arXiv:2304.09305v1 [stat.ME] 18 Apr 2023

高维　　多分类仅存在数据分类

Lili Zheng*、Garvesh Raskutti†

2023年4月20日

摘要

正类与未标记（PU）数据分类经常出现在生物信息学、临床数据和生态学研究中，因为收集负样本的成本可能高得令人难以承受。尽管先前关于正例-未标记数据的研究主要关注二分类，但本文考虑多个正标签这一实践中重要且常见的情形。我们提出了一个适用于无序标签的多项式-*PU* 模型，以及一个适用于有序标签的序数-*PU* 模型。我们提出基于近端梯度下降的算法，以最小化$\ell_{1,\ 2}$-惩罚对数似然损失，并保证其收敛至非凸目标函数的驻点。尽管仅存在数据和多类别标签所引起的非凸性带来了挑战，但我们证明，在高维情形下，真实参数邻域内驻点的统计误差界。通过仔细刻画该邻域内对数似然损失的景观，这一结果得以实现。此外，模拟和两个真实数据实验表明，与基线方法相比，我们的算法具有更好的经验表现。

关键词— 正例与未标记数据、多分类、非凸优化、高维有序回归、高维多项式回归

1 引言

正类与未标记（PU）数据，也称为仅存在数据，广泛出现于许多应用中，例如生物信息学 [Elkan 和 Noto、 2008], 物种分布的生态建模 [Ward 等人、 2009],文本挖掘 [Liu 等人， 2003], 其中只有一部分正类响应带有标签，其余均为未标记数据。例如，在生物系统工程中 [Romero 等人，2015年，Song 等人，2021], 筛选步骤可能会输出功能性蛋白质序列样本，而所有其他蛋白质序列均为未标记（其中一些具有功能，另一些则不具有功能）；在研究物种的地理分布时 [Ward 等人， 2009], 观测到的物种肯定存在，而未观测到的物种不一定不存在。在这些应用中，主要目标是基于大量特征/协变量进行分类和变量选择，例如，学习一个模型，以预测特定蛋白质序列在高温下是否具有功能。然而，直接将未标记样本视为负例会导致显著偏差和不理想的预测性能，因此需要针对仅存在数据的专门 PU 学习方法。

*电气与计算机工程系，Rice University，美国德克萨斯州休斯顿
†统计学系、计算机科学系，威斯康星大学麦迪逊分校，美国威斯康星州麦迪逊

先前研究 [Elkan 和 Noto，2008年，Ward 等人，2009年，Song 和 Raskutti， 2020] 提出了应对这一问题的方法，重点关注二分类情形（一个正类和一个负类）。具体而言，Ward 等人 [2009], Song 和 Raskutti [2020] 提出了期望最大化（EM）算法和主化-最小化（MM）算法，旨在最小化观测到的正类与未标记数据的惩罚非凸对数似然，并在高维情形下提供统计和优化保证 [Song 和 Raskutti， 2020]。然而，许多 PU 学习问题涉及多个类别。例如，在蛋白质工程中，科学家关心某些蛋白质序列在多个温度水平下是否具有功能 [Fields，2001]；在疾病诊断中，医生可以给出一个分数，表示患者疾病发展的严重程度/阶段，或疾病类型 [Bertens 等人，2016，Sivapriya 等人， 2015]；在推荐系统中，用户可能会以多种方式与推荐内容互动。考虑多分类能够大幅提升模型容量，更有效地利用现有数据，并产生信息量更丰富的预测。另一方面，现有统计方法要么假设所有观测都反映真实标签，从而导致偏差；要么只考虑二元响应，因此只能提供粗略的预测。这就需要开发能够同时考虑仅存在数据的特性和多类别标签的新模型与算法。

然而，处理 PU 设定中的多分类是一个并非易事的任务，并不是对以往二分类工作的简单扩展。当多个类别表示物种时，它们可以是无序的；而当它们编码每个蛋白质序列能够耐受的温度水平时，则可以是有序的，这需要采用两种不同的建模方法，而这两种方法各自都面临独特的技术挑战。在本文中，我们借鉴多项式和序数逻辑回归中的思想，并扩展了此前 [Song 和 Raskutti 2020]关于具有二元响应的正类与未标记数据的高维分类的工作，以解决这一问题。我们分别针对无序和有序分类响应提出了两个模型（多项式 PU 模型和序数 PU 模型）以及相应的算法。正如我们将在后文中看到的，PU 数据的对数似然损失的非凸景观在多分类设定下变得更加复杂。尽管如此，我们通过仔细刻画对数似然损失的景观，为这两种模型下的方法提供了理论保证，并通过模拟和真实数据示例展示了它们在实证方面的优点。

1.1 问题表述

符号说明： 对于任意矩阵或张量 $A \in \mathbb{R}^{p_1 \times \cdots \times p_k}$和 $1 \le l \le k$，令 $\|A\|_2 = (\sum_{i_1,\ldots,i_k} A^2_{i_1,\ldots,i_k})_{12}$.对于任意两个维度相同的张量 A 和 B，令 $\langle A, B \rangle$ 表示 A 和 B的欧几里得内积。对于任意矩阵 $A \in \mathbb{R}^{p_1 \times p_2}$，令 $\lambda_{\min}(A)$表示 A的最小特征值，并令 $A_j \in \mathbb{R}^{p_2}$ 为 A的第 j列。令 $\mathbb{1}_{\{E\}} = \begin{cases} 1, \text{若 } E \text{ true } 0, \text{否则} \end{cases}$ 表示指示函数。

在本节中，我们提出分别适用于无序和有序分类响应的多项式 PU 模型和序数 PU 模型。对于这两个模型，设 $x \in \mathbb{R}^p$ 为协变量，$y \in \{0, 1, \ldots, K\}$ 为未观测的真实分类响应，其中包含 K 个正类别。我们假设只能观测到协变量 x 和标签 $z \in \{0, 1, \ldots, K\}$，后者是真实响应$y$的噪声观测。如果标签为 $z \neq 0$，则它反映真实响应 $y = z$；否则，该样本为未标记样本，未知响应 y 可以取自 $\{0, 1, \ldots, K\}$ 中的任意值。在给定响应 y 的条件下，标签 z 的分配与协变量 x 条件独立：也就是说，真实响应 y 在观测中以仅取决于 y 的概率随机缺失，这是缺失数据问题文献中常见的假设（ [Little 和 Rubin， 2019]）。

更具体地说，为了对给定 y 的 z 条件分布进行建模，我们考虑两种常见设置：病例对照方法（ [Lancaster 和 Imbens，1996年，Ward 等人，2009年，Song 和 Raskutti， 2020] ）和单训练集情形（ [Elkan 和 Noto， 2008]）。**病例对照**设置适用于未标记样本和正样本分别抽取的情况，其中一个样本来自总体，另一个样本来自正类总体。这种设置在生物技术应用中较为常见（ [Romero 等人， 2015]）。更具体地说，在病例对照设置下，我们假设 n_u 未标记样本来自原始总体，而每个类别中的 n_i 正样本来自响应为 i 的总体 $1 \le i \le K$，总样本量为 $n = n_u + \sum_{j=1}^{K} n_j$。我们引入另一个随机变量 s，将其作为样本是否被选中的指示变量；也就是说，仅当相关指示变量 $s = 1$ 时，我们才观测到样本（x， z）。另外，设 π_1，……，$\pi_k \in (0, 1)$ 为从原始总体中观测到响应为 j:$\pi_j = P(y = j)$ 的样本的概率。在这种病例对照设置下，可以如下计算给定 y 和 $s = 1$ 的 z 条件分布：当 $y = k$ 和 $s = 1$ 时，我们有 $z \in \{0, k\}$，并且

$$\mathbb{P}(z = k|y = k, s = 1, x) = \frac{n_k}{n_k + \pi_k n_u}, \quad \mathbb{P}(z = 0|y = k, s = 1, x) = \frac{\pi_k n_u}{n_k + \pi_k n_u}; \tag{1}$$

当 $y = 0$ 和 $s = 1$ 时， $\mathbb{P}(z = 0|y = 0, \ s = 1, \ x) = 1$。有关此条件分布的更详细推导，请参见附录 E。

对于**单训练集**情形，我们首先从总体中随机抽取所有 n 个样本，抽取过程独立于真实响应 y 或协变量 x。随后，每个正样本以某个常数概率被设为未标记：

$$\mathbb{P}(z = k|y = k, s = 1, x) = \pi_k^{\mathrm{st}}, \quad \mathbb{P}(z = 0|y = k, s = 1, x) = 1 - \pi_k^{\mathrm{st}}, \tag{2}$$

其中 $\pi^{\mathrm{st}} \in (0, 1)^K$，我们用上标“st”表示单训练集情形。

下文将详细讨论本文所考虑的多项式 PU 模型和序数 PU 模型中，给定协变量 x时真实响应 y 的分布形式。

多项式 PU 模型：该模型适用于 K 个正类别无序的情况。令 $\Theta^* \in \mathbb{R}^{p \times K}$ 和 $b^* \in \mathbb{R}^K$ 分别为回归参数和偏移参数。我们假设

$$x \sim \mathbb{P}_X, \quad \mathbb{P}(y = j|x) = \begin{cases} \frac{e^{x^\top \Theta_j^* + b_j^*}}{1 + \sum_{k=1}^{K} e^{x^\top \Theta_k^* + b_k^*}}, & j > 0, \\ \frac{1}{1 + \sum_{k=1}^{K} e^{x^\top \Theta_k^* + b_k^*}}, & j = 0, \end{cases} \tag{3}$$

其中， $\Theta_j^* \in \mathbb{R}^p$ 是 Θ^*的第 j列， $b_j^* \in \mathbb{R}$ 是 b^*的第 j个条目。假设 $\{(x_i, y_i)\}_{i=1}^n$ 是根据该模型生成的独立同分布样本，并且每个 z_i 都是根据病例对照情形或单训练集情形生成的，我们希望基于 $\{(x_i, z_i)\}_{i=1}^n$ 估计未知参数 Θ^*, b^* 。

序数-PU 模型：当类别标签 0、1、……、K 具有顺序时，相较于多项式模型，采用序数建模方法更为合适。当类别标签是某个潜在变量的粗略离散化时，就可能出现这种情况，例如蛋白质能够保持具有功能的最高温度等级 [Romero 等人， 2015], 或一个美国家庭的收入水平 [McCullagh，1980]。在这种情况下，我们考虑累积对数几率序数回归模型 [McCullagh，1980，Agresti，2010] 具体如下。设 $\beta^* \in \mathbb{R}^p$ 为回归参数，$\nu^* \in \mathbb{R}^K$ 为满足 $\nu_1^* < \nu_2^* < \cdots < \nu_K^*$的偏移参数。我们假设

$$x \sim \mathbb{P}_X, \quad \mathbb{P}(y < j|x) = \frac{1}{1 + e^{-(\nu_j^* - x^\top \beta^*)}}, 1 \le j \le K. \tag{4}$$

对 ν^* 施加序数约束可能会给估计过程带来困难，因此我们考虑以下重新参数化：令 $\theta^* \in \mathbb{R}^{p+K}$ 满足

$$\theta_j = \begin{cases} \beta_j^*, & 1 \le j \le p, \\ \nu_1^*, & j = p+1, \\ \nu_{j-p}^* - \nu_{j-p-1}^*, & j > p+1. \end{cases}$$

根据定义， θ^* 可以取 $\mathbb{R}^{p+1} \times (0, \infty)^{K-1}$中的值。我们假设 $\{(x_i, y_i)\}_{i=1}^n$ 是由该模型生成的独立同分布样本，且 z_i 遵循由（1）或（2）指定的条件分布；我们的目标是根据 $\{(x_i, z_i)\}_{i=1}^n$估计未知参数 $\theta^* \in \mathbb{R}^{p+1} \times (0, \infty)^{K-1}$。

高维情形： 在许多使用仅存在数据的实际应用中，协变量的数量可能远大于样本数量，因此我们考虑高维情形。我们假设回归参数具有稀疏性或组稀疏性，并在估计量中引入 ℓ_1 或$\ell_{1,2}$ 惩罚项。更多细节见第 2 节和第 3 节。

1.2 相关工作

已有大量关于仅存在数据分析的先前研究 [Ward 等人，2009年，Elkan 和 Noto，2008年，Liu 等人，2003年，Du Plessis 等人，2015年，Song and Raskutti, 2020，Song 等人， 2020], 但这些研究均考虑二元标签，而非多项式或有序标签。具体而言，Ward 等人 [2009] 和 Song 和 Raskutti [2020] 与我们的研究最为相关；二者都考虑逻辑回归模型，该模型对应于我们模型的特殊情形，其中 $K = 1$，并假设采用病例对照方法。Ward 等人 [2009] 针对低维情形提出了期望最大化（EM）算法，而 Song 和 Raskutti [2020]针对高维情形提出了带有$\ell_{1,2}$ 惩罚项的最大化-主化（MM）算法。

考虑到高维情形，我们的工作与关于估计稀疏广义线性模型的先前研究密切相关 [Van de Geer 等人，2008年、Fan 等人，2010年、Kakade 等人，2010年、Friedman 等人，2010年、Tibshirani，1996年、Li 等人， 2020]。这些研究大多关注凸对数似然损失。而在我们的工作中，我们关注非凸对数似然损失，因此我们为惩罚损失的驻点而非全局极小值点提供统计保证。这种方法也被 Song 和 Raskutti [2020], Loh 等人 [2017], Loh 和 Wainwright [2013]采用。我们的证明的主要思想与 Song 和 Raskutti [2020], Loh 等人 [2017], 相似，而关键挑战在于为使用正例-未标记数据的多项式和有序对数似然损失建立受限强凸性。Xu 等人[2017] 也考虑了具有稀疏性或其他约束的广义线性模型的估计，其中也使用 MM 框架来处理损失函数的非凸性。然而，他们提出的到集合距离惩罚诱导了非凸性，而不像本文这样由非凸对数似然损失诱导非凸性。

我们的工作也与关于高维多项式和有序逻辑回归的既有研究密切相关 [Wurm 等人，2017，Archer 等人，2014，Krishnapuram 等人， 2005]。这些研究考虑的是完全观测的数据；而对于正类与未标记数据，应用这些方法的唯一方式是将未标记样本视为负例。我们在正类与未标记数据集的数值实验中采用了这一简单方法作为基线方法，结果表明，我们的算法在实证表现上具有显著优势。

结构安排 本文其余部分安排如下：我们将在第 2 节提出用于估计多项 PU 模型和有序 PU 模型的算法。随后，我们将讨论这两种算法的收敛保证

，并从优化和统计视角进行分析。第5节和第6节将介绍一系列基于合成数据和真实数据集开展的实证实验。最后，我们将在第 7 节进行讨论并总结全文。

2 提出的算法

本节将在高维情形下，介绍适用于多项式 PU 模型和序数 PU 模型的相应估计算法。为简便起见，我们将重点讨论用于建模 $\mathbb{P}(z_i|y_i)$ 的病例对照方法，并将单一训练集情形下的算法留待附录 F 介绍。具体而言，我们推导了两个模型的观测对数似然损失的精确形式，并提出应用近端梯度下降（PGD）算法来最小化惩罚对数似然损失。另一种方法是使用正则化 EM 算法，其两个模型的详细公式以及收敛性质也可见附录 D。鉴于 PGD 算法具有较高的计算效率，主论文将重点介绍该算法。

2.1 多项式 PU 模型

在介绍多项式 PU 模型的详细算法之前，我们首先在以下引理中写出观测数据 $\{(x_i,\ z_i,\ s_i=1)\}$ 和完整数据 $\{(x_i,\ y_i,\ z_i,\ s_i=1)\}$ 的对数似然函数，其中 s_i 是第 1.1 节中引入的选择指示变量。

引理 2.1. 观测到的仅存在数据的对数似然函数 $\{x_i, z_i, s_i=1\}_{i=1}^n$ *is*

$$\begin{aligned}&\log L^{\mathrm{MN}}(\Theta, b; \{(x_i, z_i, s_i=1)\}_{i=1}^n)\\ =&\sum_{i=1}^n\left[\sum_{k=1}^K \mathbb{1}_{\{z_i=k\}}[f(\Theta^\top x_i+b)]_k-\log\left(1+\sum_{k=1}^K e^{[f(\Theta^\top x_i+b)]_k}\right)\right],\end{aligned}$$

其中 $f\colon\ \mathbb{R}^K\to\mathbb{R}^K$ 满足 $(f(u))_k=u_k+\log\frac{n_k}{\pi_k n_u}-\log(1+\sum_{j=1}^K e^{u_j})$。完整数据的对数似然函数 $\{x_i, y_i, z_i, s_i=1\}_{i=1}^n$ 为

$$\begin{aligned}&\log L_f^{\mathrm{MN}}(\Theta, b; \{(x_i, y_i, z_i, s_i=1)\}_{i=1}^n)\\ =&\sum_{i=1}^n\Bigg[\sum_{k=1}^K \mathbb{1}_{\{y_i=k\}}(x_i^\top\Theta_k+b_k)-\log\left(1+\sum_{k=1}^K(1+\frac{n_k}{\pi_k n_u})e^{x_i^\top\Theta_k+b_k}\right)\\ &+\sum_{k=1}^K \mathbb{1}_{\{y_i=z_i=k\}}\log\frac{n_k}{\pi_k n_u}\Bigg],\end{aligned}$$

其中 Θ_k 是 Θ 的第 k 列。

引理 2.1 的证明见附录 E。对观测 PU 对数似然与完整对数似然的比较表明，函数 $f(\cdot)$ 反映了仅存在数据的特性。

用于求解惩罚最大似然估计的 PGD：在高维情形下，一种自然的模型参数估计思路是最小化惩罚对数似然损失函数，以促进回归参数的稀疏性。具体而言，设

$$\mathcal{L}_n^{\mathrm{MN}}(\Theta, b)=-\frac{1}{n}\log L^{\mathrm{MN}}(\Theta, b; \{(x_i, z_i, s_i=1)\}_{i=1}^n),$$

那么我们希望最小化

$$\mathcal{F}_n^{\mathrm{MN}}(\Theta, b) = \mathcal{L}_n^{\mathrm{MN}}(\Theta, b) + P_\lambda^{\mathrm{MN}}(\Theta),$$

其中，我们将 $P_\lambda^{\mathrm{MN}}(\Theta) = \lambda\|\Theta\|_{\omega,2,1}$ 设为组稀疏惩罚，这可以体现生物医学或遗传学应用中常见的协变量潜在组结构。具体而言，

$$\|\Theta\|_{\omega,2,1} = \sum_{j=1}^{J} \omega_j \|\Theta_{\mathcal{G}_j}\|_2. \tag{5}$$

这里，$\mathcal{G}_1, \ldots, \mathcal{G}_J$ 是满足 $\cup_j \mathcal{G}_j = [p] \times [K]$、$\mathcal{G}_j = \mathcal{R}_j \times \mathcal{C}_j$ 的不重叠组，其中某些 $\mathcal{R}_j \subset [p]$、$\mathcal{C}_j \subset [K]$ 的大小分别为 $r_j = |\mathcal{R}_j|$、$c_j = |\mathcal{C}_j|$。$\Theta_{\mathcal{G}_j}$ 是 Θ 的一个子矩阵，其行由 $\mathcal{R}_j$ 索引，列由 $\mathcal{C}_j$ 索引。当 $J = 1$ 且 $\mathcal{G}_1 = [p] \times [K]$ 时，$P_\lambda^{\mathrm{MN}}(\Theta)$ 就成为 ℓ_1 惩罚。为了最小化 $\mathcal{F}_n(\Theta, b)$，我们可以直接应用带有 $\ell_{1,2}$ 惩罚的近端梯度下降算法 [Wright 等人， 2009] 。具体而言，设 $[\Theta^m;\ b^m] \in \mathbb{R}^{(p+1)\times K}$ 由 Θ^m 和 b^m 按行拼接而成，则可在第 m 次迭代时按如下方式更新参数:

$$\begin{aligned}
&[\Theta^{m+1}; b^{m+1}] \\
=& \underset{[\Theta;b]\in\mathbb{R}^{(p+1)\times K}}{\arg\min} \left\{ \eta_m P_\lambda^{\mathrm{MN}}(\Theta) + \frac{1}{2}\|[\Theta; b] - ([\Theta^m; b^m] - \eta_m \nabla \mathcal{L}_n^{\mathrm{MN}}(\Theta^m, b^m))\|_2^2 \right\},
\end{aligned} \tag{6}$$

其中 η_m 是步长。我们可以选择初始值 $[\Theta^0; b^0]$ ，使其对应的损失函数不大于任何仅截距模型的损失: $\mathcal{F}_n^{\mathrm{MN}}(\Theta^0,\ b^0) \le \min_b \mathcal{F}_n^{\mathrm{MN}}(0_{p\times K},\ b)$，这可通过$[0_{p\times K}; \arg\min_b \mathcal{F}_n^{\mathrm{MN}}(0_{p\times K},\ b)]$满足。

2.2 有序 PU 模型

与多项式 PU 模型类似，我们首先在以下引理中给出对数似然函数。

引理 2.2. 观测到的仅存在数据 $\{x_i, z_i, s_i = 1\}_{i=1}^n$ 的对数似然函数为

$$\begin{aligned}
&\log L^{\mathrm{ON}}(\theta; \{(x_i, z_i, s_i = 1)\}_{i=1}^n) \\
=&\sum_{i=1}^{n} \left[\sum_{k=1}^{K} \mathbb{1}_{\{z_i=k\}} [f(\log r(x_i, \theta))]_k - \log\left(1 + \sum_{k=1}^{K} e^{[f(\log r(x_i,\theta))]_k}\right)\right],
\end{aligned}$$

其中 f 在引理 *2.1* 中定义，且 $r:\ \mathbb{R}^p \times \mathbb{R}^{p+K} \to \mathbb{R}^K$ 满足

$$\begin{aligned}
&r_j(x, \theta) \\
=&(1 + e^{x^\top \theta_{1:p} - \theta_{p+1}}) \left[(1 + e^{x^\top \theta_{1:p} - \sum_{l=1}^{j+1} \theta_{p+l}})^{-1} - (1 + e^{x^\top \theta_{1:p} - \sum_{l=1}^{j} \theta_{p+l}})^{-1} \right],
\end{aligned} \tag{7}$$

对于 $1 \le j < K$ 和 r_K（x、θ）$= (1 + e^{x^\top \theta_{1:p} - \theta_{p+1}})\ \left[1 - \ (1 + e^{x^\top \theta_{1:p} - \sum_{l=1}^{j} \theta_{p+l}})\ ^{-1}\right]$。完整数据 $\{x_i, y_i, z_i, s_i = 1\}_{i=1}^n$ 的对数似然函数为

$$\begin{aligned}
&\log L_f^{\mathrm{ON}}(\theta; \{(x_i, y_i, z_i, s_i = 1)\}_{i=1}^n) \\
=&\sum_{i=1}^{n} \left[\sum_{k=1}^{K} \mathbb{1}_{\{y_i=k\}} \log r_k(x_i, \theta) - \log\left(1 + \sum_{k=1}^{K} (1 + \frac{n_k}{\pi_k n_u}) r_k(x_i, \theta)\right)\right. \\
&\left. + \sum_{k=1}^{K} \mathbb{1}_{\{y_i=z_i=k\}} \log \frac{n_k}{\pi_k n_u} \right].
\end{aligned}$$

引理 2.2 的证明见附录 E。与引理 2.1 中多项式 PU 模型的对数似然相比，引理 2.2 中唯一的区别是将 $\Theta^\top x_i + b \in \mathbb{R}^K$ 替换为 $\log\ r(x_i、\ \theta)$。这种联系背后的原因是以下事实：在多项式 PU 模型和序数 PU 模型下，$\Theta_k^\top x_i + b_k$ 和 $\log\ r_k(x_i、\ \theta)$ 分别是 $\mathbb{P}(y_i = k|x_i)$ 与 $\mathbb{P}(y_i = 0|x_i)$ 之间的对数比。

用于求解惩罚最大似然估计的投影梯度下降：设

$$\mathcal{L}_n^{\mathrm{ON}}(\theta) = -\frac{1}{n}\log L^{\mathrm{ON}}(\theta; \{(x_i, z_i, s_i = 1)\}_{i=1}^n).$$

与多项情形类似，我们可以通过在惩罚损失函数 $\mathcal{F}_n^{\mathrm{ON}}(\theta) = \mathcal{L}_n^{\mathrm{ON}}(\theta) + P_\lambda^{\mathrm{ON}}(\theta_{1:p})$ 上应用近端梯度下降算法来估计 θ^*，其中 $P_\lambda^{\mathrm{ON}}(\theta_{1:p}) = \lambda\|\theta_{1:p}\|_{\omega,\ 2,\ 1} = \lambda\sum_{j=1}^J \omega_j\|\theta_{\mathcal{G}_j}\|_2$，其不相交的分组为 $\mathcal{G}_1$，……，$\mathcal{G}_J \subset [p]$，满足 $\cup_j \mathcal{G}_j = [p]$。在每次迭代 m 中，我们按如下方式更新参数：

$$\theta^{m+1} = \underset{\theta\in\mathbb{R}^{p+K}}{\arg\min}\left\{\eta_m P_\lambda^{\mathrm{ON}}(\theta_{1:p}) + \frac{1}{2}\|\theta - (\theta^m - \eta_m\nabla\mathcal{L}_n^{\mathrm{ON}}(\theta^m))\|_2^2\right\}, \tag{8}$$

其中 η_m 为步长。我们选择初始值 θ^0，使其损失函数不大于任何仅截距模型的损失：$\mathcal{F}_n^{\mathrm{ON}}(\theta^0) \le \min_{\theta_{1:p}=0_{p\times 1}} \mathcal{F}_n^{\mathrm{ON}}(\theta)$，一个简单的选择是 $\theta^0 = \arg\min_{\theta_{1:p}=0_{p\times 1}} \mathcal{F}_n^{\mathrm{ON}}(\theta)$。

3 理论保证

本节针对病例对照设置，为第 2 节提出的算法给出优化和统计保证。我们还将在附录 F 中简要介绍这些算法在单训练集情形下的理论性质。

3.1 算法收敛性

我们首先证明，将投影梯度下降算法用于最小化惩罚对数损失函数时，该算法会收敛到一个驻点。命题 3.1 针对多项式 PU 模型；对于序数 PU 模型，我们也有类似的收敛保证，详见附录 D。

命题 3.1（多项式 PU 模型算法的收敛性）。若参数迭代序列$\{\Theta^m, b^m\}_m$ 由 *PGD* 更新规则(6) 生成，且步长选择适当，则其满足以下性质：

(i) 序列 $\{(\Theta^m, b^m)\}_m$ 至少有一个极限点。*(ii)* 存在 $R^{\mathrm{MN}} > 0$，使得 $\{(\Theta^m, b^m)\}_m$ 的所有极限点都属于 Γ^{MN}，即优化问题 $\min_{\|\Theta\|_2+\|b\|_2\le R^{\mathrm{MN}}} \mathcal{F}_n^{\mathrm{MN}}(\Theta, b)$ 的一阶驻点集合。*(iii)* 函数值序列 $\{\mathcal{F}_n^{\mathrm{MN}}(\Theta^m, b^m)\}_m$ 单调不增；若 $(\Theta^m, b^m) \notin \Gamma^{\mathrm{MN}}$，则 $\mathcal{F}_n^{\mathrm{MN}}(\Theta^{m+1}, b^{m+1}) < \mathcal{F}_n^{\mathrm{MN}}(\Theta^m, b^m)$。此外，存在 $(\widetilde{\Theta}, \widetilde{b}) \in \Gamma^{\mathrm{MN}}$，使得 $\{\mathcal{F}_n^{\mathrm{MN}}(\Theta^m, b^m)\}_m$ 单调收敛至 $\mathcal{F}_n^{\mathrm{MN}}(\widetilde{\Theta}, \widetilde{b})$。

3.2 驻点的统计理论：多项式 PU 模型

在本节中，我们将为 $\mathcal{F}_n^{\mathrm{MN}}(\Theta, b) = \mathcal{L}_n^{\mathrm{MN}}(\Theta, b) + P_\lambda^{\mathrm{MN}}(\Theta)$ 的驻点提供理论保证。为简单起见，我们假设偏移参数 $b = 0_{K\times 1}$ 为零，并给出惩罚损失 $\mathcal{F}_n(\Theta, 0_{K\times 1})$ 在 Θ^* 周围某个可行域内任意驻点 Θ 的统计性质：$\Theta^* + \mathbb{B}_{1,\infty}(R_0)$，其中 $\mathbb{B}_{1,\infty}(R_0) = \{\Delta : \max_{1\le i\le K}\|\Delta_{:,i}\|_1 \le R_0\}$，而关于 $R_0 > 0$ 的条件将在假设 3.3 中说明。

多项式 PU 模型理论的关键量： 在呈现我们的理论结果之前，我们首先定义一些关键量。最大组大小参数定义为$m = \max_{1\le j\le J} c_j r_j$，其中 c_j、 r_j 是与$\ell_{1,2}$ 惩罚中的每个组 $\mathcal{G}_j$ 相关联的组大小参数。令 $R^* = \max_{1\le i\le K} \|\Theta^*_{:,i}\|_1$ 为 Θ^*的有界性参数。此外，令 $S = \{j : \Theta^*_{\mathcal{G}_j, c_j} \neq 0\}$ 以及$s = |S|$。

为推导我们的理论保证，需要以下假设。

假设 3.1 (次高斯协变量)。 $\{x_i \in \mathbb{R}^p\}_{i=1}^n$ 是相互独立、均值为零的次高斯向量，其次高斯参数为 σ，协方差矩阵为 Σ，且满足 $\sigma^2 \le C$ 最小值$\{\lambda_{\min}(\Sigma)\ 1\}$。同时，$_1|x_{ij}| \le C_x$，其中某个常数 $0 < C_x \le C\lambda^2$最小值(Σ)。

假设 3.1 中的 x_i 次高斯条件在高维统计学文献中很常见，且标准高斯向量满足该条件。假设 x_i 的每个元素都有界，以确保对于可行域中的任意 Θ， $x_i^\top\Theta$ 都有界，从而使损失函数能够适当地集中。此外，我们假设 $\sigma^2 \le C\{\lambda_{\min}(\Sigma)\ 1\}$，以便以高概率证明受限特征值条件。

假设 3.2 (速率条件)。 组稀疏惩罚中的组权重 ω 满足 $c \le$最小值$_j\,\omega_j \le$最大值$_j\,\omega_j \le C\sqrt{\ }^{n}$对数 $J+m$；有标签数据和未标记数据的样本量满足 $c \le$最小值$_j \frac{n_j}{\pi_j n_u} \le C$；$\lambda_{\max}(\Sigma) \ge c$、 K 对数 $J \le Cn$ 以及 对数 $J \ge cm$对数对数 n。

假设 3.2 中的速率条件是标准的，与高维统计学和 PU 学习领域的既有文献相当，例如 Song 和 Raskutti, 2020 年、Loh 和 Wainwright, 2013]。

我们的统计理论所依赖的另一个关键假设涉及优化问题的可行域：$\Theta \in \Theta^* + \mathbb{B}_{1,\infty}(R_0)$。正如第 4 节将更清楚地解释的那样，在 Θ 处评估的非凸损失函数可以分解为一个受限凸项和一个非凸零均值项；后者以高概率实现良好的集中。受限凸项在 Θ^* 的某个特定邻域内具有曲率下界，这促使我们对邻域半径 R_0 施加如下约束。

假设 3.3。

$$R_0 \le \frac{\min_j \frac{n_j}{\pi_j n_u} e^{-C_x R^*}}{4C_x(1+\max_j \frac{n_j}{\pi_j n_u})^2(1+1.1Ke^{C_x R^*})^3}.$$

假设 3.3 中对可行域半径 R_0 的约束取决于病例对照研究设计： $\frac{n_j}{\pi_j n_u}$ 相对于 $1 \le j \le K$ 越均衡，可行集就越大；它也取决于真实模型参数 Θ^*的大小： R^* 越小，可行集也越大。后一种关系有直观的解释：当所有标签均可观测时，原始多项式对数似然损失在零点附近的曲率更大，因此，多项式 PU 损失在零点附近的非凸性也相应较弱。尽管假设 3.3 可能显得略为严格，但正如我们将在第5节和第6节中展示的那样，在合成数据和真实数据实验中，采用最佳仅截距模型进行初始化通常都能取得良好效果。

在正式陈述理论结果之前，我们还定义函数

$$h^{\mathrm{MN}}(R) = \frac{e^{-C_x(R+R^*)} \min_j \frac{n_j}{\pi_j n_u}}{(1+\max_j \frac{n_j}{\pi_j n_u})^2 \left(1+Ke^{C_x(R+R^*)}\right)^3} - 4C_x R. \tag{9}$$

正如我们将在证明中所示，函数 $h^{\mathrm{MN}}(R)$ 决定了对数似然受限强凸性的下界；只要假设 3.3 成立， $h^{\mathrm{MN}}(R_0)$ 就保证为正（参见引理 A.3 ）。

定理 3.1。 假设假设 *3.1* 至 *3.3* 成立，并且对于某个常数 $C>0$，$\lambda \geq C\lambda^{\frac{1}{2}}\max(\Sigma)(K+C_x R^*\sqrt{K})\sqrt{\log \frac{J+m}{n}}$。若 $\widehat{\Theta}$ 是$\min_{\Theta-\Theta^*\in\mathbb{B}_{1,\infty}(R_0)} \mathcal{F}_n^{\mathrm{MN}}(\Theta, 0_{K\times 1})$的一个驻点，则

$$\|\widehat{\Theta}-\Theta^*\|_2 \leq \frac{3\|\omega_S\|_2\lambda}{h^{\mathrm{MN}}(R_0)\lambda_{\min}(\Sigma)}, \quad \|\widehat{\Theta}-\Theta^*\|_{\omega,2,1} \leq \frac{12\|\omega_S\|_2^2\lambda}{h^{\mathrm{MN}}(R_0)\lambda_{\min}(\Sigma)}, \tag{10}$$

概率至少为 $1-\exp\{-c^{K^2(\log\ J+m)\max_j r_j}\}-\exp\{-c(\log\ J+m)\}$.

备注 3.1。 当 $\omega_j \asymp C$对所有 $1\leq j\leq J$均成立，且 $\lambda \asymp \lambda^{\frac{1}{2}}\max(\Sigma)(K+C_x R^*\sqrt{K})\sqrt{\log\frac{J+m}{n}}$时，定理 *3.1* 表明，高概率下 $\|\widehat{\Theta}-\Theta^*\|_2 = O(K\sqrt{s(\log\ J+m)}n)$以及$\|\widehat{\Theta}-\Theta^*\|_{\omega,2,1} = O(Ks\sqrt{\log\frac{J+m}{n}})$。特别地，当正标签的数量为 $K\leq C$ 时，该速率与 *Song* 和 *Raskutti* *[2020]* 中推导出的速率相同。

备注 3.2。 证明定理 *3.2* 的主要挑战在于展示，在 $\Theta^*+\mathbb{B}_{1,\infty}(R_0)$ 内，损失函数 $\mathcal{L}_n^{\mathrm{MN}}(\theta)$ 集中于一个受限强凸函数附近，参见引理 *A.2*。我们的证明基于 *Song* 和 *Raskutti* *[2020]*，所发展的技术，而考虑多个正标签需要更加细致的分析，例如引理 *A.4* 的证明。此外，由于 $K>1$，我们的证明依赖于 *Rademacher* 复杂度的向量压缩不等式 *[Maurer*，其中压缩函数的定义域为向量值定义域。关于非凸性问题以及我们如何处理这些问题的更详细讨论见第 *4.1* 节。

3.3 驻点的统计理论：有序-PU 模式

类似于第 3.2 节，下面我们将给出 $\theta-\theta^*$ 的误差界，其中 θ 是损失函数 $\mathcal{F}_n^{\mathrm{ON}}(\theta)$ 在 θ^* 周围某个区域内的任意驻点： $\theta-\theta^*\in S(R_0, r_0)$。这里的集合 $S(R_0, r_0)$ 定义如下:

$$S(R_0, r_0)=\{\Delta : \|\Delta_{1:p}\|_1, \|\Delta_{(p+1):(p+K)}\|_1 \leq R_0, \min_{2\leq j\leq K}\Delta_{p+j}\geq -r_0\}, \tag{11}$$

其中 R_0， $r_0>0$ 取决于真实参数，并将在假设 3.4 中具体说明。与多项式 PU 模型不同，此处的约束集 $S(R_0,\ r_0)$ 还对 $\Delta_{(p+2):(p+K)}$施加了下界，而不仅仅约束 Δ的范数。这是由于序数逻辑回归模型的特性：我们需要确保 $\theta_{p+j}>0$ 对于 $2\leq j\leq K$，从而使 $\mathbb{P}_{\hat{\theta}}(y=k|x)>0$ 对所有 $k>0$都成立。约束集 $S(R_0,\ r_0)$ 会在 R_0 或 r_0 增大时变大。

序数 PU 模型理论中的关键量： 最大组大小参数定义为 $m=\max_{1\leq j\leq J} g_j$。令 $R^*=\max\{\|\theta^*_{1:p}\|_1, \|\theta^*_{(p+1):(p+K)}\|_1\}$， $r^*=\min_{1\leq j\leq K-1}\theta^*_{p+j+1}>0$ 为 θ^*的有界性参数。此外，令 $S=\{j : \theta^*_{\mathcal{G}_j}\neq 0\}$ 和 $s=|S|$。

对于序数 PU 模型，我们还假设协变量服从次高斯分布，并且参数的缩放方式与多项式 PU 模型相同（假设 3.1 和假设 3.2）。此外，与假设 3.3 类似，我们需要可行域 $S(R_0,\ r_0)$ 满足以下条件，以确保对数似然损失函数在 $\theta^*+S(R_0,\ r_0)$ 内具有受限强凸性。

假设 3.4 （对 R_0、r_0 的约束）。 $0<r_0<r^*$ 和

$$R_0 \leq \min\left\{80, \min_j \frac{n_j}{\pi_j n_u}\right\}\frac{\left(1+e^{(C_x+1)(R^*+0.01)}\right)^{-6}}{512(C_x+1)K^3(1+\frac{2}{r^*-r_0})^3(2+\frac{2}{r^*-r_0})} \tag{12}$$

^条件 $0 < r_0 < r^*$ 确保对于 $j \geq 2$, $\theta_{p+j} > 0$ 。此外，假设 3.4 要求R_0 不能太大，这与多项式 PU 模型的假设 3.3 类似。约束 (12) 可以通过以下任一方式放宽：减少未标记样本的数量、降低真实参数的幅度 R^* ，或减小 r_0 的取值（即放宽对Δ_{p+j}、$2 \leq j \leq K$的下界约束）。我们还定义以下关键函数 h^{ON} : $\mathbb{R}^2 \to \mathbb{R}$ ，它决定了 $\theta \in \theta^* + S$（R、r）中对数似然损失的受限曲率：

$$h^{\mathrm{ON}}(R, r) = \max\left\{ \frac{(1 + e^{(C_x+1)(R^*+R)})^2}{(r^* - r)e^{(C_x+1)(R^*+R)}}, 1 + e^{(C_x+1)(R^*+R)} \right\}, \tag{13}$$

只要 $r < r^*$ 为正，.....

