arXiv:2302.05851v1 [math.ST] 12 Feb 2023

适用于维度增长的非参数交互模型的深度神经网络

BY SOHOM BHATTACHARYA[1,a], JIANQING FAN[1,a] 和 DEBARGHYA MUKHERJEE [1,a]

[1]运筹学与金融工程系，普林斯顿大学，a

深度神经网络凭借其表示能力以及对低维结构的适应能力，取得了巨大成功。近期文献已证实，它们在估计结构化回归函数方面具有潜力。然而，大多数研究要求输入维度固定，因此忽略了维度对收敛速率的影响，也限制了其在高维大数据中的应用。本文通过分析一种 k^{th} 阶非参数交互模型，弥补了这一空白；我们同时研究维度增长情形（d随 n 增长，但增长速率较慢）和高维情形（$d \gtrsim n$）。在后一种情形下，为了获得最优收敛速率，需要满足稀疏性假设并采用相应的正则化。维度发散情形带来了一个新的挑战：计算均方误差时，估计加性分量之间的协方差项比方差项大一个数量级，若处理不当，便会损害统计性质。我们提出了一种关键的去偏技术来解决这一问题。我们证明，在某些标准假设下，去偏深度神经网络在（n，d）方面均达到极小极大最优速率。我们的证明方法关键依赖于一种新颖的去偏技术，该技术使均方误差计算中加性分量的协方差可以忽略不计。此外，我们还建立了相匹配的下界。

1. 引言。 技术的最新进展使统计学家能够收集大量解释变量的数据，以预测感兴趣的结果（Goodfellow 等人，2016；Fan 等人，2020）。这些结果与预测变量之间的关系通常高度非线性（例如，MNIST（LeCun，1998）、CIFAR（Krizhevsky 等人，2009）等图像数据），因此在实践中有必要研究多元非参数回归模型

$$Y_i = f(X_i) + \varepsilon_i, \quad i = 1, \ldots, n, \tag{1.1}$$

其中，X_i是解释变量，Y_i是响应变量，ε_i是未观测误差。这里的统计问题是在某些最小光滑度假设下恢复 f。Charles Stone 的经典结果（Stone，1982）表明，若 f 是一个 d元的(β, C) 光滑函数（其精确定义将在后文给出），则极小极大最优估计速率的阶为$n^{-\frac{2\beta}{2\beta+d}}$，这被称为“维数灾难”，也就是说，要准确估计回归函数，需要较大的样本量。特别地，当 β 固定时，很容易看出，当 $d \asymp \log n$时，无法获得极小极大一致估计量。

为了解决上述问题，需要在函数上施加低维结构 f (Kpotufe 和 Dasgupta，2012；Yang 和 Dunson，2016；Yang 和 Tokdar，2015），例如加性回归模型（Stone，1985）、投影寻踪回归模型（Fried- man 和 Stuetzle，1981）、高阶交互模型（Stone，1994）、广义高阶交互模型（Horowitz 和 Mammen，2007；Kohler 和 Krzy˙zak，2016）、单指标模型

关键词和短语：深度神经网络、高维统计学、非参数交互模型、极小极大速率、稀疏非参数分量。

以及多指标模型（Hardle 等人，1993）等；综述参见 Fan 和 Gijbels（1996）。这些简化方法背后的核心思想，是降低底层函数类的复杂度，并减轻维度对收敛速率的影响。例如，Stone（1985）证明，对于加性回归模型（当 $f(x) = \sum_{j=1}^{d} f_j(x_j)$ 时），若单变量分量具有 β-光滑性，则 f 关于平方 $L_2(P_X)$ 损失的极小极大最优估计速率为 $C_d n^{-\frac{2\beta}{2\beta+1}}$。Fan 等人（1998）进一步展示了这种加性结构的自适应性，而 Horowitz 和 Mammen（2007）则进一步研究了一类具有未知连接函数的非参数回归模型。由于每个分量函数 f_j 都是单变量函数，其估计速率为 $n^{-\frac{2\beta}{2\beta+1}}$，维度的影响仅体现为乘法常数 C_d，而不会影响 n 的幂次。随后，Stone（1994）将这一结果推广到高阶交互模型（即当 $f(x) = \sum_{J \in S} f_J(x_J)$ 时，其中 S 是 $\{1, \ldots, d\}$ 的子集的集合，且对所有 $J \in S$ 均有 $|J| \leq d^\star$），并证明极小极大最优估计速率为 $C_d n^{-\frac{2\beta}{2\beta+d^\star}}$；与前面一样，$d$ 的影响体现为乘法常数。上述结果同样都假设 d 有限，这不适用于许多现代数据科学应用。

在实践中，仅获得极小极大最优速率是不够的，还需要高效且易于计算的估计量。核、样条和小波估计量等方法（综述见 Fan 和 Gijbels (1996) 的第 2 节）、回代算法（Buja 等人，1989），以及提升法（Freund 等人，1996）、随机森林（Breiman，2001）等现成的非参数方法，都得到了广泛应用。近年来，神经网络已成为分析协变量与响应变量之间非线性关系的重要工具。深度（或多层）神经网络一直是机器学习领域令人难以置信的进步的基础，并促成了强化学习、机器人学、计算机视觉、自然语言处理以及其他统计预测任务的巨大成功（Goodfellow 等人，2016）。借助大量数字化数据的可获得性，深度学习取得了大量实证成功。然而，这些成功大多纯粹依赖人为设计（即在调节大量超参数后实现），缺乏理论保证。因此，研究人员自然希望理解这些有用且实用的估计量的统计性质。在 Kohler 和 Krzy˙zak (2005) 中，作者证明，对于标准非参数模型以及高阶交互模型，基于深度神经网络（以下简称 DNN）的估计量在对数因子意义下达到极小极大速率最优。随后，在 Kohler 和 Krzy˙zak (2016)、Bauer 和 Kohler (2019)、Schmidt-Hieber (2020)、Kohler 和 Langer (2021) 以及 Fan et al. (2022) 的一系列工作中，这些结果被推广到更一般的函数类（即广义层次交互模型，见 Bauer 和 Kohler (2019)）。

以往关于理解基于深度神经网络的估计量收敛速率的研究，假设协变量 d 的底层维度是固定的，因此忽略了 d 对估计量收敛速率的影响。然而，在大数据时代，许多实际问题中的底层协变量维度都相当高，往往高于底层样本量 n。例如，在一个简单的图像分类问题中，一幅 28×28 图像位于$\mathbb{R}^{784}$。类似地，在全基因组关联研究中，单核苷酸多态性（SNP）的数量可能远高于个体数量（Novembre etal.，2008年）。在该论文中，作者分析了一个由 1387 名个体和$\sim 2 \times 10^5$ 个SNP组成的数据集。在此类例子中，通常有必要进行更精细的分析，以量化对 d 的最优依赖关系。

本文旨在通过分析维度增长情形（当 d 随 n 增长，但 $d = o(n)$）和高维情形（当$d \gtrsim n$）下的非参数回归模型，弥合这一差距。这两种情形的区别在于，是否需要稀疏性和正则化才能实现相合估计，这与线性

维度增长模型和高维稀疏线性模型类似。我们分析以下 k阶交互模型（Stone，1994）：

$$Y_i = \sum_{l=1}^{k} \sum_{(j_1<\cdots<j_l)\in[d]^l} f_{j_1,j_2,\ldots,j_l}(X_{i,\{j_1,j_2,\ldots,j_l\}}) + \epsilon_i\,, \tag{1.2}$$

其中 $[d]^l$ 是所有长度为 l 的 $\{1,\ldots,d\}$有序子序列的集合。如前所述， $k = o(\log n)$ 是获得一致估计量的必要条件。因此，我们将采用有限的 k。此类模型包括广为研究的加性模型（Stone，1985），即 $f(x) = \sum_{j=1}^{d} f_j(x_j)$ 以及交互模型（Stone，1994）。当 $d \gtrsim n$时，即使在高维线性模型中，也有必要施加稀疏性假设。因此，我们考虑如下稀疏 k阶交互模型：

$$Y_i = \sum_{l=1}^{k} \sum_{(j_1,j_2,\ldots,j_l)\in S_l} f_{j_1,j_2,\ldots,j_l}(X_{i,\{j_1,j_2,\ldots,j_l\}}) + \epsilon_i\,, \tag{1.3}$$

其中 $S_l \subset [d]^l$ 满足 $|S_l| \ll d^l$。当每个单变量函数都被限制为线性函数时，模型（1.3）就简化为稀疏参数交互模型。当 $k = 1$时，模型（1.3）成为稀疏加性模型，该模型已在文献中得到广泛研究（Lin andZhang，2006；Koltchinskii and Yuan，2010年；Ravikumar 等，2009；Tan and Zhang，2019；Yuan and Zhou，2016），其研究建立在惩罚线性回归的最新进展之上。例如，Tan andZhang(2019）提出对每个分量 f_j 同时施加惩罚，惩罚项分别为其经验范数和函数半范数，以估计回归函数。

当 $d \ll n$时，我们通过最小化平方误差损失来估计（1.2）的均值函数（详情见第2节）。然而，当 $d \gtrsim n$时，我们需要显式地施加惩罚项以确保稀疏性。在以往关于稀疏加性模型的研究中，研究者通常施加两种惩罚项：i）一种用于确保稀疏性（例如，对分量函数的 $L_2(\mathbb{P}_n)$ 范数求和）；ii）另一种用于控制分量函数的复杂度（例如，当分量函数属于再生核希尔伯特空间时，针对再生核希尔伯特空间范数施加惩罚）。实际上，实现这种双重惩罚估计量通常在计算上颇具挑战。为克服这一困难，我们提出一种两步式基于硬阈值的估计量，其动机来源于SURE独立筛选的开创性工作（Fan和Lv，2008年；Fan和Song，2010年）以及高维线性回归模型中模型选择后的最小二乘估计（Belloni和Chernozhukov，2013年）。其思想可简述如下：首先，仅通过对分量函数的经验 $L_2(\mathbb{P}_n)$ 范数之和施加惩罚，得到均值函数的一个初始估计量（记为$\hat{}$ f^{init}）（这可以通过对神经网络的最后一层施加惩罚来实现）。类似地，这也可以被视为组套索惩罚的一种形式，其中对应于每个分量的神经网络构成一个变量组，我们对各组范数之和施加惩罚（此处我们取 $L_2(\mathbb{P}_n)$) 范数），以确保分量级稀疏性。接下来，我们进一步对$\hat{}$ f^{init} 中的非零分量进行硬阈值处理，以剔除较小的非零分量并估计活跃集。最后，仅在上一步选出的分量上，通过最小化平方误差损失执行经验风险最小化，从而得到最终估计量。我们证明，在一些相当标准的假设下，该估计量是极小极大速率最优的。现将我们的贡献总结如下。

贡献： 我们的主要贡献是对 k 阶交互模型（也称为非参数方差分析模型）在维度增长和高维设置下进行了严谨的理论分析。为实现这一目标，我们还证明了若干可能具有独立研究价值的结果。我们总结如下关键贡献：

• 我们使用基于神经网络的估计量分析非参数 k 阶交互模型，既考虑 $d=\dim(X)$ 随 n 和 $d=o(n)$ 增长的情形，也考虑 $d \gtrsim n$ 并采用上述正则化的情形。我们证明，基于神经网络的估计量在相差一个多对数因子以内达到极小极大速率最优。

• 我们提出了一种新颖的去偏技术，使加性分量之间的协方差可以忽略不计，从而降低统计误差。

• 我们通过一种2.7) 新颖的去偏技术证明了利用深度神经网络逼近光滑函数的结果（定理，这意味着即使在约束条件下（即多元函数的边际函数为0），也能以相同的速率使用神经网络逼近光滑函数。

• 我们建立了估计高维 k 阶交互模型的极小极大下界；据我们所知，现有文献中尚无此结果。

本文其余部分安排如下：第2节分析 k 阶交互模型，其中 $d \ll n$。我们将整个分析分为三个小节：小节2.1界定逼近误差，小节2.2 分析统计误差，小节2.3建立极小极大下界。第3 节分析 $d \gtrsim n$时的情形。我们的分析大致分为两部分：小节3.1 分析固定设计，小节3.2 讨论随机设计的情形。此外，小节3.3建立稀疏非参数交互模型的极小极大下界。第4 节给出结论性评述。最后，第5 节包含定理2.7的证明。其余证明见附录。

2. 低发散维度下深度神经网络的分析。本节将分析基于深度神经网络的均值函数估计量，其中维度 d 随 X 增大，但 n $d \ll n$。我们考虑一个 k 阶交互非参数回归模型。对于输入输出对 $(X,\ Y)$，其中 $X \in \mathbb{R}^d$ 且 $Y \in \mathbb{R}$，k 阶交互模型定义如下：

$$Y_i = \sum_{\substack{J \subset [d] \\ |J| \le k}} f_{0,J}(X_{i,J}) + \epsilon_i \,. \tag{2.1}$$

这里，ϵ_i 被假设为与 X 独立的中心化误差，函数 $f_{0,\ J}$：$\mathbb{R}^{|J|} \mapsto \mathbb{R}$。为简化表述，以下我们仅讨论2 阶交互模型。将分析从 2 阶交互模型推广到一般的 k 阶交互模型纯属技术性工作，相关证明思路将在相应章节末尾概述。对于 $k=2$，方程（2.1）中的模型可简化为：

$$Y_i = \sum_{j=1}^{d} f_{0,j}(X_{i,j}) + \sum_{j<k} f_{0,jk}(X_{i,j}, X_{i,k}) + \epsilon_i \equiv f_0(X_i) + \varepsilon_i \,. \tag{2.2}$$

其中，$f'_{0,\ j}s$ 是单变量函数，而 $f_{0,\ jk}$'s 是二元函数。我们对 X 和 ε作出如下假设：

假设 2.1。X 的支集为 $[0,1]^d$，且具有密度函数 p，满足上确界$_{x\in[0,1]^d} p\ (x)\ =: p_{\max} < \infty$，其中 $p_{\max}$ 与 d无关。我们假设 ϵ 服从次高斯分布，其次高斯常数为 σ_ϵ^2。

假设 2.1 是非参数回归文献中的标准假设。注意，通过中心化和缩放，假设 X 具有紧支撑，等价于假设其支撑为 $[0,1]^d$。通过截断论证，我们的许多结果可以推广到无界情形；

X 为使本文简洁，我们省略这类论证。对密度作上界假设相当自然，因为这仅排除了某些退化分布（即某个区域的密度发散至无穷大的情形）。由于我们旨在分析基于神经网络的 f 最小二乘估计量，因此对误差项作次高斯性假设是自然的。文献中已有充分研究表明，对于重尾误差，非参数函数的最小二乘估计量并非极小极大最优（例如，参见Han 和 Wellner(2019)），此时应使用 Huber 损失函数的变体。我们认为，我们的大部分分析都可以推广到这种重尾误差情形，但这超出了本文的讨论范围。近期的文章（Fan 等人，2022）为固定维度设定下重尾误差时深度神经网络的估计误差分析提供了一些启示。

在我们的模型（2.2）中，f_0 的分量函数在没有进一步假设的情况下无法识别，原因有二：
1. 所有函数都只能在相差一个常数的意义下识别，即我们无法区分$(f_{0,i_1i_2}, f_{0,j_1j_2})$与$(f_{0,i_1i_2}+c, f_{0,j_1j_2}-c)$，其中常数为 c。2. 二元函数的一元边际无法识别。例如，考虑函数子集 $\mathcal{C}_1=\{f_{0,1}, f_{0,12}, f_{0,13}, \ldots, f_{0,1d}\}$。令 $g_k(x)=\int f_{0,1k}(x,y)\ dy$ 为 $f_{0,1k}$的边际函数。那么，我们无法区分 $\mathcal{C}_1$ 与经过修改的函数集合
$\mathcal{C}_2=\{f_{0,1}+\sum_{k\geq 2} g_k, f_{0,12}-g_2, \ldots, f_{0,1d}-g_d\}$。
因此，为了识别和估计这些非参数函数，我们需要施加某些结构条件，以避免函数发生平移。为此，我们作如下假设：

假设 2.2（可识别性与有界性）。我们作出以下条件假设，以确保可识别性和有界性：

1. （2.2）中的所有单变量函数的积分均为0，即 $\int_0^1 f_{0,\ j}(x)\ dx=0$ 对于 $1\leq j\leq d$。2. 所有二元函数的0 边际函数，即

$$\int_0^1 f_{0,ij}(x,y)\ dx=\int_0^1 f_{0,ij}(x,y)\ dy=0\,.$$

3. $\|f_0\|_\infty\leq B$ 对于某个 $B>0$，其中 f_0 定义于 （2.2）。

注意，上述假设意味着 $f_{0,\ ij}$ 的总积分也为 0。假设2.2 确保了（2.2）中所有相关函数的可识别性。接下来，我们对底层分量函数作出光滑度假设：

假设 2.3. 假设函数 $\{f_{0,j}\}$ 和 $\{f_{0,ij}\}$ 是(β,L)-*Holder* 光滑的，即这些函数可微 $\lfloor\beta\rfloor$ 次，且其 $\lfloor\beta\rfloor^{th}$ 阶导数是 *Holder* 连续的，指数为 $\beta-\lfloor\beta\rfloor$，常数为 L。

这种光滑性假设在非参数回归文献中很常见，参见（Tsy-bakov，2004年，第 1 章），因为这种光滑性假设控制了底层函数类的复杂度。另一种常见假设是，所有 $f_{0,i}$ 和 $f_{0,ij}$ 都属于再生核希尔伯特空间（RKHS），例如参见Raskutti 等人 (2012), Koltchinskii andYuan (2010年)及其中的参考文献。在这些假设下研究我们结果的类似结论将会很有意思。

注释 2.4. 假设 $\{f_{0,ij}\}$ 具有不同的光滑度 β_{ij}，等价于假设它们都具有最$_{i,\ j}\beta_{i,\ j}$ 光滑度（只要关注的是估计误差界）。

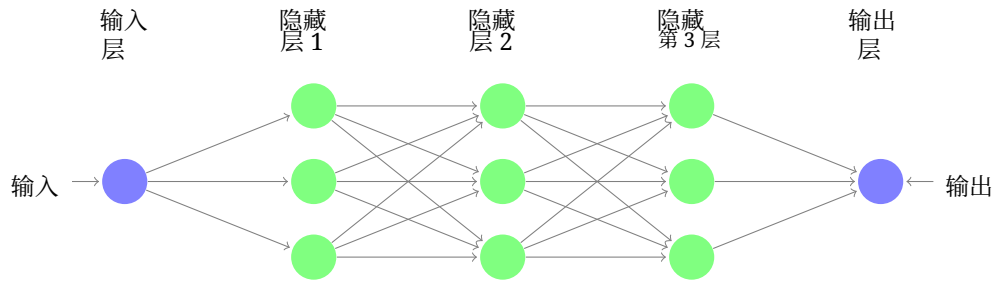

图1。深度神经网络的图形表示，包含 3 个隐藏层、一个输入层和一个输出层。

为了使用深度神经网络估计（2.2）中的均值函数，我们需要了解深度神经网络的一些性质。神经网络是一种参数化函数 f_θ，它将输入空间映射到输出空间；在我们的语境中，f_θ 将 $\mathbb{R}^d$ 映射到 $\mathbb{R}$。一个具有 浅层神经网络（即一个隐藏层）的两层神经网络（通常称为 N 个神经元和激活函数 σ 定义如下：

$$f_\theta(x) = a + \sum_{i=1}^{N} b_i \sigma(c_i^\top x + d_i)\,.$$

其中 $\theta = \{a, b_i, c_i, d_i\}$。因此， f_θ 首先通过线性变换 x 映射到 $\mathbb{R}^N$ 中$x \mapsto Cx + d$，其中 $c_i^\top$是 C 的行， d_i是 d的元素。然后，对 σ 的所有坐标应用非线性激活函数 $Cx + d$。最后，使用映射 $x \mapsto a + \langle b, x\rangle$对各坐标进行线性组合。通过增加更多隐藏层，可以将这个浅层神经网络扩展为深度神经网络。神经网络的流程图形表示见图 1。

对于一个神经网络，我们用 L 表示隐藏层的数量（称为深度），用 N表示隐藏层中神经元数量的最大值（称为宽度）。此后，我们用 $\mathcal{F}_{NN}(d,\ N,\ L,\ W,\ o)$表示所有满足以下条件的神经网络的集合：深度为 L，宽度为 N，激活权重（非零权重）的总数为 W，输入维度为 d，输出维度为 o。若上下文清楚，有时在指定神经网络类别时，我们会省略输入和输出维度。

几十年来，神经网络的表达能力一直备受关注，并在20世纪90年代初开始蓬勃发展。Hornik 等人（1989）证明，所有浅层神经网络（即只有一个隐藏层的网络）构成的集合在所有博雷尔可测函数空间中稠密，并且通过增加隐藏层中的神经元数量，可以将任意博雷尔可测函数 f 估计到任意精度。Barron （1993）量化了精度：如果函数的傅里叶变换具有有限的一阶矩，那么逼近误差以 $1/\sqrt{N}$ 的速率递减，其中 N 是隐藏神经元的数量。Mhaskar （1996）确定了多层前馈网络逼近任意光滑函数的误差，并给出了可实现最优精度的 N, L 选择。此后，关于深度神经网络在宽度、深度、权重和激活函数方面的逼近能力，研究层出不穷。近期，研究人员已精确界定了光滑函数的逼近误差与神经网络深度和宽度之间的关系，参见Yarotsky（2017）、Lu 等人（2021年）、Fan 和 Gu（2022）及其中的参考文献。例如，Lu 等人 （2021年）的定理 1.1 表明，若 $f\colon [0,1]^d \to \mathbb{R}$ 具有 β 阶有界导数，则存在一个宽度为 $c_1 N\log N$、深度为 $c_2 L\log L + d$ 的神经网络 ϕ（c_1, c_2是依赖于 β 的一些常数），其逼近误差满足：

$$\|f - \phi\|_\infty \le C(NL)^{-\frac{2\beta}{d}}\,,$$

其中常数 C 取决于(d,β)。

另一条相关研究路线深入探讨了深度神经网络如何成功适应非参数回归中均值函数未知的底层低维结构。使用合适的深度神经网络进行非参数回归，并实现极小极大最优误差界（至多相差某个多对数因子），最初始于 Kohler 和 Krzyżak（2005）的开创性工作。在标准非参数回归问题中（例如加性噪声模型 $Y = f(X) + \epsilon$），均值函数 f 的估计极小极大速率为 $n^{-2\beta/(2\beta+d)}$，其中 β 是 f 的光滑度指数（假设 2.3），而 d 是 X 的维度。估计误差会受到维数灾难的影响，也就是说，当 d 很大时，收敛速率会非常慢。然而，正如引言中所述，可以通过对 f 施加更多结构来规避这一问题，例如加性结构（见 Stone（1985））或某种高阶交互模型（见 Stone（1994））。在 Kohler 和 Krzyżak（2005）中，作者表明，如果模型是 k 阶交互模型，就可以使用基于神经网络的估计量达到速率 $n^{-2\beta/(2\beta+k)}$。使用神经网络的主要优势在于，我们只需指定其架构，即宽度和层数，而无需明确指定均值函数的具体结构，所得神经网络估计量便能达到极小极大速率最优。近期，在一系列论文中（Kohler 和 Krzyżak，2016；Schmidt-Hieber，2020；Bauerand Kohler，2019；Kohler 和 Langer，2021；Fan 等人，2022；Fan 和 Gu，2022），作者还证明，当 f_0 属于比 k 阶交互模型更一般的一类模型（称为广义层次交互模型）时，可以通过神经网络以速率 $n^{-2\beta/(2\beta+k)}$ 估计底层均值函数 f_0。

上述所有分析都假设 d 的潜在维度是固定的。这与许多现代应用不符。通常，估计速率前的常数取决于 X 的维度，即使对于简单的加性模型也是如此，因为我们需要估计 d 个单变量函数。当 d 固定时，可以采用较为粗略的界。然而，如果 $d \to \infty$，我们需要新的技术来获得关于 d 的最优依赖关系，这是本文的主要目标之一。

现在我们简要介绍我们的估计程序。考虑方程（2.2）中的模型。ˆ ˆ我们通过最小化平方误差损失来选择两个神经网络 $(\phi_1,\ \phi_2)$：

$$\left(\hat{\phi}_1, \hat{\phi}_2\right) = \arg\min_{\phi_1 \in \mathcal{F}_{NN}^1, \phi_2 \in \mathcal{F}_{NN}^2} \frac{1}{n} \sum_{i=1}^{n} (Y_i - \phi_1(X_i) - \phi_2(X_i))^2 \tag{2.3}$$

