arXiv:2602.02172v1 [stat.ME] 2 Feb 2026

神经网络机器回归（NNMR）：一种揭示高阶协同效应的深度学习框架

Jiuchen Zhang，密歇根大学生物统计系；
Ling Zhou，西南财经大学统计研究中心和统计学院；
Peter Song，密歇根大学生物统计系

2026年2月3日

摘要

我们提出了一种新的神经网络框架，称为神经网络机器回归（NNMR），该框架整合了可训练输入门控和自适应深度正则化，以端到端的方式联合进行特征选择和函数估计。通过对门控参数和冗余层同时施加惩罚，NNMR 能够生成稀疏且可解释的架构，同时捕捉由高阶协同效应驱动的复杂非线性关系。我们进一步开发了一种基于分割样本和置换检验的事后推断程序，使得无需严格的参数假设即可进行有效推断。与包括贝叶斯核机器回归和常用事后归因技术在内的现有方法相比，NNMR 能够高效处理高维特征空间，同时严格控制第一类错误。模拟研究表明，该方法具有更高的选择准确率和更可靠的推断。最后，一项实证应用揭示了与居住在墨西哥城的青少年躯体生长相关的稀疏且具有生物学意义的食物组预测因子。

关键词：选择后推断、深度正则化、输入门控

1 引言

在生物医学研究、基因组学、流行病学和环境科学等许多科学领域，研究人员处理的高维数据集往往包含远多于观测值数量的潜在解释变量。识别出相关变量的子集，对于提升模型的可解释性、减少过拟合并增强预测性能至关重要。

变量选择在线性建模中有着悠久的历史，其中诸如LASSO（Tibshirani 1996）和平滑截断绝对偏差（SCAD；Fan和Li 2001）等稀疏惩罚现已成为常规方法。然而，许多科学问题涉及非线性的、可能为高阶的交互作用，远远超出线性假设的范畴。为适应这种复杂性，研究已转向非参数框架，使数据能够揭示灵活的函数形式，同时仍能分离出有影响的预测变量。早期研究将线性模型扩展为加性结构，其中每个协变量通过一个未指定的单变量函数进入（Hastie 2017）。尽管加性模型继承了可解释性，并可配备逐分量选择规则，但除非纳入高阶加性项或交互核，否则它们可能遗漏交互效应；而这一步会大大增加估计与特征选择的复杂性。

互补策略通过筛选而非惩罚来应对“维数灾难”。确定性独立筛选（SIS；Fan 和 Lv, 2008）及其变体依据边际效用对预测变量排序，以便在开展更复杂的建模之前选定特征空间；然而，由于依赖边际信号，这些方法可能会忽略仅通过交互作用发挥作用的变量。以贝叶斯核机器回归（BKMR；Bobb等人，2015）为代表的完整贝叶斯方法，将变量选择、非线性响应面和不确定性量化整合到一个连贯的框架中，但代价是需要进行马尔可夫链蒙特卡洛计算，而这种计算在样本或

协变量达到数千个时扩展性较差。总体而言，这些进展凸显了非参数回归中始终存在的矛盾：如何平衡建模灵活性、统计效率、计算可行性和可解释性。

深度神经网络给特征选择带来了双重挑战：参数化存在大量冗余，而权重与输入之间的对应关系又高度纠缠。SHAP（Lundberg & Lee2017）、积分梯度（Sundararajan 等人 2017）和 DeepLIFT（Shrikumar 等人 2017）等事后归因工具通过将梯度或相关性值反向传播至输入层，估计事后重要性分数。这些方法只需要经过训练的模型，在实践中应用广泛，但不会在训练过程中施加稀疏性；在存在共线性时，其分数可能不稳定，也无法正式保证低分变量无关紧要——在高维生物医学应用中，这些局限尤为突出。

为在神经网络中引入稀疏性，目前已有两种嵌入式方法。深度特征选择在网络输入处插入一个稀疏线性层，并对其权重施加 ℓ_1惩罚，从而以端到端的方式联合学习表示和特征重要性（Li 等人 2016）。在此框架基础上，非线性变量选择方法引入一种连续的 ℓ_0 选择层松弛，在广义稳定 Hessian 条件下实现严格的收敛保证和选择一致性（Chen 等人 2021）。组 Lasso 正则化则将每个输入变量的所有出边连接视为一组，通过剪除整个神经元或输入，构建高度紧凑的网络，而无需进行大量人工分组（Scardapane 等人 2017）。然而，这些方法仍无法自动控制网络深度和宽度，对有效的选择后推断也支持有限。

在本文中，我们提出了一种基于神经网络的变量选择方法，旨在有效识别高维情境中的相关特征。我们的方法融入了

多项创新，以增强可解释性、可扩展性和统计严谨性。首先，我们在神经网络的输入层引入了一种可训练的特征门控机制，为每个特征赋予一个可学习的权重参数。结合类似于 LASSO（Tibshirani 1996）的 L_1惩罚项，该机制可直接诱导稀疏性，并明确识别出与结果变量关联最强的一组精简预测变量。

其次，我们的方法在训练过程中采用自适应阈值处理程序，定期移除门控参数持续保持较小值的特征。此外，如果隐藏层的贡献变得微不足道，我们会动态地将其替换为恒等映射，从而剪除冗余隐藏层。这两种处理方式促进了输入特征和网络深度的结构化稀疏性，最终得到紧凑且具有可解释性的模型。

第三，我们通过推导风险上界，为该方法建立了理论保证，并证明其在温和条件下达到极小极大最优性。结合我们提出的数据拆分方法和基于置换的推断框架，我们的方法能够在高维分析中严格控制第一类错误，有效减少假发现。

与现有方法相比，我们基于神经网络的变量选择方法具有多项优势，尤其适用于涉及大规模生物医学和遗传数据的场景。与贝叶斯核机器回归（BKMR）（Bobb等人，2015）需要进行计算开销高昂的核矩阵求逆，或传统筛选方法（Fan &Lv 2008不同，我们的框架能够高效扩展至数百甚至数千个预测变量。此外，与 SHapley Additive exPlanations（SHAP）或积分梯度等常见的事后解释方法相比——这些方法在高维场景中往往不稳定或会产生误导——我们的方法通过结构化惩罚，在模型训练过程中直接纳入可解释性。

总体而言，我们的单一该框架同时解决特征选择、自适应神经

网络优化和严格的统计推断问题，使我们的方法成为一项强有力的进展，可用于当代统计建模和深度学习应用中的可解释、可扩展且在统计上严谨的特征选择。

2 方法论

我们在神经网络框架下研究变量选择。假设我们观测到独立同分布样本 $\{(\boldsymbol{X}_i,\ Y_i)\}_{i=1}^n$，其中 $Y_i \in \mathbb{R}$ 是标量响应，$\boldsymbol{X}_i \in \mathbb{R}^d$ 是高维预测变量向量。我们的目标是同时估计预测变量与响应之间未知的关系 $g(\cdot)$，并识别出相关特征的一个稀疏子集。为此，我们将变量选择直接嵌入神经网络训练过程。具体而言，我们通过最小化经验平方误差损失来估计 g：

$$\hat{g} = \arg\min_{g \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^{n} (Y_i - g(\boldsymbol{X}_i))^2, \tag{1}$$

其中$\mathcal{G}$ 表示　　专门用于执行变量选择的神经网络类　　.

2.1 神经网络

我们将 $\mathcal{G}$ 设为一个函数类，该类由采用 ReLU 激活函数的多层神经网络组成，用于近似条件期望：

$$\mathcal{G} := \mathcal{NN}(\mathcal{W}, \mathcal{D}, \mathcal{S}, \mathcal{B}),$$

其中，输入数据是预测变量 $\boldsymbol{X}_i$，构成第一层；输出则是网络的最后一层。该网络$\mathcal{G}$ 的宽度为$\mathcal{D}$，其中包含$\mathcal{D}$ 个隐藏层，且$\mathcal{D}+2$ 总层数。令 N_l 表示层 l 的宽度，其中 $l = 0, \ldots, \mathcal{D}, \mathcal{D}+1$，即每层的节点数。具体而言，$N_0 = d$ 表示 X 的输入维度，

而 $N_{\mathcal{D}+1}=1$ 表示响应 Y。宽度$\mathcal{W}$ 是隐藏层中的最大宽度：

$$\mathcal{W}=\max(N_1,\ldots,N_{\mathcal{D}}).$$

不失一般性，本文对所有隐藏层考虑相同的宽度。规模 S 是网络 G 中的参数总数，由 S $=\sum \mathrm{D}_{l=0} N_{l+1}\times(N_l+1)$ 给出。具体而言，具有 D 个隐藏层的 ReLU 神经网络是一组映射g：形式为 $g(\boldsymbol{x}) = h_{\mathcal{D}}\circ\sigma\circ h_{\mathcal{D}-1}\circ\cdots\circ\sigma\circ h_0(\boldsymbol{x})$ 的 $\mathbb{R}^{N_0}\to\mathbb{R}^{N_{\mathcal{D}+1}}$，其中 $h_1\circ h_2(\boldsymbol{x})$: $= h_1(h_2(\boldsymbol{x}))$ 表示两个函数 h_1 与 h_2 的复合； $\sigma(\boldsymbol{x}) := \max(\boldsymbol{x}, 0)$ 是 ReLU 函数，按元素应用； $h_l(\boldsymbol{x}) := \boldsymbol{W}_l\boldsymbol{x}+c_l$ 是带有$\boldsymbol{W}_l\in\mathbb{R}^{N_{l+1}\times N_l}$ 和 $c_l\in\mathbb{R}^{N_l}$ 的仿射变换。我们假设每个函数 $g\in$G 对某个 $0<\mathrm{B}<\infty$ 满足 $\|g\|_\infty\le \mathrm{B}$，其中 $\|g\|_\infty$ 是函数 g 的上确界范数。

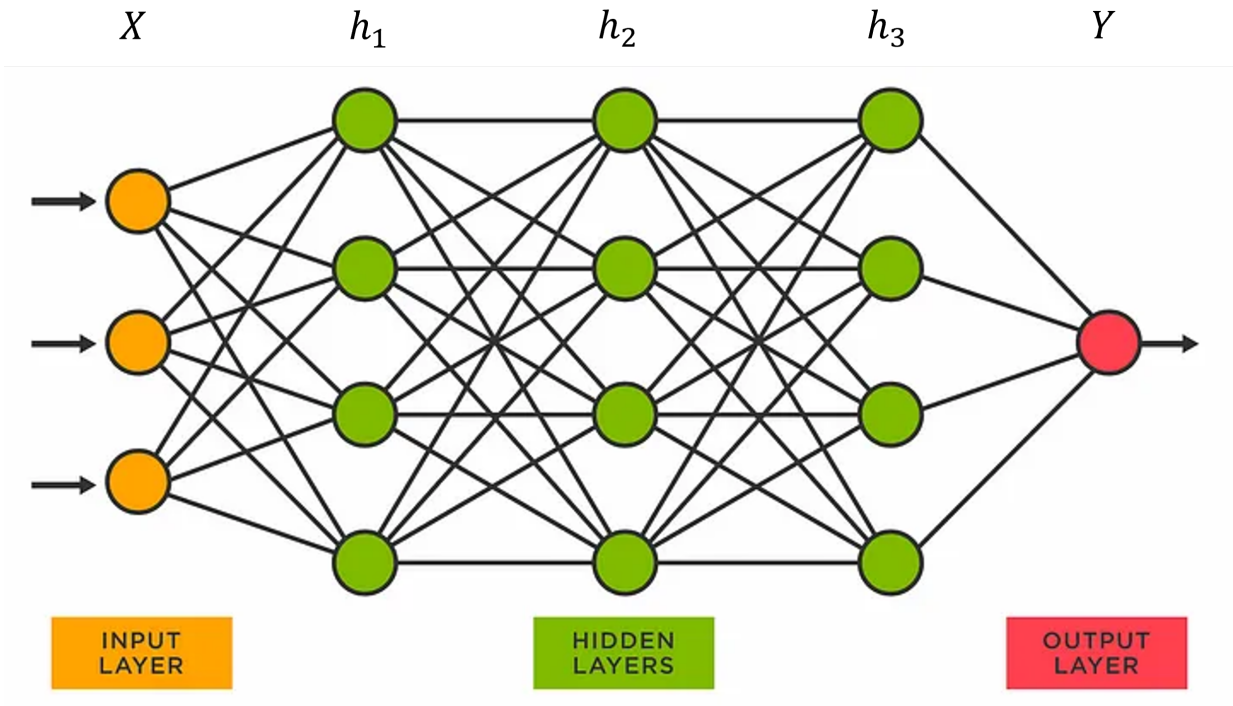

图1：简单的多层神经网络

2.2 变量选择

在高维非参数回归问题中，响应变量 Y 通常仅依赖于可用特征中的一小部分，而其余特征对预测准确率的贡献很小。这表明数据位于嵌入高维特征空间中的低维流形上。利用这一假设，我们旨在识别并仅保留最相关的特征，从而在保持预测性能的同时降低计算复杂度。

为利用这种稀疏性，我们在输入层附加一个可训练门控向量 $\boldsymbol{\alpha} = (\alpha_1, \ldots\ldots, \alpha_d)^\top \in \mathbf{R}^d$，并将重新加权后的输入馈入

$$\widetilde{\boldsymbol{X}} = \boldsymbol{\alpha} \odot \boldsymbol{X}, \qquad (\boldsymbol{\alpha} \odot \boldsymbol{X})_j = \alpha_j X_j,$$

网络。

这种变换使模型能够通过将 α 的某些项设为零来学习哪些特征至关重要，从而有效地将这些变量排除在考虑范围之外。由此产生的特征选择机制嵌入神经网络架构中，确保只有一部分输入维度会对最终模型产生贡献。

