LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

LMU MUNICH SCHOOL OF MANAGEMENT

INSTITUTE OF ARTIFICIAL INTELLIGENCE (AI) IN MANAGEMENT

Munich Center for Machine Learning

Bounds on Representation-Induced Confounding Bias for Treatment Effect Estimation

Valentyn Melnychuk, Dennis Frauen, Stefan Feuerriegel

LMU Munich & Munich Center for Machine Learning (MCML), Munich, Germany

2nd Munich CausalML Workshop

Introduction: Representation learning for CATE estimation

Why this is important?

- State-of-the-art methods for conditional average treatment effect (CATE) estimation make widespread use of representation learning
- Low-dimensional (potentially constrained) representations reduce the variance, but, at the same time lose information about covariates, including information about confounders

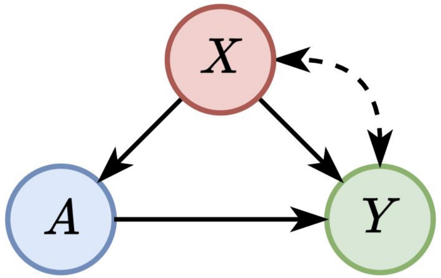
Problem formulation: representation-based CATE estimation

Given i.i.d. observational dataset $\mathcal{D} = \{X_i, A_i, Y_i\}_{i=1}^{n} \sim \mathbb{P}(X, A, Y)$

X covariates

A binary treatments

Y continuous (factual) outcomes



Representation learning methods estimate the **conditional average treatment effect (CATE)**

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

by (1) learning a low-dimensional (potentially constrained) representation $\Phi(\cdot) : X \to \Phi(X)$

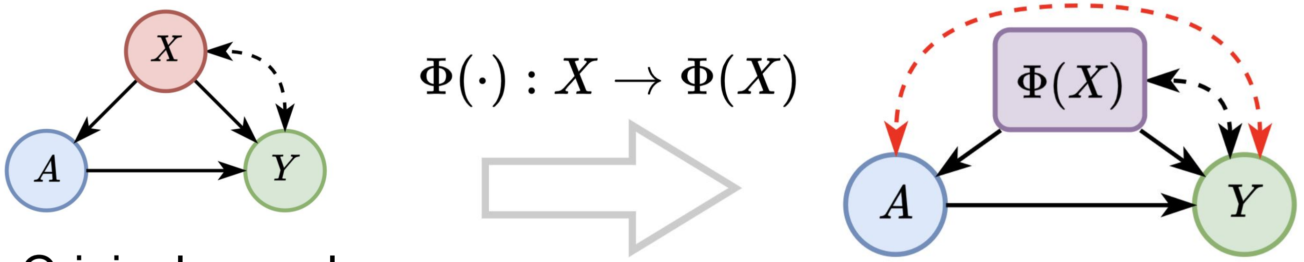
and by (2) estimating CATE wrt. representations $\mu_1^{\phi}(\phi) - \mu_0^{\phi}(\phi)$

$$\mu_a^{\phi}(\phi) = \mathbb{E}(Y \mid A = a, \Phi(X) = \phi)$$

Introduction: Representation-induced confounding bias

Problem formulation: representa-tion-induced confounding bias

- Constraints on the low-dimensional representations include:
  - treatment balancing with a probability metric: $\mathrm{dist}\left[\mathbb{P}(\Phi(X) \mid A=0), \mathbb{P}(\Phi(X) \mid A=1)\right] \approx 0$,
  - invertibility: $\Phi^{-1}(\Phi(X)) \approx X$.
- Such low-dimensional representations can lead to a **representation-induced confounding bias (RICB),** which we want to estimate / bound



Original causal diagram

Transformed causal diagram

Introduction: Task complexity – Related work

Why this is hard?

- Directly estimating RICB is (1) impractical and (2) intractable:

$$\tau^{\phi}(\Phi(x)) = \int_{\mathcal{X}_{\Delta} \times \mathcal{X}_{Y}} \tau^{x}(x)\, \mathbb{P}(X^{\Delta} = x^{\Delta}, X^{y} = x^{y} \mid \Phi(x))\, \mathrm{d}x^{\Delta}\, \mathrm{d}x^{y} \neq \tau^{x}(x)$$

- The partitioning of X is unknown as well $\{X^{\varnothing}, X^{a}, X^{y}, X^{\Delta}\}$

Related work

Table 1: Overview of key representation learning methods for CATE estimation with respect to different constraints, imposed on the representation.

Method	Treatment balancing via: probability metrics	Treatment balancing via: re-weighting	Invertibility
BNN (Johansson et al., 2016)	IPM (MMD)	–	–
TARNet (Shalit et al., 2017; Johansson et al., 2022)	–	–	–
CFR (Shalit et al., 2017; Johansson et al., 2022)	IPM (MMD, WM)	–	–
RCFR (Johansson et al., 2018; 2022)	IPM (MMD, WM)	Learnable weights	–
DACPOL (Atan et al., 2018); CRN (Bica et al., 2020); CT (Melnychuk et al., 2022)	JSD (adversarial learning)	–	–
SITE (Yao et al., 2018)	Middle point distance	–	Local similarity
CFR-ISW (Hassanpour & Greiner, 2019a)	IPM (MMD, WM)	Representation propensity	–
DR-CFR (Hassanpour & Greiner, 2019b); DeR-CFR (Wu et al., 2022)	IPM (MMD, WM)	Representation propensity	–
DKLITE (Zhang et al., 2020)	Counterfactual variance	–	Reconstruction loss
BWCFR (Assaad et al., 2021)	IPM (MMD, WM)	Covariate propensity	–

