arXiv:2305.01171v1 [stat.ME] 2 May 2023

高维数据中用于最优治疗决策的平滑一致性辅助学习

Angzhi Fan *芝加哥大学统计学系

2023年5月3日

摘要

最优治疗方案是个体化治疗决策规则，其期望治疗结果最优。治疗决策规则的一个简单形式是线性决策规则，它由其系数和阈值刻画。随着患者异质性数据不断积累，在高维情形下估计采用线性决策规则的最优治疗方案颇受关注。单时间点最优治疗方案可以通过一致性辅助学习（CAL）进行估计，该方法基于成对比较。CAL 灵活且在低维情形下效果良好。然而，由于其中包含指示函数，CAL 在高维情形下难以优化。最近，研究人员提出了一种平滑方法，利用一族累积分布函数替代指示函数。本文提出平滑一致性辅助学习（SMCAL），通过一族 Sigmoid 函数对 CAL 应用平滑方法。随后，我们针对协变量为连续变量的情形，分析近似误差和随机误差，并证明估计系数的收敛速率。我们还考虑了协变量为离散变量的情形，并建立了类似的结果。我们进行了模拟研究，展示了该方法的优势。

关键词：最优治疗方案、精准医学、平滑近似、单调单指标模型

*作者谨此感谢 Wenbin Lu 教授和 Rui Song 教授在作者 2017年访问北卡罗来纳州立大学期间给予的悉心指导。

1. 引言

为了给出精准的治疗决策，我们必须考虑患者的异质性。基于患者自身特征制定的决策规则称为个体化治疗规则。在二元治疗决策场景中， $\mathcal{A} = \{1, 0\}$ 是治疗指标，且

$$\{(Y_i, X_i, A_i), i = 1, 2, ..., n\}$$

为独立同分布观测，其中 Y_i 和 X_i 分别是受试者 i的结局和特征。假设 $Y_i^*(A_i)$ 是受试者 i 在具有特征 X_i 并接受治疗 A_i时的预期结局。如果 $Y_i^*(1) > Y_i^*(0)$，与治疗 0 相比，治疗 1 对受试者 i 更有利。假设线性决策规则 $d(X) = \mathbb{I}(\beta^T X > c)$，则治疗方案由系数 β 和阈值 c确定。平均治疗结局可以建模为

$$E[Y^*(d(X))] = E_X[E(Y|X, A = 1)d(X) + E(Y|X, A = 0)(1 - d(X))].$$

最优治疗方案是指能够产生最有利平均治疗结局的治疗方案。具有多个决策时点的最优治疗决策称为最优动态治疗方案。目前主要有两种动态治疗学习方法。第一种是Watkins（1989）提出的 Q学习。Watkins和Dayan（1992）证明了 Q学习的收敛性质。Song等人（2011）将 Q学习扩展为惩罚Q学习（PQ学习）。第二种是 Murphy（2003）提出的优势学习（A学习）。Blatt等人（2004）指出，与 Q学习相比，A学习可能具有较小的偏倚。

随着患者异质性数据的迅速累积，人们开始关注高维情境。Zhu等人（2019）将Q学习扩展为高维Q学习，重点研究两阶段动态治疗方案。Shi等人（2018）

提出了惩罚A学习（PAL），在A学习框架下应用Dantzig选择器，重点关注两阶段或三阶段情境。

对于单个时间点的治疗决策，Zhang 等人（2012年）采用双重稳健的增广逆概率加权估计量，以处理模型可能设定错误的问题。Zhao 等人（2012年）采用支持向量机框架，并提出了结果加权学习（OWL）。Song 等人（2015年）将 OWL 改进为惩罚性结果加权学习（POWL）。Lu 等人（2013年）提出了一种稳健估计量，该估计量无需估计基线函数，并且易于在高维情形下进行变量选择。Fan 等人（2017年）提出的 CAL 利用两种治疗之间的对比构造一致性函数，再利用该一致性函数估计系数 β^*。CAL 的收敛率较快，且不假设对比函数是特征的线性组合。他们假设满足 Rubin（1980年）提出的稳定单位治疗值假设（SUTVA），

$$Y_i^*(A_i) = \mathbb{I}(A_i = 0)Y_i^*(0) + \mathbb{I}(A_i = 1)Y_i^*(1), \tag{1.1}$$

以及无未测量混杂因素条件，

$$\{Y_i^*(0), Y_i^*(1)\} \perp A_i | X_i. \tag{1.2}$$

所提出的一致性函数 $C(\beta)$ 为

$$\begin{aligned} C(\beta) &= E\{[\{Y_i^*(1) - Y_i^*(0)\} - \{Y_j^*(1) - Y_j^*(0)\}]\mathbb{I}(\beta^T X_i > \beta^T X_j)\} \\ &= E\{[D(X_i) - D(X_j)]\mathbb{I}(\beta^T X_i > \beta^T X_j)\}, \end{aligned} \tag{1.3}$$

其中，对比函数 $D(X_i) = E[Y_i|A_i = 1, X_i] - E[Y_i|A_i = 0, X_i]$。在一致性函数中，一个重要的假设是 $\beta^{*T} X_i$ 与 $D(X_i)$ 一致，

这意味着

$$D(X_i) = Q(\beta^{*T} X_i), \tag{1.4}$$

其中 $Q(x)$ 是未知的单调递增函数。使用倾向评分$\pi(X_i) = P(A_i = 1|X_i)$，其 $D(X_i)$ 由其无偏估计量估计

$$w_i = \frac{\{Y_i - v(X_i, \hat{\theta})\}\{A_i - \pi(X_i)\}}{\pi(X_i)\{1 - \pi(X_i)\}},$$

ˆ其中 $v(X_{i,}\theta)$ 可以是与 A_i 无关的任意函数，但通常选择为接受治疗 0 的患者的平均响应。线性决策规则中的真实 β^* 和真实阈值 c_0 通过以下方式估计

$$\hat{\beta} = \underset{\beta:\|\beta\|_2=1}{\arg\max} \frac{1}{n(n-1)} \sum_{i \neq j} (w_i - w_j) \times_{\mathbb{I}} (\beta^T x_i > \beta^T x_j), \tag{1.5}$$

and

$$\hat{c} = \arg\max_c \frac{1}{n} \sum_{i=1}^{n} \frac{Y_{i\,\mathbb{I}}[A_i = {}_{\mathbb{I}}(\hat{\beta^T} x_i > c)]}{A_i \pi(x_i) + (1 - A_i)[1 - \pi(x_i)]}. \tag{1.6}$$

然而，在高维情境下，由于一致性函数中包含指示函数，CAL 的优化有一定难度。为了解决 CAL 的这一优化问题，Liang 等人（2018）提出了 SCAL，使用合页损失替代 $_{\mathbb{I}}(x)$，并添加了用于变量选择的 L_1 惩罚项。他们通过以下方式估计 β^* 和 c_0：

$$\hat{\beta} = \underset{\beta:\|\beta\|_2=1}{\arg\min} \frac{2}{n(n-1)} \sum_{w_i > w_j} (w_i - w_j) \times (1 - \beta^T (x_i - x_j))_+ + \lambda \|\beta\|_1,$$

and

$$\hat{c} = \arg\max_c \frac{1}{n} \sum_{i=1}^{n} w_{i\,\mathbb{I}} (\hat{\beta^T} x_i > c),$$

它实现了 $n^{1/2}$-率，但由于合页损失与原始0-1 损失之间存在差异，因此产生了相对较大的偏倚。

我们采用 Han 等人（2017年）提出的平滑方法，克服高维数据下 CAL 的优化难题。我们使用一族 Sigmoid 函数替代 CAL 中的指示函数，并加入 L_1 惩罚项。随后，我们采用坐标下降算法进行优化。我们的方法称为平滑一致性辅助学习（SMCAL）。

与 SCAL 相比，我们的 SMCAL 收敛率较慢，但偏倚显著更小。与 POWL、SCAL 和 PAL 的数值比较表明，我们的方法具有优势，尤其是在样本量相对较大时。除了 SCAL 设置中的连续协变量情形外，我们还将 SMCAL 扩展到离散协变量情形。

我们的工作与 Han 等人（2017年）有三点主要区别。首先，他们研究的是 Han（1987年）提出的最大秩相关（MRC），而我们的平滑方法应用于一致性函数。其次，他们的论文研究单调变换模型，而我们关注单调单指标模型，且我们的收敛率更快。第三，我们采用比他们针对协变量提出的正态性假设更弱的假设条件，并将我们的方法扩展到离散情形。

本文其余部分安排如下：第 2 节介绍 SMCAL，并描述坐标下降算法。第 3 节给出连续情形下的 L_2-误差率和离散情形下的 L_1-误差率，这是本文的主要结果。第 4 节将 SMCAL 与 SCAL、PAL 和 POWL 进行数值比较。第 5 节给出 STAR*D 数据集的真实数据应用。第 3 节中引理和定理的证明见补充材料。

2. 方法

2.1. 模型设定

我们仍然假设 Fan 等人（2017）提出的假设条件 (1.1)、(1.2) 和 (1.4) 成立。我们的平滑过程将其一致性函数近似为

$$C_n(\beta) = E\{[D(X_i) - D(X_j)]F(\beta^T(X_i - X_j)\alpha_n)\}, \tag{2.1}$$

其中，我们采用一族 Sigmoid 函数

$$F(\alpha_n x) = \frac{1}{1 + e^{-\alpha_n x}}$$

来替代 Fan 等人（2017）中的 $\mathbb{I}(x)$。 α_n 是一个依赖于 n 的正数常量。 $C_n(\beta)$ 的无偏估计量为

$$\hat{C_n}(\beta) = \frac{1}{n(n-1)}\sum_{i \neq j}(w_i - w_j) \times F(\beta^T(x_i - x_j)\alpha_n). \tag{2.2}$$

不失一般性，我们可以令 $\beta_1^* = \beta_1 = 1$。定义

$$\beta^*_{r(\gamma),\alpha_n} = \underset{\beta:\beta_1=1, \|\beta-\beta^*\|_2 \leq r(\gamma)}{\arg\min} (-C_n(\beta))$$

其中 $r(\gamma)$ 为 一个待确定的常数。$\beta^*_{r(\gamma),\ \alpha_n}$ 将收敛于 β^* ，当 α_n 趋于无穷大时。

定义 $\mathcal{H} = \{\beta \in \mathbb{R}^d :\ \beta_1 = 1、\ \|\beta\|_0 \leq s_n、\ \|\beta - \beta^*\|_2 \leq r\ （\gamma）\}$，其中 s_n 是一个常数，使得 $\|\beta^*_{r(\gamma),\ \alpha_n}\|_0 \leq s_{n_0}$ 注意，最大化 C_n（β）等价于最大化

$$\frac{2}{n(n-1)}\sum_{w_i > w_j}(w_i - w_j) \times F(\beta^T(x_i - x_j)\alpha_n),$$

我们可以将损失函数写为

$$l_n(\beta) = -\frac{2}{n(n-1)}\sum_{w_i > w_j}(w_i - w_j) \times F(\beta^T(x_i - x_j)\alpha_n) + \lambda_n\|\beta\|_1. \tag{2.3}$$

然后，我们使用以下两个步骤来估计 β^* 和 c。

$$\hat{\beta}_{r(\gamma),\alpha_n} = \arg\min_{\beta\in\mathcal{H}} l_n(\beta) \tag{2.4}$$

$$\hat{c} = \arg\max_{c} \frac{1}{n}\sum_{i=1}^{n} w_i \mathbb{I}(\hat{\beta}^T_{r(\gamma),\alpha_n} x_i > c) \tag{2.5}$$

