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 2, 2022

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

ÉCOLE POLYTECHNIQUE IP PARIS

TotalEnergies

Introduction

Motivation: Interpretable Machine Learning

In engineering and industry:

The prediction task is statistically achieved (high R^2, low RMSE ...)

Motivation: Interpretable Machine Learning

In engineering and industry:

The prediction task is statistically achieved (high R^2, low RMSE ...)

However

Motivation: Interpretable Machine Learning

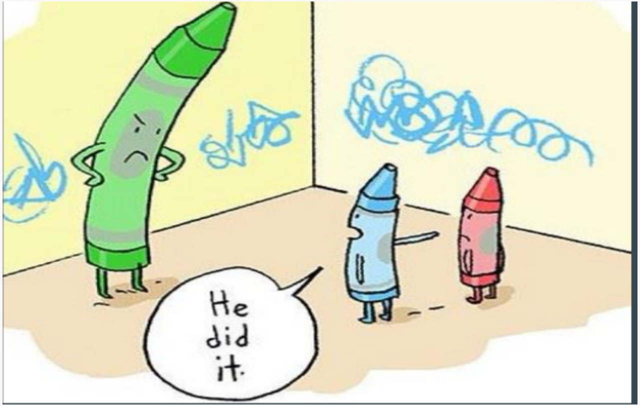
In engineering and industry:

The prediction task is statistically achieved (high R^2, low RMSE ...)

However

There is a discordance between ML predictions and what engineers and specialists expect with their physical models.

Motivation: Interpretable Machine Learning



ML Predictions: *"I saw similar scratches with the red so it's the red pen"*.

Physical models: *"The color is blue so it's the blue pen"*.

Machine Learning vs Causal Inference

Machine Learning: We want to predict what *usually happens* in a given situation.

Machine Learning vs Causal Inference

Machine Learning: We want to predict what *usually happens* in a given situation.

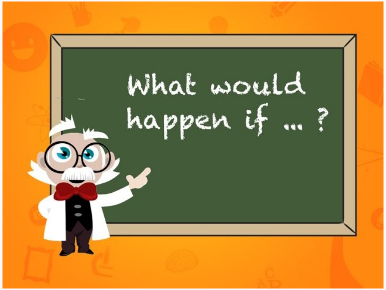
Causal inference: We want to predict what *would happen* if we change the system.

Machine Learning vs Causal Inference

Machine Learning: We want to predict what *usually happens* in a given situation.

Causal inference: We want to predict what *would happen* if we change the system.

- Causal mechanisms are more stable than correlations.
- Inferring causal effects is crucial for development of new strategies and decision making.



Examples of causal inference questions

Medicine: Was it the aspirin that stopped my headache? would I still have had the headache if I did not taken Aspirin [Dawid, 2000]

Economy: How effective are financial incentives for teachers [Imberman, 2015] ?

Sociology: Did busing programs increase the school achievement of disadvantaged minority youth [Morgan and Winship, 2014] ?

Politics: Do polls influence the electoral choice and behavior of voters [Arceneaux et al., 2006]?

Advertising/Marketing: What is the impact of promotions on user retention [Du et al., 2019]?

Potential outcome theory and Rubin Causal model

Definition of causal effect

A practical definition of causality - the counterfactual prediction - [Hernán, 2004]: The variable (treatment) T has an causal effect on the outcome Y *if and only if* changing T leads to a change in Y, while keeping everything else constant.

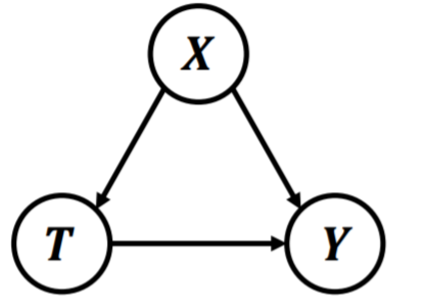
Definition of causal effect

A practical definition of causality - the counterfactual prediction - [Hernán, 2004]: The variable (treatment) T has an causal effect on the outcome Y *if and only if* changing T leads to a change in Y, while keeping everything else constant.

Keeping everything else constant: Strong requirement. All confounders need to be known or observed.

Rubin Causal Model with mutli-treatments

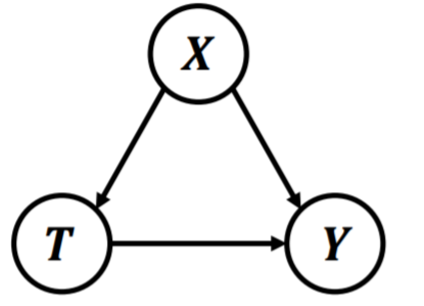
- $i = 1, \ldots, n$: an individual subject to a treatment.



Rubin Causal Model [Rubin, 1974]

Rubin Causal Model with mutli-treatments

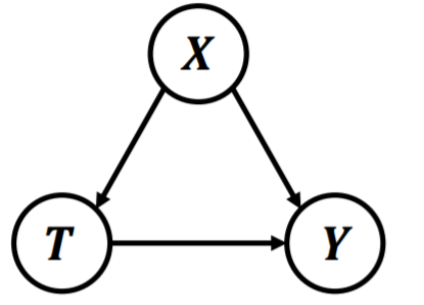
- $i = 1, \ldots, n$: an individual subject to a treatment.
- T: the treatment assignment variable.



Rubin Causal Model [Rubin, 1974]

Rubin Causal Model with mutli-treatments

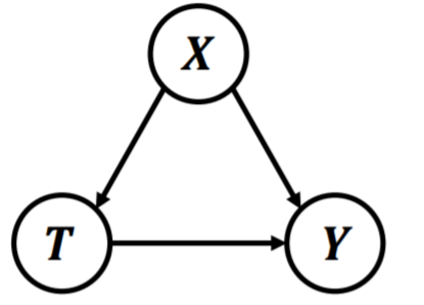
- $i = 1, \dots, n$: an individual subject to a treatment.
- T: the treatment assignment variable.
- $\mathcal{T} = \{t_0, t_1, \dots, t_K\}$: the set of possible treatments.



Rubin Causal Model [Rubin, 1974]

Rubin Causal Model with mutli-treatments

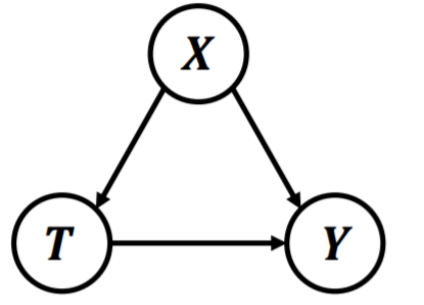
- $i = 1, \ldots, n$: an individual subject to a treatment.
- T: the treatment assignment variable.
- $\mathcal{T} = \{t_0, t_1, \ldots, t_K\}$: the set of possible treatments.
- $\boldsymbol{X} \in \mathbb{R}^d$: vector of d covariates (confounders).



Rubin Causal Model [Rubin, 1974]

Rubin Causal Model with mutli-treatments

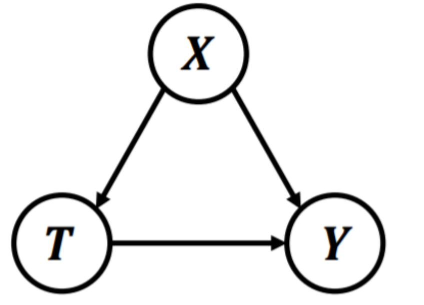
- $i = 1, \dots, n$: an individual subject to a treatment.
- T: the treatment assignment variable.
- $\mathcal{T} = \{t_0, t_1, \dots, t_K\}$: the set of possible treatments.
- $\boldsymbol{X} \in \mathbb{R}^d$: vector of d covariates (confounders).
- $Y_{\mathrm{obs}} = Y(T)$: the observed outcome corresponding to the treatment T.



Rubin Causal Model [Rubin, 1974]

Rubin Causal Model with mutli-treatments

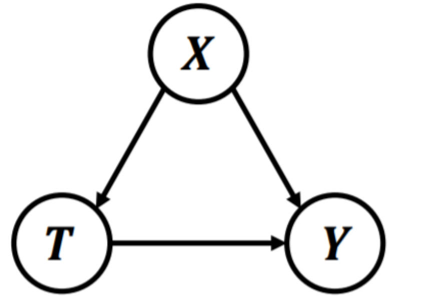
- $i = 1, \dots, n$: an individual subject to a treatment.
- T: the treatment assignment variable.
- $\mathcal{T} = \{t_0, t_1, \dots, t_K\}$: the set of possible treatments.
- $\boldsymbol{X} \in \mathbb{R}^d$: vector of d covariates (confounders).
- $Y_{\text{obs}} = Y(T)$: the observed outcome corresponding 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]

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.
- $\boldsymbol{X} \in \mathbb{R}^d$: vector of d covariates (confounders).
- $Y_{\mathrm{obs}} = Y(T)$: 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.

