

## Lecture 11 Policy learning

只需研究如何分配治疗使效益最大化，只考虑将人们分到哪一组（治疗 or 控制）

a treatment policy  $\pi(x)$  is a mapping 映射：

$$\pi : X \rightarrow \{0, 1\}$$

即具有相同特征  $X_i = x$  的个体会被治疗  $\Leftrightarrow \pi(x) = 1$

Goal：找到一个最大化预期效用的 policy：

$$V(\pi) = E [Y_i(\pi(X_i))] \quad (\text{设 } Y_i = 1 \text{ 好})$$

对任何 policy 类： $\Pi$ ，最优 policy  $\pi^*$ ：

def:

$$\pi^* = \operatorname{argmax} \{ V(\pi') : \pi' \in \Pi \}$$

★ (而不是平方误差损失)

其它 policy 的 regret 为：

$$R(\pi) = \sup \{ V(\pi') : \pi' \in \Pi \} - V(\pi)$$

目标为学习一个最坏情况下  $R(\pi)$  界限的 policy： minimax regret criterion  
保证

假设，有  $i=1, \dots, n$  i.i.d.  $(X_i, Y_i, w_i) \in X \times \mathcal{R} \times \{0, 1\}$

$$\{Y_{i(0)}, Y_{i(1)}\} \perp\!\!\!\perp w_i | X_i, \quad Y_i = Y_i(w_i)$$

$$0 < y \leq e(X_i) \leq 1 - y < 1 \quad e(x) = P[w_i = 1 | X_i = x]$$

学习出后，设置  $w_i = \hat{\pi}(x_i)$  并计算  $Y_i = Y_i(\hat{\pi}(x_i))$  (会很大)  
这里治疗是随机性，则不能再了解因果效应

No.

Date

類似:  $E[\hat{\tau}_{IPW}^*] = E\left[\frac{w_i Y_i}{e(x_i)} - \frac{(1-w_i)Y_i}{1-e(x_i)}\right]$

 $= E\left[\frac{w_i Y_{i(1)}}{e(x_i)} - \frac{(1-w_i)Y_{i(0)}}{1-e(x_i)}\right]$ 
 $= E\left[E\left[\frac{w_i Y_{i(1)}}{e(x_i)} | e(x_i)\right] - E\left[\frac{(1-w_i)Y_{i(0)}}{1-e(x_i)} | e(x_i)\right]\right]$ 
 $= E[Y_{i(1)} - Y_{i(0)}]$  unbiased  $\tau$ .

# Policy learning via empirical maximization 经验最大化

一般,

$$\pi^* = \operatorname{argmax} \{ V(\pi) : \pi \in \Pi \}$$

若有  $e(x)$ , 由 IPW:  $\hat{V}(\pi)$  可选择(无偏)

$$\hat{V}_{IPW}(\pi) = \frac{1}{n} \sum_{i=1}^n \frac{\mathbb{1}(\{w_i = \pi(x_i)\}) Y_i}{P[w_i = \pi(x_i) | x_i]}$$

$$\hat{\pi}_{IPW} = \operatorname{argmax} \{ \hat{V}_{IPW}(\pi) : \pi \in \Pi \}$$

$R^P$ , 对样本治疗  $w_i$  与 policy  $\pi(x_i)$  匹配的观测结果进行平均, 用逆  $e(x)$  加权

$$E[\hat{V}(\pi)] = E\left[ \frac{\mathbb{1}(\{w_i = \pi(x_i)\})}{P[w_i = \pi(x_i) | x_i]} \right]$$

$$= E\left[ \frac{\mathbb{1}(\{w_i = \pi(x_i)\}) Y_i(\pi(x_i))}{P[w_i = \pi(x_i) | x_i]} \right]$$

$$= E\left[ \boxed{E\left[ \frac{\mathbb{1}(\{w_i = \pi(x_i)\})}{P[w_i = \pi(x_i) | x_i]} \mid x_i \right]} E[Y_i(\pi(x_i)) | x_i] \right] = 1 \quad \text{unconfoundedness}$$

$$= E[Y_i(\pi(x_i))] = V(\pi)$$

VC 指表示  $\Pi$  的复杂程度，具体定义 重。

若有  $|Y_i| \leq M$ ,  $\eta \leq e(X_i) \leq 1-\eta$ , 以使  $\Gamma_i$  有界，

$\Pi$  有有界的 Vapnik - Chervonenkis 指数。



则 先的遗憾界限为：

$$R(\hat{\alpha}_{IPW}) = O_p\left(\frac{M}{\eta} \sqrt{\frac{VC(\Pi)}{n}}\right) \quad \hat{\alpha}_{IPW} = \arg \max_{\alpha \in \Pi} \{ \hat{A}_{IPW}(x) \}$$

最优收敛速率 其中  $VC(\Pi)$  表示  $\Pi$  的 VC 指数

也有高效率。固定  $2\eta$ ，观察  $n$  时收敛速率

可能会起  $\sqrt{\frac{VC(\Pi)}{n}}$

## Policy learning as weighted classification

价值函数本身分解:  $V(\pi) = E[Y_i(0)] + E[(Y_i(1) - Y_i(0))\pi(x_i)]$

$\rightarrow$

突出显示其依赖于基线效应, 和被  $\pi(\cdot)$

处理的人的平均治疗效应

$$\text{CATE: } A(\pi) = 2E[Y_i(\pi(x_i))] - E[Y_i(0) + Y_i(1)] \\ = E[(2\pi(x_i) - 1)\pi(x_i)]$$

$A(\pi)$  表示  $\pi(\cdot)$  的 advantage.  $\pi^*$  仍是  $A(\pi)$  最大化  $\pi \in \Pi$ .

重写 IPW:  $\hat{\pi}_{IPW}$  maximizes  $\hat{A}_{IPW}(\pi)$ . where.

$$\hat{A}_{IPW}(\pi) = 2\hat{V}_{IPW}(\pi) - \frac{1}{n} \sum_{i=1}^n \left( \frac{w_i Y_i}{e(x_i)} + \frac{(1-w_i) Y_i}{1-e(x_i)} \right) \\ = \frac{1}{n} \sum_{i=1}^n (2\pi(x_i) - 1) \left( \frac{w_i Y_i}{e(x_i)} - \frac{(1-w_i) Y_i}{1-e(x_i)} \right) \quad (11.10)$$

*unbiased  $A_{IPW}(\pi)$ .*

写为:  $\hat{A}_{IPW}(\pi) = \frac{1}{n} \sum_{i=1}^n (2\pi(x_i) - 1) \Gamma_i^{IPW}, \Gamma_i^{IPW} = \frac{w_i Y_i}{e(x_i)} - \frac{(1-w_i) Y_i}{1-e(x_i)} \quad (11.11)$

其中,  $\Gamma_i^{IPW}$  在 ATE 估计分析中的 IPW 分数, 即 ATE 的 IPW 估计是  $\hat{\pi}_{IPW} = n^{-1} \sum_{i=1}^n \Gamma_i^{IPW}$   
 $\hat{A}_{IPW}(\pi)$  类似于 ATE 的 IPW 估计器,  $\pi(x_i) = 1$  earn 并许的治疗效应  
 $\pi(x_i) = 0$  pay 并许的治疗效应.

