq\frac{2M}{n^{\nu_0/2}}$且 $\widehat{h}_{n,\ NP}$在（***1.6***）中定义。

(ii) 以至少 $1-2\eta$ 的概率，我们有

$$R\left(\widehat{h}_{n,NP}\right)-R\left(h_{\mathcal{H}_\sigma}\right)\leq\varepsilon\left(n,\eta,\nu_0\right)+\varepsilon'\left(n,\eta,\nu_0\right), \tag{4.4}$$

其中 $\varepsilon'(n,\eta,\nu_0)=(M+\theta_{\infty,\ n}\ 1\ ^2+0)\log(1/\eta)\,n^{\nu_0}$。$2n^{1-\nu_0}$

注 4.3】【。 非渐近界 ***(4.3)*** 和 ***(4.4)*** 的学习率阶为 $\mathcal{O}(n^{-1/4})$.

5 SPDNN预测器超额风险的预言机不等式

本节在 ψ and θ_∞-弱依赖条件下，给出了 SPDNN 超额风险的一些预言机不等式。当目标预测器 h^* （见 (1.2)）属于一类 H¨older 函数时，我们推导了该超额风险的一个上界。下文中，我们采用以下记号： $a\vee b=$ max$(a,\ b)$、 $a_n\lesssim b_n$ or $b_n\gtrsim a_n$ 若存在正常数 $C>0$ 使得 $a_n\leq Cb_n$ 对所有 $n\in\mathbb{N}$成立、 $a_n\asymp b_n$ if $a_n\lesssim b_n$ and $a_n\gtrsim b_n$。 在下文的定理 5.1 中，记 $C_{1,\ n}=4M_n^2\Psi(1,\ 1)L_1$和 $C_{2,\ n}=2M_nL_2\max(\ ^{2^{3+\mu}}_{\Psi(1,\ 1)},1)$，其中 Ψ是 (2.3) 中的一个函数，且满足假设 **(A2)** 和 **(A3)**。

定理 5.1. 假设 ***(A1)-(A3)*** 成立。令 $L_n\lesssim$对数 $n, N_n\lesssim n^{\nu_1}, 1\leq B_n\lesssim n^{\nu_2}, F>0$，其中某些$\nu_1>0, \nu_2>0$。则由下式给出的 *SPDNN* 估计量，

$$\widehat{h}_n=\underset{h\in\mathcal{H}_\sigma(L_n,N_n,B_n,F)}{\operatorname{argmin}}\left[\frac{1}{n}\sum_{i=1}^{n}\ell\left(h(X_i),Y_i\right)+\lambda_n\|\theta(h)\|_{clip,\tau_n}\right], \tag{5.1}$$

其中 $\lambda_n\asymp$(对数 $n)^{\nu_3}/n^{\nu_4}, \tau_n\leq\ ^{\rho_n}()((N_n+1)B_n)^{L_n+1}$ 其中 $\rho_n:=\Big($ $\frac{C_{1,\ n}/(}{4\mathcal{K}_\ell\ L_n+1}$ $2C_{2\ n}^{1/(\mu+2)}\ ^{(\mu+2)/(2\mu+3)}{}_{,\ )}\Big)$，满足

$$\mathcal{E}_{Z_0}(\widehat{h}_n)\leq 2\left\{\inf_{h\in\mathcal{H}_\sigma(L_n,N_n,B_n,F)}\left[\mathcal{E}_{Z_0}(h)+J_{\lambda_n,\tau_n}(h)\right]\right\}\vee\frac{C_\mu\left(C_{1,n}/(2C_{2,n}^{1/(\mu+2)})\right)^{(\mu+2)/(2\mu+3)}}{n^{(\mu+1)/(2\mu+3)}},$$

对所有 $n\geq N$，对于某些依赖于 $\mu>0$, $\mathcal{K}_\ell>0$ 的 $N\in\mathbb{N}, C_\mu>0$，当 $\nu_3,\nu_4>0$ 满足 $\nu_1+\nu_2+\nu_4<1$ $\mu+2$时。

下一个定理讨论 θ_∞-弱依赖条件。

定理 5.2. 假设满足 ***(A1)-(A2)*** 和 ***(A4)***。令 $L_n\lesssim\log n, N_n\lesssim n^{\nu_1}, 1\leq B_n\lesssim n^{\nu_2}, F>0$，其中 $\nu_1>0$, $\nu_2>0$。那么，*SPDNN*估计量为：

$$\widehat{h}_n=\underset{h\in\mathcal{H}_\sigma(L_n,N_n,B_n,F)}{\operatorname{argmin}}\left[\frac{1}{n}\sum_{i=1}^{n}\ell\left(h(X_i),Y_i\right)+\lambda_n\|\theta(h)\|_{clip,\tau_n}\right], \tag{5.2}$$

with $\lambda_n \asymp (\log n)^{\nu_3}/n^{\nu_4}$, $\tau_n \le \dfrac{\rho_n}{4\mathcal{K}_\ell(L_n+1)((N_n+1)B_n)^{L_n+1}}$ *where* $\rho_n = 1/n^{2\nu_6}$, $\mathcal{K}_\ell > 0$, *satisfies*

$$\mathbb{E}\Big[\mathcal{E}_{Z_0}(\widehat{h}_n)\Big] \lesssim 2\left\{\inf_{h\in\mathcal{H}_\sigma(L_n,N_n,B_n,F)}\Big[\mathcal{E}_{Z_0}(h)+J_{\lambda_n,\tau_n}(h)\Big]\right\} \vee \frac{C}{n^{2\nu_6}}, \tag{5.3}$$

对所有 $n \ge N$，存在某些 $N \in \mathbb{N}$, $C > 0$, $\nu_3 > 0$, $\nu_4 > 0$, $\nu_5 \in (0,1)$ 使得 $\nu_4 + 2\nu_6 + \nu_1 + \nu_2 < \nu_5$ 和 $\nu_6 < \dfrac{1-\nu_5}{2}$。

上述定理 5.1 和 5.2 可用于推导 SPDNN预测器的超额风险界，如推论 5.3 所示。考虑一类 H¨older 光滑函数。对于 $U \subset \mathbb{R}^d, \beta = (\beta_1, \cdots, \beta_d)^T \in \mathbb{N}^d$, $x = (x_1, \cdots, x_d)^T \in U$，令

$$|\beta| = \sum_{i=1}^{d} \beta_i \text{ and } \partial^\beta = \frac{\partial^{|\beta|}}{\partial^{\beta_1} x_1, \cdots, \partial^{\beta_d} x_d}.$$

对于任意 $s > 0$，H¨older 空间 $\mathcal{C}^s(U)$ 是由满足以下条件的函数 $h : U \to \mathbb{R}$ 组成的集合：对任意 $\beta \in \mathbb{N}^d$，满足 $|\beta| \le [s]$, $\|\partial^\beta h\|_\infty < \infty$；且对任意 $\beta \in \mathbb{N}^d$，满足 $|\beta| = [s]$, $\text{Lip}^{s-[s]}(\partial^\beta h) < \infty$（其中 $[x]$ 表示 x 的整数部分）。该空间配备的范数为

$$\|h\|_{\mathcal{C}^s(U)} = \sum_{0\le|\beta|\le[s]} \|\partial^\beta h\|_\infty + \sum_{|\beta|=[s]} \text{Lip}_{s-[s]}(\partial^\beta h).$$

对于任意 $s > 0$, $U \subset \mathbb{R}^d$ 和 $\mathcal{K} > 0$，令

$$\mathcal{C}^{s,\mathcal{K}}(U) = \{h \in \mathcal{C}^s(U),\ \|h\|_{\mathcal{C}^s(U)} \le \mathcal{K}\}.$$

推论 *5.3.* 假设 *(A1)-(A2)*、*(A4)* 成立，并且对于某个 $\mathcal{K} > 0$，有 $h^* \in \mathcal{C}^{s,\mathcal{K}}(\mathcal{X})$，其中 h^* 在 *(1.2)* 中定义。令 $L_n, N_n, B_n, F > 0, \lambda_n, \tau_n, \rho_n$ 如定理 *(5.2)* 中所给，且 $\nu_1 = \dfrac{d\nu_4}{s}(1+d/s)$ $\nu_2 = 4d\nu_4(s+1)(1+d/s)$，并且 $\nu_5 \in (0,1)$。考虑深度神经网络类 $\mathcal{H}_\sigma(L_n, N_n, B_n, F)$，其中激活函数 σ 或为分段线性函数，或为局部二次函数，并固定一个区间 $I \subseteq [0,1]$（见 *[12]*）。那么，在 *(5.2)* 中定义的 *SPDNN*估计量满足：

$$\mathbb{E}\Big[\mathcal{E}_{Z_0}(\widehat{h}_n)\Big] \lesssim \frac{2\nu_4(\log n)^{\nu_3+1}}{(1+d/s)n^{\nu_4/(1+d/s)}} \vee \frac{C}{n^{2\nu_6}}, \tag{5.4}$$

对所有 $n \ge N$，存在某些 $N \in \mathbb{N}$、 $C > 0$、 $\nu_3 > 0$、 $\nu_4 > 0$、 $\nu_4 + 2\nu_6 + \nu_1 + \nu_2 < \nu_5$ 和 $\nu_6 < \dfrac{1-\nu_5}{2}$。

由 (5.4) 可知，若目标预测器 h^* 足够光滑，使得 $s \gg d$，通过选择ν_1， ν_2 接近 0、 ν_5 接近 2/3，可得 SPDNN预测器的超额风险收敛速率接近 $\mathcal{O}(n^{-1/3})$。

6 非参数回归中的应用

在本节中，我们利用深度神经网络对一大类自回归模型进行非参数回归。

6.1 仿射因果模型

设 $(\mathcal{X}_t)_{t\in\mathbb{Z}}$ 是取值于 $\mathbb{R}^{d_x}$ $(d_x \in \mathbb{N})$ 的协变量过程。考虑具有外生协变量的仿射因果模型类（见 ([6])），其形式为

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

其中 $\mathcal{M}, f$: $\mathbb{R}^p \times \mathbb{R}^{d_x} \to \mathbb{R}, (p \in \mathbb{N})$ 是两个可测函数，而 $(\xi_t)_{t\in\mathbb{Z}}$ 是一列独立同分布的中心化随机变量，满足 $\mathbb{E}[\xi_0^r] < \infty$，其中某个 $r \geq 2$ 和 $\mathbb{E}[\xi_0^2] = 1$。为研究模型 (6.1) 的稳定性，[6] 对函数 $f, \mathcal{M}$ 或 $\mathcal{M}^2$ 设定了以下类 Lipschitz 条件。将任意向量空间中的零向量记为 0。对于 $\Psi = f$ 或 $\mathcal{M}$，考虑以下假设:

假设 A(Ψ): $|\Psi(0; 0)| < \infty$，且存在一列非负实数 $(\alpha_{k,Y}(\Psi))_{1\leq k\leq p}$ 和 $\alpha_{\mathcal{X}}(\Psi) > 0$，满足 $\sum_{k=1}^p \alpha_{k,Y}(\Psi) < \infty$, $\alpha_{\mathcal{X}}(\Psi) < \infty$；使得对任意 $(y, u), (y', u') \in \mathbb{R}^p \times \mathbb{R}^{d_x}$,

$$|\Psi(y; u) - \Psi(y'; u')| \leq \sum_{k=1}^p \alpha_{k,Y}(\Psi)|y_k - y'_k| + \alpha_{\mathcal{X}}(\Psi)|u - u'|.$$

对于 ARCH-X 类型模型中的函数 $H = \mathcal{M}^2$，我们作出以下假设。

假设 A(H): 假设 $f = 0$, $|\mathcal{M}(0;0)| < \infty$，并且存在一个非负实数序列 $(\alpha_{k,\ Y}(H))_{1\leq k\leq p}$ 和 $\alpha_{\mathcal{X}}(H) > 0$，满足 $\sum_{k=1}^p \alpha_{k,\ Y}(H) < \infty$, $\alpha_{\mathcal{X}}(H) < \infty$；使得对于任意 $(y,\ u)$, $(y',\ u') \in \mathbb{R}^p \times \mathbb{R}^{d_x}$,

$$|H(y, u) - H(y', u')| \leq \sum_{k=1}^p \alpha_{k,Y}(H)|y_k^2 - y_k'^2| + \alpha_{\mathcal{X}}(H)|u - u'|.$$

对于以下内容，采用如下约定：若 $\mathbf{A}(\mathcal{M})$ 成立，则 $\alpha_{k,Y}(H) = \alpha_{\mathcal{X}}(H) = 0$ 对所有 $1 \leq k \leq p$ 均成立；若 $\mathbf{A}(H)$ 成立，则$\alpha_{k,Y}(\mathcal{M}) = \alpha_{\mathcal{X}}(\mathcal{M}) = 0$ 对所有 $1 \leq k \leq p$ 均成立。对协变量施加自回归型结构:

$$\mathcal{X}_t = g(\mathcal{X}_{t-1}, \mathcal{X}_{t-2}, \cdots; \eta_t), \tag{6.2}$$

其中，$(\eta_t)_{t\in\mathbb{Z}}$ 是一个取值于 $\mathbb{R}^{d_\eta}$ $(d_\eta \in \mathbb{N})$ 的独立同分布随机向量序列，且 $g(u; \eta)$ 是一个取值于 $\mathbb{R}^{d_x}$ 的可测函数，满足

$$\mathbb{E}[\|g(0; \eta_0)\|^r] < \infty \text{ and } \|g(u; \eta_0) - g(u'; \eta_0)\|_r \leq \sum_{k=1}^\infty \alpha_k(g)\|u_k - u'_k\| \text{ for all } u, u' \in \left(\mathbb{R}^{d_x}\right)^\infty, \tag{6.3}$$

对某个 $r \geq 1$，以及一个满足 $\sum_{k=1}^\infty \alpha_k(g) < 1$ 的非负序列 $(\alpha_k(g))_{k\geq 1}$；其中，对于任意随机向量 U, $\|U\|_r$: $= (\mathbb{E}\|U\|^r)^{1/r}$。考虑模型 (6.1)，并假设协变量满足条件 (6.2) 和 (6.3)，其中 (6.2) 中的 $(\eta_t)_{t\in\mathbb{Z}}$ 满足：$(\eta_t,\ \xi_t)_{t\in\mathbb{Z}}$ 是一个独立同分布随机向量序列，并且我们假设对某个 $r \geq 1$,

$$\sum_{k=1}^p \max\left\{\alpha_k(g), \alpha_{k,Y}(f) + \|\xi_0\|_r \alpha_{k,Y}(\mathcal{M}) + \|\xi_0\|_r^2 \alpha_{k,Y}(H)\right\} + \sum_{k=p+1}^\infty \max\{\alpha_k(g)\} < 1. \tag{6.4}$$

在假设 $A(f)$ $A(\mathcal{M})$ $A(H)$ 和 (6.4) 下，存在一个满足 $\|(Y_0, \mathcal{X}_0)\|_r < \infty$（见 ([6])）的 τ-弱依赖平稳遍历非自适应解 $(Y_t, \mathcal{X}_t)_{t\in\mathbb{Z}}$，该解满足 (6.1)。该解是 θ-弱依赖的，其系数 $\theta(j)$ 有如下界： $j \to \infty$ by,

$$\theta(j) \leq \tau(j) = \mathcal{O}\left(\inf_{1\leq\iota\leq j}\left\{\alpha^{j/\iota} + \sum_{k\geq\iota+1}\alpha_k\right\}\right), \tag{6.5}$$

其中 $\alpha_k = \max\left\{\alpha_k(g)\,\alpha_{k,Y}(f) + \|\xi_0\|_r \alpha_{k,Y}(\mathcal{M}) + \|\xi_0\|_r^2 \alpha_{k,Y}(H)\right\}$ 当 $1 \le k \le p$ 时，$\alpha_k = \alpha_k(g)$ 当 $k \ge p+1$ 且 $\alpha = \sum_{k\ge 1}\alpha_k$ 时。该解还满足假设 **(A4)** 中的 θ_∞- 弱依赖条件，见 [2]。

