Invariant Representation Learning for Treatment Effect Estimation*

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

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

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) (Arjovsky et al., 2019) 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 adjusting for all the covariates.

1 INTRODUCTION

Consider the following causal inference problem.

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, we observe whether the drug was administered T_i^e, the patient's outcome Y_i^e, and their covariates X_i^e, which includes comprehensive health and socioeconomic information. The different hospitals serve different populations, so the distribution of the covariate X^e is different across the datasets. But the causal mechanism between sleeping pills T^e and lung disease Y^e remains the same across hospitals.

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 (Rosenbaum and Rubin, 1983; Pearl, 2000). To account for confounding, we try to find them among the covariates X and then adjust for them, e.g., using a method like G-computation (Robins, 1986), backdoor adjustment (Pearl, 2009), or inverse propensity score weighting (Austin, 2011). The selected covariates are called the adjustment set.

To ensure that we have adjusted for all confounding variables, we might include every covariate in the adjustment set. However, naively adjusting for all covariates runs the risk of including "bad controls" (Bhattacharya and Vogt, 2007; Pearl, 2009; Cinelli and Hazlett, 2020), variables that *induce* bias when they are adjusted for. In the example, a health condition caused by lung disease would be a bad control. It is causally affected by the outcome.

How can we exclude bad controls from the adjustment set? One approach is to select confounders through a causal graph (Pearl, 2009). We ask a domain expert to construct a causal graph or a class of equivalent graphs. We then select the confounders for the causal adjustment. However, in practice, we may have thousands of covariates in the dataset. It may be too difficult to construct a graph with thousands of nodes.

Another approach is to restrict the adjustment set to those that are known to be pre-treatment covariates (Rosenbaum, 2002; Rubin, 2009). However, this approach can lead us to include covariates that are predictive of treatment assignment but not the outcome. If the record is sufficiently rich, this information can lead to near-perfect prediction of treatment, which is a problem for causal inference. Specifically, this creates an apparent violation of overlap, the requirement

*Code is available at github.com/claudiashi57/nice.

that each unit had a non-zero probability of receiving treatment (D'Amour et al., 2020). Practically, near-violations of overlap can lead to unstable or high-variance estimates of treatment effects (Ding et al., 2017).

But these methods, and their challenges, suggest a new approach for causal estimation — we want a representation of the covariates that contains sufficient information for causal adjustment, excludes bad controls, and helps provide low-variance causal estimates. This paper presents a method to find such a representation.

Problem. We now state the problem plainly. We want to do causal inference with data collected from multiple environments, as in the hospitals' example above. The observed covariates are rich — including all the causal parents of the outcome. There are no unobserved confounders, but identifiability (Pearl, 2000) or strong ignorability (Rosenbaum and Rubin, 1983) is *not* guaranteed, due to the possible existence of bad controls. We do not know which covariates are safe to adjust for. The main question is: how can we use the multiple environments to find a representation of the covariates for valid causal estimation?

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 of the covariates.

NICE applies Invariant Risk Minimization (IRM) (Arjovsky et al., 2019) for causal adjustment. IRM is a framework for solving prediction problems. The goal is produce a predictor that is robust to changes in the deployment domain. The IRM procedure uses data from multiple environments to learn an *invariant representation* $\Phi(T, X)$, a function such that the outcome Y and the representation of the treatment and covariates $\Phi(T, X)$ have the same relationship in each environment. Predictors built on top of this representation will have the desired robustness.

The main insight that enables NICE is that the IRM invariant representation also suffices for causal adjustment. Informally, a representation is invariant if and only if it is informationally equivalent to the causal parents of the outcome Y (Arjovsky et al., 2019). For example, an invariant representation of the medical records will isolate the causal parents of lung disease. Assuming *no mediators* — variables on the causal path between the treatment and outcome — in the covariate set, the causal parents of Y constitute an adjustment set that suffices for causal adjustment, minimally impacts overlap, and that excludes all bad controls. Hence, adjusting for an invariant representation is a safe way to estimate the causal effect.[1]

Contributions. This paper develops NICE, an estimation procedure that leverages data from multiple environments to do causal inference. It articulates the theoretical conditions under which NICE provides unbiased causal estimates and evaluates the method on synthetic and semi-synthetic causal estimation problems.

2 RELATED WORK

Estimating the treatment effect from observational data consists of identification and estimation. The motivating problem is related to identification — we do not know what covariates to adjust for.