为了形式化这一选择过程，我们对 α加入 L_1 正则化，通过惩罚较小的系数值来促进$\widetilde{\ }$稀疏性。若 α 恰有 d_0 个非零元素，则 $\boldsymbol{X}$可视为一个 d_0维向量，使模型能够学习一个函数 $g\colon \mathbb{R}^{d_0} \to \mathbb{R}$。这引出了以下优化问题：

$$\mathcal{L}_n(g) = \frac{1}{n}\sum_{i=1}^{n}\Big\{Y_i - g(\boldsymbol{X}_i)\Big\}^2 + \lambda_1 \|\boldsymbol{\alpha}\|_1,$$

此处， $g \in \mathcal{G}(\boldsymbol{\alpha}, \boldsymbol{\theta})$ 是一个将所选特征映射到 Y 的神经网络，其中 $\mathcal{G}(\boldsymbol{\alpha}, \boldsymbol{\theta}) = \{g\colon g = \phi(\boldsymbol{\alpha} \odot \boldsymbol{X})\ \phi \in \mathcal{G}\}$， $\boldsymbol{\theta} = (\boldsymbol{W}_l, c_l, l = 0, \cdots, \mathcal{D})$，并且 $\|\cdot\|_1$用作确保 $\boldsymbol{\alpha}$ 中稀疏性的 L_1范数。该公式不仅选择信息量最大的特征

还学习这些特征与响应变量之间的非线性映射。与在单独预处理步骤中运行的传统特征选择方法不同，我们的方法将特征选择整合到模型训练过程中，使其能够根据数据动态调整所选特征。

与现有的特征选择技术相比，例如贝叶斯核机器回归（BKMR）和基于 SHAP 的方法，我们的方法在可扩展性和效率方面具有显著优势。BKMR 在建模非线性交互方面虽功能强大，但由于需要进行核矩阵求逆，计算成本高昂，因此不适用于大规模问题。同样，基于 SHAP 的方法通过事后估计特征重要性，既不强制实现稀疏性，在处理高维数据时计算成本也可能很高。

相比之下，我们的方法与 GPU 加速深度学习框架完全兼容，能够高效扩展。集成的特征选择机制无需单独进行预处理，从而减少了开销。此外，通过显式学习稀疏表示，我们的方法确保模型在保留预测能力的同时仍具可解释性。

输入门控层缩小了有效特征空间的宽度，但网络的深度保持不变。每当活动集的大小发生变化时，都重新训练一种新架构，在计算上并不可行。遵循Tan 等人（2024），当隐藏层 l 的线性映射 $h_l\left(x\right) = \boldsymbol{W}_l x$ 只是传递其输入时，该层便是冗余的，即 $\boldsymbol{W}_l = \boldsymbol{I}$ 且 $c_l = 0$，其中 $\boldsymbol{I}$ 是单位矩阵。请注意，偏置项已并入增广输入中。

这正是引入深度惩罚的动机，

$$\mathrm{DP}(\theta) = \sum_{l=1}^{\mathcal{D}} (\|\boldsymbol{W}_l - \boldsymbol{I}\|_1 + |c_l|),$$

当且仅当每个可省略层都退化为恒等映射时，该惩罚才为零。结合输入门控与深度正则化，最终的训练问题变为

$$(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\theta}}) = \arg \min_{g \in \mathcal{G}(\boldsymbol{\alpha}, \boldsymbol{\theta})} \Big\{ \frac{1}{n} \sum_{i=1}^{n} \big\{ Y_i - g(\boldsymbol{X}_i) \big\}^2 + \lambda_1 \|\boldsymbol{\alpha}\|_1 + \lambda_2 \sum_{l=1}^{\mathcal{D}} (\|\boldsymbol{W}_l - \boldsymbol{I}\|_1 + |c_l|) \Big\}, \quad (12)$$

其中， λ_1 控制稀疏性， λ_2 调节深度。在优化过程中，每个 $\boldsymbol{W}_l$ 都会被软性地收缩至单位矩阵；训练后，可剪除惩罚项收敛至零的部分，从而得到一种紧凑架构，其复杂度与选定特征子集相匹配，同时保留深层 ReLU 网络的表达能力。

2.3 变量选择后的推断程序

完成变量选择后，需要进行严格的统计推断，以评估保留预测变量的显著性。我们采用数据拆分策略（Cox 1975年、Hurvich 和 Tsai 1990年），并结合条件独立性的无模型检验（Cai 等， 2022年）。具体而言，我们将样本划分为两部分：一部分用于变量选择和条件均值函数估计，另一部分专门用于推断，从而最大限度地保留检验效能。

具体而言，我们将原始数据集划分为两个子集：一个用于变量选择和估计（记为 D_1），另一个专门用于推断（记为 D_2）。第一个子集D_1 既用于通过第 2.2 节所述的惩罚神经网络识别重要预测变量，也用于估计推断所需的条件期望。推断步骤直接使用第二个子集的全部数据 D_2 ，不再进一步拆分，从而提高统计效能。

将从 D_1 中识别出的选定变量记为 X_S。推断程序涉及检验以下原假设：

$$H_0 : E(Y \mid \boldsymbol{X}_{-j,S}) = E(Y \mid \boldsymbol{X}_S) \quad \text{versus} \quad H_1 : E(Y \mid \boldsymbol{X}_{-j,S}) \neq E(Y \mid \boldsymbol{X}_S),$$

其中， $\boldsymbol{X}_j$ 表示潜在的感兴趣变量，而 $\boldsymbol{X}_{-j,\ S}$ 表示$\boldsymbol{X}_S$ 中除 $\boldsymbol{X}_j$以外的其余变量。

按照 Cai 等人 (2022年) 的方法，结合我们改进的方法，我们使用来自 D_1 的数据拟合两个非参数模型：一个同时纳入 X_S 和 Z 的模型，记为 $\hat{}g_1(\boldsymbol{X}_S$，以及一个仅纳入 $\boldsymbol{X}_{-j,\ S}$ 的零模型，记为 $\hat{}g_0(\boldsymbol{X}_{-j,\ S})$。为评估这些模型的预测性能，我们使用推断数据集 D_2 对这两个模型的残差进行双样本比较。具体而言，我们采用双样本 t 检验（TS），其定义为两个模型的平方残差之差的平均值。检验统计量 T_{TS} 的计算方式如下：

$$T_{TS} = \frac{1}{n_2} \sum_{i \in D_2} \left[\{Y_i - \hat{g}_1(\boldsymbol{X}_{S,i})\}^2 - \{Y_i - \hat{g}_0(\boldsymbol{X}_{-j,S,i})\}^2 \right],$$

如果 T_{TS} 显著为负，则拒绝原假设 H_0 。该检验在假设数据存在二阶矩的前提下，评估两个模型的均方预测误差之差。

算法 1：采用置换检验的推断程序 输入：训练数据 D_1、推断数据 D_2、置换次数 B、通过 D_1获得的选定变量 $\boldsymbol{X}_S$ 输出：检验统计量和 p 值 1 使用 D_1估计 $\hat{}g_1(\boldsymbol{X}_S)$ 和 $\hat{}g_0(\boldsymbol{X}_{-j,S})$； 2 在 D_2 上评估两个模型并计算残差：3 $U_i = Y_i - \hat{}g_1(\boldsymbol{X}_{S,i})$, 4 $V_i = Y_i - \hat{}g_0(\boldsymbol{X}_{-j,S,i})$，对所有 $i \in D_2$； 5 定义合并集合 $S = \{U_i\} \cup \{V_i\}$； 6 使用 $\{U_i\}$、$\{V_i\}$计算观测检验统计量 T； 7 对 $b = 1$ 到 B 执行 8 将 S随机划分为两个大小相等的集合 $\{U_i^*\}$、$\{V_i^*\}$； 9 使用 $\{U_i^*\}$、$\{V_i^*\}$计算检验统计量 $T^*_{TS,b}$； 10 Com

pute p-值：$\hat{}\ p_{TS} = \frac{1}{B}\sum_{b=1}^{B} I\{T > T^*_{TS,b}\}$;

这种增量预测能力的显著性通过置换检验进行评估，具体如算法 1所述。使用 D_1 对 $\hat{}g_1(\boldsymbol{X}_S)$ 和 $\hat{}g_0(\boldsymbol{X}_{-j,\ S})$ 进行估计，可充分利用推断数据集，从而提高统计效率和检验效能，并确保在不依赖严格参数假设的情况下进行稳健推断。此外，我们的推断框架具有灵活性，既可分别对每个选定变量进行单独检验，也可根据研究目标仅关注特定感兴趣变量。

此外，在适当的条件下，可以证明，t 统计量 T_{TS} 的渐近正态性已在Lei （2020）中确立。因此， p 值计算的最后一步可以调整为：先估计$\{T^*_{TS,b}, b = 1, ..., B\}$的标准差，记为 $\hat{}\sigma_B$，再将 p 值计算为 $\Phi^{-1}(T/\hat{}\sigma_B)$，其中 Φ 是标准正态分布的累积分布函数。这样得到的 p 值将具有

相对较高的分辨率。

3 理论

本节介绍有关我们基于神经网络的变量选择方法的一致性和有效性的理论结果。我们首先正式定义本方法所依据的理论框架。

设 $X_{[j]}$ 为 $\boldsymbol{X}$ 的第 j 个分量。当且仅当满足以下条件时， $X_{[j]}$ 被定义为条件不重要

$$\mathbb{E}(Y \mid \boldsymbol{X}_{[-j]} = \boldsymbol{x}_{[-j]}) = \mathbb{E}(Y \mid \boldsymbol{X} = \boldsymbol{x}), \tag{2}$$

w这里， $\boldsymbol{X}_{[-j]} = (X_{[1]}, \cdots X_{[j-1]}, X_{[j+1]}, \cdots X_{[d]})$， $\boldsymbol{x}_{-j}$ 的定义类似 y.

我们假设 $\boldsymbol{X}$ 的支撑集是一个有界集。为简单起见，假设该有界集为 $[0,1]^d$。在本文余下部分，常数 c 表示一个正常数，其取值可能因上下文而异。我们将 $\boldsymbol{X}$ 分为 $\boldsymbol{X}^s \in [0,1]^{d_0}$ 和 $\boldsymbol{X}^c \in [0,1]^{d-d_0}$，分别表示条件重要变量和不重要变量。对于任意 $\boldsymbol{x} \in [0,1]^p$， $\mathbb{E}(Y \mid \boldsymbol{X}^s = \boldsymbol{x}^s) = \mathbb{E}(Y \mid \boldsymbol{X} = \boldsymbol{x})$。

设 g^* 为总体损失的最小化解，即

$$g^* = \arg \min_{g \in \mathcal{H}^\beta([0,1]^p, \mathcal{B})} \mathbb{E}\left(Y_i - g(\boldsymbol{X}_i)\right)^2,$$

其中，极小值是在整个空间上取得的，因此意味着 g^* 不一定属于前馈神经网络集合 $\mathcal{G}(\boldsymbol{\alpha}, \boldsymbol{\theta})$。显然，根据定义（2）， $g^*(\boldsymbol{X})$仅取决于 $\boldsymbol{X}^s$ 。为简化记号，我们假设 d 中前 $\boldsymbol{X}$ 个分量是重要的，并将 $\boldsymbol{\alpha}$ 分为两部分 $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_s, \boldsymbol{\alpha}_c)$ ，其中 $\boldsymbol{\alpha}_s \in \mathbb{R}^d$ 和 $\boldsymbol{\alpha}_c \in \mathbb{R}^{p-d}$分别对应于 $\boldsymbol{X}^s$ 和 $\boldsymbol{X}^c$ 的系数。

我们根据其超额风险建立 $\hat{g}$ 的大样本性质，其中超额风险定义为 g 与 g^*的风险之差:

$$\mathcal{R}(g) - \mathcal{R}(g^*) = \mathbb{E}\left(Y_i - g(\boldsymbol{X}_i)\right)^2 - \mathbb{E}\left(Y_i - g^*(\boldsymbol{X}_i)\right)^2.$$

我们构造了一个估计量 g^*，其位于$\mathcal{G}(\boldsymbol{\alpha}, \boldsymbol{\theta})$内，网络参数为 $\boldsymbol{\theta}$，变量重要性指数参数为 $\boldsymbol{\alpha}$。因此，参照最优网络解

$$g_{\mathcal{G}}^* = \arg\min_{g \in \mathcal{G}(\boldsymbol{\alpha}, \boldsymbol{\theta})} \mathbb{E}\left(Y_i - g(\boldsymbol{X}_i)\right)^2,$$

我们将$(\boldsymbol{\alpha}, \boldsymbol{\theta})$的集合定义为

$$\boldsymbol{\Theta} = \{(\boldsymbol{\alpha}, \boldsymbol{\theta}) : \mathcal{R}(g_{\boldsymbol{\alpha},\boldsymbol{\theta}}) = \mathcal{R}(g_{\mathcal{G}}^*), g_{\boldsymbol{\alpha},\boldsymbol{\theta}} \in \mathcal{G}(\boldsymbol{\alpha}, \boldsymbol{\theta})\}.$$