我们着重研究利用 SPDNN 对 (6.1) 中函数 f 进行非参数估计。令，

$$X_t = (Y_{t-1}, \cdots, Y_{t-p}; \mathcal{X}_{t-1}).$$

若使用二次损失函数，则对所有 $x \in \mathbb{R}^p \times \mathbb{R}^{d_x}$，目标预测器可表示为

$$h^*(x) = \mathbb{E}[Y_0 | X_0 = x] = f(x). \tag{6.6}$$

若采用 L_1 损失函数，则目标预测器可表示为

$$h^*(x) = \operatorname{med}(Y_0 | X_0 = x), \tag{6.7}$$

其中，med(V) 表示 V 的中位数，V 为任意随机变量。我们考虑 (6.1) 中由标准化均匀分布 $\mathcal{U}[-a$ 生成的创新项 $(\xi_t)_{t\in\mathbb{Z}}$，以及 (6.2) 中由标准化均匀分布 $\mathcal{U}[-a$ 生成的创新项 $(\eta_t)_{t\in\mathbb{Z}}$，分别对应某个 $a, a' > 0$ 和 $\mathcal{U}[-a'; a']$，并假设对所有 $x \in \mathbb{R}^p \times \mathbb{R}^{d_x}$，$\mathcal{M}(x) > 0$。在这种情况下，可以得到 med $(Y_0|X_0 = x) = f(x)$。因此，对于利用 SPDNN 对 (6.1) 中的 f 进行非参数估计，在条件 (6.2) 和 (6.4) 下，可以应用定理 5.1 和 5.2 的结果。此外，如果 f 属于某个 H¨older 空间 $\mathcal{C}^s(\mathcal{Z})$，其中 $s > 0$，则 $\mathcal{Z} = \mathcal{X} \times \mathcal{Y} \subset \mathbb{R}^d \times \mathbb{R}^{d_x}$ 可以取为有界，此时可以应用推论 5.3。

6.2 非线性 ARX-ARCH 模型示例

我们考虑如下定义的非线性 ARX-ARCH(1) 模型。

$$\begin{cases} Y_t &= f(Y_{t-1}, Y_{t-2}; \mathcal{X}_{t-1}) + \varepsilon_t \\ \varepsilon_t &= \xi_t \sqrt{\phi_0 + \phi_1 \varepsilon_{t-1}^2}, \end{cases} \tag{6.8}$$

其中，$(\xi_t)_{t\in\mathbb{Z}}$ 表示由标准化均匀分布 $\mathcal{U}[-2$生成的创新项；$2]$ 而 $(\mathcal{X}_t)_{t\in\mathbb{Z}}$是一个定义如下的 AR(1) 过程。

$$\mathcal{X}_t = \alpha_0 + \alpha_1 \mathcal{X}_{t-1} + \eta_t, \quad \text{with } \alpha_0 \in \mathbb{R},\ |\alpha_1| < 1, \tag{6.9}$$

由服从标准化均匀分布的创新项 $(\eta_t)_{t\in\mathbb{Z}}$ 生成 $\mathcal{U}[-2;\ 2]$，使得，$(\eta_t, \xi_t)_{t\in\mathbb{Z}}$ 构成一个独立同分布（i.i.d.）随机向量序列。可以将模型 (6.8) 写成 (6.1) 的形式：

$$Y_t = f(Y_{t-1}, Y_{t-2}; \mathcal{X}_{t-1}) + \xi_t \sqrt{\phi_0 + \phi_1 \left(Y_{t-1} - f(Y_{t-2}, Y_{t-3}; \mathcal{X}_{t-2})\right)^2}. \tag{6.10}$$

Set,

$$X_t = (Y_{t-1}, Y_{t-2}; \mathcal{X}_{t-1}).$$

我们对 f 施加一个利普希茨条件。假设存在 $\alpha_{1,\ Y}(f)\,\alpha_{2,\ Y}(f)\ \alpha_{\mathcal{X}}(f) > 0$ 使得，对于任意 $x = (y_1,\ y_2; u)$ $x' = (y_1',\ y_2';\ u') \in \mathbb{R}^2 \times \mathbb{R}^{d_x}$:

$$|f(x) - f(x')| \le \alpha_{1,Y}(f)|y_1 - y_1'| + \alpha_{2,Y}(f)|y_2 - y_2'| + \alpha_{\mathcal{X}}(f)|u - u'|. \tag{6.11}$$

Set,

$$\mathcal{M}(X_t) = \sqrt{\phi_0 + \phi_1 \left(Y_{t-1} - f(X_{t-1})\right)^2}.$$

对于所有 $x=(y_1,y_2;u)$ $x'=(y_1',y_2';u')\in\mathbb{R}^2\times\mathbb{R}^{d_x}$,

$$\begin{aligned}
|\mathcal{M}(x)-\mathcal{M}(x')| &= \left|\sqrt{\phi_0+\phi_1(y_1-f(x))^2}-\sqrt{\phi_0+\phi_1(y_1'-f(x'))^2}\right| \\
&= \frac{|\phi_1|\left|(y_1-f(x))^2-(y_1'-f(x'))^2\right|}{\sqrt{\phi_0+\phi_1(y_1-f(x))^2}+\sqrt{\phi_0+\phi_1(y_1'-f(x'))^2}} \\
&\leq \frac{|\phi_1|\,|(y_1-f(x))-(y_1'-f(x'))|}{\sqrt{\phi_1}} \\
&\leq \sqrt{\phi_1}\big(|y_1-y_1'|+\alpha_{1,Y}(f)|y_1-y_1'|+\alpha_{2,Y}(f)|y_2-y_2'|+\alpha_{\mathcal{X}}(f)|u-u'|\big) \\
&\leq \alpha_{1,Y}(\mathcal{M})|y_1-y_1'|+\alpha_{2,Y}(\mathcal{M})|y_2-y_2'|+\alpha_{\mathcal{X}}(\mathcal{M})|u-u'|, \qquad (6.12)
\end{aligned}$$

其中 $\alpha_{1,Y}(\mathcal{M})=\sqrt{\phi_1}(1+\alpha_{1,Y}(f))$ $\alpha_{2,Y}(\mathcal{M})=\sqrt{\phi_1}\alpha_{2,Y}(f)$ $\alpha_{\mathcal{X}}(\mathcal{M})=\sqrt{\phi_1}\alpha_{\mathcal{X}}(f)$。我们将重点研究（6.4），其中 $r=2$，并假设满足以下条件，

$$\max\left\{\alpha_1,\alpha_{1,Y}(f)+\alpha_{1,Y}(\mathcal{M})\right\}+\alpha_{2,Y}(f)+\alpha_{2,Y}(\mathcal{M})<1. \qquad (6.13)$$

因此，在条件（6.9）、（6.11）、（6.12）和（6.13）下，存在 $(Y_t,\mathcal{X}_t)_{t\in\mathbb{Z}}$ 是（6.8）的一个解，且满足上述假设（**A3**）和（**A4**）。

设（$Y_1,\mathcal{X}_1$） $\cdots$，（$Y_n,\mathcal{X}_n$）是满足（6.8）的过程 $(Y_t,\mathcal{X}_t)_{t\in\mathbb{Z}}$ 的一条轨迹。我们旨在基于该样本预测 Y_{n+1}。我们使用输入变量 $X_t=(Y_{t-1},Y_{t-2};\mathcal{X}_{t-1})\in\mathcal{X}\subset\mathbb{R}^3$ 和输出空间 $\mathcal{Y}\subset\mathbb{R}$，采用上述学习理论中的SPDNN预测器。我们考虑模型（6.8），其中 $\phi_0=0.25$, $\phi_1=0.1$ 属于以下情形的 f:

$$\begin{aligned}
&\text{DGP1}: f(Y_{t-1},Y_{t-2},\mathcal{X}_{t-1})=-0.75+0.1\max(Y_{t-1},0)-0.2\min(Y_{t-1},0)+0.15Y_{t-2}+0.4\sqrt{1+0.5\mathcal{X}_{t-1}^2}; \\
&\text{DGP2}: f(Y_{t-1},\mathcal{X}_{t-1})=0.4+(0.2-0.15e^{-Y_{t-1}^2})Y_{t-1}-\frac{0.5}{1+|\mathcal{X}_{t-1}|},
\end{aligned}$$

其中，$(\mathcal{X}_t)_{t\in\mathbb{Z}}$ 是满足 (6.9) 且具有 $\alpha_0=\alpha_1=0.5$ 的 AR(1) 过程。DGP1 是具有非线性协变量的阈值自回归模型，DGP2 是具有非线性协变量的指数自回归模型。

下文中，我们讨论 L_2 和 L_1 损失函数，它们各自对应的目标函数为:

$$h^*(x)=\mathbb{E}[Y_0|X_0=x]=f(x), \qquad (6.14)$$

and,

$$h^*(x)=\text{med}(Y_0|X_0=x)=f(x), \qquad (6.15)$$

对所有 $x\in\mathcal{X}$ 其中，该集合可选为 $\mathbb{R}^3$ 的紧子集，med 表示中位数 .

对于这些 DGP，我们均采用包含 2 个隐藏层的网络架构，每层有 100 个隐藏节点。隐藏层使用 ReLU 激活函数，输出层使用线性激活函数。我们使用 R 软件中的 keras 包，采用 Adam 算法（见 ([15])），以 10^{-3} 的学习率和 32 的小批量大小训练网络权重。当均方误差（ MSE ）连续 30 个周期未见改善时，我们停止训练。

我们从真实的数据生成过程（DGP）中生成了 $n=250,500,$ 和 1000 对应的轨迹 $((Y_1,\mathcal{X}_1),(Y_2,\mathcal{X}_2)\cdots,(Y_n,\mathcal{X}_n))$。式 (1.4) 中的预测量 $\widehat{h}_n$ 是通过调节参数推导得到的，其形式为 $\lambda_n=10^{-i}\log(n)/n,$ 和 $\tau_n=10^{-j}/\log(n)$，其中 $i,j=0,1,\cdots,10$，并通过最小化均方误差（MSE）进行校准，以及.....

验证数据集上的平均绝对误差（MAE）$((Y_1', \mathcal{X}_1'), (Y_2', \mathcal{X}_2') \cdots, (Y_n', \mathcal{X}_n'))$。然后，基于一条新的轨迹$((Y_1'', \mathcal{X}_1''), (Y_2'', \mathcal{X}_2'') \cdots, (Y_m'', \mathcal{X}_m''))$计算$\widehat{}$SPDNN预测器 h_n 的超额风险，其中$m = 10^4$。

图（1）和（2）展示了 SPDNN 和非惩罚 DNN（NPDNN，根据（1.4）在 $\lambda_n = 0$ 条件下得到）预测器在 100 次重复实验中的经验 L_1 超额风险（a）和经验 L_2 超额风险（b）的箱线图。对于 DGP1 和 DGP2，可以看到，对于 L_1 和 L_2 损失，经验超额风险总体上随着样本量的增加而降低，这与理论结果一致。此外，SPDNN估计量得出的结果总体上比 NPDNN 估计量更准确。这一发现表明，与 NPDNN 相比，SPDNN预测器可以提高预测准确性。

7 真实数据示例

我们考察了维多利亚大都市区的 PM_{10} 日浓度（直径小于 $10\mu m$ 的颗粒物）；数据采集时间为 2005 年 1 月 21 日，至 2006 年 3 月 4 日，。图 3 展示了研究期间 PM_{10} 的浓度和空气相对湿度（RH）。这些数据由州环境与水资源研究所提供，采集自八个监测站。我们重点分析其中的 408 条观测数据，这些数据来自一个大型数据集（可通过 https://rss.onlinelibrary.wiley. com/pb-assets/hub-assets/rss/Datasets/RSSC%2067.2/C1239deSouza-1531120585220.zip 获取），由 [25]收集，用于评估呼吸系统疾病与空气污染浓度之间的关联。

我们考虑非参数回归任务，旨在预测给定输入 $X_t = (PM_{10,t-1}, RH_{t-1})$ 时的输出变量 $Y_t = PM_{10,t}$。 [6]已在这些数据上拟合了一个含外生协变量的双自回归（DAR）模型，并获得了，

$$PM_{10,t} = 37.946 + 0.330PM_{10,t-1} - 0.210RH_{t-1} + \xi_t\sqrt{32.108 + 0.023PM_{10,t-1}^2}, \tag{7.1}$$

其中，$(\xi_t)_{t\in Z}$ 是一个中心化独立同分布随机变量，$(RH_t)_{t\in\mathbb{Z}}$ 是空气相对湿度协变量。我们考虑 L_2 损失函数，因此，根据 DAR 模型 (7.1)，目标预测器为，

$$PM_{10,t}^{DAR} = \mathbb{E}\left[PM_{10,t}|(PM_{10,t-1}, RH_{t-1})\right] = 37.946 + 0.330PM_{10,t-1} - 0.210RH_{t-1}. \tag{7.2}$$

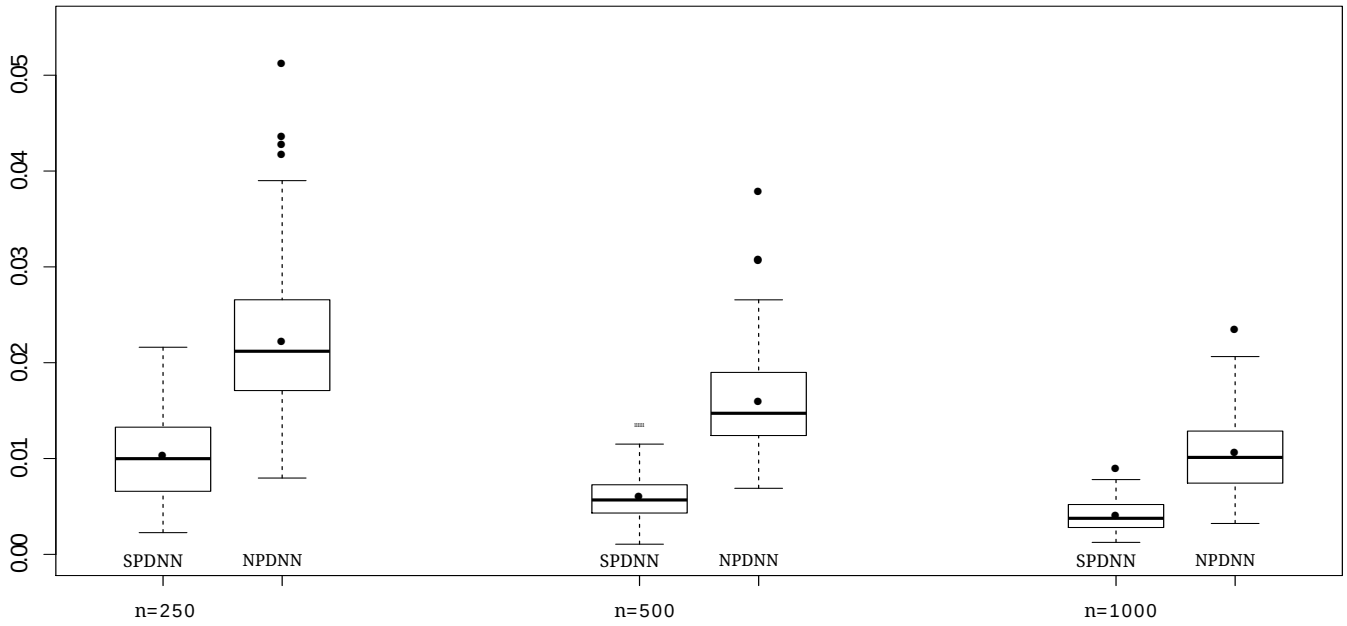
对于估计任务，我们采用与上一节相同的网络架构（2 个隐藏层，每层 100 个隐藏节点，并使用 ReLU 激活函数），输入变量为 $X_t = (PM_{10,\ t-1},\ RH_{t-1})$。