In the introduction, we discussed two widely applied adjustment approaches: selecting covariates basing on a causal graph (Pearl, 2000) and restricting to covariates that are known to be pre-treatment (Rosenbaum, 2002; Rubin, 2009). Another approach to select the adjustment set is through causal discovery.

Causal discovery methods aim to recover causal relationships or causal direction from data (Murphy, Mian, et al., 1999; Spirtes et al., 2000; Shimizu et al., 2006; Glymour et al., 2019; Shortreed and Ertefaie, 2017; Peters et al., 2016; Mooij et al., 2016; Heinze-Deml et al., 2018). In particular, NICE shares the same setup as invariance based causal discovery methods. Peters et al. (2016), Heinze-Deml et al. (2018), and Pfister et al. (2019) leverage multiple environments to find the causal predictors of the target variable in the linear, non-linear, and time series settings.

Causal discovery assumes that the observed covariates correspond to well-defined variables in the causal graph (e.g., no measurement issues). The representation learning approach of NICE does not require this assumption. Further, even in the case where this assumption holds, causal discovery methods are designed to conservatively select parents of Y. In practice, they often fail to select many actual parents. In Section 5.2, we show that while the causal discovery method (Peters et al., 2016) is better at stripping out bad controls, it also discards confounders, which leads to poor estimation quality.

With identification, we can then estimate the treatment effect. There is extensive literature on different statistical estimators (Austin, 2011; Glynn and Quinn, 2010; VanderWeele and Shpitser, 2011; Funk et al., 2011) and machine learning methods adapted for causal inference (Hill, 2011; Athey and Imbens, 2016; Beck et al., 2000; Hartford et al., 2017; Shalit et al., 2016; Louizos et al., 2017; Yoon et al., 2018; Shi et al., 2019). All these estimators and methods assume identification and focus on improving the finite sample estimation quality. In contrast, NICE considers a setting where identification is not guaranteed.

NICE uses the principle of invariance to solve a causal

[1] To keep the exposition simple, we defer the discussion of mediators to the appendix.

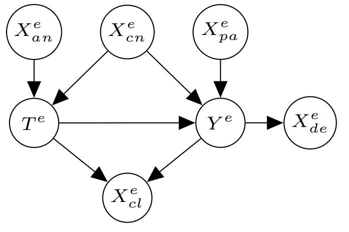


Figure 1: If the composition of X^e is unknown, the treatment effect cannot be identified. (cn=confounder, cl=collider, pa=parents, an=ancestors, de=descendants)

inference problem. A thread of related work uses the same principle to tackle different problems.

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. Conversely, if a relationship is invariant to many different perturbations, it's likely be causal (Haavelmo, 1943; Bühlmann, 2018). This principle inspired a line of causality-based domain adaptation and robust prediction work.

Rojas-Carulla et al. (2018) apply the idea for causal transfer learning, assuming the conditional distribution of the target variable given some subset of covariates is the same across domains. Magliacane et al. (2018) relax that assumption. Peters et al. (2016) and Heinze-Deml et al. (2018) apply this principle for causal variable selection from multiple environments. Zhang et al. (2020) recast the problem of domain adaptation as a problem of Bayesian inference on the graphical models. Arjovsky et al. (2019) advocate a new generalizable statistical learning principle that is based on the invariant principle. Rosenfeld et al. (2020) critically examined the generalizability of the proposed principle and its implementations.

These works focus of *robust prediction*. NICE focuses *causal estimation*. NICE is complementary as it studies the idea of applying domain adaptation methods for causal estimation. In particular, we focus on the application of IRM for treatment effect estimation.

3 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 other variables X^e, called covariates. Assume 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, i.e., it contains a set of variables sufficient for adjustment. 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 an example graph that 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., $\{X_i^e, T_i^e, Y_i^e\} \overset{\text{iid}}{\sim} P^e$. 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)[2] in each environment,

$$\begin{aligned}\psi^e \triangleq \mathbb{E}\left[Y^e \,|\, \mathrm{do}(T^e=1), T^e=1\right] \\ -\mathbb{E}\left[Y^e \,|\, \mathrm{do}(T^e=0), T^e=1\right].\end{aligned} \tag{3.1}$$

The use of do notation (Pearl, 2000) indicates that the estimand is causal. 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.

