Heterogeneous Treatment Effects Estimation: When Machine Learning meets multiple treatments regime

Causal TAU Seminar, Inria Saclay & LISN, Gif-sur-Yvette

Naoufal Acharki [1,2], Josselin Garnier[1] and Antoine Bertoncello[2]

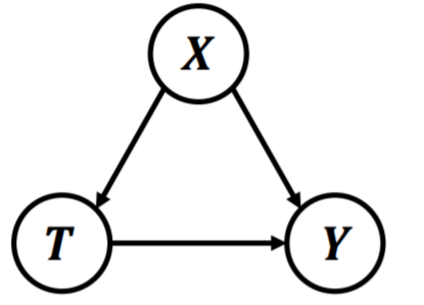
June 23, 2022

[1]Centre de Mathématiques Appliquées, Ecole Polytechnique. [2]TotalEnergies One Tech.

Potential outcome theory and Rubin Causal model

Rubin Causal Model with mutli-treatments

- $i = 1, \dots, n$: an individual subject to a treatment.
- T: the treatment assignment variable.
- $\mathcal{T} = \{t_0, t_1, \dots, t_K\}$: the set of possible treatments.
  Historically, T is binary and $\mathcal{T} = \{0, 1\}$.
- $\boldsymbol{X} \in \mathbb{R}^d$: vector of d covariates (confounders).
- $Y_{\text{obs}} = Y(T)$: the observed outcome corresponding to the treatment T.
- $Y(t)$: the counter-factual outcome that would have been observed under treatment level $t \in \mathcal{T}$.



Rubin Causal Model [Rubin, 1974]

Goal: Estimate the Causal Effect of the treatment T on the outcome Y.

Assumptions of RCM

Consistency: For an individual i, we observe the potential outcome associated to assigned treatment T_i

$$Y_{obs,i} = Y_i(T_i)$$

Unconfoundedness: Given the covariates $\boldsymbol{X}$, the treatment mechanism is unconfounded for all treatment levels

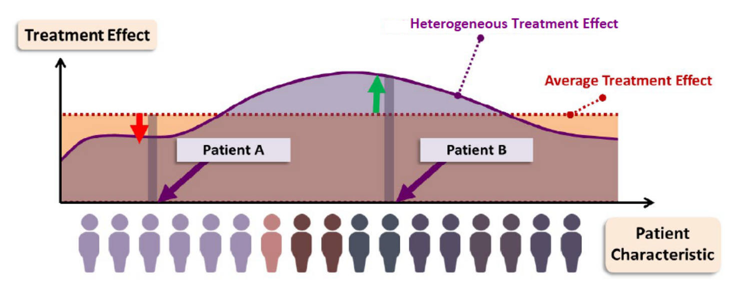
$$\forall t \in \mathcal{T},\ \mathbf{1}\{T = t\} \perp\!\!\!\perp Y(t) \mid \boldsymbol{X}$$

Positivity: Each individual has a positive probability of receiving any dose of treatment t when given the observed covariates

$$\forall t \in \mathcal{T}, \forall \boldsymbol{x} \in \mathbb{R}^d \quad 0 < \mathbb{P}(T = t | \boldsymbol{X} = \boldsymbol{x}) < 1.$$

Why Heterogeneous Treatment Effects?

Challenge 1: A treatment may affect individuals differently. We need to conduct group-level comparisons.



The treatment effect within a sub-group with covariates $\boldsymbol{x}$ is modelled by the Conditional Average Treatment Effect (CATE)

$$\tau_t(\boldsymbol{x}) = \mathbb{E}[Y(t) - Y(t_0)|\boldsymbol{X} = \boldsymbol{x}].$$

Estimation of CATEs using Machine Learning

Challenge 2: This is not a standard ML supervised learning problem

Machine Learning - Mitchell [1997]

A computer program is said to learn from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E

Experience E = Supervised Learning i.e. Regression of $Y(t) - Y(t_0)$ on the covariates $\boldsymbol{X}$.
$Y_i(t) - Y_i(t_0)$ is not observed for each unit i. This is the fundamental problem of causal inference [Holland, 1986].

Task T = Prediction of the CATE τ_t for a given sub-group with covariates $\boldsymbol{x}$.

Performance Measure P = Accuracy, Precision, RMSE etc. The counterfactual prediction is counterfactual by definition, it cannot be measured without knowing the ground truth model.

Meta-Learners for estimating the CATE

What is a meta-learner?

A Meta-learner [Künzel et al., 2019] is a statistical framework that models and estimate the CATE

$$\tau_t(\boldsymbol{x}) = \mathbb{E}[Y(t) - Y(t_0)|\boldsymbol{X} = \boldsymbol{x}]$$

No model restrictions: any supervised ML method can be used.

Direct plug-in meta-learners

Definition: Naive estimators that estimate the CATE directly by a plug-in difference.

Direct plug-in meta-learners

Definition: Naive estimators that estimate the CATE directly by a plug-in difference.

The **T-learner** (T stands for *two*):

- Consider two models models μ_t and μ_{t_0}, where $\mu_w(\boldsymbol{x}) = \mathbb{E}(Y(w)|\boldsymbol{X} = \boldsymbol{x})$ for $w \in \{t, t_0\}$
- Estimate $\hat{\mu}_t$ by regressing $Y(t)$ on $\boldsymbol{X}$ using $\mathbf{S}_t = \{i, T_i = t\}$. Do the same for $\hat{\mu}_{t_0}$.
- Compute the CATE as plug-in difference $\hat{\tau}_T(\boldsymbol{x}) = \hat{\mu}_t(\boldsymbol{x}) - \hat{\mu}_{t_0}(\boldsymbol{x})$

Direct plug-in meta-learners

Definition: Naive estimators that estimate the CATE directly by a plug-in difference.

The **T-learner** (T stands for *two*):

- Consider two models models μ_t and μ_{t_0}, where $\mu_w(\boldsymbol{x}) = \mathbb{E}(Y(w)|\boldsymbol{X} = \boldsymbol{x})$ for $w \in \{t, t_0\}$
- Estimate $\hat{\mu}_t$ by regressing $Y(t)$ on $\boldsymbol{X}$ using $\mathbf{S}_t = \{i, T_i = t\}$. Do the same for $\hat{\mu}_{t_0}$.
- Compute the CATE as plug-in difference $\hat{\tau}_T(\boldsymbol{x}) = \hat{\mu}_t(\boldsymbol{x}) - \hat{\mu}_{t_0}(\boldsymbol{x})$

The **S-learner** (S stands for *single*):

- Consider a single model μ such that $\mu(w, \boldsymbol{x}) = \mathbb{E}(Y_{\text{obs}} \mid T = w, \boldsymbol{X} = \boldsymbol{x})$.
- Estimate $\hat{\mu}$ by regressing Y_{obs} on both $\boldsymbol{X}$ and T using all observed data.
- Compute the CATE as plug-in difference $\hat{\tau}_S(\boldsymbol{x}) = \hat{\mu}(t, \boldsymbol{x}) - \hat{\mu}(t_0, \boldsymbol{x})$.