Intuition behind RCM

Reminder: If two random variables X and Y are correlated, then either X causes Y, or Y causes X, or some other variable Z causes both X and Y.

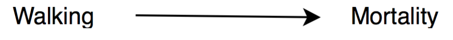
Intuition behind RCM

Reminder: If two random variables X and Y are correlated, then either $\underline{X \text{ causes } Y}$, or Y causes X, or $\underline{\text{some other variable } Z \text{ causes both } X \text{ and } Y}$.

The RCM is useful in local analysis when inferring the "*effects of causes*", not for identifying the "*causes of effects*" (i.e. causal discovery).

Intuition behind RCM: Example

You run a regression in two settings:

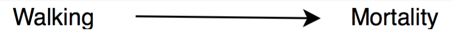


$Mortality = -10.44 \; Walking + 8.583$

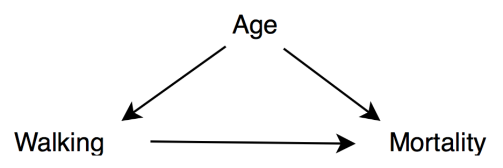
On the left: the number of steps is associated with lower mortality => Relevant as association.

Intuition behind RCM: Example

You run a regression in two settings:



$Mortality = -10.44 \ Walking + 8.583$



$Mortality = -2.401 \ Walking + 6.228 \ Age + 7.989.$

On the left: the number of steps is associated with lower mortality => Relevant as association.

On the right: The effect of walking is lower than what you expect. => *Age* Causes *Mortality* more than *Walking*.

Assumptions of RCM

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

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

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. That is,

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

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. That is,

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

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

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

Estimating Causal Effects

Ingredients: The causal assumptions of consistency, unconfoundedness and positivity.

Estimating Causal Effects

Ingredients: The causal assumptions of consistency, unconfoundedness and positivity.

Result: The counterfactual response is the conditional expectation given T and $\boldsymbol{X}$.

$$\mathbb{E}[Y(t) \mid \boldsymbol{X} = \boldsymbol{x}] = \mathbb{E}[Y_{\text{obs}} \mid T = t, \boldsymbol{X} = \boldsymbol{x}]$$

Proof: Using unconfoundedness, cf. Michèle's talk Monday.

Estimating Causal Effects

Ingredients: The causal assumptions of consistency, unconfoundedness and positivity.

Result: The counterfactual response is the conditional expectation given T and $\boldsymbol{X}$.

$$\mathbb{E}[Y(t) \mid \boldsymbol{X} = \boldsymbol{x}] = \mathbb{E}[Y_{\mathrm{obs}} \mid T = t, \boldsymbol{X} = \boldsymbol{x}]$$

Proof: Using unconfoundedness, cf. Michèle's talk Monday.

Remark: The conditional expectation $\mathbb{E}(Y_{\mathrm{obs}} \mid T = t, X_j = x_j) \neq \mathbb{E}(Y(t) | X_j = x_j)$ does not have causal interpretation since the unconfoundedness assumption can not be satisfied for X_j.

Be aware of variable selection and dimensionality reduction.

Estimating Causal Effects

This causal effect is defined as the Average Treatment Effect (ATE):

$$\mu(t) = \mathbb{E}[Y(t) - Y(t_0)] \quad \text{where } t \in \mathcal{T}/\{t_0\}.$$

Estimating Causal Effects

This causal effect is defined as the Average Treatment Effect (ATE):

$$\mu(t) = \mathbb{E}[Y(t) - Y(t_0)] \quad \text{where } t \in \mathcal{T}/\{t_0\}.$$

E.g. In the binary setting $\mathcal{T} = \{0, 1\}$

$$\tau = \mathbb{E}[Y(1) - Y(0)].$$

Estimating Causal Effects

This causal effect is defined as the Average Treatment Effect (ATE):

$$\mu(t) = \mathbb{E}[Y(t) - Y(t_0)] \quad \text{where } t \in \mathcal{T}/\{t_0\}.$$

E.g. In the binary setting $\mathcal{T} = \{0, 1\}$

$$\tau = \mathbb{E}[Y(1) - Y(0)].$$

The direct estimation of the ATE from observed data be biased due to *selection bias*. (*e.g. Sampling individuals with high walking rate means sampling indirectly younger individuals.*)

Why Heterogeneous Treatment Effects? i

Problem: A treatment may affect the individuals differently.

Purpose: Conduct group-level comparisons to personalize treatment for some units.

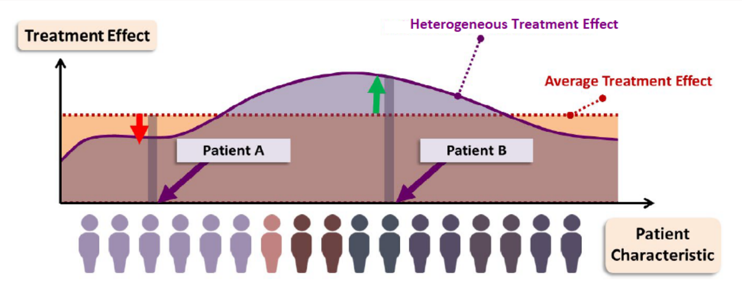


Illustration of the Average Treatment Effect vs Heterogeneous Treatment Effect

Heterogeneous Treatment Effects ii

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

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

Heterogeneous Treatment Effects ii

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

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

E.g. In the binary setting $\mathcal{T} = \{0, 1\}$

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

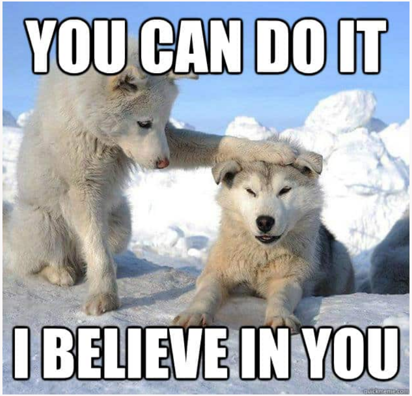
Heterogeneous Treatment Effects ii

From a statistical point of view: The inference of the CATE can be seen as a nonparametric regression problem.

Heterogeneous Treatment Effects ii

From a statistical point of view: The inference of the CATE can be seen as a nonparametric regression problem.

Me: "seems to be an easy Task, I can handle it with any supervised ML algorithm"



Machine Learning i

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. - Tom Mitchell, 1997

Machine Learning ii

Experience E = Supervised Learning i.e. Regression of $Y(t) - Y(t_0)$ on the covariates $\boldsymbol{X}$.

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

Performance Measure P = Accuracy, Precision, RMSE etc.

Machine Learning ii

Machine Learning iii

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 known as the fundamental problem of causal inference [Holland, 1986].

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 multi-treatment Heterogeneous Effects

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

Direct plug-in meta-learners: Naive estimators that estimate the CATE directly by a plug-in difference (**T- and S-learners**).

Indeed,

Direct plug-in meta-learners

Direct plug-in meta-learners: Naive estimators that estimate the CATE directly by a plug-in difference (**T- and S-learners**).

Indeed,

$$\begin{aligned}\tau_t(\boldsymbol{x}) &= \mathbb{E}[Y(t) \mid \boldsymbol{X} = \boldsymbol{x}] - \mathbb{E}[Y(t_0) \mid \boldsymbol{X} = \boldsymbol{x}] \\ &= \mathbb{E}[Y_{\text{obs}} \mid T = t, \boldsymbol{X} = \boldsymbol{x}] - \mathbb{E}[Y_{\text{obs}} \mid T = t_0, \boldsymbol{X} = \boldsymbol{x}] \\ &\qquad \text{(Identification of the counterfactual response using unconfoundedness)}\end{aligned}$$

Direct plug-in meta-learners: T-learner

The **T-learner** (T stands for *two*) is similar to the binary case.

Direct plug-in meta-learners: T-learner

The **T-learner** (T stands for *two*) is similar to the binary case.

- Consider two models models μ_t and μ_{t_0}, where $\mu_w(\boldsymbol{x}) = \mathbb{E}(Y(w)|\boldsymbol{X} = \boldsymbol{x})$ for $w \in \{t, t_0\}$

Direct plug-in meta-learners: T-learner

The **T-learner** (T stands for *two*) is similar to the binary case.

