Invariant Representation Learning for Treatment Effect Estimation

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

[1]*Columbia University*
[2]*Google Research*

Abstract

The defining challenge for causal inference from observational data is the presence of 'confounders', covariates that affect both treatment assignment and the outcome. To address this challenge, practitioners collect and adjust for the covariates, hoping that they adequately correct for confounding. However, including every observed covariate in the adjustment runs the risk of including 'bad controls', variables that *induce* bias when they are conditioned on. The problem is that we do not always know which variables in the covariate set are safe to adjust for and which are not. To address this problem, we develop Nearly Invariant Causal Estimation (NICE). NICE uses invariant risk minimization (IRM) [Arj19] to learn a representation of the covariates that, under some assumptions, strips out bad controls but preserves sufficient information to adjust for confounding. Adjusting for the learned representation, rather than the covariates themselves, avoids the induced bias and provides valid causal inferences. We evaluate NICE on both synthetic and semi-synthetic data. When the covariates contain unknown collider variables and other bad controls, NICE performs better than existing methods that adjust for all the covariates.

1 Introduction

Consider the following motivating causal inference problem.

Example 1.1. We want to estimate the effect of sleeping pills on lung disease using electronic health records, collected from multiple hospitals around the world. For each hospital e and patient i, the record contains whether the drug was administered T_i^e, the patient's outcome Y_i^e, and their medical record X_i^e, which includes comprehensive health and socioeconomic information. The different hospitals serve different populations, so the distribution of the records X^e is different across the datasets. But the causal mechanism between sleeping pills T^e and lung disease Y^e is the same.

The data in this example are observational. One challenge to causal inference from observational data is the presence of *confounding variables* that influence both T and Y [RR83; Pea00]. To account for confounding, we adjust for the collected covariates X.

If we know the true causal graph [Pea00] and we observe the nodes in the graph then the adjustment is straightforward. However, we rarely find ourselves in this setting. Causal inference through graphical models assume we observe well-defined "macro"-level causal variables. However, the high dimensional covariates we collect are often actually "micro"-level measurements of those variables [CPE14; CEP16]. In this case, simply picking out a minimal set of covariates as confounders is not possible. In practice, we often do not know what variables in the covariate set to adjust for. The motivating example is inspired by several real-world studies [KLK12; Kao+12; Wei+14; Pat+17], where differences in what variables to adjust for leads to different conclusions about the causal effect.

When our covariates are (many) imperfect measurements of the confounders, naive adjustment only identifies the causal effect if the covariates contain sufficient information to reconstruct the

Preprint. Under review.

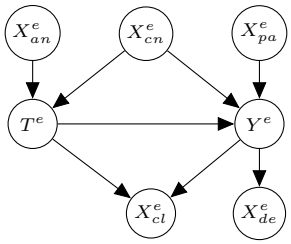


Figure 1: If the composition of X^e is unknown, the treatment effect cannot be identified. (cn=confounder, cl=collider, pa=parents, an=ancestors, de=descendants)

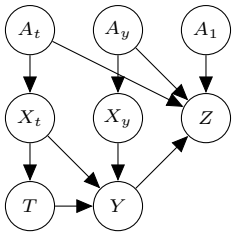


Figure 2: If the graph is observed, we can identify the treatment effect by adjusting for $\{X_t\}$, $\{X_t, A_t\}$, or $\{A_1, A_t, A_y, X_t, X_y\}$

confounders. Accordingly, to ensure that we have adjusted for all confounding variables, we might want to include every covariate in the medical record X. There are two main issues with this kitchen-sink approach. First, simply adjusting for the entire record runs the risk of including "bad controls" [BV07; Pea09; CH20], variables that *induce* bias when they are adjusted for. A health condition caused by lung disease would be a bad control, as it is causally affected by the outcome. Second, the kitchen-sink approach can lead us to include information that is predictive of treatment assignment but not the outcome. If the record is sufficiently rich, this information can lead to near-perfect treatment prediction. This creates an apparent violation of overlap, the requirement that each unit had non-zero probability of receiving treatment [D'A+20].

Problem. We want to do causal adjustment in the setting where it is unknown which covariates are safe to adjust for. In general, this problem is unsolveable without further structure. We are interested in the case where our data is collected from multiple environments. Our main question is: how can we use the multiple environments to achieve valid causal adjustment?

To address this question, we develop nearly invariant causal estimation (NICE), an estimation procedure for causal inference from observational data where the data comes from multiple datasets. The datasets are drawn from distinct environments, corresponding to distinct distributions.

NICE adapts Invariant Risk Minimization(IRM) [Arj+19] for causal adjustment. IRM is a domain generalization learning framework. It uses data from multiple environments to learn an *invariant representation* $\Phi(T, X)$, a function such that the outcome Y and the representation $\Phi(T, X)$ have the same relationship in each environment. Arjovsky et al. [Arj+19] point out that a representation is invariant if and only if it depends only on the causal parents of the outcome Y. For example, an invariant representation of the medical records will isolate the causal parents of lung disease. (⋆) NICE studies the idea that the causal parents of the outcome are sufficient for causal adjustment, satisfy overlap, and excludes bad controls. Thus adjusting for an invariant representation is a safe way to estimate the causal effect. This statement implicitly assumes *no mediators*, variables on the causal path between the treatment and outcome. To keep the exposition simple, we defer a discussion of mediators to § 3.

Contributions. This paper identifies a common setting in causal inference where the classical causal adjustment schemes are not adequate. It develops the idea and algorithmic details of NICE, an estimation procedure that leverages data from multiple environments to do causal inference in such a setting. It articulates the theoretical conditions under which it provides unbiased causal estimates and evaluates the method on synthetic and semi-synthetic causal estimation problems.

2 Nearly Invariant Causal Estimation

We observe multiple datasets. Each dataset is from an environment e, in which we observe a treatment T^e, an outcome Y^e, and covariates X^e. Each environment involves the same causal mechanism between the causal parents of Y^e and Y^e, but otherwise might be different from the others, e.g., in the distribution of X^e. Assume we have enough information in X^e to estimate the causal effect, but we do not know the status of each covariate in the causal graph. A covariate might be an ancestor, confounder, collider, parent, or descendant. Figure 1 shows the graph and defines these terms. Each environment is a data generating process (DGP) with a causal graph and an associated probability distribution P^e. The data from each environment is drawn i.i.d., $\{X^e_i, T^e_i, Y^e_i\} \overset{\text{iid}}{\sim} P^e$. The causal mechanism relating Y to T and X is assumed to be the same in each environment. In the example from the introduction, different hospitals constitute different environments. All the hospitals share the

same causal mechanism for lung disease, but they vary in the population distribution of who they serve, their propensity to prescribe sleeping pills, and other aspects of the distribution.

The goal is to estimate the *average treatment effect on the treated* (ATT) [1] in each environment,

$$\psi^e \triangleq \mathbb{E}\left[Y^e \,|\, \mathrm{do}(T^e = 1), T^e = 1\right] - \mathbb{E}\left[Y^e \,|\, \mathrm{do}(T^e = 0), T^e = 1\right]. \tag{2.1}$$

The ATT is the difference between *intervening* by assigning the treatment and intervening to prevent the treatment, averaged over the people who were actually assigned the treatment. The causal effect for any given individual does not depend on the environment. However, the ATT does depend on the environment because it averages over different populations of individuals.