ˆ ˆˆ ˆ ˆ ˆ并将估计量设为 $f_{\text{big}} := \phi_1 + \phi_2$。最后，将估计量设为 $f := sgn(f_{\text{big}})(|f_{\text{big}}| \wedge B)$，即ˆ将 f 的输出在 $[-B, B]$ 处截断，因为我们假设 $\|f_0\|_\infty \le B$（见假设2.2 的第 3 点）。这里的神经网络类 $\mathcal{F}_{NN}^1$ 和 $\mathcal{F}_{NN}^2$ 定义如下：

$$\mathcal{F}_{NN}^1 = \mathcal{F}_{NN}\left(d, c_1 d N_1 \log N_1, c_2 L_1 \log L_1, c_3 d (N_1 \log N_1)^2 L_1 \log L_1, 1\right) \tag{2.4}$$

$$\mathcal{F}_{NN}^2 = \mathcal{F}_{NN}\left(d, c_1 d^2 N_2 \log N_2, c_2 L_2 \log L_2, c_3 d^2 (N_2 \log N_2)^2 L_2 \log L_2, 1\right) \tag{2.5}$$

对于某些常数 c_1，c_2，c_3。例如，在 $\mathcal{F}_{NN}^2$ 中，所有成对交互都用作输入变量。请注意，这里我们没有使用全连接神经网络，而是分别通过宽度为 $N_1 \log N_1$（分别为 $N_2 \log N_2$）、深度为 $L_1 \log L_1$（分别为 $L_2 \log L_2$）的全连接神经网络来估计每个单变量（分别为双变量）分量函数，然后将它们相加。图2 给出了示意图。略微滥用记号，我们仍用$\mathcal{F}_{NN}^1$ 和 $\mathcal{F}_{NN}^2$ 表示这些具有特殊结构的神经网络。量 N_i 和L_i 通常取决于样本量 n，其具体形式将在后文给出。与其他任何非参数估计量一样， f 的泛化误差 可分解为两部分：

$$\|\hat{f} - f_0\|_{L_2(P_X)} \le \underbrace{\|\hat{f} - f^\star\|_{L_2(P_X)}}_{\text{Statistical error}} + \underbrace{\|f^\star - f_0\|_{L_2(P_X)}}_{\text{Approximation error}} \tag{2.6}$$

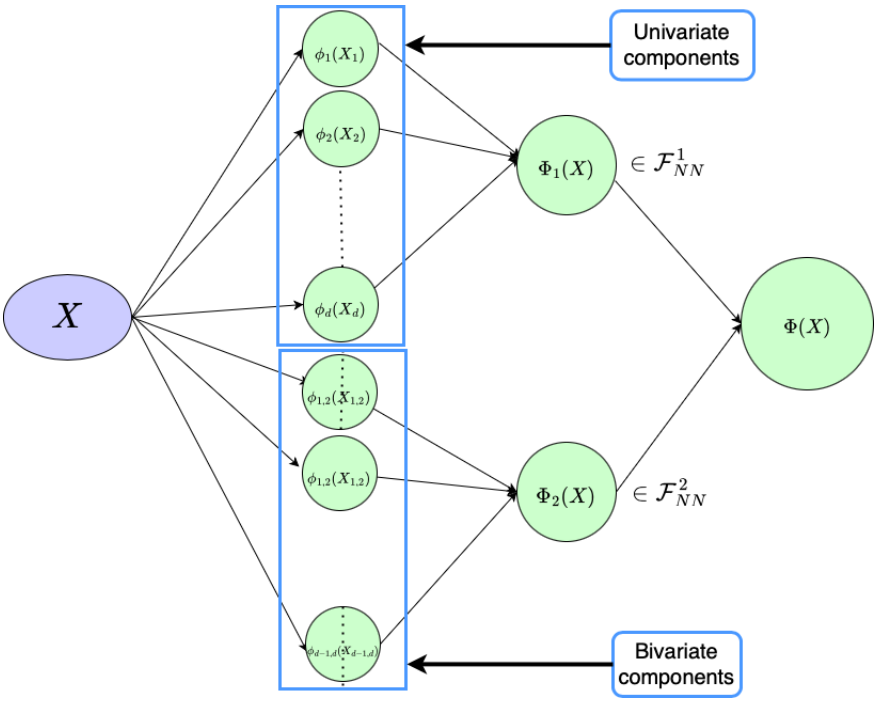

图2。用于估计含交互作用的结构化非参数回归的结构化深度神经网络示意图。第一部分由针对单变量预测因子的加性全连接神经网络组成，第二部分则由全连接双变量神经网络求和而成，用于逼近双变量交互效应。

这里，$f^\star = \phi_1^\star + \phi_2^\star$是在神经网络这一类函数中，按 $L_2(P_X)$范数衡量时对 f_0 的（近似）最佳逼近函数，即

$$\|f^\star - f_0\|_{L_2(P_X)} \leq \inf_{\phi_1 \in \mathcal{F}_{NN}^1, \phi_2 \in \mathcal{F}_{NN}^2} \|f_0 - \phi_1 - \phi_2\|_{L_2(P_X)} + \tau_n$$

对于某个稍后指定的小容差 τ_n。其余分析分为两部分：第2.1 小节界定逼近误差，第2.2 小节则从(d, N_i, L_i, n)的角度界定统计误差。最后，我们将在第2.3 小节选择$\mathcal{F}_{NN}^1$，$\mathcal{F}_{NN}^2$的架构，以平衡这些误差。这将证明，神经网络所达到的收敛速率在(n, d)方面是极小极大最优的，至多相差一个对数因子。

注释 *2.5*。需要特别指出的是，我们将 f_0 估计为两个具有不同架构的神经网络之和，而不是使用单个神经网络。为说明其必要性，注意到估计量的收敛速率为 $dn^{-2\beta_1/(2\beta_1+1)}$ 和 $d^2 n^{-\beta_2/(\beta_2+1)}$（推论 *2.10*），相差一个对数因子，其中 β_1 是单变量分量的光滑度指数，而 β_2 是 f_0的双变量分量的光滑度指数。现在注意，如果 β_2 足够大，则 $dn^{-2\beta_1/(2\beta_1+1)}$ 将是主导速率；否则，$d^2 n^{-\beta_2/(\beta_2+1)}$ 将占主导地位。如果只使用一个神经网络，则所得速率为 $d^2(n^{-2\beta_1/(2\beta_1+1)} + n^{-\beta_2/(\beta_2+1)})$，这不是极小极大最优的（定理 *2.12*），尤其是在 $\beta_2 \gg \beta_1$时。误差项之间的权衡更为复杂，因为 $d := d(n) \to \infty$，而使用两个神经网络（一个用于单变量分量，另一个用于双变量分量）可以获得最优误差界。我们猜想，不可能通过使用单个神经网络获得极小极大误差界。

2.1. 逼近误差。本节计算二阶交互模型（2.2）中均值估计的逼近误差。为此，我们调用 Lu 等人（2021）的定理 1.1，该定理指出，任意 β光滑函数 f： $[0,1]^d \to \mathbb{R}$均可由具有宽度 $O(N\log\ N)$ 和深度 $O(L\log\ L+d)$ 的深度神经网络，在 L_∞范数下以 $(NL)^{-2\beta/d}$的误差逼近 f。回顾假设 2.2， $\int f_j(x)dx=0$ ，且 f_{ij}的两个边缘积分均为 0。如上一节所述，我们将使用两个具有不同深度和宽度的独立神经网络来逼近 f_is 和 f_{ij}s。在陈述主定理之前，我们首先在以下假设中量化 d 相对于 n 的增长界：

假设 2.6. 我们假设 d 随 n 增长，且

$$\left(dn^{-\frac{2\beta_1}{2\beta_1+1}}+\binom{d}{2}n^{-\frac{\beta_2}{\beta_2+1}}\right)\log^{4.5}n=o(1)\,.$$

下面的定理通过神经网络给出了均值函数的逼近误差：

定理 2.7. 考虑方程 (2.2)中的双向交互模型，其中所有单变量分量均为 β_1 光滑，所有双变量分量均为 β_2 光滑。取 $N_1L_1=\lfloor n^{1/2(2\beta_1+1)}\rfloor$, $N_2L_2=\lfloor n^{1/2(\beta_2+1)}\rfloor$ 则有：

$$\inf_{\phi_2\in\mathcal{F}_{NN}^1,\phi_2\in\mathcal{F}_{NN}^2}\mathbb{E}\left[\|f_0(X)-\phi_1(X)-\phi_2(X)\|^2\right]$$
$$\leq C_3\left(d(N_1L_1)^{-4\beta_1}+\binom{d}{2}(N_2L_2)^{-2\beta_2}\right),$$

其中， $\mathcal{F}_{NN}^1$ 和 $\mathcal{F}_{NN}^2$ 与(2.4)和(2.5)中的定义相同。所有常数(C_1,C_2,C_3)均独立于(d,N_1,L_1,N_2,L_2)。

定理 2.7 的证明推迟至第 5节。这里我们给出证明概要。证明的核心思想是构造一个神经网络，使其不仅能够近似 f，还能够（近似地）满足可识别性约束假设 2.2，即 f_{ij} 的边际函数为 0。这将消除高阶项的影响，从而得到最优近似速率。对于单变量分量，该积分约束可以精确满足；考虑函数 f： $[0,1]\to\mathbb{R}$，其中 f 满足假设 2.3，且具有某个 β 和 $\int_0^1 f(x)\ dx=0$。那么，根据（Lu 等人，2021年，定理 1.1），存在一个神经网络 ϕ，其宽度为 $O(N_1\log N_1)$，深度为 $O(L_1\log\ L_1)$，使得

$$\|f-\phi\|_\infty\leq C(N_1L_1)^{-2\beta}$$

其中常数 C 仅取决于 β 和 f 的 Sobolev 范数。定义 $I_\phi:=\int_0^1\phi(x)\ dx$，并定义一个新的神经网络 $\tilde{\phi}=\phi-I_\phi$，以保证 $\int\tilde{\phi}=0$。由于从神经网络中减去一个常数相当于改变最后一层的偏差，因此 $\tilde{\phi}$ 和 ϕ 的架构相同。利用三角不等式，

$$\|f-\tilde{\phi}\|_\infty\leq 2C(N_1L_1)^{-2\beta}\,.$$

上述思路无法用于逼近二元函数 f_{ij}，因为我们不仅需要积分为 0，还需要边际函数为 0，即 $\int f_{ij}(x,y)\ \ dx=\int f_{ij}(x,y)\ dy=0$。为解决这一问题，我们设计了一种去偏技术，概述如下：对于任意具有有界导数且 β 次可微的二元函数 f，存在一个神经网络 $\phi(x,y)$ （使用（Lu 等人，2021年，定理 1.1）），其架构为$O(N_2\log\ N_2)$ $O(L_2\log\ L_2)$，并满足：

$$\|f-\phi\|_\infty\leq C(N_2L_2)^{-\beta}\,.$$

给定这样的神经网络 ϕ，我们构造 ϕ_1 （分别为 ϕ_2），使得 $\phi_1(x)$（分别为 $\phi_2(y)$）是一个宽度为 $O(N_2 \log N_2)$、深度为 $O(L_2 \log L_2)$ 的神经网络，用于逼近 $\int \phi(x,y)\, dy$（分别为 $\int \phi(x,y)\, dx$）。我们的最终估计量形式为 $\tilde{\phi} = \phi - \phi_1 - \phi_2 + \iint \phi\, x\, y\, dx\, dy$ $O\left(NL^{-2\beta}\right)$ $\|\tilde{\phi} - f\|_\infty = O\ NL^{-\beta}$ $N_2 L_2^{-2\beta}$ $\int \tilde{\phi}\, x\ dx = \int \tilde{\phi}\, x\, y\ dy = (,)$。我们证明了 (( )) 和 ( ) (,) ( )。这种较小的逼近误差（(( ) 而非 ( )）对于获得最优误差界至关重要。证明的详情见附录5。

2.2. 统计误差与主定理。 本节将讨论估计的统计误差界定技术，并给出界定样本外预测误差的主定理。估计量的统计误差通常取决于样本数量以及底层函数类的复杂度。回顾（2.3）可知，ˆ ˆ ˆ我们将 f_0 近似为两个神经网络之和： $\phi_1 + \phi_2$ 其中， ϕ_1 近似ˆ单变量分量，而 ϕ_2 近似双变量分量。由定理 2.7 的证明可立即看出，逼近误差取决于 N 和 L 的乘积 NL。因此，我们采用具有恒定深度的神经网络，并对宽度进行优化。所以，$\mathcal{F}_{NN}^1$ 和$\mathcal{F}_{NN}^2$ （方程（2.4）和（2.5））可修订为:

$$\mathcal{F}_{NN}^1 = \mathcal{F}_{NN}\left(d, c_1 d N_1 \log N_1, c_2, c_3 d (N_1 \log N_1)^2, 1\right), \tag{2.7}$$

$$\mathcal{F}_{NN}^2 = \mathcal{F}_{NN}\left(d, c_1 \binom{d}{2} N_2 \log N_2, c_2, c_3 \binom{d}{2} (N_2 \log N_2)^2, 1\right). \tag{2.8}$$

ˆ ˆ在实践中，人们通过梯度下降优化神经网络的权重来估计 ϕ_1， ϕ_2 （例如参见 Goodfellow 等人（2016）的第 6.2 节）。由于就神经网络的权重而言，我们的问题具有高度非凸性，因此无法立即确定此类算法是否会收敛到全局最小值。为避免这些问题，本文采用最小化经验风险的神经网络。基于梯度下降的估计量的性质留待未来研究。

本节旨在界定方差项，即 $\mathbb{E}[(\hat{f}(X) - \phi^\star(X))^2 \mid \mathsf{S}_n]$ 其中$\phi^\star$ 是均值函数 f_0 的最佳逼近函数，且 $\mathsf{S}_n := \{(X_i, Y_i)\ i \in [n]\}$ 为训练样本。预测/测试误差取决于底层函数类的复杂度与训练样本数量之间的权衡。若相对于样本数量 n而言，底层函数类过于复杂，则可能对数据过拟合，从而导致较高的泛化误差。因此，量化神经网络类的复杂度至关重要。Bartlett 等人（2019）的以下引理给出了具有 W 个有效权重和深度 L的神经网络类的 VC 维数上界:

引理 2.8（（5）， Bartlett 等人 （2019）。 设 V（W， L） 表示具有 W 个参数和 L 层的 *ReLU* 网络的最大 *VC* 维数。存在常数 c_1， $c_2 > 0$使得

$$c_1 W L \log(WL) \le V(W, L) \le c_2 W L \log WL.$$

粗略地说，如果我们有一个宽度为 N、深度为 L 的全连接神经网络，那么 $W \sim N^2 L$，因此 VC 维数为 $\sim N^2 L^2$。一旦我们得到了底层函数类复杂度的界，就可以利用经验过程理论中的技术，得到泛化误差的收敛速率。对于 VC 维数为 V 的 VC 型函数类，统计误差为 $O(\sqrt{V/n})$，相差至多一个对数因子。为说明这一假设的必要性，考虑推论 2.10 得到的收敛速率。假设 2.6 保证了估计量的一致性。另一方面，该假设在相差一个对数因子的意义下也是必要的，正如我们在第 4 节中所证明的，推论2.10

得到的收敛速率在相差一个对数因子的意义下是极小极大最优的（定理2.12）。因此，在没有其他结构性假设的情况下，该假设是关于 d 相对于 n 的增长所能作出的最弱假设。以下定理给出了总体逼近误差的一个界，同时考虑了统计误差和逼近误差。

定理 2.9（主定理）。在假设 *2.1 - 2.3* 和 *2.6*下，经验风险最小化估计量 $\hat{f}$，其中 f_0 定义于方程(2.3) 中，满足：

$$\mathbb{E}\left[\left(\hat{f}(X)-f_0(X)\right)^2 \mid \mathbf{S}_n\right]=O_p\left(\rho_n^2+\frac{V_n}{n}\log^{3/2} n\right),$$

其中

$$\rho_n^2=\textit{approximation error}\le C_1\left(dN_1^{-4\beta_1}+\binom{d}{2}N_2^{-2\beta_2}\right)$$

$$V_n=\textit{complexity}\le C_2\left(dN_1^2\log^2 N_1\log\left(dN_1\right)+\binom{d}{2}N_2^2\log^2 N_2\log\left(dN_2\right)\right).$$

该定理的证明见附录A.1。现在，我们需要谨慎选择数值N_1，N_2 以平衡逼近误差和统计误差。这导出了以下推论：

推论 2.10。在定理 $N_1=n^{1/2(2\beta_1+1)}$ 和 $N_2=n^{1/2(\beta_2+1)}$ 中取 *2.9*的值（若不是整数，则取最接近的整数），可得：

$$\mathbb{E}\left[\left(\hat{f}(X)-f_0(X)\right)^2 \mid \mathbf{S}_n\right]=O_p\left(\left(dn^{-\frac{2\beta_1}{2\beta_1+1}}+\binom{d}{2}n^{-\frac{\beta_2}{\beta_2+1}}\right)\log^{4.5} n\right).$$

在下一节中，我们将证明该收敛速率在对数因子范围内是极小极大最优的。由此分析直接引出的一个问题是，我们能否改善对数因子的依赖关系。近期文章（Fan 和 Gu，2022）提出了一种与Lu 等人 （2021年）所用方法略有不同的逼近技术，可能会降低对数因子的幂次，使其从4.5下降。然而，更有趣的问题是，我们能否彻底消除对数因子；这个问题目前仍未解决。

注释 2.11. *[向 k 阶交互模型的扩展]* 我们对二阶交互模型的分析也可原样扩展到 k 阶交互模型。下面简要说明这一扩展。为了保证模型的可识别性，我们需要假设 j 分量函数的所有边缘分布均为 0, $j\ge 2$。我们通过以下 k 个神经网络之和来估计 f_0：$\phi_1+\cdots+\phi_k$其中

$$\phi_i\in\mathcal{F}_{NN}\left(d,c_1\binom{d}{i}N_i\log N_i,c_2,\binom{d}{i}(N_i\log N_i)^2,1\right).$$

由于我们在证明中使用的去偏技术依赖于用神经网络逼近多项式，因此可将其推广到一般的 k元交互模型。进而，这控制了逼近误差。根据引理，统计误差可以类似地通过神经网络的 *VC* 维度界定*2.8*。

2.3. 极小极大下界。 本节针对非参数 2元交互模型（2.2），在协变量的维度 $d = d(n) \to \infty$， $d = o(n)$ 满足假设 2.6的条件下，建立其估计的极小极大下界。对于我们的估计问题，极小极大风险定义如下：

$$\mathfrak{M}(n, d, \mathcal{F}) = \inf_{\hat{f}} \sup_{\substack{f \in \mathcal{F} \\ X \sim P_X}} \mathbb{E}_f \left[\left(\hat{f}(X) - f(X) \right)^2 \right]$$

其中 $f(X) = \sum_{j=1}^{d} f_j(X_j) + \sum_{i<j} f_{ij}(X_i, X_j)$，且上确界是在所有 $\{f_j\}$和 $\{f_{ij}\}$ 上取的，其中分量函数属于 $\Sigma(\beta, L)$ （见假设 2.3）。

建立极小极大下界的主要难点在于纳入维度增长 d 的影响。显然，我们的问题比估计$f_1(x_1) + f_{12}(x_1, x_2)$更难，因为我们假设其他所有分量都已预先知晓。问题在于，在维度增长的 d 情形下，其他分量会如何影响下界。以下定理给出了精确答案。

定理 2.12. 考虑方程 (2.2)中定义的二阶交互模型，其中每个分量函数 $\{f_i\}_{i=1}^{d}$ 和 $\{f_{ij}\}_{1 \le i < j \le d}$ 属于 $\Sigma(\beta, L)$ （见假设 *2.3* ），并将这一均值函数集合记为 $\mathcal{F}$。那么，在假设 *2.1–2.3*下，估计的极小极大速率为：

$$\mathfrak{M}(n, d, \mathcal{F}) = \inf_{\hat{f}} \sup_{\substack{f \in \mathcal{F} \\ X \sim P_X}} \mathbb{E}_f \left[\left(\hat{f}(X) - f(X) \right)^2 \right] \ge c \left(d n^{-\frac{2\beta_1}{2\beta_1 + 1}} + \binom{d}{2} n^{-\frac{2\beta_2}{2\beta_2 + 2}} \right) .$$

对于某个与（n， d）独立的常数 C 。

该定理的证明见附录A.2。这里需要说明几点；首先，定理2.12 确立了以下事实：均值函数 f_0 的基于神经网络的估计在相差一个多对数因子的意义下是极小极大速率最优的。这一结果补充了Kohler 和 Langer （2021年）以及Schmidt-Hieber （2020）关于固定维度情形下非参数回归的研究。其次，定理2.12 是在假设2.3下推导的，其中我们假设所有单变量分量都是 β_1-光滑的，且所有双变量分量都是 β_2-光滑的。然而，我们的证明可以很容易地调整到每个函数具有不同光滑度的情形，但需要付出最低光滑度所带来的代价。参见注释2.11 ，其中我们指出了证明这一推广结果的关键步骤。

注2.13。*[k* 阶交互模型的推广*]* 二阶交互模型的极小极大下界证明可以通过类似地构造备择函数并使用 *Fano* 不等式，轻松推广到一般的 k–阶交互模型。最优速率为

$$\inf_{\hat{f}} \sup_{f \in \mathcal{F}, X \sim P_X} \mathbb{E}_{P_X} [(\hat{f}(X) - f(X))^2] \gtrsim \left(\sum_{j=0}^{k} \binom{d}{j} n^{-\frac{2\beta_j}{2\beta_j + j}} \right)$$

REMARK 2.14. 我们在本文中研究了交互模型的极小极大上界和下界中维数的影响。一个自然的下一步是求出发散维数下 k 阶交互模型的 *Pinsker* 常数，这留待未来研究。

3. $d \gg n$下 DNN 的分析。 本节介绍当 X 的潜在维度大于样本量时的分析。尽管为避免维数灾难，较简单的参数模型已得到充分研究，但关于这一情形下 k 阶交互模型（2.1）的非参数估计，相关文献相对较少。在非参数框架下，针对带有稀疏性约束的高维加性模型，已有大量研究，其考虑的模型如下：

$$Y_i = \sum_{j \in S} f_{0,j}(X_{ij}) + \epsilon_i$$

其中 $S \subseteq \{1, \ldots, d\}$ 满足 $s = |S| \ll n$。这是高维统计分析中的一个标准假设：真实信号仅依赖于少数几个协变量，但活跃集 S 事先未知。高维稀疏加性模型的研究始于Lin and Zhang (2006年)，随后在一系列文献中得到进一步完善（Koltchinskii and Yuan，2010年；Ravikumar 等，2009；Tanand Zhang，2019；Yuan and Zhou，2016）。Koltchinskiiand Yuan (2010年) 在全局有界性假设下获得了误差界，随后该假设被 Raskutti 等人(2012) 移除。最近，Tan and Zhang (2019) 在底层分量函数属于再生核希尔伯特空间时，获得了极小极大保证。以往的研究通常使用两个惩罚项来实现最优估计：一个用于控制底层函数类的复杂度，另一个用于控制稀疏性。然而，对于基于神经网络的估计量而言，只需使用一个惩罚项来诱导稀疏性即可，因为复杂度可以通过网络的宽度和深度加以控制。

尽管针对稀疏线性模型的研究已十分广泛，但关于含有 k阶交互作用（$k > 1$）的高维线性模型的研究却很少，更不用说 k阶非参数交互作用了。关于高维线性交互模型中的变量选择，读者可参阅Zhao 等 (2009)、Bien 等 (2013年)、Haoand Zhang (2014) 及其中的参考文献。本节将通过研究神经网络估计量的渐近性质来弥补这一空白。

与前文一样，我们仅详细分析二阶交互模型，并简要讨论如何将分析推广到一般的 k阶交互模型。二阶稀疏加性模型定义如下：

$$Y_i = \sum_{j \in S_1} f_{0,j}(X_{ij}) + \sum_{(k,l) \in S_2} f_{0,kl}(X_{ik}, X_{il}) + \varepsilon_i \equiv f_0(X_i) + \varepsilon_i, \tag{3.1}$$

其中 $S_1 \subset [d]$, $S_2 \subset [d] \times [d]$ 满足 S_1，S_2 稀疏，即 $s_i := |S_i| \ll n$ 对 $i = 1$，2成立。用 $\mathcal{F}_{\text{sp}}$ 表示这一函数类。这里我们同样假设单变量分量是β_1 光滑的，双变量分量是 β_2 光滑的。根据定理2.9，存在$\{\phi_j^\star\}_{j \in S_1}$，$\{\phi_{kl}^\star\}_{(j,k) \in S_2}$, $j \in S_1$，$(j,\ k) \in S_2$，使得 $\phi_j^\star \in \mathcal{F}_{NN,\ 1}$, $\phi_{jk}^\star \in \mathcal{F}_{NN,\ 2}$并且

$$\|f_{0,j} - \phi_j^\star\|_\infty \le C_1 N_1^{-2\beta_1} \ \forall\, j \in S_1\,,$$
$$\|f_{0,kl} - \phi_{kl}^\star\|_\infty \le C_2 N_2^{-\beta_2} \ \forall\, (k < l) \in S_2\,.$$