在数据结构和网络类满足适当条件时，$g_{\mathcal{G}}^*$ 可以足够接近 g^*，使得 $\boldsymbol{\alpha}_c = 0$ 且 $\boldsymbol{\alpha}_s \neq 0$。也就是说，对于任意$(\boldsymbol{\alpha}, \boldsymbol{\theta}) \in \boldsymbol{\Theta}$，存在一个正常数 $c > 0$ 使得$\min_{1 \le j \le d} |\alpha_j| \ge c$。此外，存在一个解 $\boldsymbol{\alpha}_s \neq 0$，$\boldsymbol{\alpha}_c = \mathbf{0}$使得$(\boldsymbol{\alpha}_s, \mathbf{0}, \boldsymbol{\theta}) \in \boldsymbol{\Theta}$ 仍然成立。关于 $g_{\mathcal{G}}^*$和 $\boldsymbol{\Theta}$ 的这些性质见补充材料中的引理7.1。

在实践中，在式（12）中得到的网络参数（$\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\theta}}$）位于 $\boldsymbol{\Theta}$之外。为了度量$(\boldsymbol{\alpha}, \boldsymbol{\theta})$与 $\boldsymbol{\Theta}$之间的距离，我们定义度量 $d((\boldsymbol{\alpha}, \boldsymbol{\theta}), \boldsymbol{\Theta}) := \min_{\boldsymbol{\vartheta} \in \boldsymbol{\Theta}} \|\binom{\boldsymbol{\alpha}}{\boldsymbol{\theta}} - \boldsymbol{\vartheta}\|_2^2$。显然，$d((\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\theta}}), \boldsymbol{\Theta}) \to 0$ 足以保证$\hat{\boldsymbol{\alpha}}_s$远离零。因此，我们可以通过建立所提出的惩罚网络$\hat{g}_{\mathcal{G}}$ 关于其参数的收敛速率，来确立变量选择结果。

对于给定序列 $\boldsymbol{x} = (\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n)$，我们定义 $\mathsf{G}|_{\boldsymbol{x}} := \{g(\boldsymbol{x}_1), g(\boldsymbol{x}_2), \cdots, g(\boldsymbol{x}_n) : g \in \mathsf{G}(\boldsymbol{\alpha}, \boldsymbol{\theta})\}$，并将 $\mathsf{N}_{2n} = \sup_{\boldsymbol{x}} \mathsf{N}_{2n}(\delta, \|\cdot\|_\infty, \mathsf{G}|_{\boldsymbol{x}})$ 记为 $\mathsf{G}|_{\boldsymbol{x}}$ 在范数 $\|\cdot\|_\infty$ 下半径为 δ 的覆盖数。令 $A \preceq B$ 表示正常数 c 对应的 $A \le cB$，其中 $A \wedge B = \min(A, B)$；$\Phi_1 \circ \Phi_2 := \{\phi_1 \circ \phi_2 : \phi_1 \in \Phi_1, \phi_2 \in \Phi_2\}$ 表示两个函数类 Φ_1 和 Φ_2 的复合。

建立理论结果需要以下条件：

(C1) 条件重要变量 d 的维数固定，并且存在一个常数 $\tau_d > 0$，使得

$$\min_{1\le j\le d}\left|\mathbb{E}(Y \mid \boldsymbol{X}_{[-j]} = \boldsymbol{x}_{[-j]}) - \mathbb{E}(Y \mid \boldsymbol{X} = \boldsymbol{x})\right| \ge c\tau_d,$$

对于某个正常数 c。

(C2) 假设 Y 是一个次高斯随机变量。

(C3) $g_{\boldsymbol{\alpha},\,\boldsymbol{\theta}}$ 和 g^* 的函数类：对于任意函数 $g_{\boldsymbol{\alpha},\,\boldsymbol{\theta}} \in \mathcal{G}$ 以及真实函数 g^*，我们假设 $\|g_{\boldsymbol{\alpha},\,\boldsymbol{\theta}}\|_\infty < \mathcal{B}$ 和 $\|g^*\|_\infty < \mathcal{B}$。

定理 *3.1* 假设条件 *(C1)-(C3)* 成立。如果 $\lambda_1^2 \preceq \log^2 n \log \mathcal{N}_{nd}^{2n} + \left(\mathcal{R}(g_{\mathcal{G}}^*) - \mathcal{R}(g^*)\right)_d$ 且 $\lambda_2^2 \preceq \log^2 n \log \mathcal{N}_n^{2n} W D + \left(\mathcal{R}(g_{\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\theta}}}) - \mathcal{R}(g^*)\right) W^2 D$，则 *(12)* 中定义的 $(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\theta}})$ 满足

$$\begin{aligned}
\mathbb{E}[d((\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\theta}}), \boldsymbol{\Theta})] &\preceq \frac{\log^2 n \log \mathcal{N}_{2n}}{n} + \log^2 n \left(\mathcal{R}(g_{\mathcal{G}}^*) - \mathcal{R}(g^*)\right) \\
\mathbb{E}\|\hat{\boldsymbol{\alpha}}_c\|_2^2 &\preceq \frac{\log^2 n \log \mathcal{N}_{2n}}{n} + \log^2 n \left(\mathcal{R}(g_{\mathcal{G}}^*) - \mathcal{R}(g^*)\right),
\end{aligned}$$

其中，$\mathbb{E}$ 是相对于 $(\boldsymbol{X}_i, Y_i)_{i=1}^n$ 而言的。

接下来，我们进一步探讨误差如何取决于前馈神经网络结构以及 g^* 所属的函数类。我们考虑霍尔德类，它的范围足够广，能够涵盖大多数应用。具体而言，记 $\lceil a \rceil$ 和 $\lfloor a \rfloor$ 分别为不小于 a 的最小整数和严格小于 a的最大整数。令 $\mathbb{N}^+$ 为正整数集，$\mathbb{N}_0$ 为非负整数集。令 $\beta = s + r$, $r \in (0, 1]$ 和 $s = \lfloor \beta \rfloor \in \mathbb{N}_0$。对于有限常数 $B_0 > 0$，$H\ddot{o}lder$ 类 $\mathcal{H}^\beta([0,1]^d,\ B_0)$ 定义为

$$\begin{aligned}
\mathcal{H}_\beta([0,1]^d, B_0) = \{g : [0,1]^d \mapsto \mathbb{R}, &\max_{\|\alpha\|_1 < s} \|\partial^\alpha g\|_\infty \le B_0, \\
&\max_{\|\alpha\|_1 = s} \sup_{x \ne y} \frac{|\partial^\alpha g(x) - \partial^\alpha g(y)|}{\|x - y\|_2^r} \le B_0\},
\end{aligned}$$

其中 $\partial^\alpha = \partial^{\alpha_1} \cdots \partial^{\alpha_d}$，其中 $\alpha = (\alpha_1, \cdots \ \alpha_d)^\top \in \mathbb{N}_0^d$ 且 $\|\alpha\|_1 = \sum_{i=1}^d |\alpha_i|$.

根据 Jiao 等人 (2023) 关于前馈神经网络结构近似误差的定理 3.3，以及 Bartlett 等人 (2019年关于覆盖数界的定理 3 和定理 7，我们可以从定理3.1 中得出以下推论3.1：

推论 *3.1* 给定 $H\ddot{o}lder$ 光滑函数 $g^* \in H^\beta([0,1]^d, B_0)$，对任意 $D \in \mathbb{N}^+$、$W \in \mathbb{N}^+$，在定理 *3.1* 的条件下，以及 *Jiao* 等人（*2023*）中定理 *3.3* 的条件和 *Bartlett* 等人（*2019*年）中定理 *3* 与定理 *7* 的条件下，若采用 *ReLU* 激活函数的前馈神经网络的宽度 $W = c(\lfloor\beta\rfloor+1)^2 d^{\lfloor\beta\rfloor+1} W\lceil \log_2(8W)\rceil$ 且深度 $D = c(\lfloor\beta\rfloor+1)^2 D\lceil \log_2(8D)\rceil$且 $\lambda_1 \preceq n^{-1}SD\log(S)/d + (WD)^{-4\beta/d}/d$且 $\lambda_2 \preceq n^{-1}S\log(S)/W^2 + (WD)^{-4\beta/d}/(W^2D)$，则

$$\begin{aligned}\mathbb{E}[d((\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\theta}}),\boldsymbol{\Theta})] &\preceq (\log^2 n)n^{-1}\mathcal{S}\mathcal{D}\log(\mathcal{S}) + (WD)^{-4\beta/d}(\log^2 n).\\ \mathbb{E}\|\hat{\boldsymbol{\alpha}}_c\|_2^2 &\preceq (\log^2 n)n^{-1}\mathcal{S}\mathcal{D}\log(\mathcal{S}) + (WD)^{-4\beta/d}(\log^2 n).\end{aligned}$$

为便于阅读，文中还在引理和附录中列出了 Jiao 等人（2023）的定理 3.3，以及 Bartlett 等人（2019）的定理 3 和定理 7。在推论 3.1 中，第一项来自 G 的覆盖数，其上界由其 VC 维数给出：$\log N_{2n}(n^{-1}, \|\cdot\|_\infty, G|_{\boldsymbol{x}}) = O(SD\log(S/n^{-1}))$（Bartlett 等人，2019），其中 S 和 D 分别是参数总数和隐藏层数。第二项则来自 Jiao 等人（2023）的逼近结果，即 $\left\|g^* - g_{\mathcal{G}}^*\right\|_\infty \le 18B_0(\lfloor\beta\rfloor+1)^2 d^{\lfloor\beta\rfloor+\max\{\beta,1\}/2}(WD)^{-2\beta/d}$and $\mathbb{E}(R(g_{\mathcal{G}}^*) - R(g^*)) \simeq \mathbb{E}|g_{\mathcal{G}}^* - g^*|^2$，其中 $A \simeq B$ 表示 $A \preceq B$ 和 $B \preceq A$。

推论 *3.2* 假设推论 *3.1* 的条件成立。若 $\lambda_1 \preceq n^{-1}SD\log(S)/d + (WD)^{-4\beta/d}/d$且 $\lambda_2 \preceq n^{-1}S\log(S)/W^2 + (WD)^{-4\beta/d}/(W^2D)$，则对于任意 $j = 1,\cdots d$和$k = d+1,\cdots p$，有

$$Pr(\{|\hat{\alpha}_j| \ge \tau_d\} \cap \{|\hat{\alpha}_k| \le \tau_d\}) \to 1,$$

其中 τ_d 在条件（*C1*）中定义。

4 计算

优化复合目标函数

$$\mathcal{J}(g,\boldsymbol{\alpha}) = \frac{1}{n}\sum_{i=1}^{n}\Big\{Y_i - g(\boldsymbol{\alpha}\odot\boldsymbol{X}_i)\Big\}^2 + \lambda_1\|\boldsymbol{\alpha}\|_1 + \lambda_2\sum_{l=1}^{\mathcal{D}}(\|\boldsymbol{W}_l - \boldsymbol{I}\|_1 + |c_l|) \tag{11}$$

具有挑战性，因为 $\boldsymbol{\alpha}$ 和每个 $\boldsymbol{W}_l - \boldsymbol{I}$ 上的 L_1 惩罚项在零点处不可微，但我们需要精确的零值来识别无关输入和冗余层。因此，我们采用周期性硬阈值化（截断）。通过这些梯度下降步骤进行连续收缩，确保较小的系数趋近于零，同时每隔 K 次迭代通过截断强制得到精确的零值。

在截断步骤中，任何 α_j 满足 $|\alpha_j| \le \tau_1$ 的都会被设为零，而任何满足 l 条件的 $\|\boldsymbol{W}_l - \boldsymbol{I}\|_1 \le \tau_2$ 层都会通过重置 $\boldsymbol{W}_l = \boldsymbol{I}$而被折叠。阈值 τ_1 根据验证性能选定；我们遵循 Scardapane 等人 (2017) 的做法，将 $\tau_2 = 10^{-2}$ 固定用于合成数据研究，将 10^{-3} 固定用于真实数据。通过将次梯度收缩与周期性截断相结合，该算法实现了稳定收敛，在 $\boldsymbol{\alpha}$ 和网络深度方面均具有精确稀疏性，并且完全兼容 GPU 加速训练流水线。

算法 2： 截断过程

输入： 门控向量 $\boldsymbol{\alpha}$；权重矩阵 $\{\boldsymbol{W}_l\}$；阈值 τ_1，τ_2

输出： $\boldsymbol{\alpha}_{\text{trunc}}$，$\{\boldsymbol{W}_l^{\text{trunc}}\}$

$\boldsymbol{\alpha}_{\text{trunc}} \leftarrow \boldsymbol{\alpha}$

对于 $j = 1$ **到** d，**执行**

  如果 $|\alpha_j| \le \tau_1$，**则**

    $\alpha_j \leftarrow 0$

对于 $l = 1$ **到** D，**执行**

  如果 $\|\boldsymbol{W}_l - \boldsymbol{I}\|_1 \le \tau_2$ **那么**

    $\boldsymbol{W}_l \leftarrow \boldsymbol{I}$

返回 $\boldsymbol{\alpha}_{\text{trunc}}$，$\{\boldsymbol{W}_l^{\text{trunc}}\}$

5 模拟

我们开展模拟研究，以评估所提出的基于神经网络的变量选择方法的经验性能。具体而言，我们评估该方法识别真实活跃预测变量集合的能力，并将其与几种广泛使用的方法进行比较：贝叶斯核机器回归（BKMR）、SHapley Additive exPlanations（SHAP）和 DeepLIFT。评估重点在于复杂非线性模型中稀疏信号下的变量选择准确度。

5.1 变量选择的一致性

数据由非线性模型生成

$$Y = X_0^3 \left(X_2^2 + X_5\right) \; - \; |X_7| \cos(X_8) \; + \; \varepsilon,$$

其中 $\varepsilon \sim N$ 在 (0,1) 内，每个预测变量 $X_j \sim N$ 对 $j = 0, \ldots,$199 也独立地取值于 (0, 1)。只有五个坐标 {0,2,5,7,8} 进入真实模型，其余 $d - 5 = 195$ 个变量