出于优化目的, 写为

$$\hat{A}_{IPW}(\pi) = \underbrace{\frac{1}{n} \sum_{i=1}^n (2\pi(x_i) - 1) \text{sign}(\Gamma_i) |\Gamma_i|}_{\text{分类目标 classification objective}} \quad (11.12)$$

换个说法,  $\max \hat{A}_{IPW}(\pi)$  相当于优化一个加权分类目标

实践中, 常常通过加权最小化类似的替代损失来

No.

Date

以上后果，加上经验过程集中性的论证，暗示使用 AIPW 评分进行 policy learning 的 regret 有界：

$$R(\hat{\pi}_{AIPW}) = O_p\left(\sqrt{\frac{V^* \cdot V(\pi)}{n}}\right)$$

$$\hat{\pi}_{AIPW} = \arg \max_{\pi \in \Pi} \{ \hat{A}_{AIPW}(x) \}.$$

$$V^* = E[\tau^2(x_i)] + E\left[\frac{\text{Var}[Y_i(0)|X_i]}{1 - e(X_i)}\right] + E\left[\frac{\text{Var}[Y_i(1)|X_i]}{e(X_i)}\right].$$

详见 Wager 2017,

上述界限在一个治疗效应刚从噪声中凸显出来  
的情境中是最优的 不大利...

## Efficient scoring rules for policy learning

有无更好的学习方法？

$$\text{注意到: } A(\pi) = 2E[Y_i(\pi(x_i))] - E[Y_i(0) + Y_i(1)] \\ = E[Y_i(\pi(x_i))] - E[Y_i(1 - \pi(x_i))].$$

即  $A(\pi)$  是在一个实验中比较部署  $\pi(\cdot)$  和始终部署  $\pi(\cdot)$  反策略的 ATE

我们已知 oracle IPW 是可行的但不是高效的

Oracle AIPW 估计量  $\hat{A}_{AIPW}^*(\pi)$ , 通过一个有效的得分来估计  $A(\pi)$ , 达了半参数效率界.

Bp:

$$\hat{A}_{AIPW}^*(\pi) = \frac{1}{n} \sum_{i=1}^n (2\pi(x_i) - 1) \hat{\Gamma}_i^*,$$

$$\hat{\Gamma}_i^* := \mu_{(0)}(x_i) - \mu_{(1)}(x_i) + w_i \frac{Y_i - \mu_{(0)}(x_i)}{e(x_i)} - (1-w_i) \frac{Y_i - \mu_{(1)}(x_i)}{1-e(x_i)}$$

进一步假设存在  $op(n^{-1/4})$ -consistent regression adjustments for  $\hat{\mu}(\omega)(x)$  and  $\hat{e}(x)$

$n \rightarrow \infty$ , 两式误差  $op(n^{-1/4})$

构建高效 oracle 的双重 robust 估计量

$$\star \quad \hat{A}_{AIPW}(\pi) = \frac{1}{n} \sum_{i=1}^n (2\pi(x_i) - 1) \hat{\Gamma}_i,$$

$$\hat{\Gamma}_i := \mu_{(0)}^{(-k(i))}(x_i) - \hat{\mu}_{(1)}^{(-k(i))}(x_i)$$

$$+ w_i \frac{Y_i - \hat{\mu}_{(0)}^{(-k(i))}(x_i)}{\hat{e}^{(-k(i))}(x_i)} - (1-w_i) \frac{Y_i - \hat{\mu}_{(1)}^{(-k(i))}(x_i)}{1-\hat{e}^{(-k(i))}(x_i)}.$$

这也是一个加权分位数

目标

由 L3 已知,  $\hat{A}_{AIPW}(\pi)$  是逐点渐近最优  $\hat{A}_{AIPW}^*(\pi)$  faster than  $\frac{1}{\sqrt{n}}$ .

if  $\Pi$  is a VC class:

$$\sqrt{n} \sup \left\{ |\hat{A}_{AIPW}(\pi) - \hat{A}_{AIPW}^*(\pi)| : \pi \in \Pi \right\} \rightarrow_P 0 \quad (11.17)$$

## The role of the policy class $\Pi$ .

从一个非参数模型开始 ( $\mu(w)(x)$  和  $e(x)$  可以是任意的)

这种情况下，

贝叶斯最优的治疗分配规则为：  $\pi_{\text{bayes}}(x) = \mathbb{I}\{z(x) > 0\}$

然而，我们的目标并不是找一种近似  $\pi_{\text{bayes}}(x)$  的方法，而是另给一个指定的

$\text{policy}$  类  $\Pi$ . (例如  $\Pi$  可能由浅性决策规则,  $k$ -稀疏决策规则、

深度为  $l$  的决策树组成等).

从未设  $\pi_{\text{bayes}}(\cdot) \in \Pi$ .

原因： ①  $X_i$  实现 unconfoundedness:  $(Y_{i(0)}, Y_{i(1)}) \perp W_i | X_i$

一般而言，我们能访问的预处理变量越多，unconfoundedness 越有可能。

为了模型可靠，最好用多样的特征为  $e(x)$ ,  $\mu(w)(x)$  构建灵活的非参数模型。

②  $\pi(\cdot)$  有具体形式。(政策思...)

例： 不应使用某些特征、预算限制、边际预算限制、 $\pi(\cdot)$  功能形式限制

等...

No.

Date

$$\begin{array}{c} \text{1} \quad \dots \quad T \\ \text{2} \\ \vdots \quad \vdots \\ w_1 \quad w_2 \Rightarrow 2^T \text{ 个 } Y \end{array}$$

一个  $t$  次判断，先  $t$  次判断，依赖  $t-1$

## ② 完全随机化

$\{(all\ potential\ outcomes)\} \uparrow w_{1:T}$

不合理

## Lecture 12 Evaluating Dynamic Policies

(强化学习)

评估动态政策 Policy 随时间变化

$n$  个  $i=1, \dots, n$  IID  $t=1, \dots, T$

在每个时间点，观察到一组协变量  $X_{it}$  (随时间变化)