其中

$$\mathcal{F}_{NN,1} = \mathcal{F}_{NN}(1, c_1 N_1 \log N_1, c_2, c_3 N_1^2 \log^2 N_1, 1) \tag{3.2}$$

$$\mathcal{F}_{NN,2} = \mathcal{F}_{NN}(2, c_4 N_2 \log N_2, c_5, c_6 N_2^2 \log^2 N_2, 1)\,. \tag{3.3}$$

最后，定义

$$\phi^\star = \sum_{j \in S_1} \phi_j^\star + \sum_{(k<l) \in S_2} \phi_{kl}^\star. \tag{3.4}$$

根据假设2.6，我们对稀疏度参数 s_1, s_2作出如下增长条件假设：

假设 3.1. 与单变量分量和双变量分量相对应的稀疏度参数 s_1, s_2 满足以下条件：

$$s_1\left(n^{-\frac{2\beta_1}{2\beta_1+1}}\log^4 n+\frac{\log d}{n}\right)+s_2\left(n^{-\frac{\beta_2}{\beta_2+1}}\log^4 n+\frac{\log d}{n}\right)=o(1)\,.$$

假设3.1 的原因与假设2.6类似，即如果没有这一假设，就不存在一致估计量（忽略对数因子），因为该速率是极小极大最优估计速率。高维模型的分析并非对 $d=o(n)$时所用技术的直接扩展。为理解其中缘由，请回想，即使在一个简单的稀疏线性模型中，采用 ℓ_1 惩罚项，也无法在缺少某种形式的受限强凸性假设时获得极小极大速率最优估计量；而当维度增长慢于样本量时，通常并不需要这一假设。因此，为便于阐述，我们将整个分析分为三部分：第3.1 节给出固定设计模型的收敛速率，即 X_i是 $\mathbb{R}^p$中的一些固定点。第3.2 节讨论随机设计，即假设 X 是一个随机变量。最后，在第3.3 节中，我们给出极小极大下界，以证明我们的基于神经网络的估计量在对数因子范围内达到极小极大速率最优。

3.1. 固定设计模型的分析。 本节中，我们对固定设计模型进行分析，并假设 X_i 为固定值。与（2.3）类似，我们使用具有不同架构的神经网络估计单变量分量和双变量分量，同时使用 ℓ_1-范数的 $L_2(\mathbb{P}_n)$ 惩罚项来实现稀疏性。关于利用此类方法选择变量组的思路，参见 Antoniadis 和 Fan （2001）第 966 页以及 Yuan 和 Lin （2006）。我们的 ^ ^ ^估计量 $f=\sum_j\phi_j+\sum_{k<l}\phi_{kl}$，其分量定义为

$$\left\{\hat\phi_j\right\},\left\{\hat\phi_{kl}\right\}=\underset{\substack{\phi_j\in\mathcal{F}_{NN,1}\\ \phi_{kl}\in\mathcal{F}_{NN,2}}}{\arg\min}\left[\frac{1}{n}\sum_i\left(Y_i-\sum_j\phi_j(X_{ij})-\sum_{k<l}\phi_{kl}(X_{ik},X_{il})\right)^2\right.$$
$$\left.+\left(\lambda_{n,1}\sum_j\|\phi_j\|_n+\lambda_{n,2}\sum_{k<l}\|\phi_{kl}\|_n\right)\right] \tag{3.5}$$

其中 $\mathcal{F}_{NN,1}$ 和 $\mathcal{F}_{NN,2}$ 分别在（3.2）和（3.3）中定义，$\|\cdot\|_n$ 为 $L_2(\mathbb{P}_n)$ 范数，而 $\lambda_{n,1}$,$\lambda_{n,2}$ 将在后文说明。

现在对估计程序作几点说明。首先，尽管我们对分量函数施加L_2（$\mathbb{P}_n$）惩罚项以促进稀疏性，但也可以对权重的最后一层施加 ℓ_1 惩罚项（该层将分量神经网络相加）。尽管这种惩罚项具有计算优势，但从理论上来说很难进行分析。其次，由于我们不一定假设 $\beta_1=\beta_2$，因此需要两个正则化参数$\lambda_{n,\,1}$ 和$\lambda_{n,\,2}$。我们还需要采用不同的架构来适应光滑度，正如备注2.5所强调的那样。此外，为了技术上的简便，我们假设以下 l^∞ 界：

$$\|f_j\|_\infty\le B,\qquad \|f_{jk}\|_\infty\le B. \tag{3.6}$$

因此，为了保证数值稳定性，我们将在水平 B 处截断分量神经网络（即ϕ_j、ϕ_{jk}）。以下定理给出了收敛速率。

定理3.2. 在假设 *2.1-2.3* 和*3.1* 以及 (3.6)的条件下，式(3.5) 中得到的估计量满足

$$\|\hat f-f_0\|_n^2=O_p\left(s_1\left(\rho_{n,1}^2+\lambda_{n,1}+\frac{\lambda_{n,1}^2}{2}\right)+s_2\left(\rho_{n,2}^2+\lambda_{n,2}+\frac{\lambda_{n,2}^2}{2}\right)\right)$$

其中 $\rho_{n,\ 1}$ (分别为 $\rho_{n,\ 2}$) 是均值函数的单变量（分别为双变量）分量由神经网络逼近时的逼近误差，其上界由式(3.8)给出，前提是

$$\lambda_{n,1} = C_3\sqrt{\frac{V_{n,1}\log n}{n} + \frac{2\log d}{n}}, \quad \lambda_{n,2} = C_4\sqrt{\frac{V_{n,2}\log n}{n} + \frac{3\log d}{n}}, \tag{3.7}$$

其中 $V_{n,\ 1} = N_1^2\log^3 N_1$ 以及 $V_{n,\ 2} = N_2^2\log^3 N_2$。

让我们尝试理解并简化定理 3.2 得到的收敛速率。回顾定理2.7，我们有：

$$\rho_{n,1} \le C_1 N_1^{-2\beta_1}, \quad \rho_{n,2} \le C_2 N_2^{-\beta_2}, \tag{3.8}$$

其中常数为 $C_1, C_2 > 0$。此外，我们的证明（见 (A.19) 和 (A.20)）表明，由 (3.7) 给出的 λ_n,1 和 λ_n,2 是最优选择，其中 V_n,1 和 V_n,2 分别是 $\mathcal{F}_{NN}$,1 和 $\mathcal{F}_{NN}$,2 的 VC 维数阶。现在，通常 $\lambda_{n,i} = o(1)$（只要 $V_{n,i} = o(n/\log\ n)$ 就成立，而这是相合性所必需的条件），因此 $\lambda_{n,i}^2 = o(\lambda_{n,i})$。因此，定理 3.2 中的收敛速率可简化为

$$\|\hat{f} - f_0\|_n^2 = O_p\left(s_1\left(\rho_{n,1}^2 + \lambda_{n,1}\right) + s_2\left(\rho_{n,2}^2 + \lambda_{n,2}\right)\right).$$

现在我们选择 N_1， N_2 以平衡 $\rho_{n,\ i}^2$ 和 $\lambda_{n,\ i}$，从而得到以下推论：

推论 3.3. 选择 $N_1 = n^{1/(2(1+4\beta_1))}$ 和 $N_2 = n^{1/(2(1+2\beta_2))}$ （若它们不是整数，则取最接近的整数），我们得到：

$$\|\hat{f} - f_0\|_n^2 = O_p\left(s_1\left(n^{-\frac{2\beta_1}{1+4\beta_1}}\log^2 n + \sqrt{\frac{\log d}{n}}\right) + s_2\left(n^{-\frac{\beta_2}{1+2\beta_2}}\log^2 n + \sqrt{\frac{\log d}{n}}\right)\right).$$

c正如我们将在第 3.3 小节中看到的，这一速率并非极小极大最优。原因与高维稀疏回归模型类似：如果不对损失函数的曲率作任何假设（例如受限等距性或受限特征值类型的条件），那么就不可能得到任何能够在多项式时间内计算的极小极大最优估计量，正如 Zhang 等人 (2014) 所证明的。因此，我们需要假设某种形式的 受限强凸性 (RSC)，以获得极小极大最优误差界。RSC 类型的假设由 Bickel 等人 (2010年)；Candès 和 Tao (2007) 推广开来。稀疏加性回归模型的分析中也采用了类似的假设，例如参见（ Tanand Zhang，2019，假设 3）。具体而言，我们作出如下假设：

假设 3.4. 存在常数 κ_1， $\kappa_2 > 0$，使得对于任何满足以下条件的函数 $\phi = \sum_j \phi_j + \sum_{k<l}\phi_{kl}$：

$$\begin{aligned}
&\lambda_{n,1}\sum_{j\in S_1^c}\|\phi_j - \phi_j^\star\|_n + \lambda_{n,2}\sum_{(k<l)\in S_2^c}\|\phi_{kl} - \phi_{kl}^\star\|_n \\
&\le 4(s_1\rho_{n,1}^2 + s_2\rho_{n,2}^2) + 3\lambda_{n,1}\sum_{j\in S_1}\|\phi_j - \phi_j^\star\|_n + 3\lambda_{n,2}\sum_{(k<l)\in S_2}\|\phi_{kl} - \phi_{kl}^\star\|_n \\
&\quad + s_1\lambda_{n,1}^2 + s_2\lambda_{n,2}^2
\end{aligned} \tag{3.9}$$

也满足：

$$\kappa_1^2\sum_{j\in S_1}\|\phi_j - \phi_j^\star\|_n^2 + \kappa_2^2\sum_{(k<l)\in S_2}\|\phi_{kl} - \phi_{kl}^\star\|_n^2 \le \|\phi - \phi^\star\|_n^2 \tag{3.10}$$

其中 $\phi^\star$ 由 (3.4) 定义，$\rho_{n,1}$ 和 $\rho_{n,2}$ 在定理 *3.2*中定义，而 $\lambda_{n,1}$ 和 $\lambda_{n,2}$ 定义于(3.7)。

在进一步讨论细节之前，我们先将我们的假设与高维线性模型的标准受限特征值（RE）条件进行比较Bickel 等人 (2009)。粗略地说，RE 条件假设：

$$\sum_{j\in S}(\beta_j-\beta_{0,j})^2 \lesssim \frac{1}{n}\|X(\beta-\beta_0)\|_2^2 \tag{3.11}$$

对于满足以下条件的任意 β：

$$\sum_{j\in S^c}|\beta_j| \lesssim \sum_{j\in S}|\beta_j-\beta_{0,j}|\,, \tag{3.12}$$

其中 S 是真实活跃集。现在考虑非参数加性模型，其中 X_j通过 Y影响 $f_j(X_j)$。此外，我们通常使用 $L_2(\mathbb{P}_n)$ 范数之和作为惩罚函数，以替代 L_1 范数来实现稀疏性。暂时忽略逼近误差（即假设真实的 f_j 属于我们进行优化所依据的函数类），条件（3.11）的一个自然类比是：

$$\sum_{j\in S}\|f_j-f_{0,j}\|_n^2 \lesssim \|\hat{f}-f_0\|_n^2 \tag{3.13}$$

对于所有 $f=\sum_j f_j$ 满足以下条件的（类似于（3.12））

$$\sum_{j\in S^c}\|f_j\|_n \lesssim \sum_{j\in S}\|f_j-f_{0,j}\|_n\,. \tag{3.14}$$

这正是 Tan and Zhang（2019，假设 3）在分析高维加性模型时所采用的条件。假设 3.4 与该条件几乎相同，只是由于我们通过神经网络逼近分量函数，因此可能存在非零逼近误差，需要进行一些修改。忽略双变量分量，式（3.10）与式（3.13）完全相同，只需将 f 替换为 ϕ，将 f_0 替换为 $\phi^\star$；后者是从神经网络函数类中对 f_0 的最佳逼近函数。唯一的区别在于式（3.9）：由于需要考虑逼近误差，它与式（3.14）有所不同。这正是为什么式（3.9）的右侧会出现项 $s_1\rho_{n,1}^2+s_2\rho_{n,}^2{}_2$ 的原因。另一个附加项 $s_1\lambda_{n,1}^2+s_2\lambda_{n,}^2{}_2$ 则源于证明中的某种数学伪影，并且可以通过增大阶数使其任意变小。为简便起见，如果暂时忽略双变量分量和这一附加项，则式（3.9）可简化为：

$$\lambda_{n,1}\sum_{j\in S_1^c}\|\phi_j-\phi_j^\star\|_n \le 4s_1\rho_{n,1}^2+3\lambda_{n,1}\sum_{j\in S_1}\|\phi_j-\phi_j^\star\|_n\,. \tag{3.15}$$

如果函数类设定正确（例如线性回归或Tan and Zhang （2019）中的加性模型），则可去掉右侧的第一项。因此，两边消去$\lambda_{n,\ 1}$ 后，我们的条件就与Tan and Zhang（2019）中使用的条件（3.13）相同。现在给出统计误差定理：

ˆ定理3.5. 设函数 ϕ 由(3.5)定义。那么，在与定理*3.2* 相同的假设以及假设*3.4*下，我们有：

$$\|\hat{f}-f_0\|_n^2 = O_p\left(s_1(\rho_{n,1}^2+\lambda_{n,1}^2)+s_2(\rho_{n,2}^2+\lambda_{n,2}^2)\right)\,. \tag{3.16}$$

该定理的证明见附录 A.3。定理 3.2 与定理3.5 所得收敛速率的关键区别在于，后者没有 $\lambda_{n,i}$，因此在假设 3.4下可获得更快的收敛速率。这里也需要像定理 3.2 之后那样进行讨论；我们需要选择 N_1, N_2，以平衡 $\rho_{n,i}^2$ 和 $\lambda_{n,i}^2$。由此得到以下推论：

推论 3.6. 选择 $N_1 = n^{1/(2(1+2\beta_1))}$ 和 $N_2 = n^{1/(2(1+\beta_2))}$（若它们不是整数，则取最接近的整数），可得：

$$\|\hat{f} - f_0\|_n^2 = O_p\left(s_1\left(n^{-\frac{2\beta_1}{1+2\beta_1}}\log^4 n + \frac{\log d}{n}\right) + s_2\left(n^{-\frac{\beta_2}{1+\beta_2}}\log^4 n + \frac{\log d}{n}\right)\right).$$

在第 3.3 节中，我们证明了一个匹配的下界，从而得出该速率实际上在对数因子范围内是极小极大最优的。

3.2. 随机设计设定。 在上一节中，我们分析了固定设计模型，即假设 X_i 为固定值。此处，我们将分析扩展到随机设计模型，也就是说， X_i 是满足假设2.1所述分布的随机变量。由于假设 X_i 为随机变量，我们从期望平方误差损失的角度分析估计量的表现，该损失也称为泛化误差/样本外误差。先来说明本小节与上一节的区别。在固定设计设定中，整个分析都基于给定的 $(X_1, \ldots, X_n)$，也就是说，我们无需评估预测器在某些未观测到的 X 上的表现。然而，当假设 X 为随机变量时，就必须评估其在未见过的 X 上的表现，因为未来我们很可能（事实上，如果 X 的某个分量是连续的，则该概率为 1）需要对新的观测值进行预测。因此，在这种设定下，我们需要界定期望平方误差损失，即 $\|\hat{} f - f_0\|_{L_2(P_X)}^2$，而不是经验平方误差损失，即 $\|\hat{} f - f_0\|_n^2$。

要控制这一泛化误差，一个关键步骤是控制相应经验过程的波动，从而确保在训练样本上表现良好的预测器，在此前未观测到的测试样本上同样表现良好。通常，仅使用 L_1-范数和 $L_2(\mathbb{P}_n)$ 惩罚项很难实现这种保证，因为该惩罚项只能控制函数在观测样本上的复杂度。尽管可以通过对分量函数同时使用 L_∞ 惩罚项和L_1-范数 $L_2(\mathbb{P}_n)$ 惩罚项来获得极小极大最优估计量，但优化过程在计算上颇具挑战。为规避这一问题，我们提出了一种计算高效的两步法；该方法只需稍强的假设，即可使我们对真实均值函数进行最优估计。我们通过算法 1 展示了我们的估计量。

其核心思想如下：首先，我们将样本分成两个大小（几乎）相等的部分。基于样本的第一部分，我们采用固定设计情形下使用的相同惩罚方法来估计分量函数（见（3.5））。接下来，我们对单变量和双变量分量函数的估计量进行硬阈值处理，以估计活跃集（算法 ^ ^1 的步骤 3）。分别记估计得到的活跃集为 S_1 和 S_2。注意，我们的阈值水平与（3.5）中所用的惩罚项成正比。常数 c_1、c_2 将在^ ^证明中明确说明。得到 S_1、S_2后，我们基于数据的第二部分，仅在活跃集 上求解无惩罚最小二乘问题（算法 1 的步骤 4），以获得最终估计值。

我们的算法主要受到 SURE独立筛选（Fan和Lv，2008年；Fan和Song，2010年）以及通过 LASSO 或其他惩罚回归（Belloni和Chernozhukov，2013年）选出活跃集后进行最小二乘回归这一思想的启发。SURE独立筛选通过对加性模型进行边际回归来筛选活跃变量；而对于稀疏线性回归，Belloni和Chernozhukov （2013年）提出先用 LASSO 选出活跃集，再对选定子集进行普通最小二乘法（OLS），以减少偏差。我们在算法 1 中结合了这两种思想。首先，我们估计单变量和^ ^双变量分量的活跃子集 S_1, S_2（算法 1 的步骤 2-3），并确保以下结论以高概率成立：^ ^i） $S_i \subseteq S_i$，以及 ii）$|S_i \cap S_i^c| = O(|S_i|)$。

我们需要 i）避免因忽略活跃分量而可能产生的任何偏差；ii）防止出现假阳性，即不要选择过多的非活跃变量。为此，我们需要以下假设：

算法 1： 随机设计设定下的估计

输入： 常数 c_1，c_2 和惩罚参数 $\lambda_{n,\ 1}, \lambda_{n,\ 2}$.**输出：** $\hat{}\ f$：基于神经网络的估计量。 **数据：** 数据集 $\{(X_1,\ Y_1)\ldots,(X_n,\ Y_n)\}$。

1 将数据集分成两半，每组各含 $n/2$ 个数据（如果 n 为奇数，则第一半取 $(n+1)/2$ 个数据，第二半取 $(n-1)/2$ 个数据）。将这两半记为 $\mathcal{D}_1$ 和 $\mathcal{D}_2$。

2 使用 $\mathcal{D}_1$，估计分量函数 $\{\hat{}\phi_{\text{init}_j}\ \}$ 和 $\{\hat{}\phi_{\text{init}_{jk}}\}$，通过求解（3.5）。

3 Set $\hat{S}_1 = \{j : \|\hat{\phi}_j^{\text{init}}\|_n \geq c_1\lambda_{n,1}\}$ and $\hat{S}_2 = \{(j<k) : \|\hat{\phi}_{jk}^{\text{init}}\|_n \geq c_2\lambda_{n,2}\}$.

$\hat{}\ \hat{}$4 在估计得到的活跃集 S_1 和 S_2 上重新估计分量函数，方法是最小化（未加惩罚的）平方误差损失：

$$\{\hat{\phi}_j^{\text{final}}\}_{j\in\hat{S}_1}, \{\hat{\phi}_{jk}^{\text{final}}\}_{(j,k)\in\hat{S}_2} = \arg\min_{\substack{\phi_j\in\mathcal{F}_{NN,1}\\ \phi_{j,k}\in\mathcal{F}_{NN,2}}} \frac{1}{n}\sum_i \left(Y_i - \sum_{j\in\hat{S}_1}\phi_j(X_{ij}) - \sum_{(j,k)\in\hat{S}_2}\phi_{jk}(X_{ij},X_{ik})\right)^2. \tag{3.17}$$

5 返回最终估计值 $\hat{}\ f\text{final} = \sum_{j\in\hat{}S_1}\hat{}\phi\text{final}_j \quad + \sum_{(j<k)\in\hat{}S_2}\hat{}\phi\text{final}_{jk}$。

假设 3.7. 假设所有分量函数都具有最小信号强度r_n，即 $\min_{j\in\mathcal{S}_1}\|f_j^0\|_2 \geq r_n$，$\min_{(j,k)\in\mathcal{S}_2}\|f_{j,\ k}^0\|_2 \geq r_n$ 其中：

$$r_n \gtrsim \sqrt{\left(s_1\left(n^{-\frac{2\beta_1}{1+2\beta_1}}\log^4 n + \frac{\log d}{n}\right) + s_2\left(n^{-\frac{\beta_2}{1+\beta_2}}\log^4 n + \frac{\log d}{n}\right)\right)}.$$

注意，r_n 恰好是推论3.6中得到的收敛速率。假设3.7确保活跃分量中的信号足够强，从而能够以高概率检测到（确保真实活跃集是估计活跃集的子集）。此外，通过选择最优惩罚项$\lambda_{n,\ i}$我们控制假阳性的数量。最后，在算法1的第 4 步中，我们在选定子集上优化平方误差损失，以得到$\hat{}$最终估计值。以下定理给出了通过算法 f_{final}1获得的估计量的收敛速率：

定理 3.8. 假设受限强凸性假设（假设3.4）在 $\{X_1,\ldots,X_n\}$ 上以高概率成立。那么，在假设3.7下，估计量$\hat{}\ f^{final}$ 满足以下关系（忽略对数因子）：

$$\|\hat{f}^{final} - f_0\|_2^2 = O_p\left(s_1\lambda_{n,1}^2 + s_2\lambda_{n,2}^2\right).$$

其中 $\lambda_{n,\ 1}$ 和 $\lambda_{n,\ 2}$ 与定理3.5相同。

该定理的证明将推迟至附录A.4。

3.3. 极小极大下界。 本节将建立高维双向交互模型收敛速率的极小极大下界。如前所述，我们的证明可以推广到一般的 k 阶交互模型。主定理如下：

定理 3.9. 假设我们有 n 个观测值，来自模型 (3.1)。那么在假设2.2、2.1下，我们有：

$$\inf_{\hat{f}} \sup_{\substack{f\in\mathcal{F}_{sp}\\ X\sim P_X}} \|\hat{f}-f\|_2^2 \gtrsim \left(s_1\left(n^{-\frac{2\beta_1}{2\beta_1+1}} \vee \frac{\log(d/s_1)}{n}\right) + s_2\left(n^{-\frac{2\beta_2}{2\beta_2+2}} \vee \frac{\log\left(d^2/s_2\right)}{n}\right)\right)$$

显然，由定理 3.9 和定理 3.8 可知，基于神经网络的均值函数估计量确实在对数因子范围内达到极小极大最优。

尽管定理 3.9 和定理2.12的证明技巧相似，但由于此处需要施加稀疏性约束，备择函数的构造有一些重要变化。我们还想指出，我们的证明不同于Raskutti 等人 (2012)的证明，因为我们不假设均值函数的各分量属于再生核希尔伯特空间。下面简要概述证明的主要思路：假设每个单变量分量函数都是 β_1-光滑的，并且每个双变量分量函数都是 β_2-光滑的。我们首先为这些分量构造一组备择函数，然后对它们进行稀疏组合。最后，通过构造备择函数并恰当选择相关超参数，得到下界。证明详见附录。

4. 结论。 深度神经网络在非参数函数估计中取得了巨大成功。然而，由于“维数灾难”固有的困难，在不对函数类施加过多限制的情况下，它们只能处理低维非参数函数。这促使人们研究低维非参数交互模型。与此同时，现代大数据应用通常涉及包含大量预测变量的非参数回归。然而，大多数关于神经网络的统计理论仅关注有限维回归。因此，迫切需要研究发散维数下的非参数交互模型，并理解维数对这类结构化非参数模型的影响。

本文对理解发散维数下神经网络在低阶交互模型中的性能作出了重要贡献。我们的研究得出的一个重要结论是，估计出的分量应具有较小的偏差，以避免不必要的误差累积；我们通过新提出的去偏技术实现了这一点。对于缓慢发散的维数问题，无需额外的正则化。经过适当的去偏后，结构化神经网络上的直接最小二乘估计被证明能够达到与新建立的极小极大下界相匹配的收敛速率，即达到了最优。在高维情形下，对交互项作出稀疏性假设是必要的。我们采用惩罚最小二乘估计和筛选技术（针对随机设计），并通过建立相匹配的下界，证明所得程序是极小极大最优的。我们的结果全面展示了发散维数下神经网络对于结构化非参数模型的性能；这对现代数据科学至关重要，也是神经网络取得成功所必需的。