- 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 $\mathbf{S}_t = \{i, T_i = t\}$.

Direct plug-in meta-learners: T-learner

The **T-learner** (T stands for *two*) is similar to the binary case.

- 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 $\hat{\mu}_{t_0}$.

Direct plug-in meta-learners: T-learner

The **T-learner** (T stands for *two*) is similar to the binary case.

- 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: T-learner

The **T-learner** (T stands for *two*) is similar to the binary case.

- 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 T-learner regresses on X for t fixed.

The T-learner does not account for the interaction between treatment T and the outcome Y and create different models for different treatments.

Direct plug-in meta-learners: T-learner

Challenge 1: You may have a small sample of observations while estimating $\hat{\mu}_t$ and $\hat{\mu}_{t_0}$.

Direct plug-in meta-learners: T-learner

Challenge 1: You may have a small sample of observations while estimating $\hat{\mu}_t$ and $\hat{\mu}_{t_0}$.

Unfortunately, there is no solution, maybe data-augmentation?

Direct plug-in meta-learners: T-learner

Challenge 1: You may have a small sample of observations while estimating $\hat{\mu}_t$ and $\hat{\mu}_{t_0}$.

Unfortunately, there is no solution, maybe data-augmentation?

Challenge 2: The T-learner approach may suffer from selection bias *i.e.* μ_w are estimated with respect to the wrong distribution when sampling $\mathbf{S}_w = \{i, T_i = w\}$.

$$\mathbb{E}_{\boldsymbol{X} \sim \mathbb{P}(\cdot)}\left[(\widehat{\mu}_t(\boldsymbol{X}) - \mu_t(\boldsymbol{X}))^2\right] \neq \mathbb{E}_{\boldsymbol{X} \sim \mathbb{P}(\cdot \mid T=t)}\left[(\widehat{\mu}_t(\boldsymbol{X}) - \mu_t(\boldsymbol{X}))^2\right] \quad (1)$$

Direct plug-in meta-learners: T-learner

Challenge 1: You may have a small sample of observations while estimating $\hat{\mu}_t$ and $\hat{\mu}_{t_0}$.

Unfortunately, there is no solution, maybe data-augmentation?

Challenge 2: The T-learner approach may suffer from selection bias *i.e.* μ_w are estimated with respect to the wrong distribution when sampling $\mathbf{S}_w = \{i, T_i = w\}$.

$$\mathbb{E}_{\boldsymbol{X}\sim\mathbb{P}(\cdot)}\left[(\widehat{\mu}_t(\boldsymbol{X}) - \mu_t(\boldsymbol{X}))^2\right] \neq \mathbb{E}_{\boldsymbol{X}\sim\mathbb{P}(\cdot\,|\,T=t)}\left[(\widehat{\mu}_t(\boldsymbol{X}) - \mu_t(\boldsymbol{X}))^2\right] \tag{1}$$

Solution: μ_w should be estimated by minimizing the expected squared error on the nominal *weighted* distribution $\mathbb{P}(T = t)/r(t, \boldsymbol{X})$.

Direct plug-in meta-learners: T-learner

Proof: Let $p(\boldsymbol{x})$ denotes the PDF of $\boldsymbol{X}$ under $\mathbb{P}(\cdot)$, $p(\boldsymbol{x} \mid T = t)$ the PDF of the conditional law of $\boldsymbol{X}|T$ and $R_t = \int (\widehat{\mu}_t(\boldsymbol{x}) - \mu_t(\boldsymbol{x}))^2 p(\boldsymbol{x} \mid T = t) d\boldsymbol{x}$ and let $r(t, \boldsymbol{X}) = \mathbb{P}(T = t \mid \boldsymbol{X} = \boldsymbol{x})$ the Generalized Propensity Score. We consider the expected squared error of $\widehat{\mu}_t$

Direct plug-in meta-learners: T-learner

Proof: Let $p(\boldsymbol{x})$ denotes the PDF of $\boldsymbol{X}$ under $\mathbb{P}(\cdot)$, $p(\boldsymbol{x} \mid T = t)$ the PDF of the conditional law of $\boldsymbol{X}|T$ and $R_t = \int (\widehat{\mu}_t(\boldsymbol{x}) - \mu_t(\boldsymbol{x}))^2 p(\boldsymbol{x} \mid T = t) d\boldsymbol{x}$ and let $r(t, \boldsymbol{X}) = \mathbb{P}(T = t \mid \boldsymbol{X} = \boldsymbol{x})$ the Generalized Propensity Score. We consider the expected squared error of $\widehat{\mu}_t$

$$\mathbb{E}_{\boldsymbol{X} \sim \mathbb{P}(\cdot)}\left[(\widehat{\mu}_t(\boldsymbol{X}) - \mu_t(\boldsymbol{X}))^2\right] = \int (\widehat{\mu}_t(\boldsymbol{x}) - \mu_t(\boldsymbol{x}))^2 p(\boldsymbol{x}) d\boldsymbol{x}$$

Direct plug-in meta-learners: T-learner

Proof: Let $p(\boldsymbol{x})$ denotes the PDF of $\boldsymbol{X}$ under $\mathbb{P}(\cdot)$, $p(\boldsymbol{x} \mid T = t)$ the PDF of the conditional law of $\boldsymbol{X}|T$ and $R_t = \int (\widehat{\mu}_t(\boldsymbol{x}) - \mu_t(\boldsymbol{x}))^2 p(\boldsymbol{x} \mid T = t) d\boldsymbol{x}$ and let $r(t, \boldsymbol{X}) = \mathbb{P}(T = t \mid \boldsymbol{X} = \boldsymbol{x})$ the Generalized Propensity Score. We consider the expected squared error of $\widehat{\mu}_t$

$$\mathbb{E}_{\boldsymbol{X}\sim\mathbb{P}(\cdot)}\left[(\widehat{\mu}_t(\boldsymbol{X}) - \mu_t(\boldsymbol{X}))^2\right] = \int (\widehat{\mu}_t(\boldsymbol{x}) - \mu_t(\boldsymbol{x}))^2 p(\boldsymbol{x}) d\boldsymbol{x}$$

$$= \mathbb{P}(T = t) \int (\widehat{\mu}_t(\boldsymbol{x}) - \mu_t(\boldsymbol{x}))^2 p(\boldsymbol{x} \mid T = t) d\boldsymbol{x} + \sum_{t' \neq t} \mathbb{P}(T = t') \int (\widehat{\mu}_t(\boldsymbol{x}) - \mu_t(\boldsymbol{x}))^2 p(\boldsymbol{x} \mid T = t') d\boldsymbol{x}$$

Direct plug-in meta-learners: T-learner

Proof: Let $p(\boldsymbol{x})$ denotes the PDF of $\boldsymbol{X}$ under $\mathbb{P}(\cdot)$, $p(\boldsymbol{x} \mid T = t)$ the PDF of the conditional law of $\boldsymbol{X}|T$ and $R_t = \int (\widehat{\mu}_t(\boldsymbol{x}) - \mu_t(\boldsymbol{x}))^2 p(\boldsymbol{x} \mid T = t) d\boldsymbol{x}$ and let $r(t, \boldsymbol{X}) = \mathbb{P}(T = t \mid \boldsymbol{X} = \boldsymbol{x})$ the Generalized Propensity Score. We consider the expected squared error of $\widehat{\mu}_t$

$$\mathbb{E}_{\boldsymbol{X} \sim \mathbb{P}(\cdot)}\left[(\widehat{\mu}_t(\boldsymbol{X}) - \mu_t(\boldsymbol{X}))^2\right] = \int (\widehat{\mu}_t(\boldsymbol{x}) - \mu_t(\boldsymbol{x}))^2 p(\boldsymbol{x}) d\boldsymbol{x}$$

$$= \mathbb{P}(T = t) \int (\widehat{\mu}_t(\boldsymbol{x}) - \mu_t(\boldsymbol{x}))^2 p(\boldsymbol{x} \mid T = t) d\boldsymbol{x} + \sum_{t' \neq t} \mathbb{P}(T = t') \int (\widehat{\mu}_t(\boldsymbol{x}) - \mu_t(\boldsymbol{x}))^2 p(\boldsymbol{x} \mid T = t') d\boldsymbol{x}$$

$$= \mathbb{P}(T = t) R_t + \sum_{t' \neq t} \mathbb{P}(T = t') \int (\widehat{\mu}_t(\boldsymbol{x}) - \mu_t(\boldsymbol{x}))^2 \frac{p(\boldsymbol{x} \mid T = t')}{p(\boldsymbol{x} \mid T = t)} p(\boldsymbol{x} \mid T = t) d\boldsymbol{x}$$