Pseudo-outcome meta-learners

Definition: Learners that target the CATE directly by regressing a pseudo-outcome Z_t on $\boldsymbol{X}$.
Here $r(t, \boldsymbol{X}) = \mathbb{P}(T = t \mid \boldsymbol{X})$ is the GPS and $\mu_t = \mathbb{E}(Y(t) \mid \boldsymbol{X})$.

Pseudo-outcome meta-learners

Definition: Learners that target the CATE directly by regressing a pseudo-outcome Z_t on $\boldsymbol{X}$. Here $r(t, \boldsymbol{X}) = \mathbb{P}(T = t \mid \boldsymbol{X})$ is the GPS and $\mu_t = \mathbb{E}(Y(t) \mid \boldsymbol{X})$.

M-learner: (M stands for *modified*)

$$Z_t^M = \frac{\mathbf{1}\{T = t\}}{r(t, \boldsymbol{X})} Y_{\text{obs}} - \frac{\mathbf{1}\{T = t_0\}}{r(T = t_0, \boldsymbol{X})} Y_{\text{obs}}.$$

DR-learner: (DR stands for *Doubly-Robust*)

$$Z_t^{DR} = \frac{Y_{\text{obs}} - \mu_T(\boldsymbol{X})}{r(T = t, \boldsymbol{X})} \mathbf{1}\{T = t\} - \frac{Y_{\text{obs}} - \mu_T(\boldsymbol{X})}{r(t_0, \boldsymbol{X})} \mathbf{1}\{T = t_0\} + \mu_t(\boldsymbol{X}) - \mu_{t_0}(\boldsymbol{X}).$$

X-learner: (X stands for *Cross* estimation procedure)

$$Z_t^X = \mathbf{1}\{T = t\}(Y_{\text{obs}} - \mu_{t_0}(\boldsymbol{X})) + \sum_{t' \neq t} \mathbf{1}\{T = t'\}(\mu_t(\boldsymbol{X}) - Y_{\text{obs}}) + \sum_{t' \neq t} \mathbf{1}\{T = t'\}(\mu_{t'}(\boldsymbol{X}) - \mu_{t_0}(\boldsymbol{X})).$$

Neyman orthogonality based learners: R-learner

Definition: Learners that use the Neyman-Orthogonality and the Robinson [1988] decomposition to address a minimization problem with respect to a causal component.

R-Learner: Estimate all $K-1$ CATE models $\{\tau_t\}_{t\neq 0}$ by addressing:

$$\{\widehat{\tau}_t^{(\mathrm{R})}\}_{t\neq t_0\in\mathcal{T}} = \underset{\{\tau_t\}_{t\neq t_0}\in\mathcal{F}}{\arg\min} \frac{1}{n}\sum_{i=1}^{n}\Big[\big(Y_{\mathrm{obs},i} - m(\boldsymbol{X}_i)\big) - \sum_{t\neq t_1\in\mathcal{T}}\big(\mathbf{1}\{T_i = t\} - r(t,\boldsymbol{X}_i)\big)\tau_t(\boldsymbol{X}_i)\Big]^2$$

where $r(t,\boldsymbol{x}) = \mathbb{P}(T = t \mid \boldsymbol{X} = \boldsymbol{x})$, $m(\boldsymbol{x}) = \mathbb{E}(Y_{\mathrm{obs}} \mid \boldsymbol{X} = \boldsymbol{x})$ and $\mathcal{F}$ is the space of candidate models (e.g. linear models).

Comparison of meta-learners

Meta-learner	Advantages	Disadvantages
T-learner	✓ Simple approach	✗ Selection bias ✗ Low sample regime
S-learner	✓ Simple approach	✗ Confounding effects ✗ Regularization bias
M-learner	✓ Consistency*	✗ High variance
DR-learner	✓ Consistency* ✓ Doubly Robust	✗ High variance
X-learner	✓ Consistency* ✓ Low variance	✗ Too complex
R-learner	✓ Flexible representation	✗ Heavy problem ✗ Consistency?

The consistency of pseudo-outcome meta-learners

A pseudo-outcome meta-learner is said to be *consistent if* $\mathbb{E}(Z_t \mid \boldsymbol{X} = \boldsymbol{x})$ gives an unbiased estimation of the CATE $\tau_t(\boldsymbol{x})$.

The pseudo-outcome random Z_t incorporate the GPS r and the outcome model $\mu_{\cdot}$. they are called *nuisance parameters*.

The consistency of pseudo-outcome meta-learners

A pseudo-outcome meta-learner is said to be *consistent if* $\mathbb{E}(Z_t \mid \boldsymbol{X} = \boldsymbol{x})$ gives an unbiased estimation of the CATE $\tau_t(\boldsymbol{x})$.

The pseudo-outcome random Z_t incorporate the GPS r and the outcome model $\mu_{\cdot}$. they are called *nuisance parameters*.

In reality, you need to first the nuisance parameters (now $\widehat{r}$ and $\widehat{\mu}_{\cdot}$) to have the pseudo-outcome vector $\boldsymbol{z}_t = (Z_{t,i})_{i=1}^{n}$ and regress it on $\boldsymbol{X}$.

The consistency of these meta-learners is achieved if the nuisance parameters are well-specified.

The consistency of pseudo-outcome meta-learners

One key element to prove the consistency of pseudo-outcome meta-learners is the assumption of Unconfoundedness.

Indeed,

The consistency of pseudo-outcome meta-learners

One key element to prove the consistency of pseudo-outcome meta-learners is the assumption of Unconfoundedness.

Indeed,

$$\mathbb{E}(\mathbf{1}\{T = t\} Y_{\mathrm{obs}} \mid \boldsymbol{X} = \boldsymbol{x}) = \mathbb{E}(\mathbf{1}\{T = t\} Y(t) \mid \boldsymbol{X} = \boldsymbol{x})$$

The consistency of pseudo-outcome meta-learners

One key element to prove the consistency of pseudo-outcome meta-learners is the assumption of Unconfoundedness.

Indeed,

$$\begin{aligned}\mathbb{E}(\mathbf{1}\{T = t\}Y_{\text{obs}} \mid \boldsymbol{X} = \boldsymbol{x}) &= \mathbb{E}(\mathbf{1}\{T = t\}Y(t) \mid \boldsymbol{X} = \boldsymbol{x}) \\ &= \mathbb{E}(\mathbf{1}\{T = t\} \mid \boldsymbol{X} = \boldsymbol{x})\mathbb{E}(\mathbf{1}\{T = t\} \mid \boldsymbol{X} = \boldsymbol{x})\end{aligned}$$

The consistency of pseudo-outcome meta-learners

One key element to prove the consistency of pseudo-outcome meta-learners is the assumption of Unconfoundedness.