5. 定理2.7的证明。假设 $f\colon [0,1]^d \to \mathbb{R}$ 是 β-光滑的，即它可微 $\lfloor\beta\rfloor$ 次且具有有界导数。于是，通过在某个点x_0附近进行泰勒级数近似，我们得到：

$$\begin{aligned} f(\mathbf{x}) &= \sum_{|\boldsymbol{\alpha}|\le\beta} \frac{1}{\boldsymbol{\alpha}!}\frac{\partial^{\boldsymbol{\alpha}} f(\mathbf{x}_0)}{\partial x^{\boldsymbol{\alpha}}}(\mathbf{x}-\mathbf{x}_0)^{\boldsymbol{\alpha}} + \sum_{|\boldsymbol{\alpha}|=\beta} \frac{1}{\boldsymbol{\alpha}!}\frac{\partial^{\boldsymbol{\alpha}} f(\mathbf{x}_0+\lambda_{\boldsymbol{\alpha}}\mathbf{x})}{\partial x^{\boldsymbol{\alpha}}}(\mathbf{x}-\mathbf{x}_0)^{\boldsymbol{\alpha}} \\ &:= T(\mathbf{x}) + R(\mathbf{x}) \end{aligned}$$

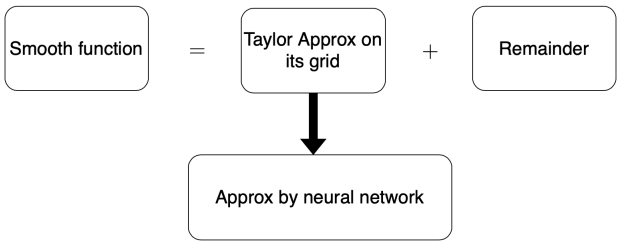

图 3 3。用神经网络逼近光滑函数

其中我们使用标准的多维泰勒级数记号：对于任意 $\boldsymbol{\alpha}=(\boldsymbol{\alpha}_1,\ldots,\boldsymbol{\alpha}_d)$，令 $|\boldsymbol{\alpha}|=\sum_{j=1}^{d}\boldsymbol{\alpha}_j$。令 $\partial^{\boldsymbol{\alpha}}/\partial x^{\boldsymbol{\alpha}}$ 表示 $\partial^{\boldsymbol{\alpha}_1+\cdots+\boldsymbol{\alpha}_d}/\partial x_1^{\boldsymbol{\alpha}_1}\partial x_2^{\boldsymbol{\alpha}_2}\ldots\partial x_d^{\boldsymbol{\alpha}_d}$。对于任意向量 $\mathbf{v}$，令 $\mathbf{v}^{\boldsymbol{\alpha}}$ 表示 $v_1^{\boldsymbol{\alpha}_1}\ldots v_d^{\boldsymbol{\alpha}_d}$。用神经网络逼近光滑函数的基本思路包括以下三个关键步骤：

1. 将定义域 $[0,1]^d$ 划分为小网格。2. 在每个网格上，围绕某个固定点（例如网格角点的中点）进行泰勒级数展开。3. 用神经网络逼近泰勒多项式。

图 3 展示了这三个步骤的流程图。按照 Lu 等人(2021年) 的记号，最佳神经网络逼近器可写为：

$$\phi(x)=\sum_{|\alpha|\leq\beta}\varphi\left(\frac{\phi_\alpha\left(\boldsymbol{\Psi}(x)\right)}{\alpha!},\mathbf{P}_\alpha(x-\boldsymbol{\Psi}(x))\right) \tag{5.1}$$

其中，$\boldsymbol{\Psi}(x)$ 将点 x 映射到其所在网格的角点，ϕ_α 逼近导数的值，P_α 逼近多项式 $\mathbf{x}^{\boldsymbol{\alpha}}$，最后 φ（外层函数）通过 $\phi(x,y)$ 逼近运算 xy。接下来，我们将说明如何用神经网络进行加法。给定 k 个神经网络，每个网络的宽度为 N、深度为 L，可以按以下方式将它们相加：首先将 x 输入到 $2k$ 个神经网络 $\{\phi_1,\phi_2,\ldots,\phi_k\}$ 和 $\{-\phi_1,\ldots,-\phi_k\}$中。然后，最终输出为 $\sum_{j=1}^{k}(\sigma(\phi_j(x))-\sigma(-\phi_j(x)))$。由于 $\sigma(\cdot)$ 是 ReLU，我们有恒等式 $\sigma(x)-\sigma(-x)=x$。关于图示说明，请参见图 2。因此，可以使用宽度为 $2kN$、深度为 $L+1$的神经网络来完成求和。如果 k 很大，则可以通过宽度为 $kN\vee 2k$ 且宽度为 $L+2$的神经网络执行相同的运算。在 (5.1) 中，k 的值为 β^d 阶，这是固定的，因此宽度和深度的阶数保持不变。

如（5.1）所示，神经网络逼近包含五个误差来源（表 5）：i) E_1，在网格上用泰勒逼近 f 所产生的误差；ii) E_2，逼近 $\boldsymbol{\Psi}$(x)的误差；iii) E_3，逼近 ϕ_α的误差；iv) E_4，用 P_α 逼近 $\mathbf{x}^{\boldsymbol{\alpha}}$ 所产生的误差；最后是 v) E_5，用 $\phi(x,y)$逼近 xy 所产生的误差。对于任意给定的 N 和 L，假设我们将 $[0,1]$ 划分为 K 个长度为$1/K$的等间距区间，其中 $K=\lfloor N^{1/d}\rfloor^2\lfloor L^{2/d}\rfloor$。如果我们使用宽度为 $C_1N\log N$ 和宽度为$C_2L\log L$的神经网络，其中 C_1、$C_2>0$为某些给定值，则会产生如下逼近误差：因此，对于 $\beta\geq 1$ 和 $d\geq 1$，唯一受到维度影响的项是泰勒逼近误差，其为 E_1。NN $\boldsymbol{\Psi}$ 是一个阶梯函数，它将 x 映射到其网格的角点，且不会产生任何误差。ϕ_α 是一个点拟合函数，将网格的角点映射到 f 的 $\boldsymbol{\alpha}^{th}$ 阶导数，并产生阶为$(NL)^{-2\beta}$ 的误差，该误差与其维度无关。$\mathrm{P}_{\boldsymbol{\alpha}}$ 是一个逼近 $\mathbf{x}^{\boldsymbol{\alpha}}$ 的神经网络，其误差阶为$(N+1)^{-7\beta L}$；同样，神经网络 φ 逼近二元函数 $(x,y)\rightarrow xy$。

误差	阶数
E_1	$K^{-\beta} \sim (NL)^{-2\beta/d}$
E_2	0
E_3	$2(NL)^{-2\beta}$
E_4	$9\beta(N+1)^{-7\beta L}$
E_5	$216(N+1)^{-7\beta L}$

表 1
神经网络的逼近误差（参见（*Lu* 等人，*2021*年，表 *2*））

步骤 1：（逼近一维分量）对于每个一维分量 f_j，存在一个神经网络 ϕ_j（通过上述构造得到），其宽度为$C_1 N_1 \log N_1$、深度为 $C_2 L_1 \log L_1$，使得对于某个 $C>0$,

$$\|f_j - \phi_j\|_\infty \le C(N_1 L_1)^{-2\beta_1}.$$

定义 $I_j = \int_0^1 \phi_j(x)\,dx$，以及另一个神经网络 $\widetilde{\phi}_j(x) = \phi_j(x) - I_j$。由于减去一个常数等价于改变最后一层的偏差，因此不会改变架构。因此，$\widetilde{\phi}_j$ 的宽度和深度与 ϕ_j相同。此外，我们有：

$$\|f_j - \widetilde{\phi}_j\|_\infty = \left\| f_j - \int_0^1 f_j - \phi_j - \int_0^1 \phi_j \right\|_\infty \le 2\|f_j - \phi_j\|_\infty \le 2C(N_1 L_1)^{-2\beta_1}.$$

步骤 2：（逼近二维分量）现在考虑一个二维函数，例如 f_{12}。和之前一样，根据 Lu 等人（2021年），我们可以构造一个神经网络ϕ_{12}，其宽度为 $O(N_2 \log N_2)$，深度为 $O(L_2 \log L_2)$，使得

$$\|f_{12} - \phi_{12}\|_\infty \le C(N_2 L_2)^{-\beta_2}. \tag{5.2}$$

然而，根据（5.1），ϕ_{12} 并不直接逼近 f_{12}，它只逼近泰勒展开式 $T_{12}(\mathbf{x})$。因此，存在某个 $L^\star(\beta) > 0$，使得对于任意 $L \ge L^\star(\beta)$，

$$\|f_{12}(\mathbf{x}) - \phi_{12}(\mathbf{x})\|_\infty \le \underbrace{\|T_{12}(\mathbf{x}) - \phi_{12}(\mathbf{x})\|_\infty}_{O((N_1 L_1)^{-2\beta_2})} + \underbrace{\|R_{12}(\mathbf{x})\|_\infty}_{O((N_2 L_2)^{-\beta_2})}. \tag{5.3}$$

注意，$T_{12}(\mathbf{x})$ 是一个二元多项式。尽管神经网络的总体逼近误差为 $O((N_2 L_2)^{-\beta_2})$，但 T_{12} 由 ϕ_{12} 得到的逼近误差要快得多，并且当 $L \ge L^\star(\beta_2)$ 时，可以假设其快于 $(N_2 L_2)^{-2\beta_2}$，其中 $L^\star(\beta_2)$ 是仅依赖于 β_2 的某个常数。若将 $T_1(x)$（分别将 $T_2(y)$）定义为 $T_1(x) = \int_0^1 T_{12}(x,y)\,dy$（分别为 $T_2(y) = \int_0^1 T_{12}(x,y)\,dx$），那么 T_1 和 T_2 都是次数为 $\le \beta_2 - 1$ 的一元多项式，因此β_2 光滑。因此，我们可以构造宽度为 $O(N_2 \log N_2)$、深度为 $O(L_2 \log L_2)$ 的神经网络 ξ_1 和 ξ_2，使得：

$$\|T_1(x) - \xi_1(x)\|_\infty \le C(N_2 L_2)^{-2\beta_2}, \tag{5.4}$$

$$\|T_2(y) - \xi_2(y)\|_\infty \le C(N_2 L_2)^{-2\beta_2}. \tag{5.5}$$

我们定义常数 $I_{12} = \iint \phi_{12}(x,y)\,dx\,dy$。最后，我们将估计量定义为

$$\widetilde{\phi}_{12}(x,y) := \phi(x,y) - \xi_1(x) - \xi_2(y) + I_{12}。$$

下面我们讨论 ϕ_{12} 的架构。如前所述，添加 I_{12} 不会改变架构。此外，由于 ϕ_{12}, ξ_1, ξ_2 具有宽度 $O(N_2 \log N_2)$ 和深度 $O(L_2 \log L_2)$，ϕ_{12} 也具有宽度 $O(N_2 \log N_2)$ 和深度 $O(L_2 \log L_2)$（见图 2）。接下来，我们证明 ϕ_{12} 与 f_{12} 的 L_∞ 距离仍然

阶为 $(N_2L_2)^{-\beta_2}$。回顾一下， f_{12} 的边缘分布为 0。利用三角不等式，我们有：

$$\begin{aligned}
\|f_{12}-\widetilde{\phi}_{12}\|_\infty &= \|f_{12}-\phi_{12}+\xi_1+\xi_2-I_{12}\|_\infty \\
&\le \|f_{12}-\phi_{12}\|_\infty + \left\|\int_0^1 f_{12}\,dy-\xi_1\right\|_\infty \\
&\qquad + \left\|\int_0^1 f_{12}\,dx-\xi_2\right\|_\infty + \left\|\iint_{[0,1]^2} f_{12}\,dx\,dy-I_{12}\right\|_\infty \\
&\le 2\|f_{12}-\phi_{12}\|_\infty + \left\|T_1+\int_0^1 R_{12}\,dy-\xi_1\right\|_\infty \\
&\qquad + \left\|T_2+\int_0^1 R_{12}\,dx-\xi_2\right\|_\infty \quad \left[\text{As } I_{12}=\iint \phi_{12}\right] \\
&\le 2\|f_{12}-\phi_{12}\|_\infty + \|T_1-\xi_1\|_\infty + \|T_2-\xi_2\|_\infty \\
&\qquad + \left\|\int_0^1 R_{12}\,dy\right\|_\infty + \left\|\int_0^1 R_{12}\,dx\right\|_\infty \\
&\le 2\|f_{12}-\phi_{12}\|_\infty + \|T_1-\xi_1\|_\infty + \|T_2-\xi_2\|_\infty + 2\|R_{12}\|_\infty \\
&\le C(N_2L_2)^{-\beta_2} + 2C(N_2L_2)^{-2\beta_2} + 2C(N_2L_2)^{-\beta_2} = 5C(N_2L_2)^{-\beta_2}.
\end{aligned}$$

这里，第一个误差项由方程（5.2）得出，第二个和第三个误差项由方程（5.4）得出，最后一个误差项由方程（5.3）得出。
接下来我们证明 $\left\|\int_0^1\left(\widetilde{\phi}_{12}-f_{12}\right)\,dy\right\|_\infty \lesssim (NL)^{-2\beta}$。由于 $\int_0^1 f_{12}\,dy=0$，该结论成立，因为

$$\begin{aligned}
\left\|\int_0^1 \widetilde{\phi}_{12}\,dy\right\|_\infty &= \left\|\int_0^1 \phi_{12}\,dy-\xi_1-\int_0^1 \xi_2\,dy+I_{12}\right\|_\infty \\
&\le \left\|\int_0^1 (\phi_{12}-T_{12})\,dy+T_1-\xi_1\right\|_\infty \\
&\qquad + \left\|\int_0^1 (\xi_2-T_2)\,dy+\int_0^1 T_2\,dy-I_{12}\right\|_\infty \\
&\le \|\phi_{12}-T_{12}\|_\infty + \|\xi_1-T_1\|_\infty + \|\xi_2-T_2\|_\infty \\
&\qquad + \left\|\iint_{[0,1]^2} T_{12}\,dx\,dy-\iint_{[0,1]^2}\phi\,dx\,dy\right\|_\infty \\
&\le 2\|\phi_{12}-T_{12}\|_\infty + \|\xi_1-T_1\|_\infty + \|\xi_2-T_2\|_\infty \le 4C(N_2L_2)^{-2\beta_2}
\end{aligned}$$

这里，第一个不等式由方程（5.3）得出，第二个和第三个不等式由（5.4）得出。

步骤 3：（合并所有界）在此，我们合并一维~和二维分量的界，以证明定理 2.7。定义函数 ϕ 如下：

$$\widetilde{\phi}=\sum_{j=1}^{d}\widetilde{\phi}_j+\sum_{k<l}\widetilde{\phi}_{kl}.$$

˜回到我们对多个神经网络求和的论证， $\sum_j \phi_j$ 的宽度为∼$O(dN_1$ 对数 $N_1)$，深度为 $O(L_1$ 对数 $L_1)$；$\sum_{k<l} \phi_{kl}$ 的宽度为 $O((d(d-1)/2)N_2$ 对数 $N_2)$，深度为 $O(L_2$ 对数 $L_2)$。此外，我们有：

$$
\begin{aligned}
&\mathbb{E}\left[\left(f_0(X)-\widetilde{\phi}(X)\right)^2\right] \le p_{\max}\int_{[0,1]^d}\left(f_0(\mathbf{x})-\widetilde{\phi}(\mathbf{x})\right)^2 d\mathbf{x}\\
&= p_{\max}\int_{[0,1]^d}\left(\sum_j f_{0,j}(x_j)+\sum_{i<j} f_{0,ij}(x_i,x_j)-\sum_j \widetilde{\phi}_j(x_j)-\sum_{i<j}\widetilde{\phi}_{ij}(x_i,x_j)\right)^2 d\mathbf{x}\\
&\le 2p_{\max}\int_{[0,1]^d}\left(\sum_j f_{0,j}(x_j)-\sum_j \widetilde{\phi}_j(x_j)\right)^2 d\mathbf{x}+2p_{\max}\int_{[0,1]^d}\left(\sum_{i<j} f_{0,ij}(x_i,x_j)-\sum_{i<j}\widetilde{\phi}_{ij}(x_i,x_j)\right)^2 d\mathbf{x}\\
&= 2p_{\max}\sum_{j=1}^d\int_0^1\left(f_{0,j}(x_j)-\widetilde{\phi}_j(x_j)\right)^2 dx_j+2p_{\max}\sum_{i<j}\int_{[0,1]^2}\left(f_{0,ij}(x_i,x_j)-\widetilde{\phi}_{ij}(x_i,x_j)\right)^2 dx_i\,dx_j\\
&+2p_{\max}\sum_{j\neq j'}\int_0^1\left(f_{0,j}(x_j)-\widetilde{\phi}_j(x_j)\right)dx_j\int_0^1\left(f_{0,j'}(x_{j'})-\widetilde{\phi}_{j'}(x_{j'})\right)^2 dx_{j'}\\
&+2p_{\max}\int_{[0,1]^4}\sum_{(i,j)\neq(k,l)}\left(f_{0,ij}(x_i,x_j)-\widetilde{\phi}_{ij}(x_i,x_j)\right)\left(f_{0,kl}(x_k,x_l)-\widetilde{\phi}_{kl}(x_k,x_l)\right)dx_i dx_j dx_k dx_l\\
&\le 2p_{\max}\left[\sum_{j=1}^d\|f_{0,j}-\widetilde{\phi}_j\|_\infty^2+\sum_{i<j}\|f_{0,ij}-\widetilde{\phi}_{ij}\|_\infty^2\right]\\
&\quad+2p_{\max}\sum_{\substack{i\neq(k,l)\\ j\neq(k,l)}}\int_{[0,1]^2}\widetilde{\phi}_{ij}\,dx_i\,dx_j\int_{[0,1]^2}\widetilde{\phi}_{kl}\,dx_k\,dx_l+\\
&\quad+2p_{\max}\sum_{i,j,k}\int_0^1\left(\int_0^1\widetilde{\phi}_{ij}\,dx_j\right)\left(\int_0^1\widetilde{\phi}_{ik}\,dx_k\right)dx_i\\
&\le 2C^2p_{\max}\left[d(N_1L_1)^{-4\beta_1}+\binom{d}{2}(N_2L_2)^{-2\beta}+\binom{d}{4}(N_2L_2)^{-4\beta}+\binom{d}{3}(N_2L_2)^{-4\beta}\right]\\
&\le C_3\left(d(N_1L_1)^{-4\beta_1}+\binom{d}{2}(N_2L_2)^{-2\beta_2}\right),
\end{aligned}
$$

对于某些 C，$C_3>0$ 由于 $d^2(N_2L_2)^{-2\beta_2}\to 0$ 根据假设 2.6。这里倒数第二个不等式由结合步骤 1 和步骤 2 的结果得到。证明至此完成。

资助。本文得到 ONR N00014-22-1-2340 以及 NSF 基金 DMS-2052926、DMS-2053832 和 DMS-2210833 的资助。

参考文献

ANTONIADIS, A. 和 FAN, J.（2001）。小波逼近的正则化。《美国统计学会杂志》**96** 939–967。
BARRON, A. R.（1993）。S 形函数叠加的通用逼近界。*IEEE*信息论汇刊**39** 930–945。

P. L. Bartlett、N. Harvey、C. Liaw 和 A. Mehrabian（2019）。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 2285 2301–。B. Bauer 和 M. Kohler（2019）。深度学习如何应对非参数回归中的维数灾难。《统计年刊》47 2261–2285。Belloni, A. 和 Chernozhukov, V.（2013）。在高维稀疏模型中进行模型选择后的最小二乘法。《伯努利》19 521–547。P. J. Bickel、Y. Ritov 和 A. B. Tsybakov（2009）。LASSO 和 Dantzig 选择器的同时分析。P. J. Bickel、Y. Ritov、A. B. Tsybakov 等（2010）。稀疏高维回归中的变量层次选择。IMS 文集 6 28。J. Bien、J. Taylor 和 R. Tibshirani（2013）。用于层次交互的 LASSO。《统计年刊》41 1111。Breiman（2001）。随机森林。《机器学习》45 5–32。A. Buja、T. Hastie 和 R. Tibshirani（1989）。线性平滑器与加性模型。《统计年刊》453–510。E. Candès 和 T. Tao（2007）。Dantzig 选择器：当 p 远大于 n 时的统计估计。《统计年刊》35 2313–2351。Chernozhukov, V.、Chetverikov, D. 和 Kato, K.（2014）。经验过程上确界的高斯近似。《统计年刊》42 1564–1597。J. Fan 和 I. Gijbels（1996）。局部多项式建模及其应用。Chapman & Hall。J. Fan 和 Y. Gu（2022）。用于高维回归的因子增强稀疏吞吐深度 ReLU 神经网络。arXiv 预印本 arXiv:2210.02002。J. Fan、Y. Gu 和 W.-X. Zhou（2022）。噪声尾部如何影响深度 ReLU 网络？arXiv 预印本 arXiv:2203.10418。J. Fan、W. Härdle 和 E. Mammen（1998）。加性模型中低维分量的直接估计。《统计年刊》26 943–971。J. Fan、R. Li、C.-H. Zhang 和 H. Zou（2020）。数据科学的统计基础。Chapman and Hall/CRC。J. Fan 和 J. Lv（2008）。超高维特征空间的确定性独立筛选。《皇家统计学会杂志：B辑（统计方法学）》70 849–911。J. Fan 和 R. Song（2010）。具有 np 维数的广义线性模型中的确定性独立筛选。《统计年刊》38 3567–3604。Y. Freund、R. E. Schapire 等（1996）。一种新提升算法的实验。载于国际机器学习会议，第 96 卷。Citeseer。J. H. Friedman 和 W. Stuetzle（1981）。投影寻踪回归。《美国统计学会杂志》76 817–823。I. Goodfellow、Y. Bengio 和 A. Courville（2016）。深度学习。麻省理工学院出版社。Q. Han 和 J. A. Wellner（2019）。重尾误差下最小二乘回归估计量的收敛速率。《统计年刊》47 2286–2319。N. Hao 和 H. H. Zhang（2014）。超高维数据的交互筛选。《美国统计学会杂志》109 1285–1301。W. Härdle、P. Hall 和 H. Ichimura（1993）。单指标模型中的最优平滑。《统计年刊》21 157–178。K. Hornik、M. Stinchcombe 和 H. White（1989）。多层前馈网络是通用逼近器。《神经网络》2 359–366。J. L. Horowitz 和 E. Mammen（2007）。具有未知连接函数的一般非参数回归模型的速率最优估计。《统计年刊》35 2589–2619。M. Kohler 和 KRZY ́ZAK, A.（2005）。使用多层前馈神经网络进行自适应回归估计。《非参数统计》17 891–913。M. Kohler 和 KRZY ́ZAK, A.（2016）。基于层次交互模型的非参数回归。IEEE信息论汇刊 63 1620–1630。M. Kohler 和 S. Langer（2021）。全连接深度神经网络回归估计量的收敛速率。《统计年刊》49 2231–2249。V. Koltchinskii 和 M. Yuan（2010 38 3660 3695）。多核学习中的稀疏性。《统计年刊》–。M. R. Kosorok（2008）。经验过程与半参数推断导论。施普林格。S. Kpotufe 和 S. Dasgupta（2012）。一种可适应内在维数的基于树的回归器。《计算机与系统科学杂志》78 1496–1515。A. Krizhevsky、G. Hinton 等（2009）。从微小图像中学习多层特征。