3.1 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 (Rosenbaum and Rubin, 1983). Let $Z(X)$ be an *admissible* subset of X—it contains no descendants of Y and blocks all "backdoor paths" between Y and T (Pearl and Paz, 2014). 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,

$$\begin{aligned}\psi = \mathbb{E}_X[\mathbb{E}_Y\left[Y \,|\, T=1, Z(X)\right] \\ - \mathbb{E}_Y\left[Y \,|\, T=0, Z(X)\right] \,|\, T=1].\end{aligned} \tag{3.2}$$

We estimate ψ in two stages. First, we fit a model $\hat{Q}$ for the conditional expectation $Q(T, Z(X)) = \mathbb{E}_Y\left[Y \,|\, T, Z(X)\right]$. Second, we use Monte Carlo to approximate the expectation over X,

$$\hat{\psi} = \frac{1}{\sum_i t_i} \sum_{i:t_i=1} \left(\hat{Q}(1, Z(X_i)) - \hat{Q}(0, Z(X_i))\right), \tag{3.3}$$

[2] For simple exposition, we focus on the ATT estimation. The method can also be applied to conditional average treatment effect or average treatment estimations.

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 (3.2) 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.[3] Given the representation, we estimate the conditional expectations $\mathbb{E}_Y [Y \mid \Phi(T, X)]$ and proceed to estimate the causal effect.

3.2 INVARIANT RISK MINIMIZATION

To learn an admissible representation, we use IRM. IRM is a framework for learning predictors that perform well across many environments. We first 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 3.1 (Valid environment Arjovsky et al., 2019)**.** 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. The intervention can be either atomic or stochastic. An intervention is valid with respect to $(\mathcal{G}, P)$ if (i) $\mathbb{E}_{P^e} [Y^e|Pa(Y)] = \mathbb{E}_P [Y|Pa(Y)]$, 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 3.2 (Invariant representation)**.** A representation $\Phi(T, X)$ is invariant with respect to environments $\mathcal{E}$ if and only if $\mathbb{E} [Y^{e_1}|\Phi(T^{e_1}, X^{e_1}) = \pi] = \mathbb{E} [Y^{e_2}|\Phi(T^{e_2}, X^{e_2}) = \pi]$ for all $e_1, e_2 \in \mathcal{E}$.

Arjovsky et al. (2019) 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 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 3.3 (Invariant representation via predictor Arjovsky et al., 2019)**.** 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,

$$w \in \underset{\bar{w}:\mathcal{H}\to\mathcal{Y}}{\arg\min}\, R^e(\bar{w} \circ \Phi) \quad \text{for all } e \in \mathcal{E}, \tag{3.4}$$

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

The invariant representations in Definitions 3.2 and 3.3 align if we choose a loss function for which the minimizer of the associated risk in (3.4) is a conditional expectation. (Examples include squared loss and cross entropy loss.) 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 (3.4) for both w and Φ.

However, the general formulation of (3.4) is computationally intractable, Arjovsky et al. (2019) introduce IRMv1 as a practical alternative.

Definition 3.4 (IRMv1Arjovsky et al., 2019)**.** IRMv1 is:

$$\hat{\Phi} = \underset{\Phi}{\arg\min} \sum_{e\in\mathcal{E}} R^e(1.0{\cdot}\Phi) + \lambda \parallel \nabla_{w|w=1.0} R^e(w{\cdot}\Phi) \parallel^2 . \tag{3.5}$$

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.[4]

In practice, we parameterize Φ with a neural network that takes $\{t_i^e, x_i^e\}$ as input and outputs a real number. Let ℓ be a loss function, such as squared error or cross entropy, and 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:

$$\hat{R}^e(Q) = \frac{1}{n_e} \sum_i \ell(y_i^e, Q(t_i^e, x_i^e)). \tag{3.6}$$

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

[3] An admissible representation is analogous to an 'admissible set' (Pearl, 2000), which is a valid adjustment set.

[4] For details on IRMv1, see (Arjovsky et al., 2019, section 3.1)

3.3 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 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 $\mathrm{Pa}(Y)$ 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, as just explained, if $\Phi(T, X)$ is invariant over all valid environments, then $\mathbb{E}\left[Y|T, \mathrm{Pa}(Y) \setminus \{T\}\right] = \mathbb{E}\left[Y|\Phi(T, X)\right]$. Second, $\mathrm{Pa}(Y) \setminus \{T\}$ suffices for causal adjustment. That is, $\mathrm{Pa}(Y) \setminus \{T\}$ blocks any backdoor paths and does not include bad controls. Following (3.2),

$$\begin{aligned} \psi = \mathbb{E}[&\mathbb{E}\left[Y \mid T=1, \mathrm{Pa}(Y) \setminus \{T\}\right] \\ &- \mathbb{E}\left[Y \mid T=0, \mathrm{Pa}(Y) \setminus \{T\})\right] \mid T=1] \end{aligned} \tag{3.7}$$

