CI : ESTIMATING HETEROGENEOUS EFFECTS

• Set-up: observe (X_i, Y_i, W_i) with $X_i \in \mathbb{R}^p$, $Y_i = Y_i(W_i)$ (no spillover), $W_i \in \{0,1\}$

and assume unconfoundedness $\{Y_i(0), Y_i(1)\} \perp W_i \mid X_i$.
Up to now, the goal was the estimation of the ATE $\Delta^\infty = \mathbb{E}\{Y_i(1) - Y_i(0)\}$; which was achieved using several intermediate quantities:

↘ $e(x) = \mathbb{P}(W=1 \mid X=x)$ (propensity score)

↘ $\mu_{(w)}(x) = \mathbb{E}(Y_i(w) \mid X=x)$

For example, when $X \in \{1, .., K\}$ = discrete set, we saw that the ADM estimator $\hat{\Delta}_{ADM}$ satisfies a CLT $n^{1/2}(\hat{\Delta}_{ADM} - \Delta^\infty) \xrightarrow{d} \mathcal{N}(0, V_{ADM})$, where

$$V_{ADM} = \text{var}\{\Delta(X)\} + \mathbb{E}\left(\frac{\text{var}(Y_i(0) \mid X_i)}{1-e(X_i)} + \frac{\text{var}(Y_i(1) \mid X_i)}{e(X_i)}\right)$$

where
$$\Delta^*(X) = \mu_{(1)}(X) - \mu_{(0)}(X) = \mathbb{E}\{Y_i(1) - Y_i(0) \mid X\}$$

(p. 3/4 in CI: UNCONFOUNDEDNESS).

When estimating heterogeneous effects, $\Delta^*(X)$ is of primary interest and called the Conditional ATE

$$\boxed{\text{CATE}: \quad \Delta^*(x) = \mathbb{E}\{Y_i(1) - Y_i(0) \mid X = x\}}$$

I. INTRODUCTION TO META-LEARNERS

Meta-learners denote a family of algorithms that use ML estimators (base learners) to estimate the CATE. We introduce in this section three simple approaches:

- the S-learner
- the T-learner
- the X-learner

I.1. The S-learner.

In the S-learner, the treatment indicator is included as a predictor just like the other covariates X. We estimate $\mu(x, w) = \mathbb{E}(Y \mid X=x, W=w)$. The CATE estimator is then

$$\boxed{\hat{\Delta}_S(x) = \hat{\mu}(x, 1) - \hat{\mu}(x, 0).}$$

I.2. The T-learner

The T-learner estimates two separate conditional means $\mu_{(w)}(x) = \mathbb{E}(Y \mid X = x, W = w)$

The treatment group is used to estimate $\mu_{(1)}(x)$ and the control group is used to estimate $\mu_{(0)}(x)$.

$$\boxed{\hat{\Delta}_T(x) = \hat{\mu}_{(1)}(x) - \hat{\mu}_{(0)}(x)}$$

"T" for "two" learners.

Remark: Any imbalance in the treatment and control samples may lead to different levels of regularisation for estimating the $\mu_{(w)}(x)$ and a poor estimate of $\overset{*}{\Delta}(x)$. See for example Künzel et al (2017).

I.3. The X-learner.

The X-learner is an extension of the T-learner which addresses some of the regularisation problems mentioned above. It consists in the following steps:

(i) Produce estimates $\hat{\mu}_{(w)}(x)$ of $\mu_{(w)}(x)$ separately for $w=0$ and 1 (common step with the T-learner)

(ii) Compute
$$\begin{cases} \Delta_i^{(1)} = Y_i - \hat{\mu}_{(0)}(X_i) \text{ if } W_i = 1 \\ \Delta_i^{(0)} = \hat{\mu}_{(1)}(X_i) - Y_i \text{ if } W_i = 0 \end{cases}$$

trt effect for unit i

More trust on $\hat{\mu}_{(0)}(X)$ for values of X such that $e(X)$ is large

Put $\Delta^{(w)}(x) = \mathbb{E}\{Y_i(1) - Y_i(0) \mid X=x, W=w\}$

Then

$\{(X_i, \Delta_i^{(1)})\}$ = learning sample used to estimate $\Delta^{(1)}(x)$

$\{(X_i, \Delta_i^{(0)})\}$ = —— " —— $\Delta^{(0)}(x)$

More trust on the estimator $\hat{\Delta}^{(1)}(x)$ of $\Delta^{(1)}(x)$ for values of x such that $e(x)$ is large.