Direct plug-in meta-learners: T-learner

Proof: Let $p(\boldsymbol{x})$ denotes the PDF of $\boldsymbol{X}$ under $\mathbb{P}(\cdot)$, $p(\boldsymbol{x} \mid T = t)$ the PDF of the conditional law of $\boldsymbol{X}|T$ and $R_t = \int (\widehat{\mu}_t(\boldsymbol{x}) - \mu_t(\boldsymbol{x}))^2 p(\boldsymbol{x} \mid T = t) d\boldsymbol{x}$ and let $r(t, \boldsymbol{X}) = \mathbb{P}(T = t \mid \boldsymbol{X} = \boldsymbol{x})$ the Generalized Propensity Score. We consider the expected squared error of $\widehat{\mu}_t$

$$\mathbb{E}_{\boldsymbol{X}\sim\mathbb{P}(\cdot)}\left[(\widehat{\mu}_t(\boldsymbol{X}) - \mu_t(\boldsymbol{X}))^2\right] = \int (\widehat{\mu}_t(\boldsymbol{x}) - \mu_t(\boldsymbol{x}))^2 p(\boldsymbol{x}) d\boldsymbol{x}$$

$$= \mathbb{P}(T = t) \int (\widehat{\mu}_t(\boldsymbol{x}) - \mu_t(\boldsymbol{x}))^2 p(\boldsymbol{x} \mid T = t) d\boldsymbol{x} + \sum_{t' \neq t} \mathbb{P}(T = t') \int (\widehat{\mu}_t(\boldsymbol{x}) - \mu_t(\boldsymbol{x}))^2 p(\boldsymbol{x} \mid T = t') d\boldsymbol{x}$$

$$= \mathbb{P}(T = t) R_t + \sum_{t' \neq t} \mathbb{P}(T = t') \int (\widehat{\mu}_t(\boldsymbol{x}) - \mu_t(\boldsymbol{x}))^2 \frac{p(\boldsymbol{x} \mid T = t')}{p(\boldsymbol{x} \mid T = t)} p(\boldsymbol{x} \mid T = t) d\boldsymbol{x}$$

$$= \mathbb{P}(T = t) R_t + \sum_{t' \neq t} \mathbb{P}(T = t') \int (\widehat{\mu}_t(\boldsymbol{x}) - \mu_t(\boldsymbol{x}))^2 \frac{\frac{\mathbb{P}(T=t' | \boldsymbol{x}) p(\boldsymbol{x})}{\mathbb{P}(T=t')}}{\frac{\mathbb{P}(T=t | \boldsymbol{x}) p(\boldsymbol{x})}{\mathbb{P}(T=t)}} p(\boldsymbol{x} \mid T = t) d\boldsymbol{x} \qquad \text{(Bayes rule)}$$

Direct plug-in meta-learners: T-learner

$$= \mathbb{P}(T = t)R_t + \mathbb{P}(T = t)\sum_{t' \neq t} \int (\widehat{\mu}_t(\boldsymbol{x}) - \mu_t(\boldsymbol{x}))^2 \frac{\mathbb{P}(T = t' \mid \boldsymbol{x})}{\mathbb{P}(T = t \mid \boldsymbol{x})} p(\boldsymbol{x} \mid T = t) d\boldsymbol{x}$$

Direct plug-in meta-learners: T-learner

$$= \mathbb{P}(T = t)R_t + \mathbb{P}(T = t)\sum_{t' \neq t}\int (\widehat{\mu}_t(\boldsymbol{x}) - \mu_t(\boldsymbol{x}))^2 \frac{\mathbb{P}(T = t' \mid \boldsymbol{x})}{\mathbb{P}(T = t \mid \boldsymbol{x})} p(\boldsymbol{x} \mid T = t) d\boldsymbol{x}$$

$$= \mathbb{P}(T = t)R_t + \mathbb{P}(T = t)\int (\widehat{\mu}_t(\boldsymbol{x}) - \mu_t(\boldsymbol{x}))^2 \frac{\sum_{t' \neq t}\mathbb{P}(T = t' \mid \boldsymbol{x})}{\mathbb{P}(T = t \mid \boldsymbol{x})} p(\boldsymbol{x} \mid T = t) d\boldsymbol{x}$$

Direct plug-in meta-learners: T-learner

$$
\begin{aligned}
&= \mathbb{P}(T=t)R_t + \mathbb{P}(T=t)\sum_{t'\neq t}\int (\widehat{\mu}_t(\boldsymbol{x}) - \mu_t(\boldsymbol{x}))^2 \frac{\mathbb{P}(T=t' \mid \boldsymbol{x})}{\mathbb{P}(T=t \mid \boldsymbol{x})} p(\boldsymbol{x} \mid T=t) d\boldsymbol{x} \\
&= \mathbb{P}(T=t)R_t + \mathbb{P}(T=t)\int (\widehat{\mu}_t(\boldsymbol{x}) - \mu_t(\boldsymbol{x}))^2 \frac{\sum_{t'\neq t}\mathbb{P}(T=t' \mid \boldsymbol{x})}{\mathbb{P}(T=t \mid \boldsymbol{x})} p(\boldsymbol{x} \mid T=t) d\boldsymbol{x} \\
&= \mathbb{P}(T=t)R_t + \mathbb{P}(T=t)\int \frac{1-r(t,\boldsymbol{x})}{r(t,\boldsymbol{x})} (\widehat{\mu}_t(\boldsymbol{x}) - \mu_t(\boldsymbol{x}))^2 p(\boldsymbol{x} \mid T=t) d\boldsymbol{x}
\end{aligned}
$$

Direct plug-in meta-learners: T-learner

$$= \mathbb{P}(T=t)R_t + \mathbb{P}(T=t)\sum_{t'\neq t}\int (\widehat{\mu}_t(\boldsymbol{x}) - \mu_t(\boldsymbol{x}))^2 \frac{\mathbb{P}(T=t' \mid \boldsymbol{x})}{\mathbb{P}(T=t \mid \boldsymbol{x})} p(\boldsymbol{x} \mid T=t)d\boldsymbol{x}$$

$$= \mathbb{P}(T=t)R_t + \mathbb{P}(T=t)\int (\widehat{\mu}_t(\boldsymbol{x}) - \mu_t(\boldsymbol{x}))^2 \frac{\sum_{t'\neq t}\mathbb{P}(T=t' \mid \boldsymbol{x})}{\mathbb{P}(T=t \mid \boldsymbol{x})} p(\boldsymbol{x} \mid T=t)d\boldsymbol{x}$$

$$= \mathbb{P}(T=t)R_t + \mathbb{P}(T=t)\int \frac{1-r(t,\boldsymbol{x})}{r(t,\boldsymbol{x})} (\widehat{\mu}_t(\boldsymbol{x}) - \mu_t(\boldsymbol{x}))^2 p(\boldsymbol{x} \mid T=t)d\boldsymbol{x}$$

$$= \mathbb{P}(T=t)\int \left(1 + \frac{1-r(t,\boldsymbol{x})}{r(t,\boldsymbol{x})}\right) (\widehat{\mu}_t(\boldsymbol{x}) - \mu_t(\boldsymbol{x}))^2 p(\boldsymbol{x} \mid T=t)d\boldsymbol{x}$$

Direct plug-in meta-learners: T-learner

$$
\begin{aligned}
&= \mathbb{P}(T=t)R_t + \mathbb{P}(T=t)\sum_{t'\neq t}\int (\widehat{\mu}_t(\boldsymbol{x}) - \mu_t(\boldsymbol{x}))^2 \frac{\mathbb{P}(T=t' \mid \boldsymbol{x})}{\mathbb{P}(T=t \mid \boldsymbol{x})} p(\boldsymbol{x} \mid T=t) d\boldsymbol{x} \\
&= \mathbb{P}(T=t)R_t + \mathbb{P}(T=t)\int (\widehat{\mu}_t(\boldsymbol{x}) - \mu_t(\boldsymbol{x}))^2 \frac{\sum_{t'\neq t}\mathbb{P}(T=t' \mid \boldsymbol{x})}{\mathbb{P}(T=t \mid \boldsymbol{x})} p(\boldsymbol{x} \mid T=t) d\boldsymbol{x} \\
&= \mathbb{P}(T=t)R_t + \mathbb{P}(T=t)\int \frac{1-r(t,\boldsymbol{x})}{r(t,\boldsymbol{x})} (\widehat{\mu}_t(\boldsymbol{x}) - \mu_t(\boldsymbol{x}))^2 p(\boldsymbol{x} \mid T=t) d\boldsymbol{x} \\
&= \mathbb{P}(T=t)\int \left(1 + \frac{1-r(t,\boldsymbol{x})}{r(t,\boldsymbol{x})}\right) (\widehat{\mu}_t(\boldsymbol{x}) - \mu_t(\boldsymbol{x}))^2 p(\boldsymbol{x} \mid T=t) d\boldsymbol{x} \\
&= \mathbb{E}_{\boldsymbol{X}\sim\mathbb{P}(\cdot\mid T=t)}\left[\frac{\mathbb{P}(T=t)}{r(t,\boldsymbol{X})} (\widehat{\mu}_t(\boldsymbol{X}) - \mu_t(\boldsymbol{X}))^2\right].
\end{aligned}
$$

Direct plug-in meta-learners: S-learner

The **S-learner** (S stands for *single*) is similar to the binary case.