以及一个  $W_{it} \in \{0,1\}$ ,

最后，一旦到达了时间  $T$ ，我们还观察到一个结果  $Y_i \in \mathbb{R}$

为了反映问题的动态结构，我们让任何随时间变化的观察依赖于所有过去的治疗分配。因此，对于每一个  $X_{it} \in \mathcal{X}_t$ ，定义  $2^{t-1}$  个潜在结果  $X_{it}(w_{1:t-1})$

$$\text{s.t. } X_{it} = X_{it}(w_{1:(t-1)})$$

而对于  $Y_i$  有  $2^T$  个潜在结果  $Y_i(w_{1:T})$  s.t.  $Y_i = Y_i(w_{1:(t-1)})$

同理：  $W_{it} = W_{it}(w_{1:(t-1)})$  ,  $W_{it}$  依赖于  $X_{it(t)}$  与过去的治疗值。

定义一个 estimand 估计量 2 种

① 评估固定的治疗选择：  $V(w) = E[Y_i(w)]$  ,  $w \in \{0,1\}^T \leftarrow T$  维。

② 评估一个 policy :  $\pi_t : X_t \rightarrow \{0,1\}$  ,  $W_{it} = \pi_t(X_{it})$

$$\text{则 } V(\pi) = E[Y_i(\pi_1(X_{i1}), \pi_2(\underline{X_{i2}}, \pi_1(X_{i1})), \underline{X_{i2}}(\pi_1(X_{i1})), \dots)]$$

冗长， $\pi$  的协变量依赖于过去的治疗  $\xrightarrow{\text{依赖 } X_{i(1:t)}} \xrightarrow{\text{依赖过去的治疗值}}$

几个问题用以辅助定义估计量。(包括用于收集数据的治疗分配分布的扰动)

几种 natural unconfoundedness-type assumptions :

① 顺序非混杂性 (顺序可忽略性) sequential unconfoundedness (sequential ignorability)  
 $\{( \text{potential outcomes after time } t )\} \perp\!\!\!\perp W_{it} | \{(\text{History at time } t)\}$ .

假设  $W_{it}$  始终在收集到  $t$  为止的数据上是  
非混杂的

## Treatment-confounder feedback 治疗 - 混杂因素之间的反馈

$X_{i1}$  全为 0

错层案例 :  $X_{i1} = 0 \rightarrow P[W_{i1}=1] = 0.5$

$$\begin{cases} X_{i2} = 1 & P[W_{i2}=1] = 0.8 \\ X_{i2} = 0 & P[W_{i2}=1] = 0.4 \end{cases}$$

估计 :  $\tau = E[Y_{(1)} - Y_{(2)}]$

分析可见 :  $E[Y_i | W_{i1}=0] = E[Y_i | W_{i1}=1] = 0$

$E[Y_i | W_{i2}=0, W_{i1}=w_1, X_{i2}=x] = E[Y_i | W_{i2}=1, W_{i1}=w_1, X_{i2}=x]$  for 所有  $w_1, x$ .

第一种错误 : 忽略自适应抽样  $\hat{\tau} = \hat{E}[Y|W=1] - \hat{E}[Y|W=0]$  (弃善求)

第二种错误 : 以  $X_{i2}$  分层 :

$$\hat{\tau}_0 = E[Y|W=1, X_{i2}=0] - E[Y|W=0, X_{i2}=0]$$

$$\hat{\tau}_1 = E[Y|W=1, X_{i2}=1] - E[Y|W=0, X_{i2}=1]$$

$$\hat{\tau} = \frac{n_0 \hat{\tau}_0 + n_1 \hat{\tau}_1}{n_0 + n_1}$$

该方法 : 已知  $Y_i(\dots)$  及  $W_{i2}|X_{i2}$  为以上做法

但实际需要分层 :  $Y_i(\dots)$  及  $(W_{i1}, W_{i2})|X_{i2}$

例:	$W_{i1}=0$	$W_{i1}=1$	类似依从性:	
			$E[Y W=1, X_{i2}=0]$ 为①②均, $E[Y W=0, X_{i2}=0]$ 为①	
			$E[Y W=1, X_{i2}=1]$ 为③, $E[Y W=0, X_{i2}=1]$ 为②③均	
			即 在顺序随机化试验中, 分层不能控制混杂因素.	
stable	$X_{i2}=0$	$X_{i2}=0$	①	
responder	$X_{i2}=1$	$X_{i2}=0$	②	
acute	$X_{i2}=1$	$X_{i2}=1$	③	

$$V(\pi) = E[Y_i(\pi_1(X_{i1}), \pi_2(X_{i2}, \pi_1(X_{i1})), X_{i2}(\pi_1(X_{i1})), \dots)] \quad (12-2)$$

Sequential inference for sequential ignorability

顺序推断，用于顺序可忽略性。

阶层不起作用，专注于估计  $V(\pi)$ .

定义：

- $F_t = \sigma(X_1, W_1, X_2, W_2, \dots, W_{t-1}, X_t)$  表示包含所有信息的过滤器。  
直到选择时期  $t$  的治疗

- $E_\pi$  作为简写，表示不设置治疗的期望

$$(12-2) \text{ 变为 } V(\pi) = E_\pi[Y]$$

- 定义价值函数：

$$V_{\pi,t}(X_1, W_1, \dots, W_{t-1}, X_t) = E_\pi[Y | F_t]$$

这个函数衡量了，当从  $F_t$  状态，开始遵循  $\pi$ ，我们预计会获得的回报  
 $policy \pi$  整体价值为  $V_{\pi,0}$ .

由链式法则：

$$\begin{aligned} \underline{E_\pi[V_{\pi,t+1}(X_1, W_1, \dots, W_t, X_{t+1}) | F_t]} &= E_\pi[E_\pi[Y | F_{t+1}] | F_t] \\ \text{已知 } F_t \text{ 在下, } V_{\pi,t+1} \text{ 的 } E &= E_\pi[Y | F_t] \\ &= V_{\pi,t}(X_1, W_1, \dots, W_{t-1}, X_t) \end{aligned}$$

即 当已知  $F_t$ ，从  $t$  或  $t+1$  或以后的预期一直不变 “R 末尾方程”

$$V(s) = R(s) + \gamma \sum_{s'} P(s'|s, a) V(s')$$

若有  $V_{\pi,t+1}$  的好估计，我们能获得  $V_{\pi,t}$  的好估计，一直回溯到  $V(\pi)$ .

$(X_1, W_1, \dots, X_T, W_T, X_{T+1})$  的联合分布因子分解为：  $(Y = X_{T+1})$

$$\begin{aligned} P_\pi[X_1, W_1, \dots, X_T, W_T, Y] \\ = P_\pi[X_1] \prod_{t=1}^T P_\pi[W_t | F_t] P_\pi[X_{t+1} | F_t, W_t]. \end{aligned}$$

其中，unconfoundedness，意味着分解中，不积分  $W_t$  的项不依赖于  $\pi$ .

$$P_\pi[X_1] = P[X_1]$$

$$P_\pi[X_{t+1} | F_t, W_t] = P[X_{t+1} | F_t, W_t]. \quad \text{for all } \pi.$$

WT 对 XT 的影响与 π无关

$$E_{\pi} [V_{\pi, t+1}(x_1, w_1, \dots, w_t, x_{t+1}) | F_t] = E_{\pi} [E_{\pi} [Y | F_{t+1}] | F_t]$$

$$= E_{\pi} [Y | F_t] = V_{\pi, t}(x_1, w_1, \dots, w_{t-1}, x_t). \quad (12.6)$$

No.