ˆ我们的 $\beta_{r(\gamma),\alpha_n}$ 与 $\beta^*_{r(\gamma),\alpha_n}$ 不同，这是由随机因素和 L_1 惩罚项造成的。ˆ 我们将 $\beta^*_{r(\gamma),\alpha_n} - \beta^*$ 称为近似误差，将 $\beta_{r(\gamma),\alpha_n} - \beta^*_{r(\gamma),\alpha_n}$ 称为随机误差。

2.2. 坐标下降算法

我们可以在每个坐标上迭代使用近端梯度下降来求解优化问题。步长 t是一个固定常数。

算法 1 坐标下降算法

1: 初始化 $\beta = (1, 0, ..., 0)^T$ 或 $\beta = (-1, 0, ..., 0)^T$。重复以下步骤，直到收敛：对于 $k = 2$、3、. . .、 d: 2: 计算 $g_k = {}^{-\alpha}_{n\ n-1()} \sum_{i\neq.j} (w_i - w_j)(x_{j,k} - x_{j,k})F'((x_i - x_j)^T\beta\alpha)$。3: 更新 $\beta_k = S_{t\lambda}(\beta_k - tg_k)$，其中 $S_{t\lambda}(x) = \begin{cases} x - t\lambda & x > t\lambda \\ 0 & -t\lambda \le x \le t\lambda \\ x + t\lambda & x < -t\lambda \end{cases}$ 是软阈值算子。

考虑 $-{}^{2}_{n\ (n-1)} \sum (w_i - w_j)F((x_i - x_j)^T\beta\alpha_n)$。当一对 (i, j) 满足 $w_i > w_j$,$^{w_i>w_j}$时，我们很可能有 $\overline{D(X_i)} > D(X_j)$，这意味着 $(x_i - x_j)^T\beta^* > 0$。因为 $-F(x)$ 是凸的

当 $x > 0$ 时，我们知道，当 β 接近 β^* 时，我们的优化问题很可能是凸的。 λ 和 α 可以通过交叉验证来选择。

3. Con估计系数的收敛速率

ˆ 在本节中，我们建立 $\beta_{r(\gamma),\alpha_n} - \beta^*$ 在连续协变量情形以及随后 ˆ 离散协变量情形下的误差率。令 $\lambda_n \geq 2||\nabla L_n(\beta^*_{r,\alpha_n})||_\infty$。假设真实系数的稀疏性为 β^*。令 S 表示 β^* 中的非零索引， $s = |S|$，并定义 $S^c = \{1, 2, ...d\} \setminus S$。令 $X_{i,I}$ 表示 X_i 中索引属于集合 I 的所有条目。符号 . 表示，如果 $a_n \lesssim b_n$，或等价地 $b_n \gtrsim a_n$，则存在 $C > 0$ 和 $N_0 > 0$，使得对于所有 $n > N_0$，都有 $|a_n| \leq C|b_n|$。如果 $a_n \lesssim b_n$ 且 $b_n \lesssim a_n$，则记为 $a_n \asymp b_n$。

3.1. 连续情形

3.1.1. 近似误差

为分析近似误差，我们作出如下假设。 s.

假设1.

(A1) $Q(x)$ 二阶可微，且 $|Q(x)| < M_1$。(A2) 对所有 β： $\beta_1 = 1$、$||\beta - \beta^*||_2 \leq r(\gamma)$ 和 $t \in \mathbb{R}$，均有 $|\frac{\partial}{\partial t} f_{X^T\beta}(t)| < M_3$。(A3) 令 $\Gamma_{(d-1)\times(d-1)} = -\nabla^2 C(\beta^*)$，其中导数是相对于除第一个以外的每个坐标求得，因为 $\beta^*_1 = 1$ 是固定的。假设 $0 < c_3 \leq \lambda_{min}(\Gamma) \leq \lambda_{max}(\Gamma) \leq c_4$。

注 1. 对许多 X 的分布而言，(A2) 都成立。例如，容易证明，如果$X \sim \mathcal{N}(\mu, \Sigma_{d\times d})$ 且 $0 < c_1 \leq \lambda_{min}(\Sigma)$，那么 (A2) 就成立。

注 2. 由于 β^* 是 $C(\beta)$的最大化者， Γ 是半正定的。Han 等人（2017）在第 4.1.1 节中也提出了类似于 (A3) 的假设。但在离散情形下，我们将采用另一种方法构造误差上界，因此不再需要这一假设。

在上述假设条件下，以下引理 1 衡量一致性函数 $C(\beta)$ 与其近似 $C_n(\beta)$ 之间的接近程度。

引理 1.

$$\sup_{\beta\in\mathbb{R}^d:\beta_1=1,\|\beta^*-\beta\|_2\leq r(\gamma)} |C(\beta) - C_n(\beta)| \lesssim \alpha_n^{-2} \tag{3.1}$$

假设 1 中的凸性假设 (A3) 意味着 $\|\beta^*_{r(\gamma),\ \alpha_n} - \beta^*\|_2^2$ 和$|C(\beta^*) - C(\beta^*_{r(\gamma),\ \alpha_n})|$ 具有相同的收敛速率，因此可以通过分析 $|C(\beta^*) - C(\beta^*_{r(\gamma),\ \alpha_n})|$来界定近似误差；对此，我们可以应用引理 1 并得到定理 2。定理 2（近似误差率）。

$$\|\beta^*_{r(\gamma),\alpha_n} - \beta^*\|_2^2 \lesssim \alpha_n^{-2} \tag{3.2}$$

注 3。Han 等人（2017）中的近似误差为

$$\|\beta^*_{r(\gamma),\alpha_n} - \beta^*\|_2^2 \lesssim \alpha_n^{-1} \cdot \sup_{\beta:\beta_1=1} \frac{1}{\sqrt{2\beta^T\Sigma\beta}}.$$

我们有一个 f 更快的收敛率，因为我们使用一致性函数估计量。

$$\frac{1}{n(n-1)}\sum_{i\neq j}(w_i - w_j)\times_{\mathbb{I}}(\beta^T x_i > \beta^T x_j)$$

这与最大秩相关（MRC）估计量不同。

$$\frac{1}{n(n-1)}\sum_{i\neq j}\text{sign}(w_i - w_j)\times\text{sign}(\beta^T x_i - \beta^T x_j)$$

在他们的论文中。

3.1.2. 随机误差

ˆ 在研究随机误差 $\beta_{r(\gamma),\alpha_n} - \beta^*_{r(\gamma),\alpha_n}$ 时，我们遵循 Han 等人（2017）第 4.2 节的框架。为与他们使用的符号保持一致，定义

$$\hat{L}_n(\beta) = -\hat{C}_n(\beta) = -\frac{1}{n(n-1)}\sum_{i\neq j}(w_i - w_j)\times F(\beta^T(X_i - X_j)\alpha_n),$$

$$L(\beta) = -C(\beta), \quad L_n(\beta) = -C_n(\beta),$$

$$\hat{\Delta} = \hat{\beta}_{r(\gamma),\alpha_n} - \beta^*_{r(\gamma),\alpha_n},$$

且 $P(\cdot)$ 表示 L_1 惩罚项， $P^*(\cdot)$ 是 $P(\cdot)$ 的对偶范数。本小节需要以下假设条件。

假设 2。

(A4) $|w_i| \leq M_2$.

(A5) $\exists M_4,\ s.t.\ -M_4 \leq \|X_i\|_\infty \leq M_4$.

(A6) $\|\beta^*_{r(\gamma),\alpha_n}\|_0 \leq s_n$.

注 4. 我们的假设 (A4) 与假设1中的 (A1) 非常相似，但 (A1) 不包含任何随机因素。

注 5. 尽管在我们的最优治疗决策问题中提出 (A5) 是合理的，但 (A5) 不可避免地排除了多元正态分布。

注 6. (A6) 曾作为假设 (A0) 出现在 Han 等人（2017）的第 4.2.2 节中。正如我们¯将在定理 7 的证明中看到的，在某些条件下，我们有 $\beta^*_{r(\gamma),\ \alpha_n,\ S^c} = 0$。因此 s_n可以相对较小。

为应用ˆ Han 等人（2017年）中的定理 4.8，我们给出两个必要步骤，即引理 3 和引理 4。引理 3 用于界定项 $\langle \nabla L_n(\beta^*_{r(\gamma),\alpha_n})\ \Delta \rangle$。

引理 3。

$$\|\nabla \hat{L}_n(\beta^*_{r(\gamma),\alpha_n})\|_\infty \overset{p}{\lesssim} \alpha_n \sqrt{\frac{log(d)}{n}} \tag{3.3}$$

ˆ 在 Han 等人（2017年）中，他们引入了参数 $\kappa_n = E \sup_{\beta\in\mathcal{H}} |L_n(\beta) - L_n(\beta)|$，并指出当固定 d 时，$\kappa_n = O(n^{-1/2})$。然而，在我们的情况下，d 可能远大于 n，因此我们将采用不同的方法，即引理 4，来分析 ˆsup$_{\beta\in\mathcal{H}} |L_n(\beta) - L_n(\beta)|$ 的速率。定义

$$B(h) = |\hat{L}_n(\beta^*_{r(\gamma),\alpha_n} - h) - L_n(\beta^*_{r(\gamma),\alpha_n} - h) - \hat{L}_n(\beta^*_{r(\gamma),\alpha_n}) + L(\beta^*_{r(\gamma),\alpha_n})|,$$

下面的引理 4 以高概率给出了 $B(h)$ 的一个界。

引理 4.

$$\sup_{\|h\|_0\le 2s_n, \|h\|_2\le 2r(\gamma)} \frac{B(h)}{\|h\|_2} \overset{p}{\lesssim} \frac{\sqrt{s_n^2 log(d)}}{\sqrt{n}} \alpha_n \tag{3.4}$$

ˆ ˆ ˆ$B(h)$ 用于界定 $L_n(\beta^*_{r(\gamma),\alpha_n}+\Delta) - L_n(\beta^*_{r(\gamma),\alpha_n})$ 与 $L_n(\beta^*_{r(\gamma),\alpha_n}-$ˆ$\Delta) - L_n(\beta^*_{r(\gamma),\alpha_n})$ 之间的差值。通过采用与 Han 等人（2017）中的定理 4.8 和引理4.10 类似的论证，我们可以证明下述定理所述的随机误差率。

定理5（随机误差率）。

$$\|\hat{\beta}_{r(\gamma),\alpha_n} - \beta^*_{r(\gamma),\alpha_n}\|_2^2 \overset{p}{\lesssim} \alpha_n^2 log(d) s_n^2/n + \alpha_n^{-2} \tag{3.5}$$

由定理2和定理5，我们得出 $\|$ˆ$\beta_{r(\gamma),\ \alpha_n} - \beta^*\|_2^2 \overset{p}{\lesssim} \alpha_n^2 log(d) s_n^2/n + \alpha_n^{-2}$。通过选择 $\alpha_n \asymp (\ log\ ^n_{d()s_n^2})^{\frac{1}{4}},$，我们得到总体 L_2 误差率为

定理6（总体误差率）。

$$\|\hat{\beta}_{r(\gamma),\alpha_n} - \beta^*\|_2 \overset{p}{\lesssim} (log(d) s_n^2/n)^{\frac{1}{4}} \tag{3.6}$$

