经验似然校准的主动统计推断

Dec.27 2025

Contents

1 实验代码需修改内容

1. 数据集拆分出训练集（比例？），剩余部分再拆分出一个有偏数据集（拆分方法？拆分比例？）
2. 确定新的方差估计（包括新估计量的方差估计、有偏数据集下原方差估计也失效）
3. 增加p_i估计函数（确定所需距离函数以及所需算法）
4. 改变$\pi_i^{(1)}$的估计方法以及对应新的估计量形式
5. 考虑实验设计的改变，需要对比的内容有哪些
   （数据分割方法？数据集偏离程度（可类似于讨论τ、以及偏向low、high、extreme）？
   baseline选择？metric选择？budget是不是需要改变？）

2 初步实验结果

数据集：Friedman(synthetic)
拆分方法：0.2比例用于估计模型（视为有标签数据集），
　　剩余0.8按照**数据集分割**拆分出0.4样本量构建无标签且特征有偏数据集。给定的均值为0.8的无标签数据集

在数据集偏离总体均值较低时， **偏向极端值**近似于没有无偏的设定，
此时的结果基本同traditional_active方法，也与正常分割数据集的结果相近。
　　但朝**单向偏离**时，很明显traditional_active方法的方差估计已经失效，
同时估计偏离总体标签均值很大，而新提出的方法EL_active的方差估计依旧有效，估计稳健。

在数据集偏离总体均值逐渐提高时， 面对**偏向极端值**的子数据集，各其余方法的方差估计也逐渐失效，
此时新提出的方法依旧有效，估计稳健，且RMSE较低，估计更稳健。
　　而对于**单向偏离**，很明显EL_active方法的方差估计也逐渐失效，但RMSE依旧很低
同时估计偏离总体标签均值很大，而新提出的方法EL_active的方差估计依旧有效，估计稳健。

2.1 偏离强度0.2

偏高

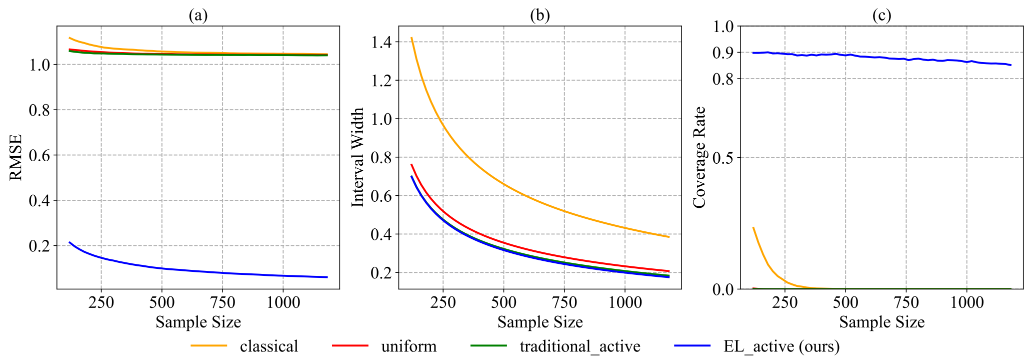

Figure 1: (a) RMSE, (b) 90% confidence interval width, and (c) empirical coverage rate

偏低

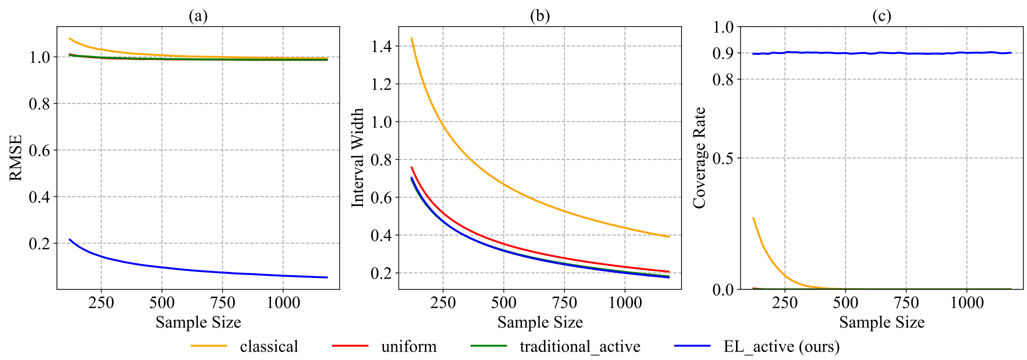

Figure 2: (a) RMSE, (b) 90% confidence interval width, and (c) empirical coverage rate