2.1 Causal estimation

For the moment, consider one environment. In theory, we can estimate the effect by adjusting for the confounding variables that influence both T and Y [RR83]. Let $Z(X)$ be an *admissible* subset of X—it contains no descendants of Y and blocks all "backdoor paths" between Y and T [PP14]. An admissible subset in Figure 1 is any that includes X_{cn} but excludes X_{cl} and X_{de}. Using $Z(X)$, the causal effect can be expressed as a function of the observational distribution,

$$\psi = \mathbb{E}_X\left[\mathbb{E}_Y\left[Y \,|\, T = 1, Z(X)\right] - \mathbb{E}_Y\left[Y \,|\, T = 0, Z(X)\right] \,|\, T = 1\right]. \tag{2.2}$$

We estimate ψ in two stages. First, we fit a model $\hat{Q}$ for the conditional expectation $Q(T, Z(X)) = \mathbb{E}_Y\left[Y \,|\, T, Z(X)\right]$. Then, we estimate the expectation by plugging in the fitted model,

$$\hat{\psi} = \frac{1}{\sum_i t_i} \sum_{i:t_i=1} \left(\hat{Q}(1, Z(X_i)) - \hat{Q}(0, Z(X_i))\right), \tag{2.3}$$

The function $\hat{Q}$ can come from any model that predicts Y from $\{T, Z(X)\}$.

If the causal graph is known, then the admissible set $Z(X)$ can be easily selected and the estimation in (2.2) is straightforward. But here we do not know the status of each covariate—if we inadvertently include bad controls in $Z(X)$ then we will bias the estimate. To solve this problem, we develop a method for learning an *admissible representation* $\Phi(T, X)$, which is learned from datasets from multiple environments. An admissible representation is a function of the full set of covariates but one that captures the confounding factors and excludes the bad controls, i.e., the descendants of the outcome that can induce bias[2]. Given the representation, we estimate the conditional expectations $\mathbb{E}_Y\left[Y \,|\, \Phi(T, X)\right]$ and proceed to estimate the causal effect.

2.2 Invariant Risk Minimization

IRM is a framework for learning predictors that perform well across many environments. We review the main ideas of IRM and then adapt it to causal estimation.

Each environment is a causal structure and probability distribution. Informally, for an environment to be valid, it must preserve the causal mechanism relating the outcome and the other variables.

Definition 2.1 (Valid environment [Arj+19])**.** Consider a causal graph $\mathcal{G}$ and a distribution $P(X, T, Y)$ respecting $\mathcal{G}$. Let $\mathcal{G}_e$ denote the graph under an intervention and $P^e = P(X^e, T^e, Y^e)$ be the distribution induced by the intervention. An intervention is valid with respect to $(\mathcal{G}, P)$ if (i) $\mathbb{E}_{P^e}\left[Y^e | Pa(Y)\right] = \mathbb{E}_P\left[Y | Pa(Y)\right]$, and (ii) $V(Y^e | Pa(Y))$ is finite. An environment is valid with respect to $(\mathcal{G}, P)$ if it can be created by a valid intervention.

Given this definition, a natural notion of an invariant representation is one where the conditional expectation of the outcome is the same regardless of the environment.

Definition 2.2 (Invariant representation)**.** A representation $\Phi(T, X)$ is invariant with respect to environments $\mathcal{E}$ if and only if $\mathbb{E}\left[Y^{e_1} | \Phi(T^{e_1}, X^{e_1})\right] = \mathbb{E}\left[Y^{e_2} | \Phi(T^{e_2}, X^{e_2})\right]$ for all $e_1, e_2 \in \mathcal{E}$.

Arjovsky et al. [Arj+19] recast the problem of finding an invariant representation as one about prediction. In this context, the goal of IRM is to learn a representation such that there is a single

[1] For simple exposition, we focus on the ATT estimation. NICE can also be applied to CATE or ATE estimations.

[2] An admissible representation is analogous to an 'admissible set' [Pea00], where conditional on an admissible set renders a causal effect identifiable.

classifier w that is optimal in all environments. Thus IRM seeks a composition $w \circ \Phi(T^e, X^e)$ that is a good estimate of Y^e in the given set of environments. This estimate is composed of a representation $\Phi(T, X)$ and a classifier w that estimates Y from the representation.

Definition 2.3 (Invariant representation via predictor [Arj+19])**.** A data representation $\Phi : \mathcal{X} \to \mathcal{H}$ elicits an invariant predictor across environments $\mathcal{E}$ if there is a classifier $w : \mathcal{H} \to \mathcal{Y}$ that is simultaneously optimal for all environments. That is,

$$w \in \underset{\bar{w}:\mathcal{H}\to\mathcal{Y}}{\arg\min}\, R^e(\bar{w} \circ \Phi) \quad \text{for all } e \in \mathcal{E}, \tag{2.4}$$

where R^e is the risk of the training objective in environment e.

The invariant representations in Definitions 2.2 and 2.3 align if we choose a loss function for which the minimizer of the associated risk in (2.4) is a conditional expectation. (Examples of such loss functions include squared error and cross entropy.) In this case, we can find an invariant predictor $Q^{\text{inv}} = w \circ \Phi(T^e, X^e) = \mathbb{E}[Y \mid \Phi(T, X)]$ by solving (2.4) for both w and Φ.

However, the general formulation of (2.4) is computationally intractable, so Arjovsky et al. [Arj+19] introduce IRMv1 as a practical alternative.

Definition 2.4 (IRMv1[Arj+19])**.** IRMv1 is:

$$\hat{\Phi} = \underset{\Phi}{\arg\min} \sum_{e\in\mathcal{E}} R^e(1.0 \cdot \Phi) + \lambda \parallel \nabla_{w|w=1.0} R^e(w \cdot \Phi) \parallel^2 . \tag{2.5}$$

Notice here, IRMv1 fixes the classifier to the simplest possible choice: multiplication by the scalar constant $w = 1.0$. The task is then to learn a representation Φ such that $w = 1.0$ is the optimal classifier in all environments. In effect, Φ becomes the invariant predictor, as $Q^{\text{inv}} = 1.0 \cdot \Phi$. The gradient norm penalizes model deviations from the optimal classifier in each environment e, enforcing the invariance. The hyperparameter λ controls the trade-off between invariance and predictive accuracy.[3]

In practice, we parameterize Φ with a neural network that takes T, X as input and outputs a real number. Let l be a loss function such as squared error and cross entropy, n_e be the number of units sampled in environment e. Then, we learn $\hat{\Phi}$ by solving IRMv1 where each environment risk is replaced with the corresponding empirical risk:

$$\hat{R}^e(Q) = \frac{1}{n_e} \sum_i l(y_i^e, Q(t_i^e, x_i^e)). \tag{2.6}$$

Now, $\hat{Q}^{\text{inv}} = 1.0 \cdot \hat{\Phi}$ is an empirical estimate of $\mathbb{E}[Y|\Phi(T, X)]$.