注 7. 为作比较，Han 等人（2017）证明了 $\|\beta^*_{r(\gamma),\alpha_n} - \beta^*\|_2^2 \lesssim \alpha_n^{-1}$和$\|$ˆ$\beta_{r(\gamma),\alpha_n} - \beta^*_{r(\gamma),\alpha_n}\|_2^2 \overset{p}{\lesssim} \alpha_n^2 log(d) s_n/n + \alpha_n^{-1} + \kappa_n$。选择 $\alpha_n \asymp (\ log\ ^n_{d()s_n})_{\ 1\ 3}$，则在他们的论文中 $\|$ˆ$\beta_{r(\gamma),\ \alpha_n} - \beta^*\|_2 \overset{p}{\lesssim} (log(d) s_n/n)^{\ \frac{1}{6}} + \kappa_n^{\frac{1}{2}}$。

3.2. 离散情形

在本节讨论离散协变量情形时，我们将采用一些不同的假设条件。离散情形与连续情形的主要区别在于，在离散情形中，$-C(\beta)$ 可能在 β^* 附近为常数。

假设3。

(B1) 存在一个矩阵 $\mathcal{M} \in \mathbb{R}^{(d-s)\times(s-1)}$，使得条件分布 $(X_i - X_j)_{S^c} - \mathcal{M}(X_i - X_j)_{S\setminus 1}|(X_i - X_j)_S$是一个以原点为中心的对称分布。(B2) $|w_i| \le M_2$。(B3) $-M_4 \le \|X_i\|_\infty \le M_4$。

注 8. 在连续协变量情形下，某些多元正态分布满足 (B1)；此时，我们可以认为 S 略大于非零索引集。我们可以令 S 包含与第一个特征相关的特征。

注 9. (B2) 和 (B3) 在连续情形中也会出现，并用于引理 3 和 4 的证明。

在离散情形下，我们同时对近似误差和随机误差进行界定。引理 3 和 4 对离散情形也成立，证明几乎相同。此外，我们有如下不同的主定理：

定理 7（离散情形下的误差率）。只要 $s^3\log(d)/n \to 0$，对于所有满足 $(x_i^T - x_j^T)\beta^* > 0$ 的 (x_i, x_j)，我们有 $(x_i^T - x_j^T)_S \hat{\beta}_{r,\alpha_n,S} \overset{p}{>} \frac{\epsilon}{\alpha_n}$，并且 $\|\hat{\beta}_{r,\alpha_n,S^c}\|_1 \overset{p}{\lesssim} (s^3\log(d)\, n)^{1/2}$。注 10。在证明中，我们选择 $\alpha_n = max(1, \frac{log(c_{11})}{-log(\delta)+log(3)}\varepsilon)$。利用 (B3)，该定理意味着，对于任意 i, j, $|\hat{\beta}_{r,\alpha_n,S^c}(X_i^T - X_j^T)_{S^c}| \overset{p}{\lesssim} (s^3\log(d)\, n)^{1/2}$。~~如果 $(x_i^T - x_j^T)\beta^* > 0$~~，那么 $(x_i^T - x_j^T)_S \hat{\beta}_{r,\alpha_n,S} \overset{p}{>} \frac{\varepsilon}{\alpha_n} \gtrsim (s^3\log(d)\, n)^{1/2} \overset{p}{\gtrsim} |\hat{\beta}_{r,\alpha_n,S^c}(x_i^T - x_j^T)_{S^c}|$。~~因此，当~~

$\frac{s^3\log(d)}{n} \longrightarrow 0$，我们得出结论：一旦 $(x_i^T - x_j^T)\beta^* > 0$，那么 $(x_i^T - x_j^T)\hat{\beta}_{r,\ \alpha_n} \overset{p}{\geq} 0$。换句话说，如果对比函数 $D(x_i) > D(x_j)$，那么 $x_i^T\hat{\beta}_{r,\ \alpha_n} \geq x_j^T\hat{\beta}_{r,\ \alpha_n}$，这意味着我们很有可能按照正确的顺序对受试者进行排序。

4. 模拟研究

本节比较四种方法：Song 等人（2015）的 POWL、Shi 等人（2018）的 PAL、Liang 等人（2018）的 SCAL，以及 SMCAL。POWL 方法最小化

$$\frac{1}{n}\sum_{i=1}^{n}\frac{Y_i \cdot [1-(2A_i-1)g(X_i)]_+}{A_i\pi(X_i)+(1-A_i)[1-\pi(X_i)]} + \lambda\sum_{i=1}^{d}|\beta_i| \tag{4.1}$$

其中，$g(X_i) = \beta^T \cdot X_i + \beta_0$ 是决策函数，调节参数 λ 通过使逆概率加权估计量最大的方式选取。PAL 是一种动态治疗方案方法，但由于 PAL 首先估计最后一个阶段，因此可用于单阶段问题。PAL 使用惩罚A学习进行变量选择，然后使用非惩罚A学习求解系数。我们使用 PAL 作者提供的 R软件包获取其结果。

为评估估计系数，我们报告归一化后的均方误差（MSE）。为评估变量选择的准确性，我们报告错误零值（真实系数为零，但估计值非零）和正确零值（真实系数为零，且估计值为零）。为评估估计的治疗方案，我们报告正确决策百分比（PCD），以及遵循估计治疗方案时的平均响应（估计值）。估计值通过抽取 1000 个新样本计算得出。PCD 和估计值的标准误列于括号中。

4.1. 低维情形

4.1.1. 线性情形

第一个示例见 Zhao 等人（2012年）以及 Liang 等人（2018年）。设$d = 50$、 X_{i1}、X_{i2}、...、 X_{i50}为相互独立的变量，在连续情形下均由 $U(-1, 1)$生成。对于所有 $i = 1, ...n$。 $Y_i|X_i$、 $A_i \sim N(Q_0(X_i), 1)$，其中 $Q_0(X_i) = 1 + 2X_{i1} + X_{i2} + 0.5X_{i3} + 0.442(1 - X_{i1} - X_{i2})(2A_i - 1)$。分别针对 $n = 30$、100 和 200 进行了 100 次模拟。$\pi(X_i)$均取 0.5，2，

在离散情形下，除从 $\{-0.9, -0.7, -0.5, -0.3, -0.1, 0.1, 0.3, 0.5, 0.7, 0.9\}$ 上的均匀分布生成 X_{i1}、 X_{i2}、...、 X_{i50} 外，其他均保持不变。

数值比较结果汇总于表 1 中。我们仅对 SMCAL 和 PAL 进行了模拟，SCAL 和 POWL 的结果来自 Liang 等人（2018）。

一般而言，POWL 会得到许多估计值为零的结果。我们的均方误差大于 SCAL，但小于 POWL。事实上，我们的理论结果表明，其收敛率比 SCAL 更慢。当样本量为 200 时，我们的 PCD 和估计值显著高于 SCAL 中的相应结果，这可能是因为与 SCAL 相比，SMCAL 的偏倚更小。我们还注意到，200 个样本的 SCAL 的 PCD 异常地小于 100 个样本的 SCAL 的 PCD。这可能是因为合页损失引入了偏倚，导致 SCAL 无法收敛到真实系数。总之，我们认为，当样本量较小时，SCAL 优于 SMCAL；但随着样本量增大，SMCAL 将越来越优于 SCAL。

在离散情形中，均方误差、变量选择、PCD 和估计值呈现出相似的模式。

在这个例子中，PAL 的表现最佳。100 个样本的 PAL 已经大幅

表 1：POWL、PAL、SCAL 和 SMCAL 的比较：低维情形。

	n	MSE	Incorr0(0)	Corr0(48)	PCD	估计值
POWL	30	1.60	1.70	42.23	0.615(0.02)	1.09(0.02)
	100	1.27	1.94	46.64	0.768(0.02)	1.27(0.02)
	200	1.09	1.99	47.78	0.786(0.02)	1.30(0.03)
PAL	30	1.83	1.76	46.23	0.631(0.012)	1.17(0.015)
	100	1.01	0.92	46.53	0.808(0.009)	1.37(0.009)
	200	0.32	0.17	47.00	0.903(0.005)	1.45(0.006)
SCAL	30	1.40	0.73	35.79	0.659(0.01)	1.16(0.01)
	100	0.52	0.11	41.97	0.764(0.01)	1.31(0.01)
	200	0.19	0.01	46.03	0.749(0.01)	1.32(0.01)
SMCAL	30	0.93	0.82	43.09	0.677(0.014)	1.21(0.017)
	100	0.81	0.75	44.11	0.735(0.010)	1.28(0.013)
	200	0.69	0.57	45.42	0.788(0.007)	1.34(0.009)
SMCAL-离散	30	0.95	0.89	43.12	0.653(0.014)	1.20(0.017)
	100	0.79	0.74	44.15	0.723(0.009)	1.28(0.011)
	200	0.70	0.50	43.78	0.764(0.006)	1.33(0.008)

表2：PAL与SMCAL的比较：非线性情形。

	n	MSE	Incorr0(0)	Corr0(46)	PCD	估计值
PAL	30	1.44	3.12	44.30	0.579(0.007)	10.42(0.431)
	100	0.82	2.22	45.35	0.666(0.006)	14.67(0.486)
	200	0.58	1.80	45.83	0.706(0.005)	16.39(0.552)
SMCAL	30	0.90	1.03	33.53	0.689(0.006)	17.45(0.464)
	100	0.52	0.05	29.44	0.814(0.006)	18.56(0.319)
	200	0.28	0.00	30.66	0.888(0.004)	19.92(0.510)

在 PCD 和估计值方面优于 200 个样本的 SCAL 和 SMCAL。在此例中，对比函数 $D(X_i) = 0.884(1 - X_{i1} - X_{i2})$ 是特征的线性组合。在这种线性情形下，基于成对比较的方法（如 CAL、SCAL 和 SMCAL）似乎不如 PAL 高效。但 CAL、SCAL 和 SMCAL 仅假设 $D(X_i) = Q(\beta^{*T} X_i)$，这种假设较为灵活，也可应用于非线性情形。

4.1.2. 非线性情形

第二个例子（见表2）比较了在对比函数关于特征呈非线性时 PAL 和 SMCAL 的表现。我们设定 $d = 50, n = 30, 100, 200, X_i \sim \mathcal{N}(0, I)$, $\pi(X_i) = 0.5$ 和 $Y_i = 1 + X_{i,\ 1} - X_{i,\ 2} + X_{i,\ 3} + X_{i,\ 4} + A \cdot (exp(1 + X_{i,\ 1} + X_{i,\ 2} - X_{i,\ 3} + X_{i,\ 4}) - exp(1)) + \epsilon$，其中 $\epsilon \sim \mathcal{N}(0, 0.5)$。我们的均方误差远小于 PAL；当样本量相对较大时，我们还能成功筛选出重要特征。我们的 PCD 和估计值也远优于 PAL。PAL 似乎不适用于此类非线性情形，但 SMCAL 表现相当不错。

4.2. 高维

我们针对以下六种高维模型进行模拟，这些模型见 Liang 等人（2018）的第 4.2 节。我们统一设定样本量为 $n = 100$，维度为$d = 500$。治疗类型为 A 的交互项为 $X\beta$ 或 $(X\beta)^3$,，这符合我们的单调单指标模型假设。然而，基线函数可以非常灵活。以下模型采用线性、多项式，甚至三角正弦函数作为基线函数。我们设定 $X \sim \mathcal{N}(0, I)$，$\pi(X_i) = 0.5$ 和 $\varepsilon \sim \mathcal{N}(0, 1)$。