偏向极端值

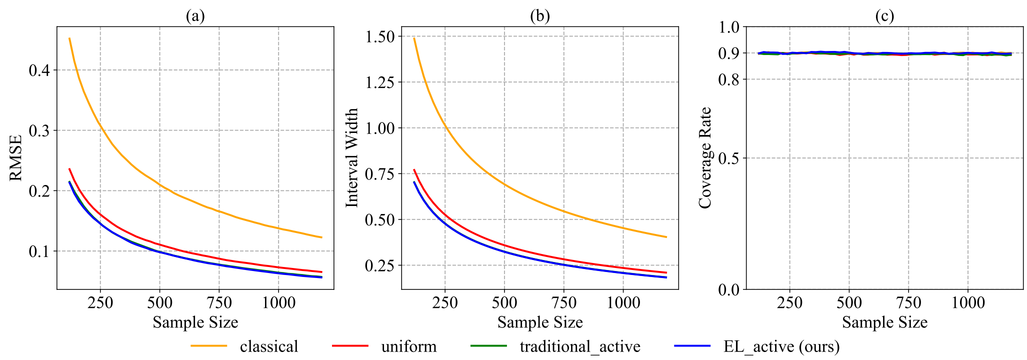

Figure 3: (a) RMSE, (b) 90% confidence interval width, and (c) empirical coverage rate

2.2 偏离强度0.5

偏高

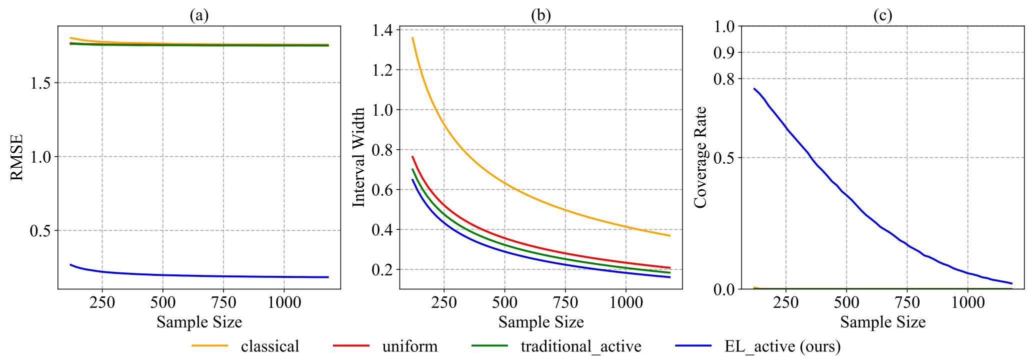

Figure 4: (a) RMSE, (b) 90% confidence interval width, and (c) empirical coverage rate

偏低

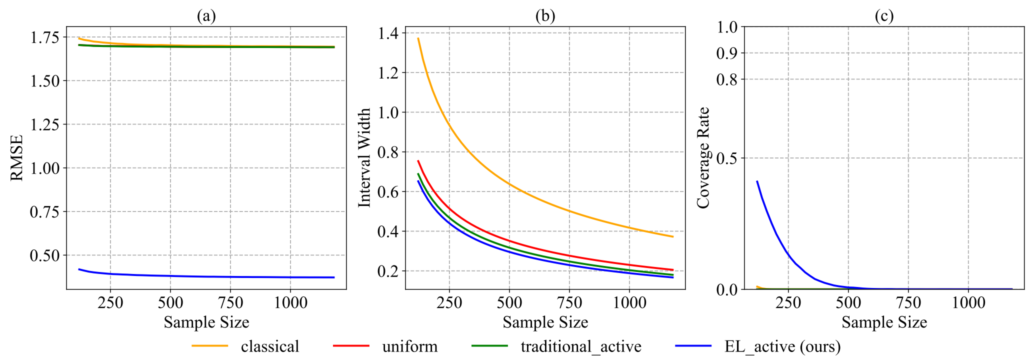

Figure 5: (a) RMSE, (b) 90% confidence interval width, and (c) empirical coverage rate