Date

### Inverse-propensity weighting

为了让(12.6)有用，而进行逆概率变换

寻求“off policy”分布的期望，实际数据与 $\pi(\cdot)$ 不完全一致

$$V_{\pi, t}(x_1, w_1, \dots, w_{t-1}, x_t)$$

$$= E_{\pi} [V_{\pi, t+1}(x_1, w_1, \dots, w_t, x_{t+1}) | F_t]$$

从 $\pi$ 到 $\pi$

$\Rightarrow$ 更大

$$= E \left[ \frac{P_{\pi}[w_t, x_{t+1} | F_t]}{P[w_t, x_{t+1} | F_t]} V_{\pi, t+1}(x_1, w_1, \dots, w_t, x_{t+1}) | F_t \right]$$

$$= E \left[ \frac{P_{\pi}[w_t | F_t] P_{\pi}[x_t | F_t, w_t]}{P[w_t | F_t] P[x_t | F_t, w_t]} V_{\pi, t+1}(x_1, w_1, \dots, w_t, x_{t+1}) | F_t \right]$$

$$= E \left[ \frac{\mathbb{1}(\{w_t = \pi_t(\dots, x_t)\})}{P[w_t = \pi_t(\dots, x_t) | F_t]} V_{\pi, t+1}(x_1, w_1, \dots, w_t, w_{t+1}) | F_t \right]$$

由逆概率变换：成为 $V(\pi)$ 的估计量，对每个 $t=1, \dots, T$ 写出变换。

$$V(\pi) = E[V_{\pi, 1}(x_1)]$$

$$V_{\pi, 1}(x_1) = E \left[ \frac{\mathbb{1}(\{w_1 = \pi_1(x_1)\})}{P[w_1 = \pi_1(x_1)]} V_{\pi, 2}(x_1, w_2, x_2) | F_1 \right]$$

$|F_1?$

用反向替换 最后只剩下  $V_{\pi, T+1}(\cdots) = E_{\pi} [Y | F_{T+1}] = Y$

$$V(\pi) = E \left[ \prod_{t=1}^T \frac{\mathbb{1}(\{w_t = \pi_t(\dots, x_t)\})}{P[w_t = \pi_t(\dots, x_t) | F_t]} Y \right]$$

则有一种IPW估计量：

$$\hat{V}_{IPW}(\pi) = \frac{1}{n} \sum_{i=1}^n \gamma_i(\pi) Y_i$$

该估计量：平均了匹配的结果

并用IPW校正由  
观察到的（时间变化）  
的混杂因素而产生的  
的选择效应

$$\gamma_i(\pi) = \gamma_{i(t-1)}(\pi) \frac{\mathbb{1}(\{w_t = \pi_t(\dots, x_t)\})}{P[w_t = \pi_t(\dots, x_t) | F_t]},$$

$$\gamma_{i0}(\pi) = 0.$$

若已知 $\gamma_{iT}$ ，即 每个 $t$ 的  $P[w_t = \pi_t(\dots, x_t) | F_t]$ ， $n_{IPW}$  为无偏

KOKUYO

$$E_{\pi} [V_{\pi,t+1}(X_1, W_1), \dots, W_t, X_{t+1}) | F_t]$$

$$= E_{\pi} [E_{\pi} [Y | F_{t+1}] | F_t] = E_{\pi} [Y | F_t] = V_{\pi,t}(X_1, W_1, \dots, W_{t-1}, X_t).$$

(12.6)

Backwards regression adjustment.

回归调整，利用(12.6)和序贯无偏性：

反向迭代：

对于  $t=T$ ：

$$\begin{aligned} V_{\pi,T}(X_1, W_1, \dots, X_T) &= E_{\pi} [Y | F_T] \\ &= E_{\pi} [Y | F_T, W_T = \pi_T(X_1, W_1, \dots, X_T)] \\ &= E [Y | F_T, W_T = \pi_T(X_1, W_1, \dots, X_T)]. \end{aligned}$$

非参数回归，直接学习  $\hat{V}_{n,T}(X_1, W_1, \dots, X_T)$

若我们有一个合理的  $\hat{V}_{n,t+1}$  的估计，则

$$V_{\pi,t}(X_1, W_1, \dots, X_t) \approx E [\hat{V}_{n,t+1}(\dots, X_{t+1}) | F_t, W_t = \pi_t(\dots, X_t)].$$

继续反向递归，直至恢复  $V_{\pi,0}$  的估计。

(12.12)

与 IPW 不同，需要量化回归误差在 反向迭代时如何传播，Q-learning. ✓

A doubly robust estimator 双重稳健估计器.

已有 IPW 和 回归.

在回归 12.12 中  $V_{z,t}(x_1, w_1, \dots, x_t) \approx E[\hat{V}_{\pi,t+1}(\dots, x_{t+1}) | F_t, w_t = z_t(\dots, x_t)]$ .

$\Downarrow$  得一个良好的价值估计  $\hat{V}_{\pi,1}(x_1)$

$$\hat{V}_{\text{REG}}(x) = \frac{1}{n} \sum_{i=1}^n \hat{V}_{\pi,1}(x_{i1})$$

若, 我们对  $\hat{V}_{\pi,2}$  比  $\hat{V}_{\pi,1}$  更信任: 用  $\hat{V}_{\pi,2}$  去偏

考虑使用:

$$\hat{V}(x) = \frac{1}{n} \sum_{i=1}^n \left( \hat{V}_{\pi,1}(x_{i1}) + r_{i1}(x) \underbrace{\left( \hat{V}_{\pi,2}(x_{i1}, w_{i1}, x_{i2}) - \hat{V}_{\pi,1}(x_{i1}) \right)}_{\text{校正项}} \right)$$

$\Downarrow$

直观  $V_{\pi,T+1} = Y$ :

$$\hat{V}_{\text{AIPW}}(x) = \frac{1}{n} \sum_{i=1}^n \left( Y_i + \sum_{t=1}^T (\delta_{i(t-1)}(x) - \delta_{it}(x)) \hat{V}_{\pi,t}(x_{i1}, \dots, x_{it}) \right)$$

任一  $\hat{V}_{\pi,i}$  无偏,  $\hat{V}_{\text{AIPW}}$  无偏,

非动态情况下 AIPW 估计器的一个  
泛化

## Lecture 13 Structural Equation Modeling

SEM. 结构方程模型 (之前  $Y_i = \alpha + w_i z + \varepsilon_i$ , 现 set  $w_i = w$ )

then observe  $Y_i(w_i=w) = \alpha + w z + \varepsilon_i$

Non-parametric SEM. 定义:

设  $(X_1, \dots, X_p)$  为  $p$  个随机变量. 具有联合分布  $IP$

该  $IP$  总是可有一个有向无环图 (DAG. Directed Acyclic Graph)  $G$  来表示,

意味着  $IP$  可分解为:

$$IP[X_1, \dots, X_p] = \prod_{j=1}^p IP[X_j | p_{aj}] \quad (13.2)$$

其中  $p_{aj}$  表示  $G$  中  $X_j$  的父节点.