模型1：$Y = 3+X\gamma_1+X\beta A+\varepsilon$，其中 $\gamma_1 = (-1, 1, \mathbf{0}_{d-2})^T$, $\beta = (2, 1.8, 0, 0, 0, -1.6, \mathbf{0}_{d-6})^T$。

模型2：$Y = 3 - 0.5(X\gamma_1)^2 + 0.625(X\gamma_2)^2 + X\beta A + \varepsilon$，其中 $\gamma_1 = (1, 0.5, \mathbf{0}_{d-2})^T$, $\gamma_2 = (0, 1, \mathbf{0}_{d-2})^T$, $\beta = (2, 1.8, 0, 0, 0, -1.6, \mathbf{0}_{d-6})^T$。

模型3：$Y = 1 - sin(X\gamma_1) + sin(X\gamma_2) + X\beta A + \varepsilon$，其中 $\gamma_1 = (1, \mathbf{0}_{d-1})^T$, $\gamma_2 = (0, 1, \mathbf{0}_{d-2})^T$, $\beta = (2, 1.8, 0, 0, 0, -1.6, \mathbf{0}_{d-6})^T$。

模型4：$Y = 3+X\gamma_1+(X\beta)^3A+\varepsilon$，其中 $\gamma_1 = (-1, 1, \mathbf{0}_{d-2})^T$, $\beta = (1, 0.9, 0, 0, 0, -0.8, \mathbf{0}_{d-6})^T$。

模型5：$Y = 3 - 0.5(X\gamma_1)^2 + 0.625(X\gamma_2)^2 + (X\beta)^3A + \varepsilon$，其中 $\gamma_1 = (1, 0.5, \mathbf{0}_{d-2})^T$, $\gamma_2 = (0, 1, \mathbf{0}_{d-2})^T$, $\beta = (1, 0.9, 0, 0, 0, -0.8, \mathbf{0}_{d-6})^T$。

模型6：$Y = 1 - sin(X\gamma_1) + sin(X\gamma_2) + (X\beta)^3A + \varepsilon$，其中 $\gamma_1 = (1, \mathbf{0}_{d-1})^T$, $\gamma_2 = (0, 1, \mathbf{0}_{d-2})^T$, $\beta = (1, 0.9, 0, 0, 0, -0.8, \mathbf{0}_{d-6})^T$。

Liang 等人（2018年）报告了 SCAL 的结果，我们则针对 SMCAL 进行了模拟，所有结果汇总于表3。结果显示，我们的模型选择和估计值与 SCAL 的结果相近。均方误差和 PCD 的表现不如 SCAL，这可能是因为 SCAL 的收敛率更快，因此在高维情况下表现优于 SMCAL。

表3：SCAL 和 SMCAL 的模拟结果：高维情形。

模型	MSE		Incorr0(0)		Corr0(497)		PCD		估计值	
	SCAL	SMCAL	SCAL	SMCAL	SCAL	SMCAL	SCAL	SMCAL	SCAL	SMCAL
模型 1	0.61	0.76	0.75	0.86	482.62	490.81	0.744(0.01)	0.732(0.005)	3.80(0.02)	3.82(0.016)
模型 2	0.56	0.71	0.57	0.55	485.34	492.69	0.763(0.01)	0.749(0.006)	3.79(0.03)	3.87(0.015)
模型 3	0.44	0.69	0.49	0.47	488.12	491.59	0.786(0.01)	0.751(0.005)	1.92(0.02)	1.88(0.015)
模型4	0.35	0.62	0.35	0.21	486.81	481.69	0.801(0.01)	0.752(0.006)	5.67(0.04)	5.69(0.043)
模型5	0.32	0.56	0.25	0.18	487.00	485.41	0.810(0.01)	0.740(0.008)	5.63(0.05)	5.62(0.057)
模型6	0.29	0.55	0.20	0.05	485.33	485.09	0.820(0.01)	0.774(0.007)	3.74(0.04)	3.78(0.040)

5. 真实数据分析

STAR*D研究聚焦于重性抑郁障碍（MDD），招募了4,000多名年龄在18至75岁之间的门诊患者。治疗级别共分四级：2、2A、3和4。在每个治疗级别，患者都被随机分配到不同的治疗组。研究记录了各种临床和社会经济因素以及治疗结果。更多详情请参见Fava等（2003年）。

我们聚焦于 STAR*D 研究的第 2 级数据。该数据集包含 319 个样本，每个样本包含患者ID、治疗类型、治疗结局以及另外 305 个临床或遗传特征。共有两种不同的治疗类型：安非他酮（BUP）和舍曲林（SER）。我们使用 0 标记 SER，使用 1 标记 BUP。16 项抑郁症状快速评估量表临床医师评定版（QIDS-C16）的评分范围为 0 至 24，评分越低表示结局越好。在我们的设定中，数值越大表示治疗效果越好，因此我们将 QIDS-C16 的负值作为治疗结局。

在选出的319名患者中，166名接受了SER，153名

表4：POWL、SCAL 和 SMCAL 的估计值。

治疗方案	估计值	差值	差值的 95%置信区间
最优方案（SCAL）	-6.77		
最优方案（SMCAL）	-6.90	0.13	(-0.46,0.69)
最优方案 (PAL)	-8.15	1.38	(0.71,2.02)
最优方案 (POWL)	-9.46	2.69	(1.18,4.24)
BUP	-10.50	3.75	(2.38,5.19)
SER	-10.72	3.97	(2.57,5.50)

接受了BUP。得到估计的治疗方案后，我们使用Zhang 等人（2012年）提出的逆概率加权估计量（IPW）

$$\frac{1}{n}\sum_{i=1}^{n}\frac{Y_i\mathbb{I}[A_i=\mathbb{I}(\beta^T\hat{X_i}>c)]}{A_i\pi(X_i)+(1-A_i)[1-\pi(X_i)]}$$

以计算估计值。我们进行了1000次自助法抽样，每次抽取1000个样本，以获得估计值和95%置信区间。

我们对 PAL 和 SMCAL 进行了模拟。Liang 等人（2018）报告了 SCAL、POWL 和非动态治疗方案的结果。这些结果均汇总于表4。以 SCAL 为基准，表中的差值表示各方法的估计值与 SCAL 估计值之间的差异。根据表4，最优治疗方案均优于非动态治疗方案。SCAL 和 SMCAL 得出的最优治疗方案具有最佳估计值。在这个真实数据示例中，PAL 和 POWL 的表现不如 SCAL 和 SMCAL。

已接受的治疗与估计的最优治疗方案的比较

表 5：SMCAL 推荐的治疗。

	推荐：SER	推荐：BUP
随机分配的治疗：SER	75	91
随机分配的治疗：BUP	68	85

见表5。随机治疗是指STAR*D研究中实际实施的治疗。估计治疗是指估计的最优治疗方案所建议的治疗。从表5可以看出，在接受BUP的153人中，SMCAL给出的最优治疗方案建议其中77人接受SER，76人继续接受BUP。这是一个相当均衡的治疗方案。

6. 结论

本文提出了 SMCAL，该方法基于 Han 等人（2017年）提出的一致性辅助学习（CAL）框架和平滑过程，旨在解决 CAL 在高维数据中的优化问题。我们分别针对连续协变量和离散协变量情形建立了收敛结果。所提出的 SMCAL 可成功应用于对比函数相对于特征呈非线性的情形，且与 SCAL 相比不会引入较大的偏倚。

补充材料

7. 证明

我们首先给出引理 8，该引理将在引理 4 的证明中用到。

引理 8. 存在 M_5，使得 $\sup_{h\in\mathbb{R}^d,\ \|h\|_0\le 4s_n} \frac{h^T D^T D h}{\frac{n(n-1)}{2}\|h\|_2^2} \le M_5 s_n$，其中

$$D = \begin{pmatrix} X_1^T - X_2^T \\ X_1^T - X_3^T \\ \cdots \\ X_1^T - X_n^T \\ X_2^T - X_3^T \\ \cdots \end{pmatrix}_{\frac{n(n-1)}{2}\times d}$$

证明 o 明。假设条件 2 中的 (A5) 或假设3中的 (B3) 意味着

$$\sup_{h\in\mathbb{R}^d:\|h\|_0\le 4s_n,\|h\|_2=1} ((X_i - X_j)^T h)^2 \le 4M_4^2 \sup_{h\in\mathbb{R}^d:\|h\|_0\le 4s_n,\|h\|_2=1} \|h\|_1^2 \le 16 M_4^2 s_n.$$

因此,

$$\begin{aligned} \sup_{h\in\mathbb{R}^d:\|h\|_0\le 4s_n} \frac{h^T D^T D h}{\frac{n(n-1)}{2}\|h\|_2^2} &\le \frac{2}{n(n-1)} \sup_{h\in\mathbb{R}^d:\|h\|_0\le 4s_n} \frac{h^T(\sum_{1\le i<j\le n}(X_i - X_j)(X_i - X_j)^T)h}{\|h\|_2^2} \\ &\le \frac{2}{n(n-1)} \sum_{1\le i<j\le n} \sup_{h\in\mathbb{R}^d:\|h\|_0\le 4s_n} \frac{((X_i - X_j)^T h)^2}{\|h\|_2^2} \\ &< \sup_{h\in\mathbb{R}^d:\|h\|_0\le 4s_n,\|h\|_2=1} ((X_i - X_j)^T h)^2 \le 16 M_4^2 s_n, \end{aligned} \tag{7.1}$$

这意味着，如果我们令 $M_5 = 16M_4^2$，那么

$$\sup_{h\in\mathbb{R}^d,\|h\|_0\le 4s_n} \frac{h^T D^T D h}{\frac{n(n-1)}{2}\|h\|_2^2} \le M_5 s_n.$$

□

7.1. 第 3.1 节的证明

引理 1 的证明。定义

$$W(v,\beta)=E[(Q(X_i^T\beta^*)-Q(X_j^T\beta^*))|(X_i^T-X_j^T)\beta=v]f_{(X_i^T-X_j^T)\beta}(v),$$

and

$$T=\frac{1}{2}(X_i^T+X_j^T)\beta,\quad U=X^T\beta^*-X^T\beta,$$

then

$$\begin{aligned}
&W(v,\beta)\\
=&\int_{\mathbb{R}}\int_{\mathbb{R}}\int_{\mathbb{R}}(Q(t+\frac{v}{2}+u_1)-Q(t-\frac{v}{2}+u_2))f_{(X^T\beta,U)}(t+\frac{v}{2},u_1)f_{(X^T\beta,U)}(t-\frac{v}{2},u_2)du_1du_2dt\\
=&\int_{\mathbb{R}}\int_{\mathbb{R}}Q(t+\frac{v}{2}+u_1)f_{(X^T\beta,U)}(t+\frac{v}{2},u_1)f_{X^T\beta}(t-\frac{v}{2})du_1dt\\
&-\int_{\mathbb{R}}\int_{\mathbb{R}}Q(t-\frac{v}{2}+u_2)f_{X^T\beta}(t+\frac{v}{2})f_{(X^T\beta,U)}(t-\frac{v}{2},u_2)du_2dt\\
=&\int_{\mathbb{R}}\int_{\mathbb{R}}Q(t+\frac{v}{2}+u_1)f_{(X^T\beta,U)}(t+\frac{v}{2},u_1)f_{X^T\beta}(t-\frac{v}{2})du_1dt\\
&-\int_{\mathbb{R}}\int_{\mathbb{R}}Q(t+\frac{v}{2}+u_1)f_{X^T\beta}(t+\frac{3v}{2})f_{(X^T\beta,U)}(t+\frac{v}{2},u_1)du_1dt\\
=&\int_{\mathbb{R}}\int_{\mathbb{R}}Q(t+\frac{v}{2}+u_1)f_{(X^T\beta,U)}(t+\frac{v}{2},u_1)(f_{X^T\beta}(t-\frac{v}{2})-f_{X^T\beta}(t+\frac{3v}{2}))du_1dt \qquad (7.2)
\end{aligned}$$