IPM: integral probability metric; MMD: maximum mean discrepancy; WM: Wasserstein metric; JSD: Jensen-Shannon divergence

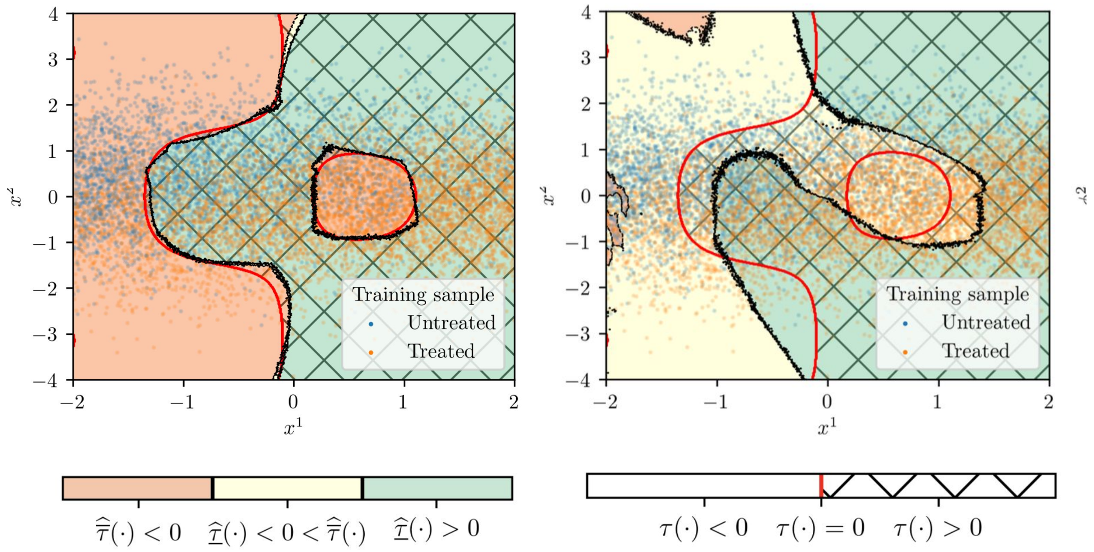
Introduction: Research gap – Our contributions

Research gap

- No work has studied the confounding bias (RICB) in low-dimensional (constrained) representations for CATE estimation

Our contributions

- We formalize the representation-induced confounding bias (RICB)
- We propose a neural framework for estimating bounds based on the **Marginal Sensitivity Model**, which can be seen as a **refutation method** for representation learning CATE estimators
- We show that the estimated bounds are highly effective for the CATE-based decision-making



Representation learning for CATE estimation: Assumptions

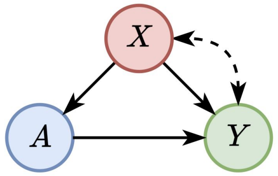
Identifiability assumptions

- Potential outcomes framework (Neuman-Rubin):
  - **(i) Consistency**. If $A = a$ is a treatment for some patient, then $Y = Y[a]$
  - **(ii) Positivity (Overlap).** There is always a non-zero probability of receiving/not receiving any treatment, conditioning on the covariates: $\epsilon > 0, \mathbb{P}(1-\epsilon \geq \pi_a(X) \geq \epsilon) = 1$
  - **(iii) Exchangeability (Ignorability).** Current treatment is independent of the potential outcome, conditioning on the covariates $A \perp\!\!\!\perp Y[a] \mid X \text{ for all } a.$
- Under assumptions (i)–(iii) CATE is identifiable

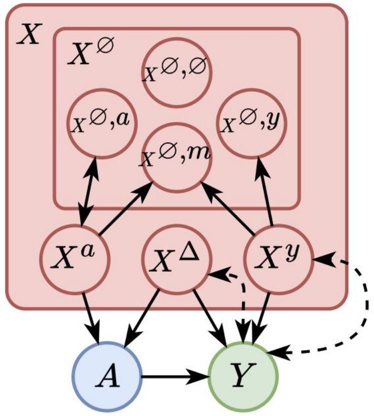
$$\tau^x(x) = \mu_1^x(x) - \mu_0^x(x) \qquad \mu_a^x(x) = \mathbb{E}(Y \mid A = a, X = x)$$

Implicit partitioning assumption

- We assume an implicit partitioning (clustering) of X on $\{X^{\varnothing}, X^a, X^y, X^{\Delta}\}$
  (1) noise
  (2) instruments
  (3) outcome-predictive covariates
  (4) confounders