2.3 Nearly Invariant Causal Estimation

We now introduce nearly invariant causal estimation (NICE). NICE is a causal estimation procedure that uses data collected from multiple environments. NICE exploits invariance across the environments to perform causal adjustment without detailed knowledge of which covariates are bad controls.

Informally, the key connection between causality and invariance is that if a representation is invariant across all valid environments then the information in that representation is exactly the information in the causal parents of Y. Since the causal structure relevant to the outcome is invariant across environments, a representation capturing only the causal parents will also be invariant. We can see that $\text{Pa}(Y)$ is the minimal information required for invariance. A representation that is invariant over all valid environments will be minimal; hence, an invariant representation must capture only the parents of Y.

NICE is based on two insights. First, if $\Phi(T, X)$ is invariant over all valid environments, then $\mathbb{E}[Y|T, \text{Pa}(Y) \setminus \{T\}] = \mathbb{E}[Y|\Phi(T, X)]$. Second, $\text{Pa}(Y) \setminus \{T\}$ suffices for causal adjustment. That is, $\text{Pa}(Y) \setminus \{T\}$ blocks any backdoor paths and doesn't include bad controls. Following (2.2),

$$\begin{aligned} \psi &= \mathbb{E}\left[\mathbb{E}[Y \mid T = 1, \text{Pa}(Y) \setminus \{T\}] - \mathbb{E}[Y \mid T = 0, \text{Pa}(Y) \setminus \{T\})\right] \mid T = 1] && (2.7) \\ &= \mathbb{E}\left[\mathbb{E}[Y \mid \Phi(1, X)] - \mathbb{E}[Y \mid \Phi(0, X)] \mid T = 1\right]. && (2.8) \end{aligned}$$

Recall the invariant predictor $Q^{\text{inv}}(T, X) = \mathbb{E}[Y \mid \Phi(T, X)]$. The NICE procedure is

[3]For more details about the intuition of IRMv1, see section 3.1 [Arj+19].

1. Input: multiple datasets $D_e := \{(X_i^e, Y_i^e, T_i^e)\}_{i=1}^{n_e}$.
2. Estimate the invariant predictor $\hat{Q}^{\text{inv}} = 1.0 \cdot \hat{\Phi}$ using IRMv1.
3. Compute $\hat{\psi}^e = \frac{1}{\sum_i t_i^e} \sum_{i:t_i^e=1} \hat{Q}^{\text{inv}}(1, x_i^e) - \hat{Q}^{\text{inv}}(0, x_i^e)$ for each environment e.

Note that $\hat{Q}^{\text{inv}}$ can be any class of predictors. In § 5, we use the TARNet [SJS16] and Dragonnet architectures [SBV19]. We call the procedure 'nearly' invariant as we only ever have access to a limited number of environments, so we cannot be certain that we'll achieve invariance across all possible environments.

3 Justification

We now establish the validity of NICE as a causal estimation procedure. All proofs to the appendix.

First consider the case where we observe data from a sufficiently diverse set of environments that the learned representation is invariant across all possible valid environments. We prove that conditioning on a fully invariant representation is the same as conditioning on the parents of Y.

Lemma 3.1. *Suppose that* $\mathbb{E}[Y \mid \text{Pa}(Y) = a] \neq \mathbb{E}[Y \mid \text{Pa}(Y) = a']$ *whenever* $a \neq a'$. *Then a representation* Φ *is invariant across all valid environments if and only if* $\mathbb{E}[Y^e | \Phi(T^e, X^e)] = \mathbb{E}[Y | \text{Pa}(Y)]$ *for all valid environments.*

Lemma 3.1 helps show that a representation that elicits an invariant predictor suffices for adjustment.

Theorem 3.2. *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 environments. Assuming there is no mediators between the treatment and the outcome, then* $\psi^e = \mathbb{E}\left[Q^{\text{inv}}(1, X^e) - Q^{\text{inv}}(0, X^e) | T^e = 1\right]$.

Theorem 3.2 shows that the NICE estimand is equal to the ATT as long as invariance over the observed environments guarantees invariance on any valid environment (total invariance). Invariance across a limited set of diverse environments may already suffice. Assuming a linear DGP, Arjovsky et al. [Arj+19] establish sufficient conditions on the number and diversity of the training environments such that the learned representation generalizes to all valid environments. In the non-linear case, there are no known sufficiency results. However, Arjovsky et al. [Arj+19] give empirical evidence that access to even a few environments may suffice.

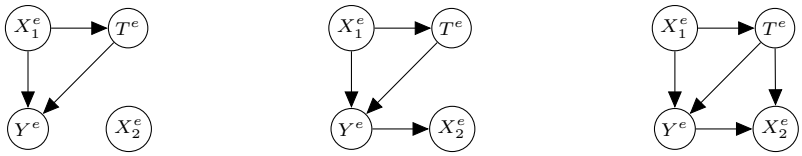
In addition to identifiability, non-parametric estimation of treatment effects with finite data, i.e., (2.3), requires 'positivity' or 'overlap' – both treatment and non-treatment have a non-zero probability for all levels of the confounders [RR83; Imb04]. Let $\Phi(X^e)$ be the covariate representation, i.e., $\Phi(X^e) = \{\Phi(T^e = 1, X^e), \Phi(T^e = 0, X^e)\}$, in the following theorem, we establish that if the covariate set X is sufficient for overlap, then $\Phi(X^e)$ is sufficient for overlap.

Theorem 3.3. *Suppose* $0 < P(T^e = 1 | X^e) < 1$ *with probability 1, then* $0 < P(T^e = 1 | \Phi(X^e)) < 1$ *with probability 1.*

The intuition is that the richer the covariate set is, the more likely it is to predict the treatment assignment accurately [D'A+20]. 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.

Even when total invariance is not achieved, NICE may still improve the estimation quality when there are possible colliders in the adjustment set. Invariance across a subset of environments should remove at least some (if not all) collider dependence. Intuitively, conditioning on a subset of collider information should reduce bias in the resulting estimate. We show a sample theorem in the appendix that this intuition holds for at least one illustrative causal structure. A fully general statement remains open.

The case of mediators. So far, we assumed no mediators between T and Y. What happens to the interpretation of the learned parameter if the adjustment set contains mediators? Intuitively, NICE captures the information in the direct link between T and Y. Concretely, 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 [Pea00]. If there are mediators and confounders, we define the parameters as the natural direct effect on the treated (NDET). We continue the discussion of mediators in the Appendix.



(a) Noise (b) Descendant (c) Collider

Figure 3: In (a) X_2 has no relation with other variables. Conditioning on X_2 doesn't induce additional bias asymptotically. In (b) and (c), X_2 is downstream of Y. Conditioning on X_2 induces bias.