Since $\mathbb{E}\left[Y|T, \mathrm{Pa}(Y) \setminus \{T\}\right] = \mathbb{E}\left[Y|\Phi(T, X)\right]$,

$$\psi = \mathbb{E}\left[\mathbb{E}\left[Y \mid \Phi(1, X)\right] - \mathbb{E}\left[Y \mid \Phi(0, X)\right] \mid T=1\right]. \tag{3.8}$$

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

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

Similar to the function $\hat{Q}$ in (3.2), $\hat{Q}^{\mathrm{inv}}$ can come from any prediction model that uses an invariant objective. In Section 5, we use linear regression, TARNet (Shalit et al., 2016), and Dragonnet (Shi et al., 2019).

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 valid environments.

4 JUSTIFICATION OF NICE

We now establish the validity of NICE as a causal estimation procedure. All proofs are in 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 valid environments. We prove that conditioning on a fully invariant representation is the same as conditioning on the parents of Y.

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.*

Lemma 4.1 helps show that a representation that elicits an invariant predictor suffices for adjustment.

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 environments. Assuming* X^e *does not contain mediators between the treatment and the outcome, then* $\psi^e = \mathbb{E}\left[Q^{\mathrm{inv}}(1, X^e) - Q^{\mathrm{inv}}(0, X^e)|T^e = 1\right]$.

Theorem 4.2 shows that the NICE estimand is equal to the ATT as long as the predictor Q^{inv} is invariant across all valid environments.

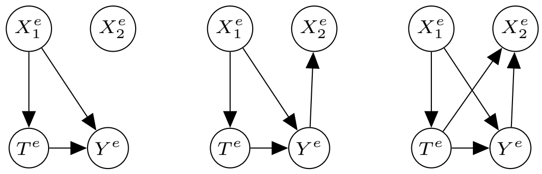
In practice, if a predictor is invariant across a limited set of diverse environments, it may generalize to all valid environments. Assuming a linear data generating process, Arjovsky et al. (2019) establish sufficient conditions on the number and diversity of the training environments such that the learned representation generalizes to all valid environments. In the non-linear case, there are no known sufficiency results. However, Arjovsky et al. (2019) give empirical evidence that access to even a few environments may suffice.[5]

In addition to identifiability, non-parametric estimation of treatment effects with finite data, i.e., (3.3), requires 'positivity' or 'overlap' – both treatment and non-treatment have a non-zero probability for all levels of the confounders (Rosenbaum and Rubin, 1983; Imbens, 2004). 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 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.*

The intuition is that the richer the covariate set is, the more likely it is to accurately predict the treatment assignment

[5] Establishing the sufficiency result of IRM is an open question.



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

Figure 2: We observe $\{X_1, X_2\}$, but do not know its composition. In (a) $\{X_1, X_2\}$ is a valid adjustment. In (b) and (c), X_2 is downstream of Y, so $\{X_1, X_2\}$ is not a valid adjustment.

(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.

Even when invariance across all valid environments is not guaranteed, NICE may still improve the estimation quality when there are possible colliders in the adjustment set. If the observed environments are induced by valid interventions, either atomic or stochastic, on the bad controls, an invariant representation over these environments can also exclude bad controls. Even when the representation does not exclude the bad controls, invariance may remove at least some (if not all) collider dependence. Intuitively, conditioning on a subset of collider information should reduce bias in the resulting estimate. Theorem 7.1 in the appendix shows 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 the observed covariate set X does not contain mediators between T and Y. What happens to the interpretation of the learned parameter, $\hat{\psi}^e$, 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 (Pearl, 2000). If there are mediators and confounders, we define the parameters as the natural direct effect on the treated (NDET). The mathematics definitions are in the appendix.

5 EMPIRICAL STUDIES

We study the performance of NICE with three experiments. We are interested in three empirical questions: (1) Does NICE strip out bad controls in practice? (2) Is NICE "costless" when there are no bad controls? (3) What is the effect of different amounts of environmental variation on NICE's performance.

We find that (1) when there are bad controls in the adjustment set, NICE can reduce bias induced by the bad controls. (2) When there are no bad controls in the adjustment set, NICE does not hurt the estimation quality. (3) Whether NICE can strip out bad controls depends on the diversity of the environments. The more diverse the environments, the more likely it is that NICE can strip out the bad controls.

