arXiv:2302.00333v1 [stat.ML] 1 Feb 2023

# 面向 $\psi$-弱依赖过程的深度学习

2023年2月2日

William Kengne [1] 和 Modou Wade [2]

*THEMA，CY 塞尔吉巴黎大学，33 Boulevard du Port，95011 塞尔吉-蓬图瓦兹 Cedex，法国* 电子邮箱：*william.kengne@cyu.fr*；*modou.wade@cyu.fr*

**摘要**：本文采用深度神经网络，对 $\psi$-弱依赖过程进行学习。这种弱依赖性质涵盖了一类弱依赖条件，例如混合、关联，$\cdots$而本文考虑的情形包括许多常见场景，例如：回归估计、时间序列预测、时间序列分类，$\cdots$本文证明了深度神经网络预测器类中经验风险最小化算法的一致性。我们得到了泛化界，并获得了一个学习率，该学习率对所有 $\alpha > 2$ 均小于 $\mathcal{O}(n^{-1/\alpha})$。本文还将方法应用于二元时间序列分类，以及带外生协变量的仿射因果模型中的预测。文中给出了一些模拟结果，并将方法应用于美国经济衰退数据。

关键词：深度神经网络，$\psi$-弱依赖，经验风险最小化原则，泛化界，一致性。

## 1 引言

深度学习在文献中备受关注，并已在多个领域取得巨大成功，例如图像处理 [27], 语音识别 [19]以及一般而言，在产业中发挥人工智能（AI）作用。人们对深度学习感兴趣，是因为深度神经网络（DNN）算法在众多应用中具有很高的准确率，尽管这些算法在许多情形下的理论性质尚未得到充分研究。深度神经网络的一项重要性质是，它们可以用于恰当地逼近多类函数，包括单变量函数和多变量函数，例如参见 [32], [31] 或 [33]。过去几年里，许多研究者致力于理解深度神经网络的理论优势。对于独立同分布（i.i.d.）观测值的一些研究结果，除其他论文外，可参见 [17],[31], [4], [34], [33], [24]。然而，独立同分布假设在许多现实应用中并不成立，例如市场预测、信号处理、气象观测、医疗预约预测，...关于依赖观测值或非独立同分布观测值下的深度学习，已有许多研究成果，例如参见 [5], [26], [28], [30]及其中的参考文献。

[1]本项目由 ANRBREAKRISK：ANR-17-CE26-0001-01 和 CY 卓越计划（“未来投资”项目，ANR-16-IDEX-0008）资助，项目“EcoDep” PSI-AAP2020-0000000013[2]由 MME-DII 卓越中心（ANR-11-LABEX-0023-01）支持

我们考虑来自平稳遍历过程 $\{Z_t = (X_t, Y_t)\ t \in \mathbb{Z}\}$ 的 $D_n = \{Z_1 = (X_1, Y_1) \cdots, Z_n = (X_n, Y_n)\}$（训练样本），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。我们研究一类深度神经网络预测函数 $\mathcal{H}_\sigma(L, B, N, F, S)$（见（2.4））以及损失函数 $\ell\colon \mathbb{R} \times \mathcal{Y} \to [0, \infty)$。对于任意假设函数 $h \in \mathcal{H}_\sigma(L, B, N, F, S)$，定义风险如下：

$$R(h) = \mathbb{E}[\ell(h(X_0), Y_0)],$$

以及（相对于 $D_n$的）经验风险，

$$\widehat{R}_n(h) = \frac{1}{n} \sum_{i=1}^{n} \ell(h(X_i), Y_i). \tag{1.1}$$

目标是根据观测值 $D_n$ 构建一个深度神经网络预测器 $h_n \in \mathcal{H}_\sigma(L, N, B, F, S)$，使其风险较低；也就是说，对于任意 $t \in \mathbb{Z}$，$h_n(X_t)$ 的平均值都与 $Y_t$“接近”。下文中，对所有 $z = (x, y) \in \mathcal{Z} = \mathcal{X} \times \mathcal{Y}$，我们令 $\ell(h, z) = \ell(h(x) y)$

设 $h_{\mathcal{H}_\sigma(L, N, B, F, S)}$ 为一个目标神经网络（假定其存在），其定义如下：

$$h_{\mathcal{H}_\sigma(L,N,B,F,S)} = \underset{h \in \mathcal{H}_\sigma(L,N,B,F,S)}{\operatorname{argmin}} R(h)\ , \tag{1.2}$$

and $\widehat{h}_n$ 由经验风险最小化（ERM）算法得到的神经网络 ，

$$\widehat{h}_n = \underset{h \in \mathcal{H}_\sigma(L,N,B,F,S)}{\operatorname{argmin}} \widehat{R}_n(h)\ . \tag{1.3}$$

我们考虑经验风险最小化原则，并旨在研究通过 $h_n$ 对 $h_{\mathcal{H}_\sigma(L,N,B,F,S)}$ 进行估计的适用性。也就是说，通过研究 $R(h_n)$ 与 $R(h_{\mathcal{H}_\sigma(L,N,B,F,S)})$ 的接近程度，来考察经验风险最小化算法的泛化能力。如果 $R(h_n) - R(h_{\mathcal{H}_\sigma(L,N,B,F,S)}) = o_P(1)$，则称该经验风险最小化算法在深度神经网络预测器类 $\mathcal{H}_\sigma(L, N, B, F, S)$ 中是一致的。

对于一种学习算法，我们希望针对任意固定的 $n$ 校准泛化误差的界（非渐近性质），并研究其一致性（渐近性质）。如上所述，已有许多工作基于对独立同分布观测值下深度神经网络理论性质的研究。然而，迄今为止，针对相关观测值的理论研究仍然很少。 [28] 在某些混合条件下，研究了用于非参数时间序列回归的非惩罚和稀疏惩罚深度神经网络估计量。 [30] 已得到了针对 $\alpha$-混合观测值的深度神经网络预测误差的一致性速率。上述工作是在混合类型条件下针对时间序列回归开展的，并未考虑更一般的情形，例如模式识别。

这项新工作研究了用于学习取值于 $\mathcal{Z} = \mathcal{X} \times \mathcal{Y} \subset \mathbb{R}^d \times \mathbb{R}$（其中 $d \in \mathbb{N}$）的 $\psi$-弱依赖过程 $\{Z_t = (X_t,\ Y_t)\ t \in \mathbb{Z}\}$ 的深度神经网络，其依据是训练样本 $D_n = \{Z_1 = (X_1,\ Y_1) \cdots,\ Z_n = (X_n,\ Y_n)\}$；并探讨以下问题。

**(i) 深度神经网络类上经验风险最小化算法的一致性**。我们针对具有广泛激活函数类别的深度神经网络预测器，建立了使用一般损失函数的 EMR 算法在学习 $\psi$-弱依赖观测值时的一致性。许多经典模型，例如 ARMAX、TARX、GARCH-X、...（见第 6 节）或计数时间序列（例如，见 [11] 中的命题 1），都满足这种弱

依赖结构。此外，这种依赖概念比混合条件更一般，因为众所周知，许多 $\psi$-弱依赖过程并不满足混合条件，例如参见[7]。从这个意义上说，本文得到的结果比文献中已有的结果更具一般性。

**(ii) 深度神经网络的泛化界与收敛速率**。推导了深度神经网络预测器类上的泛化界以及学习率。对于所有 $\alpha > 2$，该速率小于 $\mathcal{O}(n^{-1/\alpha})$；而后者接近于独立同分布情形的一般设定下所获得的通常 $\mathcal{O}(n^{-1/2})$。

**(iii) 应用于时间序列分类**。针对一类弱依赖过程的二分类进行了应用研究。为此，我们处理一个利普希茨替代损失函数，并建立了深度神经网络预测器类上经验风险最小化算法的一致性；该类预测器享有所获得的泛化界。此外，还考虑了对美国经济衰退数据的真实数据应用。

**(iv) 应用于具有外生协变量的仿射因果模型**。该类模型包括许多经典的自回归模型，例如 ARMAX、TARX、GARCH-X、APARCH-X（见 [18]）。对于预测问题，结果表明，深度神经网络中的经验风险最小化原则的一致性结果以及泛化界可应用于该类模型。

本文其余部分安排如下。第 2 节介绍本文将考虑的深度神经网络类别。第 3 节给出一些符号与假设。第 4 节探讨深度神经网络预测器类别中经验风险最小化算法的一致性，以及泛化界及其收敛速率。第 5 节将其应用于一类弱依赖过程的二分类；第 6 节则考虑将其应用于具有外生协变量的仿射因果模型。第 7 节展示一些数值结果，第 8 节给出主要结果的证明。

## 2 深度神经网络

拟合深度神经网络需要选择一个激活函数 $\sigma: \mathbb{R} \to \mathbb{R}$，以及网络架构 $(L, \mathbf{p})$，其中 $L \in \mathbb{N}$ 表示隐藏层的数量，$\mathbf{p} = (p_0, \cdots, p_{L+1}) \in \mathbb{N}^{L+2}$ 是一个宽度向量。具有 $(L, \mathbf{p})$ 网络架构的深度神经网络，是指任意一个由以下形式定义的函数 $h$：

$$h : \mathbb{R}^{p_0} \to \mathbb{R}^{p_{L+1}},\ y \mapsto h(y) = A_{L+1} \circ \sigma_L \circ A_L \circ \sigma_{L-1} \circ \cdots \circ \sigma_1 \circ A_1(y), \tag{2.1}$$

其中 $A_j: \mathbb{R}^{p_{j-1}} \to \mathbb{R}^{p_j}$ 是由 $A_j(y) := W_j y + \mathbf{b}_j$，定义的线性仿射映射，其中给定权重矩阵 $p_{j-1} \times p_j$ $W_j$ 和偏移向量 $\mathbf{b}_j \in \mathbb{R}^{p_j}$，且 $\sigma_j: \mathbb{R}^{p_j} \to \mathbb{R}^{p_j}$ 是由 $\sigma_j(z) = (\sigma(z_1) \cdots, \sigma(z_{p_j}))'$ 定义的非线性激活映射。$p_0$、$p_{L+1}$ 分别是输入维度和输出维度。下文中，我们假设激活函数是 $C_\sigma$-利普希茨的，其中 $C_\sigma > 0$ 为某个常数；也就是说，对所有 $y_1, y_2 \in \mathbb{R}$，都有 $|\sigma(y_2) - \sigma(y_1)| \le C_\sigma |y_2 - y_1|$。对于按式 (2.1) 定义的深度神经网络 $h$，令

$$\theta(h) := \left(vec(W_1)', \mathbf{b}_1', \cdots, vec(W_{L+1})', \mathbf{b}_{L+1}'\right)', \tag{2.2}$$

其中 $vec(W)$ 通过拼接各列向量，将矩阵 $W$ 转换为相应的向量。

令 $\mathcal{H}_{\sigma,p_0,p_{L+1}}$ 表示激活函数为 $\sigma: \mathbb{R} \to \mathbb{R}$、将 $p_0$ 维输入映射为 $p_{L+1}$ 维输出的深度神经网络预测器类。由于过程 $\{Z_t = (X_t, Y_t)\ t \in \mathbb{Z}\}$ 取值于