(iii) Since

$$\begin{aligned} \overset{*}{\Delta}(x) &= \mathbb{E}\{Y_i(1) - Y_i(0) \mid X=x\} \\ &= \mathbb{E}\,\mathbb{E}\{ \text{—"—} \mid X=x, W\} \\ &= \mathbb{P}(W=1 \mid X=x)\,\mathbb{E}\{\text{—"—} \mid X=x, W=1\} \\ &\quad + \mathbb{P}(W=0 \mid X=x)\,\mathbb{E}\{\text{—"—} \mid X=x, W=0\} \\ &= e(x)\,\Delta^{(1)}(x) + (1-e(x))\,\Delta^{(0)}(x), \end{aligned}$$

Put

$$\hat{\Delta}_X(x) = \hat{e}(x)\boxed{\hat{\Delta}^{(1)}(x)} + (1-\hat{e}(x))\boxed{\hat{\Delta}^{(0)}(x)}.$$

More weight is placed on the base learnes where more training data is available.

The X-learner was introduced by Künzel et al (2017).

II. ROBINSON'S LEGACY

Throughout this section we assume unconfoundedness:
$$\{Y_i(0), Y_i(1)\} \perp W_i \mid X_i. \qquad (*)$$

This allows us to write

$$\begin{aligned} \mu_{(w)}(x) &= \mathbb{E}\{Y_i(w) \mid X_i = x\} \\ &= \mathbb{E}\{Y_i(w) \mid X_i = x, W_i = w\} \\ &= \mathbb{E}\{Y_i \mid X_i = x, W_i = w\} \end{aligned}$$

$(*)$

consistency

$\Leftrightarrow\ Y_i = \mu_{(W_i)}(X_i) + \varepsilon_i(W_i)$ with $\mathbb{E}(\varepsilon_i(W_i) \mid X_i, W_i) = 0$

Together with $\Delta^*(x) = \mu_{(1)}(x) - \mu_{(0)}(x)$, we may write

$$\boxed{Y_i = \mu_{(0)}(X_i) + W_i\,\Delta^*(X_i) + \varepsilon_i(W_i)}$$

Called a Partially Linear Model (PLM)

We may center the outcome variable and consider

$$m(x) := \mathbb{E}(Y_i \mid X_i = x) = \mu_{(0)}(x) + e(x)\,\Delta^*(x)$$

$$\Rightarrow \boxed{Y_i - m(X_i) = (W_i - e(X_i))\,\Delta^*(X_i) + \varepsilon_i} \quad (**)$$

$\varepsilon_i = \varepsilon_i(W_i)$

This class of problems was studied by Robinson (1988) and is the starting point of many modern techniques for estimating the CATE:

- ↘ Double ML of Chernozhukov et al (2018)
- ↘ R-learners of Nie & Wager (2020)
- ↘ Causal Forests of Athey, Tibshiranie & Wager (2019)

II.1. Double ML

Expression (**) is the starting point of the 3 papers mentioned above for estimating the CATE. When $\Delta^*(x) \equiv \Delta^* \equiv$ constant (no treatment heterogeneity), the relation $Y_i - m_i(X_i) = (W_i - e(X_i))\,\Delta^* + \varepsilon_i$ suggests regressing $Y_i - m(X_i)$ on $W_i - e(X_i)$ to get an estimate of Δ^* (the oracle). The difficulty is that we do not know $m(x)$ and $e(x)$. These must be estimated from the data. Unfortunately, a direct estimation of m and e that are then plugged back into (**) typically leads to estimators of Δ that are heavily biased. Chernozhukov et al (2018) showed however that the use of cross-fitting can be used to emulate the oracle. The set of algorithms making use of (**) together with cross-fitting are referred to as Double ML (DML) by the authors.

The discussion above for a constant $\Delta^*(x) = \Delta^*$ extends to CATE provided we assume a linear (parametric) model for $\Delta^*(x) := x^t \beta$. $\beta \in \mathbb{R}^p$

or making use of a basis function $\varphi(x) \in \mathbb{R}^p$.