假设 1 中的 (A1) 和 (A2) 意味着 $|Q(t+\frac{v}{2}+u_1)|\le M_1$ 以及 $|f_{X^T\beta}(t-\frac{v}{2})-f_{X^T\beta}(t+\frac{3v}{2})|\le 2|v|M_3$。因此

$$|W(v,\beta)|\le 2M_1M_3|v|.$$

注意到

$$\begin{aligned}
&E[(Q(X_i^T\beta^*)-Q(X_j^T\beta^*))(1-F((X_i^T-X_j^T)\beta\alpha_n))_{\mathbb{I}}((X_i^T-X_j^T)\beta>0)]\\
=&E[(Q(X_i^T\beta^*)-Q(X_j^T\beta^*))F((X_j^T-X_i^T)\beta\alpha_n)_{\mathbb{I}}((X_j^T-X_i^T)\beta<0)]\\
=&E[(Q(X_j^T\beta^*)-Q(X_i^T\beta^*))F((X_i^T-X_j^T)\beta\alpha_n)_{\mathbb{I}}((X_i^T-X_j^T)\beta<0)],
\end{aligned}$$

我们有

$$\begin{aligned}|C(\beta)-C_n(\beta)| =&|E[(Q(X_i^T\beta^*)-Q(X_j^T\beta^*))(_{\mathbb{I}}((X_i^T-X_j^T)\beta>0)-F((X_i^T-X_j^T)\beta\alpha_n))]|\\ =&|E[(Q(X_i^T\beta^*)-Q(X_j^T\beta^*))(1-F((X_i^T-X_j^T)\beta\alpha_n))_{\mathbb{I}}((X_i^T-X_j^T)\beta>0)\\ &-(Q(X_i^T\beta^*)-Q(X_j^T\beta^*))F((X_i^T-X_j^T)\beta\alpha_n)_{\mathbb{I}}((X_i^T-X_j^T)\beta<0)]|\\ =&2|E[(Q(X_j^T\beta^*)-Q(X_i^T\beta^*))F((X_i^T-X_j^T)\beta\alpha_n)_{\mathbb{I}}((X_i^T-X_j^T)\beta<0)]|.\end{aligned}$$

因此,

$$\begin{aligned}&\sup_{\beta\in\mathbb{R}^d:\beta_1=1,\|\beta^*-\beta\|_2\le r(\gamma)}|C(\beta)-C_n(\beta)|\\ =&2\sup_{\beta\in\mathbb{R}^d:\beta_1=1,\|\beta^*-\beta\|_2\le r(\gamma)}|\int_{-\infty}^{0}W(v,\beta)\frac{1}{1+e^{-\alpha_n v}}dv|\\ \le&4M_1M_3|\int_0^{\infty}\frac{v}{1+e^{\alpha_n v}}dv|\\ =&4M_1M_3\alpha_n^{-2}|\int_0^{\infty}\frac{v}{1+e^{v}}dv|\lesssim\alpha_n^{-2}.\end{aligned}\tag{7.3}$$

□

定理2的证明。注意到

$$C(\beta^*_{r(\gamma),\alpha_n})-C_n(\beta^*_{r(\gamma),\alpha_n})\le C(\beta^*_{r(\gamma),\alpha_n})-C_n(\beta^*)\le C(\beta^*)-C_n(\beta^*),$$

我们有

$$\begin{aligned}|C(\beta^*)-C(\beta^*_{r(\gamma)\ \alpha_n})|\le&|C_n(\beta^*)-C(\beta^*_{r(\gamma)\ \alpha_n})|+|C(\beta^*)-C_n(\beta^*)|\\ \le&2\sup_{\beta\in\mathbb{R}^d:\beta_1=1,\|\beta^*-\beta\|_2\le r(\gamma)}|C(\beta)-C_n(\beta)|\lesssim\alpha_n^{-2}.\end{aligned}\tag{7.4}$$

最后一步依据引理 1 得出。

与 Han 等人（2017年）类似，我们可以选取一组正数常量 $\gamma=\{\gamma_1,\gamma_2\}$，使 $\frac{\gamma_2}{\gamma_1}$ 接近 1，从而对于某个足够小的 $r(\gamma)>0$、$\beta_1=\beta_1^*=1$ 和 $\|\beta-\beta^*\|_2\le r(\gamma)$，有

$$\gamma_1\lambda_{min}(\Gamma)\|\beta-\beta^*\|_2^2\le C(\beta^*)-C(\beta)\le\gamma_2\lambda_{max}(\Gamma)\|\beta-\beta^*\|_2^2,\tag{7.5}$$

于是根据假设1中的 (A3)，我们知道 λ_{min}（Γ） $\geq c_3$，因此

$$\| \beta^*_{r(\gamma),\alpha_n} - \beta^* \|_2^2 \leq \frac{C(\beta^*) - C(\beta^*_{r(\gamma),\alpha_n})}{\gamma_1 \lambda_{min}(\Gamma)} \lesssim \alpha_n^{-2}.$$

□

引理 3 的证明。设

$$h_k(x_i, x_j) = \frac{\partial}{\partial \beta_k}(w_i - w_j)F((x_i - x_j)^T \beta \alpha_n)|_{\beta=\beta^*_{r(\gamma),\alpha_n}},$$

其中 $k = 2$、3、...、d。于是， $Eh_k(X_i, X_j) = 0$，因为 $\beta^*_{r(\gamma),\alpha_n}$ 是 $C_n(\beta)$ 的局部最大值。

$$F^{'}(x) = 1/(e^x + e^{-x} + 2) \leq \frac{1}{4},$$

在假设 2 或假设3下，我们都有

$$\begin{aligned} |h_k(X_i, X_j)| =& |(W_i - W_j)F^{'}((X_i - X_j)^T \beta^*_{r(\gamma),\alpha_n}\alpha_n)\alpha_n(X_{i,k} - X_{j,k})| \\ \leq& \frac{1}{4}\alpha_n |(W_i - W_j)(X_{i,k} - X_{j,k})| \leq \alpha_n M_2 M_4. \end{aligned}$$

因此，利用 U 统计量的 Hoeffding 界（可见于例如 Pitcan（2017）），

$$\begin{aligned} P\{\|\nabla \hat{L}_n(\beta^*_{r(\gamma),\alpha_n})\|_\infty \geq t\} &= P\{\|\nabla \hat{C}_n(\beta^*_{r(\gamma),\alpha_n})\|_\infty \geq t\} \\ &\leq d \cdot P\{|\sum_{1\leq i<j\leq n} \frac{2}{n(n-1)} h_k(x_i, x_j)| \geq t\} \\ &\leq d \cdot 2exp\{-\frac{\lfloor n/2 \rfloor t^2}{2\alpha_n^2 M_2^2 M_4^2}\} \end{aligned} \tag{7.6}$$

选取 $t = \sqrt{8}\alpha_n M_2 M_4 \sqrt{\frac{log_{(d)}\, n}{}}$，证明完成。

□

引理 4 的证明。定义

$$\begin{aligned}
B(h) =&|\hat{L}_n(\beta^*_{r(\gamma),\alpha_n} - h) - L_n(\beta^*_{r(\gamma),\alpha_n} - h) - \hat{L}_n(\beta^*_{r(\gamma),\alpha_n}) + L(\beta^*_{r(\gamma),\alpha_n})| \\
=&|\frac{1}{n(n-1)}\sum_{i\neq j}(w_i - w_j)F((x_i - x_j)^T(\beta^*_{r(\gamma),\alpha_n} - h)\alpha_n) \\
&- E[(W_i - W_j)F((X_i - X_j)^T(\beta^*_{r(\gamma),\alpha_n} - h)\alpha_n)] \\
&- \frac{1}{n(n-1)}\sum_{i\neq j}(w_i - w_j)F((x_i - x_j)^T\beta^*_{r(\gamma),\alpha_n}\alpha_n) \\
&+ E[(W_i - W_j)F((X_i - X_j)^T\beta^*_{r(\gamma),\alpha_n}\alpha_n)]|
\end{aligned}$$

在假设 2 下，如果 $\|h\|_0 \leq 2s_n$,

$$\begin{aligned}
&|(w_i - w_j)F((X_i^T - X_j^T)(\beta^*_{r(\gamma),\alpha_n} - h)\alpha_n) - (w_i - w_j)F((X_i^T - X_j^T)\beta^*_{r(\gamma),\alpha_n}\alpha_n)| \\
\leq&\frac{1}{4}|w_i - w_j| \cdot |(X_i^T - X_j^T)h| \cdot \alpha_n \leq M_2M_4\alpha_n\|h\|_1 \leq M_2M_4\alpha_n\sqrt{2s_n}\|h\|_2.
\end{aligned}$$

根据 U 统计量的 Hoeffding 界,

$$P(B(h) \geq t) \overset{p}{\leq} 2\exp\{-\frac{\lfloor n/2\rfloor t^2}{4M_2^2M_4^2\alpha_n^2 s_n\|h\|_2^2}\}.$$

令 $t = \frac{\|h\|_2 t'}{\sqrt{n}}$ 和 $M_6 = 8M_2^2M_4^2$，则

$$P(\frac{B(h)}{\|h\|_2} \geq \frac{t'}{\sqrt{n}}) \overset{p}{\leq} 2\exp\{-\frac{t'^2}{s_n\alpha_n^2 M_6}\}.$$

取 $t' = C\sqrt{2s_n\ \overline{(s_n+1)\ log\ (d)}\ }\alpha_n$，则

$$P(\frac{B(h)}{\|h\|_2} \geq \frac{C\sqrt{2s_n(s_n+1)log(d)}\alpha_n}{\sqrt{n}}) \overset{p}{\leq} 2d^{-\frac{2(s_n+1)C^2}{M_6}}. \tag{7.7}$$

令 N_1 为 $\{h \in \mathbb{R}^d$ 的 ϵ-覆盖： $\|h\|_0 \leq 2s_n, \|h\|_2 = 2r(\gamma)\}$，其中我们约束 $N_1 \subset \{h \in \mathbb{R}^d$： $\|h\|_0 \leq 2s_n, \|h\|_2 = 2r(\gamma)\}$ ，且 ϵ 是一个待选取的小正数常量。

由于 $\{u \in \mathbb{R}^{2s_n}$ 的覆盖数: $\|u\|_2 = 2r(\gamma)\}$ 不应超过打包数，而打包数小于

$$(\frac{1+\frac{\epsilon}{2r(\gamma)}}{\frac{\epsilon}{2r(\gamma)}})^{2s_n},$$

我们可以找到一个 N_1，使得

$$|N_1| \leq \binom{d}{2s_n}(\frac{4r(\gamma)}{\epsilon})^{2s_n}.$$

考虑

$$N = \bigcup_{i=1}^{\lceil 2r(\gamma)/\epsilon \rceil} \frac{i}{\lceil 2r(\gamma)/\epsilon \rceil} \cdot N_1, \tag{7.8}$$

then

$$\begin{aligned} &P(\sup_{h\in N}\frac{B(h)}{\|h\|_2} \geq \frac{C\sqrt{2s_n(s_n+1)log(d)}\alpha_n}{\sqrt{n}}) \\ \overset{p}{\leq}&\frac{4r(\gamma)}{\epsilon}\binom{d}{2s_n}(\frac{4r(\gamma)}{\epsilon})^{2s_n}2d^{-\frac{2(s_n+1)C^2}{M_6}} \leq 2(\frac{4r(\gamma)}{\epsilon}d^{1-\frac{C^2}{M_6}})^{2s_n+1}. \end{aligned} \tag{7.9}$$

