Asymmetrical Latent Representation for Individual Treatment Effect Modeling

Michèle Sebag

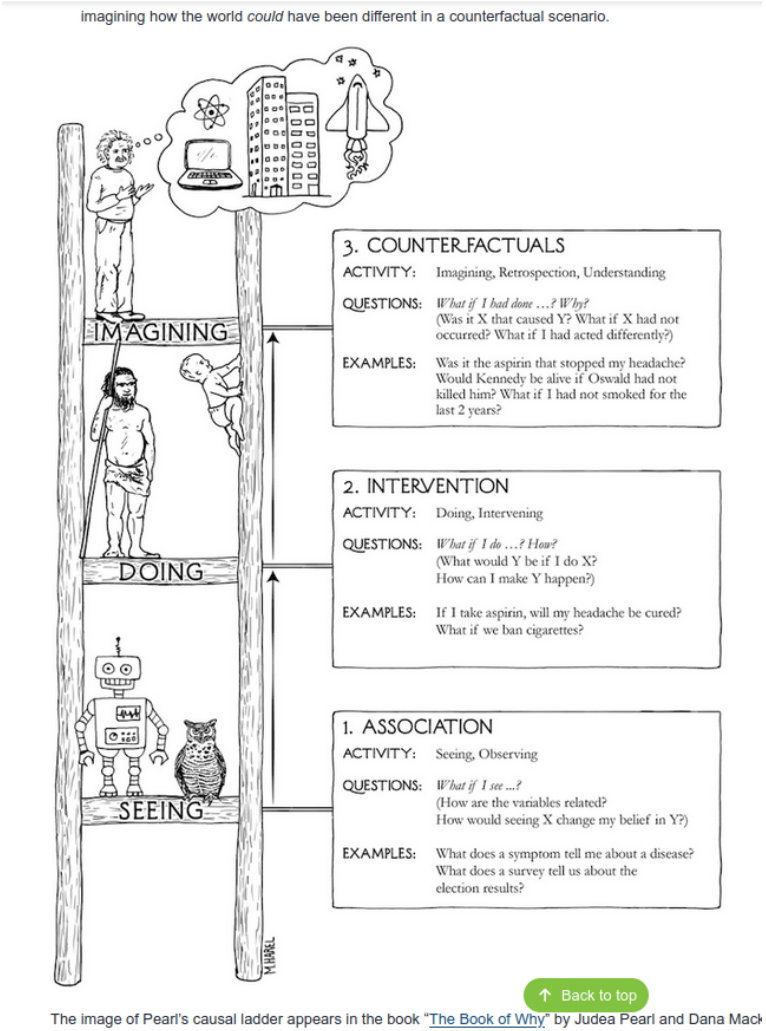
CNRS – INRIA – LISN, UMR 9015, U. Paris-Saclay

March 24th, 2026

Joint work: Armand Lacombe

The causal ladder

Pearl & McKenzie 2018



Seing: *"given the symptoms, what is the patient's disease?"*

Doing: *"will aspirin cure their headache?"*

Imagining: *"had they taken an aspirin one hour ago, would their headache have been cured?"*

Intervening, Imagining

If Cleopatra's nose had been shorter, the whole face of the earth would have changed.

Notations

- X description of the sample (covariates)
- T (in 0/1) treated / control sample
- Y^T outcome

Goal: Estimate

- Average Treatment Effect (ATE)

$$ATE = \mathbf{E}[Y^1] - \mathbf{E}[Y^0]$$

- Conditional Average Treatment Effect (CATE)

$$\tau(x) = \mathbf{E}[Y^1|X = x] - \mathbf{E}[Y^0|X = x]$$

Data

- $\mathcal{D} = \{(x_i, t_i, y_i), i = 1 \ldots n\}$

Counterfactual reasoning

Estimate what would have been the outcome

if things had been otherwise

- Given (x_i, t_i, y_i)
- ... find $(x_i, 1 - t_i, \ ??)$

Why this is not ML

- In supervised ML, given $\{(x_i, y_i)\}$, find $(x', y' = ?)$
- In counter-factual reasoning, you know (x_i, t_i, y_i), you never know $(x_i, 1 - t_i, ??)$

Assumptions

Rosenbaum-Rubin 1983

1. **Conditional exchangeability**
   No hidden confounding variable.
   Formally, $Y^t \perp\!\!\!\perp T|X, \ \forall t \in \{0,1\}$

   Conditional exchangeability is **untestable** (a leap of faith)

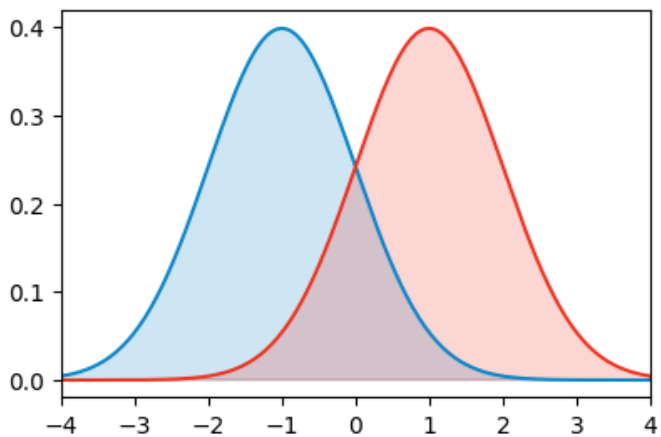
Assumptions, 2/3

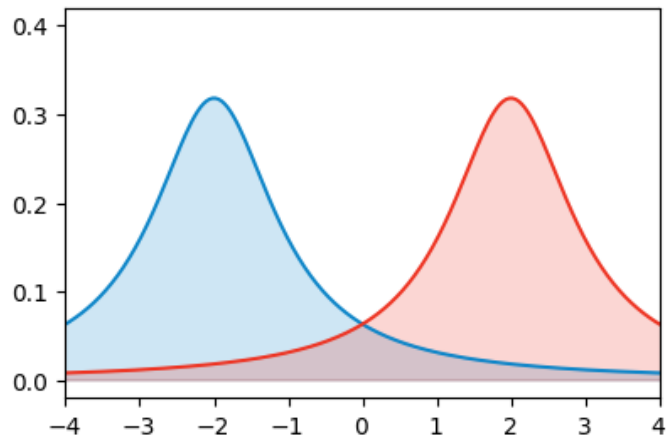
2. Positivity

Any treatment might be assigned to any individual:

$$\forall\, A \subset \mathcal{X}\ ,\ \mathbb{P}(X \in A) > 0 \implies \begin{cases} \mathbb{P}(T = 1 | X \in A) > 0 \\ \mathbb{P}(T = 0 | X \in A) > 0 \end{cases}$$

The more features, the more likely conditional exchangeability holds, but the less likely positivity does.





Assumptions, 3/3

Rubin, 1980

3. Stable Unit Treatment Value Assumption (SUTVA)

- **No interaction**: the outcome of any individual should not interfere with that of other individuals;
- **Consistency**: the observed outcome y_i corresponds to the potential outcome associated with t_i

$$Y = TY^1 + (1 - T)Y^0$$

SUTVA: does not hold if individuals compete for a shared resource, or in settings where contagion and herd immunity are possible.