$\mathcal{Z} = \mathcal{X} \times \mathcal{Y} \subset \mathbb{R}^d \times \mathbb{R}$，我们将处理满足 $p_0 = d$ 和 $p_{L+1} = 1$ 的类 $\mathcal{H}_{\sigma,p_0,p_{L+1}}$。对于深度神经网络 $h$，分别用 depth($h$) 和 width($h$) 表示 $h$ 的深度和宽度，也就是说，如果 $h$ 是一个架构为 $(L, \mathbf{p})$ 的网络，那么 depth($h$)=$L$，width($h$) $=\max_{1\le j\le L} p_j$。对于任意正常数 $L, N, B, F$ 和 $S$，我们定义

$$\mathcal{H}_\sigma(L, N, B) := \big\{h \in \mathcal{H}_{\sigma,d,1},\ \text{depth}(h) \le L, \text{width}(h) \le N, \|\theta(h)\|_\infty \le B\big\}, \tag{2.3}$$

$$\mathcal{H}_\sigma(L, N, B, F) := \big\{h \in H_\sigma(L, N, B),\ \|h\|_\infty \le F\big\}, \tag{2.4}$$

以及稀疏度为 $S > 0$ 的稀疏度约束深度神经网络类，由

$$\mathcal{H}_\sigma(L, N, B, F, S) := \big\{h \in H_\sigma(L, N, B, F),\ \|\theta(h)\|_0 \le S\big\}, \tag{2.5}$$

其中，$\|x\|_\infty = \max_{1\le i\le q} |x_i|$, $\|x\|_0 = \sum_{i=1}^q \mathbb{1}_{x_i\neq 0}$ 对所有 $x = (x_1, \dots, x_q)' \in \mathbb{R}^q$, $q \in \mathbb{N}$均成立，$\|h\|_\infty$ 表示函数 $h$ 的上确界范数（见下文）。在下文中，$\mathcal{H}_\sigma(L, N, B, F, S)$（其中 $L, N, B, F, S \ge 0$）表示式（2.4）中定义的深度神经网络类。

# 3 符号与假设

设 $E_1, E_2$ 是两个配备范数 $\|\cdot\|_{E_1}$ 和 $\|\cdot\|_{E_2}$ 的可分巴拿赫空间。下面先引入一些记号。

- 对于所有 $x \in \mathbb{R}$，$(x)_+ = \max(x, 0)$。
- 对于任意函数 $h\colon E_1 \to E_2$ 和 $U \subseteq E_1$,

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

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

- 对于任意 $\mathcal{K} > 0$ 和 $\alpha \in [0, 1]$, $\Lambda_{\alpha,\mathcal{K}}(E_1, E_2)$（当 $E_2 \subseteq \mathbb{R}$ 时，简记为 $\Lambda_{\alpha,\mathcal{K}}(E_1)$）表示函数集合$h$：$E_1^u \to E_2$ 对某个 $u \in \mathbb{N}$ 成立，且 $\text{Lip}_\alpha(h) \le \mathcal{K}$。当 $\alpha = 1$ 时，我们令 $\text{Lip}_1(h) = \text{Lip}(h)$ 且 $\Lambda_1(E_1) = \Lambda_{1,1}(E_1, \mathbb{R})$。
- $\mathcal{F}(E_1, E_2)$ 表示从 $E_1$ 到 $E_2$ 的可测函数集合。
- 对于任意 $h \in \mathcal{F}(E_1, E_2)$ 和 $\epsilon > 0$, $B(h, \epsilon)$ 表示以 $h$ 为中心、半径为 $\epsilon$ 的 $\mathcal{F}(E_1, E_2)$ 球，即 $B(h, \epsilon) = \big\{f \in \mathcal{F}(E_1, E_2)\ \ \|f - h\|_\infty \le \epsilon\big\}$。
- 对于任意 $\mathcal{H} \subset \mathcal{F}(E_1, E_2)$，$\mathcal{H}$ 的 $\epsilon$-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，

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

定义弱依赖结构，参见 [14] 和 [7]。考虑一个可分巴拿赫空间 $E$。

定义 3.1 *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* $s_1\le\cdots\le s_u\le s_u+r\le t_1\le\cdots\le t_v$, *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，采用 $\psi$ 的以下取值，可得到一些众所周知的弱依赖结构。

- $\psi(\mathrm{Lip}(g_1),\mathrm{Lip}(g_2)\ u,v)=v\mathrm{Lip}(g_2)$: $\theta$-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，记为 te $\epsilon(r)=\lambda(r)$。

接下来，对于上文 $\psi$ 的四种选择，分别设为:

$$\Psi(u,v)=2v,\ \Psi(u,v)=u+v,\ \Psi(u,v)=uv,\ \text{and}\ \Psi(u,v)=(u+v+uv)/2. \tag{3.1}$$

我们考虑取值于 $\mathcal{Z}=\mathcal{X}\times\mathcal{Y}\subset\mathbb{R}^d\times\mathbb{R}$ 的过程过程 $\{Z_t=(X_t,\ Y_t)\ t\in\mathbb{Z}\}$、损失函数 $\ell$: $\mathbb{R}\times\mathcal{Y}\to[0,\infty)$、激活函数为 $\sigma\colon\mathbb{R}\to\mathbb{R}$ 的深度神经网络类 $\mathcal{H}_\sigma(L,\ N,\ B,\ F,\ S)$，并作出以下假设。

**(A1)**: 存在常数 $C_\sigma>0$，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 $\ell\in\Lambda_{1,\ \mathcal{K}_\ell}(\mathbb{R}\times\mathcal{Y})$ 且 $M=\sup_{h\in\mathcal{H}_\sigma(L,\ N,\ B,\ F,\ S)}\sup_{z\in\mathcal{Z}}|\ell(h,\ z)|<\infty$。

在**(A2)**条件下，很容易看出,

$$G:=\sup_{h_1,h_2\in\mathcal{H}_\sigma(L,N,B,F,S),h_1\neq h_2}\sup_{z\in\mathcal{Z}}\frac{|\ell(h_1,z)-\ell(h_2,z)|}{||h_1-h_2||_\infty}<\infty. \tag{3.2}$$

现在我们来设定弱依赖假设。

**(A3)**: 令 $\Psi\colon[0,\infty)^2\times\mathbb{N}^2\to[0,\infty)$ 为 (3.1) 中的一个选项。过程 $\{Z_t=(X_t,Y_t)\ t\in\mathbb{Z}\}$ 是平稳遍历且 $(\Lambda_1(\mathcal{Z})\ \psi,\epsilon)$-弱依赖的，并且存在 $L_1,\ L_2,\ \mu\ge0$ 满足

$$\sum_{j\ge0}(j+1)^k\epsilon_j\le L_1L_2^k(k!)^\mu\ \text{for all}\ k\ge0. \tag{3.3}$$

假设 **(A1)–(A3)** 适用于许多经典模型。详情见第5节和第6节。

# 4 EMR算法的一致性与泛化界

## 4.1 EMR算法的一致性

下述命题给出了风险与其经验版本之间的集中不等式，该不等式在整个类别 $\mathcal{H}_\sigma(L,N,B,F,S)$ 上一致成立。本命题是 [13] 中定理 3.2 和定理 3.4 的一个应用。

**命题 4.1**

1. 假设条件 ***(A1)-(A3)*** 成立。对于所有 $\varepsilon>0$、 $n\in\mathbb{N}$，我们有

$$P\Big\{\sup_{h\in\mathcal{H}_\sigma(L,N,B,F,S)}[R(h)-\widehat{R}_n(h)]>\varepsilon\Big\}\leq\mathcal{N}\Big(\mathcal{H}_\sigma(L,N,B,F,S),\frac{\varepsilon}{4G}\Big)\exp\Big(-\frac{n^2\varepsilon^2/8}{A_n+B_n^{1/(\mu+2)}(n\varepsilon/2)^{(2\mu+3)/(\mu+2)}}\Big), \tag{4.1}$$

*for any real numbers* $A_n$ *and* $B_n$ *satisfying,* $A_n\geq\mathbb{E}\Big[\Big(\sum_{i=1}^n\Big(\ell(h(X_i),Y_i)-\mathbb{E}[\ell(h(X_0),Y_0)]\Big)\Big)^2\Big]$ *and* $B_n=2ML_2\max(\frac{2^{4+\mu}nM^2L_1}{A_n},1)$.

2. 假设 ***(A1)-(A2)*** 成立，并且 $(Z_t)_{t\in\mathbb{Z}}$ 是$(\Lambda_1(\mathcal{X}\times\mathcal{Y})\ \psi,\epsilon)$-弱依赖的，具有 $\epsilon_j=\mathcal{O}(j^{-2})$。对于任意 $\nu\in[0,1]$，当 $n$ 足够大时，对于所有 $\varepsilon>0$，我们有

$$P\left\{\sup_{h\in\mathcal{H}_\sigma(L,N,B,F,S)}[R(h)-\widehat{R}_n(h)]>\varepsilon\right\}\leq C_3\mathcal{N}\left(\mathcal{H}_\sigma(L,N,B,F,S),\frac{\varepsilon}{4G}\right)\exp\left(\log\log n-\frac{n^2\varepsilon^2/4}{A_n'+B_n'(n\varepsilon/2)^\nu}\right). \tag{4.2}$$

*with* $A_n'\geq\mathbb{E}\Big[\Big(\sum_{i=1}^n\Big(\ell(h(X_i),Y_i)-\mathbb{E}[\ell(h(X_0),Y_0)]\Big)\Big)^2\Big]$, $B_n'=\frac{n^{3/4}\log n}{A_n'}$ *and some constant* $C_3>0$.

如注记3.3中的情况以及例如上述命题 4.1 的第二种尝试，可以得到:

$$P\left\{\sup_{h\in\mathcal{H}_\sigma(L,N,B,F,S)}[\widehat{R}_n(h)-R(h)]>\varepsilon\right\}\leq C_3\mathcal{N}\left(\mathcal{H}_\sigma(L,N,B,F,S),\frac{\varepsilon}{4G}\right)\exp\left(\log\log n-\frac{n^2\varepsilon^2/4}{A_n'+B_n'(n\varepsilon/2)^\nu}\right),$$

并且,

$$P\left\{\sup_{h\in\mathcal{H}_\sigma(L,N,B,F,S)}|R(h)-\widehat{R}_n(h)|>\varepsilon\right\}\leq 2C_3\mathcal{N}\left(\mathcal{H}_\sigma(L,N,B,F,S),\frac{\varepsilon}{4G}\right)\exp\left(\log\log n-\frac{n^2\varepsilon^2/4}{A_n'+B_n'(n\varepsilon/2)^\nu}\right). \tag{4.3}$$

此外，在命题 4.1 的第二种尝试中，若对于某个 $\gamma>3$，有 $\epsilon_j=\mathcal{O}(j^{-\gamma})$，且根据 [21],，则对所有 $h\in\mathcal{H}_\sigma(L,N,B,F,S)$,

$$\mathbb{E}\Big[\Big(\sum_{i=1}^n\Big(\ell(h(X_i),Y_i)-\mathbb{E}[\ell(h(X_0),Y_0)]\Big)\Big)^2\Big]\leq Cn, \tag{4.4}$$

其中 $C>0$ 是一个常数。因此，通过选择 $A_n'=nC$，由（4.3）和 [31] 中的命题 1 可得:

$\widehat{\ }\sup|R(h)-R_n(h)|=o_P(1)$。这证明了在比 （**A3**） 更弱的条件下，EMR算法对于弱依赖过程，在 DNN 深度神经网络 $\mathcal{H}_\sigma(L,N,B,F,S)$ 的 $h\in\mathcal{H}_\sigma(L,N,B,F,S)$ 类中的一致性。

## 4.2 泛化界

下面我们来推导弱依赖条件下，基于深度神经网络预测器类 $\mathcal{H}_\sigma(L,N,B,F,S)$ 的 EMR算法泛化界。在下面的定理 4.2 中，记 $C_1=4M^2\Psi(1,1)L_1$和 $C_2=2ML_2\max(\frac{2^{3+\mu}}{\Psi(1,1)},1)$，其中 $\Psi$ 是在假设(**A3**)下，(3.1)中的一个函数。

**定理 4.2** 假设 ***(A1)-(A3)*** 成立。令 $\eta \in (0,\ 1)$ 且 $\alpha > 2$。假设

$$n \geq \max\left( n_0(L,S,\alpha,M,\mu,C_4), \left(\frac{C_4}{2M^2}(C_6 - \log(\eta) - 2L(S+1)\log(2M))_+\right)^{\frac{\alpha(\mu+2)}{\alpha-2}} \right)$$

其中 $C_4 = 4C_1 + 8C_2^{1/(\mu+2)} M^{(2\mu+3)/(\mu+2)}$、 $C_6 = 2L(S+1)\log(4GC_\sigma L(N+1)(B\vee 1))$ 以及 $n_0(L, S, \alpha, M, \mu, C_4) \in \mathbb{N}$定义于 ***(8.10)***。

(i) 以至少 $1-\eta$ 的概率，我们有

$$R(\widehat{h}_n) - \widehat{R}_n(\widehat{h}_n) \leq \varepsilon_1(n,\eta,\alpha), \tag{4.5}$$

其中，对于所有 $\alpha > 2$，$\varepsilon_1(n,\eta,\alpha) < \frac{2M_1}{n^{\alpha/(\mu+2)}}$。

(ii) 以至少 $1-2\eta$ 的概率，我们有

$$R(\widehat{h}_n) - R(h_{\mathcal{H}_\sigma(L,N,B,F,S)}) \leq \varepsilon_1(n,\eta,\alpha) + \varepsilon_1'(n,\eta), \tag{4.6}$$

其中

$$\varepsilon_1'(n,\eta) = \left[\frac{\log(1/\eta)}{C_{n,1}}\right]^{\mu+2} \text{ and } C_{n,1} = \frac{n^2}{4C_1 n + 8C_2^{1/(\mu+2)}(nM)^{(2\mu+3)/(\mu+2)}}.$$

式 (4.5) 中的界估计了 $R(\widehat{h}_n)$ 与 $R_n(\widehat{h}_n)$ 之间的差异，而式 (4.6) 中的界则衡量了风险 $R(\widehat{h}_n)$ 与目标神经网络在 $\mathcal{H}_\sigma(L,N,B,F,S)$ 下的风险有多接近。对于所有 $\alpha > 2$，式 (4.5) 和 (4.6) 中界的学习率均小于 $\mathcal{O}(n^{-\alpha/(\mu+2)})$。下面的定理在比~~(A3)~~ 更弱的条件下，给出了更快的收敛率。

**定理 4.3** 假设 ***(A1)-(A2)*** 成立，并且$(Z_t)_{t\in\mathbb{Z}}$是$(\Lambda_1(\mathcal{X},\mathcal{Y})\ \psi,\epsilon)-$ 弱依赖的，且满足$\epsilon_j = \mathcal{O}(j^{-\gamma})$，其中某个 $\gamma > 3.$。令 $\eta \in (0,1)$、 $\nu \in (0,\ 1)$，并且 $\alpha > 2$。假设

$$n > \left(\frac{C_5}{2M^2}(C_6 - \log(\eta/C_3) - 2L(S+1)\log(2M))_+\right)^{\alpha/(\alpha-2)} \tag{4.7}$$

其中 $C_3$ 在 ***(4.2)*** 中给出， $C_5 = 4(C + M^\nu/C)$，常数 $C$ 在 ***(4.4)*** 中定义。

(i) 当 ***n*** 足够大时，以至少 $1-\eta$ 的概率，有

$$R(\widehat{h}_n) - \widehat{R}_n(\widehat{h}_n) \leq \varepsilon_2(n,\eta,\nu,\alpha), \tag{4.8}$$

其中 $\varepsilon_2(n,\eta,\nu,\alpha) < \frac{2M_1}{n^\alpha}$，对所有 $\alpha > 2$ 均成立。

(ii) 当 ***n*** 足够大时，以至少 $1-2\eta$ 的概率，有

$$R(\widehat{h}_n) - R(h_{\mathcal{H}}) \leq \varepsilon_2(n,\eta,\nu,\alpha) + \varepsilon_2'(n,\eta,\nu), \tag{4.9}$$

其中

$$\varepsilon_2'(n,\eta,\nu) = \left[\frac{\log(C_3\log n/\eta)}{C_{n,2}'}\right]^{\frac{1}{2}} \text{ and } C_{n,2}' = \frac{n^2}{nC + \log n\, n^{\nu-1/4}(2M)^\nu/C}.$$

界 (4.8) 和 (4.9) 中的学习率对所有 $\alpha > 2$ 均小于 $\mathcal{O}(n^{-1/\alpha})$。

**注 4.4** 在 *[13]* 中推导出的泛化界在此处不适用，因为类别 $\mathcal{H}_\sigma(L,N,B,F,S)$通常不满足这些作者提出的条件 ***(A4)*** ；除非 $\mathcal{X}$ 是紧致的，且我们讨论的是具有 $s=1$ 的 ***H¨older*** 空间 $\mathcal{C}^s$。在这种情况下， *[13]* 中得到的收敛速率为 $\mathcal{O}(n^{-1/(2+2d)})$。因此，上文得到的速率更优，且不依赖于输入维度。

# 5 二分类的应用

## 5.1 弱依赖过程的二分类

设 $(X_1,\ Y_1)$ ...，$(X_n,\ Y_n)$ 是平稳遍历过程 $Z_t = \{(X_t,\ Y_t)\ t \in \mathbb{Z}\}$ 的一条轨迹，其中 $X_t \in \mathcal{X} \subset \mathbb{R}^d$ 是输入向量，$Y_t \in \{-1, 1\}$ 是类别标签。我们关注分类器 $h:\ \mathcal{X} \to \mathbb{R}$，它根据 $\text{sign}(h(X_t))$ 预测 $Y_t \in \mathcal{Y} = \{-1, 1\}$，其中对所有 $x \in \mathbb{R}$，$\text{sign}(x) = \mathbb{1}_{\{x \geq 0\}} - \mathbb{1}_{\{x<0\}}$。目标是构造一个基于 0-1损失、使分类风险最小化的预测函数，

$$R_{01}(h) := P\big(\text{sign}(h(X_0) \neq Y_0)\big) = E\big[\mathbb{1}_{\{Y_0 \cdot \text{sign}(h(X_0)) \leq 0\}}\big]. \tag{5.1}$$

但是，由于 0-1损失的非凸性，式 (1.3) 中采用 0-1损失的优化问题在计算上非常困难。一种替代方法是使用基于间隔的替代损失 $\phi:\ \mathbb{R} \to [0, \infty)$，并着重于最小化替代风险，其定义如下：

$$R_\phi(h) = E[\phi(Y_0 h(X_0))]. \tag{5.2}$$

替代经验风险定义如下：

$$\widehat{R}_{n,\phi}(h) = \frac{1}{n}\sum_{i=1}^{n} \phi(Y_i h(X_i))$$

基于该风险的深度神经网络估计量定义如下：

$$\widehat{h}_{n,\phi} = \underset{h \in \mathcal{H}_\sigma(L,N,B,F,S)}{\text{argmin}}\ \widehat{R}_{n,\phi}(h). \tag{5.3}$$

注意，如果替代损失 $\phi$ 是凸且利普希茨的，那么损失函数 $\ell(u, y) = \phi(uy)$也具有这些性质。在这种情况下，一个广泛使用的例子是合页损失，其定义为 $\phi(z) = \max(1 - z, 0)$，它是凸且利普希茨的。

## 5.2 二元时间序列预测示例

设 $(Y_1, \mathcal{X}_1)\ \cdots$、$(Y_n, \mathcal{X}_n)$ 是平稳遍历过程 $(Y_t, \mathcal{X}_t)_{t\in\mathbb{Z}}$ 的一条轨迹，其中 $(\mathcal{X}_t)_{t\in\mathbb{Z}}$ 是一个取值于 $\mathbb{R}^{d_x}$，$d_x \in \mathbb{N}$ 的协变量过程。目标是根据观测值 $(Y_n, \mathcal{X}_n)\ \cdots$、$(Y_{n-p+1}, \mathcal{X}_{n-p+1})$，在 $p \in \mathbb{N}$ 的条件下预测 $Y_{n+1}$。我们使用上文构建的深度神经网络函数进行学习理论分析，其中 $X_t = ((Y_{t-1}, \mathcal{X}_{t-1})\ \cdots, (Y_{t-p}, \mathcal{X}_{t-p}))$ $\mathcal{Y} = \{-1, 1\}$，$\mathcal{X} \subset (\mathbb{R} \times \mathbb{R}^{d_x})^p$，并采用基于间隔的损失函数 $\ell:\ \mathbb{R} \times \mathcal{Y} \to [0, \infty)$，该函数由合页损失 $\phi$定义。我们重点研究深度神经网络预测器类$\mathcal{H}_\sigma(L, N, B, F, S)$，其中 $L, N, B, F, S \geq 0$。

记 $\mathcal{F}_{t-1} = \sigma\{Y_{t-1}, \cdots;\ \mathcal{X}_{t-1},\ \cdots\}$ 为时刻 $t-1$ 整个过去所生成的 $\sigma$–场，并假设，

$$Y_t|\mathcal{F}_{t-1} \sim 2\mathcal{B}(p_t) - 1 \text{ with } 2p_t - 1 = E[Y_t|\mathcal{F}_{t-1}] = f(Y_{t-1}, Y_{t-2}, \cdots; \mathcal{X}_{t-1}, \mathcal{X}_{t-2}, \cdots), \tag{5.4}$$

其中，$f$是一个取值于 $[-1$的可测非负函数；$1]$、$p_t = P(Y_t = 1|\mathcal{F}_{t-1})$和 $\mathcal{B}(p_t)$是参数为 $p_t$的伯努利分布。采用替代损失函数 $\phi$ 的EMR算法所得到的预测函数为：

$$\widehat{h}_{n,\phi} = \underset{h \in \mathcal{H}_\sigma(L,N,B,F,S)}{\text{argmin}}\ \widehat{R}_{n,\phi}(h).$$

$\widehat{\quad}$ $\widehat{\quad}$ $\widehat{\quad}$ $\widehat{\quad}$ $\widehat{\quad}$对于任意 $t \in \mathbb{Z}$，基于$h_{n,\ \phi}$ 的 $Y_t$的预测为 $Y_t = \text{sign}(h_{n,\ \phi}(X_t))$，因此，$Y_{n+1} = \text{sign}(h_{n,\ \phi}(X_{n+1}))$。现在，我们对协变量施加自回归型结构：

$$\mathcal{X}_t = g(\mathcal{X}_{t-1}, \mathcal{X}_{t-2}, \cdots; \eta_t), \tag{5.5}$$

