arXiv:2305.04267v1 [cs.LG] 7 May 2023

通过 LASSO 正则化实现两层 ReLU 神经网络的可证明的可辨识性

Gen Li、Ganghua Wang、Jie Ding

摘要——LASSO 正则化是一种广受欢迎的回归工具，通过 ℓ_1 惩罚进行变量选择，从而提高统计模型的预测准确率；它最初是为线性模型及其变体提出的。本文将 LASSO 的应用范围拓展至两层 ReLU 神经网络这一时兴且强大的非线性回归模型。具体而言，对于一个输出 y 仅依赖于一小部分输入 $\boldsymbol{x}$（记为 $\mathcal{S}^\star$）的神经网络，我们证明，当样本数量随输入维度按对数级增长时，LASSO 估计量能够稳定地重构该神经网络并识别 $\mathcal{S}^\star$ 。对于线性模型，这一颇具挑战性的情形已得到充分研究，而在神经网络领域却鲜有探讨。我们的理论基于针对两层 ReLU 神经网络的扩展型受限等距性质（RIP）分析框架，该框架或可独立应用于其他 LASSO 或神经网络问题。基于这一结果，我们提出一种基于神经网络的变量选择方法。在模拟数据集和真实世界数据集上的实验表明，与现有技术相比，该变量选择方法表现良好。

索引词——LASSO、可辨识性、神经网络、非线性回归、变量选择。

I. 引言

给定 n 个观测值 $(y_i, \boldsymbol{x}_i)$, $i = 1, \ldots, n$，我们通常用回归形式 $y_i = f(\boldsymbol{x}_i) + \xi_i$ 对其建模，其中 f 是未知函数，$\boldsymbol{x}_i \in \mathbb{R}^p$ 是输入变量，ξ_i 表示统计误差。一般而言，我们希望通过 $\widehat{\ }$ 估计回归函数 f_n，使其尽可能接近 f，以用于预测或解释。当输入维度 p 与数据量 n 相当，甚至远大于数据量时，这一问题颇具挑战性。对于线性回归，即 $f(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{x}$，最小绝对收缩和选择算子（LASSO）[1] 正则化已成为估计 f 的标准工具。LASSO 也已成功应用于并用于研究许多非线性模型，例如广义线性模型 [2]，、比例风险模型 [3]，和神经网络 [4]。对于经 LASSO 正则化的神经网络，已有研究探讨了其不同性质，例如训练收敛性 [5]，、模型剪枝 [6]，[7]，和特征选择 [8]，[9]。LASSO 正则化也已被加入许多开源库的标准深度学习工具箱中，例如 TensorFlow [10] 和 PyTorch [11]。

尽管 LASSO 在提升神经网络的泛化能力和稀疏化方面取得了实践上的成功，但能否使用 LASSO 识别显著变量仍是研究尚不充分。对于线性模型，变量选择问题在不同文献中也称为支持集恢复或特征选择。选择一致性要求，当 $n \to \infty$时，$\text{supp}(\widehat{\boldsymbol{w}}) = \text{supp}(\boldsymbol{w})$的概率趋于 1。选择简约子模型的标准方法是求解惩罚回归问题，或迭代地选取显著变量 [12]。现有方法的差异在于如何纳入独特的领域知识（例如稀疏性、多重共线性、组行为），或旨在实现何种性质（例如系数估计的一致性、变量选择的一致性）[13]。例如，关于 LASSO 方法 [1] 在估计显著变量方面的一致性，已有大量研究，并考察了各种技术条件，包括稀疏性、互相干性 [14],受限等距性 [15]，不可表示条件 [16]，以及受限特征值 [17]。

许多关于神经网络的理论研究都聚焦于泛化能力。例如，研究者建立了通用逼近定理，表明任意连续多元函数都可以由参数数量为多项式级的两层网络精确表示 [18]。随后研究表明，任意连续函数都可以由具有Sigmoid激活函数的两层感知器任意逼近 [19]，，并建立了使用双层神经网络拟合任意光滑函数的逼近误差界 [20]，[21]。在统计方面，双层神经网络 [21]，[22] 和多层网络[23]–[25] 的泛化误差界已得到发展。从优化角度来看，神经网络的参数估计被转化为张量分解问题，并可获得可证明的全局最优解 [26]–[28]。最近，研究者提出了一种与维度无关的Rademacher复杂度，用于界定深度ReLU神经网络的泛化误差，从而避免维数灾难 [29]。研究证明，某些非零网络参数较少的深度神经网络能够达到极小极大收敛速率 [30]。通过假设数据由广义层次交互模型生成，研究者建立了一个与输入维度无关的紧误差界 [31]。

本研究从理论上探讨神经网络的可辨识性，并将其用于变量选择。具体而言，假设观测值由一个仅有少数非零系数的神经网络生成。可辨识性关注的是能否识别这些系数，进而可用于确定一组稀疏的输入变量，其中的变量与响应变量确实相关。在这方面，已有研究主张采用 LASSO 及其变体组LASSO [32] 对神经网络进行正则化，以实现变量选择，

G. Li 隶属于宾夕法尼亚大学沃顿商学院统计与数据科学系。G. Wang 和 J. Ding 隶属于明尼苏达大学统计学院。

实践（例如，[6], [8], [9], [33]）。

本文中，我们考虑以下用于回归的两层 ReLU 神经网络类别。

$$\mathcal{F}_r = \Big\{ f : \boldsymbol{x} \mapsto f(\boldsymbol{x}) = \sum_{j=1}^{r} a_j \mathrm{relu}(\boldsymbol{w}_j^\top \boldsymbol{x} + b_j), \quad \text{where } a_j, b_j \in \mathbb{R}, \boldsymbol{w}_j \in \mathbb{R}^p \Big\}.$$

这里， p 和 r 分别表示输入维度和神经元数量。我们假设数据由 $\mathcal{F}_r$ 中的一个回归函数生成，并受到一个小项的扰动。我们将研究以下两个问题。

首先，如果底层回归函数 f 具有一种简约表示，使得仅有一小组输入变量 $\mathcal{S}^\star$是相关的，那么在给定可能含噪的测量值（y_i, $\boldsymbol{x}_i$）的情况下，我们能否以高概率识别出它们，其中$i = 1$、.....、n? 其次，这样的 $\mathcal{S}^\star$ 是否可估计，也就是说，即使在小-n 和大-p 的情形下，能否以高概率通过一个优化问题求解?

为了解决上述问题，我们将通过考虑问题 $\min_{\boldsymbol{W},\boldsymbol{a},\boldsymbol{b}} \|\boldsymbol{W}\|_1$ 来建立带有 LASSO 正则化的神经网络理论，其约束条件为

$$\frac{1}{n}\sum_{i=1}^{n}\Big(y_i - \sum_{j=1}^{r} a_j \mathrm{relu}(\boldsymbol{w}_j^\top \boldsymbol{x}_i + b_j)\Big)^2 \le \sigma^2,$$

这是一种替代版本的 ℓ_1 正则化。更多符号说明将在第二节 B 小节中介绍。

我们的理论对上述问题给出了肯定的答案。我们从理论上证明，LASSO 型估计量能够稳定地识别具有稀疏输入信号的 ReLU 神经网络，且识别结果至多相差隐藏神经元的置换。我们仅关注变化的 n 和 p，并隐含地假设 $\boldsymbol{W}^\star$的稀疏性以及神经元数量 r 是固定的。虽然这并未涵盖宽型神经网络，但我们认为它仍然对应于一个非平凡且有趣的函数类。例如，当输入信号有界时，该函数类包含线性函数。我们的结果相当一般，因为它适用于y 的带噪观测值，以及样本量 n 远小于输入变量数量 p的维度情形。该理论建立在新的集中界和函数分析之上，而这些结果本身也可能具有一定的独立研究价值。

受所发展的理论启发，我们还提出了一种基于神经网络的变量选择方法。其思路是利用神经系统作为工具来建模非线性并提取显著变量。通过多项实验研究，我们展示了该技术在高维数据中识别稀疏显著变量集方面的良好表现，即使底层数据并非由神经网络生成。与基于树集成和线性 LASSO 的常用方法相比，所提出的方法适用于从非线性、高维且低噪声的系统中进行变量选择。

本文其余部分安排如下：第二节介绍主要理论结果，并提出一种用于变量选择的算法。第三节利用模拟数据集和真实世界数据集验证所提出的理论和算法。第四节对全文进行总结。

II. 主要结果

A. 符号说明

令 $\boldsymbol{u}_{\mathcal{S}}$ 表示一个向量，其在集合 $\mathcal{S}$ 中索引对应的元素与 $\boldsymbol{u}$ 中的元素保持不变，其余元素均为零。对于任意矩阵 $\boldsymbol{W} \in \mathbb{R}^{p\times r}$，我们定义

$$\|\boldsymbol{W}\|_1 = \sum_{1\le k\le p, 1\le j\le r} |w_{kj}|, \ \|\boldsymbol{W}\|_{\mathrm{F}} = \Big(\sum_{1\le k\le p, 1\le j\le r} w_{kj}^2\Big)^{1/2}.$$

向量也采用类似的记号。两个向量的内积记为 $\langle \boldsymbol{u}, \boldsymbol{v}\rangle$。令 $\boldsymbol{w}_j$ 表示 $\boldsymbol{W}$ 的第 j 列。矩阵 $\boldsymbol{W}$ 的稀疏性指 $\boldsymbol{W}$ 中非零元素的数量。令 $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_p)$ 表示标准的 p 维高斯分布，$\mathbb{I}(\cdot)$ 表示示性函数。修正线性单元（ReLU）函数定义为 $\mathrm{relu}(v) = \max\{v, 0\}$，对所有 $v \in \mathbb{R}$ 均成立。

B. 问题表述

考虑 n 个独立同分布（i.i.d.）的观测值 $\{\boldsymbol{x}_i, y_i\}_{1\le i\le n}$，它们满足

$$y_i = \sum_{j=1}^{r} a_j^\star \cdot \mathrm{relu}(\boldsymbol{w}_j^{\star\top} \boldsymbol{x}_i + b_j^\star) + \xi_i \tag{1}$$

其中 $\boldsymbol{x}_i \sim \mathcal{N}\ \boldsymbol{0}\ \boldsymbol{I}_p$ r $a_j^\star \in (,)$，这 $\{1,-1\}$ $\boldsymbol{w}_j^\star \in \mathbb{R}^p$ $b_j^\star \in \mathbb{R}$ ξ_i 里 表示神经元的数量，，，而 表示满足以下条件的随机噪声或近似误差

$$\frac{1}{n}\sum_{i=1}^{n}\xi_i^2 \le \sigma^2. \tag{2}$$

在上述问题表述中，假设 $a_j^\star \in \{1,-1\}$ 不会失去一般性，因为对于任意$c > 0$， $a \cdot \mathrm{relu}(b) = ac \cdot \mathrm{relu}(b/c)$。不等式（2）的设定是为了简化表述。如果$\xi_i$是无界随机变量，即将介绍的理论结果仍然成立，更多解释见附录。 ξ_i不一定是独立同分布的，并且允许 σ 为零，此时问题退化为无噪声情形。

令 $\boldsymbol{W}^\star = [\boldsymbol{w}_1^\star, \ldots, \boldsymbol{w}_r^\star] \in \mathbb{R}^{p\times r}$ 表示数据生成系数。需要解决的主要问题是：在 $\boldsymbol{W}^\star$ 的大多数元素为零的情况下，能否稳定地识别出那些非零元素？从可辨识性的角度研究神经网络至关重要。与研究机器学习模型预测性能的泛化问题不同，可辨识性可用于解释建模结果，并帮助科学家做出可信赖的决策。为阐明这一点，我们将在II-E小节提出使用神经网络进行变量选择。

为回答上述问题，我们提出研究以下 LASSO型优化。设 $(\widehat{\boldsymbol{W}}、\widehat{\boldsymbol{a}}、\widehat{\boldsymbol{b}})$ 是以下优化问题的一个解，

$$\min_{\boldsymbol{W},\boldsymbol{a},\boldsymbol{b}} \|\boldsymbol{W}\|_1 \tag{3}$$
$$\text{subject to } \frac{1}{n}\sum_{i=1}^{n}\Big(y_i - \sum_{j=1}^{r} a_j \cdot \mathrm{relu}(\boldsymbol{w}_j^\top \boldsymbol{x}_i + b_j)\Big)^2 \le \sigma^2,$$

在可行范围内满足 $\boldsymbol{a} \in \{1,-1\}^r$、 $\boldsymbol{W} \in \mathbb{R}^{p\times r}$ 和 $\boldsymbol{b} \in \mathbb{R}^r$。直观而言，优化是在以下条件下进行的

约束：训练误差不能太大，并且目标函数倾向于使 $\boldsymbol{W}$ 稀疏。在某些正则性条件下，我们将证明该解确实是稀疏的，并且接近数据生成过程。

C. 主要结果

我们作出如下技术假设。

假设1. 对于某个常数 $B \geq 1$,

$$1 \leq \|\boldsymbol{w}_j^\star\|_2 \leq B \quad and \quad |b_j^\star| \leq B \quad \forall 1 \leq j \leq r. \tag{4}$$

此外，对于某个常数 $\omega > 0$,

$$\max_{j,k=1,\ldots,r,j\neq k} \frac{\left|\langle \boldsymbol{w}_j^\star, \boldsymbol{w}_k^\star \rangle\right|}{\|\boldsymbol{w}_j^\star\|_2 \|\boldsymbol{w}_k^\star\|_2} \leq \frac{1}{r^\omega}. \tag{5}$$

注释1（关于假设1的讨论）。式 (4) 中的条件仅为技术上的便利而进行归一化，因为我们可以按比例重新缩放 $\boldsymbol{w}_j, b_j, y_i, \sigma$ ，而不失一般性。尽管该条件隐含地要求对所有 $j = 1, \ldots, r$， $\boldsymbol{w}_j^\star \neq \boldsymbol{0}$，但这是合理的，因为这意味着神经元$j$ 未被使用/激活。式 (5) 中的条件要求任意两个不同系数向量之间的夹角不能太小。该条件类似于线性回归问题中常用的有界特征值条件，其中每个 $w_j^\star$都被视为设计矩阵中的一列。这个条件绝不算宽松，在实践中也不容易验证。然而，由于我们关注的情形是 p、n 较大而 r 较小，我们认为式 (5) 中的条件仍然合理。例如，当 $r = 2$ 时，该条件仅要求 $w_1^\star \neq w_2^\star$。

我们的主要结果表明，如果 $\boldsymbol{W}^\star$ 是稀疏的，那么当样本数量（n）随输入维度（p）呈对数级增长时，就可以稳定地重建神经网络。持怀疑态度的读者可能会问，常数具体如何依赖于稀疏性和 r。我们将在 II-D小节中给出更为详细的结果，并在该节中介绍证明。