Indeed,

$$\begin{aligned}\mathbb{E}(\mathbf{1}\{T = t\} Y_{\text{obs}} \mid \boldsymbol{X} = \boldsymbol{x}) &= \mathbb{E}(\mathbf{1}\{T = t\} Y(t) \mid \boldsymbol{X} = \boldsymbol{x}) \\ &= \mathbb{E}(\mathbf{1}\{T = t\} \mid \boldsymbol{X} = \boldsymbol{x}) \mathbb{E}(\mathbf{1}\{T = t\} \mid \boldsymbol{X} = \boldsymbol{x}) \\ &= r(t, \boldsymbol{x}) \mu_t(\boldsymbol{x})\end{aligned}$$

and so on..

The consistency of pseudo-outcome meta-learners

One key element to prove the consistency of pseudo-outcome meta-learners is the assumption of Unconfoundedness.

Indeed,

$$\begin{aligned}\mathbb{E}(\mathbf{1}\{T=t\}Y_{\mathrm{obs}} \mid \boldsymbol{X}=\boldsymbol{x}) &= \mathbb{E}(\mathbf{1}\{T=t\}Y(t) \mid \boldsymbol{X}=\boldsymbol{x}) \\ &= \mathbb{E}(\mathbf{1}\{T=t\} \mid \boldsymbol{X}=\boldsymbol{x})\mathbb{E}(\mathbf{1}\{T=t\} \mid \boldsymbol{X}=\boldsymbol{x}) \\ &= r(t,\boldsymbol{x})\mu_t(\boldsymbol{x})\end{aligned}$$

and so on..

All you need to have is $\widehat{r}=r$ and/or $\widehat{\mu_t}=\mu_t$ to obtain $\mathbb{E}(Z_t \mid \boldsymbol{X}=\boldsymbol{x})$.

The bias-variance analysis of pseudo-outcome meta-learner

Assumption A1. We assume that the outcomes $Y(t)$ are generated from a function f such that

$$Y(t) = f(t, \boldsymbol{X}) + \epsilon \text{ with } \epsilon \sim \mathcal{N}(0, \sigma^2)$$

Assumption A2. We assume the existence of $\beta_t^* \in \mathbb{R}^p$ such that $f(t, \boldsymbol{x}) = \sum_{j=0}^{p-1} \beta_{t,j}^* f_j(\boldsymbol{x})$.

Assumption A3. We assume the positivity of the GPS $0 < r_{\min} \leq r(t, \boldsymbol{X})$, and we assume that f and μ_t are bounded i.e. there exists $C > 0$ such that $|\mu_t(\boldsymbol{x})|, |f(t, \boldsymbol{x})| \leq C$ for all $t \in \mathcal{T}$ and $\boldsymbol{x} \in \mathbb{R}^d$.

The bias-variance analysis of pseudo-outcome meta-learner

Consider the pseudo-outcome random Variable Z_t such that

$$Z_t = A_t(T, \boldsymbol{X}) Y_{\mathrm{obs}} + B_t(T, \boldsymbol{X})$$

where $A_t(T, \boldsymbol{X})$ and $B_t(T, \boldsymbol{X})$ are given for each pseudo-outcome meta-learner.

The bias-variance analysis of pseudo-outcome meta-learner

Consider the pseudo-outcome random Variable Z_t such that

$$Z_t = A_t(T, \boldsymbol{X}) Y_{\mathrm{obs}} + B_t(T, \boldsymbol{X})$$

where $A_t(T, \boldsymbol{X})$ and $B_t(T, \boldsymbol{X})$ are given for each pseudo-outcome meta-learner.

The regression coefficient $\widehat{\beta}_t$ are given by the Ordinary Least Squares (OLS) method

$$\widehat{\beta}_t = \left(\mathbf{H}^\top \mathbf{H}\right)^{-1} \mathbf{H}^\top \boldsymbol{z}_t,$$

where $\boldsymbol{z}_t = (Z_{t,i})_{1 \leq i \leq n}$ and $\mathbf{H} = (\mathbf{H}_{ij}) \in \mathbb{R}^{n \times p}$ is the regression matrix.

The bias-variance analysis of pseudo-outcome meta-learner

Theorem

Under Assumptions (A1-A3), the OLS estimator $\widehat{\beta}_t$ *has bias* $\mathbb{B}(\widehat{\beta}_t) = \mathbb{E}(\widehat{\beta}_t - \beta_t^*) = 0$ *if the nuisance parameters are well-specified, and a covariance matrix* $\mathbb{V}(\widehat{\beta}_t) = 1/n\ \mathbf{C}$*, whose terms* $\mathbf{C}_{ij}$ *are bounded by:*

$$|\mathbf{C}_{ij}| \leq \begin{cases} \mathcal{E}^{M} = \mathcal{O}\left(\frac{1}{r_{\min}^{1+\epsilon}}\right) & \text{for the M-learner} \\ \mathcal{E}^{DR} = \mathcal{O}\left(\frac{\mathrm{err}(\widehat{\mu}_t)+\mathrm{err}(\widehat{\mu}_{t_0})}{r_{\min}^{1+\epsilon}}\right) & \text{for the DR-learner} \\ \mathcal{E}^{X} = \mathcal{O}\left(K^2 \sum_{t' \neq t} \mathrm{err}(\widehat{\mu}_{t'})\right) & \text{for the X-learner} \end{cases}$$

Where $\mathrm{err}(\widehat{\mu}_t) = \mathbb{E}\left[f(t, \boldsymbol{X}) - \widehat{\mu}_t(\boldsymbol{X})\right]^2$ *is the estimation error of* $\widehat{\mu}_t$*.*

Sketches of the proof: the general case

Step 1: We write $\widehat{\beta}_t$ as function of β_t^*.

$$\widehat{\beta}_t = \left(\mathbf{H}^\top \mathbf{H}\right)^{-1} \mathbf{H}^\top \boldsymbol{z}_t$$

Sketches of the proof: the general case

Step 1: We write $\widehat{\beta}_t$ as function of β_t^*.