5.1 EXPERIMENTAL SETUP

We construct three experiments corresponding to different settings. We first consider the setting where NICE is theoretically guaranteed to strip out bad controls. In Section 5.2, the data are collected from diverse environments, and the DGPs are linear. In the non-linear setting, there are no known sufficiency results for the generalizability of IRM. Therefore, there is no theoretical guarantee that NICE can strip out bad controls.

To study whether NICE can reduce bias from bad controls empirically, we validate NICE using non-linear semi-synthetic benchmark datasets in Section 5.3. Furthermore, we study the effect of different amounts of environmental variation on NICE's performance in Section 5.4.

Causal Estimands & 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), $\tau(x_i) = Q(1, Z(x_i)) - Q(0, Z(x_i))$ (Imbens, 2004). 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) $\epsilon_{\text{PEHE}} = \frac{1}{n} \sum_0^n (\hat{\tau}(x_i) - \tau(x_i))^2$ (Hill, 2011). PEHE reflects the ability to capture individual variation in treatment effects. The main paper shows the MAE of the SATT averaged across environments. For the evaluation of CATE, see the appendix.

Predictor Choices. Under the NICE procedure, the invariant predictor $\hat{Q}^{inv}$ can be any class of predictor trained with an IRMv1 objective. In the linear settings, we use OLS-2 as the predictor. OLS-2 is linear regression with two separate regressors for the treated and the control population.

In the nonlinear settings, we consider two neural network models similar to the structure of TARNet (Shalit et al., 2016) and Dragonnet (Shi et al., 2019). TARNet is a two-headed model with a shared representation $Z(X) \in R^p$, 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 (Kingma and Ba, 2014) as the optimizer, set the learning rate as 0.001, and an l2 regularization rate of 0.0001. For Dragonnet, there's an additional treatment head, which makes treatment prediction

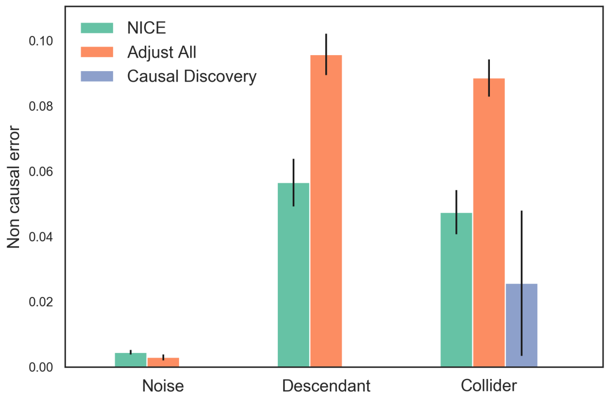


Figure 3: NICE strips out bad controls, which leads to better downstream treatment effect estimation. ICP (causal discovery) strips out bad controls, but also useful confounders (see figure 4). The non-causal error is measured by the mean square error of the weights on X_2. Lower is better.

from the shared representation. For the hyper-parameter λ used in IRMv1, we use $\lambda = 10$ in the linear settings and $\lambda = 100$ in the non-linear settings.

Since we use the same predictor across different DGPs, the hyper-parameters are chosen arbitrarily. We use data from all environments to train and evaluate the predictor. The main paper presents results using TARnet. Results derived from Dragonnet are in the appendix.

Adjustment Schemes. We compare estimation quality produced by the following adjustment schemes: (1) adjusting for all covariates, (2) NICE, and (3) causal variable selection. Under (1), we pool the data across environments to fit a predictor $\hat{Q}$ and compute SATT using (3.2). Under (3), we first use Invariant Causal Prediction (ICP) (Peters et al., 2016) to select an adjustment set. ICP is a variable selection method that identifies the target variable's causal parents by leveraging data from multiple environments. We then pool data across environments, use the adjustment set to fit a predictor and compute SATT using (3.2). The estimation procedure of NICE is described in Section 3.3.

5.2 NICE IN LINEAR SETTINGS

We simulate data with the three causal graphs in Figure 2. 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. The DGPs are:

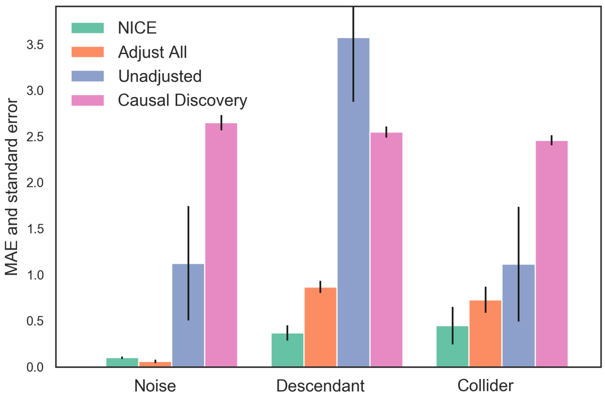


Figure 4: NICE reduces bias when the adjustment set contains bad controls and does not hurt if the adjustment set is valid. We use ICP for the causal discovery method, which is often too conservative. When ICP returns an empty set, estimated causal effect is zero. The figure reports average MAE and standard error of the SATT over 10 simulations.