定理 3.2。 存在仅取决于 R^*、R_0、r^*、r_0、C_x 的 L_0、$\gamma_0 > 0$，使得若假设 *3.1*、*3.2* 和 *3.4* 成立，$\lambda \geq CL_0\lambda^{\frac{1}{2}}$最大值$(\Sigma)\sqrt{K}$(对数 $J+m)n$，~~那么对于~~ 最小值$_{\theta-\theta^* \in S(R_0,\ r_0)}$ $\mathcal{F}_n^{\mathrm{ON}}(\theta)$ 的任意驻点 $\widehat{\theta}$，我们有

$$\|\widehat{\theta} - \theta^*\|_2 \leq \frac{\max_j \frac{\pi_j n_u}{n_j} K}{\gamma_0 \min\{\lambda_{\min}(\Sigma), 1\}} \left(3\|\omega_S\|_2 \lambda + CL_0 K \sqrt{\frac{(\log J + m + \log K)}{n}} \right), \tag{14}$$

以至少 $1 - \exp\{-^{cK}{}_{\log_{C_x^2 m+1}{}^J}\} - \sout{3\exp\{-(\log J + m)\}}$ 的概率。

注释 3.3. 当对所有 $1 \leq j \leq J$ 都有 $\omega_j \asymp C$，且对某个常数 C 有 $K \leq J^C$，并且 $\lambda \asymp L_0\lambda^{\frac{1}{2}}max(\Sigma)\sqrt{K(\log}$ $_{J+m)n}$ 时，定理 *3.2* 表明，以高概率有 $\|\widehat{\theta} - \theta^*\|_2 = O(K^{\frac{3}{2}}(\sqrt{s} + \sqrt{K})\sqrt{\log_n^{J+m}})$。当 K 为常数时，我们得到与 *Song* 和 *Raskutti* *[2020]* 相同的误差界；否则，收敛速率关于 K 呈多项式级缩放。

注释 3.4。 在 ℓ_2 误差界中，这里出现的是因子 $\sqrt{s} + \sqrt{K}$，而非定理 *3.1* 中的 $\sqrt{s}$，因为除了估计 s-稀疏回归^参数 $\theta_{1:p}$之外，还要同时估计 K维偏移参数 $\theta^*_{(p+1):(p+K)}$ 。出于同样的原因，这里不提供 $\|\theta_{1:p} - \theta^*_{1:p}\|_{\omega,2,1}$ 的误差界，因为其形式更为复杂，尽管其证明与定理 *3.1* 类似。

注释 3.5。 正如我们的证明所揭示的， γ_0 和 L_0 的具体定义如下：

$$\gamma_0 = \frac{e^{2(C_x+1)(R^*+R_0)}}{32h^{\mathrm{ON}}(R_0, r_0)(1 + e^{(C_x+1)(R^*+R_0)})^4},$$

$$L_0 = (\frac{4}{3}(C_x + 1)R_0 + 1)h^{\mathrm{ON}}(R_0, r_0) + \frac{(C_x + 1)R_0}{4}(h^{\mathrm{ON}}(R_0, r_0))^2.$$

可以看出，较大的 R^*、R_0、r_0，以及较小的 r^*、$\lambda_{\min}(\Sigma)$ 会导致较大的 L_0 和较小的 γ_0，进而意味着 $\|\widehat{\theta} - \theta^*\|_2$的估计误差界更大。这是因为在较大的 $R^* + R_0$、较小的 $r^* - r_0$ 以及较小的 $\lambda_{\min}(\Sigma)$下， $S(R_0, r_0)$ 中的 $\mathcal{L}_n^{\mathrm{ON}}$曲率较小。

注释 3.6。 与定理 *3.1* 的证明类似，证明定理 *3.2* 的主要挑战在于证明损失函数 $\mathcal{L}_n^{\mathrm{ON}}(\theta)$ 在 $\theta^* + S(R_0, r_0)$ 内集中于一个受限强凸函数附近，参见引理 *4.4*。特别是，考虑具有多个正标签的序数 *PU* 模型，会比多项式 *PU* 模型产生复杂得多的对数似然损失：回顾一下，引理 *2.1* 中的对数比项 $\Theta^\top x_i + b$ 在引理 *2.2* 中变为 $\log r(x_i, \theta)$。这需要对序数对数似然损失的导数给出新的下界和上界（参见第 *4.2* 节）。

4 证明概要

遵循高维 M 估计量统计误差的传统证明策略 [Negahban 等人， 2009], 我们证明中的一个主要组成部分，是证明我们的对数似然损失在适当区域内具有受限强凸性。如前所述，主要的理论挑战在于处理由仅存在数据导致的观测对数似然损失的非凸景观。多类别标签放大了技术难点，尤其是在序数-PU 场景下。在本节中，我们将解释非凸性的来源，然后简要概述证明思路，说明我们处理这种非凸性的主要方法。完整证明均推迟至附录。

4.1 对数似然损失非凸性的说明

我们首先从多项式 PU 模型的对数似然损失入手进行说明，随后讨论两个模型的证明思路。为简化起见，假设多项式 PU 模型的偏移参数为零，则我们的对数似然损失具有以下形式:

$$\mathcal{L}_n^{\mathrm{MN}}(\Theta) = \frac{1}{n}\sum_{i=1}^{n}\left[A(f(\Theta^\top x_i)) - \delta_i^\top f(\Theta^\top x_i)\right] \tag{15}$$

其中， $A:\ \mathbb{R}^K \to \mathbb{R}$ 是满足 $A(u) = \log(1+\sum_{k=1}^{K} e^{u_k})$ 的多项式对数配分函数; $\delta_i \in \mathbb{R}^K$ 是样本 i 的观测标签对应的指示向量: $\delta_i = \begin{cases} e_k & z_i = k\text{，如果} \\ 0 & \text{，如果} \end{cases}$，而函数 $f:\ \mathbb{R}^K \to \mathbb{R}^K$ 满足

$$(f(u))_k = u_k + \underbrace{\log\frac{n_k}{\pi_k n_u} - \log(1+\sum_{j=1}^{K} e^{u_j})}_{\text{Effect of PU-modeling}}. \tag{16}$$

这里，仅阳性数据生成机制被编码在上述定义的非线性函数$f(\cdot)$的最后两项中，从而在损失函数中引入了非凸性。作为比较，如果不考虑 PU 学习场景，并假设所有观测标签都是真实标签，那么对数似然损失也将具有式（15）的形式，但其中的函数 $f(\cdot)$ 被恒等链接函数所替代。为了更具体地说明非凸性以及我们处理该问题的直观思路，我们将损失函数重写为 $\mathcal{L}_n^{\mathrm{MN}}(\Theta) = \frac{1}{n}\sum_{i=1}^{n} g_i(\Theta^\top x_i)$，其中 $g_i:\ \mathbb{R}^K \to \mathbb{R}$ 满足 $g_i(u) = A(f(u)) - \delta_i^\top f(u)$。为了展示 $\mathcal{L}_n^{\mathrm{MN}}(\Theta)$ 的一种受限强凸性，关键步骤之一是刻画 g_i 在 $\Theta^\top x_i$ 的潜在取值范围内的函数形态/曲率。

具体来说，设 $u_i = \Theta^\top x_i \in \mathbb{R}^K$、 $u_i^* = \Theta^{*\top} x_i \in \mathbb{R}^K$，那么我们希望证明如下形式的下界:

$$\langle \nabla g_i(u_i) - \nabla g_i(u_i^*), u_i - u_i^* \rangle \geq c\|u_i - u_i^*\|_2^2,$$

对于某个较小的常数 $c > 0$。另一方面，通过一些计算并应用中值定理，我们知道

$$\langle \nabla g_i(u_i) - \nabla g_i(u_i^*), u_i - u_i^* \rangle = \underbrace{(u_i - u_i^*)^\top G_i (u_i - u_i^*)}_{\mathrm{I}} + \underbrace{(u_i - u_i^*)^\top h_i}_{\mathrm{II}}, \tag{17}$$

其中 $h_i \in \mathbb{R}^K$ 是一个向量

$$h_i = (\nabla f(u_i) - \nabla f(u_i^*))^\top (\nabla A(f(u_i^*)) - \delta_i),$$

且 $G_i \in \mathbb{R}^{K\times K}$ 是一个矩阵:

$$G_i = \nabla f(u_i)^\top \nabla^2 A(f(u_i^t)) \nabla f(u_i^t),$$

对于某个介于 u_i 和 u_i^* 之间的 $u_i^t = tu_i + (1-t)u_i^*$，其中 $t \in (0,1)$。对于任意向量 $u \in \mathbb{R}^K$，梯度 $\nabla f(u) \in \mathbb{R}^{K\times K}$,$\nabla A(f(u)) \in \mathbb{R}^K$，海森矩阵 $\nabla^2 A(f(u)) \in \mathbb{R}^{K\times K}$。我们注意到，式 (17) 中的第一项 I 呈二次型，第二项 II 的均值为零：根据指数族随机变量的性质，有

$$\mathbb{E}[A(f(u_i^*)) - \delta_i | x_i] = 0.$$

因此，这启发了我们克服非凸性问题的主要证明思路：首先对I^{MN}给出下界，然后将$\mathrm{II}^{\mathrm{MN}}$ 控制在零附近，使其不会过度影响曲率。这种解决 PU 学习中非凸性问题的证明思路并非新颖；事实上，Song 和 Raskutti [2020] 也采用了类似思路，为使用正例-未标记数据进行二分类建立统计理论。然而，在二元情形[下，Song 和 Raskutti 2020] 中，$K = 1$矩阵 G_i 会化为一个正的标量 $f'(u_i)f'(u_i^t)A''(f(u_i^t)) > 0$，它可以很容易地表示为模型参数的一个下界。然而，对于一般的 $K > 1$ 情形，这一策略，尤其是对 I 给出下界，要困难得多。在下一小节中，我们将讨论如何应对这一挑战，涵盖多项式模型和序数模型。

4.2 受限强凸性的证明概要

如前所述，我们证明中的一项主要挑战在于对 I 求下界。I 是一个依赖于非对称参数矩阵 $G_i \in \mathbb{R}^{K\times K}$ 的二次型。我们可以考虑通过 $\frac{1}{2}(G_i + G_i^\top)$ 将其对称化；但甚至无法确定 $\frac{1}{2}(G_i + G_i^\top)$ 是否正定。因此，我们转而采用以下策略：首先，将矩阵 G_i 分解为一个半正定矩阵与一个误差矩阵之和，该误差矩阵取决于 Θ 与 Θ^* 的差异：

$$G_i = \underbrace{\nabla f(u_i^t)^\top \nabla^2 A(f(u_i^t))\nabla f(u_i^t)}_{G_i^{(1),\mathrm{MN}}} + \underbrace{[\nabla f(u_i) - \nabla f(u_i^t)]^\top \nabla^2 A(f(u_i^t))\nabla f(u_i^t)}_{G_i^{(2),\mathrm{MN}}}, \tag{18}$$

其中 $G^{(1),\mathrm{MN}}_i$ 为对称半正定矩阵，而 $G^{(2),\mathrm{MN}}_i$ 的大小取决于 $u_i = \Theta^\top x_i$ 与 $u_i^t = (t\Theta + (1-t)\Theta^*)$ 之间的差异。随后，我们（i）给出 $G^{(1),\mathrm{MN}}_i$ 的最小特征值的下界；（ii）证明 $G^{(2),\mathrm{MN}}_i$ 的最大奇异值的上界，该上界取决于 Θ 与 Θ^*之间的距离，从而得到在真实参数周围适当邻域内，我们的对数似然损失满足受限强凸性（参见假设 3.3 和 3.4）。此外，对于 $G_i^{(1)}$ 的最小特征值，证明其具有正下界仍然并非易事，因为其各个元素都以非线性方式依赖于模型参数。

多项式 PU 模型的证明概要。对于多项式 PU 模型，我们发现，$\nabla^2 A(f(u)) \in \mathbb{R}^{K\times K}$和 $\nabla f(u) \in \mathbb{R}^{K\times K}$ 都可以分解为一个对角矩阵与一个秩一矩阵之和，因而很容易求逆。于是，求最小特征值下界的问题可转化为求其逆矩阵最大特征值的上界。由此得到以下引理：

引理 4.1。对于任意$\Theta \in \Theta^* + \mathbb{B}_{1,\infty}(R)$,

$$\lambda_{\min}(\nabla f(\Theta^\top x_i)^\top \nabla^2 A(f(\Theta^\top x_i))\nabla f(\Theta^\top x_i)) \geq h^{\mathrm{MN}}(R) + 4C_x R, \tag{19}$$

其中，$h^{\mathrm{MN}}(R)$ 定义见(9)，而 $C_x > 0$ 是假设 *3.1* 中的一个常数。

引理 A.4 的详细证明见附录C.1。基于这一结果以及对 $\nabla f(u)$ 的光滑性保证（$G^{(2)}$的上界，$_i^{\mathrm{MN}}$），再结合经验过程中的一些技术工具，我们可以证明 $\mathcal{L}_n^{\mathrm{MN}}(\Theta)$ 满足受限强凸性（详见引理 A.2）。

序数-PU 模型的证明概要。另一方面，序数-PU 模型的证明要复杂得多。为说明其中的差异，我们可以先写出对数似然损失如下：

$$\mathcal{L}_n^{\mathrm{ON}}(\theta) = \frac{1}{n}\sum_{i=1}^{n} A(f^{\mathrm{ON}}(u(x_i,\theta))) - \delta_i^\top f^{\mathrm{ON}}(u(x_i,\theta)).$$

其中，函数 $f^{\mathrm{ON}}\colon \mathbb{R}^K \to \mathbb{R}^K$ 和 $u\colon \mathbb{R}^p \times \mathbb{R}^{p+K} \to \mathbb{R}^K$ 满足以下条件：

$$(f^{\mathrm{ON}}(u))_j = \begin{cases} \log\frac{n_j}{\pi_j n_u} + \log\left[(1+e^{\sum_{l=1}^{j+1} u_l})^{-1} - (1+e^{\sum_{l=1}^{j} u_l})^{-1}\right], & 1 \le j < K, \\ \log\frac{n_K}{\pi_K n_u} + \log\left[1 - (1+e^{\sum_{l=1}^{K} u_l})^{-1}\right], & j = K, \end{cases} \tag{20}$$

$$u(x_i,\theta) = (x_i^\top \theta_{1:p} - \theta_{p+1}, -\theta_{p+2}, \ldots, -\theta_{p+K})^\top.$$

与多项式情形中的损失函数 $\mathcal{L}_n^{\mathrm{MN}}(\theta)$相比（见（15）），$\mathcal{L}_n^{\mathrm{ON}}(\theta)$涉及比（16）中定义的函数 $f(\cdot)$更复杂的非线性函数 $f^{\mathrm{ON}}(\cdot)$，并需要新的证明技术。具体而言，类似于多项式情形中通过在（18）中对半正定矩阵取下界并对误差矩阵取上界来处理 $K \times K$ 矩阵 G_i，这里我们希望对.....的最小特征值取下界

$$G_i^{(1),\mathrm{ON}} := \nabla f^{\mathrm{ON}}(u_i^t)^\top \nabla^2 A(f^{\mathrm{ON}}(u_i^t)) \nabla f^{\mathrm{ON}}(u_i^t), \tag{21}$$

并给出最大奇异值的上界的上界

$$G_i^{(2),\mathrm{ON}} := [\nabla f^{\mathrm{ON}}(u_i) - \nabla f^{\mathrm{ON}}(u_i^t)]^\top \nabla^2 A(f^{\mathrm{ON}}(u_i^t)) \nabla f^{\mathrm{ON}}(u_i^t). \tag{22}$$

然而，矩阵 $\nabla f^{\mathrm{ON}}(u)$ 不像在多项式情形中那样容易求逆。相反，我们将其分解为若干常数可逆矩阵与一个三对角矩阵的乘积，而该三对角矩阵的元素以高度非线性的方式依赖于所有模型参数。随后，我们利用序数模型中特殊的序贯减法结构，通过仔细分析将该问题转化为求裂项和的下界。接着，我们给出关于 $f^{\mathrm{ON}}(\cdot)$ 的以下结果：

引理 4.2。令 R_0、$r_0 > 0$ 为任意正数常数。对于任意 $\theta \in \theta^* + S(R_0, r_0)$，其中 $S(R_0, r_0)$ 如 ***(11)*** 中所定义，若 $u_i = u(x_i, \theta)$，则有

$$\lambda_{\min}(\nabla f^{\mathrm{ON}}(u_i)^\top \nabla f^{\mathrm{ON}}(u_i)) \ge \frac{e^{2(C_x+1)(R^*+R_0)}}{4K(1+e^{(C_x+1)(R^*+R_0)})^4}.$$

引理 4.2 的详细证明见附录 C.2。结合 $\nabla^2 A(\cdot)$ 的最小特征值保证，引理 4.2 可以导出在 (21) 中定义的 $G_i^{(1),\mathrm{ON}}$ 的最小特征值下界。据我们所知，先前研究尚未给出有序对数似然损失曲率的此类刻画。现有分析有序模型统计性质的研究，仅关注低维情形下最大似然估计 [McCullagh, 1980] 的唯一性，或者假设曲率下界为常数 [Lee和Wang, 2020]，而没有将该下界刻画为模型参数的函数。然而，如第 4.1 节所讨论的，为了刻画 PU 模型中非凸对数似然损失的受限凸区域，这些先前结果不足以实现我们的目标。此外，对于在 (22) 中定义的第二个误差项 $G_i^{(2),\mathrm{ON}}$，我们还将在下面的引理中证明 $\nabla f^{\mathrm{ON}}(u)$ 的 Lipschitz 性质：

引理 4.3. 令 R_0、$r_0 > 0$ 为任意正数常数。对于任意 $\theta \in \theta^* + S(R_0, r_0)$，其中 $S(R_0, r_0)$ 的定义见 (11)，若 $u_i = u(x_i, \theta)$、$u_i^* = u(x_i, \theta)$，则有

$$\|\nabla f^{\mathrm{ON}}(u_i) - \nabla f^{\mathrm{ON}}(u_i^*)\| \le \sqrt{5}[(h^{\mathrm{ON}}(R_0, r_0))^2 + h^{\mathrm{ON}}(R_0, r_0)](C_x+1)R_0 K,$$

其中，$h^{\mathrm{ON}}(\cdot)$ 的定义见（13）。

引理 4.3 的详细证明也可见附录 C.2。

基于引理 4.2、引理 4.3、一些其他支持性结果和概率集中界，我们随后可以在一个取决于模型参数的区域内，证明序数-PU 模型具有受限强凸性。

引理 4.4 （序数-PU 模型下的受限强凸性）。 如果数据集 $\{（x_i、 z_i）\}_{i=1}^n$ 是在病例对照设置下根据序数-*PU* 模型生成的，且假设 *3.1*、*3.2* 和 *3.4* 成立，那么以至少 $1-\exp\{-^{cK}{}_{\log_{C_x^2 m+1}}{}^{J}\}$ 的概率，

$$\begin{aligned}&\langle\nabla\mathcal{L}_n^{\mathrm{ON}}(\theta)-\nabla\mathcal{L}_n^{\mathrm{ON}}(\theta^*),\theta-\theta^*\rangle\\ \geq&\alpha\|\theta-\theta^*\|_2^2-C\alpha\frac{\log J+m}{n}\|\theta-\theta^*\|_{\omega,2,1}^2-CL\tau(\theta-\theta^*),\end{aligned}\tag{23}$$

对于任意 $\theta\in\theta^*+S（R_0、 r_0）$ </style}，均成立，其中 、 是取决于模型参数的正数常数，其具体形式可见附录 B，且函数 $（\Delta）=\sqrt{}^{\lambda_{\max}（\Sigma）（\log J+m）n}\|\Delta_{1:p}\|_{\omega,2,1}+\sqrt{}^{K（\log J+\log(2K)）}{}_n\|\Delta\|_2$。

在适当的速率条件下，可以证明，当 $\theta-\theta^*$ 属于受限锥时，(23) 右侧的第一项是占主导地位的正项，因此最终得到定理 3.2 中的统计误差界。

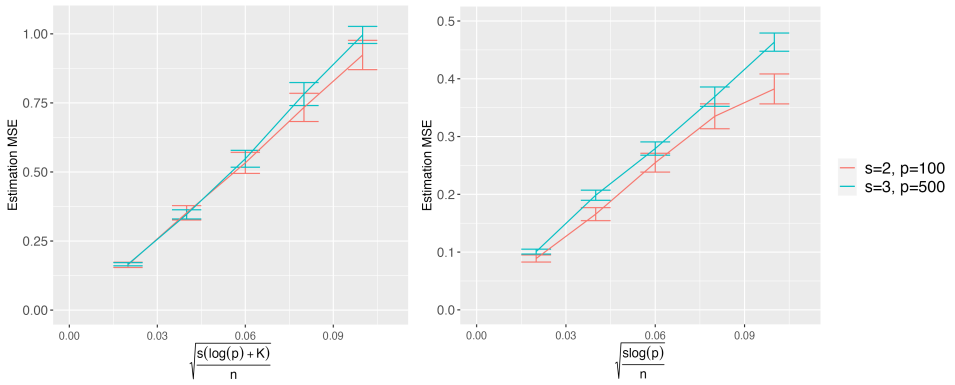
5 模拟研究

在本节中，我们通过模拟实验验证理论结果，并评估我们针对多项式-PU模型和序数-PU模型提出的算法。在模拟实验中，我们重点关注病例对照设置，而在真实数据实验中，我们则研究单训练集情形。

5.1 验证理论保证

首先，我们针对稀疏性 s、维度 p、类别数量 K 和样本量 n，实验验证第 3 节中所给出的均方误差界的理论尺度。对于多项式模型，我们考虑组稀疏性，其中每一行 $\Theta_{j,:}\in\mathbb{R}^K$ 由一个组组成；对于有序模型，我们关注元素级稀疏性。给定稀疏性水平 s，我们随机选择支撑集，并从 U($[-1,-0.5]\cup[0.51]$) 中抽取非零参数。截距的选择使得仅截距模型为每个类别分配相等的概率。我们从独立同分布的标准高斯分布中抽取特征向量 X 。随后，我们通过从总体中随机抽取 $n_u=\frac{n}{2}$ 个未标记样本，并从真实标签为 j、 $1\leq j\leq K$ 的总体中抽取 $n_j=\frac{n}{2K}$ 个已标记样本，生成正例 - 未标记响应 z_i， $1\leq i\leq n$ 。对于每个 s 和 K，调节参数设置为 $\lambda=c\sqrt{}_{\log_n}{}^{p}$，其中 c 为适当选择的常数。初始化选择为仅截距模型的最大似然估计，这些模型具有闭式解。

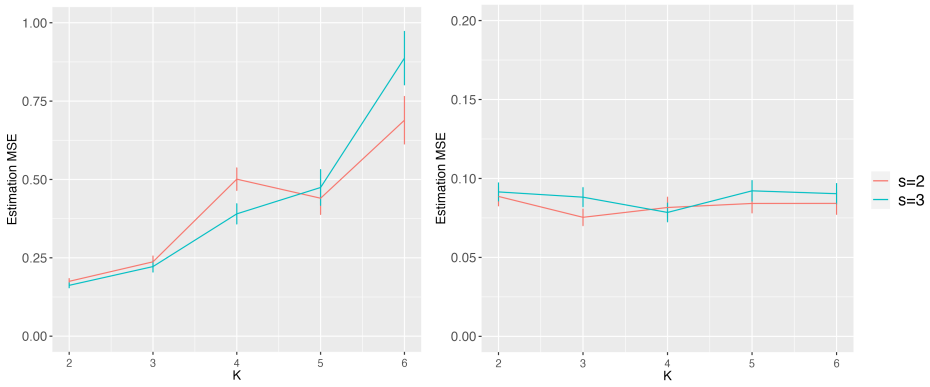
图1展示了两种模型在不同 s,p,n 且 $K=2$ 条件下的估计均方误差。 x-轴表示定理 3.1 和 3.2 中关于 s,p,n 的理论标度：对于具有组稀疏性的多项参数，组的大小为 K，标度为 $\sqrt{s\log\frac{p+K}{n}}$；对于稀疏性为 s 的序数参数，标度为 $\sqrt{s\log_n p}$。直线且彼此接近的曲线验证了这些标度。我们还研究了均方误差如何随正类别数 K 变化。我们关注 $p=100$、 $s=2$,3，以及满足 $\sqrt{s\log_n p}=0.02$ 的样本量 n ，结果见图2。结果表明，尽管理论标度



(a) 多项式　　(b) 有序

^^^^ 图1：多项式模型的均方误差 $E\|\Theta-\Theta^*\|_F$或序数模型的均方误差 $E\|\theta-\theta^*\|_2$，相对于理论收敛率 $\sqrt{s(\log\ p+K)n}$ 或 $\sqrt{s\log\ p/n}$绘制。两条线对应不同的稀疏度和维度，且看起来彼此吻合良好，验证了我们的理论收敛率。

对于两种模型关于 K 都是多项式的，但这可能仅适用于多项式模型，因为参数总数为 Kp，而序数模型的均方误差在 K 的不同取值下似乎相近。



(a) 多项式　　(b) 有序

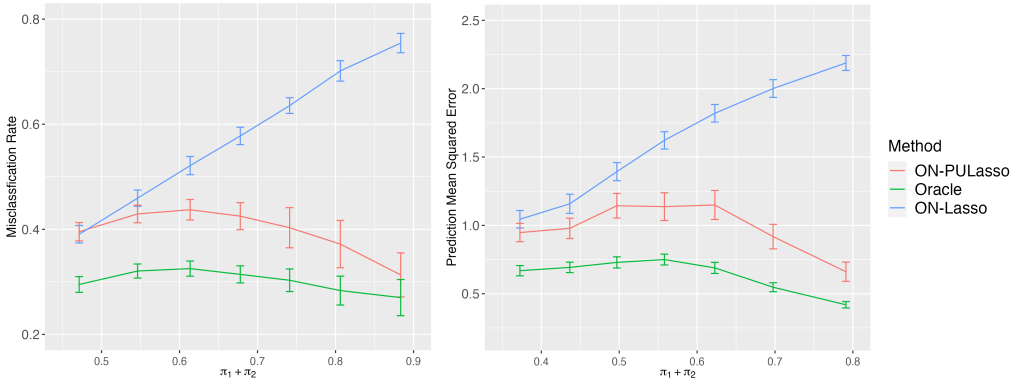
^^^^ 图2：多项式模型的均方误差 $E\|\Theta-\Theta^*\|_F$或序数模型的均方误差 $E\|\theta-\theta^*\|_2$，相对于 K绘制。 $p=100$和 n的取值满足 $\sqrt{s\log_n p}=0.02$。我们可以看到，对于多项式模型，估计误差随着 K 的增加而稳步增长；而对于序数模型，估计误差变化不大。

5.2 比较研究

在此，我们使用本文方法与相应的基线方法比较预测性能。基线方法假设所有未标记数据都是真负例，并直接最小化经 ℓ_1 惩罚的多项式和有序对数似然。我们将本文方法分别称为 *MN-PULasso* 和*ON-PULasso*，将基线方法分别称为*MN-Lasso* 和*ON-Lasso*。此外，我们还给出预言机预测误差作为参考，这些误差基于真实模型参数。具体而言，对于以下对象，我们比较拟合模型 / 真实模型的预测误差：

预测大小为 100 的测试数据集中的真实标签，并研究总体中每个类别的真实患病率$\pi_j = \mathbb{P}(y = j)$ $j = 0, \ldots, K$ 以及抽样比例 $\frac{n_u}{n}, \frac{n_j}{n}$, $j = 1, \ldots, K$ 的影响。模型参数设置基本与第 5.1 节中的描述相同，但我们将截距参数设置为实现不同的患病率 π_j，并且现在从方差为 4 的独立同分布高斯分布中抽取每个特征 X_j。我们重点考察 $s = 2$, $K = 2$, $p = 400$, $n = 200$，所有调节参数均通过五折交叉验证进行选择。

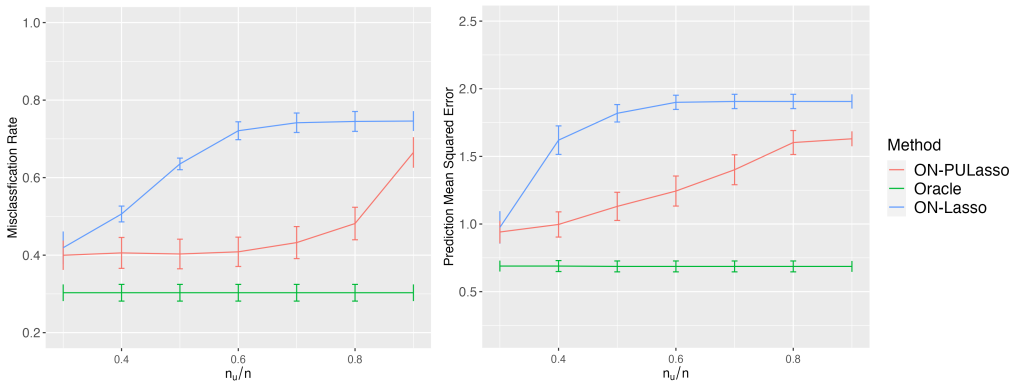
我们首先固定 $n_u = \frac{n}{2} = 200$, $n_j = \frac{n}{2K} = 50$，对于 $j > 0$，同时改变不同类别的真实患病率。然而，由于患病率以复杂的形式依赖于模型参数，因此很难直接为真实患病率设定具体数值。我们改为仔细设置截距参数，使真实患病率处于适当范围内。更多细节请参见补充材料。我们根据模拟数据估算得到的患病率，并将正样本的患病率$\pi_1 + \pi_2 = 1 - \pi_0$作为图 3 的横轴。较大的 $\pi_1 + \pi_2$（较小的 π_0）意味着事实上有更多未标记样本为正例而非负例，因此，将这些未标记样本视为负例的基线方法会受到显著偏差的影响。我们还发现，当不同类别的患病率变得更加极端时，预言者和我们方法的预测误差都会变小，因为在不平衡测试数据上，预测问题会变得更容易。



(a) 多项式　　(b) 有序

图 3：比较我们的方法（*MN-PULasso* 和 *ON-PULasso*）、基线方法（*MN-Lasso* 和 *ON-Lasso*）与基于真实模型的预言家预测性能，并绘制其随正类别真实患病率的变化。对于多项式模型，y轴表示误分类率 $\mathbb{P}_n(\hat{y} \neq y)$；对于序数模型，$y$轴表示预测均方误差 $\mathbb{E}_n(\hat{y} - y)^2$。可以看出，我们的方法优于将未标记数据简单视为零的基线方法，尤其是在正类别的真实患病率上升时。

此外，我们还研究了抽样比例的影响。我们适当设置两个模型的截距，使所有类别的真实患病率相近（平衡数据）。更多细节请参见补充材料。我们将 $\frac{n_u}{n}$ 从 0.3 变化到 0.9，其中 $n_1 = n_2 = \frac{n - n_u}{2}$。$\frac{n_u}{n}$ 越大，未标记样本就越多，因此估计问题也越具挑战性。图 4 表明，随着 $\frac{n_u}{n}$ 增大，我们的方法和基线方法的预测误差都会增大，但基线方法受到的负面影响更大。



(a) 多项式　　(b) 有序

图 4：我们的方法（*MN-PULasso* 和*ON-PULasso*）、基线方法（*MN-Lasso* 和 *ON-Lasso*）以及预言家预测的预测性能比较，横轴为未标记样本的比例。未标记样本越多，问题越具挑战性；而与基线方法相比，我们的方法的预测误差增长缓慢得多。

6 真实数据实验

本节通过两项真实数据实验验证我们的方法。一个数据集包含无序标签，适用于多项式 PULasso 方法；另一个包含有序标签，用于验证有序PUlasso 方法。

6.1 多项式应用：数字识别

数字识别是计算机视觉中的一个重要问题，旨在根据图像对数字 0 到 9 进行分类。然而，获取人工标注的样本通常既昂贵又耗时，因此需要采用 PU 学习方法进行多分类。在此，我们希望探究多项式 PULasso 方法解决这一问题的潜力，因为不同数字的图像没有顺序模式。我们从 UCI 机器学习仓库 [Dua 和 Graff, 2017], 获取了公开可用的“多特征数据集”* ，其中包含来自荷兰公用事业地图的 $n = 2000$ 手写数字。该数据集是平衡的（即数字 0 到 9 每种各有 200 个）。从图像中提取了 $p = 649$ 个特征，包括 76 个字符形状的傅里叶系数、216 个轮廓相关系数、64 个 Karhunen-Love 系数、 2×3 个窗口中的 240 个像素平均值、47 个 Zernike 矩，以及 6 个形态学特征。此数据集中的所有样本均已正确标记，但为了验证我们的方法在 PU 场景下的表现，我们手动污染训练集，使其成为正类与未标记数据。具体而言，我们污染训练集（占完整数据集的 70%），以模拟实际应用中可能遇到的正例-未标记数据，并在污染后的数据上分别应用我们的多项式 PULasso 方法和基线多项式 Lasso 方法；随后，我们在干净的测试数据上比较这两种方法的预测性能。

更具体地说，以概率 $1 - \pi_i$，训练集中的每个数字 i 都被替换为 0。这就是附录 2.1 中描述的单训练集情形，并使用其中提出的多项式 PUlasso 算法。为了研究 π_i 及其错误设定的影响，我们在此进行实验

*https://archive.ics.uci.edu/ml/datasets/Multiple+特征

，分别采用不同的 π_i 值进行数据生成，并采用不同的 $\hat{}\pi_i$ 值进行模型拟合。我们将所有 π_i ($\hat{}\pi_i$) 在 $i \in \{1, \ldots, 9\}$之间设为相同值。所有正则化参数均通过训练集上的 5折交叉验证进行选择。然后，我们通过将每个测试样本预测为具有最高预测概率的类别，在测试数据集上测量预测准确率，并在表 1 中报告正确预测百分比。我们可以看到：(i) 当观察到带标签样本的概率较低时（$\pi_i = 0.4$，0.6），即使该概率在我们的算法中设定错误，MN-PULasso 的表现也优于基线MN-Lasso；(ii) 对于不同水平的 π_i，采用正确 $\hat{}\pi_i = \pi_i$ 的 MN-PULasso 始终表现最佳。

	MN-Lasso	MN-PULasso				
		$\hat{\pi}_i = 0.5$	$\hat{\pi}_i = 0.6$	$\hat{\pi}_i = 0.7$	$\hat{\pi}_i = 0.8$	$\hat{\pi}_i = 0.9$
$\pi_i = 0.4$	26.17%	79.83%	66.50%	52.00%	43.50%	36.50%
$\pi_i = 0.6$	66.17%	88.50%	98.00%	91.33%	82.00%	72.00%
$\pi_i = 0.8$	92.33%	47.00%	82.50%	89.00%	97.33%	94.33%

表 1：数字识别数据实验：多项式 PULasso 和多项式 Lasso 在测试数据集（30%）上的准确率，两者均使用训练数据集（70%）训练。对于训练集中标签为 i （$i > 0$）的任一正样本，其未标记的概率为 $1-\pi_i$。我们的 MN-PULasso 方法的输入为 $\hat{}\pi_i$。我们观察到，输入正确观测概率 $\hat{}\pi_i = \pi_i$ 的 MN-PULasso 始终表现最佳；此外，它对 π_i 的设定错误也具有相当的鲁棒性。

6.2 序数应用：阴道镜主观性

阴道镜检查用于检查宫颈、阴道和外阴，通常在巴氏涂片发现异常时进行。与其让医学专家实时检查患者，更有效的做法是录制视频（数字阴道镜检查），供医学专家稍后查看，这样医疗从业者就能腾出时间接诊更多患者。一个重要的问题是，数字阴道镜检查的质量是否足以让医学专家做出诊断。