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 $g(x;\eta)$ 是一个取值于 $\mathbb{R}^{d_x}$的可测函数，满足

$$\mathbb{E}[\|g(0;\eta_0)\|^r] < \infty \text{ and } ||g(x;\eta_0 - g(x';\eta_0))||_r \leq \sum_{k=1}^{\infty} \alpha_k(g)||x_k - x'_k|| \text{ for all } x, x' \in (\mathbb{R}^{d_x})^\infty, \tag{5.6}$$

对于某个 $r \geq 1$，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 $U$，$||U||_r := (\mathbb{E}||U||^r)^{1/r}$。我们考虑模型 (5.4)、(5.5)、(5.6)，并作出以下假设。

- 对于任意 $(y,x) \in \mathbb{R}^\infty \times (\mathbb{R}^{d_x})^\infty$，$|f(y;\ x)| < 1$，且存在两个非负实数序列 $(\alpha_{k,Y}(f))_{k\geq 1}$ 和 $(\alpha_{k,\mathcal{X}}(f))_{k\geq 1}$，满足 $\sum_{k=1}^{\infty} \alpha_{k,\ Y}(f) < 1$ 以及 $\sum_{k=1}^{\infty} \alpha_{k,\ \mathcal{X}}(f) < \infty$；使得对于所有 $(y,x)$、$(y',x') \in \mathbb{R}^\infty \times (\mathbb{R}^{d_x})^\infty$，

$$|f(y;x) - f(y';x')| \leq \sum_{k=1}^{\infty} \alpha_{k,Y}(f)|y_k - y'_k| + \sum_{k=1}^{\infty} \alpha_{k,\mathcal{X}}(f)\|x_k - x'_k\|, \tag{5.7}$$

  其中，$\|\cdot\|$ 表示 $\mathbb{R}^{d_x}$ 中的任意向量范数。

- (5.6) 和 (5.7) 成立，其中，

$$\sum_{k=1}^{\infty} \max\{\alpha_k(g), \alpha_{k,Y}(f)\} < 1. \tag{5.8}$$

由于分布 $2\mathcal{B}(p) - 1$ 属于单参数指数族，与 [29], 中命题 3.1 的证明相同，我们可以证明，存在一个 $\tau$-弱依赖的平稳遍历解 $(Y_t,\ \mathcal{X}_t)_{t\in\mathbb{Z}}$，满足 (5.4) 和 $\|(Y_0,\ \mathcal{X}_0)\|_r < \infty$。该 $\tau$-弱依赖结构蕴含 $\eta$，$\theta$-弱依赖。事实上，对于所有 $j \geq 0$，都有 $\eta(j) \leq \theta(j) \leq \tau(j)$，见 [7]。因此，存在一个解 $(Y_t,\ \mathcal{X}_t)_{t\in\mathbb{Z}}$，它是 $\theta$-弱依赖的。此外，由 [16], 可得 $j \to \infty$,

$$\theta(j) \leq \tau(j) = \mathcal{O}\left(\inf_{1\leq\iota\leq j}\left\{\alpha^{j/\iota} + \sum_{k\geq\iota+1}\alpha_k\right\}\right), \tag{5.9}$$

其中 $\alpha_k = \max\{\alpha_k(g)\, \alpha_{k,\ Y}(f)\}$ 以及 $\alpha = \sum_{k\geq 1}\alpha_k$。

我们假设激活函数 $\sigma$ 是利普希茨连续的；例如，ReLU（修正线性单元）：$\sigma(x) = \max(0,\ x)$，或 S形函数：$\sigma(x) = 1/(1 + e^{-x})$。下面检查类模型（5.4）是否满足定理 4.2 和定理 4.3 的其他假设。

（i）如果替代损失函数是利普希茨连续且有界的，那么（**A2）</style}成立。例如，合页损失() max(0,** $1 - z$))满足这些条件。

(ii) 在条件 (5.8) 下，过程 $(Y_t, \mathcal{X}_t)_{t\in\mathbb{Z}}$ 是 $\theta$-弱依赖的，其系数 $\theta(j)$ 满足 (5.9) 中的界。容易得出，$(X_t, Y_t)_{t\in\mathcal{Z}}$ 也是 $\theta$-弱依赖的，其系数为 $\theta(j)$。

我们来考虑以下几种情形。

- 几何情形。假设,

$$\alpha_{k,Y}(f)+\alpha_k(g)=\mathcal{O}(a^k) \text{ for some } a\in[0,1).$$

由 (5.9) 可容易得到（参见 [16]） $\theta(j) \le \tau(j) = \mathcal{O}\ddot{A}\exp(-\sqrt{\log(\alpha)\log(a)j})\ddot{a}$ ，其中 $\alpha$ 由 (5.9) 给出。因此，(3.3) 在 $\mu=2$ 条件下成立（另见 [15]中的命题 8）。因此，定理 4.2 中的条件 (**A3**) 成立。

- 黎曼情形。假设,

$$\alpha_{k,Y}(f)+\alpha_k(g)=\mathcal{O}(k^{-\gamma}) \text{ for some } \gamma>1.$$

由 (5.9) 可容易得到（另见 [16]） $\theta(j) \le \tau(j) = \mathcal{O}\left(\ddot{A}_{\frac{\log j}{j}}\ddot{a}^{\gamma-1}\right)$. 因此，若 $\gamma>3$，则满足定理 4.3 中的条件 $\epsilon_j=\mathcal{O}(j^{-2})$。

# 6 仿射因果模型在外生协变量中的应用

令 $(\mathcal{X}_t)_{t\in\mathbb{Z}}$ 为取值于 $\mathbb{R}^{d_x}, d_x\in\mathbb{N}$ 的协变量过程。我们考虑具有外生协变量的仿射因果模型类（见 [12]），其定义如下:

**类** $\mathcal{AC}-X(\mathcal{M},f)$: 若过程 $\{Y_t, t\in\mathbb{Z}\}$ 满足以下条件，则称其属于 $\mathcal{AC}-X(\mathcal{M},f)$:

$$Y_t=\mathcal{M}(Y_{t-1},Y_{t-2},\cdots;\mathcal{X}_{t-1},\mathcal{X}_{t-2},\cdots)\xi_t+f(Y_{t-1},Y_{t-2},\cdots;\mathcal{X}_{t-1},\mathcal{X}_{t-2},\cdots), \tag{6.1}$$

其中 $\mathcal{M}, f$: $\mathbb{R}^{\mathbb{N}}\times(\mathbb{R}^{d_x})^{\mathbb{N}}\to\mathbb{R}$ 是两个可测函数，$(\xi_t)_{t\in\mathbb{Z}}$ 是一个中心化独立同分布随机变量序列，满足 $\mathbb{E}(\xi_0^r)<\infty$，其中某个 $r\ge 2$ 满足该条件，且 $\mathbb{E}(\xi_0^2)=1$。类 $\mathcal{AC}-X(\mathcal{M},f)$ 包含 [3], [2], [23], [1], [22] 研究的、不含外生协变量的经典仿射因果模型；该模型可由 (6.1) 在某个常数 $C$ 满足 $\mathcal{X}_t\equiv C$ 时得到。ARMAX、TARX、GARCH-X、ARMAX-GARCH、APARCH-X 等知名模型（见 [18]）均属于类 $\mathcal{AC}-X(\mathcal{M},f)$。[12] 已在半参数设定下，基于拟似然估计量，对类 $\mathcal{AC}-X(\mathcal{M},f)$ 中的推断进行了研究。

为了研究模型 (6.1) 的稳定性， [12]在函数 $f,\mathcal{M}$ or $\mathcal{M}^2$ 上设定以下 Lipschitz 型条件。本节中，0 表示任意向量空间的零向量。对于 $\Psi=f$ or$\mathcal{M}$，考虑以下假设。

假设 $A(\Psi)$: $|\Psi(0;0)|<\infty$，且存在两个非负实数序列 $(\alpha_{k,Y}(\Psi))_{k\ge1}$和 $(\alpha_{k,\mathcal{X}}(\Psi))_{k\ge1}$，满足 $\sum_{k=1}^{\infty}\alpha_{k,Y}(\Psi)<\infty$, $\sum_{k=1}^{\infty}\alpha_{k,\mathcal{X}}(\Psi)<\infty$；使得对于任意 $(y,x)$、$(y',x')\in\mathbb{R}^{\infty}\times(\mathbb{R}^{d_x})^{\infty}$,

$$|\Psi(y;x)-\Psi(y';x')|\le\sum_{k=1}^{\infty}\alpha_{k,Y}(\Psi)|y_k-y'_k|+\sum_{k=1}^{\infty}\alpha_{k,\mathcal{X}}(\Psi)\|x_k-x'_k\|.$$

对于 ARCH-X 型模型，下一条假设针对 $H=\mathcal{M}^2$ 设定。

假设 $A(H)$: 假设 $f=0$, $|\mathcal{M}(0;0)|<\infty$，且存在两个非负实数序列 $(\alpha_{k,Y}(H))_{k\ge1}$ 和 $(\alpha_{k,\mathcal{X}}(H))_{k\ge1}$，满足 $\sum_{k=1}^{\infty}\alpha_{k,Y}(H)<\infty$, $\sum_{k=1}^{\infty}\alpha_{k,\mathcal{X}}(H)<\infty$；使得对于任意 $(y,x)$,$(y',x')\in\mathbb{R}^{\infty}\times(\mathbb{R}^{d_x})^{\infty}$,

$$|H(y;x)-H(y';x')|\le\sum_{k=1}^{\infty}\alpha_{k,Y}(H)|y_k^2-{y'}_k^2|+\sum_{k=1}^{\infty}\alpha_{k,\mathcal{X}}(H)\|x_k-x'_k\|.$$

为后文起见，约定若**A**($\mathcal{M}$)成立，则对所有 $k \geq 1$ 均有$\alpha_{k,Y}(H) = \alpha_{k,\mathcal{X}}(H) = 0$；若**A**($H$)成立，则对所有 $k \geq 1$ 均有$\alpha_{k,Y}(\mathcal{M}) = \alpha_{k,\mathcal{X}}(\mathcal{M}) = 0$。我们考虑满足协变量条件 (5.5)、(5.6) 的模型 (6.1)，其中 (5.5) 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；并假设对于某个 $r \geq 1$,

$$\sum_{k=1}^{\infty} \max\left\{\alpha_k(g), \alpha_{k,Y}(f) + \|\xi_0\|_r \alpha_{k,Y}(\mathcal{M}) + \|\xi_0\|_r^2 \alpha_{k,Y}(H)\right\} < 1. \tag{6.2}$$

在假设 **A**($f$)、**A**($\mathcal{M}$) 或 **A**($\mathcal{H}$) 以及 (6.2) 下，存在一个 $\tau -$ 弱依赖、平稳、遍历且非自适应的解 $(Y_t, \mathcal{X})_{t\in\mathbb{Z}}$，满足 (6.1) 和 $||(Y_0, \mathcal{X}_0)||_r < \infty$（见 [12]）。该解是$\theta$ –弱依赖的，其系数 $\theta(j)$ 的界如 (5.9) 所示，其中

$$\alpha_k = \max\left\{\alpha_k(g), \alpha_{k,Y}(f) + \|\xi_0\|_r \alpha_{k,Y}(\mathcal{M}) + ||\xi_0||_r^2 \alpha_{k,Y}(H)\right\} < 1 \text{ and } \alpha = \sum_{k\geq 1} \alpha_k.$$

