Lecs 21-22

11/8/17 11:40 AM

Recap: Last time talked about estimating ATT using optimal balancing weights

$$\hat{\tau}_W = \frac{1}{n_1}\Sigma_{i\in\mathcal{I}_1} Y_i - \boxed{\Sigma_{i\in\mathcal{I}_0} W_i Y_i} = \frac{1}{n_0'}\Sigma_{i\in\mathcal{I}_0'} Y_i$$

$\mathcal{I}_0' \subseteq \mathcal{I}_0$

$n_0' = |\mathcal{I}_0'|$

$$\hat{\tau}_{W,\hat{f}_0} = \frac{1}{n}\Sigma_{i\in\mathcal{I}_1}(Y_i - \hat{f}_0(X_i)) - \Sigma_{i\in\mathcal{I}_0} W_i(Y_i - \hat{f}_0(X_i))$$

W optimizes the worst case (sometimes posterior)

$$\text{CMSE} = \mathbb{E}[(\hat{\tau} - \text{SATT})^2 / X_{1:n}, Z_{1:n}]$$

$$W \in \underset{W\in\mathcal{W}}{\operatorname{argmin}}\ \mathfrak{B}^2(W; \|\cdot\|) + \lambda\|W\|_2^2$$

$\mathcal{W}^{\text{simplex}} = \{W \geq 0 : \Sigma_i W_i = 1\}$

$\mathcal{W}^{n_0'\text{-subset}} = \mathcal{W}^{\text{simplex}} \cap \{0, \frac{1}{n_0'}\}^{n_0}$

$\mathcal{W}^{\text{subsets}} = \bigcup_{n_0'} \mathcal{W}^{n_0'\text{-subset}}$

$= \{W \in \{0, \alpha\}^{n_0} : \begin{matrix}\alpha > 0 \\ \Sigma_i W_i = 1\end{matrix}\}$

$\mathcal{W}^{\text{subset}} \subseteq \mathcal{W}^{\text{simplex}}$

Thm Suppose weak mean ignorability $(\mathbb{E}[Y(0)|X] = \mathbb{E}[Y(0)|X, Z])$ & Strong overlap $(\mathbb{P}(Z=1|X) \leq 1-\eta < 1)$

Let $\|\cdot\|$ be given by an RKHS with kernel $\mathcal{K}$. Suppose that:

1) For each n, W is given by $\text{KOM}(\mathcal{W}, \mathcal{K}, \lambda_n)$
2) $\lambda_n \in [\underline{\lambda}, \overline{\lambda}] \subsetneq (0, \infty)$
3) $\mathcal{W}^{\text{subsets}} \subseteq \mathcal{W}$ ← immediate if $\mathcal{W} = \mathcal{W}^{\text{simplex}}$
4) $\mathbb{E}[\mathcal{K}(X, X) | Z=1] < \infty$ ← immediate if $\mathcal{K}(x, x) \leq M$ e.g. Gaussian & Matérn kernels $e^{-\|x-x'\|}$
5) $\text{Var}(Y(0)|X)$ is a.s. bounded

Then: If $\|[f_0]\| < \infty$: $\hat{\tau}_W - \text{SATT} = O_p(1/\sqrt{n})$

If $\mathcal{K}$ is C_0-univ : $\hat{\tau}_W - \text{SATT} = o_p(1)$

we will show → (first); HW → (second)

Lemma: For random variables $Z_n \geq 0$ & any sigma algebra $\mathcal{G}$ then:

$$\mathbb{E}[Z_n|\mathcal{G}] = O_p(1) \Rightarrow Z_n = O_p(1)$$

$$\mathbb{E}[Z_n|\mathcal{G}] = o_p(1) \Rightarrow Z_n = o_p(1)$$

Simple cases: $\mathcal{G}$ = nothing $\Rightarrow$ Markov's thm

$\mathcal{G}$ = everything $\Rightarrow$ tautology

Pf of thm: We're going to exhibit a $W' \in \mathcal{W}^{\text{subsets}}$ s.t. $\mathfrak{B}(W'; \|\cdot\|) \to 0$ and $\|W'\|_2 \to 0$