定理1. 在假设1下，存在一些仅取决于 $\boldsymbol{W}^\star$ 的稀疏性（且为多项式关系）的常数$c_1, c_2, c_3 > 0$ ，使得对于任意 $\delta > 0$，以至少 $1 - \delta$ 的概率，久久久久久

$$\widehat{\boldsymbol{a}} = \boldsymbol{\Pi}\boldsymbol{a}^\star \; and \; \|\widehat{\boldsymbol{W}} - \boldsymbol{W}^\star\boldsymbol{\Pi}^\top\|_{\mathrm{F}} + \|\widehat{\boldsymbol{b}} - \boldsymbol{\Pi}\boldsymbol{b}^\star\|_2 \leq c_1\sigma \tag{6}$$

对于某个置换矩阵 $\boldsymbol{\Pi}$，满足的是

$$n > c_2 \log^4 \frac{p}{\delta} \qquad and \qquad \sigma < c_3. \tag{7}$$

注释2（定理1的解释）。置换矩阵 $\boldsymbol{\Pi}$ 是必要的，因为所考虑的神经网络在隐藏神经元进行任何置换时，都会产生相同的预测分布（ y 条件 $\boldsymbol{x}$）。该结果表明，即使样本量 n 远小于变量个数 p，底层神经系数仍可得到稳定估计。此外，估计误差界为 σ 的量级，即（2）中指定的噪声水平。

假设我们将信噪比（SNR）定义为 $\mathbb{E}\|\boldsymbol{x}\|^2/\sigma^2$。对该定理的另一种解读是，较大的信噪比可以确保全局极小值以较高概率接近真实值。有人可能会想知道如果不满足 $\sigma < c_3$ 条件会怎样。我们注意到，如果 σ 过大，（6）中的误差界会较为宽松，而且本身也没什么实际意义。换言之，如果信噪比较小，我们可能无法稳定地估计参数。这一点将在第三节的实验研究中展示。

定理1中的估计结果可以转化为变量选择结果，如下述推论1所示。二者之间的联系基于这样一个事实：如果第 i个变量是冗余的，那么与之相关的潜在系数应为零。令$\boldsymbol{w}_{i,\cdot}^\star$ 表示 $\boldsymbol{W}^\star$的第 i行。那么，

$$\mathcal{S}^\star = \{1 \leq i \leq p : \|\boldsymbol{w}_{i,\cdot}^\star\|_2 > 0\}$$

刻画了“显著变量”。推论1指出，系数估计值非零的变量集合包含所有显著变量。该推论还表明，通过对系数估计值进行适当收缩，可以实现变量选择一致性。

推论1（变量选择）。令 $\mathcal{S}_{c_1\sigma} \subseteq \{1, \ldots, p\}$表示满足$\|\widehat{\boldsymbol{w}}_{i,\cdot}\|_2 > c_1\sigma$的 i集合。在与定理1相同的假设下，且 $\min_{1\leq i\leq r}\|\boldsymbol{w}_{i,\cdot}^\star\|_2 > 2c_1\sigma$，对于任意 $\delta > 0$，有

$$\mathbb{P}(\mathcal{S}^\star = \widehat{\mathcal{S}}_{c_1\sigma}) \geq 1 - \delta,$$

只要 $n > c_2 \log^4 \frac{p}{\delta}$ 和 $\sigma < c_3$。

考虑无噪声情形 $\sigma = 0$，定理1还推出以下推论。

推论2（唯一简约表示）。在与定理1相同的假设下，存在一个常数 $c_2 >0$，它仅取决于 $\boldsymbol{W}^\star$ 的稀疏性，使得对于任意$\delta > 0$，以至少 $1 - \delta$ 的概率，有

$$\widehat{\boldsymbol{a}} = \boldsymbol{\Pi}\boldsymbol{a}^\star, \quad \widehat{\boldsymbol{W}} = \boldsymbol{W}^\star\boldsymbol{\Pi}^\top, \quad and \quad \widehat{\boldsymbol{b}} = \boldsymbol{\Pi}\boldsymbol{b}^\star$$

对于某个置换矩阵 $\boldsymbol{\Pi}$，只要 $n > c_2 \log^4 \frac{p}{\delta}$。

推论2指出，在（1）中的所有可能表示$\boldsymbol{W}$ （其中 $\xi_i = 0$）中，以高概率，具有最小 ℓ_1-范数的表示必定与 $\boldsymbol{W}^\star$ 相同（至多相差一个列置换）。换言之，两层 ReLU 神经网络在 ℓ_1 范数意义下的最简约表示是唯一的。这一观察回答了第 I 节提出的问题。

值得一提的是，由于神经网络的权重矩阵 $\boldsymbol{W}$ 是行稀疏的，Group-LASSO 是 LASSO 的一种合适替代方案。我们将 Group-LASSO 的分析留待未来研究。

D. 主要结果的详细阐述

假设 $\boldsymbol{W}^\star$ 至多有 s 个非零条目。以下定理是定理1的更详细版本。

定理2。存在一些常数 c_1、c_2、$c_3 > 0$ ，使得对于任意 $\delta > 0$，以至少 $1 - \delta$的概率，

$$\widehat{\boldsymbol{a}} = \boldsymbol{\Pi}\boldsymbol{a}^\star \; and \; \|\widehat{\boldsymbol{W}} - \boldsymbol{W}^\star\boldsymbol{\Pi}^\top\|_{\mathrm{F}} + \|\widehat{\boldsymbol{b}} - \boldsymbol{\Pi}\boldsymbol{b}^\star\|_2 \leq c_1\sigma \tag{8}$$

对于某个置换 $\mathbf{\Pi} \in \{0,1\}^{r\times r}$，前提是假设 1 成立且

$$n > c_2 s^3 r^{13} \log^4 \frac{p}{\delta} \quad \textit{and} \quad \sigma < \frac{c_3}{r}. \tag{9}$$

注3（定理1的证明概要）。定理1的证明并不简单，详见附录。下面简要说明证明思路。首先，我们将定义我们所说的 $(\boldsymbol{W}, \boldsymbol{a}, \boldsymbol{b})$ 与 $(\boldsymbol{W}^\star, \boldsymbol{a}^\star, \boldsymbol{b}^\star)$ 之间的 D_1-距离和 D_2-距离。这些距离可视为两个向量之间经典 ℓ_1 和 ℓ_2 距离的对应项，但允许神经元在任意置换下保持不变（注2）。接下来，我们令

$$\Delta_n \coloneqq \frac{1}{n}\sum_{i=1}^{n}\Bigg[\sum_{j=1}^{r} a_j \mathrm{relu}(\boldsymbol{w}_j^\top \boldsymbol{x}_i + b_j) - \sum_{j=1}^{r} a_j^\star \mathrm{relu}(\boldsymbol{w}_j^{\star\top} \boldsymbol{x}_i + b_j^\star)\Bigg]^2,$$

其中 $(\boldsymbol{W}, \boldsymbol{a}, \boldsymbol{b})$ 是问题 (3) 的解，并推导出其如下上界和下界：

$$\Delta_n \le c_6 \left(\frac{r}{S} + \frac{r\log^3 \frac{p}{n\delta}}{n}\right) D_1^2 + c_6\sigma^2 \quad \textit{and}$$
$$\Delta_n \ge c_4 \min\left\{\frac{1}{r}, D_2^2\right\} \tag{10}$$

以至少 $1-\delta$ 的概率成立，条件是 $n \ge c_5 S^3 r^4 \log^4 \frac{p}{\delta}$，其中常数 c_4，c_5，c_6，以及 S 待定。这里，上界将通过一系列基本不等式推导得出。下界让人联想到线性模型的受限等距性质（RIP）*[15]*。我们将通过集中论证，由总体对应项的下界推导得出该下界，即

$$\mathbb{E}\left[\sum_{j=1}^{r} a_j \mathrm{relu}(\boldsymbol{w}_j^\top \boldsymbol{x} + b_j) - \sum_{j=1}^{r} a_j^\star \mathrm{relu}(\boldsymbol{w}_j^{\star\top} \boldsymbol{x} + b_j^\star)\right]^2 \ge c \min\left\{\frac{1}{r}, D_2^2\right\},$$

对某个常数 $c > 0$。式 (10) 中的界表明，以高概率，

$$c_4 \min\left\{\frac{1}{r}, D_2^2\right\} \le c_6 \left(\frac{r}{S} + \frac{r\log^3 \frac{p}{n\delta}}{n}\right) D_1^2 + c_6\sigma^2,$$

利用这一点以及连接 D_1 和 D_2 的不等式，我们可以证明最终结果。

注4（替代假设与结果）。我们给出一个定理2的替代结果。考虑以下假设1'，以替代假设1。假设1'：对某个常数 $B > 0$,

$$\|\boldsymbol{w}_j^\star\|_2 \le B \quad \textit{and} \quad |b_j^\star| \le B \quad \textit{for all } 1 \le j \le r.$$

此外，

$$\mathbb{E}\Big[\langle \boldsymbol{a}, \mathrm{relu}(\boldsymbol{W}^\top \boldsymbol{x} + \boldsymbol{b})\rangle - \langle \boldsymbol{a}^\star, \mathrm{relu}(\boldsymbol{W}^{\star\top} \boldsymbol{x} + \boldsymbol{b}^\star)\rangle\Big]^2 \ge \psi D_2\left[(\boldsymbol{W}, \boldsymbol{a}, \boldsymbol{b}), (\boldsymbol{W}^\star, \boldsymbol{a}^\star, \boldsymbol{b}^\star)\right]^2, \tag{11}$$

and

$$n > \frac{c_2}{\psi} s^3 r^3 \log^4 \frac{p}{\delta}. \tag{12}$$

用假设1'代替假设1，仍然可以推导出与定理2相同的结果。上述结果的证明与定理2的证明类似，区别在于，我们将不等式（11）代入附录A中的（21），而不是不等式（22）。

E. 变量选择

为在实践中求解优化问题 (3)，我们考虑以下替代问题，

$$\min_{\boldsymbol{W}\in\mathbb{R}^{p\times r}, \boldsymbol{a}\in\mathbb{R}^r, \boldsymbol{b}\in\mathbb{R}^r} \left\{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{r} a_j \cdot \mathrm{relu}(\boldsymbol{w}_j^\top \boldsymbol{x}_i + b_j)\right)^2 + \lambda\|\boldsymbol{W}\|_1\right\}. \tag{13}$$

已有实证结果表明，随机梯度下降等算法能够为上述优化问题找到一个良好的近似解 [34], [35]。接下来，我们将讨论使用 LASSO 正则化神经网络进行变量选择的一些细节。

调优参数。在实际应用中，给定一个带标签的数据集，我们需要调整超参数，包括惩罚项λ、神经元数量 r、学习率和训练轮数。我们建议采用常见的方法，将可用数据划分为训练集和验证集。使用训练数据，根据一组候选超参数估计神经网络。根据模型在验证集上的预测性能，选出最合适的候选方案。

变量重要性。受推论1的启发，我们将 $\widehat{\boldsymbol{w}}_i$ 的 ℓ_2-范数, $\cdot$ 解释为第 i 个变量的重要性，其中 $i = 1, \ldots, p$。推论1表明，如果我们正确设置截断值 $c_1\sigma$，则可以以较高的概率准确识别 $\mathcal{S}^\star$ 中的所有显著变量。

设置截断值。我们可以优先选择重要性较高的变量，而非重要性接近于零的变量。这启发我们根据变量的重要性将其分为两组。这里，我们提出两种可能的方法。第一种是采用数据驱动的方法，例如 k-means 和高斯混合模型（GMM）。第二种是根据对重要变量数量的领域知识手动设置阈值。

扩展至深度神经网络。受文献 (13) 启发，我们可以直观地将所提出的方法推广到深度神经网络：对输入层中权重矩阵的 ℓ_1-范数进行惩罚。尽管我们尚未针对这一更广泛的设定进行理论分析，但数值研究表明，该方法仍然有效。

III. 实验

我们通过实验研究展示了所提出的变量选择方法的良好性能。我们比较了所提出的算法（“NN”）与若干其他方法在变量选择准确率和预测性能方面的表现。

基线方法，包括LASSO（‘LASSO’）、正交匹配追踪（‘OMP’）、随机森林（‘RF’）、梯度提升（‘GB’）、采用组LASSO的神经网络（‘GLASSO’）[5], 组稀疏正则化（‘GSR’） [8], 以及LNET（‘LNET’） [9]。需要搜索的‘NN’超参数包括惩罚项 $\lambda \in \{0.1, 0.05, 0.01, 0.005, 0.001\}$、神经元数量 $r \in \{20, 50, 100\}$、学习率 $\{0.01, 0.005, 0.001\}$以及训练轮数$\{100, 200, 500\}$。此外，我们将‘NN’扩展为一个包含额外的十个神经元隐藏层的神经网络。我们分别用‘NN-2’和‘NN-3’区分所提出的双层和三层神经网络方法。更多实验细节请参见附录E。

A. 合成数据集

1) NN生成的数据集：第一个实验使用根据公式（1）生成的数据，其中包含 $p = 100$ 个变量和 $r = 16$ 个神经元。神经系数的前 10 行 $\boldsymbol{W}$ 独立地从标准均匀分布中生成，其余各行为零，代表 10 个显著变量。神经元偏置 $\boldsymbol{b}$ 也从标准均匀分布中生成。神经元的符号 $\boldsymbol{a}$ 遵循相互独立的伯努利分布。训练集大小为 $n = 500$，测试集大小为 2000。噪声为均值为零的高斯噪声，其标准差 σ 分别设为 0 、 0.5 、 1 和 5。对于每个 σ，我们评估其在测试数据集上的均方误差，以及用于变量选择的三个指标：正确选出的变量数（“TP”，越大越好）、错误选出的变量数（“FP”，越小越好）和曲线下面积得分（“AUC”，越大越好）。这里，“AUC”根据各方法给出的变量重要性进行评估，详情见附录E。该过程独立重复 20 次。