考虑基于观测值 $(Y_1, \mathcal{X}_1) \cdots, (Y_n, \mathcal{X}_n)$ 对模型 (6.1) 进行预测的问题，这些观测值构成平稳遍历过程 $(Y_t, \mathcal{X}_t)_{t\in\mathbb{Z}}$ 的一条轨迹，该过程满足 (6.1) 和 (5.5)。目标是根据观测值 $(Y_n, \mathcal{X}_n) \cdots, (Y_{n-p+1}, \mathcal{X}_{n-p+1})$，对 $Y_{n+1}$ 进行预测，其中 $p \in \mathbb{N}$。上述针对深度神经网络预测器所发展的学习理论，采用 $X_t = ((Y_{t-1}, \mathcal{X}_{t-1}) \cdots, (Y_{t-p}, \mathcal{X}_{t-p}))$ $\mathcal{Y} \subset \mathbb{R}$, $\mathcal{X} \subset (\mathbb{R} \times \mathbb{R}^{d_x})^p$ 和损失函数 $\ell\colon \mathbb{R} \times \mathcal{Y} \to [0, \infty)$。我们关注深度神经网络估计器类 $\mathcal{H}_\sigma(L, N, B, F, S)$，其中$L, N, B, F, S \geq 0$ 为某些参数。通过经验风险最小化算法得到的预测函数定义如下：

$$\widehat{h}_n = \underset{h\in\mathcal{H}_\sigma(L,N,B,F,S)}{\operatorname{argmin}} \widehat{R}_n(h), \text{ with } \widehat{R}_n(h) = \frac{1}{n-p} \sum_{i=p+1}^{n} \ell(h(Y_{i-1}, \cdots, Y_{i-p}), Y_i). \tag{6.3}$$

我们假设激活函数 $\sigma$ 是利普希茨连续的；例如 ReLU 或 S形函数。我们来检查定理 4.2 和定理 4.3 对类模型 $\mathcal{AC}-X(\mathcal{M}, f)$ 的其他假设。

(i) 在条件 (6.2) 下，如果向量 $(\eta_0, \xi_0)$ 有界 $a.s.$，那么可以证明（例如参见 [13]） $Y_0$ 是$a.s.$ 有界的，且 $\mathcal{Y}$ 可以取为有界。因此，例如对于平方损失，(**A2**) 成立。

(ii) 在假设 **A**($f$)、**A**($\mathcal{M}$) 或 **A**($\mathcal{H}$) 以及 (6.2) 下，过程 $(Y_t, \mathcal{X}_t)_{t\in\mathbb{Z}}$ 是 $\theta$-弱依赖的，其系数 $\theta(j)$ 按照 (5.9) 有界。可以得到与第 5.2 小节类似的条件，具体如下：

- 几何情形下，

$$\alpha_{k,Y}(f) + \alpha_{k,Y}(\mathcal{M}) + \alpha_{k,Y}(H) + \alpha_k(g) = \mathcal{O}(a^k) \text{ for some } a \in [0, 1),$$

以及：

- 黎曼情形下，

$$\alpha_{k,Y}(f) + \alpha_{k,Y}(\mathcal{M}) + \alpha_{k,Y}(H) + \alpha_k(g) = \mathcal{O}(k^{-\gamma}) \text{ for some } \gamma > 1.$$

因此，如果激活函数和损失函数是利普希茨连续的，并且创新项（$\eta_t$, $\xi_t$）$_{t\in\mathbb{Z}}$ 是$a.s.$ 有界的，那么，用于 ARMAX、TARX、GARCH-X、APARCH-X、ARMAX-GARCH、typemodels 中预测问题的 EMR算法，在深度神经网络类上是一致的。

# 7 数值结果

本节考虑利用深度神经网络预测二元时间序列。

## 7.1 模拟研究

我们考虑一个二元 INGARCHX-X 过程（$Y_t$， $\mathcal{X}_t$）$_{t\in\mathbb{Z}}$，其取值于 $\{-1,1\}\times\mathbb{R}$ 并满足

$$Y_t|\mathcal{F}_{t-1}\sim 2\mathcal{B}(p_t)-1 \text{ with } 2p_t-1=E[Y_t|\mathcal{F}_{t-1}]=f(Y_{t-1},Y_{t-2},\cdots;\mathcal{X}_{t-1},\mathcal{X}_{t-2},\cdots), \tag{7.1}$$

其中 $\mathcal{F}_{t-1}=\sigma\{Y_{t-1},\cdots;\ \mathcal{X}_{t-1},\ \cdots\}$ 且 $f$取值于 $[-1;\ 1]$。基于观测值（$Y_1$， $\mathcal{X}_1$） $\cdots$，（$Y_n$， $\mathcal{X}_n$），我们的目标是利用第 5.2 和 5.1 小节所述的深度神经网络估计量来预测 $Y_{n+1}$。在下面的 DGP1 中，该过程取 $p=1$、$X_t=Y_{t-1}$；在 DGP2 中取 $p=2$、 $X_t=$（$Y_{t-1}$， $Y_{t-2}$， $\mathcal{X}_{t-1}$）。我们考虑（7.1）中的以下情形:

$$\begin{aligned}
&\text{DGP1}: \quad f(Y_{t-1},\ldots;\mathcal{X}_{t-1},\ldots)=-0.25+0.6Y_{t-1}\\
&\text{DGP2}: \quad f(Y_{t-1},\ldots;\mathcal{X}_{t-1},\ldots)=0.1-0.15\max(Y_{t-1},0)+0.25\min(Y_{t-1},0)+0.15Y_{t-2}+0.2\frac{1}{1+\mathcal{X}_{t-1}^2},
\end{aligned}$$

其中（$\mathcal{X}_t$）$_{t\in\mathbb{Z}}$ 是一阶自回归过程。DGP1 与下文小节中的真实数据示例相关。在下文中， $h_0$ 表示由真实数据生成过程得到的预测函数；即,

$$h_0(X_t)=\text{sign}(f(X_t)) \text{ for all } t\in\mathbb{Z}. \tag{7.2}$$

注意， $h_0$是根据 $0-1$ 损失函数定义的贝叶斯分类器；也就是说， $R_{01}(h_0)=\inf_{h\in\mathcal{F}(\mathcal{X},\mathcal{Y})}R_{01}(h)$，其中 $R_{0,1}$ 在 (5.1) 中定义，而 $\mathcal{F}(\mathcal{X},\ \mathcal{Y})$ 是从 $\mathcal{X}$到 $\mathcal{Y}$的可测函数集合。

对于这两个数据生成过程，我们采用了包含 2 个隐藏层的网络架构，每层有 16 个隐藏节点，并使用 ReLU 激活函数（$\sigma(x)=\max(x,\ 0)$）。网络权重在 R 包 Keras 中使用 Adam 算法（[25]）进行训练，学习率为 $10^{-3}$，小批量大小为 32。当准确率连续 30 个训练轮次未提升时，我们停止训练。使用合页损失（$\phi(z)=\max(1-z,\ 0)$），并将 $tanh$ 激活函数用于输出层。如果输出为负，则预测为 -1；否则预测为 1。用 $\mathcal{H}_\sigma$ 表示所考虑的网络架构集合，用 $\Theta$ 表示其参数空间。

首先，对于所考虑的每个数据生成过程，生成一个样本 $(Y_0',\ \mathcal{X}_0')$，$(Y_1',\ \mathcal{X}_1')$，$(Y_2',\ \mathcal{X}_2')\ldots,\ (Y_m',\ \mathcal{X}_m')$，其中 $m=10^4$，并通过 $h$ 估计目标网络 $h_{\mathcal{H}_\sigma}$（见 (1.2)），其中,

$$\theta(\widetilde{h})=\underset{\theta(h)\in\Theta}{\text{argmin}}\widetilde{R}_{\phi,1}(h), \text{ with } \widetilde{R}_{\phi,1}(h):=\frac{1}{m}\sum_{i=1}^{m}\max\left(1-Y_i'h(X_i'),0\right) \text{ and } X_i'=(Y_{i-1}',\mathcal{X}_{i-1}').$$

~ ~ ~因此， $R_\phi(h_{\mathcal{H}_\sigma})$ 的估计值为 $R_{\phi,\ 1}(h)$。此外，$h_0$ 的风险（见 (7.2)）估计为 $R_{\phi,\ 1}(h_0)$。其次，对于 $n=200,\ 220,\ 240,\ \ldots,\ 2000$，从真实数据生成过程生成一条轨迹 $((Y_1,\ \chi_1),\ (Y_2,\ \mathcal{X}_2)\ldots,\ (Y_n,\ \mathcal{X}_n))$。$h_{n,\ \phi}$ 通过以下方式估计

$$\theta(\widehat{h}_{n,\phi})=\underset{\theta(h)\in\Theta}{\text{argmin}}\widehat{R}_{n,\phi}(h).$$

^ ~ ^因此， $R(h_{n,\phi})$ 的估计值为 $R_2(h_{n,\phi})$，其中,

$$\widetilde{R}_{\phi,2}(h)=\frac{1}{n}\sum_{i=1}^{n}\max\left(1-Y_i''h(X_i''),0\right) \text{ for } \theta(h)\in\Theta \text{ and } X_i''=(Y_{i-1}'',\mathcal{X}_{t-1}'),$$

其中，$((Y_0'', \mathcal{X}_0''), (Y_1'', \mathcal{X}_1''), (Y_2'', \mathcal{X}_2'') \ldots, (Y_n'', \mathcal{X}_n''))$ 是由真实的数据生成过程生成的，并且独立于用于估计$h_{n,\phi}$的$\widehat{}$样本。对于每个固定的 $n$， $R(h_{n,\phi})$ 的蒙特卡罗估计基于 500 次重复。图1展示了数据生成过程1和数据生成过程2中点 $\left(n, R_\phi(\widehat{h}_{n,\phi}) - R_\phi(h_{\mathcal{H}_\sigma})\right)$ 以及 $\left(n, R_\phi(\widehat{h}_{n,\phi}) - R_\phi(h_0)\right)$的曲线。