By strong overlap $\exists \alpha > 0$ s.t.

$$q(X) = \frac{\alpha e(X)}{1 - e(X)} \in (0, 1) \quad \text{a.s.}$$

For each $i = 1, 2, \ldots$ let $\tilde{W}_i' = 1$ if $Z_i = 1$

$\tilde{w}_i' \sim \mathrm{Ber}(q(X_i))$ if $Z_i = 0$

$I_0' = \{i \in I_0 : \tilde{w}_i' = 1\}$

$n_0' = \sum_{i \in I_0} \tilde{w}_i'$

let $w_i' = 1/n_1$ if $Z_i = 1$

$= \tilde{w}_i' / \sum_{i \in I_0} \tilde{w}_i'$ if $Z_i = 0$

$w' \in \mathbb{W}$ subsets

$$X_i \Big|_{\substack{Z_i=0 \\ \tilde{w}_i'=1}} \sim X_i \,\big|\, Z_i = 1$$

Let $X_i^{(1)} \sim (X \mid Z = 1)$ be iid replicates
(ghost sample)

$$\mathbb{B}(w', \|\cdot\|) = \sup_{\|f\| \le 1} \left(\frac{1}{n_1} \sum_{i \in I_1} f(X_i) - \frac{1}{n_0'} \sum_{i \in I_0'} f(X_i) \right)$$

$\tilde{\xi}_i(f) = w_i'(f(X_i) - \mathbb{E}f(X_i^{(1)}))$

$$= \sup_{\|f\| \le 1} \Bigg(\frac{1}{n_1} \sum_{i \in I_1} \left(f(X_i) - \mathbb{E}[f(X_i^{(1)})] \right) - \frac{1}{n_0'} \sum_{i \in I_0'} \left(f(X_i) - \mathbb{E}[f(X_i^{(1)})] \right) \Bigg)$$

$$= \sup_{\|f\| \le 1} \Bigg(\underbrace{\frac{1}{n_1} \sum_{i \in I_1} \tilde{\xi}_i(f)}_{A_1(f)} - \frac{n_0}{n_0'} \underbrace{\frac{1}{n_0} \sum_{i \in I_0} \tilde{\xi}_i(f)}_{A_0(f)} \Bigg)$$

$$\le \underbrace{\|A_1\|_{\mathcal{H}}}_{\to 0} + \underbrace{\frac{n_0}{n_0'}}_{\to \text{const}} \underbrace{\|A_0\|_{\mathcal{H}}}_{\to 0}$$

$$\mathbb{E}\left[\|A_z\|_{\mathcal{H}}^2 \,\middle|\, Z_{1:n}\right] = \mathbb{E}\left[\left(\sup_{\|f\| \le 1} \frac{1}{n_z} \sum_{i \in I_z} w_i' \left(f(X_i) - \mathbb{E}\{f(X_i^{(1)})\} \right) \right)^2 \,\middle|\, Z_{1:n} \right]$$

Jensen's ineq

$$\le \mathbb{E}\left[\left(\sup_{\|f\| \le 1} \frac{1}{n_z} \sum_{i \in I_z} \overbrace{w_i' \left(f(X_i) - f(X_i^{(1)}) \right)}^{\xi_i} \right)^2 \,\middle|\, Z_{1:n} \right]$$

$$= \mathbb{E}\boxed{\left[\left\| \frac{1}{n_z} \sum_{i \in I_z} \xi_i \right\|_{\mathcal{H}}^2 \,\middle|\, Z_{1:n} \right]}$$

