Invariant Representation Learning for Treatment Effect Estimation — Supplementary Material

Claudia Shi[1] **Victor Veitch**[2,3] **David M. Blei**[1]

[1]Columbia University
[2]Google Research
[3]The University of Chicago

7 APPENDIX

7.1 PROOFS OF THEOREMS

Lemma 4.1. *Suppose that $\mathbb{E}\left[Y \mid \mathrm{Pa}(Y)=a\right] \neq \mathbb{E}\left[Y \mid \mathrm{Pa}(Y)=a'\right]$ whenever $a \neq a'$. Then a representation Φ is invariant across all valid environments if and only if $\mathbb{E}\left[Y^e|\Phi(T^e, X^e)\right] = \mathbb{E}\left[Y|\mathrm{Pa}(Y)\right]$ for all valid environments.*

Proof. The if direction is immediate.

To establish the only if direction, we first show that Φ must contain at least $\mathrm{Pa}(Y)$, in the sense $\mathbb{E}\left[Y \mid \Phi(X)\right] = \mathbb{E}\left[Y \mid \mathrm{Pa}(Y) \cup Z\right]$ for some set Z. We proceed with proof by contradiction. Suppose that conditioning on Φ is equivalent to conditioning on only $\mathrm{Pa}(Y) \setminus \{P\} \cup Z$, where P is a parent of Y. We now create two environments by setting $P = p$ and $P = p'$. Since P is a parent of Y this follows from the second rule of do calculus (Pearl, 2000),

$$\begin{aligned}&\mathbb{E}\left[Y \mid \mathrm{Pa}(Y) \setminus \{P\} \cup Z; do(P=p)\right] \\ &= \mathbb{E}\left[Y \mid \mathrm{Pa}(Y) \setminus \{P\} \cup Z, P=p\right]\end{aligned}$$

and

$$\begin{aligned}&\mathbb{E}\left[Y \mid \mathrm{Pa}(Y) \setminus \{P\} \cup Z; do(P=p')\right] \\ &= \mathbb{E}\left[Y \mid \mathrm{Pa}(Y) \setminus \{P\} \cup Z, P=p'\right].\end{aligned}$$

The equality $\mathbb{E}\left[Y \mid \mathrm{Pa}(Y) \setminus \{P\} \cup Z, P=p\right] = \mathbb{E}\left[Y \mid \mathrm{Pa}(Y) \setminus \{P\} \cup Z, P=p'\right]$ holds only if P is conditionally independent of Y given $\mathrm{Pa}(Y) \setminus \{P\} \cup Z$. Since P is a parent of Y, by the first assumption of the lemma, the equality does not hold. It follows that $\mathbb{E}\left[Y \mid \mathrm{Pa}(Y) \setminus \{P\} \cup Z\,;\, do(P=p)\right] \neq \mathbb{E}\left[Y \mid \mathrm{Pa}(Y) \setminus \{P\} \cup Z\,;\, do(P=p')\right]$. That is, if conditioning on Φ was equivalent to conditioning on less information than $\mathrm{Pa}(Y) \cup Z$, then Φ would not be invariant across all valid environments.

It remains to show that Φ does not contain any more information than $\mathrm{Pa}(Y)$.

Φ cannot contain any descendants of the outcome. Suppose that Φ depends on some descendant D of Y in the sense that there is at least one environment and $d \neq d'$ where $\mathbb{E}\left[Y \mid \Phi(X \setminus D, D=d)\right] \neq \mathbb{E}\left[Y \mid \Phi(X \setminus D, D=d')\right]$. Then, construct a new environment e by randomly intervening and setting $\mathrm{do}(D=d)$ or $\mathrm{do}(D=d')$, each with probability 0.5. In this new environment, there is no relationship between Y and D. Accordingly, $\mathbb{E}\left[Y^e \mid \Phi(X^e \setminus D^e, D^e=d)\right] = \mathbb{E}\left[Y^e \mid \Phi(X^e \setminus D^e, D^e=d')\right]$. Thus, the conditional expectations are not equal (as functions of d) in the two environments—a contradiction.

Next, we show that, Φ needs not contain the non-parent ancestors A of the outcome, because $\mathbb{E}\left[Y \mid \{A\} \cup \mathrm{Pa}(Y)\right] = \mathbb{E}\left[Y \mid \mathrm{Pa}(Y)\right]$ by the Markov property of the causal graph, where A is any non-ancestor variables. Since Φ contains $\mathrm{Pa}(Y)$, it follows that Φ does not depend on any non-parent ancestor A.

For expository purposes, the proof is done with do calculus (Pearl, 2000) for atomic interventions. If the environments are generated with stochastic interventions, we can use the same proof strategy with σ calculus (Correa and Bareinboim, 2020).