偏向极端值

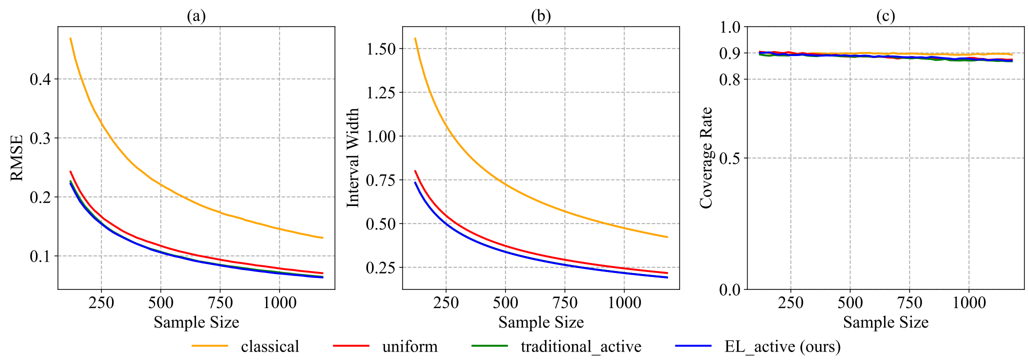

Figure 6: (a) RMSE, (b) 90% confidence interval width, and (c) empirical coverage rate

2.3 偏离强度1

偏高

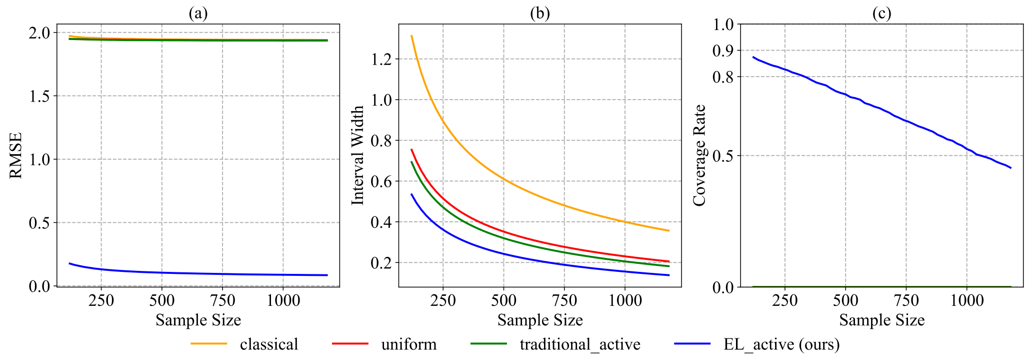

Figure 7: (a) RMSE, (b) 90% confidence interval width, and (c) empirical coverage rate

偏低

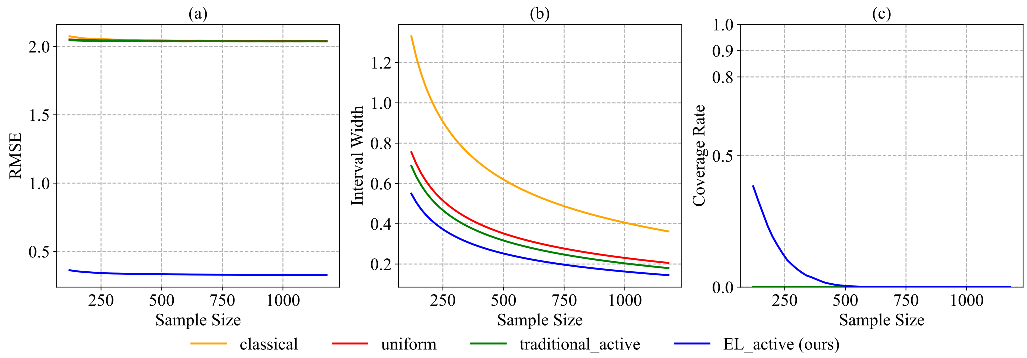

Figure 8: (a) RMSE, (b) 90% confidence interval width, and (c) empirical coverage rate