$\forall f$

$\mathbb{E}[\xi_i(f) \mid Z_{1:n}] = 0 \leftarrow$ by IPW magic

$$\mathbb{E}\left[\|\xi_i\|_X^2 \,\middle|\, Z_{1:n}\right] \le 4\, \mathbb{E}\left[\sup_f f(X_i) \,\middle|\, Z_i = 1\right]$$

$$= 4\, \mathbb{E}\left[\langle K(X_i, \cdot), K(X_i, \cdot) \rangle \,\middle|\, Z_i = 1\right]$$

$$= 4\, \mathbb{E}\left[K(X_i, X_i) \,\middle|\, Z_i = 1\right]$$

$$\le M < \infty$$

Let ξ_i' be iid replicates of ξ_i

$$\mathbb{E}\left[\|A_z\|_{\mathcal{H}}^2 \,\middle|\, Z_{1:n}\right] \le \mathbb{E}\left[\left(\sup_{\|f\| \le 1} \frac{1}{n_z} \sum_{i \in I_z} \xi_i(f) \right)^2 \,\middle|\, Z_{1:n} \right]$$

$$= \mathbb{E}\left[\left(\sup_{\|f\| \le 1} \frac{1}{n_z} \sum_{i \in I_z} \left(\xi_i(f) - \mathbb{E}\xi_i'(f) \right) \right)^2 \,\middle|\, Z_{1:n} \right]$$

Jensen's $$\leq \mathbb{E}\left[\left(\sup_{\|f\|\leq 1} \frac{1}{n_z} \sum_{i\in I_z}\left(\xi_i(f) - \xi_i'(f)\right)\right)^2 \Big| Z_{1:n}\right]$$

let ρ_i be iid Rademacher rvs

$$= \mathbb{E}\left[\left(\sup_{\|f\|\leq 1} \frac{1}{n_z} \sum_{i\in I_z} \rho_i\left(\xi_i(f) - \xi_i'(f)\right)\right)^2 \Big| Z_{1:n}\right]$$

$$\leq 4\, \mathbb{E}\left[\left(\sup_{\|f\|\leq 1} \frac{1}{n_z} \sum_{i\in I_z} \rho_i \xi_i(f)\right)^2 \Big| Z_{1:n}\right]$$

$$= \frac{4}{n_z^2} \mathbb{E}\left[\left\| \sum_{i\in I_z} \rho_i \xi_i \right\|^2 \Big| Z_{1:n}\right]$$

Key: Hilbert space

parallelogram law $$= \frac{4}{n_z^2} \cdot \sum_{i\in I_z} \mathbb{E}\left[\|\xi_i\|^2 \,\big|\, Z_{1:n}\right]$$

if not a hilbert space, then the relationship b/w these is the Rademacher type of the Banach space

$$\leq \frac{4M}{n_z} = O_p(1/n)$$

$$\Rightarrow \|A_z\|_{\mathcal{X}} = O_p(1/\sqrt{n})$$

$$n_0'/n_0 \longrightarrow \alpha/\mathbb{P}(Z=1) < \infty$$

$$\Rightarrow \mathbb{B}(w';\|\cdot\|) = O_p(1/\sqrt{n})$$

$$\|w'\|_2 = \sqrt{\sum_{i\in I_0'} \left(\frac{1}{n_0'}\right)^2} = 1/\sqrt{n_0'} = O_p(1/\sqrt{n})$$

B/c w' is feasible & w is optimal

$$\mathbb{B}^2(w;\|\cdot\|) + \lambda_n \|w\|_2^2 \leq \mathbb{B}^2(w';\|\cdot\|) + \lambda \|w'\|_2^2$$
$$= O_p(1/n)$$

$$\Rightarrow \mathbb{B}^2(w;\|\cdot\|) = O_p(1/n)$$

$$\|w\|_2^2 = \frac{1}{\lambda} O_p(1/n) = O_p(1/n)$$

Assumption: $\mathrm{Var}(Y(0)|X) \leq B$

In the case of $\|C[f_0]\| < \infty$,

$$\mathrm{CMSE}(\hat{\tau}_w) \leq \|C[f_0]\|\, \mathbb{B}^2(w;\|\cdot\|) + B \cdot \|w\|_2^2$$
$$= O_p(1/n)$$

By Lemma, $\hat{\tau}_w - \mathrm{SATT} = O_p(1/\sqrt{n})$. $\square$

Heterogeneous Effects & Estimating Them

If effects are very heterogeneous then is SATT/ATT/ATE/SATE still interesting?