Original causal diagram

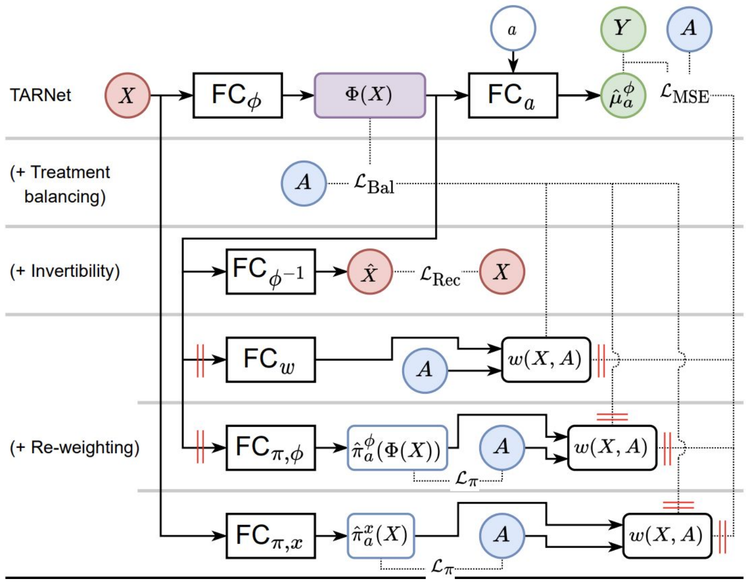


Clustered causal diagram

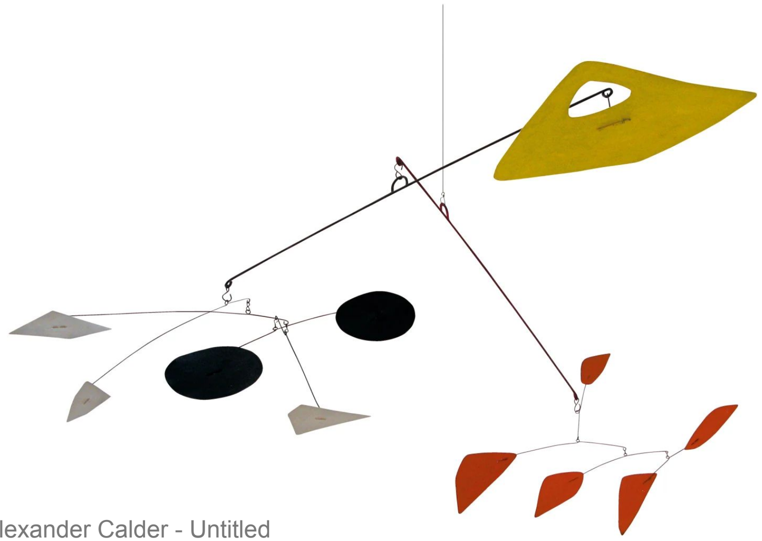
Representation learning for CATE estimation: Methods

- Meta-learners (DR-learner, R-learner, etc.) can obtain the best **asymptotic performance and other properties** by fitting several models (nuisance functions and pseudo-outcome regression)
- Representation-based CATE estimators aim at **best-in-class estimation** with one model, but contain many trade-offs
- In low-sample regime, there is no universally best solution[1]

Meta-learners vs. representation-based CATE estimators



=



Alexander Calder - Untitled

[1] Alicia Curth and Mihaela van der Schaar. Nonparametric estimation of heterogeneous treatment effects: From theory to learning algorithms. In International Conference on Artificial Intelligence and Statistics, 2021.

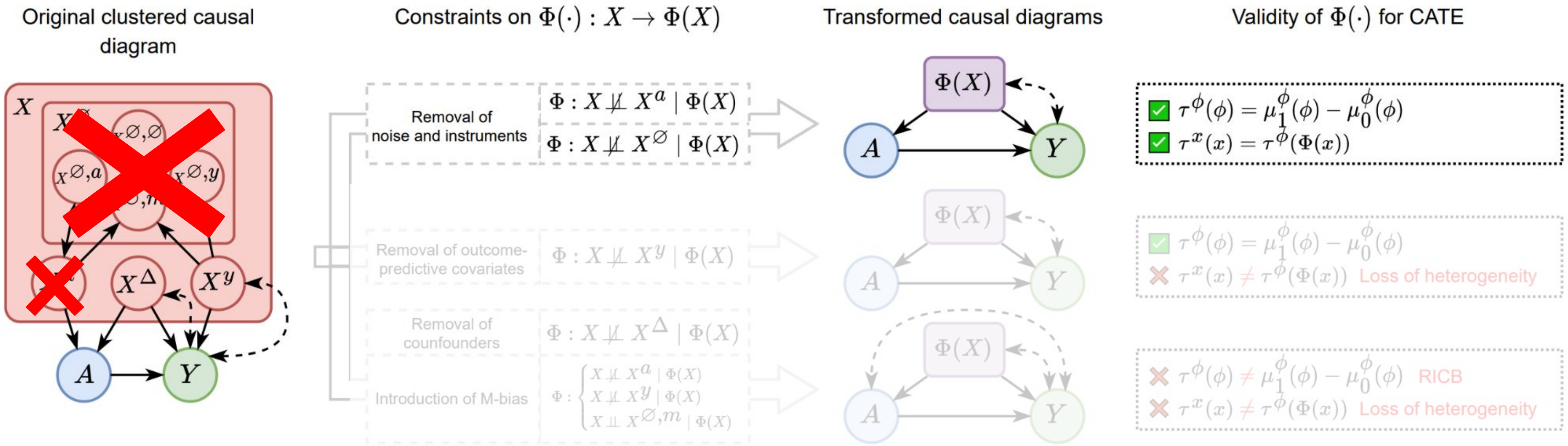
Types of representations: Valid representations