表I 神经网络生成数据上的性能比较，具体包括正确（“真阳性”）、错误（“假阳性”）选定特征的数量，以及不同 σ 的 AUC 分数。标准误位于括号内。

方法	指标	$\sigma = 0$	$\sigma = 0.5$	$\sigma = 1$	$\sigma = 5$
LASSO	TP	8.30（2.57）	9.10 (1.22)	9.20 (1.03)	6.70 (3.68)
	FP	10.40 (6.99)	10.70 (6.09)	13.35 (6.17)	8.60 (8.13)
	AUC	0.96 (0.06)	0.96（0.05）	0.96 (0.05)	0.87（0.14）
OMP	TP	8.45（1.53）	8.30（1.35）	8.00 (2.17)	6.00 (2.12)
	FP	0.10 (0.30)	0.15 (0.36)	0.25 (0.70)	0.65 (1.06)
	AUC	0.92 (0.08)	0.91 (0.07)	0.91 (0.09)	0.80 (0.11)
RF	TP	6.95 (3.12)	5.55 (3.29)	5.75 (3.27)	4.20 (2.60)
	FP	0.45 (0.67)	0.40 (0.66)	0.35 (0.57)	1.20 (2.18)
	AUC	0.99 (0.02)	0.97 (0.03)	0.95 (0.04)	0.86 (0.12)
GB	TP	6.85 (2.97)	7.15 (3.09)	5.75 (3.69)	5.65 (3.20)
	FP	1.35 (1.46)	1.60（1.91）	2.10 (2.07)	5.20 (6.69)
	AUC	0.98 (0.02)	0.97 (0.03)	0.97 (0.03)	0.88 (0.11)
图形套索	TP	9.35 (1.42)	9.80 (0.51)	9.45 (0.59)	6.35 (2.22)
	FP	0.10 (0.44)	0.65 (0.96)	1.00 (1.38)	8.50（11.45）
	AUC	1.00 (0.00)	1.00 (0.00)	0.99（0.02）	0.84 (0.14)
GSR	TP	9.55 (1.96)	10.00 (0.00)	9.90 (0.30)	7.90 (2.81)
	FP	1.25 (2.05)	0.70 (1.19)	1.35 (2.26)	22.45 (29.67)
	AUC	1.00 (0.00)	1.00 (0.00)	1.00 (0.01)	0.84 (0.18)
LNET	TP	10.00 (0.00)	10.00 (0.00)	9.95 (0.22)	5.10 (2.91)
	FP	66.40（11.15）	59.00 (14.64)	41.60（21.62）	8.05 (14.41)
	AUC	0.63 (0.06)	0.67 (0.08)	0.77 (0.12)	0.83 (0.12)
NN-2	TP	10.00 (0.00)	9.85 (0.36)	9.80 (0.51)	7.80 (1.96)
	FP	0.75 (1.13)	1.25 (2.09)	2.55 (5.80)	12.30 (10.57)
	AUC	1.00（0.00）	1.00（0.01）	1.00 (0.01)	0.88 (0.13)
NN-3	TP	9.75 (0.16)	9.85 (0.08)	9.10 (0.45)	6.95 (0.44)
	FP	0.75 (0.34)	0.65 (0.26)	0.45 (0.18)	2.50 (0.73)
	AUC	1.00（0.00）	1.00 (0.00)	0.99 (0.00)	0.85 (0.03)

表 II 神经网络生成的数据上的性能比较，以不同 σ 下的平均均方误差衡量。

方法	$\sigma = 0$	$\sigma = 0.5$	$\sigma = 1$	$\sigma = 5$
LASSO	5.15 (0.64)	6.49 (1.08)	4.25 (0.49)	7.42 (0.79)
OMP	5.08 (0.72)	6.35 (1.06)	5.48 (1.14)	6.41 (0.63)
RF	12.62 (2.57)	7.07 (1.34)	9.89 (2.19)	15.94 (3.60)
GB	8.19 (2.20)	4.06 (0.70)	7.08 (1.17)	10.60 (1.56)
图形套索	1.09 (0.06)	1.19 (0.07)	1.79 (0.08)	9.14 (0.32)
GSR	0.57 (0.03)	0.64 (0.04)	0.95 (0.04)	5.43 (0.37)
LNET	0.51 (0.02)	0.63 (0.02)	0.99 (0.03)	5.10 (0.34)
NN-2	0.67 (0.04)	0.74 (0.04)	1.06 (0.05)	3.91 (0.22)
NN-3	0.62 (0.03)	0.77 (0.04)	0.87 (0.04)	4.05 (0.25)

结果见表I和表 II，这表明“NN”在变量选择和预测方面的整体表现最佳。具体而言，“NN-2”和“NN-3”在所有情况下的表现几乎相同，这从经验上证明，所提出的方法同样适用于更深层的神经网络。有意思的是，可以比较“NN”和“LNET”：当噪声水平较低时，“NN”的测试误差略高于“LNET”，但其假正例率远低于“LNET”，AUC 分数则更高。这表明“NN”在变量选择方面更准确，而“LNET”倾向于选择过多变量以提高预测准确率。此外，随着噪声水平 σ 增加，所有方法的表现都会变差。

2）线性数据集：本实验考虑由线性模型 $y = \boldsymbol{x}^\top \boldsymbol{\beta} + \xi$生成的数据，其中 $\boldsymbol{\beta}$ =(3, 1.5, 0, 0, 2, 0, 0, 0)$^\top$、 $\xi \sim \mathcal{N}(0, \sigma^2)$，且 $\boldsymbol{x}$服从多元高斯分布，其第 (i, j) 个相关系数为$0.5^{|i-j|}$。在 $p = 8$ 个特征中，只有三个是显著的。训练集大小为 $n = 60$，测试集大小为 200。其他设置与III-A1小节相同。结果见表III和表IV。

结果表明，基于线性模型的方法“LASSO”和“OMP”总体表现最佳，这符合预期，因为底层数据来自线性模型。所提出的“神经网络”方法几乎与线性方法不相上下。请注意，在这种情况下，“NN-3”的表现优于“NN-2”。一种可能的解释是，深层神经网络的表达能力远强于双层网络。另一方面，基于树的方法“随机森林”和“梯度提升”的表现则明显较差。这可能是因为样本量$n = 60$ 相对较小，导致基于树的方法方差较大。同时，“神经网络”采用 ℓ_1 惩罚来缓解过参数化问题，从而识别出相关变量。此外，“神经网络”的变量选择准确率与预测性能之间呈正相关，而基于树的方法则没有这种关联。

3）Friedman 数据集：本实验使用 Friedman 数据集，其非线性数据生成过程如下：$y = 10\sin(\pi x_1 x_2) + 20(x_3 - 0.5)^2 + 10x_4 + 5x_5 + \xi$。我们生成维度为$p = 50$的标准高斯预测变量 $\boldsymbol{x}$ 。训练集大小为 $n = 500$ ，测试集大小为 2000。其他设置与之前相同。结果汇总于表V和表VI。对于这个非线性数据集，“NN”和“GB”能够准确找出显著变量并排除冗余变量，而线性方法未能选出

表III 线性数据上的性能比较：不同 σ 的正确（“真阳性”）、错误（“假阳性”）选定特征数量及 AUC 分数。

方法		$\sigma=0$	$\sigma=1$	$\sigma=3$	$\sigma=5$
LASSO	TP	3.00 (0.00)	3.00 (0.00)	2.85 (0.11)	2.05 (0.18)
	FP	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.50 (0.19)
	AUC	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	0.90 (0.03)
OMP	TP	3.00 (0.00)	2.90 (0.10)	2.95 (0.05)	2.15 (0.18)
	FP	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.30 (0.12)
	AUC	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	0.88 (0.03)
RF	TP	1.15 (0.08)	1.30 (0.12)	1.25 (0.12)	1.60 (0.16)
	FP	0.00 (0.00)	0.00 (0.00)	0.05 (0.05)	0.20 (0.11)
	AUC	1.00 (0.00)	0.99 (0.01)	0.99 (0.00)	0.83 (0.03)
GB	TP	1.35 (0.16)	1.35 (0.16)	1.30 (0.14)	1.90 (0.20)
	FP	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.30 (0.17)
	AUC	1.00 (0.00)	0.99 (0.01)	0.99 (0.01)	0.91 (0.02)
图形套索	TP	2.80 (0.13)	2.70 (0.16)	2.40 (0.19)	1.95 (0.18)
	FP	0.05 (0.05)	0.05 (0.05)	0.05 (0.05)	0.70 (0.20)
	AUC	1.00 (0.00)	1.00 (0.00)	0.99 (0.00)	0.80 (0.04)
GSR	TP	2.90 (0.10)	2.90 (0.07)	2.80 (0.13)	1.90 (0.16)
	FP	0.00 (0.00)	0.10 (0.10)	0.00 (0.00)	0.55 (0.15)
	AUC	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	0.84 (0.03)
LNET	TP	3.00 (0.00)	3.00 (0.00)	2.85 (0.15)	1.70 (0.19)
	FP	0.00 (0.00)	0.20 (0.09)	0.95 (0.26)	0.55 (0.23)
	AUC	1.00 (0.00)	0.98 (0.01)	0.88 (0.03)	0.80 (0.04)
NN-2	TP	2.50 (0.17)	2.40 (0.17)	2.55 (0.14)	2.25 (0.17)
	FP	0.05 (0.05)	0.20 (0.15)	0.25 (0.12)	0.75 (0.18)
	AUC	0.99 (0.00)	0.99 (0.00)	0.99 (0.00)	0.89 (0.2)
NN-3	TP	3.00 (0.00)	2.65（0.15）	2.90（0.07）	2.10（0.20）
	FP	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.35 (0.15)
	AUC	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	0.92 (0.02)

表 IV 线性数据上的性能比较，以不同 σ 的平均均方误差数量为指标。

方法	$\sigma=0$	$\sigma=1$	$\sigma=3$	$\sigma=5$
LASSO	0.00 (0.00)	0.04 (0.01)	0.17 (0.03)	4.09（0.45）
OMP	0.00 (0.00)	0.02 (0.00)	0.09 (0.01)	4.19 (0.44)
RF	3.57 (0.22)	3.58 (0.21)	3.52（0.16）	7.89 (0.55)
GB	2.45（0.15）	3.04（0.17）	3.01（0.17）	11.59 (0.81)
图形套索	0.10 (0.01)	0.21 (0.02)	0.34 (0.03)	4.95 (0.32)
GSR	0.09 (0.01)	0.18 (0.02)	0.30 (0.04)	4.16 (0.34)
LNET	0.18 (0.02)	0.17 (0.02)	0.34 (0.04)	3.70 (0.59)
NN-2	0.09 (0.01)	0.19 (0.02)	0.37 (0.04)	3.96 (1.37)
NN-3	0.03 (0.00)	0.10 (0.02)	0.17 (0.02)	3.17 (0.45)

二次因子 x_3。就预测性能而言，基于神经网络的方法优于其他方法。具体来说，“NN”优于“GLASSO”和“GSR”，而“LNET”表现出更好的预测性能和较差的变量选择性能，这与此前的实验结果一致。

B. BGSBoy 数据集

BGSBoy 数据集包含来自伯克利指导研究（BGS）的 66 名男孩，他们于 1928-29 年出生在加利福尼亚州伯克利[36]。该数据集包括他们在不同年龄（2、9、18 岁）的身高（‘HT’）、体重（‘WT’）、腿围（‘LG’）、力量（‘ST’）以及身体质量指数（‘BMI18’）。我们选择‘BMI18’作为响应变量，其定义如下。

$$\mathrm{BMI18} = \mathrm{WT18}/(\mathrm{HT18}/100)^2, \tag{14}$$

其中，WT18 和 HT18 分别表示 18 岁时的体重和身高。换言之，对于 $p=10$ 个变量而言，‘WT18’ 和 ‘HT18’ 足以用于对响应变量进行建模。

表 V Friedman 数据的性能比较：正确（‘TP’）、错误（‘FP’）选定特征的数量，以及不同 σ的 AUC 分数。

方法		$\sigma=0$	$\sigma=0.5$	$\sigma=1$	$\sigma=5$
LASSO	TP	4.05 (0.05)	4.00 (0.00)	4.10 (0.07)	4.00 (0.10)
	FP	1.45 (0.47)	1.85 (0.49)	2.00 (0.42)	3.10 (0.61)
	AUC	0.91 (0.01)	0.91 (0.01)	0.91 (0.01)	0.90 (0.01)
OMP	TP	4.00 (0.00)	4.00 (0.00)	4.00 (0.00)	3.80 (0.15)
	FP	0.10 (0.07)	0.10 (0.07)	0.10 (0.07)	0.05 (0.05)
	AUC	0.90 (0.00)	0.90 (0.00)	0.90 (0.00)	0.89 (0.01)
RF	TP	4.60 (0.27)	4.60 (0.27)	4.80 (0.19)	4.10 (0.29)
	FP	0.10 (0.07)	0.00 (0.00)	0.05 (0.05)	0.25 (0.12)
	AUC	1.00 (0.00)	1.00 (0.00)	1.00（0.00）	1.00（0.00）
GB	TP	4.70（0.21）	4.90（0.10）	4.80（0.19）	4.30（0.25）
	FP	0.00 (0.00)	0.00 (0.00)	0.05 (0.05)	0.40 (0.18)
	AUC	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)
图形套索	TP	4.80 (0.15)	4.80 (0.09)	4.35 (0.20)	3.80 (0.09)
	FP	0.05 (0.05)	0.25 (0.12)	0.10 (0.10)	0.95 (0.37)
	AUC	1.00 (0.00)	0.99（0.01）	0.99（0.00）	0.88 (0.01)
GSR	TP	4.20（0.22）	4.60 (0.18)	4.70（0.16）	4.00（0.07）
	FP	0.15（0.11）	0.35 (0.11)	0.25（0.10）	2.30（0.52）
	AUC	1.00 (0.00)	1.00 (0.00)	1.00（0.00）	0.91（0.01）
LNET	TP	4.75 (0.24)	4.75 (0.24)	5.00（0.00）	3.45 (0.17)
	FP	21.20 (2.00)	25.90 (1.97)	31.35 (1.55)	2.15 (1.24)
	AUC	0.74 (0.02)	0.69（0.02）	0.65 (0.02)	0.91（0.02）
NN-2	TP	4.80（0.13）	4.40 (0.20)	4.80（0.13）	4.20（0.13）
	FP	0.50 (0.22)	0.50（0.26）	0.60 (0.29)	1.25 (0.32)
	AUC	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	0.95（0.01）
NN-3	TP	4.85（0.08）	4.90（0.07）	4.85（0.08）	4.45（0.11）
	FP	0.35（0.16）	0.25（0.12）	0.55（0.22）	3.85 (0.70)
	AUC	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	0.96 (0.01)

表 VI 不同 σ 的 Friedman 数据性能比较（以平均均方误差计）。

方法	$\sigma=0$	$\sigma=0.5$	$\sigma=1$	$\sigma=5$
LASSO	6.16（0.05）	6.17（0.06）	6.07 (0.05)	6.97 (0.12)
OMP	5.95（0.07）	6.01（0.07）	6.04 (0.06)	6.41 (0.16)
RF	5.21 (0.05)	5.19 (0.09)	5.30 (0.07)	7.81 (0.15)
GB	2.70（0.05）	2.74（0.04）	2.82（0.06）	6.33（0.15）
图形套索	4.49 (0.12)	4.64 (0.17)	5.43 (0.14)	9.89 (0.23)
GSR	1.57 (0.05)	1.73 (0.07)	2.31 (0.11)	7.49 (0.19)
LNET	1.05 (0.19)	1.29 (0.26)	1.52 (0.12)	9.08 (0.47)
NN-2	1.58 (0.04)	1.71 (0.06)	2.14 (0.06)	5.86 (0.11)
NN-3	1.44 (0.04)	1.62 (0.04)	1.92 (0.04)	6.02 (0.16)

其他变量彼此相关但存在冗余。训练集大小为 $n=44$，测试集大小为 22。其他设置与之前相同。我们比较预测性能，并探究每种方法最常选取的三个特征。结果汇总于表 VII。

从结果来看，线性方法和基于神经网络的方法都能在所有重复实验中识别出‘WT18’和‘HT18’。与此同时，基于树的方法可能会漏掉‘HT18’，转而选择‘LG18’；后者仅与响应变量相关。有趣的是，我们发现线性方法在本次实验中仍然预测良好。一种可能的原因是，公式（14）在‘HT18’约为 180 时可通过一阶泰勒展开很好地近似，而且‘HT18’的取值范围集中在 180 附近的一个较小区间内。