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

In (a) $X_2^e \leftarrow \mathcal{N}(0,1)$, in (b) $X_2^e \leftarrow e * Y^e + \mathcal{N}(0,1)$, and in (c) $X_2^e \leftarrow e * Y^e + T^e + \mathcal{N}(0,1)$.

For evaluation, following (Arjovsky et al., 2019), 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 consider two types of variations: (1) whether 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) whether the treatment effects are heteroskedastic across environments. In the heteroskedastic setting $\tau \leftarrow 5 + \mathcal{N}(0, e^2)$. In the environment-level homoskedastic setting $\tau \leftarrow 5 + \mathcal{N}(0, 1)$.

We compare the estimation quality produced by four different adjustment approaches: (1) adjusting for all covariates, (2) causal variable selection, (3) NICE, and (4) No adjustment. The results in Figure 4 and Figure 3 are under the unscrambled and heteroskedastic variant. The results of the other variants are in the appendix.

Analysis. Figure 4 reports the average of the MAE of SATT estimates over all three environments. We observe that when the covariate set does not include bad controls—simulation setting (a)— NICE performs as well as adjusting for all covariates. When the covariate set includes bad controls that are closely related to the outcome, that is (b) and (c), NICE can help reduce the estimation bias.

To understand why NICE reduces the estimation bias, we look at the weights of the control predictor. Ideally, the weights that correspond to the bad controls should be 0. As shown in Figure 3, the predictor trained an IRMv1 objective places less weight on the bad controls than the predictor using empirical risk minimization. We observe ICP successfully strips out most of the bad controls. However, it produces worse causal estimates as it also strips out confounders. 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. A similar observation is made in Zhao et al. (2016), where slight perturbations on ICP's assumptions might lead to poor performance.

5.3 NICE IN NON-LINEAR SETTINGS

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.

Table 1: If the adjustment set is valid, NICE does not hurt the estimation performance. The table reports average MAE and bootstrap standard deviations of the SATT estimation.

Valid Adjustment		ϵ_{att}			
		Mod1	Mod2	Mod3	Mod4
Low	Adjust All	$.04 \pm .08$	$.05 \pm .09$	$.07 \pm .09$	$.01 \pm .01$
	NICE	$.07 \pm .03$	$.02 \pm .01$	$.09 \pm .03$	$.04 \pm .02$
Med	Adjust All	$.07 \pm .10$	$.05 \pm .05$	$.04 \pm .04$	$.07 \pm .08$
	NICE	$.05 \pm .02$	$.04 \pm .03$	$.05 \pm .03$	$.03 \pm .02$
High	Adjust All	$.07 \pm .07$	$.06 \pm .05$	$.06 \pm .07$	$.04 \pm .04$
	NICE	$.02 \pm .01$	$.06 \pm .03$	$.04 \pm .02$	$.07 \pm .04$

The ACIC simulations are designed to assess the estimation quality of predictors and estimators. They 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 x and divide them into three equal sized environments. For each DGP, we draw 10 bootstrap samples. To simulate bad controls, we included 20 copies of a collider in the adjustment set:

$X_{co}^{e} = T^{e} + Y^{e} + \mathcal{N}(0, e^2)$, where $e \in \{0.01, 0.2, 1\}$.

Table 2: NICE reduces estimation bias in the presence of bad controls. The table reports the average MAE and bootstrap standard deviation of SATT.

Bad Controls in Adjustment Set		ϵ_{att}			
		Mod1	Mod2	Mod3	Mod4
low	Adjust All	$.26 \pm .09$	$.42 \pm .03$	$.34 \pm .08$	$.46 \pm .09$
	NICE	$.09 \pm .07$	$.03 \pm .01$	$.11 \pm .04$	$.08 \pm .04$
med	Adjust All	$.38 \pm .10$	$.35 \pm .06$	$.40 \pm .17$	$.3 \pm .09$
	NICE	$.06 \pm .03$	$.06 \pm .03$	$.06 \pm .02$	$.03 \pm .03$
high	Adjust All	$.32 \pm .14$	$.38 \pm .09$	$.42 \pm .05$	$.28 \pm .05$
	NICE	$.05 \pm .03$	$.11 \pm .03$	$.16 \pm .05$	$.11 \pm .05$

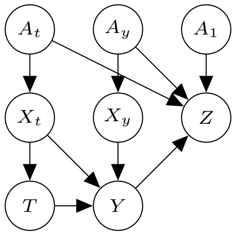


Figure 5: The DGP for Section 5.4. The adjustment set {X, A} is valid. Adjustment set {X, A, Z} is not valid.