LECUN, Y.（1998）。手写数字的 MNIST 数据库。*http://yann. lecun. com/exdb/mnist/* 。LIN, Y. 和 ZHANG, H. H.（2006）。多元非参数回归中的分量选择与平滑。《统计年刊》**34** 2272–2297。LU, J., SHEN, Z., YANG, H. 和 ZHANG, S.（2021）。光滑函数的深度网络逼近。《*SIAM* 数学分析杂志》**53** 5465–5506。MHASKAR, H. N.（1996）。用于光滑函数和解析函数最优逼近的神经网络。《神经计算》**8** 164–177。MUKHERJEE, D., BANERJEE, M. 和 Y. Ritov（2021）。增长维数下离散选择模型的最优线性判别器。《统计年刊》**49** 3324–3357。NOVEMBRE, J., JOHNSON, T., BRYC, K., KUTALIK, Z., BOYKO, A. R., AUTON, A., INDAP, A., KING, K. S., BERGMANN, S., NELSON, M. R. 等（2008年）。基因映照欧洲的地理分布。《自然》**456** 98–101。RASKUTTI, G., J WAINWRIGHT, M. 和 YU, B.（2012）。通过凸规划求解核类上稀疏加性模型的极小极大最优速率。《机器学习研究杂志》**13**。RAVIKUMAR, P., LAFFERTY, J., LIU, H. 和 WASSERMAN, L.（2009）。稀疏加性模型。《皇家统计学会杂志：*B*辑（统计方法学）》**71** 1009–1030。SCHMIDT-HIEBER, J.（2020）。使用具有 ReLU 激活函数的深度神经网络进行非参数回归。《统计年刊》**48** 1875–1897。STONE, C. J.（1982）。非参数回归的最优全局收敛速率。《统计年刊》1040–1053。STONE, C. J.（1985）。加性回归及其他非参数模型。《统计年刊》**13** 689–705。STONE, C. J.（1994）。多元函数估计中多项式样条及其张量积的应用。《统计年刊》**22** 118–171。TAN, Z. 和 C.-H. Zhang（2019）。高维数据加性回归中的双重惩罚估计。《统计年刊》**47** 2567–2600。A. B. Tsybakov（2004）。非参数估计导论，2009。*URL https://doi. org/10.1007/b13794.* 修订并扩展自 **9**。VAART, A. V. D. 和 J. A. Wellner（1997）。弱收敛与经验过程及其在统计学中的应用。《皇家统计学会杂志：*A*辑（社会统计学）》**160** 596–608。VAN DER VAART, A. W. 和 WELLNER, J. A.（1996）。弱收敛。载于《弱收敛与经验过程》。施普林格，16–28。YANG, Y. 和 Dunson, D. B.（2016）。贝叶斯流形回归。《统计年刊》**44** 876–905。YANG, Y. 和 Tokdar, S. T.（2015）。高维下极小极大最优的非参数回归。《统计年刊》**43** 652–674。YAROTSKY, D.（2017）。深度 ReLU 网络逼近的误差界。神经网络**94** 103–114。YUAN, M. 和 LIN, Y.（2006）。含分组变量的回归中的模型选择与估计。《皇家统计学会杂志：*B*辑（统计方法学）》**68** 49–67。YUAN, M. 和 ZHOU, D.-X.（2016）。高维加性模型估计的极小极大最优速率。《统计年刊》**44** 2564–2593。ZHANG, Y., WAINWRIGHT, M. J. 和 JORDAN, M. I.（2014）。稀疏线性回归中多项式时间算法性能的下界。载于学习理论会议。PMLR。ZHAO, P., ROCHA, G. 和 YU, B.（2009）。用于分组变量选择和分层变量选择的复合绝对惩罚族。

附录 A：证明

A.1. 定理 2.9 的证明。我们希望应用（Vaart 和 Wellner，1997，定理 3.2.5）。对于任意 $\phi=\phi_1+\phi_2$（其中 $\phi_1\in\mathcal{F}_{NN}^1,\phi_2\in\mathcal{F}_{NN}^2$），定义总体风险函数 $R(\phi)$（相应地，经验风险 $R_n(\phi)$）为

$$R(\phi)=\mathbb{E}[(Y-\phi(X))^2]\quad\left(\text{resp. }\hat{R}(\phi)=\frac{1}{n}\sum_i(Y_i-\phi(X_i))^2\right)$$

由于 ϵ 与 X 独立，且均值为0，因此显然有：

$$R(\phi)-R(f_0)=\mathbb{E}[(f(X)-f_0(X))^2]\triangleq d^2(\phi,f_0)\,.$$

定义 $\mathcal{F}_n:=\{\phi=\phi_1+\phi_2\colon\phi_1\in\mathcal{F}_{NN}^1,\phi_2\in\mathcal{F}_{NN}^2\}$（即 $\mathcal{F}_n=\mathcal{F}_{NN}^1+\mathcal{F}_{NN}^2$）以及 $\omega_n=d(\phi^\star,f_0)$，即神经网络函数类的逼近误差。显然有：

$$\mathcal{F}_n\subseteq\mathcal{F}_{NN}\left(d,c_1\left(dN_1\log N_1+\tbinom{d}{2}N_2\log N_2\right),c_2,W_n,1\right)$$

with

$$W_n=c_3\left(d(N_1\log N_1)^2+\binom{d}{2}(N_2\log N_2)^2\right),$$

其中某个常数 c_1,c_2,c_3 与 (n,d) 无关。因此，$\mathcal{F}_n$ 的 VC 维数小于或等于 $V_n\triangleq c_4W_n\log W_n$，其中 $c_4>0$ 为某个常数。我们将在后续分析中使用这一结果。设 $\zeta_n(\delta)$ 满足：

$$\mathbb{E}\left[\sup_{\phi\in\mathcal{F}_n:d(\phi,f_0)\le\delta}|(R_n(\phi)-R(\phi))-(R_n(f_0)-R(f_0))|\right]\lesssim\frac{\zeta_n(\delta)}{\sqrt{n}}\,.\tag{A.1}$$

这个函数 ζ_n 称为连续模，用于界定经验过程在真函数邻域内相对于总体过程的波动。$\zeta_n(\delta)$ 本质上取决于底层函数类的复杂度。我们通过以下极大不等式（例如，Chernozhukov 等人 (2014）中的定理 5.2）量化这种依赖关系。为方便读者，现将其陈述如下：

引理 A.1（Chernozhukov 等人（2014）中的定理 5.2）。定义算子 $\mathbb{G}_n=\sqrt{n}(\mathbb{P}_n-P)$。考虑函数集合 $\mathcal{F}$，其包络函数为 F。假设 $F\in L_2(P)$。则：

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}}|\mathbb{G}_n(f)|\right]\lesssim J(r,\mathcal{F},F)\|F\|_2+\frac{\|M\|_2J^2(r,\mathcal{F},F)}{r^2\sqrt{n}}$$

对于 $r=\sigma/\|F\|_2$，其中

$$J(\tau,F,\mathcal{F})=\int_0^\tau\sup_Q\sqrt{1+\log N\left(\epsilon\|F\|_2,\mathcal{F},L_2(Q)\right)}d\epsilon$$

$$M=\max_{1\le i\le n}F(X_i),\qquad\sigma^2=\sup_{f\in\mathcal{F}}Pf^2\,.$$

我们使用引理 A.1 来求函数 $\zeta_n(\delta)$。对于任意 ϕ，在 (X,ϵ) 空间上定义函数 $g\equiv g(\phi)$ 如下：

$$\begin{aligned}g(X,\epsilon)&=(f_0(X)-\phi(X)+\epsilon)^2-\epsilon^2\\&=(\phi(X)-f_0(X))^2+2(\phi(X)-f(X))\epsilon.\end{aligned}\tag{A.2}$$

将 $\mathcal{G}$ 的集合定义为 $\mathcal{G}=\{g(\phi)\colon \phi\in\mathcal{F}_n\}$。此定义意味着：

$$\mathbb{E}\left[\sup_{\phi:d(\phi,f_0)\le\delta}|(R_n(\phi)-R(\phi))-(R_n(f_0)-R(f_0))|\right]=\frac{1}{\sqrt{n}}\mathbb{E}\left[\sup_{g\in\mathcal{G}_\delta}|\mathbb{G}_n(g)|\right].$$

在此，我们将 $\mathcal{G}_\delta$ 定义为所有 $g(\phi)$的集合，其中 $d(\phi,f_0)\le\delta$。为应用引理A.1，我们需要：i）量化 σ^2，ii）求出 $\mathcal{G}_\delta$ 的包络函数，以及 iii）界定覆盖数的对数。对于 i），注意到：

$$\sup_{g\in\mathcal{G}:d(f,f_0)\le\delta}\mathbb{E}[g(X)^2]\le\delta^2(2+8\sigma_\epsilon^2)\triangleq C_{\sigma_\epsilon}\delta^2.$$

接下来，为构造包络函数，注意到（A.2）蕴含：

$$|g(X,\epsilon)|\le 9B^2+6B\epsilon.$$

对所有(X,ϵ)均成立。这是因为 $\|\phi\|_\infty\le 2B$ 和 $\|f_0\|_\infty\le B$。因此，我们可以将包络函数 G 取为 $G(X,\epsilon)=9B^2+6B\epsilon$。为界定 $\mathcal{G}_\delta$的覆盖数对数，我们使用Kosorok（2008年）的引理 9.9。注意，每个 $g=g_1+g_2+hg_3$，其中 $g_1=(\phi-f_0)_+^2$、$g_2=(\phi-f_0)_-^2$、$g_3=2(\phi-f_0)$ 以及 $h(x,\epsilon)=\epsilon$。我们计算每类函数的 VC 维数：

- $\mathcal{G}_1=\{g_1(\phi)\colon \phi\in\mathcal{F}_{NN}\}$ 的 VC 维数为 $\le V_n$。为证明这一点，写成 $g_1(x,\epsilon)=m_1\circ(\phi-f_0)\circ m_2(x,\epsilon)$，其中 $m_2(x,\epsilon)=x$ 且 $m_1(x)=x_+^2$。由于与固定函数 ϕ复合、减去固定函数 f_0 以及与单调函数复合都不会改变 VC 维数，因此结论成立。类似的论证可以证明，$\mathcal{G}_2=\{g_2(f)\colon f\in\mathcal{F}\}$ 的 VC 维数为 $\le V_n$。
- $\mathcal{G}_3$（g_3h 的集合）的 VC 维数为 $\lesssim V_n$。为证明这一点，注意到g_3 的 VC 维数为 $\le V_n$。如前所述，$g_3=2(\phi-f_0)\circ\psi$，因为我们已经提到，与 ψ复合、减去 f_0 以及乘以一个固定函数（此处为 $2h$）都不会改变 VC 维数。

因此，我们已经证明 $\mathcal{G}_i$ 的 VC 维数为 $\lesssim V_n$。根据 Haussler 界（Van Der Vaart 和 Wellner，1996，定理 2.6.7）：

$$\sup_Q\log N(\epsilon\|G_i\|_2,\mathcal{G}_i,L_2(Q))\lesssim\log KV_n+2V_n\log\left(\frac{2\sqrt{2e}}{\epsilon}\right).$$

由于 $\mathcal{G}_\delta\subseteq\mathcal{G}_1+\mathcal{G}_2+\mathcal{G}_3$，我们有：

$$\begin{aligned}\log N(\epsilon\|G\|_2,\mathcal{G}_\delta,L_2(Q))&\lesssim\sum_{j=1}^3\log N(\epsilon\|G_i\|_2,\mathcal{G}_i,L_2(Q))\\&\le 3\log KV_n+6V_n\log\left(\frac{2\sqrt{2e}}{\epsilon}\right)\\&\le 9V_n\log\left(\frac{2\sqrt{2e}}{\epsilon}\right),\end{aligned}$$

只要$\log KV_n\le V_n$ 和 $\epsilon\le 2\sqrt{2/e}$，最后一个不等式就成立。因此，对于所有足够小的 ϵ 以及所有足够大的 V_n，我们都可以使用最后一个不等式。令 $r=\sqrt{C_{\sigma_\epsilon}}\delta/\|G\|_2$，得到：

$$J(r,\mathcal{G}_\delta,G)\le\int_0^r\sqrt{1+9V_n\log\left(\frac{2\sqrt{2e}}{\epsilon}\right)}\,d\epsilon$$

$$\leq 4\sqrt{V_n}\int_0^r \sqrt{\log\left(\frac{2\sqrt{2e}}{\epsilon}\right)}\,d\epsilon$$

$$= 8\sqrt{2eV_n}\int_0^{r/2\sqrt{2e}} \sqrt{\log\left(\frac{1}{\epsilon}\right)}\,d\epsilon$$

$$\leq 4r\sqrt{V_n}\sqrt{\log\left(\frac{2\sqrt{2e}}{r}\right)}.$$

这里，只要 $V_n \log(2\sqrt{2}/\epsilon) \geq 1/7$，即 $\epsilon < 2\sqrt{2}\exp(-1/(7V_n))$，第二个不等式就成立。对于所有足够大的 n，只要 $r < 1$即可。因此，根据引理A.1，我们得出：

$$\mathbb{E}\left[\sup_{g\in\mathcal{G}_\delta}|\mathbb{G}_n(g)|\right] \lesssim \delta\sqrt{V_n\log\left(\frac{1}{\delta}\right)} + \frac{V_n}{\sqrt{n}}\log\left(\frac{1}{\delta}\right)\sqrt{\mathbb{E}\left[\max_i(1+2|\epsilon_i|)^2\right]}$$

$$\lesssim \delta\sqrt{V_n\log\left(\frac{1}{\delta}\right)} + \frac{V_n\sqrt{\log n}}{\sqrt{n}}\log\left(\frac{1}{\delta}\right) \triangleq \phi_n(\delta)\,.$$

现在我们建立收敛速率。令 $r_n = \left(\sqrt{\frac{V_n}{n}\log\frac{n}{V_n}}\sqrt{\log n} \pm \omega_n\right)^{-1}$，其中$\omega_n = d(\phi^\star,\ f_0)$是前文定义的逼近误差。注意到 r_n 满足方程$\phi_n(r_n^{-1}) \lesssim r_n^{-2}\sqrt{n}$，因为

$$r_n^2\phi_n\left(\frac{1}{r_n}\right) \lesssim r_n\sqrt{V_n\log r_n} + r_n^2\frac{V_n\sqrt{\log n}}{\sqrt{n}}\log r_n$$

$$\leq \sqrt{n}\left[\left(\frac{1}{2}\frac{\log\frac{n}{V_n} - \log\log\frac{n}{V_n} - \frac{1}{2}\log\log n}{\log\frac{n}{V_n}\sqrt{\log n}}\right)^{1/2} + \left(\frac{1}{2}\frac{\log\frac{n}{V_n} - \log\log\frac{n}{V_n} - \frac{1}{2}\log\log n}{\log\frac{n}{V_n}}\right)\right]$$

$$\lesssim \sqrt{n}$$

现在我们使用分层论证来建立收敛速率。固定 $t > 1$，并定义$A_j := \{\mathcal{F}_{NN}: 2^{j-1}t \leq r_n d(f_0,\ \phi) \leq 2^j t\}$。根据经验风险最小化的定义，我们有

$$\mathbb{P}\left(r_n d(\hat{\phi}, f_0) \geq t\right) = \mathbb{P}\left(\sup_{\substack{\phi\in\mathcal{F}_{NN}\\ r_n d(f_0,\phi)\geq t}} \hat{R}_n(\phi^\star) - \hat{R}_n(\phi) \geq 0\right)$$

$$\leq \sum_{j=1}^{\infty}\mathbb{P}\left(\sup_{\phi\in A_j}\hat{R}_n(\phi^\star) - \hat{R}_n(\phi) \geq 0\right)$$

$$\leq \sum_{j=1}^{\infty}\mathbb{P}\left(\sup_{\phi\in A_j}(R_n(\phi^\star) - R(\phi^\star)) - (R_n(\phi) - R(\phi)) \geq \inf_{\phi\in A_j}R(\phi) - R(\phi^\star)\right)$$

$$\leq \sum_{j=1}^{\infty}\frac{\mathbb{E}\left[\sup_{\phi\in A_j}|(R_n(\phi^\star) - R(\phi^\star)) - (R_n(\phi) - R(\phi))|\right]}{\inf_{\phi\in A_j}R(\phi) - R(\phi^\star)} \tag{A.3}$$

我们分别对分子和分母进行界定。对于分母，注意：

$$\inf_{f\in A_j}R(\phi) - R(\phi^\star) \geq \inf_{f: d_{f,f_0}\geq 2^{j-1}tr_n^{-1}}R(\phi) - R(f_0) + R(f_0) - R(\phi^\star)$$

$$\geq 2^{2j-2}t^2 r_n^{-2} - \omega_n^2$$

对于分子：

$$\mathbb{E}\left[\sup_{\phi\in A_j} |(R_n(\phi^\star) - R(\phi^\star)) - (R_n(\phi) - R(\phi))|\right]$$

$$\leq \mathbb{E}\left[\sup_{f:d(f,f_0)\leq 2^j t r_n^{-1}} |(R_n(\phi^\star) - R(\phi^\star)) - (R_n(\phi) - R(\phi))|\right] \lesssim \frac{\zeta_n\left(2^j t r_n^{-1}\right)}{\sqrt{n}}.$$

此处，倒数第二个等式成立，是因为 $d(\phi^\star,\ f_0) = \omega_n \leq 2^j t r_n^{-1}$因为 $r_n^{-1} \geq \omega_n$。将此界代入方程 (A.3)，得到：

$$\begin{aligned}
\mathbb{P}\left(r_n d(\hat{f}, f_0) \geq t\right) &\leq \sum_{j=1}^{\infty} \frac{\phi_n\left(2^j t r_n^{-1}\right)}{\sqrt{n}\left(2^{2j-2}t^2 r_n^{-2} - \omega_n^2\right)} \\
&\leq \sum_{j=1}^{\infty} \frac{2^j t \phi_n\left(r_n^{-1}\right)}{\sqrt{n}\left(2^{2j-2}t^2 r_n^{-2} - \omega_n^2\right)} = \sum_{j=1}^{\infty} \frac{2^j t r_n^2 \phi_n(r_n^{-1}) n^{-1/2}}{2^{2j-2}t^2 - \omega_n^2 r_n^2} \\
&\lesssim \sum_{j=1}^{\infty} \frac{2^j t}{2^{2j-2}t^2 - 1} \qquad [\text{since } r_n^2\phi_n(r_n^{-1}) \lesssim \sqrt{n},\ \ \omega_n^2 r_n^2 \leq 1] \\
&= \frac{1}{t}\sum_{j=1}^{\infty} \frac{2^j}{2^{2j-2} - \frac{1}{t^2}} \leq \frac{c}{t}
\end{aligned}$$

这就证明了：

$$r_n\, d\left(\hat{\phi}, f_0\right) = O_p(1).$$

i.e.

$$\|\hat{f} - f_0\|^2_{L_2(P_X)} = O_p\left(\omega_n^2 + \frac{V_n}{n}\log^{3/2} n\right).$$

现在，为了平衡 ω_n 和 V_n，我们选择 $N_1 = \lfloor n^{1/2(2\beta_1+1)}\rfloor$ 和 $N_2 = \lfloor n^{1/2(\beta_2+1)}\rfloor$。这意味着：

$$\begin{aligned}
\omega_n^2 &\lesssim d n^{-\frac{2\beta_1}{2\beta_1+1}} + \binom{d}{2} n^{-\frac{\beta_2}{\beta_2+1}} \\
\frac{V_n}{n} &\lesssim \frac{W}{n}\log W \lesssim \left(d n^{-\frac{2\beta_1}{2\beta_1+1}} + \binom{d}{2} n^{-\frac{\beta_2}{\beta_2+1}}\right)\log^3 n,
\end{aligned}$$

由此可得

$$\|\hat{f} - f_0\|^2_{L_2(P_X)} = O_p\left(\left(d n^{-\frac{2\beta_1}{2\beta_1+1}} + \binom{d}{2} n^{-\frac{\beta_2}{\beta_2+1}}\right)\log^{4.5} n\right).$$

证明完毕。

A.2. 定理2.12的证明。 这里，我们建立以下双向交互模型的下界：

$$Y_i = \mu + \sum_{j=1}^{d} f_j(X_{ij}) + \sum_{k<l} f_{kl}(X_{ik}, X_{il}) + \epsilon_i \triangleq f_0(X_i) + \epsilon_i.$$

我们证明:

$$\inf_{\hat{f}} \sup_{f\in\mathcal{F}, X\sim P_X} \mathbb{E}_{P_X}[(\hat{f}(X)-f(X))^2] \geq c\left(dn^{-\frac{2\beta_1}{2\beta_1+1}} + \binom{d}{2} n^{-\frac{\beta_2}{\beta_2+1}}\right).$$

其中, $\mathcal{F}$ 是所有满足以下条件的加性函数的集合: I) 一维分量是 β_1光滑的, $[0,1]$ 且积分为 0; II) 二维分量是 β_2光滑的, 支撑集为 $[0,1]^2$ 且边际函数为 0。我们的证明基于 (Tsybakov, 2004, 第 2 节 2.6.1) 中介绍的技术。在整个证明过程中, 假设 $X_{ij} \overset{\text{iid}}{\sim} Unif(0,1)$, $i \leq n$, $j \leq d$且 $\varepsilon_i \overset{\text{iid}}{\sim} N(0,1)$, $i \leq n$。关键思路是仔细选取 $\mathcal{F}$的一个子集, 并应用 Fano 不等式。遵循 Tsybakov (2004), 固定以下记号:

$$m_i = \lfloor c_0 n^{\frac{1}{2\beta_i+1}} \rfloor, \quad h_i = m_i^{-1}, \quad i=1,2. \tag{A.4}$$

我们分两步构造备择函数, 分别针对一维分量和二维分量。

一维分量的备择函数: 首先将 $[0,1]$ 划分为 m_1 个网格$[(k-1)/m_1, k/m_1]$, $1 \leq k \leq m_1$。对于每个 k, 定义 $x_k = (k-0.5)/m_1$。选取一个支撑集为 $(-1/2, 1/2)$ 的光滑核$K \in C^\infty(\mathbb{R})$, $\int K(u)\ du = 0$, 并定义$\int K^2(u)\ du \triangleq \|K\|_2^2$。定义函数 $\phi_k(x)$, $k \leq m_1$ 如下:

$$\phi_k(x) := Lh_1^{\beta_1} K\left(\frac{x-x_k}{h_1}\right) \tag{A.5}$$

常数 L 的选取使得 ϕ_k 具有有界导数并满足假设 2.3。对于每个 k, 函数 ϕ_k 的支撑为 $[(k-1)/m,\ k/m]$, 且 ϕ_k 的积分为 0, 因为

$$\begin{aligned}
\int \phi_k(x)\,dx &= Lh_1^{\beta_1} \int K\left(\frac{x-x_k}{h_1}\right) dx \\
&= Lh_1^{\beta_1} \int_{x_k-\frac{h}{2}}^{x_k+\frac{h}{2}} K\left(\frac{x-x_k}{h_1}\right) dx \quad \text{[Since } K(z)>0 \text{ when } |z|\geq 1/2\text{]} \\
&= Lh_1^{\beta_1+1} \int_{-\frac{1}{2}}^{\frac{1}{2}} K(z)\,dz = 0\,.
\end{aligned}$$

对于 $j \neq k$, ϕ_j和 ϕ_k的支撑互不相交, 因此任意 ϕ_j与 ϕ_k之间的 $L_2([0,1])$ 内积均为 0。设 $M_1 := dm_1$, 并设 $\Omega_1 := \{0,1\}^{M_1}$。对于每个 $\boldsymbol{\omega}_1 \in \Omega_1$,, 将其各项定义为 $\{\boldsymbol{\omega}_{1,jk}\}_{1\leq j\leq d,\ 1\leq k\leq m_1}$。定义函数 $f_{(1)\omega_1}$如下:

$$f^{(1)}_{\boldsymbol{\omega}_1}(X) := \sum_{j=1}^{d}\sum_{k=1}^{m} \boldsymbol{\omega}_{1,jk}\phi_k(X_j)\,,$$

这将构成一维备择函数。

二维分量的备择函数: 二维备择函数的构造与一维分量的构造类似。和之前一样, 将 $[0,1]^2$划分为 m_2^2个矩形 $[(k-1)/m_2, k/m_2] \times [(l-1)/m_2, l/m_2]$, $1 \leq k,l \leq m_2$。设 $x_k := (k-0.5)/m_2$。按照之前相同的方式, 定义 ϕ_{kl}, $1 \leq k,l \leq m_2$ 为

$$\phi_{kl}(x,y) := Lh_2^{\beta_2} K\left(\frac{x-x_k}{h}\right) K\left(\frac{y-x_l}{h}\right) \tag{A.6}$$

注意，每个 ϕ_{kl} 的支撑集均包含于 $[(k-1)/m_2, k/m_2] \times [(l-1)/m_2, l/m_2]$。因此， $L_2([0,1]^2)$ 中任意两个 ϕ_{k_1,l_1} 和 ϕ_{k_2,l_2} 的内积在 $(k_1,l_1) \neq (k_2,l_2)$ 时为 0。此外， ϕ_{kl} 的边缘分布为 0，因为 $\int K((x-x_k)/h)\ dx = 0$。设 $M_2 := d(d-1)m_2^2/2$ 且 $\Omega_2 := \{0,1\}^{M_2}$。对于每个 $\boldsymbol{\omega}_2 \in \Omega_2$，将其写成张量形式 $\boldsymbol{\omega}_{2,ijkl}$，其中 $1 \le i < j \le d$ 且 $1 \le k \le l \le m_2$，并将 $f^{(2)}_{\boldsymbol{\omega}_2}$ 定义为:

$$f^{(2)}_{\boldsymbol{\omega}_2}(X) = \sum_{i<j} \sum_{k,l} \boldsymbol{\omega}_{2,ijkl} \phi_{kl}(X_i, X_j)\,.$$