C. UJIIndoorLoc 数据集

UJIIndoorLoc 数据集旨在通过 WiFi 指纹定位以及建筑物和楼层编号等其他变量，解决室内定位问题。详细描述

表 VII BGSBOY 数据集上不同方法的实验结果。均方根误差：均方根误差的均值（标准误）。排名前 3 的特征：特征名称（20 次中被选中的次数）。

方法	LASSO	**OMP**	**RF**	**GB**	图形套索	**GSR**	**LNET**	**NN-2**	**NN-3**
RMSE	0.05 (0.00)	0.04 (0.00)	3.10 (0.37)	2.32 (0.30)	0.15 (0.06)	0.09 (0.03)	0.03 (0.00)	0.06 (0.02)	0.12 (0.02)
排名前三的高频选中特征	WT18 (20) HT18（20） HT9 (16)	WT18 (20) HT18（20） ST18 (7)	WT18（20） LG18（20） HT18 (10)	WT18（20） LG18（20） HT18 (19)	WT18（20） HT18（20） LG18 (8)	WT18（20） HT18 (20) LG18 (12)	WT18（20） HT18 (20) HT9 (9)	WT18（20） HT18 (20) HT9 (12)	WT18（20） HT18 (20) LG18 (17)

可见于 [37]。具体而言，我们拥有 520 个无线接入点（WAP）的信号（属于连续变量），以及“楼层”、“建筑物”、“空间编号”、“相对位置”、“用户编号”和“手机编号”等分类变量。响应变量是用户的经度（“经度”）。该数据集包含 19937 个观测值。我们随机抽取 3000 个观测值，并将其中 $n = 2000$ 个用于训练，1000 个用于测试。在预处理过程中，我们为分类变量创建二元虚拟变量，因此总共得到 $p = 681$ 个变量。我们探究每种方法最常选出的十个特征，并将第十大的变量重要性设为截断值。该过程独立重复 100 次。结果报告于表 VIII。

根据结果，‘神经网络’的预测性能最佳，显著优于其他方法。在变量选择方面，基于我们的领域知识，‘建筑物’对位置有很大影响，因此它在每次重复实验中都被所有方法选中，这并不令人意外。然而，除‘建筑物’外，不同方法选择的变量各不相同，但也有一些重叠，例如，‘GLASSO’和‘梯度提升’选中了‘手机编号_14’，‘神经网络’和‘LASSO’选中了‘用户编号_16’，这些变量可能具有重要性。‘LNET’再次比其他方法选中了更多变量。每次重复实验中，‘LNET’选中的变量都接近 60 个。尽管如此，这些方法在变量选择上并未达成一致。‘神经网络’表明，所有无线接入点的信号都很弱，而类别变量则提供了更多有关用户位置的信息。鉴于无线接入点信号的缺失率极高（平均为 97%，如 [37] 所报告），‘神经网络’的解释似乎合理。

D. 总结

实验结果表明以下几点。首先，“神经网络”能够稳定地识别重要变量，并且与基线方法相比具有竞争力的预测性能。其次，噪声水平的增加会同时阻碍变量选择和预测性能的提升。第三，LASSO 正则化对于神经网络避免过拟合至关重要，尤其是在小样本数据中。使用组LASSO或混合类型的惩罚项，其性能与神经网络相近，而LNET则倾向于过度选择重要变量。第四，对于神经网络而言，变量选择性能与预测性能通常呈正相关，但对于基线方法而言，情况可能并非如此。

IV. 总结

我们建立了在双层 ReLU 神经网络中使用 LASSO 的理论。具体而言，我们证明了在合理条件下，LASSO 估计量能够稳定地重构神经网络系数，并识别关键的底层变量。我们还提出了一种实用的方法来求解优化问题并进行变量选择。下面简要讨论一些值得关注的后续研究方向。首先，本研究的局限之一是只考虑了较小的 r。一个有趣的未来研究问题是探讨可能随 p 和 n 快速增长的 r。其次，我们的实验表明，该算法可以扩展到更深层的神经网络。将主定理推广到多层情形将是很有意义的。

附录包含证明和实验细节。

致谢

作者谨向匿名审稿人和副编辑致谢，感谢他们提出的建设性审稿意见。本研究部分得到了美国国家科学基金会资助，项目编号为 DMS-2134148 和 CNS- 2220286。

附录 A 分析：定理 2 的证明

令 $\mathcal{S}$ 为指标集，其基数为 S，由 $\boldsymbol{W}^\star$ 的支撑集和 $\boldsymbol{W}$ 的最大项组成，其中 S 将在下文给出。定义

$$\boldsymbol{W} := \widehat{\boldsymbol{W}}_{\mathcal{S}} \in \mathbb{R}^{p\times r},$$

$\widehat{\ }$和 $a_j = \widehat{a}_j, \quad b_j = b_{j\circ}$ 定义

$$d_1(\boldsymbol{w}_1, a_1, b_1, \boldsymbol{w}_2, a_2, b_2) = \begin{cases} \|\boldsymbol{w}_1 - \boldsymbol{w}_2\|_1 + |b_1 - b_2| & \text{if } a_1 = a_2; \\ \|\boldsymbol{w}_1\|_1 + \|\boldsymbol{w}_2\|_1 + |b_1| + |b_2| & \text{if } a_1 \neq a_2, \end{cases} \tag{15}$$

and

$$d_2(\boldsymbol{w}_1, a_1, b_1, \boldsymbol{w}_2, a_2, b_2) = \begin{cases} \sqrt{\|\boldsymbol{w}_1 - \boldsymbol{w}_2\|_2^2 + |b_1 - b_2|^2} & \text{if } a_1 = a_2; \\ 1 & \text{if } a_1 \neq a_2. \end{cases} \tag{16}$$

此外，对于置换 π $[r]$，令

$$D_1 := \min_{\pi} \sum_{j=1}^{r} d_1(\boldsymbol{w}_{\pi(j)}, a_{\pi(j)}, b_{\pi(j)}, \boldsymbol{w}_j^\star, a_j^\star, b_j^\star), \tag{17a}$$

$$D_2 := \min_{\pi} \sqrt{\sum_{j=1}^{r} d_2(\boldsymbol{w}_{\pi(j)}, a_{\pi(j)}, b_{\pi(j)}, \boldsymbol{w}_j^\star, a_j^\star, b_j^\star)^2} \tag{17b}$$

表 VIII UJIIndoor 数据集上不同方法的实验结果。均方根误差：均方根误差的均值（标准误）。前 10 个特征：特征名称（20 次中被选中的次数）。

方法	LASSO	**OMP**	**RF**	**GB**	
RMSE	14.20 (0.08)	16.75 (0.12)	9.60 (0.12)	11.02 (0.09)	
前10名频繁已选中特征	建筑物ID_2 (20) 建筑物ID_1 (20) 用户编号_9 (20) 用户编号_16 (20) 用户编号_10 (18) WAP099（18） 用户编号_7 (14) WAP121 (10) WAP118 (8) WAP039 (7)	建筑物ID_2 (20) 建筑物ID_1 (20) WAP099 (17) 用户编号_10 (17) 用户编号_16 (14) 用户编号_7 (13) 用户编号_9 (10) WAP120 (8) WAP124 (8) WAP101 (8)	建筑物ID_1 (20) 建筑物ID_2 (20) WAP141 (16) WAP120 (16) WAP117（14） WAP035 (13) WAP173 (13) WAP167 (10) WAP118 (8) WAP113 (8)	建筑物ID_2 (20) WAP120 (20) 建筑物ID_1 (20) WAP141（17） WAP099 (16) 手机编号_14 (14) WAP113 (13) WAP114 (12) WAP117 (12) WAP140 (9)	
方法	图形套索	**GSR**	**LNET**	**NN-2**	**NN-3**
RMSE	11.25 (0.14)	11.44 (0.18)	11.19 (0.09)	8.86（0.09）	8.86 (0.09)
前10名频繁已选中特征	建筑物ID_1 (20) 建筑物ID_2 (20) 手机编号_14 (18) 空间编号_202 (17) 用户编号_8 (16) 用户编号_16 (16) 楼层_3 (12) 用户编号_9 (9) WAP478 (8) WAP099 (7)	建筑物ID_1 (20) 建筑物ID_2 (20) 用户编号_16 (17) 空间编号_202 (15) 手机编号_6 (14) 空间编号_203 (14) 用户编号_7 (13) 空间编号_201 (12) WAP141 (6) WAP176 (6)	建筑物ID_1 (20) 建筑物ID_2 (20) 手机编号_22 (20) 手机编号_6 (20) 手机编号_8 (20) 空间编号_103 (20) 空间编号_136 (20) 空间编号_201 (20) 空间编号_202 (20) 空间编号_203 (20)	空间编号_202 (20) 建筑物ID_1 (20) 建筑物ID_2 (20) 用户编号_16 (19) 空间编号_203 (16) 空间编号_201 (14) WAP140 (7) WAP121 (5) WAP030 (5) 空间编号_224 (5)	空间编号_202 (20) 建筑物ID_1 (20) 建筑物ID_2 (20) 用户编号_16 (17) 空间编号_203 (13) 空间编号_201 (13) 用户编号_9 (8) WAP121 (8) WAP030 (7) 空间编号_224 (5)

分别表示 $(\boldsymbol{W}, \boldsymbol{a}, \boldsymbol{b})$ 与 $(\boldsymbol{W}^\star, \boldsymbol{a}^\star, \boldsymbol{b}^\star)$ 之间的 D_1-距离和 D_2-距离。于是，有如下界。

引理 1. 对任意满足 $\|\boldsymbol{W}\|_0 \le S$ 的 $\boldsymbol{W} \in \mathbb{R}^{p\times r}$，存在一些普适常数 c_4、$c_5 > 0$，使得

$$\frac{1}{n}\sum_{i=1}^{n}\left[\sum_{j=1}^{r} a_j \mathrm{relu}(\boldsymbol{w}_j^\top \boldsymbol{x}_i + b_j) - \sum_{j=1}^{r} a_j^\star \mathrm{relu}(\boldsymbol{w}_j^{\star\top}\boldsymbol{x}_i + b_j^\star)\right]^2 \ge c_4 \min\left\{\frac{1}{r}, D_2^2\right\} \tag{18}$$

以至少 $1-\delta$ 的概率成立，前提是

$$n \ge c_5 S^3 r^4 \log^4 \frac{p}{\delta}. \tag{19}$$

引理 2。因此，存在一个普适常数 $c_6 > 0$，使得

$$\frac{1}{n}\sum_{i=1}^{n}\left[\sum_{j=1}^{r} a_j \mathrm{relu}(\boldsymbol{w}_j^\top \boldsymbol{x}_i + b_j) - \sum_{j=1}^{r} a_j^\star \mathrm{relu}(\boldsymbol{w}_j^{\star\top}\boldsymbol{x}_i + b_j^\star)\right]^2 \le c_6\left(\frac{r}{S} + \frac{r\log^3\frac{p}{n\delta}}{n}\right) D_1^2 + c_6\sigma^2$$

以至少 $1-\delta$ 的概率成立。

通过比较引理 1 和引理 2 给出的界，可得

$$c_4 \min\left\{\frac{1}{r}, D_2^2\right\} \le c_6\left(\frac{r}{S} + \frac{r\log^3\frac{p}{n\delta}}{n}\right) D_1^2 + c_6\sigma^2,$$

前提是

$$n > c_5 S^3 r^4 \log^4 \frac{p}{\delta}.$$

令 $\mathcal{S}^\star$ 为基数为 $2s$ 的指标集，由 $\boldsymbol{W}^\star$ 的支撑集和 $\boldsymbol{W}$ 的前几个元素组成。此外，令$D_1^\star$和 $D_2^\star$分别表示 $\left(\widehat{\boldsymbol{W}}_{\widehat{\mathcal{S}}^\star}, \widehat{\boldsymbol{a}}, \widehat{\boldsymbol{b}}\right)$ 与 $(\boldsymbol{W}^\star, \boldsymbol{a}^\star, \boldsymbol{b}^\star)$ 之间的 D_1-距离和 D_2-距离，定义方式与 (17) 类似。注意到，对于 $S \ge 2s$，有 $\mathcal{S}^\star \subset \mathcal{S}^\star \subset \mathcal{S}$，因此我们有

$$\|\boldsymbol{w}_j - \boldsymbol{w}_j^\star\|_2 \ge \|\boldsymbol{w}_{j,\widehat{\mathcal{S}}^\star} - \boldsymbol{w}_j^\star\|_2 = \|\widehat{\boldsymbol{w}}_j - \boldsymbol{w}_j^\star\|_2,$$

经过某个置换后，然后

$$D_2^\star \le D_2.$$

此外，经过某个置换后，我们有

$$D_1^\star \ge \left\|\widehat{\boldsymbol{W}}_{\widehat{\mathcal{S}}^\star} - \boldsymbol{W}^\star\right\|_1 \ge \|\boldsymbol{W}^\star\|_1 - \left\|\widehat{\boldsymbol{W}}_{\mathcal{S}^\star}\right\|_1 \qquad \|\boldsymbol{W}\|_1 \le \left\|\widehat{\boldsymbol{W}}\right\|_1 \le \|\boldsymbol{W}^\star\|_1$$

和。于是，

$$D_1 \le D_1^\star + \left\|\widehat{\boldsymbol{W}}_{\mathcal{S}} - \widehat{\boldsymbol{W}}_{\widehat{\mathcal{S}}^\star}\right\|_1 \le D_1^\star + \left\|\widehat{\boldsymbol{W}}\right\|_1 - \left\|\widehat{\boldsymbol{W}}_{\mathcal{S}^\star}\right\|_1 \le 2D_1^\star.$$

结合附录 D 中的引理 3，上述结果给出

$$D_2^\star \le \frac{2c_6}{c_4}\sigma,$$

前提是对于某个常数 $c_7 > 0$

$$n \ge c_5 S^3 \log^4 \frac{p}{\delta} \qquad \text{with } S \ge c_7 sr,$$

使得

$$c_6\left(\frac{r}{S} + \frac{r\log^3\frac{p}{n\delta}}{n}\right) D_1^{\star 2} \le \frac{c_4}{8} D_2^{\star 2}.$$

然后，经过适当的置换后，我们即可得出证明。

$$\|\widehat{\boldsymbol{W}}-\boldsymbol{W}^\star\|_{\mathrm{F}} \le 2\|\widehat{\boldsymbol{W}}_{\widehat{\mathcal{S}}^\star}-\boldsymbol{W}^\star\|_{\mathrm{F}}.$$

Appendix B
Proof of Lemma 1 (lower bound)

这可由以下三个性质看出。

- 考虑这种情况:

$$D_1 \le \epsilon = \frac{\delta}{4nr}\sqrt{\frac{\pi}{\log\frac{4pn}{\delta}}}.$$

以至少 $1-\delta$ 的概率,

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^{n}\Big[\sum_{j=1}^{r}a_j\mathrm{relu}(\boldsymbol{w}_j^\top\boldsymbol{x}_i+b_j)\\&\qquad-\sum_{j=1}^{r}a_j^\star\mathrm{relu}(\boldsymbol{w}_j^{\star\top}\boldsymbol{x}_i+b_j^\star)\Big]^2\\&=\frac{D_1^2}{\epsilon^2}\frac{1}{n}\sum_{i=1}^{n}\Big[\sum_{j=1}^{r}a_j\mathrm{relu}(\widetilde{\boldsymbol{w}}_j^\top\boldsymbol{x}_i+\widetilde{b}_j)\\&\qquad-\sum_{j=1}^{r}a_j^\star\mathrm{relu}(\boldsymbol{w}_j^{\star\top}\boldsymbol{x}_i+b_j^\star)\Big]^2,\end{aligned}\tag{20}$$