$$
\begin{aligned}
\widehat{\beta}_t &= \left(\mathbf{H}^\top \mathbf{H}\right)^{-1} \mathbf{H}^\top \boldsymbol{z}_t \\
&= \left(\mathbf{H}^\top \mathbf{H}\right)^{-1} \mathbf{H}^\top \left(A_t(T_i, \boldsymbol{X}_i) Y_{\mathrm{obs},i} + B_t(T_i, \boldsymbol{X}_i)\right)_{i=1}^n \\
&= \ldots \text{ replace } Y_{\mathrm{obs}} \text{ by } f(T, X) + \epsilon \ldots \\
&= \ldots \text{ Add and subtract } \tau_t(\boldsymbol{X}) \ldots \\
&= \ldots \text{ Replace } \tau_t(\boldsymbol{X}) \text{ by } \mathbf{H}\beta_t^* \text{ i.e. assumption made on } \tau_t \ldots
\end{aligned}
$$

Sketches of the proof: the general case

Step 1: We write $\widehat{\beta}_t$ as function of β_t^*.

$$
\begin{aligned}
\widehat{\beta}_t &= \left(\mathbf{H}^\top \mathbf{H}\right)^{-1} \mathbf{H}^\top \boldsymbol{z}_t \\
&= \left(\mathbf{H}^\top \mathbf{H}\right)^{-1} \mathbf{H}^\top \left(A_t(T_i, \boldsymbol{X}_i) Y_{\mathrm{obs},i} + B_t(T_i, \boldsymbol{X}_i)\right)_{i=1}^{n} \\
&= \ldots \text{ replace } Y_{\mathrm{obs}} \text{ by } f(T, X) + \epsilon \ldots \\
&= \ldots \text{ Add and subtract } \tau_t(\boldsymbol{X}) \ldots \\
&= \ldots \text{ Replace } \tau_t(\boldsymbol{X}) \text{ by } \mathbf{H}\beta_t^* \text{ i.e. assumption made on } \tau_t \ldots \\
&= \ldots \text{ Gather terms of } \beta_t^* \text{ and residuals } \ldots
\end{aligned}
$$

Sketches of the proof: the general case

Step 1: We write $\widehat{\beta}_t$ as function of β_t^*.

$$\begin{aligned}
\widehat{\beta}_t &= \left(\mathbf{H}^\top \mathbf{H}\right)^{-1} \mathbf{H}^\top \boldsymbol{z}_t \\
&= \left(\mathbf{H}^\top \mathbf{H}\right)^{-1} \mathbf{H}^\top \left(A_t(T_i, \boldsymbol{X}_i) Y_{\mathrm{obs},i} + B_t(T_i, \boldsymbol{X}_i)\right)_{i=1}^{n} \\
&= \dots \text{ replace } Y_{\mathrm{obs}} \text{ by } f(T, X) + \epsilon \dots \\
&= \dots \text{ Add and subtract } \tau_t(\boldsymbol{X}) \dots \\
&= \dots \text{ Replace } \tau_t(\boldsymbol{X}) \text{ by } \mathbf{H}\beta_t^* \text{ i.e. assumption made on } \tau_t \dots \\
&= \dots \text{ Gather terms of } \beta_t^* \text{ and residuals} \dots \\
&= \beta_t^* + \left(\mathbf{H}^\top \mathbf{H}\right)^{-1} \mathbf{H}^\top \tilde{\boldsymbol{\epsilon}}_t
\end{aligned}$$

where $\tilde{\epsilon}_i = \psi_t(T_i, \boldsymbol{X}_i) + A_t(T_i, \boldsymbol{X}_i)\epsilon_i$ and $\psi_t(T_i, \boldsymbol{X}_i) = A_t(T_i, \boldsymbol{X}_i) f(T_i, \boldsymbol{X}_i) - \tau_t(\boldsymbol{X_i}) + B_t(T_i, \boldsymbol{X}_i)$

Sketches of the proof: the general case

Step 1: We write $\widehat{\beta}_t$ as function of β_t^*.

$$
\begin{aligned}
\widehat{\beta}_t &= \left(\mathbf{H}^\top \mathbf{H}\right)^{-1} \mathbf{H}^\top \boldsymbol{z}_t \\
&= \left(\mathbf{H}^\top \mathbf{H}\right)^{-1} \mathbf{H}^\top \left(A_t(T_i, \boldsymbol{X}_i) Y_{\text{obs},i} + B_t(T_i, \boldsymbol{X}_i)\right)_{i=1}^n \\
&= \dots \text{ replace } Y_{\text{obs}} \text{ by } f(T, X) + \epsilon \dots \\
&= \dots \text{ Add and subtract } \tau_t(\boldsymbol{X}) \dots \\
&= \dots \text{ Replace } \tau_t(\boldsymbol{X}) \text{ by } \mathbf{H}\beta_t^* \text{ i.e. assumption made on } \tau_t \dots \\
&= \dots \text{ Gather terms of } \beta_t^* \text{ and residuals } \dots \\
&= \beta_t^* + \left(\mathbf{H}^\top \mathbf{H}\right)^{-1} \mathbf{H}^\top \tilde{\boldsymbol{\epsilon}}_t
\end{aligned}
$$

where $\tilde{\epsilon}_i = \psi_t(T_i, \boldsymbol{X}_i) + A_t(T_i, \boldsymbol{X}_i)\epsilon_i$ and $\psi_t(T_i, \boldsymbol{X}_i) = A_t(T_i, \boldsymbol{X}_i) f(T_i, \boldsymbol{X}_i) - \tau_t(\boldsymbol{X}_{\boldsymbol{i}}) + B_t(T_i, \boldsymbol{X}_i)$

Here, $\mathbb{E}(\tilde{\epsilon}) = \mathbb{E}(\psi_t(T, \boldsymbol{X})) = 0$ if the nuisance parameters in A_t and B_t are well-specified.

The bias-variance analysis of pseudo-outcome meta-learner

Step 2: We consider the random variables $\boldsymbol{Z}_t^{(n)}$ of mean $\boldsymbol{m}$ and covariance $\mathbf{C}_z$ such that

$$\boldsymbol{Z}_t^{(n)} = \Big(\frac{1}{n}(\mathbf{H}^\top\tilde{\epsilon})_1, \dots, \frac{1}{n}(\mathbf{H}^\top\tilde{\epsilon})_p, \frac{1}{n}(\mathbf{H}^\top\mathbf{H})_{11}, \dots, \frac{1}{n}(\mathbf{H}^\top\mathbf{H})_{pp}\Big)^\top \in \mathbb{R}^{p+p^2}$$

We write the residual term as function of β_t^* and $\boldsymbol{Z}_t^{(n)}$:

$$\begin{aligned}\widehat{\beta}_t &= \beta_t^* + \left(\mathbf{H}^\top\mathbf{H}\right)^{-1}\mathbf{H}^\top\tilde{\epsilon}_t = \beta_t^* + \left(\frac{1}{n}\mathbf{H}^\top\mathbf{H}\right)^{-1}\left(\frac{1}{n}\mathbf{H}^\top\tilde{\epsilon}_t\right)\\ &= \beta_t^* + \phi(\boldsymbol{Z}_t^{(n)}) = \beta_t^* + \Phi(\boldsymbol{S}^{(n)}, \boldsymbol{m})\end{aligned}$$

where $\Phi : \mathbb{R}^{p+p^2} \times \mathbb{R}^{p+p^2} \to \mathbb{R}^p$ and $\phi : \mathbb{R}^{p+p^2} \to \mathbb{R}^p$ are $\mathcal{C}^1$-functions and $\boldsymbol{S}^{(n)} = \sqrt{n}\big(\boldsymbol{Z}_t^{(n)} - \boldsymbol{m}\big)$.