这些函数将作为我们备择函数的二维分量。

最后一步： 现在，我们将上述构造的一维分量和二维分量结合起来，构造备择函数。首先，设真实均值函数为 $f_0 = 0$。接着，我们谨慎选取 $S_1 \subset \Omega_1$ 和 $S_2 \subset \Omega_2$，然后构造备择函数如下:

$$\mathcal{F}_S = \left\{ f_{\boldsymbol{\omega}} = f^{(1)}_{\boldsymbol{\omega}_1} + f^{(2)}_{\boldsymbol{\omega}_2},\ \boldsymbol{\omega}_1 \in S_1, \boldsymbol{\omega}_2 \in S_2 \right\}$$

为选择 S_1 和 S_2，我们使用 Varshamov-Gilbert 定理，参见（Tsybakov，2004，引理 2.9）。为方便读者，现将该定理陈述如下:

命题 A.2 （Varshamov-Gilbert）。对于任意 $M \ge 8$，存在 $S \subseteq \{0,1\}^M$，满足 $|S| \ge 2^{M/8}$，使得对于任意 $\boldsymbol{\omega} \neq \boldsymbol{\omega}' \in S$， $\rho(\boldsymbol{\omega}, \boldsymbol{\omega}') \ge M/8$，其中 ρ 为汉明距离。

根据命题 A.2，我们可以选取 S_1 和 S_2，使得对于任意 $\boldsymbol{\omega}_1, \boldsymbol{\omega}_1' \in S_1$，都有$\rho(\boldsymbol{\omega}_1, \boldsymbol{\omega}_1') \ge M_1/8$；并且对于任意 $\boldsymbol{\omega}_2, \boldsymbol{\omega}_2' \in S_2$, $\rho(\boldsymbol{\omega}_2, \boldsymbol{\omega}_2') \ge M_2/8$。注意，对于任意$\boldsymbol{\omega} = (\boldsymbol{\omega}_1, \boldsymbol{\omega}_2) \neq \boldsymbol{\omega}' = (\boldsymbol{\omega}_1', \boldsymbol{\omega}_2')$，都有:

$$\begin{aligned}
d^2(f_{\boldsymbol{\omega}}, f_{\boldsymbol{\omega}'}) &= \int_{[0,1]^d} (f_{\boldsymbol{\omega}}(\mathbf{x}) - f_{\boldsymbol{\omega}'}(\mathbf{x}))^2\, d\mathbf{x} \\
&= \int_{[0,1]^d} \left(\sum_{j=1}^{d} \sum_{k=1}^{m} (\boldsymbol{\omega}_{1,jk} - \boldsymbol{\omega}'_{1,jk}) \phi_k(X_j) + \sum_{i<j} \sum_{k,l} (\boldsymbol{\omega}_{2,ijkl} - \boldsymbol{\omega}'_{2,ijkl}) \phi_{kl}(X_i, X_j) \right)^2 dx \\
&= \int_{[0,1]^d} \left(\sum_{j=1}^{d} \sum_{k=1}^{m} (\boldsymbol{\omega}_{1,jk} - \boldsymbol{\omega}'_{1,jk}) \phi_k(X_j) \right)^2 d\mathbf{x} + \left(\sum_{i<j} \sum_{k,l} (\boldsymbol{\omega}_{2,ijkl} - \boldsymbol{\omega}'_{2,ijkl}) \phi_{kl}(X_i, X_j) \right)^2 d\mathbf{x} \\
&\qquad + 2 \int_{[0,1]^d} \left(\sum_{j=1}^{d} \sum_{k=1}^{m} (\boldsymbol{\omega}_{1,jk} - \boldsymbol{\omega}'_{1,jk}) \phi_k(X_j) \right) \left(\sum_{i<j} \sum_{k,l} (\boldsymbol{\omega}_{2,ijkl} - \boldsymbol{\omega}'_{2,ijkl}) \phi_{kl}(X_i, X_j) \right) d\mathbf{x} \\
&\triangleq T_1 + T_2 + 2T_3\,.
\end{aligned}$$

现在我们分别分析每个 T_i。

$$\begin{aligned}
T_1 &= \int_{[0,1]^d} \left(\sum_{j=1}^{d} \sum_{k=1}^{m} (\boldsymbol{\omega}_{1,jk} - \boldsymbol{\omega}'_{1,jk}) \phi_k(X_j) \right)^2 d\mathbf{x} \\
&= \sum_{j=1}^{d} \sum_{k=1}^{m} (\boldsymbol{\omega}_{1,jk} - \boldsymbol{\omega}'_{1,jk})^2 \int_0^1 \phi_k^2(X_j)\, dX_j
\end{aligned}$$

$$= L^2 h_1^{2\beta_1} \sum_{j=1}^{d} \sum_{k=1}^{m} (\boldsymbol{\omega}_{1,jk} - \boldsymbol{\omega}'_{1,jk})^2 \int_0^1 K^2 \left(\frac{X_j - x_k}{h_1} \right) dX_j$$

$$= L^2 h_1^{2\beta_1+1} \sum_{j=1}^{d} \sum_{k=1}^{m} (\boldsymbol{\omega}_{1,jk} - \boldsymbol{\omega}'_{1,jk})^2 \int_{-\frac{1}{2}}^{\frac{1}{2}} K^2 (z) \, dz$$

$$= L^2 h_1^{2\beta_1+1} \|K\|_2^2 \sum_{j=1}^{d} \sum_{k=1}^{m} (\boldsymbol{\omega}_{1,jk} - \boldsymbol{\omega}'_{1,jk})^2 = L^2 h_1^{2\beta_1+1} \|K\|_2^2 \, \rho(\boldsymbol{\omega}_1, \boldsymbol{\omega}'_1)$$

对于 T_2:

$$T_2 = \int_{[0,1]^d} \left(\sum_{i<j} \sum_{k,l} (\boldsymbol{\omega}_{2,ijkl} - \boldsymbol{\omega}'_{2,ijkl}) \phi_{kl}(X_i, X_j) \right)^2 d\mathbf{x}$$

$$= \sum_{i<j} \sum_{k,l} (\boldsymbol{\omega}_{2,ijkl} - \boldsymbol{\omega}'_{2,ijkl})^2 \int_{[0,1]^2} \phi_{kl}^2(X_i, X_j) \, dX_i \, dX_j$$

$$= L^2 h_2^{2\beta_2} \sum_{i<j} \sum_{k,l} (\boldsymbol{\omega}_{2,ijkl} - \boldsymbol{\omega}'_{2,ijkl})^2 \int_{[0,1]^2} K^2 \left(\frac{X_i - x_j}{h_2} \right) K^2 \left(\frac{X_j - x_k}{h_2} \right) dX_i \, dX_j$$

$$= L^2 h_2^{2\beta_2+2} \sum_{i<j} \sum_{k,l} (\boldsymbol{\omega}_{2,ijkl} - \boldsymbol{\omega}'_{2,ijkl})^2 \left(\int_{-\frac{1}{2}}^{\frac{1}{2}} K^2(z) \, dz \right)^2$$

$$= L^2 h_2^{2\beta_2+2} \|K\|_2^4 \sum_{i<j} \sum_{k,l} (\boldsymbol{\omega}_{2,ijkl} - \boldsymbol{\omega}'_{2,ijkl})^2 = L^2 h_2^{4\beta_2+2} \|K\|_2^4 \rho(\boldsymbol{\omega}_2, \boldsymbol{\omega}'_2) \, .$$

最后我们断言 $T_3 = 0$。这是由以下事实得出的：如果 i、j、k 互不相同，那么对于任意 $1 \le v_1 \le m_1$ 和 $1 \le v_2$、$v_3 \le m_2$:

$$\int_{[0,1]^3} \phi_{v_1}(X_i) \phi_{v_2,v_3}(X_j, X_k) \, dX_i \, dX_j \, dX_k$$

$$= \int_{[0,1]} \phi_{v_1}(X_i) \, dX_i \int_{[0,1]^2} \phi_{v_2,v_3}(X_j, X_k) \, dX_j \, dX_k = 0$$

因为 ϕ_{v_1} 和 ϕ_{v_2}、$_{v_3}$ 的积分根据构造均为 0。当它们不全互异时，设$i = j \neq k$。那么:

$$\int_{[0,1]^2} \phi_{v_1}(X_i) \phi_{v_2,v_3}(X_i, X_k) \, dX_i \, dX_k$$

$$= \int_{[0,1]} \phi_{v_1}(X_i) \left(\int_{[0,1]} \phi_{v_2,v_3}(X_i, X_k) \, dX_k \right) dX_i = 0$$

因为作为 ϕ_{v_2}、$_{v_3}$ 的边缘分布，根据我们的构造其值为 0。因此 $T_3 = 0$。结合 T_i 的表达式，我们得到:

$$d^2(f_{\boldsymbol{\omega}}, f_{\boldsymbol{\omega}'}) = L^2 h_1^{2\beta_1+1} \|K\|_2^4 \, \rho(\boldsymbol{\omega}_1, \boldsymbol{\omega}'_1) + L^2 h_2^{2\beta_2+2} \|K\|_2^4 \, \rho(\boldsymbol{\omega}_2, \boldsymbol{\omega}'_2) \tag{A.7}$$

其中，ρ 是汉明距离。根据命题A.2，我们有:

$$d^2(f_{\boldsymbol{\omega}}, f_{\boldsymbol{\omega}'}) \ge L_1^2 h_1^{2\beta_1+1} \|K\|_2^2 \, \frac{dm_1}{8} + L_2^2 h_2^{2\beta_2+2} \|K\|_2^4 \, \frac{d(d-1)m_2^2}{16}$$

$$= \frac{L_1^2\|K\|_2^2}{8} d m_1^{-2\beta_1} + \frac{L_2^2\|K\|_2^4}{8}\binom{d}{2} m_2^{-2\beta_2} := 4\delta^2 . \tag{A.8}$$

这里，第二个等式是由以下事实得出的：$h_1 = m_1^{-1}$和$h_2 = m_2^{-1}$，第三个不等式则根据 m_1的定义得出，m_2。对集合Mukherjee 等人于2021年的证明（定理2.18的证明）应用 Fano 不等式，我们得到：$\mathcal{F}_S$

$$\inf_{\hat{f}} \sup_{f, P_X} \mathbb{E}_{P_X}[(\hat{f}(X) - f(X))^2] \geq \delta^2 \left(1 - \frac{\frac{n}{|S|^2}\sum_{\omega \neq \omega' \in S} KL\left(\mathbb{P}_\omega | \mathbb{P}_{\omega'}\right) + \log 2}{\log\left(|S| - 1\right)}\right) \tag{A.9}$$

由于误差 ϵ_i'服从正态分布，我们有：

$$\begin{aligned}
KL\left(\mathbb{P}_\omega | \mathbb{P}_{\omega'}\right) &= \frac{1}{2}\mathbb{E}\left[\left(f_\omega(X) - f_{\omega'}(X)\right)^2\right] = \frac{1}{2} d^2\left(f_{\boldsymbol{\omega}}, f_{\boldsymbol{\omega}'}\right) \\
&= L^2 h_1^{2\beta_1+1}\|K\|_2^2\, \rho(\boldsymbol{\omega}_1, \boldsymbol{\omega}_1') + L^2 h_2^{2\beta_2+2}\|K\|_2^4\, \rho(\boldsymbol{\omega}_2, \boldsymbol{\omega}_2') \\
&\leq L^2 h_1^{2\beta_1+1}\|K\|_2^2\, M_1 + L^2 h_2^{2\beta_2+2}\|K\|_2^4\, M_2 \\
&= L^2 h_1^{2\beta_1+1}\|K\|_2^2\, d m_1 + L^2 h_2^{2\beta_2+2}\|K\|_2^4 \binom{d}{2} m_2^2 \\
&= L^2\|K\|_2^2 d m_1^{-2\beta_1} + L^2\|K\|_2^4\binom{d}{2} m_2^{-2\beta_2} \leq 32\delta^2 .
\end{aligned}$$

此外，$|S| = |S_1 \times S_2| \geq 2^{(M_1+M_2)/8}$ 根据命题 A.2。将界代入（A.9），我们有：

$$\inf_{\hat{f}} \sup_{f, P_X} \mathbb{E}_{P_X}[(\hat{f}(X) - f(X))^2] \geq \delta^2 \left(1 - 9\frac{32 n\delta^2 + \log 2}{M_1 + M_2}\right) \tag{A.10}$$

现在，根据 m_1，m_2 的定义，我们有：

$$\begin{aligned}
n\delta^2 &= n\left(\frac{L_1^2\|K\|_2^2}{32} d m_1^{-2\beta_1} + \frac{L_2^2\|K\|_2^4}{32}\binom{d}{2} m_2^{-2\beta_2}\right) \\
&\leq Cn\left(\frac{L_1^2\|K\|_2^2}{32} d n^{-\frac{2\beta_1}{2\beta_1+1}} + \frac{L_2^2\|K\|_2^4}{32}\binom{d}{2} n^{-\frac{2\beta_2}{2\beta_2+2}}\right) \\
&= C\left(\frac{L_1^2\|K\|_2^2}{32} \vee \frac{L_2^2\|K\|_2^4}{32}\right)\left(d n^{\frac{1}{2\beta_1+1}} + \binom{d}{2} n^{\frac{2}{2\beta_2+2}}\right) \\
&\triangleq C\left(\frac{L_1^2\|K\|_2^2}{32} \vee \frac{L_2^2\|K\|_2^4}{32}\right)\psi_n .
\end{aligned}$$

另一方面：

$$M_1 + M_2 = d m_1 + \binom{d}{2} m_2^2 \geq c\left(d n^{\frac{1}{2\beta_1+1}} + \binom{d}{2} n^{\frac{2}{2\beta_2+2}}\right) = c\psi_n .$$

将这些界限代入（A.10）可得：

$$\begin{aligned}
\inf_{\hat{f}} \sup_{f, P_X} \mathbb{E}_{P_X}[(\hat{f}(X) - f(X))^2] &\geq \delta^2\left(1 - 9\frac{C\left(\frac{L_1^2\|K\|_2^2}{32} \vee \frac{L_2^2\|K\|_2^4}{32}\right)\psi_n + \log 2}{c\psi_n}\right) \\
&= \delta^2\left(1 - 9\frac{C\left(\frac{L_1^2\|K\|_2^2}{32} \vee \frac{L_2^2\|K\|_2^4}{32}\right) + \frac{\log 2}{\psi_n}}{c}\right)
\end{aligned}$$

常数 C，c通过 m_1，m_2的定义依赖于 c_1，c_2。适当选择这些常数，并利用 $\psi_n \uparrow \infty$ 当 $n \uparrow \infty$时这一事实，我们可以使

$$9\frac{C\left(\frac{L_1^2\|K\|_2^2}{32} \vee \frac{L_2^2\|K\|_2^4}{32}\right) + \frac{\log 2}{\psi_n}}{c} \leq \frac{1}{2}.$$

这意味着：

$$\inf_{\hat{f}} \sup_{f, P_X} \mathbb{E}_{P_X}[(\hat{f}(X) - f(X))^2] \geq \delta^2$$
$$\geq \left(\frac{L_1^2\|K\|_2^2}{8} \wedge \frac{L_2^2\|K\|_2^4}{8}\right)\left(dn^{\frac{-2\beta_1}{2\beta_1+1}} + \binom{d}{2} n^{\frac{-2\beta_2}{2\beta_2+2}}\right).$$

证明至此完成。

A.3. 定理 3.2 和 3.5 的证明。回顾一下，对于任意 $\phi = \sum_j \phi_j + \sum_{k<l} \phi_{kl}$，我们定义$\|\phi\|_{n,1} := \sum_j \|\phi_j\|_n + \sum_{j<k} \|\phi_{jk}\|_n$和$\|\phi\|_n^2 = \sum_{i=1}^n \phi^2(X_i)/n$。进一步定义$\|\phi\|_{n,lin} := \sum_j \|\phi_j\|_n$和$\|\phi\|_{n,quad} := \sum_{k<l} \|\phi_{kl}\|_n$。根据定理 2.9，存在$\{\phi_j^\star\}_{j\in S_1}$，$\{\phi_{kl}^\star\}_{(j,k)\in S_2}$，$j \in S_1$，$(j,k) \in S_2$，使得，$\phi_j^\star \in \mathcal{F}_{NN,1}$，$\phi_{jk}^\star \in \mathcal{F}_{NN,2}$和

$$\|f_{0,j} - \phi_j^\star\|_\infty \leq C_1(N_1 L_1)^{-2\beta_1} \; \forall \, j \in S_1, \tag{A.11}$$

$$\|f_{0,kl} - \phi_{kl}^\star\|_\infty \leq C_2(N_2 L_2)^{-\beta_2} \; \forall \, (k<l) \in S_2. \tag{A.12}$$

并且，$\phi_j^\star$的积分为 0，$\phi_{kl}^\star$的边缘分布为 0。为简化证明的其余部分，令 L_i 为常数阶。因此，我们在此使用的架构如下：拟合单变量函数时，使用来自类$\mathcal{F}(1, c_1 N_1 \log N_1, c_2, 1)$的深度神经网络；拟合二元函数时，使用来自类$\mathcal{F}(1, c_1 N_2 \log N_2, c_3, 1)$的深度神经网络。因此，定义 $\phi^\star = \sum_{j\in S_1} \phi_j^\star + \sum_{(j,k)\in S_2} \phi_{jk}^\star$，由定理 2.7 可得：

$$\begin{aligned}\mathbb{E}\left[(f_0(X) - \phi^\star(X))^2\right] &\leq C_1\left(s_1(N_1)^{-4\beta_1} + s_2(N_2)^{-2\beta_2}\right) \\ &\triangleq C_1(s_1\rho_{n,1}^2 + s_2\rho_{n,2}^2). \end{aligned} \tag{A.13}$$

由于 ϕ 是神经网络类中惩罚损失函数的最小化者，因此其表现优于 $\phi^\star$。令 $pen(\phi) := \lambda_{n,1}\|\phi\|_{n,lin} + \lambda_{n,2}\|\phi\|_{n,quad}$，则有：

$$\begin{aligned}&\frac{1}{2}\|Y - \hat{\phi}\|_n^2 + pen(\hat{\phi}) \leq \frac{1}{2}\|Y - \phi^\star\|_n^2 + pen(\phi^\star) \\ \implies &\frac{1}{2}\|\hat{\phi} - \phi^\star\|_n^2 + pen(\hat{\phi}) \leq \langle f_0 - \phi^\star, \hat{\phi} - \phi^\star\rangle_n + \langle \epsilon, \hat{\phi} - \phi^\star\rangle_n + pen(\phi^\star) \\ \implies &\frac{1}{2}\|\hat{\phi} - \phi^\star\|_n^2 + pen(\hat{\phi}) \leq \|f_0 - \phi^\star\|_n^2 + \frac{1}{4}\|\hat{\phi} - \phi^\star\|_n^2 + \langle \epsilon, \hat{\phi} - \phi^\star\rangle_n + pen(\phi^\star) \\ \implies &\frac{1}{4}\|\hat{\phi} - \phi^\star\|_n^2 + pen(\hat{\phi}) \leq \|f_0 - \phi^\star\|_n^2 + \langle \epsilon, \hat{\phi} - \phi^\star\rangle_n + pen(\phi^\star). \end{aligned} \tag{A.14}$$

首先，我们将经验误差 $\|f_0 - \phi^\star\|_n^2$ 表示为其总体对应量 $\|f_0 - \phi^\star\|_2^2$的形式。根据切比雪夫不等式，对于 $t > 0$，有：

$$\mathbb{P}\left(\left|\|f_0 - \phi^\star\|_n^2 - \|f_0 - \phi^\star\|_2^2\right| > t\right) \leq \frac{\mathbb{E}\left[(f_0 - \phi^\star)^4\right]}{nt^2} \tag{A.15}$$

展开四阶矩可得:

$$\mathbb{E}\left[(f_0-\phi^\star)^4\right] \le 8\left\{\mathbb{E}\left[\left(\sum_{j\in S_1}(f_{0,j}-\phi_j^\star)\right)^4\right]+\mathbb{E}\left[\left(\sum_{(j,k)\in S_2}(f_{0,jk}-\phi_{jk}^\star)\right)^4\right]\right\}$$

$$\le 8p_{\max}\left\{\int_{[0,1]^d}\left(\sum_{j\in S_1}(f_{0,j}(x_j)-\phi_j^\star(x_j))\right)^4 d\mathbf{x}+\int_{[0,1]^d}\left(\sum_{(j,k)\in S_2}(f_{0,jk}(x_j,x_k)-\phi_{jk}^\star(x_j,x_k))\right)^4 d\mathbf{x}\right\}$$

$\triangleq 8p_{\max}(T_1+T_2)$。

现在我们分别分析 T_1 和 T_2。对于 T_1:

$$\begin{aligned}T_1 &= \int_{[0,1]^d}\left(\sum_{j\in S_1}(f_{0,j}(x_j)-\phi_j^\star(x_j))\right)^4 d\mathbf{x}\\ &= \sum_{j\in S_1}\int_0^1 (f_{0,j}(x_j)-\phi_j^\star(x_j))^4\,dx_j+\sum_{j\ne k\in S_1}\int_0^1 (f_{0,j}(x_j)-\phi_j^\star(x_j))^2\,dx_j\int_0^1 (f_{0,j}(x_k)-\phi_k^\star(x_k))^2\,dx_k\\ &\le C^4 s_1^2(N_1L_1)^{-8\beta}.\end{aligned}$$

其中，最后一行由以下事实得出：$\|f_{0,j}-\phi_j^\star\|_\infty \le C(N_1L_1)^{-2\beta}$。对 T_2 进行类似分析可得:

$$T_2=\int_{[0,1]^d}\left(\sum_{(j,k)\in S_2}(f_{0,jk}(x_j,x_k)-\phi_{jk}^\star(x_j,x_k))\right)^4 d\mathbf{x}\le C^4 s_2^2(N_2L_2)^{-4\beta}.$$

反复使用$\|f_{0,jk}-\phi_{jk}^\star\|_\infty \le C(N_2L_2)^{-\beta}$，并使用假设3.1。利用 T_1 关于 T_2的界，我们有:

$$\begin{aligned}\mathbb{E}\left[(f_0-\phi^\star)^4\right] &\le 8C^4p_{\max}\left(s_1^2(N_1L_1)^{-8\beta}+s_2^2(N_2L_2)^{-4\beta}\right)\\ &\le 8C^4p_{\max}\left(s_1^2\rho_{n,1}^4+s_2^2\rho_{n,2}^4\right).\end{aligned}$$

因此，取 $t_0=C^2\sqrt{\frac{8p_{\max}(s_1^2\rho_{n,1}^4+s_2^2\rho_{n,2}^4)\log n}{n}}$，由（A.15）可得:

$$\mathbb{P}\left(\left|\|f_0-\phi^\star\|_n^2-\|f_0-\phi^\star\|_2^2\right|>t_0\right)\le\frac{1}{\log n} \tag{A.16}$$

定义事件 $\Omega_{n,1}$ 如下:

$$\Omega_{n,1}=\left\{\left|\|f_0-\phi^\star\|_n^2-\|f_0-\phi^\star\|_2^2\right|\le t_0\right\}.$$

注意到 $t_0 \ll s_1\rho_{n,1}^2+s_2\rho_{n,2}^2$，根据（A.13），我们在 $\Omega_{n,1}$ 上有

$$\text{(A.14)} \Longrightarrow \frac{1}{4}\|\hat\phi-\phi^\star\|_n^2+pen(\hat\phi)\le C_3(s_1\rho_{n,1}^2+s_2\rho_{n,2}^2)+\langle\epsilon,\hat\phi-\phi^\star\rangle_n+pen(\phi^\star)$$

接下来，我们对经验误差进行界定，即计算 ϵ 与 $\phi-\phi^\star$ 的内积。首先注意到，根据三角不等式:

$$\left|\langle\epsilon,\hat\phi-\phi^\star\rangle_n\right|\le\sum_{j=1}^d\left|\frac{1}{n}\sum_i\epsilon_i\left(\hat\phi_j(X_{ij})-\phi_j^\star(X_{ij})\right)\right|$$

$$+\sum_{k<l}\left|\frac{1}{n}\sum_{i}\epsilon_i\left(\hat{\phi}_{kl}(X_{ik},X_{il})-\phi^{\star}_{kl}(X_{ik},X_{il})\right)\right|$$