4 Related Work

To address the motivating causal adjustment question, one possible solution is to restrict to all the pre-treatment covariates [Ros02; Rub09]. However, three issues arise. 1) If the record is sufficiently rich, the record might predict the treatment perfectly [D'A+20]. This violates overlap, a requirement for non-parametric estimation [RR83]. 2) In some cases, including non-confounders that are predictive of Y may lower the variance of the causal estimates [RJ91; ML09]. 3) Pre-treatment variables do not necessarily mean pre-treatment measurements. "Micro"-level measurements [CPE14; CEP16] collected after the treatment may still be reflective of the pre-treatment variables. In the motivating example, a measurement, such as "BMI", that was collected days after the treatment may still be a good measurement of the pre-treatment variable of interest — health status.

Another possible solution is to select the adjustment set through causal discovery [MM+99; Spi+00; Shi+06; GZS19] or variable selection [SE17; PBM16; HDPM18]. Causal discovery methods aim to recover causal relationships by analyzing purely observational data. Peters et al. [PBM16] and Heinze-Deml et al. [HDPM18] leverage the multiple environments setup to find a set of covariates that are causal parents of the outcome. However, these methods are designed to discover robust causal relationships, not finding an adjustment set for the downstream inference. The 'robustness' feature may be desirable for discovering important features; it is less desirable for downstream inference, as it may throw out features relevant to the inference. As an illustration, Zhao et al. [Zha+16] showed that slight perturbations on model assumptions might lead to a model's poor performance.

NICE uses the principle of invariance to solve an estimation problem. The principle of invariance is: if a relationship between X and Y is causal, then it is invariant to perturbations that changes the distributions of X and Y. Conversely, if a relationship is invariant to many different perturbations, it must be causal [Haa43; Büh18]. This principle inspired a line of causality-based domain adaptation work.

Bareinboim and Pearl [BP14] provide theory for identification across different domains. Rojas-Carulla et al. [RC+18] and Magliacane et al. [Mag+18] apply the idea for causal transfer learning. Peters et al. [PBM16] and Heinze-Deml et al. [HDPM18] apply this principle for causal variable selection from multiple environments under the linear and non-linear settings. Zhang et al. [Zha+20] recast the problem of domain adaptation as a problem of Bayesian inference on the graphical models. Arjovsky et al. [Arj+19] advocate a new generalizable statistical learning principle that is based on the invariant principle. NICE is complementary as it studies the idea of using domain adaptation methods for causal estimations. In particular, we focus on the application of IRM for treatement effect estimation.

5 Empirical Studies

We perform three experiments to assess the estimation quality of NICE. We find that: i) When total invariance is guaranteed, NICE captures the information relevant for the adjustment. ii) When total invariance is not guaranteed, near invariance still improves the estimation quality. iii) NICE reduces biases when there are bad controls in the adjustment set. iv) NICE improves finite sample estimation quality than alternative adjustment schemes suggested in the introduction and related work.

Setup. Ground truth for counterfactuals is not accessible in real-world data. Therefore, empirical evaluation of causal estimation methods generally relies on synthetic or semi-synthetic datasets. Recovering the parents of Y is shown to be achievable with a linear DGP and well-specified models [Arj+19]. Therefore, in the first experiment, we consider three variants of a simple linear DGP, illustrated in Figure 3, as a minimal proof of idea. In the second experiment, we use SpeedDating, a benchmark dataset simulated as part of the 2019 Atlantic causal inference competition (ACIC) [Gru+]. In the third experiment, we construct a non-linear DGP, illustrated in Figure 2, to assess the finite sample performance of NICE in comparison to alternative adjustment schemes.

Evaluation metrics. We consider two estimands: the sample average treatment effect on the treated (SATT), $\psi_s = \frac{1}{\sum_i t_i} \sum_{i:t_i=1} (Q(1, Z(x_i)) - Q(0, Z(x_i)))$ and the conditional average treatment effect (CATE), $\tau(x_i) = Q(1, Z(x_i)) - Q(0, Z(x_i))$ [Imb04]. For the SATT, the evaluation metric is the mean absolute error (MAE), $\epsilon_{att} = |\hat{\psi}_s - \psi_s|$. For the CATE, the metric is the Precision in Estimation of Heterogeneous Effect (PEHE) $\epsilon_{\text{PEHE}} = \frac{1}{n} \sum_0^n (\hat{\tau}(x_i) - \tau(x_i))^2$ [Hil11]. PEHE reflects the ability to capture individual variation in treatment effects.

5.1 Linear synthetic dataset

We simulate data with the three causal graphs in Figure 3. With a slight abuse of notation, each intervention e generates a new environment e with interventional distribution $P(X^e, T^e, Y^e)$. T^e is the binary treatment and Y^e is the outcome. X^e is a 10-dimensional covariate set that differs across DGPs. $X^e = (X_1^e, X_2^e)$, where X_1^e is a five-dimensional confounder. X_2^e is either noise, a descendant, or a collider in each DGP.

For evaluation, following [Arj+19], we create three environments $\mathcal{E} = \{0.2, 2, 5\}$. We ran 10 simulations. In each simulation we draw 1000 samples from each environment. We compare against Invariant Causal Prediction (ICP) [PBM16], which selects a subset of the covariates as causal parents, and use the subset for causal adjustment. We also compare against linear regression with separate regressors for the treated and the control population (OLS-2). We examine the models' performance under two types of variations: 1) whether the observed covariates are scrambled versions of the true covariates. 2) whether the treatment effects are heteroskedastic across environments. The result in Figure 4 is under the scrambled and heteroskedastic variant. The results of the other variants and details of each dgp are in the appendix.

Analysis. Figure 4 shows that when the model is well-specified—simulation setting (a)— NICE performs well relative to the OLS-2. When the covariate set includes bad controls that are closely related to the outcome, OLS-2 relies on the spurious correlations but NICE discards them. We find that ICP often recovers none or a very limited amount of the covariates. We believe that this is because 1) the amount of noise in the DGP is non-trivial and 2) in some settings the observed covariates are scrambled versions of the true covariates. The result suggests that while ICP is a robust causal discovery method, it should not be used for downstream estimation.

5.2 Semi-synthetic dataset

We validate NICE for the non-linear case on a benchmark dataset, SpeedDating. SpeedDating was collected to study the gender difference in mate selection [Fis+06]. 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 the treatment T and 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.

Table 1: NICE performs well relative to the baselines if the adjustment set does not contain bad controls. The left table reports MAE and bootstrap standard deviation of the SATT estimation. The model is trained and evaluated on all three environments. The right table reports PEHE and bootstrap standard deviation of the out-of-distribution CATE estimation. The model is trained on two environments and evaluated on the third.