Analysis. We compare two adjustment schemes: adjusting for all covariates and NICE. We first consider the setting where there are no bad controls. Table 1 reports the average SATT MAE and standard deviations over 10 bootstraps under two adjustment schemes. We observe that NICE does not hurt the estimation quality in comparison to adjusting for all covariates.

We also consider the setting where there is a strong collider. 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 potential reason is that the predictor is not optimal.

5.4 THE EFFECT OF ENVIRONMENT VARIATIONS ON NICE'S PERFORMANCE

In this experiment, we examine the effect of environment variations on NICE's performance. We simulate non-linear data using the causal graph illustrated in Figure 5. The details of the data simulation are in the appendix.

We first draw three source environments $\{P^{e_1}, P^{e_2}, P^{e_3}\}$ that are diverse. To control the level of environment variation, we construct three new environments $\{P^{e_1'}, P^{e_2'}, P^{e_3'}\}$ that are mixtures of the three source environments. Respectively, $P^{e_1'}, P^{e_2'}, P^{e_3'}$ draw (p_1, p_2, p_3) proportions from P^{e_1}, (p_2, p_3, p_1) proportions from P^{e_2}, and (p_2, p_3, p_1) proportions from P^{e_3}. The proportions (p_1, p_2, p_3) sum to one.

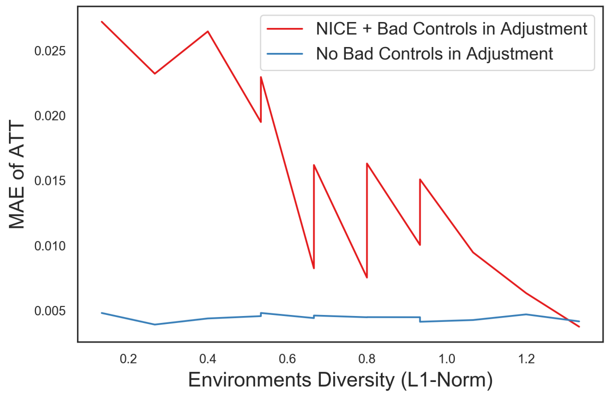


Figure 6: NICE mitigates bad controls more with access to more diverse environments. The x-axis is the environmental diversity. The y-axis is the average MAE of the SATT.

We approximate the diversity of the environments by the diversity of the proportions. The diversity measure is: $\frac{1}{3}\sum_{ij}|p_i - p_j|$. We consider 14 set of new environments, induced by different combination of the mixture probabilities. We compare the estimation quality of NICE when given a covariate set that include bad controls $\{X, A, Z\}$ against adjusting for a valid covariate set $\{X, A\}$.

As shown in Figure 6, the more diverse the environments, the more likely that NICE can strip out bad controls and reduce bias. When environments are sufficiently diverse, the learned representation is equivalent to a valid adjustment set.

6 DISCUSSION

NICE lives at the intersection of representation learning and causal inference, demonstrating how representation learning ideas can be harnessed to improve causal estimation. Here we have examined the causal setup where it's unknown which covariates are safe to adjust for. One important direction for future work is to expand this setting to one where we combine partial causal knowledge with representation learning for estimating effects in more general scenarios.

Acknowledgements

This work is funded by ONR N00014-17-1-2131, ONR N00014-15-1-2209, NIH 1U01MH115727-01, NSF CCF-1740833, DARPA SD2 FA8750-18-C-0130, Amazon, and Simons Foundation. We thank Ryan Carey, Gemma Moran, Kevin Murphy, Dhanya Sridhar, and anonymous reviewers for helpful comments and feedback.

References

Arjovsky, M., L. Bottou, I. Gulrajani, and D. Lopez-Paz (2019). "'Invariant Risk Minimization.'" In: *arXiv preprint arXiv:1907.02893*.

Athey, S. and G. Imbens (2016). "'Recursive partitioning for heterogeneous causal effects.'" In: *Proceedings of the National Academy of Sciences* 27.