We can never estimate $TE_i = Y_i(1) - Y_i(0)$

But we can narrow down our average effects:

$$CATE(x) = \mathbb{E}[TE \mid X = x]$$
$$= \mathbb{E}[Y_i(1) - Y_i(0) \mid X_i = x]$$

Notice $ATE = \mathbb{E}[TE] = \mathbb{E}[\mathbb{E}[TE \mid X]] = \mathbb{E}\, CATE(x)$

Different ways to estimate:

Regression: $CATE(x) = f_1(x) - f_0(x)$

$$= \mathbb{E}[Y(1) \mid X] - \mathbb{E}[Y(0) \mid X]$$

under weak mean ignorability $\rightarrow$ $= \mathbb{E}[Y \mid X, Z=1] - \mathbb{E}[Y \mid X, Z=0]$

estimate & plug in

$$\widehat{CATE}(x) = \hat{f}_1(x) - \hat{f}_0(x)$$

Refresher: weak ignorability: $Y_i(z) \perp\!\!\!\perp Z_i \mid X_i$ $\leftarrow$ $\mathbb{P}(Y_i(z) \mid Z_i, X_i) = \mathbb{P}(Y_i(z) \mid X_i)$

weak mean ignorability: $\mathbb{E}[Y_i(z) \mid Z_i, X_i] = \mathbb{E}[Y_i(z) \mid X_i]$

Outcom transformation:

$$Y_i^* = Y\left(\frac{Z}{e(x)} - \frac{1-Z}{1-e(x)}\right) = \frac{ZY(1)}{e(x)} - \frac{(1-Z)Y(0)}{1-e(x)}$$

$$\mathbb{E}[Y_i^* \mid X] = \mathbb{E}\left[\frac{Y(1)Z}{e(x)} \,\Big|\, X\right] - \mathbb{E}\left[\frac{Y(0)(1-Z)}{1-e(x)} \,\Big|\, X\right]$$

$$= \frac{\mathbb{E}[Y(1) \mid X]\,\overbrace{\mathbb{E}[Z \mid X]}^{e(x)}}{e(x)} - \frac{\mathbb{E}[Y(0) \mid X]\,\overbrace{\mathbb{E}[1-Z \mid X]}^{1-e(x)}}{1-e(x)}$$

$$= \mathbb{E}[Y(1) \mid X] - \mathbb{E}[Y(0) \mid X]$$
$$= CATE(x)$$

Fit $\hat{e}$, let $\tilde{Y}_i = Y_i\left(\frac{Z_i}{\hat{e}(X_i)} - \frac{1-Z_i}{1-\hat{e}(X_i)}\right)$,

regress $\tilde{Y}_i$ on X_i using your favorite algo

Regression approach to CATE:

$$\widehat{CATE}(x) = \underset{\text{fit on treated}}{\hat{f}_1(x)} - \underset{\text{fit on contro}}{\hat{f}_0(x)}$$

Same as fitting $Y_i \sim Z_i f_1(x_i) - (1-Z_i) f_0(x_i) + \varepsilon_i$

can instead fit the model

$$Y_i \sim \underset{\text{complex}}{g(x_i)} + \frac{Z_i - (1-Z_i)}{2} \underset{\text{simple}}{\delta(x_i)} + \varepsilon_i$$

$$\widehat{CATE} = \hat{\delta}$$

Different ways this manifests:

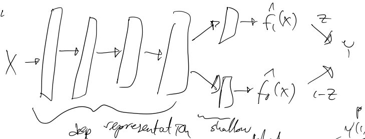
- g is linear, δ is constant
  $\Rightarrow$ OLS on Z_i, x_i
- g is very complicated but w/ limited capacity, δ is linear

  $\{\delta(x) < 0\}$ is a halfspace where we prefer to give $Z=0$

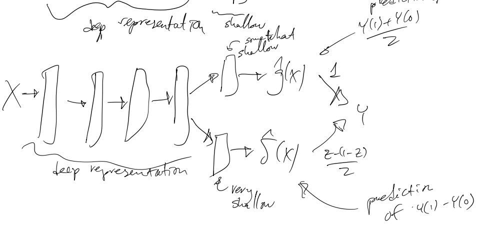
  $\{\delta(x) > 0\}$ —— // —— $Z=1$

- If using a neural net

a multitask learning approach



even better



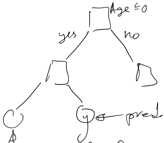
A targetted recursive partitioning approach to estimating CATE.

Maybe baseline is complex
but effect is simple, like piece-wise constant

So don't want diff of CART on each of control & treatment samples

Instead want a tree regression algo to learn CATE directly.

→ Causal Trees (Athey & Imbens '16)



mean diff of outcomes inside this partition

prediction for CATE

each leaf is same region in $\mathcal{X}$ space. Togther they from a partition of $\mathcal{X}$.

Consider any part $A \subset \mathcal{X}$

$$\hat{\tau}(x \in A) = \frac{\sum_{i \in I_1 : x_i \in A} Y_i}{\sum_{i \in I_1 : x_i \in A} 1} - \frac{\sum_{i \in I_0 : x_i \in A} Y_i}{\sum_{i \in I_0 : x_i \in A} 1}$$

Want to find a partition that has small

$$\mathbb{E}[(TE - \hat{\tau}(X))^2]$$

Let's assume that A is small enough s.t. $\boxed{X \perp\!\!\!\perp Z \mid X \in A}$

(Thinking about a tree with leaves

$A_1, \dots, A_\ell \subset \mathcal{X}$ $\quad A_i \cap A_j = \emptyset$ if j $\quad \left.\right\}$ partition

$A_1 \cup \dots \cup A_\ell = \mathcal{X}$

s.t. $\text{leafid}(x) = i$ s.t. $x \in A_i$

is a balancing score.)

then:
$$\begin{array}{cc} \mathbb{E}[\hat{\tau}(X) \mid X \in A] & \mathbb{E}[CATE(X) \mid X \in A] \\ \| & \| \\ \mathbb{E}[TE \mid X \in A] & = \mathbb{E}[\mathbb{E}[TE \mid X] \mid X \in A] \end{array}$$

$MSE_A =$
$$\mathbb{E}[(TE - \hat{\tau}(X))^2 \mid X \in A] =$$

$$= \mathbb{E}[((TE - CATE(X)) + (CATE(X) - \hat{\tau}(X)))^2 \mid X \in A]$$

$$= \mathbb{E}[(TE - CATE(X))^2 \mid X \in A] \leftarrow \text{does not depend on } \hat{\tau}$$

$$+ \mathbb{E}[\hat{\tau}^2(X) \mid X \in A] \leftarrow \text{does depend on } \hat{\tau}$$

So can estimate MSE_A (up to constant factors) inside A as

$$J(A) = \sum_{i : x_i \in A} \hat{\tau}^2(x_i)$$

$$= \hat{\tau}_A^2 \cdot |\{i : X_i \in A\}|$$

$$= n_A \cdot \left(\frac{\sum_{i \in J_1 : X_i \in A} Y_i}{n_{A,1}} - \frac{\sum_{i \in J_0 : X_i \in A} Y_i}{n_{A,0}} \right)^2$$

$\frac{1}{n} \sum_{j=1}^{\ell} I(A_j)$ ⟵ MSE of our tree regression for CATE (if sufficiently fine)

Causal Tree approach:
Recursive partition $\mathcal{X}$ so to minimize

Policy Learning

Why are we estimating CATE?
So to make personalized intervention decisions.