其中

$$\widetilde{\boldsymbol{w}}_j = \boldsymbol{w}_j^\star+\tfrac{\epsilon}{D_1}(\boldsymbol{w}_j-\boldsymbol{w}_j^\star)\qquad \widetilde{b}_j = b_j^\star+\tfrac{\epsilon}{D_1}(b_j-b_j^\star)$$

且。

- 对于任意 $\epsilon>0$ 和

$$D_1 \ge \frac{\epsilon}{\sqrt{\frac{S}{n}\log\frac{pr}{S}\log\frac{BS}{\epsilon\delta}}},$$

存在某个普适常数 $C_1>0$，使得以至少 $1-\delta$ 的概率,

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^{n}\Big[\sum_{j=1}^{r}a_j\mathrm{relu}(\boldsymbol{w}_j^\top\boldsymbol{x}_i+b_j)\\&\qquad-\sum_{j=1}^{r}a_j^\star\mathrm{relu}(\boldsymbol{w}_j^{\star\top}\boldsymbol{x}_i+b_j^\star)\Big]^2\\&\ge\mathbb{E}\left[\sum_{j=1}^{r}a_j\mathrm{relu}(\boldsymbol{w}_j^\top\boldsymbol{x}+b_j)-\sum_{j=1}^{r}a_j^\star\mathrm{relu}(\boldsymbol{w}_j^{\star\top}\boldsymbol{x}+b_j^\star)\right]^2\\&\quad-C_1D_1^2\log\frac{pn}{\delta}\sqrt{\frac{S}{n}\log\frac{pr}{S}\log\frac{BS}{\epsilon\delta}}.\end{aligned}\tag{21}$$

- 对于某个普适常数 $C_2>0$

$$\begin{aligned}&\mathbb{E}\left[\sum_{j=1}^{r}a_j\mathrm{relu}(\boldsymbol{w}_j^\top\boldsymbol{x}+b_j)-\sum_{j=1}^{r}a_j^\star\mathrm{relu}(\boldsymbol{w}_j^{\star\top}\boldsymbol{x}+b_j^\star)\right]^2\\&\ge C_2\min\left\{\frac{1}{r},D_2^2\right\}.\end{aligned}\tag{22}$$

综合以上内容。令

$$\epsilon = C_3\frac{\delta}{nr}\sqrt{\frac{S}{n}\log\frac{BnS}{\delta}},$$

对于某个普适常数 $C_3>0$，使得

$$\frac{\epsilon}{\sqrt{\frac{S}{n}\log\frac{pr}{S}\log\frac{BS}{\epsilon\delta}}} < \frac{\delta}{4nr}\sqrt{\frac{\pi}{\log\frac{4pn}{\delta}}}.$$

将 (22) 代入 (21) 可得

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^{n}\left[\sum_{j=1}^{r}a_j\mathrm{relu}(\boldsymbol{w}_j^\top\boldsymbol{x}_i+b_j)-\sum_{j=1}^{r}a_j^\star\mathrm{relu}(\boldsymbol{w}_j^{\star\top}\boldsymbol{x}_i+b_j^\star)\right]^2\\&\ge C_2\min\left\{\frac{1}{r},D_2^2\right\}-C_1D_1^2\log\frac{pn}{\delta}\sqrt{\frac{S}{n}\log\frac{pr}{S}\log\frac{BS}{\epsilon\delta}}\\&\ge\frac{C_2}{2}\min\left\{\frac{1}{r},D_2^2\right\},\end{aligned}\tag{23}$$

以至少 $1-\delta$ 的概率成立，前提是对于某个常数 $C_4>0$

$$n \ge C_4S^3r^4\log\frac{pr}{S}\log\frac{BS}{\epsilon\delta}\log^2\frac{pn}{\delta}\quad\text{and}$$
$$D_1 \ge \frac{\delta}{4nr}\sqrt{\frac{\pi}{\log\frac{4pn}{\delta}}}.$$

这里，最后一行由引理 3 得出，并且我们假设 $\max\{\|\boldsymbol{W}\|_\infty,\|\boldsymbol{b}\|_\infty\}$ 被某个常数所界定。另一方面，如果

$$D_1 < \frac{\delta}{4nr}\sqrt{\frac{\pi}{\log\frac{4pn}{\delta}}},$$

由 (20) 和 (23) 可知

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^{n}\left[\sum_{j=1}^{r}a_j\mathrm{relu}(\boldsymbol{w}_j^\top\boldsymbol{x}_i+b_j)-\sum_{j=1}^{r}a_j^\star\mathrm{relu}(\boldsymbol{w}_j^{\star\top}\boldsymbol{x}_i+b_j^\star)\right]^2\\&\ge\frac{D_1^2}{\epsilon^2}\frac{C_2}{2}\min\left\{\frac{1}{r},\widetilde{D}_2^2\right\}=\frac{C_2}{2}D_2^2.\end{aligned}$$

综上，通过验证下文中的 (20)、(21) 和 (22)，即可完成证明。

A. (20) 的证明

由于 $D_1 \le \epsilon = \frac{\delta}{4nr}\sqrt{\frac{\pi}{\log\frac{4pn}{\delta}}}$，不失一般性，假设 $a_j=a_j^\star$ for $1\le j\le r$，且

$$D_1=\sum_{j=1}^{r}\left(\|\boldsymbol{w}_j-\boldsymbol{w}_j^\star\|_1+|b_j-b_j^\star|\right)\le\epsilon.$$

通过取并集界，以至少 $1-\frac{\delta}{2}$ 的概率，对所有 $1\le i\le n$ 和 $1\le j\le r$，都有:

$$\left|\boldsymbol{w}_j^{\star\top}\boldsymbol{x}_i+b_j^\star\right|>\frac{\delta}{2nr}\sqrt{\frac{\pi}{2}},$$

由于 $\|\boldsymbol{w}_j^\star\|_2\ge1$ 和 $\boldsymbol{x}_i\sim\mathcal{N}(\boldsymbol{0},\boldsymbol{I})$。此外，对于所有 $1\le i\le n$ 和 $1\le j\le r$,

$$\begin{aligned}\left|\boldsymbol{w}_j^\top\boldsymbol{x}_i+b_j-\boldsymbol{w}_j^{\star\top}\boldsymbol{x}_i-b_j^\star\right|&\le\|\boldsymbol{w}_j-\boldsymbol{w}_j^\star\|_1\|\boldsymbol{x}_i\|_\infty+|b_j-b_j^\star|\\&\le\epsilon\sqrt{2\log\frac{4pn}{\delta}}\end{aligned}$$

以至少 $1-\frac{\delta}{2}$ 的概率成立。这里，最后一个不等式是由以下事实得出的：以至少 $1-\frac{\delta}{2}$ 的概率，

$$\|\boldsymbol{x}_i\|_\infty \le \sqrt{2\log\frac{4pn}{\delta}} \qquad \text{for all } 1\le i\le n. \tag{24}$$

综上，我们以至少 $1-\delta$ 的概率有，

$$u(\boldsymbol{w}_j^\top \boldsymbol{x}_i + b_j) = u(\boldsymbol{w}_j^{\star\top}\boldsymbol{x}_i + b_j^\star), \tag{25}$$

前提是

$$\epsilon \le \frac{\delta}{4nr}\sqrt{\frac{\pi}{\log\frac{4pn}{\delta}}}.$$

注意，当 $x>0$ 时，$u(x)=1$；当 $x\le 0$ 时，$u(x)=0$。那么结合 $\widetilde{\boldsymbol{w}}_j$ 和 b_j 的定义，上述性质可得搭配

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^n\left[\sum_{j=1}^r a_j\mathrm{relu}(\boldsymbol{w}_j^\top\boldsymbol{x}_i+b_j)-\sum_{j=1}^r a_j^\star\mathrm{relu}(\boldsymbol{w}_j^{\star\top}\boldsymbol{x}_i+b_j^\star)\right]^2\\
&=\frac{1}{n}\sum_{i=1}^n\left[\sum_{j=1}^r a_j^\star u(\boldsymbol{w}_j^{\star\top}\boldsymbol{x}_i+b_j^\star)(\boldsymbol{w}_j^\top\boldsymbol{x}_i+b_j-\boldsymbol{w}_j^{\star\top}\boldsymbol{x}_i-b_j^\star)\right]^2\\
&=\frac{D_1^2}{\epsilon^2}\frac{1}{n}\sum_{i=1}^n\Bigg[\sum_{j=1}^r a_j^\star u(\boldsymbol{w}_j^{\star\top}\boldsymbol{x}_i+b_j^\star)\\
&\qquad\qquad\times(\widetilde{\boldsymbol{w}}_j^\top\boldsymbol{x}_i+\widetilde{b}_j-\boldsymbol{w}_j^{\star\top}\boldsymbol{x}_i-b_j^\star)\Bigg]^2\\
&=\frac{D_1^2}{\epsilon^2}\frac{1}{n}\sum_{i=1}^n\Bigg[\sum_{j=1}^r a_j\mathrm{relu}(\widetilde{\boldsymbol{w}}_j^\top\boldsymbol{x}_i+\widetilde{b}_j)\\
&\qquad\qquad-\sum_{j=1}^r a_j^\star\mathrm{relu}(\boldsymbol{w}_j^{\star\top}\boldsymbol{x}_i+b_j^\star)\Bigg]^2,
\end{aligned}$$

并且该结论得证。这里，最后一个等式成立是因为 (25) 和 $a_j=a_j^\star$ 对于 $j=1$、.....、r。

B. 式 (21) 的证明

注意到

$$\begin{aligned}
&\left|a_j\mathrm{relu}(\boldsymbol{w}_j^\top\boldsymbol{x}+b_j)-a_j^\star\mathrm{relu}(\boldsymbol{w}_j^{\star\top}\boldsymbol{x}_i+b_j^\star)\right|\\
&\le\begin{cases}\|\boldsymbol{w}_j-\boldsymbol{w}_j^\star\|_1\|\boldsymbol{x}\|_\infty+|b_j-b_j^\star| & \text{if } a_j=a_j^\star,\\ \left(\|\boldsymbol{w}_j\|_1+\|\boldsymbol{w}_j^\star\|_1\right)\|\boldsymbol{x}\|_\infty+|b_j|+|b_j^\star| & \text{if } a_j\ne a_j^\star,\end{cases}
\end{aligned}$$

从而得到

$$\begin{aligned}
&\left|\sum_{j=1}^r a_j\mathrm{relu}(\boldsymbol{w}_j^\top\boldsymbol{x}+b_j)-\sum_{j=1}^r a_j^\star\mathrm{relu}(\boldsymbol{w}_j^{\star\top}\boldsymbol{x}+b_j^\star)\right|\\
&\le D_1\max\left\{\|\boldsymbol{x}\|_\infty,1\right\}.
\end{aligned}\tag{26}$$

对于任意固定的 $(\boldsymbol{W},\boldsymbol{a},\boldsymbol{b})$，令

$$z_i \coloneqq \sum_{j=1}^r a_j\mathrm{relu}(\boldsymbol{w}_j^\top\boldsymbol{x}_i+b_j)-\sum_{j=1}^r a_j^\star\mathrm{relu}(\boldsymbol{w}_j^{\star\top}\boldsymbol{x}_i+b_j^\star),$$

并定义以下事件集合

$$\mathcal{E}\coloneqq\left\{\|\boldsymbol{x}_i\|_\infty\le\sqrt{2\log\frac{4pn}{\delta}} \qquad \text{for all } 1\le i\le n\right\}.$$

那么，以至少 $1-\delta$ 的概率，

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^n\left(z_i^2-\mathbb{E}\left[z_i^2\right]\right) &(27)\\
&=\frac{1}{n}\sum_{i=1}^n\left\{z_i^2{}_{\mathbb{I}}(\mathcal{E})-\mathbb{E}\left[z_i^2{}_{\mathbb{I}}(\mathcal{E})\right]-\mathbb{E}\left[z_i^2{}_{\mathbb{I}}(\overline{\mathcal{E}})\right]\right\}\\
&\ge -4D_1^2\log\frac{4pn}{\delta}\sqrt{\frac{1}{n}\log\frac{2}{\delta}}-D_1^2\frac{\delta}{n}\\
&\ge -5D_1^2\log\frac{4pn}{\delta}\sqrt{\frac{1}{n}\log\frac{2}{\delta}}. &(28)
\end{aligned}$$

这里，第一行由 (24) 得出；最后一行由 Hoeffding不等式以及以下事实得出：

$$\begin{aligned}
\left|\mathbb{E}\left[z_i^2{}_{\mathbb{I}}(\overline{\mathcal{E}})\right]\right| &\le D_1^2\left|\mathbb{E}\left[\|\boldsymbol{x}_i\|_\infty^2{}_{\mathbb{I}}(\|\boldsymbol{x}_i\|_\infty>\sqrt{2\log\frac{4pn}{\delta}})\right]\right|\\
&\le D_1^2\int_{\sqrt{2\log\frac{4pn}{\delta}}}^\infty x^2 d\mathbb{P}(\|\boldsymbol{x}_i\|_\infty<x)\\
&\le D_1^2\int_{\sqrt{2\log\frac{4pn}{\delta}}}^\infty 4xp\exp(-\frac{x^2}{2})dx\le D_1^2\frac{\delta}{n}.
\end{aligned}$$

此外，考虑以下 ϵ-net

$$\begin{aligned}
\mathcal{N}_\epsilon\coloneqq\Bigg\{&(\boldsymbol{W},\boldsymbol{a},\boldsymbol{b}):|W_{ij}|\in\frac{\epsilon}{r+S}\left[\left\lceil\frac{B(r+S)}{\epsilon}\right\rceil\right],\\
&\|\boldsymbol{W}\|_0\le S,|b_j|\in\frac{\epsilon}{r+S}\left[\left\lceil\frac{B(r+S)}{\epsilon}\right\rceil\right],|a_j|=1\Bigg\},
\end{aligned}$$

其中 $[n]\coloneqq\{1,2,\dots,n-1\}$。那么，对于所有满足$\|\boldsymbol{W}\|_1\le B$且 $\|\boldsymbol{b}\|_1\le B$ 的 $(\boldsymbol{W},\boldsymbol{a},\boldsymbol{b})$，在 $\mathcal{N}_\epsilon$ 中存在一点（记为 $\left(\widetilde{\boldsymbol{W}},\widetilde{\boldsymbol{a}},\widetilde{\boldsymbol{b}}\right)$），其与 $(\boldsymbol{W},\boldsymbol{a},\boldsymbol{b})$ 的 D_1-距离小于 ϵ。为简便起见，定义

$$\begin{aligned}
z_i&\coloneqq\sum_{j=1}^r a_j\mathrm{relu}(\boldsymbol{w}_j^\top\boldsymbol{x}_i+b_j)-\sum_{j=1}^r a_j^\star\mathrm{relu}(\boldsymbol{w}_j^{\star\top}\boldsymbol{x}_i+b_j^\star),\\
\widetilde{z}_i&\coloneqq\sum_{j=1}^r \widetilde{a}_j\mathrm{relu}(\widetilde{\boldsymbol{w}}_j^\top\boldsymbol{x}_i+\widetilde{b}_j)-\sum_{j=1}^r a_j^\star\mathrm{relu}(\boldsymbol{w}_j^{\star\top}\boldsymbol{x}_i+b_j^\star).
\end{aligned}$$