Direct plug-in meta-learners: S-learner

The **S-learner** (S stands for *single*) is similar to the binary case.

- Consider a single model μ such that $\mu(w, \boldsymbol{x}) = \mathbb{E}(Y_{\text{obs}} \mid T = w, \boldsymbol{X} = \boldsymbol{x})$.

Direct plug-in meta-learners: S-learner

The **S-learner** (S stands for *single*) is similar to the binary case.

- Consider a single model μ such that $\mu(w, \boldsymbol{x}) = \mathbb{E}(Y_{\mathrm{obs}} \mid T = w, \boldsymbol{X} = \boldsymbol{x})$.
- Estimate $\hat{\mu}$ by regressing Y_{obs} on both $\boldsymbol{X}$ and T using all observed data.

Direct plug-in meta-learners: S-learner

The **S-learner** (S stands for *single*) is similar to the binary case.

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

Direct plug-in meta-learners: S-learner

The **S-learner** (S stands for *single*) is similar to the binary case.

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

The S-learner regresses on both X and T.

Challenge: You have no idea how the model deals with the confounding between T and $\boldsymbol{X}$.

We are unable, at the moment, to understand why and how to regularize the S-learner (cf. results).

Pseudo-outcome meta-learners

Pseudo-outcome meta-learners: Learners that build a pseudo-outcome random variable Z_t such that $\mathbb{E}(Z_t \mid \boldsymbol{X} = \boldsymbol{x}) = \tau_t(\boldsymbol{x})$ (**M-, DR- and X-learners**).

Pseudo-outcome meta-learners

Pseudo-outcome meta-learners: Learners that build a pseudo-outcome random variable Z_t such that $\mathbb{E}(Z_t \mid \boldsymbol{X} = \boldsymbol{x}) = \tau_t(\boldsymbol{x})$ (**M-, DR- and X-learners**).

The pseudo-outcome approach is a tentative to

1. Learn CATEs on the whole sample (rather than $\mathbf{S}_w$).
2. Mitigate the *selection bias* while learning the outcome Y_{obs}.

Pseudo-outcome meta-learners: M-learner

M-learner: (M- stands for modified) is Inspired from the Inverse Propensity Weighting (IPW) [Horvitz and Thompson, 1952].

Pseudo-outcome meta-learners: M-learner

M-learner: (M- stands for modified) is Inspired from the Inverse Propensity Weighting (IPW) [Horvitz and Thompson, 1952].

Consider the pseudo-outcome Z_t^M such that

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

where $\widehat{r}$ is an estimator of the GPS $r(t, \boldsymbol{x}) = \mathbb{P}(T = t \mid \boldsymbol{X} = \boldsymbol{x})$.

Pseudo-outcome meta-learners: M-learner

M-learner: (M- stands for modified) is Inspired from the Inverse Propensity Weighting (IPW) [Horvitz and Thompson, 1952].

Consider the pseudo-outcome Z_t^M such that

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

where $\widehat{r}$ is an estimator of the GPS $r(t, \boldsymbol{x}) = \mathbb{P}(T = t \mid \boldsymbol{X} = \boldsymbol{x})$.

Inconvenient: suffers from high variance because of $\widehat{r}$ in the denominator.

Pseudo-outcome meta-learners: M-learner

Proof: Relies mainly on the fact that $\mathbf{1}\{T = t\}Y_{\text{obs}} = \mathbf{1}\{T = t\}Y(t)$ and the Unconfoundedness Assumption.

Consider the first term $Y_t^M = \mathbf{1}\{T = t\}/r(t, \boldsymbol{X})Y_{\text{obs}}$

Pseudo-outcome meta-learners: M-learner

Proof: Relies mainly on the fact that $\mathbf{1}\{T = t\}Y_{\mathrm{obs}} = \mathbf{1}\{T = t\}Y(t)$ and the Unconfoundedness Assumption.

Consider the first term $Y_t^M = \mathbf{1}\{T = t\}/r(t, \boldsymbol{X})Y_{\mathrm{obs}}$

$$\mathbb{E}(Y_t^M \mid \boldsymbol{X} = \boldsymbol{x}) = \mathbb{E}\left[\frac{\mathbf{1}\{T = t\}}{r(t, \boldsymbol{X})}Y_{\mathrm{obs}} \mid \boldsymbol{X} = \boldsymbol{x}\right]$$

Pseudo-outcome meta-learners: M-learner

Proof: Relies mainly on the fact that $\mathbf{1}\{T = t\}Y_{\text{obs}} = \mathbf{1}\{T = t\}Y(t)$ and the Unconfoundedness Assumption.

Consider the first term $Y_t^M = \mathbf{1}\{T = t\}/r(t, \boldsymbol{X})Y_{\text{obs}}$

$$\begin{aligned}
\mathbb{E}(Y_t^M \mid \boldsymbol{X} = \boldsymbol{x}) &= \mathbb{E}\left[\frac{\mathbf{1}\{T = t\}}{r(t, \boldsymbol{X})} Y_{\text{obs}} \mid \boldsymbol{X} = \boldsymbol{x}\right] \\
&= \frac{1}{r(t, \boldsymbol{x})} \mathbb{E}\left[\mathbf{1}\{T = t\}Y(t) \mid \boldsymbol{X} = \boldsymbol{x}\right]
\end{aligned}$$

Pseudo-outcome meta-learners: M-learner

Proof: Relies mainly on the fact that $\mathbf{1}\{T = t\} Y_{\mathrm{obs}} = \mathbf{1}\{T = t\} Y(t)$ and the Unconfoundedness Assumption.

Consider the first term $Y_t^M = \mathbf{1}\{T = t\}/r(t, \boldsymbol{X}) Y_{\mathrm{obs}}$

$$\begin{aligned}
\mathbb{E}(Y_t^M \mid \boldsymbol{X} = \boldsymbol{x}) &= \mathbb{E}\left[\frac{\mathbf{1}\{T = t\}}{r(t, \boldsymbol{X})} Y_{\mathrm{obs}} \mid \boldsymbol{X} = \boldsymbol{x}\right] \\
&= \frac{1}{r(t, \boldsymbol{x})} \mathbb{E}\left[\mathbf{1}\{T = t\} Y(t) \mid \boldsymbol{X} = \boldsymbol{x}\right] \\
&= \frac{1}{r(t, \boldsymbol{x})} \mathbb{E}\big[\mathbf{1}\{T = t\} \mid \boldsymbol{X} = \boldsymbol{x}\big] \mathbb{E}\left[Y(t) \mid \boldsymbol{X} = \boldsymbol{x}\right] \qquad \text{(by Unconfoundedness)}
\end{aligned}$$

Pseudo-outcome meta-learners: M-learner

Proof: Relies mainly on the fact that $\mathbf{1}\{T=t\}Y_{\text{obs}} = \mathbf{1}\{T=t\}Y(t)$ and the Unconfoundedness Assumption.