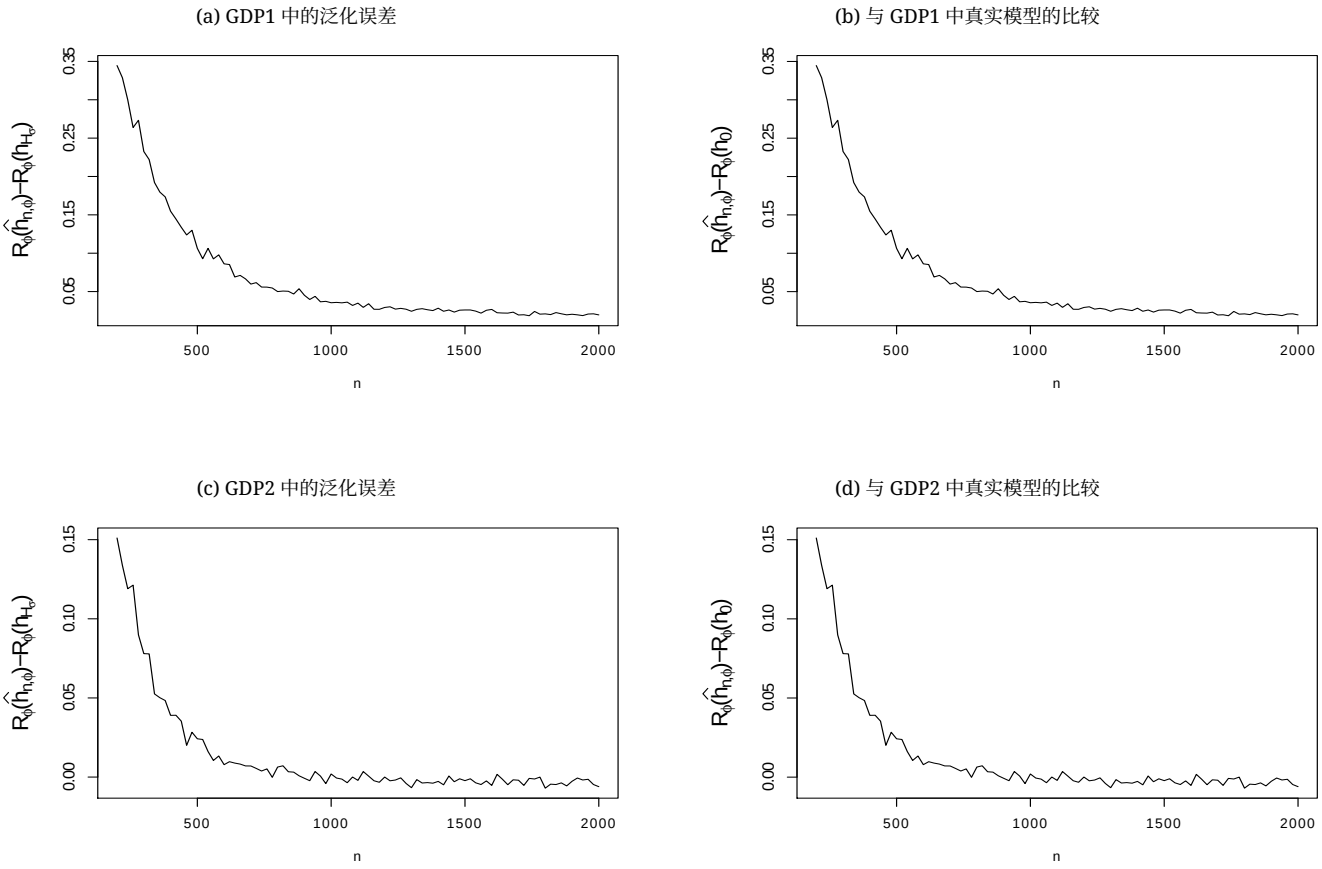


图1： $\left(n、\ R_\phi(\widehat{h}_{n,\ \phi}) - R_\phi(h_{\mathcal{H}_\sigma})\right)$ 和 $\left(n、\ R_\phi(\widehat{h}_{n,\ \phi}) - R_\phi(h_0)\right)$，其中 $n$ 的取值范围为 *200* 至 *2000*，数据生成过程为 *GDP1*（*a*）、（*b*）和 *GDP2*（*c*）、（*d*）。

$\widehat{}$对于 DGP1 和数据生成过程2，可以看到， $R_\phi(h_{n,\ \phi}) - R_\phi(h_{\mathcal{H}_\sigma})$ 随着样本量 $n$ 增加而趋近于零。这些数值结果与深度神经网络预测器类中经验风险最小化算法的一致性，以及定理 4.2 和 4.3 中的泛化界相符。还可以看到$\widehat{}$，随着 $n$ 增加，神经网络预测函数$h_{n,\ \phi}$ 的风险趋近于由真实模型得到的预测函数 $h_0$ 的风险。

## 7.2 美国经济衰退数据的应用

我们考察美国 1933—2022 年的季度衰退数据序列，共有 360 个观测值。这些数据可从 https://fred.stlouisfed.org/series/USRECQ 获取，代表一个二元变量：若该季度至少有一个月处于衰退期，则取值为 1；否则取值为 0。我们已经

将原始数据重新编码为：若该季度至少有一个月处于衰退期，则取值为 1；否则取值为 $-1$ ，见图2。

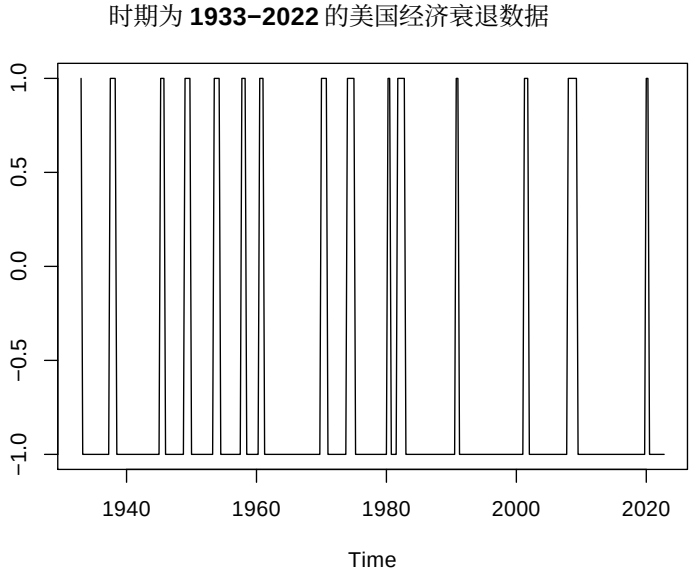


图2： 美国 *1933—2022* 年的季度衰退数据。

已有多位作者分析过这些数据，例如 [20], [8], [10]。后面这两项研究发现，一阶自回归模型能够很好地拟合这些数据。也就是说，模型（5.4），其中

$$f(Y_{t-1}) = \alpha_0 + \alpha_1 Y_{t-1},$$

其中 $|\alpha_0|+|\alpha_1| < 1$。参数 $\theta = (\alpha_0,\ \alpha_1)\widehat{}$的最大似然估计（例如参见 [6], [9]）为 $\theta = (-0.248,\ 0。660)$。这个真实数据示例与上文小节中的 DGP1 类似。对于预测问题，我们采用上一小节中的网络架构（2 个隐藏层，每层 16 个隐藏节点，并使用 ReLU 激活函数），输入变量为 $X_t = Y_{t-1}$。网络使用前一半数据进行训练，后一半数据（作为测试数据）用于评估其风险和准确率。表 1 给出了根据测试数据得到的混淆矩阵。风险（相对于合页损失

| | | 预测值 | |
|---|---|---|---|
| | | -1 | 1 |
| 实际值 | -1 | 153 | 6 |
| | 1 | 5 | 15 |

表 1： *Confusion matrix obtained from the test data.*

）准确率分别为 0.2445 和 0.9385（注意，在 DGP1 中，目标神经网络 $h_{\mathcal{H}_\sigma}$ 和预测函数 $h_0$ 在 (7.2) 处的风险分别为 0.2288979 和 0.2288229）。还可以看到，经济衰退期的预测准确率为 0.75。这些结果表明，神经网络

总体上适用于这些数据。因此，如果网络经过良好训练，仅凭本季度的数据，就能以相当令人满意的准确率预测下一季度是否会进入经济衰退期。

# 8 主要结果的证明

## 8.1 命题 4.1 的证明

考虑一类深度神经网络预测器 $\mathcal{H}_\sigma(L,N,B,F,S)$，其中 $L,N,B,F,S\geq 0$。只需证明，对于类 $h\in\mathcal{H}_\sigma(L,N,B,F,S)$、损失函数 $\ell:\ \mathbb{R}\times\mathcal{Y}\to[0,\infty)$ 以及过程 $\{Z_t=(X_t,Y_t)\ t\in\mathbb{Z}\}$，[13] 中的定理 3.2 和 3.4 的条件均得到满足。根据集合 $\mathcal{H}_\sigma(L,N,B,F,S)$ 的定义，有：

上确界 $\|h\|_\infty\leq F<\infty$。因此，根据假设 **(A2)** 和 **(A3)**，还需证明$h\in\mathcal{H}_\sigma(L,\ N,\ B,\ F,\ S)$

即存在 $\mathcal{K}_{\mathcal{H}_\sigma}>$，使得对于所有 $h\in\mathcal{H}_\sigma$（$L,\ N,\ B,\ F,\ S$），$h$ 是 $\mathcal{K}_{\mathcal{H}_\sigma}$-利普希茨的。

令 $h\in\mathcal{H}_\sigma$（$L,\ N,\ B,\ F,\ S$），对于所有 $x,\ x'\in\mathbb{R}^d$，由 （**A1**），我们有

$$\begin{aligned}\|h(x)-h(x')\| &= \|A_{L+1}\circ\sigma_L\circ A_L\circ\sigma_{L-1}\circ\cdots\circ\sigma_1\circ A_1(x)-A_{L+1}\circ\sigma_L\circ A_L\circ\sigma_{L-1}\circ\cdots\circ\sigma_1\circ A_1(x')\|\\ &\leq C_\sigma^L\prod_{i=1}^{L+1}\|W_i\|\,\|x-x'\|\\ &\leq C_\sigma^L B^{L+1}\|x-x'\|,\end{aligned}$$

其中，$\|\cdot\|$ 表示任意向量的矩阵范数。取 $\mathcal{K}_{\mathcal{H}_\sigma}=C_\sigma^L B^{L+1}$，[13] 中的定理 3.2 和定理 3.4 的条件成立，并且命题 4.1 得证。

## 8.2 定理 4.2 的证明

（i）令 $\varepsilon>0$ 并设 $\delta=\frac{\varepsilon}{4G}>0$。由于激活函数 $\sigma$ 是 $C_\sigma$-利普希茨的，根据 [31], 中的命题 1，我们有

$$\log\mathcal{N}(\mathcal{H}_\sigma(L,N,B,F,S),\frac{\varepsilon}{4G})\leq 2L(S+1)\log\left(\frac{4G}{\varepsilon}C_\sigma L(N+1)(B\vee 1)\right). \tag{8.1}$$

这等价于

$$\mathcal{N}(\mathcal{H}_\sigma(L,N,B,F,S),\frac{\varepsilon}{4G})\leq\exp\left(2L(S+1)\log\left(\frac{4G}{\varepsilon}C_\sigma L(N+1)(B\vee 1)\right)\right). \tag{8.2}$$

注意，若 $\varepsilon>2M$，则 $P¶\sup_{h\in\mathcal{H}_\sigma(L,N,B,F)}[R(h)-\widehat{R}_n(h)]>\varepsilon©=0.$对于 $\varepsilon\in(0,2M]$，根据 [13], 中的命题 3.1 和注记 3.3，取 $A_n=2nM^2\Psi(1,1)L_1$ 以及 $B_n=2ML_2\max(2^{3+\mu}/\Psi(1,1),1)$，可得，

$$P\left\{\sup_{h\in\mathcal{H}_\sigma(L,N,B,F,S)}[R(h)-\widehat{R}_n(h)]>\varepsilon\right\}\leq\exp\left(-\frac{n^2\varepsilon^2/4}{C_1n+2C_2^{1/(\mu+2)}(n\varepsilon/2)^{(2\mu+3)/(\mu+2)}}\right). \tag{8.3}$$