为研究这一问题，我们分析了数字阴道镜检查质量评估数据集 [Fernandes 等人， 2017]，该数据集来自 UCI 机器学习仓库 [Dua 和 Graff， 2017]。该数据集包含$n = 287$ 用于筛查宫颈癌的数字阴道镜图像，并从每张数字阴道镜图像中提取了 $p = 62$ 特征。6 位医学专家查看了这些数字阴道镜图像，并将其质量评为差或好。随后，我们根据将阴道镜图像评为优质的医学专家人数，为每张阴道镜图像赋予一个序数排名；该数据集中的排名范围为 0 至 5，因此形成了 5 个正类和 1 个负类。本研究旨在开发一种模型，利用 p特征预测有多少位专家认为图像质量良好。

与数字识别应用一样，这个特定的阴道镜检查数据集并不是正例-未标记数据。然而，让医学专家标注阴道镜检查的质量既耗时又昂贵，这可能导致实际应用中出现大量未标记样本。因此，为该应用中的正类与未标记数据和序数数据开发一种方法是一项重要任务，我们通过手动屏蔽数据，使样本变为正类与未标记，来研究我们的方法在这一情形下的潜力。具体而言，以概率 $1 - \pi_i$，训练集（数据集的 70%）中每个由 i 位专家评定为质量良好的阴道镜检查都被修改为其污染标签为 0，从而使 0 成为未标记类，而所有其他类均有标签。与之前一样，这一操作仅针对训练数据进行，因此我们仍然可以使用测试集（数据集剩余的 30%）来计算准确率。这就是附录 2.2 中描述的单训练集情形，并使用其中提出的有序 PUlasso 算法。此外，我们考虑不同的

用于生成正例-未标记数据的 π_i 值，以及训练算法中使用的 $\hat{}\pi_i$ ，以研究它们的影响。为简单起见，所有 π_i 和 $\hat{}\pi_i$ 对所有 $0 \leq i \leq 6$ 均设置为相同值。

我们在训练集上采用5折交叉验证进行训练。由于标签是有序的，将真实标签为6的样本预测为0，比预测为5更差。因此，我们报告均方误差来衡量预测性能，而不是报告分类正确的百分比。如表2所示，当正确观测到正标签数据的概率 π_i 介于0.6和0.8之间时，即使π_i 存在中度设定偏差，Ordinal PUlasso的表现始终优于将所有未标记样本都视为负例的Ordinal Lasso。

	ON-Lasso	ON-PULasso				
		$\hat{\pi}_i = 0.5$	$\hat{\pi}_i = 0.6$	$\hat{\pi}_i = 0.7$	$\hat{\pi}_i = 0.8$	$\hat{\pi}_i = 0.9$
$\pi_i = 0.6$	10.70	7.27	3.79	4.72	6.92	8.60
$\pi_i = 0.7$	9.36	7.27	7.27	5.53	6.28	9.01
$\pi_i = 0.8$	7.97	6.41	4.51	7.27	4.99	5.93

表2：阴道镜检查质量评估数据实验：有序PUlasso和有序Lasso在测试数据集（30%）上的预测均方误差，二者均在训练数据集（70%）上训练。对于训练集中任何标签为 i （$i > 0$）的正样本，其为未标记样本的概率为$1 - \pi_i$。我们的ON-PULasso方法的输入为$\hat{}\pi_i$。我们观察到，对于合理选择的$\hat{}\pi_i$，我们的ON-PULasso算法的预测始终比基线方法ON-Lasso更准确。

7 讨论

本文聚焦于正例与未标记（PU）数据下的高维分类问题，其中只有部分正样本带有标签，这是许多应用中常见的情况。由于对数似然损失函数具有非凸景观，此类数据集带来了独特的优化和统计挑战。不同于先前聚焦于二分类的研究，本文关注包含多个正类别的情形，该情形更为一般，但也加剧了非凸性问题。具体而言，我们分别针对无序标签和有序标签提出了多项PU模型和有序PU模型，并配套提出用于估计这些模型的算法。尽管这些问题具有颇具挑战性的非凸性，尤其是有序模型，我们仍然证明了算法收敛性，并在合理初始化的条件下刻画了统计误差界。一系列模拟实验和真实数据研究表明，我们提出的模型和方法具有实际应用价值和应用潜力。

此外，还有一些值得未来研究的开放问题。尽管我们的理论分析和实证研究已验证了非凸方法的有效性，但开发凸方法并将其性能与我们的方法进行比较，或许仍然值得探讨；这可能需要借鉴 Song 等人提出的矩方法思路 [2020]。我们目前的模型都是参数模型，但也可以考虑将其扩展为半参数模型，或纳入更灵活的非参数机器学习算法，例如深度神经网络。此外，我们还可以将这一框架推广到生物医学工程及其他领域的诸多应用中。

A Proofof定理 3.1

为简化表述，下文省略 $\mathcal{L}_n^{\mathrm{MN}}(\Theta,\ 0_{K\times 1})$，简记为 $\mathcal{L}_n^{\mathrm{MN}}(\Theta)$。我们首先给出用于证明定理 3.1 的两个主要辅助引理。

引理 A.1（多项式 PU 模型下的偏差界）。 如果数据集 $\{(x_i, z_i)\}_{i=1}^n$ 是在病例对照设置下由多项式 *PU* 模型生成的，并且假设 *3.1–3.3* 成立，则

$$\|\nabla\mathcal{L}_n^{\mathrm{MN}}(\Theta^*)\|_{\omega^{-1},2,\infty} \le C\lambda_{\max}^{\frac{1}{2}}(\Sigma)\max_{1\le j\le J}\sqrt{\frac{c_j(m+\log J)}{n}}, \tag{24}$$

概率至少为 $1-\exp\{-c(\log J+m)\}$。

引理 A.2（多项式 PU 模型下的受限强凸性）。 如果数据集 $\{(x_i, z_i)\}_{i=1}^n$是由病例对照设置下的多项式 *PU* 模型生成的，并且假设 *3.1–3.3* 成立，那么其概率至少为

$$\begin{aligned}
&1-\exp\{-cR^{*2}K\lambda_{\max}(\Sigma)\min_j r_j^{-1}(\log J+m)\},\\
&\big\langle\nabla\mathcal{L}_n^{\mathrm{MN}}(\Theta)-\nabla\mathcal{L}(\Theta^*),\Theta-\Theta^*\big\rangle\\
\ge&\alpha\|\Theta-\Theta^*\|_2^2-\tau_1\frac{\log J+m}{n}\|\Theta-\Theta^*\|_{\omega,2,1}^2-\tau_2\sqrt{\frac{\log J+m}{n}}\|\Theta-\Theta^*\|_{\omega,2,1}
\end{aligned} \tag{25}$$

对于任意 $\Theta\in\{\Theta'：\ 最大值_{1\le i\le K}\|\Theta^*_{:,\ i}-\Theta'_{:,\ i}\|_1\le R_0\}$，其中 $\alpha=\frac{h^{\mathrm{MN}}(R_0)\ \lambda_{\min}(\Sigma)}{2}$，$\tau_1=C\alpha$，$\tau_2=C(K+C_xR^*\sqrt{K})\lambda^{\frac{1}{2}}$最大值（$\Sigma$）。

为进一步利用引理 A.2，我们在此给出另一个引理，该引理保证了曲率项$\alpha>0$，并表明占主导地位的松弛项为 $\tau_2\sqrt{\log\frac{J+m}{n}}\|\Theta-\Theta^*\|_{\omega,\ 2,\ 1}$。

引理 A.3。 在假设 *3.1* 和 *3.3* 下，对于任意 $\Theta\in\Theta^*+\mathbb{B}_{1,\ \infty}(R_0)$，我们有

$$h^{\mathrm{MN}}(R_0)>0,\quad \tau_1\frac{\log J+m}{n}\|\Theta-\Theta^*\|_{\omega,2,1}^2\le\tau_2\sqrt{\frac{\log J+m}{n}}\|\Theta-\Theta^*\|_{\omega,2,1}.$$

引理 A.1-A.3 的证明将在 A.1 节中给出。现在我们准备证明定理 3.1。

^定理 *3.1* 的证明。 由于 Θ是一个驻点

$$\min_{\Theta-\Theta^*\in\mathbb{B}_{1,\infty}(R_0)}\mathcal{L}_n^{\mathrm{MN}}(\Theta)+\lambda\|\Theta\|_{\omega,2,1},$$

对于任意满足 $\max_{1\le i\le K}\|\Theta_{:,\ i}-\Theta^*_{:,\ i}\|_1\le R_0$的 $\Theta\in\mathbb{R}^{p\times K}$，我们有

$$\langle\nabla\mathcal{L}_n^{\mathrm{MN}}(\widehat{\Theta})+\nabla P_\lambda^{\mathrm{MN}}(\widehat{\Theta}),\Theta-\widehat{\Theta}\rangle\ge 0, \tag{26}$$

其中，$\nabla\lambda\|\Theta\|_{\omega,\ 2,\ 1}$ 是 $\lambda\|\Theta\|_{\omega,\ 2,\ 1}$的任意一个次梯度。由于 Θ^* 也属于可行集 ，

$$\begin{aligned}
&\langle\nabla\mathcal{L}_n^{\mathrm{MN}}(\widehat{\Theta})+\nabla P_\lambda^{\mathrm{MN}}(\widehat{\Theta}),\Theta^*-\widehat{\Theta}\rangle\ge 0,\\
&\langle\nabla\mathcal{L}_n^{\mathrm{MN}}(\widehat{\Theta})-\nabla\mathcal{L}_n^{\mathrm{MN}}(\Theta^*),\widehat{\Theta}-\Theta^*\rangle\le\langle-\nabla\mathcal{L}_n^{\mathrm{MN}}(\Theta^*)-\nabla P_\lambda^{\mathrm{MN}}(\widehat{\Theta}),\widehat{\Theta}-\Theta^*\rangle.
\end{aligned} \tag{27}$$

^ ^令 $\Delta=\Theta-\Theta^*$, $S=\{j:\ \|\Theta_{\mathcal{G}_j}\|>0\}$, $\mathcal{G}_S=\cup_{j\in S}\mathcal{G}_j$。 根据引理 A.1、引理 A.2,

$$\begin{aligned}
&\alpha\|\widehat{\Delta}\|_2^2-\tau_1\frac{\log J+m}{n}\|\widehat{\Delta}\|_{\omega,2,1}^2-\tau_2\sqrt{\frac{\log J+m}{n}}\|\widehat{\Delta}\|_{\omega,2,1}\\
\le&C\tau_2\sqrt{\frac{\log J+m}{n}}\|\widehat{\Delta}\|_{\omega,2,1}+\langle\nabla P_\lambda^{\mathrm{MN}}(\widehat{\Theta}),\Theta^*-\widehat{\Theta}\rangle.
\end{aligned} \tag{28}$$

由引理 A.3 可得 $\tau_1 \underline{对数}_n^{J+m} \|\widehat{\Delta}\|_{\omega,2,1}^2 \le \tau_2 \sqrt{\underline{对数}_n^{J+m}} \|\widehat{\Delta}\|_{\underline{\omega,2,1}}$。另请注意，

$$\begin{aligned}\langle \nabla P_\lambda^{\mathrm{MN}}(\widehat{\Theta}), \Theta^* - \widehat{\Theta} \rangle \le & P_\lambda^{\mathrm{MN}}(\Theta^*) - P_\lambda^{\mathrm{MN}}(\widehat{\Theta}) \\ \le & \lambda \|\widehat{\Delta}_{\mathcal{G}_S}\|_{\omega_S,2,1} - \lambda \|\widehat{\Delta}_{\mathcal{G}_{S^c}}\|_{\omega_{S^c},2,1},\end{aligned} \tag{29}$$

这意味着

$$\begin{aligned}\alpha \|\widehat{\Delta}\|_2^2 \le & \frac{3\lambda}{2} \|\widehat{\Delta}_{\mathcal{G}_S}\|_{\omega_S,2,1} - \frac{\lambda}{2} \|\widehat{\Delta}_{\mathcal{G}_S^c}\|_{\omega_{S^c},2,1}, \\ \|\widehat{\Delta}\|_2 \le & \frac{3\|\omega_S\|_2 \lambda}{h^{\mathrm{MN}}(R_0)\lambda_{\min}(\Sigma)}, \\ \|\widehat{\Delta}\|_{\omega,2,1} \le & 4\|\widehat{\Delta}_{\mathcal{G}_S}\|_{\omega_S,2,1} \le \frac{12\|\omega_S\|_2^2 \lambda}{h^{\mathrm{MN}}(R_0)\lambda_{\min}(\Sigma)}.\end{aligned} \tag{30}$$

只要 $\lambda \ge C\tau_2 \sqrt{\log \underline{\frac{J+m}{n}}}$，其概率至少为

$$1 - \exp\{-cR^{*2} K \lambda_{\max}(\Sigma) \min_j r_j^{-1} (\log J + m)\} - \exp\{-c(\log J + m)\}. \tag{31}$$

□

A.1 引理 A.1、A.2 和 A.3 的证明

引理 *A.1* 的证明。 首先注意，对于任意 $1 \le j \le p, 1 \le k \le K$，

$$\begin{aligned}(\nabla \mathcal{L}_n^{\mathrm{MN}}(\Theta^*))_{jk} = & -\frac{1}{n} \sum_{i=1}^n \left[\sum_{k'=1}^K (\delta_{ik'} - \mathbb{E}(\delta_{ik'}|x_i)) \frac{\partial \eta_{ik'}}{\partial \Theta_{jk}^*} \right] \\ = & -\frac{1}{n} \sum_{i=1}^n X_{ij} \left[\sum_{k'=1}^K \epsilon_{ik'} \left(\mathbb{1}_{\{k'=k\}} - \frac{e^{x_i^\top \Theta_k}}{1 + \sum_{m=1}^K e^{x_i^\top \Theta_m}} \right) \right],\end{aligned} \tag{32}$$

其中 $\delta_{ik} = \mathbb{1}_{\{z_i=k\}}$, $\epsilon_{ik'} = \delta_{ik'} - \mathbb{E}(\delta_{ik'}|x_i)$。令

$$\widetilde{\epsilon}_{ik} = \sum_{k'=1}^K \epsilon_{ik'} \left(\mathbb{1}_{\{k'=k\}} - \frac{e^{x_i^\top \Theta_k}}{1 + \sum_{m=1}^K e^{x_i^\top \theta_m}} \right),$$

则 $\{\widetilde{\epsilon}_{ik}\}_{i=1}^n$ 是独立零均值随机变量，取值于 (−2,2) 上。这里我们有 $|\widetilde{\epsilon}_{ik}| \le 2$，因为

$$\left| \sum_{k'=1}^K \delta_{ik'} \left(\mathbb{1}_{\{k'=k\}} - \frac{e^{x_i^\top \Theta_k}}{1 + \sum_{m=1}^K e^{x_i^\top \theta_m}} \right) \right| \le 1.$$

因此，我们可以写成

$$\left\| (\nabla \mathcal{L}_n^{\mathrm{MN}}(\Theta^*))_{\mathcal{R}_j, \mathcal{C}_j} \right\|_F = \left\| \frac{1}{n} X_{\mathcal{R}_j}^\top \widetilde{\epsilon}_{:,\mathcal{C}_j} \right\|_F \le \frac{\sqrt{c_j}}{n} \max_{k \in \mathcal{C}_j} \left\| X_{\mathcal{R}_j}^\top \widetilde{\epsilon}_{:,k} \right\|_2. \tag{33}$$

现在，对于每个特定的 $k \in \mathcal{C}_j$，在给定 X 的条件下，以高概率（w.h.p.）我们有 $\|X_{\mathcal{R}_j}^\top \widetilde{\epsilon}_{:,k}\|_2$。首先注意到，

$$\begin{aligned}\mathbb{E}(\|X_{\mathcal{R}_j}^\top \widetilde{\epsilon}_{:,k}\|_2 | X) \le & \left(\mathbb{E}(\|X_{\mathcal{R}_j}^\top \widetilde{\epsilon}_{:,k}\|_2^2 | X) \right)^{\frac{1}{2}} \\ = & \left(\sum_{i=1}^n \mathbb{E}(\widetilde{\epsilon}_{i,k}^2 | X) (X_{\mathcal{R}_j} X_{\mathcal{R}_j}^\top)_{ii} \right)^{\frac{1}{2}} \\ \le & 2\sqrt{\mathrm{tr}(X_{\mathcal{R}_j} X_{\mathcal{R}_j}^\top)} \\ \le & 2\sqrt{r_j} \|X_{\mathcal{R}_j}\|.\end{aligned} \tag{34}$$

同时，设 $g(u) = \|X_{\mathcal{R}_j}^\top u\|_2$ 是一个从 $\mathbb{R}^n$ 到 $\mathbb{R}$ 的函数，则 g 是凸的，且为 $\|X_{\mathcal{R}_j}\|$-Lipschitz：

$$|g(u) - g(v)| \le \|X_{\mathcal{R}_j}^\top (u - v)\|_2 \le \|X_{\mathcal{R}_j}\| \|u - v\|_2. \tag{35}$$

Ap应用 Talagrand 压缩不等式（Vershynin [2018] 中的定理 5.2.16）可得

$$\left\| \|X_{\mathcal{R}_j}^\top \widetilde{\epsilon}_{:,k}\|_2 - \mathbb{E}(\|X_{\mathcal{R}_j}^\top \widetilde{\epsilon}_{:,k}\|_2 | X) \right\|_{\psi_2} \le C\|X_{\mathcal{R}_j}\|, \tag{36}$$

这进一步意味着

$$\mathbb{P}(\|X_{\mathcal{R}_j}^\top \widetilde{\epsilon}_{:,k}\|_2 - \mathbb{E}(\|X_{\mathcal{R}_j}^\top \widetilde{\epsilon}_{:,k}\|_2) > C\|X_{\mathcal{R}_j}\|\sqrt{n} t_j) \le \exp\{-nt_j^2\}, \tag{37}$$

对于任意 $t_j > 0$，结合 (36)、(37)，并对 $k \in \mathcal{C}_j$ 和 $j \in [J]$取并集界，可以证明以至少 $1 - \sum_{j=1}^J c_j \exp\{-nt_j^2\}$ 的概率，

$$\begin{aligned} \|\nabla \mathcal{L}_n^{\mathrm{MN}}(\Theta)\|_{\omega^{-1},2,\infty} &= \max_{1\le j\le J} \frac{1}{\omega_j} \|(\nabla \mathcal{L}_n^{\mathrm{MN}}(\Theta^*))_{\mathcal{R}_j,\mathcal{C}_j}\|_F \\ &\le \max_{1\le j\le J} \frac{\sqrt{c_j}\|X_{\mathcal{R}_j}\|}{\omega_j \sqrt{n}} \left(2\sqrt{\frac{r_j}{n}} + Ct_j\right) \end{aligned} \tag{38}$$

[1]现在我们给出 $\|X_{R_j}\|$ 的概率界。根据假设 3.1，$X_{\mathcal{R}_j} \Sigma_{\mathcal{R}_j,\ \mathcal{R}_j}^{-2} \in \mathbb{R}^{n\times r_j}$ 具有相互独立的次高斯行，其次高斯参数为 $\sigma\lambda^{-2}{}_{\min}(\Sigma_{R_j,\ R_j}) \le C$，协方差为 I_{r_j}。可以对 $X_{\mathcal{R}_j} \Sigma_{\mathcal{R}_j,\ \mathcal{R}_j}^{-2}$，$1 \le j \le J$应用具有独立各向同性次高斯行的矩阵谱范数上界（Vershynin [2010] 中的定理 5.39），从而得到：

$$\|X_{\mathcal{R}_j}\| \le \lambda_{\max}^{\frac{1}{2}}(\Sigma) \|X_{\mathcal{R}_j} \Sigma_{\mathcal{R}_j,\mathcal{R}_j}^{-\frac{1}{2}}\| \le \lambda_{\max}^{\frac{1}{2}}(\Sigma) C(\sqrt{n} + \sqrt{r_j}) \le C\lambda_{\max}^{\frac{1}{2}}(\Sigma)\sqrt{n}, \tag{39}$$

以至少 $1 - 2J\exp\{-cn\} \ge 1 - \exp\{-cn\}$ 的概率对 $1 \le j \le J$ 成立，前提是按照（39）适当选择 $C > 0$。令 $t_j = \sqrt{m + \log(2c_j J)/n}$ 为 (38) 中的值，则有

$$\begin{aligned} \|\nabla \mathcal{L}_n^{\mathrm{MN}}(\Theta)\|_{\omega^{-1},2,\infty} &\le C\lambda_{\max}^{\frac{1}{2}}(\Sigma) \max_{1\le j\le J} \frac{\sqrt{c_j}}{\omega_j} \sqrt{\frac{m + \log(2c_j J)}{n}} \\ &\le C\lambda_{\max}^{\frac{1}{2}}(\Sigma) \max_{1\le j\le J} \sqrt{\frac{c_j(m + \log J)}{n}}, \end{aligned} \tag{40}$$

其成立的概率至少为

$$1 - \exp\{-(\log J + m)\} - \exp\{-cn\} \ge 1 - \exp\{-c(\log J + m)\},$$

其中应用了假设 3.2：$\log J, m \le Cn$。 □

引理 *A.2* 的证明。令 $\Delta = \Theta - \Theta^*$。首先注意到

$$\nabla \mathcal{L}_n^{\mathrm{MN}}(\Theta) = \frac{1}{n} \sum_{i=1}^n x_i (\nabla A(\eta_i) - \delta_i)^\top \nabla f(\Theta^\top x_i), \tag{41}$$

其中 $\delta_i \in \mathbb{R}^K$，且 $\delta_{ik} = \mathbb{1}_{\{z_i = k\}}$。因此，

$$\begin{aligned} &\left\langle \nabla \mathcal{L}_n^{\mathrm{MN}}(\Theta) - \nabla \mathcal{L}(\Theta^*), \Theta - \Theta^* \right\rangle \\ =&\frac{1}{n} \sum_{i=1}^n x_i^\top (\Theta - \Theta^*) \nabla f(\Theta^\top x_i)^\top \left[\nabla A(f(\Theta^\top x_i)) - \nabla A(f(\Theta^{*\top} x_i))\right] \\ &- \frac{1}{n} \sum_{i=1}^n x_i^\top (\Theta - \Theta^*) \left(\nabla f(\Theta^\top x_i) - \nabla f(\Theta^{*\top} x_i)\right)^\top \epsilon_i \\ :=&\mathrm{I}(\Delta) - \mathrm{II}(\Delta), \end{aligned} \tag{42}$$

其中 $\epsilon_i = \delta_i - \nabla A(f(\Theta^{*\top} x_i))$. 我们将给出 I(Δ) 的下界，并使 II(Δ) 集中于 0 附近 .

1. 求 I(Δ) 的下界。

下面的引理给出了关于 $\frac{1}{n}\sum_{i=1}^{n}\|\Delta^\top x_i\|_2^2$ 的 I(Δ) 下界 – .

引理 A.4. 只要 $\max_{1\le i\le K}\|\Theta_{:,i}-\Theta^*_{:,i}\|_1\le R_0$ 和 $h^{\mathrm{MN}}(R_0)>0$，其中 h 在 (9) 中定义，就可以保证

$$\mathrm{I}(\Delta)\ge\frac{h^{\mathrm{MN}}(R_0)}{n}\sum_{i=1}^{n}\|\Delta^\top x_i\|_2^2. \tag{43}$$

下面我们证明

$$\frac{h^{\mathrm{MN}}(R_0)}{n}\sum_{i=1}^{n}\|\Delta^\top x_i\|_2^2\ge\alpha\|\Delta\|_2^2-\tau_1\frac{\log J+m}{n}\|\Delta\|_{\omega,2,1}^2. \tag{44}$$

对于 $\alpha=\frac{h^{\mathrm{MN}}(R_0)\lambda_{\min}(\Sigma)}{2}$，$\tau_1=C\alpha$ 其中 $C>0$ 是一个常数。如果 $\|\Delta\|_{\omega,2,1}\ge\sqrt{\frac{\alpha n}{\tau_1(\log J+m)}}\|\Delta\|_2$，那么 (44) 的右侧非正，因此 (44) 显然成立。因此，只需证明

$$\inf_{\|\Delta\|_{\omega,2,1}\le\sqrt{\rho},\|\Delta\|_2=1}\frac{1}{n}\sum_{i=1}^{n}\|\Delta^\top x_i\|_2^2\ge\frac{\lambda_{\min}(\Sigma)}{2}, \tag{45}$$

其中 $\rho=\frac{\alpha n}{\tau_1(\log J+m)}$。由于 $\mathbb{E}(\frac{1}{n}\sum_{i=1}^{n}\|\Delta^\top x_i\|_2^2)\ge\lambda_{\min}(\Sigma)\|\Delta\|_2^2$，可由此推出 (45)

$$\sup_{\Delta\in\mathbb{B}_{\omega,2,1}(\sqrt{\rho})\cap\mathbb{B}_2(1)}\left|\frac{1}{n}\sum_{i=1}^{n}\left(\|\Delta^\top x_i\|_2^2-\mathbb{E}(\|\Delta^\top x_i\|_2^2)\right)\right|\le\frac{\lambda_{\min}(\Sigma)}{2}. \tag{46}$$

下面的引理说明了 $\mathbb{B}_{\omega,2,1}(\sqrt{\rho})\cap\mathbb{B}_2(1)$ 与稀疏集之间的联系： –

引理 A.5. 对于任意 $\rho>0$,

$$\mathbb{B}_{\omega,2,1}(\sqrt{\rho})\cap\mathbb{B}_2(1)\subset\left(1+\frac{2}{\min_j\omega_j}\right)\mathrm{cl}\{\mathrm{conv}\{\mathbb{B}_{\mathcal{G},0}(\lfloor\rho\rfloor)\cap\mathbb{B}_2(1)\}\}, \tag{47}$$

其中 $\mathbb{B}_{\mathcal{G},0}(\lfloor\rho\rfloor)=\{U\in\mathbb{R}^{p\times K}:\ |\{j:\ \|U_{\mathcal{R}_j,\mathcal{C}_j}\|_2>0\}|\le\lfloor\rho\rfloor\}$。

根据引理 A.5，可以证明

$$\begin{aligned}&\sup_{\Delta\in\mathbb{B}_{\omega,2,1}(\sqrt{\rho})\cap\mathbb{B}_2(1)}\left|\frac{1}{n}\sum_{i=1}^{n}\left(\|\Delta^\top x_i\|_2^2-\mathbb{E}(\|\Delta^\top x_i\|_2^2)\right)\right|\\&\le\left(1+\frac{2}{\min_j\omega_j}\right)\sup_{\Delta\in\mathrm{conv}\{\mathbb{B}_{\mathcal{G},0}(\lfloor\rho\rfloor)\cap\mathbb{B}_2(1)\}}\left|\mathrm{tr}\left[\Delta^\top\left(\frac{1}{n}X^\top X-\Sigma\right)\Delta\right]\right|.\end{aligned} \tag{48}$$

对于任意 $\Delta\in\mathrm{conv}\{\mathbb{B}_{\mathcal{G},0}(\lfloor\rho\rfloor)\cap\mathbb{B}_2(1)\}$，存在 $\Delta_1,\ \Delta_2,\ \ldots,\ \Delta_k\in\mathbb{B}_{\mathcal{G},\,0}(\lfloor\rho\rfloor)\cap\mathbb{B}_2(1)$，$\beta_1,\ \ldots,\ \beta_k\ge0$ 以及 $\sum_{i=1}^{k}\beta_i=1$，使得 $\Delta=\sum_{i=1}^{k}\beta_i\Delta_i$。定义函数 $g(U,\ U)=\mathrm{tr}\left[U^\top\left(\frac{1}{n}X^\top X-\Sigma\right)U\right]$。我们

可以证明,

$$
\begin{aligned}
&|g(\Delta,\Delta)| \\
=&\left|g(\sum_{i=1}^{k}\beta_i\Delta_i,\sum_{i=1}^{k}\beta_i\Delta_i)\right| \\
=&\left|\sum_{i,j=1}^{k}\beta_i\beta_j\mathrm{tr}\left[\Delta_i^\top\left(\frac{1}{n}X^\top X-\Sigma\right)\Delta_j\right]\right| \\
\leq&\left|\frac{1}{2}\sum_{i,j=1}^{k}\beta_i\beta_j\left(g(\Delta_i+\Delta_j,\Delta_i+\Delta_j)-g(\Delta_i,\Delta_i)-g(\Delta_j,\Delta_j)\right)\right| \\
\leq&3\sup_{U\in\mathbb{B}_{\mathcal{G},0}(2\lfloor\rho\rfloor)\cap\mathbb{B}_2(1)}|g(U,U)|\sum_{i,j=1}^{k}\beta_i\beta_j \\
=&3\sup_{U\in\mathbb{B}_{\mathcal{G},0}(2\lfloor\rho\rfloor)\cap\mathbb{B}_2(1)}|g(U,U)|\,.
\end{aligned}
\tag{49}
$$

因此，现在我们只需对 $\sup_{\Delta\in\mathbb{B}_{\mathcal{G},0}(2\lfloor\rho\rfloor)\cap\mathbb{B}_2(1)}|g(\Delta,\Delta)|$求上界。对于任意满足 $|S|=2\lfloor\rho\rfloor$的 $S\subset[J]$，令 $\mathcal{K}(\mathcal{G}_S)=\{U\in\mathbb{R}^{p\times K}:\ \|U\|_2=1,\ \|U_{\mathcal{G}_j}\|=0$，其中 $j\notin S\}$。于是可以证明

$$
\sup_{\Delta\in\mathbb{B}_{\mathcal{G},0}(2\lfloor\rho\rfloor)\cap\mathbb{B}_2(1)}|g(\Delta,\Delta)|=\sup_{S\subset[J],|S|=2\lfloor\rho\rfloor}\sup_{\Delta\in\mathcal{K}(\mathcal{G}_S)}|g(\Delta,\Delta)|\,,
\tag{50}
$$

对于任意这样的 S，根据球面的覆盖数上界（Vershynin [2010]中的引理 5.2），存在一个关于 $\mathcal{K}(\mathcal{G}_S)$ 的 $\frac{1}{10}$-网 $\mathcal{N}_S$，相对于 ℓ_2 范数，其大小为 $|\mathcal{N}_S|\leq 21^{\sum_{j\in S}c_jr_j}\leq 21^{2\rho m}$。我们可以将 $\sup_{\Delta\in\mathcal{K}(\mathcal{G}_S)}|g(\Delta,\Delta)|$ 的上界用 $\sup_{\Delta\in\mathcal{N}_S}|g(\Delta,\Delta)|$ 表示如下：$\forall U\in\mathcal{K}(\mathcal{G}_S)$，$\exists U'\in\mathcal{N}_S$，$\varepsilon=U-U'$，使得 $\|\varepsilon\|_2\leq\frac{1}{10}$，因此

$$
\begin{aligned}
|g(U,U)|=&|g(U'+\epsilon,U+\epsilon)| \\
\leq&|g(U',U')|+|g(\varepsilon,\varepsilon)|+2\left|\mathrm{tr}\left[U'^\top(\frac{1}{n}X^\top X-\Sigma)\varepsilon\right]\right| \\
\leq&\frac{5}{4}|g(U',U')|+5|g(\varepsilon,\varepsilon)| \\
&+\left|\mathrm{tr}\left[(\frac{1}{2}U'+2\varepsilon)^\top(\frac{1}{n}X^\top X-\Sigma)(\frac{1}{2}U'+2\varepsilon)\right]\right| \\
\leq&\frac{5}{4}\sup_{\Delta\in\mathcal{N}_S}|g(\Delta,\Delta)|+\frac{27}{50}\sup_{\Delta\in\mathcal{K}(\mathcal{G}_S)}|g(\Delta,\Delta)|\,,
\end{aligned}
\tag{51}
$$

这意味着

$$
\begin{aligned}
(1-\frac{27}{50})\sup_{\Delta\in\mathcal{K}(\mathcal{G}_S)}|g(\Delta,\Delta)|\leq&\frac{5}{4}\sup_{\Delta\in\mathcal{N}_S}|g(\Delta,\Delta)| \\
\sup_{\Delta\in\mathcal{K}(\mathcal{G}_S)}|g(\Delta,\Delta)|\leq&3\sup_{\Delta\in\mathcal{N}_S}|g(\Delta,\Delta)|\,.
\end{aligned}
\tag{52}
$$

因此，（45）可由下式推出

$$
\sup_{S\subset[J],|S|=2\lfloor\rho\rfloor}\sup_{\Delta\in\mathcal{N}_S}|g(\Delta,\Delta)|\leq\frac{\lambda_{\min}(\Sigma)}{18(1+2\max_j\omega_j^{-1})}.
\tag{53}
$$

以下引理以高概率给出 $|g(\Delta,\Delta)|$ 的上界。

引理 A.6。 对于任意满足 $\|\Delta\|_2 = 1$ 的 $\Delta \in \mathbb{R}^{p\times K}$，

$$\mathbb{P}\left(|g(\Delta,\Delta)| \geq t\right) \leq 2K \exp\left\{-cn \min\left\{\frac{t^2}{\sigma^4}, \frac{t}{\sigma^2}\right\}\right\}. \tag{54}$$

对于任意 $t > 0$。

令引理 A.6 中的 $t = \frac{\lambda_{\min}(\Sigma)}{18(1+2\max_j \omega_j^{-1})}$，并对 $\Delta \in \mathcal{N}_S, S \subset [J], |S| = 2\lfloor\rho\rfloor$ 取并集界，可以看出 (53) 以至少.....的概率成立。

$$\begin{aligned}
&1 - 2K(21)^{2\rho m}\binom{J}{2\lfloor\rho\rfloor}\exp\{-cn\} \\
\geq &1 - 2K\exp\{\rho[2\log(21)m + 3\log(eJ/2\lfloor\rho\rfloor)] - cn\} \\
\geq &1 - 2K\exp\{\rho[2\log(21)m + 3\log J] - cn\} \\
\geq &1 - 2K\exp\{-\frac{cn}{2}\}
\end{aligned} \tag{55}$$

其中，第三行在 $\rho = \frac{\alpha n}{\tau_1 (\log J + m) = c n \log J + m} \geq 2$ 时成立，而最后一行成立是因为 $\rho(\log J + m) = \frac{\alpha n}{\tau_1} \leq Cn$。

2. 将 II(Δ) 集中于 0 附近。

首先回顾一下，

$$\mathrm{II}(\Delta) = \frac{1}{n}\sum_{i=1}^{n} x_i^\top \Delta \left(\nabla f(\Theta^{*\top} x_i + \Delta^\top x_i) - \nabla f(\Theta^{*\top} x_i)\right)^\top \epsilon_i. \tag{56}$$

令 $\phi(\Delta, \{x_i\}_{i=1}^n, \{\epsilon_i\}_{i=1}^n) = \mathrm{II}(\Delta)\|\Delta\|_{\omega,2,1}$，那么我们的目标是为.....提供一个上界。

$$\sup_{\Delta\in\mathbb{B}_{1,\infty}(R_0)} |\phi(\Delta, \{x_i\}_{i=1}^n, \{\epsilon_i\}_{i=1}^n)|.$$

我们首先对其取上界

$$g_t^{(1)}((x_1,\epsilon_1),\ldots,(x_n,\epsilon_n)) := \sup_{\frac{t}{2}\leq\|\Delta\|_{\omega,2,1}\leq t} \phi(\Delta, \{x_i\}_{i=1}^n, \{\epsilon_i\}_{i=1}^n)$$

对于任意 $t > 0$，并采用类似的论证，我们得到以下上界

$$g_t^{(2)}((x_1,\epsilon_1),\ldots,(x_n,\epsilon_n)) : \sup_{\frac{t}{2}\leq\|\Delta\|_{\omega,2,1}\leq t} -\phi(\Delta, \{x_i\}_{i=1}^n, \{\epsilon_i\}_{i=1}^n).$$

然后，我们应用剥离论证，将上界扩展到 $\mathbb{B}_{1,\infty}(R_0)$。

引理 A.7 (对称化定理)。 设 $U_1, \ldots, U_n$ 是取值于 $\mathcal{U}$ 的独立随机变量，$\{\varepsilon_i\}$ 是一个独立同分布的 *Rademacher* 变量序列，每个变量以 1/2 的概率取值 ± 1。设 Γ 是 $\mathcal{U}$ 上实值函数的一个类，则

$$\mathbb{E}\left(\sup_{\gamma\in\Gamma}\sum_{i=1}^{n}(\gamma(U_i) - \mathbb{E}(\gamma(U_i)))\right) \leq 2\mathbb{E}\left(\sup_{\gamma\in\Gamma}\sum_{i=1}^{n}\varepsilon_i\gamma(U_i)\right).$$

文献中常见的对称化定理 [Vaart 和 Wellner, 1997, Wainwright, 2019]考虑的是绝对值 $|\gamma(U_i) - \mathbb{E}(\gamma(U_i))|$，而不是 $\gamma(U_i) - \mathbb{E}(\gamma(U_i))$。为完整起见，我们也将给出引理 A.7 的证明，尽管该证明与带绝对值的证明基本相同，而且可能已在以往文献中出现过。

我们通过令 $U_i = (\epsilon_i,\ x_i)$ 来应用引理 A.7,

$$\gamma(x_i, \epsilon_i) = \frac{1}{n\|\Delta\|_{\omega,2,1}} x_i^\top \Delta(\nabla f((\Theta^* + \Delta)^\top x_i) - \nabla f(\Theta^{*\top} x_i))^\top \epsilon_i.$$

于是我们有

$$\begin{aligned} &\mathbb{E} g_t^{(1)}((x_1, \epsilon_1), \ldots, (x_n, \epsilon_n)) \\ \leq &\frac{2}{n}\mathbb{E} \sup_{\frac{t}{2} \leq \|\Delta\|_{\omega,2,1} \leq t} \frac{\sum_{i=1}^n h_i(\Delta^\top x_i)\varepsilon_i}{\|\Delta\|_{\omega,2,1}} \\ \leq &\frac{4}{nt}\mathbb{E} \sup_{\frac{t}{2} \leq \|\Delta\|_{\omega,2,1} \leq t} \sum_{i=1}^n h_i(\Delta^\top x_i)\varepsilon_i, \end{aligned} \tag{57}$$