Consider the first term $Y_t^M = \mathbf{1}\{T=t\}/r(t,\boldsymbol{X})Y_{\text{obs}}$

$$\begin{aligned}
\mathbb{E}(Y_t^M \mid \boldsymbol{X}=\boldsymbol{x}) &= \mathbb{E}\left[\frac{\mathbf{1}\{T=t\}}{r(t,\boldsymbol{X})}Y_{\text{obs}} \mid \boldsymbol{X}=\boldsymbol{x}\right] \\
&= \frac{1}{r(t,\boldsymbol{x})}\mathbb{E}\left[\mathbf{1}\{T=t\}Y(t) \mid \boldsymbol{X}=\boldsymbol{x}\right] \\
&= \frac{1}{r(t,\boldsymbol{x})}\mathbb{E}\left[\mathbf{1}\{T=t\} \mid \boldsymbol{X}=\boldsymbol{x}\right]\mathbb{E}\left[Y(t) \mid \boldsymbol{X}=\boldsymbol{x}\right] \qquad \text{(by Unconfoundedness)} \\
&= \frac{1}{r(t,\boldsymbol{x})}r(t,\boldsymbol{x})\mathbb{E}\left[Y(t) \mid \boldsymbol{X}=\boldsymbol{x}\right]
\end{aligned}$$

Pseudo-outcome meta-learners: M-learner

Proof: Relies mainly on the fact that $\mathbf{1}\{T = t\} Y_{\text{obs}} = \mathbf{1}\{T = t\} Y(t)$ and the Unconfoundedness Assumption.

Consider the first term $Y_t^M = \mathbf{1}\{T = t\}/r(t, \boldsymbol{X}) Y_{\text{obs}}$

$$\begin{aligned}
\mathbb{E}(Y_t^M \mid \boldsymbol{X} = \boldsymbol{x}) &= \mathbb{E}\left[\frac{\mathbf{1}\{T = t\}}{r(t, \boldsymbol{X})} Y_{\text{obs}} \mid \boldsymbol{X} = \boldsymbol{x}\right] \\
&= \frac{1}{r(t, \boldsymbol{x})} \mathbb{E}\left[\mathbf{1}\{T = t\} Y(t) \mid \boldsymbol{X} = \boldsymbol{x}\right] \\
&= \frac{1}{r(t, \boldsymbol{x})} \mathbb{E}\left[\mathbf{1}\{T = t\} \mid \boldsymbol{X} = \boldsymbol{x}\right] \mathbb{E}\left[Y(t) \mid \boldsymbol{X} = \boldsymbol{x}\right] \qquad \text{(by Unconfoundedness)} \\
&= \frac{1}{r(t, \boldsymbol{x})} r(t, \boldsymbol{x}) \mathbb{E}\left[Y(t) \mid \boldsymbol{X} = \boldsymbol{x}\right] \\
&= \mathbb{E}(Y(t) \mid \boldsymbol{X} = \boldsymbol{x})
\end{aligned} \tag{2}$$

Pseudo-outcome meta-learners: DR-learner

DR-learner: (DR- stands for Doubly-Robust) is inspired from Augmented Inverse Probability Weighting [Robins et al., 1994] to overcome the problem of model's misspecification.

Pseudo-outcome meta-learners: DR-learner

DR-learner: (DR- stands for Doubly-Robust) is inspired from Augmented Inverse Probability Weighting [Robins et al., 1994] to overcome the problem of model's misspecification.

Consider the pseudo-outcome Z_t^{DR} such that

$$Z_t^{DR} = \frac{Y_{\mathrm{obs}} - \widehat{\mu}_T(\boldsymbol{X})}{\widehat{r}(t, \boldsymbol{X})}\mathbf{1}\{T = t\} - \frac{Y_{\mathrm{obs}} - \widehat{\mu}_T(\boldsymbol{X})}{\widehat{r}(t_0, \boldsymbol{X})}\mathbf{1}\{T = t_0\} + \widehat{\mu}_t(\boldsymbol{X}) - \widehat{\mu}_{t_0}(\boldsymbol{X})$$

where $(\widehat{\mu}_t)_{t\in\mathcal{T}}$ are estimators of the outcome models $(\mu_t)_{t\in\mathcal{T}}$.

Pseudo-outcome meta-learners: DR-learner

DR-learner: (DR- stands for Doubly-Robust) is inspired from Augmented Inverse Probability Weighting [Robins et al., 1994] to overcome the problem of model's misspecification.

Consider the pseudo-outcome Z_t^{DR} such that

$$Z_t^{DR} = \frac{Y_{\text{obs}} - \widehat{\mu}_T(\boldsymbol{X})}{\widehat{r}(t, \boldsymbol{X})} \mathbf{1}\{T = t\} - \frac{Y_{\text{obs}} - \widehat{\mu}_T(\boldsymbol{X})}{\widehat{r}(t_0, \boldsymbol{X})} \mathbf{1}\{T = t_0\} + \widehat{\mu}_t(\boldsymbol{X}) - \widehat{\mu}_{t_0}(\boldsymbol{X})$$

where $(\widehat{\mu}_t)_{t \in \mathcal{T}}$ are estimators of the outcome models $(\mu_t)_{t \in \mathcal{T}}$.

Proof: Similar to the M-learner.

Pseudo-outcome meta-learners: DR-learner

DR-learner: (DR- stands for Doubly-Robust) is inspired from Augmented Inverse Probability Weighting [Robins et al., 1994] to overcome the problem of model's misspecification.

Consider the pseudo-outcome Z_t^{DR} such that

$$Z_t^{DR} = \frac{Y_{\text{obs}} - \widehat{\mu}_T(\boldsymbol{X})}{\widehat{r}(t, \boldsymbol{X})} \mathbf{1}\{T = t\} - \frac{Y_{\text{obs}} - \widehat{\mu}_T(\boldsymbol{X})}{\widehat{r}(t_0, \boldsymbol{X})} \mathbf{1}\{T = t_0\} + \widehat{\mu}_t(\boldsymbol{X}) - \widehat{\mu}_{t_0}(\boldsymbol{X})$$

where $(\widehat{\mu}_t)_{t \in \mathcal{T}}$ are estimators of the outcome models $(\mu_t)_{t \in \mathcal{T}}$.

Proof: Similar to the M-learner.

Advantage (Double Robustness): The estimator is consistent if one of the models is correct.

Pseudo-outcome meta-learners: DR-learner

DR-learner: (DR- stands for Doubly-Robust) is inspired from Augmented Inverse Probability Weighting [Robins et al., 1994] to overcome the problem of model's misspecification.

Consider the pseudo-outcome Z_t^{DR} such that

$$Z_t^{DR} = \frac{Y_{\text{obs}} - \widehat{\mu}_T(\boldsymbol{X})}{\widehat{r}(t, \boldsymbol{X})}\mathbf{1}\{T = t\} - \frac{Y_{\text{obs}} - \widehat{\mu}_T(\boldsymbol{X})}{\widehat{r}(t_0, \boldsymbol{X})}\mathbf{1}\{T = t_0\} + \widehat{\mu}_t(\boldsymbol{X}) - \widehat{\mu}_{t_0}(\boldsymbol{X})$$

where $(\widehat{\mu}_t)_{t \in \mathcal{T}}$ are estimators of the outcome models $(\mu_t)_{t \in \mathcal{T}}$.

Proof: Similar to the M-learner.

Advantage (Double Robustness): The estimator is consistent if one of the models is correct.

Inconvenient: May also suffer from high variance because of $\widehat{r}$ in the denominator.

Pseudo-outcome meta-learners: X-learner

X-learner: (X- stands for *cross* procedure estimation) is based on *Regression-Adjustment* formulas.

Pseudo-outcome meta-learners: X-learner

X-learner: (X- stands for *cross* procedure estimation) is based on *Regression-Adjustment* formulas.

Consider the pseudo-outcome Z_t^X such that

$$Z_t^X = \mathbf{1}\{T = t\}(Y_{\mathrm{obs}} - \widehat{\mu}_{t_0}(\boldsymbol{X})) + \sum_{t' \neq t} \mathbf{1}\{T = t'\}(\widehat{\mu}_t(\boldsymbol{X}) - Y_{\mathrm{obs}}) + \sum_{t' \neq t} \mathbf{1}\{T = t'\}(\widehat{\mu}_{t'}(\boldsymbol{X}) - \widehat{\mu}_{t_0}(\boldsymbol{X}))$$

where $(\widehat{\mu}_t)_{t \in \mathcal{T}}$ are estimator of the outcome models $(\mu_t)_{t \in \mathcal{T}}$

Pseudo-outcome meta-learners: X-learner

X-learner: (X- stands for *cross* procedure estimation) is based on *Regression-Adjustment* formulas.

Consider the pseudo-outcome Z_t^X such that

$$Z_t^X = \mathbf{1}\{T = t\}(Y_{\text{obs}} - \widehat{\mu}_{t_0}(\boldsymbol{X})) + \sum_{t' \neq t} \mathbf{1}\{T = t'\}(\widehat{\mu}_t(\boldsymbol{X}) - Y_{\text{obs}}) + \sum_{t' \neq t} \mathbf{1}\{T = t'\}(\widehat{\mu}_{t'}(\boldsymbol{X}) - \widehat{\mu}_{t_0}(\boldsymbol{X}))$$

where $(\widehat{\mu}_t)_{t \in \mathcal{T}}$ are estimator of the outcome models $(\mu_t)_{t \in \mathcal{T}}$

Proof: direct calculations

Pseudo-outcome meta-learners: X-learner

X-learner: (X- stands for *cross* procedure estimation) is based on *Regression-Adjustment* formulas.