我们考虑基于 SPDNN、NPDNN估计量和 DAR 预测器（7.2）的预测。对于 SPDNN 和 NPDNN 预测器，使用前 308 个观测值训练网络，其余观测值用作测试数据。对于 SPDNN、NPDNN 和 DAR 这几种预测器，平均绝对预测误差和平均相对预测误差分别定义为，

$$\text{mean abs.pred.error} = \frac{1}{100}\sum_{t=1}^{100}\left|PM_{10,t} - \widehat{PM}_{10,t}\right|$$

$$\text{mean relat.pred.error} = \frac{1}{100}\sum_{t=1}^{100}\frac{\left|PM_{10,t} - \widehat{PM}_{10,t}\right|}{PM_{10,t}},$$

(a) DGP1 中经验 L_1 超额风险的箱线图



(b) DGP1 中经验 L_2 超额风险的箱线图

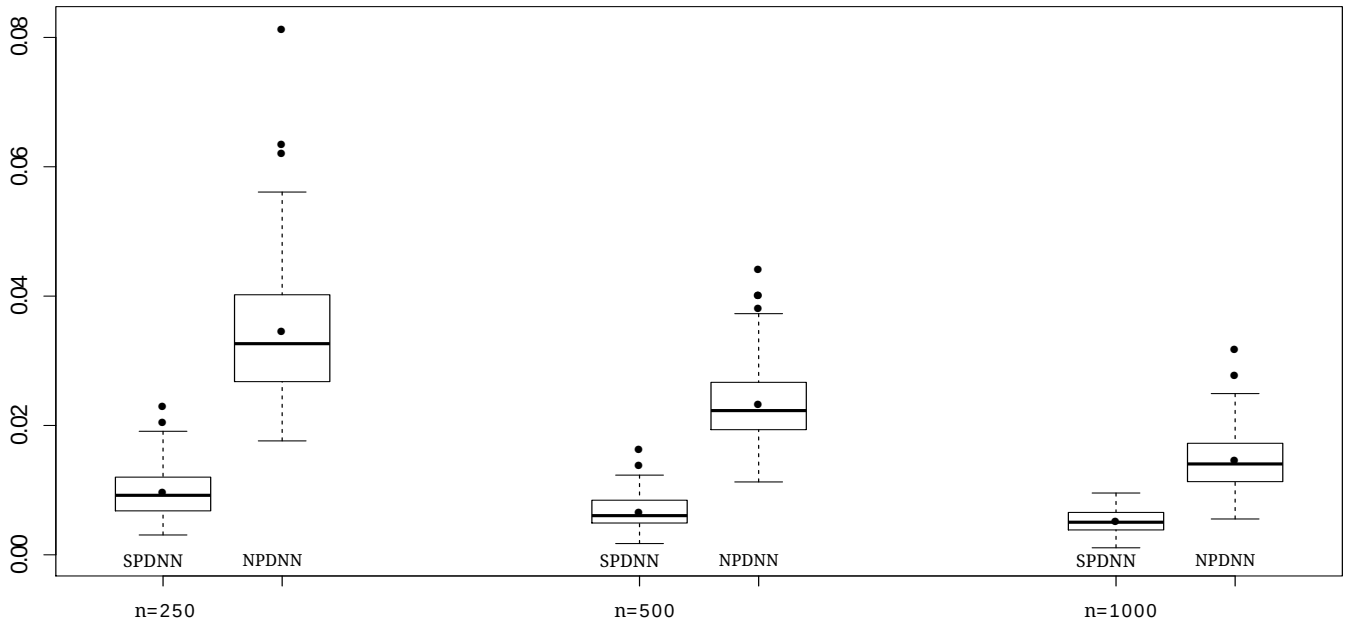
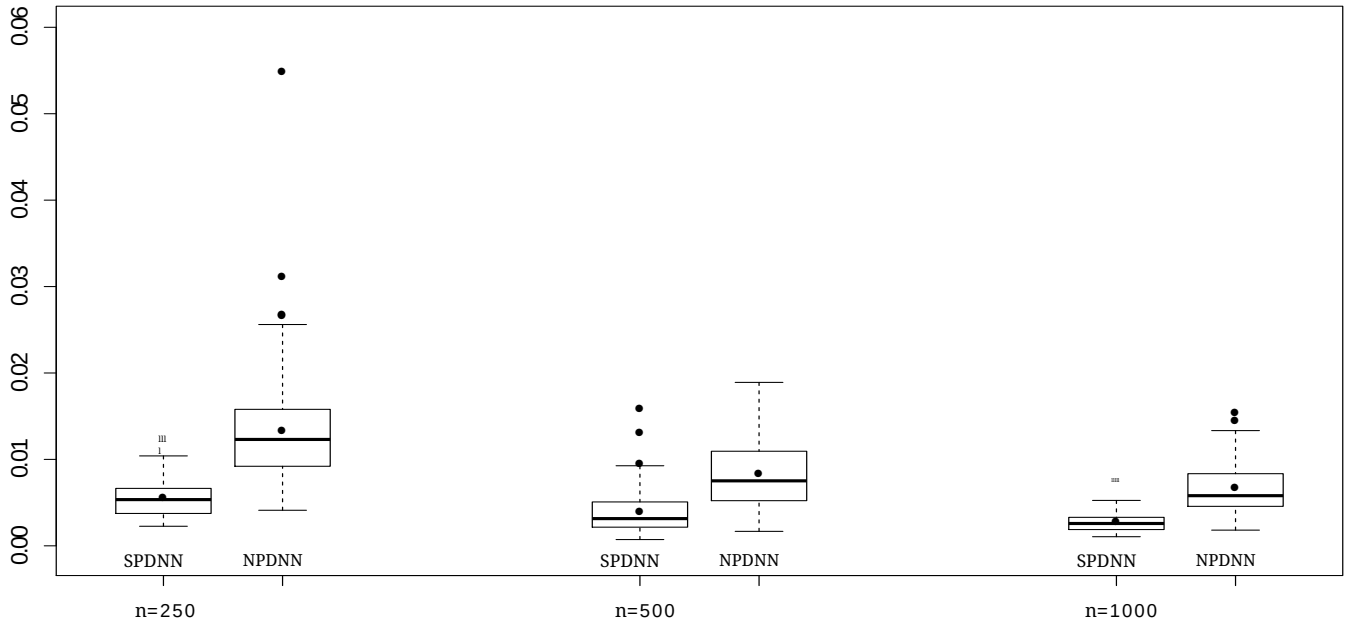


图 1: *SPDNN* 和 *NPDNN* 预测器在 *DGP1* 中的经验 L_1 超额风险（*a*）和经验 L_2 超额风险（*b*）的箱线图，其中 $n = 250$、500 和 *1000*。

(a) DGP2 中经验 L_1 超额风险的箱线图



(b) DGP2 中经验 L_2 超额风险的箱线图

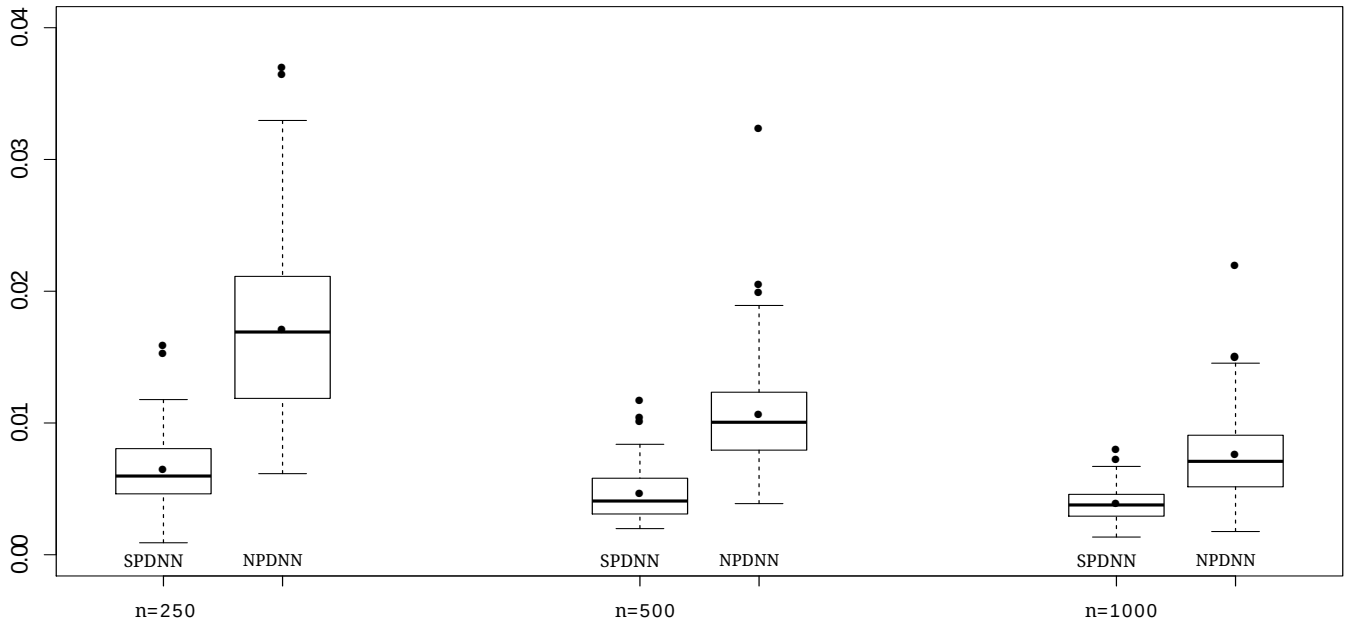


图 2：*DGP2* 中 *SPDNN* 和 *NPDNN* 预测器在 $n = 250$、500 和 *1000* 时的经验 L_1 超额风险（*a*）和经验 L_2 超额风险（*b*）的箱线图。

预测变量	平均相对预测误差（%）	平均绝对预测误差
SPDNN	5.75	18.50
NPDNN	6.03	19.55
DAR	6.01	19.79

表 1：SPDNN、NPDNN 和 DAR 预测器的部分性能

$\widehat{}$是在此处计算的，其中 $PM_{10,\ t}$ 表示时间 t 的预测值。图4 (a) 和 (b) 展示了绝对预测误差和相对预测误差的箱线图。表1列出了预测器 SPDNN、NPDNN 和 DAR 的平均绝对预测误差和平均相对预测误差。从表1可以看出，SPDNN预测器的平均相对预测误差和平均绝对预测误差均略优于 NPDNN 和 DAR 预测器。尽管如此，基于 SPDNN 估计量得到的预测属于样本外预测，而基于 DAR 模型得到的预测则属于样本内预测。

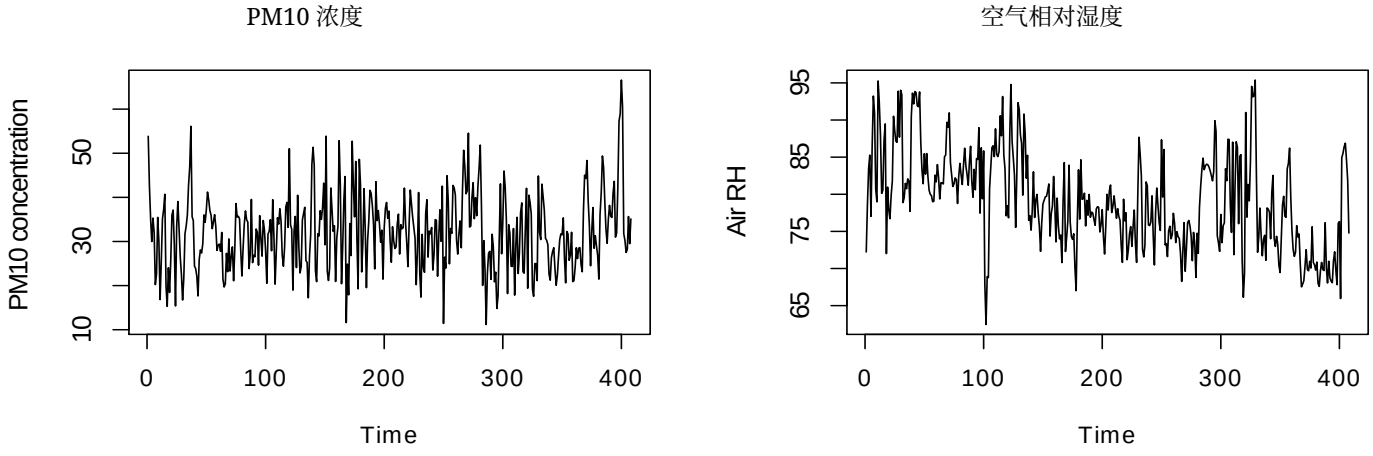
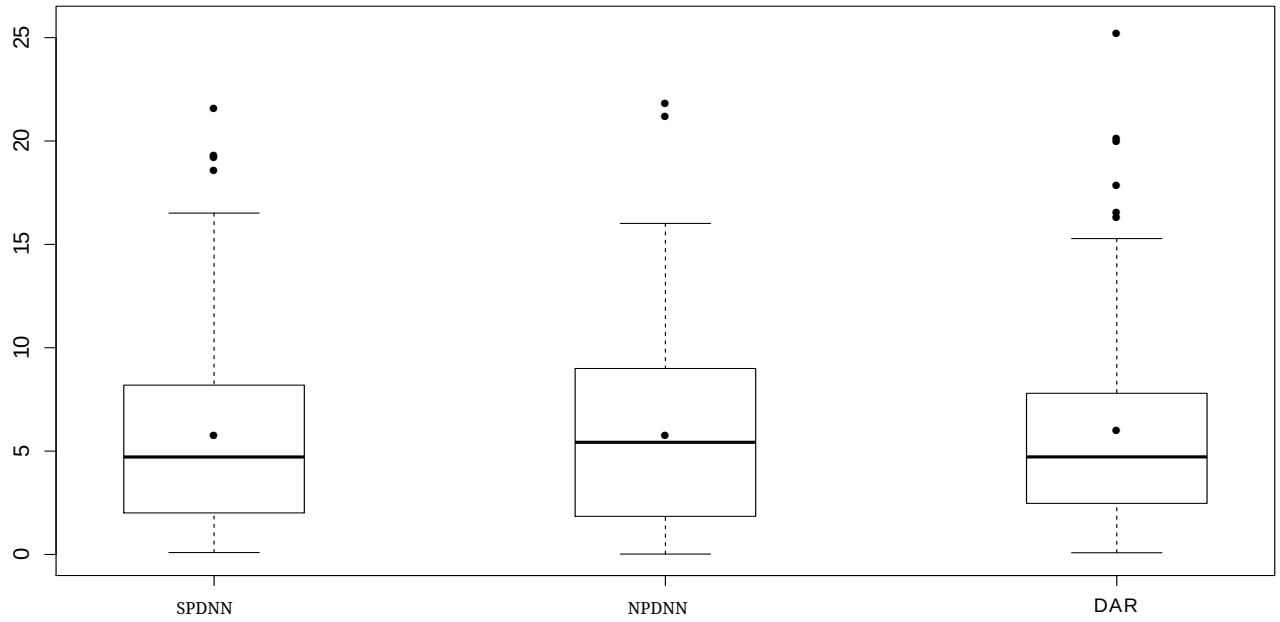


图 3：*2005* 年 *1* 月 *21* 日至 *2006* 年 *3* 月 *4* 日期间，*Vit´oria* 大都市区的 *PM10* 日浓度和空气相对湿度。