均为噪声。我们取 $n = 1000$ 个样本，并将整个模拟重复 100 次，以评估变异性。

我们将本方法与五种既有方法进行比较。贝叶斯核机器回归（BKMR）（Bobb等人 2015）拟合高斯过程型模型，并使用后验包含概率进行变量选择。SHapley Additive exPlanations（SHAP）（Lundberg& Lee 2017）在预训练神经网络上事后计算 Shapley值，以对特征重要性进行排序。DeepLIFT（Shrikumar 等人 2017）是一种基于反向传播的归因方法，将激活与参考值进行比较，并为每个输入分配贡献得分。GLNN（Scardapane 等人 2017）对神经网络应用组 Lasso 正则化，对各组权重进行惩罚。深度特征选择（DPS）（Li 等人 2016，Chen 等人 2021）通过嵌入深度架构中的选择层纳入稀疏性约束。对于SHAP和DeepLIFT，我们采用与本方法相同的网络架构，但不附加自适应阈值或深度调整；而对于GLNN和DPS，我们采用其原始论文中指定的架构。对于除BKMR和所提出方法之外的所有竞争方法，均按权重大小选取前五个特征。选择准确率由以下指标量化：

$$\text{Precision} = \frac{|\hat{S}_n \cap S^\star|}{|\hat{S}_n|}, \quad \text{Recall} = \frac{|\hat{S}_n \cap S^\star|}{|S^\star|}, \quad F_1 = \frac{2\,\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}.$$

表1 报告了 100 次模拟重复中这些指标的均值和标准差。

表1：各竞争方法的变量选择性能。

方法	精确率	召回率	F1 分数
所提方法	0.927 (0.150)	0.880 (0.183)	0.879 (0.135)
BKMR	0.025 (0.000)	1.000 (0.000)	0.049 (0.000)
SHAP	0.780 (0.060)	0.780 (0.060)	0.780 (0.060)
DeepLIFT	0.720 (0.133)	0.720 (0.133)	0.720 (0.133)
GLNN	0.460 (0.100)	0.460 (0.100)	0.460 (0.100)
DPS	0.244 (0.083)	0.244 (0.083)	0.244 (0.083)

如表1所示，我们提出的方法取得了 F_1 约 0.88 的得分，显著优于其他竞争方法。BKMR 尽管召回率达到完美水平，但精确率极低，因此总体选择表现不佳（$F_1 \approx 0.05$），表明存在严重的过度选择。SHAP 和 DeepLIFT 的选择准确率中等，但逊于我们提出的方法，这可能是因为它们采用事后归因方式，在训练过程中忽略了结构化稀疏性。GLNN 和 DPS 的表现明显较弱，反映出它们在深度有限或稀疏性约束过于严格的情况下，难以选择高度相关的非线性特征。这些发现凸显了将自适应稀疏性约束和自动深度选择直接融入神经网络训练过程的优势，从而在高维非线性建模场景中显著提升变量选择准确度和可解释性。

5.2 选择后推断

在证明我们的变量选择程序具有一致性后，我们考察其对选择后推断的影响。为此，我们检验原假设

$$H_0 : \ E(Y \mid X_1, Z) = E(Y \mid Z),$$

其中， Z 表示作为条件的协变量集合。在“无选择”情景下， Z包含除 X_1之外的所有预测变量；在选择后情景下， Z 仅包含变量选择程序保留的特征（不包括 X_1）。随后，我们在两种情景下，对推断数据 D_2 应用第2.3 节所述的基于置换的检验，并在表2中报告经验第一类错误率。

表2：在样本量为 $n = 1000$ 且进行 500 次模拟重复时，对原假设 H_0：$E(Y \mid X_1,\ Z) = E(Y \mid Z)$ 进行检验所得的经验第一类错误率。结果展示了在推断前是否进行变量选择时的推断结果。

推断设置	第一类错误
不进行选择	0.13
选择后	0.01

表2 表明，未进行事先选择的推断会将第一类错误率抬高至 13%，而以数据驱动子集为条件则可将其恢复至名义水平 1%。这证实，即使在特征选择之后，我们的选择后推断程序仍能维持有效的错误率。

6 真实数据应用

真实数据分析采用早期生命环境毒物暴露（ELEMENT）项目的数据。ELEMENT由1994年至2005年在墨西哥城招募的三个母子队列组成。共有2,000多名女性及其子女入组，并从孕期随访至青春期。我们使用这些数据评估一种具有统计推断功能的非参数神经网络机器回归（NNMR）方法。我们的主要目标是确定哪些产前暴露与婴儿

生长的关联最强。婴儿生长通过考虑年龄和性别的世界卫生组织人体测量学Z分数进行衡量。测量在出生至五岁期间定期进行，同时也在胆固醇子研究和P20随访期间进行。分析纳入了母亲和儿童的高维饮食模式面板，并按食物类别分组。我们剔除Z分数缺失或不合理的记录，并在建模前对所有暴露变量进行标准化。

我们评估了神经网络机器回归（NNMR）方法在高维饮食数据中进行变量选择的性能，并以婴儿生长z评分作为结果变量。我们将NNMR与第5节提到的竞争方法进行了比较。所有方法均应用于一个包含儿童和母亲膳食暴露的数据集，并通过在独立的训练数据集和测试数据集上计算的赤池信息准则（AIC）评估性能。

表3：应用于ELEMENT饮食暴露数据的六种竞争方法的平均AIC和计算时间（秒）。报告值为均值，括号内为标准差，基于100次重复的测试集计算得出。

方法	AIC	时间（秒）
NNMR	2164.98（23.66）	6.74（1.20）
BKMR	2230.27 (23.71)	78.21 (20.31)
SHAP	2166.99 (25.33)	87.01 (8.86)
DeepLIFT	2162.45 (23.33)	3.66 (0.56)
组套索	2195.56 (41.16)	2.68 (0.43)
DFS	2167.62 (23.36)	7.69 (1.59)

表 3 显示，与 DeepLIFT 相比，NNMR 的 AIC 值和计算时间具有竞争力；其表现优于 SHAP、DFS 和组套索，并且显著

优于 BKMR。BKMR 表现较差，反映出它倾向于选择规模过大的模型，通常几乎纳入所有变量。DeepLIFT 和 SHAP 的预测误差与 NNMR 相当，因为它们采用相同的底层模型，但 DeepLIFT 倾向于选择最少的变量。组套索的表现不稳定，有时几乎选择所有变量，有时只选择一两个变量。相比之下，NNMR 在不同数据划分间的 AIC 波动较小，体现出稳定的泛化能力。

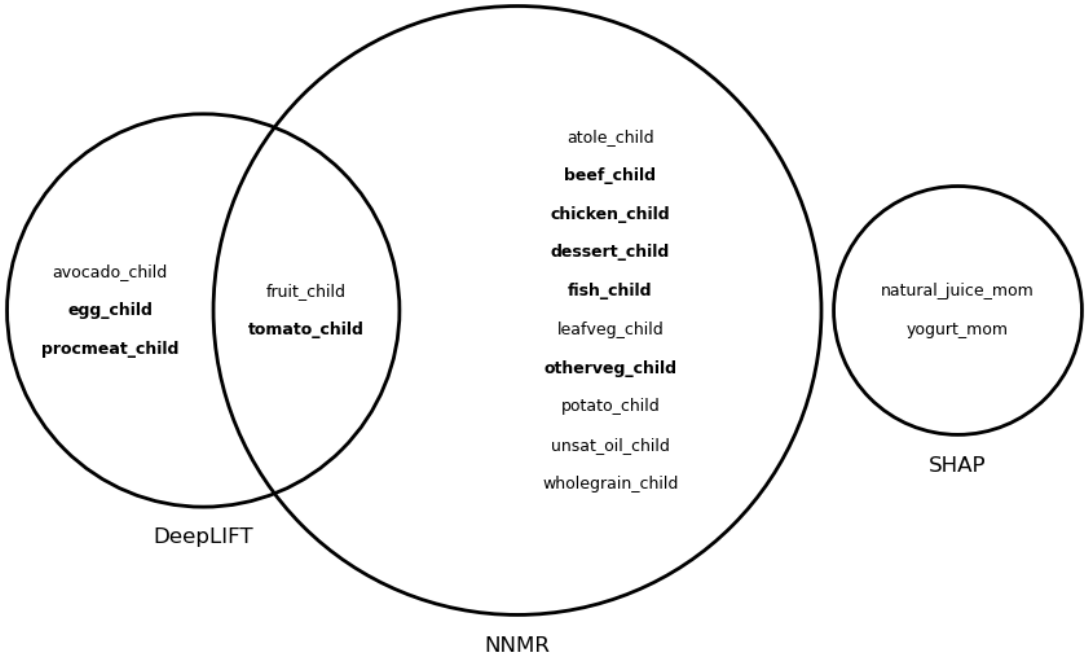

图2：NNMR、SHAP 和 DeepLIFT 在 ELEMENT研究 中选出的膳食暴露的维恩图。图中显示的是在 100 次重复中选择频率超过 20% 的变量。粗体变量表示应用选择后推断程序后仍具有统计学显著性的变量。

图 2 通过维恩图总结了变量选择结果，对比 NNMR、SHAP 和 DeepLIFT。为提高可解释性，我们仅展示在 100 次重复中选择频率超过 20% 的变量。完整结果见补充材料

材料。与先前观测结果一致，BKMR 几乎选择了所有暴露因素，而组套索的表现不稳定。DFS 仅选择了 `milk_child`，而 SHAP 强调了母亲的食物组。相比之下，NNMR 和 DeepLIFT 一致地突出显示了儿童膳食暴露，其选择频率更高且更为一致。

随后，我们将第2.3 节中的推断程序应用于 NNMR、DFS、DeepLIFT 和 SHAP 所识别变量的并集，并使用未用于变量筛选的推断数据划分。图2 中以红色标出的变量被发现是生长z评分的统计显著预测变量，其中大多数由 NNMR 优先选出。

推断后仅由 NNMR 保留的五个变量，与营养流行病学文献中的既有发现高度吻合。婴儿期摄入较多牛肉、锌和胆碱，与 3 至 5 岁儿童更好的抑制控制和注意力相关（Wilk 等人 2022年）。研究表明，在低收入环境中，营养敏感型农业项目中的鸡肉生产和消费有益于儿童生长（Passarelli 等人 2020）。鱼类是一种营养丰富的蛋白质来源，富含维生素、矿物质和必需脂肪酸（Yılmaz 等人 2018）。频繁摄入甜点与超重风险增加相关，而超重是生长z评分的一个重要决定因素（Barroso 等人 2016）。最后， `otherveg` 类别包含西葫芦、黄瓜和四季豆，体现了膳食多样性对儿童发育的重要性（Arimond 和 Ruel 2004，Thorne-Lyman 等人 2019年）。总的来说，这些结果凸显了 NNMR 所识别预测变量的生物学合理性及其与生命早期生长的相关性。

7 讨论

我们介绍了神经网络机器回归（NNMR），这是一种集成式神经网络框架，旨在高维数据分析中同时实现特征选择、非线性函数估计和严格的选择后推断。它通过嵌入一个

可训练门控层，并结合 L_1 正则化策略和自适应阈值机制，NNMR 可直接对输入特征和隐藏网络层施加稀疏性约束，从而生成高度紧凑且易于解释的模型。

我们的理论分析为相关特征集的一致恢复提供了保证，并在温和假设下给出了关于风险的极小极大最优上界。此外，我们整合的分样本置换检验方法能够稳健地控制 I 类错误率，降低高维推断中常见的假阳性风险。

大量模拟的实证结果表明，在变量选择的精确率、召回率和总体准确率（F_1 分数）方面，NNMR 优于 BKMR、SHAP、DeepLIFT、GLNN 和 DPS 等成熟方法。BKMR 以严重的过度选择为代价实现了较高的召回率，SHAP 和 DeepLIFT 则只能达到中等准确率；相比之下，NNMR 在识别真实预测变量和控制选择错误方面展现出明显优越的性能。GLNN 和 DPS 的准确率显著较低，凸显了传统稀疏性方法在捕捉复杂非线性关系方面的局限性。真实数据应用进一步验证了 NNMR 的实用价值，表明其能够在复杂的生物医学数据集中精准识别具有意义且可解释的预测变量。不同于在特征共线性下往往不稳定的事后归因方法，NNMR 内置的稀疏性能够准确识别真正不活跃的特征。对不必要层进行动态剪枝，还有助于实现适用于 GPU 加速的高效资源利用。

总之，NNMR 通过在统一的训练框架中融入统计上严谨的变量选择、自适应模型优化和有效推断，弥合了可解释性与预测建模能力之间的差距。这种整合式方法使 NNMR 成为现代高维统计学和深度学习应用中稳健且灵活的解决方案。

参考文献

Anthony, M., Bartlett, P. L., Bartlett, P. L. 等（1999年），神经网络学习：理论基础，第 9 卷，剑桥大学出版社，剑桥。

Arimond, M. & Ruel, M. T. (2004), '膳食多样性与儿童营养状况相关：来自11项人口与健康调查的证据', 《营养学杂志》134(10), 2579–2585.

Barroso, C. S., Roncancio, A., Moramarco, M. W., Hinojosa, M.B., Davila, Y. R., Mendias, E. & Reifsnider, E. (2016), '粮食安全、母亲喂养实践与儿童体重身长比', 《应用护理研究》29, 31–36.

Bartlett, P. L., Harvey, N., Liaw, C. & Mehrabian, A. (2019年), '分段线性神经网络的近乎紧致 VC 维数和伪维数界', 《机器学习研究杂志》20(1), 2285–2301.