Theorem 4.2. *Let L be a loss function such that the minimizer of the associated risk is a conditional expectation, and let Φ be a representation that elicits a predictor Q^{inv} that is invariant for all valid distributions. Assuming there is no mediators between the treatment and the outcome, then $\psi^e = \mathbb{E}\left[Q^{\mathrm{inv}}(1, X^e) - Q^{\mathrm{inv}}(0, X^e)|T^e = 1\right]$.*

Proof. We assume the technical condition of Lemma 4.1, that $\mathbb{E}\left[Y^e \mid \mathrm{Pa}(Y^e)=a\right] \neq \mathbb{E}\left[Y^e \mid \mathrm{Pa}(Y^e)=a'\right]$ whenever $a \neq a'$. This is without loss of generality because violations of this condition will not lead to different causal

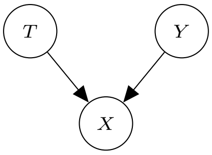

Figure 1: V-structure graph. We denote the bias induced by conditioning on X as V-bias.

effects.

By the assumption on the loss function, the elicited invariant predictor is $\mathbb{E}\left[Y^e \mid \Phi(T^e, X^e)\right]$. Lemma 4.1 shows that $\mathbb{E}\left[Y^e \mid \Phi(T^e, X^e)\right] = \mathbb{E}\left[Y^e \mid \mathrm{Pa}(Y^e)\right]$. We further observe that the non-treatment parents of Y^e are sufficient to block backdoor paths. It follows the ATT can be expressed as the following.

$$\begin{aligned}\psi^e &= \mathbb{E}[\mathbb{E}\left[Y^e \mid T^e = 1, \mathrm{Pa}(Y^e) \setminus \{T^e\}\right] \\ &\quad - \mathbb{E}\left[Y^e \mid T^e = 0, \mathrm{Pa}(Y^e) \setminus \{T^e\}\right)] \mid T^e = 1] \\ &= \mathbb{E}\left[\mathbb{E}\left[Y^e \mid \Phi(1, X^e)\right] - \mathbb{E}\left[Y^e \mid \Phi(0, X^e)\right] \mid T^e = 1\right]\end{aligned}$$

Theorem 4.3. *Suppose $\epsilon \leq P(T^e = 1|X^e) \leq 1 - \epsilon$ with probability 1, then $\epsilon \leq P(T^e = 1|\Phi(X^e)) \leq 1 - \epsilon$ with probability 1.*

Proof. The proof follows directly from Theorem 1 in D'Amour et al., 2020. The intuition is that the richer the covariate set is, the more likely it is to predict the treatment assignment accurately (D'Amour et al., 2020). The covariate representation $\Phi(X^e)$ by definition contains less information than X^e, therefore $\Phi(X^e)$ satisfies overlap if X^e satisfies overlap.

7.2 THE CASE OF COLLIDERS

Consider the DGP with binary variables $\{X, Y, T\}$ illustrated in Figure 1, where X is causally influence by Y and T.

Theorem 7.1. *Let cov denote the covariance between two variables, we define collider bias at $X = c$ as $\Delta(X = c) = cov(T, Y|X = c) - cov(T, Y)$, and collider bias of X as $\Delta(X) = |P(X = 1)\Delta(X = 1) + P(X = 0)\Delta(X = 0)|$. Let $\Phi(T, X)$ be a random variable, where $P(\Phi(T, X) = X) \geq 0.5$. Suppose $P(X = 1) = 0.5$, and $\Delta(X = 1)$ has the same sign as $\Delta(X = 0)$, conditioning on X induce more collider bias than conditioning its coarsening $\Phi(T, X)$:*

$$\Delta(\Phi(T, X)) \leq \Delta(X)$$

Proof. The proof follows corollary 2.1 in Nguyen et al., 2019.

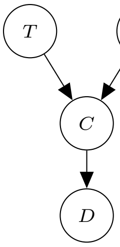

Figure 2: Y-structure graph. We denote the bias induced by conditioning on D as Y-bias.

Corollary 2.1. *We refer to collider bias in the V substructure embedded in the Y structure as 'embedded V-bias' and denote it as $\Delta(C = c)$. For the covariance effect scale, Y-bias $\Delta(D = d)$ relates to embedded V-bias through the following formula:*

$$\begin{aligned}&\Delta(D = d) \\ &= \frac{p(D = d \,|\, C = 1) - p(D = d \,|\, C = 0)}{\{\mathrm{P}(D = d)\}^2} \\ &\cdot \begin{bmatrix} p(D = d \,|\, C = 1)\{\mathrm{P}(C = 1)\}^2 \cdot \Delta(C = 1) - \\ p(D = d \,|\, C = 0)\{\mathrm{P}(C = 0)\}^2 \cdot \Delta(C = 0) \end{bmatrix}.\end{aligned}$$