(a) 绝对预测误差的箱线图



(b) 相对预测误差的箱线图

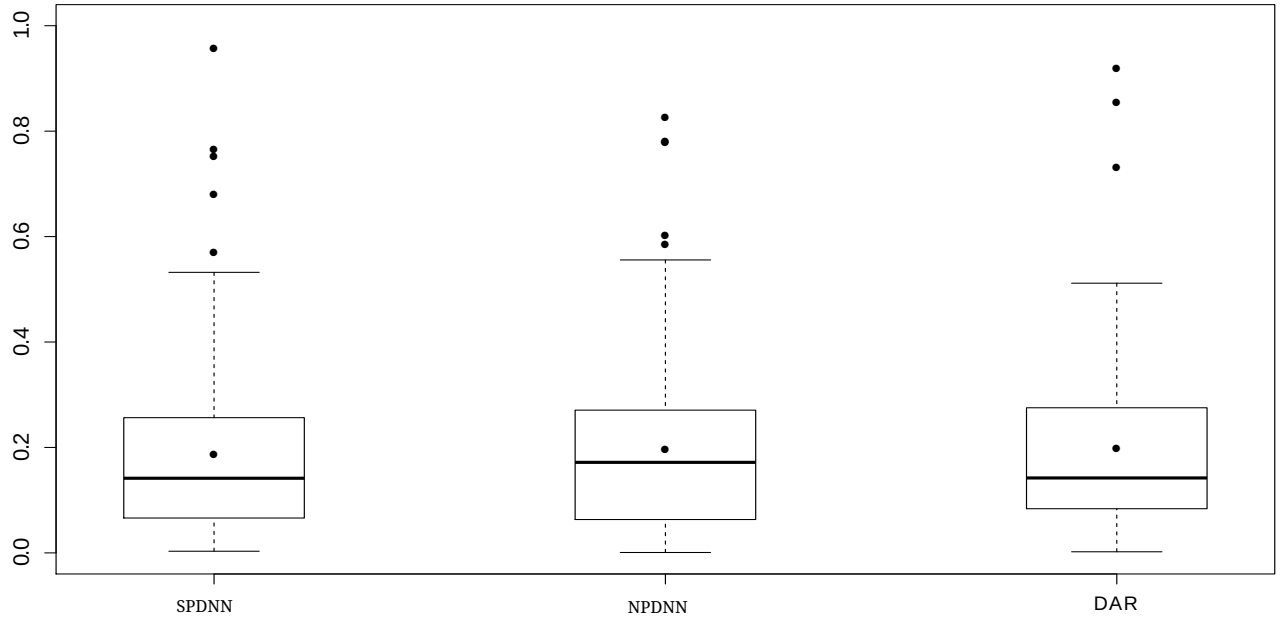


图4：*SPDNN*、*NPDNN* 和*DAR*预测器的绝对预测误差（a）和相对预测误差（b）的箱线图。

8 主要结果的证明

8.1 定理 4.1 的证明

对于所有 $h \in \mathcal{H}_\sigma(L,\ N,\ B,\ F,\ S)$ $\varepsilon > 0$, $\nu_0 \in (0,\ 1]$, $\lambda = \dfrac{1}{n^{1-\nu_0}}$，以及任意 $n \in \mathbb{N}$，根据 [1] 和马尔可夫不等式，我们有

$$\begin{aligned} P\Big\{R(h) - \widehat{R}_n(h) > \varepsilon\Big\} &= P\Big\{\Big(\mathbb{E}[\ell(h(X_0), Y_0)] - \sum_{i=1}^{n} \ell(h(X_i), Y_i)\Big) > n\varepsilon\Big\} \\ &\leq P\Big\{\lambda\Big|\Big(\mathbb{E}[\ell(h(X_0), Y_0)] - \sum_{i=1}^{n} \ell(h(X_i), Y_i)\Big)\Big| > n\lambda\varepsilon\Big\} \\ &\leq P\Big\{e^{\lambda|(\mathbb{E}[\ell(h(X_0),Y_0)] - \sum_{i=1}^{n} \ell(h(X_i),Y_i))|} > e^{n\lambda\varepsilon}\Big\} \\ &\leq \exp\Big(-n^{\nu_0}\varepsilon + \frac{n^{2\nu_0-1}(M + \theta_{\infty,n}(1))^2}{2}\Big). \end{aligned} \tag{8.1}$$

令 $l = \mathcal{N}\Big(\mathcal{H}_\sigma(L, N, B, F, S),\ \dfrac{\varepsilon}{4G}\Big)$。由（8.1）可得，且在 [7] 中，对 $n \in \mathbb{N}$,

$$\begin{aligned} P\Big\{\sup_{h\in\mathcal{H}_\sigma(L,N,B,F,S)}\Big[R(h) - \widehat{R}_n(h)\Big] > \varepsilon\Big\} &\leq \sum_{j=1}^{l} \exp\Big(-n^{\nu_0}\varepsilon + \frac{n^{2\nu_0-1}(M + \theta_{\infty,n}(1))^2}{2}\Big) \\ &\leq \mathcal{N}\Big(\mathcal{H}_\sigma(L, N, B, F, S), \frac{\varepsilon}{4G}\Big) \exp\Big(-n^{\nu_0}\varepsilon + \frac{n^{2\nu_0-1}(M + \theta_{\infty,n}(1))^2}{2}\Big). \end{aligned} \tag{8.2}$$

至此，定理 4.1 得证。 ■

8.2 定理 4.2 的证明

(i) 对于 $\varepsilon \in (0, 2M]$，利用 (8.2) 以及 [19] 中的下列不等式

$$\mathcal{N}\Big(\mathcal{H}_\sigma\left(L, N, B, F, S\right), \frac{\varepsilon}{4G}\Big) \leq \exp\left(2L(S+1)\log\Big(\frac{4G}{\varepsilon}C_\sigma L(N+1)(B \vee 1)\Big)\right). \tag{8.3}$$

我们有

$$\begin{aligned} &P\Big\{\sup_{h\in\mathcal{H}_\sigma(L,N,B,F,S)}\Big[R(h) - \widehat{R}_n(h)\Big] > \varepsilon\Big\} \\ &\leq \exp\left(2L(S+1)\log\Big(\frac{4G}{\varepsilon}C_\sigma L(N+1)(B \vee 1)\Big)\right) \exp\Big(-n^{\nu_0}\varepsilon + \frac{n^{2\nu_0-1}(M + \theta_{\infty,n}(1))^2}{2}\Big) \\ &\leq \exp\left(2L(S+1)\log\Big(\frac{4G}{\varepsilon}C_\sigma L(N+1)(B \vee 1)\Big) - n^{\nu_0}\varepsilon + \frac{n^{2\nu_0-1}(M + \theta_{\infty,n}(1))^2}{2}\right) = \eta. \end{aligned} \tag{8.4}$$

令 $0 < \eta < 1$。我们考虑关于 ε 的以下方程:

$$\exp\left(2L(S+1)\log\Big(\frac{4G}{\varepsilon}C_\sigma L(N+1)(B \vee 1)\Big) - n^{\nu_0}\varepsilon + \frac{n^{2\nu_0-1}(M + \theta_{\infty,n}(1))^2}{2}\right) = \eta,$$

i.e.

$$2L(S+1)\log\Big(\frac{4G}{\varepsilon}C_\sigma L(N+1)(B \vee 1)\Big) - n^{\nu_0}\varepsilon + \frac{n^{2\nu_0-1}(M + \theta_{\infty,n}(1))^2}{2} = \log(\eta). \tag{8.5}$$

i.e.

$$2L(S+1)\log\left(4GC_\sigma L(N+1)(B\vee 1)\right)-2L(S+1)\log(\varepsilon)-n^{\nu_0}\varepsilon+\frac{n^{2\nu_0-1}(M+\theta_{\infty,n}(1))^2}{2}-\log(\eta)=0.$$

Set

$$C_{0,n}=n^{\nu_0},\ C_1=2L(S+1)\log\left(4GC_\sigma L(N+1)(B\vee 1)\right).$$

我们有

$$2L(S+1)\log(\varepsilon)+C_{0,n}\varepsilon-\frac{n^{2\nu_0-1}(M+\theta_{\infty,n}(1))^2}{2}+\log(\eta)-C_1=0. \tag{8.6}$$

考虑函数

$$\varphi(\varepsilon)=2L(S+1)\log(\varepsilon)+C_{0,n}\varepsilon-\frac{n^{2\nu_0-1}(M+\theta_{\infty,n}(1))^2}{2}+\log(\eta)-C_1 \text{ for } \varepsilon\in(0,2M].$$

我们有， $\varphi(\varepsilon)\to-\infty$ ，其中 $\varepsilon\to 0$。此外，

$$\begin{aligned}
\varphi(2M)>0 &\Rightarrow 2L(S+1)\log(2M)+2C_{0,n}M-\frac{n^{2\nu_0-1}(M+\theta_{\infty,n}(1))^2}{2}+\log(\eta)-C_1>0\\
&\Rightarrow 2C_{0,n}M-\frac{n^{2\nu_0-1}(M+\theta_{\infty,n}(1))^2}{2}>C_1-2L(S+1)\log(2M)-\log(\eta)\\
&\Rightarrow C_{0,n}-\frac{n^{2\nu_0-1}(M+\theta_{\infty,n}(1))^2}{4M}>\frac{1}{2M}\Big(C_1-2L(S+1)\log(2M)-\log(\eta)\Big)\\
&\Rightarrow n^{2\nu_0}-\frac{n^{2\nu_0-1}(M+\theta_{\infty,n}(1))^2}{4M}>\frac{1}{2M}\Big(C_1-2L(S+1)\log(2M)-\log(\eta)\Big)\\
&\Rightarrow n^{2\nu_0}\left(1-\frac{(M+\theta_{\infty,n}(1))^2}{4Mn}\right)>\frac{1}{2M}\Big(C_1-2L(S+1)\log(2M)-\log(\eta)\Big).
\end{aligned} \tag{8.7}$$

由于 $\left(\frac{M+\theta_{\infty,\ \ n}\ 1}{4Mn}\right)^{\ 2}\longrightarrow 0$( ) 作为 $n\longrightarrow\infty$，存在 $n_0(\nu_0,\theta_{\infty,n},M)$，使得，

$$n>n_0(\nu_0,\theta_{\infty,n},M)\Rightarrow\frac{(M+\theta_{\infty,n}(1))^2}{4Mn}<\frac{1}{2}. \tag{8.8}$$

因此，为了得到 (8.7)，对于 $n>n_0(\nu_0,\ \theta_{\infty,\ n},\ M)$，只需满足，

$$\frac{n^{2\nu_0}}{2}>\frac{1}{2M}\Big(C_1-2L(S+1)\log(2M)-\log(\eta)\Big). \tag{8.9}$$

回顾一下，对于所有 $x\in\mathbb{R}$，$(x)_+=\max(x,0)$。也就是说，

$$n>\left(\frac{1}{M}\Big(C_1-2L(S+1)\log(2M)-\log(\eta)\Big)_+\right)^{1/(2\nu_0)}. \tag{8.10}$$

因此，函数 $\varepsilon\mapsto\varphi(\varepsilon)$严格递增，且在条件 (8.10) 下，当 $\varepsilon\to 0$时，我们有 $\varphi(\varepsilon)\to-\infty$以及 $\varphi(2M)>0$。因此，存在唯一的 $\varepsilon(n,\eta)\in(0,2M)$，使得 $\varphi(\varepsilon(n,\eta))=0$。令 $\varepsilon_0=\frac{2M}{n^{\nu_0/2}}$, $\nu_0<1$，则有

$$\begin{aligned}
\varphi(\varepsilon_0)>0 &\Rightarrow 2L(S+1)\log(2M)-L(S+1)\nu_0\log n+\frac{2MC_{0,n}}{n^{\nu_0/2}}-\frac{n^{2\nu_0-1}(M+\theta_{\infty,n}(1))^2}{2}+\log(\eta)-C_1>0\\
&\Rightarrow 2Mn^{\nu_0/2}-L(S+1)\nu_0\log n-\frac{n^{2\nu_0-1}(M+\theta_{\infty,n}(1))^2}{2}>C_1-2L(S+1)\log(2M)-\log(\eta)\\
&\Rightarrow 2Mn^{2\nu_0}\left(1-L(S+1)\nu_0\frac{\log n}{2Mn^{2\nu_0}}-\frac{(M+\theta_{\infty,n}(1))^2}{4Mn}\right)>C_1-2L(S+1)\log(2M)-\log(\eta).
\end{aligned} \tag{8.11}$$

由于，当 $L(S+1)\nu_0 \underbrace{\log 2Mn^{\nu_0/2}}_{n} + \underbrace{(M+\theta_{\infty,n}1^2}_{4Mn} \longrightarrow 0$ ( ) 趋于 $n\to\infty$ 时，存在 $n_0(L,S,M,\nu_0,\theta_{\infty,n})$，使得，

$$n > n_0\left(L,S,M,\nu_0,\theta_{\infty,n}\right) \Rightarrow L(S+1)\nu_0\frac{\log n}{2Mn^{\nu_0/2}} + \frac{(M+\theta_{\infty,n}(1))^2}{4Mn} < \frac{1}{2}. \tag{8.12}$$

因此，要得到 (8.11)，只需满足 $n > n_0(L,\ S,\ M,\ \nu_0,\ \theta_{\infty,\ n})$,

$$Mn^{2\nu_0} > C_1 - 2L(S+1)\log(2M) - \log(\eta).$$

i.e.

$$n > \Big(\frac{1}{M}\left(C_1 - 2L(S+1)\log(2M) - \log(\eta)\right)_+\Big)^{1/2\nu_0}. \tag{8.13}$$

因此，在条件 (8.13) 下，$\varphi(\varepsilon(n,\ \eta)) = 0$ 的唯一解 $\varepsilon_1(n,\ \eta,\ \nu_0)$ 满足 $\varepsilon_1(n,\ \eta,\ \nu_0) < \frac{2M}{n^{\nu_0/2}} < 2M$ for $\overline{\nu_0 < 1}$)。因此，由 (8.4) 可知，以至少 $1-\eta$ 的概率，

$$\sup_{h\in\mathcal{H}_\sigma(L,N,B,F,S)}\Big[R(h) - \widehat{R}_n(h)\Big] \le \varepsilon_1\left(n,\eta,\nu_0\right);$$

这意味着，

$$R(\widehat{h}_{n,NP}) - \widehat{R}_n(\widehat{h}_{n,NP}) \le \varepsilon_1\left(n,\eta,\nu_0\right). \tag{8.14}$$

定理 4.2 的初步证明成立。

(ii) 由 (8.1) 可得，

$$P\Big\{\widehat{R}_n(h_{\mathcal{H}_\sigma}) - R(h_{\mathcal{H}_\sigma}) > \varepsilon\Big\} \le \exp\Big(-n^{\nu_0}\varepsilon + \frac{n^{2\nu_0-1}(M+\theta_{\infty,n}(1))^2}{2}\Big) = \eta. \tag{8.15}$$

令 $0<\eta<1$。考虑关于 ε 的方程，

$$\exp\Big(-n^{\nu_0}\varepsilon + \frac{n^{2\nu_0-1}(M+\theta_{\infty,n}(1))^2}{2}\Big) = \eta.$$

i.e.

$$-n^{\nu_0}\varepsilon + \frac{n^{2\nu_0-1}(M+\theta_{\infty,n}(1))^2}{2} - \log(\eta) = 0. \tag{8.16}$$

该方程的一个解为，

$$\varepsilon_1'(n,\eta,\nu_0) = \frac{(M+\theta_{\infty,n}(1))^2}{2n^{1-\nu_0}} - \frac{\log(\eta)}{n^{\nu_0}}.$$

因此，根据（8.15），我们以至少 $1-\eta$ 的概率有：

$$\widehat{R}_n(h_{\mathcal{H}_\sigma}) - R(h_{\mathcal{H}_\sigma}) \le \varepsilon_1'(n,\eta,\nu_0). \tag{8.17}$$

$\widehat{\ }$ $\widehat{\ }$回顾可知 $h_n =$极小化 $R_n(h)$ 。由于 $h_{\mathcal{H}_\sigma}\in\mathcal{H}_\sigma(L,\ N,\ B,$ $F,\ S)$，我们有$_{h\in\mathcal{H}_\sigma(L,\ N,\ B,\ F,\ S)}$

$$\widehat{R}_n(\widehat{h}_{n,NP}) \le \widehat{R}_n(h_{\mathcal{H}_\sigma}). \tag{8.18}$$