- We call a representation $\Phi(\cdot)$ valid for CATE if it satisfies the following two equalities: $\boxed{\tau^x(x) \overset{(i)}{=} \tau^\phi(\Phi(x)) \quad and \quad \tau^\phi(\phi) \overset{(ii)}{=} \mu_1^\phi(\phi) - \mu_0^\phi(\phi)}$

  with $\mu_a^\phi(\phi) = \mathbb{E}(Y \mid A = a, \Phi(X) = \phi)$

- Examples of valid representations:
  - Invertible representations (still help to reduce the variance when balanced)[1]
  - Removal of noise and instruments (achieved via balancing or lowering d_ϕ)

Valid representations



[1] Fredrik D. Johansson, Uri Shalit, Nathan Kallus, and David Sontag. Generalization bounds and representation learning for estimation of potential outcomes and causal effects. Journal of Machine Learning Research, 23:7489–7538, 2022.

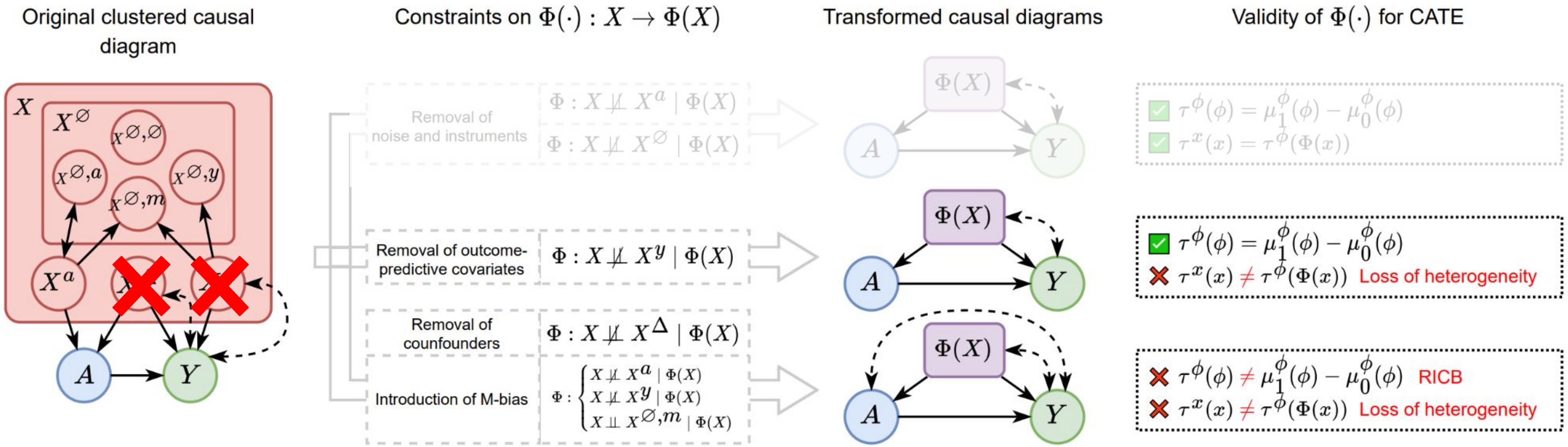
Types of representations: Loss of heterogeneity

(i) Loss of heterogeneity: the treatment effect at the covariate (individual) level is different from the treatment effect at the representation (aggregated) level:

$$\tau^x(x) \neq \tau^\phi(\Phi(x))$$

- Happens whenever some information about X^Δ or X^y is lost in the representation. E.g., propensity score is such a representation.
- Reasons: too low d_ϕ, too large balancing

Invalid representations



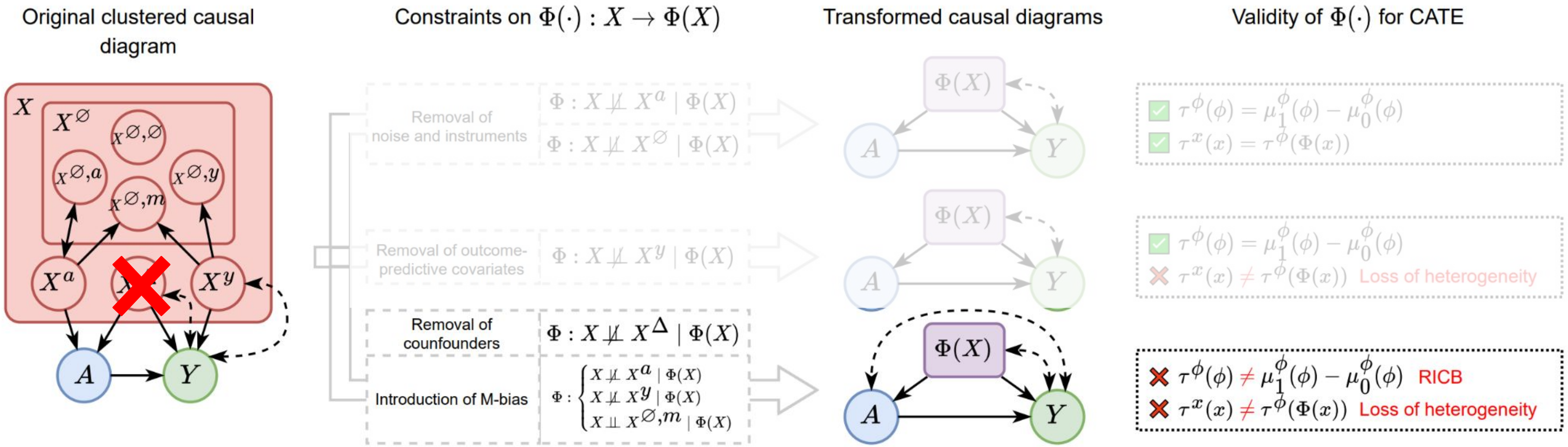
Types of representations: RICB

(i) Representation-induced confounding bias (RICB): CATE wrt. representations is non-identifiable from observational data $\mathbb{P}(\Phi(X), A, Y)$