▷ How to directly learn such decision policies.
Instead of basing them on a first stage estimate $\widehat{CATE}$.

m discrete treatments

Want: find a policy $\pi : \mathcal{X} \to [m] = \{1, \dots, m\}$
that's good
Convention: larger outcomes = better
average reward due to π

$$R(\pi) = \mathbb{E}[\, Y(\pi(X)) \,]$$

Want to find $\hat{\pi}_n$ w/ $R(\hat{\pi}_n)$ high
from observational data $X_{1:n}, Z_{1:n}, Y_{1:n}$

Policy evaluation: Estimate $R(\pi)$ for given π
Policy learning: Find a good $\hat{\pi}_n$
Former usually leads to latter
via estimated reward maximization

Consider the case of two treatments: ± 1

$$\pi : \mathcal{X} \to \{\pm 1\}$$

Maximizing $R(\pi)$ is same a maximizing

$$\bar{R}(\pi) = R(\pi) - \left(\frac{R(x \mapsto 1) + R(x \mapsto -1)}{2} \right)$$

$$= \mathbb{E}[Y(\pi(X))] - \frac{1}{2}\mathbb{E}[Y(+1) + Y(-1)]$$

$$= \mathbb{E}[Y(\pi(X))] - \frac{1}{2}\mathbb{E}[Y(\pi(X)) + Y(-\pi(X))]$$

$$= \frac{1}{2}\mathbb{E}[Y(\pi(X)) - Y(-\pi(X))] = \frac{1}{2}(R(\pi) - R(-\pi))$$

$$= \frac{1}{2}\mathbb{E}[\pi(X)\,(Y(+1) - Y(-1))]$$

$$= \frac{1}{2}\mathbb{E}[\mathbb{E}[\pi(X)\,(Y(+1) - Y(-1)) \mid X]]$$

$$= \frac{1}{2}\mathbb{E}[\pi(X)\, CATE(X)]$$

For policy learning via estimated reward maximization need to estimate:

$$\bar{R}(\pi) = \frac{1}{2}\mathbb{E}[\pi(X)\, CATE(X)]$$

for all π's of interest $\pi \in \Pi$

Let's consider a few cases:

① If $\Pi = \{x \mapsto +1,\ x \mapsto -1\}$

Then estimating $\bar{R}(\pi)$ is just estimating ATE

$\bar{R}(x \mapsto +1) = \frac{1}{2} ATE$ $\quad \bar{R}(x \mapsto -1) = -\frac{1}{2} ATE$

② If $\Pi = [X \to \{\pm 1\}]$ all functions

Then having uniformly good estimates for $\bar{R}(\pi)$ for all $\pi \in \Pi$

$\Longleftrightarrow$

having uniformly good estimates for CATE(x) at all $x \in X$

Dangerous – no restrictions on class leads to overfitting

- might want to omit some variables from π, e.g. any var that's necessary for ignorability but that's not available at test time
- need to learn CATE well everywhere
  - susceptible to misspecification e.g. linear policy vs linear outcomes
- Usually simple policies work best

③ If Π has limited capacity $VC(\Pi) < \infty$
then try to learn $\hat{\pi}_n$ directly
via policy learning.

Same conclusions are going to hold
for several treatments

$$R(\pi) = \mathbb{E}[Y(\pi(X))] = \mathbb{E}[f_{\pi(x)}(x)]$$

$$f_z(x) = \mathbb{E}[Y(z)|X]$$

How to learn $\pi \in \Pi$?