Bobb, J. F., Valeri, L., ClausHenn, B., Christiani, D. C., Wright, R. O., Mazumdar, M., Godleski, J. J. & Coull, B. A. (2015), '用于估计多污染物混合物健康效应的贝叶斯核机器回归', 《生物统计学》16(3), 493–508.

Cai, Z., Lei, J. & Roeder, K.（2022年），'应用于基因组学数据的无模型预测检验', 《美国国家科学院院刊》119(34), e2205518119.

Chen, Y., Gao, Q., Liang, F. & Wang, X.（2021），'通过深度神经网络进行非线性变量选择'，《计算与图形统计学杂志》30(2), 484–492.

Cox, D. R.（1975年），'关于用于评估显著性水平的数据拆分的一则说明'，《生物统计学》第441–444页.

Fan, J. & Li, R. (2001), '通过非凹惩罚似然进行变量选择及其预言机性质', 《美国统计学会会刊》96(456),1348–1360.

Fan, J. & Lv, J. (2008), ‘超高维特征空间的确定性独立筛选’,《皇家统计学会会刊*B*辑：统计方法学》70(5), 849–911.

Hastie, T. J.（2 017），《广义加性模型》，*S*中的统计模型，第249–307页。

Hurvich, C. M. & Tsai, C.（1990），《模型选择对线性回归推断的影响》，《美国统计学家》44（3），214–217。

Jiao, Y., Shen, G., Lin, Y. & Huang, J.（2023），《近似流形上的深度非参数回归：带多项式前因子的非渐近误差界》，《统计年鉴》51（2），691–716。

Lei, J.（2020），《带置信度的交叉验证》，《美国统计学会会刊》115（532），1978–1997。

Li, Y., Chen, C.-Y. & Wasserman, W. W.（2016），《深度特征选择：理论及其在识别增强子和启动子中的应用》，《计算生物学杂志》23（5），322–336。

Lundberg, S. M. & Lee, S.-I. (2017), ‘一种统一的模型预测解释方法’，神经信息处理系统进展30。

Passarelli, S., Ambikapathi, R., Gunaratna, N. S., Madzorera, I., Canavan, C. R., Noor, A. R., Worku, A., Berhane, Y., Abdelmenan, S., Sibanda, S. 等（2020），‘鸡生产干预及额外的营养行为改变组成部分促进了埃塞俄比亚儿童生长：一项整群随机试验’，《营养学杂志》150(10), 2806–2817。

Scardapane, S., Comminiello, D., Hussain, A. & Uncini, A. (2017), ‘深度神经网络的组稀疏正则化’，《神经计算》241, 81–89。

Shrikumar, A.、Greenside, P. 与 Kundaje, A.（2017），通过传播激活差异学习重要特征

发表于《国际机器学习大会》，机器学习研究会论文集，第3145–3153页。

Sundararajan, M.、Taly, A. 与 Yan, Q.（2017），深度网络的公理化归因，载于《国际机器学习会议》，机器学习研究会论文集，第3319–3328页。

Tan, Z.、Zhou, L. 与 Lin, H.（2024），《具有最优输入维度及其自适应生成器架构的生成对抗学习》，*arXiv* 预印本 *arXiv:2405.03723* 。

Thorne-Lyman, A. L.、Shrestha, M.、Fawzi, W. W.、Pasqualino, M.、Strand, T. A.、Kvestad, I.、Hysing, M.、Joshi, N.、Lohani, M. 与 Miller, L. C.（2019），《尼泊尔远西地区的膳食多样性与儿童发育：一项队列研究》，《营养素》11（8），1799。

Tibshirani, R.（1996），《通过LASSO进行回归收缩与选择》，《皇家统计学会会刊 *B*辑：统计方法学》58（1），267–288。

Wilk, V. C., McGuire, M. K. & Roe, A. J. (2022年), '早期牛肉消费模式与1–5岁儿童的认知结果相关：一项探索性研究', 《营养素》14(21), 4497.

Yılmaz, E.,Aydın, M.,Yıldırım, A.& Şahin, P. (2018),'鱼肉消费在幼儿期对健康发育的重要性', 苏莱曼·德米雷尔大学埃伊尔迪尔水产学院学报14(4), 357–364.

附录

ELEMENT研究中的变量选择结果完整列表

表4 列出了所提出的神经网络机器回归（NNMR）、SHAP、DeepLIFT 和 DFS 方法在重复数据划分中各变量的完整列表及其选择频率。为提高可读性，SHAP 仅显示选择频率大于

0.05 的变量。由于 BKMR 和组套索始终选中几乎所有变量，其选择频率分别接近 1.0 和 0.45，因此为简洁起见，表中省略了它们的结果。这样的呈现方式突出了那些由筛选行为更具选择性的方法最稳定地识别为重要的变量。

表4：NNMR、SHAP（仅显示 ≥ 0.05）、DeepLIFT 和 DFS 在重复划分中选择频率非零的变量。BKMR 和组套索几乎选择了所有变量（频率为 ≈ 1.0 和 0.45），为简洁起见，此处省略。

特征	NNMR	SHAP	DeepLIFT	DFS
$atole_child$	0.38	–	–	–
$avocado_child$	0.15	–	0.25	–
$beef_child$	0.41	–	0.19	–
$chicken_child$	0.43	–	0.15	–
$chili_child$	0.09	–	–	–
$chips_child$	0.02	–	–	–
$corn_tortilla_child$	0.01	0.08	–	–
$corncob_child$	0.13	–	–	–
$corncob_mom$	–	0.06	–	–
$cruveg_child$	0.01	–	–	–
$dessert_child$	0.31	0.06	–	–
egg_child	0.16	–	0.30	–
$fish_child$	0.23	–	0.03	–
$fish_mom$	–	0.06	–	–

续见下页

特征	NNMR	SHAP	DeepLIFT	DFS
fruit_child	0.38	–	0.74	–
fruit_mom	–	0.07	–	–
hf_dairy_child	0.01	–	0.16	–
jam_child	0.01	–	–	–
leafveg_child	0.39	–	–	–
legumes_child	0.03	–	–	–
milk_child	0.02	–	0.19	1.00
natural_juice_mom	–	0.48	–	–
organmeat_child	0.16	–	0.03	–
otherveg_child	0.35	–	–	–
pork_child	0.04	–	0.09	–
potato_child	0.29	–	0.04	–
procmeat_child	0.02	0.06	0.23	–
refgrain_child	0.13	0.12	–	–
refgrain_mom	–	0.08	–	–
soup_child	0.18	–	–	–
sugar_beverages_child	0.05	0.07	–	–
sugar_beverages_mom	–	0.13	–	–
tomato_child	0.39	–	0.22	–
unsat_oil_child	0.43	–	–	–
wholegrain_child	0.26	–	–	–
yeveg_child	0.02	0.09	–	–

续见下页

特征	NNMR	SHAP	DeepLIFT	DFS
yogurt_child	0.01	–	0.15	–
yogurt_mom	–	0.32	–	–

定理的证明

记 $\mathcal{R}_n(g_{\boldsymbol{\alpha},\boldsymbol{\theta}}) = \frac{1}{n}\sum_{i=1}^{n}\Big(Y_i - g(\boldsymbol{X}_i)\Big)^2$，其中 $g \in \mathcal{G}(\boldsymbol{\alpha},\boldsymbol{\theta})$。对于样本量为 n 的任意独立同分布（i.i.d.）样本 $D_n = \{\boldsymbol{X}_i, Y_i\}_{i=1}^n$。定义

$$S(g_{\boldsymbol{\alpha},\boldsymbol{\theta}}, \boldsymbol{X}_i) = \Big(Y_i - g(\boldsymbol{X}_i)\Big)^2 - \Big(Y_i - g^*(\boldsymbol{X}_i)\Big)^2.$$

令 $D_n' = \{\boldsymbol{X}_i', Y_{i}\}$ 为另一个独立于 D_n 的样本，并记

$$L(g_{\boldsymbol{\alpha},\boldsymbol{\theta}}, \boldsymbol{X}_i') = \mathbb{E}_{D_n'}\left(S(g_{\boldsymbol{\alpha},\boldsymbol{\theta}}, \boldsymbol{X}_i')\right) - 2S(g_{\boldsymbol{\alpha},\boldsymbol{\theta}}, \boldsymbol{X}_i).$$

回忆一下，$g_{\boldsymbol{\theta}}(\boldsymbol{\alpha}\odot\cdot) \in \mathcal{G}(\boldsymbol{\alpha},\boldsymbol{\theta})$，并且

$$(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\theta}}) \in \arg\min \mathcal{L}_n(g_{\boldsymbol{\alpha},\boldsymbol{\theta}}) = \arg\min_{(\boldsymbol{\alpha},\boldsymbol{\theta})}\left[\mathcal{R}_n(g_{\boldsymbol{\alpha},\boldsymbol{\theta}}) + \lambda_1\|\boldsymbol{\alpha}\|_1 + \lambda_2\sum_{l=1}^{\mathcal{D}}\|\boldsymbol{W}_l - \boldsymbol{I}\|_1\right].$$

覆盖数。给定 G 的一个 δ-一致覆盖，我们用 $g_q, q = 1, \cdots, \mathsf{N}_{2n}$ 表示各球的中心，其中 $\mathsf{N}_{2n} = \sup_{\boldsymbol{x}} \mathsf{N}_{2n}(\delta, \|\cdot\|_\infty, \mathsf{G}|_{\boldsymbol{x}})$ 是在范数 $\|\cdot\|_\infty$ 下半径为 δ 的一致覆盖数。根据覆盖的定义，存在一个 q^*，使得 $\|g_{\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\theta}}} - g_{q^*}\|_\infty \le \delta$ on $\boldsymbol{x} \in (\boldsymbol{X}_1, \cdots \boldsymbol{X}_n, \boldsymbol{X}_1', \cdots \boldsymbol{X}_n')$。令 $A \preceq B$ 表示正数 c 的 $A \le cB$。

定理3.1的证明。设$(\check{\boldsymbol{\alpha}}, \check{\boldsymbol{\theta}}) \in \boldsymbol{\Theta}$ 满足

$$(\check{\boldsymbol{\alpha}}, \check{\boldsymbol{\theta}}) \in \arg\min_{(\boldsymbol{\alpha},\boldsymbol{\theta})\in\boldsymbol{\Theta}} \mathbb{E}\left[d\left(\left(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\theta}}\right), \boldsymbol{\Theta}\right)\right].$$

不失一般性，稍微滥用一下记号，将 $\sum_{l=1}^{\mathcal{D}}\|\boldsymbol{W}_l-\boldsymbol{I}\|_1$ 记作$\sum_{l=1}^{\mathcal{D}}(\|\boldsymbol{W}_l-\boldsymbol{I}\|_1+|c_l|)$，其中截距项被吸收到神经网络的权重矩阵 $\boldsymbol{W}_l$ 中。于是可得

$$\mathcal{R}_n(g_{\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\theta}}})+\lambda_1\|\hat{\boldsymbol{\alpha}}\|_1+\lambda_2\sum_{l=1}^{\mathcal{D}}\|\hat{\boldsymbol{W}}_l-\boldsymbol{I}\|_1\leq\mathcal{R}_n(g_{\check{\boldsymbol{\alpha}},\check{\boldsymbol{\theta}}})+\lambda_1\|\check{\boldsymbol{\alpha}}\|_1+\lambda_2\sum_{l=1}^{\mathcal{D}}\|\check{\boldsymbol{W}}_l-\boldsymbol{I}\|_1.$$

对于期望超额风险，我们有如下分解：

$$\begin{aligned}
&\mathbb{E}\Big(\mathcal{R}(g_{\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\theta}}})-\mathcal{R}(g_{\check{\boldsymbol{\alpha}},\check{\boldsymbol{\theta}}})\Big)\\
=\;&\mathbb{E}\left(\mathcal{R}(g_{\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\theta}}})-\mathcal{R}_n(g_{\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\theta}}})\right)+\mathbb{E}\left(\mathcal{R}_n(g_{\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\theta}}})-\mathcal{R}_n(g_{\check{\boldsymbol{\alpha}},\check{\boldsymbol{\theta}}})\right) \qquad (3)\\
&+\mathbb{E}\Big(\mathcal{R}_n(g_{\check{\boldsymbol{\alpha}},\check{\boldsymbol{\theta}}})-\mathcal{R}(g_{\check{\boldsymbol{\alpha}},\check{\boldsymbol{\theta}}})\Big)\\
\leq\;&I_1+\mathbb{E}_{D_n}\left[\lambda_1\left(\|\check{\boldsymbol{\alpha}}\|_1-\|\hat{\boldsymbol{\alpha}}\|_1\right)+\lambda_2\sum_{l=1}^{\mathcal{D}}\left(\|\check{\boldsymbol{W}}_l-\boldsymbol{I}\|_1-\|\hat{\boldsymbol{W}}_l-\boldsymbol{I}\|_1\right)\right]. \qquad (4)
\end{aligned}$$