Why these assumptions ?

$$\begin{aligned}\mathbb{E}[Y^0|X=x] &= \mathbb{E}[Y^0|X=x, T=0] \textbf{ conditional exchangeability}\\ &= \mathbb{E}[Y^T|X=x, T=0]\\ &= \mathbb{E}[Y\ \ |X=x, T=0] \textbf{ SUTVA - consistency}\end{aligned}$$

and **positivity** makes estimating $\mathbb{E}[Y|X=x, T=0]$ possible.

(Same for $T=1$).

Therefore *ATE* and *CATE* can be estimated

$$\begin{aligned}ATE &= \mathbb{E}_X\left[\mathbb{E}[Y|X, T=1] - \mathbb{E}_X[Y|X, T=0]\right]\\ \tau(x) &= \mathbb{E}[Y|X=x, T=1] - \mathbb{E}[Y|X=x, T=0]\end{aligned}$$

Randomized Controlled Trial: in an ideal world

No selection bias

- $T \perp\!\!\!\perp X$
- $P_1(X) = P(X|T=1)$ treated distribution
- $P_0(X) = P(X|T=0)$ control distribution
- and

$$P_1(X) = P_0(X)$$

Hence

$$\begin{aligned} ATE =& \mathbb{E}[Y^1] - \mathbb{E}[Y^0] \\ =& \mathbb{E}[Y^1|T=1] - \mathbb{E}[Y^0|T=0] \\ =& \underbrace{\mathbb{E}[Y|T=1]}_{\text{may be estimated}} - \underbrace{\mathbb{E}[Y|T=0]}_{\text{may be estimated}} \end{aligned}$$

In practice

RCT with selection biases

- In medicine: drug (treated) if in serious health condition
- In education: curriculum (treated) depending on background

RCT, limitations

- limitations: Cost, Ethics, Feasibility

Observational causal studies

- Treated and Control data (selection biases; hidden confounders)

$$P_1(X) \neq P_0(X)$$

ML and Counter-factual reasoning

- Learning with missing values
- Missing not at random !

State of the art

$\mapsto$2016 ML-based methods, typically trees, where T is a variable among others;

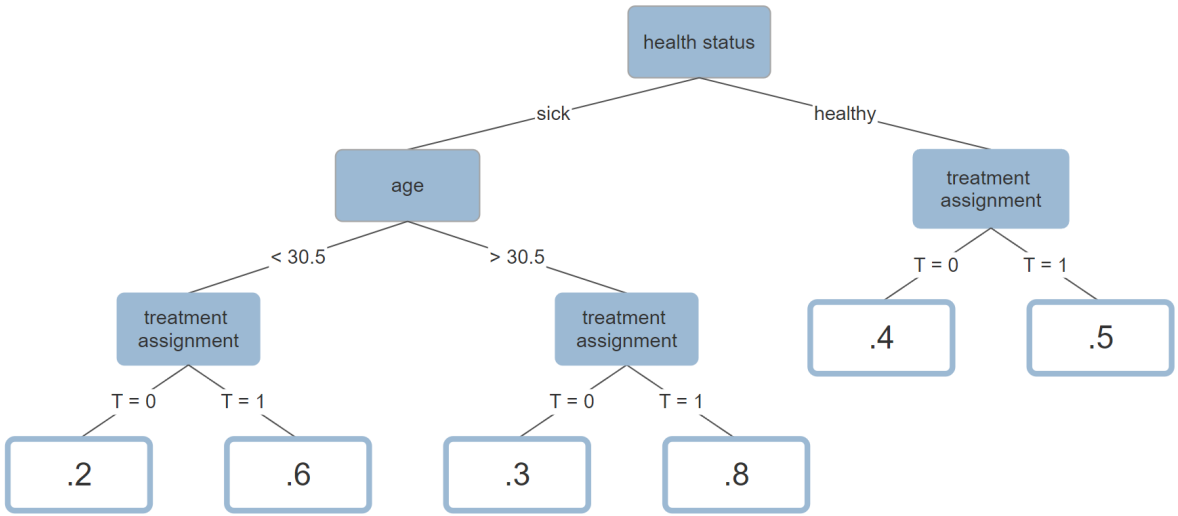
2016 neural network-based method, T among others;

2017$\mapsto$ neural network-based methods, special role of T;

2019$\mapsto$ neural network-based methods, special role of T, latent space disentanglement

Causal Forests

Athey Imbens 2016; Athey Tibshirani Wager 2019

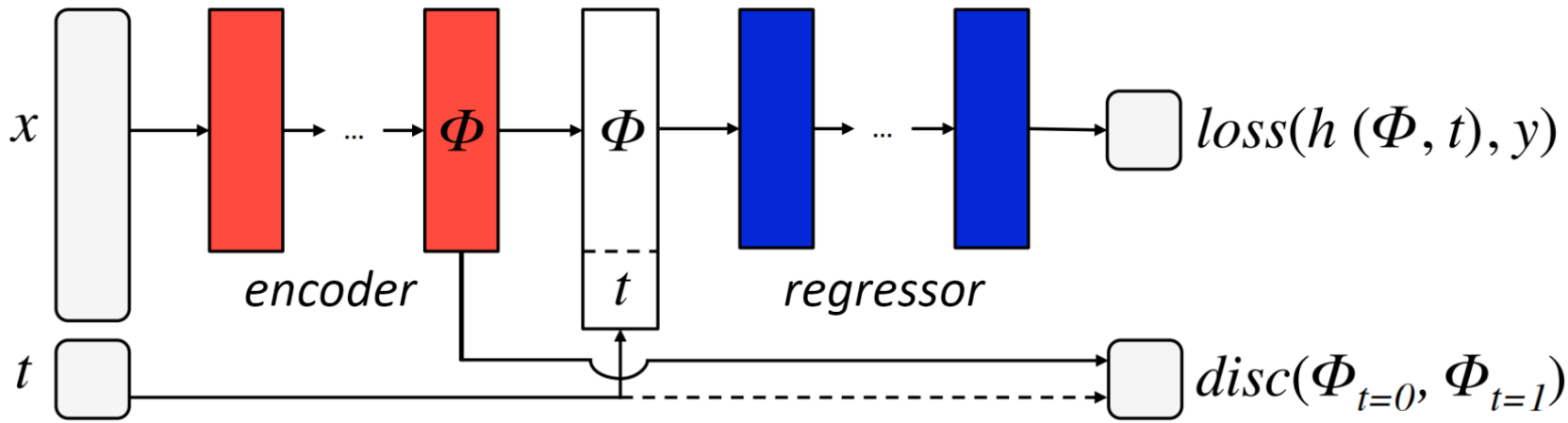


A causal tree

Causal Forests: random regression forest, built on Causal Trees. Equivalent to regression trees whose deepest leaf split is treatment assignment.

Balancing Neural Network

Johansson et al. 2016



Architecture of *BNN*, with mapping ϕ and outcome model h

BNN: one-headed, based on neural networks. Constrained imbalance between the control and treatment latent populations.