除命题 4.1 和式 (8.2) 外，还有：

$$\begin{aligned}&P\left\{\sup_{h\in\mathcal{H}_\sigma(L,N,B,F,S)}[R(h)-\widehat{R}_n(h)]>\varepsilon\right\}\\ &\qquad\leq\exp\left(2L(S+1)\log\left(\frac{4G}{\varepsilon}C_\sigma L(N+1)(B\vee 1)\right)-\frac{n^2\varepsilon^2/4}{C_1n+2C_2^{1/(\mu+2)}(nM)^{(2\mu+3)/(\mu+2)}}\right),\end{aligned} \tag{8.4}$$

其中 $C_1 = 4M^2\Psi(1,1)L_1$且 $C_2 = 2ML_2\max(\frac{2^{3+\mu}}{\Psi(1,1)}, 1)$。令 $0 < \eta < 1$。考虑关于 $\varepsilon$的下列方程:

$$\exp\left(2L(S+1)\log\left(\frac{4G}{\varepsilon}C_\sigma L(N+1)(B\vee 1)\right) - \frac{n^2\varepsilon^2/4}{C_1 n + 2C_2^{1/(\mu+2)}(nM)^{(2\mu+3)/(\mu+2)}}\right) = \eta.$$

这等价于,

$$2L(S+1)\log\left(\frac{4G}{\varepsilon}C_\sigma L(N+1)(B\vee 1)\right) - \frac{n^2\varepsilon^2/4}{C_1 n + 2C_2^{1/(\mu+2)}(nM)^{(2\mu+3)/(\mu+2)}} + \log(1/\eta) = 0. \tag{8.5}$$

Set,

$$C_{n,1} = \frac{n^2}{4C_1 n + 8C_2^{1/(\mu+2)}(nM)^{(2\mu+3)/(\mu+2)}}.$$

我们得到,

$$2L(S+1)\log\left(\frac{4G}{\varepsilon}C_\sigma L(N+1)(B\vee 1)\right) - C_{n,1}\varepsilon^2 + \log(1/\eta) = 0.$$

i.e.

$$-2L(S+1)\log\varepsilon + C_6 - C_{n,1}\varepsilon^2 - \log(\eta) = 0,$$

其中 $C_6 = 2L(S+1)\log(4GC_\sigma L(N+1)(B\vee 1))$. 即

$$2L(S+1)\log\varepsilon + C_{n,1}\varepsilon^2 + \log(\eta) - C_6 = 0.$$

考虑函数 $\phi(\varepsilon) = 2L(S+1)\log\varepsilon + C_{n,1}\varepsilon^2 + \log(\eta) - C_6$, 对于 $\varepsilon \in (0, 2M)$。我们有, $\phi(\varepsilon) \to -\infty$当 $\varepsilon \to 0$ 时。

$$\begin{aligned}\phi(2M) > 0 &\Rightarrow 2L(S+1)\log(2M) + 4C_{n,1}M^2 + \log(\eta) - C_6 > 0\\ &\Rightarrow C_{n,1} > \frac{1}{4M^2}(C_6 - 2L(S+1)\log(2M) - \log(\eta)).\end{aligned} \tag{8.6}$$

由于 $(2\mu+3)/(\mu+2) > 1$, , 很容易看出

$$C_{n,1} \geq \frac{n^2}{n^{(2\mu+3)/(\mu+2)}(4C_1 + 8C_2^{1/(\mu+2)}M^{(2\mu+3)/(\mu+2)})} = \frac{n^{\frac{1}{\mu+2}}}{C_4}, \tag{8.7}$$

取 $C_4 = 4C_1 + 8C_2^{1/(\mu+2)}M^{(2\mu+3)/(\mu+2)}$。因此, 要得到 (8.6), 只需满足

$$\frac{n^{\frac{1}{\mu+2}}}{C_4} > \frac{1}{4M^2}(C_6 - 2L(S+1)\log(2M) - \log(\eta))_+ \geq \frac{1}{4M^2}(C_6 - 2L(S+1)\log(2M) - \log(\eta)).$$

回忆一下, 对于所有 $x \in \mathbb{R}$, $(x)_+ = \max(x,0)$。也就是说,

$$n > \left(\frac{C_4}{4M^2}(C_6 - 2L(S+1)\log(2M) - \log(\eta))_+\right)^{\mu+2}. \tag{8.8}$$

函数 $\varepsilon \mapsto \phi(\varepsilon)$ 严格递增, 并且在条件 (8.8) 下, 我们得到 $\phi(\varepsilon) \to_{2M} -\infty$ 当 $\varepsilon \to 0$ 且$\phi(2M) > 0$.因此, 存在唯一的 $\varepsilon_1(n, \eta) \in (0, 2M)$, 使得 $\phi(\varepsilon(n, \eta)) = 0$。令 $\varepsilon_0 = \frac{}{n^{\alpha^{\frac{1}{(\mu+2)}}}}$为$\alpha > 2$。除 (8.7) 外, 我们还有,

$$\begin{aligned}\phi(\varepsilon_0) &= 2L(S+1)\log\varepsilon_0 + C_{n,1}\varepsilon_0^2 + \log(\eta) - C_6\\ &\geq 2L(S+1)\log(2M) - \frac{2L(S+1)}{\alpha(\mu+2)}\log n + \frac{4M^2}{C_4}n^{\frac{1}{\mu+2}(1-\frac{1}{\alpha})} + \log(\eta) - C_6\\ &\geq 2L(S+1)\log(2M) - \frac{2L(S+1)}{\alpha(\mu+2)}\log n + \frac{4M^2}{C_4}n^{\frac{\alpha-2}{\alpha(\mu+2)}} + \log(\eta) - C_6.\end{aligned}$$

要得到 $\phi(\varepsilon_0) > 0$，只需满足 $2L(S+1)\log(2M) - \frac{2L}{(S+1)\alpha(\mu+2)}\log n + \frac{4M^2}{C_4} n^{\frac{\alpha-2}{\alpha(\mu+2)}} + \log(\eta) - C_6 > 0$，i.e.

$$n^{\frac{\alpha-2}{\alpha(\mu+2)}} \frac{4M^2}{C_4}\left(1 - \frac{2C_4L(S+1)}{\alpha 4M^2(\mu+2)} \frac{\log n}{n^{\frac{\alpha-2}{\alpha(\mu+2)}}}\right) > C_6 - \log(\eta) - 2L(S+1)\log(2M) \tag{8.9}$$

由于，

$$\frac{2C_4L(S+1)}{\alpha 4M^2(\mu+2)} \frac{\log n}{n^{\frac{\alpha-2}{\alpha(\mu+2)}}} \underset{n\to\infty}{\longrightarrow} 0,$$

存在 $n_0(L, S, \alpha, M, \mu, C_4) > 0$ 使得，

$$n > n_0(L, S, \alpha, M, \mu, C_4) \Rightarrow \frac{2C_4L(S+1)}{\alpha 4M^2(\mu+2)} \frac{\log n}{n^{\frac{\alpha-2}{\alpha(\mu+2)}}} < \frac{1}{2}. \tag{8.10}$$

因此，要得到 (8.9)，对于 $n > n_0(L, S, \alpha, M, \mu, C_4)$，只需满足，

$$n^{\frac{\alpha-2}{\alpha(\mu+2)}} \frac{4M^2}{2C_4} > C_6 - \log(\eta) - 2L(S+1)\log(4M).$$

也就是说，

$$n > \left(\frac{C_4}{2M^2}(C_6 - \log(\eta) - 2L(S+1)\log(2M))_+\right)^{\frac{\alpha(\mu+2)}{\alpha-2}}.$$

因此，对于充分大的 $n$，$\phi(\varepsilon(n, \eta)) = 0$ 的唯一解 $\varepsilon_1(n, \eta, \alpha)$ 满足 $\varepsilon_1(n, \eta, \alpha) < \frac{2M}{n^{\frac{1}{\alpha(\mu+2)}}} < 2M$ 对于 $\alpha > 2$。因此，由 (8.3) 可知，以至少 $1-\eta$，的概率，

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

这意味着，

$$R(\widehat{h}_n) - \widehat{R}_n(\widehat{h}_n) \le \varepsilon_1(n, \eta, \alpha). \tag{8.11}$$

这就证明了定理 4.2 的第一部分。

(ii) 根据 命题 3.1，在 [13]，中，对于所有 $h \in \mathcal{H}_\sigma(L, N, B, F, S)$，我们有以下不等式，

$$P\left\{\widehat{R}_n(h) - R(h) > \varepsilon\right\} \le \exp\left(-\frac{n^2\varepsilon^2/2}{A_n + B_n^{1/(\mu+2)}(n\varepsilon)^{(2\mu+3)/(\mu+2)}}\right). \tag{8.12}$$

因此，取上述的 $A_n = 2nM^2\Psi(1,1)L_1$ 和 $B_n = 2ML_2\max(2^{3+\mu}/\Psi(1,1), 1)$，并令目标神经网络为 $h_{\mathcal{H}_\sigma(L,N,B,F,S)}$，可得，

$$\begin{aligned} P\left\{\widehat{R}_n(h_{\mathcal{H}_\sigma(L,N,B,F,S)}) - R(h_{\mathcal{H}_\sigma(L,N,B,F,S)}) > \varepsilon\right\} &\le \exp\left(-\frac{n^2\varepsilon^2/2}{A_n + B_n^{1/(\mu+2)}(n\varepsilon)^{(2\mu+3)/(\mu+2)}}\right) \\ &\le \exp\left(-\frac{n^2\varepsilon^2}{C_1 n + 2C_2^{1/\mu+2}(2nM)^{(2\mu+3)/(\mu+2)}}\right) \\ &\le \exp(-C_n'\varepsilon^2), \end{aligned} \tag{8.13}$$

其中 $C_n' = \frac{n^2}{C_1 n + 2C_2^{1/\mu+2}(2nM)^{(2\mu+3)/(\mu+2)}}$。考虑关于 $\varepsilon$ 的方程，

$$\exp\left(-C_n'\varepsilon^2\right) = \eta.$$

该方程的一个解为：

$$\varepsilon_1'(n, \eta) = \left[\frac{\log(1/\eta)}{C_n'}\right]^{\frac{1}{2}}.$$

因此，由（8.12）可得:

$$\widehat{R}_n(h_{\mathcal{H}_\sigma(L,N,B,F,S)}) - R(h_{\mathcal{H}_\sigma(L,N,B,F,S)}) \leq \varepsilon_1'(n,\eta)$$

以至少 $1-\eta$. 的概率。由于 $R_n(h_n) \leq R_n(h_{\mathcal{H}_\sigma(L,\ N,\ B,\ F,\ S)})$（根据 $h_n = \text{argmin}_{h\in\mathcal{H}_\sigma(L,\ N,\ B,\ F,\ S)}[R_n(h)]$的定义），
我们推得

$$\widehat{R}_n(\widehat{h}_n) - R(h_{\mathcal{H}_\sigma(L,N,B,F,S)}) \leq \varepsilon_1'(n,\eta). \tag{8.14}$$