With the corollary above, let D denote $\Phi(T, X)$, let C denote the collider X in Figure 1. The bias induced by conditioning on D is less than the bias induced by conditioning on C.

$$\begin{aligned}\Delta(D = 1) &= \frac{2\alpha - 1}{0.25}(0.25\alpha \cdot \Delta(C = 1) \\ &\quad - 0.25(1 - \alpha) \cdot \Delta(C = 0)) \\ &= (2\alpha - 1)(\alpha \cdot \Delta(C = 1) \\ &\quad - (1 - \alpha) \cdot \Delta(C = 0)) \\ \Delta(D = 0) &= \frac{1 - 2\alpha}{0.25}(0.25(1 - \alpha) \cdot \Delta(C = 1) \\ &\quad - 0.25\alpha \cdot \Delta(C = 0)) \\ &= (1 - 2\alpha)((1 - \alpha) \cdot \Delta(C = 1) \\ &\quad - \alpha \cdot \Delta(C = 0)) \\ \Delta(C) &= |0.5 \cdot \Delta(C = 0) + 0.5 \cdot \Delta(C = 1)| \\ \Delta(D) &= |0.5 \cdot \Delta(D = 0) + 0.5 \cdot \Delta(D = 1)| \\ \Delta(D) &= |0.5(2\alpha - 1)^2 \cdot \Delta(C = 1) \\ &\quad + 0.5(2\alpha - 1)^2 \cdot \Delta(C = 0)| \\ &\leq \Delta(C)\end{aligned}$$

7.3 THE CASE OF MEDIATORS

In the main paper, we assumed the covariate set X contains no mediators between treatment and outcome. What happens to the interpretation of the learned parameter if the adjustment set contains mediators? Intuitively, NICE retains the direct link between the treatment and the outcome. Specifically, if there are no mediators, the parameter reduces to ATT. If there are mediators but no confounders, the parameter reduces to the Natural Direct Effect (Pearl, 2000). If there are mediators and confounders, the NICE estimand is a non-standard causal target that we call the natural direct effect on the treated (NDET).

Conceptually, NDET describes the expected change in outcome Y for the treated population, induced by changing the value of T, while keeping all mediating factors M, constant at whatever value they would have obtained under $\mathrm{do}(t)$. The main point is that NDET provides answers to questions such as, "does this treatment have a substantial direct effect on this outcome?". Substantively, NDET is the natural direct effect, adjusted for confounders.

Formally, NDET for environment e is

$$\begin{aligned}\psi^e = &\,\mathbb{E}_{M^e \mid T^e=1}[\mathbb{E}\,[Y^e \mid M^e\,;\,\mathrm{do}(T^e=1)] \\ &- \mathbb{E}\,[Y^e \mid M^e\,;\,\mathrm{do}(T^e=0)]\,|T^e=1].\end{aligned} \tag{7.1}$$

With adjustment set W^e, the causal effect can be expressed through a parameter of the observational distribution:

$$\begin{aligned}\psi^e = &\,\mathbb{E}_{M^e,W^e}[\mathbb{E}\,[Y^e \mid T^e=1, M^e, W^e] \\ &- \mathbb{E}\,[Y^e \mid T^e=0, M^e, W^e]\;|\,T^e=1].\end{aligned} \tag{7.2}$$

Importantly, the mediators M^e and the confounders W^e show up in the same way in (7.2). Accordingly, we don't need to know which observed variables are mediators and which are confounders to compute the parameter. Under the NICE procedure, we condition on all parents of Y^e, including possible mediators. Thus, the NICE estimand is the NDET in each environment.

7.4 DETAILS OF THE EXPERIMENTS

7.4.1 Experiment 1

We evaluate the treatment effect estimation of various adjustment schemes using four variants of the data simulations. The variants are generated according to the following: 1) The observed covariates $S(X)$ are scrambled versions of the true covariates X. If scrambled, S is an orthogonal matrix. If not scrambled, S is an identity matrix. 2) In the heteroskedastic setting $\tau \leftarrow 5 + \mathcal{N}(0, e^2)$. In the environment-level homoskedastic setting $\tau \leftarrow 5 + \mathcal{N}(0, 1)$. The results are illustrated in Figure 3, Figure 4, Figure 5, and Figure 6.