BNN: the relationship with Domain Adaptation

Domain adaptation

train $\neq$ test

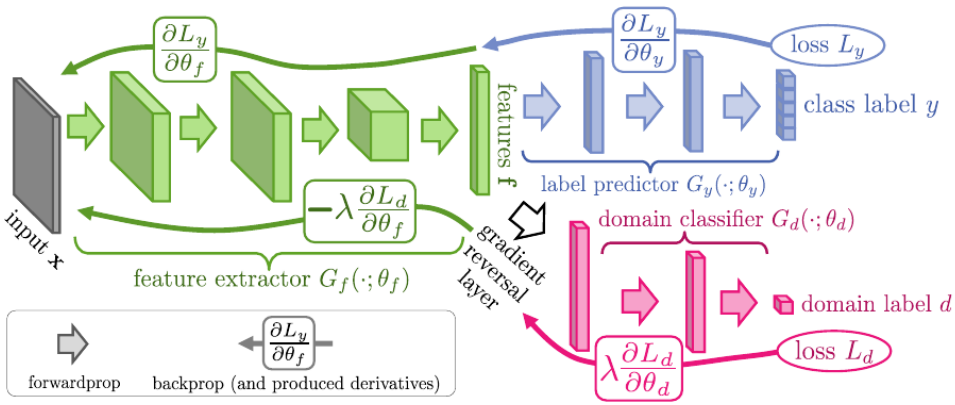
BNN: the relationship with Domain Adaptation

Domain adaptation

train $\neq$ test

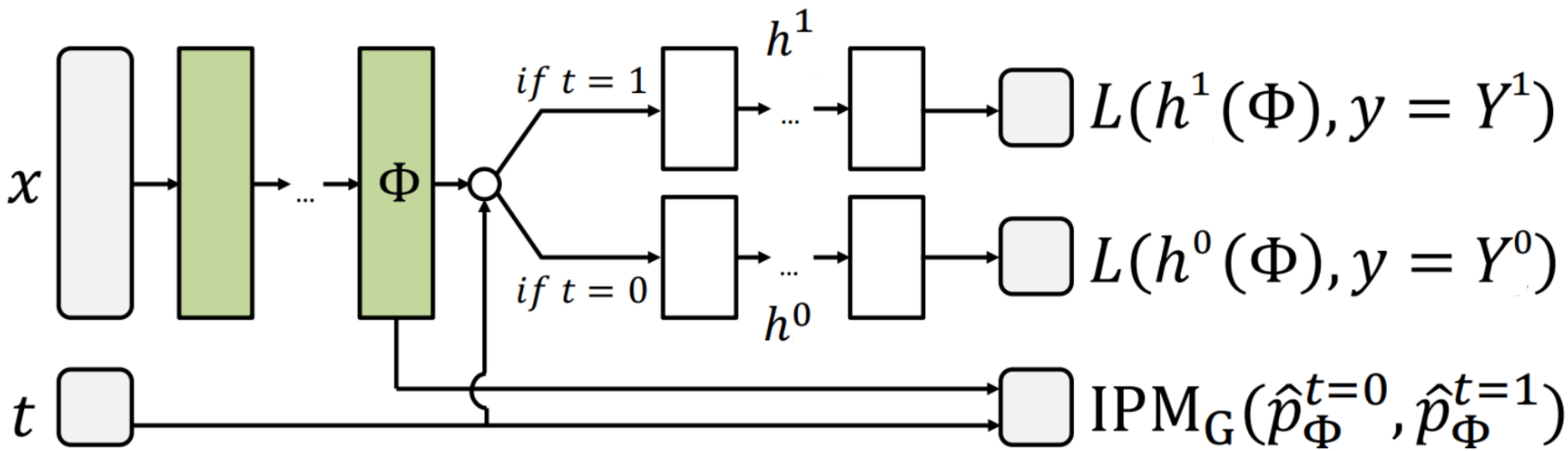
Finding a common latent space

Ganin et al. 2016



CounterFactual Regression

Shalit et al. 2017



Architecture of *CFR*, with mapping ϕ and outcome models h_0, h_1

CFR relies on a two-headed (h^0, h^1) architecture. The latent space imbalance is also constrained.

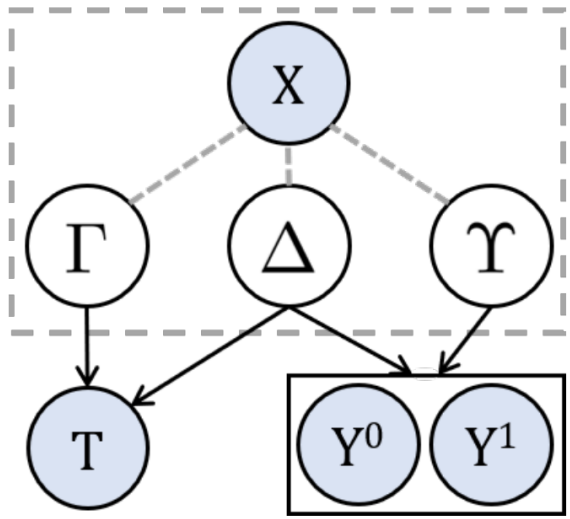
Disentangling the latent space

Hassanpour and Greiner, 2019
Zhang et al. 2021
Chauhan et al., 2023

Exploiting prior knowledge: if you know that

- (Γ) cause T only;
- (Δ) cause T and (Y^0, Y^1);
- (Υ) cause (Y^0, Y^1) only;

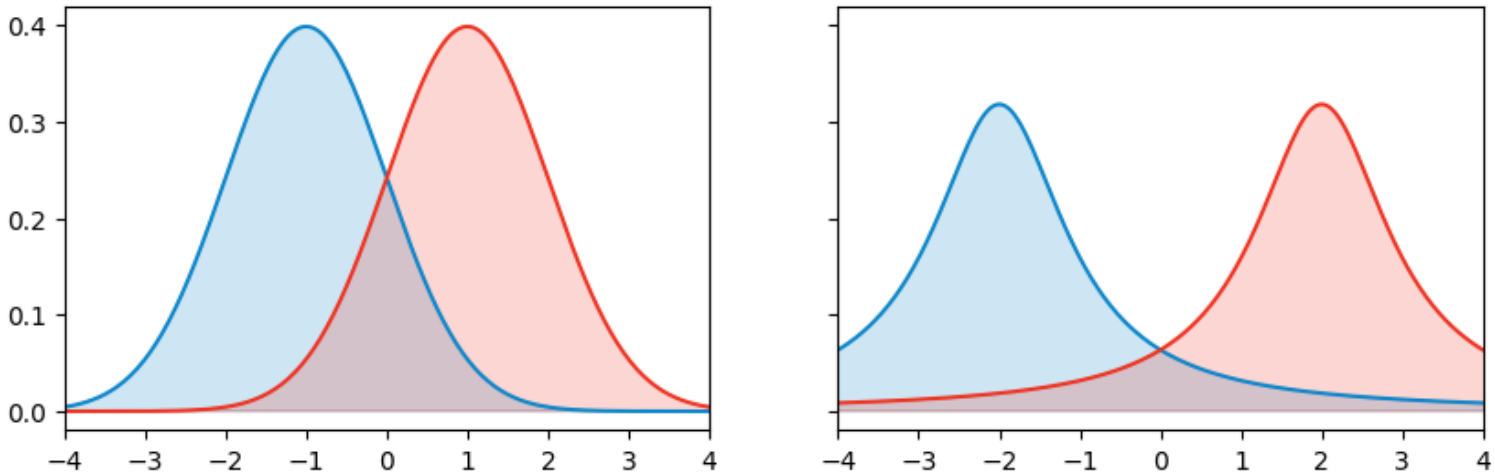
Focus: splitting the latent space into three subspaces.