偏向极端值

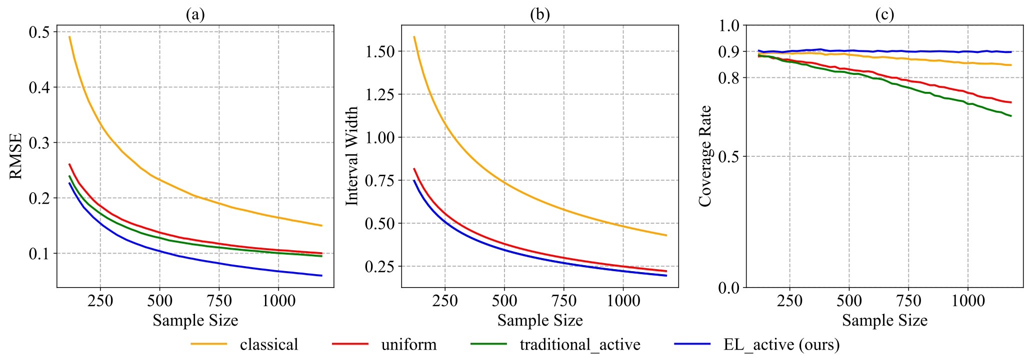

Figure 9: (a) RMSE, (b) 90% confidence interval width, and (c) empirical coverage rate

3 初步实验尝试

在此部分我主要添加了下面两个方面的代码用于新方法实验。其余设计基本同traditional_active（即active statistical inference）。

3.1 估计校准权重p_i

目标： 基于经验似然方法，求解满足校准约束的权重 $\{p_i\}_{i=1}^n$。

符号定义：

- $X_i \in \mathbb{R}^m$：第 i 个样本的特征向量
- $\mu_X \in \mathbb{R}^m$：已知的总体特征均值
- $d_i > 0$：初始设计权重（默认取 $d_i = 1$）
- $d_i^* = d_i / \sum_{j=1}^n d_j$：归一化设计权重

Step 1：数据中心化 定义中心化向量：

$$u_i = X_i - \mu_X, \quad i = 1, \ldots, n \tag{1}$$

此变换将原约束 $\sum_{i=1}^n p_i X_i = \mu_X$ 转化为 $\sum_{i=1}^n p_i u_i = 0$。

Step 2：经验似然优化问题 最大化经验似然函数：

$$\max_{p_1,\ldots,p_n} \sum_{i=1}^n d_i^* \ln\left(\frac{p_i}{d_i^*}\right) \tag{2}$$

约束条件：

$$\sum_{i=1}^n p_i = 1, \quad \sum_{i=1}^n p_i u_i = 0, \quad p_i > 0 \tag{3}$$

Step 3：Lagrangian 构造

$$\mathcal{L} = \sum_{i=1}^n d_i^* \ln(p_i) + \gamma\left(1 - \sum_{i=1}^n p_i\right) + \lambda^\top\left(-\sum_{i=1}^n p_i u_i\right) \tag{4}$$

其中 $\lambda \in \mathbb{R}^m$ 为拉格朗日乘数向量。

Step 4：一阶条件 对 p_i 求导：

$$\frac{\partial \mathcal{L}}{\partial p_i} = \frac{d_i^*}{p_i} - \gamma - \lambda^\top u_i = 0 \tag{5}$$

解得：

$$p_i = \frac{d_i^*}{\gamma + \lambda^\top u_i} \tag{6}$$

由约束 $\sum_{i=1}^n p_i = 1$ 可确定 $\gamma = 1$，因此：

$$p_i = \frac{d_i^*}{1 + \lambda^\top u_i} \tag{7}$$

Step 5：求解拉格朗日乘数 λ 将 p_i 代入约束 $\sum_{i=1}^n p_i u_i = 0$，定义:

$$g(\lambda) = \sum_{i=1}^{n} \frac{d_i^* u_i}{1 + \lambda^\top u_i} = 0 \tag{8}$$

Newton-Raphson 迭代 ：

梯度（Jacobian）：

$$\nabla g(\lambda) = -\sum_{i=1}^{n} \frac{d_i^* u_i u_i^\top}{(1 + \lambda^\top u_i)^2} \tag{9}$$