其中，多元函数 $h_i : \mathbb{R}^K \to \mathbb{R}$ 定义为

$$h_i(u) = u^\top \left(\nabla f(\Theta^{*\top} x_i + u) - \nabla f(\Theta^{*\top} x_i)\right)^\top \epsilon_i. \tag{58}$$

同时，对 $-\gamma(x_i,\ \epsilon_i)$ 应用引理 A.7 也可得到

$$\mathbb{E} g_t^{(2)}((x_1, \epsilon_1), \ldots, (x_n, \epsilon_n)) \leq \frac{4}{nt}\mathbb{E} \sup_{\frac{t}{2} \leq \|\Delta\|_{\omega,2,1} \leq t} \sum_{i=1}^n h_i(\Delta^\top x_i)\varepsilon_i.$$

T以下引理表明， $h_i(u)$ 是 L-Liptchitz 的，其中 $L = K + \sqrt{KR^*C_x + \overline{2_\circ}}$

引理 A.8。对于任意 $u \in \mathbb{R}^K,\ 1 \leq i \leq n,\ \|\nabla h_i\ (u)\ \|_2 \leq L = K + \sqrt{\overline{K}R^*C_x + 2_\circ}$

Maurer [2016] 证明了以下引理，该引理给出了当压缩函数具有向量值定义域时 Rademacher 平均的压缩不等式:

引理 A.9 (Maurer [2016])。 对于任意可数集 $\mathcal{S}$ 以及满足下列条件的函数 $\psi_i :\ \mathcal{S} \to \mathbb{R},\ \phi_i :\ \mathcal{S} \to \ell_2 1 \leq i \leq n$

$$\forall s, s' \in \mathcal{S}, \psi_i(s) - \psi_i(s') \leq \|\phi_i(s) - \phi_i(s')\|,$$

我们有

$$\mathbb{E} \sup_{s \in \mathcal{S}} \sum_{i=1}^n \varepsilon_i \psi_i(s) \leq \sqrt{2}\mathbb{E} \sup_{s \in \mathcal{S}} \sum_{i,k} \widetilde{\epsilon}_{ik} \phi_i(s)_k,$$

其中 ℓ_2 是实数平方可和序列构成的希尔伯特空间， $\varepsilon_i,\ \widetilde{\varepsilon}_{ik}$ 是关于 $1 \leq i \leq n$ 和 $1 \leq k < \infty$相互独立的 *Rademacher* 随机变量，而 $\phi_i(s)_k$ 是 k-*th* 坐标 *of*$\phi_i(s)_k$。

我们通过令 $\mathcal{S} = \{\Delta \in \mathbb{Q}^{p \times K} : \frac{t}{2} \leq \|\Delta\|_{\omega,2,1} \leq t\}$, $\psi_i(s) = h_i(s^\top x_i)$, $(\phi_i(s))_{1:K} = Ls^\top x_i$ 和 $(\phi_i(s))_{(K+1):\infty} = 0$ 来应用引理 A.9。然后可以证明

$$\begin{aligned} \mathbb{E} \sup_{\Delta \in \mathcal{S}} \sum_{i=1}^n h_i(\Delta^\top x_i)\varepsilon_i \leq &\sqrt{2}L\mathbb{E} \sup_{\Delta \in \mathcal{S}} \sum_{i=1}^n \sum_{k=1}^K \Delta_{:,k}^\top x_i \widetilde{\varepsilon}_{ik} \\ \leq &\sqrt{2}Lt\mathbb{E} \left\| \sum_{i=1}^n x_i \widetilde{\varepsilon}_i^\top \right\|_{\omega^{-1},2,\infty}. \end{aligned}$$

注意， $\mathcal{S}$ 在 $\{\Delta \in \mathbb{R}^{p\times K} : \frac{t}{2} \le \|\Delta\|_{\omega,2,1} \le t\}$ 中稠密，且 $\sum_{i=1}^n h_i(\Delta^\top x_i)\varepsilon_i$ 是关于 Δ的连续函数，因此对于 $j = 1,2$，有

$$\begin{aligned}
&\mathbb{E}g_t^{(j)}((x_1,\epsilon_1),\dots,(x_n,\epsilon_n))\\
=&\frac{4}{nt}\mathbb{E}\sup_{\frac{t}{2}\le\|\Delta\|_{\omega,2,1}\le t}\sum_{i=1}^n h_i(\Delta^\top x_i)\varepsilon_i\\
=&\frac{4}{nt}\mathbb{E}\sup_{\Delta\in\mathcal{S}}\sum_{i=1}^n h_i(\Delta^\top x_i)\varepsilon_i\\
\le&\frac{4\sqrt{2}L}{n}\mathbb{E}\left\|\sum_{i=1}^n x_i\widetilde{\varepsilon}_i^\top\right\|_{\omega^{-1},2,\infty}.
\end{aligned} \tag{59}$$

下面的引理为 $\mathbb{E}\left\|\sum_{i=1}^n x_i\widetilde{\varepsilon}_i^\top\right\|_{\omega^{-1},2,\infty}$ 提供了一个上界。

引理 A.10。

$$\mathbb{E}\left\|\sum_{i=1}^n x_i\widetilde{\varepsilon}_i^\top\right\|_{\omega^{-1},2,\infty} \le C\lambda_{\max}^{\frac{1}{2}}(\Sigma)\sqrt{(m+\log J)n}.$$

回到（59），我们知道对于 $j = 1,2$,

$$\mathbb{E}g_t^{(j)}((x_1,\epsilon_1),\dots,(x_n,\epsilon_n)) \le CL\lambda_{\max}^{\frac{1}{2}}(\Sigma)\sqrt{\frac{m+\log J}{n}}. \tag{60}$$

注意到

$$\begin{aligned}
&\frac{\left|x_i^\top\Delta(\nabla f(\Theta^{*\top}x_i+\Delta^\top x_i)-\nabla f(\Theta^{*\top}x_i))\epsilon_i\right|}{n\|\Delta\|_{\omega,2,1}}\\
\le&\frac{\left|x_i^\top\Delta(\nabla A(\Theta^{*\top}x_i+\Delta^\top x_i)-\nabla A(\Theta^{*\top}x_i))\right|}{n\|\Delta\|_{\omega,2,1}}\\
\le&\frac{1}{n}\|x_i(\nabla A(\Theta^{*\top}x_i+\Delta^\top x_i)-\nabla A(\Theta^{*\top}x_i))^\top\|_{\omega^{-1},2,\infty}\\
\le&\frac{C}{n}\max_j\|x_{i,\mathcal{R}_j}\|_2\|\nabla A(\Theta^{*\top}x_i+\Delta^\top x_i)_{\mathcal{C}_j}-\nabla A(\Theta^{*\top}x_i)_{\mathcal{C}_j}\|_2\\
\le&\frac{CC_x\max_j\sqrt{r_j}}{n},
\end{aligned} \tag{61}$$

这意味着对于任意 $\{(x_i,\epsilon_i)\}_{i=1}^n$、$(x_i',\epsilon_i')$，$j = 1,2$,

$$\begin{aligned}
&\Big|g_t^{(j)}((x_1,\epsilon_1),\dots,(x_i,\epsilon_i),\dots,(x_n,\epsilon_n))\\
&-g_t^{(j)}((x_1,\epsilon_1),\dots,(x_i',\epsilon_i'),\dots,(x_n,\epsilon_n))\Big|\\
\le&\frac{CC_x\max_j\sqrt{r_j}}{n}.
\end{aligned}$$

因此，我们可以应用有界差分不等式 [McDiarmid, 1989] 并得到以下结果:

$$\sup_{\frac{t}{2}\le\|\Delta\|_{\omega,2,1}\le t}|\phi(\Delta,\{x_i\}_{i=1}^n,\{\epsilon_i\}_{i=1}^n)| \le CL\lambda_{\max}^{\frac{1}{2}}(\Sigma)\sqrt{\frac{m+\log J}{n}}, \tag{62}$$

以至少 $1-2\exp\{-cL^2C_x^{-2}\lambda_{\max}(\Sigma)\min_j r_j^{-1}(m+\log J)\}$ 的概率。现在，我们应用剥离论证，将概率界 (62) 扩展到所有 $\Delta \in \mathbb{B}_{1,\infty}(R_0)$。由于$\|\Delta\|_{\omega,2,1} \le \max_j \omega_j K\|\Delta\|_{1,\infty}$,

$\mathbb{B}_{1,\ \infty}(R_0) \subset \mathbb{B}_{\omega,\ 2,\ 1}(\max_j \omega_j K R_0)$。令 $C_n = \frac{C}{C_x^2} L \lambda_{-}^{\frac{1}{2}} \max(\Sigma) \min_j \omega_j^2 \sqrt{\frac{\log J+m}{m^2 n}}$, $N = \log_2(\frac{\max_j \omega_j K R_0}{C_n})$，那么可以证明

$$\begin{aligned}\mathbb{B}_{1,\infty}(R_0) \subset & \cup_{k=1}^{N} \{\Delta : 2^{k-1} C_n \leq \|\Delta\|_{\omega,2,1} \leq 2^k C_n\} \\ & \cup \{\Delta : 0 \leq \|\Delta\|_{\omega,2,1} \leq C_n\}.\end{aligned} \tag{63}$$

首先，我们考虑如何对 $\{\Delta$ 一致地建立上界：$0 \leq \|\Delta\|_{\omega,\ 2,\ 1} \leq C_n\}$。对于任意满足 $\|d\|_{\omega,\ 2,\ 1} = 1$ 的 $r > 0$, $d \in \mathbb{R}^{p \times K}$，令 $\widetilde{\phi}(r,\ d,\ \{x_i\}_{i=1}^n,\ \{\epsilon_i\}_{i=1}^n) = \phi(rd,\ \{x_i\}_{i=1}^n,\ \{\epsilon_i\}_{i=1}^n)$。首先注意到

$$\begin{aligned}& \left| \frac{\partial \widetilde{\phi}(r, d, \{x_i\}_{i=1}^n, \{\epsilon_i\}_{i=1}^n)}{\partial r} \right| \\ =& \left| \frac{1}{n} \sum_{i=1}^n x_i^\top d \nabla^2 A(\Theta^{*\top} x_i + r d^\top x_i) d^\top x_i 1_K^\top \epsilon_i \right| \\ \leq & \frac{1}{n} \sum_{i=1}^n \|x_i^\top d\|_2^2 \\ \leq & C_x^2 \|d\|_1^2 \\ \leq & C_x^2 \max_j \omega_j^{-2} m,\end{aligned} \tag{64}$$

这意味着，对于任意 $\|\Delta\|_{\omega,\ 2,\ 1} \leq C_n$,

$$\begin{aligned}& |\phi(\Delta, \{x_i\}_{i=1}^n, \{\epsilon_i\}_{i=1}^n)| \\ =& \left| \widetilde{\phi}(\|\Delta\|_{\omega,2,1}, \frac{\Delta}{\|\Delta\|_{\omega,2,1}}, \{x_i\}_{i=1}^n, \{\epsilon_i\}_{i=1}^n) \right| \\ \leq & \left| \widetilde{\phi}(C_n, \frac{\Delta}{\|\Delta\|_{\omega,2,1}}, \{x_i\}_{i=1}^n, \{\epsilon_i\}_{i=1}^n) \right| + C_x^2 \max_j \omega_j^{-2} m C_n \\ \leq & \sup_{\|\Delta\|_{\omega,2,1} = C_n} |\phi(\Delta, \{x_i\}_{i=1}^n, \{\epsilon_i\}_{i=1}^n)| + C L \lambda_{\max}^{\frac{1}{2}}(\Sigma) \sqrt{\frac{\log J + m}{n}}.\end{aligned}$$

因此,

$$\begin{aligned}& \sup_{\Delta \in \mathbb{B}_{1,\infty}(R_0)} |\phi(\Delta, \{x_i\}_{i=1}^n, \{\epsilon_i\}_{i=1}^n)| \\ \leq & \max_{0 \leq k \leq N} \sup_{2^{k-1} C_n \leq \|\Delta\|_{\omega,2,1} \leq 2^k C_n} |\phi(\Delta, \{x_i\}_{i=1}^n, \{\epsilon_i\}_{i=1}^n)| \\ & + C L \lambda_{\max}^{\frac{1}{2}}(\Sigma) \sqrt{\frac{\log J + m}{n}}.\end{aligned} \tag{65}$$

现在注意到

$$\begin{aligned}N = & \log_2 \frac{C_x^2 R_0 K m \max_j \omega_j}{L \lambda_{\max}^{\frac{1}{2}}(\Sigma) \min_j \omega_j^2} \sqrt{\frac{n}{\log J + m}} \\ \leq & \frac{1}{2} \log_2 n + \log_2 m + \log_2 K + \log_2 \frac{C_x \max_j \omega_j}{\lambda_{\max}^{\frac{1}{2}}(\Sigma)} + C \\ \leq & C \log n,\end{aligned}$$

其中我们应用了以下事实:

$$R_0 \leq \frac{\min_j \frac{n_j}{\pi_j n_u}}{4 C_x (1 + \max_j \frac{n_j}{\pi_j n_u})^2} \leq \frac{1}{16 C_x},$$

$L \geq 1$，在第二行取$_j\, \omega_j \geq c$ 的最小值，并应用 K，$m \leq Cn$，

$$C_x \max_j \omega_j \lambda_{\max}^{-\frac{1}{2}}(\Sigma) \leq n^C$$

在第三行中。对 $0 \leq k \leq N$ 取并集界，这意味着

$$\sup_{\Delta \in \mathbb{B}_{1,\infty}(R_0)} |\phi(\Delta, \{x_i\}_{i=1}^n, \{\epsilon_i\}_{i=1}^n)| \leq CL\lambda_{\max}^{\frac{1}{2}}(\Sigma)\sqrt{\frac{\log J + m}{n}}, \tag{66}$$

以至少为此的概率

$$\begin{aligned} & 1 - C\log n \exp\{-cL^2 C_x^{-2}\lambda_{\max}(\Sigma) \min_j r_j^{-1}(\log J + m)\} \\ \geq & 1 - \exp\{C\log\log n - cL^2 C_x^{-2}\lambda_{\max}(\Sigma) \min_j r_j^{-1}(\log J + m)\} \\ \geq & 1 - \exp\{-cK^2 \min_j r_j^{-1}(\log J + m)\}, \end{aligned} \tag{67}$$

其中最后一行是由于

$$\begin{aligned} & L^2 C_x^{-2}\lambda_{\max}(\Sigma) \min_j r_j^{-1}(\log J + m) \\ \geq & cK^2 \min_j r_j^{-1}(\log J + m) \\ \geq & C\log\log n. \end{aligned}$$

□

引理 *A.3* 的证明。回顾一下，假设 3.3 要求 $R_0 \leq$最小值 $\frac{{}_j\, \frac{n_j}{\pi_j n_u} e^{-C_x R^*}(\text{最大值}_j\, \frac{n_j}{\pi_j n_u})(1+1.1Ke^{C_x R^*})^3}{2}$。我们首先证明，在此约束下，对于 R_0，有 $h^{\mathrm{MN}}(R_0) > 0$。注意到 $\frac{4C_x\ 1+}{\text{最小值}_j\, \frac{n_j}{\pi_j n_u}(1+\text{最大值}_j\, \frac{n_j}{\pi_j n_u})^2} < 1$、$e^{-C_x R^*} \leq 1$、$1 + 1.1Ke^{C_x R^*} \geq 3.2$，我们知道 $e^{C_x R_0} < e^{\frac{1}{4\times 3.2^3}} < 1.01$。因此，

$$\begin{aligned} h^{\mathrm{MN}}(R_0) &= \frac{e^{-C_x(R_0+R^*)} \min_j \frac{n_j}{\pi_j n_u}}{(1+\max_j \frac{n_j}{\pi_j n_u})^2 \left(1 + Ke^{C_x(R_0+R^*)}\right)^3} - 4C_x R_0 \\ &> \frac{e^{-C_x R^*} \min_j \frac{n_j}{\pi_j n_u}}{1.01(1+\max_j \frac{n_j}{\pi_j n_u})^2 \left(1 + 1.01Ke^{C_x R^*}\right)^3} - 4C_x R_0 \\ &\geq \frac{\min_j \frac{n_j}{\pi_j n_u} e^{-C_x R^*}}{(1+\max_j \frac{n_j}{\pi_j n_u})^2 \left(1 + 1.1Ke^{C_x R^*}\right)^3} - 4C_x R_0 \\ &\geq 0, \end{aligned}$$

其中最后一行是由于假设 3.3，倒数第二行是由于以下事实：

$$\frac{1 + 1.1Ke^{C_x R^*}}{1 + 1.01Ke^{C_x R^*}} \geq 1 + \frac{0.09Ke^{C_x R^*}}{1 + 1.01Ke^{C_x R^*}} \geq 1 + \frac{0.09Ke^{C_x R^*}}{1.51Ke^{C_x R^*}} = 1.059 > (1.01)^{1/3},$$

我们还应用了上述不等式中的事实 $K \geq 2$。

至于引理 A.3 中的第二个结论，我们注意到先前的论证表明，对于某个常数 $c > 0$，$C_x R_0 < c$。此外，根据 $h^{\mathrm{MN}}(R_0)$的定义，不难看出 $0 < h^{\mathrm{MN}}(R_0) < 1$。进一步地，假设 3.1 和 3.2 表明 $C\overline{\lambda^{\frac{1}{2}}_{\text{最小值}}(\Sigma)} \leq C_x$ 且最大值$_j\, \omega_j \leq C\sqrt{{}^{n}{}_{\text{对数}}\ J+m}$，这意味着

$$R_0 \leq \frac{c\lambda_{\max}^{\frac{1}{2}}(\Sigma)(K + C_x R^* \sqrt{K})}{h^{\mathrm{MN}}(R_0)\lambda_{\min}(\Sigma) \max_j \omega_j K} \sqrt{\frac{n}{\log J + m}}.$$

因此，

$$\text{因此 } \tau_1 \log_n^{J+m} \|\widehat{\Delta}\|_{\omega,\ 2,\ 1}^2 \leq \tau_2 \sqrt{\log_n^{J+m}} \|\widehat{\Delta}\|_{\omega,\ 2,\ 1}。 \quad \|\widehat{\Delta}\|_{\omega,2,1} \leq K \max_j \omega_j R_0 \leq \frac{\tau_2}{\tau_1} \sqrt{\frac{n}{\log J + m}}, \tag{68}$$

□

B 定理 3.2 的证明

第 4 节中的引理 4.4 以及以下引理是证明定理 3.2 的关键构件。

引理 B.1 (序数-PU 模型下的偏差界)。 若数据集 $\{(x_i, z_i)\}_{i=1}^n$ 是在病例对照设置下由序数 *PU* 模型生成的，则

$$\begin{aligned} \|(\nabla \mathcal{L}_n^{\mathrm{ON}}(\Theta^*))_{1:p}\|_{\omega^{-1},2,\infty} \leq & C h^{\mathrm{ON}}(0,0) \lambda_{\max}^{\frac{1}{2}}(\Sigma) \sqrt{\frac{\log J + m}{n}}, \\ \|(\nabla \mathcal{L}_n^{\mathrm{ON}}(\theta^*))_{(p+1):(p+K)}\|_\infty \leq & C h^{\mathrm{ON}}(0,0) \sqrt{\frac{\log J + m + \log K}{n}}, \end{aligned} \tag{69}$$

其概率至少为 $1 - 3\exp\{-(\log J + m)\}$。其中， $h^{\mathrm{ON}}(\cdot,\cdot)$ 定义如式 (13)所示。

而对于引理 4.4，常数 α 和 L 的具体形式如下： $\alpha = \frac{1}{2}\gamma \min\{\lambda_{\min}(\Sigma)\ 1\}$，其中 –

$$\gamma = \min_j \frac{n_j}{\pi_j n_u} \frac{e^{2(C_x+1)(R^*+R_0)}}{32 h^{\mathrm{ON}}(R_0, r_0)(1 + e^{(C_x+1)(R^*+R_0)})^4 K},$$

以及 $L = (\frac{4}{3}(C_x+1)R_0 + 1)\sqrt{K}\overline{h^{\mathrm{ON}}}(R_0,\ r_0) + \frac{(C_x+1)}{4} R_0 \sqrt{K}(h^{\mathrm{ON}}(R_0,\ r_0))^2$。

现在，我们可以基于引理 B.1 和引理 4.4 来证明定理 3.2。

定理 *3.2* 的证明。 与定理 3.1 的证明类似，可以证明

$$\begin{aligned} &\langle \nabla \mathcal{L}_n^{\mathrm{ON}}(\widehat{\theta}) + \lambda \nabla \|\widehat{\theta}_{1:p}\|_{\omega,2,1}, \theta^* - \widehat{\theta} \rangle \geq 0, \\ &\langle \nabla \mathcal{L}_n^{\mathrm{ON}}(\widehat{\theta}) - \nabla \mathcal{L}_n^{\mathrm{ON}}(\theta^*), \widehat{\theta} - \theta^* \rangle \leq \langle -\nabla \mathcal{L}_n^{\mathrm{ON}}(\theta^*) - \lambda \nabla \|\widehat{\theta}_{1:p}\|_{\omega,2,1}, \widehat{\theta} - \theta^* \rangle. \end{aligned} \tag{70}$$

$\widehat{\ }$ $\widehat{\ }$其中 $\nabla\|\theta_{1:p}\|_{\omega,\ 2,\ 1}$ 是 $\|\theta_{1:p}\|_{\omega,\ 2,\ 1}$ 关于 θ 的任意次梯度。令 $\Delta = \theta - \theta^*$，且 $\mathcal{G}_S = \cup_{j \in S} \mathcal{G}_j$其中 S 如前所定义，即 $S = \{j : \theta^*_{\mathcal{G}_j} \neq 0\}$。根据引理 B.1 和引理 4.4,

$$\begin{aligned} &\alpha \|\widehat{\Delta}\|_2^2 - C\alpha \frac{\log J + m}{n} \|\widehat{\Delta}_{1:p}\|_{\omega,2,1}^2 \\ &- CL \left(\lambda_{\max}^{\frac{1}{2}}(\Sigma) \sqrt{\frac{\log J + m}{n}} \|\widehat{\Delta}_{1:p}\|_{\omega,2,1} + \sqrt{\frac{K(\log J + \log(2K))}{n}} \|\widehat{\Delta}\|_2 \right) \\ \leq & C h^{\mathrm{ON}}(0,0) \lambda_{\max}^{\frac{1}{2}}(\Sigma) \sqrt{\frac{\log J + m}{n}} \|\widehat{\Delta}_{1:p}\|_{\omega,2,1} \\ &+ C h^{\mathrm{ON}}(0,0) \sqrt{\frac{K(\log J + m + \log K)}{n}} \|\widehat{\Delta}\|_2 + \lambda \langle \nabla \|\widehat{\theta}_{1:p}\|_{\omega,2,1}, \theta^* - \widehat{\theta} \rangle. \end{aligned} \tag{71}$$

根据 L 的定义， $L \geq h^{\mathrm{ON}}(R_0,\ r_0)\sqrt{K} \geq \overline{h}^{\mathrm{ON}}(0,\ 0)$，因此上述不等式可转化为

$$\begin{aligned} \alpha \|\widehat{\Delta}\|_2^2 \leq & C\alpha \frac{\log J + m}{n} \|\widehat{\Delta}_{1:p}\|_{\omega,2,1}^2 + CL \lambda_{\max}^{\frac{1}{2}}(\Sigma) \sqrt{\frac{\log J + m}{n}} \|\widehat{\Delta}_{1:p}\|_{\omega,2,1} \\ &+ CL \sqrt{\frac{K(\log J + m + \log K)}{n}} \|\widehat{\Delta}\|_2 + \lambda \langle \nabla \|\widehat{\theta}_{1:p}\|_{\omega,2,1}, \theta^* - \widehat{\theta} \rangle. \end{aligned}$$

同时，根据 α，L 的定义、假设 3.2 以及 $h^{\mathrm{ON}}(R_0,\ r_0) \geq 1$ 这一事实，可以证明 $\alpha \leq C$，$L \geq 1$，$\max_j \omega_j \leq \sqrt{n \log J + m}$。此外，根据假设 3.1 和 3.2，有 $\lambda_{\max}(\Sigma) \leq C\sigma^2 \leq C \min\{\lambda_{\min}(\Sigma)\ 1\}$，这意味着 $0 < C_1 \leq \lambda_{\min}(\Sigma) \leq \lambda_{\max}(\Sigma) \leq C_2$。假设 3.4 还意味着，对于某个常数 $C > 0$，$R_0 \leq C$。因此我们知道

$$\begin{aligned}\|\widehat{\Delta}_{1:p}\|_{\omega,2,1} \leq & \max_j \omega_j \|\widehat{\Delta}_{\mathcal{G}_j}\|_1 \\ \leq & \max_j \omega_j R_0 \\ \leq & \frac{C}{\alpha} L \lambda_{\max}^{\frac{1}{2}}(\Sigma) \sqrt{\frac{n}{\log J + m}},\end{aligned}$$

这进一步意味着 $\alpha \sqrt{\text{对数} \frac{J+m}{n}} \|\widehat{\Delta}_{1:p}\|_{\omega,\ 2,\ 1}^2 \leq CL\lambda^2_{\text{最大值}}(\Sigma) \sqrt{\text{对数} \frac{J+m}{n}} \|\widehat{\Delta}_{1:p}\|_{\omega,\ 2,\ 1}$。此外，

$$\begin{aligned}\langle \nabla \|\widehat{\theta}_{1:p}\|_{\omega,2,1}, \theta^* - \widehat{\theta} \rangle \leq & \|\theta^*_{1:p}\|_{\omega,2,1} - \|\widehat{\theta}_{1:p}\|_{\omega,2,1} \\ \leq & \|\widehat{\Delta}_{\mathcal{G}_S}\|_{\omega_S,2,1} - \|\widehat{\Delta}_{\mathcal{G}_S^c}\|_{\omega_{S^c},2,1}.\end{aligned}$$

因此，可以证明

$$\begin{aligned}\alpha \|\widehat{\Delta}\|_2^2 \leq & \frac{3\lambda}{2} \|\widehat{\Delta}_{\mathcal{G}_S}\|_{\omega_S,2,1} + CL\sqrt{\frac{K(\log J + m + \log K)}{n}} \|\widehat{\Delta}\|_2 \\ \leq & \left(\frac{3\lambda \|\omega_S\|_2}{2} + CL\sqrt{\frac{K(\log J + m + \log K)}{n}} \right) \|\widehat{\Delta}\|_2, \\ \|\widehat{\Delta}\|_2 \leq & \frac{3\|\omega_S\|_2 \lambda}{\gamma \min\{\lambda_{\min}(\Sigma), 1\}} + \frac{CL}{\gamma \min\{\lambda_{\min}(\Sigma), 1\}} \sqrt{\frac{K(\log J + m + \log K)}{n}},\end{aligned}$$

其概率至少为 $1 - \exp\{-\text{对数}_{C_x^2 m+1}^{cK} J\} = 3\exp\{-(\log\ J + m)\}$。令 $\gamma_0 = \max_j \frac{\pi_j n_u}{n_j} K\gamma$ 和 $L_0 = \frac{L}{\sqrt{K}}$，即可得到最终结果。 □

B.1 引理 B.1 和引理 4.4 的证明

回顾函数的定义 ： $\mathbb{R}^K \to \mathbb{R}^K$ 和 u： $\mathbb{R}^p \times \mathbb{R}^{p+K} \to \mathbb{R}^K$，首次定义于第 4 节 .2:

$$(f^{\mathrm{ON}}(u))_j = \begin{cases} \log \frac{n_j}{\pi_j n_u} + \log \left[(1 + e^{\sum_{l=1}^{j+1} u_l})^{-1} - (1 + e^{\sum_{l=1}^{j} u_l})^{-1} \right], & 1 \leq j < K, \\ \log \frac{n_K}{\pi_K n_u} + \log \left[1 - (1 + e^{\sum_{l=1}^{K} u_l})^{-1} \right], & j = K, \end{cases}$$

$$u(x_i, \theta) = (x_i^\top \theta_{1:p} - \theta_{p+1}, -\theta_{p+2}, \ldots, -\theta_{p+K})^\top.$$

对数似然损失 $\mathcal{L}_n^{\mathrm{ON}}(\theta)$ 可写为:

$$\mathcal{L}_n^{\mathrm{ON}}(\theta) = \frac{1}{n} \sum_{i=1}^n \log \left(1 + \sum_{k=1}^K e^{f^{\mathrm{ON}}(u(x_i,\theta))} \right) - \sum_{k=1}^K \mathbb{1}_{\{z_i = k\}} f^{\mathrm{ON}}(u(x_i, \theta)).$$

Al那么回顾一下首次在第 C.3 节中定义的函数 p: $\mathbb{R}^K \to \mathbb{R}^K$ 和 α： $\mathbb{R}^K \to \mathbb{R}^{K+1}$:

$$\alpha_j(u) = \begin{cases} e^{\sum_{l=1}^{j} u_l} \left(1 + e^{\sum_{l=1}^{j} u_l} \right)^{-2}, & j \leq K, \\ 0, & j = K+1, \end{cases}$$

and

$$p_j(u) = \begin{cases} (1 + e^{\sum_{l=1}^{j+1} u_l})^{-1} - (1 + e^{\sum_{l=1}^{j} u_l})^{-1}, & 1 \leq j < K, \\ 1 - (1 + e^{\sum_{l=1}^{j} u_l})^{-1}, & j = K. \end{cases}$$

这两个函数决定了函数 f^{ON} 的导数，因此会影响我们的偏差界以及曲率下界。引理 C.2 给出了 $p(\cdot)$ $\alpha(\cdot)$ 的取值范围以及 $f^{\mathrm{ON}}(\cdot)$ 的导数，这将有助于我们证明引理 B.1 和引理 4.4。

引理 *B.1* 的证明。首先注意到，

$$\begin{aligned}\nabla\mathcal{L}_n^{\mathrm{ON}}(\theta^*) =& -\frac{1}{n}\sum_{i=1}^{n}\sum_{k'=1}^{K}(\delta_{ik'}-\mathbb{E}(\delta_{ik'}|x_i))\frac{\partial f_{k'}(u(x_i,\theta^*))}{\partial\theta^*}\\ =& -\frac{1}{n}\sum_{i=1}^{n}\begin{pmatrix} x_i & 0 & \dots & 0\\ -1 & 0 & \dots & 0\\ 0 & -1 & \dots & 0\\ 0 & 0 & \ddots & 0\\ 0 & 0 & \dots & -1\end{pmatrix}(\nabla f^{\mathrm{ON}}(u_i^*))^\top\epsilon_i,\end{aligned} \tag{72}$$

其中 $\delta_{ik}=\mathbb{1}_{\{z_i=k\}}$, $\epsilon_{ik'}=\delta_{ik'}-\mathbb{E}(\delta_{ik'}|x_i)$, $\epsilon_i=(\epsilon_{i1},\dots,\epsilon_{iK})^\top$, $u_i^*=u(x_i,\theta^*)\in\mathbb{R}^K$。如第 C.3 节所示，对于任意 $1\le j,k\le K$,

$$(\nabla f^{\mathrm{ON}}(u_i^*))_{jk}=p_j(u_i^*)^{-1}\left[\alpha_j(u_i^*)\mathbb{1}_{\{k\le j\}}-\alpha_{j+1}(u_i^*)\mathbb{1}_{\{k\le j+1\}}\right],$$

其中函数 $p(\cdot)$ 和 $\alpha(\cdot)$ 定义于 (123) 和 (122)，而 $p_j(\cdot)$ $\alpha_j(\cdot)$ 分别表示函数 $p(\cdot)$ 和 $\alpha(\cdot)$ 的第 j 个坐标。令 $\widetilde{\epsilon}\in\mathbb{R}^K$，其中 $\widetilde{\epsilon}_{ik}=(\nabla f^{\mathrm{ON}}(u_i^*)_{:,k})^\top\epsilon_i$, $1\le k\le K$。于是可以证明

$$\begin{aligned}|\widetilde{\epsilon}_{i1}|\le&\left|\sum_{j=1}^{K}p_j(u_i^*)^{-1}[\alpha_j(u_i^*)-\alpha_{j+1}(u_i^*)]\delta_{ij}\right|\\ &+\left|\sum_{j=1}^{K}p_j(u_i^*)^{-1}[\alpha_j(u_i^*)-\alpha_{j+1}(u_i^*)]\mathbb{E}(\delta_{ij}|x_i)\right|\\ \le&2\max_j\left|p_j(u_i^*)^{-1}[\alpha_j(u_i^*)-\alpha_{j+1}(u_i^*)]\right|\\ \le&\frac{1}{2}\max_j p_j(u_i^*)^{-1},\end{aligned} \tag{73}$$

其中，最后一行是因为如引理 C.2 所示，$0<\alpha_j(u_i^*)<\frac{1}{4}$。而对于 $2\le k\le K$,

$$\begin{aligned}|\widetilde{\epsilon}_{ik}|\le&\left|\sum_{j=k}^{K}p_j(u_i^*)^{-1}(\alpha_j(u_i^*)-\alpha_{j+1}(u_i^*))\delta_{ij}-p_{k-1}(u_i^*)^{-1}\alpha_k(u_i^*)\delta_{i,k-1}\right|\\ &+\left|\sum_{j=k}^{K}p_j(u_i^*)^{-1}(\alpha_j(u_i^*)-\alpha_{j+1}(u_i^*))\mathbb{E}\delta_{ij}-p_{k-1}(u_i^*)^{-1}\alpha_k(u_i^*)\mathbb{E}\delta_{i,k-1}\right|\\ \le&\frac{1}{2}\max_j p_j(u_i^*)^{-1}.\end{aligned} \tag{74}$$

因此，根据引理 C.2，$\|\widetilde{\epsilon}_i\|_\infty\le\frac{1}{2}h^{\mathrm{ON}}(0,0)$。因此，

$$\|(\nabla\mathcal{L}_n^{\mathrm{ON}}(\theta^*))_{(p+1):(p+K)}\|_\infty=\max_k\left|\frac{1}{n}\sum_{i=1}^{n}\widetilde{\epsilon}_{ik}\right|\le h^{\mathrm{ON}}(0,0)\sqrt{\frac{\log J+m+\log K}{2n}},$$

以至少 $1-K\exp\{-(\log J+m+\log K)\}\ge 1-\exp\{-(\log J+m)\}$ 的概率成立。对于界定$\|(\nabla\mathcal{L}_n^{\mathrm{ON}}(\theta^*))_{1:p}\|_{\omega,2,1}$，可以证明，对于任意 $1\le j\le J$,

$$\left\|\left(\nabla\mathcal{L}_n^{\mathrm{ON}}(\theta^*)\right)_{\mathcal{G}_j}\right\|_2=\left\|\frac{1}{n}X_{\mathcal{G}_j}^\top\widetilde{\epsilon}_{:,1}\right\|_2.$$

这里我们采用引理 A.1 的证明中用于界定 $\left\|X_{\mathcal{G}_j}^\top \widetilde{\epsilon}_{:,1}\right\|_2$ 的类似论证。首先注意到

$$\begin{aligned}\mathbb{E}(\|X_{\mathcal{G}_j}^\top \widetilde{\epsilon}_{:,1}\|_2|X) \leq & \left(\mathbb{E}(\|X_{\mathcal{G}_j}^\top \widetilde{\epsilon}_{:,1}\|_2^2|X)\right)^{\frac{1}{2}} \\ = & \left(\sum_{i=1}^n \mathbb{E}(\widetilde{\epsilon}_{i,1}^2|X)\left(X_{\mathcal{G}_j}X_{\mathcal{G}_j}^\top\right)_{ii}\right)^{\frac{1}{2}} \\ \leq & \frac{1}{2}h^{\mathrm{ON}}(0,0)\sqrt{\mathrm{tr}(X_{\mathcal{G}_j}X_{\mathcal{G}_j}^\top)} \\ \leq & \frac{1}{2}h^{\mathrm{ON}}(0,0)\sqrt{m}\|X_{\mathcal{G}_j}\|.\end{aligned} \tag{75}$$

设 $g(u) = \|X_{\mathcal{G}_j}^\top u\|_2$ 是从 $\mathbb{R}^n$ 到 $\mathbb{R}$ 的函数，则 g 是凸函数且为 $\|X_{\mathcal{G}_j}\|$-Lipschitz。应用 Talagrand 压缩不等式（Vershynin [2018] 中的定理 5.2.16）可得

$$\left\|\|X_{\mathcal{G}_j}^\top \widetilde{\epsilon}_{:,1}\|_2 - \mathbb{E}(\|X_{\mathcal{G}_j}^\top \widetilde{\epsilon}_{:,1}\|_2|X)\right\|_{\psi_2} \leq Ch^{\mathrm{ON}}(0,0)\|X_{\mathcal{G}_j}\|, \tag{76}$$

这进一步意味着

$$\mathbb{P}(\|X_{\mathcal{G}_j}^\top \widetilde{\epsilon}_{:,1}\|_2 > 2h^{\mathrm{ON}}(0,0)\|X_{\mathcal{G}_j}\|(\sqrt{m} + C\sqrt{nt}) \leq \exp\{-nt^2\}, \tag{77}$$

对任意 $t > 0$ 均成立。由引理 A.1 的证明中类似的论证可得：

$$\|X_{\mathcal{G}_j}\| \leq C\lambda_{\max}^{\frac{1}{2}}(\Sigma)\sqrt{n}, \tag{78}$$