To understand why NICE reduces the estimation bias, we measure the weight of the control predictor. The non-causal

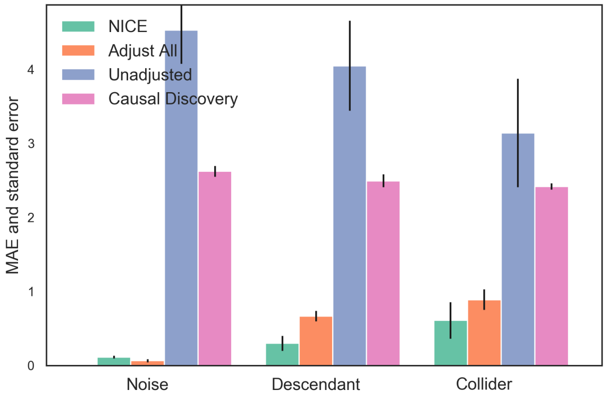

Figure 3: models performance under the scrambled and heteroskedastic setting

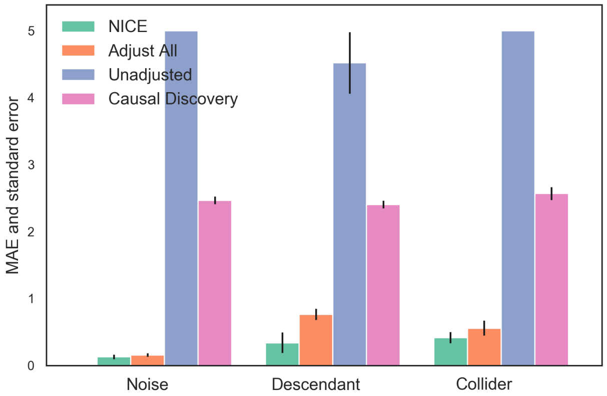

Figure 4: models performance under the scrambled and homoscedastic setting

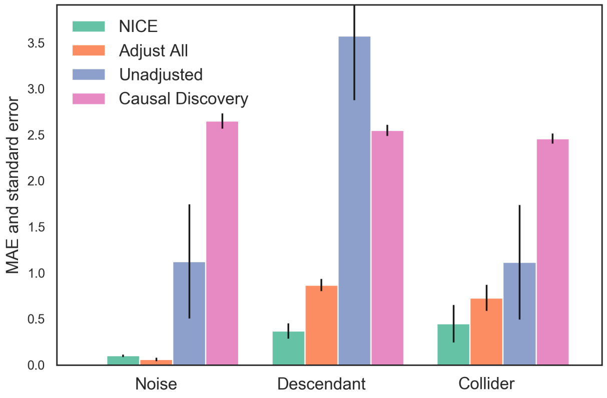

Figure 5: models performance under the unscrambled and heteroskedastic setting

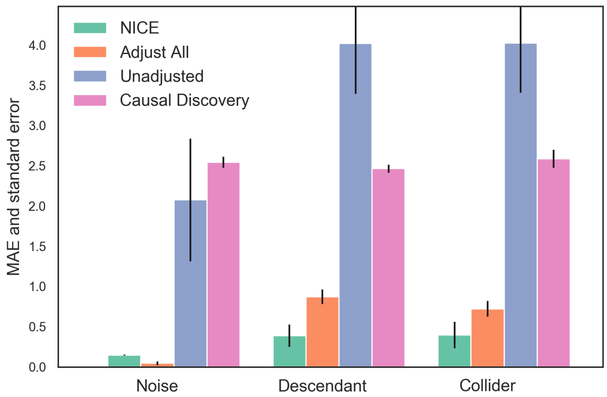

Figure 6: models performance under the unscrambled and homoscedastic setting

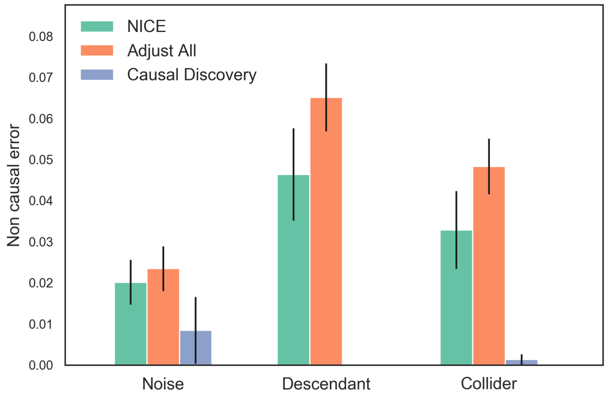

Figure 7: non causal error under the scrambled and heteroskedastic setting

error is measured by the mean square error of the weight on X_2. The results are illustrated Figure 7, Figure 8, Figure 9, and Figure 10.

7.4.2 Experiment 2

We validate NICE for the non-linear case on a benchmark dataset, SpeedDating. SpeedDating was collected to study the gender difference in mate selection (Fisman et al., 2006). The study recruited university students to participate in speed dating, and collected objective and subjective information such as 'undergraduate institution' and 'perceived attractiveness'. It has 8378 entries and 185 covariates. ACIC 2019's simulation samples subsets of the covariates to simulate binary treatment T and binary outcome Y. Specifically, it provides four modified DGPs: Mod1: parametric models; Mod2: complex models; Mod3: parametric models with poor overlap; Mod4: complex models with treatment heterogeneity. Each modification includes three versions: low, med, high, indicating an increasing number of covariates included in the models for T and Y.