(即  $p_{aj} = \{X_i : E_{ij}=1\}$ ),  $E_{ij}$  表示  $G$  中从  $X_i \rightarrow X_j$  边的存在

更进一步假设, 存在确定性函数  $f_j(\cdot)$ ,  $j=1, \dots, p$

分解 (13.2).

(13.3)

$$x_j = f_j(p_{aj}, \varepsilon_j) \quad \text{其中 } \varepsilon_j \sim F_j \text{ 是相互独立的噪声项.}$$

就一个 SEM, DAG, 一点由其父节点和噪声决定  
即使上游的  $p_{aj}$  改变, 但  $f_j(\cdot)$  不变, SEM.

(为具体化, 可看为  $f_j$  根据  $p_{aj}$  对  $x_j$  的分布率引)

考虑: 因果查询 (causal query). 外生地设置  $G$  中某些节点的值, 并观察其如何影响其它节点的分布.

给两个不相交的节点集合:  $W, Y \subset X$ , 将  $W$  设置为  $w$  对  $Y$  的因果效应写作,

$$IP[Y | do(W=w)]$$

对应于 (13.3), 相当于删除所有与  $W$  相关的方程, 并在其余方程中用  $w$  替代  $W$ . ✓

例: 对  $X_j$  干预

$$P[X | do(X_j=x_j)] = \begin{cases} P[X] / P[X_j=x_j | p_{aj}] & \text{若 } X_j=x_j \\ 0 & \text{else} \end{cases}$$

# The do calculus      do计算

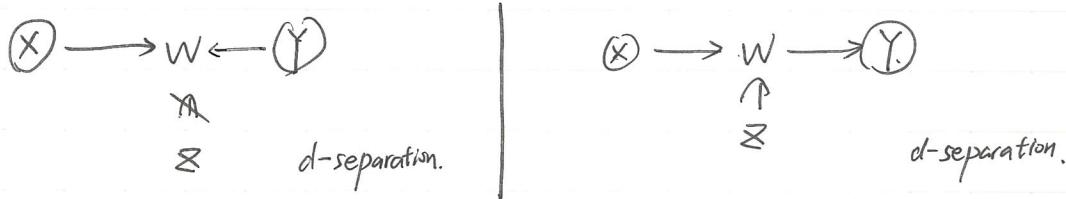
定义 d-separation d-分离. ( $X, Y, Z$  为不相关的节点集合)

$p$ , 为  $X$  中一个节点到  $Y$  中一个节点的任意无向路径

若  $p$  上有一个节点  $W$ , s.t.

- (1)  $W$  在  $p$  上是一个碰撞点 (collider), 且  $W$  及其任何后代都不在  $Z$  中
- (2)  $W$  不是碰撞点, 且  $W$  在  $Z$  中

我们称:  $Z$  阻断了  $X$  和  $Y$  之间的一条路径,  $Z$  d-separates  $X$  and  $Y$ .



当  $Z$  d 分离  $X, Y$ , 可以从 (13.2) 导出  $X \perp\!\!\!\perp Y | Z$

写作  $(X \perp\!\!\!\perp Y | Z)_G$ .

定义:  $G_{\bar{X}}$  为删去指向  $X$  的边的  $G$  的子图.

$G_X$  从  $X$  出发

$G_{X\bar{Z}}$  从  $X$  出发和指向  $Z$

No.

Date

三个式中都有  $do(X=x)$ .

简记:  $\bar{X}$  提供 =  $do(X=x)$  条件.

之后, ①  $Z, Y|Z \Rightarrow do Z = control Z$ . 解.

②  $\bar{Z}(W), Y|Z \Rightarrow do Z$  相当于设  $do$ .

$Z(W)$  下游无  $W$

$do(X=x)$  相当于删去入边:  $\bar{X}$

$G\bar{X}$  为删指  $\rightarrow X$  ,  $G\bar{x}$  为删指出  $\leftarrow \bar{x}$

对任何不相交的点集  $X, Y, Z, W$ , 以下等式成立:

(1). 插入/删除观察 Insertion/deletion of observations:

if  $(Y \sqcup Z | X, W)_{G\bar{X}}$ :

$$P[Y | do(X=x), Z=z, W=w] = P[Y | do(X=x), W=w].$$

(2). 动作/观察交换 Action/observation exchange

if  $(Y \sqcup Z | X, W)_{G\bar{X}\bar{Z}}$ :

$$P[Y | do(X=x), do(Z=z), W=w] = P[Y | do(X=x), Z=z, W=w].$$

(3) Insertion / deletion of actions: 插入/删除动作

if  $(Y \sqcup Z | X, W)_{G\bar{X}\bar{Z}(W)}$  where  $Z(W)$  是  $G\bar{X}$  中不是  $W$  节点祖先的  $Z$  节点集合

$$P[Y | do(X=x), do(Z=z), W=w] = P[Y | do(X=x), W=w]$$

我们要使用  $do$  运算, 将因果查询简化为关于  $P$  的可观察矩的查询,

即不涉及  $do$ -操作符并且仅依赖于观察到的  
随机变量的条件期望.

$do$  运算是完全的, 若我们不能用  $do$  计算来简化因果查询, 那么它在

SEM 的术语中就不是非参数可识

别的.

X满足Y-W后门准则, X为Y-W父  
w] Ydown 可用 X 父.

Example 1. Back-door criterion. 后门准则.

三组不相邻节点  $X \perp\!\!\! \perp W$ , 想要 query  $P[Y | do(W=w)]$

假设,  $X$ 不包含  $W$  的下游节点; 且在阻断所有  $W$  的下游边后,  $X$  与  $W$  和  $Y$

$$\text{即: } (Y \perp\!\!\! \perp w | X)_{Gw} \quad (13.8)$$

则有:

$$P[Y | do(W=w)] = \sum_x P[X=x | do(W=w)] P[Y | X=x, do(W=w)]$$

$$= \sum_x P[X=x] P[Y | X=x, do(W=w)] \quad \text{因为 } X \text{ 在 } W \text{ 上游. } (X \perp\!\!\! \perp W)_{Gw}$$

$$= \sum_x P[X=x] P[Y | X=x, W=w] \quad \# \text{ Rule 2} \quad (13.9)$$

13.8: 考虑,  $Y, W$  都为单例, 且  $W$  在  $G$  中除  $Y$  以外无下游变量,

然后, 阻断从  $W$  出发的下游箭头  $\Rightarrow$  使  $W$  对  $Y$  的影响不明确.