The bias-variance analysis of pseudo-outcome meta-learner

Step 3: We apply on $\boldsymbol{S}^{(n)}$ the multivariate Central Limit Theorem (CLT):

$$\sqrt{n}\big(\boldsymbol{S}^{(n)} - \mathbf{0}\big) \xrightarrow{\mathcal{L}} \mathcal{N}(\mathbf{0}, \mathbf{C}_z)$$

The bias-variance analysis of pseudo-outcome meta-learner

Step 3: We apply on $\boldsymbol{S}^{(n)}$ the multivariate Central Limit Theorem (CLT):

$$\sqrt{n}\big(\boldsymbol{S}^{(n)} - \mathbf{0}\big) \xrightarrow{\mathcal{L}} \mathcal{N}(\mathbf{0}, \mathbf{C}_z)$$

and the Delta method

$$\sqrt{n}\Big[\Phi(S^{(n)}, \boldsymbol{m}) - \Phi(\mathbf{0}, \boldsymbol{m})\Big] \xrightarrow{\mathcal{L}} \mathcal{N}\left(\mathbf{0}, J_{\Phi}^{(1)}(\mathbf{0}, \boldsymbol{m})^{\top} \mathbf{C}_z J_{\Phi}^{(1)}(\mathbf{0}, \boldsymbol{m})\right),$$

The bias-variance analysis of pseudo-outcome meta-learner

Step 3: We apply on $\boldsymbol{S}^{(n)}$ the multivariate Central Limit Theorem (CLT):

$$\sqrt{n}\big(\boldsymbol{S}^{(n)} - \mathbf{0}\big) \xrightarrow{\mathcal{L}} \mathcal{N}(\mathbf{0}, \mathbf{C}_z)$$

and the Delta method

$$\sqrt{n}\Big[\Phi(S^{(n)}, \boldsymbol{m}) - \Phi(\mathbf{0}, \boldsymbol{m})\Big] \xrightarrow{\mathcal{L}} \mathcal{N}\left(\mathbf{0}, J_{\Phi}^{(1)}(\mathbf{0}, \boldsymbol{m})^{\top} \mathbf{C}_z J_{\Phi}^{(1)}(\mathbf{0}, \boldsymbol{m})\right),$$

and we get

$$\widehat{\beta}_t = \beta_t^* + \Phi(\boldsymbol{S}_n, \boldsymbol{m}) \approx \beta_t^* + \Phi(\mathbf{0}, \boldsymbol{m}) + \boldsymbol{g}_n/\sqrt{n}.$$

where $\boldsymbol{g}_n$, a Gaussian noise with covariance matrix of $J_{\Phi}^{(1)}(\mathbf{0}, \boldsymbol{m})^{\top} \mathbf{C} J_{\Phi}^{(1)}(\mathbf{0}, \boldsymbol{m})$.

The bias-variance analysis of pseudo-outcome meta-learner

Step 4: We get the expression of the bias and variance of $\widehat{\beta}_t$

For n big enough :

$$\mathbb{E}(\widehat{\beta}_t) = \beta_t^* + \Phi(\mathbf{0}, \boldsymbol{m}).$$

and,

$$\mathbb{V}(\widehat{\beta}_t) \approx \frac{1}{n} \, J_{\Phi}^{(1)}(\mathbf{0}, \boldsymbol{m})^\top \mathbf{C} J_{\Phi}^{(1)}(\mathbf{0}, \boldsymbol{m}).$$

The bias-variance analysis of pseudo-outcome meta-learner

Step 4: We get the expression of the bias and variance of $\widehat{\beta}_t$

For n big enough :

$$\mathbb{E}(\widehat{\beta}_t) = \beta_t^* + \Phi(\mathbf{0}, \boldsymbol{m}).$$

and,

$$\mathbb{V}(\widehat{\beta}_t) \approx \frac{1}{n}\, J_\Phi^{(1)}(\mathbf{0}, \boldsymbol{m})^\top \mathbf{C} J_\Phi^{(1)}(\mathbf{0}, \boldsymbol{m}).$$

Here, $\mathbb{B}(\widehat{\beta}_t) = \mathbb{E}(\widehat{\beta}_t) - \beta_t^*) = \Phi(\mathbf{0}, \boldsymbol{m})$ in the general case, and $\mathbb{B}(\widehat{\beta}_t) = 0$ in the specific case of well-specified nuisance parameters.

The bias-variance analysis of pseudo-outcome meta-learner

The specific case: Assume that nuisance parameters in A_t and B_t are well-specified.

The bias-variance analysis of pseudo-outcome meta-learner

The specific case: Assume that nuisance parameters in A_t and B_t are well-specified.

By Slutsky's theorem:

$$\begin{aligned}\sqrt{n}\big(\widehat{\beta}_t - \beta_t^*\big) &= n\big(\mathbf{H}^\top\mathbf{H}\big)^{-1} \cdot 1/\sqrt{n}\ \mathbf{H}^\top\tilde{\epsilon} \\ &\xrightarrow{\mathcal{L}} \mathcal{N}(\mathbf{0}, \mathbf{F}^{-1}\mathbf{\Sigma}\mathbf{F}^{-1})\end{aligned}$$

where $\mathbf{F} = \lim_{n\to+\infty} 1/n\ \big(\mathbf{H}^\top\mathbf{H}\big)$ and $\mathbf{\Sigma}$ is a covariance matrix with entries

$$\mathbf{\Sigma}_{ij} = \mathbb{E}\big[f_i(\boldsymbol{X})f_j(\boldsymbol{X})\psi_t^2(T, \boldsymbol{X})\big) + \sigma^2\mathbb{E}\big(f_j(\boldsymbol{X})f_{j'}(\boldsymbol{X})A_t^2(T, \boldsymbol{X})\big].$$

The bias-variance analysis of pseudo-outcome meta-learner

Thus

$$\mathbb{B}(\widehat{\beta}_t) = \mathbb{E}(\widehat{\beta}_t - \beta_t^*) = 0,$$
$$\mathbb{V}(\widehat{\beta}_t) \approx \frac{1}{n}\, \mathbf{F}^{-1}\boldsymbol{\Sigma}\mathbf{F}^{-1}.$$