Consider the pseudo-outcome Z_t^X such that

$$Z_t^X = \mathbf{1}\{T = t\}(Y_{\mathrm{obs}} - \widehat{\mu}_{t_0}(\boldsymbol{X})) + \sum_{t' \neq t} \mathbf{1}\{T = t'\}(\widehat{\mu}_t(\boldsymbol{X}) - Y_{\mathrm{obs}})$$
$$+ \sum_{t' \neq t} \mathbf{1}\{T = t'\}(\widehat{\mu}_{t'}(\boldsymbol{X}) - \widehat{\mu}_{t_0}(\boldsymbol{X}))$$

where $(\widehat{\mu}_t)_{t \in \mathcal{T}}$ are estimator of the outcome models $(\mu_t)_{t \in \mathcal{T}}$

Proof: direct calculations

Advantage: Has the lowest variance because of the Regression-Adjustment formula.

Pseudo-outcome meta-learners: X-learner

X-learner: (X- stands for *cross* procedure estimation) is based on *Regression-Adjustment* formulas.

Consider the pseudo-outcome Z_t^X such that

$$Z_t^X = \mathbf{1}\{T = t\}(Y_{\text{obs}} - \widehat{\mu}_{t_0}(\boldsymbol{X})) + \sum_{t' \neq t} \mathbf{1}\{T = t'\}(\widehat{\mu}_t(\boldsymbol{X}) - Y_{\text{obs}})$$
$$+ \sum_{t' \neq t} \mathbf{1}\{T = t'\}(\widehat{\mu}_{t'}(\boldsymbol{X}) - \widehat{\mu}_{t_0}(\boldsymbol{X}))$$

where $(\widehat{\mu}_t)_{t \in \mathcal{T}}$ are estimator of the outcome models $(\mu_t)_{t \in \mathcal{T}}$

Proof: direct calculations

Advantage: Has the lowest variance because of the Regression-Adjustment formula.

Inconvenient: You need to have a well estimation of $(\widehat{\mu}_t)_{t \in \mathcal{T}}$.

Pseudo-outcome meta-learners: X-learner

In the binary setting, we had the original version of X-learner by Künzel et al. [2019]

$$Z_t^X = e(\boldsymbol{X})(Y_{\mathrm{obs}} - \widehat{\mu}_0(\boldsymbol{X})) + (1 - e(\boldsymbol{X}))(\widehat{\mu}_1(\boldsymbol{X}) - Y_{\mathrm{obs}})$$

Pseudo-outcome meta-learners: X-learner

In the binary setting, we had the original version of X-learner by Künzel et al. [2019]

$$Z_t^X = e(\boldsymbol{X})(Y_{\mathrm{obs}} - \widehat{\mu}_0(\boldsymbol{X})) + (1 - e(\boldsymbol{X}))(\widehat{\mu}_1(\boldsymbol{X}) - Y_{\mathrm{obs}})$$

and developed later by Curth and van der Schaar [2021] into

$$Z_t^X = T(Y_{\mathrm{obs}} - \widehat{\mu}_0(\boldsymbol{X})) + (1 - T)(\widehat{\mu}_1(\boldsymbol{X}) - Y_{\mathrm{obs}})$$

Pseudo-outcome meta-learners: X-learner

In the binary setting, we had the original version of X-learner by Künzel et al. [2019]

$$Z_t^X = e(\boldsymbol{X})(Y_{\mathrm{obs}} - \widehat{\mu}_0(\boldsymbol{X})) + (1 - e(\boldsymbol{X}))(\widehat{\mu}_1(\boldsymbol{X}) - Y_{\mathrm{obs}})$$

and developed later by Curth and van der Schaar [2021] into

$$Z_t^X = T(Y_{\mathrm{obs}} - \widehat{\mu}_0(\boldsymbol{X})) + (1 - T)(\widehat{\mu}_1(\boldsymbol{X}) - Y_{\mathrm{obs}})$$

Random reviewer: The proposed extension seems to be complicated and not intuitive! why did you not consider the naive extension at t and t_0?

Pseudo-outcome meta-learners: X-learner

In the binary setting, we had the original version of X-learner by Künzel et al. [2019]

$$Z_t^X = e(\boldsymbol{X})(Y_{\mathrm{obs}} - \widehat{\mu}_0(\boldsymbol{X})) + (1 - e(\boldsymbol{X}))(\widehat{\mu}_1(\boldsymbol{X}) - Y_{\mathrm{obs}})$$

and developed later by Curth and van der Schaar [2021] into

$$Z_t^X = T(Y_{\mathrm{obs}} - \widehat{\mu}_0(\boldsymbol{X})) + (1 - T)(\widehat{\mu}_1(\boldsymbol{X}) - Y_{\mathrm{obs}})$$

Random reviewer: The proposed extension seems to be complicated and not intuitive! why did you not consider the naive extension at t and t_0?

To be discussed... In the next talk.

Neyman orthogonality based learners

Neyman orthogonality based learners: use the Robinson [1988] decomposition and the Neyman-Orthogonality and address a minimization problem with respect to a causal component.

Neyman orthogonality based learners

Neyman orthogonality based learners: use the Robinson [1988] decomposition and the Neyman-Orthogonality and address a minimization problem with respect to a causal component.

Advantage: Flexible representation of CATEs estimation problem.

Inconvenient: Solving this problem is very challenging.

Neyman orthogonality based learners

Let $\epsilon = Y_{\text{obs}} - \mu_T(\boldsymbol{X})$ be the counterfactual model error.

Neyman orthogonality based learners

Let $\epsilon = Y_{\mathrm{obs}} - \mu_T(\boldsymbol{X})$ be the counterfactual model error. Then ϵ satisfies Neyman-Orthogonality propriety $\mathbb{E}(\epsilon \mid T, \boldsymbol{X}) = 0$ and the generalized Robinson [1988] decomposition

$$\epsilon = Y_{\mathrm{obs}} - m(\boldsymbol{X}) - \sum_{t \neq t_0} \big(\mathbf{1}\{T = t\} - r(t, \boldsymbol{X})\big)\tau_t(\boldsymbol{X})$$

where $r(t, \boldsymbol{x}) = \mathbb{P}(T = t \mid \boldsymbol{X} = \boldsymbol{x})$ and $m(\boldsymbol{x}) = \mathbb{E}(Y_{\mathrm{obs}} \mid \boldsymbol{X} = \boldsymbol{x})$.

Neyman orthogonality based learners

Let $\epsilon = Y_{\text{obs}} - \mu_T(\boldsymbol{X})$ be the counterfactual model error. Then ϵ satisfies Neyman-Orthogonality propriety $\mathbb{E}(\epsilon \mid T, \boldsymbol{X}) = 0$ and the generalized Robinson [1988] decomposition

$$\epsilon = Y_{\text{obs}} - m(\boldsymbol{X}) - \sum_{t \neq t_0} \big(\mathbf{1}\{T = t\} - r(t, \boldsymbol{X})\big)\tau_t(\boldsymbol{X})$$

where $r(t, \boldsymbol{x}) = \mathbb{P}(T = t \mid \boldsymbol{X} = \boldsymbol{x})$ and $m(\boldsymbol{x}) = \mathbb{E}(Y_{\text{obs}} \mid \boldsymbol{X} = \boldsymbol{x})$.

Sketch of the Proof: Step 1: Neyman-Orthogonality propriety

Neyman orthogonality based learners

Let $\epsilon = Y_{\text{obs}} - \mu_T(\boldsymbol{X})$ be the counterfactual model error. Then ϵ satisfies Neyman-Orthogonality propriety $\mathbb{E}(\epsilon \mid T, \boldsymbol{X}) = 0$ and the generalized Robinson [1988] decomposition

$$\epsilon = Y_{\text{obs}} - m(\boldsymbol{X}) - \sum_{t \neq t_0} \big(\mathbf{1}\{T = t\} - r(t, \boldsymbol{X})\big)\tau_t(\boldsymbol{X})$$

where $r(t, \boldsymbol{x}) = \mathbb{P}(T = t \mid \boldsymbol{X} = \boldsymbol{x})$ and $m(\boldsymbol{x}) = \mathbb{E}(Y_{\text{obs}} \mid \boldsymbol{X} = \boldsymbol{x})$.