$$(13.8) \text{ 变为 } F_Y(w) \perp\!\!\! \perp w | X \quad (13.10)$$

$$\text{where } F_Y(w) = f_Y(w, X, \varepsilon_Y)$$

除  $w$  以外所有因素在

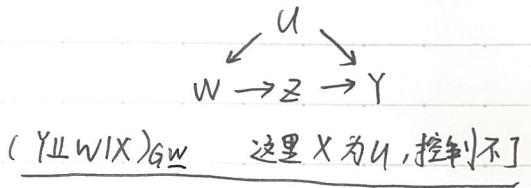
$x_j = f_j(paj, ej)$  都未确定

可用图推断  
出类似 13.10  
的条件独立性  
声明

Example 2: Front-door criterion

例: 在右图中计算  $P[Y|do(W=w)]$

但  $U$  无法观测, 不可用后门准则



但若存在一个  $Z$ , 其完全中介了  $W$  对  $Y$  的影响, 且未受  $U$  影响,

$$\begin{aligned} P[Y|do(W=w)] &= \sum_z P[Z=z] \underbrace{P[Y|Z=z, do(W=w)]}_{\downarrow W \text{ 和 } Z \text{ 的关系, 不存在 } W-Z \text{ 后门}} \\ &= \sum_z P[Z=z|W=w] P[Y|Z=z, do(W=w)] \end{aligned}$$

$$\text{而 } P[Y|Z=z, do(W=w)] = P[Y|do(Z=z), do(W=w)] \quad \text{Rule 2}$$

$$= P[Y|do(Z=z)] \quad \text{Rule 3}$$

$$= \sum_{w'} P[W=w'] P[Y|Z=z, W=w'] \quad \text{后门准则. } \checkmark$$

$$\text{综上 } P[Y|do(W=w)] = \sum_{w'} P[Z=z|W=w] \sum_{w'} P[W=w'] P[Y|Z=z, W=w']. \quad (13.11)$$

$W-Y$

$W-Z$

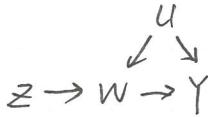
$Z-Y$

前门准则: 尽管其是通过  $do(W=w)$  干预的, 但其仍需要整合观察不到的

$P[W=w']$  分布

Example 3: Instrumental variables.

希望估计  $P[Y|do(w=w)]$ .



$do$  演示在这种情况下无用.

IV 无法帮助, 增加节点只会使满足 do-calculus 所用的 d-分离条件更复杂

在用 IV 时, 我们的假设不可仅为 SEM, (例如还带单调...)

而且即使这样, 也只能识别非标准的因素量, 如 LATE.

而且如插入单调性至 SEM 中, 也很复杂

SEMs 和 Potential Outcome.

结构方程模型、潜在结果模型

在简单的两节点图,  $w \rightarrow y$ , 二者相同.  $y(w)$  和 SEM 的  $f_y(w, \epsilon_y)$

可一映射

更一般的, SEM 和 Potential outcome 不同

Potential outcome: 只能提出, 将 ... 变量设为什么什么值, Y 会是 ...

SEM: 灵活

$w \rightarrow z$



$z(w) := f_z(w, \epsilon_z)$

$P[y|do(w=0), do(z=z(w=1))]$

$= f_y(w=0, z=f_z(w=1, \epsilon_z), \epsilon_y)$

可解?

不同时有关也可放一起求解

Potential outcome 中不行, 除非扩围

一些观点: SEM 可识别 不对立操作的因果效应 为 SEM 缺陷.

P·O... 不 - - -

为 P·O... 局限性

叙述：有  $T$  个相似独立个体，

对每个（都一样）

其有  $k$  个处理方法，对  $\underbrace{\text{这 } k \text{ 个潜在结果}}_{\substack{W_t \\ (\text{从 } K \text{ 个选})}}$

$y_t(k)$  表示了

第  $t$  个对象的结果，

观察只能观察到：

$$Y_t = Y_t(W_t)$$

$\mu$  为收益均值：

$\mu_k$  为第  $k$  种处理的收益期望 = mean.

→ 都是一个处理。

$R_T$  为总遗憾，为  $T$  个对象 所选处理与最佳处理之差 的和。

# Lecture 14 Adaptive Experiments 自适应实验 (动态调整实验方案)

之前的固定实验，

在IID环境下，这个条件有时过于严格

Setting and notation.

分析： $k = 1, 2, \dots, K$  个 actions

$t = 1, \dots, T$  个 主体

每个主体者有独立同分布IID的潜在结果

我们观察： $Y_t(k) \sim F_k$ ,  $Y_t = Y_t(w_t)$

其中， $w_t$  是  $t$  采取的动作， $F_k$  是第  $k$  个臂的潜在结果分布

$\mu_k = E[Y_t(k)]$  表示  $F_k$  的均值：

定义 Regret:

$$R_T = \sum_{t=1}^T (\mu^* - \mu_{w_t}), \quad \mu^* = \sup \{ \mu_k : 1 \leq k \leq K \}$$

$\mu^*$  是所有可能行动中

Regret 为一系列 actions 奖励 不足部分的 (expected)

预期奖励的最大值。

$$n_{k,t} = \sum_{j=1}^t \mathbf{1}(\{w_j=k\}) \Rightarrow \text{第 } k \text{ 个臂被拉次数}$$

从 1...t 的  
进程量。

$$\hat{\mu}_{k,t} = \frac{1}{n_{k,t}} \sum_{j=1}^t \mathbf{1}(\{w_j=k\}) Y_j \Rightarrow \text{第 } k \text{ 个臂的平均奖励}$$

显然，在随机试验中： $w_t$  均匀（且非自适应）分布在  $\{1, \dots, K\}$  上

Regret 和  $T$  成线性比例： $R_T \sim T \sum_{k=1}^K (\mu^* - \mu_k) / K$ . 1 个体的损失期望

↑  
线性

自适应实验：第一目标：实现 子线性遗憾  $\Rightarrow$  Regret 不随实验次数或时间

成性增长，

而是越来越接近最优选择。



exploration  
先试

exploitation  
利用好的

## Optimism in the face of uncertainty

平衡 explore - exploit

不确定性

UCB 算法 upper confidence band (置信区间上界):

首先, 用  $t_0$  次抽样来初始化每个臂: 先  $K$  次抽样: 选择并观察  $t_0$  次;

然后: 1. 在每个时间  $t = K_0 + 1, K_0 + 2, \dots$ , 基于时间  $t-1$  收集到的数据

为  $\mu_k$  构建一个置信区间  $\hat{U}_{k,t}$

2. 选择与有最大上端点的  $\hat{U}_{k,t}$  对应的动作  $w_t$ , 观察  $y_t = Y_t(w_t)$

每过一个  $t$ , 只多了一个数据点。

UCB 控制 Regret 的原理: 考虑  $Y_{t|K} \sim F_K$ ,  $y_t = Y_t(w_t)$ ,  $F_K$  高斯分布

简化:

$$Y_{t|K} \sim N(\mu_k, \sigma^2) \quad (14.4)$$

$\sigma^2$  已知.  $\hookrightarrow$  "已知" 是为了简化计算

则可使用该式计算:

$$\hat{U}_{k,t} = \hat{\mu}_{k,t-1} \pm \sigma \sqrt{4 \log(T) / n_{k,t-1}} \quad (14.5)$$