Latent space splitting

Domain adaptation $\neq$ CATE

Situation of latent control and latent treated

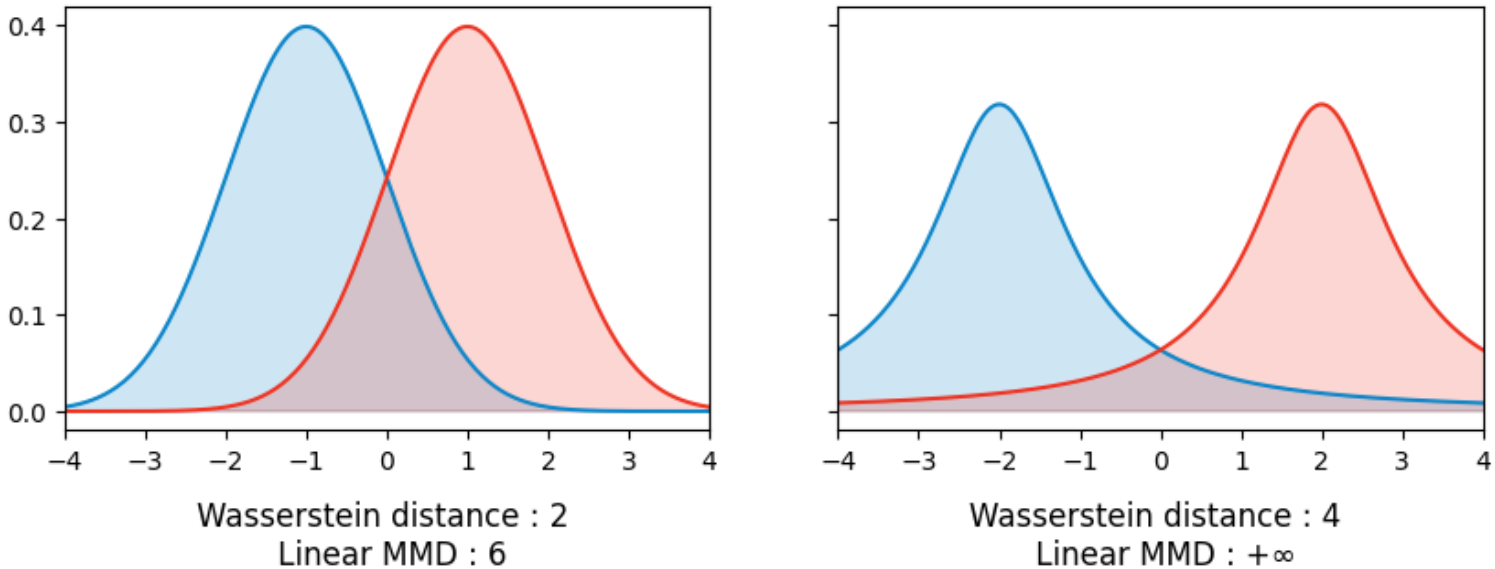


Common discrepancy measures in two latent distribution settings

Which situation is best ?

Domain adaptation $\neq$ CATE

Situation of latent control and latent treated



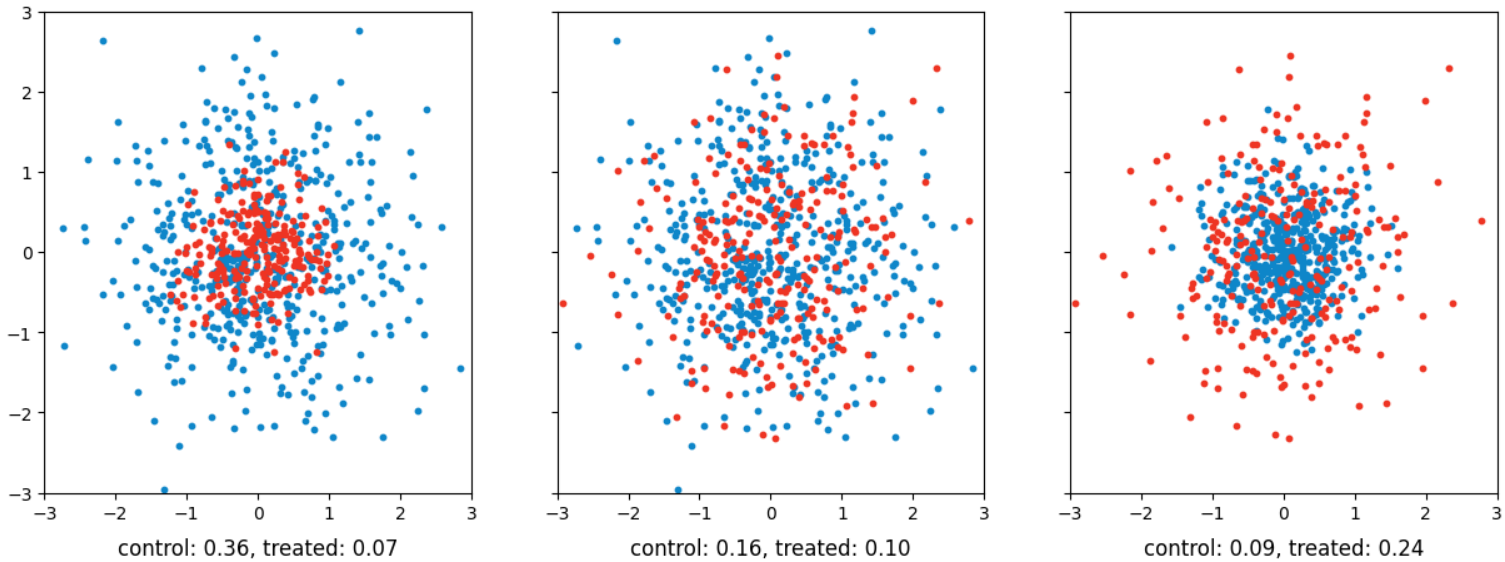
Common discrepancy measures in two latent distribution settings

Which situation is best ?

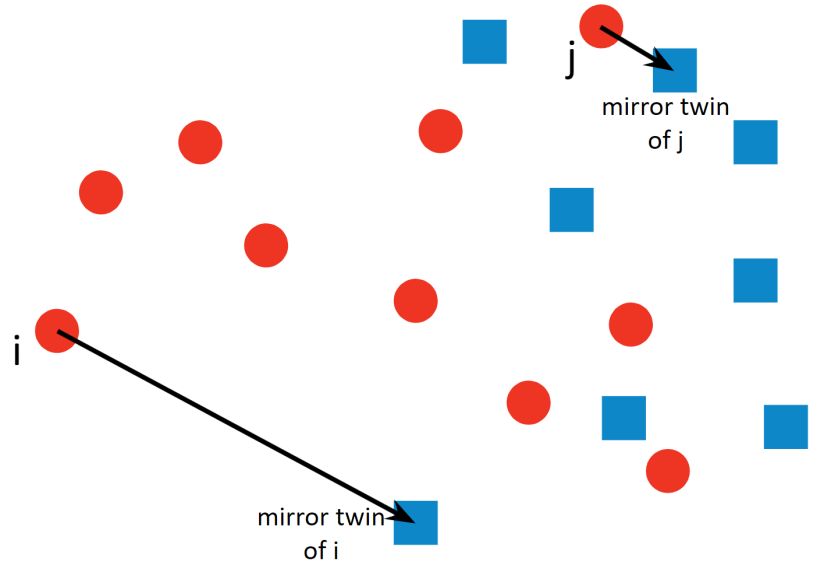
- For domain adaptation, left
- For CATE, right

What matters ?