为了界定右侧的每个求和项，我们先对求和中的每一项分别进行界定。为此，我们使用 Fan 和 Gu 的引理 (2022)，该引理指出，若 $\mathcal{G}_n$ 是一个 VC 维数为 V_n 的函数类，则对于任意固定的 $g_0\in\mathcal{G},\epsilon>0,t>0$，以概率 $\geq 1-\log(1/\epsilon)e^{-t}$：

$$\frac{1}{n}\left|\sum_{i=1}^{n}\epsilon_i(g(X_i)-g_0(X_i))\right|\leq C_4\left(\|g-g_0\|_n+\epsilon\right)\sqrt{\frac{V_{n,1}\log n}{n}+\frac{t}{n}}$$

对于某个普适常数 $c_{\circ}$ 注意，$\phi_j\in\mathcal{F}_{NN}(1,c_1N_1\log N_1,c_2,c_3(N_1\log N_1)^2,1)$，其 VC 维数为 $V_{n,1}\leq C_5N_1^2\log N_1$）（见引理 2.8）。现在，我们将此引理应用于每个分量函数，其中 $t=2\log d$，并利用并集界得出：

(A.17)
$$\sum_{j=1}^{d}\left|\frac{1}{n}\sum_{i}\epsilon_i\left(\hat{\phi}_j(X_{ij})-\phi^{\star}_j(X_{ij})\right)\right|\leq C_4\sqrt{\frac{V_{n,1}\log p}{n}+\frac{2\log d}{n}}\left(\sum_{j=1}^{d}\|\hat{\phi}_j-\phi^{\star}_{n,j}\|_n+d\epsilon_1\right).$$

上述事件发生的概率为 $\geq 1-\log(1/\epsilon_1)e^{\log d-2^{\log d}}=1-\log(1/\epsilon_1)/d_{\circ}$ 对双变量分量进行类似计算可得：

$$\sum_{k<l}\left|\frac{1}{n}\sum_{i}\epsilon_i\left(\hat{\phi}_{kl}(X_{ik},X_{il})-\phi^{\star}_{kl}(X_{ik},X_{il})\right)\right|$$

(A.18)
$$\leq C_5\sqrt{\frac{V_{n,2}\log n}{n}+\frac{3\log d}{n}}\left(\sum_{k<l}\|\hat{\phi}_{kl}-\phi^{\star}_{kl}\|_n+\frac{d(d-1)}{2}\epsilon_2\right),$$

其发生概率为 $\geq 1-\log(1/\epsilon_2)e^{2\log d-3\log d}=1-\log(1/\epsilon_2)/d_{\circ}$ 在接下来的分析中，将惩罚参数 $\lambda_{n,1},\lambda_{n,2}$ 和 ϵ_1,ϵ_2 定义为：

(A.19)
$$\lambda_{n,1}=2C_4\sqrt{\frac{V_{n,1}\log p}{n}+\frac{2\log d}{n}}$$

(A.20)
$$\lambda_{n,2}=2C_5\sqrt{\frac{V_{n,2}\log n}{n}+\frac{3\log d}{n}}$$

(A.21)
$$\epsilon_i=\frac{s_i\lambda_{n,i}}{\binom{d}{i}}\text{ for }i=1,2.$$

结合公式（A.17）和（A.18），我们得到：

$$\left|\langle\epsilon,\hat{\phi}-\phi^{\star}\rangle_n\right|\leq\frac{\lambda_{n,1}}{2}\sum_{j=1}^{p}\|\hat{\phi}_j-\phi^{\star}_{n,j}\|_{n,lin}+\frac{\lambda_{n,2}}{2}\sum_{k<l}\|\hat{\phi}_{kl}-\phi^{\star}_{kl}\|_{n,quad}$$

(A.22)
$$+\frac{s_1\lambda_{n,1}^2}{2}+\frac{s_2\lambda_{n,2}^2}{2}.$$

定义事件（A.22）为 $\Omega_{n,\,2}$，稍后我们将证明，对 $(N_1,\ N_2)$ 的选择确保了 $\mathbb{P}(\Omega_{n,\,2})\to 1_{\circ}$ 在事件 $\Omega_{n,\,1}\cap\Omega_{n,\,2}$ 上，有：

$$\frac{1}{4}\|\hat{\phi}-\phi^{\star}\|_n^2+\lambda_{n,1}\|\hat{\phi}\|_{n,lin}+\lambda_{n,2}\|\hat{\phi}\|_{n,quad}$$

$$\leq C_3(s_1\rho_{n,1}^2 + s_2\rho_{n,2}^2) + \frac{\lambda_{n,1}}{2}\sum_{j=1}^{p}\|\hat{\phi}_j - \phi_{n,j}^\star\|_{n,lin}$$

$$+ \frac{\lambda_{n,2}}{2}\sum_{k<l}\|\hat{\phi}_{kl} - \phi_{kl}^\star\|_{n,quad} + \frac{s_1\lambda_{n,1}^2}{2} + \frac{s_2\lambda_{n,2}^2}{2} + \lambda_{n,1}\|\hat{\phi}\|_{n,lin} + \lambda_{n,2}\|\hat{\phi}\|_{n,quad}$$

经过一些简单的代数运算可得:

$$\frac{1}{4}\|\hat{\phi} - \phi^\star\|_n^2 + \frac{\lambda_{n,1}}{2}\|\hat{\phi}\|_{n,lin} + \frac{\lambda_{n,2}}{2}\|\hat{\phi}\|_{n,quad}$$

$$\leq C_3(s_1\rho_{n,1}^2 + s_2\rho_{n,2}^2) + \frac{3\lambda_{n,1}}{2}\sum_{j\in S_1}\|\hat{\phi}_j - \phi_{n,j}^\star\|_{n,lin} + \frac{3\lambda_{n,2}}{2}\sum_{k<l\in S_2}\|\hat{\phi}_{kl} - \phi_{kl}^\star\|_{n,quad}$$

(A.23)

$$+ \frac{s_1\lambda_{n,1}^2}{2} + \frac{s_2\lambda_{n,2}^2}{2}.$$

如果我们不对底层函数类$\hat{}$（即在我们的情形中，具有某种预先指定架构的深度神经网络）假设任何受限强凸性（RSC），那么我们可以利用事实 $\|\phi_j - \phi_{n,j}^\star\|_n \leq 2B$ for j，从而得到:

$$\frac{1}{4}\|\hat{\phi} - \phi^\star\|_n^2 + \frac{\lambda_{n,1}}{2}\|\hat{\phi}\|_{n,lin} + \frac{\lambda_{n,2}}{2}\|\hat{\phi}\|_{n,quad}$$

(A.24)

$$\leq C_3(s_1\rho_{n,1}^2 + s_2\rho_{n,2}^2) + \frac{3}{2}(s_1\lambda_{n,1} + s_2\lambda_{n,2}) + \frac{s_1\lambda_{n,1}^2}{2} + \frac{s_2\lambda_{n,2}^2}{2}.$$

至此，定理 3.2的证明。对于第二部分，我们假设满足受限强凸性条件假设（3.4）。由（A.23），我们得到:

$$\frac{1}{4}\|\hat{\phi} - \phi^\star\|_n^2 + \frac{\lambda_{n,1}}{2}\|\hat{\phi}\|_{n,lin} + \frac{\lambda_{n,2}}{2}\|\hat{\phi}\|_{n,quad}$$

$$\leq C_3(s_1\rho_{n,1}^2 + s_2\rho_{n,2}^2) + \frac{3\lambda_{n,1}}{2}\sum_{j\in S_1}\|\hat{\phi}_j - \phi_{n,j}^\star\|_{n,lin} + \frac{3\lambda_{n,2}}{2}\sum_{k<l\in S_2}\|\hat{\phi}_{kl} - \phi_{kl}^\star\|_{n,quad}$$

$$+ \frac{s_1\lambda_{n,1}^2}{2} + \frac{s_2\lambda_{n,2}^2}{2}$$

$$\leq C_3(s_1(\rho_{n,1}^2 + \lambda_{n,1}^2) + s_2(\rho_{n,2}^2 + \lambda_{n,2}^2)) + \frac{3}{2}\lambda_{n,1}\sqrt{s_1}\sqrt{\sum_{j\in S_1}\|\hat{\phi}_j - \phi_{n,j}^\star\|_{n,lin}^2}$$

$$+ \frac{3}{2}\lambda_{n,2}\sqrt{s_2}\sqrt{\sum_{k<l\in S_2}\|\hat{\phi}_{kl} - \phi_{kl}^\star\|_{n,quad}^2}$$

$$\leq C_3(s_1(\rho_{n,1}^2 + \lambda_{n,1}^2) + s_2(\rho_{n,2}^2 + \lambda_{n,2}^2)) + \frac{9s_1\lambda_{n,1}^2}{2\kappa_1^2} + \frac{\kappa_1^2}{8}\sum_{j\in S_1}\|\hat{\phi}_j - \phi_{n,j}^\star\|_{n,lin}^2$$

$$+ \frac{9s_2\lambda_{n,2}^2}{2\kappa_2^2} + \frac{\kappa_2^2}{8}\sum_{k<l\in S_2}\|\hat{\phi}_{kl} - \phi_{kl}^\star\|_{n,quad}^2$$

$$\leq C_3(s_1(\rho_{n,1}^2 + \lambda_{n,1}^2) + s_2(\rho_{n,2}^2 + \lambda_{n,2}^2)) + \frac{9s_2\lambda_{n,2}^2}{2\kappa_2^2} + \frac{9s_1\lambda_{n,1}^2}{2\kappa_1^2} + \frac{1}{8}\|\hat{\phi} - \phi^\star\|_n^2$$

这意味着：

$$\frac{1}{8}\|\hat{\phi}-\phi^\star\|_n^2+\frac{\lambda_{n,1}}{2}\sum_{j\in S_1^c}\|\hat{\phi}_j\|_n+\frac{\lambda_{n,2}}{2}\sum_{(j<k)\in S_2^c}\|\hat{\phi}_{jk}\|_n \quad \le C_6\left(s_1(\rho_{n,1}^2+\lambda_{n,1}^2)+s_2(\rho_{n,2}^2+\lambda_{n,2}^2)\right)$$

现在，我们通过平衡 $\rho_{n,i}$ 和 $\lambda_{n,i}$，选取 N_i 的最优值。首先取 $i=1$。回顾一下，我们有：

$$\rho_{n,1}^2=N_1^{-4\beta_1},\ \text{and}\ \lambda_{n,1}^2=4C_4^2\left(\frac{V_{n,1}\log n}{n}+\frac{2\log d}{n}\right)\le C_7\left(\frac{N_1^2\log^3 N_1\log n}{n}+\frac{\log d}{n}\right)$$

取 $N_1=\lfloor n^{\frac{2}{2(2\beta_1+1)}}\rfloor$，得到：

$$\rho_{n,1}^2\sim n^{-\frac{2\beta_1}{2\beta_1+1}}, \tag{A.25}$$

$$\lambda_{n,1}^2\sim n^{-\frac{2\beta_1}{2\beta_1+1}}\log^4 n+\frac{\log d}{n}. \tag{A.26}$$

特别地，我们有：

$$\rho_{n,1}^2+\lambda_{n,1}^2\le C_8\left(n^{-\frac{2\beta_1}{2\beta_1+1}}\log^4 n+\frac{\log d}{n}\right).$$

对 $i=2$ 进行类似计算可知，取 $N_2=\lfloor n^{\frac{2}{2(\beta_2+1)}}\rfloor$ 时，我们有：

$$\rho_{n,2}^2+\lambda_{n,2}^2\le C_9\left(n^{-\frac{\beta_2}{\beta_2+1}}\log^4 n+\frac{\log d}{n}\right).$$

因此，上述惩罚参数的选择给出：

$$\begin{aligned}\frac{1}{8}\|\hat{\phi}-\phi^\star\|_n^2+\frac{\lambda_{n,1}}{2}\sum_{j\in S_1^c}\|\hat{\phi}_j\|_n+\frac{\lambda_{n,2}}{2}\sum_{(j<k)\in S_2^c}\|\hat{\phi}_{jk}\|_n&\\ \le C_{10}\left(s_1\left(n^{-\frac{2\beta_1}{2\beta_1+1}}\log^4 n+\frac{\log d}{n}\right)+s_2\left(n^{-\frac{\beta_2}{\beta_2+1}}\log^4 n+\frac{\log d}{n}\right)\right).&\end{aligned}$$

证明至此完成。

A.4. 定理 3.8 的证明。为简化记号，定义 $r_n^2=\tilde{C}(s_1\lambda_{n,1}^2+s_2\lambda_{n,2}^2)$，其中 C 是某个常数，其具体取值将在证明后文明确给出。首先，我们证明对 $i\in\{1,2\}$ 进行硬阈值处理（算法 1 的步骤 3）得到的集合 S_i 以高概率（whp）满足 $S_i\supset S_i$。回顾一下，对所有 $j\in S_1$ 和 $(j<k)\in S_2$，有 $\|f_j^0\|_2>r_n$ 和 $\|f_{jk}^0\|_2>r_n$。根据 Lu 等人（2021），存在深度神经网络 $\{\phi_j^\star\}_j$ 和 $\{\phi_{jk}^\star\}$，使得 $\|\phi_j^\star\|_2>r_n/2$ 且 $\|\phi_{jk}^\star\|_2>r_n/2$；根据 r_n 的定义，对所有 $j\in S_1$ 和 $(j<k)\in S_2$，每个分量的逼近误差为 $<r_n/2$。我们首先证明 $S_1\subseteq S_1$ 以高概率成立。关键思路如下：若 $S_1\nsubseteq S_1$，则存在 $j\in S_1$ 使得 $j\notin S_1$，即 $\|\phi_j\|_n<C_2\lambda_n$，1。由于 $\|\phi_j^\star\|_2\ge r_n/2$，我们有 $\|\phi_j^\star\|_n\ge r_n/4$ 以高概率成立（这同样可由 $nr_n^2\to\infty$ 得出，细节稍后说明）。若 $\|\phi_j\|_n<C_2\lambda_n$，1，则

$$C(s_1\lambda_{n,1}^2+s_2\lambda_{n,2}^2)\ge\frac{1}{8}\|\hat{\phi}-\phi^\star\|_n^2\ge\frac{1}{8}\|\hat{\phi}_j-\phi_j^\star\|_n^2\ge\frac{1}{8}\left(\frac{r_n}{4}-C_2\lambda_{n,1}\right)^2\ge\frac{r_n^2}{2^7}. \tag{A.27}$$

一旦 $\tilde{C}^2\ge 2^7C$，便会得出矛盾。现在我们将这一直觉严格化。定义以下事件：

1. $\Omega_{n,1}=\{\sum_{j\in S_1}\|\hat{\phi}_j-\phi_j^\star\|_n\le C(s_1\lambda_{n,1}+s_2\lambda_{n,2})\}$.

2. $\Omega_{n,2} = \left\{ \left| \|\phi_j^\star\|_n - \|\phi_j^\star\|_2 \right| \le \frac{\|\phi_j^\star\|_2^2}{2} \ \forall\ j \in S_1 \right\}$.
3. $\Omega_{n,3} = \left\{ \left| \|\phi_{jk}^\star\|_n - \|\phi_{jk}^\star\|_2 \right| \le \frac{\|\phi_{jk}^\star\|_2^2}{2} \ \forall\ (j<k) \in S_2 \right\}$.

利用定理 2.9 $\mathbb{P}(\Omega_{n,1}) \to 1$ 作为 $n \to \infty$。对于 $\Omega_{n,2}$，注意到：

$$\begin{aligned}
\sum_{j\in S_1} \mathbb{P}\left(\left| \|\phi_j^\star\|_n^2 - \|\phi_j^\star\|_2^2 \right| \ge \frac{\|\phi_j^\star\|_2^2}{2} \right) &\le \sum_{j\in S_1} \frac{4\mathsf{var}\left(\left(\phi_j^\star(X)\right)^2 \right)}{n\|\phi_j^\star\|_2^4} \\
&\le \sum_{j\in S_1} \frac{4\mathbb{E}\left(\left(\phi_j^\star(X)\right)^4 \right)}{n\|\phi_j^\star\|_2^4} \\
&\le \sum_{j\in S_1} \frac{4B^2\mathbb{E}\left(\left(\phi_j^\star(X)\right)^2 \right)}{n\|\phi_j^\star\|_2^4} \qquad [\text{As } \|\phi_j^\star\|_\infty \le B] \\
&= \frac{4s_1B^2}{n\|\phi_j^\star\|_2^2} \le \frac{4s_1B^2}{nr_n^2}.
\end{aligned}$$

根据我们对 r_n 的选择，使用 (3.7) 可立即得到 $\frac{nr_n^2}{s_1} \gtrsim \lambda_{n,1}^2 \gtrsim$ 对数 $n \to \infty$。这意味着 $\mathbb{P}(\Omega_{n,2}^c) \to 0$。类似的计算可得 $\mathbb{P}(\Omega_{n,3}^c) \to 0$。现在，在事件 $\Omega_{n,1} \cap \Omega_{n,2} \cap \Omega_{n,3}$ 上，由 (A.27) 可知，对于 $i=1,2$，$S_i \subseteq S_i$。

在接下来的计算中，定义事件 $\mathcal{T} = \{S_i \subseteq S_i, i=1,2\}$。我们已证明 $\mathbb{P}(\mathcal{T}) = 1 - o(1)$。在接下来的分析中，我们假设 $\mathcal{T}$ 发生。定义 $|S_1 \cap S_1^c| = \gamma_1$ 和 $|S_2 \cap S_2^c| = \gamma_2$。因此，$|S_1| = s_1 + \gamma_1$ 和 $|S_2| = s_2 + \gamma_2$。注意，γ_1，γ_2 的取值满足：

$$\begin{aligned}
2C_{10}\left(\gamma_1\lambda_{n,1}^2 + \gamma_2\lambda_{n,2}^2\right) &\le \frac{\lambda_{n,1}}{2}\sum_{j\in S_1^c}\|\hat\phi_j\|_n + \frac{\lambda_{n,2}}{2}\sum_{(j<k)\in S_2^c}\|\hat\phi_{jk}\|_n \\
&\le 2C_{10}\left(s_1\lambda_{n,1}^2 + s_2\lambda_{n,2}^2\right)
\end{aligned}$$

i.e.

$$\gamma_1\lambda_{n,1}^2 + \gamma_2\lambda_{n,2}^2 \le s_1\lambda_{n,1}^2 + s_2\lambda_{n,2}^2. \tag{A.28}$$

根据算法 1，我们使用数据的后半部分来估计均值函数 $\hat{}$，仅限于 S_1 和 S_2，并最小化（未惩罚的）平方误差损失，即我们的最终估计值为：

$$\hat f^{\text{final}}(x) = \sum_{j\in\hat S_1}\hat\phi_j^{\text{final}}(x_j) + \sum_{(j<k)\in\hat S_2}\hat\phi_{jk}^{\text{final}}(x_j,x_k),$$

其中，分量函数的估计方式为：

$$\{\hat\phi_j^{\text{final}}\}, \{\hat\phi_{jk}^{\text{final}}\} = \arg\min_{\phi_j,\phi_{jk}} \frac{1}{n}\sum_i \left(Y_i - \sum_{j\in\hat S_1}\phi_j(X_{ij}) - \sum_{(j<k)\in\hat S_2}\phi_{jk}(X_{ij},X_{ik}) \right)^2.$$

证明的其余部分与定理2.9类似。根据（Fan 和 Gu，2022，定理 4.8），我们知道，若 f 是一个 d 元 β 光滑函数，则存在神经网络 g，其

宽度为 N，且深度恒定（其中常数取决于 d，β），使得

$$\|f-g\|_\infty \le cN^{-\frac{2\beta}{d}}.$$

对于某个常数 $c=c(\beta,\ d)>0$。这里我们分别拟合单变量函数和二元函数。假设我们使用宽度为 N_1（且深度恒定为 c_1）的神经网络拟合单变量分量，并使用宽度为 N_2（且深度恒定为 $\hat{}c_2$）的神经网络拟合双变量分量。于是，为了拟合加性函数，我们需要 $|S_1|$ 个这样的单变量分量和 $\hat{}|S_2|$ 个双变量分量。因此，拟合此类均值函数所需的权重总数（活跃参数数量）为 $W=C_1(|\hat{}S_1|N_1^2+|\hat{}S_2|N_2^2)$。根据 Bartlett 等人 (2019)，此类神经网络的 VC 维数为 $V_n \le C_2 W\log W$（因为深度为 $O(1)$）。偏差-方差分解给出：

$$\|\hat{f}^{\text{final}}-f_0\|_{L_2(P_X)}^2 \le 2\left(\underbrace{\|\phi^\star-f_0\|_{L_2(P_X)}^2}_{\text{bias}}+\underbrace{\|\hat{f}^{\text{final}}-\phi^\star\|_{L_2(P_X)}^2}_{\text{variance}}\right)$$

其中 $\phi^\star$ 是 f_0 在我们进行优化的神经网络类中的最佳逼近函数，即由 $\hat{}\ \hat{}$ 个宽度为 $|S_1|$ 的单变量网络之和，以及 $|S_2|\hat{}$ 个宽度为 N_2的双变量分量之和组成的函数。由于 $S_i\subseteq S_i$ 对于 $i\in\{1,2\}$，根据定理 3.5 的证明（见公式 A.13）可知，

$$\|\phi^\star-f_0\|_{L_2(P_X)}^2 \le C_1\left(s_1N_1^{-4\beta_1}+s_2N_2^{-2\beta_2}\right).$$

现在，为了界定方差项，我们采用与定理 2.9 的证明类似的 VC 维数技术。对于任意选定的分量函数，我们可以将总体和 $\sum_{j\in\hat{}S_1}\phi_j+\sum_{(j<k)\in S_2}\phi_{jk}$ 视为来自一个 VC 类的函数，其 VC-维数为 $\le C_2W\log W$。因此，根据定理 2.9 的证明，我们有：

$$\begin{aligned}
\|\hat{f}^{\text{final}}-f_0\|_{L_2(P_X)}^2 &= O_p\left(\frac{V_n\log n}{n}+s_1N_1^{-4\beta_1}+s_2N_2^{-2\beta_2}\right)\\
&= O_p\left(\frac{W\log W\log n}{n}+s_1N_1^{-4\beta_1}+s_2N_2^{-2\beta_2}\right)\\
&= O_p\left(\frac{(s_1+\gamma_1)N_1^2+(s_2+\gamma_2)N_2^2\log\left((s_1+\gamma_1)N_1^2+(s_2+\gamma_2)N_2^2\right)\log n}{n}\right.\\
&\qquad\left.+s_1N_1^{-4\beta_1}+s_2N_2^{-2\beta_2}\right)\\
&= O_p\left((s_1+\gamma_1)\lambda_{n,1}^2+(s_2+\gamma_2)\lambda_{n,2}^2\right)=O_p\left(s_1\lambda_{n,1}^2+s_2\lambda_{n,2}^2\right),
\end{aligned}$$

其中倒数第二行由 λ_n 的定义得出，1 以及 λ_n，2 (参见(3.7))；最后一行由 (A.28) 得出。证明完毕。

A.5. 定理 3.9 的证明。 在此，我们将定理2.12 的证明扩展到稀疏交互模型。这里的主要技术变化是纳入稀疏性。为此，我们需要对（m_1，m_2 来自A.4）采用略有不同的定义。我们将 m_1 定义为以下方程的解：

$$m_1^{-2\beta_1}=c_1\left(n^{-\frac{2\beta_1}{2\beta_1+1}}\vee 8\frac{\log\left(\frac{2d}{s_1}-2\right)}{n}\right)=\frac{c_1}{n}\left(n^{\frac{1}{2\beta_1+1}}\vee 8\log\left(\frac{2d}{s_1}-2\right)\right) \tag{A.29}$$

并将 m_2 定义为以下方程的解：

(A.30)

$$m_2^{-2\beta_2} = c_2\left(n^{-\frac{2\beta_2}{2\beta_2+2}} \vee 8\frac{\log\left(\frac{d(d-1)}{s_2}-2\right)}{n}\right) = \frac{c_2}{n}\left(n^{-\frac{2\beta_2}{2\beta_2+2}} \vee 8\log\left(\frac{d(d-1)}{s_2}-2\right)\right)$$