(Assume ignorability throughout)

① Direct Method

$$R(\pi) = \mathbb{E}[f_{\pi(x)}(X)]$$

So given estimates $\hat{f}_z(x)$ of $\mathbb{E}[Y|X=x, Z=z]$
can form

$$\hat{R}_n^{DM}(\pi) = \frac{1}{n}\sum_i \hat{f}_{\pi(x_i)}(x_i)$$

In the case of binary ±1 treatments

$$\hat{R}_n^{DM}(\pi) = \frac{1}{n}\sum_i \pi(x_i)\left(\underbrace{\frac{\hat{f}_{+1}(x_i) - \hat{f}_{-1}(x_i)}{2}}_{\widehat{CATE}(x_i)}\right)$$

② IPW

let $e(z|x) = \mathbb{P}(Z=z|X=x)$

Then $e(Z,X)$ is called the
generalized propensity score (GPS)

for binary treatments

$$1/e(Z,X) = Z/e(X) + 1-Z/(1-e(X))$$

$$R(\pi) = \mathbb{E}[Y(\pi(X))]$$

$$= \mathbb{E}[\mathbb{E}[Y(\pi(X))|X]]$$

$$= \mathbb{E}\left[\mathbb{E}\left[Y(\pi(X))\,\middle|\,\begin{matrix}X\\ Z=\pi(X)\end{matrix}\right]\right]$$

$$= \mathbb{E}[\mathbb{E}[Y|X, Z=\pi(X)]]$$

$$= \mathbb{E}\left[\mathbb{E}\left[\frac{\mathbb{I}[Z=\pi(X)]\,Y}{\mathbb{P}(Z=\pi(X)|X)}\,\middle|\,X\right]\right]$$

$$= \mathbb{E}\left[\mathbb{E}\left[\frac{\mathbb{I}[Z=\pi(X)]\,Y}{e(\pi(X)|X)}\,\middle|\,X\right]\right]$$

$$= \mathbb{E}\left[\mathbb{E}\left[\frac{\mathbb{I}[Z=\pi(X)]\,Y}{e(Z|X)}\,\middle|\,X\right]\right]$$

$$= \mathbb{E}\left[\mathbb{I}[Z=\pi(X)]\,Y / e(Z|X)\right]$$

So given an estimate $\hat{e}$ of e can form

$$\hat{R}_n^{IPW}(\pi) = \frac{1}{n}\sum_i \frac{Y_i\,\mathbb{I}[Z_i=\pi(X_i)]}{\hat{e}(Z|X)}$$

In case of binary treatments ±1

$$\hat{R}_n^{IPW}(\pi) = \frac{1}{n}\sum_i \frac{1}{2}\cdot \underbrace{\frac{Y}{\hat{e}(Z|X)}}_{\frac{Y(1)Z_i}{\hat{e}(X_i)} - \frac{Y(0)(1-Z_i)}{1-\hat{e}(X_i)}} \cdot Z_i\,\pi(X_i)$$

$\leftarrow$ just like Y_i^* from the CATE transformation

③ AIPW same as DR approach for ATE

$$\hat{R}_n^{AIPW}(\pi) = \frac{1}{n}\sum_i\left(\hat{f}_{\pi(X_i)}(X_i) + \frac{\mathbb{I}[\pi(X_i)=Z_i]}{\hat{e}(Z_i|X_i)}\overbrace{\left(Y_i - \hat{f}_{Z_i}(X_i)\right)}^{\text{residuals from fitting } \hat{f}_1,\dots,\hat{f}_m}\right)$$

For binary treatments

$$\hat{R}_n^{AIPW}(\pi) = \frac{1}{n}\sum_i \frac{1}{2}\pi(X_i)\cdot\left(\hat{f}_{+1}(X_i) - \hat{f}_{-1}(X_i) + \frac{Z_i}{\hat{e}(Z_i|X_i)}\left(Y_i - \hat{f}_{Z_i}(X_i)\right)\right)$$