以至少 $1 - 2J\exp\{-cn\} \geq 1 - \exp\{-cn\}$ 的概率对 $1 \leq j \leq J$ 成立；若在 (78) 中适当选择 $C > 0$。现在令 $t = \sqrt{2\log\frac{J+m}{n}}$，代入 (77)，则有

$$\begin{aligned}\|(\nabla\mathcal{L}_n^{\mathrm{ON}}(\theta))_{1:p}\|_{\omega^{-1},2,\infty} \leq & \max_j \omega_j^{-1}\left\|\frac{1}{n}X_{\mathcal{G}_j}^\top \widetilde{\epsilon}_{:,1}\right\|_2 \\ \leq & Ch^{\mathrm{ON}}(0,0)\lambda_{\max}^{\frac{1}{2}}(\Sigma)\sqrt{\frac{\log J + m}{n}},\end{aligned} \tag{79}$$

以至少 $1 - \exp\{-(\log J + m)\} - \exp\{-cn\}$ 的概率成立。 □

引理 *4.4* 的证明。根据 $\mathcal{L}_n^{\mathrm{ON}}(\theta)$的定义，可以证明

$$\begin{aligned}&\nabla\mathcal{L}_n^{\mathrm{ON}}(\theta) \\ =&\frac{1}{n}\sum_{i=1}^n \begin{pmatrix} x_i & 0 & \cdots & 0 \\ -1 & 0 & \cdots & 0 \\ 0 & -1 & \cdots & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & \cdots & -1 \end{pmatrix} \left(\nabla f^{\mathrm{ON}}(u(x_i,\theta))\right)^\top \left(\nabla A(f^{\mathrm{ON}}(u(x_i,\theta))) - \delta_i\right),\end{aligned} \tag{80}$$

其中 $\delta_i \in \mathbb{R}^K$ 满足 $\delta_{ik} = \mathbb{1}_{\{z_i=k\}}$。令 $\Delta = \theta - \theta^*$，则有

$$\begin{aligned}
&\langle \nabla \mathcal{L}_n^{\mathrm{ON}}(\theta) - \nabla \mathcal{L}_n^{\mathrm{ON}}(\theta^*), \theta - \theta^* \rangle \\
=&\frac{1}{n}\sum_{i=1}^{n} \left(\nabla A(f^{\mathrm{ON}}(u(x_i,\theta))) - \nabla A(f^{\mathrm{ON}}(u(x_i,\theta^*)))\right)^\top \nabla f^{\mathrm{ON}}(u(x_i,\theta)) \begin{pmatrix} x_i^\top \Delta_{1:p} - \Delta_{p+1} \\ -\Delta_{p+2} \\ \vdots \\ -\Delta_{p+K} \end{pmatrix} \\
&- \frac{1}{n}\sum_{i=1}^{n} \epsilon_i^\top \left(\nabla f^{\mathrm{ON}}(u(x_i,\theta)) - \nabla f^{\mathrm{ON}}(u(x_i,\theta^*))\right) \begin{pmatrix} x_i^\top \Delta_{1:p} - \Delta_{p+1} \\ -\Delta_{p+2} \\ \vdots \\ -\Delta_{p+K} \end{pmatrix} \\
=:&\mathrm{I}(\Delta) + \mathrm{II}(\Delta),
\end{aligned} \tag{81}$$

其中 $\epsilon_i = \delta_i - \nabla A(f^{\mathrm{ON}}(u(x_i,\ \theta^*)))$. 下文将给出 I(Δ) 的下界，并在可行集上对 II(Δ) 进行一致控制。

1. I(Δ) 的下界

引理 B.2. 对于任意 $\Delta \in S(R_0,\ r_0)$,

$$\mathrm{I}(\Delta) \geq \frac{\gamma}{n}\sum_{i=1}^{n}(x_i^\top \Delta_{1:p} - \Delta_{p+1})^2 + \gamma \|\Delta_{(p+2):(p+K)}\|_2^2, \tag{82}$$

其中 $\gamma = \text{最小值}_j \frac{n_j}{\pi_j n_+ 32 h^{\mathrm{ON}} R_0 r_0} \frac{e^{2(C_x+1)(R^*+R_0)}}{(,\ )(1+e^{(C_x+1)(R^*+R_0)})^4 K}$。

由引理 B.2,

$$\mathrm{I}(\Delta) \geq \frac{\gamma}{n}\sum_{i=1}^{n}\left((x_i^\top \Delta_{1:p} - \Delta_{p+1})^2 + \sum_{j=2}^{K}\Delta_{p+j}^2\right).$$

令 $\widetilde{\Delta} = \Delta_{1:(p+1)}$, $\widetilde{X} = \begin{pmatrix} x_1^\top & -1 \\ & \\ x_n^\top & -1 \end{pmatrix}$ …… 。下文将证明

$$\frac{\gamma}{n}\|\widetilde{X}\widetilde{\Delta}\|_2^2 \geq \alpha\|\widetilde{\Delta}\|_2^2 - C\alpha\frac{\log J + m}{n}\|\Delta_{1:p}\|_{\omega,2,1}^2, \tag{83}$$

其中 $\alpha = \frac{1}{2}\gamma \min\{\lambda_{\min}(\Sigma)\ 1\}$。由于 (83) 的左端非负，当 $\sim \quad \sim \|\Delta_{1:p}\|_{\omega,\ 2,}$ $_1 \geq \sqrt{\rho}\|\Delta\|_2$，其中 $\rho = C \quad ^n{}_{(\log\ J+m)}$ 时，(83) 显然成立。同时，由于

$$\frac{1}{n}\mathbb{E}\|\widetilde{X}\widetilde{\Delta}\|_2^2 = \widetilde{\Delta}^\top \begin{pmatrix} \Sigma & 0 \\ 0 & 1 \end{pmatrix} \widetilde{\Delta} \geq \min\{\lambda_{\min}(\Sigma), 1\}\|\widetilde{\Delta}\|_2^2,$$

我们只需证明

$$\sup_{\|\widetilde{\Delta}_{1:p}\|_{\omega,2,1} \leq \sqrt{\rho}, \|\widetilde{\Delta}\|_2 \leq 1} \frac{1}{n}\left|\|\widetilde{X}\widetilde{\Delta}\|_2^2 - \mathbb{E}\|\widetilde{X}\widetilde{\Delta}\|_2^2\right| \leq \frac{\min\{\lambda_{\min}(\Sigma), 1\}}{2}. \tag{84}$$

定义函数

$$g(\widetilde{\Delta},\widetilde{\Delta})=\frac{1}{n}\|\widetilde{X}\widetilde{\Delta}\|_2^2-\frac{1}{n}\mathbb{E}\|\widetilde{X}\widetilde{\Delta}\|_2^2,$$

然后，沿用引理 A.2 证明中的类似论证，可以得到

$$\begin{aligned}&\sup_{\|\widetilde{\Delta}_{1:p}\|_{\omega,2,1}\le\sqrt{\rho},\|\widetilde{\Delta}\|_2\le 1}\left|g(\widetilde{\Delta},\widetilde{\Delta})\right|\\ \le&9(1+2\max_j\omega_j^{-1})\sup_{S\subset[J],|S|=2\lfloor\rho\rfloor}\sup_{\widetilde{\Delta}\in\mathcal{N}_S}\left|g(\widetilde{\Delta},\widetilde{\Delta})\right|,\end{aligned}$$

其中 $\mathcal{N}_S$ 是 $\mathcal{K}(\mathcal{G}_S)=\{\widetilde{\Delta}:\|\widetilde{\delta}\|_2=1,\ \widetilde{\Delta}_{\mathcal{G}_j}=0$（若 $j\notin S\}$），且 $|\mathcal{N}_S|\le 21^{2\rho m+1}$ 的 $\frac{1}{10}$-网。下面的~~引理以高概率集中于每个固定的 Δ 的 $g(\Delta,\Delta)$。

引理 B.3. 对于任意满足 $\|\widetilde{\Delta}\|_2=1$ 的 $\widetilde{\Delta}\in\mathbb{R}^{p+1}$，

$$\mathbb{P}\left(\left|g(\widetilde{\Delta},\widetilde{\Delta})\right|\ge t\right)\le 2\exp\left\{-cn\min\left\{\frac{t^2}{\sigma^4},\frac{t^2}{\sigma^2},\frac{t}{\sigma^2}\right\}\right\}. \tag{85}$$

对于任意 $t>0$。

我们将引理 B.3 应用于 $t=\min\{\lambda_{\min}(\Sigma),1\}(\max_j\omega_j^{-1})$，并对 $\mathcal{N}_S$ 和 $S\subset[J]$ 取并集界，随后可证明 (84) 以至少.....的概率成立

$$\begin{aligned}&1-2(21)^{2\rho m+1}\binom{J}{2\lfloor\rho\rfloor}\exp\{-cn\}\\ \ge&1-2\exp\{(2\rho m+1)\log(21)+3\rho\log(eJ/2\lfloor\rho\rfloor)]-cn\}\\ \ge&1-2\exp\{(2\log(21)+3)\rho(m+\log J)+\log(21)-3\rho\log(2\lfloor\rho\rfloor)-cn\}\\ \ge&1-2\exp\{-\frac{cn}{2}\}\end{aligned}$$

其中，最后一行成立是因为 $\rho=\frac{cn}{(\log J+m)}$，因此只要将该常数 C 选得大于 $4\log(21)+6)$和 $\rho\ge 2$，就能保证 $(2\log(21)+3)\rho(m+\log J)\le\frac{c}{2}n$ 且 $\log(21)-3\rho\log(2\lfloor\rho\rfloor)$。因此，概率至少为 $1-2\exp\{-cn\}$，

$$\mathrm{I}(\Delta)\ge\alpha\|\Delta\|_2^2-C\alpha\frac{\log J+m_1}{n}\|\Delta_{1:p}\|_{\omega,2,1}^2.$$

2. 集中 II(Δ)

我们沿用引理 A.2 证明第二部分中的类似论证。首先定义

$$\phi(\Delta,\{x_i\}_{i=1}^n,\{\epsilon_i\}_{i=1}^n)=\frac{\mathrm{II}(\Delta)}{\tau(\Delta)},$$

and

$$\begin{aligned}g_t^{(1)}((x_1,\epsilon_1),\dots,(x_n,\epsilon_n))&=\sup_{\substack{\frac{t}{2}\le\tau(\Delta)\le t\\ \Delta\in\mathcal{S}(R_0,r_0)}}\phi(\Delta,\{x_i\}_{i=1}^n,\{\epsilon_i\}_{i=1}^n),\\ g_t^{(2)}((x_1,\epsilon_1),\dots,(x_n,\epsilon_n))&=\sup_{\substack{\frac{t}{2}\le\tau(\Delta)\le t\\ \Delta\in\mathcal{S}(R_0,r_0)}}-\phi(\Delta,\{x_i\}_{i=1}^n,\{\epsilon_i\}_{i=1}^n),\end{aligned}$$

对任意 $t>0$。我们将给出 $g_t^{(j)}$ 的上界，然后应用剥离论证。根据引理 A.7 ，

$$\mathbb{E}g_t^{(j)}((x_1,\epsilon_1),\dots,(x_n,\epsilon_n))\le\frac{4}{nt}\mathbb{E}\sup_{\substack{\frac{t}{2}\le\tau(\Delta)\le t\\ \Delta\in\mathcal{S}(R_0,r_0)}}\sum_{i=1}^n q_i(u_i(\Delta))\varepsilon_i,$$

其中 $q_i(v)=\epsilon_i^\top\left[\nabla f^{\mathrm{ON}}(u_i(\theta^*)+v)-\nabla f_{\mathrm{ON}}(u_i(\theta^*))\right]v$，为简洁起见，我们使用 $u_i(\theta)$ 表示 $u(x_i,\theta)$。下面的引理表明，$q_i(\cdot)$ 在我们关注的区域内是 L-Lipschitz 的。

引理 B.4。 对于任意满足 $\Delta \in \mathcal{S}(R_0, r_0)$的向量 $\Delta \in \mathbb{R}^{p+K}$， $1 \le i \le n$，

$$\|\nabla q_i(u_i(\Delta))\|_2 \le L,$$

其中 $L = (\frac{4}{3}(C_x+1)R_0+1)\sqrt{K}h^{\mathrm{ON}}(R_0,\ r_0) + \frac{(C_x+1)R_0}{4}\sqrt{K}(h^{\mathrm{ON}}(R_0,\ r_0))^2$。

然后根据引理 A.9 以及引理 A.2 的证明中所采用的类似论证，可以证明

$$\mathbb{E}g_t^{(j)}((x_1,\epsilon_1),\ldots,(x_n,\epsilon_n)) \le \frac{4\sqrt{2}L}{nt}\mathbb{E}\sup_{\substack{\frac{t}{2}\le\tau(\Delta)\le t\\ \Delta\in\mathcal{S}(R_0,r_0)}}\sum_{i=1}^n u_i(\Delta)^\top\widetilde{\varepsilon}_i, \tag{86}$$

其中 $\widetilde{\varepsilon}_i \in \mathbb{R}^K$和 $\widetilde{\varepsilon}_{ij}$、 $1 \le i \le n$、 $1 \le j \le K$是独立的 Rademacher 随机变量。可以证明

$$\begin{aligned}
&\frac{4\sqrt{2}L}{nt}\mathbb{E}\sup_{\substack{\frac{t}{2}\le\tau(\Delta)\le t\\ \Delta\in\mathcal{S}(R_0,r_0)}}\sum_{i=1}^n u_i(\Delta)^\top\widetilde{\varepsilon}_i\\
=&\frac{4\sqrt{2}L}{nt}\mathbb{E}\sup_{\substack{\frac{t}{2}\le\tau(\Delta)\le t\\ \Delta\in\mathcal{S}(R_0,r_0)}}\Delta^\top\sum_{i=1}^n\begin{pmatrix}x_i\widetilde{\varepsilon}_{i1}\\ -\widetilde{\varepsilon}_i\end{pmatrix}\\
\le&\frac{4\sqrt{2}L}{n}\mathbb{E}\left[\sqrt{\frac{n}{\lambda_{\max}(\Sigma)(\log J+m)}}\left\|\sum_{i=1}^n x_i^\top\widetilde{\varepsilon}_{i1}\right\|_{\omega^{-1},2,\infty}\right.\\
&\left.+\sqrt{\frac{n}{K(\log J+\log(2K))}}\left\|\sum_{i=1}^n\widetilde{\varepsilon}_i\right\|_2\right]
\end{aligned} \tag{87}$$

由于我们对协变量 x_i 作出了与多项情形相同的次高斯假设，沿用引理 A.10 证明中的相同论证可得

$$\mathbb{E}\max_{1\le j\le J}\omega_j^{-1}\|X_{\mathcal{G}_j}^\top\widetilde{\epsilon}_{:,1}\|_2 \le C\lambda_{\max}^{\frac{1}{2}}(\Sigma)\sqrt{(m+\log J)n}. \tag{88}$$

现在我们给出 $\mathbb{E}\max_k|\sum_{i=1}^n\widetilde{\epsilon}_{ik}|$的上界。由于 $\{\widetilde{\epsilon}_{ik}\}_{i=1}^n$是具有常数参数的独立次高斯随机变量，应用霍夫丁型不等式，并对 $1 \le k \le K$ 取并集界，可得

$$\mathbb{P}\left(\max_{1\le k\le K}\left|\sum_{i=1}^n\widetilde{\epsilon}_{ik}\right| > t\right) \le 2Ke^{-\frac{Ct^2}{n}},$$

这进一步意味着

$$\begin{aligned}
\mathbb{E}\max_{1\le k\le K}\left|\sum_{i=1}^n\widetilde{\epsilon}_{ik}\right| =& \int_0^\infty\mathbb{P}\left(\max_{1\le k\le K}\left|\sum_{i=1}^n\widetilde{\epsilon}_{ik}\right| > t\right)\mathrm{d}t\\
\le&\sqrt{\frac{\log 2K}{C}}n + \int_{\sqrt{\frac{\log 2K}{C}}n}^\infty e^{-\frac{C}{n}(t-\sqrt{\frac{\log 2K}{C}}n)^2}\mathrm{d}t\\
\le&C\sqrt{\log(2K)n}.
\end{aligned} \tag{89}$$

结合式（86）、（87）、（88）和（89），可得对于 $j = 1,\ 2,$

$$\mathbb{E}g_t^{(j)}((x_1,\epsilon_1),\ldots,(x_n,\epsilon_n)) \le CL.$$

此外，可以证明，对于 $1 \le i \le n$ 以及任意 $\Delta \in S(R_0,\ r_0)$,

$$
\begin{aligned}
&\left|\epsilon_i^\top(\nabla f^{\mathrm{ON}}(u_i(\theta^*+\Delta)) - \nabla f^{\mathrm{ON}}(u_i(\theta^*)))\begin{pmatrix} x_i, 0 \\ -I_K \end{pmatrix}^\top \Delta\right| \\
\le& \|x_i\epsilon_i^\top(\nabla f^{\mathrm{ON}}(u_i(\theta^*+\Delta)) - \nabla f^{\mathrm{ON}}(u_i(\theta^*)))_{:,1}\|_{\omega^{-1},2,\infty}\|\Delta_{1:p}\|_{\omega,2,1} \\
&+ \|(\nabla f^{\mathrm{ON}}(u_i(\theta^*+\Delta)) - \nabla f^{\mathrm{ON}}(u_i(\theta^*)))^\top\epsilon_i\|_2\|\Delta\|_2 \\
\le& \|(\nabla f^{\mathrm{ON}}(u_i(\theta^*+\Delta)) - \nabla f^{\mathrm{ON}}(u_i(\theta^*)))^\top\epsilon_i\|_\infty \\
&\cdot (\max_j \omega_j^{-1}\|(x_i)_{\mathcal{G}_j}\|_2\|\Delta_{1:p}\|_{\omega,2,1} + \sqrt{K}\|\Delta\|_2) \\
\le& Ch^{\mathrm{ON}}(R_0,r_0)(C_x\sqrt{m}\|\Delta_{1:p}\|_{\omega,2,1} + \sqrt{K}\|\Delta\|_2),
\end{aligned}
$$

其中，最后一行源于假设 3.1、3.2 以及以下事实:

$$
\begin{aligned}
\|\nabla(f^{\mathrm{ON}}(u_i(\theta^*))\epsilon_i\|_\infty &\le \frac{1}{2}h^{\mathrm{ON}}(0,0) \le \frac{1}{2}h^{\mathrm{ON}}(R_0,r_0), \\
\sup_{\Delta\in\mathcal{S}(R_0,r_0)} \|\nabla(f^{\mathrm{ON}}(u_i(\theta^*+\Delta)))\epsilon_i\|_\infty &\le \frac{1}{2}h^{\mathrm{ON}}(R_0,r_0),
\end{aligned} \tag{90}
$$

这由引理 C.2 和 $\|\epsilon_i\|_1 \le 2$ 推出。因此,

$$
\begin{aligned}
&\sup_{\Delta\in\mathcal{S}(R_0,r_0)} \frac{\left|\epsilon_i^\top(\nabla f^{\mathrm{ON}}(u_i(\theta^*+\Delta)) - \nabla f^{\mathrm{ON}}(u_i(\theta^*)))\begin{pmatrix} x_i, 0 \\ -I_K \end{pmatrix}\Delta\right|}{n\tau(\Delta)} \\
\le& Ch^{\mathrm{ON}}(R_0,r_0)\left\{C_x\sqrt{\frac{m}{\lambda_{\max}(\Sigma)(\log J+m)n}} + \sqrt{\frac{1}{(\log J+\log(2K))n}}\right\} \\
\le& Ch^{\mathrm{ON}}(R_0,r_0)\left(C_x\sqrt{\frac{m}{\lambda_{\max}(\Sigma)}}+1\right)\frac{1}{\sqrt{n\log J}}.
\end{aligned} \tag{91}
$$

以下有界差异结果可直接由（91）推出:

$$
\begin{aligned}
&\Big|g_t^{(j)}((x_1,\epsilon_1),\dots,(x_i,\epsilon_i),\dots,(x_n,\epsilon_n)) \\
&- g_t^{(j)}((x_1,\epsilon_1),\dots,(x_i',\epsilon_i'),\dots,(x_n,\epsilon_n))\Big| \\
\le& Ch^{\mathrm{ON}}(R_0,r_0)\left(C_x\sqrt{\frac{m}{\lambda_{\max}(\Sigma)}}+1\right)\frac{1}{\sqrt{n\log J}},
\end{aligned}
$$

因此，应用有界差分不等式 [McDiarmid, 1989] upon$g_t^{(j)}$(((x_1, ϵ_1).....(x_n, ϵ_n)) 将得到

$$
\sup_{\substack{\frac{t}{2}\le\|\Delta\|_{\omega,2,1}\le t \\ \Delta\in\mathcal{S}(R_0,r_0)}} |\phi(\Delta,\{x_i\}_{i=1}^n,\{\epsilon_i\}_{i=1}^n)| \le CL, \tag{92}
$$

概率至少为

$$
\begin{aligned}
&1 - 2\exp\left\{-\frac{cL^2\log J}{(h^{\mathrm{ON}}(R_0,r_0))^2(C_x^2\lambda_{\max}^{-1}(\Sigma)m+1)}\right\} \\
\ge& 1 - 2\exp\{-\frac{cK\log J}{C_x^2 m+1}\},
\end{aligned}
$$

其中我们应用了这一事实： $L^2 \geq K(h^{\mathrm{ON}}(R_0, r_0))^2$ 且 $\lambda_{\max}(\Sigma) \geq c_\circ$ 现在，我们应用剥离论证，将上述上界扩展到 $\Delta \in \mathcal{S}(R_0, r_0)$。首先注意，对于任意 $\Delta \in \mathcal{S}(R_0, r_0)$,

$$\begin{aligned}\tau(\Delta) \leq & \left(\max_j \omega_j \sqrt{\frac{\lambda_{\max}(\Sigma)(\log J + m)}{n}} + 2\sqrt{\frac{K(\log J + \log(2K))}{n}} \right) R_0 \\ =: & C_n^{(1)}\end{aligned}$$

定义

$$C_n^{(2)} = \frac{CL \log J \min\{C_x^{-2}, 1\}}{(\frac{4}{3} h^{\mathrm{ON}}(R_0, r_0) + \frac{1}{4}(h^{\mathrm{ON}}(R_0, r_0))^2) mn}$$

$N = \log_2(\frac{C_n^{(1)}}{C_n^{(2)}})$，则可以证明

$$\begin{aligned}\mathcal{S}(R_0, r_0) \subset & \cup_{k=1}^N \{\Delta : 2^{k-1} C_n^{(2)} \leq \|\Delta\|_{\omega,2,1} \leq 2^k C_n^{(2)}\} \\ & \cup \{\Delta : 0 \leq \|\Delta\|_{\omega,2,1} \leq C_n^{(2)}\}.\end{aligned}$$

类似于引理 A.2 的证明，我们考虑函数

$$\widetilde{\phi}(r, d, \{x_i\}_{i=1}^n, \{\epsilon_i\}_{i=1}^n) = \phi(rd, \{x_i\}_{i=1}^n, \{\epsilon_i\}_{i=1}^n)$$

对于任意满足 $\tau(d) = 1$ 的 $r \geq 0$ 和 $d \in \mathbb{R}^{p+K}$。经过一些计算可得

$$\begin{aligned}& \left| \frac{\partial \widetilde{\phi}(r, d, \{x_i\}_{i=1}^n, \{\epsilon_i\}_{i=1}^n)}{\partial r} \right| \\ = & \left| \frac{1}{n} \sum_{i=1}^n \left\langle \nabla^2 f^{\mathrm{ON}}(u_i(\theta^* + rd)), \epsilon_i \otimes \left(\begin{pmatrix} x_i, 0 \\ -I_K \end{pmatrix}^\top d \right) \otimes \left(\begin{pmatrix} x_i, 0 \\ -I_K \end{pmatrix}^\top d \right) \right\rangle \right| \\ \leq & \max_i \|\nabla^2 f^{\mathrm{ON}}(u_i(\theta^* + rd))\|_\infty \|\epsilon_i\|_1 \left\| \begin{pmatrix} x_i, 0 \\ -I_K \end{pmatrix}^\top d \right\|_1^2 \\ \leq & 2 \max\{C_x^2, 1\} \|d\|_1^2 \max_i \|\nabla^2 f^{\mathrm{ON}}(u_i(\theta^* + rd))\|_\infty,\end{aligned}$$

其中，最后一行是因为 $\|\epsilon_i\|_1 \leq \|\delta_i\|_1 + \|\mathbb{E}(\delta_i | x_i)\|_1 \leq 2$，并且

$$\left\| \begin{pmatrix} x_i, 0 \\ -I_K \end{pmatrix}^\top d \right\|_1 \leq |x_i^\top d_{1:p}| + \|d_{(p+1):(p+K)}\|_1 \leq \max\{C_x, 1\} \|d\|_1.$$

注意

$$\begin{aligned}\|d\|_1 \leq & \max_j \omega_j^{-1} \sqrt{m} \|d_{1:p}\|_{\omega,2,1} + \sqrt{K} \|d_{(p+1):(p+K)}\|_2 \\ \leq & \max_j \omega_j^{-1} \sqrt{\frac{mn}{\lambda_{\max}(\Sigma)(\log J + m)}} + \sqrt{\frac{n}{\log J + \log(2K)}} \\ \leq & C \sqrt{\frac{mn}{\log J}},\end{aligned}$$

根据引理 C.2，我们有

$$\begin{aligned}& \left| \frac{\partial \widetilde{\phi}(r, d, \{x_i\}_{i=1}^n, \{\epsilon_i\}_{i=1}^n)}{\partial r} \right| \\ \leq & C \max\{C_x^2, 1\} (\frac{4}{3} h^{\mathrm{ON}}(R_0, r_0) + \frac{1}{4}(h^{\mathrm{ON}}(R_0, r_0))^2 \frac{mn}{\log J},\end{aligned}$$

对于任意 $rd \in S(R_0,\ r_0)$。因此,

$$\begin{aligned}
&\sup_{\substack{0\le\|\Delta\|_{\omega,2,1}\le C_n^{(2)}\\ \Delta\in\mathcal{S}(R_0,r_0)}} |\phi(\Delta,\{x_i\}_{i=1}^n,\{\epsilon_i\}_{i=1}^n)|\\
\le &\sup_{\substack{\|\Delta\|_{\omega,2,1}=C_n^{(2)}\\ \Delta\in\mathcal{S}(R_0,r_0)}} |\phi(\Delta,\{x_i\}_{i=1}^n,\{\epsilon_i\}_{i=1}^n)|\\
&+C\max\{C_x^2,1\}(\frac{4}{3}h^{\mathrm{ON}}(R_0,r_0)+\frac{1}{4}(h^{\mathrm{ON}}(R_0,r_0))^2\frac{mn}{\log J}C_n^{(2)}\\
\le &\sup_{\substack{\|\Delta\|_{\omega,2,1}=C_n^{(2)}\\ \Delta\in\mathcal{S}(R_0,r_0)}} |\phi(\Delta,\{x_i\}_{i=1}^n,\{\epsilon_i\}_{i=1}^n)|+CL.
\end{aligned}$$

现在我们对 $t = C_n^{(2)},\ 2C_n^{(2)},\ .....,\ 2^{N-1}C_n^{(2)}$应用概率界 (92), 并取并集界, 得到如下结果:

$$\sup_{\Delta\in\mathcal{S}(R_0,r_0)} |\phi(\Delta,\{x_i\}_{i=1}^n,\{\epsilon_i\}_{i=1}^n| \le CL,$$

其概率至少为 $1-2N\exp\{-\frac{cK}{\log_{C_x^2m+1}J}\}$。现在注意到

$$\begin{aligned}
N=&\log_2\left\{\left[\max_j\omega_j m\sqrt{\frac{\lambda_{\max}(\Sigma)(\log J+m)n}{(\log J)^2}}+m\sqrt{\frac{K(\log J+\log(2K))n}{(\log J)^2}}\right]\right.\\
&\left.\cdot\frac{R_0(\frac{4}{3}h^{\mathrm{ON}}(R_0,r_0)+\frac{1}{4}(h^{\mathrm{ON}}(R_0,r_0))^2)}{C\min\{C_x^{-1},1\}L}\right\}\\
\le&\log_2\left\{Cm\sqrt{\frac{n}{\log J}}(\max_j\omega_j\lambda_{\max}^{\frac{1}{2}}(\Sigma)+\sqrt{K})\frac{R_0(\frac{4}{3}h^{\mathrm{ON}}(R_0,r_0)+\frac{1}{4}(h^{\mathrm{ON}}(R_0,r_0))^2)}{C\min\{C_x^{-1},1\}L}\right\}.
\end{aligned}$$

下 述引理表明, R_0 还具有可由假设 3.4 推出的另一项性质。

引理 B.5。 假设 $R_0>0$ 满足假设 *3.4*, 则

$$\begin{aligned}
R_0\le&\min_j\frac{n_j}{\pi_j n_u}\frac{\left(1+e^{(C_x+1)(R^*+R_0)}\right)^{-6}}{512(C_x+1)K^3(1+\frac{2}{r^*-r_0})^3(2+\frac{2}{r^*-r_0})}\\
\le&\min_j\frac{n_j}{\pi_j n_u}\frac{\left(1+e^{(C_x+1)(R^*+R_0)}\right)^{-2}}{512(C_x+1)K^3(h^{\mathrm{ON}}(R_0,r_0))^3(1+h^{\mathrm{ON}}(R_0,r_0))}.
\end{aligned}$$

引理 B.5 意味着

$$\begin{aligned}
&R_0(\frac{4}{3}h^{\mathrm{ON}}(R_0,r_0)+\frac{1}{4}(h^{\mathrm{ON}}(R_0,r_0))^2)\\
\le&\min_j\frac{n_j}{\pi_j n_u}\frac{\left(1+e^{(C_x+1)(R^*+R_0)}\right)^{-2}}{372(C_x+1)K^3(h^{\mathrm{ON}}(R_0,r_0))^2}\\
\le&\min_j\frac{n_j}{\pi_j n_u}\frac{3e^{2(C_x+1)(R^*+C)}(1+e^{(C_x+1)(R^*+C)})^{-4}}{512(C_x+1)(h^{\mathrm{ON}}(C,r_0))^2}\\
\le&\frac{C}{C_x+1}.
\end{aligned}$$

上述不等式结合 $L\ge1$ 这一事实可得 $\frac{R_0}{(\frac{4}{3}h^{\mathrm{ON}}(R_0,\ r_0)+\frac{1}{4}(h^{\mathrm{ON}}(R_0,\ r_0))^2)C\min\{C_x^{-1},1\}L^-}\le C$。同时, 由于 $m\le Cn$、$\max_j\omega_j\le C\sqrt{\frac{\log K}{n}}\le Cn$, 且 $\lambda_{\max}(\Sigma)\le C\sigma^2\le C$, 显然有 $N\le C\log_2 n$。因此,

$$|\mathrm{II}(\Delta)|\le CL\tau(\Delta)$$

以至少.....的概率，对所有 $\Delta \in S(R_0,\ r_0)$ 均成立

$$\begin{aligned}1 - C\log_2 n \exp\{-\frac{cK\log J}{C_x^2 m+1}\} \geq & 1-\exp\{\log(C\log n) - \frac{cK\log J}{C_x^2 m+1}\}\\ \geq & 1-\exp\{-\frac{cK\log J}{C_x^2 m+1}\}.\end{aligned}$$

□

C 定理 3.1 和定理 3.2 的证明所需引理

C.1 多项式 PU 模型的辅助引理

Proof of 引理 *4.1.* 回顾 $A(\cdot)$ 的定义，可以证明对于任意 $\eta \in \mathbb{R}^K$,

$$(\nabla A(\eta))_i = \frac{e^{\eta_i}}{1+\sum_{k=1}^K e^{\eta_k}}, 1 \leq i \leq K, \tag{93}$$

and

$$\begin{aligned}(\nabla^2 A(\eta))_{jk} =& \frac{e^{\eta_j}}{1+\sum_{l=1}^K e^{\eta_l}}\left(\mathbb{1}_{\{j=k\}} - \frac{e^{\eta_k}}{1+\sum_{l=1}^K e^{\eta_l}}\right)\\ =& (\nabla A(\eta))_j\left(\mathbb{1}_{\{j=k\}} - (\nabla A(\eta))_k\right).\end{aligned} \tag{94}$$

因此， $\nabla^2 A(\eta)$ 的最小特征值具有如下下界:

$$\begin{aligned}\lambda_{\min}(\nabla^2 A(\eta)) =& \inf_{\|v\|_2=1} v^\top \nabla^2 A(\eta) v\\ =& \inf_{\|v\|_2=1} \sum_{j=1}^K (\nabla A(\eta))_j v_j^2 - \left(\sum_{j=1}^K (\nabla A(\eta))_j v_j\right)^2\\ \geq& \inf_{\|v\|_2=1} \sum_{j=1}^K (\nabla A(\eta))_j v_j^2 - \sum_{j=1}^K (\nabla A(\eta))_j v_j^2 \sum_{j=1}^K (\nabla A(\eta))_j\\ \geq& \min_j (\nabla A(\eta))_j \left(1 - \sum_{j=1}^K (\nabla A(\eta))_j\right)\\ =& \frac{\min_j e^{\eta_j}}{\left(1+\sum_{l=1}^K e^{\eta_l}\right)^2}.\end{aligned} \tag{95}$$

回顾一下， $R^* = \max_{1\leq i\leq K} \|(\Theta^*)_{:,\ i}\|_1$ 和 $\Theta \in \Theta^* + \mathbb{B}_{1,\ \infty}(R)$。于是我们有

$$\exp\{[f(\Theta^\top x_i)]_k\} = \frac{\frac{n_k}{\pi_k n_u}\exp\{\Theta_{:,k}^\top x_i\}}{1+\sum_{j=1}^K \exp\{\Theta_{:,j}^\top x_i\}} \geq \frac{\min_j \frac{n_j}{\pi_j n_u} e^{-C_x(R+R^*)}}{1+Ke^{C_x(R+R^*)}},$$

$$\sum_{j=1}^K \exp\{[f(\Theta^\top x_i)]_j\} \leq \max_j \frac{n_j}{\pi_j n_u}.$$

因此我们有

$$\begin{aligned}\lambda_{\min}(\nabla^2 A(f(\Theta^\top x_i)) \geq & \frac{\exp\{\min_j (f(\Theta^\top x_i))_j\}}{\left[1+\sum_{j=1}^K \exp\{(f(\Theta^\top x_i))_j\}\right]^2} \\ \geq & \frac{e^{-C_x(R+R^*)} \min_j \frac{n_j}{\pi_j n_u}}{(1+\max_j \frac{n_j}{\pi_j n_u})^2 (1+Ke^{C_x(R+R^*)})}\end{aligned} \tag{96}$$

现在我们给出的最小特征值下界为

$$\nabla f(v)^\top \nabla f(v) = (I_K - 1_K \nabla A(v)^\top)^\top (I_K - 1_K \nabla A(v)^\top).$$

对于任意向量 $z \in \mathbb{R}^K$，令 $y = \nabla f(v) z = z - (\nabla A(v)^\top z) 1_K$。则可以证明

$$\nabla A(v)^\top y = \nabla A(v)^\top z (1 - \nabla A(v)^\top 1_K) = (1+\sum_{j=1}^K e^{v_j})^{-1} \nabla A(v)^\top z,$$

这意味着 $y = z - (1+\sum_{j=1}^K e^{v_j}) \nabla A(v)^\top y 1_K$，因此

$$\|z\|_2 \leq \|y\|_2 + \sqrt{K} \left(1+\sum_{j=1}^K e^{v_j}\right) \|\nabla A(v)\|_2 \|y\|_2. \tag{97}$$

因此，

$$\begin{aligned}\lambda_{\min}(\nabla f(v)^\top \nabla f(v)) = & \inf_{z \in \mathbb{R}^K} \frac{\|\nabla f(v) z\|_2^2}{\|z\|_2^2} \\ \geq & \left[1+\sqrt{K}\left(1+\sum_{j=1}^K e^{v_j}\right) \|\nabla A(v)\|_2\right]^{-2} \\ = & \left[1+\sqrt{K}\left(\sum_{j=1}^K e^{2v_j}\right)^{\frac{1}{2}}\right]^{-2},\end{aligned} \tag{98}$$

and

$$\lambda_{\min}(\nabla f(\Theta^\top x_i)^\top \nabla f(\Theta^\top x_i)) \geq \left[1+Ke^{C_x(R+R^*)}\right]^{-2}. \tag{99}$$

结合（96）和（99）即完成证明。 □

引理 *A.4* 的证明。根据泰勒定理，存在 $t \in [0,1]$ 使得

$$\begin{aligned}\mathrm{I}(\Delta) = & \frac{1}{n}\sum_{i=1}^n (x_i^\top \Delta) \nabla f(\Theta^\top x_i)^\top \left[\nabla A(f(\Theta^\top x_i)) - \nabla A(f(\Theta^{*\top} x_i))\right] \\ = & \frac{1}{n}\sum_{i=1}^n (\Delta^\top x_i)^\top \nabla f(\Theta^\top x_i)^\top \nabla^2 A(f(\Theta_t^\top x_i)) \nabla f(\Theta_t^\top x_i) (\Delta^\top x_i) \\ \geq & \frac{1}{n}\sum_{i=1}^n (\Delta^\top x_i)^\top \nabla f(\Theta_t^\top x_i)^\top \nabla^2 A(f(\Theta_t^\top x_i)) \nabla f(\Theta_t^\top x_i) (\Delta^\top x_i) \\ & - \frac{1}{n}\sum_{i=1}^n \|\Delta^\top x_i\|_2^2 \left\|\nabla f(\Theta_t^\top x_i)^\top \nabla^2 A(f(\Theta_t^\top x_i)) \left[\nabla f(\Theta_t^\top x_i) - \nabla f(\Theta^\top x_i)\right]\right\|,\end{aligned} \tag{100}$$