当 $\|h_1\|_2 = \|h_2\|_2$ 时，

$$\begin{aligned} &|\frac{B(h_1)}{\|h_1\|_2} - \frac{B(h_2)}{\|h_2\|_2}| = \frac{1}{\|h_1\|_2}|B(h_1) - B(h_2)| \\ =&\frac{1}{\|h_1\|_2}|\frac{1}{n(n-1)}\sum_{i\neq j}(w_i - w_j)F((x_i - x_j)^T(\beta^*_{r(\gamma),\alpha_n} - h_1)\alpha_n) \\ &- E[(W_i - W_j)F((X_i - X_j)^T(\beta^*_{r(\gamma),\alpha_n} - h_1)\alpha_n)] \\ &- \frac{1}{n(n-1)}\sum_{i\neq j}(w_i - w_j)F((x_i - x_j)^T(\beta^*_{r(\gamma),\alpha_n} - h_2)\alpha_n) \\ &+ E[(W_i - W_j)F((X_i - X_j)^T(\beta^*_{r(\gamma),\alpha_n} - h_2)\alpha_n)]| \\ \leq&\frac{M_2\alpha_n}{2n(n-1)}\|D(\frac{h_1}{\|h_1\|_2} - \frac{h_2}{\|h_2\|_2})\|_1 + \frac{M_2\alpha_n}{2}\cdot E|(X_i^T - X_j^T)(\frac{h_1}{\|h_1\|_2} - \frac{h_2}{\|h_2\|_2})| \\ \leq&\frac{\frac{1}{4}M_2\alpha_n}{\sqrt{\frac{n(n-1)}{2}}}\|D(\frac{h_1}{\|h_1\|_2} - \frac{h_2}{\|h_2\|_2})\|_2 + M_2\alpha_n \cdot E|X_i^T(\frac{h_1}{\|h_1\|_2} - \frac{h_2}{\|h_2\|_2})|, \end{aligned} \tag{7.10}$$

其中 D 在引理 8 中定义。根据引理 8，我们有

$$\begin{aligned}
&\sup_{\|h_1-h_2\|_0\le 4s_n,\|h_1\|_2=\|h_2\|_2,\|\frac{h_1}{\|h_1\|_2}-\frac{h_2}{\|h_2\|_2}\|_2\le\epsilon} |\frac{B(h_1)}{\|h_1\|_2}-\frac{B(h_2)}{\|h_2\|_2}| \\
\le&\frac{1}{4}M_2\alpha_n\epsilon\cdot(\sqrt{M_5 s_n}+4E[\sup_{h\epsilon\mathbb{R}^d:\|h\|_0\le 4s_n,\|h\|_2=1}|X_i^T h|]) \\
\le&\frac{1}{4}M_2\alpha_n\epsilon\cdot(\sqrt{M_5 s_n}+8M_4\sqrt{s_n}). \qquad (7.11)
\end{aligned}$$

它界定了当 $\|h_1\|_2=\|h_2\|_2$ 时，${}^{B}{}_{(h_1)}\overline{\|h_1\|_2}$ 与 ${}^{B}{}_{(h_2)}\overline{\|h_2\|_2}$ 之间的差值。接下来，考虑方向相同但 $\|h_1\|_2\neq\|h_2\|_2$ 的 h_1 和 h_2。若 $\|h\|_2=1$ 且 $\|h\|_0\le 2s_n$，则有

$$\begin{aligned}
&|(\frac{1}{t}[F((x_i-x_j)^T(\beta^*_{r(\gamma),\alpha_n}-th)\alpha_n)-F((x_i-x_j)^T\beta^*_{r(\gamma),\alpha_n}\alpha_n)])'| \\
=&|\frac{-1}{t}F'((x_i-x_j)^T(\beta^*_{r(\gamma),\alpha_n}-th)\alpha_n)\alpha_n(x_i-x_j)^T h \\
&-\frac{1}{t^2}[F((x_i-x_j)^T(\beta^*_{r(\gamma),\alpha_n}-th)\alpha_n)-F((x_i-x_j)^T\beta^*_{r(\gamma),\alpha_n}\alpha_n)]| \\
=&|\frac{-1}{t}F'((x_i-x_j)^T(\beta^*_{r(\gamma),\alpha_n}-th)\alpha_n)\alpha_n(x_i-x_j)^T h \\
&-\frac{1}{t^2}[F'((x_i-x_j)^T(\beta^*_{r(\gamma),\alpha_n}-th)\alpha_n)(-t(x_i-x_j)^T h\alpha_n) \\
&+\frac{1}{2}F''((x_i-x_j)^T(\beta^*_{r(\gamma),\alpha_n}-\xi h)\alpha_n)\cdot(-t(x_i-x_j)^T h\alpha_n)^2]| \\
=&|\frac{1}{2}F''((x_i-x_j)^T(\beta^*_{r(\gamma),\alpha_n}-\xi h)\alpha_n)\cdot\alpha_n^2((x_i-x_j)^T h)^2|\le c_6\alpha_n^2M_4^2\|h\|_1^2\le 2c_6\alpha_n^2M_4^2 s_n,
\end{aligned}$$

因此,

$$
\begin{aligned}
&|\frac{B(t_1h)}{t_1}-\frac{B(t_2h)}{t_2}|\\
\leq&\frac{2M_2}{n(n-1)}\sum_{i\neq j}|[F((x_i-x_j)^T(\beta^*_{r(\gamma),\alpha_n}-t_1h)\alpha_n)-F((x_i-x_j)^T\beta^*_{r(\gamma),\alpha_n}\alpha_n)]/t_1\\
&-[F((x_i-x_j)^T(\beta^*_{r(\gamma),\alpha_n}-t_2h)\alpha_n)-F((x_i-x_j)^T\beta^*_{r(\gamma),\alpha_n}\alpha_n)]/t_2|\\
&+2M_2\cdot E|[F((X_i-X_j)^T(\beta^*_{r(\gamma),\alpha_n}-t_1h)\alpha_n)-F((X_i-X_j)^T\beta^*_{r(\gamma),\alpha_n}\alpha_n)]/t_1\\
&-[F((X_i-X_j)^T(\beta^*_{r(\gamma),\alpha_n}-t_2h)\alpha_n)-F((X_i-X_j)^T\beta^*_{r(\gamma),\alpha_n}\alpha_n)]/t_2|\leq 8c_6M_2M_4^2s_n\alpha_n^2|t_1-t_2|
\end{aligned}
\tag{7.12}
$$

对于任意 $h\in\{h:\|h\|_0\leq 2s_n,\|h\|_2\leq 2r(\gamma)\}$，我们可以找到一个整数 $i\in\{1,2,...,\lceil 2r(\gamma)/\epsilon\rceil\}$以及 $h_0\in N$，使得 $|\|h\|_2-\frac{2ir(\gamma)}{\lceil 2r(\gamma)/\epsilon\rceil}|\leq\epsilon$, $\|h_0\|_2=\frac{2ir(\gamma)}{\lceil 2r(\gamma)/\epsilon\rceil}$ 以及 $\|h_0-\frac{2ir(\gamma)h}{\lceil 2r(\gamma)/\epsilon\rceil\|h\|_2}\|_2\leq\epsilon$，其中$N$ 定义见 (7.8)。

由于

$$
|\frac{B(h)}{\|h\|_2}-\frac{B(h_0)}{\|h_0\|_2}|\leq|\frac{B(h)}{\|h\|_2}-\frac{B(\frac{2ir(\gamma)h}{\lceil 2r(\gamma)/\epsilon\rceil\|h\|_2})}{\|\frac{2ir(\gamma)h}{\lceil 2r(\gamma)/\epsilon\rceil\|h\|_2}\|_2}|+|\frac{B(\frac{2ir(\gamma)h}{\lceil 2r(\gamma)/\epsilon\rceil\|h\|_2})}{\|\frac{2ir(\gamma)h}{\lceil 2r(\gamma)/\epsilon\rceil\|h\|_2}\|_2}-\frac{B(h_0)}{\|h_0\|_2}|,
$$

由（7.11）和（7.12）可得

$$
\sup_{\|h\|_0\leq 2s_n,\|h\|_2\leq 2r(\gamma)}\frac{B(h)}{\|h\|_2}\leq\sup_{h\in N}\frac{B(h)}{\|h\|_2}+c_7M_2\alpha_n^2\epsilon,
$$

其中 $c_7=\frac{1}{4}\sqrt{M_5s_n}+2M_4\sqrt{s_n}+8M_4^2s_nc_6$。

令 $\epsilon=\sqrt{\frac{2s_n}{(s_n+1)log(d)\sqrt{n}c_7\alpha_n}}$，并利用 (7.9)，可得

$$
\begin{aligned}
&P(\sup_{\|h\|_0\leq 2s_n,\|h\|_2\leq 2r(\gamma)}\frac{B(h)}{\|h\|_2}\geq\frac{(C-1)\sqrt{2s_n(s_n+1)log(d)}M_2\alpha_n}{\sqrt{n}})\\
\leq&2(\frac{4r(\gamma)}{\epsilon}d^{1-\frac{C^2}{2M_6}})^{2s_n+1}=2(\frac{\sqrt{n}c_7\alpha_n4r(\gamma)}{\sqrt{2s_n(s_n+1)log(d)}}d^{1-\frac{C^2}{2M_6}})^{2s_n+1},
\end{aligned}
\tag{7.13}
$$

并令 $\alpha_n \ll n$。取足够大的 C，可得

$$\sup_{\|h\|_0 \le 2s_n, \|h\|_2 \le 2r(\gamma)} \frac{B(h)}{\|h\|_2} \overset{p}{\lesssim} \frac{\sqrt{s_n^2 log(d)}}{\sqrt{n}} \alpha_n.$$

T假设3下离散情形的证明基本相同 e. □

定理5的证明。定理5的证明与 Han 等人（2017）中引理4.10的证明非常相似，因此这里仅给出证明概要。根据引理 1,

$$L_n(\beta^*_{r(\gamma),\alpha_n} + \hat{\Delta}) - L_n(\beta^*_{r(\gamma),\alpha_n}) \ge L(\beta^*_{r(\gamma),\alpha_n} + \hat{\Delta}) - L(\beta^*_{r(\gamma),\alpha_n}) - c_5 \alpha_n^{-2}. \tag{7.14}$$

ˆ根据 $L(\beta^*_{r(\gamma),\ \alpha_n}) - L(\beta^*)$ 和 $L(\beta^*_{r(\gamma),\ \alpha_n} + \Delta) - L(\beta^*)$ 的泰勒展开,

$$L(\beta^*_{r(\gamma),\alpha_n} + \hat{\Delta}) - L(\beta^*_{r(\gamma),\alpha_n}) \ge \frac{1}{2}\hat{\Delta}^T \Gamma \hat{\Delta}(1 + o(1)) - \hat{\Delta}^T \Gamma(\beta^* - \beta^*_{r(\gamma),\alpha_n})(1 + o(1)), \tag{7.15}$$