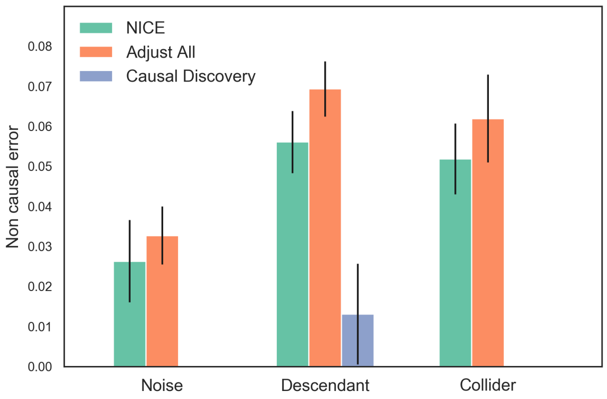

Figure 8: non causal error under the scrambled and homoscedastic setting

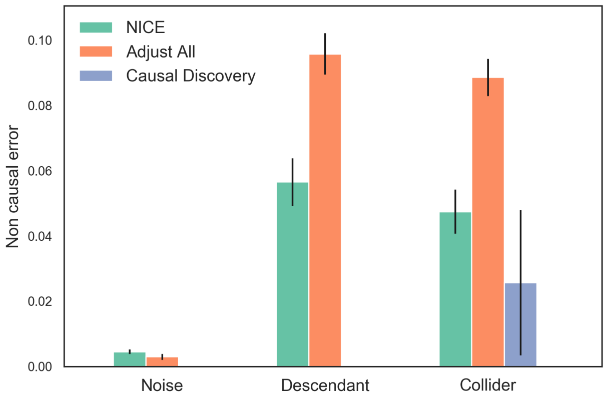

Figure 9: non causal error under the unscrambled and heteroskedastic setting

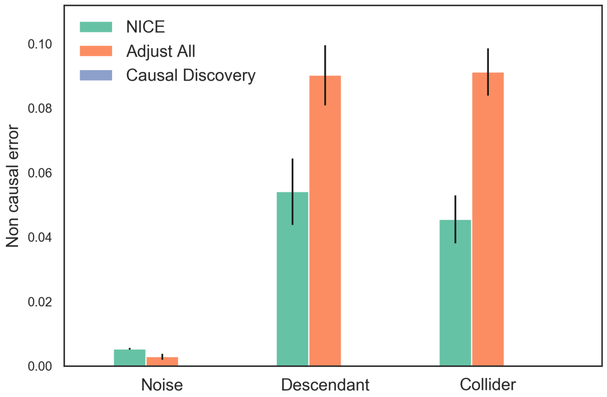

Figure 10: non causal error under the scrambled and homoscedastic setting

We compare the estimation quality of the SATT and CATE over 10 bootstraps. We use two predictors: TARnet and Dragonnet. The main paper report the estimation quality of SATT using TARnet. We now report the estimation quality of CATE and SATT using Dragonnet, as well as CATE using TARnet.

The DGPs are recorded in the R file under SpeedDating folder. The original ACIC DGPs can be downloaded using the following link.

```
drive.google.com/file/d/
1Qqgmb3R9Vt9KTx6t8i_5IbFenylsPfrK/view
```

We made several modifications to the original DGPs. (1) ACIC competitions are usually designed to evaluate model performances using ATE or ATT. The simulation study provided ATE values but not ITE. In the SpeedDating DGPs, the treatment is binary, and the outcome is also binary. To calculate ITE, we take the difference between the propensity of the outcome of the treated predictor and the propensity of the outcome of the control predictor. (2) Unfortunately, ACIC datasets do not come in multiple environments, nor do the covariates include bad controls. To create multiple environments, we draw 6000 samples and select a covariate x that's not the causal parent of Y. We sort the samples based on the the covariates value and divide them into three equal sized environments. For each DGP, we draw 10 bootstrap samples.

Table 3 and Table 4 show the average of the MAE of SATT estimates over three environments. The predictor uses the architecture of Dragonnet. We observe that NICE does not hurt the estimation quality in comparison to adjusting for all the covariates. When there is a strong collider in the adjustment set, NICE reduces collider bias across simulation setups.

Table 1 and Table 5 show the PEHE of the CATE, with TARnet and dragonnet, respectively. We find that NICE improves the CATE estimates in Mod1, Mod2, and Mod3. This is surprising, as we expect NICE to perform equally well as adjusting for all covariates when the adjustment set is valid. To understand this phenomenon better, we examine the CATE estimates across simulation settings.