Newton 更新:

$$\lambda^{(t+1)} = \lambda^{(t)} - [\nabla g(\lambda^{(t)})]^{-1} g(\lambda^{(t)}) \tag{10}$$

令 $a_i = 1 + \lambda^\top u_i$，则:

$$D_1 = g(\lambda) = \sum_{i=1}^{n} \frac{d_i^* u_i}{a_i} \tag{11}$$

$$D_D = \nabla g(\lambda) = -\sum_{i=1}^{n} \frac{d_i^* u_i u_i^\top}{a_i^2} \tag{12}$$

Newton 方向：$D_2 = D_D^{-1} D_1$

线搜索（步长控制）： 为确保更新后所有权重为正（即 $1+(\lambda-D_2)^\top u_i > 0$），需进行步长回退:

$$\text{若 } \min_i \left(1 + (\lambda - D_2)^\top u_i\right) \le 0, \text{ 则 } D_2 \leftarrow D_2/2 \tag{13}$$

算法：经验似然校准权重（多变量）

输入： 特征矩阵 $X \in \mathbb{R}^{n\times m}$，总体均值 $\mu_X \in \mathbb{R}^m$，设计权重 $\{d_i\}_{i=1}^n$，最大迭代次数 $T_{\max}$，收敛阈值 ϵ
输出： 校准权重 $\{p_i\}_{i=1}^n$

1. **初始化：**
   - 归一化设计权重：$d_i^* \leftarrow d_i / \sum_{j=1}^n d_j$
   - 中心化：$u_i \leftarrow X_i - \mu_X$，$i = 1, \dots, n$
   - 初始拉格朗日乘数：$\lambda^{(0)} \leftarrow 0 \in \mathbb{R}^m$
2. **Newton-Raphson 迭代：** 对 $t = 0, 1, \dots, T_{\max} - 1$：
   (a) 计算分母：$a_i \leftarrow 1 + (\lambda^{(t)})^\top u_i$，$\forall i$
   (b) 检查可行性：若 $\exists i$ 使 $a_i \leq 0$，则报错退出
   (c) 计算梯度：$D_1 \leftarrow \sum_{i=1}^n \frac{d_i^* u_i}{a_i}$
   (d) 计算 Hessian：$D_D \leftarrow -\sum_{i=1}^n \frac{d_i^* u_i u_i^\top}{a_i^2}$
   (e) 求解 Newton 方向：$D_2 \leftarrow D_D^{-1} D_1$
   (f) 收敛检查：若 $\|D_2\|_\infty < \epsilon$，则收敛，跳出循环
   (g) 线搜索：
   - **while** $\min_i(1 + (\lambda^{(t)} - D_2)^\top u_i) \leq 0$ **do**
   - $D_2 \leftarrow D_2/2$
   - **end while**

   (h) 更新：$\lambda^{(t+1)} \leftarrow \lambda^{(t)} - D_2$
3. **计算权重：**
$$p_i \leftarrow \frac{d_i^*}{1 + (\lambda^*)^\top u_i}, \quad i = 1, \dots, n$$
4. **归一化：** $p_i \leftarrow p_i / \sum_{j=1}^n p_j$（数值保护）

收敛性与可行性条件：

- **可行性**：总体均值 μ_X 必须位于样本 $\{X_i\}_{i=1}^n$ 的凸包内部，否则约束无解。
- **收敛判据**：$\|D_2\|_\infty < \epsilon$，即 Newton 更新量的最大分量小于阈值。
- **数值稳定性**：线搜索步骤确保每次迭代后所有权重保持正定。

权重解释： 最终权重形式为：

$$p_i = \frac{d_i^*}{1 + \lambda^\top (X_i - \mu_X)} \tag{14}$$

当 $d_i = 1$（均匀初始权重）时，简化为：

$$p_i = \frac{1}{n(1 + \lambda^\top (X_i - \mu_X))} \tag{15}$$

此权重通过调整有偏样本的相对贡献，使加权样本均值与已知总体均值 μ_X 对齐，从而校正分布偏移。

3.2 估计抽样概率$\pi_i^{(1)}$

目标： 基于校准权重 $\{p_i\}_{i=1}^n$ 和不确定性估计 $\{\hat{u}(X_i)\}_{i=1}^n$，计算抽样概率。

符号定义：

- p_i：第一步得到的校准权重
- $\hat{u}(X_i)$：预训练的不确定性预测模型输出
- n_b：抽样预算（期望采集标签数量）
- $B = n_b/n$：抽样预算率
- $\tau \in [0,1]$：混合参数，控制主动抽样与均匀抽样的权衡