由此可得，

$$\widehat{R}_n(h_{\mathcal{H}_\sigma}) - R(h_{\mathcal{H}_\sigma}) \le \varepsilon_1'(n,\eta,\nu_0). \tag{8.19}$$

结合 (8.14) 和 (8.19)，以至少 $1-2\eta$ 的概率，我们得到

$$R(\widehat{h}_{n,NP}) - R(h_{\mathcal{H}_\sigma}) \le \varepsilon_1(n,\eta,\nu_0) + \varepsilon_1'(n,\eta,\nu_0). \tag{8.20}$$

由此完成定理 4.2 的证明。 ■

8.3 定理 5.1 的证明

我们考虑以下分解:

$$R(\widehat{h}_n) - R(h^*) \coloneqq B_{1,n} + B_{2,n}.$$

其中,

$$\begin{aligned} B_{1,n} &= [R(\widehat{h}_n) - R(h^*)] - 2[\widehat{R}_n(\widehat{h}_n) - \widehat{R}_n(h^*)] - 2J_{\lambda_n,\tau_n}(\widehat{h}_n); \\ B_{2,n} &= 2[\widehat{R}_n(\widehat{h}_n) - \widehat{R}_n(h^*)] + 2J_{\lambda_n,\tau_n}(\widehat{h}_n). \end{aligned}$$

对于界$B_{1,\ n}$，采用与 [13] 中类似的方法。令

$$\Delta(h)(Z_0) \coloneqq \ell(h(X_0), Y_0) - \ell(h^*(X_0), Y_0) \ \text{ with } Z_0 \coloneqq (X_0, Y_0).$$

定义

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

因此，对于 $\rho > 0$，我们有

$$\begin{aligned} P(B_{1,n} > \rho) &= P\Big(\mathbb{E}[\ell(\widehat{h}_n(X_0), Y_0)] - \mathbb{E}[\ell(h^*(X_0), Y_0)] - \frac{2}{n}\sum_{i=1}^n \left[\ell(\widehat{h}_n(X_i), Y_i) - \ell((h^*(X_i), Y_i)\right] - 2J_{\lambda_n,\tau_n}(\widehat{h}_n) > \rho\Big) \\ &\le P\Big(\mathbb{E}[\ell(h(X_0), Y_0)] - \mathbb{E}[\ell(h^*(X_0), Y_0)] - \frac{1}{n}\sum_{i=1}^n [\ell(h(X_i), Y_i) - \ell(h^*(X_i), Y_i)] > \frac{1}{2}\Big(\mathbb{E}[\ell(h(X_0), Y_0)] \\ &\qquad - \mathbb{E}[\ell(h^*(X_0), Y_0)] + 2J_{\lambda_n,\tau_n}(h) + \rho\Big)\Big) \\ &\le P\Big(\exists h \in \mathcal{H}_\sigma(L_n, N_n, B_n, F) : \mathbb{E}[\Delta(h)(Z_0)] - \frac{1}{n}\sum_{i=1}^n \mathbb{E}[\Delta(h)(Z_i)] > \frac{1}{2}\Big(\rho + 2J_{\lambda_n,\tau_n}(h) + \mathbb{E}[\Delta(h)(Z_0)]\Big)\Big) \\ &\le \sum_{j=0}^\infty P\left(\sup_{h\in\mathcal{H}_{n,j,\rho}} \frac{\mathbb{E}[\Delta(h)(Z_0)] - \frac{1}{n}\sum_{i=1}^n \mathbb{E}[\Delta(h)(Z_i)]}{\rho + 2J_{\lambda_n,\tau_n}(h) + \mathbb{E}[\Delta(h)(Z_0)]} > \frac{1}{2}\right). \end{aligned} \tag{8.21}$$

由于,

$$\mathbb{E}[\Delta(h)(Z_0)] \ge 0, \ \rho > 0, \ \text{and for } j \neq 0, 2J_{\lambda_n,\tau_n} \ge 2^j\rho.$$

因此我们有

$$P(B_{1,n} > \rho) \le \sum_{j=1}^\infty P\Big(\sup_{h\in\mathcal{H}_{n,j,\rho}} \mathbb{E}[\Delta(h)(Z_0)] - \frac{1}{n}\sum_{i=1}^n \mathbb{E}[\Delta(h)(Z_i)] > \frac{2^j\rho}{2}\Big).$$

令其定义为

$$\mathcal{G}_{n,j,\rho} \coloneqq \left\{ \Delta(h) : \mathbb{R}^d \times \mathcal{Y} \to \mathbb{R} : h \in \mathcal{H}_{n,j,\rho} \right\}. \tag{8.22}$$

由于，$h \in \mathcal{H}_\sigma(L_n,\ N_n,\ B_n,\ F)$ $h^* \in \mathcal{C}^{s,\ \mathcal{K}}$ 均满足利普希茨连续性，其利普希茨系数分别记为

$\mathcal{K}_{\mathcal{H}_\sigma(L_n,\ N_n,\ B_n,\ F)}$ $\mathcal{K}_{\mathcal{C}^{s,\ \kappa}} > 0$。令 $g(x,\ y) = \Delta(h)(x,\ y)$，可以看出，对于 $(x,\ y)$，$(x',\ y') \in \mathbb{R}^d \times \mathbb{R}$，

$$\begin{aligned}
|g(x,y) - g(x',y')| &= |\Delta(h)(x,y) - \Delta(h)(x',y')| = |[\ell(h(x),y) - \ell(h(x'),y')] - [\ell(h^*(x),y) - \ell(h^*(x'),y')]| \\
&\leq |\ell(h(x),y) - \ell(h(x'),y')| + |\ell(h^*(x),y) - \ell(h^*(x'),y')| \\
&\leq \mathcal{K}_\ell \Big(|(h(x),y) - (h(x'),y')| + |(h^*(x),y) - (h^*(x'),y')| \Big) \\
&\leq \mathcal{K}_\ell \Big(|h(x) - h(x')| + |h^*(x) - h^*(x')| + 2|y - y'| \Big) \\
&\leq \mathcal{K}_\ell \Big(\mathcal{K}_{\mathcal{H}_\sigma(L_n,N_n,B_n,F)} |x - x'| + \mathcal{K}_{\mathcal{C}^{s,\delta}} |x - x'| + 2|y - y'| \Big) \\
&\leq 2\mathcal{K}_\ell \max(\mathcal{K}_{\mathcal{H}_\sigma(L_n,N_n,B_n,F)} + \mathcal{K}_{\mathcal{C}^{s,\kappa}}, 1) \Big(\|x - x'\| + |y - y'| \Big). \qquad (8.23)
\end{aligned}$$

因此，函数 g 是 $2\mathcal{K}_\ell \max(\mathcal{K}_{\mathcal{H}_\sigma(L_n,N_n,B_n,F)} + \mathcal{K}_{\mathcal{C}^{s,\kappa}}, 1)$-Lipschitz 的。因此，过程 $\{g(Z_t) := \ell(h(X_t)\ Y_t)\ t \in \mathbb{Z}\}$ 是 ψ-弱依赖的。因此，我们有

$$\begin{aligned}
P(B_{1,n} > \rho) &\leq \sum_{j=1}^{\infty} P\left(\sup_{h \in \mathcal{H}_{n,j,\rho}} \mathbb{E}[g(Z_0)] - \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[g(Z_i)] > \frac{2^j \rho}{2} \right) \\
&\leq \sum_{j=1}^{\infty} P\left(\sup_{g \in \mathcal{G}_{n,j,\rho}} \mathbb{E}[g(Z_0)] - \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[g(Z_i)] > \frac{2^j \rho}{2} \right). \qquad (8.24)
\end{aligned}$$

可以看出

$$\begin{aligned}
P\Big\{ \mathbb{E}[g(Z_0)] - \frac{1}{n} \sum_{i=1}^{n} g(Z_i) > \varepsilon \Big\} &= P\Big\{ \Big(\mathbb{E}[g(X_0,Y_0)] - \sum_{i=1}^{n} g(X_i,Y_i) \Big) \geq n\varepsilon \Big\} \\
&\leq P\Big\{ \Big| \Big(\mathbb{E}[g(X_0,Y_0)] - \sum_{i=1}^{n} g(X_i,Y_i) \Big) \Big| \geq n\varepsilon \Big\}. \qquad (8.25)
\end{aligned}$$

令 $l = \mathcal{N}(\varepsilon, \mathcal{G}_{n,j,\rho}, \|\cdot\|_\infty)$。回顾条件

$$A_n \geq \mathbb{E}\Big[\Big(\sum_{i=1}^{n} \Big(\ell(h(X_i),Y_i) - \mathbb{E}[\ell(h(X_0),Y_0)] \Big) \Big)^2 \Big], \quad B_n = 2ML_2 \max\left(\frac{2^{4+\mu} n M^2 L_1}{A_n}, 1 \right).$$

可以在 [9] 和 [12]中看到

$$P\Big\{ \sup_{g \in \mathcal{G}_{n,j,\rho}} \Big[\mathbb{E}[g(Z_0)] - \frac{1}{n} \sum_{i=1}^{n} g(Z_i) \Big] > \varepsilon \Big\} \leq \mathcal{N}(\varepsilon, \mathcal{G}_{n,j,\rho}, \|\cdot\|_\infty) \exp\left(- \frac{n^2\varepsilon^2/8}{A_n + B_n^{1/(\mu+2)} (n\varepsilon/2)^{(2\mu+3)/(\mu+2)}} \right). \qquad (8.26)$$

在 [20], 中，我们有

$$\mathcal{N}(\varepsilon, \mathcal{G}_{n,j,\rho}, \|\cdot\|_\infty) \leq \mathcal{N}\left(\frac{\varepsilon}{\mathcal{K}_\ell}, \mathcal{H}_{n,j,\rho}, \|\cdot\|_\infty \right). \qquad (8.27)$$

不难看出,

$$\mathcal{H}_{n,j,\rho} \subset \left\{ h \in \mathcal{H}_\sigma(L_n, N_n, B_n, F, \frac{2^j \rho}{\lambda_n}) : \|\theta(h)\|_{\text{clip},\tau_n} \leq \frac{2^j \rho}{\lambda_n} \right\}. \qquad (8.28)$$

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

$$\begin{aligned}
\|\Delta(h_1)(x,y) - \Delta(h_2)(x,y)\|_\infty &= |\ell(h_1(x),y) - \ell(h_2(x),y)| \\
&\leq \mathcal{K}_\ell |h_1(x) - h_2(x)|. \qquad (8.29)
\end{aligned}$$

从 [20] 可以看出，以下不等式成立

$$\begin{aligned}\mathcal{N}\left(\varepsilon, \mathcal{G}_{n,j,\rho}, \|\cdot\|_\infty\right) &\leq \mathcal{N}\Big(\frac{\varepsilon}{\mathcal{K}_\ell}, \mathcal{H}_{n,j,\rho}, \|\cdot\|_\infty\Big) \leq \mathcal{N}\Big(\frac{\varepsilon}{\mathcal{K}_\ell}, \mathcal{H}_\sigma(L_n, N_n, B_n, F, \frac{2^j\rho}{\lambda_n}), \|\cdot\|_\infty\Big) \\ &\leq \exp\left(2\frac{2^j\rho}{\lambda_n}(L_n+1)\log\left(\frac{(L_n+1)(N_n+1)B_n}{\frac{\varepsilon}{\mathcal{K}_\ell} - \tau_n(L_n+1)((N_n+1)B_n)^{L_n+1}}\right)\right). \qquad (8.30)\end{aligned}$$

因此，

$$\begin{aligned}&P\Big\{\sup_{g\in\mathcal{G}_{n,j,\rho}}\Big[\mathbb{E}[g(Z_0)] - \frac{1}{n}\sum_{i=1}^n g(Z_i)\Big] > \varepsilon\Big\} \\ &\leq \exp\left(2\frac{2^j\rho}{\lambda_n}(L_n+1)\log\left(\frac{(L_n+1)(N_n+1)B_n}{\frac{\varepsilon}{\mathcal{K}_\ell} - \tau_n(L_n+1)((N_n+1)B_n)^{L_n+1}}\right)\right)\exp\left(-\frac{n^2\varepsilon^2/8}{A_n + B_n^{1/(\mu+2)}(n\varepsilon/2)^{(2\mu+3)/(\mu+2)}}\right) \\ &\leq \exp\left(2\frac{2^j\rho}{\lambda_n}(L_n+1)\log\left(\frac{(L_n+1)(N_n+1)B_n}{\frac{\varepsilon}{\mathcal{K}_\ell} - \tau_n(L_n+1)((N_n+1)B_n)^{L_n+1}}\right) - \frac{n^2\varepsilon^2/8}{A_n + B_n^{1/(\mu+2)}(n\varepsilon/2)^{(2\mu+3)/(\mu+2)}}\right). \qquad (8.31)\end{aligned}$$

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

$$\begin{aligned}&P\Big\{\sup_{g\in\mathcal{G}_{n,j,\rho}}\Big[\mathbb{E}[g(Z_0)] - \frac{1}{n}\sum_{i=1}^n g(Z_i)\Big] > \frac{2^j\rho}{2}\Big\} \\ &\leq \exp\left(2\frac{2^j\rho}{\lambda_n}(L_n+1)\log\left(\frac{(L_n+1)(N_n+1)B_n}{\frac{2^j\rho}{2\mathcal{K}_\ell} - \tau_n(L_n+1)((N_n+1)B_n)^{L_n+1}}\right) - \frac{n^2(2^j\rho)^2/32}{A_n + B_n^{1/(\mu+2)}(n2^j\rho/4)^{(2\mu+3)/(\mu+2)}}\right). \qquad (8.32)\end{aligned}$$

取 A_n，B_n，$C_{1,\,n}$，$C_{2,\,n}$ 如 [12] 中所示，我们有

$$\begin{aligned}&P\Big\{\sup_{g\in\mathcal{G}_{n,j,\rho}}\Big[\mathbb{E}[g(Z_0)] - \frac{1}{n}\sum_{i=1}^n g(Z_i)\Big] > \frac{2^j\rho}{2}\Big\} \\ &\leq \exp\left(2\frac{2^j\rho}{\lambda_n}(L_n+1)\log\left(\frac{(L_n+1)(N_n+1)B_n}{\frac{2^j\rho}{2\mathcal{K}_\ell} - \tau_n(L_n+1)((N_n+1)B_n)^{L_n+1}}\right) - \frac{n^2(2^j\rho)^2/16}{C_{1,n}n + 2C_{2,n}^{1/(\mu+2)}(n2^j\rho/4)^{(2\mu+3)/(\mu+2)}}\right). \qquad (8.33)\end{aligned}$$

Let,

$$\begin{aligned}&2C_{2,n}^{1/(\mu+2)}(n2^j\rho/4)^{(2\mu+3)/(\mu+2)} > C_{1,n}n, \\ &\qquad\Longrightarrow \rho > \frac{\left(C_{1,n}/(2C_{2,n}^{1/(\mu+2)})\right)^{(\mu+2)/(2\mu+3)}}{n^{(\mu+1)/(2\mu+3)}} \coloneqq \rho_n. \qquad (8.34)\end{aligned}$$

由（8.34）可知

$$-\frac{n^2(2^j\rho)^2/16}{C_{1,n}n + 2C_{2,n}^{1/(\mu+2)}(n2^j\rho/4)^{(2\mu+3)/(\mu+2)}} \leq -\frac{n^2(2^j\rho)^2/16}{4C_{2,n}^{1/(\mu+2)}(n2^j\rho/4)^{(2\mu+3)/(\mu+2)}}.$$