	Within-sample				**out-of-sample**			
	ϵ_{att}				$\sqrt{\epsilon_{\text{PEHE}}}$			
	MOD1	MOD2	MOD3	MOD4	MOD1	MOD2	MOD3	MOD4
TARNet	$.08 \pm .06$	$.14 \pm .09$	$.08 \pm .05$	$.01 \pm .01$	$.15 \pm .03$	$.14 \pm .06$	$.21 \pm .02$	$.04 \pm .00$
+NICE	$.05 \pm .05$	$.04 \pm .02$	$.07 \pm .02$	$.03 \pm .05$	$.07 \pm .02$	$.08 \pm .08$	$.09 \pm .01$	$.07 \pm .12$
Dragon	$.09 \pm .06$	$.14 \pm .09$	$.08 \pm .05$	$.05 \pm .02$	$.21 \pm .07$	$.13 \pm .06$	$.25 \pm .04$	$.05 \pm .02$
+NICE	$.06 \pm .06$	$.04 \pm .02$	$.07 \pm .02$	$.02 \pm .02$	$.06 \pm .02$	$.06 \pm .04$	$.08 \pm .02$	$.05 \pm .02$

We compare NICE against two neural network models similar to the structure of TARNet [SJS16] and Dragonnet [SBV19]. TARNet is a two-headed model with a shared representation $Z(X) \in R^p$,

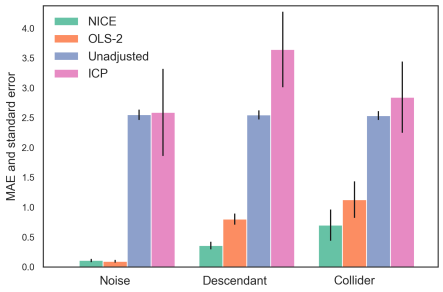


Figure 4: NICE yields better estimates if X contains bad controls, performs equally well to OLS-2 otherwise. When ICP returns an empty set, estimated causal effect as zero. The figure reports MAE and standard error of the SATT over 10 simulations.

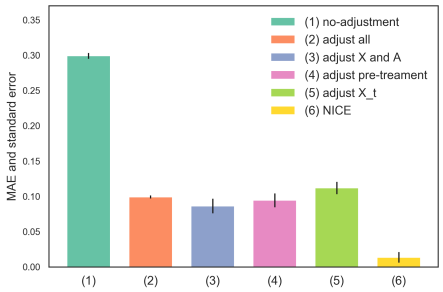


Figure 5: NICE yields better estimates than adjusting for the minimal adjustment set or the pre-treatment variables, performs comparably to adjusting for X and A. The figure reports MAE and standard deviation of the SATT over 10 simulations.

and two heads for the treated and control representation. The network has 4 layers for the shared representation and 3 layers for each expected outcome head. The hidden layer size is 250 for the shared representation layers and 100 for the expected outcome layers. We use Adam [KB14] as the optimizer, set learning rate as 0.001, and an l2 regularization rate of 0.0001. For Dragonnet, an additional 'treatment head' for the treatment prediction is added.

Analysis. We intervene on a variable to generate three environments and draw 2000 samples from each environment. We compare the estimation quality of the within-sample SATT and out-of-sample CATE over 10 bootstraps in Table 1. Since the DGPs do not include bad controls in the covariate set, NICE performs as well as existing methods, though factors such as optimization difficulty may lead to variation.

To examine whether NICE helps reduce collider bias, we simulated 20 copies of a collider: $X^e_{co} = T^e + Y^e + \mathcal{N}(0, e^2)$, where $e \in \{0.01, 0.2, 1\}$ and include it in the covariate set. As shown in Table 2, NICE reduces collider bias across simulation setups. However, we also observe that while it reduces the collider bias, it does not eliminate it completely. One possible reason is that the predictor is not optimal[4].

Table 2: NICE reduces estimation bias in the presence of colliders. The table reports the MAE and bootstrap standard deviation of SATT. The model is trained and evaluated on three environments.

Within-sample

		ϵ_{att}			
		MOD1	MOD2	MOD3	MOD4
low	TARNet	$.31 \pm .11$	$.39 \pm .14$	$.32 \pm .07$	$.44 \pm .07$
	+NICE	$.10 \pm .03$	$.08 \pm .07$	$.15 \pm .04$	$.14 \pm .08$
med	TARNet	$.37 \pm .12$	$.29 \pm .13$	$.37 \pm .12$	$.25 \pm .08$
	+NICE	$.09 \pm .3$	$.17 \pm .10$	$.09 \pm .03$	$.05 \pm .03$
high	TARNet	$.28 \pm .09$	$.35 \pm .09$	$.49 \pm .16$	$.33 \pm .18$
	+NICE	$.09 \pm .06$	$.15 \pm .08$	$.10 \pm .08$	$.14 \pm .07$

5.3 Finite sample performance

In the previous experiments, we consider the setting where the status of any covariate is unknown. In this experiment, in addition to the motivating setting, we consider the case the causal graph and the pre-treatment variables are known. We simulate non-linear data according to Figure 2. The details of the simulation is in the supplementary material. we consider the following adjustment schemes: (1) no adjustment; (2) adjusting for all variables; (3) adjusting for A:$\{A_1, A_t, A_y\}$ and X:$\{X_t, X_y\}$, variables that are safe to adjust for; (4) adjusting for the pre-treatment variables $\{X_t, A_t\}$; (5) adjusting for the minimal adjustment set, X_t; (6) adjustment via NICE.

Analysis. As shown in figure 5, NICE yields better estimates than adjusting for X_t, the minimal adjustment, and adjusting for all pre-treatment variables. NICE performs comparably to adjusting for variables that are safe to adjust for (all X and A). While adjustment scheme (3), (4), (5), and (6) are all valid, the result suggests including the covariates that are predictive of the outcome improve the estimation quality. In at least one case, NICE may still improve the finite sample estimation quality over alternative adjustment schemes.

[4]Comparison of Dragonnet with Dragonnet + NICE is in the Appendix.

References

[Arj+19] M. Arjovsky, L. Bottou, I. Gulrajani, and D. Lopez-Paz. "Invariant risk minimization". In: *arXiv preprint arXiv:1907.02893* (2019).

[BP14] E. Bareinboim and J. Pearl. "Transportability from multiple environments with limited experiments: completeness results". In: *Advances in neural information processing systems*. 2014.

[BV07] J. Bhattacharya and W. B. Vogt. *Do instrumental variables belong in propensity scores?* Tech. rep. National Bureau of Economic Research, 2007.

[Büh18] P. Bühlmann. "Invariance, causality and robustness". In: *arXiv preprint arXiv:1812.08233* (2018).

[CEP16] K. Chalupka, F. Eberhardt, and P. Perona. "Multi-level cause-effect systems". In: *Artificial Intelligence and Statistics*. 2016.

[CPE14] K. Chalupka, P. Perona, and F. Eberhardt. "Visual causal feature learning". In: *arXiv preprint arXiv:1412.2309* (2014).

[CH20] C. Cinelli and C. Hazlett. "Making sense of sensitivity: extending omitted variable bias". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 1 (2020).

[D'A+20] A. D'Amour, P. Ding, A. Feller, L. Lei, and J. Sekhon. "Overlap in observational studies with high-dimensional covariates". In: *Journal of Econometrics* (2020).

[Fis+06] R. Fisman, S. S. Iyengar, E. Kamenica, and I. Simonson. "Gender differences in mate selection: evidence from a speed dating experiment". In: *The Quarterly Journal of Economics* 2 (2006).

[GZS19] C. Glymour, K. Zhang, and P. Spirtes. "Review of causal discovery methods based on graphical models". In: *Frontiers in genetics* (2019).

[Gru+] S. Gruber, G. Lefebvre, A. PichÃľ, and T. Schuster. *Data Challenge*. URL: `https://sites.google.com/view/acic2019datachallenge`.