Step 1：计算加权不确定性 定义加权不确定性：

$$w_i = p_i \cdot |\hat{u}(X_i)|, \quad i = 1, \ldots, n \tag{16}$$

Step 2：计算抽样概率 基于方差最小化原则，抽样概率与加权不确定性成正比：

$$\pi_i^{\text{active}} = \frac{n_b \cdot w_i}{\sum_{j=1}^n w_j} = \frac{n_b \cdot p_i|\hat{u}(X_i)|}{\sum_{j=1}^n p_j|\hat{u}(X_j)|} \tag{17}$$

Step 3：定义均匀抽样概率 均匀抽样作为基准：

$$\pi_i^{\text{uniform}} = B = \frac{n_b}{n}, \quad \forall i \tag{18}$$

Step 4：τ-混合策略 为提高数值稳定性并避免极端抽样概率，采用混合策略（**依旧需要混合策略**）：

$$\pi_i^{(1)} = \tau \cdot \pi_i^{\text{active}} + (1-\tau) \cdot \pi_i^{\text{uniform}} \tag{19}$$

展开为：

$$\pi_i^{(1)} = \tau \cdot \frac{n_b \cdot p_i|\hat{u}(X_i)|}{\sum_{j=1}^n p_j|\hat{u}(X_j)|} + (1-\tau) \cdot \frac{n_b}{n} \tag{20}$$

算法：校准加权主动抽样概率

输入： 校准权重 $\{p_i\}_{i=1}^n$，不确定性估计 $\{\hat{u}(X_i)\}_{i=1}^n$，预算率 $B = n_b/n$，混合参数 $\tau \in [0,1]$，截断阈值 $\epsilon > 0$
输出： 抽样概率 $\{\pi_i^{(1)}\}_{i=1}^n$

1. **计算加权不确定性：**
$$w_i \leftarrow p_i \cdot |\hat{u}(X_i)|, \quad i = 1, \dots, n$$
2. **计算归一化因子：**
$$W \leftarrow \sum_{j=1}^n w_j = \sum_{j=1}^n p_j |\hat{u}(X_j)|$$
数值保护：若 $W < \epsilon$，则 $W \leftarrow \epsilon$
3. **计算主动抽样概率：**
$$\pi_i^{\text{active}} \leftarrow \frac{n \cdot B \cdot w_i}{W}, \quad i = 1, \dots, n$$
4. **计算均匀抽样概率：**
$$\pi_i^{\text{uniform}} \leftarrow B, \quad i = 1, \dots, n$$
5. **τ-混合：**
$$\pi_i^{(1)} \leftarrow \tau \cdot \pi_i^{\text{active}} + (1-\tau) \cdot \pi_i^{\text{uniform}}, \quad i = 1, \dots, n$$

混合参数 τ 的作用：

- $\tau = 1$：纯主动抽样，完全依赖加权不确定性分配抽样概率
- $\tau = 0$：纯均匀抽样，退化为简单随机抽样
- $0 < \tau < 1$：混合策略，在效率与稳定性之间权衡