步骤 1: $\rho > 1 > \rho_n$。回顾条件

$$\tau_n \leq \frac{1}{4\mathcal{K}_\ell(L_n+1)((N_n+1)B_n)^{L_n+1}}.$$

我们有

$$P\Big\{\sup_{g\in\mathcal{G}_{n,j,\rho}}\Big[\mathbb{E}[g(Z_0)]-\frac{1}{n}\sum_{i=1}^{n}g(Z_i)\Big]>\frac{2^j\rho}{2}\Big\}$$
$$\leq \exp\left(2\frac{2^j\rho}{\lambda_n}(L_n+1)\log\left(\frac{(L_n+1)(N_n+1)B_n}{\frac{2^j\rho}{2\mathcal{K}_\ell}-\tau_n(L_n+1)((N_n+1)B_n)^{L_n+1}}\right)-\frac{n^2(2^j\rho)^2/16}{4C_{2,n}^{1/(\mu+2)}(n2^j\rho/4)^{(2\mu+3)/(\mu+2)}}\right). \tag{8.35}$$

由于 $2^j\rho>1$

$$\frac{2^j\rho}{4\mathcal{K}_\ell}>\frac{1}{4\mathcal{K}_\ell}\Longrightarrow\frac{2^j\rho}{2\mathcal{K}_\ell}-\frac{1}{4\mathcal{K}_\ell}>\frac{2^j\rho}{2\mathcal{K}_\ell}-\frac{2^j\rho}{4\mathcal{K}_\ell}=\frac{2^j\rho}{4\mathcal{K}_\ell}$$
$$\Longrightarrow\frac{1}{\frac{2^j\rho}{2\mathcal{K}_\ell}-\frac{1}{4\mathcal{K}_\ell}}<\frac{4\mathcal{K}_\ell}{2^j\rho}\Longrightarrow\frac{(L_n+1)(N_n+1)B_n}{\frac{2^j\rho}{2\mathcal{K}_\ell}-\frac{1}{4\mathcal{K}_\ell}}<\frac{4\mathcal{K}_\ell(L_n+1)(N_n+1)B_n}{2^j\rho}. \tag{8.36}$$

因此，由（8.36），并应用不等式 $\log x < m(x^{1/m}-1)$，对所有 $x>0,\quad m>0$，我们有

$$P\Big\{\sup_{g\in\mathcal{G}_{n,j,\rho}}\Big[\mathbb{E}[g(Z_0)]-\frac{1}{n}\sum_{i=1}^{n}g(Z_i)\Big]>\frac{2^j\rho}{2}\Big\}$$
$$\leq \exp\left(2\frac{2^j\rho}{\lambda_n}(L_n+1)m\left(\left(\frac{4\mathcal{K}_\ell(L_n+1)(N_n+1)B_n}{2^j\rho}\right)^{1/m}-1\right)-\frac{n^2(2^j\rho)^2/16}{4C_{2,n}^{1/(\mu+2)}(n2^j\rho/4)^{(2\mu+3)/(\mu+2)}}\right)$$
$$\leq \exp\left(2\frac{(L_n+1)m}{\lambda_n}(2^j\rho)^{1-1/m}\left(4\mathcal{K}_\ell(L_n+1)(N_n+1)B_n\right)^{1/m}-2\frac{2^j\rho}{\lambda_n}(L_n+1)m-\frac{n^{1/(\mu+2)}(2^j\rho)^{1/(\mu+2)}/16}{4C_{2,n}^{1/\mu+2}(1/4)^{(2\mu+3)/(\mu+2)}}\right). \tag{8.37}$$

Let,

$$1-\frac{1}{m}=\frac{1}{\mu+2}\Longleftrightarrow m=\frac{\mu+2}{\mu+1}.$$

因此,

$$P\Big\{\sup_{g\in\mathcal{G}_{n,j,\rho}}\Big[\mathbb{E}[g(Z_0)]-\frac{1}{n}\sum_{i=1}^{n}g(Z_i)\Big]>\frac{2^j\rho}{2}\Big\}$$
$$\leq \exp\left(2\frac{(L_n+1)(\mu+2)}{\lambda_n(\mu+1)}(2^j\rho)^{1/(\mu+2)}\left(4\mathcal{K}_\ell(L_n+1)(N_n+1)B_n\right)^{\mu+1/(\mu+2)}-2\frac{2^j\rho(L_n+1)(\mu+2)}{\lambda_n(\mu+1)}\right.$$
$$\left.-\frac{n^{1/(\mu+2)}(2^j\rho)^{1/(\mu+2)}/16}{4C_{2,n}^{1/\mu+2}(1/4)^{(2\mu+3)/(\mu+2)}}\right)$$
$$\leq \exp\left(2\frac{(L_n+1)(\mu+2)}{\lambda_n(\mu+1)}(2^j\rho)^{1/(\mu+2)}\left(4\mathcal{K}_\ell(L_n+1)(N_n+1)B_n\right)^{\mu+1/(\mu+2)}-\frac{n^{1/(\mu+2)}(2^j\rho)^{1/(\mu+2)}/16}{4C_{2,n}^{1/\mu+2}(1/4)^{(2\mu+3)/(\mu+2)}}\right)$$
$$\leq \exp\left((2^j\rho)^{1/(\mu+2)}\left(2\frac{(L_n+1)(\mu+2)}{\lambda_n(\mu+1)}\left(4\mathcal{K}_\ell(L_n+1)(N_n+1)B_n\right)^{\mu+1/(\mu+2)}-\frac{n^{1/(\mu+2)}/16}{4C_{2,n}^{1/\mu+2}(1/4)^{(2\mu+3)/(\mu+2)}}\right)\right). \tag{8.38}$$

回顾以下假设

$$L_n \lesssim \log n,\ N_n \lesssim n^{\nu_1}, B_n \lesssim n^{\nu_2},\ \lambda_n \asymp (\log n)^{\nu_3}/n^{\nu_4},\ \nu_1+\nu_2+\nu_4 < \frac{1}{\mu+2} \quad \text{for some } \nu_1,\ \nu_2,\ \nu_3,\ \nu_4 > 0. \tag{8.39}$$

在上述关于 (8.39) 的条件下，存在 $N_1 \in \mathbb{N}$，使得若 $n > N_1$，则有

$$2\frac{(L_n+1)(\mu+2)}{\lambda_n(\mu+1)}\left(4\mathcal{K}_\ell(L_n+1)(N_n+1)B_n\right)^{\mu+1/(\mu+2)} < \frac{n^{1/(\mu+2)}/16}{8C_{2,n}^{1/\mu+2}(1/4)^{(2\mu+3)/(\mu+2)}}. \tag{8.40}$$

因此，我们有

$$P\Big\{\sup_{g\in\mathcal{G}_{n,j,\rho}}\Big[\mathbb{E}[g(Z_0)] - \frac{1}{n}\sum_{i=1}^n g(Z_i)\Big] > \frac{2^j\rho}{2}\Big\} \le \exp\left(-(2^j\rho)^{1/(\mu+2)}\frac{n^{1/(\mu+2)}/16}{8C_{2,n}^{1/\mu+2}(1/4)^{(2\mu+3)/(\mu+2)}}\right). \tag{8.41}$$

$$\sum_{j=1}^\infty P\Big\{\sup_{g\in\mathcal{G}_{n,j,\rho}}\Big[\mathbb{E}[g(Z_0)] - \frac{1}{n}\sum_{i=1}^n g(Z_i)\Big] > \frac{2^j\rho}{2}\Big\} \le \sum_{j=1}^\infty \exp\left(-(2^{1/(\mu+2)})^j\frac{n^{1/(\mu+2)}/16}{8C_{2,n}^{1/\mu+2}(1/4)^{(2\mu+3)/(\mu+2)}}\rho^{1/(\mu+2)}\right). \tag{8.42}$$

Let,

$$\beta_n \coloneqq \frac{n^{1/(\mu+2)}/16}{8C_{2,n}^{(1/\mu+2)}(1/4)^{(2\mu+3)/(\mu+2)}}\rho^{1/(\mu+2)},$$

我们有

$$\sum_{j=0}^\infty \exp\left(-\beta_n(2^{1/(\mu+2)})^j\right) \le \exp(-\beta_n)\sum_{j=0}^\infty\left((2^{1/(\mu+2)})^{-\frac{\beta_n}{\log 2}}\right)^j \le \exp(-\beta_n)\left(\frac{1}{1-(2^{1/(\mu+2)})^{-\frac{\beta_n}{\log 2}}}\right) \le 2\exp(-\beta_n). \tag{8.43}$$

因此，

$$P(B_{1,n} > \rho) \le 2\exp\left(-\frac{n^{1/(\mu+2)}/16}{8C_{2,n}^{1/\mu+2}(1/4)^{(2\mu+3)/(\mu+2)}}\rho^{1/(\mu+2)}\right). \tag{8.44}$$

因此，对于 $\rho > 1$，我们有

$$\int_1^\infty P(B_{1,n} > \rho)d\rho \le 2\int_1^\infty \exp\left(-\frac{n^{1/(\mu+2)}/16}{8C_{2,n}^{1/\mu+2}(1/4)^{(2\mu+3)/(\mu+2)}}\rho^{1/(\mu+2)}\right)d\rho.$$

Let

$$x = \frac{n^{1/(\mu+2)}/16}{8C_{2,n}^{1/\mu+2}(1/4)^{(2\mu+3)/(\mu+2)}}\rho^{1/(\mu+2)},\ g(x) \coloneqq x - 2(\mu+2)\log(x).$$

令 $x_0 = (2(\mu+2))^3$。容易看出 $g(x_0) > 0$，因此只需证明

$$x > x_0 \text{ that is } \quad \frac{n^{1/(\mu+2)}/16}{8C_{2,n}^{1/\mu+2}(1/4)^{(2\mu+3)/(\mu+2)}}\rho^{1/(\mu+2)} > (2(\mu+2))^3, \quad \forall \rho > 1,$$

即可，

$$n > \left(128C_{2,n}^{1/\mu+2}(1/4)^{(2\mu+3)/(\mu+2)}(2(\mu+2))^3\right)^{\mu+2}.$$

可以注意到，对于任意

$$x > x_0, \quad \text{we have} \quad \exp(-x) < \frac{1}{x^{2(\mu+2)}}. \tag{8.45}$$

由（8.45）可知

$$\begin{aligned}\int_1^\infty P(B_{1,n}>\rho)d\rho &\le 2\int_1^\infty \exp\left(-\frac{n^{1/(\mu+2)}/16}{8C_{2,n}^{1/\mu+2}(1/4)^{(2\mu+3)/(\mu+2)}}\rho^{1/(\mu+2)}\right)d\rho\\ &\le 2\int_1^\infty \frac{1}{\left(\frac{n^{1/(\mu+2)}/16}{8C_{2,n}^{1/\mu+2}(1/4)^{(2\mu+3)/(\mu+2)}}\rho^{1/(\mu+2)}\right)^{2(\mu+2)}}d\rho\\ &\le 2\frac{(128)^{(2(\mu+2))}C_{2,n}^2(1/4)^{(2(2\mu+3))}}{n^2}\int_1^\infty \frac{1}{\rho^2}d\rho\\ &\le 2\frac{(128)^{(2(\mu+2))}(1/4)^{(2(2\mu+3))}C_{2,n}^2}{n^2}. \end{aligned} \tag{8.46}$$

步骤 2: $\rho_n<\rho<1$。回顾该条件

$$\tau_n \le \frac{\rho_n}{4\mathcal{K}_\ell(L_n+1)((N_n+1)B_n)^{L_n+1}}.$$

由于 $\rho>\rho_n$，我们还有

$$\tau_n \le \frac{\rho}{4\mathcal{K}_\ell(L_n+1)((N_n+1)B_n)^{L_n+1}}.$$

Thus

$$\begin{aligned}&P\Big\{\sup_{g\in\mathcal{G}_{n,j,\rho}}\Big[\mathbb{E}[g(Z_0)]-\frac{1}{n}\sum_{i=1}^n g(Z_i)\Big]>\frac{2^j\rho}{2}\Big\}\\ &\le \exp\left(2\frac{2^j\rho}{\lambda_n}(L_n+1)\log\left(\frac{(L_n+1)(N_n+1)B_n}{\frac{2^j\rho}{2\mathcal{K}_\ell}-\tau_n(L_n+1)((N_n+1)B_n)^{L_n+1}}\right)-\frac{n^2(2^j\rho)^2/16}{4C_{2,n}^{1/\mu+2}(n2^j\rho/4)^{(2\mu+3)/(\mu+2)}}\right)\\ &\le \exp\left(2\frac{2^j}{\lambda_n}(L_n+1)\log\left(\frac{(L_n+1)(N_n+1)B_n}{\frac{\rho}{2\mathcal{K}_\ell}\left(2^j-\frac{1}{2}\right)}\right)-\frac{n^{1/(\mu+2)}(2^j\rho)^{1/(\mu+2)}/16}{4C_{2,n}^{1/\mu+2}(1/4)^{(2\mu+3)/(\mu+2)}}\right). \end{aligned} \tag{8.47}$$

因此我们有了

$$\begin{aligned}&P\Big\{\sup_{g\in\mathcal{G}_{n,j,\rho}}\Big[\mathbb{E}[g(Z_0)]-\frac{1}{n}\sum_{i=1}^n g(Z_i)\Big]>\frac{2^j\rho}{2}\Big\}\\ &\le \exp\left(2\frac{2^j}{\lambda_n}(L_n+1)\log\left(\frac{4\mathcal{K}_\ell(L_n+1)(N_n+1)B_n}{2^j\rho}\right)-\frac{n^{1/(\mu+2)}(2^j\rho)^{1/(\mu+2)}/16}{4C_{2,n}^{1/\mu+2}(1/4)^{(2\mu+3)/(\mu+2)}}\right)\\ &\le \exp\left(2\frac{2^j}{\lambda_n}(L_n+1)m\left(\left(\frac{4\mathcal{K}_\ell(L_n+1)(N_n+1)B_n}{2^j\rho}\right)^{1/m}-1\right)-\frac{n^{1/(\mu+2)}(2^j\rho)^{1/(\mu+2)}/16}{4C_{2,n}^{1/\mu+2}(1/4)^{(2\mu+3)/(\mu+2)}}\right)\\ &\le \exp\left(2\frac{(2^j\rho)^{1-1/m}}{\lambda_n}(L_n+1)m\left(4\mathcal{K}_\ell(L_n+1)(N_n+1)B_n\right)^{1/m}-\frac{n^{1/(\mu+2)}(2^j\rho)^{1/(\mu+2)}/16}{4C_{2,n}^{1/\mu+2}(1/4)^{(2\mu+3)/(\mu+2)}}\right). \end{aligned} \tag{8.48}$$

设 $1-\dfrac{1}{\cancel{m}}=\dfrac{1}{\cancel{\mu+2}}$。我们有

$$
\begin{aligned}
&P\Big\{\sup_{g\in\mathcal{G}_{n,j,\rho}}\Big[\mathbb{E}[g(Z_0)]-\frac{1}{n}\sum_{i=1}^{n}g(Z_i)\Big]>\frac{2^j\rho}{2}\Big\}\\
&\leq\exp\left((2^j\rho)^{1/(\mu+2)}\left(\frac{2(L_n+1)(\mu+2)}{\lambda_n(\mu+1)}\left(4\mathcal{K}_\ell(L_n+1)(N_n+1)B_n\right)^{(\mu+1)/(\mu+2)}-\frac{n^{1/(\mu+2)}/16}{4C_{2,n}^{1/\mu+2}(1/4)^{(2\mu+3)/(\mu+2)}}\right)\right).
\end{aligned}
\tag{8.49}
$$

令 $\nu_4+\nu_1+\nu_2<1/(\mu+2)$。在满足上述 (8.39) 条件的情况下，存在 $N_2\in\mathbb{N}$，使得如果 $n\geq N_2$，则有

$$
\frac{2(L_n+1)(\mu+2)}{\lambda_n(\mu+1)}4\mathcal{K}_\ell(L_n+1)(N_n+1)B_n\leq\frac{n^{1/(\mu+2)}/16}{8C_{2,n}^{1/\mu+2}(1/4)^{(2\mu+3)/(\mu+2)}}.
$$