[Haa43] T. Haavelmo. "The statistical implications of a system of simultaneous equations". In: *Econometrica, Journal of the Econometric Society* (1943).

[HDPM18] C. Heinze-Deml, J. Peters, and N. Meinshausen. "Invariant causal prediction for nonlinear models". In: *Journal of Causal Inference* 2 (2018).

[Hil11] J. Hill. "Bayesian nonparametric modeling for causal inference". In: *Journal of Computational and Graphical Statistics* (2011).

[Imb04] G. W. Imbens. "Nonparametric estimation of average treatment effects under exogeneity: a review". In: *Review of Economics and statistics* 1 (2004).

[Kao+12] C.-H. Kao, L.-M. Sun, J.-A. Liang, S.-N. Chang, F.-C. Sung, and C.-H. Muo. "Relationship of zolpidem and cancer risk: a taiwanese population-based cohort study". In: *Mayo clinic proceedings*. 5. Elsevier. 2012.

[KB14] D. P. Kingma and J. Ba. "Adam: a method for stochastic optimization". In: *arXiv preprint arXiv:1412.6980* (2014).

[KLK12] D. F. Kripke, R. D. Langer, and L. E. Kline. "Hypnotics' association with mortality or cancer: a matched cohort study". In: *BMJ open* 1 (2012).

[Mag+18] S. Magliacane, T. van Ommen, T. Claassen, S. Bongers, P. Versteeg, and J. M. Mooij. "Domain adaptation by using causal inference to predict invariant conditional distributions". In: *Advances in Neural Information Processing Systems*. 2018.

[ML09] K. L. Moore and M. J. van der Laan. "Covariate adjustment in randomized trials with binary outcomes: targeted maximum likelihood estimation". In: *Statistics in medicine* 1 (2009).

[MM+99] K. Murphy, S. Mian, et al. *Modelling gene expression data using dynamic Bayesian networks*. Tech. rep. Citeseer, 1999.

[NDO19] T. Q. Nguyen, A. Dafoe, and E. L. Ogburn. "The magnitude and direction of collider bias for binary variables". In: *Epidemiologic Methods* 1 (2019).

[Pat+17] E. Patorno, R. J. Glynn, R. Levin, M. P. Lee, and K. F. Huybrechts. "Benzodiazepines and risk of all cause mortality in adults: cohort study". In: *bmj* (2017).

[Pea00] J. Pearl. *Causality: models, reasoning and inference*. 2000.

[Pea09] J. Pearl. *Causality*. 2009.

[PP14] J. Pearl and A. Paz. "Confounding equivalence in causal inference". In: *Journal of Causal Inference J. Causal Infer.* 1 (2014).

[PBM16] J. Peters, P. Bühlmann, and N. Meinshausen. "Causal inference by using invariant prediction: identification and confidence intervals". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* (2016).

[RJ91] L. D. Robinson and N. P. Jewell. "Some surprising results about covariate adjustment in logistic regression models". In: *International Statistical Review/Revue Internationale de Statistique* (1991).

[RC+18] M. Rojas-Carulla, B. Schölkopf, R. Turner, and J. Peters. "Invariant models for causal transfer learning". In: *The Journal of Machine Learning Research* 1 (2018).

[Ros02] P. R. Rosenbaum. "Overt bias in observational studies". In: *Observational studies*. 2002.

[RR83] P. R. Rosenbaum and D. B. Rubin. "The central role of the propensity score in observational studies for causal effects". In: *Biometrika* (1983).

[Rub09] D. B. Rubin. "Should observational studies be designed to allow lack of balance in covariate distributions across treatment groups?" In: *Statistics in Medicine* 9 (2009).

[SJS16] U. Shalit, F. D. Johansson, and D. Sontag. "Estimating individual treatment effect: generalization bounds and algorithms". In: *arXiv e-prints arXiv:1606.03976* (2016).

[SBV19] C. Shi, D. M. Blei, and V. Veitch. "Adapting neural networks for the estimation of treatment effects". In: *Advances in neural information processing systems* (2019).

[Shi+06] S. Shimizu, P. O. Hoyer, A. Hyvärinen, and A. Kerminen. "A linear non-gaussian acyclic model for causal discovery". In: *Journal of Machine Learning Research* Oct (2006).

[SE17] S. M. Shortreed and A. Ertefaie. "Outcome-adaptive lasso: variable selection for causal inference". In: *Biometrics* 4 (2017).

[Spi+00] P. Spirtes, C. Glymour, R. Scheines, S. Kauffman, V. Aimale, and F. Wimberly. "Constructing bayesian network models of gene expression networks from microarray data". In: (2000).

[Wei+14] S. Weich, H. L. Pearce, P. Croft, S. Singh, I. Crome, J. Bashford, and M. Frisher. "Effect of anxiolytic and hypnotic drug prescriptions on mortality hazards: retrospective cohort study". In: *Bmj* (2014).

[Zha+20] K. Zhang, M. Gong, P. Stojanov, B. Huang, and C. Glymour. "Domain adaptation as a problem of inference on graphical models". In: *arXiv preprint arXiv:2002.03278* (2020).

[Zha+16] Q. Zhao, C. Zheng, T. Hastie, and R. Tibshirani. "Comment on causal inference using invariant prediction". In: *arXiv preprint arXiv: 1501.01332* (2016).

6 Appendix

6.1 Proofs of theorems

Lemma 3.1. *Suppose that* $\mathbb{E}[Y \mid \mathrm{Pa}(Y) = a] \neq \mathbb{E}[Y \mid \mathrm{Pa}(Y) = a']$ *whenever* $a \neq a'$. *Then a representation* Φ *is invariant across all valid environments if and only if* $\mathbb{E}[Y^e|\Phi(T^e, X^e)] = \mathbb{E}[Y|\mathrm{Pa}(Y)]$ *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 $\mathbb{E}[Y \mid \Phi(X)] = \mathbb{E}[Y \mid \mathrm{Pa}(Y) \cup Z]$ for some set Z. We proceed 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 [Pea00],

$$\mathbb{E}[Y \mid \mathrm{Pa}(Y) \setminus \{P\} \cup Z; do(P = p)] = \mathbb{E}[Y \mid \mathrm{Pa}(Y) \setminus \{P\} \cup Z, P = p]$$

and

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

The equality $\mathbb{E}[Y \mid \mathrm{Pa}(Y) \setminus \{P\} \cup Z, P = p] = \mathbb{E}[Y \mid \mathrm{Pa}(Y) \setminus \{P\} \cup Z, P = p']$ 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 $\mathbb{E}[Y \mid \mathrm{Pa}(Y) \setminus \{P\} \cup Z\,;\, do(P = p)] \neq \mathbb{E}[Y \mid \mathrm{Pa}(Y) \setminus \{P\} \cup Z\,;\, do(P = p')]$. 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 $\mathbb{E}[Y \mid \Phi(X \setminus D, D = d)] \neq \mathbb{E}[Y \mid \Phi(X \setminus D, D = d')]$. 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, $\mathbb{E}[Y^e \mid \Phi(X^e \setminus D^e, D^e = d)] = \mathbb{E}[Y^e \mid \Phi(X^e \setminus D^e, D^e = d')]$. Thus, the conditional expectations are not equal (as functions of d) in the two environments—a contradiction.