当  $t_0=1$ : 用(14.5)的UCB Regret 如下:

UCB 的 Regret 只以对数增长

$$R_T = \sum_{k \neq k^*} \frac{16\sigma^2 \log(T)}{\mu_{k^*} - \mu_k} + (\mu_{k^*} - \mu_k) \quad \text{with prob. at least } 1 - K/T$$

其中  $k^*$  表示最优的 arm  $(14.6)$

臂与最优臂越接近,

UCB 越需要更长的时间分辨; 很烂的 arm 可快速丢弃

$\mu_{k^*} - \mu_k$  很小时可能会使  $R_T$  无界  $\rightarrow \infty$ ; 实际上  $\begin{cases} \text{effect 强: UCB 专用 (arm 间差异)} \\ \text{effect 弱: 上界为 } T(\mu_{k^*} - \mu_k) \end{cases}$

综上. Regret 总有界, 所有  $\mu_k$  在最坏情况下, Regret 上界也在  $K\sqrt{T \log(T)}$  上

也有说：几乎肯定  $z_i \in [a, b]$

Hoeffding's inequality:



一系列随机变量  $z_1, \dots, z_n, z_i \in [a, b] \text{ for all } i$ .

有: for all  $t > 0$

$$P\left(\frac{1}{n} \sum_{i=1}^n (z_i - E[z_i]) \geq t\right) \leq \exp\left(-\frac{2nt^2}{(b-a)^2}\right)$$

$$\text{或 } \leq -t$$

$$\text{或 } | \geq t \leq 2 \exp$$

Union bound:  $n$  个事件  $A_1, \dots, A_n \Rightarrow$  放缩  $\Rightarrow 1 - \frac{k}{T}$

$$P(A_1 \cup A_2 \cup \dots \cup A_n) \leq P(A_1) + P(A_2) + \dots + P(A_n)$$

计算?

猜证, 只需证  $P\left(M_k - \frac{1}{j} \sum_{i=1}^j Y_i'(k) - b\sqrt{4\log(T)}j \geq 0\right) \leq \frac{1}{T}$

而: 左边  $P\left(\frac{1}{j} \sum_{i=1}^j Y_i'(k) - M_k \leq -b\sqrt{4\log(T)}j\right) \leftarrow Y_i'(k) \text{ 无界, } k \text{ 是否写错?}\right.$

$$\leq \exp\left(-\frac{2j b^2 4 \log(T) j}{(b-a)^2}\right)$$

$$= \exp\left[-\frac{8b^2}{(b-a)^2} \log(T)\right]$$

$$\leq \frac{1}{T}$$

不知道怎么化的,

可能有未说明的假设

## A regret Bound of UCB

证明  $R_T = \sum_{k \neq k^*} \frac{16B^2 \log(T)}{\mu_{k^*} - \mu_k} + (\mu_{k^*} - \mu_k)$  with prob. at least  $1 - K/T$ .

$R_T$  可写为:  $R_T = \sum_{k \neq k^*} n_{k,T} (\mu_{k^*} - \mu_k)$ .  
T时, k的次数

主要任务是限制  $n_{k,T}$ , 即 UCB 拉动任何次优臂的次数. ← 也是 UCB 设计的理据,  
以这个为目的的设计

用  $\zeta_{kj}$  表示第  $k$  个 arm 第  $j$  次被拉动的时间. 对  $t=k+1, \dots, T$

$$\begin{aligned} & P \left[ \sup_{K < t \leq T} \left\{ \mu_k - \hat{\mu}_{k,t-1} - 6\sqrt{4\log(T)/n_{k,t-1}} \geq 0 \right\} \right] \\ & \quad \xrightarrow{\text{范围} \uparrow \sup \uparrow} \text{这个概率的上确界 } P[\sup[1]] \\ & \leq P \left[ \sup_{1 \leq j \leq n_{k,T}} \left\{ \mu_k - \hat{\mu}_k, \zeta_{kj} - 6\sqrt{4\log(T)/j} \geq 0 \right\} \right] \\ & \quad \xrightarrow{\text{同分布}} \\ & = P \left[ \sup_{1 \leq j \leq n_{k,T}} \left\{ \mu_k - \frac{1}{j} \sum_{l=1}^j Y_l'(0) - 6\sqrt{4\log(T)/j} \geq 0 \right\} \right] \\ & \quad \xrightarrow{\text{范围} \uparrow \sup \uparrow} \\ & \leq P \left[ \sup_{1 \leq j \leq T} \left\{ \mu_k - \frac{1}{j} \sum_{l=1}^j Y_l'(0) - 6\sqrt{4\log(T)/j} \geq 0 \right\} \right] \\ & \leq 1/T \quad \textcircled{1} \end{aligned}$$

其中  $Y_l'(0)$  是从  $N(\mu_k, \sigma^2)$  独立抽取的

则有 对每一个 arm  $k=1, \dots, K$ , 有  $1 - \frac{1}{T}$  的概率.

$$\mu_k \leq \hat{\mu}_{k,t-1} + 6\sqrt{4\log(T)/n_{k,t-1}} \quad \textcircled{1}' \quad \text{for all } t = k+1, \dots, T$$

所有边界 (求和)  $1 - K/T$

prob.  $(-\frac{1}{T}, k=1 \dots K, t=k+1 \dots T)$

$$\text{当 } \mu_k \leq \hat{\mu}_{k,t-1} + b\sqrt{4\log(T)/n_{k,t-1}} \quad (14.8)$$

当 14.8 对所有 arm 成立，

$$\text{时 } w_t = k \Rightarrow \hat{\mu}_{k,t-1} + b\sqrt{4\log(T)/n_{k,t-1}} \geq \hat{\mu}_{k^*,t-1} + b\sqrt{4\log(T)/n_{k^*,t-1}}$$

拉动 k:

$$K \text{ 这个臂的上界必须} \Rightarrow \hat{\mu}_{k,t-1} + b\sqrt{4\log(T)/n_{k,t-1}} \geq \mu_{k^*}$$

大于等于  $\mu_{k^*}$

$$\Rightarrow \mu_k + 2b\sqrt{4\log(T)/n_{k,t-1}} \geq \mu_{k^*}$$

$$\Rightarrow n_{k,t-1} \leq \underbrace{16L^2 \log(T)}_{(1)} / (\mu_{k^*} - \mu_k)^2. \quad (2)$$



当  $n_{k,t-1}$  超过某个界值

$n_{k,t-1}$  增长比  $\log(T)$  快很多，

拉动非  $k^*$  的 arm，不可用

将这个界代入 Regret 定义

$$RT = \sum_{k \neq k^*} (\mu_{k^*} - \mu_k) n_{k,T}$$

界，在界后全为  $k^*$ .

$$RT = \sum_{k \neq k^*} \underbrace{\frac{16L^2 \log(T)}{\mu_{k^*} - \mu_k}}_{t=k+1, \dots, T} + (\mu_{k^*} - \mu_k) \quad (2)$$

- 开始如 = 1 累积的 RT

方案

Adaptive randomization schemes.