Thus

$$
P\Big\{\sup_{g\in\mathcal{G}_{n,j,\rho}}\Big[\mathbb{E}[g(Z_0)]-\frac{1}{n}\sum_{i=1}^{n}g(Z_i)\Big]>\frac{2^j\rho}{2}\Big\}\leq\exp\left(-\left(2^{1/(\mu+2)}\right)^j\frac{n^{1/(\mu+2)}/16}{8C_{2,n}^{1/\mu+2}(1/4)^{(2\mu+3)/(\mu+2)}}\rho^{1/(\mu+2)}\right).
\tag{8.50}
$$

Let

$$
\beta_n\coloneqq\frac{n^{1/(\mu+2)}/16}{8C_{2,n}^{1/\mu+2}(1/4)^{(2\mu+3)/(\mu+2)}}\rho^{1/(\mu+2)}.
$$

采用与 (8.43) 类似的论证，可得

$$
P(B_{1,n}>\rho)\leq 2\exp\left(-\frac{n^{1/(\mu+2)}/16}{8C_{2,n}^{1/\mu+2}(1/4)^{(2\mu+3)/(\mu+2)}}\rho^{1/(\mu+2)}\right).
\tag{8.51}
$$

因此，采用与**步骤 1** 类似的论证，我们有

$$
\begin{aligned}
\int_{\rho_n}^{1}P(B_{1,n}>\rho)d\rho&\leq 2\int_{\rho_n}^{1}\exp\left(-\frac{n^{1/(\mu+2)}/16}{8C_{2,n}^{1/\mu+2}(1/4)^{(2\mu+3)/(\mu+2)}}\rho^{1/(\mu+2)}\right)d\rho\\
&\leq 2\frac{(128)^{(2(\mu+2))}(1/4)^{(2(2\mu+3))}C_{2,n}^2}{n^2}\int_{\rho_n}^{1}\frac{1}{\rho^2}d\rho\\
&\leq\frac{2(128)^{(2(\mu+2))}(1/4)^{(2(2\mu+3))}C_{2,n}^2}{\left(C_{1,n}/(2C_{2,n}^{1/(\mu+2)})\right)^{(\mu+2)/(2\mu+3)}n^{(3\mu+5)/(2\mu+3)}}-2\frac{(128)^{(2(\mu+2))}(1/4)^{(2(2\mu+3))}C_{2,n}^2}{n^2}\\
&\leq\frac{(128)^{(2(\mu+2))}(1/4)^{(2(2\mu+3))}C_{2,n}^2}{\left(C_{1,n}/(2C_{2,n}^{1/(\mu+2)})\right)^{(\mu+2)/(2\mu+3)}n^{(3\mu+5)/(2\mu+3)}}.
\end{aligned}
\tag{8.52}
$$

Let,

$$
\begin{aligned}
&2C_{2,n}^{1/\mu+2}(n2^j\rho/4)^{(2\mu+3)/(\mu+2)}<C_{1,n}n\\
&\qquad\Rightarrow\rho<\frac{\left(C_{1,n}/(2C_{2,n}^{1/(\mu+2)})\right)^{(\mu+2)/(2\mu+3)}}{n^{\mu+1/(2\mu+3)}}\coloneqq\rho_n.
\end{aligned}
\tag{8.53}
$$

由 (8.53) 可知

$$-\frac{n^2(2^j\rho)^2/16}{C_{1,n}n+2C_{2,n}^{1/\mu+2}(n2^j\rho/4)^{(2\mu+3)/(\mu+2)}}\leq-\frac{n^2(2^j\rho)^2/16}{2C_{1,n}}.$$

第 3 步: $0<\rho<\rho_n$。

$$\int_0^{\rho_n}P(B_{1,n}>\rho)d\rho\leq\int_0^{\rho_n}d\rho\leq\rho_n=\frac{\left(C_{1,n}/(2C_{2,n}^{1/(\mu+2)})\right)^{(\mu+2)/(2\mu+3)}}{n^{\mu+1/(2\mu+3)}}.$$

因此，令 $N=\max(N_1,\ N_2)$，这意味着

$$\mathbb{E}[B_{1,n}]\leq\int_0^{\infty}P(B_{1,n}>\rho)d\rho\leq\frac{\left(C_{1,n}/(2C_{2,n}^{1/(\mu+2)})\right)^{(\mu+2)/(2\mu+3)}}{n^{\mu+1/(2\mu+3)}}+\frac{(128)^{2(\mu+2)}C_{2,n}^2}{\left(C_{1,n}/(2C_{2,n}^{1/(\mu+2)})\right)^{(\mu+2)/(2\mu+3)}n^{3\mu+5/(2\mu+3)}}+2\frac{(128)^{2(\mu+2)}C_{2,n}^2}{n^2}.\tag{8.54}$$

对于有界的 $B_{2,\ n}$，考虑 $h_n^\circ\in\mathcal{H}_\sigma(L_n,\ N_n,\ B_n,\ F)$，使得

$$R(h_n^\circ)+J_{\lambda_n,\tau_n}(h_n^\circ)\leq\inf_{h\in\mathcal{H}_{\sigma,n}}[R(h)+J_{\lambda_n,\tau_n}(h)]+\frac{1}{n}.\tag{8.55}$$

回顾不等式:

$$\widehat{R}_n(\widehat{h}_n)+J_{\lambda_n,\tau_n}(\widehat{h}_n)\leq\widehat{R}_n(h)+J_{\lambda_n,\tau_n}(h),\ \text{for all}\ h\in\mathcal{H}_\sigma(L_n,N_n,B_n,F).\tag{8.56}$$

Thus

$$\begin{aligned}B_{2,n}&=2[\widehat{R}_n(\widehat{h}_n)-\widehat{R}_n(h^*)]+2J_{\lambda_n,\tau_n}(\widehat{h}_n)\\&=2[\widehat{R}_n(\widehat{h}_n)+J_{\lambda_n,\tau_n}(\widehat{h}_n)-\widehat{R}_n(h_n^\circ)]+2[\widehat{R}_n(h_n^\circ)-\widehat{R}_n(h^*)]\\&\leq2J_{\lambda_n,\tau_n}(h_n^\circ)+2[\widehat{R}_n(h_n^\circ)-\widehat{R}_n(h^*)].\end{aligned}$$

因此,

$$\mathbb{E}t[B_{2,n}t]\leq2J_{\lambda_n,\tau_n}(h_n^\circ)+2\mathcal{E}_{Z_0}(h_n^\circ)\leq2\inf_{h\in\mathcal{H}_\sigma(L_n,N_n,B_n,F)}\Big[\mathcal{E}_{Z_0}(h)+J_{\lambda_n,\tau_n}(h)\Big]+\frac{1}{n}.\tag{8.57}$$

其中 $\mathcal{E}_{Z_0}(h)=R(h)-R(h^*)$。因此,

$$\begin{aligned}\mathcal{E}_{Z_0}(\widehat{h}_n)&\leq2\Big\{\inf_{h\in\mathcal{H}_\sigma(L_n,N_n,B_n,F)}\Big[\mathcal{E}_{Z_0}(h)+J_{\lambda_n,\tau_n}(h)\Big]\Big\}\\&\vee\left(\frac{\left(C_{1,n}/(2C_{2,n}^{1/(\mu+2)})\right)^{(\mu+2)/(2\mu+3)}}{n^{\mu+1/(2\mu+3)}}+\frac{(128)^{2(\mu+2)}C_{2,n}^2}{\left(C_{1,n}/(2C_{2,n},n^{1/(\mu+2)})\right)^{(\mu+2)/(2\mu+3)}n^{3\mu+5/(2\mu+3)}}+2\frac{(128)^{2(\mu+2)}C_{2,n}^2}{n^2}\right)\\&\lesssim2\Big\{\inf_{h\in\mathcal{H}_{\sigma,n}}\Big[\mathcal{E}_{Z_0}(h)+J_{\lambda_n,\tau_n}(h)\Big]\Big\}\vee\frac{C_\mu\left(C_{1,n}/(2C_{2,n}^{1/(\mu+2)})\right)^{(\mu+2)/(2\mu+3)}}{n^{\mu+1/(2\mu+3)}},\end{aligned}$$

其中正数常数 C_μ 取决于 μ。至此，定理得证。 ■

8.4 定理 5.2 的证明

我们考虑以下分解:

$$R(\widehat{h}_n) - R(h^*) \coloneqq B_{1,n} + B_{2,n}.$$

其中,

$$\begin{aligned} B_{1,n} &= [R(\widehat{h}_n) - R(h^*)] - 2[\widehat{R}_n(\widehat{h}_n) - \widehat{R}_n(h^*)] - 2J_{\lambda_n,\tau_n}(\widehat{h}_n); \\ B_{2,n} &= 2[\widehat{R}_n(\widehat{h}_n) - \widehat{R}_n(h^*)] + 2J_{\lambda_n,\tau_n}(\widehat{h}_n). \end{aligned}$$

对于界$B_{1,\ n}$，我们使用与定理 5.1 证明类似的论证。

令 $l = \mathcal{N}(\varepsilon,\ \mathcal{G}_{n,\ j,\ \rho},\ \|\cdot\|_\infty)$。由 (8.2) 可得,

$$P\Big\{\sup_{g\in\mathcal{G}_{n,j,\rho}}\Big[\mathbb{E}[g(Z_0)] - \frac{1}{n}\sum_{i=1}^{n} g(Z_i)\Big] > \varepsilon\Big\} \le \mathcal{N}\left(\varepsilon, \mathcal{G}_{n,j,\rho}\right)\exp\Big(-n^{\nu_5}\varepsilon + \frac{n^{2\nu_5-1}(M+\theta_{\infty,n}(1))^2}{2}\Big). \tag{8.58}$$

在 [20], 中，我们有

$$\mathcal{N}(\varepsilon, \mathcal{G}_{n,j,\rho}, \|\cdot\|_\infty) \le \mathcal{N}\left(\frac{\varepsilon}{\mathcal{K}_\ell}, \mathcal{H}_{n,j,\rho}, \|\cdot\|_\infty\right). \tag{8.59}$$

显然,

$$\mathcal{H}_{n,j,\rho} \subset \left\{h \in \mathcal{H}_\sigma(L_n, N_n, B_n, F, \frac{2^j\rho}{\lambda_n}) : \|\theta(h)\|_{\text{clip},\tau_n} \le \frac{2^j\rho}{\lambda_n}\right\}. \tag{8.60}$$

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

$$\begin{aligned} \|\Delta(h_1)(x,y) - \Delta(h_2)(x,y)\|_\infty &= |\ell(h_1(x), y) - \ell(h_2(x), y)| \\ &\le \mathcal{K}_\ell\,|h_1(x) - h_2(x)| \quad \text{from } \textbf{(A3)}. \end{aligned} \tag{8.61}$$

因此，在 [20], 中，我们有以下不等式

$$\begin{aligned} \mathcal{N}(\varepsilon, \mathcal{G}_{n,j,\rho}, \|\cdot\|_\infty) &\le \mathcal{N}\left(\frac{\varepsilon}{\mathcal{K}_\ell}, \mathcal{H}_{n,j,\rho}, \|\cdot\|_\infty\right) \\ &\le \mathcal{N}\left(\frac{\varepsilon}{\mathcal{K}_\ell}, \mathcal{H}_\sigma(L_n, N_n, B_n, F, \frac{2^j\rho}{\lambda_n}), \|\cdot\|_\infty\right) \\ &\le \exp\left(2\frac{2^j\rho}{\lambda_n}(L_n+1)\log\left(\frac{(L_n+1)(N_n+1)B_n}{\frac{\varepsilon}{\mathcal{K}_\ell} - \tau_n(L_n+1)((N_n+1)B_n)^{L_n+1}}\right)\right). \end{aligned} \tag{8.62}$$

Thus

$$\begin{aligned} &P\Big\{\sup_{g\in\mathcal{G}_{n,j,\rho}}\Big[\mathbb{E}[g(Z_0)] - \frac{1}{n}\sum_{i=1}^{n} g(Z_i)\Big] > \varepsilon\Big\} \\ &\le \exp\left(2\frac{2^j\rho}{\lambda_n}(L_n+1)\log\left(\frac{(L_n+1)(N_n+1)B_n}{\frac{\varepsilon}{\mathcal{K}_\ell} - \tau_n(L_n+1)((N_n+1)B_n)^{L_n+1}}\right)\right)\exp\left(-n^{\nu}\varepsilon + \frac{n^{2\nu-1}(M+\theta_{\infty,n}(1))^2}{2}\right) \\ &\le \exp\left(2\frac{2^j\rho}{\lambda_n}(L_n+1)\log\left(\frac{(L_n+1)(N_n+1)B_n}{\frac{\varepsilon}{\mathcal{K}_\ell} - \tau_n(L_n+1)((N_n+1)B_n)^{L_n+1}}\right) - n^{\nu_5}\varepsilon + \frac{n^{2\nu_5-1}(M+\theta_{\infty,n}(1))^2}{2}\right). \end{aligned} \tag{8.63}$$

令 $\varepsilon = \dfrac{2^j\rho}{2}$，我们有

$$P\Big\{\sup_{g\in\mathcal{G}_{n,j,\rho}}\Big[\mathbb{E}[g(Z_0)]-\frac{1}{n}\sum_{i=1}^{n}g(Z_i)\Big]>\frac{2^j\rho}{2}\Big\}$$
$$\leq \exp\left(2\frac{2^j\rho}{\lambda_n}(L_n+1)\log\left(\frac{(L_n+1)(N_n+1)B_n}{\frac{2^j\rho}{2\mathcal{K}_\ell}-\tau_n(L_n+1)((N_n+1)B_n)^{L_n+1}}\right)-\frac{2^j\rho}{2}n^{\nu_5}+\frac{n^{2\nu_5-1}(M+\theta_{\infty,n}(1))^2}{2}\right). \tag{8.64}$$

步骤 1：　$\rho>1$。

回顾假设

$$\tau_n\leq\frac{1}{4\mathcal{K}_\ell(L_n+1)((N_n+1)B_n)^{L_n+1}}.$$

由于 $2^j\rho>1$，我们有

$$P\Big\{\sup_{g\in\mathcal{G}_{n,j,\rho}}\Big[\mathbb{E}[g(Z_0)]-\frac{1}{n}\sum_{i=1}^{n}g(Z_i)\Big]>\frac{2^j\rho}{2}\Big\}$$
$$\leq \exp\left(2\frac{2^j\rho}{\lambda_n}(L_n+1)\log\left(4\mathcal{K}_\ell(L_n+1)(N_n+1)B_n\right)-\frac{2^j\rho}{2}n^{\nu_5}+2^j\rho\frac{n^{2\nu_5-1}(M+\theta_{\infty,n}(1))^2}{2}\right) \tag{8.65}$$
$$\leq \exp\left(2^j\rho\left(\frac{2}{\lambda_n}(L_n+1)\log\left(4\mathcal{K}_\ell(L_n+1)(N_n+1)B_n\right)+\frac{n^{2\nu_5-1}(M+\theta_{\infty,n}(1))^2}{2}-\frac{n^{\nu_5}}{2}\right)\right). \tag{8.66}$$

回顾条件

$$L_n\lesssim\log\ n,\quad N_n\lesssim n^{\nu_1},\ B_n\lesssim n^{\nu_2},\quad \lambda_n\asymp(\log\ n)^{\nu_3}/n^{\nu_4},\quad \nu_1+\nu_2+\nu_4<\nu_5\qquad \text{对于某些 } \nu_1,\ \nu_2,\ \nu_3,\ \nu_4>0,\ \nu_5<1. \tag{8.67}$$

在上述条件 (8.67) 下，存在 $N_3\in\mathbb{N}$，使得若 $n\geq N_3$，则有

$$\delta_n\coloneqq\frac{2}{\lambda_n}(L_n+1)\log\left(4\mathcal{K}_\ell(L_n+1)(N_n+1)B_n\right)+\frac{n^{2\nu_5-1}(M+\theta_{\infty,n}(1))^2}{2}<\frac{n^{\nu_5}}{4}.$$