We proceed as follows:

(i) Divide the data into K folds.
Compute estimators $\hat{m}^{(-k)}(x)$ and $\hat{e}^{(-k)}(x)$ by regressing $Y \sim X$ and $W \sim X$ non-parametrically, excluding the k-th fold.

(ii) Define the transformed features
$\tilde{Y}_i = Y_i - \hat{m}^{(-k(i))}(X_i)$, $\tilde{W}_i = X_i\big(W_i - \hat{e}^{(-k(i))}(X_i)\big)$

where $k(i)$ = mapping taking observation i and placing it into the k-th fold.

(iii) Estimate $\hat{\beta} \leftarrow OLS(\tilde{Y}_i \sim \tilde{W}_i)$

Denoting $\beta^* \leftarrow OLS\left(Y_i - m(X_i) \sim (W_i - e(X_i))X_i\right)$,

the oracle, since it uses the true $m(x)$ and $e(x)$.

Then one can show that $n^{1/2}(\beta^* - \beta) \xrightarrow{d} \mathcal{N}(0, V)$ for some covariance matrix V. If all non-parametric regressions satisfy

$$(\circ) \begin{cases} \left[\mathbb{E}(\hat{m}(X) - m(X))^2\right]^{1/2} = o_p(n^{-1/4}) \\ \left[\mathbb{E}(\hat{e}(X) - e(X))^2\right]^{1/2} = o_p(n^{-1/4}), \end{cases}$$

then cross-fitting emulates the oracle:

$n^{1/2}(\hat{\beta} - \beta^*) \xrightarrow{P} 0$,

which ensures that $n^{1/2}(\hat{\beta} - \beta) \xrightarrow{d} \mathcal{N}(0, V)$ as well. We will revisit conditions of the form $(\circ)$ when discussing causal forests. For more details regarding DML and alternative approaches, see Chernozhukov et al (2018).

II.2. R learners

Chernozhukov et al (2018)'s contribution is to show how ML models can be succesfully applied for estimating nuisance parameters ($m(x)$ and $e(x)$) for semi-parametric inference. However, their approach requires a parametric model for the CATE. Nie and Wager ('20) use Robinson's transformation $(**)$ differently. They note that $(**)$ can be equivalently expressed as:

$$\Delta^*(\cdot) = \underset{\Delta}{\operatorname{argmin}} \left\{ \mathbb{E}\left([Y_i - m(X_i)] - [W_i - e(X_i)]\Delta(X_i)\right)^2 \right\}$$

definition of a loss function $\Rightarrow$ no need for a parametric model for $\Delta(x)$.
$\Delta(x)$ can be estimated via empirical loss minimization

$$\hat{\Delta}^*_R(\cdot) = \underset{\Delta}{\operatorname{argmin}} \left\{ \frac{1}{n}\sum_{i=1}^{n}\left([Y_i - m(X_i)] - [W_i - e(X_i)]\Delta(X_i)\right) + \Lambda_n(\Delta(\cdot)) \right\}$$

The R-learner

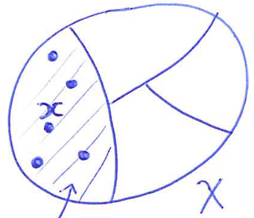
Regularizer on the complexity of $\Delta(\cdot)$.

This approach can be implemented using many possible variants = kernel ridge regression, boosting, deep learning.

In practice, we do not know the nuisance parameters $m(\cdot)$ and $e(\cdot)$. Instead, we consider plug-in alternatives with cross-fitting. Quasi-oracle properties are established in Nie and Wager (2020).

II. 3. Causal Forests.

Causal trees (forests) split the feature space $\mathcal{X}$ into regions where the CATE is believed constant.