$$\tau^{\phi}(\phi) \neq \mu_1^{\phi}(\phi) - \mu_0^{\phi}(\phi)$$

- Happens whenever some information about X^{Δ} is lost in the representation or when M-bias is induced (this is rather a theoretic concept)
- Reasons: too low d_{ϕ}, too large balancing

Invalid representations



Types of representations: Takeaways

Takeaways

- The minimal sufficient and valid representation would aim to remove only the information about noise and instruments
- The loss of heterogeneity does not introduce bias but can only make CATE less individualized, namely, suitable only for subgroups
- The RICB automatically implies a loss of heterogeneity => We consider the RICB to be the main problem in representation learning methods for CATE
- RICB is an **infinite-sample confounding bias** (not a low-sample bias), present in the representations

Partial identification of CATE under the RICB: MSM

Marginal sensitivity model

- Our idea is to employ a Marginal sensitivity model (MSM)[1] to perform the partial identification of the CATE (= find bounds on the RICB):

$$\Gamma(\phi)^{-1} \leq \left(\pi_0^{\phi}(\phi)/\pi_1^{\phi}(\phi)\right)\left(\pi_1^{x}(x)/\pi_0^{x}(x)\right) \leq \Gamma(\phi) \quad \text{for all } x \in \mathcal{X} \text{ s.t. } \Phi(x) = \phi,$$

  where the sensitivity parameters can be **estimated from the combined data** $\mathbb{P}(X, \Phi(X), A, Y)$

- Under the sensitivity constraint, the bounds on the RICB are given by

$$\underline{\tau^{\phi}}(\phi) = \underline{\mu_1^{\phi}}(\phi) - \overline{\mu_0^{\phi}}(\phi) \qquad \text{and} \qquad \overline{\tau^{\phi}}(\phi) = \overline{\mu_1^{\phi}}(\phi) - \underline{\mu_0^{\phi}}(\phi)$$

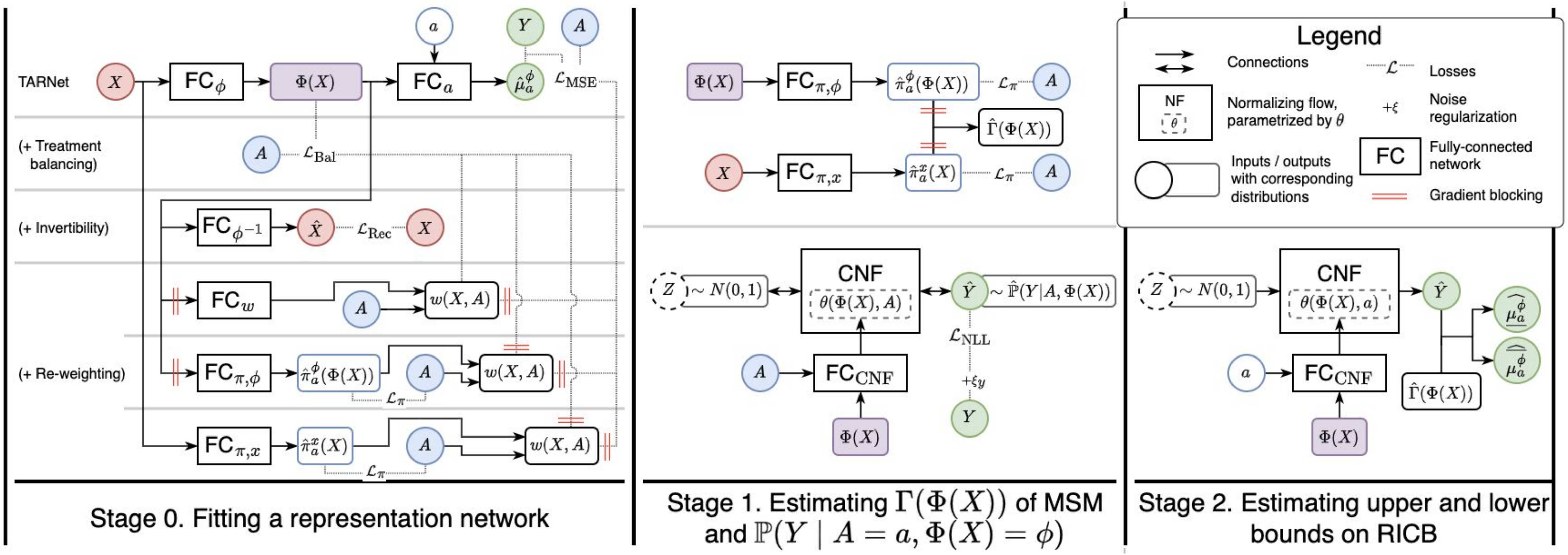
$$\underline{\mu_a^{\phi}}(\phi) = \frac{1}{s_-(a,\phi)} \int_{-\infty}^{\mathbb{F}^{-1}(c_-|a,\phi)} y\,\mathbb{P}(Y = y \mid a, \phi)\,\mathrm{d}y + \frac{1}{s_+(a,\phi)} \int_{\mathbb{F}^{-1}(c_-|a,\phi)}^{+\infty} y\,\mathbb{P}(Y = y \mid a, \phi)\,\mathrm{d}y,$$