The bias-variance analysis of pseudo-outcome meta-learner

Thus

$$\mathbb{B}(\widehat{\beta}_t) = \mathbb{E}(\widehat{\beta}_t - \beta_t^*) = 0,$$
$$\mathbb{V}(\widehat{\beta}_t) \approx \frac{1}{n}\,\mathbf{F}^{-1}\mathbf{\Sigma}\mathbf{F}^{-1}.$$

Comparing the errors bounds of each meta-learner is equivalent to compare the terms $|\mathbf{\Sigma}_{ij}|$

The bias-variance analysis of pseudo-outcome meta-learner

Thus

$$\mathbb{B}(\widehat{\beta}_t) = \mathbb{E}(\widehat{\beta}_t - \beta_t^*) = 0,$$
$$\mathbb{V}(\widehat{\beta}_t) \approx \frac{1}{n}\, \mathbf{F}^{-1}\boldsymbol{\Sigma}\mathbf{F}^{-1}.$$

Comparing the errors bounds of each meta-learner is equivalent to compare the terms $|\boldsymbol{\Sigma}_{ij}|$

Here, after some *long* calculations + Minkowski + Holder (see Appendix B in the paper).

$$|\boldsymbol{\Sigma}_{ij}| \leq \begin{cases} \mathcal{E}^M = \mathcal{O}\left(\frac{1}{r_{\min}^{1+\epsilon}}\right) & \text{for the M-learner} \\ \mathcal{E}^{DR} = \mathcal{O}\left(\frac{\mathrm{err}(\widehat{\mu}_t)+\mathrm{err}(\widehat{\mu}_{t_0})}{r_{\min}^{1+\epsilon}}\right) & \text{for the DR-learner} \\ \mathcal{E}^X = \mathcal{O}\left(K^2 \sum_{t' \neq t} \mathrm{err}(\widehat{\mu}_{t'})\right) & \text{for the X-learner} \end{cases}$$

Description of the semi-synthetic dataset

A semi-synthetic dataset simulating the heat extraction performance Q_{well} delivered by a multistage Enhanced Geothermal System (EGS) following the physical model:

$$Q_{well} = Q_{fracture} \times \ell_L / d \times \eta_d .$$

Description of the semi-synthetic dataset

A semi-synthetic dataset simulating the heat extraction performance Q_{well} delivered by a multistage Enhanced Geothermal System (EGS) following the physical model:

$$Q_{well} = Q_{fracture} \times \ell_L / d \times \eta_d .$$

where

Description of the semi-synthetic dataset

A semi-synthetic dataset simulating the heat extraction performance Q_{well} delivered by a multistage Enhanced Geothermal System (EGS) following the physical model:

$$Q_{well} = Q_{fracture} \times \ell_L / d \times \eta_d .$$

where

- $Q_{fracture}$ is the *unknown* heat extraction performance from a single fracture.

Description of the semi-synthetic dataset

A semi-synthetic dataset simulating the heat extraction performance Q_{well} delivered by a multistage Enhanced Geothermal System (EGS) following the physical model:

$$Q_{well} = Q_{fracture} \times \ell_L / d \times \eta_d .$$

where

- $Q_{fracture}$ is the *unknown* heat extraction performance from a single fracture.
- $\ell_L \in [2000, 14000]$ is the lateral length of the well.

Description of the semi-synthetic dataset

A semi-synthetic dataset simulating the heat extraction performance Q_{well} delivered by a multistage Enhanced Geothermal System (EGS) following the physical model:

$$Q_{well} = Q_{fracture} \times \ell_L / d \times \eta_d .$$

where

- $Q_{fracture}$ is the *unknown* heat extraction performance from a single fracture.
- $\ell_L \in [2000, 14000]$ is the lateral length of the well.
- $d \in [100, 500]$ is the average spacing between two fractures.

Description of the semi-synthetic dataset

A semi-synthetic dataset simulating the heat extraction performance Q_{well} delivered by a multistage Enhanced Geothermal System (EGS) following the physical model:

$$Q_{well} = Q_{fracture} \times \ell_L / d \times \eta_d .$$

where

- $Q_{fracture}$ is the *unknown* heat extraction performance from a single fracture.
- $\ell_L \in [2000, 14000]$ is the lateral length of the well.
- $d \in [100, 500]$ is the average spacing between two fractures.
- η_d, *known* function of d, is the stage efficiency penalizing the individual contribution when fractures are close to each other.

Description of the semi-synthetic dataset ii

$Q_{fracture}$ is *simulated* (with a numerical emulator) using fracture's length, height, width and permeability (fracture design), reservoir's porosity, permeability and pore pressure (reservoirs characteristics).

Description of the semi-synthetic dataset ii

$Q_{fracture}$ is *simulated* (with a numerical emulator) using fracture's length, height, width and permeability (fracture design), reservoir's porosity, permeability and pore pressure (reservoirs characteristics).

A full factorial DoE dataset of $n = \underbrace{10 \times 10 \times 2 \times 3}_{design} \times \underbrace{3 \times 3 \times 3}_{reservoir} = 16200$ observations covering all possible scenarios of a fracture in a reservoir is created.

Description of the semi-synthetic dataset ii

$Q_{fracture}$ is *simulated* (with a numerical emulator) using fracture's length, height, width and permeability (fracture design), reservoir's porosity, permeability and pore pressure (reservoirs characteristics).

A full factorial DoE dataset of $n = \underbrace{10 \times 10 \times 2 \times 3}_{design} \times \underbrace{3 \times 3 \times 3}_{reservoir} = 16200$ observations covering all possible scenarios of a fracture in a reservoir is created.

The final dataset containing Q_{well} is obtained after defining *your own* well characteristics (lateral lengths ℓ_L and fracture spacing d).

Creation of biased dataset

Purpose: Emulate observational data found in real-world situations.

Creation of biased dataset

Purpose: Emulate observational data found in real-world situations.

How: - *Preferential selection* strategy - selecting preferentially only observations with certain characteristics

Creation of biased dataset

Purpose: Emulate observational data found in real-world situations.

How: - *Preferential selection* strategy - selecting preferentially only observations with certain characteristics

Example: Geothermal wells with larger lateral lengths are likely to have more fractures (expensive wells are located in better geological areas).

Creation of biased dataset

Purpose: Emulate observational data found in real-world situations.

How: - *Preferential selection* strategy - selecting preferentially only observations with certain characteristics

Example: Geothermal wells with larger lateral lengths are likely to have more fractures (expensive wells are located in better geological areas).

Consequence: Low (under-estimated) heat performance for small wells and high (over-estimated) heat performance for large wells.