$\Delta(x) =$ constant for all x belonging in this region (even if we expect globally some heterogeneity)

Robinson's transformation (**) simplifies to

$$\boxed{Y_i - m(X_i) = (W_i - e(X_i))\,\Delta^* + \varepsilon_i}$$

If we knew the nuisance parameters $m(x)$ and $e(x)$, we could compute the oracle OLS estimator

$$\hat{\Delta}^* \leftarrow \text{OLS}\Big(\underbrace{Y_i - m(X_i)}_{\text{residual}} \sim \underbrace{W_i - e(X_i)}_{\text{residual}}\Big)$$

In practice, we consider the plug-in alternative

$$\hat{\Delta} \leftarrow \text{OLS}\big(Y_i - \hat{m}(X_i) \sim W_i - \hat{e}(X_i)\big)$$

$$= \frac{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{m}(X_i))(W_i - \hat{e}(X_i))}{\frac{1}{n}\sum_{i=1}^{n}(W_i - \hat{e}(X_i))^2}$$

While $n^{1/2}(\hat{\Delta}^* - \Delta^*) \xrightarrow{d} \mathcal{N}(0, V_{PLM})$, a direct plug-in of the non-parametric estimates $\hat{m}$ and $\hat{e}$ yield $n^{1/2}(\hat{\Delta} - \hat{\Delta}^*) \not\xrightarrow{P} 0$ and a CLT does not hold for the feasible estimator $\hat{\Delta}$.

As usual, cross-fitting will help here. We provide some details next.

Cross-fitting (simple case)

(i) Split the sample $\{1, \dots, n\}$ into I_1 and I_2.

(ii) Estimate $\hat{m}_{I_1}(x)$ by predicting Y from X on I_1

(iii) Estimate $\hat{m}_{I_2}(x)$ by predicting Y from X on I_2

(iv) Similarly for $\hat{e}_{I_1}(x)$ and $\hat{e}_{I_2}(x)$.

(v) Compute

$$\hat{\Delta} = \frac{\frac{1}{n/2}\sum_{i\in I_1}(Y_i - \hat{m}_{I_2}(X_i))(W_i - \hat{e}_{I_2}(X_i)) + \frac{1}{n/2}\sum_{i\in I_2}(Y_i - \hat{m}_{I_1}(X_i))(W_i - \hat{e}_{I_1}(X_i))}{\frac{1}{n/2}\sum_{i\in I_1}(W_i - \hat{e}_{I_2}(X_i))^2 + \frac{1}{n/2}\sum_{i\in I_2}(W_i - \hat{e}_{I_1}(X_i))^2}$$

Take one sample to fit and one sample to evaluate.

- **Claim** If $\sqrt{\mathbb{E}(\hat{m}(X) - m(X))^2} = o_p(n^{-1/4})$
  & $\sqrt{\mathbb{E}(\hat{e}(X) - e(X))^2} = o_p(n^{-1/4})$
  Then $n^{1/2}(\hat{\Delta} - \hat{\Delta}^*) = o_p(1)$
  & $n^{1/2}(\hat{\Delta} - \Delta^*) \xrightarrow{d} \mathcal{N}(0, V_{PLM})$

A CLT holds under relatively weak / general conditions on the accuracy of $\hat{m}(\cdot)$ and $\hat{e}(\cdot)$. Conditions stated here are sufficient. We can improve on them.

(11)

proof : We focus first on the numerator

$$\frac{2}{n}\sum_{i\in I_1}\Big\{(Y_i-\hat m_{I_2}(X_i))(W_i-\hat e_{I_2}(X_i)) - (Y_i-m(X_i))(W_i-e(X_i))\Big\}$$

quantities with hats compared with the oracle.

we want to show that this quantity is $o_p(n^{-1/2})$. We decompose it into three terms :

$$= \frac{2}{n}\sum_{i\in I_1}(Y_i-m(X_i))(e(X_i)-\hat e_{I_2}(X_i)) \quad \text{——— (A)}$$

$$+ \frac{2}{n}\sum_{i\in I_1}(m(X_i)-\hat m_{I_2}(X_i))(W_i-e(X_i)) \quad \text{——— (B)}$$

$$+ \frac{2}{n}\sum_{i\in I_1}(\hat m_{I_2}(X_i)-m(X_i))(\hat e_{I_2}(X_i)-e(X_i)) \quad \text{— (C)}$$