与 (26) 类似，我们可以推导出

$$\begin{aligned}
&\left|\sum_{j=1}^r a_j\mathrm{relu}(\boldsymbol{w}_j^\top\boldsymbol{x}+b_j)-\sum_{j=1}^r \widetilde{a}_j\mathrm{relu}(\widetilde{\boldsymbol{w}}_j^\top\boldsymbol{x}+\widetilde{b}_j)\right|\\
&\le\epsilon\max\left\{\|\boldsymbol{x}\|_\infty,1\right\},
\end{aligned}$$

这意味着

$$\left|z_i^2-\widetilde{z}_i^2\right|\le\epsilon\left(\epsilon+D_1\right)\max\left\{\|\boldsymbol{x}_i\|_\infty^2,1\right\},$$

进而，以至少 $1-\delta$ 的概率，

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^n\left(z_i^2-\mathbb{E}\left[z_i^2\right]\right)-\frac{1}{n}\sum_{i=1}^n\left(\widetilde{z}_i^2-\mathbb{E}\left[\widetilde{z}_i^2\right]\right)\\
&\ge-4\epsilon\left(\epsilon+D_1\right)\log\frac{4pn}{\delta}.
\end{aligned}\tag{29}$$

此外，可以验证

$$\log|\mathcal{N}_\epsilon|\le C_5 S\log\frac{pr}{S}\log\frac{BS}{\epsilon},\tag{30}$$

对于某个普适常数 $C_5 > 0$。结合 (28)、(29) 和 (30) 可得

$$\frac{1}{n}\sum_{i=1}^{n}\left(z_i^2 - \mathbb{E}\left[z_i^2\right]\right) \geq -5\left(\epsilon + D_1\right)^2 \log\frac{4pn}{\delta}\sqrt{\frac{1}{n}\log\frac{2\left|\mathcal{N}_\epsilon\right|}{\delta}} - 4\epsilon\left(\epsilon + D_1\right)\log\frac{4pn}{\delta}.$$

因此，（21）成立。

C. （22）的证明

我们首先考虑一种简单情形，即 $b_j = 0$ 和 $b_j^\star = 0$ 对于 $1 \leq j \leq r$，并证明对于某个足够小的常数 $C_6 > 0$,

$$\mathbb{E}\left[\sum_{j=1}^{r} a_j \mathrm{relu}(\boldsymbol{w}_j^\top \boldsymbol{x}) - \sum_{j=1}^{r} a_j^\star \mathrm{relu}(\boldsymbol{w}_j^{\star\top}\boldsymbol{x})\right]^2 \geq C_6 \min\left\{\frac{1}{r}, D_2^2\right\}. \tag{31}$$

接下来，我们假设

$$\mathbb{E}\left[\sum_{j=1}^{r} a_j \mathrm{relu}(\boldsymbol{w}_j^\top \boldsymbol{x}) - \sum_{j=1}^{r} a_j^\star \mathrm{relu}(\boldsymbol{w}_j^{\star\top}\boldsymbol{x})\right]^2 \leq \frac{C_6}{r}.$$

否则，不等式（31）已经成立。根据引理4，对于任意常数 $k \geq 0$，存在某个常数 $\alpha_k > 0$ ，使得

$$\begin{aligned}&\mathbb{E}\left[\sum_{j=1}^{r} a_j \mathrm{relu}(\boldsymbol{w}_j^\top \boldsymbol{x}) - \sum_{j=1}^{r} a_j^\star \mathrm{relu}(\boldsymbol{w}_j^{\star\top}\boldsymbol{x})\right]^2 \\ &\geq \alpha_k \left\| \sum_{j=1}^{r} a_j \|\boldsymbol{w}_j\|_2 \big(\frac{\boldsymbol{w}_j}{\|\boldsymbol{w}_j\|_2}\big)^{\otimes 2k} - \sum_{j=1}^{r} a_j^\star \|\boldsymbol{w}_j^\star\|_2 \big(\frac{\boldsymbol{w}_j^\star}{\|\boldsymbol{w}_j^\star\|_2}\big)^{\otimes 2k} \right\|_F^2.\end{aligned} \tag{32}$$

假设1告诉我们，对于任意整数 $k \geq \frac{2}{\omega}$，

$$\left|\langle \boldsymbol{v}_{j_1}^\star, \boldsymbol{v}_{j_2}^\star\rangle\right| \leq \frac{1}{r^2}. \tag{33}$$

其中

$$\boldsymbol{v}_j \coloneqq \mathrm{vec}\left(\big(\frac{\boldsymbol{w}_j}{\|\boldsymbol{w}_j\|_2}\big)^{\otimes k}\right) \quad \text{with} \quad \beta_j \coloneqq a_j\|\boldsymbol{w}_j\|_2,$$

and

$$\boldsymbol{v}_j^\star \coloneqq \mathrm{vec}\left(\big(\frac{\boldsymbol{w}_j^\star}{\|\boldsymbol{w}_j^\star\|_2}\big)^{\otimes k}\right) \quad \text{with} \quad \beta_j^\star \coloneqq a_j^\star\|\boldsymbol{w}_j^\star\|_2.$$

于是，由（32）可得

$$\mathbb{E}\left[\sum_{j=1}^{r} a_j \mathrm{relu}(\boldsymbol{w}_j^\top \boldsymbol{x}) - \sum_{j=1}^{r} a_j^\star \mathrm{relu}(\boldsymbol{w}_j^{\star\top}\boldsymbol{x})\right]^2 \geq \alpha_{3k}\left\|\sum_{j=1}^{r}\beta_j\boldsymbol{v}_j^{\otimes 6} - \sum_{j=1}^{r}\beta_j^\star\boldsymbol{v}_j^{\star\otimes 6}\right\|_F^2.$$

定义

$$\mathbb{S}_+ \coloneqq \mathrm{span}\left\{\boldsymbol{v}_j\right\}_{j:\beta_j>0} \qquad \mathbb{S}_- \coloneqq \mathrm{span}\left\{\boldsymbol{v}_j\right\}_{j:\beta_j<0},$$

and

$$\mathbb{S}_+^\star \coloneqq \mathrm{span}\left\{\boldsymbol{v}_j^\star\right\}_{j:\beta_j^\star>0} \qquad \mathbb{S}_-^\star \coloneqq \mathrm{span}\left\{\boldsymbol{v}_j^\star\right\}_{j:\beta_j^\star<0}.$$

令 $\boldsymbol{P}_{\mathbb{S}}$ 和 $\boldsymbol{P}_{\mathbb{S}}^\perp$ 分别表示投影到子空间 $\mathbb{S}$ 上和与子空间 $\mathbb{S}$ 垂直的分量。注意到，对于满足 $\beta_j < 0$ 的 j，有 $\boldsymbol{P}_{\mathbb{S}_-}^\perp \boldsymbol{v}_j = \boldsymbol{0}$; 对于满足 $\beta_j^\star > 0$ 的 j，有 $\boldsymbol{P}_{\mathbb{S}_+^\star}^\perp \boldsymbol{v}_j^\star = \boldsymbol{0}$，因此可得

$$\begin{aligned}&\left\|\sum_{j=1}^{r}\beta_j\boldsymbol{v}_j^{\otimes 6} - \sum_{j=1}^{r}\beta_j^\star\boldsymbol{v}_j^{\star\otimes 6}\right\|_F^2 \\ &\geq \Big\| \sum_{j:\beta_j>0}\beta_j\big(\boldsymbol{P}_{\mathbb{S}_-}^\perp\boldsymbol{v}_j\big)^{\otimes 2}\otimes\big(\boldsymbol{P}_{\mathbb{S}_+^\star}^\perp\boldsymbol{v}_j\big)^{\otimes 4} \\ &\qquad - \sum_{j:\beta_j^\star<0}\beta_j^\star\big(\boldsymbol{P}_{\mathbb{S}_-}^\perp\boldsymbol{v}_j^\star\big)^{\otimes 2}\otimes\big(\boldsymbol{P}_{\mathbb{S}_+^\star}^\perp\boldsymbol{v}_j^\star\big)^{\otimes 4}\Big\|_F^2 \\ &\geq \sum_{j:\beta_j^\star<0}\Big\|\beta_j^\star\big(\boldsymbol{P}_{\mathbb{S}_-}^\perp\boldsymbol{v}_j^\star\big)^{\otimes 2}\otimes\big(\boldsymbol{P}_{\mathbb{S}_+^\star}^\perp\boldsymbol{v}_j^\star\big)^{\otimes 4}\Big\|_F^2 \\ &\geq \frac{1}{2}\sum_{j:\beta_j^\star<0}\left\|\boldsymbol{P}_{\mathbb{S}_-}^\perp\boldsymbol{v}_j^\star\right\|_2^4,\end{aligned}$$

其中倒数第二个不等式成立，是因为每一对项之间的内积均为正; 最后一个不等式则由以下事实得出：$|\beta_j^\star| \geq 1$ 以及（33）。

此外，通过算术-几何平均不等式和（33），可以看出。

$$\begin{aligned}\sum_{j:\beta_j^\star<0}\left\|\boldsymbol{P}_{\mathbb{S}_-}^\perp\boldsymbol{v}_j^\star\right\|_2^4 &\geq \frac{1}{r}\Big(\sum_{j:\beta_j^\star<0}\left\|\boldsymbol{P}_{\mathbb{S}_-}^\perp\boldsymbol{v}_j^\star\right\|_2^2\Big)^2 \\ &= \frac{1}{r}\left\|\boldsymbol{P}_{\mathbb{S}_-}^\perp\left[\boldsymbol{v}_j^\star\right]_{j:\beta_j^\star<0}\right\|_F^4 \\ &\geq \frac{1}{2r}\left\|\boldsymbol{P}_{\mathbb{S}_-}^\perp\boldsymbol{P}_{\mathbb{S}_-^\star}\right\|_F^4.\end{aligned}$$

然后结合（31）、上述结果以及对应的 $\beta_j^\star > 0$ 情形，可得

$$\dim(\mathbb{S}_-) \geq \dim(\mathbb{S}_-^\star) \quad \text{and} \quad \dim(\mathbb{S}_+) \geq \dim(\mathbb{S}_+^\star),$$

由此可得

$$\dim(\mathbb{S}_-) = \dim(\mathbb{S}_-^\star) \quad \text{and} \quad \dim(\mathbb{S}_+) = \dim(\mathbb{S}_+^\star).$$

此外，对于某个较小的常数 $C_6 > 0$，我们有

$$\mathrm{dist}(\mathbb{S}_-, \mathbb{S}_-^\star) \leq C_6 \quad \text{and} \quad \mathrm{dist}(\mathbb{S}_+, \mathbb{S}_+^\star) \leq C_6.$$

令 $\boldsymbol{P}_i^\perp$ 表示垂直于.....的投影

$$\mathrm{span}\left\{\boldsymbol{v}_j^\star\right\}_{j\neq i:\beta_j^\star>0},$$

and

$$\gamma_j \coloneqq \frac{\beta_j\langle\boldsymbol{P}_{\mathbb{S}_-}^\perp\boldsymbol{v}_j, P_{\mathbb{S}_-}^\perp\boldsymbol{v}_i^\star\rangle^2\langle\boldsymbol{P}_i^\perp\boldsymbol{v}_i, P_{\mathbb{S}_-}^\perp\boldsymbol{v}_i^\star\rangle^2}{\left\|\boldsymbol{P}_{\mathbb{S}_-}^\perp\boldsymbol{v}_i^\star\right\|_2^2\left\|\boldsymbol{P}_i^\perp\boldsymbol{v}_i^\star\right\|^2}.$$

那么，对于任意 i,

$$\begin{aligned}
&\left\|\sum_{j=1}^{r}\beta_j\boldsymbol{v}_j^{\otimes 6}-\sum_{j=1}^{r}\beta_j^\star\boldsymbol{v}_j^{\star\otimes 6}\right\|_F^2\\
&\geq\Big\|\sum_{j:\beta_j>0}\beta_j\big(\boldsymbol{P}_{\mathbb{S}_-}^{\perp}\boldsymbol{v}_j\big)^{\otimes 2}\otimes\boldsymbol{v}_j^{\otimes 4}\\
&\qquad-\sum_{j=1}^{r}\beta_j^\star\big(\boldsymbol{P}_{\mathbb{S}_-}^{\perp}\boldsymbol{v}_j^\star\big)^{\otimes 2}\otimes\boldsymbol{v}_j^{\star\otimes 4}\Big\|_F^2\\
&\geq\frac{1}{2}\Big\|\sum_{j:\beta_j>0}\beta_j\big(\boldsymbol{P}_{\mathbb{S}_-}^{\perp}\boldsymbol{v}_j\big)^{\otimes 2}\otimes\boldsymbol{v}_j^{\otimes 4}\\
&\qquad-\sum_{j:\beta_j^\star>0}\beta_j^\star\big(\boldsymbol{P}_{\mathbb{S}_-}^{\perp}\boldsymbol{v}_j^\star\big)^{\otimes 2}\otimes\boldsymbol{v}_j^{\star\otimes 4}\Big\|_F^2\\
&\geq\frac{1}{2}\Big\|\sum_{j:\beta_j>0}\beta_j\big(\boldsymbol{P}_{\mathbb{S}_-}^{\perp}\boldsymbol{v}_j\big)^{\otimes 2}\otimes\big(\boldsymbol{P}_i^{\perp}\boldsymbol{v}_i\big)^{\otimes 2}\otimes\boldsymbol{v}_j^{\otimes 2}\\
&\qquad-\beta_i^\star\big(\boldsymbol{P}_{\mathbb{S}_-}^{\perp}\boldsymbol{v}_i^\star\big)^{\otimes 2}\otimes\big(\boldsymbol{P}_i^{\perp}\boldsymbol{v}_i^\star\big)^{\otimes 2}\otimes\boldsymbol{v}_i^{\star\otimes 2}\Big\|_F^2\\
&\geq\frac{1}{2}\left\|\sum_{j:\beta_j>0}\gamma_j\boldsymbol{v}_j^{\otimes 2}-\beta_i^\star\big\|\boldsymbol{P}_{\mathbb{S}_-}^{\perp}\boldsymbol{v}_i^\star\big\|_2^2\big\|\boldsymbol{P}_i^{\perp}\boldsymbol{v}_i^\star\big\|^2\boldsymbol{v}_i^{\star\otimes 2}\right\|_F^2,
\end{aligned}$$

结合 (31)，可知存在某个 j，使得

$$\|\sqrt{\beta_j}\boldsymbol{v}_j-\sqrt{\beta_i^\star}\boldsymbol{v}_i^\star\|_2^2\leq\frac{1}{r}.$$