与传统主动推断的对比： 传统主动推断的抽样概率为：

$$\pi_i^{\text{trad}} \propto |\hat{u}(X_i)| \tag{21}$$

而本方法的抽样概率为：

$$\pi_i^{(1)} \propto p_i \cdot |\hat{u}(X_i)| \tag{22}$$

关键区别在于引入了校准权重 p_i 作为加权因子。由于 p_i 校正了有偏数据集的分布偏移，因此：

- 在过度代表区域（$p_i < 1/n$）：降低抽样概率
- 在代表不足区域（$p_i > 1/n$）：提高抽样概率

这使得抽样策略同时考虑了不确定性和分布校正，实现了更高效的预算分配。

4 问题设置

4.1 可用数据

- **有标签小数据集**：$\mathcal{D}_l = \{(X_j, Y_j)\}_{j=1}^m$，i.i.d. 来自分布 $P = P_X \times P_{Y|X}$
- **无标签大数据集**：$\mathcal{D}_u = \{X_i\}_{i=1}^n$ **有偏**
- **抽样预算**：n_b（期望采集标签数量）
- **辅助信息**：μ_X（已知大数据集的特征总体均值）
- **已训练模型**：
  - $\hat{f}(\cdot)$：标签预测模型
  - $\hat{u}(\cdot)$：残差/不确定性预测模型，近似 $|Y - \hat{f}(X)|$

4.2 估计目标

$$\theta^* = \mathbb{E}[Y] \tag{23}$$

4.3 传统主动推断估计量

$$\hat{\theta}_{\text{active}} = \frac{1}{n}\sum_{i=1}^{n}\left[\hat{f}(X_i) + \frac{\xi_i}{\pi_i}(Y_i - \hat{f}(X_i))\right] \tag{24}$$

其中 $\pi_i \propto |\hat{u}(X_i)|$，$\xi_i \sim \text{Bern}(\pi_i)$ 独立。

4.4 新估计量形式

$$\hat{\theta}_{\text{new}} = \sum_{i=1}^{n} p_i \left[\hat{f}(X_i) + \frac{\xi_i^{(1)}}{\pi_i^{(1)}}(Y_i - \hat{f}(X_i))\right] \tag{25}$$

其中:

- p_i：校准权重，满足 $\sum p_i = 1$ 且 $\sum p_i X_i = \mu_X$
- $\pi_i^{(1)}$：优化后的抽样概率
- $\xi_i^{(1)} \sim \text{Bern}(\pi_i^{(1)})$

5 第一步：估计校准权重 p_i

5.1 目标

目标： 找到权重 $\{p_i\}_{i=1}^n$ 使得：满足校准约束 $\sum_{i=1}^n p_i X_i = \mu_X$

经验似然方法： 最大化经验似然函数

$$\max_{p_1,\dots,p_n} \prod_{i=1}^n p_i = \max \sum_{i=1}^n \ln p_i \tag{26}$$

约束条件：

$$\sum_{i=1}^n p_i = 1, \quad \sum_{i=1}^n p_i X_i = \mu_X, \quad p_i > 0 \tag{27}$$

5.2 求解方法

Lagrangian：

$$\mathcal{L} = \sum_{i=1}^n \ln p_i + \gamma \left(1 - \sum_{i=1}^n p_i\right) + \lambda' \left(\mu_X - \sum_{i=1}^n p_i X_i\right) \tag{28}$$

一阶条件：

$$\frac{\partial \mathcal{L}}{\partial p_i} = \frac{1}{p_i} - \gamma - \lambda' X_i = 0 \tag{29}$$

解的形式：

$$p_i = \frac{1}{\gamma + \lambda' X_i} = \frac{1}{n(1 + \lambda'(X_i - \bar{X}_n))} \tag{30}$$

其中 $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$，$\lambda$ 由校准方程确定。

5.3 算法：Newton-Raphson迭代

Step 1： 定义

$$\phi(\lambda) = \sum_{i=1}^n p_i(\lambda) X_i - \mu_X = \sum_{i=1}^n \frac{X_i}{n(1 + \lambda'(X_i - \bar{X}_n))} - \mu_X \tag{31}$$

Step 2： 计算Jacobian

$$\phi'(\lambda) = -\sum_{i=1}^n \frac{(X_i - \bar{X}_n)(X_i - \bar{X}_n)'}{n(1 + \lambda'(X_i - \bar{X}_n))^2} \tag{32}$$

Step 3： Newton迭代

$$\lambda^{(t+1)} = \lambda^{(t)} - [\phi'(\lambda^{(t)})]^{-1} \phi(\lambda^{(t)}) \tag{33}$$