因此，根据 (8.11) 和 (8.14)，以至少 $1-2\eta$ 的概率，有:

$$R(\widehat{h}_n) - R(h_{\mathcal{H}_\sigma(L,N,B,F,S)}) \leq \varepsilon_1(n,\eta,\alpha) + \varepsilon_1'(n,\eta).$$

定理的证明至此完成。

## 8.3 定理 4.3 的证明

(i) 根据 (4.4)，取 $A_n' = nC$，其中 $B_n' = \log\binom{n}{n^{1/4}C}$。因此，对于 $\varepsilon \in (0,\ 2M]$，利用 (4.2) 和 (8.2)，当 $n$ 足够大时，我们得到，

$$P\left\{\sup_{h\in \mathrm{H}_\sigma(L,N,B,F,S)}[R(h) - \widehat{R}_n(h)] > \varepsilon\right\} \leq C_3 \exp\left(2L(S+1)\log\left(\frac{4G}{\varepsilon}C_\sigma L(N+1)(B\vee 1)\right)\right) \times \exp\left(\log\log n - \frac{n^2\varepsilon^2/4}{nC + \log n\ n^{\nu-1/4}M^\nu/C}\right). \tag{8.15}$$

考虑关于 $\varepsilon$ 的如下方程:

$$C_3 \exp\left(2L(S+1)\log\left(\frac{4G}{\varepsilon}C_\sigma L(N+1)(B\vee 1)\right) + \log\log n - \frac{n^2\varepsilon^2/4}{nC + \log n\ n^{\nu-1/4}M^\nu/C}\right) = \eta,$$

i.e.

$$2L(S+1)\log\left(\frac{4G}{\varepsilon}C_\sigma L(N+1)(B\vee 1)\right) + \log\log n - \frac{n^2\varepsilon^2/4}{nC + \log n\ n^{\nu-1/4}M^\nu/C} - \log(\eta) + \log C_3 = 0,$$

i.e.

$$-2L(S+1)\log\varepsilon + 2L(S+1)\log\left(4GC_\sigma L(N+1)(B\vee 1)\right) + \log\log n - \frac{n^2\varepsilon^2/4}{nC + \log n\ n^{\nu-1/4}M^\nu/C} - \log(\eta) + \log C_3 = 0,$$

i.e.

$$-2L(S+1)\log\varepsilon + 2L(S+1)\log\left(4GC_\sigma L(N+1)(B\vee 1)\right) + \log\left(C_3\log n/\eta\right) - \frac{n^2\varepsilon^2/4}{nC + \log n\ n^{\nu-1/4}M^\nu/C} = 0. \tag{8.16}$$

Set,

$$C_{n,2} = \frac{n^2/4}{nC + \log n\ n^{\nu-1/4}M^\nu/C}.$$

由 (8.16) 可得,

$$2L(S+1)\log\varepsilon + C_{n,2}\varepsilon^2 + \log(\eta/C_3) - C_6 - \log(n) = 0,$$

其中 $C_6 = 2L(S+1)$ 对数为 $(4GC_\sigma L(N+1)(B\vee 1))$。考虑函数

$$\varphi(\varepsilon) = 2L(S+1)\log\varepsilon + C_{n,2}\varepsilon^2 + \log(\eta/C_3) - C_6 - \log(n) \text{ for } \varepsilon \in (0, 2M].$$

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

$$\begin{aligned}\varphi(2M) > 0 &\Rightarrow 2L(S+1)\log(2M) + 4C_{n,2}M^2 + \log(\eta/C_3) - C_6 - \log(n) > 0 \\ &\Rightarrow C_{n,2} - \frac{\log(n)}{4M^2} > \frac{1}{4M^2}\left(C_6 - 2L(S+1)\log(2M) - \log(\eta/C_3)\right) \end{aligned} \tag{8.17}$$

注意，由于 $\nu \in (0,1)$，当 $n$ 足够大时，我们有

$$C_{n,2} \geq \frac{n^2/4}{C_1' n + C_2' n M^\nu} \geq \frac{n}{C_5} \text{ and } \frac{\log(n)}{4M^2} < \frac{n}{2C_5} \tag{8.18}$$

其中 $C_5 = 4(C + M^\nu/C)$。因此，要得到 (8.17)，只需满足：

$$\frac{n}{2C_5} > \frac{1}{4M^2}\left(C_6 - 2L(S+1)\log(2M) - \log(\eta/C_3)\right)_+ \geq \frac{1}{4M^2}\left(C_6 - 2L(S+1)\log(2M) - \log(\eta/C_3)\right).$$

也就是说，

$$n > \frac{C_5}{8M^2}\left(C_6 - 2L(S+1)\log(2M) - \log(\eta/C_3)\right)_+. \tag{8.19}$$

因此，函数 $\varepsilon \mapsto \varphi(\varepsilon)$ 严格递增，并且对于充分大的 $n$ 且满足 (8.19)，我们得到 $\varphi(\varepsilon) \to -\infty$（当 $\varepsilon \to 0$ 时）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 $\varepsilon_0' = \frac{2M}{1}$（当 $\alpha > 2$ 时）。与定理 4.2 第一部分的证明相同，可以发现，要得到 $n\varphi(\varepsilon_0') > 0$，只需满足：

$$n > \left(\frac{C_5}{2M^2}(C_6 - \log(\eta/C_3) - 2L(S+1)\log(2M))_+\right)^{\alpha/(\alpha-2)}. \tag{8.20}$$

因此，对于充分大的 $n$ 且满足 (8.20) 的情形，$\varphi(\varepsilon_0') > 0$。因此，方程 $\varphi(\varepsilon)$ 的唯一解 $\varepsilon_2(n, \eta, \nu, \alpha)$ 满足 $\varepsilon_2(n, \eta, \nu, \alpha) < \frac{2M}{1}$，其中 $\alpha > 2$。

根据 (8.15)，我们推得以至少 $1-\eta$ 的概率，

$$R(\widehat{h}_n) - \widehat{R}(\widehat{h}_n) \leq \varepsilon_2(n, \eta, \nu, \alpha). \tag{8.21}$$

因此，(i) 部分得证。

(ii) 同理，我们可以沿用 (8.14) 中的类似思路（参见定理 4.2(ii) 的证明）证明，以至少 $1-\eta$ 的概率

$$\widehat{R}_n(\widehat{h}_n) - R(h_{\mathcal{H}_\sigma(L,N,B,F)}) \leq \varepsilon_2'(n, \eta, \nu), \tag{8.22}$$

其中，

$$\varepsilon_2'(n,\eta,\nu) = \left[\frac{\log(C_3 \log n/\eta)}{C_{n,2}'}\right]^{\frac{1}{2}} \text{ with } C_{n,2}' = \frac{n^2}{nC + \log n\, n^{\nu-1/4} M^\nu/C}.$$

结合 (8.21) 和 (8.22)，以至少 $1-2\eta$ 的概率，我们得到

$$R(\widehat{h}_n) - R(h_{\mathcal{H}_\sigma(L,N,B,F,S)}) \leq \varepsilon_2(n,\eta,\nu,\alpha) + \varepsilon_2'(n,\eta,\nu)$$

至此，定理得证。

# 参考文献

[1] Bardet, J.-M., Kamila, K., and Kengne, W. 常见时间序列模型的一致模型选择准则与拟合优度检验。《电子统计学期刊》14, 1 (2020), 2009–2052.[2] Bardet, J.-M., Kengne, W., and Wintenberger, O. 使用惩罚准似然对一般因果时间序列进行多重断点检测。《电子统计学期刊》6 (2012), 435–477.[3] Bardet, J.-M., and Wintenberger, O. 多维因果过程的准最大似然估计量的渐近正态性。《统计年鉴》37, 5B (2009), 2730–2759.[4] Bauer, B., and Kohler, M. 关于将深度学习作为非参数回归中维数灾难的解决方法。《统计年鉴》47, 4 (2019), 2261–2285.[5] 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.[6] Davis, R. A., and Liu, H. 一类非线性模型的理论与推断及其在计数时间序列中的应用。《中华统计学志》(2016), 1673–1707.[7] 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。[8] Diop, M. L., and Kengne, W. 一般整数值时间序列中的参数变化检验。《时间序列分析杂志》38, 6 (2017), 880–894.[9] DIOP, M. L., and KENGNE, W. 具有多元外生协变量的一般整数值时间序列的密度幂散度。arXiv预印本 arXiv:2006.11948 (2020).[10] Diop, M. L., and Kengne, W. 一般整数值时间序列的分段自回归。《统计规划与推断杂志》211 (2021), 271–286.[11] Diop, M. L., and Kengne, W. 多元时间序列变点检测的一般程序。《TEST》(2022), 1–33.[12] 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。《电子统计学期刊》16, 1 (2022), 116–157.[13] Diop, M. L., and Kengne, W. $\psi$-弱依赖过程的统计学习。arXiv预印本 arXiv:2210.00088 (2022).[14] 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.[15] 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.[16] Doukhan, P., and Wintenberger, O. 具有无限记忆的弱依赖链。《随机过程及其应用》118, 11 (2008), 1997–2013。

[17] Dziugaite, G. K., and Roy, D. M. 为参数数量远超训练数据的深度（随机）神经网络计算非平凡泛化界。arXiv预印本 arXiv:1703.11008（2017）。[18] Francq, C., et al. 带协变量的波动率模型的QML推断。《计量经济学理论》35，1（2019），37–72。[19] 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. 用于语音识别声学建模的深度神经网络：四个研究团队的共同观点。IEEE信号处理杂志29，6（2012），82–97。[20] Hudecová, Š. 二元时间序列自回归模型中的结构变化。《统计规划与推断杂志》143，10（2013），1744–1752。[21] Hwang, E., and Shin, D. W. 弱依赖条件下矩不等式的研究。韩国统计学会杂志42，1（2013），133–141。[22] Kengne, W. 一般因果时间序列的强一致模型选择。统计与概率快报171（2021），109000。[23] Kengne, W. C. 一般因果时间序列模型参数恒定性的检验。《时间序列分析杂志》33，3（2012），503–518。[24] 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。[25] Kingma, D. P., and Ba, J. Adam：一种随机优化方法。arXiv预印本 arXiv:1412.6980（2014）。[26] Kohler, M., and Krzyzak, A. 相关数据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）。[27] Krizhevsky, A., Sutskever, I., and Hinton, G. E. 使用深度卷积神经网络进行ImageNet分类。ACM通讯60，6（2017），84–90。
[28] 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）。[29] Lamine DIOP, M., and KENGNE, W. 具有多元外生协变量的一般整数值时间序列的密度幂散度。arXiv电子预印本（2020），arXiv–2006。[30] Ma, M., and Safikhani, A. 时间相关观测下深度神经网络的理论分析。arXiv预印本 arXiv:2210.11530（2022）。[31] Ohn, I., and Kim, Y. 使用一般激活函数的深度神经网络进行光滑函数逼近。《熵》21，7（2019），627。
[32] Schmidt-Hieber, J. 深度ReLU网络对流形上函数的逼近。arXiv预印本 arXiv:1908.00695（2019）。

[33] Schmidt-Hieber, J. 使用ReLU激活函数的深度神经网络进行非参数回归。《统计年鉴》*48*，4（2020），1875–1897。

[34] Valle-Pérez, G., and Louis, A. A. 深度学习的泛化界。*arXiv* 预印本 *arXiv:2012.04115* (2020)。