In Figure 11, Figure 12, Figure 13, and Figure 14 illustrate the estimation quality of CATE under the “med" simulation setting. we compare the ground truth, the CATE estimates using NICE and Adjusting for all covariates across settings. Recall NICE uses an predictor trained with IRMv1 objective, adjusting for all uses a predictor that's trained with ERM objective. In Mod1, Mod2, and Mod3, there are little heterogeneity of the treatment effects. In Mod4, the treatment effects are more spread out and heterogeneous. We observe that in Mod1, Mod2, and Mod3, the ERM predictors produce extreme CATE estimates. In contrast, using an IRMv1 objective, the CATE estimates are less extreme. In Mod4, where the CATE varies drastically, both IRMv1 and ERM predictors were able to capture the heterogeneity.

To examine whether the difference is due to over-fitting, we use two environments as training and one environment as testing. Figure 15 and Figure 16 show the corresponding training and testing accuracy across experiment setups. We observe the ERM predictors have similar training and testing accuracy. This suggests the model is not overfitting in the robust prediction sense. We suspect that the IRMv1 penalty term becomes a regularization term that restrict the model to simpler solutions. Prior work (Janzing, 2019) has shown that regularizing terms in linear regression settings not only help against over-fitting finite data, but sometimes also produce better causal models in the infinite sample settings. There are no known results in the non-linear settings.

Note that the ACIC competitions are *not* designed for evaluating CATE performance. Estimating CATE when the outcomes are binary are difficult, especially given a flexible neural network model. The analysis above is about model specification for estimation, in settings *without* bad controls. In this paper, we consider the problem of causal adjustment. We focus on finding a causal representation that strips out bad controls. Does an invariant predictor produce better causal estimate, even when there are no bad controls? We defer this question to future work.

valid adjustment				ϵ_{pehe}	
	TARnet	Mod1	Mod2	Mod3	Mod4
low	Adjust All	.16 ± .06	.13 ± .05	.22 ± .18	.05 ± .01
	NICE	.06 ± .02	.05 ± .02	.06 ± .02	.05 ± .01
med	Adjust All	.14 ± .06	.15 ± .02	.12 ± .02	.06 ± .02
	NICE	.05 ± .01	.05 ± .01	.07 ± .03	.04 ± .01
high	Adjust All	.13 ± .07	.12 ± .02	.16 ± .09	.06 ± .01
	NICE	.05 ± .01	.05 ± .01	.07 ± .02	.04 ± .01

Table 1

bad controls in adjustment				ϵ_{pehe}	
	TARnet	Mod1	Mod2	Mod3	Mod4
low	Adjust All	.35 ± .12	.17 ± .02	.30 ± .04	.24 ± .04
	NICE	.08 ± .04	.05 ± .01	.07 ± .02	.07 ± .01
med	Adjust All	.26 ± .02	.27 ± .05	.39 ± .07	.13 ± .01
	NICE	.04 ± .01	.05 ± .02	.06 ± .01	.04 ± .01
high	Adjust All	.31 ± .04	.23 ± .06	.31 ± .07	.12 ± .01
	NICE	.07 ± .02	.07 ± .03	.14 ± .07	.06 ± .03

Table 2

valid adjustment				ϵ_{att}	
	Dragonnet	Mod1	Mod2	Mod3	Mod4
low	Adjust All	.09 ± .08	.05 ± .03	.04 ± .04	.03 ± .01
	NICE	.06 ± .03	.03 ± .01	.09 ± .02	.02 ± .02
med	Adjust All	.06 ± .09	.06 ± .05	.11 ± .11	.07 ± .05
	NICE	.07 ± .02	.06 ± .03	.08 ± .03	.04 ± .03
high	Adjust All	.04 ± .04	.05 ± .08	.03 ± .02	.02 ± .02
	NICE	.02 ± .01	.07 ± .02	.06 ± .02	.08 ± .05