$$\overline{\mu_a^{\phi}}(\phi) = \frac{1}{s_+(a,\phi)} \int_{-\infty}^{\mathbb{F}^{-1}(c_+|a,\phi)} y\,\mathbb{P}(Y = y \mid a, \phi)\,\mathrm{d}y + \frac{1}{s_-(a,\phi)} \int_{\mathbb{F}^{-1}(c_+|a,\phi)}^{+\infty} y\,\mathbb{P}(Y = y \mid a, \phi)\,\mathrm{d}y,$$

- The bounds are **valid** wrt. the original CATE and **sharp** wrt. the sensitivity constraint
- The bounds are still conservative, i.e., they do not distinguish instruments from confounders (but to do that we would need the original CATE)
- Yet, other sensitivity models, e.g., outcome sensitivity model, are impractical

[1] Zhiqiang Tan. A distributional approach for causal inference using propensity scores. Journal of the American Statistical Association, 101(476):1619–1637, 2006.

Partial identification of CATE under the RICB: Neural framework



$\hat{\Gamma}(\phi_i)$ is a maximum over all $\hat{\Gamma}(\Phi(x_j))$, where $\Phi(x_j)$ are the representations of the training sample in δ-ball around ϕ_i. δ is the only hyper-parameter

Experiments: Baselines – Evaluation – Datasets

Baselines

Evaluation

Datasets

- We evaluate our refutation framework together with SOTA representation-based CATE estimators: TARNet, BNN, CFR, InvTARNet, RCFR, CFR-ISW, BWCFR
- To compare our bounds with the point estimates, we employ an error rate of the policy (ER):
  - a policy based on the point estimate of the CATE applies a treatment whenever the CATE is positive:

$$\hat{\pi}(\phi) = \mathbb{1}\{\widehat{\tau^{\phi}}(\phi) > 0\}$$

  - a policy based on the bounds on the RICB has three decisions:
    - (1) to treat $\underline{\widehat{\tau^{\phi}}}(\phi) > 0$
    - (2) to do nothing $\overline{\widehat{\tau^{\phi}}}(\phi) < 0$
    - (3) to defer a decision, otherwise
- We used 1 synthetic and 2 semi-synthetic datasets (IHDP100, HC-MNIST)

Experiments: Results

- Our framework achieves clear improvements in the error rate among all the baselines, without deferring too many patients

Results

	ER_{out} (Δ ER_{out})	
d_ϕ	1	2
TARNet	30.79% (−12.89%)	9.82% (−3.73%)
BNN (MMD; $\alpha = 0.1$)	34.32% (−15.41%)	16.15% (−4.19%)
CFR (MMD; $\alpha = 0.1$)	35.01% (−14.27%)	11.92% (−5.54%)
CFR (MMD; $\alpha = 0.5$)	35.79% (−11.43%)	17.89% (−7.27%)
CFR (WM; $\alpha = 1.0$)	34.97% (−14.27%)	10.88% (−7.97%)
CFR (WM; $\alpha = 2.0$)	35.18% (−13.63%)	13.19% (−6.28%)
InvTARNet	29.51% (−0.95%)	5.64% (−0.02%)
RCFR (WM; $\alpha = 1.0$)	33.02% (−3.58%)	8.00% (−4.27%)
CFR-ISW (WM; $\alpha = 1.0$)	35.00% (−9.43%)	7.27% (−1.86%)
BWCFR (WM; $\alpha = 1.0$)	34.97% (−10.02%)	7.44% (−4.57%)

Classical CATE estimators: k-NN: 8.18%; BART: 17.37%; C-Forest: 16.10%

	ER_{out} (Δ ER_{out})		
d_ϕ	7	39	78
TARNet	11.21% (−2.59%)	10.91% (−3.34%)	11.01% (−2.62%)
BNN (MMD; $\alpha = 0.1$)	12.00% (−4.50%)	11.37% (−5.29%)	20.78% (−2.01%)
CFR (MMD; $\alpha = 0.1$)	11.40% (−1.89%)	11.05% (−3.13%)	11.73% (−4.67%)
CFR (MMD; $\alpha = 0.5$)	16.01% (+19.25%)	12.55% (−4.95%)	12.90% (−5.25%)
CFR (WM; $\alpha = 1.0$)	24.55% (−10.42%)	27.87% (−10.18%)	31.19% (−11.53%)
CFR (WM; $\alpha = 2.0$)	31.71% (−10.34%)	30.77% (−7.22%)	31.83% (−11.91%)
InvTARNet	12.18% (−1.29%)	11.38% (−3.98%)	11.55% (−4.34%)
RCFR (WM; $\alpha = 1.0$)	21.51% (−9.17%)	26.97% (−6.17%)	30.14% (−14.26%)
CFR-ISW (WM; $\alpha = 1.0$)	32.64% (−10.32%)	26.66% (−11.30%)	30.02% (−13.31%)
BWCFR (WM; $\alpha = 1.0$)	13.62% (−3.96%)	28.18% (+0.24%)	32.54% (−6.75%)