其中 $\Theta_t = t\Theta + (1-t)\Theta^*$ 位于 Θ 和 Θ^*之间。根据引理 4.1，我们有

$$\begin{aligned}&(\Delta^\top x_i)^\top \nabla f(\Theta_t^\top x_i)^\top \nabla^2 A(f(\Theta_t^\top x_i))\nabla f(\Theta_t^\top x_i)(\Delta^\top x_i)\\ \geq&(h^{\mathrm{MN}}(R) + 4C_x R)\|\Delta^\top x_i\|_2^2.\end{aligned} \tag{101}$$

另一方面，由于 $\nabla f(v) = I_K - 1_K \nabla A(v)^\top$，可以证明

$$\begin{aligned}&\left\|\nabla f(\Theta_t^\top x_i)^\top \nabla^2 A(f(\Theta_t^\top x_i))\left[\nabla f(\Theta_t^\top x_i) - \nabla f(\Theta^\top x_i)\right]\right\|\\ \leq&\left\|\nabla f(\Theta_t^\top x_i)^\top \nabla^2 A(f(\Theta_t^\top x_i))1_K\right\|_2 \left\|\nabla A(\Theta_t^\top x_i) - \nabla A(\Theta^\top x_i)\right\|\\ \leq&2\left\|\nabla A(\Theta_t^\top x_i) - \nabla A(\Theta^\top x_i)\right\|\\ \leq&4C_x R,\end{aligned} \tag{102}$$

其中第三行是由以下事实得出的：

$$\begin{aligned}&\left\|\nabla f(\Theta_t^\top x_i)^\top \nabla^2 A(f(\Theta_t^\top x_i))1_K\right\|_2\\ \leq&\|\nabla^2 A(f(\Theta_t^\top x_i))1_K\|_2 + 1_K^\top \nabla^2 A(f(\Theta_t^\top x_i))1_K\|\nabla A(\Theta_t^\top x_i)\|_2\\ \leq&1 + \|\nabla A(\Theta_t^\top x_i)\|_2\\ \leq&2,\end{aligned} \tag{103}$$

最后一行是由以下事实得出的：

$$\begin{aligned}&\left\|\nabla A(\Theta_t^\top x_i) - \nabla A(\Theta^\top x_i)\right\|_2\\ \leq&\sup_{\eta\in\mathbb{R}^K}\left[\sum_{i=1}^K \|(\nabla^2 A(\eta))_{i,:}\|_1^2\right]^{\frac{1}{2}} \|(\Theta_t - \Theta)^\top x_i\|_\infty\\ \leq&2\|(\Theta_t - \Theta)^\top x_i\|_\infty\\ \leq&2C_x R.\end{aligned} \tag{104}$$

因此，

$$\mathrm{I}(\Delta) \geq \frac{h^{\mathrm{MN}}(R_0)}{n}\sum_{i=1}^n \|\Delta^\top x_i\|_2^2, \tag{105}$$

这里用到了这样一个事实： $R \leq R_0$ 和 $h^{\mathrm{MN}}(\cdot)$ 是递减函数。 □

引理 *A.5* 的证明。令

$$A_1 = \mathbb{B}_{\omega,2,1}(\sqrt{\rho}) \cap \mathbb{B}_2(1) \tag{106}$$

and

$$A_2 = \left(1 + \frac{2}{\min_j \omega_j}\right)\mathrm{cl}\{\mathrm{conv}\{\mathbb{B}_{\mathcal{G},0}(\lfloor\rho\rfloor) \cap \mathbb{B}_2(1)\}\}. \tag{107}$$

由于 A_1和 A_2都是凸集，只需证明对于任意$U \in \mathbb{R}_{p\times K}$，支撑函数 $\phi_{A_1}(U) \leq \phi_{A_2}(U)$，其中 $\phi_{A_1}(U) = \sup_{V\in A_1}\langle V, U\rangle$，且 $\phi_{A_2}(U) = \sup_{V\in A_2}\langle V, U\rangle$。令 $\mathcal{G}_{\lfloor\rho\rfloor}$为满足$\|U_{\mathcal{R}_j\cdot} c_j\|_2$ 最大条件的 $\lfloor\rho\rfloor$ $\mathcal{G}_j$ 的并集。然后

$$\begin{aligned}\phi_{A_1}(U) \leq&\|U_{\mathcal{G}_{\lfloor\rho\rfloor}}\|_2 + \sqrt{\rho}\|U_{\mathcal{G}^c_{\lfloor\rho\rfloor}}\|_{\omega^{-1},2,\infty}\\ \leq&\|U_{\mathcal{G}_{\lfloor\rho\rfloor}}\|_2 + \frac{\sqrt{\rho}}{\sqrt{\lfloor\rho\rfloor}\min_j \omega_j}\|U_{\mathcal{G}_{\lfloor\rho\rfloor}}\|_2\\ \leq&\left(1 + \frac{2}{\min_j \omega_j}\right)\|U_{\mathcal{G}_{\lfloor\rho\rfloor}}\|_2.\end{aligned} \tag{108}$$

另一方面，也可以证明

$$\begin{aligned}\phi_{A_2}(U) &= \left(1+\frac{2}{\min_j \omega_j}\right) \sup_{V\in\mathbb{B}_{\mathcal{G},0}(\lfloor\rho\rfloor)\cap\mathbb{B}_2(1)} \langle V, U\rangle \\ &= \left(1+\frac{2}{\min_j \omega_j}\right) \|U_{\mathcal{G}_{\lfloor\rho\rfloor}}\|_2.\end{aligned} \tag{109}$$

因此，我们有 $A_1 \subset A_2$。 □

引理 *A.6* 的证明。 根据假设 3.1， $\{(x_i\Delta_{:,\ j}\}_{i=1}^n$ 是具有参数 $\sigma\|\Delta_{:,\ j}\|_2$的独立次高斯随机变量，这意味着 $\|(x_i^\top\Delta_{:,\ j})^2\|_{\psi_1} \le C\sigma^2\|\Delta_{:,\ j}\|_2^2$。 因此

$$\mathbb{P}\left(\left|\Delta_{:,j}^\top\left(\frac{1}{n}X^\top X-\Sigma\right)\Delta_{:,j}\right| \ge t\|\Delta_{:,j}\|_2^2\right) \le 2\exp\left\{-cn\min\left\{\frac{t^2}{\sigma^4},\frac{t}{\sigma^2}\right\}\right\}. \tag{110}$$

由于 $g(\Delta,\ \Delta) = \sum_{j=1}^K \Delta_{:,\ j}^\top\left(\frac{1}{n}X^\top X-\Sigma\right)\Delta_{:,\ j}$， (110) 意味着

$$\mathbb{P}\left(|g(\Delta,\Delta)| \ge t\right) \le 2K\exp\left\{-cn\min\left\{\frac{t^2}{\sigma^4},\frac{t}{\sigma^2}\right\}\right\}. \tag{111}$$

□

~引理 *A.7* 的证明。 设 $\{U_i\}_{i=1}^n$ 是 $\{U_i\}_{i=1}^n$的独立副本，且 $\{\varepsilon_i\}_{i=1}^n$ 是独立的 Rademacher 随机变量。于是可以证明

$$\begin{aligned}&\mathbb{E}_U\left(\sup_{\gamma\in\Gamma}\sum_{i=1}^n(\gamma(U_i)-\mathbb{E}(\gamma(U_i)))\right)\\ =&\mathbb{E}_U\left(\sup_{\gamma\in\Gamma}\sum_{i=1}^n(\gamma(U_i)-\mathbb{E}_{\widetilde{U}}(\gamma(\widetilde{U}_i)))\right)\\ =&\mathbb{E}_U\left(\sup_{\gamma\in\Gamma}\mathbb{E}_{\widetilde{U}}\left(\sum_{i=1}^n(\gamma(U_i)-\gamma(\widetilde{U}_i))|U\right)\right)\\ \le&\mathbb{E}_{U,\widetilde{U}}\left(\sup_{\gamma\in\Gamma}\sum_{i=1}^n(\gamma(U_i)-\gamma(\widetilde{U}_i))\right)\\ =&\mathbb{E}_\varepsilon\mathbb{E}_{U,\widetilde{U}}\left(\sup_{\gamma\in\Gamma}\sum_{i=1}^n\varepsilon_i(\gamma(U_i)-\gamma(\widetilde{U}_i))\right),\end{aligned}$$

其中第 4 行是由于

$$\sum_{i=1}^n(\gamma(U_i)-\gamma(\widetilde{U}_i)) \le \sup_{\gamma\in\Gamma}\sum_{i=1}^n(\gamma(U_i)-\gamma(\widetilde{U}_i))$$

~ ~对于任意 $\gamma\in\Gamma$，最后一行是由于 U_i， U_i， $i=1$、.....、n 彼此独立，且 U_i和 U_i同分布。此外，我们有

$$\begin{aligned}&\mathbb{E}_\varepsilon\mathbb{E}_{U,\widetilde{U}}\left(\sup_{\gamma\in\Gamma}\sum_{i=1}^n\varepsilon_i(\gamma(U_i)-\gamma(\widetilde{U}_i))\right)\\ \le&\mathbb{E}\left(\sup_{\gamma\in\Gamma}\sum_{i=1}^n\varepsilon_i\gamma(U_i)\right)+\mathbb{E}\left(\sup_{\gamma\in\Gamma}\sum_{i=1}^n(-\varepsilon_i)\gamma(\widetilde{U}_i)\right)\\ =&2\mathbb{E}\left(\sup_{\gamma\in\Gamma}\sum_{i=1}^n\varepsilon_i\gamma(U_i)\right),\end{aligned}$$

其中我们利用了 ε_i 的对称性。 □

引理 *A.8* 的证明。首先注意到

$$\begin{aligned}\nabla h_i(u) = &\left[\nabla A(\Theta^{*\top} x_i) - \nabla A(\Theta^{*\top} x_i + u)\right] 1_K^\top \epsilon_i \\ &- \nabla^2 A(\Theta^{*\top} x_i + u)^\top u 1_K^\top \epsilon_i,\end{aligned} \tag{112}$$

以及 $|1_K^\top \epsilon_i| = |1_K^\top \delta_i - 1_K^\top \mathbb{E}\delta_i| \le 1$。回顾 $A(u) = \log(1 + \sum_{k=1}^K e^{u_i})$ 的定义，可以证明 $\|\nabla A(u)\|_2 \le 1$。如引理 A.4 的证明所示，$\nabla^2 A(u)$ 是正定的，且每个元素

$$(\nabla^2 A(u))_{jk} = \frac{e^{u_j}}{1 + \sum_{l=1}^K e^{u_l}} \left(\mathbb{1}_{\{j=k\}} - \frac{e^{u_k}}{1 + \sum_{l=1}^K e^{u_l}} \right).$$

Thus

$$\begin{aligned}\lambda_{\max}(\nabla^2 A(u)) = &\sup_{\|v\|_2=1} v^\top \nabla^2 A(u) v \\ \le &\sup_{\|v\|_2=1} \sum_{j=1}^K \frac{v_j^2 e^{u_j}}{1 + \sum_{l=1}^K e^{u_l}} \\ \le &1,\end{aligned}$$

这意味着，对于任意 $u,\ v \in \mathbb{R}^K$，$\|\nabla^2 A(u) v\|_2 \le \|v\|_2$。同时，对于任意 $u \in \mathbb{R}^K$，

$$\begin{aligned}\|\nabla^2 A(u) u\|_2^2 = &\sum_{i=1}^K \left(\frac{u_i e^{u_i}}{1 + \sum_{k=1}^K e^{u_k}} - \frac{\sum_{j=1}^K u_j e^{u_i + u_j}}{(1 + \sum_{k=1}^K e^{u_k})^2} \right)^2 \\ = &\frac{\sum_{i=1}^K \left(u_i e^{u_i} + \sum_{j=1}^K (u_i - u_j) e^{u_i + u_j} \right)^2}{(1 + \sum_{k=1}^K e^{u_k})^4} \\ \le &\frac{K^2 \sum_{i=1}^K e^{4u_i}}{(1 + \sum_{k=1}^K e^{u_k})^4} \\ \le &K^2,\end{aligned}$$

其中在第三行用到了如下事实：对于任意 $x \in \mathbb{R}$，$x \le e^x$。因此，

$$\begin{aligned}\|\nabla h_i(u)\|_2 \le &2 + \|\nabla^2 A(\Theta^{*\top} x_i + u)^\top u\|_2 \\ \le &2 + \|\nabla^2 A(\Theta^{*\top} x_i + u)^\top \Theta^* x_i\|_2 \\ &+ \|\nabla^2 A(\Theta^{*\top} x_i + u)^\top (\Theta^* x_i + u)\|_2 \\ \le &2 + \|\Theta^* x_i\|_2 + K \\ \le &K + \sqrt{K} R^* C_x + 2,\end{aligned} \tag{113}$$

其中在最后一行用到了如下事实：$\|\Theta^*\|_{1,\ \infty} \le R^*$ 以及 $\|x_i\|_\infty \le C_x$。 □

引理 *A.10* 的证明。令 $E \in \mathbb{R}^{n \times K}$ 使得 $E_{ij} = \widetilde{\varepsilon}_{ij}$，则

$$\mathbb{E} \left\| \sum_{i=1}^n x_i \widetilde{\varepsilon}_i^\top \right\|_{\omega^{-1},2,\infty} = \mathbb{E} \max_{1 \le j \le J} \frac{1}{\omega_j} \|X_{\mathcal{R}_j}^\top E_{\mathcal{C}_j}\|_2.$$

首先注意到，利用引理 A.1 证明中的类似论证，可以得出

$$\mathbb{E}\left[\|X_{\mathcal{R}_j}^\top E_{\mathcal{C}_j}\|_2 | X \right] \le \left(c_j \mathrm{tr}(X_{\mathcal{R}_j} X_{\mathcal{R}_j}^\top) \right)^{\frac{1}{2}} \le \sqrt{r_j c_j} \|X_{\mathcal{R}_j}\|. \tag{114}$$

与此同时，应用 Talagrand 压缩不等式（参见 Vershynin [2018] 中的定理 5.2.16）可得

$$\mathbb{P}\left(\|X_{\mathcal{R}_j}^\top E_{\mathcal{C}_j}\|_2 > C\|X_{\mathcal{R}_j}\|(\sqrt{r_j c_j}+t_j)|X\right) \le \exp\{-ct_j^2\}. \tag{115}$$

为了在不对 X 进行条件化的情况下得到尾概率界，我们将具有独立各向同性次高斯行的矩阵[1]的谱范数界应用于 $X_{\mathcal{R}_j}\Sigma_{\mathcal{R}_j^2,\ \mathcal{R}_j}$，于是对于任意 $t_j>0$,

$$\mathbb{P}\left(\|X_{\mathcal{R}_j}\| > \lambda_{\max}^{\frac{1}{2}}(\Sigma)(C\sqrt{n}+t_j)\right) \le 2\exp\{-ct_j^2\}. \tag{116}$$

结合 (115) 和 (116)，可以证明

$$\begin{aligned}&\mathbb{P}\left(\|X_{\mathcal{R}_j}^\top E_{\mathcal{C}_j}\|_2 > C\lambda_{\max}^{\frac{1}{2}}(\Sigma)(\sqrt{r_j c_j n}+(\sqrt{n}+\sqrt{r_j c_j})t_j+t_j^2)\right)\\ \le& 3\exp\{-nt_j^2\},\end{aligned} \tag{117}$$

这等价于

$$\begin{aligned}&\mathbb{P}\left(\|X_{\mathcal{R}_j}^\top E_{\mathcal{C}_j}\|_2 > C\lambda_{\max}^{\frac{1}{2}}(\Sigma)\sqrt{r_j c_j n}+t_j\right)\\ \le& 3\exp\left\{-\min\left\{\frac{t_j^2}{\lambda_{\max}(\Sigma)(n+r_j c_j)}, \frac{t_j}{\lambda_{\max}^{\frac{1}{2}}(\Sigma)}\right\}\right\}.\end{aligned} \tag{118}$$

因此，我们可以将 $\mathbb{E}\max_{1\le j\le J}\frac{1}{\omega_j}\|X_{\mathcal{R}_j}^\top E_{\mathcal{C}_j}\|_2$ 界定如下：

$$\begin{aligned}&\mathbb{E}\max_{1\le j\le J}\frac{1}{\omega_j}\|X_{\mathcal{R}_j}^\top E_{\mathcal{C}_j}\|_2\\ \le&\int_0^\infty \mathbb{P}\left(\max_{1\le j\le J}\frac{1}{\omega_j}\|X_{\mathcal{R}_j}^\top E_{\mathcal{C}_j}\|_2 > t\right)\mathrm{d}t\\ \le& C\lambda_{\max}^{\frac{1}{2}}(\Sigma)\sqrt{mn}+C\lambda_{\max}^{\frac{1}{2}}(\Sigma)\sqrt{\log J(m+n)}\\ &+\int_{C\lambda_{\max}^{\frac{1}{2}}(\Sigma)\sqrt{\log J(m+n)}}^\infty 3J\exp\left\{-c\min\left\{\frac{t}{\lambda_{\max}^{\frac{1}{2}}(\Sigma)}, \frac{t^2}{\lambda_{\max}(\Sigma)(m+n)}\right\}\right\}\mathrm{d}t\\ \le& C\lambda_{\max}^{\frac{1}{2}}(\Sigma)\sqrt{mn}+C\lambda_{\max}^{\frac{1}{2}}(\Sigma)\sqrt{\log J(m+n)}\\ &+\int_0^\infty 3\exp\left\{-c\min\left\{\frac{t}{\lambda_{\max}^{\frac{1}{2}}(\Sigma)}, \frac{t^2}{\lambda_{\max}(\Sigma)(m+n)}\right\}\right\}\mathrm{d}t\\ \le& C\lambda_{\max}^{\frac{1}{2}}(\Sigma)(\sqrt{mn}+\sqrt{(\log J(m+n)}+\sqrt{m+n})\\ \le& C\lambda_{\max}^{\frac{1}{2}}(\Sigma)\sqrt{(m+\log J)n},\end{aligned} \tag{119}$$

其中最后一行用到了假设 3.2。 □

C.2 序数-PU 模型的辅助引理

引理 *B.2* 的证明。令 $u_i = u(x_i,\ \theta)$, $u_i^* = u(x_i,\ \theta^*)$。首先注意到 $\exists t\in[0,1]$ 使得

$$\begin{aligned}&\mathrm{I}(\Delta)\\ =&\frac{1}{n}\sum_{i=1}^n\begin{pmatrix}x_i^\top\Delta_{1:p}-\Delta_{p+1}\\ -\Delta_{p+2}\\ \vdots\\ -\Delta_{p+K}\end{pmatrix}^\top\nabla f^{\mathrm{ON}}(u_i^t)^\top\nabla^2 A(f^{\mathrm{ON}}(u_i^t))\nabla f^{\mathrm{ON}}(u_i)\begin{pmatrix}x_i^\top\Delta_{1:p}-\Delta_{p+1}\\ -\Delta_{p+2}\\ \vdots\\ -\Delta_{p+K}\end{pmatrix},\end{aligned}$$

其中 $u_i^t = u_i^* + t(u_i - u_i^*)$。然后经过一些计算可得

$$\begin{aligned}\mathrm{I}(\Delta) \geq & \frac{1}{n}\sum_{i=1}^{n}\Bigg[\lambda_{\min}\left(\nabla f^{\mathrm{ON}}(u_i^t)^\top \nabla^2 A(f^{\mathrm{ON}}(u_i^t))\nabla f^{\mathrm{ON}}(u_i^t)\right) \\ & - \left\|\nabla f^{\mathrm{ON}}(u_i^t)^\top \nabla^2 A(f^{\mathrm{ON}}(u_i^t))(\nabla f^{\mathrm{ON}}(u_i^t) - \nabla f^{\mathrm{ON}}(u_i))\right\|\Bigg] \\ & \left((x_i^\top \Delta_{1:p} - \Delta_{p+1})^2 + \sum_{j=2}^{K}\Delta_{p+j}^2\right).\end{aligned}$$

为了进一步求得上述项的下界，这里有两个关键步骤：(i) 求 $\nabla f_{\mathrm{ON}}(u_i^t)^\top \nabla f^{\mathrm{ON}}(u_i^t)$ 的最小特征值的下界；(ii) 求 $\|\nabla f^{\mathrm{ON}}(u_i^t) - \nabla f^{\mathrm{ON}}(u_i)\|$ 的上界。

第二个关键步骤是说明为什么需要在引理 4.4 中约束 θ 与 θ^* 之间的距离：通过关注某个 $R_0, r_0 > 0$ 的 $\theta \in \theta^* + S(R_0, r_0)$，我们可以适当地给 $\|\nabla f^{\mathrm{ON}}(u_i^t) - \nabla f^{\mathrm{ON}}(u_i)\|$ 上界，从而证明对于某个 $\gamma > 0$，$\mathrm{I}(\Delta) \geq \frac{\gamma}{n}\sum_{i=1}^{n}\left((x_i^\top \Delta_{1:p} - \Delta_{p+1})^2 + \sum_{j=2}^{K}\Delta_{p+j}^2\right)$。引理 4.2 和引理 C.1 的证明包含在第 C.3 节中。如引理 A.4 的证明所示，对于任意向量 $\eta \in \mathbb{R}^K$，

$$\lambda_{\min}(\nabla^2 A(\eta)) \geq \frac{\min_j e^{\eta_j}}{\left(1 + \sum_{l=1}^{K} e^{\eta_l}\right)^2}.$$

回忆一下 $\Delta \in S(R_0, r_0)$，这意味着 $\max\{\|\Delta_{1:p}\|_1, \|\Delta_{(p+1):(p+K)}\|_1\} \leq R_0$，且 $-\min_{2\leq j\leq K}\Delta_{p+j} \leq r_0 < r^*$。由于 $\exp\{f_j(u_i^t)\} = \frac{n_j}{\pi_j n_u} p_j(u_i^t)$，其中 p_j 的定义见 (123)，根据引理 C.2，可以证明

$$\min_j \exp\{f_j(u_i^t)\} = \min_j \frac{n_j}{\pi_j n_u} p_j(u_i^t) \geq \min_j \frac{n_j}{\pi_j n_u}(h^{\mathrm{ON}}(R_0, r_0))^{-1}.$$

因此我们有

$$\lambda_{\min}(\nabla^2 A(f^{\mathrm{ON}}(u_i^t))) \geq \min_j \frac{n_j}{4\pi_j n_u h^{\mathrm{ON}}(R_0, r_0)}.$$

现在回顾主论文中的引理 4.2 和引理 4.3。这里我们对引理 4.3 稍作修改，使其更适合我们的用途：

引理 C.1。对于任意 $\Delta \in S(R_0, r_0)$，

$$\|\nabla f^{\mathrm{ON}}(u_i^t) - \nabla f^{\mathrm{ON}}(u_i)\| \leq \sqrt{5}[(h^{\mathrm{ON}}(R_0, r_0))^2 + h^{\mathrm{ON}}(R_0, r_0)](C_x + 1)R_0 K,$$

其中 $h^{\mathrm{ON}}(\cdot)$ 的定义见 (13)。

根据引理 4.2，

$$\begin{aligned}& \lambda_{\min}(\nabla f^{\mathrm{ON}}(u_i^t)^\top \nabla^2 A(f^{\mathrm{ON}}(u_i^t))\nabla f^{\mathrm{ON}}(u_i^t)) \\ \geq & \lambda_{\min}(\nabla f^{\mathrm{ON}}(u_i^t)^\top \nabla f^{\mathrm{ON}}(u_i^t))\lambda_{\min}(\nabla^2 A(f^{\mathrm{ON}}(u_i^t))) \\ \geq & \min_j \frac{n_j}{\pi_j n_u} \frac{e^{2(C_x+1)(R^*+R_0)}}{16Kh^{\mathrm{ON}}(R_0, r_0)(1 + e^{(C_x+1)(R^*+R_0)})^4}.\end{aligned} \tag{120}$$

现在我们给出 $\left\|\nabla f_{\mathrm{ON}}(u_i^t)\nabla^2 A(f_{\mathrm{ON}}(u_i^t))(\nabla f_{\mathrm{ON}}(u_i^t) - \nabla f^{\mathrm{ON}}(u_i))\right\|$ 的上界。根据（125），

$$\begin{aligned}
\|\nabla f^{\mathrm{ON}}(u_i^t)v\|_2^2 =& \sum_{j=1}^{K-1} \frac{(\alpha_j(u_i^t)\widetilde{v}_j - \alpha_{j+1}(u_i^t)\widetilde{v}_{j+1})^2}{p_j^2(u_i^t)} + \frac{\alpha_K^2(u_i^t)\widetilde{v}_K^2}{p_K^2(u_i^t)} \\
\leq& \sum_{j=1}^{K-1} \frac{2\alpha_j^2(u_i^t)\widetilde{v}_j^2 + 2\alpha_{j+1}^2(u_i^t)\widetilde{v}_{j+1}^2}{p_j^2(u_i^t)} + \frac{\alpha_K^2(u_i^t)\widetilde{v}_K^2}{p_K^2(u_i^t)} \\
\leq& \frac{2(1+e^{(C_x+1)(R^*+R_0)})^4}{(r^*-r_0)^2 e^{2(C_x+1)(R^*+R_0)}} \sum_{j=1}^{K-1}(\widetilde{v}_j^2 + \widetilde{v}_{j+1}^2) \\
&+ (1+e^{(C_x+1)(R^*+R_0)})^2\widetilde{v}_K^2 \\
\leq& (h^{\mathrm{ON}}(R_0,r_0))^2 \left(2\sum_{j=1}^{K-1}(\widetilde{v}_j^2 + \widetilde{v}_{j+1}^2) + \widetilde{v}_K^2\right) \\
\leq& (h^{\mathrm{ON}}(R_0,r_0))^2 \left(2\sum_{j=1}^{K-1}(j\sum_{l=1}^{j} v_l^2 + (j+1)\sum_{l=1}^{j+1} v_l^2) + K\sum_{l=1}^{K} v_l^2\right) \\
\leq& (h^{\mathrm{ON}}(R_0,r_0))^2 \sum_{l=1}^{K}\left(2l + 2\sum_{j=l}^{K-1}(2j+1) + K\right) v_l^2 \\
\leq& 3(h^{\mathrm{ON}}(R_0,r_0))^2 K^2 \|v\|_2^2,
\end{aligned}$$

这意味着

$$\|\nabla f^{\mathrm{ON}}(u_i^t)\| \leq \sqrt{3}h^{\mathrm{ON}}(R_0,r_0)K.$$

此外，根据引理 C.1,

$$\|\nabla f^{\mathrm{ON}}(u_i^t) - \nabla f^{\mathrm{ON}}(u_i)\| \leq \sqrt{5}((h^{\mathrm{ON}}(R_0,r_0))^2 + h^{\mathrm{ON}}(R_0,r_0))(C_x+1)R_0K.$$

而对于 $\|\nabla^2 A(f^{\mathrm{ON}}(u_i^t))\|$，注意，对于任意 $v \in \mathbb{R}^K$,

$$\begin{aligned}
v^\top \nabla^2 A(f^{\mathrm{ON}}(u_i^t))v =& \sum_{j=1}^{K}(\nabla A(f^{\mathrm{ON}}(u_i^t)))_j v_j^2 - \left(\sum_{j=1}^{K}(\nabla A(f^{\mathrm{ON}}(u_i^t)))_j v_j\right)^2 \\
\leq& \sum_{j=1}^{K}(\nabla A(f^{\mathrm{ON}}(u_i^t)))_j v_j^2 \\
\leq& \max_j \frac{e^{f^{\mathrm{ON}}(u_{ij}^t)}}{1+\sum_{k=1}^{K} e^{f^{\mathrm{ON}}(u_{ik}^t)}}\|v\|_2^2 \\
\leq& \|v\|_2^2,
\end{aligned}$$

因此 $\|\nabla^2 A(f^{\mathrm{ON}}(u_i^t))\| = \lambda_{\max}(\nabla^2 A(f^{\mathrm{ON}}(u_i^t))) \leq 1$。因此,

$$\begin{aligned}
&\|\nabla f^{\mathrm{ON}}(u_i^t)^\top \nabla^2 A(f^{\mathrm{ON}}(u_i^t))(\nabla f^{\mathrm{ON}}(u_i^t) - \nabla f^{\mathrm{ON}}(u_i))\| \\
\leq& 4((h^{\mathrm{ON}}(R_0,r_0))^3 + (h^{\mathrm{ON}}(R_0,r_0))^2)(C_x+1)R_0K^2.
\end{aligned}$$

根据引理 B.5，我们知道

$$R_0 \leq \min_j \frac{n_j}{\pi_j n_u} \frac{\left(1+e^{(C_x+1)(R^*+R_0)}\right)^{-2}}{512(C_x+1)K^3(h^{\mathrm{ON}}(R_0,r_0))^3(1+h^{\mathrm{ON}}(R_0,r_0))},$$

这进而意味着

$$\begin{aligned}
&\|\nabla f^{\mathrm{ON}}(u_i^t)^\top \nabla^2 A(f^{\mathrm{ON}}(u_i^t))(\nabla f^{\mathrm{ON}}(u_i^t) - \nabla f^{\mathrm{ON}}(u_i))\| \\
\leq &\min_j \frac{n_j}{\pi_j n_u} \frac{\left(1 + e^{(C_x+1)(R^*+R_0)}\right)^{-2}}{128Kh^{\mathrm{ON}}(R_0, r_0)} \\
\leq &\min_j \frac{n_j}{\pi_j n_u} \frac{e^{2(C_x+1)(R^*+R_0)}}{32Kh^{\mathrm{ON}}(R_0, r_0)\left(1 + e^{(C_x+1)(R^*+R_0)}\right)^4},
\end{aligned}$$

其中，最后一行是由于以下事实： $e^{(C_x+1)(R^*+R_0)} \geq 1$。因此，

$$\begin{aligned}
&\lambda_{\min}(\nabla f^{\mathrm{ON}}(u_i^t)^\top \nabla^2 A(f^{\mathrm{ON}}(u_i^t))\nabla f^{\mathrm{ON}}(u_i^t)) \\
&- \|\nabla f^{\mathrm{ON}}(u_i^t)^\top \nabla^2 A(f^{\mathrm{ON}}(u_i^t))(\nabla f^{\mathrm{ON}}(u_i^t) - \nabla f^{\mathrm{ON}}(u_i))\| \\
\geq &\min_j \frac{n_j}{\pi_j n_u} \frac{e^{2(C_x+1)(R^*+R_0)}}{32Kh^{\mathrm{ON}}(R_0, r_0)(1 + e^{(C_x+1)(R^*+R_0)})^4},
\end{aligned}$$

至此，证明完成。 □

引理 *B.3* 的证明。根据假设 3.1， $\{(x_i^\top \widetilde{\Delta}_{1:p}\}_{i=1}^n$ 是独立次高斯随机变量~，其参数为 $\sigma\|\Delta_{1:p}\|_2$，这意味着

$$\|(x_i^\top \widetilde{\Delta}_{1:p})^2\|_{\psi_1} \leq C\sigma^2 \|\widetilde{\Delta}_{1:p}\|_2^2 \leq C\sigma^2,$$

and

$$\|x_i^\top \widetilde{\Delta}_{1:p}|\widetilde{\Delta}_{p+1}|\|_{\psi_2} \leq \sigma|\widetilde{\Delta}_{p+1}|\|\widetilde{\Delta}_{1:p}\|_2 \leq \frac{\sigma}{2}.$$

与此同时，注意到

$$g(\widetilde{\Delta}, \widetilde{\Delta}) = \frac{1}{n}\sum_{i=1}^n (x_i^\top \widetilde{\Delta}_{1:p})^2 - \widetilde{\Delta}_{1:p}^\top \Sigma \widetilde{\Delta}_{1:p} - 2\widetilde{\Delta}_{p+1}\widetilde{\Delta}_{1:p}^\top x_i$$

Thus

$$\begin{aligned}
\mathbb{P}\left(|g(\widetilde{\Delta}, \widetilde{\Delta})| \geq t\right) \leq &\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n (x_i^\top \widetilde{\Delta}_{1:p})^2 - \widetilde{\Delta}_{1:p}^\top \Sigma \widetilde{\Delta}_{1:p}\right| \geq \frac{t}{2}\right) \\
&+ \mathbb{P}\left(\left|\widetilde{\Delta}_{p+1}\widetilde{\Delta}_{1:p}^\top x_i\right| \geq \frac{t}{4}\right) \\
&\leq \exp\left\{-cn\min\left\{\frac{t^2}{\sigma^4}, \frac{t}{\sigma^2}\right\}\right\} + \exp\left\{-cn\frac{t^2}{\sigma^2}\right\} \\
&\leq 2\exp\left\{-cn\min\left\{\frac{t^2}{\sigma^4}, \frac{t^2}{\sigma^2}, \frac{t}{\sigma^2}\right\}\right\}.
\end{aligned} \tag{121}$$

□

引理 *B.4* 的证明。 对于任意 $1 \leq l \leq K$，经计算可得

$$\begin{aligned}
|(\nabla q_i(u_i(\Delta)))_l| = &\Big|\epsilon_i^\top \left(\nabla f^{\mathrm{ON}}(u_i(\theta^*) + u_i(\Delta)) - \nabla f^{\mathrm{ON}}(u_i(\theta^*))\right)_{:,l} \\
&+ \epsilon_i^\top (\nabla^2 f^{\mathrm{ON}}(u_i(\theta^*) + u_i(\Delta)))_{:,:,l} u_i(\Delta)\Big| \\
\leq &2\left\|\nabla f^{\mathrm{ON}}(u_i(\theta^*) + u_i(\Delta)) - \nabla f^{\mathrm{ON}}(u_i(\theta^*))\right\|_\infty \\
&+ 2\|u_i(\Delta)\|_1 \left\|\nabla^2 f^{\mathrm{ON}}(u_i(\theta^*) + u_i(\Delta))\right\|_\infty.
\end{aligned}$$

根据引理 C.2，我们有

$$|(\nabla q_i(u_i(\Delta)))_l| \le h^{\mathrm{ON}}(R_0, r_0) + 2\|u_i(\Delta)\|_1 \left[\frac{2}{3}h^{\mathrm{ON}}(R_0, r_0) + \frac{1}{8}(h^{\mathrm{ON}}(R_0, r_0))^2\right].$$

同时，由于 $\|u_i(\Delta)\|_1 \le C_x\|\Delta_{1:p}\|_1 + \|\Delta_{(p+1):(p+K)}\|_1 \le (C_x+1)R_0$，我们得到

$$\begin{aligned}\|\nabla q_i(u_i(\Delta))\|_2 \le& \sqrt{K}h^{\mathrm{ON}}(R_0, r_0) + 2\sqrt{K}(C_x+1)R_0\left[\frac{2}{3}h^{\mathrm{ON}}(R_0, r_0) + \frac{1}{8}(h^{\mathrm{ON}}(R_0, r_0))^2\right]\\ =&(\frac{4}{3}(C_x+1)R_0+1)\sqrt{K}h^{\mathrm{ON}}(R_0, r_0) + \frac{(C_x+1)R_0}{4}\sqrt{K}(h^{\mathrm{ON}}(R_0, r_0))^2.\end{aligned}$$

□

引理 *B.5* 的证明。 首先，我们注意到假设 3.4 表明

$$\begin{aligned}R_0 &\le 80\frac{\left(1+e^{(C_x+1)(R^*+0.01)}\right)^{-6}}{512(C_x+1)K^3(1+\frac{2}{r^*-r_0})^3(2+\frac{2}{r^*-r_0})}\\ &\le \frac{80}{1024K^3}\\ &< 0.01,\end{aligned}$$

其中最后一行应用了 $K \ge 2$ 。因此，假设 3.4 还意味着

$$R_0 \le \min_j \frac{n_j}{\pi_j n_u}\frac{\left(1+e^{(C_x+1)(R^*+R_0)}\right)^{-6}}{512(C_x+1)K^3(1+\frac{2}{r^*-r_0})^3(2+\frac{2}{r^*-r_0})}.$$

另一方面，我们知道 $h^{\mathrm{ON}}(R_0,\ r_0) \le (1+e^{(C_x+1)(R^*+R_0)})(1+\frac{2}{r^*-r_0})$。该上界随即意味着引理 B.5 中的第二个不等式。 □

C.3 序数-PU 模型中 f^{ON} 的性质证明

在本节中，我们证明引理 4.2 和引理 C.1，它们处理了序数对数似然损失的复杂形式带来的若干主要挑战。通过一些计算可以得到，对于任意向量 $u \in \mathbb{R}^K$，$\nabla f^{\mathrm{ON}}(u) \in \mathbb{R}^{K\times K}$ 满足

$$\nabla f^{\mathrm{ON}}(u)_{jk} = \frac{\alpha_j(u)\mathbb{1}_{\{k\le j\}} - \alpha_{j+1}(u)\mathbb{1}_{\{k\le j+1\}}}{p_j(u)}.$$

其中，函数 $p:\mathbb{R}^K \to \mathbb{R}^K$ 和 $\alpha:\ \mathbb{R}^K \to \mathbb{R}^{K+1}$ 定义如下:

$$\alpha_j(u) = \begin{cases} e^{\sum_{l=1}^j u_l}\left(1+e^{\sum_{l=1}^j u_l}\right)^{-2}, & j \le K,\\ 0, & j = K+1,\end{cases} \tag{122}$$

and

$$p_j(u) = \begin{cases} (1+e^{\sum_{l=1}^{j+1} u_l})^{-1} - (1+e^{\sum_{l=1}^j u_l})^{-1}, & 1 \le j < K,\\ 1-(1+e^{\sum_{l=1}^j u_l})^{-1}, & j = K,\end{cases} \tag{123}$$