Table 3

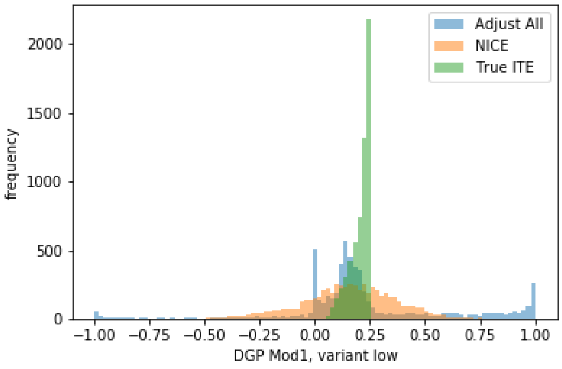

Figure 11: Mod1: parametric models

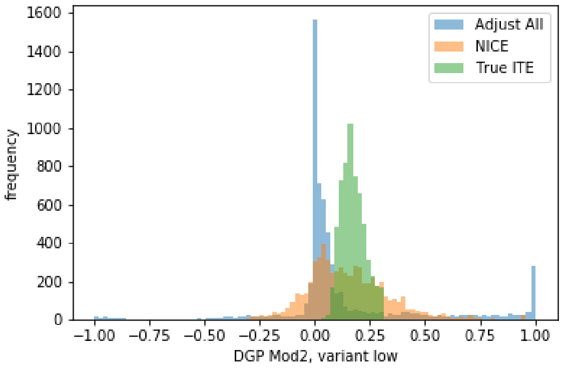

Figure 12: Mod2: complex models

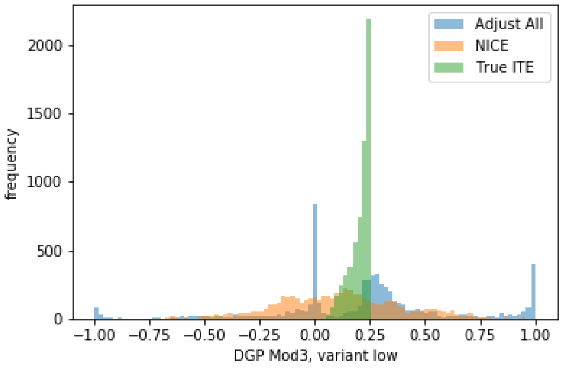

Figure 13: Mod1: parametric models with poor overlap

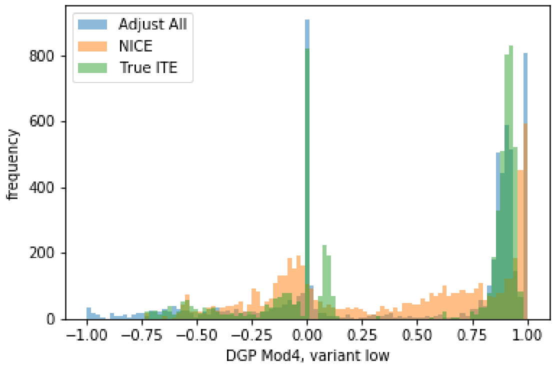

Figure 14: Mod4: complex models with treatment heterogeneity.

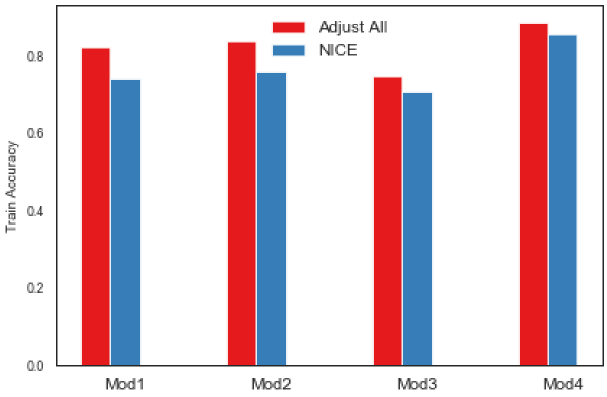

Figure 15: The training accuracy of the predictors

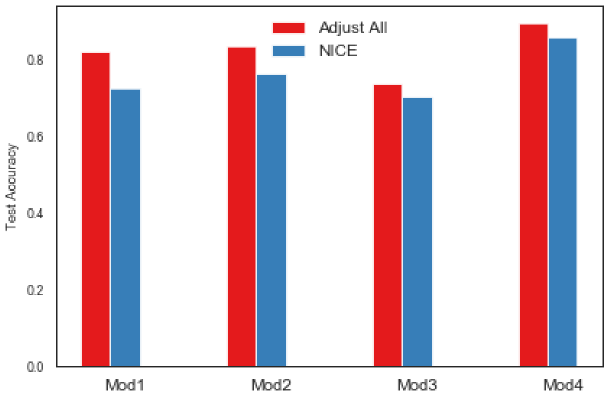

Figure 16: The testing accuracy of the predictors

bad controls in adjustment ϵ_{att}

	Dragonnet	Mod1	Mod2	Mod3	Mod4
low	Adjust All	.31 ± .12	.52 ± .16	.39 ± .06	.75 ± .11
	NICE	.17 ± .10	.03 ± .02	.24 ± .03	.11 ± .06
med	Adjust All	.54 ± .09	.47 ± .15	.57 ± .24	.31 ± .13
	NICE	.13 ± .10	.25 ± .05	.15 ± .08	.05 ± .03
high	Adjust All	.50 ± .06	.58 ± .05	.54 ± .05	.43 ± .13
	NICE	.07 ± .04	.21 ± .06	.17 ± .05	.08 ± .03