To estimate Y^0 for a **treated sample** $(x_i, t_i = 1, y_i)$,
this sample must be close to **control samples**



Definitions



Definition: Mirror twin

$$mt(x_i, t_i, y_i) = argmin\{d(x_i, x_j),\ (x_j, t_j = 1 - t_i, y_j) \in \mathcal{D}\}$$

$$d(x_i, x_j) = \|\phi(x_i) - \phi(x_j)\|$$

Definition: Insulation

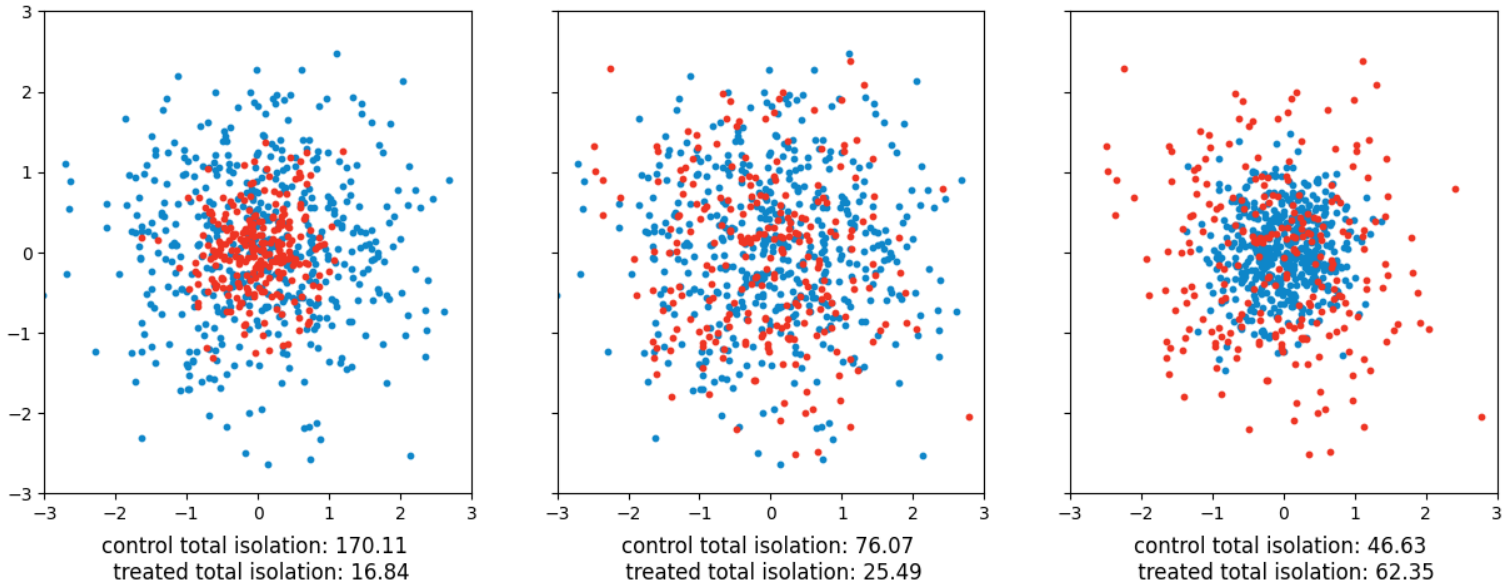
Given (x_i, t_i, y_j) and $mt_i = (x_j, t_j, y_j)$,

$$Isol(x_i) = d(x_i, x_j)$$

Conjecture

If $Isol(x_i)$ small, $PEHE(x_i)$ small too

CATE: Two estimation problems

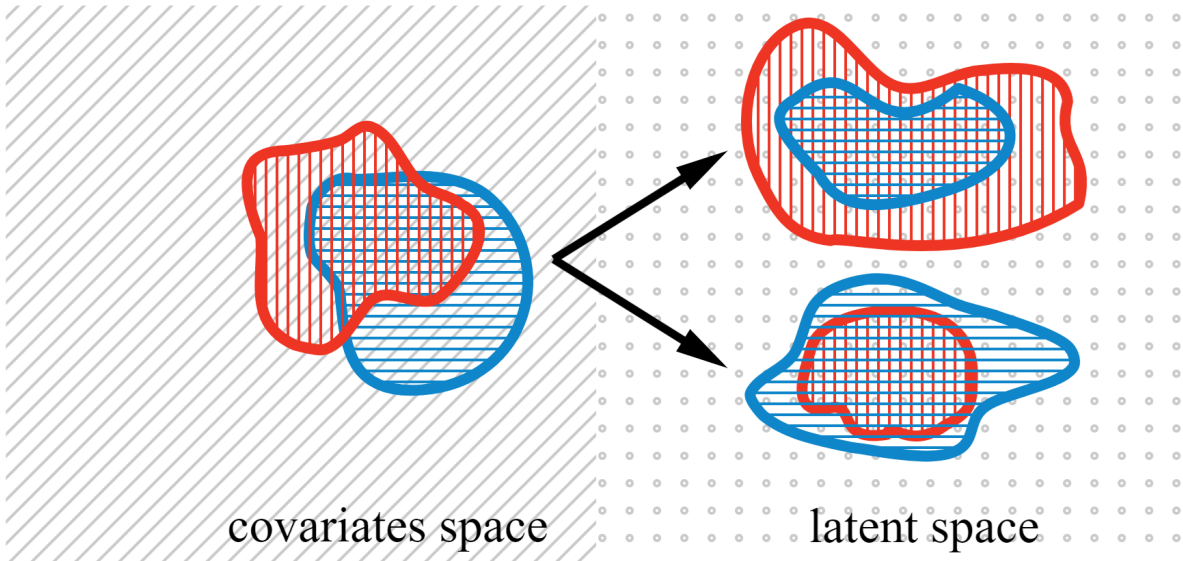


treated and control distributions isolation

We need

- one latent space enforcing low isolation for **treated samples** (left)
- one latent space enforcing low isolation for **control samples** (right)

A two-pipeline architecture



Top: pipeline $\mathcal{P}_0$; bottom: pipeline $\mathcal{P}_1$

Two latent spaces.