注意，第一项具有如下上界：

$$\begin{aligned}
I_1\;:=\;&\mathbb{E}\left(\mathcal{R}(g_{\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\theta}}})-\mathcal{R}_n(g_{\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\theta}}})\right)\\
=\;&\mathbb{E}_{D_n}\left\{n^{-1}\sum_{i=1}^{n}\left[\mathbb{E}_{D_n'}\left(n^{-1}\sum_{i=1}^{n}\left(Y_i'-g_{\hat{\boldsymbol{\theta}}}(\hat{\boldsymbol{\alpha}}\odot\boldsymbol{X}_i')\right)^2\right)\right.\right. \qquad (5)\\
&\left.\left.-\left(Y_i-g_{\hat{\boldsymbol{\theta}}}(\hat{\boldsymbol{\alpha}}\odot\boldsymbol{X}_i)\right)^2\right]\right\}\\
=\;&\mathbb{E}_{D_n}\left\{n^{-1}\sum_{i=1}^{n}\left[\mathbb{E}_{D_n'}\left(n^{-1}\sum_{i=1}^{n}S(g_{\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\theta}}},\boldsymbol{X}_i')\right)-2S(g_{\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\theta}}},\boldsymbol{X}_i')\right]\right\}\\
&+\mathbb{E}_{D_n}\left\{n^{-1}\sum_{i=1}^{n}\left[\left(Y_i-g_{\hat{\boldsymbol{\theta}}}(\hat{\boldsymbol{\alpha}}\odot\boldsymbol{X}_i)\right)^2-\left(Y_i-g_{\check{\boldsymbol{\theta}}}(\check{\boldsymbol{\alpha}}\odot\boldsymbol{X}_i)\right)^2\right]\right\}\\
&+\mathbb{E}_{D_n}\left\{n^{-1}\sum_{i=1}^{n}S(g_{\check{\boldsymbol{\theta}},\check{\boldsymbol{\alpha}}},\boldsymbol{X}_i')\right\}\\
\leq\;&I_{11}+\mathbb{E}_{D_n}\left[\lambda_1\left(\|\check{\boldsymbol{\alpha}}\|_1-\|\hat{\boldsymbol{\alpha}}\|_1\right)+\lambda_2\sum_{l=1}^{\mathcal{D}}\left(\|\check{\boldsymbol{W}}_l-\boldsymbol{I}\|_1-\|\hat{\boldsymbol{W}}_l-\boldsymbol{I}\|_1\right)\right]\\
&+\left(\mathcal{R}(g_{\check{\boldsymbol{\alpha}},\check{\boldsymbol{\theta}}})-\mathcal{R}(g^*)\right). \qquad (6)
\end{aligned}$$

接下来，我们将给出 I_{11} 的上界，并通过截断和经典链式法来处理它

经验过程技巧。根据 $S(g_{\boldsymbol{\alpha},\,\boldsymbol{\theta}},\ \boldsymbol{X}_i)$ 的定义，我们有

$$\begin{aligned}
&\left|S(g_{\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\theta}}},\boldsymbol{X}_i')-S(g_{q^*},\boldsymbol{X}_i')\right| \\
=\ &\left|2Y_i\Big(g_{\hat{\boldsymbol{\theta}}}(\hat{\boldsymbol{\alpha}}\odot\boldsymbol{X}_i)-g_{q^*}(\boldsymbol{X}_i)\Big)+\Big(g^2_{\hat{\boldsymbol{\theta}}}(\hat{\boldsymbol{\alpha}}\odot\boldsymbol{X}_i)-g^2_{q^*}(\boldsymbol{X}_i)\Big)\right| \\
\leq\ &(2|Y_i|+2\mathcal{B})\delta,
\end{aligned}$$

and

$$\begin{aligned}
&\left|L(g_{\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\theta}}},\boldsymbol{X}_i')-L(g_{q^*},\boldsymbol{X}_i')\right| \\
\leq\ &\mathbb{E}_{D_n'}\left(\left|S(g_{\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\theta}}},\boldsymbol{X}_i')-S(g_{q^*},\boldsymbol{X}_i')\right|\right)+2\left|S(g_{\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\theta}}},\boldsymbol{X}_i')-S(g_{q^*},\boldsymbol{X}_i')\right| \\
\leq\ &3(\mathbb{E}|Y_i|+\mathcal{B}+|Y_i|)\delta.
\end{aligned}$$

于是可得

$$\mathbb{E}_{D_n}\left(n^{-1}\sum_{i=1}^{n}\left|L(g_{\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\theta}}},\boldsymbol{X}_i')-L(g_{q^*},\boldsymbol{X}_i')\right|\right)\leq 6(\mathbb{E}|Y_i|+\mathcal{B})\delta,$$

由此得到

$$\mathbb{E}_{D_n}\left(n^{-1}\sum_{i=1}^{n}L(g_{\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\theta}}},\boldsymbol{X}_i')\right)\leq\mathbb{E}_{D_n}\left(n^{-1}\sum_{i=1}^{n}L(g_{q^*},\boldsymbol{X}_i')\right)+6(\mathbb{E}|Y_i|+\mathcal{B})\delta. \tag{7}$$

设 $0<\beta_n$ 为一个正数，它可能依赖于样本量 n。若 $Y\leq\beta_n$，记为 $T_{\beta_n}Y=Y$；否则记为 $T_{\beta_n}Y=\beta_n$。定义函数 $g^*_{\beta_n}$ 为

$$g^*_{\beta_n}(\boldsymbol{x})=\arg\min_{g:\|g\|_\infty<\mathcal{B}}\mathbb{E}\left[\Big(T_{\beta_n}Y_i-g(\boldsymbol{X}_i)\Big)^2\mid\boldsymbol{X}_i=\boldsymbol{x}\right].$$

对于任意 $g\in\mathcal{G}(\boldsymbol{\alpha},\ \boldsymbol{\theta})$，令 $S_{\beta_n}(g_{\boldsymbol{\alpha},\,\boldsymbol{\theta}},\ \boldsymbol{X}_i')=\Big(T_{\beta_n}Y_i-g_{\boldsymbol{\theta}}(\boldsymbol{\alpha}\odot\boldsymbol{X}_i)\Big)^2-\Big(T_{\beta_n}Y_i-g^*_{\beta_n}(\boldsymbol{X}_i)\Big)^2$。那么，

我们有

$$\begin{aligned}
\mathbb{E}\Big(S(g_{\alpha,\boldsymbol{\theta}}, \boldsymbol{X}_i')\Big) &= \mathbb{E}\Big(S_{\beta_n}(g_{\alpha,\boldsymbol{\theta}}, \boldsymbol{X}_i')\Big) + \mathbb{E}\left[\Big(Y_i - g_{\boldsymbol{\theta}}(\boldsymbol{\alpha}\odot\boldsymbol{X}_i)\Big)^2 - \Big(T_{\beta_n}Y_i - g_{\boldsymbol{\theta}}(\boldsymbol{\alpha}\odot\boldsymbol{X}_i)\Big)^2\right] \\
&\quad -\mathbb{E}\left[\Big(Y_i - g^*(\boldsymbol{X}_i)\Big)^2 - \Big(T_{\beta_n}Y_i - g^*(\boldsymbol{X}_i)\Big)^2\right] \\
&\quad -\mathbb{E}\left[\Big(T_{\beta_n}Y_i - g^*(\boldsymbol{X}_i)\Big)^2 - \Big(T_{\beta_n}Y_i - g^*_{\beta_n}(\boldsymbol{X}_i)\Big)^2\right] \\
&\leq \mathbb{E}\Big(S_{\beta_n}(g_{\alpha,\boldsymbol{\theta}}, \boldsymbol{X}_i')\Big) + \mathbb{E}\Big(Y_i^2 - (T_{\beta_n}Y_i)^2 - 2(Y_i - T_{\beta_n}Y_i)g_{\boldsymbol{\theta}}(\boldsymbol{\alpha}\odot\boldsymbol{X}_i)\Big) \\
&\quad -\mathbb{E}\Big(Y_i^2 - (T_{\beta_n}Y_i)^2 - 2(Y_i - T_{\beta_n}Y_i)g^*(\boldsymbol{X}_i)\Big) \\
&\leq \mathbb{E}\Big(S_{\beta_n}(g_{\alpha,\boldsymbol{\theta}}, \boldsymbol{X}_i')\Big) + 4\mathbb{E}(|Y_i|I(Y_i > \beta_n))\mathcal{B},
\end{aligned}$$

and

$$\begin{aligned}
\mathbb{E}\Big(S_{\beta_n}(g_{\alpha,\boldsymbol{\theta}}, \boldsymbol{X}_i')\Big) &= \mathbb{E}\Big(S(g_{\alpha,\boldsymbol{\theta}}, \boldsymbol{X}_i')\Big) - \mathbb{E}\left[\Big(Y_i - g_{\boldsymbol{\theta}}(\boldsymbol{\alpha}\odot\boldsymbol{X}_i)\Big)^2 - \Big(T_{\beta_n}Y_i - g_{\boldsymbol{\theta}}(\boldsymbol{\alpha}\odot\boldsymbol{X}_i)\Big)^2\right] \\
&\quad -\mathbb{E}\left[\Big(T_{\beta_n}Y_i - g^*_{\beta_n}(\boldsymbol{X}_i)\Big)^2 - \Big(Y_i - g^*_{\beta_n}(\boldsymbol{X}_i)\Big)^2\right] \\
&\quad -\mathbb{E}\left[\Big(Y_i - g^*_{\beta_n}(\boldsymbol{X}_i)\Big)^2 - \Big(Y_i - g^*(\boldsymbol{X}_i)\Big)^2\right] \\
&\leq \mathbb{E}\Big(S(g_{\alpha,\boldsymbol{\theta}}, \boldsymbol{X}_i')\Big) - \mathbb{E}\Big(Y_i^2 - (T_{\beta_n}Y_i)^2 - 2(Y_i - T_{\beta_n}Y_i)g_{\boldsymbol{\theta}}(\boldsymbol{\alpha}\odot\boldsymbol{X}_i)\Big) \\
&\quad +\mathbb{E}\Big(Y_i^2 - (T_{\beta_n}Y_i)^2 - 2(Y_i - T_{\beta_n}Y_i)g^*_{\beta_n}(\boldsymbol{X}_i)\Big) \\
&\leq \mathbb{E}\Big(S(g_{\alpha,\boldsymbol{\theta}}, \boldsymbol{X}_i')\Big) + 4\mathbb{E}(|Y_i|I(Y_i > \beta_n))\mathcal{B},
\end{aligned}$$

这导致

$$\left|\mathbb{E}\Big(S(g_{\alpha,\boldsymbol{\theta}}, \boldsymbol{X}_i') - S_{\beta_n}(g_{\alpha,\boldsymbol{\theta}}, \boldsymbol{X}_i')\Big)\right| \leq 4\mathbb{E}(|Y_i|I(Y_i > \beta_n))\mathcal{B}.$$

然后,

$$\begin{aligned}
&\left|\mathbb{E}_{D_n}\left[\frac{1}{n}\sum_{i=1}^{n}\Big(L(g_{q^*}, \boldsymbol{X}_i') - L_{\beta_n}(g_{q^*}, \boldsymbol{X}_i')\Big)\right]\right| \\
\leq\ &\Big|\mathbb{E}_{D_n'}\Big(S(g_{q^*}, \boldsymbol{X}_i') - S_{\beta_n}(g_{q^*}, \boldsymbol{X}_i')\Big)\Big| + 2\Big|\mathbb{E}_{D_n}\Big(S(g_{q^*}, \boldsymbol{X}_i') - S_{\beta_n}(g_{q^*}, \boldsymbol{X}_i')\Big)\Big| \\
\leq\ &12\mathbb{E}(|Y_i|I(Y_i > \beta_n))\mathcal{B}. \qquad (8)
\end{aligned}$$

另一方面，对于任意的 $g \in \mathcal{G}(\boldsymbol{\alpha}, \boldsymbol{\theta})$，我们有

$$|S_{\beta_n}(g, \boldsymbol{X}_i')| \le 5(\beta_n + \mathcal{B})^2,$$

$$\sigma_S^2(g) := \mathrm{Var}(S_{\beta_n}(g, \boldsymbol{X}_i')) \le \mathbb{E}\{S_{\beta_n}^2(g, \boldsymbol{X}_i')\} \le 5(\beta_n + \mathcal{B})^2 \mathbb{E}(S_{\beta_n}(g, \boldsymbol{X}_i')).$$

根据伯恩斯坦不等式，对于任意 $t > 0$，令 $u = t/2 + \sigma_S^2(g)/(10(\beta_n + \mathcal{B})^2)$，则有

$$\begin{aligned}
& P\left\{ n^{-1} \sum_{i=1}^n L_{\beta_n}(g_q, \boldsymbol{X}_i') > t \right\} \\
= \;& P\left\{ \mathbb{E}_{D_n'}\left(S_{\beta_n}(g_q, \boldsymbol{X}_i')\right) - n^{-1} 2 \sum_{i=1}^n S_{\beta_n}(g_q, \boldsymbol{X}_i') > t \right\} \\
= \;& P\left\{ \mathbb{E}_{D_n'}\{S_{\beta_n}(g_q, \boldsymbol{X}_i')\} - \frac{1}{n}\sum_{i=1}^n S_{\beta_n}(g_q, \boldsymbol{X}_i') > \frac{t}{2} + \frac{1}{2}\mathbb{E}_{D_n'}\{S_{\beta_n}(g_q, \boldsymbol{X}_i')\} \right\} \\
\le \;& P\left\{ \mathbb{E}_{D_n'}\{S_{\beta_n}(g_q, \boldsymbol{X}_i')\} - \frac{1}{n}\sum_{i=1}^n S_{\beta_n}(g_q, \boldsymbol{X}_i') > \frac{t}{2} + \frac{1}{2}\frac{\sigma_S^2(g)}{5(\beta_n + \mathcal{B})^2} \right\} \\
\le \;& \exp\left(-\frac{nu^2}{2\sigma_S^2(g) + 20u(\beta_n + \mathcal{B})^2/3}\right) \\
\le \;& \exp\left(-\frac{nu^2}{20u(\beta_n + \mathcal{B})^2 + 20u(\beta_n + \mathcal{B})^2/3}\right) \\
\le \;& \exp\left(-\frac{1}{20 + 20/3}\frac{nu}{(\beta_n + \mathcal{B})^2}\right) \\
\le \;& \exp\left(-\frac{1}{40 + 40/3}\frac{nt}{(\beta_n + \mathcal{B})^2}\right) \\
= \;& \exp\left(-\frac{Cnt}{(\beta_n + \mathcal{B})^2}\right).
\end{aligned}$$