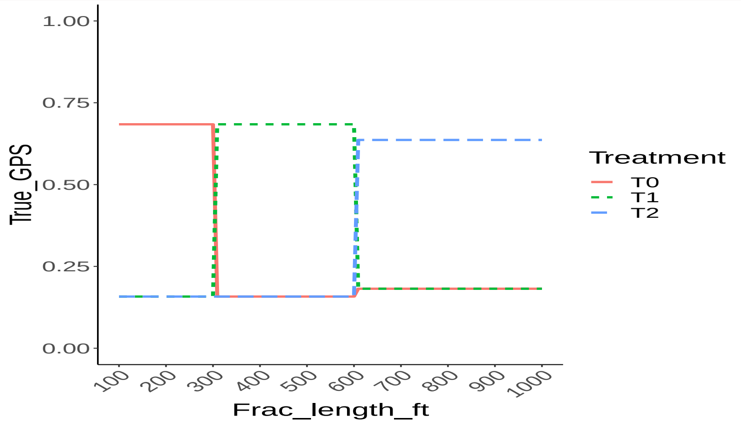
Example of preferential selection

Consider three-level treatments $T \in \{0, 1, 2\}$ (e.g. lateral length) and **discrete** covariate X is uniformly distributed $X \sim \mathcal{U}(100, 1000)$ (e.g. fracture length).

$$\text{In } \mathbf{D}_0,\ \ T_i = 0 \text{ and the } X_i \text{ are i.i.d uniformly distributed over } [100, 300] = I_0.$$
$$\text{In } \mathbf{D}_1,\ \ T_i = 1 \text{ and the } X_i \text{ are i.i.d uniformly distributed over } (300, 600] = I_1.$$
$$\text{In } \mathbf{D}_2,\ \ T_i = 2 \text{ and the } X_i \text{ are i.i.d uniformly distributed over } (600, 1000] = I_2.$$
$$\text{In } \mathbf{D}_3,\ \text{the treatment } T_i \text{ is assigned randomly (RCT setting) to } X_i$$

Example of preferential selection ii

This is a observational setting where T is confounded X (e.g. the larger X is, the more likely we have chance to receive the treatment $T = 2$). The Generalized Propensity Score r satisfies:



What can we do with this dataset?

You can have more fun by manipulating the dataset.

- Introduce more selection bias in the dataset.
- Remove some observations (causal inference with missing data).
- Remove some covariates (causal inference with unobserved confounders)
- Change the distribution of "controlled" covariates (Lateral length and average spacing)
- ... any other suggestion?

Availability: The semi-synthetic dataset is available at this link.

It will be available *soon* on my Github (with the code and the biased dataset).

Application on the estimation of multi-valued CATEs i

We consider the lateral length $T = \ell_L$ as treatment, $Y = \log\left(Q_{well}\right)$ as outcome and $\boldsymbol{X}$ are the rest of parameters. We want to estimate CATEs of the lateral length such that

$$\tau_{\ell_L}(\boldsymbol{x}) = \mathbb{E}\left[\log\left(Q_{well}(\ell_L)\right) - \log\left(Q_{well}(\ell_0)\right) \mid \boldsymbol{X} = \boldsymbol{x}\right] = \log(\ell_L) - \log(\ell_0)$$

i.e. the expected improvement of $\log\left(Q_{well}\right)$ *compared to baseline well of* ℓ_0.

Application on the estimation of multi-valued CATEs i

We consider the lateral length $T = \ell_L$ as treatment, $Y = \log(Q_{well})$ as outcome and $\boldsymbol{X}$ are the rest of parameters. We want to estimate CATEs of the lateral length such that

$$\tau_{\ell_L}(\boldsymbol{x}) = \mathbb{E}\left[\log\left(Q_{well}(\ell_L)\right) - \log\left(Q_{well}(\ell_0)\right) \mid \boldsymbol{X} = \boldsymbol{x}\right] = \log(\ell_L) - \log(\ell_0)$$

i.e. the expected improvement of $\log(Q_{well})$ *compared to baseline well of* ℓ_0.

Observational biased dataset. A sample of $n = 10000$ units such that Wells with high lateral length ℓ_L are likely to have larger fractures ℓ_F (and therefore better heat Q_{well}) and vice versa.
Confounder variable: fracture length

Application on the estimation of multi-valued CATEs i

We consider the lateral length $T = \ell_L$ as treatment, $Y = \log(Q_{well})$ as outcome and $\boldsymbol{X}$ are the rest of parameters. We want to estimate CATEs of the lateral length such that

$$\tau_{\ell_L}(\boldsymbol{x}) = \mathbb{E}\left[\log\left(Q_{well}(\ell_L)\right) - \log\left(Q_{well}(\ell_0)\right) \mid \boldsymbol{X} = \boldsymbol{x}\right] = \log(\ell_L) - \log(\ell_0)$$

i.e. the expected improvement of $\log(Q_{well})$ *compared to baseline well of* ℓ_0.

Observational biased dataset. A sample of $n = 10000$ units such that Wells with high lateral length ℓ_L are likely to have larger fractures ℓ_F (and therefore better heat Q_{well}) and vice versa. *Confounder variable: fracture length*

Goal. Know which meta-learners perform better to estimate the true CATEs τ_{ℓ_L}?

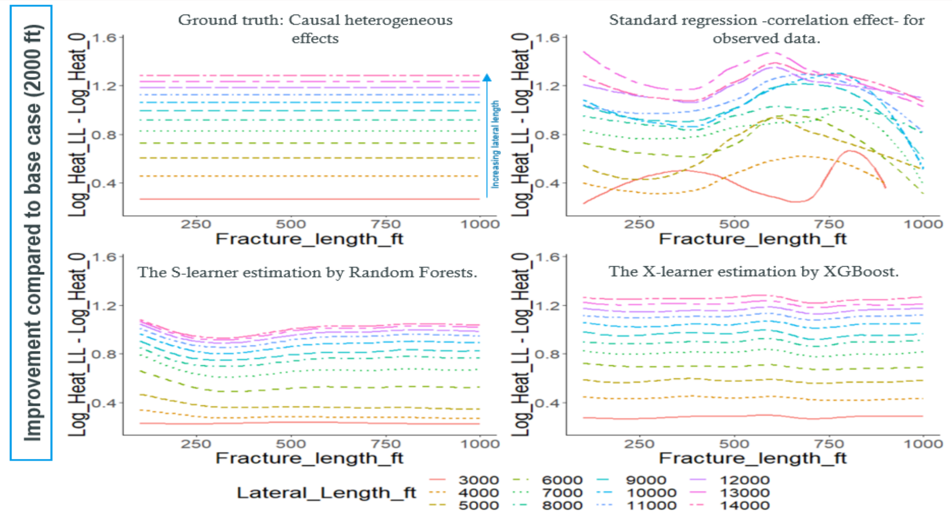
Application on the estimation of multi-valued CATEs i

Here, the GPS satisfies (proof in the paper for the generalized case with K treatment)

$$\mathbb{P}(\text{Lateral\_length} = \ell_L \mid \boldsymbol{X} = \boldsymbol{x}) = \begin{cases} \frac{14}{26} & \text{if frac\_length} \in [h(\ell_L), h(\ell_L) + 1000], \\ \frac{1}{26} & \text{otherwise.} \end{cases}$$

Consequence: low (under-estimated) heat performance for small wells and high (over-estimated) heat performance for large wells.