Sketch of the Proof: Step 1: Neyman-Orthogonality propriety

$$\begin{aligned}
\mathbb{E}\big[\epsilon \mid T = t, \boldsymbol{X} = \boldsymbol{x}\big] &= \mathbb{E}\big[Y_{\text{obs}} - \mu_T(\boldsymbol{X}) \mid T = t, \boldsymbol{X} = \boldsymbol{x}\big] \\
&= \mathbb{E}\big[Y(t) - \mu_T(\boldsymbol{X}) \mid T = t, \boldsymbol{X} = \boldsymbol{x}\big] \qquad \text{(by Unconfoundedness)} \\
&= \mu_t(\boldsymbol{x}) - \mu_t(\boldsymbol{x}) = 0.
\end{aligned}$$

Neyman orthogonality based learners

Step 2: The observed outcome model satisfies

$$
\begin{aligned}
m(\boldsymbol{X}) &= \mathbb{E}(Y_{\text{obs}} \mid \boldsymbol{X} = \boldsymbol{x}) \\
&= \mathbb{E}\Big[\epsilon + \sum_{t \in \mathcal{T}} \mathbf{1}\{T = t\} \mu_t(\boldsymbol{X}) \mid \boldsymbol{X} = \boldsymbol{x}\Big] \\
&= \sum_{t \in \mathcal{T}} \mathbb{E}\Big[\mathbf{1}\{T = t\} \mid \boldsymbol{X} = \boldsymbol{x}\Big] \mu_t(\boldsymbol{x}) \\
&= \text{...direct calculations...} \\
&= \mu_{t_0}(\boldsymbol{x}) + \sum_{t \neq t_0 \in \mathcal{T}} r(t, \boldsymbol{x}) \tau_t(\boldsymbol{x}).
\end{aligned}
$$

Neyman orthogonality based learners

Step 3: Gather both terms to obtain

$$\begin{aligned} Y_{\text{obs}} - m(\boldsymbol{X}) &= \mu_T(\boldsymbol{X}) - m(\boldsymbol{X}) + \epsilon \\ &= \sum_{t \in \mathcal{T}} \mathbf{1}\{T = t\}\mu_t(\boldsymbol{X}) - \mu_{t_0}(\boldsymbol{X}) - \sum_{t \neq t_0 \in \mathcal{T}} r(t, \boldsymbol{X})\tau_t(\boldsymbol{X}) + \epsilon \\ &= \text{...direct calculations...} \\ &= \sum_{t \neq t_0 \in \mathcal{T}} \left[\mathbf{1}\{T = t\} - r(t, \boldsymbol{X})\right]\tau_t(\boldsymbol{X}) + \epsilon. \end{aligned}$$

Neyman orthogonality based learners: R-learner

R-Learner: Estimate all $K-1$ CATE models $\{\tau_t\}_{t\neq 0}$ by minimizing the error $\boldsymbol{\epsilon} = (\epsilon_i^2)_{i=1}^n$ and address the problem:

Neyman orthogonality based learners: R-learner

R-Learner: Estimate all $K-1$ CATE models $\{\tau_t\}_{t\neq 0}$ by minimizing the error $\boldsymbol{\epsilon} = (\epsilon_i^2)_{i=1}^n$ and address the problem:

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

where $\mathcal{F}$ is the space of candidate models (e.g. linear models).

Evaluation on a semi-synthetic dataset

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).

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.

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.

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

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

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.

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

Application on the estimation of multi-valued CATEs iii

mPEHE for XGBoost and RandomForest

Meta-learner	XGBoost	RandomForest
T-learner	0.167	0.154
Reg T-Learner	0.153	**0.153**
S-learner	0.101	0.216
M-learner	1.05	0.907
DR-learner	0.100	0.162
X-learner	**0.095**	0.175
RLin-learner	0.336	0.338

mPEHE $= \frac{1}{K-1}\sum_{t\neq t_0}\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left[\widehat{\tau}_t(\boldsymbol{X}_i)-\tau_t(\boldsymbol{X}_i)\right]^2}$: The mean of the Precision in Estimation of Heterogeneous Effect [Shalit et al., 2017] over all possible treatment levels.

Conclusion

Conclusion

Theory and Numerical evaluation:

- The extension of Heterogeneous Treatment Effects to the multi-valued treatment setting.
- Development of the X- and R-learners in the multi-valued treatment setting.
- Conception and creation of a semi-synthetic dataset for validating causal inference methods.

Next talk: Discussion about theory, limits and perspectives

- Comparison of the errors bounds of the pseudo-outcome meta-learners.
- Are the extensions proposed to X- and R-learners worthy?
- Sample-Splitting for CATEs estimation.
- Discussion about the numerical results: S-learner and over-fitting.

Questions?

References

K. Arceneaux, A. S. Gerber, and D. P. Green. Comparing experimental and matching methods using a large-scale voter mobilization experiment. *Political Analysis*, 14(1):37–62, 2006.

A. Curth and M. van der Schaar. Nonparametric estimation of heterogeneous treatment effects: From theory to learning algorithms. In A. Banerjee and K. Fukumizu, editors, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 1810–1818. PMLR, 13–15 Apr 2021. URL `https://proceedings.mlr.press/v130/curth21a.html`.

A. P. Dawid. Causal inference without counterfactuals. *Journal of the American Statistical Association*, 95(450):407–424, 2000. doi: 10.1080/01621459.2000.10474210.

S. Du, J. Lee, and F. Ghaffarizadeh. Improve user retention with causal learning. In *Proceedings of Machine Learning Research*, volume 104 of *Proceedings of Machine Learning Research*, pages 34–49. PMLR, 05 Aug 2019.

M. A. Hernán. A definition of causal effect for epidemiological research. *Journal of Epidemiology & Community Health*, 58(4):265–271, 2004. ISSN 0143-005X. doi: 10.1136/jech.2002.006361.

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

D. G. Horvitz and D. J. Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):663–685, 1952. ISSN 01621459.

S. A. Imberman. How effective are financial incentives for teachers. *The IZA World of Labor*, pages 158–158, 2015.

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.

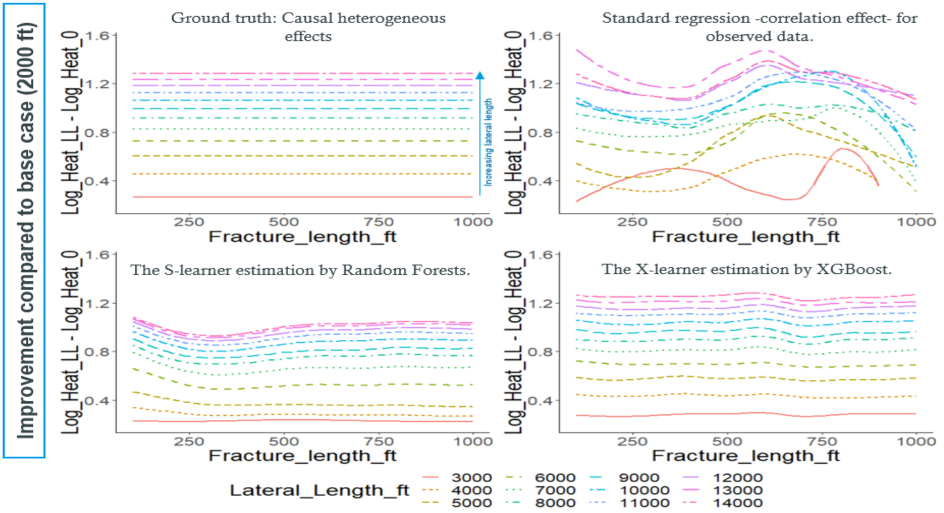
S. L. Morgan and C. Winship. *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. Analytical Methods for Social Research. Cambridge University Press, 2 edition, 2014. doi: 10.1017/CBO9781107587991.

J. M. Robins, A. Rotnitzky, and L. P. Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427): 846–866, 1994.

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.


Improvement compared to base case (2000 ft)
Ground truth: Causal heterogeneous effects
Standard regression -correlation effect- for observed data.
The S-learner estimation by Random Forests.
The X-learner estimation by XGBoost.
Log_Heat_LL - Log_Heat_0
Fracture_length_ft
Increasing lateral length
1.6
1.2
0.8
0.4
250
500
750
1000
Lateral_Length_ft
3000
4000
5000
6000
7000
8000
9000
10000
11000
12000
13000
14000