Austin, P. C. (2011). "'An introduction to propensity score methods for reducing the effects of confounding in observational studies.'" In: *Multivariate behavioral research* 3.

Beck, N., G. King, and L. Zeng (2000). "'Improving quantitative studies of international conflict: A conjecture.'" In: *American Political science review*.

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

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

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

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

Ding, P., T. VanderWeele, and J. M. Robins (2017). "'Instrumental variables as bias amplifiers with general outcome and confounding.'" In: *Biometrika* 2.

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

Funk, M. J., D. Westreich, C. Wiesen, T. Stürmer, M. A. Brookhart, and M. Davidian (2011). "'Doubly robust estimation of causal effects.'" In: *American journal of epidemiology* 7.

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

Glynn, A. N. and K. M. Quinn (2010). "'An introduction to the augmented inverse propensity weighted estimator.'" In: *Political analysis*.

Haavelmo, T. (1943). "'The statistical implications of a system of simultaneous equations.'" In: *Econometrica, Journal of the Econometric Society*.

Hartford, J., G. Lewis, K. Leyton-Brown, and M. Taddy (2017). "'Deep IV: A flexible approach for counterfactual prediction.'" In: *International Conference on Machine Learning*. PMLR.

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

Hill, J. (2011). "'Bayesian Nonparametric Modeling for Causal Inference.'" In: *Journal of Computational and Graphical Statistics*.

Imbens, G. W. (2004). "'Nonparametric estimation of average treatment effects under exogeneity: A review.'" In: *Review of Economics and statistics* 1.

Kingma, D. P. and J. Ba (2014). "'Adam: A method for stochastic optimization.'" In: *arXiv preprint arXiv:1412.6980*.

Louizos, C., U. Shalit, J. M. Mooij, D. Sontag, R. Zemel, and M. Welling (2017). "'Causal effect inference with deep latent-variable models.'" In: *NEURIPS*.

Magliacane, S., T. van Ommen, T. Claassen, S. Bongers, P. Versteeg, and J. M. Mooij (2018). "'Domain adaptation by using causal inference to predict invariant conditional distributions.'" In: *Advances in Neural Information Processing Systems*.

Mooij, J. M., J. Peters, D. Janzing, J. Zscheischler, and B. Schölkopf (2016). "'Distinguishing cause from effect using observational data: methods and benchmarks.'" In: *The Journal of Machine Learning Research* 1.

Murphy, K., S. Mian, et al. (1999). *Modelling gene expression data using dynamic Bayesian networks*. Tech. rep. Citeseer.

Pearl, J. (2000). *Causality: models, reasoning and inference*.

– (2009). *Causality*.

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

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

Pfister, N., P. Bühlmann, and J. Peters (2019). "'Invariant causal prediction for sequential data.'" In: *Journal of the American Statistical Association* 527.

Robins, J. (1986). "'A new approach to causal inference in mortality studies with a sustained exposure period–application to control of the healthy worker survivor effect.'" In: *Mathematical modelling* 9-12.

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

Rosenbaum, P. R. (2002). "'Overt bias in observational studies.'" In: *Observational studies*.

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

Rosenfeld, E., P. Ravikumar, and A. Risteski (2020). "'The Risks of Invariant Risk Minimization.'" In: *arXiv preprint arXiv:2010.05761*.

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

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

Shi, C., D. M. Blei, and V. Veitch (2019). "'Adapting Neural Networks for the Estimation of Treatment Effects.'" In: *Advances in neural information processing systems*.

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

Shortreed, S. M. and A. Ertefaie (2017). "'Outcome-adaptive lasso: Variable selection for causal inference.'" In: *Biometrics* 4.

Spirtes, P., C. Glymour, R. Scheines, S. Kauffman, V. Aimale, and F. Wimberly (2000). "'Constructing Bayesian network models of gene expression networks from microarray data.'" In.

VanderWeele, T. J. and I. Shpitser (2011). "'A new criterion for confounder selection.'" In: *Biometrics* 4.

Yoon, J., J. Jordon, and M. van der Schaar (2018). "'GANITE: Estimation of individualized treatment effects using generative adversarial nets.'" In.

Zhang, K., M. Gong, P. Stojanov, B. Huang, and C. Glymour (2020). "'Domain Adaptation As a Problem of Inference on Graphical Models.'" In: *arXiv preprint arXiv:2002.03278*.

Zhao, Q., C. Zheng, T. Hastie, and R. Tibshirani (2016). "'Comment on Causal inference using invariant prediction.'" In: *arXiv preprint arXiv: 1501.01332*.