- By cross-fitting, in (A)

$$\mathbb{E}\{(Y_i-m(X_i))(e(X_i)-\hat e_{I_2}(X_i)) \mid I_2\} = 0$$

$$\Rightarrow \mathbb{E}(A \mid I_2) = 0 \text{ and}$$

$$\mathbb{E}A^2 = \mathbb{E}\,\mathbb{E}(A^2 \mid I_2)$$

zero mean cond on I_2

$$= \mathbb{E}\,\mathrm{var}(A \mid I_2)$$

uncorrelated

$$= \frac{4}{n^2}\sum_{i\in I_1}\mathbb{E}\mathrm{var}\{(Y_i-m(X_i))(e(X_i)-\hat e_{I_2}(X_i)) \mid I_2\}$$

key-step: we earn a factor $1/n$ since iid obs.

$$= \frac{2}{n}\mathbb{E}\mathrm{var}\{(Y_i-m(X_i))(e(X_i)-\hat e_{I_2}(X_i)) \mid I_2\}$$

(12)

$$= \frac{2}{n}\mathbb{E}\,\mathbb{E}\{(Y_i-m(X_i))^2(e(X_i)-\hat e_{I_2}(X_i))^2 \mid I_2\}$$

$$\leq \frac{2\sigma^2}{n}\mathbb{E}(e(X_i)-\hat e_{I_2}(X_i))^2$$

$$= o_p(n^{-3/2})$$

It follows from Markov-Chebyshev that $(A) = o_p(n^{-0.75})$

More than we need (so there is room for errors)

- The term (B) is treated similarly.
- Regarding (C), we use Cauchy-Schwartz :

$$\frac{2}{n}\sum_{i\in I_1}(\hat m_{I_2}(X_i)-m(X_i))(\hat e_{I_2}(X_i)-e(X_i))$$

$$\leq \sqrt{\frac{2}{n}\sum_{i\in I_1}(\hat m_{I_2}(X_i)-m(X_i))^2} \times \sqrt{\frac{2}{n}\sum_{i\in I_1}(\hat e_{I_2}(X_i)-e(X_i))^2}$$

$$= o_p(n^{-1/2})$$

- The denominator is treated similarly:

$$\frac{2}{n}\sum_{i\in I_1}\{(W_i-\hat e_{I_2}(X_i))^2-(W_i-e(X_i))^2\}$$

$$= \underbrace{\frac{2}{n}\sum_{i\in I_1}(e(X_i)-\hat e_{I_2}(X_i))^2}_{=o_p(n^{-1/2})} + \underbrace{\frac{2}{n}\sum_{i\in I_1}(W_i-e(X_i))(e(X_i)-\hat e_{I_2}(X_i))}_{\text{treated as on page 11 } = o_p(n^{-0.75})}$$

× **Remark** = An active area of research is refining (C) and throw away Cauchy-Schwartz [CS is very crude, and provide a good upper bound when the errors $(\hat{m} - m)$ and $(\hat{e} - e)$ are aligned]. Note that for terms (A) and (B), we only need consistency of $\hat{m}$ and $\hat{e}$ due to the extra averaging $1/n$.

• **General notation:** If K folds for cross-filting $I_1, \dots, I_K$, denote by $\hat{m}^{(-k)}$ and $\hat{e}^{(-k)}$ the estimate of m and e trained over all folds excluding I_k.

× **Meta-Algorithm**

(i) Estimate $m(x) = \mathbb{E}(Y \mid X = x)$
$e(x) = \mathbb{P}(W = 1 \mid X = x)$
with cross-filting; producing K estimates if K folds $I_1, \dots, I_K$ are used

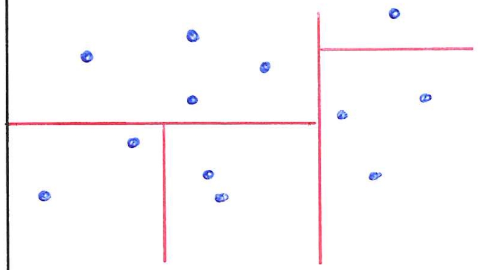
(ii) Cut the space into neighborhoods: for each $x \in \mathcal{X}$, define a neighborhood $\mathcal{N}(x)$.