Next, we show that, Φ need not to contain the non-parent ancestors A of the outcome, because $\mathbb{E}[Y \mid \{A\} \cup \mathrm{Pa}(Y)] = \mathbb{E}[Y \mid \mathrm{Pa}(Y)]$ 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.

Theorem 3.2. *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 environments. 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 3.1, that $\mathbb{E}[Y \mid \mathrm{Pa}(Y) = a] \neq \mathbb{E}[Y \mid \mathrm{Pa}(Y) = a']$ whenever $a \neq a'$. This is without loss of generality because violations of this condition will not lead to different causal effects.

By the assumption on the loss function, the elicited invariant predictor is $\mathbb{E}[Y \mid \Phi(T, X)]$. Lemma 3.1 shows that $\mathbb{E}[Y \mid \Phi(T, X)] = \mathbb{E}[Y \mid \mathrm{Pa}(Y)]$. We further observe that the non-treatment parents of Y are sufficient to block backdoor paths. It follows the ATT can be expressed as the following.

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

Theorem 3.3. *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.*

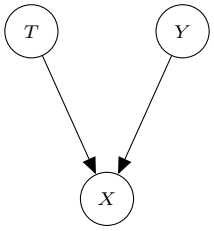


Figure 6: V-structure graph. We denote the bias induced by conditioning on X as V-bias.

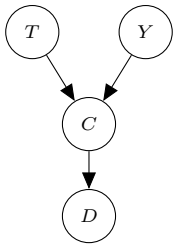


Figure 7: Y-structure graph. We denote the bias induced by conditioning on D as Y-bias.

Proof. The proof follows directly from Theorem 1 in [D'A+20]. The intuition is that the richer the covariate set is, the more likely it is to predict the treatment assignment accurately [D'A+20]. 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.

Consider the DGP with binary variables $\{X, Y, T\}$ illustrated in figure 6, where X is causally influence by Y and T.

Theorem 6.1. *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 [NDO19].

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:*

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

With the corollary above, let D denote $\Phi(T, X)$, let C denote the collider X in figure 6. The bias induced by conditioning on D is less than the bias induced by conditioning on C.

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

6.2 The Case of Mediators

In the most part of the paper, we assumed 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 [Pea00]. 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

$$\begin{aligned}\psi^e = \mathbb{E}_{M^e \mid T^e=1}[\mathbb{E}\left[Y^e \mid M^e\,;\, \mathrm{do}(T^e=1)\right] \\ - \mathbb{E}\left[Y^e \mid M^e\,;\, \mathrm{do}(T^e=0)\right] \mid T^e=1].\end{aligned} \tag{6.1}$$

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

$$\begin{aligned}\psi^e = \mathbb{E}_{M^e,W^e}[\mathbb{E}\left[Y^e \mid T^e=1, M^e, W^e\right] \\ - \mathbb{E}\left[Y^e \mid T^e=0, M^e, W^e\right] \mid T^e=1].\end{aligned} \tag{6.2}$$

Importantly, the mediators M^e and the confounders W^e show up in the same way in (6.2). 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.

6.3 Details of the experiments

Experiment 1

We simulate data with three causal graphs in Figure 3. With a slight abuse of notation, each intervention e generates a new environment e with interventional distribution $P(X^e, T^e, Y^e)$. T^e is the binary treatment and Y^e is the outcome. X^e is a 10-dimensional covariate set that differs across DGPs. $X^e = (X_1^e, X_2^e)$, where X_1^e is a five-dimensional confounder. X_2^e is either noise, a descendant, or a collider in each DGP. We examine the models' performance under two types of variations: 1) whether the observed covariates are scrambled versions of the true covariates. 2) whether the treatment effects are heteroskedastic across environments. The data generating process is illustrated below.

$$\begin{aligned}X_1^e &\leftarrow \mathcal{N}(0, e^2) \\ P^e &\leftarrow \text{sigmoid}(X_1^e \cdot w_{xt^e} + \mathcal{N}(0,1)) \\ T^e &\leftarrow Bern(P^e) \\ \tau &\leftarrow 5 + \mathcal{N}(0, \sigma^2) \\ Y^e &\leftarrow X_1^e \cdot w_{xy^e} + T^e \cdot \tau + \mathcal{N}(0, e^2)\end{aligned}$$

X_2^e equals $\mathcal{N}(0, e^2)$ in setting a), X_2^e equals $e * Y^e + \mathcal{N}(0,1)$ in setting b), and X_2^e equals $e{*}Y^e{+}T^e{+}\mathcal{N}(0,1)$ in setting c). For the four variants, in the scrambled setting $\mathcal{N}(0,\sigma^2) = \mathcal{N}(0,e^2)$, in the un-scrambled setting $\mathcal{N}(0,\sigma^2) = \mathcal{N}(0,1)$. In the environment-level heteroskedastic setting $\tau \leftarrow 5 + \mathcal{N}(0,e^2)$. In the environment-level homoscedastic setting $\tau \leftarrow 5 + \mathcal{N}(0,1)$. The performance under the four variants are illustrated in Figure 8, Figure 9, Figure 10, and Figure 11.

Experiment 2

We validate NICE for the non-linear case on the benchmark dataset SpeedDating. SpeedDating was collected to study the gender difference in mate selection [Fis+06]. 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 the treatment T and outcome Y. Specifically, it provides four modified DGPs: Mod1: parametric models; Mod2: complex models;

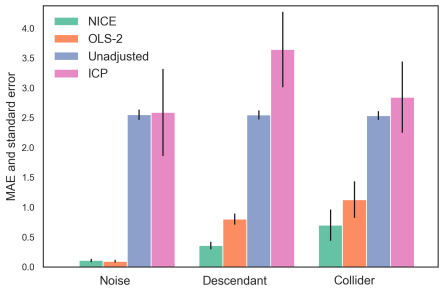


Figure 8: Models performance under the scrambled and heteroskedastic setting

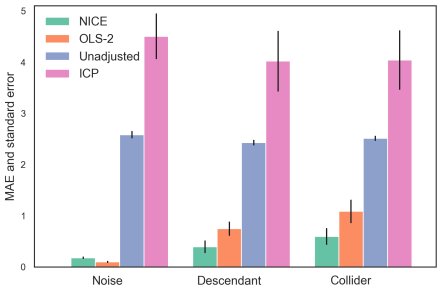


Figure 9: Models performance under the scrambled and homoscedastic setting

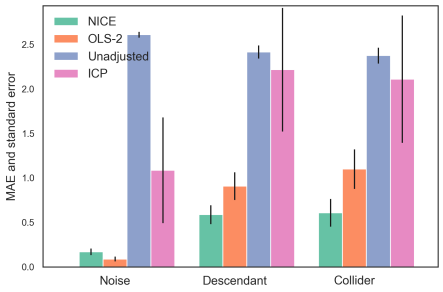


Figure 10: Models performance under the unscrambled and heteroskedastic setting

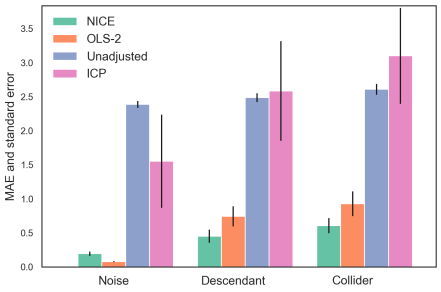


Figure 11: Models performance under the ubscrambled and homoscedastic setting

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.