ˆ这里的 Δ 和 $\beta^* - \beta^*_{r(\gamma),\ \alpha_n}$ 不包括第一个坐标。然后利用定理2和假设1，我们有

$$L_n(\beta^*_{r(\gamma),\alpha_n} + \hat{\Delta}) - L_n(\beta^*_{r(\gamma),\alpha_n}) \ge c_8 \|\hat{\Delta}\|_2^2 - c_9 \alpha_n^{-1} \|\hat{\Delta}\|_2 - c_{10} \alpha_n^{-2} \tag{7.16}$$

其中 c_8、c_9、c_{10} 是一些适当的常数。

ˆ 与 Han 等人（2017）中定理 4.8 的证明类似，只要 $\lambda_n \ge 2P^*(\nabla L_n(\beta^*_{r(\gamma),\alpha_n}))$ ˆ和 $L_n(\cdot)$ 局部凸且可微，其中 $P^*(\cdot)$ 是 $P(\cdot)$ 的对偶范数，我们就有

$$\|\hat{\Delta}_{S_n^c}\|_1 \le 3\|\hat{\Delta}_{S_n}\|_1$$

那么，根据引理 3、引理 4 和 (7.16)，定义

$$
\begin{aligned}
&\delta\hat{L}_n(\hat{\Delta}, \beta^*_{r(\gamma),\alpha_n}) := \hat{L}_n(\beta^*_{r(\gamma),\alpha_n} + \hat{\Delta}) - \hat{L}_n(\beta^*_{r(\gamma),\alpha_n}) - \langle \nabla\hat{L}_n(\beta^*_{r(\gamma),\alpha_n}), \hat{\Delta}\rangle \\
\geq & c_8\|\hat{\Delta}\|_2^2 - c_9\alpha_n^{-1}\|\hat{\Delta}\|_2 - c_{10}\alpha_n^{-2} - \sup_{\hat{\beta}\in\mathcal{H}} B(\hat{\Delta}) - \|\hat{\Delta}\|_1\|\nabla\hat{L}_n(\beta^*_{r(\gamma),\alpha_n})\|_\infty \\
\overset{p}{\geq} & c_8\|\hat{\Delta}\|_2^2 - c_9\alpha_n^{-1}\|\hat{\Delta}\|_2 - c_{10}\alpha_n^{-2} - 2(C-1)\alpha_n\frac{\sqrt{s_n^2 log(d)}}{\sqrt{n}}\|\hat{\Delta}\|_2 - 2\sqrt{2}M_2M_4\|\hat{\Delta}\|_1\alpha_n\sqrt{\frac{log(d)}{n}} \\
\geq & c_8\|\hat{\Delta}\|_2^2 - c_9\alpha_n^{-1}\|\hat{\Delta}\|_2 - c_{10}\alpha_n^{-2} - 2(C-1)\alpha_n\frac{\sqrt{s_n^2 log(d)}}{\sqrt{n}}\|\hat{\Delta}\|_2 - 8\sqrt{2}M_2M_4\|\hat{\Delta}_{S_n}\|_1\alpha_n\sqrt{\frac{log(d)}{n}} \\
\geq & c_8\|\hat{\Delta}\|_2^2 - \|\hat{\Delta}\|_2(c_9\alpha_n^{-1} + (2C - 2 + 8\sqrt{2}M_2M_4)\alpha_n\sqrt{\frac{log(d)s_n^2}{n}}) - c_{10}\alpha_n^{-2}.
\end{aligned}
\tag{7.17}
$$

由于 $\Psi(\bar{S}_n) = \sup_{v\in\bar{S}_n\setminus 0} \frac{P(v)}{\|v\|_2} \leq \sqrt{s_n}$，我们可以验证 Han 等人（2017）中的定理 4.8 的假设条件均得到满足。令 $\lambda_n \asymp \alpha_n\sqrt{log_{(d)}\ n}$。Han 等人（2017）中的定理 4.8 表明 $\|\hat{\Delta}\|_2^2 \overset{p}{\leq} (2\lambda_n\Psi(\bar{S}_n) + c_9\alpha_n^{-1} + (2C - 2 + 8\sqrt{2}M_2M_4)\alpha_n\sqrt{log_{(d\ s_n^2)}/n})^2/c_8^2 + 2c_{10}\alpha_n^{-2}/c_8 \overset{p}{\lesssim} \alpha_n^2 log(d)s_n^2/n + \alpha_n^{-2}$。

□

7.2. 第 3.2 节的证明

定理 7 的证明。回顾假设3中 $\mathcal{M}$ 的定义。定义

$$
\varepsilon = \min_{x_i^T\beta^* > x_j^T\beta^*} (x_i - x_j)^T\beta^*.
$$

$\mathcal{U}_t = \{\beta:\ \beta_1 = 1、\ \forall x_i、\ x_j,\ 若\ (x_i^T - x_j^T)\ \beta^* > 0,\ 则\ (x_i^T - x_j^T)_S\beta_S + (x_i^T - x_j^T)_{S\setminus 1}\mathcal{M}^T\beta_{S^c} > t\}$。令 $\alpha_n \geq 1$，则根据 ϵ的定义，$\beta^* \in \mathcal{U}_{\frac{\epsilon}{\alpha_n}}$，因此 $\mathcal{U}_{\frac{\epsilon}{\alpha_n}}$ 非空。接下来分三步证明该定理。

首先，我们证明：对于足够大的 α_n， $\beta^*_{r,\ \alpha_n} \in \mathcal{U}_{\frac{\epsilon}{\alpha_n}}$。显然

$$\begin{aligned} C_n(\beta) &= E[(Q(X_i^T\beta^*) - Q(X_j^T\beta^*))(2F((X_i^T - X_j^T)\beta\alpha_n) - 1)_{\mathbb{I}}((X_i^T - X_j^T)\beta^* > 0)] \\ &\leq E[(Q(X_i^T\beta^*) - Q(X_j^T\beta^*)) \cdot 1 \cdot {}_{\mathbb{I}}((X_i^T - X_j^T)\beta^* > 0)] = C(\beta^*), \end{aligned} \tag{7.18}$$

并且存在一个正数常量 c_{11}，使得

$$\begin{aligned} C(\beta^*) - C_n(\beta^*) =& 2E[(Q(X_j^T\beta^*) - Q(X_i^T\beta^*))F((X_i^T - X_j^T)\beta^*\alpha_n)_{\mathbb{I}}((X_i^T - X_j^T)\beta^* < 0)] \\ \leq& 2E[(Q(X_j^T\beta^*) - Q(X_i^T\beta^*))_{\mathbb{I}}(X_j^T\beta^* > X_i^T\beta^*)]/(1 + e^{\alpha_n\varepsilon}) \leq c_{11}e^{-\varepsilon\alpha_n}. \end{aligned} \tag{7.19}$$

若 $\beta \notin \mathcal{U}_{\frac{\epsilon}{\alpha_n}}$，则 $\exists x_i^T, x_j^T, s.t. (x_i^T - x_j^T)\beta^* > 0$，但 $(x_i^T - x_j^T)_S\beta_S + (x_i^T - x_j^T)_{S\setminus 1}\mathcal{M}^T\beta_{S^c} < \frac{\varepsilon}{\alpha_n}$。定义 $\delta = \min_{x_i^T\beta^* > x_j^T\beta^*}(1 - F(\varepsilon))(Q(x_i^T\beta^*) - Q(x_j^T\beta^*))p(x_{i,\ s})p(x_{j,\ s})$，其中 $p(X_S)$ 是 X_S 的概率质量函数（p.m.f.），则 $\delta > 0$ 独立于 d。利用假设3中的 (B1)，我们有

$$\begin{aligned} & C(\beta^*) - C_n(\beta) \\ =& E[(Q(X_i^T\beta^*) - Q(X_j^T\beta^*))(2 - 2F((X_i^T - X_j^T)\beta\alpha_n))_{\mathbb{I}}((X_i^T - X_j^T)\beta^* > 0)] \\ =& E[(Q(X_i^T\beta^*) - Q(X_j^T\beta^*))(2 - 2F((X_i - X_j)_S^T\beta_S\alpha_n + (X_i - X_j)_{S\setminus 1}^T\mathcal{M}^T\beta_{S^c}\alpha_n \\ & + [(X_i - X_j)_{S^c} - \mathcal{M}(X_i - X_j)_{S\setminus 1}]^T\beta_{S^c}\alpha_n))_{\mathbb{I}}((X_i^T - X_j^T)\beta^* > 0)] \\ \geq& E[(Q(X_i^T\beta^*) - Q(X_j^T\beta^*))(1 - F((X_i - X_j)_S^T\beta_S\alpha_n + (X_i - X_j)_{S\setminus 1}^T\mathcal{M}^T\beta_{S^c}\alpha_n)) \\ & {}_{\mathbb{I}}((X_i^T - X_j^T)\beta^* > 0)] \geq p(x_{i,S})p(x_{j,S})(Q(x_i^T\beta^*) - Q(x_j^T\beta^*))(1 - F(\varepsilon)) \geq \delta. \end{aligned} \tag{7.20}$$

设 $\alpha_n > \frac{log_{(c_{11}) - log(\delta)\varepsilon}}{}$，则 $C(\beta^*) - C_n(\beta^*) \leq c_{11}e^{-\varepsilon\alpha_n} < \delta \leq C(\beta^*) - C_n(\beta)$，这意味着 $\beta \notin \mathcal{U}_{\frac{\epsilon}{\alpha_n}}$ 不是 $C_n(\cdot)$ 的最大化点。因此 $\beta^*_{r,\ \alpha_n} \in \mathcal{U}_{\frac{\epsilon}{\alpha_n}}$。

Sec其次，我们证明 $\beta^*_{r,\ \alpha_n,\ S^c}$ 的各项均为 0。根据假设3中的 (B1) ，

$$
\begin{aligned}
&C_n(\beta)\\
=&E[(Q(X_i^T\beta^*)-Q(X_j^T\beta^*))(2F((X_i-X_j)^T\beta\alpha_n)-1)_{\mathbb{I}}((X_i-X_j)^T\beta^*>0)]\\
=&E[(Q(X_i^T\beta^*)-Q(X_j^T\beta^*))(2F((X_i-X_j)_S^T\beta_S\alpha_n+(X_i-X_j)_{S\setminus 1}^T\mathcal{M}^T\beta_{S^c}\alpha_n\\
&+[(X_i-X_j)_{S^c}-\mathcal{M}(X_i-X_j)_{S\setminus 1}]^T\beta_{S^c}\alpha_n)+2F((X_i-X_j)_S^T\beta_S\alpha_n+(X_i-X_j)_{S\setminus 1}^T\mathcal{M}^T\beta_{S^c}\alpha_n\\
&-[(X_i-X_j)_{S^c}-\mathcal{M}(X_i-X_j)_{S\setminus 1}]^T\beta_{S^c}\alpha_n)-2)\cdot{}_{\mathbb{I}}((X_i-X_j)^T\beta^*>0)\cdot\\
&{}_{\mathbb{I}}([(X_i-X_j)_{S^c}-\mathcal{M}(X_i-X_j)_{S\setminus 1}]^T\beta_{S^c}>0)].
\end{aligned}
\tag{7.21}
$$