初始值 ：$\lambda^{(0)} = 0$

收敛判据 ：$\|\phi(\lambda^{(t)})\| < \epsilon$

Step 4: 得到权重

$$p_i = \frac{1}{n(1 + (\lambda^*)'(X_i - \bar{X}_n))} \tag{34}$$

其他距离函数： 选择其他距离函数，可得到不同形式的权重

Table 1: 不同校准方法对比

方法	**距离函数** $G(w,d)$	**权重形式** $F(u)$
线性 (GREG)	$(w-d)^2/2d$	$1+u$
指数	$w\ln(w/d) - w + d$	$\exp(u)$
逻辑	有界约束	$\frac{L(U-1)+U(1-L)e^{Au}}{(U-1)+(1-L)e^{Au}}$

6 第二步：估计抽样概率 $\pi_i^{(1)}$

6.1 问题设置

给定： 第一阶段得到的校准权重 $\{p_i\}_{i=1}^n$

估计量：

$$\hat{\theta} = \sum_{i=1}^n p_i \left[\hat{f}(X_i) + \frac{\xi_i^{(1)}}{\pi_i^{(1)}} (Y_i - \hat{f}(X_i)) \right] \tag{35}$$

条件方差： （给定 $\{X_i, Y_i\}$ 和 $\{p_i\}$）

$$\text{Var}(\hat{\theta} \mid X, Y, p) = \sum_{i=1}^n p_i^2 \cdot \frac{1 - \pi_i^{(1)}}{\pi_i^{(1)}} \cdot (Y_i - \hat{f}(X_i))^2 \tag{36}$$

用 $\hat{u}(X_i)^2$ 替代未知的 $(Y_i - \hat{f}(X_i))^2$）：

$$\widetilde{V}(\pi^{(1)}) = \sum_{i=1}^n p_i^2 \cdot \frac{1 - \pi_i^{(1)}}{\pi_i^{(1)}} \cdot \hat{u}(X_i)^2 \tag{37}$$

6.2 优化问题

$$\min_{\{\pi_i^{(1)}\}} \sum_{i=1}^n \frac{p_i^2 \hat{u}(X_i)^2}{\pi_i^{(1)}} \tag{38}$$

约束条件 ：

$$\sum_{i=1}^n \pi_i^{(1)} = n_b, \quad 0 < \pi_i^{(1)} \leq 1, \quad \forall i \tag{39}$$

6.3 求解

Lagrangian ：

$$\mathcal{L} = \sum_{i=1}^n \frac{p_i^2 \hat{u}(X_i)^2}{\pi_i^{(1)}} + \eta \left(\sum_{i=1}^n \pi_i^{(1)} - n_b \right) \tag{40}$$

一阶条件 ：

$$\frac{\partial \mathcal{L}}{\partial \pi_i^{(1)}} = -\frac{p_i^2 \hat{u}(X_i)^2}{(\pi_i^{(1)})^2} + \eta = 0 \tag{41}$$

解 ：

$$(\pi_i^{(1)})^2 = \frac{p_i^2 \hat{u}(X_i)^2}{\eta} \Rightarrow \pi_i^{(1)} = \frac{p_i |\hat{u}(X_i)|}{\sqrt{\eta}} \tag{42}$$

由约束确定 η ：

$$\sum_{i=1}^n \pi_i^{(1)} = n_b \Rightarrow \sqrt{\eta} = \frac{\sum_{i=1}^n p_i |\hat{u}(X_i)|}{n_b} \tag{43}$$

6.4 抽样概率公式

抽样概率

$$\pi_i^{(1)} = \frac{p_i|\hat{u}(X_i)|}{\sum_{j=1}^{n} p_j|\hat{u}(X_j)|} \cdot n_b \tag{44}$$

基于校准权重的加权归一化?

7 方差估计

traditional_active估计方差:

$$\mathrm{Var}(\tilde{\theta}) = \frac{1}{n^2}\sum_{i=1}^{n} \frac{(1-\pi(X_i))\xi_i}{\pi^2(X_i)} \cdot [Y_i - f(X_i)]^2.$$

EL_active估计方差:

$$\mathrm{Var}(\hat{\theta}) = \sum_{i=1}^{n} p_i^2 \cdot \frac{(1-\pi_i^{(1)})\xi_i^{(1)}}{(\pi_i^{(1)})^2} \cdot [Y_i - f(X_i)]^2$$

8 整体示意图