(iii) $\hat{\Delta}(x) \leftarrow \mathrm{OLS}\Big(Y_i - \hat{m}^{(-k(i))}(X_i) \sim W_i - \hat{e}^{(-k(i))}(X_i) \,;\, X_i \in \mathcal{N}(x)\Big)$

$k(i)$ = fold containing obs i

One question remains: how to pick $\mathcal{N}(x)$?

Pick neighbours adaptively, growing a CART tree



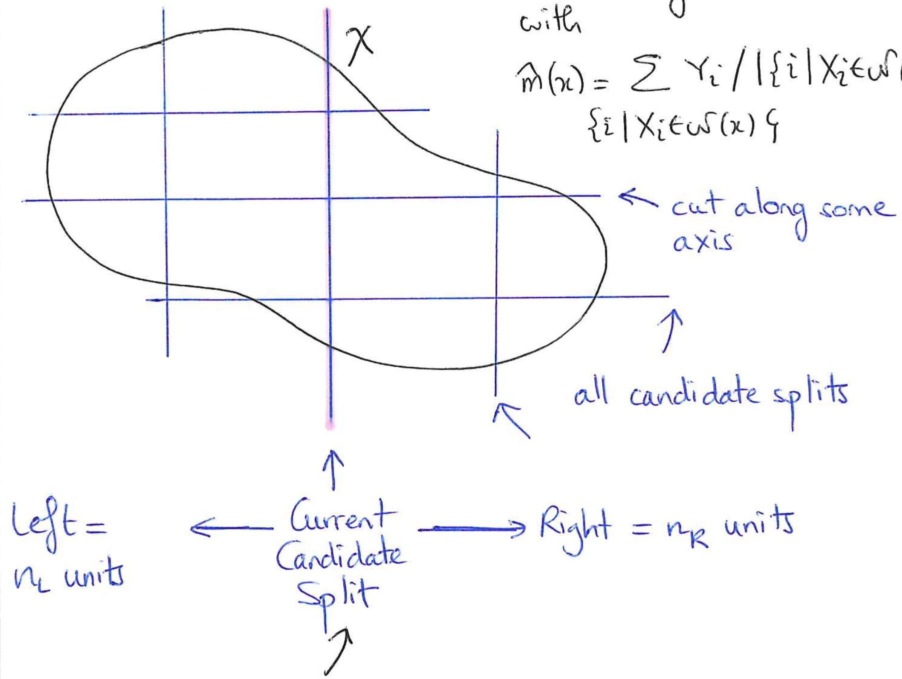
Review = CART as a usual ML algorithm, trained on $\{(X_i, Y_i)\}$ iid

feature

target

estimating $m(x) = \mathbb{E}(Y \mid X = x)$ with

$$\hat{m}(x) = \sum_{\{i \mid X_i \in \mathcal{N}(x)\}} Y_i \Big/ \big|\{i \mid X_i \in \mathcal{N}(x)\}\big|$$



Does the current candidate split represents useful heterogeneity?

On L, compute $\hat{m}_L = \bar{Y}_L$ = average outcome on the left region

On R, compute $\hat{m}_R = \bar{Y}_R$;
and choose the split that minimizes the total prediction error, across leaves:

$$\ell_{split} = \sum_{i\in L}(Y_i-\bar{Y}_L)^2 + \sum_{i\in R}(Y_i-\bar{Y}_R)^2$$

$$= \sum_{i\in L}\left\{(Y_i-\bar{Y})^2 - (\bar{Y}_L-\bar{Y})^2\right\} + \sum_{i\in R}\left\{(Y_i-\bar{Y})^2-(\bar{Y}_R-\bar{Y})^2\right\}$$

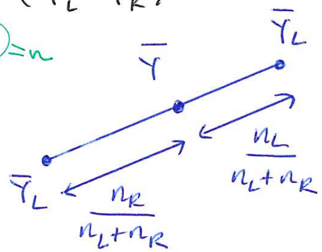
$$= \sum_{i=1}^{n}(Y_i-\bar{Y})^2 - n_L(\bar{Y}_L-\bar{Y})^2 - n_R(\bar{Y}_R-\bar{Y})^2$$

$$= \sum_{i=1}^{n}(Y_i-\bar{Y})^2 - \frac{2 n_L n_R}{\underbrace{n_L+n_R}_{=n}}(\bar{Y}_L-\bar{Y}_R)^2$$

$\Leftrightarrow$

select split such that

$$\boxed{\max\ n_L n_R(\bar{Y}_L-\bar{Y}_R)^2}$$



↑ Selecting regions that minimize heterogeneity within
≡ Selecting regions that maximize heterogeneity across.