- **Control-driven** pipeline $\mathcal{P}_0$ aims for *CATE* estimation of **control samples**
- **Treatment-driven** pipeline $\mathcal{P}_1$ aims for *CATE* estimation of **treated samples**

Detail of treatment-driven pipeline $\mathcal{P}_1$

Estimating *CATE* for **treated samples**

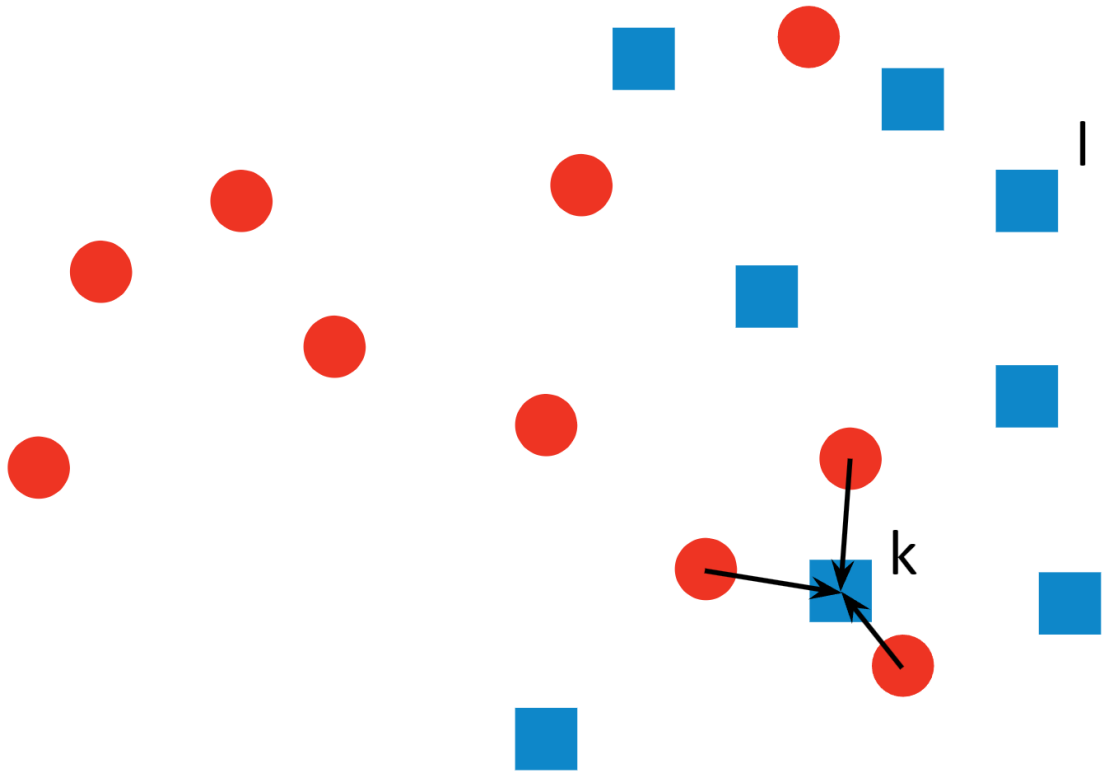
- modelling $\mathbb{E}[Y^1|T=1, X=x]$ (general)
- modelling $\mathbb{E}[Y^0|T=0, X=x]$ (general)
- achieving low isolation for treated samples (specific of $\mathcal{P}_1$)

Loss

$$\mathcal{L}_{\mathcal{P}_1} = \underbrace{\frac{1}{n_1}\sum_{t_i=1} \underset{\mathcal{P}_1}{error}(x_i, t_i, y_i)}_{\text{treated factual MSE}} + \underbrace{\frac{1}{n_0}\sum_{t_i=0} \underset{\mathcal{P}_1}{error}(x_i, t_i, y_i)}_{\text{control factual MSE}}$$

$$+\alpha_1 \underbrace{\frac{1}{n_1}\sum_{t_i=1} \underset{\mathcal{P}_1}{Isol}(i)^2}_{\text{average isolation}} + \gamma_1 \underbrace{||\mathcal{P}_1||_2^2}_{\text{regularization}}$$

Remark: Not all control samples are equal



C-weight(k)=3, *C-weight*(l)=0

C-weight(k): counter-factual weight, number of samples whose mirror twin is k. High **C-weight** implies **high importance** in the estimation of counterfactual distribution.

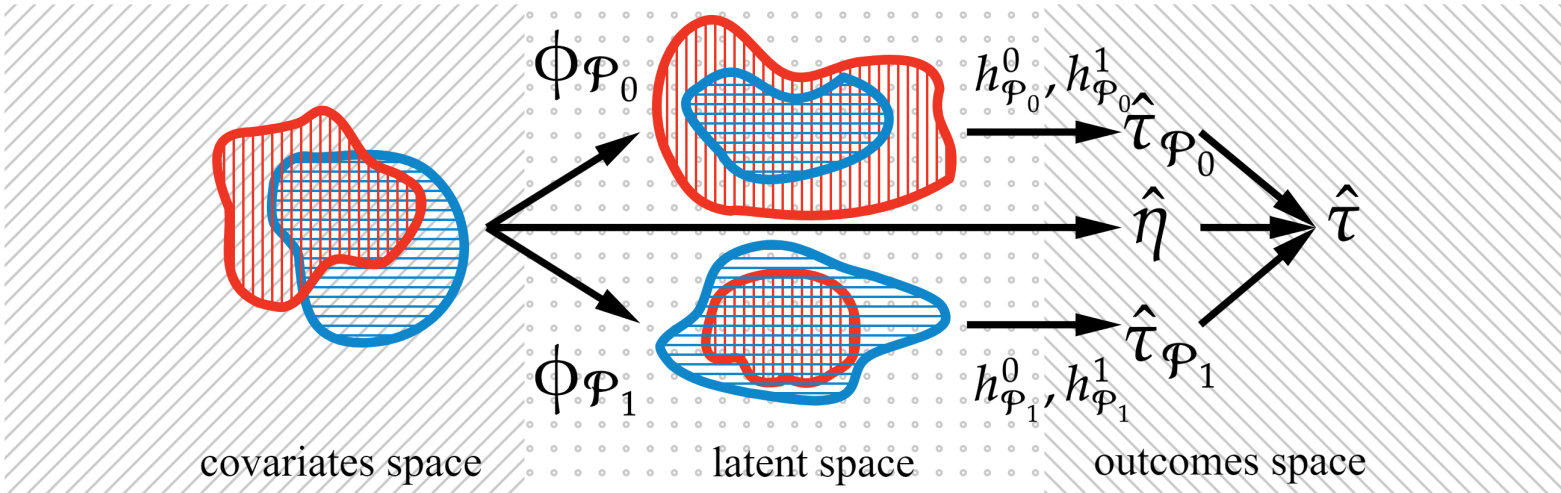
Detail of treatment-driven pipeline $\mathcal{P}_1$, followed

Loss

$$\mathcal{L}_{\mathcal{P}_1} = \frac{1}{n_1}\sum_{t_i=1} \underset{\mathcal{P}_1}{error}(x_i, t_i, y_i) + \alpha_1 \frac{1}{n_1}\sum_{t_i=1} \underset{\mathcal{P}_1}{Isol}(i)^2 + \gamma_1 ||\mathcal{P}_1||_2{}^2$$
$$+ \frac{1}{n_0 + \beta_1 n_1} \sum_{t_i=0} \underbrace{(1 + \beta_1 \underset{\mathcal{P}_1}{\textit{C-weight}}(i))}_{\textit{importance weighting}} \times \underset{\mathcal{P}_1}{error}(x_i, t_i, y_i)$$

Architecture of Alrite

Given $\mathcal{P}_1$ and $\mathcal{P}_0$



Learn propensity score

$$\eta(x) = \mathbb{E}[T|X = x]$$

Define overall CATE

$$\hat{\tau} = (1 - \hat{\eta}) \times \hat{\tau}_{\mathcal{P}_0} + \hat{\eta} \times \hat{\tau}_{\mathcal{P}_1}$$