其中， $\alpha_j(u)$ $p_j(u)$指的是 $\alpha(u)$和 $p(u)$的第 j个坐标。

引理证明 *4.2.* 根据式（20）中函数 f^{ON} 的定义，可以证明，对于任意 $1 \le j,\ k \le K$ ，

$$(\nabla f^{\mathrm{ON}}(u_i^*))_{jk} = p_j(u_i^*)^{-1}\left[\alpha_j(u_i^*)\mathbb{1}_{\{k\le j\}} - \alpha_{j+1}(u_i^*)\mathbb{1}_{\{k\le j+1\}}\right]. \tag{124}$$

对于任意向量 $v \in \mathbb{R}^K$，有

$$\begin{aligned}\nabla f^{\mathrm{ON}}(u_i^t)v &= \begin{pmatrix} \frac{\alpha_1(u_i^t)}{p_1(u_i^t)} & -\frac{\alpha_2(u_i^t)}{p_1(u_i^t)} & 0 & \cdots & 0 \\ 0 & \frac{\alpha_2(u_i^t)}{p_2(u_i^t)} & -\frac{\alpha_3(u_i^t)}{p_2(u_i^t)} & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & \cdots & \frac{\alpha_{K-1}(u_i^t)}{p_{K-1}(u_i^t)} & -\frac{\alpha_K(u_i^t)}{p_{K-1}(u_i^t)} \\ 0 & \cdots & \cdots & \cdots & \frac{\alpha_K(u_i^t)}{p_K(u_i^t)} \end{pmatrix} \begin{pmatrix} v_1 \\ v_1+v_2 \\ \sum_{j=1}^3 v_j \\ \vdots \\ \sum_{j=1}^K v_j \end{pmatrix} \\ &= \left(\frac{\alpha_1(u_i^t)\widetilde{v}_1 - \alpha_2(u_i^t)\widetilde{v}_2}{p_1(u_i^t)}, \ldots, \frac{\alpha_{K-1}(u_i^t)\widetilde{v}_{K-1} - \alpha_K(u_i^t)\widetilde{v}_K}{p_{K-1}(u_i^t)}, \frac{\alpha_K(u_i^t)\widetilde{v}_K}{p_K(u_i^t)} \right)^\top, \end{aligned}$$

其中 $\widetilde{v}_k = \sum_{j=1}^k v_j$。由此可得

$$\begin{aligned}\left\|\nabla f^{\mathrm{ON}}(u_i^t)v\right\|_2^2 &= \sum_{j=1}^{K-1} \frac{(\alpha_j(u_i^t)\widetilde{v}_j - \alpha_{j+1}(u_i^t)\widetilde{v}_{j+1})^2}{p_j^2(u_i^t)} + \frac{\alpha_K^2(u_i^t)\widetilde{v}_K^2}{p_K^2(u_i^t)} \\ &\geq \frac{1}{K}\sum_{k=1}^{K-1}\sum_{j=k}^{K-1} \frac{(\alpha_j(u_i^t)\widetilde{v}_j - \alpha_{j+1}(u_i^t)\widetilde{v}_{j+1})^2}{p_j^2(u_i^t)} + \frac{\alpha_K^2(u_i^t)\widetilde{v}_K^2}{p_K^2(u_i^t)}, \end{aligned} \tag{125}$$

因为

$$\sum_{j=1}^{K-1} \frac{(\alpha_j(u_i^t)\widetilde{v}_j - \alpha_{j+1}(u_i^t)\widetilde{v}_{j+1})^2}{p_j^2(u_i^t)} = \max_k \sum_{j=k}^{K-1} \frac{(\alpha_j(u_i^t)\widetilde{v}_j - \alpha_{j+1}(u_i^t)\widetilde{v}_{j+1})^2}{p_j^2(u_i^t)}.$$

根据柯西-施瓦茨不等式,

$$\begin{aligned}&\left(\sum_{j=k}^{K-1} \frac{(\alpha_j(u_i^t)\widetilde{v}_j - \alpha_{j+1}(u_i^t)\widetilde{v}_{j+1})^2}{p_j^2(u_i^t)} + \frac{\alpha_K^2(u_i^t)\widetilde{v}_K^2}{p_K^2(u_i^t)} \right) \left(\sum_{j=k}^{K} p_j^2(u_i^t) \right) \\ \geq &\left(\sum_{j=k}^{K-1} (\alpha_j(u_i^t)\widetilde{v}_j - \alpha_{j+1}(u_i^t)\widetilde{v}_{j+1}) + \alpha_K(u_i^t)\widetilde{v}_K \right)^2 = \alpha_k^2(u_i^t)\widetilde{v}_k^2, \end{aligned}$$

这意味着

$$\begin{aligned}\left\|\nabla f^{\mathrm{ON}}(u_i^t)v\right\|_2^2 &\geq \frac{1}{K}\left[\frac{\alpha_K^2(u_i^t)\widetilde{v}_K^2}{p_K^2(u_i^t)} + \sum_{k=1}^{K-1} \left(\sum_{j=k}^{K} p_j^2(u_i^t) \right)^{-1} \alpha_k^2(u_i^t)\widetilde{v}_k^2 \right] \\ &\geq \frac{1}{K}\sum_{k=1}^{K} \alpha_k^2(u_i^t)\widetilde{v}_k^2. \end{aligned}$$

此外，由于

$$\|v\|_2^2 = \sum_{j=1}^{K-1} (\widetilde{v}_{j+1} - \widetilde{v}_j)^2 + \widetilde{v}_1^2 \leq 4\sum_{j=1}^{K} \widetilde{v}_j^2 \leq \frac{4}{\min_{1\leq k\leq K} \alpha_k^2(u_i^t)} \sum_{j=1}^{K} \alpha_j^2(u_i^t)\widetilde{v}_j^2,$$

我们有

$$\left\|\nabla f^{\mathrm{ON}}(u_i^t)v\right\|_2^2 \geq \frac{\min_k \alpha_k^2(u_i^t)}{4K}\|v\|_2^2.$$

回想一下式（122）中 $\alpha_k(u)$ 的定义，以及以下事实:

$$\left| \sum_{j=1}^{k} u_{ij}^t \right| \leq (C_x + 1)(R^* + R_0),$$

可以证明，对于 $1 \le k \le K$,

$$\alpha_k(u_i^t) \ge \frac{e^{(C_x+1)(R^*+R_0)}}{(1+e^{(C_x+1)(R^*+R_0)})^2}.$$

因此

$$\lambda_{\min}(\nabla f^{\mathrm{ON}}(u_i^t)^\top \nabla f^{\mathrm{ON}}(u_i^t)) \ge \frac{e^{2(C_x+1)(R^*+R_0)}}{4K(1+e^{(C_x+1)(R^*+R_0)})^4}.$$

□

引理 *C.1* 的证明。至于对 $\|\nabla f^{\mathrm{ON}}(u_i^t) - \nabla f^{\mathrm{ON}}(u_i)\|$ 进行上界估计，我们首先回顾 (124) 中 ∇f_{ON} 的每个条目，并注意到

$$\begin{aligned}
&\|\nabla f^{\mathrm{ON}}(u_i^t) - \nabla f^{\mathrm{ON}}(u_i)\| \\
\le &\|\nabla f^{\mathrm{ON}}(u_i^t) - \nabla f^{\mathrm{ON}}(u_i)\|_F \\
= &\Bigg[\sum_{j=1}^{K} j\left(\frac{\alpha_j(u_i^t)}{p_j(u_i^t)} - \frac{\alpha_j(u_i)}{p_j(u_i)} - \frac{\alpha_{j+1}(u_i^t)}{p_j(u_i^t)} + \frac{\alpha_{j+1}(u_i)}{p_j(u_i)}\right)^2 \\
&+ \sum_{j=1}^{K-1}\left(\frac{\alpha_{j+1}(u_i^t)}{p_j(u_i^t)} - \frac{\alpha_{j+1}(u_i)}{p_j(u_i)}\right)^2\Bigg]^{\frac{1}{2}} \\
\le &\sqrt{2K^2+3K-1}\max_{j,k}\left|\frac{\alpha_k(u_i^t)}{p_j(u_i^t)} - \frac{\alpha_k(u_i)}{p_j(u_i)}\right|.
\end{aligned} \tag{126}$$

可以证明

$$\begin{aligned}
&\max_{j,k}\left|p_j^{-1}(u_i^t)\alpha_k(u_i^t) - p_j^{-1}(u_i)\alpha_k(u_i)\right| \\
\le &\max_j\left|p_j^{-1}(u_i^t) - p_j^{-1}(u_i)\right| + \max_{j,k} p_j^{-1}(u_i)\left|\alpha_k(u_i^t) - \alpha_k(u_i)\right|,
\end{aligned} \tag{127}$$

其中在第二行中，我们应用了这样的事实：对于任意 $w \in \mathbb{R}^K$， $0 < \alpha_k(w) < 1$。同时，对于任意$w \in \mathbb{R}^K$、$1 \le j < K$,

$$\nabla_w(p_j^{-1}(w)) = \frac{e^{\sum_{l=1}^{j+1} w_l}}{p_j^2(w)(1+e^{\sum_{l=1}^{j+1} w_l})^2}(1_{j+1}^\top, 0_{K-j-1}^\top)^\top - \frac{e^{\sum_{l=1}^{j} w_l}}{p_j^2(w)(1+e^{\sum_{l=1}^{j} w_l})^2}(1_j^\top, 0_{K-j}^\top)^\top,$$

and

$$\nabla_w(p_K^{-1}(w)) = -p_K^{-2}(w)\frac{e^{\sum_{l=1}^{K} w_l}}{(1+e^{\sum_{l=1}^{K} w_l})^2}1_K$$

由于对于 $x \in \mathbb{R}, \quad 0 < \quad {}^{e^x}\!/_{(1+e^x)_2} < 1$，因此我们有

$$\left\|\nabla_w(p_j^{-1}(w))\right\|_\infty \le p_j^{-2}(w) \le (h^{\mathrm{ON}}(R,r))^2,$$

对于位于 u_i^t 与 u_i 之间的任意 w，这意味着

$$\max_j |p_j^{-1}(u_i^t) - p_j^{-1}(u_i)| \le (h^{\mathrm{ON}}(R,r))^2\|u_i^t - u_i\|_1. \tag{128}$$

为界定 $p_j^{-1}(u_i)|\alpha_k(u_i^t) - \alpha_k(u_i)|$，我们应用以下引理，该引理给出了函数 $p(\cdot)$、 $\alpha(\cdot)$、 $f^{\mathrm{ON}}(\cdot)$ 的若干上界：

引理 C.2. 在 (123)、(122) 和 (20) 中定义的函数 $p(\cdot)$ $\alpha(\cdot)$ 和 $f^{\mathrm{ON}}(\cdot)$ 满足以下界:

$$0 < \alpha_j(u) \le \frac{1}{4}, \quad p_j(u) \ge \min\{\frac{-\max_{k>1} u_k e^{\|u\|_1}}{(1+e^{\|u\|_1})^2}, \frac{1}{1+e^{\|u\|_1}}\}, \tag{129}$$

若 $u_j < 0$ 对 $j > 1$成立。特别地，若 $u = u(x, \theta)$ 对任意 $\|x\|_\infty \le C_x$成立，且 $\theta \in \theta^* + S(R, r)$ f_o r

$$S(R,r) = \{\Delta : \|\Delta_{1:p}\|_1, \|\Delta_{(p+1):(p+K)}\|_1 \le R, \min_{2\le j\le K} \Delta_{p+j} \ge -r\},$$

then

$$\begin{aligned} p_j(u) \ge (h^{\mathrm{ON}}(R,r))^{-1}, \quad \|\nabla f^{\mathrm{ON}}(u)\|_\infty \le \frac{1}{4} h^{\mathrm{ON}}(R,r) \\ \|\nabla^2 f^{\mathrm{ON}}(u)\|_\infty \le \frac{2}{3} h^{\mathrm{ON}}(R,r) + \frac{1}{8}(h^{\mathrm{ON}}(R,r))^2. \end{aligned} \tag{130}$$

引理 C.2 将在后文证明。因此，根据引理 C.2, $p_j^{-1}(u_i) \le h^{\mathrm{ON}}(R_0, r_0)$。此外，令 $g(x) = \frac{e^x}{(1+e^x)^2}$，则 $\alpha_k(u_i^t) - \alpha_k(u_i) = g(\sum_{j=1}^k u_{ij}^t) - g(\sum_{j=1}^k u_{ij})$。由于

$$|g'(x)| = \frac{|e^x - e^{2x}|}{(1+e^x)^3} \le 1,$$

$$\begin{aligned} & p_j^{-1}(u_i)|\alpha_k(u_i^t) - \alpha_k(u_i)| \\ \le & h^{\mathrm{ON}}(R,r)\|u_i^t - u_i\|_1. \end{aligned} \tag{131}$$

由于 $\|u_i^t - u_i\|_1 \le \|u_i^* - u_i\|_1 \le (C_x+1)R$, 结合 (126)、(127)、(128) 和 (131) 可得

$$\|\nabla f^{\mathrm{ON}}(u_i^t) - \nabla f^{\mathrm{ON}}(u_i)\| \le \sqrt{5}((h^{\mathrm{ON}}(R,r))^2 + h^{\mathrm{ON}}(R,r))(C_x+1)RK,$$

□

引理 *C.2* 的证明。令 $g(x) = \frac{x}{(1+x)^2}$，则可证明，对任意 $x \ge 0$, $g(x) \le g(1) = \frac{1}{4}$，这直接推出 $0 < \alpha_j(u) \le \frac{1}{4}$。而对于 $p_j(u)$, 由 (123) 可证明

$$p_K(u) = 1 - (1 + e^{\sum_{l=1}^K u_l})^{-1} = (1 + e^{-\sum_{l=1}^K u_l})^{-1} \ge \frac{1}{1+e^{\|u\|_1}}.$$

对于 $1 \le j < K$, $p_j(u) = -u_{j+1} g(\xi)$ 对某个 $\xi \in [e^{\sum_{l=1}^{j+1} u_l}, e^{\sum_{l=1}^j u_l}]$ 成立。注意到 $g(x)$ 在 $x < 1$ 时递增、在 $x > 1$ 时递减，且 $g(\frac{1}{x}) = g(x)$，我们知道 –

$$g(\xi) \ge \min\{g(e^{\sum_{l=1}^{j+1} u_l}), g(e^{\sum_{l=1}^j u_l})\} \ge \frac{e^{\|u\|_1}}{(1+e^{\|u\|_1})^2},$$

这意味着 $p_j(u) \ge \frac{}{\text{最大值}_{k>1} u_k e^{\|u\|_1}/(1+e^{\|u\|_1})^2}$。此外，回顾我们已将 u 定义为 $u = u(x, \theta)$，并假设 $\|x\|_\infty \le C_x$, $\theta \in S(R, r)$。因此有

$$\|u(x,\theta)\|_1 \le C_x\|\theta_{1:p}\|_1 + \|\theta_{(p+1):(p+K)}\|_1 \le (C_x+1)(R^* + R),$$

并且对于 $k > 1$,

$$u_k(x,\theta) \le -\min_{k>1} \theta_{p+k} \le r - r^*.$$

因此，根据式 (13) 中 $h^{\mathrm{ON}}(\cdot,\cdot)$ 的定义,

$$p_j(u) \ge (h^{\mathrm{ON}}(R,r))^{-1}$$

对于 $1 \le j \le K$。

现在我们开始求 $\|\nabla f^{\mathrm{ON}}(u)\|_\infty$ 和 $\|\nabla^2 f^{\mathrm{ON}}(u)\|_\infty$ 的上界。计算表明，对于任意向量 $u \in \mathbb{R}^K$，$\nabla f^{\mathrm{ON}}(u) \in \mathbb{R}^{K\times K}$ 满足

$$\nabla f^{\mathrm{ON}}(u)_{jk} = \frac{\alpha_j(u)\mathbb{1}_{\{k\le j\}} - \alpha_{j+1}(u)\mathbb{1}_{\{k\le j+1\}}}{p_j(u)}.$$

由于 $0 < \alpha_j(u) \le \frac{1}{4}$，$p_j(u) \ge (h^{\mathrm{ON}}(R, r))^{-1}$，因此有 $\|\nabla f^{\mathrm{ON}}(u)\|_\infty \le \frac{1}{4}h^{\mathrm{ON}}(R, r)$。同时，可以证明

$$\begin{aligned}(\nabla^2 f^{\mathrm{ON}}(u))_{jkl} = &\left(\frac{(\nabla\alpha_j(u))_l}{p_j(u)} - \frac{\alpha_j(u)(\nabla p_j(u))_l}{p_j^2(u)}\right)\mathbb{1}_{\{k\le j\}}\\ &- \left(\frac{(\nabla\alpha_{j+1}(u))_l}{p_j(u)} - \frac{\nabla p_j(u))_l\alpha_{j+1}(u)}{p_j^2(u)}\right)\mathbb{1}_{\{k\le j+1\}},\end{aligned}$$

和 $|(\nabla p_j(u))_l| = |\alpha_j(u)\mathbb{1}_{\{l\le j\}} - \alpha_{j+1}(u)\mathbb{1}_{\{l\le j+1\}}| \le \frac{1}{4}$，-

$$|(\nabla\alpha_j(u))_l| = \frac{\left|e^{\sum_{m=1}^j u_j} - e^{2\sum_{m=1}^j u_j}\right|}{(1+e^{\sum_{m=1}^j u_j})^3}\mathbb{1}_{\{l\le j\}} \le \frac{1}{3}.$$

因此有

$$\|\nabla^2 f^{\mathrm{ON}}(u)\|_\infty \le \max_j\left(\frac{2}{3}h^{\mathrm{ON}}(R,r) + \frac{1}{8}(h^{\mathrm{ON}}(R,r))^2\right).$$

□

D 正则化 EM 算法、额外的收敛保证以及收敛证明

多项式 PU 模型的 EM 算法：另一种估计方法是 EM算法，该方法已被广泛应用于缺失/隐变量模型相关文献中。下面我们针对模型提出一种正则化 EM 算法。基于完整对数似然函数

$$\log L_f^{\mathrm{MN}}(\Theta, b; \{(x_i, y_i, z_i, s_i = 1)\}_{i=1}^n)$$

如引理 1 所述，我们在算法 1 中总结了适用于多项式 PU 模型的正则化 EM 算法。在算法 1 的 E步中，我们通过以下方式估计未观测的 $\mathbb{1}_{\{y_i=j\}}$

$$\mathbb{P}_{\Theta^m,b^m}(y_i = j|x_i, z_i, s_i = 1)$$

在第 m 次迭代中，对 $j = 0$、.....、K 进行估计。由于 $z_i = k > 0$ 意味着 $y_i = k$，可以证明

$$\mathbb{P}_{\Theta^m,b^m}(y_i = j|x_i, z_i = k, s_i = 1) = \mathbb{1}_{\{j=k\}}$$

如果 $k > 0$。另一方面，当 $z_i = 0$ 时，

$$\begin{aligned}&\mathbb{P}_{\Theta^m,b^m}(y_i = j|x_i, z_i = 0, s_i = 1)\\ =&\frac{\mathbb{P}_{\Theta^m,b^m}(y_i = j, z_i = 0|x_i, s_i = 1)}{\sum_{k=1}^K \mathbb{P}_{\Theta^m,b^m}(y_i = k, z_i = 0|x_i, s_i = 1) + \mathbb{P}_{\Theta^m,b^m}(y_i = 0, z_i = 0|x_i, s_i = 1)}\\ =&\frac{e^{x_i^\top\Theta_j^m + b_j^m}}{1 + \sum_{k=1}^K e^{x_i^\top\Theta_k^m + b_k^m}},\end{aligned}$$

其中，最后一行可由引理 2.1 中给出的完整数据对数似然函数 $\log L_f^{\mathrm{MN}}$ 推导得出。在 M 步中，我们最小化正则化完整对数似然损失

$$-\frac{1}{n}\log L_f^{\mathrm{MN}}(\Theta, b; \{(x_i, \widehat{y}_i(\Theta^m, b^m), z_i, s_i = 1)\}_{i=1}^n) + P_\lambda^{\mathrm{MN}}(\Theta),$$

其中，我们使用 $\widehat{y}_{ik}(\Theta^m,\ b^m)$ 作为 $\mathbb{1}_{\{y_i=k\}},\ k = 1, \ldots,\ K$ 的替代项。

为求解 M步，可以应用近端梯度下降算法 [Wright et al., 2009]。对于初始化，我们仍考虑任意满足正则化观测对数似然损失 $\mathcal{F}_n^{\mathrm{MN}}$ 不大于任何仅截距模型的 Θ^0, b^0 。在没有任何先验知识的情况下，可以直接选择 $(0_{p\times K},\ \arg\min_b \mathcal{F}_n^{\mathrm{MN}}(0_{p\times K},\ b))$ 作为初始值。

算法 1： 多项式 PU 模型的期望最大化

输入：Θ^0, b^0，使得 $\mathcal{F}_n^{\mathrm{MN}}(\Theta^0,\ b^0) \le \min_b \mathcal{F}_n^{\mathrm{MN}}(0_{p\times K},\ b)$

对于 $m = 0, 1, \cdots$ **执行**

E步：按如下方式计算 $\widehat{y}(\Theta^m, b^m) \in \mathbb{R}^{n\times K}$，其中每个条目 $\widehat{y}_{ij}$ 都是对 $\mathbb{1}_{\{y_i=j\}}$的估计：

$$\hat{y}_{ij}(\Theta^m, b^m) = \begin{cases} \frac{e^{x_i^\top \Theta_j^m + b_j^m}}{1+\sum_{k=1}^K e^{x_i^\top \Theta_k^m + b_k^m}} & \text{if } z_i = 0 \\ 1 & \text{if } z_i = j \\ 0 & \text{else} \end{cases}$$

M步：通过 $\Theta^{m+1},\ b^{m+1}$ 获得

$$(\Theta^{m+1}, b^{m+1}) \in \arg\min_{\Theta, b} \frac{1}{n}\sum_{i=1}^n \left[\log\left(1 + \sum_{k=1}^K (1 + \frac{n_k}{\pi_k n_u}) e^{x_i^\top \Theta_k + b_k}\right) - \sum_{j=1}^K \hat{y}_{ij}(\Theta^m, b^m)(x_i^\top \Theta_j + b_j) \right] + P_\lambda^{\mathrm{MN}}(\Theta)$$

结束循环

序数 PU 模型的 EM算法：我们还为序数 PU 模型提出了以下正则化 EM 算法。

算法2：序数-PU 模型的期望最大化

输入： θ^0 使得 $\mathcal{F}_n^{\mathrm{ON}}(\theta^0) \leq \min_{\theta_{1:p}=0} {}_{p\times 1}\, \mathcal{F}_n^{\mathrm{ON}}(\theta)$

对于 $m = 0, 1, \cdots$ **执行**

E步：按如下方式计算 $\widehat{y}(\theta^m) \in \mathbb{R}^{n\times K}$，其中每个条目 $[\widehat{y}(\theta^m)]_{ij}$ 都是对 $\mathbb{1}_{\{y_i=j\}}$ 的估计：

$$\hat{y}_{ij}(\theta^m) = \begin{cases} \frac{1}{1+e^{x^\top \theta^m_{1:p} - \sum_{l=1}^{j+1} \theta^m_{p+l}}} - \frac{1}{1+e^{x^\top \theta^m_{1:p} - \sum_{l=1}^{j} \theta^m_{p+l}}}, & \text{if } z_i = 0, \\ 1, & \text{if } z_i = j, \\ 0, & \text{else.} \end{cases}$$

M步：通过以下方式获得 θ^{m+1}

$$\theta_j^{m+1} \in \arg\min_{\theta} \frac{1}{n} \sum_{i=1}^{n} \Bigg[-\sum_{j=1}^{K} \widehat{y}_{ij}(\theta^m) \log r_j(x_i, \theta) + \log\left(1 + \sum_{k=1}^{K} \frac{n_k + \pi_k n_u}{\pi_k n_u} r_k(x_i, \theta)\right) \Bigg] + P_\lambda^{\mathrm{ON}}(\theta_{1:p}).$$

结束循环

与算法 1 类似，在E步中，我们仍通过 $\mathbb{P}_{\theta^m}(y_i = j | x_i, z_i, s_i = 1)$ 估计 $\mathbb{1}_{\{y_i=j\}}$，适用于第 m 次迭代时的 $j = 0, \ldots, K$，且满足 $\mathbb{P}_{\theta^m}(y_i = j | x_i, z_i, s_i = 1) = \mathbb{1}_{\{j=z_i\}}$（如果 $z_i > 0$）。而如果 $z_i = 0$，根据引理 2.1 中给出的完整对数似然函数 $\log L_f^{\mathrm{MN}}$，

$$\begin{aligned} &\mathbb{P}_{\theta^m}(y_i = j | x_i, z_i = 0, s_i = 1) \\ =&\frac{r_j(x_i, \theta^m)}{\sum_{k=1}^{K} r_k(x_i, \theta^m) + 1} \\ =&(1 + e^{x^\top \theta^m_{1:p} - \sum_{l=1}^{j+1} \theta^m_{p+l}})^{-1} - (1 + e^{x^\top \theta^m_{1:p} - \sum_{l=1}^{j} \theta^m_{p+l}})^{-1}. \end{aligned}$$

M步最小化

$$-\frac{1}{n} \log L_f^{\mathrm{ON}}(\theta; \{(x_i, \widehat{y}_i(\theta^m), z_i, s_i = 1)\}_{i=1}^n) + P_\lambda^{\mathrm{ON}}(\theta),$$

并且也可以通过近端梯度下降算法 [Wright 等人， 2009] 求解

收敛性质： 事实上，我们可以证明，PGD 算法和正则化 EM 算法都具有相同的收敛性质。它们最终都会收敛到相应惩罚对数似然损失函数的驻点。

用于求解惩罚最大似然估计的 PGD 算法与 EM 算法： 从理论角度来看，应用于惩罚对数似然损失的 PGD 算法与 EM 算法相似，因为它们具有相同的收敛保证。然而，正如数值实验所示，在两种模型下，PGD 算法的速度快得多，同时统计误差与 EM 算法相近。因此，我们建议实践者使用 PGD 算法。

命题 D.1 （序数 PU 模型算法的收敛性）。 若参数迭代序列 $\{\theta^m\}_m$)由近端梯度下降算法 (8) 在适当选择步长或采用算法*2*的情况下生成，则满足以下性质：

(*i*) 序列 $\{\theta^m\}_m$ 至少有一个极限点。

(ii) 存在 $R^{\mathrm{ON}}>0$，使得 $\{\theta^m\}_m$ 的所有极限点都属于 Γ^{ON}，即优化问题 $\min_{\|\theta\|_2\le R^{\mathrm{ON}}}\mathcal{F}_n^{\mathrm{ON}}(\theta)$ 的一阶驻点集合。

(iii) 函数值序列 $\{\mathcal{F}_n^{\mathrm{ON}}(\theta^m)\}_m$ 非递增，并且当 $\theta^m\notin\Gamma^{\mathrm{ON}}$ 时，$\mathcal{F}_n^{\mathrm{ON}}(\theta^{m+1})<\mathcal{F}_n^{\mathrm{ON}}(\theta^m)$ 成立。存在 $\widetilde{\theta}\in\Gamma^{\mathrm{ON}}$，使得 $\{\mathcal{F}_n^{\mathrm{ON}}(\theta^m)\}_m$ 单调收敛到 $\mathcal{F}_n^{\mathrm{ON}}(\widetilde{\theta})$。

引理 *3.1* 和引理 *D.1* 的证明。这里我们首先证明，对于两种模型和算法，函数值序列均为非递增序列。具体而言，考虑由参数迭代序列 $\{\Theta^m,b^m\}_m$ 按照算法 1 生成的序列 $\{\mathcal{F}_n^{\mathrm{MN}}(\Theta^m,b^m)\}_m$。对于任意 $m\ge0$，可以证明，

$$
\begin{aligned}
&\mathcal{L}^{\mathrm{MN}}(\Theta,b;\{(x_i,z_i)\}_{i=1}^n)\\
=&-\frac{1}{n}\sum_{i=1}^n\log\mathbb{P}(z_i|x_i,\Theta,b)\\
=&-\frac{1}{n}\sum_{i=1}^n\mathbb{E}_{y_i|x_i,z_i,\Theta^m,b^m}\left[\log\frac{\mathbb{P}(y_i,z_i|x_i,\Theta,b)}{\mathbb{P}(z_i|x_i,\Theta,b)}\right]\\
=&-\frac{1}{n}\sum_{i=1}^n\mathbb{E}_{y_i|x_i,z_i,\Theta^m,b^m}\left[\log\frac{\mathbb{P}(y_i,z_i|x_i,\Theta,b)}{\mathbb{P}(y_i|x_i,z_i,\Theta^m,b^m)}-\log\frac{\mathbb{P}(y_i|x_i,z_i,\Theta,b)}{\mathbb{P}(y_i|x_i,z_i,\Theta^m,b^m)}\right]\\
\le&-\frac{1}{n}\sum_{i=1}^n\mathbb{E}_{y_i|x_i,z_i,\Theta^m,b^m}\left[\log\frac{\mathbb{P}(y_i,z_i|x_i,\Theta,b)}{\mathbb{P}(y_i|x_i,z_i,\Theta^m,b^m)}\right]\\
=&-\frac{1}{n}\log L_f^{\mathrm{MN}}(\Theta,b;\{(x_i,\widehat{y}_i(\Theta^m,b^m),z_i)\}_{i=1}^n)\\
&+\frac{1}{n}\sum_{i=1}^n\mathbb{E}_{y_i|x_i,z_i,\Theta^m,b^m}\log\mathbb{P}(y_i|x_i,z_i,\Theta^m,b^m),
\end{aligned}
$$

其中，第五行是因为

$$
\begin{aligned}
&\mathbb{E}_{y_i|x_i,z_i,\Theta^m,b^m}\left[\log\frac{\mathbb{P}(y_i|x_i,z_i,\Theta,b)}{\mathbb{P}(y_i|x_i,z_i,\Theta^m,b^m)}\right]\\
\le&\log\mathbb{E}_{y_i|x_i,z_i,\Theta^m,b^m}\left[\frac{\mathbb{P}(y_i|x_i,z_i,\Theta,b)}{\mathbb{P}(y_i|x_i,z_i,\Theta^m,b^m)}\right]\\
=&0,
\end{aligned}
$$

且当且仅当 $\Theta=\Theta^m$、$b=b^m$ 时等号成立。为简便起见，这里省略了从数据点 $(x_i,z_i,s_i=1)$ 到 (x_i,z_i) 的部分。因此，我们有

$$
\begin{aligned}
&\mathcal{F}_n^{\mathrm{MN}}(\Theta^{m+1},b^{m+1})\\
\le&-\frac{1}{n}\log L_f^{\mathrm{MN}}(\Theta^{m+1},b^{m+1};\{(x_i,\widehat{y}_i(\Theta^m,b^m),z_i)\}_{i=1}^n)+P_\lambda^{\mathrm{MN}}(\Theta^{m+1})\\
&+\frac{1}{n}\sum_{i=1}^n\mathbb{E}_{y_i|x_i,z_i,\Theta^m,b^m}\left[\log P(y_i|x_i,z_i,\Theta^m,b^m)\right]\\
\le&-\frac{1}{n}\log L_f^{\mathrm{MN}}(\Theta^m,b^m;\{(x_i,\widehat{y}_i(\Theta^m,b^m),z_i)\}_{i=1}^n)+P_\lambda^{\mathrm{MN}}(\Theta^m)\\
&+\frac{1}{n}\sum_{i=1}^n\mathbb{E}_{y_i|x_i,z_i,\Theta^m,b^m}\left[\log P(y_i|x_i,z_i,\Theta^m,b^m)\right]\\
=&\mathcal{F}_n^{\mathrm{MN}}(\Theta^m,b^m),
\end{aligned}
$$

其中，第二个不等式源于算法 1 中的 M步。如果 $\mathcal{F}_n^{\mathrm{MN}}(\Theta^{m+1}, b^{m+1}) = \mathcal{F}_n^{\mathrm{MN}}(\Theta^m, b^m)$，那么 (Θ^m, b^m) 必须是以下表达式的极小值点

$$-\frac{1}{n}\log L_f^{\mathrm{MN}}(\Theta, b; \{(x_i, \widehat{y}_i(\Theta^m, b^m), z_i)\}_{i=1}^n) + P_\lambda^{\mathrm{MN}}(\Theta)$$

关于 $\mathbb{R}^{p\times K}\times\mathbb{R}^K$。由于

$$\begin{aligned}&\left(\nabla\mathcal{L}^{\mathrm{MN}}(\Theta, b; \{(x_i, z_i)\}_{i=1}^n)\right)|_{\Theta^m, b^m}\\ =&\left(-\frac{1}{n}\nabla\log L_f^{\mathrm{MN}}(\Theta, b; \{(x_i, \widehat{y}_i(\Theta^m, b^m), z_i)\}_{i=1}^n)\right)|_{\Theta^m, b^m},\end{aligned}$$

$\mathcal{F}_n^{\mathrm{MN}}(\Theta^{m+1}, b^{m+1}) = \mathcal{F}_n^{\mathrm{MN}}(\Theta^m, b^m)$ 意味着 $(\Theta^m, b^m) \in \Gamma^{\mathrm{MN}}$。使用相同的论证可以证明，如果 $\{\theta^m\}_m$ 由算法2生成，则 $\{\mathcal{F}_n^{\mathrm{ON}}(\theta^m)\}_m$ 也是非递增的；如果 $\theta^m \in \Gamma^{\mathrm{ON}}$，则 $\mathcal{F}_n^{\mathrm{ON}}(\theta^{m+1}) < \mathcal{F}_n^{\mathrm{ON}}(\theta^m)$。另一方面，如果 $\{\Theta^m, b^m\}_m$（分别为 $\{\theta^m\}_m$）通过近端梯度下降算法生成，且每次更新分别采用 (6)（分别为 (8)），那么已有结果 [Beck, 2017，参见引理 10。4] 表明，在适当选择步长 η_m 后，$\{\mathcal{F}_n^{\mathrm{MN}}(\Theta^m, b^m)\}$（分别为 $\{\mathcal{F}_n^{\mathrm{ON}}(\theta^m)\}$）是非递增的。如果 $\mathcal{F}_n^{\mathrm{MN}}\Theta^{m+1}(\, b^{m+1}, = \mathcal{F}_n^{\mathrm{MN}})\, \Theta^m(\, b^m, \Theta^m)$，则 $(\, b^m,)$ 必须是

$$\eta_m P_\lambda^{\mathrm{MN}}(\Theta) + \frac{1}{2}\|[\Theta; b] - ([\Theta^m; b^m] - \eta_m\nabla\mathcal{L}_n^{\mathrm{MN}}(\Theta^m, b^m))\|_2^2.$$

由于上述损失函数在 (Θ^m, b^m) 处求得的任何次梯度都满足

$$\begin{aligned}&\left(\nabla\eta_m P_\lambda^{\mathrm{MN}}(\Theta) + \frac{1}{2}\|[\Theta; b] - ([\Theta^m; b^m] - \eta_m\nabla\mathcal{L}_n^{\mathrm{MN}}(\Theta^m, b^m))\|_2^2\right)|_{\Theta^m, b^m}\\ =&\eta_m(\nabla P_\lambda^{\mathrm{MN}}(\Theta^m) + \nabla\mathcal{L}_n^{\mathrm{MN}}(\Theta^m, b^m)))\\ =&\eta_m(\nabla\mathcal{F}_n^{\mathrm{MN}}(\Theta, b))|_{\Theta^m, b^m},\end{aligned}$$

我们还有 $\mathcal{F}_n^{\mathrm{MN}}(\Theta^{m+1},\ b^{m+1}) < \mathcal{F}_n^{\mathrm{MN}}(\Theta^m,\ b^m)$，如果 $(\Theta^m,\ b^m) \notin \Gamma^{\mathrm{MN}}$。类似地，如果 $\theta^m \notin \Gamma^{\mathrm{ON}}$ 是通过近端梯度下降算法 (8) 生成的，我们有 $\mathcal{F}_n^{\mathrm{ON}}(\theta^{m+1}) < \mathcal{F}_n^{\mathrm{ON}}(\theta^m)$。

现在我们将 已知 $\lim_{\|\Theta\|_2+\|b\|_2\to+\infty}\mathcal{F}_n^{\mathrm{MN}}(\Theta,\ b) = +\infty$ 且 $\lim_{\|\theta\|_2\to+\infty}\mathcal{F}_n^{\mathrm{ON}}(\theta) = +\infty$，因此 e

$$\mathcal{F}_n^{\mathrm{MN}}(\Theta, b) \le \arg\min_b \mathcal{F}_n^{\mathrm{MN}}(0_{p\times K}, b), \quad \mathcal{F}_n^{\mathrm{ON}}(\theta) \le \arg\min_{\theta_{1:p}=0_{p\times 1}} \mathcal{F}_n^{\mathrm{ON}}(\theta)$$