令公式中的 $\beta=\beta^*_{r,\ \alpha_n}$。注意到 $\forall a>0,t>0,\frac{\partial}{\partial t}(F(a+t)+F(a-t))=e^{-a+t}\frac{(e^t+e^{-a})^2-e^{-a+t}}{(1+e^{-a+t})^2}$。由于 $e^t+e^{-a}-(1+e^{-a+t})=(e^t-1)(1-e^{-a})>0$，我们知道 $(F(a+t)+F(a-t))$ 在 $a>0$ 时于 $t=0$ 处取得最大值。因此，我们有

$$
\begin{aligned}
&C_n(\beta^*_{r,\alpha_n})\\
\le&2E[(Q(X_i^T\beta^*)-Q(X_j^T\beta^*))(2F((X_i-X_j)_S^T\beta^*_{r,\alpha_n,S}\alpha_n+(X_i-X_j)_{S\setminus 1}^T\mathcal{M}^T\beta^*_{r,\alpha_n,S^c}\alpha_n)-1)\cdot\\
&{}_{\mathbb{I}}((X_i-X_j)^T\beta^*>0)_{\mathbb{I}}([(X_i-X_j)_{S^c}-\mathcal{M}(X_i-X_j)_{S\setminus 1}]^T\beta^*_{r,\alpha_n,S^c}>0)]\\
=&E[(Q(X_i^T\beta^*)-Q(X_j^T\beta^*))(2F((X_i-X_j)_S^T\beta^*_{r,\alpha_n,S}\alpha_n+(X_i-X_j)_{S\setminus 1}^T\mathcal{M}^T\beta^*_{r,\alpha_n,S^c}\alpha_n)-1)\cdot\\
&{}_{\mathbb{I}}((X_i-X_j)^T\beta^*>0)],
\end{aligned}
\tag{7.22}
$$

这意味着，满足 $\beta_1=1,\beta_{S\setminus 1}=\beta^*_{r,\ \alpha_n,\ S\setminus 1}+\mathcal{M}^T\beta^*_{r,\ \alpha_n,\ S^c}$ 且 $\beta_{S^c}=\vec{0}$ 的 β 会使 $C_n(\beta)$ 不小于 $C_n(\beta^*_{r,\ \alpha_n})$。因此，不失一般性，我们可以设 $\beta^*_{r,\ \alpha_n,\ S^c}=\vec{0}$，因为 $\beta^*_{r,\ \alpha_n}$ 会使 $C_n(\cdot)$ 取得最大值。

^^^^^ 最后，定义 $F(\Delta)=L_n(\beta^*_{r,\alpha_n}+\Delta)+\lambda_n\|\beta^*_{r,\alpha_n}+\Delta\|_1-L_n(\beta^*_{r,\alpha_n})-\lambda_n\|\beta^*_{r,\alpha_n}\|_1$，则

$$
\begin{aligned}
F(\hat{\Delta}) \geq & C_n(\beta^*_{r,\alpha_n}) - C_n(\hat{\beta}_{r,\alpha_n}) - \sup_{\hat{\beta}\in\mathcal{H}} B(\hat{\Delta}) + \lambda_n(\|\beta^*_{r,\alpha_n} + \hat{\Delta}\|_1 - \|\beta^*_{r,\alpha_n}\|_1) \\
= & C(\beta^*) - C_n(\hat{\beta}_{r,\alpha_n}) - (C(\beta^*) - C_n(\beta^*_{r,\alpha_n})) - \sup_{\hat{\beta}\in\mathcal{H}} B(\hat{\Delta}) + \lambda_n(\|\beta^*_{r,\alpha_n} + \hat{\Delta}\|_1 - \|\beta^*_{r,\alpha_n}\|_1).
\end{aligned} \tag{7.23}
$$

若 $\hat{\beta}_{r,\alpha_n} \notin \mathcal{U}_{\frac{\epsilon}{\alpha_n}}$，利用（7.19）、（7.20）以及 $\beta^*_{r,\alpha_n,S^c} = \vec{0}$ 这一事实，

$$
\begin{aligned}
F(\hat{\Delta}) \geq & \delta - c_{11}e^{-\varepsilon\alpha_n} - \sup_{\hat{\beta}\in\mathcal{H}} B(\hat{\Delta}) + \lambda_n(\|\beta^*_{r,\alpha_n} + \hat{\Delta}\|_1 - \|\beta^*_{r,\alpha_n}\|_1) \\
\geq & \delta - c_{11}e^{-\varepsilon\alpha_n} - \sup_{\hat{\beta}\in\mathcal{H}} B(\hat{\Delta}) + \lambda_n(\|\beta^*_{r,\alpha_n,S} + \hat{\Delta}_S\|_1 + \|\hat{\Delta}_{S^c}\|_1 - \|\beta^*_{r,\alpha_n,S}\|_1),
\end{aligned} \tag{7.24}
$$

然后由引理 4,

$$
\begin{aligned}
F(\hat{\Delta}) \overset{p}{\geq} & \delta - c_{11}e^{-\varepsilon\alpha_n} + \lambda_n(-\|\hat{\Delta}_S\|_1 + \|\hat{\Delta}_{S^c}\|_1) - c_{12}\|\hat{\Delta}\|_2\alpha_n\sqrt{log(d)s^2/n} \\
\geq & -\lambda_n\|\hat{\Delta}_S\|_1 + \delta - c_{11}e^{-\varepsilon\alpha_n} - c_{12}\|\hat{\Delta}\|_2\alpha_n\sqrt{log(d)s^2/n} \\
\geq & -\lambda_n\|\hat{\Delta}\|_2\sqrt{s} + \delta - c_{11}e^{-\varepsilon\alpha_n} - c_{12}\|\hat{\Delta}\|_2\alpha_n\sqrt{log(d)s^2/n}.
\end{aligned} \tag{7.25}
$$

选择 $\lambda_n = \frac{\delta}{3r\sqrt{s}}$ 和 $\alpha_n = max\ (1, \frac{log\ (c_{11})\ -log\ (\delta)\ +log\ (3)\ \varepsilon}{log\ (d)\ n\ \varepsilon})$，则其满足 $\alpha_n \geq \frac{log\ (c_{11})}{-log\ (\delta)\ \varepsilon}$以及 $\lambda_n \geq \alpha_n\sqrt{log\ (d)\ n}$。因此

$$
F(\hat{\Delta}) \overset{p}{\geq} \delta/3 - c_{12}\|\hat{\Delta}\|_2\alpha_n\sqrt{log(d)s^2/n} \geq \delta/3 - max(1, \frac{log(3c_{11}/\delta)}{\varepsilon})c_{12}r\sqrt{log(d)s^2/n} \tag{7.26}
$$

若 $log(d)s^2/n \longrightarrow 0$，则 $F(\hat{\Delta}) \overset{p}{>} 0$，这与 $\hat{\beta}_{r,\ \alpha_n}$ 在 $\mathcal{H}$ 中使 $\hat{L}_n(\beta) + \lambda_n P(\beta)$ 最小化这一事实矛盾。因此 $P(\beta_{r,\ \alpha_n} \in \mathcal{U}_{\frac{\epsilon}{\alpha_n}}) \longrightarrow 1$。

现在考虑 β_{r,α_n,S^c}。将 β_{r,α_n} 中的 S^c 部分截断为 0，并将这个新向量记为 $\tilde{\beta}_{r,\alpha_n}$。根据定义，$L_n(\beta_{r,\alpha_n}) + \lambda_n\|\beta_{r,\alpha_n}\|_1 \leq L_n(\tilde{\beta}_{r,\alpha_n}) + \lambda_n\|\tilde{\beta}_{r,\alpha_n}\|_1$，因此 $\lambda_n\|\beta_{r,\alpha_n,S^c}\|_1 \leq L_n(\tilde{\beta}_{r,\alpha_n}) - L_n(\beta_{r,\alpha_n})$。与 (7.22) 类似，我们知道 $L_n(\tilde{\beta}_{r,\alpha_n}) - L_n(\beta_{r,\alpha_n}) \leq 0$，因此

$$
\lambda_n\|\hat{\beta}_{r,\alpha_n,S^c}\|_1 \leq B(\hat{\beta}_{r,\alpha_n} - \tilde{\hat{\beta}}_{r,\alpha_n}) \overset{p}{\lesssim} (\frac{s^2\log(d)}{n})^{1/2}, \tag{7.27}
$$

因此 $\|\hat{\beta}_{r,\ \alpha_n,\ S^c}\|_1 \overset{p}{\lesssim} (s^3 \text{对数}(d)/n)^{1/2}$。 □

参考文献

Blatt, D., Murphy, S., 和 Zhu, J.（2004）。用于近似规划的 A学习。

Fan, C., Lu, W., Song, R., 和 Zhou, Y.（2017年）。用于估计最优个体化治疗方案的一致性辅助学习。英国皇家统计学会会刊：B辑（统计方法学），79(5):1565 – 1582。

Fava, M., Rush, J., Kupfer, D. J., Trivedi, M. H., Nierenberg, A. A., Thase, M. E., Sackeim, H. A., Quitkin, F. M., Wisniewski, S., Lavori, P. W., Rosenbaum, J. F., Dunner, D. L., 和 STAR*D研究团队, G.（2003年）。抑郁症缓解序贯治疗替代方案（STAR*D）研究的背景与理论依据。药物治疗：反应预测因素，26(2)。

Han, F.、Ji, H.、Ji, Z.和Wang, H.（2017年）。一种具有可证明性的高维广义回归平滑方法及其在基因组学中的应用。《电子统计学期刊》，11(2)：4347–4403。

Han, A.、K.（1987年）。广义回归模型的非参数分析：最大秩相关估计量。《计量经济学期刊》，35(2-3)：303–316。

Liang, S.、Lu, W.、Song, R.和Wang, L.（2018年）。用于最优治疗决策的稀疏一致性辅助学习。《机器学习研究期刊》，18(154-234)：1–26。

Lu, W.、Zhang, H. 和 Zeng, D.（2013）。最优治疗决策的变量选择。《医学研究统计方法》，(5):493。

Murphy, S. A.（2003年）。最优动态治疗方案。《英国皇家统计学会杂志》B辑（统计方法学），65(2):331 – 366。Pitcan, Y.（2017年）。关于 U 统计量集中不等式的注记。Rubin, D. B.（1980年）。实验数据的随机化分析：关于 Fisher 随机化检验的评论。《美国统计学会会刊》，75(371):591 – 593。Shi, C.、Fan、Song、Lu, W.（2018年）。用于最优动态治疗方案的高维A学习。《统计学年鉴》，46(3):925–957。

Song, R., Kosorok, M., Zeng, D., Zhao, Y., Laber, E., and Yuan, M.（2015年）。关于采用惩罚性结果加权学习进行最优个体化治疗选择的稀疏表示。

Song, R., Wang, W., Zeng, D., and Kosorok, M. R.（2011年）。动态治疗方案的惩罚Q学习。Watkins, C. J. C. H.（1989年）。《从延迟奖励中学习》。博士学位论文，剑桥国王学院，英国。Watkins, C. J. C. H. and Dayan, P.（1992年）。Q学习。《机器学习》，8(3-4):279。Zhang, B., Tsiatis, A., Laber, E., and Davidian, M.（2012年）。一种估计最优治疗方案的稳健方法。《生物统计学》，68(4):1010 – 1018。Zhao, Y., Zeng, D., Rush A., J., and Kosorok Michael, R.（2012年）。利用结果加权学习估计个体化治疗规则。《美国统计学会会刊》，107(499):1106 – 1118。

Zhu, W., Zeng, D., and Song, R.（2019年）。动态治疗方案高维Q学习中价值函数的恰当推断。《美国统计学会会刊》，114(527):1404–1417。