不失一般性，假设

$$\|\sqrt{\beta_j}\boldsymbol{v}_j-\sqrt{\beta_j^\star}\boldsymbol{v}_j^\star\|_2^2\leq\frac{1}{r},\qquad\text{for all } 1\leq j\leq r.\tag{34}$$

Then

$$\begin{aligned}
&\mathbb{E}\big[\sum_{j=1}^{r}a_j\mathrm{relu}(\boldsymbol{w}_j^\top\boldsymbol{x})-\sum_{j=1}^{r}a_j^\star\mathrm{relu}(\boldsymbol{w}_j^{\star\top}\boldsymbol{x})\big]^2\\
&\geq\alpha_k\left\|\sum_{j=1}^{r}\beta_j\boldsymbol{v}_j\boldsymbol{v}_j^\top-\sum_{j=1}^{r}\beta_j^\star\boldsymbol{v}_j^\star\boldsymbol{v}_j^{\star\top}\right\|_F^2\\
&\geq\alpha_k\sum_{j=1}^{r}\big\|\beta_j\boldsymbol{v}_j\boldsymbol{v}_j^\top-\beta_j^\star\boldsymbol{v}_j^\star\boldsymbol{v}_j^{\star\top}\big\|_F^2\\
&\quad-\frac{\alpha_k}{2r}\left(\sum_{j=1}^{r}\big\|\beta_j\boldsymbol{v}_j\boldsymbol{v}_j^\top-\beta_j^\star\boldsymbol{v}_j^\star\boldsymbol{v}_j^{\star\top}\big\|_F\right)^2\\
&\geq\frac{\alpha_k}{2}\sum_{j=1}^{r}\big\|\beta_j\boldsymbol{v}_j\boldsymbol{v}_j^\top-\beta_j^\star\boldsymbol{v}_j^\star\boldsymbol{v}_j^{\star\top}\big\|_F^2.
\end{aligned}$$

这里，第一行由 (32) 得出；第二行由以下断言成立

$$\begin{aligned}
&\big|\langle\beta_{j_1}\boldsymbol{v}_{j_1}\boldsymbol{v}_{j_1}^\top-\beta_{j_1}^\star\boldsymbol{v}_{j_1}^\star\boldsymbol{v}_{j_1}^{\star\top},\beta_{j_2}\boldsymbol{v}_{j_2}\boldsymbol{v}_{j_2}^\top-\beta_{j_2}^\star\boldsymbol{v}_{j_2}^\star\boldsymbol{v}_{j_2}^{\star\top}\rangle\big|\\
&\leq\frac{1}{2r}\|\beta_{j_1}\boldsymbol{v}_{j_1}\boldsymbol{v}_{j_1}^\top-\beta_{j_1}^\star\boldsymbol{v}_{j_1}^\star\boldsymbol{v}_{j_1}^{\star\top}\|_2\|\beta_{j_2}\boldsymbol{v}_{j_2}\boldsymbol{v}_{j_2}^\top-\beta_{j_2}^\star\boldsymbol{v}_{j_2}^\star\boldsymbol{v}_{j_2}^{\star\top}\|_2
\end{aligned}$$

因为对于 $\boldsymbol{\delta}_j\coloneqq\sqrt{\beta_j}\boldsymbol{v}_j-\sqrt{\beta_j^\star}\boldsymbol{v}_j^\star$,

$$\beta_j\boldsymbol{v}_j\boldsymbol{v}_j^\top-\beta_j^\star\boldsymbol{v}_j^\star\boldsymbol{v}_j^{\star\top}=\boldsymbol{\delta}_j\boldsymbol{\delta}_j^\top+\sqrt{\beta_j^\star}\boldsymbol{\delta}_j\boldsymbol{v}_j^{\star\top}+\sqrt{\beta_j^\star}\boldsymbol{v}_j^\star\boldsymbol{\delta}_j^{\star\top}.$$

注意到这一点，结论便显而易见。

$$\big\|\beta_j\boldsymbol{v}_j\boldsymbol{v}_j^\top-\beta_j^\star\boldsymbol{v}_j^\star\boldsymbol{v}_j^{\star\top}\big\|_F\geq\|\boldsymbol{w}_j-\boldsymbol{w}_j^\star\|_2.$$

最后，我们分析包含 b_j、$b_j^\star\neq 0$ 的一般情形，其论证与上述类似。为简洁起见，这里只说明不同之处。根据引理4，对于任意常数 $k\geq 0$，存在某个常数 $\alpha_k>0$ 和某个函数 $f_k\colon\mathbb{R}\to\mathbb{R}$，使得

$$\begin{aligned}
&\mathbb{E}\left[\sum_{j=1}^{r}a_j\mathrm{relu}(\boldsymbol{w}_j^\top\boldsymbol{x}+b_j)-\sum_{j=1}^{r}a_j^\star\mathrm{relu}(\boldsymbol{w}_j^{\star\top}\boldsymbol{x}+b_j^\star)\right]^2\\
&\geq\sum_{k\geq\frac{12}{\omega}}^{\infty}\Big\|\sum_{j=1}^{r}a_jf_k(\frac{b_j}{\|\boldsymbol{w}_j\|_2})\|\boldsymbol{w}_j\|_2\big(\frac{\boldsymbol{w}_j}{\|\boldsymbol{w}_j\|_2}\big)^{\otimes k}\\
&\qquad-\sum_{j=1}^{r}a_j^\star f_k(\frac{b_j^\star}{\|\boldsymbol{w}_j^\star\|_2})\|\boldsymbol{w}_j^\star\|_2\big(\frac{\boldsymbol{w}_j^\star}{\|\boldsymbol{w}_j^\star\|_2}\big)^{\otimes k}\Big\|_F^2\\
&\gtrsim\sum_{j=1}^{r}\sum_{k\geq\frac{12}{\omega}}^{\infty}\left\|a_jf_k(\frac{b_j}{\|\boldsymbol{w}_j\|_2})\boldsymbol{w}_j-a_j^\star f_k(\frac{b_j^\star}{\|\boldsymbol{w}_j^\star\|_2})\boldsymbol{w}_j^\star\right\|_F^2\\
&\gtrsim\sum_{j=1}^{r}\inf_{R_l(\boldsymbol{x})}\mathbb{E}\Big[a_j\mathrm{relu}(\boldsymbol{w}_j^\top\boldsymbol{x}+b_j)\\
&\qquad-a_j^\star\mathrm{relu}(\boldsymbol{w}_j^{\star\top}\boldsymbol{x}+b_j^\star)-R_l(\boldsymbol{x})\Big]^2\\
&\gtrsim\sum_{j=1}^{r}\left(\|\boldsymbol{w}_j-\boldsymbol{w}_j^\star\|_2^2+|b_j-b_j^\star|^2\right).
\end{aligned}\tag{35}$$

这里，$l=\big[\frac{12}{\omega}\big]$，第二个不等式的成立方式与上述分析类似。于是，一般性结论便显而易见。

附录 C 引理 2 的证明（上界）

为简便起见，令

$$\begin{aligned}
z_i&\coloneqq\sum_{j=1}^{r}a_j\mathrm{relu}(\boldsymbol{w}_j^\top\boldsymbol{x}_i+b_j)-\sum_{j=1}^{r}a_j^\star\mathrm{relu}(\boldsymbol{w}_j^{\star\top}\boldsymbol{x}_i+b_j^\star),\\
\widehat{z}_i&\coloneqq\sum_{j=1}^{r}a_j\mathrm{relu}(\boldsymbol{w}_j^\top\boldsymbol{x}_i+b_j)-\sum_{j=1}^{r}\widehat{a}_j\mathrm{relu}(\widehat{\boldsymbol{w}}_j^\top\boldsymbol{x}_i+\widehat{b}_j).
\end{aligned}$$

回顾 $\left(\widehat{\boldsymbol{W}},\widehat{\boldsymbol{a}},\widehat{\boldsymbol{b}}\right)$ 关于问题 (3) 的最优性。根据三角不等式，可得

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}z_i^2}\leq\sqrt{\frac{1}{n}\sum_{i=1}^{n}\widehat{z}_i^2}+2\sigma.\tag{36}$$

我们可以对右边的第一项进行界定，得到

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\widehat{z}_i^2\\
&=\frac{1}{n}\sum_{i=1}^{n}\left[\sum_{j=1}^{r}a_j\left(\mathrm{relu}(\boldsymbol{w}_j^\top\boldsymbol{x}_i+b_j)-\mathrm{relu}(\widehat{\boldsymbol{w}}_j^\top\boldsymbol{x}_i+\widehat{b}_j)\right)\right]^2\\
&\le\frac{1}{n}\sum_{i=1}^{n}\left[\sum_{j=1}^{r}\left|(\boldsymbol{w}_j-\widehat{\boldsymbol{w}}_j)^\top\boldsymbol{x}_i\right|\right]^2\\
&\le\frac{r}{n}\sum_{i=1}^{n}\sum_{j=1}^{r}\left|(\boldsymbol{w}_j-\widehat{\boldsymbol{w}}_j)^\top\boldsymbol{x}_i\right|^2,
\end{aligned}$$

其中，第二行根据 ReLU函数的压缩性质成立，最后一行则由算术-几何平均不等式得出。引理5进一步给出，对于某个常数 $C_7>0$,

$$\begin{aligned}
\sum_{j=1}^{r}\frac{1}{n}\sum_{i=1}^{n}\left|(\boldsymbol{w}_j-\widehat{\boldsymbol{w}}_j)^\top\boldsymbol{x}_i\right|^2\le&C_7\sum_{j=1}^{r}\|\boldsymbol{w}_j-\widehat{\boldsymbol{w}}_j\|_2^2\\
&+C_7\frac{\log^3\frac{p}{n\delta}}{n}\sum_{j=1}^{r}\|\boldsymbol{w}_j-\widehat{\boldsymbol{w}}_j\|_1^2
\end{aligned}$$

以至少 $1-\delta$ 的概率成立。此外，

$$\begin{aligned}
\sum_{j=1}^{r}\|\boldsymbol{w}_j-\widehat{\boldsymbol{w}}_j\|_1^2&\le\left\|\boldsymbol{W}-\widehat{\boldsymbol{W}}\right\|_1^2\\
&\le\left(\|\boldsymbol{W}^\star\|_1-\|\widehat{\boldsymbol{W}}\|_1\right)^2\le D_1^2,
\end{aligned}$$

and

$$\begin{aligned}
\sum_{j=1}^{r}\|\boldsymbol{w}_j-\widehat{\boldsymbol{w}}_j\|_2^2=\left\|\boldsymbol{W}-\widehat{\boldsymbol{W}}\right\|_1^2&\le\left\|\boldsymbol{W}-\widehat{\boldsymbol{W}}\right\|_1\left\|\boldsymbol{W}-\widehat{\boldsymbol{W}}\right\|_\infty\\
&\le\frac{\left(\|\boldsymbol{W}^\star\|_1-\|\widehat{\boldsymbol{W}}\|_1\right)\left(\|\boldsymbol{W}^\star\|_1-\|\widehat{\boldsymbol{W}}^\star\|_1\right)}{S/2}\\
&\le\frac{4}{S}D_1^2.
\end{aligned}$$

这里，$\widehat{\boldsymbol{W}^\star}$ 表示 $\boldsymbol{W}$ 在 $\boldsymbol{W}^\star$的支撑集上的元素，$\widehat{\ }$，并且我们利用了以下事实：$\|\boldsymbol{W}\|_1\le\|\boldsymbol{W}^\star\|_1$ 以及

$$\left\|\boldsymbol{W}-\widehat{\boldsymbol{W}}\right\|_\infty\le\frac{\|\widehat{\boldsymbol{W}}^\star-\widehat{\boldsymbol{W}}\|_1}{S-s}\le\frac{\|\boldsymbol{W}^\star\|_1-\|\widehat{\boldsymbol{W}}^\star\|_1}{S/2}.$$

综上即可得到所需结果。

Appendix D
Technical lemmas

引理 3. 对任意满足 $\|\boldsymbol{W}\|_0+\|\boldsymbol{b}\|_0+\|\boldsymbol{W}^\star\|_0+\|\boldsymbol{b}^\star\|_0\le S$ 的 $(\boldsymbol{W},\boldsymbol{a},\boldsymbol{b})$，假设 $\|\boldsymbol{W}\|_1+\|\boldsymbol{b}\|_1\le\|\boldsymbol{W}^\star\|_1+\|\boldsymbol{b}^\star\|_1$ 且 $\|\boldsymbol{w}_j^\star\|_2^2+|b_j^\star|^2\le1$。则有

$$D_1\le2\sqrt{S}D_2,\tag{37}$$

其中，D_1、D_2 在（17）中定义。

证明。为简单起见，假设 \(\)

$$D_2^2=\sum_{j\in\mathcal{J}}\left(\|\boldsymbol{w}_j-\boldsymbol{w}_j^\star\|_2^2+|b_j-b_j^\star|^2\right)+\sum_{j\notin\mathcal{J}}\left(\|\boldsymbol{w}_j^\star\|_2^2+|b_j^\star|^2\right).$$

这里，$j\in\mathcal{J}$ 表示 $a_j=a_j^\star$ 和

$$\|\boldsymbol{w}_j-\boldsymbol{w}_j^\star\|_2^2+|b_j-b_j^\star|^2\le\|\boldsymbol{w}_j^\star\|_2^2+|b_j^\star|^2.$$

然后，根据算术-几何平均不等式，有

$$\begin{aligned}
\sqrt{S}D_2\ge&\sum_{j\in\mathcal{J}}\left(\|\boldsymbol{w}_j-\boldsymbol{w}_j^\star\|_1+|b_j-b_j^\star|\right)\\
&+\sum_{j\notin\mathcal{J}}\left(\|\boldsymbol{w}_j^\star\|_1+|b_j^\star|\right)\\
\ge&\sum_{j\in\mathcal{J}}\left(\|\boldsymbol{w}_j^\star\|_1-\|\boldsymbol{w}_j\|_1+|b_j^\star|-|b_j|\right)+\|\boldsymbol{W}^\star\|_1\\
&+\|\boldsymbol{b}^\star\|_1-\sum_{j\in\mathcal{J}}\left(\|\boldsymbol{w}_j^\star\|_1+|b_j^\star|\right)\\
\ge&\sum_{j\notin\mathcal{J}}\left(\|\boldsymbol{w}_j\|_1+|b_j|\right),
\end{aligned}$$

这意味着，

$$\begin{aligned}
2\sqrt{S}D_2\ge&\sum_{j\in\mathcal{J}}\left(\|\boldsymbol{w}_j-\boldsymbol{w}_j^\star\|_1+|b_j-b_j^\star|\right)\\
&+\sum_{j\notin\mathcal{J}}\left(\|\boldsymbol{w}_j^\star\|_1+|b_j^\star|+\|\boldsymbol{w}_j\|_1+|b_j|\right).
\end{aligned}$$