Upper-bounding the PEHE of a T-learner

- if the factual predictions are good enough;
- if the outcome functions are smooth enough;
- if the average squared isolation is small enough;

then the *PEHE* is upper-bounded.

$$M_1 = \frac{4}{n}\sum_{i=1}^{n}\left(1 + \underset{\phi}{\text{C-weight}}(i)\right)\left(\hat{\nu}^{t_i} \circ \phi(x_i) - \mu^{t_i}(x_i)\right)^2$$

$$+ 4(L^2 + \hat{L}^2)\frac{1}{n}\sum_{i=1}^{n}\underset{\phi}{Isol}(i)^2$$

Theorem 1

Premises

Embedding $\phi : \mathcal{X} \to \mathcal{Z}$ from feature space to latent space.
Assume there exist L-Lipschitz ν^0, ν^1 such that $\mu^t = \nu^t \circ \phi$.
Let $\hat{\nu}^0, \hat{\nu}^1 : \mathcal{Z} \to \mathbb{R}$ be $\hat{L}$-Lipschitz approximate models.

Then Then the empirical *PEHE* is upper bounded by

$$M_1 = \frac{4}{n} \sum_{i=1}^{n} \left(1 + \underset{\phi}{\textit{C-weight}}(i)\right) \left(\hat{\nu}^{t_i} \circ \phi(x_i) - \mu^{t_i}(x_i)\right)^2$$

$$+ \frac{4}{n} (L^2 + \hat{L}^2) \sum_{i=1}^{n} \underset{\phi}{\textit{Isol}}(i)^2$$

Upper-bounding the PEHE of Alrite

In the *within-sample* context (T and Y available), and given an ALRITE model:

Assume

- the factual predictions of high C-weight samples are good enough;
- the outcome functions are smooth enough;
- the average squared isolation is small enough;
- the noise level on Y is low enough;

Then

$$PEHE \leq M_2$$

with

$$M_2 = \frac{5}{n}\sum_{i=1}^{n} \underset{\mathcal{P}_{1-t_i}}{\text{C-weight}}(i)\left(h^{t_i}_{\mathcal{P}_{1-t_i}} \circ \phi_{\mathcal{P}_{1-t_i}}(x_i) - y_i\right)^2$$

$$+5(L^2 + \hat{L}^2)\frac{1}{n}\sum_{i=1}^{n} \underset{\mathcal{P}_{t_i}}{Isol}(i)^2 + 5\kappa_Y$$

Theorem 2

Assumptions on gt

Assume there exist L-Lipschitz $\nu^0_{\mathcal{P}_0}, \nu^1_{\mathcal{P}_0}, \nu^0_{\mathcal{P}_1}, \nu^1_{\mathcal{P}_1}$ such that $\mathbb{E}[Y^t|X = x] = \nu^t_{\mathcal{P}_0} \circ \phi_{\mathcal{P}_0} = \nu^t_{\mathcal{P}_1} \circ \phi_{\mathcal{P}_1}$.

Assumptions on learned models

Given $\mathcal{P}_0, \mathcal{P}_1$ be two pipelines.

Assume $h^0_{\mathcal{P}_0}, h^1_{\mathcal{P}_0}, h^0_{\mathcal{P}_1}, h^1_{\mathcal{P}_1}$ are $\hat{L}$-Lipschitz.

Then

$$\bar{\tau}_i = (1 - t_i)(h^1_{\mathcal{P}_0} \circ \phi_{\mathcal{P}_0}(x_i) - y_i) + t_i(y_i - h^0_{\mathcal{P}_1} \circ \phi_{\mathcal{P}_1}(x_i))$$

The within-sample *PEHE* of $\bar{\tau}$ is upper bounded by

$$M_2 = \frac{5}{n} \sum_{i=1}^{n} \underset{\mathcal{P}_{1-t_i}}{\textit{C-weight}}(i) \left(h^{t_i}_{\mathcal{P}_{1-t_i}} \circ \phi_{\mathcal{P}_{1-t_i}}(x_i) - y_i\right)^2$$

$$+5(L^2 + \hat{L}^2)\frac{1}{n} \sum_{i=1}^{n} \underset{\mathcal{P}_{t_i}}{\textit{Isol}}(i)^2 \; + 5\kappa_Y$$

where $\kappa_Y = \frac{1}{n} \sum_{i=1}^{n} \left(1 + \underset{\mathcal{P}_{t_i}}{\textit{C-weight}}(i)\right) \left(y_i - \mu^{t_i}(x_i)\right)^2$.

A non-constructive result

In the *within-sample* context (T and Y available), given an ALRITE model:

- there exist hyper-parameters $(\alpha_0, \alpha_1, \beta_0, \beta_1)$ such that, if the training losses $\mathcal{L}_{\mathcal{P}_0}$ and $\mathcal{L}_{\mathcal{P}_1}$ are low enough;

then the *PEHE* is upper-bounded by a computable quantity.

$$M_3 = 5\big(\mathcal{L}_{\mathcal{P}_0} + \mathcal{L}_{\mathcal{P}_1}\big) - 5\big(\gamma_0 ||\mathcal{P}_0||_2^2 + \gamma_1 ||\mathcal{P}_1||_2^2\big) + 5\kappa_Y$$
$$- \frac{5}{n}\sum_{i=1}^{n} \frac{1}{p^{t_i}} \big(h^{t_i}_{\mathcal{P}_{t_i}} \circ \phi_{\mathcal{P}_{t_i}}(x_i) - y_i\big)^2 + \big(h^{t_i}_{\mathcal{P}_{1-t_i}} \circ \phi_{\mathcal{P}_{1-t_i}}(x_i) - y_i\big)^2$$

Theorem 3

Let p^1 and p^0 be the fraction of treated and control samples in the training set. Then with adequate choice of hyper-parameters $(\alpha_0, \alpha_1, \beta_0, \beta_1)$, the *within-sample* empirical *PEHE* is upper bounded by