这得到 $n^{-1}\sum_{i=1}^n L_{\beta_n}(g_{q^*}, \boldsymbol{X}_i')$ 的尾概率界，即

$$P\left\{ n^{-1} \sum_{i=1}^n L_{\beta_n}(g_{q^*}, \boldsymbol{X}_i') > t \right\} \le 2\mathcal{N}_{2n} \exp\left(-\frac{Cnt}{(\beta_n + \mathcal{B})^2}\right).$$

于是对于 $a_n > 0$,

$$\begin{aligned}
\mathbb{E}_{D_n}\left[\frac{1}{n}\sum_{i=1}^n L_{\beta_n}(g_{q^*}, \boldsymbol{X}_i')\right] &\le a_n + \int_{a_n}^\infty P\left\{\frac{1}{n}\sum_{i=1}^n L_{\beta_n}(g_{q^*}, \boldsymbol{X}_i') > t\right\} dt \\
&\le a_n + \int_{a_n}^\infty 2\mathcal{N}_{2n} \exp\left(-\frac{Cnt}{(\beta_n + \mathcal{B})^2}\right) dt \\
&\le a_n + 2\mathcal{N}_{2n} \exp\left(-a_n \frac{Cn}{(\beta_n + \mathcal{B})^2}\right) \frac{(\beta_n + \mathcal{B})^2}{Cn}.
\end{aligned}$$

取 $a_n = \log 2\mathcal{N}_{2n} \frac{(\beta_n + \mathcal{B})^2}{Cn}$，上述不等式可得

$$\mathbb{E}_{D_n}\left[\frac{1}{n}\sum_{i=1}^n L_{\beta_n}(g_{q^*}, \boldsymbol{X}_i')\right] \le \frac{C(\beta_n + \mathcal{B})^2(\log 2\mathcal{N}_{2n} + 1)}{n}. \tag{9}$$

结合不等式（7）、（8）和（9），可得

$$\begin{aligned} I_{11} &= \mathbb{E}_{D_n}\left(n^{-1}\sum_{i=1}^{n} L(g_{\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\theta}}}, \boldsymbol{X}_i')\right) \\ &\leq \mathbb{E}_{D_n}\left(n^{-1}\sum_{i=1}^{n} L(g_{q^*}, \boldsymbol{X}_i')\right) + 6\left(\mathbb{E}|Y_i| + \mathcal{B}\right)\delta \\ &\leq \mathbb{E}_{D_n}\left(n^{-1}\sum_{i=1}^{n} L_{\beta_n}(g_{q^*}, \boldsymbol{X}_i')\right) + 6\left(\mathbb{E}|Y_i| + \mathcal{B}\right)\delta + 12\mathbb{E}(|Y_i|I(Y_i > \beta_n))\mathcal{B} \\ &\leq \frac{C(\beta_n + \mathcal{B})^2(\log 2\mathcal{N}_{2n} + 1)}{n} + 6\left(\mathbb{E}|Y_i| + \mathcal{B}\right)\delta + 12\mathbb{E}(|Y_i|I(Y_i > \beta_n))\mathcal{B}. \end{aligned}$$

令 $\beta_n = \log n$ 以及 $\delta = n^{-1}$。在条件 (C2) 和 (C3) 下，利用上述不等式，我们得到

$$I_{11} \preceq \frac{C\log^2 n \log \mathcal{N}_{2n}}{n}. \tag{10}$$

然后，结合不等式 (3)、(5) 和 (10)，利用 $\mathbb{E}\Big[d\Big((\boldsymbol{\alpha},\boldsymbol{\theta})\ \boldsymbol{\Theta}\Big)\Big] \leq c\mathbb{E}\Big(\mathcal{R}(g_{\boldsymbol{\alpha},\boldsymbol{\theta}})\ -\mathcal{R}(g_{\check{\boldsymbol{\alpha}},\check{\boldsymbol{\theta}}})\Big)$ 这一条件，我们可以得到

$$\begin{aligned} &\mathbb{E}\Big[d\Big((\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\theta}}),\boldsymbol{\Theta}\Big)\Big] = \mathbb{E}\Big(\big\|(\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\theta}}) - (\check{\boldsymbol{\alpha}},\check{\boldsymbol{\theta}})\big\|_2^2\Big) \ \leq\ c\,\mathbb{E}\Big\{\mathcal{R}\Big(g_{\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\theta}}}\Big) - \mathcal{R}\Big(g_{\check{\boldsymbol{\alpha}},\check{\boldsymbol{\theta}}}\Big)\Big\} \\ &\leq c\,I_1 + c\,\mathbb{E}\Big[\lambda_1\Big(\|\check{\boldsymbol{\alpha}}\|_1 - \|\hat{\boldsymbol{\alpha}}\|_1\Big) + \lambda_2\sum_{l=1}^{\mathcal{D}}\Big(\|\check{\boldsymbol{W}}_l - \boldsymbol{I}\|_1 - \|\hat{\boldsymbol{W}}_l - \boldsymbol{I}\|_1\Big)\Big] \\ &\leq c\,I_{11} + c\,\lambda_1\,\mathbb{E}[\|\check{\boldsymbol{\alpha}} - \hat{\boldsymbol{\alpha}}\|_1] + c\,\lambda_2\sum_{l=1}^{\mathcal{D}}\mathbb{E}\Big[\|\check{\boldsymbol{W}}_l - \hat{\boldsymbol{W}}_l\|_1\Big] \\ &\leq c\,\frac{C\ \log^2 n\ \log \mathcal{N}_{2n}(n^{-1}, \|\cdot\|_\infty, \mathcal{G}|_{\boldsymbol{x}})}{n} + c\,\lambda_1\sqrt{d}\,\mathbb{E}[\|\check{\boldsymbol{\alpha}} - \hat{\boldsymbol{\alpha}}\|_2] \\ &\quad + c\,\lambda_2\sum_{l=1}^{\mathcal{D}}\mathcal{W}\,\mathbb{E}\Big[\|\check{\boldsymbol{W}}_l - \hat{\boldsymbol{W}}_l\|_F\Big], \end{aligned} \tag{11}$$

为了使上述不等式的右侧达到最小值，我们有

$$\begin{aligned} \lambda_1^2 &\leq \frac{c\log^2 n\ \log \mathcal{N}_{2n}}{nd} + \frac{c\Big(\mathcal{R}(g_{\check{\boldsymbol{\alpha}},\check{\boldsymbol{\theta}}}) - \mathcal{R}(g^*)\Big)}{d}, \\ \lambda_2^2 &\leq \frac{c\log^2 n\ \log \mathcal{N}_{2n}}{n\mathcal{W}^2\mathcal{D}} + \frac{c\Big(\mathcal{R}(g_{\check{\boldsymbol{\alpha}},\check{\boldsymbol{\theta}}}) - \mathcal{R}(g^*)\Big)}{\mathcal{W}^2\mathcal{D}}, \end{aligned} \tag{12}$$

注意，利用杨氏不等式，我们有

$$\begin{aligned}\mathbb{E}\Big[c\lambda_1\sqrt{d}\,\|\check{\boldsymbol{\alpha}}-\hat{\boldsymbol{\alpha}}\|_2\Big] &\le \frac{1}{2}\mathbb{E}\Big[\|\check{\boldsymbol{\alpha}}-\hat{\boldsymbol{\alpha}}\|_2^2\Big] + \frac{c^2\lambda_1^2 d}{2}.\\ \mathbb{E}\Big[c\lambda_2\,\mathcal{W}\,\|\check{\boldsymbol{W}}_l-\hat{\boldsymbol{W}}_l\|_F\Big] &\le \frac{1}{2}\mathbb{E}\Big[\|\check{\boldsymbol{W}}_l-\hat{\boldsymbol{W}}_l\|_F^2\Big] + \frac{c^2\lambda_2^2\mathcal{W}^2}{2}.\end{aligned} \tag{13}$$

结合不等式（11）、（12）和（13），我们可以得到

$$\begin{aligned}&\mathbb{E}\Big[d\Big((\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\theta}}),\boldsymbol{\Theta}\Big)\Big]\\ &\le c\left\{\frac{\log^2 n\,\log\mathcal{N}_{2n}}{n} + \Big(\mathcal{R}(g_{\check{\boldsymbol{\alpha}},\check{\boldsymbol{\theta}}})-\mathcal{R}(g^*)\Big) + \lambda_1^2 d + \lambda_2^2\mathcal{W}^2\mathcal{D}\right\}.\\ &\le c\left\{\frac{\log^2 n\,\log\mathcal{N}_{2n}}{n} + \Big(\mathcal{R}(g_{\check{\boldsymbol{\alpha}},\check{\boldsymbol{\theta}}})-\mathcal{R}(g^*)\Big)\right\}.\end{aligned}$$

令 $(\boldsymbol{\alpha}_s^*,\boldsymbol{\theta}_s^*)\in$ 极小值点$_{\boldsymbol{\alpha}_s,\boldsymbol{\theta}_s}\mathrm{R}(g_{\boldsymbol{\alpha}_s,\boldsymbol{\theta}_s})=$ 极小值点$_{\boldsymbol{\alpha}_s,\boldsymbol{\theta}_s}\mathbb{E}\,(Y-g_{\boldsymbol{\theta}_s}(\boldsymbol{\alpha}_s\odot\boldsymbol{X}_s))^2$。定义$\boldsymbol{\Theta}_s=\{(\boldsymbol{\alpha}_s,\boldsymbol{\theta}_s):\mathrm{R}(g_{\boldsymbol{\alpha}_s,\boldsymbol{\theta}_s})=\mathrm{R}(g_{\boldsymbol{\alpha}_s^*,\boldsymbol{\theta}_s^*})\}$。定义 $d((\boldsymbol{\alpha}_s,\boldsymbol{\theta}_s)\ \boldsymbol{\Theta}_s)=\min_{(\boldsymbol{\alpha}_s,\boldsymbol{\theta}_s)\in\boldsymbol{\Theta}_s}\|(\hat{\boldsymbol{\alpha}}_s,\hat{\boldsymbol{\theta}}_s)-(\boldsymbol{\alpha}_s,\boldsymbol{\theta}_s)\|_2^2$，并记 $(\check{\boldsymbol{\alpha}}_s,\check{\boldsymbol{\theta}}_s)\in$ 极小值点$_{(\boldsymbol{\alpha}_s,\boldsymbol{\theta}_s)\in\boldsymbol{\Theta}_s}\|(\hat{\boldsymbol{\alpha}}_s,\hat{\boldsymbol{\theta}}_s)-(\boldsymbol{\alpha}_s,\boldsymbol{\theta}_s)\|_2^2$，其中 $\check{\boldsymbol{\theta}}_s=\{(\check{W}_{s,l},\check{c}_{s,l})\ l=0,\cdots L\}$。记 $\tilde{\boldsymbol{\theta}}\in$极小值点$_{\boldsymbol{\theta}}\mathrm{R}(g_{\tilde{\boldsymbol{\alpha}},\boldsymbol{\theta}})$，其中 $\tilde{\boldsymbol{\alpha}}=(\check{\boldsymbol{\alpha}}_s,0)$,~那么，由引理 7.1 可得 $(\check{\boldsymbol{\alpha}}_s,0,\boldsymbol{\theta})\in\boldsymbol{\Theta}$。

由此可得

$$\mathcal{R}_n(g_{\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\theta}}})+\lambda_1\|\hat{\boldsymbol{\alpha}}\|_1+\lambda_2\sum_{l=1}^{\mathcal{D}}\|\hat{\boldsymbol{W}}_l-\boldsymbol{I}\|_1\le\mathcal{R}_n(g_{\check{\boldsymbol{\alpha}}_s,\boldsymbol{0},\check{\boldsymbol{\theta}}})+\lambda_1\|\check{\boldsymbol{\alpha}}_s\|_1+\lambda_2\sum_{l=1}^{\mathcal{D}}\|\check{\boldsymbol{W}}_l-\boldsymbol{I}\|_1,$$

that

$$\lambda_1\|\hat{\boldsymbol{\alpha}}_c\|_1\le\mathcal{R}_n(g_{\check{\boldsymbol{\alpha}}_s,\boldsymbol{0},\check{\boldsymbol{\theta}}})-\mathcal{R}_n(g_{\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\theta}}})+\lambda_1\Big(\|\check{\boldsymbol{\alpha}}_s\|_1-\|\hat{\boldsymbol{\alpha}}_s\|_1\Big)+\lambda_2\sum_{l=1}^{\mathcal{D}}\Big(\big\|\check{\boldsymbol{W}}_l-\boldsymbol{I}\big\|_1-\big\|\hat{\boldsymbol{W}}_l-\boldsymbol{I}\big\|_1\Big). \tag{14}$$

插入并减去总体风险，得到:

$$\begin{aligned}\lambda_1\|\hat{\boldsymbol{\alpha}}_c\|_1\le&\Big[\mathcal{R}(g_{\check{\boldsymbol{\alpha}}_s,\boldsymbol{0},\check{\boldsymbol{\theta}}})-\mathcal{R}(g_{\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\theta}}})\Big]+\Big[\mathcal{R}(g_{\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\theta}}})-\mathcal{R}_n(g_{\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\theta}}})\Big]\\ &+\lambda_1\Big(\|\check{\boldsymbol{\alpha}}_s\|_1-\|\hat{\boldsymbol{\alpha}}_s\|_1\Big)+\lambda_2\sum_{l=1}^{\mathcal{D}}\Big(\big\|\check{\boldsymbol{W}}_l-\boldsymbol{I}\big\|_1-\big\|\hat{\boldsymbol{W}}_l-\boldsymbol{I}\big\|_1\Big).\end{aligned} \tag{15}$$