由此得证。 □

引理4 (定理2.1 [27]). 对于任意常数 $k\ge0$，存在某个普适函数 $f_k:\mathbb{R}\to\mathbb{R}$ 使得

$$\begin{aligned}
&\mathbb{E}\left[\sum_{j=1}^{r}a_j\mathrm{relu}(\boldsymbol{w}_j^\top\boldsymbol{x}+b_j)-\sum_{j=1}^{r}a_j^\star\mathrm{relu}(\boldsymbol{w}_j^{\star\top}\boldsymbol{x}+b_j^\star)\right]^2\\
&=\sum_{k=0}^{\infty}\left\|\sum_{j=1}^{r}a_jf_k\left(\frac{b_j}{\|\boldsymbol{w}_j\|_2}\right)\|\boldsymbol{w}_j\|_2\left(\frac{\boldsymbol{w}_j}{\|\boldsymbol{w}_j\|_2}\right)^{\otimes k}\right.\\
&\qquad\left.-\sum_{j=1}^{r}a_j^\star f_k\left(\frac{b_j^\star}{\|\boldsymbol{w}_j^\star\|_2}\right)\|\boldsymbol{w}_j^\star\|_2\left(\frac{\boldsymbol{w}_j^\star}{\|\boldsymbol{w}_j^\star\|_2}\right)^{\otimes k}\right\|_F^2,
\end{aligned}\tag{38}$$

with

$$\alpha_k:=f_{2k}(0)>0,\qquad\textit{for all } k>0.\tag{39}$$

此外，我们有

$$\begin{aligned}
&\inf_{R_l}\mathbb{E}\left[a\mathrm{relu}(\boldsymbol{w}^\top\boldsymbol{x}+b)-\sum_{j=1}^{r}a^\star\mathrm{relu}(\boldsymbol{w}^{\star\top}\boldsymbol{x}+b^\star)-R_l(\boldsymbol{x})\right]^2\\
&=\sum_{k>l}^{\infty}\left\|af_k\left(\frac{b}{\|\boldsymbol{w}\|_2}\right)\|\boldsymbol{w}\|_2\left(\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|_2}\right)^{\otimes k}\right.\\
&\qquad\left.-a^\star f_k\left(\frac{b^\star}{\|\boldsymbol{w}^\star\|_2}\right)\|\boldsymbol{w}^\star\|_2\left(\frac{\boldsymbol{w}^\star}{\|\boldsymbol{w}^\star\|_2}\right)^{\otimes k}\right\|_F^2,
\end{aligned}\tag{40}$$

其中，R_l 表示一个关于 $\boldsymbol{x}$ 的次数小于 l的多项式函数。

引理5. 存在某个普适常数 $c>0$，使得对所有 $\boldsymbol{w}\in\mathbb{R}^p$,

$$\frac{1}{n}\sum_{i=1}^{n}\left|\boldsymbol{w}^\top\boldsymbol{x}_i\right|^2\le c\|\boldsymbol{w}\|_2^2+c\frac{\log^3\frac{p}{n\delta}}{n}\|\boldsymbol{w}\|_1^2,\tag{41}$$

以至少 $1-\delta$ 的概率成立。

证明. 在继续之前，我们先介绍一些关于受限等距性质（RIP）的有用技巧。令 $\boldsymbol{X} := \frac{1}{\sqrt{n}}[\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n]$ $c_0 > 0$ $n \geq$。$c_0\left(s \quad \frac{p}{s} + \quad \frac{1}{\delta}\right)$ $1-\delta$
对于某个常数，若 log log，则以至少 的概率，

$$\left\|\boldsymbol{X}^\top \boldsymbol{w}\right\|_2^2 \leq 2\|\boldsymbol{w}\|_2^2 \tag{42}$$

对所有满足 $\|\boldsymbol{w}\|_0 \leq s$ 的 $\boldsymbol{w}$ 均成立。

我们将 $\boldsymbol{w}$ 的各个元素分成若干组 $\mathcal{S}_1 \cup \mathcal{S}_2 \cup \ldots \cup \mathcal{S}_L$，每组大小相同，为 s（$\mathcal{S}_L$ 除外），使得对于任意 $j < k$，$\mathcal{S}_j$ 中的元素均不小于 $\mathcal{S}_k$。然后，根据 (42)，可得

$$\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{w}^\top \boldsymbol{x}_i)^2 = \boldsymbol{w}^\top \boldsymbol{X}\boldsymbol{X}^\top \boldsymbol{w} = \sum_{j,k} \boldsymbol{w}_{\mathcal{S}_j}^\top \boldsymbol{X}\boldsymbol{X}^\top \boldsymbol{w}_{\mathcal{S}_k}$$

$$\leq 2\sum_{j,k}\|\boldsymbol{w}_{\mathcal{S}_j}\|_2\|\boldsymbol{w}_{\mathcal{S}_k}\|_2 = 2\Big(\sum_{l=1}^{L}\|\boldsymbol{w}_{\mathcal{S}_l}\|_2\Big)^2.$$

此外，$\boldsymbol{w}_{\mathcal{S}_l}$ 的顺序给出，对于 $l > 1$,

$$\|\boldsymbol{w}_{\mathcal{S}_l}\|_2 \leq \sqrt{s}\|\boldsymbol{w}_{\mathcal{S}_l}\|_\infty \leq \frac{1}{(l-1)\sqrt{s}}\|\boldsymbol{w}\|_1,$$

由此可得

$$\Big(\sum_{l=1}^{L}\|\boldsymbol{w}_{\mathcal{S}_l}\|_2\Big)^2 \leq 2\|\boldsymbol{w}_{\mathcal{S}_1}\|_2^2 + 2\Big(\sum_{l=2}^{L}\frac{1}{(l-1)\sqrt{s}}\|\boldsymbol{w}\|_1\Big)^2$$

$$\leq 2\|\boldsymbol{w}\|_2^2 + \frac{2\log^2 L}{s}\|\boldsymbol{w}\|_1^2.$$

通过结合上述不等式，我们完成了证明。

□

附录E
进一步实验细节

第三节中使用的超参数总结于表 IX。

我们简要说明变量选择过程。首先，我们获得一个变量重要性向量。对于“LASSO”和“OMP”，我们使用估计系数的绝对值作为变量重要性；对于“神经网络”、“GLASSO”和“GSR”，我们对神经网络输入层中的权重矩阵按行应用 ℓ_2-范数来获得重要性；对于“随机森林”、“梯度提升”和“LNET”，我们使用这些方法生成的重要性。获得重要性向量后，我们通过改变截断阈值，得到合成数据集的受试者工作特征（ROC）曲线，并计算 AUC 分数。至于变量选择，我们对合成数据集的重要性向量应用由两个混合成分构成的 GMM。重要性较高的簇中的变量被视为显著变量。随后，我们据此计算正确或错误选择的变量数量。对于 BGSBoy 和 UJIIndoorLoc 数据集，我们分别选择重要性最高的三个和十个变量。

参考文献

[1] R. Tibshirani，《通过套索进行回归收缩与变量选择》，《皇家统计学会会刊B辑》，第58卷，第1期，第267–288页，1996年。[2] S. A. Van de Geer等，《高维广义线性模型与LASSO》，《统计年鉴》，第36卷，第2期，第614–645页，2008年。

[3] R. Tibshirani，“用于 Cox 模型变量选择的套索方法”，《医学统计学》，第 16 卷，第 4 期，第 385–395 页，1997年。[4] I. Goodfellow、Y. Bengio、A. Courville 和 Y. Bengio，《深度学习》。麻省理工学院出版社，剑桥，2016年，第 1 卷，第 2 期。[5] J. Wang、Q. Cai、Q. Chang 和 J. M. Zurada，“通过组套索正则化对稀疏前馈神经网络进行收敛性分析”，《信息科学》，第 381 卷，第 250–269 页，2017年。[6] W. Yang、G. Wang、J. Ding 和 Y. Yang，“从稀疏线性近似角度理解神经网络压缩的理论基础”，arXiv 预印本 arXiv:2206.05604，2022年。[7] E. Diao、G. Wang、J. Zhan、Y. Yang、J. Ding 和 V. Tarokh，“从稀疏性视角对深度神经网络进行剪枝”，发表于 ICLR 会议论文集，2023年。[8] S. Scardapane、D. Comminiello、A. Hussain 和 A. Uncini，“深度神经网络的组稀疏正则化”，《神经计算》，第 241 卷，第 81–89 页，2017年。[9] I. Lemhadri、F. Ruan 和 R. Tibshirani，“Lassonet：具有特征稀疏性的神经网络”，发表于人工智能与统计学国际会议论文集。机器学习研究会议论文集，2021年，第 10–18 页。[10] M. Abadi、P. Barham、J. Chen、Z. Chen、A. Davis、J. Dean、M. Devin、S. Ghemawat、G. Irving 和 M. Isard，“TensorFlow：用于大规模机器学习的系统”，发表于 USENIX 会议论文集，2016年，第 265–283 页。[11] A. Paszke、S. Gross、F. Massa、A. Lerer、J. Bradbury、G. Chanan、T. Killeen、Z. Lin、N. Gimelshein、L. Antiga 等，“PyTorch：一种命令式风格的高性能深度学习库”，NeurIPS 会议论文集，第 32 卷，2019年。[12] J. Ding、V. Tarokh 和 Y. Yang，“模型选择技术：综述”，《IEEE信号处理杂志》，第 35 卷，第 6 期，第 16–34 页，2018年。[13] J. Zhang、Y. Yang 和 J. Ding，“模型选择的信息准则”，《威利跨学科评论：计算统计学》，第 e1607 页，2023年。[14] D. L. Donoho 和 X. Huo，“不确定性原理与理想原子分解”，《IEEE信息论汇刊》，第 47 卷，2001年，第 2862–2845 页，第 7 期。[15] E. Candes 和 T. Tao，“线性规划解码”，《IEEE信息论汇刊》，第 51 卷，第 12 期，第 4203–4215 页，2005年。[16] P. Zhao 和 B. Yu，“关于 LASSO 的模型选择一致性”，《机器学习研究杂志》，第 7 卷，第 11 期，第 2541–2563 页，2006年。[17] P. J. Bickel、Y. Ritov、A. B. Tsybakov 等，“套索与 Dantzig选择器的同时分析”，《统计年鉴》，第 37 卷，2009年，第 1732–1705 页，第 4 期。[18] A. N. Kolmogorov，“通过一个变量的连续函数的叠加与加法表示多变量连续函数”，载于《科学院通报》，第 114 卷，第 5 期。俄罗斯科学院，1957年，第 953–956 页。[19] G. Cybenko，“通过 S形函数的叠加进行逼近”，《数学、控制与信号系统》，第 2 卷，第 183–192 页，1989年。[20] A. R. Barron，“S形函数叠加的通用逼近界”，《IEEE信息论汇刊》，第 39 卷，第 3 期，第 930–945 页，1993年。[21] ——，“人工神经网络的逼近界与估计界”，《机器学习》，第 14 卷，第 1 期，第 115–133 页，1994年。[22] G. Li、Y. Gu 和 J. Ding，“双层神经网络中的 L1正则化”，《IEEE信号处理快报》，第 29 卷，第 135–139 页，2021年。[23] B. Neyshabur、R. Tomioka 和 N. Srebro，“神经网络中基于范数的容量控制”，学习理论会议论文集，第 1376–1401 页，2015年。[24] N. Golowich、A. Rakhlin 和 O. Shamir，“与规模无关的神经网络样本复杂度”，发表于学习理论会议论文集。机器学习研究会议论文集，2018年，第 297–299 页。[25] G. Li 和 J. Ding，“变差约束深度神经网络的收敛速率”，《IEEE信号处理汇刊》，第 71 卷，第 1 期，第 631–640 页，2023年。[26] M. Janzamin、H. Sedghi 和 A. Anandkumar，“克服非凸性的困境：利用张量方法保证神经网络训练”，arXiv 预印本 arXiv:1506.08473，2015年。[27] R. Ge、J. D. Lee 和 T. Ma，“通过景观设计学习单隐层神经网络”，NeurIPS 会议论文集，2018年。[28] M. Mondelli 和 A. Montanari，“双层神经网络学习与张量分解之间的联系”，发表于人工智能与统计学国际会议论文集。机器学习研究会议论文集，2019年，第 1051–1060 页。[29] A. R. Barron 和 J. M. Klusowski，“利用总路径变差的深度网络的复杂度、统计风险与度量熵”，《统计》，第 1050 卷，第 6 页，2019年。[30] J. Schmidt-Hieber，“使用 ReLU激活函数的深度神经网络进行非参数回归”，《统计年鉴》，第 48 卷，第 4 期，第 1875–1897 页，2020年。[31] B. Bauer 和 M. Kohler，“关于深度学习作为非参数回归中维数灾难补救方法的研究”，《统计年鉴》，第 47 卷，第 4 期，第 2261–2285 页，2019年。

表IX 我们实验中使用的超参数。

数据集	线性	神经网络生成	弗里德曼	BGSBoy	UJIINdoorLoc
训练轮数	{100,200,500}				
批量大小	32	32	32	8	128
优化器	ADAM				
学习率	{0.001, 0.005, 0.01}				
调度器	N/A				

[32] M. Yuan 和 Y. Lin，“具有分组变量的回归中的模型选择与估计”，J. Roy. Statist. Soc. Ser. B，第 68 卷，2006年，第 67–49 页，第 1 期。[33] V. Dinh 和 L. S. T. Ho，“解析深度神经网络的一致特征选择”，NeurIPS 会议论文集，2020年。[34] L. Bottou，“利用随机梯度下降进行大规模机器学习”，发表于 COMPSTAT 会议论文集。Springer，2010年，第 177–186 页。[35] D. P. Kingma 和 J. Ba，“Adam：一种随机优化方法”，ICLR 会议论文集，2015年。[36] R. D. Tuddenham，“加利福尼亚男孩和女孩从出生到十八岁的身体生长”，Univ. Calif. Publ. Child Dev.，第 1 卷，第 183–364 页，1954年。[37] J. Torres-Sospedra、R. Montoliu、A. Martínez-Usó、J. P. Avariento、T. J. Arnau、M. Benedito-Bordonau 和 J. Huerta，“UJIIndoorLoc：一个新的用于基于 WLAN 指纹的室内定位问题的多建筑、多楼层数据库”，发表于 IPIN 会议论文集。IEEE，2014年，第 261–270 页。

Gen Li 目前是宾夕法尼亚大学沃顿商学院统计与数据科学系的博士后研究员。他于2021年获得清华大学电子工程博士学位，并获得清华大学电子工程和数学理学学士学位。

2016年毕业于清华大学。他的研究兴趣包括强化学习、高维统计学、机器学习、信号处理和数学优化。他曾获得清华大学优秀毕业生奖和优秀论文奖。

Ganghua Wang 于 2019 年获得中国北京北京大学理学学士学位。自 2019 年起，他一直是美国明尼苏达州双城明尼苏达大学统计学院的博士生。他的研究兴趣包括机器学习理论基础和机器学习安全。

Jie Ding 于 2017 年在剑桥的哈佛大学获得工程科学博士学位，2012 年在北京的清华大学获得理学学士学位。2018 年，他加入明尼苏达大学双城分校，现任统计学院助理教授，并受聘于电气与计算机工程系研究生院。他的研究方向涵盖机器学习、统计学、信号处理和信息论。