We compare the estimation quality of the within-sample SATT and out-of-sample CATE over 10 bootstraps in Table 3 and Table 4. The main paper reports the models performance under the low setting. We now report results for the med and high setting.

Within-sample

	ϵ_{att}			
	MOD1	MOD2	MOD3	MOD4
med				
TARNet	$.04 \pm .02$	$.09 \pm .1$	$.23 \pm .16$	$.12 \pm .09$
+NICE	$.02 \pm .02$	$.06 \pm .03$	$.07 \pm .04$	$.02 \pm .02$
Dragon	$.12 \pm .12$	$.09 \pm .1$	$.23 \pm .16$	$.13 \pm .07$
+NICE	$.04 \pm .02$	$.06 \pm .03$	$.07 \pm .04$	$.07 \pm .08$
high				
TARNet	$.06 \pm .05$	$.18 \pm .12$	$.10 \pm .08$	$.03 \pm .03$
+NICE	$.01 \pm .01$	$.06 \pm .03$	$.05 \pm .02$	$.05 \pm .06$
Dragon	$.18 \pm .12$	$.18 \pm .12$	$.10 \pm .08$	$.05 \pm .03$
+NICE	$.02 \pm .01$	$.06 \pm .03$	$.05 \pm .02$	$.09 \pm .09$

Table 3: NICE performs well relative to the baselines if the adjustment set does not contain bad controls. The table reports MAE and bootstrap standard deviation of the SATT estimation. The model is trained and evaluated on all three environments.

To examine whether NICE helps reduce collider bias, we simulated 20 copies of a collider: $X_{co}^e = T^e + Y^e + \mathcal{N}(0, e^2)$, where $e \in \{0.01, 0.2, 1\}$ and include it in the covariate set. table 5 compares Dragonnet trained under standard empirical risk minimization framework and trained under NICE. NICE reduces collider bias across simulation setups.

Out-of-sample

	$\sqrt{\epsilon_{\text{PEHE}}}$			
	MOD1	MOD2	MOD3	MOD4
med				
TARNet	$.14 \pm .04$	$.13 \pm .03$	$.11 \pm .05$	$.09 \pm .06$
+NICE	$.06 \pm .02$	$.06 \pm .01$	$.08 \pm .05$	$.07 \pm .02$
Dragon	$.15 \pm .03$	$.13 \pm .03$	$.22 \pm .15$	$.07 \pm .02$
+NICE	$.04 \pm .01$	$.04 \pm .01$	$.08 \pm .02$	$.07 \pm .04$
high				
TARNet	$.14 \pm .06$	$.14 \pm .08$	$.13 \pm .03$	$.08 \pm .10$
+NICE	$.05 \pm .01$	$.07 \pm .01$	$.06 \pm .01$	$.07 \pm .08$
Dragon	$.14 \pm .04$	$.14 \pm .05$	$.15 \pm .04$	$.09 \pm .08$
+NICE	$.06 \pm .01$	$.06 \pm .02$	$.05 \pm .01$	$.05 \pm .06$

Table 4: NICE performs well relative to the baselines if the adjustment set does not contain bad controls. The table reports PEHE and bootstrap standard deviation of the out-of-distribution CATE estimation. The model is trained on two environments and evaluated on the third.

Within-sample

	ϵ_{att}			
	MOD1	MOD2	MOD3	MOD4
low				
Dragon	$.32 \pm .16$	$.39 \pm .14$	$.32 \pm .02$	$.50 \pm .04$
+NICE	$.11 \pm .05$	$.08 \pm .07$	$.15 \pm .04$	$.08 \pm .05$
med				
Dragon	$.39 \pm .15$	$.29 \pm .13$	$.37 \pm .13$	$.27 \pm .10$
+NICE	$.08 \pm .04$	$.17 \pm .10$	$.09 \pm .03$	$.06 \pm .04$
high				
Dragon	$.36 \pm .11$	$.35 \pm .09$	$.49 \pm .16$	$.28 \pm .06$
+NICE	$.09 \pm .06$	$.15 \pm .08$	$.10 \pm .08$	$.14 \pm .09$

Table 5: NICE reduces estimation bias in the presence of colliders. The table reports the MAE and bootstrap standard deviation of SATT. The model is trained and evaluated on three environments.

Out-of-sample

	ϵ_{pehe}			
	MOD1	MOD2	MOD3	MOD4
low				
TARNet	$.18 \pm .05$	$.42 \pm .03$	$.25 \pm .04$	$.36 \pm .12$
+NICE	$.08 \pm .02$	$.07 \pm .01$	$.08 \pm .02$	$.08 \pm .03$
Dragon	$.25 \pm .06$	$.49 \pm .05$	$.29 \pm .06$	$.45 \pm .06$
+NICE	$.09 \pm .01$	$.09 \pm .03$	$.09 \pm .03$	$.09 \pm .04$
med				
TARNet	$.41 \pm .08$	$.28 \pm .08$	$.35 \pm .06$	$.21 \pm .03$
+NICE	$.09 \pm .02$	$.08 \pm .02$	$.07 \pm .02$	$.09 \pm .01$
Dragon	$.40 \pm .08$	$.32 \pm .04$	$.47 \pm .11$	$.25 \pm .06$
+NICE	$.07 \pm .01$	$.10 \pm .05$	$.08 \pm .03$	$.07 \pm .01$
high				
TARNet	$.24 \pm .06$	$.26 \pm .10$	$.28 \pm .08$	$.25 \pm .10$
+NICE	$.06 \pm .02$	$.09 \pm .02$	$.07 \pm .02$	$.10 \pm .05$
Dragon	$.34 \pm .14$	$.36 \pm .03$	$.37 \pm .08$	$.28 \pm .07$
+NICE	$.09 \pm .03$	$.13 \pm .04$	$.07 \pm .01$	$.12 \pm .05$

Table 6: NICE reduces estimation bias in the presence of colliders. The table reports the CATE and standard deviation of CATE. The model is trained on two environments and evaluated on a third environment.

Experiment 3

In the third experiment, we consider data generated according figure 2. Notably, in the setup, we observe $P(A, X, T, Y, Z)$, where $A = \{A_1, A_t, A_y\}$ and $X = \{X_t, X_y\}$. Here A is a 50 dimensional covariate, X a 30 dimensional covariate, and Z a 50 dimensional covariate. Z is causally affected by A and Y.

We compare NICE against a neural network model similar to the structure of TARNet. The model architecture is the same as the models in the SpeedDating experiment, except the hidden layer size is 200 for the shared representation. For the exact data generating process and the detailed implementation of the models, see the associated codebase.