由于 $(\check{\boldsymbol{\alpha}}_s、\mathbf{0}、\check{\boldsymbol{\theta}}) \in \Theta$我们有 $\mathcal{R}(g_{\check{\boldsymbol{\alpha}}_s,\ \mathbf{0}、\check{\boldsymbol{\theta}}}) = \mathcal{R}(g^*_{\mathcal{G}}) \leq \mathcal{R}(g_{\hat{\boldsymbol{\alpha}},\ \hat{\boldsymbol{\theta}}})$，因此式（15）中的第一个括号项非正，可以舍去：

$$\begin{aligned}\lambda_1\|\hat{\boldsymbol{\alpha}}_c\|_1 \leq & \Big[\mathcal{R}(g_{\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\theta}}}) - \mathcal{R}_n(g_{\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\theta}}})\Big] + \lambda_1\Big(\|\check{\boldsymbol{\alpha}}_s\|_1 - \|\hat{\boldsymbol{\alpha}}_s\|_1\Big) \\ & + \lambda_2\sum_{l=1}^{\mathcal{D}}\Big(\big\|\check{\boldsymbol{W}}_l - \boldsymbol{I}\big\|_1 - \big\|\hat{\boldsymbol{W}}_l - \boldsymbol{I}\big\|_1\Big). \end{aligned} \tag{16}$$

对于门控项，$\|\check{\boldsymbol{\alpha}}_s\|_1 - \|\hat{\boldsymbol{\alpha}}_s\|_1 \leq \|\check{\boldsymbol{\alpha}}_s - \hat{\boldsymbol{\alpha}}_s\|_1 \leq \sqrt{d}\,\|\check{\boldsymbol{\alpha}}_s - \hat{\boldsymbol{\alpha}}_s\|_2$。对于深度项，根据三角不等式和 ℓ_1–ℓ_2,

$$\sum_{l=1}^{\mathcal{D}}\Big|\big\|\check{\boldsymbol{W}}_l - \boldsymbol{I}\big\|_1 - \big\|\hat{\boldsymbol{W}}_l - \boldsymbol{I}\big\|_1\Big| \ \leq\ \sum_{l=1}^{\mathcal{D}}\big\|\check{\boldsymbol{W}}_l - \hat{\boldsymbol{W}}_l\big\|_1 \ \leq\ \sqrt{S}\,\|\check{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}\|_2,$$

其中 $S = \mathcal{W}^2\mathcal{D}$ 是 θ中的标量参数总数。因此，式（16）可得 s

$$\lambda_1\|\hat{\boldsymbol{\alpha}}_c\|_1 \leq \Big[\mathcal{R}(g_{\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\theta}}}) - \mathcal{R}_n(g_{\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\theta}}})\Big] + \lambda_1\sqrt{d}\,\|\check{\boldsymbol{\alpha}}_s - \hat{\boldsymbol{\alpha}}_s\|_2 + \lambda_2\sqrt{S}\,\|\check{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}\|_2. \tag{17}$$

根据杨氏不等式，我们有：

$$\lambda_1\|\hat{\boldsymbol{\alpha}}_c\|_1 \leq \Big[\mathcal{R}(g_{\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\theta}}}) - \mathcal{R}_n(g_{\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\theta}}})\Big] + \tfrac{1}{2}\|\check{\boldsymbol{\alpha}}_s - \hat{\boldsymbol{\alpha}}_s\|_2^2 + \tfrac{1}{2}\lambda_1^2 d + \tfrac{1}{2}\|\check{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}\|_2^2 + \tfrac{1}{2}\lambda_2^2 S. \tag{18}$$

取期望：

$$\mathbb{E}[\lambda_1\|\hat{\boldsymbol{\alpha}}_c\|_1] \leq \mathbb{E}\Big[\mathcal{R}(g_{\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\theta}}}) - \mathcal{R}_n(g_{\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\theta}}})\Big] + \tfrac{1}{2}\,\mathbb{E}\Big[\|(\hat{\boldsymbol{\alpha}}_s, \hat{\boldsymbol{\theta}}) - (\check{\boldsymbol{\alpha}}_s, \check{\boldsymbol{\theta}})\|_2^2\Big] + \tfrac{1}{2}\lambda_1^2 d + \tfrac{1}{2}\lambda_2^2 S. \tag{19}$$

由此得到

$$\mathbb{E}(\|\hat{\boldsymbol{\alpha}}_c\|_2^2) \leq \mathbb{E}(\|\hat{\boldsymbol{\alpha}}_c\|_1^2) \leq \frac{c\log^2 n\log\mathcal{N}_{2n}}{n} + c\left(\mathcal{R}(g_{\check{\boldsymbol{\alpha}}_s,\mathbf{0},\check{\boldsymbol{\theta}}}) - \mathcal{R}(g^*)\right).$$

引理

引理 7.1 在条件（*C1*）下，若 $p = O(log^c n)$ 对于某个正常数 c，则对于$(\boldsymbol{\alpha}^*, \boldsymbol{\theta}^*) \in$极小值点$_{g\in\mathcal{G}(\boldsymbol{\alpha},\boldsymbol{\theta})}\,\mathbb{E}(Y_i - g(\boldsymbol{X}_i))^2$, 可得

(i) $|\alpha_j^*| \geq \tau_d$，对于 $\forall j = 1, \cdots\ d$ 以及某个正常数 c;

(ii) 存在一个解 $(\boldsymbol{\alpha}^*, \boldsymbol{\theta}^*)$，使得 $\boldsymbol{\alpha}_c^* = 0$。

引理的证明7.1。

(i) 假设存在一个$(\boldsymbol{\alpha}^*, \boldsymbol{\theta}^*)$，使得对于至少一个$j \in \{1, \cdots \ d\}$，$|\alpha_j^*| < c\tau_d$成立。那么，对于任意随机向量 $\boldsymbol{X} \in [0,1]^p$，我们构造向量 $\boldsymbol{X}_{[j]}(0) = (X_{[1]}, \cdots \ X_{[j-1]}, 0, X_{[j+1]}, \cdots \ X_{[p]})$，显然，$|g_{\boldsymbol{\theta}^*}\left(\boldsymbol{\alpha}^* \odot \boldsymbol{X}_{[j]}(0)\right) - g_{\boldsymbol{\theta}^*}\left(\boldsymbol{\alpha}^* \odot \boldsymbol{X}\right)| \leq c\tau_d$，其中某个正常数为 c。根据 $g_{\mathcal{G}}^* =$ 极小值点$_{g \in \mathcal{G}(\boldsymbol{\alpha}, \boldsymbol{\theta})} \mathbb{E}(Y_i - g_{\boldsymbol{\theta}}(\boldsymbol{\alpha} \odot \boldsymbol{X}_i))^2$,的定义，

$$g_{\boldsymbol{\theta}^*}\left(\boldsymbol{\alpha}^* \odot \boldsymbol{X}_{[j]}(0)\right) = \mathbb{E}(Y \mid \boldsymbol{X}_{[-j]} = \boldsymbol{x}_{[-j]}), \ \text{and} \ g_{\boldsymbol{\theta}^*}(\boldsymbol{\alpha}^* \odot \boldsymbol{X}) = \mathbb{E}(Y \mid \boldsymbol{X} = \boldsymbol{x}).$$

这与条件(C2)矛盾。因此，对于任意 $j = 1, \cdots \ d$，$|\alpha_j^*| \geq \tau_d$，其中某个正常数为 c。

(ii) 令 $(\boldsymbol{\alpha}_s^*, \boldsymbol{\theta}_s^*) \in \arg\min_{\boldsymbol{\alpha}_s, \boldsymbol{\theta}_s} \mathrm{R}(g_{\boldsymbol{\alpha}_s, \boldsymbol{\theta}_s}) = \arg\min_{\boldsymbol{\alpha}_s, \boldsymbol{\theta}_s} \mathbb{E}(Y - g_{\boldsymbol{\theta}_s}(\boldsymbol{\alpha}_s \odot \boldsymbol{X}_s))^2$。定义 $\boldsymbol{\Theta}_s = \{(\boldsymbol{\alpha}_s, \boldsymbol{\theta}_s) : \mathrm{R}(g_{\boldsymbol{\alpha}_s, \boldsymbol{\theta}_s}) = \mathrm{R}(g_{\boldsymbol{\alpha}_s^*, \boldsymbol{\theta}_s^*})\}$。定义 $d((\boldsymbol{\alpha}_s, \boldsymbol{\theta}_s) \ \boldsymbol{\Theta}_s) = \min_{(\boldsymbol{\alpha}_s, \boldsymbol{\theta}_s) \in \boldsymbol{\Theta}_s} \|(\hat{\boldsymbol{\alpha}}_s, \hat{\boldsymbol{\theta}}_s) - (\boldsymbol{\alpha}_s, \boldsymbol{\theta}_s)\|_2^2$，并记 $(\check{\boldsymbol{\alpha}}_s, \check{\boldsymbol{\theta}}_s) \in \arg\min_{(\boldsymbol{\alpha}_s, \boldsymbol{\theta}_s) \in \boldsymbol{\Theta}_s} \|(\hat{\boldsymbol{\alpha}}_s, \hat{\boldsymbol{\theta}}_s) - (\boldsymbol{\alpha}_s, \boldsymbol{\theta}_s)\|_2^2$，其中 $\check{\boldsymbol{\theta}}_s = \{(\check{A}_{s,l}, \check{c}_{s,l}) \ l = 0, \cdots \ L\}$。记 $\tilde{\boldsymbol{\theta}} \in \arg\min_{\boldsymbol{\theta}} \mathrm{R}(g_{\tilde{\boldsymbol{\alpha}}, \boldsymbol{\theta}})$，其中 $\tilde{\boldsymbol{\alpha}} = (\check{\boldsymbol{\alpha}}_s, 0)$。令 $\check{\boldsymbol{\theta}} = \{(\tilde{A}_l, \tilde{c}_l) \ l = 0, \cdots \ L\}$，其中 $\tilde{c}_l = \check{c}_{s,l}$ for $l = 0, \cdots \ L$, $\tilde{A}_l = \check{A}_{s,l}$ for $l = 1, \cdots \ L$，且 $\tilde{A}_0 = (\check{A}_s, 0, A_c, 0)$。然后，根据定义 (2)，容易证明

$$\mathcal{R}(g_{\tilde{\boldsymbol{\alpha}}, \tilde{\boldsymbol{\theta}}}) \leq \mathcal{R}(g_{\tilde{\boldsymbol{\alpha}}, \check{\boldsymbol{\theta}}}) = \mathcal{R}(g_{\check{\boldsymbol{\alpha}}_s, \check{\boldsymbol{\theta}}_s}) = \mathcal{R}(g_{\boldsymbol{\alpha}^*, \boldsymbol{\theta}^*}).$$

~显然，（$\tilde{\boldsymbol{\alpha}}$， $\boldsymbol{\theta}$）是解。

引理 7.2（逼近误差（见 *Jiao 等人（2023）* 中的定理 *3.3*））

给定 *Hölder* 光滑函数 $g^* \in \boldsymbol{H}^{\beta}([0,1]^d, B_0)$，对于任意 $D \in \mathbb{N}^+$ 和 $W \in \mathbb{N}^+$，存在一个函数 $g_{\mathcal{G}}^*$，可由宽度为$\boldsymbol{W} = 38(\lfloor\beta\rfloor + 1)^2 d^{\lfloor\beta\rfloor+1} W \lceil\log_2(8W)\rceil$、深度为$\boldsymbol{D} = 21(\lfloor\beta\rfloor + 1)^2 D \lceil\log_2(8D)\rceil$ 的***ReLU*** 前馈神经网络实现，使得

$$\left|g^* - g_{\mathcal{G}}^*\right| \leq 18 B_0 (\lfloor\beta\rfloor + 1)^2 d^{\lfloor\beta\rfloor + \max\{\beta, 1\}/2} (WD)^{-2\beta/d},$$

对所有 $x \in [0,1]^d \mid \Omega([0,1]^d, K, \delta)$，其中

$$\Omega([0,1]^d, K, \delta) = \bigcup_{i=1}^{d} \{x = [x_1, \cdots, x_d]^T : x_i \in \bigcup_{k=1}^{K-1} (k/K - \delta, k/K)\},$$

其中 $K = \lfloor WD \rfloor$ 和 δ 是 $(0, 1/(3K))$中的任意数。

引理 7.3 *(界定覆盖数，（Anthony 等人，1999年中的定理 12.2）和（Bartlett 等人，2019年中的定理 3 和定理 7)*

设 *ReLU* 前馈神经网络 G 是从定义域 X 到有界区间 $[0,B]$ 的实值函数集合。存在一个通用常数 C，使得以下结论成立。给定任意 D、S，且 $S > CD > C^2$，存在一个网络类 G，其具有 $\leq D$ 层和 $\leq S$ 个参数，VC-维数为 $\geq SD\log(S/D)/C$，并且给定 $\delta > 0$

$$\sup_{\boldsymbol{x}} \log \mathcal{N}_{2n}(\delta, \|\cdot\|_\infty, \mathcal{G}|_{\boldsymbol{x}}) = O\Big(\mathcal{S}\mathcal{D}\log(\mathcal{S}/\delta)\Big).$$