Lower = better. Improvement over the baseline in green, worsening of the baseline in red
Classical CATE estimators: k-NN: 22.34%; BART: 17.51%; C-Forest: 17.65%

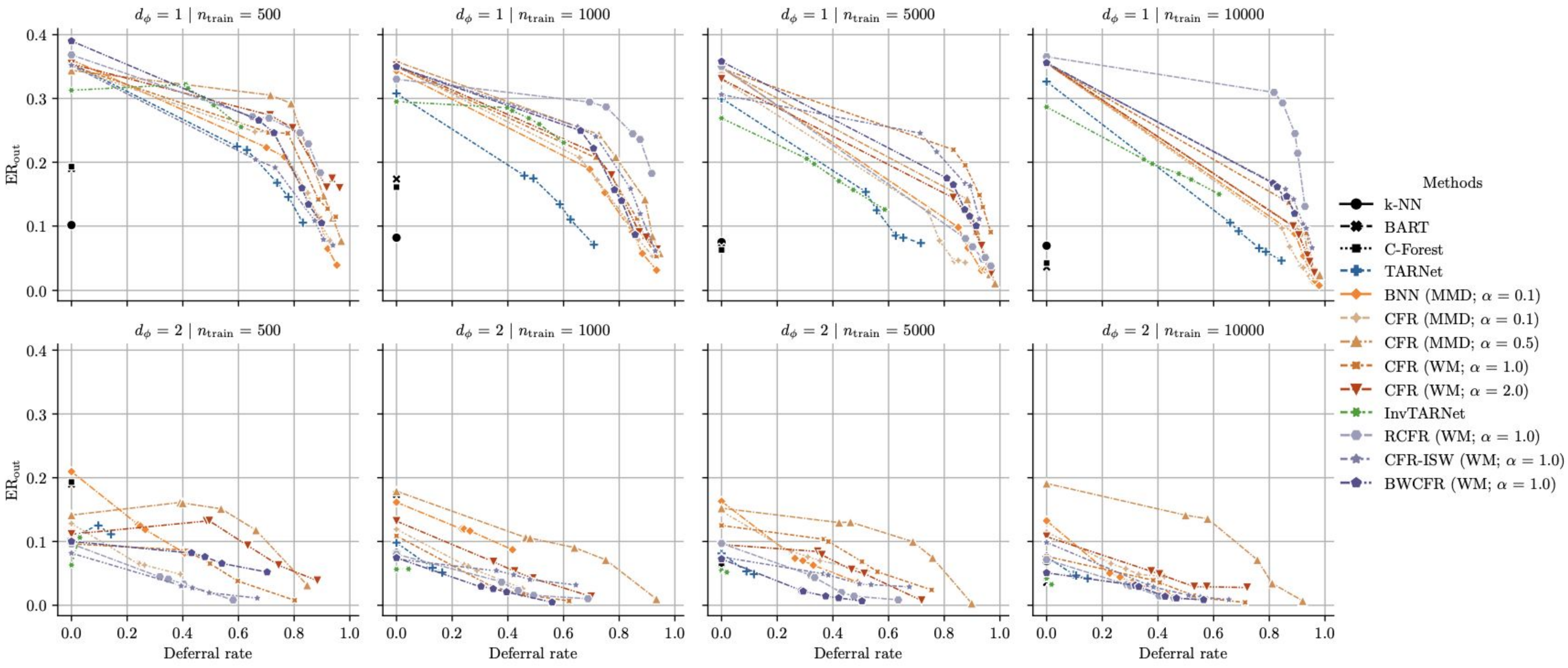
	ER_{out} (Δ ER_{out})				
d_ϕ	5	10	15	20	25
TARNet	3.17% (−2.65%)	2.88% (−2.30%)	3.28% (−2.74%)	3.23% (−2.52%)	2.89% (−2.37%)
BNN (MMD; $\alpha = 0.1$)	2.32% (−1.49%)	2.43% (−1.40%)	2.59% (−2.03%)	2.43% (−1.87%)	2.29% (−1.16%)
CFR (MMD; $\alpha = 0.1$)	1.77% (−0.89%)	2.09% (−1.03%)	2.23% (−1.63%)	1.88% (−0.48%)	2.04% (−1.46%)
CFR (MMD; $\alpha = 0.5$)	2.07% (−1.46%)	2.00% (+3.98%)	2.68% (+1.89%)	2.36% (+6.37%)	2.17% (+3.41%)
CFR (WM; $\alpha = 1.0$)	1.93% (−0.89%)	1.75% (−0.25%)	1.83% (−1.24%)	1.83% (−0.49%)	1.80% (−0.20%)
CFR (WM; $\alpha = 2.0$)	1.97% (−0.04%)	2.17% (−1.49%)	2.05% (−1.21%)	2.08% (−1.29%)	2.09% (−1.36%)
InvTARNet	2.52% (−1.95%)	3.11% (−2.47%)	2.99% (−2.51%)	2.79% (−2.41%)	2.83% (−2.28%)
RCFR (WM; $\alpha = 1.0$)	3.36% (−2.84%)	3.45% (−1.52%)	2.67% (−1.57%)	4.69% (−3.83%)	1.95% (+1.06%)
CFR-ISW (WM; $\alpha = 1.0$)	2.24% (−0.96%)	1.93% (−0.68%)	1.71% (−1.18%)	1.85% (−1.54%)	1.88% (−0.19%)
BWCFR (WM; $\alpha = 1.0$)	3.57% (−1.49%)	3.52% (−2.16%)	3.88% (−1.10%)	3.80% (−2.38%)	4.07% (−1.18%)