Back to our estimation of $\Delta(x)$, compute the following quantities for each candidate split:

$$\hat{\Delta}_R \leftarrow OLS\left(Y_i - \hat{m}^{(-k(i))}(X_i) \sim W_i - \hat{e}^{(-k(i))}(X_i)\ ;\ X_i \in R\right)$$

$$\hat{\Delta}_L \leftarrow OLS\left(\text{—"—} \sim \text{—"—}\ ;\ X_i \in L\right)$$

& Choose the split = $\arg\max\left(n_L n_R(\hat{\Delta}_R-\hat{\Delta}_L)^2\right)$;
& recurse
& stop eventually.

× Remark = Initially, there will be a lot of heterogeneity across the L/R regions; while the OLS assumes constant CATE in both L and R. One can show that when $\Delta(x)$ is not constant, the OLS estimator converges to

$$\frac{\mathbb{E}\{e(X)(1-e(X))\Delta(X)\}}{\mathbb{E}\{e(X)(1-e(X))\}}$$

↑ a weighted version of $\mathbb{E}\{\Delta(X)\}$.

Li, Morgan, Zaslavsky (2018)
& p.12 in CI: UNCONFOUNDEDNESS

To turn a single causal tree into a forest, let

$$\alpha_{b,i}(x) = \frac{\mathbb{1}\{X_i \in L_b(x)\}}{|\{i \mid X_i \in L_b(x)\}|}$$

← $L_b(x)$ = leaf of the tree where x is located.

b-th tree

In region $L_b(x)$, the CATE $\Delta(x)$ is constant and estimated to be

$$\hat{\Delta}_L(x) \leftarrow OLS\left(Y_i - \hat{m}^{(-k(i))}(X_i) \sim W_i - \hat{e}^{(-k(i))} \ \middle|\ \text{weight obs } i = \alpha_{b,i}(x)\right)$$

constant $\forall i$ in leaf $L_b(x)$

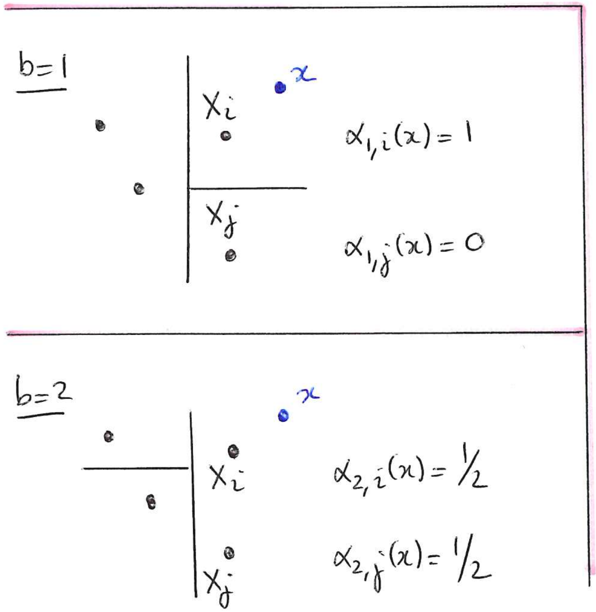
$\hat{\Delta}_L(x)$ = solution of a LS weighted problem

(since the weights are constant within a region, we could as well have used $\text{weight}_i = 1 \quad \forall i \in L_b(x)$)

$\searrow$ But this representation allows us to generalize to a forest. Simply consider

$$\text{weight}_i = \alpha_i(x) = \underbrace{\frac{1}{B}}_{\text{grow } B \text{ trees}} \sum_{b=1}^{B} \alpha_{b,i}(x).$$

R library = grf



$\alpha_i(x)$ = how often observation i falls in the same leaf as x

$$\Rightarrow \begin{cases} \alpha_i(x) = 3/4 \\ \alpha_j(x) = 1/4 \\ \alpha_k(x) = 0 \qquad k \neq i, j. \end{cases}$$