Table 4

valid adjustment ϵ_{pehe}

	Dragonnet	Mod1	Mod2	Mod3	Mod4
low	Adjust All	.20 ± .06	.12 ± .02	.19 ± .02	.06 ± .01
	NICE	.06 ± .01	.04 ± .01	.06 ± .01	.04 ± .01
med	Adjust All	.15 ± .06	.16 ± .01	.16 ± .07	.06 ± .01
	NICE	.04 ± .01	.03 ± .01	.07 ± .02	.03 ± .01
high	Adjust All	.12 ± .02	.14 ± .05	.13 ± .02	.06 ± .00
	NICE	.05 ± .01	.06 ± .01	.07 ± .02	.04 ± .02

Table 5

7.4.3 Experiment 3

In the third experiment, we examine the effect of environment variations on NICE's performance. We simulate nonlinear data using the causal graph illustrated in figure 5. We draw three source environments $\{P^{e_1}, P^{e_2}, P^{e_3}\}$, where $e_1 = 0.2, e_2 = 1, e_3 = 5$. In each source environments, we draw 900 samples. Figure 6 reports the average MAE of SATT over 5 simulations.

$$\begin{aligned}
A^e &\leftarrow \mathcal{N}(0, e^2)\\
X^e &\leftarrow A^e \cdot w_{ax^e}\\
X^e_t &\leftarrow X^e_{\{1\ldots12\}}\\
X^e_y &\leftarrow X^e_{\{13\ldots30\}}\\
p^e_t &\leftarrow \text{sigmoid}(f(X^e_t))\\
T^e &\leftarrow Bern(p^e_t)\\
p^e_y &\leftarrow \text{sigmoid}(g(X^e_t, X^e_y, T^e))\\
Y^e &\leftarrow B(n, p^e_y)\\
Z^e &\leftarrow Y^e + T^e + \mathcal{N}(0, 1)
\end{aligned}$$

$f(X^e_t) = X^e_t \cdot w_{xt^e} + h(X^e_t) \cdot w_{xt^{e'}}$, where $h(X^e_t)$ is implemented as

```
[X_t[:, :1] * X_t[:, 1:2],
X_t[:, 1:2] * X_t[:, 2:4],
X_t[:, 2:3] * X_t[:, 3:]
/ np.square(X_t).mean()]
```

$g(X^e_t, X^e_y, T^e) = 1.25 * T^e + X_t \cdot w_{xy^e} + 2 * p^e_t + m(x^e_y) \cdot w_{xy^{e'}}$

Here $m(x^e_y)$ is implemented as

```
[X_y[:, :1] * X_y[:, 4:5],
X_y[:, 1:2] * X_y[:, 3:4],
X_y[:, 1:2] * X_y[:, 2:]
/ np.square(X_y).mean()]
```

The complete data generating code is under

```
diverse_environments/gen_dat.py
```

New environments $P^{e'_1}, P^{e'_2}, P^{e'_3}$ are mixtures of the three source environments $P^{e_1}, P^{e_2}, P^{e_3}$. Respectively, $P^{e'_1}, P^{e'_2}, P^{e'_3}$ draw (p_1, p_2, p_3) proportions from from P^{e_1}, (p_2, p_3, p_1) proportions from from P^{e_2}, and (p_2, p_3, p_1) proportions from from P^{e_3}.The proportions (p_1, p_2, p_3) sum to one.

The mixing proportions we considered are: $(0, 0, 1)$, $(0, 0.1, 0.9)$, $(0, 0.2, 0.8)$, $(0, 0.3, 0.7)$, $(0, 0.4, 0.6)$, $(0, 0.5, 0.5)$,$(0.1, 0.1, 0.8)$,$(0.1, 0.2, 0.7)$,$(0.1, 0.3, 0.6)$, $(0.1, 0.4, 0.5)$,$(0.2, 0.2, 0.6)$, $(0.2, 0.3, 0.5)$,$(0.2, 0.4, 0.4)$, $(0.3, 0.3, 0.4)$.

bad controls in adjustment ϵ_{pehe}

	Dragonnet	Mod1	Mod2	Mod3	Mod4
low	Adjust All	.47 ± .05	.23 ± .05	.37 ± .05	.46 ± .14
	NICE	.16 ± .08	.05 ± .02	.18 ± .04	.05 ± .02
med	Adjust All	.32 ± .05	.37 ± .06	.55 ± .18	.15 ± .02
	NICE	.07 ± .03	.13 ± .04	.11 ± .05	.04 ± .01
high	Adjust All	.38 ± .06	.29 ± .05	.39 ± .05	.14 ± .02
	NICE	.09 ± .03	.11 ± .04	.13 ± .05	.04 ± .01

Table 6