这意味着对于某些 $R^{\mathrm{MN}} > 0$、$R^{\mathrm{ON}} > 0$，$\|\Theta\|_2 + \|b\|_2 \le R^{\mathrm{MN}}$ 和 $\|\theta\|_2 \le R^{\mathrm{ON}}$ 成立。根据我们的近端梯度下降算法和 EM算法（算法 1、2）的初始化条件，以及 $\{\mathcal{F}_n^{\mathrm{MN}}(\Theta^m, b^m)\}$ 和 $\{\mathcal{F}_n^{\mathrm{ON}}(\theta^m)\}$的非递增性质，可以保证所有参数迭代序列 $\{(\Theta^m, b^m)\}_m$都属于集合 $\mathcal{C}^{\mathrm{MN}} = \{[\Theta; b]\colon \|\Theta\|_2 + \|b\|_2 \le R^{\mathrm{MN}}\}$，且 $\{\theta^m\}_m$ 属于 $\mathcal{C}^{\mathrm{ON}} = \{\theta\colon \|\theta\|_2 \le R^{\mathrm{ON}}\}$。$\mathcal{C}^{\mathrm{MN}}$ 和 $\mathcal{C}^{\mathrm{ON}}$ 均为紧集，这意味着引理 3.1 和引理 D.1 中的 (i) 成立。

由于我们的损失函数 $\mathcal{F}_n^{\mathrm{MN}}$、$\mathcal{F}^{\mathrm{ON}}$ 连续，且所有参数迭代序列都属于紧集 $\mathcal{C}^{\mathrm{MN}}$、$\mathcal{C}^{\mathrm{ON}}$，因此我们可以应用 Zangwill [1969] 中的全局收敛定理，得到命题 3.1和命题 D.1中的 (ii) 和 (iii)。

□

E 病例对照设定下对数似然函数的推导

首先，我们推导病例对照设置下给定 x、y 和 $s = 1$ 时 z 的条件分布，该分布已在第 1.1 节中介绍。接着，我们将给出引理 2.1 和引理 2.2 的证明。

首先注意，在病例对照设置下， n_u 样本是从总体中随机抽取的，而 n_k 样本则是从标签为 k 的总体中抽取的，其中 $1 \le k \le K$。因此，我们有

$$\frac{\mathbb{P}(z=k|s=1)}{\mathbb{P}(z=0|s=1)} = \frac{n_k}{n_u},$$

这意味着

$$\frac{\mathbb{P}(z=k, y=k|s=1)}{\mathbb{P}(z=0, y=k|s=1)} = \frac{\mathbb{P}(z=k|s=1)}{\mathbb{P}(z=0|s=1)\mathbb{P}(y=k)} = \frac{n_k}{\pi_k n_u}.$$

因此，

$$\mathbb{P}(z=k|y=k, s=1) = \frac{\mathbb{P}(z=k, s=1|y=k)}{\mathbb{P}(s=1|y=k)} = \frac{n_k}{n_k + \pi_k n_u}, \quad \mathbb{P}(z=0|y=k, s=1) = \frac{\pi_k n_u}{n_k + \pi_k n_u}.$$

引理 *2.1* 的证明。首先，我们推导给定 x 和 $s=1$时任意数据点 $(x,\ y,\ z,\ s=1)$ 的条件概率质量函数。注意，$(y,\ z)$ 有 $2K+1$ 种可能的组合： $y=z=0$、 $y=z=j$，以及$y=j,\ z=0$ （对于 $1 \le j \le K$）。随后可以证明，对于任意 $1 \le j \le K$，

$$\begin{aligned}
\mathbb{P}(y=j, z=j|x, s=1) =& \frac{\mathbb{P}(y=j|x)\mathbb{P}(z=j, s=1|y=j)}{\sum_{k=1}^K \mathbb{P}(y=k|x)\mathbb{P}(z=k, s=1|y=k) + \mathbb{P}(z=0, s=1|x)} \\
=& \frac{n_j/\pi_j e^{x^\top \Theta_j + b_j}}{n_u + \sum\limits_{k=1}^K (n_k/\pi_k + n_u) e^{x^\top \Theta_k + b_k}},
\end{aligned} \tag{132}$$

其中我们用到了以下事实： $\mathbb{P}(y=j|x) = \frac{e^{x^\top \Theta_j + b_j}}{1 + \sum\limits_{k=1}^K e^{x^\top \Theta_k + b_k}}$ 和

$$\frac{\mathbb{P}(z=j, s=1|y=j)}{\mathbb{P}(z=k, s=1|y=k)} = \frac{n_j/\pi_j}{n_k/\pi_k}, \quad \frac{\mathbb{P}(z=j, s=1|y=j)}{\mathbb{P}(z=0, s=1|x)} = \frac{n_j/\pi_j}{n_u}.$$

应用上述事实还可得

$$\begin{aligned}
\mathbb{P}(y=j, z=0|x, s=1) =& \frac{\mathbb{P}(y=j|x)\mathbb{P}(z=0, s=1|y=j)}{\sum_{k=1}^K \mathbb{P}(y=k|x)\mathbb{P}(z=k, s=1|y=k) + \mathbb{P}(z=0, s=1|x)} \\
=& \frac{n_u e^{x^\top \Theta_j + b_j}}{n_u + \sum\limits_{k=1}^K (n_k/\pi_k + n_u) e^{x^\top \Theta_k + b_k}}, \\
\mathbb{P}(y=0, z=0|x, s=1) =& \frac{\mathbb{P}(y=0|x)\mathbb{P}(z=0, s=1|y=0))}{\sum_{k=1}^K \mathbb{P}(y=k|x)\mathbb{P}(z=k, s=1|y=k) + \mathbb{P}(z=0, s=1|x)} \\
=& \frac{n_u}{n_u + \sum\limits_{k=1}^K (n_k/\pi_k + n_u) e^{x^\top \Theta_k + b_k}}.
\end{aligned}$$

因此， $\{x_i, y_i, z_i, s_i=1\}_{i=1}^n$ 的完整对数似然函数如下:

$$\begin{aligned}
&\log L_f^{\mathrm{MN}}(\Theta, b; \{x_i, y_i, z_i, s_i=1\}_{i=1}^n) \\
=& \sum_{i=1}^n \left[\sum_{k=1}^K \mathbb{1}_{\{y_i=k\}} (x_i^\top \Theta_k + b_k) - \log\left(1 + \sum_{k=1}^K (1 + \frac{n_k}{\pi_k n_u}) e^{x_i^\top \Theta_k + b_k}\right) \right] \\
&+ \sum_{i=1}^n \sum_{k=1}^K \mathbb{1}_{\{y_i=z_i=k\}} \log \frac{n_k}{\pi_k n_u}.
\end{aligned} \tag{133}$$

为展示观测数据 $\{x_i, z_i, s_i = 1\}_{i=1}^n$ 的对数似然函数，请注意，给定 x， $s = 1$ 时， z 的条件概率质量函数如下:

$$\begin{aligned}
\mathbb{P}(z = j|x, s = 1) =& \mathbb{P}(y = j, z = j|x, s = 1) \\
=& \frac{\frac{n_j}{\pi_j n_u} e^{x^\top \Theta_j + b_j}}{1 + \sum\limits_{k=1}^K (1 + \frac{n_k}{\pi_k n_u}) e^{x^\top \Theta_k + b_k}}, \quad j > 0, \\
\mathbb{P}(z = 0|x, s = 1) =& \sum_{j=1}^K \mathbb{P}(y = j, z = 0|x, s = 1) + \mathbb{P}(y = 0, z = 0|x, s = 1) \\
=& \frac{1 + \sum_{j=1}^K e^{x^\top \Theta_j + b_j}}{1 + \sum\limits_{k=1}^K (1 + \frac{n_k}{\pi_k n_u}) e^{x^\top \Theta_k + b_k}}.
\end{aligned} \tag{134}$$

因此，观测数据 $\{x_i, z_i, s_i = 1\}_{i=1}^n$ 的对数似然函数为

$$\begin{aligned}
& \log L^{\mathrm{MN}}(\Theta, b; \{x_i, z_i, s_i = 1\}_{i=1}^n) \\
=& \sum_{i=1}^n \left[\sum_{j=1}^K \mathbb{1}_{\{z_i = j\}} (x_i^\top \Theta_j + b_j + \log(\frac{n_j}{\pi_j n_u})) + \mathbb{1}_{\{z_i = 0\}} \log(1 + \sum_{j=1}^K e^{x_i^\top \theta_j + b_j}) \right. \\
& \left. - \log \left(1 + \sum_{k=1}^K (1 + \frac{n_k}{\pi_k n_u}) e^{x_i^\top \Theta_k + b_k} \right) \right] \\
=& \sum_{i=1}^n \left[\sum_{j=1}^K \mathbb{1}_{\{z_i = j\}} (x_i^\top \Theta_j + b_j + \log \frac{n_j}{\pi_j n_u} - \log(1 + \sum_{j=1}^K e^{x_i^\top \Theta_j + b_j})) \right. \\
& \left. - \log \left(1 + \frac{\sum\limits_{k=1}^K \frac{n_k}{\pi_k n_u} e^{x_i^\top \Theta_k + b_k}}{1 + \sum_{k=1}^K e^{x_i^\top \Theta_k + b_k}} \right) \right].
\end{aligned} \tag{135}$$

□

引理证明 *2.2.* 同样，根据前述设定，我们可以推导出在给定 x a 的情况下 y， z 的分布 s

$$\begin{aligned}
\mathbb{P}(y = j, z = j|x, s = 1) =& \frac{\mathbb{P}(y = j, z = j, s = 1|x)}{\mathbb{P}(s = 1|x)} \\
=& \frac{\frac{n_j}{\pi_j n_u} r_j(x, \theta)}{1 + \sum\limits_{k=1}^K (\frac{n_k + \pi_k n_u}{\pi_k n_u}) r_k(x, \Theta)}, \quad j > 0, \\
\mathbb{P}(y = j, z = 0|x, s = 1) =& \frac{r_j(x, \theta)}{1 + \sum\limits_{k=1}^K (\frac{n_k + \pi_k n_u}{\pi_k n_u}) r_k(x, \theta)}, \\
\mathbb{P}(y = 0, z = 0|x, s = 1) =& \frac{1}{1 + \sum\limits_{k=1}^K (\frac{n_k + \pi_k n_u}{\pi_k n_u}) r_k(x, \theta))},
\end{aligned} \tag{136}$$

其中

$$\begin{aligned}
r_j(x, \theta) =& \frac{\mathbb{P}_\theta(y = j|x)}{\mathbb{P}_\theta(y = 0|x)} \\
=& (1 + e^{x^\top \theta_{1:p} - \theta_{p+1}}) \left[(1 + e^{x^\top \theta_{1:p} - \sum_{l=1}^{j+1} \theta_l})^{-1} - (1 + e^{x^\top \theta_{1:p} - \sum_{l=1}^{j} \theta_l})^{-1} \right].
\end{aligned} \tag{137}$$

另一方面，给定 x 时， z 的分布为

$$\begin{aligned}\mathbb{P}(z=j|x,s=1)=&\frac{\frac{n_j}{\pi_j n_u}r_j(x,\theta)}{1+\sum\limits_{k=1}^{K}(\frac{n_k+\pi_k n_u}{\pi_k n_u})r_k(x,\theta)},\quad j>0,\\ \mathbb{P}(z=0|x,s=1)=&\frac{1+\sum_{k=1}^{K}r_k(x,\theta)}{1+\sum\limits_{k=1}^{K}(\frac{n_k+\pi_k n_u}{\pi_k n_u})r_k(x,\theta)}.\end{aligned}\tag{138}$$

因此， $\{(x_i,\ y_i,\ z_i,\ s_i=1)\}_{i=1}^n$ 的完整对数似然函数如下：

$$\begin{aligned}&\log L_f^{\mathrm{ON}}(\theta;\{(x_i,y_i,z_i,s_i=1)\}_{i=1}^n)\\ =&\sum_{i=1}^{n}\left[\sum_{j=1}^{K}\mathbb{1}_{\{y_i=j\}}\log r_j(x_i,\theta)-\log\left(1+\sum_{k=1}^{K}\frac{n_k+\pi_k n_u}{\pi_k n_u}r_k(x_i,\theta)\right)\right]\\ &+\sum_{i=1}^{n}\sum_{j=1}^{K}\mathbb{1}_{\{y_i=z_i=j\}}\log\frac{n_j}{\pi_j n_u},\end{aligned}\tag{139}$$

观测数据 $\{(x_i,\ z_i,\ s_i=1)\}_{i=1}^n$ 的对数似然为

$$\begin{aligned}&\log L^{\mathrm{ON}}(\theta;\{(x_i,z_i,s_i=1)\}_{i=1}^n)\\ =&\sum_{i=1}^{n}\left[\sum_{j=1}^{K}\mathbb{1}_{\{z_i=j\}}(\log r_j(x_i,\theta)-\log(\frac{n_j}{\pi_j n_u}))+\mathbb{1}_{\{z_i=0\}}\log(1+\sum_{j=1}^{K}r_j(x_i,\theta))\right.\\ &\left.-\log\left(1+\sum_{k=1}^{K}\frac{n_k+\pi_k n_u}{\pi_k n_u}r_k(x_i,\theta)\right)\right]\\ =&\sum_{i=1}^{n}\left[\sum_{k=1}^{K}\mathbb{1}_{\{z_i=k\}}[f(\log r(x_i,\theta))]_k-\log\left(1+\sum_{j=1}^{K}e^{[f(\log r(x_i,\theta))]_k}\right)\right],\end{aligned}$$

其中 $f\colon\ \mathbb{R}^K\to\mathbb{R}^K$ 满足 $(f^{\mathrm{ON}}(u))_k=u_k+\log\frac{n_k}{\pi_k n_u}-\log(1+\sum_{j=1}^{K}e^{u_j})$。 □

F 单训练集情形下 PU 模型的算法

这里我们考虑单训练集情形 [Elkan 和 Noto， 2008], 这与 PU 模型的病例对照情形不同。具体而言，我们仍然假设真实标记数据的多项式分布和有序分布（x_i， y_i），同时令观测 z_i 以某个常数概率处于未标记状态。形式上，对于任意 j， $k\geq 0$,

$$\mathbb{P}(z_i=k|x_i,y_i=j)=\begin{cases}\pi_j^{\mathrm{st}}, & \text{if } j=k>0,\\ 1-\pi_j^{\mathrm{st}}, & \text{if } j>0,\ k=0,\\ 1, & \text{if } j=k=0,\\ 0, & \text{otherwise,}\end{cases}\tag{140}$$

其中， $\pi^{\mathrm{st}}\in(0,1)^K$ 表示正样本被标记的概率，上标 st 表示单训练集情形。

F.1 多项式 PU 模型

下面的引理给出了多项式 PU 模型和单训练集情形下观测数据 $\{(x_i, z_i)\}$ 以及完整数据 $\{(x_i, y_i, z_i)\}$ 的对数似然函数。

引理 F.1。 当观测响应 z_i 根据 (140) 生成时，多项式模型下观测到的仅存在数据 $\{x_i, z_i\}_{i=1}^n$ 的对数似然函数为

$$\begin{aligned}&\log L^{\mathrm{MN,st}}(\Theta, b; \{(x_i, z_i)\}_{i=1}^n)\\ =&\sum_{i=1}^n \left[\sum_{k=1}^K \mathbb{1}_{\{z_i=k\}}[\widetilde{f}(\Theta^\top x_i + b + \log(1-\pi^{st}))]_k - \log\left(1+\sum_{k=1}^K e^{[\widetilde{f}(\Theta^\top x_i + b + \log(1-\pi^{st}))]_k}\right)\right],\end{aligned} \tag{141}$$

其中 $\widetilde{f}\colon \mathbb{R}^K \to \mathbb{R}^K$ 满足 $(\widetilde{f}(u))_k = u_k +$ 对数 $\frac{\pi_k^{\mathrm{st}}}{1-\pi_k^{\mathrm{st}}}$ − 对数$(1+\sum_{j=1}^K e^{u_j})$。此处 对数$(1-\pi^{\mathrm{st}})$ =(对数$(1-\pi_1^{\mathrm{st}})$.....，对数$(1-\pi_K^{\mathrm{st}})$)$^\top \in \mathbb{R}^K$。完整数据 $\{x_i, y_i, z_i\}_{i=1}^n$ 的对数似然函数为

$$\begin{aligned}&\log L_f^{\mathrm{MN,st}}(\Theta, b; \{(x_i, y_i, z_i)\}_{i=1}^n)\\ =&\sum_{i=1}^n \left[\sum_{k=1}^K \mathbb{1}_{\{y_i=k\}}(x_i^\top \Theta_k + b_k + \log(1-\pi_k^{\mathrm{st}})) - \log\left(1+\sum_{k=1}^K e^{x_i^\top \Theta_k + b_k}\right)\right.\\ &\left.+\sum_{k=1}^K \mathbb{1}_{\{y_i=z_i=k\}} \log\frac{\pi_k^{\mathrm{st}}}{1-\pi_k^{\mathrm{st}}}\right],\end{aligned} \tag{142}$$

其中 Θ_k 是 Θ 的第 k 列。

证明。 在单训练集情形下的多项式 PU 模型中，可以证明，对于任意数据点（x， y， z）和 $1 \le j \le K$,

$$\begin{aligned}\mathbb{P}(y=j, z=j|x) =&\mathbb{P}(y=j|x)\mathbb{P}(z=j|y=j) = \frac{\pi_j^{\mathrm{st}} e^{x^\top \Theta_j + b_j}}{1+\sum_{k=1}^K e^{x^\top \Theta_k + b_k}},\\ \mathbb{P}(y=j, z=0|x) =&\mathbb{P}(y=j|x)\mathbb{P}(z=0|y=j) = \frac{(1-\pi_j^{\mathrm{st}}) e^{x^\top \Theta_j + b_j}}{1+\sum_{k=1}^K e^{x^\top \Theta_k + b_k}},\\ \mathbb{P}(y=0, z=0|x) =&\mathbb{P}(y=0|x) = \frac{1}{1+\sum_{k=1}^K e^{x^\top \Theta_k + b_k}}.\end{aligned}$$

而对于观测正例-未标记数据（x， z），我们有

$$\begin{aligned}\mathbb{P}(z=j|x) =&\mathbb{P}(y=j, z=j|y=j) = \frac{\pi_j^{\mathrm{st}} e^{x^\top \Theta_j + b_j}}{1+\sum_{k=1}^K e^{x^\top \Theta_k + b_k}},\\ \mathbb{P}(z=0|x) =&\sum_{j=1}^K \mathbb{P}(y=j, z=0|y=j) + \mathbb{P}(y=0, z=0|x) = \frac{1+\sum_{j=1}^K (1-\pi_j^{\mathrm{st}}) e^{x^\top \Theta_j + b_j}}{1+\sum_{k=1}^K e^{x^\top \Theta_k + b_k}}.\end{aligned}$$

因此，样本 $\{(x_i,\ y_i,\ z_i)\}_{i=1}^n$ 的完整对数似然可写为:

$$\begin{aligned}&\log L_f^{\mathrm{MN,st}}(\Theta, b; \{(x_i, y_i, z_i)\}_{i=1}^n)\\ =&\sum_{i=1}^n \left[\sum_{j=1}^K \mathbb{1}_{\{y_i=j\}}(x_i^\top \Theta_j + b_j + \log(1-\pi_j^{\mathrm{st}})) - \log(1+\sum_{k=1}^K e^{x_i^\top \Theta_k + b_k})\right]\\ &+\mathbb{1}_{\{z_i=j\}} \log\frac{\pi_j^{\mathrm{st}}}{1-\pi_j^{\mathrm{st}}}.\end{aligned}$$

而观测正例-未标记数据 $\{(x_i,\ z_i)\}_{i=1}^n$ 的对数似然为

$$
\begin{aligned}
&\log L^{\mathrm{MN,st}}(\Theta, b; \{(x_i, z_i)\}_{i=1}^n) \\
=&\sum_{i=1}^n \left[\sum_{j=1}^K \mathbb{1}_{\{z_i=j\}}(x_i^\top \Theta_j + b_j + \log \pi_j^{\mathrm{st}} - \log(1 + \sum_{k=1}^K (1-\pi_k^{\mathrm{st}}) e^{x_i^\top \Theta_k + b_k}))) \right. \\
&\left. - \log(1 + \frac{\sum_{k=1}^K \pi_k^{\mathrm{st}} e^{x_i^\top \Theta_k + b_k}}{1 + \sum_{k=1}^K (1-\pi_k^{\mathrm{st}}) e^{x_i^\top \Theta_k + b_k}}) \right] \\
=&\sum_{i=1}^n \left[\sum_{k=1}^K \mathbb{1}_{\{z_i=k\}} [\widetilde{f}(\Theta^\top x_i + b + \log(1-\pi^{st}))]_k - \log\left(1 + \sum_{j=1}^K e^{[\widetilde{f}(\Theta^\top x_i + b + \log(1-\pi^{st}))]_k}\right) \right],
\end{aligned}
$$

□

与病例-对照情形（见第 2 节）下的估计类似，可以对

$$
\mathcal{F}_n^{\mathrm{MN,st}}(\Theta, b) = -\frac{1}{n} \log L^{\mathrm{MN,st}}(\Theta, b; \{(x_i, z_i)\}_{i=1}^n) + P_\lambda^{\mathrm{MN}}(\Theta)
$$

进行近端梯度下降算法以估计 Θ 和 b。在此设定下，我们也可以推导出 EM算法（见算法3）。

算法 hm 3:单训练集情形下多项式 PU 模型的期望最大化

输入： Θ^0、 b^0，满足 $\mathcal{F}^{\mathrm{MN}}{}_{,n}{}^{\mathrm{st}}(\Theta^0, b^0) \leq$最小值$_b\, \mathcal{F}^{\mathrm{MN}}{}_{,n}{}^{\mathrm{st}}(0_{p\times K}, b)$

针对 $m = 0, 1,$ .....**执行**

E步：按如下方式计算 $\widehat{y}(\Theta^m,\ b^m) \in \mathbb{R}^{n\times K}$，其中每个条目 $\widehat{y}_{ij}$ 都是 $\mathbb{1}_{\{y_i=j\}}$ 的估计值:

$$
\hat{y}_{ij}(\Theta^m, b^m) = \begin{cases} \frac{e^{x_i^\top \Theta_j^m + b_j^m + \log(1-\pi_j^{\mathrm{st}})}}{1 + \sum\limits_{k=1}^K e^{x_i^\top \Theta_k^m + b_k^m + \log(1-\pi_k^{\mathrm{st}})}} & \text{if } z_i = 0 \\ 1 & \text{if } z_i = j \\ 0 & \text{else} \end{cases}
$$

M步：通过以下方式获得 $\Theta^{m+1},\ b^{m+1}$

$$
\begin{aligned}
(\Theta^{m+1}, b^{m+1}) \in \arg\min_{\Theta, b} \frac{1}{n} \sum_{i=1}^n \left[\log\left(1 + \sum_{k=1}^K e^{x_i^\top \Theta_k + b_k}\right) \right. \\
\left. - \sum_{j=1}^K \hat{y}_{ij}(\Theta^m, b^m)(x_i^\top \Theta_j + b_j) \right] + P_\lambda^{\mathrm{MN}}(\Theta)
\end{aligned}
$$

结束循环

理论性质： 一个关键观察是，在单训练集情形下，多项式 PU 模型可以视为对病例对照设置的一种简单重新参数化。更具体地说，要获得单训练集情形下的对数似然函数，只需将 $\log L_f^{\mathrm{MN}}$ 和 $\log\ L^{\mathrm{MN}}$ 中的 $\frac{n_k}{\pi_k n_u}$ 替换为 $\frac{\pi_k^{\mathrm{st}}}{1-\pi_k^{\mathrm{st}}}$，并将 b_k 改为 $1 \leq k \leq K$ 对应的 $\log(1-\pi_k^{\mathrm{st}})\ b_k +$。因此，只要将关于 $\frac{n_k}{\pi_k^{\mathrm{st}} n_u}$ 的条件改为关于 $\frac{\pi_k^{\mathrm{st}}}{1-\pi_k^{\mathrm{st}}}$ 的条件，命题 3.1 和定理 3.1 中给出的所有优化结果和统计保证仍适用于这种设置。

F.2 有序-PU 模型

下面的引理给出了序数 PU 模型和单训练集情形下观测数据 $\{(x_i,\ z_i)\}$ 以及完整数据 $\{(x_i,\ y_i,\ z_i)\}$ 的对数似然函数。

引理 F.2. 当观测响应 z_i 按照(140)生成时，有序 *PU* 模型下观测到的仅存在数据 $\{x_i, z_i\}_{i=1}^n$ 的对数似然函数为

$$\begin{aligned}&\log L^{\mathrm{ON,st}}(\theta;\{(x_i,z_i)\}_{i=1}^n)\\=&\sum_{i=1}^n\left[\sum_{k=1}^K\mathbb{1}_{\{z_i=k\}}[\widetilde{f}(\log r(x_i,\theta)+\log(1-\pi^{\mathrm{st}}))]_k-\log\left(1+\sum_{k=1}^K e^{[\widetilde{f}(\log r(x_i,\theta)+\log(1-\pi^{\mathrm{st}}))]_k}\right)\right],\end{aligned}\tag{143}$$

其中，$\widetilde{f}$ 在引理 *F.1* 中定义，$r:\ \mathbb{R}^p\times\mathbb{R}^{p+K}\to\mathbb{R}^K$ 在引理 *2.2* 中定义。完整数据 $\{x_i,y_i,z_i\}_{i=1}^n$ 的对数似然函数为

$$\begin{aligned}&\log L_f^{\mathrm{ON,st}}(\theta;\{(x_i,y_i,z_i)\}_{i=1}^n)\\=&\sum_{i=1}^n\left[\sum_{k=1}^K\mathbb{1}_{\{y_i=k\}}(\log r_k(x_i,\theta)+\log(1-\pi_k^{\mathrm{st}})-\log\left(1+\sum_{k=1}^K r_k(x_i,\theta)\right)\right.\\&\left.+\sum_{k=1}^K\mathbb{1}_{\{y_i=z_i=k\}}\log\frac{\pi_k^{\mathrm{st}}}{1-\pi_k^{\mathrm{st}}}\right].\end{aligned}\tag{144}$$

证明. 与引理 F.1 的证明类似，我们首先写出 (x,y,z) 的联合概率质量，如下所示:

$$\begin{aligned}\mathbb{P}(y=j,z=j|x)=&\frac{\pi_j^{\mathrm{st}}r_j(x,\theta)}{1+\sum_{k=1}^K r_k(x,\theta)},\\\mathbb{P}(y=j,z=0|x)=&\frac{(1-\pi_j^{\mathrm{st}})r_j(x,\theta)}{1+\sum_{k=1}^K r_k(x,\theta)},\\\mathbb{P}(y=0,z=0|x)=&\frac{1}{1+\sum_{k=1}^K r_k(x,\theta)}.\end{aligned}$$

于是，(x,z) 的联合概率质量变为:

$$\begin{aligned}\mathbb{P}(z=j|x)=&\frac{\pi_j^{\mathrm{st}}r_j(x,\theta)}{1+\sum_{k=1}^K r_k(x,\theta)},\\\mathbb{P}(z=0|x)=&\frac{1+\sum_{k=1}^K(1-\pi_k^{\mathrm{st}})r_k(x,\theta)}{1+\sum_{k=1}^K r_k(x,\theta)}.\end{aligned}$$

Th因此，可以直接写出对数似然函数，如引理 F.2 所示 . □

Let

$$\mathcal{F}_n^{\mathrm{ON,st}}(\theta)=-\frac{1}{n}\log L^{\mathrm{ON,st}}(\theta;\{(x_i,z_i)\}_{i=1}^n)+P_\lambda^{\mathrm{ON}}(\theta_{1:p}),$$

然后，我们可以通过对 $\mathcal{F}^{\mathrm{ON},}{}_n{}^{\mathrm{st}}(\theta)$ 应用近端梯度下降算法来估计 θ^*。正则化的

此设置下的EM算法总结见算法4。

算法4： 单训练集情形下序数-PU模型的EM算法

输入： θ^0 使得 $\mathcal{F}^{\mathrm{ON},\ n}{}^{\mathrm{st}}(\theta^0)$ $\leq$最小值$_{\theta_{1:p}=0_{p\times 1}}\mathcal{F}^{\mathrm{ON},\ n}{}^{\mathrm{st}}(\theta)$

当 $m = 0$**时**, $1, \cdots$ **执行**

E步：计算 $\widehat{y}(\theta^m) \in \mathbb{R}^{n\times K}$ 如下，其中每个条目 $[\widehat{y}(\theta^m)]_{ij}$ 是对 $\mathbb{1}_{\{y_i=j\}}$ 的估计：

$$\hat{y}_{ij}(\theta^m) = \begin{cases} \frac{(1-\pi_j^{\mathrm{st}})r_j(x_i,\theta^m)}{1+\sum_{k=1}^K(1-\pi_k^{\mathrm{st}})r_k(x_i,\theta^m)}, & \text{if } z_i = 0, \\ 1, & \text{if } z_i = j, \\ 0, & \text{else.} \end{cases}$$

M步：通过 θ^{m+1} 获得

$$\theta_j^{m+1} \in \arg\min_{\theta} \frac{1}{n}\sum_{i=1}^n \left[-\sum_{j=1}^K \widehat{y}_{ij}(\theta^m)\log r_j(x_i,\theta) + \log\left(1+\sum_{k=1}^K r_k(x_i,\theta)\right)\right] + P_\lambda^{\mathrm{ON}}(\theta_{1:p}).$$

结束循环

理论性质： 使用与命题 D.1 的证明类似的论证，我们仍然可以证明这两种算法收敛到优化问题 $\min_{\|\theta\|_2\leq C}\mathcal{F}^{\mathrm{ON},\ n}{}^{\mathrm{st}}(\theta)$ 的驻点，其中 $C > 0$ 是某个常数。尽管在这种设定下，我们尚无关于驻点的严格统计误差界，但我们推测它们具有与定理 3.2 类似的统计性质。对这一猜想的严格证明留待未来工作。

G 数值细节

控制不同类别的真实患病率： 在我们的比较模拟研究中，我们通过适当设置截距，以近似方式控制不同类别的患病率。具体而言，对于研究患病率影响的实验，我们选择截距，以确保仅截距模型的患病率为 $\pi_0 \in \{0.1, 0.2, \ldots, 0.7\}$ 和 $\pi_1 = \pi_2 = \frac{1-\pi_0}{2}$。由于所有特征 X_j 均由均值为零的高斯分布生成，因此我们预期完整模型的患病率与仅截距模型的患病率不会有太大差异。随后，我们根据生成的数据估计完整模型的患病率，并在图中展示。对于研究抽样比例 $\frac{n_u}{n}$影响的实验，我们则确保仅截距模型中未标记数据占 30%，标签 1 和 2 的样本各占 35%。由此得到的完整多项式模型的患病率为$\pi_0 = 0.2584$， $\pi_1 = 0.3677$， $\pi_2 = 0.3739$，而完整有序模型的患病率为 $\pi_0 = 0.3753$， $\pi_1 = 0.2153$， $\pi_2 =$0.4095。因此，总体的类别分布较为均衡。

—

致谢

我们谨此感谢 David Neiman 开展了一些初步实证研究并探索真实数据。GR 和 LZ 的部分工作得到了 NSF-DMS 1811767、NIH R01 GM131381-01 的支持。GR 还部分

由 NSF-DMS 1839338 资助。

参考文献

Alan Agresti。《有序分类数据分析》，第 656 卷。John Wiley & Sons，2010 年。

Kellie J Archer、Jiayi Hou、Qing Zhou、Kyle Ferber、John G Layne 和 Amanda E Gentry。ordinalgmifs：一种用于高维数据场景下有序回归的 R 软件包。癌症信息学，13：CIN–S20806，2014。

Amir Beck。优化中的一阶方法。SIAM，2017。

Loes CM Bertens、Karel GM Moons、Frans H Rutten、Yvonne van Mourik、Arno W Hoes 和 Johannes B Reitsma。开发了一种列线图，以促进多项逻辑回归建模在诊断研究中的应用。临床流行病学杂志，71：51–57，2016。

Marthinus Du Plessis、Gang Niu 和 Masashi Sugiyama。用于从正类与未标记数据中学习的凸形式化。载于国际机器学习会议，第 1386–1394 页。PMLR，2015年。

Dheeru Dua 和 Casey Graff。UCI 机器学习仓库，2017年。URL http://archive.ics.uci.edu/ ml。

Charles Elkan 和 Keith Noto。仅使用正类与未标记数据学习分类器。载于第*14*届*ACM SIGKDD*知识发现与数据挖掘国际会议论文集，第213–220页，2008年。

Jianqing Fan、Rui Song 等。具有 np 维度的广义线性模型中的确定性独立筛选。*The Annals of Statistics*，38(6)：3567–3604，2010年。

Kelwin Fernandes、Jaime S Cardoso 和 Jessica Fernandes。应用于宫颈癌筛查的部分可观测性迁移学习。载于伊比利亚模式识别与图像分析会议，第243–250页。Springer，2017年。

Peter A Fields。极端温度下的蛋白质功能：平衡稳定性与柔性。比较生物化学与生理学*A*辑：分子与整合生理学，129(2-3)：417–431，2001年。

Jerome Friedman、Trevor Hastie 和 Robert Tibshirani。关于组套索与稀疏组套索的一则说明。*arXiv* 预印本 *arXiv:1001.0736*，2010年。

Sham Kakade、Ohad Shamir、Karthik Sindharan 和 Ambuj Tewari。高维度下的指数族学习：强凸性与稀疏性。载于第十三届人工智能与统计国际会议论文集，第381–388页。JMLR研讨会与会议论文集，2010年。

Balaji Krishnapuram、Lawrence Carin、M´ario AT Figueiredo 和 Alexander J Hartemink。稀疏多项逻辑回归：快速算法与泛化界。*IEEE*模式分析与机器智能汇刊，27(6):957–968，2005年。

TonyLancaster 和 Guido Imbens。受污染对照组的病例对照研究。计量经济学杂志，71(1-2)：145–160，1996年。

Chanwoo Lee 和 Miaoyan Wang。基于序数观测的张量去噪与补全。载于国际机器学习会议，第5778–5788页。PMLR，2020 年。Yuan Li、Benjamin Mark、Garvesh Raskutti、Rebecca Willett、Hyebin Song 和 David Neiman。具有对齐和高度相关设计的回归问题的基于图的正则化。*SIAM*数据科学数学期刊，2(2)：480–504，2020 年。Roderick JA Little 和 Donald B Rubin。缺失数据的统计分析，第793卷。John Wiley & Sons，2019。Bing Liu、Yang Dai、Xiaoli Li、Wee Sun Lee 和 Philip S Yu。使用正类与未标记样本构建文本分类器。载于第三届*IEEE*国际数据挖掘会议，第179–186页。IEEE，2003年。Po-Ling Loh 和 Martin J Wainwright。具有非凸性的正则化m估计量：局部最优解的统计与算法理论。*arXiv*预印本 *arXiv:1305.2436*，2013年。Po-Ling Loh、Martin J Wainwright 等。无需不相干性的支持恢复：非凸正则化的一个案例。统计学年鉴，45(6)：2455–2482，2017年。Andreas Maurer。Rademacher 复杂度的向量压缩不等式。载于国际算法学习理论会议，第3–17页。Springer，2016。Peter McCullagh。序数数据的回归模型。皇家统计学会杂志：*B*辑（方法论），42(2)：109–127，1980年。Colin McDiarmid。论有界差分方法。组合学综述，141(1)：148–188，1989。Sahand Negahban、Bin Yu、Martin J Wainwright 和 Pradeep K Ravikumar。具有可分解正则项的 m估计量高维分析统一框架。载于神经信息处理系统进展，第1348–1356页，2009年。Philip A Romero、Tuan M Tran 和 Adam R Abate。利用基于微流控的深度突变扫描解析酶功能。美国国家科学院院刊，112(23)：7159–7164，2015年。TR Sivapriya、AR Kamal 和 P Thangaiah。用于增强阿尔茨海默病痴呆多项分类的集成优点合并特征选择。医学计算与数学方法，2015，2015。Hyebin Song 和 Garvesh Raskutti。Pulasso：利用仅存在数据进行高维变量选择。美国统计学会杂志，115(529)：334–347，2020 年。Hyebin Song、Ran Dai、Garvesh Raskutti 和 Rina Foygel Barber。使用类条件噪声标签进行统计推断的凸方法与非凸方法。机器学习研究杂志，21(168)：1–58，2020 年。Hyebin Song、Bennett J Bremer、Emily C Hinds、Garvesh Raskutti 和 Philip A Romero。利用大规模正例-未标记学习推断蛋白质序列-功能关系。*Cell systems*，12(1)：92–101，2021。

Robert Tibshirani。通过套索进行回归收缩与选择。皇家统计学会杂志：*B*辑（方法论），58(1):267–288，1996年。AW van der Vaart 和 Jon A Wellner。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(3):596–608，1997。Sara A Van de Geer 等人。高维广义线性模型与套索。统计学年鉴，36(2):614–645，2008年。Roman Vershynin。随机矩阵非渐近分析导论。*arXiv*预印本 *arXiv:1011.3027*，2010。Roman Vershynin。高维概率：面向数据科学应用的导论，第47卷。剑桥大学出版社，2018。Martin J Wainwright。高维统计学：非渐近视角，第48卷。剑桥大学出版社，2019。Gill Ward、Trevor Hastie、Simon Barry、Jane Elith 和 John R Leathwick。仅存在数据与期望最大化算法。生物统计学，65(2):554–563，2009。Stephen J Wright、Robert D Nowak 和 M´ario AT Figueiredo。通过可分离近似进行稀疏重构。*IEEE*信号处理汇刊，57(7):2479–2493，2009。Michael J Wurm、Paul J Rathouz 和 Bret M Hanlon。正则化有序回归与 ordinalnet R软件包。*arXiv*预印本 *arXiv:1706.05003*，2017。Jason Xu、Eric C Chi 和 Kenneth Lange。到集合距离惩罚下的广义线性模型回归。*arXiv*预印本 *arXiv:1711.01341*，2017。Willard I Zangwill。非线性规划：统一方法，第52卷。Prentice-Hall，恩格尔伍德克利夫斯，新泽西州，1969。