Application on the estimation of multi-valued CATEs iii



Answers to some questions about our work i

Q2: Do the « Double Machine Learning » and « Doubly robust learning » approaches fall in the same classe of Meta-learners?

A: The DR-learner is inspired from « Doubly Robust learning » approach.

1. You estimate the outcome model μ_t and the propensity score r
2. You build the pseudo-outcome Z_t
3. You regress Z_t on $\boldsymbol{X}$ to estimate CATEs.

But the « Double Machine Learning » is quite a different approach, but similar somehow to the R-learner).

1. You assume a structural equation on the outcome Y and T given $\boldsymbol{X}$ and the CATE τ
2. You estimate the structural components of this structural equation
3. You estimate the CATEs τ by minimizing the residuals errors

Answers to some questions about our work ii

Q3: Random Forest and XGBoost are interpolating models unlike linear models, don't you think that maybe the reason why linear model performs better ?

A: Excellent remark ! We have doubt the problem of *overfitting*, we will try extrapolating models and investigate their results.

Q4: Is the estimation of CATEs an interpolation or extrapolation problem?

A: For $\boldsymbol{X}$ it is an interpolation problem whereas for T it is extrapolation problem. We can describe it as *interpolation problem with missing data*.

Answers to some questions about our work iii

Q6: Can you comment more about the mPEHE metric?

A: The mPEHE is the mean of PEHE over all treatments. It is an extension of the PEHE [Hill, 2011, Shalit et al., 2017, Curth et al., 2021], which is an equivalent to RMSE in the binary case.

Q7: Don't you think that the mPEHE metric is the adapted one?

A1: Well spotted, mPEHE is a combinaison of norms ℓ_2 and norm ℓ_1. It could be more interesting to take $norm\ell_2$ or $norm\ell_1$ over all treatments.

A2: The mPEHE treats all treatment equally, one may think of a weighted metric that penalizes more or less certain treatments.

Answers to some questions about our work iv

Q9: Did you try to run different simulations with the same selection bias?

A: No, our simulations were run on a fixed seed. We will try to run different simulations and inspect the results.

Q9 bis: Did you change the sample size n and see what happens ?

A: Yes, we did in Appendix D5. Increasing the sample size n improves the quality of the meta-learner's estimation (expect for the M-learner)

Answers to some questions about our work v

Q10: What about The conditional independence testing?

A: Unfortunately, no use at this stage. Maybe it can be used to regularized meta-learners?

Q12: The Generalized Propensity Score appears less in the paper, why?

A1: The nature of the problem require the estimation of the CATE at specific sub-groups of units.

A2: Unlike the ATE, conditioning on the covariates $\boldsymbol{X}$ is much stronger than conditioning on the GPS r.

A3: The GPS is use to regularize the T-learner and to define pseudo-outcome variables that target the CATE.

Answers to some questions about our work vi

Q12: At which level you may need the assumption 3.1 of Unconfoundedness?

A: To guaranty the identification of the CATE and the consistency of meta-learners

Q13: Don't you need the Do-calculus in your work?

A: The graph of the Rubin Causal Model is known, no collider, no mediator, only a confounder $\boldsymbol{X}$.

In the context of counterfactual prediction with RCM, intervening on $\boldsymbol{X}$ is equivalent to conditioning on $\boldsymbol{X}$.

$$p(Y(t) \mid do(\boldsymbol{X} = \boldsymbol{x})) = p(Y(t) \mid \boldsymbol{X} = \boldsymbol{x}). \tag{1}$$

Answers to some questions about our work vii

Q15: You considered discrete treatment with $K = 10$, what would be the result if $K \to +\infty$?

A1: On-going work.. but some preliminary results indicate that:

- The performances of the R-learner increase.
- The performances of the T-learner decrease.
- The performances of the X- and S-learners are similar
- Maybe the X-learner is equivalent to S-learner for $K \to +\infty$.

A2: The generalization to continuous treatment would require kernel methods and the estimation of conditional distribution (generalized propensity score).

Answers to some questions about our work viii

Q17: Imagine that we want to be more precise certain treatments and make the errors smaller than other treatments? What should we do and how?

A1: The X-learner may be a solution. It incorporates information from other treatments to predict the CATE at specific level. This claim is to be verified numerically. More numerical experiments are needed.

A2: Maybe we should suggest a weighted metric ?

Answers to some questions about our work ix

Q18: Do you have any reasons why to select a specific model or approach while estimating CATEs ?

A1: The notion of meta-learners does not require specific model (i.e. model-free approach). We had the freedom to use any base-learner for prediction (Neural Network will be also included later).

A2: One of our perspectives is the further investigation of the so-called "Sample Fitting" strategies [Okasa, 2022]: Cross-validation, Train-test split etc.

Q20: Does the treatment change the distribution of $p(Y(t))$?

A: We did not consider these issues on our work.

References

A. Curth, D. Svensson, J. Weatherall, and M. van der Schaar. Really doing great at estimating CATE? a critical look at ML benchmarking practices in treatment effect estimation. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.

J. L. Hill. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240, 2011. doi: 10.1198/jcgs.2010.08162. URL https://doi.org/10.1198/jcgs.2010.08162.

P. W. Holland. Statistics and causal inference. *Journal of the American Statistical Association*, 81(396):945–960, 1986. ISSN 01621459.

S. R. Künzel, J. S. Sekhon, P. J. Bickel, and B. Yu. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences*, 116(10):4156–4165, Feb 2019. ISSN 1091-6490. doi: 10.1073/pnas.1804597116. URL http://dx.doi.org/10.1073/pnas.1804597116.

T. M. Mitchell. *Machine Learning*. McGraw-Hill, New York, 1997. ISBN 978-0-07-042807-2.

G. Okasa. Meta-learners for estimation of causal effects: Finite sample cross-fit performance, 2022.

P. M. Robinson. Root-n-consistent semiparametric regression. *Econometrica*, 56(4):931–954, 1988.

D. Rubin. Estimating causal effects if treatment in randomized and nonrandomized studies. *J. Educ. Psychol.*, 66, 01 1974.

U. Shalit, F. D. Johansson, and D. Sontag. Estimating individual treatment effect: Generalization bounds and algorithms. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 3076–3085. JMLR.org, 2017.