Lower = better. Improvement over the baseline in green, worsening of the baseline in red
Classical CATE estimators: k-NN: 7.47%; BART: 5.07%; C-Forest: 6.28%

Experiments: Results

Results

- Ablation study on δ shows, that the bounds remain valid under different values



Conclusion

We studied the validity of representation learning for CATE estimation. The validity may be violated due to low-dimensional representations as these introduce a **representation-induced confounding bias**.

As a remedy, we introduced a novel, **representation-agnostic refutation framework** that estimates bounds on the RICB and thus improves the reliability of their CATEs.

ArXiv Paper:
arxiv.org/abs/2311.11321

Appendix: Johansson et al., 2022

Generalization bounds for the counterfactual risk

Theorem 2. *Given is a sample* $(x_1, t_1, y_1), \ldots, (x_n, t_n, y_n) \overset{i.i.d.}{\sim} p(X, T, Y)$ *with empirical measure* $\hat{p}$. *Assume that ignorability (Assumption 1) holds w.r.t.* X. *Suppose that* Φ *is a twice-differentiable, invertible representation, that* $h_t(\Phi)$ *is a hypothesis on* $\mathcal{Z}$, *and* $f_t = h_t(\Phi(x)) \in \mathcal{H}$. *Let* $\ell_{\Phi,h_t}(z) := \mathbb{E}_Y[L(h_t(z), Y(t)) \mid X = \Psi(z)]$ *where* $L(y, y') = (y - y')^2$. *Further, let* A_Φ *be a constant such that* $\forall z \in \mathcal{Z} : A_\Phi \geq |J_\Psi(z)|$, *where* $J_\Psi(z)$ *is the Jacobian of the representation inverse* Ψ, *and assume that there exists a constant* $B_\Phi > 0$ *such that, with* $C_\Phi := A_\Phi B_\Phi$, $\ell_{\Phi,h_t}/C_\Phi \in \mathcal{L}$, *where* $\mathcal{L}$ *is a reproducing kernel Hilbert space of a kernel,* k *such that* $k(x, x) < \infty$. *Finally, let* w *be a valid re-weighting of* $p_{\Phi,t}$. *Then, with probability at least* $1 - 2\delta$,

$$R_{1-t}(f_t) \leq \hat{R}_t^w(f_t) + C_\Phi \cdot \mathrm{IPM}_{\mathcal{L}}(\hat{p}_{\Phi,1-t}, \hat{p}_{\Phi,t}^w) + V_{p_t}(w, \ell_{f_t}) \frac{\mathcal{C}_{n_t,\delta}^{\mathcal{H}}}{n_t^{3/8}} + \mathcal{D}_{n_0,n_1,\delta}^{\Phi,\mathcal{L}} \left(\frac{1}{\sqrt{n_0}} + \frac{1}{\sqrt{n_1}} \right) + \sigma_{Y(t)}^2 \tag{20}$$

where $\mathcal{C}_{n,\delta}^{\mathcal{H}}$ *is a function of the pseudo-dimension of* $\mathcal{H}$, $\mathcal{D}_{n_0,n_1,\delta}^{\mathcal{L}}$ *is a function of the kernel norm of* $\mathcal{L}$ *(see Lemma 5), both only with logarithmic dependence on* n *and* m, $\sigma_{Y(t)}^2$ *is the expected variance in* $Y(t)$, *and* $V_p(w, \ell_f) = \max\left(\sqrt{\mathbb{E}_p[w^2 \ell_f^2]}, \sqrt{\mathbb{E}_{\hat{p}}[w^2 \ell_f^2]}\right)$. *A similar bound exists where* $\mathcal{L}$ *is the family of functions Lipschitz constant at most 1 and* $\mathrm{IPM}_{\mathcal{L}}$ *the Wasserstein distance, but with worse sample complexity.*

Appendix: Meta-learners

Meta-learners comparison

	Plug-in learners		Two-step learners			
	S-learner	T-learner	RA-learner	IPW-learner	DR-learner	R-learner
a. Large sample size? —Yes→	=	=	=	=	⇑	⇑
b. Need to regularize CATE predictor? —Yes→	×	×	⇑	⇑	⇑	⇑
c. Strong overlap violations (e.g., patients with certain covariates are never treated)? —Yes→	=	=	=	⇓	⇓	=
d. Propensity score is known (e.g., RCT) or easy to estimate? —Yes→	=	=	=	⇑	⇑	⇑
e. Interpretability through a single model for predicting CATE and potential outcomes? —Yes→	⇑	⇑	×	×	×	×

⇑ Beneficial for relative performance

= Has no influence on relative performance

⇓ Detrimental for relative performance

× Not available