因此,

$$P\Big\{\sup_{g\in\mathcal{G}_{n,j,\rho}}\Big[\mathbb{E}[g(Z_0)]-\frac{1}{n}\sum_{i=1}^{n}g(Z_i)\Big]>\frac{2^j\rho}{2}\Big\}\leq\exp\left(-2^j\frac{n^{\nu_5}}{4}\rho\right). \tag{8.68}$$

令 $\beta_n\coloneqq\dfrac{n^{\nu_5}}{4}\rho$。利用以下不等式 $\exp\left(-\beta_n2^j\right)\leq\exp\left(-\beta_n\right)2^{-\beta_nj}{}_{\log 2}$，可得,

$$\sum_{j=1}^{n}P\Big\{\sup_{g\in\mathcal{G}_{n,j,\rho}}\Big[\mathbb{E}[g(Z_0)]-\frac{1}{n}\sum_{i=1}^{n}g(Z_i)\Big]>\frac{2^j\rho}{2}\Big\}\leq\exp\left(-\beta_n\right)\sum_{j=1}^{n}\left(2^{-\frac{\beta_n}{\log 2}}\right)^j\leq 2\exp\left(-\beta_n\right). \tag{8.69}$$

应用 (8.69) 可得

$$P\left(B_{1,n}>\rho\right)\leq 2\exp\left(-\frac{n^{\nu_5}}{4}\rho\right). \tag{8.70}$$

我们有,

$$\int_1^\infty P\left(B_{1,n}>\rho\right)d\rho\leq 2\int_1^\infty\exp\left(-\frac{n^{\nu_5}}{4}\rho\right)d\rho=\frac{8}{n^{\nu_5}}\exp\left(-\frac{n^{\nu_5}}{4}\right). \tag{8.71}$$

步骤 2: $\rho_n < \rho < 1$。回顾假设

$$\tau_n \leq \frac{\rho_n}{4\mathcal{K}_\ell(L_n+1)((N_n+1)B_n)^{L_n+1}}, \text{ with } \rho_n \coloneqq \frac{1}{n^{2\nu_6}}.$$

由于 $\rho \in (\rho_n, 1)$，我们有

$$\begin{aligned}
&P\Big\{\sup_{g\in\mathcal{G}_{n,j,\rho}}\Big[\mathbb{E}[g(Z_0)] - \frac{1}{n}\sum_{i=1}^n g(Z_i)\Big] > \frac{2^j\rho}{2}\Big\} \\
&\leq \exp\left(2\frac{2^j\rho}{\lambda_n}(L_n+1)\log\left(\frac{(L_n+1)(N_n+1)B_n}{\frac{2^j\rho}{2\mathcal{K}_\ell} - \tau_n(L_n+1)((N_n+1)B_n)^{L_n+1}}\right) - \frac{2^j\rho}{2}n^{\nu_5} + \frac{n^{2\nu_5-1}(M+\theta_{\infty,n}(1))^2}{2}\right) \\
&\leq \exp\left(2\frac{2^j\rho}{\lambda_n}(L_n+1)\log\left(4\mathcal{K}_\ell\frac{(L_n+1)(N_n+1)B_n}{\rho_n}\right) - \frac{2^j\rho}{2}n^{\nu_5} + \frac{2^j\rho n^{2\nu_5-1}(M+\theta_{\infty,n}(1))^2}{2\rho_n}\right) \\
&\leq \exp\left(2^j\rho\left(\frac{2}{\lambda_n}(L_n+1)\log\left(\frac{4\mathcal{K}_\ell(L_n+1)(N_n+1)B_n}{1/n^{2\nu_6}}\right) + \frac{n^{2\nu_5+2\nu_6-1}(M+\theta_{\infty,n}(1))^2}{2} - \frac{n^{\nu_5}}{2}\right)\right). \qquad (8.72)
\end{aligned}$$

回顾条件

$L_n \lesssim \log\ n,\ N_n \lesssim n^{\nu_1}, B_n \lesssim n^{\nu_2},\ \lambda_n \asymp (\log\ n)^{\nu_3}/n^{\nu_4},\ \nu_4+2\nu_6+\nu_1+\nu_2 < \nu_5$ 对于 $\nu_1,\ \nu_2,\ \nu_3,\ \nu_4 > 0,\ \nu_5 < 1,\ \nu_6 < \frac{1-\nu_5}{2}$.

(8.73) U

在上述条件 (8.73) 下，存在 $N_4 \in \mathbb{N}$，使得如果 $n \geq N_4$，我们有

$$\frac{2}{\lambda_n}(L_n+1)\log\left(\frac{4\mathcal{K}_\ell(L_n+1)(N_n+1)B_n}{1/n^{2\nu_6}}\right) + \frac{n^{2\nu_5+2\nu_6-1}(M+\theta_{\infty,n}(1))^2}{2} < \frac{n^{\nu_5}}{4}.$$

因此,

$$P\Big\{\sup_{g\in\mathcal{G}_{n,j,\rho}}\Big[\mathbb{E}[g(Z_0)] - \frac{1}{n}\sum_{i=1}^n g(Z_i)\Big] > \frac{2^j\rho}{2}\Big\} \leq \exp\left(-2^j\frac{n^{\nu_5}}{4}\rho\right)$$

设 $\beta_{n,2} \coloneqq \frac{n^{\nu_5}}{4}\rho$。因此,

$$P\left(B_{1,n} > \alpha\right) \leq 2\exp\left(-\frac{n^{\nu_5}}{4}\rho\right) \qquad (8.74)$$

因此,

$$\begin{aligned}
\int_{\rho_n}^1 P\left(B_{1,n} > \rho\right)d\rho &\leq 2\int_{\rho_n}^1 \exp\left(-\frac{n^{\nu_5}}{4}\rho\right)d\rho \\
&\leq \frac{8}{n^{\nu_5}}\exp\left(-\frac{n^{\nu_5-2\nu_6}}{4}\right) - \frac{8}{n^{\nu_5}}\exp\left(-\frac{n^{\nu_5}}{4}\right) \leq \frac{4}{n^{\nu_5}}\exp\left(-\frac{n^{\nu_5-2\nu_6}}{4}\right). \qquad (8.75)
\end{aligned}$$

第 3 步: $0 \leq \rho \leq \rho_n$.

我们有

$$\int_0^{\rho_n} P\left(B_{1,n} > \rho\right)d\rho \leq \int_0^{\rho_n} d\rho = \rho_n = \frac{1}{n^{2\nu_6}}. \qquad (8.76)$$

因此，令 $N = \max(N_3,\ N_4)$，由此可知

$$\begin{aligned}
\mathbb{E}[B_{1,n}] &\leq \int_0^\infty P\left(B_{1,n} > \rho\right)d\rho \\
&\leq \frac{8}{n^{\nu_5}}\exp\left(-\frac{n^{\nu_5}}{8}\right) + \frac{1}{n^{2\nu_6}} + \frac{4}{n^{\nu_5}}\exp\left(-\frac{n^{\nu_5-2\nu_6}}{4}\right). \qquad (8.77)
\end{aligned}$$

可以像定理 5.1 的证明中那样对 $B_{2,n}$ 进行界定，另见 (8.57)。因此，

$$\begin{aligned}\mathcal{E}_{Z_0}(\widehat{h}_n) &= R(\widehat{h}_n) - R(h^*) \\ &\le 2\Big\{\inf_{h\in\mathcal{H}_{\sigma,n}}\Big[\mathcal{E}_{Z_0}(h) + J_{\lambda_n,\tau_n}(h)\Big]\Big\} \vee \left(\frac{8}{n^{\nu_5}}\exp\left(-\frac{n^{\nu_5}}{8}\right) + \frac{1}{n^{2\nu_6}} + \frac{4}{n^{\nu_5}}\exp\left(-\frac{n^{\nu_5-2\nu_6}}{4}\right)\right) \\ &\lesssim 2\Big\{\inf_{h\in\mathcal{H}_{\sigma,n}}\Big[\mathcal{E}_{Z_0}(h) + J_{\lambda_n,\tau_n}(h)\Big]\Big\} \vee \frac{C}{n^{2\nu_6}}, \end{aligned} \tag{8.78}$$

其中 C 为正数。因此，定理得证。 ■

8.5 推论 5.3 的证明

设 $L_n,\ N_n,\ S_n,\ B_n,\ F_n > 0$，并考虑深度神经网络类 $\mathcal{H}_{\sigma,\,n} = \mathcal{H}_\sigma(L_n,\ N_n,\ S_n,\ B_n,\ F_n) \subset \mathcal{F}(\mathcal{X},\ \mathcal{Y})$。在 ℓ 满足 Lipschitz 性质的条件下，对于所有 $h \in \mathcal{H}_{\sigma,\,n}$，有：

$$R(h) - R(h^*) = \mathbb{E}_{Z_0}[\ell(h(X_0), Y_0)] - \mathbb{E}_{Z_0}[\ell(h^*(X_0), Y_0)] \le \mathcal{K}_\ell \mathbb{E}_{X_0}[|h(X_0) - h^*(X_0)|]. \tag{8.79}$$

由于对某个 $\mathcal{K} > 0$ 有 $h^* \in C^{s,\mathcal{K}}(\mathcal{X})$，因此，对于所有 $\epsilon > 0$，我们都可以找到正常数 L_0, N_0, S_0, B_0，使得，在♀♀♀♀♀♀

$$L_n = L_0 \log_+(1/\epsilon),\ N_n = N_0\epsilon^{-d/s},\ S_n = S_0\epsilon^{-d/s}\log_+(1/\epsilon) \text{ and } B_n = B_0\epsilon^{-4(d/(s+1))}, \tag{8.80}$$

存在一个神经网络 $h_\epsilon \in \mathcal{H}_{\sigma,\,n}$ 满足，

$$\|h_\epsilon - h^*\|_{\infty,\mathcal{X}} < \epsilon, \tag{8.81}$$

见 [12]；并且对所有 $x > 0$，$\log_+(x) = \max(1,\ \log\ x)$。因此，

$$\mathbb{E}_{X_0}\left[|h_\epsilon(X_0) - h^*(X_0)|\right] < \epsilon.$$

令 $\epsilon_n = n^{-\frac{\nu_4}{d\nu_4(1+d/s)}}$。容易看出，条件 $L_n \lesssim \log\ n,\quad N_n \lesssim n^{\nu_1},\quad 1 \le B_n \lesssim n^{\nu_2}$ 均得到满足，其中 $\nu_1 = \frac{4d\nu_4}{s}(1+d/s)\ \nu_2 = (s+1)(1+d/s)$。因此，利用定理（5.2）和（8.81），以及对于任意 $h \in \mathcal{H}_{\sigma,\,n},\quad \|\theta(h)\|_{\text{clip},\,\tau} \le \|\theta(h)\|_0$ 对任意 $\tau > 0$成立这一事实，我们有，

$$\begin{aligned}\mathbb{E}[\mathcal{E}_{Z_0}(\widehat{h}_n)] &\lesssim 2\Big\{\inf_{h\in\mathcal{H}_\sigma(L_n,N_n,S_n,B_n,F)}\Big[\mathcal{E}_{Z_0}(h) + J_{\lambda_n,\tau_n}(h)\Big]\Big\} \vee \frac{C}{n^{2\nu_6}} \\ &\lesssim 2\Big\{\inf_{h\in\mathcal{H}_\sigma(L_n,N_n,S_n,B_n,F)}\Big[\mathcal{E}_{Z_0}(h) + \lambda_n S_0\epsilon_n^{-d/s}\log(1/\epsilon_n)\Big]\Big\} \vee \frac{C}{n^{2\nu_6}} \\ &\lesssim 2\Big\{\inf_{h\in\mathcal{H}_\sigma(L_n,N_n,S_n,B_n,F)}[\mathcal{E}_{Z_0}(h)] + \lambda_n S_0\epsilon_n^{-d/s}\log(1/\epsilon_n)\Big\} \vee \frac{C}{n^{2\nu_6}} \\ &\lesssim 2\Big\{\mathcal{K}_\ell\mathbb{E}_{X_0}\left[|h_{\epsilon_n}(X_0) - h^*(X_0)|\right] + \lambda_n S_0\epsilon_n^{-d/s}\log(1/\epsilon_n)\Big\} \vee \frac{C}{n^{2\nu_6}} \\ &\lesssim 2\Big(\mathcal{K}_\ell\epsilon_n + \lambda_n S_0\epsilon_n^{-d/s}\log(1/\epsilon_n)\Big) \vee \frac{C}{n^{2\nu_6}} \\ &\lesssim \frac{2\nu_4(\log n)^{\nu_3+1}}{(1+d/s)n^{\nu_4/(1+d/s)}} \vee \frac{C}{n^{2\nu_6}}. \end{aligned} \tag{8.82}$$

参考文献

[1] Alquier, P., Li, X., and Wintenberger, O. 基于统计学习的时间序列预测：一般损失函数与快速收敛速率。《依赖性建模》1，2013（2013），65–93。[2] Alquier, P., and Wintenberger, O. 弱依赖时间序列预测的模型选择。《伯努利》18，3（2012），883–913。[3] Bauer, B., and Kohler, M. 关于深度学习如何解决非参数回归中的维数灾难。[4] 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。《熵》21，1（2019），57。[5] 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》。施普林格，2007，第9–20页。[6] 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。《电子统计学杂志》16，1（2022），116–157。[7] Diop, M. L., and Kengne, W. 针对 ψ-弱依赖过程的统计学习。arXiv 预印本 arXiv:2210.00088（2022）。[8] 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。《随机过程及其应用》84，2（1999），313–342。[9] Doukhan, P., and Neumann, M. H. 弱依赖随机变量之和的概率与矩不等式及其应用。《随机过程及其应用》117，7（2007），878–903。[10] Imaizumi, M., and Fukumizu, K. 深度神经网络在估计具有超曲面奇异性的函数方面的优势。《机器学习研究杂志》23（2022），1–54。[11] Kengne, W. 弱依赖条件下深度学习的超额风险界。arXiv 预印本 arXiv:2302.07503（2023）。[12] Kengne, W., and Modou, W. 针对 ψ-弱依赖过程的深度学习。arXiv 预印本 arXiv:2302.00333（2023）。[13] Kengne, W., and Wade, M. 弱依赖条件下的稀疏惩罚深度神经网络估计量。arXiv 预印本 arXiv:2303.01406（2023）。[14] 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。《神经网络》138（2021），179–197。[15] Kingma, D. P., and Ba, J. Adam：一种随机优化方法。arXiv 预印本 arXiv:1412.6980（2014）。[16] Kohler, M., and Krzyzak, A. 具有依赖数据的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）。

[17] 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）。[18] Ma, M., and Safikhani, A. 时间依赖观测数据的深度神经网络理论分析。arXiv 预印本 arXiv:2210.11530（2022）。[19] Ohn, I., and Kim, Y. 一般激活函数下深度神经网络对光滑函数的逼近。Entropy 21, 7（2019），627。[20] 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），476–517。[21] Rio, E. 弱相关随机过程的渐近理论，第31卷。Springer Science & Business Media，1999。[22] Rio, E. 关于依赖序列的 Lipschitz 函数的 Hoeffding 不等式。Comptes Rendus de l'Académie des Sciences-Series I-Mathematics 330, 10（2000），905–908。[23] Schmidt-Hieber，J. 流形上函数的深度 ReLU 网络逼近。arXiv 预印本 arXiv:1908.00695（2019）。[24] 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），1875–1897。[25] Souza，J. B.，Reisen，V. A.，Franco，G. C.，Ispány，M.，Bondon，P.，and Santos，J. M.结合主成分分析的广义加性模型：在呼吸系统疾病和空气污染数据时间序列中的应用。Journal of the Royal Statistical Society Series C: Applied Statistics 67, 2（2018），453–480。[26] Tsuji, K., and Suzuki, T. 变指数 Besov 空间回归问题中深度学习的估计误差分析。[27] Zhang，T. 稀疏正则化多阶段凸松弛的分析。Journal of Machine Learning Research 11, 3（2010）。