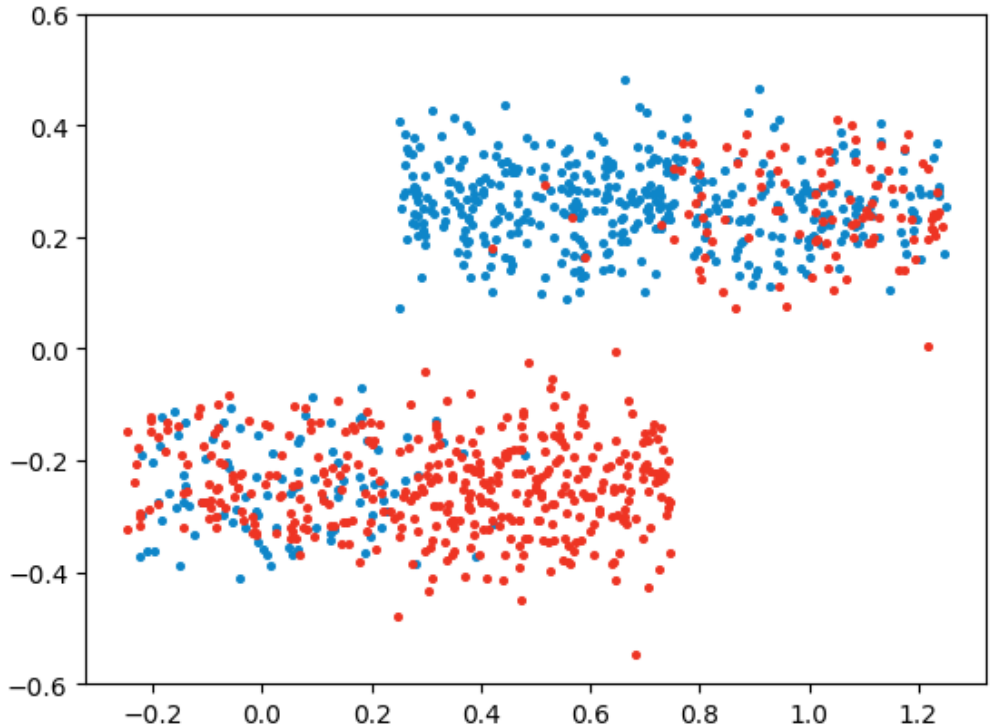
$$M_3 = 5\big(\mathcal{L}_{\mathcal{P}_0} + \mathcal{L}_{\mathcal{P}_1}\big) - 5\big(\gamma_0||\mathcal{P}_0||_2^2 + \gamma_1||\mathcal{P}_1||_2^2\big) + 5\kappa_Y$$
$$- \frac{5}{n}\sum_{i=1}^{n} \frac{1}{p^{t_i}} \big(h^{t_i}_{\mathcal{P}_{t_i}} \circ \phi_{\mathcal{P}_{t_i}}(x_i) - y_i\big)^2 + \big(h^{t_i}_{\mathcal{P}_{1-t_i}} \circ \phi_{\mathcal{P}_{1-t_i}}(x_i) - y_i\big)^2$$

Limitation 1: Potential bias towards well predicted samples

$$\mathcal{L}_{\mathcal{P}_1} = \frac{1}{n_1} \sum_{t_i=1} \underset{\mathcal{P}_1}{error}(x_i, t_i, y_i) + \frac{\alpha_1}{n_1} \sum_{t_i=1} \underset{\mathcal{P}_1}{Isol}(i)^2 + \gamma_1 ||\mathcal{P}_1||_2^2$$
$$+ \frac{1}{n_0 + \beta_1 n_1} \sum_{t_i=0} (1 + \beta_1 \underset{\mathcal{P}_1}{\text{C-weight}}(i)) \times \underset{\mathcal{P}_1}{error}(x_i, t_i, y_i)$$

A risk with the control term of $\mathcal{L}_{\mathcal{P}_1}$: instead of predicting high C-weight control samples well, allocate high importance to well predicted control samples.

Limitation 2: w.r.t. breaches of positivity



An instance space (X) bad case scenario with low overlap

If ϕ projects all samples on the first feature axis: low total isolation. But the second feature axis may be important to predict Y.

Validation: Specifics of causal inference

$\mathcal{D}_{\mathcal{V}}$: subset of the dataset held out for validation

supervised learning: compute *MSE* on $\mathcal{D}_{\mathcal{V}}$ for hyper-parameter tuning;

causal inference: no access to counterfactuals, *PEHE* **cannot be computed**

How to achieve hyper-parameter selection?

Metrics

One Nearest Neighbor imputation

Shalit et al. 17, Du et al. 21, Zhou et al. 21

PEHE approximated by (with $c(i)$: mirror twin of x_i):

$$1NNI(\hat{\tau}) = \frac{1}{|\mathcal{D}_{\mathcal{V}}|} \sum_{(x,t,y)\in\mathcal{D}_{\mathcal{V}}} \Big(\hat{\tau}(x) - (2t_i - 1)(y_i - y_{c(i)})\Big)^2$$

μ-Risk

Doutreligne & Varoquaux, 23

$$\hat{\mu}^t = (1 - \hat{\eta}) \times h^t_{\mathcal{P}_0} \circ \phi_{\mathcal{P}_0} + \hat{\eta} \times h^t_{\mathcal{P}_1} \circ \phi_{\mathcal{P}_1}$$

IHDP

Brooks-Gunn et al. 92; Hill 11

Based on *Infant Health and Development Program*.
Contains 747 individuals.

Covariates:
- 25 features describing the infants and their mothers;
- 6 continuous, 19 binary

Treatment:
- binary, child-care and home visits;
- 139 treated samples;
- imbalanced: $X \not\perp\!\!\!\perp T$

Outcome:
- simulated, enabling computation of *PEHE* on test dataset;
- available in 100 versions depending on the simulation's parameters

Results on *IHDP*

	IHDP			
	within-sample		*out-of-sample*	
	$\sqrt{PEHE}$	ϵ_{ATE}	$\sqrt{PEHE}$	ϵ_{ATE}
OLS/LR-2	$2.4 \pm .1$	$.14 \pm .01$	$2.5 \pm .1$	$.31 \pm .02$
BNN	$2.2 \pm .1$	$.37 \pm .03$	$2.1 \pm .1$	$.42 \pm .03$
CF	$3.8 \pm .2$	$.18 \pm .01$	$3.8 \pm .2$	$.40 \pm .03$
CFR-Wass	$.71 \pm .0$	$.25 \pm .01$	$.76 \pm .0$	$.27 \pm .01$
CEVAE	$2.7 \pm .1$	$.34 \pm .01$	$2.6 \pm .1$	$.46 \pm .02$
DeR-CFR	$.44 \pm .02$	$.13 \pm .02$	$.53 \pm .07$	$.15 \pm .02$
Alrite	$\mathbf{.42} \pm .03$	$.12 \pm .009$	$\mathbf{.43} \pm .02$	$.13 \pm .01$

Comparative performances averaged over the 100 versions of *IHDP*.

ACIC2016

Dorie et al. 17; Light 73

Based on the 2016 *Atlantic Causal Inference Challenge*, built on a survey regarding children's organic neurological defects.
Contains 4802 individuals.

Covariates:
- 58 features describing the children and their environment;
- 23 continuous features, 27 counts, 5 binary and 3 categorical;

Treatment:
- binary, simulated;

Outcome:
- continuous;
- simulated, enabling computation of *PEHE* on test dataset;
- available in 770 versions depending on the outcome generation protocol ($\times 77$) and parameters ($\times 10$).

Results on *ACIC2016*

	ACIC2016	
	$\sqrt{PEHE}$	
	within-sample	*out-of-sample*
CF	2.16	2.18
CFR-Wass	2.05	2.18
CEVAE	2.78	2.84
TEDVAE	**1.75**	**1.77**
Alrite	1.79	1.81

Comparative performances averaged over the 770 versions of *ACIC2016*.

Conclusion and Perspectives

Alrite

- Theoretical guarantees
- Good empirical results
- Cam be plugged on any NN method

Perspectives, short term

- co-training of both pipelines;
- extension to the multi-treatment framework;

Perspectives, medium term

- extension from mirror twin to latent K-nearest neighbors;
- adapting to the disentangled latent space framework;

Perspectives, long term

- defining overall $\hat{\tau}$ based on uncertainty rather than propensity

Thanks

Armand Lacombe