若 m_1, m_2 不是整数，我们将取最接近的整数。由于这不会影响渐近结果，为简化证明，以下假设它们是精确解。常数 $c_1, c_2 > 0$ 将在证明末尾选取。类似于 (A.4)，定义 $h_i := m_i^{-1}, i \in \{1,2\}$。考虑 $i \in \{1,2\}$ 对应的集合 $\Omega_i = \{0,1\}^{m_i}$。根据命题 A.2，我们可以找到 $S_i \subset \Omega_i$，使得 $|S_i| \geq 2^{m_i/8}$，并且对任意 $\boldsymbol{\omega}, \boldsymbol{\omega}' \in S_i$，都有 $\rho(\boldsymbol{\omega}, \boldsymbol{\omega}') \geq m_i/8$，其中 ρ 是汉明距离。和之前一样，定义 $\Omega = \Omega_1 \times \Omega_2$ 和 $S = S_1 \times S_2$。对于任意 $\boldsymbol{\omega} \in \Omega$，我们将其写作 $\boldsymbol{\omega} = (\boldsymbol{\omega}_1, \boldsymbol{\omega}_2)$，其中 $\boldsymbol{\omega}_i \in \Omega_i$。对于一维分量，给定任意 $\boldsymbol{\omega}_1 \in \Omega_1,$，我们定义函数 $f^{(1)}_{\boldsymbol{\omega}_1}$ 如下：

$$f^{(1)}_{\boldsymbol{\omega}_1}(x) = \sum_{k=1}^{m} \boldsymbol{\omega}_{1,k}\phi_k(x)\,,$$

其中 ϕ_k 在（A.5）中定义。定义 $\mathcal{F}_1 := \{f^{(1)}_{\boldsymbol{\omega}_1} \colon \boldsymbol{\omega}_1 \in S_1\}$ 和 $\Gamma_1 := |\mathcal{F}_1| \geq 2^{m_1/8}$。将 $\mathcal{F}_1$ 中的函数依次记为 $\mathcal{F}_1 = \{f^{(1)}_1, \ldots, f_{(1)_{\Gamma_1}}\}$。注意，对于任意 $i \neq j$：

$$\begin{aligned}
&\int_0^1 \left(f^{(1)}_i(x) - f^{(1)}_j(x)\right)^2 dx \\
&= \sum_{k=1}^{m} (\boldsymbol{\omega}_{i,k} - \boldsymbol{\omega}_{j,k})^2 \int_0^1 \phi_k^2(x)\, dx \quad [\text{Since } \langle \phi_j, \phi_k \rangle_{L_2} = 0 \;\forall\, j \neq k] \\
&= L_1^2 h^{2\beta_1+1} \|K\|_2^2 \rho(\boldsymbol{\omega}_i, \boldsymbol{\omega}_j) \qquad \text{(A.31)}
\end{aligned}$$

对于二维分量，给定任意 $\boldsymbol{\omega}_2 \in \Omega_2,$，我们定义 $f^{(2)}_{\boldsymbol{\omega}_2}$ 如下：

$$f^{(2)}_{\boldsymbol{\omega}_2}(x,y) = \sum_{k,l=1}^{m} \boldsymbol{\omega}_{2,k,l}\phi_{k,l}(x,y)\,,$$

其中 $\phi_{k,l}$ 由（A.6）定义。定义 $\mathcal{F}_2 := \{f^{(2)}_{\boldsymbol{\omega}_2} \colon \boldsymbol{\omega}_2 \in S_2\}$ 和 $\Gamma_2 := |\mathcal{F}_2| \geq 2^{m_2^2/8}$。将 $\mathcal{F}_2$ 中的函数依次记为 $\mathcal{F}_2 = \{f^{(2)}_1, \ldots, f_{(2)_{\Gamma_2}}\}$。对于任意 $i \neq j$：

$$\begin{aligned}
&\int_{[0,1]^2} \left(f^{(2)}_i(x,y) - f^{(2)}_j(x,y)\right)^2 dx\, dy \\
&= \sum_{1 \leq k,l \leq m_2} (\boldsymbol{\omega}_{i,k,l} - \boldsymbol{\omega}_{j,k,l})^2 \iint_0^1 \phi^2_{k,l}(x,y)\, dx \quad [\text{Since } \langle \phi_{j,k}, \phi_{l,m} \rangle_{L_2} = 0 \;\forall\, (j,k) \neq (l,m)] \\
&= L_2^2 h^{2\beta_2+2} \|K\|_4^2 \rho(\boldsymbol{\omega}_i, \boldsymbol{\omega}_j) \qquad \text{(A.32)}
\end{aligned}$$

为构造稀疏备选项集合，定义集合 U_1^* 和 U_2^* 如下：

$$U_i^* = \left\{u \in \{0,1,\ldots,\Gamma_1\}^{\binom{d}{2}} : \|u\|_0 = s_i\right\}, \qquad i = 1,2.$$

显然， $|U_i^*| = \binom{d}{i}_{s_i}\Gamma_i^{s_i}$， $i=1$， 2。选取$U_i \subset U_i^*$使得对于任意 u， $v \in U_i$， 均有 $\rho(u, v) \geq s_i/2$。根据（Raskutti 等人， 2012，引理4）的证明，我们有：

$$|U_i| \geq \frac{1}{2}\left(\frac{\binom{d}{i} - s_i}{s_i/2}\right)^{s_i/2} \Gamma_i^{s_i/2}, \qquad i=1,2.$$

对于 $i \leq 2$, $u_i \in U_i$， 定义函数 $f_{u_i}^{(i)}$如下：

$$f_{u_1}^{(1)}(X) = \sum_{j=1}^{d} f_{u_{1,j}}^{(1)}(X_j), \qquad f_{u_2}^{(2)}(X) = \sum_{1\leq i<j\leq d} f_{u_{2,i,j}}^{(2)}(X_i, X_j)$$

其中 $f_0^{(i)} = 0$， 否则从 $\mathcal{F}_i$ 中选取。现在，我们的备择函数集合为：

$$f(X) = f_{u_1}^{(1)}(X) + f_{u_2}^{(2)}(X) \;\; u_1 \in U_1, u_2 \in U_2\,.$$

定义 $\mathcal{F}^{\text{sparse}}$ 为所有此类函数 f 的集合。令 $M = |\mathcal{F}_{\text{稀疏}}|$。任取两个 f， $f' \in \mathcal{F}^{\text{sparse}}$。我们有：

$$\begin{aligned}
d^2(f,f') &= \int_{[0,1]^d} \left(\sum_{j=1}^{d}(f_{u_{1,j}}^{(1)} - f_{u'_{1,j}}^{(1)})(X_j) + \sum_{1\leq i<j\leq d}(f_{u_{2,i,j}}^{(2)} - f_{u'_{2,i,j}}^{(2)})(X_i, X_j)\right)^2 dX \\
&= \int_{[0,1]^d} \left(\sum_{j=1}^{d}(f_{u_{1,j}}^{(1)} - f_{u'_{1,j}}^{(1)})(X_j)\right)^2 dX + \int_{[0,1]^d} \left(\sum_{1\leq i<j\leq d}(f_{u_{2,i,j}}^{(2)} - f_{u'_{2,i,j}}^{(2)})(X_i, X_j)\right)^2 dX \\
&\qquad + 2\int_{[0,1]^d} \left(\sum_{j=1}^{d}(f_{u_{1,j}}^{(1)} - f_{u'_{1,j}}^{(1)})(X_j)\right)\left(\sum_{1\leq i<j\leq d}(f_{u_{2,i,j}}^{(2)} - f_{u'_{2,i,j}}^{(2)})(X_i, X_j)\right) dX \\
&\triangleq T_1 + T_2 + 2T_3
\end{aligned}$$

现在我们分别分析每个 T_i。

$$\begin{aligned}
T_1 &= \int_{[0,1]^d} \left(\sum_{j=1}^{d}(f_{u_{1,j}}^{(1)} - f_{u'_{1,j}}^{(1)})(X_j)\right)^2 dX \\
&= \sum_{j=1}^{d}\int_0^1 \left((f_{u_{1,j}}^{(1)} - f_{u'_{1,j}}^{(1)})(X_j)\right)^2 dX_j \quad \left[\text{Since} \int f_{u_1,j}^{(1)}(X_j)\, dX_j = 0\right] \\
&= L_1^2 h_1^{2\beta_1+1}\|K\|_2^2 \sum_{j=1}^{d} \rho(\boldsymbol{\omega}_{u_{1,j}}, \boldsymbol{\omega}_{u'_{1,j}})\, \mathbb{1}_{u_{1,j}\neq u'_{1,j}} \quad \text{[From (A.31)]}
\end{aligned}$$

$$\begin{aligned}
T_2 &= \int_{[0,1]^d} \left(\sum_{1\leq i<j\leq d}(f_{u_{2,i,j}}^{(2)} - f_{u'_{2,i,j}}^{(2)})(X_i, X_j)\right)^2 dX \\
&= \sum_{1\leq i<j\leq d}\int_{[0,1]^2} \left((f_{u_{2,i,j}}^{(2)} - f_{u'_{2,i,j}}^{(2)})(X_i, X_j)\right)^2 dX_i\, dX_j \quad [\text{ As marginals of } f^{(2)} \text{ are } 0] \\
&= L_2^2 h_2^{2\beta_2+2} \sum_{1\leq i<j\leq d} \rho(\boldsymbol{\omega}_{u_{2,i,j}}, \boldsymbol{\omega}_{u'_{2,i,j}})\, \mathbb{1}_{u_{2,i,j}\neq u'_{2,i,j}} \quad \text{[From (A.32)]}
\end{aligned}$$

Furt此外，由于 $f^{(2)}$ 的边缘分布为0。结合 T_i的界，我们得到 n:

$$
\begin{aligned}
d^2(f,f') &= L_1^2 h_1^{2\beta_1+1}\|K\|_2^2 \sum_{j=1}^{d} \rho(\boldsymbol{\omega}_{u_{1,j}}, \boldsymbol{\omega}_{u'_{1,j}})\, \mathbb{1}_{u_{1,j}\neq u'_{1,j}} \\
&\quad + L_2^2 h_2^{2\beta_2+2}\|K\|_2^4 \sum_{1\le i<j\le d} \rho(\boldsymbol{\omega}_{u_{2,i,j}}, \boldsymbol{\omega}_{u'_{2,i,j}})\, \mathbb{1}_{u_{2,i,j}\neq u'_{2,i,j}} \\
&\ge \frac{L_1^2\|K\|_2^2}{8} h_1^{2\beta_1+1} m_1 \sum_{j=1}^{d} \mathbb{1}_{u_{1,j}\neq u'_{1,j}} + \frac{L_2^2\|K\|_2^4}{8} h_2^{2\beta_2+2} m_2^2 \sum_{1\le i<j\le d} \mathbb{1}_{u_{2,i,j}\neq u'_{2,i,j}} \\
&\ge \frac{L_1^2\|K\|_2^2}{16} h_1^{2\beta_1+1} m_1 s_1 + \frac{L_2^2\|K\|_2^4}{16} h_2^{2\beta_2+2} m_2^2 s_2 \\
&\ge \frac{L_1^2\|K\|_2^2}{16} m_1^{-2\beta_1} s_1 + \frac{L_2^2\|K\|_2^4}{16} m_2^{-2\beta_2} s_2 := 4\delta^2 . \qquad \text{(A.33)}
\end{aligned}
$$

类似地，我们可以得到关于 δ 的上界：

$$
\begin{aligned}
d^2(f,f') &= L_1^2 h_1^{2\beta_1+1}\|K\|_2^2 \sum_{j=1}^{d} \rho(\boldsymbol{\omega}_{u_{1,j}}, \boldsymbol{\omega}_{u'_{1,j}})\, \mathbb{1}_{u_{1,j}\neq u'_{1,j}} \\
&\quad + L_2^2 h_2^{2\beta_2+2}\|K\|_2^4 \sum_{1\le i<j\le d} \rho(\boldsymbol{\omega}_{u_{2,i,j}}, \boldsymbol{\omega}_{u'_{2,i,j}})\, \mathbb{1}_{u_{2,i,j}\neq u'_{2,i,j}} \\
&\le 2L_1^2\|K\|_2^2 s_1 m_1 h_1^{2\beta_1+1} + 2L_2^2\|K\|_2^4 s_2 m_2^2 h_2^{2\beta_2+2} \qquad \text{(A.34)} \\
&= 2L_1^2\|K\|_2^2 s_1 m_1^{-2\beta_1} + 2L_2^2\|K\|_2^4 s_2 m_2^{-2\beta_2} = 128\delta^2 . \qquad \text{(A.35)}
\end{aligned}
$$

结合方程（A.33）和方程（A.34）中得到的界，对于任意两个 f， $f' \in \mathcal{F}^{\text{sparse}}$：

$$
4\delta^2 \le d^2(f,f') \le 128\delta^2 . \qquad \text{(A.36)}
$$

这进一步表明 $\mathcal{F}^{\text{sparse}}$ 是 2δ 所有可行函数集合的一个打包集。应用 Fano 不等式（Mukherjee 等人，2021年，定理2.18的证明）可得：

$$
\inf_{\hat{f}} \sup_{f, P_X} \mathbb{E}_{P_X}[(\hat{f}(X) - f(X))^2] \ge \delta^2 \left(1 - \frac{\frac{n}{M^2}\sum_{f,f'\in\mathcal{F}^{\text{sparse}}} KL\left(\mathbb{P}_f|\mathbb{P}_{f'}\right) + \log 2}{\log\left(M-1\right)}\right) \qquad \text{(A.37)}
$$

现在，由于误差服从正态分布：

$$
KL\left(\mathbb{P}_f|\mathbb{P}_{f'}\right) = \mathbb{E}\left[\left(f(X) - f'(X)\right)^2\right] = d^2(f,f')
$$

利用方程（A.36）的界，我们有 $KL\left(\mathbb{P}_f|\mathbb{P}_{f'}\right) \le 128\delta^2$ 对所有 f， f'。因此，由方程（A.37）可得：

$$
\begin{aligned}
\inf_{\hat{f}} \sup_{f, P_X} \mathbb{E}_{P_X}[(\hat{f}(X) - f(X))^2] &\ge \delta^2 \left(1 - \frac{128n\delta^2 + \log 2}{\log\left(M-1\right)}\right) \\
&\ge \delta^2 \left(1 - \frac{128n\delta^2 + \log 2}{2\log M}\right), \qquad \text{(A.38)}
\end{aligned}
$$

其中 $M = |U_1||U_2|$。证明的其余部分是平衡 $n\delta^2$ 和对数 M，也就是说，我们可以说 ~~$128n\delta^2 +$对数~~ 2对数 M $\le c^*$，其中某个 $0 < c^* < 1$。为此，只需证明 $n\delta^2$和对数 M具有

相同的阶。随后我们可以谨慎地选择常数，使比率落在$(0,\ 1)$内。根据 δ 的定义，我们有：

$$
\begin{aligned}
n\delta^2 &= \frac{L_1^2\|K\|_2^2}{64} n m_1^{-2\beta_1} s_1 + \frac{L_2^2\|K\|_2^4}{64} n m_2^{-2\beta_2} s_2 \\
&= \frac{L_1^2\|K\|_2^2}{4} n m_1^{-2\beta_1} \frac{s_1}{16} + \frac{L_2^2\|K\|_2^4}{4} n m_2^{-2\beta_2} \frac{s_2}{16} \\
&= \frac{L_1^2 c_1\|K\|_2^2}{4} \frac{s_1}{16} \left[n^{\frac{1}{2\beta_1+1}} \vee 8\log\left(\frac{2d}{s_1} - 2\right)\right] \\
&\qquad + \frac{L_2^2 c_2\|K\|_2^4}{4} \frac{s_2}{16} \left[(n^{\frac{1}{2\beta_2+2}} \vee 8\log\left(\frac{d(d-1)}{s_2} - 2\right)\right]
\end{aligned} \tag{A.39}
$$

另一方面，根据我们的构造，我们有：

$$
\begin{aligned}
\log M &= \log\ |U_1| + \log\ |U_2| \\
&\geq 2\log\frac{1}{2} + \frac{s_1}{2}\log\left(\frac{d-s_1}{s_1/2}\right) + \frac{s_1}{2}\log\Gamma_1 + \frac{s_2}{2}\log\left(\frac{d(d-1)/2 - s_2}{s_2/2}\right) + \frac{s_2}{2}\log\Gamma_2 \\
&= 2\log\frac{1}{2} + \frac{s_1}{2}\log\left(\frac{2d}{s_1} - 2\right) + \frac{s_1}{2}\log\Gamma_1 + \frac{s_2}{2}\log\left(\frac{d(d-1)}{s_2} - 2\right) + \frac{s_2}{2}\log\Gamma_2 \\
&= 2\log\frac{1}{2} + \frac{s_1}{2}\log\left(\frac{2d}{s_1} - 2\right) + \frac{s_1 m_1}{16} + \frac{s_2}{2}\log\left(\frac{d(d-1)}{s_2} - 2\right) + \frac{s_2 m_2^2}{16} \\
&= 2\log\frac{1}{2} + \frac{s_1}{16}\left[m_1 + 8\log\left(\frac{2d}{s_1} - 2\right)\right] + \frac{s_2}{16}\left[m_2^2 + 8\log\left(\frac{d(d-1)}{s_2} - 2\right)\right] \\
&\triangleq 2\log\frac{1}{2} + a_n + b_n\,.
\end{aligned} \tag{A.40}
$$

现在我们考虑四种情况：

Case 1: $n^{1/(2\beta_1+1)} \geq 8\log\left(\frac{2d}{s_1} - 2\right)$ **and** $n^{\frac{1}{2\beta_2+2}} \geq 8\log\left(\frac{d(d-1)}{s_2} - 2\right)$

利用（A.29）、（A.30），可得 $m_1 = n^{1/(2\beta_1+1)}$ 以及 $m_2 = n^{1/(2\beta_2+2)}$。因此，方程（A.39）和（A.40）可简化为：

$$
\begin{aligned}
n\delta^2 &= \frac{L_1^2 c_1\|K\|_2^2}{4}\frac{s_1}{16} n^{\frac{1}{2\beta_1+1}} + \frac{L_2^2 c_2\|K\|_2^4}{4}\frac{s_2}{16} n^{-\frac{2\beta_2}{2\beta_2+2}} \\
&\leq \frac{L_1^2 c_1\|K\|_2^2}{4}\frac{s_1}{16}\left[m_1 + 8\log\left(\frac{2d}{s_1} - 2\right)\right] + \frac{L_2^2 c_2\|K\|_2^4}{4}\frac{s_2}{16}\left[m_2^2 + 8\log\left(\frac{d(d-1)}{s_2} - 2\right)\right] \\
&\leq \frac{L_1^2 c_1\|K\|_2^2}{4} a_n + \frac{L_2^2 c_2\|K\|_2^4}{4} b_n
\end{aligned}
$$

Case 2: $n^{1/(2\beta_1+1)} \geq 8\log\left(\frac{2d}{s_1} - 2\right)$ **and** $n^{\frac{1}{2\beta_2+2}} < 8\log\left(\frac{d(d-1)}{s_2} - 2\right)$

Here, $m_1 = n^{1/(2\beta_1+1)}$ and $m_2 = \left(8\log\left(\frac{d(d-1)}{s_2}-2\right)/n\right)^{-1/2\beta_2}$ yielding

$$\begin{aligned} n\delta^2 &= \frac{L_1^2 c_1 \|K\|_2^2}{4}\frac{s_1}{16} n^{\frac{1}{2\beta_1+1}} + \frac{L_2^2 c_2 \|K\|_2^4}{4}\frac{s_2}{16} 8\log\left(\frac{d(d-1)}{s_2}-2\right) \\ &\le \frac{L_1^2 c_1 \|K\|_2^2}{4}\frac{s_1}{16}\left[m_1 + 8\log\left(\frac{2d}{s_1}-2\right)\right] + \frac{L_2^2 c_2 \|K\|_2^4}{4}\frac{s_2}{16}\left[m_2^2 + 8\log\left(\frac{d(d-1)}{s_2}-2\right)\right] \\ &\le \frac{L_1^2 c_1 \|K\|_2^2}{4} a_n + \frac{L_2^2 c_2 \|K\|_2^4}{4} b_n \end{aligned}$$

情形3： $n^{1/(2\beta_1+1)} < 8\log\left(\frac{2d}{s_1}-2\right)$ **和** $n^{\frac{1}{2\beta_2+2}} \ge 8\log\left(\frac{d(d-1)}{s_2}-2\right)$

这里 $m_1 = \left(8\log\left(\frac{2d}{s_1}-2\right)/n\right)^{-1/2\beta_1}$ 和 $m_2 = n^{\frac{1}{2\beta_2+2}}$ 由此得到

$$\begin{aligned} n\delta^2 &= \frac{L_1^2 c_1 \|K\|_2^2}{4}\frac{s_1}{16} 8\log\left(\frac{2d}{s_1}-2\right) + \frac{L_2^2 c_2 \|K\|_2^4}{4}\frac{s_2}{16} n^{\frac{2}{2\beta_2+2}} \\ &\le \frac{L_1^2 c_1 \|K\|_2^2}{4}\frac{s_1}{16}\left[m_1 + 8\log\left(\frac{2d}{s_1}-2\right)\right] + \frac{L_2^2 c_2 \|K\|_2^4}{4}\frac{s_2}{16}\left[m_2^2 + 8\log\left(\frac{d(d-1)}{s_2}-2\right)\right] \\ &\le \frac{L_1^2 c_1 \|K\|_2^2}{4} a_n + \frac{L_2^2 c_2 \|K\|_2^4}{4} b_n. \end{aligned}$$

情形4： $n^{1/(2\beta_1+1)} < 8\log\left(\frac{2d}{s_1}-2\right)$ **和** $n^{\frac{1}{2\beta_2+2}} < 8\log\left(\frac{d(d-1)}{s_2}-2\right)$

Here $m_1 = \left(8\log\left(\frac{2d}{s_1}-2\right)/n\right)^{-1/2\beta_1}$ and $m_2 = \left(8\log\left(\frac{d(d-1)}{s_2}-2\right)/n\right)^{-1/2\beta_2}$ yielding

$$\begin{aligned} n\delta^2 &= \frac{L_1^2 c_1 \|K\|_2^2}{4}\frac{s_1}{16} 8\log\left(\frac{2d}{s_1}-2\right) + \frac{L_2^2 c_2 \|K\|_2^4}{4}\frac{s_2}{16} 8\log\left(\frac{d(d-1)}{s_2}-2\right) \\ &\le \frac{L_1^2 c_1 \|K\|_2^2}{4}\frac{s_1}{16}\left[m_1 + 8\log\left(\frac{2d}{s_1}-2\right)\right] + \frac{L_2^2 c_2 \|K\|_2^4}{4}\frac{s_2}{16}\left[m_2^2 + 8\log\left(\frac{d(d-1)}{s_2}-2\right)\right] \\ &\le \frac{L_1^2 c_1 \|K\|_2^2}{4} a_n + \frac{L_2^2 c_2 \|K\|_2^4}{4} b_n. \end{aligned}$$

因此，在四种情况下，分子和分母的阶相同。因此，

$$n\delta^2 \le \frac{L_1^2 c_1 \|K\|_2^2}{4} a_n + \frac{L_2^2 c_2 \|K\|_2^4}{4} b_n\,.$$

将其代入方程（A.38），可得：

$$\begin{aligned} \inf_{\hat{f}} \sup_{f, P_X} \mathbb{E}_{P_X}[(\hat{f}(X) - f(X))^2] &\ge \delta^2 \left(1 - \frac{32\left(L_1^2 c_1 \|K\|_2^2\, a_n + L_2^2 c_2 \|K\|_2^4\, b_n\right) + \log 2}{2(2\log\frac{1}{2} + a_n + b_n)}\right) \\ &\ge \delta^2 \left(1 - \frac{33\left(L_1^2 c_1 \|K\|_2^2\, a_n + L_2^2 c_2 \|K\|_2^4\, b_n\right)}{3(a_n + b_n)}\right) \end{aligned}$$

$$\geq \delta^2 \left(1 - \frac{11\left(L_1^2 c_1 \|K\|_2^2 \, a_n + L_2^2 c_2 \|K\|_2^4 \, b_n\right)}{(a_n + b_n)}\right).$$

(A.41)

现在选择 c_1, c_2 使得 $(L_1^2 c_1 \|K\|_2^2) \vee (L_2^2 c_2 \|K\|_2^4) \leq 1/22$。于是有:

$$\begin{aligned}
&\inf_{\hat{f}} \sup_{f, P_X} \mathbb{E}_{P_X}[(\hat{f}(X) - f(X))^2] \\
&\geq \frac{1}{2}\delta^2 \geq \frac{L_1^2 \|K\|_2^2}{128} m_1^{-2\beta_1} s_1 + \frac{L_2^2 \|K\|_2^4}{128} m_2^{-2\beta_2} s_2 \\
&= \frac{L_1^2 \|K\|_2^2 c_1}{128} s_1 \left(n^{-\frac{2\beta_1}{2\beta_1+1}} \vee 8\frac{\log\left(\frac{2d}{s_1} - 2\right)}{n}\right) + \frac{L_2^2 \|K\|_2^4 c_2}{128} s_2 \left(n^{-\frac{2\beta_2}{2\beta_2+2}} \vee 8\frac{\log\left(\frac{d(d-1)}{s_2} - 2\right)}{n}\right).
\end{aligned}$$

证明至此完成。