UCB 有局限性。

很难与 IID 方法集成

泛化到更复杂的抽样设计困难

UCB:  $W_t$  依赖过去的观察

其它: 依赖随机化的 IID 方法

(统计检验、系数推导等)

需要调整  $\hat{U}_{kt}$  形式

Thompson sampling 汤普森采样 (一个代替 UCB 的方法)

建立在不同浓度不等式基础上的弱证明

基于贝叶斯更新的启发式算法

首先:

# 在潜在结果分布  $F_k$  上选择一个先验  $\pi_{k,0}$ 对每个时间点  $t=1, \dots, T$ ,① 计算 arm  $k$  是最佳 arm 的 prob.  $e_{k,t-1}$ 

$$e_{k,t-1} = P_{\pi_{k,t-1}} [\mu_k = \mu_*].$$

② 随机选一个 action  $w_t \sim \text{Multinomial}(e_{\cdot,t-1})$ ③ 观察  $y_t = y_t(w_t)$  并更新后验  $\pi_{k,t}$ .后验概率降低  $\frac{1}{t}$  以下Thompson 与 UCB 在统计行为上相似.  $\Rightarrow$  定期探索每个 arm, 直到基本确认该 arm 不好

→ 抽样中采取的行动是随机的 (具有依赖于过去数据的自适应随机概率)

与 Causal Inference 行为,

调整因素  $\pi_{k,0}$ , 以 UCB 中的  $\hat{U}_{kt}$  更易推理.

鞅: 序列  $X_1 \dots X_n$ .

1.  $\forall i, X_i$  可测, 即  $\forall x \in R, \{X_i \leq x\}$  是一个事件,  $P$  可测  
 2.  $E[X_i] < \infty$   
 3. 对所有  $i$

$$E[X_{i+1} | X_1, X_2, \dots, X_i] = X_i$$

若 鞅的性质 (1.3)   
 1.  $M_t$  仅与  $y_1, \dots, y_n$  有关  
 $E(M_t | y_1, \dots, y_n) = M_t$   
 $M_t$  为  $y_1, \dots, y_n$  的函数.

$$\hat{\mu}_k^{AW} = \sum_{t=1}^T \frac{I(\{W_t=k\})y_t}{\sqrt{e_{t,k}}} / \sum_{t=1}^T \frac{I(\{W_t=k\})}{\sqrt{e_{t,k}}} \quad (14.3)$$

验证 (14.3).

$$\hat{\mu}_k^{AW} - \mu_k = \sum_{t=1}^T \frac{I(\{W_t=k\})(y_t - \mu_k)}{\sqrt{e_{t,k}}} / \sum_{t=1}^T \frac{I(\{W_t=k\})}{\sqrt{e_{t,k}}}$$

其中分子:

$$M_t = \sum_{j=1}^t \frac{I(\{W_j=k\})(y_j - \mu_k)}{\sqrt{e_{j,k}}} \quad \text{为其部分和 (数列前段)}$$

若  $W_t$  是根据直到  $t$  的信息随机选择的.  $W_t$  在给定  $M_{1:t-1}$  的条件下与  $y_t(k)$  独立因此  $M_t$  是一个鞅:

$$H_t: E[M_t | M_{1:t-1}] = M_{t-1}$$

由加权方案

与其它 estimator 不同

$$\text{Var}[M_t | M_{1:t-1}] = b_k^2 \quad b_k^2 = \text{Var}[y_t(k)] \quad (14.8)$$

↓ 自适应加权的 CLT

鞅 CLT: 只要  $e_{t,k}$  不太快衰减. (有点矛盾, CLTS 要求抽样概率  $e_{t,k}$  的衰减速度慢于  $b_k$ ,

$$M_t / \sqrt{Tb_k^2} \Rightarrow N(0,1) \quad \begin{matrix} \text{这排除了强信号下衰弱衰减} \\ \text{反之} \end{matrix} \quad \begin{matrix} \checkmark \\ \log(T) \text{ 的抽样方案} \end{matrix}$$

↓

$$\sum_{t=1}^T \left( \frac{I(\{W_t=k\})(y_t - \hat{\mu}_k^{AW})}{\sqrt{e_{t,k}}} \right)^2 / (Tb_k^2) \rightarrow p 1$$

在鞅集中,

(前提是倾向性不太快衰减)

$$\hat{\mu}_k^{AW}$$
 和  $\hat{V}_k$  的  
分子会抵消

## Inference in adaptive experiments.

分析 UCB / Thompson 抽样所得的 data. (例 1 为底层问题参数提供置信声明)

这比 IID 中困难很多。

估计  $\mu_k$  下，自然想到两种估计量。

$$\text{样本均值估值: } \hat{\mu}_k^{\text{AVG}} = \hat{\mu}_{k,T} = \frac{1}{n_{k,T}} \sum_{j=1}^T I(\{w_j=k\}) Y_j \quad (14.11)$$

和:

Thompson 抽样下。

$$\text{IPW 估计量: } \hat{\mu}_k^{\text{IPW}} = \frac{1}{T} \sum_{t=1}^T \frac{I(\{w_t=k\}) Y_t}{e_{t,k}} \quad (14.12)$$

然而，由于自适应数据收集方案，这些估计量都没有渐近正态的极限分布，阻碍了它们制作置信区间。

可设计自适应加权的  $\mu_k$  估计量，允许了一个高斯枢轴 接近 or 渐进  $N(0,1)$

例子：

$$\hat{\mu}_k^{\text{AW}} = \frac{1}{T} \sum_{t=1}^T \frac{I(\{w_t=k\}) Y_t}{\sqrt{e_{t,k}}} \Bigg/ \sqrt{\frac{1}{T} \sum_{t=1}^T \frac{I(\{w_t=k\})}{e_{t,k}}} \quad (14.13)$$

在合理的正态性条件下：

$$\hat{V}_k^{-1/2} (\hat{\mu}_k^{\text{AW}} - \mu_k) \Rightarrow N(0,1)$$

$$\hat{V}_k = \sum_{t=1}^T \left( \frac{I(\{w_t=k\})(Y_t - \hat{\mu}_k^{\text{AW}})}{\sqrt{e_{t,k}}} \right)^2 \Bigg/ \left( \sum_{t=1}^T \frac{I(\{w_t=k\})}{\sqrt{e_{t,k}}} \right)^2 \quad (14.14)$$

这里能恢复一个 CLT (central limit Theorem) 是因为它们是方差稳定的。

即估计量方差可核对，才可满足 CLT.